Bayesian Analysis of Parton Densities

Glen Cowan **g.cowan@rhul.ac.uk**

Physics Department, Royal Holloway, University of London

In collaboration with Clare Quarman

I. Statement of the problem

II. Why not frequentist methods?

III. Bayesian analysis -- our approach to & difficulties with:

i) model uncertainties

ii) goodness-of-fit

iii) MCMC

Parton densities for statisticians

Parton densities are a set of several functions described by around n = two dozen parameters

$$\theta = (\theta_1, ..., \theta_n).$$

The functions have (roughly) the form

$$f(x) = a\, x^b\, (1 - x)^c\, (1 + \text{corrections} + ...).$$

Given a set of parton density parameters θ, we can predict the expected value of a measurable quantity $y(\theta)$ (non-trivial calculation).

The prediction cannot be calculated rigorously from the fundamental theory (QCD). Several approximations enter, e.g., perturbation theory to a limited order.

Constraining the parton densities

Many measurements have been carried out that constrain the parton densities. The existing measurements constitute a set of hundreds (perhaps even a few thousand) numbers

$$\boldsymbol{y} = (y_1, \ldots, y_m).$$

The y_i are usually taken to be normally distributed. The standard deviations σ_i are estimated for each measurement.

Gaussian model not justified in all cases (need longer tails).

Subsets of the y_i correlated (covariances available).

Systematic biases can be significant -- some info on this reported with measurement but can be very uncertain. E.g. measurements from different experiments of same quantity don't always agree.

Frequentist approach

Most parton density analyses up to now have used Method of Least Squares to estimate the parameters.

Goodness-of-fit typically poor; a number of important measurements incompatible.

Variances of estimators from e.g.

$$\chi^2(\theta) = \chi^2_{\text{min}} + 1$$

are typically small, and do not reflect full uncertainty on the parton densities. Try e.g.

$$\chi^2(\theta) = \chi^2_{\text{min}} + \text{some big number (50, 100?)}$$

Difficult to quantify model uncertainties

?Bayesian analysis.

What we want to do with parton densities

We are not primarily interested in the parton density parameters θ themselves, but rather we want to predict values for quantities that have not yet been (but will soon be) measured.

Our goal is to quantify the uncertainty in the prediction for a measurable quantity *y(*θ*)*.

Find posterior pdf for θ

$$p(\vec{\theta}) \propto L(\vec{y}|\vec{\theta})\pi(\vec{\theta})$$

For some new measurable quantity *z(*θ*)* determine *p(z)* e.g. by MCMC from *p(*θ*)*.

Glen Cowan
RHUL Physics

Quantifying model uncertainties

Some of the model uncertainties stem from using perturbation theory to a fixed order. Very roughly,

$$y = a(\theta)\ (b\alpha + c\alpha^2 + \ldots)$$

The expansion parameter α is small, itself uncertain. Here suppose b and c are known, but higher order terms not yet calculated.

Try to quantify uncertainty due to higher order terms by taking e.g.

$$y = a(\theta)\ (b\alpha + c\alpha^2 + d\,\alpha^3)$$

Assign prior $\pi(d)$ using some guesses of people who have experience computing e.g. b and c.

We can play a similar game for a variety of model uncertainties.

Assigning priors

Some rough guesses for the parton density parameters exist; use these as a guide.

Parameter set θ enlarged to include perhaps an even greater number of nuisance parameters ν:

Priors from interrogation of theoretical physicists.

Correlations for $\pi_\nu(\nu)$ non-trivial

Problem now of even higher dimension (up to, say, 100). Can we trust anyone's intuition in a high-dim. space?

For physics community, emphasise the "if-then" nature of the result:

"If your uncertainty in ν is such-and-such, then your corresponding uncertainty in the observable z is so-and-so."

