Model selection was an underdeveloped country on the map of classical statistics. The blame lay with intractable mathematics, of the sort connected with discontinuous functional forms. That excuse has worn thin in an era of virtually infinite computer power. The current article shows some progress being made through a combination of a little mathematics with a lot of computation, while the discussants offer other promising paths forward. It seems a safe bet that model selection, how to do it and what are its effects on inference, will continue to be a major topic for statisticians. Here I have touched on both aspects, bootstrap smoothing for the how-to-do-it part, and two bagging accuracy theorems on its effects.

Professor Hjort, following through on his ambitious series of articles with Claeskens, puts the problem into an asymptotic framework. This necessarily involves making the signal weaker (“γj = δj/√n”) as sample size n increases. Otherwise the model-selection aspect disappears: in terms of my schematic Figure 9, the distributional ellipses shrink to lie within a single wedge. Here, we have to worry that changing the signal strength may reduce asymptotics’ relevance to the situation at hand. No such change is necessary in the classical picture of Figure 8, where the asymptotics are inherently simpler.

My article avoids asymptotics, or at least the mention of asymptotics. Bootstrap methods are by their nature nonasymptotic, though their formal justification in the literature usually involves large-sample calculations. Professor Politis’ Model-Free Prediction Principle for regression is justified in terms of transformations that induce heteroscedastic residuals. This is similar in intention to Efron (1987), where bootstrap confidence intervals are justified by hypothetical transformations to normality, avoiding at least a direct appeal to asymptotics.

Politis’ discussion is a reminder that neither of my current theorems applies to bootstrapping residuals. On the other hand, they do apply to model-selection situations other than regression—K-means clustering for example, or data-based choice of window width in kernel density estimation.

I was reassured by the new method’s good performance in the simulation studies of Professors Wang, Sherwood, and Li, notably more ambitious than the few in my article. The questions raised of the number of bootstrap replications B is an important one. The tactic in my examples, choosing B much bigger than necessary, might be impractical in more complicated problems. Formula (3.11) provides a data-based guide to the choice, supplemented with the bias correction calculations of Remark J.

The bootstrap is fundamentally a plug-in methodology, along the lines of maximum likelihood estimation (MLE). As such I would not expect se to perform well in the “large p” context of Wang–Sherwood–Li’s second question, but I would be happy to be proven wrong.

Bootstrap smoothing, or bagging, can be thought of as a form of nonparametric MLE (Efron and Tibshirani 1997): suppose \( x = (x_1, x_2, \ldots, x_n) \) is an iid sample from distribution \( F \), and \( \hat{\theta} = r(x) \) is an unbiased estimator of a parameter of interest \( \theta = \theta(F) \),

\[
\theta(F) = E_F[\hat{\theta}].
\]

Then the nonparametric MLE of \( \theta \) is

\[
\hat{\theta} = \theta(\hat{F}) = E_{\hat{F}}[\hat{\theta}^*],
\]

where \( \hat{F} \) is the empirical distribution corresponding to \( x \); in other words \( \hat{\theta} \) is the bagged version of \( \hat{\theta} \).

Blind application of bootstrap smoothing is not going to work if \( \hat{\theta}^* = r(x^*) \) can take on infinite values. This is a sign that the statistic \( r(\cdot) \) is unstable in the relevant neighborhood of the sample space, and that some regularization, of the kind Gelman and Veharti use in the “bad” example, is called for. In other words, one should not throw out the bootstrap with the bad statistic’s bathwater.

Professors Gupta and Lahiri suggest two more approaches to the post model-selection accuracy problem. Zou’s interesting ALASSO method asymptotically selects the right model. In terms of Figure 9, the sampling ellipses around the ALASSO estimate must shrink to lie entirely within a single wedge, and in fact the correct one. In the examples I have looked at, admittedly not all that many, model selection was realistically far more random than that.

A quite different approach is suggested as the maximum frequency (MF) method: only employ those bootstrap replications that fall into the same wedge as the original data, thereby avoiding model-selection jumpiness. For the Cholesterol example of Table 3, MF gives approximate 95% interval

\[
4.71 \pm 1.96 \cdot 5.43 = [-5.93, 15.35],
\]

far to the right of the smoothed interval [−13.3, 8.0].

This raises the question of conditionality: perhaps the statistician should condition on the observed selected model (though this raises the peril of ignoring model selection effects, our original objection to classical practice).

Bayesian estimates of accuracy are automatically conditional. For model selection problems, however, they require

Bradley Efron is Professor of Statistics and Biostatistics, Department of Statistics, Stanford University, Stanford, CA 94305-4065 (E-mail: brad@stat.stanford.edu).
a fearsome amount of prior specification: prior probabilities for the different models, and then informative prior distributions within each model. In two current articles (Efron 2012, 2014), I have drawn connections between “objective” Bayes’ analysis and bootstrap estimates of variability. That line of thinking supports the unconditional kinds of bootstrap smoothing suggested in the current article, but so far it is only a suggestion.

My thanks go to the discussants and editors for an informative exchange on an important topic.

REFERENCES


