Bayesian inference: what it means and why we care

Robin J. Ryder

Centre de Recherche en Mathématiques de la Décision
Université Paris-Dauphine

6 November 2017
Mathematical Coffees

The aim of Statistics

In Statistics, we generally care about inferring information about an unknown parameter θ. For instance, we observe $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$ and wish to:

- Obtain a (point) estimate $\hat{\theta}$ of θ, e.g. $\hat{\theta} = 1.3$.
- Measure the uncertainty of our estimator, by obtaining an interval or region of plausible values, e.g. $[0.9, 1.5]$ is a 95% confidence interval for θ.
- Perform model choice/hypothesis testing, e.g. decide between $H_0 : \theta = 0$ and $H_1 : \theta \neq 0$ or between $H_0 : X_i \sim \mathcal{N}(\theta, 1)$ and $H_1 : X_i \sim \mathcal{E}(\theta)$.
- Use this inference in postprocessing: prediction, decision-making, input of another model...

Why be Bayesian?

Some application areas make heavy use of Bayesian inference, because:

- The models are complex
- Estimating uncertainty is paramount
- The output of one model is used as the input of another
- We are interested in complex functions of our parameters

Frequentist statistics

- Statistical inference deals with estimating an unknown parameter θ given some data D.
- In the frequentist view of statistics, θ has a true fixed (deterministic) value.
- Uncertainty is measured by confidence intervals, which are not intuitive to interpret: if I get a 95% CI of [80 ; 120] (i.e. 100 ± 20) for θ, I cannot say that there is a 95% probability that θ belongs to the interval [80 ; 120].
- Frequentist statistics often use the maximum likelihood estimator: for which value of θ would the data be most likely (under our model)?

$$L(\theta|D) = P[D|\theta]$$

$$\hat{\theta} = \arg\max_{\theta} L(\theta|D)$$

Bayes' rule

Recall Bayes' rule: for two events A and B, we have

$$P[A|B] = \frac{P[B|A]P[A]}{P[B]}.$$

Alternatively, with marginal and conditional densities:

$$\pi(y|x) = \frac{\pi(x|y)\pi(y)}{\pi(x)}.$$

Bayesian statistics

- In the Bayesian framework, the parameter θ is seen as inherently random: it has a distribution.
- Before I see any data, I have a *prior* distribution on $\pi(\theta)$, usually uninformative.
- Once I take the data into account, I get a *posterior* distribution, which is hopefully more informative.
  By Bayes' rule,

$$\pi(\theta|D) = \frac{\pi(D|\theta)\pi(\theta)}{\pi(D)}.$$

  By definition, $\pi(D|\theta) = L(\theta|D)$. The quantity $\pi(D)$ is a normalizing constant with respect to θ, so we usually do not include it and write instead

$$\pi(\theta|D) \propto \pi(\theta)L(\theta|D).$$

Bayesian statistics

$$\pi(\theta|D) \propto \pi(\theta)L(\theta|D)$$

- Different people have different priors, hence different posteriors. But with enough data, the choice of prior matters little.
- We are now allowed to make probability statements about θ, such as "there is a 95% probability that θ belongs to the interval [78 ; 119]" (credible interval).

Advantages and drawbacks of Bayesian statistics

- More intuitive interpretation of the results
- Easier to think about uncertainty
- In a hierarchical setting, it becomes easier to take into account all the sources of variability
- Prior specification: need to check that changing your prior does not change your result
- Computationally intensive

Example: Bernoulli

Take $X_i \sim Bernoulli(\theta)$, *i.e.*

$$P[X_i = 1] = \theta \quad P[X_i = 0] = 1 - \theta.$$

Possible prior: $\theta \sim \mathcal{U}([0,1])$: $\pi(\theta) = 1$ for $0 \leq \theta \leq 1$.
Likelihood:

$$L(\theta|X_i) = \theta^{X_i}(1-\theta)^{1-X_i}$$

$$L(\theta|X_1,\ldots,X_n) = \theta^{\sum X_i}(1-\theta)^{n-\sum X_i} = \theta^{S_n}(1-\theta)^{n-S_n}$$

Posterior, with $S_n = \sum_{i=1}^n X_i$:

$$\pi(\theta|X_1,\ldots,X_n) \propto 1 \cdot \theta^{S_n}(1-\theta)^{n-S_n}$$

We can compute the normalizing constant analytically:

$$\pi(\theta|X_1,\ldots,X_n) = \frac{(n+1)!}{S_n!(n-S_n)!}\theta^{S_n}(1-\theta)^{n-S_n}$$

Conjugate prior

Suppose we take the prior $\theta \sim Beta(\alpha, \beta)$:

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Then the posterior verifies

$$\pi(\theta|X_1,\ldots,X_n) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \theta^{S_n}(1-\theta)^{n-S_n}$$

hence

$$\theta|X_1,\ldots,X_n \sim Beta(\alpha+S_n, \beta+n-S_n).$$

Whatever the data, the posterior is in the same family as the prior: we say that the prior is *conjugate* for this model. This is very convenient mathematically.

Jeffrey's prior

Another possible default prior is Jeffrey's prior, which is invariant by change of variables.
Let ℓ be the log-likelihood and $\mathcal{I}$ be Fisher's information:

$$\mathcal{I}(\theta) = E\left[\left(\frac{d\ell}{d\theta}\right)^2 \middle| X \sim \mathcal{P}_\theta\right] = -E\left[\frac{d^2}{d\theta^2}\ell(\theta; X) \middle| X \sim \mathcal{P}_\theta\right].$$

Jeffrey's prior is defined by

$$\pi(\theta) \propto \sqrt{\mathcal{I}(\theta)}.$$

Invariance of Jeffrey's prior

Let ϕ be an alternate parameterization of the model. Then the prior induced on ϕ by Jeffrey's prior on θ is

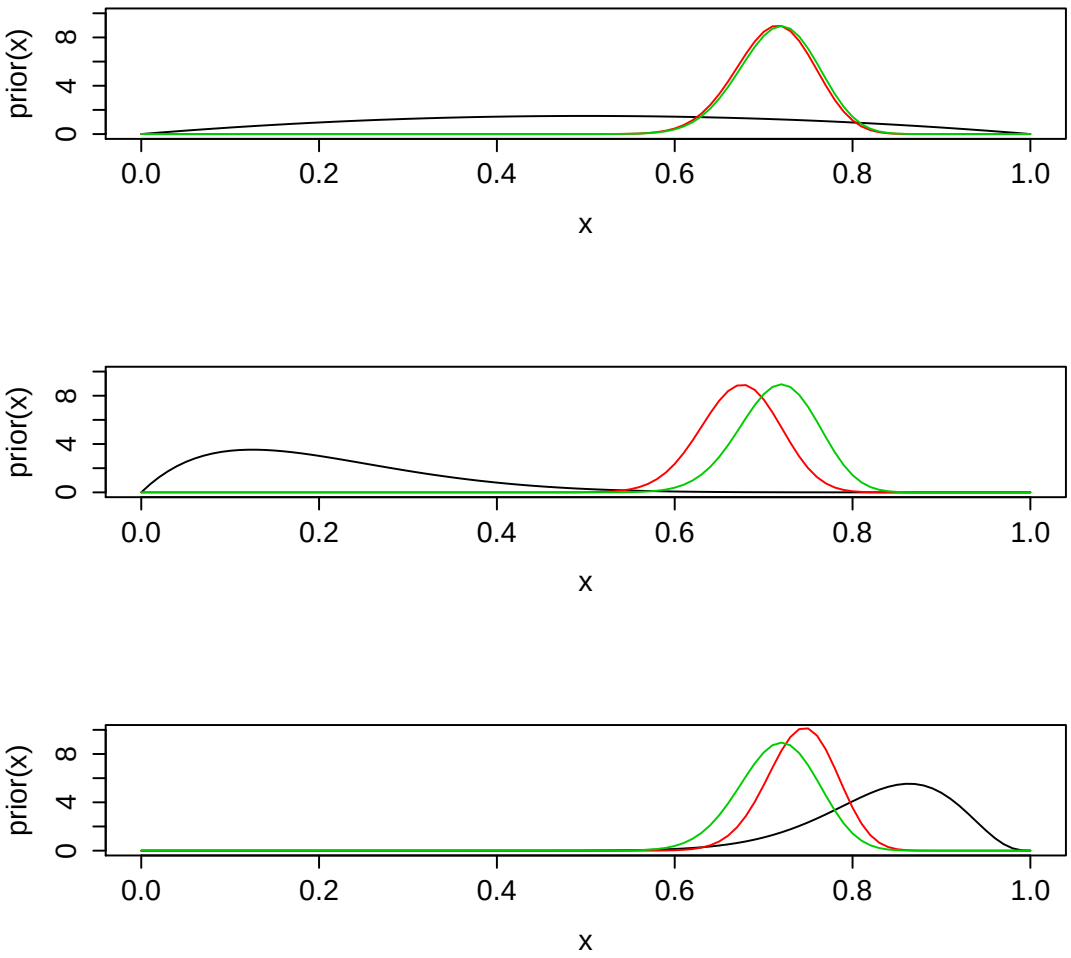
$$\begin{aligned}
\pi(\phi) &= \pi(\theta)\left|\frac{d\theta}{d\phi}\right| \\
&\propto \sqrt{\mathcal{I}(\theta)\left(\frac{d\theta}{d\phi}\right)^2} = \sqrt{E\left[\left(\frac{d\ell}{d\theta}\right)^2\right]\left(\frac{d\theta}{d\phi}\right)^2} = \sqrt{E\left[\left(\frac{d\ell}{d\theta}\frac{d\theta}{d\phi}\right)^2\right]} \\
&= \sqrt{E\left[\left(\frac{d\ell}{d\phi}\right)^2\right]} = \sqrt{\mathcal{I}(\phi)}
\end{aligned}$$

