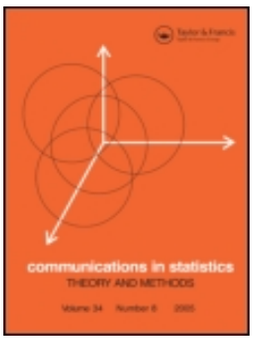

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: http://www.tandfonline.com/loi/lsta20

Kernel Density-Based Linear Regression Estimate

Weixin Yao & Zhibiao Zhao

To cite this article: Weixin Yao & Zhibiao Zhao (2013) Kernel Density-Based Linear Regression Estimate, Communications in Statistics - Theory and Methods, 42:24, 4499-4512, DOI: 10.1080/03610926.2011.650269

To link to this article: https://doi.org/10.1080/03610926.2011.650269

Communications in Statistics—Theory and Methods, 42: 4499–4512, 2013

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610926.2011.650269

Taylor & Francis
Taylor & Francis Group

Kernel Density-Based Linear Regression Estimate

WEIXIN YAO[1] AND ZHIBIAO ZHAO[2]

[1]Department of Statistics, Kansas State University, Manhattan, Kansas, USA
[2]Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania, USA

For linear regression models with non normally distributed errors, the least squares estimate (LSE) will lose some efficiency compared to the maximum likelihood estimate (MLE). In this article, we propose a kernel density-based regression estimate (KDRE) that is adaptive to the unknown error distribution. The key idea is to approximate the likelihood function by using a nonparametric kernel density estimate of the error density based on some initial parameter estimate. The proposed estimate is shown to be asymptotically as efficient as the oracle MLE which assumes the error density were known. In addition, we propose an EM type algorithm to maximize the estimated likelihood function and show that the KDRE can be considered as an iterated weighted least squares estimate, which provides us some insights on the adaptiveness of KDRE to the unknown error distribution. Our Monte Carlo simulation studies show that, while comparable to the traditional LSE for normal errors, the proposed estimation procedure can have substantial efficiency gain for non normal errors. Moreover, the efficiency gain can be achieved even for a small sample size.

Keywords EM algorithm; Kernel density estimate; Least squares estimate; Linear regression; Maximum likelihood estimate.

Mathematics Subject Classification 62F35; 62J05.

1. Introduction

Linear regression models are widely used to investigate the relationship between several variables. Suppose $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are sampled from the regression model

$$y = \mathbf{x}^T \boldsymbol{\beta} + \epsilon, \tag{1.1}$$

Received March 29, 2011; Accepted December 7, 2011
Address correspondence to Weixin Yao, Department of Statistics, Kansas State University, Manhattan, Kansas 66506, USA; E-mail: wxyao@ksu.edu

where $\mathbf{x}$ is a p-dimensional vector of covariates independent of the error ϵ with $E(\epsilon) = 0$. The well-known least squares estimate (LSE) of $\boldsymbol{\beta}$ is

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2. \tag{1.2}$$

For normally distributed errors, $\tilde{\boldsymbol{\beta}}$ is exactly the maximum likelihood estimate (MLE). However, $\tilde{\boldsymbol{\beta}}$ will lose some efficiency when the error is not normally distributed. Therefore, it is desirable to have an estimate that can be adaptive to the unknown error distribution.

The idea of adaptiveness is not new. Beran (1974) and Stone (1975) considered adaptive estimation for location model. Bickel (1982), Schick (1993), Yuan and De Gooijer (2007), and Yuan (2010) extended the adaptive idea to regression and some other models. Linton and Xiao (2007) further applied the adaptive idea to nonparametric regression estimate. Wang and Yao (2012) applied the adaptive idea to dimension reduction. Empirical likelihood techniques (Owen, 1988, 2001) have also been used for regression problems to adaptively construct the confidence intervals and testing statistics without any parametric assumption for the error density. However, empirical likelihood regression can't provide the efficient point regression estimates by adaptively using the unknown error density information.

In this article, we propose an adaptive kernel density based regression estimate (KDRE). The basic idea is to estimate the error density by kernel density estimate based on some initial parameter estimate and then estimate the regression parameters by maximizing the estimated likelihood function. Our proposed estimation procedure uses similar kernel error idea of Stone (1975) and Linton and Xiao (2007) to gain the adaptiveness based on some initial consistent estimate. However, Linton and Xiao (2007) mainly deals with nonparametric regression, the current article deals with parametric regression. We prove that the proposed estimate is asymptotically as efficient as the *oracle* MLE, which assumes the error density were known. Therefore, our proposed estimate can adapt to different error distributions. In addition, we propose a novel EM algorithm to maximize the estimated likelihood function and show that the KDRE can be viewed as an iterated weighted least squares estimate, which provides us some insights on why the KDRE can adapt to the unknown error distribution. To examine the finite sample performance, we conduct a Monte Carlo simulation study based on a wide range of error densities, including heavy-tail error, multiple-modal error, and skewed error density. Our simulation study confirms our theoretical finding. Our main claims are as follows.

1. The KDRE is comparable to the traditional LSE when the error is normal.
2. The KDRE is more efficient than the LSE when the error is not normal. The efficiency gain can be substantial even for a small sample size.

The remainder of this article is organized as follows. In Sec. 2, we introduce the new estimation procedure and prove its asymptotic oracle property. In addition, an EM type algorithm is introduced to maximize the estimated likelihood function. Numerical comparisons are conducted in Sec. 3. Summary and discussion are given in Sec. 4. Technical proofs are gathered in the Appendix.

