

**The Sequential Empirical Bayes Method: An
Adaptive Constrained-Curve Fitting
Algorithm for Lattice QCD**

χ QCD Collaboration:

Y. Chen, S.-J. Dong, T. Draper, I. Horváth, F.-X. Lee,
K.-F. Liu, N. Mathur, C. Srinivasan, S. Tamhankar and J. Zhang

OUTLINE

- Introduction
- The Sequential Empirical Bayes Method
- Some Applications
- Summary

I. Introduction

- The Formalism of Lattice QCD: Comparisons

| Experiments | Lattice QCD | Statistical Mechanics |
|---------------------------------------|---|-----------------------|
| Facilities, such as Colliders | Monte Carlo Simulation | Ensemble |
| Probes, such as Detectors, etc. | Green's Function Correlation Functions | Observable |
| Data Analysis | Data Analysis | Data Analysis |

- **Monte Carlo Estimates of Correlation Functions**

The two-point function of a hadron operator $O(t)$ is calculated by the Monte Carlo simulation,

$$\langle G(t) \rangle = \langle O(t)O(0) \rangle,$$

(here $\langle \rangle$ represents the vacuum expectation value) which can be fitted by a theoretical model,

$$G(t; w_i, m_i) = \sum_i^{\infty} w_i e^{-m_i t}$$

where m_i and w_i are the mass and the spectral weights of the i^{th} state.

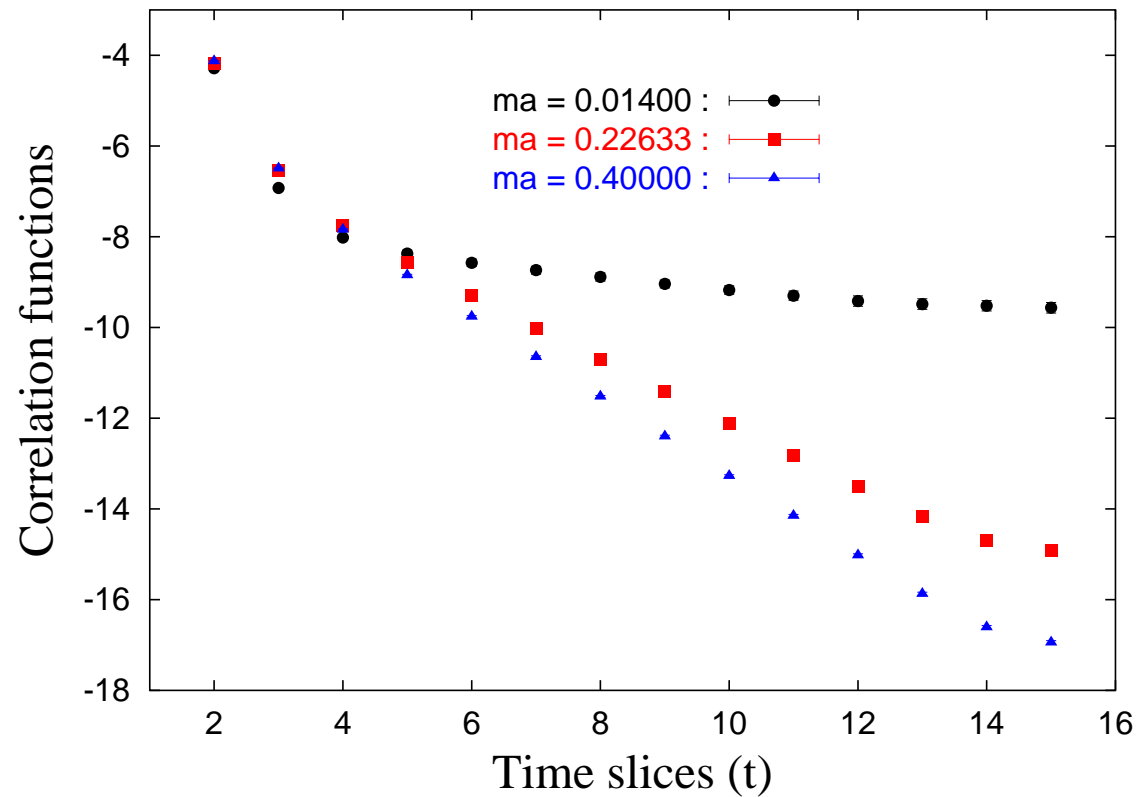


Figure 1: Two-point correlation function $\langle A_4 A_4 \rangle$ for the pion for three bare quark masses.

- Deriving Physical Quantities for MC DaTa

The maximum-likelihood fitting procedure

- (a) The Normal minimal- χ^2 method (statistically independent data)

$$\chi^2(w_i, m_i) = \sum_t \frac{(\langle G(t) \rangle - G(t; w_i, m_i))^2}{\sigma_t^2}$$

- (b) The correlated minimal- χ^2 method (statistically dependent data)

$$\chi^2(w_i, m_i) = \sum_{t,t'} (\langle G(t) \rangle - G(t; w_i, m_i)) \sigma_{t,t'}^{-2} (\langle G(t') \rangle - G(t'; w_i, m_i))$$

with covariance matrix

$$\sigma_{t,t'} = \langle G(t)G(t') \rangle - \langle G(t) \rangle \langle G(t') \rangle$$

- **Reliable Derivation of Excited States Is Desirable**

(a) Ground states can be easily derived.

When the time t is large enough that $\langle G(t) \rangle$ is dominated by the contribution of the ground state, $\langle G(t) \rangle$ can be well fitted by a function form of single exponential.

(b) The extraction of excited states is more complicated.

The fitted parameters of higher states fluctuate violently if the conventional maximum likelihood procedure is directly performed.

(c) It is desirable to derive reliable information of excited states.

- **Constrained Curve Fitting-**”*Teaching the Fitting Physics*”

(a) Adding to the χ^2 by an extra prior term

$$\chi_{\text{aug}}^2 = \chi^2 + \chi_{\text{prior}}^2, \quad \chi_{\text{prior}}^2 = \sum_i \frac{(\rho_i - \tilde{\rho}_i)^2}{\tilde{\sigma}_i^2}$$

where ρ_i denotes the collective parameters of the fit.

(b) The Bayesian priors stabilize the fitting of higher states.

The Bayesian priors, say, values of the parameters obtained from *a priori* estimates $\rho_i = \tilde{\rho}_i \pm \tilde{\sigma}_i$, will guide the fit to achieve stability. With improved stability, the data sets can be enlarged to include small t and the theory can be enlarged by including many more terms in the fit model until convergence is obtained.

(c) The systematic error associated with the choice of t_{\min} is thereby largely absorbed into the statistical error.

