Statistica Sinica **23** (2013), 1091-1116
doi:http://dx.doi.org/10.5705/ss.2012.133

LOCAL LEAST ABSOLUTE RELATIVE ERROR ESTIMATING APPROACH FOR PARTIALLY LINEAR MULTIPLICATIVE MODEL

Qingzhao Zhang and Qihua Wang

Chinese Academy of Science

Abstract: The partially linear multiplicative regression model is considered. This model, which becomes a partially linear regression model after taking logarithmic transformation, is useful in analyzing data with positive responses. Chen et al. (2010) mentioned that in many applications the size of relative error, rather than that of error itself, is the central concern of practitioners. We extend the criterion of least absolute relative error (LARE) to the partially linear multiplicative regression model by local smoothing techniques. Consistency and asymptotic normality are investigated. We utilize a random weighting method to estimate asymptotic covariance of the parameter estimator. We also propose a simple and effective method to select important variables in the linear part. The oracle property (Fan and Li (2001)) is proved. Some numerical studies are conducted to evaluate and compare the performance of the proposed estimators. The body fat dataset is analyzed for illustration.

Key words and phrases: Lasso, least absolute relative error, partially linear model, variable selection.

1. Introduction

In linear, non-linear, and semiparametric regression analysis, commonly used approaches are least squares (LS) and quantile regression (QR) methods. The LS method is sensitive to outliers, and its efficiency can be significantly improved for non-normal errors; QR estimators are more robust. However, we note that QR method requires positivity of the density of the errors at quantiles and the asymptotic relative efficiency of a single quantile to LS can be arbitrarily small. For a complete discussion on quantile regression, see Koenker (2005).

The two criteria are based on absolute errors, while in many applications the concern is with relative errors. Some papers have suggested estimation methods for linear model and non-linear model based on relative errors. See Narula and Wellington (1977), Khoshgoftaar, Bhattacharyya, and Richardson (1992), Park and Stefanski (1998), for examples. Chen et al. (2010) note that for relative error methods, consistency and asymptotic normality of their estimators have not been established for linear or nonlinear models under general regularity conditions; furthermore, the relative error in all such studies was the ratio of the error with respect to the target and that could be inadequate. They suggested instead the ratio of the error with respect to the predictor, and proposed the least absolute relative errors criterion (LARE) that used both types of relative errors for the linear multiplicative model $y_i = \exp(x_i^T\beta_0)\varepsilon_i$. They found the least absolute relative error (LARE) estimator by minimizing

$$LARE_n(\beta) = \sum_{i=1}^{n} \Big\{ \Big| \frac{y_i - \exp(x_i^T\beta)}{y_i} \Big| + \Big| \frac{y_i - \exp(x_i^T\beta)}{\exp(x_i^T\beta)} \Big| \Big\}, \tag{1.1}$$

and consistency and asymptotic normality were proved. The asymptotic properties of this estimator do not require the positivity of the density of the error over its support.

As pointed out by Chen et al. (2010), an advantage of the LARE criterion is that it is scale free; this is important for applying the LARE criterion to certain types of data, and they give some examples of this.

In many cases, the linear multiplicative model is not complex enough to capture the underlying relationship between response variables and their associated covariates. This motivates us to consider the following partially linear multiplicative model

$$Y = H(X^T\beta_0 + g(T))\varepsilon, \tag{1.2}$$

where $H(\cdot) > 0$ is a given function, Y is a scalar response variable, X is a p-dimensionial random covariate vector, T is a random variable with a bounded support Ω, $g(\cdot)$ is an unknown univariate link function on Ω, and the random error ε has $P(\varepsilon > 0) = 1$ with probability density function f. For the sake of identification, the intercept term is not included in β_0. This model reduces to the linear multiplicative model when $H(\cdot)$ is the exponential function and $g(\cdot) = 0$. It is useful and more flexible in analyzing data with positive responses, such as stock prices or life times, which are particulary common in economic and biomedical studies.

Our first aim is to extend Chen et al. (2010)'s work to Model (1.2), and to propose the *semiparametric* least absolute relative errors criterion (Semi-LARE) for estimating both parametric and nonparametric parts. This extension involves nonparametric estimation and kernel smoothing techniques. We use a random weighting method to approximate the asymptotic covariance matrix of the estimators in linear part.

A second goal is variable selection for the parametric part of (1.2). There are often many covariates in the parametric part of Model (1.2). With sparsity, variable selection can improve the estimation accuracy by effectively identifying the subset of important predictors and can enhance model interpretability with parsimonious representation (see Fan and Li (2006)). A number of variable selection methods are popular, such as the LASSO (Tibshirani (1996)), SCAD(Fan and Li (2001)), and the Adaptive LASSO (Zou (2006)). Here we give an easy and effective variable selection to select significant parametric components in Model (1.2). Our approach makes use of the LARS algorithm in Efron et al. (2004) and can be quickly implemented to obtain the solution path. Moreover, The oracle property (Fan and Li (2001)) is proved.

The rest of the paper is organized as follows. The method and its theoretical properties are investigated in Section 2. In Section 3, we address issues related to asymptotic covariance estimation. We propose variable selection for the parametric part and give its oracle properties in Section 4. Simulation studies are presented in Section 5 to show the finite sample performance of the proposed methods. A data set is analyzed in Section 6. All technical details are in the Appendix.

2. Semi-LARE Method and Its Asymptotic Properties

Let $\{x_i, t_i, y_i\},\ i = 1, \ldots, n$ be an i.i.d sample from Model (1.2). Let $K(\cdot)$ be a given kernel function and $K_h(\cdot) = K(\frac{\cdot}{h})/h$ with bandwidth h. We define our estimator in several steps.

Step 1. For t_i in the neighborhood of t, use a local linear approximation

$$g(t_i) \approx g(t) + g'(t)(t_i - t) \equiv a(t) + b(t)(t_i - t),$$

and let $\{\tilde{\beta}, \tilde{a}, \tilde{b}\}$ be the minimizer of the *local* absolute relative loss function

$$\sum_{i=1}^{n} \Big\{ \Big| \frac{y_i - H\{x_i^T\beta + a + b(t_i - t)\}}{y_i} \Big| + \Big| \frac{y_i - H\{x_i^T\beta + a + b(t_i - t)\}}{H\{x_i^T\beta + a + b(t_i - t)\}} \Big| \Big\} K_h(t_i - t). \tag{2.1}$$

Then $\tilde{g}(t) = \tilde{a}, \quad \tilde{g'}(t) = \tilde{b}$.

Step 2. Compute an improved estimator of β_0 as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \Big\{ \Big| \frac{y_i - H\{x_i^T\beta + \tilde{g}(t_i)\}}{y_i} \Big| + \Big| \frac{y_i - H\{x_i^T\beta + \tilde{g}(t_i)\}}{H\{x_i^T\beta + \tilde{g}(t_i)\}} \Big| \Big\}. \tag{2.2}$$

Step 3. Let $\{\hat{a}, \hat{b}\}$ be the minimizer of

$$\sum_{i=1}^{n} \Big\{ \Big| \frac{y_i - H\{x_i^T\hat{\beta} + a + b(t_i - t)\}}{y_i} \Big| + \Big| \frac{y_i - H\{x_i^T\hat{\beta} + a + b(t_i - t)\}}{H\{x_i^T\hat{\beta} + a + b(t_i - t)\}} \Big| \Big\} K_h(t_i - t). \tag{2.3}$$

Then $\hat{g}(t)=\hat{a}$, $\hat{g}'(t)=\hat{b}$.

We establish theoretical justifications for the these estimators. Take $J=E[\varepsilon_i H^{-1}(0)sgn(\varepsilon_i H^{-1}(0)-1)]$, $V_t=E[(x^T,1)^T(x^T,1)|T=t], \gamma=H'(0)/H(0)$, $A=E(\varepsilon_i H^{-1}(0)+\varepsilon_i^{-1}H(0))^2$, and

$$\mu_j=\int u^jK(u)du \;\text{ and }\; \nu_j=\int u^jK^2(u)du, \quad j=0,1,2.$$

Let 0_p be the $p\times 1$ zero vector. The aysmptotic properties of $\{\tilde{\beta},\tilde{g}(t)\}$ are as follows.

Theorem 1. *Under the regularity conditions in the Appendix, if $h\to 0$ and $nh\to\infty$ as $n\to\infty$, then*

$$\sqrt{nh}\Big\{\begin{pmatrix}\tilde{\beta}-\beta_0\\ \tilde{g}(t)-g(t)\end{pmatrix}-\frac{\mu_2h^2f(H(0))H(0)g''(t)}{2(J+2f(H(0))H(0))}V_t^{-1}\begin{pmatrix}E[x|T=t]\\ 1\end{pmatrix}\Big\}$$
$$\xrightarrow{d} N\Big(0,\frac{\nu_0}{4\gamma^2f_T(t)}[J+2f(H(0))H(0)]^{-2}AV_t^{-1}\Big),$$

where $f_T(t)$ is the density function of T and $f(\cdot)$ is the density function of ε.

The next result gives the asymptotic normality of $\hat{\beta}$.

Theorem 2. *Let*

$$\eta(t,x)=\frac{J}{J+2f(H(0))H(0)}E[x(0_p^T,1)|T=t]V_t^{-1}(x^T,1)^T.$$

Under the regularity conditions in the Appendix, if $nh^4\to 0$ and $nh^2/\log(1/h)\to\infty$ as $n\to\infty$, then

$$\sqrt{n}(\hat{\beta}-\beta_0)\to_d N\Big(0,\frac{1}{4}\{\gamma[J+2f(H(0))H(0)]\}^{-2}AC^{-1}\Xi C^{-1}\Big),$$

where $C=E(xx^T), \Xi=E[\{x-\eta(t,x)\}\{x-\eta(t,x)\}^T]$.

The optimal bandwidth in Theorem 1 is $h\sim n^{-1/5}$, but this bandwidth does not satisfy the condition in Theorem 2. Hence, in order to obtain the root-n consistency and asymptotic normality of $\hat{\beta}$, undersmoothing of $\tilde{g}(t)$ is necessary. This is a common requirement in semiparametric models; see Carroll et al. (1997) for a detailed discussion. We use a random weighted approach to estimate the asymptotic variance, see the next section.

Theorem 3. *Under the regularity conditions given in the Appendix, if $h\to 0$ and $nh\to\infty$ as $n\to\infty$, then*

$$\sqrt{nh}\Big\{\hat{g}(t)-g(t)-\frac{\mu_2h^2f(H(0))H(0)g''(t)}{2(J+2f(H(0))H(0))}(0_p^T,1)V_t^{-1}E[(x^T,1)^T|T=t]\Big\}$$
$$\to_d N\Big(0,\frac{\nu_0}{4\gamma^2f_T(t)}[J+2f(H(0))H(0)]^{-2}A\Big).$$

The asymptotic variance of $\hat{g}(t)$ here is smaller than that of $\tilde{g}(t)$. The proof of Theorem 3 is similar to that of Theorem 1, and is omitted.

