PROCEEDINGS OF SCIENCE

Bootstrap-determined p-values in lattice QCD

Rajiv Eranki[a,*]

[a] *Department of Physics, Columbia University, New York, NY, USA*

E-mail: re2433@columbia.edu

We present a method based on the bootstrap to determine p-values from Monte Carlo data, in particular those generated in a lattice QCD calculation, where we make no assumptions about the underlying distribution. By generating samples from the underlying data, we are able to naturally incorporate the effects of autocorrelations and non-normally-distributed samples, both of which skew the distribution away from the conventional χ^2 or t^2 distributions. Additionally, with these methods we can estimate the p-values for uncorrelated fits and more elaborate fitting procedures for which these analytical distributions are also unsuitable. (This talk summarizes work published in 10.1103/PhysRevD.111.074514) [2]

The 42nd International Symposium on Lattice Field Theory (LATTICE2025)
2-8 November 2025
Tata Institute of Fundamental Research, Mumbai, India

*Speaker

https://pos.sissa.it/

PoS(LATTICE2025)045

1. Introduction and Motivation

Lattice quantum chromodynamics (QCD) utilizes Monte Carlo evaluation of the QCD path integral in order to statistically sample correlation functions of the theory. We may then fit to parameters of a theoretical ansatz describing this correlation function, where some parameters correspond to physical quantities of interest, making it possible to extract these quantities. In fitting to this ansatz, we would like to achieve two main goals: 1) we'd like the fitted function to interpolate the average correlator values well, taking into account statistical uncertainties; 2) we'd like to evaluate how good our fit was.

Using the conventional fitting approach, described in more detail below, one minimizes a standard goodness-of-fit metric, often called χ^2. In our work we call the metric q^2 to include the more general case in which we may want to modify the form of the metric (say, to massage the covariance matrix).

If the underlying data are statistically independent samples and normally distributed, it is possible to convert this χ^2 score into a p-value, knowing only the number of parameters in the ansatz, the number of correlator timeslices, and the number of data points. The p-value tells us the likelihood of the data having been generated by the ansatz, assuming that the ansatz was correct (the *null hypothesis*). In this frequentist approach to hypothesis testing, a larger p-value, indicating that data that were quite plausibly generated from the ansatz, implies a good fit.

The issue with using this approach in lattice QCD is that the core assumptions of the independence and normality of the underlying data are violated. Furthermore, the covariance matrix, whose inverse is used in the conventional q^2 metric, may be very close to singular and we may want to modify it (for instance, by, removing the off-diagonal components in an uncorrelated fit), invaliding the conventional mapping to p-values.

We propose a method based on the bootstrap [1] to overcome these limitations. By repeatedly resampling the ensemble of data we have obtained from a lattice simulation, and obtaining a histogram of goodness-of-fit scores, we can infer a p-value from where our fit score lies in this distribution. In the toy models we have tested this on, for which we know the true distribution of these scores, the bootstrap shows remarkable agreement with the true distribution. This holds even in the presence of autocorrelations and modifications of the covariance matrix, for which no analytical distribution is available.

2. Conventional Fitting Approach

The usual definition of the goodness-of-fit uses the sample mean and covariance matrix estimated from the data. In the following $C_i(t)$ refers to a correlator at timeslice $t = 1, \ldots, T$ at the measurement index $i = 1, \ldots, N$. The fit function is denoted by f, and the parameters to be fit are

PoS(LATTICE2025)045

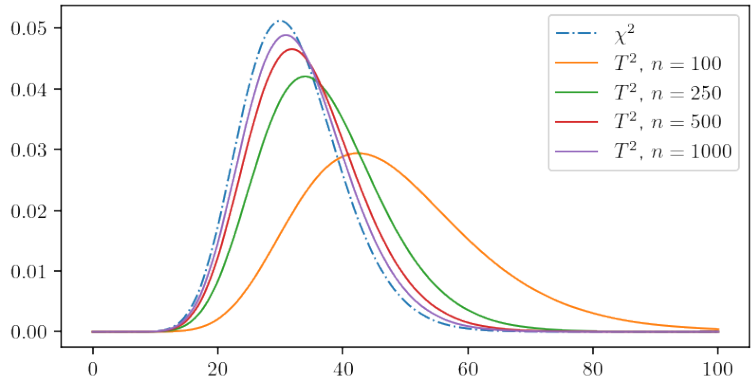


Figure 1: χ^2 vs. T^2 distributions for $K = 32$ degrees of freedom

in vector $\vec{a}$:

