# Bayesian curve fitting for lattice gauge theorists

**Colin Morningstar**

**Carnegie Mellon University**

**Cairns, Australia, July 2001**

# Overview

- “New” approach to extracting physical observables from stochastically-determined correlation functions
  - Use of Bayesian statistical inference
- Advantages of new approach:
  - Uses all of your simulation data
  - Incorporates uncertainties associated with choice of fit ranges and systematic effects
  - Incorporates prior knowledge from physical constraints and other calculations
- Work in progress
  - In collaboration with P. Lepage, P. Mackenzie, K. Hornbostel, and B. Clark
  - Still struggling with integration techniques

# Outline

- Standard analysis applied to three example correlation functions
  - Few parameters
  - Temporal range limited – information discarded
- Exploiting information in correlation functions at small temporal separations
  - More parameters
  - Instabilities
- Bayesian statistics: overcoming the instabilities
  - Bayes' theorem
  - Bayesian regression and parameter estimation
  - Choosing the *prior*
  - Integration challenges
- Conclusions and outlook

# Extracting physics

- Correlation functions *C(t)* computed directly in Monte Carlo simulations
- Physical observables (masses, couplings, matrix elements) must be inferred from *C(t)*
  - Energies $E_n$ from single correlator (neglecting boundaries)

$$C(t) = \sum_{n=0}^{\infty} A_n \exp(-E_n t), \qquad 0 \leq A_n \leq 1, \quad E_{n+1} > E_n$$

  - Extrapolations to continuum or chiral limit
- Maximize likelihood:

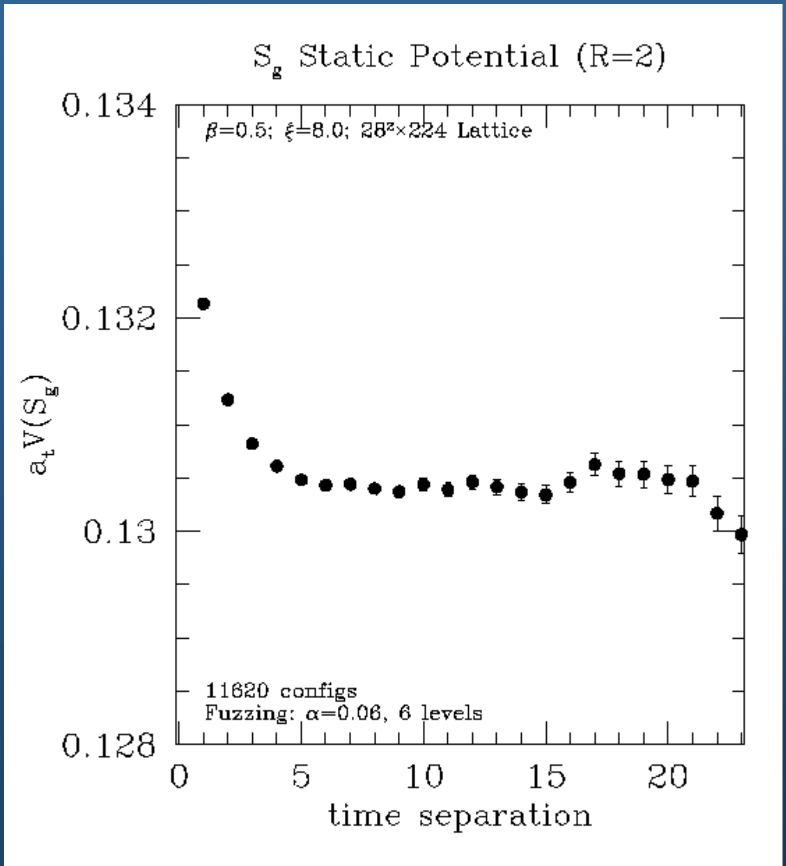
$$P(D \mid M \cap I) \propto \exp(-\chi^2)$$

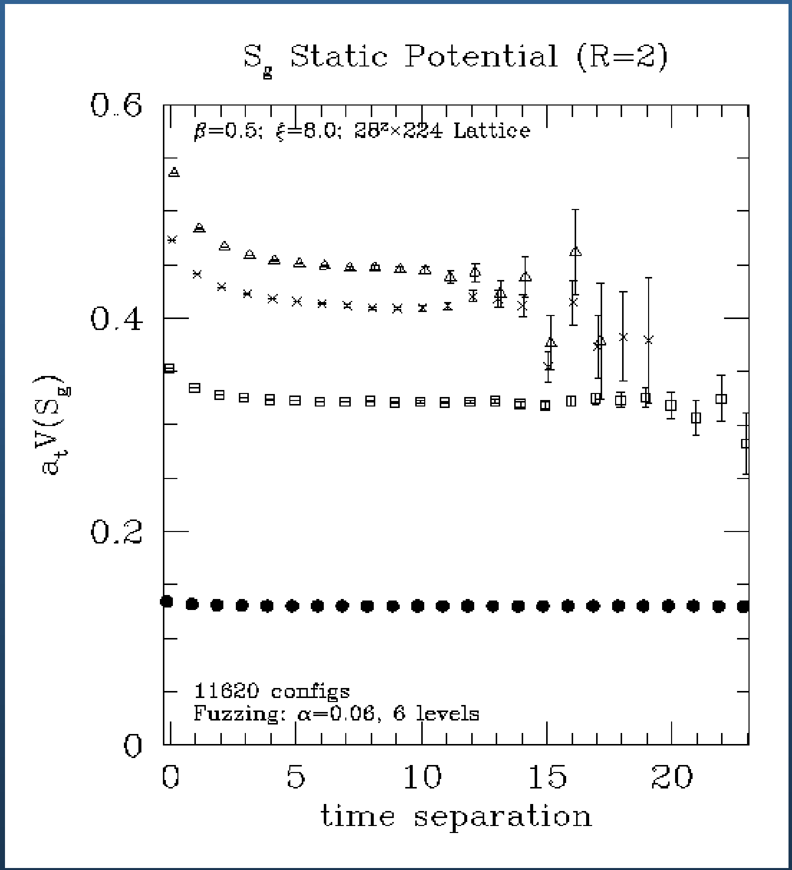
$$\chi^2 = \sum_{ij} [M_i(u) - d_i] C_{ij}^{-1} [M_j(u) - d_j]$$

- Illustrate method using three examples

# Example 1 -- excellent plateaus

Effective masses for static potential in compact U(1) in 2+1 dimensions



# Example 1 – standard analysis

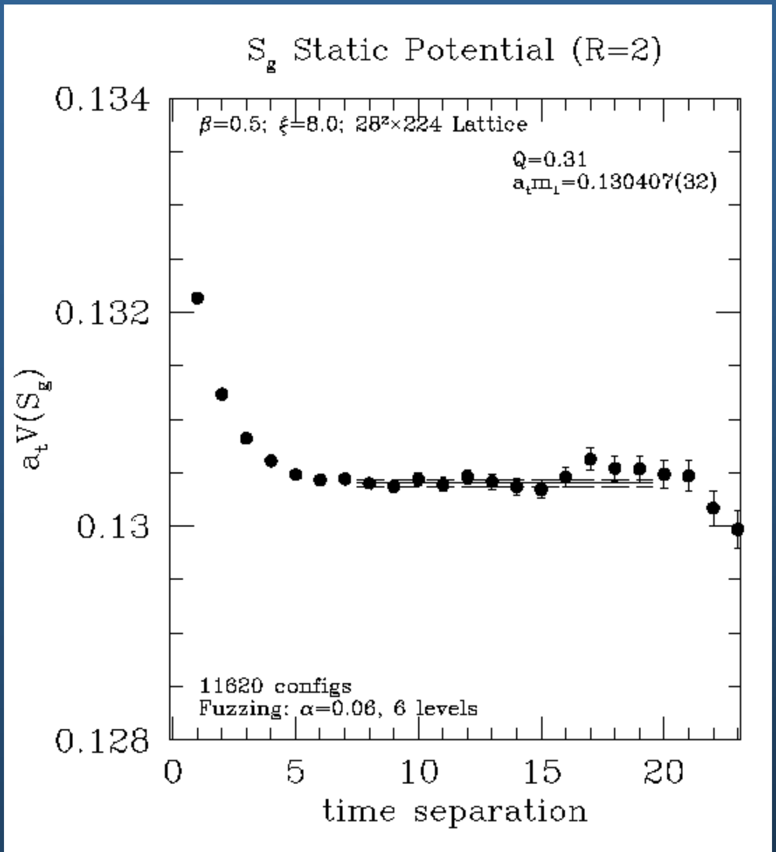
- Allow temporal separation *t* to increase to *tmin* where excited-state contamination is sufficiently suppressed
- Simulation data for *t*<*tmin* discarded
- Fit *C(t)* to single exponential in range *tmin* to *tmax*
- For *N* correlators, do simultaneous fits to *N* functions, each a sum of *N* exponentials