3. Asymptotic Covariance Estimate

We look to the estimation of the asymptotic variance of $\hat{\beta}$. It involves the density function of the error terms and cannot be properly estimated using plug-in rules (Chen et al. (2010)). To avoid density estimation, we apply a random weighting method in our two-step semiparametric inference.

For $m = 1, \ldots, M$,

Step 1. Generate i.i.d nonnegative random variables $w_1^m, \ldots, w_n^m$, that have mean and variance equal to 1;

Step 2. Let $\{\tilde{\beta}^m, \tilde{a}^m, \tilde{b}^m\}$ be the minimizer of

$$\sum_{i=1}^{n} w_i^m \Big\{ \Big| \frac{y_i - H\{x_i^T\beta + a + b(t_i - t)\}}{y_i} \Big| + \Big| \frac{y_i - H\{x_i^T\beta + a + b(t_i - t)\}}{H\{x_i^T\beta + a + b(t_i - t)\}} \Big| \Big\} K_h(t_i - t). \tag{3.1}$$

Then $\tilde{g}^m(t) = \tilde{a}^m, \quad \tilde{g}'^m(t) = \tilde{b}^m$.

Step 3. Compute $\hat{\beta}^m$ by minimizing

$$\sum_{i=1}^{n} w_i^m \Big\{ \Big| \frac{y_i - H\{x_i^T\beta + \tilde{g}^m(t_i)\}}{y_i} \Big| + \Big| \frac{y_i - H\{x_i^T\beta + \tilde{g}^m(t_i)\}}{H\{x_i^T\beta + \tilde{g}^m(t_i)\}} \Big| \Big\}. \tag{3.2}$$

Similar to Jin, Ying, and Wei (2001), we find that the distribution of $\sqrt{n}(\hat{\beta} - \beta_0)$ can be approximated by the resampling distribution of $\sqrt{n}(\hat{\beta}^m - \hat{\beta})$, the covariance matrix of $\hat{\beta}$ consistently estimated by

$$\hat{\Sigma} = \frac{1}{M} \sum_{m=1}^{M} (\hat{\beta}^m - \hat{\beta})(\hat{\beta}^m - \hat{\beta})^\top.$$

Under the regularity conditions of Theorem 2, it can be verified that

$$n\hat{\Sigma} \to_p \frac{1}{4} \big\{ \gamma[J + 2f(H(0))H(0)] \big\}^{-2} AC^{-1} \Xi C^{-1},$$

where J, A, C, and Ξ are defined as in Section 2. Our simulation also illustrates that the method can produce a good covariance estimator for $\sqrt{n}(\hat{\beta} - \beta_0)$.

4. Variable Selection

4.1. Variable selection method

We propose a simple variable selection approach. The main idea is to separate the process into steps: (1) construct U, linearly dependent on predictors X;

(2) apply the variable selection methods for linear models to the linear regression of U on X. Oracle properties (Fan and Li (2001)) are proved.

Set

$$u_i = x_i^T\hat{\beta}, \quad i = 1, \ldots, n, \tag{4.1}$$

where $\hat{\beta}$ is obtained by minimizing (2.2). Then define

$$H_n(\beta) = \sum_{i=1}^{n}(u_i - x_i^T\beta)^2 + \lambda_n \sum_{j=1}^{p}\omega_j|\beta_j|, \tag{4.2}$$

where λ_n is a tuning parameter and the weight is $\omega_j = |\hat{\beta}|^{-\tau}$ with $\tau > 0$. We define a new estimator as $\hat{\beta}^{\lambda_n} = \mathrm{argmin}_\beta H_n(\beta)$.

Let $\beta_0 = (\beta_{01}, \beta_{02}, \ldots, \beta_{0p})^T$ and take $\aleph = \{j : \beta_{0j} \neq 0\}$. Similarly, let $\hat{\aleph} = \{j : \hat{\beta}_j^{\lambda_n} \neq 0\}$.

Theorem 4. *If the regularity conditions of Theorem* 2 *hold, and if* $\lambda_n/\sqrt{n} \to 0$ *and* $\lambda_n n^{(\tau-1)/2} \to \infty$ *as* $n \to \infty$, *then*

(i) $P(\hat{\aleph} = \aleph) \to 1$;

(ii) $\sqrt{n}(\hat{\beta}^{\lambda_n} - \beta_0) \to_d N(0, (1/4)\{\gamma[J + 2f(H(0))H(0)]\}^{-2} A\Pi_{\aleph\aleph})$, *where* $\Pi_{\aleph\aleph} = C_{\aleph\aleph}^{-1}\Xi_{\aleph\aleph}C_{\aleph\aleph}^{-1}$ *whose entries correspond to the variables in* $\aleph$.

It is natural to conduct variable selection by minimizing $T_n(\beta)$ given by

$$\sum_{i=1}^{n}\Big\{\Big|\frac{y_i - H\{x_i^T\beta + \tilde{g}(t_i)\}}{y_i}\Big| + \Big|\frac{y_i - H\{x_i^T\beta + \tilde{g}(t_i)\}}{H\{x_i^T\beta + \tilde{g}(t_i)\}}\Big|\Big\} + \lambda_n^* \sum_{j=1}^{p}\omega_j|\beta_j|, \tag{4.3}$$

where λ_n^* is a tuning parameter. We define a new estimator as $\hat{\beta}^{\lambda_n^*} = \underset{\beta}{\mathrm{argmin}}\, T_n(\beta)$ and $\hat{\aleph}^* = \{j : \hat{\beta}_j^{\lambda_n^*} \neq 0\}$.

Theorem 5. *If the regularity conditions of Theorem* 2 *hold, and if* $\lambda_n^*/\sqrt{n} \to 0$ *and* $\lambda_n^* n^{(\tau-1)/2} \to \infty$ *as* $n \to \infty$, *then*

(i) $P(\hat{\aleph}^* = \aleph) \to 1$;

(ii) $\sqrt{n}(\hat{\beta}^{\lambda_n^*} - \beta_0) \to_d N(0, \frac{1}{4}\{\gamma[J + 2f(H(0))H(0)]\}^{-2} A\Pi_{\aleph\aleph})$.

The properties of the estimators that minimize (4.2) and (4.3) are identical. However, it is time-consuming to obtain the solution path of the minimizer of (4.3), while the solution path of our proposed estimator can be computed by the LARS algorithm.

4.2. Tuning selection

The features of Theorem 4 depend on the appropriate choice of the tuning parameter λ_n. Various techniques have been proposed. For example, Golub, Heath, and Wahba (1979) used GCV method to estimate the ridge parameter; Wang, Li, and Tsai (2007) proposed a BIC tuning parameter selector that was able to identify the true model consistently. We propose two approaches to selecting the tuning parameter.

We are motivated by Wang and Leng (2007) to propose the criterion

$$WIC(\lambda_n) = (\hat{\beta}^{\lambda_n} - \hat{\beta})^{\top}(n\hat{\Sigma}_n)^{-1}(\hat{\beta}^{\lambda_n} - \hat{\beta}) + df^{\lambda_n}\frac{\log(n)}{n},$$

where $n\hat{\Sigma}_n$ is the estimator for the asymptotic covariance of $\sqrt{n}(\hat{\beta} - \beta_0)$ given in Section 3, and df^{λ_n} corresponds to the number of non-zero coefficients in the parametric part of the fitting model. Take $\hat{\lambda}_{WIC} = \operatorname{argmin} WIC(\lambda_n)$.

In addition, as for estimators are based on semi-LARE, we use the BIC-type criterion $BIC(\lambda_n)$, which is defined as

$$\log\Big(\frac{1}{n}\sum_{i=1}^{n}\Big\{\Big|\frac{y_i - H\{x_i^T\hat{\beta}^{\lambda_n} + \tilde{g}(t_i)\}}{y_i}\Big| + \Big|\frac{y_i - H\{x_i^T\hat{\beta}^{\lambda_n} + \tilde{g}(t_i)\}}{H\{x_i^T\hat{\beta}^{\lambda_n} + \tilde{g}(t_i)\}}\Big|\Big\}\Big) + df^{\lambda_n}\frac{\log(n)}{n},$$

where $\tilde{g}(t_i)$ are obtained from the full model at Step 1, Section 2. Take $\hat{\lambda}_{BIC} = \operatorname{argmin} BIC(\lambda_n)$.

The estimator $\hat{\beta}^{\lambda_n}$ defines a candidate model $\aleph^{\lambda_n} = \{j : \hat{\beta}_j^{\lambda_n} \neq 0\}$ and the selection consistency of the proposed criteria are as follows.

Theorem 6. *Assume the regularity conditions of Theorem* 2. *The tuning parameters are selected by the WIC and BIC criteria satisfy* $P(\aleph^{\hat{\lambda}_{WIC}} = \aleph) \to 1$ *and* $P(\aleph^{\hat{\lambda}_{BIC}} = \aleph) \to 1$ *as* $n \to \infty$.

The proof is similar to that of Theorem 4 in Wang and Leng (2007). The finite sample performance of WIC and BIC are illustrated in the next section.

5. Numerical Study

We conducted a simulation study to investigate the finite-sample performances of our proposed method. In all examples, we fixed the kernel function to be the Epanechnikov kernel $K(u) = (3/4)(1 - u^2)_+$. We set the bandwidth h_1 to be $n^{-1/3}$ for Step 1 and $h_2 = n^{-1/5}$ for Step 3. The selection of h_1 might not be so critical in terms of the $\sqrt{n}$-rate asymptotic normality of $\hat{\beta}$, and a proper choice of h_1 depends on only the second order term of the mean square error of $\hat{\beta}$. For a detailed discussions, see Remark 3.3 in Wang and Rao (2002). Here

$h_1 = n^{-1/3}$ satisfies the conditions in Theorem 2, while $h_2 = n^{-1/5}$, selected by the rule of thumb, is required for achieving the optimal rate of convergence from the term of mean squares error. In our study we repeated the simulation 500 times.

Example 1. We considered two situations with sample sizes of $n = 60, n = 120$, and $n = 200$. We considered

$$Y = \exp\{X^T\beta + 8T(1-T)\}\varepsilon, \tag{5.1}$$

where the predictor X is independently generated from a p-dimensional normal distribution with mean 0 and covariances $cov(X_{ij}, X_{ik}) = 0.5^{|j-k|}$. The covariate T is uniformly distributed on $[0, 1]$. Here we fixed $p = 3$ and set $\boldsymbol{\beta}_0 = (2, -1, 0.5)^\top$ to compare the estimation efficiency of semi-LARE with that of least squares (LS) and least absolute deviation (LAD). We got the LS and LAD estimators by minimizing L_2 and L_1-norm absolute errors based on the data set $\{log(Y), X, T\}$, respectively. We considered two error distributions: $\log(\varepsilon) \sim N(0, 1)$ and $\log(\varepsilon) \sim U(-2, 2)$.