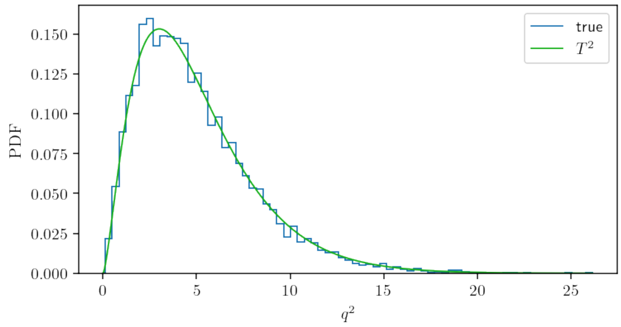
$$\begin{aligned}
\overline{C}(t) &\equiv \frac{1}{N}\sum_{i=1}^{N} C_i(t) \\
\Sigma(t,t') &\equiv \frac{1}{N-1}\left\langle \left(C(t)-\overline{C}(t)\right)\left(C(t')-\overline{C}(t')\right)\right\rangle \\
q^2 &= \sum_{t,t'} \left[\overline{C}(t) - f(t;\vec{a})\right] \Sigma^{-1}(t,t') \left[\overline{C}(t') - f(t';\vec{a})\right].
\end{aligned}$$

The fitting procedure is to tune the parameters $\vec{a}$ such that q^2 is minimized. In the case where we expect the data is independent and normally distributed, statistical theory provides an analytic form for the distribution of q^2 scores under the null hypothesis. In the simplest form, depending only on the number degrees of freedom $K = T - \text{length}(\vec{a})$, where T is the total number of correlator time separations measured, we expect that the q^2 scores follow the χ^2 distribution. If in addition we supply the number of data points N, we obtain the T^2 distribution (see [9]), which is exact in the case where the data follow the assumptions. Both distributions are shown in Figure (1), where we can also see that as $n \to \infty$, the T^2 distribution tends toward the χ^2 distribution. In order to produce a p-value, then, we integrate the tail of this distribution past the score that we obtained from our fit. This tells us, if the data were, in fact, generated by a process reflecting the ansatz but with statistical errors that are normally distributed and independent, that we would expect this score or worse p percent of the time.

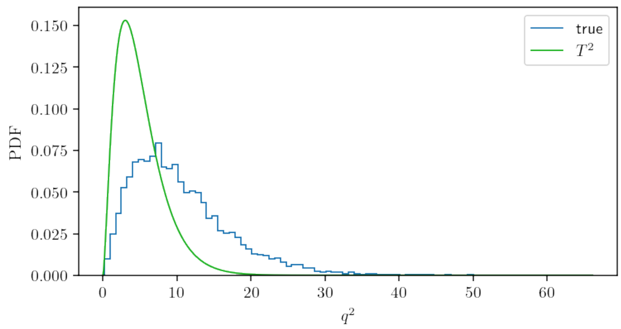
3. Toy Model

As mentioned in the introduction, the assumptions underlying the analytical distributions above are invalid for Monte Carlo lattice data. We will refer only to the T^2 distribution for comparison now, being the more accurate distribution. Our approach to generate data is described in detail in [2], and uses simple stochastic processes to generate "correlator" data. For the data described in this document, $T = 6$ (the number of timeslices) and $N = 800$ (the number of samples). We generate many ensembles of data, and for each of these, we perform a fit of the ansatz by minimizing q^2. The resulting histogram of q^2 scores is plotted in comparison with T^2.

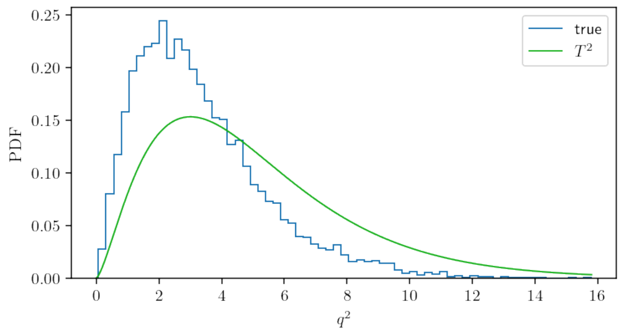
If we generate data with Gaussian errors (see [2] for the specific parameters) distributed around the ansatz function, we get a histogram of q^2 scores shown in Figure 2a, in good agreement with the T^2 distribution. However, if there are even small autocorrelations present in the data, we can see in Figure 2b that the distribution of q^2 scores is skewed. Similarly, if we perform an uncorrelated fit, we see in Figure 2c that the distribution is also skewed. In both of these last cases, reporting a p-value obtained from T^2 will be inaccurate.



