Table 2. Coverage accuracy and average lengths of MFB CIs for $\mu_1$ for the CP, the AIC, and the BIC methods based on 500 simulation runs and 1000 bootstrap replicates. The nominal confidence level is 95

| | Length of CIs | Coverage probability |
|---|---|---|
| AIC | 3.140 | 0.927 |
| $C_p$ | 3.161 | 0.960 |
| BIC | 3.182 | 0.980 |

of bootstrap replicates (among $B$ many) that resulted in selection of the model $j_0$. Then, the MFB method makes use of the subcollection

$$\{\mathcal{D}_n^{*b} : b \in \mathcal{B}_0\}$$

of resamples to carry out bootstrap-based inference. For example, bootstrap CIs for linear combinations of the regression parameter vector can be obtained by using the bootstrap-$t$ method applied only to the resamples $\{\mathcal{D}_n^{*b} : b \in \mathcal{B}_0\}$. Since all replicates in this collection correspond to a single model, the extra variability that results from the model selection step in different resamples is eliminated. In fact, this MFB approach was used for constructing percentile-$t$ CIs for the parameter $\mu_1$ in Section 3. Although the respective model selection methods have considerable variability in selecting the true model among $B$ resamples, the empirical coverage accuracy of the MFB approach reported therein appears reasonable for each of the three model selection methods. Theoretical properties of the MFB method is currently under investigation.

## 3. NUMERICAL RESULTS

Here, we report results from a small simulation study on the MFB method. We consider model (1) with $p = 10$ and $p_0 = 3$ (a cubic model), where $\beta_0 = 1$, $\beta_1 = 0.5$, $\beta_2 = 0.4$, $\beta_3 = 5.0$, and $\beta_i = 0$ for all $i = 4, \ldots, 10$. We generated the variables $(c_i, \epsilon_i)$ as iid bivariate normal vectors with zero mean vector and identity covariance matrix. The sample size considered was $n = 200$. The MFB method was used to construct bootstrap CIs for the parameter $\mu_1 = E(y_1|c_1)$ where the model selection was performed with the CP, the Akaike information criterion (AIC), and the Bayesian information criterion (BIC) methods. The results from the model selection step applied to the bootstrap resamples are summarized in Table 1. The first three rows of the table give the frequencies of the different models, which were selected by each of the three methods over 600 simulation runs. The last three rows give the associated standard deviations. It is evident from the table that except for the BIC, which is known to be consistent for model selection, the other two methods selected the true model with low empirical probability. As a result, the use of either of these model methods in the naive approach would produce very distorted results. However, by using the MFB approach, even in such situations, we are able to identify the true model. The empirical coverage accuracy and the average lengths of a nominal 95% CI for $\mu_1$ are reported in Table 2. The coverage is evidently very good irrespective of the model selection performance of the three model selection methods.

## REFERENCES

Bühlmann, P., and Yu, B. (2002), "Analyzing Bagging," *The Annals of Statistics*, 30, 927–961. [1013]

Chatterjee, A., and Lahiri, S. N. (2013), "Rates of Convergence of the Adaptive LASSO Estimators to the Oracle Distribution and Higher Order Refinements by the Bootstrap," *The Annals of Statistics*, 41, 1055–1692. [1014]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1014]

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [1014]

# Comment

Andrew GELMAN and Aki VEHTARI

## 1. ACCOUNTING FOR MODEL SELECTION IN STATISTICAL INFERENCE

How can one proceed with predictive inference and assessment of model accuracy if we have selected a single model from some collection of models? Selecting a single model instead of model averaging can be useful as it makes the model easier to explain, and in some cases that single model gives similar predictions as the model averaging.

The selection process, however, causes overfitting and biased estimates of prediction error; thus much work has gone into estimating predictive accuracy given available data (e.g., Gelman, Hwang, and Vehtari 2013). In Efron's article, bagging is used to average over different models, and the main contribution is providing a useful new formula estimating the accuracy of bagging in this situation.

It makes sense that bagging should work for the smooth unstable ("jumpy") estimates in the examples shown. Full Bayesian inference should also be able to handle these problems, but it can be useful to have different approaches based on different principles.

Andrew Gelman is Professor, Department of Statistics, Columbia University, New York, NY 10027 (E-mail: *gelman@stat.columbia.edu*). Aki Vehtari is Adjunct Professor, Department of Biomedical Engineering and Computational Science, Aalto University, Espoo, Finland (E-mail: *aki.vehtari@aalto.fi*).


Journal of the American Statistical Association
September 2014, Vol. 109, No. 507, Theory and Methods
DOI: 10.1080/01621459.2014.906153

One of the appeals of the bootstrap is its generality (as, in a completely different way, with Bayes; see Gelman 2011). Any estimate can be bootstrapped; all that is needed are an estimate and a sampling distribution. The very generality of the boostrap creates both opportunity and peril, allowing researchers to solve otherwise intractable problems but also sometimes leading to an answer with an inappropriately high level of certainty.

