# SUPPLEMENT TO
# ‘BAYESIAN LOGICAL DATA ANALYSIS
# FOR THE PHYSICAL SCIENCES’

P. C. Gregory

*Department of Physics and Astronomy, University of British Columbia*

23 Jan. 2018

Supplemental Chapter 1 deals with my Fusion Markov chain Monte Carlo (FMCMC) algorithm, an extension of the work introduced in the latter part of Chapter 12 of the textbook. FMCMC is a special version of the Metropolis algorithm that incorporates parallel tempering, genetic crossover operations, and an automatic simulated annealing. Each of these features facilitate the detection of a global minimum in chi-squared in a highly multi-modal environment. By combining all three, the algorithm greatly increases the probability of realizing this goal. It is a powerful general purpose tool for Bayesian model fitting which I have been using to great success in the arena of extra-solar planets. The FMCMC is controlled by a unique adaptive control system that automates the tuning of the MCMC proposal distributions for efficient exploration of the model parameter space even when the parameters are highly correlated.

Chapter 1 also includes important new work on Bayesian model comparison and introduces a new marginal likelihood estimator, called Nested Restricted Monte Carlo (NRMC), used in the calculation of Bayes' factors. Its performance is compared with two other marginal likelihood estimators that depend on the posterior MCMC samples. There are three supporting appendices to Chapter 1.

Supplemental Chapter 2 provides an introduction to hierarchical or multilevel Bayes. It examines how to handle hidden variables and missing data. There are numerous examples of the use of fusion MCMC in Bayesian regression and regression involving selection effects, providing an important enhancement to the coverage of the textbook.

These supplemental chapters are intended to maintain and strengthen the book's appeal as a text for graduate-level courses in the physical sciences. Although the main textbook provides some examples using Mathematica commands, the supplemental chapters are standalone and do not make use of any specific computer language.

# Contents

# 1

# Fusion Markov chain Monte Carlo: A Powerful Tool for Bayesian Data Analysis

## 1.1 Introduction

The purpose of this chapter is to introduce a new tool for Bayesian data analysis called fusion Markov chain Monte Carlo (FMCMC), a new general purpose tool for nonlinear model fitting and regression analysis. It is the outgrowth of an earlier attempt to achieve an automated MCMC algorithm discussed in Section 12.8 of my book "Bayesian Logical Data Analysis for the Physical Sciences," Cambridge University Press (2005, 2010) [24].

FMCMC is a special version of the Metropolis MCMC algorithm that incorporates parallel tempering, genetic crossover operations, and an automatic simulated annealing. Each of these features facilitate the detection of a global minimum in chi-squared in a highly multi-modal environment. By combining all three, the algorithm greatly increases the probability of realizing this goal.

The FMCMC is controlled by a unique adaptive control system that automates the tuning of the MCMC proposal distributions for efficient exploration of the model parameter space even when the parameters are highly correlated. This controlled statistical fusion approach has the potential to integrate other relevant statistical tools as desired. The FMCMC algorithm is implemented in *Mathematica* using parallized code to take advantage of multiple core computers.

The power of this approach will be illustrated by considering a particular model fitting problem from the exciting area of extra-solar planet (exoplanets) research. In a subsequent chapter we will examine other applications in multilevel (hierarchical) Bayesian analysis. The next section of this chapter provides some background information on exoplanet discoveries to motivate the Bayesian analysis. Section 1.3 provides a brief self contained introduction to Bayesian logical data analysis applied to the Kepler exoplanet problem leading to the MCMC challenge. For more details on Bayesian logical data analysis please see my book [24].

The different components of Fusion MCMC are described in Section 1.4. In

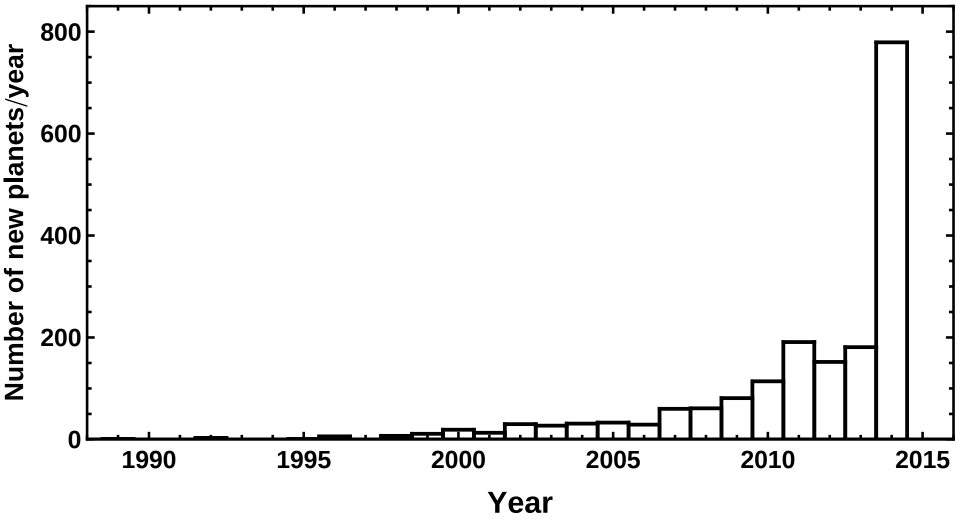


Figure 1.1 The pace of exoplanet discoveries as of Dec. 2014.

Section 1.5, the performance of the algorithm is illustrated with some exoplanet examples. Section 1.6 deals with the challenges of Bayesian model comparison and describes a new method for computing marginal likelihoods called Nested Restricted Monte Carlo (NRMC). Several appendices provide important details on FMCMC control system, and details behind the choices of priors.

## 1.2 Exoplanets

A remarkable array of new ground based and space based astronomical tools are providing astronomers access to other solar systems. Close to 2000 planets have been discovered to date, starting from the pioneering work of [9, 66, 46, 45]. Fig. 1.1 illustrates the pace of discovery[1] up to Oct. 2014.

A wide range of techniques including radial velocity, transiting, gravitational micro-lensing, timing, and direct imaging, have contributed exoplanet discoveries. Because a typical star is approximately a billion times brighter than a planet, only a small fraction of the planets have been detected by direct imaging. The majority of the planets have been detected through transits or the reflex motion of the star caused by the gravitational tug of unseen planets, using precision radial velocity (RV) measurements. There are currently 1832 planets in 1145 planetary systems (as of 17 October 2014). 469 planets are known to be in multiple planet systems, the largest of which has seven planets [44]. Many additional candidates from NASA's

[1] Data from the Extrasolar planet Encyclopedia, Exoplanet.eu. [57]

Kepler transit detection mission are awaiting confirmation. One recent analysis of the Kepler data [49] estimates that the occurrence rate of Earth-size planets (radii between 1-2 times that of the Earth) in the habitable zone (where liquid water could exist) of sun like stars is 22 ± 8%.

These successes on the part of the observers has spurred a significant effort to improve the statistical tools for analyzing data in this field (e.g., [42, 41, 11, 24, 25, 19, 20, 21, 70, 12, 13, 17, 1, 5, 60, 35]. Much of this work has highlighted a Bayesian MCMC approach as a way to better understand parameter uncertainties and degeneracies and to compute model probabilities. MCMC algorithms provide a powerful means for efficiently computing the required Bayesian integrals in many dimensions (e.g., an 8 planet model has 42 unknown parameters). More on this below.

## 1.3 Bayesian primer

*What is Bayesian probability theory (BPT)?*

BPT = extended logic.

Deductive logic is based on axiomatic knowledge. In science we never know any theory of nature is absolutely 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 in his book[2], "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. Below are some examples of propositions.

- $A \equiv$ "Theory X is correct."
- $\bar{A} \equiv$ "Theory X is not correct."
- $D \equiv$ "The measured redshift of the galaxy is 0.15 ± 0.02."
- $B \equiv$ "The star has 5 planets."

[2] The book was published 5 year after Jaynes' death through the efforts of a former graduate student Dr. G. L. Bretthorst.

- $A \equiv$ "The orbital frequency is between $f$ and $f + df$."

In the last example, $f$ is a continuous parameter giving rise to a continuous hypothesis space. Negation of a proposition is indicated by a bar over top, i.e, $\bar{A}$.

The propositions that we want to reason about in science are commonly referred to as hypotheses or models. We will need to consider compound propositions like $A, B$ which asserts that propositions $A$ and $B$ are both true conditional on the truth of another proposition $C$. This is written $A, B|C$.

*Rules for manipulating probabilities and Bayes' theorem*

There are only two rules for manipulating probabilities.

$$\text{Sum rule : } p(A|C) + p(\bar{A}|C) = 1 \tag{1.1}$$

$$\begin{aligned}\text{Product rule : } \; p(A, B|C) &= p(A|C) \times p(B|A, C) \\ &= p(B|C) \times p(A|B, C)\end{aligned} \tag{1.2}$$

Bayes' theorem is obtained by rearranging the two right hand sides of the product rule.

$$\text{Bayes}' \text{ theorem : } p(A|B, C) = \frac{p(A|C) \times p(B|A, C)}{p(B|C)} \tag{1.3}$$

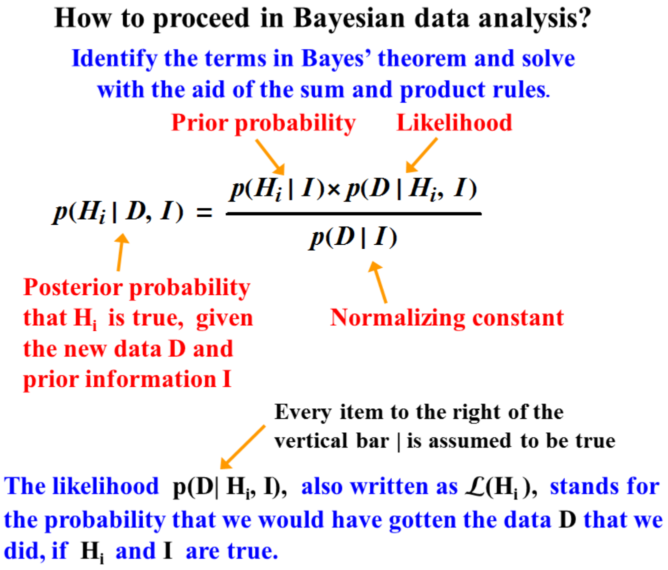
Figure 1.2 shows Bayes' theorem in its more usual form for data analysis purposes. In a well posed Bayesian problem the prior information, $I$, specifies the hypothesis space of current interest (range of models actively under consideration) and the procedure for computing the likelihood. The starting point is always Bayes' theorem. In any given problem the expressions for the prior and likelihood may be quite complex and require repeated applications of the sum and product rule to obtain an expression that can be solved. This will be more apparent in Chapter **??** dealing with hidden and missing variables.

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.

### *1.3.1 Two common inference problems*

1. **Model comparison (discrete hypothesis space):** Which one of 2 or more models (hypotheses) is most probable given our current state of knowledge? Examples:

   - Hypothesis or model $M_0$ asserts that the star has no planets.



**How to proceed in Bayesian data analysis?**

Identify the terms in Bayes' theorem and solve with the aid of the sum and product rules.

Prior probability

Likelihood

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

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

Figure 1.2 How to proceed in Bayesian data analysis.

- Hypothesis $M_1$ asserts that the star has one 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. Example:

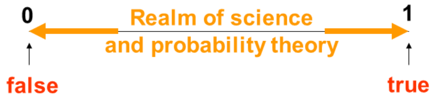
- Hypothesis $P$ asserts that the orbital period is between $P$ and $P + dP$.

For a continuous hypothesis space the same symbol $P$ is often employed in two different ways. When it appears as an argument of a probability, e.g., $p(P|D, I)$ it acts as a proposition (obeying the rules of Boolean algebra) and asserts that the true value of the parameter lies in the infinitesmal numerical range $P$ to $P + dP$. In other situations it acts as an ordinary algebraic variable standing for possible numerical values.

Figure 1.3 illustrates the significance of the extended logic provided by Bayesian probability theory. Deductive logic is just a special case of Bayesian probability theory in the idealized limit of complete information. For demonstration of this see Section 2.5.4 of my book [24].

**Significance of this development**

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



0

Realm of science and probability theory

1

false

true

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.

Figure 1.3 Significance of the extended logic provided by Bayesian probability theory.

### *Calculation of a simple likelihood $p(D|M,I)$*

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

$$d_i = f_i(X) + e_i, \tag{1.4}$$

where $X$ represents the set of model parameters and $e_i$ represents our knowledge of the measurement error which can be different for each data point (heteroscedastic data).

If the prior information $I$ indicates that distribution of the measurement errors are independent Gaussians, then

$$\begin{aligned} p(D_i|M,X,I) &= \frac{1}{\sigma_i\sqrt{2\pi}}\,Exp[-\frac{e_i^2}{2\sigma_i^2}] \\ &= \frac{1}{\sigma_i\sqrt{2\pi}}\,Exp\left[-\frac{(d_i-f_i(X))^2}{2\sigma_i^2}\right]. \end{aligned} \tag{1.5}$$

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

$$\begin{aligned} p(D|M,X,I) &= (2\pi)^{N/2}\left\{\prod_{i=1}^{N}\sigma_i^{-1}\right\}\,Exp\left[-0.5\sum_{i=1}^{N}\frac{(d_i-f_i(X))^2}{\sigma_i^2}\right] \\ &= (2\pi)^{N/2}\left\{\prod_{i=1}^{N}\sigma_i^{-1}\right\}\,Exp\left[-0.5\,\chi^2\right], \end{aligned} \tag{1.6}$$

where the summation within the square brackets is the familiar $\chi^2$ statistic that is minimized in the method of least-squares. Thus maximizing the likelihood corresponds to minimizing $\chi^2$.

Recall: Bayesian posterior ∝ prior × likelihood.

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

In the exoplanet problem the prior range for the unknown orbital period $P$ is very large from 1/2 day to 1000 yr (upper limit set by perturbations from neighboring stars). Suppose we assume a uniform prior probability density for the $P$ parameter. According to Equation 1.7, 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 decade from 1 to 10 d.

$$\frac{\int_{10^4}^{10^5} p(P|M,I)dP}{\int_1^{10} p(P|M,I)dP} = 10^4 \tag{1.7}$$

Usually, expressing great uncertainty in some quantity corresponds more closely to a statement of scale invariance or equal probability per decade (uniform on ln$P$. A scale invariant prior has this property. A recent analysis of the occurrence rate of transiting planets [49] versus orbital period found that the occurrence rate is constant, within 15%, between 12.5 and 100 d.

$$p(\ln P|M,I)\ d\ln P = \frac{d\ \ln P}{\ln(P_{max}/P_{min})} \tag{1.8}$$

### *1.3.2 Marginalization: an important Bayesian tool*

Suppose our model parameter set, $X$, consists of two continuous parameters $\theta$ and $\phi$. In parameter estimation, we are often interested in the implications of our current state of knowledge data $D$ and prior information $I$ for each parameter separately, independent of the values of the other parameters. As shown in Section 1.5 of my book [24], we can write

$$p(\theta|D,I) = \int d\phi\ p(\theta,\phi|D,I). \tag{1.9}$$

This can be expanded using Bayes' theorem. If our prior information for $\phi$ is independent of $\theta$, this yields

$$p(\theta|D,I) \propto p(\theta|I) \int d\phi\ p(\phi|I)p(D|\theta,\phi,I). \tag{1.10}$$

This gives the marginal[3] posterior distribution $p(\theta|D,I)$, in terms of the weighted average of the likelihood function, $p(D|\theta,\phi,I)$, weighted by $p(\phi|I)$, the prior probability density function for $\phi$. This operation marginalizes out the $\phi$ parameter. For exoplanet detection, the need to be to fit at least an 8 planet model to the data requires integration in 42 dimensions. The integral in Equation (1.9) can sometimes be evaluated analytically which can greatly reduce the computational aspects of the problem especially when many parameters are involved. If the joint posterior in $\theta,\phi$ is non Gaussian then the marginal distribution $p(\theta|D,I)$ can look very different from the projection of the joint posterior onto the $\theta$ axis as Figure 11.3 in my book [24] clearly demonstrates. This is because the marginal, $p(\theta|D,I)$, for any particular choice of $\theta$, is proportional to the integral over $\phi$ which depends both on width of the distribution in the $\phi$ coordinate as well as the height.

In Bayesian model comparison, we are interested in the most probable model, independent of the model parameters (i.e., marginalize out all parameters). This is illustrated in the equation below for model $M_2$.

$$p(M_2|D,I) = \int_{\Delta X} dX\; p(M_2,X|D,I), \tag{1.11}$$

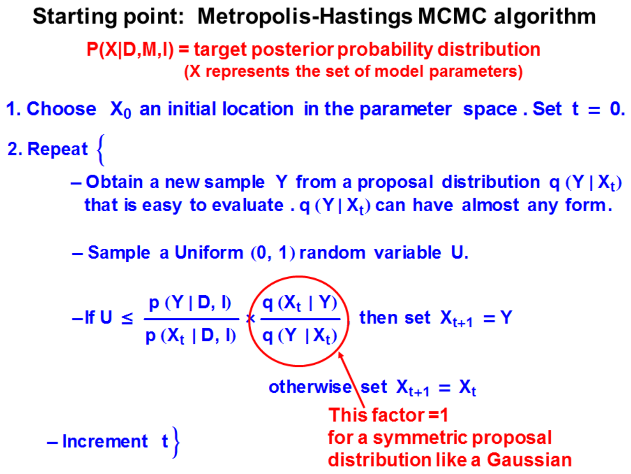
where $\Delta X$ designates the appropriate range of integration for the set of model parameters designated by $X$, as specified by our prior information, $I$.

### *Integration and Markov chain Monte Carlo (MCMC)*

It should be clear from the above that a full Bayesian analysis involves integrating over model parameter spaces. 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 lengthy additional calculations, i.e., Monte Carlo simulations. Fortunately the Markov chain Monte Carlo (MCMC) algorithms [48] provide a powerful means for efficiently computing integrals in many dimensions to within a constant factor. This factor is not required for parameter estimation.

The output at each iteration of the MCMC is a sample from the model parameter space of the desired joint posterior distribution (called the target distribution). All MCMC algorithms generate the desired samples by constructing a kind of random walk in the model parameter space. The random walk is accomplished using a Markov chain, whereby the new sample, $X_{t+1}$ , depends on previous sample $X_t$ according to an entity called the transition probability or transition kernel, $p(X_{t+1}|X_t)$. The transition kernel is assumed to be time independent. Each sample is correlated

[3] Since a parameter of a model is not a random variable, the frequentist statistical approach is denied the concept of the probability of a parameter.



**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 $q(Y|X_t)$
      that is easy to evaluate. $q(Y|X_t)$ can have almost any form.
    – Sample a Uniform (0, 1) random variable U.
    – If $U \le \frac{p(Y|D,I)}{p(X_t|D,I)} \times \frac{q(X_t|Y)}{q(Y|X_t)}$ 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

Figure 1.4 The Metropolis-Hastings algorithm. In this example the same Gaussian proposal distribution is used for both parameters.

with nearby samples[4]. The remarkable property of $p(X_{t+1}|X_t)$ is that after an initial burn-in period (which is discarded) it generates samples of $X$ with a probability density equal to the desired joint posterior probability distribution of the parameters.

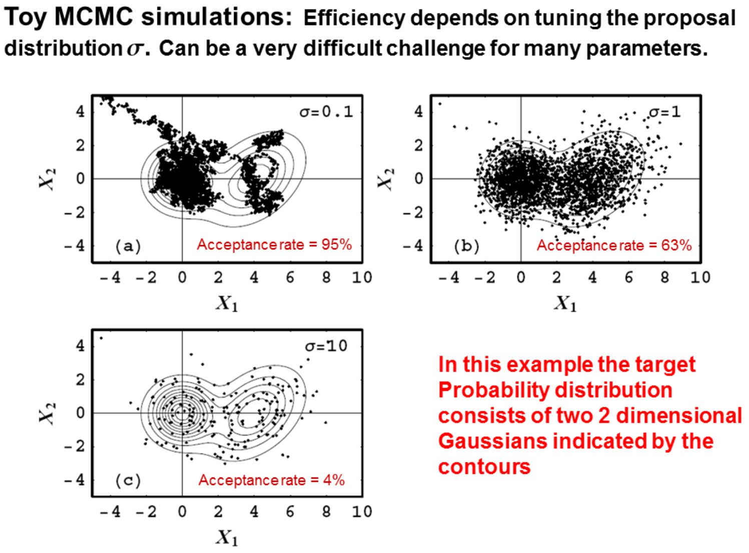
The marginal posterior probability density function (PDF) for any single parameter is given by a histogram of that component of the sample for all post burn-in iterations. Because the density of MCMC samples is proportional to the joint posterior probability distribution, it doesnt waste time exploring regions where the joint posterior density is very small in contrast to straight Mont Carlo integration.

In general the target distribution is complex and difficult to draw samples from. Instead new samples are drawn from a distribution which is easy to sample from, like a multivariate Normal with mean equal to the current $X_t$. Figure 1.4 shows the operation of a Metropolis-Hastings MCMC algorithm[5]. In this example the same Gaussian proposal distribution is used for both parameters. For details on why the MetropolisHastings algorithm works, see Section 12.3 of my book [24].

Figure 1.5 shows the behavior of the Metropolis-Hastings algorithm for a simple toy target distribution consisting of two 2 dimensional Gaussians. A single Gaussian proposal distribution (with a different choice of $\sigma$ in each panel) was em-

[4] Care must be taken if independent samples are desired (typically by thinning the resulting chain of samples by only taking every $n^{\text{th}}$ value, e.g. every 100th value).

[5] Gibbs sampling is a special case of the MetropolisHastings algorithm which is frequently employed. Gibbs sampling is applicable when the joint distribution is not known explicitly or is difficult to sample from directly, but the conditional distribution of each variable is known and is relatively easy to sample from. One advantage is that it does not require the tuning of proposal distributions.



**Toy MCMC simulations: Efficiency depends on tuning the proposal distribution $\sigma$. Can be a very difficult challenge for many parameters.**

**In this example the target Probability distribution consists of two 2 dimensional Gaussians indicated by the contours**

Figure 1.5 Toy Metopolis-Hastings MCMC simulations for a range of poposal distribution $\sigma$.

ployed for both parameters, $X_1, X_2$. Each simulation started from the same location (top left) and used the same number of iterations. For $\sigma = 0.1$ the acceptance rate is very high at 95% and the samples are strongly correlated. For $\sigma = 1$ the acceptance rate is 63% and the correlations much weaker. One can barely detect the burn-in samples in this case. For $\sigma = 10$ the acceptance rate is only 4%. In this case most samples were rejected resulting in many repeats of the current sample. Simulation (b) is close to ideal, while the other two would have to be run much longer to achieve reasonable sampling of the underlying target distribution. Based on empirical studies, Roberts, Gelman, and Gilks [52] recommend calibrating the acceptance rate to about 25% for high-dimensional models and to about 50% for models of one or two dimensions.

## 1.4 Fusion MCMC

Frequently, MCMC algorithms have been augmented with an additional tool such as parallel tempering, simulated annealing or differential evolution depending on the complexity of the problem. The exoplanet detection problem is particularly challenging because of the large search range in period space coupled with the sparse sampling in time which gives rise to many peaks in the target probability distribution. The goal of fusion MCMC (FMCMC) [31] has been to fuse together the advantages of all of the above tools together with a genetic crossover operation in a single automated MCMC algorithm to facilitate the detection of a global minimum in $\chi^2$ (maximum posterior probability in the Bayesian context).

To achieve this, a unique multi-stage adaptive control system was developed that automates the tuning of the proposal distributions for efficient exploration of the model parameter space even when the parameters are highly correlated. The FMCMC algorithm is currently implemented in *Mathematica* using parallelized code and run on an 8 core PC. When implemented with a multi-planet Kepler model[6], it is able to identify any significant periodic signal component in the data that satisfies Kepler's laws and function as a multi-planet Kepler periodogram[7].

The adaptive FMCMC is intended as a very general Bayesian nonlinear model fitting program. After specifying the model, $M_i$, the data, $D$, and priors, $I$, Bayes' theorem dictates the target joint probability distribution for the model parameters which is given by

$$p(X|D, M_i, I) = C\ p(X|M_i, I) \times p(D|M_i, X, I). \tag{1.12}$$

where $C$ is the normalization constant which is not required for parameter estimation purposes and $X$ represent the set of model parameters. The term, $p(X|M_i, I)$, is the prior probability distribution of $X$, prior to the consideration of the current data $D$. The term, $p(D|, M_i, I)$, is called the likelihood and it is the probability that we would have obtained the measured data $D$ for this particular choice of parameter vector $X$, model $M_i$, and prior information $I$. At the very least, the prior information, $I$, must specify the class of alternative models being considered (hypothesis space of interest) and the relationship between the models and the data (how to compute the likelihood). In some simple cases the log of the likelihood is simply proportional to the familiar $\chi^2$ statistic. For further details of the likelihood function for this type of problem see Gregory [25].

### 1.4.1 Parallel tempering

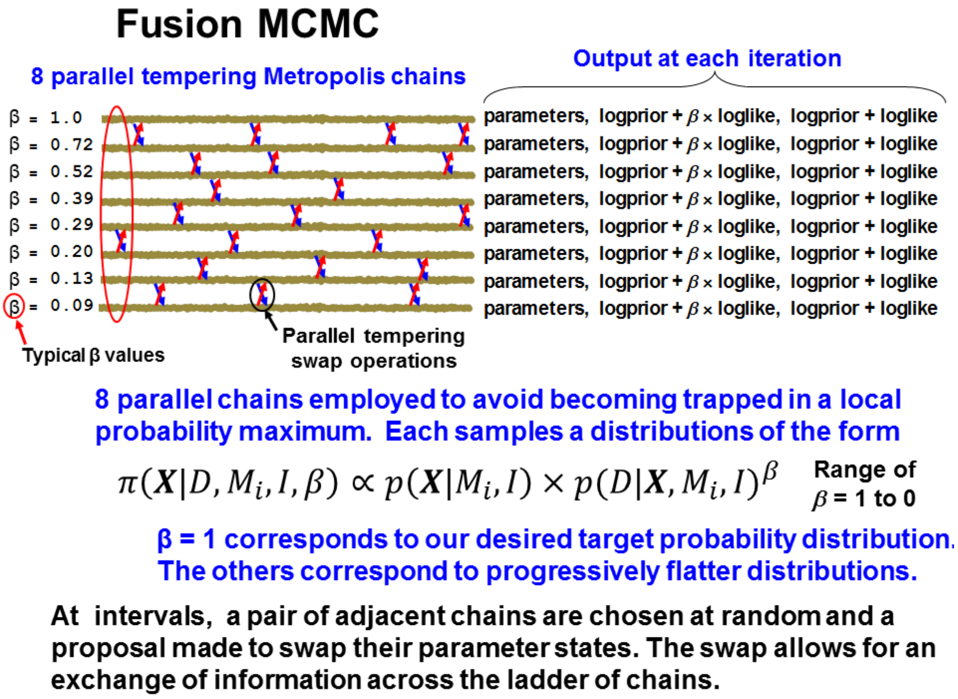
An important feature that prevents the fusion MCMC from becoming stuck in a local probability maximum is *parallel tempering* [23] (and re-invented under the name *exchange Monte Carlo* [36]). Multiple MCMC chains are run in parallel. The joint distribution for the parameters of model $M_i$, for a particular chain, is given by

$$\pi(X|D, M_i, I, \beta) \propto p(X|M_i, I) \times p(D|X, M_i, I)^{\beta}. \tag{1.13}$$

Each MCMC chain corresponding to a different $\beta$, with the value of $\beta$ ranging from zero to 1. When the exponent $\beta = 1$, the term on the LHS of the equation is the

[6] For multiple planet models, there is no analytic expression for the exact radial velocity perturbation. In many cases, the radial velocity perturbation can be well modeled as the sum of multiple independent Keplerian orbits which is what has been assumed in this work.

[7] Following on from the pioneering work on Bayesian periodograms by [38, 7]


Fusion MCMC
8 parallel tempering Metropolis chains
Output at each iteration
β = 1.0
β = 0.72
β = 0.52
β = 0.39
β = 0.29
β = 0.20
β = 0.13
β = 0.09
parameters, logprior + β × loglike, logprior + loglike
parameters, logprior + β × loglike, logprior + loglike
parameters, logprior + β × loglike, logprior + loglike
parameters, logprior + β × loglike, logprior + loglike
parameters, logprior + β × loglike, logprior + loglike
parameters, logprior + β × loglike, logprior + loglike
parameters, logprior + β × loglike, logprior + loglike
parameters, logprior + β × loglike, logprior + loglike
Typical β values
Parallel tempering swap operations

8 parallel chains employed to avoid becoming trapped in a local probability maximum. Each samples a distributions of the form

$$\pi(\boldsymbol{X}|D, M_i, I, \beta) \propto p(\boldsymbol{X}|M_i, I) \times p(D|\boldsymbol{X}, M_i, I)^{\beta}$$

Range of $\beta$ = 1 to 0

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

At intervals, a pair of adjacent chains 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 chains.

Figure 1.6 Parallel tempering schematic.

target joint probability distribution for the model parameters, $p(X|D, M_i, I)$. For $\beta \ll 1$, the distribution is much flatter.

In Equation 1.13, an exponent $\beta = 0$ yields a joint distribution equal to the prior. The reciprocal of $\beta$ is analogous to a temperature, the higher the temperature the broader the distribution. For parameter estimation purposes 8 chains were employed. A representative set of $\beta$ values is shown in Fig. 1.6. At an interval of 10 to 40 iterations, a pair of adjacent chains on the tempering ladder are chosen at random and a proposal made to swap their parameter states. A Monte Carlo acceptance rule determines the probability for the proposed swap to occur (e.g., Gregory [24], Equation 12.12). This swap allows for an exchange of information across the population of parallel simulations. In low $\beta$ (higher temperature) simulations, radically different configurations can arise, whereas in higher $\beta$ (lower temperature) states, a configuration is given the chance to refine itself. The lower $\beta$ chains can be likened to a series of scouts that explore the parameter terrain on different scales. The final samples are drawn from the $\beta = 1$ chain, which corresponds to the desired target probability distribution. The choice of $\beta$ values can be checked by computing the swap acceptance rate. When they are too far apart the swap rate drops to very low values. In this work a typical swap acceptance rate of $\approx 30\%$ was

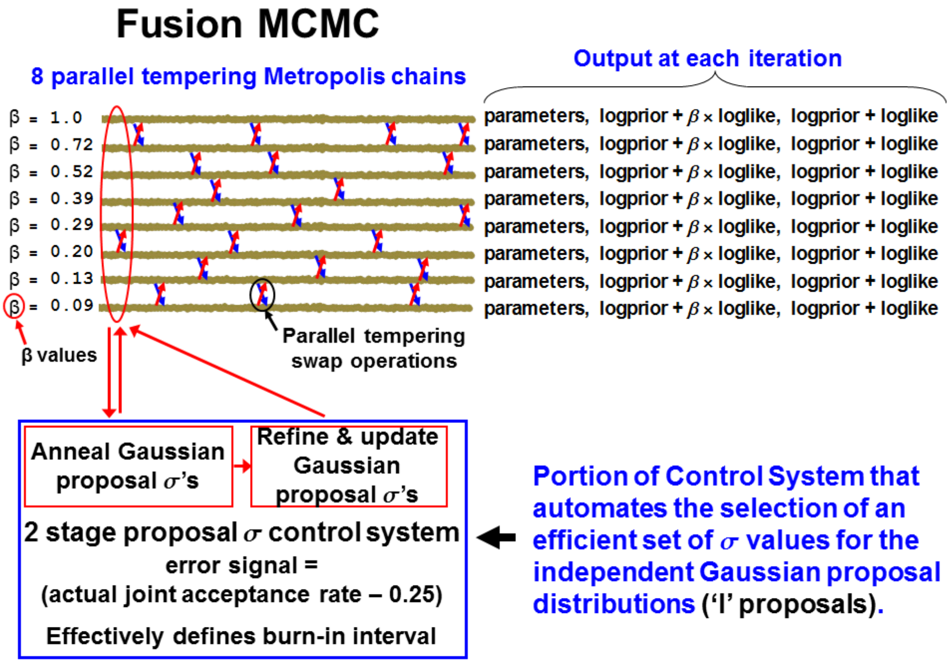

Fusion MCMC
8 parallel tempering Metropolis chains
Output at each iteration
β = 1.0
β = 0.72
β = 0.52
β = 0.39
β = 0.29
β = 0.20
β = 0.13
β = 0.09
parameters, logprior + β × loglike, logprior + loglike
β values
Parallel tempering swap operations
Anneal Gaussian proposal σ's
Refine & update Gaussian proposal σ's
2 stage proposal σ control system
error signal = (actual joint acceptance rate – 0.25)
Effectively defines burn-in interval
Portion of Control System that automates the selection of an efficient set of σ values for the independent Gaussian proposal distributions ('I' proposals).

Figure 1.7 First two stages of the adaptive control system.

employed but rates in a broad range from 0.15 to 0.5 were deemed acceptable as they did not exhibit any clear differences in performance. For a swap acceptance rate of 30%, jumps to adjacent chains will occur at an interval of ~ 230 to 920 iterations while information from more distant chains will diffuse much more slowly. Recently, Atchade et al. (2010) have shown that under certain conditions, the optimal swap acceptance rate is 0.234. An extension to the control system to automate the selection of an optimal set of $\beta$ values is described in Section A.5.

### 1.4.2 Fusion MCMC adaptive control system

At each iteration, a single joint proposal to jump to a new location in the parameter space is generated from independent Gaussian proposal distributions (centered on the current parameter location), one for each parameter. In general, the values of $\sigma$ for these Gaussian proposal distributions are different because the parameters can be very different entities. If the values of $\sigma$ are chosen too small, successive samples will be highly correlated and will require many iterations to obtain an equilibrium set of samples. If the values of $\sigma$ are too large, then proposed samples

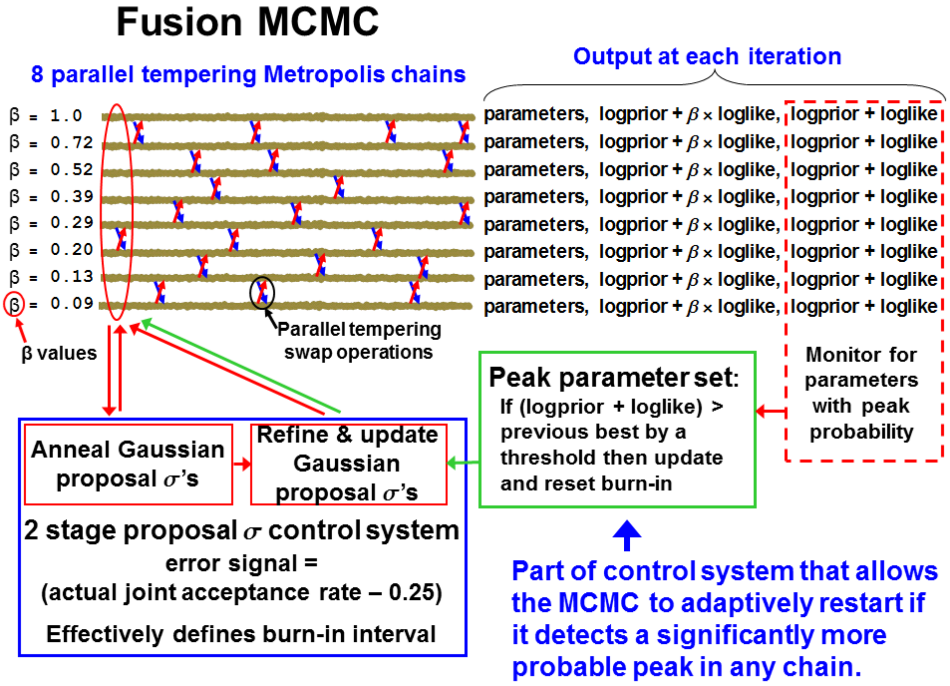

Fusion MCMC
8 parallel tempering Metropolis chains
Output at each iteration
β = 1.0
β = 0.72
β = 0.52
β = 0.39
β = 0.29
β = 0.20
β = 0.13
β = 0.09
parameters, logprior + β × loglike, logprior + loglike
β values
Parallel tempering swap operations
Monitor for parameters with peak probability
Peak parameter set:
If (logprior + loglike) > previous best by a threshold then update and reset burn-in
Anneal Gaussian proposal σ's
Refine & update Gaussian proposal σ's
2 stage proposal σ control system
error signal = (actual joint acceptance rate – 0.25)
Effectively defines burn-in interval
Part of control system that allows the MCMC to adaptively restart if it detects a significantly more probable peak in any chain.

Figure 1.8 Schematic illustrating how the second stage of the control system is restarted if a significantly more probable parameter set is detected.

will very rarely be accepted. The process of choosing a set of useful proposal values of $\sigma$ when dealing with a large number of different parameters can be very time consuming. In parallel tempering MCMC, this problem is compounded because of the need for a separate set of Gaussian proposal distributions for each tempering chain. This process is automated by an innovative statistical control system [27, 29] in which the error signal is proportional to the difference between the current joint parameter acceptance rate and a target acceptance rate [52], $\lambda$ (typically $\lambda \sim 0.25$). A schematic of the first two stages of the adaptive control system (CS) is shown[8] in Fig. 1.7. Details on the operation of the control system are given in Appendix A.

The adaptive capability of the control system can be appreciated from an examination of Fig. 1.8. The upper left portion of the figure depicts the FMCMC iterations from the 8 parallel chains, each corresponding to a different tempering level $\beta$ as indicated on the extreme left. One of the outputs obtained from each chain at every iteration (shown at the far right) is the log prior + log likelihood. This information is continuously fed to the CS which constantly updates the most

[8] The interval between tempering swap operations is typically much smaller than is suggested by this schematic.

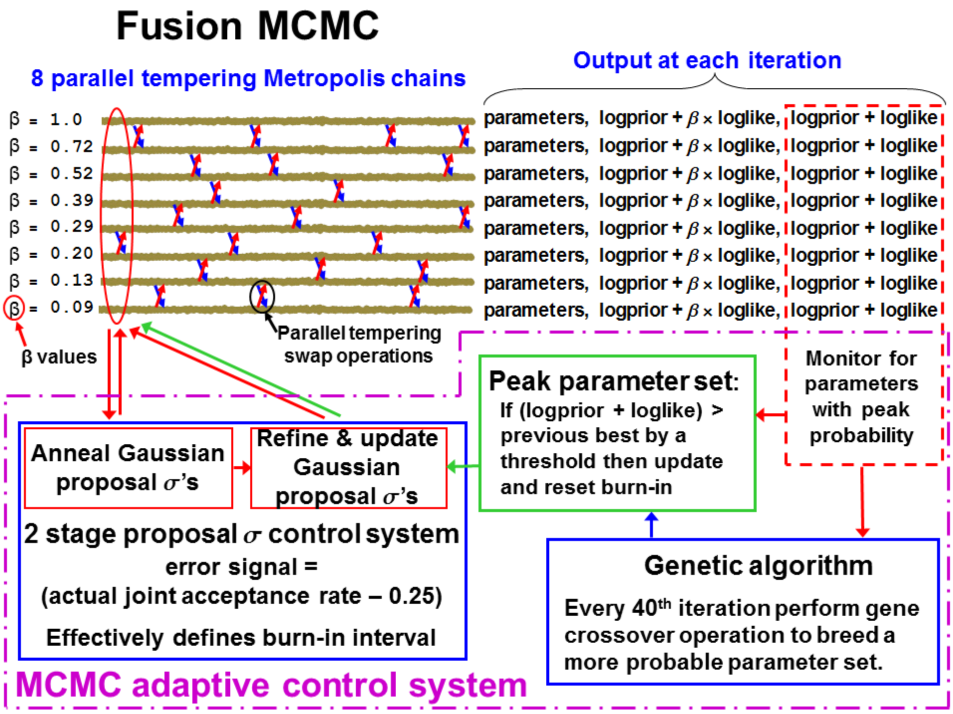

Fusion MCMC
8 parallel tempering Metropolis chains
Output at each iteration
β = 1.0
β = 0.72
β = 0.52
β = 0.39
β = 0.29
β = 0.20
β = 0.13
β = 0.09
parameters, logprior + β × loglike, logprior + loglike
parameters, logprior + β × loglike, logprior + loglike
parameters, logprior + β × loglike, logprior + loglike
parameters, logprior + β × loglike, logprior + loglike
parameters, logprior + β × loglike, logprior + loglike
parameters, logprior + β × loglike, logprior + loglike
parameters, logprior + β × loglike, logprior + loglike
parameters, logprior + β × loglike, logprior + loglike
β values
Parallel tempering swap operations
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 σ's
Refine & update Gaussian proposal σ's
2 stage proposal σ control system
error signal = (actual joint acceptance rate – 0.25)
Effectively defines burn-in interval
Genetic algorithm
Every 40th iteration perform gene crossover operation to breed a more probable parameter set.
MCMC adaptive control system

Figure 1.9 This schematic shows how the genetic crossover operation is integrated into the adaptive control system.

probable parameter combination regardless of which chain the parameter set occurred in. This is passed to the ‘Peak parameter set’ block of the CS. Its job is to decide if a significantly more probable parameter set has emerged since the last execution of the second stage CS. If so, the second stage CS is re-run using the new more probable parameter set which is the basic adaptive feature of the existing CS[9]. Fig. 1.8 illustrates how the second stage of the control system is restarted if a significantly more probable parameter set is detected regardless of which chain it occurs in. This also causes the burn-in phase to be extended.

The control system also includes a genetic algorithm block which is shown in the bottom right of Fig. 1.9. The current parameter set can be treated as a set of genes. In the present version, one gene consists of the parameter set that specify one orbit. On this basis, a three planet model has three genes. At any iteration there exist within the CS the most probable parameter set to date $X_{\text{max}}$, and the current

[9] *Mathematica* code that implements a recent version of fusion MCMC is available on the Cambridge University Press web site for my textbook [24], ‘Bayesian Logical data Analysis for the Physical Sciences’, in the free resources section. There you will also find ‘Additional book examples with *Mathematica* 8 tutorial’. Non *Mathematica* users can download a free Wolfram CDF Player to view the resource material.

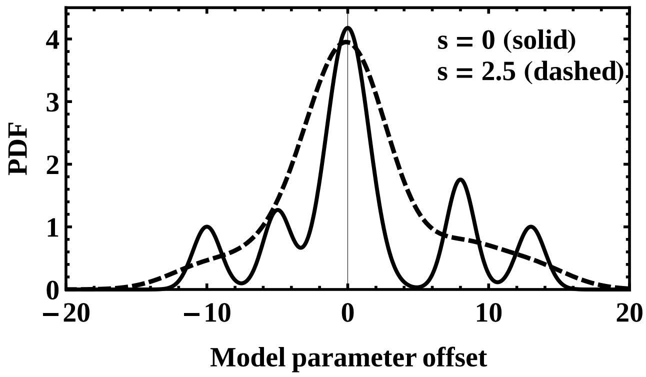


Figure 1.10 A simulated toy posterior probability distribution (PDF) for a single parameter model with (dashed) and without (solid) an extra noise term $s$.

iteration most probable parameter set of the 8 chains, $X_{\rm cur}$. At regular intervals (user specified) each gene from $X_{\rm cur}$ is swapped for the corresponding gene in $X_{\rm max}$. If either substitution leads to a higher probability it is retained and $X_{\rm max}$ updated. The effectiveness of this operation was tested by comparing the number of times the gene crossover operation gave rise to a new value of $X_{\rm max}$ compared to the number of new $X_{\rm max}$ arising from the normal parallel tempering MCMC operations. The gene crossover operations proved to be very effective, and gave rise to new $X_{\rm max}$ values $\approx$ 1.7 times more often than MCMC operations. It turns out that individual gene swaps from $X_{\rm cur}$ to $X_{\rm max}$ are much more effective (in one test by a factor of 17) than the other way around (reverse swaps). Since it costs just as much time to compute the probability for a swap either way we no longer carry out the reverse swaps. Instead, we have extended this operation to swaps from $X_{\rm cur2}$, the parameters of the second most probable current chain, to $X_{\rm max}$. This gave rise to new values of $\rm X_{max}$ at a rate $\sim$ 70% that of swaps from $X_{\rm cur}$ to $X_{\rm max}$. Crossover operations at a random point in the entire parameter set did not prove as effective except in the single planet case where there is only one gene.

### *1.4.3 Automatic simulated annealing*

The annealing of the proposal $\sigma$ values occurs while the MCMC is homing in on any significant peaks in the target probability distribution. Concurrent with this, another aspect of the annealing operation takes place whenever the Markov chain is started from a location in parameter space that is far from the best fit values. This automatically arises because all the models considered incorporate an extra additive noise [25], whose probability distribution is independent and identically distributed (IID) Gaussian with zero mean and with an unknown standard devia-

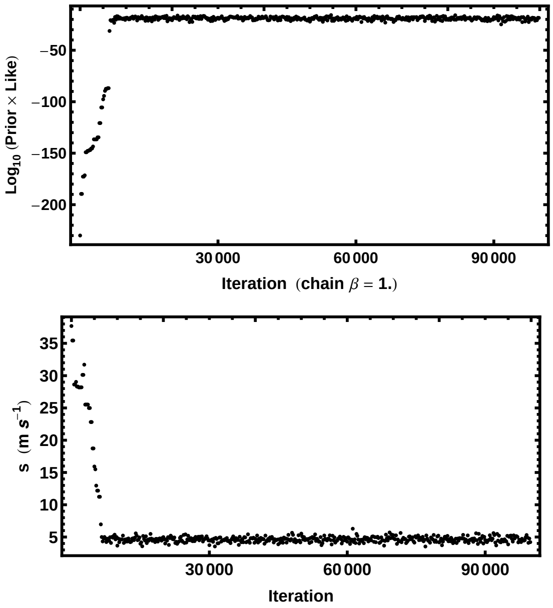
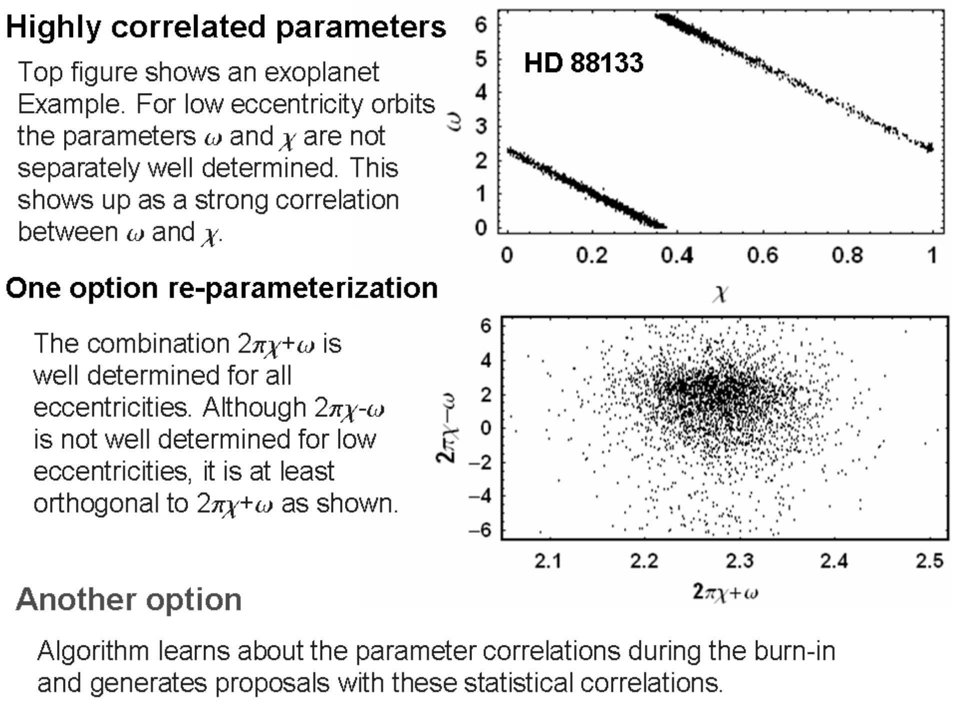


Figure 1.11 The upper panel is a plot of the Log$_{10}$[Prior × Likelihood] versus MCMC iteration. The lower panel is a similar plot for the extra noise term $s$. Initially $s$ is inflated and then rapidly decays to a much lower level as the best fit parameter values are approached.

tion $s$. When the $\chi^2$ of the fit is very large, the Bayesian Markov chain automatically inflates $s$ to include anything in the data that cannot be accounted for by the model with the current set of parameters and the known measurement errors. This results in a smoothing out of the detailed structure in the $\chi^2$ surface and, as pointed out by [20], allows the Markov chain to explore the large scale structure in parameter space more quickly. This is illustrated in Figure 1.10 which shows a simulated toy posterior probability distribution (PDF) for a single parameter model with (dashed) and without (solid) an extra noise term $s$. Figure 1.11 shows the behavior of Log$_{10}$[Prior × Likelihood] and $s$ versus MCMC iteration for a some real data. In the early stages $s$ is inflated to around 38 m s$^{-1}$ and then decays to a value of $\approx$ 4 m s$^{-1}$ over the first 9,000 iterations as Log$_{10}$[Prior × Likelihood] reaches a maximum. This is similar to simulated annealing, but does not require choosing a cooling scheme.

