

Comparison of Conformal Prediction and Bootstrapping on a Regression Problem with Random Forests

732A76 Research Project

David Hrabovszki
Linköping University

23 December 2020

Abstract

Introduction

Explain coverage rate, mean interval sizes (size of prediction region), runtime.

Related Work

Methods

Random Forest

Probably can skip this part.

Conformal Prediction

Bootstrap

Empirical Evaluation

Experimental Setup

Data Simulation

Empirical experiments are often conducted on real world datasets, but because this research focuses on the parametric version of bootstrapping, we opted for an artificially simulated dataset instead. This means that we can be certain about the distribution of the response variable, which is necessary for parametric bootstrap. The implementation was based on [3] that uses an additive error model:

$Y = m(X) + \epsilon$, where $X \sim \mathcal{N}(0, \Sigma_{50})$ and Σ_{50} is an AR(1) covariance matrix with $\rho = 0.6$ and diagonal values equal to 1. The error term ϵ is Gaussian noise and the $m(X)$ mean function is nonlinear with interaction between the predictors to make the data more complex. 100,000 instances were simulated from this model. The response variable is normalized, so that the resulting interval sizes can be more easily interpreted. An interval size of 0.5 means that the region covers 50% of the range the variable can take.

Data Splits

The experiments use the holdout method for evaluation with 70% - 30% split between training and test sets. The first 70,000 instances were used as the pool for training and the last 30,000 for testing. For each different data size, the appropriate number of observations were sampled from these pools randomly to ensure that no test instance has been used for training or for validation while optimizing the parameters. It is important to note that while bootstrapping uses all training examples for training the random forest, conformal prediction utilizes only a part of it for proper training and the rest for calibration. A possible approach would have been to make the number of instances used for training equal for both methods, but since calibration is an important step in obtaining prediction intervals, I made the decision to use a subset of the training data for calibration. The number of calibration instances was set to

$$q = 100 * \left\lfloor \frac{|Z|}{400} \right\rfloor - 1$$

according to [1], where Z is the full training set. This means that around 20 - 25% of the full training set was used for calibration.

Computer Specifications

The experiments were run on a notebook with an Intel Core i5-8250U processor (4-core 1.6 GHz base frequency) and 12GB of RAM. The conformal predictions were run on a single core, but the bootstrapping utilized all cores in parallel, which is straightforward to implement with the `boot()` function on a Linux system.

Chosen Parameters

Three parameters were optimized using grid search: the number of trees in a random forest (`ntree`), the sensitivity of the nonconformity measure in conformal prediction (`beta`) and the number of bootstrap replicates (`R`). 10,000 instances were used from the training pool for the grid search, where `ntree` was chosen to be **125**, because this value resulted in a “low enough” mean squared error. A higher number of trees would make the predictions slightly more accurate, but only at the a significantly larger computation cost. β was chosen to be **0.01**, because the mean of conformal prediction region sizes was the smallest with this value, and the runtime was slightly shorter than with others. The confidence level set for the interval was barely affected by `beta`. The number of bootstrap replicates should be as large as possible, but increasing this parameter increases the runtime linearly, so the optimal value was set to a low value (**`R` = 200**) that still results in reasonably good coverage rates and mean region sizes.

Results

Prediction intervals were constructed using conformal prediction and bootstrapping at 95% confidence level on artificially simulated datasets of sizes 1,000 - 10,000. In each run, we monitor the coverage rate, the mean interval size and the time required to train the models.

Theoretically, the coverage rate of the conformal prediction should be around the confidence level [1], because the prediction intervals produced are always valid. This is exactly what we observe here as well, the coverage rate is slightly larger than 0.95, but it seems to converge to it as the number of observations grows. Bootstrapping, however, resulted in much lower coverage rates in all the runs, and there seems to be no indication that a larger dataset would solve this problem (Figure 1).

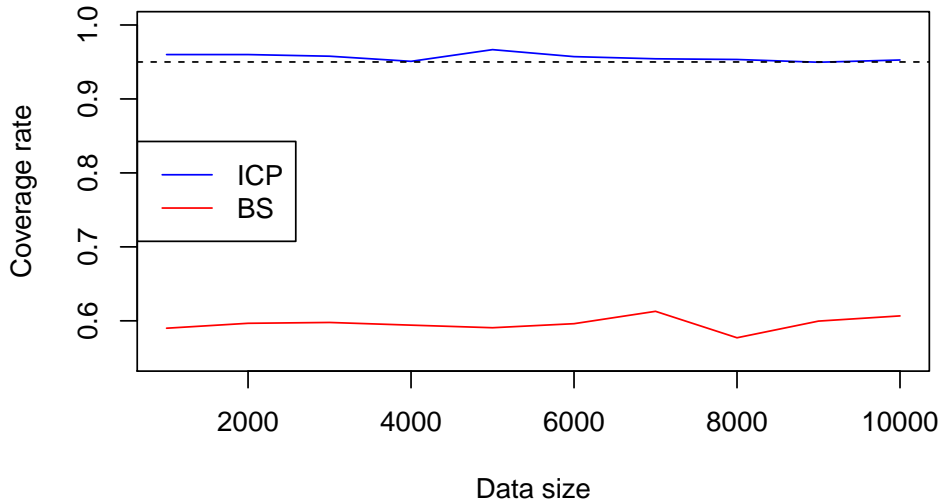


Figure 1: Coverage rates

The mean size of the prediction regions decreases in both cases as the dataset grows, but the bands resulting from bootstrapping are much narrower, around half the size of the conformal prediction bands. The prediction regions cover around 10% of the whole range with bootstrapping, and 20% - 25% with conformal prediction (Figure 2).

