A General Bootstrap Performance Diagnostic

Ariel Kleiner
Computer Science Division
UC Berkeley
akleiner@cs.berkeley.edu

Ameet Talwalkar
Computer Science Division
UC Berkeley
ameet@cs.berkeley.edu

Sameer Agarwal
Computer Science Division
UC Berkeley
sameerag@cs.berkeley.edu

Ion Stoica
Computer Science Division
UC Berkeley
istoica@cs.berkeley.edu

Michael I. Jordan
Computer Science Division
UC Berkeley
jordan@cs.berkeley.edu

ABSTRACT

As datasets become larger, more complex, and more available to diverse groups of analysts, it would be quite useful to be able to automatically and generically assess the quality of estimates, much as we are able to automatically train and evaluate predictive models such as classifiers. However, despite the fundamental importance of estimator quality assessment in data analysis, this task has eluded highly automatic solutions. While the bootstrap provides perhaps the most promising step in this direction, its level of automation is limited by the difficulty of evaluating its finite sample performance and even its asymptotic consistency. Thus, we present here a general diagnostic procedure which directly and automatically evaluates the accuracy of the bootstrap's outputs, determining whether or not the bootstrap is performing satisfactorily when applied to a given dataset and estimator. We show that our proposed diagnostic is effective via an extensive empirical evaluation on a variety of estimators and simulated and real datasets, including a real-world query workload from Conviva, Inc. involving 1.7TB of data (i.e., approximately 0.5 billion data points).

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: *nonparametric statistics, statistical computing*

Keywords

bootstrap; performance; diagnostic; estimator quality assessment

1. INTRODUCTION

Modern datasets are growing rapidly in size and are increasingly subjected to diverse, rapidly evolving sets of complex and exploratory queries, often crafted by non-statisticians. These developments render generic applicability and automation of data analysis methodology particularly desirable, both to allow the statistician to work more efficiently and to allow the non-statistician to correctly and effectively utilize more sophisticated inferential techniques. For example, the development of generic techniques for training classifiers and evaluating their generalization ability has allowed this methodology to spread well beyond the boundaries of the machine learning and statistics research community, to great practical benefit. More generally, estimation techniques for a variety of settings have been rendered generically usable. However, except in some restricted settings, the fundamental inferential problem of assessing the quality of estimates based upon finite data has eluded a highly automated solution.

Assessment of an estimate's quality—for example, its variability (e.g., in the form of a confidence region), its bias, or its risk—is essential to both its interpretation and use. Indeed, such quality assessments underlie a variety of core statistical tasks, such as calibrated inference regarding parameter values, bias correction, and hypothesis testing. Beyond simply enabling other statistical methodology, however, estimator quality assessments can also have more direct utility, whether by improving human interpretation of inferential outputs or by allowing more efficient management of data collection and processing resources. For instance, we might seek to collect or process only as much data as is required to yield estimates of some desired quality, thereby avoiding the cost (e.g., in time or money) of collecting or processing more data than is necessary. Such an approach in fact constitutes an active line of work in research on large database systems, which seeks to answer queries on massive datasets quickly by only applying them to subsamples of the total available data [1, 15]. The result of applying a query to only a subsample is in fact an estimate of the query's output if applied to the full dataset, and effective implementation of a system using this technique requires an automatic ability to accurately assess the quality of such estimates for generic queries.

In recent decades, the bootstrap [7, 9] has emerged as a powerful and widely used means of assessing estimator quality, with its popularity due in no small part to its relatively generic applicability. Unlike classical methods—which have generally relied upon analytic asymptotic approximations requiring deep analysis of specific classes of estimators in specific settings [17]—the bootstrap can be straightforwardly applied, via a simple computational mechanism, to

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

ACM 978-1-4503-2174-7/13/08 ...$15.00.

a broad range of estimators. Since its inception, theoretical work has shown that the bootstrap is broadly consistent [3, 10, 20] and can be higher-order correct [11]. As a result, the bootstrap (and its various relatives and extensions) provides perhaps the most promising avenue for obtaining a generically applicable, automated estimator quality assessment capability.

Unfortunately, however, while the bootstrap is relatively automatic in comparison to its classical predecessors, it remains far from being truly automatically usable, as evaluating and ensuring its accuracy is often a challenge even for experts in the methodology. Indeed, like any inferential procedure, despite its excellent theoretical properties and frequently excellent empirical performance, the bootstrap is not infallible. For example, it may fail to be consistent in particular settings (i.e., for particular pairs of estimators and data generating distributions) [19, 4]. While theoretical conditions yielding consistency are well known, they can be non-trivial to verify analytically and provide little useful guidance in the absence of manual analysis. Furthermore, even if consistent, the bootstrap may exhibit poor performance on finite samples.

Thus, it would be quite advantageous to have some means of diagnosing poor performance or failure of the bootstrap in an automatic, data-driven fashion, without requiring significant manual analysis. That is, we would like a diagnostic procedure which is analogous to the manner in which we evaluate performance in the setting of supervised learning (e.g., classification), in which we directly and empirically evaluate generalization error (e.g., via a held-out validation set or cross-validation). Unfortunately, prior work on bootstrap diagnostics (see [5] for a comprehensive survey) does not provide a satisfactory solution, as existing diagnostic methods target only specific bootstrap failure modes, are often brittle or difficult to apply, and generally lack substantive empirical evaluations. For example, a theoretical result of Beran regarding bootstrap asymptotics has been proposed as the basis of a diagnostic for bootstrap inconsistency [2]; however, it is unclear how to reliably construct and interpret the diagnostic plots required by this proposal, and the limited existing empirical evaluation reveals it to be of questionable practical utility [5]. Other work has sought to diagnose bootstrap failure specifically due to incorrect standardization of the quantity being bootstrapped (which could occur if an estimator's convergence rate is unknown or incorrectly determined), use of an incorrect resampling model (if, for example, the data has a correlation structure that is not fully known a priori), or violation of an assumption of pivotality of the quantity being bootstrapped [5]. Additionally, jackknife-after-bootstrap and bootstrap-after-bootstrap calculations have been proposed as a means of evaluating the stability of the bootstrap's outputs [8, 5]; while such procedures can be useful data analysis tools, their utility as the basis of a diagnostic remains limited, as, among other things, it is unclear whether they will behave correctly in settings where the bootstrap is inconsistent.

