Generalized Partially Linear Single-Index Models

R. J. CARROLL, Jianqing FAN, Irène GIJBELS, and M. P. WAND

The typical generalized linear model for a regression of a response Y on predictors $(\mathbf{X}, \mathbf{Z})$ has conditional mean function based on 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. We call these *generalized partially linear single-index models* (GPLSIM). The models include the "single-index" models, which have $\boldsymbol{\beta}_0 = 0$. Using local linear methods, we propose estimates of the unknown parameters $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ and the unknown function $\eta_0(\cdot)$ and obtain their asymptotic distributions. Examples illustrate the models and the proposed estimation methodology.

KEY WORDS: Asymptotic theory; Generalized linear models; Kernel regression; Local estimation; Local polynomial regression; Nonparametric regression; Quasi-likelihood.

1. INTRODUCTION

1.1 Motivation

The Framingham Heart Study (Kannel et al. 1986) comprises a series of exams taken 2 years apart. For the purpose of illustration, we use Exam #3 as the baseline. The dataset includes 1,615 men age 31–65, with the outcome indicating the occurrence of coronary heart disease (CHD) within an 8-year period following Exam #3; there were 128 such cases of CHD. Predictors used in this example are patient's age, smoking status, and serum cholesterol level, 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 $\mathbf{X}$, consisting of the three variables X_1 (age of patient), X_2 $(= \log(\text{SPB} - 25))$, and X_3 $(= \log(\text{cholesterol level}))$. An ordinary logistic regression model says that the logit of CHD probabilities satisfies

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

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

As we discuss in Section 3.2, some curvature is not captured by this linear-logistic model. This article is concerned with simple semiparametric alternatives to the fully parametric model (1) that 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 $\alpha_0^T\mathbf{X}$ 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

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

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 quasi-likelihood models, describe a way to fit such models, and derive an asymptotic theory that allows inference about the parameters (α_0, β_0). In the rest of this section, we describe the general class of models of interest to us here, which we call *generalized partially linear single-index models* (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. In particular, we present a class of asymptotically optimal estimators of the unknown parameters.

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. 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

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

for known functions $\mathcal{B}$ and $\mathcal{C}$. In parametric generalized linear models, the unknown regression function $\mu(\mathbf{x}, \mathbf{z}) = E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = \mathcal{B}'\{\theta(\mathbf{x}, \mathbf{z})\}$ is modeled linearly via a link function g by

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

R. J. Carroll is Professor of Statistics, Nutrition and Toxicology, Department of Statistics, Texas A&M University, College Station, TX 77843. Jianqing Fan is Associate Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599. Irène Gijbels is Associate 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 National Cancer Institute grant CA-57030. Fan's research was supported by National Science Foundation (NSF) grant DMS-9203135 and an NSF postdoctoral fellowship. Gijbels' research was supported by the Programme d'Action de Recherche Concertée, No. 93-98-164. The authors thank the referee and associate editor for many helpful comments that significantly improved the article.

If $g = (\mathcal{B}')^{-1}$ (the inverse function of $\mathcal{B}'$), then g is the canonical link function (see McCullagh and 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 GPLSIM,

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

The restriction $\|\boldsymbol{\alpha}_0\| = 1$ is required for identifiability.

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 and Welsh (1994). In other contexts, when only the mean function is specified, this problem is known as the nonparametric single-index model (Härdle, Hall, and 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 (albeit at the cost of some loss in flexibility). Other recent work on estimation in the framework of single-index models was done by Bonneu, Delecroix, and Hristache (1995).

The meaning of the single-index parameter $\boldsymbol{\alpha}_0$ deserves a short explanation. Here we basically follow the lead of Li (1991), who noted three points:

a. Clearly, as a practical matter, lowering dimensionality before fitting data is important (Li's remark 1.2 goes even further and suggests that in many cases this is the crucial step), and the appeal of single-index models is that they provide a readily interpretable means of performing this reduction.
b. If $\eta_0(\cdot)$ is monotone, then $\boldsymbol{\alpha}$ takes on the same general meaning as "effect" parameters as would occur in ordinary linear models.
c. Given an estimated "direction" $\boldsymbol{\alpha}_0$, model criticism becomes a more manageable proposition.

Severini and Staniswalis (1994) considered model (5) but with $\eta_0(\boldsymbol{\alpha}_0^T \mathbf{x})$ replaced by $\gamma(\mathbf{x})$, a p-variate function. Hunsberger (1994) considered model (5) but with $\mathbf{X}$ scalar, so that $p = 1$ and $\boldsymbol{\alpha}_0 = 1$. In this case model (5) becomes

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

Model (6) is particularly popular in the spline literature. (See, e.g., Chen 1988, Cuzick 1992, Heckman 1986, Speckman 1988, and Wahba 1984, where it is called the partial spline model or the partially linear model.) Recently, Mammen and van de Geer (1995) studied penalized quasi-likelihood estimation in partially linear models.

A different approach to modeling (and coping with the "curse of dimensionality") is through generalized additive models (GAM's) (see Hastie and 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 GPLSIM (5) have the obvious advantage of being more parsimonious, although they are clearly more difficult to compute given the existence of commercial software for GAM's. We have in our own work combined the two to fit models of the form (5), with an estimator $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}_0$ obtained using our techniques and then GAM applied to $\mathbf{Z}$ and $\hat{\boldsymbol{\alpha}}^T \mathbf{X}$. 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 use parsimonious parametric models (see Royston and Altman 1994, and the discussions therein). GPLSIM fall somewhere between the fully parametric flexible models of Royston and Altman (1994) and the almost fully nonparametric models of Hastie and Tibshirani (1990).

1.3 Aim and Outline

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 methods previously cited in that we use local linear rather than simple kernel regression methods. Our aim 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 model and the partially linear model. Our work also applies to quasi-likelihood 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 log-likelihood $\ln f_{Y|\mathbf{X},\mathbf{Z}}(y|\mathbf{x}, \mathbf{z})$ by a *quasi-likelihood function* $Q\{\mu(\mathbf{x}, \mathbf{z}), y\}$. If the conditional variance is modeled as $\text{var}(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = \sigma^2 V\{\mu(\mathbf{x}, \mathbf{z})\}$ for some known positive function V, then the corresponding quasi-likelihood function $Q(w, y)$ satisfies

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

(McCullagh and Nelder 1989, chap. 9). The quasi-score (7) possesses properties similar to those of the usual likelihood score function.

In Section 2 we propose estimation procedures, and in Section 3 we illustrate their performance via simulation and examples. In Sections 4 and 5 we describe distribution theory. In Section 6 we present the result showing asymptotic efficiency of the parametric estimators (in the semiparametric sense). In Section 7 we provide methods for estimating the standard errors of the parametric and nonparametric parts of the model. The usual method for estimating standard errors is to derive a formula for the asymptotic covariance matrix, and then plug into this formula to obtain an estimated covariance matrix. Unfortunately, as a general principle this has the drawback that the formula for the asymptotic covariance matrix requires additional nonparametric regression. We derive consistent covariance matrix

estimates that avoid these additional nonparametric regressions. We give some implementation details in Sections 3.2 and 8, and discuss the issue of incorporating interactions in the model in Section 9. Proofs are given in the Appendix.

2. MAXIMUM QUASI-LIKELIHOOD

2.1 The Estimation Method

Under model (5), the primary interest is to estimate $\boldsymbol{\alpha}_0$, $\boldsymbol{\beta}_0$, and $\eta_0(\cdot)$. Because $\eta_0(\cdot)$ is modeled nonparametrically, it is natural to consider *local* quasi-likelihood. However, efficient estimation of the global parameters $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ requires using all data points and hence should rely on the *global* quasi-likelihood. In local quasi-likelihood, we 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 let $K_h(t) = K(t/h)/h$ be a rescaling of K. The function K is usually called a kernel function, and the parameter h is called the bandwidth. For $i = 1, \ldots, n$, a sample $(Y_i, \mathbf{X}_i, \mathbf{Z}_i)$ is observed. The local quasi-likelihood is really a weighted quasi-likelihood, with weights $K_h(\boldsymbol{\alpha}^T\mathbf{X}_i - u)$.

The estimation procedure for estimating $\boldsymbol{\alpha}_0$, $\boldsymbol{\beta}_0$ and $\eta_0(\cdot)$ is as follows:

Step 0 (Initialization step). Fit a parametric generalized linear model to obtain initial values $(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\beta}})$, and set $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_1/\|\hat{\boldsymbol{\alpha}}_1\|$.

Step 1. Find $\hat{\eta}(u; h, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \hat{a}$ by maximizing the local quasi-likelihood

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

with respect to a and b. 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 $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ by maximizing

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

with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

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 $\hat{\eta}(u; h, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \hat{a}$, where $(\hat{a}, \hat{b})$ is obtained by maximizing (8). 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.