To assess performance, we calculated mean of the biases (Bias), absolute bias (AB), and standard errors(SE) of the three estimators. Moreover, we set M=500 and used the proposed method in Section 3 to obtain standard error estimators(SEE) of semi-LARE estimator. The simulation results are reported in Tables 1.

Both the sample bias and SE decreased as n increased, which is expected. When $\log(\varepsilon) \sim N(0, 1)$, semi-LARE did well compared to the LS which is efficient, while LAD showed poorly, in terms of AB and SE. For $\log(\varepsilon) \sim U(-2, 2)$, semi-LARE performed much better than LS and LAD. These results are coincide with that of Chen et al. (2010). Moreover, SEE and SE were close when n increased.

Similar to Xue and Wang (2012), we computed root mean squared error (RMSE) for the functional component estimator $\hat{g}(\cdot)$,

$$RMSE = \left[n_{grid}^{-1} \sum_{k=1}^{n_{grid}} \{\hat{g}(t_k) - g(t_k)\}^2\right]^{1/2},$$

where $\{t_k, k = 1, \ldots, n_{grid}\}$ were regular grid points. Here $n_{grid} = 100$. The boxplots for 500 RMSEs under log-norm and log-uniform error distributions are shown in Figure 1 and 2. From Figures 1 and 2, one gets conclusions similar to those of the estimates of β in terms of median RMSEs.

Table 1. Summary of Bias and Standard Deviation of Example 1.

n	Method		$\log(\varepsilon) \sim U(-2,2)$			$\log(\varepsilon) \sim N(0,1)$		
			β_1	β_2	β_3	β_1	β_2	β_3
60	LARE	Bias	-0.0138	0.0094	-0.0051	0.0066	-0.0028	0.0057
		AB	0.1470	0.1631	0.1430	0.1371	0.1523	0.1408
		SE	0.1839	0.2080	0.1783	0.1720	0.1892	0.1744
		SEE	0.1693	0.1893	0.1713	0.1572	0.1733	0.1552
	LAD	Bias	-0.0111	0.0116	-0.0076	0.0104	-0.0079	-0.0013
		AB	0.2228	0.2588	0.2271	0.1642	0.1773	0.1690
		SE	0.2743	0.3194	0.2795	0.2071	0.2229	0.2102
	LS	Bias	-0.0104	0.0087	-0.0070	0.0084	-0.0033	0.0049
		AB	0.1516	0.1664	0.1476	0.1305	0.1441	0.1338
		SE	0.1876	0.2105	0.1842	0.1645	0.1813	0.1678
120	LARE	Bias	-0.0048	0.0002	0.0048	0.0012	-0.0015	-0.0127
		AB	0.0944	0.1085	0.0982	0.0905	0.1054	0.0950
		SE	0.1213	0.1371	0.1222	0.1132	0.1323	0.1147
		SEE	0.1163	0.1303	0.1160	0.1078	0.1200	0.1064
	LAD	Bias	-0.0105	0.0015	0.0088	0.0056	-0.0062	-0.0047
		AB	0.1638	0.1765	0.1602	0.1077	0.1150	0.1088
		SE	0.2045	0.2189	0.2024	0.1347	0.1440	0.1362
	LS	Bias	-0.0060	-0.0017	0.0074	0.0007	-0.0009	-0.0102
		AB	0.0997	0.1145	0.1026	0.0863	0.0997	0.0895
		SE	0.1271	0.1434	0.1278	0.1080	0.1254	0.1095
200	LARE	Bias	0.0019	0.0007	-0.0034	-0.0015	-0.0064	0.0031
		AB	0.0784	0.0828	0.0731	0.0686	0.0802	0.0691
		SE	0.0966	0.1030	0.0926	0.0862	0.1003	0.0879
		SEE	0.0887	0.0987	0.0886	0.0833	0.0926	0.0829
	LAD	Bias	0.0065	-0.0030	-0.0087	-0.0015	-0.0055	0.0002
		AB	0.1375	0.1414	0.1250	0.0816	0.0928	0.0819
		SE	0.1664	0.1784	0.1597	0.1024	0.1169	0.1033
	LS	Bias	0.0019	0.0009	-0.0048	-0.0002	-0.0088	0.0038
		AB	0.0828	0.0876	0.0784	0.0660	0.0784	0.0658
		SE	0.1019	0.1087	0.0989	0.0828	0.0972	0.0828

Example 2. Here we evaluated the performance of the variable selection proposed in Section 4. The setup is same as that of Example 1, except $p = 8$ and $\beta_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. Moreover, we set $\tau = 2$. Since the proposed method is based on initial estimators, we used semi-LARE to generate the estimator. We used WIC and BIC to select the tuning parameter.

In Table 2, we report the proportion of under-fitted(PU), the proportion of over-fitted(PO), the proportion of correct-fitted(PC), the average number of

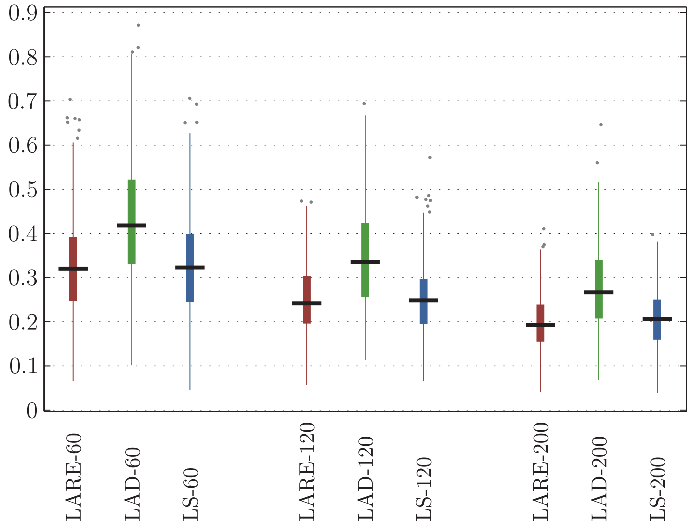

Figure 1. The boxplots of 500 RMSEs based on LARE, LAD, and LS criteria under log-uniform error distribution. Note that LARE-60 means LARE estimate under sample size 60.

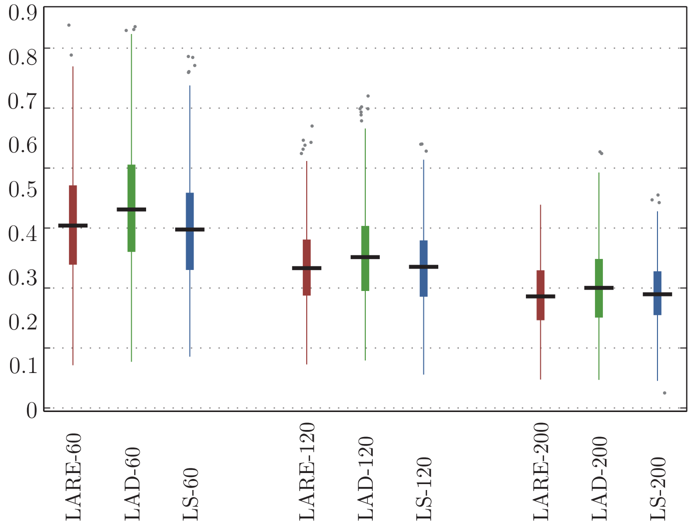

Figure 2. The boxplots of 500 RMSEs based on LARE, LAD, and LS criteria under log-norm error distribution.

correct zeros (C), and the average size of model selected(Msize). As to performance of the variable selection estimation, following Wang and Xia (2009), we

Table 2. Summary of Example 2.

		Proportion of models			model size		$MRGMSE_\beta$		$MREE_g$	
Method	n	PU	PO	PC	C	Msize	unpenalized	oracle	unpenalized	oracle
$\log(\varepsilon) \sim U(-2,2)$										
WIC	60	0.0020	0.2460	0.7520	4.6380	3.3600	0.4479	1.2735	0.9619	1.0144
	120	0.0000	0.1740	0.8260	4.7880	3.2120	0.4175	1.1024	0.9822	1.0008
	200	0.0000	0.1280	0.8720	4.8500	3.1500	0.4125	1.0708	0.9891	0.9999
BIC	60	0.0060	0.1720	0.8220	4.7520	3.2320	0.4088	1.1815	0.9558	1.0096
	120	0.0000	0.0820	0.9180	4.9140	3.0860	0.3804	1.0778	0.9793	0.9999
	200	0.0000	0.0540	0.9460	4.9420	3.0580	0.3571	1.0475	0.9882	0.9997
$\log(\varepsilon) \sim N(0,1)$										
WIC	60	0.0020	0.2700	0.7280	4.6280	3.3700	0.4390	1.2681	0.9702	1.0029
	120	0.0000	0.1720	0.8280	4.7880	3.2120	0.4219	1.1273	0.9807	1.0016
	200	0.0000	0.1220	0.8780	4.8500	3.1500	0.4020	1.0276	0.9863	1.0015
BIC	60	0.0040	0.2200	0.7760	4.7140	3.2820	0.4183	1.2226	0.9702	1.0037
	120	0.0000	0.1200	0.8800	4.8540	3.1460	0.3645	1.0977	0.9816	1.0015
	200	0.0000	0.0860	0.9140	4.9100	3.0900	0.3763	1.0222	0.9874	1.0016

computed $RGMSE_\beta$ and the relative estimation error (REE_g),

$$RGMSE_\beta = \frac{(\hat{\beta} - \beta_0)^T E(XX^T)(\hat{\beta} - \beta_0)}{(\bar{\beta} - \beta_0)^T E(XX^T)(\bar{\beta} - \beta_0)}, \quad REE_g = \frac{\sum_{k=1}^{n_{grid}} |\hat{g}(t_k) - g(t_k)|}{\sum_{k=1}^{n_{grid}} |\bar{g}(t_k) - g(t_k)|},$$

where $\bar{\beta}, \bar{g}(\cdot)$ are either the unpenalized estimators or the oracle estimators. The median of $RGMSE_\beta, REE_g$ values (labeled as $MRGMSE_\beta, MREE_g$) are listed.

Several observations can be made from the tables. Performance gets better with sample size n as expected: the proportion of the correctly fitted models increases for every model. This also confirms that the BIC and WIC criteria proposed in Section 4.2 can indeed identify the true model consistently. Note that variable selection based on BIC has slightly higher probability of correct-select than that based on WIC. This is clearest in small samples, and may be due to the poor performance of the estimate of asymptotic variance when the sample is small. Implementation of WIC is lengthy due to the estimation of the asymptotic covariance matrix, so we recommend BIC for selecting the tuning parameter. The median $RGMSE_\beta$ of our estimators to that of the unpenalized estimators based on full model are much less than 1, and close to those of the oracle estimators based on the true model.

