Covariance Fitting of Highly-correlated Data in Lattice QCD

Boram YOON and Yong-Chull JANG
Lattice Gauge Theory Research Center, FPRD, and CTP,
Department of Physics and Astronomy, Seoul National University, Seoul 151-747, Korea

Chulwoo JUNG
Physics Department, Brookhaven National Laboratory, Upton, NY 11973, USA

Weonjong LEE∗
Lattice Gauge Theory Research Center, FPRD, and CTP,
Department of Physics and Astronomy, Seoul National University, Seoul 151-747, Korea

(Received 13 March 2013, in final form 18 April 2013)

We address a frequently-asked question on the covariance fitting of highly-correlated data such as our $B_K$ data based on the SU(2) staggered chiral perturbation theory. Basically, the essence of the problem is that we do not have a fitting function accurate enough to fit extremely precise data. When eigenvalues of the covariance matrix are small, even a tiny error in the fitting function yields a large chi-square value and spoils the fitting procedure. We have applied a number of prescriptions available in the market, such as the cut-off method, modified covariance matrix method, and Bayesian method. We also propose a brand new method, the eigenmode shift (ES) method, which allows a full covariance fitting without modifying the covariance matrix at all. We provide a pedagogical example of data analysis in which the cut-off method manifestly fails in fitting, but the rest work well. In our case of the $B_K$ fitting, the diagonal approximation, the cut-off method, the ES method, and the Bayesian method work reasonably well in an engineering sense. However, interpreting the meaning of $\chi^2$ is easier in the case of the ES method and the Bayesian method in a theoretical sense aesthetically. Hence, the ES method can be a useful alternative optional tool to check the systematic error caused by the covariance fitting procedure.

PACS numbers: 11.15.Ha, 12.38.Gc, 12.38.Aw
Keywords: Lattice QCD, $B_K$, CP violation
DOI: 10.3938/jkps.63.145

I. INTRODUCTION

We have reported results of $B_K$ calculated using improved staggered fermions with $N_f = 2 + 1$ flavors in Ref. [1,2]. We refer to Refs. [1] and [2] as SW-1 and SW-2 afterwards. In SW-1, we use three different lattice spacings to control the discretization errors. The dominant error in this result comes from the uncertainties in the matching factors. One hidden uncertainty is that we use the diagonal approximation (uncorrelated fitting) instead of the full covariance fitting in SW-1. In fact, one of the most frequently-asked questions on SW-1 is why we do the uncorrelated fitting (i.e., use the diagonal approximation) instead of the full covariance fitting. Here, we would like to address that issue with respect to the covariance fitting and the diagonal approximation.

A significant difficulty in fitting highly-correlated data has been pointed out in the literature such as Refs. [3–7]. In the literature, they have proposed a number of prescriptions such as the diagonal approximation [6], the modified covariance matrix [7], and the cut-off method under the popular name of a singular value decomposition (SVD) [3–5]. The weakness of these approaches is that all try to modify the covariance matrix one way or another. Hence, we lose the true meaning of $\chi^2$, and we do not know the quality of the fitting. Therefore, we propose a new method, the eigenmode shift (ES) method, which does not modify the covariance matrix but only use our freedom to modify the fitting functional form based on the Bayesian method. The ES method allows for a probability interpretation of the quality of the fitting based on the Bayesian $\chi^2$ distribution [8]. An alternative approach is the orthodox Bayesian method. In this approach, we add higher order terms to the fitting function with proper constraints until it fits the data. This also turns out to be another good solution to the problem.

∗E-mail: wlee@snu.ac.kr
The paper is organized as follows. In Sec. II, we review the covariance fitting process and give a physical meaning of the covariance matrix. In Sec. III, we address the problem with small eigenvalues of the covariance matrix. In Sec. IV, we list the possible solutions to the problem and discuss their pros and cons. Here, we explain the ES method. In Sec. V, we give a theoretical background for the pdf (probability distribution function) of the eigenvalues of the covariance matrix. In Sec. VI, we close with some concluding remarks.

II. REVIEW OF COVARIANCE FITTING

Let us consider $N$ samples of unbiased estimates of the quantity $y_i$ with $i = 1, 2, 3, \ldots, D$. Here, the data set is $\{y_i(n)\}_{n=1}^N$. Let us assume that the samples $y_i(n)$ are statistically independent in $n$ for fixed $i$ but are substantially correlated in $i$. For example, a similar situation occurs in lattice gauge theory calculations where there are $N$ independent gauge configurations and $B_K$ functions measured over the gauge configurations. An introduction to this subject is given in Ref. [9–12].

