ANNUAL REVIEWS

Annual Review of Astronomy and Astrophysics

Markov Chain Monte Carlo Methods for Bayesian Data Analysis in Astronomy

Sanjib Sharma

Sydney Institute for Astronomy, School of Physics, University of Sydney, NSW 2006, Australia; email: sanjib.sharma@gmail.com

Annu. Rev. Astron. Astrophys. 2017. 55:213–59

First published as a Review in Advance on June 21, 2017

The *Annual Review of Astronomy and Astrophysics* is online at astro.annualreviews.org

https://doi.org/10.1146/annurev-astro-082214-122339

Keywords

methods: data analysis, numerical statistical

Abstract

Markov chain Monte Carlo–based Bayesian data analysis has now become the method of choice for analyzing and interpreting data in almost all disciplines of science. In astronomy, over the past decade, we have also seen a steady increase in the number of papers that employ Monte Carlo–based Bayesian analysis. New, efficient Monte Carlo–based methods are continuously being developed and explored. In this review, we first explain the basics of Bayesian theory and discuss how to set up data analysis problems within this framework. Next, we provide an overview of various Monte Carlo–based methods for performing Bayesian data analysis. Finally, we discuss advanced ideas that enable us to tackle complex problems and thus hold great promise for the future. We also distribute downloadable computer software (**https://github.com/sanjibs/bmcmc**) `Python` that implements some of the algorithms and examples discussed here.

Contents

1. INTRODUCTION

Markov chain Monte Carlo (MCMC) and Bayesian statistics are two independent disciplines, the former being a method to sample from a distribution while the latter is a theory to interpret observed data. When these two disciplines are combined together, the effect is so dramatic and powerful that it has revolutionized data analysis in almost all disciplines of science, and astronomy is no exception. This review explores the power of this combination.

What is so special about MCMC-based Bayesian data analysis? The usefulness of Bayesian methods in science and astronomy is easy to understand. In many situations, it is easy to predict the outcome given a cause. But in science, most often, we are faced with the opposite question. Given the outcome of an experiment, what are the causes or what is the probability of a cause

as compared with some other cause? If we have some prior information, how does that help us? This opposite problem is more difficult to solve. The power of Bayesian theory lies in the fact that it provides a unified framework to quantitatively answer such questions. Hence it has become indispensable for science. As opposed to deductive logic, Bayesian theory provides a framework for plausible reasoning, a concept that is more powerful and general and an idea championed by Jaynes (2003) in his book.

The question now is how does one solve a problem that has been set up using Bayesian theory. This mostly involves computing the probability distribution function (pdf) of some parameters given the data and is written as $p(\theta|D)$. Here, θ need not be a single parameter; in general, it represents a set of parameters. Usually here and elsewhere, such functions do not have analytical solutions, and so we need methods to numerically evaluate the distribution. This is where MCMC methods come to the rescue. They provide an efficient and easy way to sample points from any given distribution, which is analogous to evaluating the distribution.

Bayesian data analysis (Jeffreys 1939) and MCMC (Metropolis et al. 1953) techniques have existed for more than 50 years. Their tremendous increase in popularity over the past decade is due to an increase in computational power that has made it affordable to do such computations.

The simplest and the most widely used MCMC algorithm is the "random walk" Metropolis algorithm (Section 3.2). However, the efficiency of this algorithm depends upon the "proposal distribution," which the user has to supply. This means that there is some problem-specific fine tuning to be done by the user. The problem to find a suitable proposal distribution becomes worse as the dimensionality of the space over which the sampling is done increases. Correlations and degeneracies between the coordinates further exacerbate the problem. Many algorithms have been proposed to solve this problem, and it remains an active area of research. Some algorithms work better for specific problems and under special conditions, but algorithms that work well in general are in high demand. Multimodal distributions pose additional problems for MCMC algorithms. In such situations, an MCMC chain can easily get stuck at a local density maximum. To overcome this, algorithms like simulated tempering and parallel tempering have been proposed (Section 3.8). Hence discussion of efficient MCMC algorithms is one focus of this review.

Given its general applicability, the Bayesian framework can be used in almost any field of astronomy. Hence, it is not possible to discuss all its applications. However, there are many examples in which alternatives either do not exist or are inferior. The aim of this review is to specifically discuss such cases where Bayesian-MCMC methods have enjoyed great success. The Bayesian framework by itself is very simple. The difficult part when attempting to solve a problem is to express the problem within this framework and then to choose the appropriate MCMC method to solve it. The best way to master this is by studying a diverse set of applications, and we aim to provide this in our review (Section 5). Finally, we also discuss a few advanced topics like nonparametric models and hierarchical Bayesian models (Section 4) that are not yet mainstream in astronomy but are very powerful and allow one to solve complex problems.

To summarize, the review has three main aims. The first is to explain the basics of Bayesian theory using simple familiar problems, e.g., fitting a straight line to a set of data points with errors both in the coordinates and in the presence of outliers. This is targeted at readers who are new to the topic. The second goal is to provide a concise overview of recent developments. This will benefit people who are familiar with Bayesian data analysis but are interested in learning more. The final aim is to discuss emerging ideas that hold great promise in the future. We also develop and distribute downloadable software (available at **https://github.com/sanjibs/bmcmc** or by running the command "pip install bmcmc") implementing some of the algorithms and examples that we discuss.

1.1. The Rise of MCMC-Based Bayesian Methods in Astronomy and Science

The emergence of Bayesian statistics has a long and interesting history dating back to 1763 when Thomas Bayes laid down the basic ideas of his new probability theory (Bayes & Price 1763, published posthumously by Richard Price). It was rediscovered independently by Laplace (de Laplace 1774) and used in a wide variety of contexts, e.g., celestial mechanics, population statistics, reliability, and jurisprudence. However, after that it was largely ignored. A few scientists, like Bruno de Finetti and Harold Jeffreys, kept the Bayesian theory alive in the first half of the twentieth century. Harold Jeffreys published the book *Theory of Probability* (Jeffreys 1939), which for a long time remained the main reference for using Bayes's theorem. Bayes's theorem was used in the Second World War at Bletchley Park, United Kingdom, for cracking the German Enigma code, but its use remained classified for many years afterward. From 1950 onward, the tide turned toward Bayesian methods. However, the lack of proper tools to do Bayesian inference remained a challenge. The frequentist methods in comparison were simpler to implement, which made them more popular. [A recent statement by the American Statistical Association (Wasserstein & Lazar 2016) warning on the misuse of P values is another example of the superiority of the Bayesian methods of hypothesis testing.]

Interestingly, efficient methods like MCMC to sample distributions had been invented by 1954 in the context of solving problems in statistical mechanics (Metropolis et al. 1953). [The brand name Monte Carlo was coined by Metropolis & Ulam (1949), where they discussed a stochastic method making use of random numbers to solve a class of problems in mathematical physics that are difficult to solve owing to the large number of dimensions.] Such problems typically involve N interacting particles. A single configuration ω of such a system is fully specified by giving the position and velocity of all the particles; i.e., ω can be defined by a point in $\mathcal{R}^{2N}$ space, also known as the configuration space Ω. The total energy is a function of the configuration $E(\omega)$. For a system in equilibrium, the probability of a configuration is given by $\exp[-E(\omega)/kT]$, where k is the Boltzmann constant and T is the temperature of the system. Computing any thermodynamic property of the system, e.g., pressure or energy, typically involves computing integrals of the form

$$\bar{F} = \frac{\int F(\omega)\exp[-E(\omega)/kT]\mathrm{d}\omega}{Z}, \tag{1}$$

for which $Z = \int \exp[-E(\omega)/kT]\mathrm{d}\omega$ is known as the partition function. The integrals over ω are in most cases analytically and computationally intractable. The idea of Metropolis and colleagues was to start with an arbitrary configuration of N particles and then move each particle by a random walk. If $\Delta E < 0$, the move is always accepted; otherwise, it is accepted stochastically with probability $\exp(-\Delta E/kT)$, which is the ratio of the probability of the new configuration with respect to the old. The method ends up choosing a configuration ω sampled from $\exp[-E(\omega)/kT]$. The method immediately became popular in the statistical physics community.

However, the fact that the same method can be used for sampling an arbitrary pdf $p(\omega)$ by simply replacing $E(\omega)/kT$ with $\ln[p(\omega)]$ had to wait till the important paper by Hastings (1970). He generalized the work of Metropolis and colleagues and derived the essential condition for the acceptance ratio that a Markov chain ought to satisfy in order to sample the target distribution. The generalized algorithm is now known as the Metropolis–Hastings (MH) algorithm. Later, Hastings's student Peskun showed that, among the available choices, the one by Metropolis and colleagues was the most efficient (Peskun 1973). Despite their introduction to the statistical community, the ideas remained dormant till 1980.

Around 1980 things suddenly changed, and a few influential algorithms appeared. Simulated annealing was presented by Kirkpatrick et al. (1983) to solve combinatorial optimization problems using the MH algorithm in conjunction with ideas of annealing from solid-state physics. It is

especially useful for situations in which we have multiple maxima and applies to any setting where we have to minimize an objective function $C(\omega)$. This is done by sampling $\exp[-C(\omega)/T]$, with progressively decreasing T to allow annealing and selection of a globally optimum solution. A year later Geman & Geman (1984) introduced what we currently know as Gibbs sampling in the context of image restoration. This was the first proper use of MCMC techniques to solve a problem set up in a Bayesian framework, in the sense that simulating from conditional distributions is the same as simulating from the joint distribution. However, there exists earlier work related to Gibbs sampling; the Hammersley–Clifford theorem, which was developed in the early 1970s, and the work by Besag (1974).

At about this time, one of the most influential methods of the twentieth century emerged—the expectation maximization (EM) algorithm by Dempster et al. (1977). This provided a way to deal with missing data and hidden variables and vastly increased the range of problems that can be addressed by Bayesian methods. The EM algorithm is deterministic and has some sensitivity to the starting configuration. To address this, stochastic versions were developed (Celeux & Diebolt 1985), quickly followed by the data augmentation (DA) algorithm (Tanner & Wong 1987).

The watershed moment in the field of statistics is largely credited to the paper by Gelfand & Smith (1990) that unified the ideas of Gibbs sampling, the DA algorithm, and the EM algorithm (Tanner & Wong 2010, Robert & Casella 2011). It firmly established that Gibbs sampling and MH-based MCMC algorithms can be used to solve a wide class of problems that fall into the category of hierarchical Bayesian models. The citation history of the famous Metropolis et al. (1953) paper shown in **Figure 1** corroborates the historical narrations on this topic. In physics, the MH algorithm was well known during the period 1970–1990, but this was not so in statistics

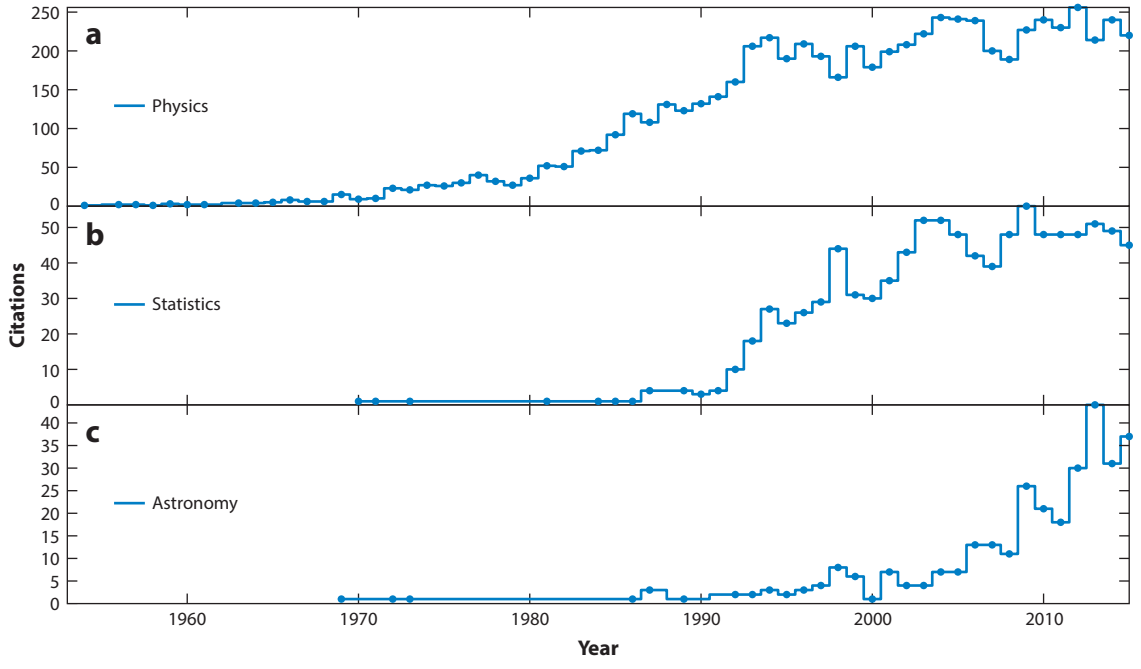

Figure 1

Citation history for the Metropolis et al. (1953) paper for three different subject areas.

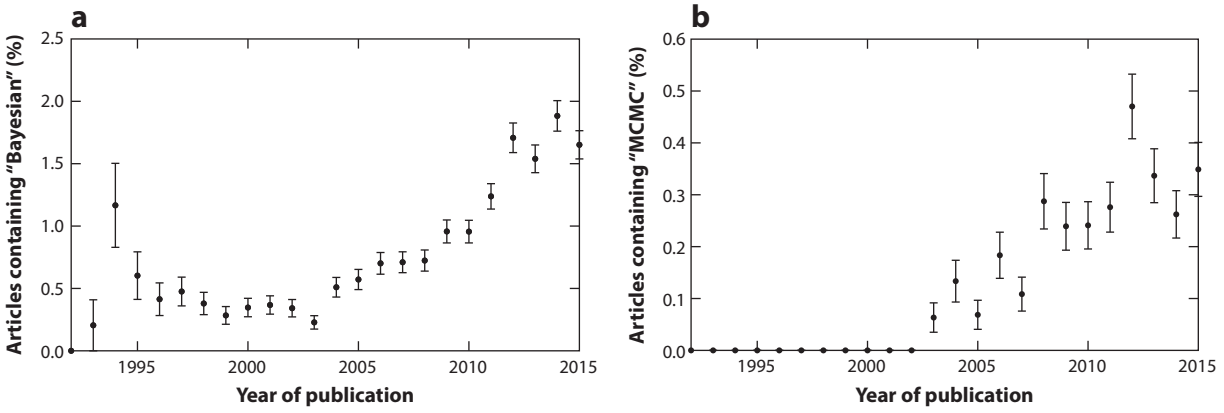

Figure 2

Percentage of articles in ArXiv astro-ph abstracts containing the terms (*a*) Bayesian and (*b*) MCMC. Computed using the code *arxiv.py*; courtesy of Dustin Lang.

or astronomy. In astronomy, a watershed moment can be seen in 2002; this is visible more clearly in **Figure 2** in which we track the usage of the terms MCMC and Bayesian.

But prior to 2002, the Bayesian-MCMC technique was not unknown to the astronomy community. We can see its use in Saha & Williams (1994) who applied it to extract galaxy kinematics from absorption line spectra. Further seeds were planted down the line by Christensen & Meyer (1998) while studying gravitational wave radiation, and then by Christensen et al. (2001) and Knox et al. (2001) in the context of cosmological parameter estimation using cosmic microwave background data. Inspired by these papers, Lewis & Bridle (2002) more than any other paper seems to have galvanized the astronomy community in the use of Bayesian and MCMC techniques. They laid out in detail the Bayesian-MCMC framework, applied it to one of the most important data sets of the time (cosmic background radiation) and used it to address a significant scientific question—the fundamental parameters of our Universe. Additionally, they made their MCMC code publicly available, which was instrumental in lowering the barrier for new entrants to the field.

2. BAYESIAN DATA ANALYSIS

In this section, we briefly review the basics of Bayesian theory. We start with Bayes's theorem and then use it to set up the problem of fitting a model to data. This is followed by a discussion of the role of priors in Bayesian analysis. Next, the Bayesian solution of fitting a straight line is discussed in detail to illustrate the ideas discussed. Finally, we show how to perform model selection. To further explore the topics discussed here, many excellent resources are available. A stimulating discussion on Bayesian theory can be found in Jaynes (2003). Gregory (2005), Sivia & Skilling (2006), Gelman et al. (2013), and Andreon & Weaver (2015) are excellent textbooks with a good emphasis on applications in science. Hogg et al. (2010) provides lucid tutorial on fitting models to data. A fascinating discussion on Bayesian versus frequentist approaches to solving problems can be found in Loredo (1990). A review with emphasis on cosmology is given by Trotta (2008).

2.1. Bayes's Theorem

Cox (1946) showed that the rules of Bayesian probability theory can be derived from just two basic rules:

$$p(H\,|I) + p(\bar{H}\,|I) = 1 \qquad \text{Sum Rule,} \tag{2}$$

$$p(H, D|I) = p(H\,|D, I)p(D|I) = p(D|H, I)p(H\,|I) \quad \text{Product Rule.} \tag{3}$$

Here H stands for some proposition being true and D stands for some other proposition being true, and $\bar{H}$ means the proposition H is false. So the sum rule just states that the probability of a proposition being true plus the probability of it being false is unity. The product rule expresses the joint probability of two propositions being true in terms of conditional probabilities, one being true given the other is true. The vertical bar | is a conditioning symbol and means "given." I denotes relevant background information that is used to construct the probabilities. The product rule leads to Bayes's theorem,

$$p(H\,|D, I) = \frac{p(D|H, I)p(H\,|I)}{p(D|I)}, \quad \text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}, \tag{4}$$

where we identify H with the hypothesis and D with the data. The $p(D|H, I)$ is the probability of observing the data D if the hypothesis is true and is known as the likelihood. The quantity $p(H\,|I)$ is the prior and specifies our prior knowledge of H being true. The $p(H\,|D, I)$, known as posterior, expresses our updated belief about the truth of the hypothesis in light of the data D. The quantity $p(D|I)$ is a constant and serves the purpose of normalizing $\int p(H\,|D, I)\,\mathrm{d}H$ to 1. It is known as the evidence.

Another important result that can be derived from the sum rule and the product rule is the marginalization equation,

$$p(X\,|I) = \int p(X, Y\,|I)\mathrm{d}Y = \sum_i p(X, Y_i|I). \tag{5}$$

First let us write the sum rule in an alternate form. Instead of considering just Y and $\bar{Y}$, we consider a set of possibilities $\{Y_i\}$ that are mutually exclusive.

$$\sum p(Y_i|I) = 1, \quad \text{Extended Sum Rule.} \tag{6}$$

Now, making use of the product rule and the sum rule we get

$$\sum_i p(X, Y_i|I) = \sum_i p(Y_i|X, I)p(X\,|I) \qquad \text{using product rule} \tag{7}$$

$$= p(X\,|I)\sum_i p(Y_i|X, I) = p(X\,|I) \qquad \text{using sum rule.} \tag{8}$$

2.2. Fitting a Model to Data

Typically, we have some data and we want to use them for scientific inference. One of the most effective approaches to dealing with such problems is to develop a model that describes how the data were created. Let θ be the set of parameters of the model and x^t a data point generated by the model according to $f(x^t|\theta)$. The observed data points x can have some measurement errors, described by a parameter σ_x. The probability of the observed value is then given by $p(x|x^t, \sigma_x)$, which could be $\mathcal{N}(x|x^t, \sigma_x^2)$ for Gaussian errors; hereafter, $\mathcal{N}(.|\mu, \sigma^2)$ refers to a normal distribution with mean μ and variance σ^2. The probability of observed data point x given a model and an error is then

$$p(x|\theta, \sigma_x) = \int f(x^t|\theta)p(x|x^t, \sigma_x)\mathrm{d}x^t. \tag{9}$$

We have integrated over true values x^t, which are unknown.

If we have reason to believe that there are outliers in the data, e.g., a fraction of points are not described by the model, we can supplement a background model $f_b(x^t|\theta_b)$ with probability P_b and

parameters θ_b (Press 1997, Hogg et al. 2010). The probability of the observed data points can then be written as

$$p(x|\theta, \theta_b, P_b, \sigma_x) = \int \left[P_b\, f_b(x^t|\theta_b) + (1 - P_b) f(x^t|\theta)\right] p(x|x^t, \sigma_x) \mathrm{d}x^t \tag{10}$$

$$= p(x|\theta_b) P_b + p(x|\theta)(1 - P_b). \tag{11}$$

The total probability for a set of N data points $X = \{x_1, \ldots, x_N\}$ is then

$$p(X|\theta, \theta_b, P_b, \sigma_x) = \prod_{i=1}^{N} p(x_i|\theta, \theta_b, P_b, \sigma_{x,i}). \tag{12}$$

To infer the model parameters, one uses Bayes's theorem and computes

$$p(\theta, \theta_b, P_b|X, \sigma_x) \propto p(X|\theta, \theta_b, P_b, \sigma_x) p(\theta, \theta_b, P_b). \tag{13}$$

Here, $p(\theta, \theta_b, P_b)$ represents our prior knowledge about the parameters. We discuss this in detail in the next section.