### *1.4.4 Highly correlated parameters*

For some models the data is such that the resulting estimates of the model parameters are highly correlated and the MCMC exploration of the parameter space can



**Highly correlated parameters**

Top figure shows an exoplanet Example. For low eccentricity orbits the parameters $\omega$ and $\chi$ are not separately well determined. This shows up as a strong correlation between $\omega$ and $\chi$.

**One option re-parameterization**

The combination $2\pi\chi+\omega$ is well determined for all eccentricities. Although $2\pi\chi-\omega$ is not well determined for low eccentricities, it is at least orthogonal to $2\pi\chi+\omega$ as shown.

**Another option**

Algorithm learns about the parameter correlations during the burn-in and generates proposals with these statistical correlations.

Figure 1.12 An example of two highly correlated parameters and possible ways of dealing with this issue which includes a transformation to more orthogonal parameter set.

be very inefficient. Fig. 1.12 shows an example of two highly correlated parameters and possible ways of dealing with this issue which includes a transformation to more orthogonal parameter set. It would be highly desirable to employ a method that automatically samples correlated parameters efficiently. One potential solution in the literature is Differential Evolution Markov Chain (DE-MC) [6]. DE-MC is a population MCMC algorithm, in which multiple chains are run in parallel, typically from 15 to 40. DE-MC solves an important problem in MCMC, namely that of choosing an appropriate scale and orientation for the jumping distribution.

For the fusion MCMC algorithm, I developed and tested a new method [30], in the spirit of DE, that automatically achieves efficient MCMC sampling in highly correlated parameter spaces without the need for additional chains. The block in the lower left panel of Fig. 1.13 automates the selection of efficient proposal distributions when working with model parameters that are independent or transformed to new independent parameters. New parameter values are jointly proposed based on independent Gaussian proposal distributions (‘I’ scheme), one for each parameter. Initially, only this ‘I’ proposal system is used and it is clear that if there are strong

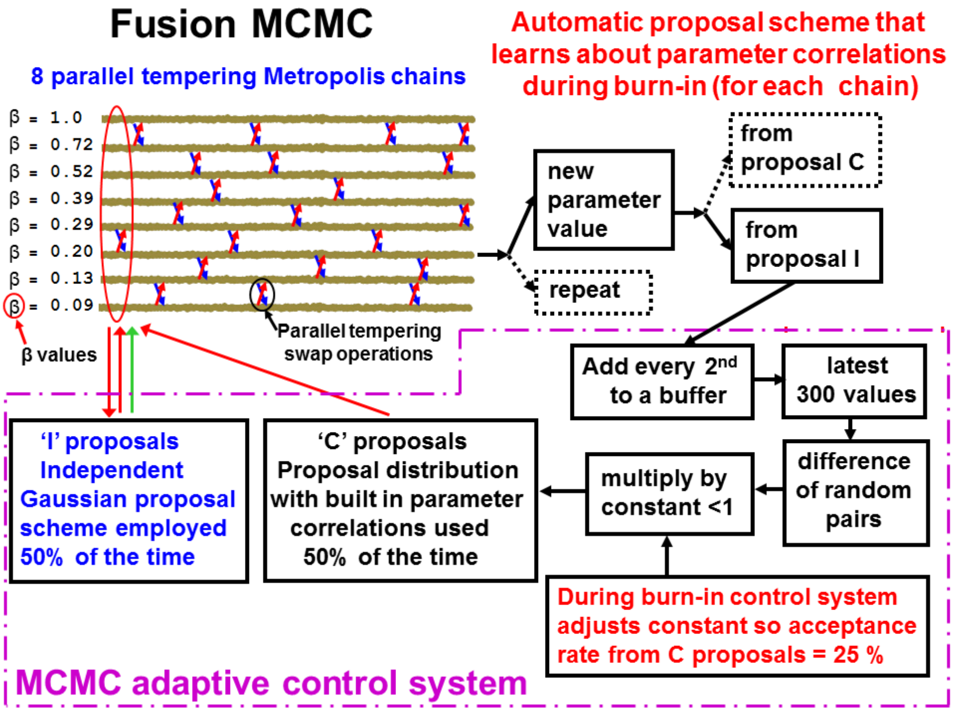

Fusion MCMC
8 parallel tempering Metropolis chains
β = 1.0
β = 0.72
β = 0.52
β = 0.39
β = 0.29
β = 0.20
β = 0.13
β = 0.09
β values
Parallel tempering swap operations
Automatic proposal scheme that learns about parameter correlations during burn-in (for each chain)
new parameter value
from proposal C
from proposal I
repeat
Add every 2nd to a buffer
latest 300 values
difference of random pairs
multiply by constant <1
'I' proposals Independent Gaussian proposal scheme employed 50% of the time
'C' proposals Proposal distribution with built in parameter correlations used 50% of the time
During burn-in control system adjusts constant so acceptance rate from C proposals = 25 %
MCMC adaptive control system

Figure 1.13 This schematic illustrates the automatic proposal scheme for handling correlated ('C') parameters.

correlations between any parameters the $\sigma$ values of the independent Gaussian proposals will need to be very small for any proposal to be accepted and consequently convergence will be very slow. However, the accepted 'I' proposals will generally cluster along the correlation path. In the optional third stage of the control system every second [10] accepted 'I' proposal is appended to a correlated sample buffer. There is a separate buffer for each parallel tempering level. Only the 300 most recent additions to the buffer are retained. A 'C' proposal is generated from the difference between a pair of randomly selected samples drawn from the correlated sample buffer for that tempering level, after multiplication by a constant. The value of this constant (for each tempering level) is computed automatically [30] by another control system module which ensures that the 'C' proposal acceptance rate is close to 25%. With very little computational overhead, the 'C' proposals provide the scale and direction for efficient jumps in a correlated parameter space.

The final proposal distribution is a random selection of 'I' and 'C' proposals such that each is employed 50% of the time. The combination ensures that the

[10] Thinning by a factor of 10 has already occurred meaning only every tenth iteration is recorded.

whole parameter space can be reached and that the FMCMC chain is aperiodic. The parallel tempering feature operates as before to avoid becoming trapped in a local probability maximum.

Because the 'C' proposals reflect the parameter correlations, large jumps are possible allowing for much more efficient movement in parameter space than can be achieved by the 'I' proposals alone. Once the first two stages of the control system have been turned off, the third stage continues until a minimum of an additional 300 accepted 'I' proposals have been added to the buffer and the 'C' proposal acceptance rate is within the range $\geq 0.22$ and $\leq 0.28$. At this point further additions to the buffer are terminated and this sets a lower bound on the burn-in period.

*Tests of the 'C' proposal scheme*

Gregory [30] carried out two tests of the 'C' proposal scheme using (a) simulated exoplanet astrometry data, and (b) a sample of real radial velocity data. In the latter test we analyzed a sample of seventeen HD 88133 precision radial velocity measurements [18] using a single planet model in three different ways. Fig. 1.14 shows a comparison of the resulting post burn-in marginal distributions for two correlated parameters $\chi$ and $\omega$, together with a comparison of the autocorrelation functions. The black trace corresponds to a search in $\chi$ and $\omega$ using only 'I' proposals. The red trace corresponds to a search in $\chi$ and $\omega$ with 'C' proposals turned on. The green trace corresponds to a search in the transformed orthogonal coordinates $\psi = 2\pi\chi + \omega$ and $\phi = 2\pi\chi - \omega$ using only 'I' proposals. It is clear that a search in $\chi$ and $\omega$ with 'C' proposals turned on achieves the same excellent results as a search in the transformed orthogonal coordinates $\psi$ and $\phi$ using only 'I' proposals.

## 1.5 Exoplanet applications

As previously mentioned the FMCMC algorithm is designed to be a very general tool for nonlinear model fitting. When implemented with a multi-planet Kepler model it is able to identify any significant periodic signal component in the data that satisfies Kepler's laws and is able to function as a multi-planet Kepler periodogram. This approach leads to the detection of planetary candidates. One reason to think of them is planetary candidates is because it is known that stallar activity (spots, and larger scale magnetically active regions) can lead to RV artifacts signals (e.g., [50, 53]). A great deal of attention is now being focussed on correlating stellar activity signals with those found in the RV data. Also it is necessary to carry out N-body simulations to establish the long term stability of the remaining candidate planets.

In this section we describe the model fitting equations and the selection of priors for the model parameters. For a one planet model the predicted radial velocity is

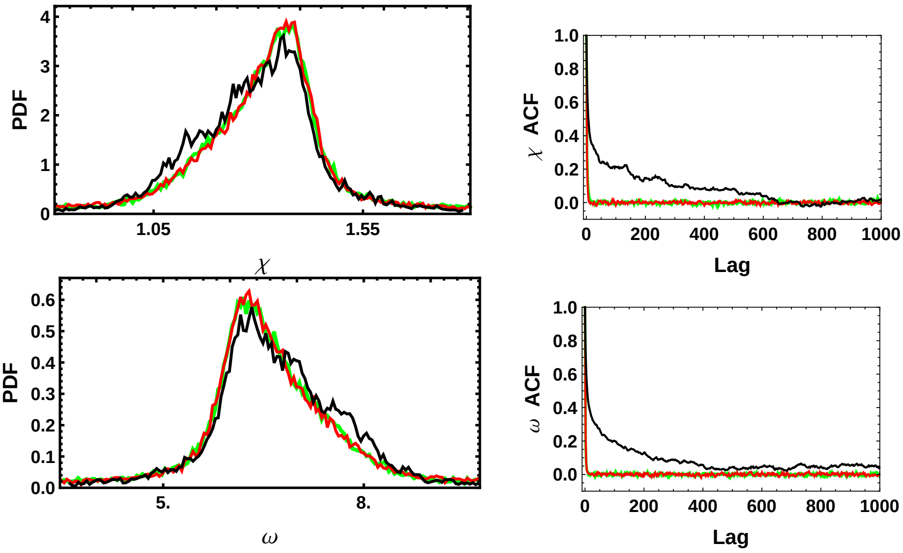


Figure 1.14 The two panels on the left show a comparison of the post burn-in marginal distributions for $\chi$ and $\omega$. The two panels on the right show a comparison of their MCMC autocorrelation functions. The black trace corresponds to a search in $\chi$ and $\omega$ using only 'I' proposals. The red trace corresponds to a search in $\chi$ and $\omega$ with 'C' proposals turned on. The green trace corresponds to a search in the transformed orthogonal coordinates $\psi = 2\pi\chi + \omega$ and $\phi = 2\pi\chi - \omega$ using only 'I' proposals.

given by

$$f(t_i) = V + K[\cos\{\theta(t_i + \chi P) + \omega\} + e\cos\omega], \tag{1.14}$$

and involves the 6 unknown parameters

- $V$ = a constant velocity.
  $K$ = velocity semi-amplitude = $\frac{2\pi}{P}\frac{a\ \sin i}{\sqrt{1-e^2}}$,
  where $a$ = semi-major axis and $i$ = inclination.
- $P$ = the orbital period.
- $e$ = the orbital eccentricity.
- $\omega$ = the longitude of periastron.
- $\chi$ = the fraction of an orbit, prior to the start of data taking, that periastron occurred at. Thus, $\chi P$ = the number of days prior to $t_i = 0$ that the star was at periastron, for an orbital period of P days.
- $\theta(t_i{+}\chi P)$ = the true anomaly, the angle of the star in its orbit relative to periastron at time $t_i$.

We utilize this form of the equation because we obtain the dependence of $\theta$ on $t_i$

by solving the conservation of angular momentum equation

$$\frac{d\theta}{dt} - \frac{2\pi[1 + e\cos\theta(t_i + \chi\, P)]^2}{P(1 - e^2)^{3/2}} = 0. \tag{1.15}$$

Our algorithm is implemented in *Mathematica* and it proves faster for *Mathematica* to solve this differential equation than solve the equations relating the true anomaly to the mean anomaly via the eccentric anomaly. *Mathematica* generates an accurate interpolating function between $t$ and $\theta$ so the differential equation does not need to be solved separately for each $t_i$. Evaluating the interpolating function for each $t_i$ is very fast compared to solving the differential equation. Details on how Equation 1.15 is implemented are given in the Appendix C [31].

We employed a re-parameterization of $\chi$ and $\omega$ to improve the MCMC convergence speed motivated by the work of Ford (2006). The two new parameters are $\psi = 2\pi\chi + \omega$ and $\phi = 2\pi\chi - \omega$. Parameter $\psi$ is well determined for all eccentricities. Although $\phi$ is not well determined for low eccentricities, it is at least orthogonal to the $\psi$ parameter. We use a uniform prior for $\psi$ in the interval 0 to $4\pi$ and uniform prior for $\phi$ in the interval $-2\pi$ to $+2\pi$. This insures that a prior that is wraparound continuous in $(\chi, \omega)$ maps into a wraparound continuous distribution in $(\psi, \phi)$. To account for the Jacobian of this re-parameterization it is necessary to multiply the Bayesian integrals by a factor of $(4\pi)^{-nplan}$, where $nplan$ = the number of planets in the model. By utilizing the orthogonal combination $(\psi, \phi)$, there is less need to make use of the 'C' proposal scheme outlined in Section 1.4.4 but to allow for other possible correlations (e.g., a planet with a period greater than the data duration) it is safest to always make use of the 'C' proposal scheme as well.

### *1.5.1 Exoplanet priors*

In a Bayesian analysis we need to specify a suitable prior for each parameter. These are tabulated in Table 1.1. For the current problem, the prior given in Equation 1.13 is the product of the individual parameter priors. Detailed arguments for the choice of each prior are given in Appendix B [26, 29].

As mentioned in Section 1.4.3, all of the models considered in this paper incorporate an extra noise parameter, $s$, that can allow for any additional noise beyond the known measurement uncertainties [11]. We assume the noise variance is finite and adopt a Gaussian distribution with a variance $s^2$. Thus, the combination of the known errors and extra noise has a Gaussian distribution with variance = $\sigma_i^2 + s^2$, where $\sigma_i$ is the standard deviation of the known noise for ith data point. For example,

[11] In the absence of detailed knowledge of the sampling distribution for the extra noise, we pick a Gaussian because for any given finite noise variance it is the distribution with the largest uncertainty as measured by the entropy, i.e., the maximum entropy distribution [37] and [24] (section 8.7.4.)

Table 1.1 *Prior parameter probability distributions.*

| Parameter | Prior | Lower bound | Upper bound |
|---|---|---|---|
| Orbital frequency | $p(\ln f_1, \ln f_2, \cdots \ln f_n \vert M_n, I) = \frac{n!}{[\ln(f_H/f_L)]^n}$<br>($n$ =number of planets) | 1/0.5 d | 1/1000 yr |
| Velocity $K_i$ (m s$^{-1}$) | Modified scale invariant[a]<br>$\frac{(K+K_0)^{-1}}{\ln\left[1+\frac{K_{\max}}{K_0}\left(\frac{P_{\min}}{P}\right)^{1/3}\frac{1}{\sqrt{1-e^2}}\right]}$ | 0 (K$_0$ = 1) | $K_{\max}\left(\frac{P_{\min}}{P}\right)^{1/3}\frac{1}{\sqrt{1-e^2}}$<br>$K_{\max} = 2129$ |
| V (m s$^{-1}$) | Uniform | $-K_{\max}$ | $K_{\max}$ |
| $e$ Eccentricity | $3.1(1-e)^{2.1}$ | 0 | 0.99 |
| $\chi$ orbit fraction | Uniform | 0 | 1 |
| $\omega$ Longitude of periastron | Uniform | 0 | $2\pi$ |
| $s$ Extra noise (m s$^{-1}$) | $\frac{(s+s_0)^{-1}}{\ln\left(1+\frac{s_{\max}}{s_0}\right)}$ | 0 (s$_0$ = 1 to 10) | $K_{\max}$ |

[a] Since the prior lower limits for $K$ and $s$ include zero, we used a modified scale invariant prior of the form

$$p(X|M,I) = \frac{1}{X+X_0}\,\frac{1}{\ln\left(1+\frac{X_{\max}}{X_0}\right)} \tag{1.16}$$

For $X \ll X_0$, $p(X|M,I)$ behaves like a uniform prior and for $X \gg X_0$ it behaves like a scale invariant prior. The $\ln\left(1+\frac{X_{\max}}{X_0}\right)$ term in the denominator ensures that the prior is normalized in the interval 0 to $X_{\max}$.

suppose that the star actually has two planets, and the model assumes only one is present. In regard to the single planet model, the velocity variations induced by the unknown second planet acts like an additional unknown noise term. Other factors like star spots and chromospheric activity can also contribute to this extra velocity noise term which is often referred to as stellar jitter. In general, nature is more complicated than our model and known noise terms. Marginalizing $s$ has the desirable effect of treating anything in the data that can't be explained by the model and known measurement errors as noise, leading to conservative estimates of orbital parameters. See Sections 9.2.3 and 9.2.4 of [24] for a tutorial demonstration of this point. If there is no extra noise then the posterior probability distribution for $s$ will peak at $s = 0$. The upper limit on $s$ was set equal to $K_{\max}$. This is much larger than the estimates of stellar jitter for individual stars based on statistical correlations with observables (e.g., [54, 55, 67]). In our Bayesian analysis, $s$ serves two purposes. First it allows for an automatic simulated annealing operation as described in Section 1.4.3 and for this purpose it is desirable to have a much larger range. The final $s$ value after the annealing is complete provides a crude measure of the residuals that can't be accounted for by the model and known measurement uncertainties. Of course, the true residuals will exhibit correlations if there are additional planets present not specified by the current model. In addition, stellar activity RV artifacts can lead to correlated noise and a number of attempts are being explored to jointly model the planetary signals and stellar activity diagnostics (e.g., [51, 1, 60, 35]). These correlations are not accounted for by this simple additional noise term. We use the same prior range for $s$ for all the models ranging from the zero planet case to the many planet case. We employed a modified scale invariant prior for $s$ with a knee, $s_0$ in the range $1 - 10$m s$^{-1}$, according to Equation (1.15).

### *1.5.2 HD 208487 example*

In 2007, Gregory [26], using an automatic Bayesian multi-planet Kepler periodogram, found evidence for a second planetary candidate with a period of $\sim$ 900 d in HD 208487. We use this as an example data set to illustrate a number of issues that can arise in the analysis using an FMCMC powered multi-planet Kepler periodogram. Figure 1.15 shows sample FMCMC traces for the two planet fit to the 35 radial velocity measurements [61] for HD 208487 based on our latest version of the FMCMC algorithm and employing the updated eccentricity prior. The top left panel is a display of the Log$_{10}$ Prior × Likelihood versus FMCMC iteration number. In total $5 \times 10^5$ iterations were executed and only every 10$^{\text{th}}$ value saved. It is clear from this trace that the burn-in period is very short. In this example the control system ceased tuning 'I' and 'C' proposal distributions at iteration 3220 iterations. The top right panel shows the trace for the extra noise parameter. During the automatic annealing operation it dropped from a high around 18 m s$^{-1}$ to an equilibrium value of around 1 m s$^{-1}$ within the burn-in period. As explained in Appendix B.2 it is more efficient to allow the individual orbital frequency parameters to roam over the entire frequency space and re-label afterwards so that parameters associated with the lower frequency are always identified with planet one and vice versa. In this approach nothing constrains $f_1$ to always be below $f_2$ so that degenerate parameter peaks often occur. This behavior can be seen clearly in the panels for $P_1$ and $P_2$, where there are frequent transitions between the 130 and 900 d periods. The lower four panels show corresponding transitions occurring for the other orbital parameters. The traces shown in Figure 1.15 are before relabeling and those in Figure 1.16 afterwards.

Figure 1.17 illustrates a variety of different types of two planet periodogram plots for the HD208487 data. The top left shows the evolution of the two period parameters (after relabeling) from their starting values marked by the two dots that occur before the zero on the iteration axis. With the default scale invariant orbital frequency prior only two periods were detected. The top right panel shows a sam-

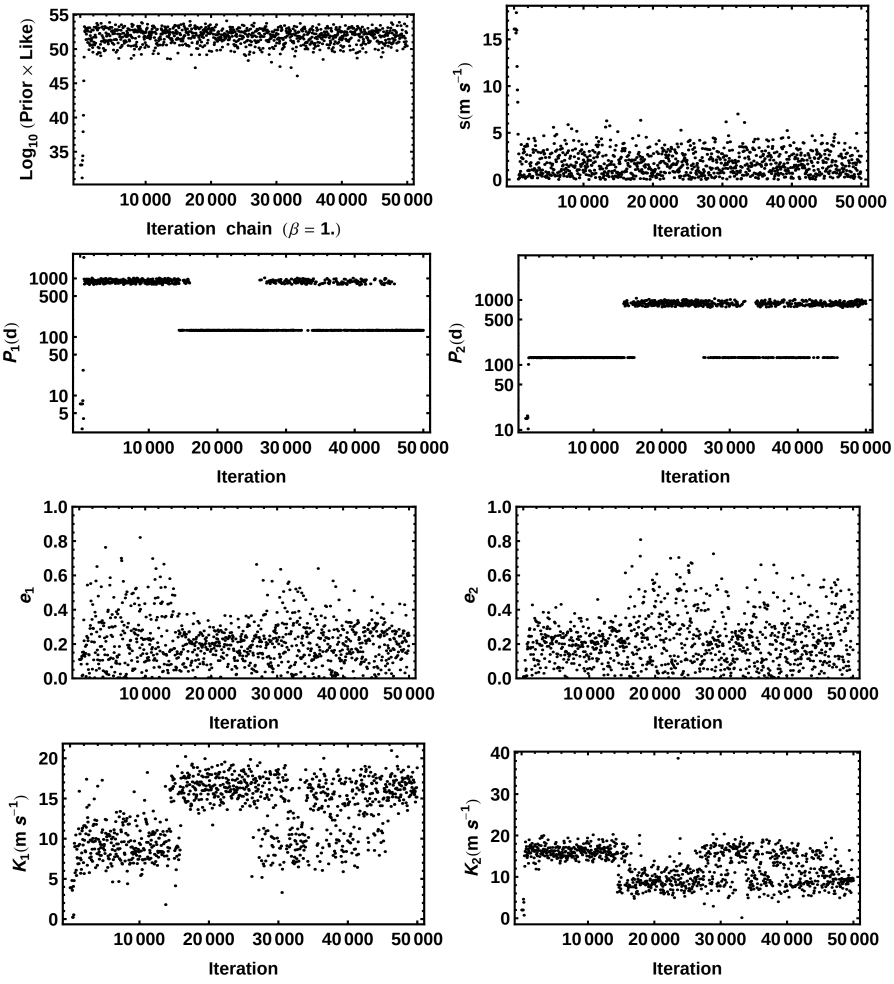


Figure 1.15 Sample FMCMC traces for a two planet fit to HD 208487 radial velocity data before relabeling.

ple of the two period parameter values versus a normalized value of Log$_{10}$[Prior × Likelihood]. The bottom left shows a plot of eccentricty verus period and the bottom right $K$ versus eccentricity. It is clear from the latter plot that eccentricity values are heavily concentrated around a value of 0.2 for the 130 d signal. The distribution of eccentricity for the secondary period is broader but the greatest concentration is towards low values. The MAP values are shown by the filled black circles. The combination of the lower two panels indicates that the eccentricity of the secondary period is lowest towards the 800 d end of the period range.

Figure 1.18 shows a plot of $K$ versus eccentricity for a one planet fit to the HD

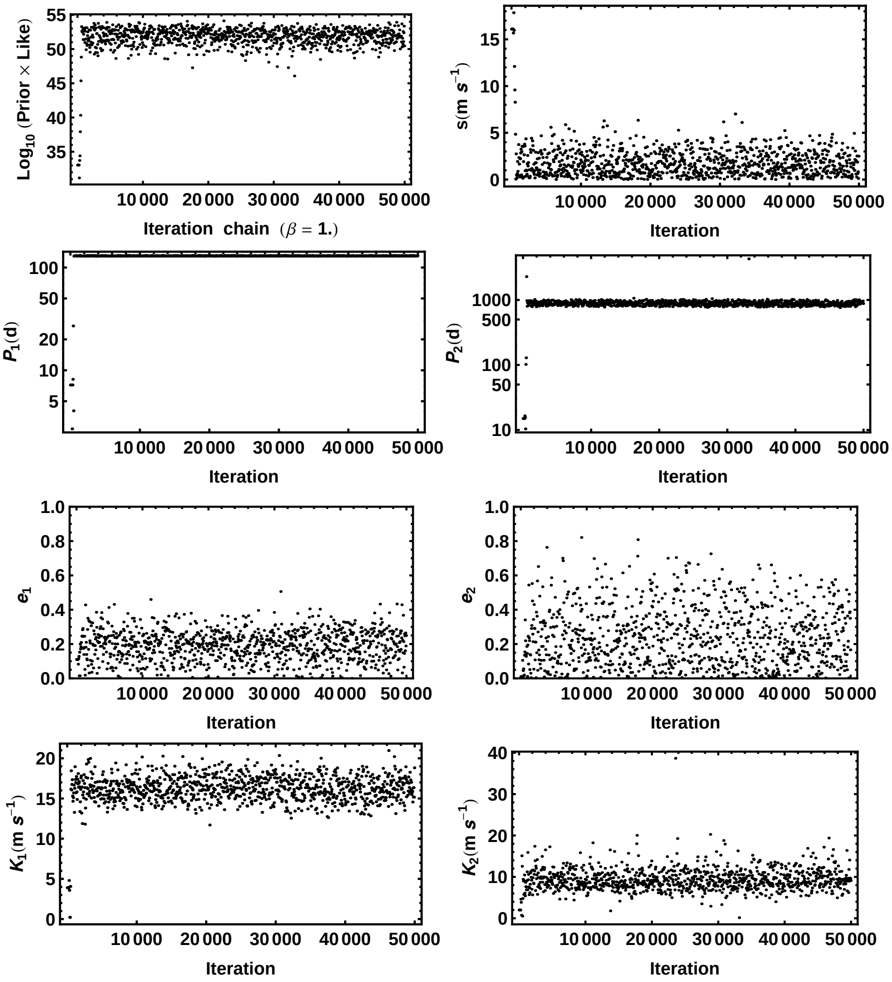


Figure 1.16 Sample FMCMC traces for a two planet fit to HD 208487 radial velocity data after relabeling.

208487 data for comparison. Clearly the single planet fit finds a larger $K$ value for the dominant 130 d period. The MAP solution is shown by the filled black circle. Even for the single planet fit there is a preference for an eccentrincity of $\sim$ 0.2.

Panel (a) of Figure 1.19 shows the radial velocity data [61]. Panel (b) and (c) show the two planet fit to the data and the fit residuals, respectively. Figure 1.20 shows the FMCMC marginal distributions for a subset of the two planet fit parameters for HD 208487. The dominant 130 d peak prefers a modest eccentricity of $\sim$ 0.2. The secondary period of 900 d exhibits a broader eccentricity range with a

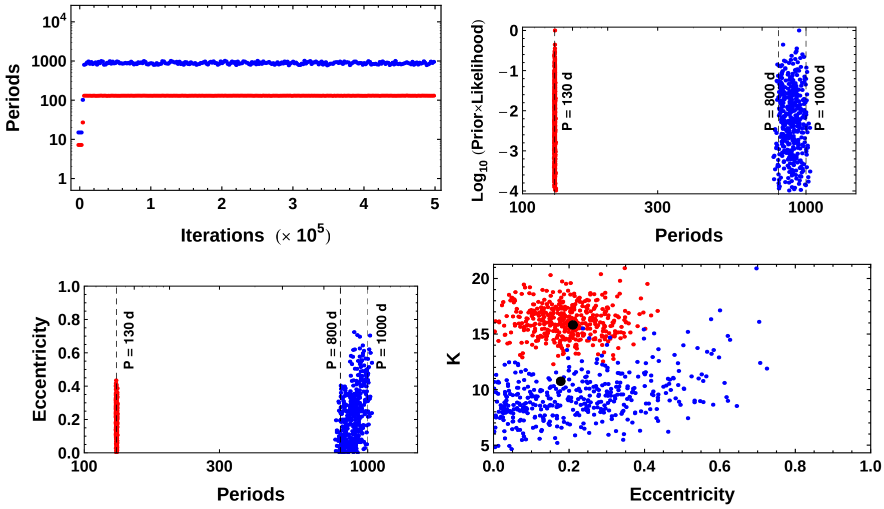


Figure 1.17 A variety of two planet periodogram plots for HD 208487.

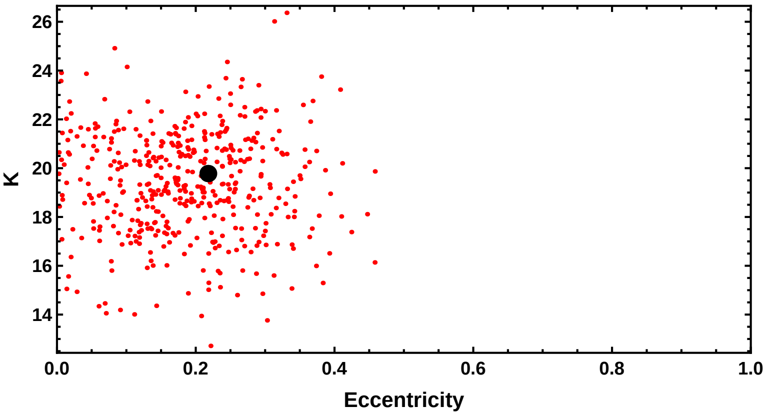


Figure 1.18 A *K* versus eccentricty plot for a one planet fit for HD 208487.

preference for low values. The marginal for the extra noise parameter peaks near zero.

Independent analysis of this data by Wright et al (2007) [68] found two possible solutions for the second period at 27 and 1000 d. Restricting the period range for the second period in our analysis, to a range that excluded the 900 d period, confirmed a feature near to their shorter period value. We subsequently investigated the effect

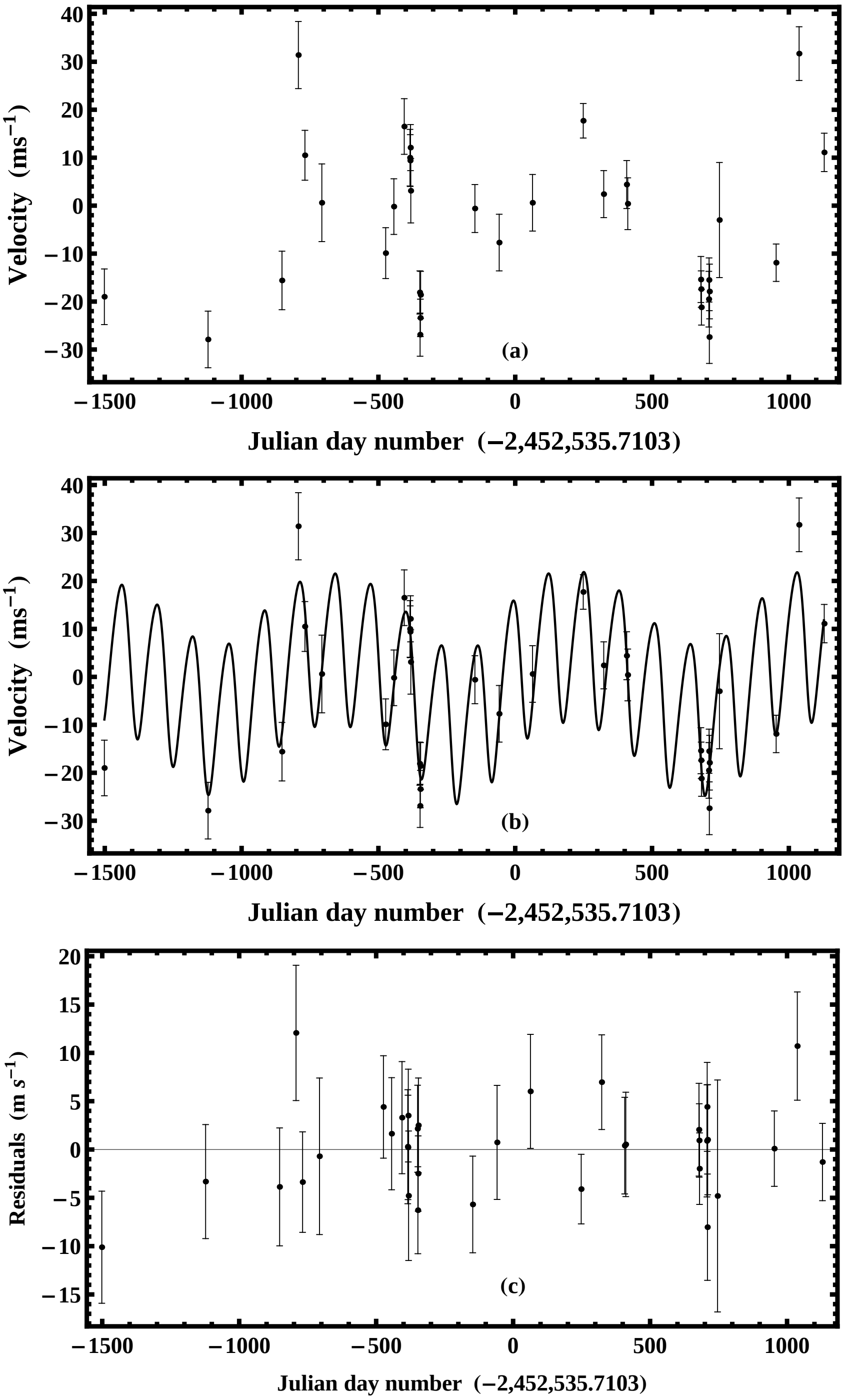


Figure 1.19 Panel (a) shows the radial velocity data [61]. Panel (b) and (c) show the two planet fit to the data and the fit residuals, respectively.

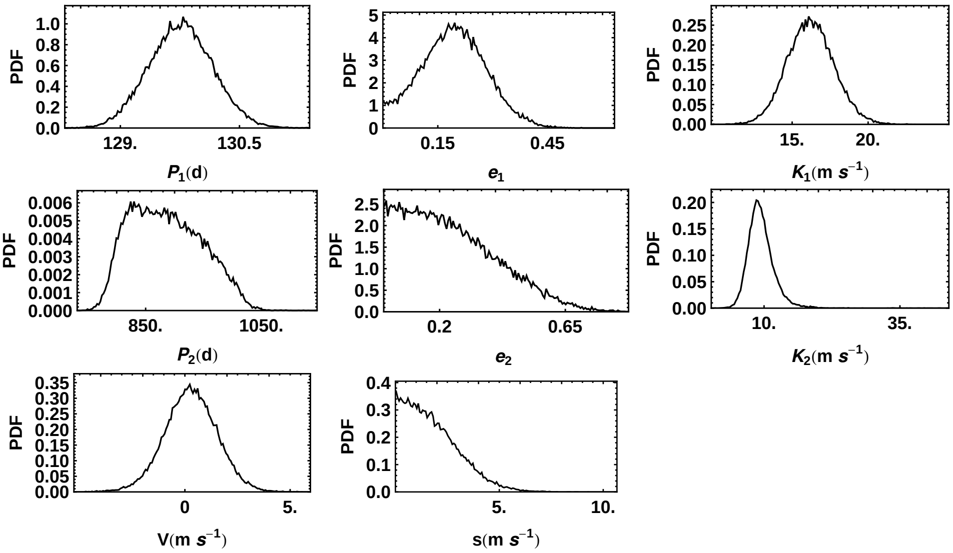


Figure 1.20 A plot of a subset of the FMCMC parameter marginal distributions for the two planet fit of the HD 208487 data.

of redoing our analysis with a frequency prior $p(f|M,I) \propto 1/\sqrt{f}$ to give more weight to shorter period signals and this resulted in the parallel tempering jumping between 28 and 900 d periods for the secondary period. The periodogram plots for the frequency prior $p(f|M,I) \propto 1/\sqrt{f}$ are shown in Figure 1.21. Because there are now three periods present, the dominant signal ($P$ = 130 d) does not have a unique color as it pairs first with the longer period signal ($p$ = 900 d) and then with the shorter period ($p$ = 28 d). We now find it useful to employ both choices of priors during the search for interesting cantidate planetary systems. It is only in the model comparison phase that we strictly employ a scale invariant frequency prior to allocate the relative probabilities of two different 2 planet models with different combinations of periods.

In the right panel of Figure 1.21, the red stars are samples of the 28d period and the black stars are the corresponding samples of the dominant 130 d period. Similarly, the blue boxes are samples of the 900 d period and the small boxes are the corresponding samples for the 130 d period. The MAP values are shown by the filled black circles. Both the 28 and 900 d samples have their highest concentration at low values of eccentricity where the average $K$ values for the 28 and 900 d periods are $\sim$ 8 m s$^{-1}$ and $\sim$ 9 m s$^{-1}$, respectively.

The extra noise parameter for the two planet fit peaks at zero which indicates that there is no additional signal to be accounted for. For consistency purposes, a three planet model was run using a frequency prior $\propto 1/\sqrt{f}$. Both the 130 d and 900 d

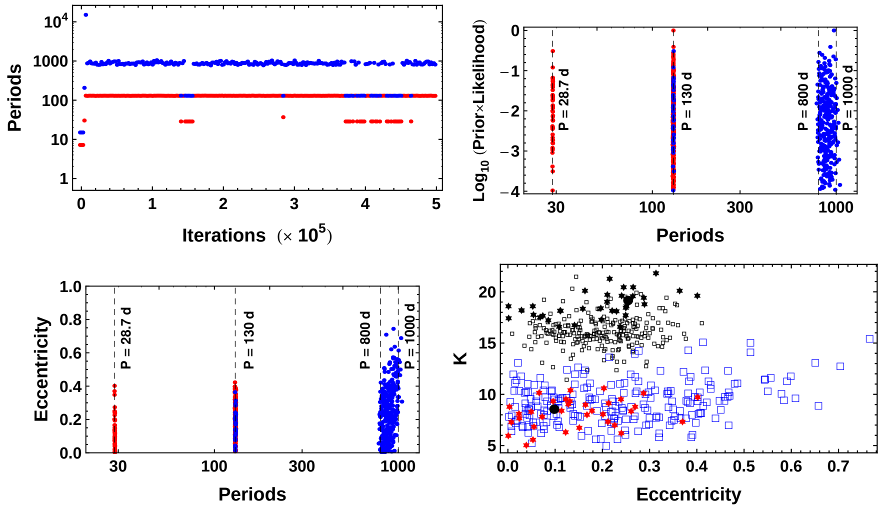


Figure 1.21 Two planet periodogram plots for HD 208487 using a frequency prior $\propto 1/\sqrt{f}$.

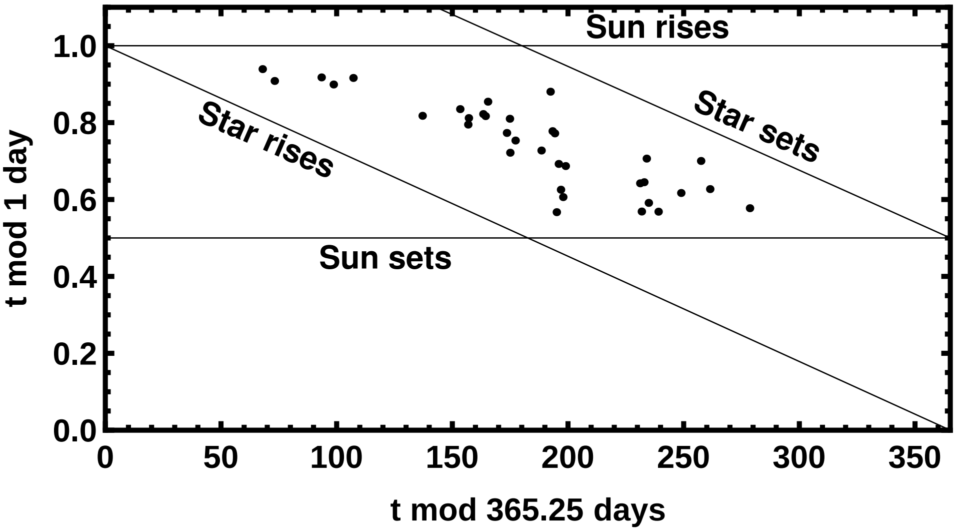


Figure 1.22 Times of HD 208487 radial velocity observations folded modulo the time of day and time of year. We used $t = JD - 2,451,224.19$ for convenience.

signals were clearly detected. A wide range of third period options were observed but did not include a clear detection of the 28 d signal.

### 1.5.3 Aliases

Dawson and Fabrycky (2010) [13] drew attention to the importance of aliases in the analysis of radial velocity data even for nonuniform sampling. Although the sampling is nonuniform when the star is observable, the restrictions imposed by the need to observe between sunset and sun rise, and when the star is above the horizon means that there are periodic intervals of time where no sample is possible. These periodicities give rise to aliases which we investigate in this section. Figure 1.22 shows the location of the HD 208487 radial velocity samples [61] modulo time of day and time of year using $t = JD - 2,451,224.19$ for convenience.

Deeming [14] [15] showed that a discrete Fourier transform ($F_N$) can be defined for arbitrary sampling of a deterministic time series (including a periodic function) and that $F_N$ is equal to the convolution[12] of the true Fourier transform (FT) with a spectral window function

$$F_N(f) = F(f) * w(f) \equiv \int_{-\infty}^{\infty} F(f - f')w(f')df', \tag{1.17}$$

where $F(f)$, true FT of the continuous time series is given by

$$F(f) = \int_{-\infty}^{\infty} f(t)e^{i2\pi ft}. \tag{1.18}$$

If $f(t)$ is a pure cosine function of frequency $f_0$, then $F(f)$ will be a pair of Dirac delta functions $\delta(f \pm f_0)$. $w(f)$ is the spectral window function given by

$$w(f) = \sum_{k=1}^{N} e^{i2\pi f t_k}. \tag{1.19}$$

It is also evident that $w(-f) = w^*(f)$. In the limit of continuous sampling in the time interval $(-T/2, +T/2)$, $w(f)$ tends to the Dirac delta funtion $\delta(0)$ as $t \longrightarrow \infty$. In general, the sampling will not be continuous so $w(f)$ may be significantly different from zero at frequencies other than $f = 0$. For $N$ evenly sampled data, unless the only physically possible frequencies are less than the Nyquist frequency $= N/(2T)$, the frequencies cannot be unambiguously determined. An important point is that $w(f)$ can be calculated directly from the data spacing alone.

It is common practice to use a normalized spectral window function

$$W(f) = N^{-1} \sum_{k=1}^{N} e^{i2\pi f t_k}, \tag{1.20}$$

normalized so $W(0) = 1$. In this case, the convolution Equation (1.17) becomes

$$\frac{1}{N} F_N(f) = F(f) * W(f) \equiv \int_{-\infty}^{\infty} F(f - f')W(f')df'. \tag{1.21}$$

[12] Such a convolution is nicely illustrated in Figures 13.6 and B.2 of my book [24].

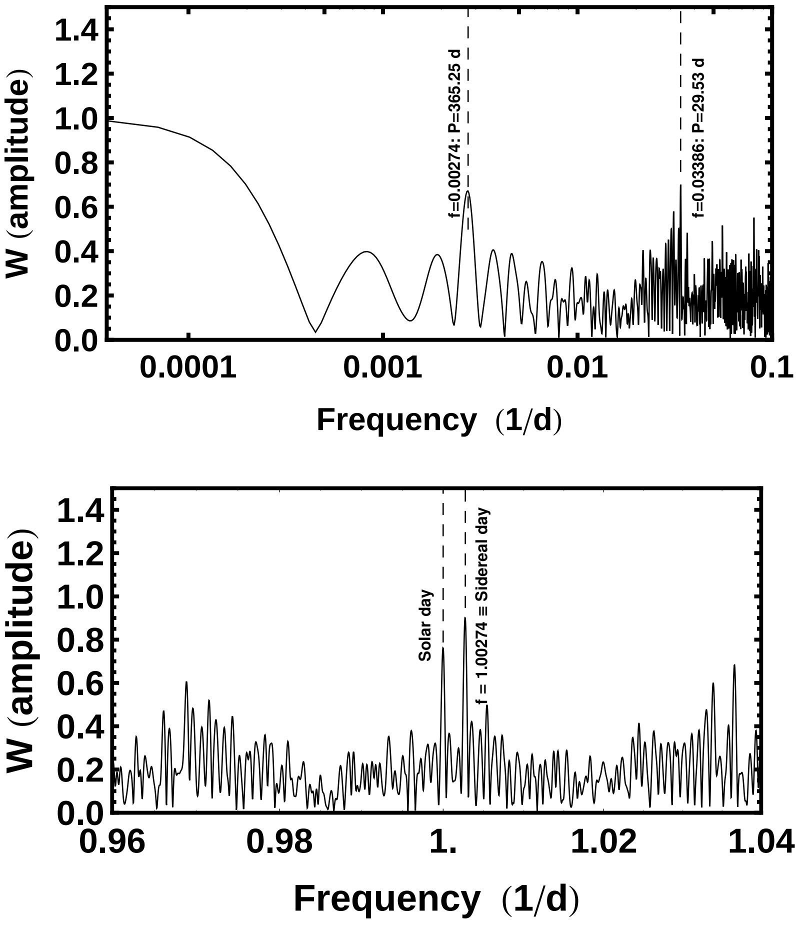


Figure 1.23 Amplitude of the spectral window function for the HD 208487 radial velocity measurements.

In general both $F(f)$ and $W(f)$ are complex so it is usual to examine the amplitude [13] of $W(f)$. Also, Dawson and Fabricky [13] show how the phase information is sometimes useful to distinquish between an alias and a physical frequency. Figure 1.23 shows a plot of the amplitude of the spectral window function, $W(f)$. Clearly with such sparce sampling there are many peaks in $W(f)$.

[13] Deeming [14] shows that for a deterministic signal (as opposed to a stochastic signal)

$$\begin{aligned}\frac{1}{N^2}|F_N(f)|^2 = \frac{1}{N^2}F_N(f)F_N^*(f) &= [F(f) * W(f)][F^*(f) * W^*(f)]\\ &\neq [F(f)F^*(f)] * [W(f)W^*(f)].\end{aligned} \tag{1.22}$$

It is thus not meaningful to plot $|W(f)|^2$.

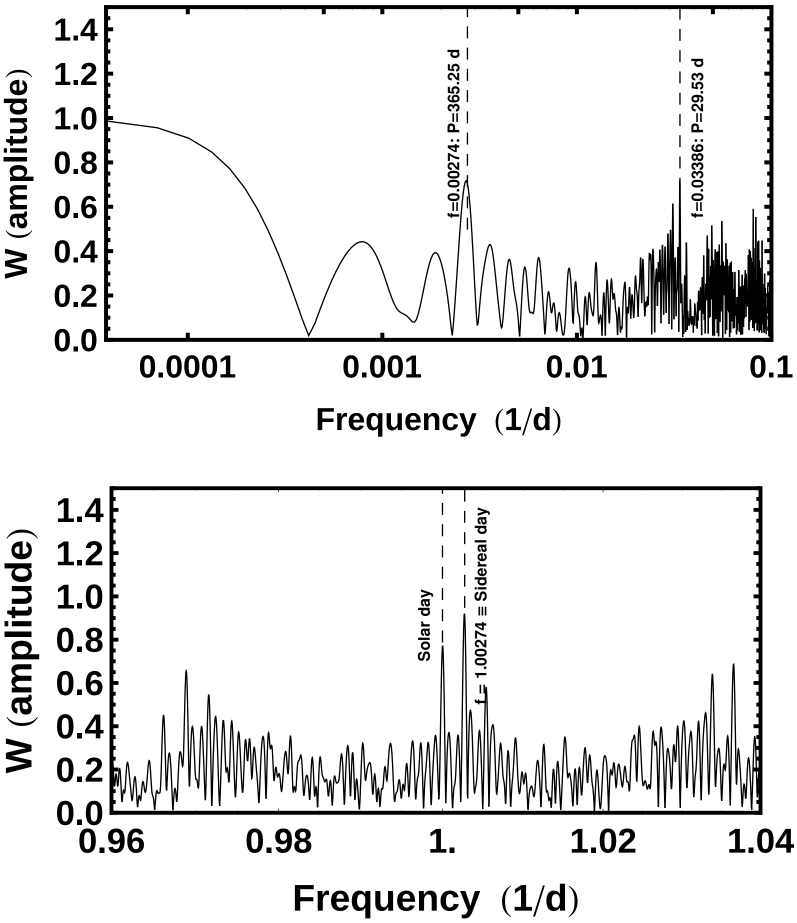


Figure 1.24 Amplitude of the weighted spectral window function for the HD 208487 radial velocity measurements.

