

Model Selection, Model Adequacy, and Big Data

Bob Thomson

Dept. of Biology
University of Hawaii
thomsonr@hawaii.edu



Topics for today

- Morning 1 - Selection of substitution models
- Morning 2 - Model Complexity & Selection of partition models
- Afternoon 1 - Model Adequacy
- Afternoon 2 - Thoughts and discussion on 'big data'

Model Selection and Testing

General Introduction to Model selection

Comparing relative model fit with Bayes factors

Model selection of common substitution models for one locus

Comparing relative model fit with Bayes factors

Model averaging of substitution models

Reversible-jump MCMC over substitution models

Model selection of partition models

Comparing relative model fit with Bayes factors

Assessing Phylogenetic Reliability Using RevBayes and P^3

Model adequacy testing using posterior prediction (Data Version).

Assessing Phylogenetic Reliability Using RevBayes and P^3

Model adequacy testing using posterior prediction (Inference Version).

<https://revbayes.github.io/tutorials/>

Topics for today

- Morning 1 - Selection of substitution models
- Morning 2 - Model Complexity & Selection of partition models
- Afternoon 1 - Model Adequacy
- Afternoon 2 - Thoughts and discussion on 'big data'

Model Selection and Testing

General Introduction to Model selection

Comparing relative model fit with Bayes factors

Model selection of common substitution models for one locus

Comparing relative model fit with Bayes factors

Model averaging of substitution models

Reversible-jump MCMC over substitution models

Model selection of partition models

Comparing relative model fit with Bayes factors

Assessing Phylogenetic Reliability Using RevBayes and P^3

Model adequacy testing using posterior prediction (Data Version).

Assessing Phylogenetic Reliability Using RevBayes and P^3

Model adequacy testing using posterior prediction (Inference Version).

Morning 1

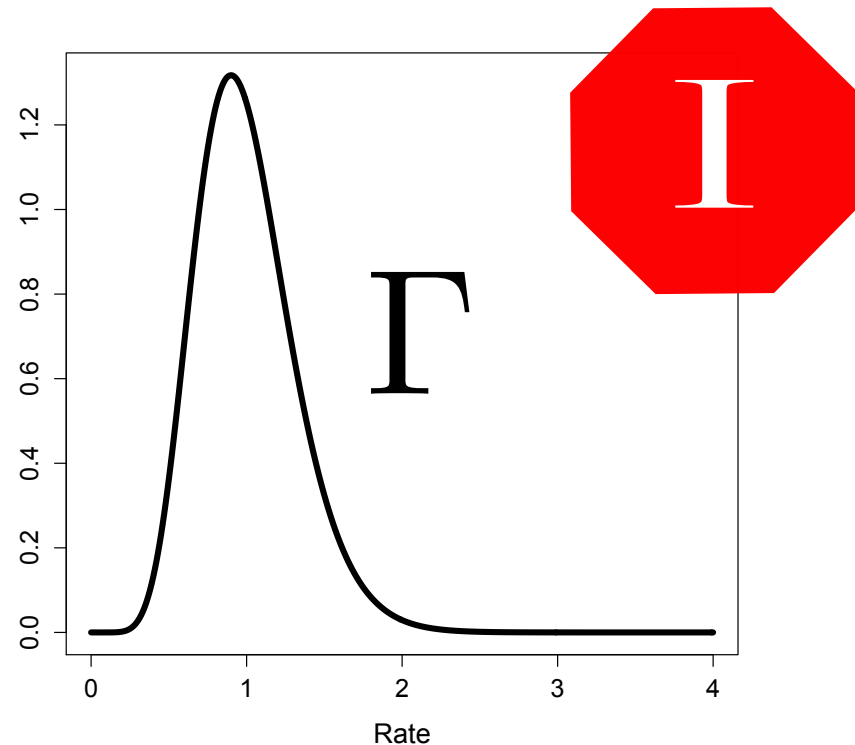
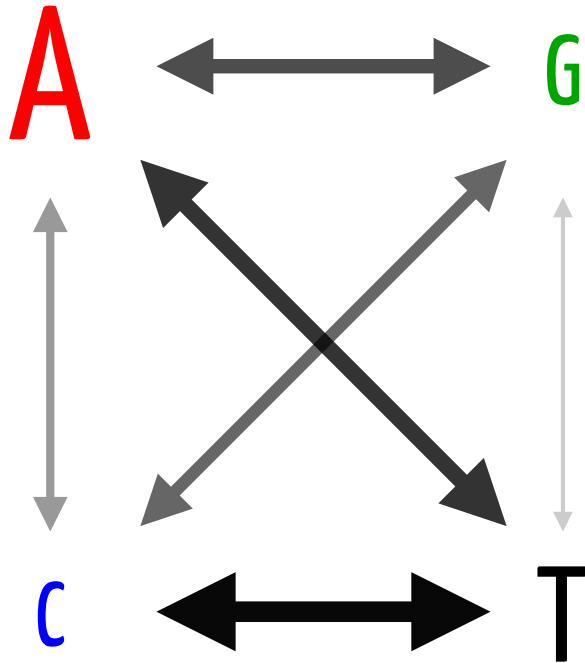
Morning 2

Afternoon 1

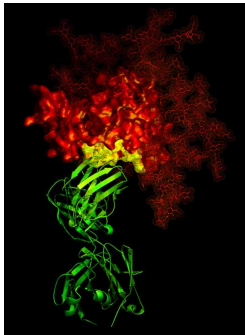
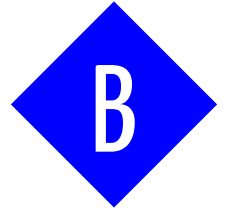
<https://revbayes.github.io/tutorials/>

So...genomes, eh?

- GTR+I+ Γ seems pretty complicated!
- 10 parameters to describe change in 4 nucleotides
- Surely that's enough.



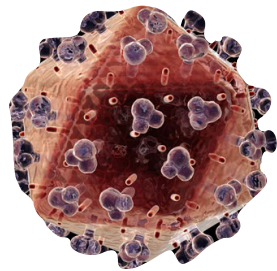
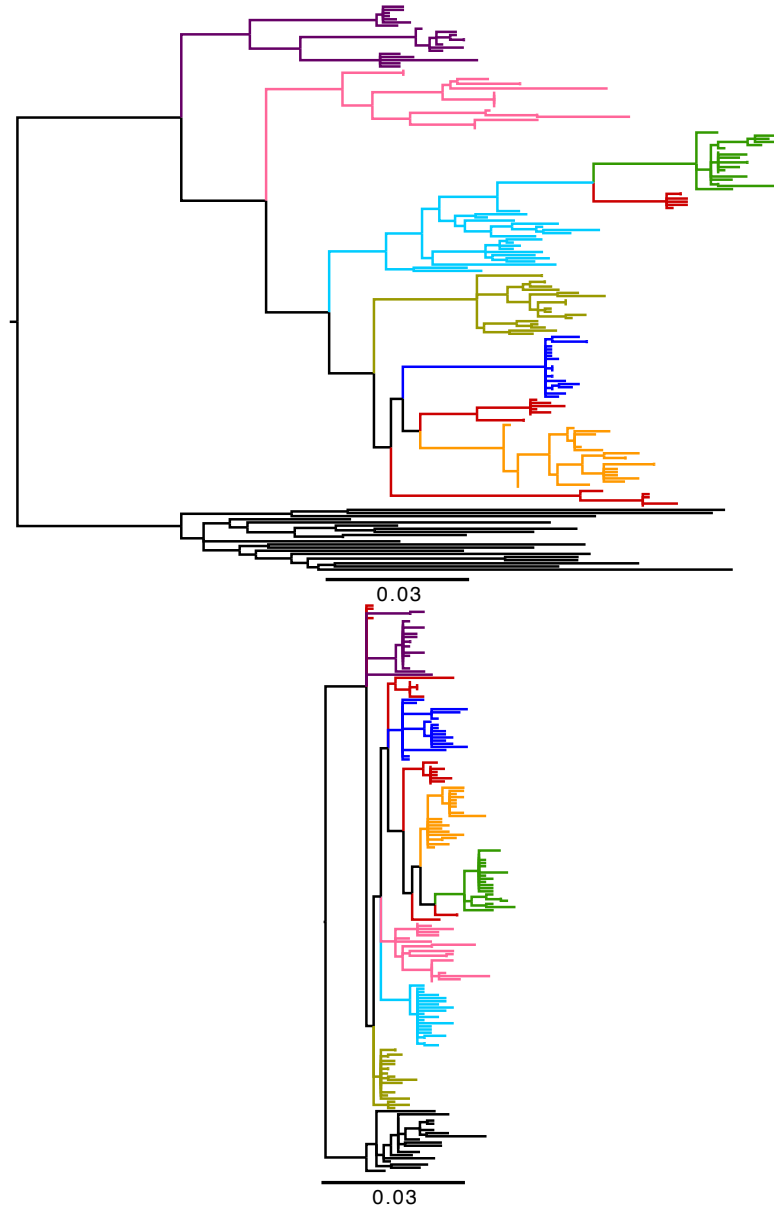
Challenge 1: Genes Vary in Rate



Envelope

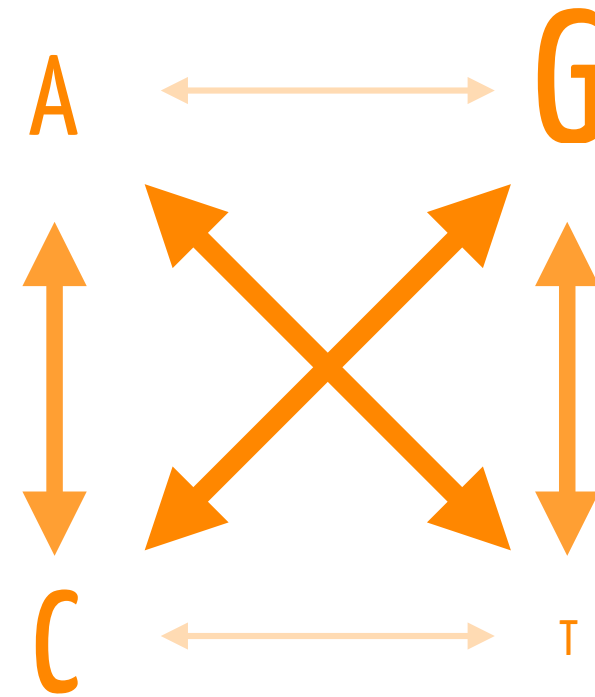
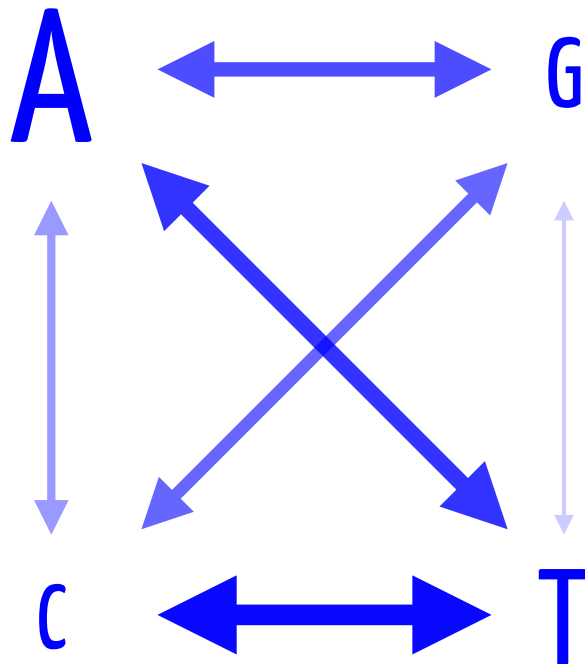
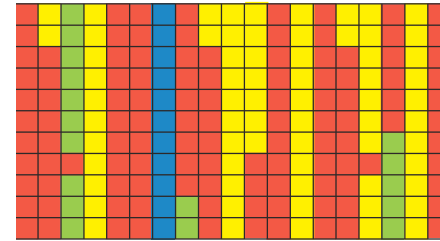
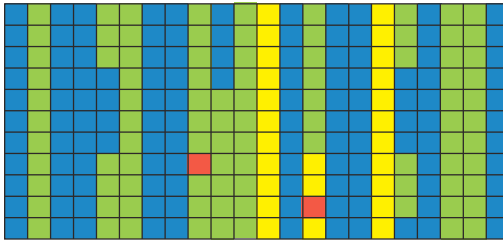
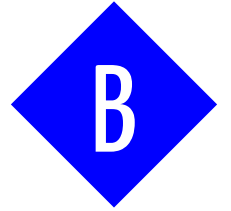


Reverse Transcriptase

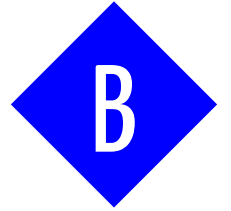


HIV

Challenge 2: Genes Vary in Model/Parameters



Challenge 3: Genes Vary in Topology



- Incomplete Lineage Sorting
- Gene Duplication
- Horizontal Gene Transfer

Challenge 4: Variation in Gene-Model Fit



- Genes and models should fit together like a hand in a glove. A glove abstracts a hand, but in a useful way.
- When fit is poor, the glove may not function properly.



Challenge 4: Variation in Gene-Model Fit



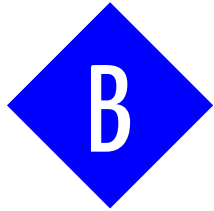
- Nearly all of our models (or at least the ones we usually consider) still assume a lot of things:
 - Independence of sites
 - Constant site rates across the tree
 - Constant base frequencies across the tree

Challenge 5: Non-homology of sites and genes



- An alignment is a statement of homology.
- We are saying that we are **certain** that nucleotides in a column have a common ancestor that diverged due to a speciation event (usually).
- This is commonly violated in at least two circumstances:
 - Alignments can be uncertain
 - Paralogy (can exert undue influence)

Types of Variation Across Genes



Rate

Topology

Model Parameters
(evolutionary dynamics)



Absolute Model Fit

Incorrect Homology

So...how do we deal with this variation?

- We develop elegant models that relax these assumptions!
- Now we do 3 things with our models:
 - Select the best available model (model selection)

So...how do we deal with this variation?

- We develop elegant models that relax these assumptions!
- Now we do 3 things with our models:
 - Select the best available model (model selection)
 - Critically evaluate the fit of this model (model adequacy)

So...how do we deal with this variation?

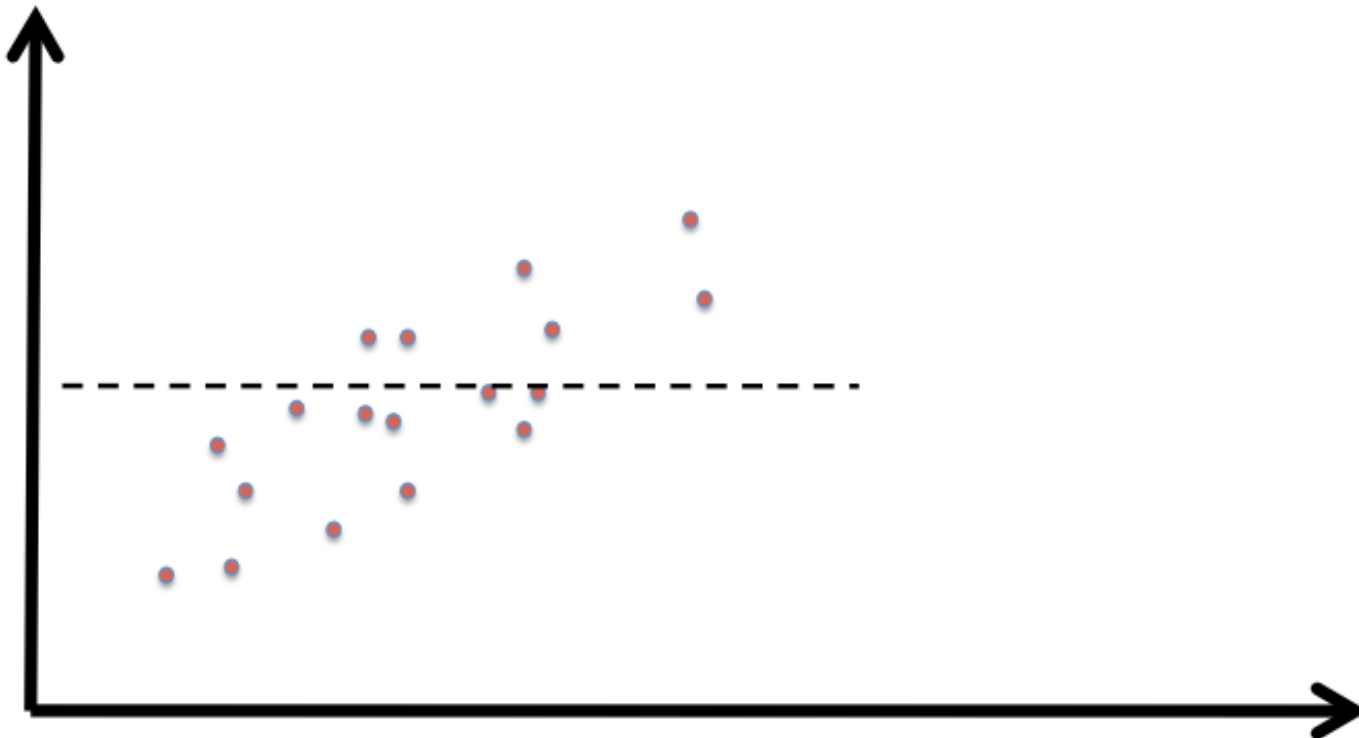
- We develop elegant models that relax these assumptions!
- Now we do 3 things with our models:
 - Select the best available model (model selection)
 - Critically evaluate the fit of this model (model adequacy)
 - Accept, refine, or reject (the art)

Model Selection

- To do statistical inference we must have a model
 - What model should that be?
 - Our goal should be to have a model that is complex enough to capture the “important” variation in the data, but not be more complex than it needs to be.

