Bayesian Methods for Parameter Estimation from Lattice Simulations

Ethan T. Neil (Colorado/RIKEN-BNL)

QCDNA VIII, Yale University

June 21, 2014

Outline

- Some basics (telling you things you know in my notation)
- Fitting correlation functions
  - “Constrained curve fitting”
  - Two-stage constrained fitting
- Model uncertainty and marginalization
  - A Bayesian EFT-inspired approach
  - Example: continuum extrapolation

A word on the Bayesian approach

- This is a road that can lead to heated philosophical debates about the use of prior information…but I don’t want to go there!

- Bayesian methods here are used:

  - As a numerical tool, to improve stability of optimization algorithms etc (without changing the results);

  - As a formal framework (Bayes' theorem), to derive some useful probability formulas

- Everything I want to talk about here works more or less independently of choice of prior information

- Bonus disclaimer: I am a physicist, not a statistician

Bayes' theorem

$$p(A)p(B|A) = p(B)p(A|B)$$

- In the context of fitting a data set **D** to a model **M**, this formula is typically written in a slightly different way:

"likelihood"

"prior"

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)}$$

"posterior" or "evidence"

usually ignored

Fitting the model: least squares

- Fix the functional form of the model M (for now), in terms of a set of model parameters **a**.

- Least-squares minimization to find the best fit from the data. E.g. for a simple data series of points (x_i,y_i,σ_i), minimize

$$\chi^2 = \sum_i \left(\frac{y_i - f(x_i, \mathbf{a})}{\sigma_i} \right)^2$$

- Chi-squared determines the likelihood function:

$$p(D|M) \propto \exp\left(-\frac{\chi^2}{2} \right)$$

Parameter estimation

- In Bayesian terms, expectation values of the model parameters are determined by the posterior PDF:

$$\langle f(\mathbf{a})\rangle = \int_{-\infty}^{\infty} d\mathbf{a}\ f(\mathbf{a})p(M|D)$$

- Assuming our prior function p(M) is constant, then the posterior p(M|D) is proportional to the likelihood p(D|M)!

Taming instability in multi-exponential fits

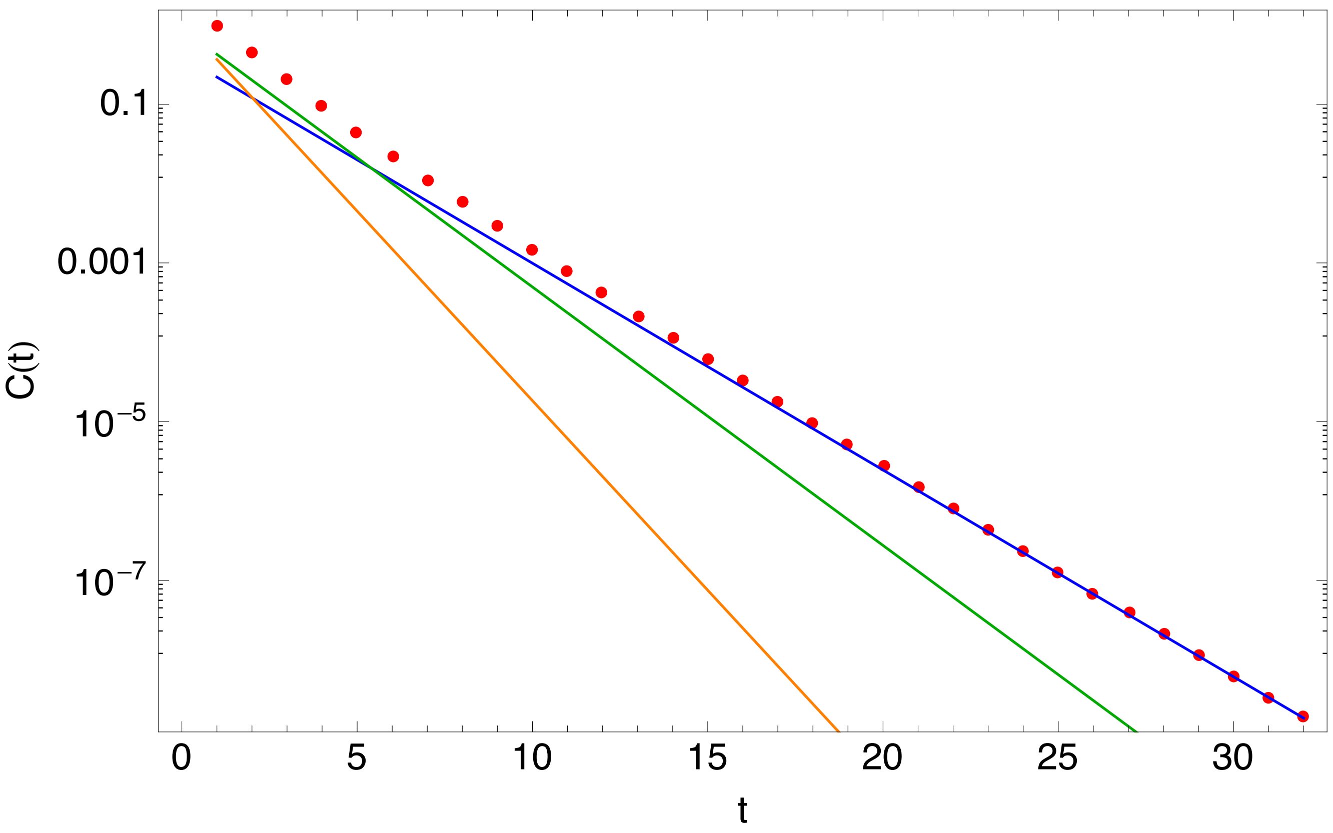
- A simple source of instability in multi-exponential fits is from ordering ambiguity: if our model function is just a sum of exponentials, who says "E_0" has to be the ground-state energy?

