Discussion on "Model Confidence Bounds for Variable Selection" by Yang Li, Yuetian Luo, Davide Ferrari, Xiaonan Hu, and Yichen Qin

Leeb, Hannes and Pötscher, Benedikt M. and Kivaranovic, Danijel

Department of Statistics, University of Vienna

July 2018

Discussion on "Model Confidence Bounds for Variable Selection" by Yang Li, Yuetian Luo, Davide Ferrari, Xiaonan Hu, and Yichen Qin

Hannes Leeb[*,+]**, Benedikt M. Pötscher**[*]**, and Danijel Kivaranovic**[*]

[*]Department of Statistics, University of Vienna

[+]DataScience@UniVienna

e-mail: benedikt.poetscher@univie.ac.at

We congratulate the authors of Li et al. (2018) for their interesting paper, and we thank the Editor for the opportunity to comment on it. We shall offer some criticism of the proposed model confidence set in the following. Unless noted otherwise, we use the same notation as in Li et al. (2018). [Li et al. (2018) are not explicit about the nature of the regressor variables. We read their paper as considering nonstochastic regressors, and we note that similar results are possible for models with stochastic regressors. Also, a full column-rank assumption on the regressor matrix X seems to be missing, without which the symbol m^* is not well-defined. We hence add this assumption.]

Let us first strengthen Theorem 1, the main result of Li et al. (2018), a little bit. Recall that $\hat{r}(m_1, m_2) = B^{-1} \sum_{b=1}^{B} I(m_1 \subseteq \hat{m}^{(b)} \subseteq m_2)$ and define $\tilde{r}(m_1, m_2) = P\left(m_1 \subseteq \hat{m}^{(b)} \subseteq m_2 \| Y\right)$, where Y are the data. That is, $\tilde{r}(m_1, m_2)$ is the analogue of $\hat{r}(m_1, m_2)$, computed from the exact bootstrap distribution rather than from an approximation to it based on an i.i.d. bootstrap sample of size B (note that $\tilde{r}(m_1, m_2) = E(\hat{r}(m_1, m_2) \| Y)$, and thus is the quantity that $\hat{r}(m_1, m_2)$ is trying to approximate with the help of the bootstrap sample). Let $\tilde{m}_L$ and $\tilde{m}_U$ now be obtained from the exact bootstrap, i.e., $\tilde{m}_L$ and $\tilde{m}_U$ are defined by program (2) in Li et al. (2018), but with $\tilde{r}(m_1, m_2)$ replacing $\hat{r}(m_1, m_2)$ everywhere. To exclude trivial cases, we assume that the nominal confidence level satisfies $0 < 1 - \alpha < 1$. [Recall that $\hat{m}_L$ and $\hat{m}_U$ are defined by program (2) using $\hat{r}(m_1, m_2)$ as given in the paper.] We note that inspection of the proof of Theorem 1 in Li et al. (2018) reveals that this proof actually seems to be given for $\tilde{m}_L$ and $\tilde{m}_U$ rather than for $\hat{m}_L$ and $\hat{m}_U$ (although the theorem also holds for the latter as will transpire from the subsequent result). We are now ready for the improved version of Theorem 1 in Li et al. (2018).

PROPOSITION 1: Assume: (A.1) (Model selection consistency) $P(\hat{m} \neq m^*) = o(1)$; (A.2) (Bootstrap validity) For the re-sampled model $\hat{m}^{(b)}$, assume $P(\hat{m}^{(b)} \neq \hat{m}) = o(1)$.

(a) Then $P(\tilde{m}_L \subseteq m^* \subseteq \tilde{m}_U) = 1 + o(1)$ and $P(|\tilde{m}_U| - |\tilde{m}_L| = 0) = 1 + o(1)$; in fact, $P(\tilde{m}_U = \tilde{m}_L = \hat{m} = m^*) = 1 + o(1)$ holds. Moreover, assuming (A.1) only, this statement trivially continues to hold if $\tilde{m}_L$ and $\tilde{m}_U$ are both replaced by $\hat{m}$.