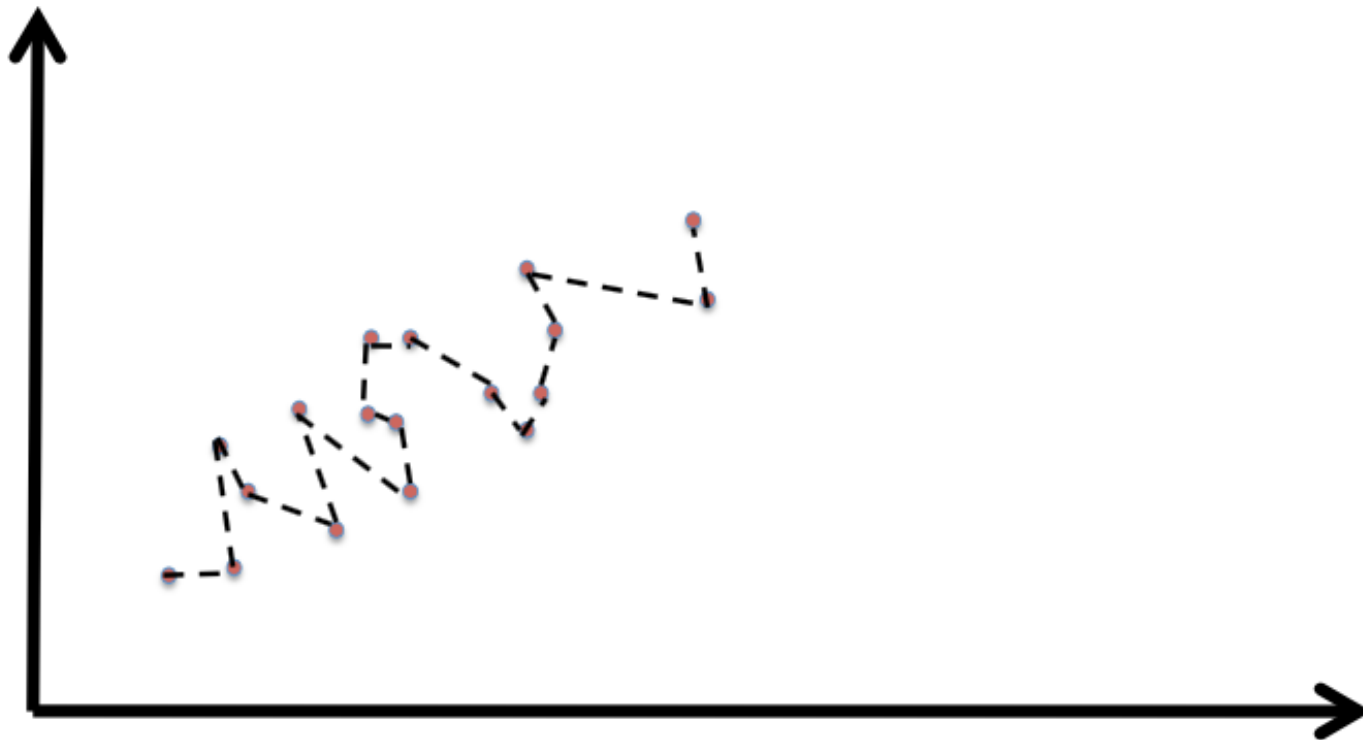
Model Selection

- Underfitting model: does not capture important variation in the data



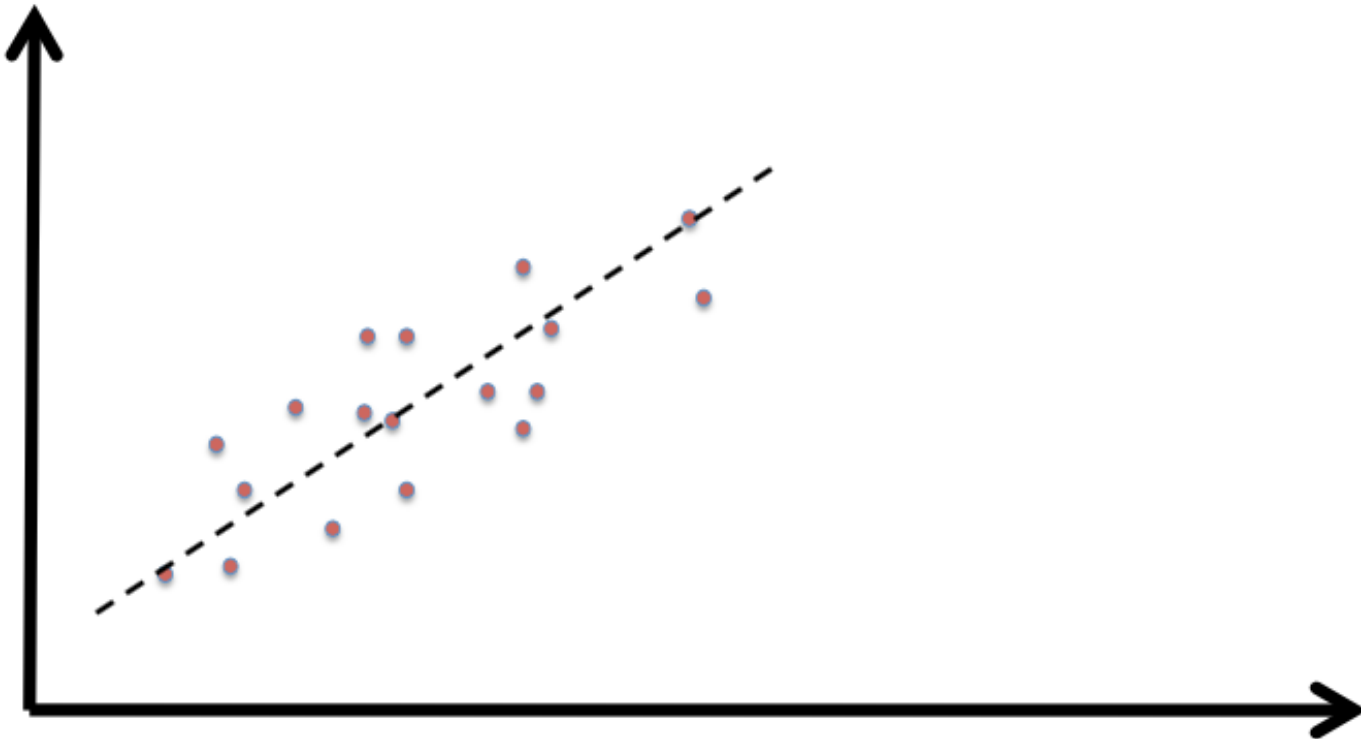
Model Selection

- Overfitting model: model captures all variation in the data, but is not a realistic description of the underlying process

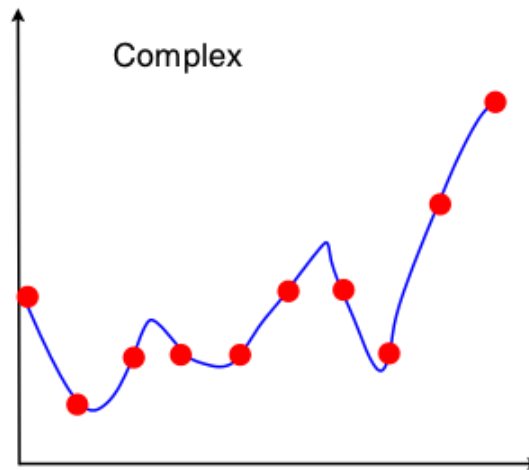
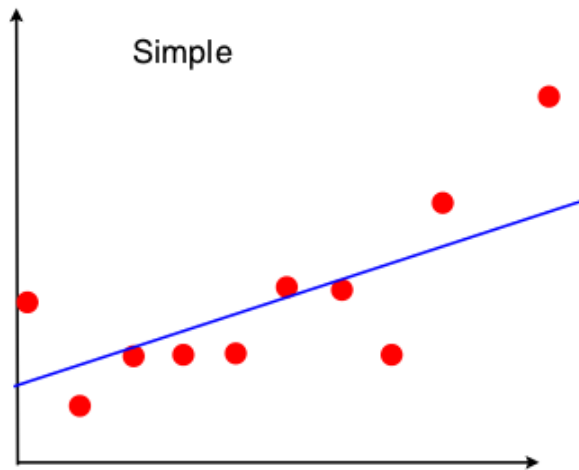
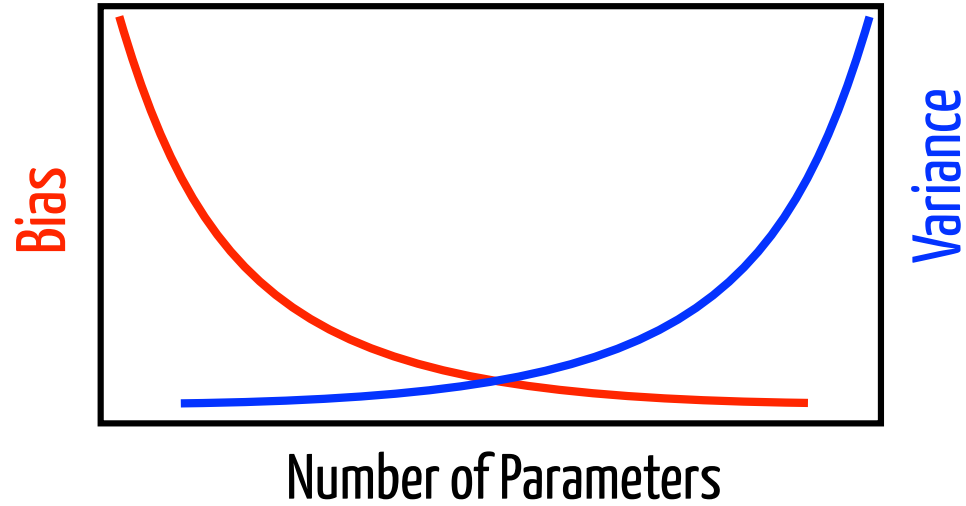


