A Brief Introduction to Bayesian Model Selection and Validation

Mike May
Department of Evolution & Ecology
University of California, Davis
CoME, 2022

Outline

I. Model selection

What is the relative fit of the candidate models (hypotheses) to our data?

Bayesian methods for selecting among candidate models (hypotheses)

Outline

I. Model selection

What is the <u>relative</u> fit of the candidate models (hypotheses) to our data?

Bayesian methods for selecting among candidate models (hypotheses)

II. Model adequacy

What is the <u>absolute</u> fit of the candidate models (hypotheses) to our data?

Bayesian methods for assessing model adequacy of candidate models (hypotheses)

Outline

I. Model selection

What is the <u>relative</u> fit of the candidate models (hypotheses) to our data?

Bayesian methods for selecting among candidate models (hypotheses)

II. Model adequacy

What is the <u>absolute</u> fit of the candidate models (hypotheses) to our data?

Bayesian methods for assessing model adequacy of candidate models (hypotheses)

III. Model averaging

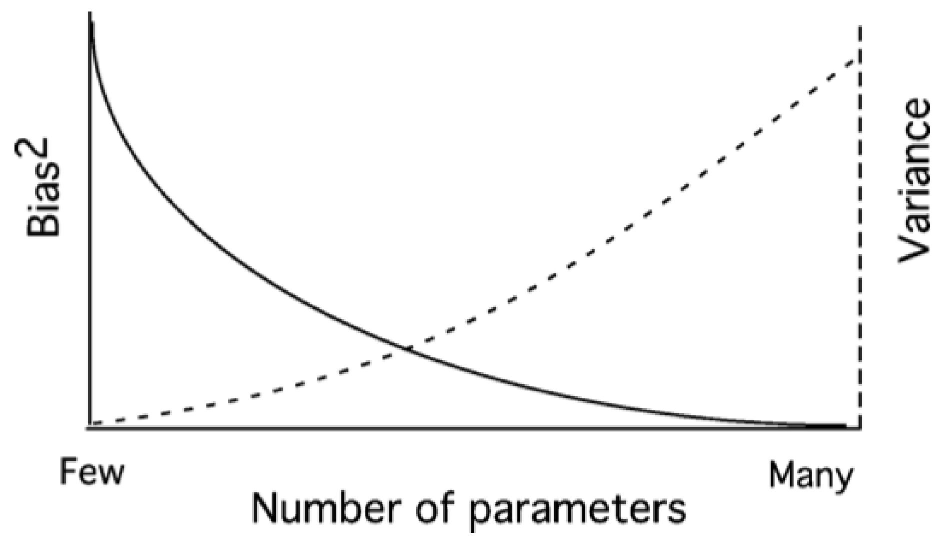
How do we accommodate uncertainty in the choice among candidate models?

Bayesian methods for averaging over candidate models (hypotheses)

Model Specification Issues

Model-based inference is based on the model

All of the parameters of our model (even 'nuisance' parameters) are critical

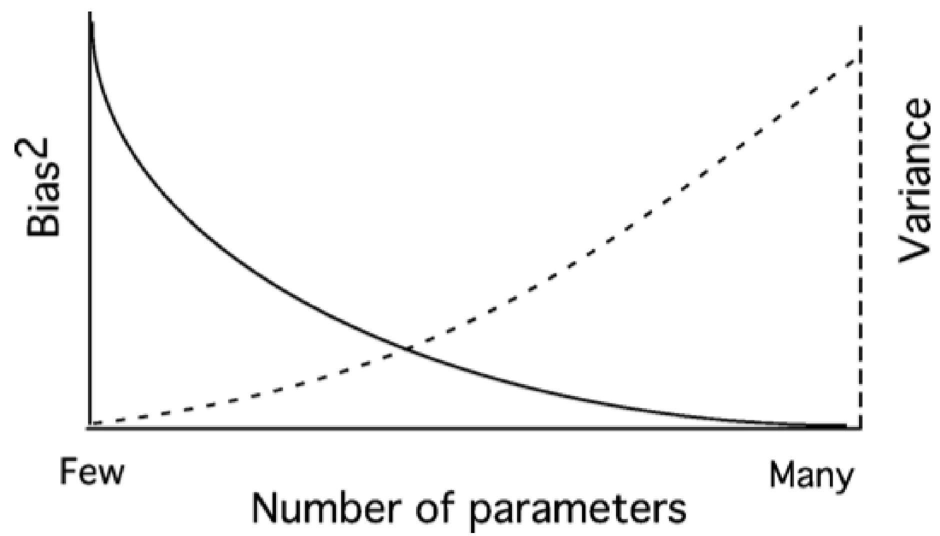


Model Specification Issues

Model-based inference is based on the model

All of the parameters of our model (even 'nuisance' parameters) are critical

- the model collectively describes the process that gave rise to our data

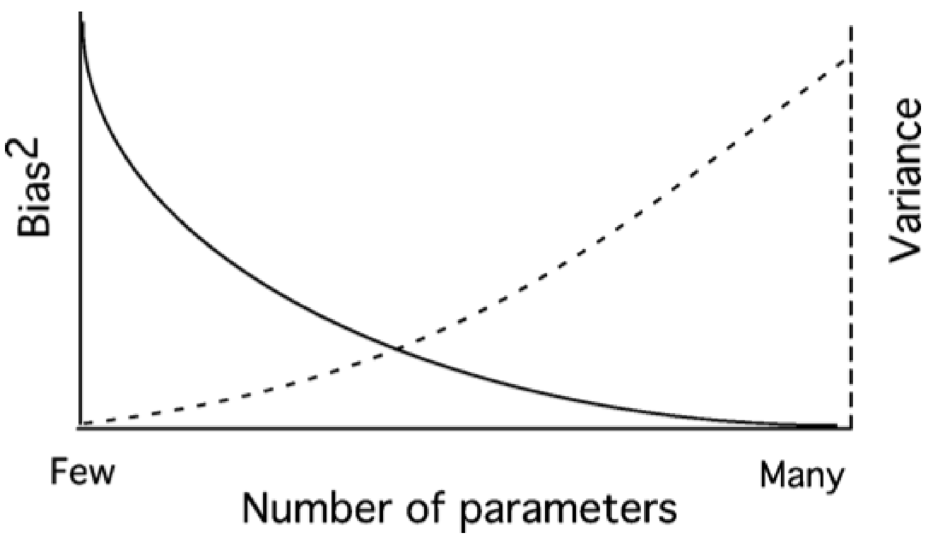


Model Specification Issues

Model-based inference is based on the model

All of the parameters of our model (even 'nuisance' parameters) are critical

- the model collectively describes the process that gave rise to our data
- an under-specified model will provide systematically biased parameter estimates

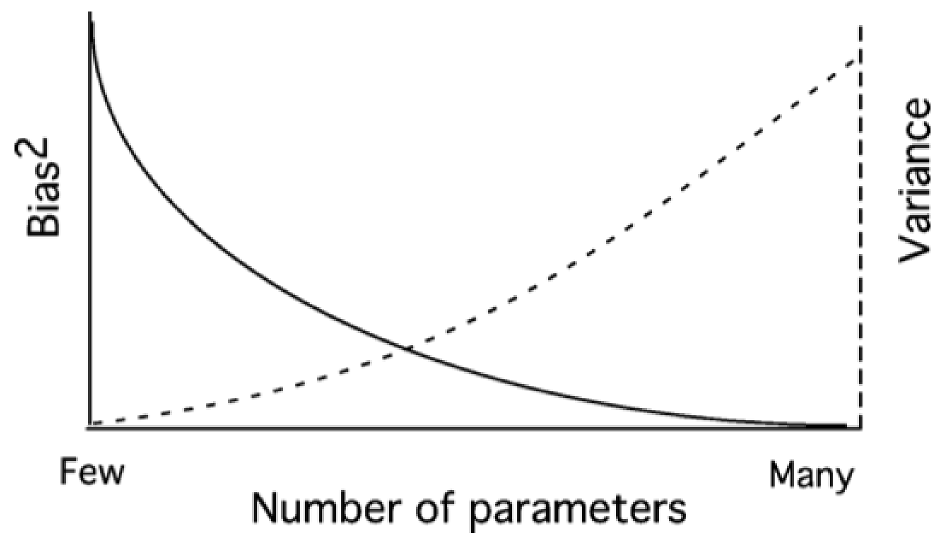


Model Specification Issues

Model-based inference is based on the model

All of the parameters of our model (even 'nuisance' parameters) are critical

- the model collectively describes the process that gave rise to our data
- an under-specified model will provide systematically biased parameter estimates
- an over-specified model will inflate the error variance of parameter estimates



Model Selection

What is the relative fit (rank) of the candidate models to the dataset?

Assessing the fit of our data to competing models is critical and useful:

Model Selection

What is the relative fit (rank) of the candidate models to the dataset?

Assessing the fit of our data to competing models is critical and useful:

- we need to identify which model provides the best fit to our data in order to obtain reliable parameter estimates

Model Selection

What is the relative fit (rank) of the candidate models to the dataset?

Assessing the fit of our data to competing models is critical and useful:

- we need to identify which model provides the best fit to our data in order to obtain reliable parameter estimates
- comparing the relative fit of two (or more) competing models is how we learn from our data

Model Selection

Data

sequence
alignment

Model

tree with
branch lengths

substitution
model

Model 1

Inference Method

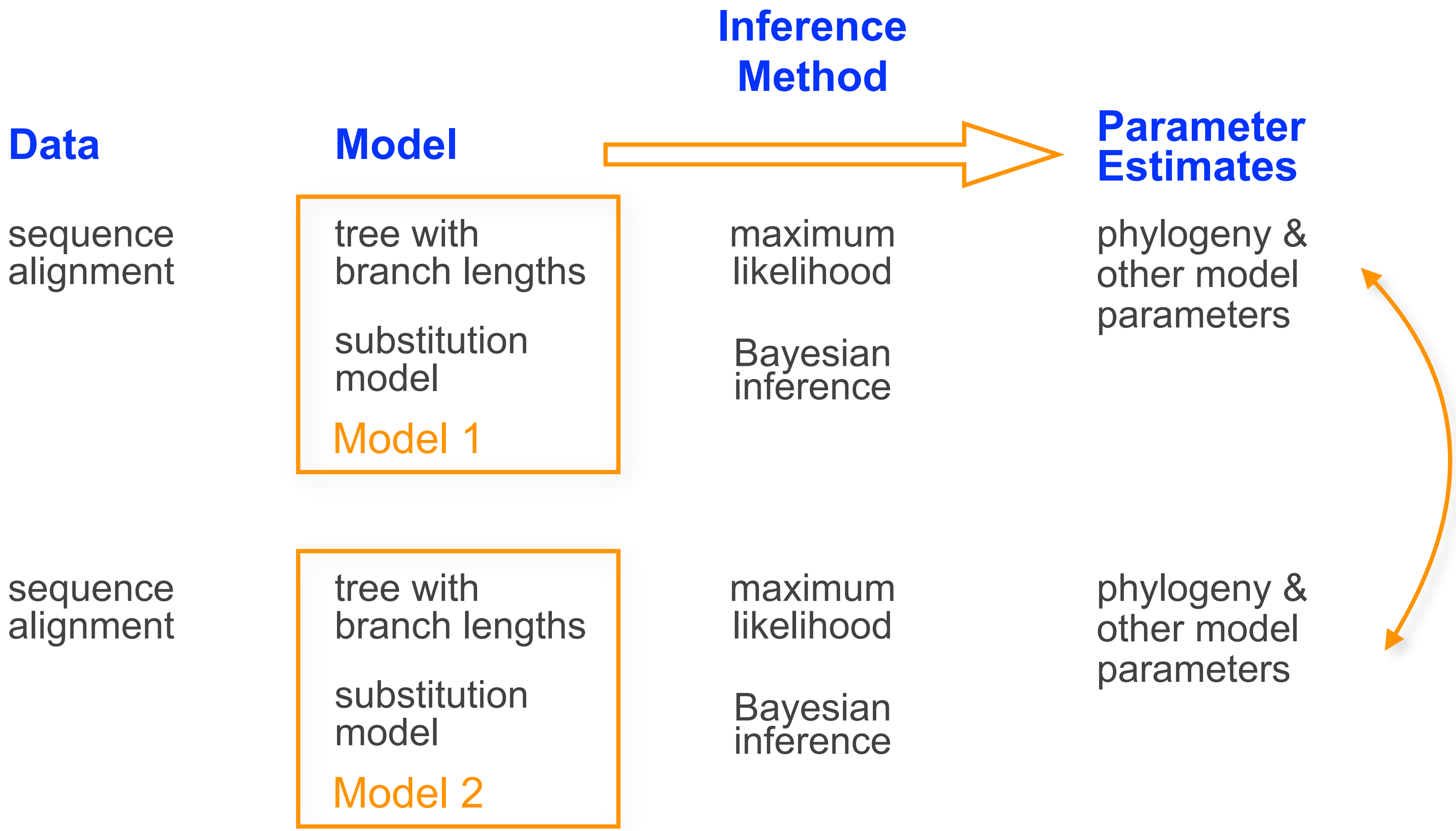
maximum
likelihood

Bayesian
inference

Parameter Estimates

phylogeny &
other model
parameters

Model Selection



Model Selection

Data	Model	Inference Method	Parameter Estimates
sequence alignment	tree with branch lengths substitution model Model 1	maximum likelihood Bayesian inference	phylogeny & other model parameters
sequence alignment	tree with branch lengths substitution model Model 2	maximum likelihood Bayesian inference	phylogeny & other model parameters

Model comparison is critical for obtaining unbiased parameter estimates

Model Selection

Data	Model	Inference Method	Parameter Estimates
sequence alignment	tree with branch lengths substitution model Model 1	maximum likelihood Bayesian inference	phylogeny & other model parameters
sequence alignment	tree with branch lengths substitution model Model 2	maximum likelihood Bayesian inference	phylogeny & other model parameters

Model comparison is critical for obtaining unbiased parameter estimates

Model comparison is the means by which we test hypotheses about our data

Model Selection: Bayesian Methods

Bayes factors

The marginal likelihood is the weighted sum over the possible discrete parameter values:

$$\overbrace{P(\theta_i \mid \mathbf{X})}^{\text{posterior probability}} = \frac{\overbrace{P(\mathbf{X} \mid \theta_i)}^{\text{likelihood}}\overbrace{P(\theta_i)}^{\text{prior}}}{\underbrace{\sum_{j=1}^{N} P(\mathbf{X} \mid \theta_j) P(\theta_j)}_{\text{marginal likelihood}}}$$

Model Selection: Bayesian Methods

Bayes factors

The marginal likelihood is the weighted sum over the possible discrete parameter values:

$$\overbrace{\Pr[\text{Biased} \mid \text{⚃}, \text{⚅}]}^{\text{posterior probability}} = \frac{\overbrace{\Pr[\text{⚃}, \text{⚅} \mid \text{Biased}]}^{\text{likelihood}} \times \overbrace{\Pr[\text{Biased}]}^{\text{prior probability}}}{\underbrace{\Pr[\text{⚃}, \text{⚅} \mid \text{Biased}] \times \Pr[\text{Biased}] + \Pr[\text{⚃}, \text{⚅} \mid \text{Fair}] \times \Pr[\text{Fair}]}_{\text{marginal likelihood}}}$$

Model Selection: Bayesian Methods

Bayes factors

The marginal likelihood is the weighted integral over the possible continuous parameter values:

$$\overbrace{P(\theta \mid \mathbf{X})}^{\text{posterior probability}} = \frac{\overbrace{P(\mathbf{X} \mid \theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{\underbrace{\int_\theta P(\mathbf{X} \mid \theta) P(\theta) d\theta}_{\text{marginal likelihood}}}$$

Model Selection: Bayesian Methods

Bayes factors

More generally, the marginal likelihood does not depend on any particular parameter values:

$$\overbrace{P(\theta \mid \mathbf{X})}^{\text{posterior probability}} = \frac{\overbrace{P(\mathbf{X} \mid \theta)}^{\text{likelihood}}\,\overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(\mathbf{X})}_{\text{marginal likelihood}}}$$

Model Selection: Bayesian Methods

Bayes factors

More generally, the marginal likelihood does not depend on any particular parameter values:

$$\overbrace{P(\theta \mid \mathbf{X})}^{\text{posterior probability}} = \frac{\overbrace{P(\mathbf{X} \mid \theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(\mathbf{X})}_{\text{marginal likelihood}}}$$

$$P(\mathbf{X}) = \int_\theta P(\mathbf{X} \mid \theta) P(\theta) d\theta$$

Model Selection: Bayesian Methods

Bayes factors

What it *does* depend on are the model and priors (the “full Bayesian model”), M_i

$$\overbrace{P(\theta \mid \mathbf{X}, M_i)}^{\text{posterior probability}} = \frac{\overbrace{P(\mathbf{X} \mid \theta, M_i)}^{\text{likelihood}}\overbrace{P(\theta \mid M_i)}^{\text{prior}}}{\underbrace{P(\mathbf{X} \mid M_i)}_{\text{marginal likelihood}}}$$

Model Selection: Bayesian Methods

Bayes factors

Bayesian model comparison is based on the *average* fit of the model to the data

- the marginal likelihood is the likelihood of the data under the model averaged over the joint prior probability of the model parameters

More complex models are automatically penalized by virtue of the corresponding priors

- model comparison is (intentionally) sensitive to the assumed priors

Model Selection: Bayesian Methods

Bayes factors

Bayes factors are computed based on the marginal likelihoods of competing models:

$$\mathrm{BF}_{01} = \frac{P(\mathbf{X} \mid M_0)}{P(\mathbf{X} \mid M_1)}$$

Model Selection: Bayesian Methods

Bayes factors

Bayes factors are computed based on the marginal likelihoods of competing models:

$$\mathrm{BF}_{01} = \frac{P(\mathbf{X} \mid M_0)}{P(\mathbf{X} \mid M_1)}$$

$$2\ln \mathrm{BF}_{01} = 2\left[\ln P(\mathbf{X} \mid M_0) - \ln P(\mathbf{X} \mid M_1)\right]$$

Model Selection: Bayesian Methods

Bayes factors

Bayes factors are computed based on the marginal likelihoods of competing models:

$$\mathrm{BF}_{01} = \frac{P(\mathbf{X} \mid M_0)}{P(\mathbf{X} \mid M_1)}$$

$$2\ln \mathrm{BF}_{01} = 2\left[\ln P(\mathbf{X} \mid M_0) - \ln P(\mathbf{X} \mid M_1)\right]$$

- models are preferred by virtue of their relative ability to predict the data

Model Selection: Bayesian Methods

Bayes factors

Bayes factors are computed based on the marginal likelihoods of competing models:

$$\mathrm{BF}_{01} = \frac{P(\mathbf{X} \mid M_0)}{P(\mathbf{X} \mid M_1)}$$

$$2\ln \mathrm{BF}_{01} = 2\left[\ln P(\mathbf{X} \mid M_0) - \ln P(\mathbf{X} \mid M_1)\right]$$

- models are preferred by virtue of their relative ability to predict the data
- $BF_{01} > 1$ supports model M_0, $BF_{01} < 1$ supports model M_1

BF_{01}	$2\ln BF_{01}$	Support for model M_0
1 to 3	0 to 2	Not worth more than a bare mention
3 to 20	2 to 6	Positive
20 to 150	6 to 10	Strong
> 150	> 10	Very strong

Model Selection: Bayesian Methods

Bayes factors

Nice properties of Bayes factors for model selection/hypothesis testing:

- can compare non-nested models

Model Selection: Bayesian Methods

Bayes factors

Nice properties of Bayes factors for model selection/hypothesis testing:

- can compare non-nested models
- accommodates uncertainty in parameter estimates

Model Selection: Bayesian Methods

Bayes factors

Nice properties of Bayes factors for model selection/hypothesis testing:

- can compare non-nested models
- accommodates uncertainty in parameter estimates

Nasty properties of Bayes factors for model selection/hypothesis testing:

- it is sensitive to the assumed priors

Model Selection: Bayesian Methods

Bayes factors

Nice properties of Bayes factors for model selection/hypothesis testing:

- can compare non-nested models
- accommodates uncertainty in parameter estimates

Nasty properties of Bayes factors for model selection/hypothesis testing:

- it is sensitive to the assumed priors
- Bayesian inference via MCMC avoids estimating marginal likelihoods

Model Selection: Bayesian Methods

Bayes factors

Nice properties of Bayes factors for model selection/hypothesis testing:

- can compare non-nested models
- accommodates uncertainty in parameter estimates

Nasty properties of Bayes factors for model selection/hypothesis testing:

- it is sensitive to the assumed priors
- Bayesian inference via MCMC avoids estimating marginal likelihoods
- this requires estimation of the marginal likelihoods of candidate models

Model Selection: Bayesian Methods

Bayes factors

Recall Bayes theorem:

$$P(\theta_1, \theta_2, \ldots, \theta_k \mid \mathbf{X}, M_i) = \frac{P(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_k, M_i) P(\theta_1, \theta_2, \ldots, \theta_k \mid M_i)}{P(\mathbf{X} \mid M_i)}$$

Model Selection: Bayesian Methods

Bayes factors

Recall Bayes theorem:

$$P(\theta_1, \theta_2, \ldots, \theta_k \mid \mathbf{X}, M_i) = \frac{P(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_k, M_i) P(\theta_1, \theta_2, \ldots, \theta_k \mid M_i)}{P(\mathbf{X} \mid M_i)}$$

The *marginal likelihood* is the likelihood of the data averaged over the joint prior distribution of *all* model parameters:

$$P(\mathbf{X} \mid M_i) = \int_{\theta_1} \int_{\theta_2} \ldots \int_{\theta_k} P(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_k, M_i) P(\theta_1, \theta_2, \ldots, \theta_k \mid M_i) d\theta_1 \, d\theta_2 \, \ldots d\theta_k$$

Model Selection: Bayesian Methods

Bayes factors

The marginal likelihood is a *very* ugly multidimensional integral that cannot be calculated, which is what motivated the Metropolis–Hastings algorithm:

$$R = \min\left[1, \underbrace{\boxed{\frac{\Pr(X \mid \theta')}{\Pr(X \mid \theta)}}}_{\text{likelihood ratio}} \times \underbrace{\boxed{\frac{\Pr(\theta')}{\Pr(\theta)}}}_{\text{prior ratio}} \times \underbrace{\boxed{\frac{\Pr(\theta' \to \theta)}{\Pr(\theta \to \theta')}}}_{\text{proposal ratio}}\right]$$

Model Selection: Bayesian Methods

Bayes factors

The marginal likelihood is a *very* ugly multidimensional integral that cannot be calculated, which is what motivated the Metropolis–Hastings algorithm:

$$R = \min\left[1, \frac{\Pr(X \mid \theta')}{\Pr(X \mid \theta)} \times \frac{\Pr(\theta')}{\Pr(\theta)} \times \frac{\Pr(\theta' \to \theta)}{\Pr(\theta \to \theta')}\right]$$

This allows us to estimate the posterior probability density while avoiding computation of the marginal likelihood:

$$P(\theta_1, \theta_2, \ldots, \theta_k \mid \mathbf{X}, M_i) = \frac{P(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_k, M_i) P(\theta_1, \theta_2, \ldots, \theta_k \mid M_i)}{}$$

Model Selection: Bayesian Methods

Bayes factors

The marginal likelihood is a *very* ugly multidimensional integral that cannot be calculated, which is what motivated the Metropolis–Hastings algorithm:

$$R = \min\left[1, \frac{\Pr(X \mid \theta')}{\Pr(X \mid \theta)} \times \frac{\Pr(\theta')}{\Pr(\theta)} \times \frac{\Pr(\theta' \to \theta)}{\Pr(\theta \to \theta')}\right]$$

This allows us to estimate the posterior probability density while avoiding computation of the marginal likelihood:

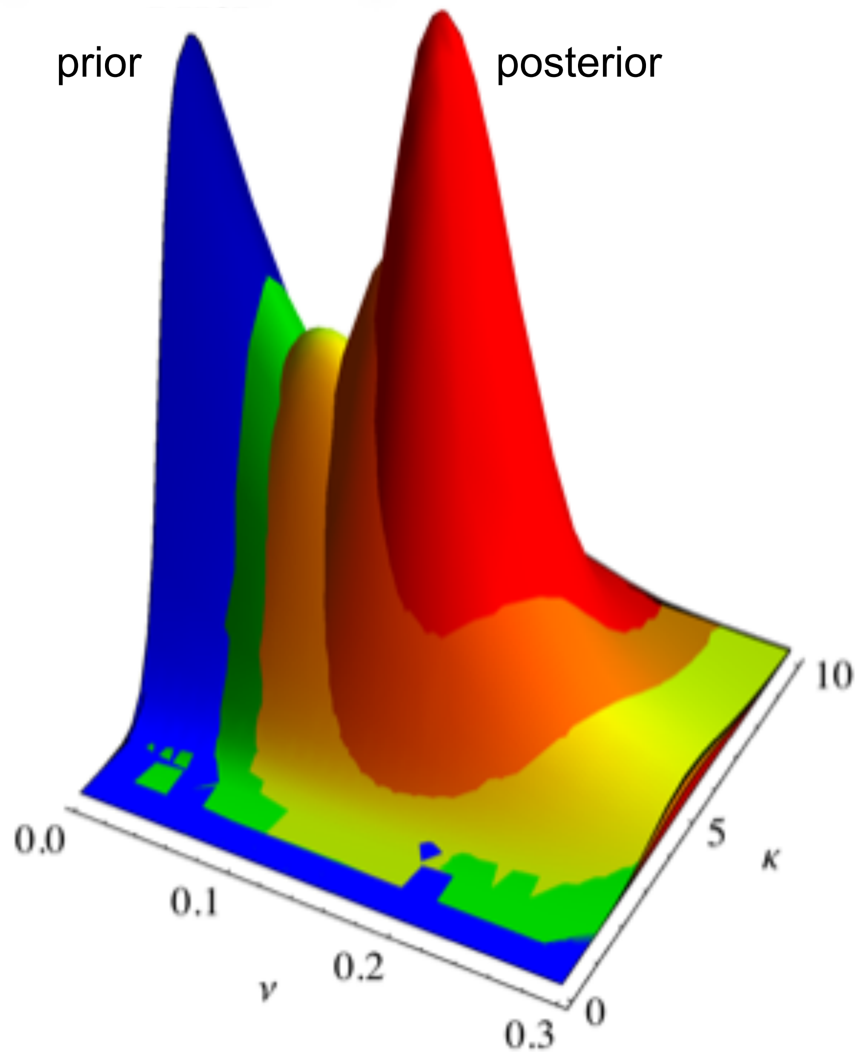
$$P(\theta_1, \theta_2, \ldots, \theta_k \mid \mathbf{X}, M_i) = \frac{P(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_k, M_i) P(\theta_1, \theta_2, \ldots, \theta_k \mid M_i)}{}$$

But, to compare models (test hypotheses), we need the marginal likelihoods!

