Title

stata.com

Intro — Introduction to Bayesian analysis

Description Remarks and examples References Also see

Description

This entry provides a software-free introduction to Bayesian analysis. See [BAYES] **Bayesian commands** for an overview of the software for performing Bayesian analysis and for an overview example.

For Bayesian model averaging, which is Bayesian analysis that averages over multiple plausible models, see [BMA] **Intro**.

Remarks and examples

stata.com

Remarks are presented under the following headings:

What is Bayesian analysis?
Bayesian versus frequentist analysis, or why Bayesian analysis?
How to do Bayesian analysis
Advantages and disadvantages of Bayesian analysis
Brief background and literature review
Bayesian statistics
- *Posterior distribution*
- *Selecting priors*
- *Point and interval estimation*
- *Comparing Bayesian models*
- *Posterior prediction*

Bayesian computation
Markov chain Monte Carlo methods
- *Metropolis–Hastings algorithm*
- *Adaptive random-walk Metropolis–Hastings*
- *Blocking of parameters*
- *Metropolis–Hastings with Gibbs updates*
- *Convergence diagnostics of MCMC*

Summary
Video examples

The first five sections provide a general introduction to Bayesian analysis. The remaining sections provide a more technical discussion of the concepts of Bayesian analysis.

What is Bayesian analysis?

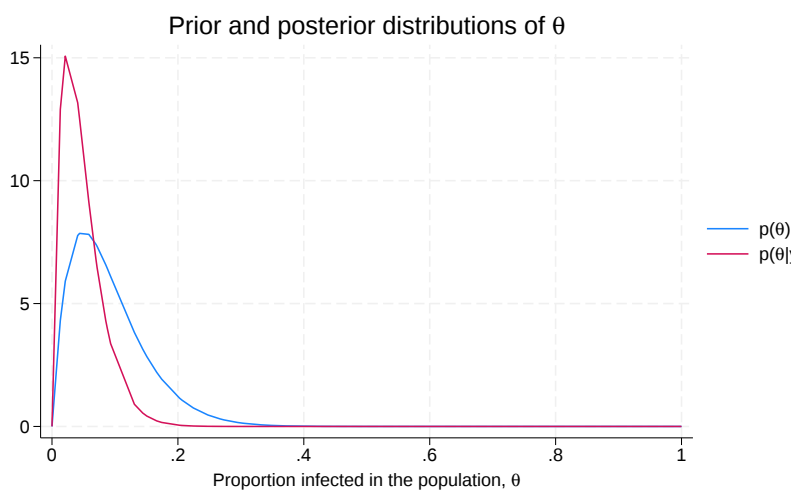
Bayesian analysis is a statistical analysis that answers research questions about unknown parameters of statistical models by using probability statements. Bayesian analysis rests on the assumption that all model parameters are random quantities and thus can incorporate prior knowledge. This assumption is in sharp contrast with the more traditional, also called frequentist, statistical inference where all parameters are considered unknown but fixed quantities. Bayesian analysis follows a simple rule of probability, the Bayes rule, which provides a formalism for combining prior information with evidence from the data at hand. The Bayes rule is used to form the so called posterior distribution of model parameters. The posterior distribution results from updating the prior knowledge about model

parameters with evidence from the observed data. Bayesian analysis uses the posterior distribution to form various summaries for the model parameters including point estimates such as posterior means, medians, percentiles, and interval estimates such as credible intervals. Moreover, all statistical tests about model parameters can be expressed as probability statements based on the estimated posterior distribution.

As a quick introduction to Bayesian analysis, we use an example, described in Hoff (2009, 3), of estimating the prevalence of a rare infectious disease in a small city. A small random sample of 20 subjects from the city will be checked for infection. The parameter of interest $\theta \in [0,1]$ is the fraction of infected individuals in the city. Outcome y records the number of infected individuals in the sample. A reasonable sampling model for y is a binomial model: $y|\theta \sim \text{Binomial}(20,\theta)$. Based on the studies from other comparable cities, the infection rate ranged between 0.05 and 0.20, with an average prevalence of 0.10. To use this information, we must conduct Bayesian analysis. This information can be incorporated into a Bayesian model with a prior distribution for θ, which assigns a large probability between 0.05 and 0.20, with the expected value of θ close to 0.10. One potential prior that satisfies this condition is a $\text{Beta}(2,20)$ prior with the expected value of $2/(2+20) = 0.09$. So, let's assume this prior for the infection rate θ, that is, $\theta \sim \text{Beta}(2,20)$. We sample individuals and observe none who have an infection, that is, $y = 0$. This value is not that uncommon for a small sample and a rare disease. For example, for a true rate $\theta = 0.05$, the probability of observing 0 infections in a sample of 20 individuals is about 36% according to the binomial distribution. So, our Bayesian model can be defined as follows:

$$y|\theta \sim \text{Binomial}(20,\theta)$$
$$\theta \sim \text{Beta}(2,20)$$

For this Bayesian model, we can actually compute the posterior distribution of $\theta|y$, which is $\theta|y \sim \text{Beta}(2+0, 20+20-0) = \text{Beta}(2,40)$. The prior and posterior distributions of θ are depicted below.



The posterior density (shown in red) is more peaked and shifted to the left compared with the prior distribution (shown in blue). The posterior distribution combined the prior information about θ with the information from the data, from which $y = 0$ provided evidence for a low value of θ and shifted the prior density to the left to form the posterior density. Based on this posterior distribution, the

posterior mean estimate of θ is $2/(2+40) = 0.048$ and the posterior probability that, for example, $\theta < 0.10$ is about 93%.

If we compute a standard frequentist estimate of a population proportion θ as a fraction of the infected subjects in the sample, $\overline{y} = y/n$, we will obtain 0 with the corresponding 95% confidence interval $(\overline{y} - 1.96\sqrt{\overline{y}\,(1-\overline{y})/n}, \overline{y} + 1.96\sqrt{\overline{y}\,(1-\overline{y})/n})$ reducing to 0 as well. It may be difficult to convince a health policy maker that the prevalence of the disease in that city is indeed 0, given the small sample size and the prior information available from comparable cities about a nonzero prevalence of this disease.

We used a beta prior distribution in this example, but we could have chosen another prior distribution that supports our prior knowledge. For the final analysis, it is important to consider a range of different prior distributions and investigate the sensitivity of the results to the chosen priors.

For more details about this example, see Hoff (2009). Also see *Beta-binomial model* in [BAYES] **bayesmh** for how to fit this model using `bayesmh`.

Rabe-Hesketh and Skrondal (2022, chap. 16) and Cameron and Trivedi (2022, chap. 28) provide introductions to Bayesian analysis with Stata-specific examples.

Bayesian versus frequentist analysis, or why Bayesian analysis?

Why use Bayesian analysis? Perhaps a better question is when to use Bayesian analysis and when to use frequentist analysis. The answer to this question mainly lies in your research problem. You should choose an analysis that answers your specific research questions. For example, if you are interested in estimating the probability that the parameter of interest belongs to some prespecified interval, you will need the Bayesian framework, because this probability cannot be estimated within the frequentist framework. If you are interested in a repeated-sampling inference about your parameter, the frequentist framework provides that.

Bayesian and frequentist approaches have very different philosophies about what is considered fixed and, therefore, have very different interpretations of the results. The Bayesian approach assumes that the observed data sample is fixed and that model parameters are random. The posterior distribution of parameters is estimated based on the observed data and the prior distribution of parameters and is used for inference. The frequentist approach assumes that the observed data are a repeatable random sample and that parameters are unknown but fixed and constant across the repeated samples. The inference is based on the sampling distribution of the data or of the data characteristics (statistics). In other words, Bayesian analysis answers questions based on the distribution of parameters conditional on the observed sample, whereas frequentist analysis answers questions based on the distribution of statistics obtained from repeated hypothetical samples, which would be generated by the same process that produced the observed sample given that parameters are unknown but fixed. Frequentist analysis consequently requires that the process that generated the observed data is repeatable. This assumption may not always be feasible. For example, in meta-analysis, where the observed sample represents the collected studies of interest, one may argue that the collection of studies is a one-time experiment.

Frequentist analysis is entirely data-driven and strongly depends on whether or not the data assumptions required by the model are met. On the other hand, Bayesian analysis provides a more robust estimation approach by using not only the data at hand but also some existing information or knowledge about model parameters.

In frequentist statistics, estimators are used to approximate the true values of the unknown parameters, whereas Bayesian statistics provides an entire distribution of the parameters. In our example of a prevalence of an infectious disease from *What is Bayesian analysis?*, frequentist analysis produced one point estimate for the prevalence, whereas Bayesian analysis estimated the entire posterior distribution of the prevalence based on a given sample.

Frequentist inference is based on the sampling distributions of estimators of parameters and provides parameter point estimates and their standard errors as well as confidence intervals. The exact sampling distributions are rarely known and are often approximated by a large-sample normal distribution. Bayesian inference is based on the posterior distribution of the parameters and provides summaries of this distribution including posterior means and their MCMC standard errors (MCSE) as well as credible intervals. Although exact posterior distributions are known only in a number of cases, general posterior distributions can be estimated via, for example, Markov chain Monte Carlo (MCMC) sampling without any large-sample approximation.

Frequentist confidence intervals do not have straightforward probabilistic interpretations as do Bayesian credible intervals. For example, the interpretation of a 95% confidence interval is that if we repeat the same experiment many times and compute confidence intervals for each experiment, then 95% of those intervals will contain the true value of the parameter. For any given confidence interval, the probability that the true value is in that interval is either zero or one, and we do not know which. We may only infer that any given confidence interval provides a plausible range for the true value of the parameter. A 95% Bayesian credible interval, on the other hand, provides a range for a parameter such that the probability that the parameter lies in that range is 95%.