The plots show well defined peak at $f = 0$ with a width of order $T^{-1}$ as well as strong peaks at 1 sidereal day, 1 solar day, the synodic month (29.53 d), and 32.13 = 1/(1/29.53 – 1/365.25) d.

Most exoplanet periodogram analysis makes use of the data weighted by $\mathrm{wt}_k = 1/\sigma_k^2$ so it is more appropriate to employ a spectral window function of the form

$$W(f) = wt_{\mathrm{sum}}^{-1} \sum_{k=1}^{N} wt_k \, e^{i2\pi f t_k}, \tag{1.23}$$

where $wt_{\mathrm{sum}} = \sum_{k=1}^{N} wt_k$. The dominant spectral features at 1 day, the synodic month and one year remain but the contrast is reduced in the weighted case.

Now that we understand the $W(f)$ for HD 208487, can we see how the 28 d period could be an alias of the 900 d period and vice versa? If the true spectrum contains a physical signal of frequency $f$, then when convolved with $W(f)$ we expect to see aliases in the observed spectrum at $f \pm f_w$. If $f_w > f$, we will still see a positive frequency alias at $|f - f_w|$ but to see this you need to recall that $W(f)$ has negative frequency peaks that satisfy $W(-f) = W^*(f)$. Also, the true spectrum of a periodic signal likewise has positive and negative frequencies when we use the common exponential notation as employed in the Fourier transform defined by Equation 1.17.

Suppose the 28 d period is the true signal. The 68% credible region boundaries of the 28 d peak extend from 28.58 to 28.70 d. One of the aliases produced by the synodic month feature (29.53 d) in $W(f)$ would be expected to give rise to an alias somewhere in the range $1/(1/28.58 - 1/29.53) = 890$ d to $1/(1/28.70 - 1/29.53) = 1026$ d. This range nicely overlaps the 68% credible region of 804 to 940 d of the 900 d peak. Similarly, if the 900 d signal is the true signal, its synodic month alias should be found in the 68% credible range 28.48 to 28.63 d, which it does.

### *1.5.4 Which is the alias?*

In this section we attempt to answer the question of which of the two secondary Kepler solutions at 28 and 900 d is a real physical signal. Below we consider some criteria that have proven useful:

1. Dawson and Fabrycky [13] outline a method for helping to distinguish a physical signal from an alias which makes use of the Generalized Lomb-Scargle (GLS) periodogram [70]. The method involves comparing the periodogram of the true residuals of an $n$ signal fit to perodograms of noise free simulations of possible choices of the $n + 1$ signal and includes comparison of the GLS phase of each spectral peak as well. GLS improves on the Lomb-Scargle method by allowing for a floating offset and weights.

   We illustrate the general idea of the Dawson-Fabrcyky method in Figure 1.25 as it applies to our HD 208487 analysis for aliases arising from the strongest peaks in the window function. The top row shows three portions of GLS periodogram of the one planet MAP fit residuals with phase circle above several peaks of interest. In the first two columns, dashed lines show the locations of the 854 and 28 d candidate signals together with one year aliases and the the dotted lines show one synodic month aliases of the 28 d candidate signal (overlaps the 854 d signal) and 854 d signal. In the third column, the dashed lines show the aliases of the 1 solar day and one sidereal day widow function peaks with the 28 d candidate signal and the dotted lines are for the 854 d candidate signal.

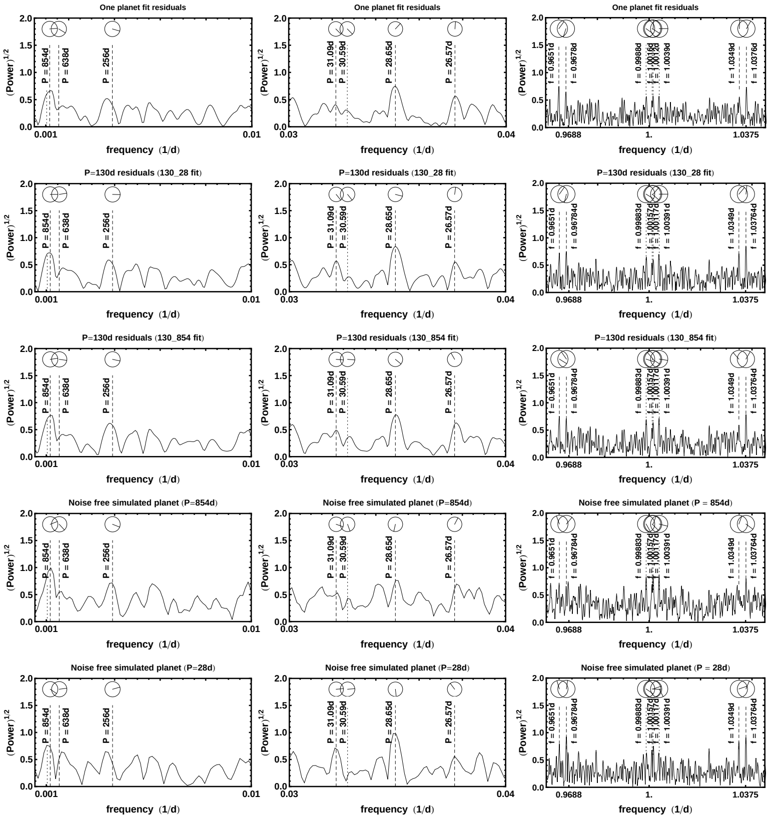


Figure 1.25 The Dawson-Fabrcyky method for distinguishing an alias from a physical signal. Top row shows 3 portions of GLS periodogram of the one planet MAP fit residuals with a phase circle above the peaks of interest. In the first two columns, dashed lines show the locations of the 854 and 28 d candidate signals together with one year aliases and the the dotted lines show one synodic month aliases of the 28 d candidate signal (overlaps the 854 d signal) and 854 d signal. In the third column, the dashed lines show the aliases of the 1 solar day and one sidereal day widow function peaks with the 28 d candidate signal and the dotted lines are for the 854 d candidate signal. The second row shows 3 portions of GLS periodogram of the $P = 130$ d planet residuals obtained from a two planet ($P = 130$ d and $P = 28$ d) fit. The third row is a similar GLS periodogram of the $P = 130$ d planet residuals obtained from a two planet ($P = 130$ d and $P = 854$ d) fit. Row 4 is a GLS periodogram of a noise free simulation of the 854 d signal. Row 5 is a GLS periodogram of a noise free simulation of the 28 d signal.

In the left two columns, the strongest peaks are for our two candidates for the physical signal, ~ 854 and 28.65 d, respectively. If the 28 d peak corresponds to the physical signal, then when convolved with the spectral window function[14] shown in Figures 1.23 and 1.24 additional aliases would be expected at $26.57 = 1/(1/28.65 + 1/365.25)$ and $31.09 = 1/(1/28.65 - 1/365.25)$. The alias at 26.57 d is clearly present but not the expected peak at 31.09. If the 854 d peak corresponds to the physical signal, then additional aliases would be expected near $256 = 1/(1/365.25 + 1/854)$, $638 = 1/(1/365.25 - 1/854)$ and $30.59 = 1/(1/29.53 - 1/854)$. The peak near to 256 d is at $267 = 1/(1/365.25 + 1/1000)$ or just within the period uncertainty. There is no clear evidence for a peak near 638 d. In the third column the dashed lines show the four aliases of the 28 d candidate signal, two for the 1 solar day and two for the 1 sidereal day peaks in the window function. Smaller peaks are just discernible for aliases of the 854 d candidate signal.

As shown in Figures 1.17 and 1.18, the typical $K$ value of the dominant 130 d signal in the $K$ versus eccentricity plot for the one planet fit is considerably higher than for the two planet fits regardless of whether the second real physical signal is 28 or 854 days. This suggested it might be useful to constructed two other typical 130 d planet residuals as show in rows 2 and 3. The second row shows the same two portions of GLS periodogram of the $P = 130$ d planet residuals obtained from a two planet ($P = 130$ d and $P = 28$ d) fit. The third row is for the $P = 130$ d planet residuals obtained from a two planet ($P = 28$ d and $P = 854$ d) fit. In the second row, a feature corresponding to the expected alias at 31.09 d has increased in height lending more support to 28 d as the physical signal. In the third row, the feature nearest the expected 256 d alias of a possible 854 d physical signal is stronger, peaking at $\sim 1/(1/365.25 + 1/940)$. However, there is still no clear feature near 638 d.

Row 4 is a GLS periodogram of a noise free simulation of the 854 d signal and row 5 is a GLS periodogram of a noise free simulation of the 28 d signal. An examination of all the rows in Figure 1.25 indicates that the 28 day aliases are slightly stronger. This analysis does not lead to a definite conclusion as to which is the true signal but slightly favors the 28 d candidate.

2. Both exhibit a preference for low eccentricities and the $K$ value of the 900 d signal is slightly stronger. On this grounds it is more likely to be the real signal. However, noise may add coherently to the alias causing the alias to be stronger.
3. In Section 1.6 we show how to do Bayesian model comparison using the Bayes factor. The Bayes factor favors a two planet Kepler model with periods of 130 and 900 d by a factor of 9.0 over the 28, 130 d combination. We also show that the Bayesian false alarm probability for a two planet model (regardless of the true second period) is $4.4 \times 10^{-3}$, but that the false alarm probability for the best

[14] The dominant peaks in the spectral window function are at 1 sidereal day, one synodic month (29.53 d) and one year.

candidate two planet model (130 & 900 d) is too high at 0.10 to conclude that the 900 d signal is the correct second signal.

4. Can we form a long term stable two planet system together with the 130 d Kepler orbit? Clearly if only one choice is viable this argues it is the real signal. An approximate Lagrange stability [15] analysis was carried out for both HD 208487 solutions, following the work of Tuomi [62] which in turn is based on the work of Barnes and Greenberg [4]. This analysis indicates that both choices appear to be long term stable but full-scale numerical integrations are needed to confirm this.

### 1.5.5 Gliese 581 example

In Section 1.6 we will intercompare three different Baysian methods for model comparison making use of our FMCMC model fits to precision radial velocity data for a range of models from 1 to 5 planets. In anticipation of this we re-analyzed the latest HARPS [22] data for Gliese 581 (Gl 581) for models spaning the range 3 to 6 planets using our latest version of FMCMC together with the priors discussed in Section 1.5.1. Gl 581 is an M dwarf with a mass of 0.31 times the mass of the sun at a distance of 20 light years which received a lot of attention because of the possibility of two super-earths in the habitable zone where liquid water could exist [64]. Our earlier Bayesian analysis [31] of the HARPS [47] and HIRES data [64] did not support the detection of a second habitable zone planet known at the time as Gl 581g. Subsequent analysis of a larger sample of HARPS data [22] failed to detect more than 4 planets. Recent analysis of $H_\alpha$ stellar activity for Gliese 581 indicate a correlation between the RV and stellar activity which leads to the conclusion that the 67 d signal (Gl 581d) is not planetary in origin [53]. In the context of comparing marginal likelihood estimators, we will not be concerned about the origin of the signals but only on how many signals are significant on the basis of the RV data and Keplerian models.

Our current analysis clearly detects the earlier periods of 3.15, 3.56, 12.9 and 67 days and only hints at a fifth signal with a period of 192 d. Still it is an interesting model comparison challenge to quantify the probability of this 5 signal model. With this in mind we show a variety of periodogram results for the 4 and 5 Keplerian signal models.

Figure 1.26 shows a variety of 4 planet periodogram plots for the GL 581 data for a scale invariant orbital frequency prior $\propto f^{-1}$. The top right shows the $\mathrm{Log}_{10}$[Prior

[15] Work in the 1970s and 1980s showed that the motions of a system of a star with two planets (not involved in a low-order mean motion resonance) would be bounded in some situations. Two dominant definitions of stability emerged, Hill stability and Lagrange stability. In Hill stability the ordering of the two planets in terms of distance from the central star is conserved. In addition, for Lagrange stability the planets remain bound to the star and the semimajor axis and eccentricity remain bounded.

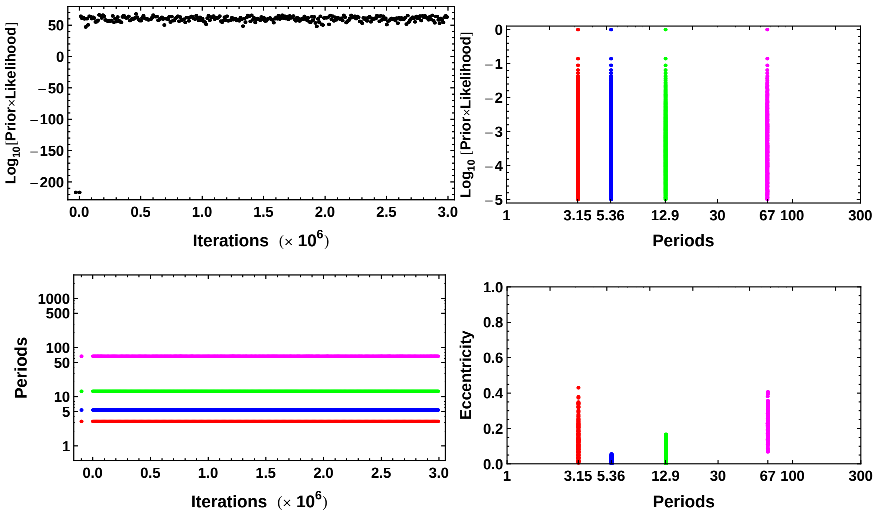


Figure 1.26 A variety of 4 planet periodogram plots for GL 581.

× Likelihood] versus FMCMC iteration for every 100th point. The bottom left shows the evolution of the 4 period parameters from their starting values marked by the four dots that occur before the zero on the iteration axis. It is clear that the FMCMC did not make transitions to any other peaks. The top right panel shows a sample of the 4 period parameter values versus a normalized value of Log$_{10}$[Prior × Likelihood]. The bottom right shows a plot of eccentricty verus period.

Figures 1.27 and 1.28 shows the 5 planet Kepler periodogram results using two different orbital frequency priors. The latter is scale invariant and the former employs a frequency prior $\propto 1/\sqrt{f}$, which helps with the detection of shorter period signals. The best set of parameters from the 4 planet fit were used as start parameters. The starting period for the fifth period was set = 30 d and the dominant fifth period found in both trials was ∼ 192 d, on the basis of the number of samples. As illustrated in these example, the parallel tempering feature identifies not only the strongest peak but other potential interesting ones as well. In Fig. 1.28 the MAP value of the fifth period is 192 d which is > 10 times larger than the next strongest with a period of 45 d. Two other peaks at 72 and 90 are consistent with one year aliases of each other. For the scale invariant trial shown in Fig. 1.28, 89% of the samples include the 192 d peak. The previous 4 periods in the 4 planet fit are clearly present in both trials.

Looking at the eccentricity distributions of fifth period candidate signals, it is clear that only the 192 d peak favors low eccentricities. Fig. 1.29 shows a plot of a subset of the FMCMC parameter marginal distributions for the 5 signal fit of the

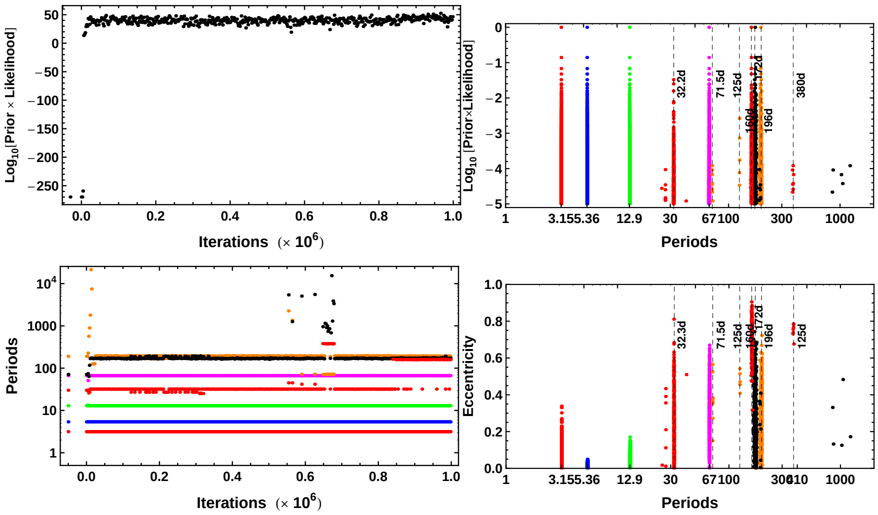


Figure 1.27 A variety of 5 planet periodogram plots for GL 581 for an orbital frequency prior $\propto 1/\sqrt{f}$.

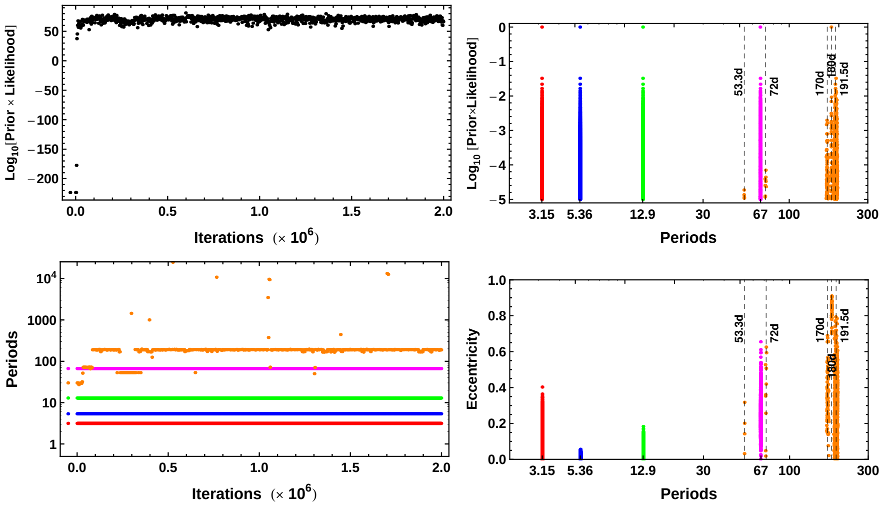


Figure 1.28 A variety of 5 planet periodogram plots for GL 581 for a scale invariant orbital frequency prior $\propto f^{-1}$.

HARPS data after filtering out the post burn-in FMCMC iterations that correspond to the 5 dominant period peaks at 3.15, 5.37, 12.9, 66.9, and 192 d. Still, on the basis of this data, the author is not inclined to claim this as a likely candidate planet. The main point of this exercise is taken up in the next section where we

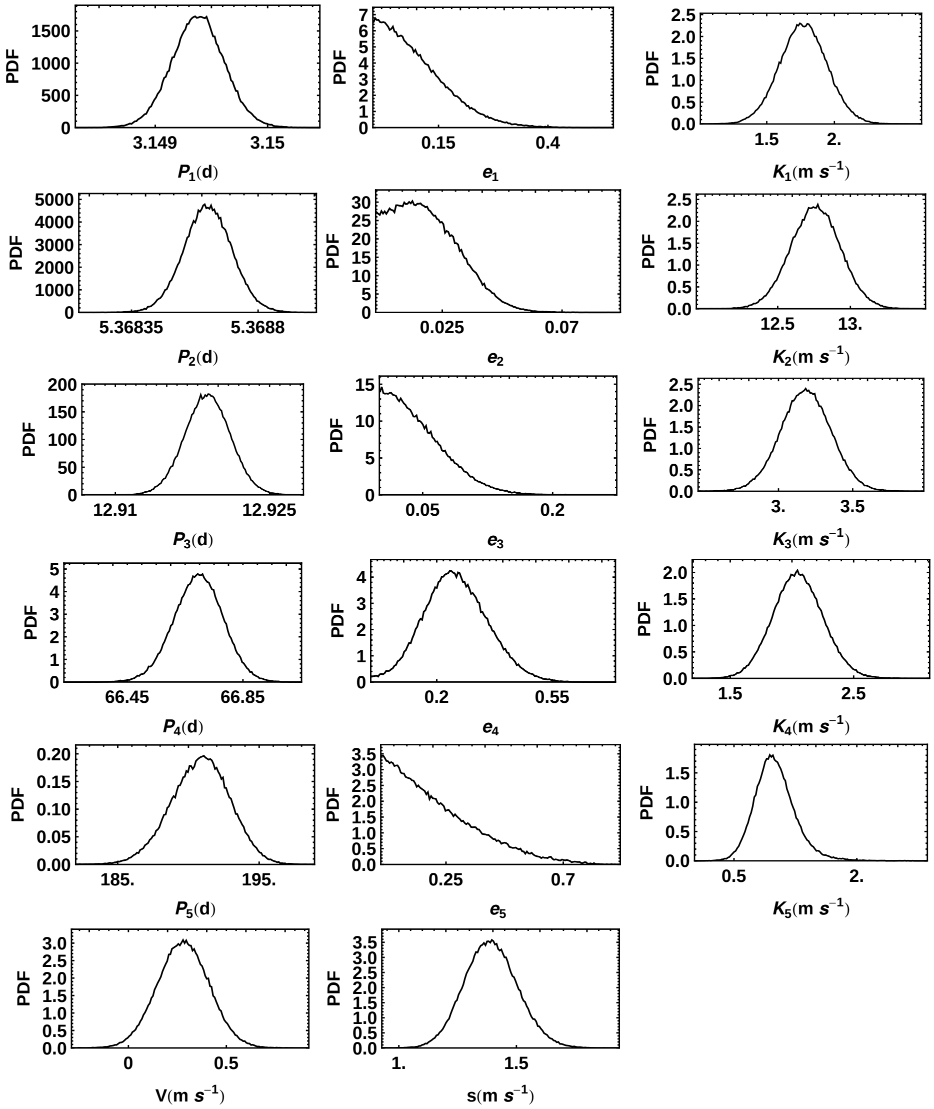


Figure 1.29 FMCMC parameter marginal distributions for the 5 planet fit of the HARPS data after filtering out the post burn-in FMCMC iterations that correspond to the 5 dominant period peaks at 3.15, 5.37, 12.9, 66.9, and 192 d

will see what probability theory has to say about the relative probability of this particular 5 signal model to the 4 signal model, for our choice of priors.

It is sometimes useful to explore the option for additional Keplerian-like signals beyond the point at which the false alarm probability starts to increase. This is because the presence of other signals not accounted for in the model can give rise to

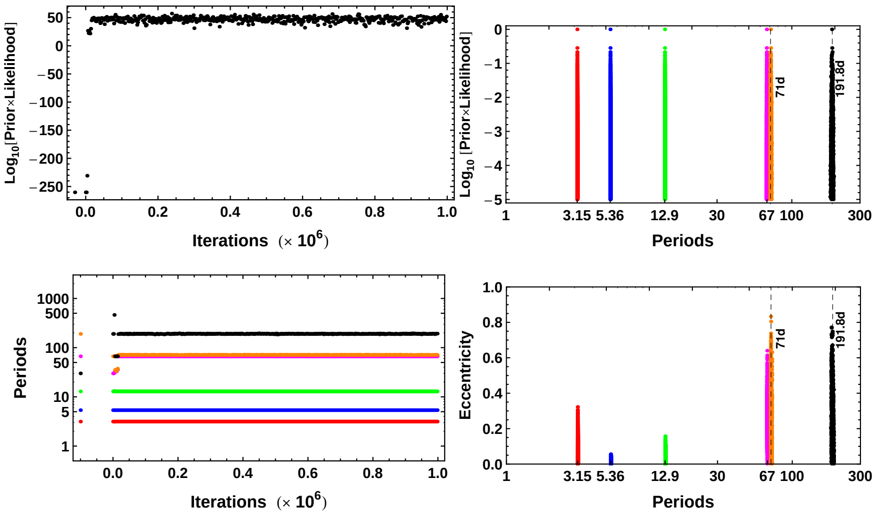


Figure 1.30 A variety of 6 planet periodogram plots for GL 581. Orbital frequency prior $\propto \nu^{-1/2}$

an effective correlated noise that once removed can sometimes lead to significantly improved detections. Figure 1.30 shows the results for a 6 signal fit. As will be shown in Section 1.6.5, the false alarm probability of the 6 signal case is lower than for the 5 signal case, but is still too high to be considered significant. Both 7 signal and 8 signal models were run but resulted in large false alarm probabilities and are not shown.

## 1.6 Model Comparison

One of the great strengths of Bayesian analysis is the built-in Occam's razor. More complicated models contain larger numbers of parameters and thus incur a larger Occam penalty, which is automatically incorporated in a Bayesian model comparison analysis in a quantitative fashion (see for example, Gregory [24], p. 45). The analysis yields the relative probability of each of the models explored.

To compare the posterior probability of the i$^{\text{th}}$ planet model[16] to the one planet model we need to evaluate the odds ratio, $O_{i1} = p(M_i|D,I)/p(M_1|D,I)$, the ratio of the posterior probability of model $M_i$ to model $M_1$. Application of Bayes' theorem leads to,

$$O_{i1} = \frac{p(M_i|I)}{p(M_1|I)} \frac{p(D|M_i,I)}{p(D|M_1,I)} \equiv \frac{p(M_i|I)}{p(M_1|I)} B_{i1} \tag{1.24}$$

[16] More accurately these are models that assume different numbers of Kepler-like signals. As previously mentioned stellar activity can also generate Kepler-like signals which need to be ruled out before the signal is ascribed to a planetary candidate.

where the first factor is the prior odds ratio, and the second factor is called the *Bayes factor*, $B_{i1}$. The Bayes factor is the ratio of the marginal (global) likelihoods of the models. The marginal likelihood for model $M_i$ is given by

$$p(D|M_i, I) = \int dX p(X|M_i, I) \times p(D|, M_i, I). \tag{1.25}$$

Thus Bayesian model comparison relies on the ratio of marginal likelihoods, not maximum likelihoods. The marginal likelihood is the weighted average of the conditional likelihood, weighted by the prior probability distribution of the model parameters and $s$. This procedure is referred to as marginalization.

The marginal likelihood can be expressed as the product of the maximum likelihood and the Occam penalty (e.g., see Gregory [24], page 48). The Bayes factor will favor the more complicated model only if the maximum likelihood ratio is large enough to overcome this penalty. In the simple case of a single parameter with a uniform prior of width $\Delta X$, and a centrally peaked likelihood function with characteristic width $\delta X$, the Occam factor is $\approx \delta X/\Delta X$. If the data is useful then generally $\delta X \ll \Delta X$. For a model with $m$ parameters, each parameter will contribute a term to the overall Occam penalty. The Occam penalty depends not only on the number of parameters but also on the prior range of each parameter (prior to the current data set, $D$), as symbolized in this simplified discussion by $\Delta X$. If two models have some parameters in common then the prior ranges for these parameters will cancel in the calculation of the Bayes factor. To make good use of Bayesian model comparison, we need to fully specify priors that are independent of the current data $D$. The sensitivity of the marginal likelihood to the prior range depends on the shape of the prior and is much greater for a uniform prior than a scale invariant prior (e.g., see Gregory [24], page 61). In most instances we are not particularly interested in the Occam factor itself, but only in the relative probabilities of the competing models as expressed by the Bayes factors. Because the Occam factor arises automatically in the marginalization procedure, its effect will be present in any model comparison calculation. Note: no Occam factors arise in parameter estimation problems. Parameter estimation can be viewed as model comparison where the competing models have the same complexity so the Occam penalties are identical and cancel out.

The MCMC algorithm produces samples which are in proportion to the posterior probability distribution which is fine for parameter estimation but one needs the proportionality constant for estimating the model marginal likelihood. Clyde et al. [10] reviewed the state of techniques for model comparison from a statistical perspective and Ford and Gregory [21] have evaluated the performance of a variety of marginal likelihood estimators in the exoplanet context.

Nested Sampling, developed by Skilling [59], is another powerful way for calculating model marginal likelihoods (referred to as the evidence by Skilling). Nested sampling reverses the historical approach. The marginal likelihood (evidence) is now the prime target, with representative posterior samples available as an optional by-product. An invariance (over monotonic relabelling) allows Nested Sampling to deal with a class of phase-change problems which Skilling argues are a problem for thermal annealing methods like thermodynamic integration (see Section 1.6.1). A variant of Nested Sampling, called MultiNest [17], has been employed in the analysis of RV data.

Other techniques for computing marginal likelihoods that have recently been proposed include: Nested Restricted Monte Carlo (NRMC) [29], Annealing Adaptive Importance Sampling (AAIS) [43], and Reversible Jump Monte Carlo using a kD-tree [16].

For one planet models with 7 parameters, a wide range of techniques perform satisfactorily. The challenge is to find techniques that handle high dimensions. A six planet model has 32 parameters and one needs to develop and test methods of handling at least 8 planets with 42 parameters. At present there is no widely accepted method to deal with this challenge.

In this work we will compare the results from three marginal likelihood estimators: (a) Parallel Tempering, (b) the Ratio Estimator, and (c) Nested Restricted Monte Carlo. A brief outline of each method is presented in Sections 1.6.1, 1.6.2, and 1.6.3. A comparison of the three methods is given in Section 1.6.4.

### *1.6.1 Parallel tempering estimator*

The MCMC samples from all ($n_\beta$) simulations can be used to calculate the marginal likelihood of a model according to Equation (1.26) [24]. This method of estimating the marginal likelihood is commonly referred to as thermodynamic integration.

$$\ln[p(D|M_i, I)] = \int d\beta \langle \ln[p(D|M_i, X, I)] \rangle_\beta, \tag{1.26}$$

where $i = 0, 1, \cdots, m$ corresponds to the number of planets, and $X$ represent the set of the model parameters which includes the extra Gaussian noise parameter $s$. In words, for each of the $n_\beta$ parallel simulations, compute the expectation value (average) of the natural logarithm of the likelihood for post burn-in MCMC samples. It is necessary to use a sufficient number of tempering levels that we can estimate the above integral by interpolating values of

$$\langle \ln[p(D|M_i, X, I)] \rangle_\beta = \frac{1}{n} \sum_t \ln[p(D|M_i, X_{t,\beta}, I)], \tag{1.27}$$

in the interval from $\beta = 0$ to 1, from the finite set and where $n$ is the number of post burn-in samples in each set. For this problem we used 44 tempering levels in the

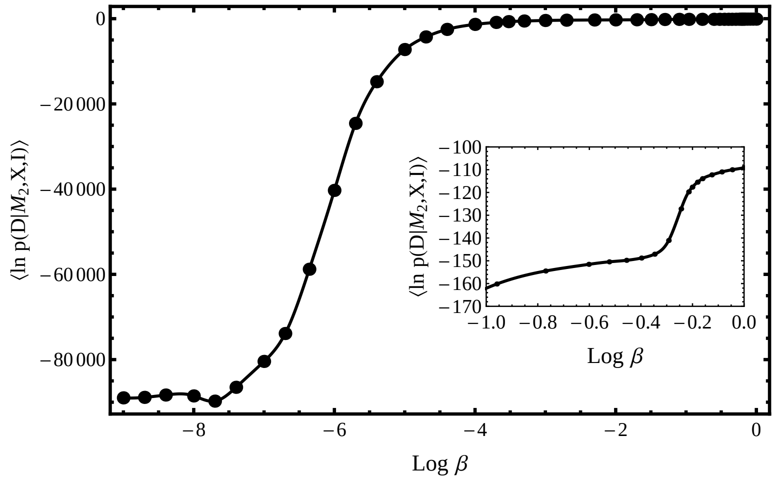


Figure 1.31 A plot of $\langle \ln[p(D|M_2, X, I)] \rangle_\beta$ versus $\beta$ for the two planet HD 208487 model results. The inset shows a blow-up of the range $\beta = 0.01$ to 1.0.

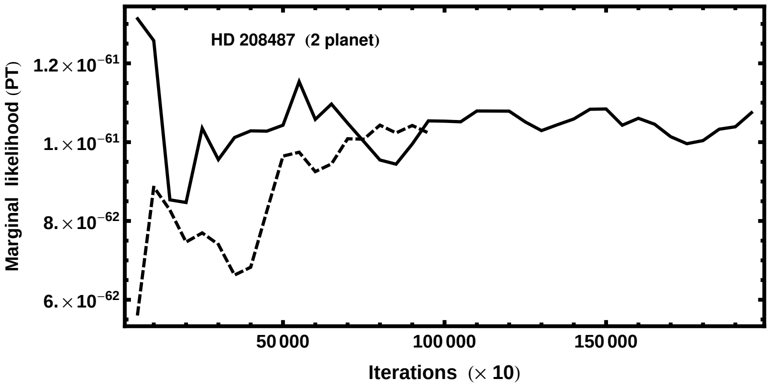


Figure 1.32 A plot of the marginal likelihood, $p(D|M_2, X, I)_{PT}$, versus FMCMC iteration for the two planet HD 208487 model results for two trials.

range $\beta = 10^{-9}$ to 1.0. Figure 1.31 shows a plot of $\langle \ln[p(D|M_2, X, I)] \rangle_\beta$ versus $\beta$ for a two planet model fit to the HD 208487 radial velocity data of Tinney (2005) [61]. The inset shows a blow-up of the range $\beta = 0.1$ to 1.0.

The relative importance of different ranges of $\beta$ can be judged from Table 1.2. The first column gives the range of $\beta$ included. The second column gives the estimated marginal likelihood, $p(D|M_2, I)$ for that range. The third column gives the fractional error relative to the $p(D|M_2, I)$ value derived from the largest $\beta$ range extending from $10^{-9}$ to 1. Thus, if we neglected the contribution from the $\beta$ range extending from $10^{-6}$ to $10^{-9}$ this would result in a fractional error of 0.06. The

Table 1.2 *Parallel tempering marginal likelihood estimate, $p(D|M_2, I)_{PT}$, and fractional error versus $\beta$ range for the two planet HD 208487 model results.*

| $\beta$ range | $p(D\|M_2, I)_{PT}$ | Fractional error |
|---|---|---|
| $10^{-1}$ – 1.0 | $3.290 \times 10^{-52}$ | $3 \times 10^{9}$ |
| $10^{-2}$ – 1.0 | $4.779 \times 10^{-60}$ | 43 |
| $10^{-3}$ – 1.0 | $2.817 \times 10^{-61}$ | 2.6 |
| $10^{-4}$ – 1.0 | $1.635 \times 10^{-61}$ | 0.51 |
| $10^{-5}$ – 1.0 | $1.306 \times 10^{-61}$ | 0.21 |
| $10^{-6}$ – 1.0 | $1.148 \times 10^{-61}$ | 0.06 |
| $10^{-7}$ – 1.0 | $1.091 \times 10^{-61}$ | 0.008 |
| $10^{-8}$ – 1.0 | $1.083 \times 10^{-61}$ | 0.0008 |
| $10^{-9}$ – 1.0 | $1.082 \times 10^{-61}$ | 0.0 |

fractional error falls rapidly with each decade. If we wanted an answer good to $\sim$ 20% then a $\beta$ range from $10^{-5}$ to 1 would suffice. An earlier one planet model results for HD 188133 yielded a fractional error of 0.26 for $\beta = 10^{-5}$ to 1. Later we will see from a comparison of three different Bayesian marginal likelihood methods that differences of order of a factor of two are not uncommon so a $\beta$ range of $10^{-5}$ – 1.0 will generally be sufficient.

Figure 1.32 show the dependence of the parallel tempering (PT) marginal likelihood estimate versus FMCMC iteration number for the two planet HD 208487 model results for two trials. Only every tenth iteration is saved so the true number of iterations is a factor of 10 larger.

Figure 1.33 compares marginal likelihood estimates for 1 to 5 planet radial velocity fits. The one and two planet fits are to the HD 208487 data, the 3, 4 and 5 planet fits are for Gliese 581 data. The left hand column of plots show PT marginal likelihood estimates versus iteration. For the one and two planet cases, where repeats were carried out, the the agreement was good, to within 10%. For the 4 planet case, it is clear that the parallel tempering derived marginal likelihood results did not reach an equilibrium value in $2.5 \times 10^{6}$ iterations. A PT derived marginal likelihood for a 5 planet model was not attempted. The right hand column of plots are for the marginal likelihoods derived from the ratio estimator (RE) and nested restricted Monte Carlo (NRMC) methods which are discussed in the next two sections.

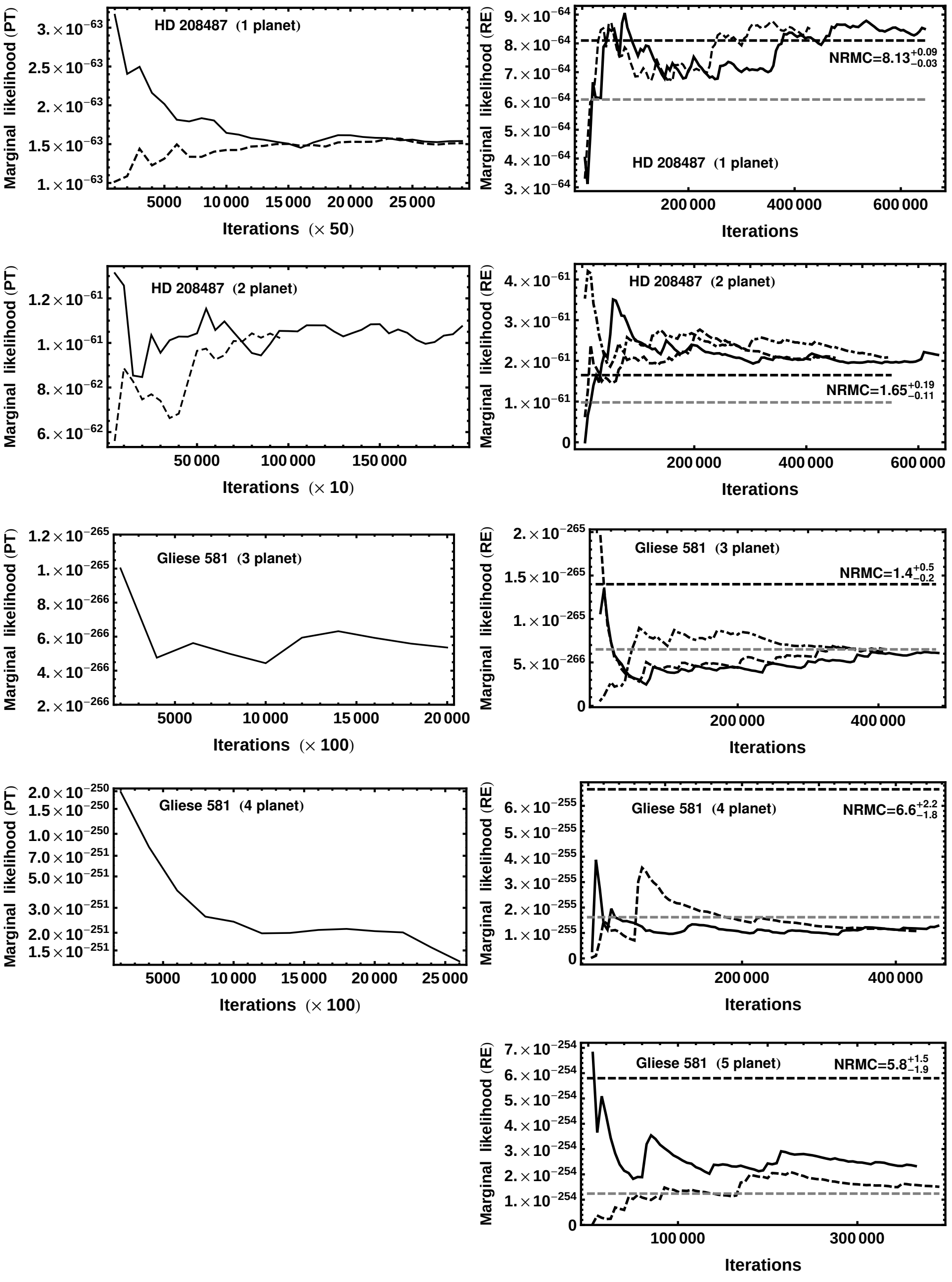


Figure 1.33 Comparisons of three different marginal likelihood estimators versus iteration for one to five planet radial velocity model fits. The left hand column of plots show parallel tempering marginal likelihood values versus FMCMC iteration number. The curves in the right hand column of panels show ratio estimator marginal likelihood values versus iteration. The horizontal black dashed lines are the marginal likelihoods from the NRMC method together with the numerical value of the mean and range of 5 repeats. The horizontal gray dashed lines are the NRMC marginal likelihood value within the 95% credible region of the model parameters.

### 1.6.2 Marginal likelihood ratio estimator

Our second method [17] was introduced by Ford and Gregory (2007) [21]. It makes use of an additional sampling distribution $h(X)$. Our starting point is Bayes' theorem

$$p(X|D, M_i, I) = \frac{p(X|M_i, I)p(D|M_i, X, I)}{p(D|M_i, I)}. \tag{1.28}$$

Re-arranging the terms and multiplying both sides by $h(X)$ we obtain

$$\begin{aligned} p(D|M_i, I)p(X|D, M_i, I)h(X) = \\ p(X|M_i, I)p(D|M_I, X, I)h(X). \end{aligned} \tag{1.29}$$

Integrate both sides over the prior range for $X$.

$$\begin{aligned} p(D|M_i, I)_{RE} \int p(X|D, M_i, I)h(X)dX = \\ \int p(X|M_i, I)p(D|M_I, X, I)h(X)dX. \end{aligned} \tag{1.30}$$

The ratio estimator of the marginal likelihood, which we designate by $p(D|M_i, I)_{RE}$, is given by

$$p(D|M_i, I)_{RE} = \frac{\int p(X|M_i, I)p(D|M_i, X, I)h(X)dX}{\int p(X|D, M_i, I)h(X)dX}. \tag{1.31}$$

To obtain the marginal likelihood ratio estimator, $p(D|M_i, I)_{RE}$, we approximate the numerator by drawing samples $\tilde{X}^1, \tilde{X}^2, \cdots, \tilde{X}^{n'_s}$ from $h(X)$ and approximate the denominator by drawing samples $X^1, X^2, \cdots, X^{n_s}$ from the $\beta = 1$ MCMC post burn-in iterations.

$$p(D|M_i, I)_{RE} = \frac{\frac{1}{n'_s} \sum_{i=1}^{n'_s} p(\tilde{X}^i|M_i, I)p(D|M_i, \tilde{X}^i, I)}{\frac{1}{n_s} \sum_{i=1}^{n_s} h(X^i)}. \tag{1.32}$$

The arbitrary function $h(X)$ was set equal to a multivariate normal distribution (multinormal) with a covariance matrix equal to twice the covariance matrix computed from a sample of the $\beta = 1$ MCMC output. We used [18] $n'_s = 10^5$ and $n_s$ from $10^4$ to $2 \times 10^5$. Some of the samples from a multinormal $h(X)$ can have non physical parameter values (e.g. $K < 0$). Rejecting all non physical samples corresponds to sampling from a truncated multinormal. The factor required to normalize the truncated multinormal is just the ratio of the total number of samples from the full multinormal to the number of physical valid samples. Of course we need to

[17] Initially proposed by J. Berger, at an Exoplanet Workshop sponsored by the Statistical and Applied Mathematical Sciences Institute in Jan. 2006

[18] According to [21], the numerator converges more rapidly than the denominator.

use the same truncated multinormal in the denominator of Equation (1.31) so the normalization factor cancels.

### *Mixture model*

It is clear that a single multinormal distribution cannot be expected to do a very good job of representing the correlation between the parameters that is frequently evident. Following [21], we improved over the single multinormal by using a mixture of multivariate normals by setting

$$h(X) = \frac{1}{n_c} \sum_{j=1}^{n_c} h_j(X). \tag{1.33}$$

We chose each mixture component to be a multivariate normal distribution, $h_j(X) = N(X|X_j, \Sigma_j)$, and determined a covariance matrix for each $h_j(X)$ using the posterior sample. As a first step, compute $\vec{\rho}$, defined to be a vector of the sample standard deviations for each of the components of $X$, using the posterior sample [19]. Next, define the distance between the posterior sample $X_i$ and the center of $h_j(X)$, $d_{ij}^2 = \sum_k \left(X_{ki} - X_{kj}\right)^2 / \rho_k^2$, where $k$ indicates the element of $X$ and $\vec{\rho}$. Now draw another random subset of $100n_c$ samples [20] from the original posterior sample (without replacement), select the 100 posterior samples closest to each mixture component and use them to calculate the covariance matrix, $\Sigma_j$, for each mixture component. To compute the covariance matrix for each component we adopted the following approach. A random pair of the 100 posterior samples was selected with the intention of constructing a difference vector. For the angular parameters $\psi$ and $\phi$, we compute both the straight difference ($d_1$) and the difference of these components ($d_2$) after adding $2\pi$ to each. The smaller of these two differences avoids the wrap around problem mentioned in the previous footnote. This process of selecting random pairs and computing their difference vector is repeated until we have 100 difference vectors. The covariance matrix for this mixture component is then computed from this set of difference vectors. Actually, it proves useful to employ this difference covariance matrix with components that are twice those of the true covariance matrix. Since the posterior sample is assumed to have fully explored the posterior, $h(X)$ should explore in all regions of significant probability, provided that we use enough mixture components.

In the case of the five planet Kepler model fit to GL 581, the FMCMC analysis leads to multiple choices of five signal configurations. For both the RE and NRMC

[19] The angular parameters need to be treated in a special way because the PDF can pile-up at both ends of the range with a big gap in the middle. The two ends of the PDF are actually close to one another in a wrap around sense because they are angular coordinates. Without allowing for this a simple variance calculation can lead to a misleadingly large value.

[20] This needs to be increased to $200n_c$ samples for a ≥ 5 planet model.

methods, it is possible to filter the FMCMC samples to select the individual signal configurations separately allowing for a calculation of their relative probability. For five planet fit to GL 581, the results reported here are only for the 3.15, 5.36, 12.9, 67, 192 d period configuration. For the PT method this would not be possible, only the global marginal likelihood of the model can be evaluated.

The right hand column of plots in Figure 1.33 show RE marginal likelihoods versus FMCMC iteration for one to five planet model fits using a mixture model with 150 centers. The RE curve was computed twice (solid and dashed) to demonstrate the level of repeatability. In the worst case (five planets) the agreement was within a factor of 2. Agreement with the other two marginal likelihood estimators was best when the RE method was used with FMCMC data which was thinned sufficiently (by a factor of 50 to 100) that the samples were essentially independent .

### *1.6.3 Nested Restricted Monte Carlo*

Straight Monte Carlo (MC) integration can be very inefficient because it involves random samples drawn from the prior distribution to sample the whole prior volume. The fraction of the prior volume of parameter space containing significant probability rapidly declines as the number of dimensions increase. For example, if the fractional volume with significant probability is 0.1 in one dimension then in 32 dimensions the fraction might be of order $10^{-32}$. In restricted MC integration (RMC) this problem is reduced because the volume of parameter space sampled is greatly restricted to a region delineated by the outer borders of the marginal distributions of the parameters for the particular model. However, in high dimensions most of the MC samples will fall near the outer boundaries of that volume and so the sampling could easily under sample interior regions of high probability leading to an underestimate of the marginal likelihood.

In NRMC integration [29, 33], multiple boundaries of a restricted hypercube in parameter space are constructed based on credible regions ranging from 30% to $\geq$ 99%, as needed. To construct the $X\%$ hypercube we compute the $X\%$ credible region of the marginal distribution for each parameter of the particular model. The hypercube is delineated by the the $X\%$ credible range of the marginal for each parameter. Note that the fraction of the total probability of the joint posterior distribution contained within the hypercube will be greater than $X\%$, in part because the marginal distributions of the parameters will be broadened by any parameter correlations.

The next step is to compute the contribution to the total NRMC integral from each nested interval and sum these contributions. For example, for the interval between the 30% and 60% hypercubes, we generate random parameter samples within the 60% hypercube and reject any sample that falls within the 30% hypercube. Using the remaining samples we can compute the contribution to the NRMC integral from that interval. For NRMC, if there is more than one peak in the joint probability of the parameters that emerge from the FMCMC analysis, then NRMC must be performed for each peak separately.

The left panel of Figures 1.34 through 1.38 shows the NRMC contributions to the marginal likelihood from the individual intervals for five repeats of 1 and 2 planet fits to the HD 208487 data and 3, 4, and 5 planet fits to the GL 581 data. The right panel shows the summation of the individual contributions versus the volume of the credible region. The credible region encoded as 9995% is defined as follows. Let $X_{U99}$ and $X_{L99}$ correspond to the upper and lower boundaries of the 99% credible region, respectively, for any of the parameters. Similarly, $X_{U95}$ and $X_{L95}$ are the upper and lower boundaries of the 95% credible region for the parameter. Then $X_{U9995} = X_{U99} + (X_{U99} - X_{U95})$ and $X_{L9995} = X_{L99} + (X_{L99} - X_{L95})$. Similarly[21], $X_{U9984} = X_{U99} + (X_{U99} - X_{U84})$. For the 3 planet fit the spread in results is within ±23% of the mean. For each credible region interval, 80,000 MC samples were used (4 repeats of 20,000 samples each). The *Mathematica* code parallelizes the computation. The mean value of the prior × likelihood within the 30% credible region is a factor of $2{\times}10^5$ larger than the mean in the shell between the 97 and 99% credible regions. However, the volume of parameter space in the shell between the 97 and 99% credible regions is a factor of $8{\times}10^{11}$ larger than the volume within the 30% credible region so the contribution from the latter to the marginal likelihood is negligible.