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

A reliable (but computationally) marginal-likelihood estimator

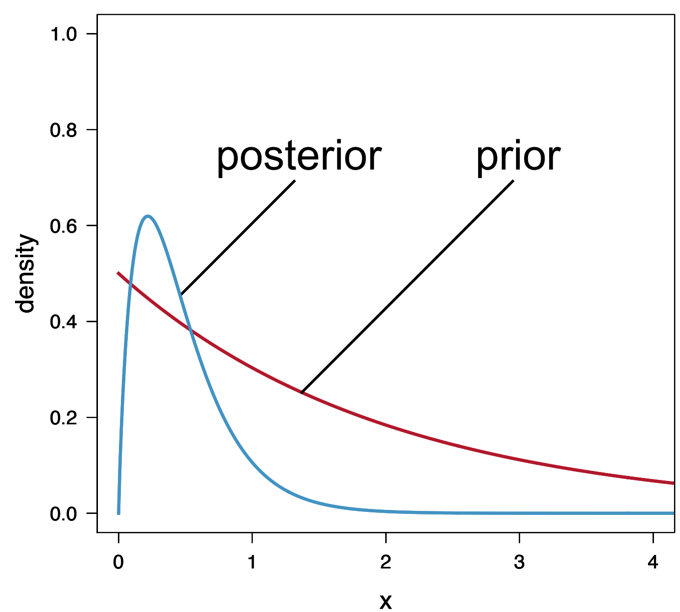


Xie *et al.* 2010

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

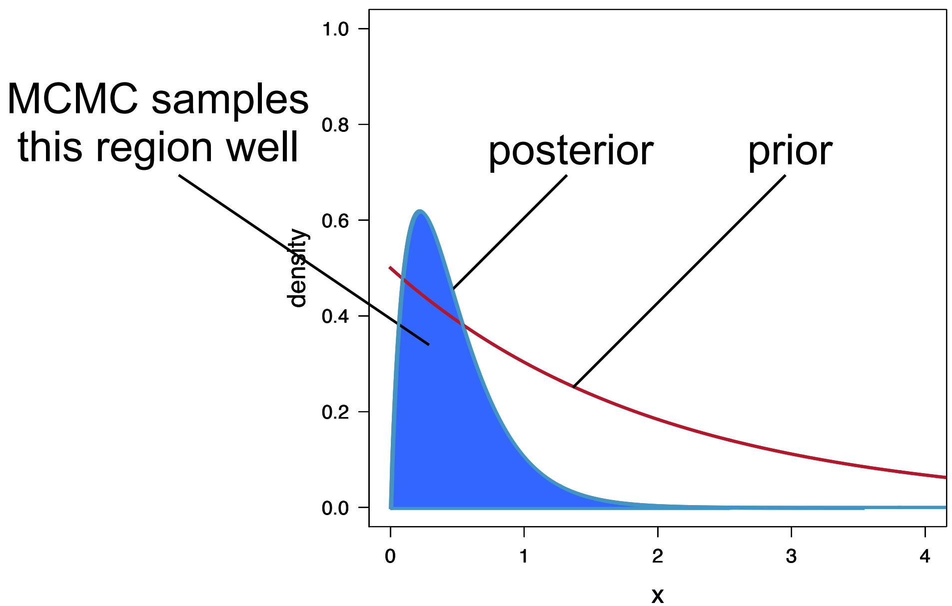
It is difficult to estimate the marginal likelihood from MCMC because the posterior is (hopefully) very focused compared to the prior (so we cannot accurately estimate the likelihood in large regions of parameter space).



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

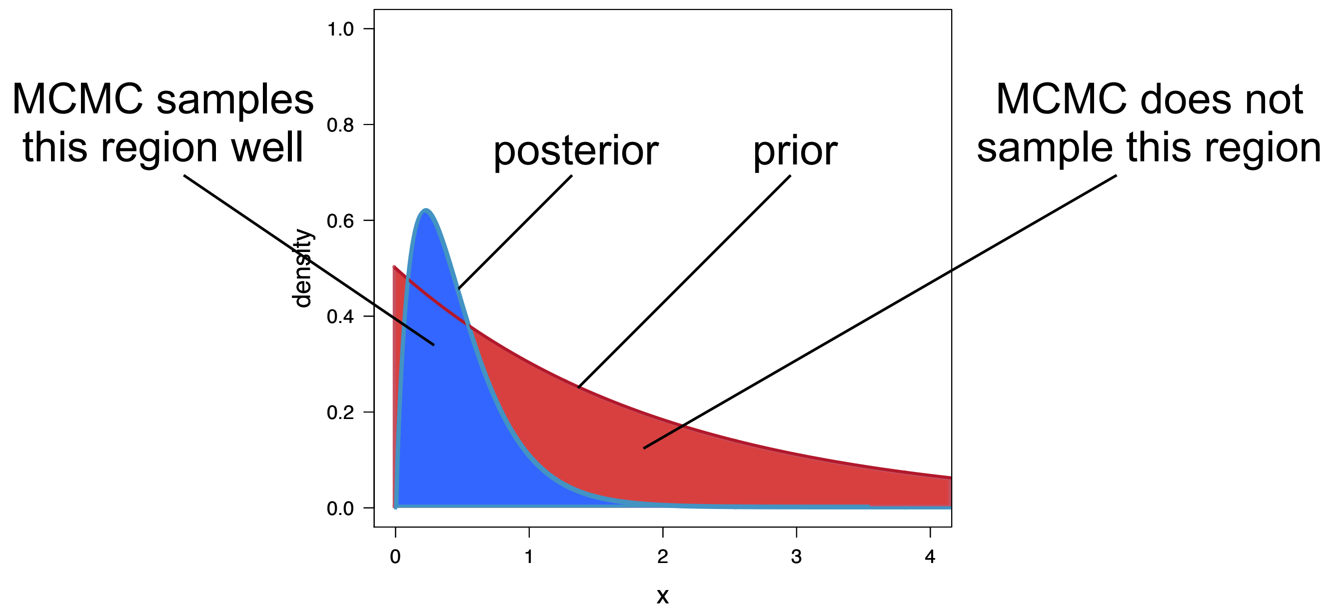
It is difficult to estimate the marginal likelihood from MCMC because the posterior is (hopefully) very focused compared to the prior (so we cannot accurately estimate the likelihood in large regions of parameter space).



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

It is difficult to estimate the marginal likelihood from MCMC because the posterior is (hopefully) very focused compared to the prior (so we cannot accurately estimate the likelihood in large regions of parameter space).



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

Stepping-stone sampling forces the MCMC to sample from power posterior densities ranging from the prior to the posterior.

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

Stepping-stone sampling forces the MCMC to sample from power posterior densities ranging from the prior to the posterior.

$$P(\theta \mid \mathbf{X})_{\beta_i} = \frac{P(\mathbf{X} \mid \theta)^{\beta_i} P(\theta)}{P(\mathbf{X})_{\beta_i}}$$

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

Stepping-stone sampling forces the MCMC to sample from power posterior densities ranging from the prior to the posterior.

power posterior

$$P(\theta \mid \mathbf{X})_{\beta_i} = \frac{P(\mathbf{X} \mid \theta)^{\beta_i} P(\theta)}{P(\mathbf{X})_{\beta_i}}$$

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

Stepping-stone sampling forces the MCMC to sample from power posterior densities ranging from the prior to the posterior.

power posterior

$$P(\theta \mid \mathbf{X})_{\beta_i} = \frac{P(\mathbf{X} \mid \theta)^{\beta_i} P(\theta)}{P(\mathbf{X})_{\beta_i}}$$

When $\beta = 1$, the MCMC targets the joint posterior probability (as usual).

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

Stepping-stone sampling forces the MCMC to sample from power posterior densities ranging from the prior to the posterior.

power posterior

$$P(\theta \mid \mathbf{X})_{\beta_i} = \frac{P(\mathbf{X} \mid \theta)^{\beta_i} P(\theta)}{P(\mathbf{X})_{\beta_i}}$$

When $\beta = 1$, the MCMC targets the joint posterior probability (as usual).

When $\beta = 0$, the MCMC targets the joint prior probability.

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

Stepping-stone sampling forces the MCMC to sample from power posterior densities ranging from the prior to the posterior.

power posterior

$$P(\theta \mid \mathbf{X})_{\beta_i} = \frac{P(\mathbf{X} \mid \theta)^{\beta_i} P(\theta)}{P(\mathbf{X})_{\beta_i}}$$

When $\beta = 1$, the MCMC targets the joint posterior probability (as usual).

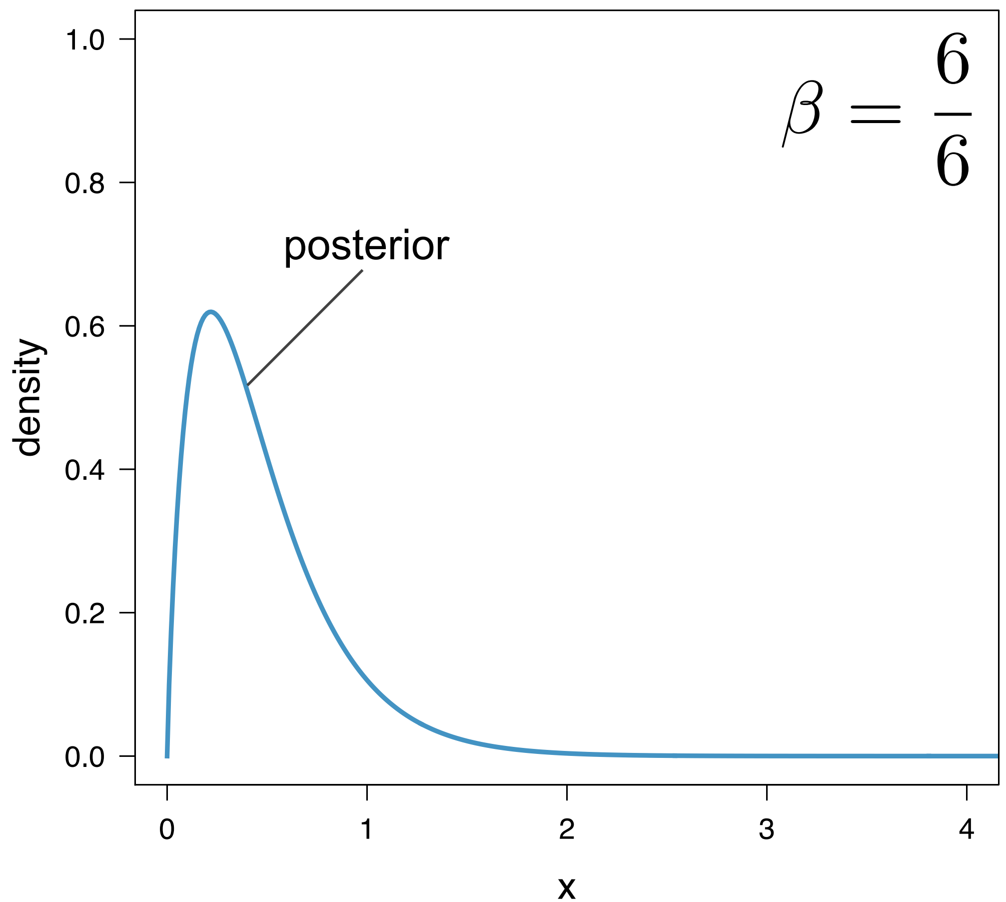
When $\beta = 0$, the MCMC targets the joint prior probability.

We run an MCMC simulation across many power posteriors from $\beta = 0$ to $\beta = 1$ to more accurately characterize the marginal likelihood.

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

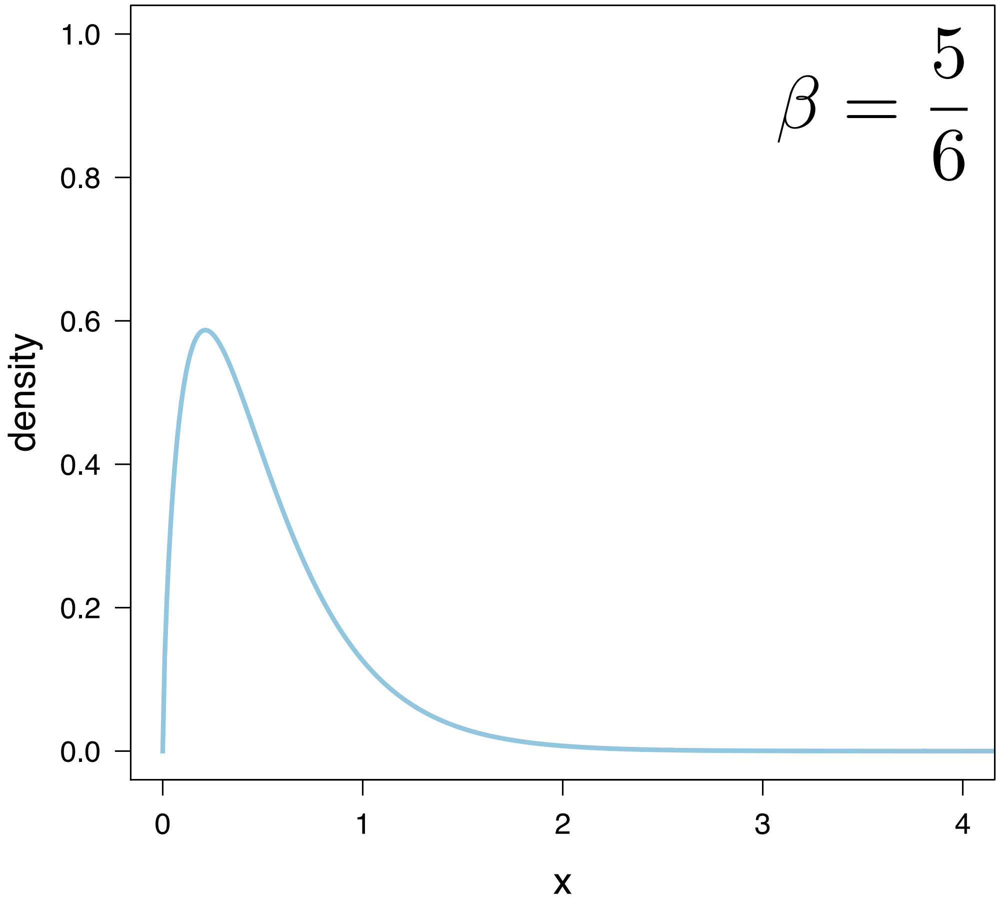
We simulate a Markov chain across a series of steps from the posterior to the prior, where each step corresponds to a power posterior, β.



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

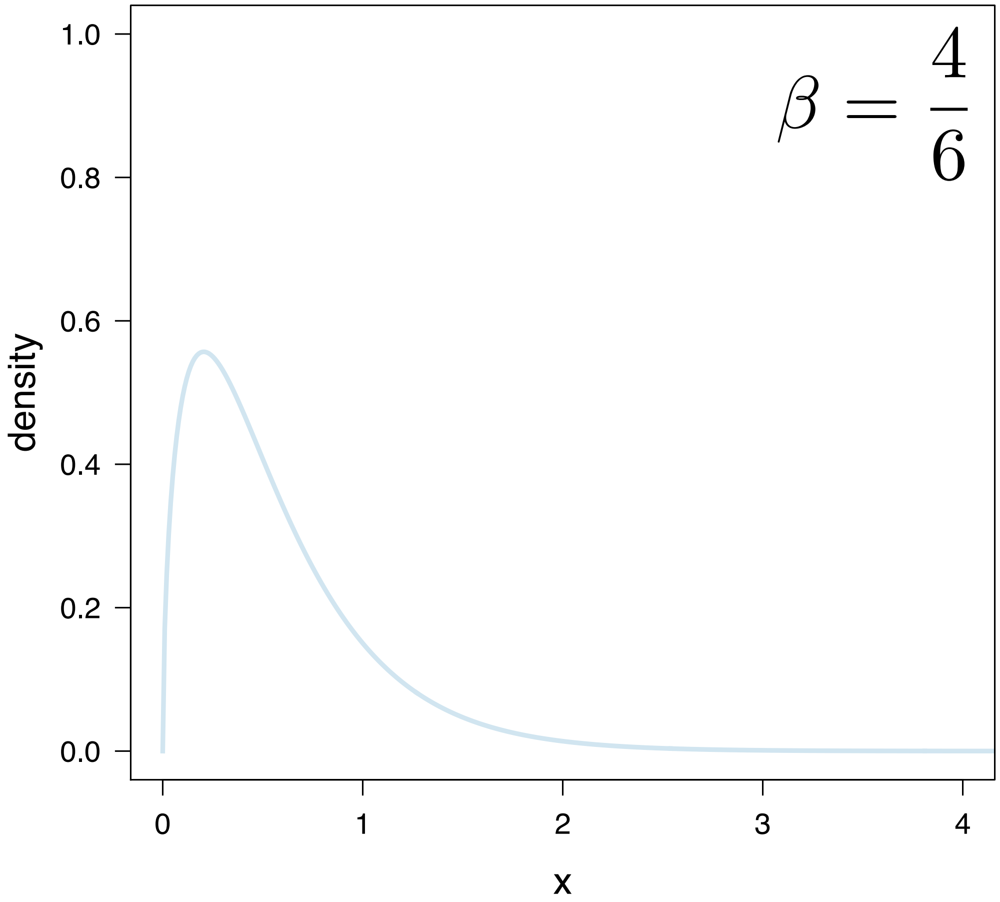
We simulate a Markov chain across a series of steps from the posterior to the prior, where each step corresponds to a power posterior, β.



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

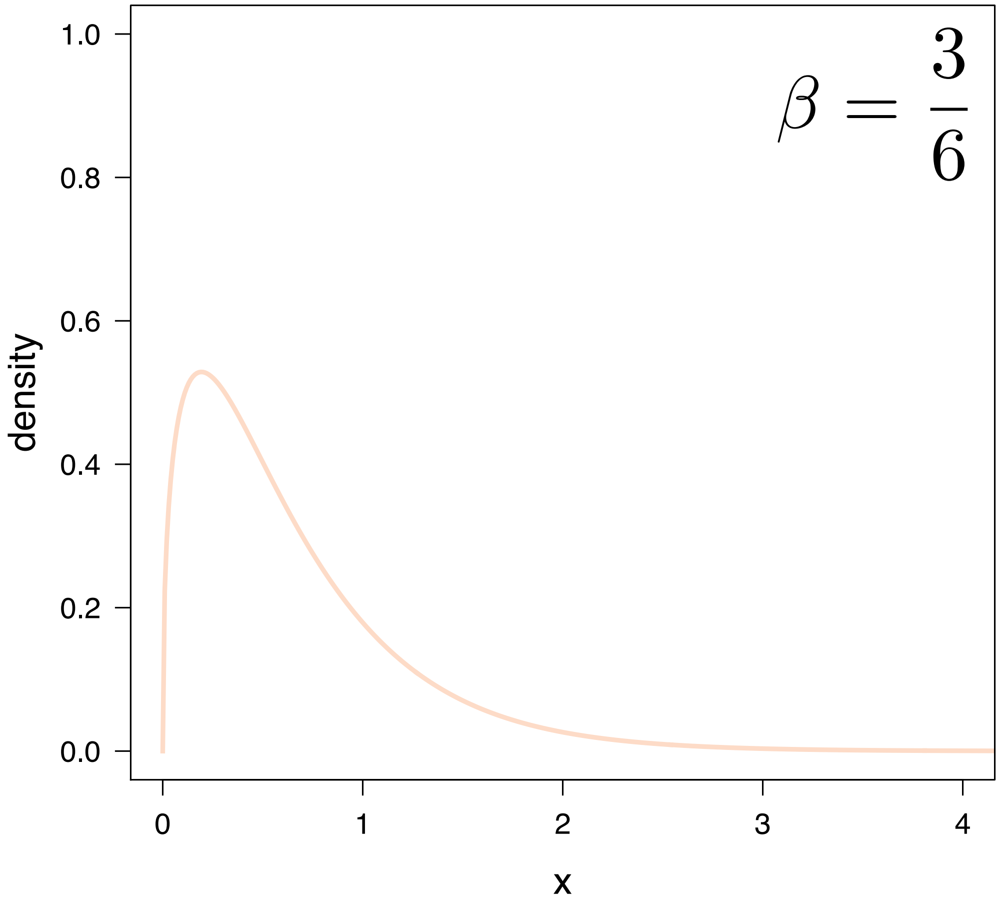
We simulate a Markov chain across a series of steps from the posterior to the prior, where each step corresponds to a power posterior, β.



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

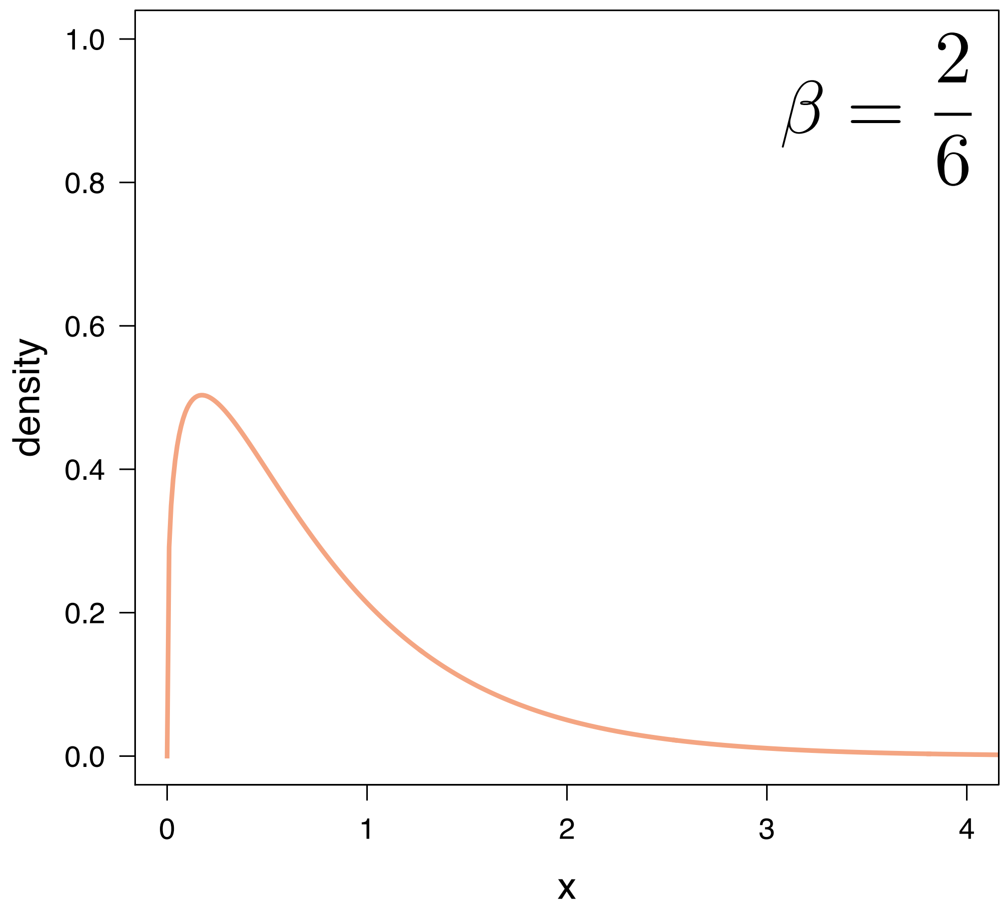
We simulate a Markov chain across a series of steps from the posterior to the prior, where each step corresponds to a power posterior, β.