Frequentist hypothesis testing is based on a deterministic decision using a prespecified significance level of whether to accept or reject the null hypothesis based on the observed data, assuming that the null hypothesis is actually true. The decision is based on a p-value computed from the observed data. The interpretation of the p-value is that if we repeat the same experiment and use the same testing procedure many times, then given our null hypothesis is true, we will observe the result (test statistic) as extreme or more extreme than the one observed in the sample $(100 \times p\text{-value})\%$ of the times. The p-value cannot be interpreted as a probability of the null hypothesis, which is a common misinterpretation. In fact, it answers the question of how likely are our data given that the null hypothesis is true, and not how likely is the null hypothesis given our data. The latter question can be answered by Bayesian hypothesis testing, where we can compute the probability of any hypothesis of interest.

How to do Bayesian analysis

Bayesian analysis starts with the specification of a posterior model. The posterior model describes the probability distribution of all model parameters conditional on the observed data and some prior knowledge. The posterior distribution has two components: a likelihood, which includes information about model parameters based on the observed data, and a prior, which includes prior information (before observing the data) about model parameters. The likelihood and prior models are combined using the Bayes rule to produce the posterior distribution:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

If the posterior distribution can be derived in a closed form, we may proceed directly to the inference stage of Bayesian analysis. Unfortunately, except for some special models, the posterior distribution is rarely available explicitly and needs to be estimated via simulations. MCMC sampling can be used to simulate potentially very complex posterior models with an arbitrary level of precision. MCMC methods for simulating Bayesian models are often demanding in terms of specifying an efficient sampling algorithm and verifying the convergence of the algorithm to the desired posterior distribution. See [BAYES] **Bayesian estimation**.

Inference is the next step of Bayesian analysis. If MCMC sampling is used for approximating the posterior distribution, the convergence of MCMC must be established before proceeding to inference (see, for example, [BAYES] **bayesgraph** and [BAYES] **bayesstats grubin**). Point and interval estimators

are either derived from the theoretical posterior distribution or estimated from a sample simulated from the posterior distribution. Many Bayesian estimators, such as posterior mean and posterior standard deviation, involve integration. If the integration cannot be performed analytically to obtain a closed-form solution, sampling techniques such as Monte Carlo integration and MCMC and numerical integration are commonly used. See [BAYES] **Bayesian postestimation** and [BAYES] **bayesstats**.

Another important step of Bayesian analysis is model checking, which is typically performed via posterior predictive checking. The idea behind posterior predictive checking is the comparison of various aspects of the distribution of the observed data with those of the replicated data. Replicated data are simulated from the posterior predictive distribution of the fitted Bayesian model under the same conditions that generated the observed data, such as the same values of covariates, etc. The discrepancy between the distributions of the observed and replicated data is measured by test quantities (functions of the data and model parameters) and is quantified by so-called posterior predictive p-values. See [BAYES] **bayesstats ppvalues** and [BAYES] **bayespredict**.

Bayesian hypothesis testing can take two forms, which we refer to as interval-hypothesis testing and model-hypothesis testing. In an interval-hypothesis testing, the probability that a parameter or a set of parameters belongs to a particular interval or intervals is computed. In model hypothesis testing, the probability of a Bayesian model of interest given the observed data is computed. See [BAYES] **bayestest**.

Model comparison is another common step of Bayesian analysis. The Bayesian framework provides a systematic and consistent approach to model comparison using the notion of posterior odds and related to them Bayes factors. See [BAYES] **bayesstats ic** for details.

Finally, prediction of some future unobserved data may also be of interest in Bayesian analysis. The prediction of a new data point is performed conditional on the observed data using the so-called posterior predictive distribution, which involves integrating out all parameters from the model with respect to their posterior distribution. Again, Monte Carlo integration is often the only feasible option for obtaining predictions. Prediction can also be helpful in estimating the goodness of fit of a model. See [BAYES] **bayespredict**.

Advantages and disadvantages of Bayesian analysis

Bayesian analysis is a powerful analytical tool for statistical modeling, interpretation of results, and prediction of data. It can be used when there are no standard frequentist methods available or the existing frequentist methods fail. However, one should be aware of both the advantages and disadvantages of Bayesian analysis before applying it to a specific problem.

The universality of the Bayesian approach is probably its main methodological advantage to the traditional frequentist approach. Bayesian inference is based on a single rule of probability, the Bayes rule, which is applied to all parametric models. This makes the Bayesian approach universal and greatly facilitates its application and interpretation. The frequentist approach, however, relies on a variety of estimation methods designed for specific statistical problems and models. Often, inferential methods designed for one class of problems cannot be applied to another class of models.

In Bayesian analysis, we can use previous information, either belief or experimental evidence, in a data model to acquire more balanced results for a particular problem. For example, incorporating prior information can mitigate the effect of a small sample size. Importantly, the use of the prior evidence is achieved in a theoretically sound and principled way.

By using the knowledge of the entire posterior distribution of model parameters, Bayesian inference is far more comprehensive and flexible than the traditional inference.

Bayesian inference is exact, in the sense that estimation and prediction are based on the posterior distribution. The latter is either known analytically or can be estimated numerically with an arbitrary

precision. In contrast, many frequentist estimation procedures such as maximum likelihood rely on the assumption of asymptotic normality for inference.

Bayesian inference provides a straightforward and more intuitive interpretation of the results in terms of probabilities. For example, credible intervals are interpreted as intervals to which parameters belong with a certain probability, unlike the less straightforward repeated-sampling interpretation of the confidence intervals.

Bayesian models satisfy the likelihood principle (Berger and Wolpert 1988) that the information in a sample is fully represented by the likelihood function. This principle requires that if the likelihood function of one model is proportional to the likelihood function of another model, then inferences from the two models should give the same results. Some researchers argue that frequentist methods that depend on the experimental design may violate the likelihood principle.

Finally, as we briefly mentioned earlier, the estimation precision in Bayesian analysis is not limited by the sample size—Bayesian simulation methods may provide an arbitrary degree of precision.

Despite the conceptual and methodological advantages of the Bayesian approach, its application in practice is still considered controversial sometimes. There are two main reasons for this—the presumed subjectivity in specifying prior information and the computational challenges in implementing Bayesian methods. Along with the objectivity that comes from the data, the Bayesian approach uses potentially subjective prior distribution. That is, different individuals may specify different prior distributions. Proponents of frequentist statistics argue that for this reason, Bayesian methods lack objectivity and should be avoided. Indeed, there are settings such as clinical trial cases when the researchers want to minimize a potential bias coming from preexisting beliefs and achieve more objective conclusions. Even in such cases, however, a balanced and reliable Bayesian approach is possible. The trend in using noninformative priors in Bayesian models is an attempt to address the issue of subjectivity. On the other hand, some Bayesian proponents argue that the classical methods of statistical inference have built-in subjectivity such as a choice for a sampling procedure, whereas the subjectivity is made explicit in Bayesian analysis.

Building a reliable Bayesian model requires extensive experience from the researchers, which leads to the second difficulty in Bayesian analysis—setting up a Bayesian model and performing analysis is a demanding and involving task. This is true, however, to an extent for any statistical modeling procedure.

Lastly, one of the main disadvantages of Bayesian analysis is the computational cost. As a rule, Bayesian analysis involves intractable integrals that can only be computed using intensive numerical methods. Most of these methods such as MCMC are stochastic by nature and do not comply with the natural expectation from a user of obtaining deterministic results. Using simulation methods does not compromise the discussed advantages of Bayesian approach, but unquestionably adds to the complexity of its application in practice.

For more discussion about advantages and disadvantages of Bayesian analysis, see, for example, Thompson (2012), Bernardo and Smith (2000), and Berger and Wolpert (1988).

Brief background and literature review

The principles of Bayesian analysis date back to the work of Thomas Bayes, who was a Presbyterian minister in Tunbridge Wells and Pierre Laplace, a French mathematician, astronomer, and physicist in the 18th century. Bayesian analysis started as a simple intuitive rule, named after Bayes, for updating beliefs on account of some evidence. For the next 200 years, however, Bayes's rule was just an obscure idea. Along with the rapid development of the standard or frequentist statistics in 20th century, Bayesian methodology was also developing, although with less attention and at a slower pace. One of the obstacles for the progress of Bayesian ideas has been the lasting opinion among mainstream

statisticians of it being subjective. Another more-tangible problem for adopting Bayesian models in practice has been the lack of adequate computational resources. Nowadays, Bayesian statistics is widely accepted by researchers and practitioners as a valuable and feasible alternative.

Bayesian analysis proliferates in diverse areas including industry and government, but its application in sciences and engineering is particularly visible. Bayesian statistical inference is used in econometrics (Poirier [1995]; Chernozhukov and Hong [2003]; Kim, Shephard, and Chib [1998], Zellner [1997]); education (Johnson 1997); epidemiology (Greenland 1998); engineering (Godsill and Rayner 1998); genetics (Iversen, Parmigiani, and Berry 1999); social sciences (Pollard 1986); hydrology (Parent et al. 1998); quality management (Rios Insua 1990); atmospheric sciences (Berliner et al. 1999); and law (DeGroot, Fienberg, and Kadane 1986), to name a few.

The subject of general statistics has been greatly influenced by the development of Bayesian ideas. Bayesian methodologies are now present in biostatistics (Carlin and Louis [2009]; Berry and Stangl [1996]); generalized linear models (Dey, Ghosh, and Mallick 2000); hierarchical modeling (Hobert 2000); statistical design (Chaloner and Verdinelli 1995); classification and discrimination (Neal [1996]; Neal [1999]); graphical models (Pearl 1998); nonparametric estimation (Müller and Vidakovic [1999]; Dey, Müller, and Sinha [1998]); survival analysis (Barlow, Clarotti, and Spizzichino 1993); sequential analysis (Carlin, Kadane, and Gelfand 1998); predictive inference (Aitchison and Dunsmore 1975); spatial statistics (Wolpert and Ickstadt [1998]; Besag and Higdon [1999]); testing and model selection (Kass and Raftery [1995]; Berger and Pericchi [1996]; Berger [2006]); and time series (Pole, West, and Harrison [1994]; West and Harrison [1997]).