Fig. 1.39 shows the fraction of the total NRMC marginal likelihood within the 95% and 99% credible regions versus the number of planets. The contribution to the marginal likelihood from a region bounded by the 95% credible region decreases systematically from 74% for a one planet fit to 22% for a 5 planet fit. The same trend is evident at a lower level for the region bounded by the 99% region with the exception of the last point.

What about the repeatability of the NRMC results? The 5 repeats span ± $1, 9, 23, 30, 30\%$ for the 1, 2, 3, 4, and 5 planet fit, respectively. The biggest contribution to the spread in repeated NRMC marginal likelihood estimates comes

[21] **Import details:**
Test that the extended credible region outer boundary (like 9930) for each period parameter does not overlap the credible region of an adjacent period parameter in a multiple planet fit. Also, in the case of a probability distribution with multiple peaks it is advisable to define cutoffs in period parameter space about each peak to prevent this overlap from happening. Even in the case of a single peak it is useful to define period cutoffs as follows. Note the combination of upper and lower period parameter values that just contain all the MCMC samples. Then define cutoff period intervals that are approximately 1% larger. In high dimensions this translates to a significant increase in parameter space volume.

Determining the marginal PDF boundaries of angular parameters needs to be treated in a special way because the PDF can pile-up at both ends of the range with a big gap in the middle. The two ends of the PDF are actually close to one another in a wrap around sense because they are angular coordinates.

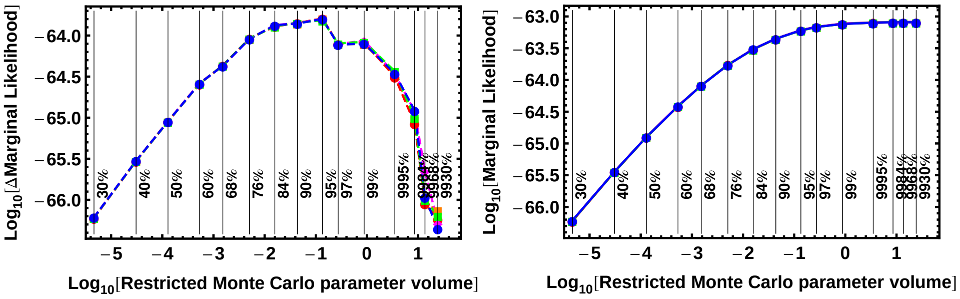


Figure 1.34 Left panel shows the contribution of the individual nested intervals to the NRMC marginal likelihood (for five repeats) based on a 1 planet model fit to the HD 208487 data. The right panel shows the sum of these contributions versus the parameter volume of the credible region. See the text for meaning of the XXXX% boundary (e.g. 9995%).

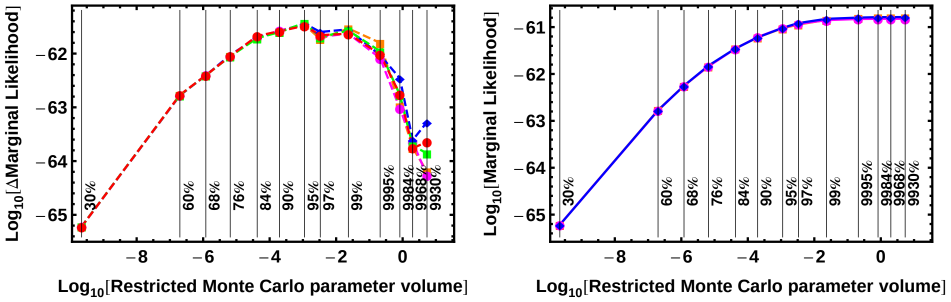


Figure 1.35 Left panel shows the contribution of the individual nested intervals to the NRMC marginal likelihood (for five repeats) based on a 2 planet model fit to the HD 208487 data. The right panel shows the sum of these contributions versus the parameter volume of the credible region.

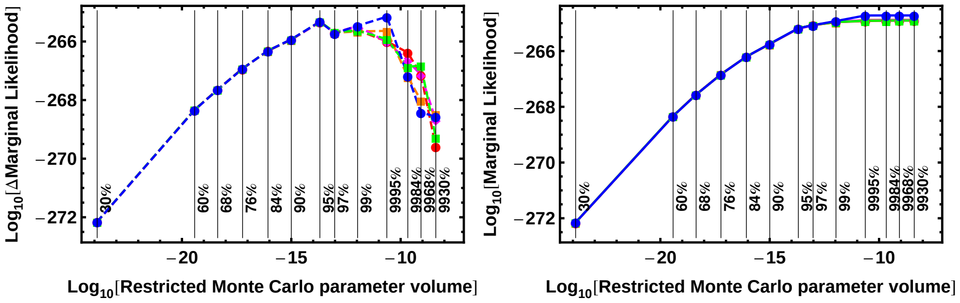


Figure 1.36 Left panel shows the contribution of the individual nested intervals to the NRMC marginal likelihood (for five repeats) based on a 3 planet model fit to the Gliese 581 data. The right panel shows the sum of these contributions versus the parameter volume of the credible region.

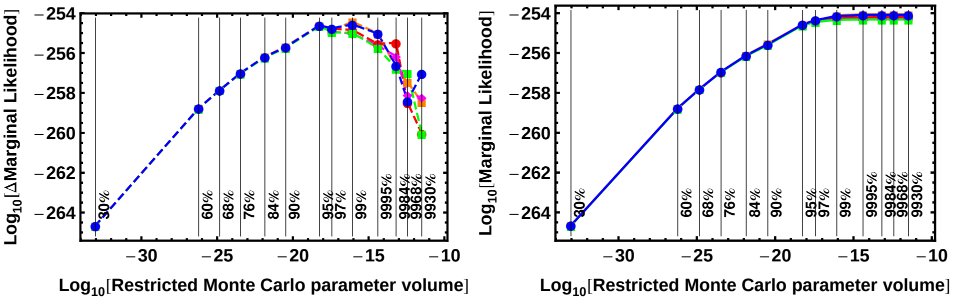


Figure 1.37 Left panel shows the contribution of the individual nested intervals to the NRMC marginal likelihood (for five repeats) based on a 4 planet model fit to the Gliese 581 data. The right panel shows the sum of these contributions versus the parameter volume of the credible region.

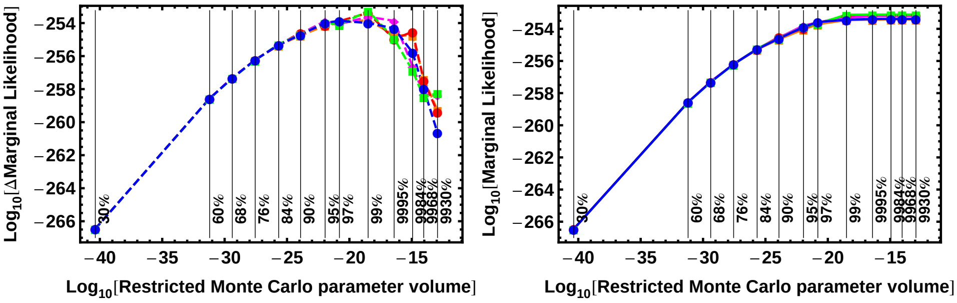


Figure 1.38 Left panel shows the contribution of the individual nested intervals to the NRMC marginal likelihood (for five repeats) based on a 5 planet model fit to the Gliese 581 data. The right panel shows the sum of these contributions versus the parameter volume of the credible region.

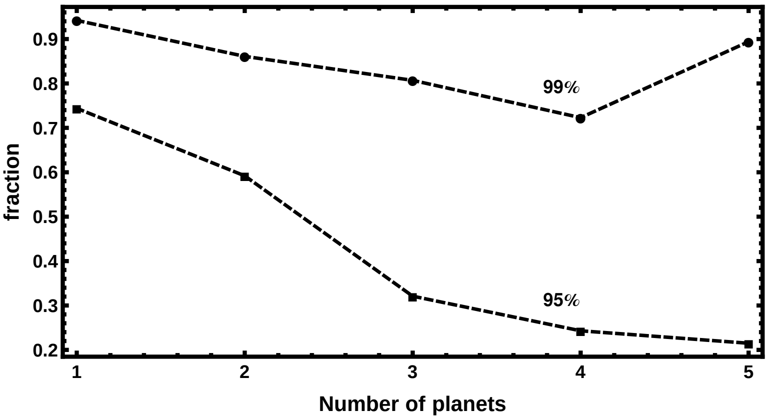


Figure 1.39 The fraction of the total NRMC marginal likelihood within the MCMC 95% and 99% credible regions versus the number of planets.

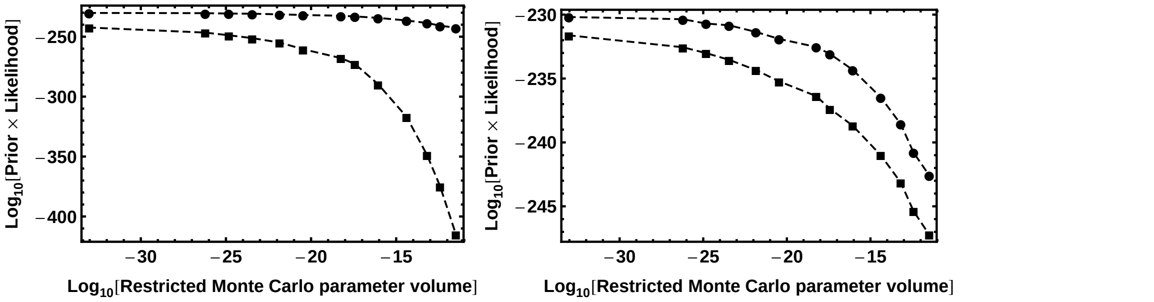


Figure 1.40 Left panel shows the maximum and min values of the Log$_{10}$[prior × likelihood] for each interval of credible region versus parameter volume for the NRMC 4 planet fit samples. The right panel shows the maximum and mean values of the Log$_{10}$[prior × likelihood] versus the parameter volume.

from the outer credible region intervals starting around 99%. The reason for the increased scatter in the Log$_{10}$[Δ Marginal Likelihood] is apparent when we examine the NRMC samples. Fig 1.40 shows plots of the maximum and minimum values (Left) and maximum and mean values (Right) derived from the NRMC samples, for the 4 planet model, of the Log$_{10}$[prior × likelihood] for each interval of credible region, versus the parameter volume. The range of Log$_{10}$[prior × likelihood] values increase rapidly with increasing parameter volume starting around the 99% credible region boundary. This makes the Monte Carlo evaluation of the mean value more difficult in these outer intervals. Based on Figure 1.39 for the 4 planet model, the fraction of the total marginal likelihood estimate that arises for the intervals beyond the 99% credible region is 28%. This fraction increases to 76% for all intervals beyond the 95% credible region.

What about the efficiency of the NRMC method? For example, when we sample the volume within the 95% credible region boundaries and reject all samples within the 90% we need a significant portion to fall in the region between the two boundaries. We also want some samples to be rejected to insure that the sampling between the two boundaries extends throughout. This efficiency was examined as a function of the number of planets fit for the particular choice of boundaries used in the study. The average efficiency ranged from 60% for a one planet fit to 95% for 5 planet fit which is quite reasonable. Extending this analysis to many more planets will likely require a finer grained selection of boundaries to avoid 100% efficiency.

### 1.6.4 Comparison of marginal likelihood methods

In my earlier discusion of the RMC method [28], I indicated that the method is expected to underestimate the marginal likelihood in higher dimensions and this underestimate is expected to become worse the larger the number of model parameters, i.e. increasing number of planets. This is because, in high dimensions most of the random MC samples will fall close the outer boundaries of that volume and so the sampling can easily under sample interior regions of high probability leading to an underestimate of the marginal likelihood. Nested RMC overcomes this problem [22] and in addition extends the outer credible region boundary allowing for a growing contribution (with increasing model complexity) to the marginal likelihood from lower probability density regions as demonstrated in Figure 1.39. Further confidence in this assertion comes from the comparison between the RE, NRMC, and PT marginal likelihood estimators shown in Figure 1.33. For the one planet case the RE, NRMC, PT yielded values in the ratio 1.0 : 0.96 : 1.82, respectively. For the two planet case the values were in the ratio 1.0 : 0.75 : 0.52. For three planets the values were in the ratio 1.0 : 2.22 : 0.94. For four planets the values for the RE, NRMC methods were in the ratio 1.0 : 5.6, respectively. For five planets the values for the RE, NRMC methods were in the ratio 1.0 : 2.4, respectively. So for three planets and beyond the NRMC is yielding higher values by up to a factor of nearly 6 in the 4 planet case. What is going on here?

To help understand this, the right hand column of plots in Figure 1.33 also shows a dashed horizontal gray bar which corresponds to the NRMC marginal likelihood contribution within the 95% credible region. For the 3 to 5 planet case this provides much better agreement with the RE estimate. The RE estimator depends on the MCMC samples. The MCMC sample density is proportional to the probability density of the target probability distribution, i.e., proportional to the $\mathrm{Log}_{10}$[prior × likelihood]. As expected, we found that the range of MCMC $\mathrm{Log}_{10}$[prior × likelihood] values was significantly reduced when a one tenth subset of iterations was extracted. This suggests that dynamic range issues can limit marginal likelihood estimators that depend on the MCMC samples, like the RE estimator. In NRMC we only use the MCMC samples as a guide for setting up nested hypercubes. The exact values of the credible region associated with any given hypercube is not important. We proceed to generate additional nested intervals until the contribution to the marginal likelihood is negligible. The contribution of lower probability density regions to the NRMC estimator is responsible for the larger marginal likelihood values.

It is clear from Figures 1.34 to 1.38, and 1.39, that the contribution from very low probability density regions is much lower for the one and two planet cases so the agreement between the NRMC and RE methods is better.

What lessons can be learned from this?

[22] In several subsequent publications (e.g., [33]) I mistakenly claimed that the NRMC method, like the RMC method, would be prone to underestimate the true marginal likelihood. The current analysis indicates that NRMC only underestimates because we are neglecting the possible contribution to the marginal likelihood from regions of parameter space outside those shown to be significant from the Fusion MCMC exploration.

- The NRMC method is not limited by the dynamic range of methods relying on the MCMC posterior samples (e.g., RE method), leading to a better estimate of the marginal likelihood enabling shorter MCMC simulations.
- Of the three methods the NRMC is not only conceptually simpler to appreciate but also much faster to compute making it quick to run multiple trials to examine repeatability.
- In some situations the MCMC analysis leads to a multiple choice of signal configurations. For both the RE and NRMC methods, it is possible to filter the MCMC samples to select the individual signal configurations separately allowing for a calculation of their relative probability. For the PT method this is not possible and only the global marginal likelihood of that model can be evaluated.
- The PT method is really only computationally feasible for up to and including three planets. The PT method must also suffer from dynamic range limitations for a finite number of MCMC samples.
- The NRMC method works for even larger number of planets but care must be taken in the choice of credible region boundaries to insure that some samples are always being rejected to insure good sampling of each volume shell. To date, the author has successfully employed NRMC on up to 8 planet models involving 43 parameters.
- As a general rule the repeatability spread increases with additional parameters suggesting the need for more samples when dealing with larger numbers of parameters. However, keep in mind that in some cases we only need marginal likelihoods that are accurate to a factor of 2 because we usually require Bayes factors of $> 100$ to achieve sufficiently low Bayesian false alarm probabilities (see Sec 1.6.5) to justify the more complicated model.

In the following sections our model comparison conclusions will be based on Bayes factors computed from NRMC marginal likelihood estimates.

### *1.6.5 Bayesian false alarm probability*

We can readily convert the Bayes factors, which was introduced in Equation 1.24, to a Bayesian False Alarm Probability (FAP) which we define in Equation 1.34. For example, in the context of claiming the detection of $m$ planets the FAP$_m$ is the probability that there are actually fewer than $m$ planets, i.e., $m - 1$ or less.

$$\mathrm{FAP}_m = \sum_{i=0}^{m-1} (\text{prob.of i planets}) \tag{1.34}$$

If we assume *a priori* that all models under consideration are equally likely, then the probability of each model is related to the Bayes factors by

$$p(M_i \mid D, I) = \frac{B_{i1}}{\sum_{j=0}^{N} B_{j1}} \tag{1.35}$$

where $N$ is the maximum number of planets in the hypothesis space under consideration, and of course $B_{11} = 1$. For the purpose of computing FAP$_m$ we set $N = m$. Suppose $m = 2$ then Equation 1.34 gives

$$\mathrm{FAP}_2 = \frac{(B_{01} + B_{11})}{\sum_{j=0}^{2} B_{j1}} \tag{1.36}$$

Lets now evaluate the Bayes factors and false alarm probabilities for our two example data sets, HD 208487 and Gliese 581.

### *HD 208487*

Table 1.3 summarizes the NRMC marginal likelihood estimates for the models under consideration and the corresponding Bayes factors relative to model 1. Initially, we are interested in whether there is a single planet ($m = 1$) which yields a very small false alarm probability of $1.4 \times 10^{-4}$. The question then shifts to false alarm probability of two planets ($m = 2$).

Table 1.3 *HD 208487 NRMC marginal likelihood estimates, Bayes factors relative to model 1, and false alarm probabilities. The quoted errors are the spread in the results for 5 repeats, not the standard deviation.*

| Model | Periods<br>(d) | Marginal<br>Likelihood | Bayes factor<br>nominal | False Alarm<br>Probability |
|---|---|---|---|---|
| $M_0$ | | $1.44 \times 10^{-68}$ | $1.77 \times 10^{-5}$ | |
| $M_1$ | (130) | $(8.13^{+0.09}_{-0.03}) \times 10^{-64}$ | 1 | $1.4 \times 10^{-4}$ |
| $M_{2a}$ | (29, 130) | $(1.83^{+0.05}_{-0.03}) \times 10^{-62}$ | 22.5 | 0.90 |
| $M_{2b}$ | (130, 900) | $(1.65^{+0.19}_{-0.11}) \times 10^{-61}$ | 203 | 0.10 |
| $M_2$ | (29, 130) or (130, 900) | $(1.83^{+0.19}_{-0.11}) \times 10^{-61}$ | 225 | $4.4 \times 10^{-3}$ |

For HD 208487 there are two possible choices of two planet models which we label 2a, for the 29, 130 d period combination, and 2b, for the 130, 900 d combination. In the case of model 2b, Equation 1.36 becomes

$$\mathrm{FAP}_2b = \frac{(B_{01} + B_{11} + B_{2a1})}{(B_{01} + B_{11} + B_{2a1} + B_{2b1})} = 0.10 \tag{1.37}$$

We will use the label 2 to represent a two planet model regardless of which of the two period configurations is true. In this case we can rewrite Equation 1.36 as

$$\mathrm{FAP}_2 = \frac{(B_{01} + B_{11})}{(B_{01} + B_{11} + B_{2a1} + B_{2b1})} = 4.4 \times 10^{-3} \tag{1.38}$$

On this basis of the false alarm probailities, there is significant evidence for a second planet like signal but at present not enough data to determine which of the two period combinations is correct although the evidence currently favors the 130, 900 d combination.

### *Gliese 581*

Table 1.4 *Gliese 581 NRMC marginal likelihood estimates, Bayes factors relative to model 4, and false alarm probabilities. The quoted errors are the spread in the results for 5 repeats, not the standard deviation.*

| Model | Periods<br>(d) | Marginal<br>Likelihood | Bayes factor<br>nominal | False Alarm<br>Probability |
|---|---|---|---|---|
| $M_0$ | | $5.32 \times 10^{-393}$ | $7.9 \times 10^{-139}$ | |
| $M_1$ | (5.37) | $(1.45 \pm 0.004) \times 10^{-295}$ | $2.2 \times 10^{-41}$ | $3.7 \times 10^{-98}$ |
| $M_2$ | (5.37, 12.9) | $(5.55^{+0.26}_{-0.09}) \times 10^{-273}$ | $2.6 \times 10^{-19}$ | $2.6 \times 10^{-23}$ |
| $M_3$ | (5.37, 12.9, 66.9) | $(1.40^{+0.5}_{-0.15}) \times 10^{-265}$ | $2.1 \times 10^{-11}$ | $3.9 \times 10^{-8}$ |
| $M_4$ | (3.15, 5.37, 12.9, 66.9) | $(6.7^{+2.2}_{-1.8}) \times 10^{-255}$ | 1.0 | $2.1 \times 10^{-11}$ |
| $M_{5a}$ | (3.15, 5.37, 12.9, 66.9, 192) | $(5.8^{+1.5}_{-1.9}) \times 10^{-254}$ | 8.7 | 0.19 |
| $M_{5b}$ | (3.15, 5.37, 12.9, 66.9,not 192) | $(0.7^{+0.18}_{-0.22}) \times 10^{-254}$ | 1.0 | 0.90 |
| $M_5$ | (3.15, 5.37, 12.9, 66.9,all) | $(6.5^{+1.7}_{-2.1}) \times 10^{-254}$ | 9.7 | 0.093 |
| $M_6$ | (3.15, 5.37, 12.9, 66.9, 71, 190) | $(5.21^{+1.5}_{-1.7}) \times 10^{-252}$ | 778 | 0.014 |

Table 1.4 summarizes the NRMC marginal likelihood estimates for the models under consideration and the corresponding Bayes factors relative to model 4. The false alarm probability makes a good case for up to 4 signals[23]. For a 5 signal case, we consider two alternatives. Case 5a corresponds to a fifth period of 192 days for which the contribution to the marginal likelihood for a 5 planet model was computed to be $(5.8^{+1.5}_{-1.9}) \times 10^{-254}$ using the NRMC method. Other choices for a fifth period were found (see Figure 1.27) and we represent them collectively as case 5b, and designate the period as "not 192 d." We estimate their collective contribution

[23] Recent analysis of $H_\alpha$ stellar activity for Gliese 581 indicate a correlation between the RV and stellar activity which leads to the conclusion that the 67 d signal (Gl 581d) is not planetary in origin [53] (4 planet candidates).

to the marginal likelihood for a 5 planet model by multiplying the ratio of the post burn-in FMCMC samples that were not in the 192 d peak region compared to the number in the 192 d peak times the 192 d marginal likelihood contribution for the scale invariant prior case. The final total 5 planet model marginal likelihood is then the sum of these two contributions. The false alarm probability of our ”preferred” 192 d fifth period candidate is high at 0.19 so there no reasonable case to be made for such a signal on the basis of our current state of knowledge. Also, the case for a five planet model of any period in our prior range does not pass our minimum threshold of a false alarm probability of ≤ 0.01 to be considered significant.

It is sometimes useful to explore the option for additional Keplerian-like signals beyond the point at which the false alarm probability starts to increase. Additional models ranging from 6 to 8 signals were also considered, e.g., see Figure 1.30. The false alarm probability of the 6 signal case has decreased, compared to the 5 signal case, to 0.014. This is still too high to be considered significant. Both 7 signal and 8 signal models were run but resulted in large false alarm probabilities and are not listed.

## 1.7 Impact of stellar activity on RV

In several earlier sections we have referred to the challenges posed by stellar activity that can induce line shape variations that mimic Keplerian RV signals on time scales similar to planetary signals. Many of the presentations at the “Towards Other Earth II” conference in Porto, Portugal (Sept. 2014) concerned methods to deal with activity-induced RV variability. Currently a wide range of different approaches are being explored. They range from the simplest independent Gaussian noise stellar jitter approach to methods that allow for the natural signal correlation in time that results from stellar rotational modulation and the intrinsic evolution of magnetized regions. At the meeting, Xavier Dumusque (Harvard-Smithsonian) proposed a blind competition using realistic fake RV data plus photometry and diagnostics, to allow the community to determine the best strategy to distinguish real planetary signals from stellar activity-induced signals. The results of the competition are to be presented at the ”Extreme Precision Radial Velocity” workshop at Yale, New Haven (6-8 July 2015).

## 1.8 Conclusions

The main focus of this chapter has been on a new fusion MCMC approach to Bayesian nonlinear model fitting. In fusion MCMC the goal has been to develop an automated MCMC algorithm which is well suited to exploring multi-modal probability distributions such as those that occur in the arena of exoplanet research.

This has been accomplished by the fusion of a number of different statistical tools. At the heart of this development is a sophisticated control system that automates the selection of efficient MCMC proposal distributions (including for highly correlated parameters) in a parallel tempering environment. It also adapts to any new significant parameter set that is detected in any of the parallel chains or is bred by a genetic crossover operation. This controlled statistical fusion approach has the potential to integrate other relevant statistical tools as required. A scheme to automate the selection of an efficient set of $\beta$ values used in the parallel tempering is included in Appendix A.

For some special applications it is possible to develop a faster more specialized MCMC algorithm, perhaps for dealing with real time analysis situations. In the current development of fusion MCMC, the primary focus has not been speed but rather to see how powerful a general purpose MCMC algorithm we could develop and automate. That said, the *Mathematica* code does implement parallel processing on as many cores as are available. In real life applications to challenging multi-modal exoplanet data, fusion MCMC is proving to be a powerful tool. One can anticipate that this approach will also allow for the joint analysis of different types of data (e.g., radial velocity, astrometry, and transit information) giving rise to statistical fusion and data fusion algorithms.

In this document, considerable space has been devoted to Bayesian model comparison. In particular, significant new testing and comparison has been carried out for three Bayesian marginal likelihood estimators: (1) parallel tempering (PT), (2) the ratio estimator (RE), and nested restricted Monte Carlo (NRMC). All three are shown to be in good agreement for up to 17 parameters (3 planet model). PT ceased to be computationally practical for 4 or more planet models. Comparison between RE and NRMC was extended to the 5 planet case (27 parameters). On the basis of this comparison we recommend the NRMC method. The NRMC method is not limited by the dynamic range of methods relying on the MCMC posterior samples, leading to a better estimate of the marginal likelihood particular for models with larger numbers of parameters. Of the three, NRMC is not only conceptually simpler to appreciate but also much faster to compute making it quick to run multiple trials to examine repeatability. The NRMC method works for even larger number of planets but care must be taken in the choice of credible region boundaries to insure that some samples are always being rejected to insure good sampling of each volume shell.

# References

[1] Aigrain, S., Pont, F., Zucker, S.: A simple method to estimate radial velocity variations due to stellar activity using photometry, MNRAS, 419, 3147-3158 (2012)

[2] Anglada-Escudé, G., Tuomi, M., Gerlach, E., Barnes, R., Heller, R., Jenkins, J. S., Wende, S., Vogt, S. S., Butler, R. P., Reiners, A. and Hugh R. A. Jones, H. R. A.: A dynamically-packed planetary system around GJ 667C with three super-Earths in its habitable zone, A&A, 556, A126 (2013)

[3] Atchadé, Y. F., Roberts, G. O., & Rosenthal, J. S.: Towards optimal scaling of metropolis-coupled Markov chain Monte Carlo, Stat. Comput., 21(4):555-568 (2011)

[4] Barnes, R. & Grenberg, R.: Stability Limits in Extrasolar Planetary Systems, ApJ, 647, L-163-L166 (2006)

[5] Baluev, R. V.: The impact of red noise in radial velocity planet searches: only three planets orbiting GJ 581?, MNRAS, 429, 2052-2068 (2013)

[6] Ter Braak, C. J. F.: A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces, Statistical Computing. 16, 239-249 (2006)

[7] Bretthorst, G. L.: Bayesian Spectrum Analysis and Parameter Estimation, New York: Springer-Verlag (1988)

[8] Butler, R. P., Wright, J. T., Marcy, G. W., Fischer, D. A., Vogt, S. S., Tinney, C. G., Jones, H. R. A., Carter, B. D., Johnson, J. A., McCarthy, C., and Penny, A. J.: Catalog of Nearby Exoplanets, ApJ, 646, 505-522 (2006)

[9] Campbell, B., Walker, G. A. H., & Yang, S.: A Search for Substellar Companions to Solar-type Stars, ApJ, 331, 902-921 (1988)

[10] Clyde, M. A., Berger, J. O., Bullard, F., Ford, E. B., Jeffreys, W. H., Luo, R., Paulo, R., Loredo, T.: Current Challenges in Bayesian Model Choice. In ‘Statistical Challenges in Modern Astronomy IV,’ G. J. Babu and E. D. Feigelson (eds.), ASP Conf. Ser., 371, 224-240 (2007)

[11] Cumming, A.:Detectability of extrasolar planets in radial velocity surveys , MNRAS, 354, 1165-1176 (2004)

[12] Cumming, A., Dragomir, D.: An Integrated Analysis of Radial Velocities in Planet Searches, MNRAS, 401, 1029-1042 (2010)

[13] Dawson, R. I., & Fabrycky, D. C.: Radial Velocity Planets De-aliased: A New, Short Period for Super-Earth 55 Cnc e, ApJ, 722, 937-953 (2010)

[14] Deeming, T. J.: Fourier Analysis With Unequally-spaced Data, Astrophysics & Space Science, 36, 137-158 (1975)

[15] Deeming, T. J.: Fourier Analysis With Unequally-spaced Data, Astrophysics & Space Science, 42, 257- (1976)

[16] Farr, W.M., Sravan, N., Cantrell, A., Kreidberg, L., Bailyn, C. D., Mandel, I., and Kalogera, V.: The Mass Distribution of Stellar-Mass Black Holes, ApJ, 741, 103-122, (2011)

[17] Feroz, F., Balan, S. T., Hobson, M. P.: Detecting extrasolar planets from stellar radial velocities using Bayesian evidence, MNRAS, 415, 3462-3472 (2011)

[18] Fischer,D. A., Laughlin, G. L., Butler, R. P., Marcy, G. W., Johnson, J., Henry, G.,Valenti,J., Vogt, S. S., Ammons, M., Robinson, S., Spear, G., Strader, J., Driscoll, P., Fuller, A., Johnson, T., Manrao, E., McCarthy, C., Munõz, M., Tah, K. L., Wright, J., Ida, S., Sato, B., Toyota, E., and Minniyi, D.: The N2K Consortium. I. A Hot Saturn Planet Orbiting HD 88133, ApJ, 620, 481-486 (2005)

[19] Ford, E. B.: Quantifying the Uncertainty in the Orbits of Extrasolar Planets, AJ, 129, 1706-1717 (2005)

[20] Ford, E. B.: Improving the Efficiency of Markov Chain Monte Carlo for Analzing the Orbits of Extrasolar Planets. ApJ, 642, 505-522 (2006)

[21] Ford, E. B., & Gregory, P. C.:Bayesian Model Selection and Extrasolar Planet Detection. In 'Statistical Challenges in Modern Astronomy IV,' G. J. Babu and E. D. Feigelson (eds.), ASP Conf. Ser., 371, 189-204 (2007)

[22] Forveille, T., Bonfils, X., Delfosse, X., Alonso, R., Udry, S., Bouchy, F., Gillon, M., Lovis, C., Neves, V., Mayor, M., Pepe, F., Queloz, D., Santos, N. C., Ségransan, D., Almenara, J. M., Deeg, H. J. and Rabus, M.: The HARPS search for southern extra-solar planets ? XXXII. Only 4 planets in the Gl 581 system, 2011, arXiv:1109.2505v1

[23] Geyer, C. J.: Marlov Chain Monte Carlo, in E. M. Keramidas (ed.),'Computing Science and Statistics: 23rd Symposium on the Interface, Interface Foundation, Fairfax Station, 156-163,(1991)

[24] Gregory, P. C.: Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with *Mathematica* Support, Cambridge University Press (2005)

[25] Gregory, P. C.: A Bayesian Analysis of Extra-solar Planet Data for HD 73526. ApJ, 631, 1198-1214 (2005)

[26] Gregory, P. C.: A Bayesian Kepler Periodogram Detects a Second Planet in HD 208487. MNRAS, 374, 1321-1333 (2007)

[27] Gregory, P. C.: A Bayesian Re-analysis of HD 11964: Evidence for Three Planets, in 'Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 27th International Workshop', Saratoga Springs, eds. K. H. Knuth, A. Caticha, J. L. Center, A,Ġiffin, C. C. Rodrguez, AIP Conference Proceedings, 954, 307-314 (2007)

[28] Gregory, P. C.: A Bayesian Periodogram Finds Evidence for Three Planets in HD 11964. MNRAS, 381, 1607-1619 (2007)

[29] Gregory, P. C., and Fischer, D. A.: A Bayesian Periodogram Finds Evidence for Three Planets in 47 Ursae Majoris. MNRAS, 403, 731-747, (2010)

[30] Gregory, P. C.: Bayesian Exoplanet Tests of a New Method for MCMC Sampling in Highly Correlated Parameter Spaces. MNRAS, 410, 94-110 (2011)

[31] Gregory, P. C.: Bayesian Re-analysis of the Gliese 581 Exoplanet System. MNRAS, 415, 2523-2545 (2011)

[32] Gregory, P. C.: Discussion on paper by Martin Weinberg regarding Bayesian Model Selection and Parameter Estimation, in Statistical Challenges in Modern Astronomy V, Eric Feigelson and Jogesh Babu,(eds.), Springer-Verlag (in press 2011)

[33] Gregory, P. C.: Extrasolar Planets via Fusion MCMC, in Astrostatistical Challenges for the New astronomy, Joseph M. Hilbe (ed.), Springer Series in Astrostatistics 1,121-148 (2013)

[34] Gregory, P. C, Lawler, S. M., Gladman, B.: Additional Keplerian Signals in the HARPS data for Gliese 667C: Further Analysis, in Exploring the Formation and Evolution of Planetary Systems, Proc of IAU Symp. 299, B. Matthews and J. Graham (eds.), Cambridge University Press (in press)

[35] Haywood, R. D., Collier Cameron, A., Queloz, D., Barros, S. C. C., Deleuil, M., Fares, R., Gillon, M., Lanza, A. F., Lovis, C., Moutou, C., Pepe, F., Pollacco, D., Santerne, A., Sgransan, D., Unruh, Y. C.: Planets and stellar activity: hide and seek in the CoRoT-7 system, MNRAS, 443, 2517-2531 (2014)

[36] Hukushima, K., and Nemoto, K.: Exchange Monte Carlo Method and Application to Spin Glass Simulations, Journal of the Physical Society of Japan, 65(4), 1604-1608 (1996)

[37] Jaynes, E. T.: How Does the Brain Do Plausible Reasoning?, Stanford University Microwave Laboratory Report 421, 1957, Reprinted in 'Maximum Entropy and Bayesian Methods in Science and Engineering', G. J. Erickson and C. R. Smith, (eds.), Dordrecht: Kluwer Academic Press, 1-29 (1988)

[38] Jaynes, E.T.: Bayesian Spectrum & Chirp Analysis, in *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*, C.R. Smith and G.L. Erickson, D., (eds.), Reidel, Dordrecht, 1-37 (1987)

[39] Jefferys, W. H.: Discusion on "Current Challenges in Bayesian Model Choice" by Clyde et al. In 'Statistical Challenges in Modern Astronomy IV,' G. J. Babu and E. D. Feigelson (eds.), ASP Conf. Ser., 371, 241–244 (2007)

[40] Kipping, D. M.: Parameterizing the exoplanet eccentricity distribution with the Beta distribution, MNRAS, 434, L51-55 (2013)

[41] Loredo, T., 2004: Bayesian Adaptive Exploration, in 'Bayesian Inference And Maximum Entropy Methods in Science and Engineering: 23rd International Workshop', G.J. Erickson & Y. Zhai, (eds.), AIP Conf. Proc. 707, 330-346 (2004)

[42] Loredo, T. L. and Chernoff, D.: Bayesian Adaptive Exploration, in 'Statistical Challenges in Modern Astronomy III', E. D. Feigelson and G. J. Babu (eds.) , Springer-Verlag, New York, 57-69 (2003)

[43] Loredo, T. L., Berger, J. O., Chernoff, D. F., Clyde, M. A., Liu, B.: Bayesian Methods for Analysis and Adaptive Scheduling of Exoplanet Observations, Statistical Methodology, 9, 101-114 (2011)

[44] Lovis, C. et al.: The HARPS search for southern extra-solar planets. XXVIII. Up to seven planets orbiting HD 10180: probing the architecture of low-mass planetary systems, A&A, 528, 112L (2011)

[45] Marcy G. W., Butler R. P.: A Planetary Companion to 70 Virginis, ApJ, 464, L147-151 (1996)

[46] Mayor M., Queloz D., Nature, 378, 355 (1995)

[47] Mayor, M., Bonfils, X., Forveille, T., Delfosse2, X, Udry, S., Bertaux,J.-L., Beust, H., Bouchy, F., Lovis, C., Pepe, F., Perrier, C., Queloz, D., and Santos, N. C., A&A, 507, 487 (2009)

[48] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E.: Equation of state calculation by fast computing machines, Journal of Chemical Physics, 21, 1087-1092 (1953)

[49] Petigura, E. A., Marcy, G. W., Howard, A. W.: Prevalence of Earth-size planets orbiting Sun-like stars, Proc. Natl. Acad. Sc. USA, www.pnas.org/cgi/doi/10.1073/pnas.1321363110

[50] Queloz, D., Henry, G.W., Sivan, J. P., Baliunas, S. L., Beuzit, J. L., Donahue, R. A., Mayor, M., Naef, D., Perrier, C., and Udry, S.: No planet for HD 166435, A&A, 379, 279-287 (2001)

[51] Queloz, D., Bouchy, F., Moutou, C., Hatzes, A., Hbrard, G., Alonso, R., Auvergne, M., Baglin, A., Barbieri, M., Barge, P., Benz, W., Bord, P., Deeg, H. J., Deleuil, M., Dvorak, R., Erikson, A., Ferraz Mello, S., Fridlund, M., Gandolfi, D., Gillon, M., Guenther, E., Guillot, T., Jorda, L., Hartmann, M., Lammer, H., Lger, A., Llebaria, A., Lovis, C., Magain, P., Mayor, M., Mazeh, T., Ollivier, M., Ptzold, M., Pepe, F., Rauer, H., Rouan, D., Schneider, J., Segransan, D., Udry, S., Wuchterl, G.: The CoRoT-7 planetary system: two orbiting super-Earths, A&A, 506, 303-319 (2009)

[52] Roberts, G. O., Gelman, A. and Gilks, W. R.: Weak convergence and optimal scaling of random walk Metropolis algorithms, Annals of Applied Probability, 7, 110-120 (1997)

[53] Robertson, P., Mahadevan, S., Endl, M., Roy, A.: Stellar activity masquerading as planets in the habitable zone of the M dwarf Gliese 581, Science, 345, 440-444 (2014)

[54] Saar, S. H., & Donahue, R. A.: Activity-related Radial Velocity Variation in Cool Stars, ApJ, 485, 319-327 (1997)

[55] Saar, S. H., Butler, R. P., & Marcy, G. W.: Magnetic Activity-related Radial Velocity Variations in Cool Stars: First Results from the Lick Extrasolar Planet Survey, ApJ, 498, L153-157 (1998)

[56] Scargle, J. D.: Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data, ApJ, 263, 835-853 (1982)

[57] Schneider, J., Dedieu, C.,Le Sidaner, P. , Savalle, R., Zolotukhin, I.: Defining and cataloging exoplanets: the exoplanet.eu database, A&A 532, A79-89 (2011)

[58] Shen,Y., and Turner, E. L.: On the Eccentricity Distribution of Exoplanets from Radial Velocity Surveys, ApJ, 685, 553 (2008)

[59] Skilling, J.: Nested Sampling for General Bayesian Computation, Bayesian Analysis 4, pp. 833-860 (2006)

[60] Tuomi, M., Jones, H. R. A., Jenkins, J. S., Tinney, C. G., Butler, R. P., Vogt, S. S., Barnes, J. R., Wittenmyer, R. A., O'Toole, S., Horner, J., Bailey, J., Carter, B. D., Wright, D. J., Salter, G. S., Pinfield, D.: Signals embedded in the radial velocity noise. Periodic variations in the  Ceti velocities, A&A, 551, A79, 21 pages (2013)

[61] Tinney, C. G., Butler, R. P., Marcy, G. W., Jones, H. R. A., Penny, A. J., McCarthy, C., Carter, B. D., & Fischer, D. A.: Three Low-Mass Planets from the Anglo-Australian Planet Search, ApJ, 623, 1171-1179 (2005)

[62] Tuomi, M.: Evidence for nine planets in the HD 10180 system, A&A 543, 52-63 (2012)

[63] Udry, S., Bonfils, X., Delfosse, X., Forveille, T., Mayor, M., Perrier, C., Bouchy, F., Lovis, C., Pepe, F., Queloz, D., and Bertaux, J.-L., A&A, 469, L43 (2007)

[64] Vogt, S.S, Butler, R. P., Rivera, E. J., Haghighipour, N., Henry, G. W., and Williamson, M. H.: The Lick-Carnegie Exoplanet Survey: A 3.1 Earth Mass Planet in the Habitable Zone of the Nearby M3V Star Gliese 581 ApJ, 723, 954-965 (2010)

[65] Wittenmyer, R. A., Endl, M., Cochran, W. D., Levison, H. F., Henry, G. W.: A Search for Multi-Planet Systems Using the Hobby-Eberly Telescope , ApJ Supplement, 182, 97-119 (2009)

[66] Wolszczan, A., & Frail, D.: A Planetary System Around the Millisecond Pulsar PSR1257 + 12, Nature,355, 145-147 (1992)

[67] Wright, J. T.: Radial Velocity Jitter in Stars from the California and Carnegie Planet Search at Keck Observatory, PASP, 117, 657-664 (2005)

[68] Wright, J. T.: Four New Exoplanets and Hints of Additional Substellar Companions to Exoplanet Host Stars, Wright J. T., Marcy G. W., Fischer D. A., Butler R. P., Vogt S. S., Tinney C. G., Jones H. R. A., Carter B. D., Johnson J. A., McCarthy C., Apps K., ApJ, 657, 533-545 (2007)

[69] Zakamska, N. L., Pan, M., Ford, E. B.:Observational biases in determining extrasolar planet eccentricities in single-planet systems, MNRAS, 410, 1895 (2011)

[70] Zechmeister, M. & Kürster, M.: The generalised Lomb-Scargle periodogram. A new formalism for the floating-mean and Keplerian periodograms , A&A, 496, 577-584 (2009)

# 2
# Hidden variables, missing data and multilevel (hierarchical) Bayes

## 2.1 Introduction

In previous chapters we have been concerned with fitting models to data sets with measurement errors in the dependent variable, $y$. In this chapter we will learn how to deal with data subject to measurement errors in the independent $x$ variable (covariate) as well. Thus each independent variable has a true "hidden" value, which we represent by $x_{ti}$. All we know is the set of measured values which we represent by $\{x_i\}$, $\{y_i\}$, or in vector form by the bold faced symbols, $\mathbf{x}, \mathbf{y}$, or more consisely by the proposition $D$, plus any prior information about the distribution of measurement errors. To solve the model parameter estimation problem it is necessary to extend the conversation [1] to include propositions about the hidden independent variables (sometimes called latent variables) by treating them as additional parameters.

To arrive at the Bayesian posterior distribution of the model parameters, conditional on the measurements, we need to integrate over the numerous hidden $\mathbf{x_t}$ parameters/variables after specifying a prior for each $x_{ti}$. As we shall see, in some cases we can carry out this integration analytically. In other cases the integration can still be performed with MCMC. In both cases it proves useful to employ an informative prior for the hidden parameters, one which itself has parameters and these in turn have priors. You can begin to see multiple levels of parameters and priors entering the discussion which is generally referred to as hierarchical Bayes or multilevel modeling [2] (MLM). Hierarchical models can have enough parameters to fit the data well, while the choice of an informative prior for the hidden parameters, structures some dependence into these parameters thereby avoiding problems of over fitting.

It turns out that we can introduce another complication which does not signif-

[1] See Section 4.4 of my book for details about extending the conversation.

[2] According to Loredo and Hendry (2010) [15], MLM is a relatively recent term that underlies several important statistical innovations of the latter twentieth century, including empirical and hierarchical Bayes methods, random effects and latent variable models, shrinkage and ridge regression.

icantly complicate the equations for model fitting with errors in both variables, but greatly extend the reach of the analysis. The complication is to allow for an intrinsic scatter in the relationship between the true values of the dependent and independent variables which takes us into the arena of regression analysis with measurement errors in both variables. Useful results are derived for the specific case of linear regression. Multilevel or hierarchical Bayesian regression can yield representative samples of the underlying regression, effectively deconvolving the blurring effect of the measurement errors. This is generalized to linear regression with multiple independent variables.

The next step is to allow for selection effects which cause some potential data to be missed but we would still like to allow for the effect of this "missing" data on our regression model parameters.

## 2.2 Fitting a straight line with errors in both coordinates

We start by assuming that there is a precise deterministic linear relationship between two variables $x_t, y_t$ of the form $y_t = \alpha + \beta x_t$. We will refer to $x_t, y_t$ as the true variables where $x_t$ is the independent variable and $y_t$ the dependent variable.

Now let's distinguish between the measured value $y_i$ and the corresponding true value $y_{ti}$, where $y_i = y_{ti} + e_{y,i}$, where $e_{y,i}$ represents an unknown error component. Based on our prior knowledge $I$ for this particular problem, the error is assumed to be Gaussian distributed with mean zero and variance $\sigma^2_{y,i}$, commonly written in the statistical literature [3] as $e_{y,i} \sim N(0, \sigma^2_{y,i})$. Substituting for $y_t$ from above gives

$$y_i = \alpha + \beta\, x_{ti} + e_{y,i}. \tag{2.1}$$

When we allow for the measurement error in the independent variable, the measured value $x_i$ is related to the true value by $x_i = x_{ti} + e_{x,i}$, where $e_{x,i} \sim N(0, \sigma^2_{x,i})$ and again the prior information for this particular problem assumes that $\sigma_{x,i}$ will be provided with the data.

This is a common problem in the physical sciences and we can readily apply probability theory to make inferences about $\alpha$ and $\beta$. The starting point is the joint probability distribution $p(\alpha,\beta, \{x_{ti}\}, \{x_i\}, \{y_i\}|I)$ which can be written as $p(\alpha,\beta, \mathbf{x_t}, \mathbf{x}, \mathbf{y}|I)$ or more concisely as $p(\alpha,\beta, \mathbf{x_t}, D|I)$, where $\alpha$ and $\beta$ are the model parameters. Next expand the joint probability in two different ways using the prod-

[3] **Note on notation:** in statistics the symbol "$\sim$" means "is drawn from" or "is distributed as" and should not be confused with the common usage of implying "similar to". For example, $y \sim N(0, \sigma^2)$ means that $y$ has a normal distribution with mean 0 and variance $\sigma^2$. Note: the prior information for this particular problem does not include the precise values of the $\sigma_{y,i}$ and $\sigma_{x,i}$, but rather assumes that they will be provided in the data set together with the measured $x_i, y_i$ values.

uct rule.

$$p(\alpha,\beta,\mathbf{x_t},D,|I) = p(\alpha,\beta,\mathbf{x_t}|I)p(D|\alpha,\beta,\mathbf{x_t},I)$$
$$= p(D|I)p(\alpha,\beta,\mathbf{x_t}|D,I) \tag{2.2}$$

Reorganizing the terms gives us Bayes theorem for this problem

$$p(\alpha,\beta,\mathbf{x_t}|D,I) \propto p(\alpha,\beta,\mathbf{x_t}|I)p(D|\alpha,\beta,\mathbf{x_t},I)$$
$$= p(\alpha,\beta|I)p(\mathbf{x_t}|\alpha,\beta,I)p(D|\alpha,\beta,\mathbf{x_t},I). \tag{2.3}$$

To obtain $p(\alpha,\beta|D,I)$, the marginal probability distribution[4] for $\alpha,\beta$, integrate over $\mathbf{x_t}$.

$$p(\alpha,\beta|D,I) \propto p(\alpha,\beta|I)\int p(\mathbf{x_t}|\alpha,\beta,I)p(D|\alpha,\beta,\mathbf{x_t},I)d\mathbf{x_t}$$
$$= p(\alpha,\beta|I)\int p(D,\mathbf{x_t}|\alpha,\beta,I)d\mathbf{x_t}$$
$$= p(\alpha,\beta|I)p(D|\alpha,\beta,I), \tag{2.4}$$

where $p(D|\alpha,\beta,I) = \int p(D,\mathbf{x_t}|\alpha,\beta,I)d\mathbf{x_t}$. The second and third line of equation (2.4) demonstrate an alternate way of proceeding (introducing $\mathbf{x_t}$ through the likelihood term), obtained by using the product rule to form the joint probability of $D,\mathbf{x_t}$. In what follows we will use this alternate approach to further extend the conversation through the likelihood term.