which is Jeffrey's prior on ϕ.

Effect of prior

Example: Bernoulli model (biased coin). θ=probability of success.
Observe $S_n = 72$ successes out of $n = 100$ trials.
Frequentist estimate: $\hat{\theta} = 0.72$
95% confidence interval: [0.63 0.81].
Bayesian estimate: will depend on the prior.

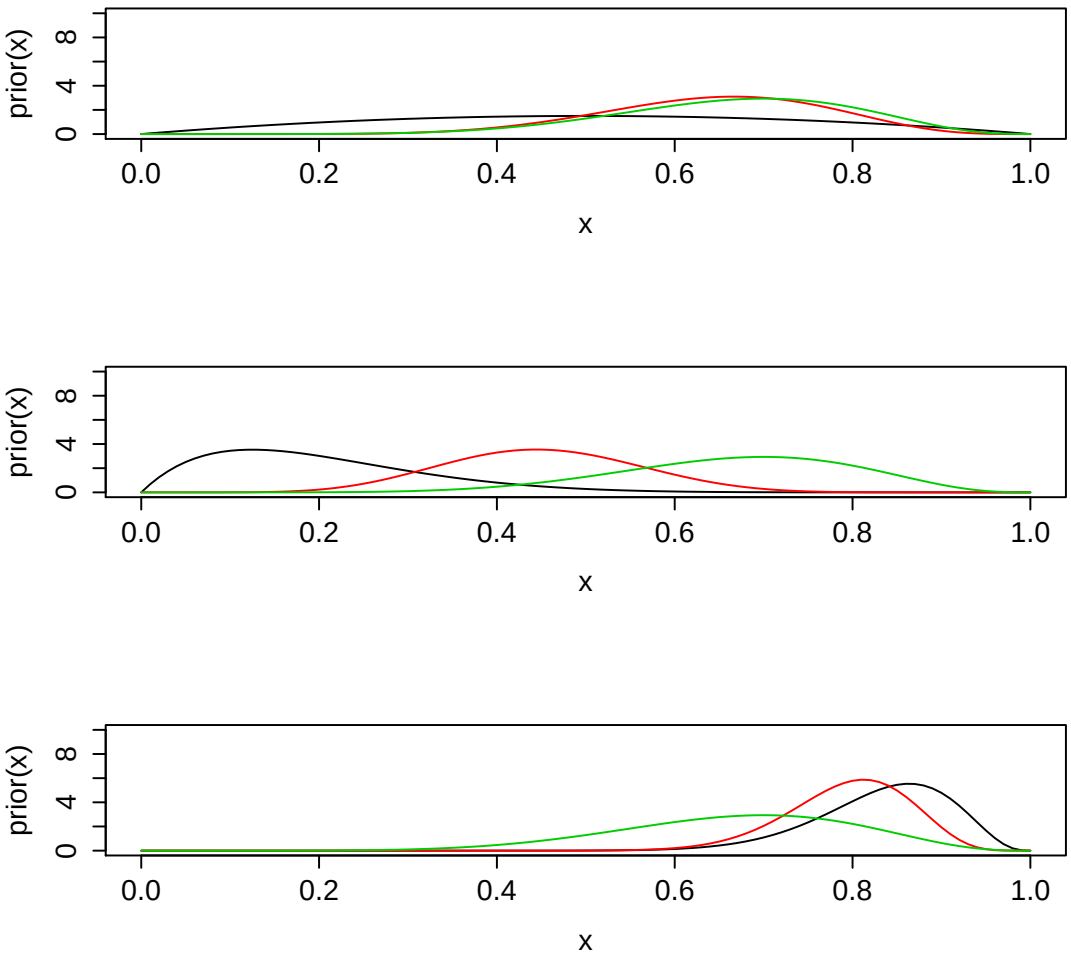
Effect of prior



$S_n = 72,\ n = 100$

Black:prior; green: likelihood; red: posterior.

