Supplement to "Efficiency and Consistency for Regularization Parameter Selection in Penalized Regression: Asymptotics and Finite-Sample Corrections"

November 2, 2011

This document contains the technical proofs for the theoretical results included in the manuscript "Efficiency and Consistency for Regularization Parameter Selection in Penalized Regression: Asymptotics and Finite-Sample Corrections".

Proofs of Theorems 1, 2, and 3

Before proving Theorems 1, 2, and 3, we start by establishing the following two lemmas.

Lemma 1. *Assume that (A1)-(A4) hold and that* $d_n/n \to 0$ *as* $n \to \infty$. *Then*

$$\sup_{\lambda \in [0, \lambda_{max}]} \frac{d_n(\alpha_\lambda)|\hat{\sigma}_n^2(\lambda) - \sigma^2|}{nL(\hat{\beta}_n(\lambda))} \to_p 0.$$

Proof. The technique used to prove this result is similar to the proof of Theorem 2 in Shibata

(1981). First consider

$$
\begin{aligned}
|\hat{\sigma}_n^2(\lambda) - \sigma^2| &= \left| \frac{||\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n(\lambda)||^2}{n} - \sigma^2 \right| \\
&\leq \left| \frac{||\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda)||^2}{n} - \sigma^2 \right| + \frac{||\hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda) - \hat{\boldsymbol{\mu}}_n(\lambda)||^2}{n} \\
&= \left| \frac{||\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda)||^2 + 2\boldsymbol{\varepsilon}_n^T(\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda)) + ||\boldsymbol{\varepsilon}_n||^2}{n} - \sigma^2 \right| + \frac{||\hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda) - \hat{\boldsymbol{\mu}}_n(\lambda)||^2}{n} \\
&\leq L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda)) + 2\left| \frac{\boldsymbol{\varepsilon}_n^T(\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda))}{n} \right| + \left| \frac{||\boldsymbol{\varepsilon}_n||^2}{n} - \sigma^2 \right| + \frac{||\hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda) - \hat{\boldsymbol{\mu}}_n(\lambda)||^2}{n}.
\end{aligned}
$$

Applying the Cauchy-Schwarz inequality, it follows that

$$
|\hat{\sigma}_n^2(\lambda) - \sigma^2| \leq L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda)) + 2||\boldsymbol{\varepsilon}_n|| \frac{||\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda)||}{n} + \left| \frac{||\boldsymbol{\varepsilon}_n||^2}{n} - \sigma^2 \right| + \frac{||\hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda) - \hat{\boldsymbol{\mu}}_n(\lambda)||^2}{n}.
$$

Then

$$
\begin{aligned}
\frac{|\hat{\sigma}_n^2(\lambda) - \sigma^2| d_n(\alpha_\lambda)}{nL(\hat{\boldsymbol{\beta}}_n(\lambda))} &\leq \frac{d_n(\alpha_\lambda)}{n} \left[\frac{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))} \right] \\
&\quad + \frac{2}{\sigma} \left[\frac{d_n(\alpha_\lambda)}{n} \frac{||\boldsymbol{\varepsilon}_n||^2}{n} \right]^{1/2} \left[\frac{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))} \right]^{1/2} \left[\frac{\sigma^2 d_n(\alpha_\lambda)}{n\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))} \frac{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))} \frac{\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))} \right]^{1/2} \\
&\quad + \left[\frac{\sigma^2 d_n(\alpha_\lambda)}{n\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))} \frac{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))} \frac{\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))} \right] \left| \frac{||\boldsymbol{\varepsilon}_n||^2}{n\sigma^2} - 1 \right| + \frac{d_n(\alpha_\lambda)}{n} \frac{||\hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda) - \hat{\boldsymbol{\mu}}_n(\lambda)||^2}{nL(\hat{\boldsymbol{\beta}}_n(\lambda))}.
\end{aligned}
$$

By definition, $\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda)) \geq \sigma^2 d_n(\alpha_\lambda)/n$. Thus

$$
\begin{aligned}
\frac{|\hat{\sigma}_n^2(\lambda) - \sigma^2| d_n(\alpha_\lambda)}{nL(\hat{\boldsymbol{\beta}}_n(\lambda))} &\leq \sup_{\lambda \in [0, \lambda_{max}]} \frac{d_n(\alpha_\lambda)}{n} \left[\frac{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))} \right] \\
&\quad + \frac{2}{\sigma} \left[\frac{d_n(\alpha_\lambda)}{n} \frac{||\boldsymbol{\varepsilon}_n||^2}{n} \right]^{1/2} \left[\frac{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))} \right]^{1/2} \left[\frac{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))} \frac{\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))} \right]^{1/2} \\
&\quad + \left[\frac{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))} \frac{\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))} \right] \left| \frac{||\boldsymbol{\varepsilon}_n||^2}{n\sigma^2} - 1 \right| + \frac{d_n(\alpha_\lambda)}{n} \frac{||\hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda) - \hat{\boldsymbol{\mu}}_n(\lambda)||^2}{nL(\hat{\boldsymbol{\beta}}_n(\lambda))}.
\end{aligned}
$$

