Published in final edited form as:
Ann Stat. 2008 ; 36(1): 261–286. doi:10.1214/009053607000000604.

Variable Selection in Semiparametric Regression Modeling[1]

RUNZE LI and
Department of Statistics, Pennsylvania State University, University Park, PA16802-2111, rli@stat.psu.edu

HUA LIANG
Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105, hua.liang@stjude.org

Abstract

In this paper, we are concerned with how to select significant variables in semiparametric modeling. Variable selection for semiparametric regression models consists of two components: model selection for nonparametric components and select significant variables for parametric portion. Thus, it is much more challenging than that for parametric models such as linear models and generalized linear models because traditional variable selection procedures including stepwise regression and the best subset selection require model selection to nonparametric components for each submodel. This leads to very heavy computational burden. In this paper, we propose a class of variable selection procedures for semiparametric regression models using nonconcave penalized likelihood. The newly proposed procedures are distinguished from the traditional ones in that they delete insignificant variables and estimate the coefficients of significant variables simultaneously. This allows us to establish the sampling properties of the resulting estimate. We first establish the rate of convergence of the resulting estimate. With proper choices of penalty functions and regularization parameters, we then establish the asymptotic normality of the resulting estimate, and further demonstrate that the proposed procedures perform as well as an oracle procedure. Semiparametric generalized likelihood ratio test is proposed to select significant variables in the nonparametric component. We investigate the asymptotic behavior of the proposed test and demonstrate its limiting null distribution follows a chi-squared distribution, which is independent of the nuisance parameters. Extensive Monte Carlo simulation studies are conducted to examine the finite sample performance of the proposed variable selection procedures.

Keywords

Nonconcave penalized likelihood; SCAD; efficient score; local linear regression; partially linear model; varying coefficient models

1 Introduction

Semiparametric regression models retain the virtues of both parametric and nonparametric modeling. To retain the flexibility of nonparametric models and the explanatory power of generalized linear models, various semiparametric regression models have been proposed in the literature. Partially linear models have been extensively studied (Engle, *et al*., 1986; Heckman, 1986; Chen, 1988; Robin, 1988; Speckman, 1988 and among others). Härdle, Liang and Gao (2000) gave a systematic study for the partially linear models. Generalized partially

[1] Li's research was supported by NSF grant DMS-0348869. Liang's research was partially supported by the American Lebanese Syrian Associated Charities and NIH/NAIAD grant R01 AI62247-01.

linear models were proposed in Severini and Stanliswalis (1994) and Hunsberger (1994), and generalized partially linear single-index models have been studied by many authors (Carroll, *et al*., 1997; Yu and Ruppert, 2002; Liang and Wang, 2003 and among others). Ruppert, Wang and Carroll (2003) and Yatchew (2003) present diverse semiparametric regression models, and their inference procedures and applications. The goal of this paper is to develop effective model selection procedures for a new class of semiparametric regression models, which include many existing semiparametric models as special cases thereof.

Let Y be a response variable and $\{U, \mathbf{X}, \mathbf{Z}\}$ its associated covariates. Denote $\mu(u, \mathbf{x}, \mathbf{z}) = E(Y \mid U = u, \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$. The generalized varying-coefficient partially linear model (GVCPLM) assumes that

$$g\{\mu(u,\mathbf{x},\mathbf{z})\}=\mathbf{x}^{\mathrm{T}}\alpha(u)=\mathbf{z}^{\mathrm{T}}\beta, \tag{1.1}$$

where $g(\cdot)$ is a known link function, $\boldsymbol{\beta}$ is an unknown regression coefficients and $\boldsymbol{\alpha}(\cdot)$ is a vector consisting of unspecified, smoothing regression coefficient functions. Model (1.1) is a semiparametric model, $\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}$ is referred to as parametric component, and $\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}(u)$ as nonparametric component as $\boldsymbol{\alpha}(\cdot)$ is nonparametric. This semiparametric model retains the flexibility of a nonparametric regression model and has the explanatory power of a generalized linear regression model. Many existing semiparametric or nonparametric regression models are special cases of model (1.1). For instance, partially linear models, generalized partially linear models, semi-varying coefficient models (Zhang, Lee, Song, 2002; Xia, Zhang and Tong, 2004; Fan and Huang, 2004), varying coefficient models (Hastie and Tibshirani, 1993; Cai, Fan and Li, 2000) can be written in the form of (1.1). Thus, the newly proposed procedures provides a general framework of model selection for these existing models.

Variable selection is fundamental in statistical modeling. In practice, a number of variables are available to include into an initial analysis, but many of them may not be significant and should be excluded from the final model in order to increase the accuracy of prediction. Variable selection for the GVCPLM is challenging because it includes selection of significant variables in nonparametric component as well as identification of significant variables in parametric component. Traditional variable selection procedures such as stepwise regression and the best subset variable selection for linear models may be extend to the GVCPLM, but it poses great challenges because, for each submodel, it may need to choose smoothing parameters for the nonparametric component. This will dramatically increase computational burden. Furthermore, the traditional variable selection procedures ignore the stochastic error inherited in the selection course, therefore, it is difficult to establish the sampling properties of the resulting estimate, and hard to understand the behavior of the final model. As analyzed by Breiman (1996), the stepwise regression and the best subset selection suffer from several drawbacks, the most severe one of which is the lack of stability, namely, a small perturbation on data may yield a very different model.

In an attempt to select significant variables and estimate unknown regression coefficients simultaneously, Fan and Li (2001) proposed a family of variable selection procedures for parametric models via nonconcave penalized likelihood. This family for linear regression models includes bridge regression (Frank and Friedman, 1993) and LASSO (Tibshirani, 1996). It has been demonstrated that with proper choice of penalty function and regularization parameters, the nonconcave penalized likelihood estimator performs as well as an oracle estimator (Fan and Li, 2001). This encourages us to adopt this methodology for semiparametric regression models. In this paper, we propose a class of variable selection procedures for the parametric component of the GVCPLM. We also study the asymptotic properties of the resulting estimator. We illustrate how the rate of convergence of the resulting estimate depends on the regularization parameters. We further establish the oracle properties of the resulting estimate. Extensive Monte Carlo simulation studies are conducted to assess the finite sample performance of the proposed procedures, and test the accuracy of the standard error formula derived by using sandwich formula.

To select significant variables in the nonparametric component of the GVCPLM, we extend generalized likelihood ratio tests (GLRT, Fan, *et al*., 2001) from fully nonparametric models to semiparametric models. We unveil the Wilks phenomenon in semiparametric modeling: the limiting null distribution of the proposed GLRT does not depend unknown nuisance parameter, and it follows a chi-square distribution with a diveraging degrees of freedom. This allows us to easily obtain critical values for the GLRT either using the asymptotic chi-squares distribution or using bootstrap method.

The paper is organized as follows. In Section 2, we first propose a class of variable selection procedures for the parametric component via nonconcave penalized likelihood approach, and then study the sampling properties of the proposed procedures. In Section 3, variable selection procedures are proposed for the nonparametric component using generalized likelihood ratio (GLR) test. The limiting null distribution of the GLR test is derived. Monte Carlo studies and real data applications are presented in Section 4. Regularity conditions and technical proofs are presented in Section 5.

2 Select significant variables in parametric component

Suppose that $\{U_i, \mathbf{X}_i, \mathbf{Z}_i, Y_i\}$, $i = 1, \cdots, n$, be independent and identically distributed sample, and conditionally on $\{U_i, \mathbf{X}_i, \mathbf{Z}_i\}$, the conditional log-likelihood of Y_i is $\ell_i\{\mu(U_i, \mathbf{X}_i, \mathbf{Z}_i), Y_i\}$. Throughout this paper, $\mathbf{X}_i$ is p-dimensional, $\mathbf{Z}_i$ is d-dimensional, and U is univariate. The methods can be extended for multivariate U in a similar way without essential difficulty. However, the extension may not be very useful in practice due to the "curse of dimensionality".

