Bayesian Phylogenetic Inference using RevBayes:

Model Selection

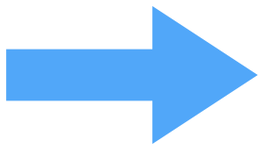
Sebastian Höhna

Division of Evolutionary Biology
Ludwig-Maximilians Universität, München

Model Testing

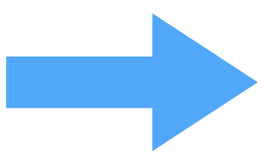
Two main purposes:

a) Select the best model from a set of candidate models:



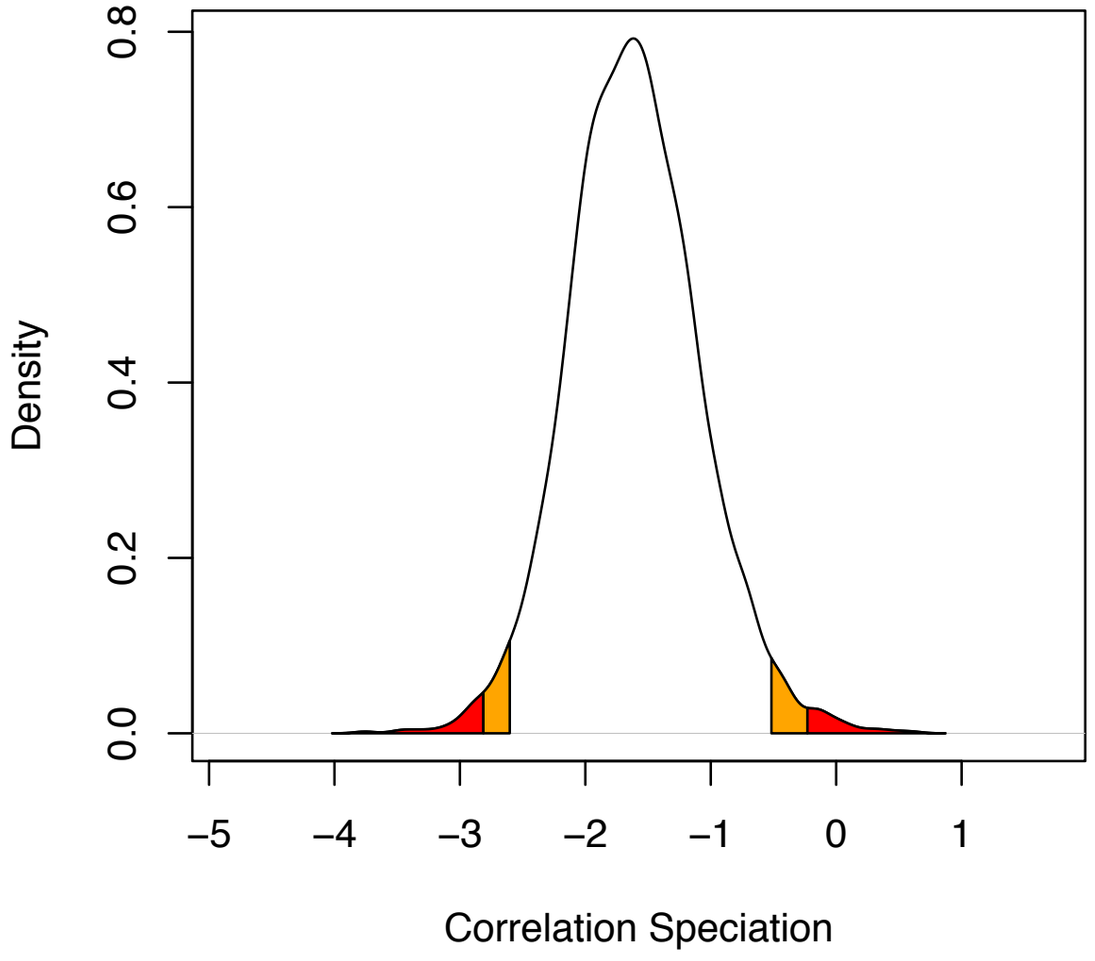
Model selection and hypothesis testing.

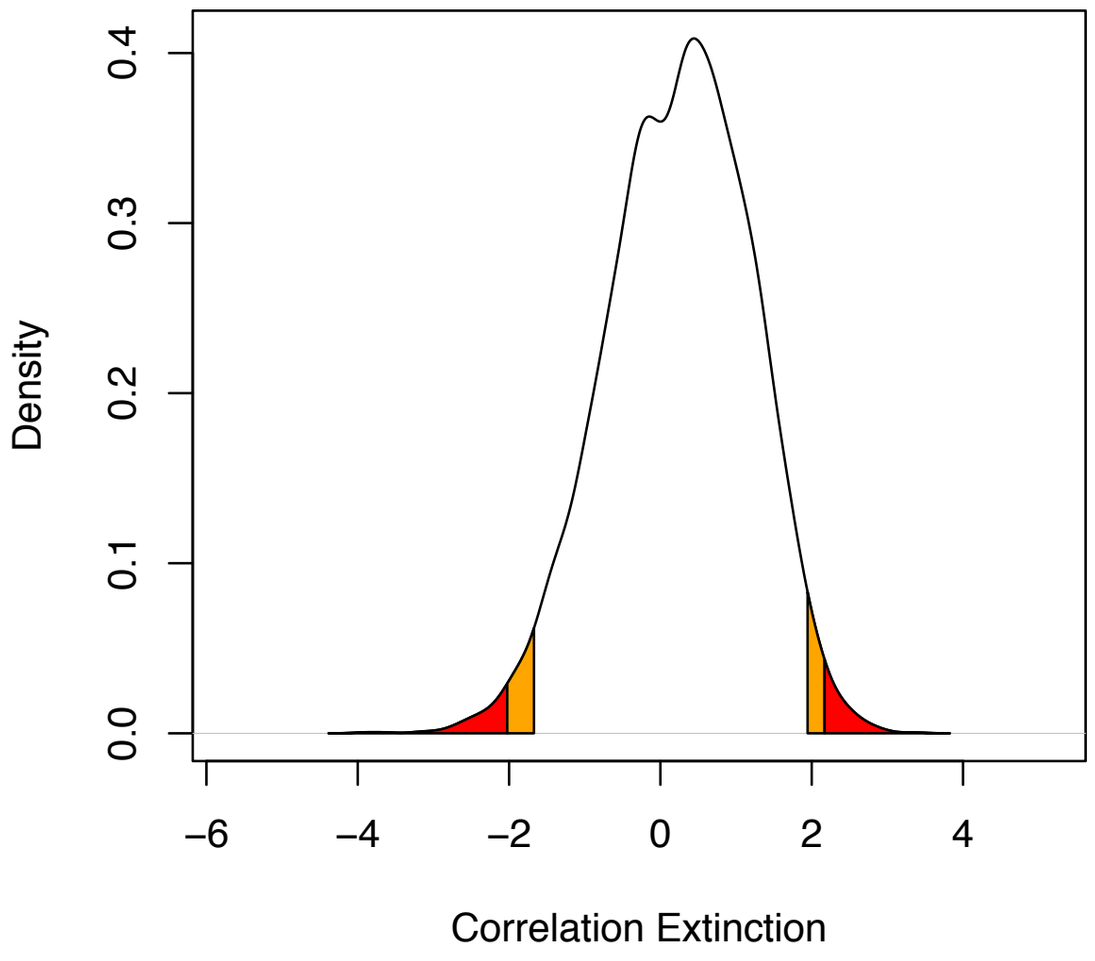
b) How good does a model fit to the data:



Model adequacy testing.

