Cultural evolution

Remco R. Bouckaert
r.bouckaert@auckland.ac.nz

Centre of Computational Evolution, University of Auckland

Online, June 2021

Topics

- **Cultural evolution**
- Phylogeography
- Model selection
- Model comparison

Encoding word lists as cognates

word list

language	hand	mother	father	...
English	hand	mother	father	...
Dutch	hand	moeder	vader	...
German	hand	mutter	vater	...
French	main	mère	père	...
Spanish	mano	madre	padre	...
Dhudhuroa	?	papa	mama	...

Encoding word lists as cognates

word list

language	hand	mother	father	...
English	hand	mother	father	...
Dutch	hand	moeder	vader	...
German	hand	mutter	vater	...
French	main	mère	père	...
Spanish	mano	madre	padre	...
Dhudhuroa	?	papa	mama	...

cognate list

language	hand	mano	mother	papa	father	mama	...
English	1	0	1	0	1	0	...
Dutch	1	0	1	0	1	0	...
German	1	0	1	0	1	0	...
French	0	1	1	0	1	0	...
Spanish	0	1	1	0	1	0	...
Dhudhuroa	?	?	0	1	0	1	...

Adding ascertainment correction

$$P(D|T\&D \neq 0) =$$

Adding ascertainment correction

$$P(D|T\&D \neq 0) = \frac{P(D\&D \neq 0|T)}{P(D \neq 0|T)}$$

Adding ascertainment correction

$$P(D|T\&D \neq 0) = \frac{P(D\&D \neq 0|T)}{P(D \neq 0|T)} = \frac{P(D|T)}{1 - P(D = 0|T)}$$

Adding ascertainment correction

$$P(D|T\&D \neq 0) = \frac{P(D\&D \neq 0|T)}{P(D \neq 0|T)} = \frac{P(D|T)}{1 - P(D = 0|T)}$$

cognate list

language	hand	mano	mother	papa	father	mama	...
English	1	0	1	0	1	0	...
Dutch	1	0	1	0	1	0	...
German	1	0	1	0	1	0	...
French	0	1	1	0	1	0	...
Spanish	0	1	1	0	1	0	...
Dhudhuroa	?	?	0	1	0	1	...

cognate list + ascertainment

language	ascertainment	hand	mano	ascertainment	mother	papa	ascertainment	father	mama
English	0	1	0	0	1	0	0	1	0
Dutch	0	1	0	0	1	0	0	1	0
German	0	1	0	0	1	0	0	1	0
French	0	0	1	0	1	0	0	1	0
Spanish	0	0	1	0	1	0	0	1	0
Dhudhuroa	?	?	?	0	0	1	0	0	1

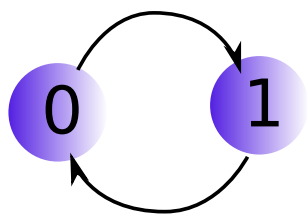
Cognate substitution models: CTMC

Simplest model: continuous time Markov chain model

$$\begin{matrix} 0 \\ 1 \end{matrix} \begin{pmatrix} - & 1 \\ 1 & - \end{pmatrix} \times \begin{pmatrix} f_0 \\ f_1 \end{pmatrix} = \begin{pmatrix} - & f_1 \\ f_0 & - \end{pmatrix} = Q$$

f_0, f_1 equilibrium frequency of a 0 or 1 respectively

$$P(x_i = j | x_{\pi_i} = j, t, \theta) = e^{tQ}_{j,k}$$



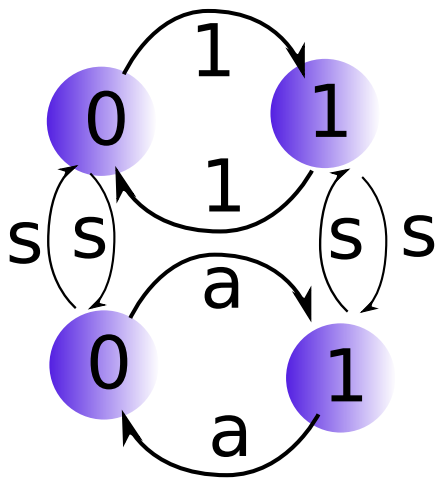
Cognate substitution models: Covarion

0 in alignment is either slow 0 or fast 0

1 in alignment is either slow 1 or fast 1

$$\begin{array}{l} \text{fast} \begin{cases} 0: \\ 1: \end{cases} \\ \text{slow} \begin{cases} 0: \\ 1: \end{cases} \end{array} \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{pmatrix} - & 1 & s & 0 \\ 1 & - & 0 & s \\ s & 0 & - & \alpha \\ 0 & s & \alpha & - \end{pmatrix} \times \begin{pmatrix} f_0 \\ f_1 \\ f_0 \\ f_1 \end{pmatrix} = \begin{pmatrix} - & f_1 & sf_0 & 0 \\ f_0 & - & 0 & sf_1 \\ sf_0 & 0 & - & \alpha f_1 \\ 0 & sf_1 & \alpha f_0 & - \end{pmatrix} = Q$$

- f_0, f_1 equilibrium frequency of a 0 or 1 respectively
- s switch rate between fast and slow
- α slow mutation rate



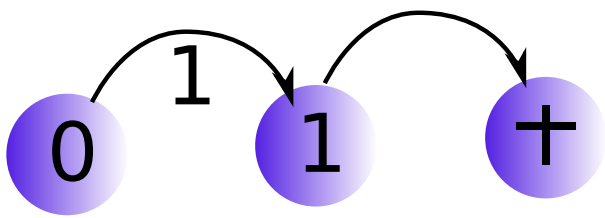
Cognate substitution models: Stochastic Dollo

Dollo principle: every trait appears only once, but can die out many times

New features appear according to a Poisson process with rate r

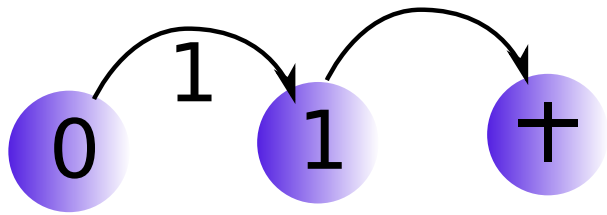
$$\begin{array}{c} \\ 0: \\ 1: \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} - & 0 \\ \mu & - \end{bmatrix} \end{array} = Q$$

μ rate of extinction



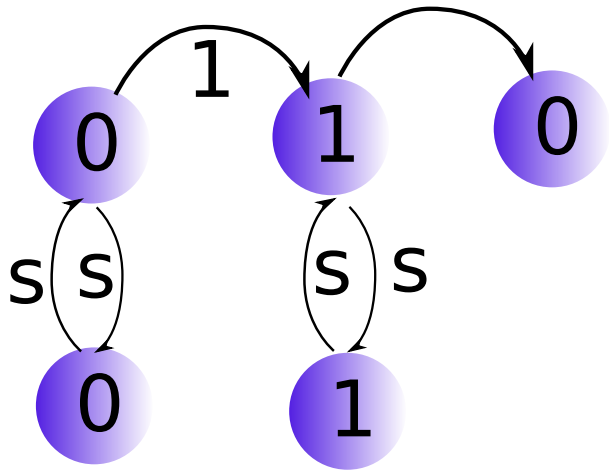
Cognate substitution models: Pseudo Dollo

As Stochastic Dollo, but allow multiple birth events



Cognate substitution models: Pseudo Dollo Covarion

As Pseudo Dollo, but add slow-0 and slow-1 state



Rate heterogeneity

Meaning classes with many cognates require more mutations than cognate classes with fewer mutations => higher rates

Common options:

- all meaning classes have same rate
- all meaning classes have own rate
  - potentially many parameters
  - use model selection
- use gamma rate heterogeneity
  - Note: makes no sense with covarion

Potentially: share relative rate or gamma rate among meaning classes (as in SubstBMA)

Multi state encoding

word list

language	hand	mother	father	...
English	hand	mother	father	...
Dutch	hand	moeder	vader	...
German	hand	mutter	vater	...
French	main	mère	père	...
Spanish	mano	madre	padre	...
Dhudhuroa	?	papa	mama	...
Albanian	dorë”	nëna	babai	...

Multi state encoding

word list

language	hand	mother	father	...
English	hand	mother	father	...
Dutch	hand	moeder	vader	...
German	hand	mutter	vater	...
French	main	mère	père	...
Spanish	mano	madre	padre	...
Dhudhuroa	?	papa	mama	...
Albanian	dorë"	nëna	babai	...

encoding

language	hand	mother	father	
English	1	0	1	...
Dutch	1	0	1	...
German	1	0	1	...
French	0	0	1	...
Spanish	0	0	1	...
Dhudhuroa	?	1	0	...
Albanian	2	2	2	...

Might want to add "out of data set" state, i.e. 4 states for the example above

Multi state substitution models

Mk model (MM package)

- For structural data (linguistics), morphological data (biology)
- Generalisation of Jukes Cantor: for a trait with k traits, the $k \times k$ rate matrix has rates all equal
- MKv model: MK with ascertainment correction for the trait being present at least once

Ordinal (Babel package)

- For ordered data
- Rate matrix: all off-diagonal rates are 1, the remainder 0

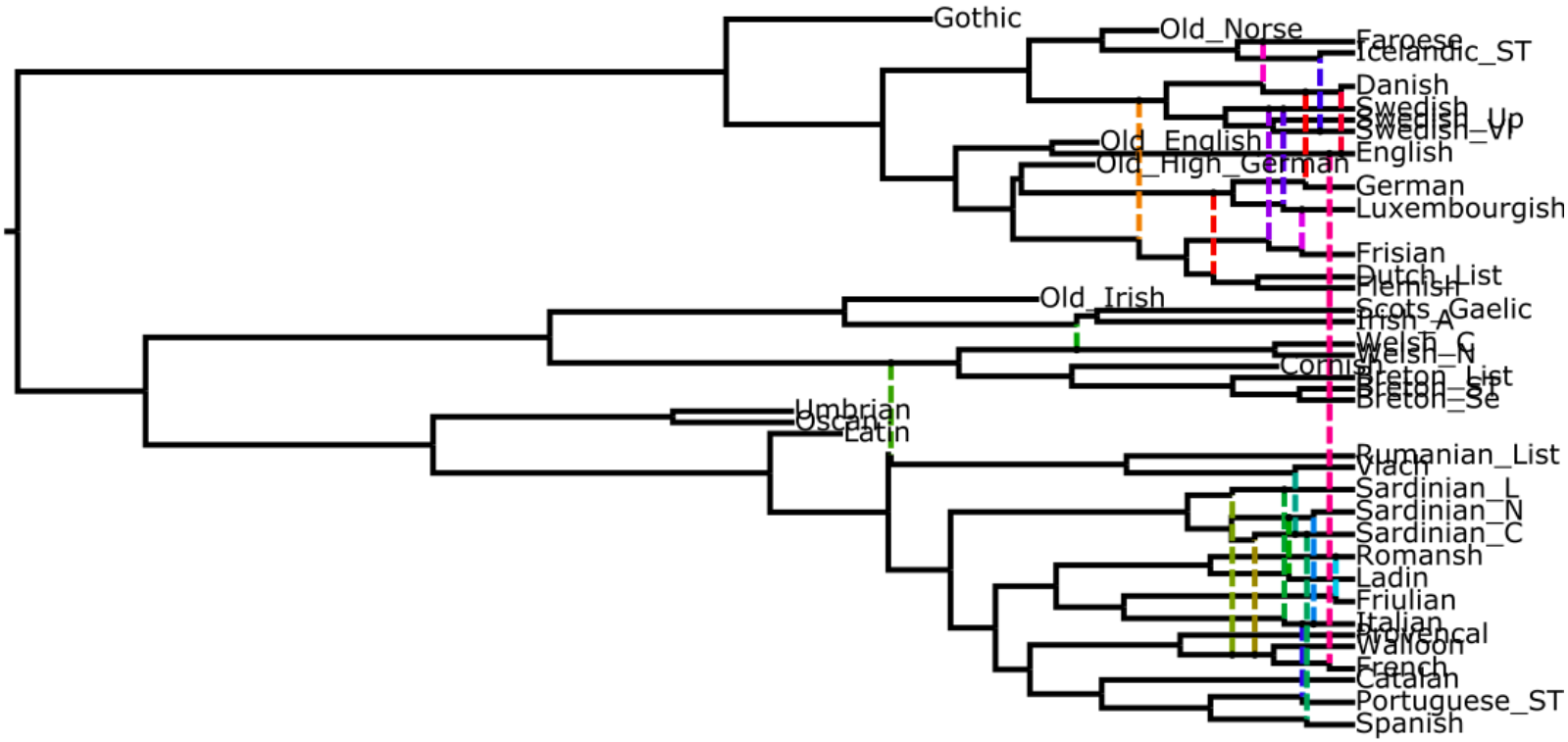
Clock model

- Strict clock
- Optimised Relaxed Clock (ORC)

- Only use ORC when sufficient amount of data is available
- Use model selection to find best clock model

Dealing with borrowing

- Ignore
- Contact trees Nico Neureiter, 2021?



Cultural evolution summary

Encoding:

- cognate classes vs multi-state
- ascertainment correction

Models:

- CTMC
- Covarion
- Stochastic Dollo
- Pseudo Dollo (covarion)

With/without rate heterogeneity per meaning class

Phylogeography

Remco R. Bouckaert
r.bouckaert@auckland.ac.nz

Centre of Computational Evolution, University of Auckland

Online, June 2021

Topics

- Cultural evolution
- **Phylogeography**
- Model selection
- Model comparison

Phylogeography

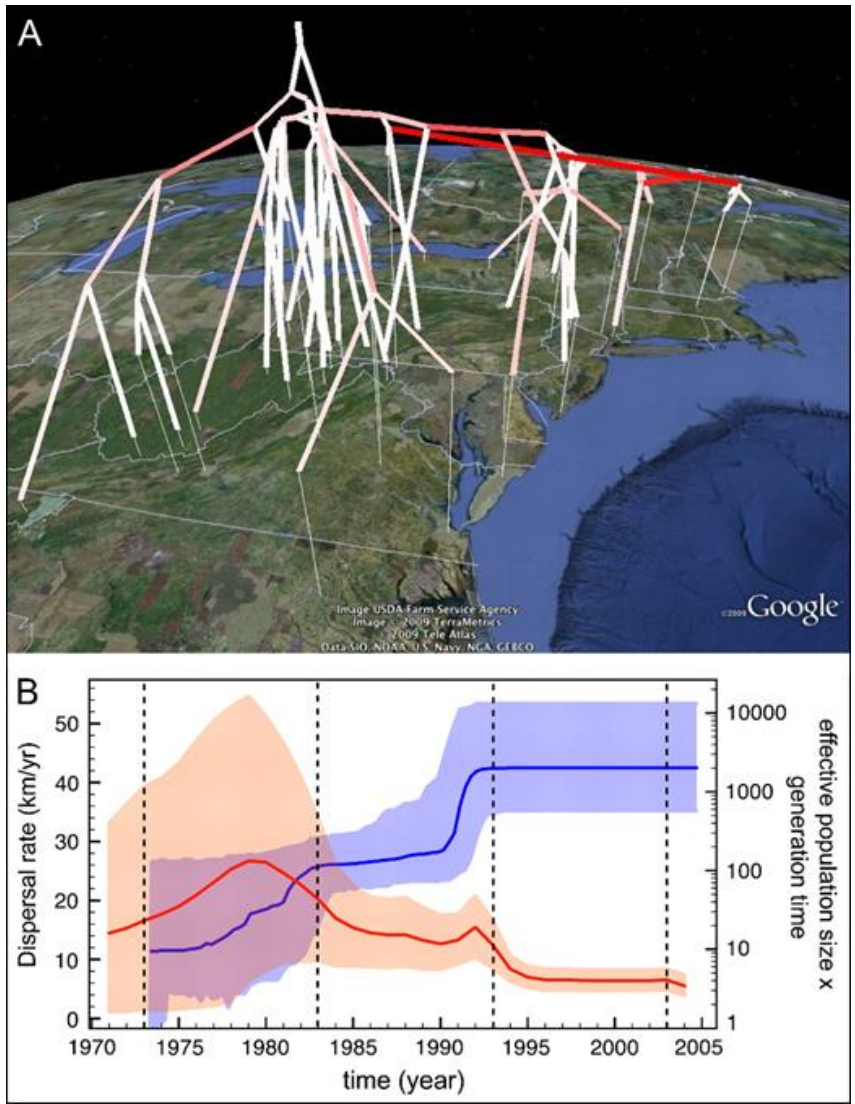
Discrete state vs Continuous state (latitude/longitude locations)