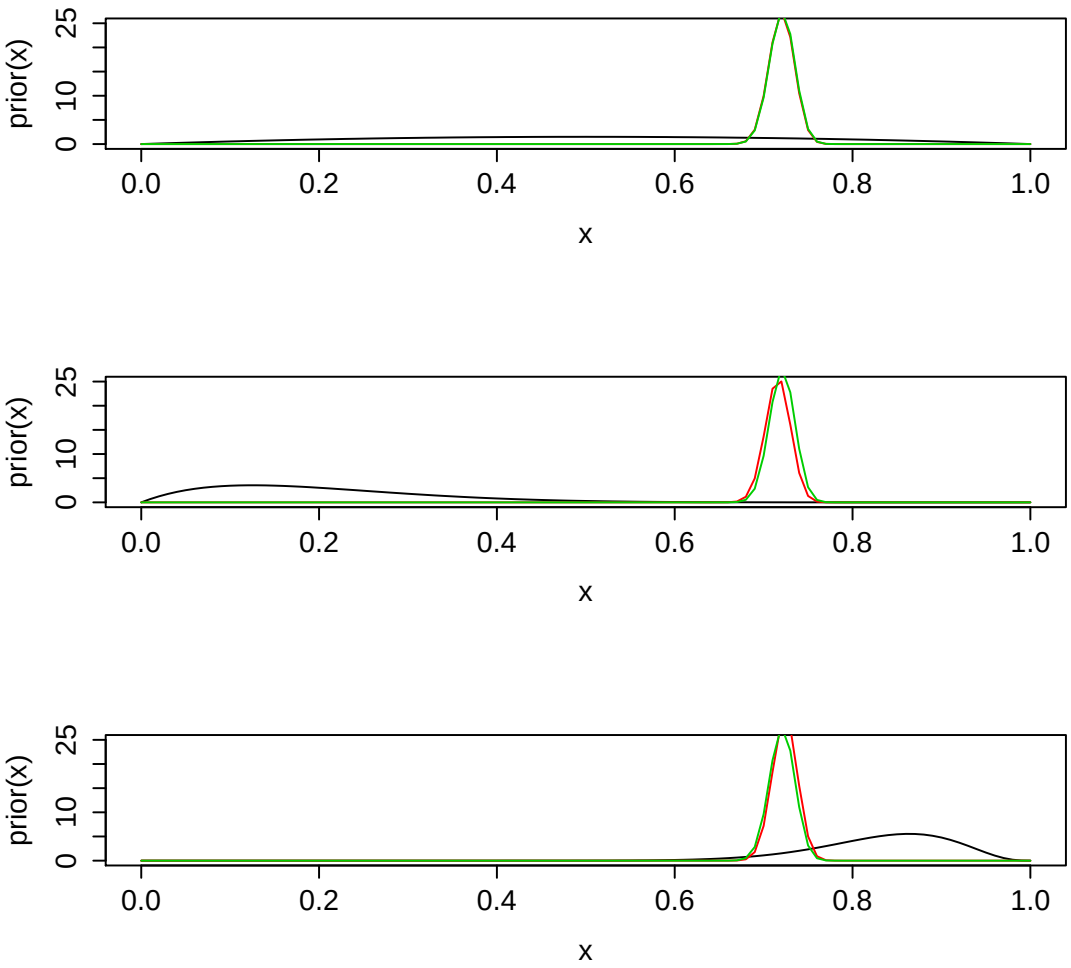
Effect of prior



$S_n = 7,\ n = 10$

Black:prior; green: likelihood; red: posterior.

Effect of prior



$S_n = 721,\ n = 1000$

Black:prior; green: likelihood; red: posterior.

Choosing the prior

The choice of the prior distribution can have a large impact, especially if the data are of small to moderate size. How do we choose the prior?

- Expert knowledge of the application
- A previous experiment
- A conjugate prior, *i.e.* one that is convenient mathematically, with moments chosen by expert knowledge
- A non-informative prior
- ...

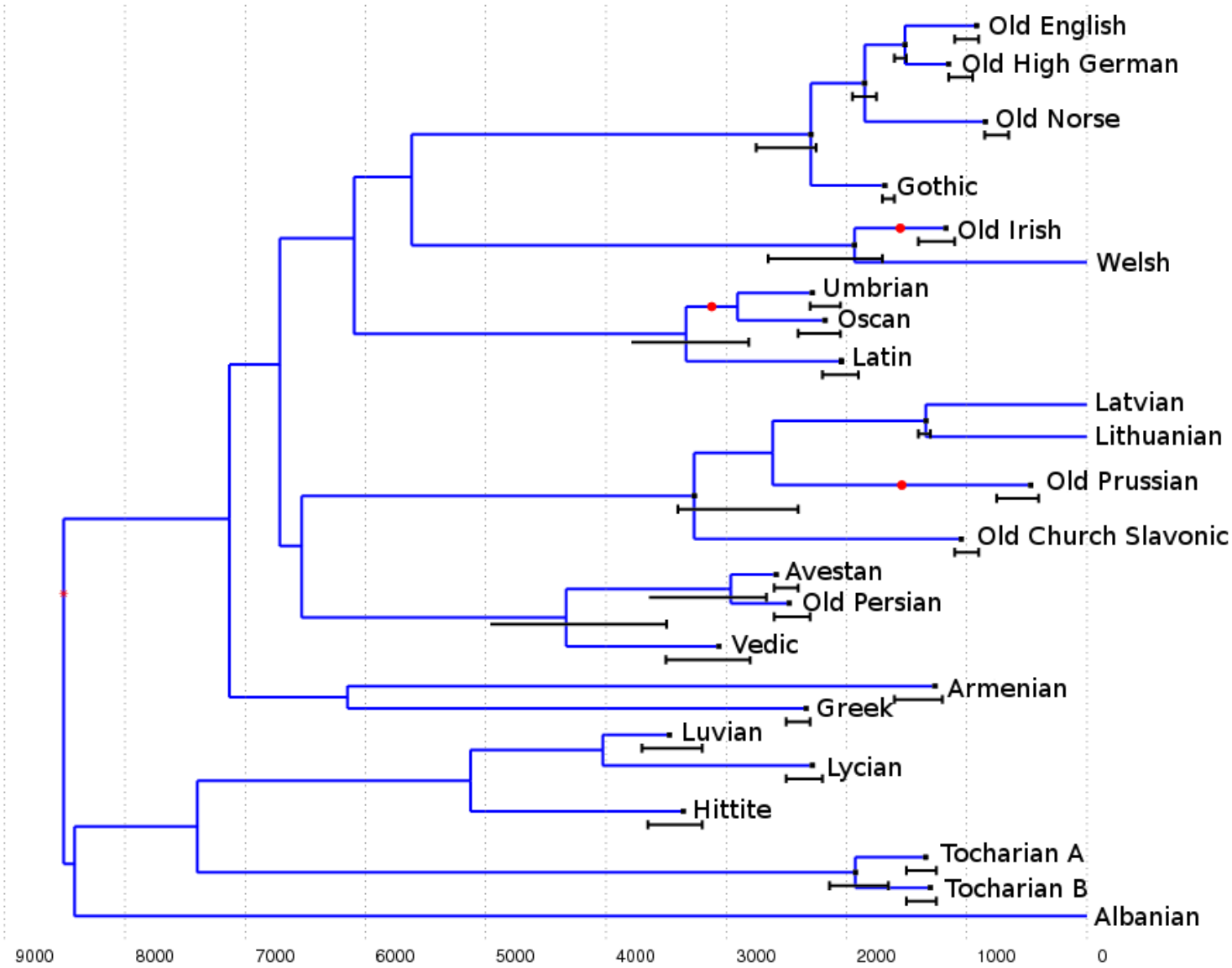
In all cases, the best practice is to try several priors, and to see whether the posteriors agree: would the data be enough to make agree experts who disagreed a priori?

Example: phylogenetic tree

Example from Ryder & Nicholls (2011).
Given lexical data, we wish to infer the age of the Most Recent Common Ancestor to the Indo-European languages.
Two main hypotheses:

- Kurgan hypothesis: root age is 6000-6500 years Before Present (BP).
- Anatolian hypothesis: root age is 8000-9500 years BP

Example of a tree



Why be Bayesian in this setting?

- Our model is complex and the likelihood function is not pleasant
- We are interested in the marginal distribution of the root age
- Many nuisance parameters: tree topology, internal ages, evolution rates...
- We want to make sure that our inference procedure does not favour one of the two hypotheses a priori
- We will use the output as input of other models

For the root age, we choose a prior $\mathcal{U}([5000, 16000])$. Prior for the other parameters is out of the scope of this talk.

Model parameters

Parameter space is large:

- Root age R
- Tree topology and internal ages g (complex state space)
- Evolution parameters λ, μ, ρ, κ
- ...

The posterior distribution is defined by

$$\pi(R, g, \lambda, \mu, \rho, \kappa | D) \propto \pi(R)\pi(g)\pi(\lambda, \mu, \kappa, \rho)L(R, g, \lambda, \mu, \kappa, \rho | D)$$