Model Selection: Example 1

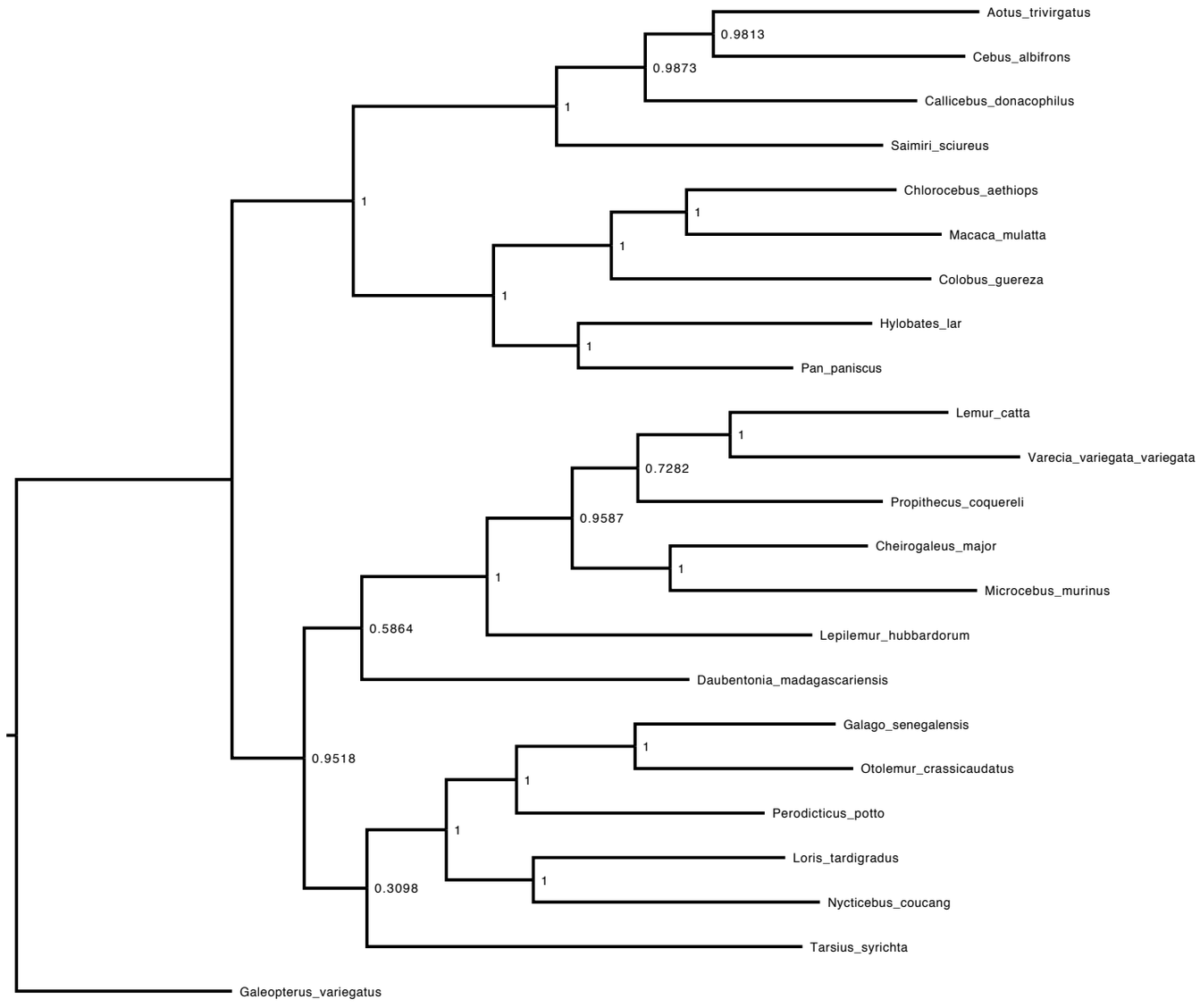




We want to test if diversification rates are correlated to environment CO_2.

Hypthesis testing

Model Selection: Example 2

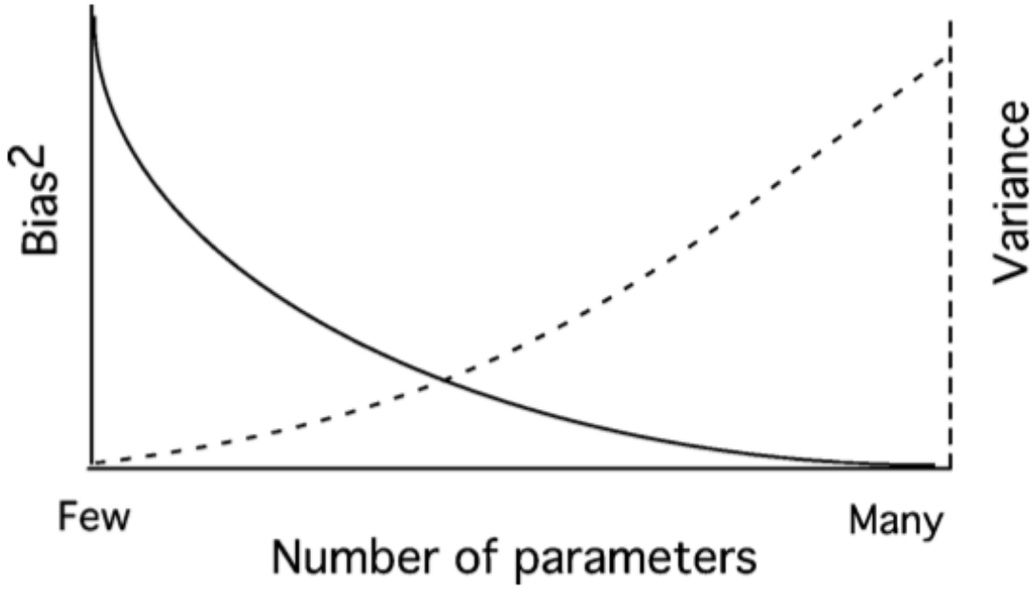


We want to estimate a phylogeny; which substitution model to use?

Model Specification Issues

Model selection, adequacy, and related issues

The model is central to parametric estimation of phylogeny: an under parameterized model will cause estimates to be biased (e.g., under estimation of branch lengths, topological error, inflated estimates of nodal support...); however, an over parameterized model will inflate estimation error (error variance, etc.).

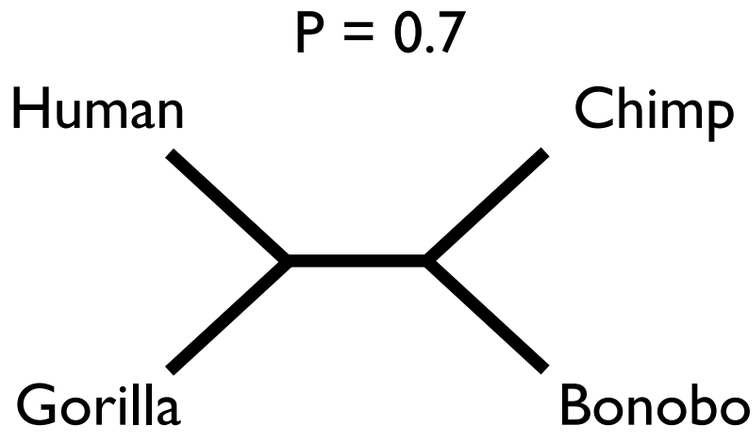


Outline

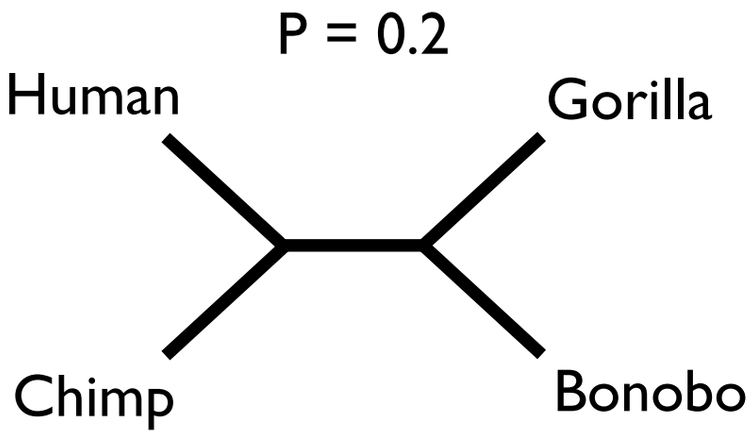
- Model Selection:
  - Frequentist Inference:
    - Likelihood Ratio Tests
    - AIC (Akaike's Information Criterion)
    - BIC (Bayesian Information Criterion)
  - Bayesian Inference:
    - Bayes Factors
    - Computing Marginal Likelihoods

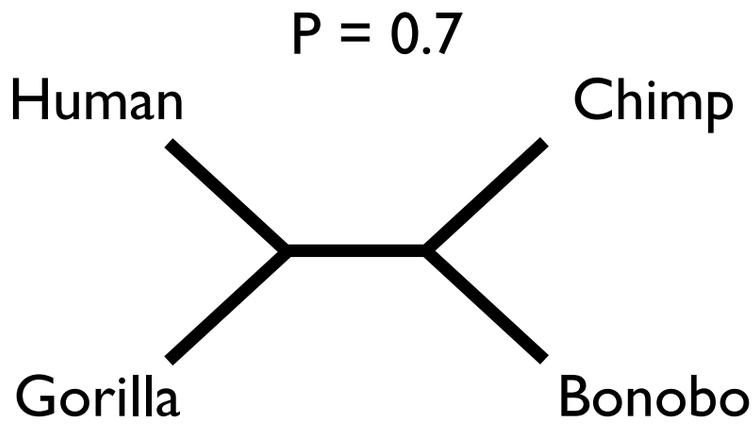
- Model Adequacy Testing:
  - Posterior Predictive Testing

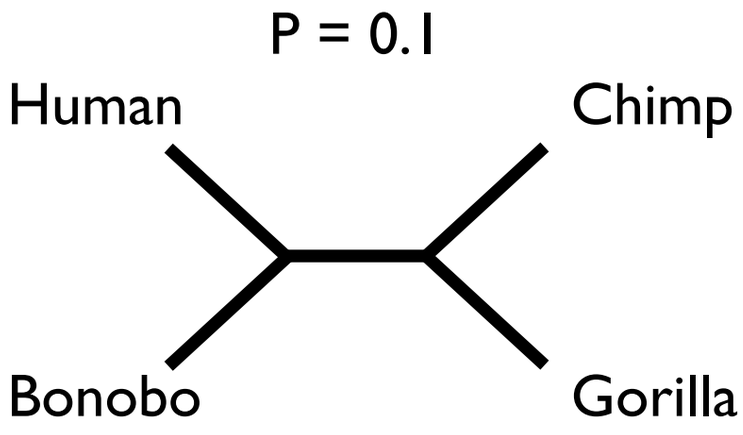
Model 1: Constrained (Chimp - Bonobo)