If the measurement uncertainties are independent then the likelihood is given by

$$p(D|\alpha,\beta,\mathbf{x_t},I) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_{x,i}}\exp[-\frac{(x_i - x_{ti})^2}{2\sigma_{x,i}^2}]$$
$$\times\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_{y,i}}\exp[-\frac{(y_i - \alpha - \beta\, x_{ti})^2}{2\sigma_{y,i}^2}], \tag{2.5}$$

Its clear from equation (2.4), that to complete the integral over $\mathbf{x_t}$ we need to specify a prior that characterizes our prior knowledge about possible $\mathbf{x_t}$ values. We will assume our prior knowledge about $\mathbf{x_t}$ is independent of our prior knowledge of $\alpha$ and $\beta$ so that $p(\mathbf{x_t}|\alpha,\beta,I) = p(\mathbf{x_t}|I)$. As a first guess we might assume an independent flat distribution for each $p(x_{ti}|I)$ and specify some large prior range between $-\lambda_i$ and $+\lambda_i$ that we are confident the $\mathbf{x_t}$ values fall within. Suppose that the first $n$ samples of $x_i$ fall within the much smaller range $-0.01\lambda_i$ and $0.1\lambda_i$. Do we really believe the next $x_{ti}$ is likely to be anywhere in the range $-\lambda_i$ and $+\lambda_i$? Another choice for $p(\mathbf{x_t}|I)$ is to choose what we will refer to as an informative prior like a Gaussian and learn about the mean and variance of the $x_t$ values from the measured

[4] Alternatively, we might be interested in $p(\mathbf{x_t}|D,I)$, the marginal distribution of the hidden variables $\mathbf{x_t}$ in which case we need to integrate over $\alpha$ and $\beta$.

sample [5] and avoid the biased estimates of the intercept and slope common to ordinary least-squares analysis of this situation (Kelly, 2007, Gull 1989). It is termed a hierarchical prior because it depends on two more parameters, the mean $\mu$ and variance $\tau^2$ according to

$$p(x_{ti}|\mu, \tau, I) = \frac{1}{\sqrt{2\pi\tau^2}} \exp[-\frac{1}{2}\frac{(x_{ti} - \mu)^2}{\tau^2}]. \tag{2.6}$$

One can well imagine situations where $p(\mathbf{x_t}|I)$ is very different from a single Gaussian. For the moment we will employ a single Gaussian but in Section 2.5 we will consider a mixture or sum of Gaussians to better model $p(\mathbf{x_t}|I)$. A mixture of Gaussians is flexible enough to model a wide variety of distributions and simplifies the integration over $\mathbf{x_t}$ which we will need to carry out.

Figure 2.1 illustrates graphically the components of the Bayesian calculation. This is a simple example of what is commonly referred to as ‘multilevel Bayes’ or ‘hierarchical Bayes’ as it involves more than two levels relating the parameters of interest to the observed data.

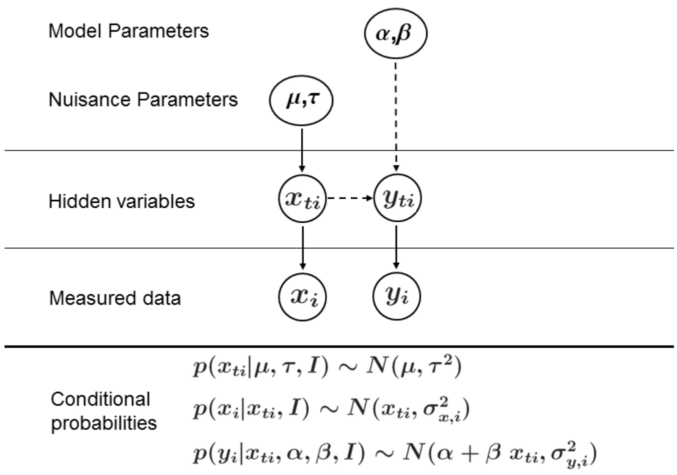

Model Parameters
$\alpha,\beta$
Nuisance Parameters
$\mu,\tau$
Hidden variables
$x_{ti}$
$y_{ti}$
Measured data
$x_i$
$y_i$
Conditional probabilities
$p(x_{ti}|\mu, \tau, I) \sim N(\mu, \tau^2)$
$p(x_i|x_{ti}, I) \sim N(x_{ti}, \sigma^2_{x,i})$
$p(y_i|x_{ti}, \alpha, \beta, I) \sim N(\alpha + \beta\, x_{ti}, \sigma^2_{y,i})$

Figure 2.1 Graphical model of the multilevel Bayes calculation. The solid lines connecting nodes denote conditional dependencies with the conditional probabilities listed below. The absence of a connection denotes conditional independence. The dashed lines represent deterministic conditionals.

[5] This leads to a probabilistic dependence among the $x_{ti}$ values that implements a pooling of information that can improve the accuracy of inference. Each $x_i$ measurement bears on the estimation of the unknown mean and variance of the population of $x_t$ values, and thus indirectly, each measured $x_i$ bears on the estimation of every other $x_{ti}$, via a kind of adaptive bias correction. This is referred to as borrowing strength from each other.

We need to specify priors for $\mu$ and $\tau$ and integrate over these nuisance parameters. The likelihood $p(D|\alpha,\beta,I)$ appearing on the last line of equation (2.5) becomes

$$p(D|\alpha,\beta,I) = \int\int\int p(D,\mathbf{x_t},\mu,\tau|\alpha,\beta,I)\, d\mathbf{x_t}\, d\mu\, d\tau. \tag{2.7}$$

We expand equation (2.7) using the product rule and Replace $D$ on the R.H.S. by $\mathbf{x},\mathbf{y}$.

$$\begin{aligned}
p(D|\alpha,\beta,I) &= \int\int\int p(\mathbf{x_t},\mu,\tau|\alpha,\beta,I)p(\mathbf{x},\mathbf{y}|\alpha,\beta,\mathbf{x_t},\mu,\tau,I)\, d\mathbf{x_t}\, d\mu\, d\tau \\
&= \int\int\int\ p(\mu,\tau|I)\ p(\mathbf{x_t}|\mu,\tau,I) \\
&\times\ p(\mathbf{x}|\mathbf{x_t},I)\ p(\mathbf{y}|\alpha,\beta,\mathbf{x_t},I)\, d\mathbf{x_t}\, d\mu\, d\tau \\
&= \int\int\ d\mu\, d\tau\ p(\mu,\tau|I) \\
&\times \left[\int p(\mathbf{x_t}|\mu,\tau,I)p(\mathbf{x}|\mathbf{x_t},I)p(\mathbf{y}|\alpha,\beta,\mathbf{x_t},I)\, d\mathbf{x_t}\right] \\
&= \int\int\ d\mu\, d\tau\ p(\mu,\tau|I)\ p(D|\alpha,\beta,\mu,\tau,I)
\end{aligned} \tag{2.8}$$

where, in component form,

$$p(D|\alpha,\beta,\mu,\tau,I) = \left[\prod_{i=1}^{n}\int p(x_{ti}|\mu,\tau,I)p(x_i|x_{ti},I)p(y_i|\alpha,\beta,x_{ti},I)\, dx_{ti}\right]. \tag{2.9}$$

Integrating $x_{ti}$ from $-\infty$ to $+\infty$ yields [6] an analytic solution [11] for $p(D|\alpha,\beta,\mu,\tau,I)$ given by

$$p(D|\alpha,\beta,\mu,\tau,I) = \prod_{i=1}^{n}\frac{1}{2\pi\sqrt{\det\mathbf{V}_i}}\exp\left[-\frac{1}{2}(z_i-\zeta)^T\mathbf{V}_i^{-1}(z_i-\zeta)\right], \tag{2.10}$$

where

$$z_i = \begin{pmatrix} y_i \\ x_i \end{pmatrix},\ \ \zeta = \begin{pmatrix} \alpha+\beta\mu \\ \mu \end{pmatrix},\ \ (z_i-\zeta) = \begin{pmatrix} y_i-\alpha-\beta\mu \\ x_i-\mu \end{pmatrix}, \tag{2.11}$$

and

$$\mathbf{V}_i = \begin{pmatrix} \beta^2\tau^2+\sigma_{y,i}^2 & \beta\tau^2 \\ \beta\tau^2 & \tau^2+\sigma_{x,i}^2 \end{pmatrix}, \tag{2.12}$$

and $(z_i-\zeta)^T$ is the transpose of $(z_i-\zeta)$.

[6] If the $-\infty$ to $+\infty$ limits are not a reasonable assumption then the result given in equation (2.9) needs to be multiplied by the difference of two error functions (erf). These erf functions can be determined by evaluating the integral with *Mathematica*.

Substituting equation (2.8) into equation (2.5) we have

$$p(\alpha,\beta|D,I) \propto p(\alpha,\beta|I) \int\int d\mu\, d\tau\; p(\mu,\tau|I)\; p(D|\alpha,\beta,\mu,\tau,I). \tag{2.13}$$

The remaining integrals can be evaluated numerically. We will find it convenient to rewrite equation (2.13) as

$$p(\alpha,\beta,\mu,\tau|D,I) \propto p(\alpha,\beta,\mu,\tau|I)\; p(D|\alpha,\beta,\mu,\tau,I). \tag{2.14}$$

This gives the joint posterior distribution for $\alpha,\beta,\mu,\tau$. A useful way of computing the marginal distribution for any of these parameters is with a Markov chain Monte Carlo approach like fusion MCMC (FMCMC) method described in Chapter 1 of this supplement. We will do that shortly after first introducing a further complication in Section 2.3.

### *2.2.1 Correlated xy data errors*

Following Kelly (2007), we can generalize the result to allow for correlations in the measurement errors of **x** and **y**. In this case we write $y_i, x_i|y_{ti}, x_{ti} \sim N_2([y_{ti}, x_{ti}], \mathbf{\Sigma}_i)$, where $N_2$ is a bivariate normal density distribution with a two element mean and a 2 × 2 covariance matrix $\mathbf{\Sigma}$ given by

$$\mathbf{\Sigma}_i = \begin{pmatrix} \sigma_{y,i}^2 & \sigma_{xy,i} \\ \sigma_{xy,i} & \sigma_{x,i}^2 \end{pmatrix}. \tag{2.15}$$

The only change is to introduces a term $\sigma_{xy,i}$ into the $\mathbf{V}_i$ matrix which becomes.

$$\mathbf{V}_i = \begin{pmatrix} \beta^2\tau^2 + \sigma_{y,i}^2 & \beta\tau^2 + \sigma_{xy,i} \\ \beta\tau^2 + \sigma_{xy,i} & \tau^2 + \sigma_{x,i}^2 \end{pmatrix}. \tag{2.16}$$

## 2.3 Linear regression with errors in both coordinates

In this section, we generalize our previous results for fitting a straight line with errors in both coordinates by adding another feature which greatly extends the reach of the solution to the arena of linear regression following the treatment of Kelly (2007, 2013) [11] [13] and Gelman et al. (2004) [7]. Kelly (2007) [11] extended the statistical model of Carroll et al. (1999) [3] to allow for measurement errors of different magnitudes (i.e., heteroscedastic errors), non detections, and selection effects, so long as the selection function can be modeled mathematically. Linear regression is one of the most common techniques used in data analysis. We hasten to add that the Bayesian approach developed here is not limited to linear models.

It is a common problem in astronomy to explore whether there is a correlation (i.e., a straight line relationship, commonly referred to as the regression line), between the dependent and independent variables. Initially we will consider a single independent variable. There is often some intrinsic scatter about the regression line. The intrinsic scatter arises from variations in the physical properties that are not completely captured by the independent variables included in the regression, i.e., in this case the single independent variable $x_t$. We can allow for an intrinsic scatter in the relationship between $x_t$ and $y_t$ according to the additive noise model

$$y_{ti} = \alpha + \beta\, x_{ti} + \epsilon_i, \tag{2.17}$$

where $\epsilon_i \sim N(0, \sigma^2)$ and $\sigma$ is unknown and treated as a model parameter.

We now allow for independent measurement errors in both the dependent and independent variables according to

$$x_i = x_{ti} + e_{x,i}, \tag{2.18}$$

where $e_{x,i} \sim N(0, \sigma^2_{x,i})$ and $\sigma_{x,i}$ is assumed known. Also

$$y_i = y_{ti} + e_{y,i}, \tag{2.19}$$

where $e_{y,i} \sim N(0, \sigma^2_{y,i})$ and $\sigma_{y,i}$ is assumed known. For the moment we are ignoring possible correlated errors. Now both $\mathbf{x_t}$ and $\mathbf{y_t}$ are hidden variables that we need to marginalize over. A graphical model of this multilevel analysis is shown in Figure 2.2 along with the conditional probabilities.

We could use as our starting point the joint probability distribution

$$p(\alpha, \beta, \mu, \tau, \sigma, \{x_{ti}\}, \{y_{ti}\}, \{x_i\}, \{y_i\}|I)$$

which can be written as

$$p(\alpha, \beta, \mu, \tau, \sigma, \mathbf{x_t}, \mathbf{y_t}, \mathbf{x}, \mathbf{y}|I)$$

or more concisely as

$$p(\alpha, \beta, \mu, \tau, \sigma, \mathbf{x_t}, \mathbf{y_t}, D|I)$$

and follow the steps outlined in equations (2.2) to (2.4). Instead we will use the equivalent approach mentioned following the discussion of equation (2.4) to introduce the additional hidden variables, $\mathbf{y_t}$, and the intrinsic scatter $\sigma$ parameter into our likelihood equation as we did in equations (2.7) to (2.9) for $\mathbf{x_t}, \mu, \tau$. We need to provide probability distributions that characterize our prior knowledge of $\mathbf{y_t}$ and $\sigma$ and then integrate over these distribution[7].

[7] Alternatively, we might be interested in $p(\mathbf{x_t}, \mathbf{y_t}|D, I)$, the marginal distribution of the hidden variables in which case we need to integrate over $\alpha, \beta, \mu, \tau, \sigma$.

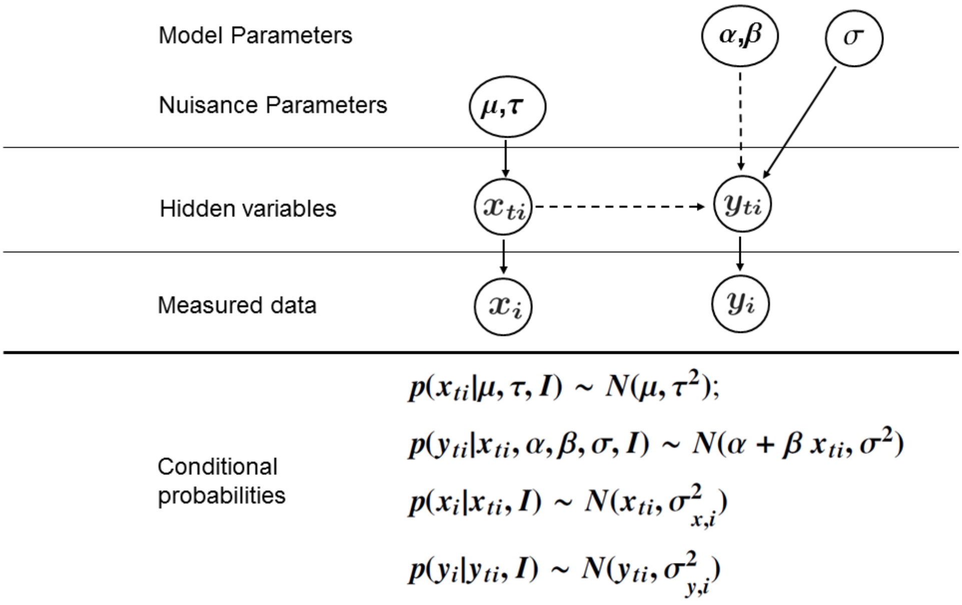

Model Parameters
$\alpha,\beta$
$\sigma$
Nuisance Parameters
$\mu,\tau$
Hidden variables
$x_{ti}$
$y_{ti}$
Measured data
$x_i$
$y_i$
Conditional probabilities
$p(x_{ti}|\mu,\tau,I) \sim N(\mu,\tau^2)$;
$p(y_{ti}|x_{ti},\alpha,\beta,\sigma,I) \sim N(\alpha+\beta\, x_{ti},\sigma^2)$
$p(x_i|x_{ti},I) \sim N(x_{ti},\sigma^2_{x,i})$
$p(y_i|y_{ti},I) \sim N(y_{ti},\sigma^2_{y,i})$

Figure 2.2 Graphical model for a simple multilevel Bayesian regression analysis. The solid lines connecting nodes denote conditional dependencies with the conditional probabilities listed below. The absence of a connection denotes conditional independence. The dashed lines represent deterministic conditionals.

Starting from equation (2.7) we can add the additional terms as follows

$$\begin{aligned} p(D|\alpha,\beta,I) &= \int\int\int\int\int p(D,\mathbf{x_t},\mathbf{y_t},\mu,\tau,\sigma|\alpha,\beta,I)\, d\mathbf{x_t}\, d\mu\, d\tau\, d\mathbf{y_t}\, d\sigma \\ &= \int\int\int\int\int p(\mathbf{x_t},\mathbf{y_t},\mu,\tau,\sigma|\alpha,\beta,I) \\ &\quad \times p(D|\alpha,\beta,\mathbf{x_t},\mathbf{y_t},\mu,\tau,\sigma,I)\, d\mathbf{x_t}\, d\mu\, d\tau\, d\mathbf{y_t}\, d\sigma. \end{aligned} \tag{2.20}$$

Expanding equation (2.20) further using the product rule, and recognizing that $p(\mathbf{x_t},\mathbf{y_t}|\alpha,\beta,\mu,\tau,\sigma,I) = p(\mathbf{x_t}|\mu,\tau,I)\, p(\mathbf{y_t}|\alpha,\beta,\mathbf{x_t},\sigma,I)$, yields

$$\begin{aligned} p(D|\alpha,\beta,I) &= \int\int\int\int\int p(\mu,\tau,\sigma|I)\, p(\mathbf{x_t}|\mu,\tau,I)\, p(\mathbf{y_t}|\alpha,\beta,\mathbf{x_t},\sigma,I) \\ &\quad \times p(\mathbf{x}|\mathbf{x_t},I)\, p(\mathbf{y}|\mathbf{y_t},I)\, d\mathbf{x_t}\, d\mu\, d\tau\, d\mathbf{y_t}\, d\sigma \\ &= \int\int\int p(\mu,\tau,\sigma|I)\, d\mu\, d\tau\, d\sigma \end{aligned}$$

$$\times \int \left\{ \int p(\mathbf{y}|\mathbf{y_t}, I)\; p(\mathbf{y_t}|\alpha, \beta, \mathbf{x_t}, \sigma, I)\; d\mathbf{y_t} \right\}$$
$$\times\; p(\mathbf{x}|\mathbf{x_t}, I)\; p(\mathbf{x_t}|\mu, \tau, I)\; d\mathbf{x_t}. \tag{2.21}$$

Expanding equation (2.21) in terms of components yields

$$\begin{aligned} p(D|\alpha,\beta, I) &= \int \int \int d\mu\; d\tau\; d\sigma\; p(\mu, \tau, \sigma|I) \\ &\times \left[ \prod_{i=1}^{n} \int \mathcal{I}_{y_i}\; p(x_{ti}|\mu, \tau, I)\; p(x_i|x_{ti}, I)\; dx_{ti} \right] \\ &= \int \int \int d\mu\; d\tau\; d\sigma\; p(\mu, \tau, \sigma|I)\; p(D|\alpha,\beta, \sigma, \mu, \tau, I), \end{aligned} \tag{2.22}$$

where

$$\begin{aligned} \mathcal{I}_{y_i} &= \int_{-\infty}^{+\infty} p(y_i|y_{ti}, I) p(y_{ti}|x_{ti}, \alpha, \beta, \sigma, I)\; dy_{ti} \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_{y,i}} \exp[-\frac{(y_i - y_{ti})^2}{2\sigma_{y,i}^2}] \\ &\times \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(y_{ti} - \alpha - \beta x_{ti})^2}{2\sigma^2}]\; dy_{ti} \\ &= \frac{1}{\sqrt{2\pi\; (\sigma^2 + \sigma_{yi}^2)}} \exp[-\frac{(y_i - \alpha - \beta x_{ti})^2}{2(\sigma^2 + \sigma_{yi}^2)}]. \end{aligned} \tag{2.23}$$

Inserting the result for $\mathcal{I}_{y_i}$ from equation (2.23) into equation (2.22), and integrating $x_{ti}$ from $-\infty$ to $+\infty$, yields an analytic solution for $p(D|\alpha,\beta, \sigma, \mu, \tau, I)$ given by

$$p(D|\alpha,\beta, \sigma, \mu, \tau, I) = \prod_{i=1}^{n} \frac{1}{2\pi\; \sqrt{\det \mathbf{V}_i}} \exp\left[ -\frac{1}{2}(z_i - \zeta)^T \mathbf{V}_i^{-1} (z_i - \zeta) \right], \tag{2.24}$$

where

$$z_i = \begin{pmatrix} y_i \\ x_i \end{pmatrix}, \;\; \zeta = \begin{pmatrix} \alpha + \beta\mu \\ \mu \end{pmatrix}, \;\; (z_i - \zeta) = \begin{pmatrix} y_i - \alpha - \beta\mu \\ x_i - \mu \end{pmatrix}, \tag{2.25}$$

and

$$\mathbf{V}_i = \begin{pmatrix} \beta^2\tau^2 + \sigma^2 + \sigma_{y,i}^2 & \beta\tau^2 \\ \beta\tau^2 & \tau^2 + \sigma_{x,i}^2 \end{pmatrix}. \tag{2.26}$$

Comparing these last three equations to our earlier result in Section 2.2 (i.e., equations (2.10) to (2.12)), we see that the only change is the addition of a $\sigma^2$ (intrinsic scatter) to equation (2.12).

To obtain the marginal distributions of any of the regression parameters, $\alpha, \beta, \sigma$, or nuisance parameters, $\mu, \tau$, we can integrate over the remaining parameters numerically, perhaps by exploiting a MCMC approach like fusion MCMC (FMCMC), as described in Chapter 1 of this supplement. The results of such an analysis are illustrated in the sample problem below.

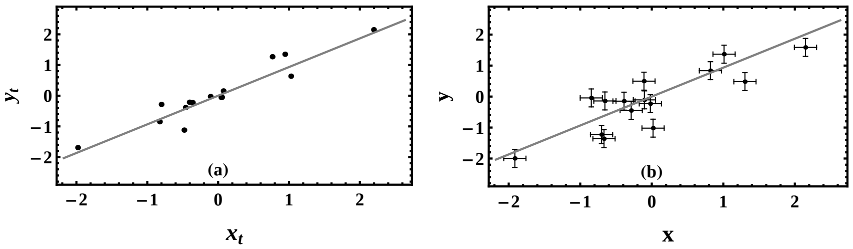


Figure 2.3 Panel (a) shows the true regression line (gray) and the simulated true values (black points) which include the intrinsic scatter. Panel (b) shows the same regression line (gray) and the simulated measured values which include both the intrinsic scatter and measurement errors in both coordinates.

### 2.3.1 Example 1

| $x$ | $y$ | $x_t$ | $y_t$ |
|---|---|---|---|
| 0.0182045 | -1.01896 | 0.0428525 | -0.0574915 |
| -0.387978 | -0.147827 | -0.404814 | -0.212603 |
| -0.108916 | 0.497445 | 0.053096 | -0.0516395 |
| -0.65517 | -0.141506 | -0.460119 | -0.384669 |
| 0.819814 | 0.834607 | 0.765019 | 1.26963 |
| 1.01058 | 1.36803 | 0.943613 | 1.35037 |
| -1.91422 | -1.99596 | -1.98054 | -1.69163 |
| 1.30227 | 0.481718 | 1.02763 | 0.635988 |
| -0.701947 | -1.22742 | -0.825174 | -0.847801 |
| -0.103098 | -0.106871 | 0.0740896 | 0.154979 |
| -0.845334 | -0.0436895 | -0.801533 | -0.283528 |
| -0.0200126 | -0.230387 | -0.10878 | -0.0262748 |
| -0.667872 | -1.3606 | -0.479666 | -1.12028 |
| 2.14888 | 1.58631 | 2.19575 | 2.14965 |
| -0.287609 | -0.450218 | -0.360076 | -0.226541 |

Table 2.1 *The table contains 15 pairs of measured $x, y$ values and true $x_t, y_t$ values.*

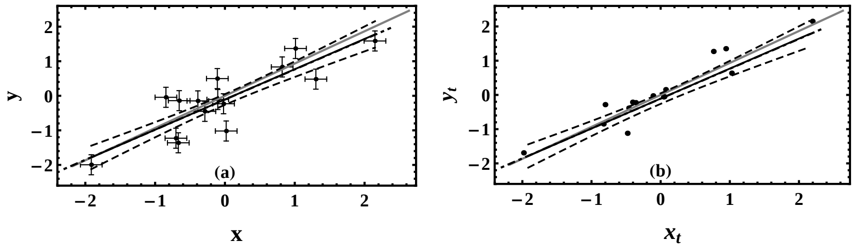


Figure 2.4 Panel (a) shows the true regression line (gray), the simulated measured values including intrinsic scatter and measurement errors in both coordinates, the mean fitted line (black) and the MAP fit (dot-dashed) derived from the FMCMC iterations. The dashed lines are the mean fitted line ±1 standard deviation in the FMCMC fit uncertainty. The error bars indicate the 1 sigma IID measurement errors. Panel (b) is the same as (a) with the measured values replaced by the simulated true values including only the intrinsic scatter (black points).

Panel (a) of Figure 2.3 shows the true regression line (gray) and the simulated true values (black points) after including the intrinsic scatter. Panel (b) shows the same regression line (gray) and the simulated measured values which were derived from the simulated true values by adding measurement errors in both coordinates. The data set is given in Table 2.1. The error bars indicate the 1 sigma IID measurement errors which are $\sigma_x = 0.155$ and $\sigma_y = 0.290$. The values of the parameters employed in this simulation were $\alpha = 0.0, \beta = 0.933, \sigma = 0.290, \mu = 0.0, \tau = 1.086$. The true regression line equation is

$$y_{ti} = 0.0 + 0.933\, x_{ti} + \epsilon_i, \tag{2.27}$$

where $\epsilon_i \sim N(0, \sigma^2)$ and $\sigma = 0.290$.

Our starting point is the joint posterior distribution for $\alpha, \beta, \sigma, \mu, \tau$.

$$p(\alpha, \beta, \sigma, \mu, \tau | D, I) \propto p(\alpha, \beta, \sigma, \mu, \tau | I)\ p(D | \alpha, \beta, \sigma, \mu, \tau, I), \tag{2.28}$$

where $p(D|\alpha, \beta, \sigma, \mu, \tau, I)$ is given by equations (2.24) to (2.26). We assume independent prior distributions for $\alpha, \beta, \sigma, \mu, \tau$ so

$$p(\alpha, \beta, \sigma, \mu, \tau | I) = p(\alpha|I) p(\beta|I) p(\sigma|I) p(\mu|I) p(\tau|I). \tag{2.29}$$

Flat priors were adopted for $\alpha, \beta, \mu$ and normalized modified scale invariant priors for both $\tau$ and $\sigma$ of the form

$$p(\tau|I) = \frac{(\tau + \tau_0)^{-1}}{\ln(1 + \frac{\tau_{max}}{\tau_0})}. \tag{2.30}$$

For $\tau < \tau_0$, $p(\tau|I)$ behaves like a flat prior and for $\tau > \tau_0$ behaves like a scale invariant prior. We designate this prior by FTSI for flat to scale invariant [8]. The break point $\tau_0$ was set = Mean[$\sigma_{x,i}$]. Similarly for $\sigma$,

$$p(\sigma|I) = \frac{(\sigma + \sigma_0)^{-1}}{\ln(1 + \frac{\sigma_{max}}{\sigma_0})}. \tag{2.31}$$

The break point $\sigma_0$ was set = 2 × Mean[$\sigma_{y,i}$].

We used the fusion Markov chain Monte Carlo (FMCMC) method to explore $p(\alpha, \beta, \sigma, \mu, \tau|D, I)$, compute the marginal distributions and the mean and MAP regression lines. Panel (a) of Figure 2.4 includes: (1) the measured values together with the true regression line (gray), (2) the mean fitted line (black) and the maximum *a posterior* (MAP) fit derived from the FMCMC iterations (dot-dashed), and (3) the mean fitted line ±1 standard deviation in the FMCMC fit uncertainty (dashed curves). The fit uncertainty is computed as follows. Each post burn-in FMCMC iteration yields an intercept and slope. We compute a set of model $y$ predictions for a uniform grid of $x$ values for that particular intercept and slope. This is repeated for each FMCMC iteration. At each $x$ grid point the mean and standard deviation of the corresponding y values are computed. The fit uncertainty curves are then plots of this grid of mean ± 1 standard deviation values.

Panel (b) of Figure 2.4 shows the same information as in panel (a) with the measured data replaced by the true values (black points) of the simulated data set. The maximum *a posterior* (MAP) fit coincides with the mean fit line to within the line width.

### *Alternate approach*

Next we investigate an alternate way of proceeding with a Bayesian regression analysis which has some advantages and disadvantages. Our starting point is the joint probability distribution for $\alpha, \beta, \sigma, \mu, \tau, \mathbf{x_t}, \mathbf{y_t}$ where we include the hidden variables as additional parameters to explore with MCMC.

$$\begin{aligned}
&p(\alpha, \beta, \sigma, \mu, \tau, \mathbf{x_t}, \mathbf{y_t}|D, I) \\
&\propto p(\alpha, \beta, \sigma, \mu, \tau, \mathbf{x_t}, \mathbf{y_t}|I)\; p(D|\alpha, \beta, \sigma, \mu, \tau, \mathbf{x_t}, \mathbf{y_t}, I) \\
&= p(\alpha, \beta, \sigma, \mu, \tau|I) p(\mathbf{x_t}|\mu, \tau, I) p(\mathbf{y_t}|\alpha, \beta, \sigma, \mathbf{x_t}, I) \\
&\times p(\mathbf{x}|\mathbf{x_t}, I) p(\mathbf{y}|\mathbf{y_t}, I)
\end{aligned} \tag{2.32}$$

In our previous analysis we analytically integrated over the hidden variables which was possible because we assumed Gaussian distributions for the intrinsic scatter, $\sigma$, and the measurement errors. In other situations the relevant distributions might be known but non-Gaussian and not amenable to an analytic integration. We can

[8] If we are uncertain about the scale of a parameter it is common to employ a scale invariant prior. This has an unfortunate singularity at 0 which the FTSI version overcomes by introducing a break point which is set initially to a simple multiple of the mean measurement error in that coordinate.

still explore the joint distribution within the Bayesian MCMC framework but now need to treat $\mathbf{x_t}$ and $\mathbf{y_t}$ as additional parameters to explore with the MCMC. Another advantage of this approach is that we can expose representative samples of the true coordinates $\mathbf{x_t}, \mathbf{y_t}$ of the underlying regression. In effect we are de-convolving the blurring effect of the measurement errors.

What are the disadvantages? As the number of measurements increases so does the number of parameters that must be computationally explored with the MCMC. The total number of parameters has risen to 35 in this example. What information do we have to constrain these parameters? We have 30 pieces of information represented by $D$, namely, the 15 values of $x_i$ and 15 values of $y_i$. In addition, accomanying $D$ are the 15 values of $\sigma_{x,i}$, and the same again for $\sigma_{y,i}$. It is well to remember that we also have a great deal of prior information on the parameters and their relationships in the form of the deterministic and conditional probabilities illustrated in the graphical model of the multilevel analysis shown in Figure 2.2. Multilevel models can have enough parameters to fit the data well, while the choice of an informative prior for the hidden parameters structures some dependence into these parameters thereby avoiding problems of over fitting. The value of an informative prior was discussed earlier in Section 2.1.

Another disadvantage is the joint distribution might contain multiple peaks. One of these might have the highest peak probability by a significant factor but the integrated probability associated with this peak might be negligible compared to a much broader lower probability peak. The presence of this dominant peak with negligible integrated probability can complicate the implementation of an MCMC algorithm as we shall see in Section 2.3.2.

In the current problem $p(\mathbf{x_t}|\mu, \tau, I)p(\mathbf{y_t}|\alpha, \beta, \sigma, \mathbf{x_t}, I)p(\mathbf{x}|\mathbf{x_t}, I)p(\mathbf{y}|\mathbf{y_t}, I)$ that appears in equation (2.32) are given by

$$\prod_{i=1}^{n}[\frac{1}{\sqrt{2\pi}\tau}\exp[-\frac{(x_{ti}-\mu)^2}{2\tau^2} \times \frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{(y_{ti}-\alpha-\beta x_{ti})^2}{2\sigma^2}$$

$$\times \frac{1}{\sqrt{2\pi}\sigma_{x,i}}\exp[-\frac{(x_i-x_{ti})^2}{2\sigma_{x,i}^2} \times \frac{1}{\sqrt{2\pi}\sigma_{y,i}}\exp[-\frac{(y_i-y_{ti})^2}{2\sigma_{y,i}^2}]. \qquad (2.33)$$

The other priors in equation (2.32) are the same as before.

We again used the fusion Markov chain Monte Carlo (FMCMC) method to explore $p(\alpha, \beta, \sigma, \mu, \tau, \mathbf{x_t}, \mathbf{y_t}|D, I)$, compute the marginal distributions and mean and MAP regression lines. Figure 2.5 shows our initial results compared to our previous results using analytic integration over $\mathbf{x_t}, \mathbf{y_t}$. Panel (a) shows $\log_{10}$[prior × likelihood] versus $\sigma$ (the regression intrinsic scatter parameter) for a sample of the FMCMC iterations that employed analytic integration over the hidden variables $\mathbf{x_t}, \mathbf{y_t}$. Panel (b) shows the same plot for the second FMCMC run in which the

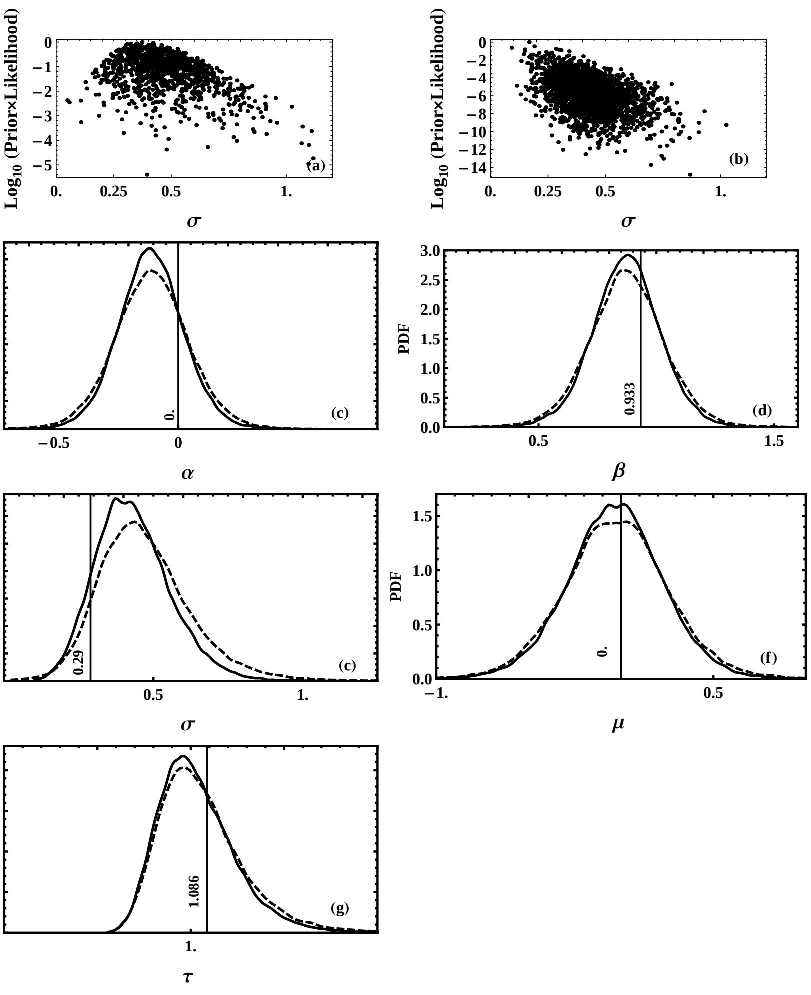


Figure 2.5 First attempt. Panel (a) shows $\log_{10}$[prior × likelihood] versus $\sigma$ (the regression intrinsic scatter parameter) for a sample of the FMCMC iterations with analytic integration over the hidden variables $\mathbf{x_t}, \mathbf{y_t}$. Panel (b) shows the same plot for the second FMCMC run in which the hidden variables were treated as parameters in the joint distribution. Panels (c) to (g) compares the posterior marginal distributions for the regression line parameters $\alpha,\beta,\sigma$ and nuisance parameters $\mu, \tau$ for the two cases. The dashed curves are for analytic integration over the hidden variables $\mathbf{x_t}, \mathbf{y_t}$ and the solid curves apply for the hidden variables treated as additional parameters in the FMCMC.

hidden variables were treated as parameters in the joint distribution. Although the bulk of the samples fall in the same range from about 0.2 to 0.8 as in panel (a), the $\log_{10}$[prior × likelihood] is much higher at small values of $\sigma$ compared to analytic case [9]. This arises when $\mathbf{x_t}, \mathbf{y_t}$ are treated as parameters in the FMCMC because there are two peaks in the full joint probability distribution. The highest occurs for a $\sigma$ near 0 but has a negligible integrated probability as evidenced from the first analysis where the hidden variables are integrated analytically. The second peak in $\sigma$ coincides with the peak found in the analytic version of FMCMC.

Panels (c) to (g) compares the posterior marginal distributions for the regression line parameters $\alpha, \beta, \sigma$ and nuisance parameters $\mu, \tau$ for the two cases. The dashed curves are for analytic integration over the hidden variables $\mathbf{x_t}, \mathbf{y_t}$ and the solid curves apply for the hidden variables treated as additional parameters in the FMCMC. The true parameter values are indicated by the vertical lines.

In the normal mode of operation the FMCMC control system anneals the Metropolis proposal distributions with respect to the highest peak in the distribution. Because the highest peak coincided with low $\sigma$ value, the width of the $\sigma$ proposal distribution for 'I' proposals [10] was driven to low values making it more difficult for FMCMC to explore larger values of $\sigma$ with the 'I' proposals. The 'C' proposals continued to function effectively in this case.

### *Second run*

Another option of FMCMC is to fix the width of the 'I' proposal distributions. For this case we set the width of the 'I' proposal distributions to those found from the analytic FMCMC run and set the width's for the $\mathbf{x_t}, \mathbf{y_t}$ parameters to $0.3\times$ Mean$[\sigma_{x,i}]$, $0.5\times$ Mean$[\sigma_{y,i}]$. Figure 2.6 shows the results for this second run. Again panel (a) shows a sample of the FMCMC iterations with analytic integration over the hidden variables, $\mathbf{x_t}, \mathbf{y_t}$, for $\log_{10}$[prior × likelihood] versus $\sigma$, the regression intrinsic scatter parameter. Panel (b) shows the same plot for the second FMCMC run in which the hidden variables were treated as parameters in the joint distribution. In this case the two distributions are much more similar and the marginal distributions shown in panels (c) to (g) are essentially identical. Again, the true parameter values are indicated by the vertical lines.

One advantage of treating $\mathbf{x_t}, \mathbf{y_t}$ as parameters in a multilevel (hierarchical) Bayesian MCMC regression regression analysis is that it yields representative samples of the underlying regression, effectively de-convolving the blurring effect of the mea-

[9] Loredo and Hendry (2010) [15] note that a log-flat $\sigma$ prior can lead to an un-normalizable posterior in some MLM problems. When we treat $\mathbf{x_t}, \mathbf{y_t}$ as additional parameters in the MCMC analysis, there is the potential for a singularity at $\sigma = 0$ leading to an un-normalizable posterior. Instead, priors flat in $\sigma$ or $\sigma^2$ (among others) are advocated. The prior for $\sigma$ used in our analysis changes from flat for $\sigma < 2 \times \sigma_y$ to log-flat for $\sigma > 2 \times \sigma_y$. Kelly (2007) employed a flat prior for $\sigma^2$ which corresponds to a $\sigma$ prior $\propto \sigma$. There is considerable research on the topic of prior distributions for variance parameters in hierarchical models, e.g., [2] [8].

[10] FMCMC employs two different proposal schemes, the 'I' proposals and the 'C' proposals as discussed in Appendix A. The 'I' scheme is ideally suited for the exploration of independent parameters while the 'C' scheme is well suited to dealing with correlated parameters. Each scheme is employed 50% of the time and the two are designed to work together.

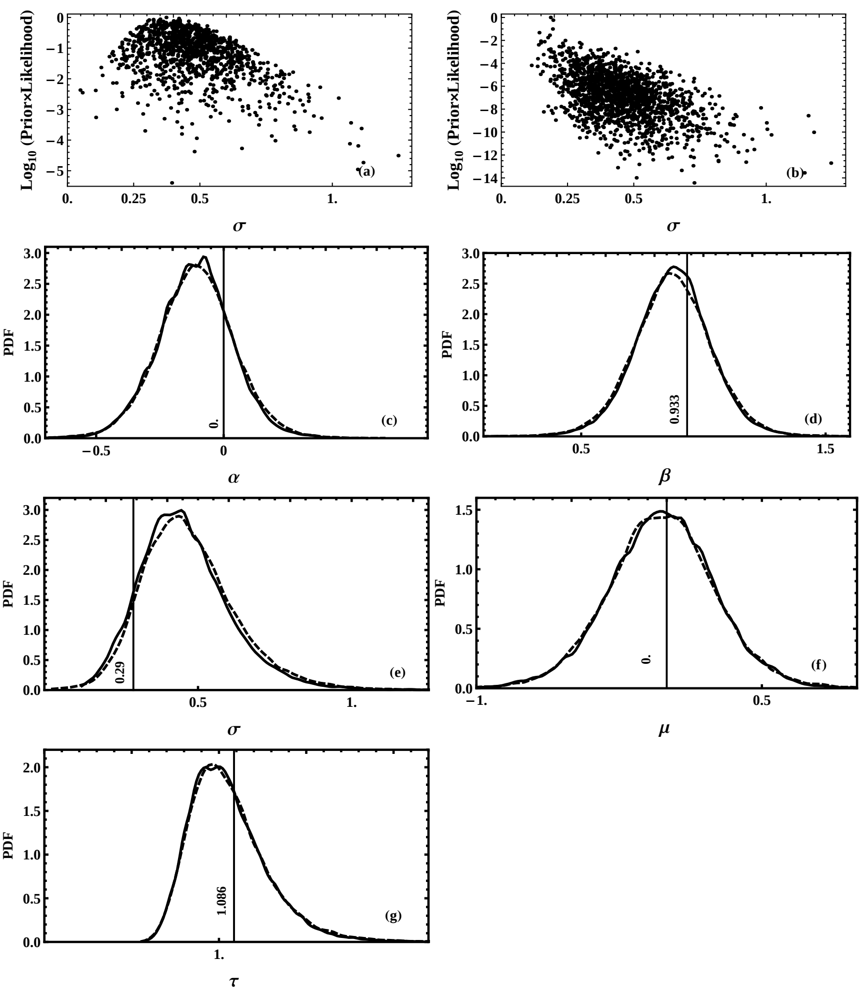


Figure 2.6 Second run. Panel (a) shows $\log_{10}$[prior × likelihood] versus $\sigma$ (the regression intrinsic scatter parameter) for a sample of the FMCMC iterations with analytic integration over the hidden variables $\mathbf{x_t}, \mathbf{y_t}$. Panel (b) shows the same plot for the second FMCMC run in which the hidden variables were treated as parameters in the joint distribution. Panels (c) to (g) compares the posterior marginal distributions for the regression line parameters $\alpha, \beta, \sigma$ and nuisance parameters $\mu, \tau$ for the two cases. The dashed curves are for analytic integration over the hidden variables $\mathbf{x_t}, \mathbf{y_t}$ and the solid curves apply for the hidden variables treated as additional parameters in the FMCMC.

surement errors. This is illustrated in Figure 2.7. Panel (a) includes: (1) the true regression line (gray) and measured values, (2) the mean fitted line (black) and the maximum *a posterior* (MAP) fit (dot-dashed) derived from the FMCMC iterations,

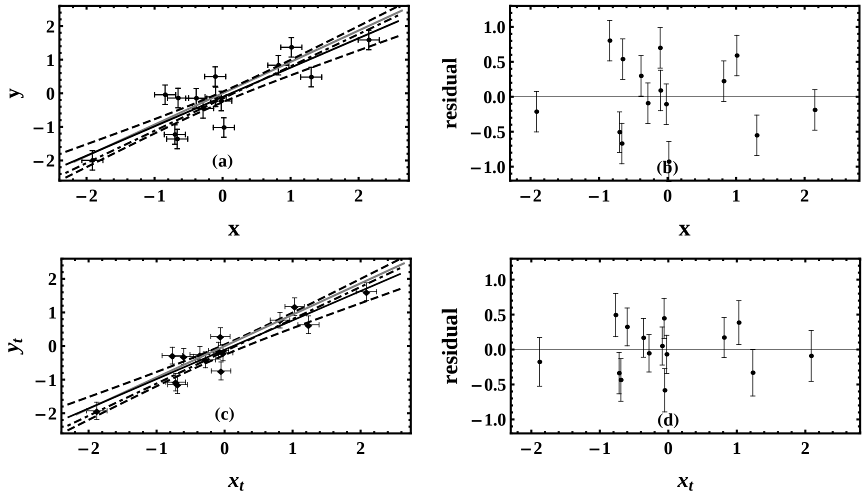


Figure 2.7 FMCMC analysis which includes the hidden variables as parameters. Panel (a) shows the true regression line (gray) and measured values, the mean fitted line (black) and MAP fit (dot-dashed) derived from the FMCMC iterations, and the mean fitted line ±1 standard deviation in the FMCMC fit uncertainty (dashed curves). Panel (b) shows the residuals from the mean fit. In panel (c), the points and error bars are the mean and standard deviation of the MCMC estimates of the true coordinates and the other lines are the same as in (a). Panel (d) shows the residuals of the mean estimates of the true coordinates from the mean fit.

and (3) the mean fitted line ±1 standard deviation in the FMCMC fit uncertainty (dashed curves). Panel (b) shows the residuals from the mean fit. In panel (c), the points and error bars are the mean and standard deviation of the FMCMC estimates of the true coordinates and the other lines are the same as in (a). The FMCMC starting values for $\mathbf{x_t}, \mathbf{y_t}$ were the measured data set. Panel (d) shows the residuals of the mean estimates of the true coordinates from the mean fit. The residuals in panel (d) are significantly smaller than in panel (b).

### *2.3.2 Example 2*

In this example we increased the measurement errors in both coordinates compared to the previous example and increased the number of data points from 15 to 22. As with the previous example, we carried out the FMCMC regression analysis in two different ways: (a) analytic integration over the hidden variables $\mathbf{x_t}, \mathbf{y_t}$ and (b) treating $\mathbf{x_t}, \mathbf{y_t}$ as additional parameters. We again employed a FTSI (flat to scale invariant) prior for the $\sigma$ and $\tau$ parameters and flat priors for $\alpha, \beta$ and $\mu$. We set the width of the 'I' proposal distributions to those found from the analytic FMCMC run

and set the width's for the $\mathbf{x_t}, \mathbf{y_t}$ parameters to 0.3× Mean[$\sigma_{x,i}$], 0.5× Mean[$\sigma_{y,i}$]. The values of the parameters employed in this simulation were
$\alpha = 0.0, \beta = 0.911, \sigma = 0.251, \mu = 0.0, \tau = 0.963$. The measurement errors were $\sigma_x = 0.275$ and $\sigma_y = 0.501$.

Panel (a) of Figure 2.8 shows a sample of the FMCMC iterations with analytic integration over the hidden variables $\mathbf{x_t}, \mathbf{y_t}$. It shows a plot of $\log_{10}$[prior × likelihood] versus $\sigma$, the regression intrinsic scatter parameter. The larger measurement errors result in significant probability extending to a value of $\sigma = 0$, although the marginal for $\sigma$ shown by the dashed curve in panel (e) peaks close to the true value of $\sigma$ of 0.251. Panel (b) shows the corresponding plot for the FMCMC run in which the hidden variables were treated as parameters in the joint distribution. The $\log_{10}$[prior × likelihood] exhibits a narrow strong peak at a value of $\approx 0.07$ while the marginal distribution for $\sigma$ exhibits a broad peak centered on the true value as shown by the solid curve in panel (e). For the other parameters the marginal distributions shown in panels (c), (d), (f) and (g) are essentially identical for the two cases.

## 2.4 Effect of measurement error on correlation and regression

In the previous two sections we carefully developed the equations for fitting a straight line and conducting Bayesian linear regression when there are measurement errors in both coordinates. In this section we will briefly explore the potential consequences of ignoring these measurement errors as is commonly done in ordinary least-squares regression.