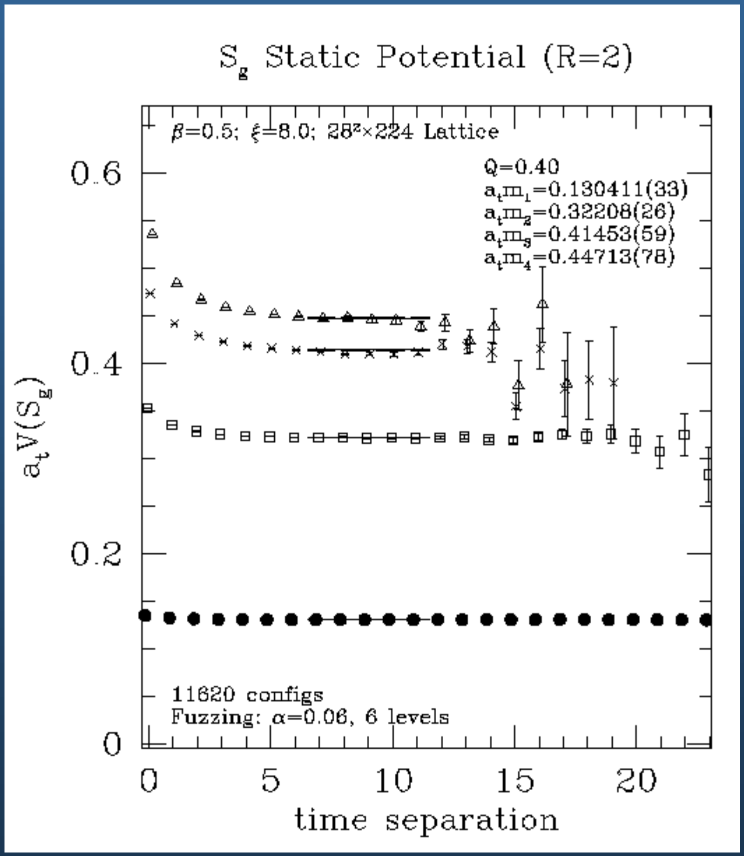
$$C_i(t) = \sum_{n=0}^{N} A_n^{(i)} \exp(-E_n t), \qquad \begin{array}{l} 0 \le A_n^{(i)} \le 1, \\ E_{n+1} > E_n \end{array}$$

- For excellent data
  - choice of *tmin* and *tmax* easily guided by fit quality *Q*
  - results, including bootstrap error, reasonably insensitive to minor changes in fit range

# Example 1 – standard analysis (continued)

- Typical fit results



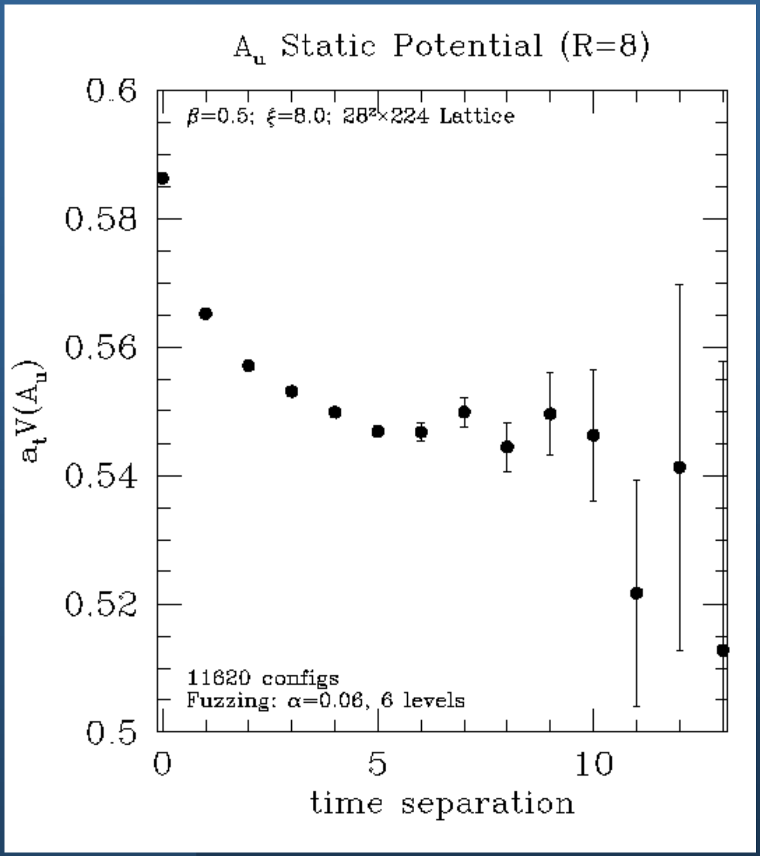
# Example 1 – standard analysis (continued)

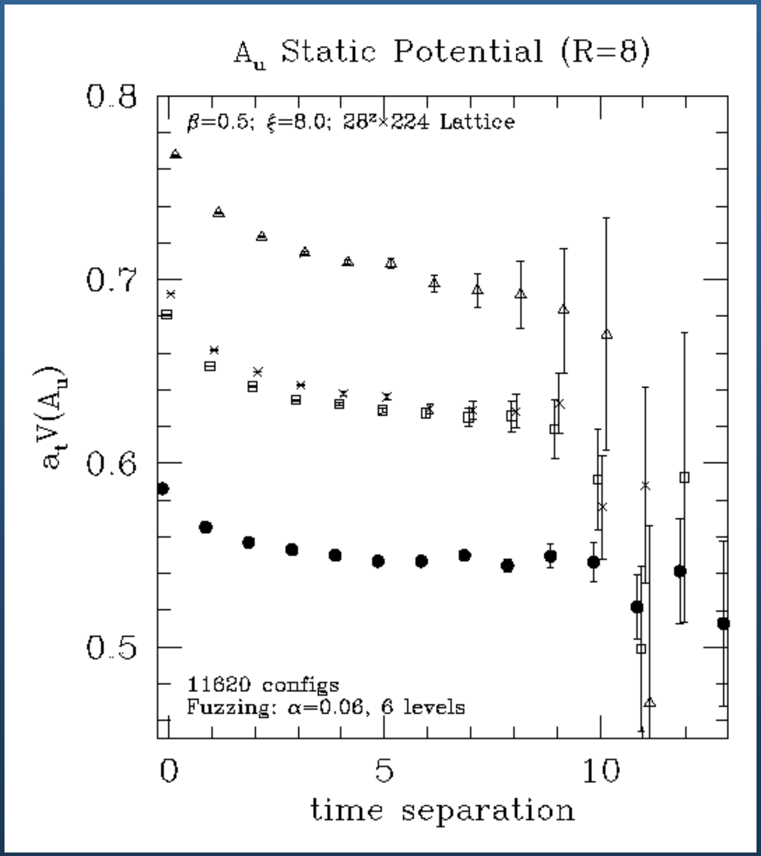
- Fits to single correlator

| *tmin* | *tmax* | *Q* | $E_0$ |
|---|---|---|---|
| 2 | 20 | 0 | 0.130907(25) |
| 3 | 20 | 0 | 0.130659(22) |
| 4 | 20 | 0 | 0.130530(27) |
| 5 | 20 | 0.06 | 0.130469(23) |
| 6 | 20 | 0.16 | 0.130445(29) |
| 7 | 20 | 0.17 | 0.130431(30) |
| 8 | 20 | 0.31 | 0.130407(32) |
| 9 | 20 | 0.24 | 0.130406(36) |
| 10 | 20 | 0.39 | 0.130431(40) |