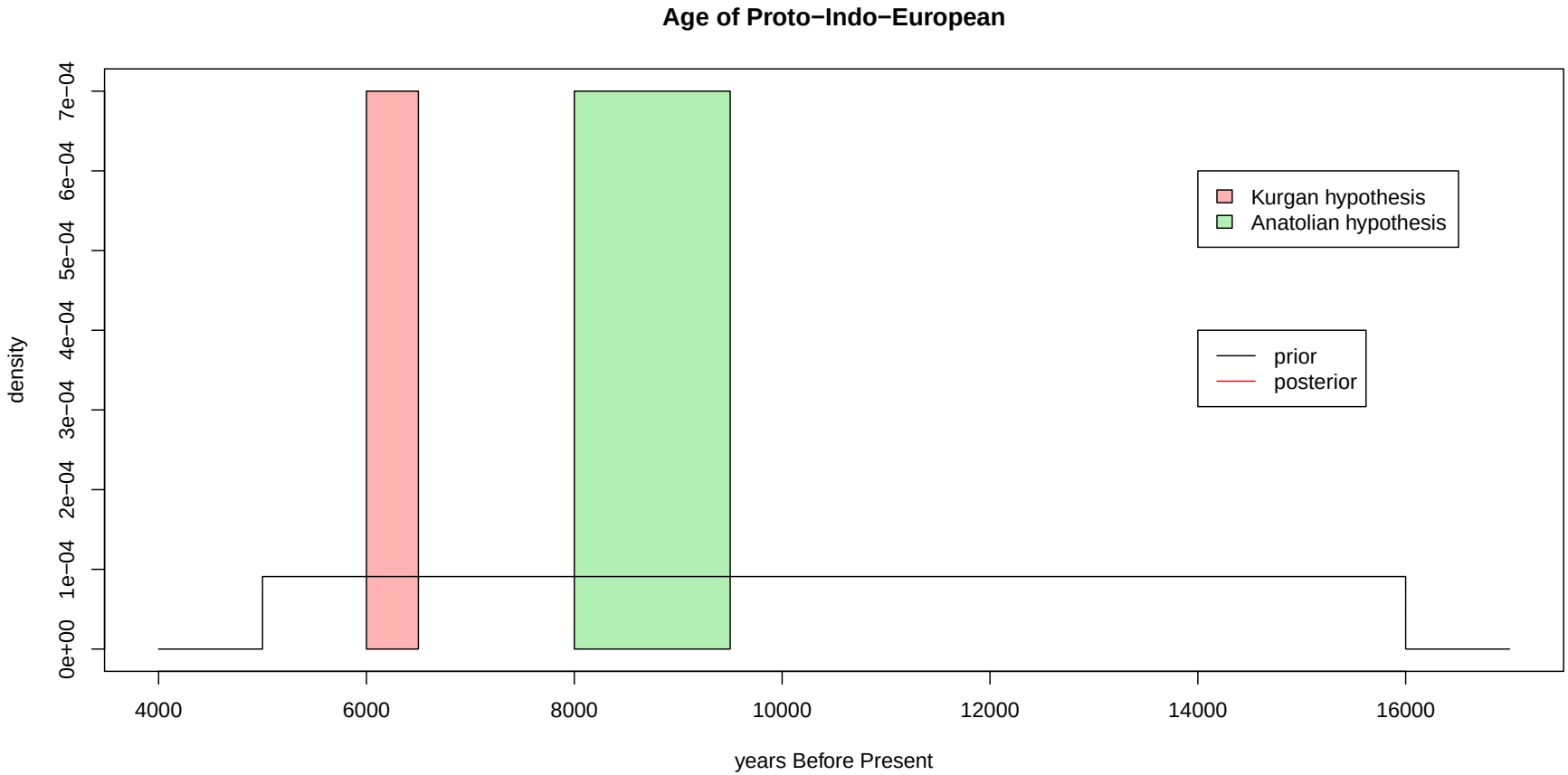
We are interested in the marginal distribution of R given the data D:

$$\pi(R|D) = \int \pi(R, g, \lambda, \mu, \rho, \kappa | D)\, dg\, d\lambda\, d\mu\, d\rho\, d\kappa.$$

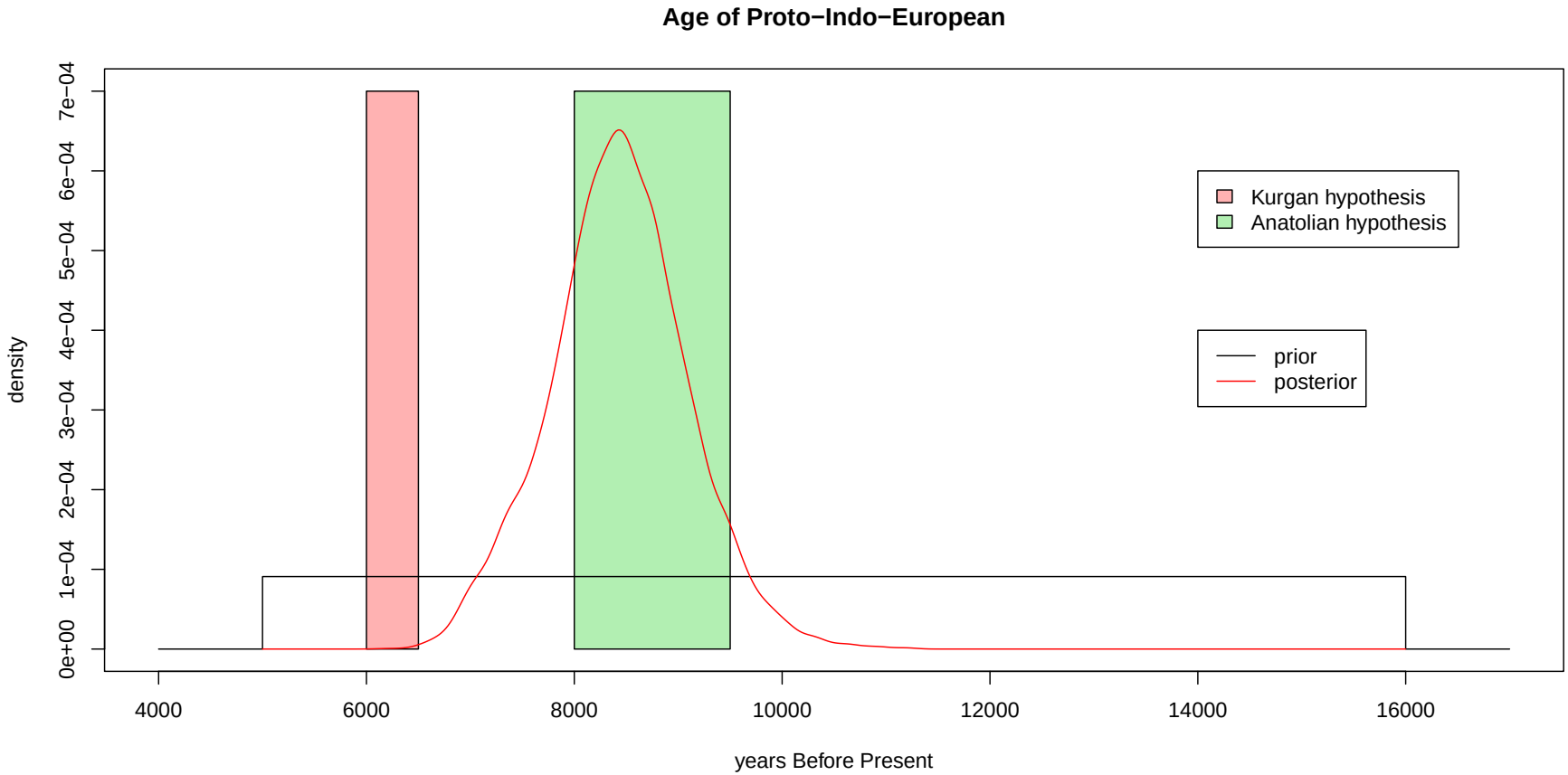
Computation

This distribution is not available analytically, nor can we sample from it directly.
But we can build a Markov Chain Monte Carlo scheme (see Jalal's talk) to get a sample from the joint posterior distribution of $(R, g, \lambda, \mu, \rho, \kappa)$ given D.
Then keeping only the R component gives us a sample from the marginal posterior distribution of R given D.

Root age: prior



Root age: posterior



Phylogenetics tree of languages: conclusions

- Strong support for Anatolian hypothesis; no support for Kurgan hypothesis
- Measuring the uncertainty of the root age estimate is key
- We integrate out the uncertainty of the nuisance parameters
- This setting is much easier to handle in the Bayesian setting than in the frequentist setting
- Computational aspects are complex

Air France Flight 447

This section and its figures are after Stone et al. (Statistical Science 2014)
Air France Flight 447 disappeared over the Atlantic on 1 June 2009, en route from Rio de Janeiro to Paris; all 228 people on board were killed.
The first three search parties did not succeed at retrieving the wreckage or flight recorders.
In 2011, a fourth party was launched, based on a Bayesian search.