2.1 Penalized likelihood

Denote by $\ell(\boldsymbol{\alpha}, \boldsymbol{\beta})$ the quasi-likelihood of the collected data $\{(U_i, \mathbf{X}, \mathbf{Z}_i, Y_i), i = 1, \cdots, n\}$. That is,

$$\ell(\alpha,\beta)=\sum_{i=1}^{n}\ell_i[g^{-1}\{\mathbf{X}_i^{\mathrm{T}}\alpha(U_i)+\mathbf{Z}_i^{\mathrm{T}}\beta\},Y_i].$$

Following Fan and Li (2001), define the penalized quasi-likelihood as

$$\mathscr{L}(\alpha,\beta)=\ell(\alpha,\beta)-n\sum_{j=1}^{d}p_{\lambda_j}(|\beta_j|), \tag{2.1}$$

where $p_{\lambda_j}(\cdot)$ is a prespecified penalty function with a regularization parameter λ_j, which can be chosen by a data-driven criterion, such as cross-validation (CV) and generalized cross-validation (GCV, Craven and Wahba, 1979). Note that the penalty functions and regularization parameters are not necessarily the same. For example, we wish to keep some important variables in the final model, and therefore do not want to penalize their coefficients.

Before we pursue further, let us briefly discuss how to select the penalty functions. Various penalty function have been used in the literature of variable selection for linear regression models. Take the penalty function to be the L_0 penalty, namely, $p_{\lambda_j}(|\beta|)=\frac{1}{2}\lambda_j^2 I(|\beta|\neq 0)$, where $I(\cdot)$ is the indicator function. Note that $\sum_{j=1}^{d} I(|\beta_j|\neq 0)$ equals the number of nonzero regression coefficients in the model. Hence many popular variable selection criteria, such as the C_p

(Mallows, 1973), AIC (Akaike, 1974), BIC (Schwarz, 1978), φ-criterion (Hannan and Quinn, 1979, and Shibata, 1984), and RIC (Foster and George, 1994), can be derived from a penalize least squares with the L_0 penalty by choosing different values of λ_j, although these criteria were motivated from different principles. Since the L_0 penalty is discontinuous, it requires an exhaustive search over all possible subsets of predictors to find the solution. This approach is very expensive in computational cost when the dimension d is large. Furthermore, the best subset variable selection suffers from other drawbacks, the most severe of which is its lack of stability as analyzed, for instance, by Breiman (1996).

To avoid the drawbacks of the best subset selection, expensive computational cost and the lack of stability, Tibshirani (1996) proposed the LASSO, which can be viewed as the solution of penalized least squares with the L_1 penalty, defined by $p_{\lambda_j}(|\beta|) = \lambda_j|\beta|$. He further demonstrated that LASSO retains the virtues of both best subset selection and ridge regression. Frank and Friedman (1993) considered the L_q penalty, $p_{\lambda_j}(|\beta|) = \lambda_j|\beta|^q$, $(0 < q < 1)$, which yields a “bridge regression”. The issue of selection penalty function has been studied in depth by various authors, for instance, Antoniadis and Fan (2001). Fan and Li (2001) suggested the use of the smoothly clipped absolute deviation (SCAD) penalty, defined by

$$p'_{\lambda_j}(\beta)=\lambda_j\{I(\beta \le \lambda_j)+\frac{(a\lambda_j-\beta)_+}{(a-1)\lambda_j}I(\beta>\lambda_j)\} \text{ for some } a>2 \text{ and } \beta>0,$$

with $p_{\lambda_j}(0)= 0$. This penalty function involves two unknown parameters λ_j and a. Justifying from a Bayesian statistical point of view, Fan and Li (2001) suggested using $a = 3.7$. The Bayes risk cannot be reduced much with other choices of a, and simultaneous data-driven selection of a and λ_j does not have any significant improvements from our experience.

Since $\boldsymbol{\alpha}(\cdot)$ consists of nonparametric functions, (2.1) is not ready for optimization. We first use local likelihood techniques (Fan and Gijbels, 1996) to estimate $\boldsymbol{\alpha}(\cdot)$, then substitute the resulting estimate into (2.1), and finally maximize (2.1) with respect to $\boldsymbol{\beta}$. Thus, we can obtain a penalized likelihood estimate for $\boldsymbol{\beta}$. With specific choices of penalty functions, the resulting estimate of $\boldsymbol{\beta}$ will contain some exact zero coefficients. This is equivalent to excluding the corresponding variables from the final model. Thus, we achieve the purpose of variable selection.

Specifically, we linearly approximate $\boldsymbol{\alpha}_j(\upsilon)$ for υ in a neighborhood of u by

$$\alpha_j(v) \approx \alpha_j(u)+\alpha'_j(u)(v-u) \equiv a_j+b_j(v-u),$$

Denote $\mathbf{a} = (a_1, \cdots, a_p)^{\mathrm{T}}$ and $\mathbf{b} = (b_1, \cdots, b_p)^{\mathrm{T}}$. Local likelihood method is to maximize the local likelihood function:

$$\sum_{i=1}^{n}\ell_i\left[g^{-1}\left\{\mathbf{a}^T\mathbf{X}_i+\mathbf{b}^T\mathbf{X}_i(U_i-u)+\mathbf{Z}_i^T\beta\right\},Y_i\right]K_h(U_i-u), \tag{2.2}$$

with respect to $\mathbf{a}$, $\mathbf{b}$ and $\boldsymbol{\beta}$, where $K(\cdot)$ is a kernel function, and $K_h(t) = K(t/h)/h$ be a rescaling of K with bandwidth h. Let $\{\tilde{\mathbf{a}}, \tilde{\mathbf{b}}, \tilde{\boldsymbol{\beta}}\}$ be the solution of maximizing (2.2). Then

$$\tilde{\alpha}(u)= \tilde{\mathbf{a}}\ .$$

The local likelihood approach provides us a good estimate for $\boldsymbol{\alpha}$ (See Theorem 1 below), but not for $\boldsymbol{\beta}$. It is usual, the resulting estimate $\tilde{\boldsymbol{\beta}}$ does not have root n convergent rate as $\boldsymbol{\beta}$ was estimated locally. To improve efficiency, $\boldsymbol{\beta}$ should be estimated using global likelihood.

Substituting $\boldsymbol{\alpha}$ in (2.1) by its estimate, we obtain a penalized likelihood

$$\mathscr{L}_P(\beta)=\sum_{i=1}^{n}\ell_i\{g^{-1}(\mathbf{X}_i^{\mathrm{T}}\ \tilde{\alpha}\ (U_i)+\mathbf{Z}_i^{\mathrm{T}}\beta),Y_i\}-n\sum_{j=1}^{d}p_{\lambda_j}(|\beta_j|). \tag{2.3}$$

Maximizing $\mathscr{L}_P(\boldsymbol{\beta})$ results in a penalized likelihood estimator $\hat{\boldsymbol{\beta}}$. The proposed approach is in the same spirit of one-step back-fitting algorithm estimate, although one may further employ back-fitting algorithm method with a full iteration or profile likelihood approach to improve efficiency. Next theorem demonstrates $\hat{\boldsymbol{\beta}}$ performs as well as an oracle estimator in asymptotic sense. Compared with fully iterated back-fitting algorithms and profile likelihood estimate, the newly proposed method is much less computation cost and easily implemented.

2.2 Sampling properties

We first study the sampling properties of the local likelihood estimate $\tilde{\mathbf{a}}$, $\tilde{\mathbf{b}}$ and $\tilde{\boldsymbol{\beta}}$. Let $\boldsymbol{\alpha}_0(\cdot)$ and $\boldsymbol{\beta}$ denote the true value of $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}$, respectively. Define $\kappa_j = \int t^j K(t)dt$, $\nu_j = \int t^j K^2(t)dt$ for $j = 0, 1, 2$, (**Hua: need to modify the notation below by using likelihood, rather than quasi-likelihood**.) $\rho_k(t)=\left\{\frac{dg^{-1}(t)}{dt}\right\}^k/[\sigma^2V\{g^{-1}(t)\}]$, $k = 1,2, q_1(x,y) = \{y - g^{-1}(x)\}\rho_1(x)$, $m_i=m_i(U_i,\mathbf{X}_i)=\alpha_0^T(U_i)\mathbf{X}_i+\mathbf{Z}_i^T\beta_0$ and

$$\sum(u)=E\left[\rho_2\left\{\alpha_0^T(U)\mathbf{X}+\beta_0^T\mathbf{Z}\right\}\begin{pmatrix}\mathbf{XX}^T & \mathbf{XZ}^T\\ \mathbf{ZX}^T & \mathbf{ZZ}^T\end{pmatrix}|U=u\right], \tag{2.4}$$

