Data Analysis 1

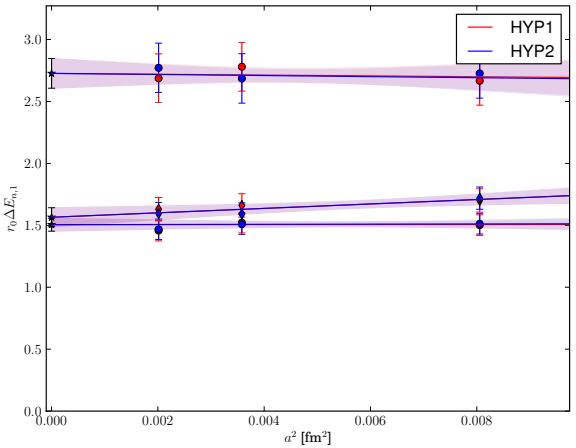
Georg von Hippel

Institut für Kernphysik,
Johannes-Gutenberg-Universität Mainz

Lattice Practices 2011
DESY Zeuthen, 9-11 March 2011

Introduction

After running a simulation you have bits and bytes on disk.

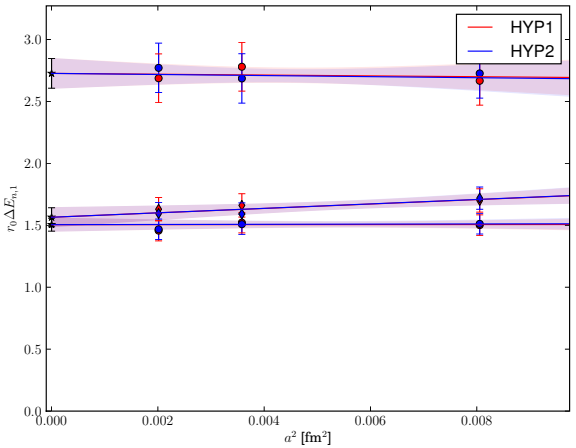


You want predictions for physical quantities (with errors!).

Introduction

After running a simulation you have bits and bytes on disk.

Data analysis bridges the gap.



You want predictions for physical quantities (with errors!).

Two important tasks:

- Reliably estimate size of statistical errors
  - Correlations between observables
  - Autocorrelations between configurations
- Systematically control systematic errors
  - Lattice spacing
  - Finite volume
  - Unphysical pion mass
  - Excited state contaminations

Two important tasks:

- Reliably estimate size of statistical errors: data analysis
  - Correlations between observables
  - Autocorrelations between configurations
- Systematically control systematic errors
  - Lattice spacing: improvement, continuum extrapolation
  - Finite volume: $M_\pi L \gtrsim 4$, different volumes
  - Unphysical pion mass: $M_\pi \leq M_K^{\text{phys}}$, χPT
  - Excited state contaminations

Two important tasks:

- Reliably estimate size of statistical errors: data analysis
  - Correlations between observables
  - Autocorrelations between configurations
- Systematically control systematic errors
  - Lattice spacing: improvement, continuum extrapolation
  - Finite volume: $M_\pi L \gtrsim 4$, different volumes
  - Unphysical pion mass: $M_\pi \leq M_K^{\mathrm{phys}}$, χPT
  - Excited state contaminations: data analysis!

Autocorrelations

Subsequent (in simulation time) measurements $\{\alpha_i\}$ are generally not fully independent

Autocorrelation function

$$C_\alpha(t) = \langle \alpha_{i+t}\alpha_i \rangle - \langle \alpha_{i+t} \rangle \langle \alpha_i \rangle$$

Normalised autocorrelation function

$$\Gamma_\alpha(t) = \frac{C_\alpha(t)}{C_\alpha(0)}$$

For large t,

$$\Gamma_\alpha(t) \sim \mathrm{e}^{-t/\tau_{\alpha,\mathrm{exp}}}$$

Exponential autocorrelation time

$$\tau_{\mathrm{exp}} = \sup_\alpha \tau_{\alpha,\mathrm{exp}}$$

Define estimators for mean and variance

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\alpha_i$$

$$\hat{\sigma}_\alpha^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\alpha_i - \hat{\alpha})^2$$