In contrast to prior work, we present here a general bootstrap performance diagnostic which does not target any particular bootstrap failure mode but rather directly and automatically determines whether or not the bootstrap is performing satisfactorily (i.e., providing sufficiently accurate outputs) when applied to a given dataset and estimator. The key difficulty in evaluating the accuracy of the bootstrap's (or any estimator quality assessment procedure's) outputs is the lack of ready availability of even approximate comparisons to ground truth estimate quality. While comparisons to ground truth labels are readily obtained in the case of supervised learning via use of a held-out validation set or cross-validation, comparing to ground truth in the context of estimator quality assessment requires access to the (unknown) sampling distribution of the estimator in question. We surmount this difficulty by constructing a proxy to ground truth for various small sample sizes (smaller than that of our full observed dataset) and comparing the bootstrap's outputs to this proxy, requiring that they converge to the ground truth proxy as the sample size is increased. This approach is enabled by the increasing availability of large datasets and more powerful computational resources. We show via an extensive empirical evaluation, on a variety of estimators and simulated and real data, that the resulting diagnostic is effective in determining—fully automatically—whether or not the bootstrap is performing satisfactorily in a given setting.

In Section 2, we formalize our statistical setting and notation. We introduce our diagnostic in full detail in Section 3. Sections 4 and 5 present the results of our evaluations on simulated and real data, respectively. Finally, we conclude in Section 6.

2. SETTING AND NOTATION

We assume that we observe n data points $\mathcal{D} = (X_1, \ldots, X_n)$ sampled i.i.d. from some unknown distribution P; let $\mathbb{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$ be the empirical distribution of the observed data. Based upon this dataset, we form an estimate $\hat{\theta}(\mathcal{D})$ of some parameter $\theta(P)$ of P; note that, unlike $\theta(P)$, $\hat{\theta}(\mathcal{D})$ is a random quantity due to its dependence on the data $\mathcal{D}$. We then seek to form an assessment $\xi(P, n)$ of the quality of the estimate $\hat{\theta}(\mathcal{D})$, which consists of a summary of the distribution Q_n of some quantity $u(\mathcal{D}, P)$. Our choice of summary and form for u depends upon our inferential goals and our knowledge of the properties of $\hat{\theta}$. For instance, $\xi(P, n)$ might compute an interquantile range for $u(\mathcal{D}, P) = \hat{\theta}(\mathcal{D})$, the expectation of $u(\mathcal{D}, P) = \hat{\theta}(\mathcal{D}) - \theta(P)$ (i.e., the bias), or a confidence interval based on the distribution of $u(\mathcal{D}, P) = n^{1/2}(\hat{\theta}(\mathcal{D}) - \theta(P))$. Unfortunately, we cannot compute $\xi(P, n)$ directly because P and Q_n are unknown, and so we must resort to estimating $\xi(P, n)$ based upon a single observed dataset $\mathcal{D}$.

The bootstrap addresses this problem by estimating the unknown $\xi(P, n)$ via the plug-in approximation $\xi(\mathbb{P}_n, n)$. Although computing $\xi(\mathbb{P}_n, n)$ exactly is typically intractable, we can obtain an accurate approximation using a simple Monte Carlo procedure: repeatedly form simulated datasets $\mathcal{D}^*$ of size n by sampling n points i.i.d. from $\mathbb{P}_n$, compute $u(\mathcal{D}^*, \mathbb{P}_n)$ for each simulated dataset, form the empirical distribution $\mathbb{Q}_n$ of the computed values of u, and return the desired summary of this distribution. We overload notation somewhat by referring to this final bootstrap output as $\xi(\mathbb{Q}_n, n)$, allowing ξ to take as its first argument either a data generating distribution or a distribution of u values.

For ease of exposition, we assume below that ξ is real-valued, though the proposed methodology can be straightforwardly generalized (e.g., to contexts in which ξ produces elements of a vector space).

3. THE DIAGNOSTIC

We frame the task of evaluating whether or not the bootstrap is performing satisfactorily in a given setting as a decision problem: for a given estimator, data generating distribution P, and dataset size n, is the bootstrap's output sufficiently likely to be sufficiently near the ground truth value $\xi(P, n)$? This formulation avoids the difficulty of producing uniformly precise quantifications of the bootstrap's accuracy by requiring only that a decision be rendered based upon some definition of "sufficiently likely" and "sufficiently near the ground truth." Nonetheless, in developing a diagnostic procedure to address this decision problem, we face the key difficulties of determining the distribution of the bootstrap's outputs on datasets of size n and of obtaining even an approximation to the ground truth value against which to evaluate this distribution.

Ideally, we might approximate $\xi(P, n)$ for a given value of n by observing many independent datasets, each of size n. For each dataset, we would compute the corresponding value of u, and the resulting collection of u values would approximate the distribution Q_n, which would in turn yield a direct approximation of the ground truth value $\xi(P, n)$. Furthermore, we could approximate the distribution of bootstrap outputs by simply running the bootstrap on each dataset of size n. Unfortunately, however, in practice we only observe a single set of n data points, rendering this approach an unachievable ideal.

To surmount this difficulty, our diagnostic (Algorithm 1) executes this ideal procedure for dataset sizes smaller than n. That is, for a given $p \in \mathbb{N}$ and $b \leq \lfloor n/p \rfloor$, we randomly sample p disjoint subsets of the observed dataset $\mathcal{D}$, each of size b. For each subset, we compute the value of u; the resulting collection of u values approximates the distribution Q_b, in turn yielding a direct approximation of $\xi(P, b)$, the ground truth value for the smaller dataset size b. Additionally, we run the bootstrap on each of the p subsets of size b, and comparing the distribution of the resulting p bootstrap outputs to our ground truth approximation, we can determine whether or not the bootstrap performs acceptably well at sample size b.