Li (1987) established that

$$\sup_{\alpha \in \mathcal{A}_n} \left| \frac{L(\hat{\boldsymbol{\beta}}_n^*(\alpha))}{R(\hat{\boldsymbol{\beta}}_n^*(\alpha))} - 1 \right| \to_p 0$$

and it follows that

$$\sup_{\lambda \in [0, \lambda_{max}]} \left| \frac{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))} - 1 \right| \to_p 0. \tag{1}$$

In addition, from the proof of Theorem 2 in Zhang et al. (2010) we have that

$$\sup_{\lambda \in [0, \lambda_{max}]} \left| \frac{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda)) - L(\hat{\boldsymbol{\beta}}_n(\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))} \right| \to_p 0, \tag{2}$$

and

$$\sup_{\lambda \in [0, \lambda_{max}]} \frac{||\hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda) - \hat{\boldsymbol{\mu}}_n(\lambda)||^2}{nL(\hat{\boldsymbol{\beta}}_n(\lambda))} \to_p 0. \tag{3}$$

Combining these results with the Law of Large Numbers and the assumption that $d_n/n \to 0$ as $n \to \infty$ the right-hand side converges to 0 in probability. Hence,

$$\sup_{\lambda \in [0, \lambda_{max}]} \frac{2d_n(\alpha_\lambda)|\hat{\sigma}_n^2(\lambda) - \sigma^2|}{nL(\hat{\boldsymbol{\beta}}_n(\lambda))} \to_p 0$$

as desired. □

Lemma 2. *Assume that (A1)-(A4) hold and that $d_n/n \to 0$ as $n \to \infty$. Then*

$$\sup_{\lambda \in [0, \lambda_{max}]} \frac{d_n(\alpha_\lambda)|\tilde{\sigma}_n^2 - \sigma^2|}{nL(\hat{\beta}_n(\lambda))} \to_p 0.$$

Proof. Start by noting that for all $\lambda \in [0, \lambda_{max}]$, $\Delta_n(\alpha_\lambda) \geq \Delta_n(\bar{\alpha})$ (Zhang et al., 2010). Consider

$$\frac{\tilde{R}(\hat{\boldsymbol{\beta}}_n(\bar{\alpha}))d_n(\alpha_\lambda)}{\tilde{R}(\hat{\boldsymbol{\beta}}_n(\alpha_\lambda))d_n} \leq \frac{(\Delta_n(\bar{\alpha}) + \frac{d_n)\sigma^2}{n})d_n(\alpha_\lambda)}{(\Delta_n(\bar{\alpha}) + \frac{d_n(\alpha_\lambda)\sigma^2}{n})d_n} \leq \frac{\Delta_n(\bar{\alpha})}{\Delta_n(\bar{\alpha}) + \frac{d_n(\alpha_\lambda)\sigma^2}{n}} + \frac{\frac{d_n\sigma^2}{n}d_n(\alpha_\lambda)}{\frac{d_n(\alpha_\lambda)\sigma^2}{n}d_n} \leq 2.$$

Now, from the proof of Lemma 1 we have that

$$|\tilde{\sigma}_n^2 - \sigma^2| \leq \frac{n}{n-d_n-1} L(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha})) + 2\frac{n}{n-d_n-1}||\boldsymbol{\varepsilon}_n|| \frac{||\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}^*(\bar{\alpha})||}{n} + \left| \frac{||\boldsymbol{\varepsilon}_n||^2}{n-d_n-1} - \sigma^2 \right|.$$