- continuous contains more information
- continuous makes strong assumptions on dispersal process
- discrete when random walk does not apply (e.g. virus dispersal through air travel)

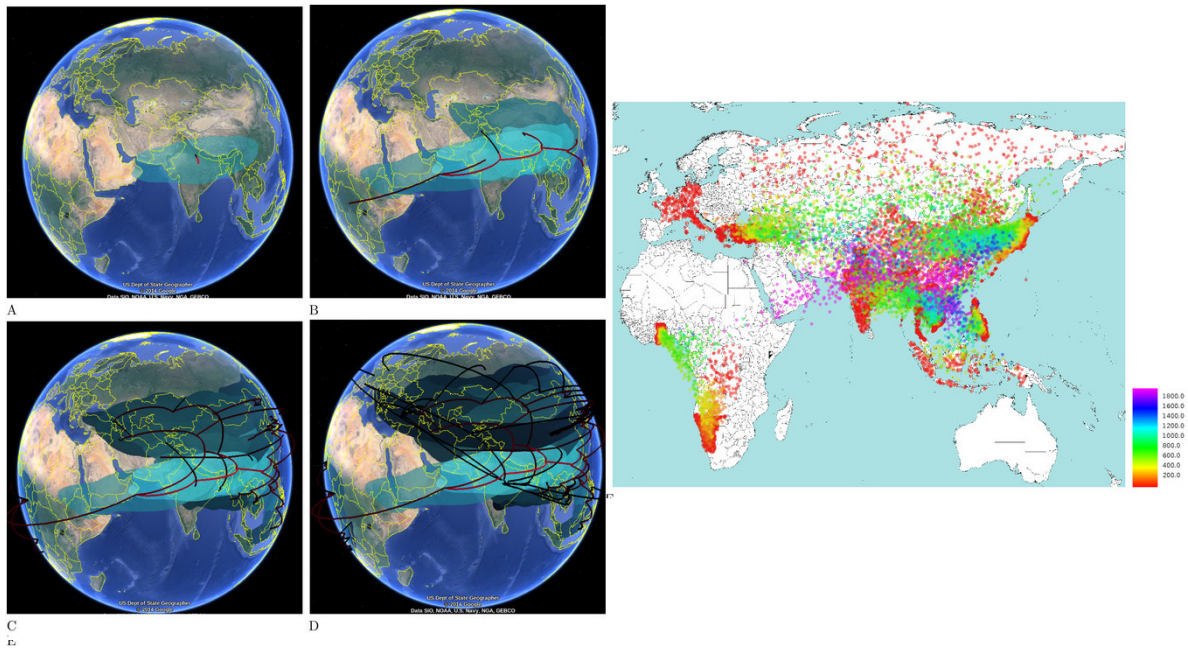
Diffusion on a plane Lemey et al, MBE, 2010

Random walk on plane

- Brownian motion
- Cauchy: big tail allows larger jumps



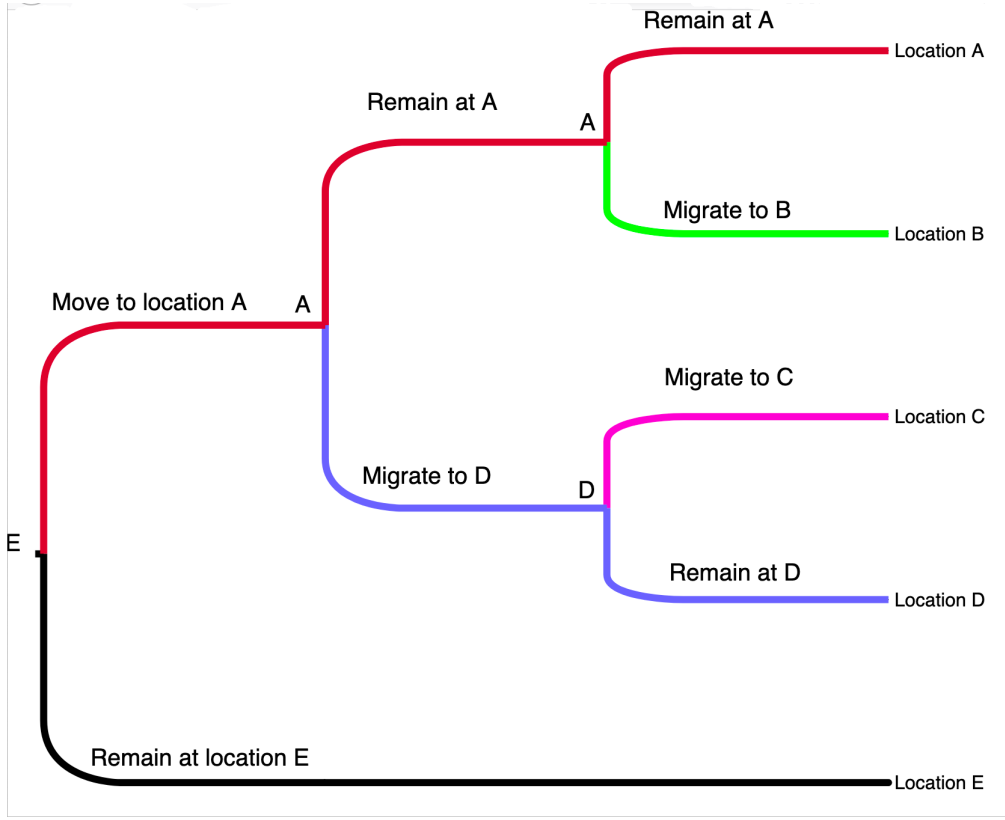
Diffusion on a sphere Bouckaert, PeerJ, 2016



- Takes curvature of earth in account
- Integrates out internal node locations: more efficient
- Allows location priors

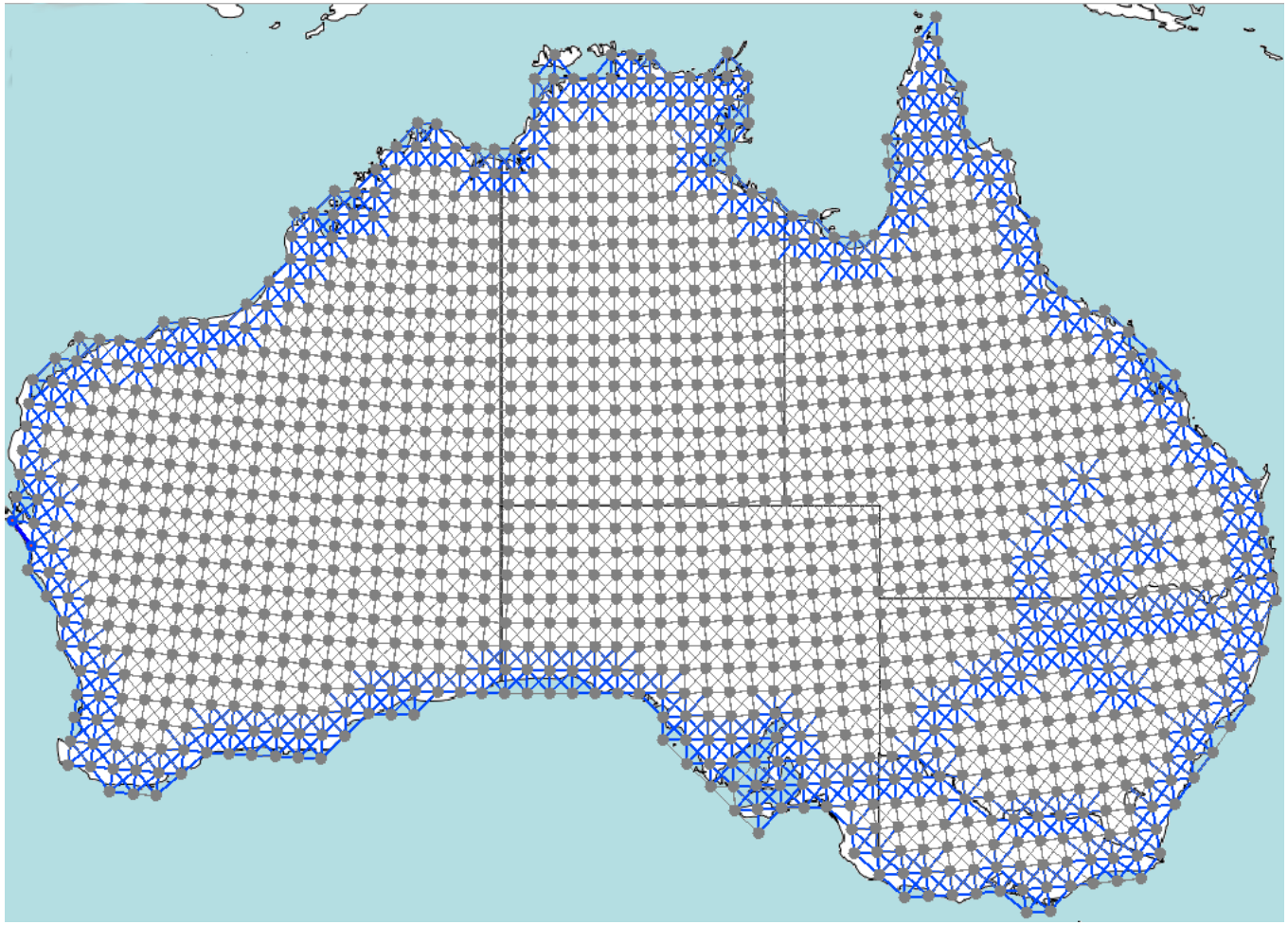
Tutorial: `beast2.org/tutorials` geography on a sphere

Break-away model Bouckaert et al, Nature E&E, 2018



howto: `http://www.beast2.org/2018/03/12/break-away-phylogeography.html`

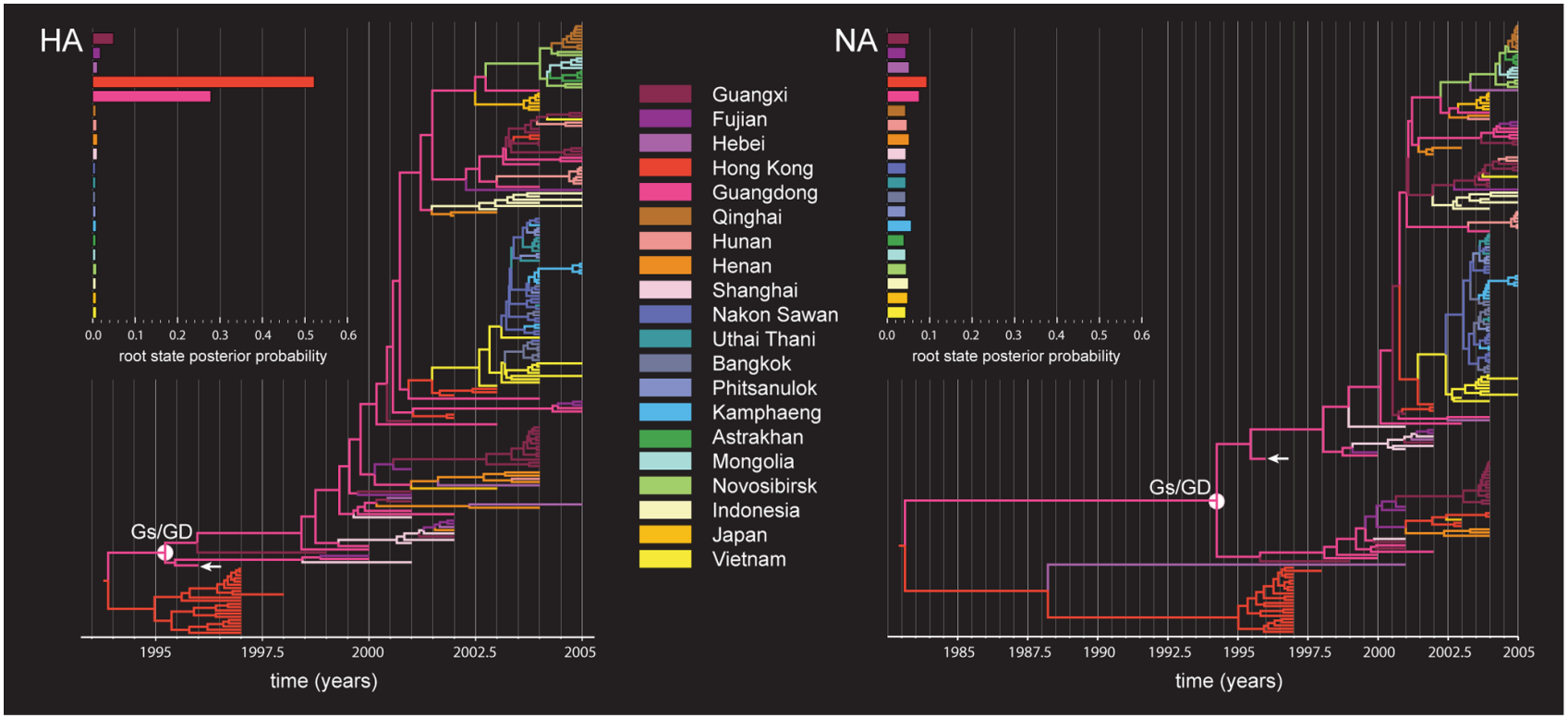
Landscape aware model Bouckaert et al, Nature E&E, 2018



Works with diffusion and break-away models

https://cpb-ap-se2.wpmucdn.com/blogs.auckland.ac.nz/dist/e/360/files/2018/02/movie-21nivdy.gif

Ancestral reconstruction Lemey et al, PloS Comp Bio, 2009



Tutorial: `http://beast2.org/tutorials` Ancestral reconstruction/discrete phylogeography

Model averaging: stochastic variable selection

Use indicator variable to select model

- Example: ancestral state reconstruction using mask matrix I and rate matrix R

$$I = \begin{pmatrix} - & i_{12} & i_{13} & i_{14} \\ i_{21} & - & i_{23} & i_{24} \\ i_{31} & i_{32} & - & i_{34} \\ i_{41} & i_{42} & i_{43} & - \end{pmatrix} \qquad R = \begin{pmatrix} - & r_{12} & r_{13} & r_{14} \\ r_{21} & - & r_{23} & r_{24} \\ r_{31} & r_{32} & - & r_{34} \\ r_{41} & r_{42} & r_{43} & - \end{pmatrix}$$

- Use r_{ij} is i_{ij} is true, but use rate 0 if i_{ij} is false
- Sample I and all rates in R throughout MCMC run.
- Use strong prior on number of $i_{ij} = \mathit{true}$ to reduce number of non-zero rates

Lemey et al, PLoS Comput Biol, 2009

Model averaging: stochastic variable selection

Use indicator variable to select model

- Example: ancestral state reconstruction using mask matrix I and rate matrix R

$$I = \begin{pmatrix} - & i_{12} & i_{13} & i_{14} \\ i_{21} & - & i_{23} & i_{24} \\ i_{31} & i_{32} & - & i_{34} \\ i_{41} & i_{42} & i_{43} & - \end{pmatrix} \qquad R = \begin{pmatrix} - & r_{12} & r_{13} & r_{14} \\ r_{21} & - & r_{23} & r_{24} \\ r_{31} & r_{32} & - & r_{34} \\ r_{41} & r_{42} & r_{43} & - \end{pmatrix}$$

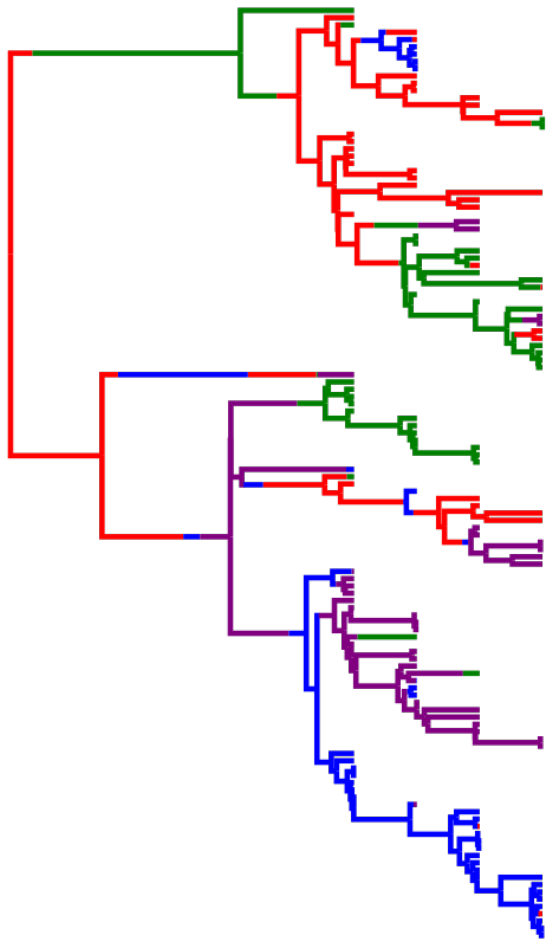
- Use r_{ij} is i_{ij} is true, but use rate 0 if i_{ij} is false
- Sample I and all rates in R throughout MCMC run.
- Use strong prior on number of $i_{ij} = \mathit{true}$ to reduce number of non-zero rates