2. Kernel Density Based Regression Estimate

2.1. *The New Estimation Method*

Let $f(t)$ be the marginal density of ϵ in (1.1). If $f(t)$ is known, instead of using the LSE, we can better estimate $\boldsymbol{\beta}$ in (1.1) by maximizing the log-likelihood

$$\sum_{i=1}^{n} \log f(y_i - \mathbf{x}_i^T \boldsymbol{\beta}). \tag{2.1}$$

In practice, however, $f(t)$ is often unknown and thus (2.1) is not directly applicable. To attenuate this, denote by $\tilde{\boldsymbol{\beta}}$ an initial estimate of $\boldsymbol{\beta}$, such as the LSE in (1.2). Based on the residuals $\tilde{\epsilon}_i = y_i - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}$, we can estimate $f(t)$ by the kernel density estimate, denoted by $\tilde{f}(t)$, as

$$\tilde{f}(t) = \frac{1}{n} \sum_{j=1}^{n} K_h(t - \tilde{\epsilon}_j), \tag{2.2}$$

where $K_h(t) = h^{-1}K(t/h)$, $K(\cdot)$ is a kernel density, and h is the tuning parameter. In this article, we use the Gaussian kernel for $K(\cdot)$. Replacing $f(\cdot)$ in (2.1) with $\tilde{f}(\cdot)$, we then propose the kernel density-based regression parameter estimate (KDRE) as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, Q(\boldsymbol{\beta}), \tag{2.3}$$

where $Q(\boldsymbol{\beta})$ is the estimated likelihood function

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log \tilde{f}(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) = \sum_{i=1}^{n} \log \left\{ \frac{1}{n} \sum_{j \neq i} K_h \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \tilde{\epsilon}_j \right) \right\}. \tag{2.4}$$

Here we use leave-one-out kernel density estimate for $f(\epsilon_i)$ to remove the estimation bias; see also Yuan and De Gooijer (2007) and Linton and Xiao (2007). The above estimation procedure can be easily extended to the nonlinear regression by replacing $\mathbf{x}_i^T \boldsymbol{\beta}$ in (2.4) with the assumed nonlinear function.

2.2. *Asymptotic Result*

Let $\boldsymbol{\beta}_0$ be the true value of $\boldsymbol{\beta}$. Then we have the following asymptotic oracle results for our proposed estimate $\hat{\boldsymbol{\beta}}$.

Theorem 2.1. *Assume that Assumptions C1–C5 in the Appendix hold. As* $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N\left(0, V^{-1}\right), \tag{2.5}$$

where $\epsilon = y - \mathbf{x}^T \boldsymbol{\beta}_0$,

$$V = \mathcal{I}_{\boldsymbol{\beta}_0} M, \quad M = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T = E(\mathbf{x}\mathbf{x}^T), \quad \mathcal{I}_{\boldsymbol{\beta}_0} = E\left\{ \frac{f'(\epsilon)^2}{f(\epsilon)^2} \right\}. \tag{2.6}$$

Remark 2.1. By the above theorem, the proposed estimate $\hat{\boldsymbol{\beta}}$ in (2.4) has root n convergence rate and its asymptotic distribution does not depend on the kernel $K(\cdot)$ or the bandwidth h, although the kernel density estimator with slower convergence rate is involved. In addition, $\hat{\boldsymbol{\beta}}$ has the same asymptotic variance as that of the infeasible oracle MLE, which assumes $f(\cdot)$ were known.

Remark 2.2. In (2.4), if we replace the objective function $\log \tilde{f}(\cdot)$ by another objective function, say $\rho(\cdot)$ with $\mathrm{E}\{\rho'(\epsilon)\} = 0$ (the LSE corresponds to $\rho(\epsilon) = \epsilon^2$), then the resulting estimate has limiting variance

$$v_\rho = \left[\frac{\mathrm{E}\{\rho'(\epsilon)^2\}}{\mathrm{E}\{\rho''(\epsilon)\}^2} M\right]^{-1}.$$

Based on the classical Cramér-Rao inequality that

$$\left[\frac{\mathrm{E}\{\rho'(\epsilon)^2\}}{\mathrm{E}\{\rho''(\epsilon)\}^2}\right]^{-1} \geq \mathcal{I}_{\boldsymbol{\beta}_0}^{-1},$$

we have $v_\rho \geq [\mathcal{I}_{\boldsymbol{\beta}_0} M]^{-1}$. Therefore, the objective functions we used in (2.4) is optimal in the sense that the proposed estimate is asymptotically efficient.

Remark 2.3. Our proposed method can also be applied to nonlinear regression model and similar oracle properties can also be established as in Theorem 2.1.

Remark 2.4. Yuan and De Gooijer (2007) proposed estimating $\boldsymbol{\beta}$ by maximizing

$$\sum_{i=1}^{n} \log\left[\frac{1}{n}\sum_{j\neq i} K_h\left\{r(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - r(y_j - \mathbf{x}_j^T\boldsymbol{\beta})\right\}\right], \tag{2.7}$$

where $r(\cdot)$ is some monotone nonlinear function, such as $r(z) = e^z/(1+e^z)$. Here, $r(\cdot)$ is used to avoid the cancelation of the intercept term in $\boldsymbol{\beta}$. Note that the asymptotic variance in (2.5) is the same as that in Yuan and De Gooijer (2007) with $r(t) = t$, which is efficient. One main advantage of their method is that it does not require an initial estimate. However, the asymptotic variance of their estimator depends on the choice of $r(\cdot)$ and generally does not reach the Cramér-Rao lower bound $[\mathcal{I}_{\boldsymbol{\beta}_0} M]^{-1}$ for a nonlinear function of $r(\cdot)$.

Note that when $r(t) = t$ in (2.7), although the intercept term, denoted by β_0, will be canceled, the slope parameter, denoted by $\boldsymbol{\beta}_1$, will remain estimable. Let $\bar{\boldsymbol{\beta}}_1$ be its estimate. In (2.5), let

$$V^{-1} = \begin{pmatrix} V^{11} & V^{12} \\ V^{21} & V_{22} \end{pmatrix},$$

where V^{11} is a scalar. Based on the result of Yuan and De Gooijer (2007), we know that $\bar{\boldsymbol{\beta}}_1$ is still an efficient estimate and has the asymptotic distribution

$$\sqrt{n}(\bar{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \xrightarrow{d} N\left(0, V^{22}\right).$$

