Glossary

a posteriori. In the context of Bayesian analysis, we use a posteriori to mean "after the sample is observed". For example, a posteriori information is any information obtained after the data sample is observed. See *posterior distribution, posterior*.

a priori. In the context of Bayesian analysis, we use a priori to mean "before the sample is observed". For example, a priori information is any information obtained before the data sample is observed. In a Bayesian model, a priori information about model parameters is specified by prior distributions.

acceptance rate. In the context of the MH algorithm, acceptance rate is the fraction of the proposed samples that is accepted. The optimal acceptance rate depends on the properties of the target distribution and is not known in general. If the target distribution is normal, however, the optimal acceptance rate is known to be 0.44 for univariate distributions and 0.234 for multivariate distributions.

adaptation. In the context of the MH algorithm, adaptation refers to the process of tuning or adapting the proposal distribution to optimize the MCMC sampling. Typically, adaptation is performed periodically during the MCMC sampling. The `bayesmh` command performs adaptation every # of iterations as specified in option `adaptation(every(#))` for a maximum of `adaptation(maxiter())` iterations. In a continuous-adaptation regimes, the adaptation lasts during the entire process of the MCMC sampling. See [BAYES] **bayesmh**.

adaptation period. Adaptation period includes all MH adaptive iterations. It equals the length of the adaptation interval, as specified by `adaptation(every())`, times the maximum number of adaptations, `adaptation(maxiter())`.

adaptive iteration. In the adaptive MH algorithm, adaptive iterations are iterations during which adaptation is performed.

Akaike information criterion, AIC. Akaike information criterion (AIC) is an information-based model-selection criterion. It is given by the formula $-2 \times \log$ likelihood $+ 2k$, where k is the number of parameters. AIC favors simpler models by penalizing for the number of model parameters. It does not, however, account for the sample size. As a result, the AIC penalization diminishes as the sample size increases, as does its ability to guard against overparameterization.

batch means. Batch means are means obtained from batches of sample values of equal size. Batch means provide an alternative method for estimating MCMC standard errors (MCSE). The batch size is usually chosen to minimize the correlation between different batches of means.

Bayes factor. Bayes factor is given by the ratio of the marginal likelihoods of two models, M_1 and M_2. It is a widely used criterion for Bayesian model comparison. Bayes factor is used in calculating the posterior odds ratio of model M_1 versus M_2,

$$\frac{P(M_1|\mathbf{y})}{P(M_2|\mathbf{y})} = \frac{P(\mathbf{y}|M_1)}{P(\mathbf{y}|M_2)}\frac{P(M_1)}{P(M_2)}$$

where $P(M_i|\mathbf{y})$ is a posterior probability of model M_i, and $P(M_i)$ is a prior probability of model M_i. When the two models are equally likely, that is, when $P(M_1) = P(M_2)$, the Bayes factor equals the posterior odds ratio of the two models.

Bayes's theorem. The Bayes's theorem is a formal method for relating conditional probability statements. For two (random) events X and Y, the Bayes's theorem states that

$$P(X|Y) \propto P(Y|X)P(X)$$

that is, the probability of X conditional on Y is proportional to the probability of X and the probability of Y conditional on X. In Bayesian analysis, the Bayes's theorem is used for combining prior information about model parameters and evidence from the observed data to form the posterior distribution.

Bayesian analysis. Bayesian analysis is a statistical methodology that considers model parameters to be random quantities and estimates their posterior distribution by combining prior knowledge about parameters with the evidence from the observed data sample. Prior knowledge about parameters is described by prior distributions and evidence from the observed data is incorporated through a likelihood model. Using the Bayes's theorem, the prior distribution and the likelihood model are combined to form the posterior distribution of model parameters. The posterior distribution is then used for parameter inference, hypothesis testing, and prediction.