II. The Sequential Empirical Bayes Method

- The Standard Constrained-Curve Fitting

The conventional χ^2 can be rewritten compactly as

$$\chi^2(\rho) = \sum_{\alpha\beta} (M_\alpha(\rho) - D_\alpha) \sigma_{\alpha\beta}^{-2} (M_\beta(\rho) - D_\beta),$$

where ρ are the collective parameters of the fit (e.g. $\rho_i = \{w_i, m_i\}$ for a sum of exponentials), the indices α, β distinguish different values of the independent variable (time for correlation function) and different interpolating fields, D are the Monte Carlo data, $M_\alpha(\rho)$ is the fit model, and $\sigma_{\alpha\beta}^2$ is the covariance matrix.

- The Standard Constrained-Curve Fitting (cont'd)

- (a) Traditional inference (*frequentist theory*)

Minimizing the χ^2 is the solution of the problem of determining the set of fit parameters ρ which maximizes

$P(D|\rho)$, the conditional probability of measuring the data D given a set of parameters,

also known as the "likelihood" of the data.

- (b) Bayesian inference

Bayesian inference asks which fit-model parameters are most likely given the data. It demands that the solution of the curve-fitting problem consist of determining the set of parameters ρ which maximizes

$P(\rho|D)$, the conditional probability that ρ is correct given the measured data D .

- The Standard Constrained-Curve Fitting (cont'd)

(c) Bayes theorem

$$P(\rho|D) = \frac{P(D|\rho)P(\rho)}{P(D)} = \frac{P(D|\rho)P(\rho)}{\int d\rho P(D|\rho)P(\rho)}$$

which follows directly from the elementary properties of probability theory

$$P(\rho|D)P(D) = P(\rho \cap D) = P(D|\rho)P(\rho)$$

(d) The likelihood $P(D|\rho)$ will have Gaussian statistics, as assured by the **Central Limit Theorem**, regardless of the statistics of the underlying distribution (the ensemble of all possible configurations),

$$P(D|\rho) \propto \exp(-\chi^2/2)$$

- The Standard Constrained-Curve Fitting (cont'd)

(e) Gaussian prior distribution

$$P(\rho) \propto \exp(-\chi_{\text{prior}}^2/2)$$

(f) Thus $P(\rho|D) \propto P(D|\rho)P(\rho)$ is maximized by minimizing the augmented χ^2

$$\chi_{\text{aug}}^2 = \chi^2 + \chi_{\text{prior}}^2$$

(g) Priors stabilize the fitting

Minimizing $\mathcal{A}(\rho) + \lambda\mathcal{B}(\rho)$,

$$\frac{\delta}{\delta\rho} (\mathcal{A}(\rho) + \lambda\mathcal{B}(\rho)) = 0.$$

If $\mathcal{A}(\rho)$ is degenerate, but $\mathcal{B}(\rho)$ is not, the degeneracy is lifted.

(i) Here comes a question:

How do we obtain the priors?

- **Overview of the Sequential Empirical Bayes Method (SEB)**

- (a) SEB is inspired by the constrained curve fitting and devote much effort to get priors.

- (b) For concreteness, consider again the fit model of the sum of decaying exponentials

$$G(t; w_i, m_i) = \sum_i^{\infty} w_i e^{-m_i t}$$

- (c) **Sequential**: *Dealing with the states one by one.*

- (d) **Empirical**: *In a manner of self-tuning.*

1. Scanning
2. Monitoring

- (e) **Bayes**: *In the spirit of constrained curve fitting.*

- **The Basic Algorithm**

- (a) Choose t_{\max} and t_{\min}
- (b) Determine the number of the terms we want to use in the fit model, and determine t_{start} as the starting point of the fit. Ensure that the number of the data points is larger than the number of the parameters to be fitted.
- (c) Choose the initial values w_1 and m_1 . For each, use an unconstrained fit on the one-mass-term model to fit the correlator data including time slices t_{start} to t_{\max} and obtain $w_1^{(1)} \pm \sigma_{w_1}^{(1)}$ and $m_1^{(1)} \pm \sigma_{m_1}^{(1)}$. Choose as input for the next step those values which yield the lowest χ^2/dof .
- (d) Using these values of $w_1 \pm \sigma_{w_1}$ and $m_1 \pm \sigma_{m_1}$ as both the priors and initial values, do a constrained curve fit (using the one-mass-term model on the data set enlarged to include $t_{\text{start}} - 1$ to obtain $w_1^{(2)} \pm \sigma_{w_1}^{(2)}$ and $m_1^{(2)} \pm \sigma_{m_1}^{(2)}$.
- (e) Loop on a wide range of trial values for w_2 and m_2 . With a two-mass-term model, constrain the first mass and weight (using the

previous output as priors and initial values) but leave the second mass and weight unconstrained. Loop on various trial values for the latter. Do this half-constrained fit on the data set enlarged to include $t_{\text{start}} - 2$ and obtain $w_2^{(3)} \pm \sigma_{w_2}^{(3)}$ and $m_2^{(3)} \pm \sigma_{m_2}^{(3)}$. Choose as input for the next step those values which yield the lowest (but reasonable) χ^2/dof .

- (f) Using these values of the first two states as both priors and initial values, do a fully-constrained fit (using the two-mass-term model on the data set enlarged to include $t_{\text{start}} - 3$) to obtain the $w_1^{(4)} \pm \sigma_{w_1}^{(4)}$, $m_1^{(4)} \pm \sigma_{m_1}^{(4)}$, $w_2^{(4)} \pm \sigma_{w_2}^{(4)}$, and $w_1^{(4)} \pm \sigma_{w_1}^{(4)}$.
- (g) Repeat the last two steps until all desired mass terms and time slices are included. One thus obtains a complete set of priors.
- (i) Add the final time slice t_{min} and do a fully-constrained fit using previously obtained values for priors and initial guesses.