Lemey et al, PLoS Comput Biol, 2009

Stochastic variable selection:

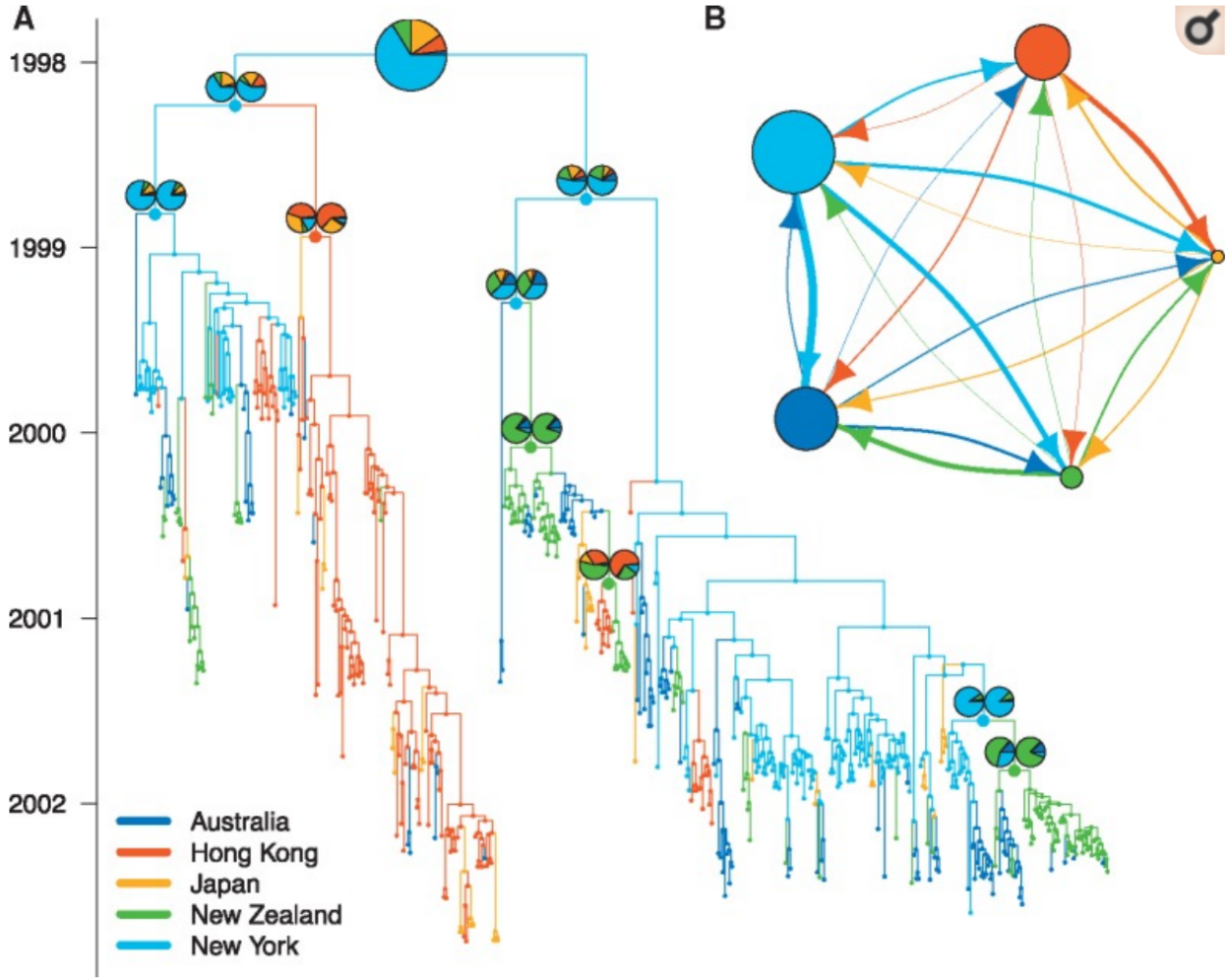
- Simple to implement
- Potentially inefficient in sampling unused parameters

Structured coalescent Vaughan et al, BioInf. 2009



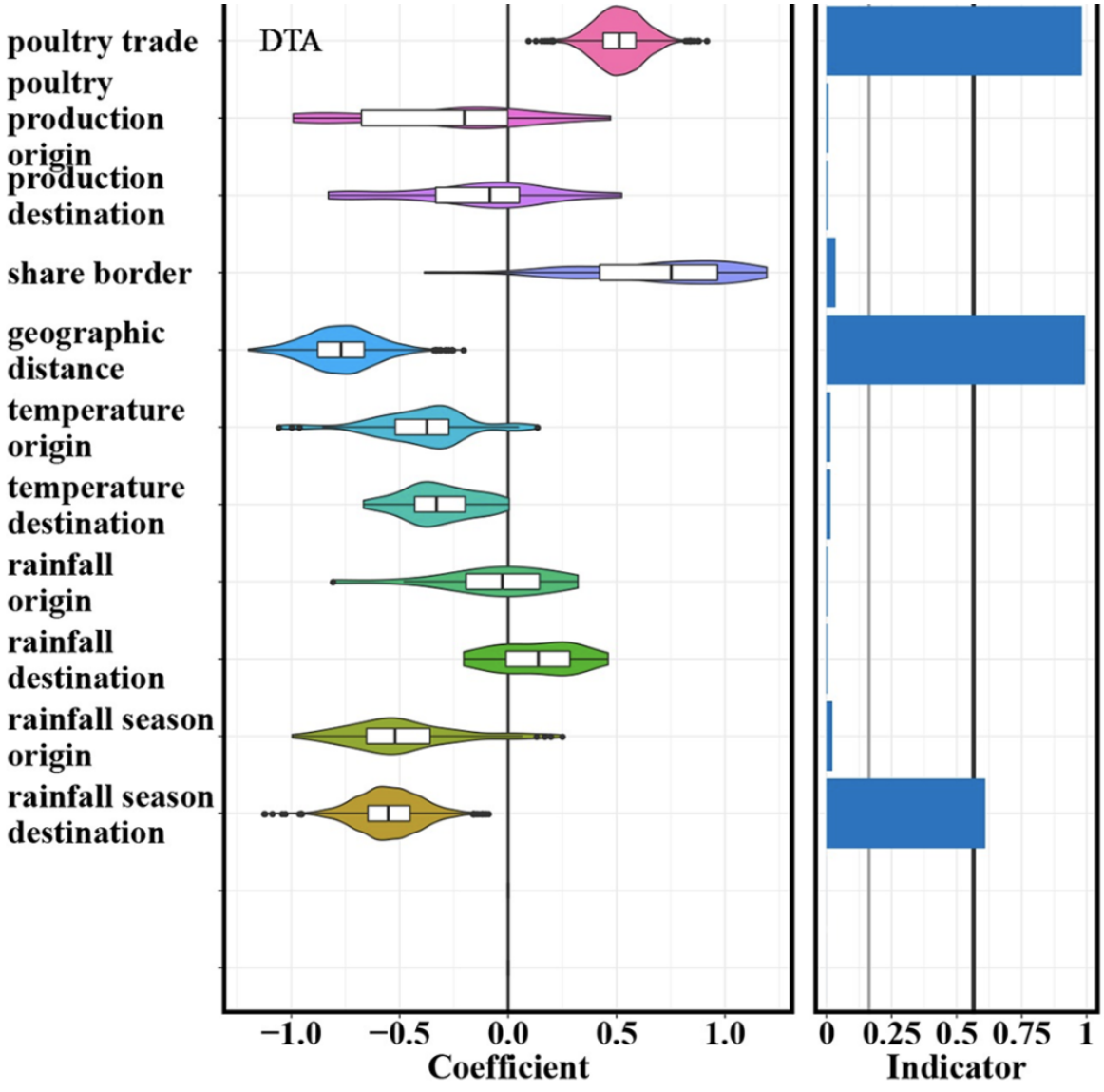
Tutorial: TTB Structured coalescent

MASCOT Muller et al, BioInf. 2018



Tutorial: TTB MASCOT

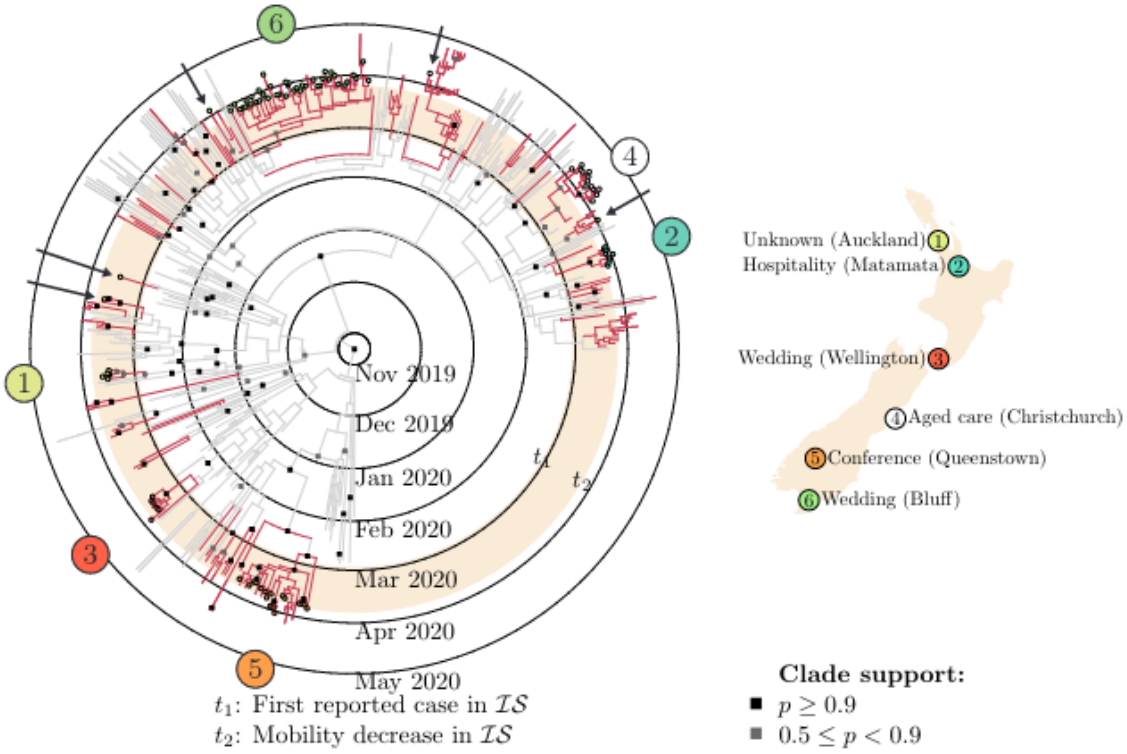
MASCOT Yang et al, PloS Comp Bio. 2019



See example XML in Mascot package

Multitype birth-death model

Structured Birth/Death model equivalent of structured coalescent
Tutorial: TTB Structured birth death model
Sensitive to priors



Douglas et al, Virus Evolution, 2021

Phylogeography summary

Continuous models

- random walk on plain/sphere
- break-away model
- easy to introduce landscape awareness
- landscape awareness requires
  - non-trivial justification
  - comparison with not landscape awareness

Discrete models

- Discrete trait – fast, simple, wrong
- Structured coalescent/birth death – slow, complex, correct

Model selection

Remco R. Bouckaert
r.bouckaert@auckland.ac.nz

Centre of Computational Evolution, University of Auckland

Online, June 2021

Topics

- Cultural evolution
- Phylogeography
- **Model selection**
- Model comparison

Model selection

When to do model selection:

- answers of interest are not robust for different models
- to test hypotheses (encoded by prior)
- not because the reviewers demand it

Evade it if you can

Model selection

Measures of fit:

- Super naive: compare posteriors
  - priors are not normalised, so posteriors cannot be compared
  - Never do this!

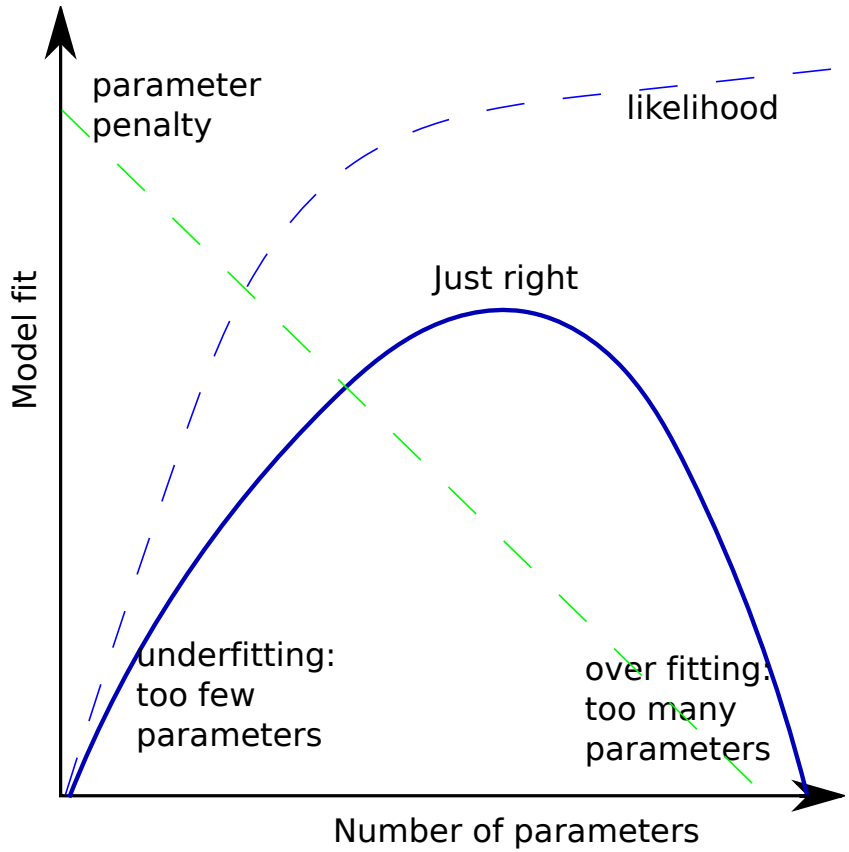
Model selection

Measures of fit:

- Super naive: compare posteriors
  - priors are not normalised, so posteriors cannot be compared
  - Never do this!
- Super naive: compare likelihoods
  - overparameterisation/overfitting cannot be detected
  - Never do this!

