Bootstrap Tilting Diagnostics

Tim C. Hesterberg
Insightful Corp., 1700 Westlake Ave. N., Suite 500
Seattle, WA 98109-3044, U.S.A.
TimH@insightful.com

Abstract
The fundamental bootstrap assumption is that the bootstrap approximates reality; that the sampling distribution of a statistic under the empirical distribution $\hat{F}$ approximates the sampling distribution under the true (unknown) distribution.

A natural way to test this is to investigate how the bootstrap distribution varies when $\hat{F}$ is replaced by other distributions. Iterated bootstrapping, jackknife-after-bootstrap (JAB), and bootstrap tilting diagnostics all do this, but tilting lets one focus on a key questions – how the sampling distribution depends on a parameter of interest – without the noise of the other procedures.

Both tilting and iterated bootstrapping may be used for calibration, and in some cases giving confidence intervals or hypothesis tests that are an order of magnitude more accurate than the uncalibrated versions. But whereas iterated bootstrapping is computationally much more expensive than ordinary bootstrapping, bootstrap tilting is less expensive – 17 to 37 times less expensive than common bootstrap confidence intervals.

Key words: bootstrap tilting, calibration.

Acknowledgements This work was partially supported by grants NIH 2R44CA67734–02 and NSF DMI-0078706. Chris Fraley, Shan Jin, and Robert Thurman have contributed to the software.

1 Introduction

The fundamental bootstrap assumption is that the sampling distribution of a statistic under the unknown true distribution F may be approximated by the sampling distribution under the empirical distribution $\hat{F}$, e.g.

$$\begin{aligned} \mathrm{Var}_F(\hat{\theta}) &\doteq \mathrm{Var}_{\hat{F}}(\hat{\theta}^*) \\ G_F(a) &\doteq G_{\hat{F}}(a) \\ G_F^{-1}(.975) &\doteq G_{\hat{F}}^{-1}(.975) \end{aligned}$$

where $\hat{\theta}$ is a parameter estimate, or G is the sampling distribution (and $G_{\hat{F}}$ the bootstrap distribution).

Curiously, that seemingly innocent introductory paragraph contains a serious error — where the bootstrap should not be used. In this article we outline diagnostic procedures which can be used to diagnose that error.

The basic theme in a variety of diagnostic procedures is to compare the sampling distribution under $\hat{F}$ with the sampling distribution under other distributions, typically with support on the observed data. Here we consider three such procedures:

- Jackknife-after-bootstrap (jackknife sample)
- Iterated bootstrapping (bootstrap sample)
- Bootstrap tilting diagnostics (tilted sample)

This article proceeds largely by examples.

2 Example 1

In this example we observe $n = 28$ observations from a distribution F, and the statistic of interest is $\theta = \mu$. The empirical and bootstrap distributions are shown in Figure 1

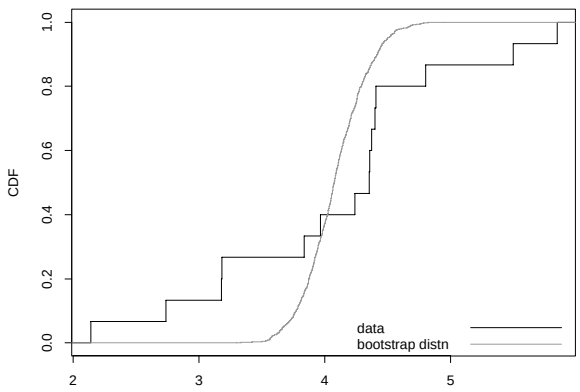

Figure 1: Empirical distribution, and bootstrap distribution of the sample mean

It is natural to attempt to use the bootstrap to estimate quantiles of the sampling distribution of $\bar{X}$. But Figure 2 shows how sensitive those quantiles are to changes in $\hat{F}$. In retrospect, it is clear that one should not do this; the quantiles of the sampling distribution of $\hat{\theta}$ are independent of F (and θ), only if $\hat{\theta}$ is worthless as an estimator of θ. Unfortunately, this is a common mistake when bootstrapping; this

diagnostic procedure could be used to help prevent this error.

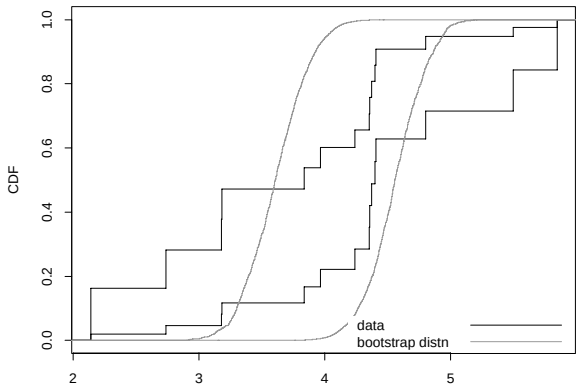

Figure 2: Sensitivity of quantiles of the bootstrap distribution to changes in the underlying distribution, using bootstrap tilting diagnostics. Shown are two weighted empirical distributions, distributions with support on the original data but unequal probabilities, and the bootstrap distributions when sampling from those weighted empirical distributions.

