Annals of the Institute of Statistical Mathematics (2020) 72:427–449
https://doi.org/10.1007/s10463-018-0699-0

Error density estimation in high-dimensional sparse linear model

Feng Zou[1] · **Hengjian Cui**[1]

Received: 19 March 2018 / Revised: 18 September 2018 / Published online: 16 November 2018

Abstract
This paper is concerned with the error density estimation in high-dimensional sparse linear model, where the number of variables may be larger than the sample size. An improved two-stage refitted cross-validation procedure by random splitting technique is used to obtain the residuals of the model, and then traditional kernel density method is applied to estimate the error density. Under suitable sparse conditions, the large sample properties of the estimator including the consistency and asymptotic normality, as well as the law of the iterated logarithm are obtained. Especially, we gave the relationship between the sparsity and the convergence rate of the kernel density estimator. The simulation results show that our error density estimator has a good performance. A real data example is presented to illustrate our methods.

Keywords High-dimensional sparse linear model · Kernel density estimation · Refitted cross-validation method · Asymptotic properties · Law of the iterated logarithm

1 Introduction

Error density estimation is a basic problem in statistical modeling. For conventional linear model, we often assume that the error satisfies Gaussian distribution with mean zero; then, ordinary least square (OLS) approach is used to estimate regression coefficients and corresponding statistical inference on coefficients could be also set up.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-018-0699-0) contains supplementary material, which is available to authorized users.

✉ Hengjian Cui
hjcui@bnu.edu.cn

Feng Zou
zoufengjiayou@163.com

[1] Department of Statistics, School of Mathematical Sciences, Capital Normal University, No.105, West Third Ring Road, Haidian District, Beijing 100048, China

Under the normality assumption, it is well known that the OLS estimator has some nice properties. Meanwhile, it is equivalent to the maximum likelihood estimator (MLE). Hence, the OLS method has occupied a prominent place in application fields for a long time. However, the normality assumption in many practical problems is unreasonable. From this point of view, it is very necessary to estimate accurately the error density function other than simple assumption. Another motivation of the problem is that we have to estimate error density when a statistical inference on regression coefficients is considered. Specifically speaking, when some observations of the response are truncated in the linear model, Powell (1984) showed that the coefficients could be estimated by a least absolute deviation criterion, and then the estimator's asymptotical normality was proved under suitable regularity conditions. Unfortunately, there is an unknown error density function in the representation of the asymptotic covariance matrix. That is, it is sometimes necessary for us to estimate error density in order to make statistical inference on regression coefficients.

In this paper, a nonparametric kernel method is proposed for the error density estimation. In the past few decades, this method has been extensively applied to statistical regression models and proved to have many large sample properties such as consistency, the asymptotic normality and the law of iterated logarithm. The case of error distribution in conventional linear regression model was studied in Chai and Li (1993), where the unknown slope parameters were estimated with OLS method; then, an estimation method based on Parzen kernel function was used for estimating error density and asymptotic theories of the estimator were also obtained. More related work could be referred to Yang (1997) and Cheng (2005) for nonparametric regression model, Liang and Hardle (1999) for semi-parametric model and so on.

Without loss of generality, the error density estimation could be accomplished by two procedures. First, estimate the unknown parameters. Second, the kernel method based on the residuals is applied to fit the error density function. However, in the high-dimensional linear model where the number of covariates is greater than the sample size, and OLS method is invalid to parameter estimation due to suffering from the singularity problem. That is, the sample covariance matrix is not invertible when the dimensionality is larger than the sample size. In another "high-dimensional" problem where the sample size is larger than dimensionality but the dimensionality adds up with the multiple of the sample, the performance of OLS method is poor. Here, we take Fig. 6 as an illustration. Obviously, the fitted kernel density curve is over-fitted, which mainly blames to the inefficiency of OLS method. In this paper, we propose and compare several methods for error density estimation in high-dimensional linear model. As well as Fan and Lv (2008), Fan et al. (2012), the sparse condition is not negligible to make the problem solvable. In a word, the number of nonzero components is small relative to the sample size, which guarantees the variable selection procedure can identify all the important predictors.

Therefore, to meet aforementioned challenges, a direct idea is to apply the kernel approach to estimate error density after model selection and estimation in high-dimensional linear model. Recently, a large number of variable selection methods have been proposed in high-dimensional data analysis, which has become more and more important in various research fields. Examples based on the specific model include the LASSO (Tibshirani 1996), smoothly clipped absolute deviation called SCAD (Fan

and Li 2001), the adaptive LASSO (Zou 2006), Dantzig selector (Candes and Tao 2007), SIS (Fan and Lv 2008) and MCP (Zhang 2010), please see the article by Fan and Lv (2010) for an overview. Furthermore, to conquer the challenge of specifying a correct model, a series of model-free sure screening procedures have been developed by many authors, for example, the DC-SIS (Li et al. 2012), the RoSIS (Zhong 2014), the MV-SIS (Cui et al. 2015) and so on. In this paper, we are concerned with high-dimensional sparse linear model.

As mentioned above, a natural idea to estimate the error density is the following two-stage procedure. In the first stage, a model selection tool such as LASSO, SCAD, SIS is applied to select a subset model, which includes all important predictors with the moderate size smaller than the sample size even if it is not exactly the true model. In the second stage, the regression coefficients are estimated by OLS method, and then kernel method is used to estimate the error density on the foundation of residuals. It is apparent that if we can recover exactly the true model in the first stage, then the two-stage procedure will work well. Unfortunately, the facts ran counter to our expectations. Intuitively, we can see that a over-fitted phenomenon is caused from Fig. 1, which can be easily illustrated by endogeneity between variables. According to Fan et al. (2012), when the number of irrelevant variables is huge, some of them have large sample correlations with the realized noises, which are called spurious variables in their paper. Along with the appearance of spurious variables, the model finally selected is over-fitted. As Fan et al. (2012) showed, the naivety of two-stage methods led to a serious underestimate of the residual variance. Similarly, it is unreasonable to estimate the error density by naive two-stage method in high-dimensional linear model. Since almost all variable selection procedure will select these spurious variables with high probability when the model is over-fitted, we will introduce refitted cross-validation (RCV) method (Fan et al. 2012) to work out our challenges. The later simulations suggest that the performance of RCV method is nice. To complete the above missions, the rest of paper is organized as follows. In Sect. 2, we make an explanation about nice properties of RCV method and introduce how to apply kernel method to estimate error density in high-dimensional linear model. In Sect. 3, we present some conditions and theoretical results. A set of simulation studies and a real data example are given in Sect. 4. In Sect. 5, we detailedly justify all the theorems.

2 Methodology for error density estimation

Consider the following linear model

$$y_i = X_i^{\mathrm{T}}\beta + e_i,\ i = 1, \ldots, n, \tag{1}$$

where $\{X_i\}_{i=1}^n$ are p-dimensional i.i.d. covarite vectors with $Cov(X_1) = \Sigma$, β is a p-dimensional regression coefficient. The error sequence $\{e_i\}_{i=1}^n$ is i.i.d.r.v.s. and independent of predictors with a common unknown density $f(x)$,

$$E(e_i) = 0,\ \mathrm{Var}(e_i) = \sigma^2 < \infty,\quad i = 1, \ldots, n. \tag{2}$$

In the setting of $p > n$, it is often assumed that only a small number of predictors contribute to the response, which amounts to say the true model $M_0 = \{j : \beta_j \neq 0\}$ is

sparse. In this paper, we assume that the number of nonzero coefficients $s = \#M_0 = O(n^{\gamma})$ with $0 \leq \gamma < 1$, in other words, s is fixed or diverging at a mild rate. With sparsity and regularity conditions, many variable selection tools such as the LASSO, SCAD, SIS and Dantzig selector have lots of excellent properties about model selection consistency, such as sure screening property, model consistency, sign consistency and the oracle property. Especially, as a crucial criterion, the sure screening property ensures that we can pick out the true sparse model with probability tending to one. It is worth noting that there are also some spurious variables in the selected variable set other than these true variables, as mentioned in the introduction.

Considering that the naive two-stage method is imprecise to error density estimation in high-dimensional linear model. Here, we attempt to work out this problem by RCV method. Fan et al. (2012) showed that RCV method improved dramatically the performance of the naive two-stage procedure. The reason lies in that the two methods both require that the model selection procedure in the first stage has a sure screening property, but RCV method can remove or reduce the influence of spurious variables in the second stage, which motivates us to apply RCV method to kernel density estimation. Next, we will illustrate that how RCV method is used to obtain the residuals and give the kernel density estimator.

For any index set $\hat{M} = \{j_1, \ldots, j_{\hat{s}}\} \subset \{1, \ldots, p\}$, $(1 \leq j_1 < j_2 < \cdots < j_{\hat{s}} \leq p)$, we may as well denote $X_{\hat{M}} = (X_{1\hat{M}}, \ldots, X_{n\hat{M}})^{\mathrm{T}}$, where $X_{i\hat{M}} = (X_{ij_1}, \ldots, X_{ij_{\hat{s}}})^{\mathrm{T}}$. To elaborate the idea of RCV method, we consider a data set with sample size n, which is randomly split into two data sets $(y^{(1)}, X^{(1)})$ and $(y^{(2)}, X^{(2)})$ with one size is $n_1 = [bn]$ and another size is $n_2 = n - n_1$, where $0 < b < 1$ and $[\cdot]$ is integer part. Usually, we take $b = 1/2$ and denote $P^{(j)}_{\hat{M}} = X^{(j)}_{\hat{M}}(X^{(j)\mathrm{T}}_{\hat{M}} X^{(j)}_{\hat{M}})^{-1} X^{(j)\mathrm{T}}_{\hat{M}}$, $(j = 1, 2)$.

