Error Density Estimation in High-Dimensional Sparse Linear Model

Feng Zou* and Hengjian Cui*

* Department of Statistics, Capital Normal University, Beijing, China

SUPPLEMENTARY MATERIAL

(I) Some explanations about $\mathscr{G}_n$

Here, we give some properties of $\mathscr{G}_n$, where

$$\mathscr{G}_n =: \{\, g_{nu} = K_1(\frac{(e+Cc_n)-u}{h_n}) - K_1(\frac{e-u}{h_n}) : u \in R\}.$$

We may as well set $\mathscr{A}_n =: \{g_{1nu} = K_1(\frac{(e+Cc_n)-u}{h_n}) : u \in R\}$, $\mathscr{B}_n =: \{g_{2nu} = K_1(\frac{e-u}{h_n}) : u \in R\}$, then $\mathscr{G}_n = \mathscr{A}_n - \mathscr{B}_n$. Here, we make an explanation about $\mathscr{A}_n$, another is the same. For fixed u, g_{1nu} is a function. When u varies in the real field, then $\mathscr{A}_n$ is a class of functions.

(i) According to the definition of $\mathscr{A}_n$, $\mathscr{B}_n$ and $\mathscr{G}_n$, we could know that they all satisfy the conditions of a permissible class of functions (Definition 1, in Appendix C, Pollard(1984)). Therefore, $\mathscr{A}_n$, $\mathscr{B}_n$ and $\mathscr{G}_n$ are all permissible classes of functions.

(ii) K_1 is monotonically increasing and bounded on the interval $[-M, M]$. By question 27, page 42 of Pollard(1984), we know that the class of graphs (please refer to line 11, page 27 of Pollard(1984)) of functions in $\mathscr{A}_n$ has polynomial discrimination (Definition 13, in Chapter 2, Pollard(1984)). Say simply, $\mathscr{A}_n$ and $\mathscr{B}_n$ are both permissible classes of functions with polynomial discrimination. By Lemma 15 in Chapter 2, Pollard(1984), we can imply that $\mathscr{G}_n$ is also a permissible class of functions with polynomial discrimination.

(iii) By Lemma 36 (ii), page 34 of Pollard(1984), the covering number $N_1(\varepsilon, Q, \mathscr{F}_n)$ (Definition 23, in Chapter 2, Pollard(1984)) satisfies $\sup_Q N_1(\epsilon, Q, \mathscr{F}_n) \leq A\epsilon^{-W}, 0 < \epsilon < 1$, with constants A and W not depending on n.

Since $K(\cdot)$ is bounded on the compact support, there exists a constant C such that $\sup_u |g_{n,u}| \leq C$. If we take $\alpha_n = \frac{\log n}{\sqrt{n\delta_n}}$, $\delta_n^2 = O(c_n \wedge h_n)$, then the

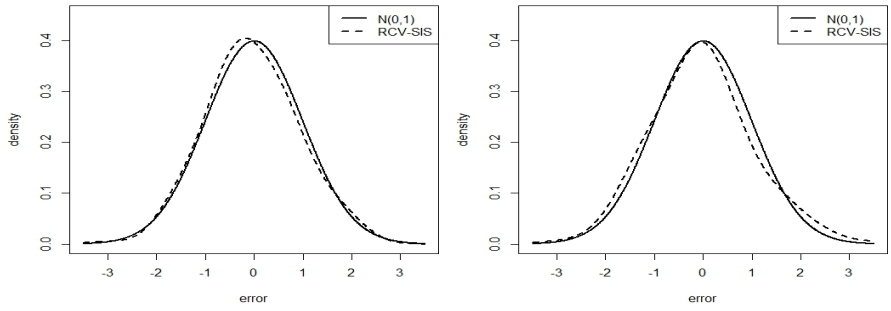
conditions of Lemma 1 holds.

(II) Some explanations about multi-split method

There are two places where multi-split method (Meinshausen et al.(2009)) are used in our paper. One is in example 2, the other is in empirical study. In example 2, since some non-zero coefficients are very small, it is hard to say some method can chose all true variables in the stage of model selection. When we conducted this simulation, we found that the density estimation curves may change due to different splits, which motivates us to remedy this deficiency by dividing the sample repeatedly. Meinshausen et al.(2009)) showed that multi-split method had better stability than a single-split. Therefore, we take their average as the final estimator to obtain a stable error density estimator in example 2. For the same purpose, we also applied multi-split method to analysis the example in empirical study, the corresponding results are presented in Figure 9. The density curves obtained by different splits are almost the same from Figure 9 (b), so we say that it may be acceptable to take one or their average as the final error density estimator. Next, we mainly explain the stability of multi-split method from two aspects.

