# Introduction to Markov chain Monte Carlo (MCMC) and its role in modern Bayesian analysis

**Phil Gregory**

**University of British Columbia**

**March 2010**

# Outline

# What is Bayesian Probability Theory? (BPT)

**BPT = a theory of extended logic**

**Deductive logic is based on Axiomatic knowledge.**

**In science we never know any theory of nature is true because our reasoning is based on incomplete information.**

**Our conclusions are at best probabilities.**

**Any extension of logic to deal with situations of incomplete information (realm of inductive logic) requires a theory of probability.**

**A new perception of probability has arisen in recognition that the mathematical rules of probability are not merely rules for manipulating random variables.**

**They are now recognized as valid principles of logic for conducting inference about any hypothesis of interest.**

**This view of, ``Probability Theory as Logic'', was championed in the late 20th century by E. T. Jaynes.**

**"Probability Theory: The Logic of Science"**
**Cambridge University Press 2003**

**It is also commonly referred to as Bayesian Probability Theory in recognition of the work of the 18th century English clergyman and Mathematician Thomas Bayes.**

**Logic is concerned with the truth of propositions.**

**A proposition asserts that something is true.**

**Examples of propositions:**

$A \equiv$ "The newly discovered radio astronomy object is a galaxy."

$B \equiv$ "The measured redshift of the object is $0.150 \pm 0.005$ ."

$A \equiv$ "Theory $X$ is correct."

$\overline{A} \equiv$ "Theory $X$ is not correct."

$A \equiv$ "The frequency of the signal is between $f$ and $f + df$."

We will need to consider compound propositions like

*A,B* which asserts that propositions *A* and *B* are true

*A,B*|*C* asserts that propositions *A* and *B* are true given that proposition C is true

Rules for manipulating probabilities

Sum rule: $p(A \mid C) + p(\overline{A} \mid C) = 1$

Product rule: $p(A, B \mid C) = p(A \mid C)\, p(B \mid A, C)$

$= p(B \mid C)\, p(A \mid B, C)$

Bayes theorem:

$$p(A \mid B, C) = \frac{p(A \mid C)\, p(B \mid A, C)}{p(B \mid C)}$$

# How to proceed in a Bayesian analysis?

**Write down Bayes' theorem, identify the terms and solve.**

**Prior probability** **Likelihood**

$$p(H_i \mid D, I) = \frac{p(H_i \mid I) \times p(D \mid H_i, I)}{p(D \mid I)}$$

**Posterior probability that $H_i$ is true, given the new data D and prior information I**

**Normalizing constant**

**Every item to the right of the vertical bar | is assumed to be true**

**The likelihood $p(D \mid H_i, I)$, also written as $\mathcal{L}(H_i)$, stands for the probability that we would have gotten the data D that we did, if $H_i$ is true.**

**As a theory of extended logic BPT can be used to find optimal answers to well posed scientific questions for a given state of knowledge, in contrast to a numerical recipe approach.**

## Two basic problems

1. **Model selection (discrete hypothesis space)**

   **"Which one of 2 or more models (hypotheses) is most probable given our current state of knowledge?"**

**e.g.**

- **Hypothesis or model $M_0$ asserts that the star has no planets.**
- **Hypothesis $M_1$ asserts that the star has 1 planet.**
- **Hypothesis $M_i$ asserts that the star has *i* planets.**

2. **Parameter estimation (continuous hypothesis)**

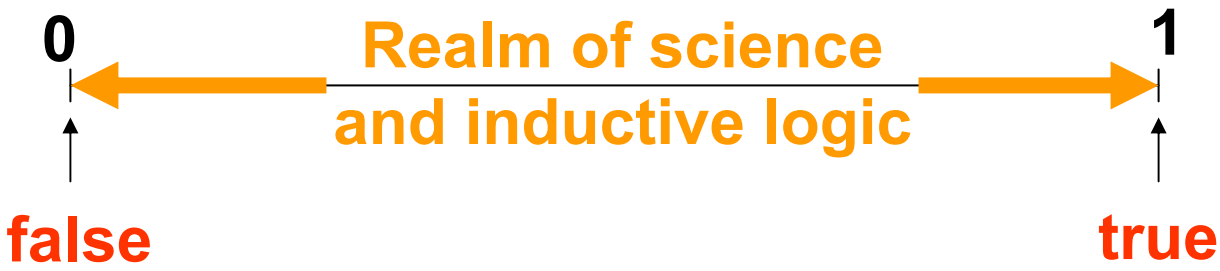
   **"Assuming the truth of $M_1$, solve for the probability density distribution for each of the model parameters based on our current state of knowledge."**

**e.g.**

- **Hypothesis H asserts that the orbital period is between P and P+dP.**

# Significance of this development

**Probabilities are commonly quantified by a real number between 0 and 1.**




**The end-points, corresponding to absolutely false and absolutely true, are simply the extreme limits of this infinity of real numbers.**

**Bayesian probability theory spans the whole range.**

**Deductive logic is just a special case of Bayesian probability theory in the idealized limit of complete information.**

# Calculation of a simple Likelihood $p\left(D \mid M, \vec{X}, I\right)$

outline

Let $d_i$ represent the $i^{th}$ measured data value . We model $d_i$ by,

$$d_i = f_i(\vec{X}) + e_i$$

Model prediction for $i^{th}$ data value for current choice of parameters $\vec{X}$

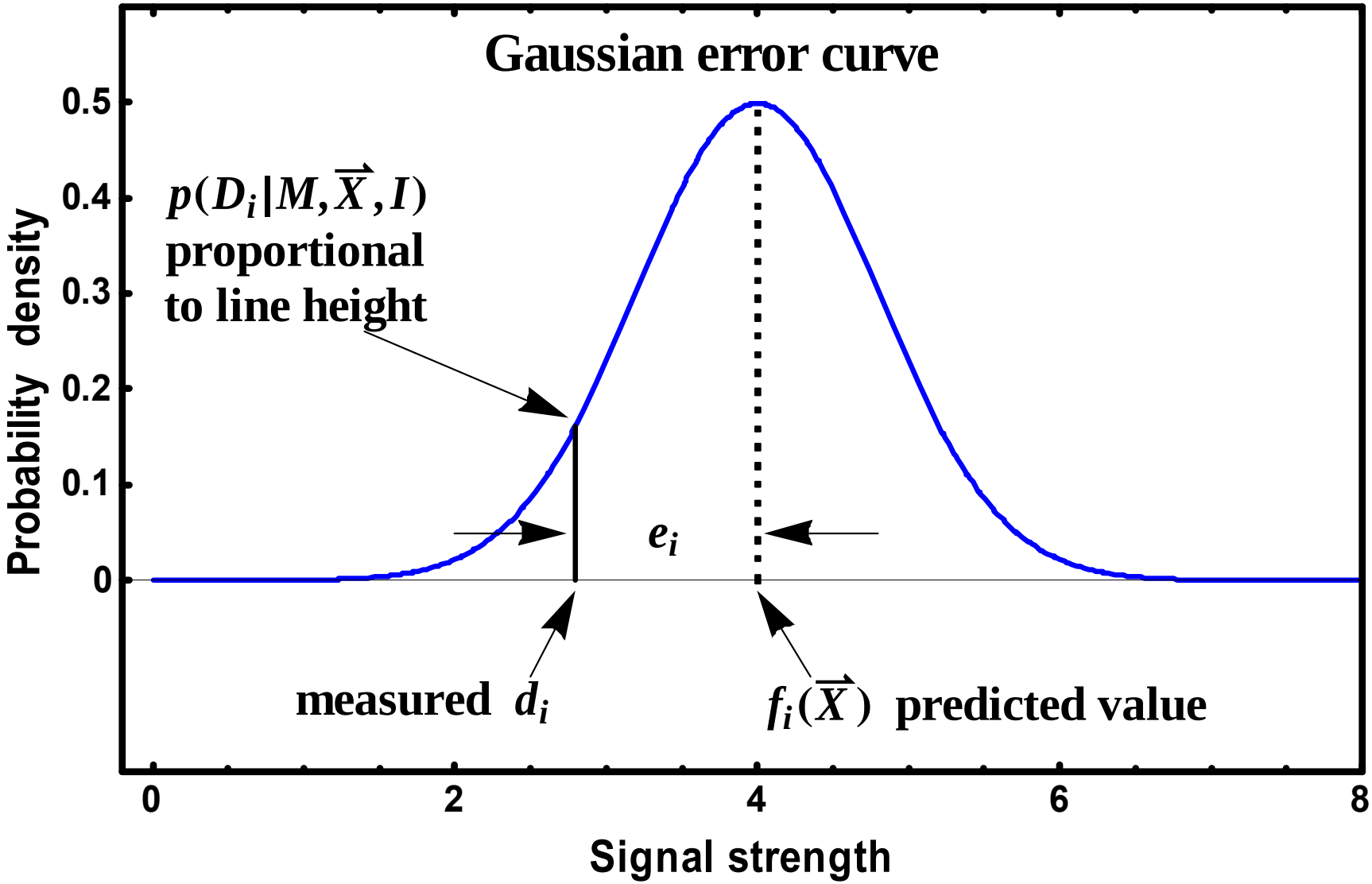
where $e_i$ represents the error component in the measurement.