Theorem 1—*Consider the maximizer of the local likelihood* (*2.2*)*. Then, as n* → ∞*, h* → 0 *and nh* → ∞*, under Condition 1 in the appendix,*

$$\sup_{u\in D}\|\ \tilde{\alpha}\ (u)-\alpha_0(u)-\frac{1}{nf(u)}\sum_{i=1}^{n}\tilde{W}_iK_h(U_i-u)\|=O_P\left\{h^2c_n+c_n^2\log^{1/2}(1/h)\right\}, \tag{2.5}$$

where $\|\cdot\|$ *is the Euclidean norm,* $f(\cdot)$ *the density of* U, $\tilde{W}_i$ *is a p-vector consisted of the first p elements of* $q_1(\overline{\alpha}_i,Y_i)\sum^{-1}(u)(\mathbf{X}_i^{\mathrm{T}},\mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}}$ *with* $\overline{\alpha}_i=\overline{\alpha}_i(u)=\mathbf{X}^{\mathrm{T}}\alpha_0(u)+\mathbf{Z}_i^{\mathrm{T}}\beta_0+(U_i-u)\mathbf{X}_i^{\mathrm{T}}\alpha_0'(u)$. What is D here? Should we use Ω for the support of U? We have used D for other notation.

We next investigate the sampling properties of $\hat{\boldsymbol{\beta}}$. Let $\beta_0=(\beta_{10},\cdots,\beta_{d0})^{\mathrm{T}}=(\beta_{10}^{\mathrm{T}},\beta_{20}^{\mathrm{T}})^{\mathrm{T}}$. For ease of presentation and without loss of generality, it is assumed that $\boldsymbol{\beta}_{10}$ consists of all nonzero components of $\boldsymbol{\beta}_0$, and $\boldsymbol{\beta}_{20} = \mathbf{0}$. Denote

$$a_n=\max_{1\le j\le d}\{|p'_{\lambda_j}(|\beta_{j0}|)|,\beta_{j0}\neq 0\},\ \text{and}\ b_n=\max_{1\le j\le d}\{|p''_{\lambda_j}(|\beta_{j0}|)|,\beta_{j0}\neq 0\}. \tag{2.6}$$

Theorem 2—*Suppose Condition 1 in Appendix holds, and* $nh^4 \to 0$ *and* $nh^2/\log(1/h) \to \infty$ *as n* → ∞*. There exists a local maximizer* $\hat{\boldsymbol{\beta}}$ *of* $\mathscr{L}_P(\boldsymbol{\beta})$ *defined in* (*2.3*) *such that its rate of convergence is* $O_P(n^{-1/2} + a_n)$*, where* a_n *is given in* (*2.6*).

Define

$$\mathbf{b}_n=\{p'_{\lambda_1}(|\beta_{10}|)\mathrm{sgn}(\beta_{10}),\cdots,p'_{\lambda_s}(|\beta_{s0}|)\mathrm{sgn}(\beta_{s0})\}^{\mathrm{T}},\text{and}\sum\nolimits_{\lambda}=\mathrm{diag}\{p''_{\lambda_1}(|\beta_{10}|),\cdots,p''_{\lambda_s}(|\beta_{s0}|)\}. \tag{2.7}$$

where s is the number of nonzero components of $\boldsymbol{\beta}_0$.

Theorem 3—*Suppose Condition 1 in Appendix holds, and* $nh^4 \to 0$ *and* $nh^2/\log(1/h) \to \infty$ *as n* → ∞*. The root n consistent estimator* $\hat{\boldsymbol{\beta}}$ *in Theorem 2 must satisfy that* $\hat{\boldsymbol{\beta}}_2 = 0$*, and*

$$\sqrt{n}(\mathbf{B}_1+\sum\nolimits_{\lambda})\{\widehat{\beta}_1-\beta_{10}+(\mathbf{B}_1+\sum\nolimits_{\lambda})^{-1}\mathbf{b}_n\}\xrightarrow{D}N(0,\sum),$$

where $\mathbf{B}_1=\left[\rho_2\{\alpha_0^T(U)\mathbf{X}+\mathbf{Z}_1^T\beta_{10}\}\mathbf{Z}_1\mathbf{Z}_1^T\right]$ *and* $\Sigma = var\{q_1(m_1, Y)\mathbf{Z}_1-\Gamma_1(U)\}$,

$$\Gamma_1(u)=\sum_{k=1}^{p}\nu_k E\left[\rho_2\{\alpha_0^T(U)\mathbf{X}+\mathbf{Z}_1^T\beta_{10}\}X_k\mathbf{Z}_1|U=u\right]$$

with ν_k *the k–th element of* $q_1(m_1,Y)\sum^{-1}(u)(\mathbf{X}_i^T,\mathbf{Z}_i^T)^T$.

Theorem 3 indicates that undersmoothing is necessary in order for $\hat{\boldsymbol{\beta}}$ to have root n consistency and asymptotic normality. This is a standard case in the generalized partially linear models. See Carroll *et al*.1997 for a detailed discussion. Thus, a special care is needed for bandwidth selection, which is discussed in next section.

2.3 Bandwidth selection

From the proof of Theorem 1, $\hat{\boldsymbol{\alpha}}(u)$ has conditional asymptotic bias

$$0.5h^2\mu_2\alpha''(u)+o_P(h^2),$$

and conditional asymptotic covariance

$$(nh)^{-1}\nu_0 D^{-1}(u)f^{-1}(u)+o_P(\frac{1}{nh}).$$

A theoretic optimal local bandwidth for estimating the elements of $\boldsymbol{\alpha}(\cdot)$ can be obtained by minimizing the conditional mean squared error (MSE) given by

$$E\{\|\widehat{\alpha}(u)-\alpha(u)\|^2\mathbf{Z},\mathbf{X}\}=\frac{1}{4}h^4\mu_2^2\|\alpha''(u)\|^2+\frac{1}{nh}\frac{\nu_0\mathrm{tr}\{D^{-1}(u)\}}{f(u)}+o_P(h^4+\frac{1}{nh}),$$

where ||·|| is the Euclidean distance. Thus, the ideal choice of a local bandwidth is

$$\widehat{h}_{opt}=\left\{\frac{\nu_0\mathrm{tr}\{D^{-1}(u)\}}{f(u)\mu_2^2\|\alpha''(u)\|^2}\right\}^{1/5}n^{-1/5}.$$

With expressions of the asymptotic bias and variance, we can also derive a theoretic or data-driven global bandwidth selector by utilizing the existing bandwidth selection techniques for the canonical univariate nonparametric model, such as substitution method (See, for instance, Ruppert, Sheather and Wand, 1995). To save space, we omit the details here.

As usual, the optimal bandwidth will be of order $n^{-1/5}$. This does not satisfy the condition in Theorems 2 and 3. A relatively appropriate bandwidth is generally generated by $\hat{h}_{opt}\times n^{-2/15} = O(n^{-1/3})$.

In order for the resulting variable selection procedures to possess an oracle property, it requires the bandwidth satisfying that $nh^8\to 0$ and $nh^2/(\log n)^2\to\infty$. The order of bandwidth aforementioned satisfies these requirement. This enables us to easily choose a bandwidth by either data-driven procedures or asymptotic theory based method.

2.4 Issues in practical implementation

Local quadratic algorithm—The penalty function $p_{\lambda_j}(|\beta_j|)$ including the L_1 penalty and the SCAD penalty is irregular at the origin and may not have the second derivative at some points. Direct implementation of the Newton-Raphson algorithm may be difficult. Following Fan and Li (2001), we locally approximate the penalty function by a quadratic function every step during the course of iteration as follows. Given an initial value $\boldsymbol{\beta}^{(0)}$ that is close to the maximizer of the penalized likelihood function, when $\beta_j^{(0)}$ is not very close to 0, the penalty $p_{\lambda_j}(|\beta_j|)$ can be locally approximated by the quadratic function as

$$[p_{\lambda_j}(|\beta_j|)]'=p'_{\lambda_j}(|\beta_j|)\mathrm{sgn}(\beta_j) \approx \{p'_{\lambda_j}(|\beta_j^{(0)}|)/\beta_j^{(0)}|\}\beta_j,$$

otherwise, set $\hat{\beta}_j = 0$. In other words,

$$p_{\lambda_j}(|\beta_j|) \approx p_{\lambda_j}(|\beta_j^{(0)}|)+\frac{1}{2}\{p'_{\lambda_j}(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}(\beta_j^2-\beta_j^{(0)2}) \ \text{for}\, \beta_j \approx \beta_j^{(0)}.$$

For instance, this local quadratic approximation for the L_1 penalty yields

$$|\beta_j| \approx \frac{1}{2}|\beta_j^{(0)}|+\frac{1}{2}\frac{\beta_j^2}{|\beta_j^{(0)}|} \ \text{for}\, \beta_j \approx \beta_j^{(0)}.$$

With the aid of the local quadratic approximation, the Newton-Raphson algorithm can be employed for searching the solution of the penalized likelihood.