The basic idea behind the foregoing algorithm is simple: estimate $\eta_0(\cdot)$ locally via (8), and then use all of the data and (9) to estimate $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, with $\hat{\eta}(\cdot)$ replacing $\eta_0(\cdot)$. We briefly discuss an alternative estimator in Section 4.1. We recommend calculating $\hat{\eta}(\cdot; h, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ at a *fixed* but fine grid of points and using linear interpolation to calculate the other values of $\hat{\eta}(\cdot; h, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ when needed.

The estimation procedure involves choosing a smoothing parameter on two quite different levels. In Steps 1 and 2 of the algorithm the aim is estimation of the parametric part $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, and hence here the bandwidth h should be optimal for this task. In Step 4, however, the goal is to estimate the nonparametric part $\eta_0(\cdot)$, and hence the bandwidth h should be optimal in this respect.

Finally, we mention that following work of Severini and Staniswalis (1994), maximizing

$$\sum_{i=1}^{n} Q[g^{-1}\{\hat{\eta}(\boldsymbol{\alpha}^T\mathbf{X}_i; h, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \boldsymbol{\beta}^T\mathbf{Z}_i\}, Y_i] \tag{10}$$

instead of (9) leads to estimates that are asymptotically equivalent to those resulting from the foregoing algorithm. We make use of this fact later, but for brevity we do not provide the calculations. The statement is true when working with the function Q as in (7), but it does not hold for completely arbitrary functions Q.

2.2 Alternatives

The algorithm suggested here uses local linear weighted fits based on kernel weights with a fixed global bandwidth. One may replace these by more sophisticated smoothers, such as those using higher-degree polynomials, locally varying bandwidths, nearest neighbor weights, and so on. Other nonkernel smoothers, such as splines, also may be used.

3. NUMERICAL EXAMPLES

3.1 Simulation

We ran a small simulation study with $n = 200$ and data generated according to the "sine-bump" model

$$Y_i = \sin\{\pi(\alpha^T X_i - A)/(B - A)\} + \beta Z_i + \varepsilon_i,$$

where the X_i are trivariate with independent uniform (0, 1) components, $Z_i = 0$ for i odd and $Z_i = 1$ for i even, and the ε_i are normally distributed with mean 0 and variance .01. The parameters were $\alpha = (1, 1, 1)/\sqrt{3}$ and $\beta = .3$. We took $A = \sqrt{3}/2 - 1.645/\sqrt{12}$ and $B = \sqrt{3}/2 + 1.645/\sqrt{12}$ to ensure that the design was relatively thick in the tails. The number of replications was 100.

In this particular simulation, the GPLSIM estimates are far more accurate than the ordinary least squares (OLS) estimates, which are badly biased, and comparable to estimates obtained using nonlinear least squares based on the sinusoidal model. Table 1 displays the results of five randomly selected outcomes of the simulations. Note that although GPLSIM estimates are asymptotically efficient in the semiparametric sense (see Sec. 6), asymptotically they are more variable than fully parametric estimators computed at the correct model (see Theorem 4, Sec. 5.2), and this intrinsic difference between semiparametric and (correctly specified) parametric modeling exhibits itself here in the coefficient for Z. Not only are the GPLSIM estimates better than the OLS estimates, but they also do a reasonably effective job of fitting the data; see Figure 1.

Finally, we evaluated the accuracy of the estimated standard errors (defined in Sec. 7). In this simulation the cov-

Table 1. Results From Five Randomly Chosen Samples From the Sine-Bump Simulation Study

	Ordinary least squares				*Nonlinear least squares*				*GPLSIM*			
	X_1	X_2	X_3	Z	X_1	X_2	X_3	Z	X_1	X_2	X_3	Z
Est.	.564	.498	.659	.403	.595	.571	.565	.286	.595	.568	.569	.274
s.e.	.361	.368	.298	.069	.012	.013	.012	.012	.013	.013	.013	.026
Est.	.218	.142	.966	.216	.568	.568	.595	.277	.563	.574	.595	.281
s.e.	.766	.781	.207	.054	.010	.009	.009	.010	.010	.010	.010	.022
Est.	−.126	−.512	−.85	.263	.579	.580	.572	.310	.581	.580	.571	.310
s.e.	1.137	.97	.596	.059	.010	.010	.009	.010	.011	.011	.010	.023
Est.	.851	−.264	−.453	.351	.567	.590	.575	.300	.565	.599	.568	.307
s.e.	.796	1.364	1.349	.068	.010	.011	.011	.011	.012	.013	.013	.023
Est.	−.881	.396	−.261	.323	.587	.574	.570	.283	.592	.569	.571	.291
s.e.	.697	1.309	1.496	.064	.010	.010	.010	.010	.011	.010	.010	.020
MSE	.67	.79	.76	3.9e−3	1.1e−4	1.2e−4	1.1e−4	1.1e−4	1.4e−4	1.6e−4	1.3e−4	2.7e−4
MAE	.69	.75	.74	5.0e−2	8.5e−3	8.6e−3	8.5e−3	9.0e−3	9.6e−3	9.9e−3	9.0e−3	1.3e−2
mdE	.69	.79	.76	3.8e−2	7.9e−3	7.5e−3	6.9e−3	8.6e−3	8.6e−3	8.6e−3	8.2e−3	1.1e−2

NOTE: Mean squared error (MSE), mean absolute error (MAE), and median absolute error (mdE) values for the whole simulation are also given.

erage probabilities for nominal 95% confidence intervals were 94%, 96%, and 98% for the three components of $\mathbf{X}$, and 94% for Z. At least for this sample size and this model, the standard error estimates seem reasonably accurate.

3.2 Example: Framingham Data

The Framingham data were described in Section 1.1; Y corresponds to incidence of CHD and Z to smoking status. In this discussion we use disease and smoker to denote these variables. For covariates we used X_1, X_2, and X_3 as described in Section 1.1, with each variable scaled to lie between 0 and 1. To avoid problems with sparse data near the boundaries, after some experimentation we used only those data with a single-index value in range [.4, 1.2] for curve estimation. This excluded 45 of the 1,615 observations. We applied our methodology to the model

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

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). Table 2 displays the results of our analysis. For the purpose of illustration, we have compared these results to those obtained by ordinary logistic regression, which in this context is simply another way of estimating the "direction" $\boldsymbol{\alpha}_0$. We made the ordinary logistic regression coefficients for age, trblood, and logchol comparable to the single-index analysis by making their Euclidean norm equal to 1.0, and adjusted their standard error estimates accordingly.

Figure 2 shows the estimates of (a) η_0 and (b) the conditional probability of heart disease for both smokers and nonsmokers. An interesting feature of this figure is the curvature of $\hat{\eta}$ when the single index becomes greater than .8. We checked this curvature 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 partially linear model to the data using the GAM procedure of S-PLUS. The resulting estimate also showed curvature, of the same form as displayed in Figure 2. We also fit an ordinary GAM with nonparametric components in age, trblood, and logchol, and found a nonlinear structure with the "flatness" of Figure 2 for age.

We compared the GPLSIM fit to others as follows. First, we formed the estimated single-index $U = \hat{\boldsymbol{\alpha}}^T\mathbf{X}$, then ran

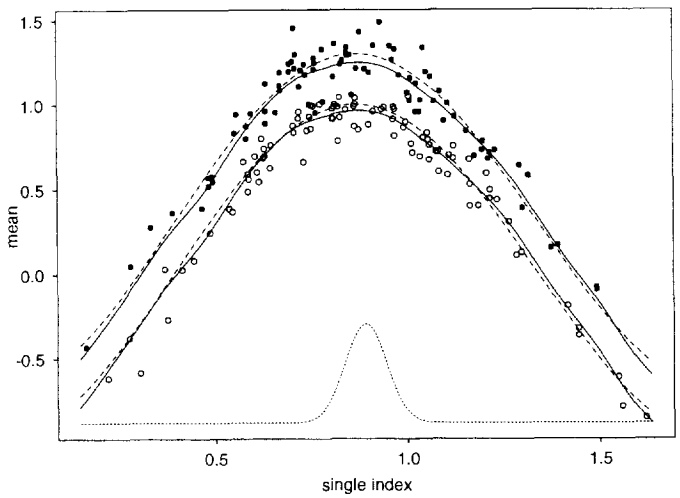

Figure 1. Curve Estimates for a Single Replication of the Sine-Bump Simulation Study. The data are shown by open circles for Z = 0 and closed circles for Z = 1. The solid curves correspond to the estimates of the underlying mean function when Z = 0 and Z = 1. The dashed curves are the true mean functions. The dotted curve is the kernel weight used in the local fitting process.

Table 2. Framingham Heart Study

	age	*trblood*	*logchol*	*smoker*
Ordinary logistic	.43	.57	.69	.57
s.e.	.10	.13	.11	.25
GPLSIM	.37	.65	.66	.59
s.e.	.086	.11	.12	.24

NOTE: "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.