Thus

$$\begin{aligned}\frac{d_n|\tilde{\sigma}_n^2 - \sigma^2|}{nL(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha}))} \leq{}& \frac{n}{n-d_n-1}\frac{d_n}{n} + \frac{2}{\sigma}\frac{n}{n-d_n-1}\left[\frac{||\boldsymbol{\varepsilon}_n||^2}{n}\frac{d_n}{n}\right]^{1/2}\left[\frac{\sigma^2 d_n}{n\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha}))}\frac{\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha}))}{L(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha}))}\right]^{1/2} \\ &+ \left[\frac{d_n\sigma^2}{n\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha}))}\frac{\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha}))}{L(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha}))}\right]\left|\frac{||\boldsymbol{\varepsilon}_n||^2}{(n-d_n-1)\sigma^2} - 1\right|.\end{aligned}$$

Under the assumption that $d_n/n \to 0$ as $n \to \infty$ it follows that

$$\frac{d_n|\tilde{\sigma}_n^2 - \sigma^2|}{nL(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha}))} \to_p 0.$$

Combining these results with (1) and (2) it follows that

$$\begin{aligned}\frac{d_n(\alpha_\lambda)|\tilde{\sigma}_n^2 - \sigma^2|}{nL(\hat{\beta}_n(\lambda))} &\leq \sup_{[0,\lambda_{max}]} \frac{d_n|\tilde{\sigma}_n^2 - \sigma^2|}{nL(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha}))}\frac{d_n(\alpha_\lambda)\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha}))}{d_n\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}\frac{L(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha}))}{\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha}))}\frac{\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}\frac{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))} \\ &\leq 2\frac{d_n|\tilde{\sigma}_n^2 - \sigma^2|}{nL(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha}))} \sup_{[0,\lambda_{max}]} \frac{L(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha}))}{\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\bar{\alpha}))}\frac{\tilde{R}(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}\frac{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))} \to_p 0.\end{aligned}$$

□

Proof of Theorem 1. As in the proofs in Zhang et al. (2010), to prove that C_{p_λ} is asymptotically loss efficient, it is sufficient to show that

$$\sup_{\lambda\in[0,\lambda_{max}]} \left| \frac{C_{p_\lambda} - ||\boldsymbol{\varepsilon}_n||^2/n - L(\hat{\boldsymbol{\beta}}_n(\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))} \right| \to_p 0. \tag{4}$$

Decomposing C_{p_λ} it can be established that

$$\begin{aligned}
C_{p_\lambda} &= \frac{||\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n(\lambda)||^2}{n} + \frac{2\tilde{\sigma}_n^2 d_n(\alpha_\lambda)}{n} \\
&= \frac{||\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda)||^2}{n} + \frac{||\hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda) - \hat{\boldsymbol{\mu}}_n(\lambda)||^2}{n} + \frac{2\tilde{\sigma}_n^2 d_n(\alpha_\lambda)}{n} \\
&= \frac{||\boldsymbol{\varepsilon}_n||^2}{n} + L(\hat{\boldsymbol{\beta}}_n(\lambda)) + (L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda)) - L(\hat{\boldsymbol{\beta}}_n(\lambda))) + \frac{||\hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda) - \hat{\boldsymbol{\mu}}_n(\lambda)||^2}{n} \\
&+ \frac{2\boldsymbol{\varepsilon}_n^T(I - \mathbf{H}_n(\alpha_\lambda))\boldsymbol{\mu}_n}{n} + \frac{2(\sigma^2 d_n(\alpha_\lambda) - \boldsymbol{\varepsilon}_n^T \mathbf{H}_n(\alpha_\lambda)\boldsymbol{\varepsilon}_n)}{n} + \frac{2(\tilde{\sigma}_n^2 - \sigma^2)d_n(\alpha_\lambda)}{n}.
\end{aligned}$$

The proof of Theorem 2 in Zhang et al. (2010) established that

$$\sup_{\lambda \in [0, \lambda_{max}]} \left| \frac{2\boldsymbol{\varepsilon}_n^T(I - \mathbf{H}_n(\alpha_\lambda))\boldsymbol{\mu}_n}{nL(\hat{\boldsymbol{\beta}}_n(\lambda))} \right| \to_p 0, \tag{5}$$

and,

$$\sup_{\lambda \in [0, \lambda_{max}]} \left| \frac{2(\sigma^2 d_n(\alpha_\lambda) - \boldsymbol{\varepsilon}_n^T \mathbf{H}_n(\alpha_\lambda)\boldsymbol{\varepsilon}_n)}{nL(\hat{\boldsymbol{\beta}}_n(\lambda))} \right| \to_p 0. \tag{6}$$

Combining these results with (1)-(3) and Lemma 2, (4) follows as desired. □

Proof of Theorem 2. The proof is the same as that of Theorem 1 except that the estimated variance is based on the candidate model rather than the full model and the result is established by using Lemma 1 in place of Lemma 2. □