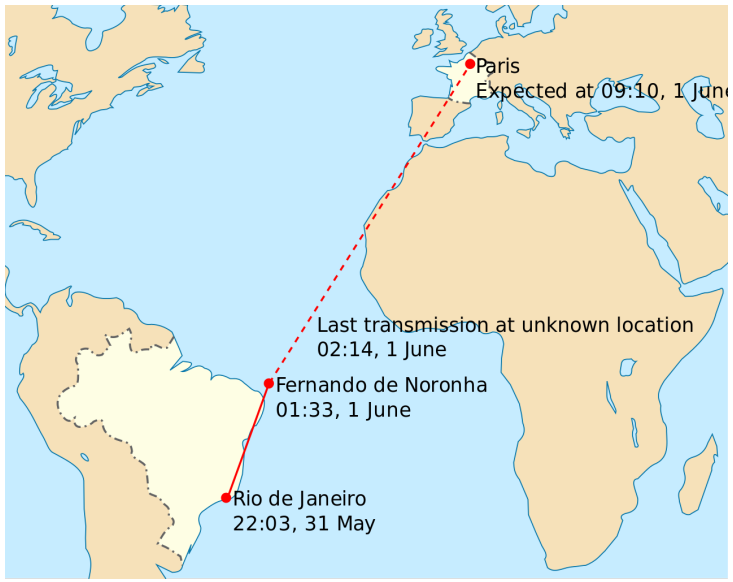
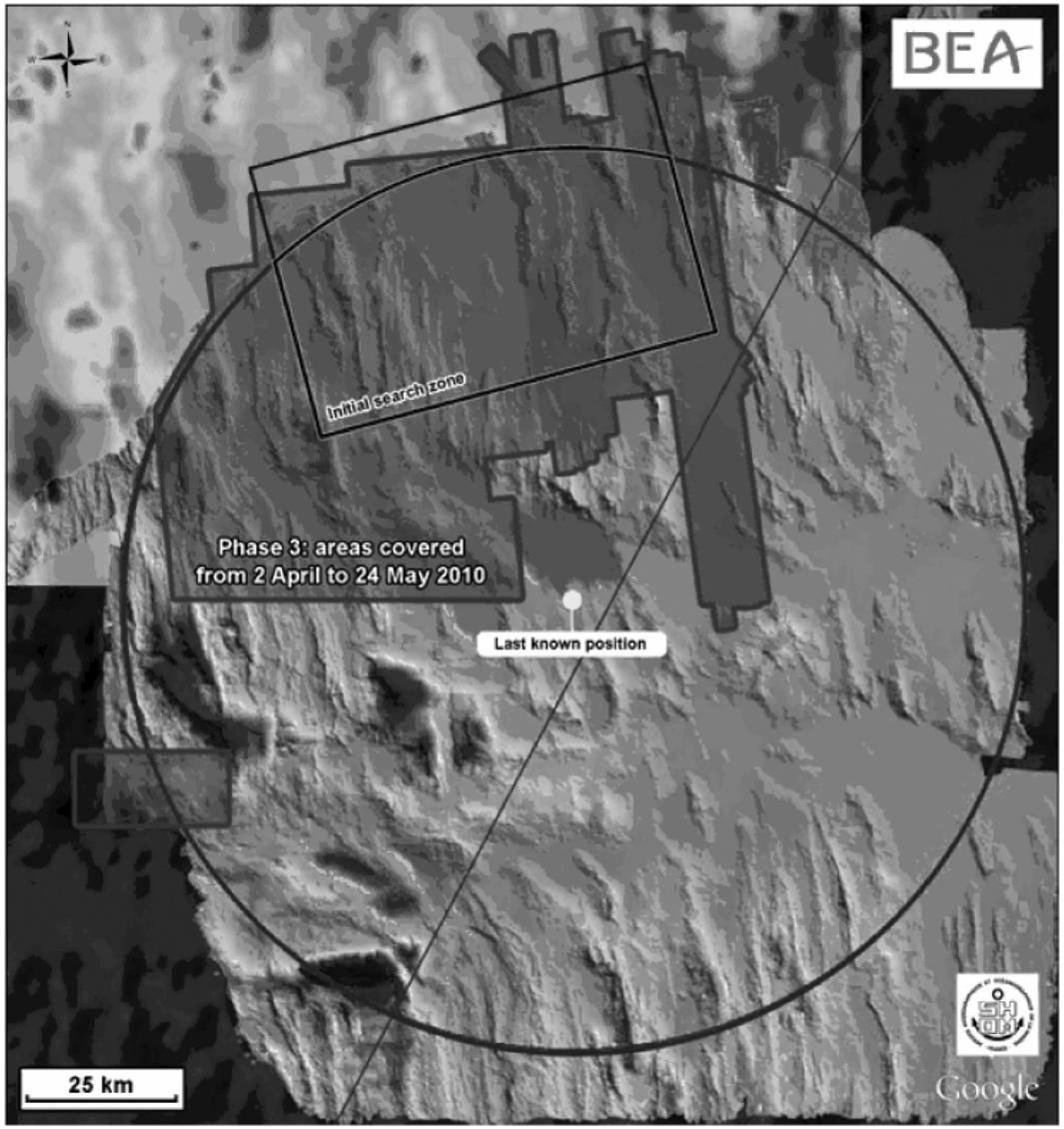


Figure : Flight route. Picture by Mysid, Public Domain.

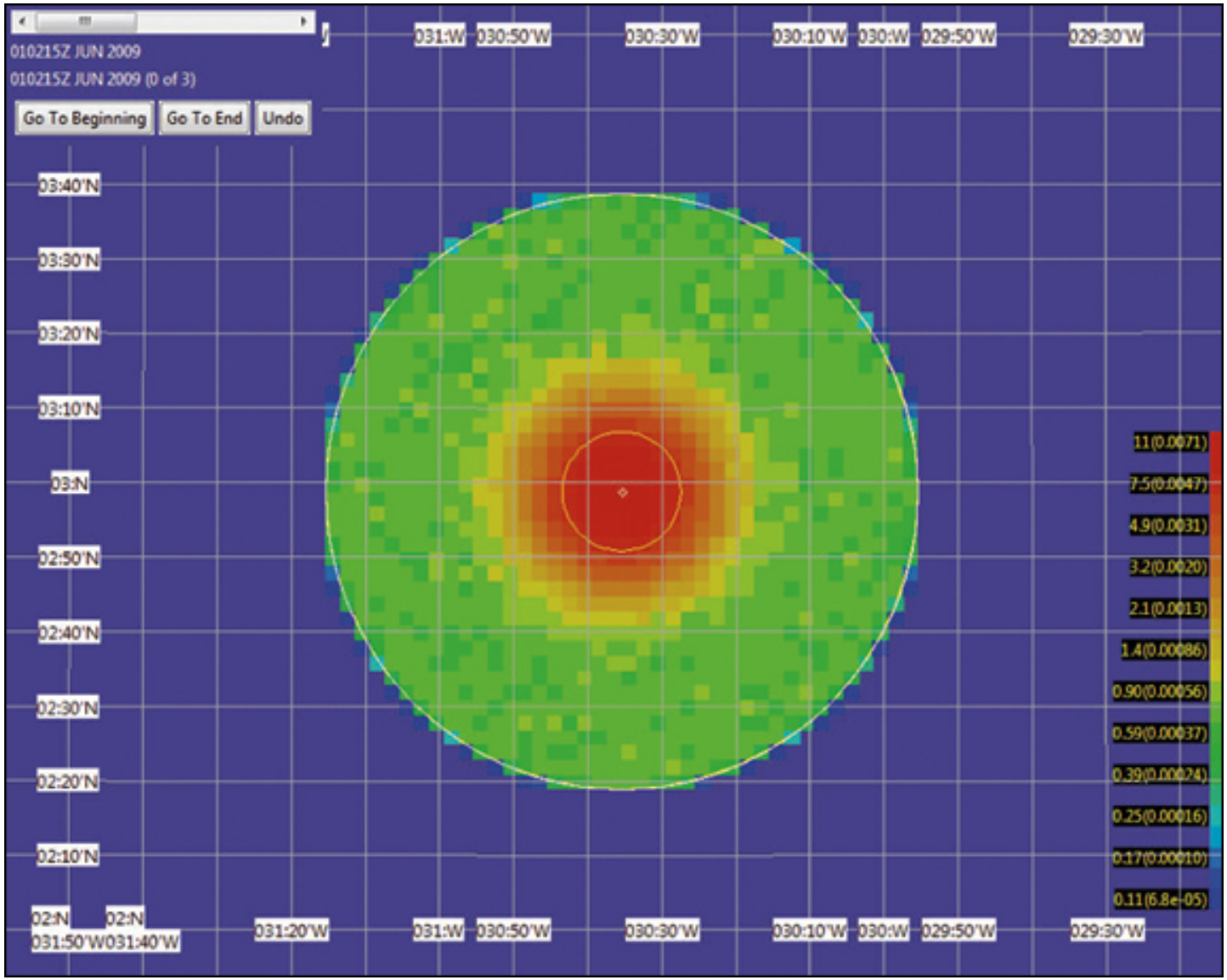
Why be Bayesian?

