Bayesian statistical methods for parton analyses

Glen Cowan
Royal Holloway, University of London

g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

In collaboration with Clare Quarman

DIS 2006
Tsukuba
22 April, 2006

Outline

I. Data analysis difficulties for prediction of LHC observables

Some problems with frequentist statistical methods

II. Bayesian statistics

Quick review of basic formalism and tools

Application to:

incompatible data sets,
model (theoretical) uncertainties.

III. Prospects for LHC predictions

Some uncertainties in predicted cross sections

I. PDFs based on fits to data with:

imperfectly understood systematics,
not all data compatible.

II. Perturbative prediction only to limited order

PDF evolution & cross sections to NLO, NNLO...

III. Modelling of nonperturbative physics

parametrization of PDF at low Q^2,
details of flavour composition, ...

LHC game plan

$$\text{predicted } \sigma \left\{ \begin{array}{c} = \\ \neq \end{array} \right\} \text{measured } \sigma \rightarrow \text{New Physics}$$

Understanding uncertainties in predicted cross sections is a recognized Crucial Issue for LHC analyses, e.g.

extra dimensions, parton substructure, $\sin^2 \theta_W$

For LHC observables we have

$$\sigma_{\text{pred}} = \sum_{i,j} \int \int f_i(x_i) f_j(x_j) \sigma_{ij} \, dx_i \, dx_j$$

uncertainties in PDFs

uncertainties in parton cross sections

PDF fit (symbolic)

Given measurements: $y_i \pm \sigma_i^{\text{stat}} \pm \sigma_i^{\text{sys}}, \quad i = 1, \ldots, n$,

and (usually) covariances: V_{ij}^{stat}, V_{ij}^{sys} .

Predicted value: $\mu(x_i; \theta)$, expectation value $E[y_i] = \mu(x_i; \theta) + b_i$

control variable — PDF parameters, α_s, etc. — bias

Often take: $V_{ij} = V_{ij}^{\text{stat}} + V_{ij}^{\text{sys}}$

Minimize $\chi^2(\theta) = (\vec{y} - \vec{\mu}(\theta))^T V^{-1} (\vec{y} - \vec{\mu}(\theta))$

Equivalent to maximizing $L(\theta)$ » $e^{-\chi^2/2}$, i.e., least squares same as maximum likelihood using a Gaussian likelihood function.

Uncertainties from PDF fits

If we have incompatible data or an incorrect model, then minimized χ^2 will be high, but this does not automatically result in larger estimates of the PDF parameter errors.

Frequentist statistics provides a rule to obtain standard deviation of estimators (1σ statistical errors):

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

but in PDF fits this results in unrealistically small uncertainties.

Try e.g. $\chi^2 = \chi^2_{\text{min}} + 50, 75, 100$?

The problem lies in the application of a rule for statistical errors to a situation dominated by systematics & model uncertainties.

→ Try Bayesian statistical approach

The Bayesian approach

In Bayesian statistics we can associate a probability with a hypothesis, e.g., a parameter value θ.

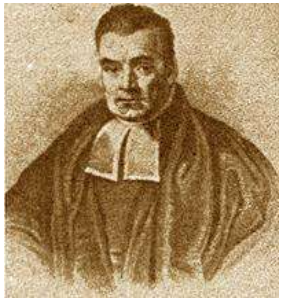
Interpret probability of θ as 'degree of belief' (subjective).

Need to start with 'prior pdf' $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Our experiment has data x, → likelihood function $L(x|\theta)$.

Bayes' theorem tells how our beliefs should be updated in light of the data x:

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta')\pi(\theta')\,d\theta'} \propto L(x|\theta)\pi(\theta)$$



Rev. Thomas Bayes

Posterior pdf $p(\theta|x)$ contains all our knowledge about θ.