The goal of regression is often to understand how one variable changes with another. If the data are measured without error, the ordinary least squares estimate of the regression slope, $\hat{\beta}_{\mathrm{OLS}}$, and the estimated correlation coefficient, $\hat{\rho}$, are

$$\hat{\beta}_{\mathrm{OLS}} = \frac{\mathrm{Cov}(\mathbf{x_t}, \mathbf{y_t})}{\mathrm{Var}(\mathbf{x_t})}, \tag{2.34}$$

$$\hat{\rho} = \frac{\mathrm{Cov}(\mathbf{x_t}, \mathbf{y_t})}{\sqrt{\mathrm{Var}(\mathbf{x_t})\mathrm{Var}(\mathbf{y_t})}} = \hat{\beta}_{\mathrm{OLS}} \sqrt{\frac{\mathrm{Var}(\mathbf{x_t})}{\mathrm{Var}(\mathbf{y_t})}}, \tag{2.35}$$

where $\mathbf{x_t} = (x_{t1}, \cdots, x_{tn})$ and $\mathbf{y_t} = (y_{t1}, \cdots, y_{tn})$ are the true values of the variables. $\mathrm{Cov}(\mathbf{x_t}, \mathbf{y_t})$ is the sample covariance [11] between $\mathbf{x_t}$ and $\mathbf{y_t}$ and $\mathrm{Var}(\mathbf{x_t})$ is the variance

[11] The random variable sample covariance is given by

$$\mathrm{Cov} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}), \tag{2.36}$$

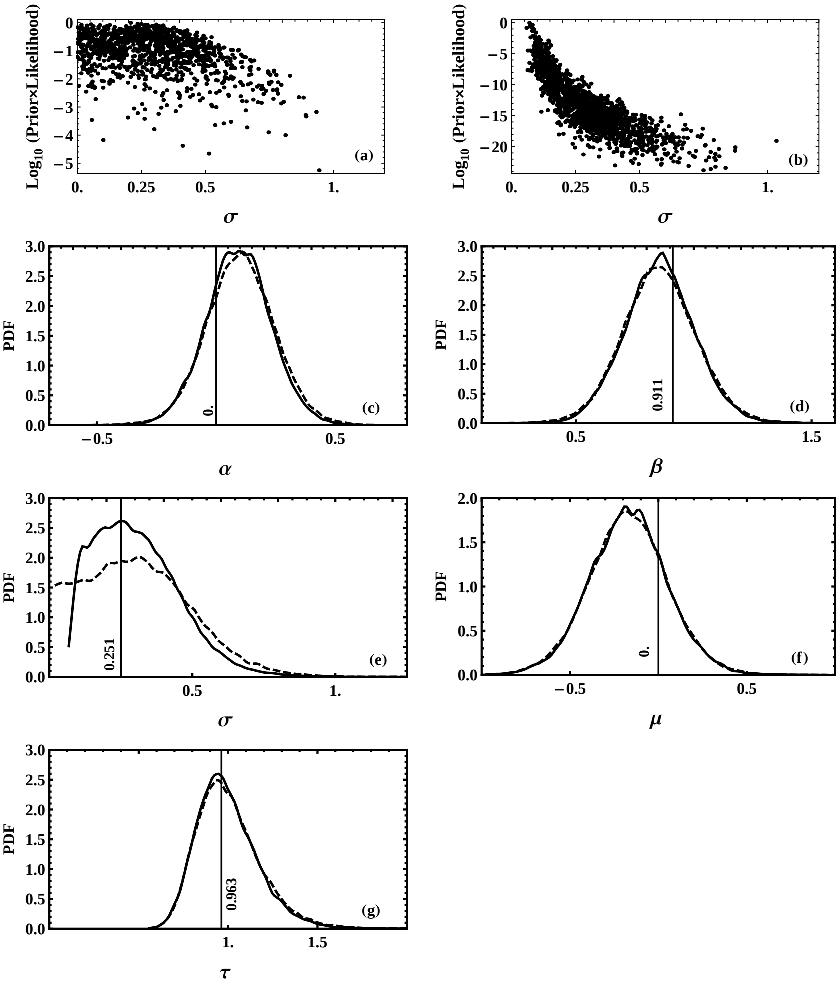


Figure 2.8 Panel (a) shows $\log_{10}$[prior × likelihood] versus $\sigma$ (the regression intrinsic scatter parameter) for a sample of the FMCMC iterations with analytic integration over the hidden variables $\mathbf{x_t}, \mathbf{y_t}$. Panel (b) shows the same plot for the second FMCMC run in which the hidden variables were treated as parameters in the joint distribution. Panels (c) to (g) compares the posterior marginal distributions for the regression line parameters $\alpha, \beta, \sigma$ and nuisance parameters $\mu, \tau$ for the two cases. The dashed curves are for analytic integration over the hidden variables $\mathbf{x_t}, \mathbf{y_t}$ and the solid curves apply for the hidden variables treated as additional parameters in the FMCMC. The vertical lines indicate the true values.

of $\mathbf{x_t}$. The square of the correlation coefficient can be shown to be proportional to

where $\bar{X}$ is the mean of $X_i$. The correlation coefficient derived from the sample covariance and sample the fraction of Var($\mathbf{y_t}$) which is accounted for by the regression and has values in the range 0 to 1, while $\rho$ can range from $-1$ to $+1$.

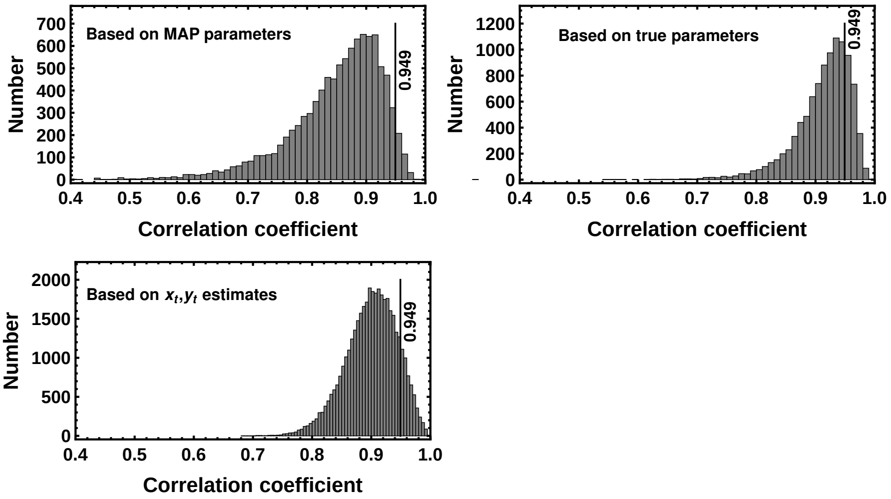


Figure 2.9 The correlation coefficient distribution derived from the Bayesian linear regression analysis of Section 2.3.1. The distributions are from (a) simulations of the hidden variables, $\mathbf{x_t}, \mathbf{y_t}$, based on the MAP parameters from the first solution (upper left), (b) from simulations of $\mathbf{x_t}, \mathbf{y_t}$ based on the true parameter set (upper right), and (c) from the estimates of the true coordinates, $\mathbf{x_t}, \mathbf{y_t}$ from the second solution (lower left). The black line indicates the correlation coefficient computed from the simulated true data set given in Table 2.1.

In Section 2.3.1 we carried out a Bayesian linear regression analysis of a simulated data set using two different approaches. The second of these treated the hidden true coordinates as additional MCMC model parameters and yielded distributions for the true coordinates. From these and equation (2.36) we estimated the distribution of the correlation coefficient ($\rho$) which is shown in Figure 2.9. The three distributions shown are from (a) simulations of the hidden variables, $\mathbf{x_t}, \mathbf{y_t}$, based on the MAP parameters from the first solution (upper left), (b) from simulations of $\mathbf{x_t}, \mathbf{y_t}$ based on the true parameter set (upper right), and (c) from the estimates of the true coordinates, $\mathbf{x_t}, \mathbf{y_t}$ from the second solution (lower left). For (a) and (c) the distributions peak around an $\rho \approx 0.9$. The black line indicates the correlation coefficient computed from the simulated true data set given in Table 2.1.

variance is given by

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \tag{2.37}$$

and is often called the Pearson product-moment correlation coefficient.

Let $\hat{b}_{\text{OLS}}$ = the estimated slope and $\hat{r}$ = the estimated correlation coefficient when there are measurement errors. Then

$$\hat{b}_{\text{OLS}} = \frac{\text{Cov}(\mathbf{x},\mathbf{y})}{\text{Var}(\mathbf{x})} = \frac{\text{Cov}(\mathbf{x_t},\mathbf{y_t}) + \sigma_{xy}}{\text{Var}(\mathbf{x_t}) + \sigma_x^2}, \tag{2.38}$$

$$\hat{r} = \frac{\text{Cov}(\mathbf{x},\mathbf{y})}{\sqrt{\text{Var}(\mathbf{x})\text{Var}(\mathbf{y})}} = \frac{\text{Cov}(\mathbf{x_t},\mathbf{y_t}) + \sigma_{xy}}{\sqrt{(\text{Var}(\mathbf{x_t}) + \sigma_x^2)(\text{Var}(\mathbf{y_t}) + \sigma_y^2)}}, \tag{2.39}$$

The second expression on the right of equations (2.38) and (2.39) relates the covariance and variances derived from the measured variables to the desired covariance and variances of the true variables designated $\mathbf{x_t}, \mathbf{y_t}$. From this it is clear that the effect of measurement error ($\sigma_x$) in the independent variable, $\mathbf{x}$, is to bias the slope towards zero and reduce the magnitude of the observed correlation. Measurement error ($\sigma_y$) in the response, $\mathbf{y}$, also reduces the magnitude of the correlation. Finally, if the measurement errors are correlated the effect depends on the sign of the correlation. If the measurement error correlation has the same sign as the intrinsic correlation between $\mathbf{x_t}$ and $\mathbf{y_t}$, then the affect is to cause a spurious increase in the observed correlation. If the sign is opposite this results in a spurious decrease in observed correlation.

Equation (2.39) suggests that one way to estimate $\rho$ using the measured coordinates, assuming $\sigma_{xy} = 0$, is given by Equation (2.40).

$$\hat{\rho} = \frac{\text{Cov}(\mathbf{x},\mathbf{y})}{\sqrt{(\text{Var}(\mathbf{x}) - \sigma_x^2)(\text{Var}(\mathbf{y}) - \sigma_y^2)}}, \tag{2.40}$$

where we replaced $\mathbf{x_t}, \mathbf{y_t}$ by the measured values, $\mathbf{x}, \mathbf{y}$, and subtracted the variance of the measured errors in the denominator [12]. Using the measured data and measurement errors from Section 2.3.1 we obtain a $\rho = 0.9$, in close agreement with

[12] **Alternative Linear Regression Approaches:** Kelly (2007 & 2013) and Carroll et al. (2006), review some of the other non Bayesian approaches to linear regression with measurement errors. In least-squares linear regression analysis, estimates of the slope, intercept and intrinsic dispersion are obtained from moments of the data. As we have just seen these yield biased estimators in the presence of measurement errors. A number of other methods, termed method of moments estimators (MM), debiase the moments by removing the contributions from the measurement errors (e.g., BCES method by Akritas & Bershady, 1996). The main advantage of MM estimators are that they do not make any assumptions about the distribution of the measurement errors, the distribution of $\mathbf{x_t}$, nor the intrinsic scatter. This makes MM estimators robust. However, because they do not make use of prior information about the distributions of the errors, intrinsic dispersion and $\mathbf{x_t}$, they are not as precise as methods that do or employ parametric models of these distributions. Another disadvantage is that the MM estimators are highly variable when the sample size is small and/or the measurement errors are large. According to Cheng & Van Ness (1999) the MM estimators can be understood as arising from the minimization with respect to $(\alpha, \beta, \sigma)$ of a modified least-squares loss function such as

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}[(y_i - \alpha - \beta x_i)^2 - \sigma_{y,i}^2 - \beta^2\sigma_{x,i}^2]. \tag{2.41}$$

the peak of the distribution calculated from the estimates of the true hidden coordinates.

## 2.5 Gaussian mixture model

We can improve upon the single Gaussian model for the intrinsic dispersion of the independent variable, $\mathbf{x_t}$, by employing a mixture of $K$ Gaussian functions. The basic idea is that with a large enough number of Gaussian functions we can more accurately approximate the dispersion, even though the individual Gaussians have no physical meaning. It is a common and well-studied model that allows flexibility when estimating a distribution and is referred to in the statistical literature as a "nonparametric" model.

$$p(x_{ti}|\mu_1,\tau_1,\pi_1,\cdots,\mu_K,\tau_K,\pi_K,I) = \sum_{i=1}^{K} \frac{\pi_k}{\sqrt{2\pi\tau_k^2}} \exp[-\frac{(x_{ti}-\mu_k)^2}{2\tau_k^2}], \qquad (2.42)$$

where $\pi_k$ is the weighting of the $k^{\text{th}}$ Gaussian such that $\sum_{k=1}^{K} \pi_k = 1$. The $\pi_k$ may be interpreted as the probability of drawing a data point $x_{ti}$ from the $k^{\text{th}}$ Gaussian function $\sim N(\mu_k, \tau_k^2)$. The mixture of Gaussians is also a conjugate distribution for the measurement errors and assumed regression relationship given in equation (2.17), thus simplifying the mathematics. To include the Gaussian mixture model [11], equations (2.24) to (2.26) needs to be replaced by

$$p(D|\alpha,\beta,\sigma,\psi,I) = \prod_{i=1}^{n}\sum_{k=1}^{K} \frac{\pi_k}{2\pi\sqrt{\det\mathbf{V}_{k,i}}} \exp\left[-\frac{1}{2}(z_i-\zeta_k)^T\mathbf{V}_{k,i}^{-1}(z_i-\zeta_k)\right], \qquad (2.43)$$

where $\psi$ is an abbreviation for the nuisance parameters $\mu_k, \tau_k, \pi_k$ for $k = 1$ to $K$, and

$$z_i = \begin{pmatrix} y_i \\ x_i \end{pmatrix}, \ \zeta_k = \begin{pmatrix} \alpha+\beta\mu_k \\ \mu_k \end{pmatrix}, \ (z_i-\zeta_k) = \begin{pmatrix} y_i-\alpha-\beta\mu_k \\ x_i-\mu_k \end{pmatrix}, \qquad (2.44)$$

and

$$\mathbf{V}_{k,i} = \begin{pmatrix} \beta^2\tau_k^2+\sigma^2+\sigma_{y,i}^2 & \beta\tau_k^2 \\ \beta\tau_k^2 & \tau_k^2+\sigma_{x,i}^2 \end{pmatrix}. \qquad (2.45)$$

We need to specify priors for the nuisance parameters, $\pi_k$, $\mu_k$ and $\tau_k$. We adopt a Dirichlet [13] prior for $\pi_i$.

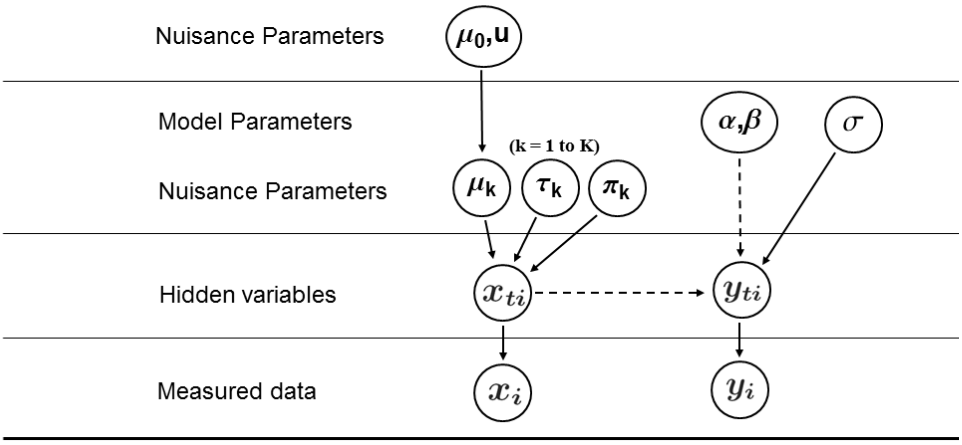
$$\pi_1,\pi_2,\cdots,\pi_k \sim \text{Dirichlet}[\gamma_1,\gamma_2,\cdots,\gamma_k], \text{ where } \gamma_i = 1 \text{ for all } i. \qquad (2.46)$$

[13] The Dirichlet distribution is a multivariate generalization of the $\beta$ distribution. The $\beta$ distribution is a conjugate prior for the binomial distribution while the Dirichlet distribution is a conjugate prior for the multinomial distribution.

This insures that $\sum_{i=1}^{K} \pi_i = 1$. The expectation value of $\pi_i$ is given by

$$\langle \pi_i \rangle = \frac{\gamma_i}{\sum_{i=1}^{K} \gamma_i} = 1/K \tag{2.47}$$

It is often sufficient to use only two Gaussian components in the mixture model, in which case $\pi_2 = 1 - \pi_1$.



Conditional probabilities

$$p(\mu_k|\mu_0, u, I) \sim N(\mu_0, u^2);\;\; p(\tau_k|I) = \frac{(\tau_k+\tau_0)^{-1}}{\ln(1+\frac{\tau_{max}}{\tau_0})}$$

$$p(\pi_k|I) \sim \mathbf{Dirichlet}(1, 1, \cdots, 1);\;\; p(u|I) = \frac{(u+u_0)^{-1}}{\ln(1+\frac{u_{max}}{u_0})}$$

$$p(x_{ti}|\mu_1, \tau_1, \cdots, \mu_K, \tau_K, I) = \sum_{i=1}^{K} \frac{\pi_k}{\sqrt{2\pi\tau_k^2}} \exp[-\frac{(x_{ti}-\mu_k)^2}{2\tau_k^2}]$$

$$p(y_{ti}|x_{ti}, \alpha, \beta, \sigma, I) \sim N(\alpha + \beta\, x_{ti}, \sigma^2)$$

$$p(x_i|x_{ti}, I) \sim N(x_{ti}, \sigma^2_{x,i});\;\; p(y_i|y_{ti}, I) \sim N(y_{ti}, \sigma^2_{y,i})$$

Figure 2.10 Graphical model of the multilevel Bayes calculation for a Gaussian mixture model. The solid lines connecting nodes denote conditional dependencies with the conditional probabilities listed below. The absence of a connection denotes conditional independence. The dashed lines represent deterministic conditionals.

We adopt a Gaussian prior for the individual $\mu_k$, with a common mean $\mu_0$ and variance $u^2$. This reflects a prior belief that the distribution of $\mathbf{x_t}$ is likely to be fairly unimodal, i.e., the individual Gaussian functions are more likely to be close together than far apart. Thus

$$\mu_k \sim N(\mu_0, u^2). \tag{2.48}$$

We will adopt a normalized modified scale invariant prior for $u$ according to

$$p(u|I) = \frac{(u + u_0)^{-1}}{\ln(1 + \frac{u_{max}}{u_0})}. \tag{2.49}$$

Based on prior knowledge of the measurement apparatus, the break point $u_0$ was set = Mean[$\sigma_{x,i}$]. For $u < u_0$, $p(u|I)$ behaves like a flat prior and for $u > u_0$ behaves like a scale invariant prior.

Figure 2.10 illustrates graphically the multilevel Bayes components of the current Bayesian calculation [14].

### *2.5.1 Example*

A two Gaussian mixture model was employed to fit a simple regression line data set, consisting of 22 points, that is given in Table 2.3 and described in more detail in example 2 of Section 2.7.4. In this case the model parameters are $\alpha, \beta, \sigma, \pi_1, \mu_1, \tau_1, \mu_2, \tau_2, \mu_0, u$. It is convenient to refer to Figure 2.10. Here $\pi_1$ is the weighting of the first Gaussian and since there are only two, $\pi_2 = 1 - \pi_1$ so $\pi_2$ is not required as a separate parameter. We assumed flat to scale invariant (FTSI) priors for the $\sigma, \tau_1, \tau_2, u$ as described in Section 2.3.2, and flat priors for the remaining parameters. The results are shown in Figure 2.11.

The marginal distributions for the two pairs of $\mu$ parameters (dashed and solid), and $\tau$ parameters are shown in panels (a) and (b) of Figure 2.12. Since both $\mu_1$ and $\mu_2$ are free to explore the entire prior range, their distributions are degenerate leading to similar multi-peak marginals as expected. The same is true of $\tau_1$ and $\tau_2$. We can re-label these subscripts for the parameter vectors by organizing them such that $\pi_1 < \pi_2$. This yields panels (c) and (d) in Figure 2.12 in which the lower weight distributions are dashed. These latter distributions are more distinctive and it is clear that the lower weight Gaussian allows for a range of narrower peaks that supplement a better defined higher weight broad Gaussian.

The same data set was re-analyzed using a single Gaussian model and yielded almost identical results to those shown in Figure 2.11. Thus the current data set does not justify a multiple Gaussian mixture model.

[14] According to Loredo (2013), “The impact of the graph structure on a model’s predictive ability becomes less intuitively accessible as complexity grows, making predictive tests of MLMs important, but also nontrivial; simple posterior predictive tests may be insensitive to significant discrepancies. An exemplary feature of the SNe Ia MLM work of Mandel et al. (2011) is the use of careful predictive checks, implemented via a frequentist cross-validation procedure, to quantitatively assess the adequacy of various aspects of the model (see Carroll et al. 2006 for intro to this topic).”

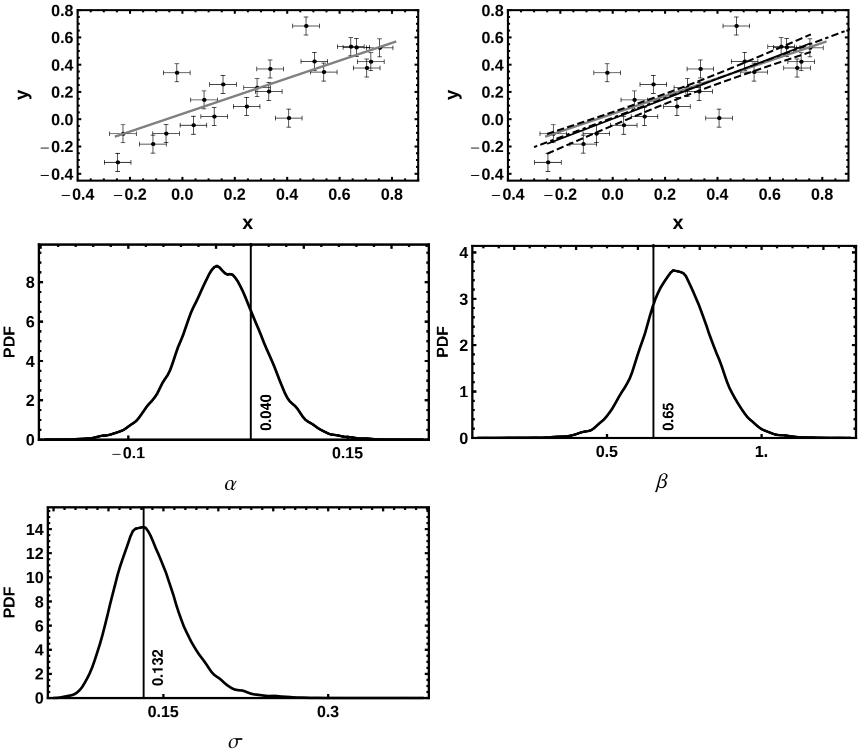


Figure 2.11 The top left panel shows the true regression line and the simulated data. The right panel shows overlaid the FMCMC mean regression line fit (black solid line) and the mean ±1$\sigma$ fit uncertainty (dashed curves) compared to the true regression line (gray). The following 3 panels compares the posterior marginal distributions for the regression line parameters $\alpha, \beta, \sigma$ to their true values.

## 2.6 Regression with multiple independent variables

The results of Section 2.3 and 2.5 assume a single independent variable. In many situations a number of different factors significantly influence the dependent variable. For example income may depend on education, sex and age. In astronomy the X-ray luminosity of a galaxy might be related to the optical luminosity and redshift. We can readily extend the formalism developed to $p$ independent variable (following Kelly 2007) [11]. Equation (2.17) becomes

$$y_{ti} = \alpha + \boldsymbol{\beta}^T \, \mathbf{x}_{ti} + \epsilon, \tag{2.50}$$

where $\boldsymbol{\beta}$ is now a $p$-element column vector and $\mathbf{x}_{ti}$ is a a $p$-element column vector containing the true values of the independent variables for the $i$th data point. For

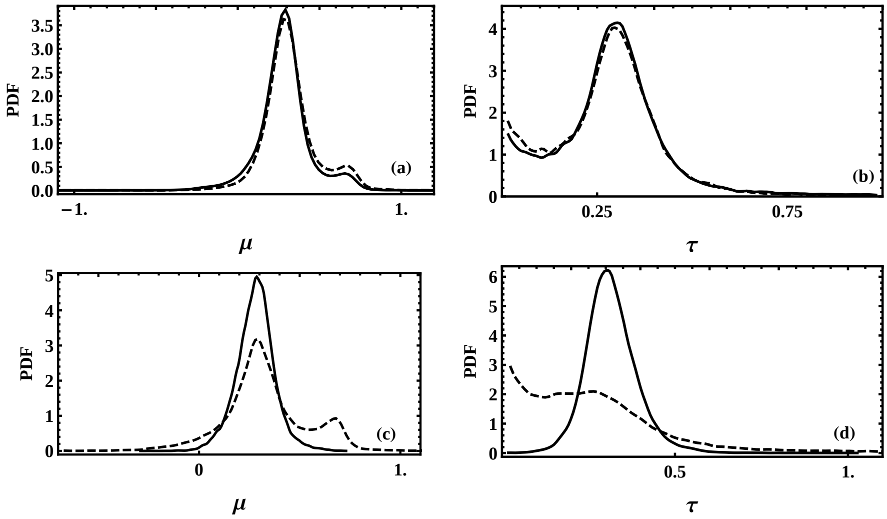


Figure 2.12 The marginal distributions for the two pairs of $\mu$ parameters (solid and dashed curves), and the two $\tau$ parameters are shown in panels (a) and (b). Panels (c) and (d) show the re-labeled $\mu$ and $\tau$ distributions (see text).

example, for $p = 2$, equation (2.50) becomes

$$y_{ti} = \alpha + \beta_1\ x_{ti,1} + \beta_2\ x_{ti,2} + \epsilon. \tag{2.51}$$

If we use a Gaussian mixture model to approximate the intrinsic dispersion of the independent variables then the $\mathbf{x}_{ti}$ is given by $K$ multivariate normal densities with $p$-element mean vectors $\boldsymbol{\mu}_k$, $p \times p$ covariance matrices $\mathbf{T}_k$, and weights $\pi_k$. Thus

$$p(\mathbf{x}_{ti}|\psi, I) = \sum_{k=1}^{K} \pi_k N_p(\boldsymbol{\mu}_k, \mathbf{T}_k), \tag{2.52}$$

where $\psi$ is an abbreviation for the nuisance parameters $\pi_k, \mu_k, \mathbf{T}_k$ for $k = 1$ to $K$. We also use the symbol $\theta$ as an abbreviation for the regression parameters $\alpha, \boldsymbol{\beta}, \sigma$. The $\mathbf{T}_k$ matrix is

$$\mathbf{T}_k = \begin{pmatrix} T_{k11} & T_{k12} & \cdot & T_{k1p} \\ T_{k21} & T_{k22} & \cdot & T_{k2p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ T_{kp1} & \cdot & \cdot & T_{kpp} \end{pmatrix} = \begin{pmatrix} \tau_{k1}^2 & T_{k12} & \cdot & T_{k1p} \\ T_{k21} & \tau_{k2}^2 & \cdot & T_{k2p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ T_{kp1} & \cdot & \cdot & \tau_{kp}^2 \end{pmatrix}, \tag{2.53}$$

The measured value of $\mathbf{x}_{ti}$ is the $p$-element vector $\mathbf{x}_i$. As before the combination of the measured dependent variable, $y_i$, and independent variable vector, $\mathbf{x}_i$ is represented by $\mathbf{z}_i$, a $p+1$ element column vector.

$$\mathbf{z}_i = \begin{pmatrix} y_i \\ x_{1i} \\ . \\ x_{(p+1)i} \end{pmatrix}, \tag{2.54}$$

At this point we can also allow for the possibility of correlations in the measurement errors by employing a $(p+1) \times (p+1)$ covariance matrix $\mathbf{\Sigma}_i$. Thus

$$p(y_i, \mathbf{x}_i | y_{ti}, \mathbf{x}_{ti}, I) = N_{p+1}([y_{ti}, \mathbf{x}_{ti}] \, |\mathbf{\Sigma}_i). \tag{2.55}$$

If the measurements are uncorrelated then $\mathbf{\Sigma}_i$ is a diagonal matrix.

The likelihood is given by

$$p(D|\theta, \psi, I) = \prod_{i=1}^{n} \sum_{k=1}^{K} \frac{\pi_k}{(2\pi)^{p+1} \sqrt{\det \mathbf{V}_{k,i}}} \exp\left[ -\frac{1}{2} (\mathbf{z}_i - \zeta_k)^T \mathbf{V}_{k,i}^{-1} (\mathbf{z}_i - \zeta_k) \right], \tag{2.56}$$

where

$$\zeta_k = \begin{pmatrix} \alpha + \boldsymbol{\beta}^T \boldsymbol{\mu}_k \\ \mu_1 \\ . \\ \mu_{(p+1)i} \end{pmatrix}, \quad (\mathbf{z}_i - \zeta_k) = \begin{pmatrix} y_i - \alpha - \boldsymbol{\beta}^T \boldsymbol{\mu}_k \\ x_{1i} - \mu_1 \\ . \\ x_{(p+1)i} - \mu_{(p+1)i} \end{pmatrix}, \tag{2.57}$$

and

$$\mathbf{V}_{k,i} = \left( \begin{array}{c|c} \boldsymbol{\beta}^T \mathbf{T}_k \boldsymbol{\beta} + \sigma^2 + \sigma_{y,i}^2 & \beta^T \mathbf{T}_k + \boldsymbol{\sigma}_{\mathbf{x}y,i}^T \\ \hline \mathbf{T}_k \beta + \boldsymbol{\sigma}_{\mathbf{x}y,i} & \mathbf{T}_k + \mathbf{\Sigma}_{\mathbf{x},i} \end{array} \right). \tag{2.58}$$

$\zeta_k$ is the mean vector of $\mathbf{z}_i$, $\mathbf{V}_{k,i}$ is the $(p+1) \times (p+1)$ covariance matrix of $\mathbf{z}_i$ for Gaussian function $k$, $\sigma_{y,i}^2$ is the variance of the measurement error for $y_i$, $\boldsymbol{\sigma}_{\mathbf{x}y,i}$ is the $p$-element vector of covariances between the measurement errors on $y_i$ and $\mathbf{x}_i$, and $\mathbf{\Sigma}_{x,i}$ is the $p \times p$ covariance matrix of measurement errors on $\mathbf{x}_i$.

### 2.6.1 Example: two independent variables

Consider the special case of $p = 2$, for two independent variables. We start with $\mathbf{\Sigma}_i$

$$\mathbf{\Sigma}_i = \left( \begin{array}{c|cc} \sigma_{y,i}^2 & \sigma_{x_1 y,i} & \sigma_{x_2 y,i} \\ \hline \sigma_{x_1 y,i} & \sigma_{x_1,i}^2 & \sigma_{x_1 x_2,i} \\ \sigma_{x_2 y,i} & \sigma_{x_1 x_2,i} & \sigma_{x_2,i}^2 \end{array} \right). \tag{2.59}$$

This can be factored as

$$\mathbf{\Sigma}_i = \left( \begin{array}{c|c} \sigma_{y,i}^2 & \boldsymbol{\sigma}_{\mathbf{x}y,i}^T \\ \hline \boldsymbol{\sigma}_{\mathbf{x}y,i} & \mathbf{\Sigma}_{\mathbf{x},i} \end{array} \right). \tag{2.60}$$

Of course, if the covariances between the measurement errors are all zero then Equation (2.59)

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma^2_{y,i} & 0 & 0 \\ 0 & \sigma^2_{x_1,i} & 0 \\ 0 & 0 & \sigma^2_{x_2,i} \end{pmatrix}. \tag{2.61}$$

The $\mathbf{T}_k$ matrix in equation (2.53) becomes

$$\mathbf{T}_k = \begin{pmatrix} \tau^2_{k1} & T_{k12} \\ T_{k12} & \tau^2_{k2} \end{pmatrix}, \tag{2.62}$$

and $\mathbf{V}_{k,i}$ from equation (2.58) becomes

$$\begin{pmatrix} \beta_1^2\tau^2_{k1} + \beta_2^2\tau^2_{k2} + \sigma^2 + \sigma^2_{y,i} & \beta_1\tau^2_{k1} + \beta_2 T_{k12} + \sigma_{x_1y,i} & \beta_1 T_{k12} + \beta_2\tau^2_{k2} + \sigma_{x_2y,i} \\ \beta_1\tau^2_{k1} + \beta_2 T_{k12} + \sigma_{x_1y,i} & \tau^2_{k1} + \sigma^2_{x1,i} & T_{k12} + \sigma_{x_1x_2,i} \\ \beta_1 T_{k12} + \beta_{k2}\tau^2_{k2} + \sigma_{x_2y,i} & T_{k12} + \sigma_{x_1x_2,i} & \tau^2_{k2} + \sigma^2_{x2,i} \end{pmatrix} \tag{2.63}$$

If the covariances between the measurement errors are all zero as well as the off diagonal elements of $\mathbf{T}_k$, then $\mathbf{V}_{k,i}$ simplifies to

$$\mathbf{V}_{k,i} = \begin{pmatrix} \beta_1^2\tau^2_{k1} + \beta_2^2\tau^2_{k2} + \sigma^2 + \sigma^2_{y,i} & \beta_1\tau^2_{k1} & \beta_2\tau^2_{k2} \\ \beta_1\tau^2_{k1} & \tau^2_{k1} + \sigma^2_{x1,i} & 0 \\ \beta_2\tau^2_{k2} & 0 & \tau^2_{k2} + \sigma^2_{x2,i} \end{pmatrix}. \tag{2.64}$$

## 2.7 Selection effects

Selection effects play an important role in many observations and measurements. For example instrumental detection systems often result in an upper and/or lower limit on what can be detected. This could result in a biased estimate of the regression parameters. Here we explore how the combination of selection effects and measurement errors affect our regression results. Suppose one collects a sample of $n$ sources out of a possible sample of $N$ sources where $N$ itself is unknown parameter in the analysis. For example, in astronomy one often performs a flux-limited survey of some sample area of the sky containing $N$ sources where $N$ is an unknown. Because of selection effects and noise we only detect $n$ which yield the observed/measured values. We would like to know what effect the missing sources would have on the regression. Let's examine how this can be achieved in the Bayesian framework (following Kelly 2007, Kelly & Fan, 2008) [11] [12] .

In Section 2.3, $\mathbf{x}, \mathbf{y}$ denoted the known observed/measured independent and dependent variables and $\mathbf{x_t}, \mathbf{y_t}$, the true but unknown values of these variables. The $\mathbf{x_t}, \mathbf{y_t}$ values were referred to as hidden variables. Here we need to introduce the idea of missing data and extend the conversation to include the missing data just as we did for the hidden variables. We first extend the meaning of $\mathbf{x}, \mathbf{y}$ to include both the $n$ observed data and the $N - n$ missing data, e.g., $\mathbf{x} = (\mathbf{x_{obs}}, \mathbf{x_{mis}})$. Similarly, we will extend the meaning of $\mathbf{x_t}, \mathbf{y_t}$ to include true coordinates of both the $n$ observed data and the $N - n$ the missing data, e.g., $\mathbf{x_t} = (\mathbf{x_{t_{obs}}}, \mathbf{x_{t_{mis}}})$. We can incorporate the sample selection into the likelihood by introducing a new variable for each source called an indicator variable $q_i$. For anyone of the $n$ detected sources $q_i = 1$, and $q_i = 0$ for any of the $N - n$ missing sources. Let $\mathbf{q}$ represent a vector of all $N$ values of $q_i$. In this analysis we will assume that the selection function, $p(\mathbf{q}|\mathbf{x}, \mathbf{y}, I)$, depends only on the measured quantities, $\mathbf{x}$ and $\mathbf{y}$.

To make the length of the equations more manageable we will let $\theta$ represent the linear regression variables $\alpha, \beta, \sigma$ and $\psi$ represent all the nuisance variables needed to define the probability of the hidden variables, $\mathbf{x_t}$. In the example shown in Figure 2.10, $\psi$ included the variables $\mu_0, u, \mu_k, \tau_k, \pi_k$. As usual, $I$ represents our prior information that includes knowledge of $\sigma_{xi}^2, \sigma_{yi}^2$, the known variances of the Gaussian measurement errors. $I$ also include any prior knowledge of the selection effect(s).

The complete likelihood of the observed data, hidden variables, missing data, and $N$ is given by

$$p(\mathbf{x}, \mathbf{y}, \mathbf{x_t}, \mathbf{y_t}, \mathbf{q}, N|\theta, \psi, I) = p(N|I)\ p(\mathbf{q}|\mathbf{x}, \mathbf{y}, I)\ p(\mathbf{x}, \mathbf{y}|\mathbf{x_t}, \mathbf{y_t}, I) \\ \times p(\mathbf{y_t}|\mathbf{x_t}, \theta, I)\ p(\mathbf{x_t}|\psi, N, I). \tag{2.65}$$

After integrating over the hidden variables (we saw how in Section 2.3) equation (2.65) becomes

$$p(\mathbf{x}, \mathbf{y}, \mathbf{q}, N|\theta, \psi, I) = p(N|I)\ p(\mathbf{x}, \mathbf{y}, \mathbf{q}|\theta, \psi, N, I). \tag{2.66}$$

Factor $p(\mathbf{x}, \mathbf{y}, \mathbf{q}|\theta, \psi, N, I)$, the second term on the right of equation (2.66), using the product rule and remove redundant conditionals.

$$p(\mathbf{x}, \mathbf{y}, \mathbf{q}|\theta, \psi, N, I) = p(\mathbf{q}|\mathbf{x}, \mathbf{y}, I)\ p(\mathbf{x}, \mathbf{y}|\theta, \psi, N, I). \tag{2.67}$$

We can marginalize over $\mathbf{q}$ in equation (2.67) as follows

$$p(\mathbf{x}, \mathbf{y}|\theta, \psi, N, I) = \sum_{\mathbf{q}} p(\mathbf{q}|\mathbf{x}, \mathbf{y}, I)\ p(\mathbf{x}, \mathbf{y}|\theta, \psi, N, I), \tag{2.68}$$

where the sum is over all possible configurations of the vector $\mathbf{q}$. Since $p(\mathbf{x}, \mathbf{y}|\theta, \psi, N, I)$ is conditional on $N$, each possible configuration of the $\mathbf{q}$ vector contains $n$ values with $q = 1$ and $N - n$ values of $q = 0$. The total number of these configurations is given by the binomial coefficient $C_n^N = N!/n!(N - n)!$. In component form equation (2.68) becomes

$$p(\mathbf{x},\mathbf{y}|\theta,\psi,N,I) = C_n^N \prod_{(i=1)_{obs}}^{n} p(q_i = 1|x_i,y_i,I)\; p(x_i,y_i|\theta,\psi,I)$$
$$\times \prod_{(j=1)_{mis}}^{N-n} p(q_i = 0|x_j,y_j,I) \times p(x_j,y_j|\theta,\psi,I), \tag{2.69}$$

In general, $N$ is an unknown parameter in the analysis which we will will eventually marginalize over. To do this we will have treat $N$ as the random variable and consider the number of sources detected, $n$, as given. The probability of N is described by a negative binomial distribution which we will get to shortly.

Now integrate this equation over $d\mathbf{x_{mis}}$ and $d\mathbf{y_{mis}}$, the missing data for which $q = 0$, to obtain the observed likelihood $p(\mathbf{x_{obs}},\mathbf{y_{obs}}|\theta,\psi,N,I)$ conditional on $N$. Recall that $\mathbf{x} = \mathbf{x_{obs}},\mathbf{x_{mis}}$ and $\mathbf{y} = \mathbf{y_{obs}},\mathbf{y_{mis}}$.

$$p(\mathbf{x_{obs}},\mathbf{y_{obs}}|\theta,\psi,N,I) = C_n^N \prod_{(i=1)_{obs}}^{n} p(q_i = 1|x_i,y_i,I)\; p(x_i,y_i|\theta,\psi,I)$$
$$\times \prod_{(j=1)_{mis}}^{N-n} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(q_i = 0|x_j,y_j,I)$$
$$\times\; p(x_j,y_j|\theta,\psi,I)dx_{mis_j}\, dy_{mis_j}, \tag{2.70}$$

As a next step, rewrite the integral in equation (2.70).

$$\prod_{(j=1)_{mis}}^{N-n} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(q_j = 0|x_j,y_j,I)\; p(x_j,y_j|\theta,\psi,I)dx_{mis_j}\, dy_{mis_j}$$
$$= \left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(q = 0|x,y,I)\; p(x,y|\theta,\psi,I)dx\, dy\right]^{N-n}$$
$$= \left[p(q = 0|\theta,\psi,I)\right]^{N-n}. \tag{2.71}$$

The quantity $p(q = 0|\theta,\psi,I)$ is the weighted average probability (over all $x, y$) of not selecting a source for a particular choice of $\theta,\psi$. The weighting function is $p(x,y|\theta,\psi,I)$. Also

$$p(q = 0|\theta,\psi,I) = 1 - p(q = 1|\theta,\psi,I), \tag{2.72}$$

where $p(q = 1|\theta,\psi,I)$ = the probability of selecting a source.

Using equation (2.71), we can rewrite equation (2.70) as

$$p(\mathbf{x_{obs}},\mathbf{y_{obs}}|\theta,\psi,N,I) = C_n^N \left[p(q = 0|\theta,\psi,I)\right]^{N-n}$$

$$\times \prod_{(i=1)_{obs}}^{n} p(q_i = 1|x_i, y_i, I)\ p(x_i, y_i|\theta, \psi, I) \tag{2.73}$$

We can simplify the term involving the product over the observed sources in equation (2.73).

$$\prod_{(i=1)_{obs}}^{n} p(q_i = 1|x_i, y_i, I)\ p(x_i, y_i|\theta, \psi, I) = \prod_{(i=1)_{obs}}^{n} p(q_i = 1|x_i, y_i, I) \times \prod_{(i=1)_{obs}}^{n} p(x_i, y_i|\theta, \psi, I. \tag{2.74}$$

The first product on the RHS does not depend on the regression parameters and is just a constant which can be dropped.

We are now ready to marginalize over $N$. First rewrite equation (2.66) for the subset of observed quantities now that we have marginalized over $q$ and the missing data.

$$p(\mathbf{x_{obs}}, \mathbf{y_{obs}}, N|\theta, \psi, I) = p(N|I)\ p(\mathbf{x_{obs}}, \mathbf{y_{obs}}|\theta, \psi, N, I). \tag{2.75}$$

To marginalize over $N$, we sum over $p(\mathbf{x_{obs}}, \mathbf{y_{obs}}, N|\theta, \psi, I)$ from $N = n$ to $\infty$, after specifying a suitable prior for $N$. We will assume a scale invariant prior for $N$ so that $p(N|I) \propto 1/N$.

$$p(\mathbf{x_{obs}}, \mathbf{y_{obs}}|\theta, \psi, I) \propto \mathcal{A} \times \sum_{N=n}^{N} \frac{1}{N} C_n^N\ \left[p(q = 0|\theta, \psi, I)\right]^{N-n}, \tag{2.76}$$

where

$$\mathcal{A} = \prod_{(i=1)_{obs}}^{n} p(x_i, y_i|\theta, \psi, I) \tag{2.77}$$

Now

$$\frac{1}{N} C_n^N = \frac{1}{N} C_{n-1}^{N-1} \frac{N}{n} = \frac{1}{n} C_{n-1}^{N-1}. \tag{2.78}$$

Substitute for $\frac{1}{N} C_n^N$ in equation (2.76) and multiply by

$$\left[p(q = 1|\theta, \psi, I)\right]^n \times \left[p(q = 1|\theta, \psi, I)\right]^{-n}$$

The RHS of equation (2.76) can be rewritten as

$$\{\sum_{N=n}^{\infty} C_{n-1}^{N-1} \left[p(q = 1|\theta, \psi, I)\right]^n \left[p(q = 0|\theta, \psi, I)\right]^{N-n}\} \times \mathcal{A}\, \frac{1}{n} \left[p(q = 1|\theta, \psi, I)\right]^{-n}. \tag{2.79}$$

Now the terms within the $\{\cdots\}$ brackets is the sum of a negative binomial distribution, so

$$\{\sum_{N=n}^{\infty} C_{n-1}^{N-1} \left[p(q = 1|\theta, \psi, I)\right]^{n} \left[p(q = 0|\theta, \psi, I)\right]^{N-n}\} = 1. \tag{2.80}$$

Within the context of this analysis, the negative binomial distribution gives the probability that the total number of sources is $N$, given that we have selected $n$ sources where the probability of selecting a single source is given by $p(q = 1|\theta, \psi, I)$. As a reminder, $p(q = 1|\theta, \psi, I)$ is the weighted average probability (over all $x, y$) of selecting a single source for a particular choice of $\theta, \psi$. The weighting function is $p(x, y|\theta, \psi, I)$.

$$p(q = 1|\theta, \psi, I) = \int\int p(q = 1|x, y, I)\, p(x, y|\theta, \psi, I) dx\, dy. \tag{2.81}$$

As a consequence equation (2.80), and ignoring the constant $1/n$ factor, equation (2.76) simplifies to

$$\begin{aligned} p(\mathbf{x_{obs}}, \mathbf{y_{obs}}|\theta, \psi, I) \propto & \left[p(q = 1|\theta, \psi, I)\right]^{-n} \\ & \times \prod_{(i=1)_{obs}}^{n} p(x_i, y_i|\theta, \psi, I), \end{aligned} \tag{2.82}$$

where $\prod_{(i=1)_{obs}}^{n} p(x_i, y_i|\theta, \psi, I) = p(D|\theta, \psi, I)$. For a single Gaussian model for the intrinsic dispersion of $\mathbf{x_t}$, $p(D|\theta, \psi, I)$ is given by equation (2.24). If we represent the intrinsic dispersion of $\mathbf{x_t}$ by a mixture of $K$ Gaussian functions then $p(D|\theta, \psi, I)$ is given by equation (2.43).

### *2.7.1 Selection on measured independent variables*

Our previous results pertain to the general case where the selection is based on measured dependent and independent variables. If the sample is selected based on the measured independent variables ($p(q|x, y, I) = p(q|x, I)$) and the measurement errors for $x$ and $y$ are independent, then inference on the regression parameters $\theta$ is unaffected by selection effects because $x$ is independent of $\theta$.

### *2.7.2 Selection on measured dependent variable*

*Sharp lower cutoff in measured dependent variable*

In this case suppose there is a sharp lower cutoff, $y_{cut}$, so $p(q_i = 1|y_i, I) = 1$ for $y_i \geq y_{cut}$ and $p(q_i = 1|y_i, I) = 0$ for $y_i < y_{cut}$. The argument of the first term in equation( 2.82) is $\left[p(q = 1|\theta, \psi, I)\right]^{-n}$, the single source selection probability averaged over all $x, y$, for a give choice of regression parameters $\theta$. For some values of $\theta$ the selection probability will be higher than for other values. By depending inversely on the selection probability, this term corrects for this effect.

A common selection effect that arises in astronomy is referred to as the Malmquist bias after Swedish astronomer Gunnar Malmquist (1893 to 1982) who discussed it in the 1920's. For a good discussion of the topic see Wall & Jenkins (2012) [17]. In a flux-limited sample there is a built in luminosity versus distance correlation because brighter objects can be detected at greater distances [15]. The threshold flux or flux density limit is usually set at a level of 3 to 5 times the measurement uncertainty in the dependent coordinate, designated $s_{\mathrm{lim}}$ where $s$ stands for flux density.

This same measurement uncertainty introduces another effect [16] that competes with the Malmquist bias because in general there are many more low luminosity objects than high luminosity objects - the number of objects as a function of the observed flux density $N(s)$ (referred to as source counts) usually rises steeply to small values of $s$. In a flux-limited survey an object is only considered detected if the flux exceed the lower limit $s_{\mathrm{lim}}$. Since the selection function depends on the measured flux density, $s_i$, some objects with true $s > s_{\mathrm{lim}}$ can be missed in the survey because of a negative noise excursion while others with a true $s < s_{\mathrm{lim}}$ can be detected because of a positive noise excursion. Since there are more low luminosity sources than high luminosity sources this can bias the faint end of the catalog towards lower luminosity objects [17].

In the following section we analyze in detail a simple problem for which the selection function is a smooth function of the measured dependent variable instead of a sharp cutoff.