Model Selection

- Proper fit: model captures important variation

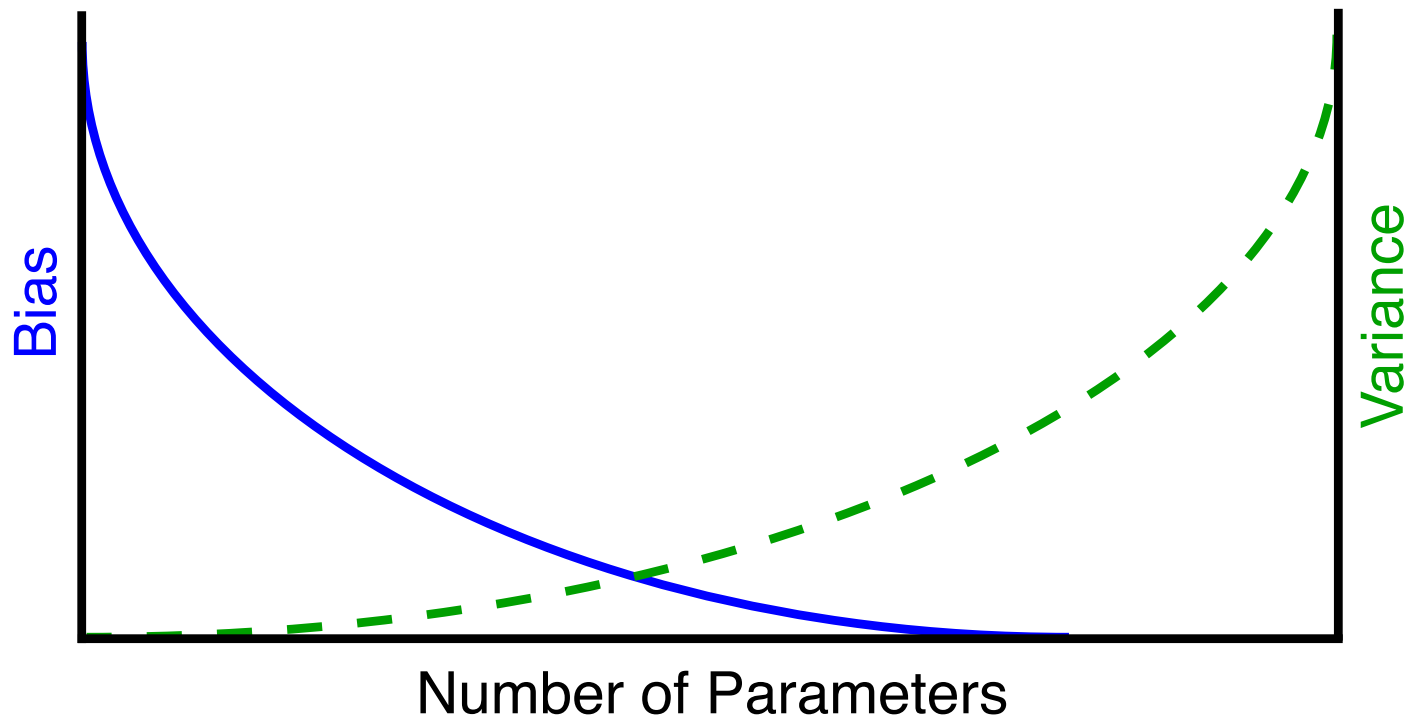


Bias Variance tradeoff



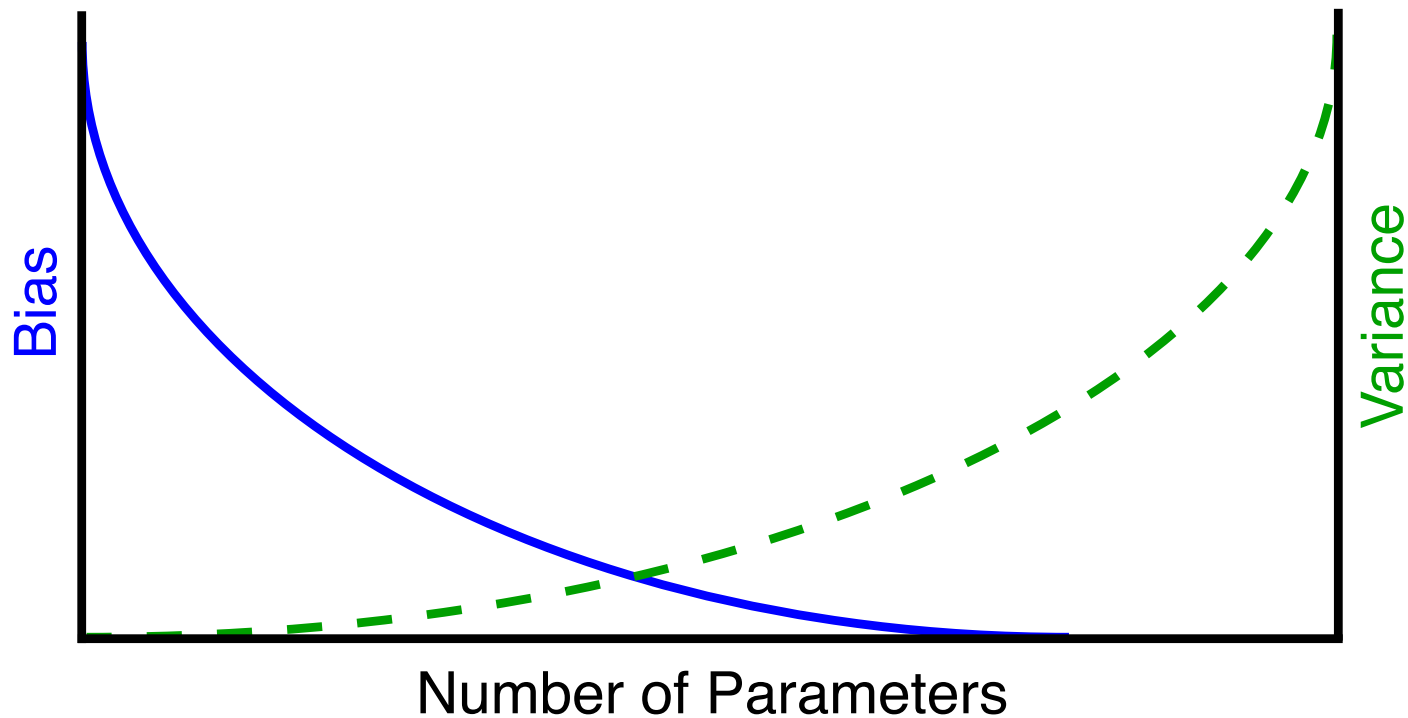
Model Selection

The Fundamental Tradeoff



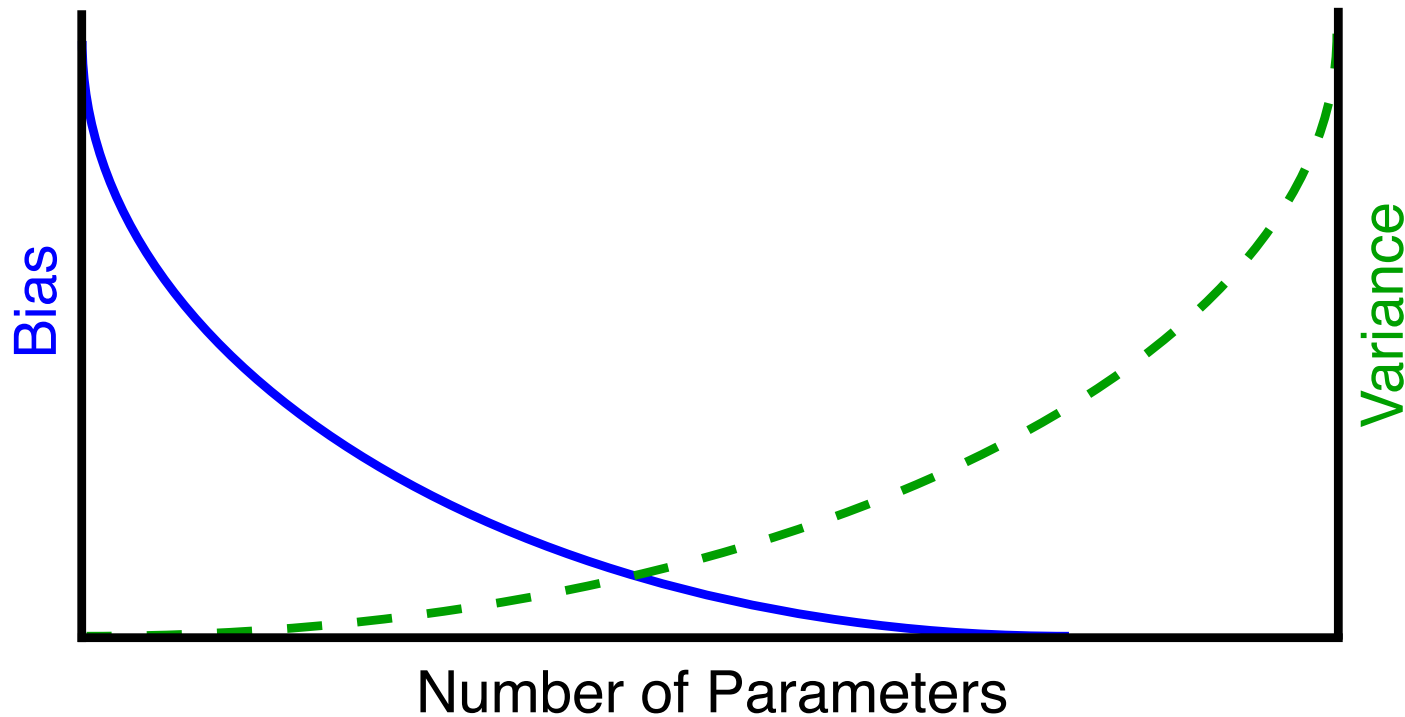
Model Selection

Model too simple!
We're misinterpreting the data.

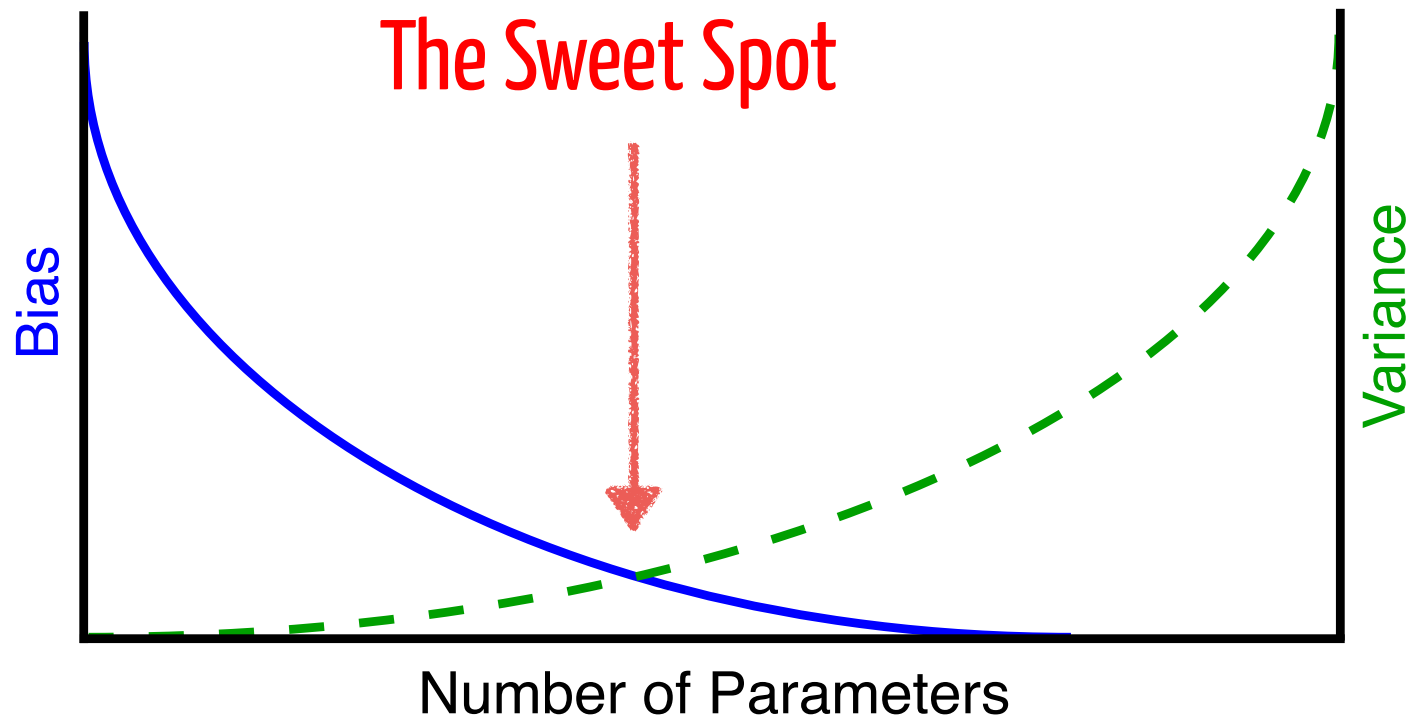


Model Selection

Model too complicated!
We don't have enough information.



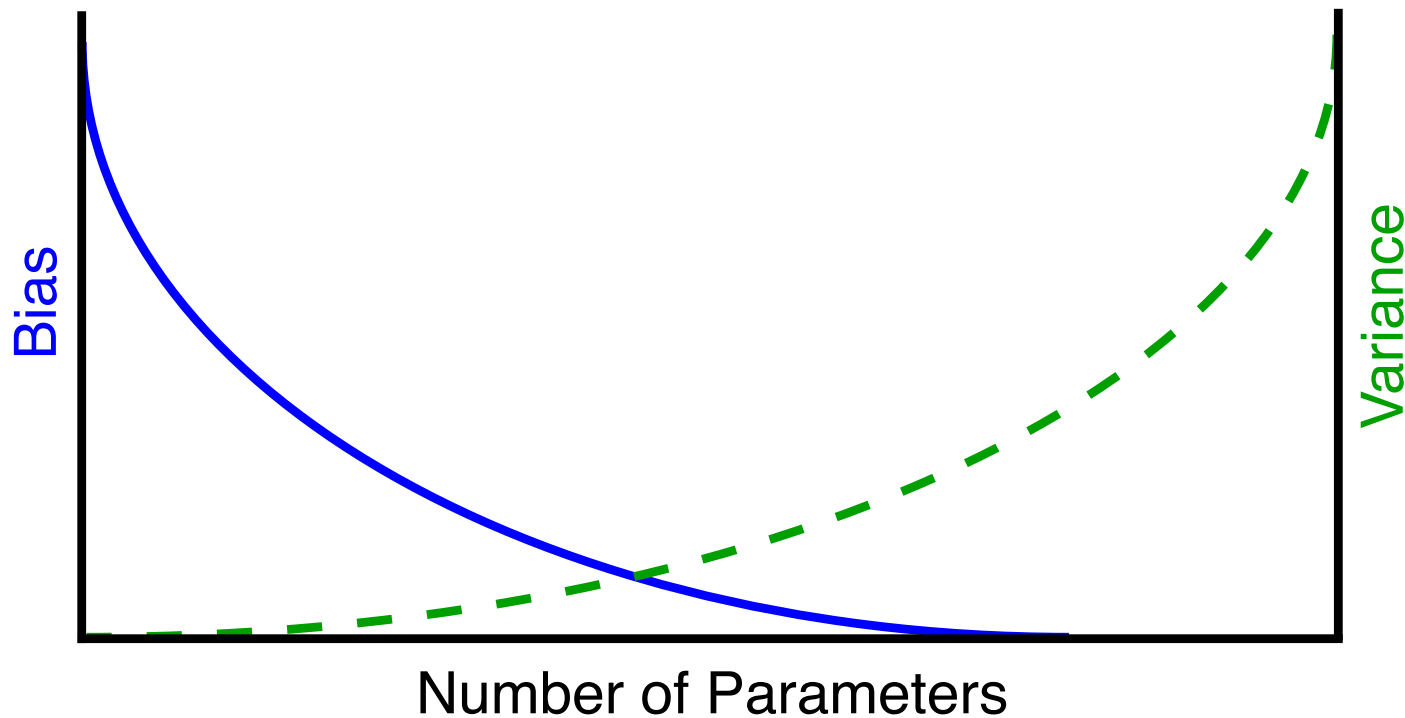
Model Selection



Model Selection

Bias and **Variance** can be traded off in different ways.

This leads to **multiple criteria** for model selection.



The Likelihood Function

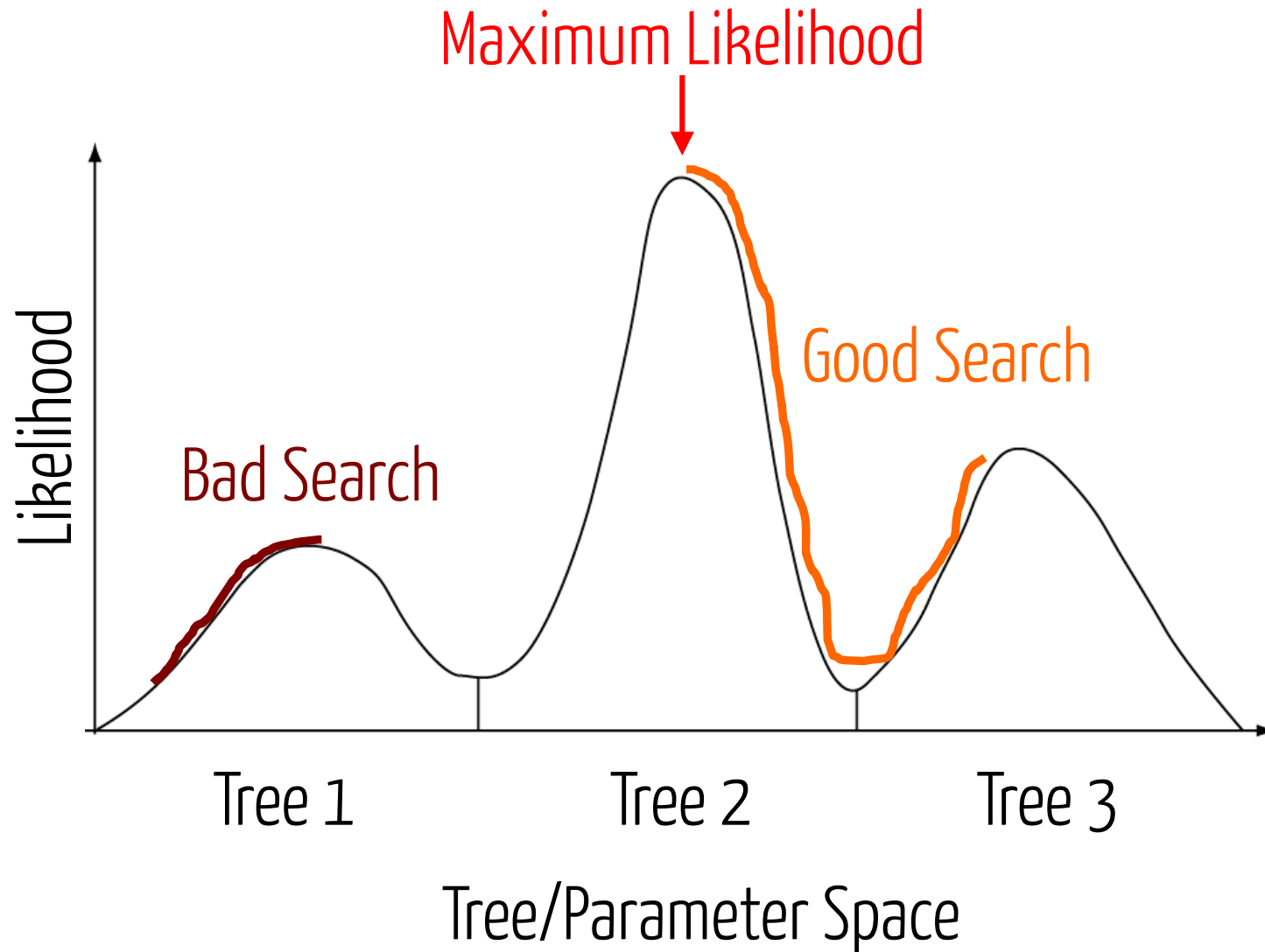
$$P(\text{sequence data} \mid \theta, \text{tree})$$

Read as “**the probability of the sequence data given a tree and model**”.

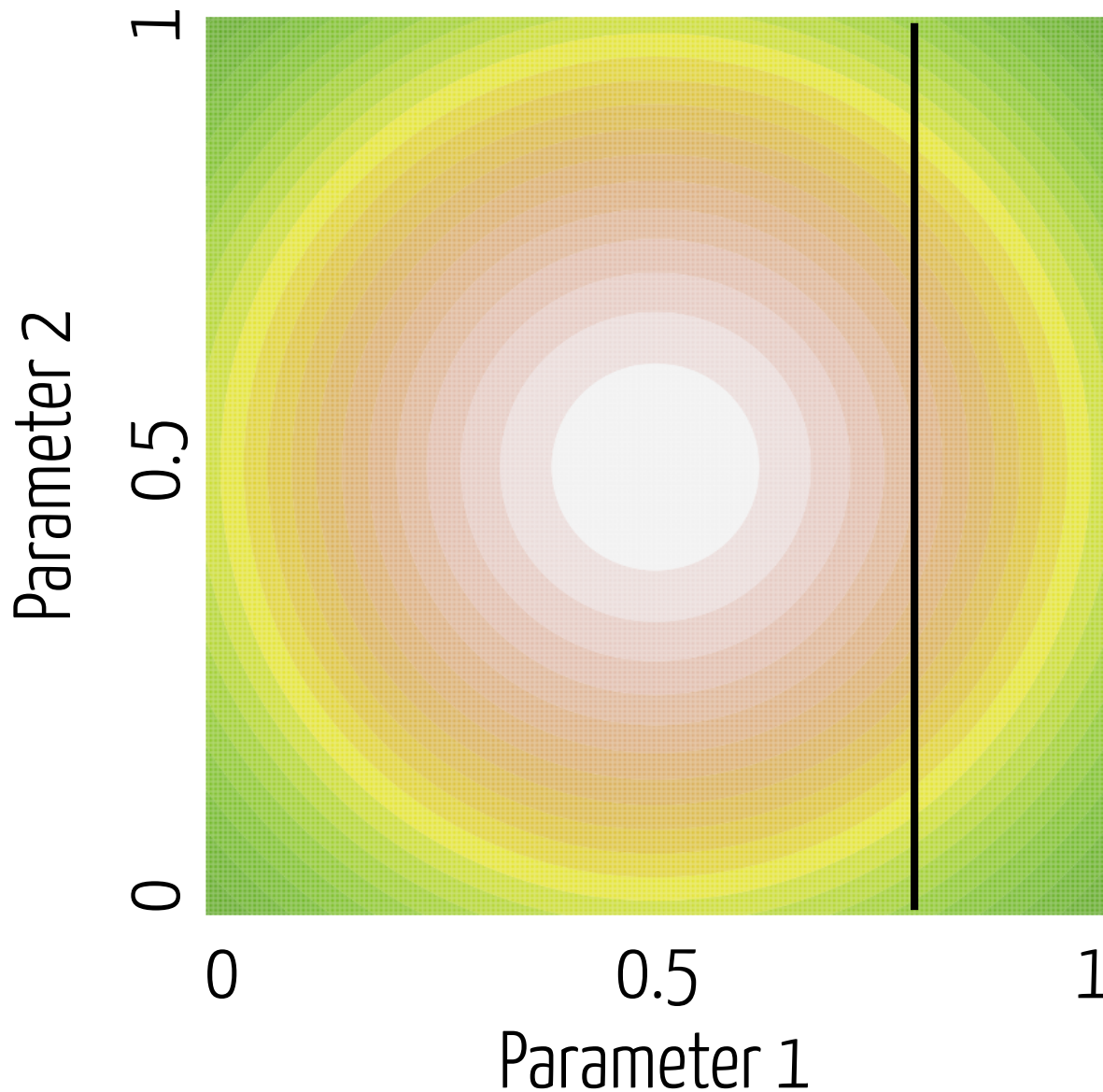
The quantity by which the data provide information.