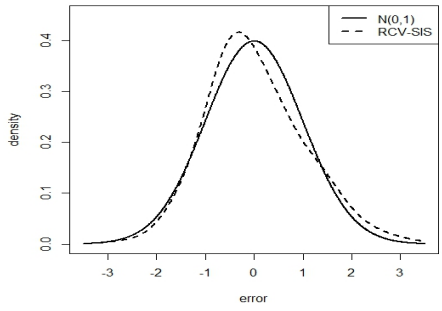
(1) When the sure screening property holds in the variable selection procedure, we first discuss theoretically the effect of different splits on the estimation results. For the sake of simplicity, we assume the true variables are "X_1, X_2". However, for two different splits, the selected variables may be "X_1, X_2, X_3" and "X_1, X_2, X_4" respectively. Therefore, there may be a slight difference among different splits. For visual intuition, we applied RCV method to example 1 again. By dividing the sample repeatedly, four representative density curves are given in Figure 10.

(2) When the sure screening property hardly holds in the variable selection procedure, our discussion is similar to (1). For example, we assume the true variables are "X_1, X_2, X_3". However, for two different splits, the selected variables are "X_1, X_2, X_5" and "X_2, X_3, X_5" respectively. Therefore, there may be

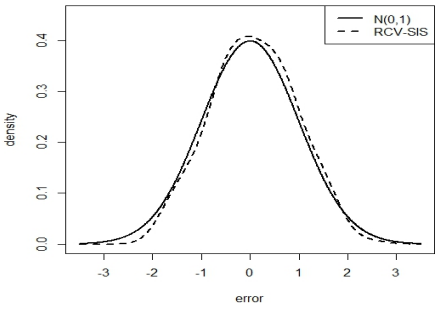


(a): split1

(b): split2



(c): split3



(d): split4

Figure 10: In the model $Y = a(X_1 + X_2 + X_3) + e$, we set $a = 1/\sqrt{3}$, $e \sim N(0,1)$ and $\{X_1, \ldots, X_p\} \sim N(0, \Sigma)$ with $\Sigma = \{\rho_{ij}\}_{i,j=1}^{p}$ where $\rho_{ii} = 1, \rho_{ij} = 0.5$ for $i \neq j$, then the simulation is conducted in the setting of $n = 200, p = 2000$.

some differences among different splits. Here, we take example 2 as an example, the simulation results are presented in Figure 11.

Through the discussion (1) and (2), we may find that the result obtained by a single-split isn't always the same due to the randomness of split. Intuitively, there may be only a slight difference among different splits when the sure screening property holds. On the contrary, different splitting results may have some differences. To sum up, the purpose of multi-split is to obtain a more stable error density estimator. A conservative approach is to divide the sample repeatedly for 3-4 times and take their average as the final estimator no matter whether these results obtained by 3-4 random splits are almost the same or not.

In our paper, our aim is to estimate the error density in the setting of $E(e_i) = 0, Var(e_i) = \sigma^2 < \infty, i = 1, \ldots, n$. As we have stated, it is the randomness of single split that leads to some differences in the results, which may be unrelated to $E[e_i^2|x_i] = \sigma^2(x_i)$.

(III) A proof about the order of $\max_{1\le i\le n} P_{ii}$

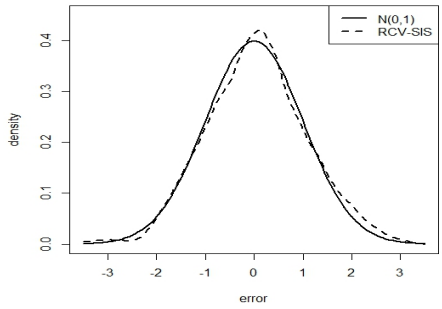
Next, we give a detailed proof. By the definition of matrix P, we have $P_{ii} = X_{i\hat{M}}^T (X_{\hat{M}}^T X_{\hat{M}})^{-1} X_{i\hat{M}}$ with $\hat{M} = \{j_1, \cdots, j_{\hat{s}}\}$, then by condition $\mathbf{C_2}$,