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

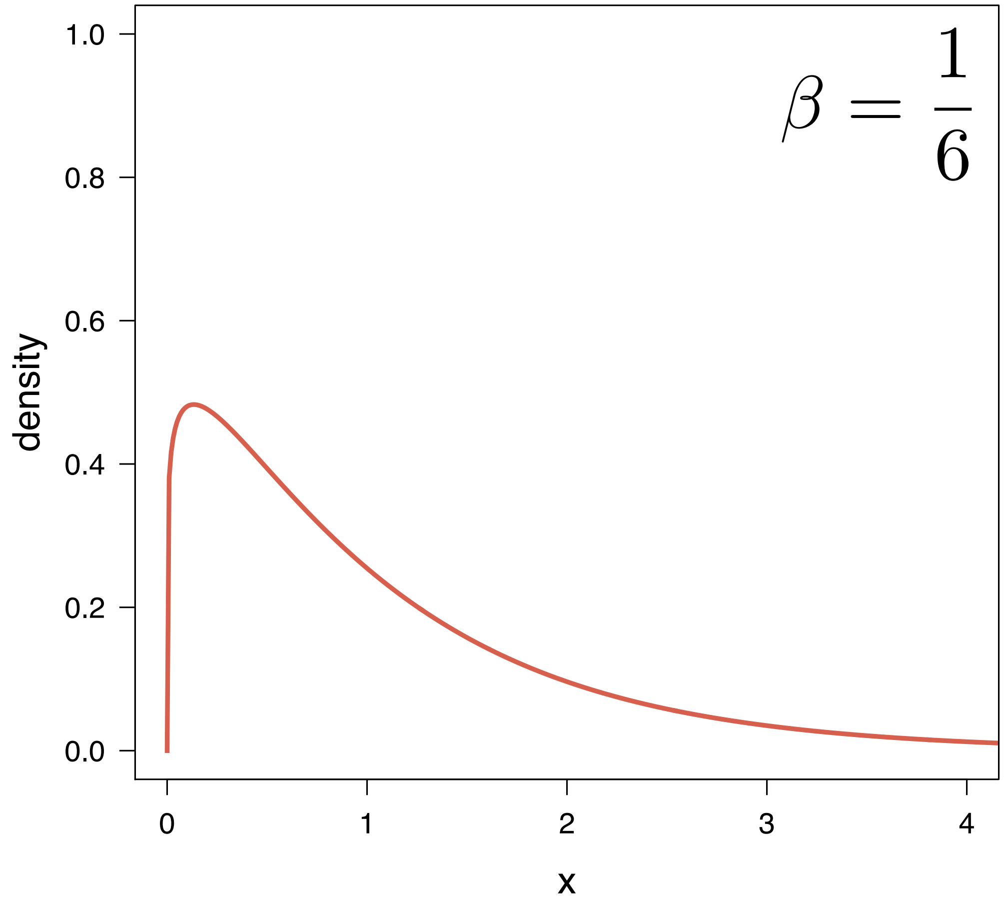
We simulate a Markov chain across a series of steps from the posterior to the prior, where each step corresponds to a power posterior, β.



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

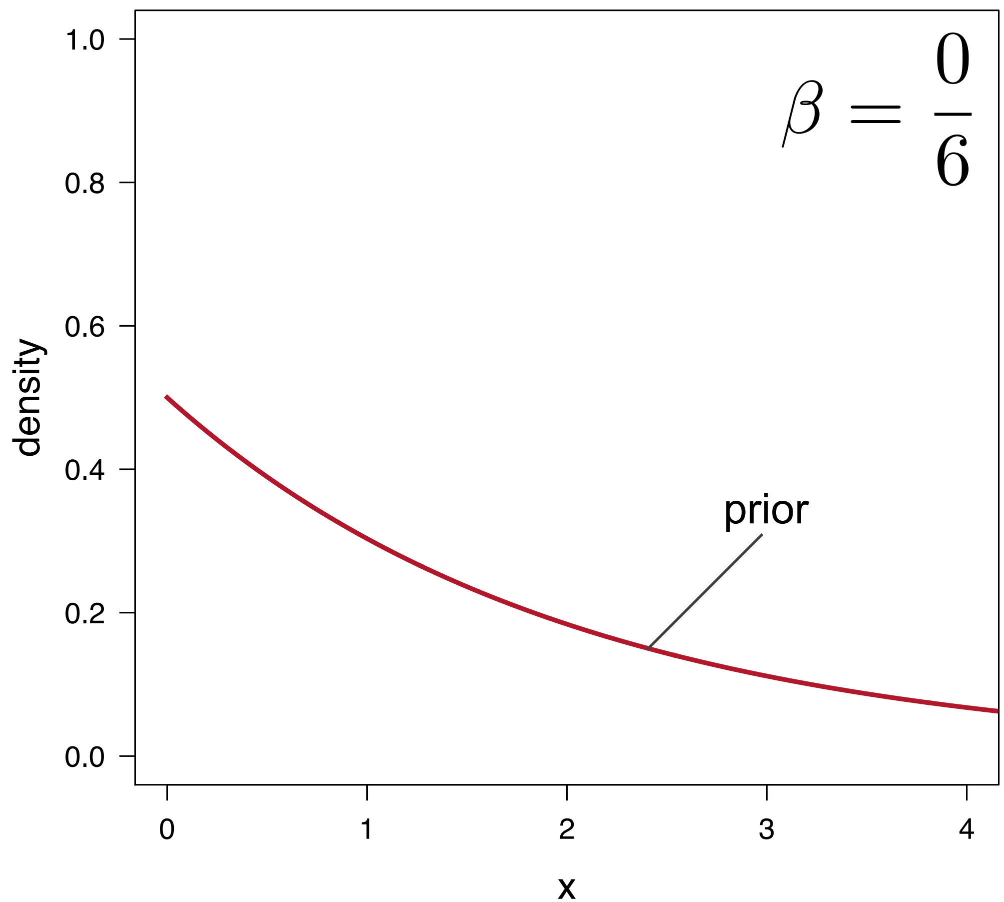
We simulate a Markov chain across a series of steps from the posterior to the prior, where each step corresponds to a power posterior, β.



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

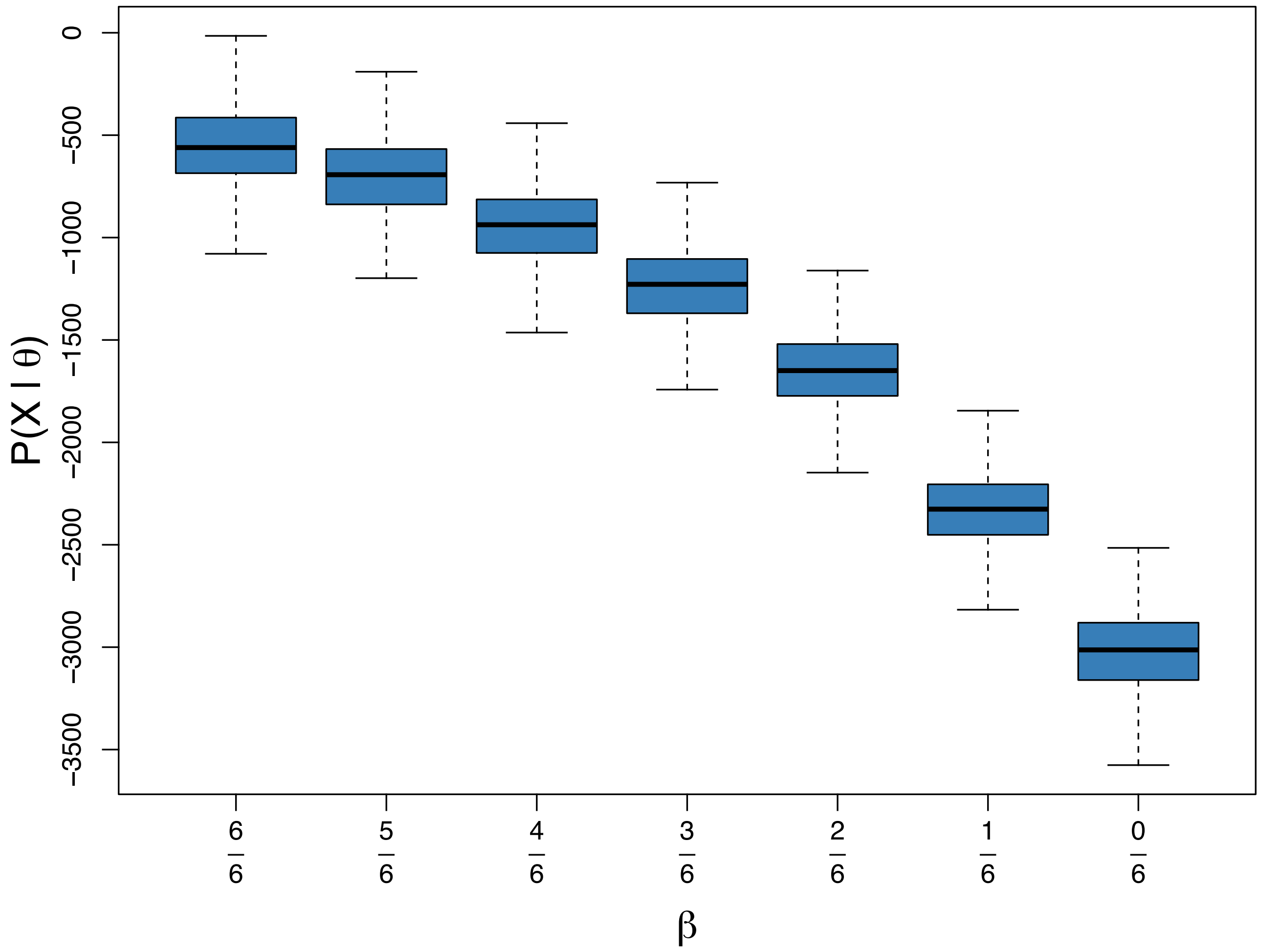
We simulate a Markov chain across a series of steps from the posterior to the prior, where each step corresponds to a power posterior, β.



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

The sampled likelihoods at each stone can be used to estimate the marginal likelihood: path samplers[1] and stepping-stone samplers[2].



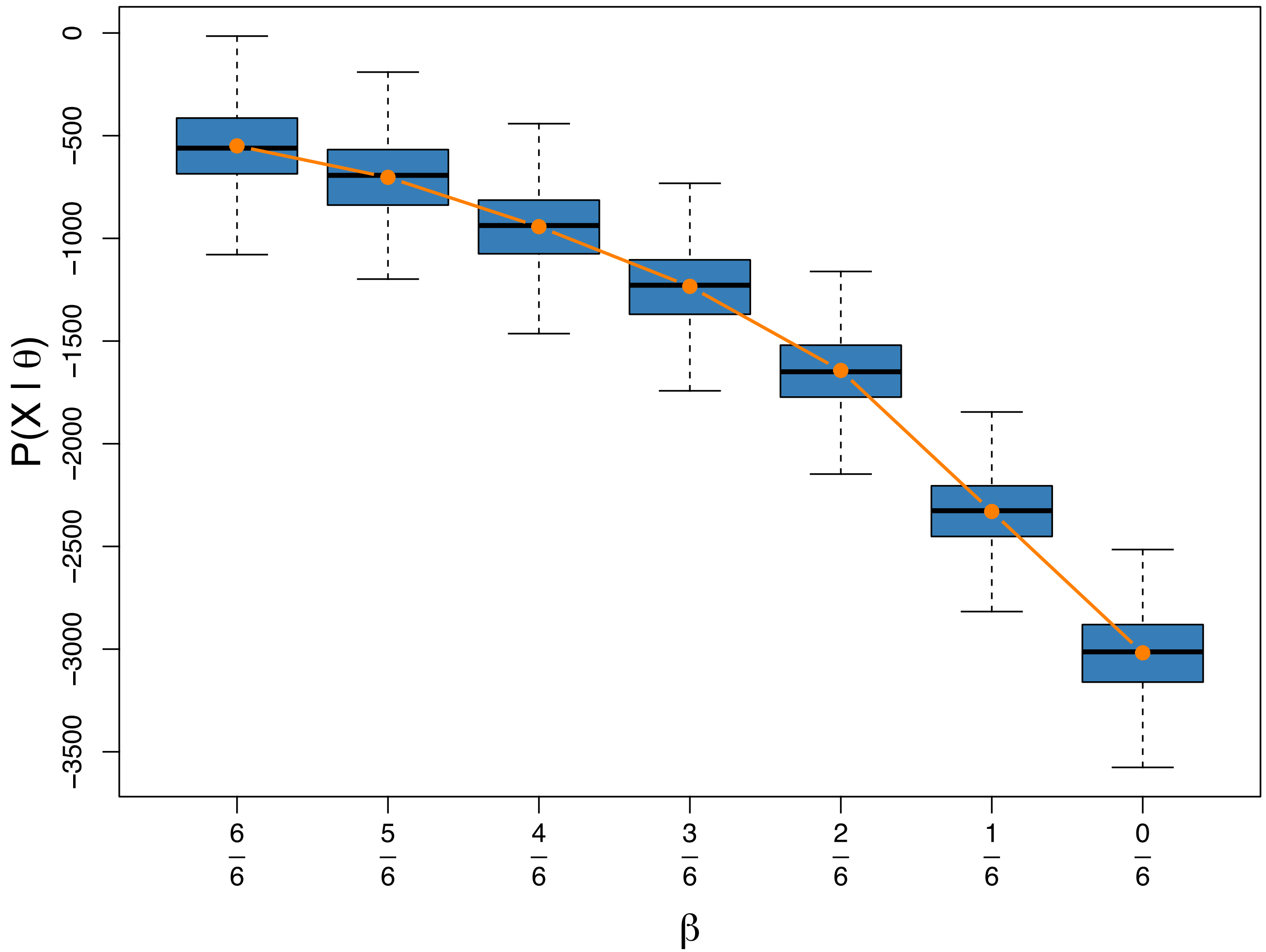
[1]Lartillot (2006)

[2]Xie *et al.* (2011)

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

The sampled likelihoods at each stone can be used to estimate the marginal likelihood: path samplers[1] and stepping-stone samplers[2].



[1]Lartillot (2006)

[2]Xie *et al.* (2011)

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

How should the stones be distributed between the posterior and the prior to efficiently estimate the marginal likelihood?

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

How should the stones be distributed between the posterior and the prior to efficiently estimate the marginal likelihood?

It is more difficult to sample adequately from the relatively diffuse prior compared to the more concentrated posterior distribution.

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

How should the stones be distributed between the posterior and the prior to efficiently estimate the marginal likelihood?

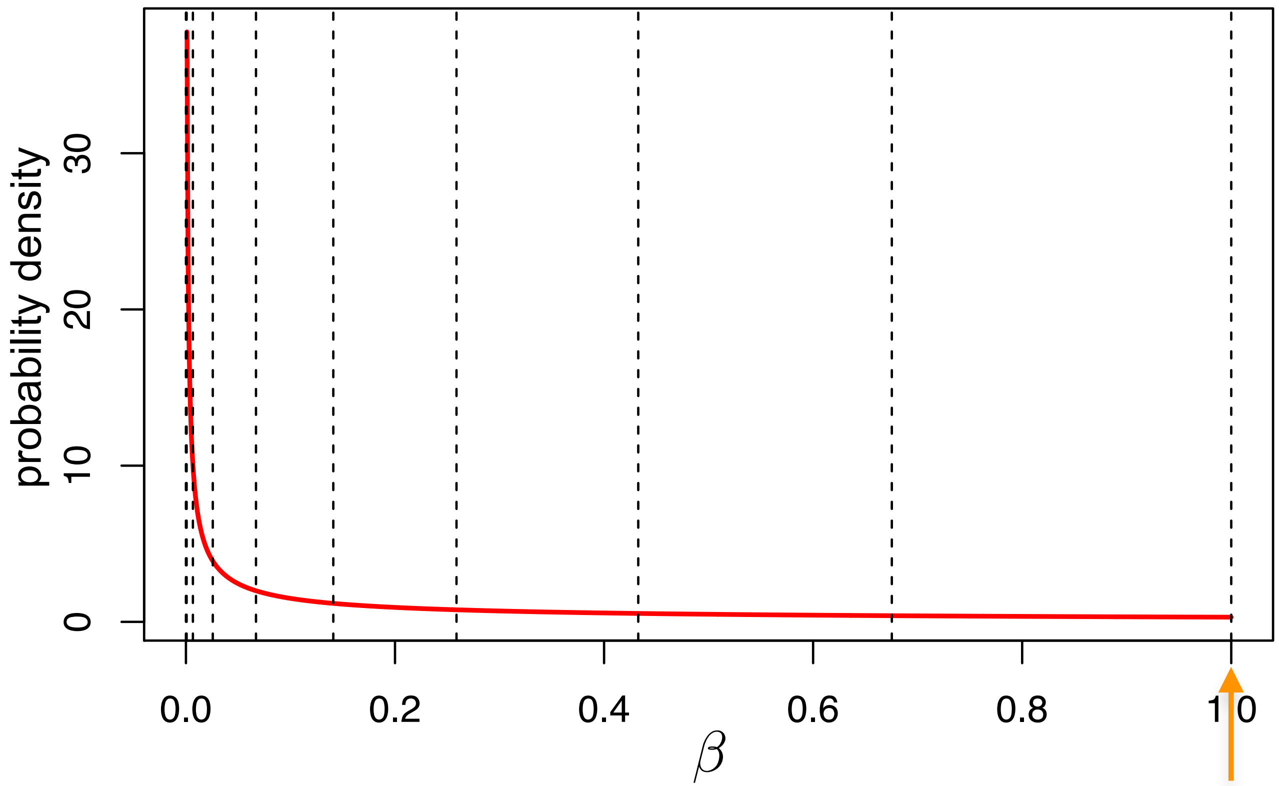
It is more difficult to sample adequately from the relatively diffuse prior compared to the more concentrated posterior distribution.

In order to improve the efficiency of the stepping-stone simulation, the stones are therefore spaced so that they are concentrated near the prior.

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

Experience suggests that spacing the stones as evenly distributed quantiles of a beta distribution works well, where by convention $\mathrm{Beta}(0.3, 1.0)$.

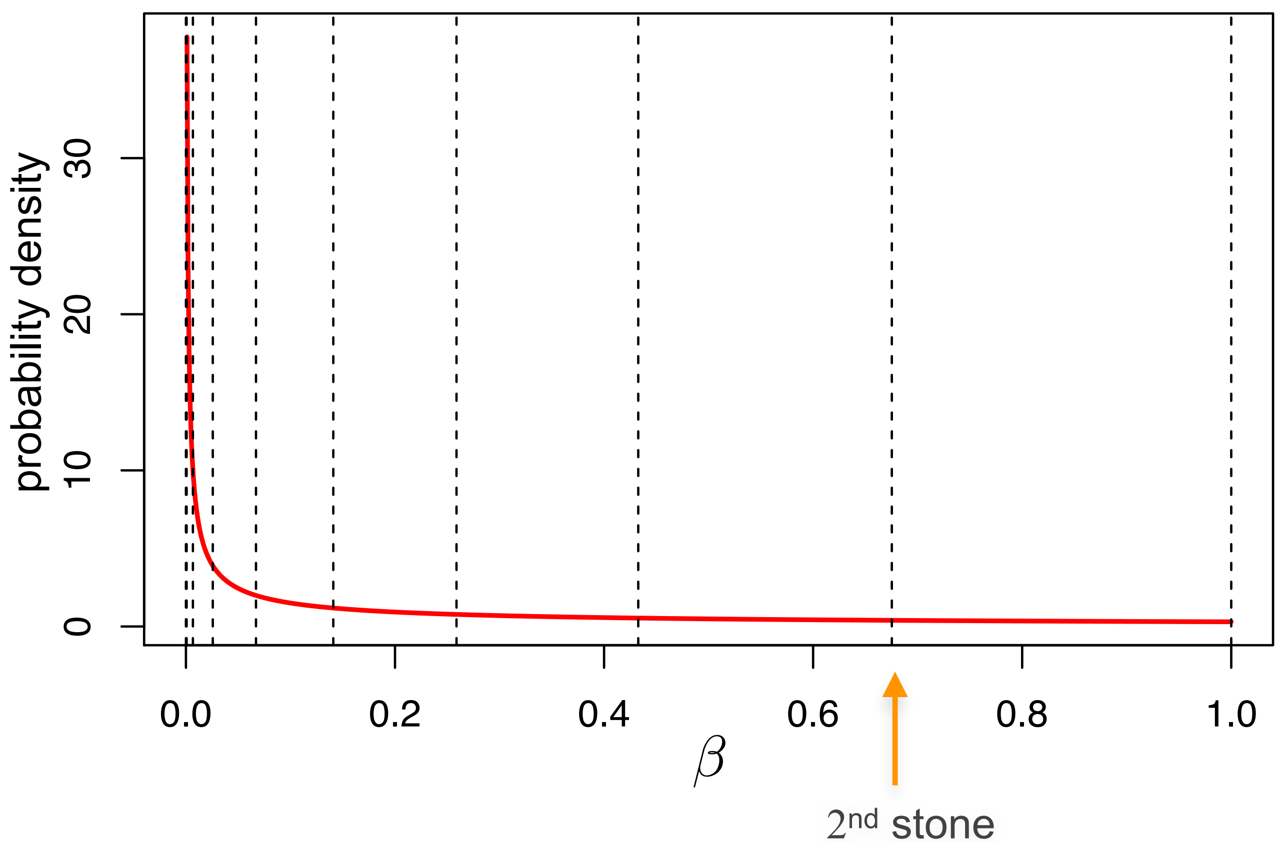


1st stone
(posterior)

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

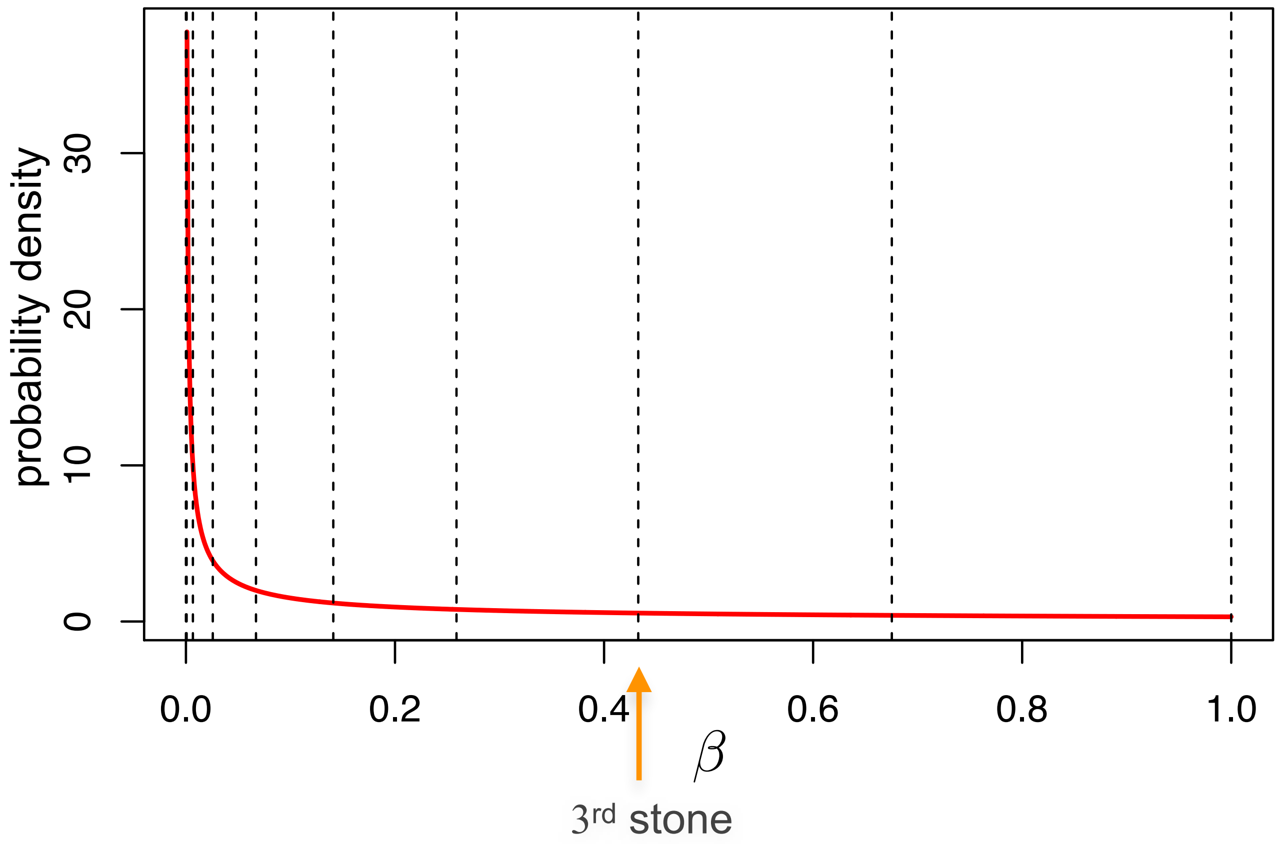
Experience suggests that spacing the stones as evenly distributed quantiles of a beta distribution works well, where by convention $\mathrm{Beta}(0.3, 1.0)$.



