Constrained Curve Fitting

P. Lepage

+ P. Mackenzie
C. Morningstar
K. Hornbostel
J. Shigemitsu
C. Davies
B. Clark

See talks: M. Wingate
M. Gürtler

The Problem

An Example

M.C. Sim'n $\overset{\text{eg}}{\Longrightarrow}$ $\overline{G}(t)$ for $t = 0, 1 \ldots 23$

Theory $\Rightarrow$ $G(t) = \sum_{m=1}^{\infty} A_m e^{-E_m t}$

24 pieces of data $\longleftrightarrow$ ∞ number of parameters to fit

How ?

↳ ~~systematic~~ errors?

↳ high-precision!

Sample Data

- 840 Υ (local-local) propagators at $\beta = 6.0$

- Fit: minimize $\chi^2(A_m, E_m)$ where

$$\chi^2 \equiv \sum_{t\,t'} \Delta G(t)\ \sigma_G^{-2}(t,t')\ \Delta G(t')$$

and

$$\Delta G(t) \equiv \overline{G}(t) - G_{th}(t; A_m, E_m)$$

$\overline{G}(t)$: M.C.

$G_{th}(t; A_m, E_m)$: $\sum_m A_m e^{-E_m t}$

$\sigma_G^2(t,t')$ ≡ covariance matrix (from MC)

Sol'n 1

Fit all t with $G(t) = \sum_{m=1}^{M} A_m e^{-E_m t}$

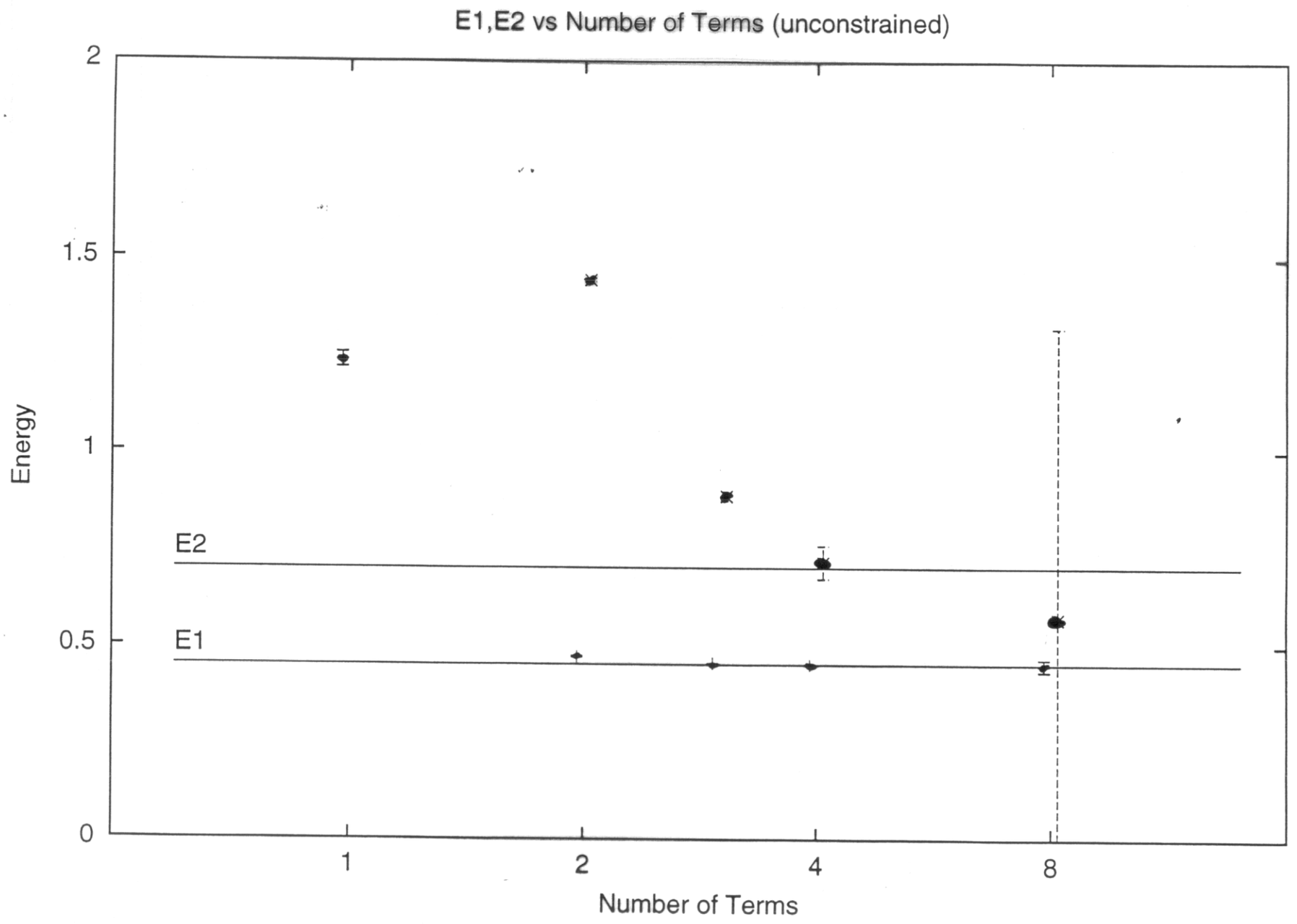
- increase M to get good fit
- Where do you stop?

M too small $\Rightarrow$ E_m's ... biased

M too large $\Rightarrow$ σ_{E_m}'s ... too large

$\to \infty$ if

$2M >$ # data points.


E1,E2 vs Number of Terms (unconstrained)
Energy
2
1.5
1
0.5
0
E2
E1
1
2
4
8
Number of Terms

Sol'n 2 (standard)

Truncate data ($t \geq t_{min}$)