Recent advances in computing allowed practitioners to perform Bayesian analysis using simulations. The simulation tools came from outside the statistics field—Metropolis et al. (1953) developed what is now known as a random-walk Metropolis algorithm to solve problems in statistical physics. Another landmark discovery was the Gibbs sampling algorithm (Geman and Geman 1984), initially used in image processing, which showed that exact sampling from a complex and otherwise intractable probability distribution is possible. These ideas were the seeds that led to the development of Markov chain Monte Carlo (MCMC)—a class of iterative simulation methods proved to be indispensable tools for Bayesian computations. Starting from the early 1990s, MCMC-based techniques slowly emerged in the mainstream statistical practice. More powerful and specialized methods appeared, such as perfect sampling (Propp and Wilson 1996), reversible-jump MCMC (Green 1995) for traversing variable dimension state spaces, and particle systems (Gordon, Salmond, and Smith 1993). Consequent widespread application of MCMC was imminent (Berger 2000) and influenced various specialized fields. For example, Gelman and Rubin (1992) investigated MCMC for the purpose of exploring posterior distributions; Geweke (1999) surveyed simulation methods for Bayesian inference in econometrics; Kim, Shephard, and Chib (1998) used MCMC simulations to fit stochastic volatility models; Carlin, Kadane, and Gelfand (1998) implemented Monte Carlo methods for identifying optimal strategies in clinical trials; Chib and Greenberg (1995) provided Bayesian formulation of a number of important econometrics models; and Chernozhukov and Hong (2003) reviewed some econometrics models involving Laplace-type estimators from an MCMC perspective. For more comprehensive exposition of MCMC, see, for example, Robert and Casella (2004); Tanner (1996); Gamerman and Lopes (2006); Chen, Shao, and Ibrahim (2000); and Brooks et al. (2011).

Bayesian statistics

Posterior distribution

To formulate the principles of Bayesian statistics, we start with a simple case when one is concerned with the interaction of two random variables, $\mathbf{A}$ and $\mathbf{B}$. Let $p(\cdot)$ denote either a probability mass function or a density, depending on whether the variables are discrete or continuous. The rule of conditional probability,

$$p(\mathbf{A}|\mathbf{B}) = \frac{p(\mathbf{A}, \mathbf{B})}{p(\mathbf{B})}$$

can be used to derive the so-called Bayes's theorem:

$$p(\mathbf{B}|\mathbf{A}) = \frac{p(\mathbf{A}|\mathbf{B})p(\mathbf{B})}{p(\mathbf{A})} \tag{1}$$

This rule also holds in the more general case when $\mathbf{A}$ and $\mathbf{B}$ are random vectors.

In a typical statistical problem, we have a data vector $\mathbf{y}$, which is assumed to be a sample from a probability model with an unknown parameter vector $\boldsymbol{\theta}$. We represent this model using the likelihood function $L(\boldsymbol{\theta}; \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$, where $f(y_i|\boldsymbol{\theta})$ denotes the probability density function of y_i given $\boldsymbol{\theta}$. We want to infer some properties of $\boldsymbol{\theta}$ based on the data $\mathbf{y}$. In Bayesian statistics, model parameters $\boldsymbol{\theta}$ is a random vector. We assume that $\boldsymbol{\theta}$ has a probability distribution $p(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$, which is referred to as a prior distribution. Because both $\mathbf{y}$ and $\boldsymbol{\theta}$ are random, we can apply Bayes's theorem (1) to derive the posterior distribution of $\boldsymbol{\theta}$ given data $\mathbf{y}$,

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} = \frac{f(\mathbf{y}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(\mathbf{y})} \tag{2}$$

where $m(\mathbf{y}) \equiv p(\mathbf{y})$, known as the marginal distribution of $\mathbf{y}$, is defined by

$$m(\mathbf{y}) = \int f(\mathbf{y}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{3}$$

The marginal distribution $m(\mathbf{y})$ in (3) does not depend on the parameter of interest $\boldsymbol{\theta}$, and we can, therefore, reduce (2) to

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto L(\boldsymbol{\theta}; \mathbf{y})\pi(\boldsymbol{\theta}) \tag{4}$$

Equation (4) is fundamental in Bayesian analysis and states that the posterior distribution of model parameters is proportional to their likelihood and prior probability distributions. We will often use (4) in the computationally more-convenient log-scale form

$$\ln\{p(\boldsymbol{\theta}|\mathbf{y})\} = l(\boldsymbol{\theta}; \mathbf{y}) + \ln\{\pi(\boldsymbol{\theta})\} - c \tag{5}$$

where $l(\cdot; \cdot)$ denotes the log likelihood of the model. Depending on the analytical procedure involving the log-posterior $\ln\{p(\boldsymbol{\theta}|\mathbf{y})\}$, the actual value of the constant $c = \ln\{m(\mathbf{y})\}$ may or may not be relevant. For valid statistical analysis, however, we will always assume that c is finite.

Selecting priors

In Bayesian analysis, we seek a balance between prior information in a form of expert knowledge or belief and evidence from data at hand. Achieving the right balance is one of the difficulties in Bayesian modeling and inference. In general, we should not allow the prior information to overwhelm the evidence from the data, especially when we have a large data sample. A famous theoretical result, the Bernstein–von Mises theorem, states that in large data samples, the posterior distribution is independent of the prior distribution and, therefore, Bayesian and likelihood-based inferences should yield essentially the same results. On the other hand, we need a strong enough prior to support weak evidence that usually comes from insufficient data. It is always good practice to perform sensitivity analysis to check the dependence of the results on the choice of a prior.

The flexibility of choosing the prior freely is one of the main controversial issues associated with Bayesian analysis and the reason why some practitioners view the latter as subjective. It is also the reason why the Bayesian practice, especially in the early days, was dominated by noninformative priors. Noninformative priors, also called flat or vague priors, assign equal probabilities to all possible states of the parameter space with the aim of rectifying the subjectivity problem. One of the disadvantages of flat priors is that they are often improper; that is, they do not specify a legitimate probability distribution. For example, a uniform prior for a continuous parameter over an unbounded domain does not integrate to a finite number. However, this is not necessarily a problem because the corresponding posterior distribution may still be proper. Although Bayesian inference based on improper priors is possible, this is equivalent to discarding the terms $\log\pi(\boldsymbol{\theta})$ and c in (5), which nullifies the benefit of Bayesian analysis because it reduces the latter to an inference based only on the likelihood. This is why there is a strong objection to the practice of noninformative priors. In recent years, an increasing number of researchers have advocated the use of sound informative priors, for example, Thompson (2014). For example, using informative priors is mandatory in areas such as genetics, where prior distributions have a physical basis and reflect scientific knowledge.

Another convenient preference for priors is to use conjugate priors. Their choice is desirable from technical and computational standpoints but may not necessarily provide a realistic representation of the model parameters. Because of the limited arsenal of conjugate priors, an inclination to overuse them severely limits the flexibility of Bayesian modeling.

Point and interval estimation

In Bayesian statistics, inference about parameters $\boldsymbol{\theta}$ is based on the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ and various ways of summarizing this distribution. Point and interval estimates can be used to summarize this distribution.

Commonly used point estimators are the posterior mean,

$$E(\boldsymbol{\theta}|\mathbf{y}) = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$$

and the posterior median, $q_{0.5}(\boldsymbol{\theta})$, which is the 0.5 quantile of the posterior; that is,

$$P\{\boldsymbol{\theta} \leq q_{0.5}(\boldsymbol{\theta}|\mathbf{y})\} = 0.5$$

Another point estimator is the posterior mode, which is the value of $\boldsymbol{\theta}$ that maximizes $p(\boldsymbol{\theta}|\mathbf{y})$.

Interval estimation is performed by constructing so-called credible intervals (CrIs). CrIs are special cases of credible regions. Let $1-\alpha \in (0,1)$ be some predefined credible level. Then, an $\{(1-\alpha)\times 100\}\%$ credible set R of θ is such that

$$\Pr(\theta \in R|\mathbf{y}) = \int_R p(\theta|\mathbf{y})d\theta = 1 - \alpha$$

We consider two types of CrIs. The first one is based on quantiles. The second one is the highest posterior density (HPD) interval.

An $\{(1-\alpha) \times 100\}\%$ quantile-based, or also known as an equal-tailed CrI, is defined as $(q_{\alpha/2}, q_{1-\alpha/2})$, where q_a denotes the ath quantile of the posterior distribution. A commonly reported equal-tailed CrI is $(q_{0.025}, q_{0.975})$.

HPD interval is defined as an $\{(1-\alpha) \times 100\}\%$ CrI of the shortest width. As its name implies, this interval corresponds to the region of the posterior density with the highest concentration. For a unimodal posterior distribution, HPD is unique, but for a multimodal distribution it may not be unique. Computational approaches for calculating HPD are described in Chen and Shao (1999) and Eberly and Casella (2003).

Comparing Bayesian models

Model comparison is another important aspect of Bayesian statistics. We are often interested in comparing two or more plausible models for our data.

