CONFIDENCE SETS FOR MODEL SELECTION BY F-TESTING

DAVIDE FERRARI AND YUHONG YANG
Universtiy of Melbourne and University of Minnesota

Supplementary Material

In this supplementary document, we give technical proofs for theorems and corollaries for the paper "Confidence sets for model selection by F-testing". The main assumptions and notations can be found in the main paper.

S1 Proof of Theorem 2.3

For the necessary condition, we just need to show that if there is a sequence of $\gamma_n \in \Gamma_u$ such that $\delta_{\gamma_n}/\sqrt{df_{\gamma_n} - df_{\gamma_f}}$ is uniformly bounded by $C > 0$, then the corresponding F-statistic stays below the cutoff value with a non-vanishing probability. Without loss of generality, assume $\sigma^2 = 1$. For the model γ_n, the F-statistic has the non-central F-distribution $F_{\nu_1,\nu_2,\delta_{\gamma_n}}$, with degrees of freedom $\nu_1 = df_{\gamma_n} - df_{\gamma_f} = p - p_{\gamma_n}$ and $\nu_2 = df_{\gamma_f} = n - p - 1$ and non-centrality parameter δ_{γ_n}. Since $RSS_{\gamma_f} \sim X^2_{\nu_2}$ and if $\nu_2 \to \infty$, we have

$$\sqrt{\frac{\nu_2}{2}}\left(\frac{RSS_{\gamma_f}}{\nu_2} - 1\right) \xrightarrow{d} N(0,1).$$

Therefore RSS_{γ_f}/ν_2 is bounded away from zero and infinity in probability. Let f^*_γ denote the cut-off point for the F-ratio $F_{\left(df_\gamma - df_{\gamma_f}\right), df_{\gamma_f}}(\alpha)$. For the numerator of the F-statistic, since $RSS_{\gamma_n} - RSS_{\gamma_f}$ follows a non-central chi-squared distribution with ν_1 degrees of freedom and non-centrality parameter δ_{γ_n}, we have $[RSS_{\gamma_n} - RSS_{\gamma_f} - (\nu_1 + \delta_{\gamma_n})]/\sqrt{2(\nu_1 + 2\delta_{\gamma_n})} \xrightarrow{d} N(0,1)$, when either ν_1 or $\delta_{\gamma_n} \to \infty$. Thus when $\nu_1 \to \infty$,

$$\frac{\nu_1}{\sqrt{2(\nu_1 + 2\delta_{\gamma_n})}}\left(\frac{RSS_{\gamma_n} - RSS_{\gamma_f}}{\nu_1} - \frac{\nu_1 + \delta_{\gamma_n}}{\nu_1}\right) \xrightarrow{d} N(0,1).$$

For the F-test, we have

$$\begin{aligned}
& P\left(\frac{(RSS_{\gamma_n} - RSS_{\gamma_f})/\left(df_{\gamma_n} - df_{\gamma_f}\right)}{RSS_{\gamma_f}/df_{\gamma_f}} \leq f^*_{\gamma_n}\right) \\
\geq\ & P\left(\left[\frac{RSS_{\gamma_n} - RSS_{\gamma_f}}{df_{\gamma_n} - df_{\gamma_f}} \leq f^*_{\gamma_n}\right] \cap \left[RSS_{\gamma_f}/df_{\gamma_f} \geq 1\right]\right) \\
=\ & P\left(\frac{\nu_1}{\sqrt{2(\nu_1 + 2\delta_{\gamma_n})}}\left(\frac{RSS_{\gamma_n} - RSS_{\gamma_f}}{\nu_1} - \frac{\nu_1 + \delta_{\gamma_n}}{\nu_1}\right) \leq \frac{\nu_1}{\sqrt{2(\nu_1 + 2\delta_{\gamma_n})}}\left(f^*_{\gamma_n} - \frac{\nu_1 + \delta_{\gamma_n}}{\nu_1}\right)\right) \\
& \times P\left(RSS_{\gamma_f}/df_{\gamma_f} \geq 1\right).
\end{aligned}$$

When ν_2 is of order n (so that ν_1 is bounded above by a multiple of ν_2), from Theorem A (due to Laurent and Massart [2000]) and Theorems 5.1 and 5.2 in Inglot [2010], it can be shown that $f^*_{\gamma_n} \geq 1 + \tau_\alpha/\sqrt{\nu_1}$ for some constant $\tau_\alpha > 0$, with $\tau_\alpha \to \infty$ as $\alpha \to 0$.