# Example 2 – fair plateaus

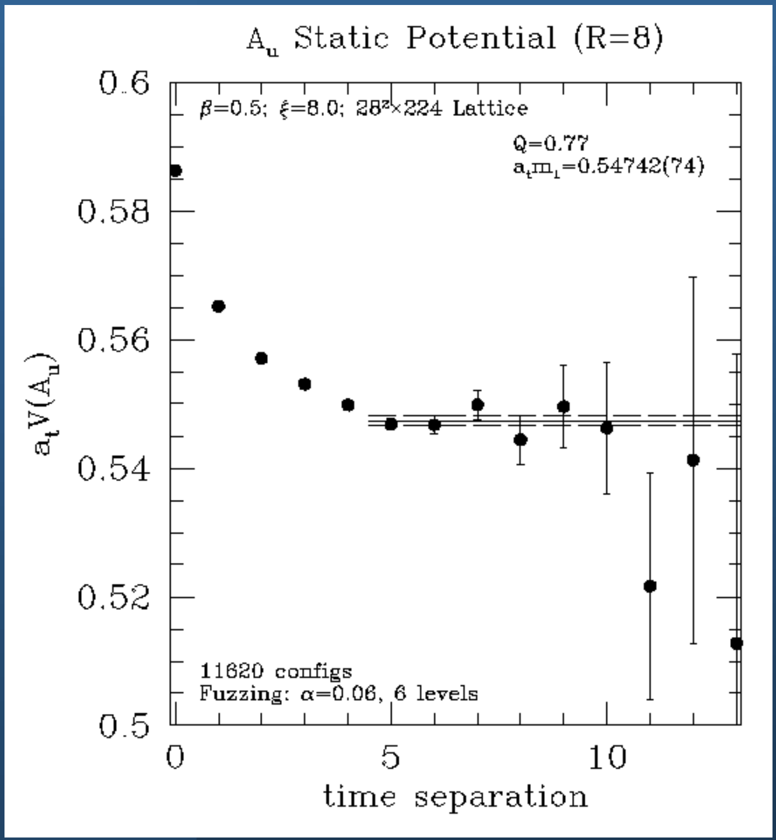
Effective masses for static potential in compact U(1) in 2+1 dimensions

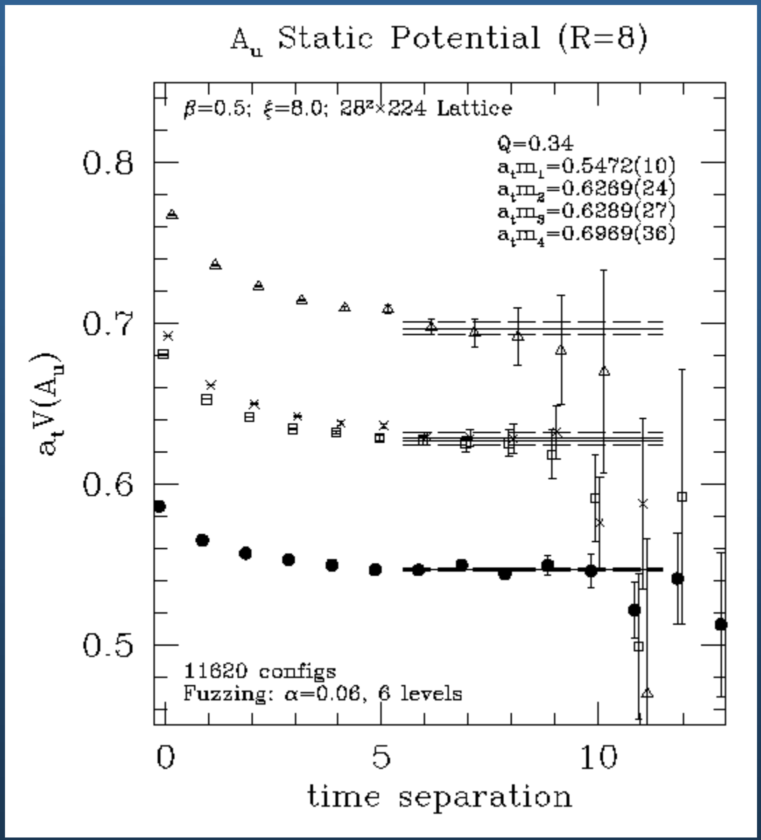


# Example 2 – standard analysis

- Typical fits for single correlator



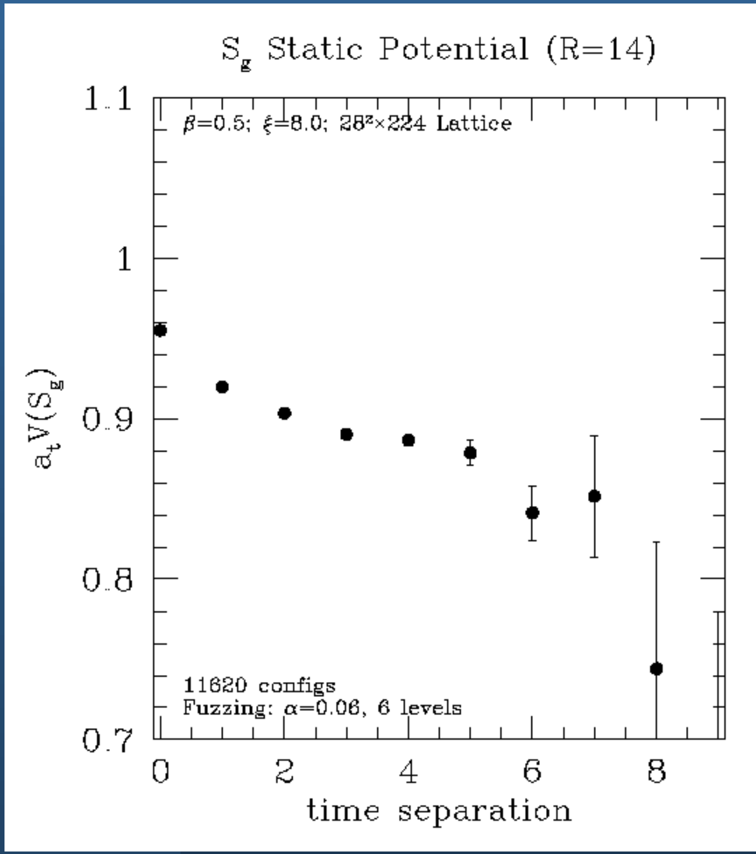
# Example 2 – standard analysis (continued)

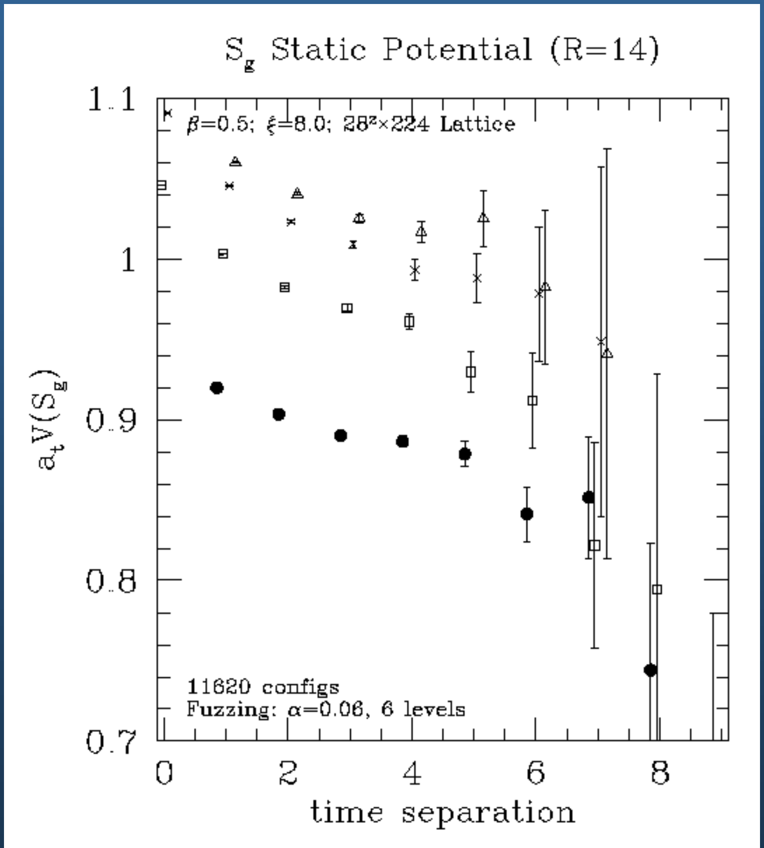
- Fits to single correlator

| *tmin* | *tmax* | *Q* | $E_0$ |
|---|---|---|---|
| 3 | 14 | 0 | 0.55203(27) |
| 4 | 14 | 0.05 | 0.54936(41) |
| 5 | 14 | 0.77 | 0.54742(74) |
| 6 | 14 | 0.67 | 0.5475(12) |
| 7 | 14 | 0.65 | 0.5489(22) |
| 8 | 14 | 0.76 | 0.5456(33) |