(b) Part (a) also holds if we replace $\tilde{m}_L$ by $\hat{m}_L$ and $\tilde{m}_U$ by $\hat{m}_U$ (where the number of bootstrap replications B may depend on n).

Proof. (a) (A.2) implies that $\tilde{r}(\hat{m}, \hat{m}) = P\left(\hat{m}^{(b)} = \hat{m} \| Y\right)$ converges to 1 in probability as $n \to \infty$. [To see this note that (A.2) implies $1 - P(\hat{m}^{(b)} = \hat{m}) = E(1 - P(\hat{m}^{(b)} = \hat{m} \| Y)) \to 0$ as $n \to \infty$. But this shows that $E(|1 - P(\hat{m}^{(b)} = \hat{m} \| Y)|)$ converges to zero, which even implies L_1-norm convergence of $P(\hat{m}^{(b)} = \hat{m} \| Y)$ to 1.] Obviously, this in turn also implies $\sum_{m \neq \hat{m}} \tilde{r}(m, m) \to 0$ in probability. Inspection of (2) in Li et al. (2018) (with $\tilde{r}(m_1, m_2)$ replacing $\hat{r}(m_1, m_2)$ everywhere) then shows that $\tilde{m}_L = \tilde{m}_U = \hat{m}$ holds on an event that has probability converging to 1. In view of (A.1), the event where $\tilde{m}_U = \tilde{m}_L = \hat{m} = m^*$ holds then also has probability converging to 1. This proves everything except for the last claim. For the last statement just note that the event $\{\hat{m} = m^*\}$ trivially has probability converging to one under (A.1).

(b) Consider first the case where B does not depend on n: Then, under (A.2), the event $A_n = \{\hat{m} = \hat{m}^{(1)} = \ldots = \hat{m}^{(B)}\}$, which is the finite intersection $\bigcap_{b=1}^{B} \{\hat{m}^{(b)} = \hat{m}\}$, has probability converging to one. Note that $\hat{r}(\hat{m}, \hat{m}) = 1$ and $\sum_{m \neq \hat{m}} \hat{r}(m, m) = 0$ hold on A_n. Inspection of (2) in Li et al. (2018) hence shows that $\hat{m}_L = \hat{m}_U = \hat{m}$ holds on A_n, which, together with (A.1), proves the first claim in Part (b) for fixed B. Consider next the case where $B = B_n$ diverges to infinity as $n \to \infty$: Similarly as in Part (a), it suffices to show that $\hat{r}(\hat{m}, \hat{m})$ converges to 1 in probability (noting that $\sum_{m \neq \hat{m}} \hat{r}(m, m) \to 0$ in probability then follows). Note that

$$
\begin{aligned}
\hat{r}(\hat{m}, \hat{m}) &= B_n^{-1} \sum_{b=1}^{B_n} \left\{ I(\hat{m}^{(b)} = \hat{m}) - E(I(\hat{m}^{(b)} = \hat{m}) \| Y) \right\} + E(I(\hat{m}^{(b)} = \hat{m}) \| Y) \\
&= B_n^{-1} \sum_{b=1}^{B_n} \left\{ I(\hat{m}^{(b)} = \hat{m}) - E(I(\hat{m}^{(b)} = \hat{m}) \| Y) \right\} + P(\hat{m}^{(b)} = \hat{m} \| Y).
\end{aligned}
$$

The final term on the right-hand side does not depend on b and coincides with $\tilde{r}(\hat{m}, \hat{m})$. It converges to 1 in probability as shown in the proof of Part (a). Now, conditional on Y, the first term on the right-hand side is the average of independent random variables that are centered versions of Bernoulli-distributed variables. Hence, conditional on Y, this term has zero expectation and variance bounded by $(4B_n)^{-1}$, which goes to zero as $n \to \infty$.