A possible Bayesian analysis

Take $L(\vec{y}|\vec{\theta},\vec{b}) \sim \exp\left[-\frac{1}{2}(\vec{y}-\vec{\mu}(\theta)-\vec{b})^T\, V_{\text{stat}}^{-1}\,(\vec{y}-\vec{\mu}(\theta)-\vec{b})\right]$

$\pi_b(\vec{b}) \sim \exp\left[-\frac{1}{2}\,\vec{b}^T\, V_{\text{sys}}^{-1}\,\vec{b}\right]$

$\pi_\theta(\theta) \sim \text{const.}$

Joint probability for all parameters

and use Bayes' theorem: $p(\theta,\vec{b}|\vec{y}) \propto L(\vec{y}|\theta,\vec{b})\pi_\theta(\theta)\pi_b(\vec{b})$

To get desired probability for θ, integrate (marginalize) over b:

$$p(\theta|\vec{y}) = \int p(\theta,\vec{b}|\vec{y})\, d\vec{b}$$

→ Posterior is Gaussian with mode same as least squares estimator, σ_θ same as from $\chi^2 = \chi^2_{\text{min}} + 1$. (Back where we started!)

Marginalizing with Markov Chain Monte Carlo

In a Bayesian analysis we usually need to integrate over some (or all) of the parameters, e.g.,

$$p(\theta|\vec{y}) = \int p(\theta, \vec{b}|\vec{y})\, d\vec{b}$$

$$p(\mu|\vec{y}) = \int \delta(\mu - \mu_{\mathsf{pred}}(\theta))\, p(\theta, \vec{b}|\vec{y})\, d\vec{b}\, d\theta$$

Probability density for prediction of observable $\mu(\theta)$

Integrals often high dimension, usually cannot be done in closed form or with acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized Bayesian computation. (Google words: Metropolis-Hastings, MCMC)

Produces a correlated sequence of points in the sampled space. Correlations here not fatal, but stat. error larger than naive $\sqrt{n}$.

Systematic uncertainty and nuisance parameters

In general we can describe the data better by including more parameters in the model (nuisance parameters), e.g.,

Low Q^2 PDF: $ax^b(1-x)^c(1+d\sqrt{x}+ex+\ldots)$
Unknown cofficients of higher order, higher twist terms, ...
Experimental biases, ...

But, more parameters → correlations → bigger errors.

Bayesian approach: include more parameters along with prior probabilities that reflect how widely they can vary.

Difficult (impossible) to agree on priors but remember 'if-then' nature of result. Usefulness to community comes from sensitivity analysis:

Vary prior, see what effect this has on posterior.

The Full Bayesian Machine

A full Bayesian PDF analysis could involve:

the usual two dozen PDF parameters,
a bias parameter for each systematic,
more parameters to quantify model uncertainties,...

as well as a meaningful assignment of priors

consultation with experimenters/theorists

and finally an integration over the entire parameter space to extract the posterior probability for a parameter of interest, e.g., a predicted cross section:

ongoing effort, primary difficulties with MCMC

The error on the error

Some systematic errors are well determined

Error from finite Monte Carlo sample

Some are less obvious

Do analysis in n 'equally valid' ways and
extract systematic error from 'spread' in results.

Some are educated guesses

Guess possible size of missing terms in perturbation series;

vary renormalization scale $(\mu/2 < Q < 2\mu$?)

Can we incorporate the 'error on the error'?

(cf. G. D'Agostini 1999; Dose & von der Linden 1999)

Motivating a non-Gaussian prior $\pi_b(b)$

Suppose now the experiment is characterized by

$$y_i, \quad \sigma_i^{\text{stat}}, \quad \sigma_i^{\text{sys}}, \quad s_i, \quad i = 1, \ldots, n \ ,$$

where s_i is an (unreported) factor by which the systematic error is over/under-estimated.

Assume correct error for a Gaussian $\pi_b(b)$ would be $s_i\sigma_i^{\text{sys}}$, so

$$\pi_b(b_i) = \int \frac{1}{\sqrt{2\pi}s_i\sigma_i^{\text{sys}}} \exp\left[-\frac{1}{2}\frac{b_i^2}{(s_i\sigma_i^{\text{sys}})^2}\right] \pi_s(s_i)\, ds_i$$

Width of $\pi_s(s_i)$ reflects 'error on the error'.