Let's assume that we have models M_j parameterized by vectors $\boldsymbol{\theta}_j$, $j = 1, \ldots, r$. We may have varying degree of belief in each of these models given by prior probabilities $p(M_j)$, such that $\sum_{j=1}^{r} p(M_j) = 1$. By applying Bayes's theorem, we find the posterior model probabilities

$$p(M_j|\mathbf{y}) = \frac{p(\mathbf{y}|M_j)p(M_j)}{p(\mathbf{y})}$$

where $p(\mathbf{y}|M_j) = m_j(y)$ is the marginal likelihood of M_j with respect to $\mathbf{y}$. Because of the difficulty in calculating $p(\mathbf{y})$, it is a common practice to compare two models, say, M_j and M_k, using the posterior odds ratio

$$\text{PO}_{jk} = \frac{p(M_j|\mathbf{y})}{p(M_k|\mathbf{y})} = \frac{p(\mathbf{y}|M_j)p(M_j)}{p(\mathbf{y}|M_k)p(M_k)}$$

If all models are equally plausible, that is, $p(M_j) = 1/r$, the posterior odds ratio reduces to the so-called Bayes factors (BF) (Jeffreys 1935),

$$\text{BF}_{jk} = \frac{p(\mathbf{y}|M_j)}{p(\mathbf{y}|M_k)} = \frac{m_j(\mathbf{y})}{m_k(\mathbf{y})}$$

which are simply ratios of marginal likelihoods.

Jeffreys (1961) recommended an interpretation of BF_{jk} based on half-units of the log scale. The following table provides some rules of thumb:

$\log_{10}(\text{BF}_{jk})$	BF_{jk}	Evidence against M_k
0 to 1/2	1 to 3.2	Bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
>2	>100	Decisive

The Schwarz criterion BIC (Schwarz 1978) is an approximation of BF in case of arbitrary but proper priors. Kass and Raftery (1995) and Berger (2006) provide a detailed exposition of Bayes factors, their calculation, and their role in model building and testing.

Posterior prediction

Prediction is another essential part of statistical analysis. In Bayesian statistics, prediction is performed using the posterior predictive distribution. The probability of observing some future data $\mathbf{y}^*$ given the observed data $\mathbf{y}$ can be obtained by the marginalization of

$$p(\mathbf{y}^*|\mathbf{y}) = \int p(\mathbf{y}^*|\mathbf{y}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

which, assuming that $\mathbf{y}^*$ is independent of $\mathbf{y}$ given $\boldsymbol{\theta}$, can be simplified to

$$p(\mathbf{y}^*|\mathbf{y}) = \int p(\mathbf{y}^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \tag{6}$$

Equation (6) is called a posterior predictive distribution and is used for Bayesian prediction. See [BAYES] **bayespredict** and [BAYES] **bayesstats ppvalues**.

Bayesian computation

An unavoidable difficulty in performing Bayesian analysis is the need to compute integrals such as those expressing marginal distributions and posterior moments. The integrals involved in Bayesian inference are of the form $E\{g(\boldsymbol{\theta})\} = \int g(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$ for some function $g(\cdot)$ of the random vector $\boldsymbol{\theta}$. With the exception of a few cases for which analytical integration is possible, the integration is performed via simulations.

Given a sample from the posterior distribution, we can use Monte Carlo integration to approximate the integrals. Let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_T$ be an independent sample from $p(\boldsymbol{\theta}|\mathbf{y})$.

The original integral of interest $E\{g(\boldsymbol{\theta})\}$ can be approximated by

$$\widehat{g} = \frac{1}{T}\sum_{t=1}^{T} g(\boldsymbol{\theta}_t)$$

Moreover, if g is a scalar function, under some mild conditions, the central limit theorem holds

$$\widehat{g} \approx N\left[E\{g(\boldsymbol{\theta})\}, \sigma^2/T\right]$$

where $\sigma^2 = \mathrm{Cov}\{g(\boldsymbol{\theta}_i)\}$ can be approximated by the sample variance $\sum_{t=1}^{T}\{g(\boldsymbol{\theta}_t) - \widehat{g}\}^2/T$. If the sample is not independent, then $\widehat{g}$ still approximates $E\{g(\boldsymbol{\theta})\}$ but the variance σ^2 is given by

$$\sigma^2 = \mathrm{Var}\{g(\boldsymbol{\theta}_t)\} + 2\sum_{k=1}^{\infty}\mathrm{Cov}\{g(\boldsymbol{\theta}_t), g(\boldsymbol{\theta}_{t+k})\} \tag{7}$$

and needs to be approximated. Moreover, the conditions needed for the central limit theorem to hold involve the convergence rate of the chain and can be difficult to check in practice (Tierney 1994).

The Monte Carlo integration method solves the problem of Bayesian computation of computing a posterior distribution by sampling from that posterior distribution. The latter has been an important problem in computational statistics and a focus of intense research. Rejection sampling techniques serve as basic tools for generating samples from a general probability distribution (von Neumann 1951). They are based on the idea that samples from the target distribution can be obtained from another, easy-to-sample distribution according to some acceptance–rejection rule for the samples from this distribution. It was soon recognized, however, that the acceptance–rejection methods did not scale well with the increase of dimensions, a problem known as the "curse of dimensionality", essentially reducing the acceptance probability to zero. An alternative solution was to use the Markov chains to generate sequences of correlated sample points from the domain of the target distribution and keeping a reasonable rate of acceptance. It was not long before Markov chain Monte Carlo methods were accepted as effective tools for approximate sampling from general posterior distributions (Tanner and Wong 1987).

Markov chain Monte Carlo methods

Every MCMC method is designed to generate values from a transition kernel such that the draws from that kernel converge to a prespecified target distribution. It simulates a Markov chain with the target distribution as the stationary or equilibrium distribution of the chain. By definition, a Markov chain is any sequence of values or states from the domain of the target distribution, such that each value depends on its immediate predecessor only. For a well-designed MCMC, the longer the chain, the closer the samples to the stationary distribution. MCMC methods differ substantially in their simulation efficiency and computational complexity.

The Metropolis algorithm proposed in Metropolis and Ulam (1949) and Metropolis et al. (1953) appears to be the earliest version of MCMC. The algorithm generates a sequence of states, each obtained from the previous one, according to a Gaussian proposal distribution centered at that state. Hastings (1970) described a more-general version of the algorithm, now known as a Metropolis–Hastings (MH) algorithm, which allows any distribution to be used as a proposal distribution. Below we review the general MH algorithm and some of its special cases.

Metropolis–Hastings algorithm

Here we present the MH algorithm for sampling from a posterior distribution in a general formulation. It requires the specification of a proposal probability distribution $q(\cdot)$ and a starting state $\boldsymbol{\theta}_0$ within the domain of the posterior, that is, $p(\boldsymbol{\theta}_0|\mathbf{y}) > 0$. The algorithm generates a Markov chain $\{\boldsymbol{\theta}_t\}_{t=0}^{T-1}$ such that at each step t 1) a proposal state $\boldsymbol{\theta}_*$ is generated conditional on the current state, and 2) $\boldsymbol{\theta}_*$ is accepted or rejected according to the suitably defined acceptance probability.

For $t = 1, \ldots, T-1$:

1. Generate a proposal state: $\boldsymbol{\theta}_* \sim q(\cdot|\boldsymbol{\theta}_{t-1})$.
2. Calculate the acceptance probability $\alpha(\boldsymbol{\theta}_*|\boldsymbol{\theta}_{t-1}) = \min\{r(\boldsymbol{\theta}_*|\boldsymbol{\theta}_{t-1}), 1\}$, where

$$r(\boldsymbol{\theta}_*|\boldsymbol{\theta}_{t-1}) = \frac{p(\boldsymbol{\theta}_*|\mathbf{y})q(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_*)}{p(\boldsymbol{\theta}_{t-1}|\mathbf{y})q(\boldsymbol{\theta}_*|\boldsymbol{\theta}_{t-1})}$$

3. Draw $u \sim \text{Uniform}(0, 1)$.
4. Set $\boldsymbol{\theta}_t = \boldsymbol{\theta}_*$ if $u < \alpha(\boldsymbol{\theta}_*|\boldsymbol{\theta}_{t-1})$, and $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$ otherwise.

We refer to the iteration steps 1 through 4 as an MH update. By design, any Markov chain simulated using this MH algorithm is guaranteed to have $p(\boldsymbol{\theta}|\mathbf{y})$ as its stationary distribution.

Two important criteria measuring the efficiency of MCMC are the acceptance rate of the chain and the degree of autocorrelation in the generated sample. When the acceptance rate is close to 0, then most of the proposals are rejected, which means that the chain failed to explore regions of appreciable posterior probability. The other extreme is when the acceptance probability is close to 1, in which case the chain stays in a small region and fails to explore the whole posterior domain. An efficient MCMC has an acceptance rate that is neither too small nor too large and also has small autocorrelation. Gelman, Gilks, and Roberts (1997) showed that in the case of a multivariate posterior and proposal distributions, an acceptance rate of 0.234 is asymptotically optimal and, in the case of a univariate posterior, the optimal value is 0.45.

A special case of MH employs a Metropolis update with $q(\cdot)$ being a symmetric distribution. Then, the acceptance ratio reduces to a ratio of posterior probabilities,

$$r(\boldsymbol{\theta}_*|\boldsymbol{\theta}_{t-1}) = \frac{p(\boldsymbol{\theta}_*|\mathbf{y})}{p(\boldsymbol{\theta}_{t-1}|\mathbf{y})}$$

The symmetric Gaussian distribution is a common choice for a proposal distribution $q(\cdot)$, and this is the one used in the original Metropolis algorithm.

Another important MCMC method that can be viewed as a special case of MH is Gibbs sampling (Gelfand et al. 1990), where the updates are the full conditional distributions of each parameter given the rest of the parameters. Gibbs updates are always accepted. If $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^d)$ and, for $j = 1 \ldots, d$, q_j is the conditional distribution of θ^j given the rest $\boldsymbol{\theta}^{\{-j\}}$, then the Gibbs algorithm is the following. For $t = 1, \ldots, T-1$ and for $j = 1, \ldots, d$: $\theta_t^j \sim q_j(\cdot|\boldsymbol{\theta}_{t-1}^{\{-j\}})$. This step is referred to as a Gibbs update.