| Step | Time Slices Fitted | Scanned Initial Values | Priors (& Other Initial Values) | Fitted Output Values |
|------|--|------------------------|--|--|
| 1 | $\{t_{\text{start}}, t_{\text{max}}\}$ | w_1, m_1 | — | $w_1^{(1)}, m_1^{(1)}, \sigma_{w_1}^{(1)}, \sigma_{m_1}^{(1)}$ |
| 2 | $\{t_{\text{start}} - 1, t_{\text{max}}\}$ | — | $\tilde{w}_1^{(2)} = w_1^{(1)}, \tilde{m}_1^{(2)} = m_1^{(1)}$ | $w_1^{(2)}, m_1^{(2)}, \sigma_{w_1}^{(2)}, \sigma_{m_1}^{(2)}$ |
| 3 | $\{t_{\text{start}} - 2, t_{\text{max}}\}$ | w_2, m_2 | $\tilde{w}_1^{(3)} = w_1^{(2)}, \tilde{m}_1^{(3)} = m_1^{(2)}$ | $w_1^{(3)}, m_1^{(3)}, \sigma_{w_1}^{(3)}, \sigma_{m_1}^{(3)}$ |
| 4 | $\{t_{\text{start}} - 3, t_{\text{max}}\}$ | | — | $\tilde{w}_1^{(4)} = w_1^{(3)}, \tilde{m}_1^{(4)} = m_1^{(3)}$ $\tilde{w}_2^{(4)} = w_2^{(3)}, \tilde{m}_2^{(4)} = m_2^{(3)}$ |
| 5 | $\{t_{\text{start}} - 4, t_{\text{max}}\}$ | w_3, m_3 | $\tilde{w}_1^{(5)} = w_1^{(4)}, \tilde{m}_1^{(5)} = m_1^{(4)}$ $\tilde{w}_2^{(5)} = w_2^{(4)}, \tilde{m}_2^{(5)} = m_2^{(4)}$ | $w_1^{(5)}, m_1^{(5)}, \sigma_{w_1}^{(5)}, \sigma_{m_1}^{(5)}$ $w_2^{(5)}, m_2^{(5)}, \sigma_{w_2}^{(5)}, \sigma_{m_2}^{(5)}$ |
| 6 | $\{t_{\text{start}} - 5, t_{\text{max}}\}$ | | — | $\tilde{w}_1^{(6)} = w_1^{(5)}, \tilde{m}_1^{(6)} = m_1^{(5)}$ $\tilde{w}_2^{(6)} = w_2^{(5)}, \tilde{m}_2^{(6)} = m_2^{(5)}$ $\tilde{w}_3^{(6)} = w_3^{(5)}, \tilde{m}_3^{(6)} = m_3^{(5)}$ |

- **A More Sophisticate Algorithm**

(a) Choose a suitable time range.

The available data are in the time range $t_{\min} - t_{\max}$.

(b) Determine the initial value of the lowest state.

From effective mass plots, choose an initial time range $t \in [t_{\min}, t_{\max}]$ to do an unconstrained one-mass fit. Use "*scanning*".

(c) Monitor the behavior of χ^2/dof when adding more data points.

Include one more time slice and repeat an (independent) unconstrained one-mass fit, and monitor the fitted parameters m_1, w_1 and the χ^2/dof .

(d) Add one more mass term to the fit model if a sharp jump of χ^2/dof is observed.

If the fitted parameters and the χ^2/dof don't change much, include one more time slice and repeat the previous step. This iteration stops if there is a noticeable change of χ^2/dof and the values of the fitted parameters indicating a breakdown of the one state model. Then set $t_1 - 1$ equal to the time at which the χ^2/dof jumps, indicating the necessity of a two-mass fit for $t < t_1$. Set the priors for the ground state mass and weight equal to the fitted values from the last low- χ^2/dof fit over $t \in [t_1, t_{\max}]$.

- (e) Do a partially constrained two-mass fit (with scanning) for $t \in [t_1 - 1, t_{\max}]$.
The ground state priors are fixed at the values determined by the previous step. The first-excited state is unconstrained but scanned.
- (f) Repeat the above steps for the third, the fourth state, and so on, until available time slices are exhausted.

Add time slices until the two-state model breaks down as indicated by a jump in the χ^2/dof . Then set $t_2 - 1$ equal to the time at which the χ^2/dof jumps, indicating the necessity of a three-mass fit for $t < t_2$. Set the priors for the ground state and first-excited state mass and weight equal to the fitted values from the last low- χ^2/dof fit over $t \in [t_2, t_{\max}]$. Note that the ground-state priors are refreshed.

- (g) The highest state in the fit model will be absorbing all the contributions from higher states in the true function, and thus its fitted parameters will differ from the true values. Thus the highest state in the fit model must be rejected.

- **An Application of SEB to an Artificial Sample**

(a) An sample of artificial data: given a function, $G(t)$, which is a sum of five decaying exponentials, with means for masses and weights and independent Gaussian error.

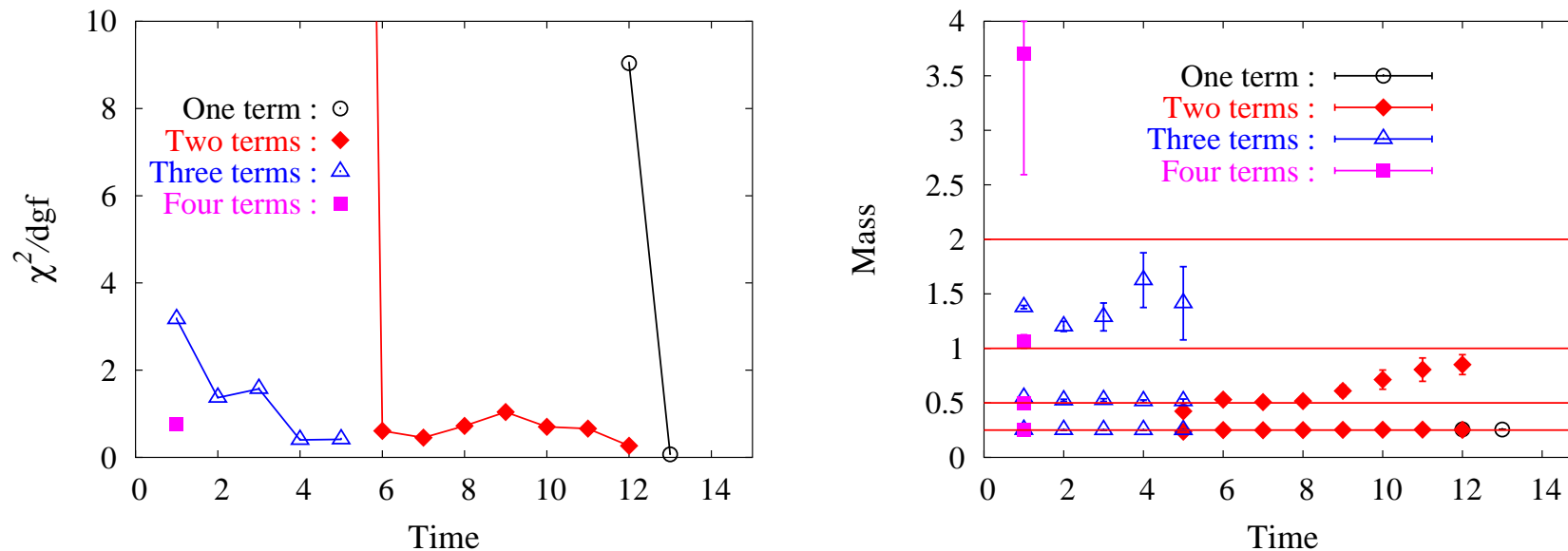


Figure 2: Left: Behavior of the χ^2/dof as earlier time slices are added to the fit range.