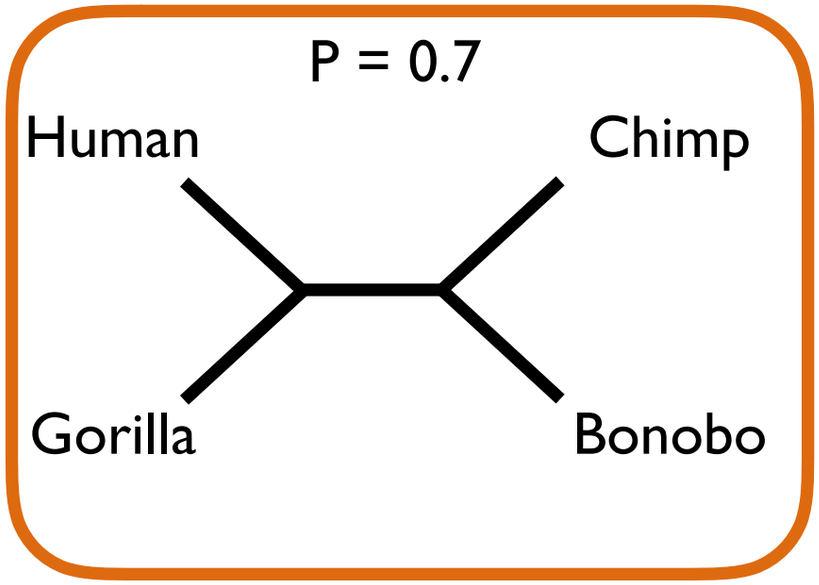
Model 2: Unconstrained







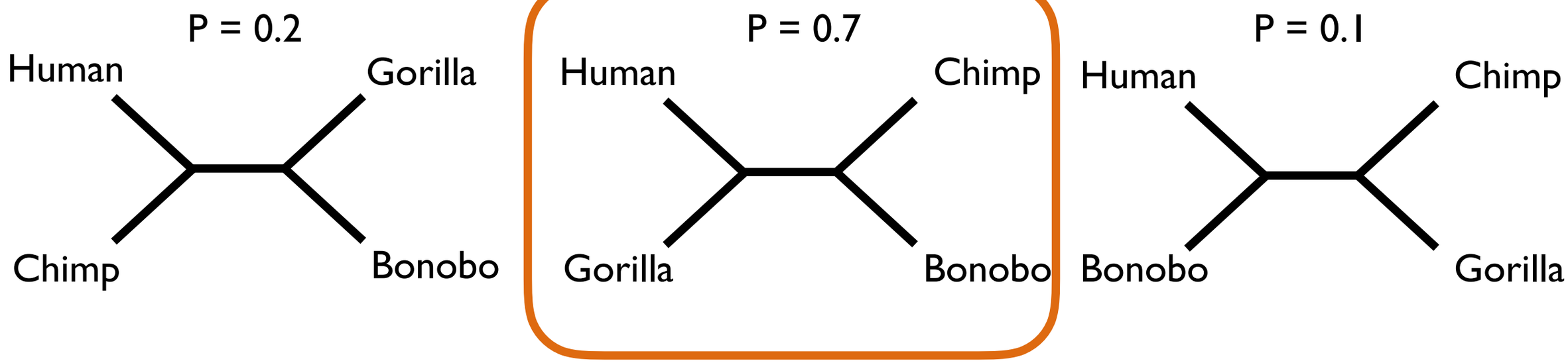
Model 1: Constrained (Chimp - Bonobo)



Maximum Likelihood

Model 2: Unconstrained

Estimates



Maximum Likelihood Model Selection

1. Likelihood Ratio Test

Compare the ratio of maximum likelihood scores under a null (restricted) model and an alternative (more general) nested model

$$\Delta = 2(\ln L_1 - \ln L_0)$$

The statistic is (approximately) distributed as a Chi-square random variable with df equal to the difference in the number of free parameters in the two nested models.

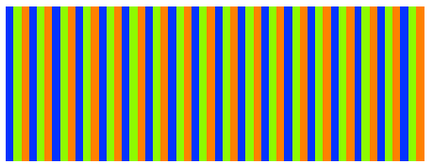
The statistic is essentially testing hypotheses about the data with respect to the alternate parameterizations under the two nested models.

Maximum Likelihood Model Selection

1. Likelihood Ratio Test

Parametric bootstrapping (Monte Carlo simulation)

Original data matrix



Candidate models

M_0

M_1

ML estimation

ML estimates

$\ln L_0$

$\ln L_1$

$hLRT_{obs}$

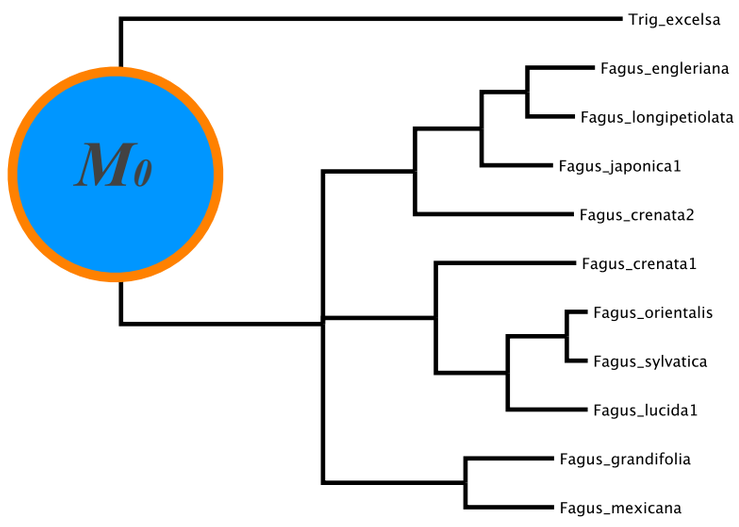
$$\Delta = 2(\ln L_1 - \ln L_0)$$

Maximum Likelihood Model Selection

1. Likelihood Ratio Test

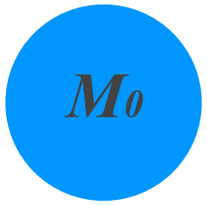
Parametric bootstrapping (Monte Carlo simulation)

ML estimates



Simulated data matrix

Candidate models



ML estimation

ML estimates

$\ln L_0$

$\ln L_1$

hLRT$_{sim}$

$$\Delta = 2(\ln L_1 - \ln L_0)$$

Reject null model
Fail to reject null model

Maximum Likelihood Model Selection

2. Akaike Information Criterion (AIC)

Estimates the expected Kullback-Leibler information distance between a given model and the true, generating model (so smaller scores are better).

$$AIC_i = -2\ln L_i + 2p_i$$

Attempts to balance model fit (the MLE under the estimation model) and error variance (the number of p parameters in model i).

Score can be computed for a single model and compared to other candidate models:

$$\Delta AIC_i = AIC_i - \min AIC$$

Enables comparison of non-nested models.

Avoids multiple test issues.

Assumes ML estimates are known without error.

Less biased toward more parameter-rich models than hLRT?

Maximum Likelihood Model Selection

3. Bayesian Information Criterion (BIC)

A (crude) approximation of the marginal likelihood under the model, measuring the relative support for the model in the data (smaller values better).

$$BIC_i = -2\ln L_i + p_i \ln n_i$$

Attempts to balance model fit (the MLE under the estimation model) and error variance (the number of p parameters in model i multiplied by the n sample size in the data set).

Score can be computed for a single model and compared to other candidate models.

Enables comparison of non-nested models.

Avoids multiple test issues.

Assumes ML estimates are known without error.

Even less biased toward more parameter-rich models?

Assumes uniform prior over models and vague priors for parameters within models.

Maximum Likelihood Model Selection

The approaches generally differ in their bias toward more parameter rich models:

$$hLRT > \Delta AIC > \Delta BIC$$

Different selection criteria may identify a different optimal model.

hLRT:	ln(L)
AIC:	- 2ln(L) + 2p
BIC:	- 2ln(L) + p ln(n)

Likelihood function:

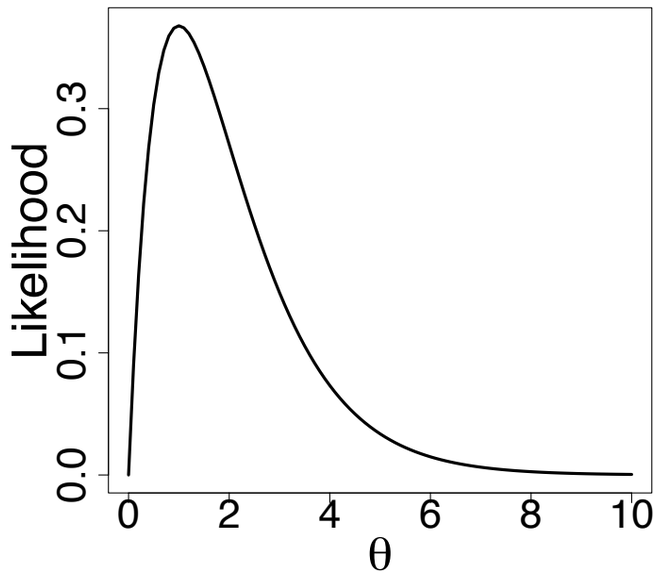
$$L(\hat{\lambda} \mid \text{TCAT / TTAT / TTGT}) = P(\text{TCAT / TTAT / TTGT} \mid \hat{\lambda}, \pi, \gamma, \lambda, \mu, \ldots)$$

Estimate: $\max_{\hat{\lambda}} L(\hat{\lambda} \mid \text{TCAT / TTAT / TTGT})$

But what is the uncertainty in our estimate?