All MCMC methods share some limitations and potential problems. First, any simulated chain is influenced by its starting values, especially for short MCMC runs. It is required that the starting point has a positive posterior probability, but even when this condition is satisfied, if we start somewhere in a remote tail of the target distribution, it may take many iterations to reach a region of appreciable probability. Second, because there is no obvious stopping criterion, it is not easy to decide for how long to run the MCMC algorithm to achieve convergence to the target distribution. Third, the observations in MCMC samples are strongly dependent and this must be taken into account in any subsequent statistical inference. For example, the errors associated with the Monte Carlo integration should be calculated according to (7), which accounts for autocorrelation.

Adaptive random-walk Metropolis–Hastings

The choice of a proposal distribution $q(\cdot)$ in the MH algorithm is crucial for the mixing properties of the resulting Markov chain. The problem of determining an optimal proposal for a particular target posterior distribution is difficult and is still being researched actively. All proposed solutions are based on some form of an adaptation of the proposal distribution as the Markov chain progresses, which is carefully designed to preserve the ergodicity of the chain, that is, its tendency to converge to the target distribution. These methods are known as adaptive MCMC methods (Haario, Saksman, and Tamminen [2001]; Giordani and Kohn [2010]; and Roberts and Rosenthal [2009]).

The majority of adaptive MCMC methods are random-walk MH algorithms with updates of the form: $\boldsymbol{\theta}_* = \boldsymbol{\theta}_{t-1} + Z_t$, where Z_t follows some symmetric distribution. Specifically, we consider a Gaussian random-walk MH algorithm with $Z_t \sim N(0, \rho^2\Sigma)$, where ρ is a scalar controlling the scale of random jumps for generating updates and Σ is a d-dimensional covariance matrix. One of the first important results regarding adaptation is from Gelman, Gilks, and Roberts (1997), where the authors derive the optimal scaling factor $\rho = 2.38/\sqrt{d}$ and note that the optimal Σ is the true covariance matrix of the target distribution.

Haario, Saksman, and Tamminen (2001) proposes Σ to be estimated by the empirical covariance matrix plus a small diagonal matrix $\epsilon \times I_d$ to prevent zero covariance matrices. Alternatively, Roberts and Rosenthal (2009) proposed a mixture of the two covariance matrices,

$$\Sigma_t = \beta\widehat{\Sigma} + (1-\beta)\Sigma_0$$

for some fixed covariance matrix Σ_0 and $\beta \in [0, 1]$.

Because the proposal distribution of an adaptive MH algorithm changes at each step, the ergodicity of the chain is not necessarily preserved. However, under certain assumptions about the adaptation procedure, the ergodicity does hold; see Roberts and Rosenthal (2007), Andrieu and Moulines (2006), Atchadé and Rosenthal (2005), and Giordani and Kohn (2010) for details.

Blocking of parameters

In the original MH algorithm, the update steps of generating proposals and applying the acceptance–rejection rule are performed for all model parameters simultaneously. For high-dimensional models, this may result in a poor mixing—the Markov chain may stay in the tails of the posterior distribution for long periods of time and traverse the posterior domain very slowly. Suboptimal mixing is manifested by either very high or very low acceptance rates. Adaptive MH algorithms are also prone to this problem, especially when model parameters have very different scales. An effective solution to this problem is called *blocking*—model parameters are separated into two or more subsets or blocks and MH updates are applied to each block separately in the order that the blocks are specified.

Let's separate a vector of parameters into B blocks: $\boldsymbol{\theta} = \{\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^B\}$. The version of the Gaussian random-walk MH algorithm with blocking is as follows.

Let T_0 be the number of burn-in iterations, T be the number of MCMC samples, and $\rho_b^2\Sigma^b$, $b = 1, \ldots, B$, be block-specific proposal covariance matrices. Let $\boldsymbol{\theta}_0$ be the starting point within the domain of the posterior, that is, $p(\boldsymbol{\theta}_0|\mathbf{y}) > 0$.

1. At iteration t, let $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$.
2. For a block of parameters $\boldsymbol{\theta}_t^b$:
   2.1. Let $\boldsymbol{\theta}_* = \boldsymbol{\theta}_t$. Generate a proposal for the bth block: $\boldsymbol{\theta}_*^b = \boldsymbol{\theta}_{t-1}^b + \epsilon$, where $\epsilon \sim N(0, \rho_b^2\Sigma^b)$.
   2.2. Calculate the acceptance ratio,
   $$r(\boldsymbol{\theta}_*|\boldsymbol{\theta}_t) = \frac{p(\boldsymbol{\theta}_*|\mathbf{y})}{p(\boldsymbol{\theta}_t|\mathbf{y})}$$
   where $\theta_* = (\boldsymbol{\theta}_t^1, \boldsymbol{\theta}_t^2, \ldots, \boldsymbol{\theta}_t^{b-1}, \boldsymbol{\theta}_*^b, \boldsymbol{\theta}_t^{b+1}, \ldots, \boldsymbol{\theta}_t^B)$.
   2.3. Draw $u \sim \text{Uniform}(0, 1)$.
   2.4. Let $\boldsymbol{\theta}_t^b = \boldsymbol{\theta}_*^b$ if $u < \min\{r(\boldsymbol{\theta}_*|\boldsymbol{\theta}_t), 1\}$.
3. Repeat step 2 for $b = 1, \ldots, B$.
4. Repeat steps 1 through 3 for $t = 1, \ldots, T + T_0 - 1$.
5. The final sequence is $\{\boldsymbol{\theta}_t\}_{t=T_0}^{T+T_0-1}$.

Blocking may not always improve efficiency. For example, separating all parameters in individual blocks (the so-called one-at-a-time update regime) can lead to slow mixing when some parameters are highly correlated. A Markov chain may explore the posterior domain very slowly if highly correlated parameters are updated independently. There are no theoretical results about optimal blocking, so

you will need to use your judgment when determining the best set of blocks for your model. As a rule, parameters that are expected to be highly correlated are specified in one block. This will generally improve mixing of the chain unless the proposal correlation matrix does not capture the actual correlation structure of the block. For example, if there are two parameters in the block that have very different scales, adaptive MH algorithms that use the identity matrix for the initial proposal covariance may take a long time to approximate the optimal proposal correlation matrix. The user should, therefore, consider not only the probabilistic relationship between the parameters in the model, but also their scales to determine an optimal set of blocks.

Metropolis–Hastings with Gibbs updates

The original Gibbs sampler updates each model parameter one at a time according to its full conditional distribution. We have already noted that Gibbs is a special case of the MH algorithm. Some of the advantages of Gibbs sampling include its high efficiency, because all proposals are automatically accepted, and that it does not require any additional tuning for proposal distributions in MH algorithms. Unfortunately, for most posterior distributions in practice, the full conditionals are either not available or are very difficult to sample from. It may be the case, however, that for some model parameters or groups of parameters, the full conditionals are available and are easy to generate samples from. This is done in a hybrid MH algorithm, which implements Gibbs updates for only some blocks of parameters. A hybrid MH algorithm combines Gaussian random-walk updates with Gibbs updates to improve the mixing of the chain.

The MH algorithm with blocking allows different samplers to be used for updating different blocks. If there is a group of model parameters with a conjugate prior (or semiconjugate prior), we can place this group of parameters in a separate block and use Gibbs sampling for it. This can greatly improve the overall sampling efficiency of the algorithm.

For example, suppose that the data are normally distributed with a known mean μ and that we specify an inverse-gamma prior for σ^2 with shape α and scale β, which are some fixed constants.

$$y \sim N(\mu, \sigma^2), \quad \sigma^2 \sim \text{InvGamma}(\alpha,\ \beta)$$

The full conditional distribution for σ^2 in this case is also an inverse-gamma distribution, but with different shape and scale parameters,

$$\sigma^2 \sim \text{InvGamma}\left\{\widetilde{\alpha} = \alpha + \frac{n}{2},\ \widetilde{\beta} = \beta + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}$$

where n is the data sample size. So, an inverse-gamma prior for the variance is a conjugate prior in this model. We can thus place σ^2 in a separate block and set up a Gibbs sampling for it using the above full conditional distribution.

See *Methods and formulas* in [BAYES] **bayesmh** for details.

Convergence diagnostics of MCMC

Checking convergence of MCMC is an essential step in any MCMC simulation. Bayesian inference based on an MCMC sample is valid only if the Markov chain has converged and the sample is drawn from the desired posterior distribution. It is important that we verify the convergence for all model parameters and not only for a subset of parameters of interest. One difficulty with assessing

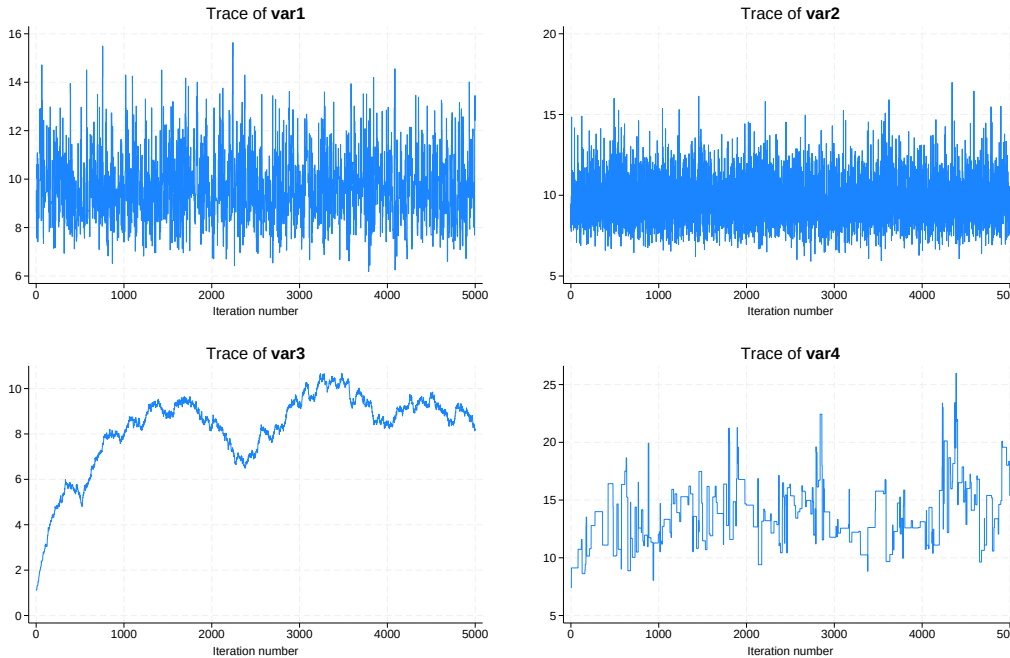
convergence of MCMC is that there is no single conclusive convergence criterion. The diagnostic usually involves checking for several necessary (but not necessarily sufficient) conditions for convergence. In general, the more aspects of the MCMC sample you inspect, the more reliable your results are.