Error of estimated mean is given by

$$\begin{aligned}\sigma_{\hat{\alpha}}^2 &= \left\langle (\hat{\alpha} - \langle\alpha\rangle)^2 \right\rangle \\ &= \frac{1}{N^2}\left\langle \sum_{i,j=1}^{N}(\alpha_i - \langle\alpha\rangle)(\alpha_j - \langle\alpha\rangle) \right\rangle \\ &= \frac{1}{N}\langle\alpha^2\rangle - \langle\alpha\rangle^2 + \frac{1}{N^2}\sum_{i\neq j}\langle\alpha_i\alpha_j\rangle\end{aligned}$$

Without autocorrelations, $\langle\alpha_i\alpha_j\rangle = \langle\alpha_i\rangle\langle\alpha_j\rangle = \langle\alpha\rangle^2$, and hence

$$\sigma_{\hat{\alpha}}^2 = \frac{1}{N}\left(\langle\alpha^2\rangle - \langle\alpha\rangle^2\right) = \frac{\sigma_\alpha^2}{N} \approx \frac{\hat{\sigma}_\alpha^2}{N}$$

For data with autocorrelations,

$$\begin{aligned}
\sigma^2_{\hat{\alpha}} &= \frac{1}{N^2} \left\langle \sum_{i,j=1}^{N} (\alpha_i - \langle\alpha\rangle)(\alpha_j - \langle\alpha\rangle) \right\rangle \\
&= \frac{1}{N^2} \sum_{i,j=1}^{N} C_\alpha(|i-j|) \\
&= \sum_{t=-N}^{N} \frac{N-|t|}{N^2} C_\alpha(|t|) \\
&= \frac{C_\alpha(0)}{N} \sum_{t=-N}^{N} \Gamma_\alpha(|t|) \left(1 - \frac{|t|}{N}\right) \\
&\approx \frac{\sigma^2_\alpha}{N} 2\tau_{\alpha,\mathrm{int}}
\end{aligned}$$

with the integrated autocorrelation time

$$\tau_{\alpha,\mathrm{int}} = \frac{1}{2} + \sum_{t=1}^{N} \Gamma_\alpha(t)$$

Binning data

Computing τ_{int} accurately can be difficult. For more details, see

1. U. Wolff, Monte Carlo errors with less errors, Comput.Phys.Commun. 156:143-153,2004; Erratum-ibid.176:383,2007 [hep-lat/0306017].
2. S. Schaefer, R. Sommer, F. Virotta, Critical slowing down and error analysis in lattice QCD simulations, Nucl.Phys. B845:93-119,2011; [arXiv:1009.5228].

Often, it is sufficient to consider blocked data

$$\beta_k = \frac{1}{W} \sum_{i=(k-1)W+1}^{kW} \alpha_i$$

Then the estimated variance rises to the true variance of the mean as $W \to \infty$ like

$$\hat{\sigma}_\beta = \sigma_{\hat{\alpha}} - \frac{\delta}{W}$$

If $W \gg \tau_{\mathrm{int}}$, the correlations between the blocks can be neglected.

Correlations between observables

Multiple observables α_k, multiple measurements $\{\alpha_{kn}\}$
Estimator for the covariance matrix

$$\hat{C}_{kl} = \frac{1}{N(N-1)} \sum_{n=1}^{N} (\alpha_{kn} - \hat{\alpha}_k)(\alpha_{ln} - \hat{\alpha}_l)$$

Diagonal elements

$$C_{kk} = \sigma^2_{\hat{\alpha}_k}$$

Off-diagonal elements contain correlations between different α_k

Correlated χ^2

Let expectation value be a function of $P < K$ parameters $\{a_p\}$,

$$\langle \alpha_k \rangle = f_k(\{a_p\})$$

Assuming means $\hat{\alpha}_k$ to follow Gaussian distribution

$$\begin{aligned} P(\hat{\alpha}_k) &\propto \exp\left(-\tfrac{1}{2} \sum_{k,l=1}^{K} (\hat{\alpha}_k - \langle \alpha_k \rangle)[C^{-1}]_{kl}(\hat{\alpha}_l - \langle \alpha_l \rangle) \right) \\ &= \exp\left(-\tfrac{1}{2} \sum_{k,l=1}^{K} (\hat{\alpha}_k - f_k(\{a_p\}))[C^{-1}]_{kl}(\hat{\alpha}_l - f_l(\{a_p\})) \right) \end{aligned}$$