An application of Chebyshev's inequality shows that this term hence converges to zero in probability. This completes the proof of the first claim in case B_n diverges to infinity. The case of a general sequence B_n now follows by a standard subsequence argument. The proof is thus complete as the last claim reduces to the last claim in Part (a). ∎

Proposition 1 obviously strengthens and extends Theorem 1 of Li et al. (2018). However, this result also raises some obvious concerns: In the limit, the coverage probability of the model confidence set (MCS) defined by $\tilde{m}_L$ and $\tilde{m}_U$ ($\hat{m}_L$ and $\hat{m}_U$, respectively) is guaranteed to equal one (and not only to be $\geqslant 1 - \alpha$), and the 'length' of this MCS equals zero. In particular, asymptotically this MCS coincides with the set $\{\hat{m}\}$ and thus provides no more information than the point-estimate $\hat{m}$, which the MCS was intended to improve. Interestingly, the fact that the MCS asymptotically reduces to $\{\hat{m}\}$ under the assumptions of Proposition 1 seems to have informally been noticed by Li et al. (2018), at least in the context of model selection by the adaptive Lasso; see the discussion in Section 1.5 of the supplementary material of that paper. Apparently, however, this did not raise any red flags. [We also note that – in light of the proof of Part (a) given above – the proof of Theorem 1 in Li et al. (2018) seems to be much too complicated; in fact, it seems to be incorrect: For example, the final inequality in the first display in the proof of Theorem 1 appears to confuse the complement of $\{\hat{m}_L \subseteq \hat{m}\}$ with $\{\hat{m}_L \not\supsetneq \hat{m}\}$.]

It turns out that a *fixed-parameter* asymptotic analysis, as used in Proposition 1 and also in Theorem 1 of Li et al. (2018), is not appropriate here. In fixed-parameter asymptotics, the true parameter θ is kept fixed while sample size n increases to infinity. At this point, it seems fitting to repeat a warning issued a while ago by Hajek (1971, p. 153):

"Especially misinformative can be those limit results that are not uniform. Then the limit may exhibit some features that are not even approximately true for any finite n . . . "

In fact, in the presence of model selection, fixed-parameter asymptotic results not only *can*, but actually often *will*, mislead as has been amply documented in Leeb and Pötscher (2005, 2006a,b, 2008a,b,c); Pötscher (2009); Pötscher and Leeb (2009); Pötscher and Schneider (2009, 2011). This is also the case here: We shall provide a *uniform* asymptotic analysis

which reveals the misleading character of the pointwise asymptotic result above and in Theorem 1 of Li et al. (2018). To this end we have to amend the notation a bit: For given sample size n, true regression parameter $\theta \in \mathbb{R}^p$, and variance parameter $\sigma^2 \in (0, \infty)$ we shall write P_{n,θ,σ^2} for P to emphasize its dependence on the quantities indicated; similarly we shall write $m^*(\theta)$ to denote the smallest correct model for the parameter θ. [That is, P_{n,θ,σ^2} and $m^*(\theta)$ replace the symbols P and m^* used in Li et al. (2018).] Conditions (A.1) and (A.2) then become $P_{n,\theta,\sigma^2}(\hat{m} \neq m^*(\theta)) = o(1)$ and $P_{n,\theta,\sigma^2}(\hat{m}^{(b)} \neq \hat{m}) = o(1)$, respectively. Compared to Proposition 1 (and Theorem 1 of Li et al. (2018)) we also add a mild condition on the design matrix X.

PROPOSITION 2:   Assume (A.1) and (A.2) hold (for every $\theta \in \mathbb{R}^p$ and every $\sigma^2 \in (0, \infty)$), and assume that the sequence of matrices $X'X/n$ is bounded.

(a) Then for every $\sigma^2 \in (0, \infty)$ we have

$$\inf_{\theta \in \mathbb{R}^p} P_{n,\theta,\sigma^2}(\tilde{m}_L \subseteq m^*(\theta) \subseteq \tilde{m}_U) \quad = \quad o(1) \tag{1}$$

as $n \to \infty$. In other words, the minimal coverage probability of the MCS defined by $\tilde{m}_L$ and $\tilde{m}_U$ converges to zero. Moreover, assuming (A.1) only, this statement continues to hold if $\tilde{m}_L$ and $\tilde{m}_U$ are both replaced by $\hat{m}$.