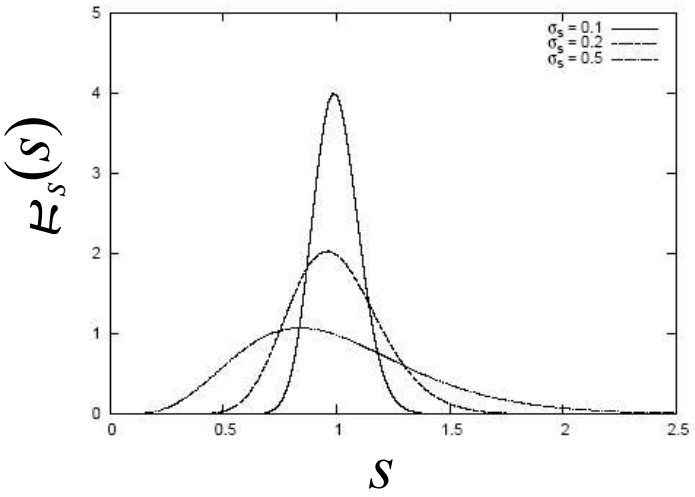
Error-on-error function $\pi_s(s)$

A simple unimodal probability density for $0 < s < 1$ with adjustable mean and variance is the Gamma distribution:

$$\pi_s(s) = \frac{a(as)^{b-1}e^{-as}}{\Gamma(b)}$$

mean = b/a
variance = b/a^2

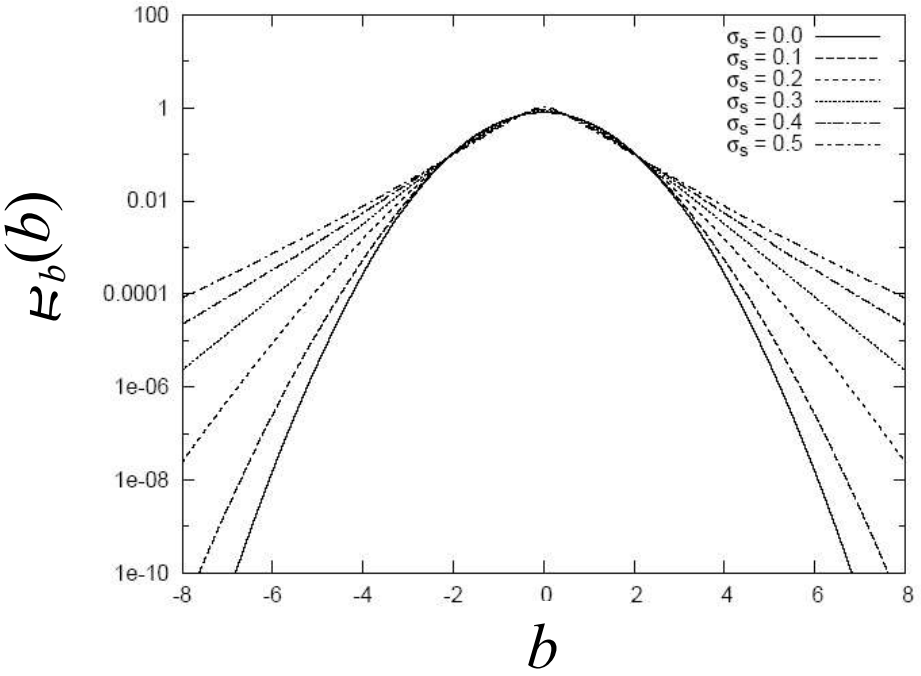
Want e.g. expectation value of 1 and adjustable standard deviation σ_s , i.e., $a = b = 1/\sigma_s^2$



In fact if we took $\pi_s(s)$ » *inverse Gamma*, we could integrate $\pi_b(b)$ in closed form (cf. D'Agostini, Dose, von Linden). But Gamma seems more natural & numerical treatment not too painful.