Thus, as long as $\delta_{\gamma_n}/\sqrt{df_{\gamma_n} - df_{\gamma_f}}$ is uniformly upper bounded, $\frac{\nu_1}{\sqrt{2(\nu_1 + 2\delta_\gamma)}}\left(f^*_{\gamma_n} - \frac{\nu_1 + \delta_\gamma}{\nu_1}\right) \geq 0$ when α is small enough. Together with that $P\left(RSS_{\gamma_f}/df_{\gamma_f} \geq 1\right)$ is bounded away from zero, regardless of whether $\nu_1 \to \infty$ or not, we know $\frac{(RSS_{\gamma_n} - RSS_{\gamma_f})/\left(df_{\gamma_n} - df_{\gamma_f}\right)}{RSS_{\gamma_f}/df_{\gamma_f}}$ has a non-vanishing probability to be smaller than $F_{\left(df_{\gamma_n} - df_{\gamma_f}\right), df_{\gamma_f}}(\alpha)$, and thus γ_n is included in the confidence set with a non-vanishing probability. This completes the proof of the necessity condition for detectability of the true terms by the ECS.

Now for the sufficient condition, let $X_{\gamma,n} = \left(RSS_\gamma - RSS_{\gamma_f}\right)/\left(df_\gamma - df_{\gamma_f}\right)$, $Y_n = RSS_{\gamma_f}/df_{\gamma_f}$ and denote by f^*_γ the cut-off point for the F-ratio. We want to show that under the condition on δ_γ, we have

$$P\left(\bigcup_{\gamma \in \Gamma_u}\left\{\frac{X_{\gamma,n}}{Y_n} \leq f^*_\gamma\right\}\right) \to 0.$$

Again, from the result in Inglot [2010], we have that

$$f^*_\gamma \leq \frac{\left[\nu_1 + 2\log\left(\frac{2}{\alpha}\right) + 2\sqrt{\nu_1 \log\left(\frac{2}{\alpha}\right)}\right]/\nu_1}{\left[\nu_2 + 2\log\left(\frac{2}{\alpha}\right) + \frac{1}{4}\sqrt{\nu_2 \log\left(\frac{2}{\alpha}\right)}\right]/\nu_2}.$$

Then,

$$P\left(\frac{X_{\gamma,n}}{Y_n} \leq f^*_\gamma\right) \leq P\left(\frac{X_{\gamma,n}}{Y_n} \leq \frac{\left[\nu_1 + 2\log\left(\frac{2}{\alpha}\right) + 2\sqrt{\nu_1 \log\left(\frac{2}{\alpha}\right)}\right]/\nu_1}{\left[\nu_2 + 2\log\left(\frac{2}{\alpha}\right) + \frac{1}{4}\sqrt{\nu_2 \log\left(\frac{2}{\alpha}\right)}\right]/\nu_2}\right) \tag{S1.1}$$

$$\leq P\left(X_{\gamma,n} \leq 1 + \frac{\tilde{\beta}_1}{\sqrt{\nu_1}} + \frac{\tilde{\beta}_2\sqrt{\eta_n}}{\sqrt{\nu_2}}\right) + P\left(Y_n \geq \frac{\nu_2 + 2\log\left(\frac{2}{\alpha}\right) + \frac{1}{4}\sqrt{\nu_2 \log\left(\frac{2}{\alpha}\right)\eta_n}}{\nu_2}\right), \tag{S1.2}$$

for any $\eta_n \to \infty$ and where $\tilde{\beta}_1$ and $\tilde{\beta}_2$ depending on α. With $\eta_n \to \infty$, the second probability in (S1.2) goes to zero as $n \to \infty$.

Next, we use a probability bound for the non-central chi-square distribution of Birgé [2001] to upper bound the probability of each event

$$\left\{X_{\gamma,n} \leq 1 + \frac{\tilde{\beta}_1}{\sqrt{\nu_1}} + \frac{\tilde{\beta}_2\sqrt{\eta_n}}{\sqrt{\nu_2}}\right\}.$$

Note that the maximum possible range of ν_1 is 1 to $n-1$, and there are no more than $\binom{p_n}{\nu_1} \leq$

$\exp\left(\nu_1 \log\left(\frac{ep_n}{\nu_1}\right)\right)$ models of size $p_n - \nu_1$ terms. From Lemma 8.1 in Birgé [2001], we have

$$\begin{aligned} &\sum_{\gamma\in\Gamma_u} P\left(X_{\gamma,n} \leq 1 + \frac{\tilde{\beta}_1}{\sqrt{\nu_1}} + \frac{\tilde{\beta}_2\sqrt{\eta_n}}{\sqrt{\nu_2}}\right) \\ &\leq \sum_{\nu_1=1}^{p-1} \exp\left\{-\min_{\gamma\in\Gamma_u, df_\gamma=n-1-(p-\nu_1)} \frac{(\delta_\gamma - \tilde{\beta}_1\sqrt{\nu_1} - \tilde{\beta}_2\sqrt{\eta_n}\nu_1/\sqrt{\nu_2})^2}{4(\nu_1+2\delta_\gamma)} + \nu_1\log\left(\frac{ep_n}{\nu_1}\right)\right\}. \end{aligned}$$

It is then sufficient to show that

$$\frac{(\delta_\gamma - \tilde{\beta}_1\sqrt{\nu_1} - \tilde{\beta}_2\sqrt{\eta_n}\nu_1/\sqrt{\nu_2})^2}{4(\nu_1+2\delta_\gamma)} \geq a\nu_1\log\left(\frac{ep_n}{\nu_1}\right) + \xi_n$$

for some constant $a > 1$ and some $\xi_n \to \infty$. When ν_2 is of order n, this requirement is satisfied if $\delta_\gamma \geq b\left(\sqrt{\nu_1\log\left(\frac{ep_n}{\nu_1}\right)} + \xi_n'\right)$ for some large enough constant $b > 0$ and some slowly increasing $\xi_n' \to \infty$.