We consider the problem of fitting a straight line with equation $y = mx + c$ to some data points $X = \{x_1, \ldots, x_N\}$ and $Y = \{y_1, \ldots, y_N\}$, with uncertainty $\sigma_{y,i}$ on the y ordinate. We generated 50 data points with $m = 2.0$ and $c = 10.0$; 20% of the data points were set as outliers and were sampled from $\mathcal{N}(30, 5^2)$. To simulate random uncertainty the y ordinate was scattered with a Gaussian function having dispersion in the range $0.25 < \sigma_y < 1.25$. The data along with the results of our fitting exercise are shown in **Figure 3**. The image shows the outliers and data sampled from a straight line. We first fitted a simple model without taking the outliers into account. Here, $\theta = \{m, b\}$, and the generative model of the data is

$$p(y_i|m, c, x_i, \sigma_{y,i}) = \frac{1}{\sqrt{2\pi}\sigma_{y,i}} \exp\left[-\frac{(y_i - mx_i - b)^2}{2\sigma_{y,i}^2}\right]. \tag{14}$$

It can be seen that the best-fit line is not a good description for the data points that were sampled from a straight line. Next, we extended the model by adding a model for the outliers as

$$p(y_i|\mu_b, \sigma_b, x_i, \sigma_{y,i}) = \frac{1}{\sqrt{2\pi(\sigma_{y,i}^2 + \sigma_b^2)}} \exp\left[-\frac{(y_i - \mu_b)^2}{2(\sigma_{y,i}^2 + \sigma_b^2)}\right]. \tag{15}$$

The full model is

$$p(Y|m, c, P_b, \mu_b, \sigma_b, X, \sigma_y) = \prod_{i=1}^{N} [p(y_i|\mu_b, \sigma_b, x_i, \sigma_{y,i})(P_b) + p(y_i|m, c, x_i, \sigma_{y,i})(1 - P_b)]. \tag{16}$$

The best-fit line resulting from this model obtained by sampling the posterior distribution using an MCMC scheme is shown in **Figure 3**. The best-fit parameters of the model resemble well the true parameters that were used to create the synthetic data set (the example is implemented in the software that we provide).

2.3. Priors

Priors are one of the most important ingredients of the Bayesian framework. Priors express our present state of knowledge about the parameters of interest, which we wish to constrain by analyzing new data. In a multidimensional parameter space, it is quite common to have degeneracies among the parameters. Here priors can play a crucial rule in restricting the posterior to a small region of the parameter space as compared with the much larger region allowed by the likelihood

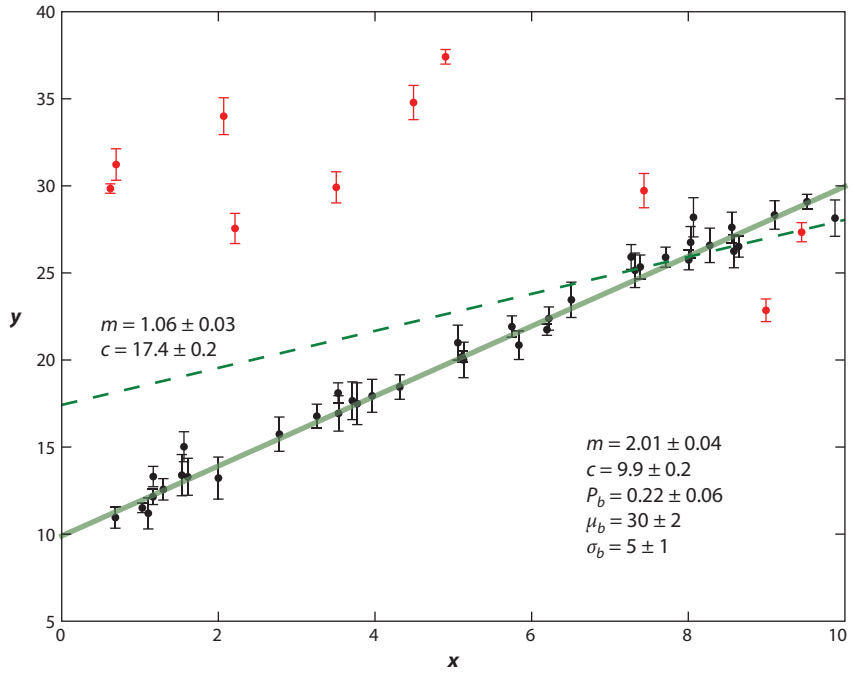

Figure 3

Fitting a straight line to data with outliers. The outliers are shown as red points. The dashed line is the best-fit line when an outlier model is not used. The solid line is the best-fit line with an outlier model. The data were generated with model parameters $m = 2$ and $c = 10$. Some 20% of the points were set as outliers and sampled from $\mathcal{N}(30, 5^2)$.

function. Priors can be broadly classified into two types, uninformative and informative. Uninformative priors express our state of ignorance and have very little restricting power. They are also known as ignorance priors. Typically their distributions are diffuse. By contrast, informative priors are typically very restricting. They might come from the analysis of some previous data. They are important when the data alone are not very informative, and without external information the data cannot adequately constrain the parameters being investigated.

Ignorance priors are used in cases where we have very little knowledge about the parameters we want to constrain, and we wish to express our ignorance by using uninformative priors. Certainly a prior with sudden jumps or oscillating features is too detailed for expressing ignorance! So smoothness is certainly an important criterion for an ideal uninformative prior. In fact, if the data are informative, almost any prior that is sufficiently smooth in the region of high likelihood will lead to very similar conclusions. Is there a formal and unique way to express our ignorance?

A number of techniques exist for constructing ignorance priors. We here discuss a few simple and commonly used techniques; for a detailed review see Kass & Wasserman (1996). The simplest is Laplace's principle of insufficient reason, which assigns equal probability to all possible values of the parameter. If the parameter space consists of a finite set of points, then it is easy to apply this principle. However, for a continuous parameter space, the prior depends upon the chosen partitioning scheme.

Ignorance priors can also be specified using the invariance of the likelihood function, $p(x'|\theta')\mathrm{d}x' = p(x|\theta)\mathrm{d}x$, under the action of a transformation group $(x', \theta') = h(x, \theta)$, e.g.,

translation, scaling, or rotation of coordinates. If the priors are really uninformative, consistency demands that we should make the same Bayesian inference, which implies that the priors should also be invariant to the transformation and satisfy $p(\theta')\mathrm{d}\theta' = p(\theta)\mathrm{d}\theta$ (Jaynes 2003). For two special types of parameters, this leads to unique choices for expressing ignorance. These are the location parameters and the scale parameters. An example is the mean μ and dispersion σ of a normal distribution $\mathcal{N}(x|\mu,\sigma^2)$, which are the location and the scale parameters, respectively. The likelihood $\mathcal{N}(x|\mu,\sigma^2)$ is invariant under transformation $(x',\mu') = (x+b,\mu+b)$, demanding that invariance for the prior leads to $p(\mu) =$ constant. Similarly, $\mathcal{N}(x|\mu,\sigma^2)$ is also invariant under $(x'-\mu',\sigma') = [a(x-\mu),a\sigma]$, which leads to $p(\sigma) \propto 1/\sigma$. In general, μ and σ are location and scale parameters, respectively, if likelihood is of the form $f[(x-\mu)/\sigma]/\sigma$.

Another commonly used technique to specify ignorance priors is the Jeffreys rule,

$$p(\theta) \propto \det[\mathcal{I}(\theta)]^{1/2}, \text{ where } [\mathcal{I}(\theta)]_{ij} = \int p(x|\theta)\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln p(x|\theta)\mathrm{d}x \tag{17}$$

is the Fisher information matrix and θ a vector of parameters. It is based on the idea that the prior should be invariant to reparameterization of θ. Applying it to the case where the likelihood is a normal distribution, $\mathcal{N}(x|\mu,\sigma^2)$, gives $p(\mu) =$ constant (for a fixed σ) and $p(\sigma) = 1/\sigma$ (for a fixed μ). However, when applied to both μ and σ together, it gives $p(\mu,\sigma) = 1/\sigma^2$. To avoid this contradiction the rule was modified to

$$p(\mu_1,\ldots,\mu_k,\theta) \propto \det[\mathcal{I}(\theta)]^{1/2}, \tag{18}$$

where μ_i are location parameters and $\mathcal{I}(\theta)$ is calculated keeping them fixed.

The principle of maximum entropy (Jaynes 1957) is also helpful for selecting priors. Suppose we are interested in knowing the pdf of a variable, e.g., the probability of a given face of a six-faced die landing up. Suppose we also have some macroscopic constraint available to us, e.g., the mean value obtained when the die is rolled a large number of times. Such a constraint cannot uniquely identify a pdf but can be used to rule out a number of pdfs. The principle says that out of all possible pdfs satisfying the constraint, the most likely one is the one having maximum entropy, where the entropy is defined as $S = -\sum p_i \log p_i$. We now use this principle to derive the most likely distribution of a variable for two common cases.

- If for a variable x we know the expectation value $\bar{x}$ and the fact that it lies in the range $[0,\infty]$, then the maximum entropy distribution of x is $p(x|\bar{x}) = \exp(-x/\bar{x})/\bar{x}$.
- If $\bar{x}$ and variance $\sigma^2 = \langle(x-\bar{x})^2\rangle$ are known, then $p(x|\bar{x},\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp[-\frac{(x-\bar{x})^2}{2\sigma^2}]$.

2.4. Fitting a Straight Line

We now consider the Bayesian solution for fitting a straight line in detail (see also Jaynes 1999, Hogg et al. 2010). We first discuss the solution for the general case in which we have uncertainties on both x and y coordinates and then discuss the case in which the uncertainties are unknown. Suppose we have a collection of points, $(X = \{x_1,\ldots,x_N\}, Y = \{y_1,\ldots,y_N\})$, with uncertainties $\mathbf{\Sigma} = \{\mathbf{\Sigma}_1,\ldots,\mathbf{\Sigma}_N\}$. Here $\mathbf{\Sigma}_i$ is the covariance matrix defined as

$$\mathbf{\Sigma}_i = \begin{bmatrix} \sigma_{x,i}^2 & \sigma_{xy,i}^2 \\ \sigma_{xy,i}^2 & \sigma_{y,i}^2 \end{bmatrix}. \tag{19}$$

We want to fit a line $y = ax + b$ to these data. For the time being, we assume $\mathbf{\Sigma}_i$ to be a diagonal matrix with $\sigma_{xy,i} = 0$. Let (x,y) be the true values corresponding to the point (x_i,y_i). Then the

probability of measuring the point (x, y) at (x_i, y_i) is

$$p(x, y|x_i, y_i, \sigma_{x,i}, \sigma_{y,i}) = \frac{1}{2\pi\sigma_{x,i}\sigma_{y,i}} \exp\left[-\frac{(x-x_i)^2}{2\sigma_{x,i}^2} - \frac{(y-y_i)^2}{2\sigma_{y,i}^2}\right]. \quad (20)$$

Let us consider a generative model for the line. We consider the pdf of a line to be described by a Gaussian with width σ_p along a direction perpendicular to the line and width σ_b along the line. Here, σ_p can be thought of as an intrinsic scatter about the linear relation that we wish to investigate. In the limit $\sigma_b \to \infty$, the probability of a point (x, y) to be sampled from this generative model is

$$p(x, y|a, b, \sigma_p) = \frac{1}{\sqrt{2\pi}\sigma_p} \exp\left\{-\frac{[y-(ax+b)]^2}{(1+a^2)2\sigma_p^2}\right\}. \quad (21)$$

Hence, the probability of (x_i, y_i) being sampled from the generative model of the line is

$$p(x_i, y_i|a, b, \sigma_{x,i}, \sigma_{y,i}, \sigma_p) = \int\int p(x, y|a, b, \sigma_p) p(x, y|x_i, y_i, \sigma_{x,i}, \sigma_{y,i})\, \mathrm{d}x\mathrm{d}y \quad (22)$$

$$= \frac{1}{\sqrt{2\pi(\sigma_{\perp,i}^2 + \sigma_p^2)}} \exp\left[-\frac{d_i^2}{2(\sigma_{\perp,i}^2 + \sigma_p^2)}\right], \quad (23)$$

where $\sigma_{\perp,i} = \sqrt{(\sigma_{y,i}^2 + a^2\sigma_{x,i}^2)/(1+a^2)}$ is the component of the error vector perpendicular to the line, and $d_i = (y_i - ax_i - b)/\sqrt{1+a^2}$ is the perpendicular distance of the point from the line. For a general matrix $\mathbf{\Sigma}_i$, $\sigma_{\perp,i} = \hat{\mathbf{u}}^T \mathbf{\Sigma}_i \hat{\mathbf{u}}$ for which $\hat{\mathbf{u}} = (-a/\sqrt{1+a^2}, 1/\sqrt{1+a^2})$ is a unit vector perpendicular to the line. For the full sample,

$$p(X, Y|\mathbf{\Sigma}, a, b, \sigma_p) = \prod_i^N \frac{1}{\sqrt{2\pi(\sigma_{\perp,i}^2 + \sigma_p^2)}} \exp\left[-\frac{d_i^2}{2(\sigma_{\perp,i}^2 + \sigma_p^2)}\right]. \quad (24)$$

If we desire to compute a and b, then

$$p(a, b|X, Y, \mathbf{\Sigma}) = Ap(a, b)p(X, Y|\mathbf{\Sigma}, a, b). \quad (25)$$

Henceforth, A is a normalization constant that may be different in different equations. The $p(a, b)$ is the prior distribution of parameters of the line. The two common choices for the prior are the uniform (flat) and Jeffreys prior. Neither is appropriate. Given the rotational symmetry in the problem, a sensible choice is to have priors that are symmetric with respect to rotation. Let $\theta = \tan^{-1}(a)$ be the angle made by the line with the x axis, and $b_\perp = b\cos(\theta)$ be the distance of the line from the origin. A uniform prior on θ and $b_\perp$ is symmetric with respect rotation (Dose 2002). This leads to

$$p(a, b)\,\mathrm{d}a\,\mathrm{d}b = \frac{\mathrm{d}\theta}{\pi}\frac{\mathrm{d}b_\perp}{2B_\perp} = \frac{1}{(1+a^2)^{3/2}}\frac{\mathrm{d}a\,\mathrm{d}b}{2B_\perp\pi}. \quad (26)$$

In **Figure 4**, we graphically show how a prior uniform in a differs from a prior uniform in $\theta = \tan^{-1}(a)$. In **Figure 4*a***, we show straight lines uniformly spaced in a. The lines tend to crowd at high values of a, and this can bias the estimate of the slope a. In **Figure 4*b***, the lines are uniformly spaced in θ, and there is no crowding effect.

The log-likelihood of the full solution after taking the prior into account is

$$\ln L = \ln p(a, b|\{x_i\}, \{y_i\}, \mathbf{\Sigma}_i, \sigma_p) \quad (27)$$

$$= K - \frac{3}{2}\ln(1+a^2) - \sum_{i=1}^N \frac{1}{2}\ln(\sigma_{\perp,i}^2 + \sigma_p^2) - \sum_{i=1}^N \frac{d_i^2}{\sigma_{\perp,i}^2 + \sigma_p^2}. \quad (28)$$

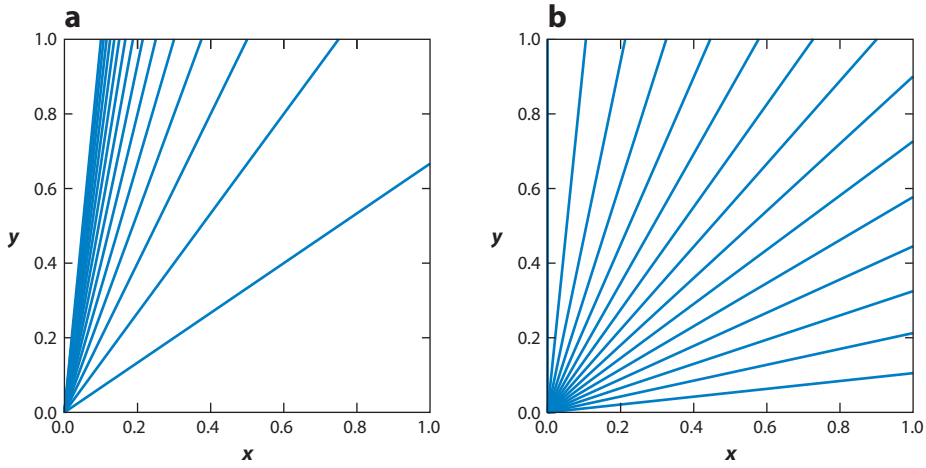

Figure 4

Prior for the slope of a straight line. (*a*) A prior uniform in slope a, and straight lines with constant interval Δa. (*b*) A prior uniform in angle $\theta = \tan^{-1}(a)$ that is symmetric with respect to rotation, and straight lines with constant interval $\delta\theta$. Adapted from Dose (2002) with permission.

We now study the case where $\mathbf{\Sigma_i}$ is unknown and $\sigma_p = 0$. For simplicity, we assume the uncertainty is the same for all data points, i.e., $\sigma_{\perp,i} = \sigma_\perp$,

$$p(a, b, \sigma_\perp | \{x_i\}, \{y_i\}) = A p(a, b, \sigma_\perp) p(\{x_i\}, \{y_i\} | \sigma_\perp, a, b) \tag{29}$$

$$= A p(a, b) p(\sigma_\perp | a, b) \frac{1}{(2\pi\sigma_\perp^2)^{N/2}} \exp\left(-\frac{\sum_i d_i^2}{2\sigma_\perp^2}\right), \tag{30}$$

and integrate over $\sigma_\perp$ using the Jeffreys prior $p(\sigma_\perp | a, b) = 1/\sigma_\perp$ to arrive at

$$p(a, b | \{x_i\}, \{y_i\}) = A p(a, b) \left(\sum_i d_i^2\right)^{-N/2}. \tag{31}$$

So, if we ignore the prior factor, the best-fit line is simply the line that minimizes the sum of the squared perpendicular distances of points from the line.

2.5. Model Comparison

When we have multiple models to explain data, we are faced with the question of which model is better. There is no unique definition of better, and depending upon what we mean by better we can come up with different criteria to compare models. We have two main schools of thought: (*a*) to compare the probability of the model given the data and (*b*) to compare the expected predictive accuracy of the model for the future data. The former is inherently a Bayesian approach and is known as Bayesian model comparison. The latter is inspired by frequentist ideas but can also be argued from a Bayesian perspective (Vehtari & Ojanen 2012, Gelman et al. 2014).

2.5.1. Bayesian model comparison. In the Bayesian formulation, the usefulness of a model is indicated by the probability of a model M given the data D:

$$p(M | D) = \frac{p(D|M) p(M)}{p(D)}. \tag{32}$$

The prior model probability $p(M)$ is generally assumed to be unity. Note in some cases it may not be so, and we might have more reason to believe one model over the other. The $p(D)$ is the same for all models, so it is irrelevant when comparing models. Thus the main thing we need to compute is the evidence $p(D|M)$ (also know as marginal likelihood). Hence, for two models M_1 and M_2, the odds ratio in favor of M_2 compared with M_1 is mainly determined by the ratio of their evidences, B_{21}, also known as the Bayes factor (for a review and a guide to interpreting the Bayes factor, see Kass & Raftery 1995),

$$\frac{p(M_2|D)}{p(M_1|D)} = \frac{p(D|M_2)p(M_2)}{p(D|M_1)p(M_1)} = \frac{p(D|M_2)}{p(D|M_1)} = B_{21}. \tag{33}$$

For some given data D and a model M parameterized by θ, we have

$$p(\theta|D, M) = \frac{p(D|\theta, M)p(\theta|M)}{p(D|M)}. \tag{34}$$

The evidence $p(D|M)$ appears as the denominator on the right-hand side and can be obtained by integrating both sides of Equation 34 over all θ. For properly normalized quantities, the left-hand side integrates to unity, leading to $p(D|M) = \int p(D|\theta, M)p(\theta|M)\mathrm{d}\theta$.

Note, the Bayes factor depends upon the adopted range of the prior, which leads to some conceptual difficulties (see the paradox in Lindley 1957). The range of priors is not an issue for parameter estimation, but it is for model selection; we cannot use improper priors. In most cases, we do have a reasonable sense of the range of priors and they are unlikely to extend to ∞. To better understand the role of priors, consider two models M_1 and M_2 in which M_2 has a free parameter θ, whereas M_1 has no free parameter (with θ being fixed to θ_0). Let $\Delta\theta_{\rm likelihood}$ be the characteristic width of the likelihood distribution and $\Delta\theta_{\rm prior}$ the range of a uniform prior that encloses the likelihood peak. The Bayes factor in favor of model M_2 as compared with M_1 is then

$$B_{21} = \frac{p(D|M_2)}{p(D|M_1)} = \frac{\int p(D|\theta)p(\theta)\mathrm{d}\theta}{L(\theta_0)} = \frac{\int p(D|\theta)\mathrm{d}\theta}{L(\theta_0)\Delta\theta_{\rm prior}} = \frac{L(\theta_{\rm max})}{L(\theta_0)}\frac{\Delta\theta_{\rm likelihood}}{\Delta\theta_{\rm prior}}. \tag{35}$$

The first term on the right-hand side will in general be greater than one and will favor M_2, as the simpler model M_1 is a special case of M_2. However, the second term penalizes M_2 if it has a large range in priors.

The conceptual difficulty associated with the dependence of the Bayes factor on the adopted prior range is alleviated if one thinks of hypothesis as a specification of a model as well as the prior on its parameters. A model M_2 with a larger range in priors allows for a larger number of possible data sets consistent with the hypothesis as compared with a simpler model M_1 with a narrow range of priors. Hence $p(D|M_2)$, being a normalized probability over possible data sets, will be lower as compared with $p(D|M_1)$ (MacKay 2003). Also, M_1 is more precise as a hypothesis as compared with M_2.

If we have more free parameters in a model, the penalty term in the Bayes factor will be higher, being of the form $\prod_{i=1}^{d} \Delta\theta_{\rm likelihood,i}/\Delta\theta_{\rm prior,i}$. In this sense, the Bayes factor has a built-in safeguard to prevent overfitting (a model with a large number of free parameters will fit a given set of data better but will perform poorly when presented with new data).

Computing the Bayes factor or the Bayesian evidence is computationally challenging. Generally, the likelihood is peaked and confined to a narrow region in the prior range but has long tails whose contributions cannot be neglected. Some commonly employed numerical techniques are (*a*) simulated annealing, (*b*) nested sampling, (*c*) Laplace's approximation, (*d*) Lebesgue integration theory (Weinberg 2012), (*e*) the Savage–Dickey density ratio (Verdinelli & Wasserman 1995),

and (f) importance sampling. For further details, see Kass & Raftery (1995) and for applications related to exoplanet detection, see Ford & Gregory (2007) and Nelson et al. (2016). Two useful approximations of the Bayes free energy $\mathcal{F} = -\ln p(D|M)$ are

$$\text{BIC}/2 = -\ln p(Y|\hat{\theta}) + (d \ln n)/2 \qquad \text{(Schwarz et al. 1978) and}$$

$$\text{WBIC}/2 = \mathbb{E}_\theta^\beta[-\ln p(Y|\theta)], \text{ where } \beta = \frac{1}{\ln n} \qquad \text{(Watanabe 2013).}$$

Here, $\mathbb{E}_\theta^\beta$ denotes expectation taken over the posterior distribution $p(\theta|Y) \propto p(Y|\theta)^\beta p(\theta)$ of θ. The case of $\beta = 1$ corresponds to the Bayesian estimation of the posterior. The posterior can be sampled using an MCMC algorithm. Assuming weak priors and that the posterior is asymptotically normal, we have $\mathcal{F} = \text{BIC} + O(1)$. Widely applicable Bayesian information criterion (WBIC) is an improved version of Bayesian information criterion (BIC), which is also applicable for singular statistical models where BIC fails. A model is singular if the Fisher information matrix is not positive definite, which typically occurs when the model contains hierarchical layers or has hidden variables.