First, a variable selection tool is performed on $(y^{(1)}, X^{(1)})$ and let $\hat{M}_1$ denote the set of variables selected. Second, the error density function $f(x)$ is estimated by kernel method on the second data set $(y^{(2)}, X^{(2)}_{\hat{M}_1})$, namely

$$\hat{f}^{(1)}_{n_2}(x) = \frac{1}{n_2 h_n} \sum_{i=1}^{n_2} K\left(\frac{\hat{e}^{(1)}_i - x}{h_n}\right),$$

where $\hat{e}^{(1)} = (I_{n_2} - P^{(2)}_{\hat{M}_1}) y^{(2)} = (I_{n_2} - P^{(2)}_{\hat{M}_1}) e^{(2)}$. Similarly, we use the second data set $(y^{(2)}, X^{(2)})$ to select the set of important variables $\hat{M}_2$ and the first data set $(y^{(1)}, X^{(1)}_{\hat{M}_2})$ for estimation of $f(x)$, resulting in

$$\hat{f}^{(2)}_{n_1}(x) = \frac{1}{n_1 h_n} \sum_{i=1}^{n_1} K\left(\frac{\hat{e}^{(2)}_i - x}{h_n}\right),$$

where $\hat{e}^{(2)} = (I_{n_1} - P^{(1)}_{\hat{M}_2}) y^{(1)} = (I_{n_1} - P^{(1)}_{\hat{M}_2}) e^{(1)}$. Then, the final estimator is defined as

$$\hat{f}_n(x) = \frac{n_2}{n} \hat{f}^{(1)}_{n_2}(x) + \frac{n_1}{n} \hat{f}^{(2)}_{n_1}(x) = \frac{1}{n h_n} \sum_{i=1}^{n} K\left(\frac{\hat{e}_i - x}{h_n}\right), \quad \hat{e} = \left(\hat{e}^{(1)}, \hat{e}^{(2)}\right)^{\mathrm{T}}. \tag{3}$$

In the above procedure, the two halves of the data set are independent. Although some extra unimportant variables other than the important ones are selected by the first data set, these extra variables will play minor roles when we estimate the error density by using the second data set along with refitting since they are just some random unrelated variables over the second data set. Furthermore, Fan et al. (2012) indicated that even when some important variables are missed in the first stage of model selection, they still have a good chance being well approximated by the other variables selected in the first stage to reduce modeling biases, which is also significant to error density estimation. Thanks to the refitting in the second stage, the best linear approximation of those selected variables is used to obtain these residuals. In a word, RCV method provides a new way for us to accurately estimate error density in high-dimensional sparse linear model. The asymptotic properties of the kernel density estimator will be presented in the next section.

3 The asymptotics for the error density estimator

To obtain our main results, these technical conditions are sufficient to facilitate the proofs, although they may be not the weakest.

- $\mathbf{C}_0$. The model selection procedure has the sure screening property, i.e., $P\{M_0 \subset \hat{M}_j\} \to 1$ and $\hat{s}_j = \#\hat{M}_j = O(s) = O(n^\gamma)\ a.s.$ with $0 \le \gamma < 1$, where M_0 is the set of true variables and $\hat{M}_j$ denotes the the number of selected variables in the first stage, $j = 1, 2$.
- $\mathbf{C}_1$. The kernel $K(\cdot)$ is a bounded variation and symmetric probability density function; moreover, it has compact support and K' is bounded except finite jump points. The bandwidth h_n satisfies $\lim_{n\to\infty} h_n = 0$ and $\lim_{n\to\infty} nh_n/\log^4 n = \infty$.
- $\mathbf{C}_2$. The covariates $\{X_{1j}\}_{j=1}^p$ satisfy $\sup_{1\le j\le p} E[X_{1j}^{4k}(\log^+(X_{1j}^2))^2] \le C < \infty$ with $k > 1$, where $C > 0$ is a constant and $h^+(\cdot) = \max\{0, h(\cdot)\}$. Moreover, $\min_{1\le \#M\le cs} \lambda_{\min}(\frac{1}{n}X_M^{\mathrm{T}}X_M) \ge \frac{\lambda_0}{2} > 0$ with $M \subset \{1, \ldots, p\}$, where $c \ge 1$, λ_0 are constants and $\lambda_{\min}(\cdot)$ stands for the minimum eigenvalue of a matrix.
- $\mathbf{C}_3$. The random error e_1 satisfies the sub-exponential assumption. That is, there exist positive constants t_0 and C such that $E\exp\{t|e_1|\} \le C < \infty, \forall\, 0 < t < t_0$.

Remark 1 Condition $\mathbf{C}_0$ is a weak condition about variable selection procedure, as stated in Fan and Lv (2008) and Fan et al. (2012). Condition $\mathbf{C}_1$ is about the assumption of the kernel and bandwidth, which is common as well as in ordinary error density estimation model. Related moment assumption about the covariates $\{X_{1j}\}_{j=1}^p$ and the minimum eigenvalue assumption about $\frac{1}{n}X_M^{\mathrm{T}}X_M$ are imposed in condition $\mathbf{C}_2$, where M denotes the set of variables selected after dimensionality reduction. In some sense, the convergence rate of the kernel density estimator $\hat{f}_n(x)$ becomes faster with the increase of k, which could be checked by Remark 2 of Theorem 3. The minimum eigenvalue assumption is similar to Assumption 2 of Fan et al. (2012). In the setting of $\lambda_{\min}(\Sigma) \ge \lambda_0 > 0$, the minimum eigenvalue inequality about $\frac{1}{n}X_M^{\mathrm{T}}X_M$ holds if $\{X_{1j}\}_{j=1}^p$ are independent with zero mean and covariance σ_j^2 or $X_1 = (X_{11}, \ldots, X_{1p})^{\mathrm{T}} \sim N(0, \Sigma)$. The two mentioned cases could be proved in the

light of Marčenko and Pastur (1967) and Bai and Yin (1993). Condition $\mathbf{C}_3$ is a tail condition on the random error, which apparently holds when the error follows normal distribution or is bounded uniformly.

Then, we give the main theorems as following:

Theorem 1 *Suppose the conditions* $\mathbf{C}_0 - C_3$ *hold,* $\gamma < 1 - 1/k$. *If* $f(u)$ *is continuous at* u, *then* $|\hat{f}_n(u) - f(u)| = o(1)\ a.s.$

Theorem 2 *Suppose the conditions in Theorem 1 hold, and* $f(u)$ *is continuous uniformly about* u *over* R, *then* $\sup_{u\in R} |\hat{f}_n(u) - f(u)| = o(1)\ a.s.$

Theorem 3 *Suppose the conditions in Theorem 1 hold, and* $f(u)$ *is the first-order Lipstchz continuous about* u *over* R, *then*

$$\sup_{u\in R} |\hat{f}_n(u) - f(u)| = O\left(\frac{\log n}{\sqrt{nh_n}} + c_n + h_n\right)\ a.s.,$$

where $c_n = n^{-\frac{1}{2}(1-1/k-\gamma)} \log n$.

Remark 2 The uniform almost sure convergence rate of $\hat{f}_n(u) - f(u)$ about the sparsity parameter s is mainly embodied in γ due to the assumption $s = O(n^{\gamma})$. To obtain the fastest convergence rate, it is worth noting that there is a trade-off between the size of $\frac{\log n}{\sqrt{nh_n}}$ and h_n. Here, we take $h_n = O(n^{-1/3}(\log n)^{2/3})$, then $\sup_{u\in R} |\hat{f}_n(u) - f(u)| = O(c_n + n^{-1/3}(\log n)^{2/3})\ a.s.$ Next, our discussion is mainly centered on the relationship between the size of $(1/k + \gamma)$ and 1/3.

(i) If $1/k + \gamma \geq 1/3$, then $\sup_{u\in R} |\hat{f}_n(u) - f(u)| = O(c_n)\ a.s.$. Especially, when the random variable sequence $\{X_{1j}\}_{j=1}^{p}$ is bounded, $\sup_{u\in R} |\hat{f}_n(u) - f(u)| = O(n^{-\frac{1}{2}(1-\gamma)} \log n)$. At this case, the fastest uniform almost sure convergence rate of $\hat{f}_n(u) - f(u)$ is $O(n^{-1/3} \log n)$ for $u \in R$, which corresponds to $\gamma = 1/3$.

