Technical Proofs for "Regression model selection - a residual likelihood approach" by P. D. Shi and C. L. Tsai

We will need the following Lemmas to show the asymptotic properties of RIC.

Lemma 1. Suppose that $T = (t_1, \cdots, t_n)'$ is a sequence of random variables with *iid* components. Assume that $E(t_1) = 0$, $var(t_1) < \infty$, and $E|t_1|^{4s} < \infty$ for some $s > 1$. Let $\hat{a} = T'BT$, where B is a sequence of nonrandom symmetric matrices. Suppose that $\lim_{n\to\infty} E(\hat{a}) = a$. Then,

1. $E|\hat{a} - E(\hat{a})|^{2s} \leq D(s)E(t_1^{4s})(\mathrm{tr}B^2)^s$, where $D(s)$ is a constant depending only on s.
2. $\lim_{n\to\infty} \hat{a} = a$ almost surely (a.s.), provided that there exist $c > 0$, $\eta > 1/s$ for which

$$n^\eta \mathrm{tr}(B^2) < c, \qquad \forall n.$$

The above Lemma is a special version of Rao and Kleffe's (1988) Equation (2.3.10) and Theorem 2.4.2. Note that E used in Lemma 1 represents the classical definition of the expectation.

Lemma 2. If the ϵ_i are *iid* and Assumptions 1, 2 and 3 are satisfied, then $\lim_n \frac{1}{n}\epsilon'\{I - H(\alpha)\}X_0\beta_0 = 0$ a.s. uniformly in $\alpha \in \mathcal{A} - \mathcal{A}^0$.

Proof. Let $\xi_n = (\xi_{1n}, \cdots, \xi_{nn})' = \{I - H(\alpha)\}X_0\beta_0/c_n$, where $c_n = [\frac{1}{n}\beta_0' X_0'\{I - H(\alpha)\}X_0\beta_0]^{1/2}$. Under Assumption 3, we have c_n is larger than some positive constant when n is large enough. Thus, we have

$$\xi_n'\xi_n = 1 \qquad \text{and} \qquad \frac{1}{n}\epsilon'\{I - H(\alpha)\}X_0\beta_0 = \frac{c_n}{n}\sum_{i=1}^n \epsilon_i\xi_{in}.$$

From Assumption 2, there exists a constant $c_0 > 0$ such that $c_0 \geq c_n$ for all n. Under Assumption 1, we apply the Chebychev's inequality to obtain that for any $\delta > 0$

$$P\{|\frac{c_n}{n}\sum_{i=1}^n \epsilon_i\xi_{in}| > \delta\} \leq (c_0/(n\delta))^4 E\left(\sum_{i=1}^n \epsilon_i\xi_{in}\right)^4 \leq Mn^{-4},$$

where M is a positive constant. This, together with the Borel-Cantelli Lemma, leads to the conclusion of Lemma 2.

A.1. Proof of theorem 1

Without loss of generality, we use y in place of $h(y,1) = y - \mathbf{1}$ in the proof of Theorem 1 below, where $\mathbf{1} = (1, \cdots, 1)'$.

STEP 1. For $\alpha \in \mathcal{A}$ and $\alpha^0 \in \mathcal{A}^0$, the corresponding model selection criteria are

$$RIC(\alpha) = n \log\{\tilde{\sigma}^2(\alpha)\} + k(\alpha)\{\log(n) - 1\} + \frac{4}{n - k(\alpha) - 2} - k(\alpha) \log\{\tilde{\sigma}^2(\alpha)\}, \quad (1)$$

and

$$RIC(\alpha^0) = n \log\{\tilde{\sigma}^2(\alpha^0)\} + k(\alpha^0)\{\log(n) - 1\} + \frac{4}{n - k(\alpha^0) - 2} - k(\alpha^0) \log\{\tilde{\sigma}^2(\alpha^0)\}, \quad (2)$$

where

$$\tilde{\sigma}^2(\alpha) = y'\{I - H(\alpha)\}y/\{n - k(\alpha)\}, \qquad \tilde{\sigma}^2(\alpha^0) = y'\{I - H(\alpha^0)\}y/\{n - k(\alpha^0)\},$$