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

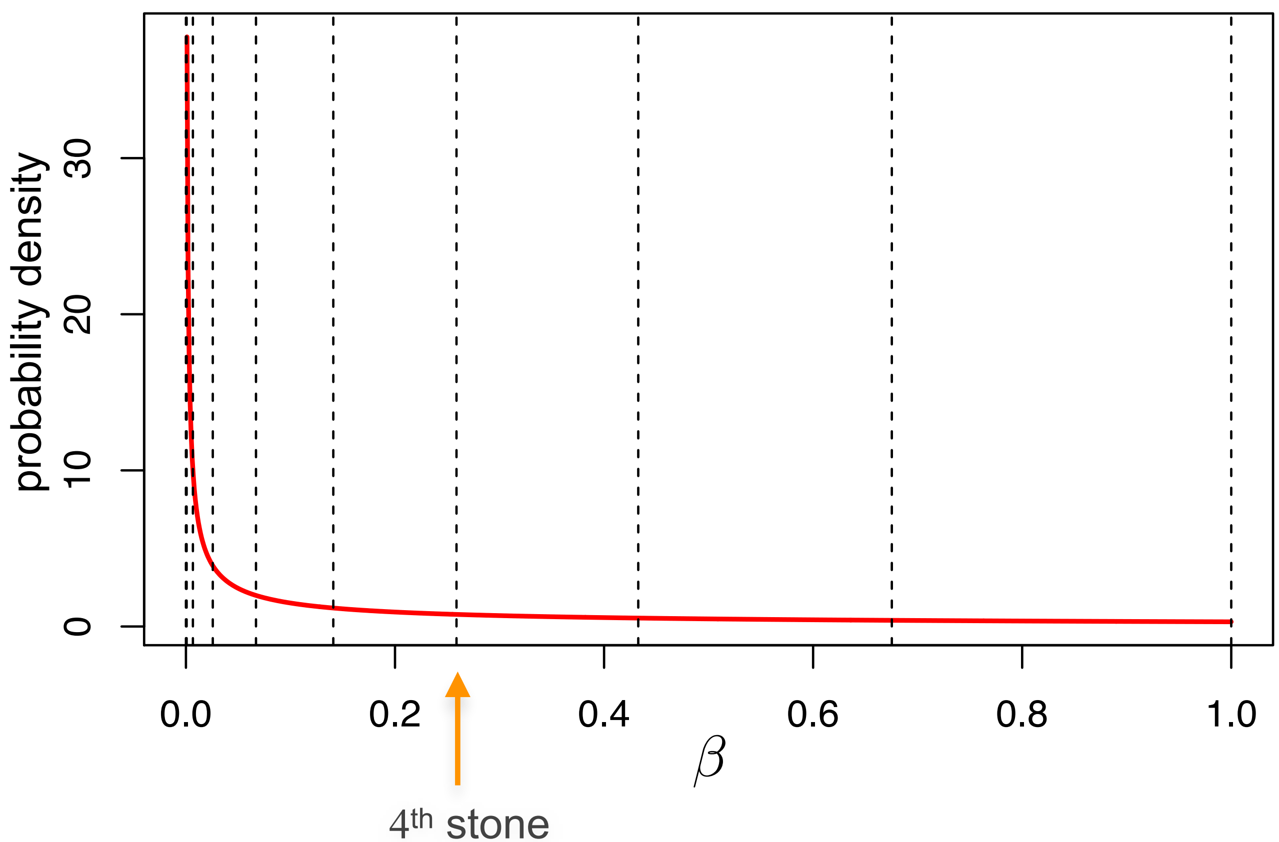
Experience suggests that spacing the stones as evenly distributed quantiles of a beta distribution works well, where by convention $\mathrm{Beta}(0.3, 1.0)$.



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

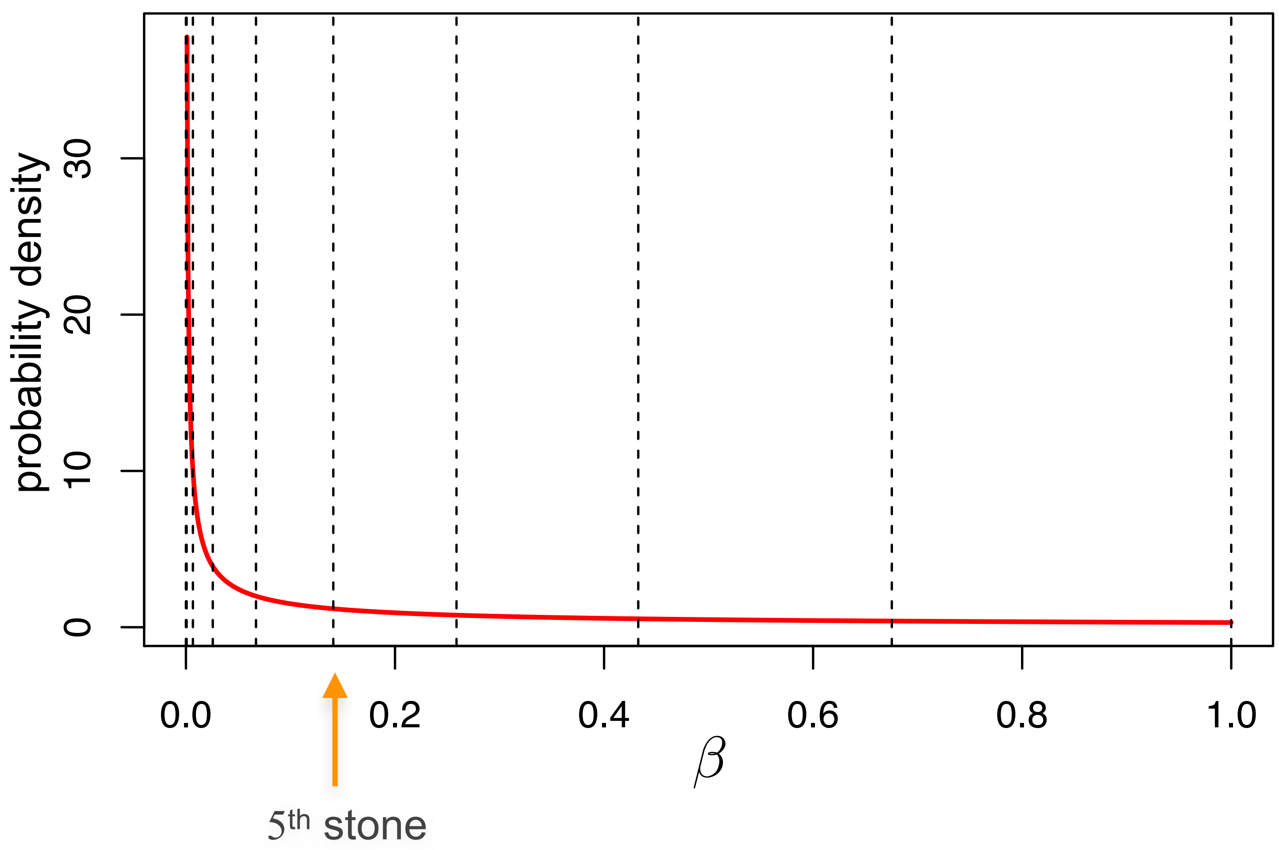
Experience suggests that spacing the stones as evenly distributed quantiles of a beta distribution works well, where by convention $\text{Beta}(0.3,1.0)$.



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

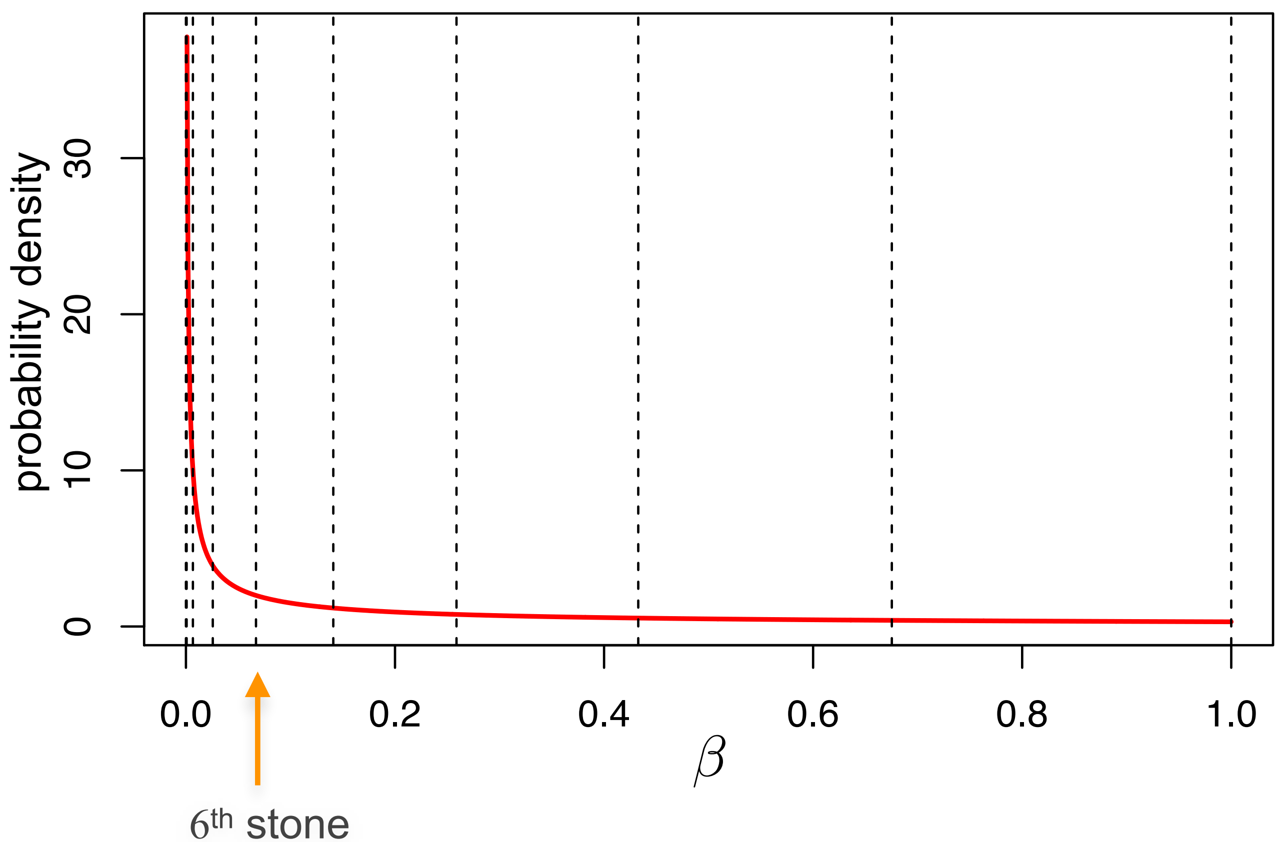
Experience suggests that spacing the stones as evenly distributed quantiles of a beta distribution works well, where by convention $\mathrm{Beta}(0.3, 1.0)$.



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

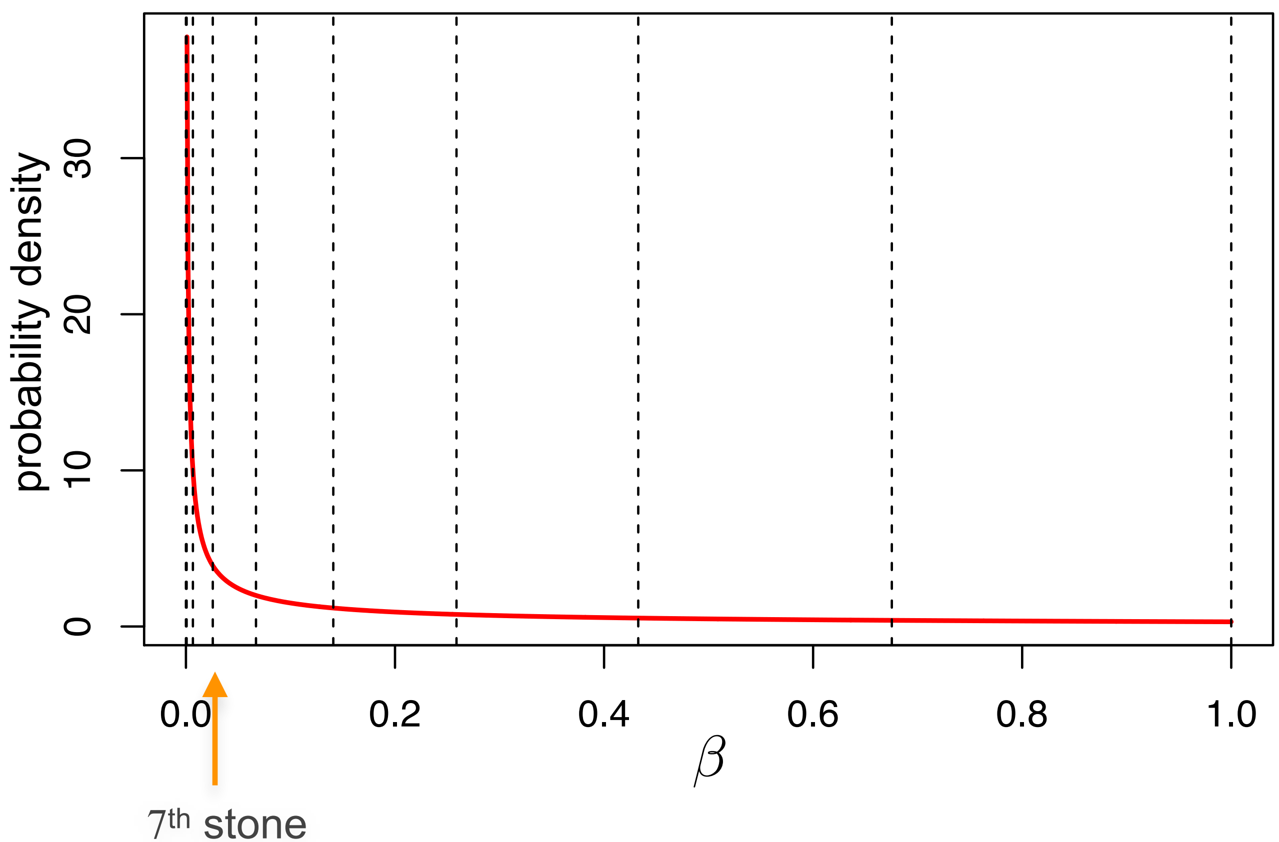
Experience suggests that spacing the stones as evenly distributed quantiles of a beta distribution works well, where by convention $\mathrm{Beta}(0.3,1.0)$.



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

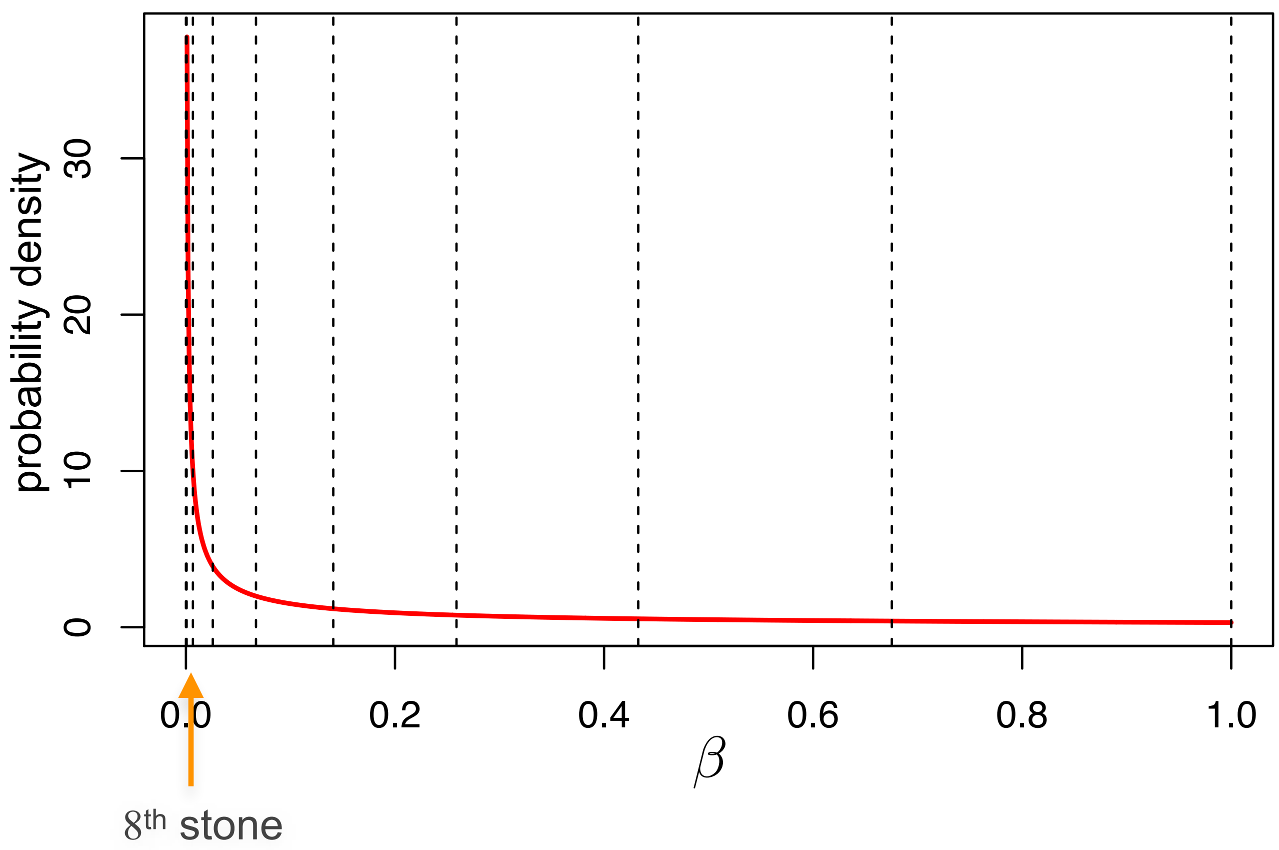
Experience suggests that spacing the stones as evenly distributed quantiles of a beta distribution works well, where by convention $\mathrm{Beta}(0.3, 1.0)$.



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

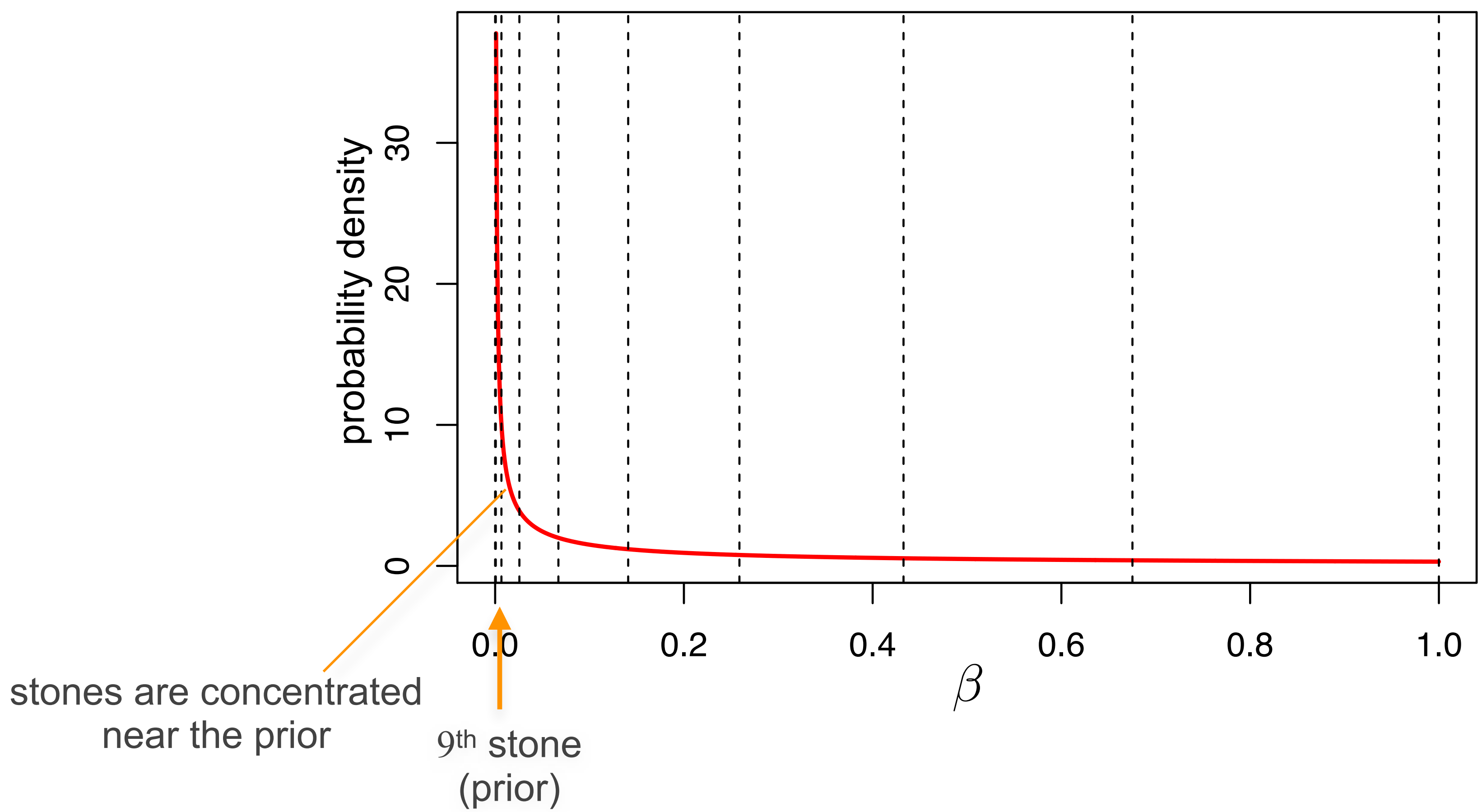
Experience suggests that spacing the stones as evenly distributed quantiles of a beta distribution works well, where by convention $\mathrm{Beta}(0.3, 1.0)$.



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

Experience suggests that spacing the stones as evenly distributed quantiles of a beta distribution works well, where by convention $\mathrm{Beta}(0.3, 1.0)$.



Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

It is important to assess the stability of the marginal-likelihood estimates:

- perform multiple replicate stepping-stone simulations

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

It is important to assess the stability of the marginal-likelihood estimates:

- perform multiple replicate stepping-stone simulations
- the marginal likelihoods from independent simulations should be ‘very similar’

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

It is important to assess the stability of the marginal-likelihood estimates:

- perform multiple replicate stepping-stone simulations
- the marginal likelihoods from independent simulations should be ‘very similar’

The *precision* of the marginal-likelihood estimates can be improved by:

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

It is important to assess the stability of the marginal-likelihood estimates:

- perform multiple replicate stepping-stone simulations
- the marginal likelihoods from independent simulations should be ‘very similar’

The *precision* of the marginal-likelihood estimates can be improved by:

- increasing the number of samples per stone

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

It is important to assess the stability of the marginal-likelihood estimates:

- perform multiple replicate stepping-stone simulations
- the marginal likelihoods from independent simulations should be ‘very similar’

The *precision* of the marginal-likelihood estimates can be improved by:

- increasing the number of samples per stone
- tuning the MCMC proposals at each stone

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

It is important to assess the stability of the marginal-likelihood estimates:

- perform multiple replicate stepping-stone simulations
- the marginal likelihoods from independent simulations should be ‘very similar’

The *precision* of the marginal-likelihood estimates can be improved by:

- increasing the number of samples per stone
- tuning the MCMC proposals at each stone

The *accuracy* of the marginal-likelihood estimates can be improved by:

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

It is important to assess the stability of the marginal-likelihood estimates:

- perform multiple replicate stepping-stone simulations
- the marginal likelihoods from independent simulations should be ‘very similar’

The *precision* of the marginal-likelihood estimates can be improved by:

- increasing the number of samples per stone
- tuning the MCMC proposals at each stone

The *accuracy* of the marginal-likelihood estimates can be improved by:

- increasing the number of stones

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

It is important to assess the stability of the marginal-likelihood estimates:

- perform multiple replicate stepping-stone simulations
- the marginal likelihoods from independent simulations should be 'very similar'

The *precision* of the marginal-likelihood estimates can be improved by:

- increasing the number of samples per stone
- tuning the MCMC proposals at each stone

The *accuracy* of the marginal-likelihood estimates can be improved by:

- increasing the number of stones
- experimenting with different spacings (*i.e.*, varying α from ~ 0.25–0.35)

Model Selection: Bayesian Methods

Estimating marginal likelihoods: power-posterior simulation

It is important to assess the stability of the marginal-likelihood estimates:

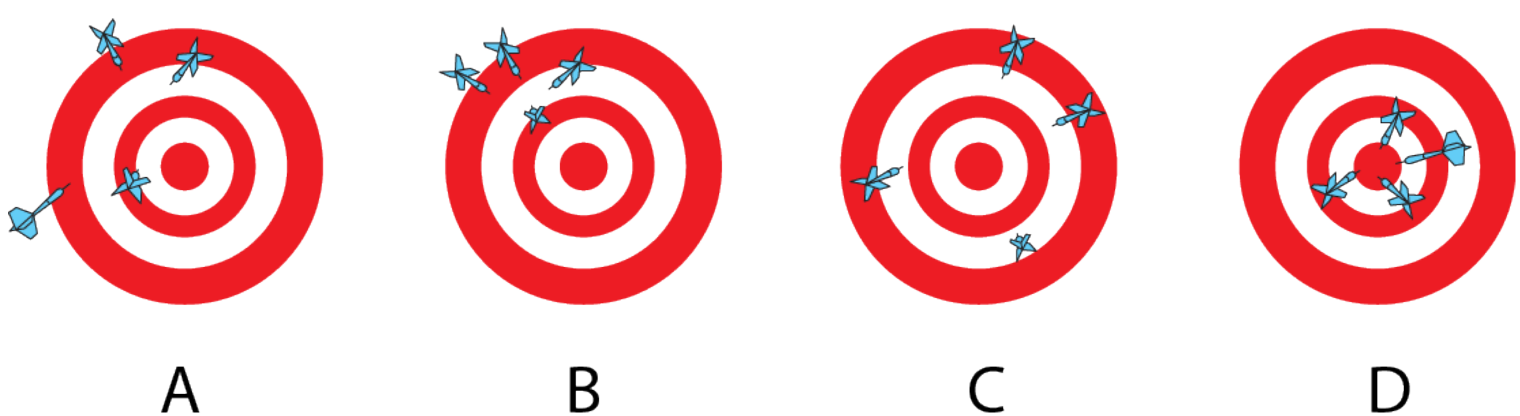
- perform multiple replicate stepping-stone simulations
- the marginal likelihoods from independent simulations should be 'very similar'

The *precision* of the marginal-likelihood estimates can be improved by:

- increasing the number of samples per stone
- tuning the MCMC proposals at each stone

The *accuracy* of the marginal-likelihood estimates can be improved by:

- increasing the number of stones
- experimenting with different spacings (*i.e.*, varying α from ~ 0.25–0.35)



Outline

I. Model selection

What is the <u>relative</u> fit of the candidate models (hypotheses) to our data?