Model selection: select model with best "fit"

Desirable model fit property 1: likelihood - parameter penalty



Desirable model fit property 2: replicability/low variance

Desirable model fit property 3: easy & cheap to calculate

(this list is not exhaustive)

Bayesian model selection: marginal likelihood

Posterior:

$$p(\theta|D,M) = \frac{\overbrace{\pi(\theta|M)}^{\text{prior}}\overbrace{L(D|M,\theta)}^{\text{likelihood}}}{\underbrace{p(D|M)}_{\text{marginal likelihood}}}$$

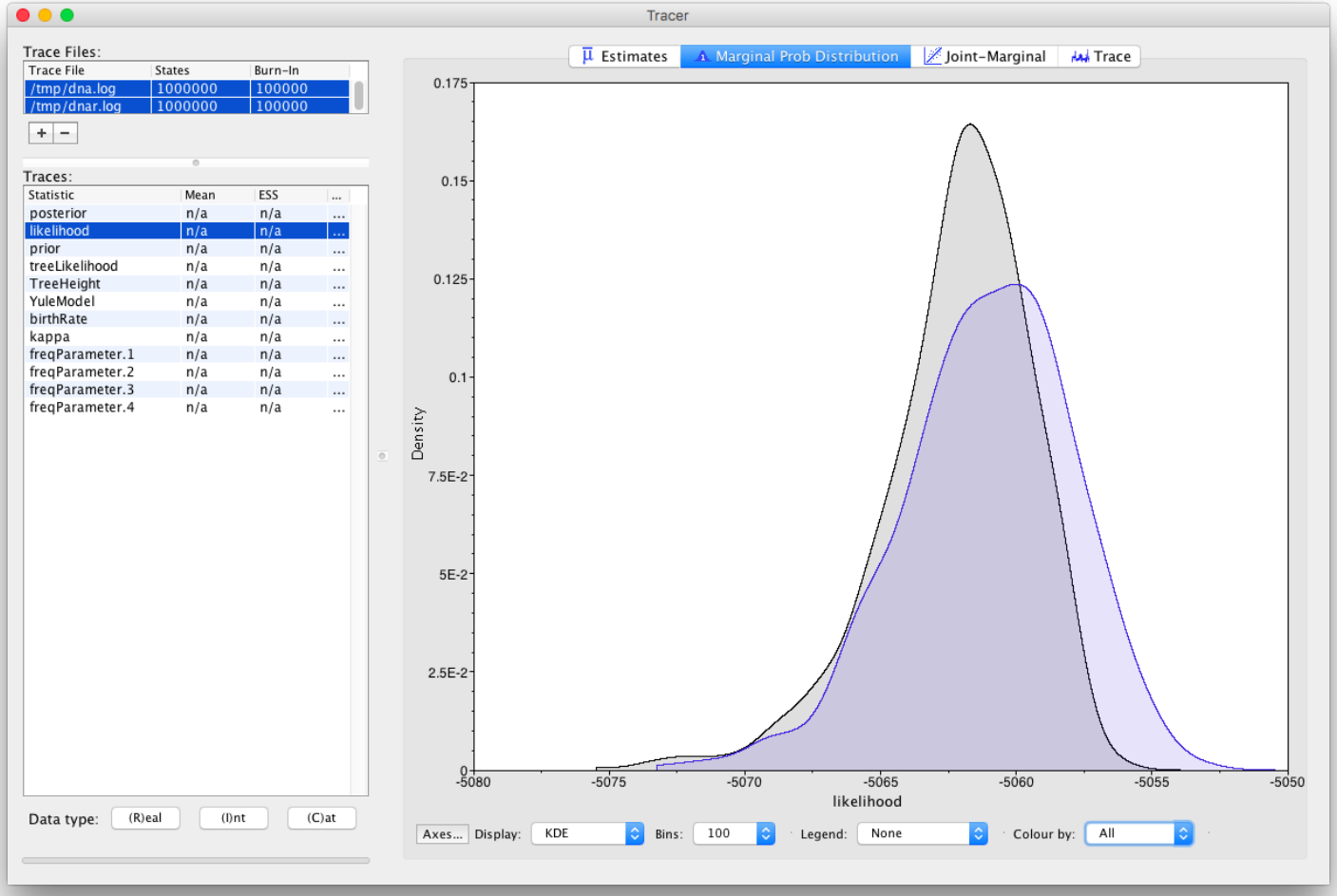
Marginal likelihood:

$$p(D|M) = \int_{\theta\in\Theta} \pi(\theta|M)L(D|M,\theta)d\theta$$

integrate/marginalise out θ
Bayes factor:

$$\frac{p(D|M_1)}{p(D|M_2)}$$

Model selection: "marginal likelihood" in Tracer



This is the likelihood averaged over samples from the posterior, not prior.

Path sampling/Stepping stone theory

- Marginal likelihood:

$$p(D) = \int_{\theta} \pi(\theta) L(D|\theta) d\theta$$

  hard to estimate directly.
- Define *power posterior* for some tractable reference distribution $p_w(\theta)$

$$P_\beta(\theta|D) = \frac{[L(D|\theta)\pi(\theta)]^\beta p_w(\theta)^{1-\beta}}{c_\beta}$$

  $P_1(\theta|D)$ is the posterior, c_1 the marginal likelihood.
  $P_0(\theta|D)$ is the reference distribution, $c_0 = 1$

Path sampling/Stepping stone theory

- Marginal likelihood:

$$p(D) = \int_{\theta} \pi(\theta) L(D|\theta) d\theta$$

  hard to estimate directly.
- Define *power posterior* for some tractable reference distribution $p_w(\theta)$

$$P_\beta(\theta|D) = \frac{[L(D|\theta)\pi(\theta)]^\beta p_w(\theta)^{1-\beta}}{c_\beta}$$

  $P_1(\theta|D)$ is the posterior, c_1 the marginal likelihood.
  $P_0(\theta|D)$ is the reference distribution, $c_0 = 1$
-

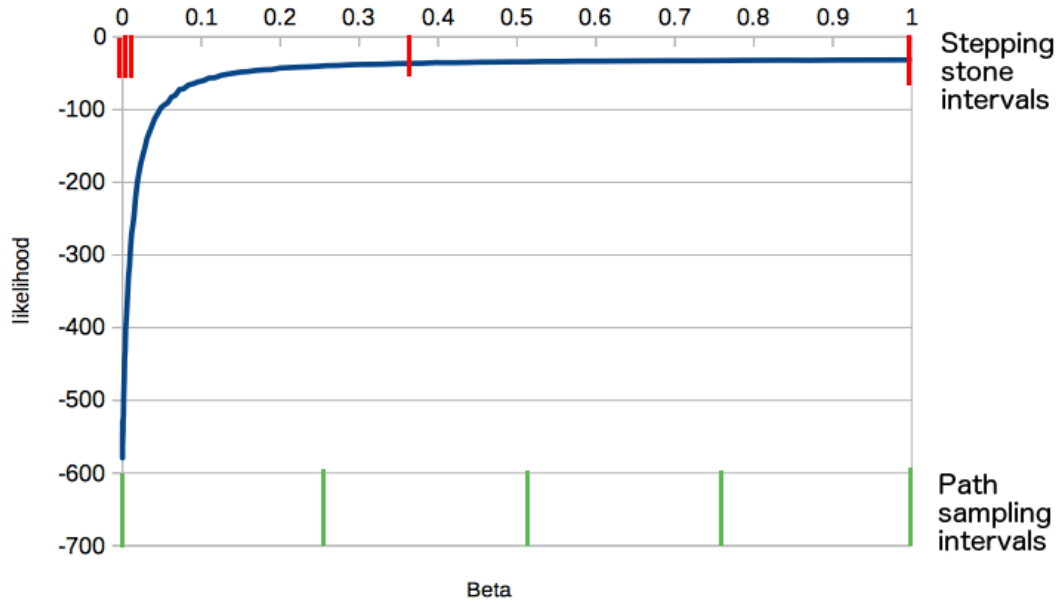
$$\frac{c_{\beta_k}}{c_{\beta_{k-1}}} \approx \frac{1}{n} \sum_{i=1}^{n} p(D|\theta_{k-1,i})^{\beta_k - \beta_{k-1}}$$

-

$$P(D|M) = \frac{c_1}{c_0} = \frac{c_1}{c_{0.3}} \frac{c_{0.3}}{c_{0.1}} \frac{c_{0.1}}{c_{0.01}} \frac{c_{0.01}}{c_0} = \frac{c_1}{\cancel{c_{0.3}}} \frac{\cancel{c_{0.3}}}{\cancel{c_{0.1}}} \frac{\cancel{c_{0.1}}}{\cancel{c_{0.01}}} \frac{\cancel{c_{0.01}}}{c_0}$$

Model selection: *Stepping stone* vs *path sampling*

Both use prior as reference distribution ($p_w(\theta) = \pi(\theta)$)



Stepping stone uses different intervals (set of β values) from path sampling, but otherwise the same

Path sampling/Stepping stone in practice

Number of steps:

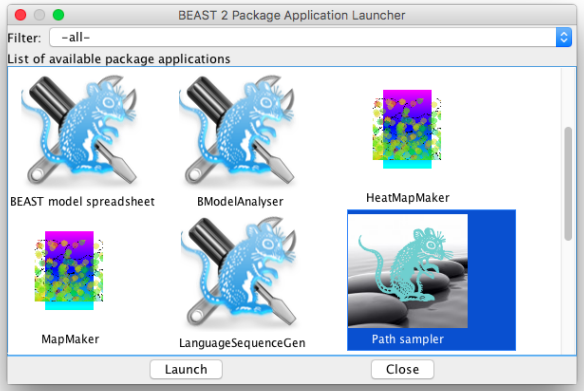
- Start with small nr of steps, say 16 and estimate ML
- increase nr of steps, estimate ML
- continue till ML estimate does not decrease any more

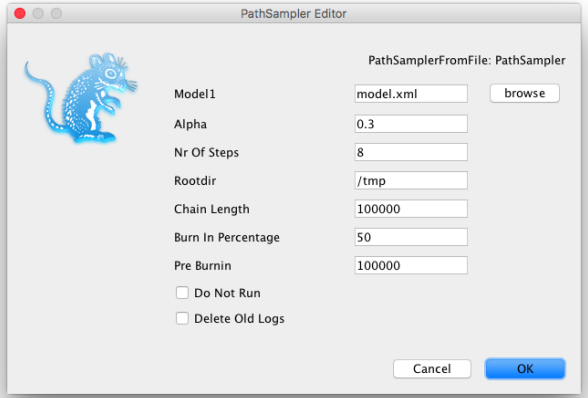
Chain length per step/ESS:

- total chain length at least as long as for posterior
- not all ESSs have to be 200 (errors cancel out)
- run different runs to get impression of variance of estimate
- use logcombiner to combine logs, for final estimate

Path sampling/Stepping stone in practice

Set up through XML or GUI





Set up through CLI:

- to list BEAST apps:
  `/path/to/applauncher -list`
- To show `PathSampler` options:
  `/path/to/applauncher PathSampler -help`
- To set up `PathSampler` analysis:
  `/path/to/applauncher PathSampler -nrOfSteps 64 -rootdir dir/withs/steps -burnInPercentage 50 -model beast.xml`

Creates subdirectory structure, one for each step containing all log files

Path sampling/Stepping stone in practice

To set up on a HPC cluster

- Set up locally, using 'doNotRun' flag = true
- Move steps to cluster, and run steps in paralel there
- Estimate ML using `PathSampleAnalyser`

```
/path/to/applauncher PathSampleAnalyser -nrOfSteps 64
-rootdir dir/withs/steps -burnInPercentage 50
```

Path sampling/Stepping stone in practice

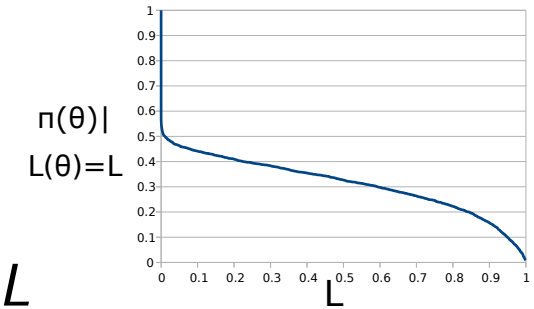
Trouble shooting

- ESS too small for a step: resume runs for that step
- Infinite likelihoods caused by numeric instability: improper priors – use proper priors instead
- -Infinite likelihoods: priors too wide – narrow priors
- Inspect log files in step directory to see which parameter escapes, so which prior to adjust

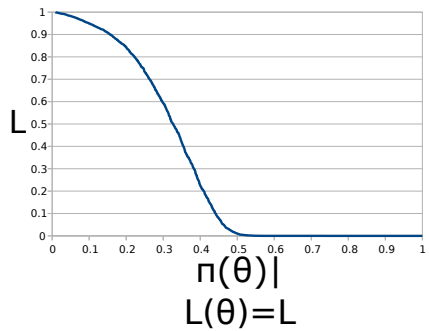
Nested Sampling Theory

$$\mathcal{Z} = \int_{\theta} \pi(\theta) L(\theta) \mathrm{d}\theta$$

$$= \int_{L=0}^{\infty} L \left(\int_{\theta, L(\theta)=L} \pi(\theta) \mathrm{d}\theta \right) \mathrm{d}L$$

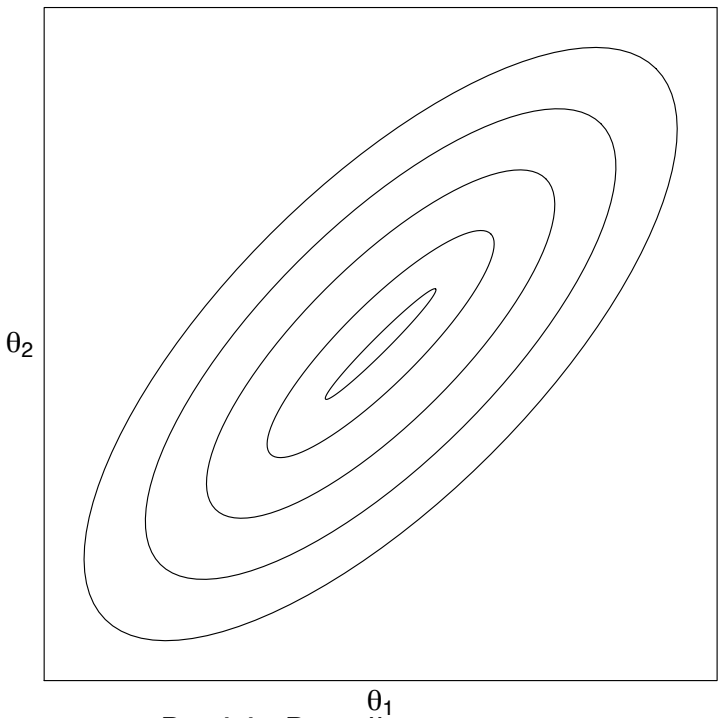


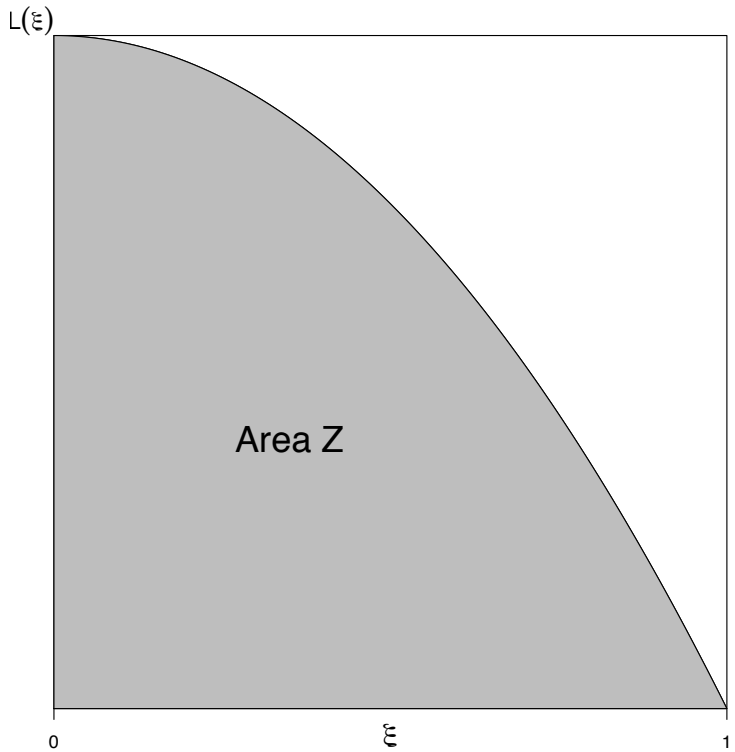
$$= \int_{X=0}^{1} \mathcal{L}(X) \mathrm{d}X$$