It then remains to use this ability to evaluate the bootstrap's performance at smaller sample sizes to determine whether or not it is performing satisfactorily at the full sample size n. To that end, we evaluate the bootstrap's performance at multiple smaller sample sizes to determine whether or not the distribution of its outputs is in fact converging to the ground truth as the sample size increases, thereby allowing us to generalize our conclusions regarding performance from smaller to larger sample sizes. Indeed, determining whether or not the bootstrap is performing satisfactorily for a single smaller sample size b alone is inadequate for our purposes, as the bootstrap's performance may degrade as sample size increases, so that it fails at sample size n despite appearing to perform sufficiently well at smaller sample size b. Conversely, the bootstrap may exhibit mediocre performance for small sample sizes but improve as it is applied to more data.

Thus, our diagnostic compares the distribution of bootstrap outputs to the ground truth approximation for an increasing sequence of sample sizes $b_1, \ldots, b_k$, with $b_k \leq \lfloor n/p \rfloor$; subsamples of each of these sizes are constructed and processed in the outer for loop of Algorithm 1. In order to conclude that the bootstrap is performing satisfactorily at sample size n, the diagnostic requires that the distribution of its outputs converges monotonically to the ground truth approximation for all of the smaller sample sizes $b_1, \ldots, b_k$. Convergence is assessed based on absolute relative deviation of the mean of the bootstrap outputs from the ground truth approximation (which must decrease with increasing sample size) and size of the standard deviation of the bootstrap outputs relative to the ground truth approximation (which must also decrease with increasing sample size). In Algorithm 1, this convergence assessment is performed by conditions (1) and (2). As a practical matter, these conditions do not require continuing decreases in the absolute relative mean deviation Δ_i or relative standard deviation σ_i when these quantities are below some threshold (given by c_1 and c_2, respectively) due to inevitable stochastic error in their estimation: when these quantities are sufficiently small, stochastic error due to the fact that we have only used p subsamples prevents reliable determination of whether or not decreases are in fact occurring. We have found that $c_1 = c_2 = 0.2$ is a reasonable choice of the relevant thresholds. Figures 1 and 2 highlight the use of conditions (1) and (2) in both positive and negative settings for the bootstrap.

Progressive convergence of the bootstrap's outputs to the ground truth is not alone sufficient, however; although the bootstrap's performance may be improving as sample size increases, a particular value of n may not be sufficiently large to yield satisfactory performance. Therefore, beyond the convergence assessment discussed above, we must also determine whether or not the bootstrap is in fact performing sufficiently well for the user's purposes at sample size n. We define "sufficiently well" as meaning that with probability at least $\alpha \in [0, 1]$, the output of the bootstrap when run on a dataset of size n will have absolute relative deviation from ground truth of at most c_3 (the absolute relative deviation of a quantity γ from a quantity γ_o is defined as $|\gamma - \gamma_o|/|\gamma_o|$); the constants α and c_3 are specified by the user of the diagnostic procedure based on the user's inferential goals. Because we can only directly evaluate the bootstrap's performance at smaller sample sizes (and not at the full sample size n), we take a conservative approach, motivated by the assumption that a false positive (incorrectly concluding that the bootstrap is performing satisfactorily) is substantially less desirable than a false negative. In particular, as embodied in condition (3) of Algorithm 1, we require that the bootstrap is performing sufficiently well under the aforementioned definition at the sample size b_k. Satisfying this condition, in conjunction with satisfying the preceding conditions indicating continuing convergence to the ground truth, is taken to imply that the bootstrap will continue to perform satisfactorily when applied to the full sample size n (in fact, the bootstrap's performance at sample size n will likely exceed that implied by α and c_3 due to the diagnostic's conservatism).

It is worth noting that this diagnostic procedure reposes on the availability in modern data analysis of both substantial quantities of data and substantial computational resources. For example, with $p = 100$ (an empirically effective choice), using $b_k = 1{,}000$ or $b_k = 10{,}000$ requires $n \geq 10^5$ or $n \geq 10^6$, respectively. Fortuitously, datasets of such sizes are now commonplace. Regarding its computational requirements, our procedure benefits from the modern shift toward parallel and distributed computing, as the vast