Let $\mathbf{x} = (1, \mathbf{x}^{*T})^T$. Based on the slope estimate $\bar{\boldsymbol{\beta}}_1$, we can simply estimate β_0 by $\bar{\beta}_0 = \bar{Y} - \mathbf{x}_i^{*T}\bar{\boldsymbol{\beta}}_1$. Note that $\bar{\beta}_0$ can be considered as an LSE for model $y_i - \mathbf{x}_i^{*T}\bar{\boldsymbol{\beta}}_1 = \beta_0 + \epsilon_i$ after we fix $\boldsymbol{\beta}_1$ at $\bar{\boldsymbol{\beta}}_1$. Denote by KDRE1 the resulting estimate $(\bar{\beta}_0, \bar{\boldsymbol{\beta}}_1)$. Based on some standard calculations (the sketchy of the proof is given at the end of Appendix), we can get the asymptotic distribution for $\bar{\beta}_0$:

$$\sqrt{n}(\bar{\beta}_0 - \beta_0) \xrightarrow{d} N(0, \sigma^2),$$

where

$$\sigma^2 = \operatorname{var}\left[\epsilon_i - \frac{f'(\epsilon_i)}{f(\epsilon_i)}\left\{\mathrm{E}(\mathbf{x}^*)^T V_{21} + \mathrm{E}(\mathbf{x}^*)^T V_{22}\mathbf{x}_i^*\right\}\right].$$

Note that generally $\bar{\beta}_0$ does not reach the Cramér-Rao lower bound and the efficiency loss depends on the true error density $f(\epsilon)$. However, one nice feature of such estimate is that it doesn't require an initial estimate. In addition, it does not require to choose a nonlinear function $r(\cdot)$.

2.3. *Computations: An EM Algorithm*

Note that the objective function (2.4) has a mixture form. In this section, we propose an EM algorithm to maximize it. The proposed EM algorithm can be similarly used to find $\bar{\boldsymbol{\beta}}_1$ by maximizing (2.7) when $r(t) = t$. Let $\boldsymbol{\beta}^{(0)}$ be an initial parameter estimate, such as the LSE. We then update the parameter estimate according to the algorithm below.

Algorithm 2.1. At $(k+1)$th step, we calculate the following E and M steps:

E-Step. Calculate the classification probabilities,

$$p_{ij}^{(k+1)} = \frac{K_h(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^{(k)} - \tilde{\epsilon}_j)}{\sum_{l\neq i} K_h(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^{(k)} - \tilde{\epsilon}_l)} \propto K_h(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^{(k)} - \tilde{\epsilon}_j), \quad j \neq i, \tag{2.8}$$

M-Step. Update $\boldsymbol{\beta}^{(k+1)}$,

$$\begin{aligned}\boldsymbol{\beta}^{(k+1)} &= \operatorname*{argmax}_{\boldsymbol{\beta}} \sum_{i=1}^{n}\sum_{j\neq i}\left\{p_{ij}^{(k+1)} \log K_h(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \tilde{\epsilon}_j)\right\} \\ &= \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n}\sum_{j\neq i}\left\{p_{ij}^{(k+1)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \tilde{\epsilon}_j)^2\right\},\end{aligned} \tag{2.9}$$

which has explicit solutions, since $K_h(\cdot)$ is a Gaussian kernel density.

From the M step (2.9), the KDRE can be considered as a weighted least squares estimate, which minimizes the weighted squared difference between the new residual $y_i - \mathbf{x}_i^T\boldsymbol{\beta}$ and the initial residual $\tilde{\epsilon}_j$ for all $1 \leq i \neq j \leq n$. Based on the weights in (2.8), one knows that if jth observation is an isolated outlier (i.e., $\tilde{\epsilon}_j$ is large), then the weights $p_{ij}^{(k+1)}$ will be small for $i \neq j$ and thus the effect of $\tilde{\epsilon}_j$ on updating $\boldsymbol{\beta}^{(k+1)}$ will be also small.

By Theorem 2.2 below, the Algorithm 2.1 is truly an EM algorithm and has the monotone property for the objective function (2.4).

Theorem 2.2. *The objective function* (2.4) *is non decreasing after each iteration of Algorithm* 2.1, *i.e.,* $Q(\boldsymbol{\beta}^{(k+1)}) \geq Q(\boldsymbol{\beta}^{(k)})$, *until a fixed point is reached.*

3. Simulation Studies

In this section, we use a simulation study to compare the proposed KDRE and KDRE1 with the traditional LSE for linear regression models with different types of error densities. For the proposed estimate, we use the rule-of-thumb bandwidth $h = 1.06n^{-1/5}\hat{\sigma}$ for the kernel density estimate of $f(\epsilon)$, where $\hat{\sigma}$ is the sample standard deviation of the initial residual $\tilde{\epsilon}_i = y_i - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ is the LSE. Better estimates might be obtained if we use some more sophisticated bandwidth for kernel density estimate; see, for example, Sheather and Jones (1991) and Raykar and Duraiswami (2006). In addition, we can also use cross validation method to selection the bandwidth, which focuses on the performance of regression estimate directly instead of density estimate.

We generate independent and identically distributed data $\{(x_i, y_i), i = 1, \ldots, n\}$ from the model

$$Y = 1 + 3X + \epsilon\,,$$

where $X \sim U(0, 1)$, the uniform distribution on $[0, 1]$. For the error density, we consider the following six choices (all have standard deviation around 1).

Case 1. $\epsilon \sim N(0, 1)$, normal error.
Case 2. $\epsilon \sim U(-2, 2)$, the uniform distribution on $[-2, 2]$, short-tail error.
Case 3. $\epsilon \sim t_3/\sqrt{3}$, t-distribution with 3 degrees of freedom, heavy-tail error.
Case 4. $\epsilon \sim 0.95N(0, 0.7^2) + 0.05N(0, 3.5^2)$, contaminated normal error. The 5% data from $N(0, 3.5^2)$ are most likely to be outliers.
Case 5. $\epsilon \sim 0.5N(-1, 0.5^2) + 0.5N(1, 0.5^2)$, multi-modal error.
Case 6. $\epsilon \sim 0.3N(-1.4, 1) + 0.7N(0.6, 0.4^2)$, skewed error.

