CDF/ANAL/TOP/GROUP/9081
Draft 1.0
October 31, 2007

The Bootstrap Technique and its Application to Analyses in the Top Mass Group

Jahred Adelman[1], Erik Brubaker[2]

University of Chicago

Abstract

Sample quantities such as the mean M_{top} residual or pull width measured in a particular Monte Carlo sample are increasingly important in top mass measurements; we use them to define, validate, and characterize our analyses. In many cases it is also essential to know the uncertainties on our estimates of these quantities due to the limited statistics of the sample. The bootstrap is a technique for estimating the statistical properties of a quantity using only the dataset from which the quantity was originally derived. It is computationally intensive, but quite general and robust. In this note we introduce the bootstrap technique, its strengths and its limitations, and as an example its application in the context of a template-based top quark mass measurement.

Contents

[1]jahred@fnal.gov
[2]brubakee@fnal.gov

1 The Problem

In many contexts in top quark mass analyses, we estimate "sample quantities," that is quantities that are properties of an entire Monte Carlo sample. Some examples are:

- The average mass residual
- The pull width
- The expected (mean, median) uncertainty

Note that all of these quantities are estimated using *ensembles* of pseudoexperiments. The relevant statistical quantity is the size of the Monte Carlo sample used. But the uncertainties on our estimates of these sample quantities are not easy to evaluate. Analytical approaches from first principles are hopeless as the techniques involved in the typical top quark mass measurement are extremely complicated. One might expect that since we generally deal with ensembles of pseudoexperiments, we can use standard methods (fits, RMS/$\sqrt{N}$, etc.) to evaluate uncertainty. But the fundamental independent unit of a Monte Carlo sample is the event, not the pseudoexperiment. Events are usually used in multiple pseudoexperiments (in different combinations), and subtle correlations are introduced. Thus many commonly used approaches to estimating the uncertainty on a sample quantity give incorrect answers.

One possible way to ensure reasonable uncertainty estimates on sample quantities is to avoid using a given event in multiple pseudoexperiments. Then the number of pseudoexperiments that can be performed is $N_{\text{indep}} = N_{\text{MC}}/N_{\text{data}}$, the number of events in the Monte Carlo sample divided by the number of events in the data—and thus in a typical pseudoexperiment. Then the usual statistical techniques (treating pseudoexperiments as independent entities) are valid. But this number N_{indep} is quite small, and one almost always gets more information by using events multiple times in different combinations.

Thus the problem becomes: how do we estimate the uncertainty on a quantity that arises due to the limited statistics of the Monte Carlo sample from which it is estimated, when the method used to estimate the quantity involves complicated operations and subtle correlations?

Finally a few words on why these uncertainties are important in top mass measurements. There are at least two contexts in which we use sample quantities.

- The performance of each analysis is validated using Monte Carlo samples that assume different values of the true top quark mass and sometimes different jet energy scales. Generally we look at the average mass residual over a range of true top quark masses, and require it to be consistent with 0; alternatively, a correction or mapping function is determined from these tests and applied to the data measurement. Similar checks are done for the pull width to characterize the validity of the reported uncertainty. Another interesting quantity, especially for comparing different analyses, is the *a priori* expected uncertainty. Finally, the

exercise is repeated using blind samples whose true mass is not known by the analysers. Again, these are all sample quantities and inaccurate uncertainties can lead either to e.g. false indications of bias (if the uncertainties are underestimated) or the hiding of real problems in the analysis (if the uncertainties are overestimated).

- Most of our systematic uncertainties are estimated using the shift method. The systematic is defined as the difference between the average mass measured in pseudoexperiments using the nominal MC sample and the average mass measured in pseudoexperiments using an MC sample "shifted" by one sigma in some parameter; both of those are sample quantities. In fact, when the uncertainty on the shift is larger than the shift itself, we conservatively use the uncertainty as the systematic. So in the current prescription, changing the uncertainty on the average mass can directly change our systematic uncertainties.

2 The Bootstrap

The bootstrap technique is a procedure, widely used in other disciplines, for approximating the sampling distribution of an observable using only the data in hand. Here we will apply the technique taking the "data" to be the Monte Carlo sample in question, and the observable to be one of the sample quantities described above. The key to the method is that it approximates sampling from the universal distribution by (re)sampling from the observed data. The canonical bootstrap references are Ref. [1, 2]; also, more general introductions can be readily found on the web.