But consider a variation of this example, where the data is the same, but the bootstrap is used to estimate the distribution of $\bar{X} - \mu$, with bootstrap analog $\bar{X}^* - \bar{X}$. We see in Figure 3 that the bootstrap distribution is now much less sensitive to changes in the underlying distribution. Hence it would be reaonable to use the bootstrap to estimate quantiles of $\bar{X} - \mu$ — at least for these data, which have little skewness. We see another example later where this is not the case.

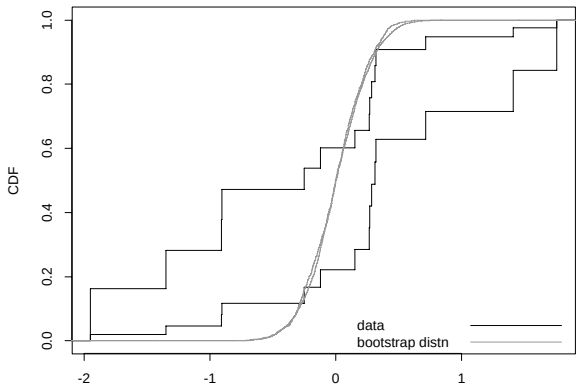

Figure 3: Shown are two weighted empirical distributions, and the corresponding bootstrap distributions of $\bar{X}^* - \bar{X}$.

Jackknife-after-bootstrap The JAB may also be used as a diagnostic procedure in this example. Here bootstrap sampling is from a weighted empirical distribution, with weight $1/(n-1)$ on all observation but one, and weight 0 on that observation. Figure 4 shows the corresponding bootstrap distributions, one for each jackknife sample. As above, we see that the distribution of $\bar{X}^*$ is very sensitive to the underlying distribution, while the distribution of $\bar{X}^* - \bar{X}$ is relatively insensitive.

Unfortunately, the JAB procedure is ineffective as a visual diagnostic procedure for large n.

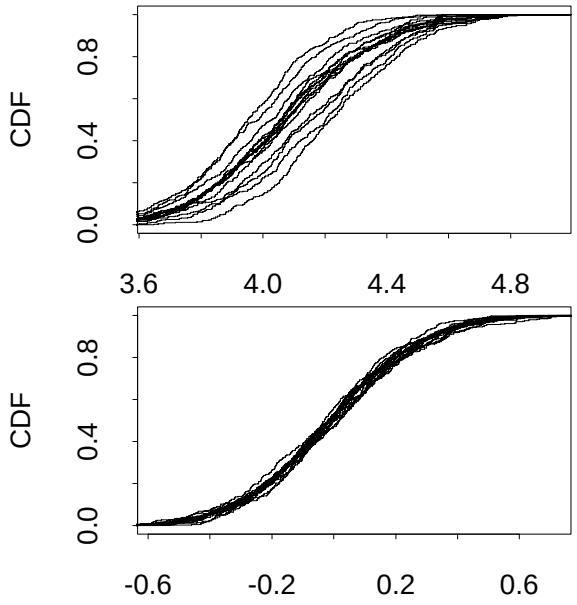

Figure 4: Jackknife-after-bootstrap sampling distribution of $\bar{X}^*$ and $\bar{X}^* - \bar{X}$.

Bootstrap-after-bootstrap The iterated bootstrap may also be used as a diagnostic procedure; here the weighted empirical distributions are bootstrap samples. Figure 2 shows the bootstrap-after-bootstrap (BAB) distributions of $\bar{X}^*$ and $\bar{X}^* - \bar{X}$. As before, the distribution of $\bar{X}^*$ is very sensitive to the underlying distribution. However, we notice here a phenomena not apparent earlier, that the distribution of $\bar{X}^* - \bar{X}$ has a variance that depends on the variance of the weighted empirical distribution. We discuss this later.

3 Bootstrap Tilting Mechanics

The weights used in the weighted empirical distributions used in Example 1 were calculated by "exponential tilting," with weights of the form

$$w_i = c \exp(\tau(x_i - \bar{x}))$$

where c normalizes the weights to sum to 1 and τ is a "tilting parameter" — positive τ tilts to the right, with larger weights on the larger observations, and negative τ tilts left. Exponential tilting may be

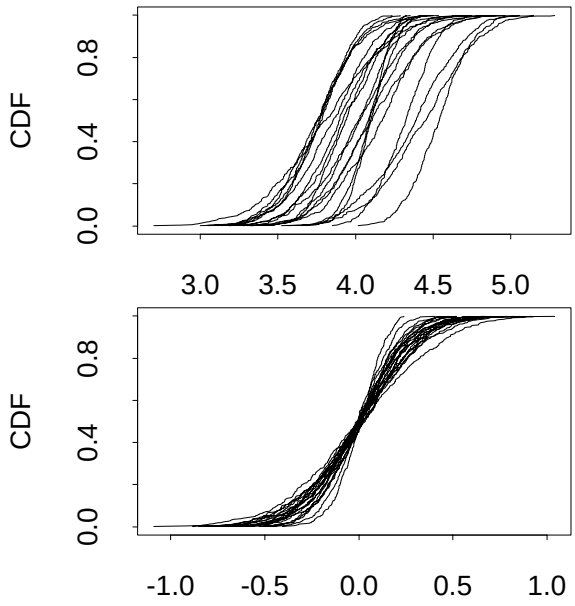

Figure 5: Bootstrap-after-bootstrap sampling distribution of $\bar{X}^*$ and $\bar{X}^* - \bar{X}$.