Right: The behavior of the fitted masses in the fitting procedure.

- An Application of SEB to an Artificial Sample (cont'd)

(b) Same as (a), but with larger errors.

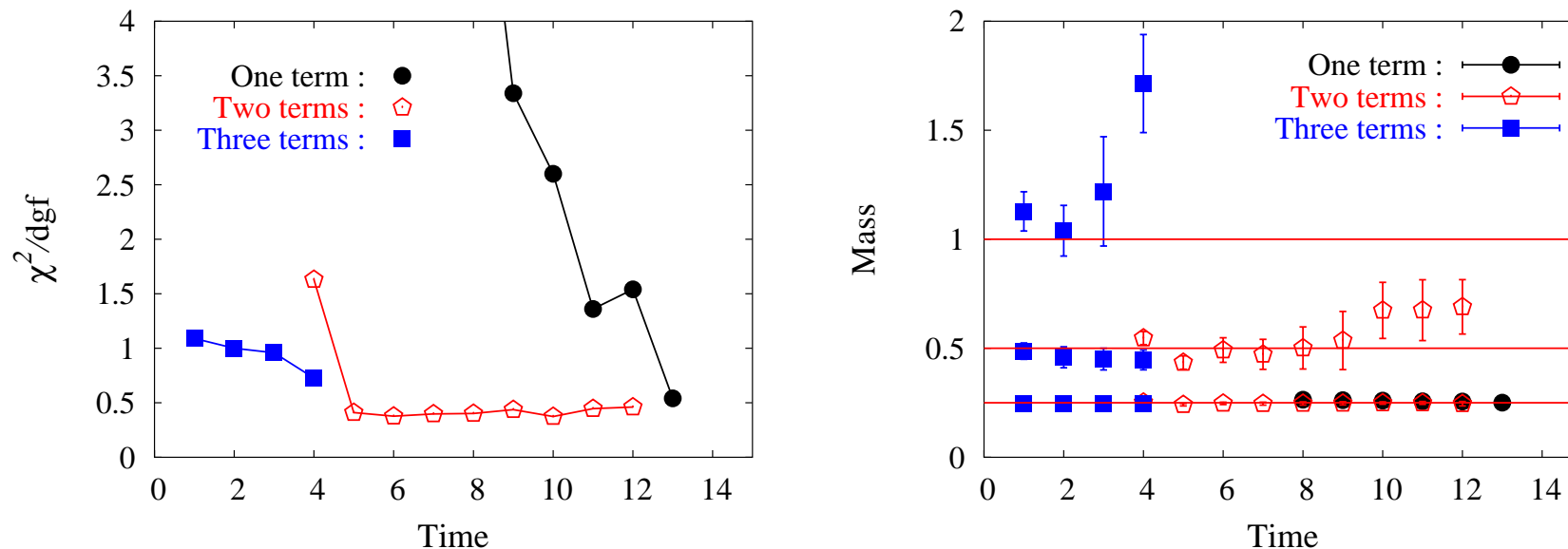


Figure 3: Left: Behavior of the χ^2/dof as earlier time slices are added to the fit range.

Right: The behavior of the fitted masses in the fitting procedure.

- An Application of SEB to an Artificial Sample (cont'd)

(b) Same as (a), but with even larger errors.

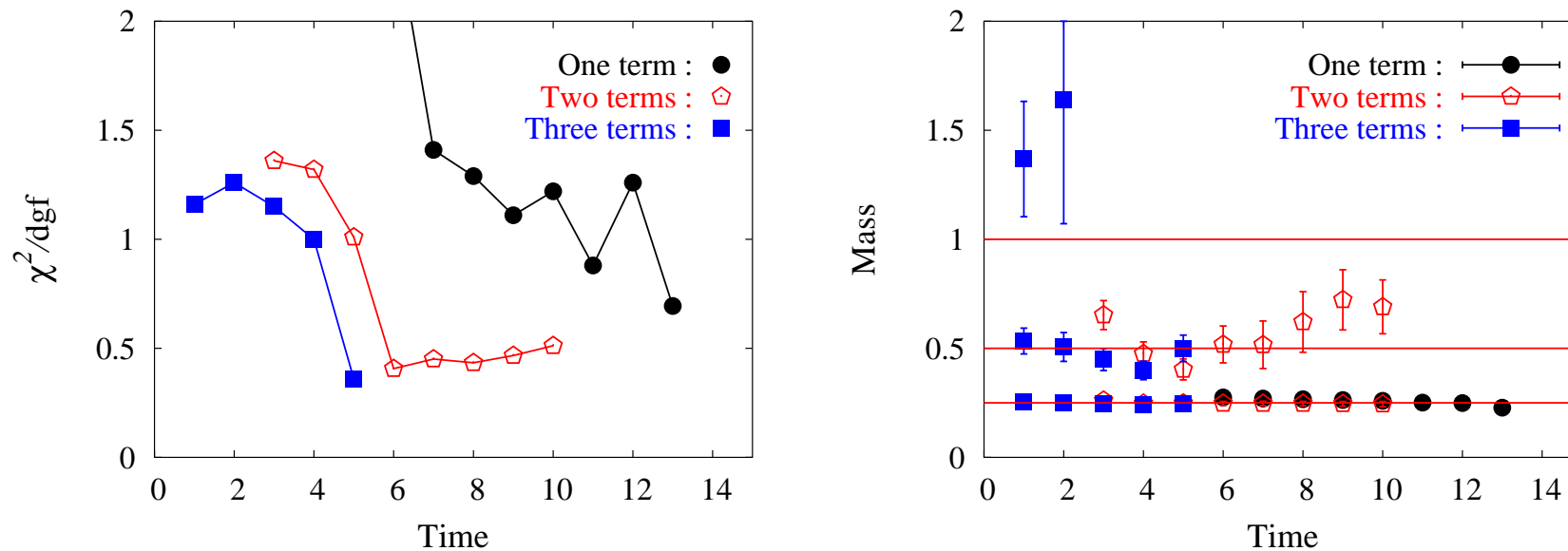


Figure 4: Left: Behavior of the χ^2/dof as earlier time slices are added to the fit range.

Right: The behavior of the fitted masses in the fitting procedure.