Bayesian estimation. Bayesian estimation consists of fitting Bayesian models and estimating their parameters based on the resulting posterior distribution. Bayesian estimation in Stata can be done using the convenient `bayes` prefix or the more general `bayesmh` command. See [BAYES] **bayesian estimation** for details.

Bayesian estimation results. Estimation results obtained after the `bayes` prefix or the `bayesmh` command.

Bayesian hypothesis testing. Bayesian hypothesis testing computes probabilities of hypotheses conditional on the observed data. In contrast to the frequentist hypothesis testing, the Bayesian hypothesis testing computes the actual probability of a hypothesis H by using the Bayes's theorem,

$$P(H|\mathbf{y}) \propto P(\mathbf{y}|H)P(H)$$

where $\mathbf{y}$ is the observed data, $P(\mathbf{y}|H)$ is the marginal likelihood of $\mathbf{y}$ given H, and $P(H)$ is the prior probability of H. Two different hypotheses, H_1 and H_2, can be compared by simply comparing $P(H_1|\mathbf{y})$ to $P(H_2|\mathbf{y})$.

Bayesian information criterion, BIC. The Bayesian information criterion (BIC), also known as Schwarz criterion, is an information based criterion used for model selection in classical statistics. It is given by the formula $-0.5 \times \log$ likelihood $+ k \times \ln n$, where k is the number of parameters and n is the sample size. BIC favors simpler, in terms of complexity, models and it is more conservative than AIC.

blocking. In the context of the MH algorithm, blocking refers to the process of separating model parameters into different subsets or blocks to be sampled independently of each other. MH algorithm generates proposals and applies the acceptance–rejection rule sequentially for each block. It is recommended that correlated parameters are kept in one block. Separating less-correlated or independent model parameters in different blocks may improve the mixing of the MH algorithm.

burn-in period. The burn-in period is the number of iterations it takes for an MCMC sequence to reach stationarity.

central posterior interval. See *equal-tailed credible interval*.

conditional conjugacy. See *semiconjugate prior*.

conjugate prior. A prior distribution is conjugate for a family of likelihood distributions if the prior and posterior distributions belong to the same family of distributions. For example, the gamma distribution is a conjugate prior for the Poisson likelihood. Conjugacy may provide an efficient way of sampling from posterior distributions and is used in Gibbs sampling.

continuous parameters. Continuous parameters are parameters with continuous prior distributions.

credible interval. In Bayesian analysis, the credible interval of a scalar model parameter is an interval from the domain of the marginal posterior distribution of that parameter. Two types of credible intervals are typically used in practice: equal-tailed credible intervals and HPD credible intervals.

credible level. The credible level is a probability level between 0% and 100% used for calculating credible intervals in Bayesian analysis. For example, a 95% credible interval for a scalar parameter is an interval the parameter belongs to with the probability of 95%.

cusum plot, CUSUM plot. The cusum (CUSUM) plot of an MCMC sample is a plot of cumulative sums of the differences between sample values and their overall mean against the iteration number. Cusum plots are useful graphical summaries for detecting early drifts in MCMC samples.

deviance information criterion, DIC. The deviance information criterion (DIC) is an information based criterion used for Bayesian model selection. It is an analog of AIC and is given by the formula $D(\bar{\theta}) + 2 \times p_D$, where $D(\bar{\theta})$ is the deviance at the sample mean and p_D is the effective complexity, a quantity equivalent to the number of parameters in the model. Models with smaller DIC are preferred.

diminishing adaptation. Diminishing adaptation of the adaptive algorithm is the type of adaptation in which the amount of adaptation decreases with the size of the MCMC chain.

discrete parameters. Discrete parameters are parameters with discrete prior distributions.

effective sample size, ESS. Effective sample size (ESS) is the MCMC sample size T adjusted for the autocorrelation in the sample. It represents the number of independent observations in an MCMC sample. ESS is used instead of T in calculating MCSE. Small ESS relative to T indicates high autocorrelation and consequently poor mixing of the chain.