where $\mathcal{L}(X)$ inverse likelihood

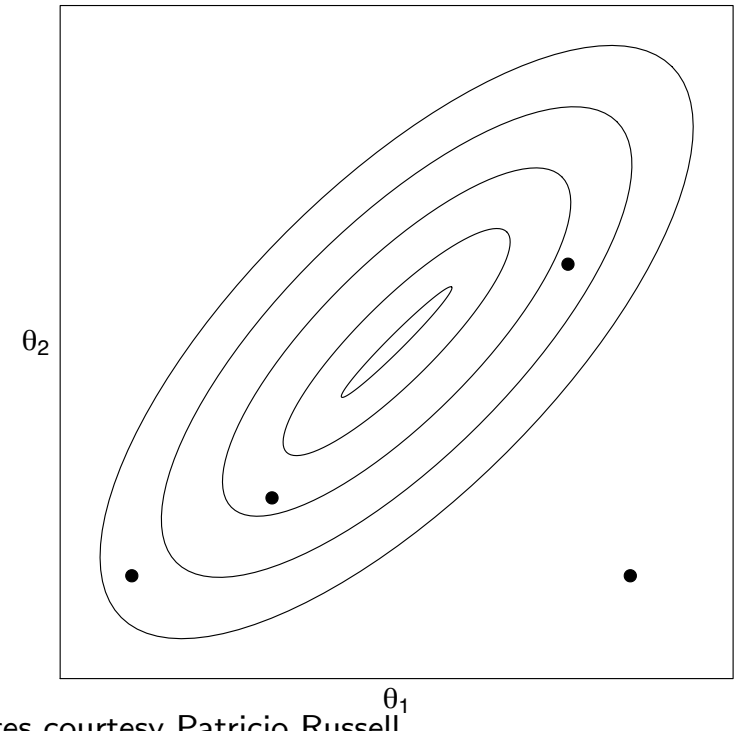
Nested sampling

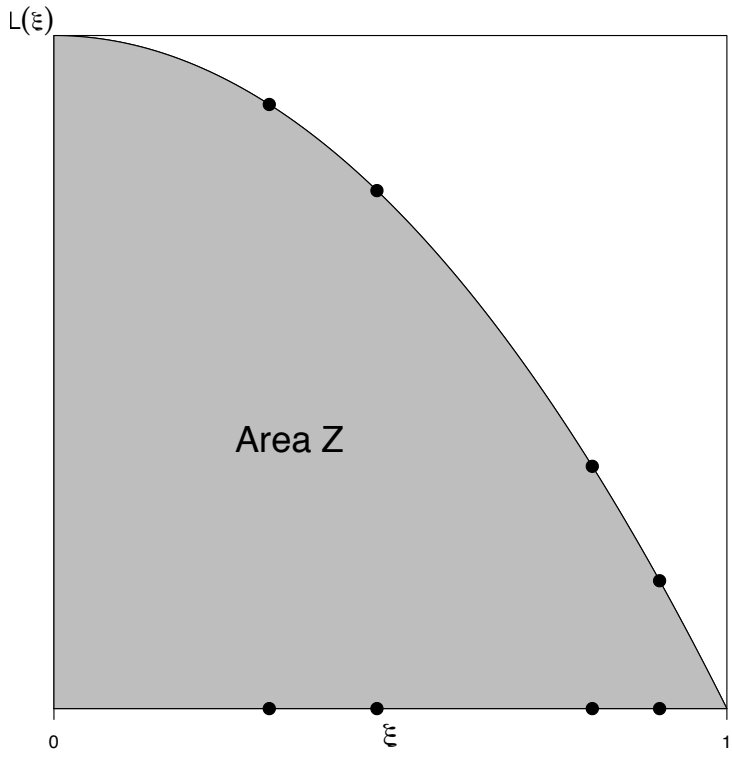




images courtesy Patricio Russell

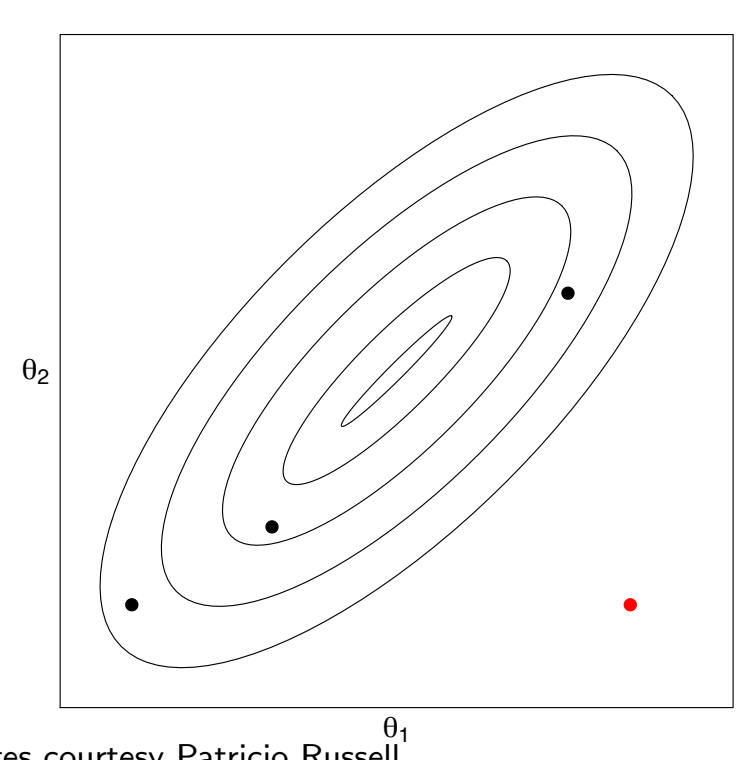
Nested sampling

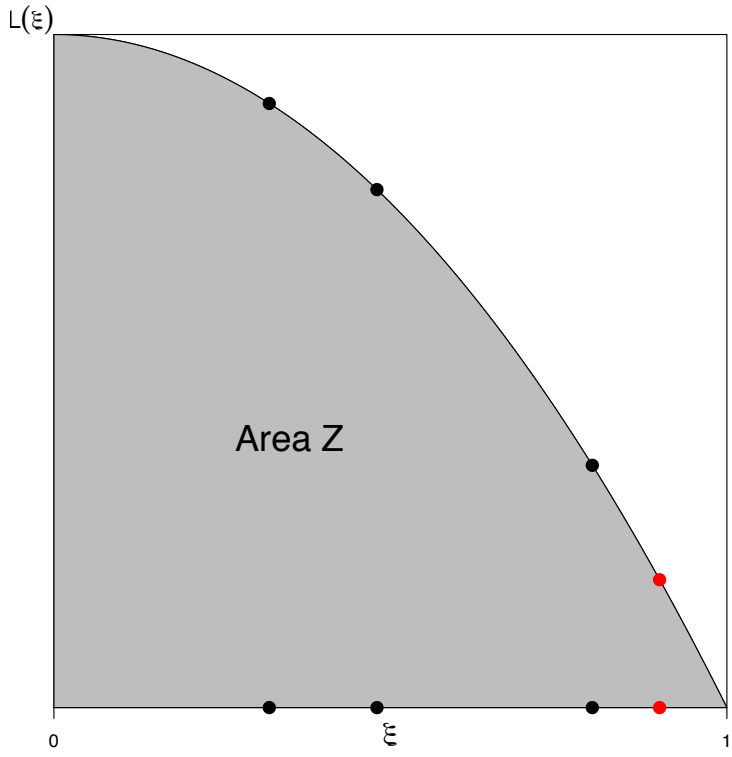




images courtesy Patricio Russell

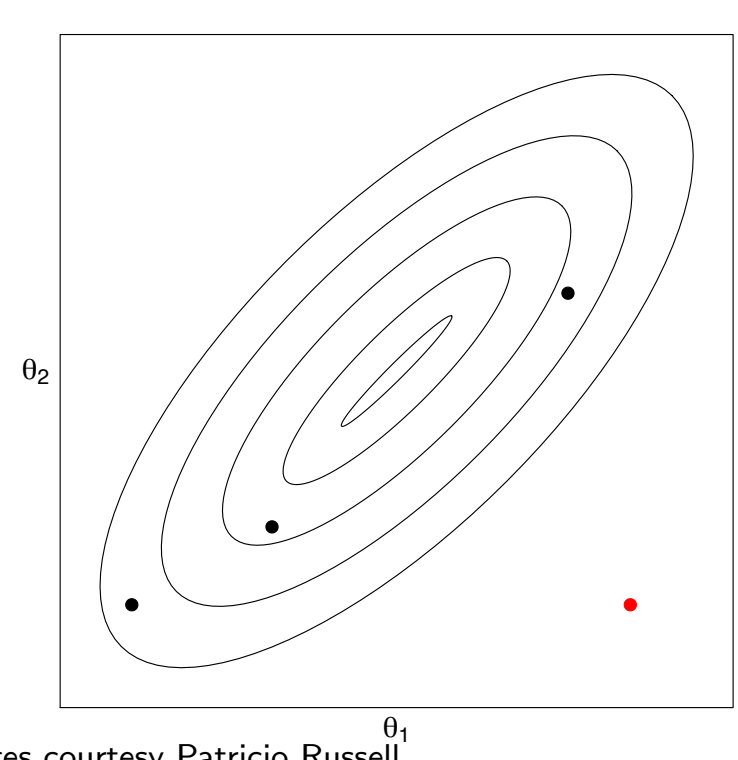
Nested sampling

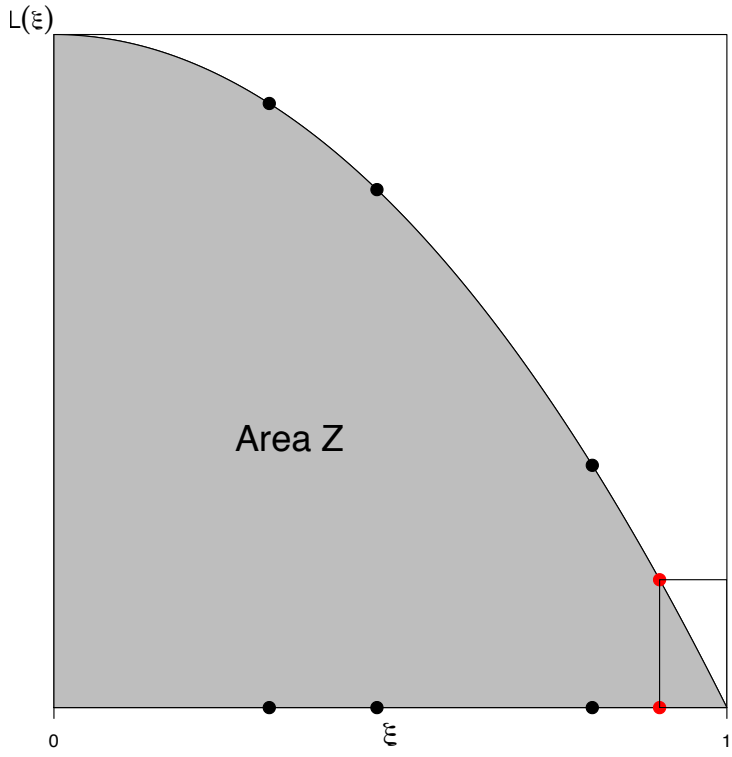




images courtesy Patricio Russell

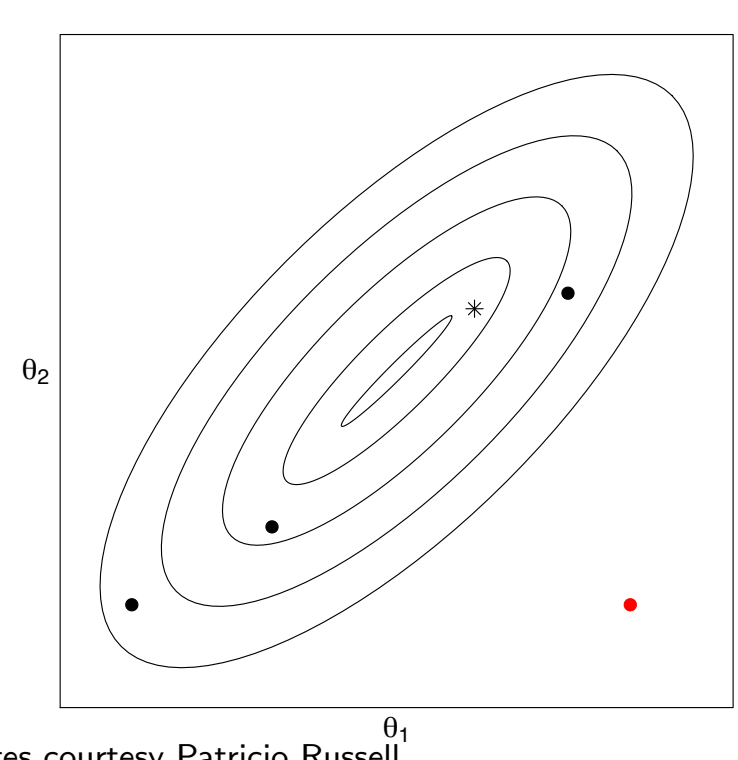
Nested sampling

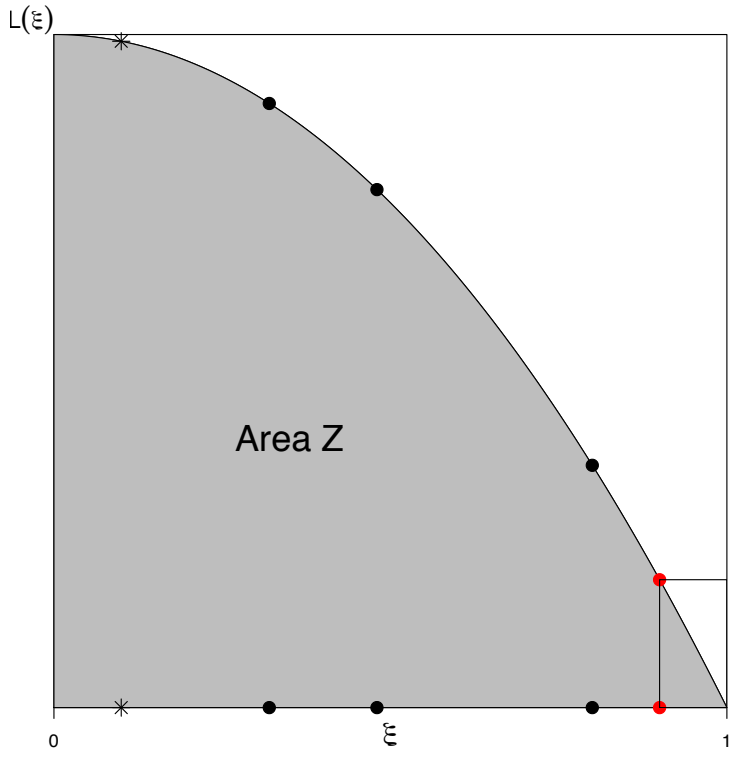




images courtesy Patricio Russell

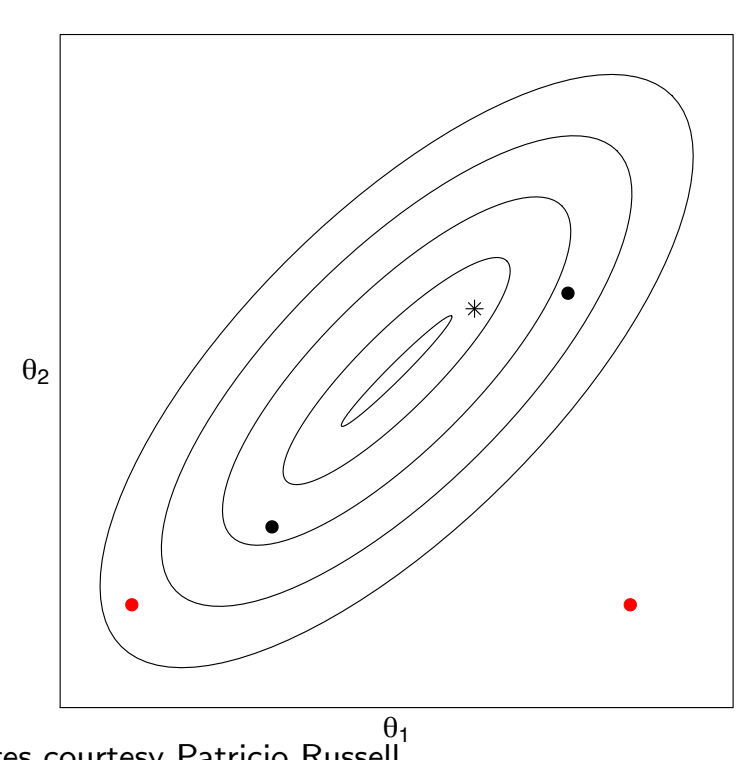
Nested sampling

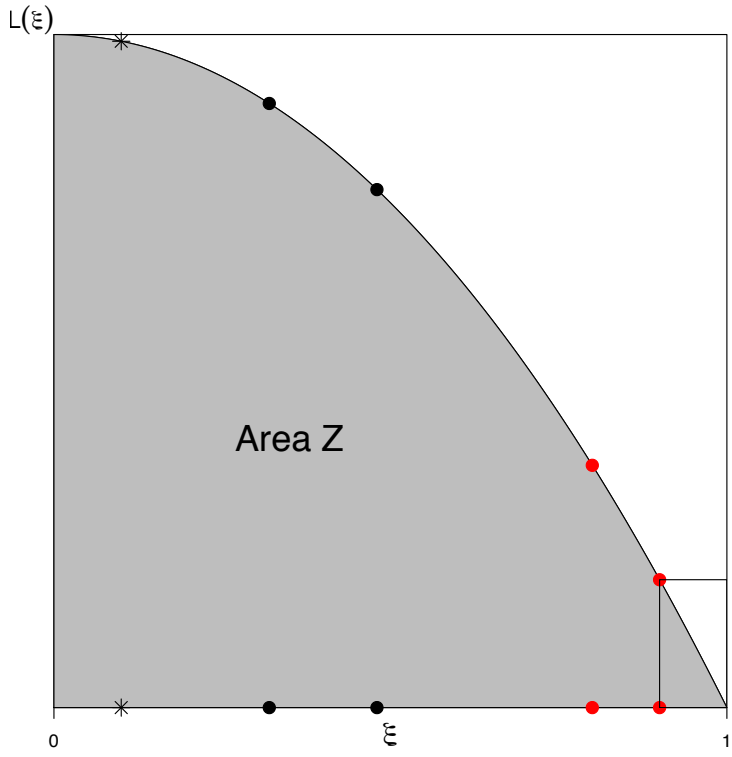




images courtesy Patricio Russell

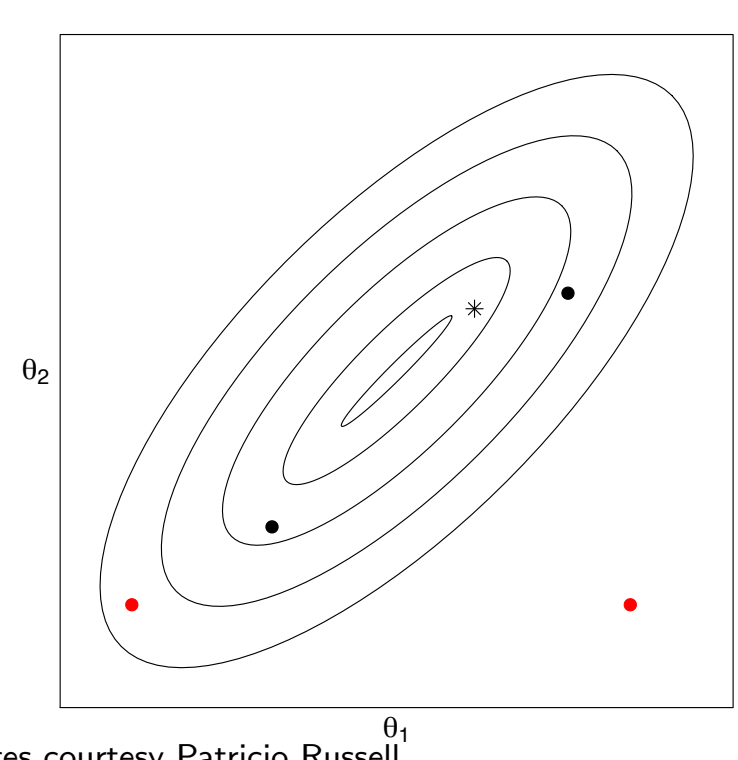
Nested sampling

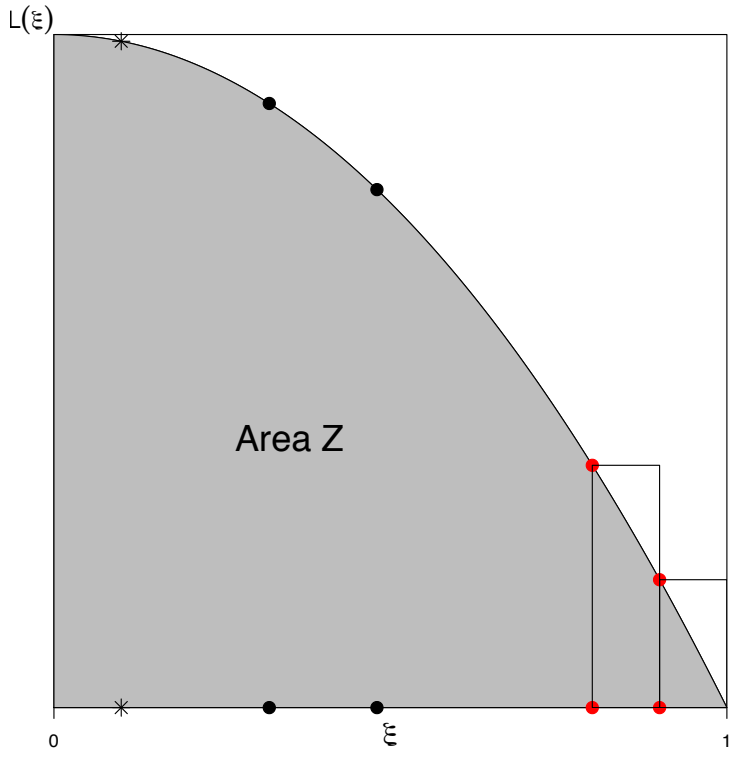




images courtesy Patricio Russell

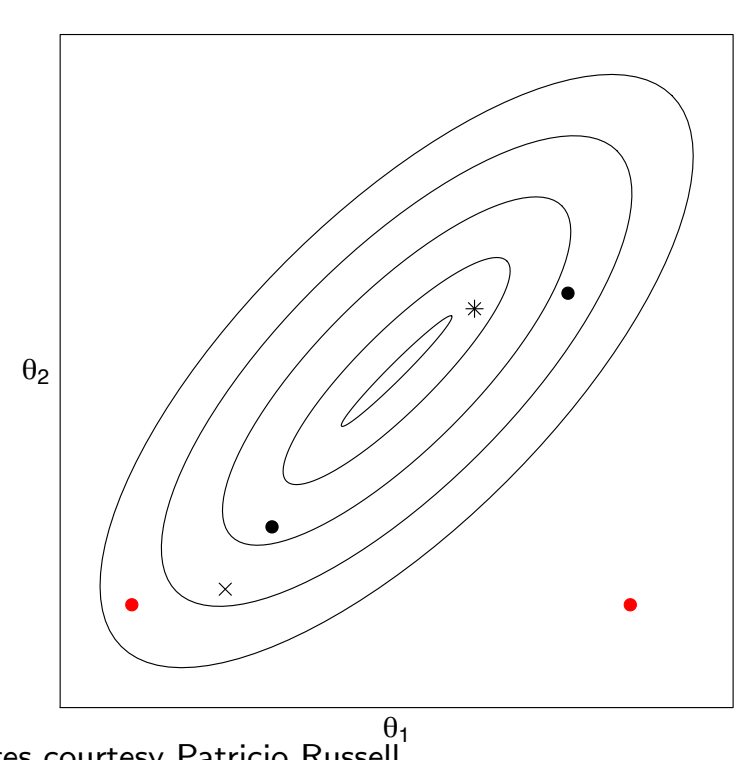
Nested sampling