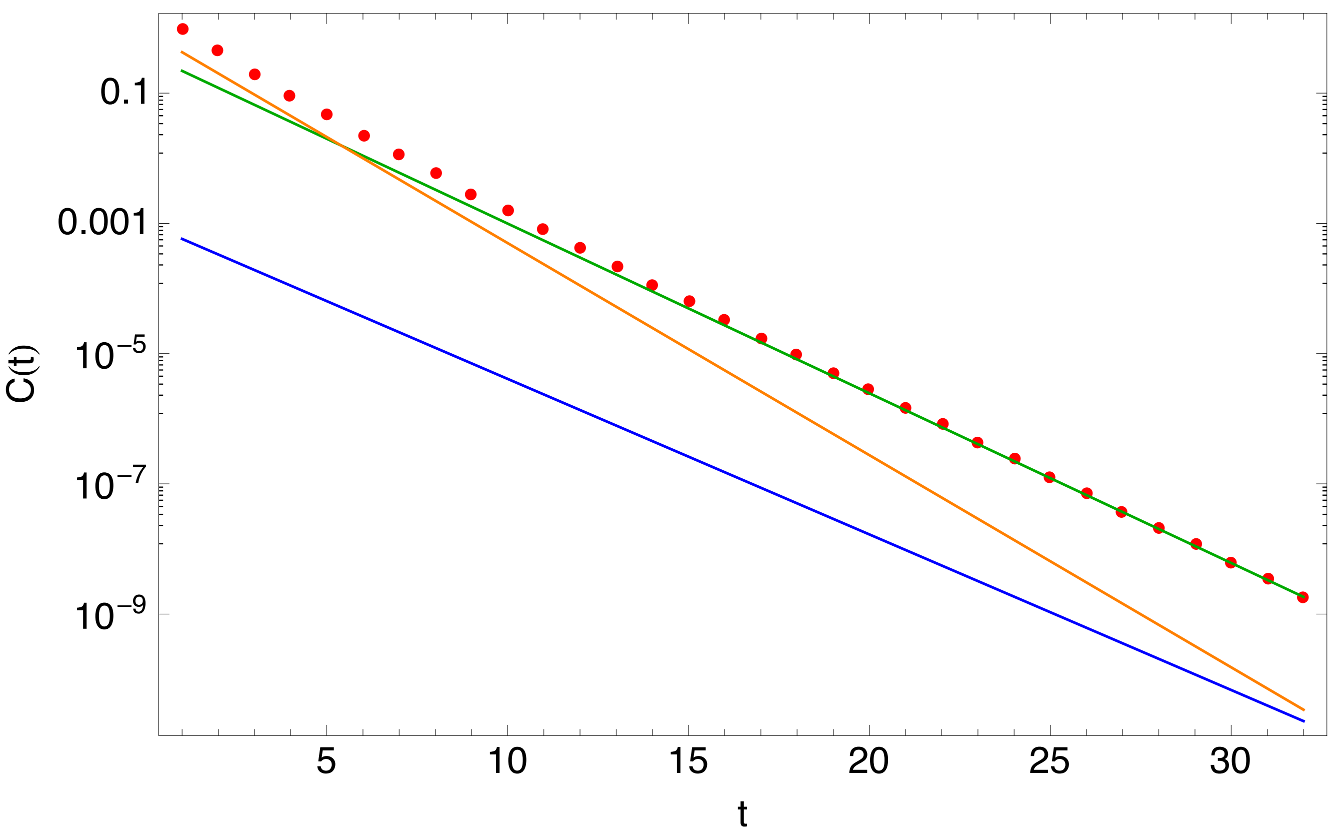
$$f(t, \vec{a}_i, \vec{E}_i) = \sum_{i=1}^{N} a_i e^{-E_i t}$$

- Mapping is a good way to resolve this: impose the ordering on the model, e.g. by using E0 and manifestly positive energy deltas as the parameters:

$$E_0, \log(E_1 - E_0), \log(E_2 - E_1), ...$$

- This leaves one ambiguity: the fitter can still attempt to model the data by making E_0 and a_0 very small (effectively zeroing out the state!) and then making state 1 the ground state. (From N! peaks to N peaks.)