- **Some Explanations**

Suppose the artificial data is created with the three-state model function

$$\begin{aligned} G^{\text{true}}(t) &= w_1 e^{-m_1 t} + w_2 e^{-m_2 t} + w_3 e^{-m_3 t} \\ &= w_1 e^{-m_1 t} \left(1 + \frac{w_2}{w_1} e^{-(m_2 - m_1)t} + \frac{w_3}{w_1} e^{-(m_3 - m_1)t} \right), \end{aligned} \quad (1)$$

with fixed values of input parameters and Gaussian statistical error $\delta G(t)$ at each time slice t .

- **Some Explanations**

Then SEB should work if the following holds:

(a) There exists a t_1 such that in the time range $t_1 < t < t_{\max}$,

$$\frac{w_2}{w_1} e^{-(m_2-m_1)t} < \delta G(t) \frac{1}{G(t)}, \quad (2)$$

so that in this range, the data can be fitted by $w_1 e^{-m_1 t}$. That is, there is a “plateau” in the effective mass plot for large time.

(b) There exists a t_2 such that in the time range $t_2 \leq t < t_1$,

$$\frac{w_3}{w_1} e^{-(m_3-m_1)t} < \delta G(t) \frac{1}{G(t)} < \frac{w_2}{w_1} e^{-(m_2-m_1)t} \quad (3)$$

so that in this range, the data can be well fitted by $w_1 e^{-m_1 t} + w_2 e^{-m_2 t}$.

(c) In the time range $t < t_2$, the third state becomes important, and the full three-state model must be used in the fit.

- **Another Toy Model**

To illustrate, consider artificial data, constructed from the following three-state toy model

$$\begin{aligned} G(t; w_i, m_i) &= \sum_{i=1}^3 w_i e^{-m_i t} \\ &= 500e^{-0.85t} + 400e^{-1.35t} + 400e^{-1.75t} \end{aligned} \tag{4}$$

in the time range $0 \leq t \leq 15$, with relative errors (uncorrelated and Gaussian) increasing with time t to mimic the actual LGT data.

1. $t \in [t_1, t_{\max}]$, **fit the ground state**;
2. $t \in [t_2, t_{\max}]$, **two-mass-term fit**;
3. $t_{\min} = 0 \leq t_2 \leq t_1 \leq t_{\max} = 15$, **three state**.

We use (t_1, t_2) to denote the specific fitting procedure.

- Another Toy Model (cont'd)

(a) Lowest precision: $\delta G(t)/G(t) \sim 0.01$ at $t = 0$, and $\delta G(t)/G(t) \sim 0.1$ at $t = 15$

| (t_1, t_2) | (10, 7) | | |
|----------------|----------|-----------|---------|
| | state 1 | state 2 | state 3 |
| mass(input) | 0.850 | 1.35 | 1.75 |
| mass(fitted) | 0.855(7) | 3.10(62) | 1.51(5) |
| weight(input) | 500 | 400 | 400 |
| weight(fitted) | 521(34) | 0.0004(1) | 769(35) |

Table 1:

- Another Toy Model (cont'd)

(c) Low precision: $\delta G(t)/G(t) \sim 0.006$ at $t = 0$, and $\delta G(t)/G(t) \sim 0.05$ at $t = 15$

| | state 1 | state 2 | state 3 |
|----------------------|----------|----------|----------|
| mass(input) | 0.850 | 1.35 | 1.75 |
| mass((10,2), 0.527) | 0.854(4) | 1.50(8) | 1.86(92) |
| mass((10,4), 0.527) | 0.853(5) | 1.39(16) | 1.65(17) |
| weight(input) | 500 | 400 | 400 |
| weight((10,2),0.527) | 518(23) | 701(63) | 69(62) |
| weight((10,4),0.527) | 511(29) | 369(188) | 408(188) |

Table 2:

- **Another Toy Model (cont'd)**

(c) **High precision:** $\delta G(t)/G(t) \sim 0.001$ at $t = 0$, and $\delta G(t)/G(t) \sim 0.01$ at $t = 15$

| | state 1 | state 2 | state 3 |
|----------------------|----------|----------|----------|
| mass(input) | 0.850 | 1.35 | 1.75 |
| mass((10,5), 0.530) | 0.854(4) | 1.38(6) | 1.77(10) |
| weight(input) | 500 | 400 | 400 |
| weight((10,5),0.530) | 504(7) | 451(141) | 343(144) |

Table 3:

- **A Few More Words about SEB**

Let $\Delta G(t)$ be the absolute value of the difference of the function $G^{\text{true}}(t)$ of input parameters and the function $G^{\text{fitted}}(t)$ of fitted parameters

$$\Delta G(t) = |G^{\text{true}}(t) - G^{\text{fitted}}(t)|. \quad (5)$$

A fit $G^{\text{fitted}}(t)$ with reasonable small χ^2/dof implies that the relation

$$\Delta G(t) < \delta G(t) \quad (6)$$

roughly holds at most times t . In other words, with this statistical error $\delta G(t)$, we cannot distinguish the fitted function $G^{\text{fitted}}(t)$ from the original $G^{\text{true}}(t)$.