Standard Error formula for $\hat{\boldsymbol{\beta}}$—The standard errors for estimated parameters can be directly obtained because we are estimating parameters and selecting variables at the same time. Following the conventional technique in the likelihood setting, the corresponding sandwich formula can be used as an estimator for the covariance matrix of the estimates $\hat{\boldsymbol{\beta}}$. Specifically, denote

$$\ell'(\widehat{\alpha},\beta)=\frac{\partial\ell(\widehat{\alpha},\beta)}{\partial\beta},\ \ell''(\widehat{\alpha},\beta)=\frac{\partial^2\ell(\widehat{\alpha},\beta)}{\partial\beta\partial\beta^{\mathrm{T}}},\ \text{and}\ \sum\nolimits_{\lambda}(\beta)=\mathrm{diag}\left\{p'_{\lambda_1}(|\beta_1|)/|\beta_1|,\ldots,p'_{\lambda_d}(|\beta_d|)/|\beta_d|\right\}.$$

Then the corresponding sandwich formula is given by

$$\widehat{\mathrm{cov}}(\widehat{\beta})=\{\ell''(\widehat{\beta})-n\sum\nolimits_{\lambda}(\widehat{\beta})\}^{-1}\widehat{\mathrm{cov}}\{\ell'(\widehat{\beta})\}\{\ell''(\widehat{\beta})-n\sum\nolimits_{\lambda}(\widehat{\beta})\}^{-1}.$$

This formula can be shown to be consistent estimator and will be shown to have good accuracy for moderate sample sizes.

Choice of $\boldsymbol{\lambda_j}$'s—We suggest selecting the tuning parameters λ_j's using data-driven approaches. Similarly to Fan & Li (2001), we will employ the generalized cross validation (GCV) to select the λ_j's. In the last step of the Newton-Raphson iteration, we may compute the effective number of parameters:

$$e(\lambda_1,\cdots,\lambda_d)=\mathrm{tr}[\left\{\ell''(\widehat{\alpha},\widehat{\beta})-n\sum\nolimits_{\lambda}(\widehat{\beta})\right\}^{-1}\ell''_P(\widehat{\alpha},\widehat{\beta})].$$

The GCV statistic is defined by

$$\mathrm{GCV}(\lambda_1,\cdots,\lambda_d)=\frac{D\left\{\widehat{\beta}(\lambda)\right\}}{n\{1-e(\lambda_1,\cdots,\lambda_d)/n\}^2}.$$

where $D\{\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})\}$ stands for sum of deviance residuals corresponding to the fitting with $\boldsymbol{\lambda}$. The minimization problem over a d-dimensional space is difficult. However, it is expected that the magnitude of λ_j should be proportional to the standard error of the unpenalized maximum pseudo-partial likelihood estimator of β_j. In practice, we suggest taking $\lambda_j=\lambda SE(\widehat{\beta}_j^u)$, where $SE(\widehat{\beta}_j^u)$ is the estimated standard error of $\widehat{\beta}_j^u$, the unpenalized likelihood estimate. Such a choice of λ_j works well from our simulation experience. Thus, the minimization problem will reduce to a one-dimensional problem, and the tuning parameter can be estimated by a grid search.

3 Statistical inferences for nonparametric components

3.1 Estimation for $\boldsymbol{\alpha}$

From the proof of Theorem 1, it has been shown that $\hat{\boldsymbol{\alpha}}(u)$ has the following asymptotic expansion:

$$\tilde{\alpha}\,(u)-\alpha_0(u)=(\kappa_2/2)\alpha_0''(u)h^2+\frac{1}{nf(u)}\sum_{i=1}^{n}W_iK_h(U_i-u)+o_P\left\{(nh)^{-1/2}+h^2\right\},$$

and hence if $nh^5 \to 0$, then

$$(nh)^{1/2}\left\{\tilde{\alpha}\,(u)-\alpha_0(u)-\frac{\kappa_2}{2}\alpha_0''(u)h^2\right\}\xrightarrow{D}N\left\{\mathbf{0},\frac{\nu_0}{f(u)}\,\tilde{D}\,(u)\right\},$$

where $\tilde{D}(u)$ is the left-upper $p \times p$ sub-matrix of the matrix $\boldsymbol{\Sigma}^{-1}(u)$. Thus, the efficiency of $\tilde{\boldsymbol{\alpha}}$ can be improved. Replacing $\boldsymbol{\beta}$ in (2.2) by its estimate $\hat{\boldsymbol{\beta}}$, we maximize the following local likelihood function

$$\sum_{i=1}^{n}\ell_i\left[g^{-1}\left\{\mathbf{a}^T\mathbf{X}_i+\mathbf{b}^T\mathbf{X}_i(U_i-u)+\mathbf{Z}_i^T\widehat{\beta}\right\},Y_i\right]K_h(U_i-u), \tag{3.1}$$

with respect to $\mathbf{a}$ and $\mathbf{b}$. Let $\{\hat{\mathbf{a}}, \hat{\mathbf{b}}\}$ be the solution of maximizing (3.1). Then

$$\widehat{\alpha}(u)=\widehat{\mathbf{a}}.$$

This provides us an efficient estimate for $\boldsymbol{\alpha}$(Cai, Fan and Li, 2000). (**Hua: I think, the asymptotic normality of Cai, Fan and Li (2000) will be valid here provided that $\hat{\boldsymbol{\beta}}$ is root n consistent**.).

3.2 Variable selection for nonparametric component

After obtaining nonparametric estimates of $\{\alpha_1(\cdot), \cdots, \alpha_p(\cdot)\}$, it is of interest to test the significance of the variable X_j. It is also of interest to test whether the coefficient $\alpha_j(u)$ varies over U or is just a constant. This leads us to consider hypothesis testing problems such as

$$H_0:\alpha_j(u)=a_j(u,\theta),\ \text{vs}\ H_1:\alpha_j(u)\neq a_j(u,\theta),$$

where $a_j(u, \theta)$ is a pre-specified parametric function with an unknown parameter vector θ. For example, $a_j(u, \theta) = 0$ for testing the significant of variable X_j, and $a_j(u, \theta) = \theta_j$ for testing whether $a_j(u, \theta)$ is a constant. For simplicity of presentation, we consider the following null hypothesis

$$H_0{:}\alpha_1(u)=\theta_1, \cdots, \alpha_d(u)=\theta_d, \text{ vs } H_1{:}\alpha_1(u) \neq \theta_1, \cdots, \alpha_d(u) \neq \theta_d,$$

where θ_j is an unknown constant. The proposed idea is also applicable for more general cases.

