VARIABLE SELECTION AND COEFFICIENT ESTIMATION VIA REGULARIZED RANK REGRESSION

Chenlei Leng

National University of Singapore

Supplemtary Materials

Relative efficiency

To compare the R^2 estimator with the MLE, the least squares estimator (LS) and the least absolute deviation estimator (LAD), we summarize the relative efficiency for various distributions in Table 1.1.

	$e(R^2, ML)$	$e(R^2, LS)$	$e(R^2, LAD)$
Normal	0.955	0.955	1.500
Logisitc	1.000	1.097	1.333
t_5	0.993	1.240	1.290
t_3	0.950	1.900	1.173
Cauchy	0.608	∞	0.750
DE	0.750	1.500	0.750
$T(0.01, 3)$	0.963	1.009	1.487
$T(0.05, 3)$	0.967	1.196	1.436
$T(0.1, 3)$	0.958	1.373	1.376

Table 1.1: The relative efficiency of the R^2. DE: double exponential, t_d: Student's t-distribution with d degrees of freedom. $T(\rho, \sigma)$: Tukey contaminated normal with cdf $F(x) = (1-\rho)\Phi(x) + \rho\Phi(x/\sigma)$ where $\Phi(\cdot)$ is the cdf of a standard normal distribution and $\rho \in [0, 1]$ is the contamination proportion.

Some Proofs

We denote the cdf and pdf of $\varepsilon_{ij} = \varepsilon_i - \varepsilon_j$ as G and g respectively. Simple algebra yields $g(s) = \int f(t)f(t-s)dt$. A proof of Theorem 1 can also be found in Hettmansperger and Mckean (1998, Theorem 3.5.4). However, for completeness and also since the proofs of Theorem 2 and Theorem 3 depend on the proof of Theorem 1, the proof of Theorem 1 is included.

Proof of Theorem 1. Denote $\mathbf{u} = \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)$. We consider

$$
\begin{aligned}
Z(\mathbf{u}) &= (2n)^{-1} \sum_{i,j} (|y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}| - |y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}^0|) \\
&= (2n)^{-1} \sum_{i,j} (|\varepsilon_{ij} - \mathbf{x}_{ij}^T \mathbf{u}/\sqrt{n}| - |\varepsilon_{ij}|)
\end{aligned}
$$

which is minimized at $\hat{\mathbf{u}} = \sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$. By Knight's identity (Knight, 1998)

$$
(|r - s| - |r|)/2 = -s(\frac{1}{2} - I(r < 0)) + \int_0^s (I(r \le t) - I(r \le 0))dt,
$$

we may write

$$
Z(\mathbf{u}) = Z_1(\mathbf{u}) + Z_2(\mathbf{u}),
$$

where

$$
\begin{aligned}
Z_1(\mathbf{u}) &= -n^{-3/2} \sum_{i,j} \mathbf{x}_{ij}^T \mathbf{u} (\frac{1}{2} - I(\varepsilon_{ij} < 0)), \\
Z_2(\mathbf{u}) &= n^{-1} \sum_{i,j} \int_0^{\mathbf{x}_{ij}^T \mathbf{u}/\sqrt{n}} (I(\varepsilon_{ij} \le t) - I(\varepsilon_{ij} \le 0))dt \triangleq \sum_{i,j} Z_{2ij}(\mathbf{u}).
\end{aligned}
$$

The limiting behavior of these two expressions is now discussed.

First note that

$$
\begin{aligned}
Z_1(\mathbf{u}) &= -n^{-3/2} \sum_{i,j} \mathbf{x}_{ij}^T \mathbf{u} (\frac{1}{2} - I(\varepsilon_{ij} < 0)) \\
&= -n^{-3/2} \sum_{i} \mathbf{x}_i^T \{2\mathrm{R}(\varepsilon_i) - (n+1)\} \mathbf{u} \\
&\triangleq W_n^T \mathbf{u},
\end{aligned}
$$

where $\mathrm{R}(\varepsilon_i)$ is the rank statistic of ε_i. By the independence between $\mathbf{x}_i$ and ε_i, $E(W_n) = 0$ and $Cov(W_n) = n^{-3} X^T Cov(\mathbf{r}) X$ for $\mathbf{r} = (2\mathrm{R}(\varepsilon_1) - (n+1), ..., \mathrm{R}(2\varepsilon_n) - (n+1))^T$. The diagonal terms of $n^{-2} Cov(\mathbf{r})$ are

$$
n^{-2} Var(r_i) = n^{-2} \sum_{i=1}^n \{2i - (n+1)\}^2 \frac{1}{n} = \frac{4(n+1)^2}{n^3} \sum_i (\frac{i}{n+1} - \frac{1}{2})^2 \to 4 \int (t - \frac{1}{2})^2 dt = \frac{1}{3},
$$

and its off-diagonal terms are

$$
n^{-2} Cov(r_i, r_j) = n^{-2} \sum_{i=1}^n \sum_{j \ne i} \{2i - (n+1)\}\{2j - (n+1)\} \frac{1}{n(n-1)} = -\frac{4(n+1)^2}{n^2(n-1)} \int (t - \frac{1}{2})^2 dt \to 0.
$$