6. Data

We now illustrate the proposed method with an application to the body fat data that is available at `http://lib.stat.cmu.edu/datasets/bodyfat`. Accu-

Table 3. The estimates from the data.

	semi-LRAE			**semi-LAD**			**semi-LS**		
	est	sd	vse	est	sd	vse	est	sd	vse
x_1	0.1476	0.0633	0	0.1115	0.0570	0	0.1562	0.0718	0
x_2	-0.3945	0.3241	-0.2885	-0.3177	0.2753	-0.2977	-0.3773	0.3003	-0.2481
x_3	0.1050	0.0979	0	0.0411	0.0806	0	0.0995	0.1062	0
x_4	-0.0660	0.0856	0	0.0117	0.0834	0	-0.0803	0.0903	0
x_6	0.8309	0.2464	0.7525	0.6325	0.1651	0.6236	0.8277	0.2542	0.7280
x_7	-0.1936	0.1791	0	-0.1555	0.1588	0	-0.2480	0.1965	0
x_8	0.1665	0.1150	0	0.1352	0.1065	0	0.2049	0.1375	0
x_9	-0.0259	0.0952	0	-0.0396	0.0954	0	0.0160	0.0961	0
x_{10}	0.0407	0.0383	0	0.0485	0.0448	0	0.0360	0.0463	0
x_{11}	0.1103	0.1062	0	0.0558	0.0855	0	0.1067	0.1035	0
x_{12}	0.0723	0.0690	0	0.0411	0.0654	0	0.0939	0.0879	0
x_{13}	-0.0860	0.0686	0	-0.0816	0.0639	0	-0.0994	0.0673	0
MAPE		2.9930			3.1461			3.0210	
MARE		0.3709			0.3657			0.3861	

rate measurement of body fat is inconvenient/costly and it is desirable to have easy methods of estimating it. The data on 252 men contains twelve baseline factors X: age (x_1), weight(x_2), height(x_3), and circumference of the skinfold measurements neck (x_4), chest (x_5), 2 abdomen (x_6), hip (x_7), thigh (x_8), knee (x_9), ankle (x_{10}), biceps (x_{11}), forearm (x_{12}), and wrist (x_{13}). The response Y is the percentage of body fat. The aim is to build a predictive model to relate the response to the covariates. We deleted possible outliers to a sample of size 250.

A descriptive analysis reveals that x_5 has a nonlinear relationship with $\log(Y)$, while other predictors are roughly linear with it. Let Z be the set of predictors $x_1, \ldots, x_4, x_6, \ldots, x_{13}$. Before applying our method, the Z predictors were transformed so that their marginal distribution was approximately $N(0,1)$. Also, the nonparametric part x_5 was transformed so that its marginal distribution was approximately $U[0,1]$. Therefore, we considered the model $Y = \exp(Z^T\beta + g(x_5))\varepsilon$. The results are presented in Table 3.

We used the first 200 samples to fit the model and to select significant variables, and then used the remaining observations to evaluate the predictive ability of the selected model. The estimate (est) and standard error (se) of β are listed. The only difference is the loss criterions, which we mark as semi-LARE, semi-LAD and semi-LS, respectively. We applied the variable selection in Section 4 and calculated the estimate of β (vse). Since WIC and BIC produced the same estimators, we show the the variable selection with the BIC selector. The results show that all estimates were similar. In particular, the predictors x_2 and x_6 were selected. The prediction performance is measured by the median absolute prediction error (MAPE) and the median absolute relative error (MARE). The predictor based on semi-LARE was meaningful and gave better predictions.

Acknowledgement

Wang's research was supported by the National Science Fund for Distinguished Young Scholars in China (10725106), the National Natural Science Foundation of China (10671198,11171331), and the National Science Fund for Creative Research Groups in China, a grant from the Key Lab of Random Complex Structure and Data Science, CAS, and the Key grant from Yunnan Province (2010CC003).

Appendix

The following technical conditions are to be assumed.

(C1) ε has a continuous density $f(\cdot)$ in a neighborhood of $H(0)$.

(C2) $P(\varepsilon > 0) = 1$.

(C3) (x_i^T, t_i) and ε_i are independent, and $\sup_t E[\|x_i\|^4|t] < \infty$.

(C4) $E(\varepsilon_i H^{-1}(0) + \varepsilon_i^{-1} H(0))^2 < \infty$, and

$$E[\{\varepsilon_i H^{-1}(0) + \varepsilon_i^{-1} H(0)\}\text{sgn}(1 - \varepsilon_i H^{-1}(0))] = 0.$$

(C5) The kernel $K(\cdot)$ is a symmetric density function with bounded support and satisfies a Lipschitz condition.

(C6) The function $H(\cdot)$ is positive and has a continuous third derivative. Moreover, for any $a > 0$,

$$a\frac{H''(s)}{H^2(s)} + \frac{H''(s)}{a} - 2a\frac{H'^2(s)}{H^2(s)} > 0\ .$$

(C7) The function $g(\cdot)$ has a bounded second derivative.

(C8) $H'(\cdot)$ and $H''(\cdot)$ satisfy a Lipschitz condition on interval $[-M, M]$ for some $M > 0$.

Remark A.1. Conditions (C1), (C2), and (C3) are regular conditions. Condition (C4) is needed for the weak consistency and identification. Condition (C5) has been used in the investigation on some nonparametric kernel estimators, e.g. in Härdle, Liang, and Gao (2000) and Mack and Silverman (1982). Condition (C6) is for the convexity of the objective function, see Lemma A.2. Obviously, $H(\cdot) = \exp(\cdot)$ satisfies the condition. Condition (C7) and $\sup_t E[x_i x_i^T|t] < \infty$ of (C3) are often seen in the literature on partially linear models.

Lemma A.1. *Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d random vectors, the y_i's univariate random variables. Assume that $E|y|^r < \infty$ and $\sup_x \int |y|^r p(x,y)dy < \infty$, $p(x,y)$ the joint density of (x,y).If K is a bounded positive function with bounded support satisfying a Lipschitz condition, then*

$$\sup_x \Big| n^{-1} \sum_{i=1}^{n} \Big\{ K_h(x_i - x)y_i - E[K_h(x_i - x)y_i] \Big\} \Big| = O_p\Big(\frac{\log^{1/2}(1/h)}{\sqrt{nh}} \Big),$$

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

Lemma A.1 is a result of Mack and Silverman (1982). We next give a modified version of Lemma 1 in Chen et al. (2010).

Lemma A.2. *Under (C6), if $\psi(x,a) = |1 - H(x)/a| + |1 - a/H(x)|$ for $a > 0$ and $x \in R$, then for any fixed $a > 0$, $\psi(x,a)$ is a strictly convex function in $x \in R$.*

Proof of Theorem 1. Recall that $\{\tilde\beta, \tilde a, \tilde b\}$ is the minimizer of (2.1). Let $\tilde\theta = (\tilde\beta^T, \tilde a, h\tilde b)^T$ with the true value $\theta_0 = (\beta_0^T, g(t), hg'(t))^T, K_i(t) = K((t_i - t)/h), X_i = (x_i^T, 1, (t_i - t)/h)^T$, and $\Delta_{it} = g(t) + g'(t)(t_i - t) - g(t_i)$. Then, $\tilde\theta$ is also the minimizer of

$$\phi_n(\theta) = \sum_{i=1}^{n} \Big\{ \Big| 1 - \frac{H\{X_i^T(\theta - \theta_0) + \Delta_{it}\}}{\varepsilon_i} \Big| + \Big| 1 - \frac{\varepsilon_i}{H\{X_i^T(\theta - \theta_0) + \Delta_{it}\}} \Big| \Big\} K_i(t).$$

We prove Theorem 1 in two steps. We first show that the estimator is $\sqrt{nh}$-consistent, then establish its asymptotic normality. For ease of presentation, we sometimes denote the matrix vv^T by v^2 for a vector v.

Step 1: consistency

It suffices to show that for any given $\epsilon > 0$, there exist a large constant $r > 0$ such that

$$P\Big(\inf_{||\boldsymbol{u}|| = r} \phi_n(\theta_0 + \boldsymbol{u}/\sqrt{nh}) > \phi_n(\theta_0) \Big) \geq 1 - \epsilon. \tag{A.1}$$

This implies that with probability at least $1 - \epsilon$, $\phi_n(\theta)$ has a local minimum $\tilde\theta$ that satisfies $\tilde\theta - \theta_0 = O_p\big((nh)^{-1/2}\big)$, see Fan and Li (2001).

By applying the identity in Knight (1998),

$$|x - y| - |x| = -y[I(x > 0) - I(x < 0)] + 2\int_0^y [I(x \leq s) - I(x \leq 0)]ds,$$

valid for $x \neq 0$, we have

$$\phi_n(\theta_0 + \frac{\boldsymbol{u}}{\sqrt{nh}}) - \phi_n(\theta_0)$$

$$
\begin{aligned}
&= -\sum_i^n \varpi_i K_i(t)\Big[I\{1-\varepsilon_i^{-1}H(\Delta_{it})>0\} - I\{1-\varepsilon_i^{-1}H(\Delta_{it})<0\}\Big] \\
&\quad +2\sum_i^n K_i(t)\int_0^{\varpi_i}\Big[I\{1-\varepsilon_i^{-1}H(\Delta_{it})\le s\} - I\{1-\varepsilon_i^{-1}H(\Delta_{it})\le 0\}\Big]ds \\
&\quad -\sum_i^n \zeta_i K_i(t)\Big[I\{1-\varepsilon_i H^{-1}(\Delta_{it})>0\} - I\{1-\varepsilon_i H^{-1}(\Delta_{it})<0\}\Big] \\
&\quad +2\sum_i^n K_i(t)\int_0^{\zeta_i}\Big[I\{1-\varepsilon_i H^{-1}(\Delta_{it})\le s\} - I\{1-\varepsilon_i H^{-1}(\Delta_{it})\le 0\}\Big]ds \\
&:= I_{1n}+I_{2n}+I_{3n}+I_{4n}, \qquad \text{(A.2)}
\end{aligned}
$$

where

$$
\begin{aligned}
\varpi_i &= \varepsilon_i^{-1}\Big\{H(\frac{X_i^T\boldsymbol{u}}{\sqrt{nh}}+\Delta_{it}) - H(\Delta_{it})\Big\}, \\
\zeta_i &= \varepsilon_i\Big\{H^{-1}(\frac{X_i^T\boldsymbol{u}}{\sqrt{nh}}+\Delta_{it}) - H^{-1}(\Delta_{it})\Big\}.
\end{aligned}
$$

Recall that $\gamma = H'(0)/H(0)$. Let $\Lambda(t)=\mathrm{diag}\{E[(x_i^T,1)^T(x_i^T,1)|T=t],\mu_2\}$ and

$$
W_n = \frac{\gamma}{\sqrt{nh}}\sum_i^n K_i(t)\{\varepsilon_i H^{-1}(0)+\varepsilon_i^{-1}H(0)\}sgn(1-\varepsilon_i H^{-1}(\Delta_{it}))X_i.
$$