The most extensive review of the methods for assessing convergence is Cowles and Carlin (1996). Other discussions about monitoring convergence can be found in Gelman et al. (2014) and Brooks et al. (2011).

There are at least two general approaches for detecting convergence issues. The first one is to inspect the mixing and time trends within the chains of individual parameters. The second one is to examine the mixing and time trends of multiple chains for each parameter. The lack of convergence in a Markov chain can be especially difficult to detect in a case of pseudoconvergence, which often occurs with multimodal posterior distributions. Pseudoconvergence occurs when the chain appears to have converged but it actually explored only a portion of the domain of a posterior distribution. To check for pseudoconvergence, Gelman and Rubin (1992) recommend running multiple chains from different starting states and comparing them; see [BAYES] **bayesstats grubin**.

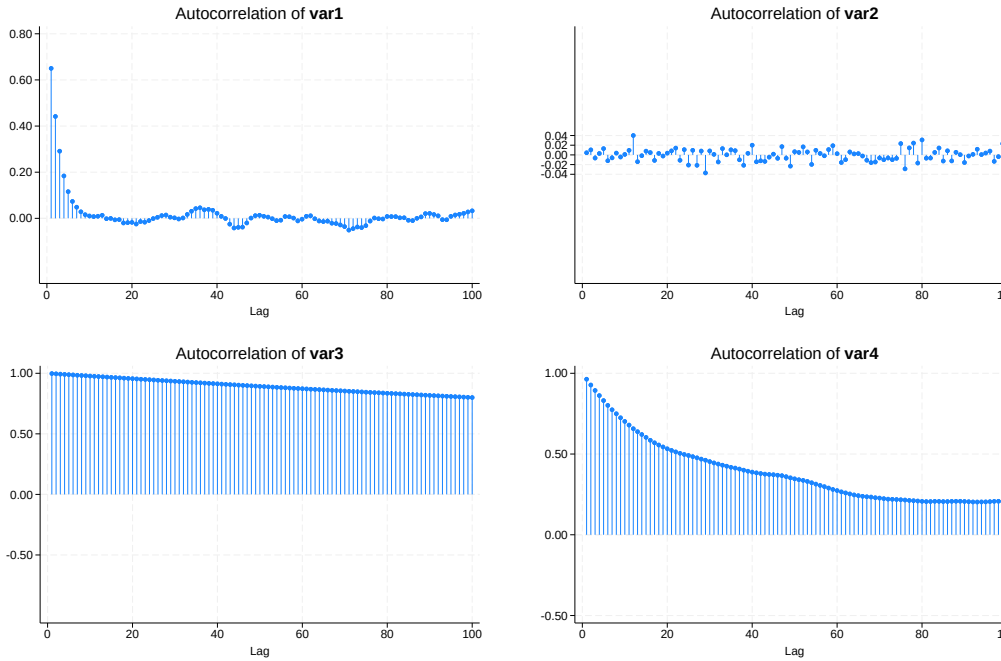
Trace plots are the most accessible convergence diagnostics and are easy to inspect visually. The trace plot of a parameter plots the simulated values for this parameter versus the iteration number. The trace plot of a well-mixing parameter should traverse the posterior domain rapidly and should have nearly constant mean and variance. See [BAYES] **bayesgraph** for details.

In the next figure, we show examples of trace plots for four parameters: `var1`, `var2`, `var3`, and `var4`. The first two parameters, `var1` and `var2`, have well-mixing chains, and the other two have poorly mixing chains. The chain for the parameter `var1` has a moderate acceptance rate, about 35%, and efficiency between 10% and 20%. This is a typical result for a Gaussian random-walk MH algorithm that has achieved convergence. The trace plot of `var2` in the top right panel shows almost perfect mixing—this is a typical example of Gibbs sampling with an acceptance rate close to 1 and efficiency above 95%. Although both chains traverse their marginal posterior domains, the right one does it more rapidly. On the downside, more efficient MCMC algorithms such as Gibbs sampling are usually associated with a higher computational cost.

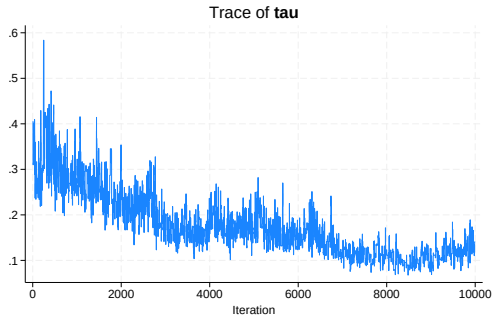


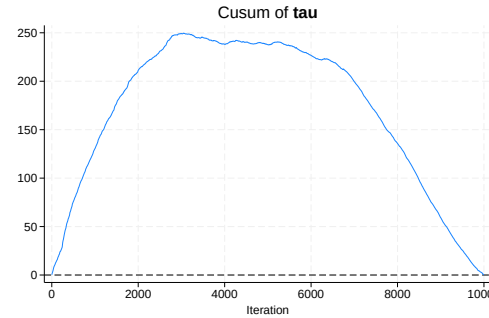
The bottom two trace plots illustrate cases of bad mixing and a lack of convergence. On the left, the chain for `var3` exhibits high acceptance rate but poor coverage of the posterior domain manifested by random drifting in isolated regions. This chain was produced by a Gaussian random-walk MH algorithm with a proposal distribution with a very small variance. On the right, the chain for the parameter `var4` has a very low acceptance rate, below 3%, because the used proposal distribution had a very large variance. In both cases, the chains do not converge; the simulation results do not represent the posterior distribution and should thus be discarded.

As we stated before, samples simulated using MCMC methods are correlated. The smaller the correlation, the more efficient the sampling process. Most of the MH algorithms typically generate highly correlated draws, whereas the Gibbs algorithm typically generates less-correlated draws. Below we show autocorrelation plots for the same four parameters using the same MCMC samples. The autocorrelation of `var1`, the one that comes from a well-mixing MH chain, becomes negligible fairly quickly, after about 10 lags. On the other hand, the autocorrelation of `var2` simulated using Gibbs sampling is essentially negligible for all positive lags. In the case of a poor mixing because of a small proposal variance (parameter `var3`), we observe very high positive correlation for at least 100 lags. The autocorrelation of `var4` is high but is lower than that of `var3`.

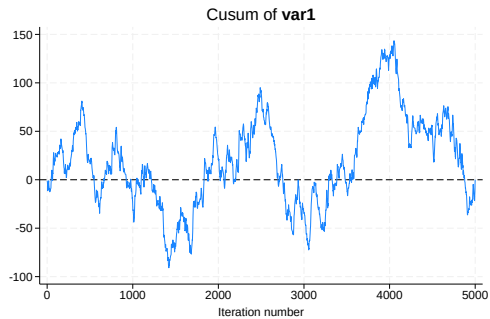


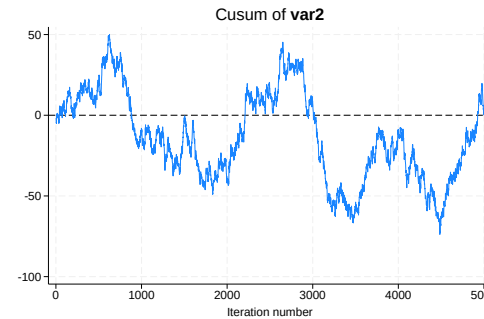
Yu and Mykland (1998) proposed a graphical procedure for assessing the convergence of individual parameters based on cumulative sums, also known as a cusum plot. By definition, any cusum plot starts at 0 and ends at 0. Cusum plots are useful for detecting drifts in the chain. For a chain without trend, the cusum plot should cross the x axis. For example, early drifts may indicate dependence on starting values. If we detect an early drift, we should discard an initial part of the chain and run it longer. Below, we show the trace plot of a poorly mixing parameter `tau` and its corresponding cusum plot on the right. There is an apparent positive drift for approximately the first half of the chain followed by the drift in the negative direction. As a result, the cusum plot has a distinctive mountain-like shape and never crosses the x axis.

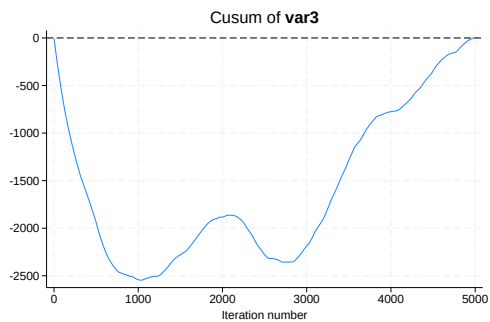


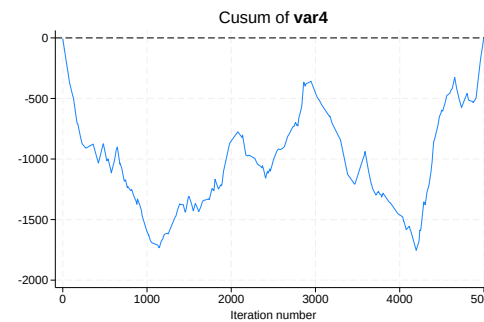


Cusum plots can be also used for assessing how fast the chain is mixing. The slower the mixing of the chain, the smoother the cusum plots. Conversely, the faster the mixing of the chain, the more jagged the cusum plots. Below, we demonstrate the cusum plots for the four variables considered previously. We can clearly see the contrast between the jagged lines of the fast mixing parameters `var1` and `var2` and the very smooth cusum line of the poorly mixing parameter `var3`.