2.1 Description

First let's review a brute force approach to understanding the uncertainty on a sample quantity: we would just generate a large number of independent samples, all of the same size, and look at the spread of our estimates of the sample quantity. For example, say we have a Monte Carlo sample with 1M $t\bar{t}$ events. We run a battery of pseudoexperiments, plot the resulting pull distribution, and find a pull width of 1.05. But what is the uncertainty on 1.05 due to the limited (1M) size of the MC sample? With unlimited computing resources, we could generate 50 more 1M samples, statistically independent but with the exact same settings, and plot the pull widths from each one. Assuming the distribution to be reasonably Gaussian, its RMS would be the uncertainty on 1.05.[3]

The bootstrap technique is used when it is not easy to generate more events from the "universe," i.e. from all possible MC events with a given M_{top}. Instead, we take the existing sample (1M events) as a reasonable approximation of the universe, and take our 50 additional samples from that original sample. Specifically, we form a "bootstrap

[3]Of course, there would be a strong temptation to use all 50M events to get a better estimate of the pull width itself; then we're back to the question of what's the uncertainty on the new, better, estimate?

sample" by sampling 1M events, with replacement, from the original 1M sample. Each event in the original sample will appear 0, 1, 2, or more times in the bootstrap sample, following a Poisson distribution.[4] Given the bootstrap sample, we determine the pull width following exactly the same procedure as in the original sample (run a battery of pseudoexperiments etc.). We repeat the procedure to form 50 bootstrap samples and 50 corresponding estimates of the pull width. As in the brute force case, the width of those 50 values is the uncertainty on our original estimate of the pull width. Note that there is no free lunch: with the bootstrap technique, we can never do a better job of estimating the sample quantity than we did with the original sample.[5] We are just approximating the sampling distribution of our estimate in order to understand its uncertainty.

The reason for choosing $N = 50$ bootstrap samples is that, in the case of a Gaussian sampling distribution, the uncertainty on the RMS is $\sigma_{\mathrm{RMS}} = \mathrm{RMS}/\sqrt{2N}$, or 10% for $N = 50$. That's good enough for the uncertainty on the uncertainty!

2.2 Limitations

We know of two limitations to the bootstrap procedure. The first arises when the original sample does not approximate the universe; thus resampling from the original sample does not approximate sampling from the universe. Usually this is a worry when the original sample size is small, so that due to fluctuations it might not include all the features of the universe. For example, imagine a bimodal distribution with 10% of events in the second peak. A sample of 10–20 events might not include any from the second peak; in that case no amount of resampling could make the missing peak appear. Since our application is to MC samples with minimum thousands of events, this is unlikely to be a problem.[6] Another possibility, again not expected to occur in top mass analyses, is pathological underlying distributions that can't be sampled correctly, such as infinite tails.

The second "limitation" is when the distribution of bootstrap estimates is not Gaussian. The problem in that case is not with the bootstrap technique itself, which just provides the aforementioned distribution. Rather it is with our notion of "uncertainty," which almost always presupposes Gaussian behavior. Even in such a case, however, the RMS of the distribution is usually the best measure of uncertainty to use.

2.3 Discussion

It is not hard to confuse the resampling involved in the bootstrap technique with the common procedure of resampling events when constructing pseudoexperiments.

[4] Actually, a multinomial, but we defy you to detect the difference with these statistics.

[5] In principle, any *bias* in the estimator due to the limited statistics of the sample will result in a shift of the distribution of the bootstrap estimates with respect to the original estimate. But this situation should be uncommon in our analyses.

[6] But be careful with multi-dimensional distributions. As the number of dimensions increases beyond one or two, you will quickly find it difficult to fully sample the space with your MC.

Please don't! The bootstrap always involves resampling, but not all resampling is part of a bootstrap. One way to keep these concepts separate is to consider that all procedures specific to a given analysis are inside a black box from the perspective of the bootstrap. The bootstrap only "knows" that for every MC sample, there exists a procedure to estimate a quantity called, e.g., "pull width." The fact that the procedure for estimating pull width involves complicated steps, including reusing MC events from the sample, is neither here nor there for the bootstrap. This very aspecificity is one reason the bootstrap is so appealing.

2.4 Application to top mass analyses

We now turn to the use of the bootstrap technique in top mass analyses at CDF.