Bayesian methods for selecting among candidate models (hypotheses)

II. Model adequacy

What is the <u>absolute</u> fit of the candidate models (hypotheses) to our data?

Bayesian methods for assessing model adequacy of candidate models (hypotheses)

III. Model averaging

How do we accommodate uncertainty in the choice among candidate models?

Bayesian methods for averaging over candidate models (hypotheses)

Outline

I. Model selection

What is the relative fit of the candidate models (hypotheses) to our data?

Bayesian methods for selecting among candidate models (hypotheses)

II. Model adequacy

What is the absolute fit of the candidate models (hypotheses) to our data?

Bayesian methods for assessing model adequacy of candidate models (hypotheses)

III. Model averaging

How do we accommodate uncertainty in the choice among candidate models?

Bayesian methods for averaging over candidate models (hypotheses)

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

Even the best candidate model may not provide a good description of the process that gave rise to our data.

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

Even the best candidate model may not provide a good description of the process that gave rise to our data.

We therefore need to assess the *absolute* fit of a given model to our data.

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

Even the best candidate model may not provide a good description of the process that gave rise to our data.

We therefore need to assess the *absolute* fit of a given model to our data.

To assess adequacy, we adopt the premise that if a given model provides an adequate description of the process that gave rise to our data, then we should be able to use that model to simulate datasets that look like our data.

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

This approach is the Bayesian analog of parametric bootstrapping.

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

This approach is the Bayesian analog of parametric bootstrapping.

1. Describe (in a single number) what our dataset “looks” like.

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

This approach is the Bayesian analog of parametric bootstrapping.

1. Describe (in a single number) what our dataset "looks" like.

   We could use many different summary statistics, such as the multinomial likelihood:

$$T(\mathbf{X}) = \ln\left(\prod_{i=1}^{n}\left(\frac{N_{\theta_i}}{N}\right)^{N_{\theta_i}}\right)$$

$\mathbf{X}$ the dataset

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

This approach is the Bayesian analog of parametric bootstrapping.

1. Describe (in a single number) what our dataset "looks" like.

We could use many different summary statistics, such as the multinomial likelihood:

$$T(\mathbf{X}) = \ln\left(\prod_{i=1}^{n}\left(\frac{N_{\theta_i}}{N}\right)^{N_{\theta_i}}\right)$$

$\mathbf{X}$ the dataset

N the total number of sites

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

This approach is the Bayesian analog of parametric bootstrapping.

1. Describe (in a single number) what our dataset “looks” like.

We could use many different summary statistics, such as the multinomial likelihood:

$$T(\mathbf{X}) = \ln\left(\prod_{i=1}^{n}\left(\frac{N_{\theta_i}}{N}\right)^{N_{\theta_i}}\right)$$

$\mathbf{X}$ the dataset

N the total number of sites

n the number of unique site patterns

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

This approach is the Bayesian analog of parametric bootstrapping.

1. Describe (in a single number) what our dataset "looks" like.

   We could use many different summary statistics, such as the multinomial likelihood:

$$T(\mathbf{X}) = \ln\left(\prod_{i=1}^{n}\left(\frac{N_{\theta_i}}{N}\right)^{N_{\theta_i}}\right)$$

$\mathbf{X}$ the dataset

N the total number of sites

n the number of unique site patterns

θ_i the i^{th} unique site pattern

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

This approach is the Bayesian analog of parametric bootstrapping.

1. Describe (in a single number) what our dataset "looks" like.

We could use many different summary statistics, such as the multinomial likelihood:

$$T(\mathbf{X}) = \ln\left(\prod_{i=1}^{n}\left(\frac{N_{\theta_i}}{N}\right)^{N_{\theta_i}}\right)$$

$\mathbf{X}$ the dataset

N the total number of sites

n the number of unique site patterns

θ_i the i^{th} unique site pattern

N_{θ_i} the number of instances of site pattern θ_i

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

This approach is the Bayesian analog of parametric bootstrapping.

1. Describe (in a single number) what our dataset “looks” like.

   We could use many different summary statistics, such as the multinomial likelihood:

$$T(\mathbf{X}) = \ln\left(\prod_{i=1}^{n}\left(\frac{N_{\theta_i}}{N}\right)^{N_{\theta_i}}\right)$$

2. Compute the summary statistic for the original dataset, T_{obs}.

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

This approach is the Bayesian analog of parametric bootstrapping.

1. Describe (in a single number) what our dataset “looks” like.

   We could use many different summary statistics, such as the multinomial likelihood:

$$T(\mathbf{X}) = \ln\left(\prod_{i=1}^{n}\left(\frac{N_{\theta_i}}{N}\right)^{N_{\theta_i}}\right)$$

2. Compute the summary statistic for the original dataset, T_{obs}.
3. Estimate the joint posterior probability density from the original dataset.

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

This approach is the Bayesian analog of parametric bootstrapping.

1. Describe (in a single number) what our dataset "looks" like.

   We could use many different summary statistics, such as the multinomial likelihood:

$$T(\mathbf{X}) = \ln\left(\prod_{i=1}^{n}\left(\frac{N_{\theta_i}}{N}\right)^{N_{\theta_i}}\right)$$

2. Compute the summary statistic for the original dataset, T_{obs}.
3. Estimate the joint posterior probability density from the original dataset.
4. Draw a sample of parameters from the joint posterior probability density and simulate a new dataset the same with N sites.

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

This approach is the Bayesian analog of parametric bootstrapping.

1. Describe (in a single number) what our dataset “looks” like.

   We could use many different summary statistics, such as the multinomial likelihood:

$$T(\mathbf{X}) = \ln\left(\prod_{i=1}^{n}\left(\frac{N_{\theta_i}}{N}\right)^{N_{\theta_i}}\right)$$

2. Compute the summary statistic for the original dataset, T_{obs}.
3. Estimate the joint posterior probability density from the original dataset.
4. Draw a sample of parameters from the joint posterior probability density and simulate a new dataset the same with N sites.
5. Compute the summary statistic for the simulated dataset, T_{sim}.

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

This approach is the Bayesian analog of parametric bootstrapping.

1. Describe (in a single number) what our dataset "looks" like.

   We could use many different summary statistics, such as the multinomial likelihood:

$$T(\mathbf{X}) = \ln\left(\prod_{i=1}^{n}\left(\frac{N_{\theta_i}}{N}\right)^{N_{\theta_i}}\right)$$

2. Compute the summary statistic for the original dataset, T_{obs}.
3. Estimate the joint posterior probability density from the original dataset.
4. Draw a sample of parameters from the joint posterior probability density and simulate a new dataset the same with N sites.
5. Compute the summary statistic for the simulated dataset, T_{sim}.
6. Repeat steps 4 and 5 many times, R.

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

This approach is the Bayesian analog of parametric bootstrapping.

1. Describe (in a single number) what our dataset "looks" like.

   We could use many different summary statistics, such as the multinomial likelihood:

$$T(\mathbf{X}) = \ln\left(\prod_{i=1}^{n}\left(\frac{N_{\theta_i}}{N}\right)^{N_{\theta_i}}\right)$$

2. Compute the summary statistic for the original dataset, T_{obs}.
3. Estimate the joint posterior probability density from the original dataset.
4. Draw a sample of parameters from the joint posterior probability density and simulate a new dataset the same with N sites.
5. Compute the summary statistic for the simulated dataset, T_{sim}.
6. Repeat steps 4 and 5 many times, R.
7. Make a histogram from the summary statistics: this is the predicted distribution.

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

This approach is the Bayesian analog of parametric bootstrapping.

1. Describe (in a single number) what our dataset "looks" like.

   We could use many different summary statistics, such as the multinomial likelihood:

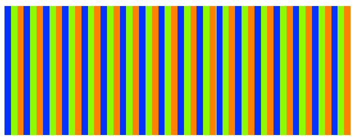
$$T(\mathbf{X}) = \ln\left(\prod_{i=1}^{n}\left(\frac{N_{\theta_i}}{N}\right)^{N_{\theta_i}}\right)$$

2. Compute the summary statistic for the original dataset, T_{obs}.
3. Estimate the joint posterior probability density from the original dataset.
4. Draw a sample of parameters from the joint posterior probability density and simulate a new dataset the same with N sites.
5. Compute the summary statistic for the simulated dataset, T_{sim}.
6. Repeat steps 4 and 5 many times, R.
7. Make a histogram from the summary statistics: this is the predicted distribution.
8. Compare T_{obs} to the distribution predicted from the posterior distribution.

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

Original data
matrix

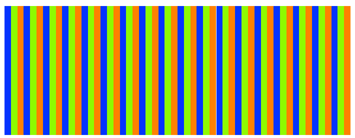


compute the summary
statistic for the observed
dataset, T_{obs}

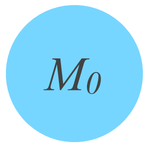
Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

Original data matrix



Candidate model



estimate the joint posterior probability distribution using MCMC

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

Original data matrix

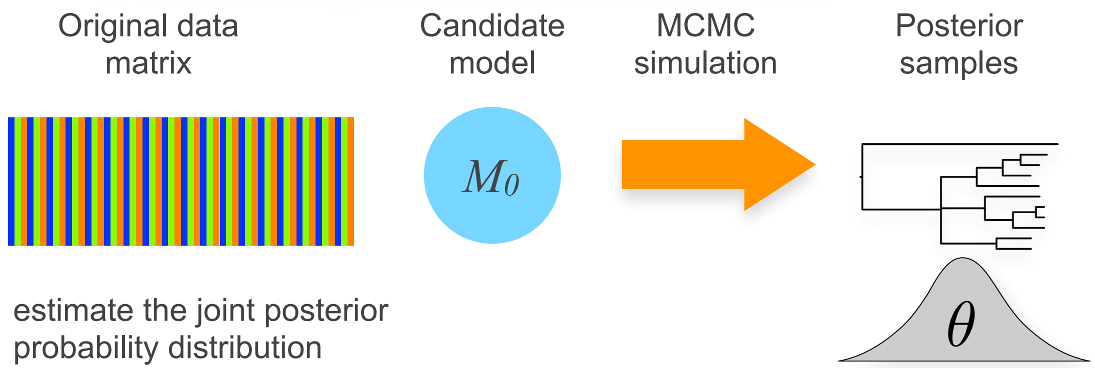
Candidate model

MCMC simulation

estimate the joint posterior
probability distribution
using MCMC

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

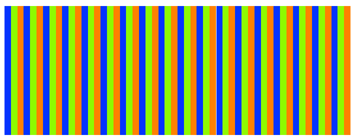


estimate the joint posterior probability distribution using MCMC

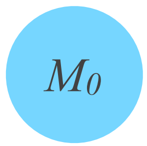
Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation

Original data matrix

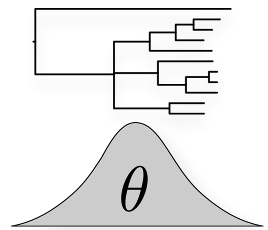


Candidate model



MCMC simulation

Posterior samples

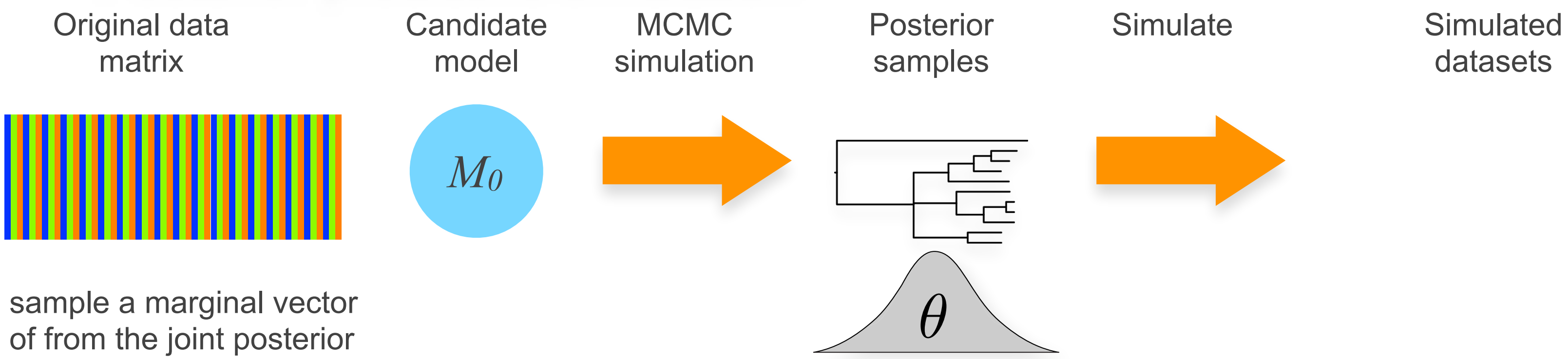


Simulate

sample a marginal vector
of from the joint posterior
and simulate a new dataset

Bayesian Evaluation of Model Adequacy

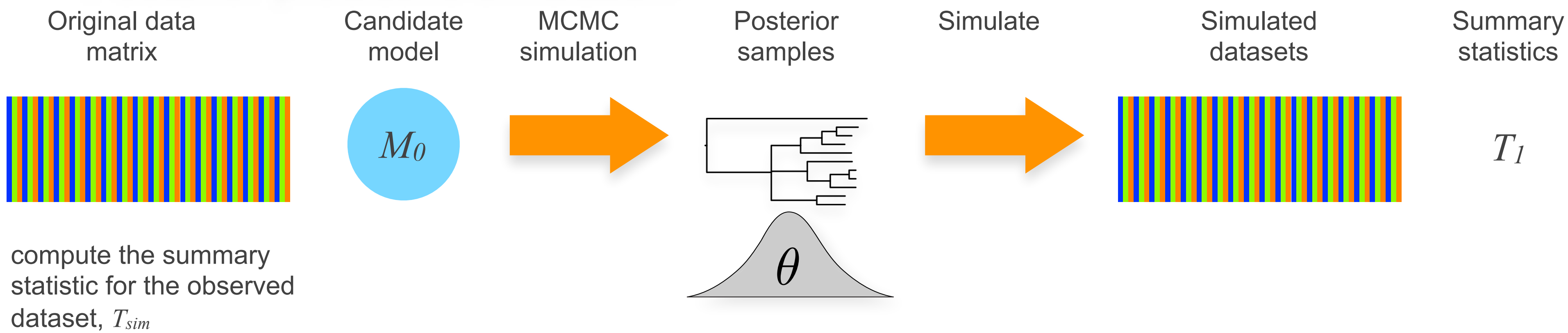
Posterior-predictive simulation



sample a marginal vector
of from the joint posterior
and simulate a new dataset

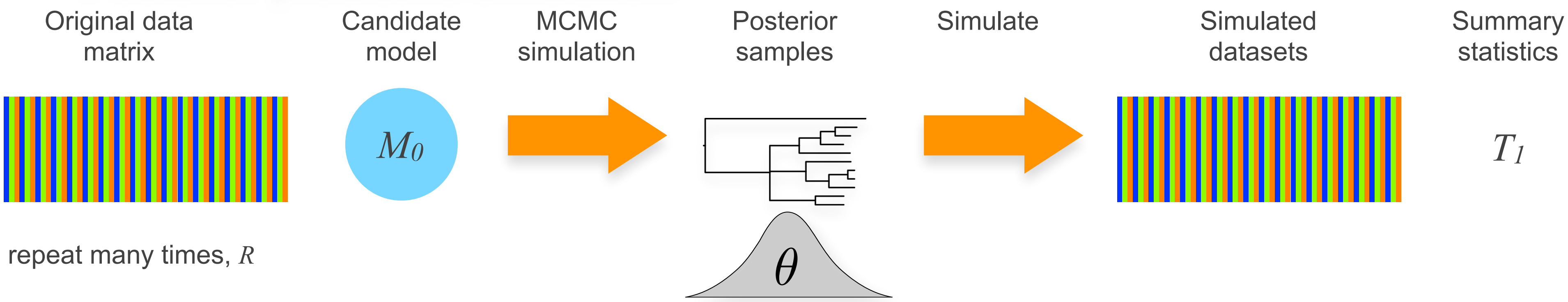
Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation



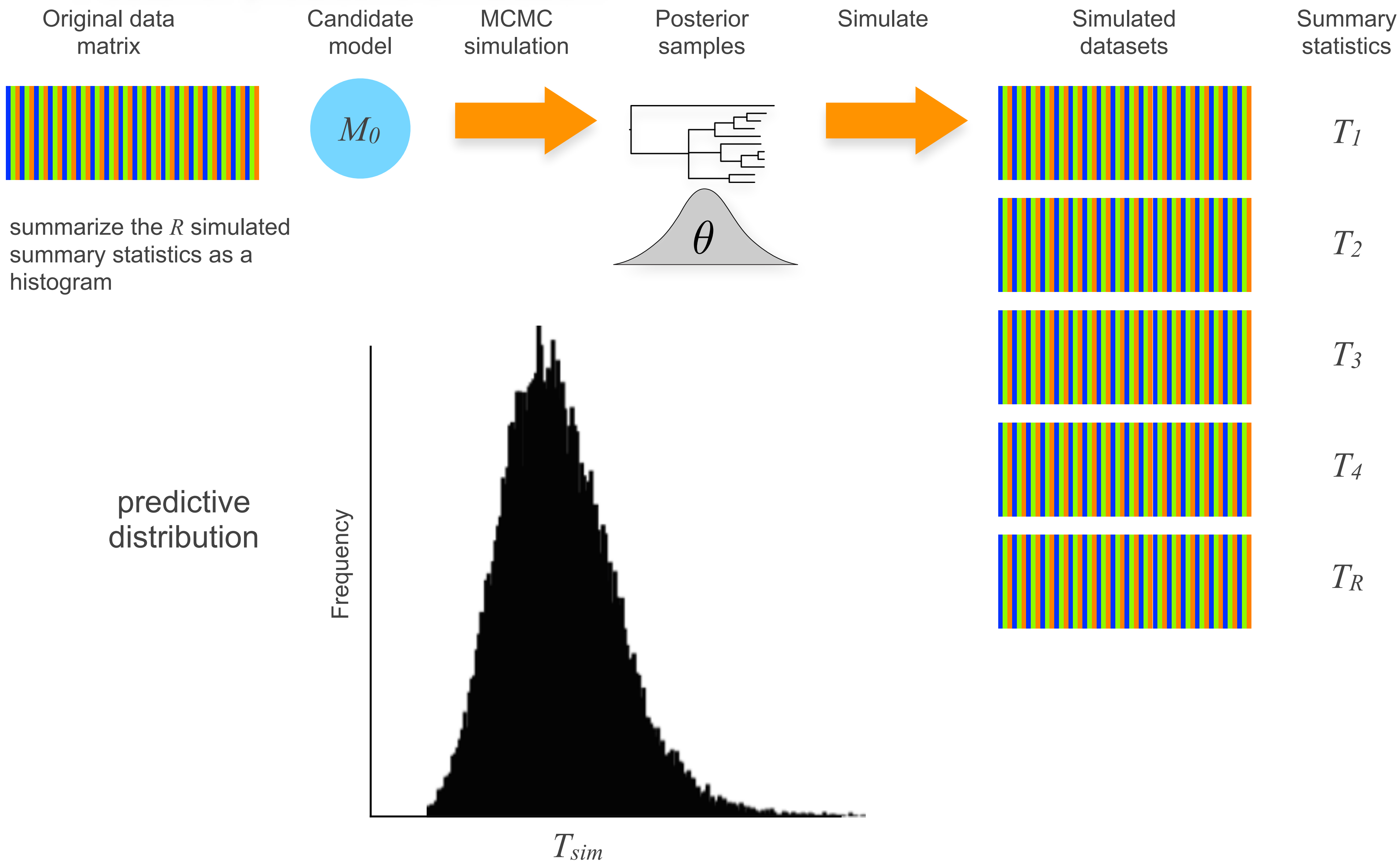
Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation



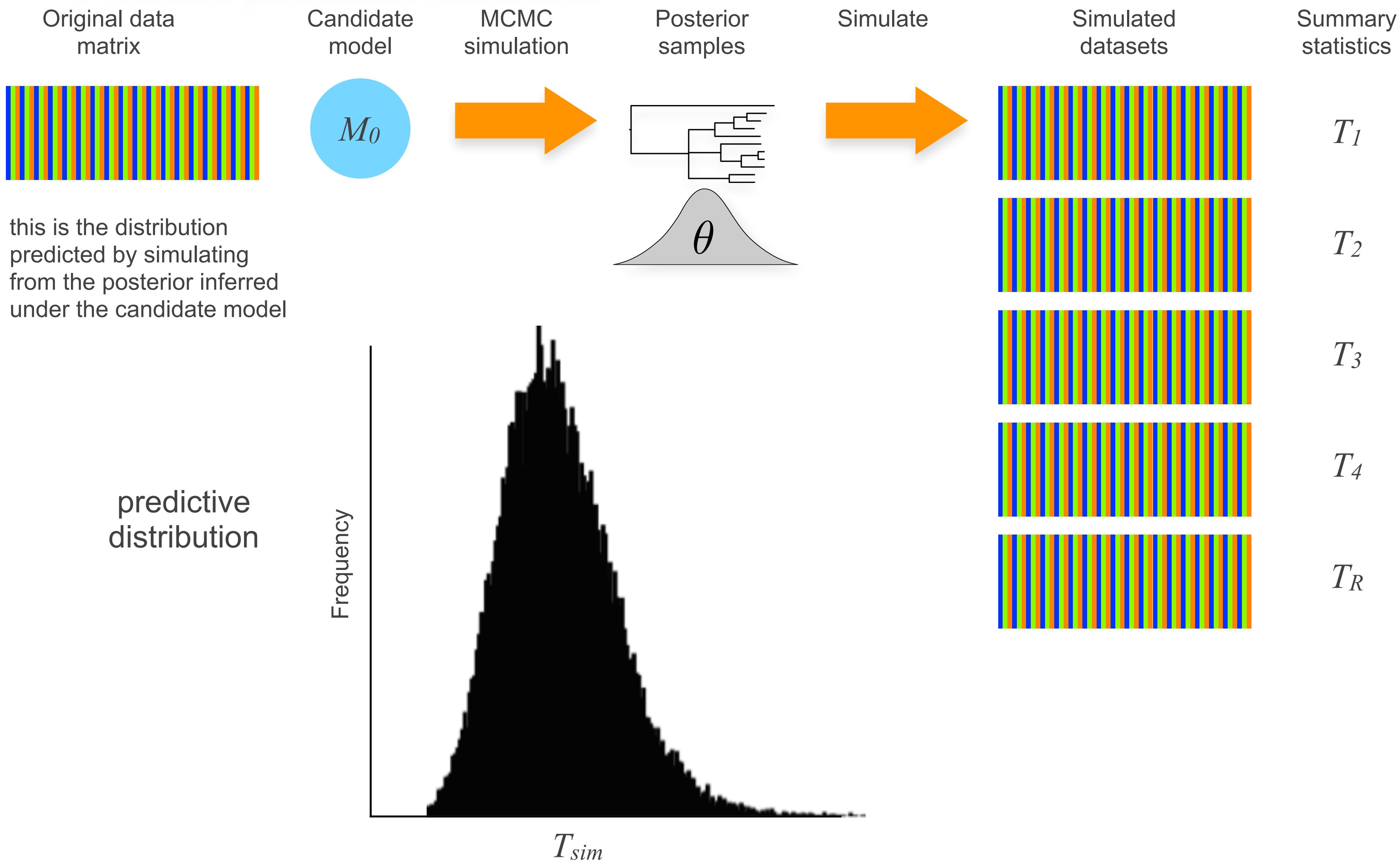
Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation



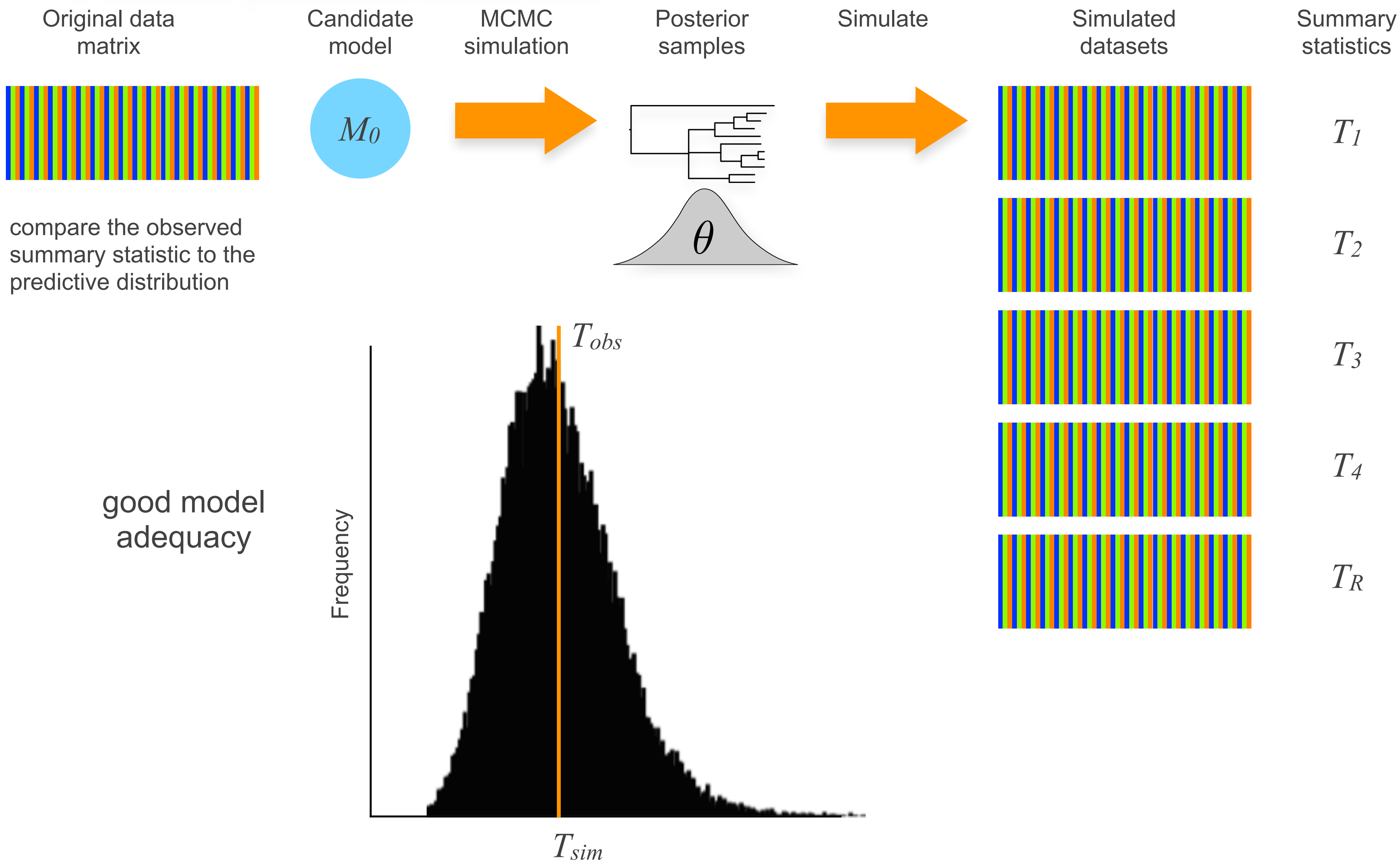
Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation



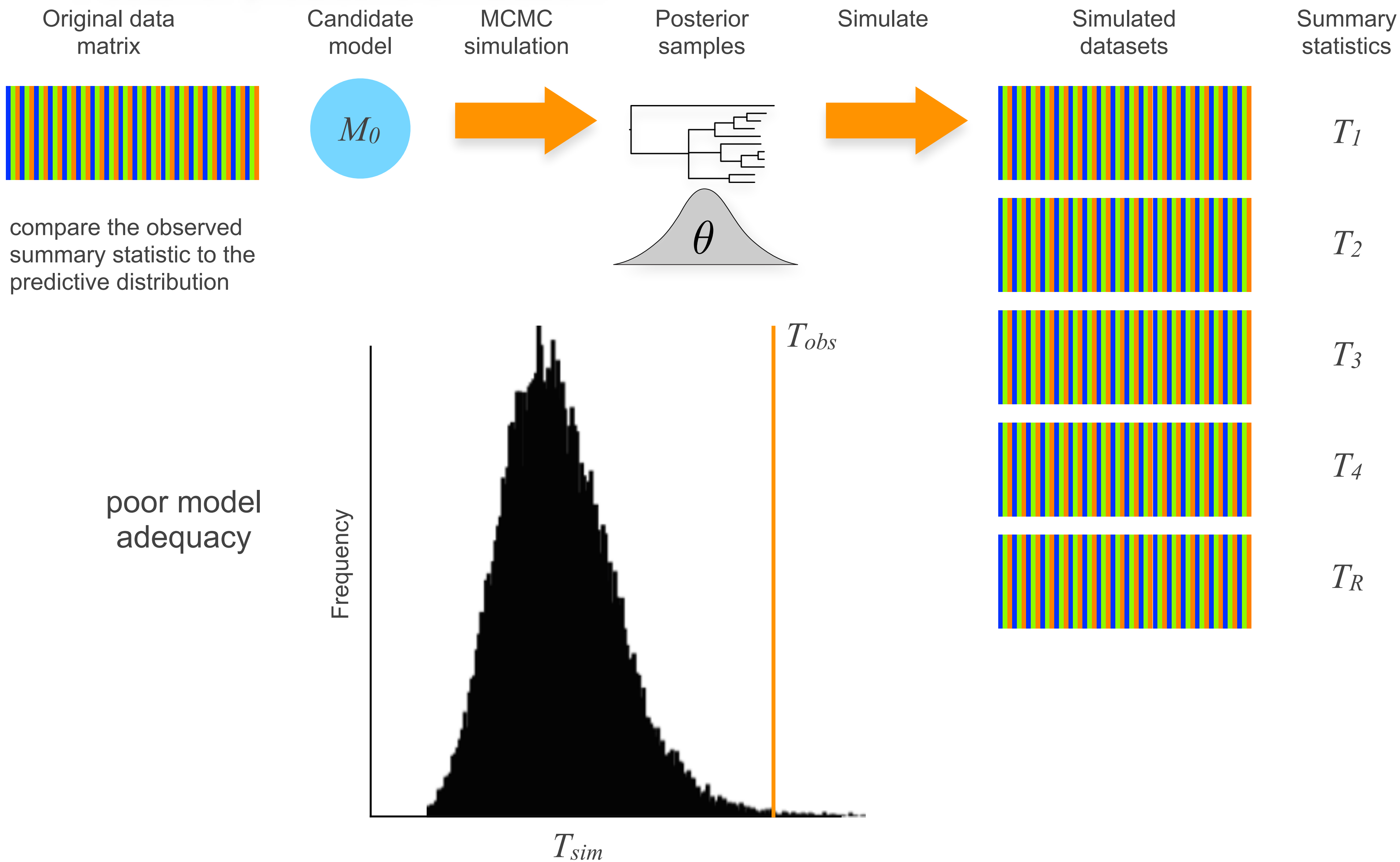
Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation



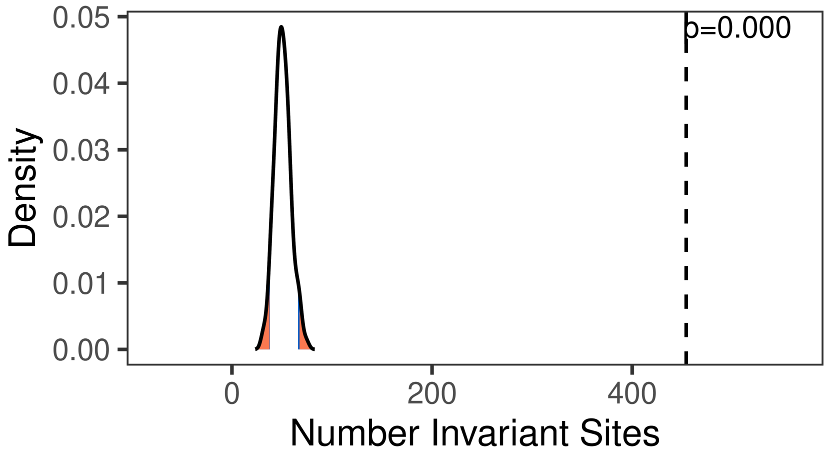
Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation



Bayesian Evaluation of Model Adequacy

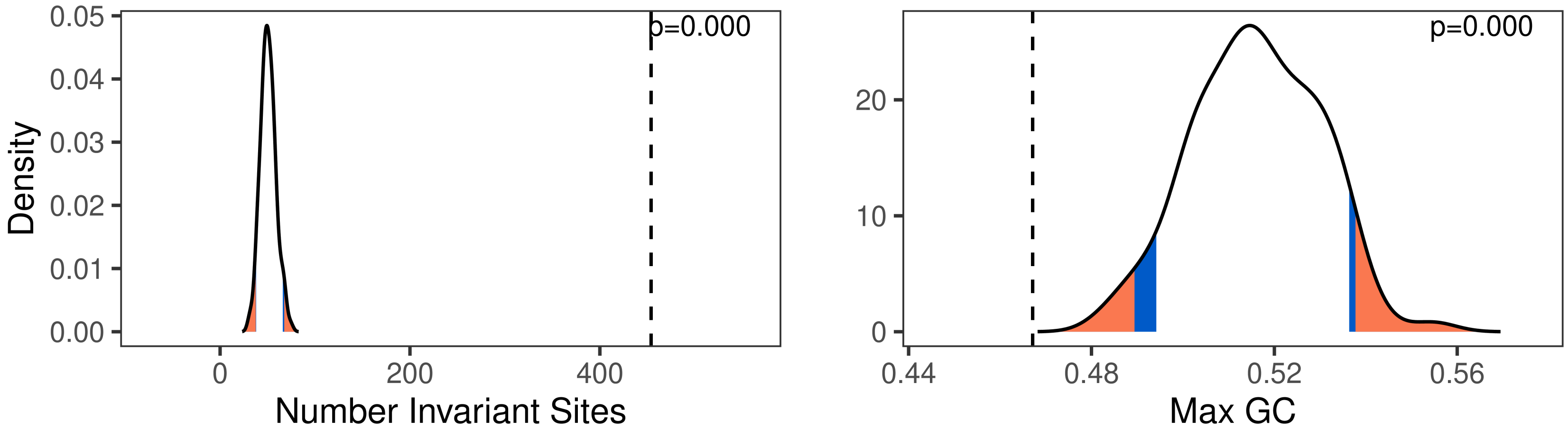
Posterior-predictive simulation



Höhna *et al.* (2018)

Bayesian Evaluation of Model Adequacy

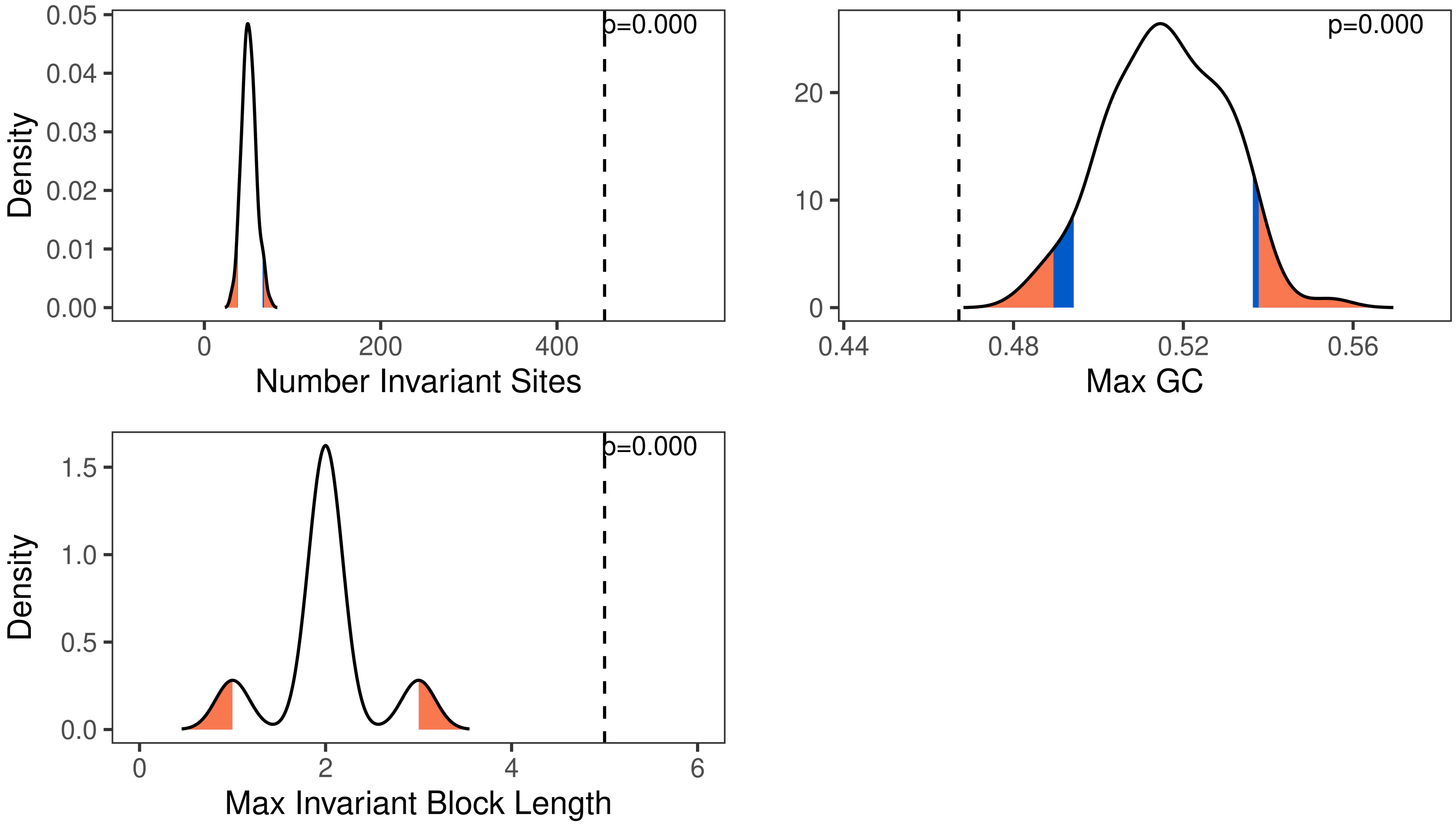
Posterior-predictive simulation



Höhna *et al.* (2018)

Bayesian Evaluation of Model Adequacy

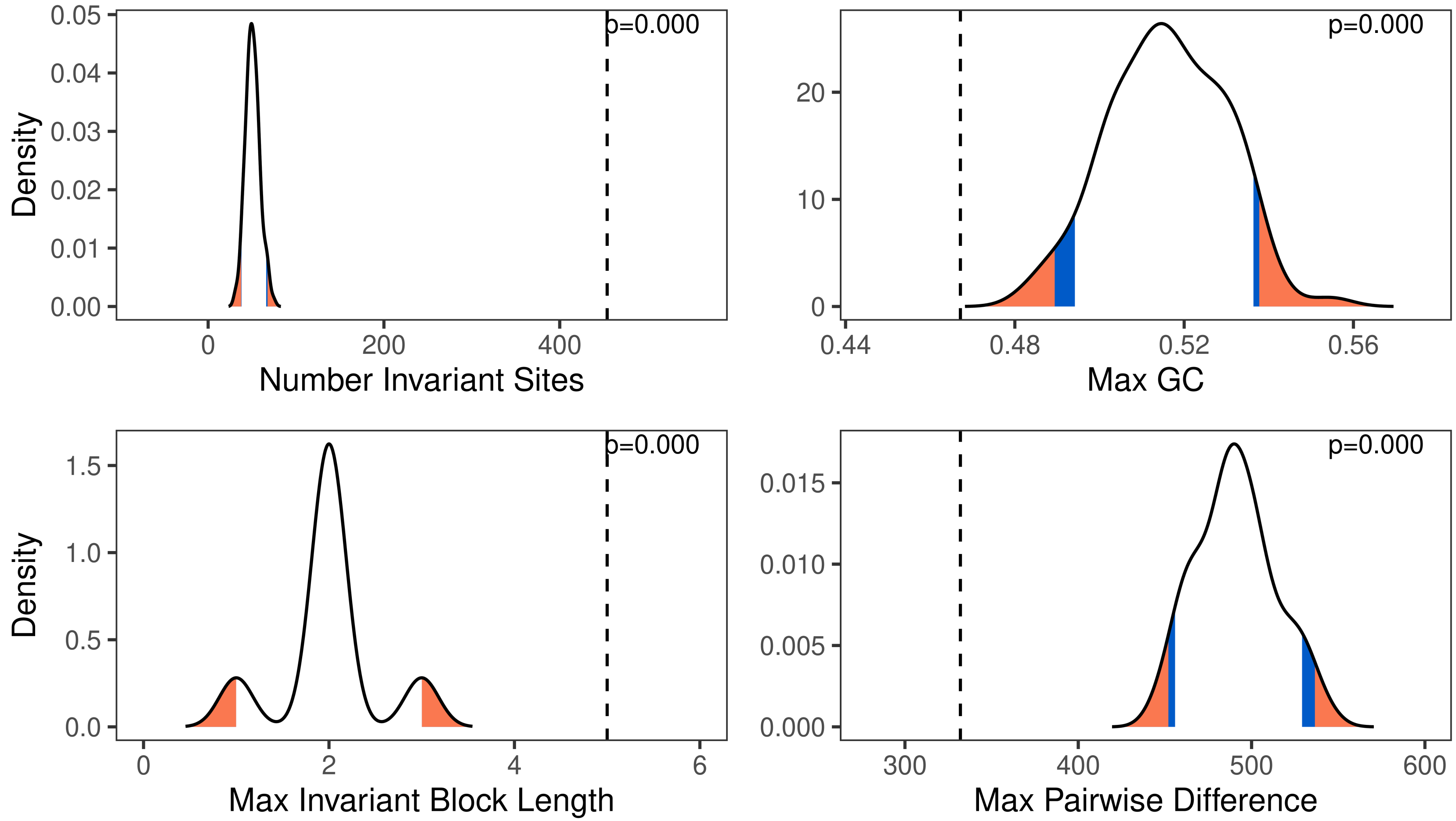
Posterior-predictive simulation



Höhna *et al.* (2018)

Bayesian Evaluation of Model Adequacy

Posterior-predictive simulation



Höhna *et al.* (2018)

Outline

I. Model selection

What is the <u>relative</u> fit of the candidate models (hypotheses) to our data?

ML and Bayesian methods for selecting among candidate models (hypotheses)

II. Model adequacy

What is the <u>absolute</u> fit of the candidate models (hypotheses) to our data?

Bayesian methods for assessing model adequacy of candidate models (hypotheses)

III. Model averaging

How do we accommodate uncertainty in the choice among candidate models?

Bayesian methods for averaging over candidate models (hypotheses)

Outline

I. Model selection

What is the <u>relative</u> fit of the candidate models (hypotheses) to our data?

ML and Bayesian methods for selecting among candidate models (hypotheses)

II. Model adequacy

What is the <u>absolute</u> fit of the candidate models (hypotheses) to our data?

Bayesian methods for assessing model adequacy of candidate models (hypotheses)

III. Model averaging

How do we accommodate uncertainty in the choice among candidate models?

Bayesian methods for averaging over candidate models (hypotheses)

Bayesian Model Selection

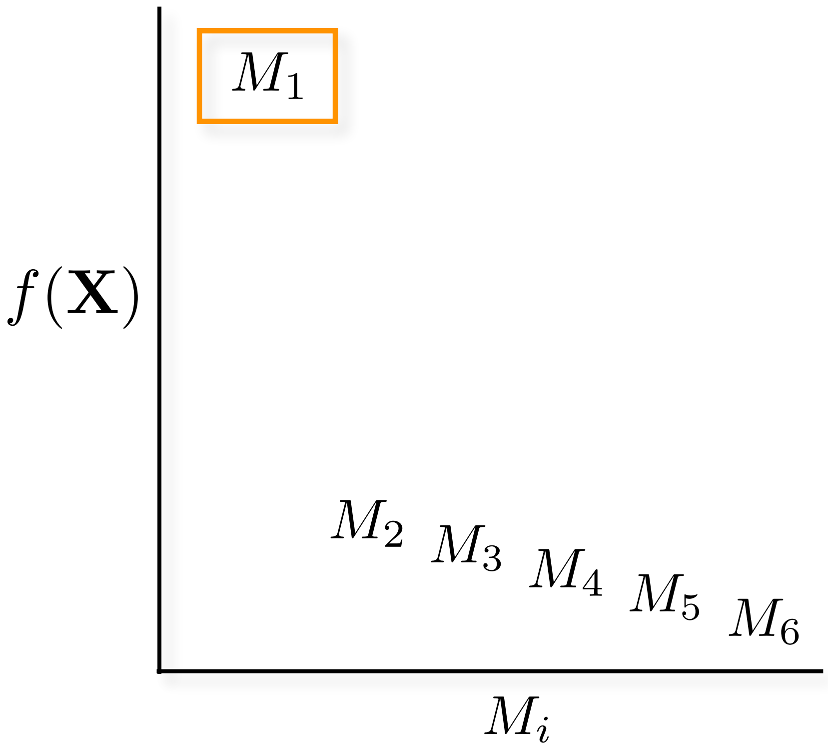
Even the best (and adequate) model might have (many) close competitors

It is possible that several (or indeed very many) mixed models provide comparable descriptions of the process that gave rise to the data.

Bayesian Model Selection

Even the best (and adequate) model might have (many) close competitors

It is possible that several (or indeed very many) mixed models provide comparable descriptions of the process that gave rise to the data.

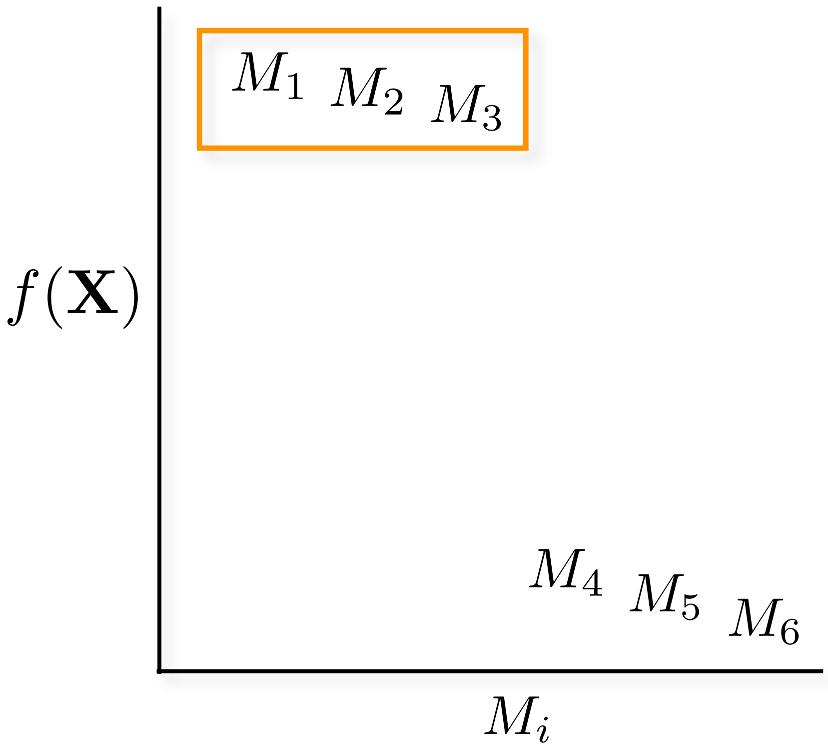


If the fit of the data to the selected mixed model is vastly better than other candidate models, conditioning on the best model may be justified.

Bayesian Model Selection

Even the best (and adequate) model might have (many) close competitors

It is possible that several (or indeed very many) mixed models provide comparable descriptions of the process that gave rise to the data.



Conversely, if the fit of the data to the selected model is only marginally better than other candidate models, it would be unwise to condition on the best model.

Bayesian Model Selection

Bayesian model averaging: reversible-jump MCMC (rjMCMC)

If we do not wish to condition the model (treat it as a fixed assumption), we could opt to treat the model itself as a *random variable*.

Bayesian Model Selection

Bayesian model averaging: reversible-jump MCMC (rjMCMC)

If we do not wish to condition the model (treat it as a fixed assumption), we could opt to treat the model itself as a *random variable*.

We do this using a numerical method that visits (jumps) between models

- within a model, it visits model parameter values in the normal way

Bayesian Model Selection

Bayesian model averaging: reversible-jump MCMC (rjMCMC)

If we do not wish to condition the model (treat it as a fixed assumption), we could opt to treat the model itself as a *random variable*.

We do this using a numerical method that visits (jumps) between models

- within a model, it visits model parameter values in the normal way
- the MCMC occasionally proposes jumps to a new model

Bayesian Model Selection

Bayesian model averaging: reversible-jump MCMC (rjMCMC)

If we do not wish to condition the model (treat it as a fixed assumption), we could opt to treat the model itself as a *random variable*.

We do this using a numerical method that visits (jumps) between models

- within a model, it visits model parameter values in the normal way
- the MCMC occasionally proposes jumps to a new model
- the models may differ in dimensionality, which must be accommodated:

$$R = \min\left[1, \underbrace{\boxed{\frac{f(\mathbf{X}|\theta')}{f(\mathbf{X}|\theta)}}}_{\text{likelihood ratio}} \cdot \underbrace{\boxed{\frac{f(\theta')}{f(\theta)}}}_{\text{prior ratio}} \cdot \underbrace{\boxed{\frac{f(\theta|\theta')}{f(\theta'|\theta)}}}_{\text{proposal ratio}}\right]$$

Bayesian Model Selection

Bayesian model averaging: reversible-jump MCMC (rjMCMC)

If we do not wish to condition the model (treat it as a fixed assumption), we could opt to treat the model itself as a *random variable*.

We do this using a numerical method that visits (jumps) between models

- within a model, it visits model parameter values in the normal way
- the MCMC occasionally proposes jumps to a new model
- the models may differ in dimensionality, which must be accommodated:

$$R = \min\left[1, \boxed{\frac{f(\mathbf{X}|\theta')}{f(\mathbf{X}|\theta)}} \cdot \boxed{\frac{f(\theta')}{f(\theta)}} \cdot \boxed{\frac{f(\theta|\theta')}{f(\theta'|\theta)}}\right]$$

likelihood ratio   prior ratio   proposal ratio

The parameter estimates are averaged over all possible models

Bayesian Model Selection

Bayesian model averaging: reversible-jump MCMC (rjMCMC)

If we do not wish to condition the model (treat it as a fixed assumption), we could opt to treat the model itself as a *random variable*.

