Lecture 3: how to speak Stan, infer parameters for ODEs and for hierarchical models

Ben Lambert[1]
ben.c.lambert@gmail.com

[1]Imperial College London

Tuesday 23rd July, 2019

Lecture outcomes

By the end of this lecture you should:

1. Know how to start coding up a model in Stan.
2. Know what to do when coding or sampling goes wrong.
3. Appreciate how to fit ODE models in a Bayesian way (using Stan).
4. Be able to perform model comparison.

Coding up MCMC algorithms

- Up until now we have coded up our own MCMC algorithms $\Longrightarrow$ (relatively) straightforward for simple models.
- However for more complex models this is a time-consuming (and frustrating) process.
- Flavours of MCMC algorithms:
  - **Random Walk Metropolis:** fairly basic but ineffectual for many real-life models.
  - **Gibbs:** a bit faster than RWM but also suffers from same issues. However can be significantly harder to code up!
  - **Hamiltonian Monte Carlo:** significantly more efficient than the aforementioned but also significantly harder to code and tune.

Introducing Stan: avoiding manual labour

Stan is an intuitive yet sophisticated programming language that does the hard work for us.
What is Stan and how do we use it?

- **Imperative** programming language like R, Python, Matlab, C++ etc. (BUGS/JAGS are a weird type of language called **declarative**.)
- Turing-complete language (unlike BUGS/JAGS) and works like most other languages: can use loops, conditional statements, and functions.
- Code up a model in Stan and then it implements Hamiltonian Monte Carlo (actually something called NUTS but similar) for us.

Why should we use Stan?

- Stan is the brainchild of Andrew Gelman at Colombia; the World's foremost Bayesian statistician.
- Stan's uses an extension of HMC called NUTS that automatically tunes. It is **fast**. Very fast $\Longrightarrow$ typically generates multiples times as many effective samples per second than BUGS/JAGS.
- Stan is **simple** to learn.
- Stan has excellent documentation (a manual full of extensive examples).
- The Stan team have translated **all** the example models from popular books; Gelman, Kruschke etc.
- **Most important:** Stan has a very active and helpful user forum and development team; for example, typical question answered in less than a couple of hours.

Why should we use Stan?

Overall: Stan makes our life easier.

- It is easy to learn; even if you are used to BUGS/JAGS or something else.
- The language is here to stay $\Longrightarrow$ all recent books use examples written in Stan rather than BUGS/JAGS.
- **Important:** it is popular $\Longrightarrow$ if you have a problem with your model you can get help, **fast**.
- The best minds in the business are working on making it even better.
- Finally, "Shiny Stan" makes it really quite fun!

How do we use it?

- Code up model in Stan code in a text editor and save as ".stan" file.
- Call Stan to run the model from:
  - R.
  - Python.
  - The command line.
  - Matlab.
  - Stata.
  - Julia.
- Use one of the above to analyse the data (of course you can export to another one.)

A straightforward example

Suppose:

- We record the height, Y_i, of 10 people.
- We want a model to explain the variation, and choose a normal likelihood:

$$Y_i \sim N(\mu, \sigma) \tag{1}$$

- We choose the following (independent) priors on each parameter:
  - $\mu \sim N(1.5, 1)$.
  - $\sigma \sim \text{gamma}(1, 1)$.

  Question: how do we code this up in Stan?

An example Stan program: heights

```
data {
    real Y[10]; // Heights for 10 people
}

parameters {
    real mu;
    real<lower=0> sigma;
}

model {
    Y ~ normal(mu,sigma); // Likelihood
    mu ~ normal(1.5,1); // Prior for mu
    sigma ~ gamma(1,1); // Prior for sigma
}
```

An example Stan program: data block

```
data {
    real Y[10]; // Heights for 10 people
}
```

- Declare all data that you will pass to Stan to estimate your model.
- Terminate all statements with a semi-colon ";".
- Use "//" for comments.

An example Stan program: data block

```
data {
    real Y[10]; // Heights for 10 people
}
```

Strongly, statically-typed language: need to tell Stan the type of data variable. For example:

- `real` for continuous data.
- `int` for discrete data.
- Arrays: above we specified `Y` as an array of continuous data of length 10.

An example Stan program: data block

```
data {
    real Y[10]; // Heights for 10 people
}
```

- Can place limits on data, for example:
  - `real<lower=0,upper=1> X`
  - `real<lower=0> Z`
- Vectors and matrices; only contain reals and can be used for matrix operations.

An example Stan program: parameter block

```
parameters {
    real mu;
    real<lower=0> sigma;
}
```

- Declare all parameters that you use in your model.
- Place limits on variables, for example `real<lower=0> sigma` above.
- A multitude of parameter types including some of the aforementioned:
  - `real` for continuous parameters.
  - Arrays of types, for example `real epsilon[12]`.

An example Stan program: parameter block

```
parameters {
    real mu;
    real<lower=0> sigma;
}
```

- `vector` or `matrix`, specified by:
  - `vector[5] beta`
  - `matrix[5,3] gamma`
- `simplex` for a parameter vector that must sum to 1.
- More exotic types like `corr_matrix` , or `ordered`.

An example Stan program: parameter block

```
parameters {
    real mu;
    real<lower=0> sigma;
}
```

Important: HMC/NUTS not developed yet to work with **discrete** parameters $\implies$ following options in Stan:

- Marginalise out the parameter. For example, if $p(\beta, \theta)$ where β is continuous and θ is discrete:

$$p(\beta) = \sum_{i=1}^{K} p(\beta, \theta_i) \tag{2}$$

- Some models can be reformulated without discrete parameters.
- Failing either of the above $\implies$ use Gibbs sampling.

An example Stan program: model block

```
model {
    Y ~ normal(mu,sigma); // Likelihood
    mu ~ normal(1.5,1); // Prior for mu
    sigma ~ gamma(1,1); // Prior for sigma
}
```

- Used to define:
  - Likelihood.
  - Priors on parameters.
- If don't specify priors on parameters $\Longrightarrow$ Stan assumes you are using flat priors (which can be improper.)

An example Stan program: model block

```
model {
    Y ~ normal(mu,sigma); // Likelihood
    mu ~ normal(1.5,1); // Prior for mu
    sigma ~ gamma(1,1); // Prior for sigma
}
```

- **Huge** range of probability distributions covered, across a range of parameterisations. For example:
  - **Discrete:** bernoulli, binomial, normal, poisson, beta-binomial, negative-binomial, categorical, multinomial.
  - **Continuous unbounded:** normal, skew-normal, student-t, cauchy, logistic.
  - **Continuous bounded:** uniform, beta, log-normal, exponential, gamma, chi-squared, inverse-chi-squared, weibull, wiener diffusion, pareto.

An example Stan program: model block

```
model {
    Y ~ normal(mu,sigma); // Likelihood
    mu ~ normal(1.5,1); // Prior for mu
    sigma ~ gamma(1,1); // Prior for sigma
}
```

- **Multivariate continuous:** normal, student-t, gaussian process.
- **Exotics:** dirichlet, LKJ correlation distribution, wishart and its inverse, von-mises.

Running Stan

Write model in a text editing program. For example:

- **emacs** - Stan syntax actually supported.
- **notepad++**.
- **Word** or **notepad**.

$\Longrightarrow$ save as "exampleModel.stan" in same directory as you run statistical software.

Running Stan in R

```
## Load packages
library(rstan)

## Generate fake data
Y = rnorm(10,mean = 0, sd = 1)

## Compile and run model, and save in fit
fit = stan(file='exampleModel.stan',data=list(Y=Y),
           iter=1000,chains=4)
```

Running Stan on example model

```
## Compile and run model, and save in fit
fit = stan(file='exampleModel.stan',data=list(Y=Y),
           iter=1000,chains=4)
```

The above R code runs NUTS for our model with the following options:

- 1000 MCMC samples of which 500 are discarded as warm-up.
- Across 4 chains.
- Using a random number seed of 1 (good to ensure you can reproduce results.)

Example model: results

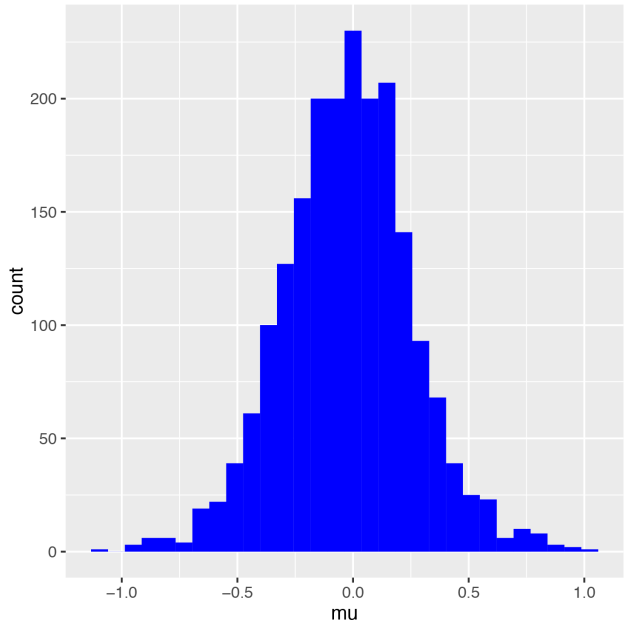
```
## Print summary statistics
print(fit,probs = c(0.25, 0.5, 0.75))

## Inference for Stan model: exampleModel.
## 4 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500
##
##         mean se_mean   sd   25%   50%   75% n_eff Rhat
## mu     -0.02    0.01 0.28 -0.20 -0.02  0.16   689    1
## sigma  0.95    0.01 0.24  0.78  0.91  1.08   696    1
## lp__  -5.26    0.04 1.00 -5.69 -4.97 -4.50   639    1
```

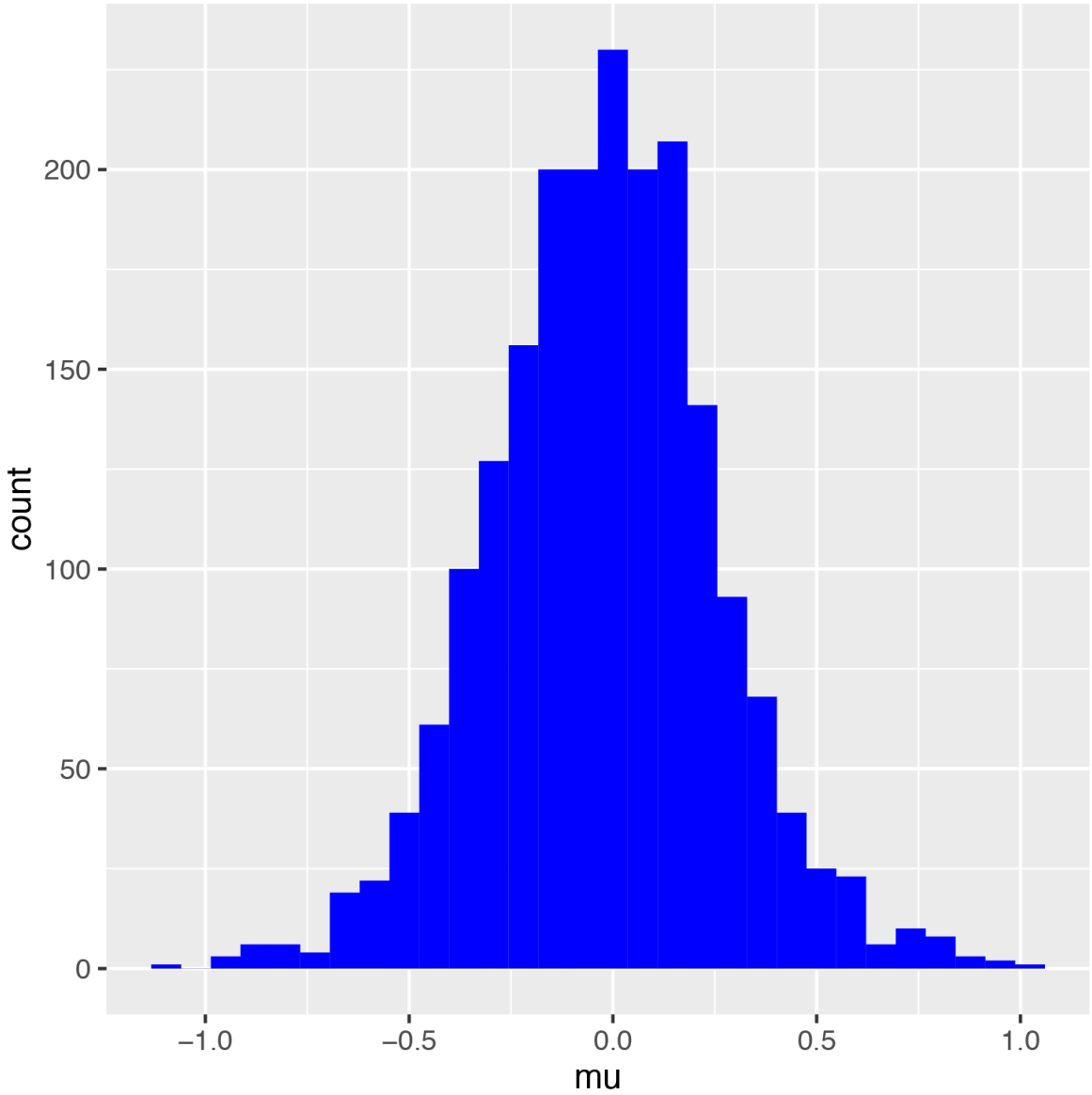
Example model: results

```
## Extract element and plot
mu = extract(fit,'mu')[[1]]

## Plot histogram
library(ggplot2)
qplot(mu, fill=I("blue"))
```