The following pseudo-code gives an idea of how to implement the bootstrap procedure. Outside of the pseudoexperiment loop, the resampling is done by storing an array of random event numbers corresponding to the original sample. Then when drawing events for a given pseudoexperiment, the random entry is mapped back into an entry in the original tree/sample. The same should be done for each signal and background sample.[7] If a dedicated random number generator is used for the resampling, with the seed set explicitly for each bootstrap sample, then the resampling will be deterministic so that the corresponding pseudoexperiments can be separated into multiple jobs.

```
<Above, set different random number seed
 for each bootstrap sample>
int treenum[100000][nsamples];
// Once for each pseudo-sample
if (bootstrapsignal) {
  int ntree = (int)chain->GetEntries();
  assert(ntree<=100000);
  for (int ievent = 0; ievent < ntree; i++) {
      treenum[ievent][isample] = myRandom->Integer(ntree);
   }
}
// Do also for background

<Now inside PE loop>
<Inside loop drawing events for some PE>
      int entrytouse = randomentry;
      if (bootstrapsignal) {
        entrytouse = treenum[entrytouse][isample];
      }
```

[7] Sometimes it is interesting to determine the bootstrap uncertainty due to the statistics of signal and background events separately. In that case the bootstrap can be applied to only the relevant samples.

A few additional comments:

- As we just discussed, in top mass analyses the typical procedure to estimate a sample quantity involves running batches of pseudoexperiments in which events from the original distribution are used multiple times. It is important that identical procedures are used for the original sample and all bootstrap samples. For example, if 3000 pseudoexperiments were used to determine the mean M_{top} residual in the original sample, then 3000 pseudoexperiments should be used in each bootstrap sample. In this way, fluctuations due to the procedure itself are built into the distribution of bootstrap estimates. Also note that only the pseudodata changes from sample to sample; the machinery of the analysis (templates, likelihood, mapping functions) is always the same.

- The bootstrap is probing an uncertainty due to limited statistics in the Monte Carlo samples. Every event was individually and independently generated by the Monte Carlo, regardless of any weighting (due to cross-sections, efficiencies, etc.) that is imposed when the event is used in an analysis. Therefore, every event should be selected into the bootstrap samples, possibly multiple times, with equal probability ($= 1/N_{MC}$). But then in the analysis context, e.g. selecting events from the (bootstrap) pseudodata to make pseudoexperiments, the event weights should be applied as usual.

- When the pseudodata is stored in and drawn from histograms, resampling with replacement reduces to fluctuating each bin of the histogram according to a Poisson distribution. This is exactly the procedure used to estimate sample quantity uncertainties in some previous top mass analyses; the bootstrap, if you like, is a generalization of this technique for pseudodata stored in ntuples.

- Although not as costly as generating additional Monte Carlo events, the bootstrap is fairly CPU-intensive due to the large number of pseudoexperiments that must be performed. We suggest generating 50 bootstrap samples for $\approx 10\%$ uncertainty on the uncertainty as discussed in Sec. 2.1. Also, the uncertainty on a sample quantity estimated using one 1M-event sample is probably valid for another 1M-event sample with slightly different M_{top}. For the current gen6 MC, we suggest performing the bootstrap on one 1M-event sample and one 4M-event sample, then using the results for any sample with roughly the same statistics.

2.5 Interpretation

Even with an accurate estimate of sample quantity uncertainties from the bootstrap technique, caveats can arise in their interpretation.

When the same background events are used in pseudoexperiments to estimate sample quantities at different values of M_{top}, the statistical uncertainties due to background statistics will be highly correlated. So to interpret consistency or trends in

the sample quantity across values of M_{top}, the signal-only bootstrap uncertainty is more appropriate; to understand the overall uncertainty on e.g. the method's bias, the signal-and-background bootstrap uncertainty is more correct.

Similarly, in 2D M_{top}-JES measurements, samples with the same M_{top} but different JES are usually not statistically independent, since they come from applying different jet energy scale assumptions to the same set of MC events. Thus even the signal-only bootstrap uncertainty is not a correct estimate of the variations along JES, since the statistical fluctuations are highly correlated.

The bootstrap tells us about the uncertainty on our *estimate* of e.g. bias arising from the limited MC statistics in the sample used to make the estimate. But remember that the method itself can be biased in strange ways due to limited statistics in the samples used to *define* the method (through templates, mapping functions, etc.).[8] We are precisely interested in an accurate estimate of the uncertainty arising from the former effect so that we can reliably uncover bias due to the latter effect, which directly affects the final measurement using real data!