Prior for bias $\pi_b(b)$ now has longer tails

$$\pi_b(b_i) = \int \frac{1}{\sqrt{2\pi} s_i \sigma_i^{\text{sys}}} \exp\left[-\frac{1}{2}\frac{b_i^2}{(s_i\sigma_i^{\text{sys}})^2}\right] \pi_s(s_i)\, ds_i$$



Gaussian ($\sigma_s = 0$) $\quad P(|b| > 4\sigma_{\text{sys}}) = 6.3 \times 10^{-5}$

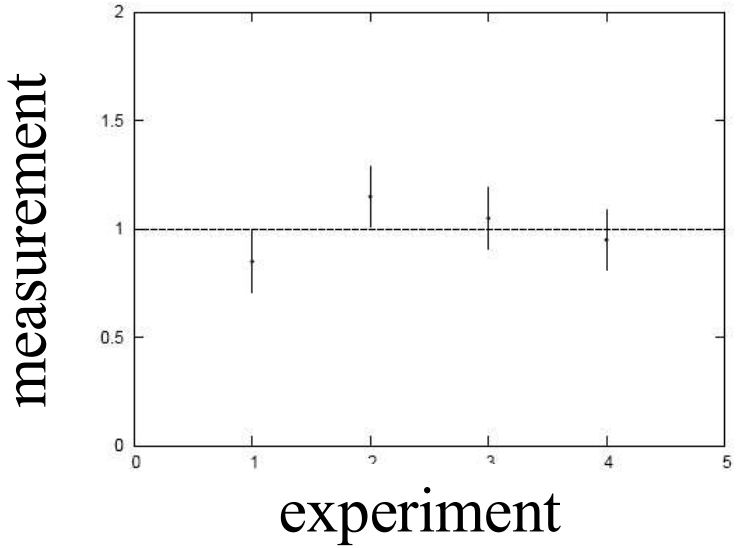
$\sigma_s = 0.5$ $\quad P(|b| > 4\sigma_{\text{sys}}) = 0.65\%$

A simple test

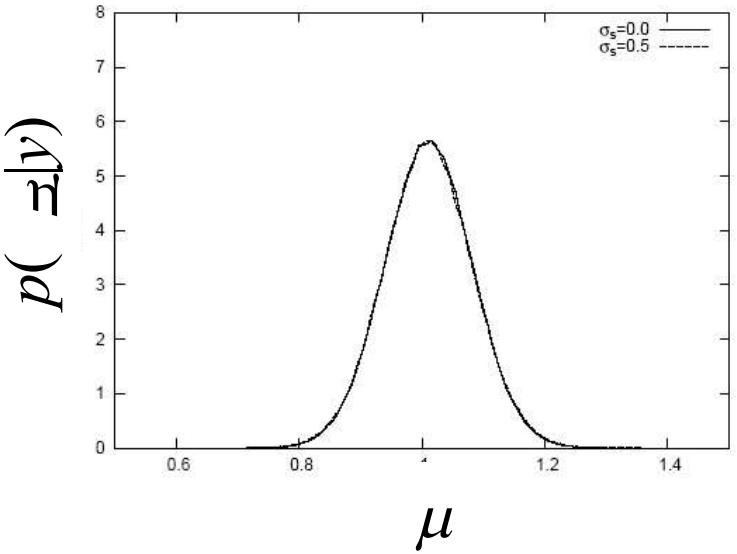
Suppose fit effectively averages four measurements.

Take $\sigma_{sys} = \sigma_{stat} = 0.1$, uncorrelated.

