Bootstrap averaging: Examples where it works and where it doesn't work*

Andrew Gelman† Aki Vehtari‡

9 Jan 2014

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 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 paper, bagging is to 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.

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 (for example, 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 (for example, 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

*To appear in *Journal of the American Statistical Association*. Discussion of "Estimation and accuracy after model selection," by Bradley Efron.

†Department of Statistics, Columbia University, New York

‡Department of Biomedical Engineering and Computational Science, Aalto University, Espoo, Finland.

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) illustrate 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 paper, 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, for example, 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. *American Statistician* **52**, 119–126.

Gelman, A. (2011). The pervasive twoishness of statistics; in particular, the sampling distribution and the likelihood are two different models, and thats a good thing. Statistical Modeling, Causal Inference, and Social Science blog, 20 June. `http://andrewgelman.com/2011/06/20/the_sampling_di_1/`

Gelman, A., Hwang, J., and Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*.

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.

Liu, Y., Gelman, A., and Zheng, T. (2013). Simulation-efficient shortest probability intervals. Technical report, Department of Statistics, Columbia University.