Approach to goodness-of-fit

One source of uncertainty is the parametric form of the parton density function, e.g.,

$$f(x) = ax^b(1-x)^c(1+d\sqrt{x}-ex)$$

Suggestion by M. Goldstein (Durham) -- take, e.g.,

$$f(x) = ax^b(1-x)^c(1+\ldots)+r(x)$$

"Residual function" *r(x)* very flexible, e.g., superposition of Bernstein polynomials, coefficients ν_i:

$$r(x) = \sum_i \nu_i B_i(x)$$

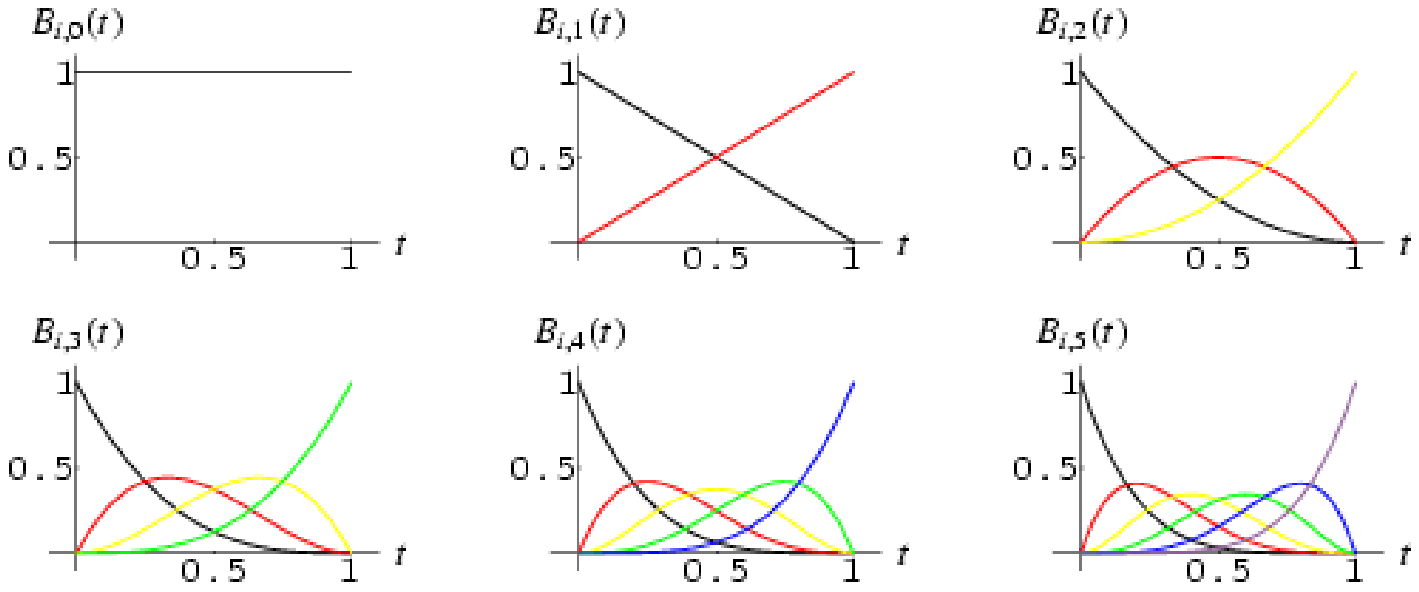
Prior for ν_i concentrated around 0, width chosen to reflect uncertainty in *f(x)* (roughly a couple of percent).

Residual function

So far residual function based on Bernstein polynomials.

$$B_{i,n}(t) = \binom{n}{i} t^i (1-t)^{n-i},$$

Test done with $B_{i,4}$, eventually need higher order (?)



mathworld.wolfram.com

Glen Cowan
RHUL Physics

Test with Bernstein polynomials

Test example with residual function

$$r(x) = \sum_{i=0}^{4} \nu_i B_{i,4}(x)$$

For prior take

$$\vec{\nu} \sim N(0, V)$$

$$V_{ij} = \sigma_\nu^2 \rho_{ij}\,, \qquad \sigma_\nu \approx 0.03$$

$$\rho_{ij} = 1 - (\Delta x_{ij}/\Delta x_{\text{max}})^\gamma = 1 - \left(\frac{|i-j|}{n\Delta x_{\text{max}}}\right)^\gamma$$

Set γ to give desired correlation length (for now $\gamma = 1$).