viewed as an approximation to "maximum likelihood tilting", with weights of the form

$$w_i = \frac{c}{1 - \tau(x_i - \bar{x})}.$$

The ML weights maximize $\prod w_i$ subject to $\sum w_1 = 1$ and $\sum w_i x_i = \mu$ (for any μ), and have nicer statistical properties, producing more conservative (and more accurate) inferences in confidence interval and hypothesis testing situations [2, 1, 6, 5], and have connections to empirical likelihood [4, 7].

In a hypothesis testing setting τ is chosen to satisfy $\sum w_i x_i = \mu_0$. In bootstrap tilting confidence intervals, τ is chosen so that $P(\bar{X}^* \leq \bar{x}) = \alpha/2$ or $(1 - \alpha/2)$. For bootstrap tilting diagnostics [3], we use the same τ values as for confidence intervals, and possibly some intermediate values, in order to investigate sensitivity over a likely range of values of θ.

Tilting can be generalized to statistics other than the mean by replacing $x_i - \bar{x}$ in the tilting formulae with the empirical influence function, and replacing $\sum w_i x_i$ with the appropriate statistic calculated for a weighted distribution.

4 Example 2

This example is similar to the first example, except now the original data are skewed. Figure 6 shows the bootstrap distributions when sampling from the empirical distribution (center) and weighted empirical distributions (using tilting). Note that in this case the variance changes as θ changes.

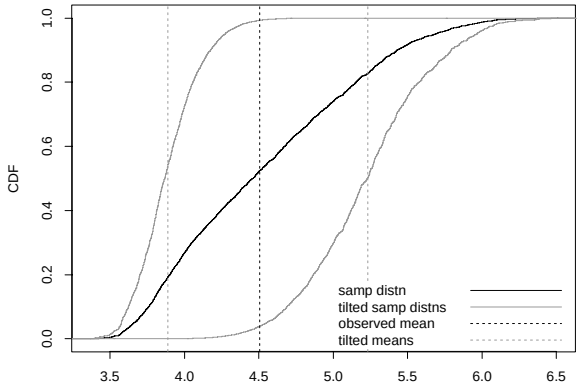

Figure 6: Bootstrap tilting diagnostics, for the distribution of $\bar{X}^*$, when sampling from a skewed dataset.

5 Relative Advantages of Diagnostic Procedures

These two examples point out a key characteristic of bootstrap tilting diagnostics — it measures the sensitivity of the sampling distribution to changes in θ. For the relatively symmetrical data in Example 1, changing the mean did not change the variance, while in Example 2 changing the mean did change the variance.

For comparison, recall the BAB results for the first example, in which the variance of the bootstrap distributions varied, but it was not clear whether there was any relationship between the variance of the distributions and the quantity of interest, the mean.

Hence bootstrap tilting diagnostics have the advantage of focusing on the quantity of interest.

Bootstrap-after-bootstrap diagnostics have the advantage of showing many possible ways in which the sampling distribution could vary. And JAB diagnostics have the advantage of showing the influence of individual observations on the sampling distribution.

JAB and bootstrap tilting have computational advantages, in that the diagnostics can be computed without actually generating bootstrap samples from the weighted empirical distributions. In the case of JAB, this is done by omitting all bootstrap samples which include the observation which is assigned zero weight. In the case of bootstrap tilting it is done by importance sampling, assigning different weights to bootstrap samples, proportional to the product of the w_i weights for the observations in the bootstrap sample. In contrast, BAB is computationally expensive, requiring a set of second-level bootstrap samples from each first-level bootstrap sample.

References

[1] T. J. DiCiccio and J. P. Romano. Nonparametric confidence limits by resampling methods and

least favorable families. *International Statistical Review*, 58(1):59–76, 1990.

[2] B. Efron. Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9:139 – 172, 1981.

[3] Tim C. Hesterberg. Bootstrap tilting inference and diagnostics. Grant application to N.S.F., November 1996.

[4] Tim C. Hesterberg. The bootstrap and empirical likelihood. In *Proceedings of the Statistical Computing Section*, pages 34–36. American Statistical Association, 1997.

[5] Tim C. Hesterberg. Bootstrap tilting confidence intervals. Research Department 84, MathSoft, Inc., 1700 Westlake Ave. N., Suite 500, Seattle, WA 98109, 1999.

[6] Tim C. Hesterberg. Bootstrap tilting confidence intervals and hypothesis tests. In K. Berk and M. Pourahmadi, editors, *Computer Science and Statistics: Proceedings of the 31st Symposium on the Interface*, volume 31, pages 389–393. Interface Foundation of North America, 1999.

[7] Art Owen. Empirical likelihood confidence regions. *Annals of Statistics*, 18:90–120, 1990.