- Many sources of uncertainties
- Subjective probabilities
- The object of interest is a distribution
- Frequentist formalism does not apply (unique event)

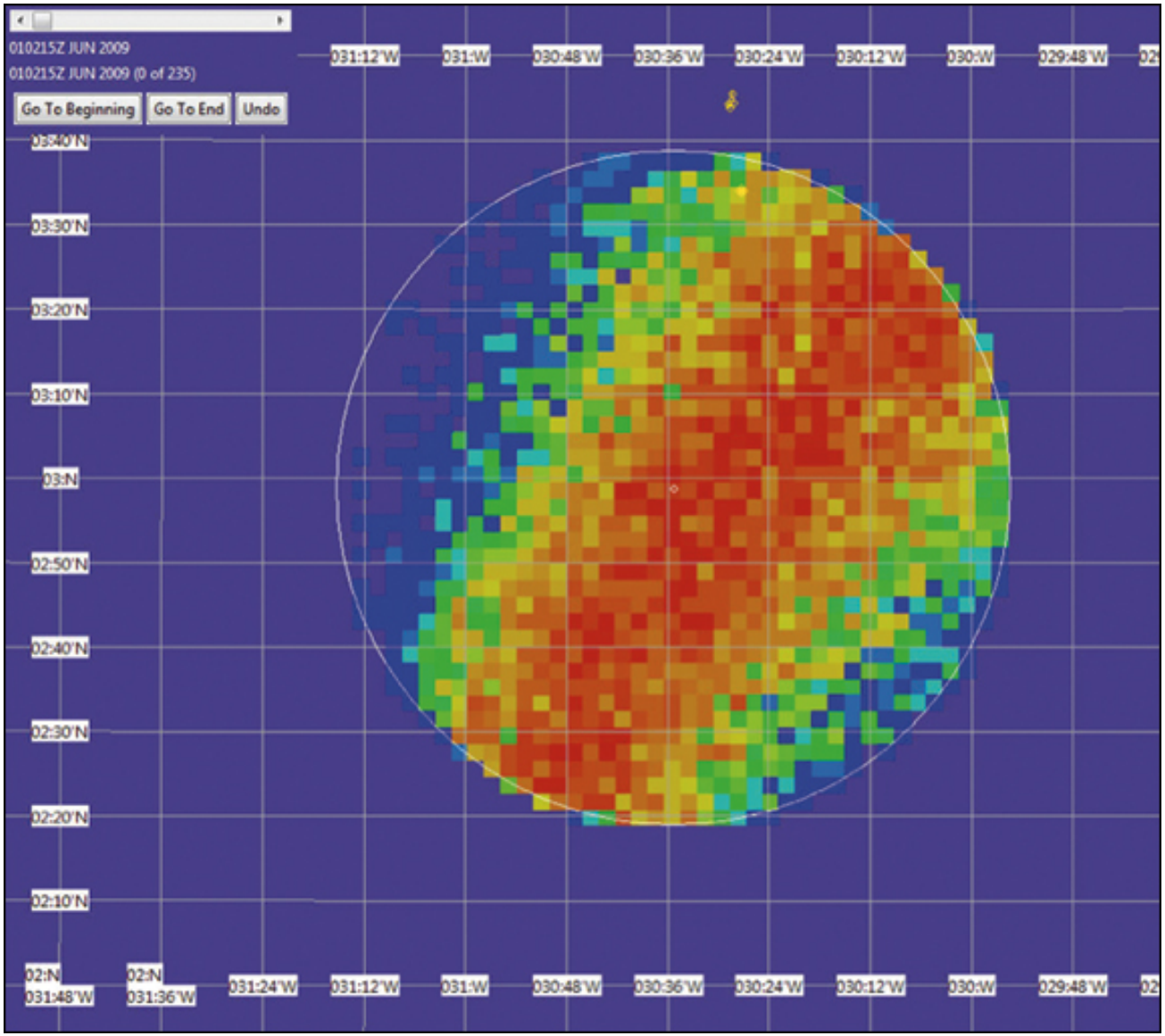
Previous searches



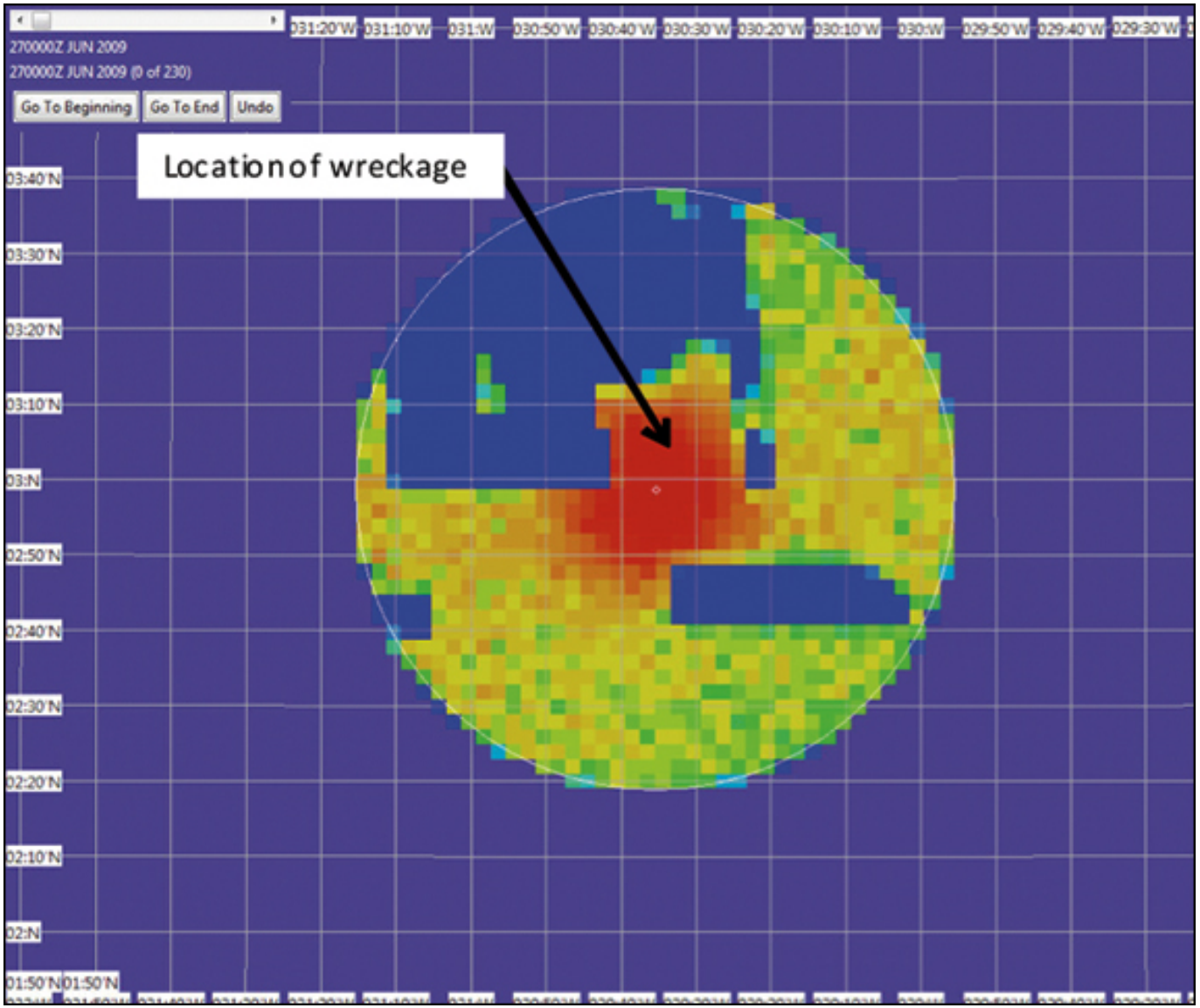
Prior based on flight dynamics



Probabilities derived from drift



Posterior


270000Z JUN 2009
270000Z JUN 2009 (0 of 230)
Go To Beginning
Go To End
Undo
Location of wreckage
031:20'W
031:10'W
031:W
030:50'W
030:40'W
030:30'W
030:20'W
030:10'W
030:W
029:50'W
029:40'W
029:30'W
03:40'N
03:30'N
03:20'N
03:10'N
03:N
02:50'N
02:40'N
02:30'N
02:20'N
02:10'N
02:N
01:50'N
01:50'N

Conclusions

- Once the posterior distribution was derived, the search was organized starting with the areas of highest posterior probability
- Actually several posteriors, because several models were considered
- The wreckage was located in one week

Point estimates

Although one of the main purposes of Bayesian inference is getting a distribution, we can also need to summarize the posterior with a point estimate.
Common choices:

- Posterior mean

$$\hat{\theta} = \int \theta \cdot \pi(\theta|D)\, d\theta$$

- Maximum a posteriori (MAP)