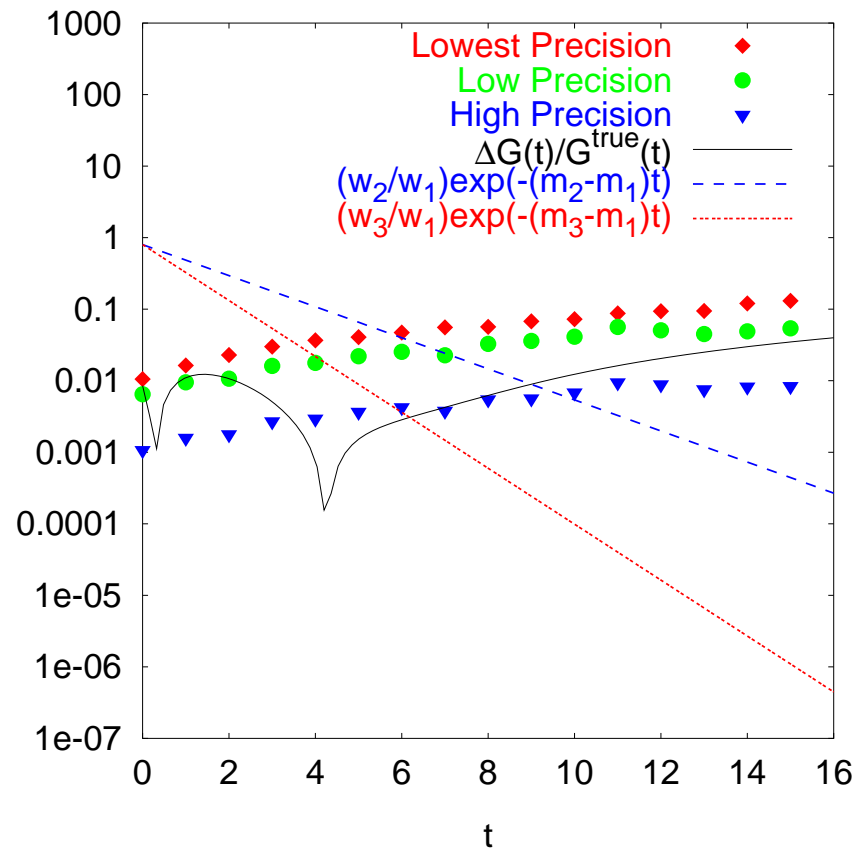


Figure 5: The relative errors of three data sets of increasing precision (“Lowest”, “Low”, and “High”) are plotted with points. The curved solid (black) curve is the plot of the function $\Delta G(t)/G^{\text{true}}(t)$, which is the relative difference between original function $G^{\text{true}}(t) = 500 \exp(-0.85t) + 400 \exp(-1.35t) + 400 \exp(-1.75t)$ and the false-positive fitted function $G^{\text{false}} = 521e^{-0.855t} + 769e^{-1.51t}$ from the data set (“Lowest Precision”).

III. SOME APPLICATIONS

- Pion Spectrum

- (a) On a 16^3 lattice, we use the overlap fermion and the Iwasaki gauge action with $\beta = 2.264$. The lattice spacing a is set to be $a = 0.200(3)$ fm by the measured pion decay constant f_π , and thus the lattice has spatial size of 3.2fm.
- (b) The correlation function $\langle A_4 A_4 \rangle$ for the pion at three bare quark masses.

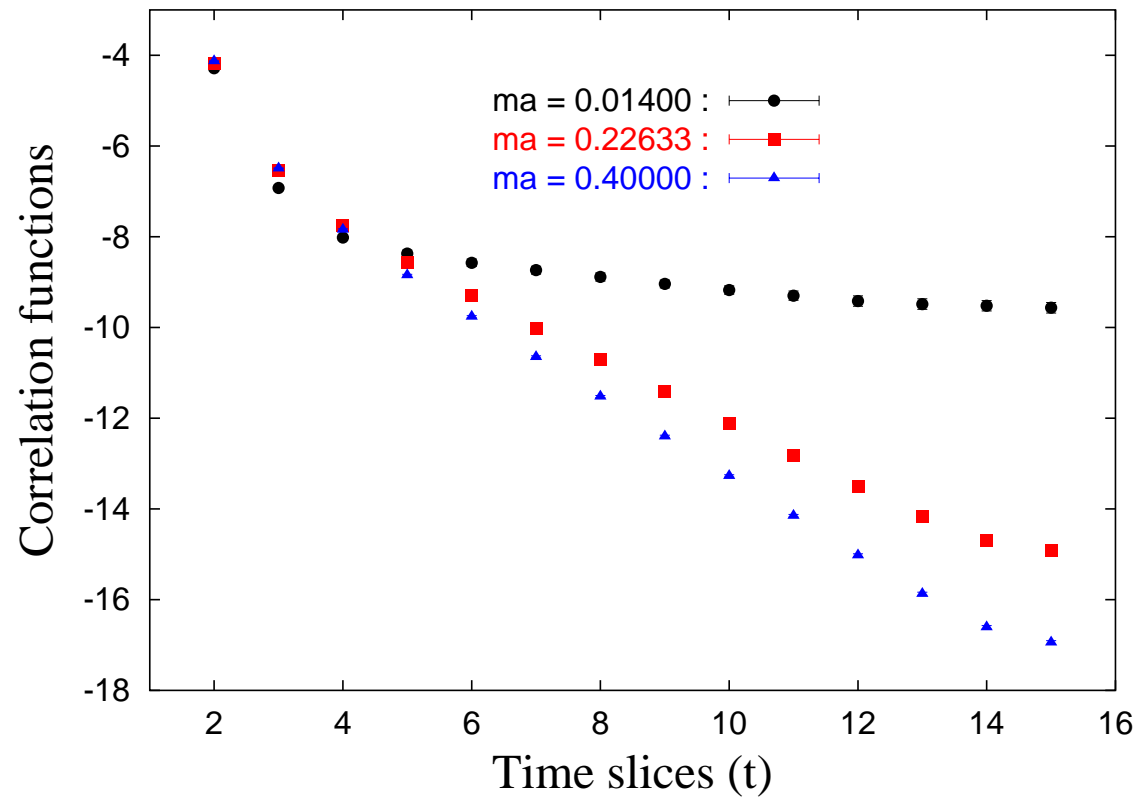


Figure 6: Two-point correlation function $\langle A_4 A_4 \rangle$ for the pion for three bare quark masses.

- Pion Spectrum (cont'd)