$H(\alpha)$ is defined in Assumption 3, and $H(\alpha^0) = X(\alpha^0)(X(\alpha^0)'X(\alpha^0))^{-1}X'(\alpha^0)$. After algebraic simplification,

$$\begin{aligned}
&\tilde{\sigma}^2(\alpha) - \tilde{\sigma}^2(\alpha^0) \\
&= \left[\frac{1}{n - k(\alpha)} - \frac{1}{n - k(\alpha^0)}\right] \sum_{i=1}^{n} \epsilon_i^2 + \frac{\varepsilon' H(\alpha^0)\varepsilon}{n - k(\alpha^0)} - \frac{\varepsilon' H(\alpha)\varepsilon}{n - k(\alpha)} \\
&\quad + \frac{2}{n - k(\alpha)} \varepsilon'\{I - H(\alpha)\}X_0\beta_0 + \frac{1}{n - k(\alpha)} \beta_0' X_0' \{I - H(\alpha)\}X_0\beta_0.
\end{aligned}$$

Observe that the first term of the right hand side of the last equation is $o(1)$. Since $H(\alpha^0)$ is an idempotent matrix and

$$n^{1/2} tr\{H^2(\alpha^0)n^{-2}\} = k(\alpha^0)n^{-3/2} \to 0,$$

we conclude, from Lemma 1 with $B = H(\alpha^0)n^{-1}$, that

$$n^{-1}\varepsilon' H(\alpha^0)\varepsilon = o(1) \quad \text{almost surely.}$$

Similarly, $n^{-1}\varepsilon' H(\alpha)\varepsilon = o(1)$ almost surely. For $\alpha \in \mathcal{A} - \mathcal{A}^0$ and when the ε_i are *iid* and satisfy Assumption 1, Lemma 2 implies that $\lim_n \frac{2}{n}\varepsilon'\{I - H(\alpha)\}X_0\beta_0 = 0$ almost surely. Thus we have

$$\liminf_n \{\tilde{\sigma}^2(\alpha) - \tilde{\sigma}^2(\alpha^0)\} \geq \liminf_n [\beta_0' X_0' \{I - H(\alpha)\}X_0\beta_0/(n - k(\alpha^0))] \geq c_1^2 > 0$$

almost surely for all $\alpha \in \mathcal{A} - \mathcal{A}^0$, where $c_1 > 0$ is a constant such that $\beta_0' X_0' \{I - H(\alpha)\} X_0 \beta_0 / (n - k(\alpha^0)) \geq c_1^2$ when n is large enough. This leads to

$$\begin{aligned} &\liminf_n \tilde{\sigma}^{-2}(\alpha^0)\{\tilde{\sigma}^2(\alpha) - \tilde{\sigma}^2(\alpha^0)\} \\ &\geq \liminf_n \{c_1^2/\tilde{\sigma}^2(\alpha^0)\} = \liminf_n c_1^2/[\varepsilon'\{I - H(\alpha^0)\}\varepsilon/(n - k(\alpha^0))] = (c_1/\sigma_0)^2, \end{aligned}$$

almost surely for all $\alpha \in \mathcal{A} - \mathcal{A}^0$. Hence,

$$\begin{aligned} &\liminf_n |n[\log\{\tilde{\sigma}^2(\alpha)\} - \log\{\tilde{\sigma}^2(\alpha^0)\}]| = \liminf_n |n \log[1 + \tilde{\sigma}^{-2}(\alpha^0)\{\tilde{\sigma}^2(\alpha) - \tilde{\sigma}^2(\alpha^0)\}]| \\ &\geq \liminf_n n\{\log(1 + (c_1/\sigma_0)^2\} > 0, \end{aligned} \tag{3}$$

almost surely for all $\alpha \in \mathcal{A} - \mathcal{A}^0$.