(ii) If $1/k + \gamma < 1/3$, then $\sup_{u\in R} |\hat{f}_n(u) - f(u)| = O(n^{-1/3}(\log n)^{2/3})\ a.s.$

Theorem 4 *Suppose the conditions in Theorem 1 hold, and* $f(u)$ *is the first-order Lipstchz continuous about* u *and* $f(u) > 0$, *if bandwidth still satisfies* $\lim_{n\to\infty} nh_n^3 = 0$, $\lim_{n\to\infty} \frac{c_n \log^2 n}{h_n} = 0$, *then*

$$(i).\ \sqrt{\frac{nh_n}{v}}[\hat{f}_n(u) - f(u)] \xrightarrow{d} N(0,1),\ v = f(u)\int K^2(y)\mathrm{d}y;$$

$$(ii).\ \limsup_n \sqrt{\frac{nh_n}{v \log\log n}}[\hat{f}_n(u) - f(u)] = \sqrt{2}\ a.s.,$$

where $\xrightarrow{d}$ *stands for the convergence in distribution.*

Theorem 5 *Denote* $T(f) = \int H(x)f(x)dx = E[H(e_1)]$, $T(\hat{f}_n) = \int H(x)\hat{f}_n(x)dx$, *where* $H(\cdot)$ *is a smooth function. Suppose the conditions in Theorem 1 hold, then*

(i). If $E\sup_{|z|\le\delta}|H'(e_i+z)| < +\infty$ *for some* $\delta > 0$, *then* $T(\hat{f}_n)$ *is a consistent estimator of* $T(f)$;

(ii). If $E\sup_{|z|\le\delta}|H''(e_i+z)| < +\infty$ *for some* $\delta > 0$, $\lim_{n\to\infty} nh_n^4 = 0$, $\gamma + 1/k < 1/2$ *and* $0 < Var(H(e_1)) < +\infty$, *then* $\sqrt{n}\,[T(\hat{f}_n) - T(f)] \xrightarrow{d} N(0, Var(H(e_1)))$.

Corollary 1 *If we take* $H(x) = x^2$ *and* $h_n = o(n^{-1/4})$, $\gamma + 1/k < 1/2$, $T(f)$, $T(\hat{f}_n)$ *are defined as Theorem 5, then*

$$\sqrt{n}\left[T(\hat{f}_n) - \sigma^2\right] \xrightarrow{d} N\left(0, E\left(e_1^4\right) - \sigma^4\right).$$

Remark 3 The similar result in Corollary 1 can be also seen in Fan et al. (2012). The condition $\gamma + 1/k < 1/2$ could be relaxed to $0 \le \gamma < 1$, and asymptotic normality still holds. In order to achieve the goal, we need to modify $\hat{f}_n(x)$ as $\hat{f}_n^*(x) = \frac{1}{(n-2\hat{s}_2)h_n}\sum_{i=1}^{n-[n/2]} K(\frac{\hat{e}_i^{(2)}-x}{h_n}) + \frac{1}{(n-2\hat{s}_1)h_n}\sum_{i=1}^{[n/2]} K(\frac{\hat{e}_i^{(1)}-x}{h_n})$, where $\hat{s}_1$ and $\hat{s}_2$, respectively, denote by the number of entry in the sets $\hat{M}_1$ and $\hat{M}_2$. The specific proof idea can be drawn from the proof of Theorem 5 in the Appendix.

4 Numeric studies

In this section, we assess the finite sample performances of the newly proposed procedures by Monte Carlo simulation. Moreover, a real data analysis is presented by our proposed procedure. In our numerical studies, we mainly report the results of RCV method with SIS to save space. Of course, LASSO, SCAD and MCP (or combining them with SIS) can be also used to make variable selection, see Fan et al. (2012). For simplicity, N-SIS and RCV-SIS, respectively, denote that SIS is employed in the model selection step for naive two-stage and refitted cross-validation methods; then, the Epanechnikov kernel function is applied to estimate the error density. Meanwhile, the bandwidth is selected by likelihood cross-validation (In Chapter 3, Section 4, Silverman 1986). All numerical studies were conducted by R code.

4.1 Simulation study

Considering that there is little work to study the error density estimation for high-dimensional linear model. This simulation is designed to compare the finite sample performances of naive two-stage method and refitted cross-validation method. As above stated, we employed SIS as our model selection tool. For SIS, the predetermined model size is always taken to be 5 in the null model as well as Fan et al. (2012). The cut value in SIS is needed to choose carefully in the sparse model, $[n/\log n]$ suggested by Fan and Lv (2008) may be used. In our simulation study, two examples were used to illustrate the performance of RCV-SIS method.

Example 1 We generated data from the following sparse linear model

$$Y = a(X_1 + X_2 + X_3) + e \tag{4}$$

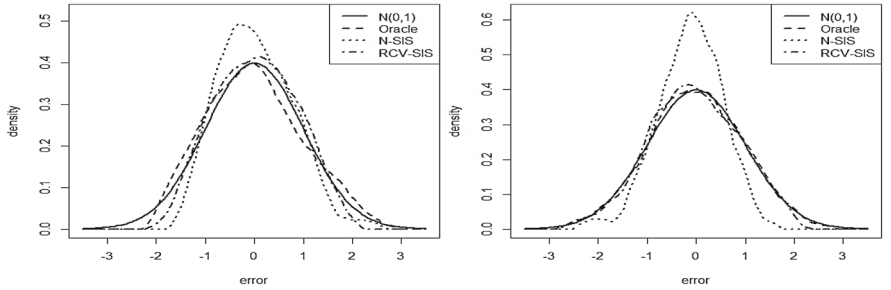

Fig. 1 Error density estimation for model (4) with $a = 0$. **a** $n = 50$, $p = 100$, **b** $n = 50$, $p = 1000$

where $e \sim N(0, 1)$, and $\{X_1, \ldots, X_p\} \sim N(0, \Sigma)$ with $\Sigma = \{\rho_{ij}\}_{i,j=1}^{p}$ where $\rho_{ii} = 1, \rho_{ij} = \rho$ for $i \neq j$. In order to examine the impact of signal-to-noise ratio (SNR) to error density estimation, we take $a = 0,\ 1/\sqrt{3}$ and $2\sqrt{3}$.

As a benchmark, the oracle or true error density is included in our simulation, where the oracle means that OLS method is used to estimate the parameters of the true model, and then kernel density method is applied to fit the error density. For $a = 0$, we only consider the case of $\rho = 0$ and let numbers of covariates vary from 100 to 1000 and the sample sizes equal 50 and 100. The corresponding results for $n = 50,\ p = 100, 1000$ are presented in Fig. 1, which shows that the over-fitted phenomenons caused by N-SIS method are becoming more and more serious with the increase of dimensionality p. To the contrary, the improved two-stage procedure RCV-SIS is comparable with the oracle and much better than N-SIS method, especially in the case of $p = 1000$.

For $a \neq 0$, our simulations are conducted in the setting of $\rho = 0.5$. For comparison, two groups of experiments are designed such as $n = 200$, $p = 2000$, $n = 300$, $p = 2000$ for $a = 2\sqrt{3}$ and $a = 1/\sqrt{3}$. These corresponding results are depicted in Figs. 2 and 3. We can clearly see that the N-SIS method and RCV-SIS methods both behave as well as the oracle from Fig. 2. However, the performance of N-SIS method is relatively poor in Fig. 3. Therefore, the two figures show that naive method depends on the SNR. In general, it performs better when the SNR is large. Furthermore, it is apparent that RCV-SIS method outperforms the N-SIS method and performs as well as the oracle procedure even if the signal is strong as well as the noise in Fig. 3. A conclusion could be drawn that RCV method is relatively insensitive to the size of SNR.

We may find that the performances of RCV-SIS are satisfactory from Example 1. Next, an additional simulation has been conducted to test the sensitivity of the RCV-SIS procedure to the final selected model size. Here, we take the case of $a = 1/\sqrt{3}, n = 200, p = 2000$ into consideration and assume the resulting model includes $\hat{s}$ predictors, and the simulation results are summarized in Fig. 4. The Figure shows that N-SIS method easily results in over-fitted phenomenon when many redundant variables are presented. To the contrary, RCV-SIS method is insensitive to model size and has a nice performance.

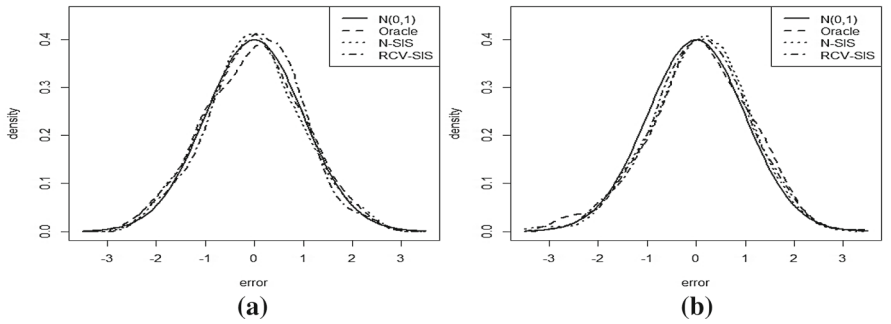

Fig. 2 Error density estimation for model (4) with $a = 2\sqrt{3}$. **a** $n = 200$, $p = 2000$, **b** $n = 300$, $p = 2000$

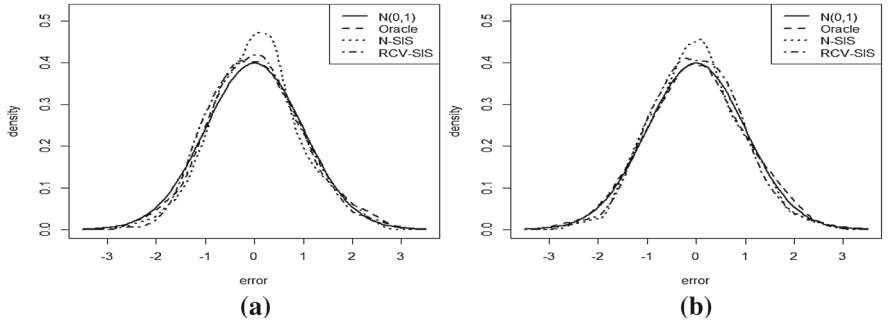

Fig. 3 Error density estimation for model (4) with $a = 1/\sqrt{3}$. **a** $n = 200$, $p = 2000$, **b** $n = 300$, $p = 2000$