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

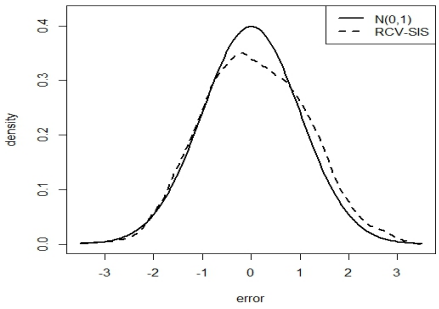
That is, we only need to compute the order of $\max_{1\le i\le n} \frac{1}{\hat{s}} \sum_{j\in\hat{M}} X_{ij}^2$. For the following random variable sequences

$$\begin{cases} X_{1j_1}\ X_{1j_2}\ \cdots\ X_{1j_{\hat{s}}} \\ X_{2j_1}\ X_{2j_2}\ \cdots\ X_{2j_{\hat{s}}} \\ \quad \cdots\ \cdots\ \cdots\ \cdots \\ X_{nj_1}\ X_{nj_2}\ \cdots\ X_{nj_{\hat{s}}}, \end{cases}$$

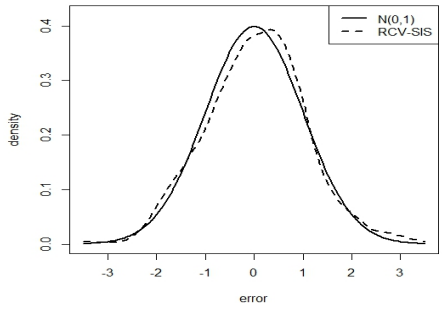
then the random sequence $\frac{1}{\hat{s}} \sum_{j\in\hat{M}} X_{ij}^2$ are i.i.d., $i = 1, \cdots, n$. For the sake of simplicity, we may as well set $Z_{in} = \frac{1}{\hat{s}} \sum_{j\in\hat{M}} X_{ij}^2$, where the subscript is



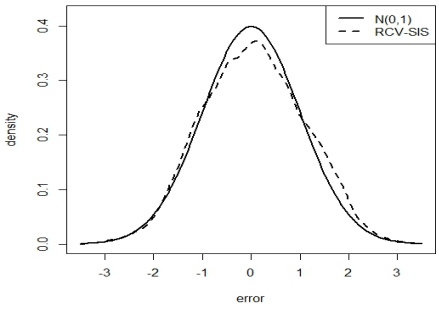
(a): split1



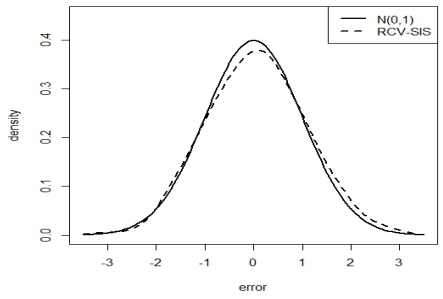
(b): split2



(c): split3



(d): split4



The average of four splits

Figure 11: In example 2, the model $Y = 1.01X_1 - 0.06X_2 + 0.72X_3 + 1.55X_5 + 2.32X_7 - 0.36X_{11} + 3.75X_{13} - 2.04X_{17} - 0.13X_{19} + 0.61X_{23} + e$, where $e \sim N(0,1)$, and $\{X_1, \ldots, X_p\} \sim N(0, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}, i, j = 1, 2, \ldots, p$, then the simulation is conducted in the setting of $n = 400, p = 10000$.

set to "in" due to $\hat{s} = \#\hat{M} = O(n^{\gamma})$ with $0 \leq \gamma < 1$. That is, the problem is turned into computing the order of $\max_{1\leq i\leq n} Z_{in}$. If $\hat{s}$ is fixed, the original condition $\sup_{1\leq j\leq p} E|X_{ij}|^{2k} < \infty$ may imply $\max_{1\leq i\leq n} \frac{1}{\hat{s}} \sum_{j\in\hat{M}} X_{ij}^2 = O(n^{1/k})$ $a.s.$. However, when the number of true variables s diverges at a mild rate, the original condition may need to be strengthened. Therefore, the original assumption $\sup_{1\leq j\leq p} E|X_{ij}|^{2k} < \infty$ in condition $\mathbf{C_2}$ has been replaced by $\sup_{1\leq j\leq p} E[X_{1j}^{4k}(\log^{+}(X_{1j}^2))^2] \leq C < \infty$. Since $Z_{in}, i = 1, \ldots, n$, are i.i.d., then

$$\begin{aligned}
&P(\max_{1\leq i\leq n} Z_{in} > A_0 n^{1/k}) \\
&= 1 - P(Z_{in} < A_0 n^{1/k})^n \\
&= 1 - [1 - P(Z_{in} > A_0 n^{1/k})]^n,
\end{aligned}$$