Compares how well different trees and models predict the observed data or as a “**measure of relative surprise**”.

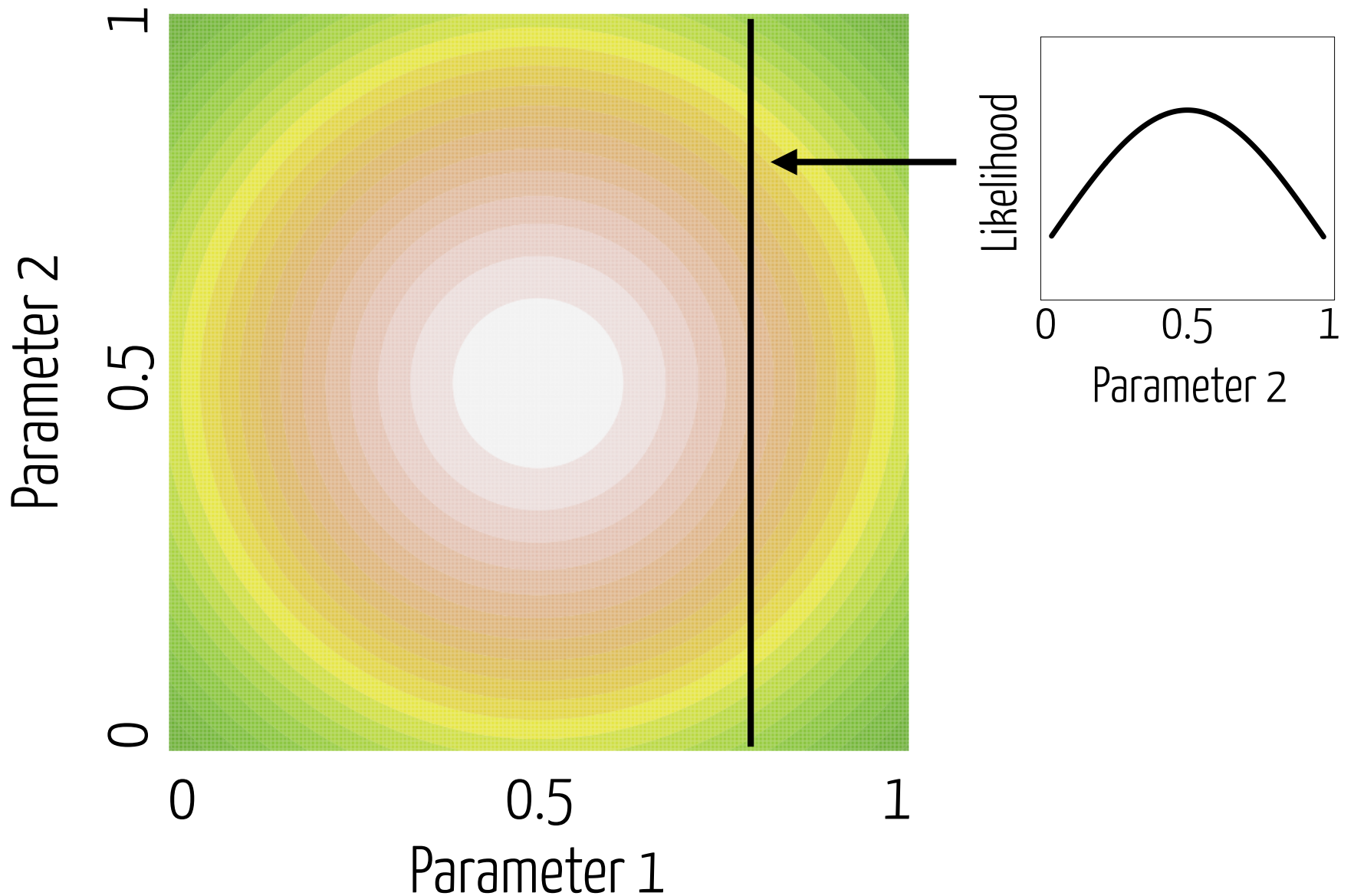
Maximum Likelihood



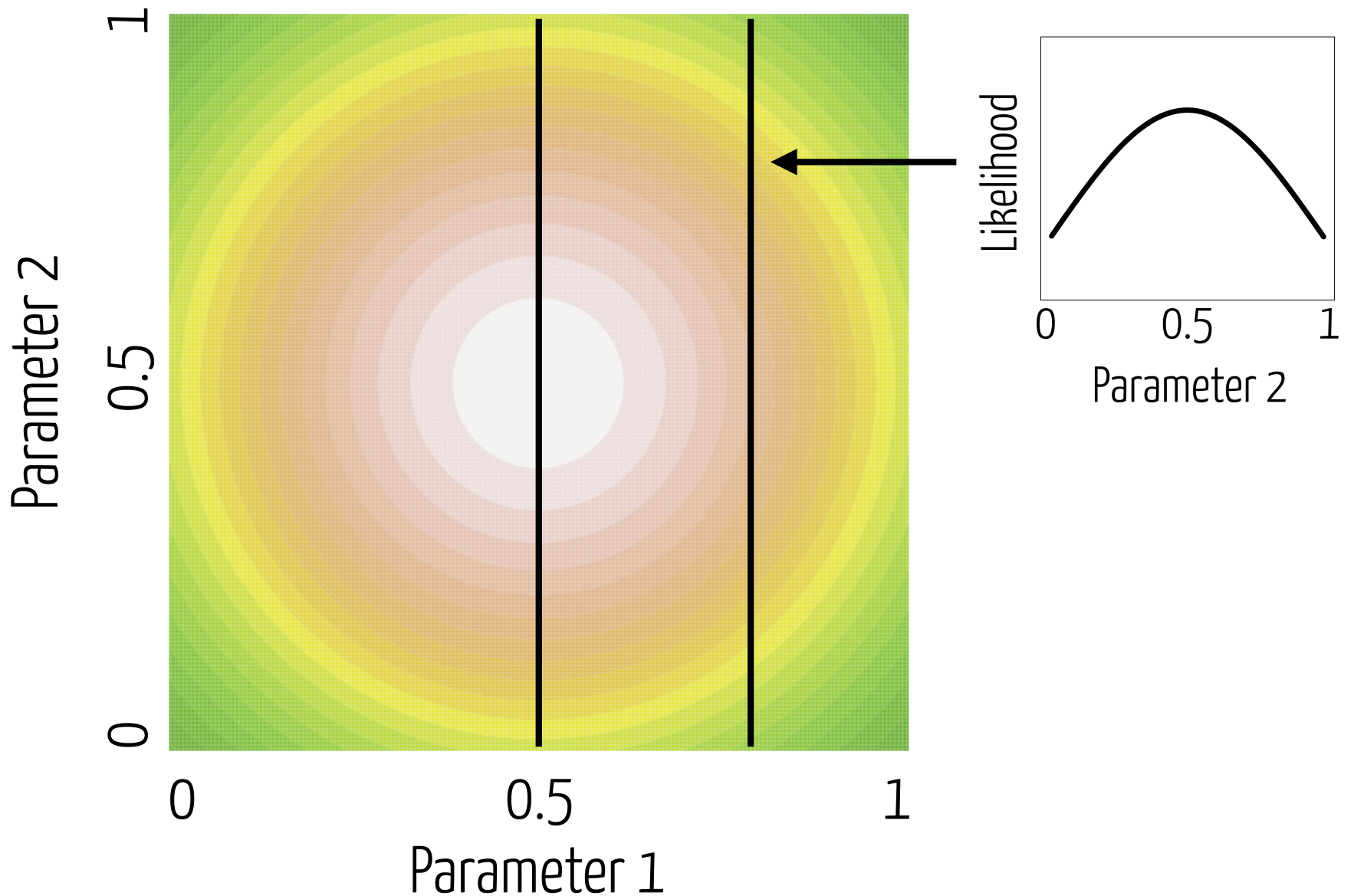
More Parameters = Better Likelihood



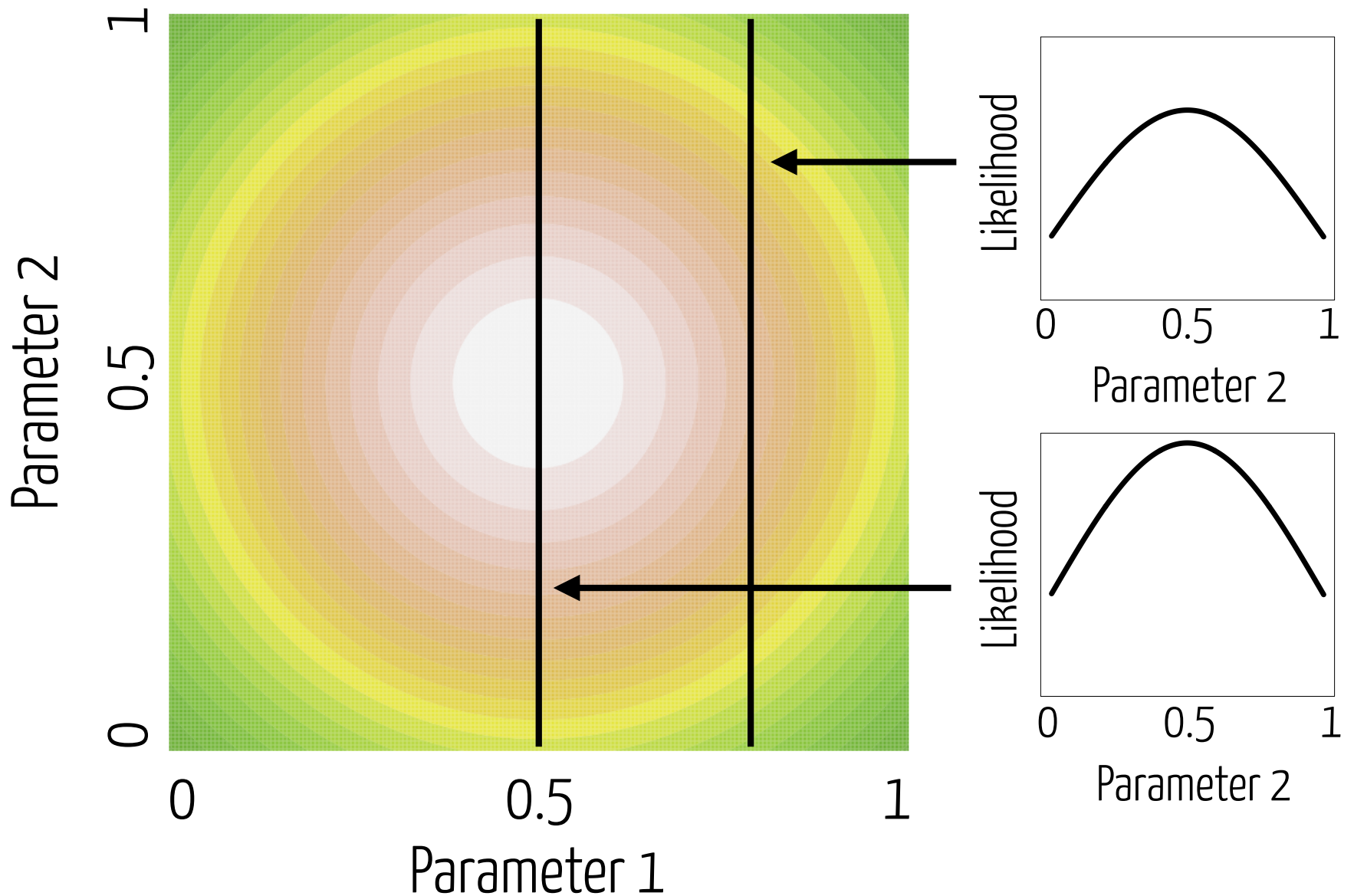
More Parameters = Better Likelihood



More Parameters = Better Likelihood

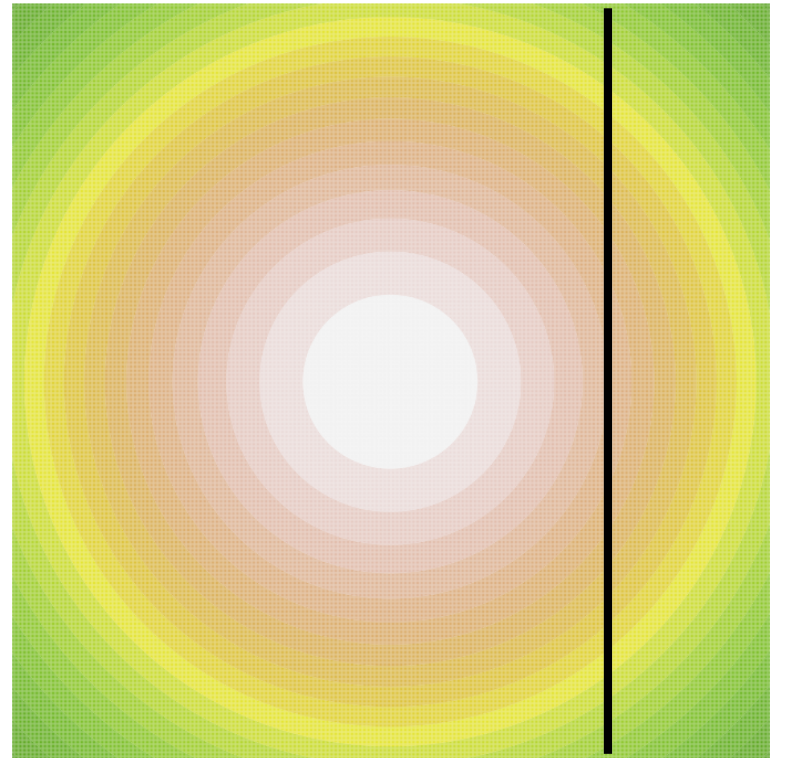


More Parameters = Better Likelihood



ML-based Model Selection

If the **more complex** model always gives a likelihood that is **at least as good** as a simpler model, even if the simpler one is true, we need ways to assess **whether it's enough better to warrant our attention**.



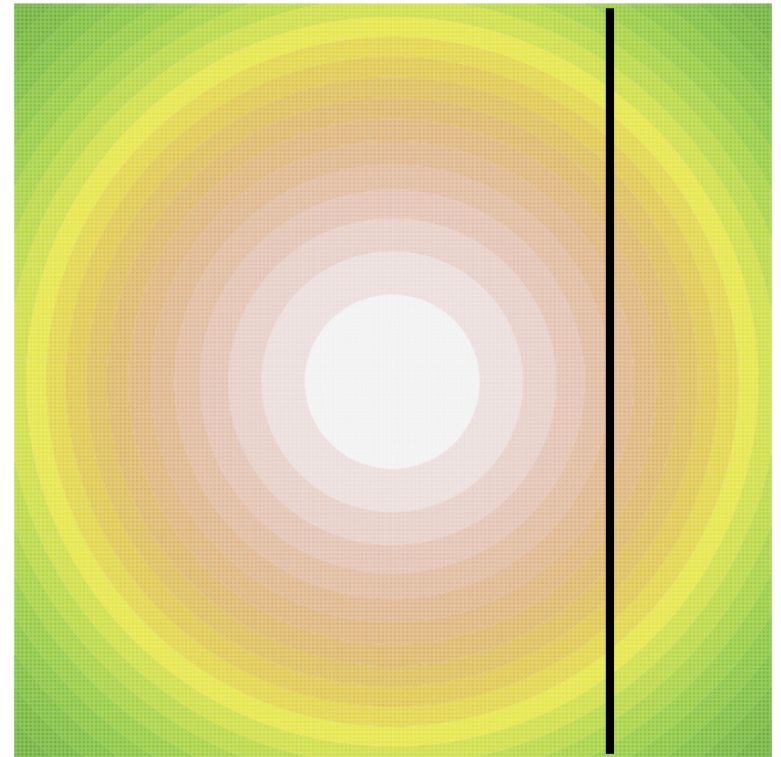
Likelihood Model Selection

- Simple comparison of likelihoods is therefore not useful
 - will typically lead us to choose overly complicated models with high error variance
- Model selection approaches are looking for a tradeoff between increase in fit and increased complexity of the model
 - incorporate a measure of each

ML-based Model Selection

If the **more complex** model always gives a likelihood that is **at least as good** as a simpler model, even if the simpler one is true, we need ways to assess **whether it's enough better to warrant our attention**.