Algorithm 1: Bootstrap Performance Diagnostic

```
Input: $\mathcal{D} = (X_1, \ldots, X_n)$: observed data
       $u$: quantity whose distribution is summarized to yield estimator quality assessments
       $\xi$: estimator quality assessment
       $p$: number of disjoint subsamples used to compute ground truth approximations (e.g., 100)
       $b_1, \ldots, b_k$: increasing sequence of subsample sizes for which ground truth approximations are computed,
                 with $b_k \leq \lfloor n/p \rfloor$ (e.g., $b_i = \lfloor n/(p2^{k-i}) \rfloor$ with $k = 3$)
       $c_1 \geq 0$: tolerance for decreases in absolute relative deviation of mean bootstrap output (e.g., 0.2)
       $c_2 \geq 0$: tolerance for decreases in relative standard deviation of bootstrap output (e.g., 0.2)
       $c_3 \geq 0, \alpha \in [0, 1]$: desired probability $\alpha$ that bootstrap output at sample size $n$ has absolute relative
                 deviation from ground truth less than or equal to $c_3$ (e.g., $c_3 = 0.5, \alpha = 0.95$)
Output: true if the bootstrap is deemed to be performing satisfactorily, and false otherwise
$\mathbb{P}_n \leftarrow n^{-1} \sum_{i=1}^{n} \delta_{X_i}$
for $i \leftarrow 1$ to $k$ do
    $\mathcal{D}_{i1}, \ldots, \mathcal{D}_{ip} \leftarrow$ random disjoint subsets of $\mathcal{D}$, each containing $b_i$ data points
    for $j \leftarrow 1$ to $p$ do
        $u_{ij} \leftarrow u(\mathcal{D}_{ij}, \mathbb{P}_n)$
        $\xi^*_{ij} \leftarrow bootstrap(\xi, u, b_i, \mathcal{D}_{ij})$
    end
    // Compute ground truth approximation for sample size $b_i$
    $\mathbb{Q}_{b_i} \leftarrow \sum_{j=1}^{p} \delta_{u_{ij}}$
    $\tilde{\xi}_i \leftarrow \xi(\mathbb{Q}_{b_i}, b_i)$
    // Compute absolute relative deviation of mean of bootstrap outputs and
    // relative standard deviation of bootstrap outputs for sample size $b_i$
    $\Delta_i \leftarrow \left| \frac{mean(\xi^*_{i1}, \ldots, \xi^*_{ip}) - \tilde{\xi}_i}{\tilde{\xi}_i} \right|$        $\sigma_i \leftarrow \left| \frac{stddev(\xi^*_{i1}, \ldots, \xi^*_{ip})}{\tilde{\xi}_i} \right|$
end
return true if all of the following hold, and false otherwise:
    $\Delta_{i+1} < \Delta_i \quad OR \quad \Delta_{i+1} \leq c_1, \quad \forall i = 1, \ldots, k, \qquad (1)$
    $\sigma_{i+1} < \sigma_i \quad OR \quad \sigma_{i+1} \leq c_2, \quad \forall i = 1, \ldots, k, \qquad (2)$
    $\frac{\# \left\{ j \in 1, \ldots, p : \left| \frac{\xi^*_{kj} - \tilde{\xi}_k}{\tilde{\xi}_k} \right| \leq c_3 \right\}}{p} \geq \alpha \qquad (3)$
```

majority of the required computation occurs in the inner for loop of Algorithm 1, the iterations of which are independent and individually process only small data subsets. Additionally, we have sought to reduce the procedure's computational costs by using an identical number of subsamples p for each subsample size $b_1, \ldots, b_k$; one could presumably improve statistical performance by using larger numbers of subsamples for smaller subsample sizes.

The guidelines given in Algorithm 1 for setting the diagnostic procedure's hyperparameters are motivated by the procedure's structure and have proven to be empirically effective. We recommend exponential spacing of the $b_1, \ldots, b_k$ to help ensure that reliable comparisons of bootstrap performance can be made across adjacent sample sizes b_i and b_{i+1}. However, by construction, setting the $b_1, \ldots, b_k$ to be too close together should primarily cause an increase in the false negative rate (the probability that the diagnostic incorrectly concludes that the bootstrap is not performing satisfactorily), rather than a less desirable increase in the false positive rate. Similarly, setting c_1 or c_2 to be too low should also primarily result in an increase in the false negative rate. Regarding c_3 and α, these hyperparameters should be determined by the user's bootstrap performance desiderata. We nonetheless expect that fairly lenient settings of c_3—such as $c_3 = 0.5$, which corresponds to allowing the bootstrap to deviate from ground truth by up to 50%—to be reasonable in many cases. This expectation stems from the fact that the actual or targeted quality of estimators on fairly large datasets is frequently high, leading to estimator quality assessments, such as interquantile ranges, which are small in absolute value; in these cases, it follows that a seemingly large relative error in bootstrap outputs (e.g., 50%) corresponds to a small absolute error.

As we demonstrate via an extensive empirical evaluation on both synthetic and real data in the following sections, our proposed bootstrap performance diagnostic is quite effective, with false positive rates that are generally extremely low or zero and false negative rates that generally approach zero as the subsample sizes $b_1, \ldots, b_k$ are increased. Of course, like any inferential procedure, our procedure does have some unavoidable limitations, such as in cases where the data generating distribution has very fine-grained adverse features which cannot be reliably observed in datasets of size b_k; we discuss these issues further below.

4. SIMULATION STUDY

We first evaluate the diagnostic's effectiveness on data generated from a variety of different synthetic distributions

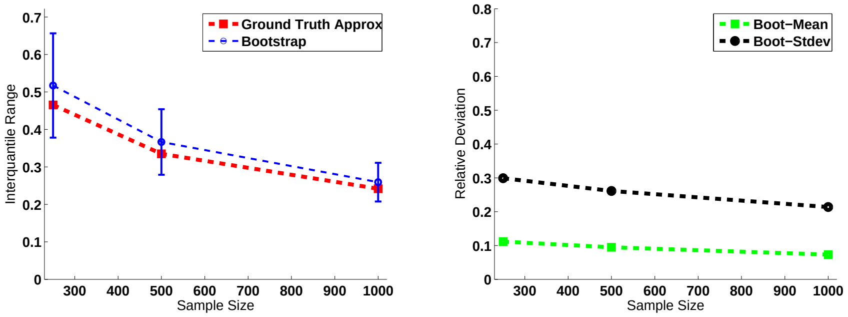

Figure 1: Illustration of the quantities computed by the diagnostic on a dataset and estimator for which the bootstrap performs satisfactorily. The left plot shows bootstrap outputs (means with standard deviations) and ground truth approximations for sample sizes $b_1, \ldots, b_3$. The right plot shows the absolute relative deviation of the mean of the bootstrap outputs and the relative standard deviation of the bootstrap outputs for each sample size.