Test with Bernstein polynomials, cont.

Generate simulated measurements y_i, from

$$y_i \sim N(\mu_i, \sigma_i^2)$$

(1) $$\mu_i = a + bx_i$$

(2) $$\mu_i = a + bx_i + cx_i^2$$

To compute likelihood, use only with linear relation, but include residual function *r(x):*

$$\mu_i = a + bx_i + r(x_i)$$

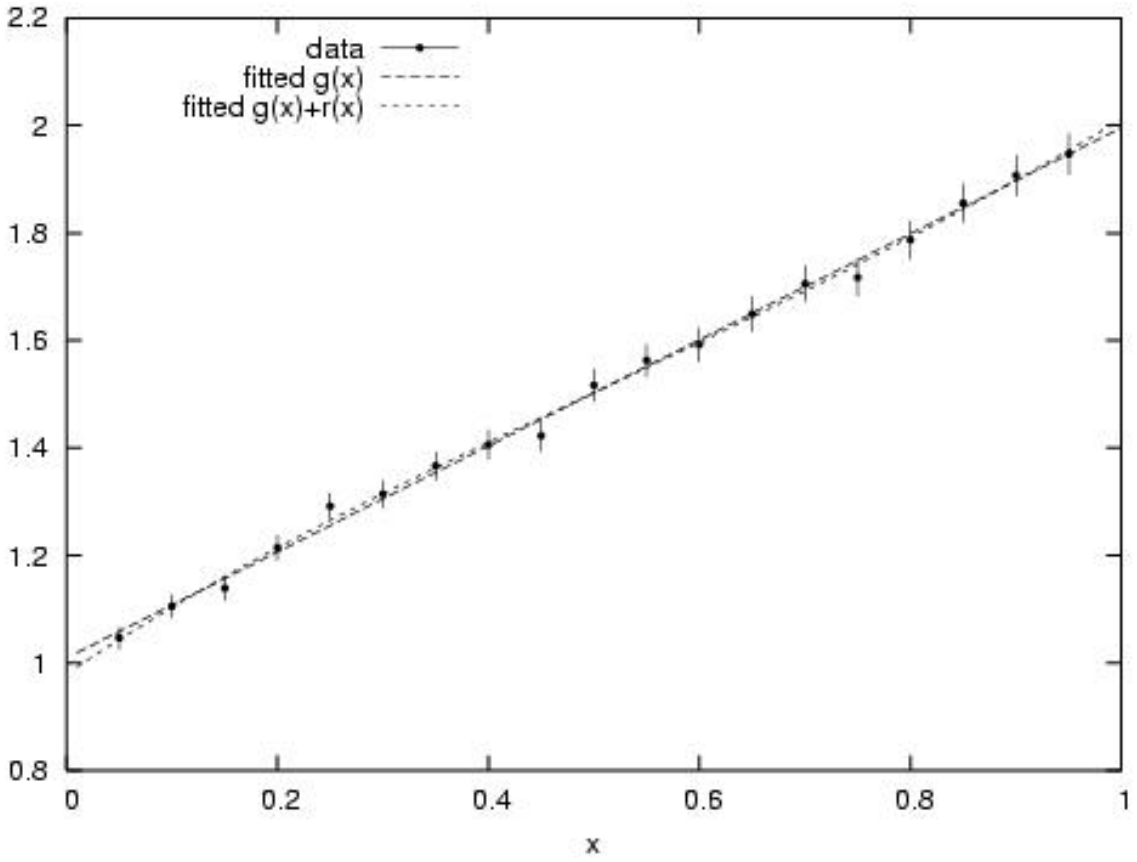
For (1) expect good fit, $p(\nu)$ should be narrower than $\pi(\nu)$.

From (2) bad fit, expect probabilities for large ν_i to grow.

Results of “good fit”

Generate points from line, model is also a line.