efficiency. In the context of MCMC, efficiency is a term used for assessing the mixing quality of an MCMC procedure. Efficient MCMC algorithms are able to explore posterior domains in less time (using fewer iterations). Efficiency is typically quantified by the sample autocorrelation and effective sample size. An MCMC procedure that generates samples with low autocorrelation and consequently high ESS is more efficient.

equal-tailed credible interval. An equal-tailed credible interval is a credible interval defined in such a way that both tails of the marginal posterior distribution have the same probability. A $\{100 \times (1 - \alpha)\}\%$ equal-tailed credible interval is defined by the $\alpha/2$th and $\{(1 - \alpha)/2\}$th quantiles of the marginal posterior distribution.

feasible initial value. An initial-value vector is feasible if it corresponds to a state with a positive posterior probability.

fixed effects. See *fixed-effects parameters*.

fixed-effects parameters. In the Bayesian context, the term "fixed effects" or "fixed-effects parameters" is a misnomer, because all model parameters are inherently random. We use this term in the context of Bayesian multilevel models to refer to regression model parameters and to distinguish them from the random-effects parameters. You can think of fixed-effects parameters as parameters modeling population averaged or marginal relationship of the response and the variables of interest.

frequentist analysis. Frequentist analysis is a form of statistical analysis where model parameters are considered to be unknown but fixed constants and the observed data are viewed as a repeatable random sample. Inference is based on the sampling distribution of the data.

full conditionals. A full conditional is the probability distribution of a random variate conditioned on all other random variates in a joint probability model. Full conditional distributions are used in Gibbs sampling.

full Gibbs sampling. See *Gibbs sampling, Gibbs sampler*.

Gibbs sampling, Gibbs sampler. Gibbs sampling is an MCMC method, according to which each random variable from a joint probability model is sampled according to its full conditional distribution.

highest posterior density credible interval, HPD credible interval. The highest posterior density (HPD) credible interval is a type of a credible interval with the highest marginal posterior density. An HPD interval has the shortest width among all other credible intervals. For some multimodal marginal distributions, HPD may not exists. See *highest posterior density region, HPD region*.

highest posterior density region, HPD region. The highest posterior density (HPD) region for model parameters has the highest marginal posterior probability among all domain regions. Unlike an HPD credible interval, an HPD region always exist.

hybrid MH sampling, hybrid MH sampler. A hybrid MH sampler is an MCMC method in which some blocks of parameters are updated using the MH algorithms and other blocks are updated using Gibbs sampling.

hyperparameter. In Bayesian analysis, hyperparameter is a parameter of a prior distribution, in contrast to a model parameter.

hyperprior. In Bayesian analysis, hyperprior is a prior distribution of hyperparameters. See *hyperparameter*.

improper prior. A prior is said to be improper if it does not integrate to a finite number. Uniform distributions over unbounded intervals are improper. Improper priors may still yield proper posterior distributions. When using improper priors, however, one has to make sure that the resulting posterior distribution is proper for Bayesian inference to be invalid.

independent a posteriori. Parameters are considered independent a posteriori if their marginal posterior distributions are independent; that is, their joint posterior distribution is the product of their individual marginal posterior distributions.

independent a priori. Parameters are considered independent a priori if their prior distributions are independent; that is, their joint prior distribution is the product of their individual marginal prior distributions.

informative prior. An informative prior is a prior distribution that has substantial influence on the posterior distribution.

interval hypothesis testing. Interval hypothesis testing performs interval hypothesis tests for model parameters and functions of model parameters.

interval test. In Bayesian analysis, an interval test applied to a scalar model parameter calculates the marginal posterior probability for the parameter to belong to the specified interval.