Y: a set $\{y_1, y_2, \ldots, y_n\}$ of independently observed data points such that $p(Y|\theta) = \prod_{i=1}^n p(y_i|\theta)$

Var_θ^1: variance taken over the posterior distribution $p(Y|\theta)p(\theta)$ of θ

2.5.2. Predictive methods for model comparison. A statistical model $p(x|\theta)$ can be thought of as an approximation of the true distribution $q(x)$ from which the observed data $Y = \{y_1, y_2, \ldots, y_n\}$ were generated. The Bayesian predictive distribution can then be defined as $p(x|Y) = \text{E}_\theta[p(x|\theta)]$, whereas the maximum likelihood estimate is given by $p[x|\hat{\theta}(Y)]$. Predictive methods judge models by their ability to fit future data $X = \{x_1, x_2, \ldots, x_n\}$, e.g., via the log-likelihood function $-\ln p(X|Y)$. Given that we do not have future data, the idea is to measure out-of-sample-prediction error from the sample at hand. Cross validation is a natural way to do this, where we divide the current data set into training and testing samples. But this is computationally costly. Hence, alternate criteria have been developed. We start by computing the training error $T_e = -\frac{1}{n}\sum_{i=1}^n \ln p(y_i|Y)$. However, this is a biased estimator of $\mathbb{E}_x[-\ln p(x|Y)]$ as the data are used twice, once to estimate the model and then to compute the log likelihood of the data. If we have more parameters in the model, it will certainly fit the given data better but will also give rise to larger variance in the estimator, and we need to penalize the model for this. This variance, which represents the effective degrees of freedom in the model, can be calculated from the data and the model. A list of some useful information criteria based on the above idea are given below. They can be easily computed using samples of θ obtained by an MCMC simulation of the posterior $p(\theta|Y)$:

$$\text{AIC}/2 = -\ln p(Y|\hat{\theta}) + d, \qquad \text{(Akaike 1974)}$$

$$\text{DIC}_1/2 = -\ln p[Y|\text{E}_\theta^1(\theta)] + 2\left\{\ln p[Y|\text{E}_\theta(\theta)] - \text{E}_\theta^1[\ln p(Y|\theta)]\right\}, \qquad \text{(Spiegelhalter et al. 2002)}$$

$$\text{DIC}_2/2 = -\ln p[Y|\text{E}_\theta^1(\theta)] + 2\text{Var}_\theta^1[\ln p(Y|\theta)], \qquad \text{(Spiegelhalter et al. 2002)}$$

$$\text{WAIC}_1/2 = -\sum_i^n \ln \text{E}_\theta^1[p(y_i|\theta)] + 2\sum_i^n \ln \text{E}_\theta^1[p(y_i|\theta)] - \text{E}_\theta^1[\ln p(y_i|\theta)], \qquad \text{(Watanabe 2010)}$$

$$\text{WAIC}_2/2 = -\sum_i^n \ln \text{E}_\theta^1[p(y_i|\theta)] + \sum_i^n \text{Var}_\theta^1[\ln p(y_i|\theta)]. \qquad \text{(Watanabe 2010)}$$

Here, Var_θ^1 denotes variance taken over the posterior distribution $p(Y|\theta)p(\theta)$ of θ. The first term is a measure of how well the model fits the observed data, whereas the second term is a penalty for the degrees of freedom d in the model.

In general, the predictive criteria have a well-defined information-theoretic interpretation. Specifically, the expected value of Akaike information criterion (AIC) and widely applicable information criterion (WAIC) is equivalent to the expected Kullback–Leibler divergence

$\int q(x) \ln[q(x)/p(x|Y)]\mathrm{d}x$ of the predictive distribution from the true distribution, the expectation is taken over random realizations of the observed data set of Y, which samples the true disribution $q(x)$. In the asymptotic limit of large sample size, both AIC and WAIC are equivalent to LOOCV (leave-one-out cross validation).

An extra parameter in a model need not necessarily contribute to extra variance in the predictive density; e.g., if we have informative priors on the parameter, the likelihood has a very weak dependence on the parameter, or if the model is hierarchical then multiple parameters might be restricted. The use of AIC can be problematic in such cases. Deviance information criterion (DIC) and WAIC overcome this problem by estimating the effective degrees of freedom directly from the likelihood function of the data and samples of θ obtained from the posterior $p(\theta|Y)$.

WAIC offers some additional advantages as compared with AIC and DIC. AIC and DIC use a point estimate for θ when computing predictive density, whereas WAIC uses the Bayesian predictive density. If a model is singular, criteria such as AIC, DIC, and BIC do not work well. In contrast, WAIC works for such cases, and in the asymptotic limit of large sample size, WAIC is always equivalent to Bayesian LOOCV (Watanabe 2010).

It is instructive to study the differences between BIC and AIC, as they represent two very different approaches to the problem of model selection (for a detailed discussion, see Burnham & Anderson 2002). Due to the presence of the $\ln n$ term, for $n > 7$ the BIC penalizes free parameters more heavily as compared with AIC. So BIC is more parsimonious or cautious when it comes to admitting new parameters in a model. In situations where two models can give rise to the same predictive distribution, BIC will favor the model with fewer degrees of freedom, whereas AIC will treat them equally. An example is nested models, where a simpler model can be considered as a special case of a complex model but with few of its parameters being fixed. Interestingly, AIC can also be argued to be using the approach of Bayes factors, but with a prior whose variance decreases with sample size n, whereas BIC would correspond to the choice of a weak prior with fixed variance (Smith & Spiegelhalter 1980).

To conclude, the Bayesian and the predictive methods both have their strengths and weaknesses. If the choice of priors is well justified, then the methods based on the Bayes factor are best suited for model selection. However, if our aim is best predictive accuracy for future data, predictive methods like WAIC are a better choice.

3. MONTE CARLO METHODS FOR BAYESIAN COMPUTATIONS

Having discussed how to set up problems in the Bayesian framework, we now discuss methods to perform the inference; i.e., we discuss how to estimate the pdf of parameters given the data. Except for some simple cases, closed-form analytical solutions are in general not available. So one makes use of Monte Carlo–based methods to sample from the desired distribution. The most popular method to do this today is the MCMC method. MCMC is a class of methods for sampling a pdf using a Markov chain whose equilibrium distribution is the desired distribution. Once we have a sample distributed according to some desired distribution, we can compute expectation values and integrals of various quantities in a process analogous to Monte Carlo integration. The term Monte Carlo in MCMC comes from the use of random numbers to drive the Markov process and the close analogy to Monte Carlo integration schemes. Note that in conventional Monte Carlo integration, the random samples are statistically independent, whereas in MCMC they are correlated. We first broach the theory behind Markov chains and then discuss specific MCMC methods based on it. To expand upon the topics discussed here, a few good resources are Robert & Casella (2004), Brooks et al. (2011), and Liang et al. (2011).

3.1. Markov Chain

A Markov chain is a sequence of random variables $X_1, \ldots, X_n$ such that, given the present state, the future and past are independent. It is formally written as

$$\text{Prob}(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \text{Prob}(X_{n+1} = x | X_n = x_n). \tag{36}$$

In other words, the conditional distribution of X_{n+1} in the future depends only upon the present state X_n. If the probability of transition is independent of n, it is a time-homogeneous chain. Such a chain is defined by specifying the probabilities of transitioning from one state to another. To simplify mathematical notation, we sometimes consider the state space to be continuous and sometimes discrete. But the presented results are equally valid for either type of space. For a continuous state space where a probability density can be defined, we can write the transition probability as

$$K(x, y) = \text{Prob}(X_{n+1} = y | X_n = x). \tag{37}$$

For a discrete state space the transition probability is a matrix and is written as K_{xy}. On a given state space, a time-homogeneous Markov chain has a stationary distribution (invariant measure) π if

$$\pi(y) = \int \mathrm{d}x\, \pi(x) K(x, y). \tag{38}$$

A Markov chain is irreducible if it can go from any state x of a discrete state space to any other state y in a finite number of steps; i.e., there exists an integer n such that $K^n_{xy} > 0$. If a chain having a stationary distribution is irreducible, the stationary distribution is unique, and the chain is positive recurrent. For an aperiodic, positive recurrent chain with stationary distribution π, the distribution is limiting (equilibrium distribution). This means if we start with any initial distribution λ (a row vector specifying probability over states of a discrete state space) and apply the transition operator K (a matrix) many times, the final distribution will approach the stationary distribution π (a row vector),

$$\lim_{n \to \infty} \| \lambda K^n - \pi \| = 0. \tag{39}$$

For an irreducible Markov chain with a unique stationary distribution π, there is a law of large numbers that says that the expectation value of a function $g(x)$ over π approaches the average taken over the output of a Markov chain,

$$E_\pi[g(x)] = \int g(x) \pi(x) \mathrm{d}x = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} g(x_i). \tag{40}$$

This property allows one to compute Monte Carlo estimates of specific quantities of interest from a Markov chain. Techniques that do this are known as MCMC.

A chain having a stationary distribution is said to be reversible if the chain starting from a stationary distribution looks the same when run forward or backward in time. In other words, if X_n has distribution π, then the pair (X_n, X_{n+1}) has the same joint distribution as (X_{n+1}, X_n):

$$\text{Prob}(X_n, X_{n+1}) = \text{Prob}(X_{n+1}, X_n). \tag{41}$$

For the transition kernel K this means

$$\pi(X_n) K(X_n, X_{n+1}) = \pi(X_{n+1}) K(X_{n+1}, X_n) \tag{42}$$

and is known as the condition of detailed balance. For a Markov chain, it is not necessary to satisfy reversibility in order to have a stationary distribution. However, reversibility guarantees the existence of a stationary distribution and is thus a stronger condition. This is the reason that most MCMC algorithms are designed to satisfy detailed balance.

3.2. Metropolis–Hastings Algorithm

The most general MCMC algorithm is the MH algorithm (Metropolis et al. 1953, Hastings 1970). Suppose we are interested in sampling a distribution $f(x)$ on a state space E, with $x \in E$. To construct a transition kernel $K(x, y)$ to go from x to y, the MH algorithm uses a two step process:

- Specify a proposal distribution $q(y|x)$.
- Accept draws from $q(y|x)$ with acceptance ratio $\alpha(x, y) = \min[1, \frac{f(y)q(x|y)}{f(x)q(y|x)}]$.

So the transition kernel is given by $K(x, y) = q(y|x)\alpha(x, y)$. The full algorithm is as follows:

Algorithm 1: Metropolis–Hastings Algorithm

Input: Starting point x_1, function $f(x)$, transition kernel function $q(y|x)$
Output: An array of N points $x_1, x_2, \ldots, x_N$
for $t = 1$ **to** $N - 1$ **do**
  Obtain a new sample y from $q(y|x_t)$;
  Sample a uniform random variable U ;
  if $U < \frac{f(y)q(x_t|y)}{f(x_t)q(y|x_t)}$ **then** $x_{t+1} = y$ **else** $x_{t+1} = x_t$;
end

The transition kernel of the MH algorithm is reversible and satisfies detailed balance, $f(x)K(x, y) = f(y)K(y, x)$. Note the reversibility condition by itself does not lead to a unique form for the acceptance ratio $\alpha(x, y)$, and alternatives exist (Barker 1965). However, the acceptance ratio of the MH algorithm results in a chain with the fastest mixing rate (Peskun 1973).

There are multiple ways to construct the proposal distribution q, each leading to a new version of the MH algorithm.

- **Symmetric Metropolis:** $q(y|x) = q(x|y)$, which simplifies the acceptance probability to $\min[1, f(y)/f(x)]$; this is the version that was proposed by Metropolis and colleagues.
- **Random walk Metropolis–Hastings (RWMH):** $q(y|x) = q(y - x)$; the direction and distance of the new point from the current point are independent of the current point. Common choices are $N(x, \sigma^2)$ and Uniform$(x - \sigma, x + \sigma)$.
- **Independence sampler:** $q(y|x) = q(y)$; i.e., the new state is drawn independent of the current state. The acceptance probability is given by $\min[1, \frac{f(y)q(x)}{f(x)q(y)}]$, a generalization of the accept–reject algorithm. The quantity $q(x)$ should resemble $f(x)$ but with longer tails.
- **Langevin algorithm:** $q(y|x) \sim N[x + \frac{\sigma^2}{2}\nabla \log f(x), \sigma^2]$; this is useful when the gradient is available.

Except when $f(y) = f(x)$ (uniform target density), the mean of the acceptance ratio α is always less than unity. Decreasing σ in the RWMH algorithm increases α but lowers the independence of the sampler. Increasing σ improves the independence but lowers α. In the Langevin algorithm, one makes use of the information in the gradient to allow faster mixing of the chain.

3.3. Gibbs Sampling

The Gibbs sampler introduced by Geman & Geman (1984) is one of the most popular computational methods for doing Bayesian computations. Suppose we want to sample $f(x)$, where $x \in \chi \subseteq \mathcal{R}^d$. In Gibbs sampling, the transition kernel $K(x, y)$ is split into multiple steps. In each step, one coordinate is advanced based on its conditional density with respect to other coordinates. The algorithm is as follows:

Algorithm 2: Gibbs Sampling Algorithm

Input: Starting point x^1, function $f(x)$
Output: An array of N points $x_1, \ldots, x_N$
for $t = 1$ **to** $N - 1$ **do**
  Sample x_{t+1}^1 from $f(x^1|x_t^2, \ldots, x_t^d)$;
  Sample x_{t+1}^2 from $f(x^2|x_{t+1}^1, x_t^3, \ldots, x_t^d)$;
  Sample x_{t+1}^d from $f(x^d|x_{t+1}^1, \ldots, x_{t+1}^{d-1})$;
end

The full transition kernel is written as,

$$\kappa_{1\to d}(x_{t+1}|x_t) = \prod_{i=1}^{d} f(x_{t+1}^i|x_{t+1}^1, \ldots, x_{t+1}^{i-1}, x_t^{i+1}, \ldots, x_t^d). \tag{43}$$

Similarly one can define a reverse move,

$$\kappa_{d\to 1}(x_t|x_{t+1}) = \prod_{i=d}^{1} f(x_t^i|x_{t+1}^1, \ldots, x_{t+1}^{i-1}, x_t^{i+1}, \ldots, x_t^d). \tag{44}$$

It can be easily shown that

$$f(x_t)\kappa_{1\to d}(x_{t+1}|x_t) = f(x_{t+1})\kappa_{d\to 1}(x_t|x_{t+1}). \tag{45}$$

Integrating both sides leads to

$$\int f(x_t)\kappa_{1\to d}(x_{t+1}|x_t)\mathrm{d}x = f(y). \tag{46}$$

Thus, f is the stationary distribution of the Markov chain formed by the transition kernel $\kappa_{1\to d}(x_{t+1}|x_t)$. Note the Gibbs sampler as given above (systematic scan) is not reversible. However, the reversible ones can easily be produced, e.g., at each iteration picking a random component to update (random scan). The random scan Gibbs sampler can be viewed as a special case of the MH sampler with acceptance ratio $\min[1, \frac{f(y)q(x|y)}{f(x)q(y|x)}] = 1$. It follows that

$$\begin{aligned} f(y)q(x|y) &= f(y^i|y_{-i})f(y^{-i})f(x^i|y^{-i}) = f(y^i|x^{-i})f(x^{-i})f(x^i|x^{-i}) \\ &= f(x^i|x^{-i})f(x^{-i})f(y^i|x^{-i}) = f(x)q(y|x). \end{aligned} \tag{47}$$

Here, $x^{-i} = \{x^1, \ldots, x^{i-1}, x^{i+1}, \ldots, x^d\}$ and $y^{-i} = x^{-i}$, as only the ith component is changed in each step.

3.4. Metropolis Within Gibbs

One problem with the Gibbs sampler is that it requires one to sample from the conditional distributions, which can be difficult. In such cases, one can replace the sampling of conditional densities with the MH step. This then becomes the Metropolis-within-Gibbs (MWG) scheme (see Müller 1991), which is shown in Algorithm 3 (it is implemented in the software that we provide). Rather than updating all the variables step by step, one can also choose to update a subset of variables together, leading to block updates. The fact that the full sampling of a complicated distribution can be broken up into a sequence of smaller and easier samplings is the main strength of the Gibbs sampler and has resulted in its widespread use.

Algorithm 3: Metropolis-within-Gibbs Algorithm

Input: Starting point x^1, function $f(x)$
Output: An array of N points $x^1, x^2, \ldots, x^N$
for $t = 1$ **to** $N - 1$ **do**
  for $i = 1$ **to** d **do**
    Generate x_*^i from $q_i(x^i|x_{t+1}^1, \ldots, x_t^{i-1}, x_t^{i+1} \ldots, x_t^d)$;
    Calculate $r = \frac{f_i(x_*^i|x_{t+1}^1,\ldots,x_t^{i-1},x_t^{i+1},\ldots,x_t^d)}{f_i(x_t^i|x_{t+1}^1,\ldots,x_t^{i-1},x_t^{i+1},\ldots,x_t^d)} \frac{q_i(x_t^i|x_{t+1}^1,\ldots,x_t^{i-1},x_t^{i+1},\ldots,x_t^d)}{q_i(x_*^i|x_{t+1}^1,\ldots,x_t^{i-1},x_t^{i+1},\ldots,x_t^d)}$;
    if $U < \text{Min}(1, r)$ **then** $x_{t+1}^i = x_*^i$ **else** $x_{t+1}^i = x_t$;
  end
end

3.5. Adaptive Metropolis

The efficiency of the MH algorithm depends crucially upon the proposal distribution. By efficiency we typically mean how independent are the samples. If the samples are not independent, then they have high correlation. For Markov chains, the correlation falls off with distance between samples. If the correlation is large, this means the mixing in the chain is slow. If the width of the proposal distribution is too small, the acceptance ratio is high, but the chain mixes very slowly. If the width of the proposal distribution is too large, the acceptance ratio is too small, and the chain again mixes slowly (see **Figure 5*a–c*** for an illustration of this effect). Gelman et al. (1996) showed that the optimal covariance matrix Σ for the RWMH algorithm using the multivariate normal distribution is $(2.38^2/\mathcal{D})\Sigma_\pi$, where $\mathcal{D}$ is the dimensionality of the space and Σ_π is the covariance matrix of the target distribution π. The optimal acceptance ratio α_{opt} is 0.44 for dimension $\mathcal{D} = 1$ and then falls off with increasing number of dimensions, reaching an asymptotic value of 0.23 for $\mathcal{D} \to \infty$. The convergence is quite fast (α =[0.441, 0.352, 0.316, 0.279, 0.275, 0.266] for $\mathcal{D}$ =[1, 2, 3, 4, 5, 6]). The efficiency as compared with independent samples is $0.331/\mathcal{D}$.

These results suggest a possible way to choose the optimal proposal distribution. Estimate the covariance matrix Σ_π by a trial run and then use it for the actual run. Even doing this is cumbersome as it is unclear how long the trial run should be. To circumvent this, Haario et al. (2001) proposed an adaptive scheme in which Σ is updated on the fly using past values. Naïvely, any scheme that uses proposals that depend upon the full past history violates the Markovian property; i.e., the future should only depend on the present and should be independent of the past. The trick is to adapt the proposal distribution in such a way that it converges to the optimal one. The resulting chain then also converges to the target distribution. Andrieu & Robert (2001) showed that such a scheme can be described as part of a more general adaptive framework.