Besides graphical convergence diagnostics, there are some formal convergence tests (Geweke [1992]; Gelman and Rubin [1992]; Heidelberger and Welch [1983]; Raftery and Lewis [1992]; Zellner and Min [1995]). See *Convergence diagnostics using multiple chains* in [BAYES] **bayesmh** and see [BAYES] **bayesstats grubin** for more details.

Summary

Bayesian analysis is a statistical procedure that answers research questions by expressing uncertainty about unknown parameters using probabilities. Bayesian inference is based on the posterior distribution of model parameters conditional on the observed data. The posterior distribution is composed of a likelihood distribution of the data and the prior distribution of the model parameters. The likelihood model is specified in the same way it is specified with any standard likelihood-based analysis. The prior distribution is constructed based on the prior (before observing the data) scientific knowledge and results from previous studies. Sensitivity analysis is typically performed to evaluate the influence of different competing priors on the results.

Many posterior distributions do not have a closed form and must be simulated using MCMC methods such as MH methods or the Gibbs method or sometimes their combination. The convergence of MCMC must be verified before any inference can be made.

Marginal posterior distributions of the parameters are used for inference. These are summarized using point estimators such as posterior mean and median and interval estimators such as equal-tailed credible intervals and highest posterior density intervals. Credible intervals have an intuitive interpretation as fixed ranges to which a parameter is known to belong with a prespecified probability. Hypothesis testing provides a way to assign an actual probability to any hypothesis of interest. A number of criteria are available for comparing models of interest. Predictions and model checking are also available based on the posterior predictive distribution.

Bayesian analysis provides many advantages over the standard frequentist analysis, such as an ability to incorporate prior information in the analysis, higher robustness to sparse data, more-comprehensive inference based on the knowledge of the entire posterior distribution, and more intuitive and direct interpretations of results by using probability statements about parameters.

Video examples

Introduction to Bayesian statistics, part 1: The basic concepts

Introduction to Bayesian statistics, part 2: MCMC and the Metropolis–Hastings algorithm

Thomas Bayes (1701(?)–1761) was a Presbyterian minister with an interest in calculus, geometry, and probability theory. He was born in Hertfordshire, England. The son of a Nonconformist minister, Bayes was banned from English universities and so studied at Edinburgh University before becoming a clergyman himself. Only two works are attributed to Bayes during his lifetime, both published anonymously. He was admitted to the Royal Society in 1742 and never published thereafter.

The paper that gives us "Bayes's Theorem" was published posthumously by Richard Price. The theorem has become an important concept for frequentist and Bayesian statisticians alike. However, the paper indicates that Bayes considered the theorem as relatively unimportant. His main interest appears to have been that probabilities were not fixed but instead followed some distribution. The notion, now foundational to Bayesian statistics, was largely ignored at the time.

Whether Bayes's theorem is appropriately named is the subject of much debate. Price acknowledged that he had written the paper based on information he found in Bayes's notebook, yet he never said how much he added beyond the introduction. Some scholars have also questioned whether Bayes's notes represent original work or are the result of correspondence with other mathematicians of the time.

Andrey Markov (1856–1922) was a Russian mathematician who made many contributions to mathematics and statistics. He was born in Ryazan, Russia. In primary school, he was known as a poor student in all areas except mathematics. Markov attended St. Petersburg University, where he studied under Pafnuty Chebyshev and later joined the physicomathematical faculty. He was a member of the Russian Academy of the Sciences.

Markov's first interest was in calculus. He did not start his work in probability theory until 1883 when Chebyshev left the university and Markov took over his teaching duties. A large and influential body of work followed, including applications of the weak law of large numbers and what are now known as Markov processes and Markov chains. His work on processes and chains would later influence the development of a variety of disciplines such as biology, chemistry, economics, physics, and statistics.

Known in the Russian press as the "militant academician" for his frequent written protests about the czarist government's interference in academic affairs, Markov spent much of his adult life at odds with Russian authorities. In 1908, he resigned from his teaching position in response to a government requirement that professors report on students' efforts to organize protests in the wake of the student riots earlier that year. He did not resume his university teaching duties until 1917, after the Russian Revolution. His trouble with Russian authorities also extended to the Russian Orthodox Church. In 1912, he was excommunicated at his own request in protest over the Church's excommunication of Leo Tolstoy.

Bruno de Finetti (1906–1985) was born in Innsbruck, Austria. He received a degree in applied mathematics from the Polytechnic University of Milan. One of his first publications was in the field of genetics, in which he introduced what is now called the de Finetti diagram. Upon graduation, he began working for the Italian Central Statistical Institute and later moved to Trieste to work as an actuary. He became a professor at the University of Trieste in 1947 and later became a professor of the theory of probability at the University of Rome "La Sapienza", a post he held for 15 years.

De Finetti made many contributions to the fields of probability and statistics. His text *Theory of Probability* helped lay the foundation for Bayesian theory. He also wrote papers on sequences of exchangeable random variables and processes with independent increments. In a paper published in 1955, de Finetti used an extension of the Lorenz–Gini concentration function to prove the Radon–Nikodym theorem. This extension has been employed in Bayesian statistics as a measure of robustness. His publications also include work on nonparametric estimation of a cumulative distribution function and group decision making, among other topics. For his many contributions, he was named a fellow of the Royal Statistical Society and the Institute of Mathematical Statistics.

David Harold Blackwell (1919–2010) was a world-renowned statistician and mathematician. At age 16, he began attending the University of Illinois, where he obtained a master's degree in mathematics and then a PhD in statistics at age 22. Shortly after, he joined Princeton University as a visiting fellow, becoming the university's first African-American faculty member and paving the way for future generations.

Blackwell is best known for developing the Rao–Blackwell theorem, used in statistics, and the Blackwell renewal theorem, used in engineering. In regard to Markov decision processes, he introduced the concepts of Blackwell optimality and positive and negative dynamic programs. His contributions also include pioneering texts, such as *Basic Statistics*, one of the first texts on Bayesian statistics, and *Theory of Games and Statistical Decisions*, which he coauthored with M. A. Girschick. Additionally, in 1949, he coauthored a paper that helped lay the groundwork for Bayesian sequential analysis. He published over 80 papers in many fields, including game theory, probability theory, and mathematical statistics.

Blackwell's contributions are also reflected in the honors bestowed upon him and in his leadership roles in prominent organizations. In 1976, he was elected an honorary fellow of the Royal Statistical Society, and in 1979, he won the John von Neumann Theory Prize. He also held 12 honorary degrees and was the first African-American man elected to the National Academy of Sciences. Additionally, he served as vice president of the American Statistical Association, American Mathematical Society, and the International Statistical Institute.

References

Aitchison, J., and I. R. Dunsmore. 1975. *Statistical Prediction Analysis*. Cambridge: Cambridge University Press.

Andrieu, C., and É. Moulines. 2006. On the ergodicity properties of some adaptive MCMC algorithms. *Annals of Applied Probability* 16: 1462–1505. https://doi.org/10.1214/105051606000000286.

Atchadé, Y. F., and J. S. Rosenthal. 2005. On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* 11: 815–828. https://doi.org/10.3150/bj/1130077595.

Barlow, R. E., C. A. Clarotti, and F. Spizzichino, ed. 1993. *Reliability and Decision Making*. Cambridge: Chapman and Hall.

Berger, J. O. 2000. Bayesian analysis: A look at today and thoughts of tomorrow. *Journal of the American Statistical Association* 95: 1269–1276. https://doi.org/10.2307/2669768.

——. 2006. "Bayes factors." In *Encyclopedia of Statistical Sciences*, edited by Kotz, S., C. B. Read, N. Balakrishnan, and B. Vidakovic. Wiley. http://onlinelibrary.wiley.com/doi/10.1002/0471667196.ess0985.pub2/abstract.

Berger, J. O., and L. R. Pericchi. 1996. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* 91: 109–122. https://doi.org/10.2307/2291387.

Berger, J. O., and R. L. Wolpert. 1988. *The Likelihood Principle: A Review, Generalizations, and Statistical Implications*. Hayward, CA: Institute of Mathematical Statistics.

Berliner, L. M., J. A. Royle, C. K. Wikle, and R. F. Milliff. 1999. Bayesian methods in atmospheric sciences. In Vol. 6 of *Bayesian Statistics: Proceedings of the Sixth Valencia International Meeting*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 83–100. Oxford: Oxford University Press.

Bernardo, J. M., and A. F. M. Smith. 2000. *Bayesian Theory*. Chichester, UK: Wiley.

Berry, D. A., and D. K. Stangl, ed. 1996. *Bayesian Biostatistics*. New York: Dekker.

Besag, J., and D. Higdon. 1999. Bayesian analysis for agricultural field experiments. *Journal of the Royal Statistical Society, Series B* 61: 691–746. https://doi.org/10.1111/1467-9868.00201.

Brooks, S. P., A. Gelman, G. L. Jones, and X.-L. Meng, ed. 2011. *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: Chapman and Hall/CRC.

Cameron, A. C., and P. K. Trivedi. 2022. *Microeconometrics Using Stata*. 2nd ed. College Station, TX: Stata Press.

Carlin, B. P., J. B. Kadane, and A. E. Gelfand. 1998. Approaches for optimal sequential decision analysis in clinical trials. *Biometrics* 54: 964–975. https://doi.org/10.2307/2533849.