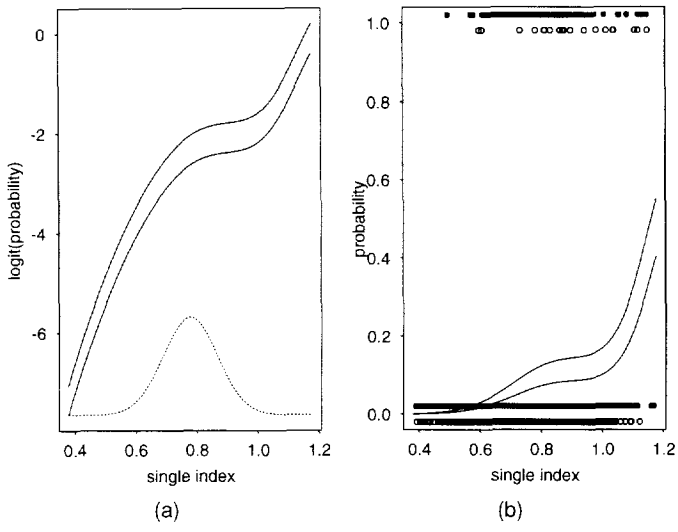

Figure 2. Curve Estimates for the Framingham Heart Study Data. (a) Solid curves correspond to estimates of logit P (heart disease) for smokers (upper curve) and nonsmokers (lower curve) against the estimated single-index described in the text. The dotted curve is the kernel weight used in the local linear fitting process. (b) Estimates of P (heart disease) for smokers (upper curve) and nonsmokers (lower curve) against the single index described in the text. The solid dot denotes smokers, the hollow dot, nonsmokers.

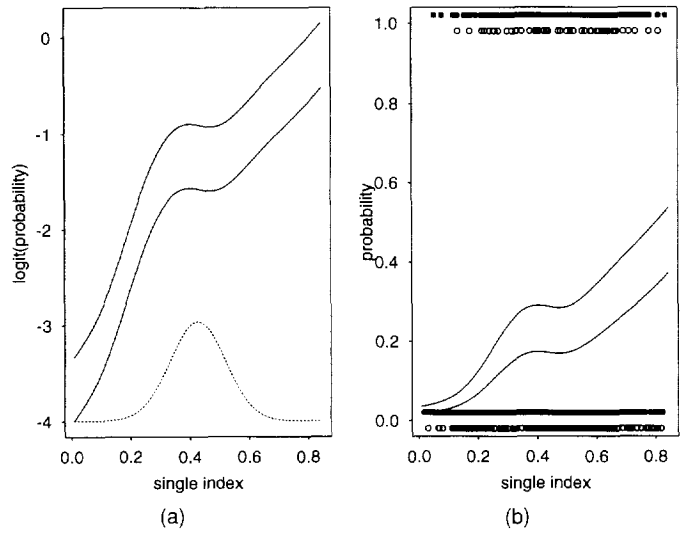

Figure 3. Curve Estimates for the Munich Dust Study Data. (a) Solid curves correspond to estimates of logit pr (Bronchitis) for smokers (upper curve) and nonsmokers (lower curve) against the estimated single-index described in the text. The dotted curve is the kernel weight used in the local linear fitting process. (b) Estimates of pr (Bronchitis) for smokers (upper curve) and nonsmokers (lower curve) against the single index described in the text. The solid dot denotes smokers; the hollow dot, nonsmokers.

a partially linear model in U (nonparametric) and Z (parametric) using the default "GAM" procedure in S-PLUS. We also ran a standard GAM with smoothers for each of X_1, X_2, and X_3, with Z entering as a parametric offset. In this case, surprisingly the GPLSIM had a *smaller* estimated deviance than the GAM, even though it had ≈ 8 more degrees of freedom.

One can also view this example as an informal model diagnostic of the logistic linear regression model via embedding it into the GPLSIM. Our result indicates certain departures from the logistic linear regression model; using the same informal method described in the previous paragraph, the linear logistic and the full GAM are not statistically significantly different, but the linear logistic and the GPLSIM are statistically significantly different.

3.3 Example: Dust Irritation Data

In occupational medicine one important issue is the assessment of the health hazard of specific harmful substances in a working area. We consider here the specific problem of estimating risk of bronchitis in a dust-burdened mechanical engineering plant in Munich.

The regressor variables $\mathbf{X}$ are X_1, the logarithm of 1.0 plus the average dust concentration in the working area over the period of time in question, and X_2, the duration of exposure. Also available was smoking status, Z. The data were described by Ulm (1991) as a possible example of a threshold regression model and were further analyzed by Küchenhoff and Carroll (1997). There were 1,246 observations. Little correlation among the variables was observed.

Table 3. Munich Dust Study

	trdust	*duration*	*smoker*
Ordinary logistic	.403	.915	.68
s.e.	.103	.045	.176
GPLSIM	.222	.975	.668
s.e.	.089	.021	.178

NOTE: "trdust" is transformed dust concentration, "duration" is the duration of exposure, and "smoker" is smoking status. The ordinary logistic coefficients for trdust and duration have been normalized to have Euclidean norm equal to 1.0, and the standard errors have been adjusted appropriately.

Table 3 displays the results of an ordinary logistic and GPLSIM fit to the data, and Figure 3 shows the logit and probability of bronchitis for smokers and nonsmokers. There is an important curvature in these data, which are not well fitted by an ordinary logistic model. As suggested by Küchenhoff and Carroll (1997), this curvature may reflect a threshold effect on concentration. The single-index model provides a slightly worse fit than a full GAM, although not a statistically significant one; we compared these using the deviances from GAM as implemented in S-PLUS, ignoring the effect of estimation of the single index. When compared to the GAM, an ordinary logistic model had an observed level of significance $< .0001$.

4. DISTRIBUTION THEORY: NONPARAMETRIC PART

4.1 Introduction

When $\boldsymbol{\alpha}_0$ is given as is the case in partially linear models or can be estimated at reasonable accuracy (e.g., by the average derivative method or sliced inverse regression), the following simple estimator is attractive from an implementation viewpoint. With the given value of $\hat{\boldsymbol{\alpha}}$, find $\hat{\eta}(u; h, \hat{\boldsymbol{\alpha}}) = \hat{a}$ by maximizing the local quasi-likelihood

$$\sum_{i=1}^{n} Q[g^{-1}\{a + b(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u) + \boldsymbol{\beta}^T\mathbf{Z}_i\}, Y_i]K_h(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u), \tag{11}$$

with respect to a, b, and $\boldsymbol{\beta}$. Because $\hat{\boldsymbol{\beta}}$ here is obtained locally, it can be improved to use all of the data, as follows.

Given $\hat{\boldsymbol{\alpha}}$ and the estimator $\hat{\eta}(u; h, \hat{\boldsymbol{\alpha}})$, one estimates $\hat{\boldsymbol{\beta}}$ by maximizing

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

with respect to $\boldsymbol{\beta}$. We call this noniterative procedure the *one-step estimator* and the algorithm of Section 2.1 the *fully iterated algorithm*. Based on he distribution theory provided in this and the next section, it is clear that both algorithms have their own merits. The fully iterated algorithm is at least as efficient as the one-step algorithm, but the one-step estimator achieves the same efficiency in some important applications with added computational convenience.

Note that in (11) we are maximizing the local quasi-likelihood with respect to $(a, b, \boldsymbol{\beta})$. This reflects the main difference from the estimation algorithm of Section 2.1, where we maximize with respect to (a, b) only. The foregoing idea can also be expanded to the case where $\boldsymbol{\alpha}$ is unknown by iteratively maximizing (11) and (12); one needs only to replace the first $\hat{\boldsymbol{\alpha}}$ in (12) by $\boldsymbol{\alpha}$ and maximize the modified (12) with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. See an earlier version of this article (Carroll, Fan, Gijbels, and Wand 1995) for details.

In this section we investigate properties of the estimators of the nonparametric part $\eta_0(\cdot)$ of (5) when α_0 is either known or estimated to the order $O_P(n^{-1/2})$ (i.e., at the usual parametric rate). The distribution theory depends on two cases: (a) the one-step approach, where $\boldsymbol{\beta}_0$ is estimated locally as in (11); and (b) the fully iterated approach (8), where $\boldsymbol{\beta}_0$ is estimated at parametric rates and thus $\eta_0(\cdot)$ can be estimated asymptotically as well as if $\boldsymbol{\beta}_0$ were known.