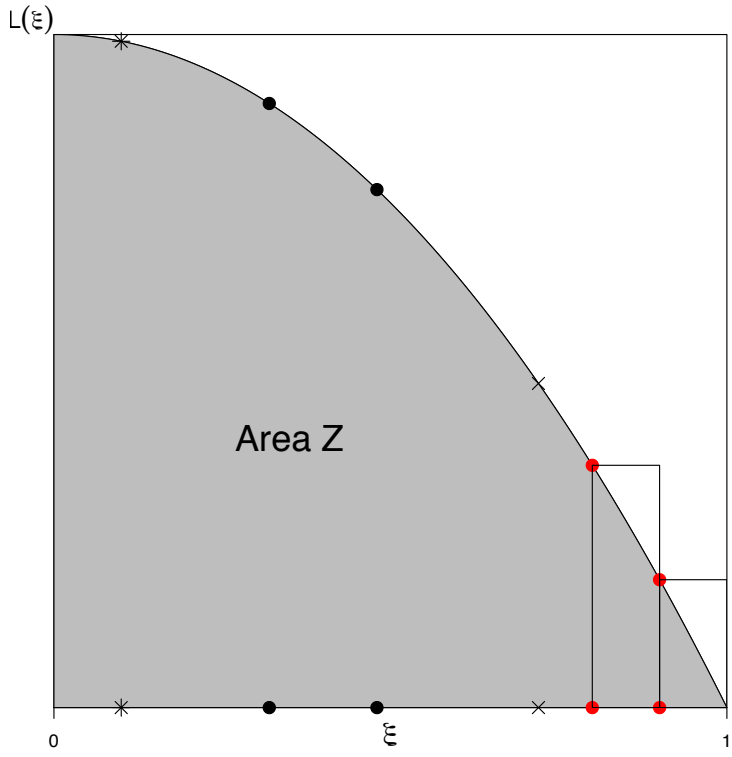




images courtesy Patricio Russell

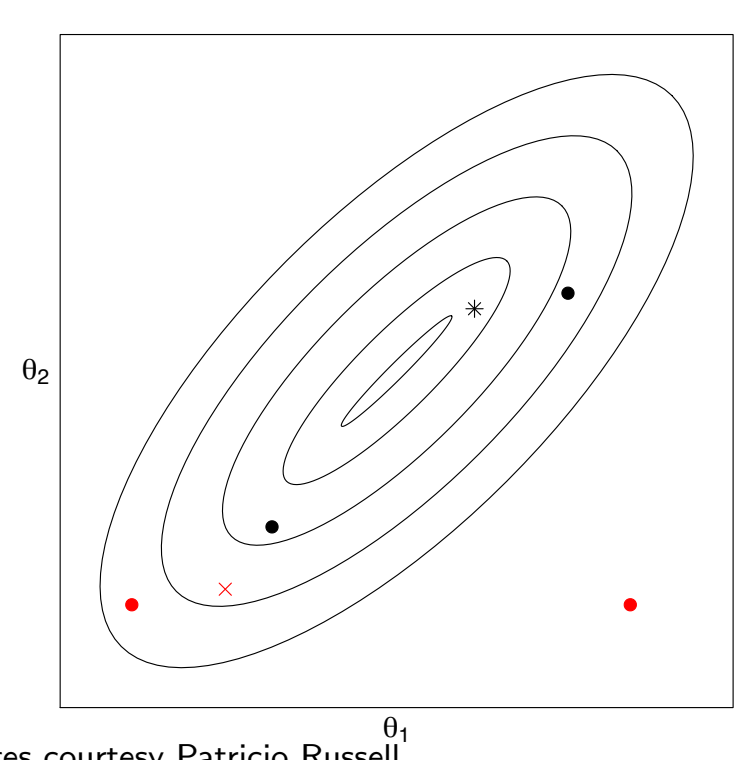
Nested sampling

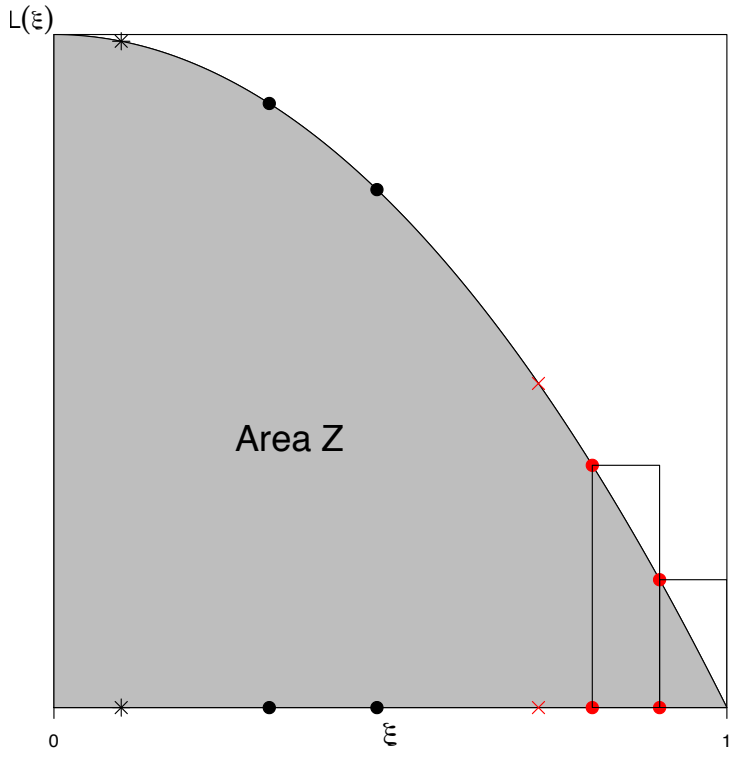




images courtesy Patricio Russell

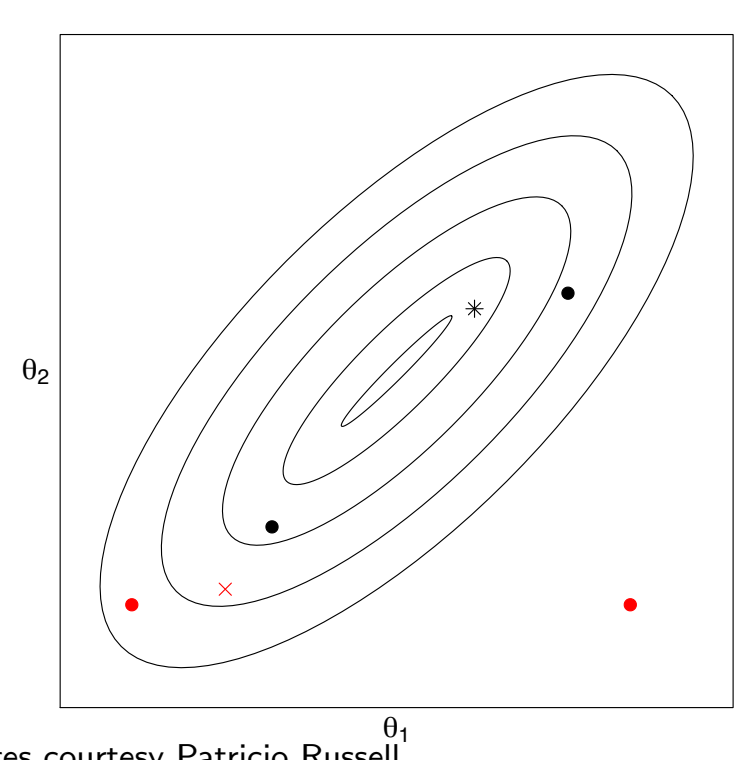
Nested sampling

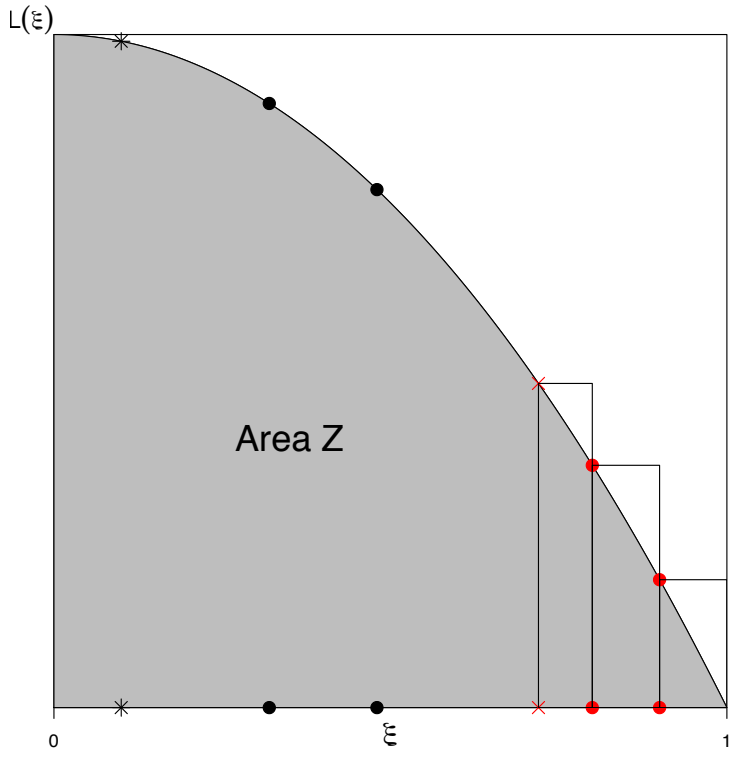




images courtesy Patricio Russell

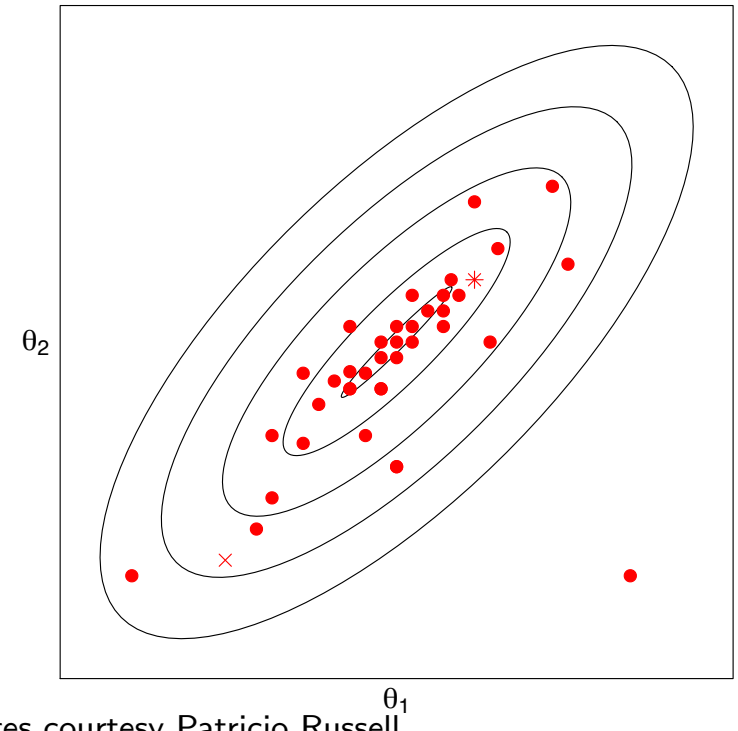
Nested sampling

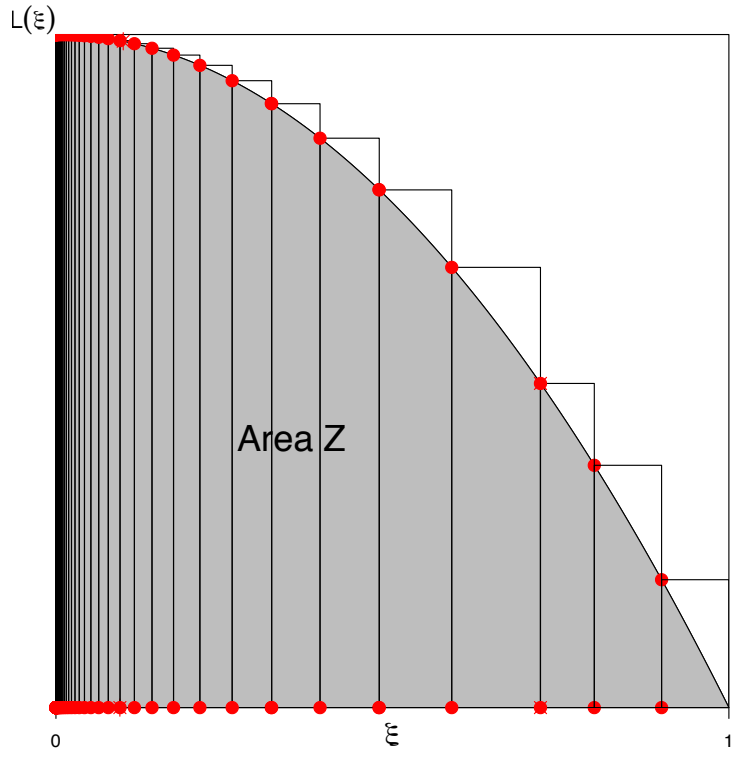




images courtesy Patricio Russell

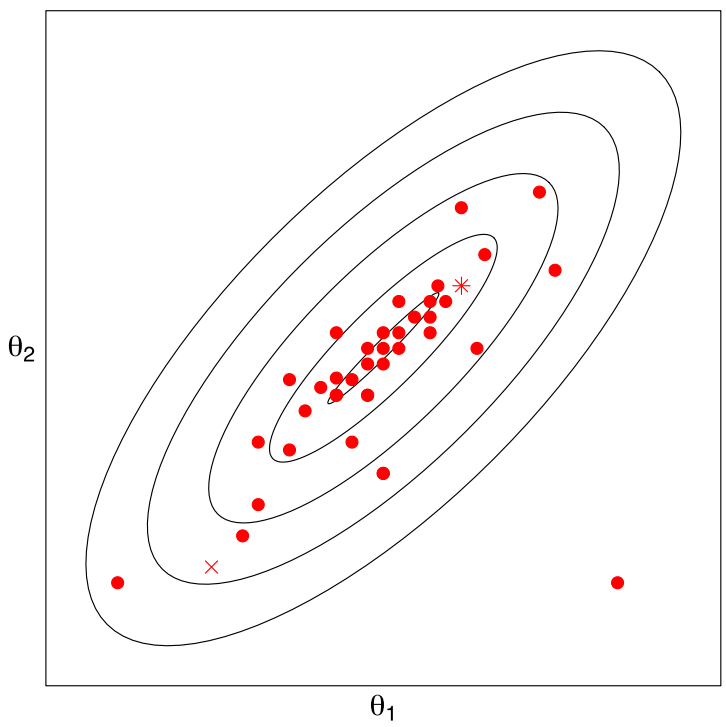
Nested sampling

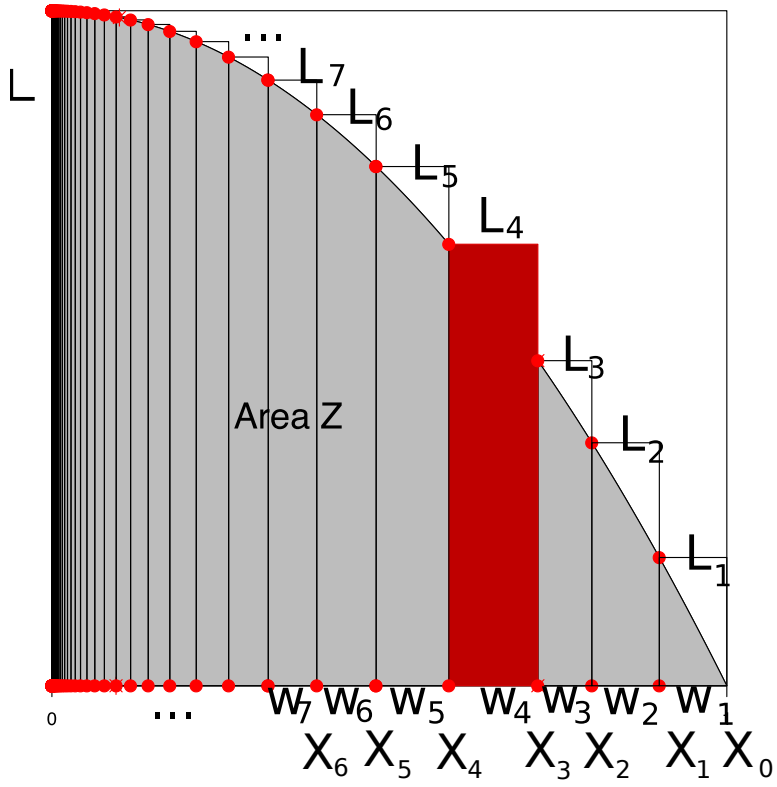




images courtesy Patricio Russell

Area under curve is ML $= \mathcal{Z}$, with N active points





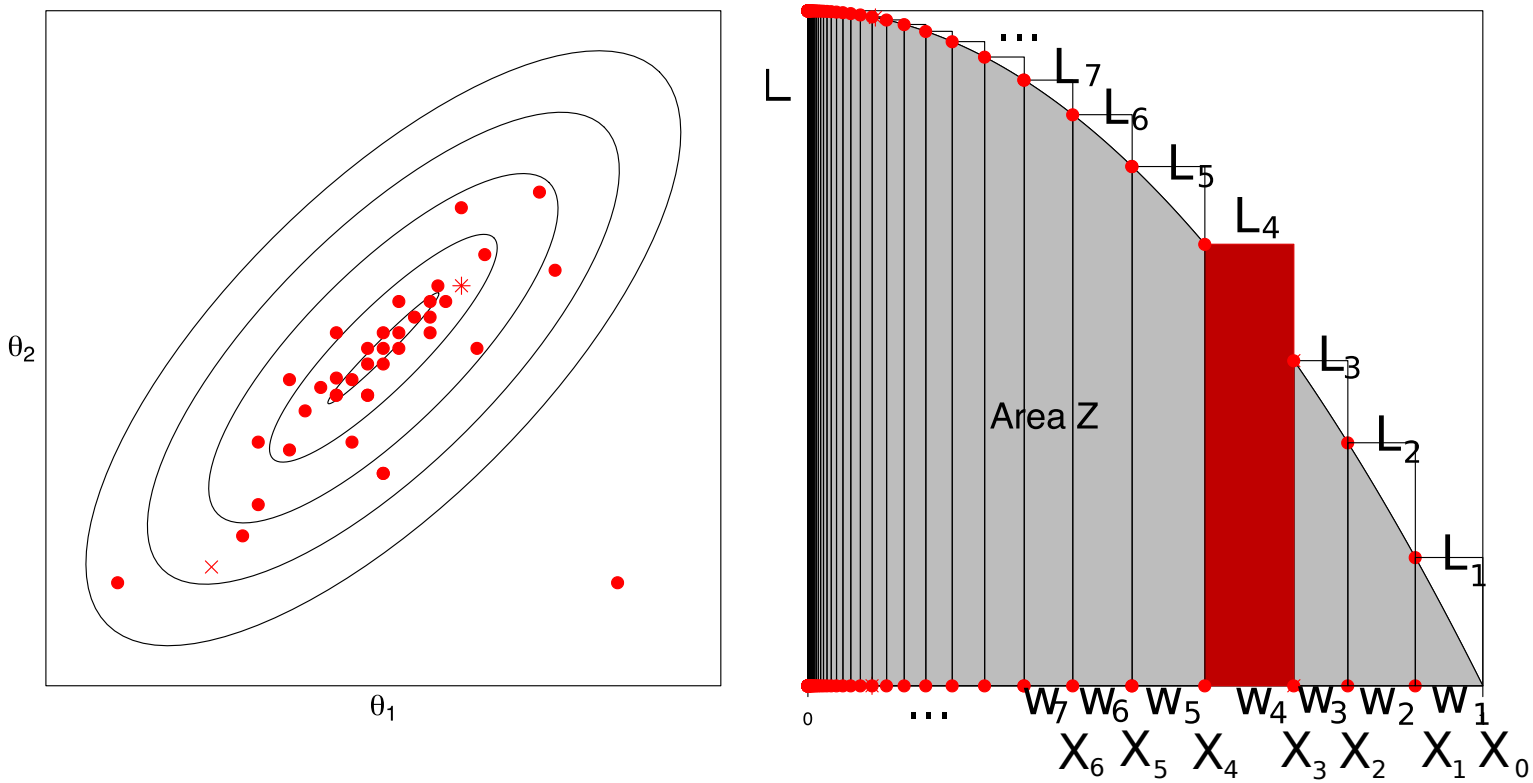
$X_0 = 1$

$X_i =$ proportion of prior mass with likelihood at least L_i

$w_i = X_i - X_{i-1}$

$\mathcal{Z} \approx \sum_i w_i L_i$

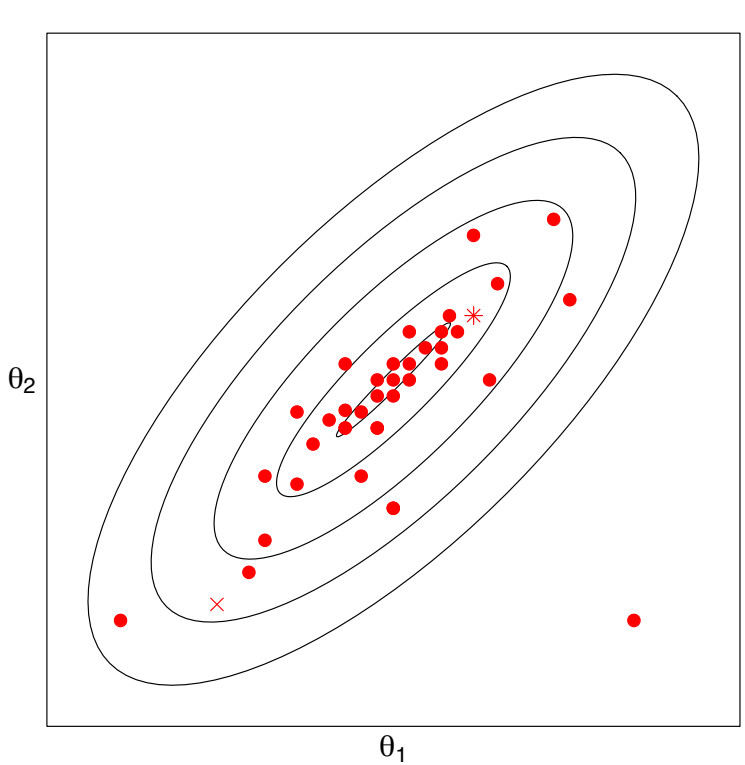
Defining X_i with N active points