Proof of Theorem 3. As in the efficiency proof for $\Gamma_n(\lambda)$, it is sufficient to show that

$$\sup_{\lambda \in [0, \lambda_{max}]} \left| \frac{\tilde{\Gamma}_n(\lambda) - ||\boldsymbol{\varepsilon}_n||^2/n - L(\hat{\boldsymbol{\beta}}_n(\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))} \right| \to_p 0.$$

to establish that $\tilde{\Gamma}_n(\lambda)$ is an asymptotically efficient selection procedure for the regularization

parameter, λ. By the definition of $\tilde{\Gamma}_n(\lambda)$ we have that,

$$\begin{aligned}\sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{\tilde{\Gamma}_n(\lambda)-||\boldsymbol{\varepsilon}_n||^2/n-L(\hat{\boldsymbol{\beta}}_n(\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))}\right| &= \sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{\delta_n(\lambda)\hat{\sigma}_n^2(\lambda)+\Gamma_n(\lambda)-||\boldsymbol{\varepsilon}_n||^2/n-L(\hat{\boldsymbol{\beta}}_n(\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))}\right| \\ &\leq \sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{\delta_n(\lambda)(\hat{\sigma}_n^2(\lambda)-\sigma^2)}{L(\hat{\boldsymbol{\beta}}_n(\lambda))}\right| + \sup_{\lambda\in[0,\lambda_{max}]}\frac{|\delta_n(\lambda)|\sigma^2}{L(\hat{\boldsymbol{\beta}}_n(\lambda))} \\ &+ \sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{\Gamma_n(\lambda)-||\boldsymbol{\varepsilon}_n||^2/n-L(\hat{\boldsymbol{\beta}}_n(\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))}\right|\end{aligned}$$

The last two terms converge to zero by (C1) and the efficiency proof for $\Gamma_n(\lambda)$. From the proof of Lemma 1 we further have that

$$\begin{aligned}\left|\frac{\delta_n(\lambda)(\hat{\sigma}_n^2(\lambda)-\sigma^2)}{L(\hat{\boldsymbol{\beta}}_n(\lambda))}\right| &\leq |\delta_n(\lambda)|\frac{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))} + 2\frac{||\boldsymbol{\varepsilon}_n||}{\sqrt{n}}\left(\frac{L(\hat{\boldsymbol{\beta}}_n^*(\alpha_\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))}\right)^{1/2}\left(\frac{|\delta_n(\lambda)|}{L(\hat{\boldsymbol{\beta}}_n(\lambda))}\right)^{1/2}(|\delta_n(\lambda)|)^{1/2} \\ &+ \frac{|\delta_n(\lambda)|}{L(\hat{\boldsymbol{\beta}}_n(\lambda))}\left|\frac{||\boldsymbol{\varepsilon}_n||^2}{n}-\sigma^2\right| + |\delta_n(\lambda)|\frac{||\hat{\boldsymbol{\mu}}_n^*(\alpha_\lambda)-\hat{\boldsymbol{\mu}}_n(\lambda)||^2}{nL(\hat{\boldsymbol{\beta}}_n(\lambda))}\end{aligned}$$

By (C1), (C2), and similar arguments as those used in the proof of Lemma 1 we have that the right hand side converges to zero in probability. Therefore, it follows that

$$\sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{\delta_n(\lambda)(\hat{\sigma}_n^2(\lambda)-\sigma^2)}{L(\hat{\boldsymbol{\beta}}_n(\lambda))}\right| \to_p 0$$

and so

$$\sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{\tilde{\Gamma}_n(\lambda)-||\boldsymbol{\varepsilon}_n||^2/n-L(\hat{\boldsymbol{\beta}}_n(\lambda))}{L(\hat{\boldsymbol{\beta}}_n(\lambda))}\right| \to_p 0.$$

□

Verifying the Conditions of Theorem 3

This following shows that AIC_λ, GCV_λ, and AIC_{c_λ} can be written in the form $\tilde{\Gamma}_n(\lambda)$ and that Conditions (C1) and (C2) of Theorem 3 are satisfied. This implies that the three methods are efficient selectors of the regularization parameter. Shibata (1981) and Hurvich and Tsai

(1989) noted that AIC and AIC_c, respectively, can be shown to satisfy these conditions. We present a detailed argument of these remarks below.