Example model: results



Quick note: what does $\sim$ actually mean?

```
model {
    Y ~ normal(mu,sigma); // Likelihood
    mu ~ normal(1.5,1); // Prior for mu
    sigma ~ gamma(1,1); // Prior for sigma
}
```

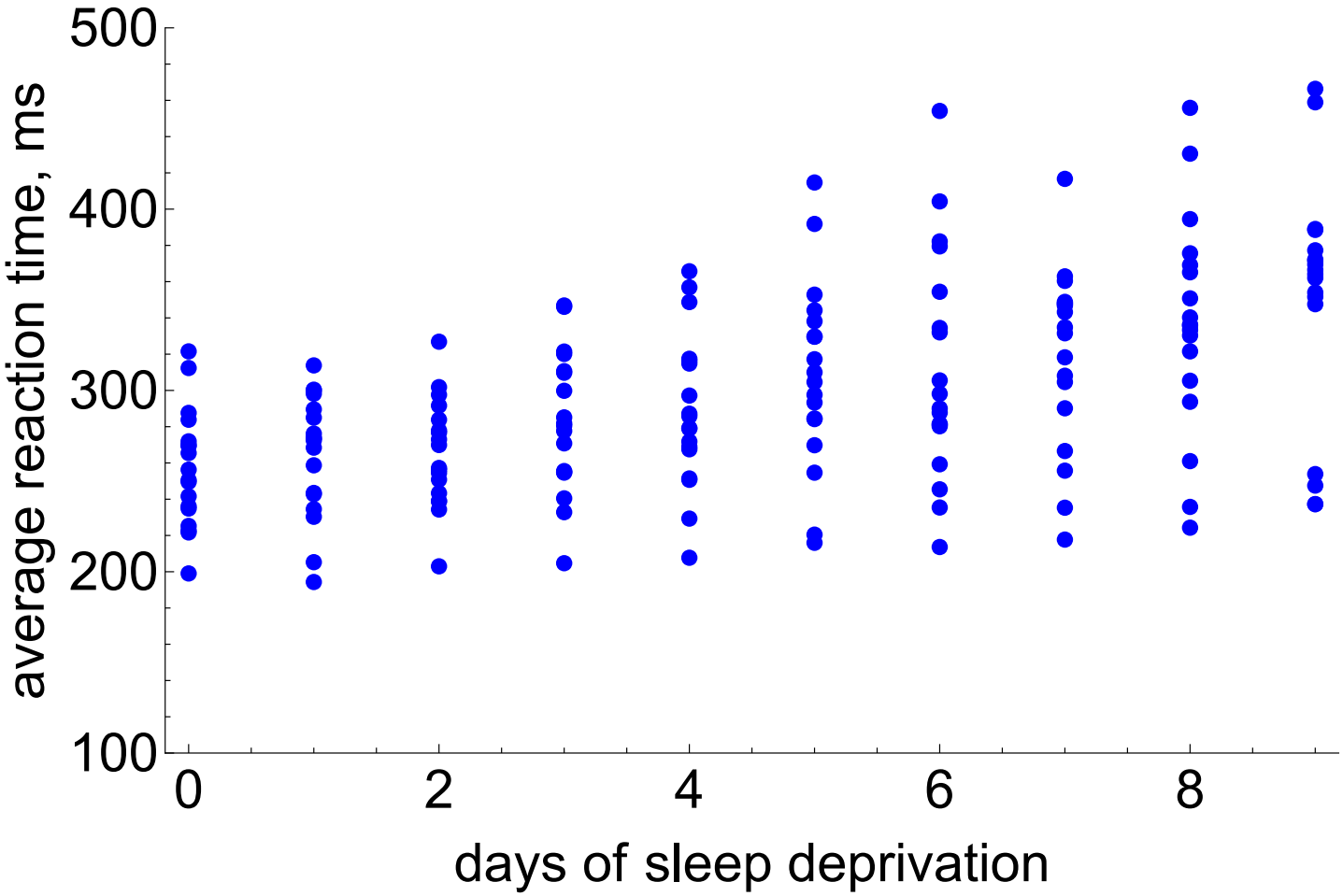
- $\sim$ doesn't mean "sampling" although for many circumstances it isn't terrible to think of it like that.
- Remember: HMC/NUTS uses the negative of the log-posterior to find potential energy of puck.
- $\implies$ logging the posterior converts it from a product into a sum.
- As such $\sim$ really means "increment log probability".

Example: sleep deprivation study

- Data from a laboratory experiment that measured the effect of sleep deprivation on cognitive performance[1].
- 18 subjects within a population of interest - long-distance lorry drivers - volunteered to participate in the 10 day experiment.
- Subjects were restricted to 3 hours of sleep per night.
- On each day the subjects' reaction time across a range of cognitive tasks were measured.

Sleep deprivation study: model aims

- Build a model to explain the effect of sleep deprivation on reaction times.



Sleep deprivation study: model

On the basis of the previous graph, assume:

$$R(t) \sim N(\alpha + \beta t, \sigma) \tag{3}$$

where $R(t)$ is the reaction time for a sleep deprivation of t days.

Question: how can we code this up in Stan?

Sleep deprivation study: 1st part of Stan code

```
data {
    int N; // number of observations
    vector[N] t; // days of sleep deprivation
    vector[N] R; // reaction times
}
parameters {
    real alpha; // reaction time at start
    real beta; // daily increment to reaction time
    real<lower=0> sigma; // variation about mean
}
```

Sleep deprivation study: 2nd part of Stan code

```
model {
    // likelihood
    for (i in 1:N){
        R[i] ~ normal(alpha + beta * t[i], sigma);
    }

    // priors
    alpha ~ normal(0,250);
    beta ~ normal(0,250);
    sigma ~ normal(0,50);
}
```

Sleep deprivation study: 2nd part of Stan code

However can write same model with faster and more efficient Stan code using vectorization.

```
data {
    int N; // number of observations
    matrix[N,2] X; // ones + days of sleep depriv.
    vector[N] R; // reaction times
 }
 parameters {
    vector[2] gamma;
    real<lower=0> sigma;
  }
model {
     // likelihood
     R ~ normal(X * gamma, sigma);
     gamma ~ normal(0,250);
}
```

Sleep deprivation study: posterior predictive distribution

- Want to carry out posterior predictive checks $\Longrightarrow$ need posterior predictive distribution.
- Use "generated quantities" block.

```
generated quantities {
    vector[N] R_sim; // Store post-pred samples
    for (i in 1:N){
        R_sim[i] = normal_rng(X[i] * gamma, sigma);
    }
}
```

Sleep deprivation study: posterior predictive distribution

```
generated quantities {
        vector[N] R_sim; // Store post-pred samples
        for (i in 1:N){
        R_sim[i] = normal_rng(X[i] * gamma, sigma);
    }
}
```

The function `normal_rng` generates a single **independent** sample from a normal distribution with parameters:

- mean $= X[i] * gamma$, where *gamma* is a sample from the estimated posterior.
- std. dev $=$ *sigma*, where *sigma* is a sample from the estimated posterior.

Sleep deprivation study: model changes

- Suppose based on posterior predictive checks we want to use a wider sampling distribution $\implies$ use a Student T.
- If we were coding this up ourselves this would involve a large structural change in the code and MCMC algorithm.

Question: how long does it take us to recode our model in Stan?

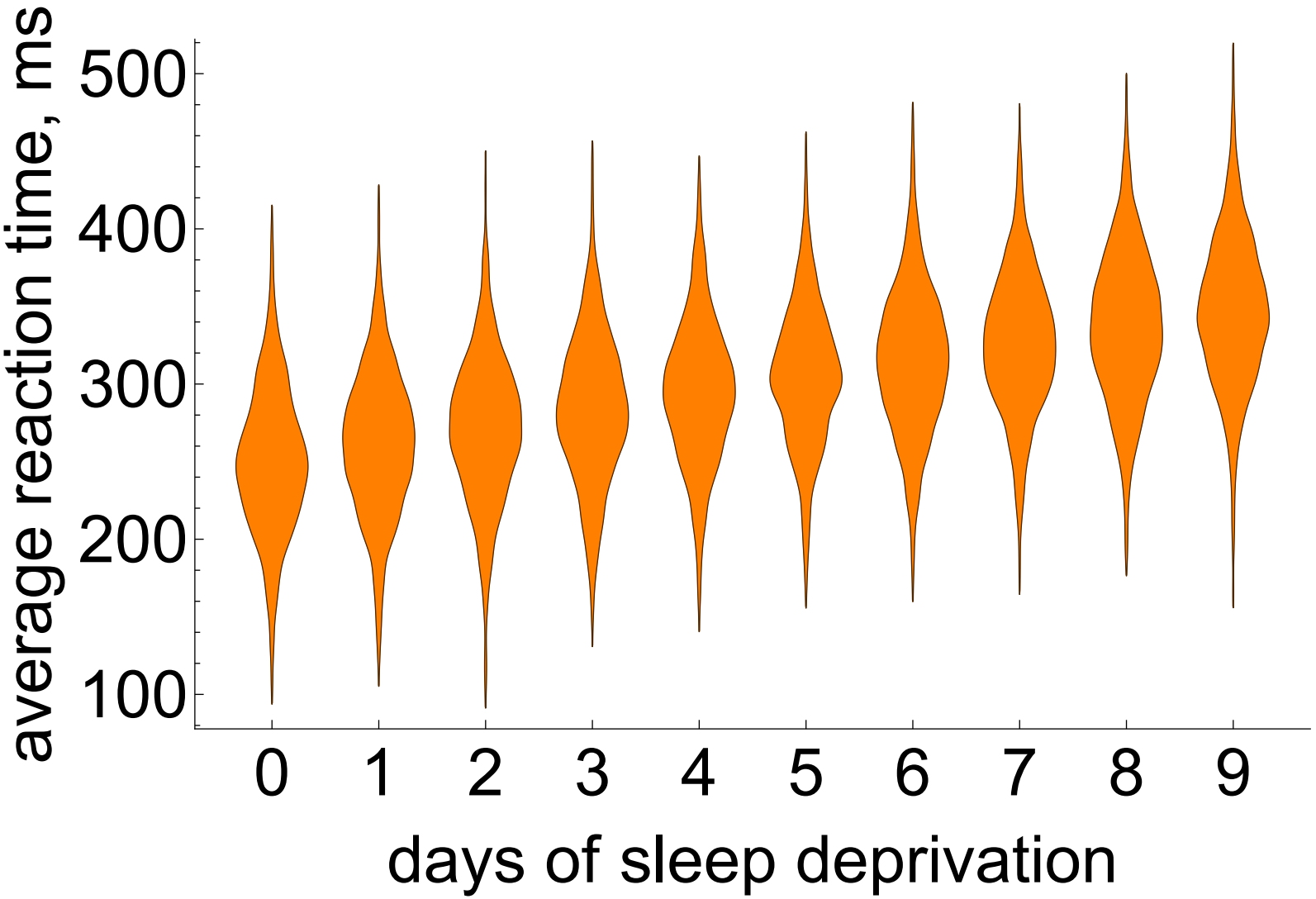
Sleep deprivation study: model changes

```
parameters {
    ...
    real<lower=0> nu;
}
model {
    // likelihood
    R ~ student_t(nu, X * gamma, sigma);
    ...
    nu ~ gamma(1,1);
}
```

$\Longrightarrow$ three changes/additions necessary.