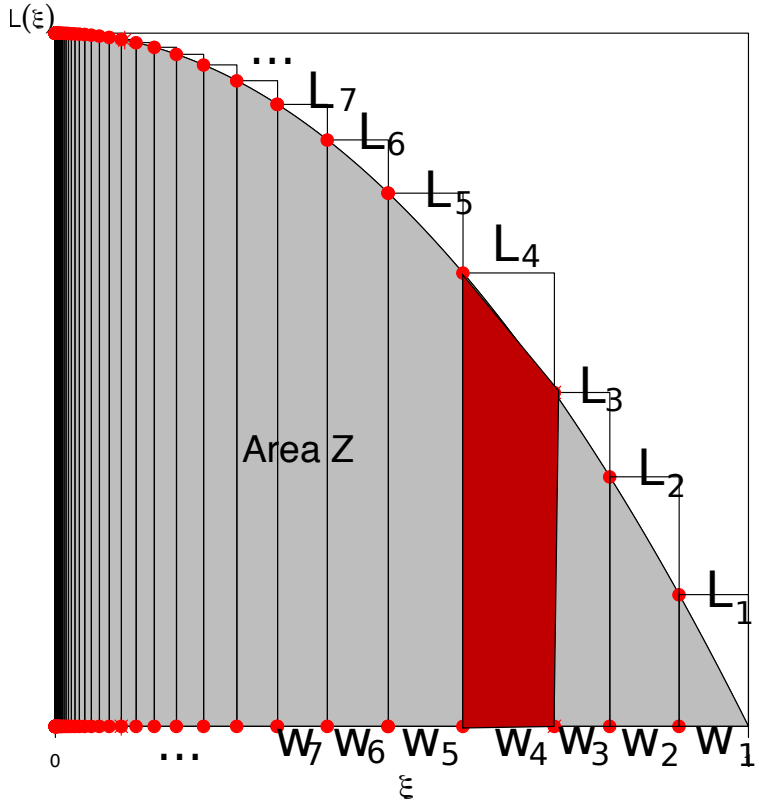


1. Arithmetic mean: $X_i = (\frac{N}{N+1})^i$
2. Geometric mean: $X_i = e^{-\frac{i}{N}}$ <= fast, most popular
3. Stochastic: $X_i = \beta(1, N)X_{i-1}$, $X_0 = 1$ <= allows SD estimate

$w_i = X_i - X_{i-1}$ $\qquad\qquad$ $\mathcal{Z} \approx \sum_i w_i L_i$

Use trapezium rule for more accurate ML estimate





$$\mathcal{Z} = \sum \ldots$$

Nested sampling with N active points

Assign weights to 'saved points'

$$E\{w_i\} = e^{-(i-1)/N} - e^{-i/N}$$

Estimate of marginal likelihood

$$\mathcal{Z} = \sum_i w_i L_i$$

Nested sampling with N active points

Assign weights to 'saved points'

$$E\{w_i\} = e^{-(i-1)/N} - e^{-i/N}$$

Estimate of marginal likelihood

$$\mathcal{Z} = \sum_i w_i L_i$$

Estimate of standard deviation of marginal likelihood

$$sd(\log \mathcal{Z}) = \sqrt{\frac{H}{N}}$$

where the information $H \approx \sum_i w_i \frac{L_i}{\mathcal{Z}} log \frac{L_i}{\mathcal{Z}}$