Here, we also used Case 6 to check how our method performed compared with LSE when the error is not symmetric. We estimate the regression parameters using KDRE, KDRE1, and the traditional LSE. Based on 1,000 replicates, Tables 1–2 report the mean squared errors (MSE) of the parameter estimates for intercept and slope, respectively, for sample size $n = 30, 100, 300$, and 600. The rightmost two columns contain the relative efficiency of KDRE and KDRE1 when compared to the LSE. For example, RE(KDRE)=MSE(LSE)/MSE(KDRE). From Cases 2–6 in Tables 1–2, we can see that KDRE and KDRE1 are much more efficient than the LSE when the error is not normal (for both symmetric and skewed error densities). Moreover, the efficiency gain can be substantial even for a small sample size. In addition, when the error is normal, KDRE is comparable to the LSE and works better than KDRE1 especially for small sample size. However, for Case 6–skewed error densities, KDRE1 works better than KDRE, although both of them have much better performance than LSE. In addition, for large sample size, the performances of KDRE and KDRE1 are almost the same, even for intercept estimate, although KDRE has some theoretical advantage over KDRE1. Note that KDRE1 is simpler without first estimating the error data.

Table 1
Simulation results for the intercept estimates

		Mean(MSE)				
Error Distribution	n	LSE	KDRE	KDRE1	RE(KDRE)	RE(KDRE1)
$N(0, 1)$	30	0.146	0.156	0.175	0.939	0.834
(Standard normal)	100	0.041	0.043	0.047	0.940	0.859
	300	0.014	0.015	0.016	0.960	0.893
	600	0.007	0.007	0.007	1.010	0.997
$U(-2, 2)$	30	0.183	0.144	0.154	1.266	1.190
(Short-tail distribution)	100	0.060	0.033	0.036	1.807	1.689
	300	0.017	0.008	0.009	2.180	1.901
	600	0.008	0.004	0.005	2.130	1.890
$t_3/\sqrt{3}$	30	0.159	0.104	0.109	1.529	1.465
(Heavy-tail distribution)	100	0.036	0.026	0.026	1.390	1.394
	300	0.112	0.009	0.009	1.315	1.329
	600	0.007	0.005	0.005	1.540	1.592
$0.95N(0, 0.7^2) + 0.05N(0, 3.5^2)$	30	0.150	0.102	0.106	1.470	1.417
(Contaminated normal)	100	0.040	0.028	0.028	1.411	1.411
	300	0.015	0.009	0.009	1.564	1.597
	600	0.008	0.005	0.005	1.513	1.438
$0.5N(-1, 0.5^2) + 0.5N(1, 0.5^2)$	30	0.180	0.122	0.111	1.477	1.598
(Multi-modal distribution)	100	0.051	0.027	0.027	1.864	1.889
	300	0.019	0.009	0.010	2.077	2.010
	600	0.009	0.005	0.005	1.918	1.825
$0.3N(-1.4, 1) + 0.7N(0.6, 0.4^2)$	30	0.182	0.115	0.088	1.593	2.083
(Skewed distribution)	100	0.053	0.028	0.022	2.005	2.412
	300	0.016	0.008	0.007	2.102	2.363
	600	0.009	0.005	0.004	1.907	2.270

4. Summary

In this article, we proposed an adaptive linear regression estimate by maximizing an estimated likelihood function, in which the error density is estimated by kernel density estimate. The proposed estimate can adapt to unknown error density and is asymptotically equivalent to the oracle MLE. Using the proposed EM algorithm, the computation is quick and stable. Our extensive simulation studies show that the proposed method outperforms the LSE in the presence of non-normal errors.

Although developed for linear regression models, the same idea can be easily extended to nonlinear regression cases. The asymptotic oracle property follows similarly. In addition, our proposed EM algorithm can be also used to estimate the adaptive nonparametric regression of Linton and Xiao (2007) and the semiparametric regression of Yuan and De Gooijer (2007). Future research directions include extensions to other regression models such as varying coefficient partially linear models and nonparametric additive models.

Table 2
Simulation results for the slope estimates

Error Distribution	n	Mean(MSE) LSE	KDRE	KDRE1	RE(KDRE)	RE(KDRE1)
$N(0, 1)$	30	0.418	0.456	0.543	0.918	0.771
(Standard normal)	100	0.119	0.128	0.144	0.933	0.826
	300	0.046	0.049	0.053	0.951	0.878
	600	0.020	0.020	0.020	1.020	0.999
$U(-2, 2)$	30	0.520	0.414	0.413	1.259	1.259
(Short-tail distribution)	100	0.169	0.088	0.081	2.001	2.093
	300	0.048	0.018	0.018	2.634	2.673
	600	0.026	0.009	0.009	3.010	3.070
$t_3/\sqrt{3}$	30	0.526	0.242	0.267	2.174	1.970
(Heavy-tail distribution)	100	0.114	0.065	0.067	1.744	1.691
	300	0.038	0.024	0.025	1.571	1.539
	600	0.018	0.009	0.009	2.020	2.024
$0.95N(0, 0.7^2) + 0.05N(0, 3.5^2)$	30	0.468	0.252	0.278	1.854	1.683
(Contaminated normal)	100	0.123	0.068	0.071	1.815	1.739
	300	0.043	0.020	0.021	2.118	2.097
	600	0.023	0.012	0.013	1.904	1.812
$0.5N(-1, 0.5^2) + 0.5N(1, 0.5^2)$	30	0.519	0.319	0.256	1.629	1.985
(Multi-modal distribution)	100	0.144	0.055	0.050	2.630	2.863
	300	0.058	0.019	0.018	3.058	3.157
	600	0.023	0.007	0.007	3.358	3.358
$0.3N(-1.4, 1) + 0.7N(0.6, 0.4^2)$	30	0.546	0.239	0.173	2.283	3.148
(Skewed distribution)	100	0.157	0.042	0.036	3.702	4.396
	300	0.046	0.012	0.011	4.007	4.153
	600	0.027	0.006	0.006	4.401	4.594

Appendix: Proofs

The following conditions are imposed to facilitate the proof.

C1. $\{\epsilon_i\}$ and $\{\mathbf{x}_i\}$ are i.i.d. and mutually independent with $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{E}(|\epsilon_i|^3) < \infty$. Additionally, the predictors $\mathbf{x}_i$ have bounded support and.