Profile Likelihood:

$$L(\curlywedge \mid \begin{smallmatrix}T&C&A&T\\T&T&A&T\\T&T&G&T\end{smallmatrix}) = \max_{\Theta}(L(\curlywedge \mid \begin{smallmatrix}T&C&A&T\\T&T&A&T\\T&T&G&T\end{smallmatrix}, \Theta))$$

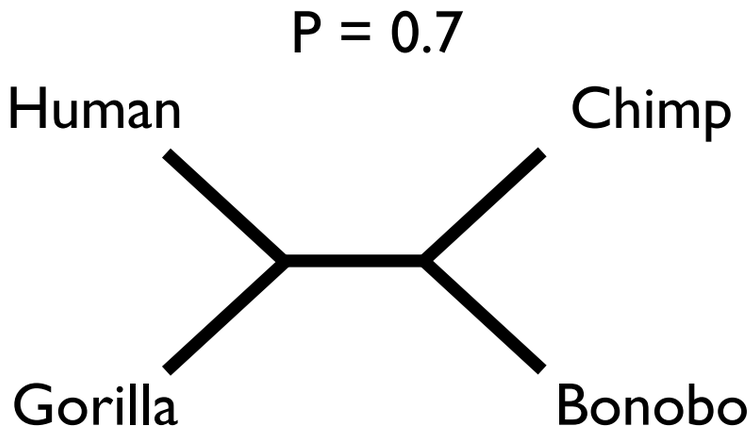


Likelihood curve of Θ for a given tree

Marginal Likelihood:

$$L(\curlywedge \mid \begin{smallmatrix}T&C&A&T\\T&T&A&T\\T&T&G&T\end{smallmatrix}) = \int L(\curlywedge \mid \begin{smallmatrix}T&C&A&T\\T&T&A&T\\T&T&G&T\end{smallmatrix}, \Theta)\, d\Theta$$

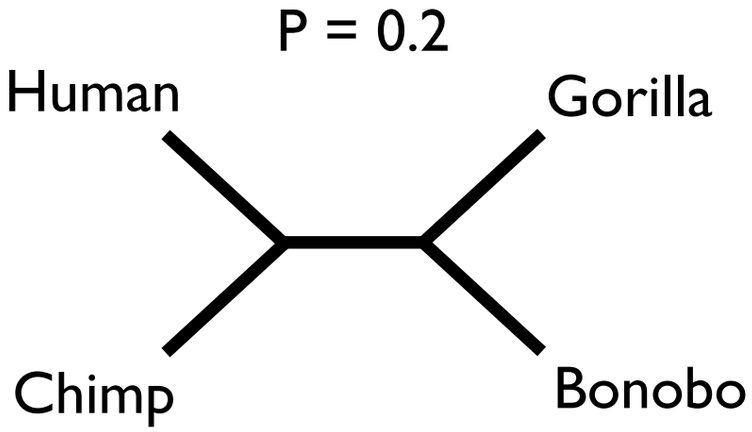
Model 1: Constrained (Chimp - Bonobo)

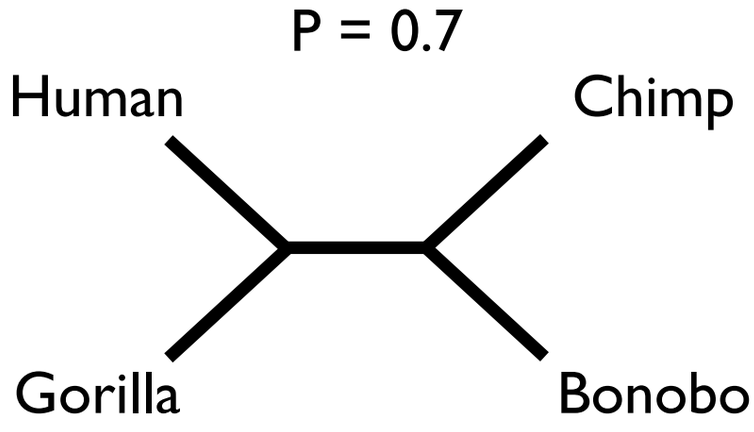


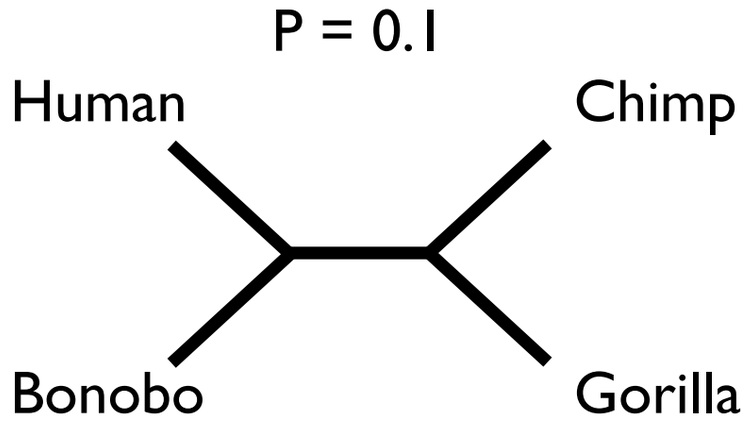
Marginal Likelihood
0.7

Model 2: Unconstrained

Marginal Likelihood
0.2/3 + 0.7/3 + 0.1/3 = 1/3







Outline

- Model Selection:
  - Frequentist Inference:
    - Likelihood Ratio Tests
    - AIC (Akaike's Information Criterion)
    - BIC (Bayesian Information Criterion)
  - Bayesian Inference:
    - Bayes Factors
    - Computing Marginal Likelihoods

- Model Adequacy Testing:
  - Posterior Predictive Testing

D = Data
θ = Model parameters

Posterior distribution

Prior distribution

”Likelihood”

$$f(\theta \mid D) = \frac{f(\theta) f(D \mid \theta)}{\int f(\theta) f(D \mid \theta)\, d\theta}$$

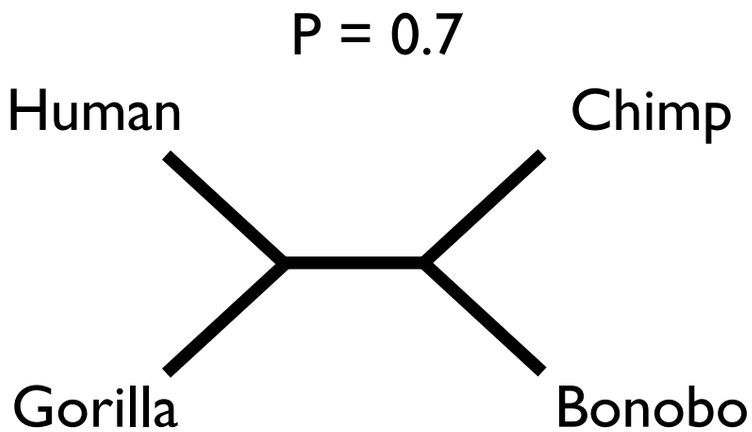
Marginal Likelihood

Bayes Factors

$$BF = \frac{P(\text{Data}|H_1)}{P(\text{Data}|H_0)}$$

$$= \frac{\text{Posterior}(H_1)}{\text{Posterior}(H_0)} \times \frac{\text{Prior}(H_0)}{\text{Prior}(H_1)}$$

Model 1: Constrained (Chimp - Bonobo)