At the heart of most adaptive algorithms is the Robbins & Monro (1951) recursion. They proposed an iterative stochastic algorithm to find roots of functions that are stochastic; i.e., their algorithm solves $M(x) = \alpha$, where instead of $M(x)$ the function available is $N(x)$, which is stochastic and is such that $\langle N(x) \rangle = M(x)$. Starting with some initial value x_0, the algorithm to get the $n + 1$th iterate is

$$x_{n+1} = x_n + \gamma_n[\alpha - N(x_n)]. \tag{48}$$

Here $\gamma_1, \gamma_2, \ldots$ is a sequence of positive steps. The x_n then converge to the true solution provided the sequence γ_n satisfies

$$\sum_{n=0}^{\infty} \gamma_n = \infty \text{ and } \sum_{n=0}^{\infty} \gamma_n^2 < \infty. \tag{49}$$

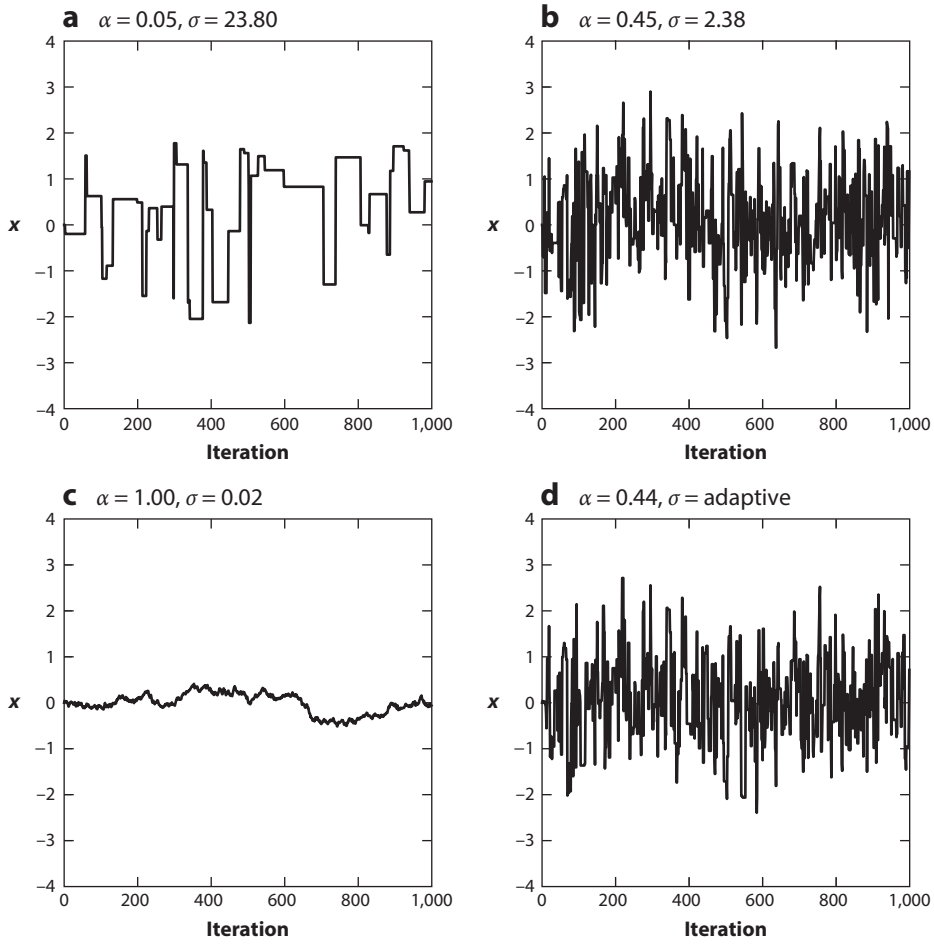

Figure 5

MCMC chains for different widths of the proposal distribution. The variable x is sampled from a Gaussian distribution $N(0.0, 1.0)$ using MCMC with different proposal distributions. The proposal distributions are also normal and are characterized by width σ. The ideal width is 2.38, and the chain for this is shown in panel *b*. In panel *a*, σ is ten times larger, whereas in panel *c* it is one-hundred times smaller. Panel *d* shows the chain when an adaptive scheme is used to adjust σ. The performance of this is the same as for the ideal case shown in panel *b*.

The first condition makes sure that, irrespective of where we start, the solution can be reached in a finite number of steps. The second condition makes sure that we do converge. A possible choice of γ_n is $\gamma_n = \gamma/n^\beta$, where $0 < \beta < 1$.

A nice description of various adaptive algorithms is given by Andrieu & Thoms (2008); below, we discuss algorithm 4 from their paper, which is quite general and is implemented in the software that we provide.

If β is too small, the convergence is too slow; if β is too large the convergence is too fast, and the simulation can quickly lean toward a wrong solution and will take a long time to get out of it. For adaptive MCMC, we find a choice of $\beta = 0.6$ to be satisfactory for most test cases.

Algorithm 4: Adaptive Symmetric Random Walk Metropolis–Hastings Algorithm

Input: Starting point x_0, μ_0, Σ_0, α^*, function $f(x)$
Output: An array of N points $x_0, x_2, \ldots, x_{N-1}$
for $i = 1$ **to** $N - 1$ **do**
  Obtain a new sample y from $N(x_i, \lambda_i \Sigma_i)$;
  Set $\alpha_i(x_i, y) = \frac{f(y)}{f(x_i)}$;
  Sample a uniform random variable U;
  if $U < \alpha_i(x_i, y)$ **then**
    $x_{i+1} = y$;
  else
    $x_{i+1} = x_i$;
  end
  $\log \lambda_{i+1} = \log \lambda_i + \gamma_{i+1}[\alpha_i(x, y) - \alpha^*]$;
  $\mu_{i+1} = \mu_i + \gamma_{i+1}(x_{i+1} - \mu_i)$;
  $\Sigma_{i+1} = \Sigma_i + \gamma_{i+1}[(x_{i+1} - \mu_i)(x_{i+1} - \mu_i)^{\mathrm{T}} - \Sigma_i]$;
end

Figure 5*d* shows an adaptive MCMC chain obtained using Algorithm 4. The adaptive chain looks very similar to the ideal case shown in **Figure 5*b***, and this demonstrates the usefulness of the adaptive MCMC scheme.

3.6. Ensemble and Affine Invariant Sampling

An elegant solution to the problem of tuning the proposal density is to use the idea of ensemble samplers (Gilks et al. 1994). Here multiple chains (walkers) are run in parallel but allowed to interact in such a way that they can adapt their proposal densities. One such general purpose algorithm is the differential evolution Markov chain, also known as DE-MC, by Ter Braak (2006). For its applications and implementations in astronomy, see, e.g., Nelson et al. (2014) and Cubillos et al. (2017) (they also provide an open source software). Another general purpose algorithm based on ensemble sampling is the affine invariant sampler by Goodman & Weare (2010) (see also Christen & Fox 2010). A `Python` implementation of this (`emcee`: the MCMC hammer, **http://dan.iel.fm/emcee/current/**) is provided by Foreman-Mackey et al. (2013) and is widely used in astronomy. We now describe this algorithm.

We saw in the previous section that adapting the proposal density can violate the Markovian property of a chain. The trick lies in using the information available in the ensemble in a way that does not violate the Markovian property. This is achieved by using the idea of partial resampling, which is a generalized version of the Gibbs sampling procedure. Let us consider an ensemble of walkers $X = (x_1, x_2, \ldots, x_L)$ and a Markov chain that walks on a product space with distribution $\Pi(X) = \pi(x_1)\pi(x_2)\cdots\pi(x_L)$. Then if x_i is updated conditional on other walkers $x_{[-i]} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_L\}$ (complementary set of walkers), but satisfying detailed balance $p(y_i|x_i, x_{[-i]}) = p(x_i|y_i, x_{[-i]})$, then each walker samples from $\pi(x)$.

One way to do this is to choose a point x_j from $x_{[-i]}$ and a scalar r with density $g(r)$ and then propose a new point y as

$$y = x_i + (r - 1)(x_i - x_j) = x_j + r(x_i - x_j). \tag{50}$$

The inverse transformation is given by $x_i = x_j + (y - x_j)/r$. Now if we want the proposal to be symmetric, then $q(y_i|x_i, x_{[-i]}) = q(x_i|y_i, x_{[-i]})$, and this implies $g(1/r) = rg(r)$. A good choice of such a function is

$$g(r) = \frac{1}{\sqrt{r}} \text{ for } r \in \left[\frac{1}{a}, a\right] \text{ and } a > 1. \tag{51}$$

To satisfy detailed balance, the acceptance probability is given by $\min[1, r^{n-1}\frac{\pi(Y)}{\pi(X_i)}]$. The factor r^{n-1} is because the proposal is restricted along a line and not the full hypersphere over the actual space. This means an appropriate Jacobian has to be calculated; for details see Gilks et al. (1994) and Roberts & Gilks (1994) [the proof is much easier when using the reversible jump MCMC formalism of Green (1995)].

Moves other than the stretch move can also be constructed, e.g., a proposal $y = x_i + W$, where W has a covariance computed from a subset of walkers in the complementary sample. It is also possible to construct algorithms that use a combination of both the stretch and the walk move. Although the Goodman & Weare (2010) affine invariant algorithm elegantly solves the problem of choosing a suitable proposal distribution, it has one drawback. The computational cost of warm-up scales linearly with the number of walkers. Note, like most other MCMC algorithms, multimodal distributions (distributions with many well-separated peaks) also pose a problem for this algorithm.

3.7. Convergence Diagnostics

Having studied MCMC methods in order to sample from distributions, we now discuss how to detect convergence; i.e., how long should we run an MCMC chain? Several convergence diagnostics have been proposed in the literature. Cowles & Carlin (1996) provide a good review of 13 convergence diagnostics. Other resources include Brooks & Gelman (1998) and Robert & Casella (2004). Unfortunately, because there is no method to detect convergence, we can only detect failure to converge. So convergence diagnostics are necessary conditions but not sufficient. Below, we present two schemes to monitor convergence. The first scheme makes use of the correlation length of the chain to compute the effective number of independent samples in a chain. The second scheme makes use of multiple chains to see if they are converging.

3.7.1. Effective sample size. Let us begin by estimating how many independent samples we need to get reliable estimates of mean and variance of a quantity. For a posterior of some variable x with standard deviation σ_x, the Monte Carlo standard error goes as $\sigma_x/\sqrt{N}$ for a sample of size N. So, to measure the mean of a quantity with about 3% error as compared with the overall uncertainty σ_x, we need $N = 1{,}000$. Raftery & Lewis (1992) showed that to measure 0.025 quantile to within ± 0.005 with probability 0.95 requires about 4,000 independent samples.

However, the MCMC is not an independent sampler. As we have seen, the points in an MCMC chain are correlated; autocorrelation provides a measure of this. Autocorrelation $\rho_{xx}(t)$ for a sequence is the correlation between two points separated by a fixed distance t; i.e.,

$$\rho_{xx}(t) = \frac{\mathbb{E}[(x_i - \bar{x})(x_{i+t} - \bar{x})]}{\mathbb{E}[(x_i - \bar{x})^2]}. \tag{52}$$

An automatic windowing procedure is discussed by Sokal (1997) for the computation of integrated autocorrelation (see also Goodman & Sokal 1989, Goodman & Weare 2010). Typically the autocorrelation falls off exponentially as $\sim \exp^{-t/\tau_x}$ and τ_x is known as the correlation time (or correlation length). The integrated autocorrelation is defined as $\tau_{\text{int},x} = (1/2)\sum_{t=-\infty}^{\infty} \rho_{xx}(t)$. The variance of the mean of x for a sample of size N can be shown to be

$$\text{Var}(\bar{x}) = (2\tau_{\text{int},x})\frac{\mathbb{E}[(x_i - \bar{x})^2]}{N}. \tag{53}$$

So, for correlated samples the variance is $2\tau_{\text{int},x}$ times larger than the variance of independent samples. Using $\tau_{\text{int},x}$, one can measure the number of effective independent samples in a correlated

chain—also known as the effective sample size (ESS)—as $N/(2\tau_{\text{int},x})$ and then use it to decide if we have enough samples (e.g., 1,000 < ESS < 4,000).

3.7.2. Variance between chains. The most widely used criterion for studying convergence was first presented by Gelman & Rubin (1992). Let us suppose we have M chains each consisting of $2N$ iterations out of which we use only the last N iterations. For any given scalar parameter of interest θ, let

$$\bar{\theta}_j = \frac{1}{n}\sum_{i=1}^{n}\theta_{i,j} \quad \text{and} \quad \bar{\theta} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_j. \tag{54}$$

The index i runs over points in a chain, and the index j runs over the chains. Then the between-chain variance and the mean within-chain variance can be written as

$$B = \frac{1}{m-1}\sum_{j=1}^{m}(\bar{\theta}_j - \bar{\theta})^2 \quad \text{and} \quad W = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n-1}\sum_{i=1}^{n}(\theta_{i,j} - \bar{\theta}_j)^2. \tag{55}$$

The total variance $\hat{\sigma}^2$ for the estimator $\bar{\theta}$ can be written as a weighted average of W and B, $\hat{\sigma}^2 = W(n-1)/n + B$. If we account for the sampling variability of the estimator $\bar{\theta}$, then this yields a pooled variance of

$$V = \hat{\sigma}^2 + \frac{B}{m} = \frac{n-1}{n}W + \frac{m+1}{m}B \tag{56}$$

for the mixture of chains. If the initial distribution is over-dispersed, then $B > \sigma^2$, and V always overestimates the true variance σ^2. For any finite n, W is expected to be less than σ^2, as individual sequences in a chain would not have had the time to explore the full target distribution. So, initially we expect $V/W > 1$. However, in the limit $n \to \infty$, the variance B between chains, which is expected to fall off as $1/n$, goes to 0 and W will approach the true variance σ^2, making V/W approach 1. Therefore, the ratio $\hat{R} = \sqrt{V/W}$, which is also known as the potential scale reduction factor, can be used to monitor the convergence.

3.7.3. Thinning. For making inferences from an MCMC chain, some algorithms use only the kth iteration of each sequence such that successive draws are approximately independent; this process is known as thinning. However, there is no additional advantage of thinning other than savings in storage. Because we are throwing away information, an estimate from a thinned chain can never be better than the original chain (Geyer 1992, MacEachern & Berliner 1994). Furthermore, it is difficult to choose an appropriate k without studying the autocorrelation of the full chain. So thinning is useful only in situations where the autocorrelation is known a priori and is known to be large. Here again k should be chosen such that it is smaller than the autocorrelation length to retain as much information as possible.

3.8. Parallel Tempering

Multimodal distributions in general pose problems for all MCMC algorithms. Parallel tempering is one way to address this problem. It is a type of ensemble sampler in which multiple chains are simulated in parallel but are allowed to exchange information. Each chain has a target distribution different from the other and is controlled by a parameter T known as the temperature. Let $\pi(x) = \exp[-H(x)]$ be the actual target distribution; then a ladder of distributions

$$\pi_i(x) = \exp[-H(x)/T_i], i = 1, \ldots, n \tag{57}$$

is created, controlled via the parameter T_i, such that $T_1 > T_2 > \cdots > T_n$. T_n is set to 1. Hence, π_n represents the target distribution. The temperature broadens the target distribution

and allows a wider exploration of the parameter space, which makes it useful to explore multimodal distributions. To exchange information between the chains, a state swapping procedure is used. A swap is proposed between a randomly chosen chain i and its neighbors $i-1$ and $i+1$ with probability $q_{i,i-1} = q_{i,i+1} = 0.5$ and $q_{1,2} = q(n, n-1) = 1$. Naïvely, accepting the swap violates the detailed balance condition. So the swap proposal is accepted with probability

$$p_{ij} = \min\left[1, \frac{\pi_i(x_j)\pi_j(x_i)}{\pi_i(x_i)\pi_j(x_j)}\right] = \min\left(1, \exp\left\{\left[H(x_i) - H(x_j)\right]\left(\frac{1}{T_i} - \frac{1}{T_j}\right)\right\}\right), \tag{58}$$

which satisfies detailed balance.

In parallel tempering the temperature ladder needs to be chosen carefully. If the neighboring temperatures are too far apart, the acceptance rate is diminished, leading to slow mixing. If the neighboring temperatures are too close, a large number of elements in the ladder are required to explore a wide range in parameter space, and this can increase the computational cost significantly. However, by exploiting the trial runs, a suitable ladder can be constructed (Liang et al. 2011).

The idea of parallel tempering can be generalized to construct evolutionary algorithms that incorporate features of genetic algorithms into the framework of MCMC. The basic idea is to have parallel chains as in parallel tempering and allow exchange of information while satisfying detailed balance on the product space defined by the chains. The exchange of information is based on ideas of mutation and crossover from genetic algorithms (Liang & Wong 2001a,b).

3.9. Monte Carlo Metropolis–Hastings

In MCMC-based Bayesian inference, we are concerned with simulating samples from some pdf $p(\theta|x) = p(x|\theta)p(\theta)$. However, there are situations when $p(x|\theta)$ cannot be easily evaluated or is not available in an analytically tractable form. In such situations one can make use of Monte Carlo–based techniques to approximately evaluate $p(\theta|x)$. More generally, the MH ratio $r = p(\theta'|x)/p(\theta|x)$ is used to update an MCMC chain. In such techniques, typically, one generates a set of auxiliary samples $Y = \{y_1, \ldots, y_m\}$ conditioned on θ and then uses them to compute $\tilde{p}(\theta|x, Y)$ [an approximation of $p(\theta|x)$] or $\tilde{r}$ (an approximation of the ratio r). However, Monte Carlo–based estimates are stochastic, and special care is needed when working with them in an MCMC scheme. An algorithm to make use of Monte Carlo–based estimates inside an MH algorithm is given in Algorithm 5 (it is implemented in the software that we provide).

Algorithm 5: Monte Carlo Metropolis–Hastings

Input: $\tilde{p}(\theta|x, Y)$, $q(\theta'|\theta)$, T and x
Output: A set of points $(\theta_1, \theta_2, \ldots, \theta_N)$ sampled approximately from $p(\theta|x)$
for $t = 1$ **to** $T - 1$ **do**
  Generate θ' from $q(.|\theta)$;
  Generate m auxiliary samples $Y = (y_1, \ldots, y_m)$ conditioned on θ;
  Obtain a Monte Carlo estimate $\tilde{r}(x, \theta, \theta', Y)$ of the MH ratio $\tilde{r} = \tilde{p}(\theta'|x, Y)/\tilde{p}(\theta|x, Y)$;
  if $U < \text{Min}(1, \tilde{r})$ **then** $\theta_{t+1} = \theta'$ **else** $\theta_{t+1} = \theta$;
end

There are many variants of Algorithm 5, depending upon how and at what stage the auxiliary sample is generated [see chapter 4 of Liang et al. (2011)]. The invariant stationary distribution of such Markov chains is not necessarily the target density $p(\theta|x)$. The characteristics of such chains and their convergence properties are discussed by Beaumont (2003) and Andrieu & Roberts (2009). In Algorithm 5, the auxiliary sample is refreshed in each iteration, and the same sample Y is used to estimate both $\tilde{p}(\theta'|x, Y)$ and $\tilde{p}(\theta|x, Y)$. This makes Algorithm 5 more robust compared to

other similar alternatives. In classical MCMC, one can reuse the previous estimate of $p(\theta|x)$ when computing r. However, when the MH ratio $\tilde{r}$ is stochastic, if $\tilde{p}(\theta|x, Y)$ is not evaluated in each iteration using a fresh sample of Y, then the MCMC chain tends to get stuck at a stochastic maximum of the estimated likelihood (Sharma et al. 2014). The smaller the size of the auxiliary sample, or the more inaccurate the Monte Carlo estimate of $\tilde{r}$, the worse is this problem. Using the same sample Y to estimate both $\tilde{p}(\theta'|x)$ and $\tilde{p}(\theta'|x)$ leads to lower noise in the estimated ratio of $\tilde{r}$. This property was also noticed by McMillan & Binney (2013) in the context of fitting models of the gravitational potential of the Milky Way to spatio-kinematic data of stars orbiting inside it. Two specific cases where the above algorithm can be used are given below.

3.9.1. Unknown normalization constant. In fitting a model to data, we are interested in sampling $p(\theta|x) = p(x|\theta)p(\theta)$. To do this, the function $p(x|\theta)$ should be properly normalized over the data space, in the sense that $\int p(x|\theta)\mathrm{d}x = 1$. However, on many occasions, we have

$$p(x|\theta) = \frac{1}{Z(\theta)} \exp[-U(x,\theta)] = \frac{1}{Z(\theta)} f(x|\theta), \tag{59}$$

where $f(x|\theta)$ is known, but the normalization constant $Z(\theta)$ is not known. An example is the problem of fitting a density profile $\rho(r|\theta)$ (r being the Galactocentric distance) to a sample of stars with Galactic latitude $b > 30°$, longitude $l > 30°$, and heliocentric distance $s < 50$ kpc. Here, we have $Z(\theta) = \int_{b=\pi/6}^{\pi/2} \mathrm{d}b \int_0^{50} \mathrm{d}s \int_{\pi/6}^{2\pi} \mathrm{d}l \rho(l, b, s|\theta) s^2 \cos(b)$.

Our aim is to compute the MH ratio $r = p(\theta'|x)/p(\theta|x) = [Z(\theta)/Z(\theta')][f(x|\theta')/f(x|\theta)]$ that is used to advance an MCMC chain, and it is the ratio $R = Z(\theta)/Z(\theta')$ that is unknown. If one can sample exactly from $f(x|\theta)$, then it is possible to cancel the normalization constant using ingenious algorithms by Møller et al. (2006) and Murray et al. (2012). However, exact sampling is not always feasible. In such cases a Monte Carlo estimate of the ratio of the unknown normalization constant $R = Z(\theta)/Z(\theta')$ can be done using samples $Y = (y_1, \ldots, y_m)$ generated from density $f(y|\theta)$, such that

$$\tilde{R}(\theta, \theta'; Y) = \frac{1}{m} \sum_{i=1}^{m} \frac{f(y_i|\theta')}{f(y_i|\theta)}. \tag{60}$$

This sampling can be done by various means, e.g., exact sampling, MCMC, and rejection sampling. If $f(y|\theta')$ is difficult to sample from, one can use so-called importance sampling by drawing samples from a distribution $g(y|\theta)$ that is easy to sample from. The required ratio of normalization constants is then given by

$$\tilde{R}(\theta, \theta'; Y) = \frac{\frac{1}{m}\sum_{i=1}^{m} f(y_i|\theta')/g(y|\theta)}{\frac{1}{m}\sum_{i=1}^{m} f(y_i|\theta)/g(y|\theta)}, \tag{61}$$

and the MH ratio is given by $\tilde{r} = \tilde{R}(\theta, \theta'; Y)[f(x|\theta')/f(x|\theta)]$.

3.9.2. Marginal inference. Here, we are interested in the marginal density $p(\theta|x) = \int p(\theta, y|x)\mathrm{d}y$, but the integral may not be analytically tractable and may also be difficult to do by deterministic schemes. In such situations, the integration can be done by Monte Carlo importance sampling using auxiliary samples Y generated from some density $g(y|\theta)$ that is easy to sample from. Thus we have

$$\tilde{p}(\theta|x, Y) = \frac{1}{m} \sum_{i=1}^{m} \frac{p(\theta, y_i|x)}{g(y_i|\theta)}. \tag{62}$$

3.10. Hamiltonian Monte Carlo

One of the attractive features of MCMC for sampling pdfs is its better performance for higher dimensions. However, for very large dimensions, traditional MCMC algorithms start running into problems. For lower dimensions, a typical set of the posterior (e.g., the region encompassing 99% of the total probability) lies close to the center, whereas for higher dimensions, a typical set lies in a shell that has a very large volume. Because a shell cannot be traversed with large step sizes, it takes a long time to explore the posterior.