Case #1: data appear compatible



Posterior $p(\mu|y)$:



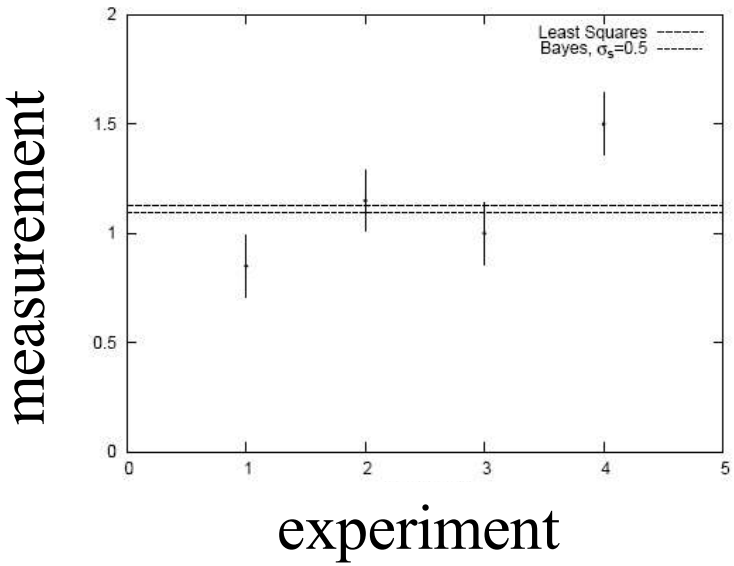
Usually summarize posterior $p(\mu|y)$ with mode and standard deviation:

$\sigma_s = 0.0: \quad \hat{\mu} = 1.000 \pm 0.071$

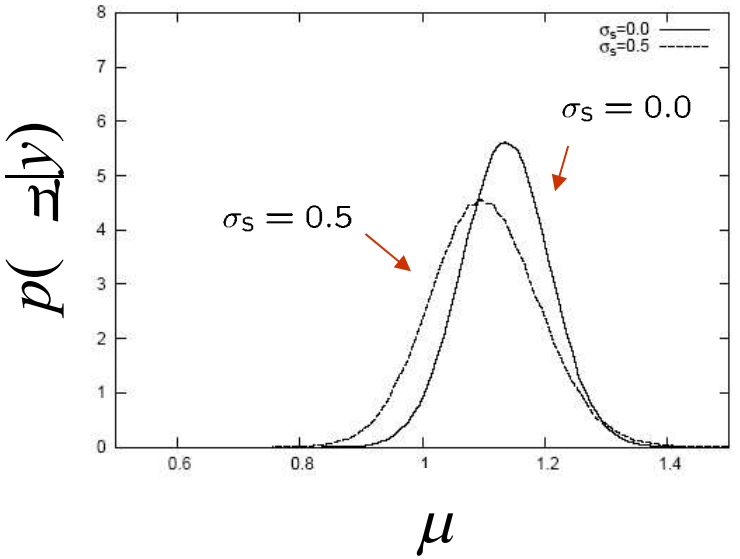
$\sigma_s = 0.5: \quad \hat{\mu} = 1.000 \pm 0.072$

Simple test with inconsistent data

Case #2: there is an outlier



Posterior $p(\mu|y)$:



$$\sigma_s = 0.0: \quad \hat{\mu} = 1.125 \pm 0.071$$

$$\sigma_s = 0.5: \quad \hat{\mu} = 1.093 \pm 0.089$$

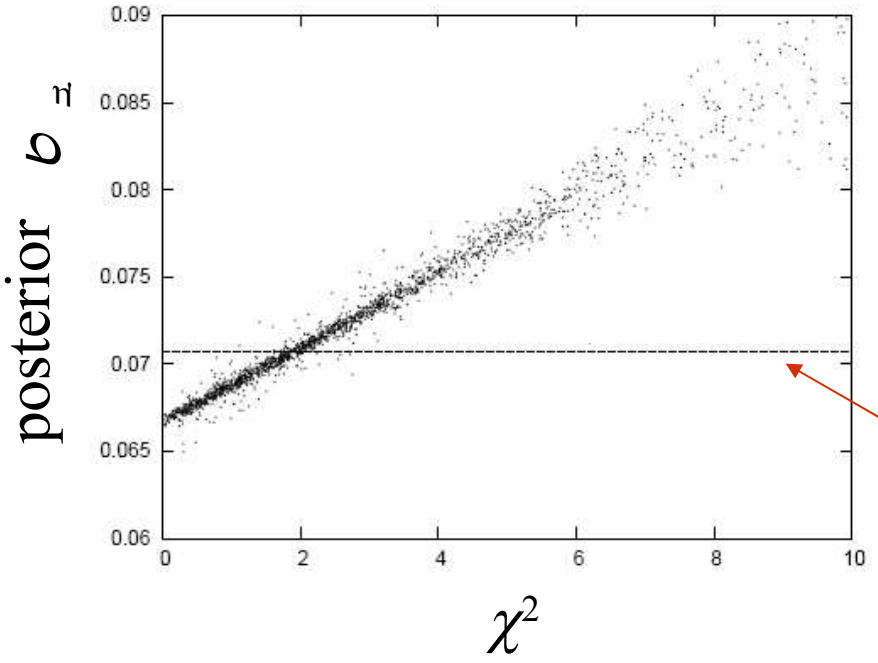
→ Bayesian fit less sensitive to outlier.