AIC_λ is Efficient

Minimizing AIC_λ is equivalent to minimizing

$$\exp\left(\frac{2d_n(\alpha_\lambda)}{n}\right)\hat{\sigma}_n^2(\lambda).$$

Taylor expanding,

$$\begin{aligned}\exp\left(\frac{2d_n(\alpha_\lambda)}{n}\right)\hat{\sigma}_n^2(\lambda) &= \sum_{k=0}^{\infty}\left(\frac{2d_n(\alpha_\lambda)}{n}\right)^k\frac{1}{k!}\\ &= 1+\frac{2d_n(\alpha_\lambda)}{n}+\sum_{k=2}^{\infty}\left(\frac{2d_n(\alpha_\lambda)}{n}\right)^k\frac{1}{k!},\end{aligned}$$

we see that AIC_λ has the same asymptotic properties as

$$\tilde{\Gamma}_n(\lambda)=\hat{\sigma}_n^2(\lambda)\left(1+2\frac{d_n(\alpha_\lambda)}{n}+\delta_n(\lambda)\right)$$

where

$$\delta_n(\lambda)=\sum_{k=2}^{\infty}\left(\frac{2d_n(\alpha_\lambda)}{n}\right)^k\frac{1}{k!}.$$

Therefore, the efficiency of AIC_λ can be established by showing that (C1) and (C2) hold. Consider

$$0<\delta_n(\lambda)=\sum_{k=2}^{\infty}\left(\frac{2d_n(\alpha_\lambda)}{n}\right)^k\frac{1}{k!}=\exp\left(\frac{2d_n(\alpha_\lambda)}{n}\right)-1-\frac{2d_n(\alpha_\lambda)}{n}$$

Therefore, under the assumption that $d_n/n \to 0$, (C1) is satisfied. Next consider

$$\begin{aligned} 0 < \frac{\delta_n(\lambda)}{\tilde{R}(\hat{\beta}_n^*(\alpha_\lambda))} &= \sum_{k=2}^{\infty} \left(\frac{2d_n(\alpha_\lambda)}{n}\right)^k \frac{1}{\tilde{R}(\hat{\beta}_n^*(\alpha_\lambda))k!} \\ &\leq \frac{2}{\sigma^2} \sum_{k=2}^{\infty} \left(\frac{2d_n(\alpha_\lambda)}{n}\right)^{k-1} \frac{1}{k!} \leq \frac{2}{\sigma^2} \sum_{k=2}^{\infty} \left(\frac{2d_n}{n}\right)^{k-1} \frac{1}{(k-1)!} \\ &= \frac{2}{\sigma^2} \sum_{k=1}^{\infty} \left(\frac{2d_n}{n}\right)^{k} \frac{1}{k!} = \frac{2}{\sigma^2} \left(\exp\left(\frac{2d_n}{n}\right) - 1\right) \to 0. \end{aligned}$$

Here the inequality on the second line follows from the fact that $R(\hat{\beta}_n^*(\alpha_\lambda)) > \sigma^2 d_n(\alpha_\lambda)/n$ and the final result follows from the assumption that $d_n/n \to 0$. Therefore,

$$\sup_{\lambda \in [0, \lambda_{max}]} \frac{|\delta_n(\lambda)|}{L(\hat{\beta}_n(\lambda))} = \sup_{\lambda \in [0, \lambda_{max}]} \left| \frac{L(\hat{\beta}_n^*(\alpha_\lambda))}{L(\hat{\beta}_n(\lambda))} \frac{\tilde{R}(\hat{\beta}_n^*(\alpha_\lambda))}{L(\hat{\beta}_n^*(\alpha_\lambda))} \frac{\delta_n(\lambda)}{\tilde{R}(\hat{\beta}_n^*(\alpha_\lambda))} \right| \to_p 0$$

so (C2) is satisfied.

GCV_λ is Efficient

Taylor expanding,

$$\frac{1}{(1 - d_n(\alpha_\lambda)/n)^2} = \sum_{k=1}^{\infty} k \left(\frac{d_n(\alpha_\lambda)}{n}\right)^{k-1} = 1 + \frac{2d_n(\alpha_\lambda)}{n} + \sum_{k=3}^{\infty} k \left(\frac{d_n(\alpha_\lambda)}{n}\right)^{k-1},$$

we see that GCV_λ has the same asymptotic properties as

$$\tilde{\Gamma}_n(\lambda) = \hat{\sigma}_n^2(\lambda) \left(1 + 2\frac{d_n(\alpha_\lambda)}{n} + \delta_n(\lambda)\right)$$

where

$$\delta_n(\lambda) = \sum_{k=3}^{\infty} k \left(\frac{d_n(\alpha_\lambda)}{n}\right)^{k-1}.$$