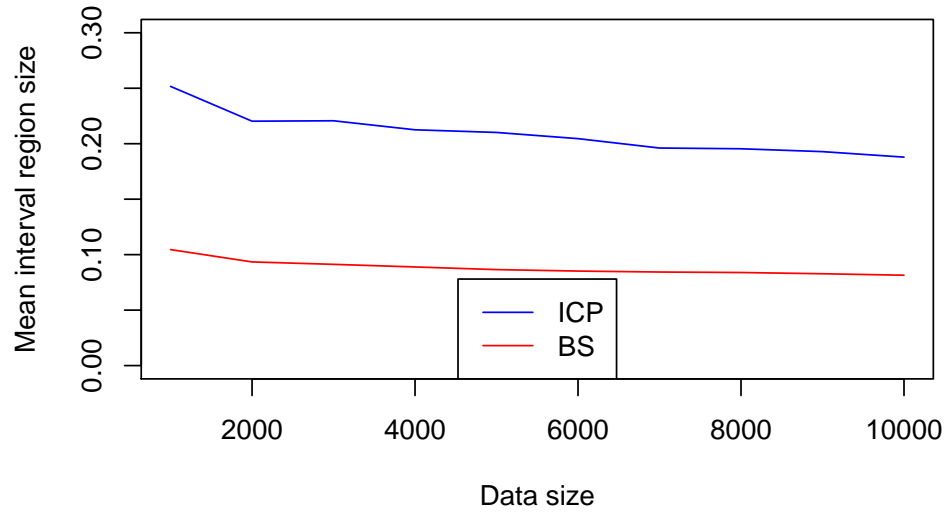


Figure 2: Mean sizes of prediction regions

There is a significant difference between the runtimes of the two methods, even though it increases linearly in both cases as we use more data. Conformal prediction in general seems to be much faster, than bootstrapping. Conformal prediction is measured in seconds, and bootstrapping is measured in minutes on Figure 3.

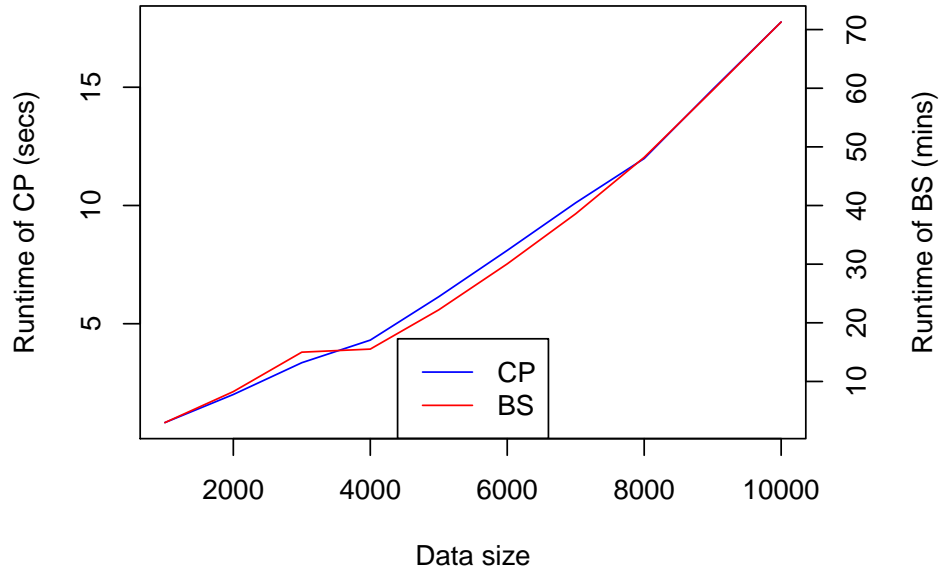


Figure 3: Runtimes

Conclusion

Future Improvements

Several modifications could be made to improve this study of comparing two different methods. The parameters of the random forest and the number of bootstrap replicates could be further optimized for different data sizes, but in this paper I intended to keep a setup unchanged for more straightforward comparison. We could look at out-of-bag predictions of the random forests instead of using a separate set for either calibration or testing, but this would not necessarily improve predictive quality. It would be worth an experiment though, because then the available dataset would be better utilized. The data simulation process could be refined, and real world datasets could be used as well to see how the two methods compare against each other in a more real world setting.

Appendix

Table 1: Results of experimental runs

Size	CP coverage rate	CP region size	CP runtime	BS coverage rate	BS region size	BS runtime
1000	0.960	0.252	0.8	0.590	0.105	178.8
2000	0.960	0.220	2.0	0.597	0.093	496.2
3000	0.958	0.221	3.4	0.598	0.091	899.4
4000	0.951	0.213	4.3	0.594	0.089	930.9
5000	0.967	0.210	6.1	0.591	0.087	1332.6
6000	0.957	0.205	8.1	0.596	0.085	1802.8
7000	0.954	0.196	10.1	0.613	0.084	2315.5
8000	0.953	0.195	12.0	0.577	0.084	2892.9
9000	0.950	0.193	14.9	0.600	0.083	3579.0
10000	0.953	0.188	17.8	0.607	0.081	4277.5