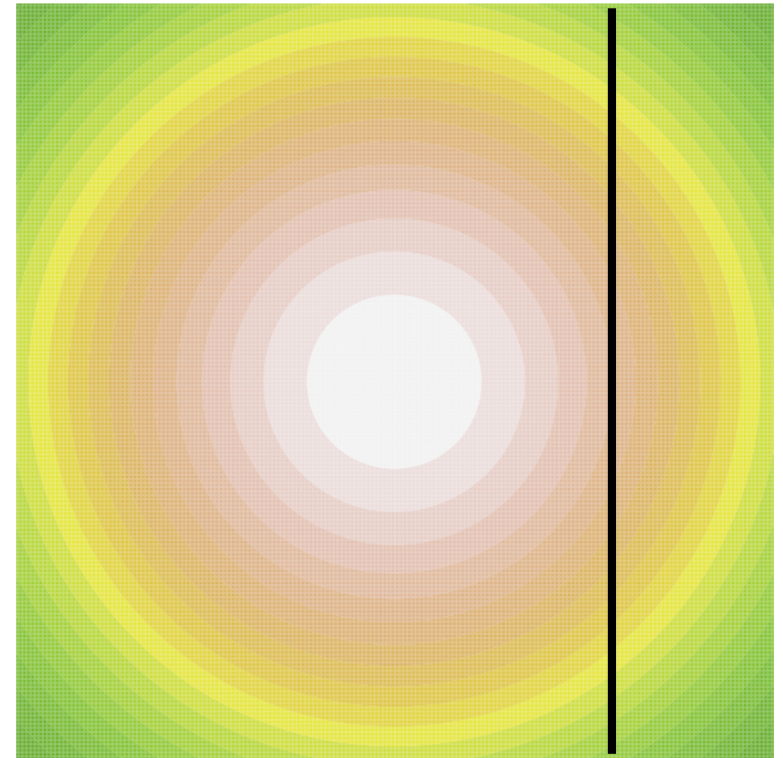
- Akaike's Information Criterion (**AIC**)
- Bayesian Information Criterion (**BIC**)
- Likelihood Ratio Test (**LRT**)



ML-based Model Selection

If the **more complex** model always gives a likelihood that is **at least as good** as a simpler model, even if the simpler one is true, we need ways to assess **whether it's enough better to warrant our attention**.

- Akaike's Information Criterion (**AIC**)
- Bayesian Information Criterion (**BIC**)
- Likelihood Ratio Test (**LRT**)



Different penalties for extra parameters.


ML-based Model Selection

Akaike's Information Criterion (**AIC**)

Minimum AIC preferred.

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

Penalty for more
parameters (k).



Likelihood term becomes
more negative when \hat{L} worse.



ML-based Model Selection

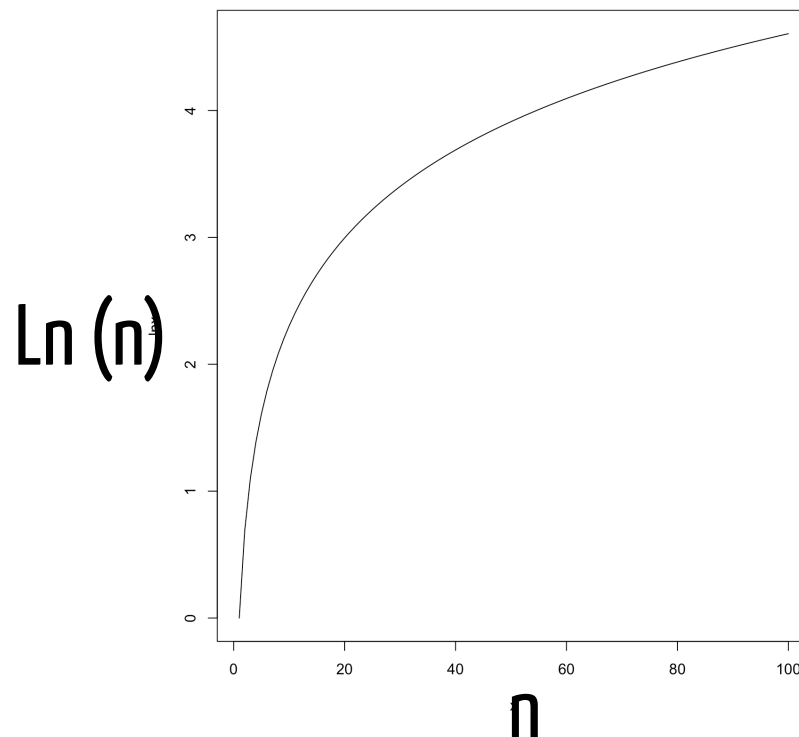
$$\text{AIC} = 2k - 2\ln(\hat{L})$$

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L})$$

ML-based Model Selection

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L})$$



Penalty term is larger
for BIC when $n > 7$

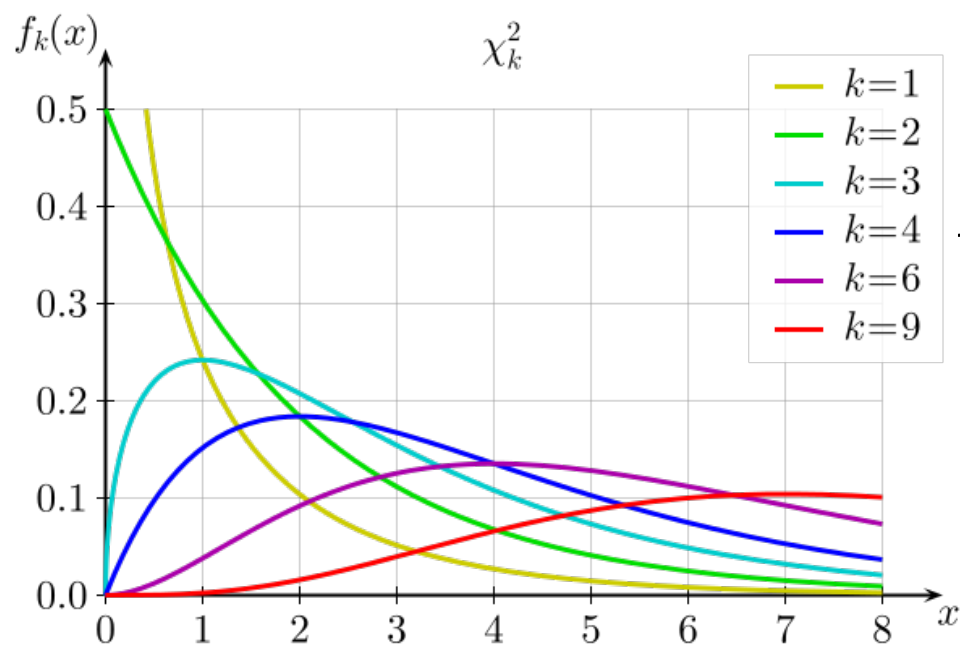
ML-based Model Selection

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L})$$

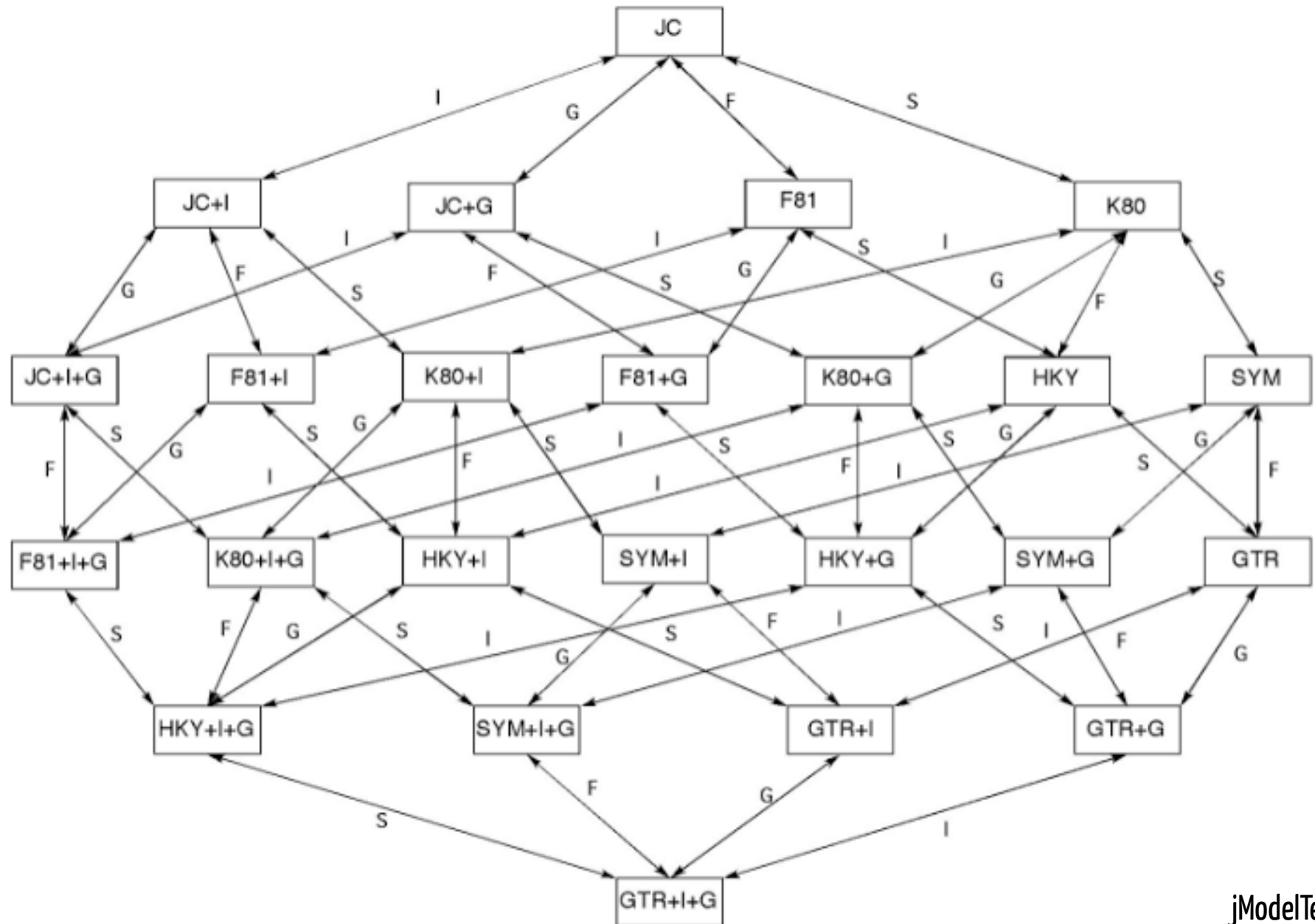
LRT
Pairwise
Hypothesis test

$$\frac{\hat{L}_o}{\hat{L}_a} \sim \chi^2$$



Difference in
free parameters

Likelihood Ratio Test - Hierarchy of Nestedness



Likelihood Model Selection

- Strength of penalty for adding extra parameters varies
- often $BIC > AIC > LRT$

Bayesian Model Selection

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

Bayesian Model Selection

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$$P(\theta|D, M) = \frac{P(D|\theta, M)P(\theta|M)}{P(D|M)}$$

Bayesian Model Selection

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$$P(\theta|D, M) = \frac{P(D|\theta, M)P(\theta|M)}{\int P(D|\theta, M)P(\theta|M)d\theta}$$

Marginal Likelihood

Probability of the data given the model, considering uncertainty in model parameters.

Bayesian Model Selection

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$$P(\theta|D, M) = \frac{P(D|\theta, M)P(\theta|M)}{\int P(D|\theta, M)P(\theta|M)d\theta}$$

Marginal Likelihood

Essentially, the **weighted average likelihood**, weighted by the prior probability of different parameter values.

Marginal Likelihood Example

Evolutionary Distance

Sp. A ————— Sp. B

Compare **JC** and **K80** models

v: edge length
estimated in both models

k: transition-transversion ratio
estimated in K80 and fixed at 1 for JC

Marginal Likelihood Example

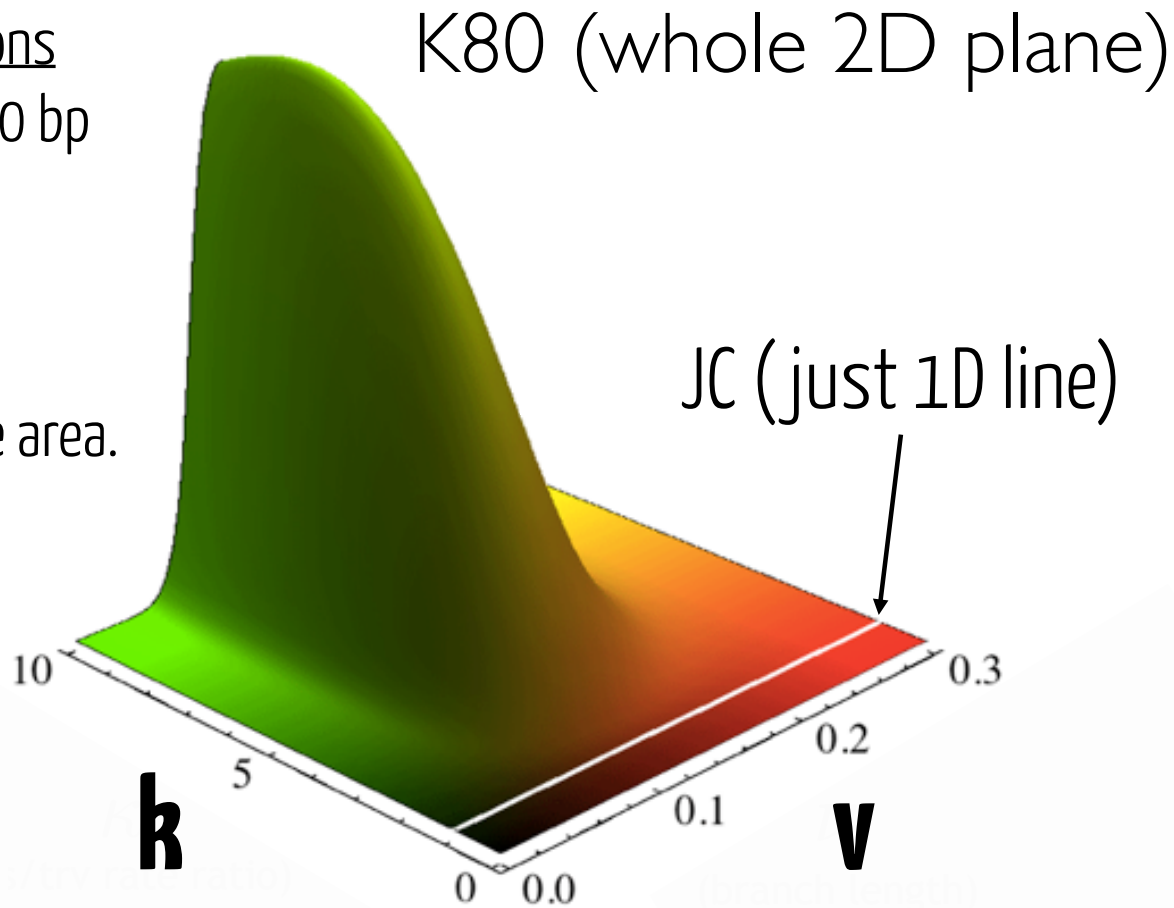
Simulation Conditions

Sequence length: 500 bp

True v : 0.15

True k : 5.0

Prior is flat over whole area.