Nested sampling with N active points

Assign weights to 'saved points'

$$E\{w_i\} = e^{-(i-1)/N} - e^{-i/N}$$

Estimate of marginal likelihood

$$\mathcal{Z} = \sum_i w_i L_i$$

Estimate of standard deviation of marginal likelihood

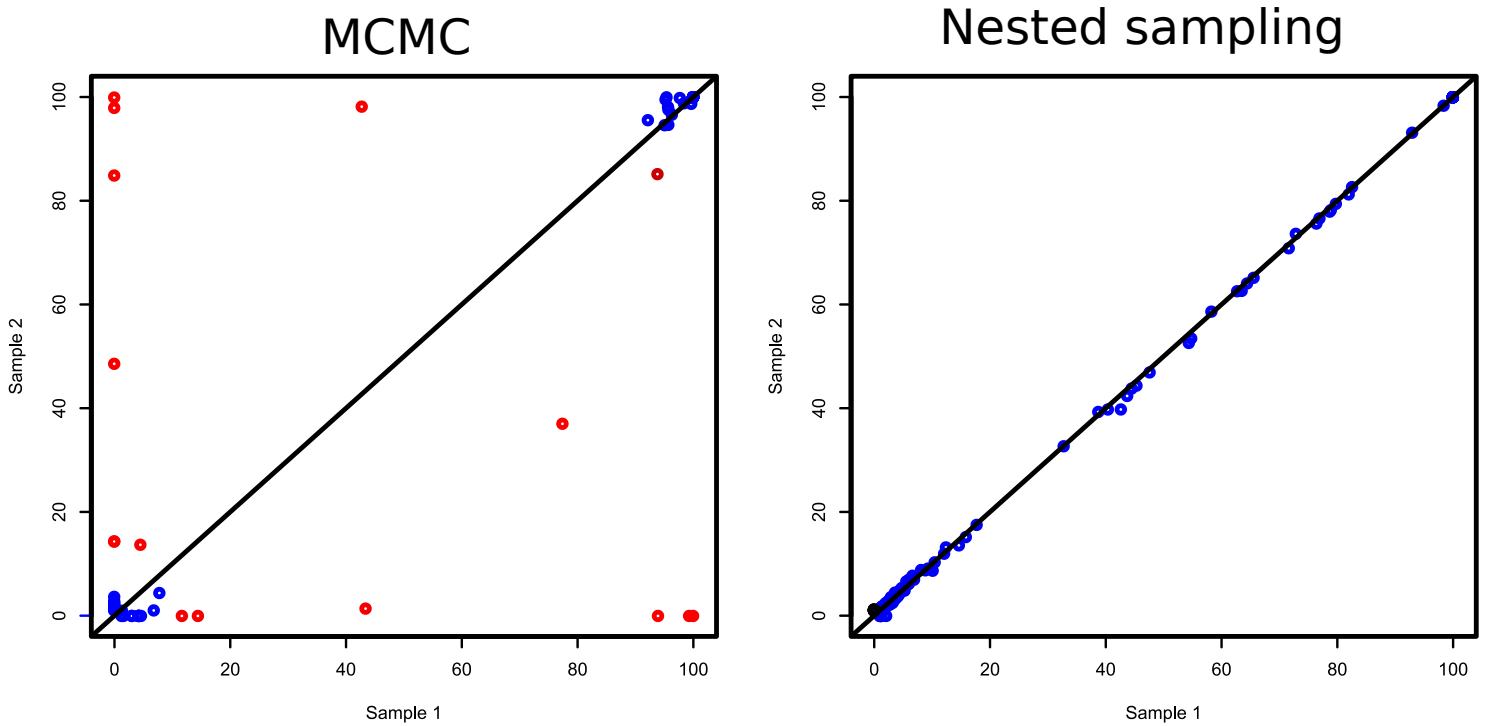
$$sd(\log \mathcal{Z}) = \sqrt{\frac{H}{N}}$$

where the information $H \approx \sum_i w_i \frac{L_i}{\mathcal{Z}} log \frac{L_i}{\mathcal{Z}}$

Sample from posterior by sampling saved points according to weights $\frac{w_i L_i}{\mathcal{Z}}$

DS1: where MCMC fails

- DS1 data set with tree islands
- MCMC has trouble moving between islands
- Consequently MCMC/stepping stone/path sampling fails



NS in practice: Setting up an analysis

Requires NS package

Set up analysis in BEAUti, then edit XML and replace

```
<run id="mcmc" spec="MCMC" chainLength="100000000">
```

with

```
<run id="mcmc" spec="beast.gss.NS" chainLength="20000"
   particleCount="1" subChainLength="5000" epsilon="1e-12">
```

More info: `https://github.com/BEAST2-Dev/nested-sampling/wiki/How-to-use-NS`

NS in practice: Are we there yet?

- Run multiple times (like MCMC).
- Check $\mathcal{Z}$ estimates are compatible

$$|\log \mathcal{Z}_1 - \log \mathcal{Z}_2| \leq 2\sqrt{(SD_1^2 + SD_2^2)}$$

- If not, run with longer sub chain length
- Check $\mathcal{Z}$ estimates are compatible with shorter runs/not systematically biased in multiple runs

Note, nested sampling under estimates $\mathcal{Z}$. Longer sub chain length results in lower $\mathcal{Z}$ if not converged yet.

NS in practice: Model selection with Nested sampling

Given M_1 and M_2: say HKY vs GTR

1. estimate log marginal likelihoods $\mathcal{Z}_1$ for M_1 and $\mathcal{Z}_2$ for M_2
   • NS provides standard deviations SD_1 and SD_2 for **log** marginal likelihoods
2. if $|\log \mathcal{Z}_1 - \log \mathcal{Z}_2| \geq 2\sqrt{SD_1^2 + SD_2^2}$ calculate Bayes factor $BF = \log \mathcal{Z}_1 - \log \mathcal{Z}_2$. Done!
3. else if $\sqrt{SD_1^2 + SD_2^2} < 3(?)$ then M_1 and M_2 cannot be distinguished. Done!
4. else, run NS with more particles. How many? Use $SD = \sqrt{\frac{H}{N}}$ so $N = SD^2/H$ for desired SD (H from NS run) and goto (2)

NS in practice: Pitfalls

- Subchain length too short – run with different length, compare whether estimates differ
- Epsilon too large, causing early stopping, underestimate of $\mathcal{Z}$
- ...

NS in practice: Parallel implementation

- Maintaining shared pool to selected starting point => behaves like single thread N particle
- Runtime scales linear with nr or particles N
- N single particle runs can be combined => embarrassingly parallel
- Little communication required, so can be forked out over different CPUs

Stepping Stone and Nested Sampling work on any model

provided the prior is proper

Model selection summary

Stepping Stone and Nested Sampling work on any model

provided the prior is proper

Improper priors do not integrate to one, e.g., 1/X, uniform(0,∞)

Model selection summary

- Path sampling/Stepping stone:
  - computationally expensive
  - most stable marginal likelihood estimation we got (so far)
  - use this if you can
- Nested sampling:
  - Provides estimate of ML + its variance
  - Computation (inverse) proportional to accuracy of estimate
  - Can choose accuracy as desired

Model comparison

Remco R. Bouckaert
r.bouckaert@auckland.ac.nz

Centre of Computational Evolution, University of Auckland

Online, June 2021

Topics

- Cultural evolution
- Phylogeography
- Model selection
- **Model comparison**

Model comparison/Bayesian hypothesis testing

- Through model selection:
  - compare Bayes factors based on ML estimates
- Through model averaging: post-hoc analysis
  - compare Bayes factors based on empirical estimates from prior and posterior samples

Bayes factor:

$$\frac{p(D|M_1)}{p(D|M_2)} \text{ estimated by } \frac{\frac{empirical\ \ posterior(M_1)}{empirical\ \ prior(M_1)}}{\frac{empirical\ \ posterior(M_2)}{empirical\ \ prior(M_2)}}$$

Obtain sample from prior for M_1 and M_2.
Obtain sample from posterior for M_1 and M_2.

Bayes Factors

BF range			$\ln(BF)$ range			$\log_{10}(BF)$ range			Interpretation
1	–	3	0	–	1.1	0	–	0.5	hardly worth mentioning
3	–	20	1.1	–	3	0.5	–	1.3	positive support
20	–	150	3	–	5	1.3	–	2.2	strong support
	$>$	150		$>$	5		$>$	2.2	overwhelming support

Kass & Raftery, JASA, 1995

Model comparison of topologies

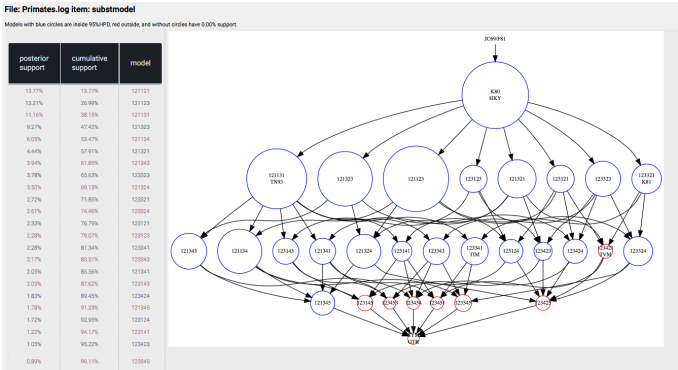
- Clade support for two alternative clades – M_1: clade A is present M_2: clade B is present
- From sample from prior: $P(M_1) = 0.3$ $P(M_2) = 0.4$
- From sample from posterior: $P(D|M_1) = 0.6$ $P(D|M_2) = 0.1$
- Bayes factor

$$\frac{p(D|M_1)}{p(D|M_2)} = \frac{\frac{posterior(M_1)}{prior(M_1)}}{\frac{posterior(M_2)}{prior(M_2)}} = \frac{\frac{0.6}{0.3}}{\frac{0.1}{0.4}} = 8$$

- Positive support for M_1 for clade A vs clade B

Model comparison of substitution models

- bModelTest – M_1: HKY vs M_2: GTR



- From sample from prior: $P(M_1) = \frac{1}{31}$ $P(M_2) = \frac{1}{31}$
- From sample from posterior: $P(D|M_1) = 0.1377$
  $P(D|M_2) = 0.0060$
- Bayes factor

$$\frac{p(D|M_1)}{p(D|M_2)} = \frac{\frac{posterior(M_1)}{prior(M_1)}}{\frac{posterior(M_2)}{prior(M_2)}} = \frac{0.1377/\frac{1}{31}}{0.0060/\frac{1}{31}} = 22.95$$

- Strong support for M_1: HKY

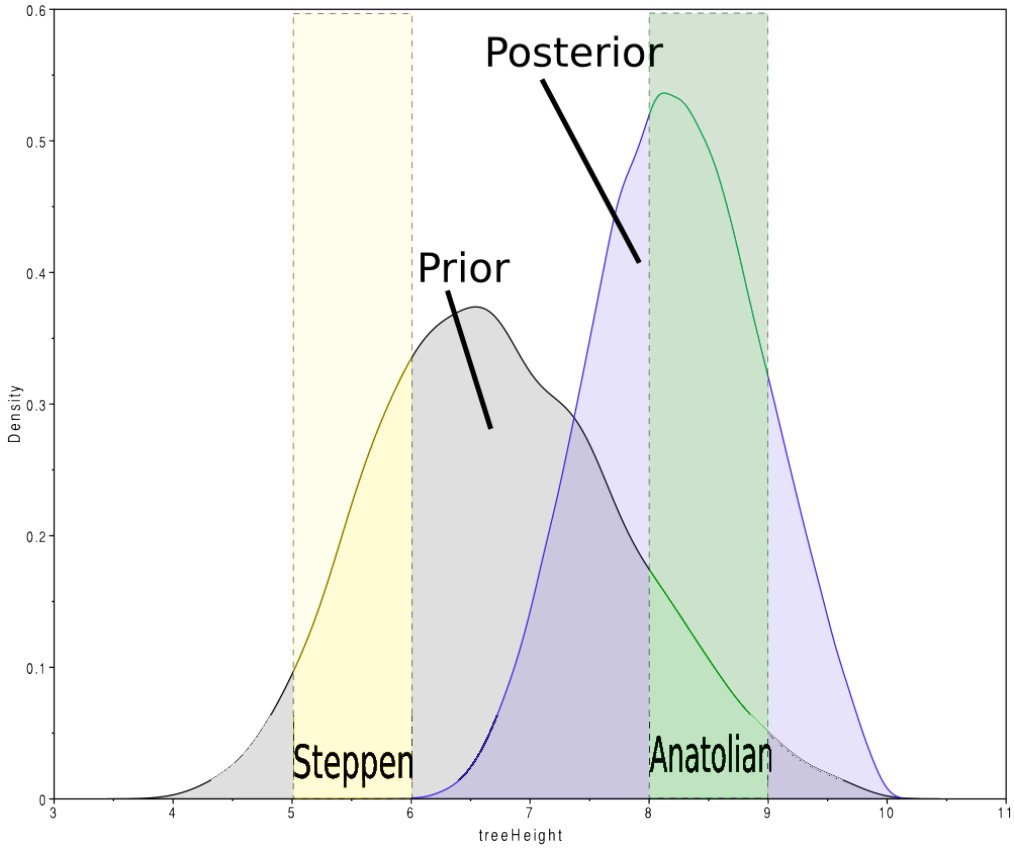
Model comparison of root age

Origin of Indo-European: Two competing theories

Steppen 5000 – 6000BP, Anatolian 8000 – 9000BP

Model comparison of root age

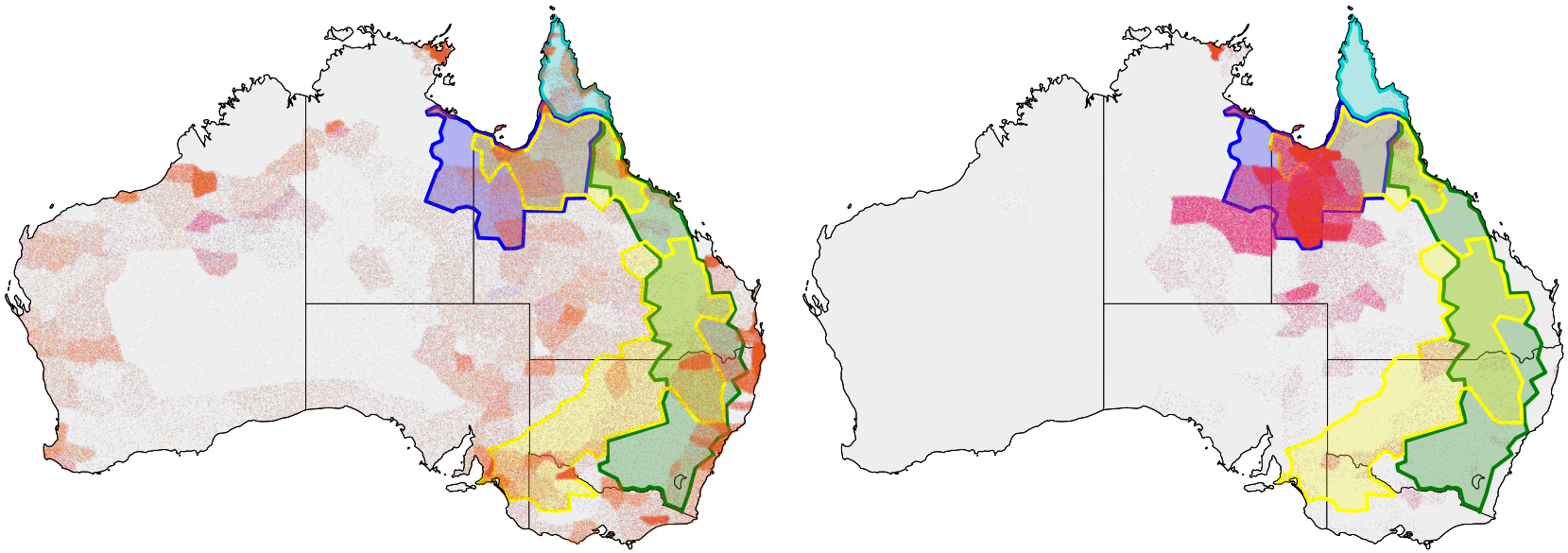
Origin of Indo-European: Two competing theories



Root height Prior = 6.8 [4.9, 8.8] Posterior 8.2 [6.9, 9.6]
Bayes Factor $\gg$ 100 in favour of Anatolian hypothesis

Model comparison of root location

Root location: Pama Nyungang M_1 M_2 M_3 M_4



Bayes Factors:

	M_1	M_2	M_3	M_4
M_1	–	6.22	76.66	486.34
M_2	0.16	–	12.33	78.23
M_3	0.01	0.08	–	6.34
M_4	0.00	0.01	0.16	–

Summary model comparison, selection

Model selection

- Path sampling
- Nested sampling

Model comparison

- Formulate hypotheses that can be measured
- Sample from prior and posterior
- Bayes factors from posterior sample

Questions?

???