Let $\hat{\alpha}(u, \beta)$ and $\hat{\beta}_{\text{GPLM}}$ be the estimators of $\alpha(u)$ and β under the alternative hypothesis, and $\hat{\theta}_j$ and $\hat{\beta}_{\text{GLM}}$ the estimators of θ_j and β under the null hypothesis. Denote

$$\mathscr{R}(H_1)=\sum_{i=1}^{n} Q\{g^{-1}(\widehat{\alpha}^{\mathrm{T}}(U_i)\mathbf{X}_i^{\mathrm{T}}+\mathbf{Z}_i^{\mathrm{T}}\widehat{\beta}_{\mathrm{GPLM}}),Y_i\}$$

and

$$\mathscr{R}(H_0)=\sum_{i=1}^{n} Q\{g^{-1}(\widehat{\theta}^{\mathrm{T}}\mathbf{X}_i+\mathbf{Z}_i^{\mathrm{T}}\widehat{\beta}_{\mathrm{GLM}}),Y_i\}$$

Following Fan, Zhang, and Zhang (2001), we define a generalized quasi-likelihood ratio test statistic

$$T_{\mathrm{GLR}}=r_K\,\{\mathscr{R}(H_1)-\mathscr{R}(H_0)\},$$

where

$$r_K+\{K(0)-0.5\int K^2(u)du\}\Big\{\int\{K(u)-0.5K*K(u)\}\,du\Big\}^{-1}.$$

Theorem 4—Suppose Condition 1 hold and $nh^8 \to 0$ and $nh^2/(\log n)^2 \to \infty$. Under H_0, the test statistic T_{GLR} has an asymptotic χ^2 distribution with δ_n degrees of freedom in the sense of Fan, Zhang, and Zhang (2001), where $\delta_n = r_K p|D|\{K(0) - 0.5\int K^2(u)\, du\}/h$, and $|D|$ stands for the length of the support of U.

Theorem 4 unveils a new Wilks phenomenon for semiparametric inference and extends the generalized likelihood ratio theory (Fan, Zhang, and Zhang 2001) for semiparametric modeling. We will also provide empirical justification to the null distribution. Similar to Cai, Fan and Li (2000), the null distribution of T_{GLR} can be estimated using Monte Carlo simulation or a bootstrap procedure. This usually provides a better estimate than the asymptotic null distribution, since the degrees of freedom tend to infinite and the results in Fan, Zhang, and Zhang (2001) give only the main order of the degrees of freedom.

4 Simulation studies and applications

In this section, we conduct extensive Monte Carlo simulations to examine finite sample performance of the proposed procedures.

Root of average square errors

The performance of estimator $\hat{\boldsymbol{\alpha}}(\cdot)$ will be assessed by using the square-Root of Average Square Errors (RASE)

$$\text{RASE}^2=n_{\text{grid}}^{-1}\sum_{j=1}^{p}\sum_{k=1}^{n_{grid}}\{\widehat{\alpha}_j(u_k)-\alpha_j(u_k)\}^2, \tag{4.1}$$

where $\{u_k, k = 1, \cdots, n_{grid}\}$ are the grid points at which the functions $\{\alpha_j(\cdot)\}$ are estimated. In our simulation, the Epanechnikov kernel $K(u) = 0.75(1 - u^2)_+$ and $n_{\text{grid}} = 200$ are used. Three bandwidths will be employed to represent widely varying degrees of smoothness. Over this range of bandwidths, we assess the performance of the estimator $\hat{\boldsymbol{\alpha}}(\cdot)$.

Prediction error, model error and generalized mean squared error

The prediction error is defined as the average error in the prediction of the dependent variable given the independent variables for future cases that are not used in the construction of a prediction equation. Let $\{U^*, \mathbf{X}^*, \mathbf{Z}^*, Y^*\}$ be a new observation from the GVCPLM model (1.1). Then the prediction error for model (1.1) is

$$\mathrm{PE}(\widehat{\alpha},\widehat{\beta})=E(Y^* - \widehat{\mu}(U^*,\mathbf{X}^*,\mathbf{Z}^*)\}^2$$

where the expectation is a conditional expectation given the data used in constructing the prediction procedure. The prediction error can be decomposed as

$$\mathrm{PE}(\widehat{\alpha},\widehat{\beta})=E\{Y^* - \mu(U^*,\mathbf{X}^*,\mathbf{Z}^*)\}^2+E\{\widehat{\mu}(U^*,\mathbf{X}^*,\mathbf{Z}^*) - \mu(U^*,\mathbf{X}^*,\mathbf{Z}^*)\}^2$$

The first component is the inherent prediction error due to noise. The second one is due to lack of fit with an underlying model. This component is termed *model error*. Note that $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$ are consistent estimate and $\mu(U^*, \mathbf{X}^*, \mathbf{Z}^*) = g^{-1}\{\mathbf{Z}^{*\mathrm{T}}\boldsymbol{\alpha}(U^*) + \mathbf{Z}^{*\mathrm{T}}\boldsymbol{\beta}\}$. By the Taylor expansion, we have the following approximation

$$\widehat{\mu}(U^*,\mathbf{X}^*,\mathbf{Z}^*) \approx \mu(U^*,\mathbf{X}^*,\mathbf{Z}^*)+\dot{g}^{-1}\{\mathbf{X}^{*\mathrm{T}}\alpha(U^*)+\mathbf{Z}^{*\mathrm{T}}\beta\}\mathbf{X}^{*\mathrm{T}}\{\widehat{\alpha}(U^*) - \alpha(U^*)\}$$
$$+\dot{g}^{-1}\{\mathbf{X}^{*\mathrm{T}}\alpha(U^*)+\mathbf{Z}^{*\mathrm{T}}\beta\}\mathbf{Z}^{*\mathrm{T}}(\widehat{\beta} - \beta),$$

where $\dot{g}^{-1}(t) = dg^{-1}(t)/dt$. Therefore the model error can be approximated by

$$E[\dot{g}^{-1}\{\mathbf{X}^{*\mathrm{T}}\alpha(U^*)+\mathbf{Z}^{*\mathrm{T}}\beta\}]^2\Big([\mathbf{X}^{*\mathrm{T}}\{\widehat{\alpha}(U^*) - \alpha(U^*)\}]^2$$
$$+[\mathbf{Z}^{*\mathrm{T}}(\widehat{\beta} - \beta)]^2+[\mathbf{X}^{*\mathrm{T}}\{\widehat{\alpha}(U^*) - \alpha(U^*)\}] \times [\mathbf{Z}^{*\mathrm{T}}(\widehat{\beta} - \beta)]\Big)$$

The first component is the inherent model error due to lack of fit of the nonparametric component $\boldsymbol{\alpha}_0(t)$, the second one is due to lack of fit of the parametric component, and the third one is the cross-product between the first two components. Thus, we define generalized mean square error (GMSE) for the parametric component as

$$\mathrm{GMSE}(\widehat{\beta})=E[\mathbf{Z}^{*\mathrm{T}}(\widehat{\beta} - \beta)]^2=(\widehat{\beta} - \beta)E(\mathbf{Z}^*\mathbf{Z}^{*\mathrm{T}})(\widehat{\beta} - \beta), \tag{4.2}$$

and use the GMSE to assess the performance of the newly proposed variable selection procedures for the parametric component.

Example 4.1—In this example, we consider semi-varying logistic regression model. Given U, $\mathbf{X}$, $\mathbf{Z}$, Y has a Bernoulli distribution with success probability $p(U, \mathbf{X}, \mathbf{Z})$, and

$$p(U,\mathbf{X},\mathbf{Z})=\exp\{\mathbf{X}^{\mathrm{T}}\alpha(U)+\mathbf{Z}^{\mathrm{T}}\beta\}/[1+\exp\{\mathbf{X}^{\mathrm{T}}\alpha(U)+\mathbf{Z}^{\mathrm{T}}\beta\}]$$

In our simulation, we take $U \sim U(0, 1)$, $\mathbf{X} = (1, X_1)^{\mathrm{T}}$ with $X_1 \sim N(0, 1)$, and the coefficient functions are given by

$$\alpha_0(u)=\exp(2u - 1), \text{ and } \quad \alpha_1(u)=8u(1 - u).$$

Furthermore, $\boldsymbol{\beta} = [3, 1.5, 0, 0, 2, 0, 0, 0]^{\mathrm{T}}$, $\mathbf{Z}$ to be a 8-dimensional multinormal distribution with zero mean and covariance matrix (σ_{ij}) with $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.5$ for $i \neq j$.