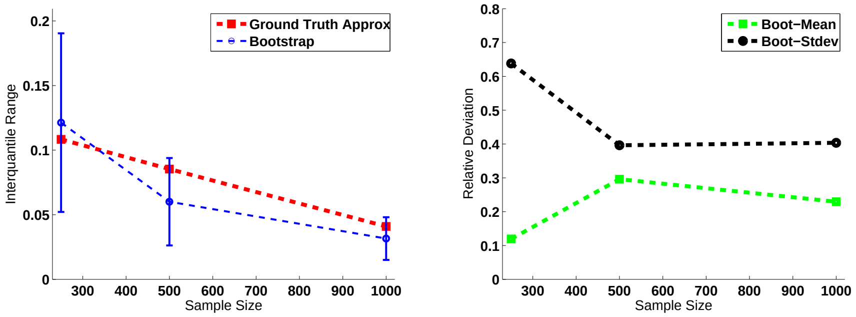

Figure 2: Illustration of the quantities computed by the diagnostic on a dataset and estimator for which the bootstrap does not perform satisfactorily. The left plot shows bootstrap outputs (means with standard deviations) and ground truth approximations for sample sizes $b_1, \ldots, b_3$. The right plot shows the absolute relative deviation of the mean of the bootstrap outputs and the relative standard deviation of the bootstrap outputs for each sample size.

paired with a variety of different estimators. Using simulated data here allows direct knowledge of the ground truth value $\xi(P, n)$, and by selecting different synthetic distributions, we can design settings that pose different challenges to the diagnostic procedure. For each distribution-estimator pair and sample size n considered, we perform multiple independent runs of the diagnostic on independently generated datasets of size n to compute the Diagnostic True Rate (DTR), the probability that the diagnostic outputs *true* in that setting. We then evaluate this DTR against the bootstrap's actual performance on datasets of size n; because the underlying data generating distributions here are known, we can also compare to known theoretical expectations of bootstrap consistency.

More precisely, we consider the following data generating distributions: Normal$(0, 1)$, Uniform$(0, 10)$, StudentT(1.5), StudentT(3), Cauchy$(0, 1)$, 0.95Normal$(0, 1)$ $+$ 0.05Cauchy$(0, 1)$, and 0.99Normal$(0, 1)$ $+$ 0.01Cauchy$(10^4, 1)$. In our plots, we denote these distributions using the following abbreviations: Normal, Uniform, StuT(1.5), StuT(3), Cauchy, Mixture1, and Mixture2. We also consider the following estimators $\hat{\theta}$ (abbreviations, if any, are given in parentheses): mean, median (med), variance (var), standard deviation (std), sample maximum (max), and 95th percentile (perc). The estimator quality assessment ξ in all experiments computes the interquantile range between the 0.025 and 0.975 quantiles of the distribution of $u(\mathcal{D}, P) = \hat{\theta}(\mathcal{D})$. For all runs of the bootstrap, we use between 200 and 500 resamples, with the precise number of resamples determined by the adaptive hyperparameter selection procedure given by [13]. All runs of the diagnostic use the hyperparameter guidelines given in Algorithm 1: $p = 100, k = 3, b_i = \lfloor n/(p2^{k-i}) \rfloor, c_1 = 0.2, c_2 = 0.2, c_3 = 0.5, \alpha = 0.95$. We consider sample sizes $n = 10^5$ and $n = 10^6$.

For each distribution-estimator pair and sample size n, we first compute the ground truth value $\xi(P, n)$ as the interquantile range of the u values for 5,000 independently generated datasets of size n. We also approximate the distribution of bootstrap outputs on datasets of size n by running the bootstrap on 100 independently generated datasets of this size. Whether or not this distribution of bootstrap