→ Error now connected to goodness-of-fit.

Goodness-of-fit vs. size of error

In LS fit, value of minimized χ^2 does not affect size of error on fitted parameter.

In Bayesian analysis with non-Gaussian prior for systematics, a high χ^2 corresponds to a larger error (and vice versa).



2000 repetitions of experiment, $\sigma_s = 0.5$, here no actual bias.

σ_μ from least squares

Is this workable for PDF fits?

Straightforward to generalize to include correlations

Prior on correlation coefficients ~ $\pi(\rho)$

(Myth: $\rho = 1$ is "conservative")

Can separate out different systematic for same measurement

Some will have small σ_s, others larger.

Remember the "if-then" nature of a Bayesian result:

We can (should) vary priors and see what effect this has on the conclusions.

Uncertainty from parametrization of PDFs

Try e.g. $xf(x) = ax^b(1-x)^c(1+d\sqrt{x}+ex)$ (MRST)

or $xf(x) = ax^b(1-x)^c e^{d\cdot x}(1+e^e x)^f$ (CTEQ)

The form should be flexible enough to describe the data;
frequentist analysis has to decide how many parameters are justified.

In a Bayesian analysis we can insert as many parameters as we want, but constrain them with priors.

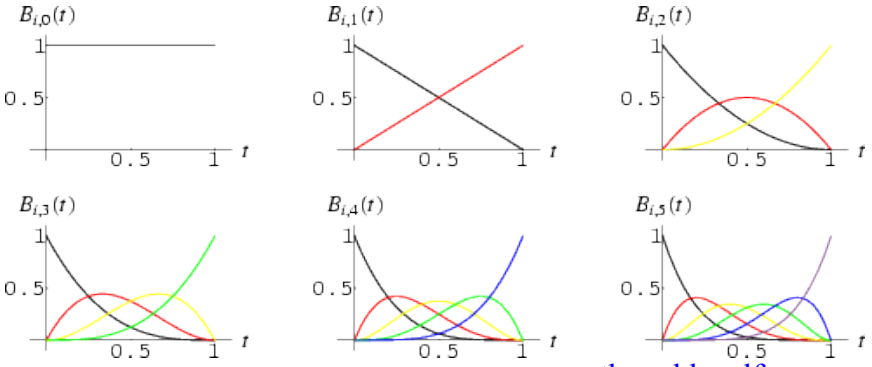
Suppose e.g. based on a theoretical bias for things not too bumpy, that a certain parametrization 'should hold to 2%'.

How to translate this into a set of prior probabilites?

Residual function

Try e.g. $xf(x) = ax^b(1-x)^c(1+\ldots) + r(x)$ ← 'residual function'

where $r(x)$ is something very flexible, e.g., superposition of Bernstein polynomials, coefficients ν_i: $r(x) = \sum_i \nu_i B_i(x)$



mathworld.wolfram.com

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

Assign priors for the ν_i centred around 0, width chosen to reflect the uncertainty in $xf(x)$ (e.g. a couple of percent).

→ Ongoing effort.

Wrapping up

A discovery at the LHC may depend crucially on assessing the uncertainty a predicted cross section.