3 Examples

Here we present three examples of the application of the bootstrap method. This will serve two purposes. First, we hope that through examples the somewhat complicated procedure will be made more clear. Second, we can provide some validation of the technique.

3.1 Toy examples

The first examples are toys in the sense that we use well understood distributions. The true values of the quantities we are trying to estimate are known.

In the first example, shown in Fig. 1, we use a triangle distribution as given in the top right plot. We will apply the bootstrap to estimate the uncertainty on the mean of a sample of $N = 1000$ from that distribution. The top left plot shows one such $N = 1000$ sample; its mean is 1.655. Each resampling of the top left plot produces another $N = 1000$ distribution, such as the one in the middle left plot. From each of 1000 resampled distributions we take the mean and those means are plotted at the middle right. This is an *estimate* of the sampling distribution derived from the single sample at the top left. From this distribution, we take the RMS of 0.074 as an estimate of the uncertainty on our original estimate of 1.655. Now, since this is a toy experiment, we can generate 1000 samples like the one on the top left from the original distribution. On the bottom left, we plot the means of those 1000 samples (1.655 is one entry in this histogram). For each such sample, we perform the entire bootstrap

[8] Subtleties arise due to the use of the same MC samples to define the method and then to test it. This interaction is worth thinking about and minimizing as much as possible, but we know of no reasonable prescription for avoiding it completely.

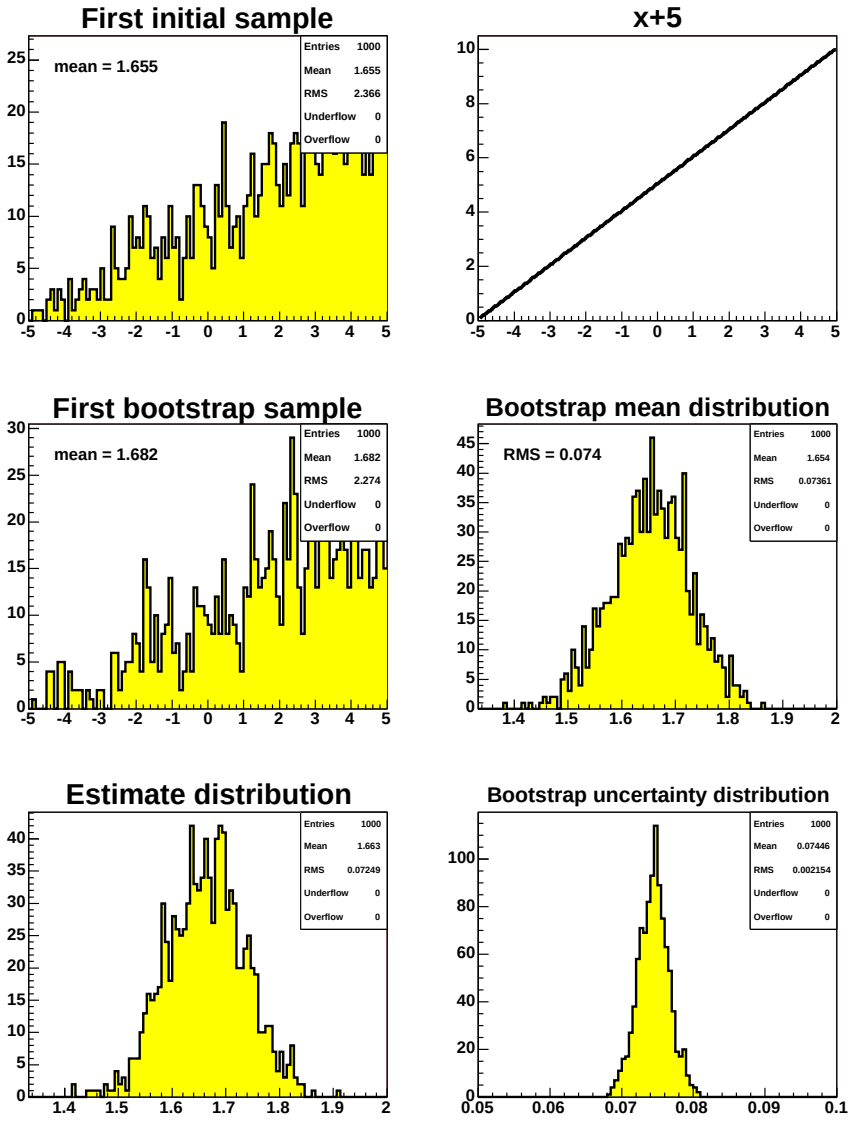

Figure 1: A toy application of the bootstrap technique. We use the bootstrap to estimate the uncertainty on the mean of a triangle distribution with $N = 1000$.