4.2 One-Step Estimate of the Nonparametric Part

Let $\rho_l(t) = \{dg^{-1}(t)/dt\}^l/[\sigma^2 V\{g^{-1}(t)\}]$, $l = 1, 2$, and denote the marginal density of $U = \boldsymbol{\alpha}_0^T\mathbf{X}$ by $f(\cdot)$. 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

$$\boldsymbol{\Sigma}(u) = E\left[\rho_2\{\eta_0(U) + \boldsymbol{\beta}_0^T\mathbf{Z}\} \times \begin{pmatrix} 1 & \mathbf{Z}^T \\ \mathbf{Z} & \mathbf{Z}\mathbf{Z}^T \end{pmatrix} \middle| U = u\right], \tag{13}$$

$$q_1(x, y) = \{y - g^{-1}(x)\}\rho_1(x),$$

$$m_i = m_i(U_i) = \eta_0(U_i) + \boldsymbol{\beta}_0^T\mathbf{Z}_i,$$

W_i = first element of the vector $q_1(m_i, Y_i)\boldsymbol{\Sigma}^{-1}(u)(1, \mathbf{Z}_i^T)^T$,

and

$$d(u) = \text{first diagonal element of the matrix } \boldsymbol{\Sigma}^{-1}(u).$$

Theorem 1. Consider the maximizer of the local quasi-likelihood (11). Then, as $n \to \infty$, $h \to 0$, and $nh \to \infty$, under Condition 1 in the Appendix,

$$(nh)^{1/2}\left(\begin{bmatrix} \hat{\eta}(u) - \eta_0(u) \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{bmatrix} - \frac{\kappa_2}{2}\eta_0''(u)h^2\boldsymbol{\Sigma}^{-1}(u)E \times \left[\rho_2\{\eta_0(U) + \boldsymbol{\beta}_0^T\mathbf{Z}\}\begin{pmatrix} 1 \\ \mathbf{Z} \end{pmatrix}\middle| U = u\right]\right) \xrightarrow{D}$$

$$\text{normal}\left[0, \frac{\nu_0}{f(u)}\boldsymbol{\Sigma}^{-1}(u)\right]. \tag{14}$$

In fact, we have the asymptotic expansion

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

and hence

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

Remark 1. Consider the situation where $\sigma^2 V(\mu) \equiv \sigma^2$ and $E(\mathbf{Z}|\mathbf{X}) = 0$. For this normal model with the identity link, the quasi-likelihood estimates are the OLS 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 $\hat{\eta}(u)$ are the same as if $\boldsymbol{\beta}_0$ were known.

Remark 2. The rate results in Theorem 1 continue to hold when the variance function is misspecified; that is, $\text{var}(Y|\mathbf{X}, \mathbf{Z}) \neq \sigma^2 V\{\mu(\mathbf{X}, \mathbf{Z})\}$. One must change the matrix $\boldsymbol{\Sigma}(u)$ to reflect the misspecification of the variance function. (See Fan, Heckman, and Wand 1995 for such a modification.)

4.3 Fully Iterated Estimate of the Nonparametric Part

For the fully iterative estimator, the parametric component can be estimated at root-n rate. Thus in Step 4 the local smoothing is carried out as if $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ were known. The results for the nonparametric component are easy: (16) continues to hold, replacing $d(u)$ by $d_*(u) = (E[\rho_2\{\eta_0(u) + \boldsymbol{\beta}_0^T\mathbf{Z}\}|U = u])^{-1}$. This result coincides with the univariate result given by Fan et al. (1995).

4.4 Bandwidth Selection

The results in the previous section suggest bandwidth estimators in the spirit of that of Ruppert, Sheather, and 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 $C(K) = (\nu_0 \kappa_2^{-2})^{1/5}$.

The Framingham example in Section 3.2 treats the case where both Y and Z are $0-1$ variables, so we briefly describe 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 $\zeta_0(\cdot)$. Let $\hat{\beta}$ 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 $\hat{\eta}(\cdot)$, $\hat{\zeta}_L(\cdot)$, and $\hat{\boldsymbol{\beta}}$. 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 approximately 10% clipped off each 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 a straight line.

A sensible rule for choice of h in Step 1 is more difficult. A relatively ad hoc possibility is

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

because this guarantees that the required bandwidth has correct order of magnitude for the conjectured optimal asymptotic performance. (See Remark 3 in Sec. 5.1 for more details.)

5. DISTRIBUTION THEORY: PARAMETRIC PARTS

We now study estimation for the parametric components $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$. We treat the one-dimensional case ($p=1$), for which $\boldsymbol{\alpha}_0 = 1$ and $\boldsymbol{\alpha}_0^T\mathbf{X} = X$, separately. Because in this case the one-step estimator has the advantage of being noniterative, we also provide its distribution theory.

5.1 The Scalar X Case: Partially Linear Models

The following theorem for the one-step estimate shows that one iteration leads already to a root-n consistent estimator.

Theorem 2. Let $\hat{\boldsymbol{\beta}}$ be the one-step estimate that maximizes the quasi-likelihood (12) with $\boldsymbol{\alpha} = 1$. Because $U = \boldsymbol{\alpha}_0^T\mathbf{X} = X$, 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 $nh^2/\log(1/h) \to \infty$,

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

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

Theorem 3. Under the conditions of Theorem 2, for the fully iterated estimator defined by (9) with $\boldsymbol{\alpha} = 1$, with $\rho_2(\cdot) = \rho_2\{\eta_0(X) + \boldsymbol{\beta}_0^T\mathbf{Z}\}$,

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

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

$$\mathbf{B}_2 = E\{\mathbf{Z}\mathbf{Z}^T\rho_2(\cdot)\} - E\left[\frac{E\{\mathbf{Z}\rho_2(\cdot)|X\}E\{\mathbf{Z}^T\rho_2(\cdot)|X\}}{E\{\rho_2(\cdot)|X\}}\right].$$

The same result holds for the estimator defined by (9) under the weaker condition that $nh^6 \to 0$.

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 to estimate $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ at the rate $n^{-1/2}$, one must undersmooth the nonparametric part $\eta_0(\cdot)$. The need to undersmooth to obtain usual rates of convergence is standard in the kernel literature and has analogs in the spline literature (Hastie and Tibshirani 1990, pp. 154–155). This undersmoothing is required for the estimator defined by (9). However, for the estimator defined by (10), in the linear regression single-index model with no $\mathbf{Z}$, ordinary bandwidth rates are permissible, as shown by Härdle et al. (1993), who suggested maximizing (10) simultaneously in the bandwidth and the parameters. Hunsberger (1994) and Severini and Staniswalis (1994) showed the same thing for the partially linear model (see also Severini and Wong 1992). Because ordinary bandwidths "work" for single-index models and also for partially linear models, it is reasonable to suppose that they also work for the combination, namely our GPLSIM's. A brief sketch of an argument was provided in an appendix of an earlier version of this article (Carroll et al. 1995), verifying that ordinary bandwidth rates are possible for full GPLSIM when (10) 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 $q(\mathbf{X}) = E(\mathbf{Z}|\mathbf{Z})$. Then $\mathbf{B}_2 = \sigma^{-2}E\{\text{var}(\mathbf{Z}|\mathbf{X})\}$, whereas the asymptotic variance (19) for the one-step estimator is

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

Because $\mathbf{B}_2^{-1} = \sigma^2\{E\mathbf{Z}\mathbf{Z}^T - q(\mathbf{X})q(\mathbf{X})^T\}^{-1}$, one can easily see that the fully iterated estimator is uniformly as efficient 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. Hence the one-step estimator is preferable when $\mathbf{X}$ and $\mathbf{Z}$ are weakly correlated, because it requires no iteration.

5.2 The Multivariate X Case: General Model

For a given $\hat{\eta}$, let $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ maximize the global quasi-likelihood (9). We assume that $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are in a $\sqrt{n}$ neighborhood of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$; that is, $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 = O_P(n^{-1/2})$ and $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O_P(n^{-1/2})$. Denote a generalized inverse of a square matrix $\mathbf{A}$ by $\mathbf{A}^{-1}$.

Theorem 4. Under Conditions 1 and 2 in the Appendix, the foregoing assumptions, and the restrictions on the bandwidths as stated in Theorem 3, for the estimators defined by (9) and (10),

$$n^{1/2}\begin{pmatrix}\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0\\ \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\end{pmatrix}\xrightarrow{D}\text{normal}(0,\mathbf{Q}^{-1}). \tag{21}$$

where, if $\rho_2(\cdot)=\rho_2\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X})+\boldsymbol{\beta}_0^T\mathbf{Z}\}$,

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

Remark 5. When $\sigma^2V(\mu)=\sigma^2$, with identity link and no $\boldsymbol{\beta}$ component, Theorem 4 reduces to the result of Härdle et al. (1993) for the single-index model.

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

6. ASYMPTOTIC EFFICIENCY IN THE SEMIPARAMETRIC SENSE

In this section we derive the information bound for the semiparametric model (3) and (5). This information bound turns out to be the matrix $\mathbf{Q}$ given in Theorem 4. Thus the estimator from Theorem 4 achieves the information lower bound and is efficient in the semiparametric sense.

To state the information bound, let us define the parameter space. Assume that η_0 is a completely unknown function with a continuous second derivative and that the joint density of $\mathbf{X}$ and $\mathbf{Z}$ with respect to some measure exists and is completely unknown.

Theorem 5. Under the foregoing assumptions, the information matrix for the semiparametric model (3) and (5) is $\mathbf{Q}$ given in Theorem 4.

7. 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 $1/V\{\hat{\mu}(\mathbf{X}_i.\mathbf{Z}_i)\}$; one can use $n-l_n-p-q$ df, 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.