Systematic uncertainties difficult to treat in frequentist statistics, often wind in with *ad hoc* recipes.

Bayesian approach tries to encapsulate the uncertainties in prior probabilities for an enlarged set of model parameters

Bayes' theorem says how where these parameters should lie in light of the data

Marginalize to give probability of parameter of interest (new tool: MCMC).

Very much still ongoing effort!

Extra slides

Some references

S.I. Alekhin, *Extraction of parton distributions and from DIS data with a Bayesian treatment of systematic errors*, Eur. Phys. J. C 10, 395-403

W. Giele, S. Keller and D. Kosower, *Parton Distribution Function Uncertainties*, hep-ph/0104052

G. D'Agostini, *Sceptical combination of experimental results: General considerations and application to ε'/ε*, CERN-EP/99-139, hep-ex/9910036

V. Dose and W. von der Linden, *Outlier tolerant parameter estimation*, in V. Dose *et al*. (eds.), Proc. of he XVIII International Workshop on Maximum Entropy and Bayesian Methods, Garching, 1998 (Kluwer Academic Publishers, Dordrecht, 1999; preprint in `www.ipp.mpg/de/OP/Datenanalyse/Publications/Papers/dose99a.ps`)

Marginalizing the posterior probability density

Bayes' theorem gives the joint probability for *all* the parameters. Crucial difference compared to freqentist analysis is ability to marginalize over the nuisance parameters, e.g.,

$$p(\theta|\vec{y}) = \int p(\theta, \vec{b}|\vec{y})\, d\vec{b}$$

Do e.g. with MC: sample full (θ, b) space and look at distribution only of those parameters of interest.

In the end we are interested not in probability of θ but of some observable $\mu(\theta)$ (e.g. a cross section), i.e.,

$$p(\mu|\vec{y}) = \int \delta(\mu - \mu_{\mathsf{pred}}(\theta))\, p(\theta, \vec{b}|\vec{y})\, d\vec{b}\, d\theta$$

Similarly do with MC: sample (θ, b) and look at distribution of $\mu(\theta)$.

Digression: marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x)\, d\theta_1 \ .$$

often high dimensionality and impossible in closed form,
also impossible with 'normal' acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized Bayesian computation.

Google for 'MCMC', 'Metropolis', 'Bayesian computation', ...

MCMC generates correlated sequence of random numbers:

cannot use for many applications, e.g., detector MC;
effective stat. error greater than $\sqrt{n}$.

Basic idea: sample multidimensional $\vec{\theta}$,
look, e.g., only at distribution of parameters of interest.

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n-dimensional pdf $p(\vec{\theta})$,

generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \ldots$

1) Start at some point $\vec{\theta}_0$

2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$

Proposal density $q(\vec{\theta}; \vec{\theta}_0)$
e.g. Gaussian centred
about $\vec{\theta}_0$

3) Form Hastings test ratio $\alpha = \min\left[1, \dfrac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)}\right]$

4) Generate $u \sim \mathsf{Uniform}[0, 1]$

5) If $u \leq \alpha$, $\vec{\theta}_1 = \vec{\theta}$, ← move to proposed point

else $\vec{\theta}_1 = \vec{\theta}_0$ ← old point repeated

6) Iterate

Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

For our purposes this correlation is not fatal, but statistical errors larger than naive $\sqrt{n}$.

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta})$

Test ratio is (*Metropolis*-Hastings): $\alpha = \min\left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)}\right]$

I.e. if the proposed step is to a point of higher $p(\vec{\theta})$, take it;

if not, only take the step with probability $p(\vec{\theta})/p(\vec{\theta}_0)$.

If proposed step rejected, hop in place.

Metropolis-Hastings caveats

Actually one can only prove that the sequence of points follows the desired pdf in the limit where it runs forever.

There may be a "burn-in" period where the sequence does not initially follow $p(\vec{\theta})$.

Unfortunately there are few useful theorems to tell us when the sequence has converged.