Next, it is also easy to see that $[\frac{4}{n-k(\alpha)-2} - k(\alpha)\log\{\tilde{\sigma}^2(\alpha)\}]/n \longrightarrow 0$, almost surely for all $\alpha \in \mathcal{A}$. This result, in conjunction with equations (1) and (2) and inequality (3), implies that

$$\begin{aligned} &\liminf_n \{RIC(\alpha) - RIC(\alpha^0)\} \\ &= \liminf_n n[\log\{1 + \tilde{\sigma}^{-2}(\alpha^0)(\tilde{\sigma}^2(\alpha) - \tilde{\sigma}^2(\alpha^0))\} \\ &\geq \liminf_n n[\log\{1 + (c_1/\sigma_0)^2\}] > 0, \end{aligned} \tag{4}$$

almost surely for all $\alpha \in \mathcal{A} - \mathcal{A}^0$.

STEP 2. Here we show the asymptotic property $\{RIC(\alpha) - RIC(\alpha^0)\}/(\log(n) - 1)$ as $\alpha \in \mathcal{A}^0$. To this end, we first note that

$$\frac{RIC(\alpha) - RIC(\alpha^0)}{\log(n) - 1} = \frac{n[\log\{\tilde{\sigma}^2(\alpha)\} - \log\{\tilde{\sigma}^2(\alpha^0)\}]}{\log(n) - 1} + k(\alpha) - k(\alpha^0) + V, \tag{5}$$

where $V = \frac{k(\alpha)\log\{\tilde{\sigma}^2(\alpha)\} - k(\alpha^0)\log\{\tilde{\sigma}^2(\alpha^0)\}}{\log(n)-1} + \frac{4\}}{\{n-k(\alpha)-2\}(\log(n)-1)} - \frac{4\}}{\{n-k(\alpha^0)-2\}(\log(n)-1)}$. In addition, Lemma 1 and Assumption 1 yield that, when $\alpha \in \mathcal{A}^0$,

$$\tilde{\sigma}^2(\alpha) = \varepsilon'\{I - H(\alpha)\}\varepsilon/(n - k(\alpha)) \to \sigma_0^2. \tag{6}$$

Hence,

$$V = o(1). \tag{7}$$

Next, we can see that whenever $|\tilde{\sigma}^{-2}(\alpha^0)\{\tilde{\sigma}^2(\alpha) - \tilde{\sigma}^2(\alpha^0)\}| \leq 1$,

$$\begin{aligned} &|\log\{\tilde{\sigma}^2(\alpha)\} - \log\{\tilde{\sigma}^2(\alpha^0)\}| = |\log[1 + \tilde{\sigma}^{-2}(\alpha^0)\{\tilde{\sigma}^2(\alpha) - \tilde{\sigma}^2(\alpha^0)\}]| \\ &\leq O\Big(\tilde{\sigma}^{-2}(\alpha^0)\Big\{\frac{k(\alpha) - k(\alpha^2)}{(n - k(\alpha))(n - k(\alpha^0))}\sum_{i=1}^n \epsilon_i^2 - \frac{\varepsilon' G \varepsilon}{n} + \frac{\varepsilon' H(\alpha^0)\varepsilon}{n(n - k(\alpha^0))} - \frac{\varepsilon' H(\alpha)\varepsilon}{n(n - k(\alpha))}\Big\}\Big), \end{aligned} \tag{8}$$

where $G = H(\alpha) - H(\alpha^0)$. Let $X(\alpha) = (X(\alpha^0), X_{2k(\alpha)})$. Then applying Chatterjee and Hadi's Property 2.4 (1988, p. 16), we have

