Leipzig University

Hauptseminar

Der Münchhausen-Trick

Pulling oneself up by one's bootstrap

by
Franz Paul Spitzner

Presentation: 21. April 2016

Contents

1 Motivation

Over the past century computer based statistical methods have evolved from fresh, new idea to a nowadays well established scientific tool, not only in physics. They are used regularly to close the gap between analytic predictions and experimental results [6]. Algorithms based on the Monte Carlo method e.g. the well known Metropolis [7], are often easier to carry out than a real life experiment, in which cumbersome external conditions need to be controlled. Analytical methods on the other hand are mostly limited to approximations and may rely on a strict range of assumptions. Computer simulations produce the desired results, are easy to control and financially more efficient than ever. Having praised the approach, there is of cause, a catch: due to the statistical sampling that is employed, arises the apparent need to evaluate the quality of the produced results - or to cite N. B. Wilding [10] *"Their value [...] does, on the other hand, agree well with our result, although since no error bars were quoted it is impossible to tell to what extent the accord is meaningful."*
The treatment of error estimations is not new, many of the (analytical) standard treatments date back to the time between 1800 and 1930 [3]. One of the biggest disadvantages at the time was the requirement of an a priori knowledge about the probability distribution from which was sampled, in order to obtain an error estimate. In 1979 Bradley Efron proposed the Bootstrap method as an easy to use general method to evaluate the statistic precision of an estimator [1]. It is similar to the Jackknife method developed in 1949 by Maurice Quenouille and follows the same idea, called resampling: The single available sample is used to create many, fictitious samples. Afterwards, the error estimates are obtained from this set of multiple samples.
Throughout the following pages the Bootstrap method will be explained and compared to the popular Jackknife. It will become apparent that, although the concept of resampling is easy to comprehend, it relies heavily on repeating tasks a multitude of times. Thereby resampling techniques are put into the computational perspective, under which they were designed.
As a model to work with, the Bivariate Gaussian will be employed, producing time series that show the characteristics desired to highlight the strengths of resampling. Although not presented, the methods shown were later tested on "real" data and are used on regular basis.

2 Preliminaries

2.1 Notation

For Monte Carlo simulations in physics, it has become common practice to treat the creation of data and its analysis separately [5]. The intermediate storage format is a time series of the desired *observable* $\mathcal{O}_i$, such as illustrated in Fig. 1. The *expectation value* is denoted by

$$\langle \mathcal{O}_i \rangle = \mathcal{O} \quad . \tag{2.1}$$

It is a real number, around which our "best guess" fluctuates. If we have taken N measurements, the expectation value can be approximated by their mean value. Hence we choose to denote the *estimator* by

$$\overline{\mathcal{O}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{O}_i \quad . \tag{2.2}$$

In order to have an expression for the magnitude of fluctuations of the estimator we also introduce the *variance*,

$$\sigma^2_{\overline{\mathcal{O}}} = \left\langle \left(\overline{\mathcal{O}} - \langle \overline{\mathcal{O}} \rangle \right)^2 \right\rangle \quad . \tag{2.3}$$

Section 3.3 will be briefly dealing with autocorrelations. Therefore the definition for the *integrated autocorrelation time* is given in accordance to [5]:

$$\tau_{int} = \frac{1}{2} + \sum_{k=1}^{N} A(k) \left(1 - \frac{k}{N} \right) \quad . \tag{2.4}$$

Where $A(k)$ is the *autocorrelation function* and k is the time separation between two measurements, e.g. $\mathcal{O}_i$ and $\mathcal{O}_{i+k}$. It evaluates how similar two measured observables are, depending on their temporal distance from each other. If $A(k)$ is equal to zero, all measurements are independent and uncorrelated. Intuitively, measurements are less likely to be similar to each other for increasing k and the autocorrelation function decays exponentially.

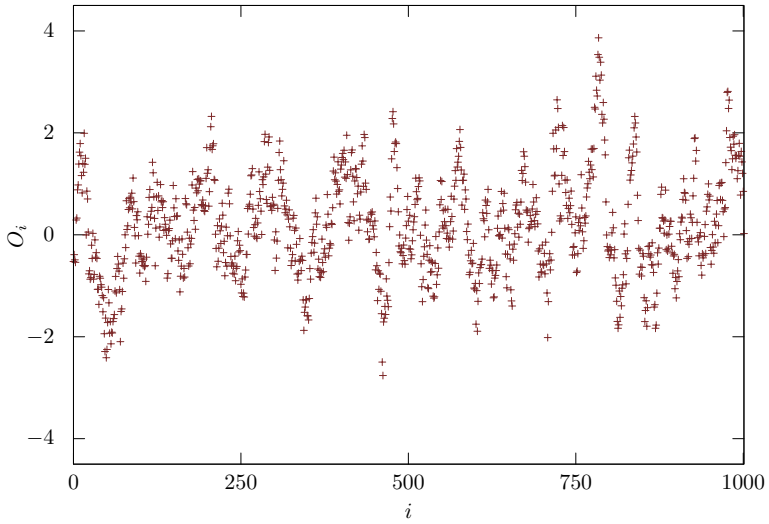

Figure 1 A time series of the Bivariate Gaussian. The observable $\mathcal{O}_i$ is obtained for discrete time steps i.