Example 4.2—In this example, we consider semi-varying Poisson regression model. Given U, $\mathbf{X}$, $\mathbf{Z}$, Y has a Poisson distribution with mean function $\mu(U, \mathbf{X}, \mathbf{Z})$, and

$$\mu(U,\mathbf{X},\mathbf{Z})=\exp\{\mathbf{X}^{\mathrm{T}}\alpha(U)+\mathbf{Z}^{\mathrm{T}}\beta\}.$$

In our simulation, U, $\mathbf{X}$, $\mathbf{Z}$ are the same as those in Example 4.1. $\boldsymbol{\beta} = [3, 1.5, 0, 0, 2, 0, 0, 0]^{\mathrm{T}}$, and the coefficient functions are given by

$$\alpha_0(u)=\exp(2u-1), \quad \text{and} \quad \alpha_1(u)=8u(1-u).$$

Example 4.3—Now we apply the methodology proposed in this paper to analyze the data set: *Burns data*, collected by General Hospital Burn Center at the University of Southern California. The binary response variable Y is 1 for those victims who survived their burns and 0 otherwise, and covariates X_1=*age*, X_2=*sex*, X_3 = log(burn area+1) and binary variable X_4 =*Oxygen* (0 if oxygen supply is normal, 1 otherwise) are considered. We are interested in studying how burn areas and the other variables affect survival probabilities for victims at different age groups.

More variable will be added.

Example 4.4—We illustrate in this example our proposed procedure via an application to the environmental data set mentioned in the introduction. Of interest is to study the association between levels of pollutants and number of total hospital admissions for circulatory and respiratory problems from January 1, 1994 to December 31, 1995 and to examine the extent to which the association varies over time. The covariates are taken as the levels of pollutants sulfur dioxide X_2 (in $\mu g/m^3$), nitrogen dioxide X_3 (in $\mu g/m^3$) and dust X_4 (in $\mu g/m^3$). Since the admissions "events" occur at certain points in time, it is reasonable to model the number of admissions as a Poisson process and use the Poisson regression model with the mean $\lambda(t, \mathbf{x})$ given by

$$\log\{\lambda(t,\mathbf{x})\}=a_1(t)+a_2(t)x_2+a_3(t)x_3+a_4(t)x_4.$$

5 Conditions and Proofs

5.1 Conditions

For simplicity of notation, in this appendix we absorb σ^2 into $V(\cdot)$, so that the variance of Y given $(U, \mathbf{X}, \mathbf{Z})$ is $V\{\mu(U, \mathbf{X}, \mathbf{Z})\}$. Denote $q_\ell(x, y) = (\partial^\ell/\partial x^\ell)Q\, g^{-1}(x), y$ for $\ell = 1, 2, 3$. Then

$$q_1(x,y)=\left\{y-g^{-1}(x)\right\}\rho_1(x) \;\text{ and }\; q_2(x,y)=\left\{y-g^{-1}(x)\right\}\rho_1'(x)-\rho_2(x), \tag{5.1}$$

where $\rho_\ell(t)=\left\{\frac{dg^{-1}(t)}{dt}\right\}^\ell/V\{g^{-1}(t)\}$ is introduced in Section 2. In Condition 1, u is a generic argument for Theorem 1, and the condition must hold *uniformly* in u for Theorems 1 – 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 random variable U has a bounded support D. the elements of function $\alpha_0''(\cdot)$ are continuous in $u \in D$.

c. The density function $f(u)$ of U has a continuous second derivative,

d. The functions $V''(\cdot)$ and $g'''(\cdot)$ are continuous.

e. With $R{=}\alpha_0^T(U)\mathbf{X}{+}\mathbf{Z}^T\beta_0, E\{q_1^2(R,Y)|U{=}u\}, E\{q_1^2(R,Y)\mathbf{Z}|U{=}u\}$, and $E\{q_1^2(R,Y)\mathbf{Z}\mathbf{Z}^T|U{=}u\}$ are twice differentiable in $u \in D$. Moreover, $E\{q_2^2(R,Y)\}{<}\infty$ and $E\left\{q_1^{2+\delta}(R,Y)\right\}{<}\infty$ 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 1 (i) is imposed so that the local likelihood is concave in the parameters, which ensures the uniqueness of the solution. 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.

5.2 Two Lemmas

Lemma A.1—(Carroll *et al.* 1997) 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 $\hat{\boldsymbol{\theta}}(\mathbf{x}) \in C$ 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

$$\sup_{\theta\in C,\mathbf{x}\in D}|f_n(\mathbf{x},\theta)-f(\mathbf{x},\theta)|\to 0,\ \text{then}\ \sup_{\mathbf{x}\in D}|\widehat{\theta}_n(\mathbf{x})-\widehat{\theta}(\mathbf{x})|\to 0,\ \text{as}\ n\to\infty.$$

Lemma A.2—(Mack and Silverman 1982) 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 $E|Y|^r < \infty$ and $\sup_{\mathbf{x}} \int |y|^r f(\mathbf{x}, y)\, dy < \infty$ 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,

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

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

5.3 Proof of Theorem 1

Throughout this proof, terms of the form $\hat{G}(u) = O_P(a_n)$ are always meant as $\sup_{u\in D}|\hat{G}(u)| = O_P(a_n)$.

Let $c_n = (nh)^{-1/2}$,

$$\mathbf{X}_i^*{=}\begin{pmatrix}\mathbf{X}_i\\(U_i-u)\mathbf{X}_i/h\\\mathbf{Z}_i\end{pmatrix},\ \widehat{\beta}^*{=}\begin{pmatrix}c_n^{-1}\{\widehat{\mathbf{a}}-\alpha_0(u)\}\\c_n^{-1}h\left\{\widehat{b}-\alpha_0'(u)\right\}\\c_n^{-1}(\widehat{\beta}-\beta_0)\end{pmatrix}.$$

If $(\hat{\mathbf{a}}, \hat{\mathbf{b}}, \hat{\boldsymbol{\beta}})^T$ maximizes (2.2) then $\hat{\boldsymbol{\beta}}^*$ maximizes

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

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

$$\ell_n(\beta^*)=\mathbf{W}_n^T\beta^*+\tfrac{1}{2}\beta^{*T}\mathbf{A}_n\beta^*\{1+o_P(1)\}, \tag{5.2}$$

where

$$\mathbf{W}_n=hc_n\sum_{i=1}^n q_1(\overline{\alpha}_i,Y_i)\mathbf{X}_i^*K_h(U_i-u) \text{ and } \mathbf{A}_n=hc_n^2\sum_{i=1}^n q_2(\overline{\alpha}_i,Y_i)\mathbf{X}_i^*\mathbf{X}_i^{*T}K_h(U_i-u).$$

Define

$$\mathbf{A}(\mathbf{X},\mathbf{Z})=\begin{pmatrix} \mathbf{X}\mathbf{X}^T & \mathbf{0}^T & \mathbf{X}\mathbf{Z}^T \\ \mathbf{0} & \kappa_2\mathbf{X}\mathbf{X}^T & \mathbf{0} \\ \mathbf{Z}\mathbf{X}^T & \mathbf{0} & \mathbf{Z}\mathbf{Z}^T \end{pmatrix};\ \mathbf{B}(\mathbf{X},\mathbf{Z})=\begin{pmatrix} \nu_0\mathbf{X}\mathbf{X}^T & \mathbf{0} & \nu_0\mathbf{X}\mathbf{Z}^T \\ 0 & \nu_2\mathbf{X}\mathbf{X}^T & 0 \\ \nu_0\mathbf{Z}\mathbf{X}^T & 0 & \nu_0\mathbf{Z}\mathbf{Z}^T \end{pmatrix}.$$

It can be shown that

$$\mathbf{A}_n=-f(u)E\left[\rho_2(\alpha_0^T(U)\mathbf{X}+\mathbf{Z}^T\beta_0)\mathbf{A}(\mathbf{X},\mathbf{Z})|U=u\right]+o_P(1)\equiv-\mathbf{A}+o_P(1). \tag{5.3}$$

Therefore, by (5.2),

$$\ell_n(\beta^*)=\mathbf{W}_n^T\beta^*-\tfrac{1}{2}\beta^{*T}\mathbf{A}\beta^*+o_P(1). \tag{5.4}$$

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 (5.4) 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

$$\sup_{u\in D}|\widehat{\beta}^*-\mathbf{A}^{-1}\mathbf{W}_n|\xrightarrow{P}0. \tag{5.5}$$

Furthermore, we have, from the definition of $\hat{\boldsymbol{\beta}}^*$, that

$$\frac{\partial}{\partial\beta^*}\ell_n(\beta^*)|_{\beta^*=\widehat{\beta}^*}=c_nh\sum_{i=1}^n q_1(\overline{\alpha}_i+c_n\widehat{\beta}^{*\mathrm{T}}\mathbf{X}_i^*,Y_i)\mathbf{X}_i^*\mathbf{X}_i^{*T}K_h(U_i-u)\widehat{\beta}^*=0.$$