0.1
0.001
10^{-5}
10^{-7}
C(t)
0
5
10
15
20
25
30
t


0.1
0.001
10^{-5}
10^{-7}
10^{-9}
C(t)
5
10
15
20
25
30
t

“Constrained curve fitting” (arXiv:hep-lat/0110175)

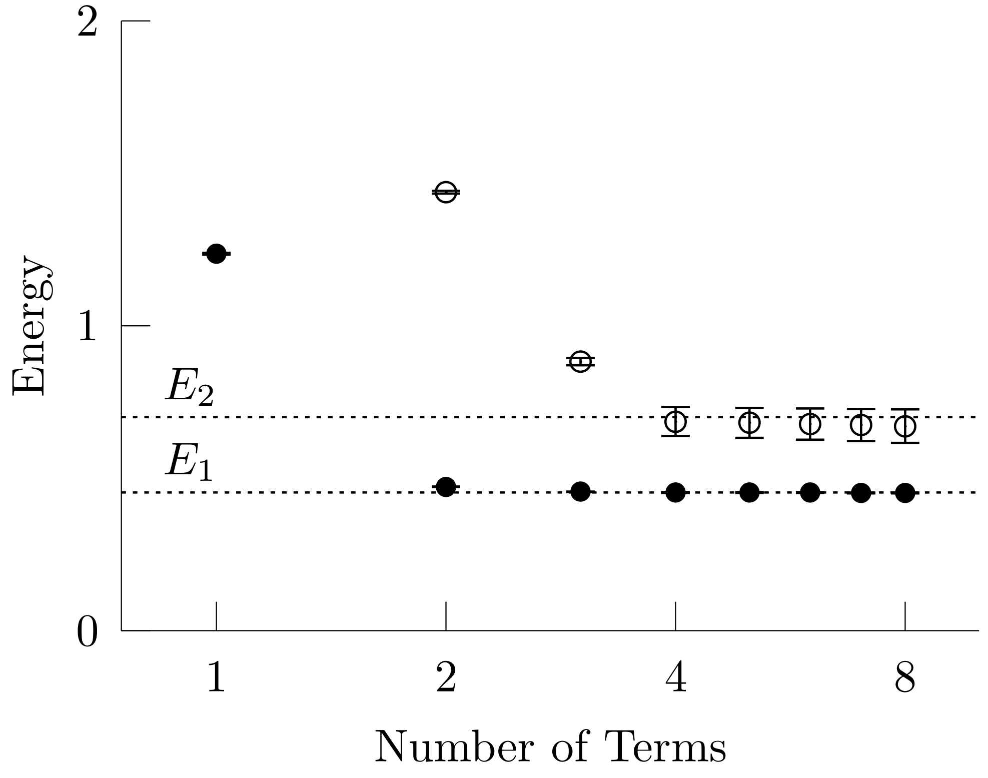
- What if we want to incorporate prior information on some or all of the parameters?

- Lepage et al, “constrained curve fitting”. Define the “augmented chi-squared” function by addition of Gaussian priors on the fit parameters:

$$\chi^2 \to \chi^2 + \chi_P^2$$

$$\chi_P^2 = \sum_j \frac{(a_j - \tilde{a_j})^2}{\tilde{\sigma_j}^2}$$

$$P(\color{blue}{D|M}\color{black})P(\color{red}{M}\color{black})$$



- Relatively mild priors can prevent optimizers from “wandering” off in a flat, unphysical direction. In practice, setting of “priors” tends to be driven by looking at a subset of the data, not really on external information…