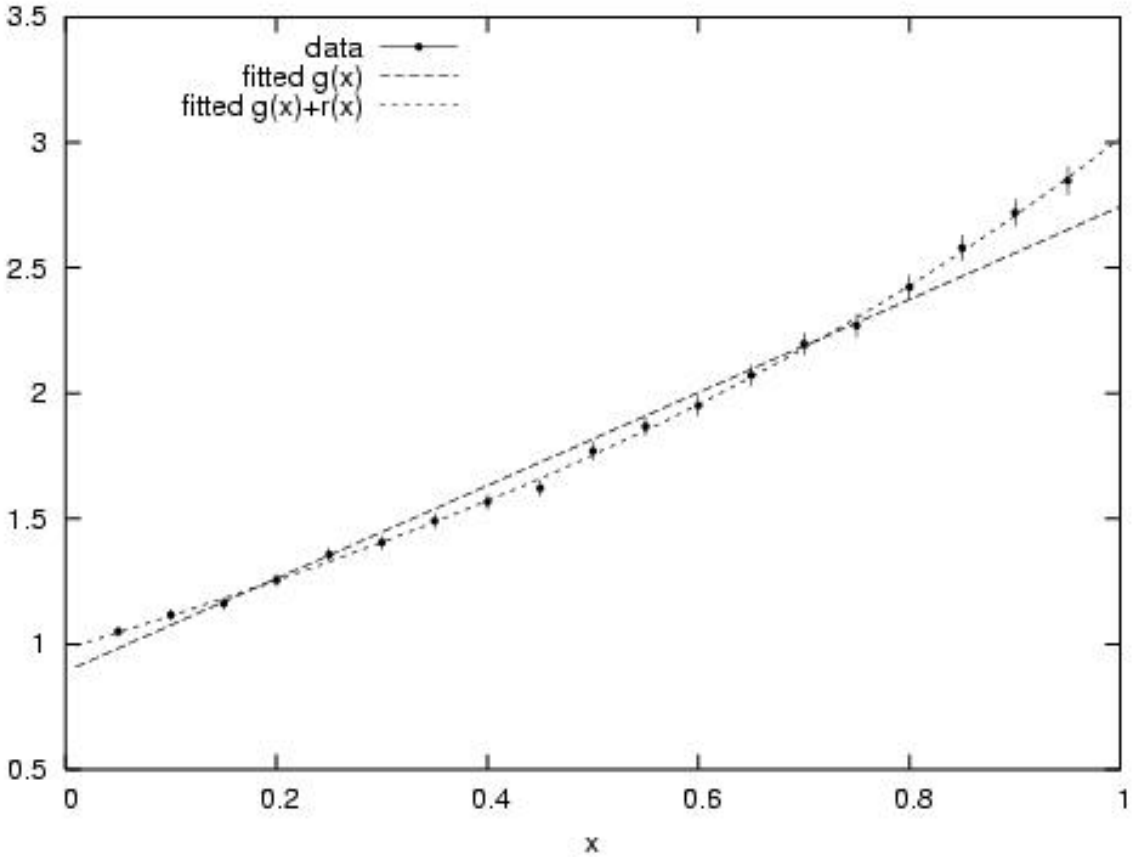
Curves shown with ML estimates.



Results of “bad fit”

Generate points from parabola, fitted model is a line.

With residual function, good fit.

