Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: www.tandfonline.com/journals/tstf20

An algorithm for distributed parameter estimation in modal regression models

Xuejun Ma & Xiaochao Xia

To cite this article: Xuejun Ma & Xiaochao Xia (2025) An algorithm for distributed parameter estimation in modal regression models, Statistical Theory and Related Fields, 9:2, 101-123, DOI: 10.1080/24754269.2025.2483553

To link to this article: https://doi.org/10.1080/24754269.2025.2483553

Published online: 03 Apr 2025.

STATISTICAL THEORY AND RELATED FIELDS
2025, VOL. 9, NO. 2, 101–123
https://doi.org/10.1080/24754269.2025.2483553

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

An algorithm for distributed parameter estimation in modal regression models

Xuejun Ma[a] and Xiaochao Xia[b]

[a]School of Mathematical Sciences, Soochow University, Suzhou, People's Republic of China; [b]School of Mathematics and Statistics, Chongqing University, Chongqing, People's Republic of China

ABSTRACT

In this paper, we propose a new algorithm to handle massive data sets, which are modelled by modal regression models. Differing from the existing methods regarding distributed modal regression, the proposed method combines the divide-and-conquer idea and a linear approximation algorithm. It is computationally fast and statistically efficient to implement. Theoretical analysis for the resultant distributed estimator under some regularity conditions is presented. Simulation studies are conducted to assess the effectiveness and flexibility of the proposed method with a finite sample size. Finally, an empirical application to the chemical sensors data is analysed for further illustration.

ARTICLE HISTORY

Received 28 December 2023
Revised 13 November 2024
Accepted 19 March 2025

KEYWORDS

Modal regression; massive data; divide-and-conquer; linear approximation

1. Introduction

With the rapid development of information science and computing techniques, massive data sets become ubiquitous in many fields. Nowadays, on the one hand, it is convenient for researchers and data analysts to collect a data set with a huge sample size (e.g. tens of GBs) from various application scenarios, such as stock data in the Shanghai exchange market. On the other hand, a massive data set may be stored in many separate machines or local workers. One of the major challenges to analysing the massive data is the difficulty in statistical computation. When using a statistical model to describe the data, the cost of calculating parameters is huge and even infeasible to accomplish the entire computation on one personal computer (PC). To solve this issue, many useful approaches have been documented in the literature.

Generally, there are three kinds of approaches that could be used to efficiently solve the statistical estimation problems associated with massive data. The first kind is the subsampling method. Ma et al. (2015) proposed a subsampling method for linear regression models and developed a two-step subsampling algorithm for which the first step is to find a weight for each data point to be sampled and the second step is to form a weighted estimator by combining the sampled data points with the weights obtained in the first step. However, Ma et al. (2015)'s approach is not optimal. In order to develop an optimal subsampling approach, Wang et al. (2018) further proposed two strategies based on minimum mean squared error (mMSE) and minimum variance-covariance (mVC), respectively. More recently, Wang and Ma (2021) considered the quantile regression for massive data using the subsampling technique. The second kind is the divide-and-conquer (DAC) method. This

CONTACT Xiaochao Xia ✉ xxc@cqu.edu.cn

method first divides the whole large data set into many disjoint subsets such that on each subdata, model parameters can be computed in a parallel way, and then aggregates the estimators via a simple averaging. Chen and Xie (2014) referred the DAC to the split-and-conquer method, and applied such an idea to generalized linear models with sparse structures. Shi et al. (2018) studied a class of M-estimators using the DAC and established a cubic convergence rate for their estimators, which is faster than that of the common M-estimators. The last kind is the approximation algorithm. Chen et al. (2022) developed an approximate Newton algorithm using stochastic subgradient. What is more, many papers have further extended the above three kinds of approaches. For instance, Zhang and Wang (2021) developed a distributed optimal subdata selection approach using the subsampling and DAC techniques. Chen et al. (2019) designed an approximation algorithm to solve quantile regression under memory constraints via DAC. Jordan et al. (2019) developed a communication-efficient distributed statistical inference for M-estimators and penalized M-estimation using a surrogate likelihood.

Even if many excellent and useful approaches were proposed in the last decade, however, most mainly focus on the mean regression or quantile regression models. In this paper, differing from the previous works, we consider the modal regression models, which can be used to capture the most likely value, rather than the mean and quantile of the response variable. To address this issue, Wang and Li (2021) developed a robust communication-efficient distributed modal regression (CDMR) by using a surrogate loss to approximate the global modal regression objective function. This work is an extension of Jordan et al. (2019)'s approach. The parameter estimation in the CDMR is mainly done on the first machine via an iterated algorithm, which involves broadcasting the gradient to the other machines. A merit of this approach is that it does not require to calculate the Hessian matrix on all the machines. Whereas, when the data on the first machine is contaminated such as outliers or heavy-tailed, the performance of CDMR may not be robust. This in part motivates us to consider a robust method for massive data. It is worth noting that the modal regression was initially studied in Yao and Li (2014), which provided an expectation-maximization (EM) algorithm. However, this algorithm is not applicable to parallel computing. Thus, it can not be directly used to handle massive data.

Inspired by the work of Chen et al. (2019), we propose a new linear-type estimator in modal regression models. Our estimation is applicable to the distributed massive data and involves a bandwidth to approximate the density of the response. Compared to the CDMR, our proposed method does not require to compute the gradient, which makes the entire computation easier. The simulation results showed that our method is computationally fast to implement and more efficient than the CDMR.

The rest of this article is organized as follows. In Section 2, we introduce the proposed method in details and provide theoretical analysis. Simulation studies are presented in Section 3. A real data set is analysed in Section 4. All technical proofs of main results are postponed to the Appendix.

2. Methodology

2.1. A linear-type estimator

Let $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ be independent and identically distributed sample data drawn from the joint distribution of population $\{(\boldsymbol{X}, Y)\}$, where $\boldsymbol{X}$ is a p-dimensional vector of covariates, and Y

represents the response, which is a scalar. We consider the following model:

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where $\boldsymbol{\beta}$ is the parameter vector in $\mathbb{R}^p$. The conditional density, $f_{\varepsilon_i \mid \boldsymbol{X}_i}(u \mid \boldsymbol{X}_i)$, of ε_i given $\boldsymbol{X}_i$ has a strictly global maximum at $\varepsilon_i = 0$, which implies $\mathrm{Mode}(Y|\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_0$. Given the data, we can obtain an estimator of $\boldsymbol{\beta}$ via maximizing a kernel-based objective function, i.e.

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} K_h\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}\right), \tag{1}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$, $h > 0$ is a bandwidth, and the kernel $K(\cdot)$ can be specified as a probability density function. In the literature, a common choice for the kernel is Gaussian kernel, i.e. the density of standard normal distribution, $\phi(u) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-u^2/2}$, suggested by Yao and Li (2014). For the sake of convenience, we use the Gaussian kernel throughout this paper.

Optimizing (1) is equivalent to solving the following estimating equation

$$\frac{1}{n} \sum_{i=1}^{n} K_h'\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}\right) \boldsymbol{X}_i = 0, \tag{2}$$

with respect to $\boldsymbol{\beta}$, where $K_h'(u) = \frac{\mathrm{d}}{\mathrm{d}u} K_h(u) = \frac{1}{h^2} K'(u/h)$. We denote by $\widehat{\boldsymbol{\beta}}$ the solution of (2) with respect to $\boldsymbol{\beta}$, and then we have

$$\frac{1}{n} \sum_{i=1}^{n} K_h'\left(Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}\right) \boldsymbol{X}_i = 0. \tag{3}$$

It follows from $K(u) = \phi(u)$ that $K'(u) = -u\phi(u)$. Then, the Equation (3) becomes

$$\frac{1}{nh} \sum_{i=1}^{n} \boldsymbol{X}_i \left(Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}\right) \phi\left(\frac{Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}}{h}\right) = 0. \tag{4}$$

Accordingly, it can be expressed as

$$\widehat{\boldsymbol{\beta}} = \left[\frac{1}{nh} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top \phi\left(\frac{Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}}{h}\right)\right]^{-1} \left[\frac{1}{nh} \sum_{i=1}^{n} \boldsymbol{X}_i Y_i^\top \phi\left(\frac{Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}}{h}\right)\right]. \tag{5}$$

However, we can see that in the formula, the term on the right-hand side of the equation still involves the estimator $\widehat{\boldsymbol{\beta}}$. In general, there is no closed form for the expression of $\widehat{\boldsymbol{\beta}}$ in the Equation (3). If we have a good initial estimator $\widehat{\boldsymbol{\beta}}^{(0)}$ in advance (e.g. a consistent estimator of the true parameter vector $\boldsymbol{\beta}_0$), we can plug it into the term on the right-hand side of Equation (5), which leads to a linear-type estimator of modal regression (LEMR):

$$\widehat{\boldsymbol{\beta}} = \left[\frac{1}{nh} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top \phi\left(\frac{Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}^{(0)}}{h}\right)\right]^{-1} \left[\frac{1}{nh} \sum_{i=1}^{n} \boldsymbol{X}_i Y_i \phi\left(\frac{Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}^{(0)}}{h}\right)\right]. \tag{6}$$

If the data size is not very large, the above estimator can be easily computed using a single PC. However, when the data size is extremely large such as having tens of GBs, storing this massive data in one PC could be very difficult under the limited memory. Thus, it may be rather computational expensive to calculate the estimator using a single machine. One solution to the problem is to divide the entire dataset into many manageable subsets such that they are processed concurrently by many computers in practice. Another feasible solution is that if the user has only one machine in hand, we can divide the data into many batches such that one batch of dataset can be uploaded to the memory when computing some quantities of interest, and then repeat the procedure over all batches and aggregate these outputs from each batch to form a final estimator.