We do this using a numerical method that visits (jumps) between models

- within a model, it visits model parameter values in the normal way
- the MCMC occasionally proposes jumps to a new model
- the models may differ in dimensionality, which must be accommodated:

$$R = \min\left[1, \boxed{\frac{f(\mathbf{X}|\theta')}{f(\mathbf{X}|\theta)}} \cdot \boxed{\frac{f(\theta')}{f(\theta)}} \cdot \boxed{\frac{f(\theta|\theta')}{f(\theta'|\theta)}}\right]$$

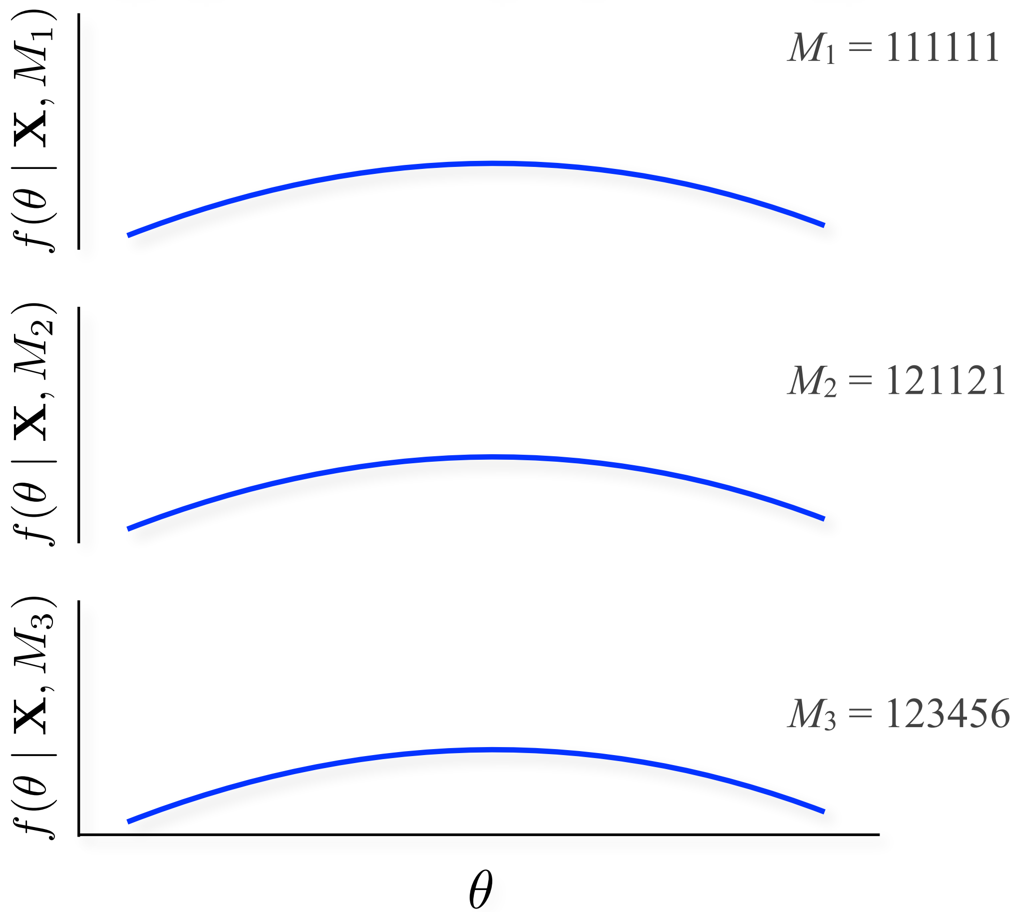
likelihood ratio prior ratio proposal ratio

The parameter estimates are averaged over all possible models

The proportion of time spent on each model is an estimate of its posterior probability

Bayesian Model Selection

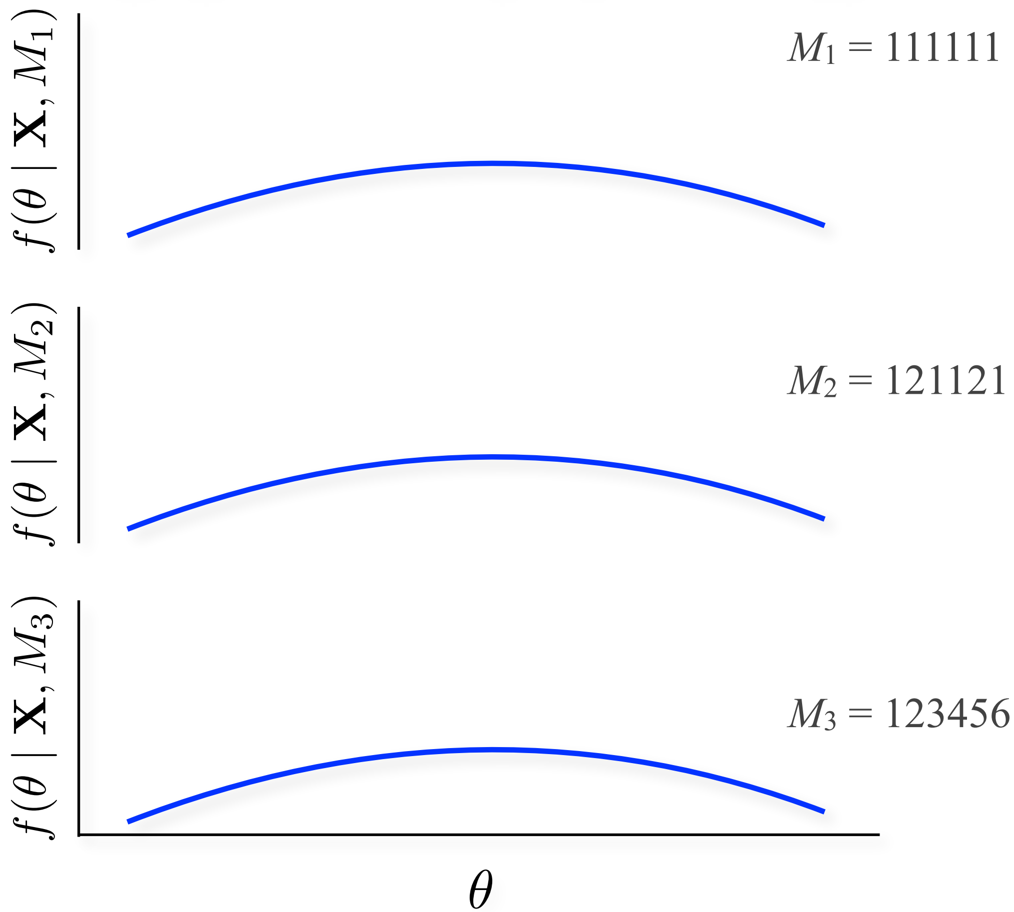
Bayesian model averaging: reversible-jump MCMC (rjMCMC)



We infer the joint posterior probability distribution by averaging over a set of models, M_i.

Bayesian Model Selection

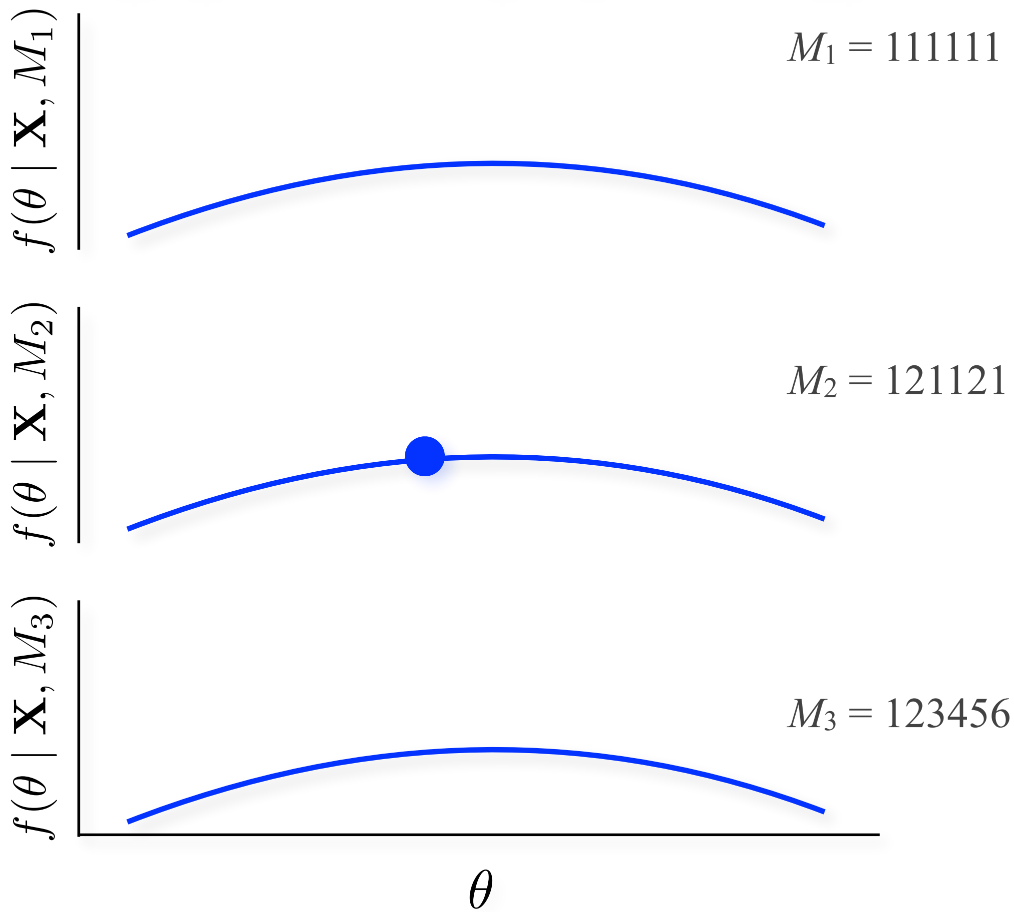
Bayesian model averaging: reversible-jump MCMC (rjMCMC)



Like other variables, we specify the prior probability distribution for the models.

Bayesian Model Selection

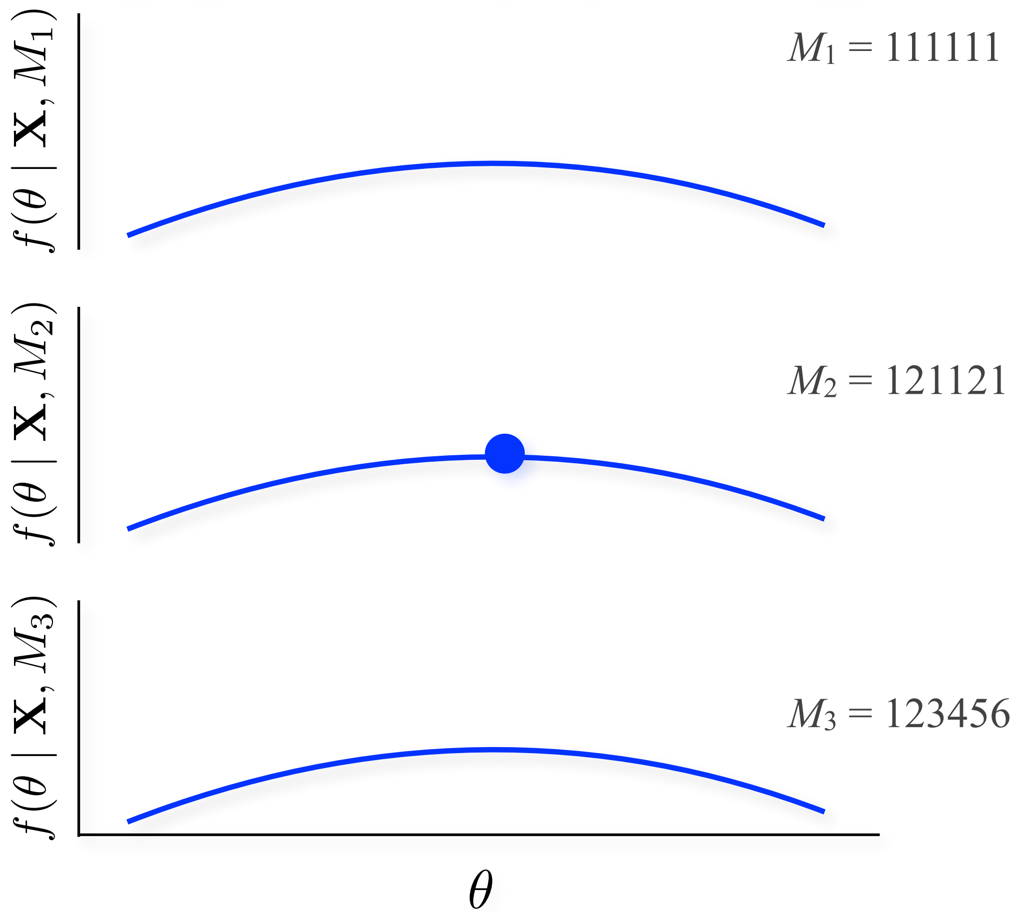
Bayesian model averaging: reversible-jump MCMC (rjMCMC)



Within a model, we sample parameter values in the normal way.

Bayesian Model Selection

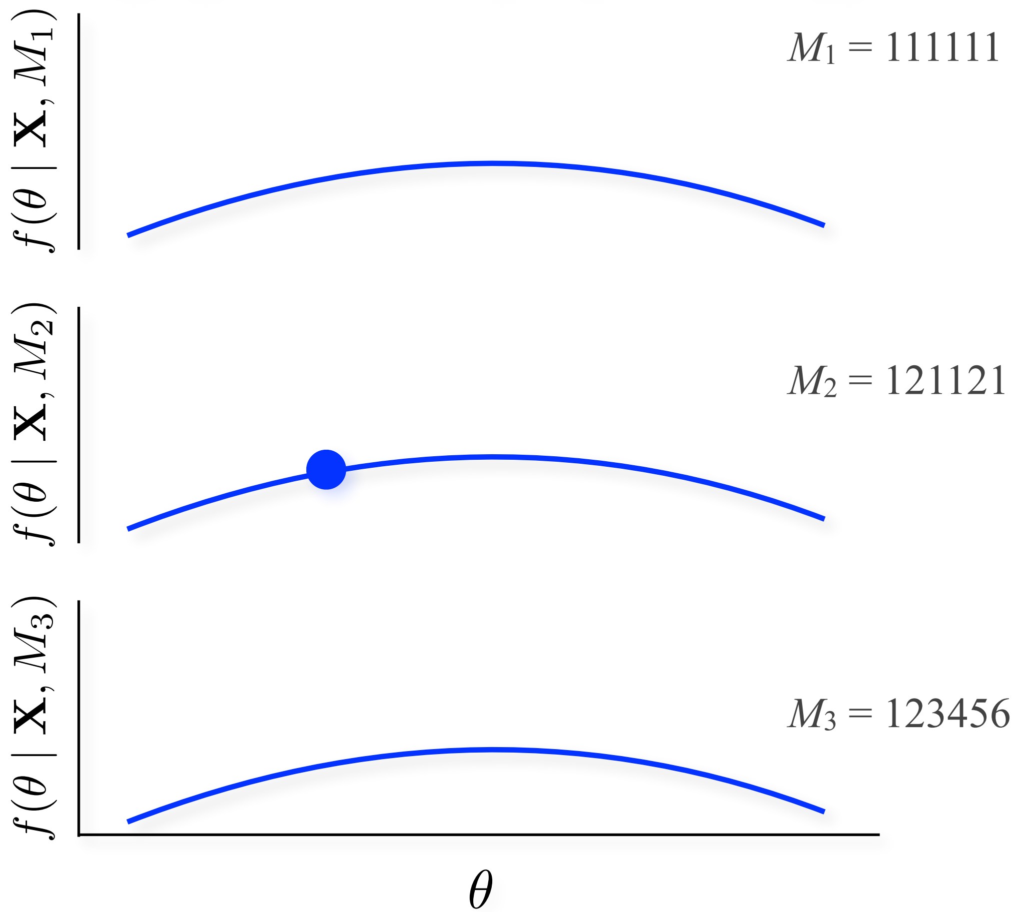
Bayesian model averaging: reversible-jump MCMC (rjMCMC)



Within a model, we sample parameter values in the normal way.

Bayesian Model Selection

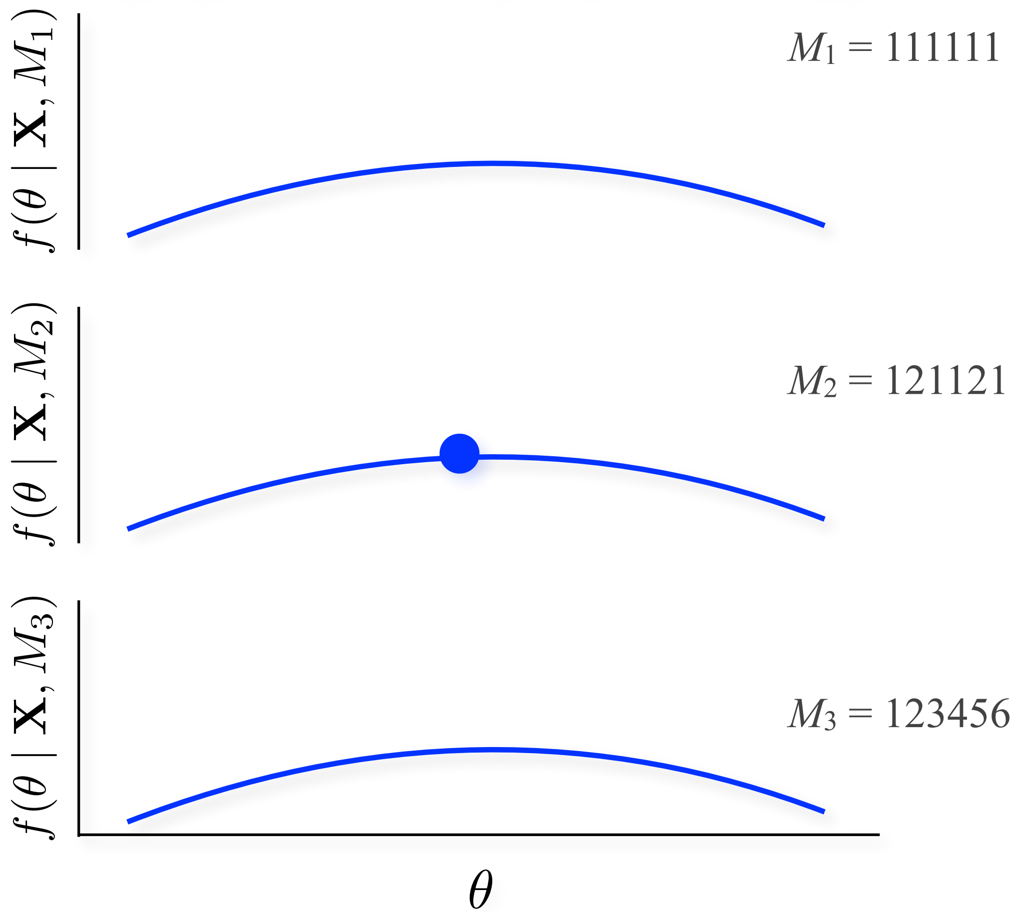
Bayesian model averaging: reversible-jump MCMC (rjMCMC)



Within a model, we sample parameter values in the normal way.

Bayesian Model Selection

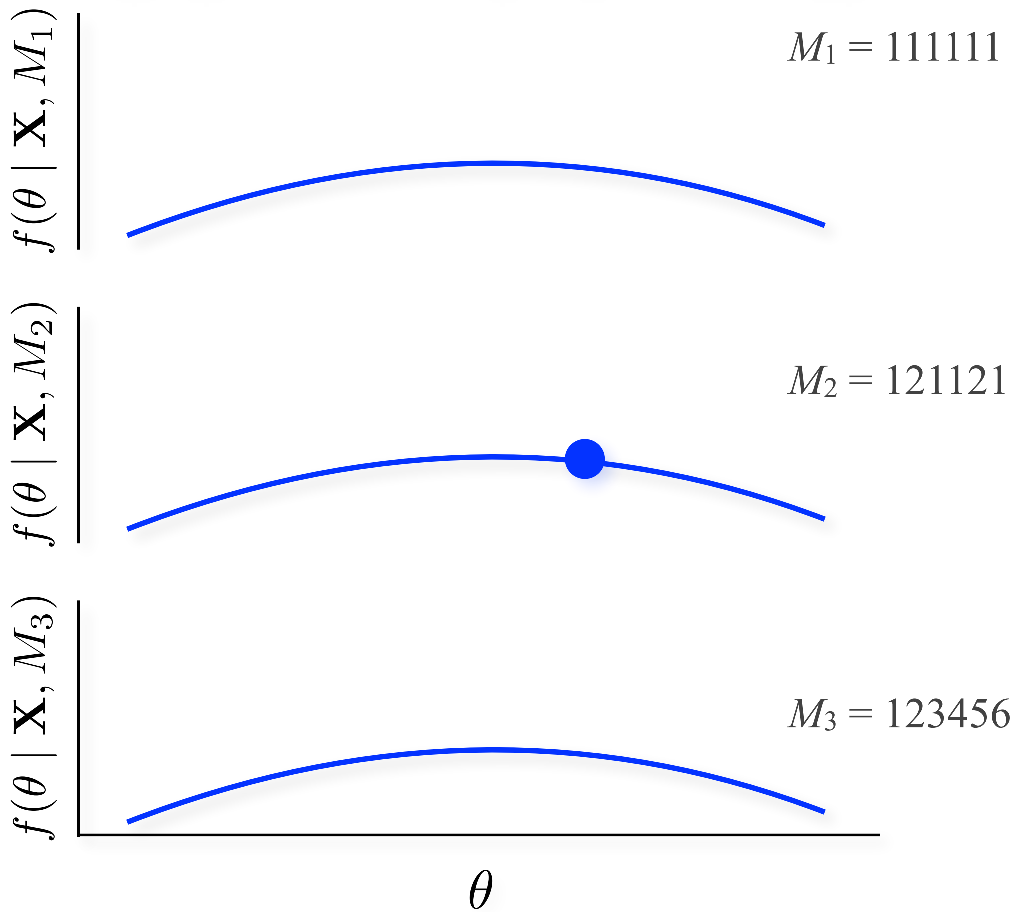
Bayesian model averaging: reversible-jump MCMC (rjMCMC)



Within a model, we sample parameter values in the normal way.

Bayesian Model Selection

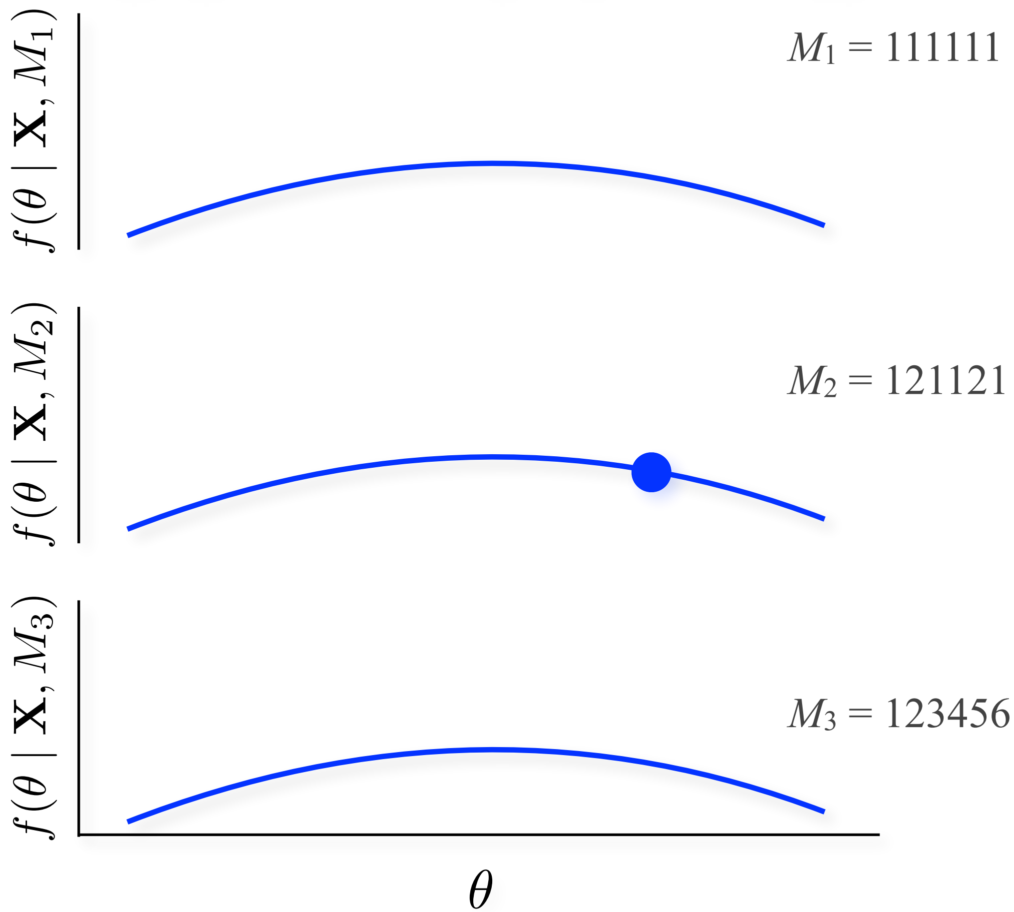
Bayesian model averaging: reversible-jump MCMC (rjMCMC)



Within a model, we sample parameter values in the normal way.