C2. The density $f(\cdot)$ of ϵ is symmetric about 0 and has bounded continuous derivatives up to order 4. Let $\ell(\epsilon) = \log f(\epsilon)$. Assume $\mathrm{E}\{\ell'(\epsilon)^2 + |\ell''(\epsilon)| + |\ell'''(\epsilon)|\} < \infty$.

C3. The kernel $K(\cdot)$ is symmetric, has bounded support, and are four times continuously differentiable.

C4. As $n \to \infty$, $nh^4 \to \infty$ and $nh^8 \to 0$.

C5. For the initial estimate $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$, assume $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O_p(n^{-1/2})$.

The condition $C1$ can guarantee that the least squares estimate is consistent and has root n convergence rate. The condition $C2$ is used to guarantee the adaptiveness of our proposed estimate. If $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \mathbf{x}_i = 0$, then the symmetric condition of $f(\epsilon)$ can be removed.

A.1 Proof of Theorem 2.1

We follow a similar strategy in Linton and Xiao (2007). Note that the maximizer $\hat{\boldsymbol{\beta}}$ in (2.3) is the solution of the score function

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{f}'(y_i-\mathbf{x}_i^T\boldsymbol{\beta})}{\tilde{f}(y_i-\mathbf{x}_i^T\boldsymbol{\beta})}\mathbf{x}_i, \tag{A.1}$$

where $\tilde{f}'(t)$ is the derivative of $\tilde{f}(t)$ in (2.2). For technical reason, we will consider another trimmed version of $\hat{\boldsymbol{\beta}}$ as the solution of

$$\widetilde{S}(\boldsymbol{\beta})=0, \quad \text{where } \widetilde{S}(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{f}'(y_i-\mathbf{x}_i^T\boldsymbol{\beta})}{\tilde{f}(y_i-\mathbf{x}_i^T\boldsymbol{\beta})}\mathbf{x}_i G_b(\tilde{f}(\boldsymbol{\epsilon}_i)). \tag{A.2}$$

Here,

$$G_b(x)=\begin{cases}0, & x<b;\\ \displaystyle\int_b^x g_b(t)dt, & b\le x\le 2b;\\ 1, & x>2b,\end{cases}$$

where $g_b(t)$ is any density function with support on $[b,2b]$ such that $G_b(t)$ is four times continuously differentiable on $[b,2b]$. In the following proof, we assume that $b=h^r$, where $0<r<1/2$. In practice, when b is small, the difference between the original estimate and the trimmed one is negligible.

By Taylor's expansion, there exists $\boldsymbol{\beta}^*$ such that $\|\boldsymbol{\beta}^*-\boldsymbol{\beta}_0\|\le\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|$ and

$$\widetilde{S}(\hat{\boldsymbol{\beta}})=\widetilde{S}(\boldsymbol{\beta}_0)+\frac{\partial\widetilde{S}(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+\frac{1}{2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\frac{\partial^2\widetilde{S}(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0).$$

The desired result then follows from Lemmas A.2–A.4 below.

Lemma A.1. *For $\tilde{f}$ in* (2.2)*, we have the uniform consistency results*

$$\sup_t|\tilde{f}(t)-f(t)|=O_p\left[h^2+\left\{\frac{\log(n)}{nh}\right\}^{1/2}\right], \tag{A.3}$$

$$\sup_t|\tilde{f}'(t)-f'(t)|=O_p\left[h^2+\left\{\frac{\log(n)}{nh^3}\right\}^{1/2}\right]. \tag{A.4}$$

Proof. Denote by $f^{(k)}$ the kth derivative of f with the convention $f^{(0)}=f$. Let

$$\check{f}^{(k)}(t)=\frac{1}{nh^{k+1}}\sum_{j=1}^{n}K^{(k)}\left(\frac{t-\boldsymbol{\epsilon}_j}{h}\right), \quad k=0,1,2,3,$$

be the traditional kernel density derivative estimator of $f^{(k)}(\cdot)$. By Silverman (1978),

$$\sup_t|\check{f}^{(k)}(t)-f^{(k)}(t)|=O_p\left\{h^2+\left\{\frac{\log(n)}{nh^{2k+1}}\right\}^{1/2}\right\}. \tag{A.5}$$

Since $\mathbf{x}_i$ has bounded support and $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O_p(n^{-1/2})$, $\tilde{\epsilon}_j - \epsilon_j = \mathbf{x}_j^T(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}) = O_p(n^{-1/2})$, uniformly over j. By Taylor's expansion, for some ϵ_j^* between ϵ_j and $\tilde{\epsilon}_j$,

$$\begin{aligned}\tilde{f}(t) - \check{f}(t) &= \frac{1}{nh^2}\sum_j K'\left(\frac{t-\epsilon_j}{h}\right)(\epsilon_j - \tilde{\epsilon}_j) + \frac{1}{2nh^3}\sum_j K''\left(\frac{t-\epsilon_j}{h}\right)(\epsilon_j - \tilde{\epsilon}_j)^2 \\ &\quad + \frac{1}{6nh^4}\sum_j K'''\left(\frac{t-\epsilon_j^*}{h}\right)(\epsilon_j - \tilde{\epsilon}_j)^3 \\ &= O_p(1/\sqrt{n}) + O_p(1/n)O_p\{1 + \sqrt{\log(n)/(nh^5)}\} + O_p(1/n^{3/2})O_p(1/h^4),\end{aligned}$$

uniformly, entailing (A.3) via Condition C4 and (A.5). Similarly, (A.4) follows. □

Lemma A.2. *Let V be defined as in* (2.6). *Then* $-\partial\widetilde{S}(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta} \xrightarrow{p} V$.