Therefore, the efficiency of GCV_λ can be established by showing that (C1) and (C2) hold. Consider

$$0 < \delta_n(\lambda) = \sum_{k=3}^{\infty} k \left(\frac{d_n(\alpha_\lambda)}{n}\right)^{k-1} = \frac{1}{(1 - d_n(\alpha_\lambda)/n)^2} - 1 - \frac{2d_n(\alpha_\lambda)}{n}$$

Therefore, under the assumption that $d_n/n \to 0$, (C1) is satisfied. Next consider,

$$\begin{aligned}
0 < \frac{\delta_n(\lambda)}{\tilde{R}(\hat{\beta}_n^*(\alpha_\lambda))} &= \sum_{k=3}^{\infty} k \left(\frac{d_n(\alpha_\lambda)}{n}\right)^{k-1} \frac{1}{\tilde{R}(\hat{\beta}_n^*(\alpha_\lambda))} \\
&\leq \frac{1}{\sigma^2} \sum_{k=3}^{\infty} k \left(\frac{d_n(\alpha_\lambda)}{n}\right)^{k-2} \\
&= \frac{1}{\sigma^2} \left(\sum_{k=3}^{\infty} (k-1) \left(\frac{d_n(\alpha_\lambda)}{n}\right)^{k-2} + \sum_{k=3}^{\infty} \left(\frac{d_n(\alpha_\lambda)}{n}\right)^{k-2} \right) \\
&= \frac{1}{\sigma^2} \left(\sum_{k=2}^{\infty} k \left(\frac{d_n(\alpha_\lambda)}{n}\right)^{k-1} + \frac{d_n(\alpha_\lambda)}{n} \sum_{k=0}^{\infty} \left(\frac{d_n(\alpha_\lambda)}{n}\right)^{k} \right) \\
&= \frac{1}{\sigma^2} \left(\frac{1}{(1 - d_n(\alpha_\lambda)/n)^2} - 1 + \frac{d_n(\alpha_\lambda)/n}{1 - d_n(\alpha_\lambda)/n} \right)
\end{aligned}$$

which converges to zero uniformly over λ under the assumption that $d_n/n \to 0$. Here, again, the inequality on the second line follows from the fact that $\tilde{R}(\hat{\beta}_n^*(\alpha_\lambda)) > \sigma^2 d_n(\alpha_\lambda)/n$. Therefore,

$$\sup_{\lambda \in [0, \lambda_{max}]} \frac{|\delta_n(\lambda)|}{L(\hat{\beta}_n(\lambda))} = \sup_{\lambda \in [0, \lambda_{max}]} \left| \frac{L(\hat{\beta}_n^*(\alpha_\lambda))}{L(\hat{\beta}_n(\lambda))} \frac{\tilde{R}(\hat{\beta}_n^*(\alpha_\lambda))}{L(\hat{\beta}_n^*(\alpha_\lambda))} \frac{\delta_n(\lambda)}{\tilde{R}(\hat{\beta}_n^*(\alpha_\lambda))} \right| \to_p 0$$

so (C2) is satisfied.

AIC_{c_λ} is Efficient

We define

$$AIC_{c_\lambda} = \log(\hat{\sigma}_n^2(\lambda)) + 2 \frac{d_n(\alpha_\lambda) + 1}{n - d_n(\alpha_\lambda) - 2}.$$

This can be equivalently defined as,

$$AIC_{c_\lambda} = \log(\hat{\sigma}_n^2(\lambda)) + 2 \frac{d_n(\alpha_\lambda) + 1}{n} + 2 \frac{(d_n(\alpha_\lambda) + 1)(d_n(\alpha_\lambda) + 2)}{n(n - d_n(\alpha_\lambda) - 2)}.$$

Based on the second definition of AIC_{c_λ} we see that the information criterion has the same asymptotic properties as,

$$\log(\hat{\sigma}_n^2(\lambda)) + 2 \frac{d_n(\alpha_\lambda)}{n} + 2 \frac{(d_n(\alpha_\lambda) + 1)(d_n(\alpha_\lambda) + 2)}{n(n - d_n(\alpha_\lambda) - 2)}.$$

because they only differ by an additive constant $(2/n)$. Therefore, AIC_{c_λ} will have the same asymptotic behavior as

$$\exp\left(2\frac{d_n(\alpha_\lambda)}{n}+2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}\right)\hat{\sigma}_n^2(\lambda).$$