By using (5.5) and a Taylor expansion, we have

$$\mathbf{W}_n+\mathbf{A}_n\widehat{\beta}^*+\frac{c_n^3h}{2}\sum_{i=1}^n q_3(\overline{\alpha}_i+\widehat{\zeta}_i,Y_i)\mathbf{X}_i^*\left\{\widehat{\beta}^{*\mathrm{T}}\mathbf{X}_i^*\right\}^2K_h(U_i-u)=0, \tag{5.6}$$

where $\hat{\zeta}_i$ is between 0 and $c_n\widehat{\beta}^{*\mathrm{T}}\mathbf{X}_i^*$. The last term in the above expression is of order $O_P(c_n\|\hat{\boldsymbol{\beta}}^*\|^2)$. Since each element in $\mathbf{A}_n$ is of a kernel form, we can deduce from Lemma A.2 that

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

Consequently, by (5.6) we obtain that

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

Hence,

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

A direct calculation yields that

$$\begin{aligned} E\mathbf{W}_n= & \ c_n^{-1}E[\,E\{q_1(\overline{\alpha},Y)\mathbf{X}^{*}K_h(U-u)|\mathbf{X},\mathbf{Z},\mathbf{X}\}] \\ = & \ c_n^{-1}\tfrac{1}{2}\alpha_0^{\prime\prime,T}(u)h^2f(u)E\left[\mathbf{X}\rho_2\left\{\alpha_0^T(U)\mathbf{X}+\mathbf{Z}^T\beta_0\right\}(\kappa_2\mathbf{X}^T,\mathbf{0},\kappa_2\mathbf{Z}^T)^T\mathbf{X}|U=u\right]+o(c_n^{-1}h^2); \end{aligned}$$

and

$$\text{var}(\mathbf{W}_n)=f(u)E\left[\rho_2\left\{\alpha_0^T(U)\mathbf{X}+\mathbf{Z}^T\beta_0\right\}B(\mathbf{X},\mathbf{Z})|U=u\right]+o(1)$$

By taking the first p elements of the expression we complete the proof of Theorem 1.

5.4 Proof of Theorem 3

The proof of (A) is similar to that of Theorem 1 of Fan and Li (2001). Let $\alpha_n = n^{-1/2} + a_n$. We show that for any given $\zeta > 0$, there exists a large constant C such that

$$P\left\{\sup_{||\mathbf{v}||=C}\mathscr{L}_P(\beta_0+\alpha_n\mathbf{v})<\mathscr{L}_P(\beta_0)\right\}\geq 1-\zeta. \tag{5.7}$$

Note that

$$\begin{aligned} D_n(\mathbf{v})\equiv\mathscr{L}_P(\beta_0= & \ \alpha_n\mathbf{v})-\mathscr{L}_P(\beta_0) \\ \leq & \ \sum_{i=1}^{n}\left[Q\{g^{-1}(\widehat{\alpha}^{\mathrm{T}}(U_i)\mathbf{X}_i+\mathbf{Z}_i^{\mathrm{T}}(\beta_0+\alpha_n\mathbf{v})),Y_i\}-Q\{g^{-1}(\widehat{\alpha}^{\mathrm{T}}(U_i)\mathbf{X}_i+\mathbf{Z}_i^{\mathrm{T}}\beta_0),Y_i\}\right] \\ & -n\sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_{j0}+\alpha_nv_j|)-p_{\lambda_n}(|\beta_{j0}|)\}, \end{aligned}$$

where s is the number of components of $\boldsymbol{\beta}_{10}$.

The second term is exactly same as in Fan and Li (2001). We deal with the first term. Let $\widehat{m}_i=\widehat{\alpha}^T(U_i)\mathbf{X}_i+\mathbf{Z}_i^T\beta_0$, and $m_i=\alpha_0^T(U_i)\mathbf{X}_i+\mathbf{Z}_i^T\beta_0$. Then, the first term equals to

$$\ell_n(\theta)=\sum_{i=1}^{n}\left[Q\left\{g^{-1}(\widehat{m}_i+\alpha_n\mathbf{v}^T\mathbf{Z}_i),Y_i\right\}-Q\left\{g^{-1}(\widehat{m}_i),Y_i\right\}\right]. \tag{5.8}$$

By Taylor’s expansion, we have

$$\ell_n(\theta)=\sum_{i=1}^{n}q_1(\widehat{m}_i,Y_i)\alpha_n\mathbf{v}^T\mathbf{Z}_i+\tfrac{1}{2}\alpha_n^2\mathbf{v}^T\mathbf{B}_n\mathbf{v}, \tag{5.9}$$

where

$$\mathbf{B}_n=\frac{1}{n}\sum_{i=1}^{n}\rho_2\left\{g^{-1}(\widehat{m}_i+\zeta_{ni})\right\}\mathbf{Z}_i\mathbf{Z}_i^T,$$

with ζ_{ni} between 0 and $\alpha_n\mathbf{v}^T\mathbf{Z}_i$, independent of Y_i. It can be shown that

$$\mathbf{B}_n=-E\rho_2\left\{\alpha_0^T(U)\mathbf{X}+\mathbf{Z}^T\beta_0\right\}\mathbf{Z}\mathbf{Z}^T+o_P(1)\equiv-\mathbf{B}+o_P(1). \tag{5.10}$$

Furthermore, we have

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n} q_1(\widehat{m}_i,Y_i)\mathbf{Z}_i = & \; n^{-1/2}\sum_{i=1}^{n} q_1(m_i,Y_i)\mathbf{Z}_i \\ & + n^{-1/2}\sum_{i=1}^{n} q_2(m_i,Y_i)\left[\{\widehat{\alpha}(U_i)-\alpha_0(U_i)\}^T\mathbf{X}_i\right]\mathbf{Z}_i + O_P(n^{1/2}\parallel\widehat{\alpha}-\alpha_0\|_\infty^2). \end{aligned}$$

By Theorem 1, the second term in the above expression can be expressed as

$$\begin{aligned} n^{-3/2} & \sum_{i=1}^{n} q_2(m_i,Y_i)f(U_i)^{-1}\sum_{j=1}^{n}(\tilde{W}_j^T\mathbf{X})K_h(U_j-U_i)\mathbf{Z}_i + O_P\left\{n^{1/2}c_n^2\log^{1/2}(1/h)\right\} \\ & \equiv T_{n1} + O_P\left\{n^{1/2}c_n^2\log^{1/2}(1/h)\right\}. \end{aligned}$$

Now define the vector $v_j = v(\mathbf{X}_j, Y_j, \mathbf{Z}_j)$ consisting of the first p elements of $q_1(m_j,Y_j)\sum^{-1}(u)(\mathbf{X}_i^T,\mathbf{Z}_j^T)^T$. Using the definition of $\bar{a}_j(U_i)$, we obtain $\bar{a}_j(U_i) - m_j = O((U_j - U_i)^2)$ and therefore

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

It can be shown via calculating the second moment that

$$T_{n2} - T_{n3} \overset{P}{\rightarrow} 0, \tag{5.11}$$

Where $T_{n3} = -n^{-1/2}\sum_{j=1}^{n}\gamma(U_j)$ With $\gamma(u_j) = \sum_{k=1}^{p} v_{jk}E\left[\rho_2\left\{\alpha_0^T(u)\mathbf{X}+\mathbf{Z}^T\beta_0\right\}X_k\mathbf{Z}|U=u_j\right]$. Combining (5.8)–(5.11) we obtain that

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

where $\Omega(U_i, Y_i, \mathbf{Z}_i) = q_1(m_i, Y_i)\mathbf{Z}_i - \boldsymbol{\gamma}(U_i)$. It follows that $\ell_n(\theta) = O_P(n^{1/2}\alpha_n) + O_P(n\alpha_n^2) + o_P(1)$.