2.2 Identities

The variance of the estimator (2.3) can be rewritten as

$$\sigma^2_{\overline{\mathcal{O}}} = \left\langle \overline{\mathcal{O}}^2 \right\rangle - \left\langle \overline{\mathcal{O}} \right\rangle^2 \quad . \tag{2.5}$$

In a similar fashion rewriting for the fluctuation of the individual measurements yields

$$\sigma^2_{\mathcal{O}_i} = \left\langle \mathcal{O}_i^2 \right\rangle - \left\langle \mathcal{O}_i \right\rangle^2 \quad . \tag{2.6}$$

As mentioned above, we have no way to calculate an expectation value. This also applies to the variance. Again using the "standard estimator" $\langle * \rangle \to \overline{*}$, we obtain an expression for an *estimator of the variance*:

$$\overline{\sigma}^2_{\mathcal{O}_i} = \overline{\mathcal{O}^2} - \overline{\mathcal{O}}^2 \tag{2.7}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left(\mathcal{O}_i - \overline{\mathcal{O}} \right)^2 \quad . \tag{2.8}$$

It can be shown, that this estimator is weakly biased. To correct this systematic (in the case at hand) underestimation, a factor can be applied, leading to the *bias corrected estimator*

$$\overline{\sigma}^2_{\mathcal{O}_i,corr} = \frac{N}{N-1} \, \overline{\sigma}^2_{\mathcal{O}_i} \quad . \tag{2.9}$$

Finally, we can look at the conventional expression for the *error*:

$$\epsilon^2_{\overline{\mathcal{O}}} = \sigma^2_{\overline{\mathcal{O}}} \quad . \tag{2.10}$$

If all N measurements are *uncorrelated*

$$\epsilon^2_{\overline{\mathcal{O}}} = \frac{\sigma^2_{\mathcal{O}_i}}{N} \quad , \tag{2.11}$$

otherwise we get mixed contributions unequal to zero, when inserting (2.2) into the second term of (2.5). These are consistently expressed in form of the autocorrelation time and we arrive at the error estimate for *correlated* measurements:

$$\epsilon^2_{\overline{\mathcal{O}}} = \frac{\sigma^2_{\mathcal{O}_i}}{N} 2\,\tau_{int} \quad . \tag{2.12}$$

For completeness it has to be mentioned, that these expressions are of cause estimators. Unfortunately, the notation of many publications neglects to clarify that the presented results are in fact estimators. Admittedly, even our own expressions e.g. (2.12) sometimes lack the overbar which would have been required for consistency.

2.3 Bivariate Gaussian Distribution

When presenting time series and their subsequent analysis in this work, it was a main concern to have control over the simulations as a necessity to highlighting key aspects of the process of error estimations. While not ideal in the sense, that it is rather "unphysical", we chose to use the Bivariate Gaussian Distribution, as it offers exactly this type of fine tuned control.
It is created from independent, uncorrelated Gaussian random variables $e^{\text{`}}_i$ with expectation value $\langle e^{\text{`}}_i \rangle = 0$ and $\langle e^{\text{`}}_i\, e^{\text{`}}_j \rangle = \delta_{ij}$. We start with

$$e_0 = e^{\text{`}}_0 \tag{2.13}$$

and apply

$$e_i = \rho\, e_{i-1} + \sqrt{1-\rho^2}\, e^{\text{`}}_i \tag{2.14}$$

recursively, to generate the now correlated e_i. Here, ρ is a correlation coefficient, determined from the autocorrelation time via $\rho = e^{-1/\tau}$.

3 Method

Let us start by going through the Bootstrap method for a simple example, where we have a time series of length $N = 2$. The set of all two measurements - the original time series, shall be denoted by

$$\widetilde{\mathcal{O}} = \{\mathcal{O}_1, \mathcal{O}_2\} \quad .$$

The resampling is realized by creating new sets from *randomly chosen* measurements of the original sample. Individual measurements can be picked multiple times (drawing with putting back) and the created samples shall have the original size $N = 2$. This way we can construct

$$N_R = N^N = 4 \tag{3.1}$$

so called *ideal* Bootsrap replicas, namely

$$\begin{aligned}
\widetilde{\mathcal{O}}_{R,1} &= \{\mathcal{O}_1, \mathcal{O}_1\} \\
\widetilde{\mathcal{O}}_{R,2} &= \{\mathcal{O}_1, \mathcal{O}_2\} \\
\widetilde{\mathcal{O}}_{R,3} &= \{\mathcal{O}_2, \mathcal{O}_1\} \\
\widetilde{\mathcal{O}}_{R,4} &= \{\mathcal{O}_2, \mathcal{O}_2\} \quad .
\end{aligned}$$

We can see from (3.1) that the amount of ideal replicas quickly exceeds even modern computer capabilities, as N is typically of order 10^6 or greater. Hence we create only a predefined number N_R of replicas, following the Monte Carlo approach. Each Bootstrap sample $\widetilde{\mathcal{O}}_{R,r}$ is created from randomly drawing a data point of the original pool with probability $p = \frac{1}{N}$.