The fitting functional form suggested by the SU(2) staggered chiral perturbation theory (SChPT) is linear as follows:

$$ f_{th}(X) = \sum_{a=1}^P c_a F_a(X), $$

where $c_a$ are the low energy constants (LECs) that we want to determine by fitting, and $F_a$ is a known function of $X$ that represents collectively $X_P$ (pion squared mass of $\bar{\pi}\pi$), $X_D$ (pion squared mass of $\bar{\pi}d$), and so forth. The details on $F_a$ and $X$ are given in SW-1. Here, we focus on the X-fit of the 4X3Y-NNLO fitting of the SU(2) SChPT, which is explained in great detail in SW-1. In this fit, we have three LECs, so $P = 3$.

We are interested in the probability distribution of the average $\bar{y}_i$ of the data $y_i(n)$:

$$ \bar{y}_i = \frac{1}{N} \sum_{n=1}^N y_i(n). $$

We assume that the measured values of $\bar{y}_i$ have a normal distribution $P(\bar{y})$, by the central limit theorem for the multivariate statistical analysis, as follows:

$$ P(\bar{y}) = \frac{1}{Z} \exp\left[-\frac{1}{2} \sum_{i,j=1}^D (\bar{y}_i - \mu_i)(N \Gamma^{-1}_{ij})(\bar{y}_j - \mu_j)\right], $$

where $\mu_i$ represents the true mean value of $y_i$, which is, in general, unknown and can be obtained as $N \to \infty$, and

$$ Z = \int [\delta \bar{y}] \exp\left[-\frac{1}{2} \sum_{i,j=1}^D (\bar{y}_i - \mu_i)(N \Gamma^{-1}_{ij})(\bar{y}_j - \mu_j)\right]. $$

Here, $\Gamma_{ij}$ is the true covariance matrix, which is, in general, unknown in our problems. The maximum likelihood estimator of $\Gamma_{ij}$ turns out to be the sample covariance matrix $S_{ij}$ defined as follows:

$$ S_{ij} = \frac{1}{N-1} \sum_{n=1}^N \left[ y_i(n) - \bar{y}_i \right] \left[ y_j(n) - \bar{y}_j \right], $$

$$ C_{ij} = \frac{1}{N} S_{ij}, $$

where $C_{ij}$ is the normalized sample covariance matrix. Here, note that the covariance matrix is a symmetric, positive semidefinite matrix that has real and non-negative eigenvalues. We assume that our theory must describe the data well. Then,

$$ \mu_1 \to \nu_1 = f_{th}(X_1) = \sum_{a=1}^P c_a F_a(X_1). $$

In other words, we want to test whether $\nu_1$ describes the data reliably from the standpoint of statistics. In this procedure, we want to determine $c_a$ (LECs) to give the best fit. Here, the best fit is defined by minimizing the $T^2$ of the numerical results $\{\bar{y}_i\}$, where the $T^2$ is defined by

$$ T^2 = \sum_{i,j=1}^D \left[ y_i - \nu_i \right] \left[ N S_{ij}^{-1} \right] \left[ y_j - \nu_j \right]. $$

We notice that $Y_i = \sqrt{N} [y_i - \nu_i]$ is distributed according to $\mathcal{N}(\rho_i, \Gamma)$ and $\rho_i = \sqrt{N} [y_i - \nu_i]$. Here, we use the same notation as in Ref. [11]. Then, we note that $(N - 1)S_{ij}$ is independently distributed as

$$ \sum_{n=1}^{N-1} Z_i(n) Z_j(n), $$

where $Z(n)$ is distributed according to $\mathcal{N}(0, \Gamma)$. In this case, $[T^2/(N - 1)]((N + d - 1)/d)$ is distributed as a non-central $F$ distribution of $F_{d,N-d}$, which is defined in Ref. [11], and its non-centrality parameter is

$$ \sum_{i,j} \rho_i \Gamma_{ij}^{-1} \rho_j = \sum_{i,j} (\mu_i - \nu_i) (N \Gamma_{ij}^{-1}) (\mu_j - \nu_j). $$

Here, $d$ is the number of degrees of freedom (dof) of the fitting and is defined by $d = D - P$. In Ref. [11], the limiting distribution of $T^2$ as $N \to \infty$ is proven to be the $\chi^2$-distribution with $d$ degrees of freedom if $\mu_i = \nu_i$.

At this point, we have to minimize $T^2$ in order to determine the LECs: $\{c_a\}$. Hence, we need to solve the following equation:

$$ \frac{\partial T^2}{\partial c_a} = 0. $$

Journal of the Korean Physical Society, Vol. 63, No. 2, July 2013