We now show (B). We first point out a fact that under the conditions of Theorem 3, with probability tending to 1, for any given $\boldsymbol{\beta}_1$ satisfying that $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ and any constant C,

$$\mathscr{L}_P\left\{\begin{pmatrix}\beta_1\\ \mathbf{0}\end{pmatrix}\right\} = \max_{\|\beta_2\|\le Cn^{-1/2}}\mathscr{L}_P\left\{\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}\right\}.$$

Its proof is similar to that of Lemma 1 of Fan and Li (2001). This statement implies that $\hat{\boldsymbol{\beta}}_2 = 0$.

Let $\hat{\boldsymbol{\theta}} = n^{1/2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})$, $\widehat{m}_{i1} = \widehat{\alpha}^T(U_i)\mathbf{X}_i + \mathbf{Z}_{i1}^T\beta_{10}$, and $m_{i1} = \alpha_0^T(U_i)\mathbf{X}_i + \mathbf{Z}_{i1}^T\beta_{10}$. Then, $\hat{\boldsymbol{\theta}}$ maximizes

$$\sum_{i=1}^{n}\left[Q\left\{g^{-1}(\widehat{m}_{i1}+n^{-1/2}\mathbf{Z}_{i1}^T\theta),Y_i\right\} - Q\left\{g^{-1}(\widehat{m}_{i1}),Y_i\right\}\right] - n\sum_{j=1}^{s}p_{\lambda_n}(\widehat{\beta}_{j1}). \tag{5.12}$$

We consider the first term, say $\ell_{n1}(\boldsymbol{\theta})$. By Taylor's expansion, we have

$$\ell_{n1}(\theta)=n^{-1/2}\sum_{i=1}^{n} q_1(\widehat{m}_{i1},Y_i)\mathbf{Z}_{i1}^T\theta+\tfrac{1}{2}\theta^T\mathbf{B}_{n1}\theta,$$

where

$$\mathbf{B}_{n1}=\frac{1}{n}\sum_{i=1}^{n}\rho_2\left\{g^{-1}(\widehat{m}_{i1}+\zeta_{ni})\right\}\mathbf{Z}_{i1}\mathbf{Z}_{i1}^T,$$

with ζ_{ni} between 0 and $n^{-1/2}\mathbf{Z}_{i1}^T\theta$, independent of Y_i. It can be shown that

$$\mathbf{B}_{n1}=-E_{\rho_2}\left\{\alpha_0^T(U)\mathbf{X}+\mathbf{Z}_1^T\beta_{10}\right\}\mathbf{Z}_1\mathbf{Z}_1^T+o_P(1)=-\mathbf{B}_1+o_P(1). \tag{5.13}$$

A similar proof for (A) yields that

$$\ell_{n1}(\theta)=n^{-1/2}\sum_{i=1}^{n}\widehat{\theta}\Omega_1(U_i,Y_i,\mathbf{Z}_{i1})-\tfrac{1}{2}\theta^T\mathbf{B}_1\theta+o_P(1),$$

where $\Omega_1(U_i, Y_i, \mathbf{Z}_{i1}) = q_1(m_{i1}, Y_i)\mathbf{Z}_{i1} - \Gamma_1(U_i)$. By the Convexity Lemma we have that

$$(\mathbf{B}_1=\textstyle\sum_\lambda)\widehat{\theta}+n^{1/2}\mathbf{b}=n^{-1/2}\displaystyle\sum_{i=1}^{n}\Omega_1(U_i,Y_i,\mathbf{Z}_{i1})+o_P(1),$$

The conclusion follows as claimed.

Proof of Theorem 4—Decompose $\mathcal{R}(H_1) - \mathcal{R}(H_0)$ as

$$\begin{aligned}&\sum_{i=1}^{n}\left[Q\{g^{-1}(\widehat{\alpha}^{\mathrm{T}}(U_i)\mathbf{X}_i+\mathbf{Z}_i^{\mathrm{T}}\widehat{\beta}_{\mathrm{GPLM}}),Y_i\}-Q\{g^{-1}(\widehat{\alpha}^{\mathrm{T}}(U_i)\mathbf{X}_i+\mathbf{Z}_i^{\mathrm{T}}\beta),Y_i\}\right]\\&+\sum_{i=1}^{n}\left[Q\{g^{-1}(\widehat{\theta}^{\mathrm{T}}\mathbf{X}_i+\mathbf{Z}_i^{\mathrm{T}}\widehat{\beta}_{\mathrm{GLM}}),Y_i\}-Q\{g^{-1}(\widehat{\theta}^{\mathrm{T}}\mathbf{X}_i+\mathbf{Z}_i^{\mathrm{T}}\beta),Y_i\}\right]\\&+\sum_{i=1}^{n}\left[Q\{g^{-1}(\widehat{\alpha}^{\mathrm{T}}(U_i)\mathbf{X}_i+\mathbf{Z}_i^{\mathrm{T}}\beta),Y_i\}-Q\{g^{-1}(\widehat{\theta}^{\mathrm{T}}\mathbf{X}_i+\mathbf{Z}_i^{\mathrm{T}}\beta),Y_i\}\right]\end{aligned}$$

An analogous proof as for Theorem 3 (B) yields the first term is $o_P(n)$. On the other hand, it is obvious that the second term is $o_P(n)$. Finally a direct application of Theorem 10 of Fan, Zhang, and Zhang (2001) derives that

$$r_K\sum_{i=1}^{n}\left[Q\left\{g^{-1}(\widehat{\alpha}^{\mathrm{T}}(U_i)\mathbf{X}_i+\mathbf{Z}_i^{\mathrm{T}}\beta),Y_i\right\}-Q(g^{-1}(\widehat{\theta}^{\mathrm{T}}\mathbf{X}_i+\mathbf{Z}_i^{\mathrm{T}}\beta),Y_i\right\}\right]\xrightarrow{D}\chi^2_{\delta_n}.$$

We complete the proof.

References

Carroll RJ, Fan J, Gijbels I, Wand MP. Generalized Partially Linear Single-Index Models. Journal of the American Statistical Association 1997;92:477–489.

Carroll RJ, Ruppert D, Welsh AH. Local Estimating Equations. Journal of the American Statistical Association 1998;93:214–227.

Fan J, Heckman NE, Wand MP. Local Polynomial Kernel Regression for Generalized Linear Models And Quasilikelihood Functions. Journal of the American Statistical Association 1995;90:141–150.

Fan J, Li R. Variable Selection Via Nonconcave Penalized Likelihood And Its Oracle Properties. Journal of the American Statistical Association 2001;96:1348–1360.

Fan J, Zhang C, Zhang J. Generalized Likelihood Ratio Statistics And Wilks Phenomenon. The Annals of Statistics 2001;29:153–193.

Härdle, W.; Liang, H.; Gao, JT. Partially Linear Models. Heidelberg: Springer Physica; 2000.

Hastie, TJ.; Tibshirani, R. Generalized Additive Models. London: Chapman and Hall; 1990.

Heckman N. Spline Smoothing in a Partly Linear Model. Journal of the Royal Statistical Society, Series B 1986;48:244–248.

Hunsberger S. Semiparametric Regression in Likelihood-Based Models. Journal of the American Statistical Association 1994;89:1354–1365.

Liang H, Härdle W, Carroll RJ. Estimation in a Semiparametric Partially Linear Errors-in- Variables Model. The Annals of Statistics 1999;27:1519–1535.

Mack YP, Silverman BW. Weak and Strong Uniform Consistency of Kernel Regression Estimates*Z. Wahrscheinlichkeitstheorie verw*. Gebiete 1982;61:405–415.

Mammen E, van de Geer S. Penalized Estimation in Partial Linear Models. The Annals of Statistics 1997;25:387–413.

Pollard D. Asymptotics for Least Absolute Deviation Regression Estimators. Econometric Theory 1991;7:186–199.

Ruppert D, Sheather SJ, Wand MP. An Effective Bandwidth Selector for Local Least Squares Regression. Journal of the American Statistical Association 1995;90:1257–1270.

Severini TA, Staniswalis JG. Quasilikelihood Estimation in Semiparametric Models. Journal of the American Statistical Association 1994;89:501–511.

Speckman P. Kernel Smoothing in Partial Linear Models. Journal of the Royal Statistical Society, Series B 1988;50:413–436.

Wahba, G. Partial Spline Models for Semiparametric Estimation of Functions of Several Variables. Statistical Analysis of Time Series; Proceedings of the Japan U.S. Joint Seminar; Tokyo. 1984. p. 319-329.