3.1 Algorithm

There are three steps to the Bootstrap method:

1. Construct N_R Bootstrap replicas $\widetilde{\mathcal{O}}_{R,r}$ as outlined above;

2. Evaluate each replica using a chosen estimator, e.g. the mean;

$$\overline{\mathcal{O}}_{R,r} = \frac{1}{N} \sum_{i=1}^{N} (\widetilde{\mathcal{O}}_{R,r})_i \tag{3.2}$$

3. Obtain the *Bootstrap estimator* $\overline{\mathcal{O}}_R$ from all the individual estimators $\overline{\mathcal{O}}_{R,r}$

$$\overline{\mathcal{O}}_R = \frac{1}{N_R} \sum_{r=1}^{N_R} \overline{\mathcal{O}}_{R,r} \tag{3.3}$$

and an according expression for the standard error

$$\sigma^2_{\overline{\mathcal{O}}_R} = \frac{1}{N_R - 1} \sum_{r=1}^{N_R} \left(\overline{\mathcal{O}}_{R,r} - \overline{\mathcal{O}}_R \right)^2 \quad , \tag{3.4}$$

where $\sigma^2_{\overline{\mathcal{O}}_R}$ is of cause an estimator;

From (3.4) it becomes clear that the Bootstrap estimator for the error is genuinely calculated from the fluctuations of all the replicas around their average. This seemingly trivial convenience actually encapsulates a great benefit of the Bootstrap: Since all the replicas follow the same empirical distribution as the original sample, no knowledge about said distribution is required.

3.2 Choosing the Number of Replicas

Efron and Tibishirani [4] give two rules of thumb for choosing N_R, which are justified by a thorough discussion. Firstly, already small $N_R \approx 50$ can be sufficient for obtaining a "good" Boostrap estimator $\overline{\mathcal{O}}_R$. Secondly, increasing N_R beyond 200 replicas is seldom needed in order to gain reliable estimates for a standard error.

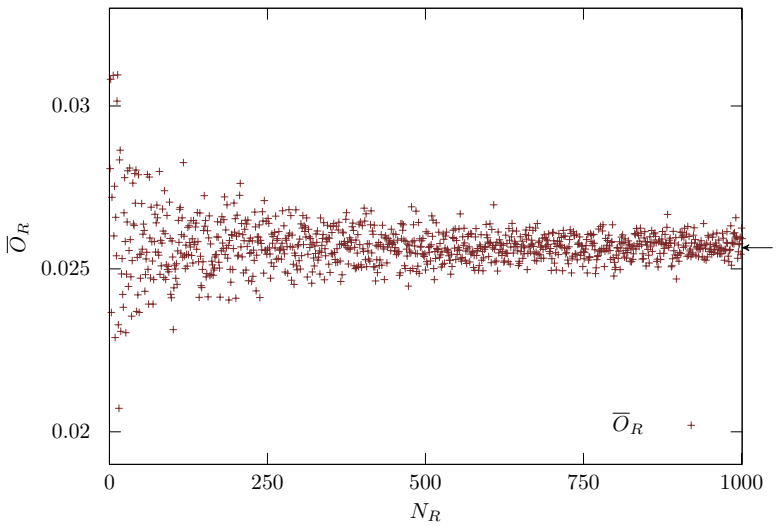

Figure 2 Bootsrap estimators calculated from increasing numbers of replicas for a time series with $N = 10^6$ measurements (Bivariate Gaussian distribution). The estimator $\overline{\mathcal{O}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{O}_i \approx 0.0256$, obtained from the original sample is indicated on the right.