2.2. DAC-LEMR

In this subsection, we present our procedure in the framework of modal regression. Let $\mathcal{S} = \{1, 2, \ldots, n\}$ denote all sample observations. We assume that the observations are distributed across K local machines with equal size $m = n/K$. Thus, $\mathcal{S}$ is divided into K distinct subsets $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_K$ such that $\mathcal{H}_j \cap \mathcal{H}_k = \emptyset, j \neq k$ and $|\mathcal{H}_j| = m$. Let $\mathcal{X}_k = \{(\boldsymbol{X}_i, Y_i), i \in \mathcal{H}_k\}$ be the sample observations distributed to the kth machine. When the kth machine receives an initial estimator $\widehat{\boldsymbol{\beta}}^{(0)}$ from the master machine, we can compute the following two quantities on the kth machine

$$\boldsymbol{V}_k = \frac{1}{nh} \sum_{i \in \mathcal{H}_k} \boldsymbol{X}_i \boldsymbol{X}_i^\top \phi\left(\frac{Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}^{(0)}}{h}\right), \tag{7}$$

and

$$\boldsymbol{U}_k = \frac{1}{nh} \sum_{i \in \mathcal{H}_k} \boldsymbol{X}_i Y_i \phi\left(\frac{Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}^{(0)}}{h}\right). \tag{8}$$

Next, all the quantities $\{\boldsymbol{V}_k, \boldsymbol{U}_k\}_{k=1}^K$ are broadcasted from local machines to the master machine, and we can do the following calculation of $\widehat{\boldsymbol{\beta}}$ in (6) on the master machine

$$\widehat{\boldsymbol{\beta}} = \left[\sum_{k=1}^K \boldsymbol{V}_k\right]^{-1} \left[\sum_{k=1}^K \boldsymbol{U}_k\right].$$

We can repeat such a procedure several rounds and obtain an efficient estimator. The details of such a procedure are summarized in Algorithm 1, where the first machine is used as the master machine. In practice, we use some stopping rule, instead of q, such as $\|\widehat{\boldsymbol{\beta}}^{(g+1)} - \widehat{\boldsymbol{\beta}}^{(g)}\| \leq 10^{-6}$. Our limited experience indicates that this algorithm has a fast convergence. Furthermore, the initial estimator $\widehat{\boldsymbol{\beta}}^{(0)}$ can be chosen as the estimator of quantile regression.

Now, we provide a bandwidth selection method for the practical use of the modal regression estimator. Following the suggestion of Yao et al. (2012), we may denote

$$F_k(h) = \frac{1}{m} \sum_{i \in \mathcal{H}_k} K_h''(\hat{\varepsilon}_i),$$

$$G_k(h) = \frac{1}{m} \sum_{i \in \mathcal{H}_k} \{K_h'(\hat{\varepsilon}_i)\}^2.$$

Algorithm 1 DACLEMR

```
Input: Data batches $\mathcal{X}_k$ for $k = 1, 2, \ldots, K$, and the number of iterations $q, \hat{h}_{\text{opt}}$ A. Initialize:
    Calculate an initial estimator based on $\mathcal{X}_1$:
        $\widehat{\boldsymbol{\beta}}^{(0)} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{m} \sum_{i\in\mathcal{H}_1} K_h\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}\right)$
    B. Distributed/Parallel computation: Distribute the computation across $K$ local machines
for $k = 1, 2, \ldots, K$ do
    Swap data $\mathcal{X}_k$ into the local machine and compute $(\boldsymbol{U}_k, \boldsymbol{V}_k)$ according to (7) and (8) using the bandwidth $\hat{h}_{\text{opt}}$.
    Send $(\boldsymbol{U}_k, \boldsymbol{V}_k)$ to the master machine.
end forC. Aggregation on master machine
for  $g = 1, 2, \ldots, q$ do
    Aggregate $(\boldsymbol{U}_k, \boldsymbol{V}_k)$ from all local machines
    Compute the global estimator $\widehat{\boldsymbol{\beta}}^{(g)}$ using the aggregated sums
        $\widehat{\boldsymbol{\beta}}^{(g)} = \left(\sum_{j=1}^{K} \boldsymbol{V}_k\right)^{-1} \left(\sum_{j=1}^{K} \boldsymbol{U}_k\right)$
end for
Output: The final estimator $\widehat{\boldsymbol{\beta}}^{(q)}$.
```

Thus, we can obtain an optimal bandwidth, $\hat{h}_{\text{opt}}$, by minimizing

$$\hat{r}(h) = \frac{1}{K} \sum_{k=1}^{K} \frac{G_k(h) F_k^{-2}(h)}{\hat{\sigma}_k^2}, \tag{9}$$

where $\hat{\varepsilon}_{i\in\mathcal{H}_k} = Y_{i\in\mathcal{H}_k} - \boldsymbol{X}_{i\in\mathcal{H}_k}^\top \tilde{\boldsymbol{\beta}}_k$ is a residual obtained by fitting the data using any robust smoothing method, such as quantile regression. The estimator $\hat{\sigma}_k$ is formed based on the pilot estimates $\hat{\varepsilon}_{i\in\mathcal{H}_k}$s. Here, $\tilde{\boldsymbol{\beta}}_k$ denotes the estimator of quantile regression for the data batch $\mathcal{X}_k$. In our numerical study, this is achieved via fitting quantile regression to ensure the robustness of the initial parameter estimator. While, in real data analysis, we use the mean regression rather than quantile regression for the initial estimator. The main reason is that the optimization of the quantile regression is quite time-consuming in computation for large-scale data. The grid search method may be useful in finding h_{opt}. Yao et al. (2012) suggested that the candidate grid points for h can be taken as $h = 0.5\hat{\sigma} \times 1.02^j$, $j = 0, 1, \ldots, l$ for some fixed l, where $\hat{\sigma} = 1/K \sum_{k=1}^{K} \hat{\sigma}_k$. We follow the suggestion of Wang and Li (2021) to set $l = 60$ in our implementation.

2.3. Theoretical analysis

In this subsection, we present some regularity conditions and related theoretical results. It can be easily derived that we have the following Bahadur representation of $\widehat{\boldsymbol{\beta}}$:

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \left[\frac{1}{nh}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\top}\phi\left(\frac{Y_i-\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}}^{(0)}}{h}\right)\right]^{-1}\left[\frac{1}{nh}\sum_{i=1}^{n}\boldsymbol{X}_iY_i\phi\left(\frac{Y_i-\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}}^{(0)}}{h}\right)\right]\\ &= \boldsymbol{\beta}_0+\boldsymbol{D}_{nh}^{-1}\boldsymbol{A}_{nh},\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{D}_{nh} &= \frac{1}{nh}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\top}\phi\left(\frac{Y_i-\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}}^{(0)}}{h}\right),\\ \boldsymbol{A}_{nh} &= \frac{1}{nh}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i\phi\left(\frac{Y_i-\boldsymbol{X}_i^{\top}\widehat{\boldsymbol{\beta}}^{(0)}}{h}\right).\end{aligned}$$

To this end, denote

$$\boldsymbol{A}_{nh}^{*} = \frac{1}{nh}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i\phi\left(\frac{\varepsilon_i}{h}\right)$$

and

$$\boldsymbol{D}_{nh}^{*} = \frac{1}{nh}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\top}\phi\left(\frac{\varepsilon_i}{h}\right).$$

We mean that for a vector $\mathbf{a}$, $\|\mathbf{a}\|$ denotes the Euclidean norm, and for a square matrix $\mathbf{A}$, $\|\mathbf{A}\|$ denotes the operator norm, i.e., $\|\mathbf{A}\| = \sqrt{\lambda_{\max}(\mathbf{A}\mathbf{A}^{\top})}$, where $\lambda_{\max}(\cdot)$ stands for the maximum eigenvalue of a matrix. We present the following conditions.

(C1) Assume that $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n}$ are iid observations drawn from the population $(\boldsymbol{X}, Y)$.

(C2) Assume that in a neighbourhood of zero, the conditional density function $f_{\varepsilon\,|\,\boldsymbol{X}}(u\,|\,\boldsymbol{X})$ of ε given $\boldsymbol{X}$ has the first, second and third derivatives $f'_{\varepsilon\,|\,\boldsymbol{X}}(u\,|\,\boldsymbol{X})$,$f''_{\varepsilon\,|\,\boldsymbol{X}}(u\,|\,\boldsymbol{X})$,$f'''_{\varepsilon\,|\,\boldsymbol{X}}(u\,|\,\boldsymbol{X})$, which are bounded uniformly in $\boldsymbol{X}$. Moreover, we assume that $f'_{\varepsilon\,|\,\boldsymbol{X}}(0\,|\,\boldsymbol{X})$ and $h = o(1)$.

(C3) Assume that the minimum and maximum eigenvalues of matrix $\boldsymbol{D} \triangleq E\{f_{\varepsilon\,|\,\boldsymbol{X}}(0\,|\,\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^{\top}\}$ satisfy $0 < c_2 \le \lambda_{\min}(\boldsymbol{D}) < \lambda_{\max}(\boldsymbol{D}) \le c_2 < \infty$ for some positive constants c_1 and c_2.

(C4) Assume that $E\{\|\boldsymbol{X}\|^4\} < \infty$.

The above conditions are standard and mild in the literature of modal regression. Condition (C1) assumes the generating process of sample data. Conditions (C2) and (C3) are similar to the assumptions (A1) and (A2) of Yao and Li (2014), respectively. Condition (C4) requires the fourth moment of covariate vector to be finite. Similar conditions are also made in Zhang et al. (2023).

The following theorem establishes the convergence rates of $\boldsymbol{A}_{nh}$ and $\boldsymbol{D}_{nh}$ given in Section 2.2, respectively.

Proposition 2.1: *Suppose that the initial estimator* $\widehat{\boldsymbol{\beta}}^{(0)}$ *satisfies* $\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0\| = O_p(\xi_n)$*. If Conditions* (C1)–(C4) *hold and* $\xi_n = O(h)$*, then we have that* (i)

$$\|\boldsymbol{A}_{nh} - \boldsymbol{A}^*_{nh} - \boldsymbol{D}(\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}_0)\| = O_p\left(\frac{\xi_n}{h}\left(\frac{\log n}{n}\right)^{1/4} + \xi_n h^2 + \xi_n^3 + h^4\right);$$

(ii)

$$\|\boldsymbol{D}_{nh} - \boldsymbol{D}^*_{nh}\| = O_p\left(\frac{\xi_n}{h^2}\left(\frac{\log n}{n}\right)^{1/4} + \xi_n^2\right);$$

(iii)

$$\|\boldsymbol{D}^*_{nh} - \boldsymbol{D}\| = O_p\left(\frac{1}{\sqrt{nh}} + h^2\right);$$

and (iv)

$$\|\boldsymbol{D}_{nh} - \boldsymbol{D}\| = O_p\left(\frac{\xi_n}{h^2}\left(\frac{\log n}{n}\right)^{1/4} + \xi_n^2 + \frac{1}{\sqrt{nh}} + h^2\right).$$

Remark 2.1: From this proposition, it is interesting to see that the convergence rate on the right-hand side of (ii) is quadratic in ξ_n, which is the rate of the initial estimator of $\boldsymbol{\beta}$. Ideally, if the initial estimator is set as the true value, $\boldsymbol{\beta}_0$, i.e., ξ_n equals zero, then we must have $\boldsymbol{D}_{nh} = \boldsymbol{D}^*_{nh}$ explicitly. This is certified in Proposition 2.1(ii) since the order is zero in this case. Further, the result (iv) follows directly from the triangle inequality as well as the results (ii) and (iii). It should be clarified that our theory analysis is in fact also related to classical nonparametric regression (Härdle, 1990; Ullah & Pagan, 1999) due to the use of kernel smoothing. While our estimator relies on the initial value, that is, $\boldsymbol{D}_{nh}$ and $\boldsymbol{A}_{nh}$ rely on the initial estimator of $\boldsymbol{\beta}$, which makes the asymptotic properties for $\widehat{\boldsymbol{\beta}}$ more difficult to investigate. To our best knowledge, the above proposition has never been stated in the existing literature.