Maximise probability by minimising

$$\chi^2(\{a_p\})) = \sum_{k,l=1}^{K} (\hat{\alpha}_k - f_k(\{a_p\}))[C^{-1}]_{kl}(\hat{\alpha}_l - f_l(\{a_p\}))$$

Problems with correlated fits

The covariance matrix is often poorly determined by the data and may be numerically singular
The smallest and least well-determined eigenmodes have the largest influence on χ^2
Possible cure:

- Compute SVD of $C = SDT$
- Omit singular values below some cut-off

$$C^{-1} \approx T^\dagger \left[D_{ii}^{-1} \theta(D_{ii} - \lambda) \right] S^\dagger$$

when computing χ^2

Resampling techniques

Crucial idea: The best estimate you have of the actual distribution of the observables and their correlations is given by the data you have measured.

Resampling: Sample from this measured distribution to estimate the (co-)variances, and thus the statistical errors

The Bootstrap

Take this idea seriously to "pull yourself up by your bootstraps"

The Bootstrap

- Given: N measurements $\{\alpha_i\}$ of observable α
- Form B synthetic data sets S_k by selecting N measurements (with repetitions allowed) for each
- Compute

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\alpha_i \qquad \alpha_k^{\mathrm{b}} = \frac{1}{N}\sum_{i\in S_k}\alpha_i \;\;, k = 1,\ldots,B$$

- Compute for $\theta = f(\alpha)$, $\hat{\theta} = f(\hat{\alpha})$

$$\tilde{\theta} = \frac{1}{B}\sum_{k=1}^{B} f(\alpha_k^{\mathrm{b}}) \qquad \sigma_{\tilde{\theta}}^2 = \frac{1}{B}\sum_{k=1}^{B}\left(f(\alpha_k^{\mathrm{b}}) - \tilde{\theta}\right)^2$$

- $\sigma_{\tilde{\theta}}$ is an estimate of the statistical error
- $\tilde{\theta} - \hat{\theta}$ is an estimate of the bias

The Jackknife

- Given: N measurements $\{\alpha_i\}$ of observable α
- Form N synthetic data sets S_k by removing the k^{th} measurement from S_k
- Compute

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\alpha_i \qquad \alpha_k^{\mathrm{J}} = \frac{1}{N-1}\sum_{i\in S_k}\alpha_i \quad , k = 1,\ldots,N$$

- Compute for $\theta = f(\alpha)$, $\hat{\theta} = f(\hat{\alpha})$

$$\tilde{\theta} = \frac{1}{N}\sum_{k=1}^{N} f(\alpha_k^{\mathrm{J}}) \qquad \sigma_{\tilde{\theta}}^2 = \frac{N-1}{N}\sum_{k=1}^{N}\left(f(\alpha_k^{\mathrm{J}}) - \hat{\theta}\right)^2$$

- $\sigma_{\tilde{\theta}}$ is an estimate of the statistical error
- $(N-1)(\tilde{\theta} - \hat{\theta})$ is an estimate of the bias

Implementation notes

- Bootstrap and Jackknife are very similar – can be implemented as subclasses of a common superclass
- Bootstrap errors can also be estimated from percentiles
- Cheaper to generate Jackknife sample means from

$$\alpha_k^{\mathrm{J}} = \frac{1}{N-1} \sum_{i \neq k} \alpha_i = \frac{1}{N-1} \left(N\hat{\alpha} - \alpha_k \right)$$

- Built-in sum(), scipy.mean(), scipy.std() are significantly faster than manual loops

Practice Problems

Let's practice!