Since $M$, $\vec{X}$ is assumed to be true, if it were not for the error $e_i$, $d_i$ would equal the model prediction $f_i$.

Now suppose prior information $I$ indicates that $e_i$ has a Gaussian probability distribution. Then

$$p\left(D_i \mid M, \vec{X}, I\right) = \frac{1}{\sigma_i \sqrt{2\pi}} Exp\left[-\frac{e_i^2}{2\sigma_i^2}\right]$$

$$= \frac{1}{\sigma_i \sqrt{2\pi}} Exp\left[-\frac{\left(d_i - f_i(\vec{X})\right)^2}{2\sigma_i^2}\right]$$




Probability of getting a data value $d_i$ a distance $e_i$ away from the predicted value $f_i$ is proportional to the height of the Gaussian error curve at that location.

# Calculation of a simple Likelihood $p\left(D \mid M, \vec{X}, I\right)$



**For independent data the likelihood for the entire data set $D=(D_1,D_2,\ldots,D_N)$ is the product of $N$ Gaussians.**

$$p\left(D \mid M, \vec{X}, I\right) = (2\pi)^{-N/2} \left\{\prod_{i=1}^{N} \sigma_i^{-1}\right\} Exp\left[-0.5 \sum_{i=1}^{N} \frac{\left(d_i - f_i(\vec{X})\right)^2}{\sigma_i^2}\right]$$

**The familiar $\chi^2$ statistic used in least-squares**

**Maximizing the likelihood corresponds to minimizing $\chi^2$**

***Recall: Bayesian posterior $\propto$ prior × likelihood***

**Thus, only for a uniform prior will a least-squares analysis yield the same solution as the Bayesian posterior.**

# Simple example of when not to use a uniform prior

**In the exoplanet problem the prior range for the unknown orbital period P is very large from ~1 day to 1000 yr (upper limit set by perturbations from neighboring stars).**

**Suppose we assume a uniform prior probability density for the P parameter. This would imply that we believed that it was ~ $10^4$ times more probable that the true period was in the upper decade ($10^4$ to $10^5$ d) of the prior range than in the lowest decade from 1 to 10 d.**

$$\frac{\int_{10^4}^{10^5} p\,(P \mid M, I)\, dP}{\int_{1}^{10} p\,(P \mid M, I)\, dP} = 10^4$$

**Usually, expressing great uncertainty in some quantity corresponds more closely to a statement of scale invariance or equal probability per decade. The Jeffreys prior has this scale invariant property.**

# Jeffreys prior (scale invariant)

$$p(P \mid M, I)\,dP = \frac{dP}{P \times \ln(P_{max}/P_{min})}$$

or equivalently $$p(\ln P \mid M, I)\,d\ln P = \frac{d\ln P}{\ln(P_{max}/P_{min})}$$

Equal probability per decade

$$\int_{1}^{10} p\,(P \mid M, I)\,dP = \int_{10^4}^{10^5} p\,(P \mid M, I)\,dP$$

Actually, there are good reasons for searching in orbital frequency f = 1/P instead of P. The form of the prior is unchanged.

$$p(\ln f \mid M, I)\,d\ln f = \frac{d\ln f}{\ln(f_{max}/f_{min})}$$

# Integration not minimization

**A full Bayesian analysis requires integrating over the model parameter space. Integration is more difficult than minimization.**

**However, the Bayesian solution provides the most accurate information about the parameter errors and correlations without the need for any additional calculations, i.e., Monte Carlo simulations.**

**Shortly discuss an efficient method for Integrating over a large parameter space called Markov chain Monte Carlo (MCMC).**

**End of Bayesian primer**

# Simple Spectral Line Problem

## Background (prior) information:

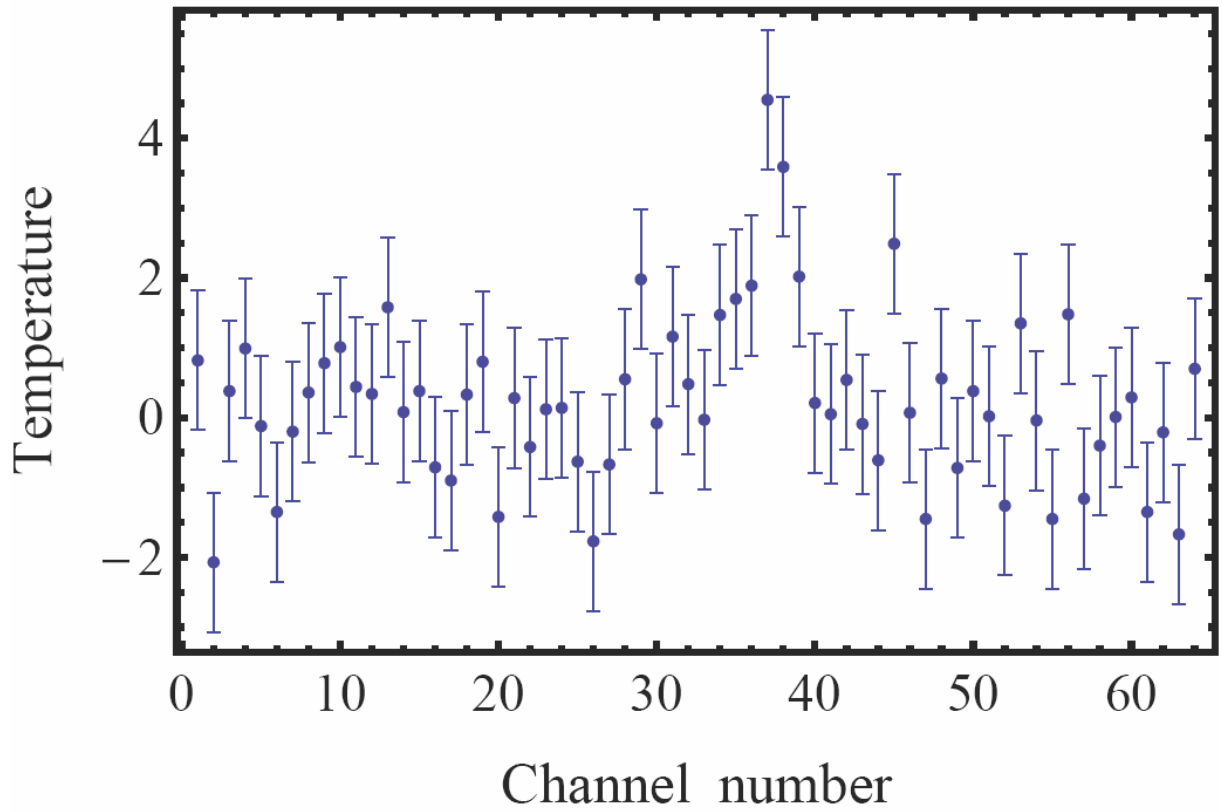
Two competing grand unification theories have been proposed, each championed by a Nobel prize winner in physics. We want to compute the relative probability of the truth of each theory based on our prior information and some new data.