(b) Part (a) also holds if we replace $\tilde{m}_L$ by $\hat{m}_L$ and $\tilde{m}_U$ by $\hat{m}_U$ (where the number of bootstrap replications B may depend on n).

Proof. (a) Without loss of generality we may assume that the probability space underlying P_{n,θ,σ^2} is given by $\mathbb{R}^n \times \mathcal{M}_{all}$, where $\mathbb{R}^n$ acts as the sample space for the data Y and $\mathcal{M}_{all}$ is the set of all models (and the σ-field is the product of the Borel-σ-field on $\mathbb{R}^n$ with the power-set of $\mathcal{M}_{all}$). The probability measure P_{n,θ,σ^2} can then be viewed as given by

$$P_{n,\theta,\sigma^2}(A) = \sum_{m \in \mathcal{M}_{all}} \int_{\mathbb{R}^n} I_A(y, m) K_n(y, m) dQ_{n,\theta,\sigma^2}(y) \tag{2}$$

where Q_{n,θ,σ^2} is the probability measure on $\mathbb{R}^n$ defined by the $N(X\theta, \sigma^2 I_n)$-distribution, $K_n(y, m)$ is the bootstrap distribution on $\mathcal{M}_{all}$ (i.e., corresponds to a Markov-kernel from $\mathbb{R}^n$ to $\mathcal{M}_{all}$), and I_A is the indicator function of A. Now, choose an arbitrary constant $\gamma > 0$ and a vector $\theta^{(0)} \in \mathbb{R}^p$ such that at least one coordinate of $\theta^{(0)}$ equals zero. By Proposition 1,

we have $P_{n,\theta^{(0)},\sigma^2}(\tilde{m}_L = \tilde{m}_U = m^*(\theta^{(0)})) = 1 + o(1)$. For each n, choose $\theta^{(n)} \in \mathbb{R}^p$ so that $m^*(\theta^{(n)}) \neq m^*(\theta^{(0)})$ and so that $\|\theta^{(n)} - \theta^{(0)}\| \leqslant \gamma/\sqrt{n}$. (E.g., obtain $\theta^{(n)}$ from $\theta^{(0)}$ by adding $\gamma/\sqrt{n}$ to one of the coordinates of $\theta^{(0)}$ that are equal to zero.) Because the sequences of measures $P_{n,\theta^{(0)},\sigma^2}$ and $P_{n,\theta^{(n)},\sigma^2}$ are mutually contiguous (see below), it follows that $P_{n,\theta^{(n)},\sigma^2}(\tilde{m}_L = \tilde{m}_U = m^*(\theta^{(0)})) = 1 + o(1)$. By construction, we have $m^*(\theta^{(0)}) \neq m^*(\theta^{(n)})$, and hence $P_{n,\theta^{(n)},\sigma^2}(\tilde{m}_L \subseteq m^*(\theta^{(n)}) \subseteq \tilde{m}_U) = o(1)$. [Mutual contiguity of $P_{n,\theta^{(0)},\sigma^2}$ and $P_{n,\theta^{(n)},\sigma^2}$ is seen as follows: $Q_{n,\theta^{(0)},\sigma^2}$ and $Q_{n,\theta^{(n)},\sigma^2}$ are well-known to be mutually contiguous, see, e.g., Lemma A.1 in Leeb and Pötscher (2006b) and note that this lemma, while given for the case $\sigma^2 = 1$, obviously extends to any $\sigma^2 > 0$. The claim then follows from Lemma 3.6 in Leeb and Pötscher (2006b).]