Jeffreys prior. The Jeffreys prior of a vector of model parameters $\boldsymbol{\theta}$ is proportional to the square root of the determinant of its Fisher information matrix $I(\boldsymbol{\theta})$. Jeffreys priors are locally uniform and, by definition, agree with the likelihood function. Jeffreys priors are considered noninformative priors that have minimal impact on the posterior distribution.

marginal distribution. In Bayesian context, a distribution of the data after integrating out parameters from the joint distribution of the parameters and the data.

marginal likelihood. In the context of Bayesian model comparison, a marginalized over model parameters $\boldsymbol{\theta}$ likelihood of data $\mathbf{y}$ for a given model M, $P(\mathbf{y}|M) = m(\mathbf{y}) = \int P(\mathbf{y}|\boldsymbol{\theta}, M)P(\boldsymbol{\theta}|M)d\boldsymbol{\theta}$. Also see *Bayes factor*.

marginal posterior distribution. In Bayesian context, a marginal posterior distribution is a distribution resulting from integrating out all but one parameter from the joint posterior distribution.

Markov chain. Markov chain is a random process that generates sequences of random vectors (or states) and satisfies the Markov property: the next state depends only on the current state and not on any of the previous states. MCMC is the most common methodology for simulating Markov chains.

matrix model parameter. A matrix model parameter is any model parameter that is a matrix. Matrix elements, however, are viewed as scalar model parameters.

Matrix model parameters are defined and referred to within the `bayesmh` command as {*param*,`matrix`} or {*eqname*:*param*,`matrix`} with the equation name *eqname*. For example, {`Sigma, matrix`} and {`Scale:Omega, matrix`} are matrix model parameters. Individual matrix elements cannot be referred to within the `bayesmh` command, but they can be referred within postestimation commands accepting parameters. For example, to refer to the individual elements of the defined above, say, 2×2 matrices, use {`Sigma_1_1`}, {`Sigma_2_1`}, {`Sigma_1_2`}, {`Sigma_2_2`} and {`Scale:Omega_1_1`}, {`Scale:Omega_2_1`}, {`Scale:Omega_1_2`}, {`Scale:Omega_2_2`}, respectively. See [BAYES] **bayesmh**.

matrix parameter. See *matrix model parameter*.

MCMC, Markov chain Monte Carlo. MCMC is a class of simulation-based methods for generating samples from probability distributions. Any MCMC algorithm simulates a Markov chain with a target distribution as its stationary or equilibrium distribution. The precision of MCMC algorithms increases with the number of iterations. The lack of a stopping rule and convergence rule, however, makes it difficult to determine for how long to run MCMC. The time needed to converge to the target distribution within a prespecified error is referred to as mixing time. Better MCMC algorithms have faster mixing times. Some of the popular MCMC algorithms are random-walk Metropolis, Metropolis–Hastings, and Gibbs sampling.

MCMC sample. An MCMC sample is obtained from MCMC sampling. An MCMC sample approximates a target distribution and is used for summarizing this distribution.

MCMC sample size. MCMC sample size is the size of the MCMC sample. It is specified in `bayesmh`'s option `mcmcsize()`; see [BAYES] **bayesmh**.

MCMC sampling, MCMC sampler. MCMC sampling is an MCMC algorithm that generates samples from a target probability distribution.

MCMC standard error, MCSE MCSE is the standard error of the posterior mean estimate. It is defined as the standard deviation divided by the square root of ESS. MCSEs are analogs of standard errors in frequentist statistics and measure the accuracy of the simulated MCMC sample.