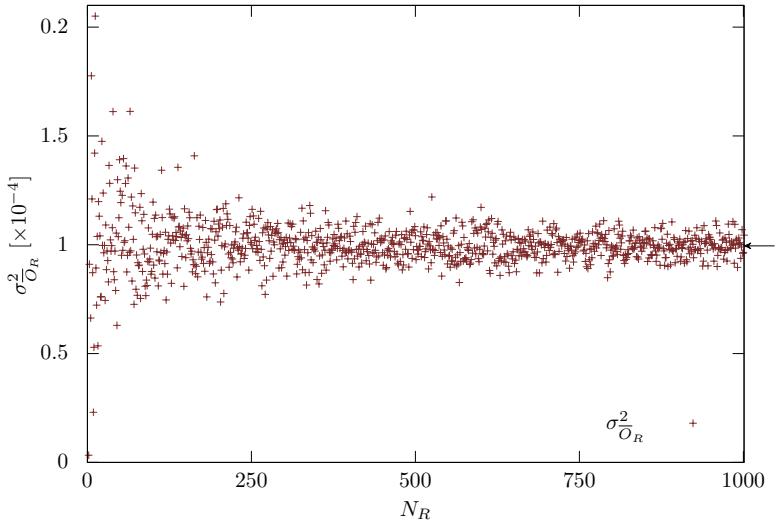

Figure 3 Variance of the Bootstrap replicas for increasing numbers of replicas for a time series with $N = 10^6$ measurements (Bivariate Gaussian distribution). The variance of the original sample $\sigma^2_{\overline{O}} \approx 0.995 \times 10^{-4}$ is again indicated on the right.

At first sight one might guess the more replicates created, the better. While Figures 2 and 3 seem to justify this assumption, one can also see that there must exist a limit on how precise Bootstrap estimates can possibly be. Since all the replicas are created from one original sample, they must have the same statistical flaws. If the data at hand - unluckily - only samples a small part of the whole distribution, so will the replicas. This outlines why the Bootstrap method is no panacea, but like other resampling techniques, a convenient way to check the *self consistency* of available data.

3.3 Autocorrelation and Binning

When dealing with Monte Carlo simulations to produce time series, it is always a difficult task to decide how long a simulation needs to run. More so in the vicinity of a phase transition, where the correlation length of the system diverges, as the transition is approached. If presented with raw data one cannot be sure if the simulation covered all phase space or just a small subsection. Running any resampling technique on such a subset would yield good errors and seemingly consistent estimates for observable - even though they might be far away from the truth. Unfortunately there is no way other than running a longer simulation, to judge if this was the case. This "sufficiently long" simulation needs to be much greater than the autocorrelation

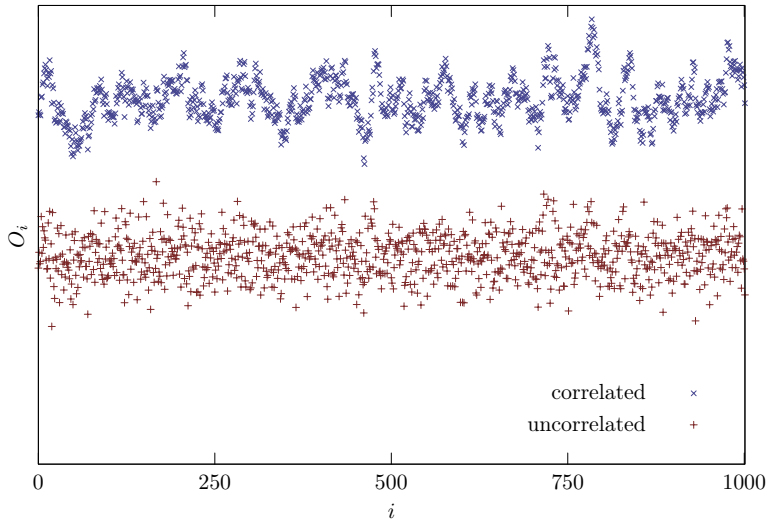

Figure 4 Qualitative plot of two time series of the Bivariate Gaussian for which the autocorrelation time can be set as a parameter. The correlated sample ($\tau = 10$) was shifted to improve readability.

time, which is beforehand unknown.
So, how can one judge if the errors calculated by resampling account for an appropriate value of the autocorrelation time? See again (2.12) and (2.4), as well as Figure 4 for a qualitative comparison between correlated and uncorrelated data. As far as the author is concerned, the most popular ansatz to examine autocorrelations is the Binning method. It is often assumed to be used in conjunction with Jackknifing but can also be combined with the Bootstrap.
Instead of resampling individual data points of the original sample with length N, said time series is grouped into N_B equally long *bins*, each containing k individual measurements:

$$k = \frac{N}{N_B} \tag{3.5}$$

Without discussing the issue in further detail, we want to mention that this ostensibly simple step is more subtle, since there is no universal way to deal with non-integer values for k. As depicted in Figure 5, each of the created bins or blocks is evaluated separately, using the chosen estimator. Ordinarily the *block average*

$$\overline{\mathcal{O}}_{B,n} = \frac{1}{k} \sum_{i=1}^{k} \mathcal{O}_{k(n-1)+i} \tag{3.6}$$

is used, where $n = 1, \ldots, N_B$. Next, we acquire an average over all the bins

$$\overline{\mathcal{O}}_B = \frac{1}{N_B} \sum_{n=1}^{N_B} \overline{\mathcal{O}}_{B,n} \quad . \tag{3.7}$$