(b) Completely analogous, where now $\mathcal{M}_{all}$ is replaced by the B-fold Cartesian product $(\mathcal{M}_{all})^B$, $I_A(y, m)$ and $K_n(y, m)$ are replaced by $I_A(y, m_1, \ldots, m_B)$ and $K_n(y, m_1, \ldots, m_B)$, respectively, and where the latter represents the bootstrap distribution of a bootstrap sample of size B. ∎

Proposition 2 shows that the proposed MCS does not perform as desired in that the worst-case coverage probability tends to zero, and not to the nominal coverage probability $1 - \alpha$. It also shows that the fixed-parameter asymptotic setting used in Theorem 1 of Li et al. (2018) and in Proposition 1 paints a misleadingly optimistic picture of the actual performance of the MCS procedure. This is in line with similar findings in the context of inference post-model-selection reported earlier in Leeb and Pötscher (2005, 2006a,b, 2008a,b,c); Pötscher (2009); Pötscher and Leeb (2009); Pötscher and Schneider (2009, 2011). In fact, the final claim in Part (a) of Proposition 2 can be easily read-off from results concerning model selection probabilities already given in these references (for a start see, e.g., Section 2.1 of Leeb and Pötscher (2005)).

The proof of Proposition 2 reveals that this result continues to hold if the infimum over $\theta \in \mathbb{R}^p$ in (1) is further restricted, even in a sample-size dependent (e.g., shrinking) way, as long as the infimum at sample size n is taken over a set containing a point $\vartheta^{(n)}$ so that the sequence $\vartheta^{(n)}$ has the same properties as the sequence $\theta^{(n)}$ that is used in the proof. Also,

under assumptions (A.1) and (A.2), a result similar to Proposition 2 continues to hold in sufficiently smooth parametric models (as long as the contiguity argument used in the proof goes through). This, of course, also covers regression models with stochastic regressors.

Proposition 2 is an asymptotic result. To illustrate its import for finite-sample situations we performed a small simulation study extending some of the results presented in Li et al. (2018). Data was generated as described in Section 5 of Li et al. (2018). In particular, we consider scenario (c) of Section 5.1 of that paper with the only difference that we consider not only one but several values for the parameter vector θ. More precisely, we set $B = 1000$, $n = 300$, $p = 15$, $p^* = 6$, $\sigma^2 = 1$, $\rho = 0.5$, $\theta_1 = \ldots, \theta_{p^*-1} = 1$, and vary θ_{p^*} through the values $\theta_{p^*} = 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 1, 2$. Note that $m^* = \{1, \ldots, p^*\}$. The nominal confidence level was set to 90%, i.e., $\alpha = 0.1$. The SCAD, Lasso, minimum BIC, and minimum AIC were used as model selectors. We used the residual bootstrap for SCAD, BIC, and AIC, and the modified residual bootstrap for Lasso. Descriptions of the bootstrap algorithms are provided in Section 1.3 of the supplementary material of Li et al. (2018).

For each of the eight different values of θ listed above and for each of the four model selectors, we generated 200 response vectors $Y^{(i)}$, obtained $(\hat{m}_L^{(i)}, \hat{m}_U^{(i)})$, and computed an estimate for the coverage probability via

$$CP_\theta = (1/200) \sum_{i=1}^{200} I(\hat{m}_L^{(i)} \subseteq m^* \subseteq \hat{m}_U^{(i)}).$$

Let CP_* denote the minimum of the CP_θ-values when θ varies through the eight values mentioned above. This results in an estimate of a (loose) upper bound for the worst-case coverage probability for each of the four model selection procedures considered. [Of course, we could have searched more thoroughly over the parameter space for θ to get a smaller upper bound for the worst-case coverage probability, but this would require a much higher investment in computational resources (including a second-stage simulation correcting downward bias resulting from searching for the minimum). Searching only over eight values of θ, as we do here, incurs only a, for our purpose, negligible bias. The rough upper bound CP_* we obtain is good enough to make our point.] We summarize the values of CP_* for the four procedures in the subsequent table:

	BIC	SCAD	AIC	LASSO
CP_*	0.45	0.16	0.92	0.94

The result for CP_*, when the consistent model selection procedure minimum BIC is used, is in good agreement with Proposition 2, as CP_* is much smaller than the nominal confidence level 0.9. [In additional simulations for model selection by minimum BIC using other values of θ we even have found values of CP_θ as small as 0.06.] Also, for SCAD the observed CP_* is much smaller than 0.9. [Following Li et al. (2018) we have chosen the tuning parameter via crossvalidation. Li et al. (2018) do not show that this results in a consistent model selection procedure, and probably it does not. However, we have not investigated this issue. If the so-tuned SCAD is not a consistent model selection procedure, then Proposition 2 strictly speaking does not apply. Nevertheless, the simulations show that the worst-case coverage probability can be way below the nominal one.] For the minimum AIC procedure, as well as for Lasso, the observed CP_* is about right. Since minimum AIC is not a consistent, but rather a conservative, model selection procedure, Proposition 2 does not apply; probably the same is true for Lasso given the tuning used. [Again the tuning parameter for Lasso has been chosen via crossvalidation. Li et al. (2018) do not show that this results in a consistent model selection procedure, and probably it does not. We have not investigated this issue any further.] Of course, this is not to say that the worst-case coverage probability for both of these procedures is in any way guaranteed to be close to the nominal one; it could also be that a more exhaustive search over the parameter space would have turned up a much lower worst-case coverage probability also for these procedures.

The above discussion shows that the proposed MCS has substantial defects when used with consistent model selection procedures. It leaves open the possibility that the MCS procedure suggested by Li et al. (2018) may have merit for some conservative model selection procedures; however, as no theoretical support for this is presented in Li et al. (2018), this still needs to be investigated. Another possible route for research is to consider alternative asymptotic scenarios where, e.g., the number of parameters increases with sample size.

Problems of prediction following model selection have been successfully tackled in such a framework, see Leeb (2008, 2009); Steinberger and Leeb (2018).

References

HAJEK, J. (1971). Limiting properties of likelihoods and inference. In *Foundations of Statistical Inference: Proceedings of the Symposium on the Foundations of Statistical Inference, University of Waterloo, Ontario, March 31 – April 9, 1970* (V. P. Godambe and D. A. Sprott, eds.). Holt, Rinehard and Winston, Toronto, Canada, 142–162.

LEEB, H. (2008). Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli* **14**, 661–690.

LEEB, H. (2009). Conditional predictive inference post model selection. *Ann. Statist.* **37**, 2838–2876.

LEEB, H. and PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21**, 21–59.

LEEB, H. and PÖTSCHER, B. M. (2006a). Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.* **34**, 2554–2591.

LEEB, H. and PÖTSCHER, B. M. (2006b). Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory* **22**, 69–97. Corrigendum ibid. **24**, 581-583, 2008.

LEEB, H. and PÖTSCHER, B. M. (2008a). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* **24**, 338–376.

LEEB, H. and PÖTSCHER, B. M. (2008b). Model selection. In *Handbook of Financial Time Series* (T. G. Andersen, R. A. Davis, J.-P. Kreiß and T. Mikosch, eds.). Springer, New York, NY, 785–821.

LEEB, H. and PÖTSCHER, B. M. (2008c). Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Econometrics* **142**, 201–211.

LI, Y., LUO, Y., FERRARI, D., HU, X. and QIN, Y. (2018). Model confidence bounds for variable selection. *Biometrics*. Forthcoming.

PÖTSCHER, B. M. (2009). Confidence sets based on sparse estimators are necessarily large. *Sankhya* **71**, 1–18.

PÖTSCHER, B. M. and LEEB, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *J. Multivariate Anal.* **100**, 2065–2082.

PÖTSCHER, B. M. and SCHNEIDER, U. (2009). On the distribution of the adaptive LASSO estimator. *J. Statist. Plann. Inference* **139**, 2775–2790.

PÖTSCHER, B. M. and SCHNEIDER, U. (2011). Distributional results for thresholding estimators in high-dimensional Gaussian regression models. *Electron. J. Statist.* **5**, 1876–1934.

STEINBERGER, L. and LEEB, H. (2018). Prediction when fitting simple models to high-dimensional data. *Ann. Statist.* Forthcoming.

July 2018