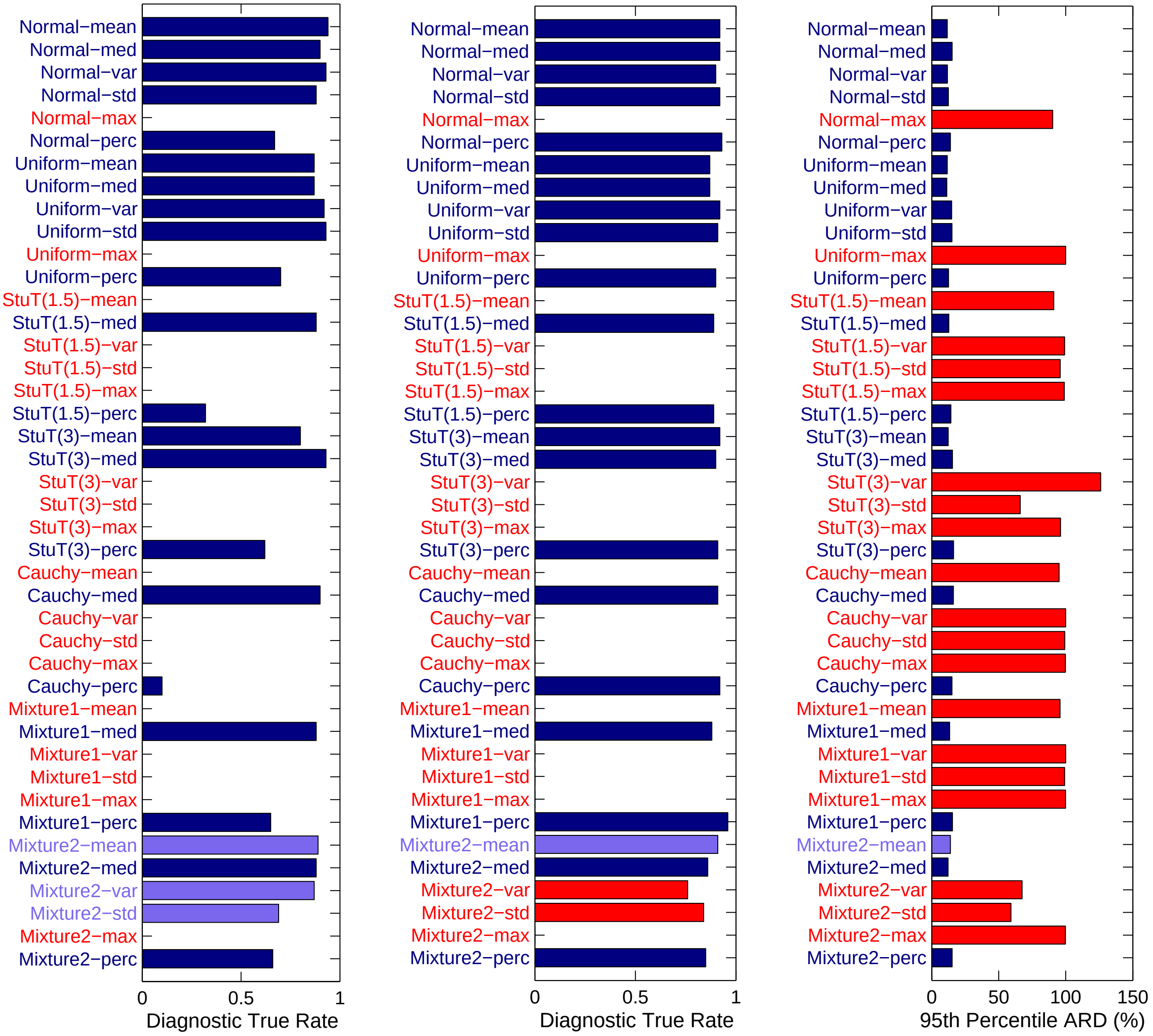

Figure 3: Diagnostic and bootstrap performance on simulated data. Dark blue indicates cases where bootstrap is performing satisfactorily on datasets of size n (based on ground truth computations) and is expected theoretically to be consistent; red indicates cases where neither of these statements is true; light purple indicates cases where bootstrap is performing satisfactorily on datasets of size n (based on ground truth computations) but is not expected theoretically to be consistent. Left and middle plots: for each distribution-estimator pair, fraction of 100 independent trials for which the diagnostic outputs *true*. For the left plot, $n = 10^5$; for the middle plot, $n = 10^6$. Right plot: for each distribution-estimator pair, 95th percentile of absolute relative deviation of bootstrap output from ground truth, over 100 independent trials on datasets of size $n = 10^6$.

outputs satisfies the performance criterion defined by c_3, α—that is, whether or not the α quantile of the absolute relative deviation of bootstrap outputs from $\xi(P, n)$ is less than or equal to c_3—determines the ground truth conclusion regarding whether or not the bootstrap is performing satisfactorily in a given setting. To actually evaluate the diagnostic's effectiveness, we then run it on 100 independently generated datasets of size n and estimate the DTR as the fraction of these datasets for which the diagnostic returns *true*. If the ground truth computations deemed the bootstrap to be performing satisfactorily in a given setting, then the DTR would ideally be 1, and otherwise it would ideally be 0.

Figure 3 presents our results for all distribution-estimator pairs and both sample sizes n considered. In these plots, dark blue indicates cases in which the ground truth computations on datasets of size n deemed the bootstrap to be performing satisfactorily and the bootstrap is expected theoretically to be consistent (i.e., the DTR should ideally be 1); red indicates cases in which neither of these statements is true (i.e., the DTR should ideally be 0); and light purple indicates cases in which the ground truth computations on datasets of size n deemed the bootstrap to be performing satisfactorily but the bootstrap is not expected theoretically to be consistent (i.e., the DTR should ideally be 1).

As seen in the lefthand and middle plots (which show DTRs for $n = 10^5$ and $n = 10^6$, respectively), our proposed diagnostic performs quite well across a range of data generating distributions and estimators, and its performance improves as it is provided with more data. For the smaller sample size $n = 10^5$, in the dark blue and light purple cases, the DTR is generally markedly greater than 0.5; furthermore, when the sample size is increased to $n = 10^6$, the DTRs in all of the dark blue and light purple cases increase to become uniformly near 1, indicating low false negative rates (i.e., the diagnostic nearly always deems the bootstrap to be performing satisfactorily when it is indeed performing satisfactorily). In the red cases, for both sample sizes, the DTR is nearly always zero, indicating that false positive rates are nearly always zero (i.e., the diagnostic only rarely deems the bootstrap to be performing satisfactorily when it is in fact not performing satisfactorily). Mixture2-var and Mixture2-std with $n = 10^6$ provide the only exceptions to this result, which is unsurprising given that Mixture2 was specifically designed to include a small heavy-tailed component which is problematic for the bootstrap but cannot be reliably detected at the smaller sample sizes $b_1, \ldots, b_k$; nonetheless, even in these cases, the righthand plot indicates that the ground truth computations very nearly deemed the bootstrap to be performing satisfactorily. Interestingly, the bootstrap's finite sample performance for the settings considered nearly always agrees with theoretical expectations regarding consistency; disagreement occurs only when Mixture2 is paired with the estimators mean, var, or std, which is again unsurprising given the properties of Mixture2.

5. REAL DATA

We next evaluate the diagnostic's effectiveness on real datasets obtained from Conviva, Inc. [6], which are routinely subjected to analysis by practitioners. Our first set of experiments pairs this real data with the (synthetic) estimators considered in the previous section; we then consider a larger dataset paired with a set of 268 production SQL queries.

Synthetic Estimators

We present here the results of experiments on three real datasets obtained from [6], which describe different attributes of large numbers of video streams viewed by Internet users. These datasets are routinely subjected to a variety of different analyses by practitioners and are the subject of ongoing efforts to improve the computational efficiency of database systems by processing only data subsamples and quantifying the resulting estimation error [1].

We designate the three (scalar-valued) datasets as follows, with their sizes (i.e., numbers of constituent data points) given in parentheses: Conviva1 (30,470,092), Conviva2 (1,111,798,565), and Conviva3 (2,952,651,449). Histograms of the three datasets are given in Figure 4; note that the datasets are heavily skewed and also contain large numbers of repeated values. Due to privacy considerations, we are unable to provide the precise values and corresponding frequencies represented in the data, but the histograms nonetheless convey the shapes of the datasets' empirical distributions.

To circumvent the fact that ground truth values for individual real datasets cannot be obtained, we do not directly apply our diagnostic to these three datasets. Rather, we treat the empirical distribution of each dataset as an underlying data generating distribution which is used to generate the datasets used in our experiments. With this setup, our experiments on these real datasets proceed identically to the experiments in Section 4 above, but now with data sampled from the aforementioned empirical distributions rather than from synthetic distributions.