Proof. For notational convenience we write $f_i = f(\epsilon_i)$, $f_i' = f'(\epsilon_i)$, $f_i'' = f''(\epsilon_i)$, $\tilde{f}_i = \tilde{f}(\epsilon_i)$, $\tilde{f}_i' = \tilde{f}'(\epsilon_i)$, $\tilde{f}_i'' = \tilde{f}''(\epsilon_i)$. Note that

$$\begin{aligned}\frac{\partial\widetilde{S}(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}} &= -\frac{1}{n}\sum_{i=1}^n \frac{\tilde{f}_i'^2}{\tilde{f}_i^2} G_b(\tilde{f}_i)\mathbf{x}_i\mathbf{x}_i^T + \frac{1}{n}\sum_{i=1}^n \frac{\tilde{f}_i''}{\tilde{f}_i} G_b(\tilde{f}_i)\mathbf{x}_i\mathbf{x}_i^T + \frac{1}{n}\sum_{i=1}^n \frac{\tilde{f}_i'^2}{\tilde{f}_i} g_b(\tilde{f}_i)\mathbf{x}_i\mathbf{x}_i^T \\ &= A + B + C.\end{aligned}$$

It suffices to prove $A \xrightarrow{p} -V$, $B \xrightarrow{p} 0$, and $C \xrightarrow{p} 0$.

First, we consider A. Let $\Delta_i = \tilde{f}_i - f_i$, $\Delta_i' = \tilde{f}_i' - f_i'$, $\delta_n = h^2 + \sqrt{\log(n)/(nh)}$, and $\delta_n' = h^2 + \sqrt{\log(n)/(nh^3)}$. By Lemma A.1, $\max_i|\Delta_i| = O_p(\delta_n)$ and $\max_i|\Delta_i'| = O_p(\delta_n')$. By definition, $\sup_x G_b(x)/x^k \le 1/b^k$, $k \ge 0$. So, by the boundedness of f_i, f_i',

$$\begin{aligned}\frac{\tilde{f}_i'^2}{\tilde{f}_i^2} G_b(\tilde{f}_i) &= \left\{\frac{f_i'^2}{f_i^2} + \frac{\Delta_i'(\tilde{f}_i' + f_i')}{\tilde{f}_i^2} + \frac{\Delta_i f_i'^2(f_i + \tilde{f}_i)}{f_i^2\tilde{f}_i^2}\right\} G_b(\tilde{f}_i) \\ &= \frac{f_i'^2}{f_i^2} G_b(\tilde{f}_i) + \frac{O_p(\delta_n')}{b^2} + \frac{O_p(\delta_n)}{b^2}\frac{f_i'^2}{f_i^2}.\end{aligned}$$

By Condition C2, $f_i'^2/f_i^2$ is integrable, so we have

$$A = -\frac{1}{n}\sum_{i=1}^n \frac{f_i'^2}{f_i^2} G_b(\tilde{f}_i)\mathbf{x}_i\mathbf{x}_i^T + O_p\left(\frac{\delta_n'}{b^2}\right). \tag{A.6}$$

By Condition C2 and the Dominated Convergence Theorem, as $b \to 0$,

$$\mathrm{E}\left\{\frac{f_i'^2}{f_i^2}(1 - G_b(f_i))\right\} \le \mathrm{E}\left\{\frac{f_i'^2}{f_i^2} I(f(\epsilon_i) < 2b)\right\} = o(1).$$

Note that $\max_{1\le i\le n}|G_b(\tilde{f}_i) - G_b(f_i)| = o_p(1)$. Therefore, by decomposing $G_b(\tilde{f}_i)$ in (A.6) into $1 + \{G_b(f_i) - 1\} + \{G_b(\tilde{f}_i) - G_b(f_i)\}$, it is easily seen that $A \xrightarrow{p} -V$.

Next, we consider B. There exists ξ between 0 and $(\tilde{f} - f)/f$ such that

$$\tilde{f}^{-1}(\epsilon) = f^{-1}(\epsilon) - (1 + \xi)^{-2} f^{-2}(\epsilon)\left\{\tilde{f}(\epsilon) - f(\epsilon)\right\}. \tag{A.7}$$

Using the latter identity, we have

$$\begin{aligned} B &= \frac{1}{n}\sum_{i=1}^{n}\frac{f''(\epsilon_i)}{f(\epsilon_i)}G_b(\tilde{f}_i)\mathbf{x}_i\mathbf{x}_i^T + \frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{f}''(\epsilon_i)-f''(\epsilon_i)}{f(\epsilon_i)}G_b(\tilde{f}_i)\mathbf{x}_i\mathbf{x}_i^T \\ &\quad - \frac{1}{n}\sum_{i=1}^{n}\frac{\{\tilde{f}(\epsilon_i)-f(\epsilon_i)\}\tilde{f}''(\epsilon_i)}{f(\epsilon_i)^2(1+\xi_i)^2}G_b(\tilde{f}_i)\mathbf{x}_i\mathbf{x}_i^T \\ &= B_1 + B_2 + B_3. \end{aligned} \tag{A.8}$$

Similar to the proof of A in (A.6), we can get $B_1 = o_p(1)$. Note that

$$B_2 \simeq \frac{1}{n^2h^4}\sum_{i=1}^{n}\frac{1}{f(\epsilon_i)}\sum_{j=1}^{n}K'''\left(\frac{\epsilon_i-\epsilon_j}{h}\right)\left(\mathbf{x}_j^T\tilde{\boldsymbol{\beta}}-\mathbf{x}_j^T\boldsymbol{\beta}_0\right)G_b(f_i)\mathbf{x}_i\mathbf{x}_i^T.$$

Elementary calculations show that

$$\mathrm{E}\left\{K^{(k)}\left(\frac{t-\epsilon}{h}\right)\right\} = h^{k+1}\int K(z)f^{(k)}(t+zh)dz, \quad k=1,2,3.$$

Let

$$k_1(\epsilon_i,\epsilon_j) = \frac{1}{h^4}\frac{1}{f(\epsilon_i)}K'''\left(\frac{\epsilon_i-\epsilon_j}{h}\right)G_b(f_i).$$

It can be easily shown that, for distinct i, j, k, l,

$$\mathrm{E}\left\{k_1(\epsilon_i,\epsilon_j)\right\} = O(b^{-1}), \mathrm{E}\left\{k_1^2(\epsilon_i,\epsilon_j)\right\} = O(b^{-2}h^{-7}), \mathrm{E}\left\{k_1(\epsilon_i,\epsilon_j)k_1(\epsilon_i,\epsilon_l)\right\} = O(b^{-2})$$