# Example 3 – questionable plateaus

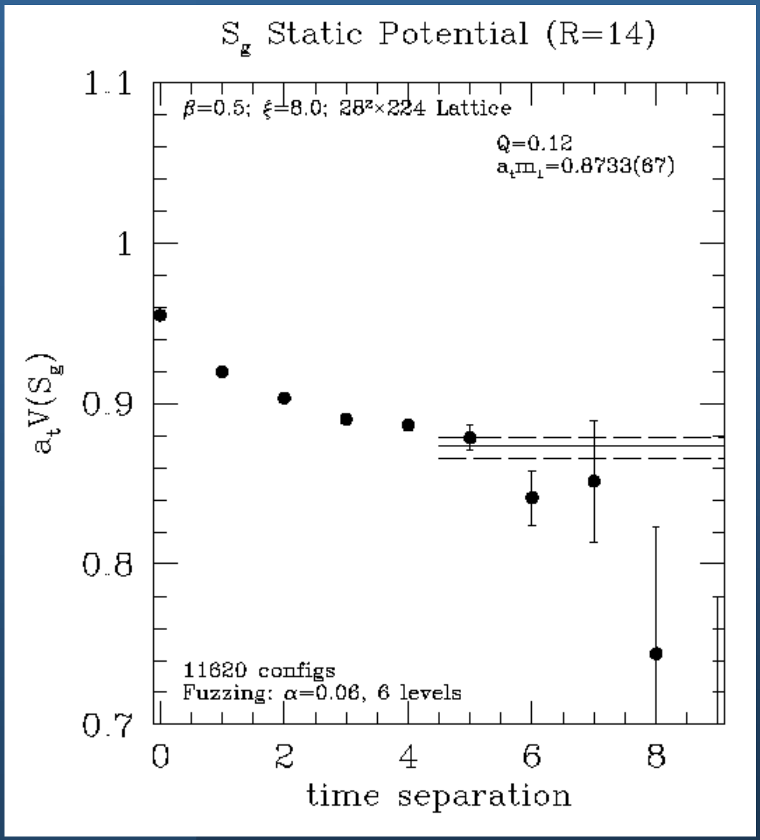
Effective masses for static potential in compact U(1) in 2+1 dimensions

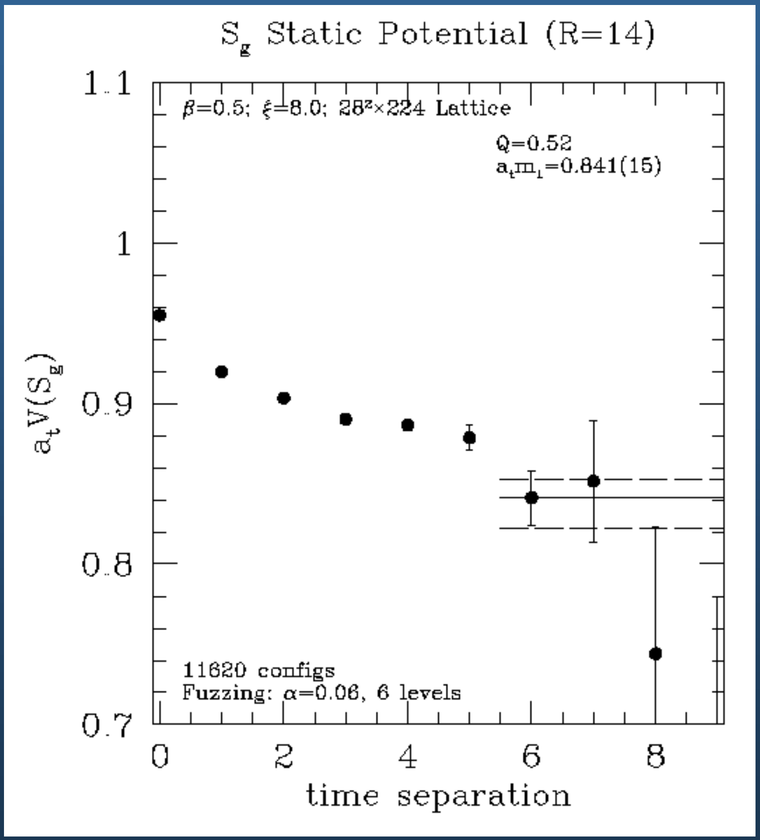


# Example 3 – standard analysis

- Which fit should I choose?



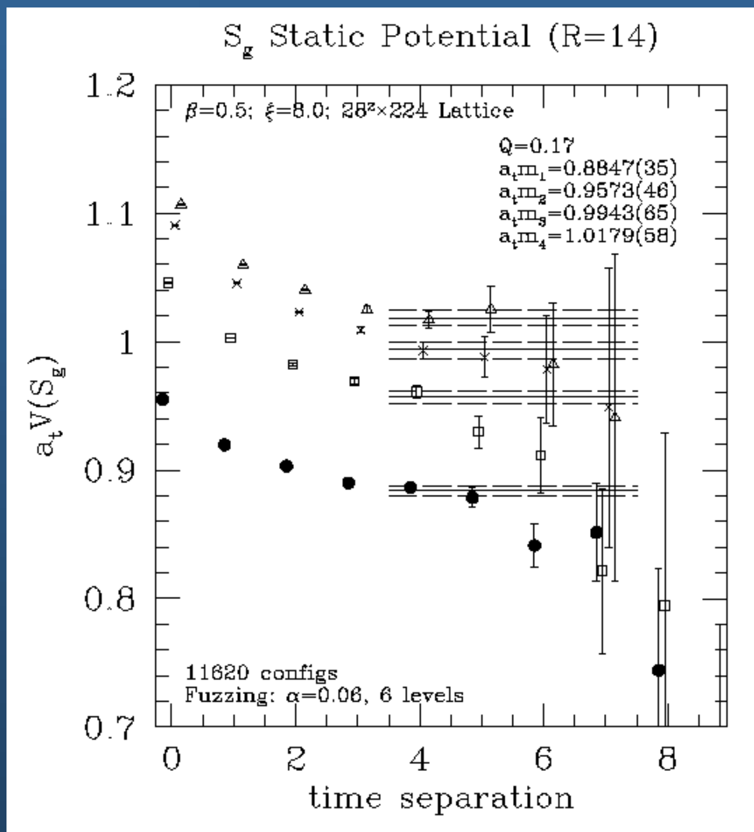
# Example 3 – standard analysis (continued)

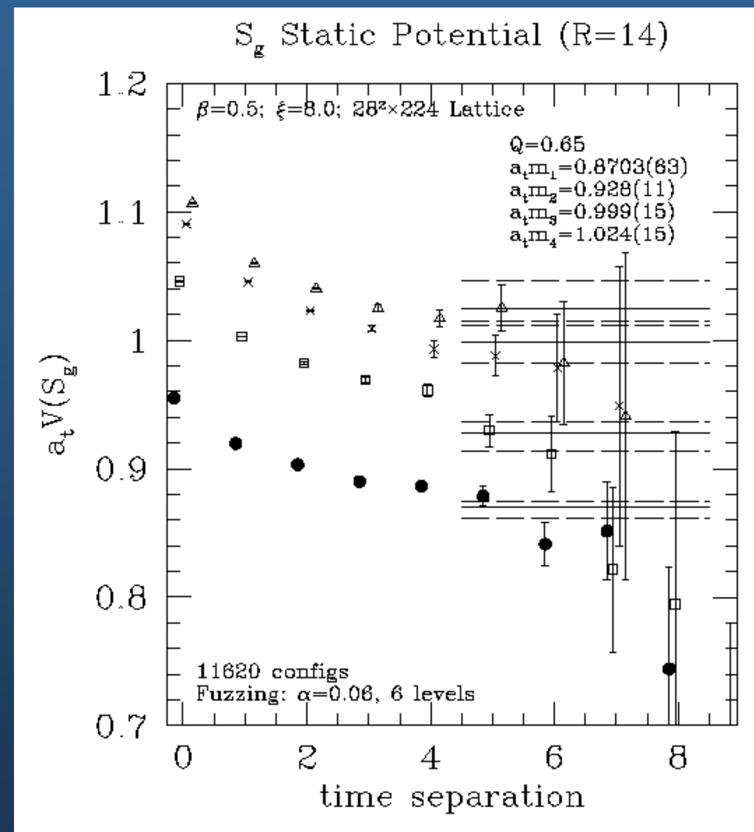
- Fits to single correlator

| *tmin* | *tmax* | *Q* | $E_0$ |
|---|---|---|---|
| 1 | 10 | 0 | 0.91630(39) |
| 2 | 10 | 0 | 0.90147(74) |
| 3 | 10 | 0.03 | 0.8902(13) |
| 4 | 10 | 0.06 | 0.8853(31) |
| 5 | 10 | 0.12 | 0.8733(67) |
| 6 | 10 | 0.52 | 0.841(15) |
| 7 | 10 | 0.32 | 0.838(37) |

# Example 3 – standard analysis (continued)

- Simultaneous fits to several correlators – which to choose?