“Constrained curve fitting”, continued

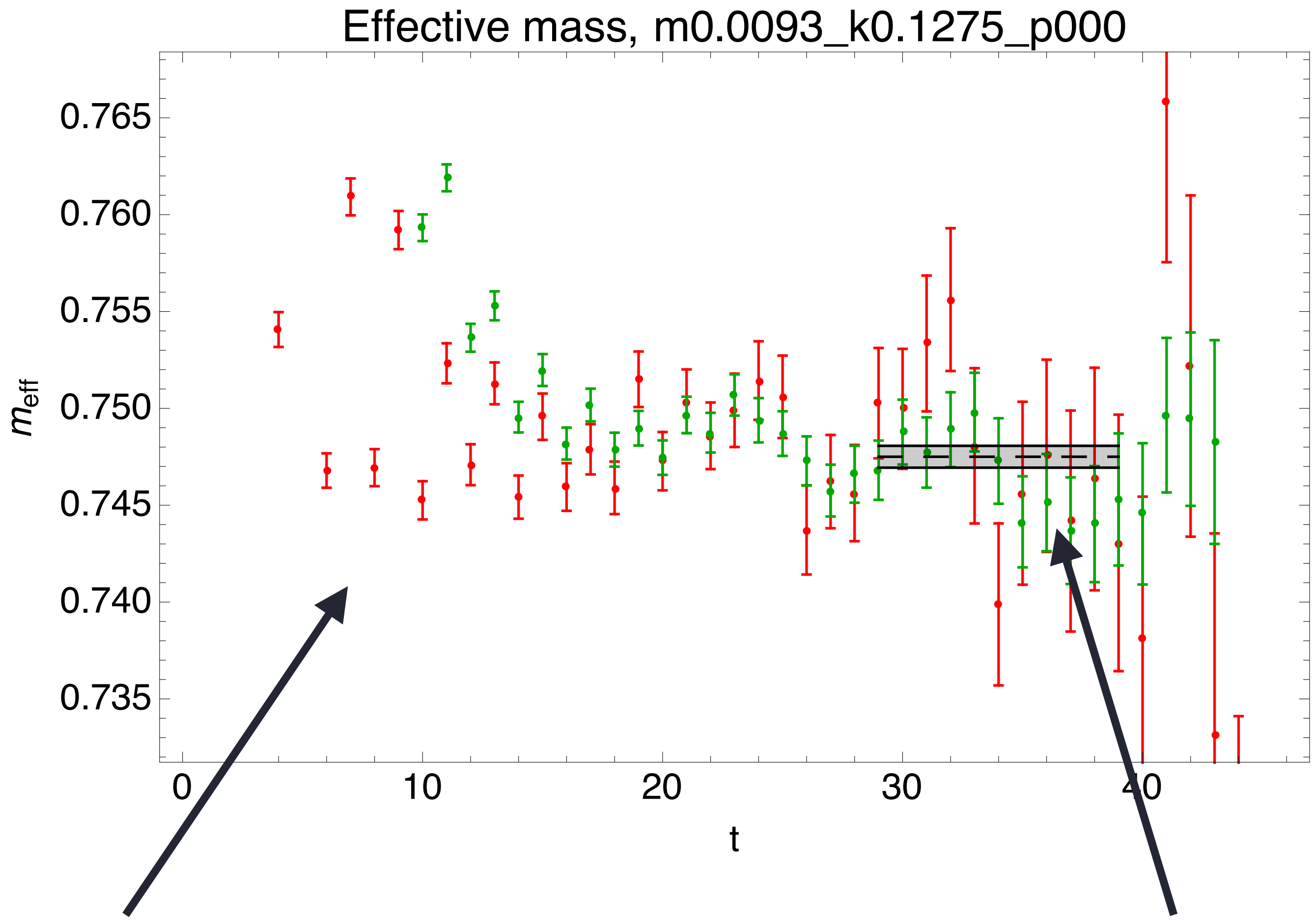
- There are other options to deal with multi-exponential fit stability:
  - Careful tuning of initial conditions (but you sort of need to know the right answer already…)
  - “Sequential” fitting algorithms (e.g. Kentucky group)
  - Multiple effective masses/Vandermonde polynomials (Fleming et al.)
  - Probably more alternatives I’ve forgotten

Two-stage constrained fits

- As long as we can maintain a good determination of the ground state, with imposed ordering the rest of the fit should be stable
- Idea: why not use the correlation function itself to constrain the ground state parameters?
    - Identify a “plateau” region, fit to estimate all ground-state parameters
    - Set priors using mean values as determined from plateau, and width = n times plateau fit uncertainty
    - Constrained fit to the full data set as normal
    - May be adaptable to be a black-box method?
- If we set the priors by looking at effective mass plots, we’re already more or less doing this in practice! Actually basing them on a fit to the same correlator can be more automated.
- (In principle, we could use n=1 and then fit the correlator below the plateau as normal, but only if the data are independent; and for a lattice correlator, they’re definitely not!)

Numerical test

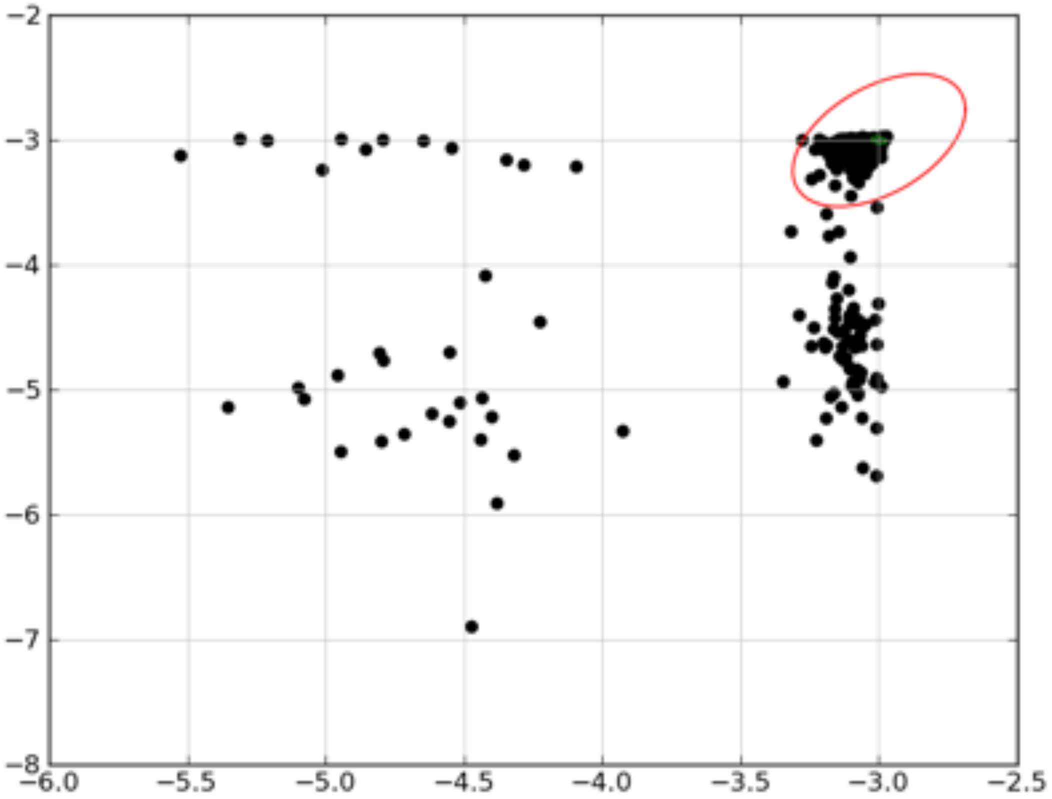
(from Fermilab/MILC heavy-light meson two-point correlator, a~0.09fm)