Data Analysis 2

Georg von Hippel

Institut für Kernphysik,
Johannes-Gutenberg-Universität Mainz

Lattice Practices 2011
DESY Zeuthen, 9-11 March 2011

The trouble with excited states

Spectral representation of a correlator (infinite time extent)

$$C(t) = \langle O(t)O(0)\rangle = \sum_{n=1}^{\infty} \mathrm{e}^{-E_n t} \left|\psi_n\right|^2 \qquad \psi_n = \langle n|\hat{O}|0\rangle\,, \quad E_n \leq E_{n+1}$$

Effective mass is contaminated by excitations

$$m_{\mathrm{eff}}(t) = \log \frac{C(t)}{C(t+1)} = E_1 + A\mathrm{e}^{-(E_2-E_1)t} + \ldots$$

- Systematic error at small t vs statistical errors at large t
- Excited states themselves may be of interest
- Multi-exponential fits tend to be ill-conditioned

The Bayesian perspective

Bayes' theorem:

$$P(A|B)P(B) = P(B|A)P(A)$$

Bayesian interpretation:

Probabilities are degrees of belief

Rearrange to get the formula that tells you how to update your beliefs given new data:

$$P(\{a_p\}|\{\hat{\alpha}_k\}) = \frac{P(\{\hat{\alpha}_k\}|\{a_p\})P(\{a_p\})}{P(\{\hat{\alpha}_k\})}$$

where the "prior" is $P(\{a_p\})$.

Bayesian fits

With

$$P(\{\hat{\alpha}_k\}|\{a_p\}) \propto \exp\left(-\tfrac{1}{2}\sum_{k,l=1}^{K}(\hat{\alpha}_k - f_k(\{a_p\}))[C^{-1}]_{kl}(\hat{\alpha}_l - f_l(\{a_p\}))\right)$$

and the mutually independent priors

$$P(a_p) \propto \exp\left(-\frac{(a_p-\bar{a}_p)^2}{2\sigma_p^2}\right)$$

we can maximise $P(\{a_p\}|\{\hat{\alpha}_k\})$ by minimising

$$\chi^2_{\text{aug}}(\{a_p\}) = \sum_{k,l=1}^{K}(\hat{\alpha}_k - f_k(\{a_p\}))[C^{-1}]_{kl}(\hat{\alpha}_l - f_l(\{a_p\})) + \sum_{p=1}^{P}\frac{(a_p-\bar{a}_p)^2}{\sigma_p^2}$$

where we may now have $P > K$.
If the fitted values are largely independent of the priors, we may take them as having been determined by the data.
Otherwise, GIGO ...

Other χ^2-based methods

Other χ^2-based proposals include

- evolutionary algorithms
- sequential empirical Bayes method (SEBM)

Limitations of χ^2-based methods:

- Resolution of near-degenerate states
- Choice of fitting range

The Generalised Eigenvalue Problem

Measure a matrix of correlation functions

$$C_{ij}(t) = \langle O_i(t) O_j(0) \rangle = \sum_{n=1}^{\infty} \mathrm{e}^{-E_n t} \psi_{ni} \psi_{nj} \,, \quad i,j = 1,\ldots,N$$

$$\psi_{ni} \equiv (\psi_n)_i = \langle n | \hat{O}_i | 0 \rangle = \psi^*_{ni} \quad E_n \leq E_{n+1}$$

and solve the GEVP(s)

$$C(t)\, v_n(t,t_0) = \lambda_n(t,t_0)\, C(t_0)\, v_n(t,t_0) \,, \quad n = 1,\ldots,N \quad t > t_0,$$

Define effective energy levels and creation operators [Blossier, GvH et al.,2008]

$$E_n^{\mathrm{eff}} = \frac{1}{a} \log \frac{\lambda_n(t,t_0)}{\lambda_n(t+a,t_0)}$$

$$\hat{\mathcal{A}}_n^{\mathrm{eff}}(t,t_0) = \mathrm{e}^{-\hat{H}t} \frac{(\hat{O}\,,\, v_n(t,t_0))}{(v_n(t,t_0)\,,\, C(t)\, v_n(t,t_0))^{-1/2}} \frac{\lambda_n(t_0+t/2,t_0)}{\lambda_n(t_0+t,t_0)}$$

such that

$$E_n^{\mathrm{eff}} = E_n + \varepsilon_n(t,t_0) \qquad \hat{\mathcal{A}}_n^{\mathrm{eff}\,\dagger}|0\rangle = |n\rangle + \sum_{n'=1}^{\infty} \pi_{nn'}(t,t_0)|n'\rangle$$