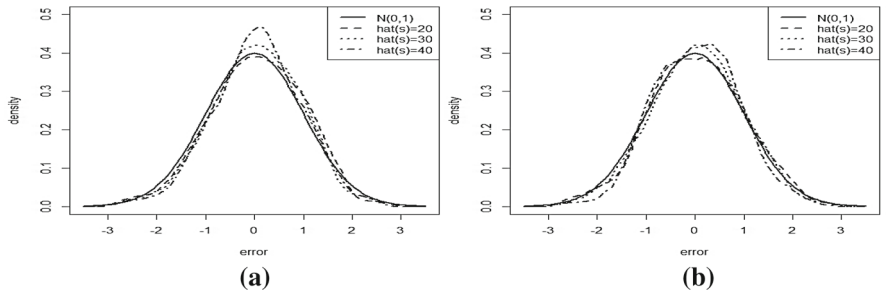

Fig. 4 Error estimation by using Naive and RCV methods with various $\hat{s}$. **a** N-SIS, **b** RCV-SIS

Example 2 We set the specific model as well as the Example 3 of Fan et al. (2012),

$$Y = 1.01X_1 - 0.06X_2 + 0.72X_3 + 1.55X_5 + 2.32X_7 - 0.36X_{11} + 3.75X_{13} - 2.04X_{17} - 0.13X_{19} + 0.61X_{23} + e,$$

where $e \sim N(0, 1)$, and $\{X_1, \ldots, X_p\} \sim N(0, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}, i, j = 1, 2, \ldots, p$.

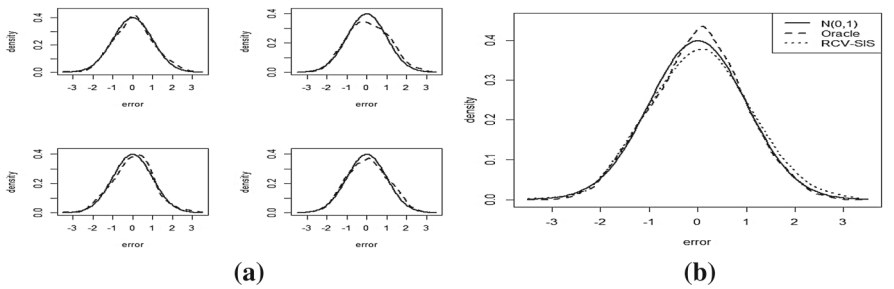

Fig. 5 Error density estimation. **a** Four arbitrary splits, **b** the average of four splits

In this example, a more realistic model with 10 important predictors is considered. Since some nonzero coefficients are very small, it is hard to say some method can chose all true variables in the stage of model selection. In this setting, two cases are considered for $n = 400,\ p = 1000$ and $n = 400,\ p = 10{,}000$. Here, we only present the simulation results of the latter given that the latter is more representative and challenging. The summarized results are shown in Fig. 5. In Fig. 5a, maybe there are some differences among different random splits. To conquer the splitting randomness, we adopted multi-split method (Meinshausen et al. 2009) and put the average result in Fig. 5b. In a small neighborhood of zero, the values of average estimator are slightly smaller than the real ones from Fig. 5b, which is corresponded to the results of Example 3 of Fan et al. (2012). Their results show that RCV methods slightly overestimate the variance when the sure screening condition is not satisfied. Hence, in some sense, Fig. 5b illustrates that RCV-SIS method still has a pretty good performance even in this case.

There is no doubt that the two examples show that RCV method greatly improves the accuracy of error density estimation in high-dimensional linear model. Furthermore, the following discussions are also very necessary to make our simulation more perfect.

RCV method relies on sample-splitting, performing variable selection and dimensionality reduction on one part of the data and OLS estimate on the remaining part. In our simulation, we randomly split the data set into two disjoint groups of equal size. However, the results may change if this split is chosen differently. An alternative to a single arbitrary split is dividing the sample repeatedly, then take the average result as the final estimator. The specific operation of this idea is embodied in Fig. 5 of Example 2. In Example 1, the simulation results are obtained by a single split. Of course, this idea of multi-split method could be also applied to Example 1. In some sense, it could be expected that the corresponding results will be more stable. About this point, more discussions could be seen in supplementary material.

Through the previous simulation results, we can clearly see that RCV method has an obvious advantage over naive method in the setting of $p > n$. Naturally, a problem is put forward about whether or not the performance of RCV method is good in the setting of $n > p$. The simulation results are depicted in Fig. 6. In this figure, it is easy to see that a over-fitted phenomenon is caused if we directly estimate the model parameters by OLS method, as we have stated in the part of introduction. However, the RCV-SIS method still performs well. Two reasons could be used to explain it. First,

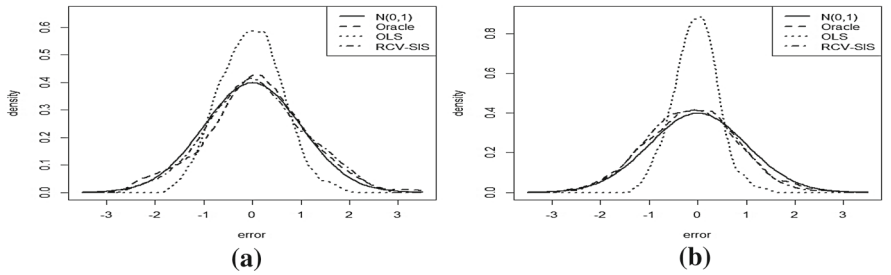

Fig. 6 Error density estimation for model (4) with $a = 1/\sqrt{3}$ and $\rho = 0.5$. **a** $n = 100$, $p = 50$, **b** $n = 100$, $p = 80$

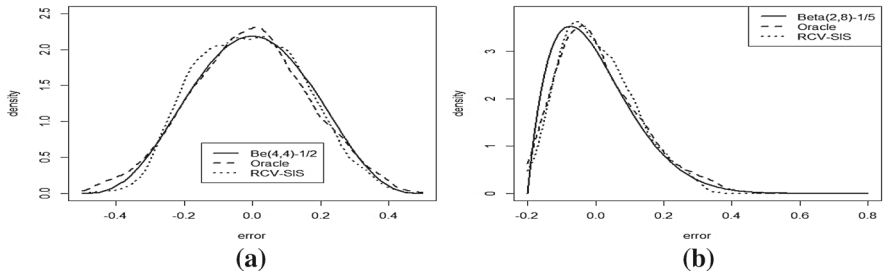

Fig. 7 Error density estimation for abnormal case. **a** $e \sim Be(4, 4) - 1/2$, **b** $e \sim Be(2, 8) - 1/5$

RCV method has a procedure of variable selection. Second, RCV method can reduce the influence of spurious variables by refitting.

To illustrate the impacts of different error distributions on the performance of RCV method, we still take model (4) as an example, where $a = 1/\sqrt{3}$, $n = 200$, $p = 2000$ and the covariates are jointly normal with equal correlation 0.5, and marginally $N(0, 1)$. Here, two different error terms are taken into consideration. Case (1) $e \sim Be(4, 4) - 1/2$; Case (2) $e \sim Be(2, 8) - 1/5$, where the error distribution is symmetric in Case (1), while the error distribution is biased at the right side in Case (2). The simulation results are presented in Fig. 7, we can see that RCV-SIS method performs slightly better than or as well as the oracle. In other words, RCV method has a certain stability for different error distributions.

4.2 Real data analysis

We use the data set reported in Scheetz et al. (2006) to illustrate the application of RCV-SIS method in error density estimation in high-dimensional linear model. Gene TRIM32 was recently found to cause Bardet–Biedl syndrome (Chiang et al. 2006), which is a genetically heterogeneous disease of multiple organ systems including the retina. Next, we make a brief description about the data set, more details could be saw in Huang et al. (2008). For this data set, 18976 probes are finally included to make a research and selected by two criteria. On the one hand, each probe should be sufficiently expressed; on the other hand, its expression values must be sufficiently

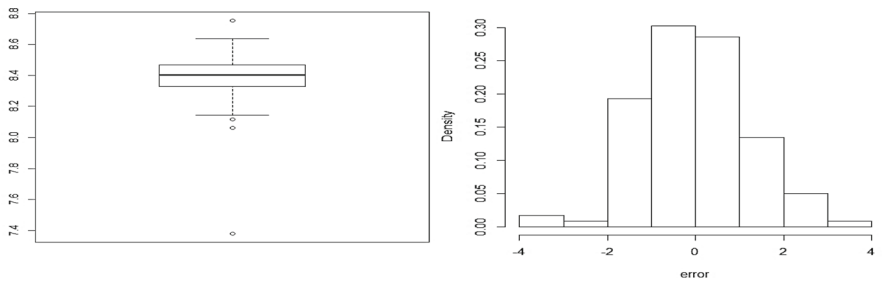

Fig. 8 Descriptive statistical analysis on the response and residuals. **a** Boxplot of Y, **b** histogram of the residuals

variable. The research shows that the probe from TRIM32 is 1389163_at. To find the probe among the remaining 18975 probes that are most related to TRIM32, Huang et al. (2008) set up a linear model to make an analysis. In their model, the sample size is 120 and the number of probes is 18975. It is expected that only a few genes are related to TRIM32. Therefore, this is a sparse, high-dimensional regression problem. In summary, the model based on the data set is $Y = X\beta + e$, where the response Y is the probe 1389163_at, the design matrix X is made up of the remaining 18975 probes, and e is the random error. In our paper, what we are concerned about is the error density estimation. Now we do as the following steps to estimate the error density:

We firstly conduct some exploratory data analysis on the response. Figure 8a depicts the boxplot of Y, where the distance the plot whiskers extend out from the box is set to be 1.5 interquartile range in order to detect the potential outliers. The boxplot shows that there is an extreme outlier in the response. Here, we eliminate it; then, the actual sample size n is 119 and the dimensionality p is 18975. Considering that $p \gg n$, then the following measures are taken,

(1) To estimate the error density by RCV-SIS method, we randomly split the sample into two independent data sets such as $(Y^{(1)}, X^{(1)})$ with size 60 and $(Y^{(2)}, X^{(2)})$ with size 59.
(2) For each data set, we follow this approach of Huang et al. (2008). First, select 3000 predictors with the largest variances, and then standardize the response and 3000 predictors so that they have mean zero and standard deviation 1.
(3) Compute the marginal correlation coefficients of the 3,000 predictors with the response and select the top 15 covariates with the largest correlation coefficients. That is, the resulting model includes $\hat{s} = 15$ predictors for each data set.
(4) The histogram of two parts of residuals is presented in Fig. 8b.

To illustrate the impact of choices of $\hat{s}$ on the performance of RCV-SIS method, we, respectively, select the top 10 or 25 covariates with the largest correlation coefficients instead of 15 at the step (3). As Fig. 9a shows that the change of $\hat{s}$ has a very little influence on the performance of RCV-SIS method. For $\hat{s} = 15$, four fitted density curves obtained by four arbitrary splits are shown in Fig. 9b. We may take one or their average as the final error density estimator.

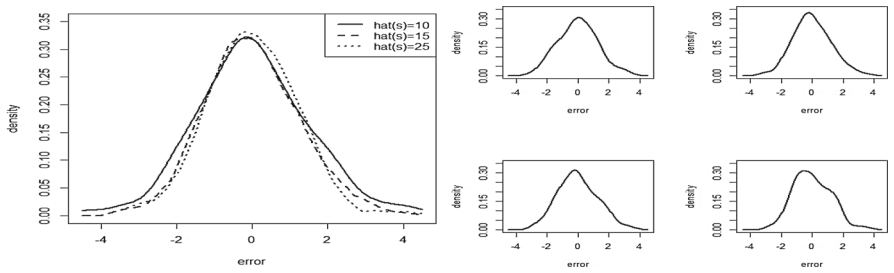

Fig. 9 Error density estimation. **a** $\hat{s} = 10, 15, 25$, **b** Four arbitrary splits for $\hat{s} = 15$

Supplementary material

Supplement to "Error Density Estimation in High-Dimensional Sparse Linear Model." This supplemental article includes four parts. Part(I) Some illustrations about $\mathscr{G}_n$; Part(II) Some explanations about multi-split method; Part(III) A proof about the order of $\max_{1\le i\le n} P_{ii}$; Part($IV$) An explanation about the minimum eigenvalue inequality in condition $\mathbf{C}_2$.

Acknowledgements This project was supported partly by the National Natural Science Foundation of China (Grant Nos. 11071022, 11231010, 11471223) and "Capacity Building for Sci-Tech Innovation-Fundamental Scientific Research Funds"(No. 025185305000/204). The authors thank the Editor, the AE and reviewers for their constructive comments, which have led to an improvement of the earlier version of this paper.

5 Appendix: Proofs of main results

Lemma 1 (Theorem 37 in Chapter 2, Pollard 1984) *For each n, let $\mathscr{F}_n$ be a permissible class of functions* (*Definition 1, in Appendix C,* Pollard 1984) *whose covering numbers* (*Definition 23, in Chapter 2,* Pollard 1984) *satisfy* $\sup_Q N_1(\epsilon, Q, \mathscr{F}_n) \le A\epsilon^{-W}$, $0 < \epsilon < 1$, *with constants A and W not depending on n. Let α_n be a non-increasing sequence of positive numbers for which $n\delta_n^2\alpha_n^2 \gg \log n$. If $|f| \le 1$, $\sqrt{P_0 f^2} \le \delta_n$, $\forall f \in \mathscr{F}_n$, then*

$$\sup_{\mathscr{F}_n} |P_n f - P_0 f| \ll \delta_n^2 \alpha_n \ a.s.,$$

where $P_0 f = \int f dP_0$, $P_n f = \int f dP_n = \frac{1}{n}\sum_{i=1}^n f(X_i)$.

For simplicity, we only prove the large sample properties of $\hat{f}_{n_1}(x)$ about $f(x)$. For $j = 1, 2$, we use n, $\hat{s}$ and $\hat{M}$ instead of n_j, $\hat{s}_j$ and $\hat{M}_j$, respectively, to make the proof look more concise and comfortable, where $\hat{s}_j$ and $\hat{M}_j$ is defined in condition $\mathbf{C_0}$. By the screening consistency, we have $\hat{s} = O(n^\gamma)$, $0 \le \gamma < 1$.

Lemma 2 *Suppose the assumptions* $\mathbf{C}_2$ *and* $\mathbf{C}_3$ *hold, then we have* $\max_{1\le i\le n} |\hat{e}_i - e_i| = o(c_n)$ *a.s., where* $c_n = n^{-\frac{1}{2}(1-\frac{1}{k}-\gamma)} \log n$ *with* $0 \le \gamma < 1 - \frac{1}{k}$.

Proof Denote $X_{\hat{M}} = (X_{1\hat{M}}, \ldots, X_{n\hat{M}})^{\mathrm{T}}$, where $X_{i\hat{M}} = (X_{ij_1}, \ldots, X_{ij_{\hat{s}}})^{\mathrm{T}}$ with $\hat{M} = \{j_1, \ldots, j_{\hat{s}}\}, (1 \le j_1 < \cdots < j_{\hat{s}} \le p)$. By the definition of P, we have $P_{ij} = X_{i\hat{M}}^{\mathrm{T}}(X_{\hat{M}}^{\mathrm{T}}X_{\hat{M}})^{-1}X_{j\hat{M}}$ and $\mathrm{Var}(\sum_{j=1}^{n} P_{ij}e_j|X_{\hat{M}}) = \sigma^2 P_{ii}$. To prove Lemma 2, the first is to compute the order of $\max_{1\le i\le n} P_{ii}$ (For more detailed proof, please see the supplementary material). By the definition of P_{ii} and the condition $\mathbf{C}_2$, we have

$$\max_{1\le i\le n} P_{ii} \le \frac{2\hat{s}}{n\lambda_0} \max_{1\le i\le n} \frac{1}{\hat{s}} \sum_{j\in\hat{M}} X_{ij}^2,$$

which is equivalent to compute the order of $\max_{1\le i\le n} \frac{1}{\hat{s}} \sum_{j\in\hat{M}} X_{ij}^2$. For the sake of simplicity, we may as well set $Z_{\mathrm{in}} = \frac{1}{\hat{s}} \sum_{j\in\hat{M}} X_{ij}^2$ due to $\hat{s} = O(n^{\gamma})$ with $0 \le \gamma < 1, i = 1, \ldots, n$. Since Z_{in} is i.i.d. random variable sequence, then

$$P\left(\max_{1\le i\le n} Z_{\mathrm{in}} > A_0 n^{1/k}\right) = 1 - P(Z_{\mathrm{in}} < A_0 n^{1/k})^n = 1 - [1 - P(Z_{\mathrm{in}} > A_0 n^{1/k})]^n,$$

where $A_0 > 2$. $\forall x > 0$, denote $h(x) = x^{2k}(\log^+(x))^2$, then we have

$$P\left(Z_{\mathrm{in}} > A_0 n^{1/k}\right) \le E\frac{h(Z_{\mathrm{in}})}{h\left(A_0 n^{1/k}\right)} \le \frac{k^2}{A_0^{2k}} \frac{\sup_{1\le j\le p} Eh\left(X_{1j}^2\right)}{n^2 \log^2 n}.$$

For fixed k and the condition $\sup_{1\le j\le p} E[X_{1j}^{4k}(\log^+(X_{1j}^2))^2] \le C < \infty$, then

$$\sum_{n=2}^{\infty} P\left(\max_{1\le i\le n} Z_{\mathrm{in}} > A_0 n^{1/k}\right) < \infty,$$

thereby we have $\max_{1\le i\le n} Z_{\mathrm{in}} = O(n^{1/k})\, a.s.$. By Cauchy's inequality, we have $\max_{i,j} |P_{ij}| \le \max_{1\le i,j\le n} \sqrt{P_{ii}P_{jj}} \le \max_{1\le i\le n} P_{ii}$. In addition, by the condition C_3, we have that $\max_{1\le i\le n} |e_i| \le c_0 \log n$ for some constant $c_0 > 0$. Denote $e_{1i} = e_i I\{|e_i| \le c_0 \log n\}$ and $e_{2i} = e_i I\{|e_i| > c_0 \log n\}$, then $E(e_{2j}) = o(n^{-3})$ and

$$\sum_{j=1}^{n} P_{ij}e_j = \sum_{j=1}^{n} P_{ij}[e_{1j} - E(e_{1j})] + \sum_{j=1}^{n} P_{ij}[e_{2j} - E(e_{2j})].$$