Theory 1 is unique in that it predicts the existence of a new short-lived baryon which is expected to form a short-lived atom and give rise to a spectral line at an accurately calculable radio wavelength.

Unfortunately, it is not feasible to detect the line in the laboratory. The only possibility of obtaining a sufficient column density of the short-lived atom is in interstellar space.

# Data

**To test this prediction, a new spectrometer was mounted on the James Clerk Maxwell telescope on Mauna Kea and the spectrum shown below was obtained. The spectrometer has 64 frequency channels.**




**All channels have Gaussian noise characterized by $\sigma = 1$ mK. The noise in separate channels is independent.**

# Simple Spectral Line Problem



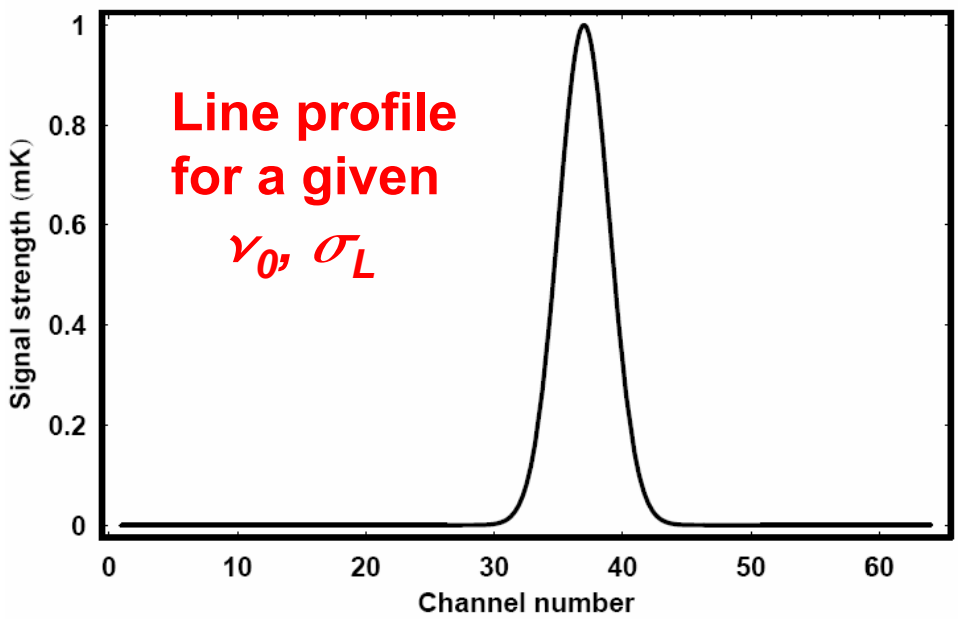
**The predicted line shape has the form**

$$T \exp\left\{\frac{-(\nu_i - \nu_o)^2}{2\sigma_L^2}\right\} \quad \text{(abbreviated by } Tf_i\text{)},$$

**where the signal strength is measured in temperature units of mK and $T$ is the amplitude of the line. The frequency, $\nu_i$ , is in units of the spectrometer channel number and the line center frequency is $\nu_0$.**

**In this version of the problem $T$, $\nu_0$, $\sigma_L$ are all unknowns with prior limits:**

**$T$ = 0.0 - 100.0**
**$\nu_0$ = 1 – 44**
**$\sigma_L$ = 0.5 – 4.0**



# Extra noise term, $\epsilon_{0i}$

We will represent the measured data by the equation

$$d_i = f_i + \epsilon_i + \epsilon_{0i}$$

$d_i = i^{th}$ measured data value

$f_i =$ model prediction

$\epsilon_i =$ component of $d_i$ which arises from measurement errors

$\epsilon_{0i} =$ any additional unknown measurement errors plus any real signal in the data that cannot be explained by the model prediction $f_i$

In the absence of detailed knowledge of the sampling distribution for $\epsilon_{0i}$, other than that it has a finite variance, the Maximum Entropy principle tells us that a Gaussian distribution is the most conservative choice (i.e., maximally non committal about the information we don't have).

We therefore adopt a Gaussian distribution for $\epsilon_{0i}$ with a variance $s^2$. Thus the combination of $\epsilon_i + \epsilon_{0i}$ has a Gaussian distribution with

$$\text{variance} = \sigma_i^2 + s^2$$

In Bayesian analysis we marginalize the unknown $s$ (integrate it out of the problem), which has the desirable effect of treating as noise anything in the data that can' t be explained by the model and known measurement errors, leading to most conservative estimates of the model parameters. Prior range for $s$ = 0 - 0.5 × data range.

# Questions of interest

Based on our current state of information, which includes just the above prior information and the measured spectrum,

1) what do we conclude about the relative probabilities of the two competing theories

*and*

2) what is the posterior PDF for the model parameters and s?

Hypothesis space of interest for model selection part:

$M_0$ ≡ "Model 0, no line exists"
$M_1$ ≡ "Model 1, line exists"

$M_1$ has 3 unknown parameters, the line temperature $T$, $\nu_0$, $\sigma_L$ and one nuisance parameter s.

$M_0$ has no unknown parameters, and one nuisance parameter s.

# Likelihood for the spectral line model

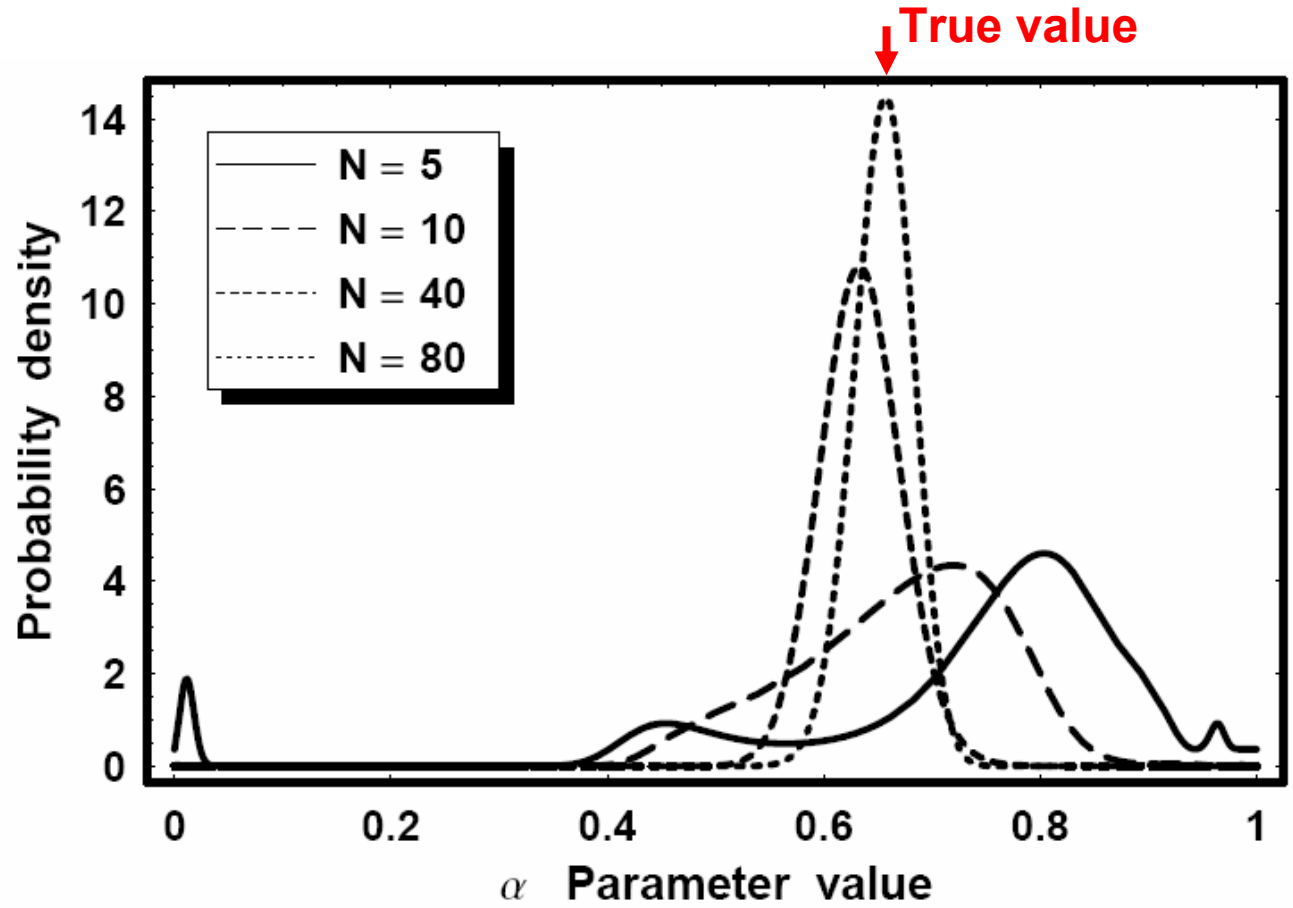
In the earlier spectral line problem which had only one unknown variable $T$ we derived the likelihood