The GEVP simplified

(Theoretically) split C_{ij} into first N states and the rest

$$C_{ij}^{(0)}(t) = \sum_{n=1}^{N} \mathrm{e}^{-E_n t} \psi_{ni} \psi_{nj} \,, \qquad C_{ij}^{(1)}(t) = \sum_{n=N+1}^{\infty} \mathrm{e}^{-E_n t} \psi_{ni} \psi_{nj}$$

The (time-independent) dual vectors are defined by

$$(u_n, \psi_m) = \delta_{mn} \,, \;\; m, n \leq N \,. \qquad (u_n, \psi_m) \equiv \sum_{i=1}^{N} (u_n)_i \, \psi_{mi}$$

One then has

$$\begin{aligned} C^{(0)}(t) u_n &= \mathrm{e}^{-E_n t} \psi_n \,, \\ C^{(0)}(t) \, u_n &= \lambda_n^{(0)}(t, t_0) \, C^{(0)}(t_0) \, u_n, \\ \lambda_n^{(0)}(t, t_0) &= \mathrm{e}^{-E_n (t - t_0)} \,, \quad v_n(t, t_0) \propto u_n \end{aligned}$$

and an orthogonality relation valid at all t

$$(u_m, C^{(0)}(t) \, u_n) = \delta_{mn} \, \rho_n(t) \,, \quad \rho_n(t) = \mathrm{e}^{-E_n t} \,.$$

The GEVP simplified

The operators

$$\hat{\mathcal{A}}_n = \sum_{i=1}^{N} (u_n)_i \hat{O}_i \equiv (\hat{O}, u_n)\,,$$

create the eigenstates of the Hamilton operator

$$|n\rangle = \hat{\mathcal{A}}_n |0\rangle\,, \hat{H}|n\rangle = E_n\,|n\rangle\,.$$

So arbitrary matrix elements can be written as

$$p_{0n} \quad = \quad \langle 0|\hat{P}|n\rangle = \langle 0|\hat{P}\hat{\mathcal{A}}_n|0\rangle$$

generalization:

$$\begin{aligned} p_{0n} \quad &= \quad \langle P(t) O_j(0)\rangle (u_n)_j = \frac{\langle P(t)\mathcal{A}_n(0)\rangle}{\langle \mathcal{A}_n(t)\mathcal{A}_n(0)\rangle^{1/2}} \mathrm{e}^{E_n\, t/2} \\ &= \quad \frac{\langle P(t) O_j(0)\rangle\, v_n(t,t_0)_j}{\left(v_n(t,t_0)\,,\; C(t)\, v_n(t,t_0)\right)^{1/2}} \frac{\lambda_n(t_0+t/2,t_0)}{\lambda_n(t_0+t,t_0)} \end{aligned}$$

Perturbation theory for the GEVP

Following [Niedermayer & Weisz, 1998, unpublished], set up a perturbative expansion for the GEVP as

$$A v_n = \lambda_n B v_n \,, \quad A = A^{(0)} + \epsilon A^{(1)} \,, \quad B = B^{(0)} + \epsilon B^{(1)} \,.$$

$$(v_n^{(0)}, B^{(0)} v_m^{(0)}) \quad = \quad \rho_n \, \delta_{nm} \,.$$

$$\begin{aligned} \lambda_n &= \lambda_n^{(0)} + \epsilon \lambda_n^{(1)} + \epsilon^2 \lambda_n^{(2)} \ \ldots \\ v_n &= v_n^{(0)} + \epsilon v_n^{(1)} + \epsilon^2 v_n^{(2)} \ \ldots \end{aligned}$$