$$\hat{\theta} = \arg\max \pi(\theta|D)$$

- Posterior median
- ...

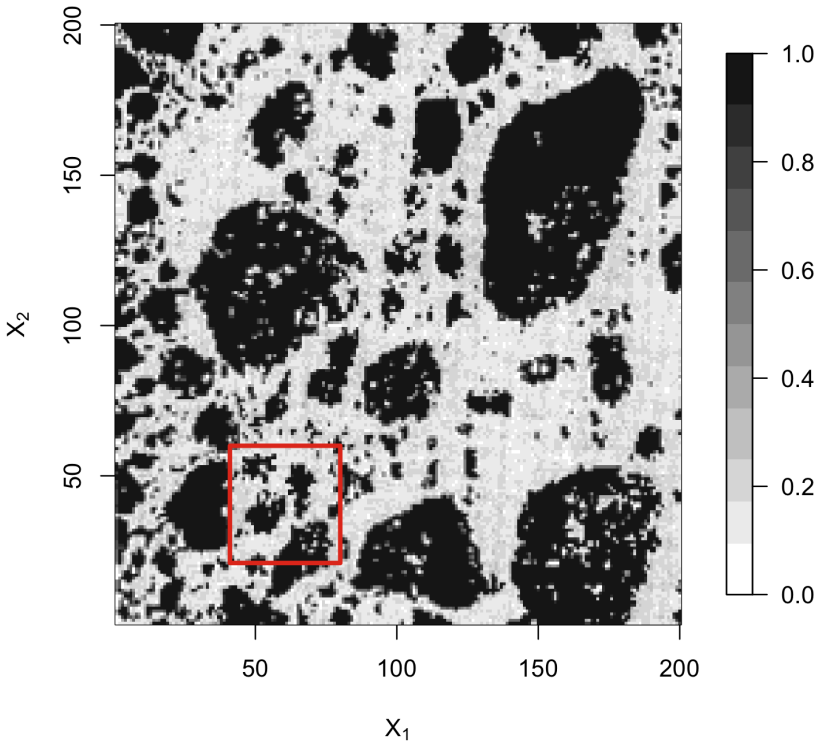
Optimality

From a frequentist point of view, the posterior expectation is optimal under a certain sense.
Let θ be the true value of the parameter of interest, and $\hat{\theta}(X)$ an estimator. Then the posterior mean minimizes the expectation of the squared error under the prior

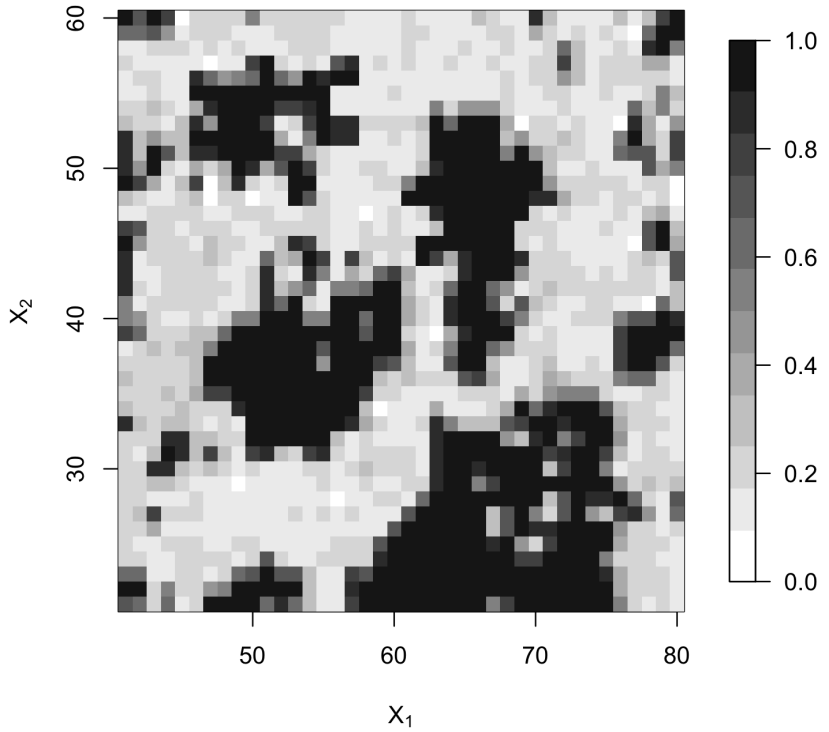
$$E_\pi \left[\|\theta - \hat{\theta}(X)\|_2^2 \right]$$

For this reason, the posterior mean is also called the minimum mean square error (MMSE) estimator.
For other loss functions, other point estimates are optimal.

2D Ising models



(a) Original Image



(b) Focused Region of Image

2D Ising model

Higdon (JASA 1998)

Target density

Consider a 2D Ising model, with posterior density

$$\pi(x|y) \propto \exp\left(\alpha \sum_i \mathbb{1}[y_i = x_i] + \beta \sum_{i \sim j} \mathbb{1}[x_i = x_j]\right)$$

with $\alpha = 1$, $\beta = 0.7$.

- The first term (likelihood) encourages states x which are similar to the original image y.
- The second term (prior) favors states x for which neighbouring pixels are equal, like a Potts model.