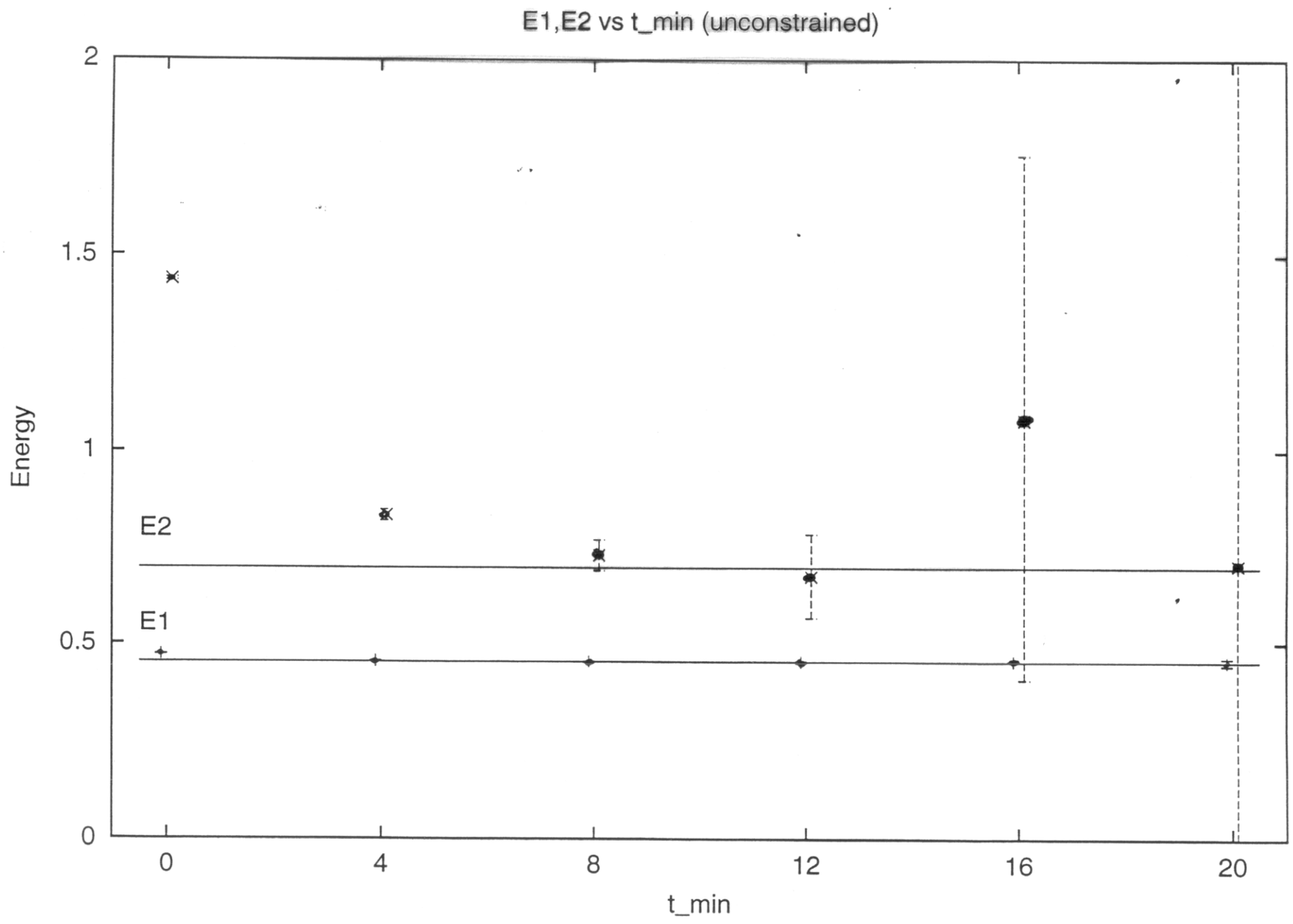
and theory (M = 2 terms)

- Increase t_{min} to get good fit
- Where do you stop?

  t_{min} too small $\Rightarrow$ E_m's ... biased

  t_{min} too large $\Rightarrow$ σ_{E_m}'s ... too large

→ usually truncate data until stat. errors large enough to swamp sys. error due to truncating theory (hopefully).


E1,E2 vs t_min (unconstrained)
Energy
t_min
E2
E1
2
1.5
1
0.5
0
0
4
8
12
16
20

Better Solution

- Want to fit $\forall$ data ($t_{min} = 0$)
with $\sum_{M}^{} A_M e^{-E_M t}$ for $M \to \infty$.

Problem: e.g., $M = 8$

$\Rightarrow$ $A_4 \sim 5$ to $10 \times A_1$

But Physics $\Rightarrow$ $A_4 \lesssim A_1$

Need to teach Physics to fitting software!

Constrained Fit

Replace $\chi^2 \rightarrow$ augmented χ^2_a :

$$\chi^2_a \equiv \chi^2 + \chi^2_p$$

where

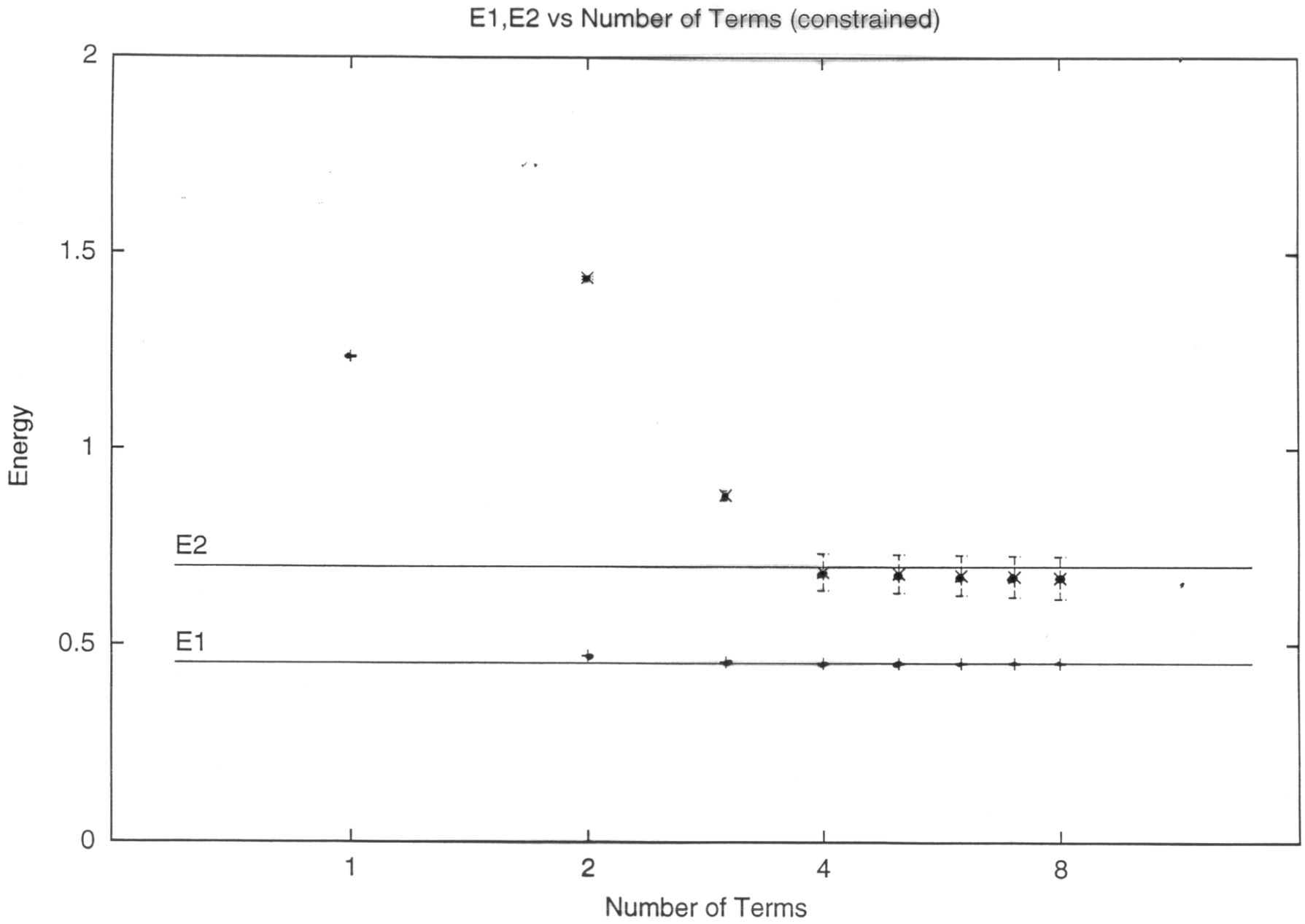
$$\chi^2_p \equiv \sum_m \frac{(A_m - \tilde{A}_m)^2}{\tilde{\sigma}^2_{A_m}} + \sum_m \frac{(E_m - \tilde{E}_m)^2}{\tilde{\sigma}^2_{E_m}}$$

→ Favors A_m in interval $\tilde{A}_m \pm \tilde{\sigma}_{A_m}$

E_m " " $\tilde{E}_m \pm \tilde{\sigma}_{E_m}$

→ $\tilde{A}_m$, $\tilde{\sigma}_{A_m}$ ... are <u>inputs</u> ;

← "PRIORS" choose reasonable values on basis of <u>prior</u> knowledge


E1,E2 vs Number of Terms (constrained)
Energy
Number of Terms
E2
E1
2
1.5
1
0.5
0
1
2
4
8

Method:

Increase M = # terms until answers (and errors) converge.