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

$$\begin{aligned}
P(Z_{in} > A_0 n^{1/k}) &\leq E\frac{h(Z_{in})}{h(A_0 n^{1/k})} \\
&\leq E\frac{\frac{1}{\hat{s}} \sum\limits_{j\in\hat{M}} h(X_{ij}^2)}{A_0^{2k}(\log^{+}(A_0) + \frac{1}{k}\log n)^2 n^2} \\
&\leq \frac{k^2}{A_0^{2k}} \frac{\sup_{1\leq j\leq p} Eh(X_{ij}^2)}{n^2 \log^2 n}.
\end{aligned}$$

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

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

thereby we have $\max_{1\leq i\leq n} Z_{in} = O(n^{1/k})$ $a.s.$.

(IV) An explanation about $\min_{1\leq \#M\leq cs} \lambda_{\min}(\frac{1}{n} X_M^T X_M) \geq \frac{\lambda_0}{2} > 0$

In our paper, the dimensionality reduces from p to $\hat{s}$ by the sure screening property of variable selection procedure, where $\hat{s} = O(s) = O(n^{\gamma})$ $a.s.$ with $0 \leq \gamma < 1$, as condition $\mathbf{C_0}$ stated in our revised manuscript. Without loss of generality, our assumption about the minimum eigenvalue inequality of $\frac{1}{n} X_M^T X_M$ is based on M, where $M \subset \{1, \ldots, p\}$ denotes the set of variables selected after dimensionality reduction and satisfies $1 \leq \#M \leq cs$ with $c \geq 1$. To illustrate

the rationality of the minimum eigenvalue assumption, we take the following two cases as examples in the setting of $\sup_{1\leq j\leq p} E[X_{1j}^{4k}(\log^+(X_{1j}^2))^2] \leq C < \infty$ with $k > 1$.

(i) $\lambda_{\min}(\Sigma) \geq \lambda_0 > 0$ and $X_1 = (X_{11}, \ldots, X_{1p})^T \sim N(0, \Sigma)$

In this case, we denote $X_M = (X_{1M}, \ldots, X_{nM})^T$ with $Cov(X_{1M}) = \Sigma_{1M}$. By the definition of M, we have $\frac{\#M}{n} \leq \frac{cs}{n} \to 0$ when n approaches to infinity. Denote $Y_M = X_M \Sigma_{1M}^{-1/2}$, then $Y_{ij}, i = 1, \ldots, n; j \in M$, are i.i.d.. Due to $\lambda_{\min}(\Sigma) \geq \lambda_0 > 0$ and the normality of $\{X_{1j}\}_{j=1}^p$, we can obtain $E|Y_{1j}|^4 < \infty$, $j = 1, \ldots, p$. By Theorems 1-2 of Bai and Yin (1993), we have $\lambda_{\min}(\frac{1}{n} Y_M^T Y_M) \to 1$ *a.s.*. That is, $\lambda_{\min}(\Sigma_{1M}^{-1/2}(\frac{1}{n} X_M^T X_M)\Sigma_{1M}^{-1/2}) \to 1$ *a.s.*. Here, we may as well set $\lambda_{\min}(\Sigma_{1M}^{-1}(\frac{1}{n} X_M^T X_M)) \geq 1/2$ holds *a.s.*. Furthermore, we have

$$\lambda_{\min}(\frac{1}{n} X_M^T X_M) \geq \frac{1}{2} \lambda_{\min}(\Sigma_{1M}) \text{ } a.s..$$

Based on the fact that Σ_{1M} is a submatrix of Σ with $\Sigma = Cov(X_1)$, where $X_1 = (X_{11}, \ldots, X_{1p})^T$, then we have $\lambda_{\min}(\Sigma_{1M}) \geq \lambda_{min}(\Sigma)$. Therefore,

$$\min_{1\leq \#M \leq cs} \lambda_{\min}(\frac{1}{n} X_M^T X_M) \geq \frac{\lambda_0}{2}.$$

(ii) $\{X_{1j}\}_{j=1}^p$ are independent with zero mean and covariance σ_j^2, where $\sigma_j^2 \geq \lambda_0 > 0$. Similar to the proof of (i), the above minimum eigenvalue inequality also holds.