$$G = \{I - H(\alpha^0)\}X_{2k(\alpha)}[X'_{2k(\alpha)}\{I - H(\alpha^0)\}X_{2k(\alpha)}]^{-1}X'_{2k(\alpha)}\{I - H(\alpha^0)\},$$

which is an idempotent matrix with trace $k(\alpha) - k(\alpha^0)$. This and Lemma 1 with $B = G$ imply that

$$\begin{aligned}&P\{|\varepsilon' G\varepsilon - \{k(\alpha) - k(\alpha^0)\}| > \delta(\log(n) - 1)\}\\ &\leq D(s)E(\varepsilon_1^{4s})\{k(\alpha) - k(\alpha^0)\}^s/\{\delta(\log(n) - 1)\}^2 \longrightarrow 0\end{aligned}$$

for any $\delta > 0$ and $s > 1$. Hence, $\varepsilon' B\varepsilon - \{k(\alpha) - k(\alpha^0)\} = o_p(\log(n) - 1)$. In conjunction with (8), we obtain

$$n[\log\{\tilde{\sigma}^2(\alpha)\} - \log\{\tilde{\sigma}^2(\alpha^0)\}] = o_p(\log(n) - 1). \tag{9}$$

Finally, combining Equations (5), (6), (7) and (9), we have, except for an event whose probability tends to zero with n, that

$$\frac{RIC(\alpha) - RIC(\alpha^0)}{\log(n) - 1} = k(\alpha) - k(\alpha^0) > 0 \tag{10}$$

for all $\alpha \in \mathcal{A}^0$ and $\alpha \neq \alpha^0$ when n is large enough.

STEP 3. From (4) and (10), we obtain that, except for an event whose probability tends to zero with n,

$$RIC(\alpha) > RIC(\alpha^0) \tag{11}$$

for all $\alpha \in \mathcal{A}$ and $\alpha \neq \alpha^0$. Since $\hat{\alpha}$ is the model satisfying

$$RIC(\hat{\alpha}) = \min_{\alpha \in \mathcal{A}} RIC(\alpha),$$

we have

$$RIC(\hat{\alpha}) \leq RIC(\alpha^0).$$

This, together with (11) implies that $\hat{\alpha} = \alpha^0$ when n is large enough, except for an event whose probability tends to zero with n.

A.2. Proof of theorem 2

Under the assumptions that the elements of $W(\theta)$ are continuous and that $\hat{\theta}$ tends to θ_0 in probability, we obtain that $\log\{|W(\hat{\theta})|/|W(\theta_0)|\} = o_p(1)$ for all $\alpha \in \mathcal{A}$. Hence, (2.10) can be rewritten as

$$\begin{aligned}RIC(\alpha, \hat{\theta}) = \ &(n-k)[\log\{\tilde{\sigma}^2(\alpha, \theta_0)\} + o_p(1)]\\ &+ \log|W_0| + k(\alpha)\{\log(n) - 1\} + \frac{4}{n - k(\alpha) - 2} + o_p(1).\end{aligned}$$

Also,

$$\begin{aligned} RIC(\alpha^0,\hat{\theta}) &= (n-k)[\log\{\tilde{\sigma}^2(\alpha^0,\theta_0)\}+o_p(1)] \\ &+\log|W_0|+k(\alpha^0)\{\log(n)-1\}+\frac{4}{n-k(\alpha^0)-2}+o_p(1). \end{aligned}$$

By using the same techniques used for the proof of Theorem 1, we can show that

$$RIC(\alpha,\hat{\theta}) > RIC(\alpha^0,\hat{\theta}) \tag{12}$$

for all $\alpha \in \mathcal{A}$ and $\alpha \neq \alpha^0$, except for an event whose probability tends to zero with n. Note that $\hat{\alpha}$ is the model satisfying

$$RIC(\hat{\alpha},\hat{\theta}) = \min_{\alpha\in\mathcal{A}} RIC(\alpha,\hat{\theta})$$

and that

$$RIC(\hat{\alpha},\hat{\theta}) \leq RIC(\alpha^0,\hat{\theta}).$$

This, together with (12) implies that $\hat{\alpha}=\alpha^0$ when n is large enough, except for an event whose probability tends to zero with n.