# Alternative approach

- Simulation data is expensive $\Longrightarrow$ discarding small time information painful
- Keep small time information $\Longrightarrow$ must retain many exponentials
- Problem:
  - Encounter fitting instabilities
  - Huge uncertainties in parameter estimates
- Source of the problem
  - Unconstrained fits allow physically insensible or impossible parameter values
- Solution of the problem
  - Introduction of constraints
- *Bayesian statistics* allows introduction of constraints in natural way
- Bayesian approach now widely used
  - economics, medical research, astrophysics, condensed matter physics,…

# The Reverend Thomas Bayes

# The Reverend Thomas Bayes

- Presbyterian minister – born 1702 London, England
  – died April 17, 1761 Tunbridge Wells
- His theory of probability described in
  - *"Essay towards solving a problem in the doctrine of chances"* published in 1763 in *Philosophical Transactions of the Royal Society*
  - Submitted posthumously by Richard Price
    - I now send you an essay which I have found among the papers of our deceased friend Mr Bayes, and which, in my opinion, has great merit... In an introduction which he has writ to this Essay, he says, that his design at first in thinking on the subject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circumstances, it has happened a certain number of times, and failed a certain other number of times.
- Theorem presented by Bayes was restricted to binomial distribution (but generality recognized by Bayes)
- Ideas in the theorem conceived by James Bernoulli in 1713
- Bayes' theorem generalized beyond binomial distribution by Laplace in 1774 (most likely independently)

# Frequentist versus Bayes

- Standard (frequentist) statistical methods were developed later than Bayesian methods
  - Linear regression – Francis Galton in late 1800's
  - Goodness of fit, correlation – Karl Pearson circa 1900
  - Field blossoms in roaring 1920's and during the Great Depression – Fisher, Neyman, Pearson
  - Flurry of research and applications during WWII
- Bayesian methods much older, but largely ignored (or actively opposed) until the 1950's
  - Championed by prominent non-statisticians, most notably physicist H. Jeffreys, economist A.Bowley
  - Popularity grows in 1970's with advent of computers
  - Beginning of the holy wars….

# Bayes' theorem

- $P(A)$ = probability of event $A$
- $P(A|B)$ = conditional probability of $B$ given $A$
- $P(A,B)$ = probability of both $A$ and $B$

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

- Rearrange to obtain Bayes' theorem:

$$P(A \mid B) = \frac{P(A) P(B \mid A)}{P(B)}$$

- Include event $C$ then theorem generalizes to

$$P(A \mid B \cap C) = \frac{P(A \mid B)\ P(C \mid B \cap A)}{P(C \mid B)}$$

- Applies to probability distributions also

# Bayesian regression

- Application of Bayes' theorem to curve fitting

| | |
|---|---|
| $D$ | data |
| $M$ | model |
| $I$ | prior information |

- Probability of *model parameters* given the *data* and *prior knowledge*

*posterior* *likelihood*

$$P(M|D \cap I) = \frac{P(D|M \cap I)\ P(M|I)}{P(D|I)}.$$

*marginal distribution of data given prior knowledge* *prior*

# Bayesian regression (continued)

- Alternative form

$$P(M|D \cap I) = \frac{P(D|M \cap I)\ P(M|I)}{\int dM\ \ P(D|M \cap I)\ P(M|I)}.$$

- Bayesian regression uses the *posterior distribution* for all statistical inference
- Estimate model parameters using your favorite *statistic* with the posterior distribution
  - Measures of central tendency: mode, mean, median
  - Measures of dispersion: variance, skewness, kurtosis
- Example: mean value and variance of a model parameter $u_j$

$$\begin{aligned} \langle u_j \rangle &= \int d\boldsymbol{u}\ u_j\ P(\ M(\boldsymbol{u})\,|D \cap I) \\ \mathrm{var}(u_j) &= \int d\boldsymbol{u}\ (u_j - \langle u_j \rangle)^2\ P(\ M(\boldsymbol{u})\,|D \cap I) \end{aligned}$$

# Likelihood

- The likelihood is the same as in the standard analysis

$$P(D|M \cap I) \propto \exp(-\chi^2/2)$$

$$\chi^2 = \sum_{ij} [M_i(u) - d_i]\; C_{ij}^{-1}\; [M_j(u) - d_j],$$

*model prediction*

*covariance matrix of data*

*data*

# The Bayesian prior

- Constraints are incorporated using the *prior* probability distribution
- Role of the prior
  - Use prior knowledge about the system to limit parameter search to the set of *feasible* solutions
  - Filters out the improbable solutions from the feasible solutions
- The prior incorporates information accumulated from
  - Past experience (previous experiments, calculations)
  - Opinions of subject-area experts
  - Theoretical constraints
- Considerations in prior construction
  - Computational ease
  - Symmetries, limiting cases
  - Avoid putting in more information than you truly know!
  - Results do and should depend on prior

# The Bayesian prior (continued)

- One common method of constructing a prior:
  - Let a monkey throw balls into bins! ⇨ maximum entropy
- True Bayesian approach:
  - Use past experience and your physical knowledge of the system
- Can you use the data?
  - Strictly speaking ⇨ no
  - Use a handful of bins to aid prior construction, then discard
  - Use of data – empirical Bayes method
- Prior: both an *opportunity* and a *nuisance*

# Our choice of prior for Examples 1—3

- Fit to $N_{cor}$ correlators using the model function $M_i(u)$

$$C_i(t) = \sum_{n=0}^{N_{\rm exp}-1} A_n^{(i)} \ \exp(-E_n t),$$

- To ensure positivity of the coefficients and to order the energies, use

$$\begin{aligned} A_n^{(i)} &= \left(b_n^{(i)}\right)^2 \\ E_n &= E_{n-1} + \epsilon_n^2 \end{aligned}$$

- Actual parameters are $E_0, \quad \epsilon_n, \quad b_n^{(i)}$
- Form of the prior $P(M|I) \propto \exp\left(-\sum_{\alpha=1}^{N_{\rm param}} \frac{(u_\alpha - \eta_\alpha)^2}{2\sigma_\alpha^2}\right)$
- Each example requires specification of $\eta_\alpha, \quad \sigma_\alpha$

# Our choice of prior (continued)

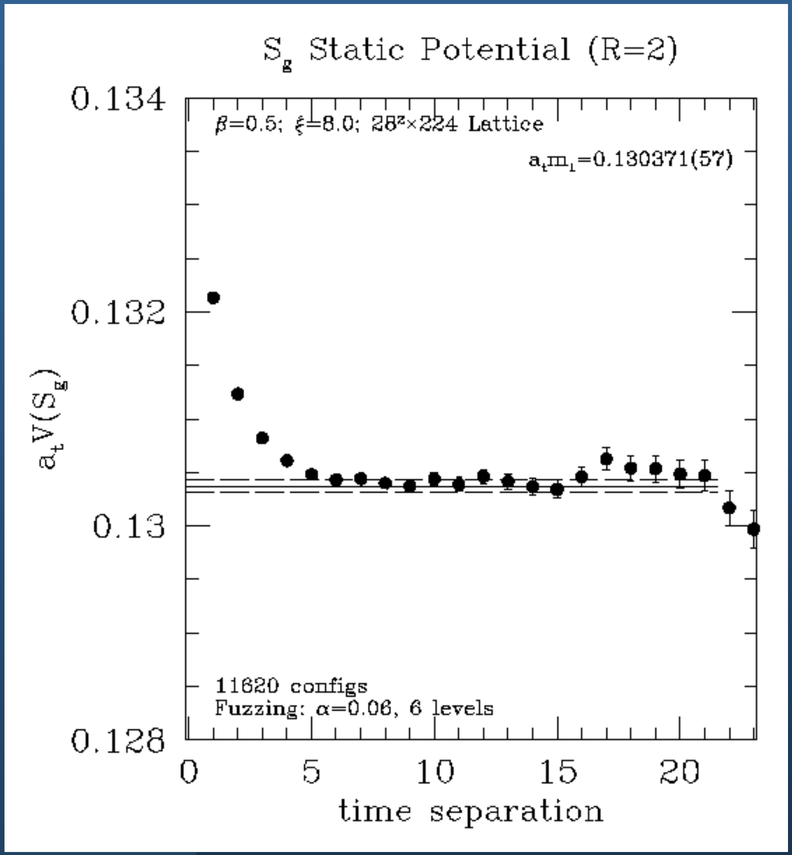
- No prior for first $N_{cor}$ energies
- Energies of excited-state contamination taken to be most likely equally spaced above the $N_{cor}$-th level
- Due to variational construction of our operators:
  - correlator $j$ dominated by $E_j$ exponential
  - all other coefficients small, taken to be most likely all equal
- Parameters in the prior:

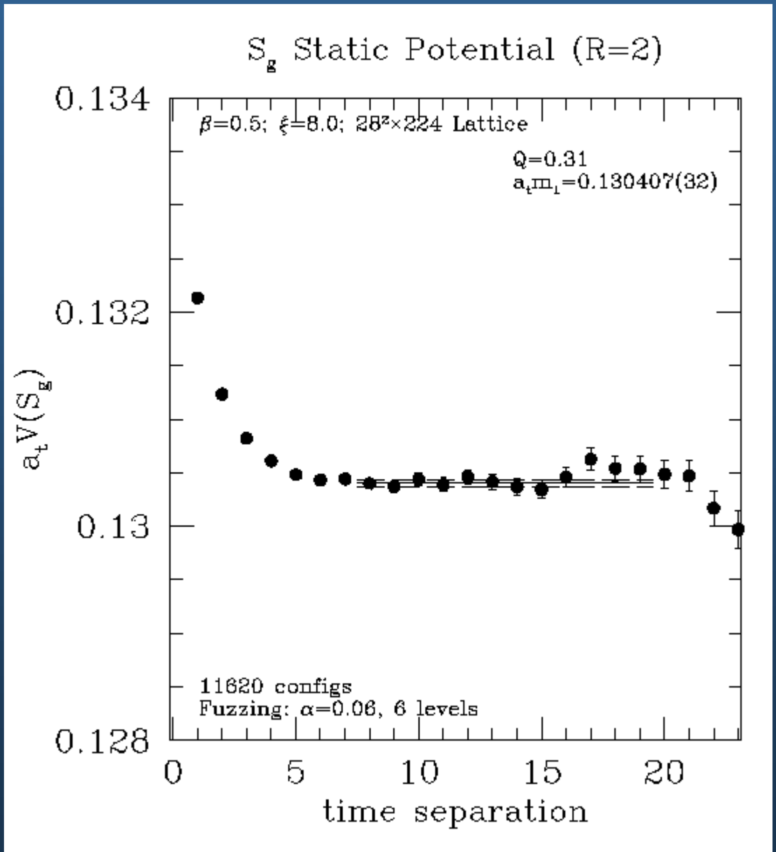
$$
\begin{aligned}
\varepsilon &\longrightarrow \epsilon_j, \quad j = N_{\mathrm{cor}} \ldots N_{\mathrm{exp}}, \\
\Gamma &\longrightarrow b_j^{(j)}, \quad j = 1 \ldots N_{\mathrm{exp}}, \\
\gamma &\longrightarrow b_j^{(i)}, \quad i \neq j.
\end{aligned}
$$

- Typical values: $\Gamma = 0.9(2)$, $\gamma = 0.05(5)$, $\varepsilon = 0.2(1)$
- Increase $N_{esc}$ until energies of interest stabilize