Taylor expanding,

$$\begin{aligned}\exp\left(2\frac{d_n(\alpha_\lambda)}{n}+2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}\right) &= \sum_{k=0}^{\infty}\left(2\frac{d_n(\alpha_\lambda)}{n}+2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}\right)^k\frac{1}{k!}\\ &= 1+\frac{2d_n(\alpha_\lambda)}{n}+2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}\\ &\quad+\sum_{k=2}^{\infty}\left(2\frac{d_n(\alpha_\lambda)}{n}+2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}\right)^k\frac{1}{k!},\end{aligned}$$

we see that AIC_{c_λ} has the same asymptotic properties as

$$\tilde{\Gamma}_n(\lambda)=\hat{\sigma}_n^2(\lambda)\left(1+2\frac{d_n(\alpha_\lambda)}{n}+\delta_n(\lambda)\right)$$

where

$$\delta_n(\lambda)=2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}+\sum_{k=2}^{\infty}\left(2\frac{d_n(\alpha_\lambda)}{n}+2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}\right)^k\frac{1}{k!}.$$

Therefore, the efficiency of AIC_{c_λ} can be established by showing that (C1) and (C2) hold. Consider

$$\begin{aligned}0<\delta_n(\lambda) &= 2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}+\sum_{k=2}^{\infty}\left(2\frac{d_n(\alpha_\lambda)}{n}+2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}\right)^k\frac{1}{k!}\\ &= 2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}+\exp\left(2\frac{d_n(\alpha_\lambda)}{n}+2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}\right)\\ &\quad-1-2\frac{d_n(\alpha_\lambda)}{n}-2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}\end{aligned}$$

which converges to zero uniformly over λ under the assumption that $d_n/n \to 0$. Thus, (C1) is satisfied. Next consider

$$
\begin{aligned}
0 < \frac{\delta_n(\lambda)}{\tilde{R}(\hat{\beta}_n^*(\alpha_\lambda))} &= 2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{\tilde{R}(\hat{\beta}_n^*(\alpha_\lambda))n(n-d_n(\alpha_\lambda)-2)} \\
&\quad + \sum_{k=2}^{\infty}\left(2\frac{d_n(\alpha_\lambda)}{n} + 2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}\right)^k \frac{1}{\tilde{R}(\hat{\beta}_n^*(\alpha_\lambda))k!} \\
&\leq 2\frac{(1+1/d_n(\alpha_\lambda))(d_n(\alpha_\lambda)+2)}{\sigma^2(n-d_n(\alpha_\lambda)-2)} \\
&\quad + \frac{n}{\sigma^2 d_n(\alpha_\lambda)}\sum_{k=2}^{\infty}\left(2\frac{d_n(\alpha_\lambda)}{n} + 2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}\right)^k \frac{1}{k!} \\
&\leq 2\frac{(1+1/d_n(\alpha_\lambda))(d_n(\alpha_\lambda)+2)}{\sigma^2(n-d_n(\alpha_\lambda)-2)} \\
&\quad + \frac{2}{\sigma^2}\left(1+\frac{(1+1/d_n(\alpha_\lambda))(d_n(\alpha_\lambda)+2)}{(n-d_n(\alpha_\lambda)-2)}\right)\sum_{k=2}^{\infty}\left(2\frac{d_n(\alpha_\lambda)}{n} + 2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}\right)^{k-1} \frac{1}{k!} \\
&\leq 2\frac{(1+1/d_n(\alpha_\lambda))(d_n(\alpha_\lambda)+2)}{\sigma^2(n-d_n(\alpha_\lambda)-2)} \\
&\quad + \frac{2}{\sigma^2}\left(1+\frac{(1+1/d_n(\alpha_\lambda))(d_n(\alpha_\lambda)+2)}{(n-d_n(\alpha_\lambda)-2)}\right)\sum_{k=1}^{\infty}\left(2\frac{d_n(\alpha_\lambda)}{n} + 2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}\right)^{k} \frac{1}{k!} \\
&= 2\frac{(1+1/d_n(\alpha_\lambda))(d_n(\alpha_\lambda)+2)}{\sigma^2(n-d_n(\alpha_\lambda)-2)} \\
&\quad + \frac{2}{\sigma^2}\left(1+\frac{(1+1/d_n(\alpha_\lambda))(d_n(\alpha_\lambda)+2)}{(n-d_n(\alpha_\lambda)-2)}\right)\left(\exp\left(2\frac{d_n(\alpha_\lambda)}{n} + 2\frac{(d_n(\alpha_\lambda)+1)(d_n(\alpha_\lambda)+2)}{n(n-d_n(\alpha_\lambda)-2)}\right) - 1\right)
\end{aligned}
$$