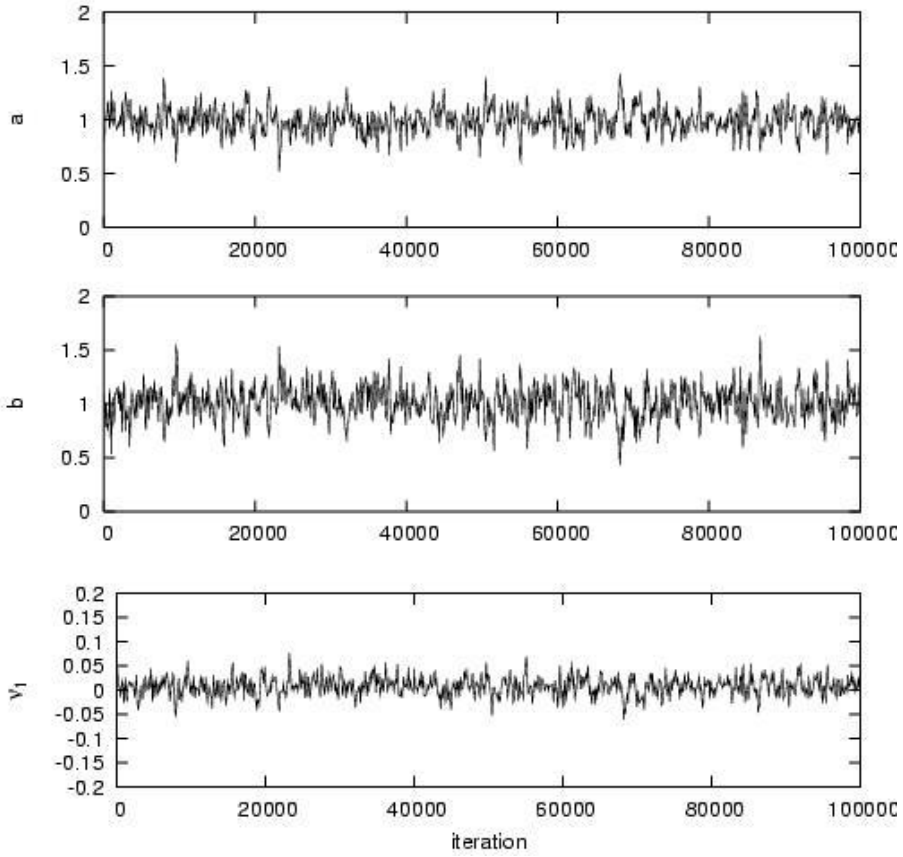


Sampling from $p(\theta)$

Sample posterior pdf with MCMC (Metropolis-Hastings):

Gaussian trial