$$p(D \mid M_1, T, I) = (2\pi)^{-\frac{N}{2}} \sigma^{-N} \, Exp\left[ -\sum_{i=1}^{N} \frac{(d_i - T f_i)^2}{2\sigma} \right]$$

Our new likelihood for the more complicated model with unknown variables $T, \upsilon_0, \sigma_L, s$

$$p(D \mid M_1, T, \upsilon_0, \sigma_L, s, I) = (2\pi)^{-\frac{N}{2}} \left(\sigma^2 + s^2\right)^{-\frac{N}{2}} \, Exp\left[ -\sum_{i=1}^{N} \frac{(d_i - T f_i(\upsilon_0, \sigma_L))^2}{2\left(\sigma^2 + s^2\right)} \right]$$

# Simple nonlinear model with a single parameter α




**The Bayesian posterior density for a nonlinear model with single parameter, α, for 4 simulated data sets of different size ranging from *N* = 5 to *N* = 80. The *N* = 5 case has the broadest distribution and exhibits 4 maxima.**

**Asymptotic theory says that the maximum likelihood estimator becomes more unbiased, more normally distributed and of smaller variance as the sample size becomes larger.**

# Integration not minimization

In Least-squares analysis we minimize some statistic like $\chi^2$.
In a Bayesian analysis we need to integrate.

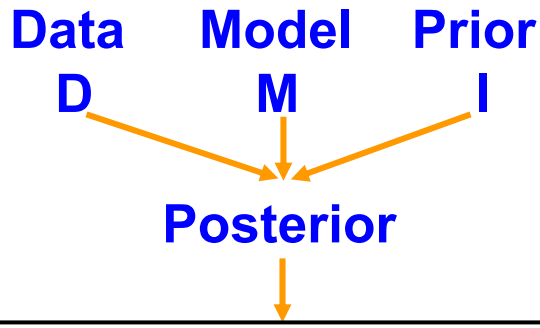
Parameter estimation: to find the marginal posterior probability density function (PDF) for the orbital period *P*, we need to integrate the joint posterior over all the other parameters.

$$\underbrace{p(T \mid D, M_1, I)}_{\text{Marginal PDF for } T} = \int d\upsilon_0\, d\sigma_L\, ds\ \underbrace{p(T, \upsilon_0, \sigma_L, s \mid D, M_1, I)}_{\text{Joint posterior probability density function (PDF) for the parameters}}$$

Integration is more difficult than minimization. However, the Bayesian solution provides the most accurate information about the parameter errors and correlations without the need for any additional calculations, i.e., Monte Carlo simulations.

Shortly discuss an efficient method for Integrating over a large parameter space called Markov chain Monte Carlo (MCMC).

# Numerical tools for Bayesian model fitting




## Linear models (uniform priors)

**Posterior has a single peak (multi-dimensional Gaussian)**

Parameters given by the normal equations of linear least-squares

No integration required
solution very fast
using linear algebra

## Nonlinear models

+ linear models (non-uniform priors)

**Posterior may have multiple peaks**

### Brute force integration

**For some parameters analytic integration sometimes possible**

### Asymptotic approx.'s

peak finding algorithms

(1) Levenberg-Marquardt

(2) Simulated annealing

(3) Genetic algorithm

Laplace approx.'s

### Moderate dimensions

quadrature

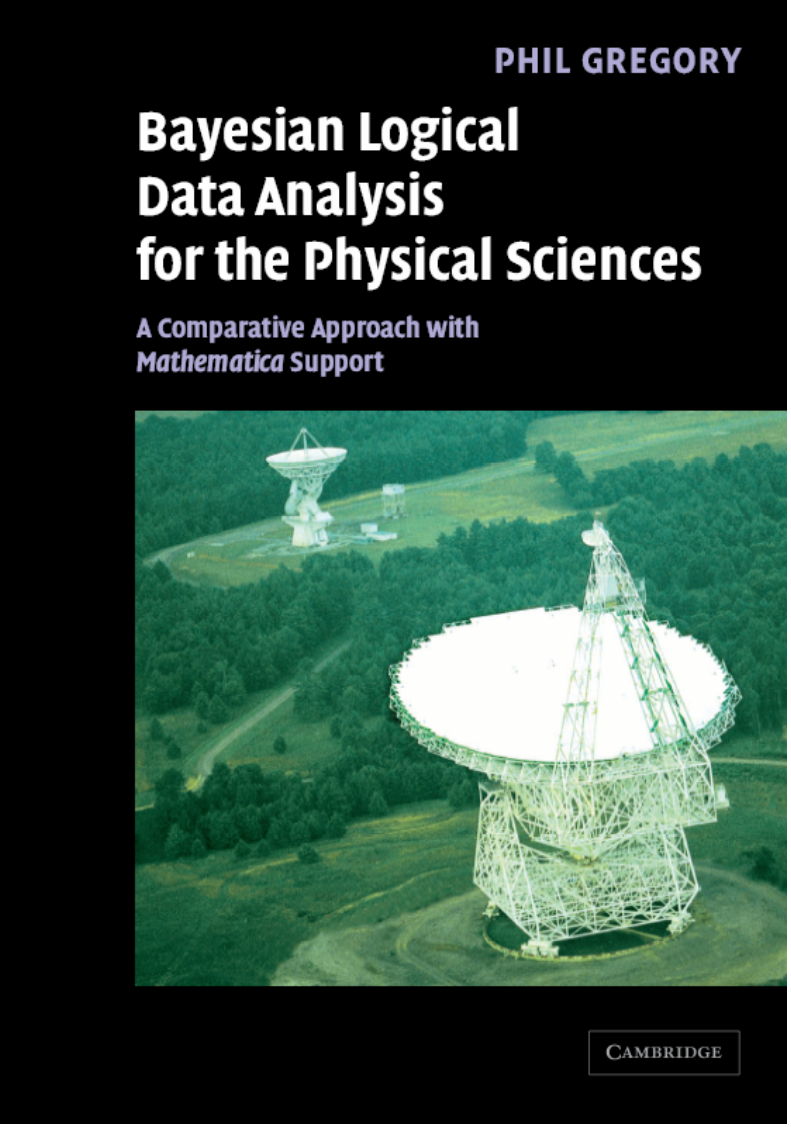
randomized quadrature

adaptive quadrature

### High dimensions

MCMC

(chapter 10) (chapter 11) (chapter 12)




# Chapters

1. Role of probability theory in science
2. Probability theory as extended logic
3. The how-to of Bayesian inference
4. Assigning probabilities
5. Frequentist statistical inference
6. What is a statistic?
7. Frequentist hypothesis testing
8. Maximum entropy probabilities
9. Bayesian inference (Gaussian errors)
10. Linear model fitting (Gaussian errors)
11. Nonlinear model fitting
12. Markov chain Monte Carlo
13. Bayesian spectral analysis
14. Bayesian inference (Poisson sampling)

**Resources and solutions**
This title has free Mathematica based support software available

Introduces statistical inference in the larger context of scientific methods, and includes 55 worked examples and many problem sets.