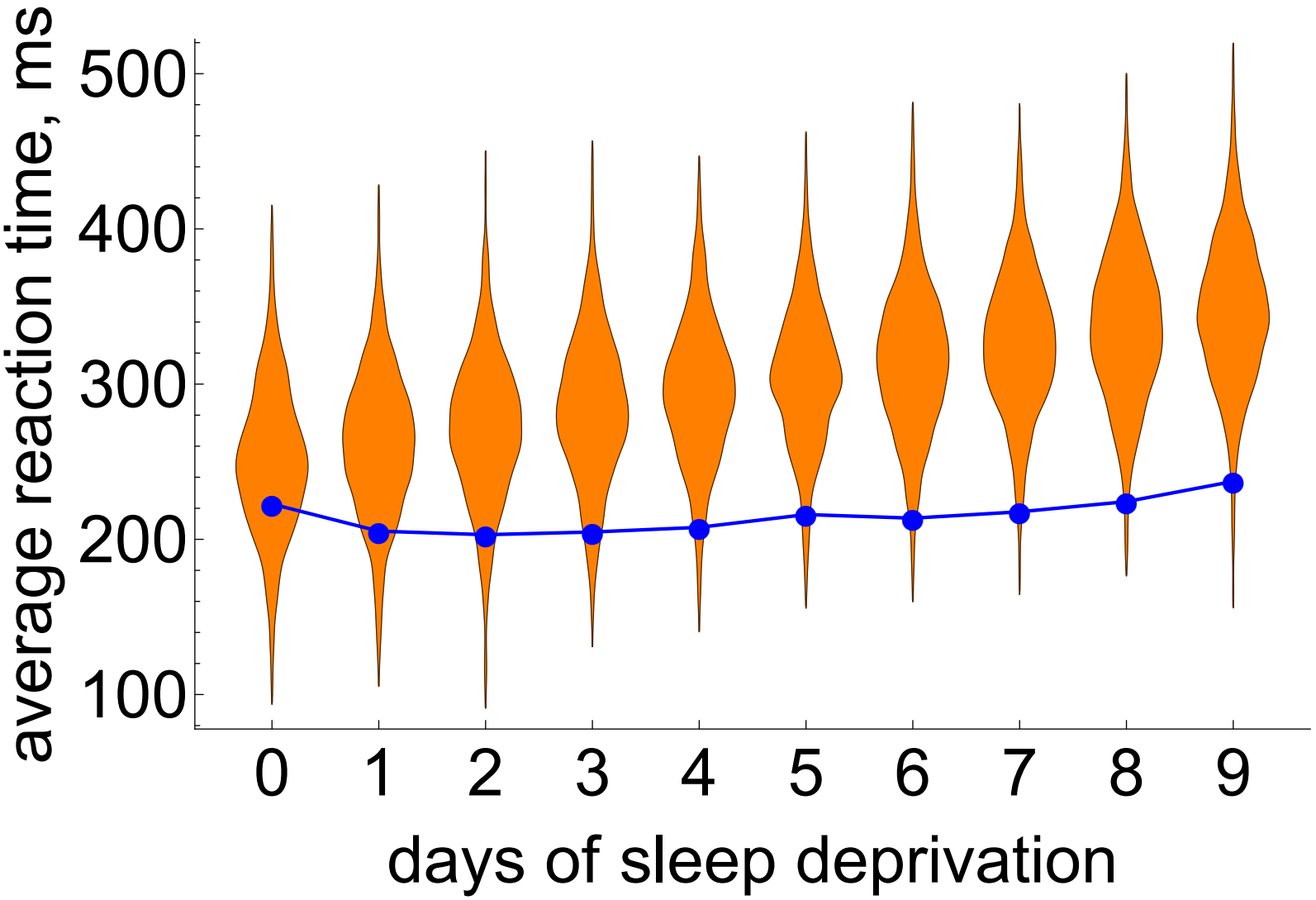
Sleep deprivation: posterior predictive checks

Posterior predictive distribution for Student T sampling.



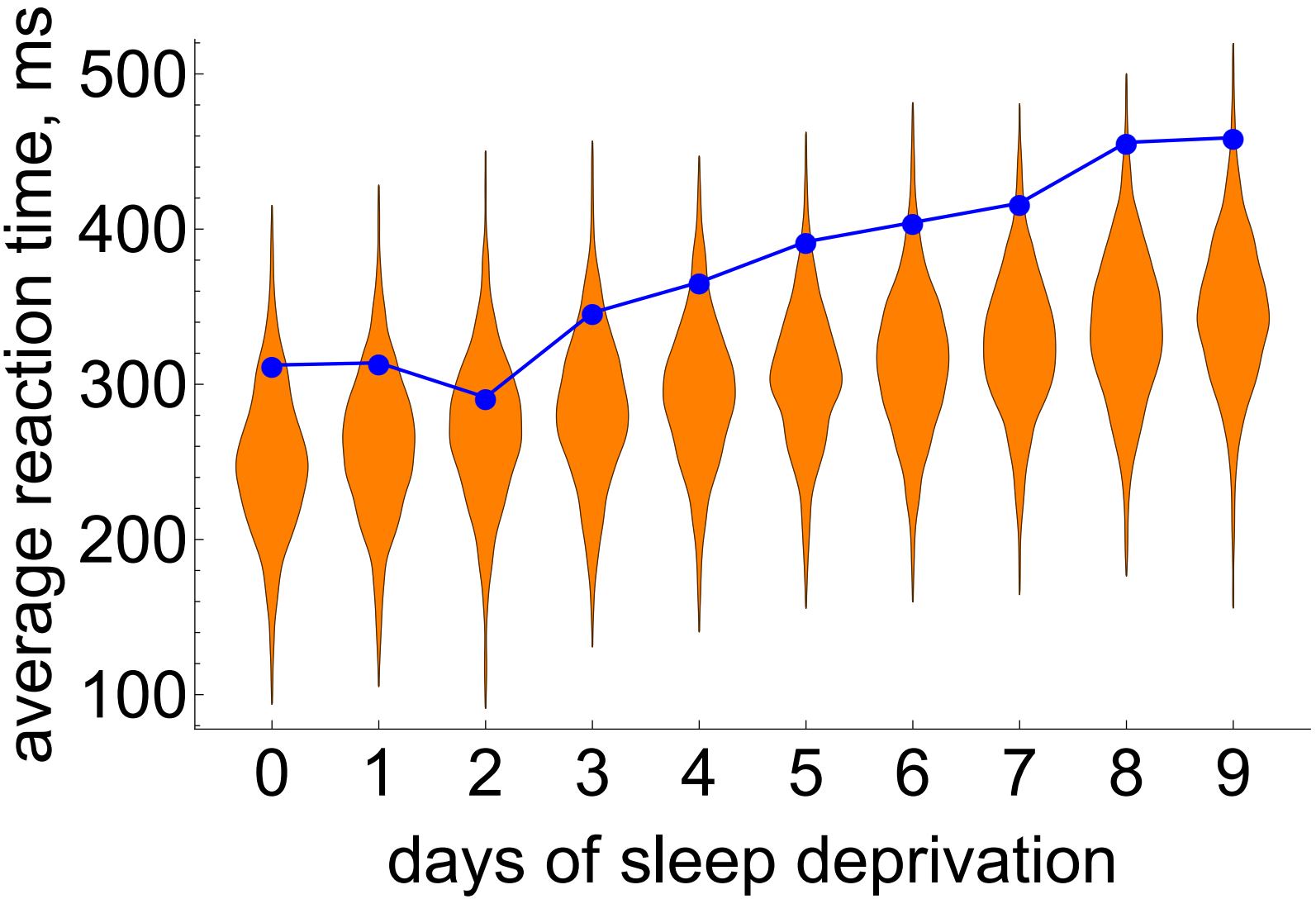
Sleep deprivation: posterior predictive checks

One participant.



Sleep deprivation: posterior predictive checks

And another.



Sleep deprivation study: model changes

Moving to a Student T distribution did not solve individual fit issues $\Longrightarrow$ try a more ambitious change:

- Go back to normal sampling model.
- Allow each subject their own response parameters.
- $\Longrightarrow$ parameters are now 18-dimensional; one per each of the 18 subjects.
- (A better way to do this is with hierarchical models.)

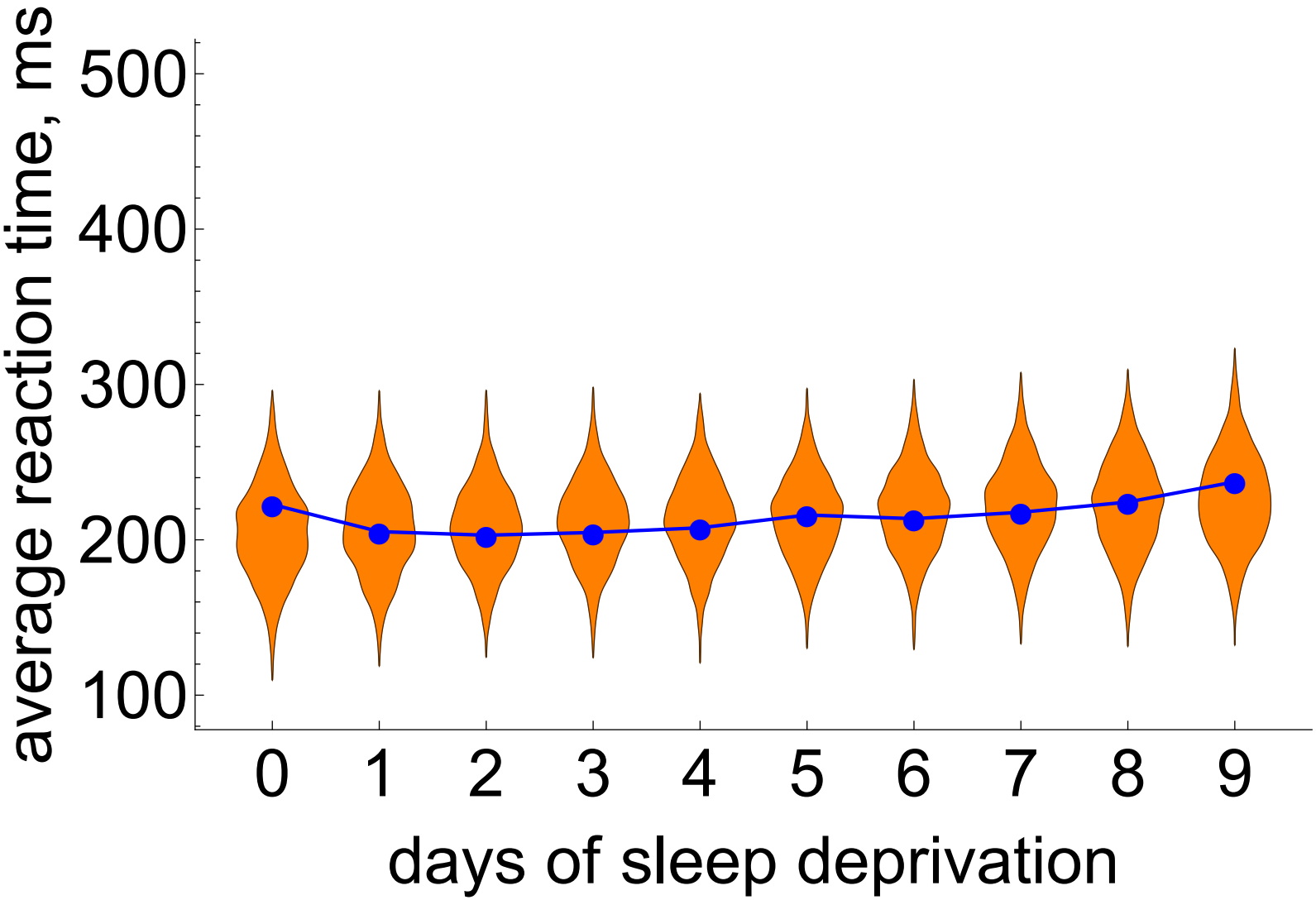
Question: how long does this modification take us?