Marginal Likelihood Example

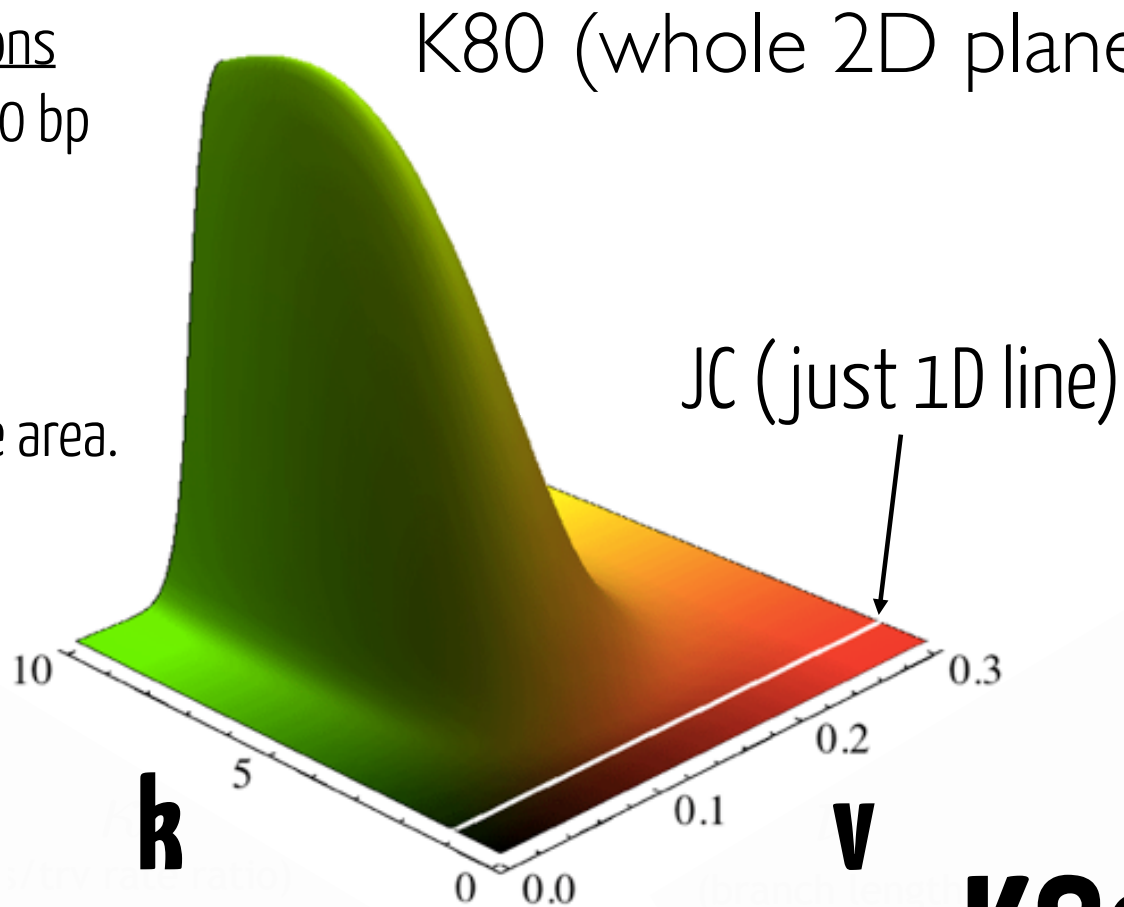
Simulation Conditions

Sequence length: 500 bp

True v : 0.15

True k : **5.0**

Prior is flat over whole area.



K80 wins!

Marginal Likelihood Example

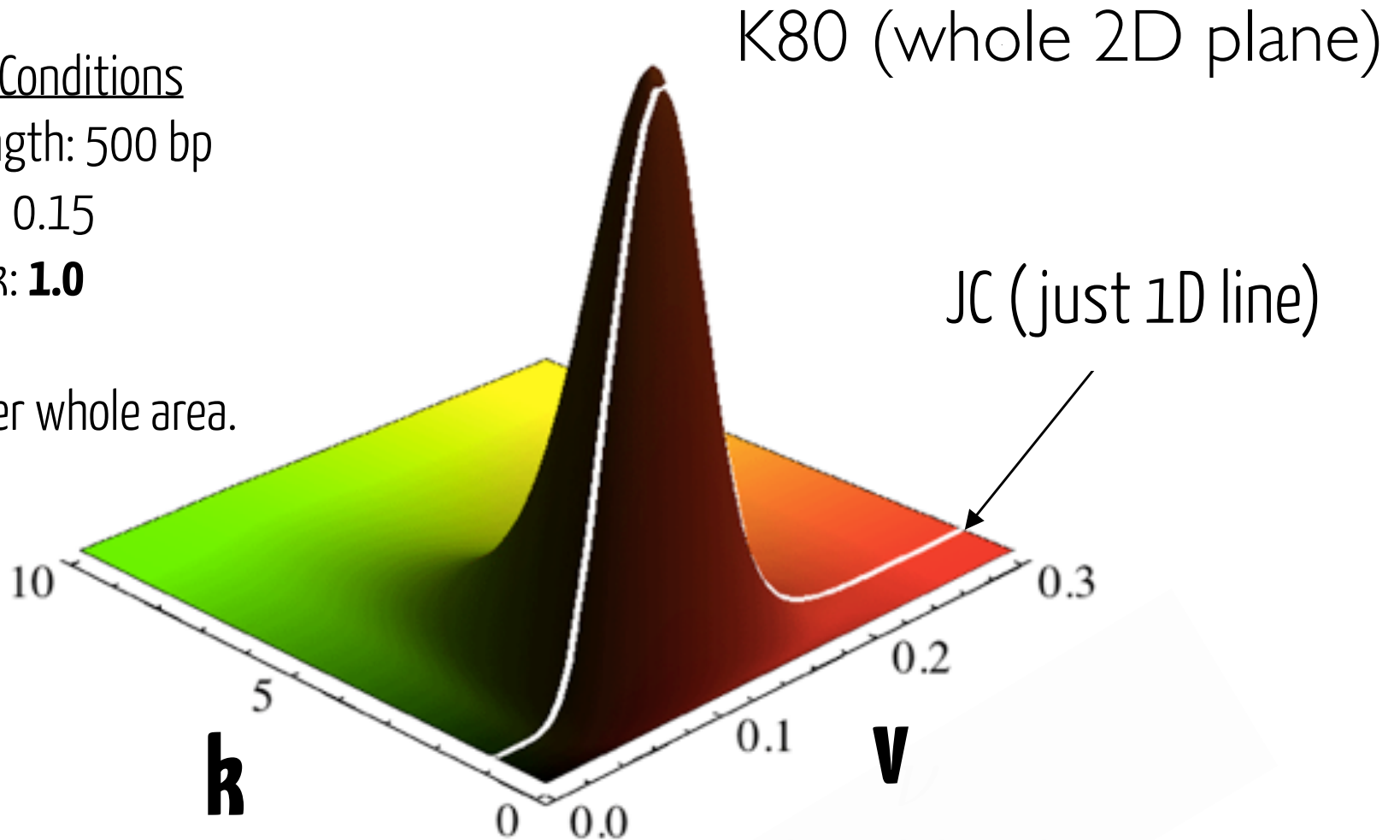
Simulation Conditions

Sequence length: 500 bp

True v : 0.15

True k : **1.0**

Prior is flat over whole area.



Marginal Likelihood Example

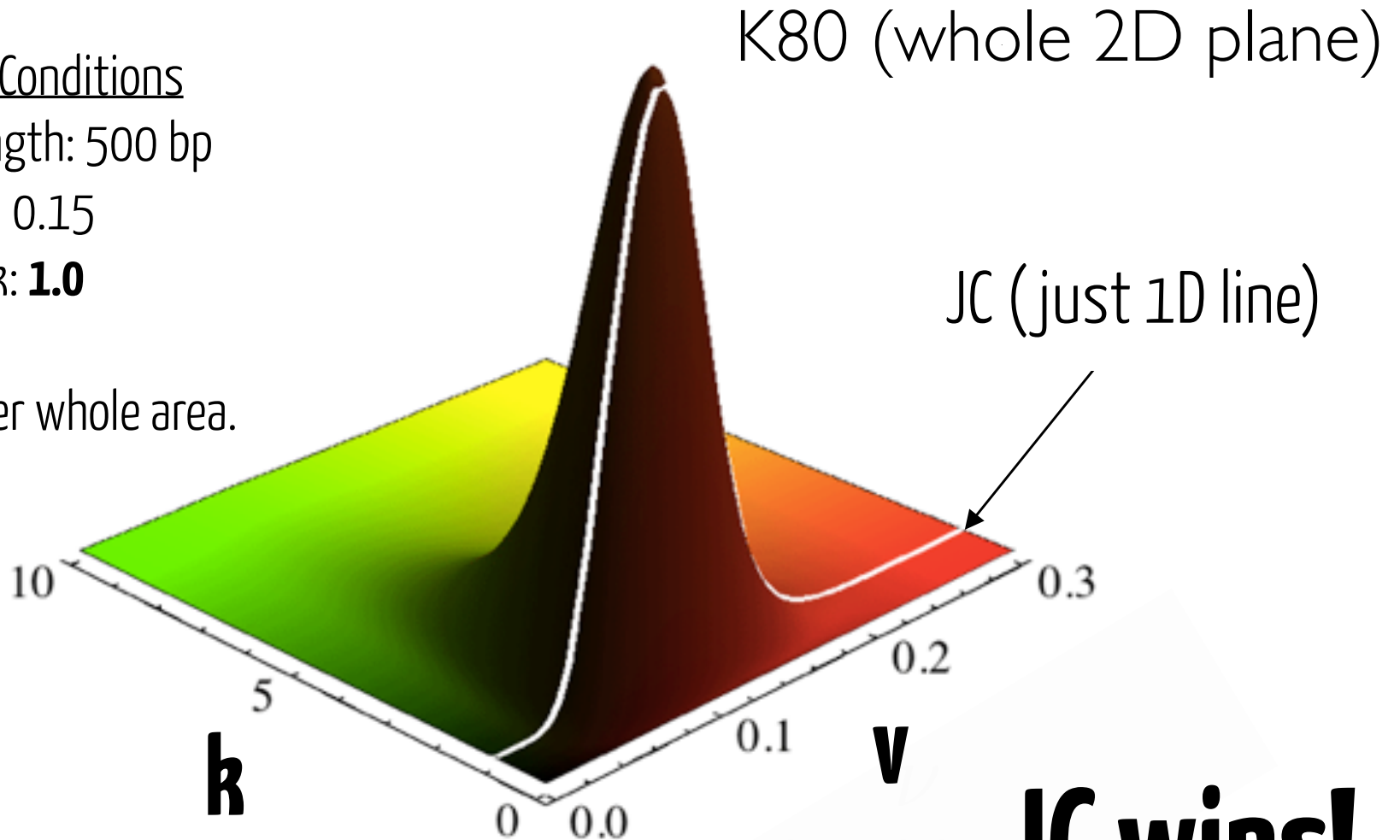
Simulation Conditions

Sequence length: 500 bp

True v : 0.15

True k : **1.0**

Prior is flat over whole area.



JC wins!

Marginal Likelihood Example

Important contrast with ML-based model selection: by marginalizing, rather than maximizing, marginal likelihoods automatically account for extra parameters.

More complicated models can have lower marginal likelihoods.

Marginal Likelihood Example

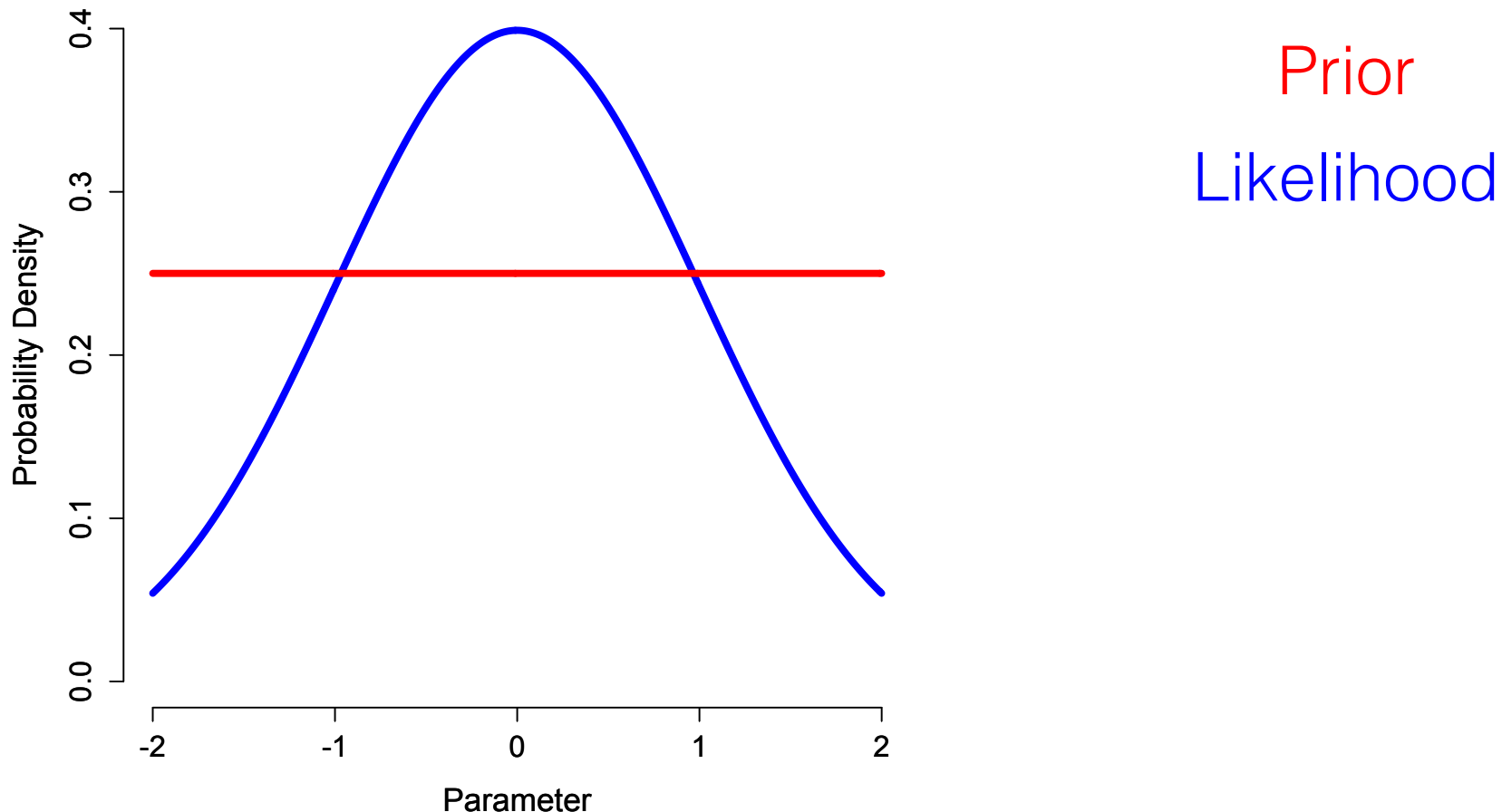
Important contrast with ML-based model selection: by marginalizing, rather than maximizing, marginal likelihoods automatically account for extra parameters.

More complicated models can have lower marginal likelihoods.

But how can we estimate them?