# MCMC for integration in large parameter spaces

Markov chain Monte Carlo (MCMC) algorithms provide a powerful means for efficiently computing integrals in many dimensions to within a constant factor. This factor is not required for parameter estimation.

After an initial burn-in period (which is discarded) , the MCMC produces an equilibrium distribution of samples in parameter space such that the density of samples is proportional to the joint posterior PDF.

It is very efficient because, unlike straight Mont Carlo integration, it doesn't waste time exploring regions where the joint posterior is very small.

The MCMC employs a Markov chain random walk, whereby the new sample in parameter space, designated $X_{\{t+1\}}$ , depends on previous sample $X_t$ according to an entity called the transition probability or kernel, $p(X_{\{t+1\}} | X_t)$. The transition kernel is assumed to be time independent.

# Starting point: Metropolis-Hastings MCMC algorithm

P(X|D,M,I) = target posterior probability distribution
(X represents the set of model parameters)

1. Choose $X_0$ an initial location in the parameter space. Set $t = 0$.

2. Repeat {

– Obtain a new sample Y from a proposal distribution $q(Y \mid X_t)$ that is easy to evaluate. $q(Y \mid X_t)$ can have almost any form.

I use a Gaussian proposal distribution. i.e., Normal distribution $N(X_t, \sigma)$

– Sample a Uniform (0, 1) random variable U.

– If $U \leq \dfrac{p(Y \mid D, I)}{p(X_t \mid D, I)} \times \dfrac{q(X_t \mid Y)}{q(Y \mid X_t)}$ then set $X_{t+1} = Y$

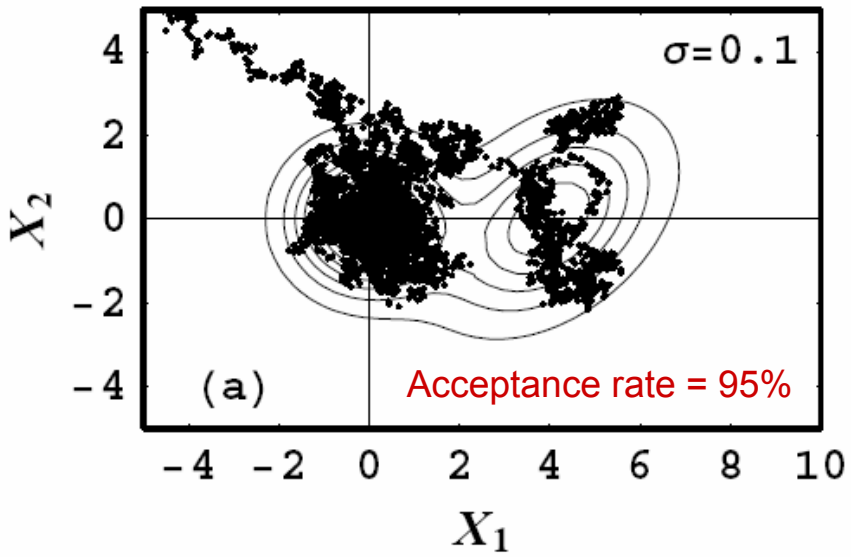
otherwise set $X_{t+1} = X_t$

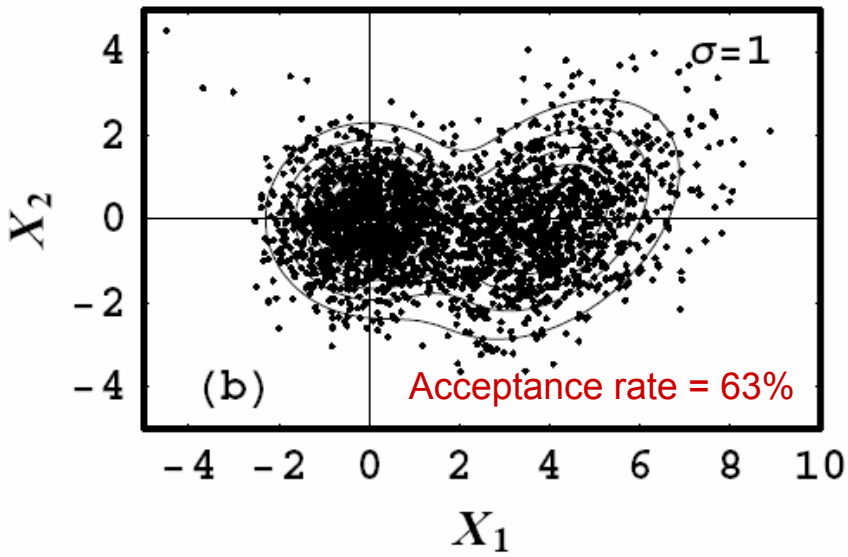
– Increment t }

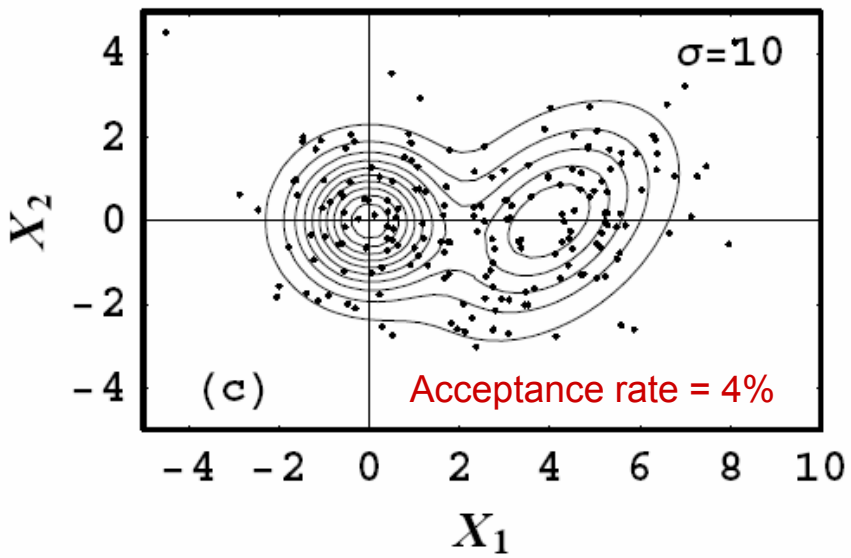
This factor =1
for a symmetric proposal
distribution like a Gaussian

return

# Toy MCMC simulations: the efficiency depends on tuning proposal distribution $\sigma$'s. Can be a very difficult challenge for many parameters.




**In this example the posterior probability distribution consists of two 2 dimensional Gaussians indicated by the contours**