### 2.7.3 Example 1: gradual cutoff in measured y

Panel (a) of Figure 2.13 shows the true regression line together with the full simulated data set after adding intrinsic scatter and measurement errors in both coordinates. The full data set is given in Table 2.2. The error bars indicate the $1\sigma$ IID measurement errors. Panel (b) shows the selection function $p(q = 1|y, I)$ which depends only on the measured value of the dependent variable $y$. The selection function employed is a cumulative normal distribution given by

$$p(q = 1|y, I) = \frac{1}{\sqrt{2\pi}\sigma_s} \int_{-\infty}^{y} \exp\left[-\frac{(t-\phi)^2}{2\sigma_s^2}\right] dt$$

[15] This also means that two statistically independent luminosities (e.g., X-ray and optical) may appear to correlate because of their mutual correlation with distance.

[16] Sometimes this effect is also confusingly referred to by the generic term Malmquist bias.

[17] This effect does not occur if the observational error is a constant fraction of the flux density, and the source counts are close to a power law [17].

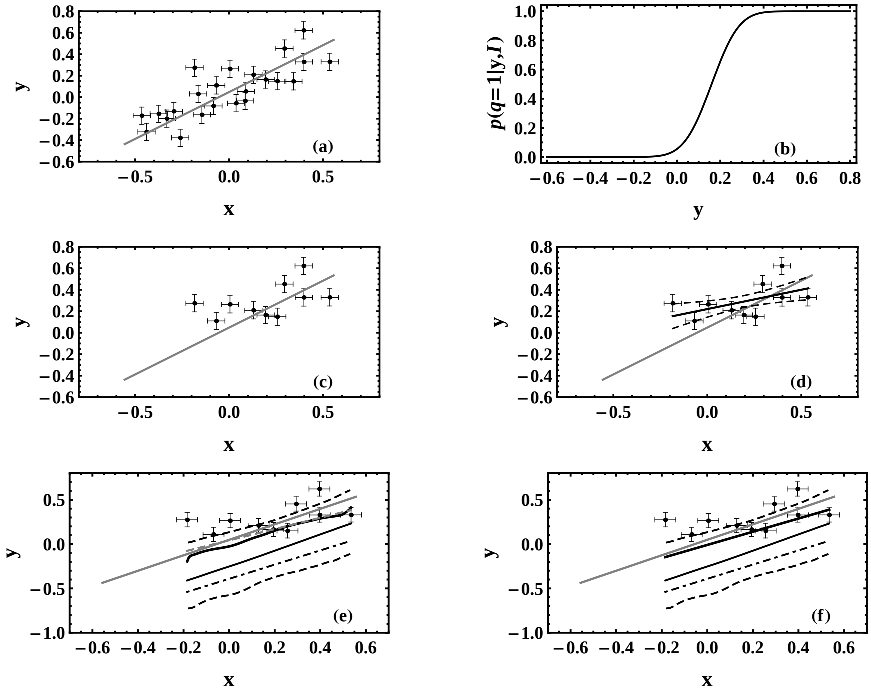


Figure 2.13 Panel (a) shows the simulated straight line together with the full simulated data set after adding intrinsic scatter and measurement errors in both coordinates. The error bars indicate the $1\sigma$ IID measurement errors. Panel (b) shows the selection function $p(q = 1|y, I)$. Panel (c) shows the same simulated straight line with only the selected data set. Panel (d) shows the FMCMC mean regression line fit (black solid line) to the selected data set, ignoring the selection function, and the mean $\pm 1\sigma$ fit uncertainty (dashed curves) compared to the simulated line (gray). Panel (e) is the same as panel (c) with four additions. The thick solid black wiggly curve is the mode of the FMCMC regression line fit distribution. The solid black line is the median of the regression line fit distribution. The dot-dashed black line is the mean of the regression line fit distribution. The dashed gray line is the MAP fit. The upper and lower dashed curves are the 68% credible region boundaries of the FMCMC regression line fit distribution. In panel (f), the mode of the fit distribution (wiggly curve in panel (e)) has been replaced by the best fitting straight line.

$$= \frac{1}{2}(1 + \mathrm{erf}[\frac{y - \phi}{\sqrt{2}\sigma_s}]), \tag{2.83}$$

where the break parameter $\phi = 0.16$ and $\sigma_s = 0.1$. The selected data is indicated in the third column of Table 2.2 by a “Yes.” The error bars indicate the $1\sigma$ IID measurement errors which are equal to $\sigma_x = 0.046$ and $\sigma_y = 0.080$. The true

| x | y | Selected |
|---|---|---|
| -0.4649 | -0.1723 | No |
| -0.3748 | -0.1549 | No |
| -0.4397 | -0.3229 | No |
| -0.3315 | -0.1996 | No |
| -0.2944 | -0.1313 | No |
| -0.2604 | -0.3780 | No |
| -0.1842 | 0.2741 | Yes |
| -0.1451 | -0.1637 | No |
| -0.1651 | 0.0305 | No |
| -0.0836 | -0.0816 | No |
| -0.0685 | 0.1097 | Yes |
| 0.0044 | 0.2641 | Yes |
| 0.0368 | -0.0563 | No |
| 0.0849 | -0.0343 | No |
| 0.0886 | 0.0538 | No |
| 0.1293 | 0.2086 | Yes |
| 0.1941 | 0.1645 | Yes |
| 0.2561 | 0.1490 | Yes |
| 0.3419 | 0.1475 | No |
| 0.2941 | 0.4528 | Yes |
| 0.3963 | 0.6220 | Yes |
| 0.3976 | 0.3283 | Yes |
| 0.5351 | 0.3292 | Yes |

Table 2.2 *The table contains 23 pairs of measured $x, y$ values.*

regression line equation is

$$y_{ti} = 0.0481 + 0.8730\, x_{ti} + \epsilon_i, \tag{2.84}$$

where $\epsilon_i \sim N(0, \sigma^2)$ and $\sigma = 0.16$. Panel (c) of Figure 2.13 shows the true regression line with only the selected points.

Our starting point is the joint posterior distribution for $\theta, \psi$.

$$\begin{aligned} p(\theta, \psi|D, I) &\propto p(\theta, \psi|I)\; p(D|\theta, \psi, I) \\ &= p(\theta, \psi|I)\; p(\mathbf{x_{obs}}, \mathbf{y_{obs}}|\theta, \psi, I), \end{aligned} \tag{2.85}$$

where

$$\begin{aligned} p(\mathbf{x_{obs}}, \mathbf{y_{obs}}|\theta, \psi, I) \propto &\left[p(q = 1|\theta, \psi, I)\right]^{-n} \\ &\times \prod_{(i=1)_{obs}}^{n} p(x_i, y_i|\theta, \psi, I). \end{aligned} \tag{2.86}$$

where $\prod_{(i=1)_{obs}}^{n} p(x_i, y_i|\theta, \psi, I)$ is given by equation (2.24) together with equa-

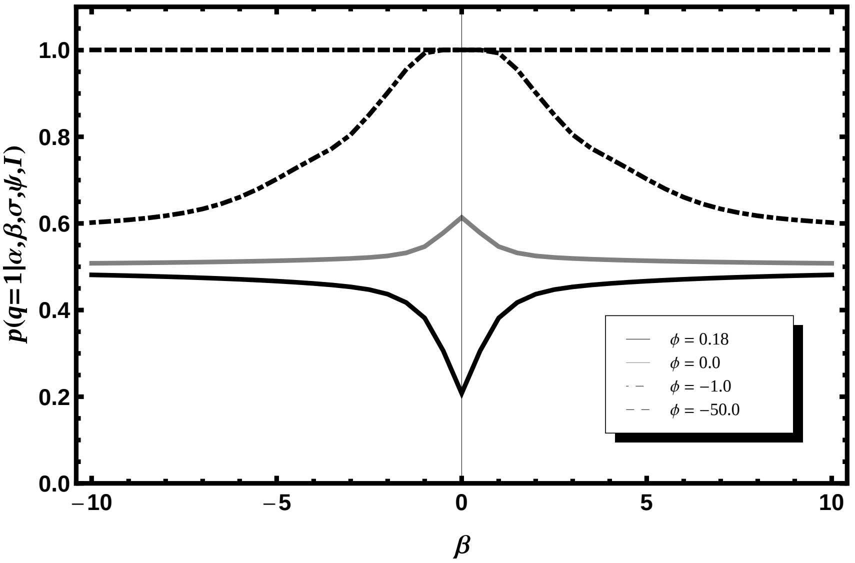


Figure 2.14 The dependence of the selection function on the regression slope parameter $\beta$ for different choices of the selection function break parameter, $\phi$,

tions (2.25) and (2.26).

Recall that $\theta$ is short for the regression parameters $\alpha, \beta, \sigma$. In this example we employ a single Gaussian model for the intrinsic dispersion of the independent variable, $\mathbf{x_t}$, so $\psi = \mu, \tau$. The first term in equation (2.86) contains the quantity $p(q = 1|\theta, \psi, I) = p(q = 1|\alpha, \beta, \sigma, \psi, I)$ which is computed from equation (2.81). It is the weighted average probability (over all $x, y$) of selecting a source for a particular choice of $\alpha, \beta, \sigma, \psi$. The weighting function is $p(x, y|\alpha, \beta, \sigma, \psi, I)$. Figure 2.14 shows some examples of the dependence of $p(q = 1|\alpha, \beta, \sigma, \psi, I)$ on the slope parameter, $\beta$, with the other parameters held constant, for several different choices of the selection function break parameter $\phi$. With $\phi = -50$, which is much less than the minimum measured $y$ value, $p(q = 1|\alpha, \beta, \sigma, \psi, I) = 1$ for all $\beta$ as would be expected. As $\phi$ increases the selection function shows a strong dependence on $\beta$.

A useful way of computing the marginal distribution for any of the parameters is with the fusion Markov chain Monte Carlo (FMCMC) method described in Chapter 1 of this supplement. We assumed flat priors for $\alpha, \beta, \mu$ and a FTSI (flat to scale invariant) prior for both $\tau$ and $\sigma$ of the form

$$p(\tau|I) = \frac{(\tau + \tau_0)^{-1}}{\ln(1 + \frac{\tau_{max}}{\tau_0})}. \tag{2.87}$$

The break point, $\tau_0$, was set equal to the characteristic measurement error in that $x$ coordinate (Mean[$\sigma_{x,i}$]), based on prior knowledge of the measurement apparatus. The break point, $\sigma_0$, was set = 2 × Mean[$\sigma_{y,i}$].

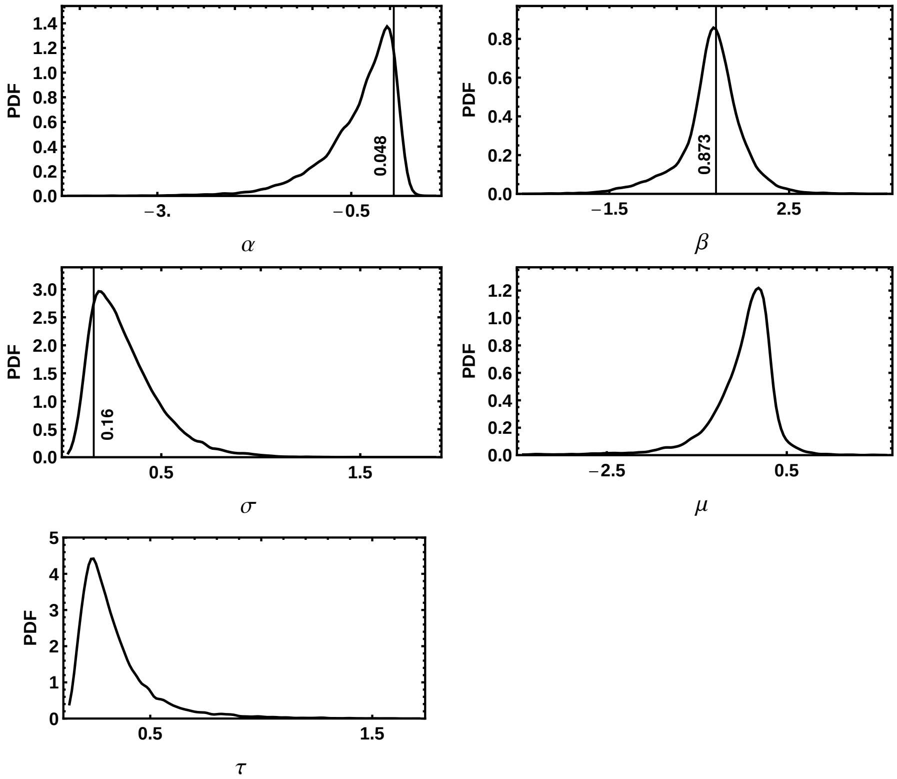


Figure 2.15 Posterior marginal distributions for the regression line parameters $\alpha, \beta, \sigma$ and nuisance parameters $\mu, \tau$.

Panel (d) of Figure 2.13 shows the FMCMC mean regression line fit (black solid line) to the selected data set, ignoring the selection function, and the mean $\pm 1\sigma$ fit uncertainty (dashed curves) compared to the true regression line (gray). The fit uncertainty is computed as follows. Each post burn-in FMCMC iteration yields an intercept and slope. We compute a set of model $y$ predictions for a uniform grid of $x$ values for that particular intercept and slope. This is repeated for each FMCMC iteration. At each $x$ grid point the mean and standard deviation of the corresponding y values are computed. The fit uncertainty curves are then plots of this grid of mean ± 1 standard deviation values.

Ignoring the selection effect, results in a good regression fit to the selected data which is much shallower than the true regression line. The maximum *a posterior* (MAP) fit for this case is identical to the mean fit within the line thickness.

In panel (e) we allow for the selection function in the analysis. Panel (e) is the same as panel (c) with four additions. The thick solid black wiggly curve is the mode of the FMCMC regression line fit distribution. The solid black line is the median of the regression line fit distribution. The dot-dashed black line is the mean of the regression line fit distribution. The dashed gray line is the MAP fit. The upper and lower dashed curves are the 68% credible region boundaries of the FMCMC regression line fit distribution. Clearly the regression line fit distribution is very asymmetic. In panel (f), the mode of the fit distribution (wiggly curve in panel (e)) has been replaced by best fitting straight line given by equation (2.88). This line is much closer to the true regression line than either the median or mean.

$$y = -0.0349 + 0.821\ x, \tag{2.88}$$

Figure 2.15 shows the marginal distributions for the regression line parameters $\alpha, \beta, \sigma$ and nuisance parameters $\mu, \tau$. The values used in the simulation for $\alpha, \beta, \sigma$, shown by the solid vertical lines, are in good agreement with the distributions.

### *2.7.4 Example 2: gradual cutoff in measured y*

The full data set for this example is given in Table 2.3 and the selected data is indicated in the third column by a “Yes.” Panel (a) of Figure 2.16 shows the true regression line together with the full simulated data set after adding intrinsic scatter and measurement errors in both coordinates. The error bars indicate the $1\sigma$ IID measurement errors which are equal to $\sigma_x = 0.051$ and $\sigma_y = 0.066$. The true regression line (gray line) equation is given by

$$y_{ti} = 0.0396 + 0.6502\ x_{ti} + \epsilon_i, \tag{2.89}$$

where $\epsilon_i \sim N(0, \sigma^2)$ and $\sigma = 0.132$. Panel (b) shows the selection function $p(q = 1|y, I)$ which depends only on the measured value of the dependent variable $y$. The selection function employed is the same cummulative normal distribution employed in the previous example, only the break parameter is $\phi = 0.18$. Panel (c) of Figure 2.16 shows the same simulated straight line with only the selected data set.

Panel (d) of Figure 2.13 shows the FMCMC mean regression line fit (black solid line) to the selected data set, ignoring the selection function, and the mean $\pm 1\sigma$ fit uncertainty (dashed curves) compared to the simulated line (gray). Ignoring the selection effect results in a good regression fit to the selected data which is much shallower than the true regression line. The maximum *a posterior* (MAP) fit for this case is identical to the mean fit within the line thickness.

In panel (e) we allow for the selection function in the analysis. The solid black

| x | y | Selected |
|---|---|---|
| -0.2268 | -0.1069 | No |
| -0.2472 | -0.3163 | No |
| -0.1124 | -0.1824 | No |
| -0.0624 | -0.1056 | No |
| -0.0213 | 0.3405 | Yes |
| 0.0421 | -0.0441 | No |
| 0.0826 | 0.1417 | No |
| 0.1212 | 0.0195 | No |
| 0.1550 | 0.2552 | Yes |
| 0.2849 | 0.2312 | No |
| 0.2451 | 0.0931 | Yes |
| 0.3298 | 0.2030 | Yes |
| 0.4059 | 0.0083 | No |
| 0.3347 | 0.3687 | Yes |
| 0.4719 | 0.6837 | Yes |
| 0.5035 | 0.4236 | Yes |
| 0.6641 | 0.5265 | Yes |
| 0.5395 | 0.3458 | Yes |
| 0.7036 | 0.3756 | Yes |
| 0.6428 | 0.5320 | Yes |
| 0.7528 | 0.5235 | Yes |
| 0.7193 | 0.4216 | Yes |

Table 2.3 *The table contains 22 pairs of measured $x, y$ values.*

wiggly curve is the mode of the FMCMC regression line fit distribution after allowing for the selection function. It is very close to the simulated line (gray). The upper and lower dashed curves are the 68% credible region boundaries of the FMCMC regression line fit distribution. The fit uncertainty is larger when we allow for the selection effects and is very asymmetric. In panel (f), the mode of the fit distribution (wiggly curve in panel (e)) has been replaced by the best fitting straight line to the mode curve given by equation (2.90). The MAP fit line is indicated by the dot-dashed line.

$$y = 0.01004 + 0.674\,x, \tag{2.90}$$

The agreement with the true regression line is very good.

Figure 2.17 shows the marginal distributions for the regression line parameters $\alpha, \beta, \sigma$ and nuisance parameters $\mu, \tau$. The values used in the simulation for $\alpha, \beta, \sigma$, shown by the solid vertical lines, are in good agreement with the distributions.

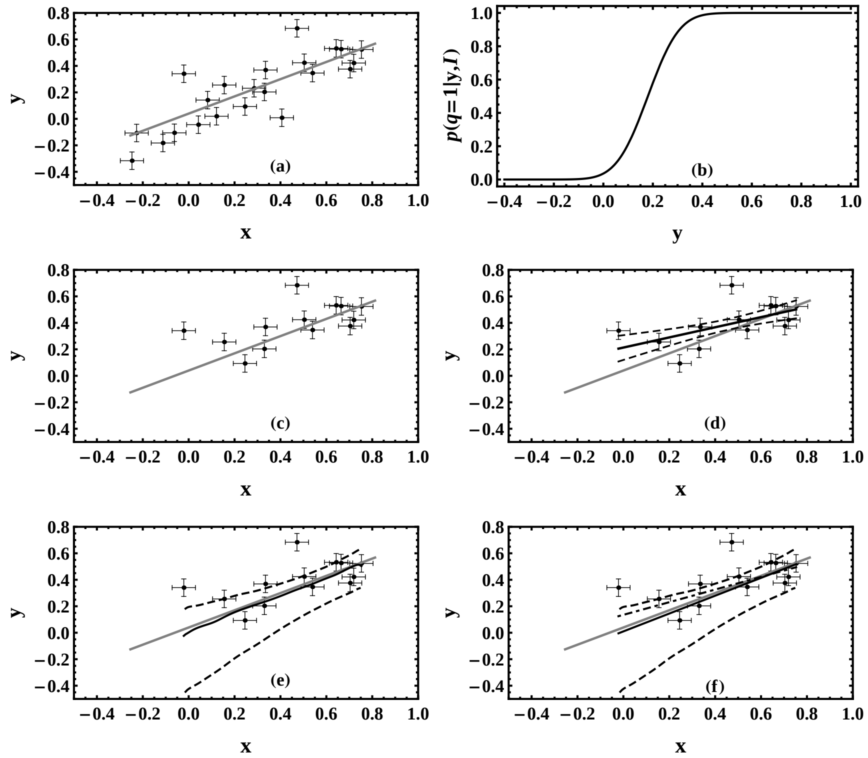


Figure 2.16 Panel (a) shows the simulated straight line together with the second example full simulated data set after adding intrinsic scatter and measurement errors in both coordinates. The error bars indicate the $1\sigma$ IID measurement errors. Panel (b) shows the selection function $p(q = 1|y, I)$. Panel (c) shows the same simulated straight line with only the selected data set. Panel (d) shows the FMCMC mean regression line fit (black solid line) to the selected data set, ignoring the selection function, and the mean $\pm 1\sigma$ fit uncertainty (dashed curves) compared to the simulated line (gray). In panel (e) we allow for the selection function in the analysis. The solid black wiggly curve is the mode of the FMCMC regression line fit distribution. The upper and lower dashed curves are the 68% credible region boundaries of the FMCMC regression line fit distribution. In panel (f), the mode of the fit distribution (wiggly curve in panel (e)) has been replaced by the best fitting straight line to the mode curve. The MAP fit line is indicated by the dot-dashed line.

## 2.8 Regression with non detections

Suppose you make measurements of some quantity $y$ at a specified set of values of the independent variable $x$. In astronomy $y$ could be the flux of a source in some frequency band and $x$ might be the optical flux or redshift. It is commonly the case that a source for which $y$ exceeds three times the background noise level is consid-

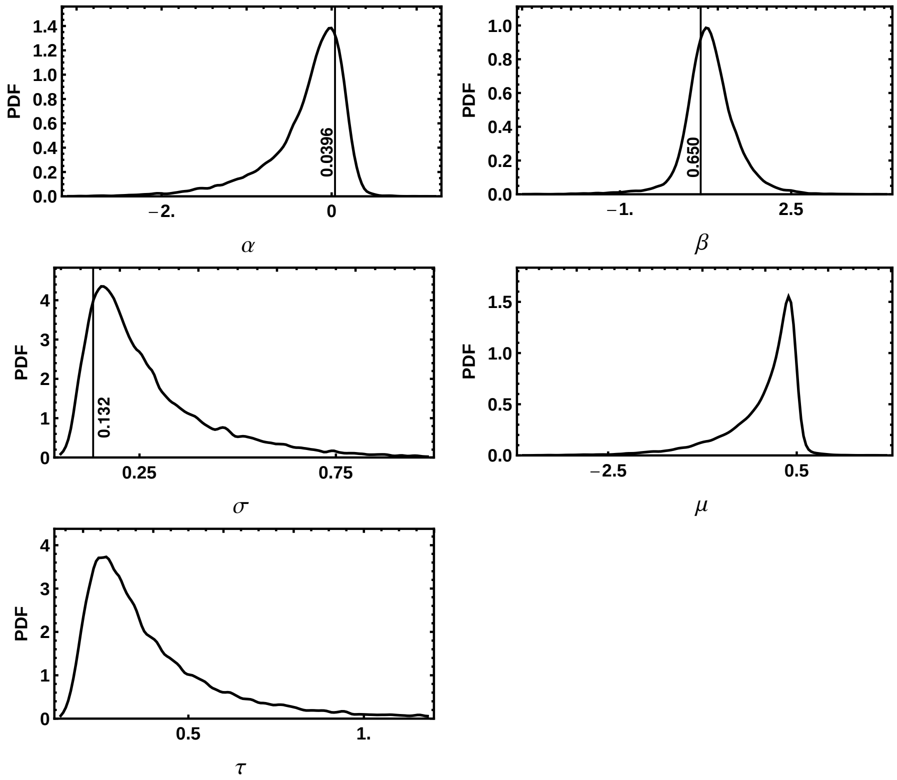


Figure 2.17 Posterior marginal distributions for the regression line parameters $\alpha, \beta, \sigma$ and nuisance parameters $\mu, \tau$.

ered detected, otherwise considered a non detection. In the case of a non detection, the $y$ value is often specified by an upper limit set = 3× the background noise level. Non detections involving an upper and/or lower limit are referred to as "censored" data. See Feigelson (1992) [6] for a review of censored data in astronomy.

The reader is referred to Kelly (2007) [13] for a Bayesian solution to regression analysis of censored data, with measurement errors in both dependent and independent coordinates, by employing a Gaussian mixture model for the true values of the independent variables.

## 2.9 Summary

In this chapter we have introduced multilevel modeling (MLM) as away to handle hidden variables and missing data problems. Our first example was concerned with fitting a straight line model to some data with measurement errors in both the dependent ($y$) and independent ($x$) variables. It was necessary to introduce additional variables $x_{ti}$ to represent the hidden true $x$ coordinates which we marginalized over. By employing an informative prior for $x_t$, we can learn about the mean and variance of the $x_t$ values and avoid the biased estimates of the intercept and slope common to ordinary least-squares analysis of this situation. A mixture of Gaussians (Gaussian mixture model) is flexible enough to model a wide variety of distributions and simplifies the integration over $x_t$. The analytic expressions were then generalized to allow for correlations between the $x, y$ measurement errors.

We extended this approach to handle linear regression problems where there is an intrinsic scatter in the relationship between the true hidden values of the dependent and independent variables. If the intrinsic scatter is modeled by a Gaussian, then the integration over the hidden variables can be evaluated analytically, greatly facilitating the calculations. An alternative is to treat the hidden true coordinates as additional parameters and use MCMC techniques to integrate over these parameters to extract the regression parameters. This is applicable even if the intrinsic scatter is non Gaussian but can still be parametrically modeled. Another advantage is that it yields the marginal distributions of the true hidden $x_t, y_t$ values to expose representative samples of the underlying regression. The multilevel (hierarchical) Bayesian regression effectively de-convolves the blurring effect of the measurement errors. We illustrated both of these approaches in a detailed analysis of two simulated data sets employing the fusion MCMC (FMCMC) algorithm. We also explored the effect of measurement errors on the regression correlation coefficient.

The analysis outlined above was then extended to regression with multiple independent variables. Section 2.7 was devoted to handling selection effects which cause some potential data to be missed giving rise to biased estimates. We learned how to extend the regression analysis to correct for the missing data and carried out a detailed analysis of two simulated data sets which had a selection function which gave rise to a gradual cutoff in the measured dependent variable. We also explored the dependence of the selection effect on the regression slope parameter for a variety of values of the selection break parameter.

A valuable reference used in preparing the chapter was Kelly (2007) [11] which also includes a discussion of how to handle Bayesian regression analysis when some of the data are only upper limits, referred to as censored data. Hopefully, the analysis in this chapter is detailed enough to enable the reader to apply these techniques to their own data and develop multilevel models (MLM) for applications in new areas. Loredo (2013) [16] provides a broader review of MLM astrostatistical applications, some of the latest of which are described the first volume of ‘Astrophysical Challenges for the New Astronomy’ [10].

# References

[1] Akritas, M. G., & Bershady, M. A. (1996), 'Linear Regression For Astronomical Data With Measurement Errors And Intrinsic Scatter', Astrophys. J., **470**, pp. 706-714.

[2] Berger, J. O., Strawderman, W. and Tang, D. (2005),'Posterior Propriety and Admissibility of Hyperpriors in Normal Hierarchical Models', Ann. Stat., 33, pp. 606-646

[3] Carroll, R. J., Roeder, K., & Wasserman, L., (1999), 'Flexible parametric measurement error models', Biometrics, **55**, pp. 44-54.

[4] Carroll, R. J., Ruppert, D., Stefanski, L. A., Crainiceanu, C. M. (2006) *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd edn., Chapman & Hall/CRC, Boca Raton.

[5] Cheng, C-L., & Van Ness, J. W. (1999), *Statistical Regression with Measurement Error: Kendalls Library of Statistics 6*, Arnold, London.

[6] Feigelson, E. D. (1992), 'Censoring in Astronomical Data Due to Nondetections', in *Statistical Challenges in Modern Astronomy*, E. Feigelson & G. Babu (Eds.) New York: Springer, pp. 221-237

[7] Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004), *Bayesian Data Analysis*, 2nd edn., Chapman & Hall/CRC, Boca Raton

[8] Gelman, A. (2006), 'Prior distributions for variance parameters in hierarchical models', Bayesian Analysis, 1, pp. 515-533

[9] Gull, S. F. (1989), 'Bayesian data Analysis - Straight Line Fitting,' in *Maximum Entropy & Bayesian Methods*, J. Skilling (Ed.), Kluwer Academic Publishers Dordrecht, pp. 511-518.

[10] *Astrostatistical Challenges for the New astronomy*, Hilbe, J. M. (ed), Springer series in astrostatistics #1, Springer Science+Business Media New York (2013)

[11] Kelly, B.C. (2007), 'Some Aspects Of Measurement Errors In Linear Regression Of Astronomical Data', Astrophys. J., **665**, pp. 1489-1506.

[12] Kelly, B.C. & Fan, X. (2008), A Flexible Method Of Estimating Luminosity Functions', Astrophys. J., **683**, pp. 874-895.

[13] Kelly, B.C. (2013), 'Measurement Models in astronomy', in *Statistical Challenges in Modern astronomy V*, E.D. Feigelson and G.J. Babu (Eds.), Vol. 209, Springer Science + Business Media, New York, pp. 147-162.

[14] Mandel, K. S., Narayan, G., & Kirshner, R. P. (2011), 'TYPE Ia SUPERNOVA LIGHT CURVE INFERENCE: HIERARCHICAL MODELS IN THE OPTICAL AND NEAR-INFRARED', Ap.J., 731, pp.120-145

[15] Loredo, T. J. & Hendry, M. A. (2010), 'Bayesian multilevel modelling of cosmological parameters' in *Bayesian Methods in Cosmology*, Hobson, M. P., Jaffe, A.

H., Liddle, A. R., Mukeherjee, P., Parkinson, D. (eds.), Cambridge University Press, Cambridge, pp. 245-264

[16] Lordeo, T. J. (2013), 'Bayesian astrostatistics: a Backward Look to the Future', in *Astrostatistical Challenges for the New astronomy*, Hilbe, J. M. (ed), Springer Series in astrostatistics #1, Springer, pp. 15-40

[17] Wall, J. V. & Jenkins, C. R. (2012), *Practical Statistics for astronomers*, Cambridge University Press.

# Appendix A

## FMCMC control system details

In this appendix we describe the details behind part of the complex control system that automates the choice of efficient proposal distributions even when the parameters exhibit strong correlations. In broad terms the fusion MCMC (FMCMC) algorithms employs two different proposal schemes, the 'I' proposals and the 'C' proposals. The 'I' scheme is ideally suited for the exploration of independent parameters while the 'C' scheme is well suited to dealing with correlated parameters. Each scheme is employed 50% of the time and the two are designed to work together.

Here we focus on the 'I' proposal scheme which employs a multiple control system (CS). Let $m$ = the number of model parameters. For the 'I' proposals we employ a separate Gaussian proposal distribution for each parameter. The challenge here is to choose a suitable $\sigma$ for each of the $m$ Gaussian distributions for each parallel chain. To tune the $\sigma$ manually involves running a series of FMCMC trials, each time varying individual $\sigma$ until the FMCMC appears to converge on an equilibrium distribution with a proposal acceptance rate that is reasonable for the number of parameters involved, e.g., approximately 25% for a large number of parameters [52]. The parallel tempering feature of our analysis compounds the difficulty because each of the parallel chains is exploring a different probability distribution. This is because the likelihood is raised to a different power, $\beta$. One of the applications of FMCMC is in the arena of exoplanet research where it is employed as a multi-planet Kepler periodgram. For an 8 planet model with 42 parameters [1] and 8 parallel chains requires 336 different $\sigma$ values. Faced with this problem the author decided to devise a method to automate the selection of $\sigma$ values.

Initially this was accomplished with two stage control system, stages 1 and 2. The CS was subsequently improved by the additional of an initial stage now referred to as stage 0. More recent studies indicate that the combination of stage 0

[1] The extra two parameters are the systematic velocity of the star and the unknown extra noise (stellar jitter) parameter.

followed by stage 2 generally performs more efficiently than the combination of stages 0, 1, and 2. Future versions of the code will likely see the complete demise of stage 1, but for the moment it remains as an option that can be turned on or off. Sections A.1, A.2 and A.3 describes these stages. The various stages can be separately turned on or off by a set of binary (0 or 1) control switches labelled sw0, sw1, and sw5 as outlined below.

## A.1 Control system stage 0

If sw0 is set = 1, then the control system begins by executing stage 0. In what follows $\sigma_{\alpha,\beta}$ represents the proposal $\sigma$ for the parameter $X_{\alpha,\beta}$, indexed by $\alpha$ for each parameter and indexed by $\beta$ for each chain. Let $\boldsymbol{X}_{i,\beta}$ represent the vector of parameter values for the $i^{\text{th}}$ iteration. As a first step we choose starting values for $\{X_{\alpha,\beta}\}$. The same initial set is used for each parallel chain. An initial sequence of $n0$ iterations (typically $n0 = 5000$) is executed to crudely home in on a set of proposal sigma values that get the joint acceptance rate close to a desirable value. Initially the proposal $\sigma$ are automatically set equal to 15% of the user specified prior range of each parameter. Keep in mind that the goal is to achieve an automatic CS that can be employed for nonlinear model fitting for a wide range of problems so we are assuming no prior experience with applying FMCMC to the particular problem in hand. The same initial set is used for each parallel tempering chain.

In the first $n0/5$ iterations the number of accepted proposals and starting proposal $\sigma$ values are recorded. This is repeated for a second $n0/5$ iterations with all the proposal $\sigma$ values reduced by a factor of 5. For the next two $n0/5$ iterations the proposal $\sigma$ values are reduced by a further factor of 5 each time and likewise for the last $n0/5$ iterations. The choice of $\sigma$ values that yield an acceptance rate closest to the desired joint acceptance rate $\approx 25\%$ are then employed as the starting $\sigma$ values for the stage 2 CS. This selection of starting $\sigma$ values is carried out separately for each parallel chain. Initially there may be only a few parameters for which the proposal $\sigma$ is seriously limiting the joint acceptance rate. We will refer to these as the critical parameters. In the exoplanet problem these are the orbital period/frequency parameters. The stage 0 procedure to bring these key parameter proposal $\sigma$ into range will very likely drive the proposal $\sigma$ for other parameters to values which are too small. This situation is eventually remedied in stage 2. A stage 1 option, which may be employed following stage 1, is described next. The focus continues to be on evolving the proposal $\sigma$ for the critical parameters while the FMCMC explores the frequently multi-modal target probability distribution.

## A.2 Control system stage 1

As mentioned in the introduction of this appendix, there no longer appears to be a need for this stage of the control system so the reader may want to skip over this section. If sw1 is set = 1, then stage 1 is execute following stage 0. The stage 1 CS contains major and minor cycles. During the major cycles the current set of $\{\sigma_{\alpha,\beta}\}$ are used for $n1$ iterations. The acceptance rate achieved during this major cycle is compared to the target acceptance rate. If the difference for the $\beta = 1$ chain is greater than a chosen threshold, $tol1$, then minor cycles are employed to explore the sensitivity of the acceptance rate to the indiviual $\sigma_{\alpha,\beta}$. Only one $\sigma_{\alpha,\beta}$ is perturbed in each minor cycle. The $\{\sigma_{\alpha,\beta}\}$ are updated and another major cycle run. The algorithm involves the following steps:

1). Start with the $\{\sigma_{\alpha,\beta}\}$ set determined from stage 0 in Section A.1.

2). Choose the maximum factor, *scmax*, to use in perturbing the current $\{\sigma_{\alpha,\beta}\}$ values. Typically $scmax = 10$.

3). **Major cycle iterations:** set the counter $nc1 = 1$ and execute $n1 = 1000$ iterations of the FMCMC with the current $\{\sigma_{\alpha,\beta}\}$ values labeled by $\sigma_{\alpha(\text{current})}$. Keep track of the number of accepted proposals for each chain which we represent by a vector quantity labeled $\boldsymbol{nc2}$. Again at iteration $i$, chain $\beta$ is in parameter state $\boldsymbol{X}_{i,\beta}$.

4). Compute an error vector term, $\boldsymbol{er}$, based on the results of the major cycle which is given by

$$\boldsymbol{er} = \boldsymbol{nc2} - \lambda \times n1. \tag{A.1}$$

The term $er_\beta$ stands for the component of $\boldsymbol{er}$ corresponding to chain $\beta$. The switch $\boldsymbol{sw}$ is a vector of $+1$ and $-1$ values, one for each chain. If $er_\beta < 0$, then set switch $sw_\beta = +1$ , otherwise set $sw_\beta = -1$. Set $\boldsymbol{X}_{i,\beta(\text{major})}$ equal to the set of parameter values at the end of the major cycle.

5). **Minor cycle iterations:** normally the major cycle is followed by a sequence of minor cycles, one for each parameter, each for $n2 = 200$ iterations. In the minor cycles each $\sigma_{\alpha,\beta}$ is separately perturbed to determine which of the $\sigma_{\alpha,\beta}$ need to be adjusted to permit the $er_\beta$ to move closer to zero. If the value of $er_\beta$ is negative for a particular $\beta$ then the $\sigma_{\alpha,\beta}$ are all decreased in the minor cycles, otherwise they are increased. The amount of the increase or decrease is discussed in item 7 below.

6). **Minor cycle data collection:** repeat the following steps for each model parameter:

{

a) Set each element of the counter vector $\boldsymbol{nc2} = 0$. This counter keeps track of the number of proposals accepted in the next minor cycle.

b) Execute $n2$ MCMC iterations, starting from $X_{\alpha,\beta}$(major), using a new value of $\sigma_{\alpha,\beta}$ given by

$$\sigma_{\alpha(\text{new}),\beta} = \sigma_{\alpha(\text{current}),\beta} \times \left(sc_\beta\right)^{\text{sw}_\beta}, \tag{A.2}$$

where

$$sc_\beta = (scmax)^{-scexp_\beta}, \tag{A.3}$$

and where $scmax$ is a constant and $scexp_\beta$ is given by

$$scexp_\beta = \text{Minimum}\left[1\ ,\ \left(\frac{|er_\beta|}{\lambda\ n1}\right)^\gamma\right], \tag{A.4}$$

where the acceptance rate, $\lambda \approx 25\%$. For the example given in the section of the text dealing with extra-solar planets, $scmax = 10$ and the value of the damping factor, $\gamma = 1.6$.

If $sw_\beta = +1$, then equation (A.2) decreases $\sigma_{\alpha,\beta}$ by a factor that depends on the size of the error $er_\beta$, or increases $\sigma_{\alpha,\beta}$ by this factor if $sw_\beta = -1$.

c) Increase the counter $nc2_\beta$ each time the new parameter set proposal is accepted.

d) At the end of $n2$ iterations store $nc2_\alpha$ into a counter array $nb_{\alpha,\beta}$ and reset $\sigma_{\alpha,\beta}$ to its original value at the start of the minor cycles.

}

7). **Minor cycle analysis:** If $sw_\beta = +1$ compute $nb_{\max,\beta}$ and $nb_{\min,\beta}$, the maximum and minimum values of the counter array $nb_{\alpha,\beta}$.

If $nb_{\min,\beta} < nb_{\max,\beta}$, only modify the $\sigma_{\alpha,\beta}$ for parameters for which $nb_{\alpha,\beta} = nb_{\max,\beta}$, according to

$$\sigma_{\alpha(\text{new}),\beta} = \sigma_{\alpha(\text{current}),\beta} \times \left(1 - gain_\beta \times \left(1 - sc_\beta\right)\right), \tag{A.5}$$

where $gain_\beta < 1$ is a simple monotonic function of $\beta$ and is discussed further at the end of this section.

Else, if $nb_{\min,\beta} = nb_{\max,\beta}$, then multiply all $\sigma_{\alpha,\beta}$ by $(1 - gain_\beta \times (1 - sc_\beta))$ for use in the next major cycle.

An earlier version of the CS changed all $\sigma_{\alpha,\beta}$ values in proportion to their ability to move the acceptance rate towards the target value. In some circumstances the stage 1 CS resulted in certain $\sigma_{\alpha,\beta}$ values being driven down to unreasonably small values. To prevent this from happening lower bounds were imposed on all $\sigma_{\alpha,\beta}$ and another vector switch labeled ***swmin*** was introduced to indicate the parameters for which this had occurred by setting that switch component = 1 from the normal setting of 0. Restricting the changes in $\sigma_{\alpha,\beta}$ to parameters corresponding to $nb_{\max,\beta}$ yielded an improved performance.

If $sw_\beta = -1$, we want to decrease the acceptance rate by increasing $\sigma_{\alpha,\beta}$.

In this case, if $nb_{\min,\beta} < nb_{\max,\beta}$, only modify the $\sigma_{\alpha,\beta}$ for parameters for which $nb_{\alpha,\beta} = nb_{\min,\beta}$, by a factor of $1 + gain_\beta \times (sc_\beta^{-1} - 1)$ for use in the next major cycle. We adjust the $\sigma_{\alpha,\beta}$ for parameters that have been shown to give the biggest reduction in the acceptance during the minor cycles, i.e., given rise to the $nb_{\min,\beta}$ values.

Otherwise, if $sw_\beta = -1$ and $nb_{\max} = nb_{\min}$, then multiply all $\sigma_{\alpha,\beta}$ by $1 + gain_\beta \times (sc_\beta^{-1} - 1)$.

8). **Termination of stage 1 CS:** The sequence of major and minor cycles is repeated unless one of two conditions is met. If the number of iterations exceeds *minCS* and $|er_{\beta=1}| < tol2$, then the stage 1 CS is terminated and normally followed by the stage 2 CS (if the appropriate switch is enabled, $sw5 = 1$). The error signal is a count which is subject to statistical fluctuations so typically $tol2$ is set to a value of $1.5\sqrt{\lambda \times n1}$. Bare in mind that while the stage 1 CS is tuning the $\{\sigma_{\alpha,\beta}\}$, the FMCMC is homing in on the most probable parameter set. Setting $minCS \sim 20,000$ iterations allows some time for the homing in process to occur. When the number of iterations exceeds *maxCS* then the stage 1 CS is always terminated and again normally followed by the stage 2 CS. If the homing in process has not happened by $maxCS \sim 60,000$ iterations, then the chances are better it will happen during the stage 2 CS.

Between *minCS* and *maxCS* the minor cycles are turned off on any occasion where the magnitude of the largest component ***er*** vector is $< tol1$. Typically $tol1 = 1.5\sqrt{\lambda \times n1}$.

The performance of the stage 1 CS can be assessed from an examination of the error signal which is shown in Figures A.1 and A.2 for a 3 planet FMCMC fit to a sample of Gliese 581 exoplanet data. The different traces correspond to the eight $\beta$ values employed during the test with

$$\beta = \{0.09, 0.13, 0.20, 0.30, 0.42, 0.55, 0.74, 1.0\}.$$

The $\beta = 1$ error signal is shown by the thicker red dashed curve. Normally, the stage 1 CS is terminated as soon as number of iterations exceeds *minCS* and $|er_{\beta=1}| < tol2$ or shut off when $maxCS = 60,000$ iterations. In this particular case we set $tol2 = 0$ and $maxCS = 300,000$ iterations to provide a longer span of stage 1 CS activity. For both figures the $gain_\beta$ was a simple monotonic function of $\beta$. In the case of Figures A.1, the $gain_\beta$ ranged from 0.6 at the lowest $\beta$ value to 0.7 at $\beta = 1$. The reduced amplitude error signal shown in Figures A.2, was achieved for a stronger taper of $gain_\beta$ ranging from 0.4 at the lowest $\beta$ value to 0.7 at $\beta = 1$. The final $gain_\beta$ equation employed is

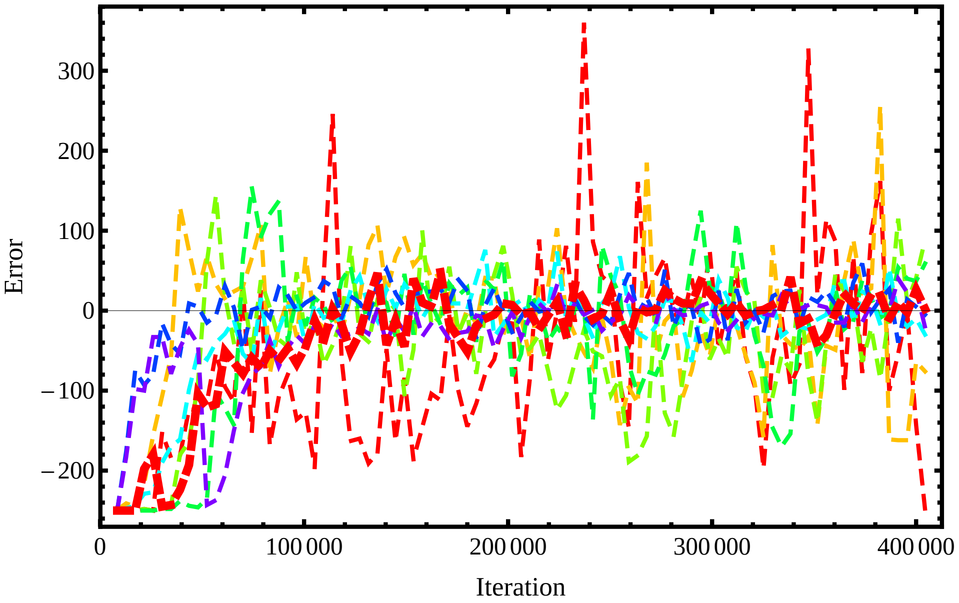


Figure A.1 The stage 1 control system (CS) error signal versus iteration for a CS $gain_\beta$ value ranging from 0.6 at the lowest $\beta$ value to 0.7 at $\beta = 1$.

$$gain_\beta = 0.361 + 0.625\beta - 0.290\beta^2. \tag{A.6}$$

### A.3 Control system stage 2

If sw5 is set = 1, then stage 2 is execute following stages 0 and 1, or if sw1=0, then stage 2 follows immediately after stage 0. For sw5=0, stage 2 is never executed. As mentioned above recent studies indicate that the combination of stage 0 followed by stage 2 performs more efficiently than the combination of stages 0, 1, and 2 above. A limitation of both stages 0 and 1 is that although the desired

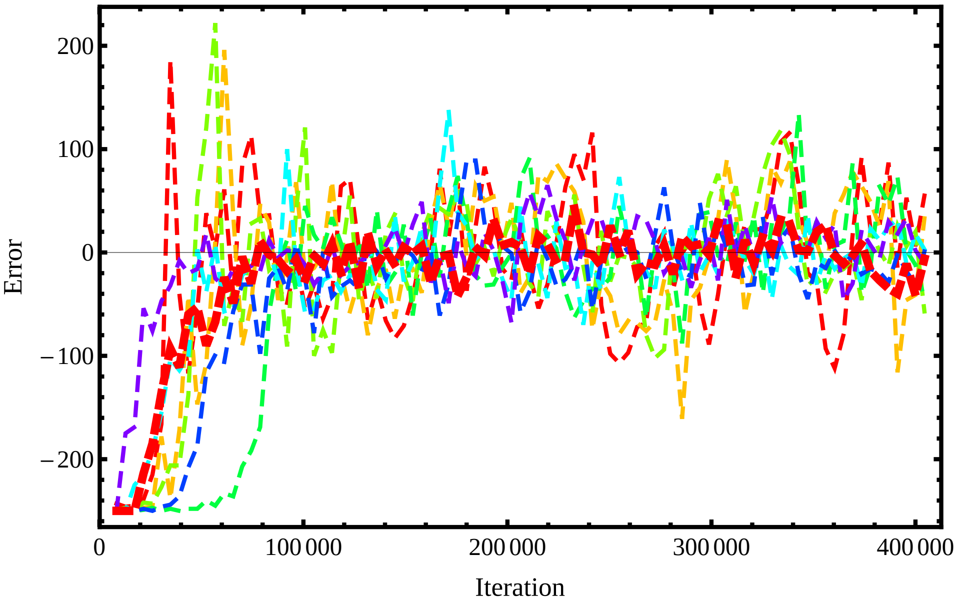


Figure A.2 The stage 1 control system (CS) error signal versus iteration for a CS $gain_{\beta}$ value ranging from 0.4 at the lowest $\beta$ value to 0.7 at $\beta = 1$.

joint acceptance rate may be achieved, typically a subset of the proposal $\sigma$'s are too small leading to an excessive autocorrelation in the MCMC iterations for these parameters. The second stage CS corrects for this as follows.

The goal of the second stage is to achieve a set of proposal $\sigma$'s that equalizes the FMCMC acceptance rates when new parameter values are proposed separately and achieves the desired acceptance rate when they are proposed jointly. Let $acc(1)$ equal the acceptance for single parameter proposals and $acc(m)$ the desired acceptance rate ($\lambda = 0.25$) for $m$ parameter joint proposals. We need to determine how $acc(1)$ depend on $m$. Initially, we might guess that for independent parameters

$$acc(m) \approx acc(1)^m, \tag{A.7}$$

The actual relationship will be modified by the Metropolis transition kernal which says that the *acceptance probability* is given by

$$acceptance\ probability = \min(1, r) = \min\left(1, \frac{p(Y|D, I)}{p(X|D, I)}\right), \tag{A.8}$$

where $Y$ is the proposed parameter set and $X$ is the current parameter set.