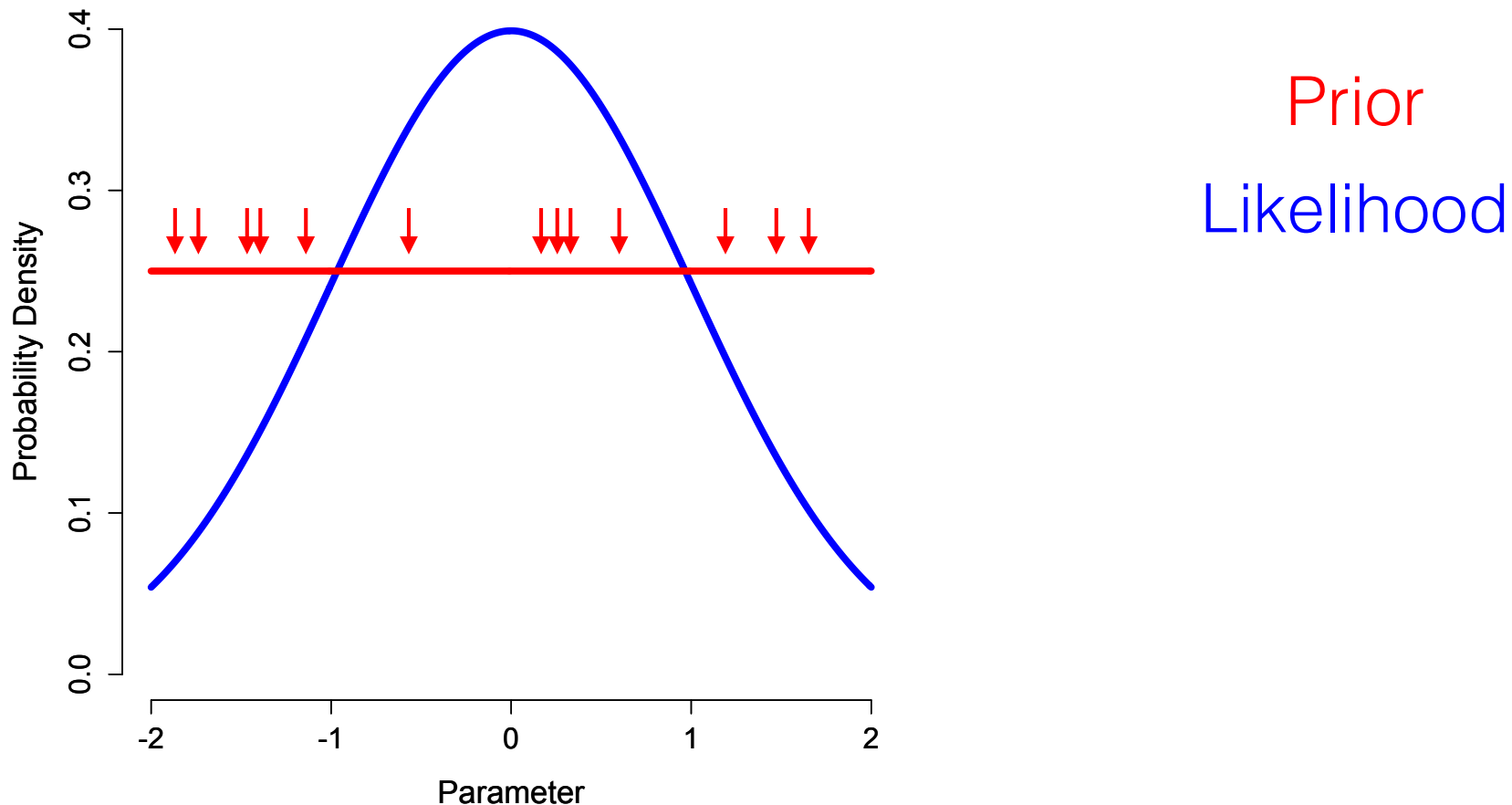
Calculating Marginal Likelihoods

Easy Approach 1 - Sample from the prior



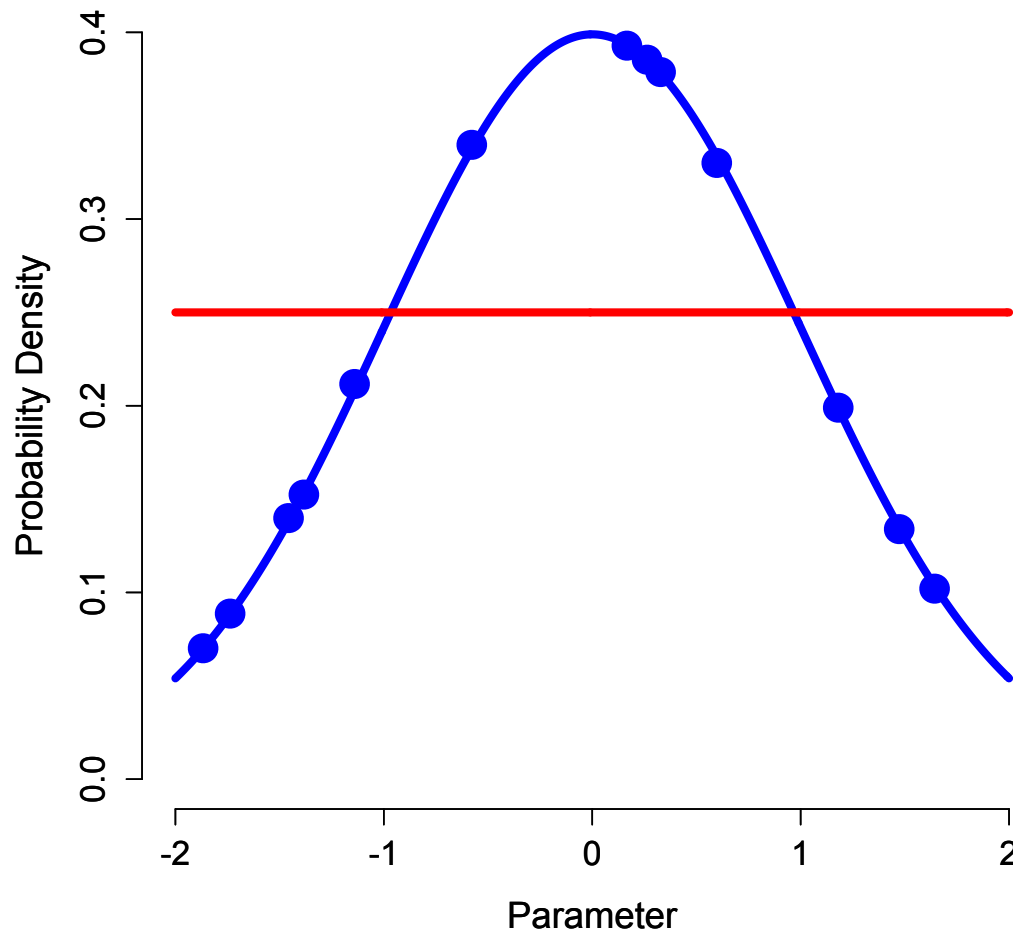
Calculating Marginal Likelihoods

Easy Approach 1 - Sample from the prior



Calculating Marginal Likelihoods

Easy Approach 1 - Sample from the prior



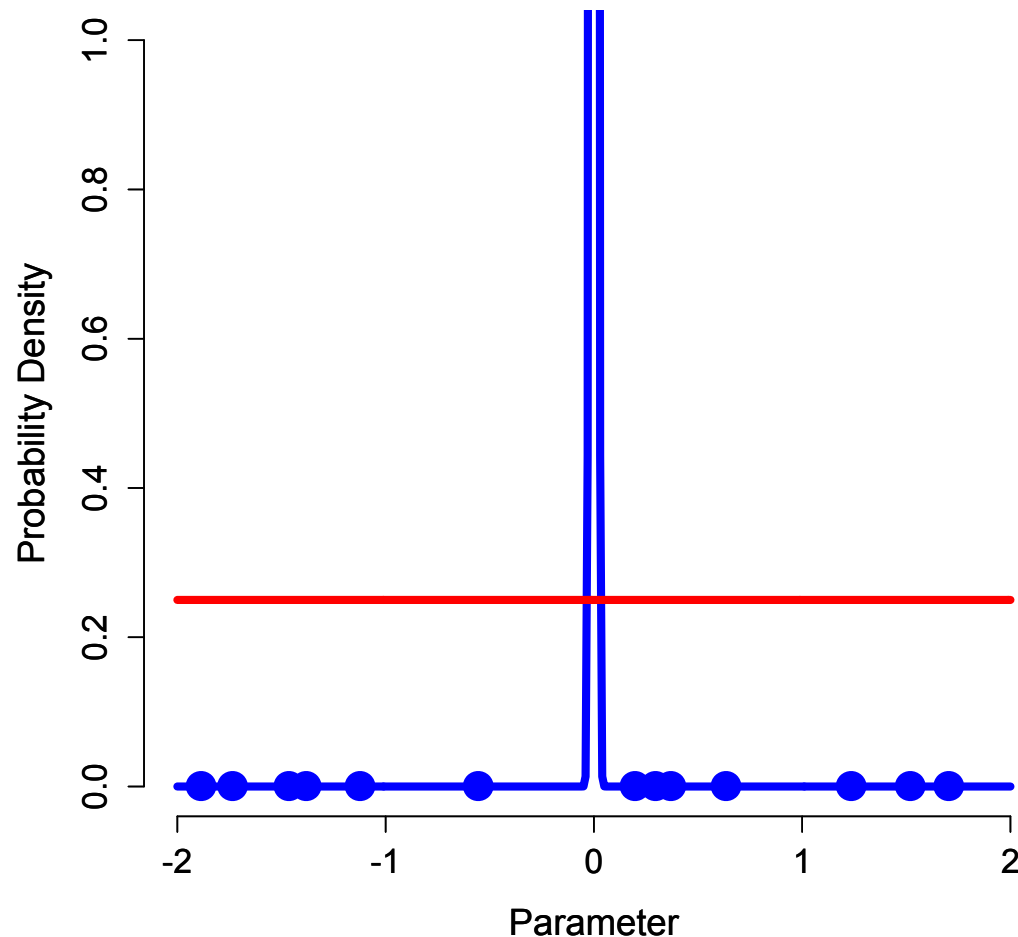
Prior

Likelihood

Take average of blue dots

Calculating Marginal Likelihoods

Easy Approach 1 - Sample from the prior



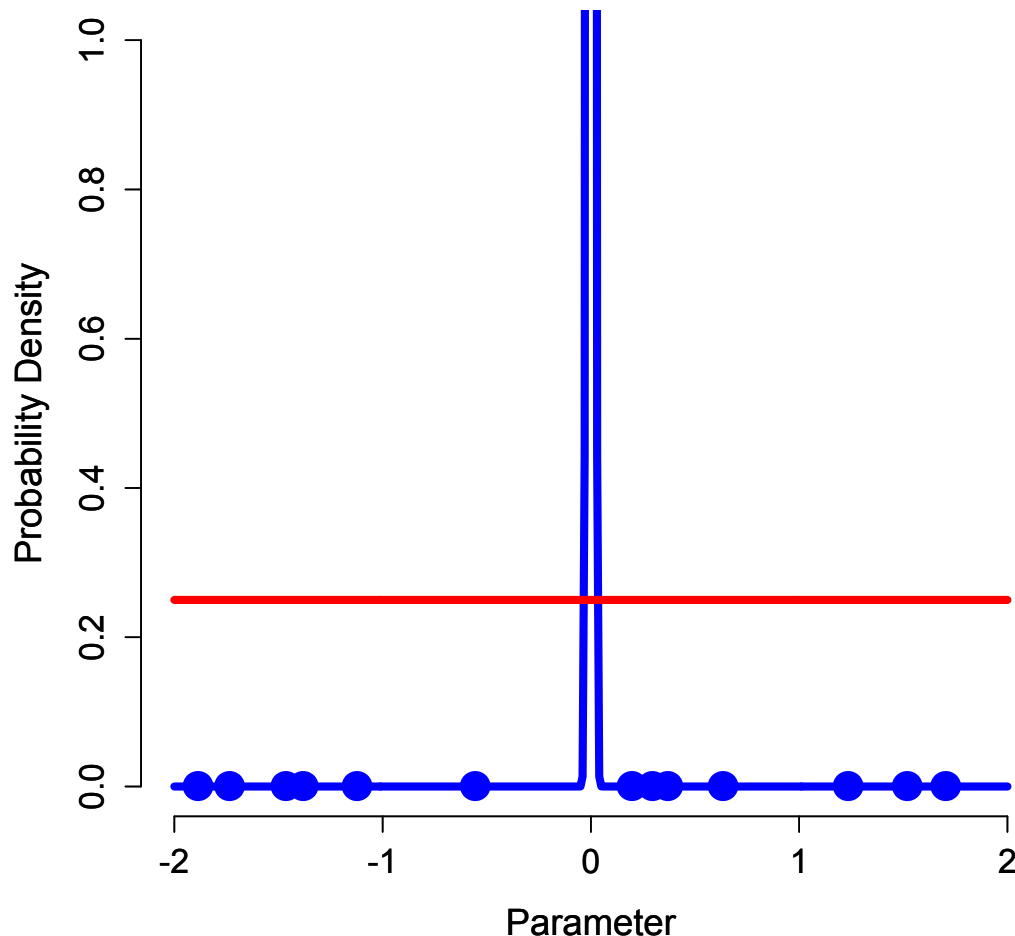
Prior

Likelihood

Take average of blue dots

Calculating Marginal Likelihoods

Easy Approach 1 - Sample from the prior



Prior

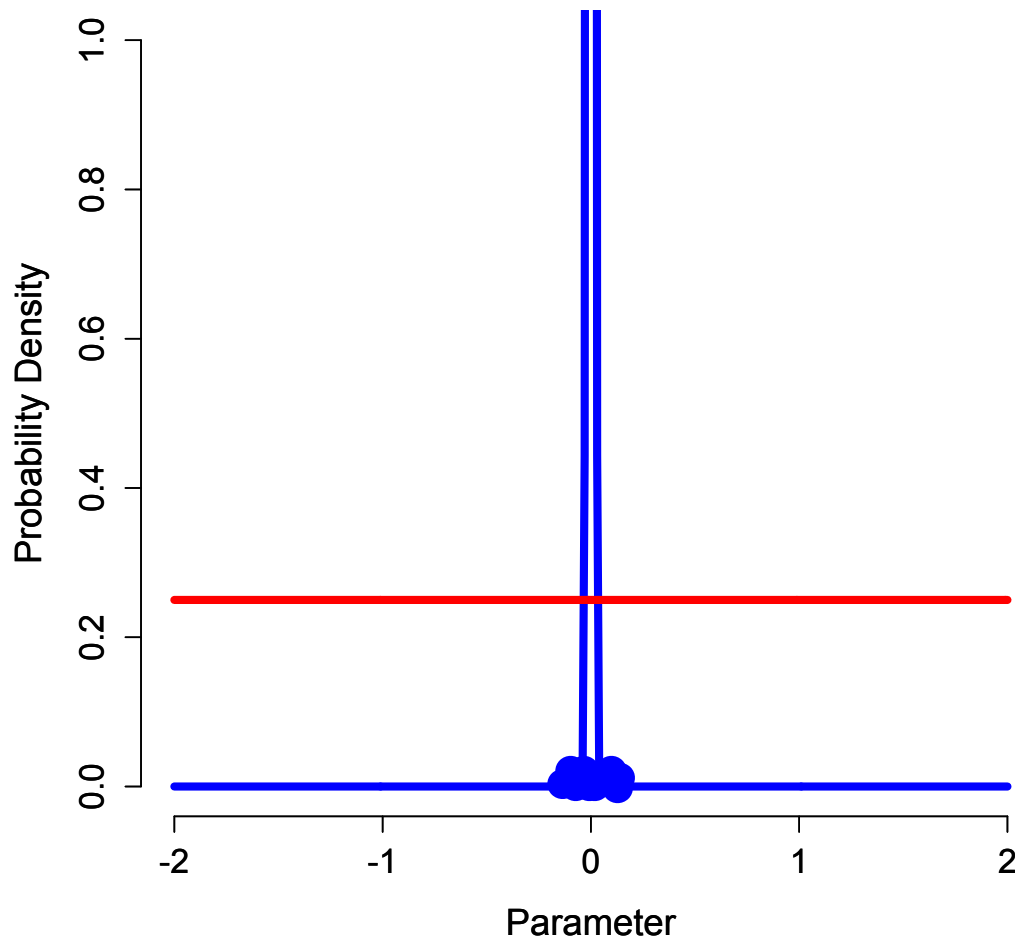
Likelihood

Take average of blue dots??!!

We'd like to make sure we're sampling high likelihood parts of space with reasonable frequency.

Calculating Marginal Likelihoods

Less-Naive Approach 2- Sample from the posterior



Prior

Posterior
(\sim Likelihood)

Since we're supposed to be integrating across the prior, we need to correct for the fact that our samples are from the posterior.

Calculating Marginal Likelihoods

Less-Naive Approach 2- Sample from the posterior

The Harmonic Mean Method

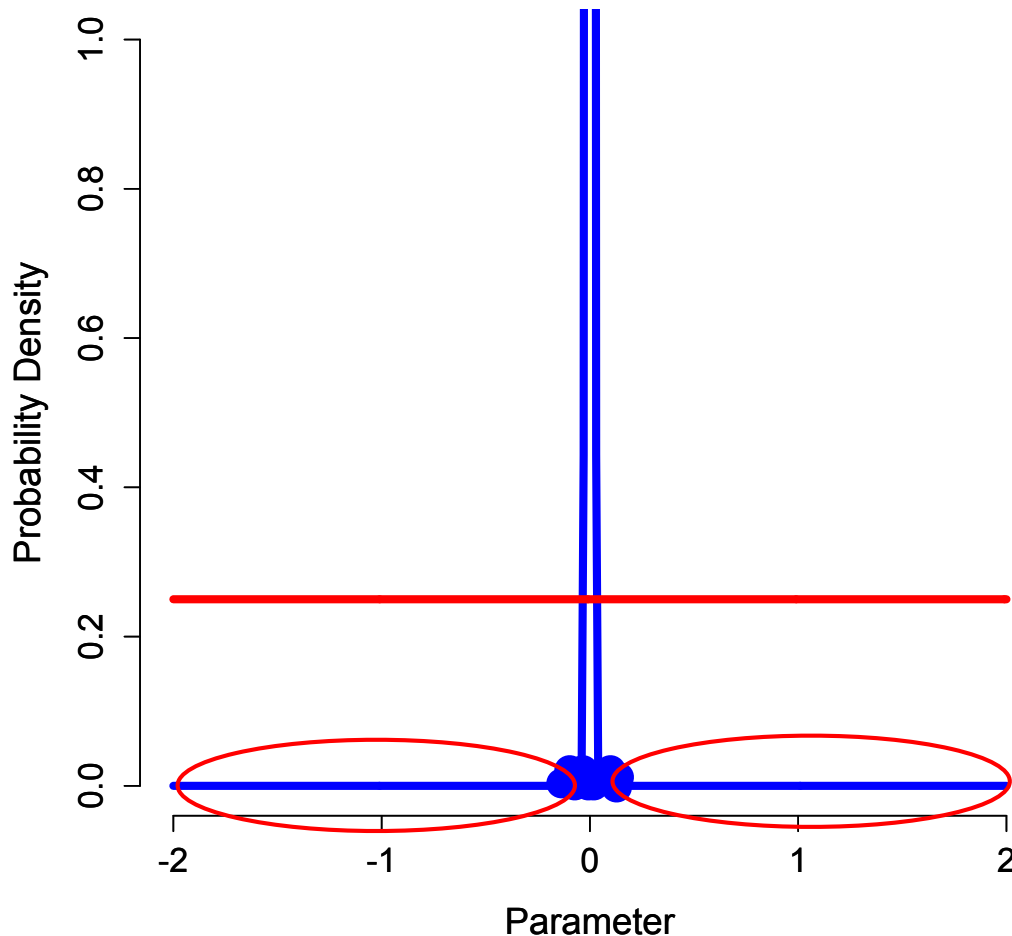
$$\frac{1}{ML} = \frac{1}{L_1} + \frac{1}{L_2} + \dots + \frac{1}{L_3}$$

What's an important property of harmonic means?