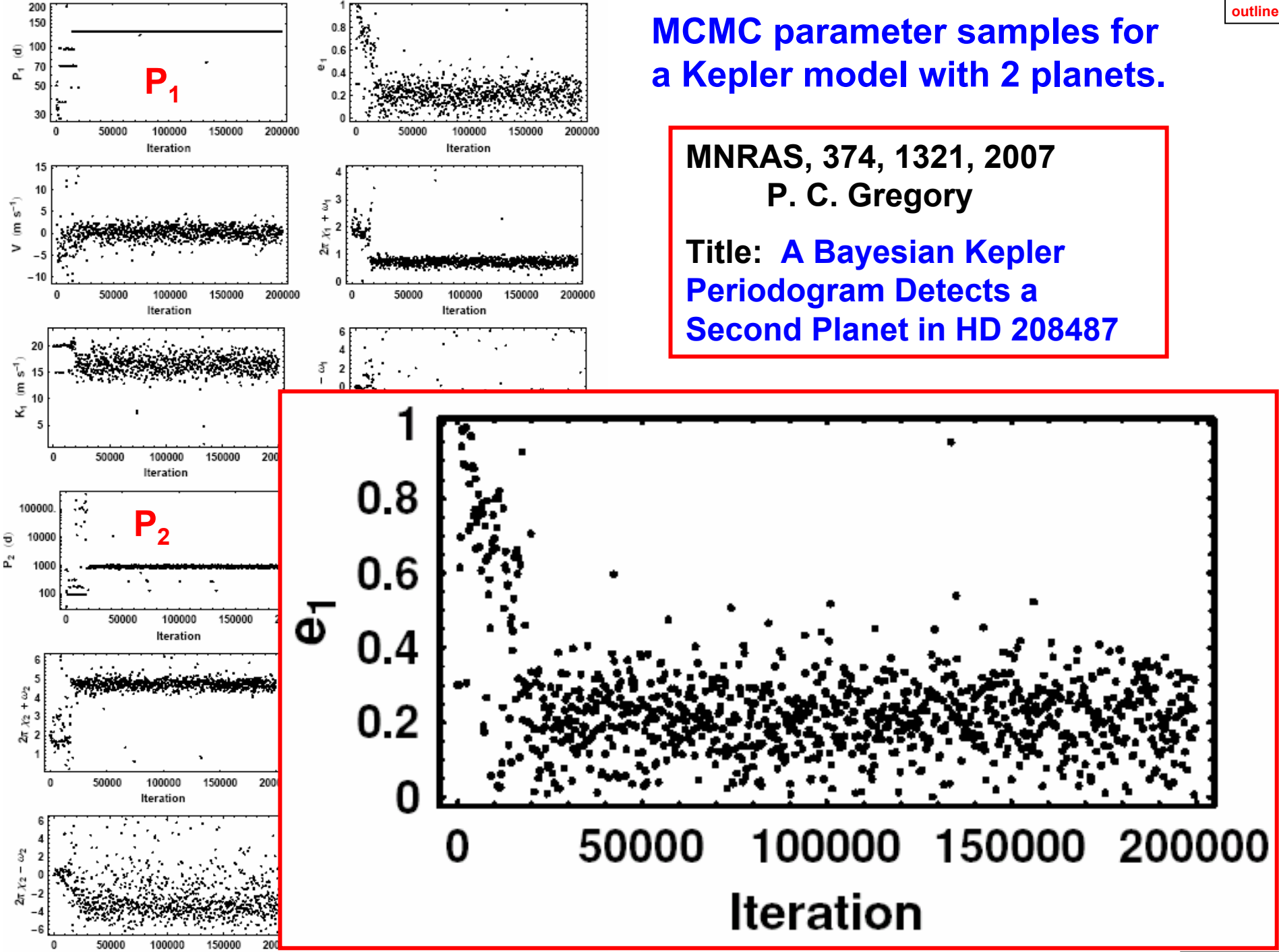
Autocorrelation

return

# MCMC parameter samples for a Kepler model with 2 planets.

MNRAS, 374, 1321, 2007
P. C. Gregory

Title: A Bayesian Kepler Periodogram Detects a Second Planet in HD 208487



# Parallel tempering MCMC

**The simple Metropolis-Hastings MCMC algorithm can run into difficulties if the probability distribution is multi-modal with widely separated peaks. It can fail to fully explore all peaks which contain significant probability, especially if some of the peaks are very narrow.**

**One solution is to run multiple Metropolis-Hastings simulations in parallel, employing probability distributions of the kind**

$$\pi(X \mid D, M, \beta, I) = p(X \mid M, I)\; p(D \mid X, M, I)^{\beta} \quad (0 < \beta \leqslant 1)$$

**Typical set of β values = 0.09,0.15,0.22,0.35,0.48,0.61,0.78,1.0**

**β = 1 corresponds to our desired target distribution. The others correspond to progressively flatter probability distributions.**

**At intervals, a pair of adjacent simulations are chosen at random and a proposal made to swap their parameter states. The swap allows for an exchange of information across the ladder of simulations.**

**In the low β simulations, radically different configurations can arise, whereas at higher β, a configuration is given the chance to refine itself.**

**Final results are based on samples from the β = 1 simulation. Samples from the other simulations provide one way to evaluate the Bayes Factor in model selection problems.**

# MCMC Technical Difficulties

1. Deciding on the burn-in period.

2. Choosing a good choice for the characteristic width of each proposal distribution, one for each model parameter.

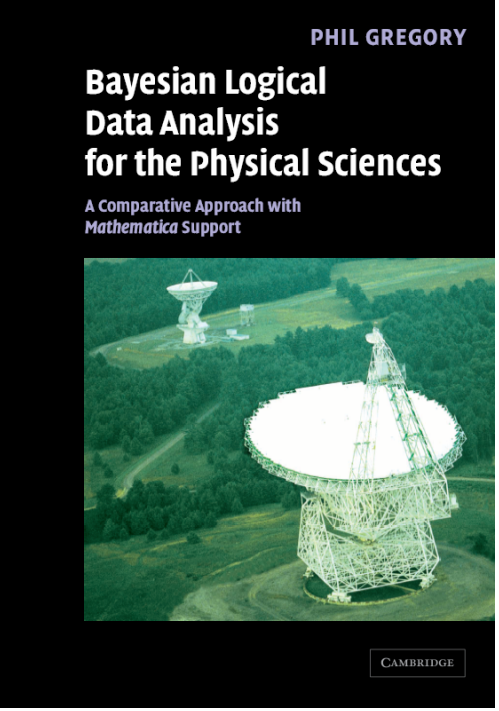
   For Gaussian proposal distributions this means picking a set of proposal σ's. This can be very time consuming for a large number of different parameters.

3. Handling highly correlated parameters.
   Ans: transform parameter set or differential MCMC

4. Deciding how many iterations are sufficient.
   Ans: use **Gelman-Rubin Statistic**

5. Deciding on a good choice of tempering levels (β values).

Gelman –Rubin statistic




## My involvement: since 2002, ongoing development of a general Bayesian Nonlinear model fitting program.

**My latest hybrid Markov chain Monte Carlo (MCMC) nonlinear model fitting algorithm incorporates:**

- **Parallel tempering**
- **Simulated annealing**
- **Genetic algorithm**
- **Differential evolution**
- **Unique control system automates the MCMC**

**Code is implemented in *Mathematica***

**Current extra-solar planet applications:**

- **precision radial velocity data – (4 new planets published to date)**
- **pulsar planets from timing residuals of NGC 6440C**
- **NASA stellar interferometry mission astrometry testing**