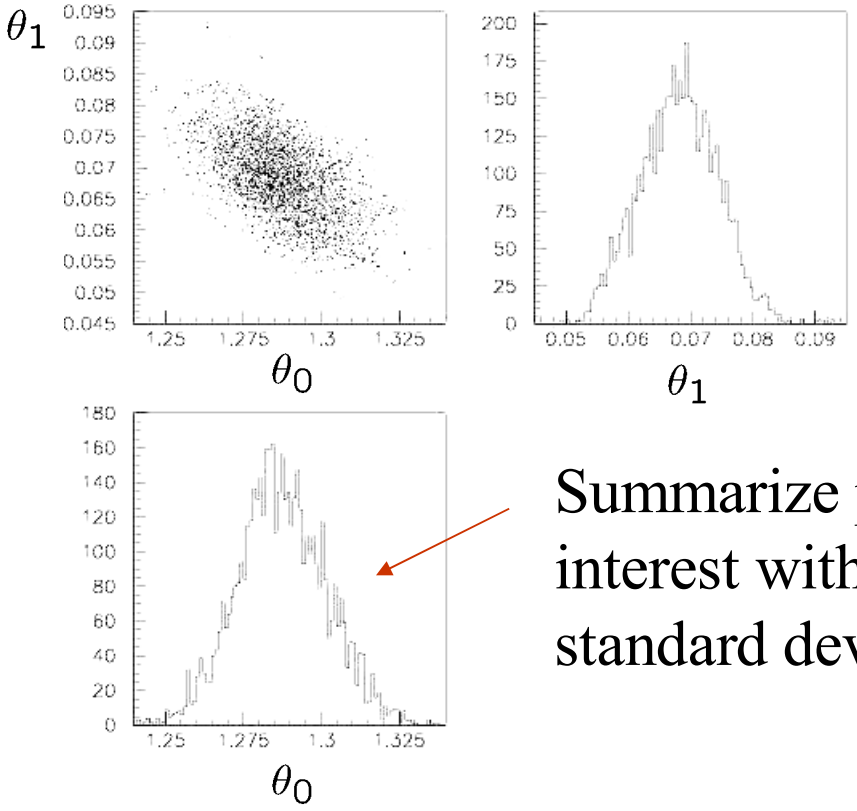
Look at trace plots, autocorrelation.

Check result with different proposal density.

If you think it's converged, try it again with 10 times more points.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:

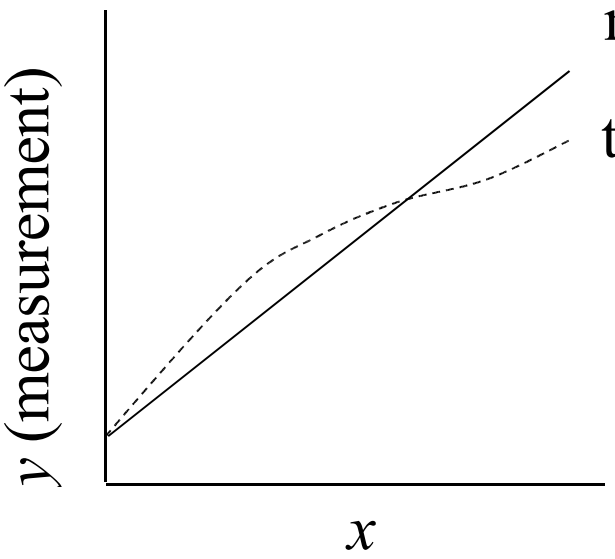


Summarize pdf of parameter of interest with, e.g., mean, median, standard deviation, etc.

Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

Bayesian approach to model uncertainty

Model can be made to approximate the truth better by including more free parameters.



model: $y = \alpha + \beta x$

truth: $y = \alpha + \beta x + \gamma x^2 + \varepsilon x^3 + \ldots$

systematic uncertainty

$\updownarrow$

nuisance parameters

In a frequentist analysis, the correlations between the fitted parameters will result in large errors for the parameters of interest.

In Bayesian approach, constrain nuisance parameters with priors.