# Example 1 – Bayesian analysis comparison

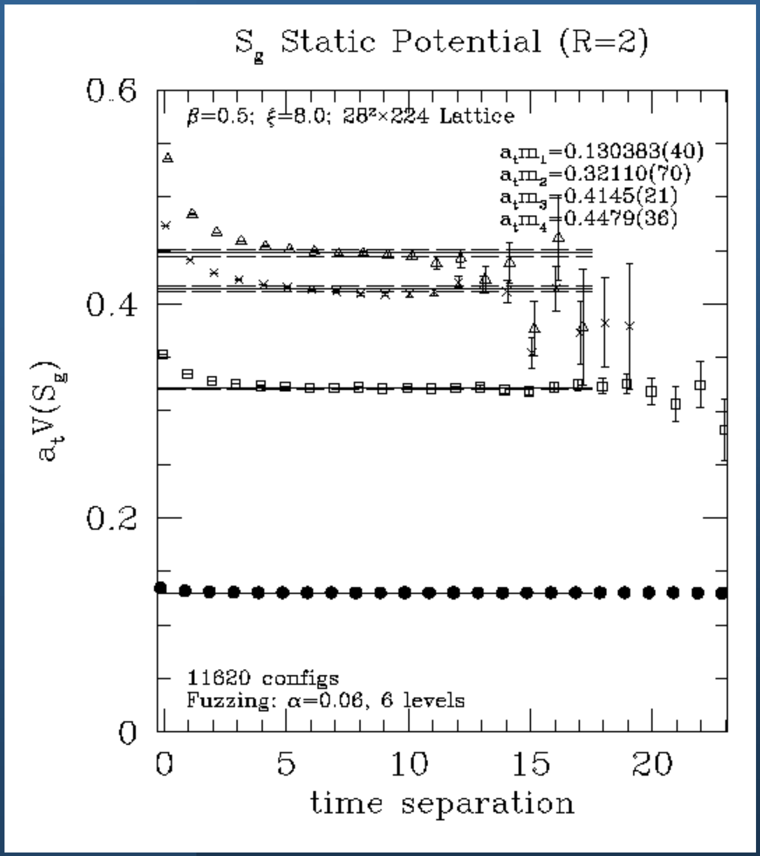
- $N_{esc} = 30$

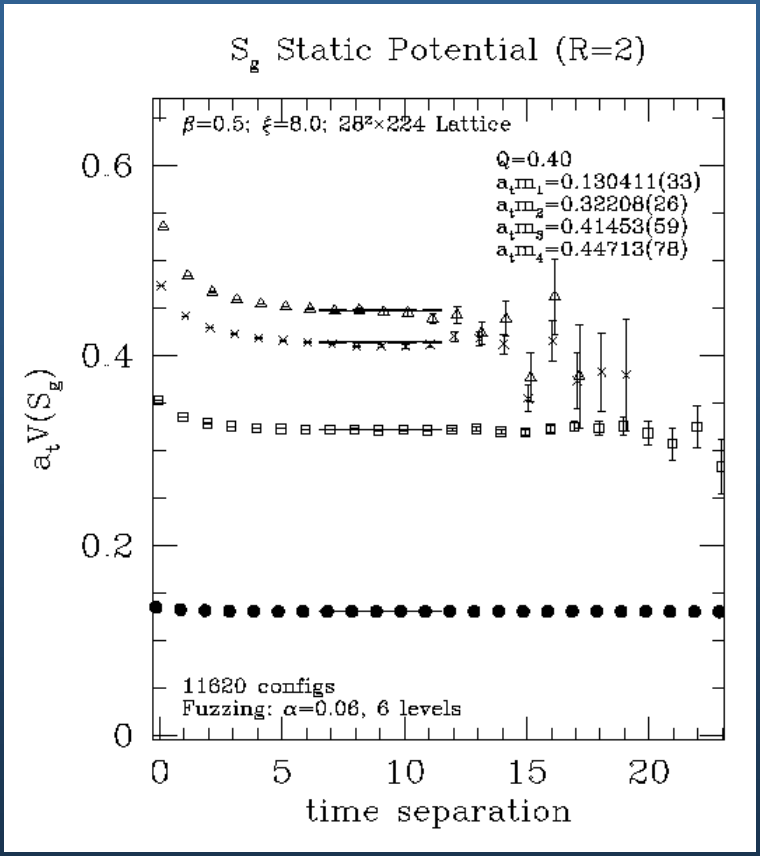


# Example 1 – comparison (continued)

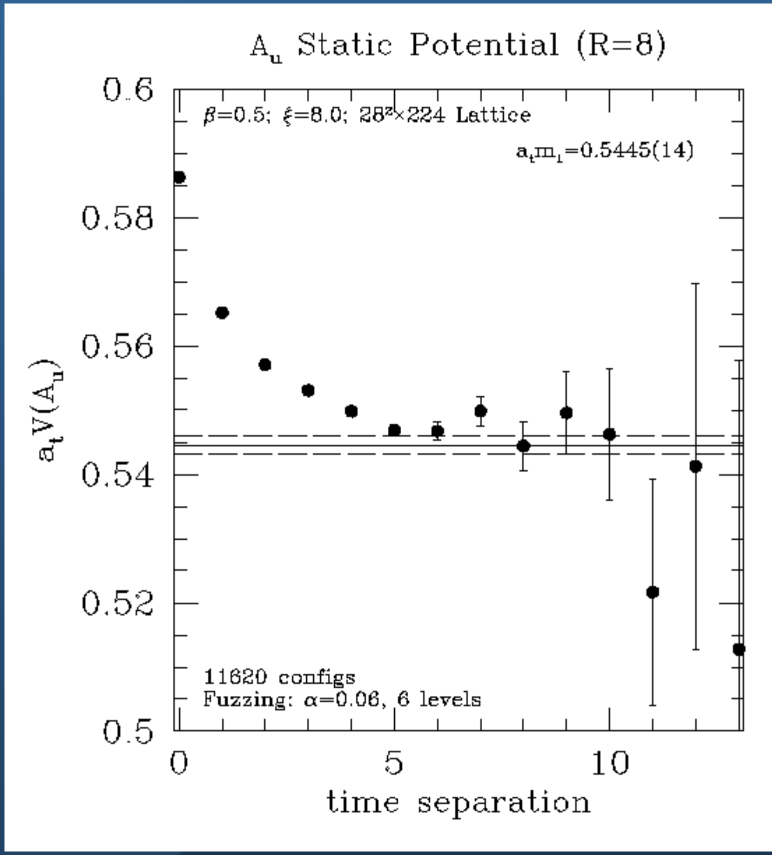
- $N_{esc} = 20$

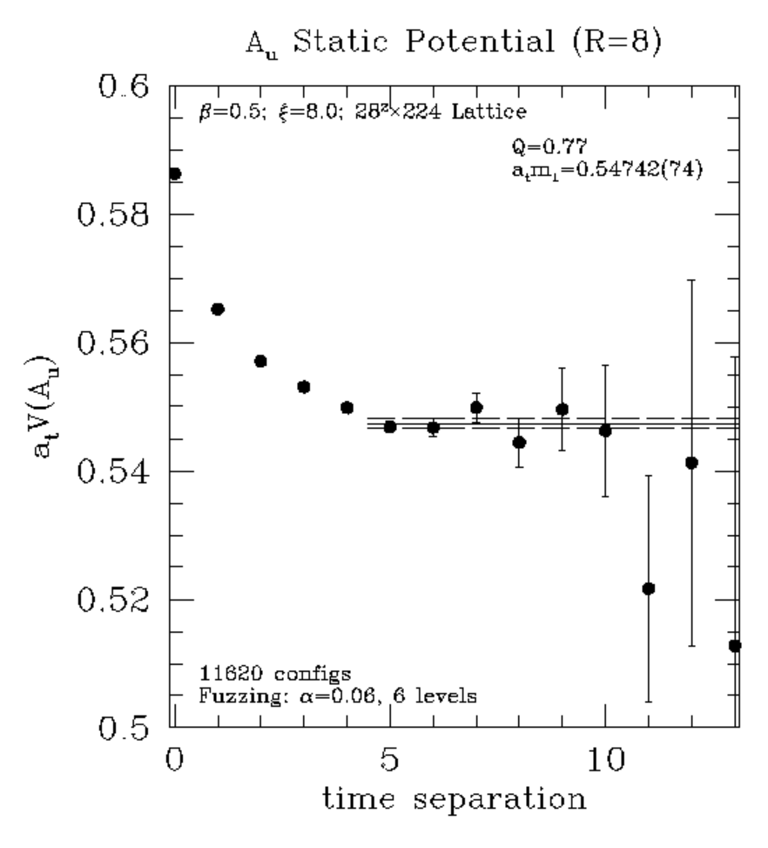


# Example 2 – Bayesian analysis comparison

- $N_{esc} = 30$



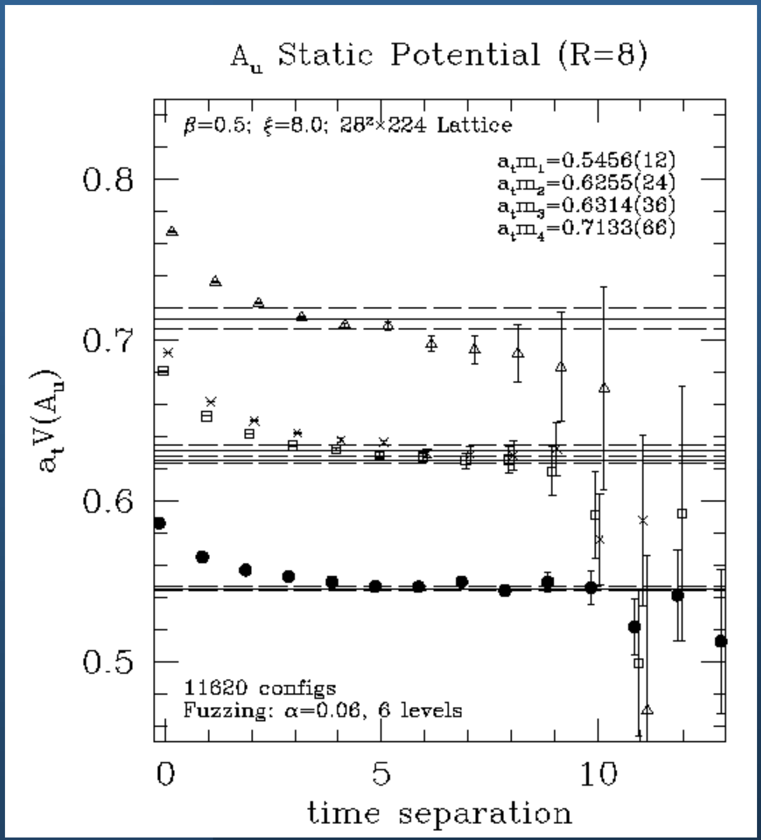
# Example 2 – sensitivity to prior

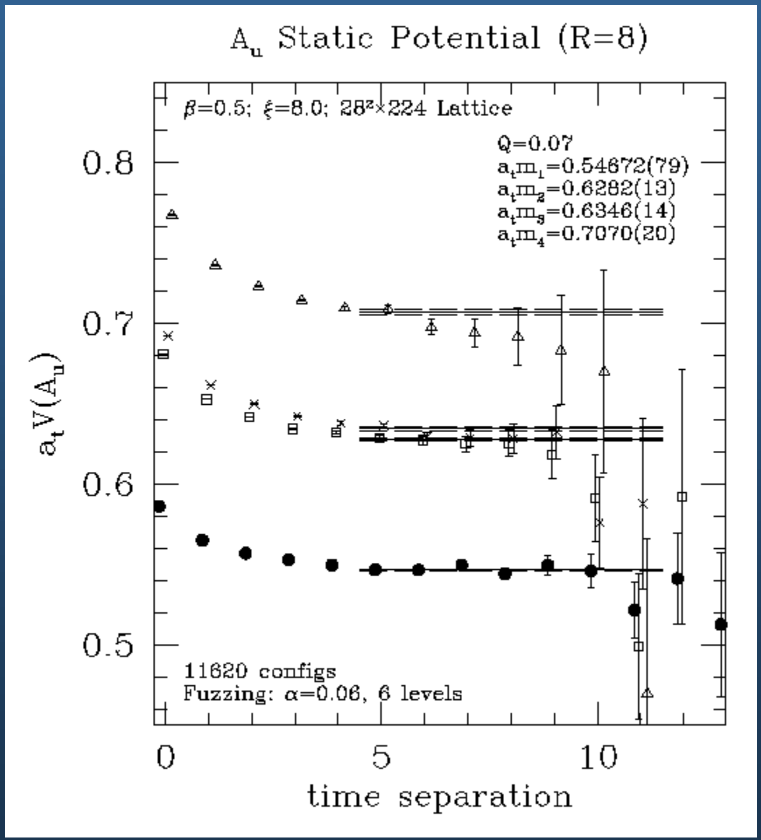
- Reference prior parameters:
  - $\Gamma = 0.9$ +/- 0.2, $\gamma = 0.05$ +/- 0.05, $\varepsilon = 0.2$ +/- 0.1
  - $N_{esc} = 50$, fit range $t = 0$ .. 14

| Change in prior | $E_0$ |
|---|---|
| Reference | 0.5446(16) |
| $\varepsilon = 0.30(15)$ | 0.5450(15) |
| $\varepsilon = 0.10(5)$ | 0.5469(12) |
| $\Gamma = 0.8(3)$, $\gamma = 0.1(1)$ | 0.5412(29) |
| $\Gamma = \gamma = 0.7(7)$ | 0.472(59) |