According to Proposition 2.1, we can obtain the following property.

Proposition 2.2: *Under the conditions of Proposition 2.1, we have*

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \boldsymbol{D}^{-1}\boldsymbol{A}^*_{nh} + O_p\left(\frac{\xi_n^2}{h^2}\left(\frac{\log n}{n}\right)^{1/4} + \xi_n + h\left(h^2 \vee \frac{1}{\sqrt{nh}}\right)^2\right).$$

Remark 2.2: This property provides a Badur convergence rate of $\widehat{\boldsymbol{\beta}}$. The rigorous proof of this result is given in the Appendix and depends on the theory of empirical processes. From the proof in the Appendix, we know $\boldsymbol{A}^*_{nh} = O_p(\sqrt{hn^{-1}} + h^4)$. Thus, the convergence rate of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ is $O_p\left(\frac{\xi_n^2}{h^2}\left(\frac{\log n}{n}\right)^{1/4} + \xi_n + h^4 + n^{-1}\right)$. If $\xi_n = h$, the bandwidth can be taken as $h = n^{-1/4}$. In our implementation, we use a grid search method as described in Section 2.2. Our code for simulation and real data analysis can be available upon request.

Table 1. The simulation results of AEE ($\times 10^{-3}$) for Example 3.1 with $\rho = 0$.

			Case 1		Case 2	
r	K	Method	$p = 2$	$p = 10$	$p = 2$	$p = 10$
0	20	LEMR	3.831	3.809	4.48	4.251
		CDMR	3.951	4.052	4.569	4.654
	50	LEMR	3.670	3.731	4.441	4.471
		CDMR	4.145	5.658	4.986	6.529
	100	LEMR	3.911	3.824	4.334	4.448
		CDMR	5.830	9.275	6.477	11.009
0.1	20	LEMR	3.893	3.793	4.594	4.538
		CDMR	8.156	6.810	8.285	7.387
	50	LEMR	3.989	3.890	4.426	4.441
		CDMR	11.305	11.123	11.621	12.371
	100	LEMR	4.078	3.760	4.455	4.417
		CDMR	13.637	17.877	14.745	20.16
0.2	20	LEMR	4.427	4.024	5.325	4.785
		CDMR	32.139	18.112	33.304	19.665
	50	LEMR	3.847	3.865	4.595	4.507
		CDMR	40.415	28.997	42.625	32.148
	100	LEMR	3.768	3.765	4.375	4.425
		CDMR	46.223	42.268	51.808	48.408
0.3	20	LEMR	5.203	4.124	6.544	5.024
		CDMR	91.262	44.551	96.012	48.666
	50	LEMR	4.166	3.957	4.834	4.606
		CDMR	110.733	68.481	119.634	77.067
	100	LEMR	3.947	3.861	4.429	4.499
		CDMR	126.781	104.127	135.898	112.426

3. Simulations

In this section, we compare our proposed method (LEMR) with the CDMR of Wang and Li (2021) in the finite sample performance. We set the number of machines $K \in \{20, 50, 100\}$. $n = 5 \times 10^4$ in Examples 3.1 and 3.2. Each experiment is repeated 200 times.

Example 3.1: Similar to Wang and Li (2021), we consider the linear model $Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i$, where we set $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top, \beta_j \sim U[1,2], \boldsymbol{X}_i = (1, \boldsymbol{Z}_i^\top)^\top$, and $\boldsymbol{Z}_i$ follows a multivariate normal distribution with $\mathrm{Cov}(Z_{ij}, Z_{ik}) = \rho^{|j-k|}(\rho = \{0.0, 0.5\})$. ε_is are simulated from two distributions: (Case 1) $N(0,1)$ and (Case 2) $t(3)$. Furthermore, in order to evaluate the robustness of the methods against outliers, we contaminate the sample observations of Y on the first machine with a proportion r% of observations from the distribution $\chi^2(5)$, i.e., chi-square distribution with 5 degrees of freedom. We here consider four settings for $r = \{0, 10, 20, 30\}$, where $f = 0$ corresponds to the situation that the data has no contamination. To assess the performance of the proposed method, we compute the average estimator error (AEE), which is defined as $1/(200(p+1)) \sum_{m=1}^{200} \sum_{j=0}^{p} |\beta_j - \widehat{\beta}_j^m|$ for 200 replications. Note that $\widehat{\beta}_j^m$ denotes the estimator of LEMR or CDMR at the mth experiment.

Tables 1 and 2 report the simulation results of Example 3.1. From this table, we can draw the following conclusions.

(1) When there is no contamination in the data set, almost LEMR and CDMR perform very comparably. Both methods are not sensitive to the values of K and p.
(2) When the data set is contaminated with outliers, our proposed LEMR performs much better than CDMR especially in the case of $r = 0.3$, in which setting, the value of AEE of

Table 2. The simulation results of AEE ($\times 10^{-3}$) for Example 3.1 with $\rho = 0.5$.

			Case 1		Case 2	
r	K	Method	$p = 2$	$p = 10$	$p = 2$	$p = 10$
0	20	LEMR	4.218	4.690	4.917	5.257
		CDMR	4.354	4.995	4.982	5.755
	50	LEMR	4.103	4.621	4.912	5.538
		CDMR	4.682	6.980	5.559	8.087
	100	LEMR	4.315	4.720	4.801	5.551
		CDMR	6.646	11.666	7.154	13.938
0.1	20	LEMR	4.309	4.719	4.972	5.656
		CDMR	8.546	8.282	9.053	8.951
	50	LEMR	4.413	4.791	4.909	5.516
		CDMR	12.219	13.539	12.640	15.244
	100	LEMR	4.482	4.556	4.787	5.488
		CDMR	15.221	21.967	16.252	24.891
0.2	20	LEMR	4.705	4.960	5.823	5.841
		CDMR	33.526	20.872	34.878	23.007
	50	LEMR	4.216	4.778	4.972	5.572
		CDMR	42.392	34.610	44.958	38.241
	100	LEMR	4.135	4.698	4.894	5.535
		CDMR	48.791	50.695	54.989	58.104
0.3	20	LEMR	5.548	5.044	6.989	6.085
		CDMR	93.795	51.071	98.915	55.523
	50	LEMR	4.592	4.775	5.342	5.723
		CDMR	114.976	80.240	123.506	89.726
	100	LEMR	4.322	4.814	4.889	5.514
		CDMR	134.045	122.913	142.939	133.418

CDMR is much larger than that of LEMR. For instance, when looking at the Case 1 with $p = 2$, $K = 50$ and $\rho = 0.0$, the AEE of CDMR is 110.733. However, the LEMR results in the AEE of value 4.166.

(3) Whether the data set is contaminated or not, our LEMR always has a stable performance, showing the robustness of our method. Because the CDMR mainly involves the calculation of the first and second gradient over the data on the first machine. When the data on this machine is contaminated, the estimated gradient would be far from normal. This leads to a poor performance of CDMR.

Furthermore, we run our R code with R version 4.2.1 on the desktop computer with AMD 2990WX 32-core 3.00 GHz processor and 32.0 GB RAM. The computational results are given in Table 3, from which, one can see that the computing time for various methods is very close.

Example 3.2: In this example, we still generate the data from the model in Example 3.1, except that we set $\boldsymbol{\beta} = (1, 0.5, 1, 1.5, 2)^\top$ and let $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})^\top$ follow a multivariate normal distribution with zero mean and covariance matrix $\mathrm{Cov}(Z_{ij}, Z_{ik}) = \rho^{|j-k|}$. Here we consider two cases for $\rho = \{0.2, 0.6\}$. To evaluate the behaviour of the methods, we also add one criteria AEE_j, the average absolute estimator error for each individual covariate including an intercept, which is defined as $200^{-1}\sum_{m=1}^{200}|\beta_j - \widehat{\beta}_j^m|$ over 200 replications. The simulation results are shown in Tables 4 –7.

According to Tables 4– 7, we can observe that LEMR clearly outperforms CDMR.

Example 3.3: Furthermore, we conducted an additional simulation study to examine the asymptotic properties of the estimator. We only consider Case 1 of Example 3.2 with $\rho = 0$

Table 3. The simulation results of computing time (seconds) for Example 3.1 with $\rho = 0$.

			Case 1		Case 2	
r	K	Method	$p = 2$	$p = 10$	$p = 2$	$p = 10$
0	20	LEMR	15.92	36.00	16.82	38.05
		CDMR	15.75	36.05	16.88	37.73
	50	LEMR	18.69	32.55	21.74	32.43
		CDMR	18.68	32.44	21.52	32.31
	100	LEMR	29.79	21.42	31.64	20.90
		CDMR	29.11	21.17	31.13	20.80
0.1	20	LEMR	16.22	36.33	16.64	33.21
		CDMR	15.98	35.77	16.50	33.47
	50	LEMR	19.00	32.75	19.12	33.37
		CDMR	18.61	32.27	19.02	32.83
	100	LEMR	28.87	21.50	28.65	21.86
		CDMR	28.47	21.30	28.13	21.25

Table 4. The simulation results of AEE_j $(\times 10^{-3})$ and AEE $(\times 10^{-3})$ for Case 1 in Example 3.2 with $\rho = 0.2$.

r	K	Method	AEE_0	AEE_1	AEE_2	AEE_3	AEE_4	AEE
0	20	LEMR	3.856	4.315	4.264	4.222	3.902	4.112
		CDMR	4.001	4.304	4.315	4.338	3.918	4.175
	50	LEMR	3.946	3.504	4.064	4.149	3.906	3.914
		CDMR	4.332	4.516	4.750	5.294	5.305	4.839
	100	LEMR	4.233	3.920	4.153	3.726	3.903	3.987
		CDMR	6.399	6.764	7.338	7.367	7.112	6.996
0.1	20	LEMR	4.096	4.545	4.117	3.762	3.994	4.103
		CDMR	13.966	6.263	5.494	6.097	6.035	7.571
	50	LEMR	3.560	3.888	4.180	4.107	3.755	3.898
		CDMR	17.462	9.172	9.282	9.586	8.675	10.836
	100	LEMR	3.781	3.723	3.747	3.663	3.980	3.779
		CDMR	21.577	14.943	13.249	14.398	14.941	15.822
0.2	20	LEMR	5.945	3.695	4.156	4.207	4.073	4.415
		CDMR	72.702	11.894	12.851	13.752	12.396	24.719
	50	LEMR	4.344	3.783	4.198	3.784	3.735	3.969
		CDMR	81.231	19.988	18.679	20.203	18.236	31.667
	100	LEMR	4.335	3.806	3.694	3.783	3.724	3.868
		CDMR	89.688	30.238	29.609	29.994	31.874	42.281
0.3	20	LEMR	8.191	3.855	3.812	3.835	4.050	4.749
		CDMR	224.474	25.802	28.156	28.958	24.816	66.441
	50	LEMR	4.483	3.601	4.034	3.653	4.070	3.968
		CDMR	249.219	40.272	45.550	48.461	44.488	85.598
	100	LEMR	4.038	3.813	4.002	3.908	3.941	3.940
		CDMR	268.772	63.647	68.260	70.605	67.374	107.731

and $r = 0$ with different values of sample size n and two different values of K to validate the asymptotic properties. We proceed with this process over 200 replications. We let the entire sample size n vary from $n \in \{4 \times 10^4, 5 \times 10^4, 6 \times 10^4, 7 \times 10^4, 8 \times 10^4, 9 \times 10^4\}$. The results are given in Table 8, from which we can see that as n increases, the AEE gradually decreases to zero, indicating that the proposed estimator has the consistency.