Combined with Assumption A1, we have $Cov(W_n) \to C/3$. Therefore, an application of the Lindeberg-Feller central limit theorem using Assumptions A1-A2, yields

$$W_n \to_d W \text{ and } Z_1(\mathbf{u}) \to_d -\mathbf{u}^T W, \text{ where } W \sim N(\mathbf{0}, C/3).$$

For $Z_2(\mathbf{u})$, we write

$$Z_2(\mathbf{u}) = \sum_{i,j} EZ_{2ij} + \sum_{i,j}(Z_{2ij} - EZ_{2ij}).$$

We have

$$\begin{aligned}
EZ_2 = \sum_{i,j} EZ_{2ij} &= n^{-1} \sum_{i,j} \int_0^{\mathbf{x}_{ij}^T \mathbf{u}/\sqrt{n}} (G(t) - G(0))dt \\
&= \frac{1}{n^{3/2}} \sum_{i,j} \int_0^{\mathbf{x}_{ij}^T \mathbf{u}} (G(s/\sqrt{n}) - G(0))ds \\
&= \frac{1}{n^2} \sum_{i,j} \int_0^{\mathbf{x}_{ij}^T \mathbf{u}} sg(0)ds + o(1) \\
&= \frac{1}{2n^2} \sum_{i,j} g(0)\mathbf{u}^T \mathbf{x}_{ij}^T \mathbf{x}_{ij} \mathbf{u} + o(1) \\
&\to g(0)\mathbf{u}^T C\mathbf{u}.
\end{aligned}$$

Here we use the fact that $\sum_{i,j} \mathbf{x}_{ij}^T \mathbf{x}_{ij} = 2nX^T X$ in the last step. By noting that

$$\begin{aligned}
V(Z_{2ij}) &= n^{-2} E\Big\{ \int_0^{\mathbf{x}_{ij}^T \mathbf{u}/\sqrt{n}} [I(\varepsilon_{ij} \le t) - I(\varepsilon_{ij} \le 0)] - [g(t) - g(0)] \Big\}^2 \\
&\le n^{-2} E\Big\{ | \int_0^{\mathbf{x}_{ij}^T \mathbf{u}/\sqrt{n}} [I(\varepsilon_{ij} \le t) - I(\varepsilon_{ij} \le 0)] - [g(t) - g(0)] | \Big\} \times 2|\frac{\mathbf{x}_{ij}^T \mathbf{u}}{\sqrt{n}}| \\
&\le n^{-2} \frac{4 \max |\mathbf{x}_{ij}^T \mathbf{u}|}{\sqrt{n}} EZ_{2ij}(\mathbf{u}),
\end{aligned}$$

we have

$$Var(Z_2(\mathbf{u})) \le n^2 \sum_{i,j} Var(Z_{2ij}) \le \frac{4 \max |\mathbf{x}_{ij}^T \mathbf{u}|}{\sqrt{n}} EZ_2 \to 0.$$

Therefore,

$$Z(\mathbf{u}) \to_d Z^0(\mathbf{u}) = -\mathbf{u}^T W + g(0)\mathbf{u}^T C\mathbf{u}.$$

Also, we have $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \rightarrow_d (2g(0))^{-1}C^{-1}W \sim N(\mathbf{0}, C^{-1}/\{12g^2(0)\})$ by the convexity of the limiting function $Z^0(\mathbf{u})$. This completes the proof.

Proof of Theorem 2. Consider

$$Q(\boldsymbol{\beta}) = (\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^T C_n(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})+\lambda\sum_{k=1}^p \lambda_k|\beta_k| - \Big\{(\boldsymbol{\beta}^0-\tilde{\boldsymbol{\beta}})^T C_n(\boldsymbol{\beta}^0-\tilde{\boldsymbol{\beta}})+\lambda\sum_{k=1}^p \lambda_k|\beta_k^0|\Big\}.$$

Denote $\mathbf{u} = \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)$. We may write $nQ(\boldsymbol{\beta})$ as

$$nQ(\mathbf{u}) = \mathbf{u}^T C_n \mathbf{u} + 2\mathbf{u}^T C_n\big[\sqrt{n}(\boldsymbol{\beta}^0 - \tilde{\boldsymbol{\beta}})\big] + n\lambda\sum_{k=1}^p \lambda_k|\beta_k| - n\lambda\sum_{k=1}^p \lambda_k|\beta_k^0|,$$

which is minimized by $\sqrt{n}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^0)$. Let

$$Z_3(\mathbf{u}) = n\lambda\sum_k \lambda_k(|\beta_k^0 + u_k/\sqrt{n}| - |\beta_k^0|).$$