We can see that

$$\overline{\mathcal{O}}_B = \overline{\mathcal{O}} \tag{3.8}$$

- under the condition, that the chosen estimators were in fact the arithmetic mean, and k was integer. Furthermore one can apply the bias corrected estimator for the variance (2.9) to obtain an error expression from the binning

$$\sigma^2_{\overline{\mathcal{O}}} = \sigma^2_{\overline{\mathcal{O}}_{B,n}} \,/\, N_B \tag{3.9}$$

$$\sigma^2_{\overline{\mathcal{O}}} = \frac{1}{N_B(N_B-1)} \sum_{n=1}^{N_B} \left(\overline{\mathcal{O}}_{B,n} - \overline{\mathcal{O}}_B\right)^2 \quad . \tag{3.10}$$

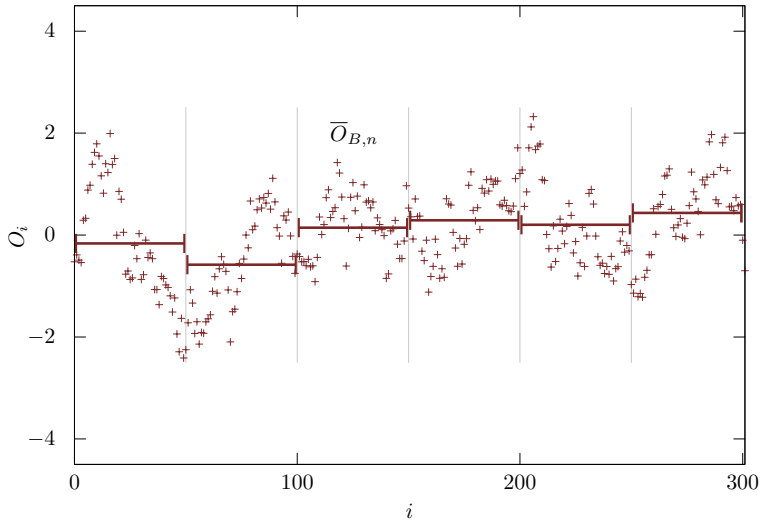

Figure 5 The time series is split into N_B blocks of length k, each one with a distinguished estimate $\overline{\mathcal{O}}_{B,n}$.

It is important to note, that (3.9) assumes, that the created bins are larger than the autocorrelation time, hence all blocks are uncorrelated, which is the only justification why the factor $2\,\tau_{int}$ does not appear. The same assumption will be made in the following section, which will show how the Binning is merged with the Jackknife and Bootstrap in comparison.

4 Binned Bootstrap and Jackknife

4.1 Comparison between both Methods

As the derived expressions are very similar we will present them at once with a consequential discussion. The starting point shall be, that the block averages $\overline{\mathcal{O}}_{B,n}$ were created successfully as in (3.6) with $n = 1, \ldots, N_B$.

For the Jackknife:

$$\overline{\mathcal{O}}_{J,n} = \frac{1}{N_B - 1} \sum_{i \neq n}^{N_B} \overline{\mathcal{O}}_{B,i} \tag{4.1}$$

$$\overline{\mathcal{O}}_{J} = \frac{1}{N_B} \sum_{n=1}^{N_B} \overline{\mathcal{O}}_{J,n} \tag{4.2}$$

$$\sigma^2_{\overline{\mathcal{O}}_J} = \frac{N_B - 1}{N_B} \sum_{n=1}^{N_B} \left(\overline{\mathcal{O}}_{J,n} - \overline{\mathcal{O}}_{J}\right)^2 \tag{4.3}$$

For the Bootstrap:

$$\overline{\mathcal{O}}_{R,r} = \frac{1}{N_B} \sum_{i=1}^{N_B} \overline{\mathcal{O}}_{B,n} \tag{4.4}$$

$$\overline{\mathcal{O}}_{R} = \frac{1}{N_R} \sum_{r=1}^{N_R} \overline{\mathcal{O}}_{R,r} \tag{4.5}$$

$$\sigma^2_{\overline{\mathcal{O}}_R} = \frac{1}{N_R - 1} \sum_{r=1}^{N_R} \left(\overline{\mathcal{O}}_{R,r} - \overline{\mathcal{O}}_{R}\right)^2 \tag{4.6}$$

From (4.1) and (4.4) we see that both, Jackknife and Bootstrap combine multiple bins into N_B larger samples, namely *Jackknife bins* and *Bootstrap replicas*, respectively. The $\overline{\mathcal{O}}_{J,n}$ are of size $N_B - 1$, as can be seen from the normalization and the $\overline{\mathcal{O}}_{R,r}$ are of size N_B. Also, while the former methodically contain every bin except one, the latter are constructed from *picking bins randomly*, as already described in Section 3. To emphasize this a bit more, it is pointed out, that a different amount of those larger blocks is