4. Real data analysis

In this section, we apply our proposed method to chemical sensors data. The data were collected at the ChemoSignals Laboratory in the BioCircuits Institute, University of California San Diego. It consists of the readings of 16 chemical sensors exposed to the mixture of Ethylene and CO at varying concentrations in air. Each measurement was constructed by the continuous acquisition of the sixteen-sensor array signals for a duration of about 12 hours

Table 5. The simulation results of AEE_j ($\times 10^{-3}$) and AEE ($\times 10^{-3}$) for Case 2 in Example 3.2 with $\rho = 0.2$.

r	K	Method	AEE_0	AEE_1	AEE_2	AEE_3	AEE_4	AEE
0	LEMR	20	4.556	4.519	4.255	4.911	4.404	4.529
	CDMR	20	4.625	4.773	4.314	5.153	4.703	4.714
	LEMR	50	4.333	4.697	4.560	4.769	4.242	4.520
	CDMR	50	5.378	5.383	5.607	5.647	5.458	5.495
	LEMR	100	4.424	4.529	4.843	4.318	4.737	4.570
	CDMR	100	6.804	8.152	8.218	7.884	7.891	7.790
0.1	LEMR	20	5.374	4.548	4.823	4.685	4.079	4.702
	CDMR	20	13.122	6.219	7.038	6.775	6.803	7.991
	LEMR	50	4.517	4.451	4.451	4.940	4.216	4.515
	CDMR	50	16.790	8.508	9.820	9.540	9.324	10.796
	LEMR	100	4.253	4.121	4.550	4.907	4.678	4.502
	CDMR	100	18.514	14.451	14.252	13.223	14.527	14.993
0.2	LEMR	20	7.526	4.129	4.837	4.250	4.375	5.023
	CDMR	20	74.185	13.732	14.091	14.324	13.700	26.006
	LEMR	50	5.110	4.403	4.533	5.202	4.518	4.753
	CDMR	50	84.757	19.947	22.380	22.784	22.810	34.536
	LEMR	100	4.030	5.131	4.751	4.402	4.357	4.534
	CDMR	100	96.863	34.707	36.354	37.620	33.037	47.716
0.3	LEMR	20	10.567	4.585	4.322	4.653	4.303	5.686
	CDMR	20	229.527	27.921	30.178	34.965	26.868	69.892
	LEMR	50	5.424	4.597	4.120	4.784	4.748	4.735
	CDMR	50	260.725	45.340	50.401	52.931	46.465	91.172
	LEMR	100	4.136	4.849	4.773	4.664	4.174	4.519
	CDMR	100	288.818	75.202	81.794	81.649	79.032	121.299

Table 6. The simulation results of AEE_j ($\times 10^{-3}$) and AEE ($\times 10^{-3}$) for Case 1 in Example 3.2 with $\rho = 0.6$.

r	K	Method	AEE_0	AEE_1	AEE_2	AEE_3	AEE_4	AEE
0	20	LEMR	3.856	5.240	5.858	5.842	4.774	5.114
		CDMR	4.001	5.246	5.957	6.063	4.799	5.213
	50	LEMR	3.949	4.240	5.895	5.919	4.786	4.958
		CDMR	4.332	5.316	6.696	7.816	6.497	6.131
	100	LEMR	4.235	4.958	5.711	5.338	4.777	5.004
		CDMR	6.399	8.611	10.090	10.096	8.710	8.781
0.1	20	LEMR	4.096	5.180	5.650	5.461	4.891	5.056
		CDMR	13.966	7.333	8.024	8.726	7.391	9.088
	50	LEMR	3.559	4.794	5.763	5.503	4.599	4.844
		CDMR	17.462	10.945	13.416	13.029	10.625	13.096
	100	LEMR	3.783	4.620	5.115	5.246	4.874	4.728
		CDMR	21.577	17.088	19.436	20.259	18.298	19.332
0.2	20	LEMR	5.941	4.645	5.918	5.868	4.989	5.472
		CDMR	72.702	14.702	18.493	19.344	15.182	28.085
	50	LEMR	4.348	4.645	5.889	5.286	4.574	4.948
		CDMR	81.231	25.379	26.673	27.603	22.334	36.644
	100	LEMR	4.335	4.316	5.236	5.214	4.560	4.732
		CDMR	89.688	36.629	41.342	42.399	39.038	49.819
0.3	20	LEMR	8.191	4.838	5.525	5.699	4.960	5.843
		CDMR	224.474	31.360	38.830	40.247	30.394	73.061
	50	LEMR	4.483	4.561	5.545	5.218	4.985	4.958
		CDMR	249.219	50.600	64.831	65.687	54.487	96.965
	100	LEMR	4.038	4.798	5.525	5.577	4.827	4.953
		CDMR	268.772	77.234	98.600	96.605	82.516	124.745

without interruption. The concentration transitions were set at random times and concentration levels. Further information about the data set can be found in Fonollosa et al. (2015). We follow the suggestion of Wang et al. (2019) to consider the last sensor as the response and other sensors as covariates. We take a log-transformation of the sensors readings. Readings from the second sensor are excluded in the analysis because there are approximately 20% of

Table 7. The simulation results of AEE$_j$ ($\times 10^{-3}$) and AEE ($\times 10^{-3}$) for Case 2 in Example 3.2 with $\rho = 0.6$.

r	K	Method	AEE$_0$	AEE$_1$	AEE$_2$	AEE$_3$	AEE4	AEE
0	20	LEMR	4.558	5.278	6.139	6.662	5.392	5.606
		CDMR	4.625	5.664	6.219	6.946	5.760	5.843
	50	LEMR	4.334	5.424	6.501	6.738	5.196	5.639
		CDMR	5.378	6.220	7.998	8.016	6.685	6.859
	100	LEMR	4.424	5.544	6.926	6.280	5.800	5.795
		CDMR	6.804	9.673	12.215	10.877	9.665	9.847
0.1	20	LEMR	5.374	5.849	6.504	6.274	4.995	5.799
		CDMR	13.122	7.949	9.687	9.162	8.332	9.650
	50	LEMR	4.517	5.493	6.410	6.587	5.163	5.634
		CDMR	16.790	10.930	13.881	13.182	11.419	13.241
	100	LEMR	4.250	5.122	6.483	6.793	5.736	5.677
		CDMR	18.514	17.252	18.627	19.139	17.792	18.265
0.2	20	LEMR	7.521	5.072	6.765	5.836	5.359	6.110
		CDMR	74.185	16.688	19.932	20.207	16.779	29.558
	50	LEMR	5.109	5.493	6.424	7.056	5.533	5.923
		CDMR	84.757	25.056	32.911	32.400	27.936	40.612
	100	LEMR	4.030	6.251	6.427	6.228	5.335	5.654
		CDMR	96.863	43.459	50.859	49.601	40.462	56.249
0.3	20	LEMR	10.567	5.459	6.173	6.293	5.270	6.753
		CDMR	229.527	36.466	43.539	46.606	32.906	77.809
	50	LEMR	5.424	5.476	6.041	6.580	5.815	5.867
		CDMR	260.725	56.020	72.670	71.130	56.908	103.490
	100	LEMR	4.136	5.867	6.731	6.483	5.112	5.666
		CDMR	288.818	100.030	117.116	116.307	96.794	143.813

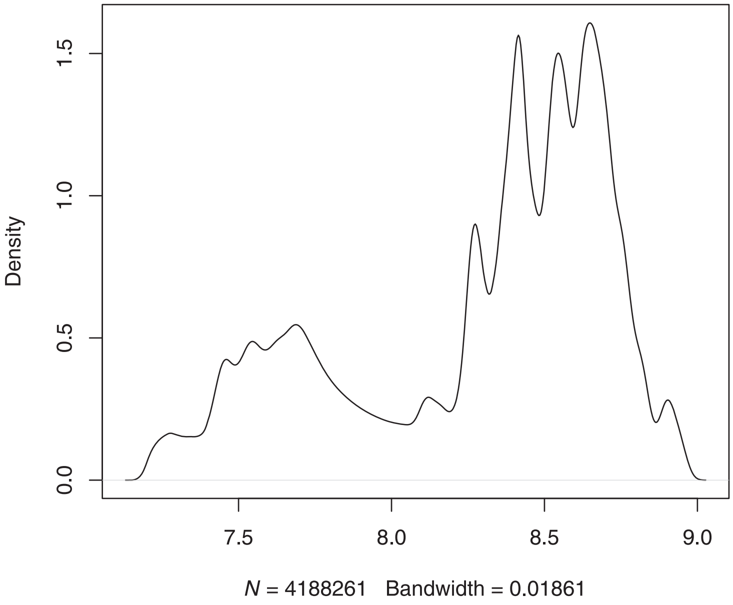

Figure 1. The kernel density estimation of the response variable.

the negative values for unknown reasons. Thus, we exclude the first 20000 data points that correspond to less than 4 minutes of system run-in time. Thus, the data to be analysed have the size of $n = 4188261$ and the dimensionality of $p = 14$.

We assume the number of machines $K = \{20, 50, 100, 150, 200\}$, and use the mean square error (MSE) defined by $n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ for a comparison with mode regression with full

Table 8. The simulation results of AEE ($\times 10^{-3}$) for Example 3.3 with Case 1, $\rho = 0$ and $r = 0$.