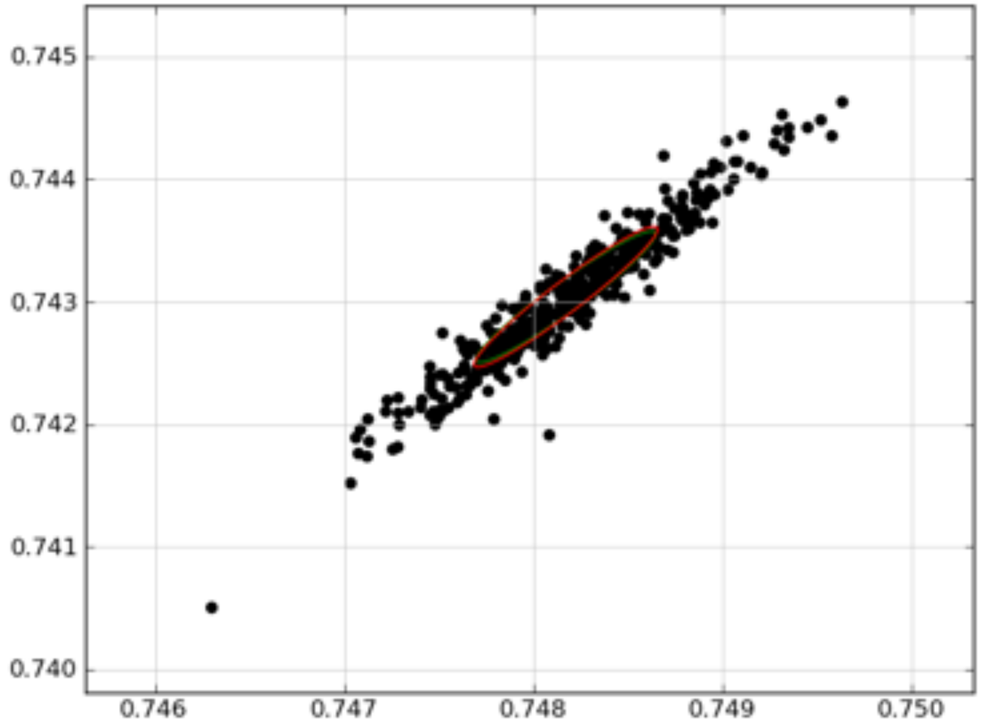


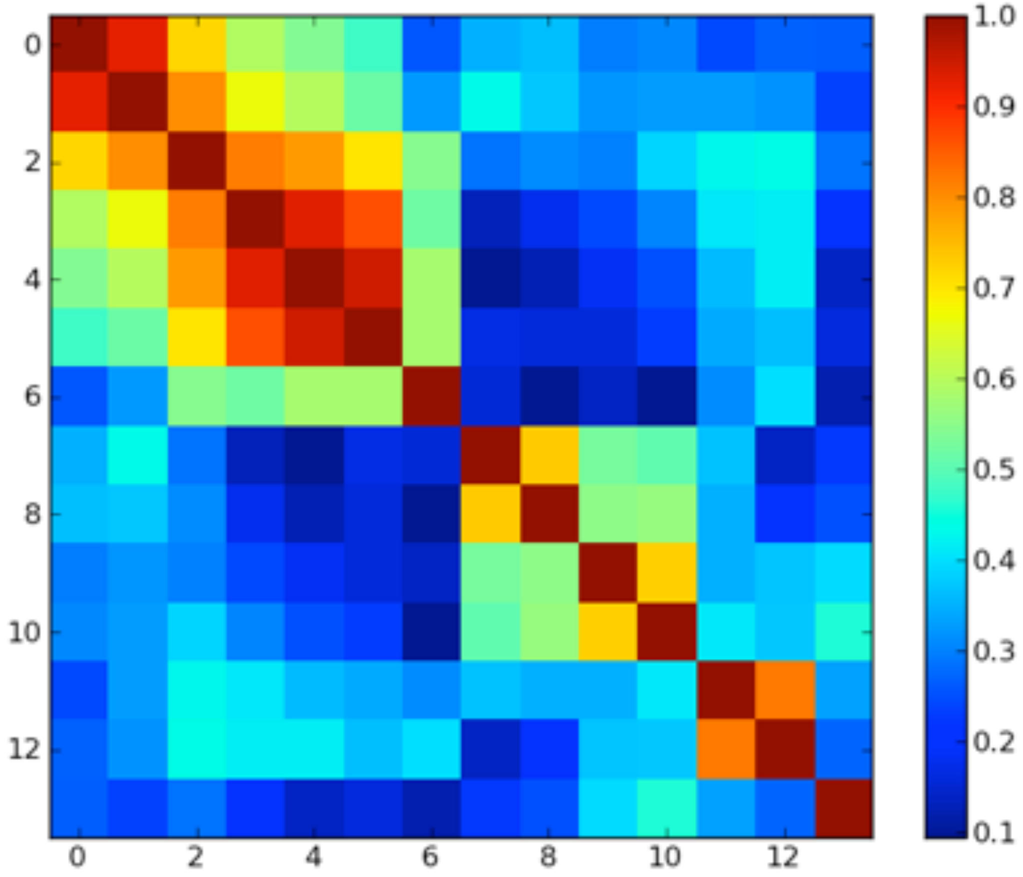
(fit full correlator out here)

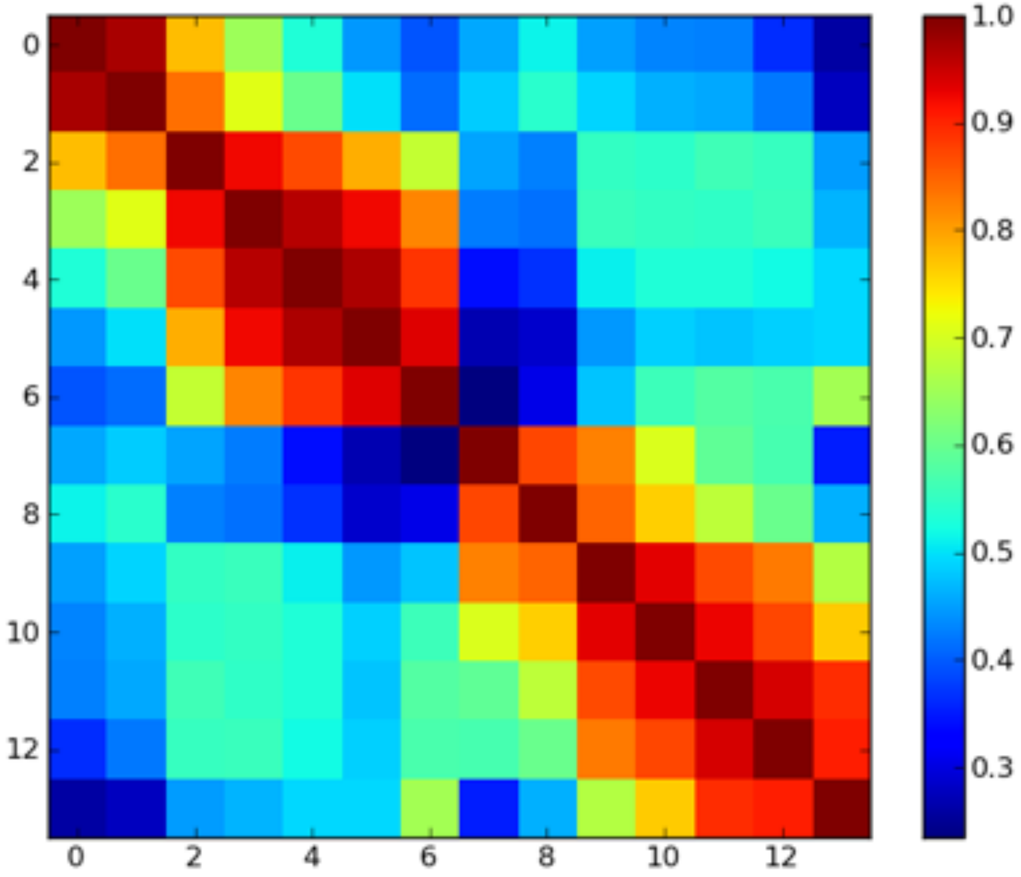
(set priors to result from out here, w/error scaled up)

Stability for replication methods (jackknife/bootstrap)!









Model selection and parameter estimation

- Parameter estimation requires a choice of model to be fit to the data

- Very often in lattice analysis (continuum extrapolation, chiral extrapolation, multi-exponential fits) our model is an effective field theory, based on an expansion with a formally infinite number of terms!

- How do we decide where to truncate? What is the error we make in doing so?