N.B.!

- Can't make M too large;
  c.f. t_{min} which must be optimized!

- Errors = statistical due to M.C.
  + systematic due to poorly constrained high-E states.

  → assumptions are explicit & intelligible

- Constraints ⇒ better numerics:
  e.g., can fit 100 term G

NB

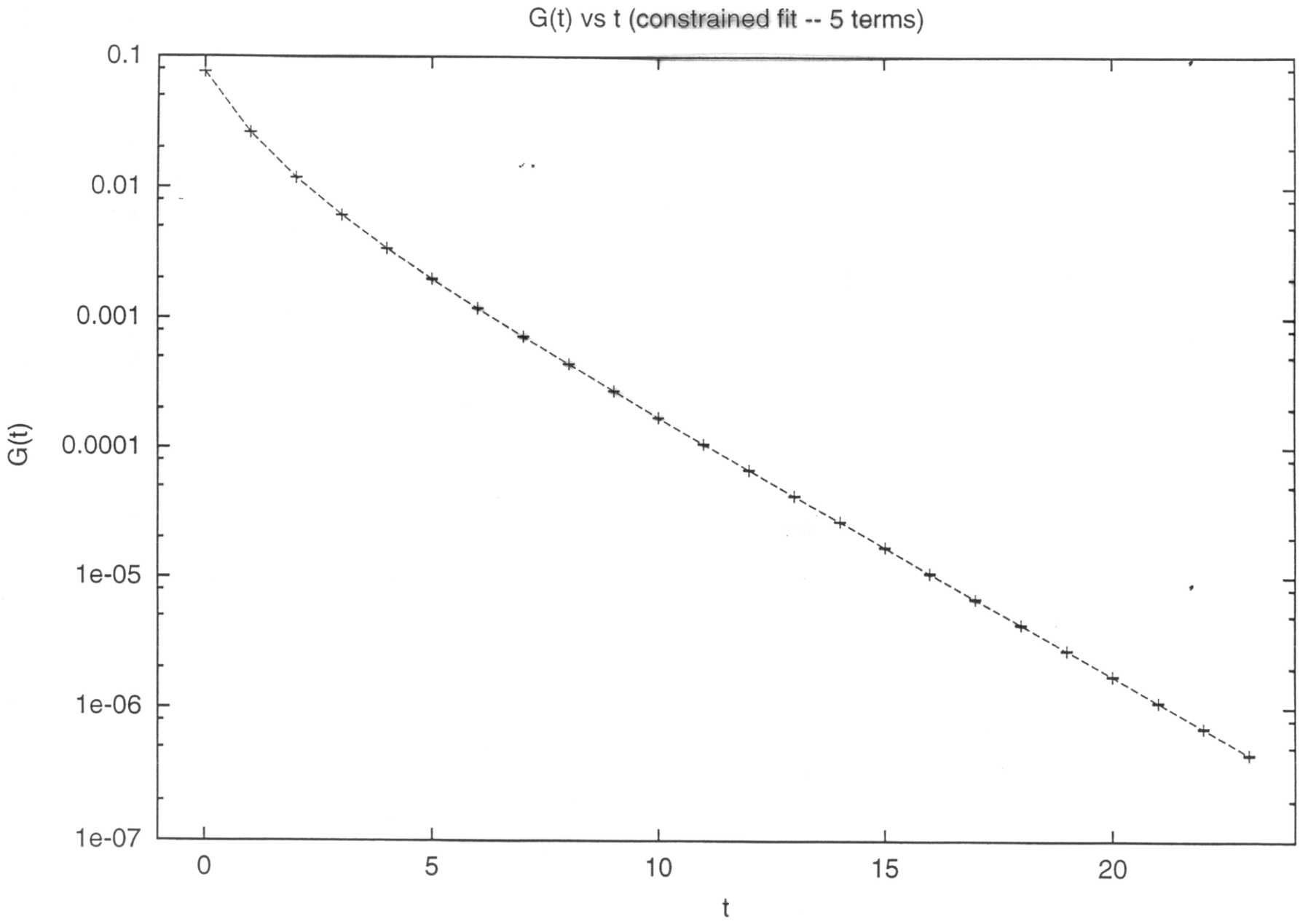
Convergence (as M increases) is key criterion.

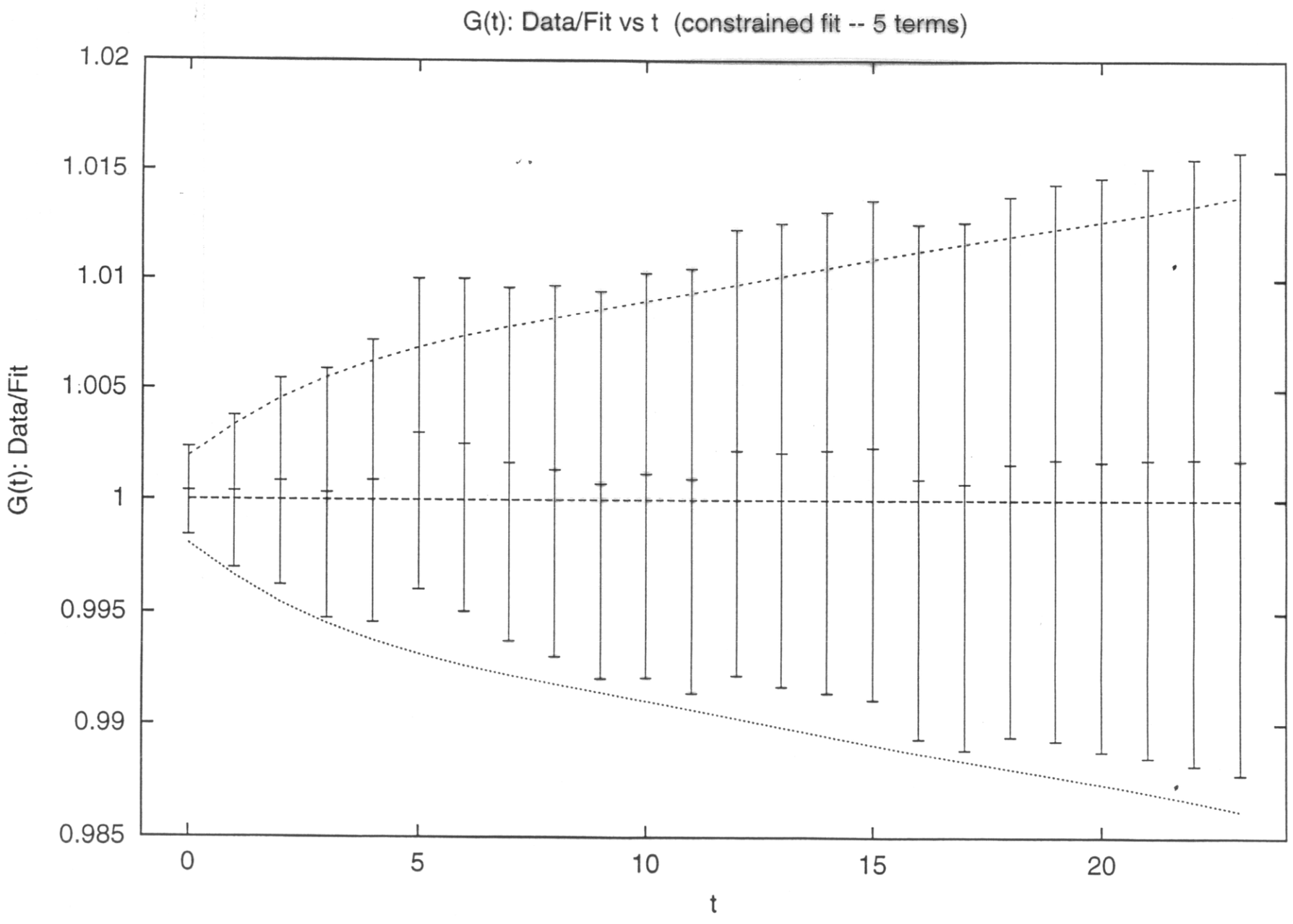
χ^2, Q ... play secondary role

Fits here have

$$\frac{\chi^2_{a,\min}}{\#\text{ Data Points}} \sim 0.7 - 0.8$$

↑ correct ratio to monitor.


