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

Gerda Claeskens
ORStat and Leuven Statistics Research Center, KU Leuven, Leuven, Belgium
gerda.claeskens@kuleuven.be

Maarten Jansen
Departments of Mathematics and Computer Science, Université libre de Bruxelles,
Brussels, Belgium
maarten.jansen@ulb.ac.be

Li, Luo, Ferrari, Hui and Qin (hereafter referred to as Li et al.) have worked on an important theme. With the growing use of variable selection methods for model search and data mining, it is important for the users to be aware of the variability involved with the use of such procedures. Since data sets might be large and might consists of a large number of variables, it is important to have fast methods that compute such model uncertainty quantifications.

1 Emphasis on the true model

A quest for truth is noble. From the start the authors make the assumption that there exists a true model and that they are so lucky as to have specified this true model within their set of possible models. Such a scenario is ideal for consistent selection methods. Indeed, such methods' driving spirit is to identify the true model with probability converging to one, or even almost surely for some situations.

This search for the truth might be narrow-minded, however, since not everyone will be so fortunate to have used the true data generating density in the likelihood construction. In terms of Li et al., this means that there is no m^* in the model set that is searched over. While the theory of the paper does not apply, it still makes sense to think about this – realistic – scenario since also in such cases, it is useful to get an idea about the variability resulting from the model selection.

Charkhi and Claeskens (2018) study the asymptotic selection behaviour of the Akaike information criterion (AIC), which is known not to be consistent. Their results explicitly use the overselection property of the AIC. Confidence intervals for model parameters or linear combinations thereof, conditional on the selected model, are constructed using certain truncated Gaussian random variables. Interestingly, only in case no true model is present and least-false (also called pseudo-true) parameters serve as limits for maximum likelihood estimators, uniformly valid results are obtained. A direct extension to build confidence sets for the model rather than for parameters in the model selected by using AIC seems tricky, though. The interpretation certainly needs care in case one cannot assume there is a true model in the model set.

2 Model selection as a one-step or a two-step procedure?

The selection of a model using an information criterion is often a two-level procedure. In low dimensional models, the first, procedural level typically consists of the maximisation of the likelihood. If the minimisation problem is numerically ill-posed or ill-conditioned, such as in high dimensional models (where the number of observations is lower than the number of parameters to be estimated), the likelihood is regularized before being optimized. Basis pursuit or lasso (Chen et al., 1998; Tibshirani, 1996) adopts an ℓ_1-regularization of the log-likelihood. In any case, the procedural level performs the actual selection and estimation of the model parameters. For lasso, this becomes, $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \{ -\log L_n(\boldsymbol{\theta}) + \lambda_n \sum_{j=2}^{p} |\theta_j| \}$, where in a regression model, typically the intercept is not penalized. The regularization parameter, λ_n, controls the size of the selected model. Finetuning of λ_n forms the second level of the variable selection. In practice, the finetuning of λ_n is equivalent to fixing the model size. It can be performed by optimization over λ_n or over the model size of an information criterion such as AIC, BIC, Mallows's C_p, cross-validation, etc. Some of these criteria, such as AIC, Mallows's C_p and cross-validation aim at optimisation of the *expected* log-likelihood with respect to the data generating process, or optimisation of a closely related objective, such as the prediction error, or the (Kullback-Leibler) distance to the data generating process. This way, the criterion typically leads to an efficient choice of the regularization parameter, as opposed to the consistent choice by methods such as BIC. As the objective function measures some distance between the model under consideration and the data generating process, its optimisation amounts to finding the best balance between (squared) bias and variance. The resulting information criterion transforms this balancing problem into a balance between the observed log-likelihood and model complexity. The model complexity can be expressed using the notion of (generalized) degrees of freedom (Ye, 1998), which for the analysis of a fixed model corresponds to the model size. In general, the degrees of freedom depend on the procedural level, and more precisely on the number and effect of false positives on the variance of the estimator in the selected model. In the lasso procedure, the false positives are numerous, but their effect is tempered by the shrinkage in the operation, thereby leading to the result that the degrees of freedom are given by the size of the selected model (Tibshirani and Taylor, 2012; Zou et al., 2007; Gao and Fang, 2011). In absence of a shrinkage, the expression for the degrees of freedom, and thus for the likelihood-complexity balance, is less straightforward (Jansen, 2014).

All of the model selection aspects should be incorporated when constructing model confidence bounds. Hence the use of 10-fold cross-validation to determine λ_n in a regularized estimation procedure might result in a different estimated/selected model as when BIC is used to specify λ_n, for example. It would be interesting to investigate whether the current approach with the incorporation of bootstrap results, can also handle such a full, two-step, selection procedure.