Since for any $\epsilon > 0$, $\{\max_{1\le i\le n} \sum_{j=1}^{n} P_{ij}e_{2j} > \epsilon c_n\} \subset \{\max_{1\le i\le n} |e_i| > c_0 \log n\}$, then $\max_{1\le i\le n} \sum_{j=1}^{n} P_{ij}e_{2j} = o(c_n)\ \ a.s.$. Therefore,

$$\max_{1\le i\le n} \sum_{j=1}^{n} P_{ij}[e_{2j} - E(e_{2j})] = o(c_n) \quad a.s. \tag{5}$$

Furthermore, for some constant $c_1 > 0$, by Bernstein's inequality, we have

$$\begin{aligned}
&P\left\{\max_{1\le i\le n}\left|\sum_{j=1}^{n} P_{ij}[e_{1j}-E(e_{1j})]\right| \ge nt, \max_{1\le i\le n} P_{ii} \le c_1\hat{s}n^{1/k-1}|X_{\hat{M}}\right\}\\
&\le \sum_{i=1}^{n} P\left\{\left|\sum_{j=1}^{n} P_{ij}[e_{1j}-E(e_{1j})]\right| \ge nt, \max_{1\le i\le n} P_{ii} \le c_1\hat{s}n^{1/k-1}|X_{\hat{M}}\right\}\\
&\le 2n\exp\left\{-\frac{n^2t^2}{2\sum_{j=1}^{n}\mathrm{Var}(P_{ij}e_j)+\frac{2}{3}c_0\max_{1\le i\le n}P_{ii}(\log n)nt}\right\}\\
&\quad I\left\{\max_{1\le i\le n} P_{ii} \le c_1\hat{s}n^{1/k-1}\right\}\\
&\le 2n\exp\left\{-\frac{n^2t^2}{2c_1\hat{s}n^{1/k-1}(\sigma^2+\frac{t}{3}c_0n\log n)}\right\}.
\end{aligned}$$

Let $t=\epsilon t_n$ with $\epsilon>0$, $t_n=\sqrt{\hat{s}}n^{-\alpha}\log n$, $\alpha=\frac{3}{2}-\frac{1}{2k}$, then

$$-\frac{n^2t^2}{2c_1\hat{s}n^{1/k-1}(\sigma^2+\frac{t}{3}c_0n\log n)}=\frac{-\epsilon^2\log^2 n}{2c_1(\sigma^2+\frac{\epsilon c_0}{3}c_n\log n)}\le-\frac{\epsilon^2}{4c_1\sigma^2}\log^2 n\le -3\log n,$$

as n is large enough. It can be derived by the Borel–Cantelli lemma that

$$\max_{1\le i\le n}\sum_{j=1}^{n} P_{ij}[e_{1j}-E(e_{1j})]=o(nt_n)=o(c_n)\quad a.s. \tag{6}$$

Then, we can easily know $\max_{1\le i\le n}|\hat{e}_i-e_i|=o(c_n)$ $a.s.$ by (5) and (6). □

Lemma 3 *Suppose the assumptions* $\mathbf{C}_0-C_3$ *hold and* $\gamma<1-1/k$, *then we have*

(i). If $f(\cdot)$ *is continuous at* u, *then* $\hat{f}_n(u)-f_n(u)=o(1)$ $a.s.$
(ii). If $f(\cdot)$ *is continuous uniformly, then*

$$\sup_u|\hat{f}_n(u)-f_n(u)|\le o\left(\frac{\log n}{\sqrt{nh_n}}\left(1\wedge\sqrt{\frac{c_n}{h_n}}\right)\right)+2\sup_u[I_{n+}(u)+I_{n-}(u)]\ a.s.,$$

where $I_{n+}(u)=\int K(y)|f(u+Cc_n+h_ny)-f(u+h_ny)|\mathrm{d}y$ *and* $I_{n-}(u)=\int K(y)|f(u-Cc_n+h_ny)-f(u+h_ny)|\mathrm{d}y$ *for some constant* $C>0$.

Proof Since $K(\cdot)$ is a bounded variation function, then K can be written as $K=K_1-K_2$ which K_1 and K_2 are two monotonically increasing functions. By the definitions of $\hat{f}_n(u)$ and $f_n(u)$, we have

$$\begin{aligned}\hat{f}_n(u) - f_n(u) &= \frac{1}{nh_n}\left[\sum_{i=1}^{n} K\left(\frac{\hat{e}_i - u}{h_n}\right) - \sum_{i=1}^{n} K\left(\frac{e_i - u}{h_n}\right)\right] \\ &= \frac{1}{nh_n}\left[\sum_{i=1}^{n} K_1\left(\frac{\hat{e}_i - u}{h_n}\right) - \sum_{i=1}^{n} K_1\left(\frac{e_i - u}{h_n}\right)\right] \\ &\quad + \frac{1}{nh_n}\left[\sum_{i=1}^{n} K_2\left(\frac{\hat{e}_i - u}{h_n}\right) - \sum_{i=1}^{n} K_2\left(\frac{e_i - u}{h_n}\right)\right] \\ &= \Delta_{1n}(u) + \Delta_{2n}(u). \end{aligned} \tag{7}$$

On the set $\{\hat{e}_i | \max|\hat{e}_i - e_i| \le Cc_n\}$ with constant $C > 0$, it can be derived that $I_{2n}(u) \le \Delta_{1n}(u) \le I_{1n}(u)$ due to the fact that K_1 is a monotonically increasing function, where

$$I_{1n}(u) = \frac{1}{nh_n}\sum_{i=1}^{n}\left[K_1\left(\frac{e_i + Cc_n - u}{h_n}\right) - K_1\left(\frac{e_i - u}{h_n}\right)\right],$$
$$I_{2n}(u) = \frac{1}{nh_n}\sum_{i=1}^{n}\left[K_1\left(\frac{e_i - Cc_n - u}{h_n}\right) - K_1\left(\frac{e_i - u}{h_n}\right)\right].$$

Let $\mathscr{G}_n =: \{ g_{nu} = K_1(\frac{(e+Cc_n)-u}{h_n}) - K_1(\frac{e-u}{h_n}) : u \in R\}$ (For more details about $\mathscr{G}_n$, please refer to supplementary material), then $\mathscr{G}_n$ is a class of permissible functions with a polynomial discriminant and

$$I_{1n}(u) = \frac{1}{h_n}[P_n g_{nu} - P_0 g_{nu}] + \frac{1}{h_n} P_0 g_{nu}.$$

Since K has compact support, we suppose that K_j has compact support $[-M, M]$ with $M > 0$ and K_j' is bounded except one jump point without loss of generality. For fixed u, we have

$$\begin{aligned}\frac{|P_0 g_{nu}|}{h_n} &= \frac{1}{h_n}\left|E\left[K_1\left(\frac{e - u + Cc_n}{h_n}\right) - K_1\left(\frac{e-u}{h_n}\right)\right]\right| \\ &= \frac{1}{h_n}\left|\int K_1\left(\frac{x - u + Cc_n}{h_n}\right) f(x)\mathrm{d}x - \int K_1\left(\frac{x-u}{h_n}\right) f(x)\mathrm{d}x\right| \\ &= \left|\int K_1(y)[f(u - Cc_n + h_n y) - f(u + h_n y)]\mathrm{d}y\right| \\ &\le \int K(y)|f(u - Cc_n + h_n y) - f(u + h_n y)|\mathrm{d}y = I_{n-}(u), \end{aligned}$$

and

$$P_0 g_{n,u}^2 = \int \left[K_1\left(\frac{x-u+Cc_n}{h_n}\right) - K_1\left(\frac{x-u}{h_n}\right)\right]^2 f(x)\mathrm{d}x$$
$$= h_n \int \left[K_1\left(y + C\frac{c_n}{h_n}\right) - K_1(y)\right]^2 f(u+h_n y)\mathrm{d}y.$$

(i). If $f(\cdot)$ is continuous at u and $h_n = o(1)$, then $f(u+h_n y) \le f(u)+1$ for $|y| \le M + Cc_n/h_n$, and $P_0 g_{n,u}^2 \le 4(f(u)+1)MC_1^2 h_n = O(h_n)$ for $c_n/h_n \ge 1$. If $c_n/h_n < 1$, let y_1 be a jump point of K_1 in $[-M, M]$, $B =: [y_1 - Cc_n/h_n, y_1 + Cc_n/h_n]$ and $B^c =: [-M - Cc_n/h_n, M + Cc_n/h_n] - B$, then

$$P_0 g_{n,u}^2 \le h_n \left[\left(\int_B + \int_{B^c}\right) \left[K_1\left(y + C\frac{c_n}{h_n}\right) - K_1(y)\right]^2 f(u+h_n y)\mathrm{d}y \right]$$
$$\le h_n \left[2C_1^2(f(u)+1)\int_B \mathrm{d}y + C_2(f(u)+1)(c_n/h_n)^2 \int_{B^c} \mathrm{d}y \right]$$
$$= (f(u)+1)O(c_n),$$

where $C_1 = \sup K, C_2 > 0$ is some constant. Thus, $P_0 g_{n,u}^2 \le (f(u)+1)O(c_n \wedge h_n)$ and $I_{n-}(u) = o(1)$. By using Bernstein's inequality, we obtain $|P_n g_{nu} - P_0 g_{nu}| = o(h_n)\ a.s.$ and $|I_{1n}(u)| \le |P_n g_{nu} - P_0 g_{nu}|/h_n + I_{n-}(u) = o(1)\ \ a.s.$ Similarly, we also have $I_{2n}(u) = o(1)\ a.s.$. It means that $\Delta_{1n}(u) = o(1)\ a.s.$. By the same derivation way, $\Delta_{2n}(u) = o(1)\ a.s.$ still holds. Therefore, we have $\hat{f}_n(u) - f_n(u) = o(1)\ \ a.s.$ by (7). □