K	n	AEE_0	AEE_1	AEE_2	AEE_3	AEE_4
50	4×10^4	4.274	4.372	4.061	4.070	4.364
	5×10^4	3.772	3.951	3.824	3.751	3.789
	6×10^4	3.493	3.454	3.550	3.280	3.444
	7×10^4	3.124	3.326	3.183	3.266	3.198
	8×10^4	3.053	2.990	2.884	2.832	3.001
	9×10^4	2.752	2.822	2.849	2.813	2.827
100	4×10^4	4.324	4.263	4.520	4.086	4.494
	5×10^4	3.834	3.762	3.970	3.642	3.702
	6×10^4	3.441	0.350	3.200	3.467	3.559
	7×10^4	3.213	3.277	3.205	3.267	3.081
	8×10^4	2.985	2.952	3.063	3.031	2.970
	9×10^4	2.777	2.655	2.968	2.976	2.890

Table 9. The results of MSE ($\times 10^{-3}$) for chemical sensors data, where the values in the parentheses correspond to optimal bandwidth involved in the methods.

	20	50	100	150	200
LEMR	0.220	0.228	0.231	0.232	0.233
MEM	0.184	0.193	0.199	0.201	0.203
(h.opt)	(11.668)	(8.977)	(7.085)	(6.519)	(6.172)

data, where $\hat{y}_i$ is the predictive value of y_i. Figure 1 shows that the distribution of the response has two peaks and is asymmetric. The CDMR works poorly with extremely large MSE, which is not displayed here. To investigate the performance of our proposal, we also compare the MEM method of Yao and Li (2014) that uses the full data. For a fair comparison, we consider the same bandwidth selected by (9). Table 9 shows the results. The MSEs of both methods are very small and generally less than 2.4×10^{-4}. The MEM has slightly better performance than LEMR. For example, we can see that when $K = 100$, the MSE for the MEM is 1.99×10^{-4}, while our method has the MSE of 2.31×10^{-4}. The relative error is only about 16%.

Acknowledgments

We acknowledge the editor and reviewers for their contributions to the improvement of our work. The authors thank Shaomin Li for sharing us the R code for the procedures proposed in their papers.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Ma’s research was supported by the National Natural Science Foundation of China [grant number 12101439). Xia’s research was supported by Fundamental Research Funds for the Central Universities [grant number 2021CDJQY-047] and National Natural Science Foundation of China [grant number 11801202].

ORCID

Xiaochao Xia http://orcid.org/0000-0002-9414-355X

References

Chen, X., Liu, W., & Zhang, Y. (2019). Quantile regression under memory constraint. *The Annals of Statistics*, *47*(6), 3244–3273.

Chen, X., Liu, W., & Zhang, Y. (2022). First-order Newton-type estimator for distributed estimation and inference. *Journal of the American Statistical Association*, *117*(540), 1858–1874. https://doi.org/10.1080/01621459.2021.1891925

Chen, X., & Xie, M. G. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, *24*, 1655–1684.

Fonollosa, J., Sheik, S., Huerta, R., & Marco, S. (2015). Reservoir computing compensates slow response of chemosensor arrays exposed to fast varying gas concentrations in continuous monitoring. *Sensors and Actuators B: Chemical*, *215*, 618–629. https://doi.org/10.1016/j.snb.2015.03.028

Härdle, W. (1990). *Applied nonparametric regression*. Cambridge University Press.

Jordan, M., Lee, J., & Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, *114*(526), 668–681. https://doi.org/10.1080/01621459.2018.1429274

Ma, P., Mahoney, M. W., & Yu, B. (2015). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, *16*, 861–911.

Shi, C., Lu, W., & Song, R. (2018). A massive data framework for M-estimators with cubic-rate. *Journal of the American Statistical Association*, *113*(524), 1698–1709. https://doi.org/10.1080/01621459.2017.1360779

Ullah, A., & Pagan, A. (1999). *Nonparametric econometrics*. Cambridge University Press.

Wang, K., & Li, S. (2021). Robust distributed modal regression for massive data. *Computational Statistics & Data Analysis*, *160*, 107225. https://doi.org/10.1016/j.csda.2021.107225

Wang, H., & Ma, Y. (2021). Optimal subsampling for quantile regression in big data. *Biometrika*, *108*(1), 99–112. https://doi.org/10.1093/biomet/asaa043

Wang, H., Yang, M., & Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, *114*(525), 393–405. https://doi.org/10.1080/01621459.2017.1408468

Wang, H., Zhu, R., & Ma, P. (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, *113*(522), 829–844. https://doi.org/10.1080/01621459.2017.1292914

Yao, W., & Li, L. (2014). A new regression model: Modal linear regression. *Scandinavian Journal of Statistics*, *41*(3), 656–671. https://doi.org/10.1111/sjos.v41.3

Yao, W., Lindsay, B., & Li, R. (2012). Local modal regression. *Journal of Nonparametric Statistics*, *24*(3), 647–663. https://doi.org/10.1080/10485252.2012.678848

Zhang, T., Kato, K., & Ruppert, D. (2023). Bootstrap inference for quantile-based modal regression. *Journal of the American Statistical Association*, *118*(541), 122–134. https://doi.org/10.1080/01621459.2021.1918130

Zhang, H., & Wang, H. (2021). Distributed subdata selection for big data via sampling-based approach. *Computational Statistics & Data Analysis*, *153*, 107072. https://doi.org/10.1016/j.csda.2020.107072

Appendix. Proofs

Before beginning with our proof, we let $\phi(u)$ denote the density function of a standard normal random variable. It can be seen that $\phi'(u) = -x\phi(u)$ and $\phi''(u) = (u^2 - 1)\phi(u)$.

Proof of Proposition 2.1: First of all, we prove the first assertion. To this end, denote

$$\Delta_{nh}(\boldsymbol{\beta}) = \frac{1}{nh}\sum_{i=1}^{n} \boldsymbol{X}_i \varepsilon_i \phi\left(\frac{Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}}{h}\right) - \frac{1}{nh}\sum_{i=1}^{n} \boldsymbol{X}_i \varepsilon_i \phi\left(\frac{\varepsilon_i}{h}\right). \tag{A1}$$

Clearly, $\Delta_{nh}(\widehat{\boldsymbol{\beta}}^{(0)}) = \boldsymbol{A}_{nh} - \boldsymbol{A}^*_{nh}$. Our aim is to derive the probabilistic order for the term $\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \le C\xi_n} \|\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_0)\|$, where C is a large constant. Define a ball $\mathcal{B} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le C\xi_n, \boldsymbol{\beta} \in \mathbb{R}^p\}$. We select a grid of points $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K$ among $\mathcal{B}$ such that $\mathcal{B}$ can be covered by the union of a set of balls $\{\mathcal{B}_k, k = 1, \ldots, K\}$, where $\mathcal{B}_k$ means a ball with a centre $\boldsymbol{\beta}_k$ and a radius $\delta_n = \xi_n/n^2$. Then, $K = O(n^{2p})$.

Denote $\boldsymbol{\theta} \triangleq \boldsymbol{\beta} - \boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_k \triangleq \boldsymbol{\beta}_k - \boldsymbol{\beta}_0$. Write

$$\begin{aligned}\Delta_{nh}(\boldsymbol{\theta}) &\triangleq \Delta_{nh}(\boldsymbol{\beta}) = \Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_0)\\ &= \frac{1}{nh}\sum_{i=1}^{n} \boldsymbol{X}_i\varepsilon_i\left[\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\right) - \phi\left(\frac{\varepsilon_i}{h}\right)\right].\end{aligned}$$

By Taylor's expansion and properties of density function of standard normal random variable, we have that for any $\|\boldsymbol{\theta}\| \le C\xi_n$,

$$\begin{aligned}&E\left\{\frac{1}{nh}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\right)\right\}\\ &= \frac{1}{nh}\sum_{i=1}^{n}E\left\{\boldsymbol{X}_i\int u\phi\left(\frac{u - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\right)f_{\varepsilon_i\,|\,\boldsymbol{X}_i}(u\,|\,\boldsymbol{X}_i)\,\mathrm{d}u\right\}\\ &= \frac{1}{n}\sum_{i=1}^{n}E\left\{\boldsymbol{X}_i\int\left(ht + \boldsymbol{X}_i^\top\boldsymbol{\theta}\right)\phi(t)f_{\varepsilon_i\,|\,\boldsymbol{X}_i}\left(ht + \boldsymbol{X}_i^\top\boldsymbol{\theta}\,|\,\boldsymbol{X}_i\right)\mathrm{d}t\right\}\\ &= \frac{1}{n}\sum_{i=1}^{n}E\left\{\boldsymbol{X}_i f_{\varepsilon_i\,|\,\boldsymbol{X}_i}(0\,|\,\boldsymbol{X}_i)\int\left(ht + \boldsymbol{X}_i^\top\boldsymbol{\theta}\right)\phi(t)\,\mathrm{d}t\right\}\\ &\quad + \frac{1}{2n}\sum_{i=1}^{n}E\left\{\boldsymbol{X}_i f''_{\varepsilon_i\,|\,\boldsymbol{X}_i}(0\,|\,\boldsymbol{X}_i)\int\left(ht + \boldsymbol{X}_i^\top\boldsymbol{\theta}\right)^3\phi(t)\,\mathrm{d}t\right\}\{1 + o(1)\}\\ &= \boldsymbol{D}\boldsymbol{\theta} + \frac{1}{2n}\sum_{i=1}^{n}E\left\{\boldsymbol{X}_i f''_{\varepsilon_i\,|\,\boldsymbol{X}_i}(0\,|\,\boldsymbol{X}_i)\left[3h^2\boldsymbol{X}_i^\top\boldsymbol{\theta} + (\boldsymbol{X}_i^\top\boldsymbol{\theta})^3\right]\right\}\{1 + o(1)\}\\ &= \boldsymbol{D}\boldsymbol{\theta} + \left[\frac{3h^2}{2}E\left\{\boldsymbol{X}\boldsymbol{X}^\top f''_{\varepsilon\,|\,\boldsymbol{X}}(0\,|\,\boldsymbol{X})\right\}\boldsymbol{\theta}\right.\\ &\quad \left. + \frac{1}{2}E\left\{\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{\theta}\right)^3 f''_{\varepsilon\,|\,\boldsymbol{X}}(0\,|\,\boldsymbol{X})\right\}\right]\{1 + o(1)\},\end{aligned}\tag{A2}$$

and similarly,

$$\begin{aligned}&E\left\{\frac{1}{nh}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i\phi\left(\frac{\varepsilon_i}{h}\right)\right\}\\ &= \frac{1}{nh}\sum_{i=1}^{n}E\left\{\boldsymbol{X}_i\int u\phi\left(\frac{u}{h}\right)f_{\varepsilon_i\,|\,\boldsymbol{X}_i}(u\,|\,\boldsymbol{X}_i)\,\mathrm{d}u\right\}\\ &= \frac{h}{n}\sum_{i=1}^{n}E\left\{\boldsymbol{X}_i\int t\phi(t)f_{\varepsilon_i\,|\,\boldsymbol{X}_i}(ht\,|\,\boldsymbol{X}_i)\,\mathrm{d}t\right\}\\ &= \frac{h}{6n}\sum_{i=1}^{n}E\left\{\boldsymbol{X}_i f'''_{\varepsilon_i\,|\,\boldsymbol{X}_i}(0\,|\,\boldsymbol{X}_i)\int t\phi(t)(ht)^3\,\mathrm{d}t\right\}\{1 + o(1)\}\\ &= \frac{h^4}{2}E\left\{\boldsymbol{X}f'''_{\varepsilon\,|\,\boldsymbol{X}}(0\,|\,\boldsymbol{X})\right\}\{1 + o(1)\}.\end{aligned}\tag{A3}$$