(c) Ground and first excited state from SEB

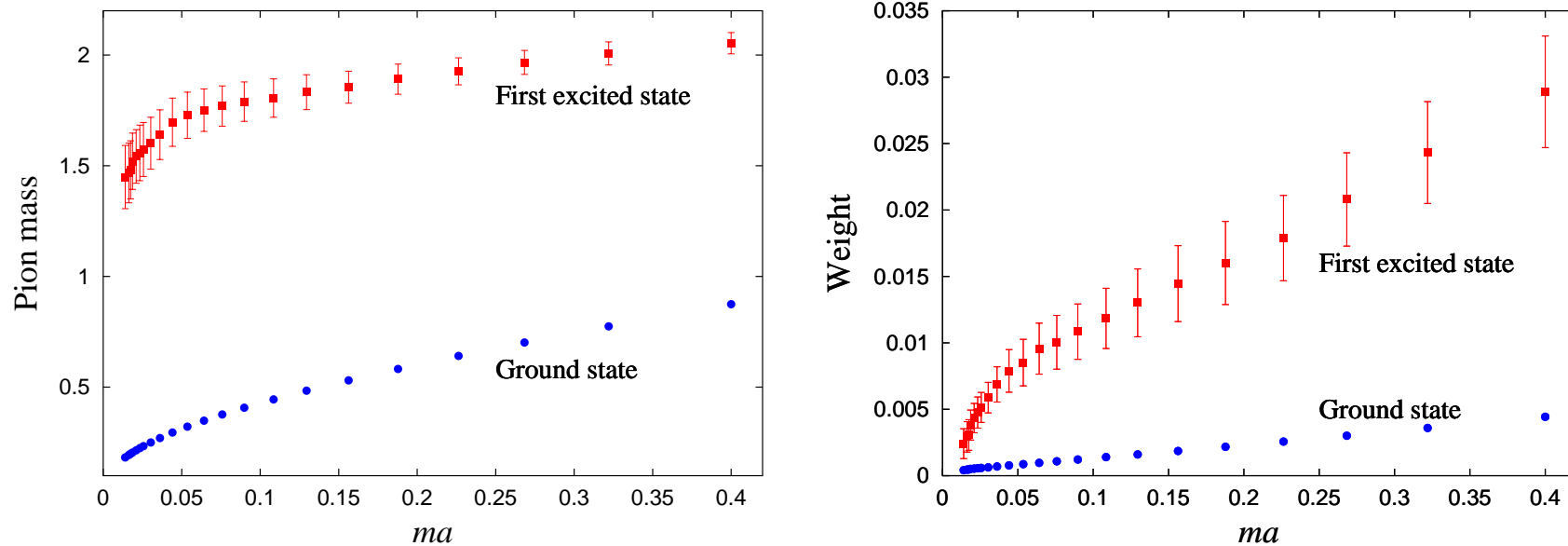


Figure 7: Ground and first-excited state pion mass $m_\pi a$, as a function of the bare quark mass ma (left). Ground and first-excited stated pion weight, as a function of the bare quark mass ma (right).

- **Handling Ghost States**

(a) Quenched artifacts associated with the absence of quark loops- "hairpin diagrams", which leads to two consequences:
chiral-log terms contributing to hadron masses;
ghost states in hadron propagators.

(b) The effect of $\eta'\pi$ ghost states on the a_0 propagator.
Ghost states can be modeled by the expression

$$G_{\text{ghost}} \sim w_{\text{ghost}}(1 + E_{\pi})e^{-m_{\eta'\pi}t}$$

where w_{ghost} is constrained to be negative, the $(1 + E_{\pi}t)$ factor reflects the double-pole nature of the hairpin diagram, and $m_{\eta'\pi}$ is the mass of the ghost state which is to be fitted.

On a finite box, E_{π} is discrete due to the discrete lattice momentum

$$p_i = \frac{2\pi n}{L},$$

$$E = \sqrt{m^2 + \sum_i \left(\frac{2}{a} \sin\left(\frac{p_i}{2}\right)\right)^2}$$

- Handling Ghost States (cont'd)

(b) The effect of $\eta'\pi$ ghost states on the a_0 propagator.

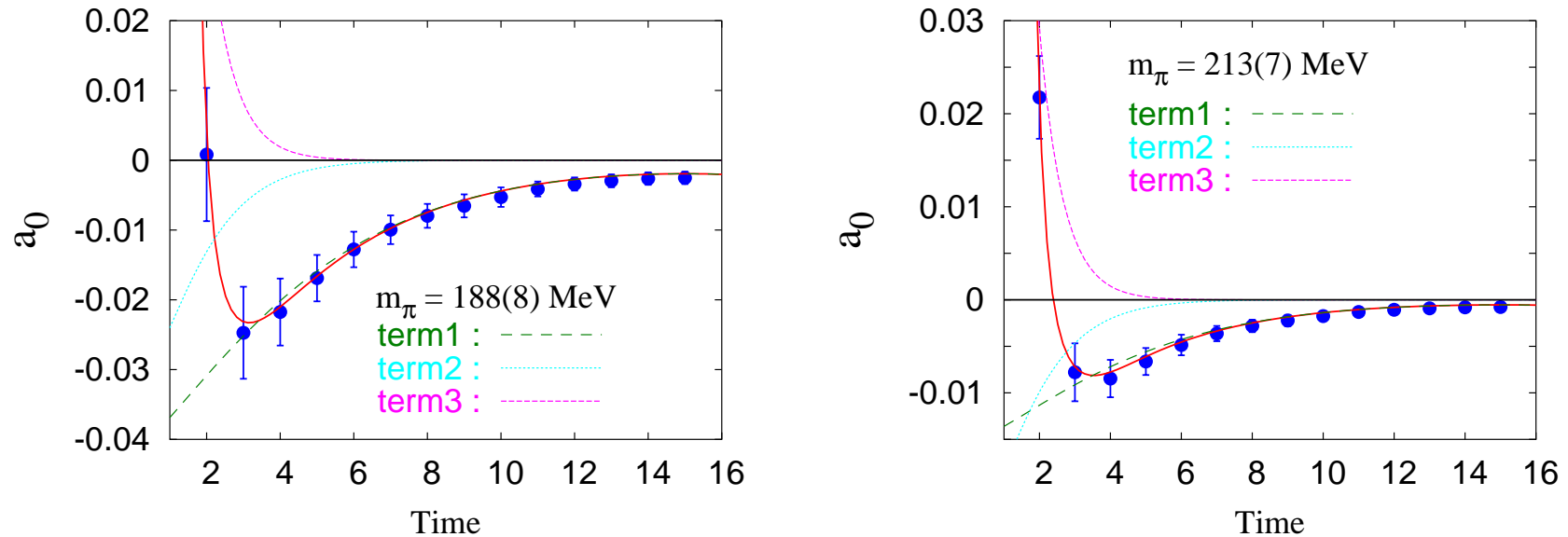


Figure 8: a_0 correlators for our lowest quark mass for which $m_\pi = 188(8)$ MeV (left) and $m_\pi = 213(7)$ MeV (right). The negative dip of the correlators is an indication of the domination of the ghost S -wave $\eta'\pi$ state over the physical a_0 . The curves are contributions of $n = 0, 1$ ghosts and the ground state to the fit model, and labeled as "term 1", "term 2" and "term 3".

- **Nucleon and Its Excited States**

- (a) Studies using standard curve fitting have heretofore failed to reliably identify the Roper resonance of the nucleon on the lattice.

- (b) We can study nucleon spectrum in the very light quark region where the lowest pion mass is as low as 180 MeV. It is found that the ghost contribution to the nucleon propagator is very important when quarks are very light.

- Nucleon and Its Excited States (cont'd)

(c) The ghost contribution to the S_{11} propagator.