Sleep deprivation study: model changes

```
data {
    ...
    // array of subject ids: 1, 2,..,18
    int subject[N];
}
parameters {
    ...
    real alpha[18]; // 18 elements of each param
    real beta[18]; // one for each subject
}
model {
    for (i in 1:N){
        R[i] ~ normal(alpha[subject[i]] +
                      beta[subject[i]] * t[i], sigma);
    }
}
```

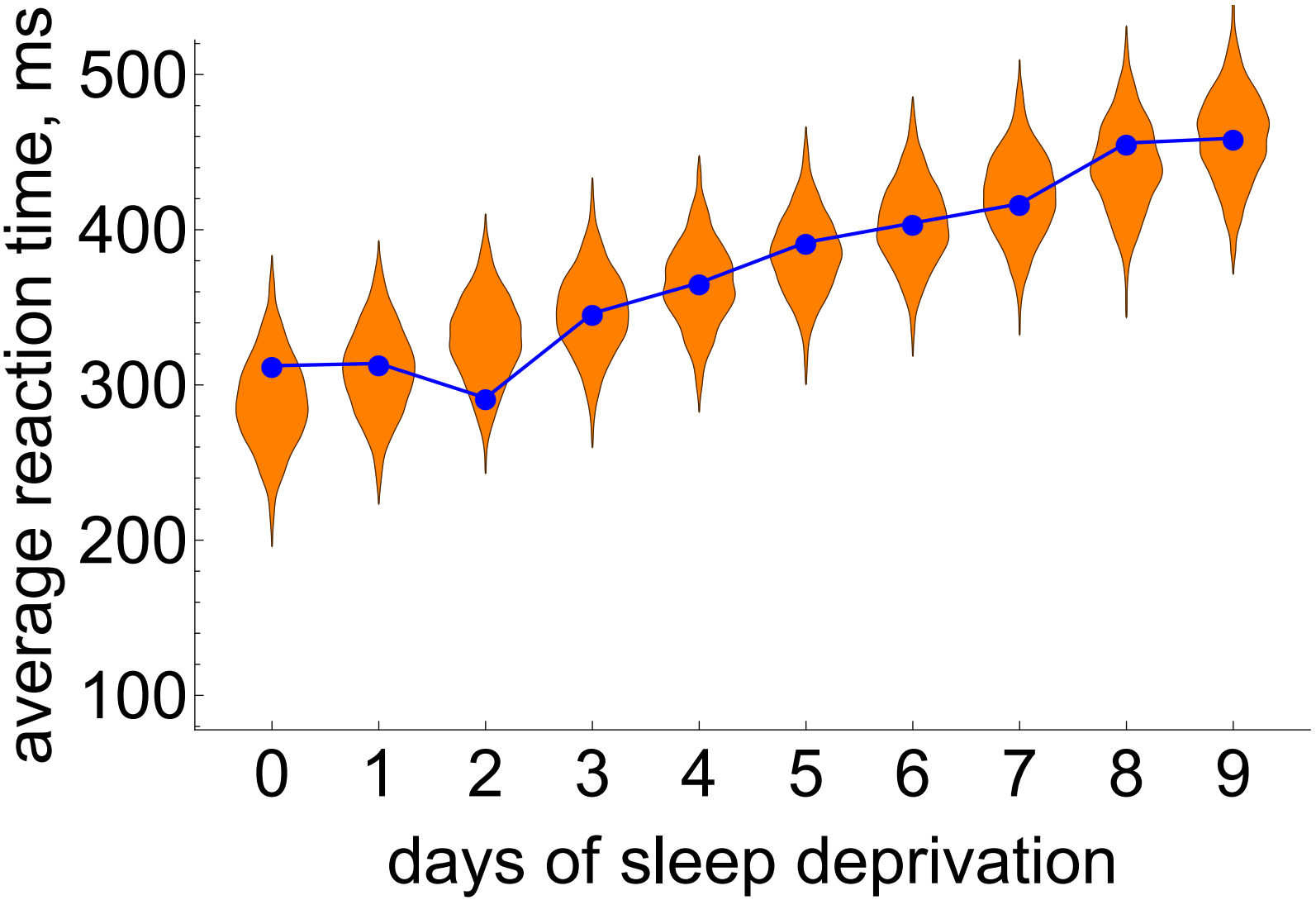
Sleep deprivation: posterior predictive checks

First problematic subject.



Sleep deprivation: posterior predictive checks

Other problematic subject.



Stan: a few of the loops and conditions

Stan has pretty much the full range of language constructs to allow pretty much any model to be coded.

```
for (i in 1:10) {something;}

while (i > 1) {something;}

if(i > 1) {something 1;}
else if (i==0) {something2;}
else {something 3;}
```

Note: this is not the case with JAGS/BUGS due to their **declarative** nature.

Stan: speed concerns

Whilst Stan is fast it pays to know the importance of each code block for efficiency.

- **data** - called once at beginning of execution.
- **parameters** - every log probability evaluation!
- **model** - every log probability evaluation!
- **generated quantities** - once per sample.

Remaining code blocks

Thus far focussed on the four main code blocks, but others exist:

- **transformed data** - carry out a data transformation in Stan; executed once after data.
- **transformed parameters** - often easier to work with transformations of original parameters; executed every log probability evaluation!
- **functions** - allows user-defined functions, must go at top of Stan program. How many times it is called depends on the function's nature. Make it possible to use any density whose log pdf can be written down (see problem set)!

Stan in parallel

In R can run chains in parallel easily using:

```
# smaller font size for chunks
library(rstan)
options(mc.cores=8)
```

Stan summary

- Stan works by default with a HMC-like algorithm called NUTS.
- The Stan language is similar in nature to other common languages with loops, conditional statements and user-definable functions (didn't cover here).
- Stan makes life easier for us than coding up the MCMC algorithms ourselves.
- Stan is typically multiple-times faster than BUGS/JAGS for generating effective samples.

How to debug a Stan model?

Two issue flavours, each with their own response.

- Coding errors.
- Sampling issues.

Coding errors

Stan error messages are generally quite informative however inevitably there are times when it is less clear why code fails

$\implies$ debug by print!

```
model {
    ...
    print(theta);
}
```

In R this prints (neatly) to the console output.

Coding errors

Important: failing a resolution via the above go to `http://mc-stan.org/` and do:

1. Look through manual for a solution.
2. Look through user forum for previous answers to similar problems.
3. Ask a question; be clear, and thorough - post as simple a model that replicates the issue.
4. Ask me!

Sampling issues

Different sort of issue to a coding error, and falls into two (often related) issues:

- **Slow convergence:** still have $\hat{R} > 1.1$ after many thousands of iterations.
- **Divergent iterations:** get a warning in output from Stan with the number of iterations where the NUTS sampler has terminated prematurely.

Sampling issues

Most important thing today: Gelman's "Folk Theorem":

"Issues with computational sampling are almost always due to problems with the underlying statistical model, **not** the algorithm."

Sampling issues: slow convergence

- Poor chain mixing is usually due to lack of **parameter identification**.
- A parameter is identified if it has some unique effect on the data generating process that can be separated from the effect of the other parameters.

Solution (very important): use **fake data** where you know the true parameter values.

$\Longrightarrow$ informative as to whether the data + priors are sufficient to estimate a parameter's value.
If possible use the most simple version of your model that replicates the error.

Sampling issues: slow convergence

A significant number of divergent transitions of NUTS are problematic:

- Indicates that the stepwise integrator used to approximate Hamiltonian dynamics has likely diverged from exact trajectory.
- Therefore these samples **cannot** be viewed as being from the posterior.
- Causes a bias away from problem area of parameter space.
- Almost always because the step size is too large relative to the curvature of the posterior.
- However can be due to placing limits on parameters that preclude an area of high probability mass.

Diagnosing problem: use Shiny Stan (or otherwise) to make pairwise plots of variables $\Longrightarrow$ look for parameters with high bivariate correlation; indicates high curvature.

Sampling issues: slow convergence

Solution: if significant number of divergent iterations do the following:

1. Lower step size and increase acceptance rate in the call to Stan from R or otherwise.
2. If above doesn't help change priors then likelihood.

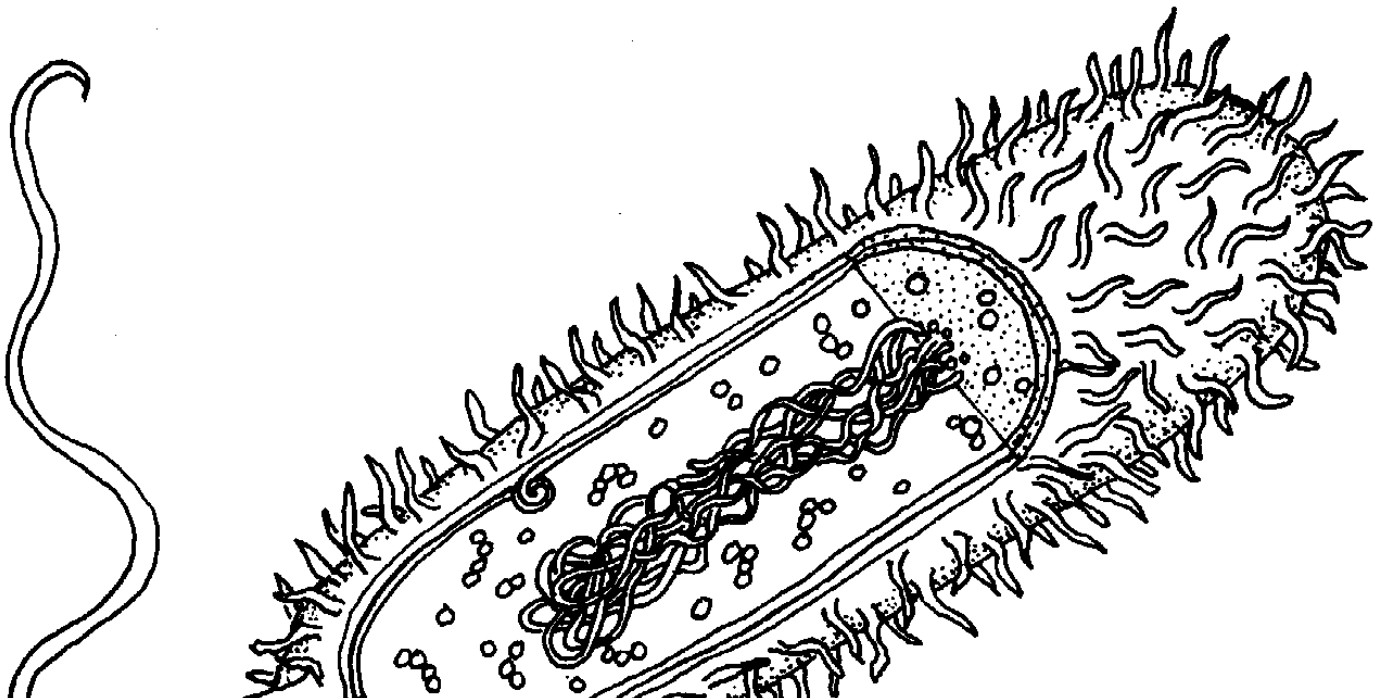
Again failing all the above look at the Stan user forums, then ask a question.

What to do when things go wrong: summary

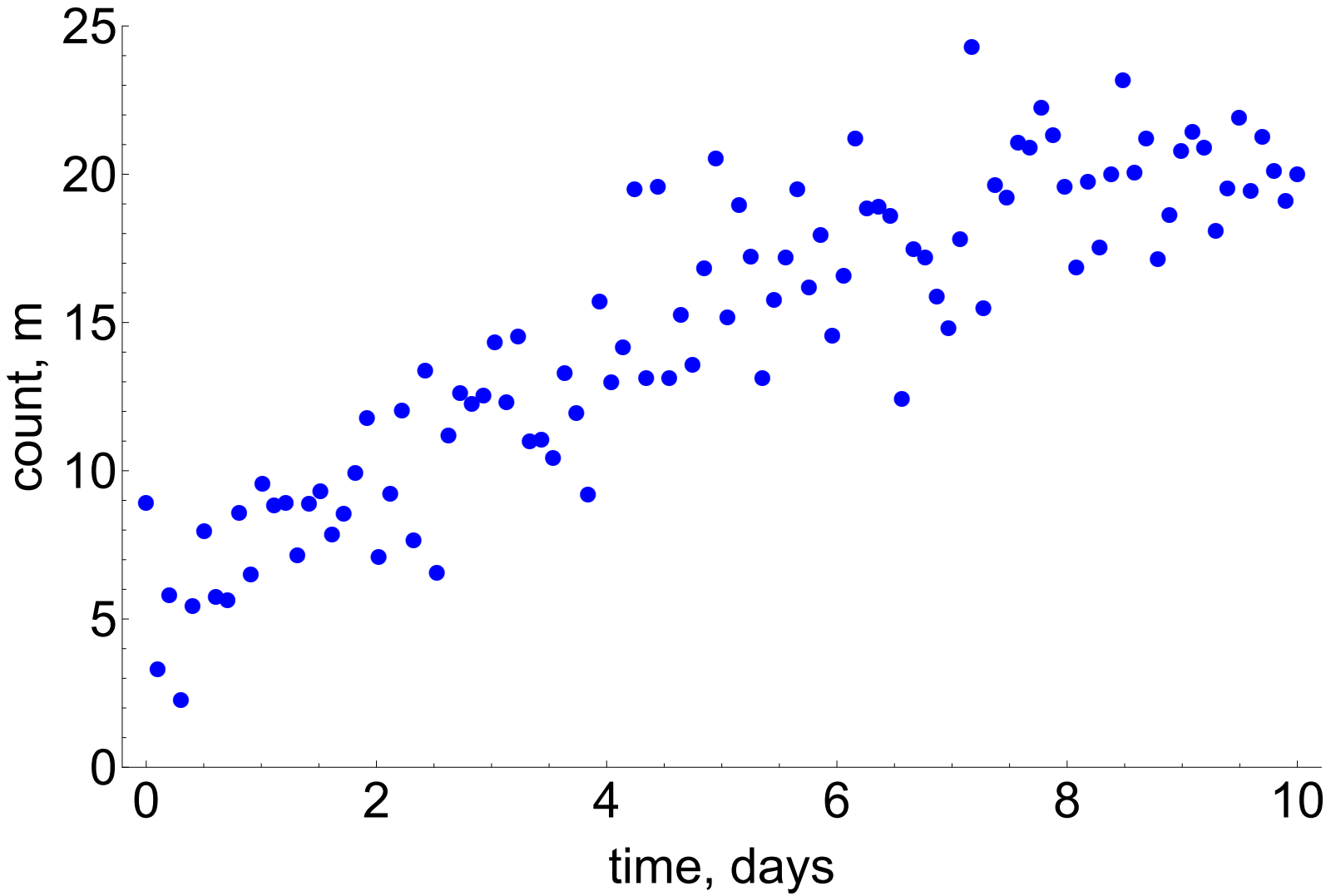
- To debug a model that fails read error messages carefully, then try "print" statements.
- Problems with sampling are almost invariably problems with the underlying model **not** the sampling algorithm per se.
- Use fake data with all models to test for parameter identification (and that you've coded up correctly.)
- Stan has an active developer and user forum, great documentation, and an extensive answer bank.
- If you ask a question on the forum include your model, or ideally a simplified version that replicates the issue.