Thus, it follows from Conditions (C2) and (C4) that

$$\|E\{\Delta_{nh}(\boldsymbol{\beta})\} - \boldsymbol{D}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\| = O(h^2\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^3 + h^4).$$

Accordingly, we have

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le C\xi_n} \|E\{\Delta_{nh}(\boldsymbol{\beta})\} - \boldsymbol{D}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\| \le O(\xi_n h^2 + \xi_n^3 + h^4). \tag{A4}$$

On the other hand, we have

$$\begin{aligned}
&\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le C\xi_n} \|[\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_0)] - E\{\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_0)\}\| \\
&\quad = \max_{1\le k\le K} \sup_{\boldsymbol{\beta}\in\mathcal{B}_k} \|\Delta_{nh}(\boldsymbol{\beta}) - E\{\Delta_{nh}(\boldsymbol{\beta})\}\| \\
&\quad \le \max_{1\le k\le K} \sup_{\boldsymbol{\beta}\in\mathcal{B}_k} \|\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_k) - E\{\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_k)\}\| \\
&\qquad + \max_{1\le k\le K} \|\Delta_{nh}(\boldsymbol{\beta}_k) - E\{\Delta_{nh}(\boldsymbol{\beta}_k)\}\| \\
&\quad =: I_{n1} + I_{n2}.
\end{aligned} \tag{A5}$$

We first consider the term I_{n1}. We can decompose

$$\begin{aligned}
&\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_k) \\
&= \frac{1}{nh}\sum_{i=1}^{n} \boldsymbol{X}_i\varepsilon_i \left[\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\right) - \phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \left[\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\right) - \frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\right)\right] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \left[\frac{\boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\right) - \frac{\boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \left[\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\right) - \frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\right)\right] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \left[\frac{\boldsymbol{X}_i^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_k)}{h}\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\right)\right] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \frac{\boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\left[\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\right) - \phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\right)\right].
\end{aligned}$$

Using the facts that both $\phi(x)$ and $x\phi(x)$ are Lipschitz continuous and $\phi(x)$ is bounded, we have

$$\|\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_k)\| \le \frac{C_2}{nh}\sum_{i=1}^{n} \|\boldsymbol{X}_i\|^2\|\boldsymbol{\beta}-\boldsymbol{\beta}_k\|\{1 + C_2' h^{-1}\|\boldsymbol{X}_i\|\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_0\|\}$$

for some positive constants C_2 and C_2'. It follows from Condition (C4) and $\xi_n = O(h)$ that

$$E\left\{\max_{1\le k\le K}\sup_{\boldsymbol{\beta}\in\mathcal{B}_k} \|\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_k)\|\right\} = O\left(\frac{\xi_n}{n^2 h}\right).$$

This together with Markov's inequality implies that

$$\begin{aligned}
I_{n1} &\le \max_{1\le k\le K}\sup_{\boldsymbol{\beta}\in\mathcal{B}_k} \|\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_k)\| \\
&\quad + \max_{1\le k\le K}\sup_{\boldsymbol{\beta}\in\mathcal{B}_k} E\{\|\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_k)\|\} \\
&= O_p\left(\frac{\xi_n}{n^2 h}\right).
\end{aligned} \tag{A6}$$

Next, we consider the term I_{n2}. Let $\psi_{ijk} \triangleq h^{-1}X_{ij}\varepsilon_i\left[\phi\left(\frac{\varepsilon_i-\boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\right)-\phi(\frac{\varepsilon_i}{h})\right]$. Then, by the Lipschitz's continuity of $x\phi(x)$ and boundedness of $\phi(x)$, we have that for any $\boldsymbol{\beta}\in\mathcal{B}_k$, there exists some positive constant C_3 such that

$$
\begin{aligned}
|\psi_{ijk}| &= \left|X_{ij}\left[\frac{\varepsilon_i-\boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\phi\left(\frac{\varepsilon_i-\boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\right)-\frac{\varepsilon_i}{h}\phi\left(\frac{\varepsilon_i}{h}\right)+\frac{\boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\phi\left(\frac{\varepsilon_i-\boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\right)\right]\right| \\
&\le |X_{ij}|\left|\frac{\varepsilon_i-\boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\phi\left(\frac{\varepsilon_i-\boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\right)-\frac{\varepsilon_i}{h}\phi\left(\frac{\varepsilon_i}{h}\right)\right|+|X_{ij}|\frac{\|\boldsymbol{X}_i\|\|\boldsymbol{\theta}_k\|}{h}\phi\left(\frac{\varepsilon_i-\boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\right) \\
&\le C_3h^{-1}|X_{ij}|\|\boldsymbol{X}_i\|\|\boldsymbol{\theta}_k\| \\
&\le C_3C\xi_nh^{-1}|X_{ij}|\|\boldsymbol{X}_i\| =: V_{ij}.
\end{aligned}
$$

Denote $e_n=\frac{\xi_n}{h}(\frac{n}{\log n})^{1/4}$ and let $\psi_{ijk}^{\dagger}=\psi_{ijk}I(V_{ij}\le e_n)$ and $\psi_{ijk}^{\ddagger}=\psi_{ijk}I(V_{ij}>e_n)$. Using the fact that $\|\mathbf{a}\|\le\|\mathbf{a}\|_1=\sum_i|a_i|$, we have

$$
\begin{aligned}
I_{n2} &= \max_{1\le k\le K}\|\Delta_{nh}(\boldsymbol{\beta}_k)-E\{\Delta_{nh}(\boldsymbol{\beta}_k)\}\| \\
&\le \max_{1\le k\le K}\sum_{j=1}^{p}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\psi_{ijk}-E\left(\psi_{ijk}\right)\right]\right| \\
&\le \max_{1\le k\le K}\sum_{j=1}^{p}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\psi_{ijk}^{\dagger}-E\left(\psi_{ijk}^{\dagger}\right)\right]\right| \\
&\quad+\max_{1\le k\le K}\sum_{j=1}^{p}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\psi_{ijk}^{\ddagger}-E\left(\psi_{ijk}^{\ddagger}\right)\right]\right| \\
&=: I_{n2,1}+I_{n2,2}.
\end{aligned}
$$

Let $\delta_n=Le_n\sqrt{\frac{\log n}{n}}$, where L denotes a sufficiently large positive constant. It follows from Boole's inequality and Hoeffding's inequality that

$$
\begin{aligned}
P\left\{I_{n2,1}>\delta_n\right\} &= P\left\{\max_{1\le k\le K}\sum_{j=1}^{p}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\psi_{ijk}^{\dagger}-E\left(\psi_{ijk}^{\dagger}\right)\right]\right|>\delta_n\right\} \\
&\le \sum_{k=1}^{K}\sum_{j=1}^{p}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\left[\psi_{ijk}^{\dagger}-E\left(\psi_{ijk}^{\dagger}\right)\right]\right|>\delta_n/p\right\} \\
&\le 2Kp\exp\left(-\frac{\delta_n^2n}{2e_n^2p^2}\right) \\
&= O(1)\exp\left(2p\log n+\log p-\frac{\delta_n^2n}{2e_n^2p^2}\right) \\
&= O(1)\exp(-(L^2/p^2-2p)\log n+\log p)=o(1)
\end{aligned}
$$

for a sufficiently large n, where the last line is because p is fixed and the constant L is large enough. This implies that

$$
I_{n2,1}=O_p\left(e_n\sqrt{\frac{\log n}{n}}\right)=O_p\left(\frac{\xi_n}{h}\left(\frac{\log n}{n}\right)^{1/4}\right). \tag{A7}
$$

On the other hand, we first note that

$$I_{n2,2} = \max_{1\le k\le K}\sum_{j=1}^{p}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\psi_{ijk}^{\ddagger} - E\left(\psi_{ijk}^{\ddagger}\right)\right]\right|$$

$$\le \max_{1\le k\le K}\sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n}|\psi_{ijk}^{\ddagger}| + \max_{1\le k\le K} E\left\{\sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n}|\psi_{ijk}^{\ddagger}|\right\}.$$

By Condition (C4), we can derive that

$$E\left\{\max_{1\le k\le K}\sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n}|\psi_{ijk}^{\ddagger}|\right\} \le E\left\{\sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n}V_{ij}I(V_{ij} > e_n)\right\}$$

$$\le \frac{1}{e_n}E\left\{\sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n}V_{ij}^2\right\}$$

$$= O\left(\frac{\xi_n^2}{e_n h^2}\right)$$

$$= O\left(\frac{\xi_n}{h}\left(\frac{\log n}{n}\right)^{1/4}\right).$$

This in conjunction with Markov's inequality leads to

$$I_{n2,2} = O_p\left(\frac{\xi_n}{h}\left(\frac{\log n}{n}\right)^{1/4}\right). \tag{A8}$$

As a result, by combining the results (A7) and (A8), we have $I_{n2} = O_p\left(\frac{\xi_n}{h}\left(\frac{\log n}{n}\right)^{1/4}\right)$. This together with (A5) and (A6) gives

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le C\xi_n}\|[\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_0)] - E\{\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_0)\}\| = O_p\left(\frac{\xi_n}{h}\left(\frac{\log n}{n}\right)^{1/4}\right).$$

This combining (A4) implies

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le C\xi_n}\|\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_0) - \boldsymbol{D}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|$$

$$\le \sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le C\xi_n}\|[\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_0)] - E\{\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_0)\}\|$$

$$+ \sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le C\xi_n}\|E\{\Delta_{nh}(\boldsymbol{\beta}) - \Delta_{nh}(\boldsymbol{\beta}_0)\} - \boldsymbol{D}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|$$

$$= O_p\left(\frac{\xi_n}{h}\left(\frac{\log n}{n}\right)^{1/4} + \xi_n h^2 + \xi_n^3 + h^4\right).$$

Thus, the assertion (i) is proved.