7.1 Estimation in Partially Linear Models: Scalar X

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. We focus on (20), for which this fairly tedious process can be replaced by a simple consistent alternative based on the usual expansions for quasi-likelihood. The derivations are based on the simple form (9), instead of taking derivatives in (10), because these are more complex to compute.

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

$$\tilde{\mathbf{S}}=\begin{bmatrix}\mathbf{e}_1^T\{\mathbf{U}(U_1)^T\tilde{\mathbf{A}}\mathbf{K}(U_1)\mathbf{U}(U_1)\}^{-1}\mathbf{U}(U_1)^T\tilde{\mathbf{A}}\mathbf{K}(U_1)\\ \vdots\\ \mathbf{e}_1^T\{\mathbf{U}(U_n)^T\tilde{\mathbf{A}}\mathbf{K}(U_n)\mathbf{U}(U_n)\}^{-1}\mathbf{U}(U_n)^T\tilde{\mathbf{A}}\mathbf{K}(U_n)\end{bmatrix}. \tag{22}$$

where $\mathbf{U}(u_0)$ is the $n\times 2$ matrix with the first column all 1's and the second column with the terms $(U_i-u_0)/h$, and $\mathbf{K}(u_0)$ is diagonal with elements $K_h(U_i-u_0)$.

Here is the motivation for $\tilde{\mathbf{S}}$. 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 quasi-likelihood regression are the iterative solutions to the equation

$$\begin{bmatrix}a(u_0)\\ hb(u_0)\end{bmatrix} = \left\{\sum_{k=1}^n\mathbf{U}_k(u_0)\mathbf{U}_k(u_0)^TK_h(U_k-u_0)A_k(u_0)\right\}^{-1} \times\sum_{k=1}^n\mathbf{U}_k(u_0)K_h(U_k-u_0)A_k(u_0) \times\{a(u_0)+b(u_0)(U_i-u_0)+(Y_i-\mu_i)/(\sigma^2V_i\rho_{1i})\}. \tag{23}$$

where $\mathbf{U}_k(u_0)=\{1,\ (U_k-u_0)/h\}^T$ and $A_k(u_0)=\rho_2\{\eta(u_0)+\boldsymbol{\beta}^T\mathbf{Z}_k\}$. Setting $u_0=U_i$ for $i=1,\ldots,n$ and multiplying both sides of (23) by $\mathbf{e}_1^T$ yields (22).

The following argument has similarities to equation (6.22) of Hastie and 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 (23) can be ignored asymptotically. This means that the local quasi-likelihood algorithm is asymptotically equivalent to solving in $\boldsymbol{\beta}$ and η the equations

$$\boldsymbol{\beta}=(\tilde{\mathbf{Z}}^T\tilde{\mathbf{A}}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathbf{A}}(\tilde{\varepsilon}-\tilde{\boldsymbol{\eta}})$$

and

$$\tilde{\eta}=\tilde{\mathbf{S}}(\tilde{\varepsilon}-\tilde{\mathbf{Z}}\boldsymbol{\beta}).$$

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

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

Because $\tilde{\varepsilon}$ has covariance matrix $\tilde{\mathbf{A}}^{-1}$, an approximate covariance matrix for $\hat{\boldsymbol{\beta}}$ is $\tilde{\mathbf{H}}_1\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{H}}_1^T$. One can show this estimate yields asymptotically consistent standard errors for $\hat{\boldsymbol{\beta}}$.

7.2 Estimation in General Models: Multivariate X

When $\boldsymbol{\alpha}_0$ is unknown, there are again two strategies: Nonparametric regression techniques can be used to estimate the terms in (21), or we can again develop directly a consistent estimate of (21). We build on the notation in Section 7.1.

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

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

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

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

where θ is a Lagrange multiplier associated with the constraint $\boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1$. Of course, the same argument used in deleting a term explained following (23) is used here. Multiplying both sides by $P^*_{\boldsymbol{\alpha}}$ and solving, we find that $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T = (\mathbf{P}^*_{\boldsymbol{\alpha}}\tilde{\mathbf{R}}^T\tilde{\mathbf{A}}\tilde{\mathbf{R}})^-\mathbf{P}^*_{\boldsymbol{\alpha}}\tilde{\mathbf{R}}^T\tilde{\mathbf{A}}(\tilde{\varepsilon} - \tilde{\boldsymbol{\eta}})$. Remembering that $\tilde{\boldsymbol{\eta}} = \tilde{\mathbf{S}}(\tilde{\varepsilon} - \tilde{\mathbf{Q}}\boldsymbol{\alpha} - \tilde{\mathbf{Z}}\boldsymbol{\beta})$, we find after some algebra that $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T = \tilde{\mathbf{H}}_2\tilde{\varepsilon}$ and

$$\tilde{\mathbf{H}}_2 = \{\mathbf{P}^*_{\boldsymbol{\alpha}}\tilde{\mathbf{R}}^T\tilde{\mathbf{A}}(\mathbf{I} - \tilde{\mathbf{S}})\tilde{\mathbf{R}}\}^-\mathbf{P}^*_{\boldsymbol{\alpha}}\tilde{\mathbf{R}}^T\tilde{\mathbf{A}}(\mathbf{I} - \tilde{\mathbf{S}}).$$

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

8. IMPLEMENTATION

To cut down on the computational labor at the curve estimation stages, we used fast binned approximations (see, e.g., Fan and Marron 1994 and Härdle and Scott 1992). Binning methods can also be used for fast computation of the standard error estimates. Details of such calculations ere given by Turlach and Wand (1995). An S-PLUS/Fortran module for fitting GPLSIM in certain special cases is available from World Wide Web site http://www.agsm.unsw.edu.au/~wand/software.html.

9. DISCUSSION

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

$$g\{\mu(\mathbf{x}, \mathbf{z})\} = \eta_{z_1}(\boldsymbol{\alpha}_0^T\mathbf{x}) + \boldsymbol{\beta}_0^T\mathbf{z}_2, \tag{24}$$

where $\mathbf{z} = (z_1, \mathbf{z}_2)$ with z_1 binary. However, our methods can be modified to handle (24). The local quasi-likelihood (8) should be replaced by

$$\begin{aligned} &\sum_{i=1}^n Q[g^{-1}\{a_0 + b_0(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u) + \hat{\boldsymbol{\beta}}^T\mathbf{Z}_{2,i}\}, Y_i] \\ &\quad \times K_{h_0}(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u)I(Z_{1,i} = 0) \\ &\quad + \sum_{i=1}^n Q[g^{-1}\{a_1 + b_1(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u) + \hat{\boldsymbol{\beta}}^T\mathbf{Z}_{2,i}\}, Y_i] \\ &\quad \times K_{h_1}(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u)I(Z_{1,i} = 1), \end{aligned}$$

where h_0 and h_1 are bandwidths for η_0 and η_1. The estimators for η_0 and η_1 are $\hat{\eta}_0(u) = \hat{a}_0$ and $\hat{\eta}_1(u) = \hat{a}_1$. One can modify the global quasi-likelihood analogously.

Model (5) also allows modeling interactions of the form

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

where $\boldsymbol{\Lambda}$ is the parameter matrix for interactions. This model is included in (5) by forming a new and longer X vector. One can also incorporate partial interaction terms in (5), which would reduce the number of effective parameters.

APPENDIX: PROOFS

Here we outline the key ideas for proving Theorems 1, 2, 4, and 5. Details can be found in an earlier draft of this article (Carroll et al. 1995). The methods for proving Theorem 3 are similar.

A.1 Conditions

For simplicity of notation, here we 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 $q_l(x, y) = (\partial^l/\partial x^l)Q\{g^{-1}(x), y\}$, $l = 1, 2, 3$. Then

$$\begin{aligned} q_1(x, y) &= \{y - g^{-1}(x)\}\rho_1(x) \\ \text{and} \quad q_2(x, y) &= \{y - g^{-1}(x)\}\rho_1'(x) - \rho_2(x), \end{aligned} \tag{A.1}$$

where $\rho_l(t) = \{dg^{-1}(t)/dt\}^l/V\{g^{-1}(t)\}$ is introduced in Section 4.2. In Condition 1, u is a generic argument for Theorem 1, and the condition must hold uniformly in u for Theorems 2–4.

Condition 1.

a. The function $q_2(x, y) < 0$ for $x \in \mathbb{R}$ and y in the range of the response variable.
b. The marginal density of $\boldsymbol{\alpha}_0^T\mathbf{X}$ is positive and continuous at the point u.
c. The function $\eta_0''(\cdot)$ is continuous at the point u.
d. $g''(\cdot)$ and $V(\cdot)$ are continuous functions.
e. With $R = \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}$, $E\{q_1^2(R, Y)|U = t\}$, $E\{q_1^2(R, Y)\mathbf{Z}|U = t\}$ and $E\{q_1^2(R, Y)\mathbf{Z}\mathbf{Z}^T|U = t\}$ are continuous in t at the point u. Moreover, $E[q_2^2\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}, Y\}] < \infty$ and $E[q_1^{2+\delta}\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}, Y\}] < \infty$, for some $\delta > 2$.
f. The kernel K is a symmetric density function with bounded support.
g. The random vector $\mathbf{Z}$ is assumed to have a bounded support.