Figure 5 presents the results of our experiments on the Conviva data. The color scheme used in these plots is identical to that in Figure 3, with the addition of magenta, which indicates cases in which the ground truth computations on datasets of size n deemed the bootstrap to not be performing satisfactorily but the bootstrap is expected theoretically to be consistent (i.e., the DTR should ideally be 0). Given that the data generating distributions used in these experiments all have finite support, the bootstrap is expected theoretically to be consistent for all estimators considered except the sample maximum. However, as seen in the righthand plot of Figure 5, the bootstrap's finite sample performance is often quite poor even in cases where consistency is expected; in this regard (as well as in other ways), the real data setting of this section differs substantially from the synthetic data setting considered in Section 4 above.

The lefthand and middle plots of Figure 5 demonstrate that our diagnostic procedure again performs quite well. Indeed, the DTR is again nearly always zero (or is quite small if positive) in the red and magenta cases, indicating false positive rates that are nearly always zero. The dark blue cases generally have DTRs markedly greater than 0.5 for $n = 10^5$ (lefthand plot), with DTRs in these cases generally increasing to become nearly 1 for $n = 10^6$, indicating low false negative rates; no light purple cases occur for the real data. Beyond these broad conclusions, it is worth noting that the Conviva2-max, Conviva2-perc, and Conviva3-med settings exhibit rather surprising behavior relative to our other results, in that the diagnostic's performance seems to degrade when the sample size is increased. We believe that this behavior is related to the particularly high redundancy (i.e., degree of repetition of values) in Conviva2 and Conviva3, and it will be the subject of future work.

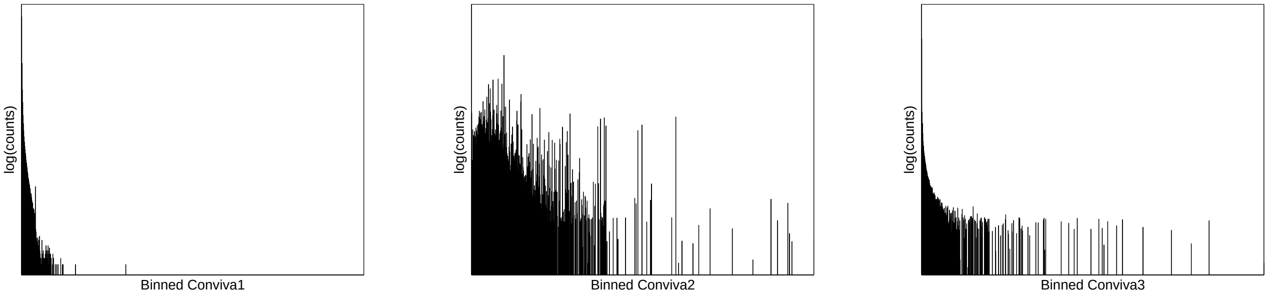

Figure 4: Histograms for the real datasets Conviva1, Conviva2, and Conviva3. Note that the y axes give frequencies on a log scale.

Production Estimators

We finally evaluate the diagnostic's effectiveness on a larger real dataset and accompanying real-world analytical workload derived from a SQL-based ad-hoc querying system at Conviva, Inc. [6]. This dataset is 1.7TB in size and contains approximately 0.5 billion records extracted from access logs describing video streams viewed by Internet users during a thirty day time span. We treat each record, which has 104 attributes (such as video genre, web browser type, request response time, etc.), as a data point.

This data is routinely processed by data analysts, who issue SQL queries which are run over the dataset to compute quantities of interest. For example, a typical query might filter the underlying data based on various attributes and then compute a quantity derived from other attributes (such as those given by Conviva1, Conviva2, and Conviva3 above). This computed quantity might correspond to a standard estimator (e.g., the sample mean or percentile), or it might be the result of an arbitrary computation performed by a User-Defined Function (UDF). We study a set of 268 queries selected randomly from the set of obtainable queries which were issued in a production setting against our dataset; 113 out of these 268 queries include UDFs which may perform arbitrary computations rather than simply using standard SQL operators.

In our experiments, we treat each query as an estimator which seeks to compute some real-valued population quantity; our estimator quality assessment ξ is again the interquantile range used in our previous experiments. To obtain ground truth conclusions regarding bootstrap performance, we randomly split the full dataset into disjoint chunks, each of size 5GB (approximately 2.5 million records), and compute both the point estimate and the bootstrap output for each chunk. We also run the diagnostic for each query; as in the experiments in preceding sections, we use the hyperparameter guidelines given in Algorithm 1 (i.e,. $p = 100, k = 3, c_1 = 0.2, c_2 = 0.2, c_3 = 0.5, \alpha = 0.95$), with the exception of now using $b_i = 10^5/2^{k-i}$ due to the large quantity of available data.

Recall that a query typically filters its input data on one or more attributes before computing its output. To address cases in which queries are highly selective and hence effectively operate on severely reduced quantities of data, we do not consider the 17 out of our 268 queries which filter out more than 99% of the data in any of the diagnostic subsamples; one might consider the diagnostic as simply not being applicable to such queries because the available subsamples are not sufficiently large after the queries apply their filters.

Of the remaining 251 queries, the diagnostic deemed the bootstrap to be performing satisfactorily on 224, with 9 false negatives and 7 false positives. Thus, on this real dataset with accompanying query workload, the diagnostic exhibited low false negative and false positive rates of 3.6% and 2.8%, respectively.

6. CONCLUSION

We have presented a general diagnostic procedure which permits automatic determination of whether or not the bootstrap is performing satisfactorily when applied to a given dataset and estimator; we have demonstrated the effectiveness of our procedure via an empirical evaluation on a variety of estimators and simulated and real data. A number of avenues of potential future work remain. For example, it would be interesting to apply our diagnostic procedure to other estimator quality assessment methods [4, 17, 13] and to devise extensions of the diagnostic which are suitable for variants of the bootstrap designed to handle non-i.i.d. data [9, 12, 14, 16, 18]. Additionally, it should be possible to characterize theoretically the consistency of our diagnostic procedure, showing that its false positive and false negative rates approach zero as $b_1, \ldots, b_k, p \to \infty$ and $c_1, c_2 \to 0$, under some assumptions (e.g., monotonicity of the bootstrap's convergence to ground truth in cases where it is performing satisfactorily). It would be interesting to make such a result precise.