For Z_3, we write $Z_{3k}(\mathbf{u}) = n\lambda\lambda_k(|\beta_k^0 + u_k/\sqrt{n}| - |\beta_k^0|)$ and then

$$Z_{3k}(\mathbf{u}) = \begin{cases} \sqrt{n}\lambda\lambda_k u_k \mathrm{Sign}(\beta_k^0), & \text{if } \beta_k^0 \neq 0, \\ \sqrt{n}\lambda\lambda_k|u_k|, & \text{if } \beta_k^0 = 0. \end{cases}$$

Now, the conditions in Theorem 2 assure the following

$$Z_{3k}(\mathbf{u}) \rightarrow P(\beta_k^0, u_k) = \begin{cases} 0 & \text{if } \beta_k^0 \neq 0, \\ 0 & \text{if } \beta_k^0 = 0 \text{ and } u_{\lambda k} = 0, \\ \infty & \text{if } \beta_k^0 = 0 \text{ and } u_{\lambda k} \neq 0. \end{cases}$$

Thus, we have

$$Q(\mathbf{u}) \rightarrow_d \mathbf{u}^T C\mathbf{u} - \frac{2}{2\omega}\mathbf{u}^T W + \sum_{k=1}^p P(\beta_k^0, u_k),$$

where W is given in Theorem 1. Applying the arguments in Knight (1998), we have

$$\begin{aligned} &\hat{u}_{\lambda\mathcal{A}^C} \rightarrow_d 0, \text{ for } \beta_k^0 = 0, \\ &\hat{u}_{\lambda\mathcal{A}} \rightarrow_d \frac{1}{2\omega}C_{\mathcal{A}\mathcal{A}}^{-1}W_{\mathcal{A}} \sim N(\mathbf{0}, 1/(12\omega^2)C_{\mathcal{A}\mathcal{A}}^{-1}). \end{aligned}$$

The asymptotic normality is established.

The consistency results can be seen as follows. Since $\hat{\boldsymbol{\beta}}_\lambda$ is root-n consistent, we have $\mathrm{P}(k \in \mathcal{S}_\lambda) \to 1$ for $k \in \mathcal{A}$, where $\mathcal{S}_\lambda$ is the model identified by $\hat{\boldsymbol{\beta}}_\lambda$. Note that if $\exists k \in \mathcal{A}^C$, such that $\hat{\beta}_{\lambda k} \neq 0$, we must have

$$\sqrt{n}\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_k}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_\lambda} = 2C_n^{(k)} \times \sqrt{n}(\hat{\boldsymbol{\beta}}_\lambda - \tilde{\boldsymbol{\beta}}) + \sqrt{n}\lambda\lambda_k \mathrm{sgn}(\hat{\boldsymbol{\beta}}_{\lambda k}),$$

where $C^{(k)}$ stands for the kth row of C. Now, the order of the first term is bounded since $C_n \to_p C$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_\lambda - \tilde{\boldsymbol{\beta}}) = O_p(1)$. On the other hand, $\sqrt{n}\lambda\lambda_k \mathrm{sgn}(\hat{\beta}_{\lambda k})$ goes to ∞ as long as n is large. Therefore, $\sqrt{n}\partial Q(\hat{\boldsymbol{\beta}}_\lambda)/\partial\beta_k$ cannot be zero for n sufficiently large. The contradiction proves the consistency of variable selection. Now, it is easy to see that once λ satisfies

$$\sqrt{n}\lambda a_n \to 0 \text{ and } \sqrt{n}\lambda b_n \to \infty,$$

the assumptions of the theorem holds.

Lemma A1. When the assumptions in Theorem 1 and 2 are satisfied, with probability tending to one,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\lambda\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) = \frac{1}{2\omega}C^{-1}_{\mathcal{A}\mathcal{A}}W_{n\mathcal{A}} + o_p(1),$$

where W_n is defined in Theorem 1.

Proof: The result follows from the proofs of Theorem 1 and Theorem 2.

Lemma A2. (Asymptotic linearity)

$$\mathrm{P}\{\sup_{\sqrt{n}\|\boldsymbol{\beta}-\boldsymbol{\beta}^0\|\le B} \|n^{-1/2}G(\boldsymbol{\beta}) - n^{-1/2}G(\boldsymbol{\beta}^0) + 2\omega C_n(\boldsymbol{\beta}-\boldsymbol{\beta}^0)\sqrt{n}\| \ge \delta\} \to 0,$$

for any fixed $B \in \mathbb{R}$ and δ.

Proof: See Sievers (1983).