Anyone remember discussing bottlenecks in pop gen?

Calculating Marginal Likelihoods

Less-Naive Approach 2- Sample from the posterior



The reverse problem to our first naive approach!

Rarely sampled low likelihoods have a big influence on estimates.

Very unstable.

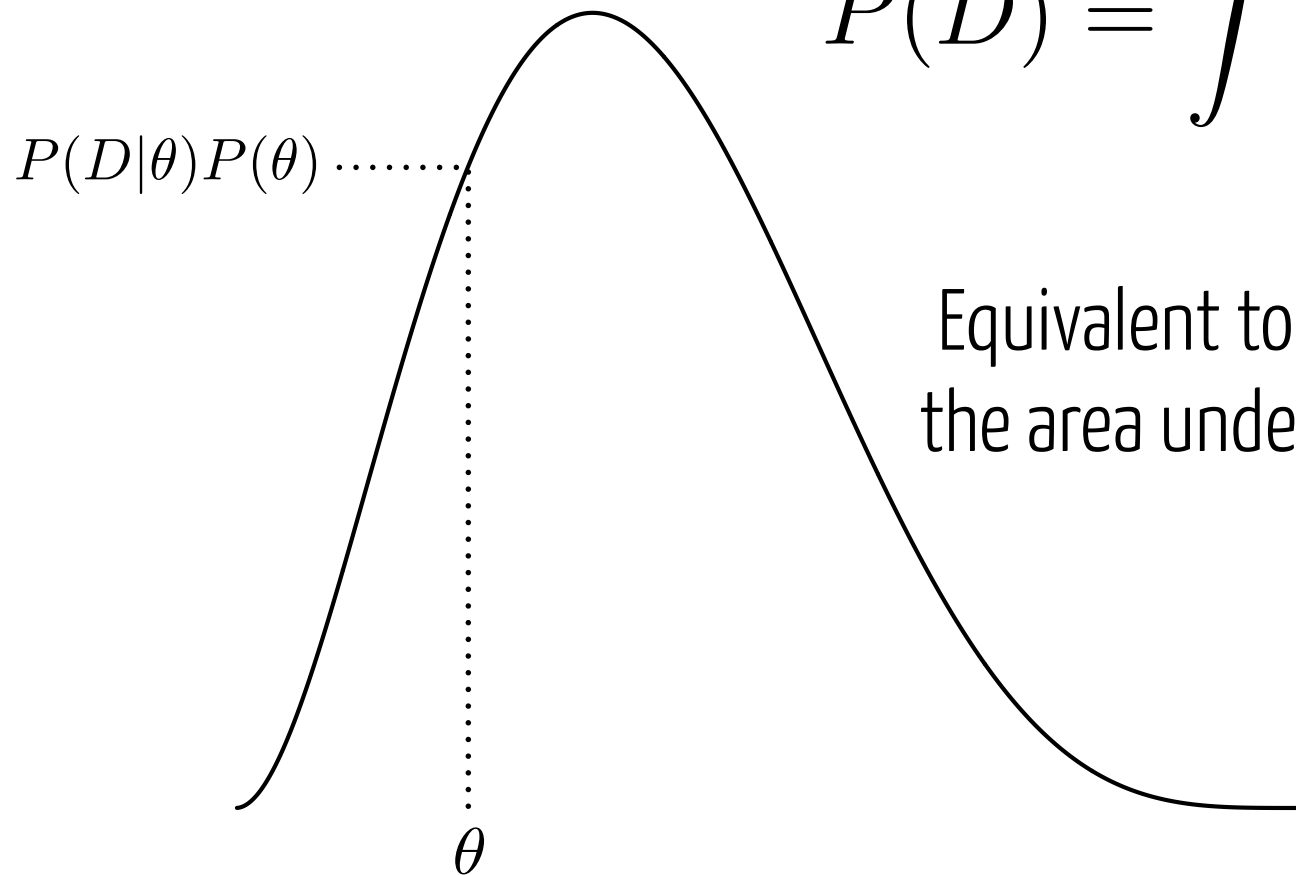
Calculating Marginal Likelihoods

Approach 3 - Sample from a series of distributions

Steppingstone or path sampling

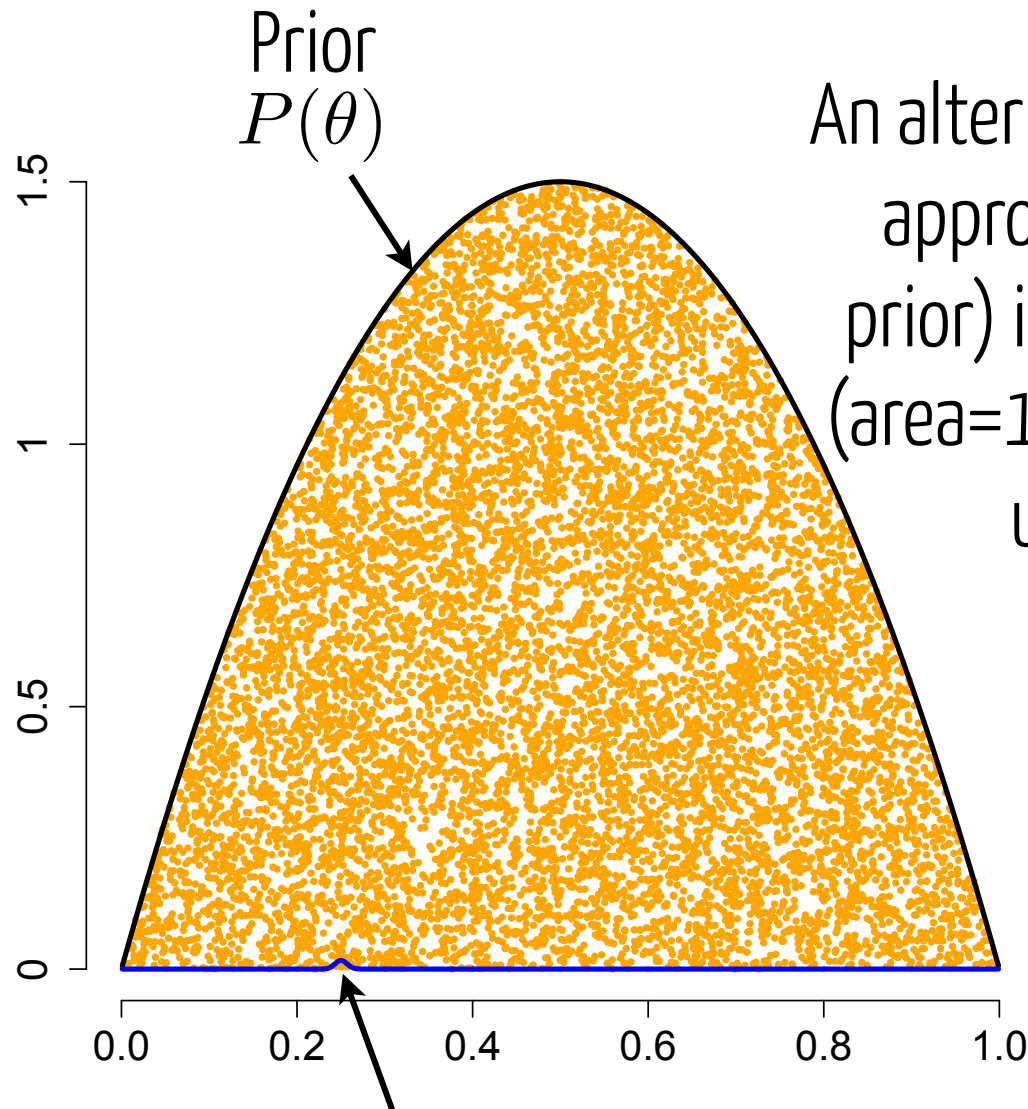
Steppingstone Sampling

$$P(D) = \int P(D|\theta)P(\theta)$$



Equivalent to estimating
the area under this curve.

Steppingstone Sampling

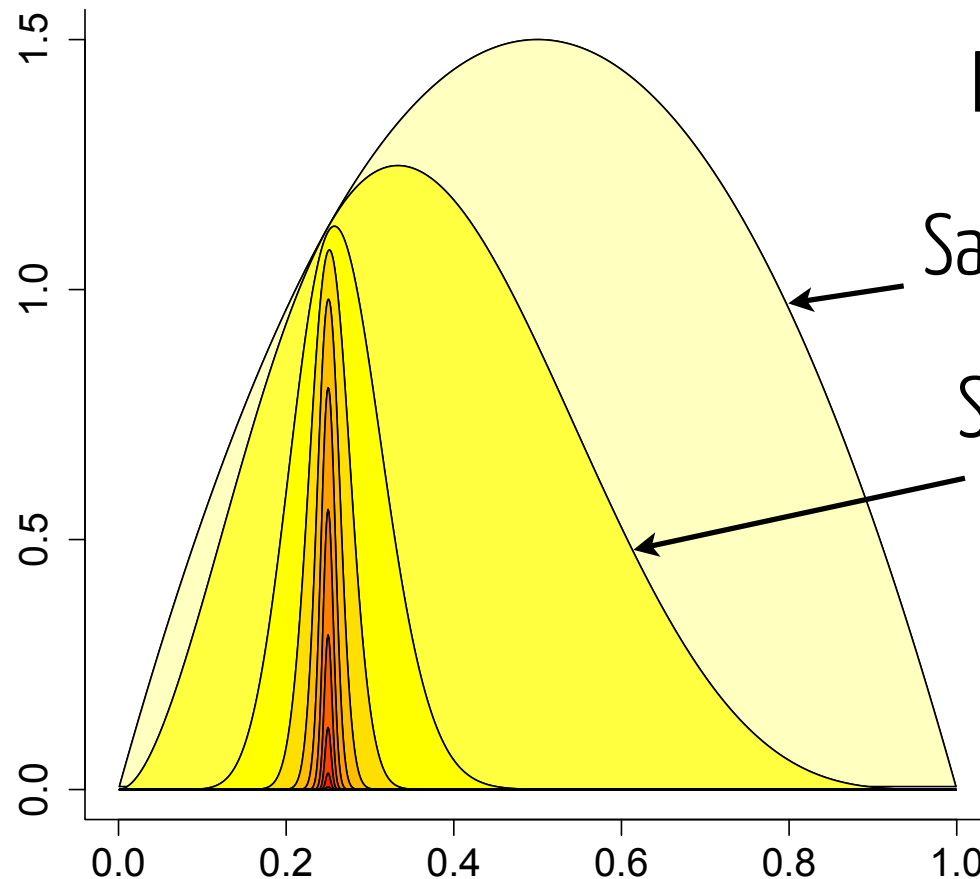


An alternative way to think about our first approach (sampling directly from the prior) is to sample points from the prior (area=1.0), then ask what proportion fall under the curve of interest.

Unfortunately, not many!
As before, this is unstable.

Unnormalized Posterior $\longrightarrow P(D|\theta)P(\theta)$

Steppingstone Sampling

 $P(D)$


Let's try it in **steps!**

Sample from this distribution

See what fraction of samples are under this curve

That fraction is an estimate of this ratio

$$\frac{c_{1.0}}{c_{0.0}} = \left(\frac{c_{1.0}}{\cancel{c_{0.9}}} \right) \left(\frac{\cancel{c_{0.9}}}{\cancel{c_{0.8}}} \right) \left(\frac{\cancel{c_{0.8}}}{\cancel{c_{0.7}}} \right) \left(\frac{\cancel{c_{0.7}}}{\cancel{c_{0.6}}} \right) \left(\frac{\cancel{c_{0.6}}}{\cancel{c_{0.5}}} \right) \left(\frac{\cancel{c_{0.5}}}{\cancel{c_{0.4}}} \right) \left(\frac{\cancel{c_{0.4}}}{\cancel{c_{0.3}}} \right) \left(\frac{\cancel{c_{0.3}}}{\cancel{c_{0.2}}} \right) \left(\frac{\cancel{c_{0.2}}}{\cancel{c_{0.1}}} \right) \left(\frac{\cancel{c_{0.1}}}{c_{0.0}} \right)$$

Power Posteriors

$$\begin{array}{ccc}
 P(D) \swarrow & & \\
 \frac{c_{1.0}}{c_{0.0}} = \left(\frac{c_{1.0}}{\cancel{c_{0.9}}} \right) \left(\frac{\cancel{c_{0.9}}}{\cancel{c_{0.8}}} \right) \left(\frac{\cancel{c_{0.8}}}{\cancel{c_{0.7}}} \right) \left(\frac{\cancel{c_{0.7}}}{\cancel{c_{0.6}}} \right) \left(\frac{\cancel{c_{0.6}}}{\cancel{c_{0.5}}} \right) \left(\frac{\cancel{c_{0.5}}}{\cancel{c_{0.4}}} \right) \left(\frac{\cancel{c_{0.4}}}{\cancel{c_{0.3}}} \right) \left(\frac{\cancel{c_{0.3}}}{\cancel{c_{0.2}}} \right) \left(\frac{\cancel{c_{0.2}}}{\cancel{c_{0.1}}} \right) \left(\frac{\cancel{c_{0.1}}}{c_{0.0}} \right) \\
 \nearrow 1 & &
 \end{array}$$

Posterior

$\beta = 1$

Prior

$\beta = 0$

$$P(D|\theta)_\beta \propto P(D|\theta)^\beta P(\theta)$$

Power Posteriors

$$\frac{c_{1.0}}{c_{0.0}} = \text{Stable estimate of marginal likelihood!}$$



But it requires a **specific type of analysis**,
independent of standard MCMC.

Bayesian Model Selection

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$$P(\theta|D, M) = \frac{P(D|\theta, M)P(\theta|M)}{\int P(D|\theta, M)P(\theta|M)d\theta}$$

Marginal Likelihood

Essentially, the **weighted average likelihood**, weighted by the prior probability of different parameter values.

The Bayes Factor

$$K = \frac{P(D|M_1)}{P(D|M_2)} = \frac{\int P(D|\theta, M_1)P(\theta|M_1)d\theta}{\int P(D|\theta, M_2)P(\theta|M_2)d\theta}$$

Ratio of the probability of the data under two models

The Bayes Factor

$$K = \frac{P(D|M_1)}{P(D|M_2)} = \frac{\int P(D|\theta, M_1)P(\theta|M_1)d\theta}{\int P(D|\theta, M_2)P(\theta|M_2)d\theta}$$

Ratio of the probability of the data under two models

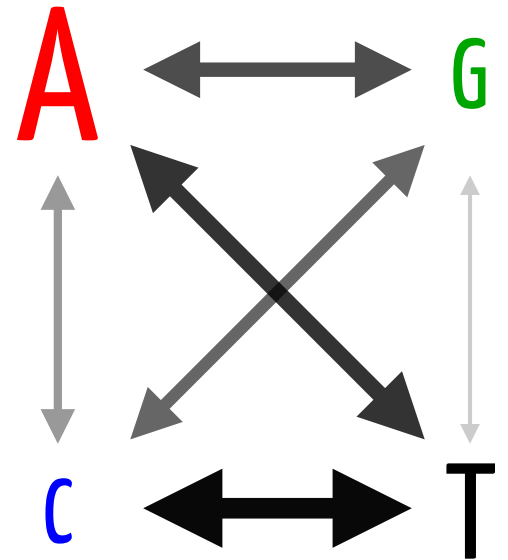
Note that this is related to the Likelihood ratio test

Bayes Factors

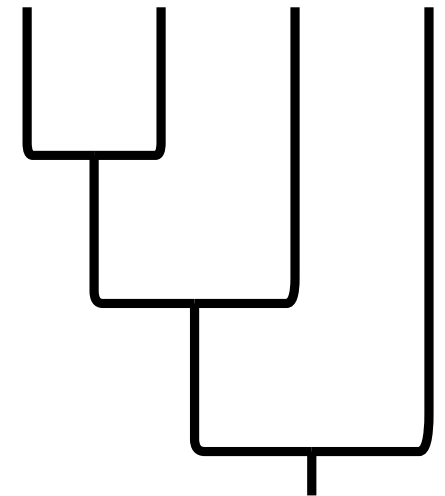
$2\ln(\text{BF})$	BF	Strength of evidence
0-2	1-3	Barely worth mentioning.
2-6	3-20	Positive
6-10	20-150	Strong
>10	>150	Very Strong

Bayes Factors

For now, we're going to use these to **compare different models of sequence evolution** as our hypotheses.



However, BFs can also be used for other hypotheses, like **partition models (later)**, topological relationships, and much more.



Or...don't choose a model!

Reversible Jump MCMC

Instead of picking a model, include MCMC moves that jump between them. Integrate out uncertainty about which model is best. This is a Bayesian form of **model averaging**.

We already do this for trees. Can also do this for models.

Or...don't choose a model!

Reversible Jump MCMC

Instead of picking a model, include MCMC moves that jump between them. Integrate out uncertainty about which model is best. This is a Bayesian form of **model averaging**.

We already do this for trees. Can also do this for models.

*Disclaimer: Setting up proper reversible jump moves can often be **very challenging**.