created in the two methods. For the Jackknife, there always exist N_B large Jackknife bins, which is directly related to the block length k. The Bootstrap on the other hand employs a second parameter N_R, the number of replicas. Consistently, the only difference in calculating the *average over all replicas* $\overline{\mathcal{O}}_J$ and $\overline{\mathcal{O}}_R$ is the normalization. These two are again the estimators approximating the expectation value of the observable we aim to find out from the original sample. Again with reference to (2.9), one can also obtain errors for the estimates. In accordance with expectations, the binned Bootstrap variance (4.6) is almost identical to the non binned expression (3.4), only the $\overline{\mathcal{O}}_{R,r}$ are now created from the bins instead the individual measurements. The Jackknife variance (4.3) however has to include the so called "inflation factor" $(N_B - 1)^2$. As mentioned earlier, the Jackknife bins $\overline{\mathcal{O}}_{J,n}$ are created in a systematic way contrasting to the random Bootstrap approach. Hence all these newly created samples are by construction very similar to each other. Taking also into account that the new sample size is exactly one block smaller than the original time series, one can deduce that the Jackknife bins follow a different - narrower - distribution. The inflation factor simply takes care of this when obtaining the variance, which is of cause the width of said distribution. Section 4.3 provides a made up example to illustrate this a bit further.

4.2 Choosing the Number of Bins

A significant question that remains to be discussed, is the choice of how many bins the initial time series is split into. Using the same example of the Bivariate Gaussian as before, Figure 6 shows how the respective error estimates from Jackknife and Bootstrap behave for increasing block sizes k. Firstly it is noteworthy that $k = 1$, equivalent to the unblocked treatment, shows matching results to Figure 3. Secondly, no qualitative difference can be seen between Jackknife and Bootstrap, assuming the number of replicas N_R was chosen large enough.
It is very characteristic that too small choices of k systematically underestimate the error. It is clear, that small blocks cannot be larger than the autocorrelation time - but which was assumed during Binning. On the other hand, increasing block sizes imply a decreasing amount of blocks N_B, hence less summands that contribute to the Jackknife bins $\overline{\mathcal{O}}_{J,n}$ and Bootstrap replicas $\overline{\mathcal{O}}_{R,r}$. The consequence are the strong fluctuations as seen on the right of the figure, which keep increasing until the scale of the actual error contributions become irrelevant.
So we can conclude that the "ideal" choice for the block size is neither 1 nor the other extreme $N - 1$, but rather a sensible intermediate. The grey

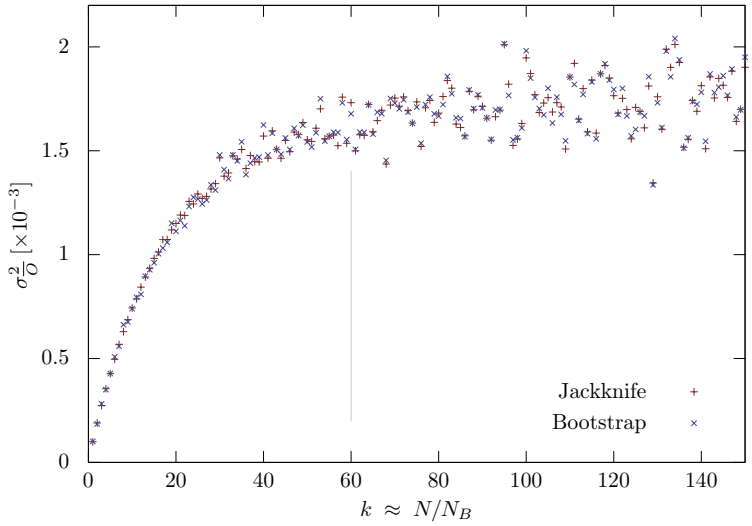

Figure 6 Behaviour of the estimates for variance from Jackknife and Bootstrap method for increasing block sizes k. The autocorrelation time was known beforehand to be $\tau = 10$, again utilizing the Bivariate Gaussian.

guide line in Figure 6 indicates $k \approx 6\,\tau$, which is only known explicitly for this example, since $\tau = 10$ was possible to be set as a parameter. For one, this choice of k is justified through another ansatz that tries to calculate the autocorrelation time directly. Usually the very long summation (2.4) is*self consistently* aborted at this point. Heuristically we see, that a steady plateau is reached here before more severe fluctuations start.
One can also estimate that the plotted value for the variance near $k \approx 60$ is approximately twenty times the value for no Binning i.e. $k = 1$, corresponding nicely to the factor $2\,\tau$ that distinguishes between correlated and uncorrelated error estimates.
For a more realistic example it is of cause more difficult to find an appropriate choice of k and N_B. While it is informative to do error analysis for different block sizes as shown above it also very expensive or labour intense, and can seldom be repeated for every simulation in a day to day workflow. A suggested compromise is to evaluate the made choice for a few limited examples of desired simulations and then implementing those consistently.