We show that

$$
\phi_n(\theta_0+\frac{\boldsymbol{u}}{\sqrt{nh}})-\phi_n(\theta_0) = W_n^T\boldsymbol{u}+\gamma^2 f_T(t)[J+2f(H(0))H(0)]\boldsymbol{u}^T\Lambda_t\boldsymbol{u}+o_p(1). \quad \text{(A.3)}
$$

To prove (A.3), we first show that

$$
I_{1n}+I_{3n} = W_n^T\boldsymbol{u}+\gamma^2 f_T(t)J\boldsymbol{u}^T\Lambda(t)\boldsymbol{u}+o_p(1). \qquad \text{(A.4)}
$$

Owing to the fact that $\|\boldsymbol{u}\| = r$, $h\to 0$, and $nh\to\infty$, with the standard Taylor expansion we have

$$
\begin{aligned}
I_{1n}+I_{3n} &= \frac{1}{\sqrt{nh}}\sum_i^n K_i(t)\big[\varepsilon_i^{-1}H'(0)+\varepsilon_i\frac{H'(0)}{H^2(0)}\big]sgn(1-\varepsilon_i H^{-1}(\Delta_{it}))X_i^T\boldsymbol{u} \\
&\quad +\frac{1}{\sqrt{nh}}\sum_i^n K_i(t)\big\{\varepsilon_i^{-1}[H'(\Delta_{it})-H'(0)]+\varepsilon_i[\frac{H'(\Delta_{it})}{H^2(\Delta_{it})}-\frac{H'(0)}{H^2(0)}]\big\} \\
&\quad \times sgn(1-\varepsilon_i H^{-1}(\Delta_{it}))X_i^T\boldsymbol{u} \\
&\quad +\boldsymbol{u}^T\Big\{\frac{1}{2nh}\sum_i^n K_i(t)\big[\varepsilon_i\frac{H''(0)}{H^2(0)}+\varepsilon_i^{-1}H''(0)-\varepsilon_i\frac{2H'^2(0)}{H^3(0)}\big]
\end{aligned}
$$

$$
\begin{aligned}
&\times sgn(1-\varepsilon_i H^{-1}(\Delta_{it}))X_iX_i^T\Big\}\boldsymbol{u}\\
&+\boldsymbol{u}^T\Big\{\frac{1}{2nh}\sum_i^n K_i(t)\{\varepsilon_i[\frac{H''(\xi_i^{[2]})}{H^2(\xi_i^{[2]})}-\frac{2H'^2(\xi_i^{[2]})}{H^3(\xi_i^{[2]})}-\frac{H''(0)}{H^2(0)}+\frac{2H'^2(0)}{H^3(0)}]\\
&+\varepsilon_i^{-1}[H''(\xi_i^{[1]})-H''(0)]\}sgn(1-\varepsilon_i H^{-1}(\Delta_{it}))X_iX_i^T\Big\}\boldsymbol{u}\\
:=\ & I_{13}^{[1]}+I_{13}^{[2]}+\boldsymbol{u}^T I_{13}^{[3]}\boldsymbol{u}+\boldsymbol{u}^T I_{13}^{[4]}\boldsymbol{u}, \qquad (A.5)
\end{aligned}
$$

where $\xi_i^{[1]}$ and $\xi_i^{[2]}$ lie between $X_i^T\boldsymbol{u}/\sqrt{nh}+\Delta_{it}$ and Δ_{it}. Obviously, $I_{13}^{[1]}=W_n^T\boldsymbol{u}$.

Together with (C5) and (C7), we have $\Delta_{it}=O(h^2)$. Clearly, by Lemma 1 and (C4) we have

$$
\begin{aligned}
I_{13}^{[3]} &= \frac{1}{2nh}\sum_i^n K_i(t)[\varepsilon_i\frac{H''(0)}{H^2(0)}+\varepsilon_i^{-1}H''(0)-\varepsilon_i\frac{2H'^2(0)}{H^3(0)}]\\
&\quad\times \mathrm{sgn}(1-\varepsilon_i H^{-1}(\Delta_{it}))X_iX_i^T\\
&=\frac{1}{2}E\Big[K_h(t_i-t)[\varepsilon_i\frac{H''(0)}{H^2(0)}+\varepsilon_i^{-1}H''(0)-\varepsilon_i\frac{2H'^2(0)}{H^3(0)}]\\
&\quad\times \mathrm{sgn}(1-\varepsilon_i H^{-1}(\Delta_{it}))X_iX_i^T\Big]+o_p(1)\\
&=\frac{1}{2}E\Big[K_h(t_i-t)[\varepsilon_i\frac{H''(0)}{H^2(0)}+\varepsilon_i^{-1}H''(0)-\varepsilon_i\frac{2H'^2(0)}{H^3(0)}]\\
&\quad\times \mathrm{sgn}(1-\varepsilon_i H^{-1}(0))X_iX_i^T\Big]+o_p(1)\\
&=\gamma^2 Jf_T(t)\ \mathrm{diag}\{E[(x_i^T,1)^T(x_i^T,1)|T=t],\mu_2\}+o_p(1). \qquad (A.6)
\end{aligned}
$$

Similarly, we can prove $I_{13}^{[2]}=o_p(1)$ and $I_{13}^{[4]}=o_p(1)$ based on (C8). Therefore, combining (A.5) and (A.6), (A.4) is proved.

Next we focus on I_{2n} and I_{4n}. Let $\ell_i(\boldsymbol{u})=H(X_i^T\boldsymbol{u}/\sqrt{nh}+\Delta_{it})-H(\Delta_{it})$. Note that

$$
\begin{aligned}
I_{2n}&=2\sum_i^n K_i(t)\int_0^{\varpi_i}\Big[I\{1-\varepsilon_i^{-1}H(\Delta_{it})\le s\}-I\{1-\varepsilon_i^{-1}H(\Delta_{it})\le 0\}\Big]ds\\
&=2\sum_i^n K_i(t)\int_0^{\ell_i(\boldsymbol{u})}\varepsilon_i^{-1}\Big[I\{\varepsilon_i\le H(\Delta_{it})+\delta\}-I(\varepsilon_i\le H(\Delta_{it})\Big]d\delta.
\end{aligned}
$$

The second equality follows the change of variable $\delta=s\varepsilon_i$. Denote its last expression by $B_n^*(\boldsymbol{u})$. Since $B_n^*(\boldsymbol{u})$ is a summation of i.i.d random variables of kernel form, it follows by Lemma 1 that

$$
B_n^*(\boldsymbol{u})=E[B_n^*(\boldsymbol{u})]+O_p\Big(\frac{\log^{1/2}(1/h)}{\sqrt{nh}}\Big).
$$

It then follows that

$$
\begin{aligned}
I_{2n} &= 2\sum_i^n K_i(t)\int_0^{\ell_i(\boldsymbol{u})} E_{\varepsilon|X}\Big\{\varepsilon_i^{-1}\Big[I\{\varepsilon_i \le H(\Delta_{it})+\delta\} - I(\varepsilon_i \le H(\Delta_{it})\Big]\Big\}d\delta + o_p(1)\\
&= 2\sum_i^n K_i(t)H^{-1}(0)\int_0^{\ell_i(\boldsymbol{u})} E_{\varepsilon|X}\Big[I\{\varepsilon_i \le H(\Delta_{it})+\delta\} - I(\varepsilon_i \le H(\Delta_{it})\Big]d\delta\\
&\quad +2\sum_i^n K_i(t)\int_0^{\ell_i(\boldsymbol{u})} E_{\varepsilon|X}\Big[(\varepsilon_i^{-1} - H^{-1}(0))\big[I\{\varepsilon_i \le H(\Delta_{it})+\delta\}\\
&\qquad -I(\varepsilon_i \le H(\Delta_{it})\big]\Big]d\delta + o_p(1)\\
&= \boldsymbol{u}^T\Big\{\frac{1}{nh}\sum_i^n K_i(t)\frac{H'^2(0)}{H(0)}f(H(0))X_iX_i^T\Big\}\boldsymbol{u}[1+o_p(1)]\\
&= \gamma^2 H(0)f(H(0))f_T(t)\boldsymbol{u}^T\Lambda\boldsymbol{u} + o_p(1). \qquad (A.7)
\end{aligned}
$$

Similarly, it can be proved that

$$
I_{4n} = \gamma^2 H(0)f(H(0))f_T(t)\boldsymbol{u}^T\Lambda\boldsymbol{u} + o_p(1). \qquad (A.8)
$$

This together with (A.4) and (A.7) proves (A.3) as follows,

$$
\phi_n(\theta_0 + \frac{\boldsymbol{u}}{\sqrt{nh}}) - \phi_n(\theta_0) = W_n^T\boldsymbol{u} + \gamma^2 f_T(t)[J + 2f(H(0))H(0)]\boldsymbol{u}^T\Lambda_t\boldsymbol{u} + o_p(1).
$$

Here the quadratic term dominates the linear term uniformly in $||\boldsymbol{u}|| = r$ for sufficient large r. Therefore, (A.1) holds for sufficient large r, which completes the proof of consistency.

Step 2: asymptotic normality

Since $\phi_n(\theta)$ is a strictly convex function of θ by Lemma 2, the local minimizer $\tilde{\theta}$ in Step 1 is also the global minimizer. Applying the epi-convergence results of Knight and Fu (2000), it can be shown that $\sqrt{nh}\{\tilde{\theta}-\theta_0\}$ has the same asymptotic distribution as

$$
-\frac{1}{2}\Big\{\gamma^2 f_T(t)[J + 2f(H(0))H(0)]\Big\}^{-1}\Lambda_t^{-1}W_n. \qquad (A.9)
$$

Let

$$
W_n^{\dagger} = \frac{\gamma}{\sqrt{nh}}\sum_i^n K_i(t)\{\varepsilon_i H^{-1}(0) + \varepsilon_i^{-1}H(0)\}sgn(1 - \varepsilon_i H^{-1}(0))X_i,
$$

and recall that $\gamma = H'(0)/H(0)$ and $A = E(\varepsilon_i H^{-1}(0) + \varepsilon_i^{-1}H(0))^2$. By (C3) and (C4), we have $E[W_n^{\dagger}] = 0$ and

$$
Var[W_n^{\dagger}] = \gamma^2 E(\varepsilon_i H^{-1}(0) + \varepsilon_i^{-1}H(0))^2 E[X_iX_i^T\frac{K_i^2(t)}{h}]
$$

$$
\begin{aligned}
&= \gamma^2 A E_{t_i} E_{X_i|t_i}[X_i X_i^T \frac{K_i^2(t)}{h}] \\
&\to \gamma^2 A f_T(t) \text{ diag}\{\nu_0 E[(x_i^T, 1)^T (x_i^T, 1)|T = t], \nu_2\}.
\end{aligned}
$$