Condition 2.

a. The marginal density of $\boldsymbol{\alpha}^T\mathbf{X}$ is positive and uniformly continuous for $\boldsymbol{\alpha}$ in a neighborhood of $\boldsymbol{\alpha}_0$. Further, $\boldsymbol{\alpha}_0^T\mathbf{X}$ has a positive density on its support D.
b. The function $\eta_0''(\cdot)$ is continuous in $u \in D$.
c. The density function of $\mathbf{X}$ has a continuous second derivative.
d. The function $V''(\cdot)$ and $g'''(\cdot)$ are continuous.
e. With $R = \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}$, $E\{q_1^2(R, Y)|\mathbf{X} = u\}$, $E\{q_1^2(\eta_0(R, Y)\mathbf{Z}|\mathbf{X} = u\}$ and $E\{q_1^2(R, Y)\mathbf{Z}\mathbf{Z}^T|\mathbf{X} = u\}$ are twice differentiable in $u \in D$,

Proof of Theorem 1

Let $c_n = (nh)^{-1/2}$, $U_i = \boldsymbol{\alpha}_0^T\mathbf{X}_i$,

$$\mathbf{X}_i^* = \begin{pmatrix} 1 \\ (U_i - u)/h \\ \mathbf{Z}_i \end{pmatrix},$$

and

$$\hat{\boldsymbol{\beta}}^* = \begin{pmatrix} c_n^{-1}\{\hat{a} - \eta_0(U)\} \\ c_n^{-1} h\{\hat{b} - \eta_0'(u)\} \\ c_n^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \end{pmatrix},$$

and let $f(\cdot)$ denote the density function of $U_i = \boldsymbol{\alpha}_0^T \mathbf{X}_i$. 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 $(\hat{a}, \hat{b}, \hat{\boldsymbol{\beta}})^T$ maximizes (11), then $\hat{\boldsymbol{\beta}}^*$ maximizes

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

with respect to $\boldsymbol{\beta}^*$. The concavity of the function $l_n(\boldsymbol{\beta}^*)$ is ensured by Condition 1a. By a Taylor expansion of the function $Q(g^{-1}(\cdot), Y_i)$ we obtain that

$$l_n(\boldsymbol{\beta}^*) = \mathbf{W}_n^T \boldsymbol{\beta}^* + \frac{1}{2} \boldsymbol{\beta}^{*T} \mathbf{A}_n \boldsymbol{\beta}^* \{1 + o_P(1)\}, \tag{A.2}$$

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

and

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

Define

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

and

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

It can be shown that $\mathbf{A}_n = -f(u) E[\rho_2(\eta_0(U) + \boldsymbol{\beta}_0^T \mathbf{Z}) \mathbf{A}(\mathbf{Z}) | U = u] + o_P(1) \equiv -A + o_P(1)$. Therefore, by (A.1),

$$l_n(\boldsymbol{\beta}^*) = \mathbf{W}_n^T \boldsymbol{\beta}^* - \frac{1}{2} \boldsymbol{\beta}^{*T} \mathbf{A} \boldsymbol{\beta}^* + o_P(1). \tag{A.3}$$

By applying the convexity lemma (see Pollard 1991), we obtain that $\hat{\boldsymbol{\beta}}^* = \mathbf{A}^{-1} \mathbf{W}_n + o_P(1)$. Hence the asymptotic normality of $\hat{\boldsymbol{\beta}}^*$ will follow from that of $\mathbf{W}_n$, which we establish next. By the definition of $\mathbf{W}_n$, it can be shown that

$$E\mathbf{W}_n = c_n^{-1} \frac{1}{2} \eta_0''(u) h^2 f(u) E \times [\rho_2\{\eta_0(U) + \boldsymbol{\beta}_0^T \mathbf{Z}\}(\kappa_2, 0, \kappa_2 \mathbf{Z}^T)^T | U = u] + o(c_n^{-1} h^2) \tag{A.4}$$

and that $\operatorname{var}(\mathbf{W}_n) = f(u) E[\rho_2\{\eta_0(U) + \boldsymbol{\beta}_0^T \mathbf{Z}\} \mathbf{B}(\mathbf{Z}) | U = u] + o(1) \equiv \mathbf{B} + o(1)$. Using Condition 1e, it can be shown that Liapounov's condition is satisfied and hence $\hat{\boldsymbol{\beta}}^*$ is asymptotically normal. This establishes Theorem 1.

Proof of Theorem 2

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

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

then

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

Proof of Theorem 2

First, we note that under Condition 2, by a result of Mack and Silverman (1982), (A.3) 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

$$\sup_{u \in D} |\hat{\boldsymbol{\beta}}^*(u) - \mathbf{A}^{-1} \mathbf{W}_n(u)| \xrightarrow{P} 0, \tag{A.5}$$

where $\hat{\boldsymbol{\beta}}^*(u)$ and $\mathbf{W}_n(u)$ are defined in the proof of Theorem 1, except that here we stress the dependence on u. So, by considering the first element of the vectors in (A.5), we have

$$\sup_{u \in D} \left| \hat{\eta}(u) - \eta_0(u) - \frac{1}{n f(u)} \sum_{i=1}^{n} W_i K_h(X_i - u) \right| = o_P(c_n),$$

where $f(u)$ is the density of X_i and W_i is the first element of the vector $q_1(\bar{\eta}_i, Y_i) \boldsymbol{\Sigma}^{-1}(u)(1, \mathbf{Z}_i^T)^T$, 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, the following stronger result holds:

$$\sup_{u \in D} \left| \hat{\eta}(u) - \eta_0(u) - \frac{1}{n f(u)} \sum_{i=1}^{n} W_i K_h(X_i - u) \right| = O_P\{h^2 c_n + c_n^2 \log^{1/2}(1/h)\}. \tag{A.6}$$

Let $\hat{\theta} = n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, $\hat{m}_i = \hat{\eta}(X_i) + \boldsymbol{\beta}_0^T \mathbf{Z}_i$, and $m_i = \eta_0(X_i) + \boldsymbol{\beta}_0^T \mathbf{Z}_i$. Then $\boldsymbol{\theta}$ maximizes

$$l_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} [Q\{g^{-1}(\hat{m}_i + n^{-1/2} \boldsymbol{\theta}^T \mathbf{Z}_i), Y_i\} - Q\{g^{-1}(\hat{m}_i), Y_i\}]. \tag{A.7}$$

By Taylor's expansion, we have

$$l_n(\boldsymbol{\theta}) = n^{-1/2} \sum_{i=1}^{n} q_1(\hat{m}_i, Y_i) \boldsymbol{\theta}^T \mathbf{Z}_i + \frac{1}{2} \boldsymbol{\theta}^T \mathbf{B}_n \boldsymbol{\theta} \tag{A.8}$$

and

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

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 can be shown that

$$\mathbf{B}_n = -E \rho_2\{\eta_0(X) + \boldsymbol{\beta}_0^T \mathbf{X}\} \mathbf{Z}\mathbf{Z}^T + o_P(1) \equiv -\mathbf{B} + o_P(1). \tag{A.9}$$

Using similar arguments as for obtaining (A.9), we get

$$n^{-1/2} \sum_{i=1}^{n} q_1(\hat{m}_i, Y_i) \mathbf{Z}_i = n^{-1/2} \sum_{i=1}^{n} q_1(m_i, Y_i) \mathbf{Z}_i + n^{-1/2} \sum_{i=1}^{n} q_2(m_i, Y_i)\{\hat{\eta}(X_i) - \eta_0(X_i)\} \mathbf{Z}_i + O_P(n^{1/2} \|\hat{\eta} - \eta_0\|_\infty^2).$$

By (A.6), the second term in the foregoing expression can be expressed as

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

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 thus

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

It can be shown via calculating the second moment that

$$T_{n2} - T_{n3} \xrightarrow{P} 0, \tag{A.10}$$

where $T_{n3} = -n^{-1/2}\sum_{j=1}^{n}\gamma(X_j)v_j$ with $\gamma(u) = E[\rho_2\{\eta_0(u) + \boldsymbol{\beta}_0^T\mathbf{Z}\}\mathbf{Z}|X = u]$. Combining (A.7)–(A.10), we obtain that $l_n(\boldsymbol{\theta}) = n^{-1/2}\sum_{i=1}^{n}\Omega(X_i, Y_i, \mathbf{Z}_i) - \frac{1}{2}\boldsymbol{\theta}^T\mathbf{B}\boldsymbol{\theta} + o_P(1)$, where $\Omega(X_i, Y_i, \mathbf{Z}_i) = q_1(m_i, Y_i)\mathbf{Z}_i - \gamma(X_i)v_i$. By the convexity lemma, we find that $\hat{\boldsymbol{\theta}} = \mathbf{B}^{-1}n^{-1/2}\sum_{i=1}^{n}\Omega(X_i, Y_i, \mathbf{Z}_i) + o_P(1)$, from which it follows that $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(0, \mathbf{B}^{-1}\boldsymbol{\Sigma}_1\mathbf{B}^{-1})$, as claimed.