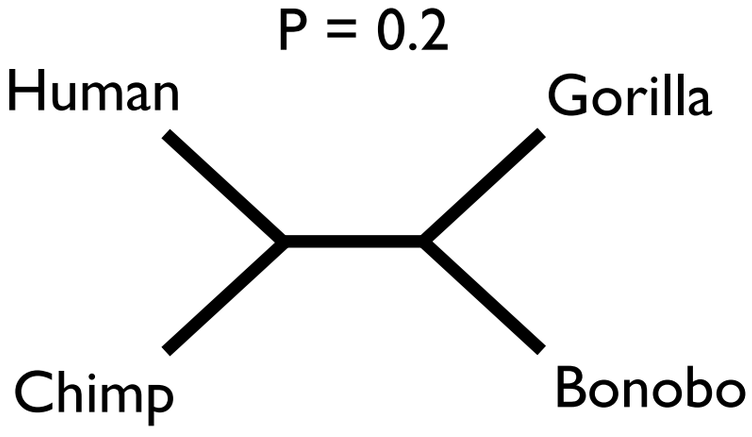


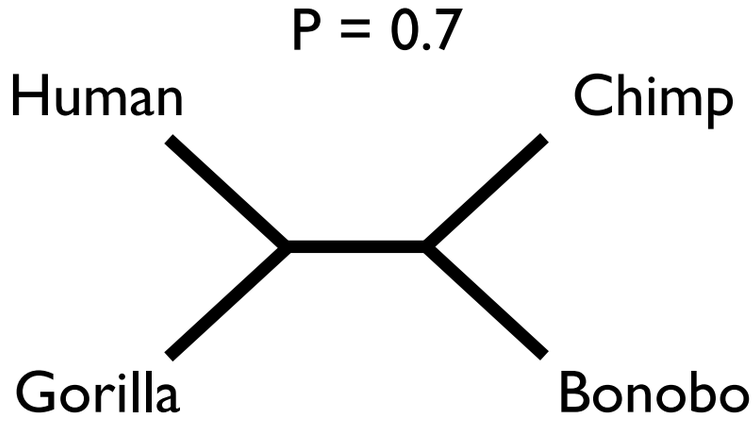
Marginal Likelihood
0.7

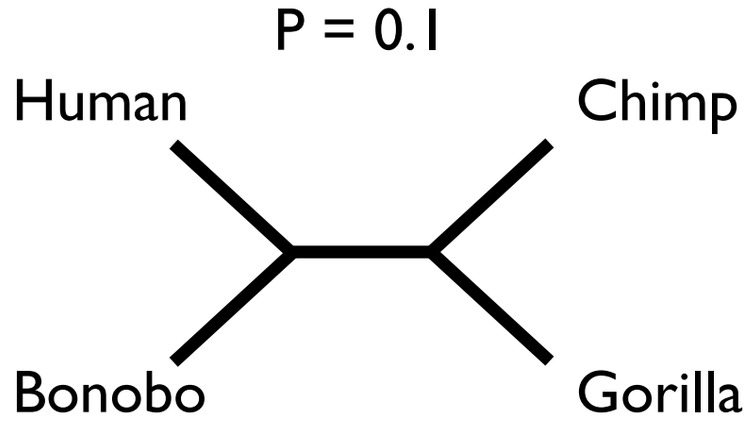
Bayes Factor: M1 / M2 = 2.1

Model 2: Unconstrained

Marginal Likelihood
0.2/3 + 0.7/3 + 0.1/3 = 1/3







Interpreting Bayes Factors

scale of evidence for Bayes factors	
Bayes factor	Interpretation
B.F. < 1/10	Strong evidence for Model 2
1/10 < B.F. < 1/3	Moderate evidence for Model 2
1/3 < B.F. < 1	Weak evidence for Model 2
1 < B.F. < 3	Weak evidence for Model 1
3 < B.F < 10	Moderate evidence for Model 1
B.F. > 10	Strong evidence for Model 1

Outline

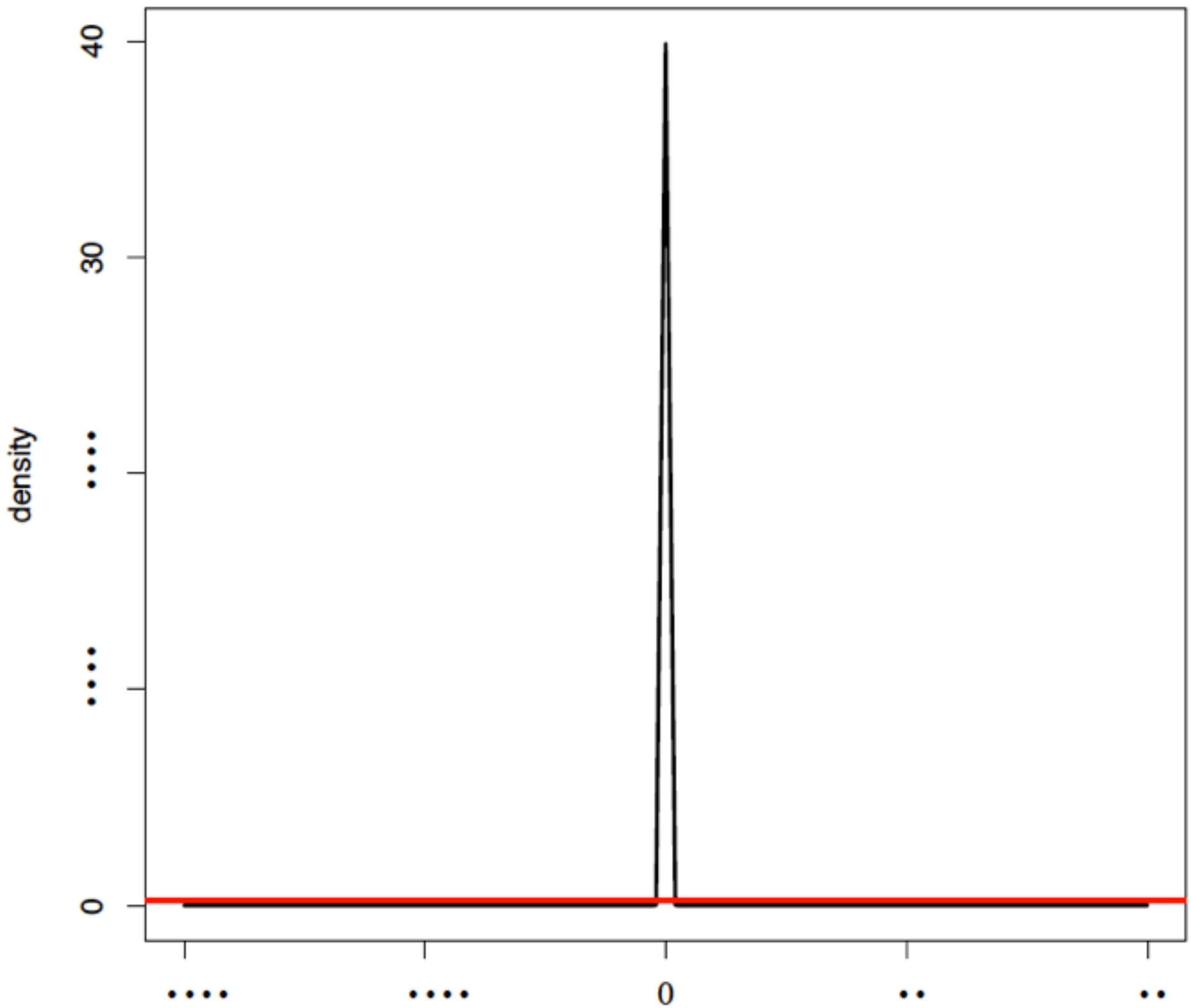
- Model Selection:
  - Frequentist Inference:
    - Likelihood Ratio Tests
    - AIC (Akaike's Information Criterion)
    - BIC (Bayesian Information Criterion)
  - Bayesian Inference:
    - Bayes Factors
    - Computing Marginal Likelihoods

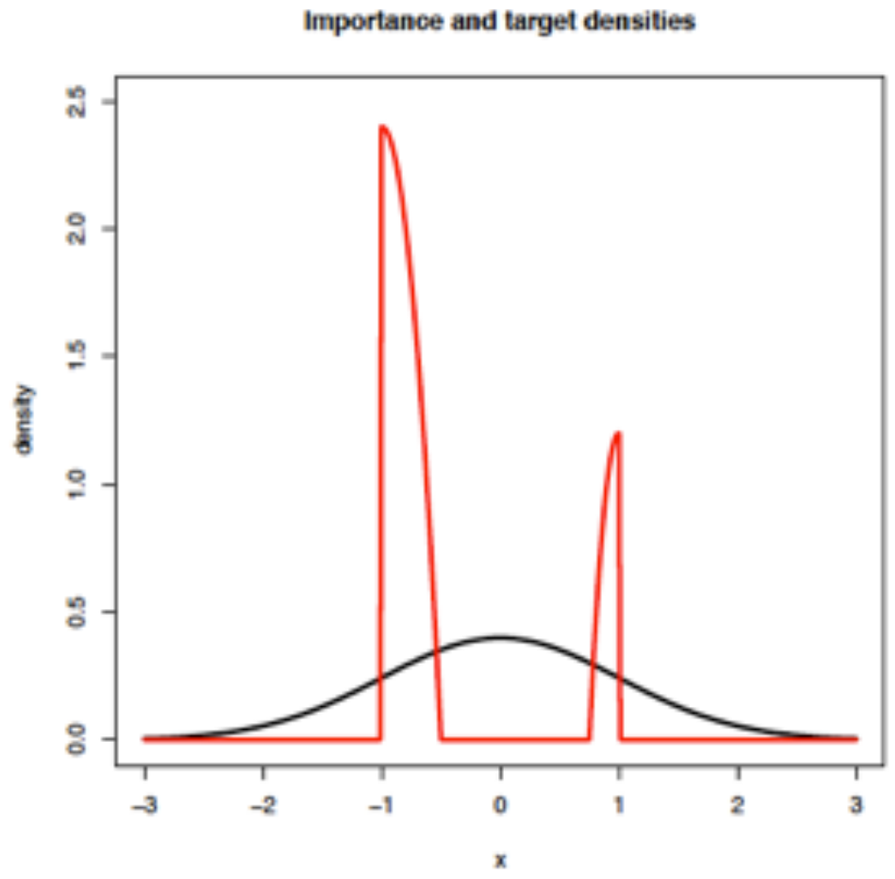
- Model Adequacy Testing:
  - Posterior Predictive Testing