A.3. Proof of theorem 3

In the case of the Box-Cox regression model, the model selection criteria RIC for $\alpha \in \mathcal{A}$ and $\alpha^0 \in \mathcal{A}^0$ are given in (1) and (2), respectively, except for replacing $\tilde{\sigma}_2(\alpha)$ and $\tilde{\sigma}_2(\alpha^0)$ by $\tilde{\sigma}^2(\alpha) = h(y,\hat{\lambda}_0)'\{I-H(\alpha)\}h(y,\hat{\lambda}_0)/(n-k(\alpha))$ and $\tilde{\sigma}^2(\alpha^0) = h(y,\hat{\lambda}_0)'\{I-H(\alpha^0)\}h(y,\hat{\lambda}_0)/(n-k(\alpha^0))$.

Using the root-n consistency of $\hat{\lambda}_0$ and Assumption 4, we obtain that

$$(h(y,\hat{\lambda})-h(y,\lambda_0))'(h(y,\hat{\lambda})-h(y,\lambda_0)) = o_p(\log(n)).$$

Applying arguments similar to those used in the proof of Theorem 1, we can verify that

$$\begin{aligned} &\liminf_n |n[\log\{\tilde{\sigma}^2(\alpha)\}-\log\{\tilde{\sigma}^2(\alpha^0)\}]| \\ &= \liminf_n |n\log[1+\tilde{\sigma}^{-2}(\alpha^0)\{\tilde{\sigma}^2(\alpha)-\tilde{\sigma}^2(\alpha^0)\}]| \\ &\geq \liminf_n n\{\log(1+(c_1/\sigma_0)^2)\} > 0, \end{aligned} \tag{13}$$

almost surely for all $\alpha \in \mathcal{A}-\mathcal{A}^0$, and

$$[\frac{4}{n-k(\alpha)-2} - k(\alpha)\log\{\tilde{\sigma}^2(\alpha)\}]/n \longrightarrow 0, \tag{14}$$

almost surely for all $\alpha \in \mathcal{A}$, where c_1 is a positive constant. Equations (13) and (14) imply that

$$\begin{aligned} &\liminf_n\{RIC(\alpha)-RIC(\alpha^0)\} \\ &= \liminf_n n[\log\{1+\tilde{\sigma}^{-2}(\alpha^0)(\tilde{\sigma}^2(\alpha)-\tilde{\sigma}^2(\alpha^0))\}+o(1)] \\ &\geq \liminf_n n[\log\{1+(c_1/\sigma_0)^2\}] > 0, \end{aligned} \tag{15}$$

almost surely for all $\alpha \in \mathcal{A} - \mathcal{A}^0$. Furthermore, using similar arguments used in the proof of (10), we can show that

$$\frac{RIC(\alpha) - RIC(\alpha^0)}{\log(n) - 1} = k(\alpha) - k(\alpha^0) > 0 \tag{16}$$

for all $\alpha \in \mathcal{A}^0$ and $\alpha \neq \alpha^0$ when n is large enough, except for an event whose probability tends to zero with n.

From (15) and (16), we obtain that, except for an event whose probability tends to zero with n,

$$RIC(\alpha) > RIC(\alpha^0) \tag{17}$$

for all $\alpha \in \mathcal{A}$ and $\alpha \neq \alpha^0$. Since $\hat{\alpha}$ is the model satisfying

$$RIC(\hat{\alpha}) = \min_{\alpha \in \mathcal{A}} RIC(\alpha),$$

we have

$$RIC(\hat{\alpha}) \leq RIC(\alpha^0).$$

This, together with (17), completes the proof of Theorem 3.