G(t) vs t (constrained fit -- 5 terms)
0.1
0.01
0.001
0.0001
1e-05
1e-06
1e-07
G(t)
0
5
10
15
20
t


G(t): Data/Fit vs t (constrained fit -- 5 terms)
G(t): Data/Fit
t
1.02
1.015
1.01
1.005
1
0.995
0.99
0.985
0
5
10
15
20

B.

- Actual fit parameters:

$$a_m \equiv \log A_m \quad \Rightarrow \quad A_m > 0$$

$$\varepsilon_m \equiv \log(E_m - E_{m-1}) \quad \Rightarrow \quad E_m > E_{m-1}$$

- Priors:

$$\left.\begin{aligned} \tilde{a}_m &= \log 0.02 \\ \tilde{\sigma}_{a_m} &= 2 \end{aligned}\right\} \Rightarrow \quad A_m = 0.02\,{}^{+0.02}_{-0.01}$$

$$\left.\begin{aligned} \tilde{\varepsilon}_m &= \log 0.2 \\ \tilde{\sigma}_{\varepsilon_m} &= 2 \end{aligned}\right\} \Rightarrow \quad E_m - E_{m-1} = 0.2\,{}^{+0.2}_{-0.1}$$

- Results insensitive to precise values of priors (within reason)?

Double $\tilde{\sigma}_{am}$ & $\tilde{\sigma}_{Em}$ $\Rightarrow$

$E_1 = 0.4526(15) \rightarrow 0.4528(14)$

$E_2 = 0.683(49) \rightarrow 0.697(50)$

$E_3 = 1.05(12) \rightarrow 1.10(21)$

$\Rightarrow$ E_1, E_2 largely determined by the M.C. data

E_3 strongly affected by priors

1B. Other methods are special cases:

$t_{min} > 0$ Method $\Rightarrow$

$$\tilde{\sigma}_{A_m}, \tilde{\sigma}_{E_m} = \begin{cases} \infty & \text{for } 1^{st} \text{ terms} \\ 0 & \text{for all rest.} \end{cases}$$

↳ forcing $A_m = 0$

$t_{min} = 0$ $M \to \infty$ Method $\Rightarrow$

$$\tilde{\sigma}_{A_m}, \tilde{\sigma}_{E_m} = \infty \quad \text{for } \forall \text{ terms}$$

$\Rightarrow$ ∞ fit errors on all A_m's & E_m's (as $M \to \infty$)

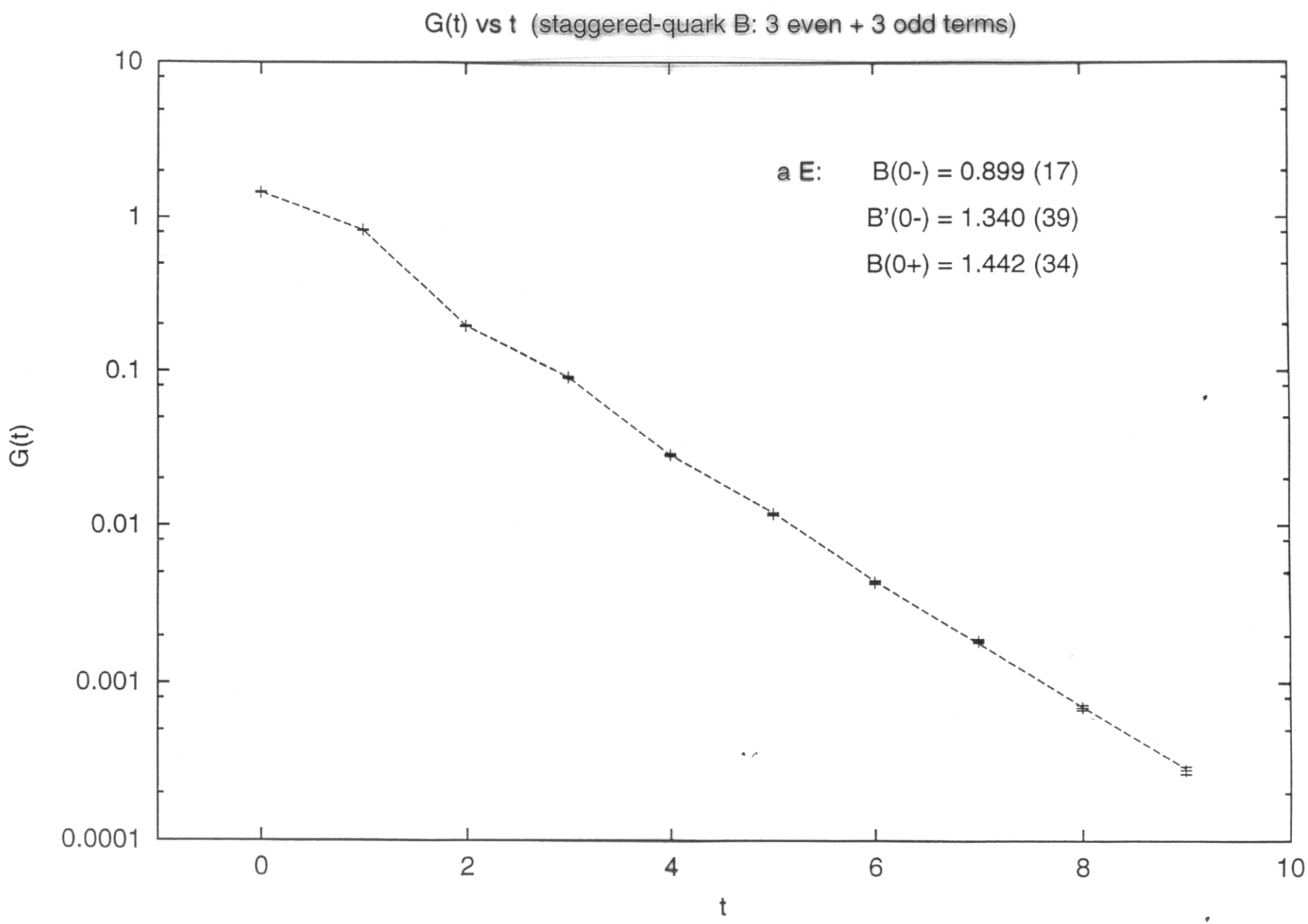
↳ Correct answer if really don't have any (prior) idea of sizes of A_m's & E_m's.

Lots of Other Fits

- Matrix G's up to 4×3 (30-40 --- parameters)
- simultaneous fits of multiple channels (eg, π & ρ)
- $V(r)$, glueballs
- staggered quarks

.
.
.