(a) T^2 vs. data obtained from multivariate Gaussian



(b) T^2 vs. data with autocorrelations $\tau_{\text{int}} \approx 1.2$



(c) T^2 vs. uncorrelated fit

Figure 2: q^2 scores: T^2 vs. true distribution from toy model data

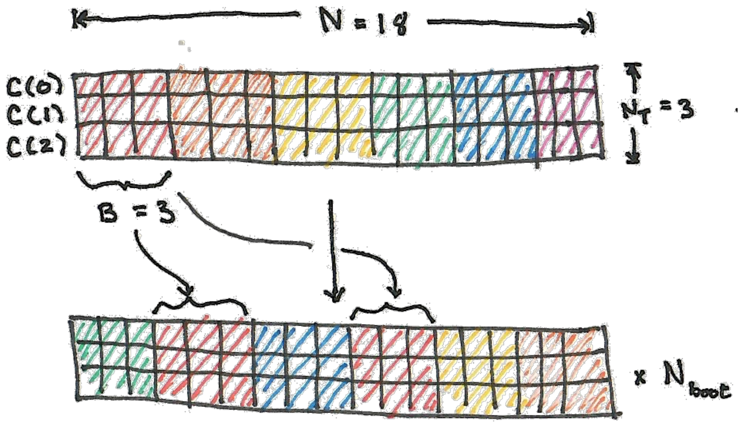


Figure 3: Bootstrap resampling

4. Bootstrap Approach

The main idea behind the bootstrap approach is to use lattice data to generate many synthetic ensembles, from which we will attempt to recreate the null distribution of q^2. We refer the reader to [2] for a detailed discussion of the rationale and algorithm, but to summarize at a high level:

1. Perform an initial fit of the data and obtain a q^2 score.

2. Draw a contiguous block of $B \approx \tau_{\text{int}}$ samples from the correlator data, across all timeslices, with replacement. That is, into a new ensemble, copy $C_i(t)$ for $i = X \ldots X + B$ and $t = 1 \ldots T$ where X is chosen randomly. Copying in blocks ensures that the autocorrelation structure is replicated in the synthetic ensemble, $C_i^b(t)$. This is depicted visually in Figure 3. Repeat N/B times to create a synthetic ensemble.

3. For each timeslice, shift the data so that the average exactly matches the ansatz function with the parameters determined by the initial data (Figure 4). This ensures that these new ensembles reflect data points generated from the ansatz function, but also reflect the underlying variance and autocorrelation structure of the underlying data:

$$C_i^b(t) \to C_i^b(t) - [\overline{C}(t) - f(t; \vec{a})]$$

4. For this new ensemble, perform a fit and record the resulting q^2 score.

5. Repeat steps [2-4] N_{boot} times, obtaining a histogram of q^2 scores.

6. Compare the original q^2 score to this histogram in order to report a p-value.

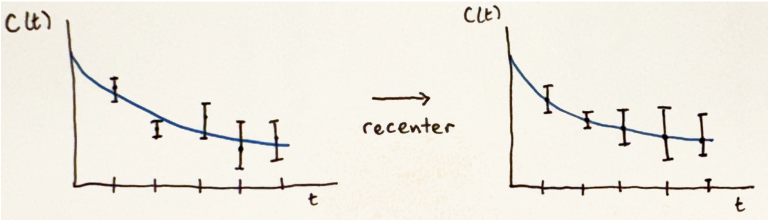


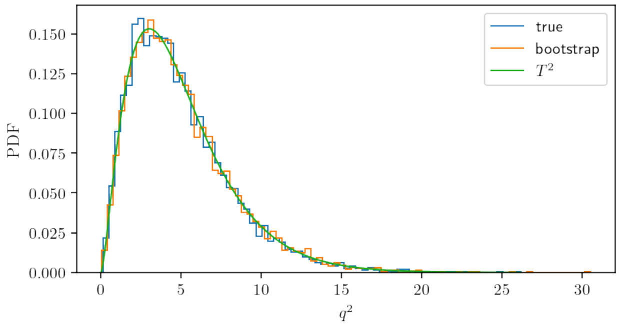
Figure 4: Recentering

5. Results

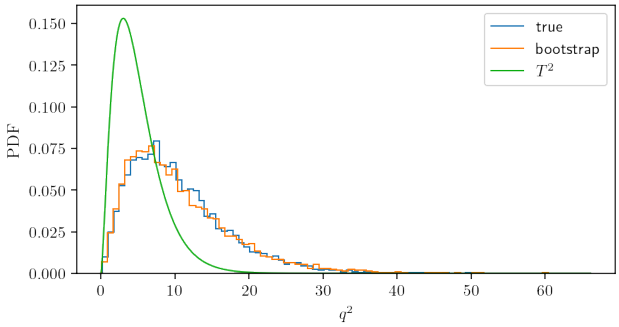
We refer the reader to [2] for more analysis and discussion, but here we highlight the results of this procedure applied to the simulated data described above. In Figure 5a, we can see that from a single initial ensemble of data we can recover the T^2 distribution with high fidelity, a reassuring result. In Figure 5b, we see that the resampled ensembles generate a histogram (orange) that is close to what the "true" distribution is obtained from many simulations of a Metropolis process that produces autocorrelations. And finally, in Figure 5c, we can see that the procedure also reproduces the true distribution of q^2 scores in the case where we do an uncorrelated fit.