Carlin, B. P., and T. A. Louis. 2009. *Bayesian Methods for Data Analysis*. 3rd ed. Boca Raton, FL: Chapman and Hall/CRC.

Chaloner, K., and I. Verdinelli. 1995. Bayesian experimental design: A review. *Statistical Science* 10: 273–304. https://doi.org/10.1214/ss/1177009939.

Chen, M.-H., and Q.-M. Shao. 1999. Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics* 8: 69–92. https://doi.org/10.2307/1390921.

Chen, M.-H., Q.-M. Shao, and J. G. Ibrahim. 2000. *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

Chernozhukov, V., and H. Hong. 2003. An MCMC approach to classical estimation. *Journal of Econometrics* 115: 293–346. https://doi.org/10.1016/S0304-4076(03)00100-3.

Chib, S., and E. Greenberg. 1995. Understanding the Metropolis–Hastings algorithm. *American Statistician* 49: 327–335. https://doi.org/10.1080/00031305.1995.10476177.

Cowles, M. K., and B. P. Carlin. 1996. Markov chain Monte Carlo convergence diagnostic: A comparative review. *Journal of the American Statistical Association* 91: 883–904. https://doi.org/10.2307/2291683.

DeGroot, M. H., S. E. Fienberg, and J. B. Kadane, ed. 1986. *Statistics and the Law*. New York: Wiley.

Dey, D. D., P. Müller, and D. Sinha, ed. 1998. *Practical Nonparametric and Semiparametric Bayesian Statistics*. New York: Springer.

Dey, D. K., S. K. Ghosh, and B. K. Mallick. 2000. *Generalized Linear Models: A Bayesian Perspective*. New York: Dekker.

Eberly, L. E., and G. Casella. 2003. Estimating Bayesian credible intervals. *Journal of Statistical Planning and Inference* 112: 115–132. https://doi.org/10.1016/S0378-3758(02)00327-0.

Gamerman, D., and H. F. Lopes. 2006. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC.

Gelfand, A. E., S. E. Hills, A. Racine-Poon, and A. F. M. Smith. 1990. Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* 85: 972–985. https://doi.org/10.2307/2289594.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2014. *Bayesian Data Analysis*. 3rd ed. Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., W. R. Gilks, and G. O. Roberts. 1997. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability* 7: 110–120. https://doi.org/10.1214/aoap/1034625254.

Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* 7: 457–472. https://doi.org/10.1214/ss/1177011136.

Geman, S., and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721–741. https://doi.org/10.1109/TPAMI.1984.4767596.

Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Vol. 4 of *Bayesian Statistics: Proceedings of the Fourth Valencia International Meeting, April 15–20, 1991*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 169–193. Oxford: Clarendon Press.

——. 1999. Using simulation methods for Bayesian econometric models: Inference, development, and communication. *Econometric Reviews* 18: 1–73. https://doi.org/10.1080/07474939908800428.

Giordani, P., and R. J. Kohn. 2010. Adaptive independent Metropolis–Hastings by fast estimation of mixtures of normals. *Journal of Computational and Graphical Statistics* 19: 243–259. https://doi.org/10.1198/jcgs.2009.07174.

Godsill, S. J., and P. J. W. Rayner. 1998. *Digital Audio Restoration*. Berlin: Springer.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith. 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings on Radar and Signal Processing* 140: 107–113. https://doi.org/10.1049/ip-f-2.1993.0015.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82: 711–732. https://doi.org/10.1093/biomet/82.4.711.

Greenland, S. 1998. Probability logic and probabilistic induction. *Epidemiology* 9: 322–332.

Haario, H., E. Saksman, and J. Tamminen. 2001. An adaptive Metropolis algorithm. *Bernoulli* 7: 223–242. https://doi.org/10.2307/3318737.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57: 97–109. https://doi.org/10.2307/2334940.

Heidelberger, P., and P. D. Welch. 1983. Simulation run length control in the presence of an initial transient. *Operations Research* 31: 1109–1144. https://doi.org/10.1287/opre.31.6.1109.

Hobert, J. P. 2000. Hierarchical models: A current computational perspective. *Journal of the American Statistical Association* 95: 1312–1316. https://doi.org/10.1080/01621459.2000.10474338.

Hoff, P. D. 2009. *A First Course in Bayesian Statistical Methods*. New York: Springer.

Iversen, E., Jr., G. Parmigiani, and D. A. Berry. 1999. Validating Bayesian Prediction Models: a Case Study in Genetic Susceptibility to Breast Cancer. In *Case Studies in Bayesian Statistics*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, vol. IV, 321–338. New York: Springer.

Jeffreys, H. 1935. Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society* 31: 203–222. https://doi.org/10.1017/S030500410001330X.

——. 1961. *Theory of Probability*. 3rd ed. Oxford: Oxford University Press.

Johnson, V. E. 1997. An alternative to traditional GPA for evaluating student performance. *Statistical Science* 12: 251–269. https://doi.org/10.1214/ss/1030037959.

Kass, R. E., and A. E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90: 773–795. https://doi.org/10.1080/01621459.1995.10476572.

Kim, S., N. Shephard, and S. Chib. 1998. Stochastic volatility: Likelihood inference and comparison with ARCH models. *Reviews of Economic Studies* 65: 361–393. https://doi.org/10.1111/1467-937X.00050.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* 21: 1087–1092. https://doi.org/10.1063/1.1699114.

Metropolis, N., and S. Ulam. 1949. The Monte Carlo method. *Journal of the American Statistical Association* 44: 335–341. https://doi.org/10.1080/01621459.1949.10483310.

Müller, P., and B. Vidakovic, ed. 1999. *Bayesian Inference in Wavelet-Based Models*. New York: Springer.

Neal, R. M. 1996. *Bayesian Learning for Neural Networks*. New York: Springer.

——. 1999. Regression and classification using gaussian process priors. In Vol. 6 of *Bayesian Statistics: Proceedings of the Sixth Valencia International Meeting*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 475–501. Oxford: Oxford University Press.

Parent, E., P. Hubert, B. Bobee, and J. Miquel. 1998. *Statistical and Bayesian Methods in Hydrological Sciences*. Paris: UNESCO Press.

Pearl, J. 1998. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann.

Poirier, D. J. 1995. *Intermediate Statistics and Econometrics: A Comparative Approach*. Cambridge, MA: MIT Press.

Pole, A., M. West, and J. Harrison. 1994. *Applied Bayesian Forecasting and Time Series Analysis*. Boca Raton, FL: Chapman and Hall.

Pollard, W. E. 1986. *Bayesian Statistics for Evaluation Research: An Introduction*. Newbury Park, CA: Sage.

Propp, J. G., and D. B. Wilson. 1996. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* 9: 223–252. https://doi.org/10.1002/(SICI)1098-2418(199608/09)9:1/2⟨223::AID-RSA14⟩3.0.CO;2-O.

Rabe-Hesketh, S., and A. Skrondal. 2022. *Multilevel and Longitudinal Modeling Using Stata*. 4th ed. College Station, TX: Stata Press.

Raftery, A. E., and S. M. Lewis. 1992. How many iterations in the Gibbs sampler? In Vol. 4 of *Bayesian Statistics: Proceedings of the Fourth Valencia International Meeting, April 15–20, 1991*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 763–773. Oxford: Clarendon Press.

Rios Insua, D. 1990. *Sensitivity Analysis in Multi-Objective Decision Making*. New York: Springer.

Robert, C. P., and G. Casella. 2004. *Monte Carlo Statistical Methods*. 2nd ed. New York: Springer.

Roberts, G. O., and J. S. Rosenthal. 2007. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability* 44: 458–475. https://doi.org/10.1239/jap/1183667414.

——. 2009. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* 18: 349–367. https://doi.org/10.1198/jcgs.2009.06134.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464. https://doi.org/10.1214/aos/1176344136.

Tanner, M. A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. 3rd ed. New York: Springer.

Tanner, M. A., and W. H. Wong. 1987. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* 82: 528–550. https://doi.org/10.2307/2289457.

Thompson, J. 2014. *Bayesian Analysis with Stata*. College Station, TX: Stata Press.

Thompson, S. K. 2012. *Sampling*. 3rd ed. Hoboken, NJ: Wiley.

Tierney, L. 1994. Markov chains for exploring posterior distributions. *Annals of Statistics* 22: 1701–1728. https://doi.org/10.1214/aos/1176325750.

von Neumann, J. 1951. Various techniques used in connection with random digits. Monte Carlo methods. *Journal of Research of the National Bureau of Standards* 12: 36–38.

West, M., and J. Harrison. 1997. *Bayesian Forecasting and Dynamic Models*. 2nd ed. New York: Springer.

Wolpert, R. L., and K. Ickstadt. 1998. Poisson/gamma random field models for spatial statistics. *Biometrika* 85: 251–267. https://doi.org/10.1093/biomet/85.2.251.

Yu, B., and P. Mykland. 1998. Looking at Markov samplers through cusum path plots: A simple diagnostic idea. *Statistics and Computing* 8: 275–286. https://doi.org/10.1023/A:1008917713940.

Zellner, A. 1997. *Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers*. Northampton, MA: Edward Elgar.

Zellner, A., and C. Min. 1995. Gibbs sampler convergence criteria. *Journal of the American Statistical Association* 90: 921–927. https://doi.org/10.1080/01621459.1995.10476591.

Also see

[BAYES] **Bayesian commands** — Introduction to commands for Bayesian analysis

[BAYES] **Glossary**

[BMA] **Intro** — Introduction to Bayesian model averaging

Stata, Stata Press, and Mata are registered trademarks of StataCorp LLC. Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations. Other brand and product names are registered trademarks or trademarks of their respective companies. Copyright © 1985–2023 StataCorp LLC, College Station, TX, USA. All rights reserved.