Secondly, we are going to verify the second assertion. To proceed, let

$$\Lambda_{nh}(\boldsymbol{\beta}) = \frac{1}{nh}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\top\left[\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\right) - \phi\left(\frac{\varepsilon_i}{h}\right)\right],$$

where $\boldsymbol{\theta} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$. Obviously, we have $\Lambda_{nh}(\widehat{\boldsymbol{\beta}}^{(0)}) = \boldsymbol{D}_{nh} - \boldsymbol{D}^*_{nh}$. Write $\psi_{nijl}(\boldsymbol{\theta}) = h^{-1}X_{ij}X_{il}[\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\right) - \phi(\frac{\varepsilon_i}{h})]$. Then, the (j, l)th element of $\Lambda_{nh}(\boldsymbol{\beta})$ is $\frac{1}{n}\sum_{i=1}^n \psi_{nijl}(\boldsymbol{\theta})$. Further, by Taylor's expansion,

$$E\left\{\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\right)\middle|\boldsymbol{X}_i\right\} = hf_{\varepsilon_i\,|\,\boldsymbol{X}_i}(0\,|\,\boldsymbol{X}_i) + \frac{h}{2}f''_{\varepsilon_i\,|\,\boldsymbol{X}_i}(0\,|\,\boldsymbol{X}_i)\left[h^2 + (\boldsymbol{X}_i^\top\boldsymbol{\theta})^2\right]\{1 + o(1)\}$$

and

$$E\left\{\phi\left(\frac{\varepsilon_i}{h}\right)\middle|\boldsymbol{X}_i\right\} = hf_{\varepsilon_i\,|\,\boldsymbol{X}_i}(0\,|\,\boldsymbol{X}_i) + \frac{1}{2}f''_{\varepsilon_i\,|\,\boldsymbol{X}_i}(0\,|\,\boldsymbol{X}_i)h^3\{1 + o(1)\}.$$

Thus, it follows that

$$\begin{aligned}E\{\psi_{nijl}(\boldsymbol{\theta})\} &= E\left\{h^{-1}X_{ij}X_{il}E\left[\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\right) - \phi\left(\frac{\varepsilon_i}{h}\right)\middle|\boldsymbol{X}_i\right]\right\}\\ &= \frac{1}{2}E\left\{X_{ij}X_{il}f''_{\varepsilon_i\,|\,\boldsymbol{X}_i}(0\,|\,\boldsymbol{X}_i)(\boldsymbol{X}_i^\top\boldsymbol{\theta})^2\right\}\{1 + o(1)\}.\end{aligned}$$

Using the fact that $\|\mathbf{A}\| \le p\|\mathbf{A}\|_\infty$ for any symmetric $p \times p$ square matrix $\mathbf{A}$, where $\|\mathbf{A}\|_\infty = \max_{j,l}|A_{jl}|$, we have

$$\begin{aligned}&\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le C\xi_n}\|\Lambda_{nh}(\boldsymbol{\beta}) - \Lambda_{nh}(\boldsymbol{\beta}_0) - E\{\Lambda_{nh}(\boldsymbol{\beta}) - \Lambda_{nh}(\boldsymbol{\beta}_0)\}\|\\ &\le \max_{1\le j\le p}\max_{1\le l\le p}\sup_{\|\boldsymbol{\theta}\|\le C\xi_n}\left|\frac{1}{n}\sum_{i=1}^n\left[\psi_{nijl}(\boldsymbol{\theta}) - E\{\psi_{nijl}(\boldsymbol{\theta})\}\right]\right|.\end{aligned} \tag{A9}$$

On the other hand, note that

$$\begin{aligned}&\sup_{\|\boldsymbol{\theta}\|\le C\xi_n}\left|\frac{1}{n}\sum_{i=1}^n\left[\psi_{nijl}(\boldsymbol{\theta}) - E\{\psi_{nijl}(\boldsymbol{\theta})\}\right]\right|\\ &\le \max_{1\le k\le K}\sup_{\boldsymbol{\beta}\in\mathcal{B}_k}\left|\frac{1}{n}\sum_{i=1}^n\left[\psi_{nijl}(\boldsymbol{\theta}) - \psi_{nijl}(\boldsymbol{\theta}_k) - E\{\psi_{nijl}(\boldsymbol{\theta}) - \psi_{nijl}(\boldsymbol{\theta}_k)\}\right]\right|\\ &\quad + \max_{1\le k\le K}\left|\frac{1}{n}\sum_{i=1}^n\left[\psi_{nijl}(\boldsymbol{\theta}_k) - E\{\psi_{nijl}(\boldsymbol{\theta}_k)\}\right]\right|\\ &\le \max_{1\le k\le K}\sup_{\boldsymbol{\beta}\in\mathcal{B}_k}\left|\frac{1}{n}\sum_{i=1}^n\left[\psi_{nijl}(\boldsymbol{\theta}) - \psi_{nijl}(\boldsymbol{\theta}_k)\right]\right|\\ &\quad + \max_{1\le k\le K}\sup_{\boldsymbol{\beta}\in\mathcal{B}_k}\left|\frac{1}{n}\sum_{i=1}^n\left[E\{\psi_{nijl}(\boldsymbol{\theta}) - \psi_{nijl}(\boldsymbol{\theta}_k)\}\right]\right|\\ &\quad + \max_{1\le k\le K}\left|\frac{1}{n}\sum_{i=1}^n\left[\psi_{nijl}(\boldsymbol{\theta}_k) - E\{\psi_{nijl}(\boldsymbol{\theta}_k)\}\right]\right|\\ &\le \max_{1\le k\le K}\sup_{\boldsymbol{\beta}\in\mathcal{B}_k}\frac{1}{n}\sum_{i=1}^n\left|h^{-1}X_{ij}X_{il}\left[\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\right) - \phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\right)\right]\right|\\ &\quad + E\left\{\max_{1\le k\le K}\sup_{\boldsymbol{\beta}\in\mathcal{B}_k}\frac{1}{n}\sum_{i=1}^n\left|h^{-1}X_{ij}X_{il}\left[\phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}}{h}\right) - \phi\left(\frac{\varepsilon_i - \boldsymbol{X}_i^\top\boldsymbol{\theta}_k}{h}\right)\right]\right|\right\}\end{aligned}$$

$$+\max_{1\le k\le K}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\psi_{nijl}(\boldsymbol{\theta}_k)-E\{\psi_{nijl}(\boldsymbol{\theta}_k)\}\right]\right|$$
$$=: T_{njl}^{(1)}+T_{njl}^{(2)}+T_{njl}^{(3)}.$$

We first consider the second term $T_{njl}^{(2)}$. Because $\phi(u)$ is Lipschitz continuous, so there exists a universal constant $C_4>0$ such that

$$\begin{aligned}\max_{1\le j\le p}\max_{1\le l\le p}T_{njl}^{(2)} &\le C_4h^{-2}\max_{1\le j\le p}\max_{1\le l\le p}E\left\{\max_{1\le k\le K}\sup_{\boldsymbol{\beta}\in\mathcal{B}_k}\frac{1}{n}\sum_{i=1}^{n}|X_{ij}||X_{il}|\|\boldsymbol{X}_i\|\|\boldsymbol{\beta}-\boldsymbol{\beta}_k\|\right\}\\ &\le C_4\frac{\xi_n}{n^2h^2}\sum_{1\le j\le p}\sum_{1\le l\le p}E\left\{\frac{1}{n}\sum_{i=1}^{n}|X_{ij}||X_{il}|\|\boldsymbol{X}_i\|\right\}\\ &\le C_4\frac{p\xi_n}{n^2h^2}\frac{1}{n}\sum_{i=1}^{n}E\left\{\|\boldsymbol{X}_i\|^3\right\}\\ &=O\left(\frac{\xi_n}{n^2h^2}\right),\end{aligned}\tag{A10}$$

where the second and third lines use the basic inequalities that $\max_{j,l}|a_{ij}|\le\sum_{i,j}|a_{ij}|$ and $(\sum_{l=1}^{p}|a_l|)^2\le p\sum_{l=1}^{p}a_l^2$, respectively, and the last line is due to Condition (C4). For the first term $T_{njl}^{(1)}$, using this result and by Markov's inequality, we can easily obtain that

$$\max_{1\le j\le p}\max_{1\le l\le p}T_{njl}^{(1)}=O_p\left(\frac{\xi_n}{n^2h^2}\right).\tag{A11}$$

We next handle the third term $T_{njl}^{(3)}$. For any $\boldsymbol{\beta}_k\in\mathcal{B}$, there exists a universal constant $C_5>0$ such that $\max_{1\le k\le K}|\psi_{nijl}(\boldsymbol{\theta}_k)|\le C_5\xi_nh^{-2}|X_{ij}||X_{il}|\|\boldsymbol{X}_i\| =: V_{ijl}$. Let $\bar{e}_n=\frac{\xi_n}{h^2}\left(\frac{\log n}{n}\right)^{-1/4}$ and write $\psi_{nijlk}^{\dagger}=\psi_{nijl}(\boldsymbol{\theta}_k)I(V_{ijl}\le\bar{e}_n)$ and $\psi_{nijlk}^{\ddagger}=\psi_{nijl}(\boldsymbol{\theta}_k)I(V_{ijl}>\bar{e}_n)$. Then, it follows that

$$\begin{aligned}\max_{1\le j\le p}\max_{1\le l\le p}T_{njl}^{(3)} &\le \max_{1\le j\le p}\max_{1\le l\le p}\max_{1\le k\le K}\left|\frac{1}{n}\sum_{i=1}^{n}[\psi_{nijlk}^{\dagger}-E\{\psi_{nijlk}^{\dagger}\}]\right|\\ &+\max_{1\le j\le p}\max_{1\le l\le p}\max_{1\le k\le K}\left|\frac{1}{n}\sum_{i=1}^{n}[\psi_{nijlk}^{\ddagger}-E\{\psi_{nijlk}^{\ddagger}\}]\right|\\ &=: II_{n1}+II_{n2}.\end{aligned}$$

Following the same arguments as used in (A7), we can derive that

$$II_{n1}=O_p\left(\bar{e}_n\left(\frac{\log n}{n}\right)^{1/2}\right)=O_p\left(\frac{\xi_n}{h^2}\left(\frac{\log n}{n}\right)^{1/4}\right).$$

Furthermore, since V_{ijl} is positive, we have

$$\begin{aligned}E\{II_{n2}\} &\le 2E\left\{\max_{1\le j\le p}\max_{1\le l\le p}\max_{1\le k\le K}\frac{1}{n}\sum_{i=1}^{n}|\psi_{nijlk}^{\ddagger}|\right\}\\ &\le 2E\left\{\max_{1\le j\le p}\max_{1\le l\le p}\frac{1}{n}\sum_{i=1}^{n}V_{ijl}I(V_{ijl}>\bar{e}_n)\right\}\\ &\le 2E\left\{V_{ijl}I(V_{ijl}>\bar{e}_n)\right\}\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{2}{\bar{e}_n} \sum_{j=1}^{p} \sum_{l=1}^{p} \frac{1}{n} \sum_{i=1}^{n} E\left\{V_{ijl}^2\right\} \\
&= \frac{2C_5^2 \xi_n^2}{\bar{e}_n h^4} \frac{1}{n} \sum_{i=1}^{n} E\left\{\|\boldsymbol{X}_i\|^4\right\} \\
&= O\left(\frac{\xi_n^2}{\bar{e}_n h^4}\right).
\end{aligned}
$$