Thus, calculating the first two moments based on the result of U-statistics, we have

$$B_2 = O_p(1/\sqrt{n})\times O_p(b^{-1})\times O_p(1/\sqrt{n^2b^2h^7}+1/\sqrt{nb^2}) = o_p(1).$$

That $B_3 = o_p(1)$ follows from

$$\max_{1\le i\le n}\left|\frac{\{\tilde{f}(\epsilon_i)-f(\epsilon_i)\}}{f(\epsilon_i)^2(1+\xi_i)^2}G_b(\tilde{f}_i)\right| = O_p\left[h^2+\left\{\frac{1}{nh}\log(1/h)\right\}^{1/2}\right]b^{-2} = o_p(1).$$

Finally, we consider C. Note that

$$\begin{aligned} C &= \frac{1}{n}\sum_{i=1}^{n}\frac{f'(\epsilon_i)^2}{f(\epsilon_i)}g_b(\tilde{f}_i)\mathbf{x}_i\mathbf{x}_i^T + \frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{f}'(\epsilon_i)^2-f'(\epsilon_i)^2}{f(\epsilon_i)}g_b(\tilde{f}_i)\mathbf{x}_i\mathbf{x}_i^T \\ &\quad - \frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{f}(\epsilon_i)-f(\epsilon_i)}{f(\epsilon_i)^2(1+\xi_i)^2}g_b(\tilde{f}_i)\mathbf{x}_i\mathbf{x}_i^T \\ &= C_1 + C_2 + C_3. \end{aligned}$$

Based on the uniform convergency results in Lemma A.1 and $g_b(\cdot) = O(b^{-1})$, we can easily get $C_2 = o_p(1)$ and $C_3 = o_p(1)$. By the Dominated Convergence Theorem,

$$\mathrm{E}\left\{\frac{f'(\epsilon_i)^2}{f(\epsilon_i)} g_b(f_i)\right\} \le \max_x\{g_b(x)x\}\mathrm{E}\left\{\frac{f'(\epsilon_i)^2}{f^2(\epsilon_i)} I(b \le f(\epsilon_i) \le 2b)\right\} \to 0,$$

which, along with the argument in the proof of A in (A.6), gives $C_1 = o_p(1)$.

Lemma A.3. *Let V be defined as in* (2.6). *Then* $\sqrt{n}\widetilde{S}(\boldsymbol{\beta}_0) \xrightarrow{d} N(0, V)$.

Proof. By (A.7),

$$\begin{aligned}\sqrt{n}\widetilde{S}(\boldsymbol{\beta}_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{f'(\epsilon_i)}{f(\epsilon_i)}\mathbf{x}_i G_b(\tilde{f}(\epsilon_i)) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\tilde{f}'(\epsilon_i) - f'(\epsilon_i)}{f(\epsilon_i)}\mathbf{x}_i G_b(\tilde{f}(\epsilon_i)) \\ &\quad - \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{f'(\epsilon_i)\tilde{f}'(\epsilon_i)\left\{\tilde{f}(\epsilon) - f(\epsilon)\right\}}{(1+\xi)^2 f(\epsilon_i)^3}\mathbf{x}_i G_b(\tilde{f}(\epsilon_i)) \\ &= J_1 + J_2 + J_3.\end{aligned}$$

By the technique in Lemma A.2 and Lemma S2 of Linton and Xiao (2007),

$$J_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{f'(\epsilon_i)}{f(\epsilon_i)}\mathbf{x}_i + o_p(1) \xrightarrow{d} N(0, V). \tag{A.9}$$

It remains to prove $J_2 \xrightarrow{p} 0$ and $J_3 \xrightarrow{p} 0$. Decompose J_2 as

$$\begin{aligned}J_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\tilde{f}'(\epsilon_i) - \check{f}'(\epsilon_i)}{f(\epsilon_i)}\mathbf{x}_i G_b(\tilde{f}(\epsilon_i)) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\check{f}'(\epsilon_i) - f'(\epsilon_i)}{f(\epsilon_i)}\mathbf{x}_i G_b(\tilde{f}(\epsilon_i)) \\ &= J_{21} + J_{22}.\end{aligned}$$

Note that

$$\begin{aligned}(J_{21})_a &\simeq \frac{1}{n\sqrt{n}h^3}\sum_{i=1}^n\sum_{j=1}^n \frac{1}{f(\epsilon_i)} K''\left(\frac{\epsilon_i - \epsilon_j}{h}\right)\mathbf{x}_j^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)X_{ia}G_b(f(\epsilon_i)) \\ &= O_p(1/\sqrt{n})\frac{1}{n\sqrt{n}h^3}\sum_{i=1}^n\sum_{j=1}^n \frac{1}{f(\epsilon_i)} K''\left(\frac{\epsilon_i - \epsilon_j}{h}\right)\mathbf{x}_j^T X_{ia}G_b(f(\epsilon_i)).\end{aligned}$$

Similar to the proof of B_2 in (A.8), by calculating the first two moments of $(J_{21})_a$ using the results of U-statistics, we have

$$E\{(J_{21})_a\} = O(h^2 b^{-1}) \quad \text{and} \quad \mathrm{var}\{(J_{21})_a\} = O(1/\sqrt{nb^4}).$$

Therefore, $(J_{21})_a = o_p(1)$. Note that

$$J_{22} \simeq \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\check{f}'(\epsilon_i) - f'(\epsilon_i)}{f(\epsilon_i)}\mathbf{x}_i G_b(f(\epsilon_i))$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(nh^2)^{-1} \sum_{j=1}^{n} \left\{ K'\left(\frac{\epsilon_i - \epsilon_j}{h}\right) - \mathrm{E}_i K'\left(\frac{\epsilon_i - \epsilon_j}{h}\right) \right\}}{f(\epsilon_i)} \mathbf{x}_i G_b(f(\epsilon_i)) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(nh^2)^{-1} \sum_{j=1}^{n} \mathrm{E}_i K'\left(\frac{\epsilon_i - \epsilon_j}{h}\right) - f'(\epsilon_i)}{f(\epsilon_i)} \mathbf{x}_i G_b(f(\epsilon_i)) \\
&= J_{22A} + J_{22B},
\end{aligned}
$$

where E_i is the conditional expectation given ϵ_i. Similar to the proof of B_2 in (A.8) and the proof techniques in the Lemma S2 of Linton and Xiao (2007), we can prove

$$\mathrm{E}(J_{22A}) = 0 \text{ and } \mathrm{var}\{J_{22A}\} = o(1).$$

Therefore, $J_{22A} = o_p(1)$. Similarly, we can prove $J_{22B} = o_p(1)$ and $J_3 = o_p(1)$.