Together with the Cramér-Wold Theorem and $V_t = E[(x_i^T, 1)^T (x_i^T, 1)|T = t]$, the CLT for $W_n^\dagger$ holds:

$$
W_n^\dagger \xrightarrow{d} N(0, \gamma^2 A f_T(t) \text{ diag}\{\nu_0 V_t, \nu_2\}).
$$

Let $\hbar_i = sgn(1 - \varepsilon_i H^{-1}(\Delta_{it})) - sgn(1 - \varepsilon_i H^{-1}(0))$. We know that if ε_i lies between $H(0)$ and $H(\Delta_{it})$, $|\hbar_i| = 2$, and otherwise $|\hbar_i| = 0$. Thus

$$
\begin{aligned}
Var(W_n - W_n^\dagger) &= Var\Big\{\frac{\gamma}{\sqrt{nh}} \sum_i^n K_i(t)\{\varepsilon_i H^{-1}(0) + \varepsilon_i^{-1} H(0)\}\hbar_i X_i\Big\} \\
&\le \frac{\gamma^2}{h} E\Big\{K_i^2(t)\{\varepsilon_i H^{-1}(0) + \varepsilon_i^{-1} H(0)\}^2 \hbar_i^2 X_i X_i^T\Big\} \\
&\le \frac{4N_0^2\gamma^2}{h} |F(H(\Delta_{it})) - F(H(0))| E\Big\{K_i^2(t) X_i X_i^T\Big\} \\
&\to \frac{4N_0^2\gamma^2}{h} f_T(t) \text{ diag}\{\nu_0 V_t, \nu_2\} f(H(0))|H'(0)||\Delta_{it}| \\
&= O(h^2) \to 0.
\end{aligned}
$$

Here the second inequality follows from (C3), and the third inequality holds due to the fact that $\Delta_{it} = O(h^2)$ and there exists a $N_0 > 0$ such that $\varepsilon_i H^{-1}(0) + \varepsilon_i^{-1} H(0) < N_0$ in the interval between $H(0)$ and $H(\Delta_{it})$.

Under the conditions in the Appendix, we have

$$
\begin{aligned}
E(\frac{W_n}{\sqrt{nh}}) &= \frac{\gamma}{h} E\Big[K_i(t)\{\varepsilon_i H^{-1}(0) + \varepsilon_i^{-1} H(0)\} sgn(1 - \varepsilon_i H^{-1}(\Delta_{it})) X_i\Big] \\
&= \gamma E\Big[K_h(t_i - t)\{\varepsilon_i H^{-1}(0) + \varepsilon_i^{-1} H(0)\} sgn(1 - \varepsilon_i H^{-1}(0)) X_i\Big] \\
&\quad + \gamma E\Big[K_h(t_i - t)\{\varepsilon_i H^{-1}(0) + \varepsilon_i^{-1} H(0)\}\hbar_i X_i\Big] \\
&= \gamma E\Big[K_h(t_i - t)\{\varepsilon_i H^{-1}(0) + \varepsilon_i^{-1} H(0)\}\hbar_i X_i\Big] \\
&= 2\gamma E\Big[K_h(t_i - t) I\{H(0) < \varepsilon_i < H(\Delta_{it})\} X_i\Big](1 + O(h^2)) \\
&= -\gamma \mu_2 h^2 f(H(0)) H'(0) g''(t) f_T(t) E[(x^T, 1, 0)^T | T = t] + o(h^2).
\end{aligned}
$$

Then using Slutsky's Theorem, we have

$$
W_n - E(W_n) \xrightarrow{d} N(0, \gamma^2 A f_T(t) \text{ diag}\{\nu_0 V_t, \nu_2\}). \tag{A.10}
$$

Thus, we obtain

$$\sqrt{nh}\{\tilde{\theta} - \theta_0 - \frac{\mu_2 h^2 f(H(0))H(0)g''(t)}{2(J + 2f(H(0))H(0))}\Lambda_t^{-1}E[(x^T, 1, 0)^T|T = t]\}$$
$$\xrightarrow{d} N\Big(0, \frac{1}{4\gamma^2 f_T(t)}[J + 2f(H(0))H(0)]^{-2}\Lambda_t^{-1}A \operatorname{diag}\{\nu_0 V_t, \nu_2\}\Lambda_t^{-1}\Big).$$

This completes the proof.

Proof of Theorem 2. Let $\rho_i = \tilde{g}(t_i) - g(t_i)$ and

$$L_n(\beta) = \sum_{i=1}^n \Big\{\Big|1 - \varepsilon_i^{-1}H\{x_i^T(\beta - \beta_0) + \rho_i\}\Big| + \Big|1 - \varepsilon_i H^{-1}\{x_i^T(\beta - \beta_0) + \rho_i\}\Big|\Big\}.$$

Consider $\theta^* = \sqrt{n}(\beta - \beta_0)$. Let

$$\delta_i = \varepsilon_i^{-1}\Big\{H(\frac{x_i^T\theta^*}{\sqrt{n}} + \rho_i) - H(\rho_i)\Big\},\ \eta_i = \varepsilon_i\Big\{H^{-1}(\frac{x_i^T\theta^*}{\sqrt{n}} + \rho_i) - H^{-1}(\rho_i)\Big\}.$$

Since the proof here is similar to that of Theorem 1, we only detail some differences.

Step 1: consistency

It suffices to show that for any given $\epsilon > 0$, there exist a large constant $r > 0$ such that

$$P\Big(\inf_{||\theta^*||=r} L_n(\beta_0 + \frac{\theta^*}{\sqrt{n}}) > L_n(\theta_0)\Big) \geq 1 - \epsilon. \tag{A.11}$$

This implies that, with probability at least $1 - \epsilon$, $L_n(\beta)$ has a local minimizer $\hat{\beta}$ satisfying $\hat{\beta} - \beta_0 = O_p\big(n^{-1/2}\big)$.

Then by the identity in Knight (1998), we have

$$\begin{aligned}
&L_n(\beta_0 + \frac{\theta^*}{\sqrt{n}}) - L_n(\beta_0)\\
&\quad = -\sum_i^n \delta_i \operatorname{sgn}\{1 - \varepsilon_i^{-1}H(\rho_i)\} - \sum_i^n \eta_i \operatorname{sgn}\{1 - \varepsilon_i H^{-1}(\rho_i)\}\\
&\qquad +2\sum_i^n \int_0^{\delta_i} \Big[I\{1 - \varepsilon_i^{-1}H(\rho_i) \leq s\} - I\{1 - \varepsilon_i^{-1}H(\rho_i) \leq 0\}\Big]ds\\
&\qquad +2\sum_i^n \int_0^{\eta_i} \Big[I\{1 - \varepsilon_i H^{-1}(\rho_i) \leq s\} - I\{1 - \varepsilon_i H^{-1}(\rho_i) \leq 0\}\Big]ds\\
&\quad := T_{1n} + T_{2n} + T_{3n} + T_{4n}.
\end{aligned} \tag{A.12}$$

Here we calculate T_{2n} in detail. Note that by a Taylor expansion, we have

$$H^{-1}(\frac{x_i^T\theta^*}{\sqrt{n}} + \rho_i) - H^{-1}(\rho_i)$$

$$
\begin{aligned}
&= -\frac{H'(0)}{H^2(0)}\frac{x_i^T\theta^*}{\sqrt{n}} - \frac{H''(\tau_1)H(\tau_{i1}) - 2H'^2(\tau_{i1})}{H^3(\tau_{i1})}\rho_i\frac{x_i^T\theta^*}{\sqrt{n}}\\
&\quad -\frac{H''(\tau_{i2})H(\tau_{i2}) - 2H'^2(\tau_{i2})}{H^3(\tau_{i2})}\frac{\theta^{*T}x_ix_i^T\theta^*}{2n}\\
&= -\frac{H'(0)}{H^2(0)}\frac{x_i^T\theta^*}{\sqrt{n}} - \frac{H''(0)H(0) - 2H'^2(0)}{H^3(0)}\rho_i\frac{x_i^T\theta^*}{\sqrt{n}}\\
&\quad -\frac{H''(0)H(0) - 2H'^2(0)}{H^3(0)}\frac{\theta^{*T}x_ix_i^T\theta^*}{2n}\\
&\quad -\Big\{\frac{H''(\tau_{i1})H(\tau_{i1}) - 2H'^2(\tau_{i1})}{H^3(\tau_{i1})} - \frac{H''(0)H(0) - 2H'^2(0)}{H^3(0)}\Big\}\rho_i\frac{x_i^T\theta^*}{\sqrt{n}}\\
&\quad -\Big\{\frac{H''(\tau_{i2})H(\tau_{i2}) - 2H'^2(\tau_{i2})}{H^3(\tau_{i2})} - \frac{H''(0)H(0) - 2H'^2(0)}{H^3(0)}\Big\}\frac{\theta^{*T}x_ix_i^T\theta^*}{2n}\\
&=: D_{i21} + D_{i22} + D_{i23} + D_{i24} + D_{i25},
\end{aligned}
$$

where τ_{i1} lies between 0 and ρ_i while τ_{i2} lies between $X_i^T\theta^*/\sqrt{n}+\rho_i$ and ρ_i. From the proof of Theorem 1, we have

$$
\sup_{t\in\Omega}\Big|\tilde{g}(t) - g(t) - \frac{1}{nQf_T(t)}\sum_{i=1}^n G_iK_h(t_i - t)\Big| = o_p(\{nh\}^{-1/2}),
$$

where $Q = 2\gamma[J + 2f(H(0))H(0)]$ and G_i is the last element of the vector $-\{\varepsilon_iH^{-1}(0) + \varepsilon_i^{-1}H(0)\}\mathrm{sgn}(1 - \varepsilon_iH^{-1}(\Delta_{it}))V_t^{-1}(x_i^T, 1)^T$.