Bayesian Model Selection

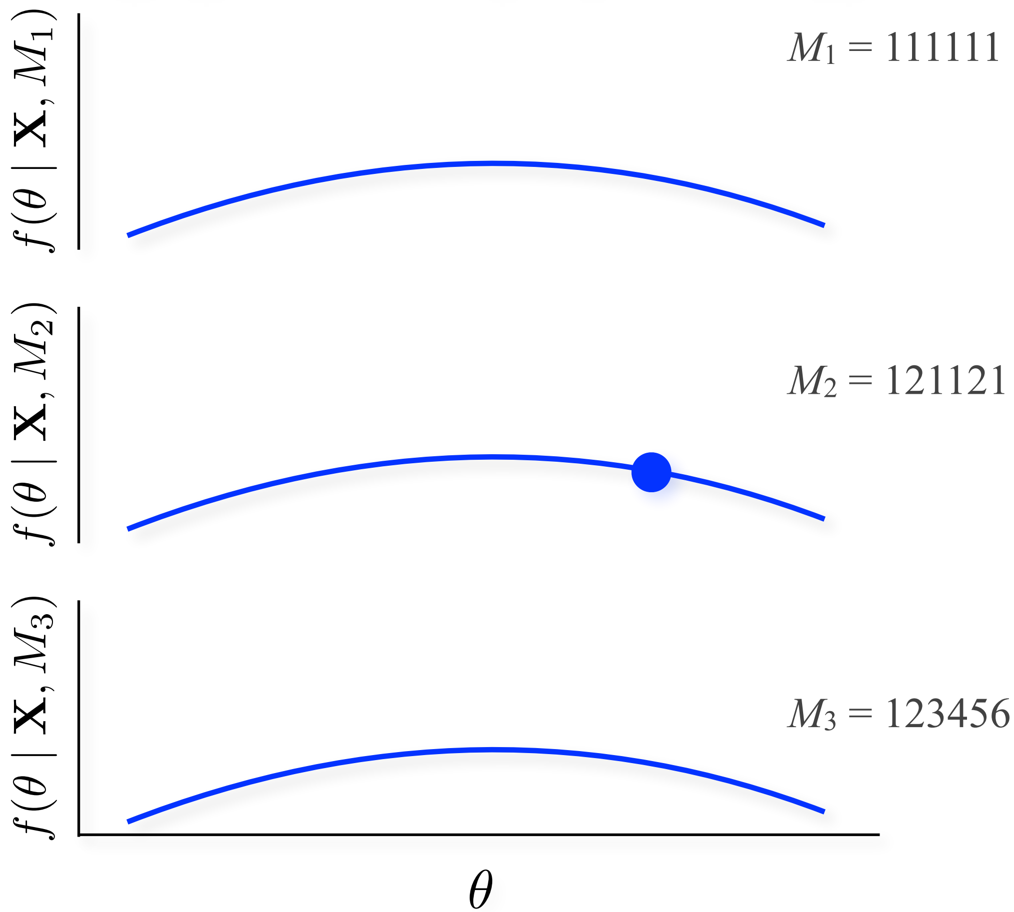
Bayesian model averaging: reversible-jump MCMC (rjMCMC)



Within a model, we sample parameter values in the normal way.

Bayesian Model Selection

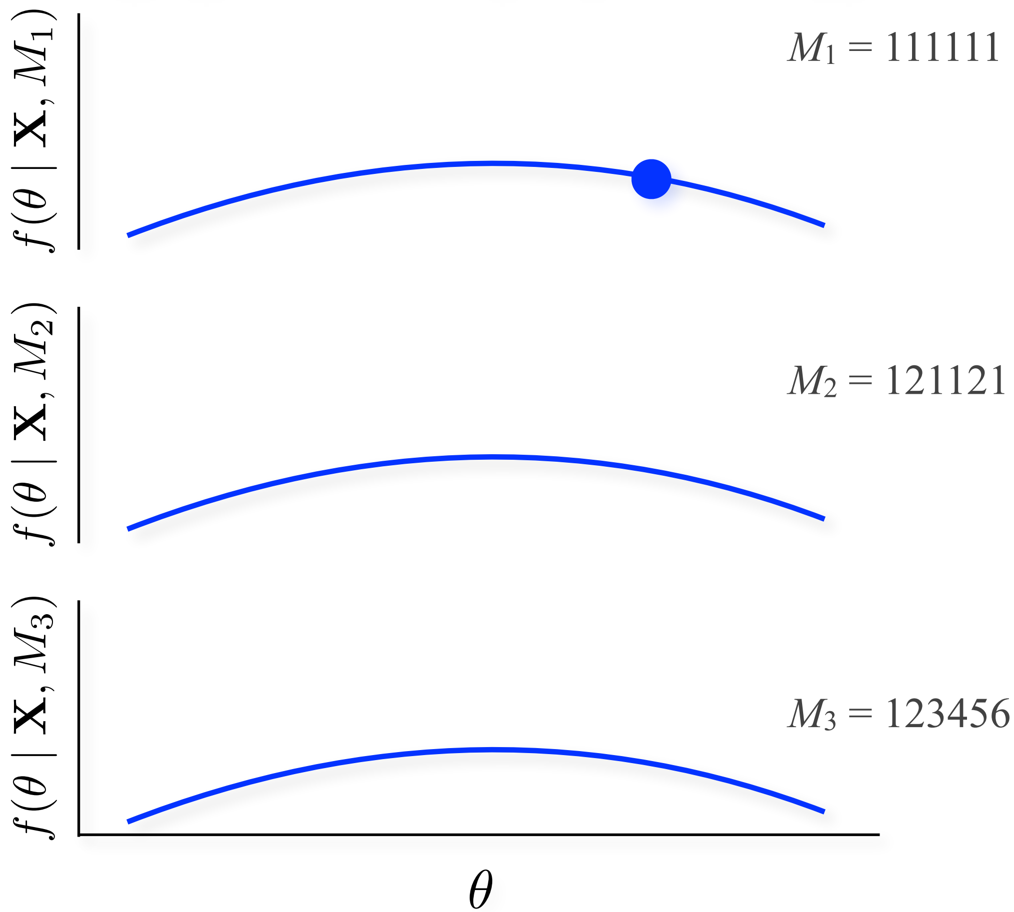
Bayesian model averaging: reversible-jump MCMC (rjMCMC)



Occasionally, we propose a move to a new model

Bayesian Model Selection

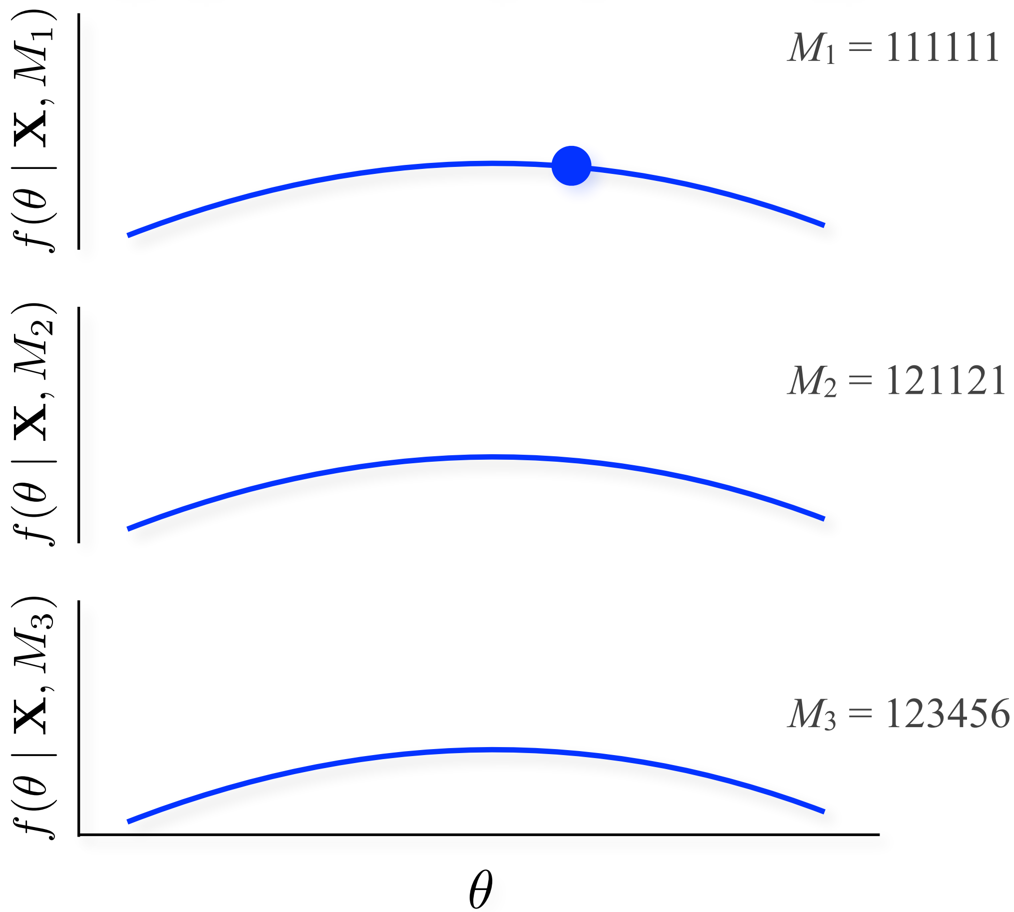
Bayesian model averaging: reversible-jump MCMC (rjMCMC)



Occasionally, we propose a move to a new model, and accept/reject in the usual way.

Bayesian Model Selection

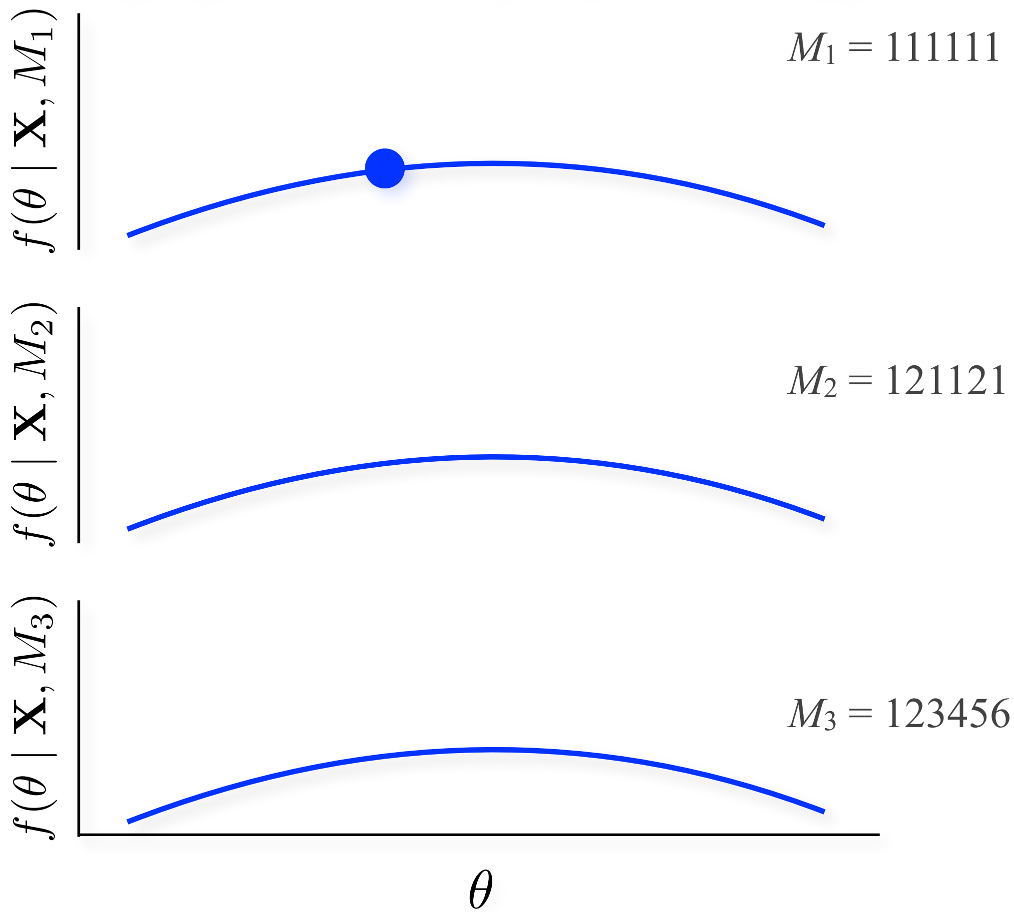
Bayesian model averaging: reversible-jump MCMC (rjMCMC)



Occasionally, we propose a move to a new model, and accept/reject in the usual way.

Bayesian Model Selection

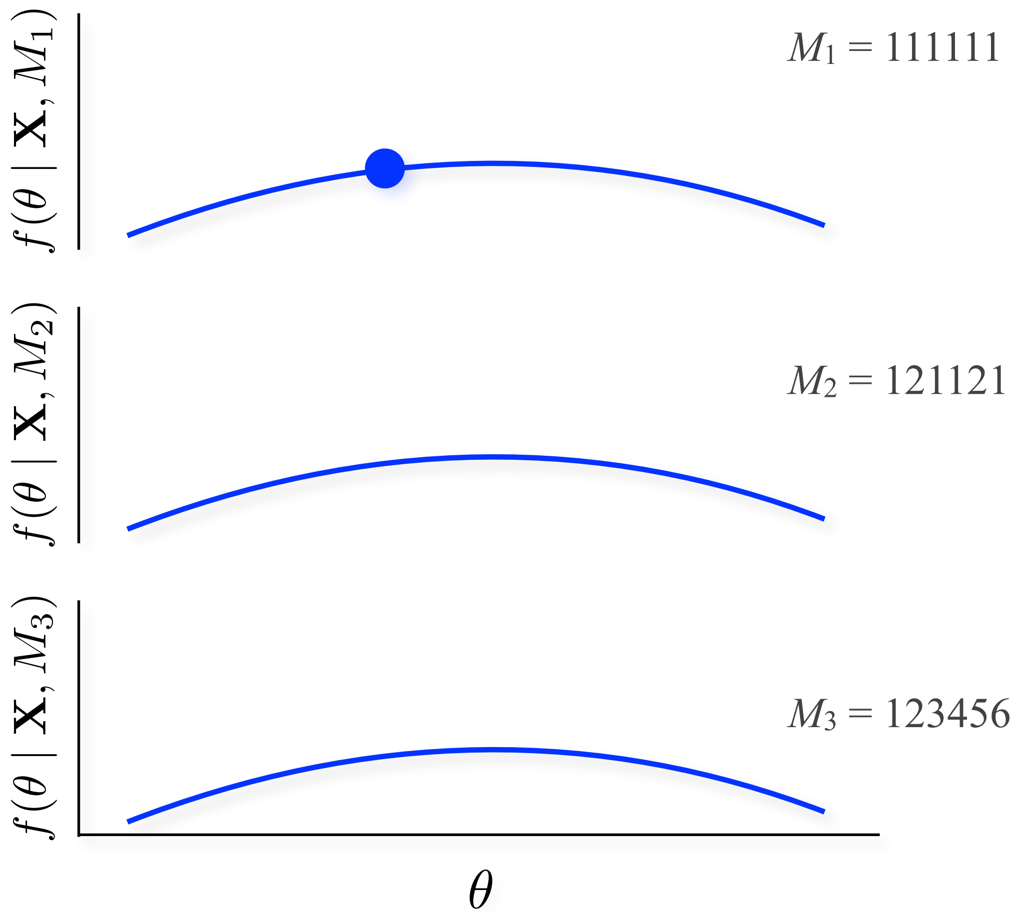
Bayesian model averaging: reversible-jump MCMC (rjMCMC)



Occasionally, we propose a move to a new model, and accept/reject in the usual way.

Bayesian Model Selection

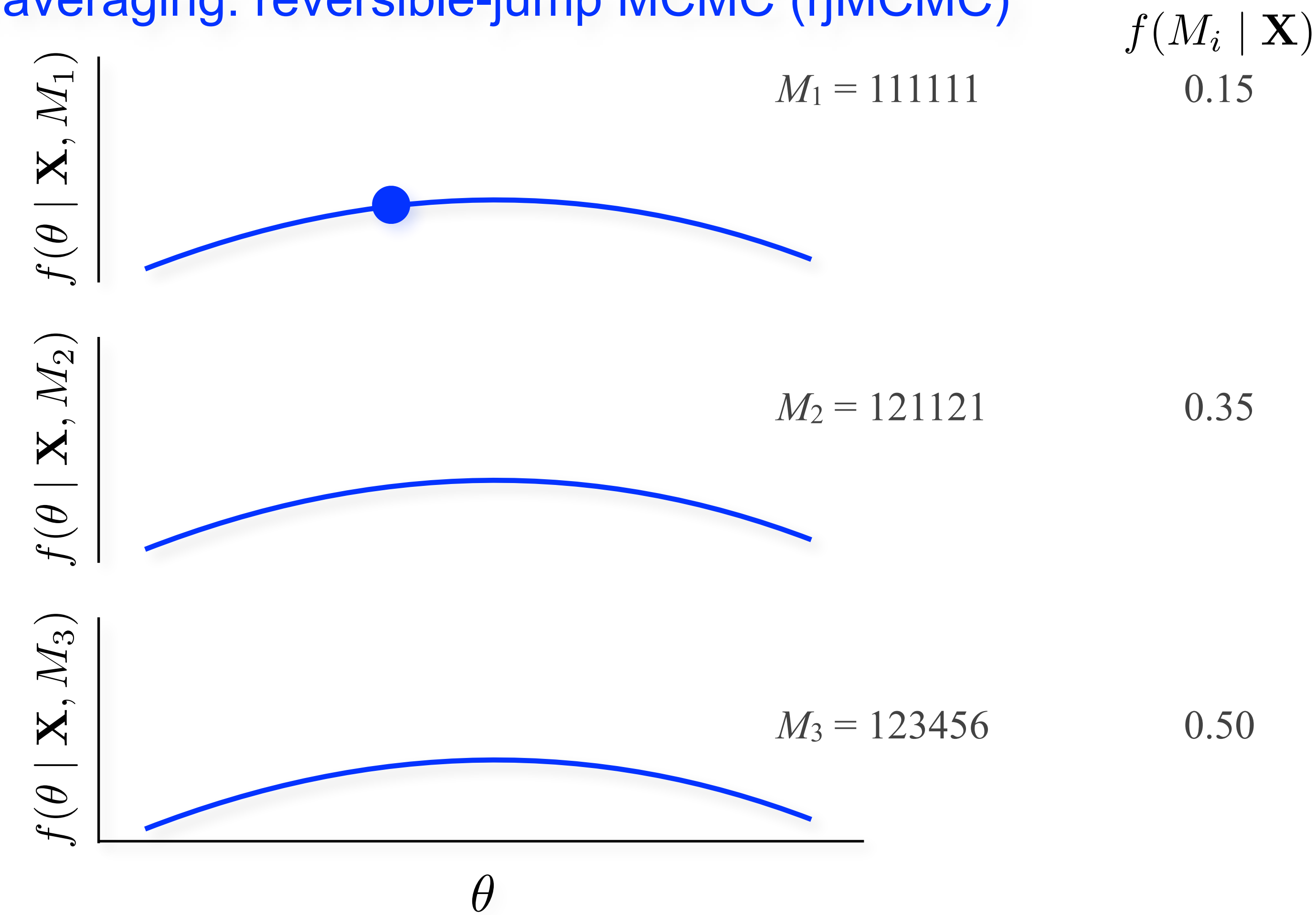
Bayesian model averaging: reversible-jump MCMC (rjMCMC)



Parameter estimates for all parameters are therefore averaged over all models.

Bayesian Model Selection

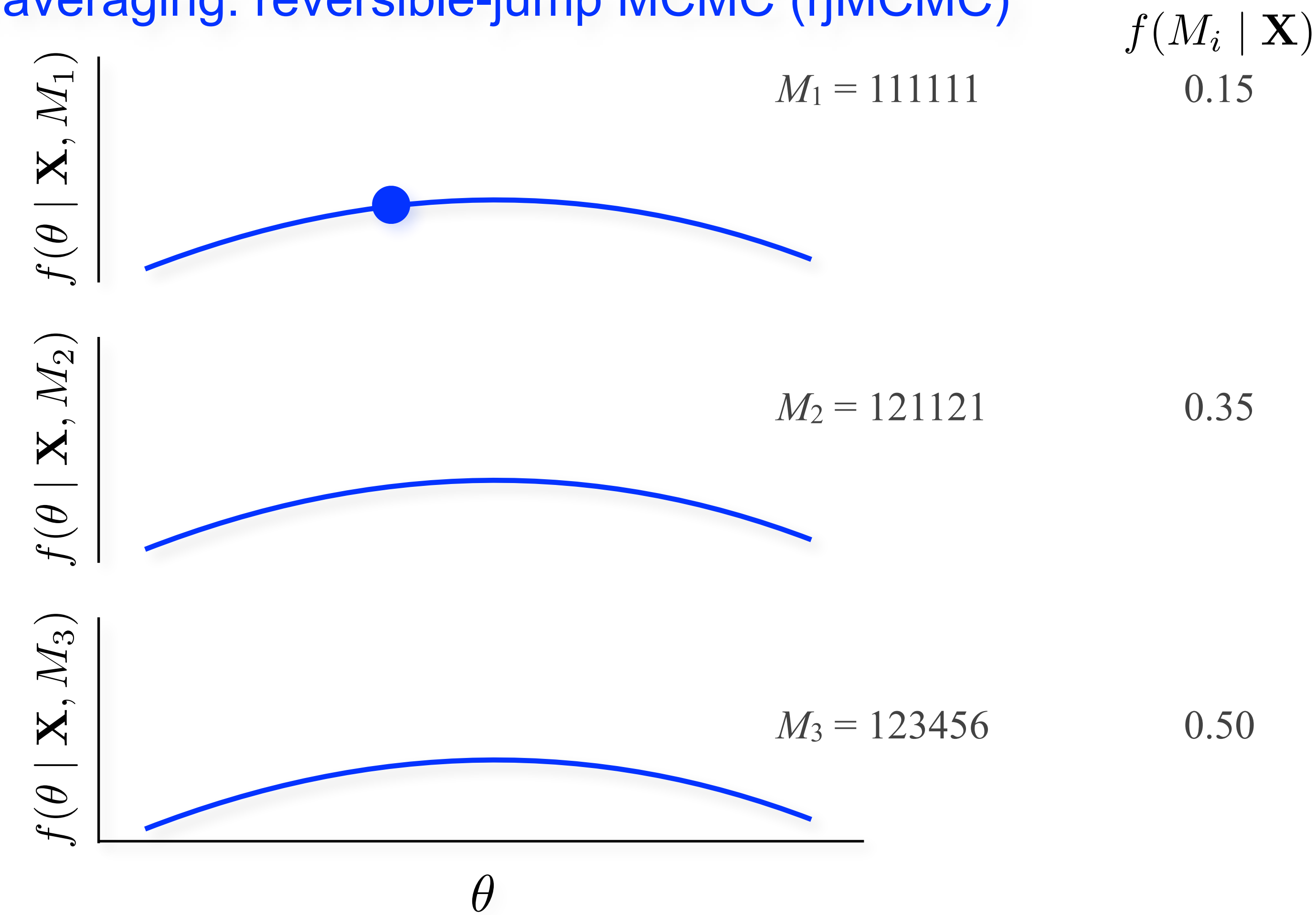
Bayesian model averaging: reversible-jump MCMC (rjMCMC)



The proportion of time the jMCMC spends visiting each model is an estimate of its marginal posterior probability.

Bayesian Model Selection

Bayesian model averaging: reversible-jump MCMC (rjMCMC)



The proportion of time the jMCMC spends visiting each model is an estimate of its marginal posterior probability, so we get model selection for free!

Model-Based Inference of Phylogeny

Model-based inference is based on the model

We have to assess our ability to estimate parameters of a given model:

- likelihood optimization: have we found the global MLE?
- MCMC simulation: have we found accurately approximated the posterior?

Model-Based Inference of Phylogeny

Model-based inference is based on the model

We have to assess our ability to estimate parameters of a given model:

- likelihood optimization: have we found the global MLE?
- MCMC simulation: have we found accurately approximated the posterior?

We have to be careful about our choice of model:

- model selection: what is the <u>relative</u> fit of candidate models to my dataset?
- model adequacy: what is the <u>absolute</u> fit of a candidate model to my dataset?
- model uncertainty/averaging: how to deal with <u>uncertainty</u> in the choice of model?

Model-Based Inference of Phylogeny

Model-based inference is based on the model

We have to assess our ability to estimate parameters of a given model:

- likelihood optimization: have we found the global MLE?
- MCMC simulation: have we found accurately approximated the posterior?

We have to be careful about our choice of model:

- model selection: what is the <u>relative</u> fit of candidate models to my dataset?
- model adequacy: what is the <u>absolute</u> fit of a candidate model to my dataset?
- model uncertainty/averaging: how to deal with <u>uncertainty</u> in the choice of model?

Comparing the (relative or absolute) fit of alternative models is how we learn

- we can test competing hypotheses by comparing the fit of competing models to our data
- we can learn what parameters are important to describe the process that gave rise to our data
- we can simultaneously improve our estimates of phylogenies and also our understanding of the factors impacting molecular evolution