**Submillimeter radio spectroscopy of galactic center methanol lines**

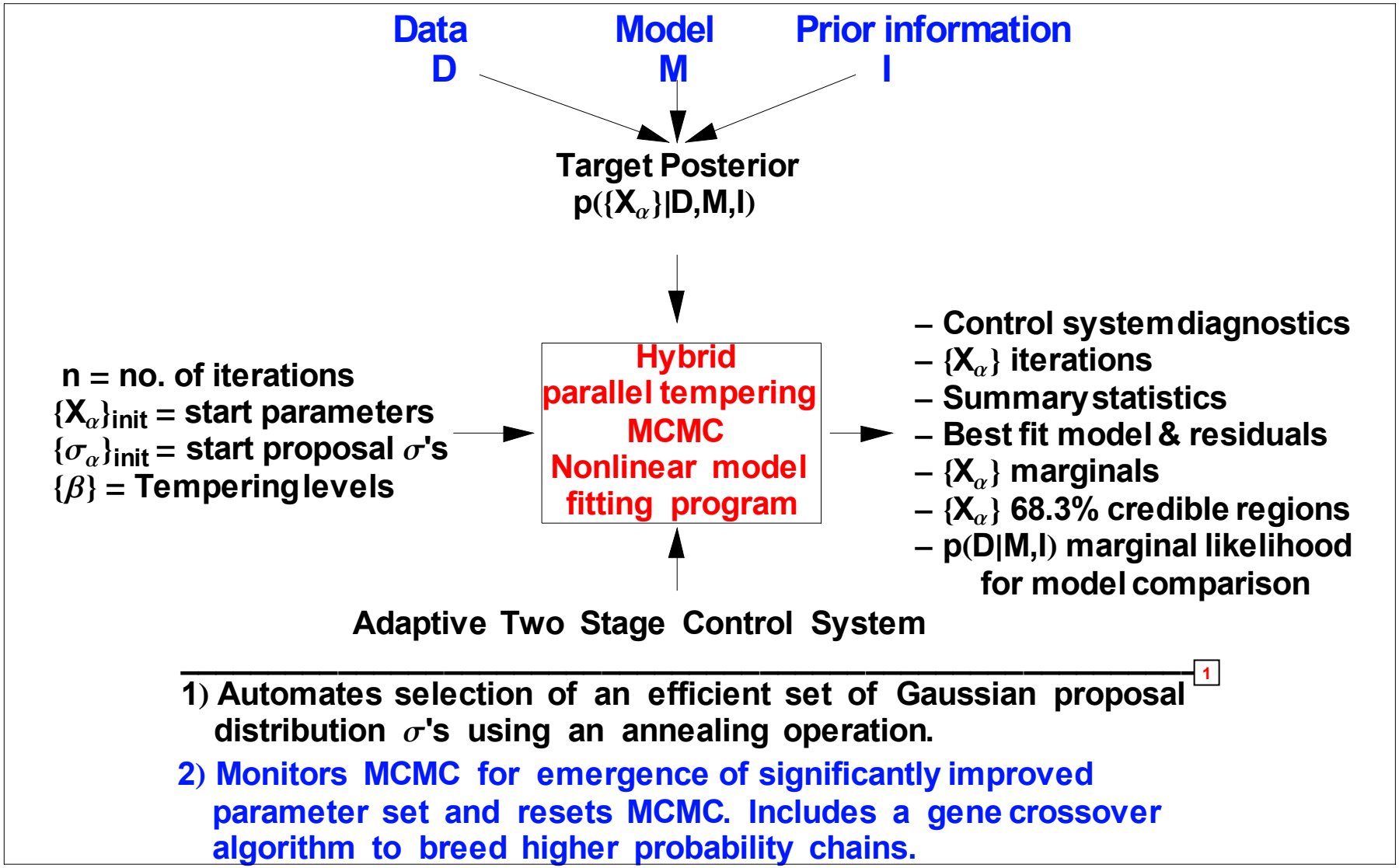
***Mathematica* 7 (latest version) provides an easy route to parallel computing.**

**I run on an 8 core PC and achieve a speed-up of 7 times.**

# Blind searches with hybrid MCMC

Parallel tempering
Simulated annealing
Genetic algorithm
Differential evolution

**Each of these methods was designed to facilitate the detection of a global minimum in $\chi^2$. By combining all four in a hybrid MCMC we greatly increase the probability of realizing this goal.**

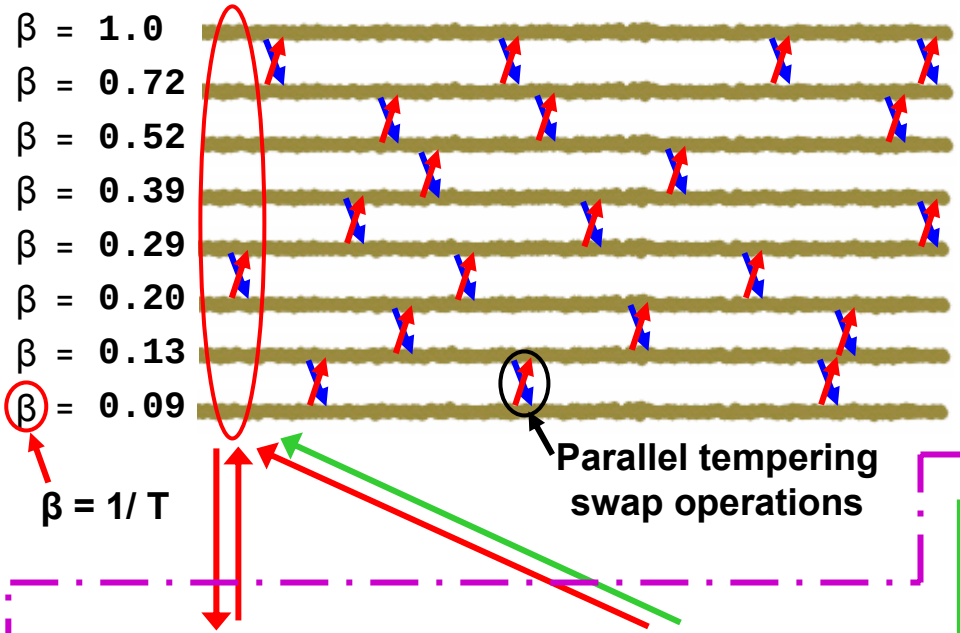



**Schematic of a Bayesian Markov chain Monte Carlo program for nonlinear model fitting. The program incorporates a control system that automates the selection of Gaussian proposal distribution σ's.**

# Adaptive Hybrid MCMC



## 8 parallel tempering Metropolis chains




## Output at each iteration

parameters, logprior + $\beta$ × loglike, logprior + loglike

parameters, logprior + $\beta$ × loglike, logprior + loglike

parameters, logprior + $\beta$ × loglike, logprior + loglike

parameters, logprior + $\beta$ × loglike, logprior + loglike

parameters, logprior + $\beta$ × loglike, logprior + loglike

parameters, logprior + $\beta$ × loglike, logprior + loglike

parameters, logprior + $\beta$ × loglike, logprior + loglike

parameters, logprior + $\beta$ × loglike, logprior + loglike

**Monitor for parameters with peak probability**

**Peak parameter set:**
If (logprior + loglike) > previous best by a threshold then update and reset burn-in

**Anneal Gaussian proposal $\sigma$'s**

**Refine & update Gaussian proposal $\sigma$'s**

**2 stage proposal $\sigma$ control system**

error signal =
(actual joint acceptance rate – 0.25)

Effectively defines burn-in interval

**Genetic algorithm**

Every 10th iteration perform gene crossover operation to breed larger (logprior + loglike) parameter set.

**MCMC adaptive control system**

Go to *Mathematica* support material

Go to *Mathematica* version of MCMC

# Calculation of $p(D|M_0,I)$

**Model $M_0$ assumes the spectrum is consistent with noise and has no free parameters so we can write**

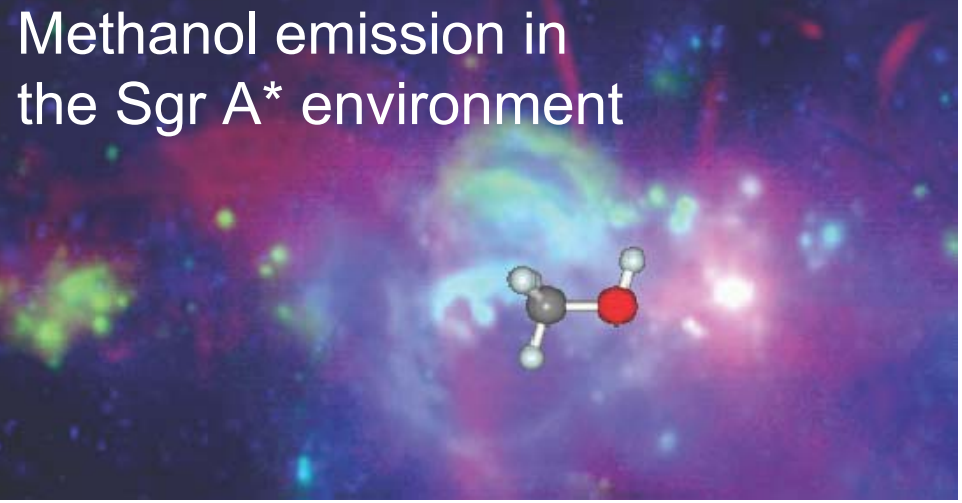
$$d_i = 0 + e_i$$

$$p(D \mid M_0, s, I) = (2\pi)^{-\frac{N}{2}} \left(\sigma^2 + s^2\right)^{-\frac{N}{2}} Exp\left[-\sum_{i=1}^{N} \frac{(d_i - 0)^2}{2\left(\sigma^2 + s^2\right)}\right]$$