Systematic error estimation

- A common approach is to pick a reasonable stopping point (e.g. NLO chiral fit), then go one step further (NNLO). The difference between the parameter estimates from NLO and NNLO is taken to estimate the systematic error. This seems conservative!

- Some collaborations have tried method of weighted sum using p-values. Less conservative, but is it right?

- We can appeal to effective field theory and naturalness, and just assume that the coefficients of the neglected terms are "order one” and try to estimate things…but we can disagree on what “order one” means too.

- Bayesian methods give us a more rigorous way to decide what to do!

Bayesian model averaging (arXiv:0808.3643)

- Divide into a family of nested models, with some common parameters $\mathbf{a}_{\mathbf{res}}$ and additional parameters $\mathbf{a}_{\mathbf{marg}}$ (within each model)

$$\mathrm{pr}(\mathbf{a}_{res}|D) = \sum_M \int d\mathbf{a}_{marg} \frac{\mathrm{pr}(D|\mathbf{a}, M)\mathrm{pr}(\mathbf{a}|M)\mathrm{pr}(M)}{\mathrm{pr}(D)}.$$

- Priors on the higher-order terms are important to avoid runaway numerics, but we can marginalize over what we mean by “order one” if we’re clever!

Parameter estimation by marginalization

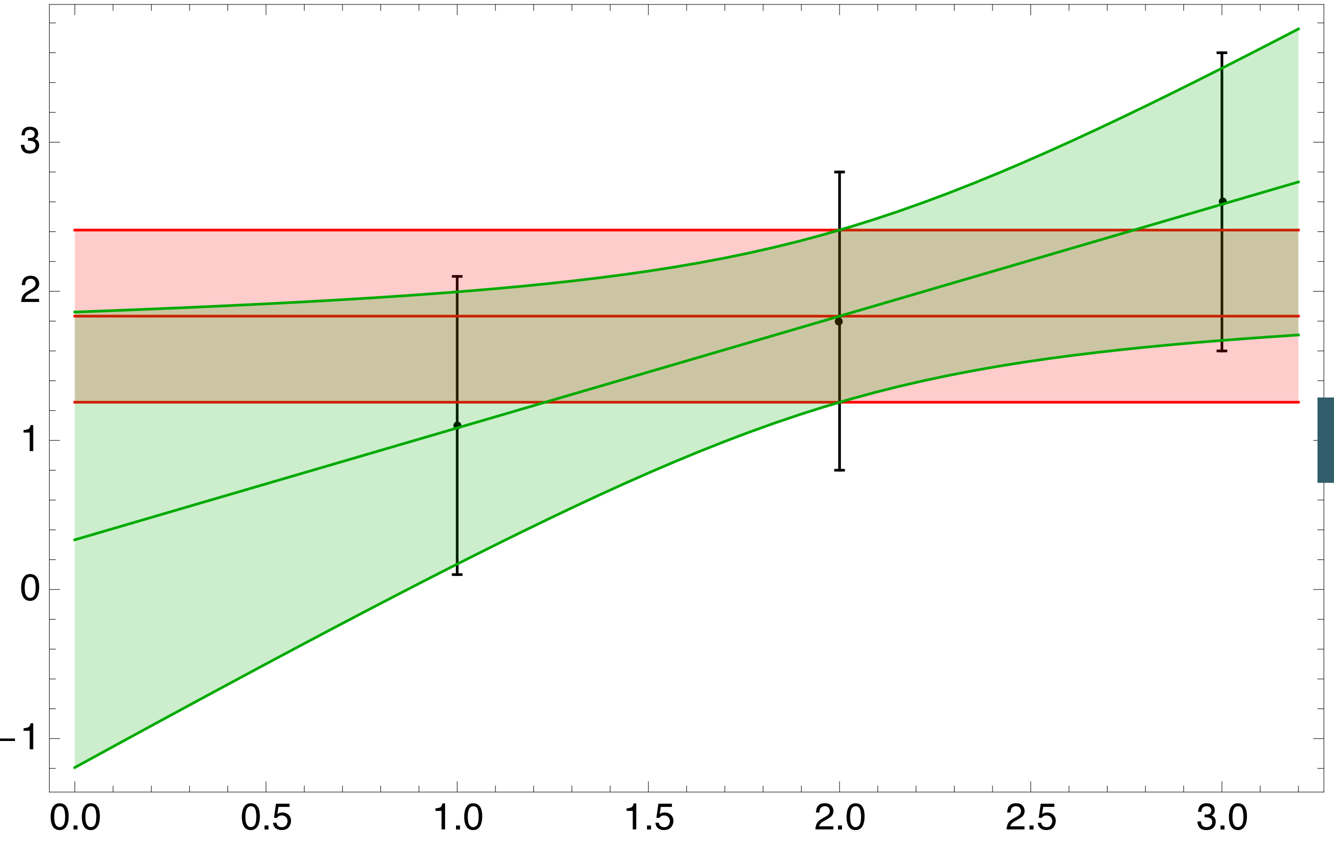
- Combination of model parameter estimates can be shown to be a weighted sum over individual expectation values, with weights given by the model evidence:

$$\langle f(a_0) \rangle = \sum_M \langle f(a_0) \rangle_M \, \mathrm{pr}(M|D).$$