Hamiltonian Monte Carlo (HMC) tries to address this problem by introducing an auxiliary variable called momentum u for each real variable x called position (Duane et al. 1987, Neal 1993). The log of posterior (target density) $\pi(x)$ is assumed to define the potential energy, $U(x) = -\ln \pi(x)$, and the momenta define the kinetic energy $K(u)$. Together they define the Hamiltonian $H(x,u) = U(x) + K(u)$, where $K(u) = u^2/2$. The distribution to be explored is

$$p(x,u) = \exp\left[-H(x,u)\right] = \exp\left[-\ln \pi(x) - \frac{1}{2}u^2\right]. \tag{63}$$

Next, principles of Hamiltonian dynamics are used to advance a given point to a new location. The point is then accepted or rejected based on the MH algorithm. The use of Hamiltonian dynamics to advance a given point allows the point to travel to locations that are far from its current location. This allows faster exploration of the parameter space.

There are two major obstacles involved with using HMC, and this has prevented its widespread use. First, it requires the gradient of the target density. Second, it requires two extra parameters to be tuned by the user: a step size ϵ to advance from the current state and the number of steps over which to evolve the Hamiltonian system. Considerable progress has been made to address both these issues.

The automatic/algorithmic differentiation can be used to accurately compute the derivatives of a given function without any user intervention (Griewank & Walther 2008). The idea is that any function written as a computer program can be described as a sequence of elementary arithmetic operations, and then by applying the chain rule of derivatives repeatedly on these operations, the derivatives can be computed. Alternatively, one can create analytical functions to approximate the target density and use these to compute the derivatives. This is because the exact Hamiltonian is only required when computing the acceptance probability, and this does not require derivatives. For simulating the trajectory, one needs derivatives, and here one can use an approximate Hamiltonian (Neal 2011). An application of HMC for fitting cosmological parameters is given by Hajian (2007) and Taylor et al. (2008). Homan & Gelman (2014) provide additional algorithms for automatic tuning of step size ϵ and the number of steps L, known as the No-U-Turn Sampler. This is used in the open-source Bayesian inference package `Stan` (available at **http://www.mc-stan.org**).

3.11. Population Monte Carlo

Population Monte Carlo is an iterative importance sampling technique that adapts itself at each iteration and produces a sample approximately simulated from the target distribution. The sample along with its importance weights can be used to construct unbiased estimates of quantities integrated over the target distribution. Suppose $h(x)$ is a quantity of interest. One of the major applications for MCMC is to compute integrals like $J = \int h(x)\pi(x)\mathrm{d}x$. In importance sampling, this is replaced by

$$J = \frac{1}{N}\sum_{i=1}^{N} h(x_i)\frac{\pi(x_i)}{q(x_i)}, \tag{64}$$

where $(x_1, \ldots, x_n)$ are sampled from a distribution $q(x)$, which is easier to sample than $\pi(x)$. The closer the importance function to the target distribution, the better the quality of the estimate (lower variance). In practice it is difficult to guess a good importance function.

The main idea in population Monte Carlo is to start with a reasonable guess of the importance function q_0 and then iteratively improve q_t by making use of the past set of samples $(x_1^{t-1}, \ldots, x_N^{t-1})$. The importance function can adapt not only in time (with each iteration), but also in space, and can be written in general as $q_t(.|x_i^{t-1})$. Suppose $X^t = \{x_1^t, \ldots, x_N^t\}$ are the set of points at iteration t. Let x_i^t be produced from importance distribution $q_t(x|x_i^{t-1})$. An estimate of J is then given by

$$J^t = \sum_{i=1}^{n} w_i^t h(x_i^t), \text{ where } \rho_i^t = \frac{\pi(x_i^t)}{q_{it}(x_i^t)} \text{ and } w_i^t = \frac{\rho_i^t}{\sum_{i=1}^{n}\rho_i^t}. \tag{65}$$

Thus the expectation value of any function $h(x)$ computed using importance sampling is unbiased; i.e.,

$$\mathbb{E}\left[h(X^t)\frac{\pi(X^t)}{q_t(X^t|X^{t-1})}\right] = \int h(x)\frac{\pi(x)}{q_t(x|y)}q_t(x|y)g(y)\mathrm{d}x\mathrm{d}y = \int h(x)\pi(x)\mathrm{d}x. \tag{66}$$

Here g is distribution of X^{t-1}, and the equality is valid for any g.

A simple choice for the importance function is to have set it as a mixture of normal or t-distributions, e.g., $q^t(x) = \sum_{d=1}^{D}\alpha_d^t\mathcal{N}(x|\mu_d^t, \Sigma_d^t)$ (Cappé et al. 2008). This has been used for cosmological parameter estimation (Wraith et al. 2009) and model comparison (Kilbinger et al. 2010).

3.12. Nested Sampling

In Section 2.5, we saw that computing the evidence is computationally challenging. Nested sampling (Skilling 2006) is designed to ease this computation. To compute the evidence, we are interested in computing quantities like

$$Z = \int L(\theta)\pi(\theta)\,\mathrm{d}\theta = \int L(\theta)\,\mathrm{d}\pi(\theta). \tag{67}$$

Integration is basically chopping up the full space into small volume elements and summing the contribution of the integrand over these cells. We are free to chop up the volume and order or label the cells as we wish. So, we divide the space by iso-likelihood contours and define a variable X to label them. A convenient choice is the prior probability mass enclosed by an iso-likelihood contour; i.e.,

$$X(L) = \int_{L(\theta)>L}\pi(\theta)\mathrm{d}\theta. \tag{68}$$

If the the prior probability is normalized, then it ranges from 0 for the highest likelihood to 1 for the lowest likelihood. Given the above definition, we can also define an inverse function $L(X)$, which is the likelihood that encloses a probability mass of X. So the integral for Z can now be written as $Z = \int L(X)\mathrm{d}X$.

Suppose we generate N samples uniformly from the prior distribution. Next, we sort them in decreasing sequence of L to give prior mass $X_i = i/N$. Then using the trapezoidal rule, one can easily perform the numerical integration. However, a significant contribution to the integral comes from a region with small prior mass X. So, the integral should be done in equal steps in

$\ln(X)$ rather than X. This can be done using an iterative procedure. We start with a set A of N points drawn from the prior. At each iteration, let L_i be the point with lowest L; we replace it in set A with a new point drawn uniformly from the prior but satisfying $L > L_i$. This generates a sequence of L_i for which the expected $X_i = \exp(-i/N)$.

Nested sampling is widely used for cosmological model selection and parameter estimation. Three publicly available packages based on nested sampling are `CosmoNest` (Mukherjee et al. 2006, Parkinson et al. 2006; see **http://www.mc-stan.org**), `MultiNest` (Feroz et al. 2009) and `DNEST` (Brewer et al. 2011; see **https://github.com/eggplantbren/DNest4**).

4. BAYESIAN HIERARCHICAL MODELING

In the simplest setting, we have some observed data Y generated by some model having parameters θ that can be inferred using Bayes's theorem as

$$p(\theta|Y) \propto p(Y|\theta)p(\theta), \tag{69}$$

where $p(\theta)$ denotes our prior knowledge or belief about θ. If the model parameters θ depend upon another set of parameters y through $p(\theta|\phi)p(\phi)$, then θ and ϕ can be inferred using

$$p(\theta, \phi|Y) \propto p(Y|\theta)p(\theta|\phi)p(\phi). \tag{70}$$

The variable ϕ is known as the hyperparameter and $p(\phi)$, the distribution of the hyperparameter, as a hyperprior. Alternatively, the observed data Y may depend upon another set of hidden variables X, which in turn depend on θ. The inference of θ and X can then be established using

$$p(\theta, X|Y) \propto p(Y|X)p(X|\theta)p(\theta). \tag{71}$$

Such situations lead to hierarchies, and Bayesian models of this type are known as hierarchical models. It turns out that hierarchies are quite common in real-world applications, often where more than two levels exist, and Bayesian hierarchical modeling (BHM) provides a framework for capturing this.

Let us conisder a simple example; for details see Gelman et al. (2013). Suppose we observe some data Y (a set of measurements of some variable y) with uncertainty σ, and we are interested in the mean $\alpha = \bar{y}$. Now suppose that the data $Y = \{y_{ij}|0 < j < J, 0 < i < n_j\}$ are grouped into J independent groups, and we have reason to believe that the group mean α_j varies from group to group. For observations within a group j, our model is

$$p(y_{.j}|\alpha_j, \sigma) = \mathcal{N}(y_{.j}|\alpha_j, \sigma^2), \tag{72}$$

where we denote by $y_{.j}$ an observation belonging to group j. We now compute the group mean $\bar{y}_{.j}$, instead of global mean $\bar{y}$, to capture the variation of mean across groups. A global mean is certainly an inaccurate description of data whenever the group mean is far away from the global mean. However, if the number of data points in a group is very small, e.g., $n_j = 2$, then the uncertainty in the group mean is large and it is much better to trust the global mean than the group mean.

BHM provides a natural way to handle the above problem of group means. It can act like a middle ground between the two extremes, global mean versus group mean. To demonstrate this, we set up the above problem using a BHM. Suppose the group means are distributed according to a normal distribution

$$p(\alpha_j|\mu, \omega) = \mathcal{N}(\alpha_j|\mu, \omega^2), \tag{73}$$

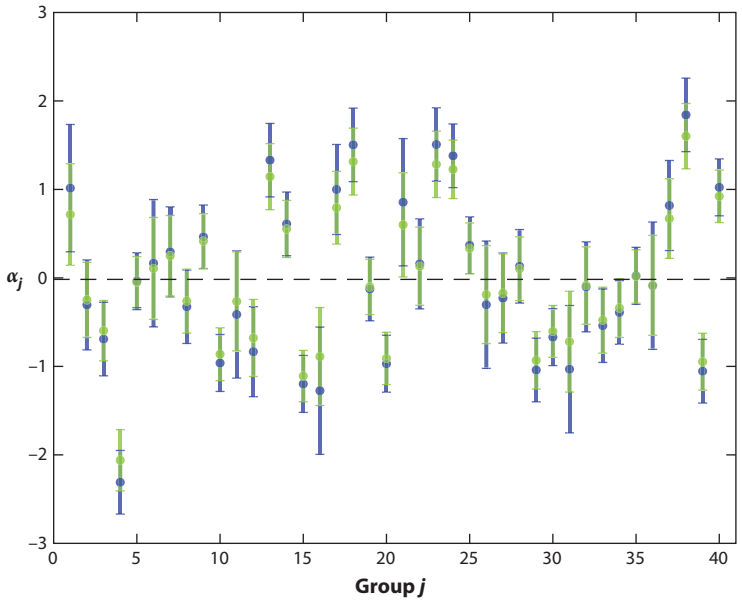

Figure 6

Analysis of group mean using a hierarchical Bayesian model. The dashed line is the global mean from all data points. The blue dots are group means computed from the data points in the group. The green points are estimates of group mean using a hierarchical Bayesian model, which makes use of the full information available. The green points have smaller error bars and are systematically closer to the global mean than the blue points.

where μ and ω are unknown parameters of the model. The μ, ω and the group means $\alpha = \{\alpha_1, \ldots, \alpha_J\}$ can then be inferred from data Y using

$$p(\alpha, \mu, \omega | Y) \propto p(Y|\alpha, \sigma) p(\alpha|\mu, \omega) p(\mu, \omega) \propto p(\mu, \omega) \prod_{j=1}^{J} p(\alpha_j|\mu, \omega) \prod_{i=1}^{n_j} p(y_{ij}|\alpha_j, \sigma). \quad (74)$$

We generated synthetic data with $\mu = 0$, $\omega = 1$, $J = 40$, $\sigma = 1$, and $2 < n_j < 10$; we then estimated α, μ and ω (assuming flat priors for μ and ω). The results are shown in **Figure 6**. The BHM based group mean estimates are systematically shifted with respect to standard group mean estimates (computed from the data points in a group). The BHM estimates are closer to the global mean than the standard estimates. The shift between the two estimates is more for cases where the error bars are large. The BHM estimates also have smaller error bars. This is because, when estimating the group mean, in addition to points within a group the BHM model also makes use of information available from other groups.

4.1. Expectation Maximization, Data Augmentation, and Gibbs Sampling

The easiest way to analyze a BHM is via Gibbs sampling, and the motivation for doing this was provided by the EM algorithm. In fact, the EM algorithm led to the development of the DA algorithm, which in turn provided the idea to use Gibbs sampling to solve BHMs.

Hence, we begin by exploring the EM algorithm (Dempster et al. 1977), which is one of the most influential algorithms in the field of statistics. Let us suppose that we have some observed

data $x = \{x_1, \ldots, x_N\}$ generated by some model $p(x|\theta)$ having parameters θ. We want to compute the most likely parameters of the model given the data; i.e., $\hat{\theta} = \text{argmax}_\theta[p(x|\theta)]$. The full model is specified by $p(x, z|\theta)$ with $p(x, z|\theta) = \prod_{i=1}^{N} p(x_i, z_i|\theta)$, where z are variables that are either missing or hidden or unobserved. The EM algorithm solves this problem as follows: The algorithm has two steps. It starts with a fiducial value of θ_0, then it does the following at every iteration t.

- E-step: Compute $Q(\theta|\theta_t, x) = \int \mathrm{d}z\ p(z|\theta_t, x) \log p(x, z|\theta)$. In other words, it computes the expectation of the log likelihood $\log p(x, z|\theta)$ with respect to $p(z|\theta_t, x)$.
- M-step: Find the value of θ that maximizes $Q(\theta|\theta_t, x)$ and set $\theta_{t+1} = \text{argmax}_\theta[Q(\theta|\theta_t, x)]$.

These steps are repeated iteratively until $\theta_{t+1} \sim \theta_t$. The proof that the EM algorithm increases the likelihood $p(x|\theta)$ at each stage is as follows: The conditional density of the missing data z given the observed data x and the model parameter θ is given by

$$p(z|\theta, x) = \frac{p(x, z|\theta)}{p(x|\theta)}. \tag{75}$$

Taking the log and then the expectation with respect to $p(z|\theta_t, x)$, we get

$$\log p(x|\theta) = \int \mathrm{d}z\ p(z|\theta_t, x) \log p(x, z|\theta) - \int \mathrm{d}z\ p(z|\theta_t, x) \log p(z|\theta, x) \tag{76}$$

$$= Q(\theta|\theta_t, x) + S(\theta|\theta_t, x), \tag{77}$$

which is valid for any θ. Using this result, we can compute the difference,

$$\log p(x|\theta_{t+1}) - \log p(x|\theta_t) = Q(\theta_{t+1}|\theta_t) - Q(\theta_t|\theta_t) + S(\theta_{t+1}|\theta_t) - S(\theta_t|\theta_t). \tag{78}$$

Due to the M-step, $Q(\theta_{t+1}|\theta_t, x) - Q(\theta_t|\theta_t, x) \geq 0$. Also, from Gibbs's inequality, $S(\theta_{t+1}|\theta_t) - S(\theta_t|\theta_t) \geq 0$. This means that each EM iteration is guaranteed to increase the marginal likelihood $p(x|\theta)$. This guarantees a convergence toward a maximum but not necessarily a global maximum. The algorithm can still get stuck at a saddle point or a local maximum.

The EM algorithm as presented above is deterministic. In general, it is not always easy to compute the expectation value, as it involves integrals over high dimensions. A general way to compute the $Q(\theta|\theta_t, x)$ would be to draw m random samples of z from the distribution $k(z|\theta_t, x)$ and take its mean. We label this stochastic estimate $Q_S(\theta|\theta_t, x)$, which in the limit $m \to \infty$ is the same as $Q(\theta|\theta_t, x)$. Having computed Q_S, the M-step can proceed as usual to maximize it and compute a new θ_{t+1}. In fact, m can be set to 1. This is the stochastic version of EM (SEM) as given by Celeux & Diebolt (1985). Because of stochasticity, one does not get a unique answer but instead a distribution. In fact, SEM generates a Markov chain, which under mild regularity conditions converges to a stationary distribution. The algorithm has an additional advantage in that it is less likely to get stuck at a local maximum.

If we now replace the M-step with a draw of θ from the $Q_S(\theta|\theta_t, x)$, this becomes a fully stochastic method; this is, as previously mentioned, the DA algorithm of Tanner & Wong (1987). This is equivalent to a two-step Gibbs sampler for sampling from

$$p(\theta, Z|X) \propto p(X, Z|\theta)p(\theta). \tag{79}$$

1. Sample Z_{t+1} from $p(Z|\theta_t, X)$.
2. Sample θ_{t+1} from $p(\theta|Z_{t+1}, X)$.

From the properties of the Gibbs sampler, we know that the sequence of (θ_t, Z_t) forms a Markov chain that samples $p(\theta, Z|X)$. Although Gibbs sampling requires sampling from the conditional distribution, the inner step can be replaced by MH sampling, leading to the MWG method as discussed in Section 3.3. This provides a completely general scheme for handling missing data.

For a few examples in astronomy in which a BHM is used and analyzed using the MWT method, see Sale (2012), Sharma et al. (2014), and Eadie et al. (2015).

Finally, the DA algorithm is not limited to just missing variables of the data, but it can also be applied to unknown parameters of the model, e.g., α in

$$p(\theta,\alpha|X) \propto p(X|\theta,\alpha)p(\theta)p(\alpha) \text{ or } p(\theta,\alpha|X) \propto p(X|\theta,\alpha)p(\theta|\alpha)p(\alpha). \quad (80)$$

Such dependencies are common in BHM. In general, the BHM provides a framework for handling marginalization in Bayesian data analysis; i.e., it provides a framework for handling parameters or variables that are either unknown or missing but are necessary to model the data.

4.2. Handling Uncertainties in Observed Data

Marginalization is not limited to handling missing data. It can also be used to handle data $X = \{x_i|i = 1,\ldots,N\}$ with uncertainty $\sigma_X = \{\sigma_{x,i}|i = 1,\ldots,N\}$. Consider

$$p(\theta|X,\sigma_X) \propto p(\theta)\prod_i \int p(x|\theta)p(x_i|x,\sigma_{x,i})\mathrm{d}x \text{ and} \quad (81)$$

$$p(\theta,X^t|X,\sigma_X) \propto p(\theta)\prod_i p(x_i^t|\theta)p(x_i|x_i^t,\sigma_{x,i}), \quad (82)$$

where $X_t = \{x_i^t|i = 1,\ldots,N\}$ are the true values of the observed data X. Here again, instead of doing an integration, one treats the true values as unknowns and samples them using the Gibbs scheme. We demonstrate this with a simple example in which $p(x_i^t|\theta) \sim \mathcal{N}(x_i^t|\mu,\sigma^2)$ is the model that generates the data, and $\theta = (\mu,\sigma)$ are the unknowns that we wish to evaluate. The data have uncertainty described by another Gaussian function $p(x_i|x_i^t,\sigma_{x,i}) \sim \mathcal{N}(x_i|x_i^t,\sigma_{x,i}^2)$. For this simple case, the integral in Equation 81 leads to an analytical expression

$$p(\theta|X,\sigma_X) \propto p(\theta)\prod_i \mathcal{N}(x_i|\mu,\sigma_{\mathrm{tot},i}^2), \text{ where } \sigma_{\mathrm{tot},i} = \sqrt{\sigma^2+\sigma_{x,i}^2}. \quad (83)$$

We used $(\mu,\sigma) = (0.0, 1.0)$ and $\sigma_x = 0.5$ to generate test data and then estimated μ and σ using two schemes: (*a*) the DA algorithm, which uses Equation 82 and treats X_t as unknown and samples from it, and (*b*) the explicit integration scheme, which uses Equation 83 in which the variable x_i^t has been integrated out of the equation. The Markov chain was run for 100,000 iterations. **Figure 7*a,b*** shows the pdf of the estimates of the two parameters. Both schemes give identical results. The autocorrelation functions for the two parameters are shown in **Figure 7*c,d***. The DA algorithm has a slightly higher autocorrelation time τ as it has to sample an extra parameter for each data point.

5. CASE STUDIES IN ASTRONOMY

In this section, we study a range of cases in astronomy where MCMC-based Bayesian analysis is making a significant impact. The emphasis is on showing how to set up a diverse range of problems within the Bayesian framework and how to solve them using MCMC techniques. The examples are intentionally chosen from different areas of astronomy to demonstrate the ubiquity of the techniques reviewed here. There is a long history of applying such techniques in the field of cosmology, and excellent reviews and books already exist here (Trotta 2008, Hobson 2010, Parkinson & Liddle 2013).