## Model selection results

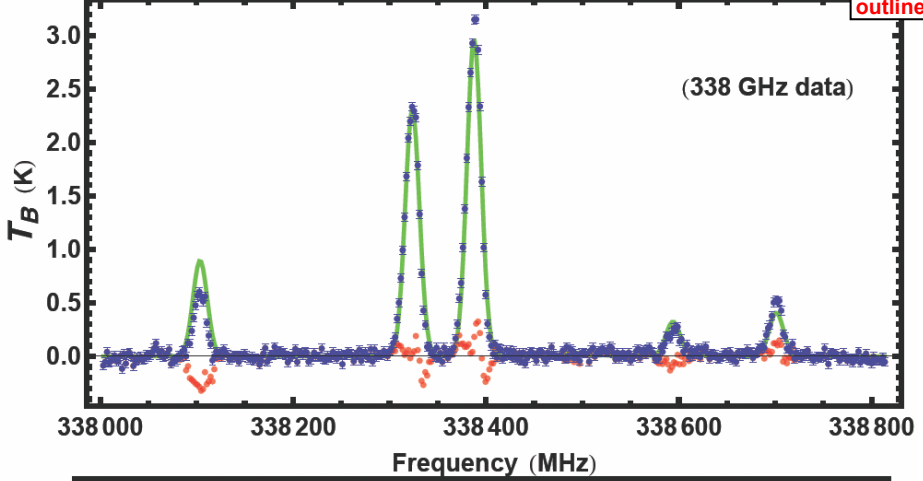
**Bayes factor =$4.5 \times 10^4$**

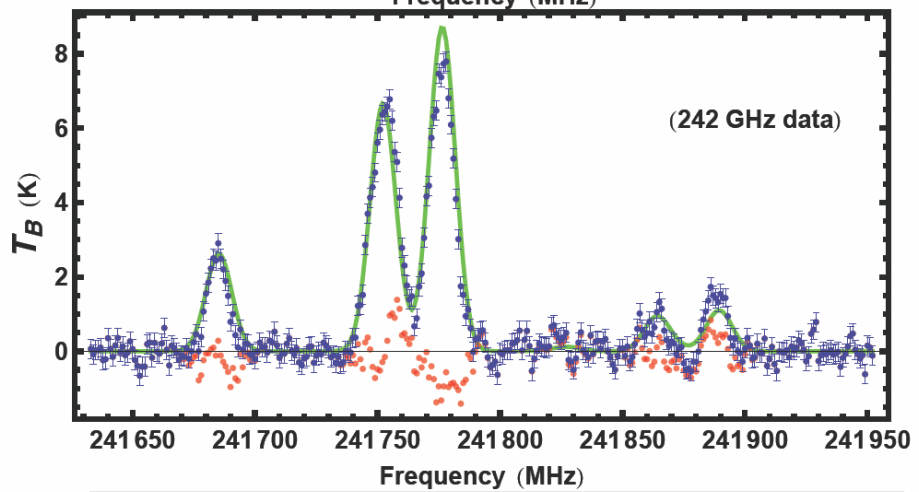
# Methanol emission in the Sgr A* environment

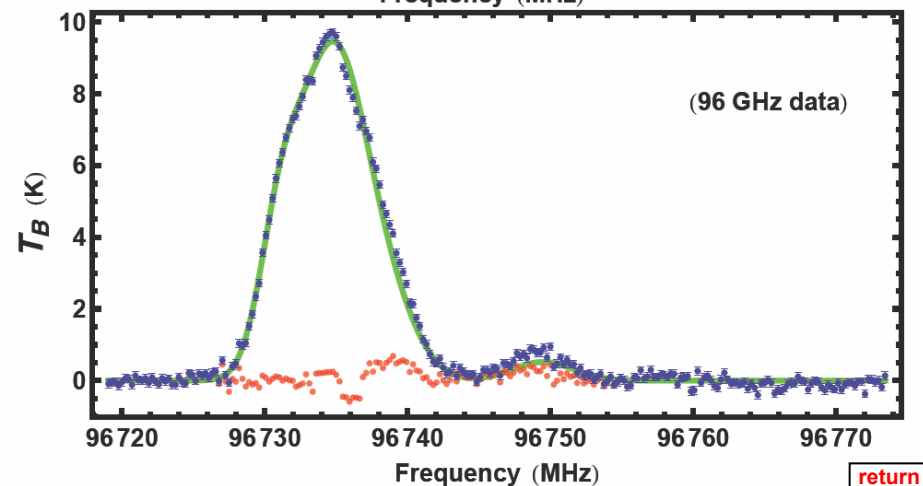


M. Stanković, E.R. Seaquist (UofT), S. Leurini (ESO), P.Gregory (UBC), S. Muehle(JIVE), K.M.Menten (MPIfR)

**Optically thin fit to 3 bands + unidentified line in 96 GHz band**




$\{v\ (\mathrm{km\ s^{-1}}),\ \mathrm{FWHM}\ (\mathrm{km\ s^{-1}}),\ T_J\ (\mathrm{K}),\ (N/Z)_A\ (\mathrm{cm^{-2}}),\ (N/Z)_A\ (\mathrm{cm^{-2}})$
$T_K\ (\mathrm{K}),\ \nu_{UL}\ (\mathrm{MHz}),\ \mathrm{FWHM}_{UL}\ (\mathrm{km\ s^{-1}}),\ T_{UL}\ (\mathrm{K}),\ ds_{96},\ ds_{242},\ s\ (\mathrm{K})\}$

$\{18.5749,\ 15.4783,\ 7.79329,\ 2.02377\times10^{14},\ 2.23857\times10^{14},$
$16.9128,\ 96\,737.,\ 10.5728,\ 3.21161,\ 0.170494,\ 0.294291,\ 0.063339\}$

$\nu_{UL}$ (MHz) is the rest frequency of the unidentied line after removal of the Doppler veocity, v (km $s^{-1}$)

outline

return

# Conclusions

1. For Bayesian parameter estimation, MCMC provides a powerful means of computing the integrals required to compute posterior probability density function (PDF) for each model parameter.
2. Even though we demonstrated the performance of an MCMC for a simple spectral line problem with only 4 parameters, MCMC techniques are really most competitive for models with a much larger number of parameters m ≥ 15.
3. Markov chain Monte Carlo analysis produces samples in model parameter space in proportion to the posterior probability distribution. This is fine for parameter estimation.

   For model selection we need to determine the proportionality constant to evaluate the marginal likelihood $p(D|Mi, I)$ for each model. This is a much more difficult problem still in search of two good solutions for large m. We need two to know if either is valid.

   One solution is to use the MCMC results from all the parallel tempering chains spanning a wide range of β values, however, this becomes computationally very intensive for m > 17.

For a copy of this talk please Google Phil Gregory

**The rewards of data analysis:**

**‘The universe is full of magical things, patiently waiting for our wits to grow sharper.’**

**Eden Philpotts (1862-1960)**

**Author and playwright**

# Gelman-Rubin Statistic

Let $\theta$ represent one of the model parameters .

Let $\theta_j^i$ represent the $i^{th}$ iteration of the $j^{th}$ of m independent simulation.

Extract the last $\eta$ post burn – in iterations for each simulation.

$$\text{Mean within chain variance } W = \frac{1}{m(\eta-1)} \sum_{j=1}^{m} \sum_{i=1}^{\eta} (\theta_j^i - \bar{\theta}_j)^2$$

$$\text{Between chain variance } B = \frac{\eta}{m-1} \sum_{j=1}^{m} (\bar{\theta}_j - \bar{\theta})^2$$

$$\text{Estimated variance } \hat{V}(\theta) = \left(1 - \frac{1}{\eta}\right) W + \frac{1}{\eta} B$$

$$\text{Gelman – Rubin statistic } = \sqrt{\frac{\hat{V}(\theta)}{W}}$$

The Gelman – Rubin statistic should be close to 1.0 (e.g. $< 1.05$) for all paramaters for convergence

Ref: Gelman, A. and D.B. Rubin (1992) ' Inference from iterative simulations using multiple sequences (with discussion) ', Statistical Science 7, pp. 457 – 511.

return