procedure to obtain an uncertainty on the mean, which is plotted at the bottom right. Finally, the pull distribution is formed (Fig. 2) by pairing each estimate in the bottom left plot with its corresponding uncertainty in the bottom right plot, and comparing to the known true answer ($= 5/3$). The well-behaved pull distribution indicates that the uncertainties determined using the bootstrap technique are correct.

The second example is shown using the same battery of plots in Fig. 3. We use a unit Gaussian distribution as given in the top right plot. We will apply the bootstrap to estimate the uncertainty on the RMS of a sample of $N = 1000$ from that distribution.

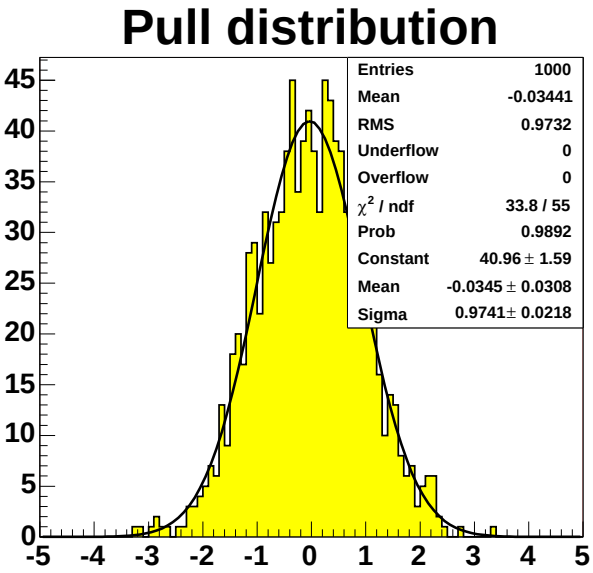

Figure 2: The pull distribution corresponding to 1000 applications of the bootstrap technique to estimate the uncertainty on the mean of a triangle distribution. The unit pull indicates we accurately estimated the uncertainty sample by sample.

The top left plot shows one such $N = 1000$ sample; its RMS is 0.966. Each resampling of the top left plot produces another $N = 1000$ distribution, such as the one in the middle left plot. From each of 1000 resampled distributions we take the RMS and those values are plotted at the middle right. This is an *estimate* of the sampling distribution derived from the single sample at the top left. From this distribution, we take the RMS of 0.023 as an estimate of the uncertainty on our original estimate of 0.966. Now, since this is a toy experiment, we can generate 1000 samples like the one on the top left from the original distribution. On the bottom left, we plot the RMS of each of those 1000 samples (0.966 is one entry in this histogram). For each such sample, we perform the entire bootstrap procedure to obtain an uncertainty on the RMS, which is plotted at the bottom right. Finally, the pull distribution is formed (Fig. 4) by pairing each estimate in the bottom left plot with its corresponding uncertainty in the bottom right plot, and comparing to the known true answer (= 1.0). The well-behaved pull distribution indicates that the uncertainties determined using the bootstrap technique are correct.

3.2 Example from real analysis

Here we present the application of the bootstrap method to an actual analysis, the measurement of the top quark mass using a template method [3].

Results of the bootstrap, implemented as sketched in Sec. 2.4, are shown in Fig. 5. Here we use `ttop73`, and we bootstrap the signal and background samples separately.