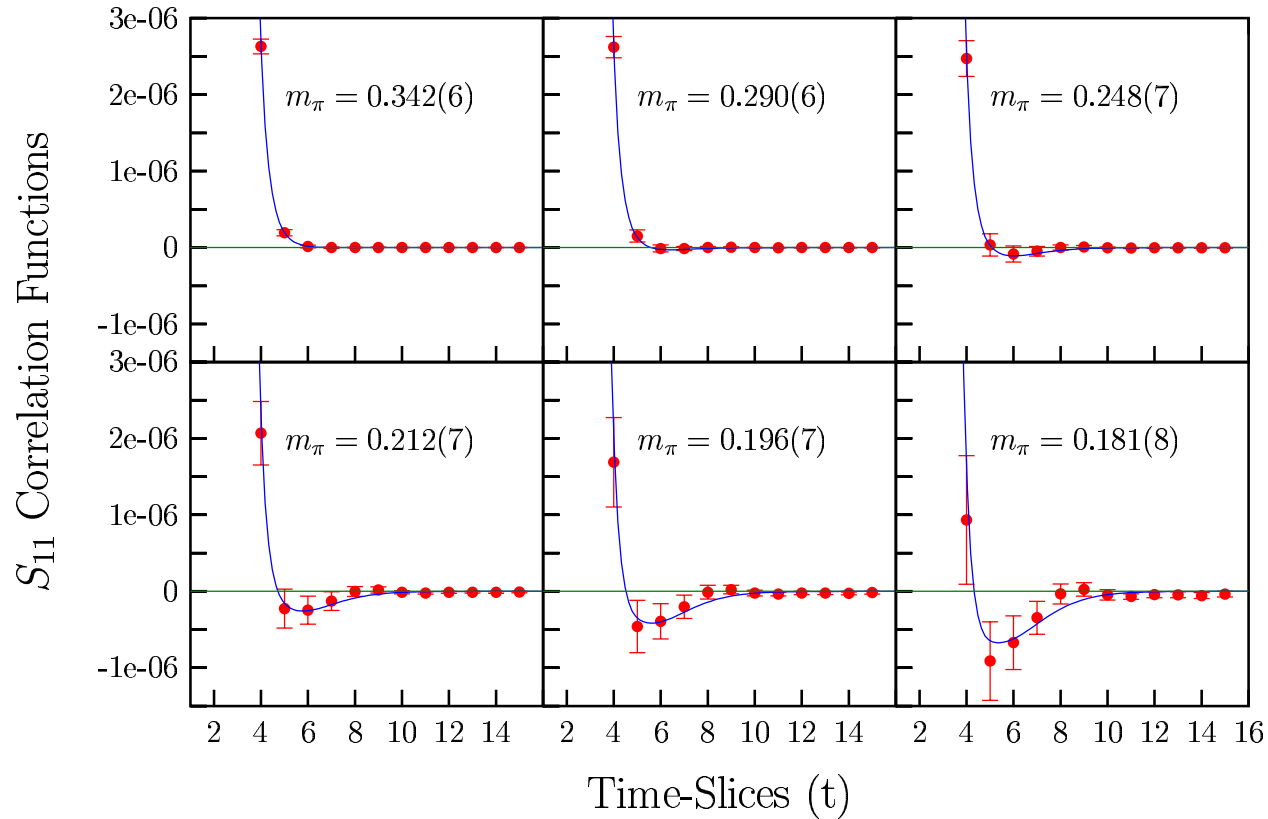


Figure 9: S_{11} correlators for six low quark masses. Negative dip of the correlators is an indication of the domination of the ghost $\eta'N$ state over the physical S_{11} . m_π are in GeV.

- Nucleon and Its Excited States (cont'd)

(d) The final result of the spectrum of nucleon and its excited states.

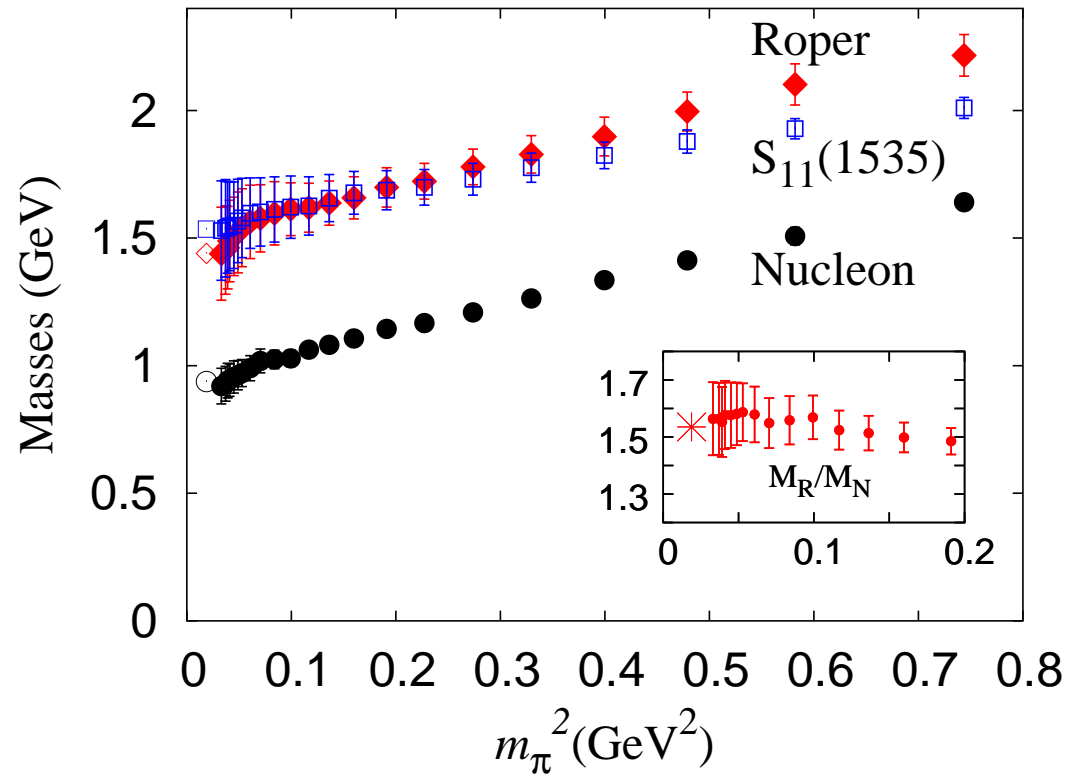


Figure 10: Nucleon, Roper, and S_{11} masses in GeV as a function of m_π^2 . The experimental values are indicated by the corresponding open symbols.

IV. Summary

- (a) We have advocated refinements of Bayesian-inspired constrained-curve fitting which better stabilizes fits.
- (b) In the "Sequential Empirical Bayes Method", we have constructed an automated and natural way of reliably obtaining the priors from naturally-nested subsets of the data.
- (c) Our algorithm can successfully recover the correct fit parameters of several artificial data sets, each of which is constructed as a sum of decaying exponentials with realistic values of the parameters.
- (d) Our method is not in strict accord with the Bayesian philosophy, as we use subsets of the data to guide the selection of the priors.
- (e) This method has been successfully applied to several realistic studies of hadron spectrum.

THANK YOU !