density, q, correlations match L (better: p)
Scale width of q to give acceptance fraction ~ 0.4

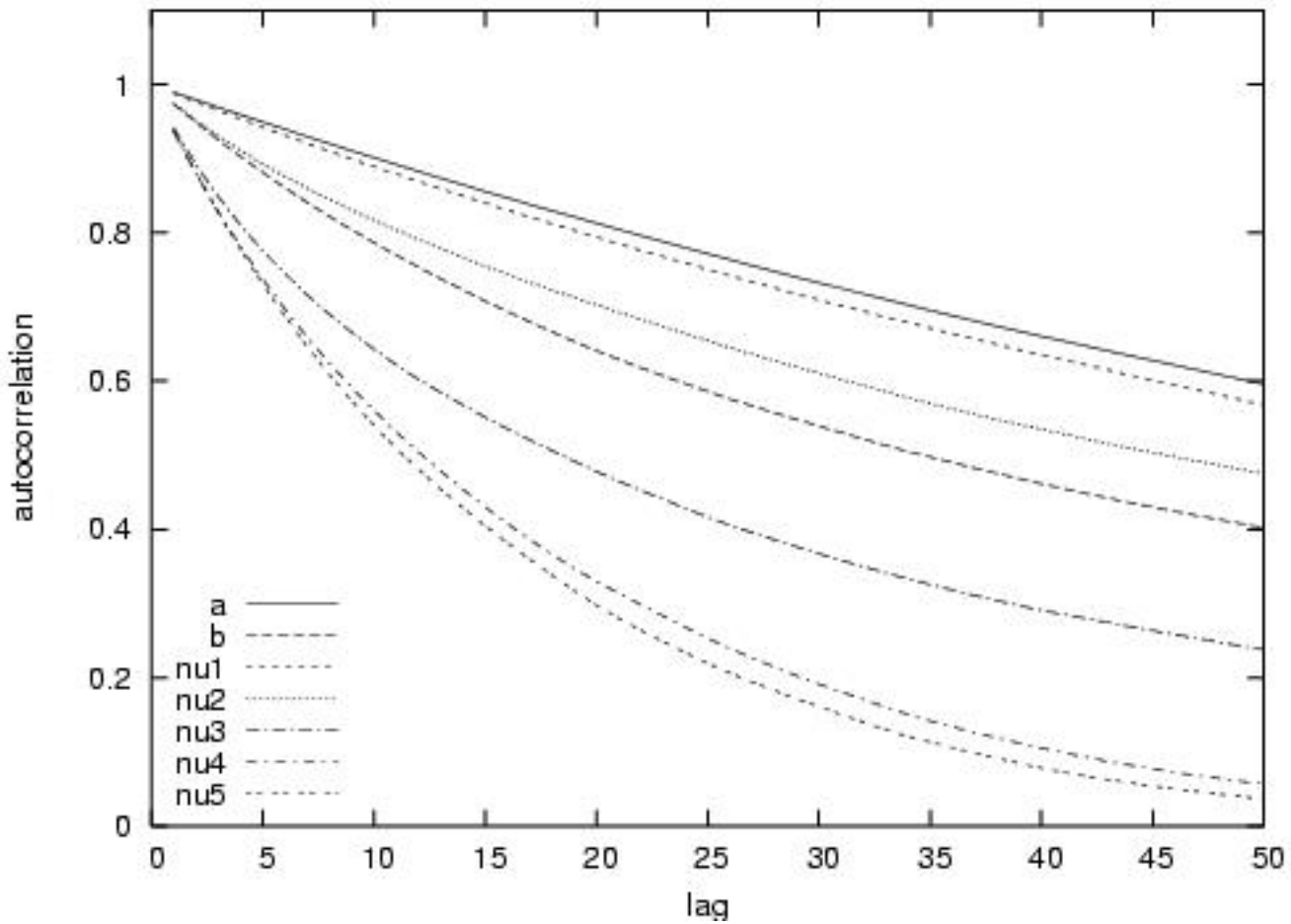
Some trace plots:



Glen Cowan
RHUL Physics

MCMC sampling of $p(\theta)$

Autocorrelation:



Glen Cowan
RHUL Physics

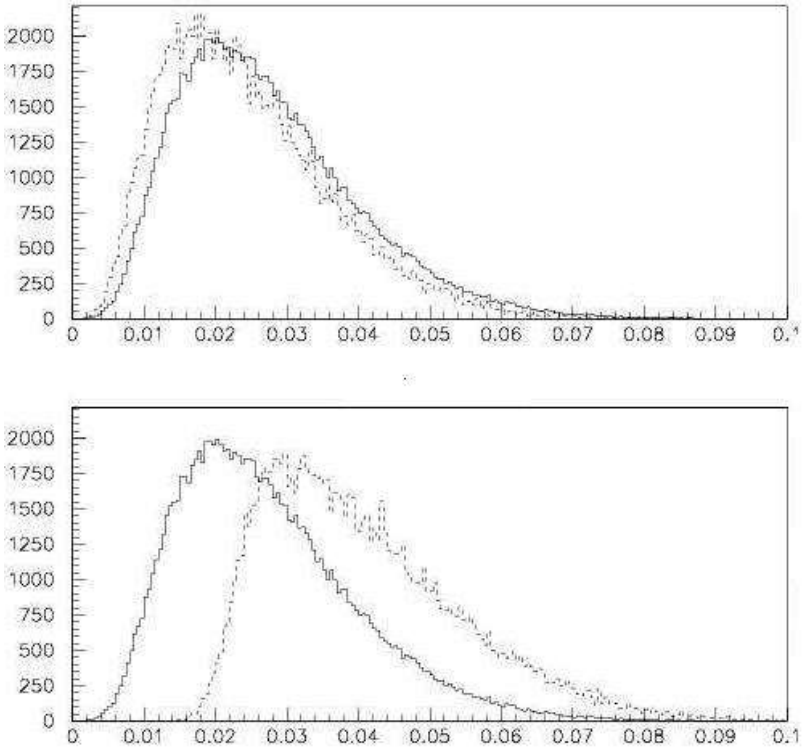
Results of test analysis

To quantify goodness-of-fit, look at prior/posterior distribution of

$$\nu_{\mathrm{rms}} = \left[\frac{1}{5}\sum_{i=0}^{4}\nu_i^2\right]^{1/2}$$

solid = prior

dashed = posterior



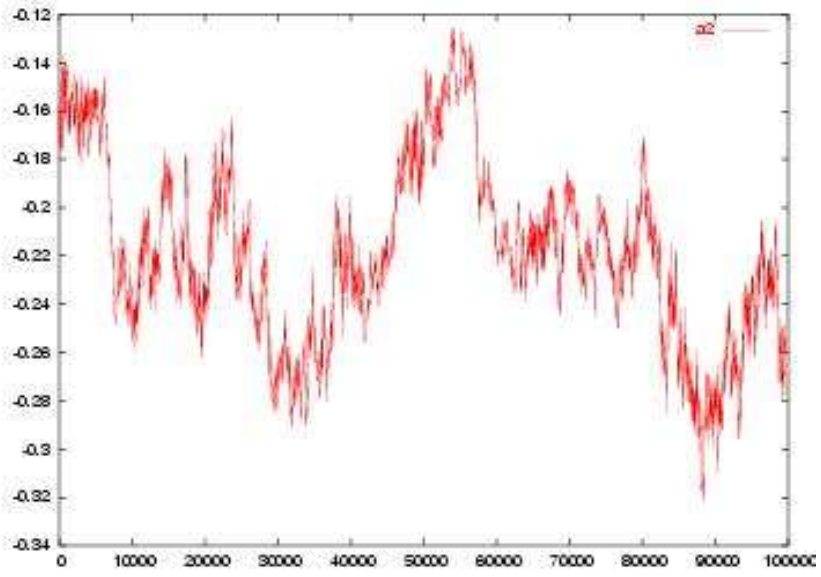
good fit

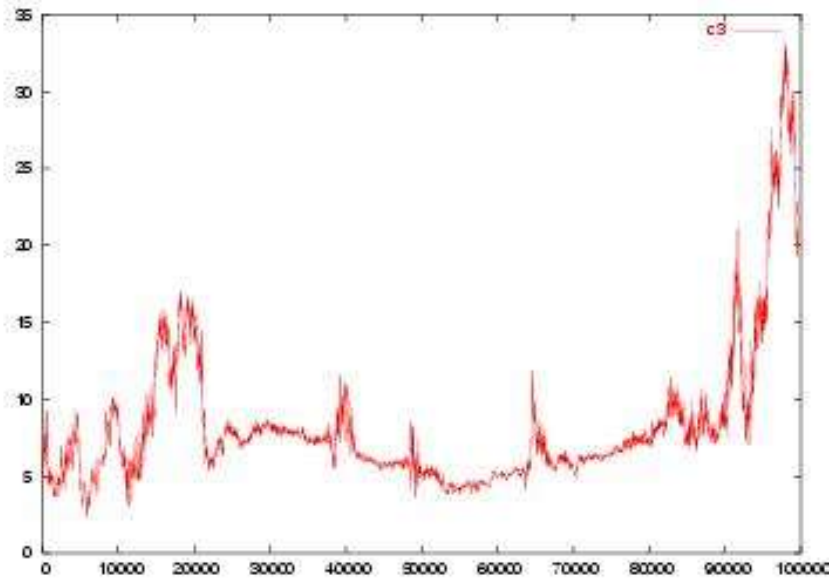
bad fit

Glen Cowan
RHUL Physics

Difficulties with MCMC

Test problems look OK. For the "real problem", MCMC does not converge well:





This took 4.5 hrs CPU, did not include residual function.
How can we improve our MCMC?

Least Squares, systematic errors, etc.

Most analyses have used Least Squares (gives maximum likelihood estimators if data Gaussian distributed).

In addition to the measurements and their standard deviations, experimenters typically report a "one standard deviation" systematic error (plus correlations, if several measurements).

This is sometimes used in a generalized Least Squares; for covariance matrix use $V = V_{\text{stat}} + V_{\text{sys}}$.

With several measurements of the same quantity, we find outliers more often than a Gaussian distribution for the data should allow. Interpretation is that the experiment underestimated its systematic error.