# Example 2 – continued

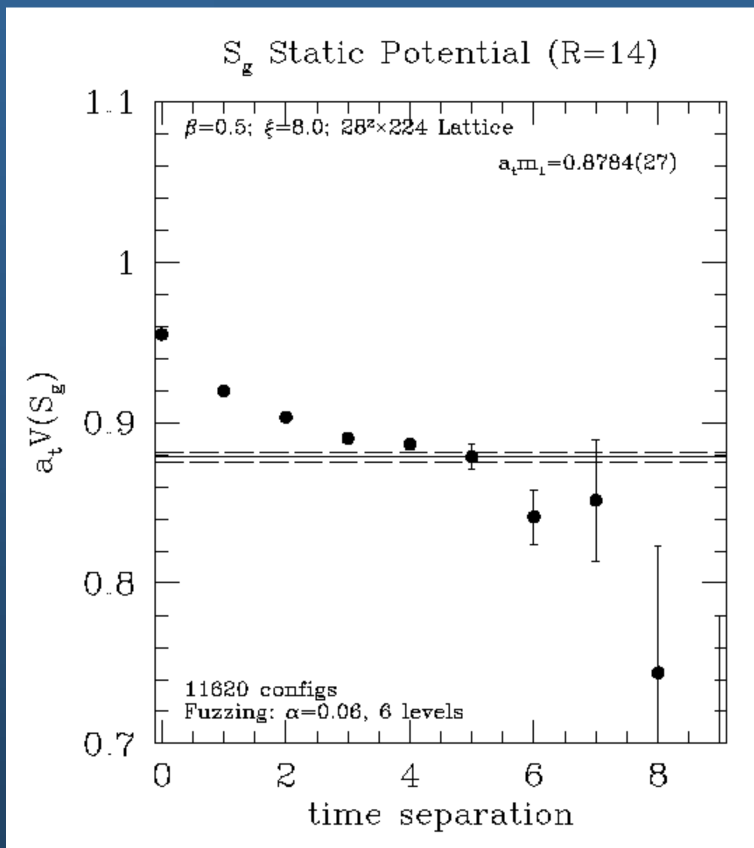
- $N_{esc} = 30$

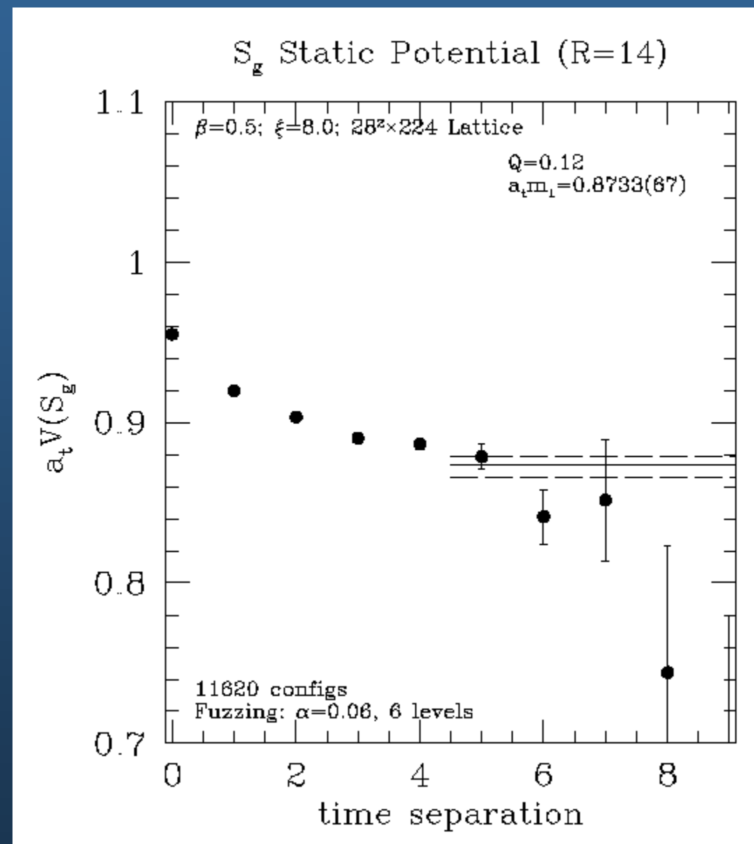


# Example 3 – Bayesian analysis comparison

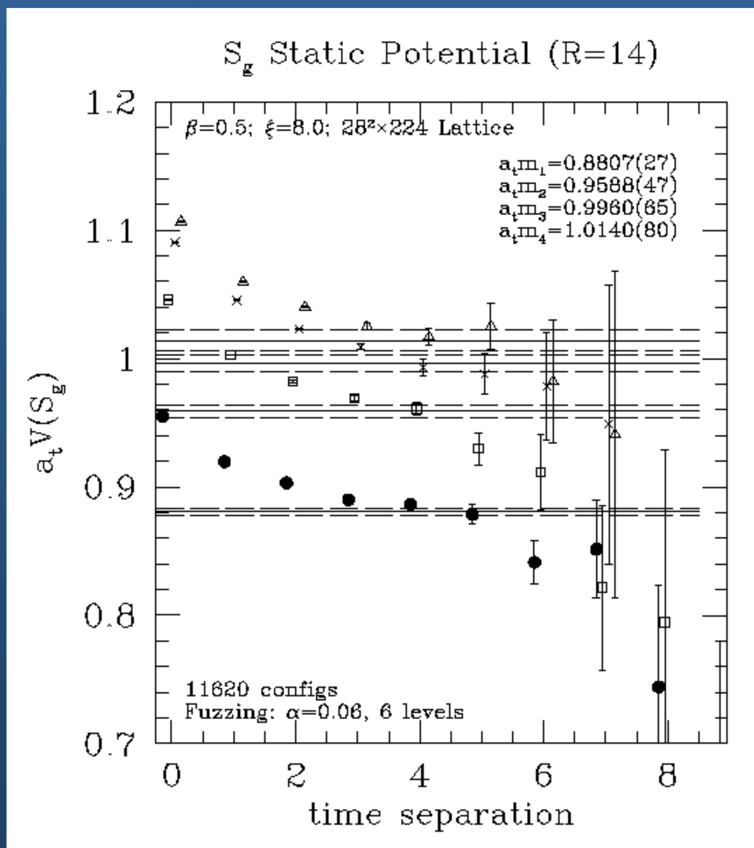
- $\varepsilon = 0.10(5)$ and $N_{esc} = 60$

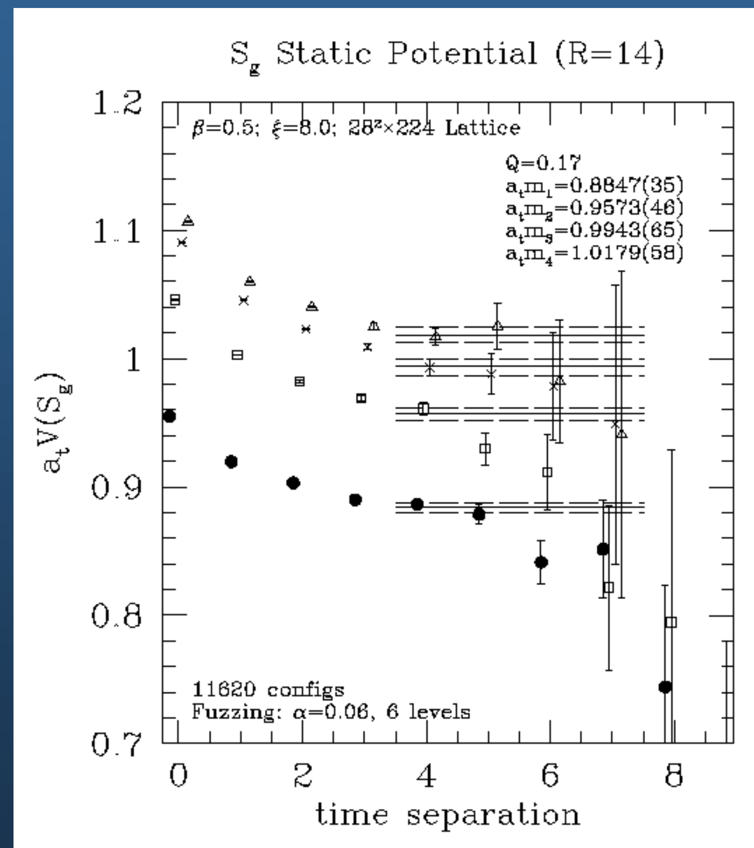


# Example 3 – continued

- ε = 0.10(5) and $N_{esc}$ = 50



# Comments on prior parameters

- Poor fits if ε set too large
- Care needed to set range of coefficients since allowing their ranges to be too large amplifies errors
- Results insensitive to moderate changes in prior parameter

# Two viewpoints

- Cautious viewpoint:
  - use prior information to constrain excited-state contamination to help extract the parameter of interest
  - example: ground state energy from single correlator
- Aggressive shoot-for-the-moon viewpoint:
  - use prior information to help extract *more* information from the data than otherwise possible
  - example: ground and first-excited state energies from a single correlator

# Bayesian tools

- Robustness: How can I tell if my prior has an undue impact on the results?
  - sensitivity analysis
    - sensitivity of results to reasonable modifications of prior
  - comparison of prior and posterior marginal distributions
- Model assessment: How can I tell if my model is providing adequate fit to the data?
  - cross-validation, model averaging,…
- Model selection: Which model(s) should I choose for final presentation of the results?
  - use of Bayes factors:
    - relative probabilities of two models $M_1$ and $M_2$

$$\frac{\int du\ P(D|M_1 \cap I)\ P(M_1|I)}{\int du\ P(D|M_2 \cap I)\ P(M_2|I)}$$

# Fly in the ointment?

- Errors shown in this talk assume strongly-peaked posterior
- Parameter estimates and errors require integrations over the parameters
  - Monte Carlo integrations techniques needed
  - Problems with autocorrelations due to "ridges"
  - Currently under investigation

# Conclusion

- Bayesian regression techniques are an alternative method for extracting physical observables from stochastically-determined correlation functions
- Uses all of your data
- Parameter errors can more easily incorporate systematics
- Takes into account prior knowledge of the system from theoretical considerations and/or previous experience
- Not a cure for bad data