5.1. Exoplanets and Binary Systems Using Radial Velocity Measurements

The presence of a planet or a companion star results in temporal variations in the radial velocity of the host star. By analyzing the radial velocity data, one can draw inferences about the ratio of masses

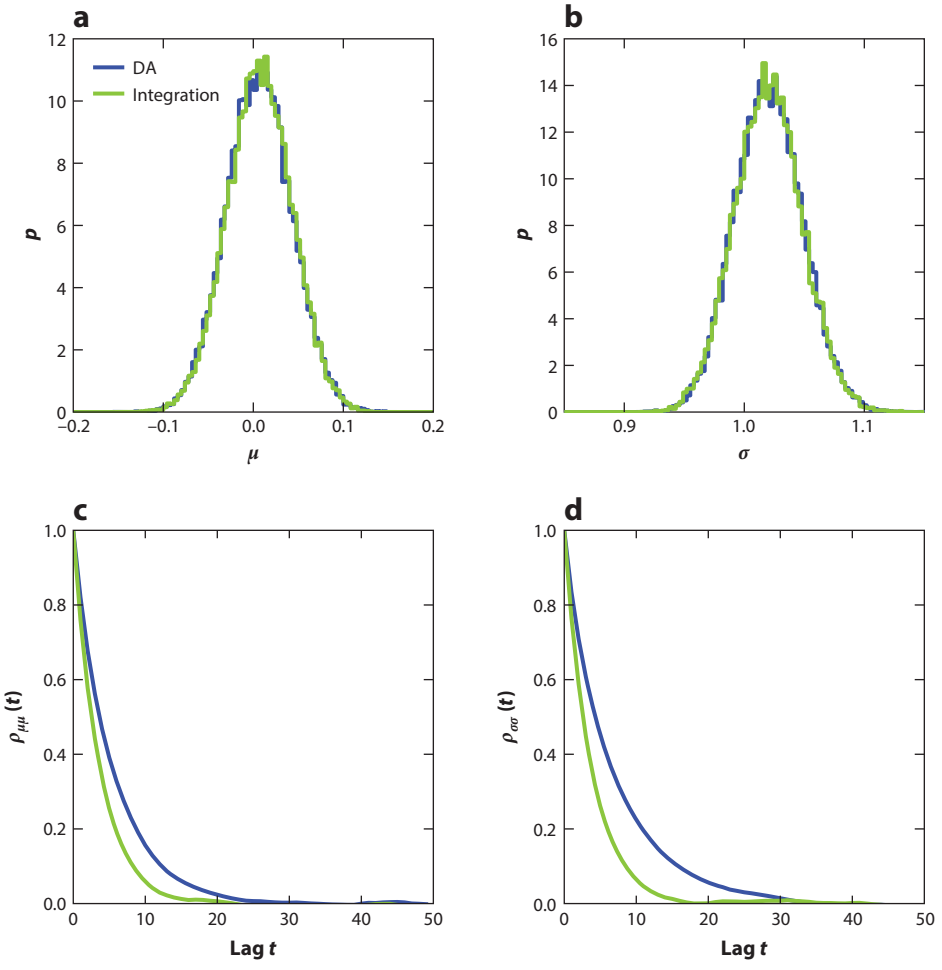

Figure 7

Comparison of two methods to handle nuisance parameters. Here the nuisance parameter is the true coordinate that is related to the observed coordinate via a given uncertainty. In the DA algorithm, the nuisance parameter is sampled alongside other parameters using Gibbs sampling on a Bayesian hierarchical model. In the other method, the nuisance parameter is marginalized via integration; an analytical form of the marginalized likelihood is used. The estimated parameters are μ and σ. Panels *a* and *b* show the probability distribution function of the parameters given the data. Panels *c* and *d* show the autocorrelation function of the two parameters in their respective Markov chains. Abbreviation: DA, data augmentation.

between the host and the companion, and orbital parameters like the period and eccentricity. We now describe how to set up the above inference problem in a Bayesian framework. We begin by describing the predictive model for the radial velocity of a star in a binary system.

The radial velocity of a star of mass M in a binary system with a companion of mass m in an orbit with time period T, inclination I, and eccentricity e is given by

$$v(t) = \kappa\,[\cos(f+\omega) + e\cos\omega] + v_0, \;\; \text{with}\; \kappa = \frac{(2\pi G)^{1/3} m \sin I}{T^{1/3}(M+m)^{2/3}\sqrt{1-e^2}}. \tag{84}$$

The true anomaly f is a function of time but depends upon e, T, and τ via

$$\tan(f/2) = \sqrt{\frac{1+e}{1-e}}\tan(u/2), \quad u - e\sin u = \frac{2\pi}{T}(t-\tau). \tag{85}$$

v_0: the mean velocity of the center of mass of the binary system

I: the inclination of the orbital plane with respect to the sky (angle between orbital angular momentum and line of sight)

ω: the angle of the pericenter measured from the ascending node (the point where the orbit intersects the plane of the sky)

τ: time of passage through the pericenter

An example of radial velocity data is shown in **Figure 8**, which shows the radial velocity for two binary systems that differ in e but have the same values for all other parameters κ, T, τ, ω, and v_0. The figure demonstrates that the radial velocity is sensitive to the eccentricity of the orbit.

The actual radial velocity data will differ from the perfect relationship given in Equation 84 due to observational uncertainty (variance σ_v^2) and the intrinsic variability of a star (variance S^2), and we can model this by a Gaussian function $\mathcal{N}(.|v, \sigma_v^2 + S^2)$. For radial velocity data D defined as a set of radial velocities $\{v_1, \ldots, v_M\}$ at various times $\{t_1, \ldots, t_M\}$, one can fit and constrain seven parameters, $\theta = (v_0, \kappa, T, e, \tau, \omega, S)$, using Bayes's theorem as shown below:

$$p(\theta|D) \propto p(D|\theta)p(\theta) \propto p(\theta)\prod_{i=1}^{M}\mathcal{N}(v_i|v(t_i;\theta), \sigma_v^2 + S^2). \tag{86}$$

We generated test data using Equation 84, and then, using the above equation, we tried to recover the parameters θ (available in the supplied software). The posterior distribution $p(\theta|D)$ was sampled using MCMC, and the results are shown in **Figure 8**. Panel *a* shows the test data along with the best fit curve. It also shows the radial velocity for the case with $e = 0$. Panel *b* shows the posterior distribution of the parameters κ, T, and e.

If we have data for a large number of binary systems, we can use them to explore the distribution of orbital parameters. A naïve way to do this would be to get a "maximum *a posteriori*" (MAP) estimate of the orbital parameters for each star and then study the population distribution by constructing histograms out of it. Such a scheme gives incorrect estimates of the population distribution as the uncertainty associated with the parameter estimates is ignored. In addition to this, as discussed by Hogg et al. (2010), the MAP estimates are in general biased. In the context of radial velocity data, the estimates of e are biased high. The problem is especially acute if the uncertainty associated with the parameters is large, which is often the case with radial velocity data from barycentric motions.

All of these problems can be avoided by setting up the problem of estimation of population distributions as a BHM. Let us suppose we have radial velocity data for N binary star systems, and we denote by y_i the radial velocity data set for the ith system. Let $x_i = (v_{0i}, \kappa_i, T_i, e_i, \tau_i, \omega_i, S_i)$ be the orbital parameters for the ith system. Finally, let α be the set of hyperparameters that govern the population distribution of the parameters x. The problem to determine α can be set up as

$$p(\alpha, \{x_i\}|\{y_i\}) \propto p(\alpha)\prod p(y_i|x_i)p(x_i|\alpha). \tag{87}$$

This is a BHM and can be sampled using the MWG scheme discussed in Section 4.1. The parameters x_i can be estimated alongside α, and to get the marginal distribution $p(\alpha|\{y_i\})$, one can simply ignore the computed x_i.

However, the above scheme is not well suited to explore a variety of population models, especially if sampling from $p(y_i|x_i)p(x_i|\alpha)$ is computationally demanding. We now show a computationally efficient scheme by Hogg et al. (2010) that can in general be applied to BHMs of two levels. The marginal distribution of hyperparameters that we are interested in is given by

$$p(\alpha|\{y_i\}) \propto p(\alpha)\prod\int \mathrm{d}x_i\, p(y_i|x_i)p(x_i|\theta). \tag{88}$$

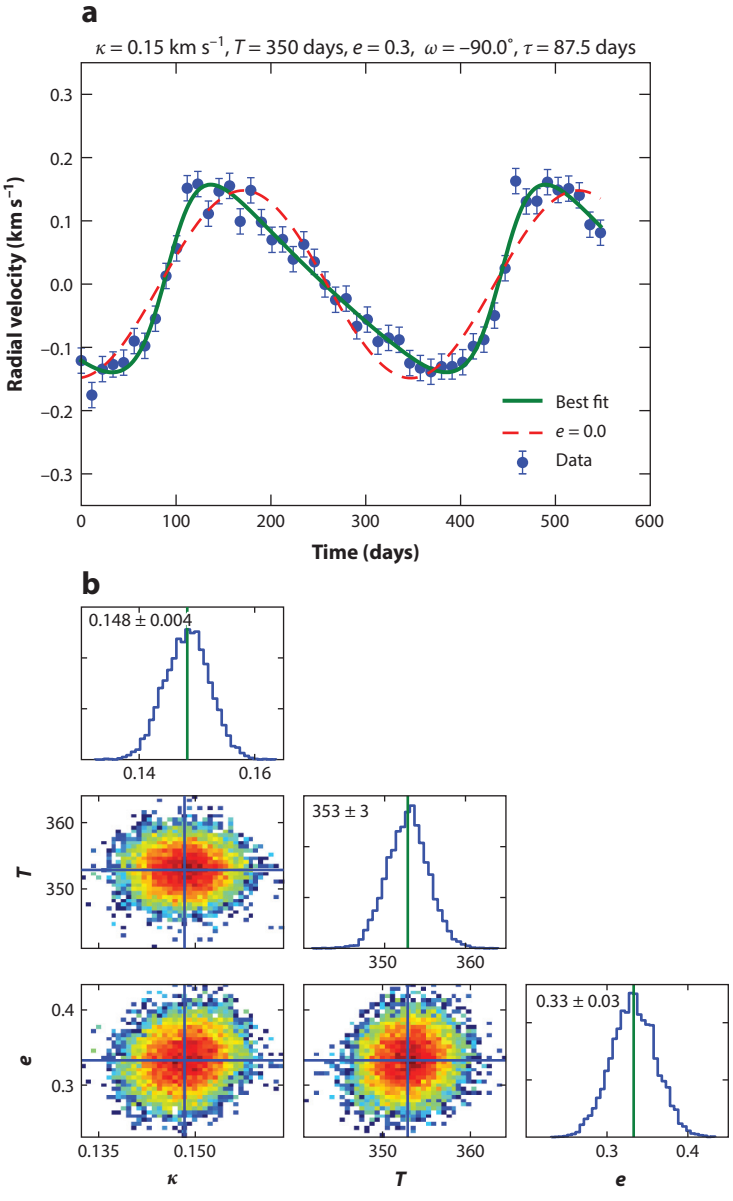

Figure 8

(*a*) Radial velocity as a function of time for a star in a binary system. The parameters of the binary system are listed on the top. The green line is the best-fit solution obtained using an MCMC simulation. The red and the green curves are generated from Equation 84 and differ only in the eccentricity e. The plot shows that the shape of the radial velocity curve depends sensitively upon the eccentricity e of the orbit. (*b*) The posterior probability distribution of parameters obtained using the MCMC simulation.

The integral on the right-hand side can be estimated using a Monte Carlo integration scheme as follows:

$$\int \mathrm{d}x_i\, p(y_i|x_i)p(x_i|\alpha) = \int \mathrm{d}x_i\, p(y_i|x_i)p(x_i)\frac{p(x_i|\alpha)}{p(x_i)} = \frac{1}{K}\sum_{k=1}^{K}\frac{p(x_{ik}|\alpha)}{p(x_{ik})}, \tag{89}$$

with x_{ik} sampled from $p(x_i|y_i) \propto p(y_i|x_i)p(x_i)$, which can be done by an MCMC scheme. For another application of BHM, see Wolfgang et al. (2016), where it is used to determine the mass–radius relationship of exoplanets.

5.2. The Data-Driven Approach to Estimation of Stellar Parameters from a Spectrum

The spectrum of a star contains information about its properties like temperature, gravity, and the abundance of different chemical elements that make up the star. Decoding information about stellar parameters from a stellar spectrum is a problem of great significance for astronomy. With the advent of large spectroscopic stellar surveys having several hundred thousand spectra, the need for fast and accurate methods to analyze the stellar spectra has gained prominence. Let us denote the stellar parameters (e.g., $T_{\rm eff}$, $\log g$, [Fe/H], and [X/Fe]) by label vector $\mathbf{x} = (x_1, \ldots, x_K)$ and the observed spectrum by vector $\mathbf{y} = \{y_1, \ldots, y_L\}$, denoting normalized flux at specific wavelengths indexed by $\lambda = (1, \ldots, L)$ (see **Figure 9**). The problem is to find $\mathbf{x}$ given $\mathbf{y}$, which using Bayes's theorem can be written down as

$$p(\mathbf{x}|\mathbf{y}, \theta) \propto p(\mathbf{y}|\mathbf{x}, \theta)p(\mathbf{x}). \tag{90}$$

Here $p(\mathbf{y}|\mathbf{x}, \theta)$ denotes a probabilistic generative model for the data, with θ being the parameters of the model. If we denote by $f_\lambda(\mathbf{x}, \theta_\lambda)$ the flux predicted by the model at wavelength λ and by s_λ^2 the variance or scatter about this relation (assuming Gaussian noise), then the probabilistic generative model for the full spectrum can be written as

$$p(\mathbf{y}|\mathbf{x}, \theta) = \prod_{\lambda=1}^{L} p(y_\lambda|\mathbf{x}, \theta_\lambda, s_\lambda) = \prod_{\lambda=1}^{L} \mathcal{N}(y_\lambda|f_\lambda(\mathbf{x}, \theta_\lambda), s_\lambda^2). \tag{91}$$

Traditionally, $f_\lambda(\mathbf{x}, \theta_\lambda)$ is calculated from first principles using a physical theory for the formation of spectral lines in a stellar atmosphere specified by stellar parameters $\mathbf{x}$. Frequently, $f_\lambda(\mathbf{x}, \theta_\lambda)$ is evaluated on a grid defined on x and then interpolation is used to get the spectrum for any arbitrary value of $\mathbf{x}$. The $f_\lambda(\mathbf{x}, \theta_\lambda)$ can also be computed by interpolating over a library of empirical spectra with predefined stellar parameters. A more refined data-driven approach to the problem using machine learning techniques was presented by Ness et al. (2015). In this approach, $f_\lambda(\mathbf{x}, \theta_\lambda)$ is approximated by a simple (linear or quadratic) function of label vector $\mathbf{x}$. Therefore,

$$f_\lambda(\mathbf{x}, \theta_\lambda) = \theta_{\lambda 0} + \sum_{i=1}^{K}\theta_{\lambda i}x_i + \sum_{i=1}^{K}\sum_{j=1}^{K}\theta_{\lambda ij}x_i x_j. \tag{92}$$

Let us consider a training set of N stars with label vectors $X = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$ and a corresponding set of fluxes at wavelength λ by $Y_\lambda = \{y_\lambda^1, \ldots, y_\lambda^N\}$. One can estimate θ_λ by sampling within MCMC such that

$$p(\theta_\lambda, s_\lambda|X, Y_\lambda) \propto p(Y_\lambda|X, \theta_\lambda, s_\lambda)p(\theta_\lambda)p(s_\lambda) \propto p(\theta_\lambda)p(s_\lambda)\prod_{i=1}^{N} p(y_\lambda^i|\mathbf{x}^i, \theta_\lambda s_\lambda). \tag{93}$$

Having obtained the model parameters $\theta = \{\theta_1, \ldots, \theta_L, s_1, \ldots, s_L\}$, one can now estimate stellar parameters $\mathbf{x}$ of a new star with given spectrum $\mathbf{y}$ using Equation 90. This is the basis of *The Cannon* algorithm (Ness et al. 2015), which is already widely used by the stellar community. The

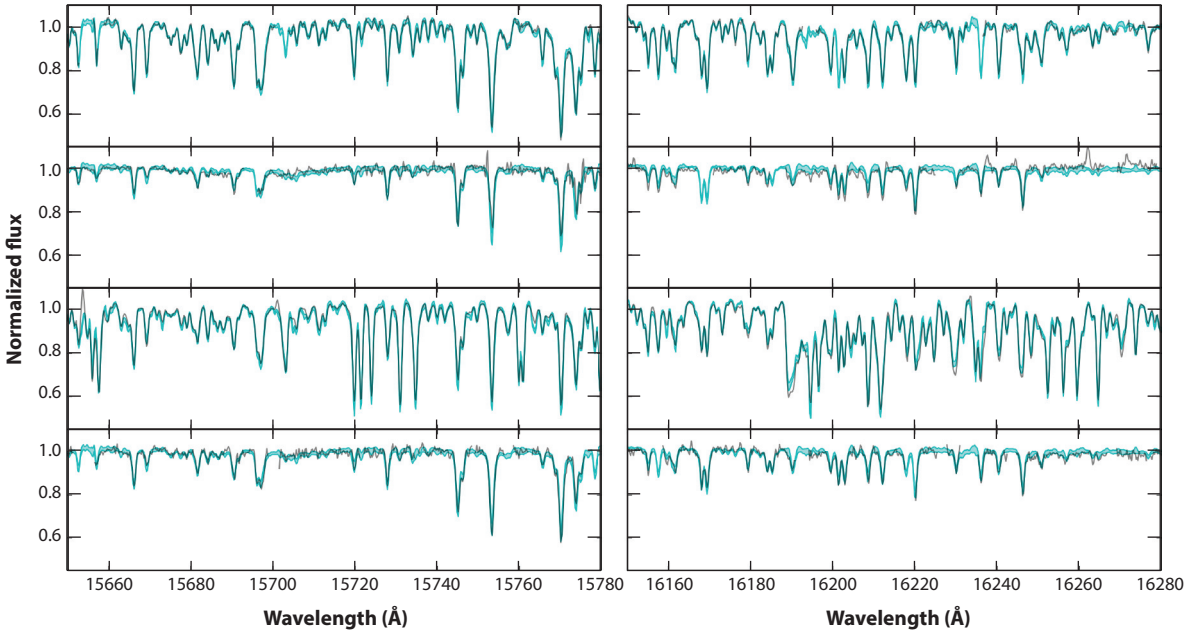

Figure 9

The APOGEE spectra of four stars (*gray line*) along with the best model spectra generated by *The Cannon* algorithm (*cyan line*) along with scatter around the fit. Each row shows the spectra of a single star in two wavelength intervals (*left* and *right*). Adapted from Ness et al. (2015) with permission.

ability of the algorithm to model the spectra is demonstrated in **Figure 9**, which shows the spectra of four stars along with the best-fit spectra for each of them.

5.3. Solar-Like Oscillations in Stars

Solar-like oscillations, which are excited and damped in the outer convective envelopes of a star, are seen in stars like the Sun and red giants. With the advent of space-based missions like *Kepler* and CoRoT that provide high-quality photometric data over a long time series, it has now become feasible to detect solar-like oscillations in tens of thousands of stars (Stello et al. 2013, 2015). Typically, the power spectrum of a star with solar-like oscillations (**Figure 10**) shows a regular pattern of modes, characterized by a large frequency separation $\Delta\nu$. The overall amplitude is modulated by a Gaussian envelope, and this is characterized by the frequency of maximum oscillation $\nu_{\rm max}$. Theory suggests that $\Delta\nu$ for a given star is related to its density (Ulrich 1986), whereas the $\nu_{\rm max}$ is related to its surface gravity and temperature (Brown et al. 1991, Kjeldsen & Bedding 1995). Using the above two relations, the mass and the radius of a star can be constrained. The mass of a red giant is sensitive to its age, and this makes asteroseismology very useful for understanding Galactic evolution (Chaplin et al. 2011, Sharma et al. 2016). For further details on solar-type oscillations see the review by Chaplin & Miglio (2013).

Bayesian-MCMC-based techniques are increasingly being adopted to extract seismic properties, e.g., $\Delta\nu$ and $\nu_{\rm max}$, by analyzing the power spectrum generated from the time series photometry of a star (Gruberbauer et al. 2009, Kallinger et al. 2010, Handberg & Campante 2011). The probability that an observed power spectrum $\mathbf{\Gamma} = \{\Gamma_1, \ldots, \Gamma_N\}$ at frequencies $\boldsymbol{\nu} = \{\nu_1, \ldots, \nu_N\}$ is

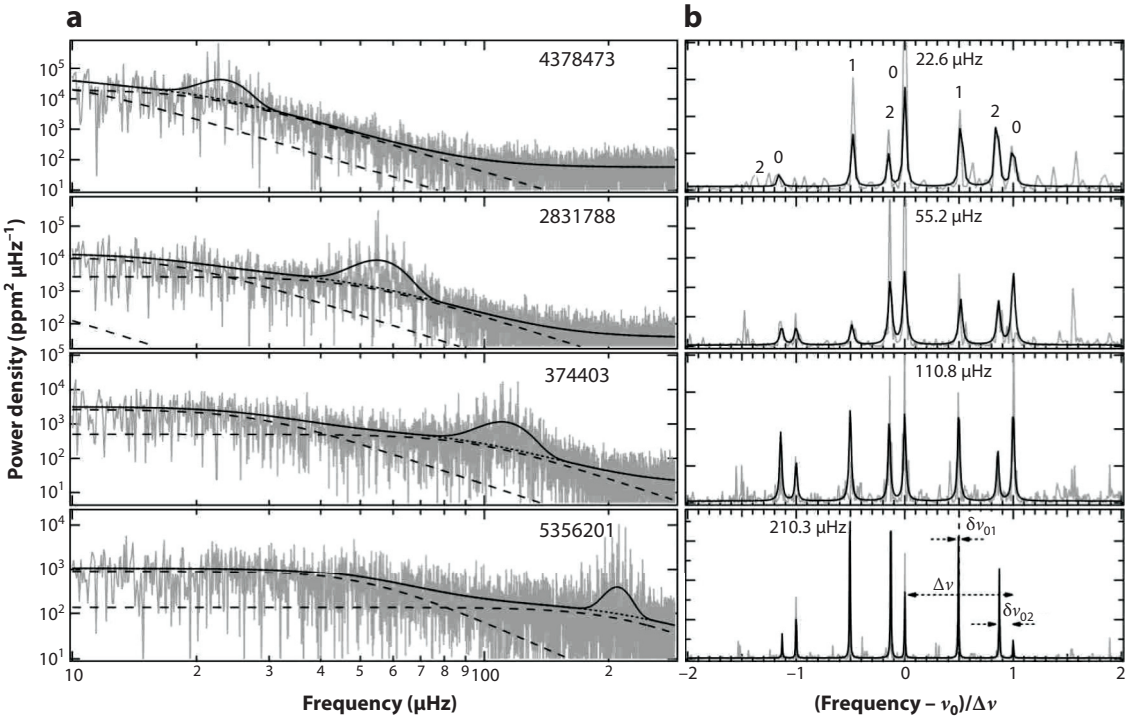

Figure 10

(*a*) The power density of four stars observed with *Kepler* showing solar-like oscillations along with a best-fit model (*solid black*). The hump is the approximate Gaussian-like envelope that modulates the power spectrum. The dashed lines are the individual super-Lorentzian profiles. The dotted line is the background model without the Gaussian component. (*b*) The spectrum region around central frequency ν_0 after subtracting the background model. Individual modes are clearly visible. The degree l of the modes is labeled in the top panel. The frequency separations $\Delta\nu$, $\delta\nu_{01}$, and $\delta\nu_{02}$ are shown using dotted lines in the bottom panel. Adapted from Kallinger et al. (2010) with permission.

produced by a model spectrum $\Gamma(\nu;\theta)$ (specified by a set of parameters θ) is given by

$$p(\mathbf{\Gamma}|\theta) = \prod_{i=1}^{N} \frac{1}{\Gamma(\nu_i;\theta)} \exp\left[-\frac{\Gamma_i}{\Gamma(\nu_i;\theta)}\right], \tag{94}$$

as shown by Duvall & Harvey (1986). This forms the basis for the Bayesian treatment of the problem of estimation of parameters θ by $p(\theta|\{\Gamma_i\}) = p(\{\Gamma_i\}|\theta)p(\theta)$. The power density is modeled as a sum of super-Lorentzian functions,

$$\Gamma(\nu;\theta) = \Gamma_{\rm wn} + \sum_k \frac{A_k}{1+(2\pi\nu\tau_k)^{\epsilon_k}} + P_{\rm g}\exp\left[\frac{-(\nu_{\rm max}-\nu)^2}{2\sigma_{\rm g}^2}\right]. \tag{95}$$

To fit the individual modes, one assumes Lorentzian profiles. Spherical harmonics are used to describe the oscillations; the modes are characterized by three wave numbers, n,l, and m. In Kallinger et al. (2010), eight main modes are fitted (three $l = 0$ and $l = 2$ and two $l = 1$), parameterized by the mode lifetime τ; the central frequency ν_0; three spacings $\Delta\nu$, $\delta\nu_{01}$, and $\delta\nu_{02}$;

and the amplitudes A_i, A_j, and A_k:

$$\Gamma(\nu) = P_{\rm wn} + \sum_{i=-1}^{1} \frac{A_i^2 \tau}{1 + 4[\nu - (\nu_0 + i\Delta\nu)]^2(\pi\tau)^2} + \sum_{j=-1}^{1} \frac{A_j^2 \tau}{1 + 4[\nu - (\nu_0 + j\Delta\nu - \delta\nu_{02})]^2(\pi\tau)^2} + \sum_{k=-1,1} \frac{A_j^2 \tau}{1 + 4[\nu - (\nu_0 + k\Delta\nu/2 - \delta\nu_{01})]^2(\pi\tau)^2}. \quad (96)$$

Figure 10 shows the result of fitting the above model to power spectra of four stars observed by the *Kepler* mission.