2D Ising models: posterior exploration

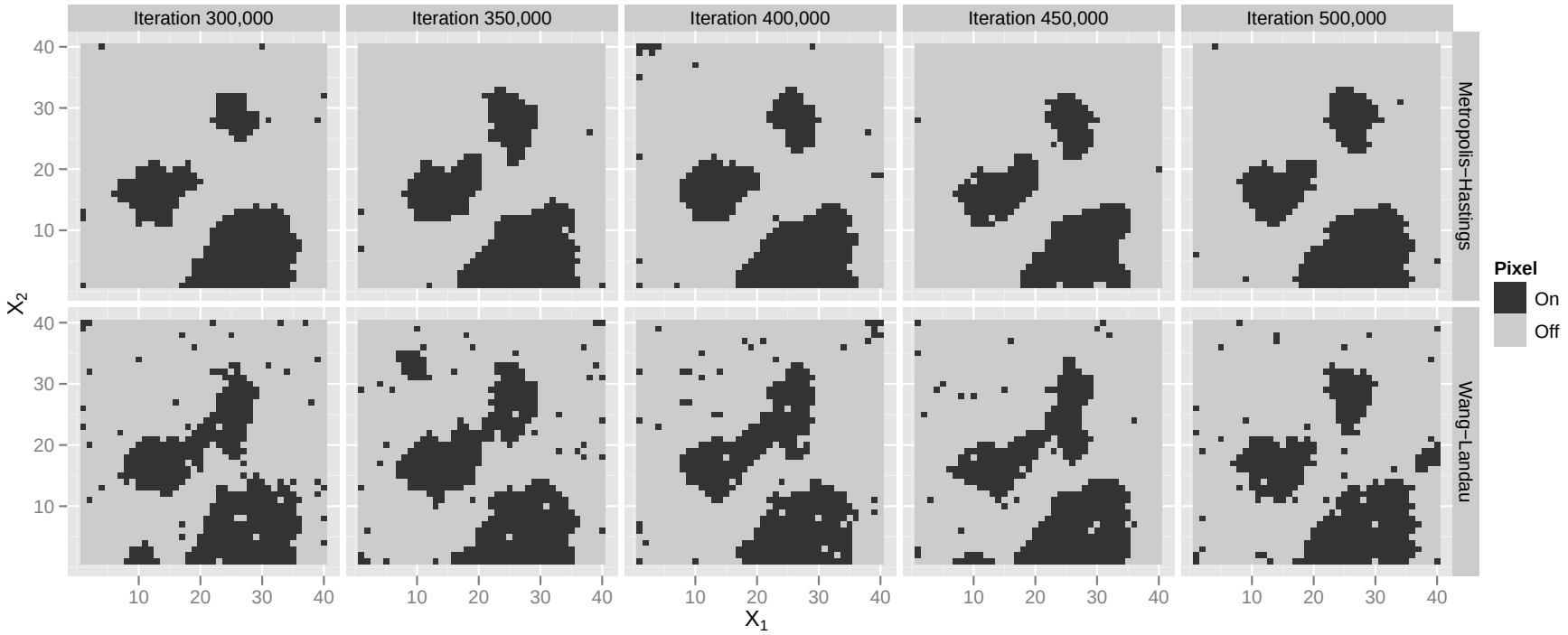


Figure : Spatial model example: states explored over 500,000 iterations for Metropolis-Hastings (top) and Wang-Landau algorithms (bottom). Figure from Bornn et al. (JCGS 2013). See also Jacob & Ryder (AAP 2014) for more on the algorithm.

2D Ising models: posterior mean

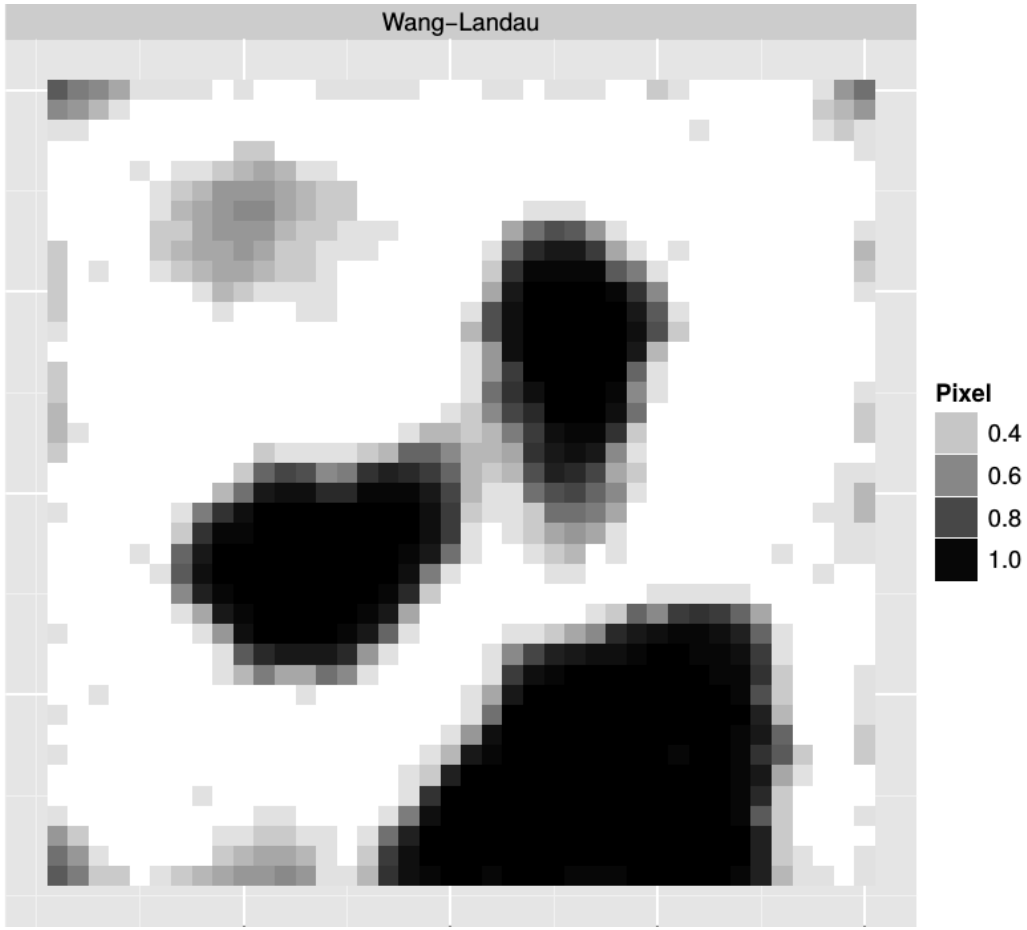


Figure : Spatial model example: average state explored with Wang-Landau after importance sampling. Figure from Bornn et al. (JCGS 2013). See also Jacob & Ryder (AAP 2014) for more on the algorithm.

Ising model: conclusions

- Problem-specific prior
- Even with a point estimate (posterior mean), we measure uncertainty
- Computational cost is very high

Several models

Suppose we have several models $m_1, m_2, \ldots, m_k$. Then the model index can be viewed as a parameter.
Take a uniform (or other) prior:

$$P[\mathcal{M} = m_j] = \frac{1}{k}.$$

The posterior distribution then gives us the probability associated with each model given the data.
We can use this for model choice (but there are other, more sophisticated, techniques) but also for estimation/prediction while integrating out the uncertainty on the model.

Example: variable selection for linear regression

A model is a choice of covariables to include in the regression. With p covariables, there are 2^p models.
Classical (frequentist) setting:

- Select variables, using your favourite penalty, thus selecting one model
- Perform estimation and prediction within that model
- If you want error bars, you can compute them, but only within that model

Bayesian setting:

- Explore space of all models
- Get posterior probabilities
- Compute estimation and prediction for each model (or, in practice, for those with non negligible probability)
- Weight these estimates/predictions by the posterior probability of each model

The uncertainty about the model is thus fully taken into account.

Conclusions

- Bayesian inference is a powerful tool to fully take into account all sources of uncertainty
- Difficulty of prior specification
- Computational issues are the main hurdle