Example: bacterial growth

- We carry out experiments where we inoculate agar plates with bacteria at time 0.
- At pre-defined time intervals we count the number of bacteria on each plate, $N(t)$.
- Suppose we want to model bacterial population growth over time.



Example: bacteria growth data



Example: bacterial growth model

- Assume the following model for bacterial population growth:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \alpha N(1 - \beta N) \tag{4}$$

where $\alpha > 0$ is the rate of growth due to bacterial cell division, and $\beta > 0$ measures the reduction in growth rate due to "crowding".

Question: how should we infer the parameters of this model?

Example: bacterial growth model

Answer: assume measurement error around true value:

$$N^*(t) \sim \text{normal}(N(t), \sigma) \tag{5}$$

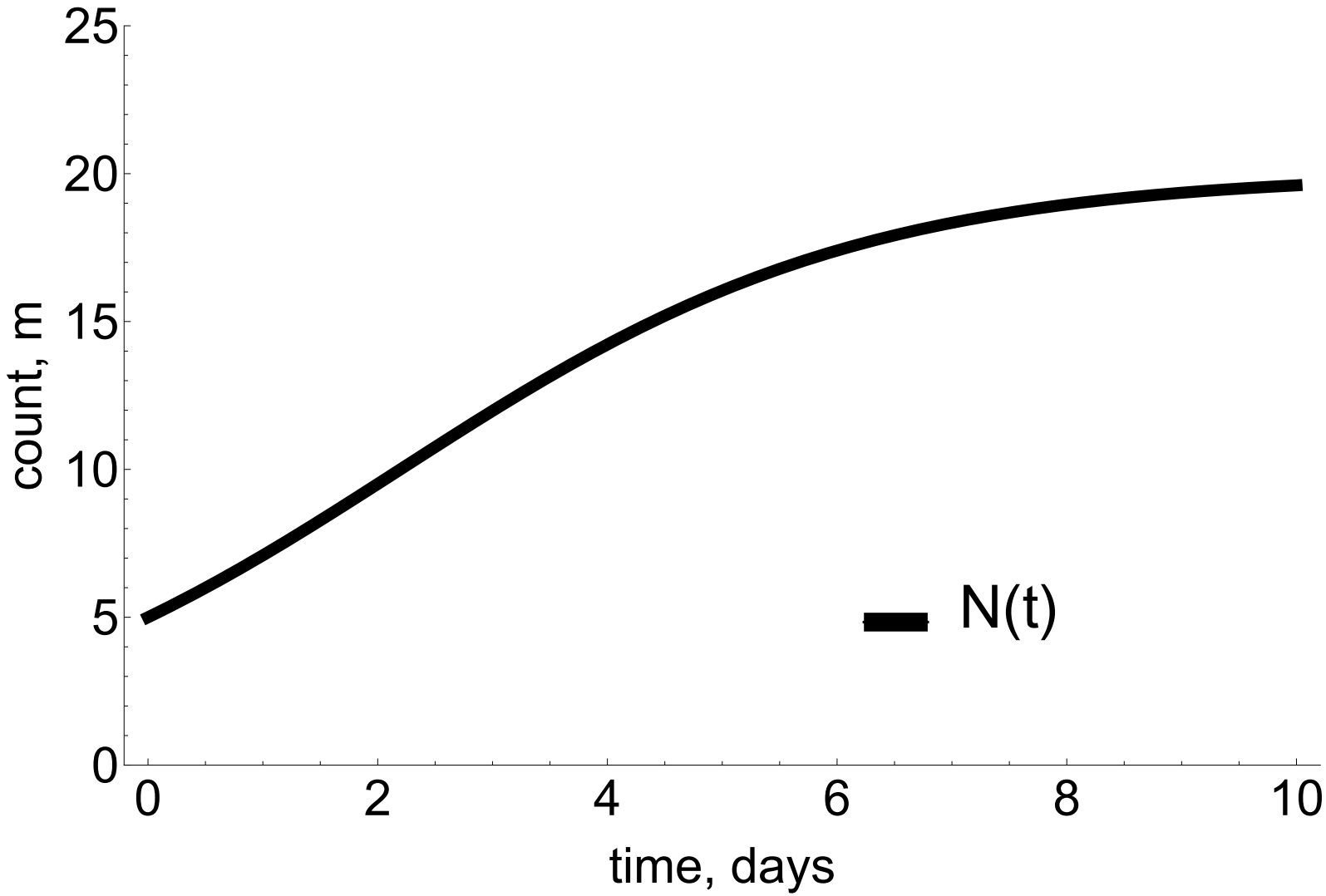
where

- $N^*(t)$ is the **measured** count of bacteria at time t.
- $N(t)$ is the solution to the ODE at time t (true number of bacteria on plate).
- $\sigma > 0$ measures the magnitude of the measurement error about the true value.

Question: how does this model work?

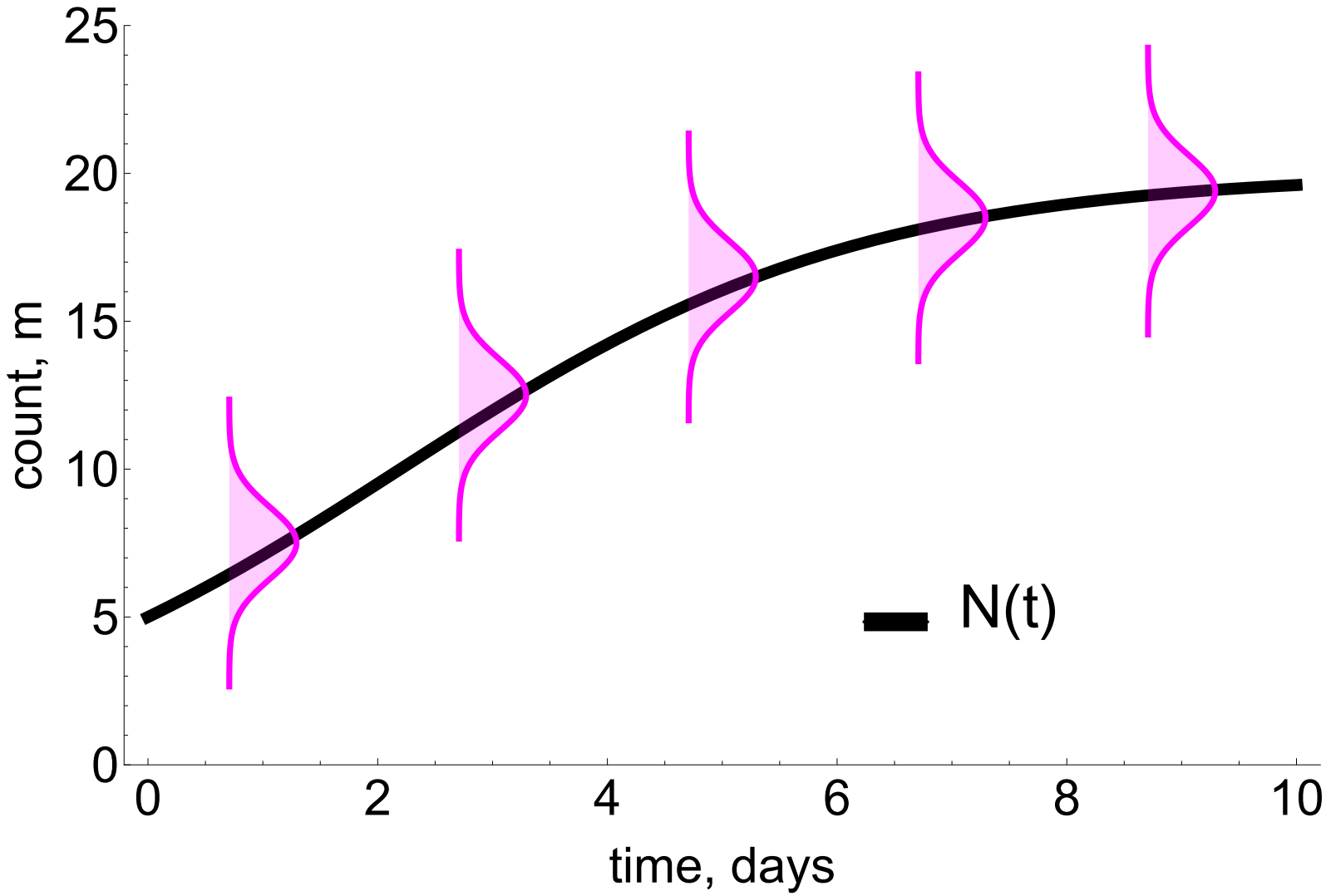
Example: bacterial growth model

Start with true number of bacterial cells, $N(t)$.



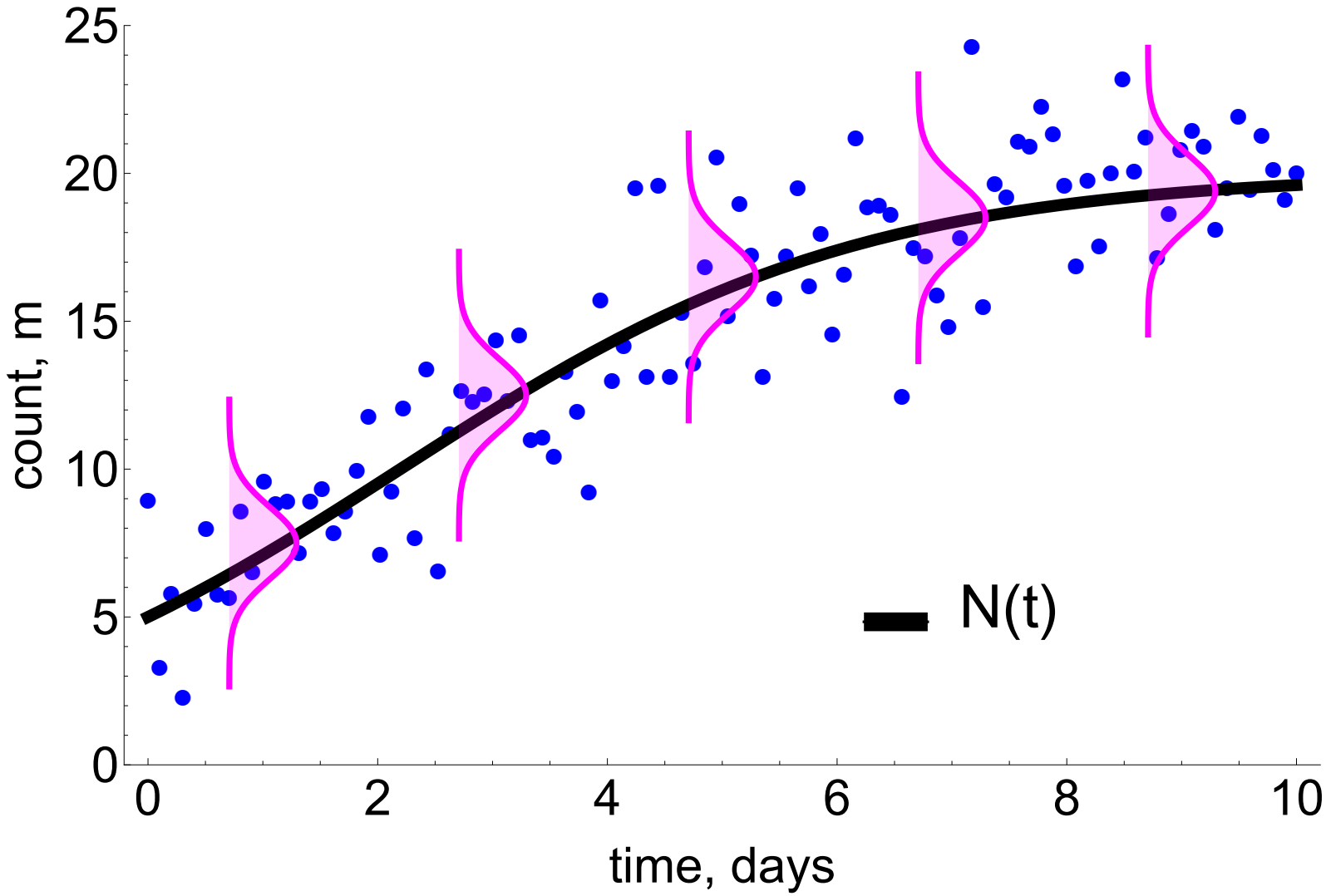
Example: bacterial growth model

Overlay sampling distribution representing measurement error.