The true relationship (shown below) was arrived at in the following way. An MCMC simulation was run on an $m$ parameter multivariate normal target probability distribution with a mean for each parameter of zero and a covariance matrix equal to the identity matrix. New parameters were proposed using another multivariate normal with mean zero and a covariance matrix equal to $\gamma^2$ times the identity matrix. Thus, $\gamma$ is the ratio of the proposal $\sigma$ to the target distribution $\sigma$ for each parameter. For each choice of $\gamma$ in the range 0.3 to 1.0, the MCMC acceptance rate for joint parameter proposals was determined as a function of $m$ in the range $m = 1$ to 60. For each $\gamma$ the acceptance rate was well fit by a function of the form

$$acc(m) = acc(1)^{m^{\alpha}}, \tag{A.9}$$

where the value of $\alpha$ depends on $\gamma$. Designate the value of $m$ at which acc$(m) = 0.25$ by $m_\alpha$. Figure (A.3) shows the acceptance rate versus number of model parameters, $m$, for $\gamma = 0.7$. In this particular case the fitted dashed line is given by $acc(m) = (acc(1))^{m^{\alpha}}$, where $\alpha = 0.703$. The horizontal line, corresponding the an acceptance rate of 25%, intersects the fitted curve at $m = 11.8$. Note, in the figure,

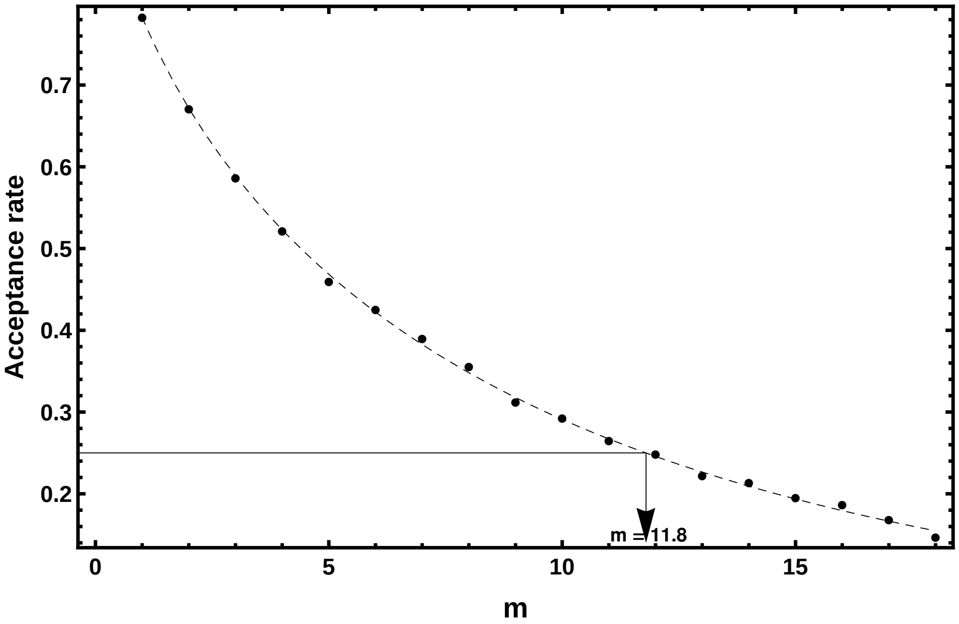


Figure A.3 The acceptance rate versus number of model parameters, $m$, for $\gamma = 0.7$. The fitted dashed line is of the form $acc(m) = (acc(1))^{m^{0.703}}$. The horizontal line, corresponding the an acceptance rate of 25%, intersects the fitted curve at $m_\alpha = 11.8$.

$acc(1)$ is the acceptance rate found for the $m = 1$ model with the same value of $\gamma$ as the models with $m > 1$. For $\gamma$ ranging from 0.3 to 1.0, $m_\alpha$ varied from 60 to 6.4 and $\alpha$ from 0.64 to 0.74. A decaying exponential provided a good fit to the $(m_\alpha, \alpha)$

pairs yielding Equ. (A.10).

$$\alpha = 0.6426 + 0.1507 \exp(-m_\alpha/15), \tag{A.10}$$

Figure (A.4) shows a plot of these values.

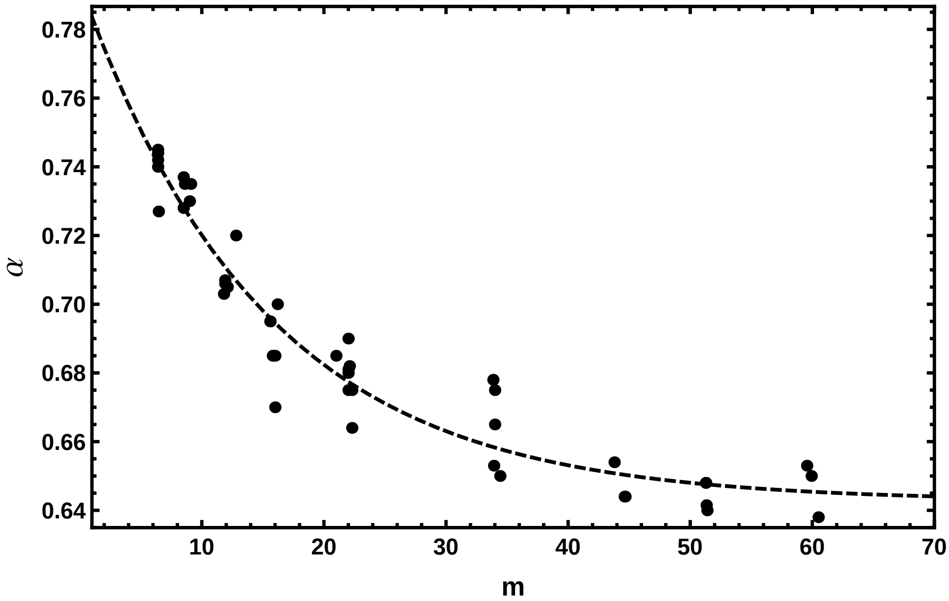


Figure A.4 A plot of computed $(m_\alpha, \alpha)$ pairs (with the $\alpha$ subscript dropped) together with the fitted exponential.

The final relationship between $acc(m)$ and $acc(1)$

$$acc(m) = acc(1)^{m^{k\,\alpha}}, \tag{A.11}$$

$$acc(1) = acc(m)^{1/m^{k\,\alpha}} = \lambda^{1/m^{k\,\alpha}}, \tag{A.12}$$

where $k$ is an empirically derived fudge factor that compensates for the fact that the true target distributions may not be well represented by a multivariate normal with a diagonal covariance matrix. A typical value of $k$ is in the range 0.91 (for $m \approx 12$) to 0.97 (for $m \approx 40$). For $m = 17$, Eq. (A.10) gives $\alpha = 0.69$. Plugging these values into Eq. (A.12) yields $acc(1) = 0.77$.

The next step is to adjust the individual parameter proposal $\sigma$'s to achieve an acceptance of $acc(1)$ given by Equ. (A.12). Using the proposal $\sigma$'s obtained from the previous stage, each parameter is allowed to vary one at a time during a minor cycle [2] of $n3 = 1000$ iterations and the acceptance rate measured. Let $acc_1$ = the

[2] No output is recorded during the stage 2 iterations except for two items: (1) the relevant statistic needed to

measured acceptance rate when the proposal $\sigma$ for the parameter in question was $\sigma_1$. We then update the proposal $\sigma$ for this parameter to $\sigma_2$ according to

$$\sigma_2 = \sigma_1 \sqrt{\frac{(acc_1 + \Delta)}{acc(1)} \frac{(1 - acc(1))}{(1 - acc_1 + \Delta)}}, \tag{A.13}$$

where we use a $\Delta = 0.01$.

If $acc_1 = acc(1)$, then Equ. (A.13) leaves the proposal $\sigma$ unchanged except for the small effect of the $\Delta$ term. The $\Delta$ term is there to handle the extremes of $acc_1 = 0$ and 1 gracefully. If $acc_1 = 1$, then we want to increase the proposal $\sigma$ for that parameter. From Equ. (A.13) and $m = 17$ parameters, $\sigma_2/\sigma_1 = 6.7$. If on the other hand $acc_1$ is too low, say $acc_1 = 0.25$, we want to decrease the size of the proposal distribution. In this case, Equ. (A.13) yields $\sigma_2/\sigma_1 = 0.39$. Equ. (A.13) can be iterated for each parameter to achieve a final set of proposal $\sigma$'s that achieve equal acceptance rates and a final joint acceptance rate of acc$(m) = \lambda$.

In general, the burn-in period occurs within the span of the CS, i.e., the significant peaks in the joint parameter probability distribution are found. The stage 2 CS improves the choice of the 'I' proposal $\sigma$ for the highest probability parameter set. Occasionally, a new higher (by a user specified threshold) target probability parameter set emerges after the CS has been turned off. The control system has the ability to detect this and re-activating the second stage. In this sense the CS is adaptive. If this happens the iteration corresponding to the end of the control system is reset, i.e, the burn-in period is reset. The useful FMCMC simulation data is obtained after the CS is switched off.

Although inclusion of the control system may result in a somewhat longer effective burn-in period, there is a huge saving in time because it eliminates many trial runs to manually establish a suitable set of proposal $\sigma$'s. When the $\sigma$'s are large all the FMCMC chains explore broadly the prior distribution and locate significant probability peaks in the joint parameter space. As the proposal $\sigma$'s are refined these peaks are more efficiently explored, especially in the higher $\beta$ chains. In the exoplanet problem, this annealing of the proposal $\sigma$'s typically takes place within the first 15,000 (unthinned) iterations for one or two planet. When there is evidence for a large number of planets the annealing can span a much larger range. This may seem like an excessive number of iterations but keep in mind that (a) we are dealing with sparse data sets that can have multiple, widely separated probability peaks, (b) the typical start location in parameter space is far from the target posterior peak, and (c) we want the FMCMC to locate the most significant probability peak before finalizing the choice of proposal $\sigma$'s. Within each chain, the CS corresponds to

compute $acc(1)$ for each parameter, and (2) any parameter vector that emerges in this stage which has a target probability higher than any previous parameter vector found (stored as *maxlogpdfAll*).

an annealing operation. Taken together with the parallel tempering, the two operations enhance the chances of detecting peaks in the target posterior compared to just implementing either one.

## A.4 Further CS studies

Here we report on some experiments to improve the CS which may help others. All the experiments were done with a blind 3 planet model fit to the HARPS Mayor 2009 data set for Gliese 581. There is good agreement that there are at least 4 Kelper-like signals. The strongest three signals occur at periods of 5.37, 12.92, 66.9 d. The presence of the other signals makes a blind 3 planet search an interesting challenge. We judged any change to the CS to be a success if it improved the speed of detection all three of the strongest signals, i.e., fewer FMCMC iterations.

1). Tried out a version of the CS that allowed one cycle of stage 2 to be run immediately following stage 0 and before stage 1. Stage 2 was run again following stage 1. The thinking here was that since many proposal $\sigma$ were driven down to very low values in stage 0, it could be advantageous to bring them back into a meaningful range (so they were influencing the joint acceptance rate). This seems to have made matters worse. Why? In the exoplanet problem the three orbital frequency/period parameters turned out to be the main focus of the action in stage 0 and 1. Maintaining the proposal $\sigma$ of the other parameters at values much smaller than optimum meant that little progress was made in refining the parameters for local probability peaks as they were encountered. The extra noise parameter, $s$, remained inflated and facilitated progress in finding the globally most probable period set.

2). Experimented with fewer iterations in each minor cycle of stage 2 (reduction from $n3 = 1000$ to 500). The potential advantage here would be to speed up the FMCMC run. Found that the success rate in correctly identifying the three periods in a 500,000 iteration run was reduced by 50%.

3). To speed up the parallel processing *npara* iterations are done at once. This also resulted in a thinning of the stored iterations by the same factor. In principle increasing *npara* should reduce the compute time. Changing *npara* from the normal value of 10 to 40 resulted in a 50% reduction in success rate.

4). Tried out a new stage 2 algorithm for adjusting the proposal $\sigma$.

$$\sigma_2 = \sigma_1 \frac{(acc_1 + 0.03)}{acc(1)} \sqrt{\frac{(1 - (acc(1))^2)}{(1 - (acc_1)^2 + 0.01)}}, \tag{A.14}$$

Preliminary results indicate that it is not quite as good as the existing algorithm.

## A.5 Choosing tempering values

For parallel tempering to work well we need to select a set of tempering $\beta$ values to achieve a desirable swap rate. We discovered a useful empirical correlation that can be used to automate this step. Figure A.5 shows a correlation between the parallel tempering swap rate between tempering levels $i$ and $i-1$ and the product of $(\beta_i - \beta_{i-1}) \times \delta\langle\text{loglike}\rangle$, where $\delta\langle\text{loglike}\rangle = \langle\text{loglike}\rangle_i - \langle\text{loglike}\rangle_{i-1}$, based on multiple 3 planet fits to Gliese 581 data. Figure A.6 shows the correlation based on multiple 5 planet fits to Gliese 581 data. $\langle\text{loglike}\rangle_i$ is the FMCMC average value of the log likelihood for the $i^{\text{th}}$ chain. Clearly the correlation is tighter for the 5 planet fit.

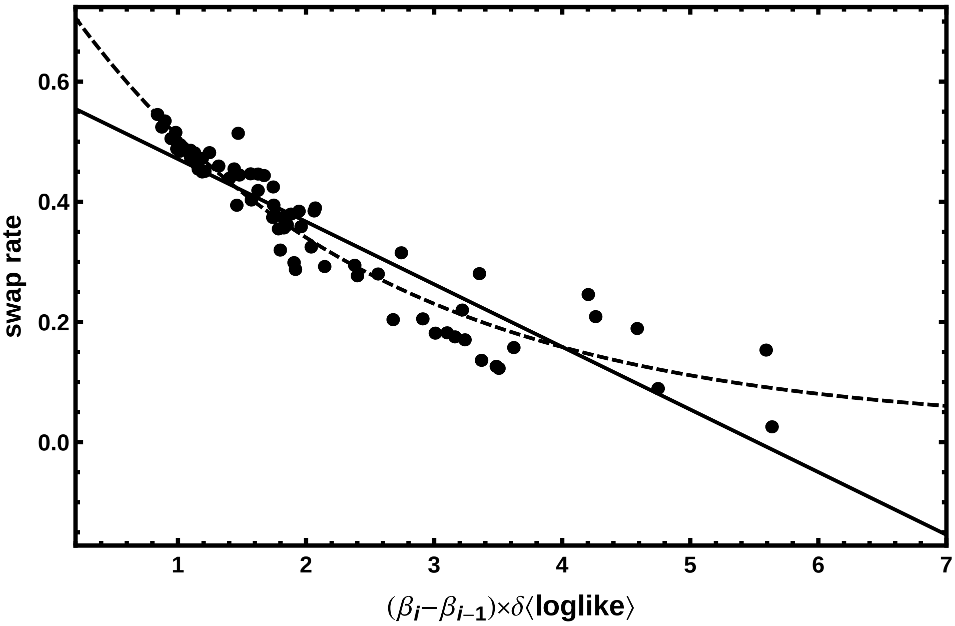


Figure A.5 Correlation between the parallel tempering swap rate and the product of $(\beta_i - \beta_{i-1}) \times \delta\langle\text{loglike}\rangle$ based on multiple 3 planet fits to Gliese 581 data.

In each figure the solid line is the best fit straight line and the dashed curve is a fit of the equation $y = A + B \times \exp[-Cx]$ where $y$ represents the swap rate and $x = (\beta_i - \beta_{i-1}) \times \delta\langle\text{loglike}\rangle$. The exponential form provides a significantly better fit. The best fitting exponentials to the 3 and 5 planet correlations are:

$$y(3\text{planet}) = 0.02235 + 0.7445 \exp[-0.4251x] \tag{A.15}$$

$$y(5\text{planet}) = 0.09137 + 0.6826 \exp[-0.5211x] \tag{A.16}$$

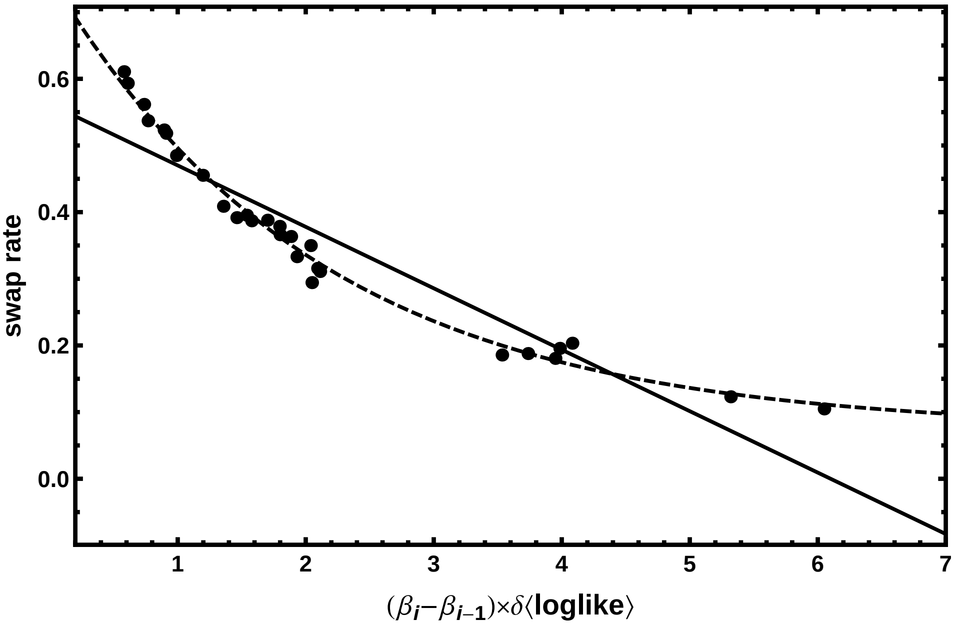


Figure A.6 Correlation between the parallel tempering swap rate and the product of $(\beta_i - \beta_{i-1}) \times \delta\langle\text{loglike}\rangle$ based on multiple 5 planet fits to Gliese 581 data.

Figure A.7 shows a comparison of equations (A.15) and (A.16). Over the range of swap rates of interest from 0.25 to 0.5 the two curves are virtually identical indicating that the correlation between swap rate and $(\beta_i - \beta_{i-1}) \times \delta\langle\text{loglike}\rangle$ is very general.

### *A.5.1 Procedure for selecting tempering values*

The empirical correlation between swap rate and $(\beta_i - \beta_{i-1}) \times \delta\langle\text{loglike}\rangle$, summarized in figure A.7, provides the basis for the following scheme for selecting $\beta$ values. Choose the desired swap rate between tempering levels, say 25% or 0.25, and figure A.7 gives a value of $(\beta_i - \beta_{i-1}) \times \delta\langle\text{loglike}\rangle = 2.8$. In practice we have obtained improved performance of the FMCMC algorithm using a value of

$$(\beta_i - \beta_{i-1}) \times \delta\langle\text{loglike}\rangle = 1.0, \tag{A.17}$$

corresponding to a swap rate of $\approx 0.5$ between the top two levels and a value of

$$(\beta_i - \beta_{i-1}) \times \delta\langle\text{loglike}\rangle = 2.0, \tag{A.18}$$

corresponding to a swap rate of $\approx 0.3$ between the other levels.

The steps in the computation of the $\beta$ values are illustrated in figure A.8. Figure A.8(a)

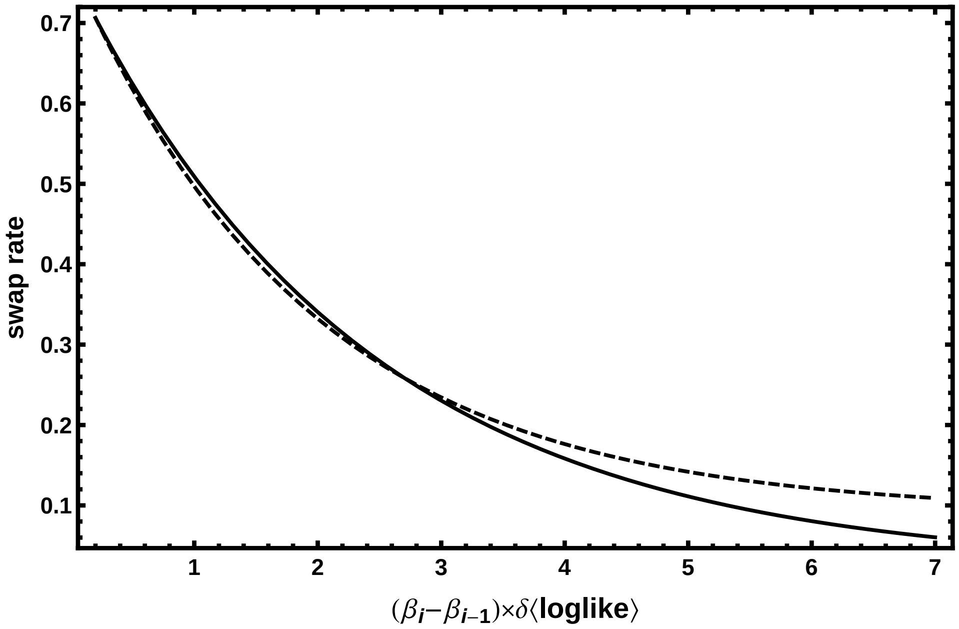


Figure A.7 Comparison of the best fit correlations between the parallel tempering swap rate and the product of $(\beta_i - \beta_{i-1}) \times \delta\langle\text{loglike}\rangle$ for the 3 planet (solid) and 5 planet (dashed) fits to Gliese 581 data.

shows the values of $\langle\text{loglike}\rangle_\beta$ obtained from an FMCMC run using an initial set of 8 tempering levels of

$$\beta = \{0.09, 0.13, 0.20, 0.30, 0.42, 0.55, 0.74, 1.0\}.$$

Figure A.8(b) shows the now familiar correlation between swap rate and $(\beta_i - \beta_{i-1}) \times \delta\langle\text{loglike}\rangle$ for this single FMCMC run. Figure A.8(c) shows a plot of $\Delta\langle\text{loglike}\rangle = (\langle\text{loglike}\rangle_{n\beta} - \langle\text{loglike}\rangle_{n\beta-i})$ versus $\Delta\beta = \beta_{n\beta} - \beta_{n\beta-i}$ for $i = 0$ to $n\beta - 1$ in steps of 1. $n\beta$ is the total number of tempering levels in this case 8 and $\beta_{n\beta} = 1$. The points are connected by dashed straight lines and the solid curve is an interpolation derived as follows. A second order spline interpolation is first fit to the Log[$\Delta\langle\text{loglike}\rangle$+0.001] versus $\Delta\beta$. The ordinate of this interpolation is transformed back by exponentiation followed by a subtraction of the small offset of 0.001. This results in an interpolation that is more constrained in its swings than one obtained using the original points. The resulting interpolation is then averaged with a straight linear interpolation of the original points to produce the final interpolation designated *interpAve*.

Figure A.8(d) shows the final interpolation (solid) and the equation $y = 1.0/\Delta\beta$ (dashed) based on equation (A.17). The intersection of the two curves yields the desired $\Delta\beta_1$ between tempering levels $n\beta$ and $n\beta - 1$. Figure A.8(e) depicts the solution for $\Delta\beta_4$. In this case the solution is given by the intersection of *interpAve*−

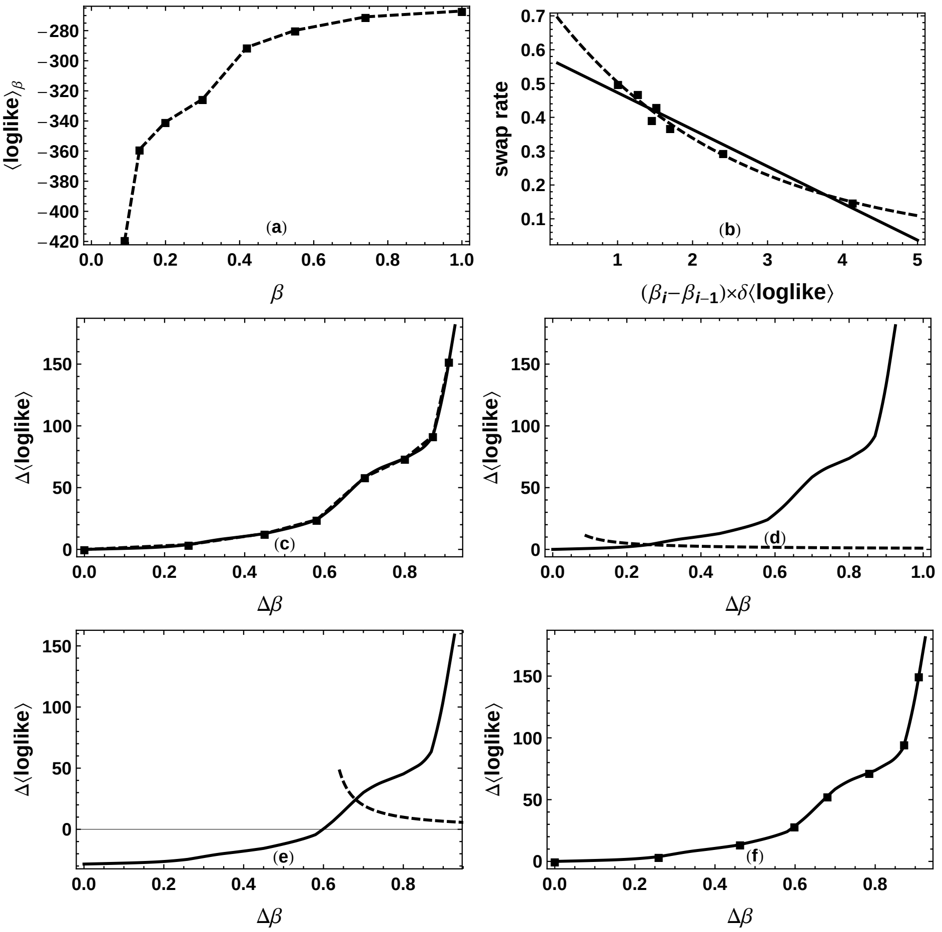


Figure A.8 Panel (a) shows the values of $\langle\text{loglike}\rangle_\beta$ obtained from an FMCMC run using an initial set of 8 tempering levels. Panel (b) shows the correlation between swap rate and $(\beta_i - \beta_{i-1}) \times \delta\langle\text{loglike}\rangle$ for this single FMCMC run. Panel (c) shows a plot of $\Delta\langle\text{loglike}\rangle$ versus $\Delta\beta$ and the interpolation. The intersection of the two curves in panel (d) yields the desired $\Delta\beta_1$ between tempering levels $n\beta$ and $n\beta - 1$. Panel (e) illustrates the intersection of the two curves described in the text that yield the desired $\Delta\beta_4$. Panel (f) shows the location of the improved set $\Delta\beta_i$ values.

*interpAve*$[\Delta\beta_3]$ and the equation $y = 2.0/(\Delta\beta - \Delta\beta_3)$ (dashed curve) based on equation (A.18). This is shown in figure A.8(d). Figure A.8(f) shows the location of the improved set $\Delta\beta_i$ values on the *interpAve* curve. They correspond to a set of tempering $\beta$ values given by

$$\beta = \{0.091, 0.13, 0.21, 0.32, 0.40, 0.54, 0.74, 1.0\}.$$

This procedure for deriving an improved set of tempering $\beta$ values from an initial FMCMC run is easily automated.

How well does this procedure work in practice? Because the shape of $\Delta\langle\text{loglike}\rangle$ versus $\Delta\beta$ curve can exhibit abrupt changes it is necessary to use more than 8 tempering levels with smaller separations in $\Delta\beta$ to more accurately define the curve if you want to do a good job. The point of going to this trouble is to attempt to pick a set of $\beta$ values that yield an optimum swap rate. In practice we did not find a strong dependence on FMCMC performance and swap rate for swap rates in the range 0.1 to 0.6.

# Appendix B
# Exoplanet priors

## B.1 Frequency search

For the Kepler model with sparse data, the target probability distribution can be very spiky. This is particularly a problem for the orbital period parameters which span roughly 6 decades [1]. The actual search in that domain is best implemented in frequency space for the following reasons. The width of a spectral peak, which reflects the accuracy of the frequency estimate, is determined by the duration of the data, the signal-to-noise (S/N) ratio and the number of data points. More precisely [24] [7], for a sinusoidal signal model, the standard deviation of the spectral peak, $\delta f$, for a S/N > 1, is given by

$$\delta f \approx \left(1.6 \frac{\mathrm{S}}{\mathrm{N}T} \sqrt{N}\right)^{-1} \quad \text{Hz}, \tag{B.1}$$

where $T$ = the data duration in s, and $N$ = the number of data points in $T$. The thing to notice is that the width of any peak is independent of the frequency of the peak. Thus the same frequency proposal distribution will be efficient for all frequency peaks. This is not the case for a period search where the width of a spectral peak is $\propto P^2$. Not only is the width of the peak independent of f, but the spacing of peaks in the spectral window function is roughly constant in frequency, which is a another motivation for searching in frequency space [56], [11]. Figure B.1 shows a section of the spectral window function, described in Section 1.5.3, for the 35 samples of radial velocity measurements for HD 208487. The true peridogram of the data is the convolution of the true spectrum with the spectral window function.

[1] With the exception of a pulsar planet (PRS 1719-14 b) with a period of 0.09d, the period range of interest is from ~ 0.2d to ~ 1000yr. A period of 1000yr corresponds roughly to a period where perturbations from passing stars and the galactic tide would disrupt the planet's orbit [21]. According to Exoplanet.eu [57], the longest period planets discovered to date are Fomalhaut b (320000d) and Oph 11 b (730000d).

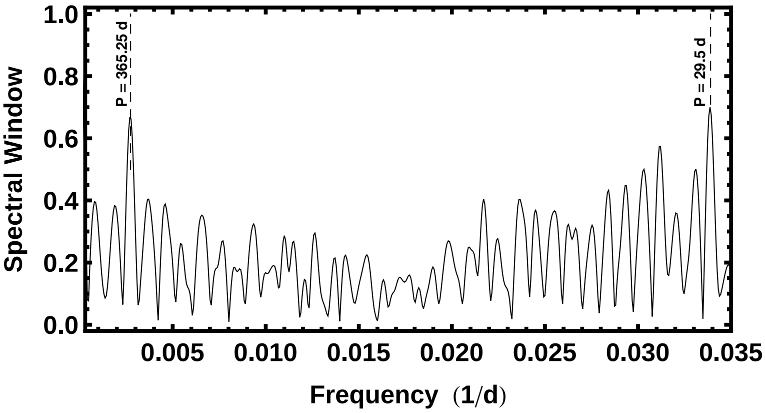


Figure B.1 A portion of the spectral window function of the radial velocity data for HD 208487 demonstrating the uniform spacing of peaks in frequency. The 29.5 d peak corresponds to the synodic month.

## B.2 Choice of frequency prior for multi-planet models

In this section we address the question of what prior to use for frequency for multi-planet models. For a single planet model we use a scale invariant prior because the prior period (frequency) range spans almost 6 decades. A scale invariant prior corresponds to a uniform probability density in $\ln f$. This says that the true frequency is just as likely to be in the bottom decade as the top. The scale invariant prior can be written in two equivalent ways.

$$p(\ln f|M_1, I)\, d\ln f = \frac{d\ln f}{\ln(f_H/f_L)} \tag{B.2}$$

$$p(f|M, I)\, df = \frac{df}{f\ \ln(f_H/f_L)} \tag{B.3}$$

What form of frequency prior should we use for a multiple planet model? We first develop the prior to be used in a frequency search strategy where we constrain the frequencies in an $n$ planet search such that $(f_L \leq f_1 \leq f_2 \cdots \leq f_n \leq f_H)$. From the product rule of probability theory and the above frequency constraints we can write

$$\begin{aligned} p(\ln f_1, \ln f_2, \cdots \ln f_n|M_n, I) &= p(\ln f_n|M_n, I) \\ &\times p(\ln f_{n-1}|\ln f_n, M_n, I) \cdots p(\ln f_2|\ln f_3, M_n, I) \\ &\times p(\ln f_1|\ln f_2, M_n, I). \end{aligned} \tag{B.4}$$

For model selection purpose we need to use a normalized prior which translates to the requirement that

$$\int_{\ln f_L}^{\ln f_H} p(\ln f_1, \ln f_2, \cdots \ln f_n | M_n, I) d \ln f_1 \cdots d \ln f_n = 1. \tag{B.5}$$

We assume that $p(\ln f_1, \ln f_2, \cdots \ln f_n | M_n, I)$ is equal to a constant $k$ everywhere within the prior volume. We can solve for $k$ from the integral equation

$$k \int_{\ln f_L}^{\ln f_H} d \ln f_n \int_{\ln f_L}^{\ln f_n} d \ln f_{n-1} \cdots \int_{\ln f_L}^{\ln f_2} d \ln f_1 = 1. \tag{B.6}$$

The solution to equation (B.6) is

$$k = \frac{n!}{[\ln(f_H/f_L)]^n}. \tag{B.7}$$

The joint frequency prior is then

$$p(\ln f_1, \ln f_2, \cdots \ln f_n | M_n, I) = \frac{n!}{[\ln(f_H/f_L)]^n} \tag{B.8}$$

Expressed as a prior on frequency, equation (B.7) becomes

$$p(f_1, f_2, \cdots f_n | M_n, I) = \frac{n!}{f_1 f_2 \cdots f_n \, [\ln(f_H/f_L)]^n} \tag{B.9}$$

We note that a similar result, involving the factor $n!$ in the numerator, was obtained by Bretthorst (2003) in connection with a uniform frequency prior.

Two different approaches to searching in the frequency parameters were employed in this work. In the first approach (a): an upper bound on $f_1 \leq f_2$ ($P_2 \geq P_1$) was utilized to maintain the identity of the two frequencies. In the second more successful approach (b): both $f_1$ and $f_2$ were allowed to roam over the entire frequency range and the parameters re-labeled afterwards. In this second approach nothing constrains $f_1$ to always be below $f_2$ so that degenerate parameter peaks can occur. For a two planet model there are twice as many peaks in the probability distribution possible compared with (a). For a $n$ planet model, the number of possible peaks is $n!$ more than in (a). Provided the parameters are re-labeled after the MCMC, such that parameters associated with the lower frequency are always identified with planet one and vice versa, the two cases are equivalent [2] and equation (B.8) is the appropriate prior for both approaches.

Approach (b) was found to be more successful because in repeated blind period searches it always converged on the highest posterior probability distribution peak, in spite of the huge period search range. Approach (a) proved to be unsuccessful in finding the highest peak in some trials and in those cases where it did find the

[2] To date this claim has been tested for $n \leq 3$.

peak it required many more iterations. Restricting $P_2 \geq P_1$ ($f_1 \leq f_2$) introduces an additional hurdle that appears to slow the MCMC period search.

## B.3 K Prior

The full set of priors used in our Bayesian calculations are given in Table 1.1. The limits on $K_i$ have evolved over time. Initially, the upper limit corresponded to the velocity of a planet with a mass = 0.01 M. in a circular orbit with a shortest period of one day or $K_{\max} = 2129$m s$^{-1}$. An upper bound of $K_{\max}\left(\frac{P_{\min}}{P_i}\right)^{1/3}$ was proposed at an exoplanet workshop at the Statistics and Applied Math Sciences Institute (spring 2006). Also an upper bound on $P_i$ of 1000 yr was suggested based on galactic tidal disruption. Previously we used an upper limit of three times the duration of the data. Again, we set $K_{\max} = 2129$m s$^{-1}$, which corresponds to a maximum planet-star mass ratio of 0.01.

Later, the upper limit on $K_i$ was set equal to

$$K_{\max}\left(\frac{P_{\min}}{P_i}\right)^{1/3}\frac{1}{\sqrt{1-e_i^2}}, \tag{B.10}$$

based on equation (B.11).

$$K = \frac{m\sin i}{M_*}\left(\frac{2\pi G M_*}{P}\right)^{1/3}\left(1+\frac{m}{M_*}\right)^{-2/3}, \tag{B.11}$$

where $m$ is the planet mass, $M_*$ is the star's mass, and $G$ is the gravitational constant. This is an improvement over $K_{\max}\left(\frac{P_{\min}}{P_i}\right)^{1/3}$ because it allows the upper limit on $K$ to depend on the orbital eccentricity. Clearly, the only chance we have of detecting an orbital period of 1000 yr with current data sets is if the eccentricity is close to one and we are lucky enough to capture periastron passage. All the calculations in this supplement are based on Equation (B.10).

## B.4 Eccentricity Prior

In the early years it was common to use a flat prior for eccentricity. It was soon realized that the effect of noise is to favour higher eccentricities. Gregory and Fischer [29] provided the following explanation of this bias. To mimic a circular velocity orbit the noise points need to be correlated over a larger fraction of the orbit than they do to mimic a highly eccentric orbit. For this reason it is more likely that noise will give rise to spurious highly eccentric orbits than low eccentricity orbits. In a related study, Shen (2008) [58] explored least-$\chi^2$ Keplerian fits to synthetic radial velocity data sets. They found that the best fit eccentricities for low signal-to-noise

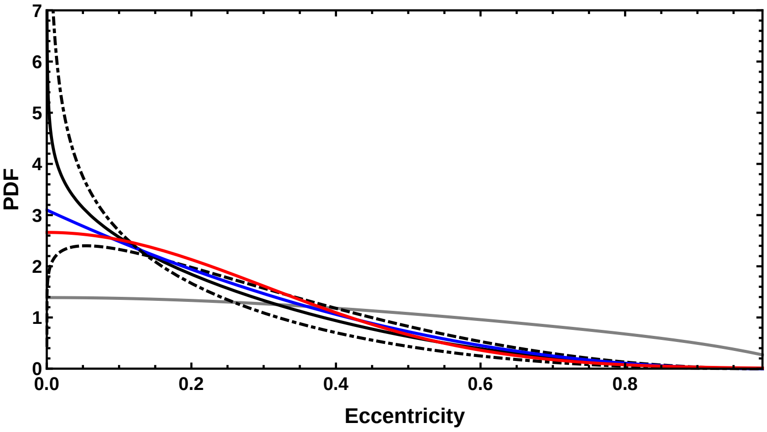


Figure B.2 Exoplanet eccentricity priors. The solid black curve is the best fit Beta distribution[40] to the eccentricity data of 396 high signal to noise exoplanets. The dashed and dot-dashed black curves are Kipping's Beta distribution fits to the subsets with periods > 382.3 d (median) and < 382.3 d, respectively. The red curve is the Gaussian eccentricity prior adopted by Tuomi et al. [62]. The gray curve is my earlier eccentricity prior which attempted a modest correction for noise induced eccentricity bias. The blue curve is eccentricity prior employed in this work.

ratio $K/\sigma \leq 3$ and moderate number of observations $\mathrm{N}_{\mathrm{obs}} \leq 60$, were systematically biased to higher values, leading to a suppression of the number of nearly circular orbits. More recently, Zakamska (2011) [69] found that eccentricities of planets on nearly circular orbits are preferentially overestimated, with typical bias of one to two times the median eccentricity uncertainty in a survey, e.g., 0.04 in the Butler et al. catalogue [8]. When performing population analysis, they recommend using the mode of the marginalized posterior eccentricity distribution to minimize potential biases.

Recently, Kipping (2013) [40] fit the eccentricity distribution of 396 exoplanets, detected through radial velocity with high signal-to-noise, with a Beta distribution. The Beta distribution can reproduce a diverse range of probability density functions (PDFs) using just two shape parameters (a and b). The black, large dash curve in Figure B.2 is the best fit Beta distribution (a = 0.867, b = 303) to all exoplanets in sample. The dot-dashed black curve is the best fit Beta distribution (a = 1.12, b =3.09) to exoplanets in sample with $P < 382.3$d. The dashed black curve is the best fit Beta distribution (a = 0.697, b = 3.27) to exoplanets in sample with $P > 382.3$d. One drawback with the first two of these Beta distributions is that they are infinite at e = 0, so it is necessary to work with a prior in the form of a cumulative density distribution (CDF) instead of a simpler PDF for MCMC work. The black curve is the eccentricity distribution utilized in this work which is another Beta Distribution (a = 1, b = 3.1) intermediate between Kipping's low and high period curves and is well behaved at e = 0. With these values the Beta distribution simplifies to

$$p(e|I) = 3.1(1 - e)^{2.1}. \tag{B.12}$$

The red curve is the Gaussian prior adopted by Tuomi et al (2012) [62]. The gray curve is my earlier eccentricity prior which attempted a modest correction for noise induced eccentricity bias.

# Appendix C

## Accuracy of *Mathematica* model Radial Velocities

As explained in Section 1.5, we convert the observation times, $t_i$, to orbital angles, $\theta_i$, by solving the conservation of angular momentum equation. As it stands equation 1.15 runs into problems when the period $P$ is small because *Mathematica's* NDSolve produces an interpolation function that spans the entire time range. If there are many cycles a very large number of iteration steps are required and the procedure slows to a crawl. To avoid this we first convert the $t_i$ to $q_i$, the corresponding fraction of one orbit, using the equation

$$q_i = mod[t_i/P + \chi, 1], \tag{C.1}$$

where again $\chi$ = the fraction of an orbit, prior to the start of data taking, that periastron occurred at. The relationship between $q_i$ and $\theta_i$ is given by

$$\frac{d\theta}{dq} - \frac{2\pi[1 + e\cos\theta(q)]^2}{(1 - e^2)^{3/2}} = 0. \tag{C.2}$$

Note: the relationship between $\theta$ and $q$ depends only on the eccentricity parameter, $e$.

*Mathematica* generates an accurate interpolating function between $q$ and $\theta$ so the differential equation does not need to be solved separately for each $q_i$. The solution of the differential equation is the largest component in the timing budget. Evaluating the interpolating function for each $q_i$ is very fast compared to solving the differential equation, so the algorithm should be able to handle much larger samples of radial velocity data than those currently available without a significant increase in computational time. For example, an increase in the sample size from 35 to 220 resulted in only an 18% increase in execution time. Of course, for a large enough sample size the interpolation operation will begin to dominate and after that the execution time will scale with the number of data points.

We now address the question of the accuracy of the model radial velocities which are limited by the accuracy of the *Mathematica* interpolating function. This was accomplished as follows:

1) Divide the interval $\theta = 0$ to $2\pi$ into $n = 10^7$ equal parts labeled $\theta_j$. Let $\theta_{acc}$ represent this set of accurate $\theta_j$ values.
2) Evaluate the corresponding accurate $RV_{acc}/K$ values, model radial velocities divided by $K$, where $RV_{acc}/K = [\cos(\theta_{acc} + \omega) + e \cos\omega]$.
3) Convert the $\theta_j$ values to $q_j$ values by computing the orbital area swept out between each pair of $\theta_j$ values. According to Kepler's Law of Areas, $dq_j$ is proportional to that area increment. The computed total area swept out in the interval $\theta = 0$ to $2\pi$ was found to agree with theory to better than 1 part in $10^{11}$.
4) The computed $q_j$ values were converted to a set of interpolated $\theta$ values ($\theta_{int}$) using *Mathematica's* solution of equation C.2.
5) Compute the corresponding set of interpolated $RV_{int}/K = [\cos(\theta_{int}{+}\omega){+}e\cos\omega]$.
6) Compute $\frac{1}{K}(RV_{int} - RV_{acc})$, the radial velocity error as a fraction of $K$.

$$\begin{aligned}\frac{1}{K}(RV_{int} - RV_{acc}) &= \cos(\theta_{int} + \omega) - \cos(\theta_{acc} + \omega) \\ &= \cos\omega\,(\cos\theta_{int} - \cos\theta_{acc}) \\ &\quad - \sin\omega\,(\sin\theta_{int} - \sin\theta_{acc})\end{aligned} \tag{C.3}$$

$$\text{Let } \theta_0 = \frac{(\theta_{int} + \theta_{acc})}{2} \text{ and } \delta\theta = (\theta_{int} - \theta_{acc}) \tag{C.4}$$

Then

$$\begin{aligned}\frac{1}{K}(RV_{int} - RV_{acc}) &= -2\ \sin\frac{\delta\theta}{2}\ \sin\theta_0\ \cos\omega \\ &\quad -2\ \sin\frac{\delta\theta}{2}\ \cos\theta_0\ \sin\omega \\ &= -2\ \sin\frac{\delta\theta}{2}\ \sin(\theta_0 + \omega) \\ &\approx -\delta\theta\ \sin(\theta_0 + \omega)\end{aligned} \tag{C.5}$$

For any $\theta_0$, the fractional model radial velocity error has a maximum positive or negative value for $\sin(\theta_0 + \omega) = \pm 1$. Fig. C.1 show plots of the $\mathrm{Log}_{10}$ of the absolute magnitude of the fractional error versus $q$ for three different values of eccentricity. The figure assumes a worst case value for $|\sin(\theta_0 + \omega)| = 1$. More realisticaly, these errors should be reduced by a factor $2/\pi$ which is expectation value of $|\sin(\theta_0 + \omega)|$. Even for $e = 0.99$ the fractional error is $< 10^{-5}$ over most of the $q$ range only rising above this towards the very end of the interpolation interval. The bottom panel shows the maximum value of the fractional error versus eccentricity. Based on this analysis, the maximum error in the *Mathematica* derived model radial velocities, expressed as a fraction of the $K$ parameter, is $\leq 2.2 \times 10^{-5}$ for $e$ values in the range 0 to 0.8. The situation degrades progressively for larger values $e$ but is still $\leq 2.8 \times 10^{-3}$ for $e = 0.98$, rising to $1.2 \times 10^{-2}$ for $e = 0.99$.

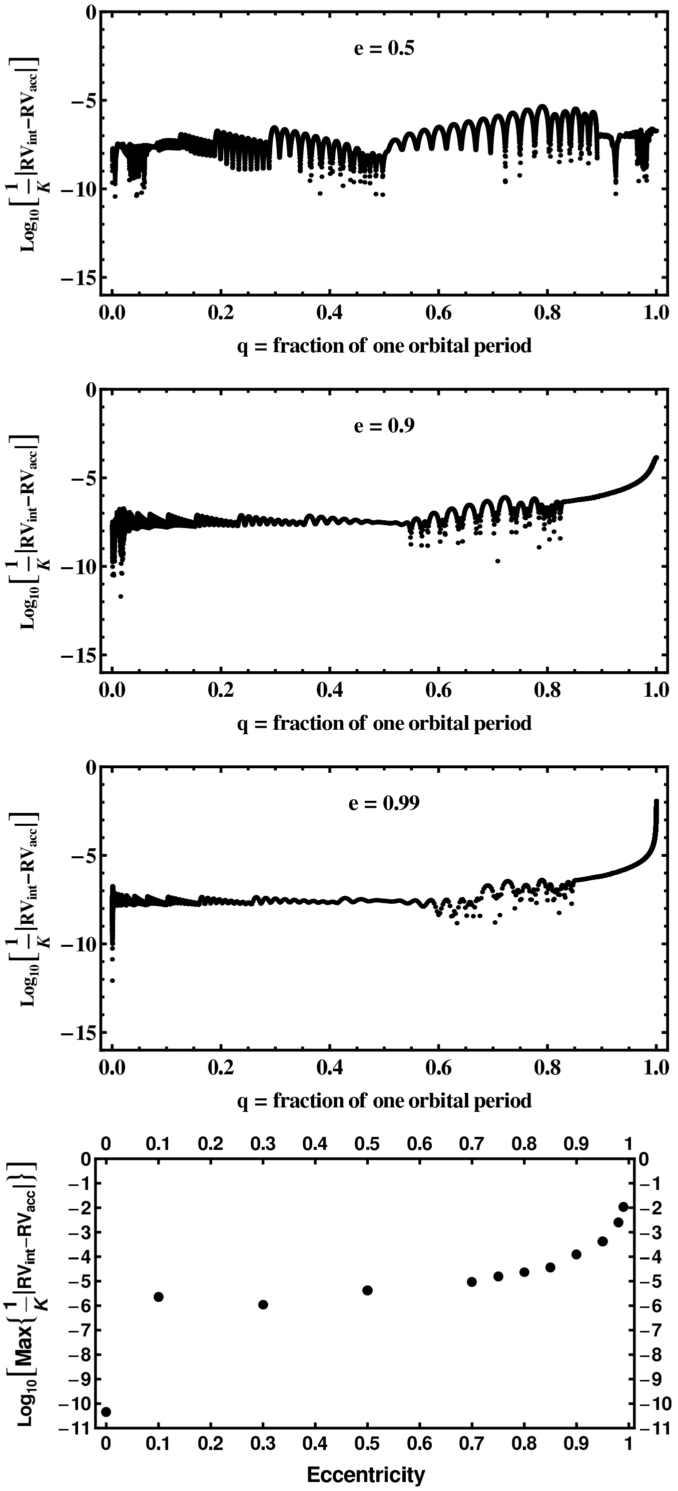


Figure C.1 The top three panels show plots of the $\text{Log}_{10}$ of the absolute magnitude of the model radial velocity error as a fraction of the $K$ value versus $q$, the fraction of the orbital period, for three different values of eccentricity. The bottom panel shows $\text{Log}_{10}$ of the maximum value of the above fractional radial velocity error versus eccentricity.

# Index