(ii). If $f(\cdot)$ is continuous uniformly, then $f(u)$ is bounded and $\sup_u P_0 g_{n,u}^2 \le 4\sup_u (f(u)+1)MC_1^2(c_n \wedge h_n) = O(c_n \wedge h_n)$. By Lemma 36 (ii) in Chapter 2, Pollard (1984), the covering numbers of $\mathscr{G}_n$ satisfy $\sup_Q N_1(\epsilon, Q, \mathscr{G}_n) \le A\epsilon^{-W}$, $0 < \epsilon < 1$, where constants A and W are not depending on n, and $\sup_u |g_{nu}| \le C_1$. Denote $\alpha_n = \frac{\log n}{\sqrt{n}\delta_n}$, $\delta_n^2 = O(c_n \wedge h_n)$, then the conditions of Lemma 1 hold. Furthermore, we have

$$\sup_u |P_n g_{nu} - P_0 g_{nu}| = o\left(\delta_n^2 \alpha_n\right)\ \ a.s.$$

Therefore, $\sup_u |I_{1n}(u)| \le o(\frac{\delta_n^2 \alpha_n}{h_n}) + \sup_u I_{n-}(u) = o(\frac{\log n}{\sqrt{nh_n}}(1 \wedge \sqrt{\frac{c_n}{h_n}})) + \sup_u I_{n-}(u)$ a.s. Similarly, we have $\sup_u |I_{2n}(u)| \le o(\frac{\log n}{\sqrt{nh_n}}) + \sup_u I_{n+}(u)\ a.s.$ and

$$\sup_u |\Delta_{1n}(u)| \le o\left(\frac{\log n}{\sqrt{nh_n}}\left(1 \wedge \sqrt{\frac{c_n}{h_n}}\right)\right) + \sup_u [I_{n+}(u) + I_{n-}(u)]\ a.s. \quad (8)$$

Similar to the proof of Eq. (8), we also have $\sup_u |\Delta_{2n}(u)| \le o(\frac{\log n}{\sqrt{nh_n}}(1 \wedge \sqrt{\frac{c_n}{h_n}})) + \sup_u [I_{n+}(u) + I_{n-}(u)]\ a.s.$. We know that $\sup_u |\hat{f}_n(u) - f_n(u)|$ can be dominated

by $\sup_u |\Delta_{1n}(u)| + \sup_u |\Delta_{2n}(u)|$ almost surely by (7). Thus, we have

$$\sup_u |\hat{f}_n(u) - f_n(u)| \le o\left(\frac{\log n}{\sqrt{nh_n}}\left(1 \wedge \sqrt{\frac{c_n}{h_n}}\right)\right) + 2\sup_u[I_{n+}(u) + I_{n-}(u)] \ \ a.s.$$

This completes the proof of Lemma 3. □

Proof of Theorem 1. Note that

$$|\hat{f}_n(u) - f(u)| \le |\hat{f}_n(u) - f_n(u)| + |f_n(u) - f(u)|.$$

By condition $\mathbf{C}_1$ and Lemma 3 (i), we have $\hat{f}_n(u) - f_n(u) = o(1)\ a.s.$ due to the continuity of f at u. For $f_n(u) - f(u) = o(1)\ a.s.$, please refer to pages 35–36 of Pollard (1984). Therefore, $|\hat{f}_n(u) - f(u)| = o(1)$ holds a.s. □

Proof of Theorem 2. According to triangle inequality,

$$\sup_u |\hat{f}_n(u) - f(u)| \le \sup_u |\hat{f}_n(u) - f_n(u)| + \sup_u |f_n(u) - f(u)|.$$

From Lemma 3 (ii),

$$\sup_u |\hat{f}_n(u) - f_n(u)| \le o\left(\frac{\log n}{\sqrt{nh_n}}\right) + 2\sup_u[I_{n+}(u) + I_{n-}(u)].$$

Due to the assumption that $f(u)$ is continuous uniformly about u and conditions $\mathbf{C}_1 - C_3$, we have $\sup_u I_{n-}(u) = \sup_u \int K(y)|f(u - Cc_n + h_n y) - f(u + h_n y)|\mathrm{d}y = o(1)$ and $\sup_u I_{n+}(u) = o(1)$. By Lemma 3 (ii), we immediately have $\sup_u |\hat{f}_n(u) - f_n(u)| = o(1)\ a.s.$. Moreover, $\sup_u |f_n(u) - f(u)| = o(1)\ a.s.$ holds according to condition $\mathbf{C}_1$. Therefore, we obtain $\sup_u |\hat{f}_n(u) - f(u)| = o(1)\ a.s.$ □

Proof of Theorem 3. According to Theorem 2 and Lemma 3 (ii),

$$\sup_u |\hat{f}_n(u) - f(u)| \le o\left(\frac{\log n}{\sqrt{nh_n}}\right) + 2\sup_u[I_{n+}(u) + I_{n-}(u)] + \sup_u |f_n(u) - f(u)|.$$

According to $Lipstchz$ condition of f, we have that $\sup_u |Ef_n(u) - f(u)| = O(h_n)$ and $\sup_u |f_n(u) - Ef_n(u)| = o(\frac{\log n}{\sqrt{nh_n}})\ a.s.$ Therefore, it is easy to know by the triangle inequality,

$$\begin{aligned}\sup_u |f_n(u) - f(u)| &\le \sup_u |f_n(u) - Ef_n(u)| + \sup_u |Ef_n(u) - f(u)| \\ &= o\left(\frac{\log n}{\sqrt{nh_n}}\right) + O(h_n)\ a.s.\end{aligned}$$

Since $f(u)$ satisfies the first-order $Lipstchz$ condition, then

$$I_{n-}(u) = \left|\int K(y)|f(u - Cc_n + h_n y) - f(u + h_n y)\right|\mathrm{d}y = O(c_n).$$

Similarly, we have $I_{n+}(u) = O(c_n)$. Thus, it can be derived that

$$\sup_u |\hat{f}_n(u) - f(u)| = O(\frac{\log n}{\sqrt{nh_n}} + c_n + h_n) \ a.s.$$

□

Lemma 4 *Assume that the conditions* $\mathbf{C}_1 - C_3$ *hold,* $f(\cdot)$ *satisfies the first-order Lipstchz condition and* $\lim_{n\to\infty} nh_n^3 = 0$*, then for fixed* $u \in R$ *and* $f(u) > 0$*, we have*

$$\sqrt{\frac{nh_n}{v}}[f_n(u) - f(u)] \xrightarrow{d} N(0,1), \ \ v = f(u)\int K^2(y)\mathrm{d}y.$$

□

Proof of Theorem 4. (i). Since $f(\cdot)$ satisfies the first-order $Lipstchz$ condition, then $\sup_u I_{n+}(u) + \sup_u I_{n-}(u) = O(c_n)$. By Lemma 3 and $c_n = o(h_n/\log^2 n)$, we know

$$\sqrt{nh_n}\,|\hat{f}_n(u) - f_n(u)| \ = \sqrt{nh_n}\left[o\left(\sqrt{\frac{c_n}{h_n}}\frac{\log n}{\sqrt{nh_n}} \wedge \frac{\log n}{\sqrt{nh_n}}\right) + O(c_n)\right] = o(1)\ a.s.$$

and

$$\begin{aligned}\sqrt{\frac{nh_n}{v}}\,[\hat{f}_n(u) - f(u)] &= \sqrt{\frac{nh_n}{v}}[\hat{f}_n(u) - f_n(u)] + \sqrt{\frac{nh_n}{v}}[f_n(u) - f(u)] \\ &= \sqrt{\frac{nh_n}{v}}[f_n(u) - f(u)] + o(1)\ a.s.\end{aligned}$$

Due to condition $\lim_{n\to\infty} nh_n^3 = 0$ and Lemma 4, then we have

$$\sqrt{\frac{nh_n}{v}}[\hat{f}_n(u) - f(u)] \xrightarrow{d} N(0,1).$$

□

(ii). By $c_n = o(h_n/\log^2 n)$, we have $\sqrt{nh_n}\,[\hat{f}_n(u) - f_n(u)] = o(1)\ a.s.$ and $Ef_n(u) - f(u) = O(h_n)$. In addition, by employing the law of the iterated logarithm (Theorem 2, Hall 1981), we have

$$\limsup_{n\to\infty} \sqrt{\frac{nh_n}{V\log\log n}}\,[f_n(u) - Ef_n(u)] = \sqrt{2}\ a.s. \tag{9}$$

Furthermore, by condition $\lim_{n\to\infty} nh_n^3 = 0$, it can be derived that

$$\sqrt{\frac{nh_n}{v\log\log n}}\,[\hat{f}_n(u) - f(u)] = \sqrt{\frac{nh_n}{v\log\log n}}\,[f_n(u) - Ef_n(u)] + o(1)\ a.s. \tag{10}$$

Finally, we have

$$\limsup_{n\to\infty}\sqrt{\frac{nh_n}{v\log\log n}}[\hat{f}_n(u)-f(u)] = \sqrt{2}\ a.s.$$

This completes the proof of Theorem 4 by (9) and (10).