We demonstrate with two examples from our own research: one problem where bootstrap smoothing was effective and led us to an improved method, and another case where bootstrap smoothing would not solve the underlying problem. Our point in these examples is not to disparage bootstrapping but rather to gain insight into where it will be more or less effective as a smoothing tool.

## 2. AN EXAMPLE WHERE BOOTSTRAP SMOOTHING WORKS WELL

Bayesian posterior distributions are commonly summarized using Monte Carlo simulations, and inferences for scalar parameters or quantities of interest can be summarized using 50% or 95% intervals. A $1-\alpha$ interval for a continuous quantity is typically constructed either as a central probability interval (with probability $\alpha/2$ in each direction) or a highest posterior density interval (which, if the marginal distribution is unimodal, is the shortest interval containing $1-\alpha$ probability). These intervals can in turn be computed using posterior simulations, either using order statistics (e.g., the lower and upper bounds of a 95% central interval can be set to the 25th and 976th order statistics from 1000 simulations) or the empirical shortest interval (e.g., the shortest interval containing 950 of the 1000 posterior draws).

For large models or large datasets, posterior simulation can be costly, the number of effective simulation draws can be small, and the empirical central or shortest posterior intervals can have a high Monte Carlo error, especially for wide intervals such as 95% that go into the tails and thus sparse regions of the simulations. We have had success using the bootstrap, in combination with analytical methods, to smooth the procedure and produce posterior intervals that have much lower mean squared error compared with the direct empirical approaches (Liu, Gelman, and Zheng 2013).

## 3. AN EXAMPLE WHERE BOOTSTRAP SMOOTHING IS UNHELPFUL

When there is separation in logistic regression, the maximum likelihood estimate of the coefficients diverges to infinity. Gelman et al. (2008) illustrated with an example of a poll from the 1964 U.S. presidential election campaign, in which none of the black respondents in the sample supported the Republican candidate, Barry Goldwater. As a result, when presidential preference was modeled using a logistic regression including several demographic predictors, the maximum likelihood for the coefficient of "black" was $-\infty$. The posterior distribution for this coefficient, assuming the usual default uniform prior density, had all its mass at $-\infty$ as well. In our article, we recommended a posterior mode (equivalently, penalized likelihood) solution based on a weakly informative Cauchy (0, 2.5) prior distribution that pulls the coefficient toward zero. Other, similar, approaches to regularization have appeared over the years. We justified our particular solution based on an argument about the reasonableness of the prior distribution and through a cross-validation experiment. In other settings, regularized estimates have been given frequentist justifications based on coverage of posterior intervals (see, e.g., the arguments given by Agresti and Coull 1998, in support of the binomial interval based on the estimate $\hat{p} = \frac{y+2}{n+4}$).

Bootstrap smoothing does not solve problems of separation. If zero black respondents in the sample supported Barry Goldwater, then zero black respondents in any bootstrap sample will support Goldwater as well. Indeed, bootstrapping can exacerbate separation by turning near-separation into complete separation for some samples. For example, consider a survey in which only one or two of the black respondents support the Republican candidate. The resulting logistic regression estimate will be noisy but it will be finite. But, in bootstrapping, some of the resampled data will happen to contain zero black Republicans, hence complete separation, hence infinite parameter estimates. If the bootstrapped estimates are regularized, however, there is no problem.

The message from this example is that, perhaps paradoxically, bootstrap smoothing can be more effective when applied to estimates that have already been smoothed or regularized.

## REFERENCES

Agresti, A., and Coull, B. A. (1998), "Approximate is Better Than Exact for Interval Estimation of Binomial Proportions," *The American Statistician*, 52, 119–126. [1016]

Gelman, A. (2011), "The Pervasive Twoishness of Statistics; in Particular, the Sampling Distribution and the Likelihood are Two Different Models, and That is a Good Thing," Statistical Modeling, Causal Inference, and Social Science Blog, 20 June. Available at *http://andrewgelman.com/2011/06/20/the_sampling_di_1/*. [1016]

Gelman, A., Hwang, J., and Vehtari, A. (2013), "Understanding Predictive Information Criteria for Bayesian Models," *Statistics and Computing*, doi: 10.1007/sl 1222-013-9416-2. [1015]

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y. S. (2008), "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models," *Annals of Applied Statistics*, 2, 1360–1383. [1016]

Liu, Y., Gelman, A., and Zheng, T. (2013), "Simulation-Efficient Shortest Probability Intervals," Technical report, Department of Statistics, Columbia University. [1016]