Metropolis–Hastings (MH) sampling, MH sampler. A Metropolis–Hastings (MH) sampler is an MCMC method for simulating probability distributions. According to this method, at each step of the Markov chain, a new proposal state is generated from the current state according to a prespecified proposal distribution. Based on the current and new state, an acceptance probability is calculated and then used to accept or reject the proposed state. Important characteristics of MH sampling is the acceptance rate and mixing time. The MH algorithm is very general and can be applied to an arbitrary target distribution. However, its efficiency is limited, in terms of mixing time, and decreases as the dimension of the target distribution increases. Gibbs sampling, when available, can provide much more efficient sampling than MH sampling.

mixing of Markov chain. Mixing refers to the rate at which a Markov chain traverses the parameter space. It is a property of the Markov chain that is different from convergence. Poor mixing indicates a slow rate at which the chain explores the stationary distribution and will require more iterations to provide inference at a given precision. Poor (slow) mixing is typically a result of high correlation between model parameters or of weakly-defined model specifications.

model hypothesis testing. Model hypothesis testing tests hypotheses about models by computing model posterior probabilities.

model parameter. A model parameter refers to any (random) parameter in a Bayesian model. Model parameters can be scalars or matrices. Examples of model parameters as defined in `bayesmh` are {`mu`}, {`scale:s`}, {`Sigma,matrix`}, and {`Scale:Omega,matrix`}. See [BAYES] **bayesmh** and,

specifically, *Declaring model parameters* and *Referring to model parameters* in that entry. Also see *Different ways of specifying model parameters* in [BAYES] **bayesian postestimation**.

model posterior probability. Model posterior probability is probability of a model M computed conditional on the observed data $\mathbf{y}$,

$$P(M|\mathbf{y}) = P(M)P(\mathbf{y}|M) = P(M)m(\mathbf{y})$$

where $P(M)$ is the prior probability of a model M and $m(\mathbf{y})$ is the marginal likelihood under model M.

noninformative prior. A noninformative prior is a prior with negligible influence on the posterior distribution. See, for example, *Jeffreys prior*.

objective prior. See *noninformative prior*.

one-at-a-time MCMC sampling. A one-at-a-time MCMC sample is an MCMC sampling procedure in which random variables are sampled individually, one at a time. For example, in Gibbs sampling, individual variates are sampled one at a time, conditionally on the most recent values of the rest of the variates.

posterior distribution, posterior. A posterior distribution is a probability distribution of model parameters conditional on observed data. The posterior distribution is determined by the likelihood of the parameters and their prior distribution. For a parameter vector $\boldsymbol{\theta}$ and data $\mathbf{y}$, the posterior distribution is given by

$$P(\boldsymbol{\theta}|\mathbf{y}) = \frac{P(\boldsymbol{\theta})P(\mathbf{y}|\boldsymbol{\theta})}{P(\mathbf{y})}$$

where $P(\boldsymbol{\theta})$ is the prior distribution, $P(\mathbf{y}|\boldsymbol{\theta})$ is the model likelihood, and $P(\mathbf{y})$ is the marginal distribution for $\mathbf{y}$. Bayesian inference is based on a posterior distribution.

posterior independence. See *independent a posteriori*.

posterior interval. See *credible interval*.

posterior odds. Posterior odds for $\boldsymbol{\theta}_1$ compared with $\boldsymbol{\theta}_2$ is the ratio of posterior density evaluated at $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ under a given model,

$$\frac{p(\boldsymbol{\theta}_1|\mathbf{y})}{p(\boldsymbol{\theta}_2|\mathbf{y})} = \frac{p(\boldsymbol{\theta}_1)}{p(\boldsymbol{\theta}_2)}\frac{p(\mathbf{y}|\boldsymbol{\theta}_1)}{p(\mathbf{y}|\boldsymbol{\theta}_2)}$$

In other words, posterior odds are prior odds times the likelihood ratio.

posterior predictive distribution. A posterior predictive distribution is a distribution of unobserved (future) data conditional on the currently observed data. Posterior predictive distribution is derived by marginalizing the likelihood function with respect to the posterior distribution of model parameters.

prior distribution, prior. In Bayesian statistics, prior distributions are probability distributions of model parameters formed based on some a priori knowledge about parameters. Prior distributions are independent of the observed data.

prior independence. See *independent a priori*.