→ All worked well


G(t) vs t (staggered-quark B: 3 even + 3 odd terms)
a E: B(0-) = 0.899 (17)
B'(0-) = 1.340 (39)
B(0+) = 1.442 (34)
G(t)
t
10
1
0.1
0.01
0.001
0.0001
0
2
4
6
8
10

Theory & Interpretation

Interpretation : Bayesian Statistics

→ Calculus of probabilities provides framework

Not'n :

$P(A|B)$ = probability of A given B (i.e. assuming B true)

$p_i = \{A_1, E_1 \dots\}$ = parameters

$\eta_i = \{\tilde{\sigma}_{A_1}, \tilde{A}_1 \dots\}$ = priors

$\bar{G}$ = M.C. data average

Assumption #1:

M.C. data set sufficiently large that $\bar{G}$ has Gaussian statistics (Central limit thm)

$\Rightarrow$ $P(\bar{G} \mid \rho) \equiv$ prob. of a particular $\bar{G}$ given a theory with param's ρ_i:

$$\propto e^{-\chi^2(\rho)/2}$$

$$\chi^2 \equiv \sum_t \frac{\left(\bar{G} - G_{th}(\rho)\right)^2}{\sigma_G^2}$$

$\bar{G}$, σ_G from M.C.

N.B. σ_G actually ρ-dependent

Goal: Don't want $P(\bar{G}|\rho)$!

Need

$P(\rho|\bar{G})$ = Prob. of particular parameter values ρ_i given $\bar{G}$

$$\langle \rho_i \rangle \equiv \int \pi d\rho \; P(\rho|\bar{G}) \; \rho_i$$

= estimate for param. ρ_i

$$\sigma_{\rho_i}^2 \equiv \langle (\rho_i - \langle \rho_i \rangle)^2 \rangle$$

= variance of estimate

$$\langle f(\rho) \rangle = \int \pi d\rho \; P(\rho|\bar{G}) \; f(\rho)$$

= estimate for fcn $f(\rho)$

⋮

Bayes' Th'm

Probab. Product Rule $\Rightarrow$

$$P(\bar{G}\,\rho) = P(\bar{G}\,|\,\rho)\,P(\rho)$$

$$= P(\rho\,|\,\bar{G})\,P(\bar{G})$$

$\Rightarrow$

$$P(\rho\,|\,\bar{G}) = \frac{P(\rho)\,P(\bar{G}\,|\,\rho)}{P(\bar{G})} \rightarrow \text{indep of } \rho$$

$$\propto P(\rho)\,P(\bar{G}\,|\,\rho)$$

$P(\rho)$ = probab. $\{\rho_i\}$ correct given no $\bar{G}$

"Prior" = Prior Assumptions

Assumption #2:

Prior dist'n $P(\rho)$ can be approximated by a Gaussian:

$$P(\rho) = e^{-\chi_p^2(\rho)/2}$$

where

$$\chi_p^2(\rho) = \sum_i \frac{(\rho_i - \tilde{\rho}_i)^2}{\tilde{\sigma}_{\rho_i}^2}$$

$\Rightarrow$ a priori assumption is ~~that~~ $\rho_i \sim \tilde{\rho}_i \pm \tilde{\sigma}_{\rho_i}$

$\Rightarrow$

$$P(\rho|\bar{G}) \propto e^{-\chi^2/2} e^{-\chi_p^2/2}$$

$$\propto e^{-\chi_a^2(\rho)/2}$$

$\hookrightarrow$ augmented χ^2 from before.

Why Gaussian's for $P(\rho)$?

- Least biased choice for $P(\rho)$ minimizes information content

$\Rightarrow$ maximizes entropy

$$S \equiv \int -P(\rho) \log P(\rho)\, d\rho$$

- Given $\langle \rho \rangle = \tilde{\rho}$, $\langle (\rho - \langle \rho \rangle)^2 \rangle = \tilde{\sigma}^2$

Compute

$$\frac{\delta}{\delta P}\left[S + \lambda \int P + \mu \int P\rho + \nu \int P\rho^2 \right] = 0$$

$$\Rightarrow \boxed{P(\rho) = \frac{1}{\sqrt{2\pi}\,\tilde{\sigma}}\, e^{-(\rho-\tilde{\rho})^2/2\tilde{\sigma}^2}}$$

$\hookrightarrow$ "Least Informative Prior"

Error Estimation

Best Fit

- From integrals :

$$\langle \rho_i \rangle = \frac{\int \Pi d\rho \; e^{-\chi_a^2/2} \rho_i}{\int \Pi d\rho \; e^{-\chi_a^2/2}}$$

$$\sigma_{\rho_i}^2 = \langle (\rho_i - \langle \rho_i \rangle)^2 \rangle \quad \ldots$$

- Evaluate using

  1) Vegas ....

  2) M.C. – Metropolis, HMC, HMD..

  ↳ $\chi_a^2 \leftrightarrow$ Action

  $\rho_i \leftrightarrow$ Field

N. B.

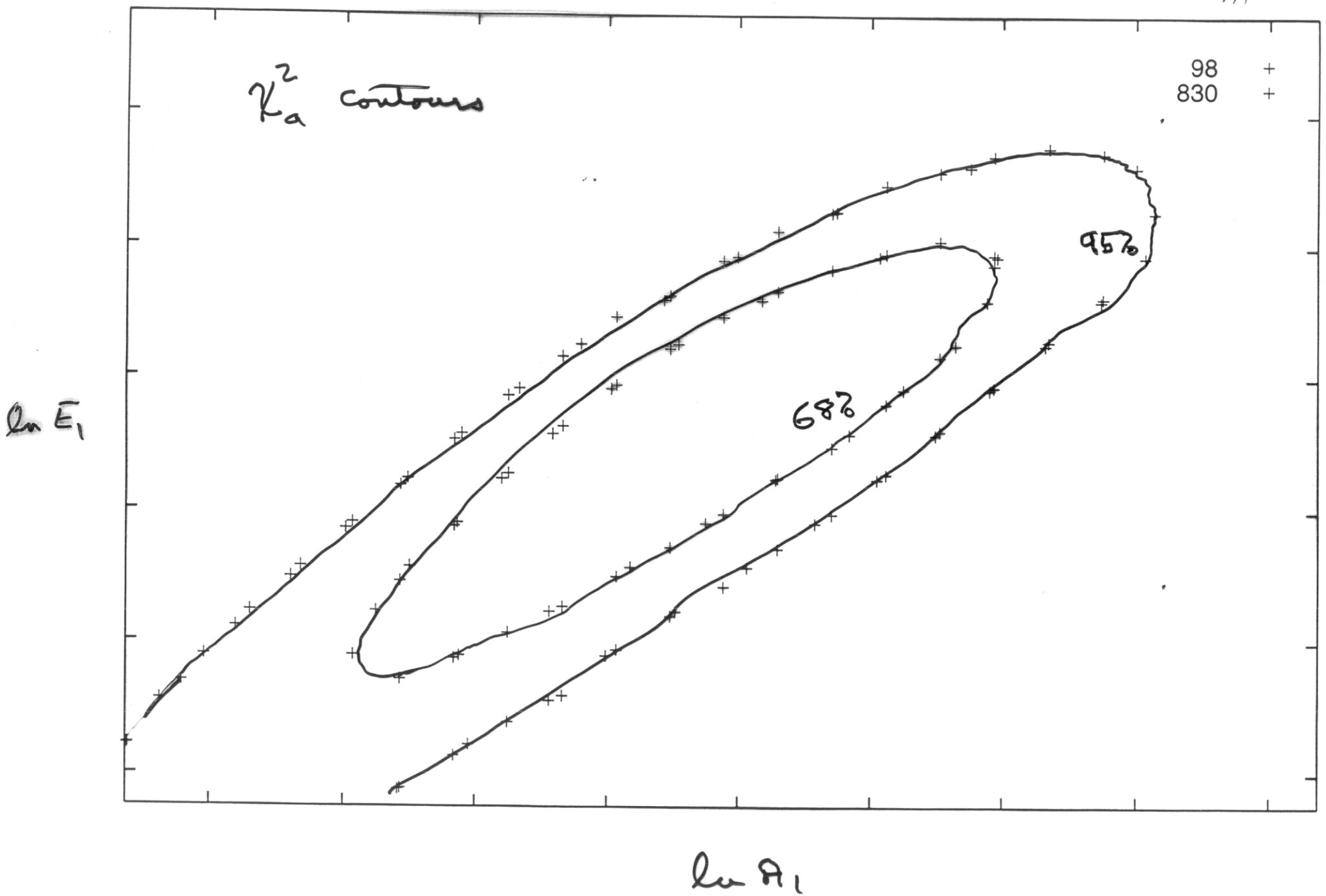
Integrals can be <u>extremely</u>
difficult & costly