Proof of Theorem 4

We use the notation $U = \boldsymbol{\alpha}_0^T\mathbf{X}$, $\hat{U} = \hat{\boldsymbol{\alpha}}^T\mathbf{X}$ and $f(\cdot)$ for the density function of U. The proof relies on two steps, which we state first and prove afterward. The first step consists of an expansion for $\hat{\eta}$ (at an argument u_0). We show that

$$\begin{aligned} & \hat{\eta}(u_0; h, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) - \eta_0(u_0) \\ & = n^{-1}\sum_{i=1}^{n} K_h(U_i - u_0) \\ & \quad \times \frac{\varepsilon_i}{f(u_0)E\{\rho_2(\cdot)|U = u_0\}} \\ & \quad - (\hat{\boldsymbol{\beta}}^T - \boldsymbol{\beta}_0^T)\frac{E\{\mathbf{Z}\rho_2(\cdot)|U = u_0\}}{E\{\rho_2(\cdot)|U = u_0\}} \\ & \quad - (\hat{\boldsymbol{\alpha}}^T - \boldsymbol{\alpha}_0^T)\frac{E\{\mathbf{X}\rho_2(\cdot)\eta_0'(\cdot)|U = u_0\}}{E\{\rho_2(\cdot)|U = u_0\}} + o_P(n^{-1/2}), \end{aligned} \tag{A.11}$$

where "$\cdot$" denotes the argument $\eta_0(U) + \boldsymbol{\beta}_0^T\mathbf{Z}$ and $\varepsilon_i = \{Y_i - \mu(\cdot)\}\rho_1(\cdot)$ with a similar convention.

The second step is as follows. Introduce the shorthand notations

$$\boldsymbol{\Lambda}_i = \begin{bmatrix} \mathbf{X}_i\eta_0'(U_i) \\ \mathbf{Z}_i \end{bmatrix}$$

and

$$\mathbf{P}\boldsymbol{\alpha} = \begin{bmatrix} \mathbf{I} - \boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^T & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} + o_P(1).$$

We show that

$$\begin{aligned} & \mathbf{P}\boldsymbol{\alpha}\mathbf{Q}n^{1/2}\begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} \\ & \quad = n^{-1/2}\textstyle\sum_{i=1}^{n}\varepsilon_i\mathbf{P}\boldsymbol{\alpha}\left[\boldsymbol{\Lambda}_i - \frac{E\{\boldsymbol{\Lambda}\rho_2(\cdot)|U_i\}}{E\{\rho_2(\cdot)|U_i\}}\right] + o_P(1). \end{aligned} \tag{A.12}$$

Because ε_i has variance ρ_{2i}, the right side of (36) has the covariance matrix $\mathbf{P}\boldsymbol{\alpha}\mathbf{Q}\mathbf{P}\boldsymbol{\alpha}$, verifying the statement of Theorem 4.

Proof of (A.11)

Let $a = \eta_0(u_0)$ and $b = h\eta'(u_0)$. The local linear estimates solve

$$0 = n^{-1}\sum_{i=1}^{n} K_h(\hat{U}_i - u_0)\begin{bmatrix} 1 \\ (\hat{U}_i - u_0)/h \end{bmatrix}\{Y_i - \hat{\mu}(\cdot)\}\hat{\rho}_1(\cdot),$$

where $\hat{\mu}(\cdot) = \mu\{\hat{a} + \hat{b}(\hat{U}_i - u_0)/h + \hat{\boldsymbol{\beta}}^T\mathbf{Z}_i\}$, and similarly for $\hat{\rho}_1(\cdot)$. Via Taylor series and using the conditions on h, we obtain

$$\begin{aligned} 0 &= n^{-1}\sum_{i=1}^{n} K_h(U_i - u_0)\begin{bmatrix} 1 \\ (U_i - u_0)/h \end{bmatrix}\{Y_i - \mu_*(\cdot)\}\rho_{1*}(\cdot) \\ &\quad - B_{n1}\begin{pmatrix} \hat{a} - a \\ \hat{b} - b \end{pmatrix} - (\hat{\boldsymbol{\beta}}^T - \boldsymbol{\beta}_0^T)B_{n2} - (\hat{\boldsymbol{\alpha}}^T - \boldsymbol{\alpha}_0^T)B_{n3} \\ &\quad + o_P(n^{-1/2}) + O_P(h^2), \end{aligned}$$

where $\mu_*(\cdot) = \mu\{a + b(U_i - u_0)/h + \boldsymbol{\beta}_0^T\mathbf{Z}_i\}$ and $\rho_{1*}(\cdot)$ is defined similarly. Here $B_{n,j}$ $(j = 1, 2, 3)$ are the resulting sample matrices of kernel form. Solving the foregoing linearized equation and substituting $B_{n,j}$ with their asymptotic counterparts, we obtain (A.11).

Proof of (A.12). Recall that (9) and (10) lead to asymptotically equivalent estimates. Consider (9) and use the expansion

$$\begin{aligned} & \hat{\eta}(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) - \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}_i) \\ & \quad = \hat{\eta}(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) - \hat{\eta}(\boldsymbol{\alpha}_0^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \\ & \qquad + \hat{\eta}(\boldsymbol{\alpha}_0^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) - \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}_i) \\ & \quad = \hat{\eta}'(\boldsymbol{\alpha}_0^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\alpha}}^T - \boldsymbol{\alpha}_0^T)\mathbf{X}_i + \hat{\eta}(\boldsymbol{\alpha}_0^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \\ & \qquad - \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}_i) + o_P(n^{-1/2}) \\ & \quad = \eta_0'(\boldsymbol{\alpha}_0^T\mathbf{X}_i)(\hat{\boldsymbol{\alpha}}^T - \boldsymbol{\alpha}_0^T)\mathbf{X}_i + \hat{\eta}(\boldsymbol{\alpha}_0^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \\ & \qquad - \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}_i) + o_P(n^{-1/2}), \end{aligned} \tag{A.13}$$

where we dropped the dependence on h for notational simplicity. The second term is handled by (A.13). With θ as the Lagrange multiplier, we know that $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ is the solution to

$$\begin{aligned} 0 &= \theta\begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ 0 \end{pmatrix} \\ &\quad + n^{-1/2}\sum_{i=1}^{n}\begin{bmatrix} \mathbf{X}_i\hat{\eta}'(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \\ \mathbf{Z}_i \end{bmatrix} \\ &\quad \times [Y_i - \mu\{\hat{\eta}(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{\beta}}^T\mathbf{Z}_i\}] \\ &\quad \times \rho_1\{\hat{\eta}(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{\beta}}^T\mathbf{Z}_i\}. \end{aligned}$$

We can expand $\hat{\eta}(\cdot)$ about $\eta_0(\cdot)$ using (A.11). Write $\mu_i = \mu\{\eta_0(U_i) + \boldsymbol{\beta}_0^T \mathbf{Z}_i)\}$, and similarly for ρ_{ji}. Make the further definition

$$A_{\boldsymbol{\alpha},\boldsymbol{\beta}} = E\left[\rho_2(\cdot)\begin{Bmatrix} \mathbf{X}\eta_0'(\cdot) \\ \mathbf{Z} \end{Bmatrix}\begin{Bmatrix} \mathbf{X}\eta_0'(\cdot) \\ \mathbf{Z} \end{Bmatrix}^T\right].$$

By the Taylor series, and using (A.13), we have that (using that $nh^4 \to 0$)

$$\begin{aligned} 0 &= \theta\begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ 0 \end{pmatrix} + n^{-1/2}\sum_{i=1}^n \boldsymbol{\Lambda}_i\varepsilon_i - n^{-1/2} \\ &\quad \times \sum_{i=1}^n \boldsymbol{\Lambda}_i(\hat{\boldsymbol{\beta}}^T - \boldsymbol{\beta}_0^T)\mathbf{Z}_i\rho_{2i}(\cdot) \\ &\quad - n^{-1/2}\sum_{i=1}^n \rho_{2i}\boldsymbol{\Lambda}_i\{\hat{\eta}(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}) - \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}_i)\} + o_P(1) \\ &= \theta\begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ 0 \end{pmatrix} + n^{-1/2}\sum_{i=1}^n \boldsymbol{\Lambda}_i\varepsilon_i - A_{\boldsymbol{\alpha},\boldsymbol{\beta}}n^{1/2}\begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} \\ &\quad - n^{-1/2}\sum_{i=1}^n \rho_{2i}\boldsymbol{\Lambda}_i\{\hat{\eta}(\boldsymbol{\alpha}_0^T\mathbf{X}_i;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}) - \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}_i)\} + o_P(1). \end{aligned}$$