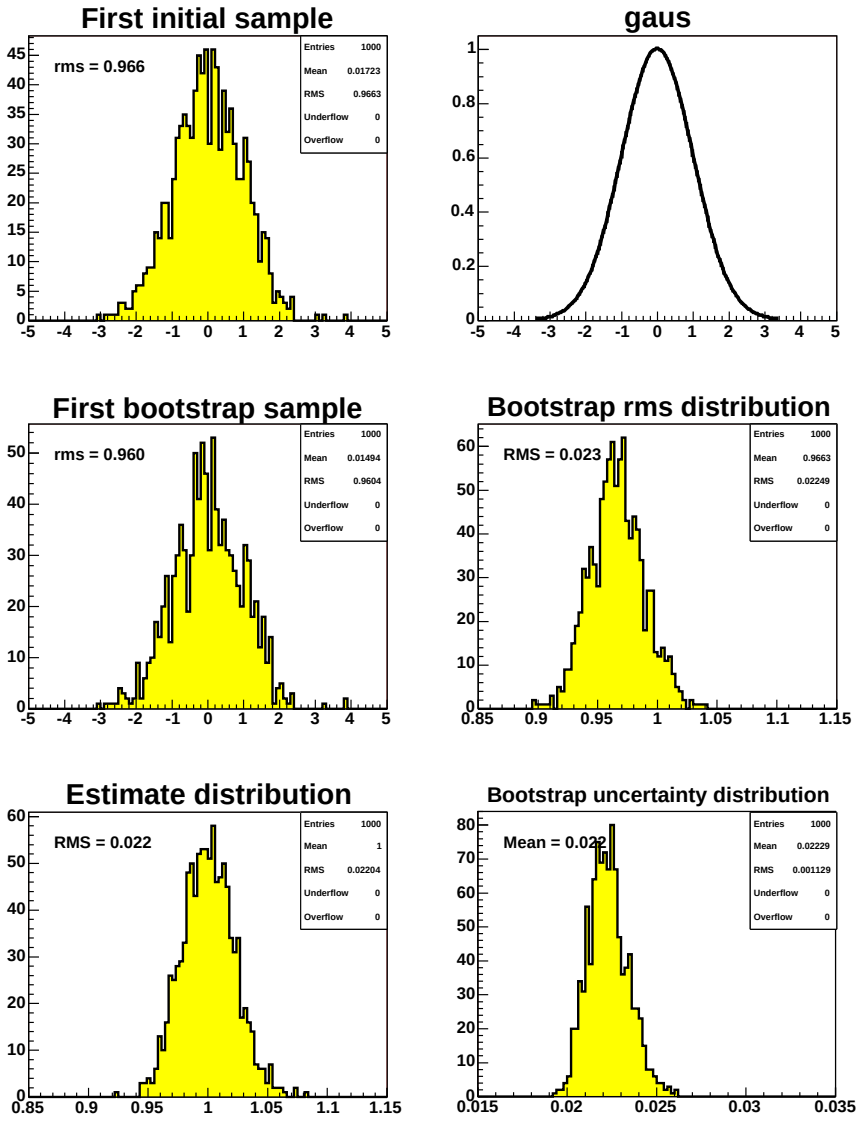

Figure 3: A toy application of the bootstrap technique. We use the bootstrap to estimate the uncertainty on the RMS of a Gaussian distribution with $N = 1000$.

We generate 50 bootstrap samples in each case; the resulting estimates of the sampling distribution for mean M_{top} residual (top plots) and M_{top} pull width (bottom plots) are shown for the signal bootstrap on the left, and the background bootstrap on the right. Since the distributions are reasonably Gaussian, we interpret the RMS of each as the uncertainty on our original estimate: thus for example signal statistics contribute a 0.23 GeV/c^2 uncertainty on our mean mass residual, and a 0.018 uncertainty on our pull width estimate. The uncertainty on these uncertainties is roughly 10%, which could be reduced if needed by analysing more bootstrap samples.

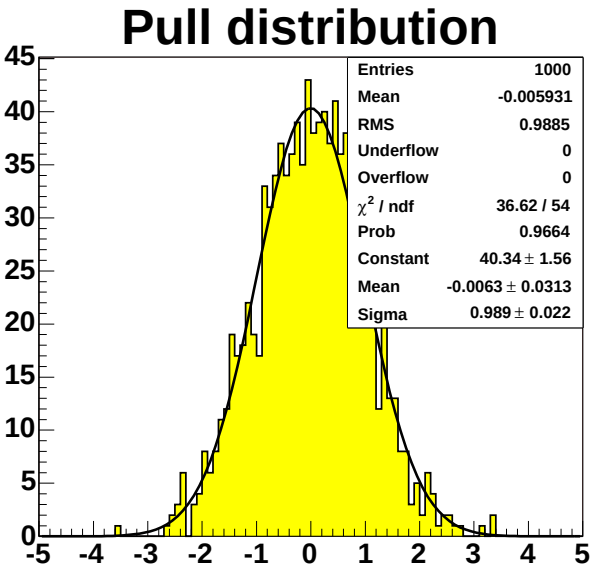

Figure 4: The pull distribution corresponding to 1000 applications of the bootstrap technique to estimate the uncertainty on the RMS of a Gaussian distribution. The unit pull indicates we accurately estimated the uncertainty sample by sample.