Example: bacterial growth model

And data generated from this process.



Example: bacteria growth model inference

Remember we are using a normal likelihood:

$$N^*(t) \sim \text{normal}(N(t), \sigma) \tag{6}$$

$\implies$ likelihood for all observations:

$$L(N(t), \sigma) = \prod_{t=t_1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[\frac{-(N^*(t) - N(t))^2}{2\sigma^2}\right] \tag{7}$$

Question: how do we calculate $N(t)$?

Example: bacteria growth model inference

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \alpha N(1 - \beta N) \tag{8}$$

- In most ODE models, the mean $N(t)$ cannot be solved for exactly so we **can't write down a "closed-form" expression for the likelihood.**
- $\implies$ approximate answer using a numerical method.
- However any solution for $N(t)$ - exact or numerical - depends on the parameters of the ODE model. For our example:

$$N(t) = f(t, \alpha, \beta) \tag{9}$$

Question: how do we do MCMC in this setting?

Example: bacteria growth model inference

For example, in Random Walk Metropolis:

- Start at random location in (α, β, σ) space.
- For t=1,...,T do:
  1. Propose a new location $(\alpha', \beta', \sigma')$ using a jumping distribution.
  2. Numerically (or analytically) integrate ODE to solve for $N(t, \alpha', \beta')$.
  3. Calculate un-normalised posterior at proposed location $\Longrightarrow$ calculate r.
  4. Based on r move to new location or stay at original.

$\Longrightarrow$ at every step we must solve ODE for $N(t)$; can be computationally expensive!

Example: bacteria growth model in Stan

- ODE models can be slow to converge - particularly when accounting for numerical solution of equations.
- Fortunately Stan has an inbuilt ODE integrator $\Longrightarrow$ can leverage the speed of HMC (NUTS) for ODE systems!
- Write a function that returns the derivative (RHS of ODE):

```
real[] bacteria_diff(real t,real[] N,
          real[] theta,real[] x_r,int[] x_i){
    real dNdt[1];
    dNdt[1] = theta[1] * N[1] *
          (1 - theta[2] * N[1]);
    return dNdt;
}
```

Example: bacteria growth model in Stan

```
model {
    sigma ~ cauchy(0,1); // Prior
    theta ~ normal(0,2); // Prior for alpha, beta
    N0 ~ normal(5,2); // Prior for initial #

    // Solve ODE at current parameter values
    real N_hat[T,1];
    N_hat = integrate_ode(bacteria_diff, N0, t0, ts,
            theta, x_r, x_i);
    // Likelihood
    for (t in 1:T) {
        N[t] ~ normal(N_hat[t,1],sigma);
    }
}
```

$\Longrightarrow$ `integrate_ode` takes the derivative function as its first input argument.

ODEs: important posterior predictive checks

- Have assumed that the errors are independent and identically distributed Gaussians $\implies$ failure of either of these conditions leads us to estimate incorrect errors. Lambert et al., (2019) (forthcoming) show:
  - Assuming independent errors when, in fact, errors are persistent leads to overconfidence in estimates.
  - Multiplicative errors leads to incorrect uncertainty; often overstating the amount of uncertainty.
  - Other miscalculations when incorrect errors assumed.

$\implies$ check for autocorrelated errors, heteroscedasticity and other regression assumptions when fitting ODEs.

Issues with inference for ODEs and PDEs

Whilst Stan's integrator makes it quite easy to use HMC, there are still problems that arise:

- ODE models are often non-identifiable $\implies$ need to reparameterise model.
- (Linked) ODE models can be slower to converge than simpler models $\implies$ need to run MCMC for longer before $\hat{R} < 1.1$ achieved.
- Gradients can be too expensive to calculate.

$\implies$ important that we "know" our model well before we start to do inference explicitly.

Also, try non-gradient based methods (see PINTS: `https://github.com/pints-team/pints`).

Inference for ODEs: summary

- ODE models are no harder to formulate than "traditional" problems.
- However for ODE models we cannot typically write down a "closed-form" expression for the likelihood.
- $\Longrightarrow$ use integrator to numerically solve for mean for each set of parameters.
- Stan has inbuilt integrator meaning we can do MCMC using Stan's fast HMC implementation.

Why do we need a measure of a model's fit?

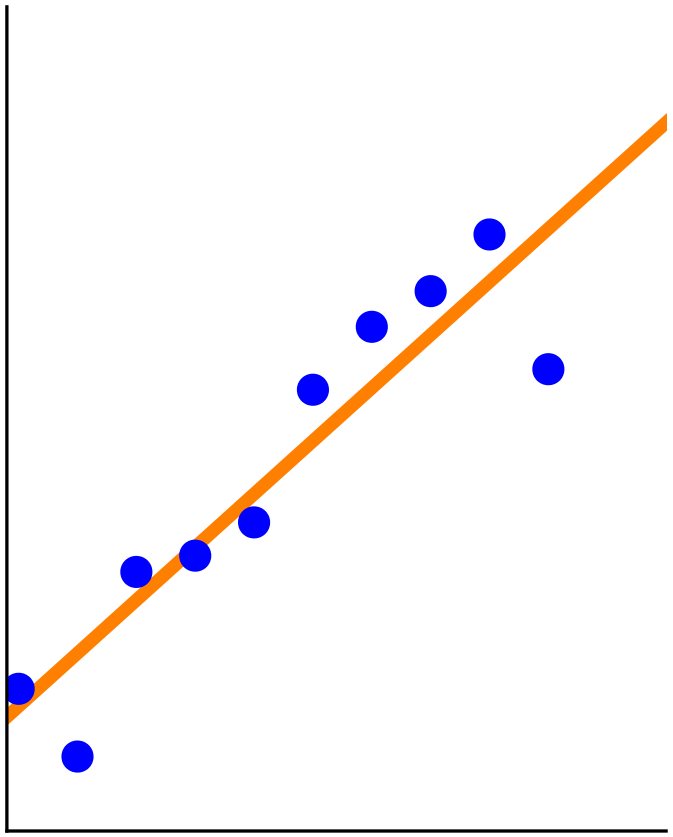
Often in modelling we are required to choose between a number of models, in order to:

- Test between rival hypotheses about the real world.
- Choose the most predictive model to make forecasts about the future.
- Avoid the computational expense of using a number of models for some future use.

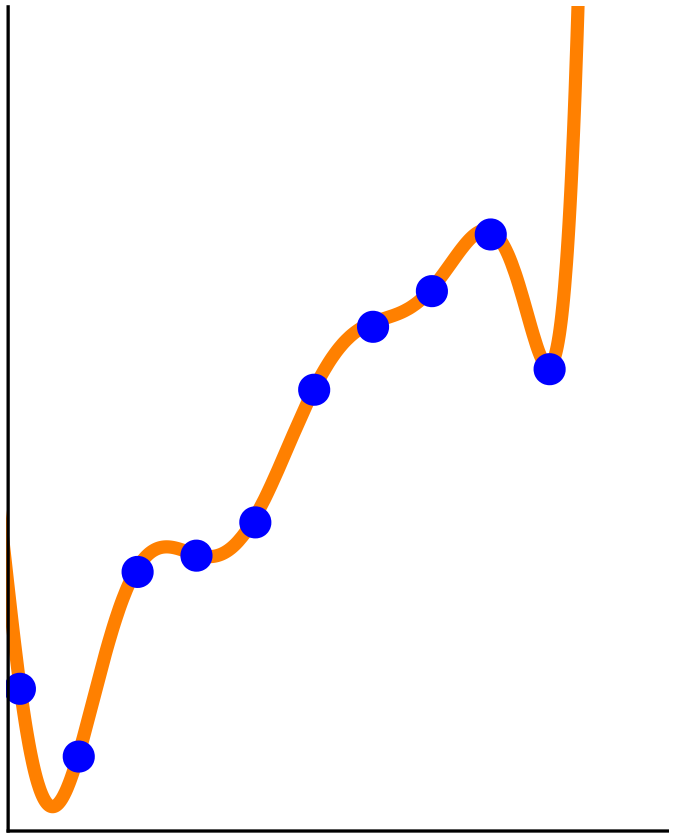
Note: model selection can sometimes be avoided by **a.** using a parameter to specify model choice or **b.** using a general model that allows all models as special cases.

Which of these models is more predictive?

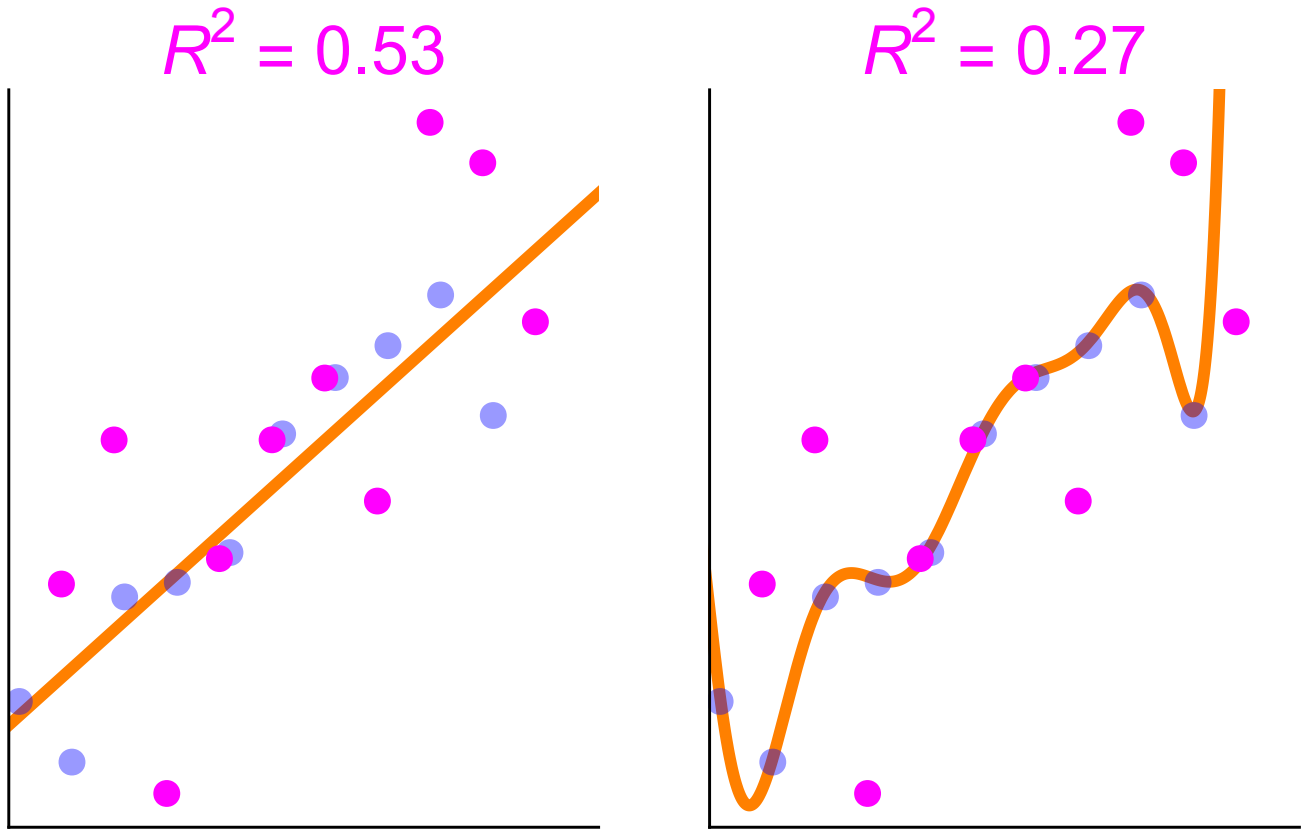
$R^2 = 0.73$



$R^2 = 0.99$



Simple models generalise better to out-of-sample data



$\Longrightarrow$ right model is **overfit** to data; it fits the **noise** not the **signal**.