prior odds. Prior odds for $\boldsymbol{\theta}_1$ compared with $\boldsymbol{\theta}_2$ is the ratio of prior density evaluated at $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ under a given model, $p(\boldsymbol{\theta}_1)/p(\boldsymbol{\theta}_2)$. Also see *posterior odds*.

proposal distribution. In the context of the MH algorithm, a proposal distribution is used for defining the transition steps of the Markov chain. In the standard random-walk Metropolis algorithm, the proposal distribution is a multivariate normal distribution with zero mean and adaptable covariance matrix.

pseudoconvergence. A Markov chain may appear to converge when in fact it did not. We refer to this phenomenon as pseudoconvergence. Pseudoconvergence is typically caused by multimodality of the stationary distribution, in which case the chain may fail to traverse the weakly connected regions of the distribution space. A common way to detect pseudoconvergence is to run multiple chains using different starting values and to verify that all of the chain converge to the same target distribution.

random effects. See *random-effects parameters*.

random-effects linear form. A linear form representing a random-effects variable that can be used in substitutable expressions.

random-effects parameters. In the context of Bayesian multilevel models, random-effects parameters are parameters associated with a random-effects variable. Random-effects parameters are assumed to be conditionally independent across levels of the random-effects variable given all other model parameters. Often, random-effects parameters are assumed to be normally distributed with a zero mean and an unknown variance–covariance matrix.

random-effects variable. A variable identifying the group structure for the random effects at a specific level of hierarchy.

reference prior. See *noninformative prior*.

scalar model parameter. A scalar model parameter is any model parameter that is a scalar. For example, `{mean}` and `{hape:alpha}` are scalar parameters, as declared by the `bayesmh` command. Elements of matrix model parameters are viewed as scalar model parameters. For example, for a 2×2 matrix parameter `{Sigma,matrix}`, individual elements `{Sigma_1_1}`, `{Sigma_2_1}`, `{Sigma_1_2}`, and `{Sigma_2_2}` are scalar parameters. If a matrix parameter contains a label, the label should be included in the specification of individual elements as well. See [BAYES] **bayesmh**.

scalar parameter. See *scalar model parameter*.

semiconjugate prior. A prior distribution is semiconjugate for a family of likelihood distributions if the prior and (full) conditional posterior distributions belong to the same family of distributions. For semiconjugacy to hold, parameters must typically be independent a priori; that is, their joint prior distribution must be the product of the individual marginal prior distributions. For example, the normal prior distribution for a mean parameter of a normal data distribution with an unknown variance (which is assumed to be independent of the mean a priori) is a semiconjugate prior. Semiconjugacy may provide an efficient way of sampling from posterior distributions and is used in Gibbs sampling.

stationary distribution. Stationary distribution of a stochastic process is a joint distribution that does not change over time. In the context of MCMC, stationary distribution is the target probability distribution to which the Markov chain converges. When MCMC is used for simulating a Bayesian model, the stationary distribution is the target joint posterior distribution of model parameters.

subjective prior. See *informative prior*.

subsampling the chain. See *thinning*.

thinning. Thinning is a way of reducing autocorrelation in the MCMC sample by subsampling the MCMC chain every prespecified number of iterations determined by the thinning interval. For example, the thinning interval of 1 corresponds to using the entire MCMC sample; the thinning interval of 2 corresponds to using every other sample value; and the thinning interval of 3 corresponds to using values from iterations 1, 4, 7, 10, and so on. Thinning should be applied with caution when used to reduce autocorrelation because it may not always be the most appropriate way of improving the precision of estimates.

vague prior. See *noninformative prior*.

valid initial state. See *feasible initial value*.

vanishing adaptation. See *diminishing adaptation*.

Zellner's g-prior. Zellner's g-prior is a form of a weakly informative prior for the regression coefficients in a linear model. It accounts for the correlation between the predictor variables and controls the impact of the prior of the regression coefficients on the posterior with parameter g. For example, $g = 1$ means that prior weight is 50% and $g \to \infty$ means diffuse prior.