We now invoke (A.11), which implies that

$$\begin{aligned} 0 &= \theta\begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ 0 \end{pmatrix} + n^{-1/2}\sum_{i=1}^n \boldsymbol{\Lambda}_i\varepsilon_i - Qn^{1/2}\begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} \\ &\quad - n^{-1/2}\sum_{i=1}^n \boldsymbol{\Lambda}_i\rho_{2i}n^{-1}\sum_{j=1}^n K_h(U_j - U_i) \\ &\quad \times \frac{Y_j - \mu\{\eta_0(U_i) + \boldsymbol{\beta}_0^T\mathbf{Z}_j\}}{f(U_i)E\{\rho_2(\cdot)|U_i\}}\rho_1\{\eta_0(U_i) + \boldsymbol{\beta}_0^T\mathbf{Z}_j\}. \quad \text{(A.14)} \end{aligned}$$

Only the last term is of interest, and hence we focus on it. Interchanging the summations, we get

$$\begin{aligned} &n^{-1/2}\sum_{i=1}^n \left[n^{-1}\sum_{j=1}^n \boldsymbol{\Lambda}_j\rho_{2j}K_h(U_j - U_i)\right. \\ &\qquad \left. \times \frac{Y_i - \mu\{\eta_0(U_j) + \boldsymbol{\beta}_0^T\mathbf{Z}_i\}}{f(U_j)E\{\rho_2(\cdot)|U_j\}}\rho_1\{\eta_0(U_j) + \boldsymbol{\beta}_0^T\mathbf{Z}_i\}\right]. \end{aligned}$$

The term in the square brackets, being a nonparametric regression, is essentially the same as

$$n^{-1/2}\sum_{i=1}^n \varepsilon_i \frac{E\{\boldsymbol{\Lambda}\rho_2(\cdot)|U_i\}}{E\{\rho_2(\cdot)|U_i\}}, \quad \text{(A.15)}$$

for a symmetric kernel. Combining (A.14) and (A.15), and multiplying by $\mathbf{P}\boldsymbol{\alpha}$, we obtain (A.11).

A.5 Proof of Theorem 5

Let $h(\mathbf{x}, \mathbf{z})$ be the joint density of $(\mathbf{X}, \mathbf{Z})$. Then, under the semiparametric model (3) and (5), the joint density of $(\mathbf{X}, Y, \mathbf{Z})$ is given by

$$f(\mathbf{x}, y, \mathbf{z}) = \exp[y\theta(\mathbf{x}, \mathbf{z}) - \mathcal{B}\{\theta(\mathbf{x}, \mathbf{z})\} + C(y)]h(\mathbf{x}, \mathbf{z}). \quad \text{(A.16)}$$

where $\theta(\mathbf{x}, \mathbf{z}) = g_0 \circ g^{-1}\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{x}) + \boldsymbol{\beta}_0^T\mathbf{z}\}$ with $\|\boldsymbol{\alpha}_0\| = 1$ and g_0 as the canonical link function. Define

$$\begin{aligned} P_1 &= \{\text{Model (A.16) with given } \eta_0(\cdot), \text{and } h\}, \\ P_2 &= \{\text{Model (A.16) with given } \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \text{and } h(\cdot)\}, \end{aligned}$$

and

$$P_3 = \{\text{Model (A.16) with given } \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0 \text{ and } \eta_0(\cdot)\}.$$

Then the score function for $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ under the parametric model P_1 is given by

$$\dot{l} = \{Y - \mu(\mathbf{X}, \mathbf{Z})\}g_1'\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}\}\begin{pmatrix} \eta_0'(\boldsymbol{\alpha}_0^T\mathbf{X})\mathbf{X} \\ \mathbf{Z} \end{pmatrix}.$$

where $g_1 = g_0 \circ g^{-1}$. The tangent space (Bickel et al. 1993, p. 50) of the nonparametric model P_2 can be shown to be $\dot{P}_2 = [\{Y - \mu(\mathbf{X}, \mathbf{Z})\}g_1'(\cdot)a(\boldsymbol{\alpha}_0^T\mathbf{X})$, for all $a \in L_2]$, and the tangent space of the nonparametric model P_3 is given by $\dot{P}_3 = [b(\mathbf{X}, \mathbf{Z}) \in L_2 : Eb(\mathbf{X}, \mathbf{Z}) = 0]$. Then, by theorem 3.4.1 of Bickel et al. (1993), the efficient score function of $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ under model (A.16) is the projection of $\dot{l}$ into the orthogonal complement of the linear space $\dot{P}_2 + \dot{P}_3$—namely, $\dot{l}^* = \dot{l} - \prod(\dot{l}|\dot{P}_2 + \dot{P}_3)$. The information matrix for $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ is just $E(\dot{l}^*)(\dot{l}^*)^T$, where $\prod(\dot{l}|\dot{P}_2 + \dot{P}_3)$ is the projection of $\dot{l}$ into $\dot{P}_2 + \dot{P}_3$. Because $\dot{P}_2 \perp \dot{P}_3$ and $\dot{l} \perp \dot{P}_3$, the projection $\prod(\dot{l}|\dot{P}_2 + \dot{P}_3) = \prod(\dot{l}|\dot{P}_2)$ is to find a vector function of form $(Y - \mu)g_1'(\cdot)a(\boldsymbol{\alpha}_0^T\mathbf{X})$ such that $E\|\dot{l} - (Y - \mu)g_1'(\cdot)a(\boldsymbol{\alpha}_0^T\mathbf{X})\|^2$ is minimized. By conditioning on $\boldsymbol{\alpha}_0^T\mathbf{X}$, one can easily find that

$$\prod(\dot{l}|\dot{P}_2) = (Y - \mu)g_1'(\cdot)\begin{bmatrix} E\{\mathbf{X}\eta_0'(U)\rho_2(\cdot)|U\}/E\{\rho_2(\cdot)|U\} \\ E\{\mathbf{Z}\rho_2(\cdot)|U\}/E\{\rho_2(\cdot)|U\} \end{bmatrix},$$

where $U = \boldsymbol{\alpha}_0^T\mathbf{X}$. Using this, it is now easy to verify that $Q = E(\dot{l}^*)(\dot{l}^*)^T$.

[Received April 1995. Revised August 1996.]

REFERENCES

Bickel, P. J., Klaassen, A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Inference in Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Bonneu, M., Delecroix, M., and Hristache, M. (1995), "Semiparametric Estimation of Generalized Linear Models and Related Models," unpublished manuscript.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1995), "Generalized Partially Linear Single-Index Models," Discussion Paper 9506, Institute of Statistics, Catholic University of Louvain, Louvain-la-Neuve, Belgium.

Chen, H. (1988), "Convergence Rates for Parametric Components in a Partly Linear Model," *The Annals of Statistics*, 16, 136–146.

Cuzick, J. (1992), "Semiparametric Additive Regression," *Journal of the Royal Statistical Society*, Ser. B, 54, 831–843.

Fan, J., Heckman, N. E., and Wand, M. P. (1995), "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-likelihood Functions," *Journal of the American Statistical Association*, 90, 141–150.

Fan, J., and Marron, J. S. (1994), "Fast Implementations of Nonparametric Curve Estimators," *Journal of Computational and Graphical Statistics*, 3, 35–56.

Härdle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing in Single-Index Models," *The Annals of Statistics*, 21, 157–178.

Härdle, W., and Scott, D. W. (1992), "Smoothing by Weighted Averaging of Rounded Points," *Computational Statistics*, 7, 97–128.

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

Heckman, N. (1986), "Spline Smoothing in a Partly Linear Model," *Journal of the Royal Statistical Society*, Ser. 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., and 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.

Küchenhoff, H., and Carroll, R. J. (1997), "Segmented Regression With Errors in Predictors," *Statistics in Medicine*, to appear.

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–342.

Mack, Y. P., and Silverman, B. W. (1982), "Weak and Strong Uniform Consistency of Kernel Regression Estimates," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 61, 405–415.

Mammen, E., and van de Geer, S. (1995), "Penalized Estimation in Partial Linear Models," Technical Report 95-05, University of Leiden, The Netherlands.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Pollard, D. (1991), "Asymptotics for Least Absolute Deviation Regression Estimators," *Econometric Theory*, 7, 186–199.

Royston, P., and Altman, D. G. (1994), "Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling," *Applied Statistics*, 43, 429–467.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257–1270.

Severini, T. A., and Staniswalis, J. G. (1994), "Quasi-likelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501–511.

Severini, T. A., and Wong, W. H. (1992), "Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics*, 20, 1768–1802.

Speckman, P. (1988), "Kernel Smoothing in Partial Linear Models," *Journal of the Royal Statistical Society*, Ser. B, 50, 413–436.

Turlach, B. A., and Wand, M. P. (1995), "Fast Computation of Auxiliary Quantities in Local Polynomial Regression," unpublished manuscript.

Ulm, K. (1991), "A Statistical Method for Assessing a Threshold in Epidemiological Studies," *Statistics in Medicine*, 10, 341–349.

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, pp. 319–329.

Weisberg, S., and Welsh, A. H. (1994), "Estimating the Missing Link Function," *The Annals of Statistics*, 22, 1674–1700.