3 Selective inference for model confidence sets?

In case of regularized estimation with a fixed value of λ_n, the estimator $\hat{\boldsymbol{\theta}}$ might have several components that are exactly equal to zero. Denote the selected model by $\hat{M}_{\text{lasso}}$. The procedure of selective inference (Lee et al., 2015) deals with the selection uncertainty by explicitly investigating the selection event $\{\hat{M}_{\text{lasso}} = M\}$. That means, it characterizes what makes that the selected model $\hat{M}_{\text{lasso}}$ is given by M. This knowledge leads to post-selection intervals for the regression parameters. See also Taylor et al. (2016) for selection events corresponding to forward selection and to using the algorithm for least angle regression. It would be interesting to investigate whether their truncated Gaussian test statistics that are conditional on the active set of coefficients, could be used to form confidence sets for the model, rather than for the parameters in the model. One possible step in this direction could be to use their conditional results in combination with a law of total probability to obtain marginal results.

4 Choice of the model set

Charkhi and Claeskens (2018) observed that confidence intervals for model parameters clearly depend on the model set $\mathcal{M}$. The more models in the model set, the more inequalities determine the selection event. For post-selection inference on the model parameters, it would be best to use as few models in the selection as possible.

A question worth considering is which influence the model set $\mathcal{M}$ has on the model confidence bounds. Would an enlargement of the set of models, which leads to an increase in selection uncertainty since more models need to be considered, also lead to different/larger model confidence sets? To gain more understanding about this question we took the diabetes data set as in Li et al., Section 4, and we have used their R code to the extended data set which also includes quadratic effects and pairwise interactions of all variables. This enlarges the design matrix from 10 columns with only main effects, to 64 columns with all additional effects added. While it turned out that Algorithm 1 still works for 10 variables, even though it takes some time, it was totally infeasible to get results with a larger number of variables due to the exponential order of calculations needed. Algorithm 2, however, works fast enough to include all 64 variables in the largest model.

We give here the results for adaptive lasso, using the unmodified R code of the authors. We consider a nominal level of 95% and take the model selection bounds for which the bootstrap coverage exceeds this value by the smallest amount.

Scenario 1: 10 main effects in the largest model. The coverage of the MCB is 98.7%.
The lower bound model contains the variables: `bmi`, `ltg`, `map`.
The upper bound model is equal to the full model, including all 10 main effects.

Scenario 2: 10 main effects and 45 pairwise interactions in the largest model. The coverage of the MCB is 95.7%.
The lower bound model contains the variable `bmi`. The upper bound model contains all variables except for `age` and the interactions between `age` and `bmi` and between `map` and `glu`. There are 52 variables in the upper bound model.

Scenario 3: 10 main effects, 9 quadratic effects and 45 pairwise interactions. The coverage of the MCB is 95.8%.
The lower bound model contains the variable `bmi`. The upper bound model contains all variables except for `age` and the interactions between `age` and `bmi` and between `map` and `glu`. Since here the full model is larger, there are 61 variables in the upper bound model.

We see that if we increase the number of variables from 10 to 55 and further to 64, the lower bound model only contains a single variable anymore (bmi), the upper bound though increases drastically: only three variables are not included in the upper bound model. Coincidently, it are the same three variables. The width increases from 7 to 51 and 60.

How should we interpret this? It makes sense that the intervals get wider since the model set $\mathcal{M}$ contains more elements. With only 10 main effects in the model, the MCB contains $2^7 = 128$ models out of the $2^{10} = 1024$ models in the model set $\mathcal{M}$, this gives a ratio of 1/8=12.5%. When adding pairwise interactions, the MCB contains the large number of 2^{51} models out of a total of 2^{55} models, which gives a ratio of 1/16=6.25%. This ratio stays the same for the last case where the MCB contains 2^{61} models, out of a total of 2^{64} possible models. These are really large sets.

Looking at the difference between the number of variables in the upper bound and the lower bound models, corresponds to the width of the MCB. Is this width representative enough for such a large set of models? While the MCBs in the last two scenarios contain a large, even a huge number of models, their relative percentage of the models in the model set is only half of what it was for the first scenario. Would this merely be an effect of the exponential number of models, or could there be another explanation?

References

Charkhi, A. and Claeskens, G. (2018). Asymptotic post-selection inference for Akaike's information criterion. *Biometrika* **105,** 645-–664.

Chen, S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* **20,** 33–61.

Gao, X. and Fang, Y. (2011). A note on the generalized degrees of freedom under the l_1 loss function. *Journal of Statistical Planning and Inference* **141,** 677–686.

Jansen, M. (2014). Information criteria for variable selection under sparsity. *Biometrika* **101,** 37–55.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2015). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* **44,** 907–927.

Taylor, J., Lockhart, R., Tibshirani, R. J., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association* **111,** 600–620.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58,** 267–288.

Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics* **40,** 1198–1232.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* **93,** 120–131.

Zou, H., Hastie, T. J., and Tibshirani, R. J. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics* **35,** 2173–2192.