5.4. Extinction Mapping and Estimation of Intrinsic Stellar Properties

Given the mass m and initial composition (e.g., metallicity [M/H]) of a star, we can use the theory of stellar evolution to predict its state and composition at a later time (age τ). However, the intrinsic parameters like mass m, [M/H], and τ are not directly observable. For most stars, we only have photometric information, apparent magnitudes in different photometric bands (for example, J, Ks, u, g, r, and i). The photometry of a star depends upon temperature $T_{\rm eff}$, gravity g, [M/H], distance s, and extinction E (proportional to the dust density integrated along the line of sight to the location of the star). If we have spectroscopy, then we can get temperature $T_{\rm eff}$, g, and even composition, but with uncertainties. From asteroseismology, we can get average seismic parameters like $\Delta\nu$ and $\nu_{\rm max}$, which are sensitive to the mass, radius, and temperature of a star. Given this state of affairs, it is quite common to ask the question that, given a certain set of observables of a star, what are the intrinsic parameters of a star or even some other set of observables? For example, given the photometry of a star, what are the distance, temperature, and gravity of a star? Or, given photometry and distance, what are the temperature and gravity of a star? Or, given photometry and spectroscopy, what is the distance? And so on. Knowing the intrinsic parameters of a star is also important for understanding the formation and evolution of the Galaxy, for example, the star-formation rate, the age-metallicity relation, and the distribution of dust in the Galaxy.

The problem of estimating intrinsic stellar parameters of a star given some observables can be formulated as follows: Let $\mathbf{y} = (J, J - Ks, J - H, T_{\rm eff}, \log g, [{\rm M/H}]_{\rm obs}, l, b)$ be a set of observables associated with a star, and $\sigma_{\mathbf{y}}$ their uncertainties. Let us denote the intrinsic variables of a star that we are interested in by $\mathbf{x} = ([{\rm M/H}], \tau, m, s, l, b, E)$. To specify prior probabilities on $\mathbf{x}$, we need a Galactic model, and we denote by θ the parameters of such a model. Typically, real catalogs have selection effects, e.g., stars selected to lie in some apparent magnitude and color range, or a set of stars with parallax error less than 10%, or stars with missing information in certain bands. To specify selection effects, we denote the event that a star exists in a catalog by S. From theoretical isochrones, we can predict $\mathbf{y}$ given $\mathbf{x}$; in other words, a function $\mathbf{y}(\mathbf{x})$ exists. However, we are interested in the inverse problem of estimating $\mathbf{x}$ given $\mathbf{y}$. A Bayesian introduction to solving such a problem was given by Pont & Eyer (2004) and Jørgensen & Lindegren (2005) in the context of estimating ages. The method was further improved and refined by Burnett & Binney (2010, 2011) and Binney et al. (2014) in the context of the estimation of distances, with a better treatment of priors and selection effects (see also Sale 2012, 2015). From Bayes's theorem, we have

$$p(\mathbf{x}|\mathbf{y}, \sigma_{\mathbf{y}}, S, \theta) \propto p(S, \mathbf{y}|\mathbf{x}, \sigma_{\mathbf{y}}) p(\mathbf{x}|\theta) \propto p(S|\mathbf{y}, \mathbf{x}, \sigma_{\mathbf{y}}) p(\mathbf{y}|\mathbf{x}, \sigma_{\mathbf{y}}) p(\mathbf{x}|\theta). \quad (97)$$

We now explain each of the terms in detail.

1. $p(\mathbf{x}|\mathbf{y}, \sigma_{\mathbf{y}}, S, \theta)$ is the posterior distribution of intrinsic parameters given the observables, the selection function, and a Galactic model.

2. $p(S|\mathbf{y}, \mathbf{x}, \sigma_\mathbf{y})$ is the selection function. This says given the observables, what is the probability that a star was observed. Typically this can be expressed as $p(S|\mathbf{y})p(S|\mathbf{x})$. The term $p(S|\mathbf{x})$ enters in situations in which the value of an observable y' is not known, but constraints on it are. Then $p(S|\mathbf{x}) = \int p(S|y')p(y'|x)\mathrm{d}y'$. For example, a parallax of a star is known to be greater than a certain limit, or the apparent magnitude of a star may be missing in a band because the star is too bright or faint (Burnett & Binney 2010, Sale 2012).
3. $p(\mathbf{y}|\mathbf{x}, \sigma_\mathbf{y})$ is the likelihood of the data given the uncertainty and the intrinsic parameters. This can be described by a Gaussian function $\mathcal{N}(y|y(\mathbf{x}), \sigma_y^2)$ for each $y \in \mathbf{y}$.
4. $p(\mathbf{x}|\theta)$ is the prior. This describes the distribution of mass, metallicity, and age, and the spatial distribution of stars in the Galaxy. More specifically it can be written as $p(x|\theta) = \sum_k p_k(m)p_k([\mathrm{M/H}])p_k(\tau)p_k(r)$, where the sum is over different Galactic components, e.g., thin disc, thick disc, bulge, and stellar halo.

We now focus on the problem of estimating distance and extinction. For simplicity, we ignore the selection effects; for an in depth discussion, see Sale (2015). By marginalizing over stellar parameters τ, m, and [M/H], one obtains $p(s, E|\mathbf{y}, \sigma_\mathbf{y}, \theta)$. If we have N stars along a line of sight, we can estimate the distance–extinction relationship $E(s_i; \alpha)$ parameterized by α as

$$p(\alpha|\{\mathbf{y}\}, \theta) = p(\alpha) \prod_{i=1}^{N} \int \mathrm{d}E_i \mathrm{d}s_i \; p(s_i, E_i|\mathbf{y}_i, \sigma_\mathbf{y}^i, \theta) \; p(E_i|s_i, \alpha). \tag{98}$$

The above method is used by Green et al. (2014, 2015) to construct 3D maps of interstellar dust reddening using Pan-STARRS (Panoramic Survey Telescope and Rapid Response System) 1 and 2MASS (Two-Micron All Sky Survey) photometry (**Figure 11**). To estimate $p(s, E|\mathbf{y}, \sigma_\mathbf{y}, \theta)$, Green et al. (2015) do a kernel density estimate over samples generated by MCMC, whereas Sale & Magorrian (2015) present a method based on the Gaussian mixture model. As described by Sale

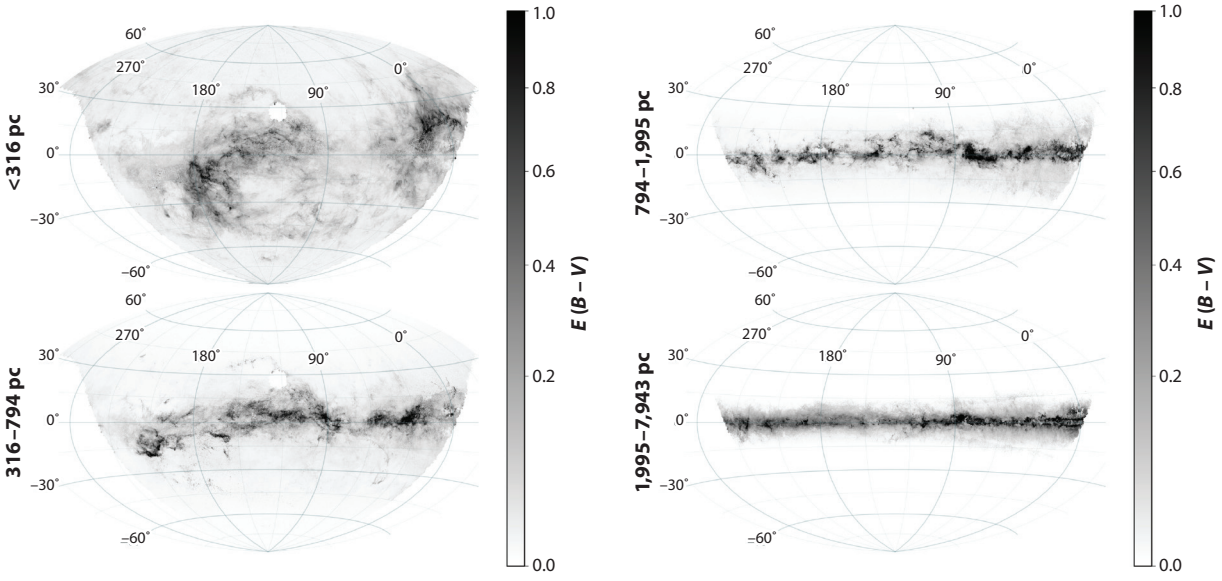

Figure 11

A 3D map of interstellar dust reddening in the Milky Way based on Pan-STARRS 1 and 2MASS photometry. Shown are the mean differential reddening in different heliocentric distance ranges. The map is adapted from Green et al. (2015) with permission and is available at **http://argonaut.skymaps.info**.

(2012), we can also directly estimate α and intrinsic parameters $\mathbf{x}$ of each star along a line of sight by setting up the problem as a BHM and sampling from the following posterior:

$$p(\alpha, \{\mathbf{x}\}|\{\mathbf{y}\}, \sigma_\mathbf{y}, \theta) \propto \mathbf{p}(\alpha) \prod_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{p}(\mathbf{y_i}|\mathbf{x_i}, \theta, \sigma_\mathbf{y}^\mathbf{i}, \alpha)\mathbf{p}(\mathbf{x_i}|\theta)\mathbf{p}(\mathbf{S}|\mathbf{y_i}). \tag{99}$$

The MWG scheme is used to accomplish this sampling.

5.5. Kinematic and Dynamical Modeling of the Milky Way

Understanding the origin and evolution of the Milky Way has received a significant boost due to the emergence of large data sets that catalog the properties of stars in the Milky Way (Binney 2011, 2013; McMillan & Binney 2012, 2013; Rix & Bovy 2013; Bland-Hawthorn & Gerhard 2016). Bayesian methods and MCMC-based schemes are now playing a prominent role in the analysis and interpretation of such large and complex data sets from, e.g., the GCS survey (Schönrich et al. 2010), the SEGUE survey (Bovy et al. 2012b), the APOGEE survey (Bovy et al. 2012a, Bovy & Rix 2013), and the RAVE survey (Piffl et al. 2014, Sharma et al. 2014, Sanders & Binney 2015). We focus on the problem of determining the mass distribution, or equivalently the gravitational potential of the Milky Way, using halo stars (Kafle et al. 2014) and disc masers (McMillan 2017).

The observational data of stars in the Milky Way is in heliocentric coordinates and is in the form of angular positions on sky (Galactic longitude ℓ and latitude b), heliocentric distance (s), heliocentric line-of-sight velocity ($v_{\rm los}$), and proper motion (tangential motion on the sky, μ_ℓ and μ_b). The velocity of halo stars can be described by a simple Gaussian model of the following form:

$$p(\mathbf{v}|\theta_v, \ell, b, s) = \mathcal{N}(v_r|0, \sigma_{vr})\mathcal{N}(v_r|0, \sigma_{v\theta})\mathcal{N}(v_\phi|v_{\rm rot}, \sigma_{v\phi}), \tag{100}$$

for which θ_v is the set of parameters that govern the velocity dispersion profiles $\sigma_{vr}, \sigma_{v\theta}$, and $\sigma_{v\phi}$. The coordinates (r, θ, ϕ) are in the Galactocentric reference frame. The observed heliocentric coordinates can be converted to Galactocentric coordinates using prior estimates of the location and the motion of the Sun. For the stellar halo stars, tangential velocities cannot be accurately determined. The distance also has some uncertainty, σ_s. Hence, we marginalize over unknown tangential velocities and true distance, s', to obtain

$$p(v_{\rm los}|\theta_v, \ell, b, s, \sigma_s) = \int\int\int p(v_\ell, v_b, v_{\rm los}|\theta_v, \ell, b, s')p(s'|s, \sigma_s)\, {\rm d}v_\ell {\rm d}v_b {\rm d}s'. \tag{101}$$

The parameters θ_v can now be estimated using the data D of multiple stars by

$$p(\theta_v|D) \propto \left(\prod_i p(v_{\rm los}|\theta_v, \ell_i, b_i, s_i, \sigma_{s,i})\right) p(\theta_v). \tag{102}$$

The marginalization in Equation 101 can be handled in various ways. One can make use of deterministic numerical integration techniques (Gaussian quadrature) or one can achieve marginalization via Monte Carlo schemes making use of importance sampling. For Monte Carlo–based integration one can make use of the Monte Carlo MH algorithm discussed in Section 3.9. Alternatively, one can treat v_ℓ, v_b, and s as unknowns by setting them up as a BHM and estimate them alongside θ by making use of the MWG scheme discussed in Section 4.1 (e.g., Sharma et al. 2014). The radial velocity dispersion profile of halo stars computed using blue horizontal branch and red giant stars in the SEGUE survey is shown in **Figure 12** (Kafle et al. 2014).

We now proceed to estimating the potential Φ. Given Φ, the density of halo stars ρ, and anisotropy $\beta = 1 - (\sigma_{v\theta}^2 + \sigma_{v\phi}^2)/(2\sigma_{vr}^2)$ as a function of distance r from the Galactic Center, one can solve for $\sigma_{vr}(r)$. Let θ be the set of parameters used to define the above profiles. So for given

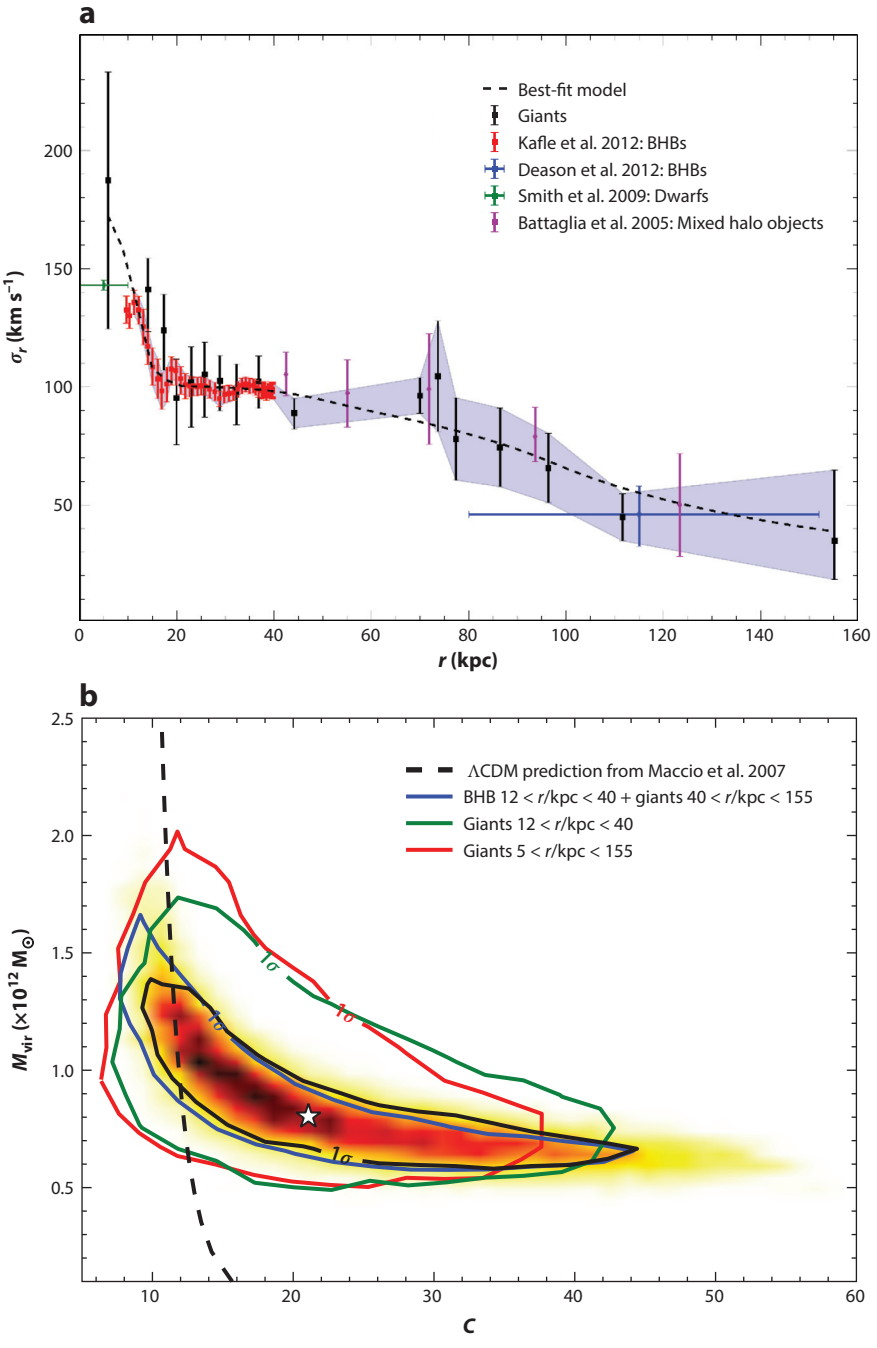

Figure 12

(*a*) Radial velocity dispersion as a function of radius for halo stars in the Milky Way. (*b*) Posterior distribution of virial mass and the concentration parameter of the Milky Way's dark matter halo. Adapted from Kafle et al. (2014) with permission. Abbreviation: BHB, blue horizontal branch.

θ, the model makes a prediction for radial velocity dispersion $\sigma_{vr}(r_i;\theta)$ at a location r_i. This can be compared with the $\sigma_{vr}(r_i)$ estimated from the observed data. The probability of model parameters θ is then given by

$$p(\theta|D) \propto p(\theta)\prod_{i=1}^{M}\mathcal{N}[\sigma_{vr}(r_i)|\sigma_{vr}(r_i;\theta),\gamma_i]. \tag{103}$$

The posterior distribution for the virial mass and the concentration parameter of the Milky Way halo using blue horizontal branch and giant stars is shown in **Figure 12** (Kafle et al. 2014).

We now discuss ways to incorporate prior information into the analysis (McMillan 2017). For example, the angular velocity of the Sun with respect to the Galactic Center ω is well constrained to be within 30.24 ± 0.12 km s^{-1} kpc^{-1} (Reid & Brunthaler 2004). The vertical force at 1.1 kpc above the Sun, in terms of surface mass density, is given by $\Sigma_{1.1,\odot} = 72 \pm 6$ (Kuijken & Gilmore 1991). Let us denote such constraints by $p(g_j(\theta)|\theta)$. Additional data sets D_k, constraining a certain subset of parameters, can also exist. For example, the tangent point velocities or terminal velocities as a function of Galactic longitude $v_{\rm term}(\ell)$ help to constrain the shape of the circular velocity curve $v_{\rm circ}(R) = \sqrt{|Rd\Phi/dR|}$. The additional priors and data all enter as multiplicative factors in the posterior, which is given by

$$p(\theta|D_1,\ldots,D_K) \propto p(\theta)\prod_{k=1}^{K}p(D_k|\theta)\prod_{j=1}^{J}p(g_j(\theta)|\theta). \tag{104}$$

The halo stars carry little information about the mass distribution close to the center and in the disc of the Milky Way. Galactic masers associated with high-mass star-forming regions are very good tracers of the Milky Way disc, which makes them excellent candidates for studying the potential of the Milky Way (Reid et al. 2009; McMillan 2011, 2017; Reid et al. 2014). Due to extremely accurate astrometric information using very long baseline interferometry, one has very accurate parallax (ϖ) and proper motion measurements. When combined with line-of-sight velocities from Doppler shift of spectral lines, one ends up with full 6D phase space information for these sources. Maser sources are young and have very little random motion, which means their orbits are highly circularized. The distribution of velocities can be described by a simple 3D Gaussian function,

$$p(v_R,v_\phi,v_z|\theta) = \mathcal{N}[v_\phi|v_{\rm circ}(R;\theta)+v_{\phi,\rm M},\sigma_{v\rm M}]\mathcal{N}(v_R|v_{R,\rm M},\sigma_{v\rm M})\mathcal{N}(v_z|v_{z,\rm M},\sigma_{v\rm M}). \tag{105}$$

Here, $\mathbf{v}_M = (v_{R,\rm M}, v_{\phi,\rm M}, v_{z,\rm M})$ is any systematic streaming velocity associated with the masers and $\sigma_{v\rm M}$ is the velocity dispersion about the mean motion. Now, we have

$$\begin{aligned} p(\mu_\alpha,\mu_\delta,v_{\rm los}|\theta) = \int {\rm d}\varpi' p(\varpi'|\varpi)\int\int\int {\rm d}\mu'_\alpha\; p(\mu'_\alpha|\mu_\alpha){\rm d}\mu'_\delta p(\mu'_\delta|\mu_\delta) \\ {\rm d}v'_{\rm los}p(v'_{\rm los}|v_{\rm los})p(\mu'_\alpha,\mu'_\delta,v'_{\rm los}|\theta,\varpi'). \end{aligned} \tag{106}$$

The last term is evaluated using Equation 105 by converting from heliocentric coordinates $(\mu'_\alpha,\mu'_\delta,v'_{\rm los},\alpha,\delta,\varpi')$ to Galactocentric coordinates $(v_R,v_\phi,v_z,R,\phi,z)$. Let D_1 denote the full data of N stars; then

$$p(D_1|\theta) = \prod_{i=1}^{N}p(\mu_{\alpha,i},\mu_{\delta,i},v_{\rm los,i}|\theta). \tag{107}$$

This, when put in Equation 104, gives the posterior distribution of model parameters.