7. ACKNOWLEDGMENTS

This research was supported in part by NSF CISE Expeditions award CCF-1139158 and DARPA XData Award FA8750-12-2-0331, and gifts from Amazon Web Services, Google, SAP, Blue Goji, Cisco, Clearstory Data, Cloudera, Ericsson, Facebook, General Electric, Hortonworks, Huawei, Intel, Microsoft, NetApp, Oracle, Quanta, Samsung, Splunk, VMware, and Yahoo!. Ameet Talwalkar was supported by NSF award No. 1122732.

8. REFERENCES

[1] S. Agarwal, A. Panda, B. Mozafari, S. Madden, and I. Stoica. Blinkdb: Queries with bounded errors and bounded response times on very large data, June 2012.

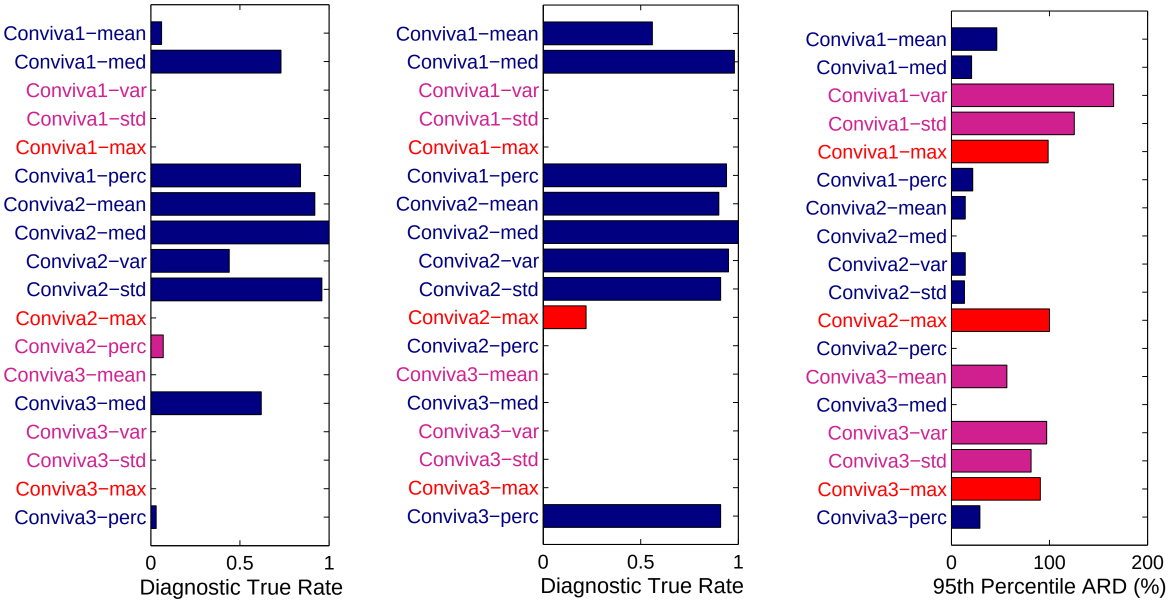

Figure 5: Diagnostic and bootstrap performance on the real datasets Conviva1, Conviva2, and Conviva3. Color scheme is identical to that in Figure 3, with the addition of magenta indicating cases where the bootstrap is not performing satisfactorily on datasets of size n (based on ground truth computations) but is expected theoretically to be consistent. Left and middle plots: for each distribution-estimator pair, fraction of 100 independent trials for which the diagnostic outputs *true*. For the left plot, $n = 10^5$; for the middle plot, $n = 10^6$. Right plot: for each distribution-estimator pair, 95th percentile of absolute relative deviation of bootstrap output from ground truth, over 100 independent trials on datasets of size $n = 10^6$.

[2] R. Beran. Diagnosing bootstrap success. *Annals of the Institute of Statistical Mathematics*, 49(1):1–24, 1997.

[3] P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9(6):1196–1217, 1981.

[4] P. J. Bickel, F. Gotze, and W. van Zwet. Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statistica Sinica*, 7:1–31, 1997.

[5] A. J. Canty, A. C. Davison, D. V. Hinkley, and V. Ventura. Bootstrap diagnostics and remedies. *The Canadian Journal of Statistics*, 34(1):5–27, 2006.

[6] Conviva, Inc. http://www.conviva.com, November 2012.

[7] B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979.

[8] B. Efron. Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society, Series B*, 54(1):83–127, 1992.

[9] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[10] E. Giné and J. Zinn. Bootstrapping general empirical measures. *Annals of Probability*, 18(2):851–869, 1990.

[11] P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag New York, Inc., 1992.

[12] P. Hall and E. Mammen. On general resampling algorithms and their performance in distribution estimation. *Annals of Statistics*, 22(4):2011–2030, 1994.

[13] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan. The big data bootstrap. In *International Conference on Machine Learning (ICML)*, 2012.

[14] H. R. Kunsch. The jacknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3):1217–1241, 1989.

[15] N. Laptev, K. Zeng, and C. Zaniolo. Early accurate results for advanced analytics on mapreduce. In *Proceedings of the VLDB Endowment*, volume 5, pages 1028–1039, 2012.

[16] R. Y. Liu and K. Singh. Moving blocks jackknife and bootstrap capture weak dependence. In R. LePage and L. Billard, editors, *Exploring the Limits of the Bootstrap*, pages 225–248. Wiley, 1992.

[17] D. Politis, J. Romano, and M. Wolf. *Subsampling*. Springer, 1999.

[18] D. N. Politis and J. P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994.

[19] H. Putter and W. R. van Zwet. Resampling: Consistency of substitution estimators. *Annals of Statistics*, 24(6):2297–2318, 1996.

[20] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag New York, Inc., 1996.