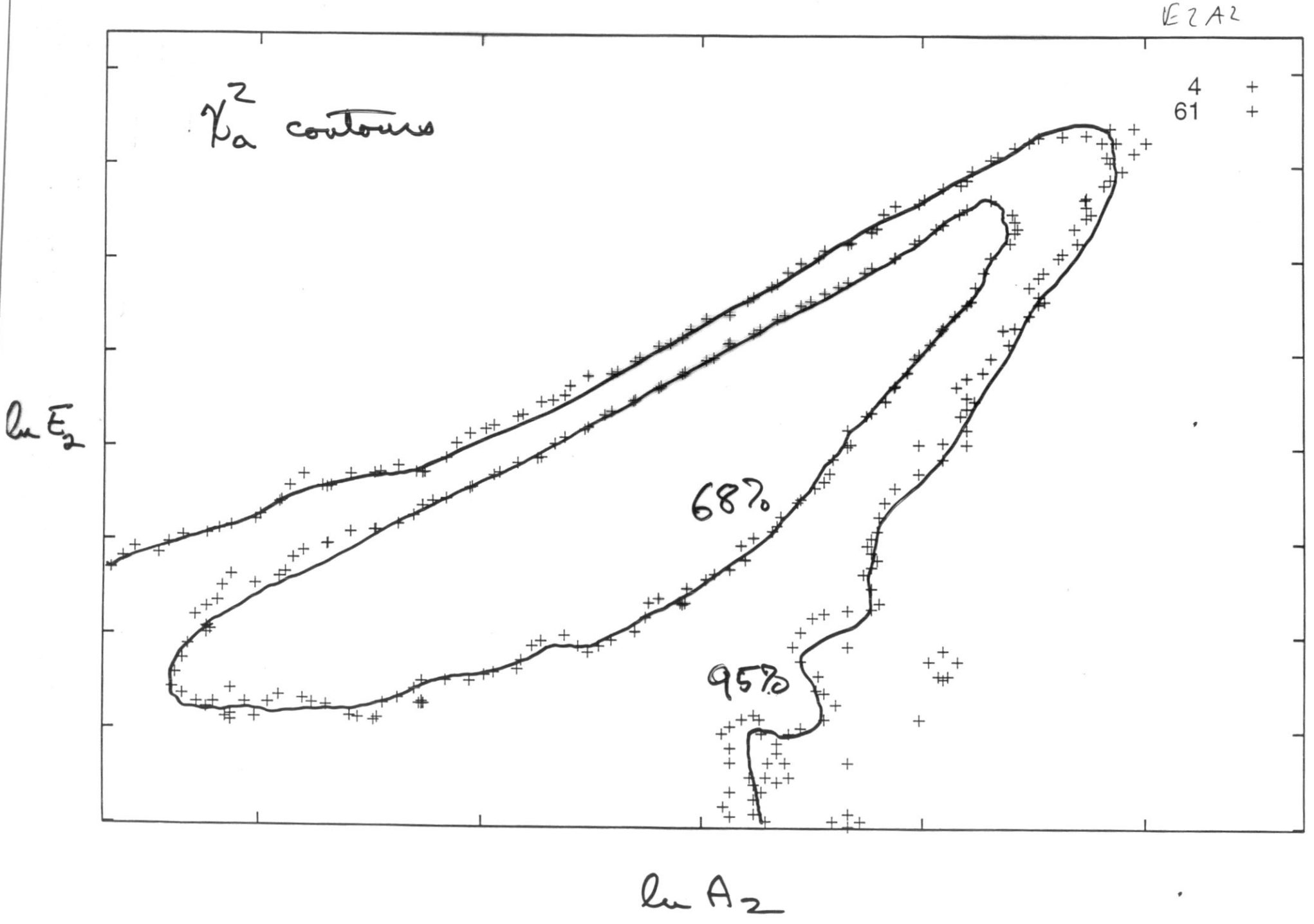
→ integrand extremely
sharply peaked

→ has long, narrow,
high wandering
ridges ⇒ huge
auto correlation times!

⇒ approx'ns useful.

E1 A1




$E_2 A_2$
χ^2_a contours
4 +
61 +
68%
95%
ln E_2
ln A_2

Gaussian Approx' (High statistics)

$$\chi^2_a(\rho) \approx \chi^2_{a,\min} + \frac{1}{2}(\rho-\rho^*)_i (\rho-\rho^*)_j \partial_i\partial_j \chi^2_a$$

$$\Rightarrow \quad \langle \rho_i \rangle \approx \rho_i^*$$

$$\sigma^2_{\rho_i} \approx C_{ii} \quad \text{where } C^{-1}_{ij} = \frac{\partial_i\partial_j \chi^2_a}{2}$$

...

$\Rightarrow$ constrained curve fitting based on χ^2_a minimization as in 1st part of talk.

Bootstrap Approx'n

Modified Bootstrap Procedure:

1) Replace $\bar{G} \to \bar{G}'$ = avg of bootstrap copy of original G's

Replace $\tilde{\rho}_i \to \tilde{\rho}_i'$ = Gauss. random # from dist'n with mean $\tilde{\rho}_i$ & s-dev $\tilde{\sigma}_{\rho_i}$.