Lemma A3. For the reference tuning parameter sequence, with probability tending to one

$$T_{\lambda_n} \to \chi^2_q,$$

where $q = p - \#\{\mathcal{A}\}$ is the number of the zero coefficients in $\boldsymbol{\beta}^0$.

Proof. Without loss of generality, assume $\mathcal{A} = \{1, ..., p-q\}$. Since λ_n satisfies the conditions in Theorem 2, we have with probability tending to one,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\lambda_n\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) = \frac{1}{2\omega}C^{-1}_{\mathcal{A}\mathcal{A}}W_{n\mathcal{A}} + o_p(1) = O_p(1)$$

and $\hat{\boldsymbol{\beta}}_{\lambda_n \mathcal{A}^C} = 0$. According to Lemma A2, we have

$$\begin{aligned}
T_{\lambda_n} &= 3n^{-1}G^T(\hat{\boldsymbol{\beta}}_{\lambda_n})C_n^{-1}G^T(\hat{\boldsymbol{\beta}}_{\lambda_n}) \\
&= 3\{n^{-1/2}G^T(\boldsymbol{\beta}^0) - 2\omega C_n(\hat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^0)\}^T C_n^{-1}\{n^{-1/2}G^T(\boldsymbol{\beta}^0) - 2\omega C_n(\hat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^0)\} \\
&= 3\{W_n^T C_n^{-1} W_n - W_{n\mathcal{A}}^T C_{n\mathcal{A}\mathcal{A}} W_{n\mathcal{A}}\} \\
&= \{\sqrt{3}C_n^{-1/2}W_n\}^T\{I - C_n^{1/2}\begin{pmatrix} C_{n\mathcal{A}\mathcal{A}}^{-1} & 0 \\ 0 & 0 \end{pmatrix} C_n^{1/2}\}\{\sqrt{3}C_n^{-1/2}W_n.\}
\end{aligned}$$

An application of Cochran's Theorem gives $T_{\lambda_n} \to \chi_q^2$ by noting $\sqrt{3}C_{n\mathcal{A}\mathcal{A}}^{-1/2}W_{n,\mathcal{A}} \to_d N(\mathbf{0}, I)$.

Proof of Consistency of SIC. We classify any $\mathcal{S}_\lambda \neq \mathcal{A}$ into two different cases according to whether the model is underfitted ($\mathbb{R}_- = \{\lambda \geq 0 : \mathcal{S}_\lambda \not\supset \mathcal{A}\}$) or overfitted ($\mathbb{R}^+ = \{\lambda \geq 0 : \mathcal{S}_\lambda \supset \mathcal{A}, \mathcal{S}_\lambda \neq \mathcal{A}\}$). In either case, we show that the theorem's conclusion is valid. Specifically,

Case 1 (Underfitted Model). Since λ_n satisfies the regularity conditions specified by Theorem 2, the resulting estimator $\hat{\beta}_{\lambda_n}$ is $\sqrt{n}$-consistent. From Lemma A3, its associated SIC value is of the order $O_p(\log\log(n))$. On the other hand, since $\mathcal{S}_\lambda \not\supset \mathcal{A}$ is an underfitted model, we know that with probability tending to one,

$$SIC_\lambda/n \geq \inf_{\boldsymbol{\beta}:\ \beta_j=0, j\notin\mathcal{S}_\lambda} 3n^{-1}G(\boldsymbol{\beta})^T C_n^{-1} G(\boldsymbol{\beta})/n > 0.$$

Hence we have $P(\inf_{\lambda\in\mathbb{R}_-} SIC_\lambda > SIC_{\lambda_n}) \to 1$.

Case 2 (Overfitted Model). For any overfitted model, we have $df_\lambda > df_{\lambda_n}$ and

$$\begin{aligned}
SIC_\lambda - SIC_{\lambda_n} &= T_\lambda - T_{\lambda_n} + (df_\lambda - df_{\lambda_n})\log\log(n) \\
&\geq \inf_{\boldsymbol{\beta}:\ \beta_j=0, j\notin\mathcal{S}_\lambda} 3n^{-1}G(\boldsymbol{\beta})^T C_n^{-1} G(\boldsymbol{\beta}) - T_{\lambda_n} + \log\log(n)
\end{aligned}$$

Now, following Lemma A3, it is easy to see that the first term converges to χ_k^2 where $k = p - \#\{\mathcal{S}_\lambda\}$. The second term is $O_p(1)$ by Lemma A3. Thus, the third term dominates the expression and we have $P(\inf_{\lambda\in\mathbb{R}^+} SIC_\lambda > SIC_{\lambda_n}) \to 1$. The proof is completed.

References

Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *The Annals of Statistics*, 26, 755-770.

Sievers, G. L. (1983). A weighted dispersion function for estimation in linear models. *Communications in Statistics Theory and Methodology*, 12, 1161-1179.