which converges to zero uniformly over λ under the assumption that $d_n/n \to 0$. Again, the inequality on the third line follows from the fact that $R(\hat{\beta}_n^*(\alpha_\lambda)) > \sigma^2 d_n(\alpha_\lambda)/n$. Therefore,

$$
\sup_{\lambda\in[0,\lambda_{max}]}\frac{|\delta_n(\lambda)|}{L(\hat{\beta}_n(\lambda))} = \sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{L(\hat{\beta}_n^*(\alpha_\lambda))}{L(\hat{\beta}_n(\lambda))}\frac{\tilde{R}(\hat{\beta}_n^*(\alpha_\lambda))}{L(\hat{\beta}_n^*(\alpha_\lambda))}\frac{\delta_n(\lambda)}{\tilde{R}(\hat{\beta}_n^*(\alpha_\lambda))}\right| \to_p 0
$$

so (C2) is satisfied.

Theorem 4

The following theorem demonstrates that, in classical regression, the probability that an information criterion selects the full model over the true model depends on the form of the

criterion, the true number of predictors, the sample size, and the number of predictors that are included in the full model.

Theorem 1. *Assume the true model is the linear model described as,*

$$\mathbf{y_n} = \mathbf{X_n}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n$$

where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. *Further assume that only* d_0 *entries of* $\boldsymbol{\beta}$ *are non-zero and that* α_0 *is the index of the true minimal model. Then, in classical regression, the probability that an information criterion,*

$$IC_n(\alpha) = \log(\hat{\sigma}_n^2(\alpha)) + pen_n(\alpha),$$

selects the full model, $\bar{\alpha}$, *over the correct model is computed as*

$$\Pr(IC_n(\alpha_0) > IC_n(\bar{\alpha})) = \Pr\left(F(d_n - d_0, n - d_n) > \frac{n - d_n}{d_n - d_0}(\exp(pen_n(\bar{\alpha}) - pen_n(\alpha_0)) - 1)\right)$$

where $F(\nu_1, \nu_2)$ *denotes an* F *random variable with* ν_1 *and* ν_2 *degrees of freedom.*

Proof of Theorem 4. Let $RSS_n(\alpha) = ||\mathbf{y}_n - \hat{\mu}_n(\alpha)||^2$. Then

$$\begin{aligned}
\Pr(\mathrm{IC}_n(\alpha_0) > \mathrm{IC}_n(\bar{\alpha})) &= \Pr\left(\log\left(\frac{\hat{\sigma}_n^2(\alpha_0)}{\hat{\sigma}_n^2(\bar{\alpha})}\right) > \mathrm{pen}_n(\bar{\alpha}) - \mathrm{pen}_n(\alpha_0)\right) \\
&= \Pr\left(\frac{\hat{\sigma}_n^2(\alpha_0)}{\hat{\sigma}_n^2(\bar{\alpha})} > \exp(\mathrm{pen}_n(\bar{\alpha}) - \mathrm{pen}_n(\alpha_0))\right) \\
&= \Pr\left(\frac{RSS_n(\alpha_0) - RSS_n(\bar{\alpha})}{RSS_n(\bar{\alpha})} > \exp(\mathrm{pen}_n(\bar{\alpha}) - \mathrm{pen}_n(\alpha_0)) - 1\right) \\
&= \Pr\left(\frac{(RSS_n(\alpha_0) - RSS_n(\bar{\alpha}))/(d_n - d_0)}{RSS_n(\bar{\alpha})/(n - d_n)} > \frac{n - d_n}{d_n - d_0}(\exp(\mathrm{pen}_n(\bar{\alpha}) - \mathrm{pen}_n(\alpha_0)) - 1)\right) \\
&= \Pr\left(F(d_n - d_0, n - d_n) > \frac{n - d_n}{d_n - d_0}(\exp(\mathrm{pen}_n(\bar{\alpha}) - \mathrm{pen}_n(\alpha_0)) - 1)\right).
\end{aligned}$$

□

References

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, 76(2):297–307.

Li, K.-C. (1987). Asymptotic Optimality for Cp, CL, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics*, 15(3):958–975.

Shibata, R. (1981). An Optimal Selection of Regression Variables. *Biometrika*, 68(1):45–54.

Zhang, Y., Li, R., and Tsai, C.-L. (2010). Regularization Parameter Selections via Generalized Information Criterion. *Journal of the American Statistical Association*, 105(489):312–323.