Thus, by Markov's inequality, we have

$$
II_{n2} = O_p\left(\frac{\xi_n^2}{\bar{e}_n h^4}\right) = O_p\left(\frac{\xi_n}{h^2}\left(\frac{\log n}{n}\right)^{1/4}\right).
$$

Hence, it follows that

$$
\max_{1\leq j\leq p} \max_{1\leq l\leq p} T_{njl}^{(3)} = O_p\left(\frac{\xi_n}{h^2}\left(\frac{\log n}{n}\right)^{1/4}\right). \tag{A12}
$$

Invoking the results (A9)–(A12), we obtain

$$
\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\leq C\xi_n} \|\Lambda_{nh}(\boldsymbol{\beta}) - \Lambda_{nh}(\boldsymbol{\beta}_0) - E\{\Lambda_{nh}(\boldsymbol{\beta}) - \Lambda_{nh}(\boldsymbol{\beta}_0)\}\| = O_p\left(\frac{\xi_n}{h^2}\left(\frac{\log n}{n}\right)^{1/4}\right). \tag{A13}
$$

By the previous arguments and Conditions (C2) and (C4), we can obtain

$$
\begin{aligned}
&\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\leq C\xi_n} \|E\{\Lambda_{nh}(\boldsymbol{\beta}) - \Lambda_{nh}(\boldsymbol{\beta}_0)\}\| \\
&\leq \max_{1\leq j\leq p} \max_{1\leq l\leq p} \sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\leq C\xi_n} \left| E\left\{\frac{1}{n}\sum_{i=1}^{n} \psi_{nijl}(\boldsymbol{\theta})\right\}\right| \\
&\leq C_7 \max_{1\leq j\leq p} \max_{1\leq l\leq p} \sup_{\|\boldsymbol{\theta}\|\leq C\xi_n} \frac{1}{n}\sum_{i=1}^{n} E\left\{|X_{ij}||X_{il}||f''_{\varepsilon_i\,|\,\boldsymbol{X}_i}(0\,|\,\boldsymbol{X}_i)|\|\boldsymbol{X}_i\|^2\right\} \|\boldsymbol{\theta}\|^2 \\
&\leq C_7 C^2 p \xi_n^2 \frac{1}{n}\sum_{i=1}^{n} E\left\{\|\boldsymbol{X}_i\|^4 |f''_{\varepsilon_i\,|\,\boldsymbol{X}_i}(0|\boldsymbol{X}_i)|\right\} \\
&\leq C_8 p \xi_n^2 E\left\{\|\boldsymbol{X}\|^4\right\} \\
&= O(\xi_n^2),
\end{aligned}
$$

where C_7 and C_8 are some positive constants. Connecting this result with (A13) yields

$$
\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\leq C\xi_n} \|\Lambda_{nh}(\boldsymbol{\beta}) - \Lambda_{nh}(\boldsymbol{\beta}_0)\| = O_p\left(\frac{\xi_n}{h^2}\left(\frac{\log n}{n}\right)^{1/4} + \xi_n^2\right).
$$

Therefore, the second assertion is proved.

Thirdly, we verify the assertion (iii). Let $\varphi_{ijl} = X_{ij}X_{il}h^{-1}\phi(\frac{\varepsilon_i}{h})$. On one hand, by Conditions (C2) and (C4) and Taylor's expansion, we can derive that for any $\delta > 0$,

$$
\begin{aligned}
P\left\{\|\boldsymbol{D}_{nh}^* - E\{\boldsymbol{D}_{nh}^*\}\| > \delta\right\} &\leq P\left\{\max_{1\leq j\leq p} \max_{1\leq l\leq p} \left|\frac{1}{n}\sum_{i=1}^{n}[\varphi_{ijl} - E\{\varphi_{ijl}\}]\right| > \delta\right\} \\
&\leq \sum_{1\leq j\leq p} \sum_{1\leq l\leq p} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}[\varphi_{ijl} - E\{\varphi_{ijl}\}]\right| > \delta\right\}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{1\leq j\leq p}\sum_{1\leq l\leq p}\delta^{-2}E\left\{\frac{1}{n^2}\sum_{i=1}^{n}X_{ij}^2X_{il}^2h^{-2}\phi^2\left(\frac{\varepsilon_i}{h}\right)\right\}\\
&\leq C_9\delta^{-2}(nh)^{-1}
\end{aligned}
$$

for some positive constant C_9. By taking $\delta = L/\sqrt{nh}$ with L being a constant, as $L \to \infty$, $P\{\|\boldsymbol{D}_{nh}^* - E\{\boldsymbol{D}_{nh}^*\}\| > L/\sqrt{nh}\} = o(1)$, which implies that $\|\boldsymbol{D}_{nh}^* - E\{\boldsymbol{D}_{nh}^*\}\| = O_P\left(\frac{1}{\sqrt{nh}}\right)$. On the other hand, it can be easily derived that

$$
\begin{aligned}
\|E\{\boldsymbol{D}_{nh}^*\} - \boldsymbol{D}\| &\leq \sum_{1\leq j\leq p}\sum_{1\leq l\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}E\left\{X_{ij}X_{il}\int\phi(t)[f_{\varepsilon_i\,|\,\boldsymbol{X}_i}(ht\,|\,\boldsymbol{X}_i) - f_{\varepsilon_i\,|\,\boldsymbol{X}_i}(0\,|\,\boldsymbol{X}_i)]\,\mathrm{d}t\right\}\right|\\
&\leq C_{10}ph^2\frac{1}{n}\sum_{i=1}^{n}E\left\{\|\boldsymbol{X}_i\|^2\right\} = O(h^2)
\end{aligned}
$$

for some positive constant C_{10}. Thus, the assertion (iii) follows by using the above arguments and the triangle inequality of norms. ■

Proof of Proposition 2.2: We first derive the probabilistic orders of $\|E\{\boldsymbol{A}_{nh}^*\}\|$ and $\|\boldsymbol{A}_{nh}^* - E\{\boldsymbol{A}_{nh}^*\}\|$, respectively. An application of Taylor's expansion yields

$$
E\{\boldsymbol{A}_{nh}^*\} = \frac{h^4}{2n}\sum_{i=1}^{n}E\{\boldsymbol{X}_i f_{\varepsilon_i\,|\,\boldsymbol{X}_i}'''(0\,|\,\boldsymbol{X}_i)\}\{1 + o(1)\},
$$

which in conjunction with Conditions (C2) and (C4) leads to

$$
\|E\{\boldsymbol{A}_{nh}^*\}\| = O(h^4). \tag{A14}
$$

Furthermore, by Taylor's expansion and Conditions (C2) and (C4), we have

$$
\begin{aligned}
E\left\{\|\boldsymbol{A}_{nh}^* - E\{\boldsymbol{A}_{nh}^*\}\|^2\right\} &= E\left\{\sum_{j=1}^{p}\left(\frac{1}{nh}\sum_{i=1}^{n}\left[X_{ij}\varepsilon_i\phi\left(\frac{\varepsilon_i}{h}\right) - EX_{ij}\varepsilon_i\phi\left(\frac{\varepsilon_i}{h}\right)\right]\right)^2\right\}\\
&\leq \frac{1}{n^2h^2}\sum_{j=1}^{p}\sum_{i=1}^{n}E\left\{X_{ij}^2\varepsilon_i^2\phi^2\left(\frac{\varepsilon_i}{h}\right)\right\}\\
&= O\left(\frac{h}{n}\right),
\end{aligned}
$$

which implies

$$
\|\boldsymbol{A}_{nh}^* - E\boldsymbol{A}_{nh}^*\| = O_P\left(\sqrt{\frac{h}{n}}\right). \tag{A15}
$$

Thus, combining (A14) and (A15) gives

$$
\begin{aligned}
\|\boldsymbol{A}_{nh}^*\| &\leq \|\boldsymbol{A}_{nh}^* - E\{\boldsymbol{A}_{nh}^*\}\| + \|E\{\boldsymbol{A}_{nh}^*\}\|\\
&= O_P\left(\sqrt{\frac{h}{n}} + h^4\right).
\end{aligned} \tag{A16}
$$

Hence, it follows from the result (A16), Proposition 1, Conditions (C3) and the fact that $\boldsymbol{D}_{nh}^{-1} - \boldsymbol{D}^{-1} = \boldsymbol{D}^{-1}(\boldsymbol{D} - \boldsymbol{D}_{nh})\boldsymbol{D}_{nh}^{-1}$ that

$$
\begin{aligned}
\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 - \boldsymbol{D}^{-1}\boldsymbol{A}_{nh}^*\| &= \|\boldsymbol{D}_{nh}^{-1}\boldsymbol{A}_{nh} - \boldsymbol{D}^{-1}\boldsymbol{A}_{nh}^*\|\\
&\leq \|(\boldsymbol{D}_{nh}^{-1} - \boldsymbol{D}^{-1})\boldsymbol{A}_{nh}\| + \|\boldsymbol{D}^{-1}(\boldsymbol{A}_{nh} - \boldsymbol{A}_{nh}^*)\|
\end{aligned}
$$

$$
\begin{aligned}
&\leq \|\boldsymbol{D}^{-1}\|\|\boldsymbol{D}-\boldsymbol{D}_{nh}\|\|\boldsymbol{D}_{nh}^{-1}\|\|\boldsymbol{A}_{nh}\| + \|\boldsymbol{D}^{-1}\|\|\boldsymbol{A}_{nh}-\boldsymbol{A}_{nh}^{*}\| \\
&\leq \|\boldsymbol{D}^{-1}\|\|\boldsymbol{D}-\boldsymbol{D}_{nh}\|\|\boldsymbol{D}_{nh}^{-1}\|\|\boldsymbol{A}_{nh}-\boldsymbol{A}_{nh}^{*}\| \\
&\quad + \|\boldsymbol{D}^{-1}\|\|\boldsymbol{D}-\boldsymbol{D}_{nh}\|\|\boldsymbol{D}_{nh}^{-1}\|\|\boldsymbol{A}_{nh}^{*}\| + \|\boldsymbol{D}^{-1}\|\|\boldsymbol{A}_{nh}-\boldsymbol{A}_{nh}^{*}\| \\
&= O_p\left(\frac{\xi_n^2}{h^2}\left(\frac{\log n}{n}\right)^{1/4} + \xi_n + h\left(h^2 \vee \frac{1}{\sqrt{nh}}\right)^2\right).
\end{aligned}
$$

■