4.3 An Example

Let us have a closer look at the error expressions introduced in Section 4.1 and the replicas created via resampling, by means of the following example. We generate a time series of length $N = 10^6$ and in an unusual manner,

repeat this "experiment" $M = 1000$ times, everyone with different random numbers. This will give as much more precise estimate of the expectation value, than ordinary available. Hence we *assume to know* $\langle \mathcal{O} \rangle$ from the many simulations $m = 1, \ldots, M$:

$$\langle \mathcal{O} \rangle = \frac{1}{M} \sum_{m=1}^{M} \overline{\mathcal{O}}_m \quad . \tag{4.7}$$

Next, we focus on only *one single* of the thousand simulations and perform the Jackknife analysis utilizing $N_B = M$ bins. Hence we obtain the Jackknife bins $\overline{\mathcal{O}}_{J,m}$ and another estimate for $\langle \mathcal{O} \rangle$, namely $\overline{\mathcal{O}}_J$:

$$\overline{\mathcal{O}}_J = \frac{1}{M} \sum_{m=1}^{M} \overline{\mathcal{O}}_{J,m} \tag{4.8}$$

Analogously the Bootstrap analysis is performed, also with $N_B = M$ but also $N_R = M$ to have the same normalization to deal with:

$$\overline{\mathcal{O}}_R = \frac{1}{M} \sum_{m=1}^{M} \overline{\mathcal{O}}_{R,m} \quad , \tag{4.9}$$

where $\overline{\mathcal{O}}_{R,m}$ are clearly the Bootstrap replicas and $\overline{\mathcal{O}}_R$ another estimator for the expectation value.
Without calculating the variances, Figure 7 depicts the explicit distribution counted for the repetition (4.7) and the Bootstrap (4.9) as a smooth histogram. To be more precise, shown are the distributions of the difference between all the replicas and the best guess available for each method, respectively.
The aim of this example is to motivate the following simplifications:

$$\sigma^2_{\overline{\mathcal{O}}_R} \;\approx\; \sigma^2_{\overline{\mathcal{O}}_m} \;=\; \epsilon^2_{\overline{\mathcal{O}}} \quad , \tag{4.10}$$

in words, we assume that the variance from the Bootstrap is a good approximation of the variance we would get from a much larger simulation. This also justifies:

$$\overline{\mathcal{O}}_R \;\approx\; \langle \mathcal{O} \rangle \quad . \tag{4.11}$$

When now repeating (4.6),

$$\sigma^2_{\overline{\mathcal{O}}_R} = \frac{1}{N_R - 1} \sum_{r=1}^{N_R} \left(\overline{\mathcal{O}}_{R,r} - \overline{\mathcal{O}}_R \right)^2 \tag{4.12}$$

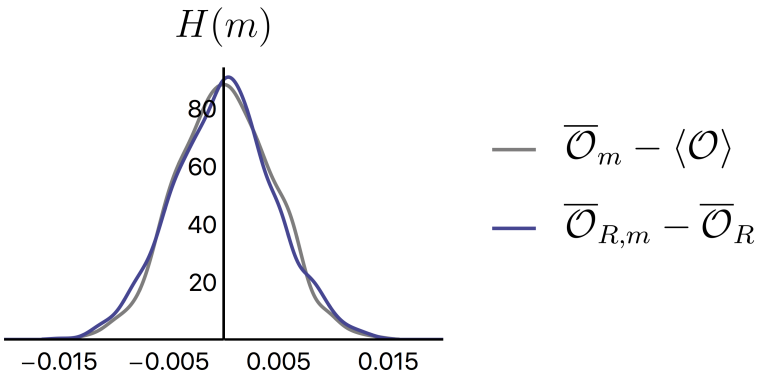

Figure 7 Histogram of the *fluctuations*: Replica estimators around the "best" estimator, for the $M = 1000$ repetitions of the simulation (grey) and the Bootstrap (blue).

we may say, that after some rearranging

$$\overline{\mathcal{O}}_{R,r} - \langle\mathcal{O}\rangle \sim \epsilon_{\overline{\mathcal{O}}} \quad . \tag{4.13}$$

Which is, as intended, the relation illustrated in Figure 7. The Bootstrap replicas have the same fluctuation and variance around the Bootstrap estimate, as a repetition of the simulation would have.
If one now is to repeat the reasoning from equation (4.10) onwards for the Jackknife, we would get:

$$\sigma^2_{\overline{\mathcal{O}}_J} \approx \sigma^2_{\overline{\mathcal{O}}_m} = \epsilon^2_{\overline{\mathcal{O}}} \tag{4.14}$$

$$\overline{\mathcal{O}}_J \approx \langle\mathcal{O}\rangle \tag{4.15}$$

$$\sigma^2_{\overline{\mathcal{O}}_J} = \frac{N_B - 1}{N_B} \sum_{r=n}^{N_B} \left(\overline{\mathcal{O}}_{J,n} - \overline{\mathcal{O}}_J\right)^2 \tag{4.16}$$

$$\overline{\mathcal{O}}_{J,n} - \langle\mathcal{O}\rangle \sim \frac{\epsilon_{\overline{\mathcal{O}}}}{\sqrt{N_B}} \quad . \tag{4.17}$$

This is illustrated in Figures 8 and 9. One can now see what has been said before, that is, that Bootstrap samples the same distribution as a longer simulation would provide. Jackknife on the other hand provides - by construction - a narrower distribution that is compensated by the inflation factor.