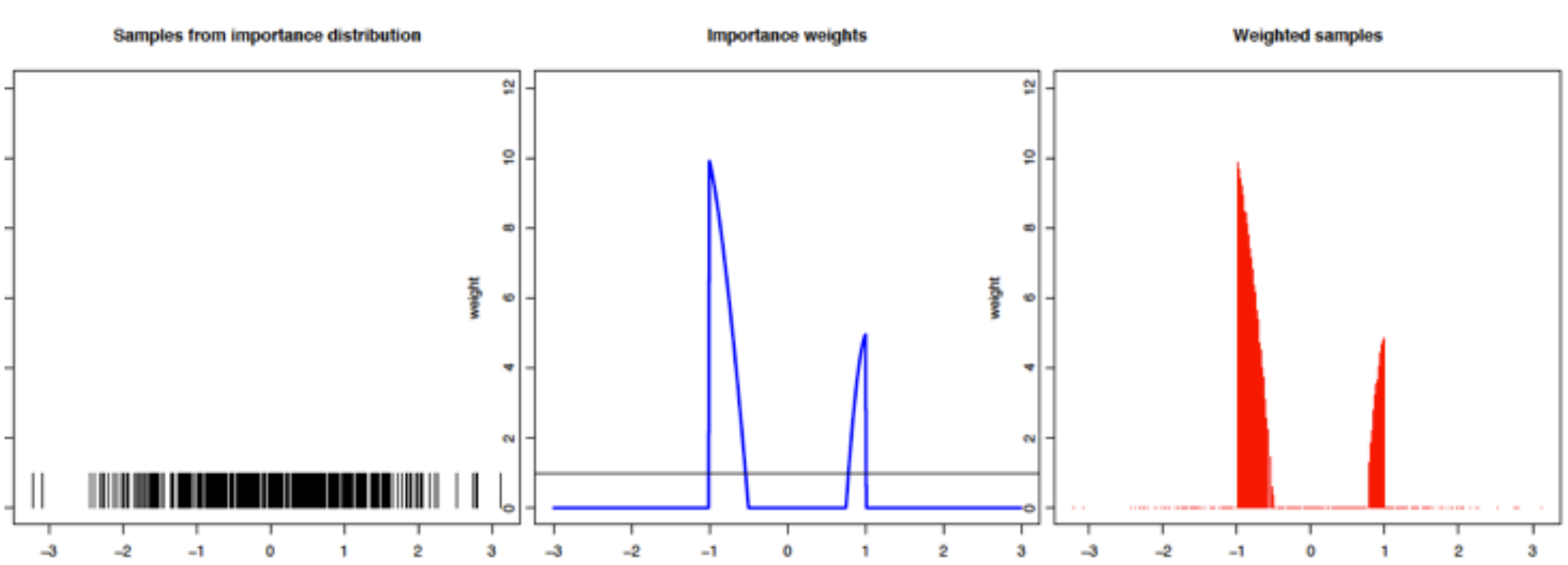
Marginal Likelihood

$$P(D|M) = \int P(D|\theta)p(\theta|M)d\theta$$

Probability of your data given the model, marginalized (integrate/summed) over all parameters.


Sharp posterior (black) and prior (red)
density
40
30
0
0


Importance and target densities
density
x


Samples from importance distribution
Importance weights
Weighted samples
weight
weight

Approximating Marginal Likelihoods

Harmonic mean estimator:

$$H = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^{-1} \right)^{-1} = \frac{1}{\frac{1}{n} \cdot \left(\frac{1}{x1} + \frac{1}{x2} + \cdots + \frac{1}{x_n} \right)} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}.$$

x_i is the likelihood for sample i

Approximating Marginal Likelihoods

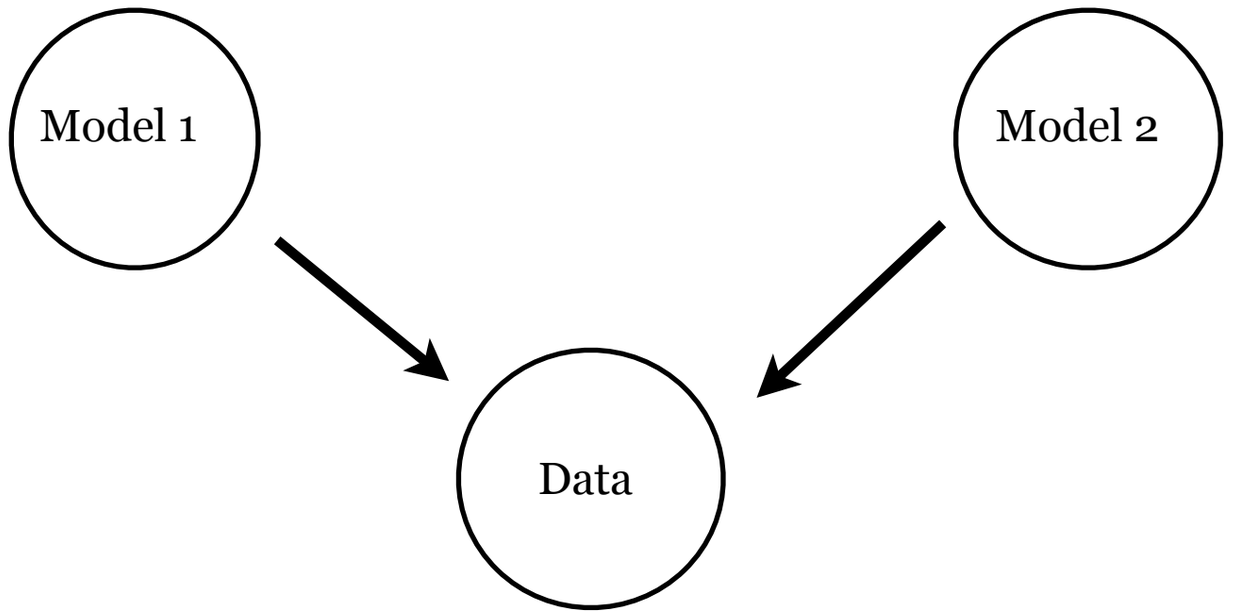
Reversible-Jump MCMC

Propose new model
with some values.
Accept/Reject using MCMC

Problem: Finding moves for proposing models is hard and model specific

Approximating Marginal Likelihoods

Mixture Models:



$P(Data|M1,M2) = 0.5*P(Data|M1) + 0.5*P(Data|M2)$

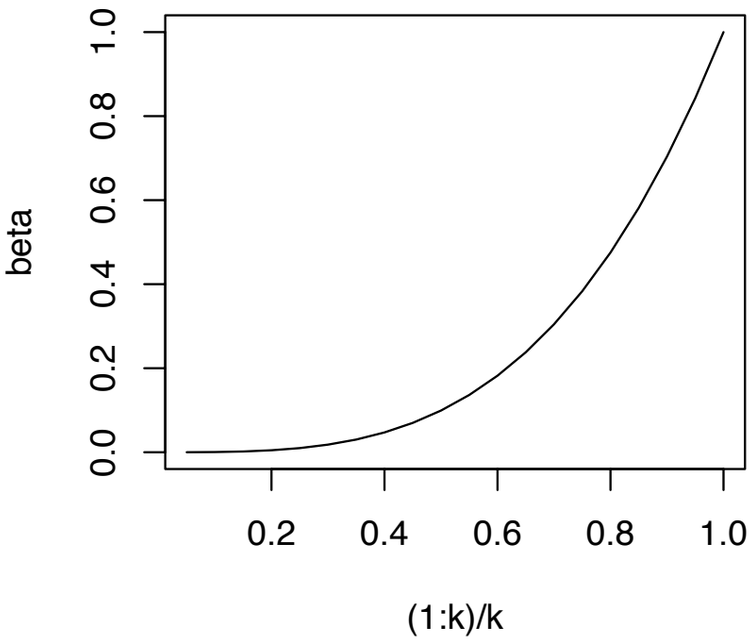
Use MCMC to accept/reject.

Sample M1 with $p = P(Data|M1) / (P(Data|M1) + P(Data|M2))$

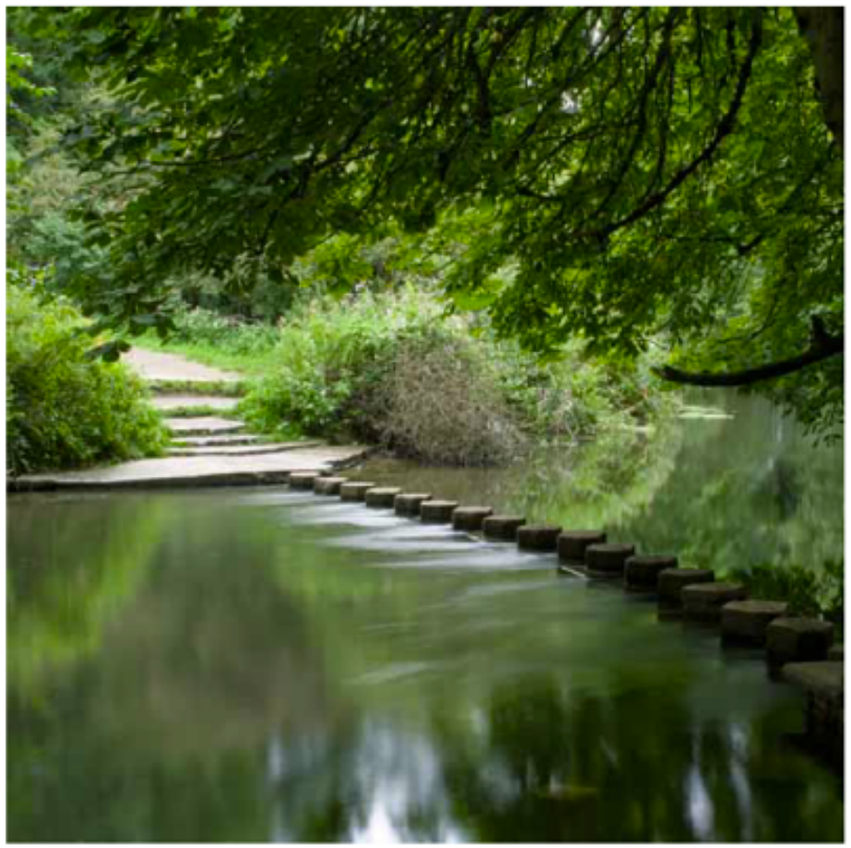
Approximating Marginal Likelihoods

Stepping-Stone-Sampling:

Run an MCMC with the likelihood to the power of beta.