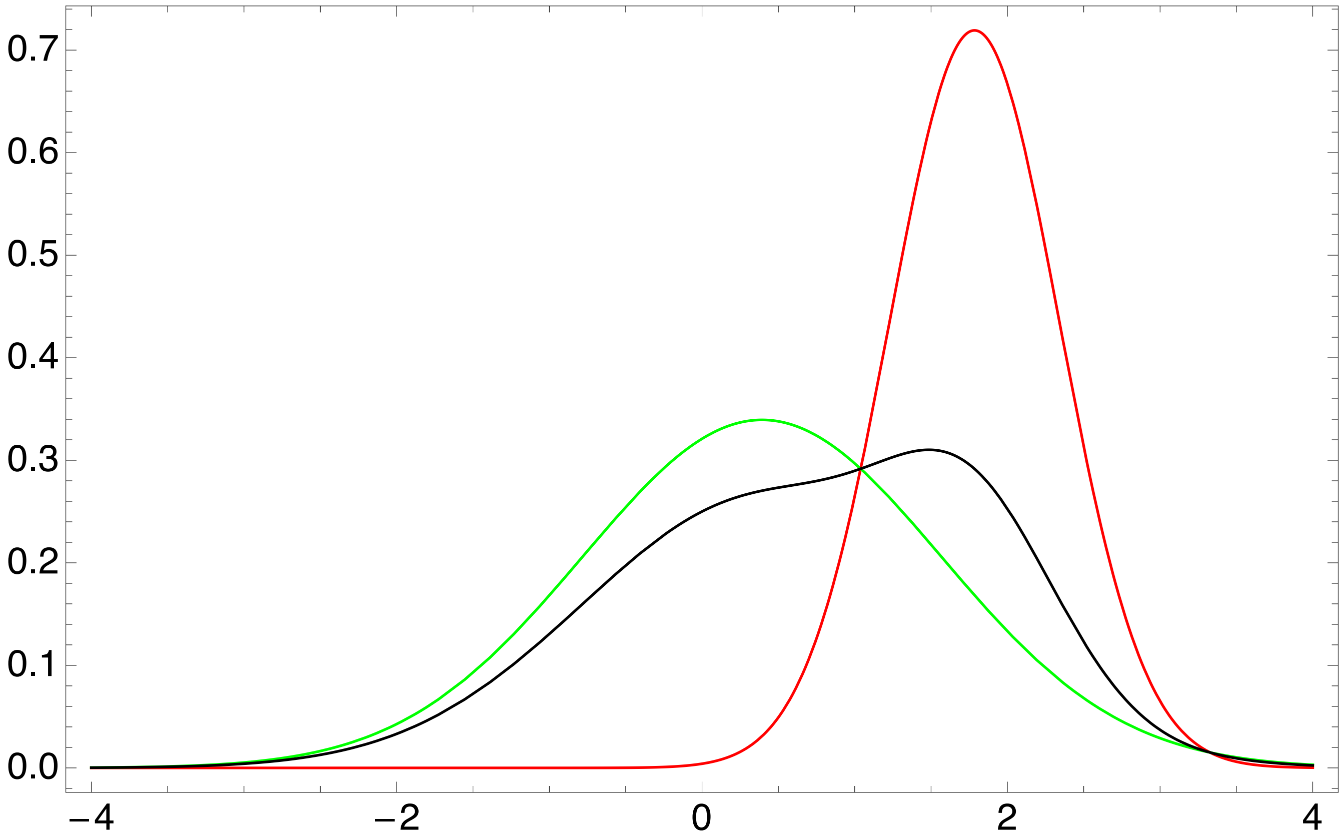
- The evidence requires computing an integral over the likelihood:

$$\mathrm{pr}(M|D) \; = \int d\mathbf{a} \; \frac{\mathrm{pr}(D|\mathbf{a}, M)\mathrm{pr}(\mathbf{a}|M)\mathrm{pr}(M)}{\mathrm{pr}(D)}.$$

- Various methods on the market for estimation of this integral, for different problems. Suggestions welcome! (An MCMC approach which could re-use samples drawn for doing the fit in the first place would be particularly nice…)



	$m = 0$	$m = 1$	marginalized
a_0	1.78(55)	0.39(1.18)	0.70(1.22)
a_1	—	0.75	—
$\mathrm{pr}(M\|D)$	0.2238	0.7762	—
χ^2_{aug}	2.02	0.23	—
p-val	0.365776	0.63424	—



Caveats

- Combining things in the first place assumes our basic statistical picture is right! As formulated here can't deal with autocorrelations (evidence factors mis-estimated). Not clear what happens yet if we try to include a deliberately pathological model in our set…

- Can we use this approach to deal with multi-exponential fits, too? Marginalization over cutoffs on data like t_{min}, t_{max}? (Important to think about for e.g. chiral perturbation theory fits with different mass ranges, for example.)