6. CONCLUDING REMARKS

The power of the Bayesian probability theory lies in the fact that it is mathematically simple, being based on just two elementary rules, and yet it is broadly applicable. However, Bayesian

calculations can be computationally demanding, and this has acted as a major bottleneck in the past. But with the increase of computational power, we have witnessed a sharp increase in the adoption of Bayesian techniques. More recently, free availability of black-box computer packages to efficiently sample from Bayesian posterior distributions has further accelerated the adoption of Bayesian techniques in astronomy.

Robust algorithms are now available to sample multidimensional and complex pdfs. The MH algorithm is still the main workhorse of MCMC methods. Good solutions now exist for the issue of application-specific tuning of the proposal distribution in the MH algorithm, e.g., adaptive Metropolis schemes and the ensemble samplers. The MH algorithm, when combined with parallel tempering, allows one to sample a wide variety of commonly occurring distributions. Situations in which the posterior is not analytically tractable can also now be solved using the Monte Carlo version of the MH algorithm. Bayesian methods also provide a framework for model comparison via the use of Bayesian evidence. However, efficient computing of evidence still remains a challenge. Various alternate criteria for comparing models exist, and importantly, these can make use of the computed MCMC chain.

BHMs further increase the usefulness of the Bayesian framework. They can solve missing data problems, marginalization over variables, convolution with observational uncertainties and so on. This makes a wide class of complex problems suddenly solvable. We showed that the MWG scheme is ideally suited for sampling posteriors generated by BHMs, and we also provide a software for doing this.

Multimodal distributions still pose a problem for most MCMC algorithms. Parallel tempering can overcome them but requires more computational time and a careful choice of ladder. If dimensionality of the space being explored is high and the distribution is complex, efficient exploration is not easy. Techniques are being developed to solve such problems that make use of derivatives of the posterior distribution, e.g., Hamiltonian Monte Carlo. However, more work is required in this area. Efficient exploration of multilevel hierarchical models will play an increasingly important role in future studies.

Communication of Bayesian results is also an area where we anticipate improvements. Traditionally, the estimates are reported by means of confidence intervals. However, there is much more information in the MCMC chain, in particular, the correlation between different variables. Also, there is an increasing need to feed results of one MCMC simulation into another. Such requirements are best addressed by reporting the full pdfs or the thinned samples from it. Other alternatives that are economical in terms of storage space are to approximate the pdf by analytical functions or to employ Gaussian mixture models. We also need better tools to visualize the Bayesian-MCMC output, especially for high-dimensional and complex hierarchical models. Such tools will allow us to understand why a model fails and how we should improve it.

There are key topics that we have not addressed here. Nonparametric Bayesian methods are becoming increasingly important, e.g., Gaussian processes (Beaumont et al. 2002) and Dirichlet process mixture models (Neal 2000). Magorrian (2014) uses this method to estimate the gravitational potential of the Milky Way.

Astronomy is no longer a data-starved science. With projects like the Large Synoptic Survey Telescope and the Square Kilometre Array, the quality and quantity of data are going to increase dramatically in the coming years. Better quality and larger quantity of data mean that we can expect our data to answer more difficult questions, which in turn means more complex models (e.g., multilevel hierarchies and a higher dimensional parameter space). Given that MCMC is a computationally expensive scheme, there will be an increasing demand for techniques that can make full use of the vast quantity of data on offer and deliver results in an affordable amount of time.

Equivalently, MCMC schemes that make use of computing environments with multiple processor and graphic processor units would also be useful. An MCMC chain is serial by nature, and it requires special care to parallelize an MCMC algorithm, e.g., use of an ensemble of chains (Foreman-Mackey et al. 2013) or parallelizing the posterior computation by splitting up the data. Relaxing the condition of reversibility can lead to MCMC algorithms with faster mixing properties (Chen et al. 1999, Diaconis et al. 2000, Girolami & Calderhead 2011). Finally, the development of approximate methods, both application specific and general, that can reduce the computational cost without significantly compromising the quality of results also holds great promise for analyzing large data sets. Approximate Bayesian computation is one such framework (Beaumont et al. 2002); see Bovy (2016) for its use in astronomy to study the chemical homogeneity of stars in open clusters.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I am indebted to my colleague Joss Bland-Hawthorn for suggesting this article and for supervising its development over the past year. I am thankful to James Binney, Jo Bovy, Brendon Brewer, Prajwal Kafle, and Prasenjit Saha for numerous suggestions and discussions from which the review has benefited significantly. I am also thankful to David Hogg for words of encouragement on the draft. I acknowledge funding from a University of Sydney Senior Fellowship made possible by the office of the Deputy Vice Chancellor of Research, and partial funding from Bland-Hawthorn's Laureate Fellowship from the Australian Research Council.

LITERATURE CITED

Akaike H. 1974. *IEEE Trans. Autom. Control* 19:716–23
Andreon S, Weaver B. 2015. *Bayesian methods for the physical sciences. Springer Series in Astrostatistics*. Dordrecht, Neth.: Springer, Cham.
Andrieu C, Robert CP. 2001. *Controlled MCMC for optimal sampling*. CiteSeerX. **http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.23.2048**
Andrieu C, Roberts GO. 2009. *Ann. Stat.* 37:697–725
Andrieu C, Thoms J. 2008. *Stat. Comput.* 18:343–73
Barker A. 1965. *Aust. J. Phys.* 18:119–34
Battaglia G, Helmi A, Morrison H, et al. 2005. *MNRAS*. 364:433
Bayes M, Price M. 1763. *Philos. Trans.* 53:370–418
Beaumont MA. 2003. *Genetics* 164:1139–60
Beaumont MA, Zhang W, Balding DJ. 2002. *Genetics* 162:2025–35
Besag J. 1974. *J. R. Stat. Soc. Ser. B (Methodol.)* 36:192–236
Binney J. 2011. *Pramana* 77:39–52
Binney J. 2013. *New Astron. Rev.* 57:29–51
Binney J, Burnett B, Kordopatis G, et al. 2014. *MNRAS* 437:351–70
Bland-Hawthorn J, Gerhard O. 2016. *Annu. Rev. Astron. Astrophys.* 54:529–96
Bovy J. 2016. *Ap. J* 817:49
Bovy J, Allende Prieto C, Beers TC, et al. 2012a. *Ap. J* 759:131
Bovy J, Rix HW. 2013. *Ap. J* 779:115
Bovy J, Rix HW, Liu C, et al. 2012b. *Ap. J* 753:148

Brewer BJ, Pártay LB, Csányi G. 2011. *Stat. Comput.* 21:649–56
Brooks S, Gelman A, Jones GL, Meng X-L. 2011. *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: Chapman and Hall/CRC
Brooks SP, Gelman A. 1998. *J. Comput. Graph. Stat.* 7:434–55
Brown TM, Gilliland RL, Noyes RW, Ramsey LW. 1991. *Ap. J* 368:599–609
Burnett B, Binney J. 2010. *MNRAS* 407:339–54
Burnett B, Binney J, Sharma S. 2011. *Astron. Astrophys.* 532:A113
Burnham KP, Anderson DR. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag. 2nd ed.
Cappé O, Douc R, Guillin A, Marin JM, Robert CP. 2008. *Stat. Comput.* 18:447–59
Celeux G, Diebolt J. 1985. *Comput. Stat. Q.* 2:73–82
Chaplin WJ, Kjeldsen H, Christensen-Dalsgaard J, et al. 2011. *Science* 332:213–16
Chaplin WJ, Miglio A. 2013. *Annu. Rev. Astron. Astrophys.* 51:353–92
Chen F, Lovász L, Pak I. 1999. In *Proc. 31st Ann. ACM Symp. Theory Comput., Atlanta, GA*, May 1–4, pp. 275–91. New York: ACM
Christen JA, Fox C. 2010. *Bayesian Anal.* 5:263–81
Christensen N, Meyer R. 1998. *Phys. Rev. D* 58:082001
Christensen N, Meyer R, Knox L, Luey B. 2001. *Class. Quantum Gravity* 18:2677–88
Cowles M, Carlin B. 1996. *J. Am. Stat. Assoc.* 91:883–904
Cox RT. 1946. *Am. J. Phys.* 14:1–13
Cubillos P, Harrington J, Loredo TJ, et al. 2017. *Astron. J.* 153:3
de Laplace P. 1774. *Mém. Acad. R. Sci. Paris* 6:353–71
Deason AJ, Belokurov V, Evans NW, et al. 2012. *MNRAS* 425:2840
Dempster AP, Laird NM, Rubin DB. 1977. *J. R. Stat. Soc. Ser. B (Methodol.)* 39:1–38
Diaconis P, Holmes S, Neal RM. 2000. *Ann. Appl. Probab.* 10:726–52
Dose V. 2002. Tech. Rep. 83, CIPS, MPI für Plasmaphysik, Garching, Ger.
Duane S, Kennedy AD, Pendleton BJ, Roweth D. 1987. *Phys. Lett. B* 195:216–22
Duvall T Jr., Harvey J. 1986. In *Seismology of the Sun and the Distant Stars. NATO Adv. Res. Workshop, Cambridge, United Kingdom*, June 17–21, pp. 105–16. Dordrecht, Neth.: Reidel
Eadie GM, Harris WE, Widrow LM. 2015. *Ap. J.* 806:54
Feroz F, Hobson MP, Bridges M. 2009. *MNRAS* 398:1601–14
Ford EB, Gregory PC. 2007. In *Statistical Challenges in Modern Astronomy IV*, ed. GJ Babu, ED Feigelson. *ASP Conf. Ser.* 371:189–205. San Francisco: ASP
Foreman-Mackey D, Hogg DW, Lang D, Goodman J. 2013. *Publ. Astron. Soc. Pac.* 125:306–12
Gelfand AE, Smith AF. 1990. *J. Am. Stat. Assoc.* 85:398–409
Gelman A, Carlin J, Stern H, et al. 2013. *Bayesian Data Analysis.* London, UK: Taylor & Francis. 3rd ed.
Gelman A, Hwang J, Vehtari A. 2014. *Stat. Comput.* 24:997–1016
Gelman A, Roberts G, Gilks W. 1996. *Bayesian Stat.* 5:599–608
Gelman A, Rubin DB. 1992. *Stat. Sci.* 7:457–72
Geman S, Geman D. 1984. *IEEE Trans. Pattern Anal. Mach. Intell.* 6:721–41
Geyer CJ. 1992. *Stat. Sci.* 7:473–83
Gilks WR, Roberts GO, George EI. 1994. *Statistician* 43:179–89
Girolami M, Calderhead B. 2011. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 73:123–214
Goodman J, Sokal AD. 1989. *Phys. Rev. D* 40:2035
Goodman J, Weare J. 2010. *Commun. Appl. Math. Comput. Sci.* 5:65–80
Green GM, Schlafly EF, Finkbeiner DP, et al. 2014. *Ap. J* 783:114
Green GM, Schlafly EF, Finkbeiner DP, et al. 2015. *Ap. J* 810:25
Green PJ. 1995. *Biometrika* 82:711–32
Gregory P. 2005. *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica® Support*. Cambridge, UK: Cambridge Univ. Press
Griewank A, Walther A. 2008. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Philadelphia, PA: Siam

Gruberbauer M, Kallinger T, Weiss WW, Guenther DB. 2009. *Astron. Astrophys.* 506:1043–53
Haario H, Saksman E, Tamminen J. 2001. *Bernoulli* 7:223–42
Hajian A. 2007. *Phys. Rev. D* 75:083525
Handberg R, Campante TL. 2011. *Astron. Astrophys.* 527:A56
Hastings WK. 1970. *Biometrika* 57:97–109
Hobson MP. 2010. *Bayesian Methods in Cosmology*. Cambridge, UK: Cambridge Univ. Press
Hogg DW, Bovy J, Lang D. 2010. arXiv:1008.4686
Hogg DW, Myers AD, Bovy J. 2010. *Ap. J* 725:2166–75
Homan MD, Gelman A. 2014. *J. Mach. Learn. Res.* 15:1593–623
Jaynes ET. 1957. *Phys. Rev.* 106:620
Jaynes ET. 1999. *Straight line fitting—a Bayesian solution.* **http://bayes.wustl.edu/etj/articles/leapz.pdf**
Jaynes ET. 2003. *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge Univ. Press
Jeffreys H. 1939. *Theory of Probability*. Oxford, UK: Oxford Univ. Press
Jørgensen BR, Lindegren L. 2005. *Astron. Astrophys.* 436:127–43
Kafle PR, Sharma S, Lewis GF, Bland-Hawthorn J. 2012. *Ap. J.* 761:98
Kafle PR, Sharma S, Lewis GF, Bland-Hawthorn J. 2014. *Ap. J.* 794:59
Kallinger T, Mosser B, Hekker S, et al. 2010. *Astron. Astrophys.* 522:A1
Kass RE, Raftery AE. 1995. *J. Am. Stat. Assoc.* 90:773–95
Kass RE, Wasserman L. 1996. *J. Am. Stat. Assoc.* 91:1343–70
Kilbinger M, Wraith D, Robert CP, et al. 2010. *MNRAS* 405:2381–90
Kirkpatrick S, Gelatt CD, Vecchi MP, et al. 1983. *Science* 220:671–80
Kjeldsen H, Bedding TR. 1995. *Astron. Astrophys.* 293:87–106
Knox L, Christensen N, Skordis C. 2001. *Ap. J. Lett.* 563:L95–98
Kuijken K, Gilmore G. 1991. *Ap. J. Lett.* 367:L9–13
Lewis A, Bridle S. 2002. *Phys. Rev. D* 66:103511
Liang F, Liu C, Carroll R. 2011. *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*, Vol. 714. Chichester, UK: John Wiley & Sons
Liang F, Wong WH. 2001a. *J. Am. Stat. Assoc.* 96:653–66
Liang F, Wong WH. 2001b. *J. Chem. Phys.* 115:3374–80
Lindley DV. 1957. *Biometrika* 44:187–92
Loredo TJ. 1990. In *Maximum Entropy and Bayesian Methods*, ed. J Skilling, pp. 81–142. Dordrecht, Neth.: Springer Sci.+Bus. Media
Macciò AV, Dutton AA, van den Bosch FC, et al. 2007. *MNRAS* 378:55
MacEachern SN, Berliner LM. 1994. *Am. Stat.* 48:188–90
MacKay DJ. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge, UK: Cambridge Univ. Press
Magorrian J. 2014. *MNRAS* 437:2230–48
McMillan PJ. 2011. *MNRAS* 414:2446–57
McMillan PJ. 2017. *MNRAS* 465:76–94
McMillan PJ, Binney J. 2012. *MNRAS* 419:2251–63
McMillan PJ, Binney JJ. 2013. *MNRAS* 433:1411–24
Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. *J. Chem. Phys.* 21:1087–92
Metropolis N, Ulam S. 1949. *J. Am. Stat. Assoc.* 44:335–41
Møller J, Pettitt AN, Reeves R, Berthelsen KK. 2006. *Biometrika* 93:451–58
Mukherjee P, Parkinson D, Liddle AR. 2006. *Ap. J. Lett.* 638:L51–54
Müller P. 1991. Tech. Rep. 91-09, Dep. Stat., Purdue Univ.
Murray I, Ghahramani Z, MacKay D. 2006. *Proc. 22nd Conf. Uncertainty Artif. Intell., Cambridge, Mass., July 13–16*, pp. 359–66. Arlington, VA: AUAI
Neal RM. 1993. Tech. Rep. CRG-TR-93-1, Dep. Comput. Sci., Univ. Toronto
Neal RM. 2000. *J. Comput. Graph. Stat.* 9:249–65
Neal RM. 2011. In *Handb. Markov Chain Monte Carlo*, ed. S Brooks, A Gelman, GL Jones, X-L Meng, pp. 113–62. Boca Raton, FL: Chapman and Hall/CRC
Nelson B, Ford EB, Payne MJ. 2014. *Ap. J. Suppl.* 210:11

Nelson BE, Robertson PM, Payne MJ, et al. 2016. *MNRAS* 455:2484–99
Ness M, Hogg DW, Rix HW, Ho AYQ, Zasowski G. 2015. *Ap. J* 808:16
Parkinson D, Liddle AR. 2013. *Stat. Anal. Data Min.* 6:3–14
Parkinson D, Mukherjee P, Liddle AR. 2006. *Phys. Rev. D* 73:123523
Peskun PH. 1973. *Biometrika* 60:607–12
Piffl T, Binney J, McMillan PJ, et al. 2014. *MNRAS* 445:3133–51
Pont F, Eyer L. 2004. *MNRAS* 351:487–504
Press WH. 1997. In *Unsolved Problems in Astrophysics*, ed. JN Bahcall, JP Ostriker, pp. 49–60. Princeton, NJ: Princeton Univ. Press
Raftery AE, Lewis S. 1992. *Bayesian Stat.* 4:763–73
Reid MJ, Brunthaler A. 2004. *Ap. J* 616:872–84
Reid MJ, Menten KM, Brunthaler A, et al. 2014. *Ap. J* 783:130
Reid MJ, Menten KM, Zheng XW, et al. 2009. *Ap. J* 700:137–48
Rix HW, Bovy J. 2013. *Astron. Astrophys. Rev.* 21:61
Robbins H, Monro S. 1951. *Ann. Math. Stat.* 22:400–7
Robert C, Casella G. 2011. In *Handb. Markov Chain Monte Carlo*, ed. S Brooks, A Gelman, GL Jones, X-L Meng, pp. 49–66. Boca Raton, FL: Chapman and Hall/CRC
Robert CP, Casella G. 2004. *Monte Carlo Statistical Methods*. New York: Springer Sci. Bus. Med. 2nd ed.
Roberts G, Gilks W. 1994. *J. Multivar. Anal.* 49:287–98
Saha P, Williams TB. 1994. *Astron. J.* 107:1295–302
Sale SE. 2012. *MNRAS* 427:2119–31
Sale SE. 2015. *MNRAS* 452:2960–72
Sale SE, Magorrian J. 2015. *MNRAS* 448:1738–50
Sanders JL, Binney J. 2015. *MNRAS* 449:3479–502
Schönrich R, Binney J, Dehnen W. 2010. *MNRAS* 403:1829–33
Schwarz G. 1978. *Ann. Stat.* 6:461–64
Sharma S, Bland-Hawthorn J, Binney J, et al. 2014. *Ap. J* 793:51
Sharma S, Stello D, Bland-Hawthorn J, Huber D, Bedding TR. 2016. *Ap. J* 822:15
Sivia D, Skilling J. 2006. *Data Analysis: A Bayesian Tutorial*. Oxford, UK: Oxford Univ. Press
Skilling J. 2006. *Bayesian Anal.* 1:833–59
Smith AF, Spiegelhalter DJ. 1980. *J. R. Stat. Soc. Ser. B* 42:213–20
Smith MC, Evans NW, Belokurov V, et al. 2009. *MNRAS* 399:1223
Sokal A. 1997. In *Functional Integration*, Vol. 361, *NATO ASI Series*, ed. C DeWitt-Morette, P Cartier, A Folacci, pp. 131–92. New York: Springer Sci.+Bus. Media
Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. 2002. *J. R. Stat. Soc. Ser. B* 64:583–639
Stello D, Huber D, Bedding TR, et al. 2013. *Ap. J. Lett.* 765:L41
Stello D, Huber D, Sharma S, et al. 2015. *Ap. J. Lett.* 809:L3
Tanner MA, Wong WH. 1987. *J. Am. Stat. Assoc.* 82:528–40
Tanner MA, Wong WH. 2010. *Stat. Sci.* 25:506–16
Taylor JF, Ashdown MAJ, Hobson MP. 2008. *MNRAS* 389:1284–92
Ter Braak CJ. 2006. *Stat. Comput.* 16:239–49
Trotta R. 2008. *Contemp. Phys.* 49:71–104
Ulrich RK. 1986. *Ap. J. Lett.* 306:L37–40
Vehtari A, Ojanen J. 2012. *Stat. Surv.* 6:142–228
Verdinelli I, Wasserman L. 1995. *J. Am. Stat. Assoc.* 90:614–18
Wasserstein RL, Lazar NA. 2016. *Am. Stat.* 70(2):129–33
Watanabe S. 2010. *J. Mach. Learn. Res.* 11:3571–94
Watanabe S. 2013. *J. Mach. Learn. Res.* 14:867–97
Weinberg MD. 2012. *Bayesian Anal.* 7:737–70
Wolfgang A, Rogers LA, Ford EB. 2016. *Ap. J.* 825:19
Wraith D, Kilbinger M, Benabed K, et al. 2009. *Phys. Rev. D* 80:023507

Contents

Annual Review of Astronomy and Astrophysics

Volume 55, 2017

Indexes

Errata

An online log of corrections to *Annual Review of Astronomy and Astrophysics* articles may be found at http://www.annualreviews.org/errata/astro