□

Proof of Theorem 5. (i)

$$\begin{aligned}
|T(f_n)-T(f)| &= \left|\int H(u)f_n(u)du - EH(e_1)\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n}\int [H(e_i+h_ny)-H(e_i)]K(y)\mathrm{d}y\right| + \left|\frac{1}{n}\sum_{i=1}^{n}H(e_i)-EH(e)\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n}\int H'(e_i+\theta_ih_ny)h_nyK(y)\mathrm{d}y\right| + o(1) \\
&\le \frac{1}{n}\sum_{i=1}^{n}\sup_{|z|\le\delta}|H'(e_i+z)|h_n\int|y|K(y)\mathrm{d}y + o(1) = o(1)\ \ a.s.
\end{aligned}$$

Next, we make some explanations about why the "$\le$" holds. For given i, $H(e_i+h_ny)-H(e_i) = H'(e_i+\theta_ih_ny)h_ny$ by Taylor's expansion of $H(\cdot)$ at e_i with $|\theta_i|\le 1$. Since $h_n\to 0$, there exists a constant $\delta>0$ such that $|\theta_ih_ny|\le Mh_n\le\delta$. Furthermore, we have $|H(e_i+h_ny)-H(e_i)|\le \sup_{|z|\le\delta}|H'(e_i+z)|h_n|y|$ as long as n is large enough. To prove $T(\hat{f}_n)$ is a consistent estimator of $T(f)$, it just needs to prove $T(\hat{f}_n)-T(f_n)=o(1)\ \ a.s.$. By Taylor's expansion,

$$\begin{aligned}
T(\hat{f}_n)-T(f_n) &= \frac{1}{n}\sum_{i=1}^{n}\int [H(\hat{e}_i+h_ny)-H(e_i+h_ny)]K(y)\mathrm{d}y \\
&= \frac{1}{n}\sum_{i=1}^{n}\int \left[H'(e_i+h_ny+\theta_i(\hat{e}_i-e_i))K(y)\mathrm{d}y(\hat{e}_i-e_i)\right] \\
&\quad I\{\max_i|\hat{e}_i-e_i|\le c_n\} + o(1) \\
&\le \frac{1}{n}\sum_{i=1}^{n}\sup_{|z|\le\delta}|H'(e_i+z)|c_n + o(1) = o(1)\ \ a.s.
\end{aligned}$$

□

(ii). By Taylor's expansion, we have

$$\begin{aligned}
&\left|\sqrt{n}[T(f_n)-T(f)] - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(H(e_i)-EH(e_1))\right| \\
&= \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int\left[H'(e_i)h_ny+\frac{1}{2}H''(e_i+\theta_ih_ny)h_n^2y^2\right]K(y)\mathrm{d}y\right|
\end{aligned}$$

$$= \left| \frac{h_n^2}{2\sqrt{n}} \sum_{i=1}^{n} \int H''(e_i + \theta_i h_n y) y^2 K(y) \mathrm{d}y \right|$$
$$\leq \frac{\sqrt{n} h_n^2}{2n} \sum_{i=1}^{n} \sup_{|z| \leq \delta} |H''(e_i + z)| \int y^2 K(y) \mathrm{d}y = o(1) \ \ a.s.,$$

and

$$\begin{aligned}
\sqrt{n} \left| T(\hat{f}_n) - T(f_n) \right| &\leq \sqrt{n} |T(\hat{f}_n) - T(f_n)| I\{\max_i |\hat{e}_i - e_i| \leq c_n\} + o(1) \\
&= \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [H(\hat{e}_i) - H(e_i)] \right| I\{\max_i |\hat{e}_i - e_i| \leq c_n\} \\
&\quad + \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [H''(\hat{e}_i + \theta_{1i} h_n y) - H''(e_i + \theta_{2i} h_n y)] h_n^2 \int y^2 K(y) \mathrm{d}y \right| \\
&\quad I\{\max_i |\hat{e}_i - e_i| \leq c_n\} + o(1) \\
&\leq \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} H'(e_i)(\hat{e}_i - e_i) \right| + \frac{1}{2\sqrt{n}} \sum_{i=1}^{n} \sup_{|z| \leq \delta} |H''(e_i + z)| c_n^2 \\
&\quad + \frac{2h_n^2}{\sqrt{n}} \sum_{i=1}^{n} \sup_{|z| \leq \delta} |H''(e_i + z)| \int y^2 K(y) \mathrm{d}y + o(1) \\
&\leq \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} H'(e_i)(\hat{e}_i - e_i) \right| + o(1) \\
&\leq \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{n} P_{ij} [H'(e_i) - E(H'(e_i))] e_j \right| + \left| \frac{E(H'(e_1)}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{n} P_{ij} e_j \right| + o(1) \\
&= \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{n} P_{ij} [H'(e_i) - E(H'(e_i))] e_j \right| + O_p(\sqrt{\hat{s}/n}) + o(1). \qquad (11)
\end{aligned}$$

Let $e_i^* = H'(e_i) - E(H'(e_i))$, then

$$\begin{aligned}
E\left\{ \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{n} P_{ij} e_i^* e_j \right]^2 | X_{\hat{M}} \right\} &= \frac{1}{n} \sum_{i=1}^{n} P_{ii}^2 E\left(e_1^{*2} e_1^2\right) \\
&\quad + \frac{1}{n} \sum_{i \neq j} P_{ii} P_{jj} E\left(e_1^* e_1\right) E\left(e_2^* e_2\right) \\
&\quad + \frac{2}{n} \sum_{i \neq j} P_{ij}^2 E\left[e_1^{*2} e_2^2\right] \\
&= O(\hat{s}/n) + O(\hat{s}^2/n) = O(\hat{s}^2/n) \ \ a.s.
\end{aligned}$$

It can be derived from (11) and condition $\gamma + 1/k < 1/2$ that $\sqrt{n}[T(\hat{f}_n) - T(f_n)] = o_p(1)$. Therefore,

$$\begin{aligned}\sqrt{n}[T(\hat{f}_n) - T(f)] &= \sqrt{n}[T(\hat{f}_n) - T(f_n)] + \sqrt{n}[T(f_n) - T(f)] \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (H(e_i) - EH(e_1)) + o_p(1) \xrightarrow{d} N(0, \mathrm{Var}(H(e_1))).\end{aligned}$$

This completes the proof of Theorem 5. □

References

Bai, Z., Yin, Y. (1993). Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix. *Annals of Probability*, *21*(3), 1275–1294.

Candes, E., Tao, T. (2007). The Danzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, *35*, 2313–2351.

Chai, G., Li, Z. (1993). Asymptotic theory for estimation of error distribution in linear model. *Science in China: Series A*, *36*, 408–419.

Cheng, F. (2005). Asymptotic distributions of error density and distribution function estimators in nonparametric regression. *Journal of Statistical Planning and Inference*, *128*, 327–349.

Chiang, A. P., Beck, J. S., Yen, H. J., Tayeh, M. K., Scheetz, T. E., Swiderski, R., Nishimura, D., Braun, T. A., Kim, K. Y., Huang, J., Elbedour, K., Carmi, R., Slusarski, D. C., Casavant, T. L., Stone, E. M., Sheffield, V. C. (2006). Homozygosity mapping with SNP arrays identifies a novel gene for Bardet–Biedl syndrome (BBS10). *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 6287–6292.

Cui, H., Li, R., Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, *110*, 630–641.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Fan, J., Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society*: *Series B*, *70*, 849–911.

Fan, J., Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*, 101–148.

Fan, J., Guo, S., Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society*: *Series B*, *74*, 37–65.

Hall, P. (1981). Laws of the iterated logarithm for nonparametric density estimators. *Probability Theory and Related Fields*, *56*, 47–61.

Huang, J., Ma, S., Zhang, C. H. (2008). Adaptive lasso for sparse high dimensional regression. *Statistica Sinica*, *18*, 1603–1618.

Liang, H., Hardle, W. (1999). Large sample theory of the estimation of the error distribution for a semiparametric model. *Communication in Statistics Theory and Methods*, *28*, 2025–2036.

Li, R., Zhong, W., Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, *107*, 1129–1139.

Marčenko, V. A., Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, *1*, 507–536.

Meinshausen, N., Meier, L., Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, *104*, 1671–1681.

Pollard, D. (1984). *Convergence of stochastic processes*. New York: Springer.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, *25*, 303–325.

Scheetz, T. E., Kim, K. Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudt- son, K. L., Dorrance, A. M., Dibona, G. F., Huang, J., Casavant, T. L. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 14429–14434.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*, 267–288.
Yang, Y. (1997). Large sample properties of estimation of the error distribution in nonparametric regression. *Acta Scientiarum Naturalium Universitatis Pekinensis*, *33*, 298–304.
Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, *38*, 894–942.
Zhong, W. (2014). Robust sure independence screening for ultrahigh dimensional non-normal data. *Acta Mathematical Sinica*, *30*, 1885–1896.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.