Need to use a sampling distribution for the data with longer tails.

A better distribution for the data

Suppose each measurement reports a result *y* and an estimate of its standard deviation σ_{stat}.

In addition the experimenter tries to report a systematic error σ_{sys} such that

$$x_i = \frac{y_i - \mu_i}{\sigma_i} \sim N(0,1)$$

with $$\sigma = \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{sys}}^2}$$

In addition to the reported numbers, each experiment is characterized by a value *s* (unreported), by which σ is mis-estimated.

Glen Cowan
RHUL Physics

Distribution of *y* and *s*

Take joint pdf for *y* and *s*

$$f(y,s) = \frac{1}{\sqrt{2\pi}s\sigma} e^{-(y-\mu)^2/2(s\sigma)^2} f_s(s)$$

$$f_s(s) = (1-q)\delta(s-1) + qg(s;\alpha,\beta)$$

where $g(s;\alpha,\beta)$ is a Gamma distribution.

With probability 1-q the experimenter reports the correct error bar.

With probability q the error bar is significantly incorrect. Choose parameters of Gamma dist. such that $V[s] = 1$, $E[s] = 2$.

Integrate to find $f_y(y) = \int f(y,s)\, ds$

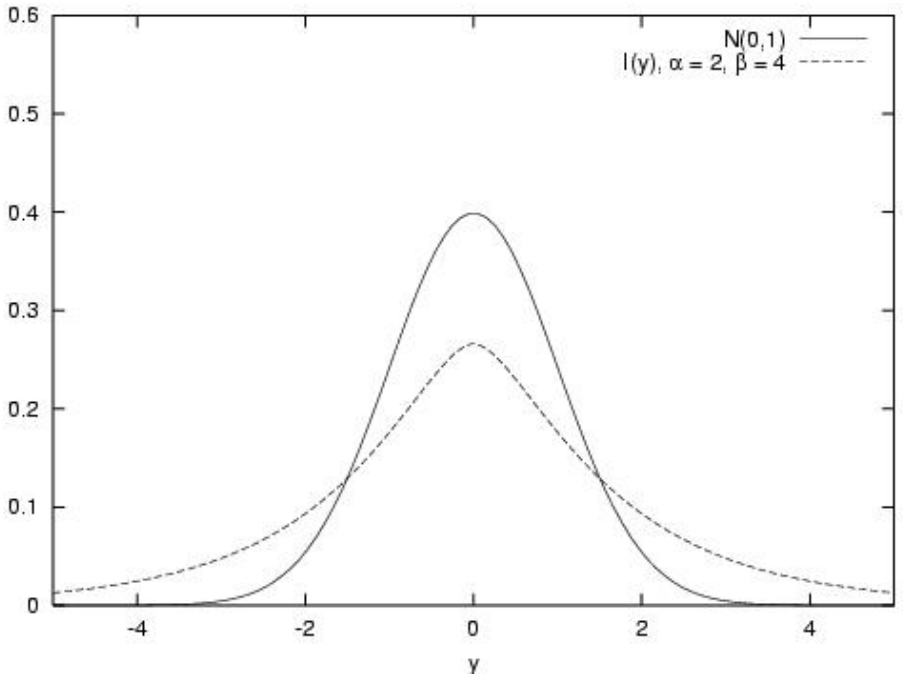
Marginal distribution of y

As s is not reported, we must find

$$f_y(y) = \int f(y, s)\, ds$$

Integrating (numerically) gives a curve with longer tails than a Gaussian:

Use in either a frequentist or Bayesian analysis.



Glen Cowan
RHUL Physics

Wrapping up

Over the past year, primary activity has been assembling the computational machinery -- still much difficulty with MCMC.

An important task for the near future will be establishing reasonable priors; rely heavily on advice of particle theorists.

Some progress on goodness-of-fit (is this a “solved problem”?)

Some progress on a more appropriate distribution for the data (longer tails than Gaussian).

“Final product” should be a software package that will allow the user to compute the posterior pdf for the prediction of a measurement; width should reflect uncertainty due to parton densities.