Write $\iota = (H''(0)H(0) - 2H'^2(0))/H^3(0)$. Since $H(\cdot)$ is second-order continuous differentiable and $H''(\cdot)$ is uniformly continuous on $[-M, M]$, we have

$$
\begin{aligned}
T_{2n} &= \Big\{-\sum_i^n\varepsilon_i\{D_{i21} + D_{i22} + D_{i23}\}\mathrm{sgn}\{1 - \varepsilon_iH^{-1}(0)\}\Big\}(1 + O_p(c_n))\\
&= \Big\{\frac{1}{\sqrt{n}}\sum_i^n\varepsilon_i\frac{H'(0)}{H^2(0)}\mathrm{sgn}\{1 - \varepsilon_iH^{-1}(0)\}x_i\Big\}^T\theta^*\\
&\quad +\iota\Big\{\frac{1}{\sqrt{n}}\sum_i^n\varepsilon_i\mathrm{sgn}\{1 - \varepsilon_iH^{-1}(0)\}\Big(\frac{1}{nQf_T(t_i)}\sum_{j=1}^n G_jK_h(t_j - t_i)\Big)x_i\Big\}^T\theta^*\\
&\quad +\iota\theta^{*T}\Big\{\frac{1}{2n}\sum_i^n\varepsilon_i\mathrm{sgn}\{1 - \varepsilon_iH^{-1}(0)\}x_ix_i^T\Big\}\theta^* + o_p(1).
\end{aligned}
$$

Denote the second term in the foregoing expression as $\iota Z_n^T\theta^*$. Then Z_n can be expressed as

$$
Z_n = \frac{1}{\sqrt{n}}\sum_i^n\varepsilon_i\mathrm{sgn}\{1 - \varepsilon_iH^{-1}(0)\}\Big(\frac{1}{nQf_T(t_i)}\sum_{j=1}^n G_jK_h(t_j - t_i)\Big)x_i
$$

$$
\begin{aligned}
&= -\frac{1}{\sqrt{n}Q}\sum_j^n \{\varepsilon_j H^{-1}(0) + \varepsilon_j^{-1}H(0)\}\text{sgn}\{1-\varepsilon_j H^{-1}(0)\}(0_p^T, 1) \\
&\quad \times\Big(\frac{1}{n}\sum_{i=1}^n \frac{1}{f_T(t_i)\varepsilon_i}\text{sgn}\{1-\varepsilon_i H^{-1}(0)\}K_h(t_i-t_j)V_{t_i}^{-1}(x_j^T,1)^T x_i\Big) + o_p(1) \\
&= \frac{JH(0)}{\sqrt{n}Q}\sum_j^n \Big(\{\varepsilon_j H^{-1}(0) + \varepsilon_j^{-1}H(0)\}\text{sgn}\{1-\varepsilon_j H^{-1}(0)\} \\
&\quad \times(0_p^T,1)V_{t_j}^{-1}(x_j^T,1)^T E[x|t_j]\Big) + o_p(1).
\end{aligned}
$$

where the third equality holds by Lemma 1.

Therefore,

$$
\begin{aligned}
T_{2n} &= \Big\{\frac{\gamma}{\sqrt{n}}\sum_i^n \varepsilon_i H^{-1}(0)\text{sgn}\{1-\varepsilon_i H^{-1}(0)\}x_i\Big\}^T\theta^* \\
&\quad + \frac{J}{Q}\Big[\frac{H''(0)}{H(0)} - 2\gamma^2\Big]\Big\{\frac{1}{\sqrt{n}}\sum_j^n \{\varepsilon_j H^{-1}(0) + \varepsilon_j^{-1}H(0)\} \qquad (A.13) \\
&\quad \times\text{sgn}\{1-\varepsilon_j H^{-1}(0)\}(0_p^T,1)V_{t_j}^{-1}(x_j^T,1)^T E[x|t_j]\Big\}^T\theta^* \\
&\quad + \Big[\frac{H''(0)}{H(0)} - 2\gamma^2\Big]\theta^{*T}\Big\{\frac{1}{2n}\sum_i^n \varepsilon_i H^{-1}(0)\text{sgn}\{1-\varepsilon_i H^{-1}(0)\}x_i x_i^T\Big\}\theta^* + o_p(1).
\end{aligned}
$$

Meanwhile, we find

$$
\begin{aligned}
T_{1n} &= \Big\{-\frac{1}{\sqrt{n}}\sum_i^n \varepsilon_i^{-1}H'(0)\text{sgn}\{1-\varepsilon_i^{-1}H(0)\}x_i\Big\}^T\theta^* \\
&\quad - \frac{J}{Q}\frac{H''(0)}{H(0)}\Big\{\frac{1}{\sqrt{n}}\sum_j^n \{\varepsilon_j^{-1}H(0) + \varepsilon_j^{-1}H(0)\} \qquad (A.14) \\
&\quad \times\text{sgn}\{1-\varepsilon_j H^{-1}(0)\}(0_p^T,1)V_{t_j}^{-1}(x_j^T,1)^T E[x|t_j]\Big\}^T\theta^* \\
&\quad + \frac{H''(0)}{H(0)}\theta^{*T}\Big\{-\frac{1}{2n}\sum_i^n \varepsilon_i^{-1}H(0)\text{sgn}\{1-\varepsilon_i^{-1}H(0)\}x_i x_i^T\Big\}\theta^* + o_p(1).
\end{aligned}
$$

Let

$$
\eta_1(t,x) = \frac{J}{J+2f(H(0))H(0)}E[x(0_p^T,1)|T=t]V_t^{-1}(x^T,1)^T.
$$

Then (A.13) and (A.14) together give

$$
T_{1n}+T_{2n} = \Big\{\frac{\gamma}{\sqrt{n}}\sum_i^n \{\varepsilon_i H^{-1}(0) + \varepsilon_i^{-1}H(0)\}\text{sgn}\{1-\varepsilon_i H^{-1}(0)\}x_i\Big\}^T\theta^*
$$

$$
\begin{aligned}
&-\frac{J\gamma}{J+2f(H(0))H(0)}\Big\{\frac{1}{\sqrt{n}}\sum_j^n\{\varepsilon_j^{-1}H(0)+\varepsilon_j^{-1}H(0)\}\\
&\quad\times\mathrm{sgn}\{1-\varepsilon_jH^{-1}(0)\}(0_p^T,1)V_{t_j}^{-1}(x_j^T,1)^TE[x|t_j]\Big\}^T\theta^*\\
&+\theta^{*T}\Big\{\frac{\gamma^2}{n}\sum_i^n\varepsilon_iH^{-1}(0)\mathrm{sgn}\{\varepsilon_iH^{-1}(0)-1\}x_ix_i^T\Big\}\theta^*+o_p(1)\\
&=S_n^T\theta^*+\theta^{*T}\gamma^2JC\theta^*+o_p(1),
\end{aligned}
$$

where

$$
S_n=\frac{\gamma}{\sqrt{n}}\sum_i^n\{\varepsilon_iH^{-1}(0)+\varepsilon_i^{-1}H(0)\}\mathrm{sgn}\{1-\varepsilon_iH^{-1}(0)\}[x_i-\eta_1(t_i,x_i)].
$$

Write $\vartheta_i=H^{-1}(x_i^T\theta^*/\sqrt{n}+\rho_i)-H^{-1}(\rho_i)$. With some calculation, we have

$$
\begin{aligned}
T_{4n}&=2H(0)\sum_i^n\int_0^{\vartheta_i}E_{\varepsilon_i|x,T}\Big[I\{1-\varepsilon_iH^{-1}(\rho_i)\le\varepsilon_i\delta\}\\
&\quad-I\{1-\varepsilon_iH^{-1}(\rho_i)\le 0\}\Big]d\delta+o_p(1)\\
&=2H(0)\sum_i^n\int_0^{\vartheta_i}\Big[f(H(\rho_i))\frac{H^2(\rho_i)\delta}{H(\rho_i)\delta+1}\{1+o(1)\}\Big]d\delta\\
&=\theta^{*T}\gamma^2f(H(0))H(0)C\theta^*+o_p(1).
\end{aligned}
$$

Similarly, we obtain $T_{3n}=\theta^{*T}\gamma^2f(H(0))H(0)C\theta^*+o_p(1)$. Therefore, $L_n(\beta_0+\theta^*/\sqrt{n})-L_n(\beta_0)$ can be represented as

$$
S_n^T\theta^*+\theta^{*T}\gamma^2[J+2f(H(0))H(0)]C\theta^*+o_p(1).
$$

Since the quadratic term dominates the linear term uniformly in $||\theta^*||=r$ for sufficient large r, (A.11) holds for sufficient large r, which completes the proof of consistency.

Step 2: asymptotic normality

Lemma 2 implies that $L_n(\beta)$ is a strictly convex function of β. Thus the local minimizer $\hat\beta$ in Step 2 is also the global minimizer. Moreover, it is easy to see that

$$
S_n\to_d S\ \text{ with }\ S\sim N\Big(0,\gamma^2AE[\{\mathbf{x}-\eta(\mathbf{t},\mathbf{x})\}\{\mathbf{x}-\eta(\mathbf{t},\mathbf{x})\}^T]\Big).
$$

By the epi-convergence results of Knight and Fu (2000), we have

$$
\sqrt{n}\{\hat\beta-\beta_0\}\to_d-\frac{1}{2}\Big\{\gamma^2[J+2f(H(0))H(0)]\Big\}^{-1}C^{-1}S. \tag{A.15}
$$

The proof is then complete.

Proof of Theorem 4. Let $\beta = \beta_0 + \xi/\sqrt{n}$ and

$$\Psi_n(\xi) = \sum_{i=1}^{n} (u_i - x_i^T(\beta_0 + \frac{\xi}{\sqrt{n}}))^2 + \lambda_n \sum_{j=1}^{p} \omega_j |\beta_0 + \frac{\xi}{\sqrt{n}}|.$$

If $\hat{\xi} = \operatorname{argmin}_\xi \Psi_n(\xi)$, then by (4.2) we have $\hat{\beta}^{\lambda_n} = \beta_0 + \hat{\xi}/\sqrt{n}$. Consider $V_n(\xi) = \Psi_n(\xi) - \Psi_n(0)$, where

$$V_n(\xi) = \xi^T C_n \xi - 2\xi^T C_n\{\sqrt{n}(\hat{\beta} - \beta_0)\} + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \omega_j \sqrt{n} \Big\{ |\beta_0 + \frac{\xi}{\sqrt{n}}| - |\beta_0| \Big\}. \tag{A.16}$$

Indeed, we have that $C_n \to C$ and $C_n\sqrt{n}(\hat{\beta} - \beta_0) \to_d W$ with

$$W \sim N\Big(0, \frac{1}{4}\big\{\gamma[J + 2f(H(0))H(0)]\big\}^{-2} A\Xi\Big)$$

by Theorem 2 and Slutsky's Lemma. Now we consider the asymptotic behavior of the third term of the right hand side of (A.16). If $\beta_{j0} = 0$, then

$$\sqrt{n}\Big(|\beta_{j0} + \frac{\xi_j}{\sqrt{n}}| - |\beta_{j0}|\Big) = sgn(\xi_j)\xi_j,$$

$$\frac{\lambda_n}{\sqrt{n}} \omega_j = \lambda_n n^{(\tau-1)/2} (|\sqrt{n}\hat{\beta}_j|)^{-\tau}.$$

By Theorem 2 we have $\sqrt{n}\hat{\beta}_j = O_p(1)$. This together with the condition that $\lambda_n n^{(\tau-1)/2} \to \infty$ gives

$$\frac{\lambda_n}{\sqrt{n}} \omega_j \sqrt{n} \Big\{ |\beta_0 + \frac{\xi}{\sqrt{n}}| - |\beta_0| \Big\} \to_p \begin{cases} 0, & \text{if } \xi_j = 0; \\ \infty, & \text{if } \xi_j \neq 0. \end{cases}$$

If $\beta_{j0} \neq 0$, then $\omega_j \to_p |\beta_{j0}|^{-\tau}$ and

$$\sqrt{n}\Big(|\beta_{j0} + \frac{\xi_j}{\sqrt{n}}| - |\beta_{j0}|\Big) = sgn(\beta_{j0})\xi_j.$$

By Slutsky's theorem, it can be shown that $V_n(\xi) \to_d V(\xi)$, where

$$V(\xi) = \begin{cases} \xi_\aleph^T C_{\aleph\aleph} \xi_\aleph - 2\xi_\aleph W_\aleph, & \text{if } \xi_j = 0 \text{ for all } j \notin \aleph, \\ \infty, & \text{otherwise.} \end{cases}.$$

Obviously, $V(\xi)$ is convex. Applying the epi-convergence results of Geyer (1994) and Knight and Fu (2000), we have

$$\hat{\xi}_\aleph \to_d C_{\aleph\aleph}^{-1} W_\aleph \quad \text{and} \quad \hat{\xi}_{\aleph^c} \to_d 0 \tag{A.17}$$

The asymptotic normality is then established.