Finally, we briefly show that the sufficient condition cannot be generally improved. Here we consider the case that $p_n \to \infty$. Recall that p_0 is the number of terms in the true model, which is assumed to satisfy that $\log p_0$ is of order $\log n$ and $p_0/p_n \to 0$. For notational ease, assume that the first p_0 terms are in the true model, and let γ_{-i} denote the model obtained from removing the i-th term in the true model for $1 \leq i \leq p_0$. Let $\nu_1 = p - p_0 + 1$ and for $1 \leq i \leq p_0$, let

$$A_i = \left\{\frac{(RSS_{\gamma_{-i}} - RSS_{\gamma_f})/\nu_1}{RSS_{\gamma_f}/\nu_2} \leq f^*_{\gamma_{-i}}\right\}.$$

To show non-detectability of the true terms, it suffices to show $P\left(\cup_{i=1}^{p_0} A_i\right)$ is bounded away from zero. Note that $P\left(\cup_{i=1}^{p_0} A_i\right)$ is lower bounded by

$$\begin{aligned} & P\left(\bigcup_{i=1}^{p_0}\left[\frac{RSS_{\gamma_{-i}} - RSS_{\gamma_f}}{\nu_1} \leq f^*_{\gamma_{-i}}\right] \cap \left[RSS_{\gamma_f}/df_{\gamma_f} \geq 1\right]\right) \\ = \quad & P\left(\bigcup_{i=1}^{p_0}\left[\frac{RSS_{\gamma_{-i}} - RSS_{\gamma_f}}{\nu_1} \leq f^*_{\gamma_{-i}}\right]\right) \times P\left(RSS_{\gamma_f}/df_{\gamma_f} \geq 1\right). \end{aligned}$$

Thus, it is sufficient to establish $P\left(\bigcap_{i=1}^{p_0}\left[\frac{RSS_{\gamma_{-i}} - RSS_{\gamma_f}}{\nu_1} \geq f^*_{\gamma_{-i}}\right]\right)$ is bounded away from 1. To proceed, consider the case that the true predictors are orthonormal with the same coefficient. Then the above probability is equal to the p_0-th power of $P\left(\left[\frac{RSS_{\gamma_{-i}} - RSS_{\gamma_f}}{\nu_1} \geq f^*_{\gamma_{-i}}\right]\right)$, which is upper bounded by $1 - P\left(RSS_{\gamma_{-i}} - RSS_{\gamma_f} \leq \nu_1\left(1 + \tau_\alpha/\sqrt{\nu_1}\right)\right)$, where $1 + \tau_\alpha/\sqrt{\nu_1}$ is a lower bound on $f^*_{\gamma_{-i}}$. Then, for $\delta_{\gamma_{-i}} = \sqrt{\nu_1\log(Cp_0)}$ for some constant $C > 0$, the moderate deviation probability $P\left(RSS_{\gamma_{-i}} - RSS_{\gamma_f} \leq \nu_1\left(1 + \tau_\alpha/\sqrt{\nu_1}\right)\right)$ is well-behaved and it is seen that the sought probability bound (away from 1) holds. This completes the proof of Theorem 2.3.

Proof of Corollary 3.1

Suppose that the ECS asymptotically detects all the true terms. Let A_n denote the event that all the models in $\widehat{\Gamma}$ are super models of γ^* (including itself). Then by the assumption, $P(A_n) \to 1$. Clearly, when $\gamma^* \in \widehat{\Gamma}$ and A_n holds, γ^* must be the unique model of $LBM(\widehat{\Gamma})$. Together with Theorem 2.1, the conclusion follows. The second statement holds similarly. This completes the proof of Corollary 3.1.

Proof of Corollary 3.4

From Theorem 2.3, we have $\liminf_{n\to\infty} P(LBM(\widehat{\Gamma}) = \{\gamma^*\}) \geq 1 - \alpha$. The statement on MEI thus holds. Also from Theorem 2.3, with probability going to 1, only the true and larger models will be included in $LBM(\widehat{\Gamma})$. Therefore, the variables in the true model will be included in all models in $LBM(\widehat{\Gamma})$ with probability going to 1. This completes the proof of Corollary 3.4.

Bibliography

L. Birgé. An alternative point of view on Lepski's method. In *State of the Art in Probability and Statistics*, pages 113–133. Institute of Mathematical Statistics, 2001.

T. Inglot. Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, 30:339–351, 2010.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.
biblio.bib