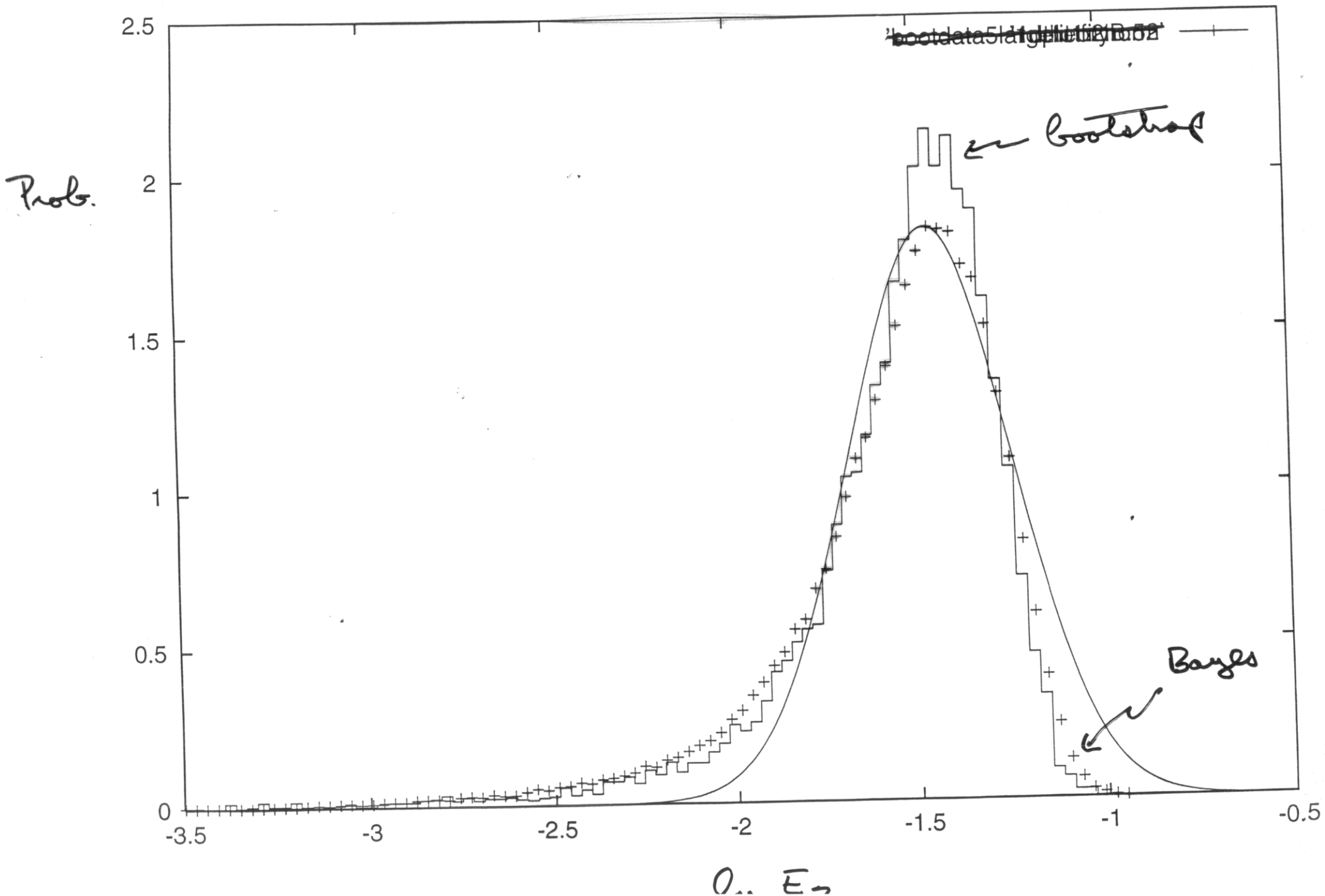
2) Compute ρ_i^* = min of χ^2_a using $\bar{G}'$, $\tilde{\rho}_i'$

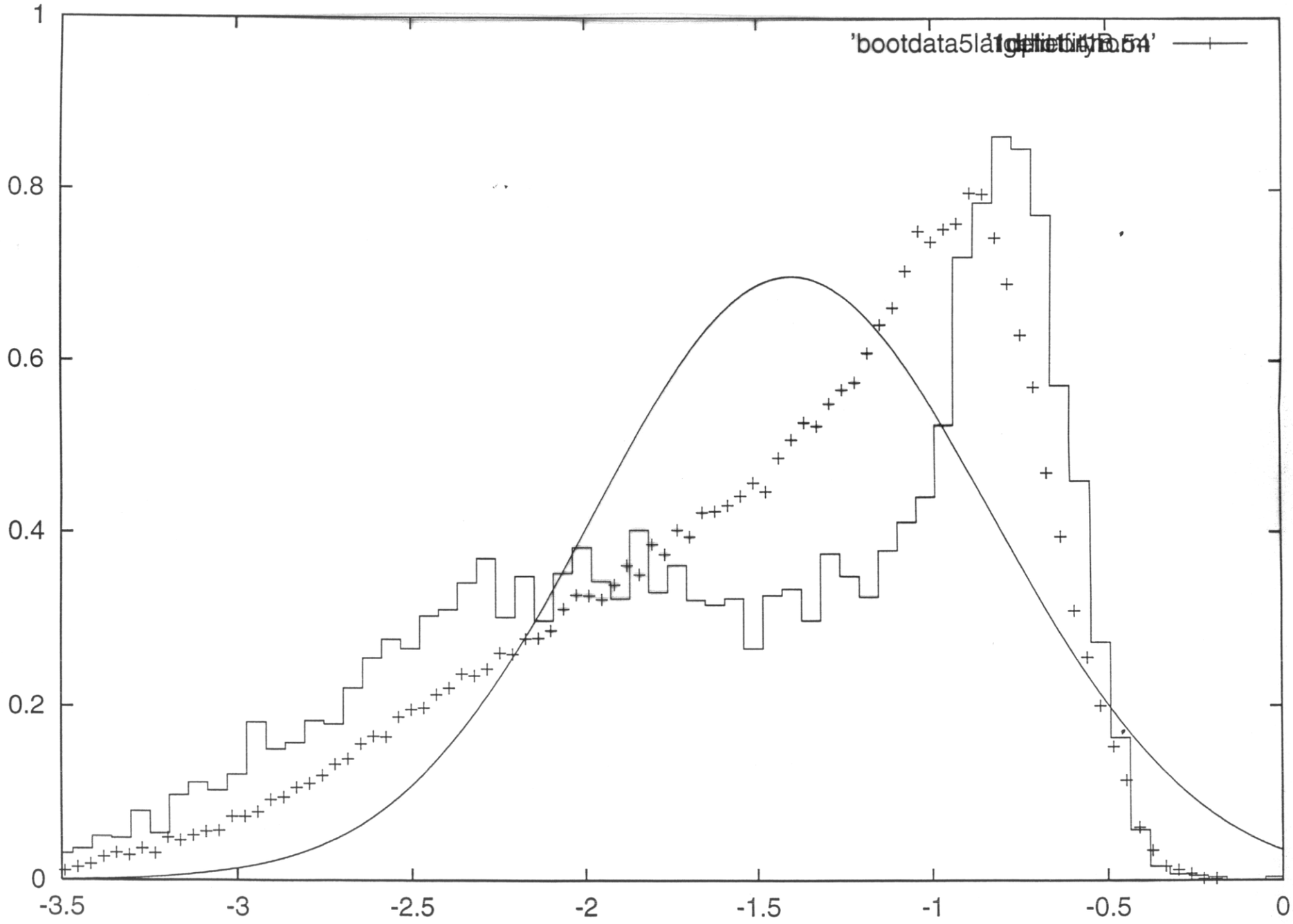
3) Repeat steps 1 & 2 100-1000 times.

$\Rightarrow$ ensemble of ρ_i^*'s

$\Rightarrow$ avg's on ρ_i^* $\approx$ Bayes avg's on ρ_i

↳ small add'l corr'n but usually not important (?)


Prob.
2.5
2
1.5
1
0.5
0
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
Bootstrap
Bayes
O.. E


'bootdata5la
1
0.8
0.6
0.4
0.2
0
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0

Other Applications

Pert'n Theory from High-β Simulations

Trottier GPL Mackenzie
Shakespeare

- Measure $W(m,m)$ for 9 lattice spacings $\Rightarrow$ $\alpha_p(3.4/a) = 0.01 - 0.07$.

- Fit M.C. values $-\dfrac{\log W(m,m)}{2m+2m}$

To $\sum_{n=1}^{M\geq 5} c_n \alpha_p^n(q^*)$ + priors $\tilde{c}_n = 0$, $\tilde{\sigma}_{c_n} = 5$

- $W(5,5)$ Results (e.g.):

		Exact
c_1 =	1.7693 (6)	1.7690
c_2 =	−1.201 (40)	−1.177
c_3 =	4.3 (7)	?