We will later set

$$\begin{aligned} A^{(0)} &= C^{(0)}(t) \,, \quad \epsilon A^{(1)} = C^{(1)}(t) \,, \\ B^{(0)} &= C^{(0)}(t_0) \,, \quad \epsilon B^{(1)} = C^{(1)}(t_0) \end{aligned}$$

Perturbation theory for the GEVP

To second order

$$A^{(0)} v_n^{(1)} + A^{(1)} v_n^{(0)} = \lambda_n^{(0)} \left[B^{(0)} v_n^{(1)} + B^{(1)} v_n^{(0)} \right] + \lambda_n^{(1)} B^{(0)} v_n^{(0)} ,$$

$$A^{(0)} v_n^{(2)} + A^{(1)} v_n^{(1)} = \lambda_n^{(0)} \left[B^{(0)} v_n^{(2)} + B^{(1)} v_n^{(1)} \right] + \lambda_n^{(1)} \left[B^{(0)} v_n^{(1)} + B^{(1)} v_n^{(0)} \right] + \lambda_n^{(2)} B^{(0)} v_n^{(0)} .$$

Solve using orthogonality $(v_n^{(0)}, B^{(0)} v_m^{(0)}) = \delta_{mn} \rho_n$

$$\lambda_n^{(1)} = \rho_n^{-1} \left(v_n^{(0)}, \Delta_n v_n^{(0)} \right) , \quad \Delta_n \equiv A^{(1)} - \lambda_n^{(0)} B^{(1)}$$

$$v_n^{(1)} = \sum_{m \neq n} \alpha_{nm}^{(1)} \rho_m^{-1/2} v_m^{(0)} , \quad \alpha_{nm}^{(1)} = \rho_m^{-1/2} \frac{\left(v_m^{(0)}, \Delta_n v_n^{(0)} \right)}{\lambda_n^{(0)} - \lambda_m^{(0)}}$$

$$\lambda_n^{(2)} = \sum_{m \neq n} \rho_n^{-1} \rho_m^{-1} \frac{\left(v_m^{(0)}, \Delta_n v_n^{(0)} \right)^2}{\lambda_n^{(0)} - \lambda_m^{(0)}} - \rho_n^{-2} \left(v_n^{(0)}, \Delta_n v_n^{(0)} \right) \left(v_n^{(0)}, B^{(1)} v_n^{(0)} \right) .$$

Also get all-orders recursion formula for the higher-order coefficients.

Perturbation theory for the GEVP

Inserting the specific case of a correlator matrix and using (for $m > n$)

$$\begin{aligned}(\lambda_n^{(0)} - \lambda_m^{(0)})^{-1} &= (\lambda_n^{(0)})^{-1}(1 - \mathrm{e}^{-(E_m - E_n)(t-t_0)})^{-1} \\ &= (\lambda_n^{(0)})^{-1} \sum_{k=0}^{\infty} \mathrm{e}^{-k(E_m - E_n)(t-t_0)}\end{aligned}$$

find

$$\begin{aligned}\varepsilon_n(t, t_0) &= \mathrm{O}(\mathrm{e}^{-\Delta E_{N+1,n}\, t}), \quad \Delta E_{m,n} = E_m - E_n\,, \\ \pi_{nn'}(t, t_0) &= \mathrm{O}(\mathrm{e}^{-\Delta E_{N+1,n}\, t_0}), \quad \text{at fixed } t - t_0\end{aligned}$$

to all orders in the perturbative expansion, giving efficient suppression of excited state contributions for large enough N.

Practice Problems

Let's practice!

Summary and Outlook

Basic toolkit to deal with

- Statistical errors:
  - Autocorrelations: blocking, τ_{int}
  - Correlations between observables: resampling, correlated χ^2
- Excited state contaminations: GEVP, Bayesian methods

More sophisticated versions

- double jackknife for correlated fits
- using estimated τ_{exp} to estimate τ_{int}
- optimising the GEVP by pruning
- fitting to the GEVP
- ...

The end

Thank you for your attention

. . . and have fun with real data!