Selection bias in model selection

The problem:

- Want to measure a model's predictive capability on an **independent** data set; i.e. one that has not been used to fit our model.
- But don't have out-of-sample data! (If we did it would form part of the "sample"!)
- If use in-sample data to gauge model fitness $\Longrightarrow$ predictive performance is upwardly-biased due to overfit.

Note: this selection bias effect $\Longrightarrow$ posterior predictive checks will fail to detect model overfit.

Selection bias in model selection

Solutions:

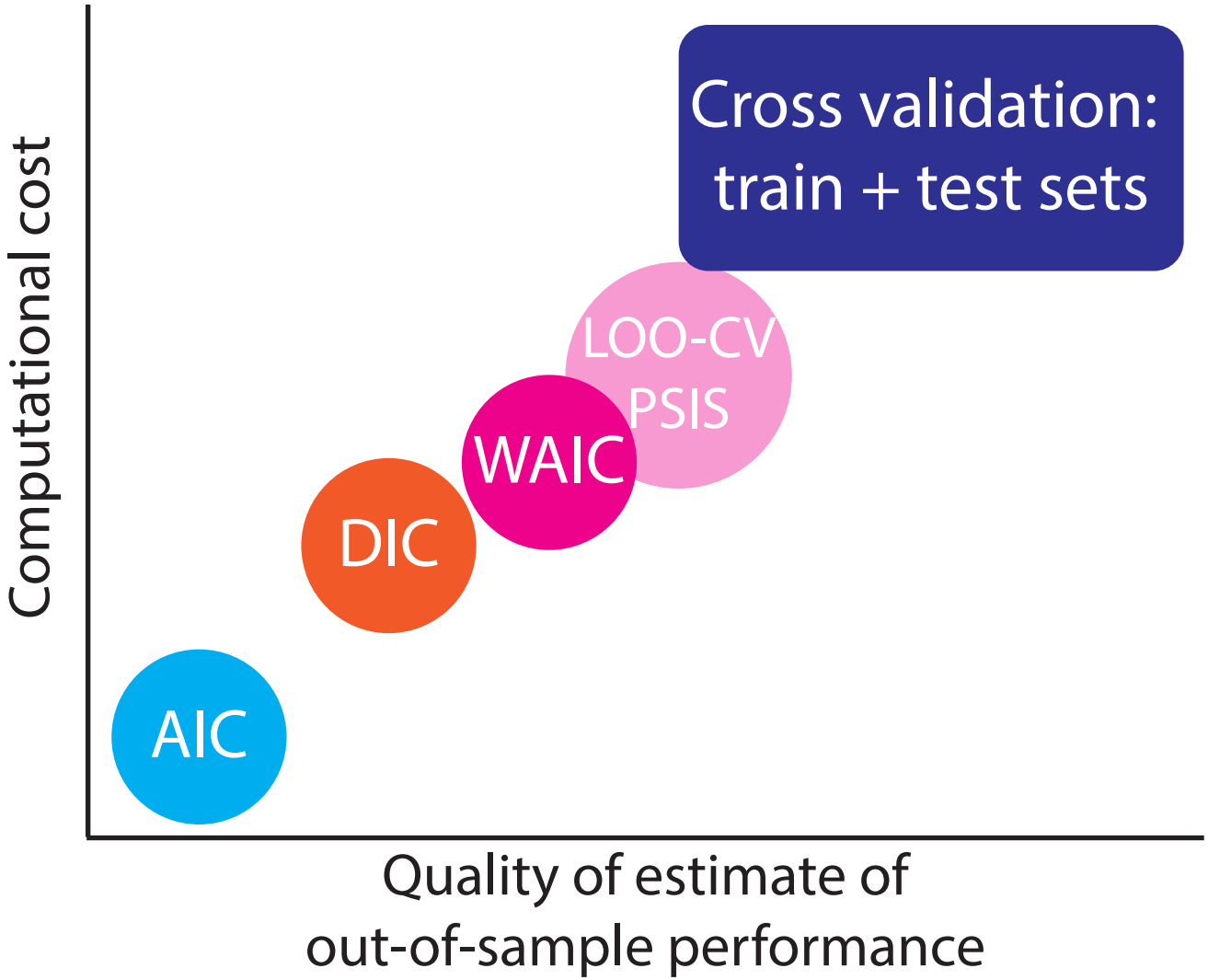
- Use **heuristics** to correct in-sample measures of fit to account for overfit.
- (Better) Use **cross validation** where data is partitioned into:
  - "Training" sets: used to fit model.
  - "Testing" sets: used to evaluate model.

Using heuristics to estimate out-of-sample predictive performance

Correct measures of fit to account for overfitting:

- $R^2 \rightarrow \overline{R^2}$; i.e. a metric that measures proportion of variance explained by model.
- Log-likelihood $\rightarrow$ AIC, DIC, or WAIC; more appropriate metric for Bayesian models due to their basis in likelihood/probability.

Ranking heuristics



Evaluating heuristics: AIC

"Akaike Information Criterion" computed by (ignoring -2 at front):

$$\text{AIC} = log\ p(X|\hat{\theta}_{MLE}) - k \tag{10}$$

where k is the number of parameters estimated in model fitting, and $\hat{\theta}_{MLE}$ is maximum likelihood (point) estimate.

- Based on asymptotic approximation to normal linear models with uniform priors.
- $\Longrightarrow$ not reasonable correction for more general models; particularly hierarchical ones.
- Ignores uncertainty in parameter estimates, and hence log-likelihood.

Evaluating heuristics: DIC

"Deviance Information Criterion" computed by (ignoring -2 at front):

$$\text{DIC} = log\ p(X|\hat{\theta}_{Bayes}) - p_{DIC} \tag{11}$$

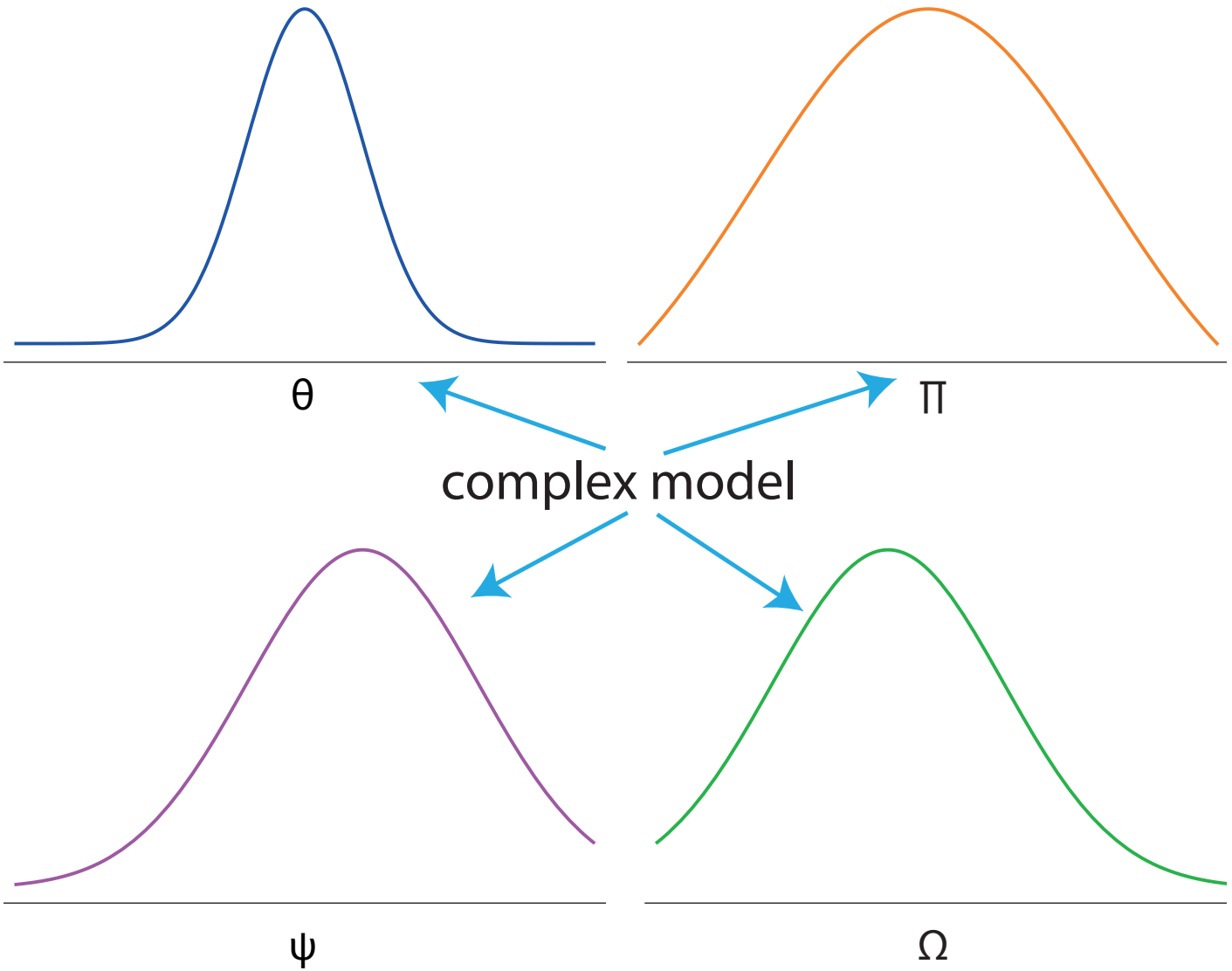
where $\hat{\theta}_{Bayes}$ is posterior mean (point) estimate and p_{DIC} is the "effective number of parameters" defined by:

$$p_{DIC} = 2var_{post}\left[log\ p(X|\theta)\right] \tag{12}$$

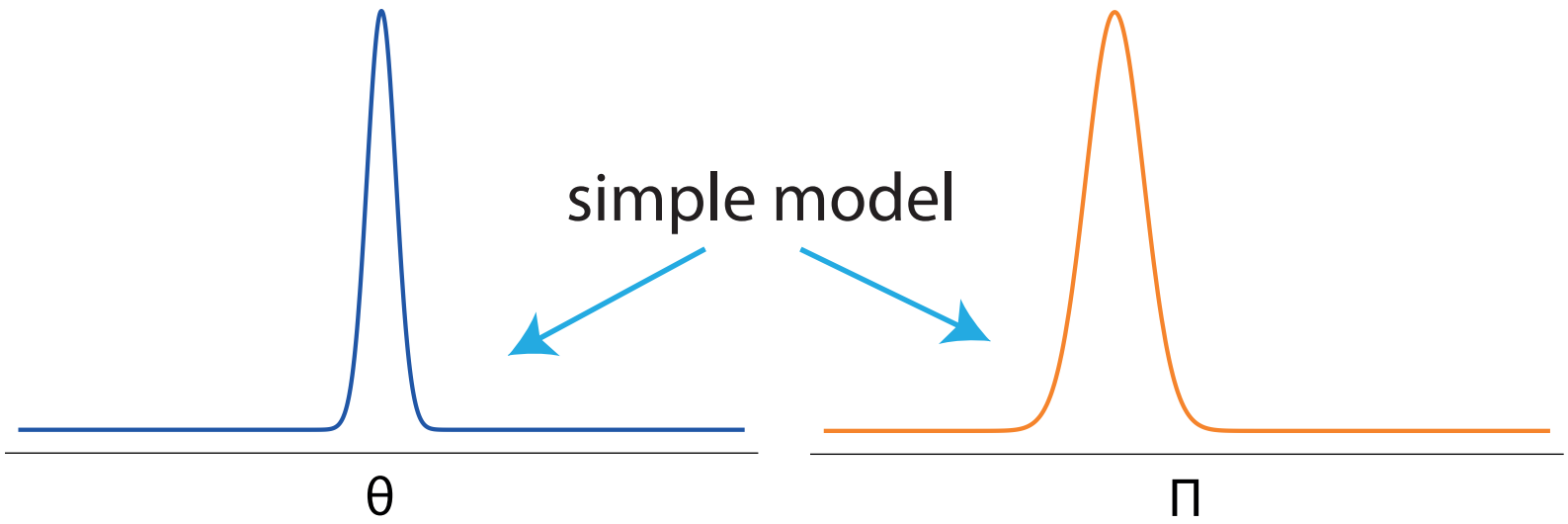
which is typically estimated across all posterior samples.