In our blessed analysis, these uncertainties were used to realistically assess the performance of our method for a range of M_{top} and JES values, and to determine the precision of our estimates of systematic uncertainties.

4 Conclusion

The bootstrap technique is an appropriate solution to our problem of accurately estimating the uncertainty on sample quantities like mean mass residual, pull width, and so on. We have described the method, shown examples of its application, and made some specific recommendations for its use in top quark mass analyses.

References

[1] Bradley Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.

[2] Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

[3] J. Adelman, E. Brubaker, W. Fedorko, Y.K. Kim, H.S. Lee, M. Shochet, S. Carron, P. Sinervo, Y.J. Lee, and G. Velev. Template-based top quark mass measurement

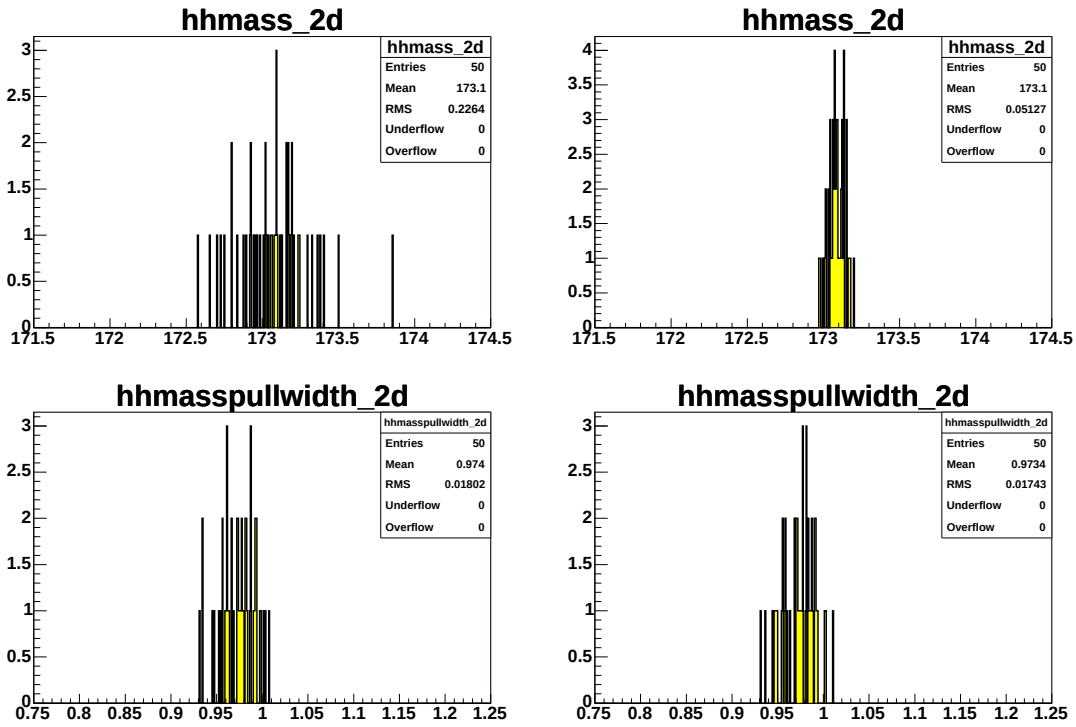

Figure 5: An example of the application of the bootstrap procedure. In the TMT-2D analysis, we separately bootstrap the signal (left) and background (right) samples. The resulting estimate of the sampling distribution is shown for the M_{top} measurement in the top plots; and for the M_{top} pull width in the bottom plots. Since the distributions are fairly Gaussian, we interpret the distributions as indicating for example a 0.23 GeV/c^2 uncertainty on the mass residual measured in this sample, and a 0.018 uncertainty on the estimated pull width.

on 1.7 fb^{-1} of data in the lepton+jets channel using kde. CDF Note 8909, Fermilab, July 2007.