→ Fix c_1, c_2 to exact val

$\Rightarrow$ $c_3 = 3.91\ (23)$

N.B.

Some c_n's fixed by MC data ($n \leq 3$); others by priors.

As statistics improve more c_n's fixed by data $\Rightarrow$

> order of fit increases automatically as warranted by statistics

↳ i.e. never agonize about linear vs. quadratic fit!
Set $M = 10$ or 20 and forget about order!

Extrapolations in Lattice Size L

Lattice ...

Same program for Polyakov line

$\Rightarrow$ static-quark energy $E_0 = \sum_m c_m \alpha_p^m(\beta^*)$

$\Rightarrow$ M.C. values for $c_3(L)$ for $L = 3, \cdots, 11$

Fit To

$$c_3(L) = c_3 + p_1 \frac{\ln^2 L^2}{L} + p_2 \frac{\ln L^2}{L} + \frac{p_3}{L} + \frac{p_4}{L^2} + \frac{p_5}{L^3} + \cdots$$

+ priors : $\tilde{\sigma}_{c_3} = \tilde{\sigma}_{p_i} \simeq 6$

and extrapolate to $L = \infty$.

Result : $c_3 = 3.8 \pm 1.1$

c.f. Burgio et al $\Rightarrow$ 3.2 ± 0.6

N.B.

$$E_0 = 1.0701\ \alpha_p(q^*) + 0.117\ \alpha_p^2 + 3.3(5)\ \alpha_p^3 + \dots$$

UV

$$E_0^{\text{unimp}} = 2.1173 \cdot \alpha_{lat} + 11.152\ \alpha_{lat}^2 + 86.3(5)\ \alpha_{lat}^3 + \dots$$

N.B.

Analogous to chiral extrapol'n

$\Rightarrow$ "all orders" extrapol'ns in m_q possible!

Variations

Empirical Bayes – Tuning Priors

priors

$P(\bar{G}|\eta) \equiv$ probab of $\bar{G}$ given $\{\eta_i\}$

$$\propto \int \Pi \frac{d\rho}{\sqrt{2\pi}\,\tilde{\sigma}_\rho} \; e^{-\chi^2_a/2}$$

$$\rightarrow \left[\prod_i \frac{1}{\tilde{\sigma}_{\rho_i}}\right] \sqrt{\det C_{ij}} \; e^{-\chi^2_{a,\min}/2}$$

in Gaussian approx'n

$\Rightarrow$ vary η_i to increase $P(\bar{G}|\eta)$

↳ using $\bar{G}$ data to suggest reasonable values for η_i

↳ OK for 1 or 2 parameters

but complete nonsense if every η_i optimized

Non-Parametric Fits – Max. Ent.

see Oevers et al hep-lat/0009031
Yamazaki et al " /0105030

Eg/ Fit $\bar{G}$ to

$$G(t) \equiv \int d\omega\, e^{-\omega t}\, \rho(\omega)$$

($\rho(\omega)$: spectral fcn)

- Fit parameters: $\rho(\omega_j)$ for $\omega_j = 0.01 \times j$ & $j = 0, 1, 2 \cdots 750$

  ⇒ "non parametric"

  ⇒ # fit parameters ≫ # data points

  ⇒ Prior is essential

  ⇒ $\rho(\omega)$ has peaks at bd state energies ....

- Max Ent ⇒ useful framework for designing priors

- Properties of $\rho(\omega)$ (source = sink)
  1) $\rho(\omega)$ has random noise (M.C.)
  2) $\rho(\omega) > 0$
  3) $\rho(\omega)\, d\omega$ = physically interesting (additive)

$\Rightarrow$ think of $\rho(\omega)$ as a probability density

- Most probable probab. density is one with <u>largest entropy</u>

"$\Rightarrow$"

$$P(\rho) \propto e^{\alpha S(\rho)}$$

α = input parameter

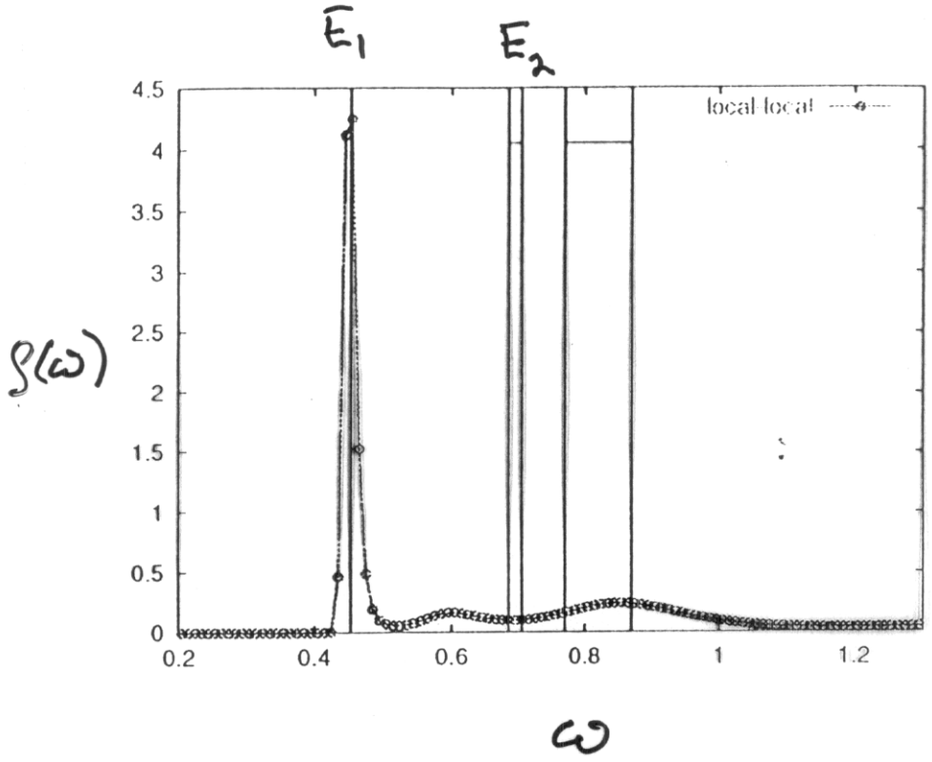
where

$$S \equiv \int d\omega \left[\rho(\omega) - m(\omega) - \rho(\omega) \ln \frac{\rho}{m} \right]$$

$\rho = m$ if no data.

↳ $m(\omega)$ = input.

$\Rightarrow$ constrained fit gives $\rho(\omega_j)$ $\forall j$.



Oewers et al

↳ same Y data as earlier

N.B.

- MaxEnt much less accurate than constrained fits to $G = \sum A_n e^{-E_n t}$.

  → uses much less prior information, but should use info if have it.

- Useful when there is no simple parametrization — e.g., in image processing

  ↳ many variations on $e^{\alpha S}$ used in applications

Conclusions

- Constrained / Bayes fits: major improvement
  - fits all data ⇒ more info, smaller σ's
  - *** includes sys. errors from ∞ of unconstrained terms
  - can combine info from other sim'ns via priors

- useful for
  - correlator analyses
  - Part'n the fits to M.C.
  - Extrapolations in M_Q or L or a ...
    - ↳ order of extrapol'n increases automatically as warranted by increased statistics.
    - ***

- Easy & Fun for Lattice experts.

Books:

D. Sivia "Data Analysis - A Bayesian Tutorial" (easy)

B. Carlin "Bayes & Empirical Bayes Methods for Data Analysis" (less easy)