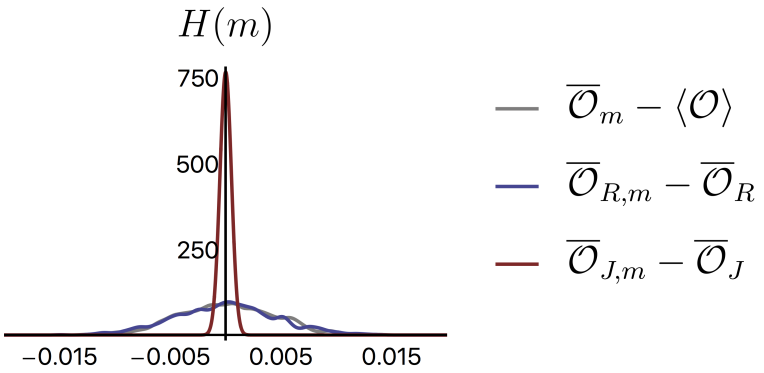

Figure 8 Histogram of the *fluctuations* of replicas: Repetition (grey), Bootstrap (blue), Jackknife (red).

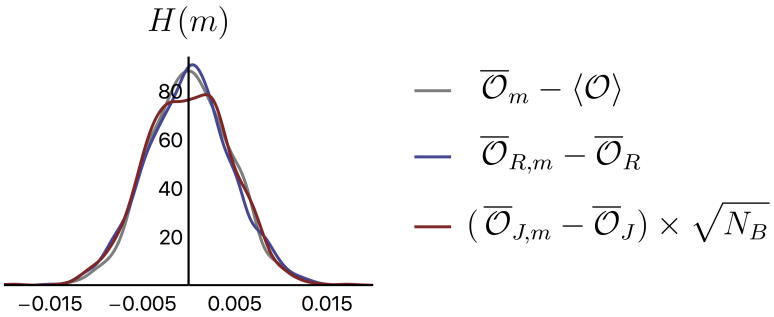

Figure 9 Histogram of the *fluctuations* of replicas: Repetition (grey), Bootstrap (blue), Jackknife with correction (red).

5 Concluding Remarks

While the exercise of the previous section only highlights an admittedly small advantage of the Bootstrap over the Jackknife, it serves to emphasize the simple, yet effective idea behind resampling schemes. Instead of rerunning a simulation thousands of times, which would be utterly expensive, the resampled series often provide an equally effective and cheaper way to obtain reliable errors.

Having said that, the main assumption highlighted multiple times, still holds: We need to make sure, that the bins - and therefore the initial simulation - are larger than the autocorrelation time. Also, while seemingly favoured throughout this work, the Bootstrap method still requires careful attention to tune the two parameters N_B and N_R, while Binning and Jackknife suffice with the former.

Finally, we want to give a few remarks about the literature used. The notation and procedure presented was inspired by - and largely based upon

- the very detailed lecture on Data Correlations and Error Estimates by W. Janke [5]. For an in-depth treatment of the Bootstrap method including bias and confidence intervals, see the book by B. Efron and R. J. Tibshirani [4]. A thorough comparison between Jackknife and Bootstrap is provided online by K.Rummukainen [9].

References

[1] B. Efron. Computers and the theory of statistics: thinking the unthinkable. *SIAM*, 21, 1979.

[2] B. Efron. The Jackknife, the Bootstrap and Other Resampling Plans. *SIAM*, 38, 1982.

[3] B. Efron and P. Diaconis. Statistik per Computer: der Münchhausen-Trick. *Spektrum der Wissenschaft*, 1983.

[4] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

[5] W. Janke. Statistical Analysis of Simulations: Data Correlations and Error Estimation. In J. Grotendorst, D. Marx, and A. Muramatsu, editors, *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*, volume 10 of *NIC Series*, pages 423–445. John von Neumann Institute for Computing, Jülich, 2002.

[6] W. Janke. Monte Carlo Simulations in Statistical Physics - From Basic Principles to Advanced Applications. In Y. Holovatch, editor, *Order, Disorder and Criticality: Advanced Problems of Phase Transition Theory*, volume 3, pages 93–166. World Scientific, 2012.

[7] N. Metropolis et al. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21:1087–1092, 1953.

[8] K. Reimer. Bootstrapping und andere Resampling-Methoden. In S. Albers et al., editors, *Methodik der empirischen Forschung*. Springer Fachmedien, 2009.

[9] K. Rummukainen. Lecture notes on Monte Carlo simulation methods, 2011.

[10] N. B. Wilding. Critical-point and coexistence-curve properties of the lennard-jones fluid: A finite-size scaling study. *Phys. Rev. E*, 52:602–611, Jul 1995.