$P(\Theta|D) = P(D|\Theta)^{(beta)} * P(\Theta)$

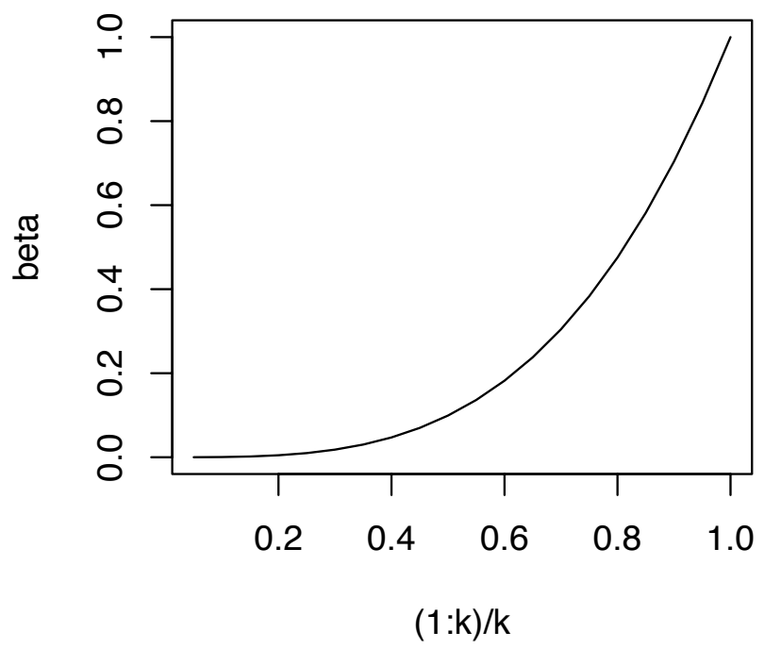


$$\mathbb{P}(D \mid M) = \left(\frac{\mathbb{P}(D|M)}{c_{0.38}}\right)\left(\frac{c_{0.38}}{c_{0.1}}\right)\left(\frac{c_{0.1}}{c_{0.01}}\right)\left(\frac{c_{0.01}}{1}\right)$$

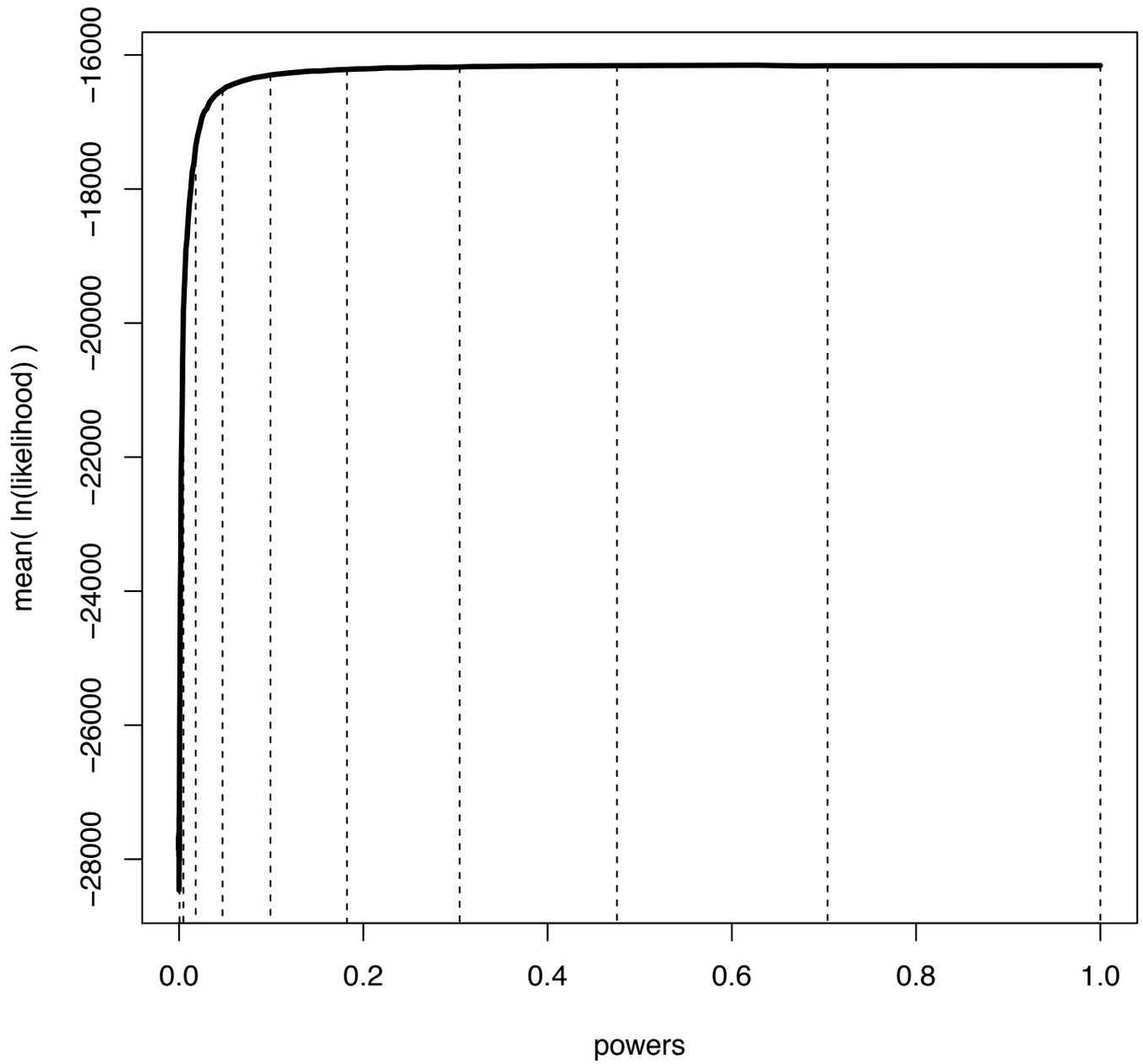
Approximating Marginal Likelihoods

Stepping-Stone-Sampling:

Run an MCMC with the likelihood to the power of beta.



$$P(\Theta|D) = P(D|\Theta)^{(beta)} * P(\Theta)$$

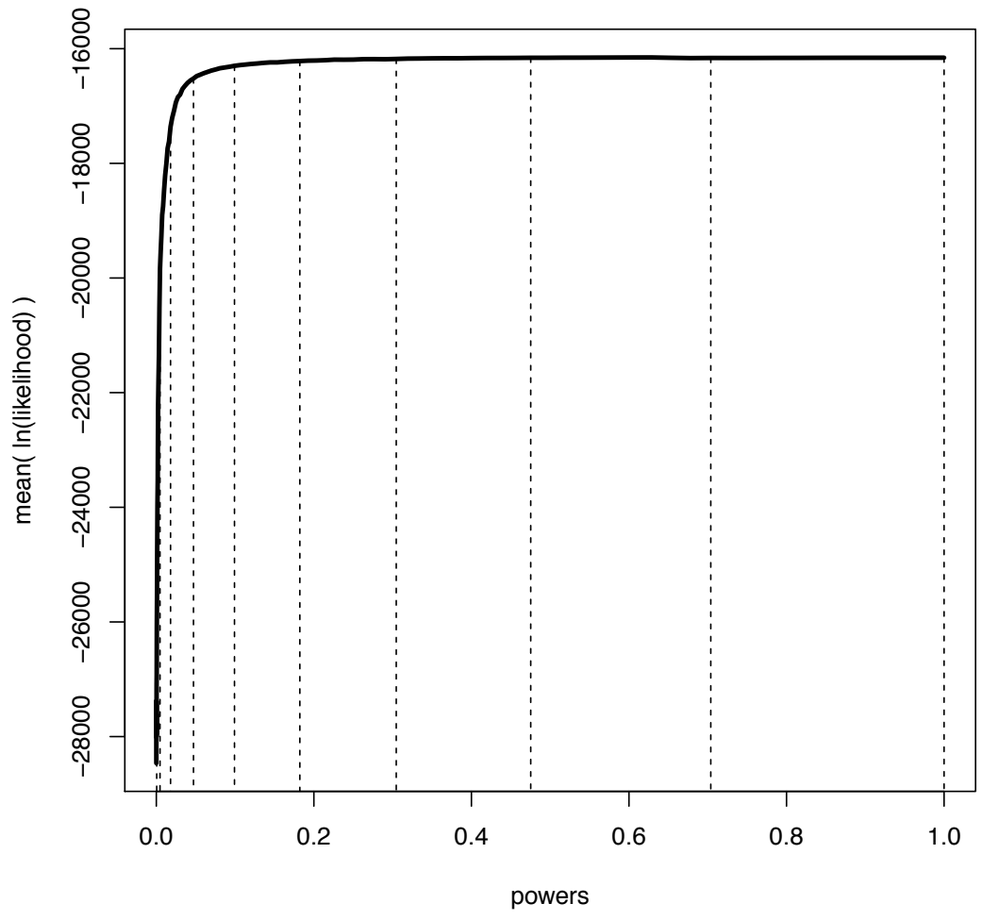


Approximating Marginal Likelihoods

Path-Sampling:

$$\ln f(D|M) = \sum_{k=1}^{K-1} \frac{\left(\frac{\sum_{i=1}^{n} \ln(l_{k-1,i})}{n} + \frac{\sum_{i=1}^{n} \ln(l_{k,i})}{n} \right) * (\beta_{k-1} - \beta_k)}{2}$$

l .. likelihood
k .. index of power
i .. sample per stone/power
β .. power

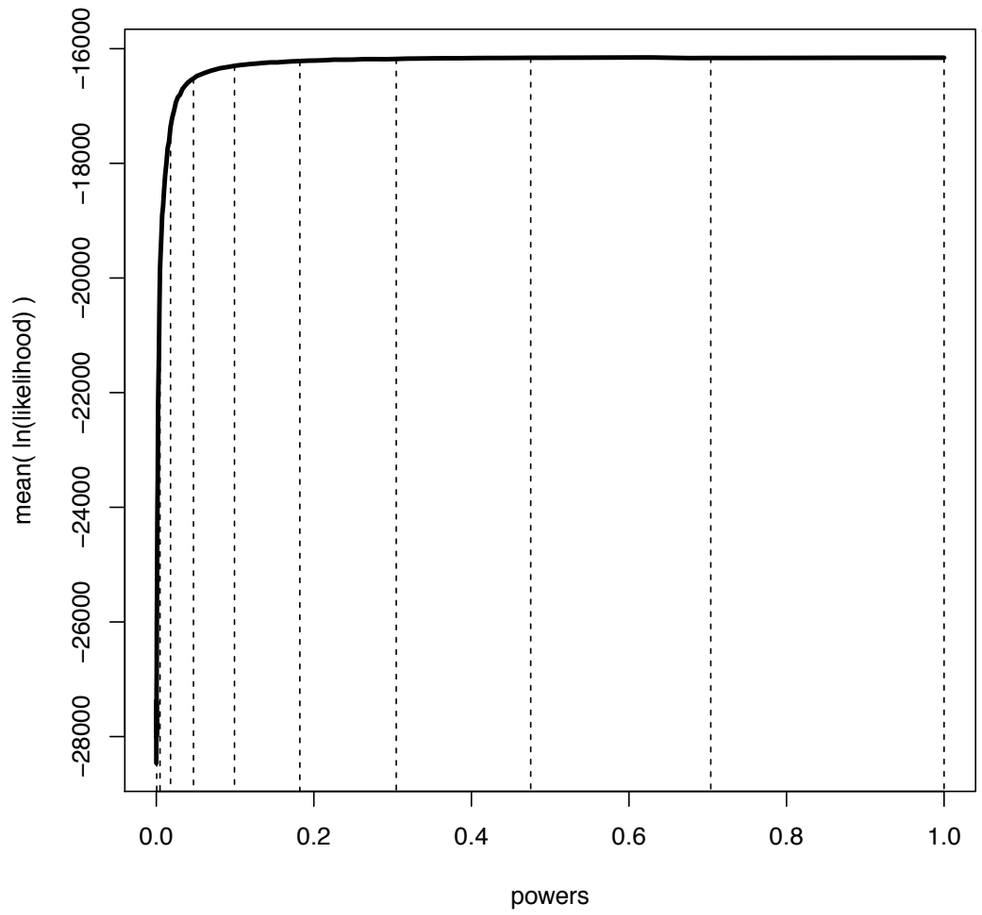


Approximating Marginal Likelihoods

Stepping-Stone—Sampling:

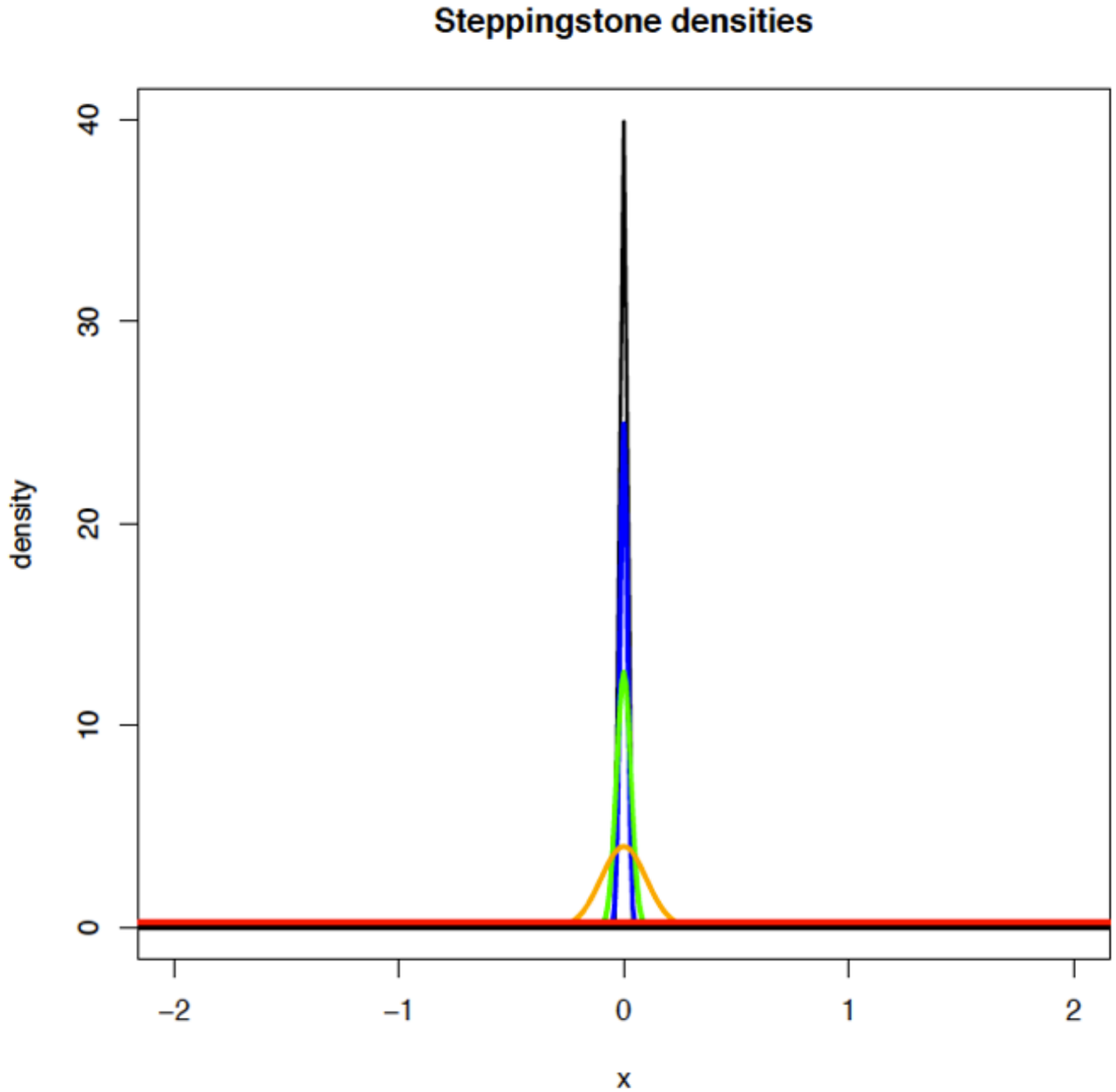
$$f(D|M) = \prod_{k=1}^{K-1} \left(\frac{1}{n} \sum_{i=1}^{n} l_{k,i}^{\beta_{k-1}-\beta_k} \right)$$

l .. likelihood
k .. index of power
i .. sample per stone/power
β .. power



Approximating Marginal Likelihoods

Path-Sampling (Stepping-Stone-Sampling):



Path Sampling

- Path-sampling (PS) and Stepping-Stone-Sampling (SSS) use the same power posterior.

- Once you have run the power-posterior mcmc you can estimate both (PS and SSS)

- SSS is slightly more robust.

- Both are time-consuming, but converge towards the true marginal likelihood.

Discussion

- LRT, AIC, BIC are unreliable
  - Monte Carlo simulations of delta-threshold

- Bayes factors (marginal likelihoods) are slow
  - Depend on the prior too

- Reversible jump - Mixture models - DPP: even slower
  - Poor MCMC mixing

- Hierarchical Models?