The consistency of variable selection can be seen as follows. Since the asymptotic property result indicates that $\hat{\beta}_j^{\lambda_n} \to_d \beta_{0j}$, $P(j \in \hat{\aleph}) \to 1$ for $\forall j \in \aleph$. Therefore, it suffices to show that $\forall j' \in \aleph^c, P(j' \in \hat{\aleph}) \to 0$.

Suppose $j' \in \hat{\aleph}$. Then

$$\begin{aligned} 0 = \frac{1}{\sqrt{n}} \frac{\partial H_n(\beta)}{\partial \beta_{j'}} &= -2\frac{1}{n}\sum_{i=1}^{n} x_{ij'} x_i^T \{\sqrt{n}(\hat{\beta} - \hat{\beta}^{\lambda_n})\} + \frac{\lambda_n}{\sqrt{n}} \omega_{j'} sgn(\hat{\beta}_{j'}^{\lambda_n}) \\ &= -2C_n^{(j')}\{\sqrt{n}(\hat{\beta} - \hat{\beta}^{\lambda_n})\} + \frac{\lambda_n}{\sqrt{n}} \omega_{j'} sgn(\hat{\beta}_{j'}^{\lambda_n}) \end{aligned}$$

where $C_n^{(j')}$ stands for the j'th row of C_n. Note that

$$\frac{\lambda_n \omega_{j'}}{\sqrt{n}} = \lambda_n n^{(\tau-1)/2} (|\sqrt{n}\hat{\beta}_{j'}|)^{-\tau} \to_p \infty,$$

$C_n^{(j')} \to C^{(j')}$, and $\sqrt{n}(\hat{\beta} - \hat{\beta}^{\lambda_n}) = O_p(1)$, by Theorem 2 and (A.17). Therefore, the right hand of $(1/\sqrt{n})\frac{\partial H_n(\beta)}{\partial \beta_{j'}}$ goes to infinity for n sufficiently large. This contradiction proves the consistency of variable selection.

Proof of Theorem 5. Let $\beta = \beta_0 + \eta/\sqrt{n}$. Then $T_n(\beta)$ at (4.3) can be rewritten as

$$\begin{aligned} \Phi_n(\eta) = \sum_{i=1}^{n} \Big\{ \Big| 1 - \varepsilon_i^{-1} H\{x_i^T \frac{\eta}{\sqrt{n}} + \rho_i\} \Big| + \Big| 1 - \varepsilon_i H^{-1}\{x_i^T \frac{\eta}{\sqrt{n}} + \rho_i\} \Big| \Big\} \\ + \lambda_n \sum_{j=1}^{p} \omega_j |\beta_0 + \frac{\eta}{\sqrt{n}}|. \end{aligned}$$

Consider $V_n(\eta) = \Phi_n(\eta) - \Phi_n(0)$. By Theorem 2 we have

$$\begin{aligned} V_n(\eta) = S_n^T \eta + \eta^\top \gamma^2 [J + 2f(H(0))H(0)] C\eta + o_p(1) \\ + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \omega_j \sqrt{n} \Big\{ |\beta_0 + \frac{\eta}{\sqrt{n}}| - |\beta_0| \Big\}, \end{aligned} \tag{A.18}$$

where $S_n \to_d S \sim N\Big(0, \gamma^2 A \Xi\Big)$. Similar to the proof of Theorem 4, $V_n(\eta) \to_d V(\eta)$, where

$$V(\eta) = \begin{cases} S_\aleph^\top \eta_\aleph + \eta_\aleph^\top \gamma^2 [J + 2f(H(0))H(0)] C_{\aleph\aleph} \eta_\aleph, & \text{if } \eta_j = 0 \text{ for all } j \notin \aleph \\ \infty, & \text{otherwise.} \end{cases}$$

Obviously, $V(\eta)$ is convex. Applying the epi-convergence results of Geyer (1994) and Knight and Fu (2000), we have

$$\hat{\eta}_\aleph \to_d \frac{1}{2} \big\{ \gamma^2 [J + 2f(H(0))H(0)] \big\}^{-1} C_{\aleph\aleph}^{-1} S_\aleph \quad \text{and} \quad \hat{\eta}_{\aleph^c} \to_d 0 \tag{A.19}$$

To prove sparsity, we only need to show that $\hat{\beta}_j = 0, j \in \aleph^c$, with probability tending to 1. Then it suffices to show that $\forall j' \in \aleph^c, P(j' \in \hat{\aleph}^*) \to 0$.

Suppose $\check{\eta}$ is the minimizer of $\Phi_n(\eta)$ with $\hat{\eta}_\kappa \neq 0$, where $\kappa \subset \aleph^c$. Define $\bar{\eta}$ with $\bar{\eta}_\aleph = \check{\eta}_\aleph, \bar{\eta}_{\aleph^c} = 0$. Then,

$$\begin{aligned}
&\Phi_n(\check{\eta}) - \Phi_n(\bar{\eta}) = (\Phi_n(\check{\eta}) - \Phi_n(0)) - (\Phi_n(\bar{\eta}) - \Phi_n(0)) \\
&= \; S_n^T \check{\eta} + \check{\eta}^\top \gamma^2 [J + 2f(H(0))H(0)] C \check{\eta} - S_n^T \bar{\eta} - \bar{\eta}^\top \gamma^2 [J + 2f(H(0))H(0)] C \bar{\eta} \\
&\quad + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \omega_j \sqrt{n} \Big\{ |\beta_0 + \frac{\check{\eta}}{\sqrt{n}}| - |\beta_0| \Big\} - \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \omega_j \sqrt{n} \Big\{ |\beta_0 + \frac{\bar{\eta}}{\sqrt{n}}| - |\beta_0| \Big\} + o_p(1) \\
&= \; S_{n\aleph^c}^\top \check{\eta}_{\aleph^c} + \gamma^2 [J + 2f(H(0))H(0)] \{ 2\check{\eta}_\aleph^\top C_{\aleph\aleph^c} \check{\eta}_{\aleph^c} + \check{\eta}_{\aleph^c}^\top C_{\aleph^c\aleph^c} \check{\eta}_{\aleph^c} \} \\
&\quad + \frac{\lambda_n}{\sqrt{n}} \sum_{j' \in \kappa} \omega_{j'} |\check{\eta}_{j'}| + o_p(1) \\
&\to_p +\infty
\end{aligned}$$

Here the second equality is due to (A.18) and the third equality follows because of the definition of $\bar{\eta}, \check{\eta}$. Since $S_n \to_d S \sim N\Big(0, \gamma^2 A \Xi\Big)$ and a positive definite C, we have

$$S_{n\aleph^c}^\top \check{\eta}_{\aleph^c} + \gamma^2 [J + 2f(H(0))H(0)] \{ 2\check{\eta}_\aleph^\top C_{\aleph\aleph^c} \check{\eta}_{\aleph^c} + \check{\eta}_{\aleph^c}^\top C_{\aleph^c\aleph^c} \check{\eta}_{\aleph^c} \} = O_p(1).$$

Together with the fact that $\lambda_n \omega_{j'} / \sqrt{n} \to_p \infty$(Theorem 4), the last $\to_p$ result holds. This contradicts the fact that $\check{\eta}$ is the minimizer of $\Phi_n(\eta)$. Thus, we claim that for all $j' \in \aleph^c, P(j' \in \hat{\aleph}^*) \to 0$. The proof is complete.

References

Chen, K., Guo, S., Lin, Y. and Ying, Z. (2010). Least absolute relative error estimation. *J. Amer. Statist. Assoc.* **105**, 1104-1112.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477-489.

Efron, Bradley, Hastie, Trevor, Johnstone, Iain, Tibshirani and Robert (2004). Least angle regression. *Ann. Statist.* **32**, 407-499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *Proceedings of the International Congress of Mathematicians* **III**, 1348-1360.

Geyer, C. J. (1994) On the asymptotics of constrained M-estimation. *Ann. Statist.* **22**, 1993–2010.

Golub, G., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215-223.

Härdle, W., Liang, H. and Gao, J. T. (2000). Partially Linear Models. Springer Physica-Verlag.

Jin, Z., Ying, Z. and Wei, L. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88**, 381-390.

Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *Ann. Statist. Assoc.* **26**, 755-770.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist. Assoc.* **28**, 1356-1378.

Koenker, R. (2005). Quantile Regression. Cambridge university press.

Khoshgoftaar, T.M., Bhattacharyya, B. B. and Richardson, G. D. (1992). Predicting software errors, during development, using nonlinear regression models: A comparative study. *IEEE Trans. Reliability* **41**, 390-395.

Mack, Y. and Silverman, B. (1982). Weak and strong uniform consistency of kernel regression estimates. *Probab. Theory Related Fields* **61**, 405-415.

Narula, S. C. and Wellington, J. F. (1977). Prediction, linear regression and the minimum sum of relative errors. *Technometrics* **19**, 185-190.

Park, H. and Stefanski, L. A. (1998). Relative-Error Prediction. *Statist. Probab. Lett.* **40**, 227-236.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Wang, H., Li, R. and Tsai, C. L (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568.

Wang, H. and Leng, C. (2007). Unified lasso estimation via least squares approximation. *J. Amer. Statist. Assoc.* **102**, 1039-1048.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *J. Amer. Statist. Assoc.* **104**, 747-757.

Wang, Q. and Rao, J. N. K. (2002). Empirical Likelihood-based inference under imputation for missing response data. *Ann. Statist.* **30**, 896-924.

Xue, L. and Wang, Q. (2012). Empirical likelihood for single-index varying-coefficient models. *Bernoulli* **18**, 836-856.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing 100190, China.

E-mail: zhangqingzhao@amss.ac.cn

Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing 100190, China.

E-mail: qhwang@amss.ac.cn

(Received April 2012; accepted October 2012)