6. Conclusion

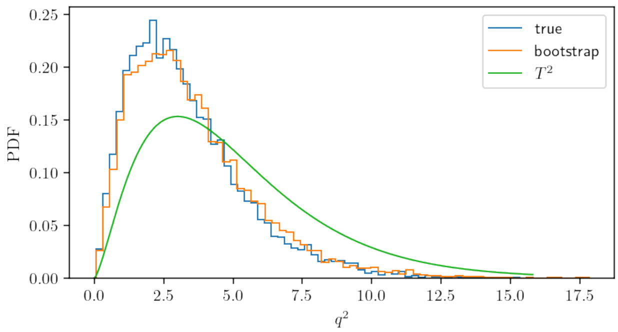
We have described a procedure to extract a p-value for a fit by bootstrapping the data, summarizing our work in [2]. This procedure is general enough to reproduce the standard analytical χ^2 or T^2 distributions in cases where they apply, but also faithfully reproduce the null distribution of goodness-of-fit scores in more realistic situations where we cannot make assumptions about the independence and normality of the statistical errors. Furthermore, when no analytic distribution is available, such as when we have modified the covariance matrix, the procedure is able to estimate the true distribution well. We believe that this work allows the lattice QCD practitioner more freedom in modifying the fitting procedure and metric to suit their needs, while also providing p-values that better reflect reality.



(a) T^2 vs. data obtained from multivariate Gaussian



(b) T^2 vs. data with autocorrelations $\tau_{\text{int}} \approx 1.2$



(c) T^2 vs. uncorrelated fit

Figure 5: q^2 scores: T^2 vs. true vs. bootstrap distributions

References

[1] B. Efron, *Bootstrap Methods: Another Look at the Jackknife*, *The Annals of Statistics* **7** (1979) 1 .

[2] N.H. Christ, R. Eranki, C. Kelly and RBC and UKQCD Collaborations, *Bootstrap-determined p values in lattice qcd*, *Physical Review D* **111** (2025) 074514.

[3] C. Kelly and T. Wang, *Update on the improved lattice calculation of direct CP-violation in K decays*, *PoS* **LATTICE2019** (2019) 129 [1911.04582].

[4] RBC, UKQCD collaboration, *Direct CP violation and the $\Delta I = 1/2$ rule in $K \to \pi\pi$ decay from the standard model*, *Phys. Rev. D* **102** (2020) 054509 [2004.09440].

[5] E. Carlstein, *The Use of Subseries Values for Estimating the Variance of a General Statistic from a Stationary Sequence*, *The Annals of Statistics* **14** (1986) 1171 .

[6] B. Efron and R. Tibshirani, *Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy*, *Statistical Science* **1** (1986) 54 .

[7] M.H. QUENOUILLE, *NOTES ON BIAS IN ESTIMATION*, *Biometrika* **43** (1956) 353 [https://academic.oup.com/biomet/article-pdf/43/3-4/353/987603/43-3-4-353.pdf].

[8] *Abstracts of Papers*, *The Annals of Mathematical Statistics* **29** (1958) 614 .

[9] H. Hotelling, *The Generalization of Student's Ratio*, *The Annals of Mathematical Statistics* **2** (1931) 360 .

[10] B. Yoon, Y.-C. Jang, C. Jung and W. Lee, *Covariance fitting of highly correlated data in lattice QCD*, *J. Korean Phys. Soc.* **63** (2013) 145 [1101.2248].

[11] K. Mardia, J. Kent and J. Bibby, *Multivariate analysis*, Probability and mathematical statistics, Acad. Press, London [u.a.] (1979).

[12] M. Bruno and R. Sommer, *On fits to correlated and auto-correlated data*, *Comput. Phys. Commun.* **285** (2023) 108643 [2209.14188].

[13] C. Michael, *Fitting correlated data*, *Phys. Rev. D* **49** (1994) 2616 [hep-lat/9310026].

[14] K.A. Bollen and R.A. Stine, *Bootstrapping goodness-of-fit measures in structural equation models*, *Sociological Methods & Research* **21** (1992) 205 [https://doi.org/10.1177/0049124192021002004].