Lemma A.4. $\partial^2 \widetilde{S}(\boldsymbol{\beta}^*)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T = o_p(\sqrt{n})$.

Proof. It follows from the same argument in Lemmas A.2–A.3 and we omit the details.

A.2 Proof of Theorem 2.2

Let $Z_i^{(k+1)}$ be a random variable such that

$$P\left\{ Z_i^{(k+1)} = K_h\left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(k+1)} - \tilde{\epsilon}_j\right) / K_h\left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(k)} - \tilde{\epsilon}_j\right) \right\} = p_{ij}^{(k+1)}, \quad j \neq i.$$

By Jensen's inequality, we have

$$
\begin{aligned}
Q(\boldsymbol{\beta}^{(k+1)}) - Q(\boldsymbol{\beta}^{(k)}) &= \sum_{i=1}^{n} \log \left\{ \frac{\sum_{j \neq i} K_h\left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(k+1)} - \tilde{\epsilon}_j\right)}{\sum_{j \neq i} K_h\left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(k)} - \tilde{\epsilon}_j\right)} \right\} \\
&= \sum_{i=1}^{n} \log \sum_{j \neq i} \left\{ p_{ij}^{(k+1)} \frac{K_h\left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(k+1)} - \tilde{\epsilon}_j\right)}{K_h\left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(k)} - \tilde{\epsilon}_j\right)} \right\} \\
&= \sum_{i=1}^{n} \log \mathrm{E}(Z_i^{(k+1)}) \geq \sum_{i=1}^{n} \mathrm{E}\left\{ \log(Z_i^{(k+1)}) \right\}.
\end{aligned}
$$

By the M-step of Algorithm 2.1, the desired result follows from

$$\sum_{i=1}^{n} \mathrm{E}\left\{ \log(Z_i^{(k+1)}) \right\} = \sum_{i=1}^{n} \sum_{j \neq i} p_{ij}^{(k+1)} \log \left\{ \frac{K_h\left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(k+1)} - \tilde{\epsilon}_j\right)}{K_h\left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(k)} - \tilde{\epsilon}_j\right)} \right\} \geq 0.$$

Sketch of the proof of asymptotic distribution of $\bar{\boldsymbol{\beta}}_0$. Let $\mathbf{x} = (1, \mathbf{x}^{*T})^T$. Note that

$$
\begin{aligned}
\bar{\beta}_0 &= \bar{y} - \bar{\mathbf{x}^*}^T \bar{\boldsymbol{\beta}}_1 = \beta_0 + \bar{\mathbf{x}^*}^T \boldsymbol{\beta}_1 + \bar{\epsilon} - \bar{\mathbf{x}^*}^T \bar{\boldsymbol{\beta}}_1 \\
&= \beta_0 + \bar{\mathbf{x}^*}^T (\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1) + \bar{\epsilon}
\end{aligned}
$$

Therefore,

$$\sqrt{n}(\bar{\beta}_0 - \beta_0) = \bar{\mathbf{x}^*}^T \sqrt{n}(\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1) + \sqrt{n}\bar{\epsilon}$$

In addition, we know

$$\sqrt{n}(\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{f'(\epsilon_i)}{f(\epsilon_i)} (V_{21} + V_{22}\mathbf{x}_i^*) + o_p(1).$$

Therefore,

$$\begin{aligned} \sqrt{n}(\bar{\beta}_0 - \beta_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \epsilon_i - \frac{f'(\epsilon_i)}{f(\epsilon_i)} (\bar{\mathbf{x}^*}^T V_{21} + \bar{\mathbf{x}^*}^T V_{22}\mathbf{x}_i^*) \right\} \\ &\xrightarrow{d} N(0, \sigma^2), \end{aligned}$$

where

$$\sigma^2 = \operatorname{var}\left[\epsilon_i - \frac{f'(\epsilon_i)}{f(\epsilon_i)} \left\{ \mathrm{E}(\mathbf{x}^*)^T V_{21} + \mathrm{E}(\mathbf{x}^*)^T V_{22}\mathbf{x}_i^* \right\} \right].$$

Acknowledgments

The authors are grateful to the editors and the referee for their insightful comments and suggestions, which greatly improved this article. In addition, the method of KDRE1 is based on the referee's suggestion.

References

Beran, R. (1978). Asymptotically efficient adaptive rank estimates in location models. *Ann. Statist.* 2:248–266.

Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* 10:647–671.

Linton, O., Xiao, Z. (2007). A nonparametric regression estimator that adapts to error distribution of unknown form. *Econometric Theory* 23:371–413.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75:237–249.

Owen, A. B. (2001). *Empirical Likelihood*. New York: Chapman & Hall/CRC.

Raykar, V. C., Duraiswami, R. (2006). Fast optimal bandwidth selection for kernel density estimation. Proc. Sixth *SIAM Int. Conf. Data Mining*, Bethesda, April, pp. 524–528.

Schick, A. (1993). On efficient estimation in regression models. *Ann. Statist.* 21:1486–1521.

Sheather, S. J., Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. B* 53:683–690.

Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of density and its derivatives. *Ann. Statist.* 6:177–184.

Stone, C. (1975). Adaptive maximum likelihood estimation of a location parameters. *Ann. Statist.* 3:267–284.

Wang, Q., Yao, W. (2012). An adaptive estimation of MAVE. *J. Multivariate Anal.* 104:88–100.

Yuan, A., De Gooijer, J. G. (2007). Semiparametric regression with kernel error model. *Scand. J. Statist.* 34:841–869.

Yuan, A. (2010). Semiparametric inference with kernel likelihood. *J. Nonparametric Statist.* 21:207–228.