$$p_{DIC} \approx 2V_{s=1}^{S} log\ p(X|\theta_s) \tag{13}$$

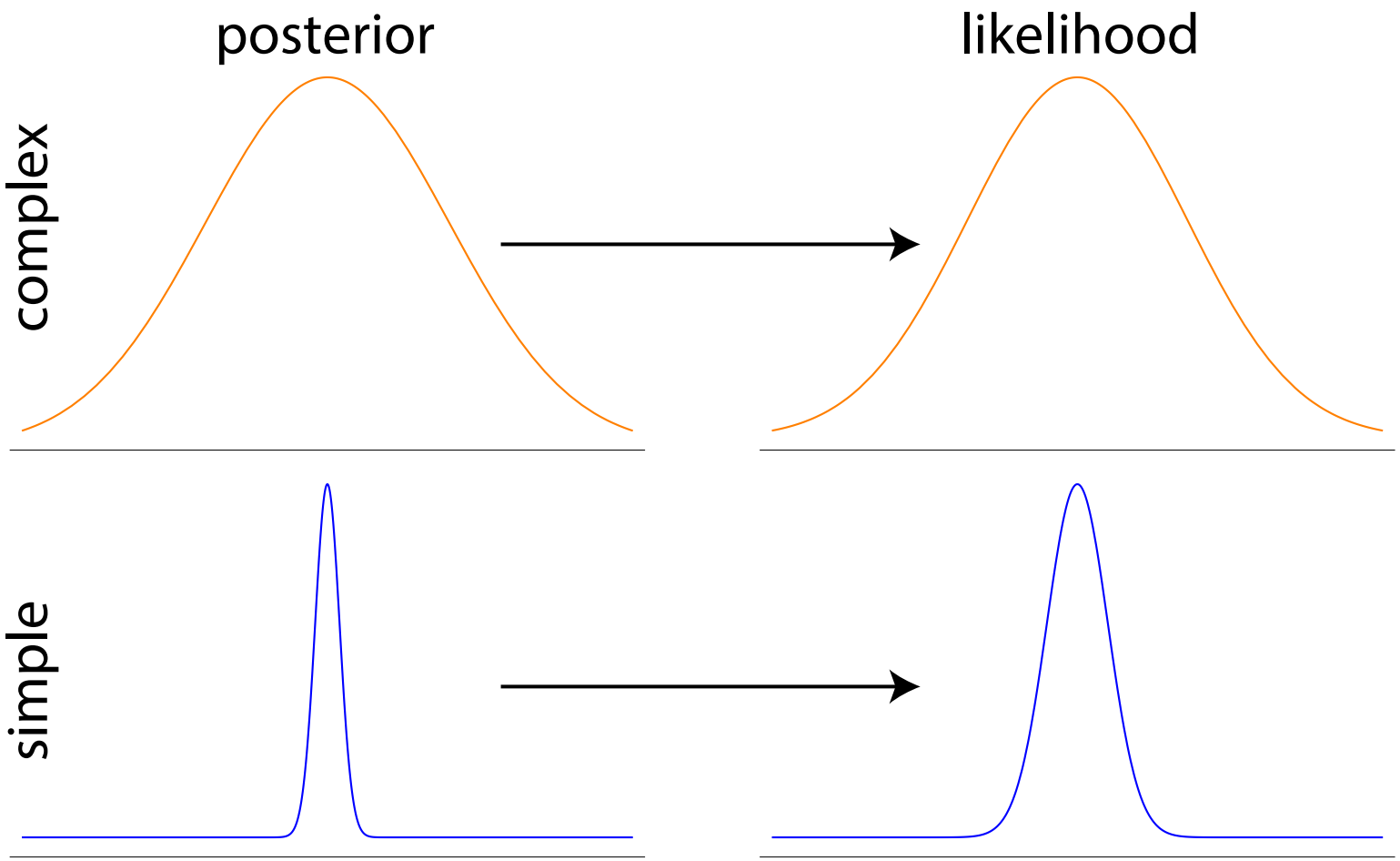
DIC complexity correction intuition: complex models have high parameter uncertainty



DIC complexity correction intuition: simpler models have lower uncertainty



DIC complexity correction intuition: lower posterior uncertainty $\Longrightarrow$ lower uncertainty in fit

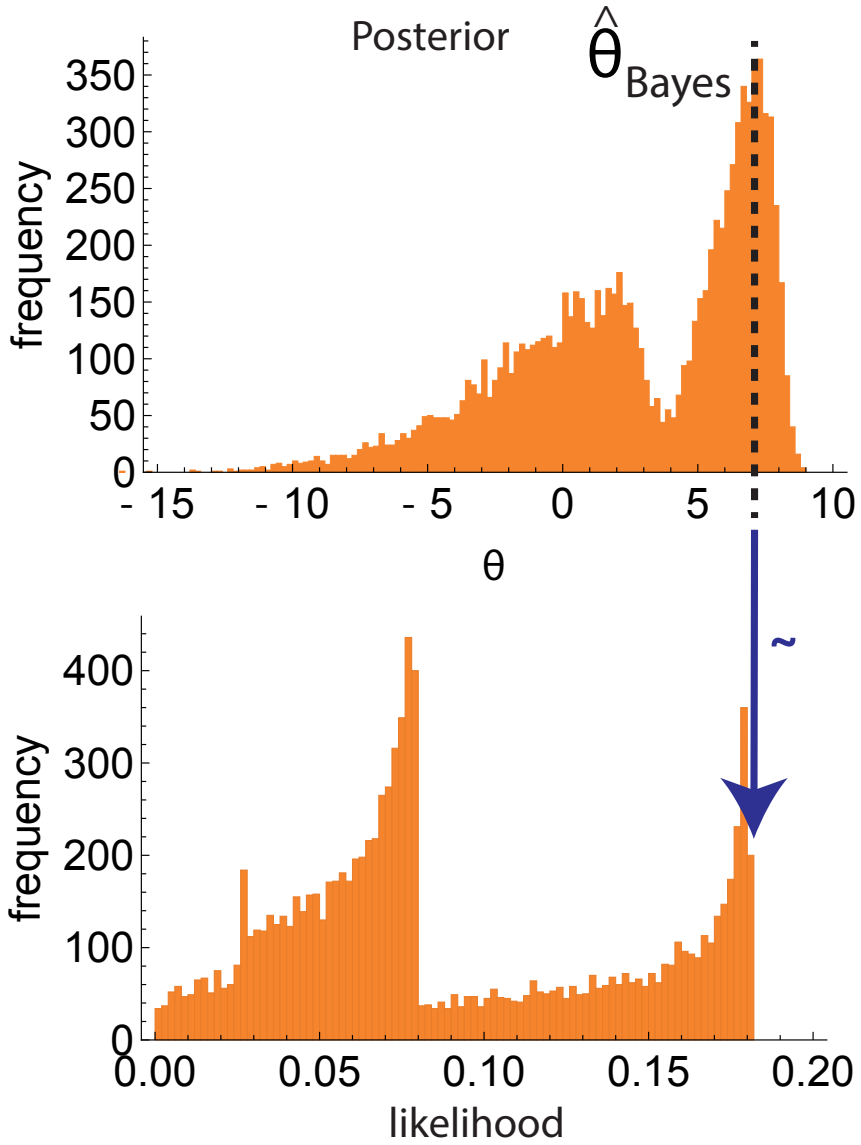


Evaluating heuristics: DIC

$$p_{DIC} = 2var_{post}\left[log\ p(X|\theta)\right] \tag{14}$$

- Spiegalhalter (2002) introduced P_{DIC} as a measure of model dimensionality.
- Overfit models have many parameters, each with high uncertainty.
- A high variance in $\theta \implies$ high variance in $log(X|\theta)$.
- $\implies$ p_{DIC} is big, so large correction.

DIC issue: uses a point estimate to determine fit



Evaluating heuristics: DIC

$$\text{DIC} = log\ p(X|\hat{\theta}_{Bayes}) - p_{DIC} \tag{15}$$

In summary:

- Uses less arbitrary notion of model dimensionality to correct measure of fit than AIC.
- $\Longrightarrow$ better for a wider class of models.
- Still does not fully address uncertainty in fit since calculates first term above with a point estimate $\hat{\theta}_{Bayes}$.

Evaluating heuristics: WAIC

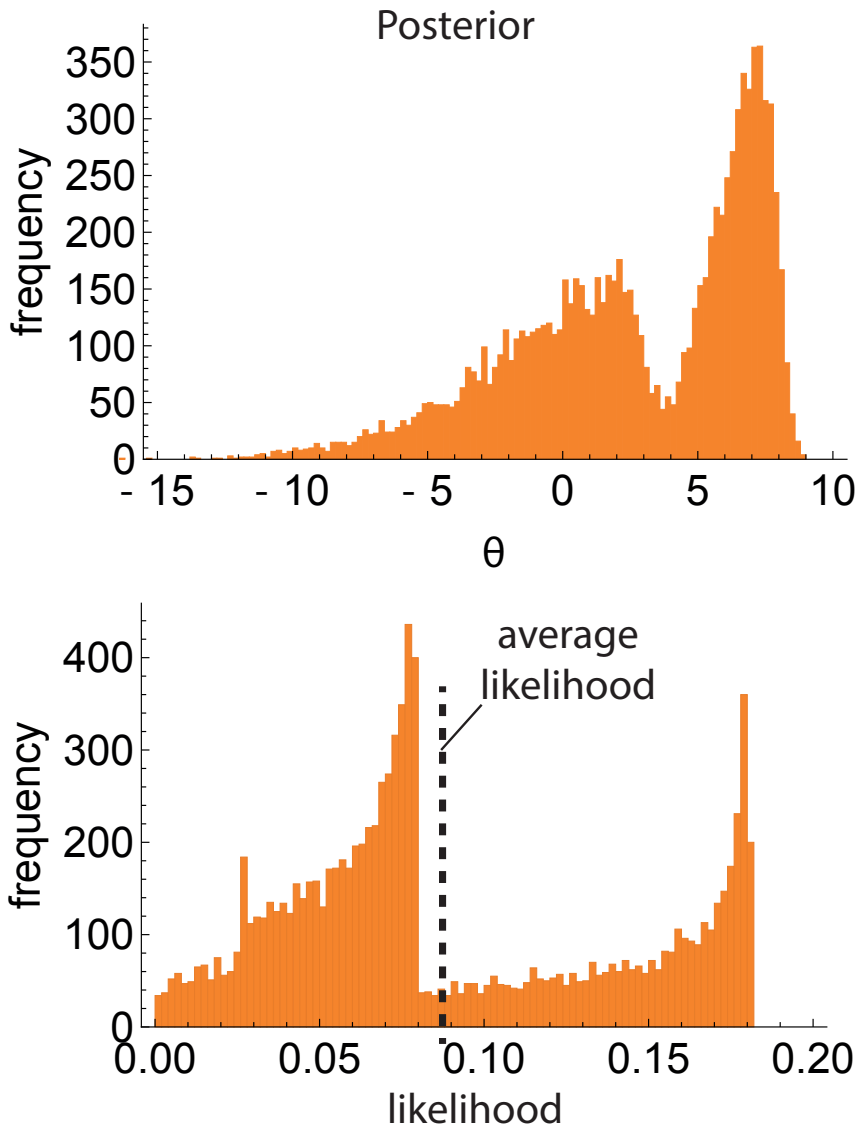
"Watanabe-Akaike Information Criterion" (Watanabe, 2010) computed by (ignoring -2 at front):

$$WAIC = \underbrace{\sum_{i=1}^{N} log\left(\frac{1}{S}\sum_{s=1}^{S} p(X_i|\theta_s)\right)}_{\text{log pointwise predictive density}} - p_{WAIC} \tag{16}$$

Since allows for uncertainty in fit $\implies$ fully-Bayesian estimate. Where p_{WAIC} is another correction factor calculated by:

$$p_{WAIC} = \sum_{i=1}^{N} var_{post}\left[log\ p(X_i|\theta)\right] \tag{17}$$

WAIC: averages likelihood over posterior to determine fit



Evaluating heuristics: WAIC

$$WAIC = \underbrace{\sum_{i=1}^{N} log\left(\frac{1}{S}\sum_{s=1}^{S} p(X_i|\theta_s)\right)}_{\text{log pointwise predictive density}} - \sum_{i=1}^{N} var_{post}\left[log\ p(X_i|\theta)\right] \tag{18}$$

In summary:

- Fully Bayesian since estimates fit across all posterior samples.
- Evaluates the log predictive density using pointwise sums.
- $\implies$ for structured models can be problematic if model not easily split into parts (can always use groups rather than individual points, though.)
- (Still an **approximation** $\implies$ ideally use proper cross validation.)

WAIC using Stan + 'loo' package

Use "generated quantities" block to store log-likelihood across all data points. For example:

```
model {
...
for (i in 1:N){
y[i] ~ normal(mu,sigma);
}
}
generated quantities{
vector logLikelihood[N];
for (i in 1:N){
logLikelihood[i] <- normal_log(y[i],mu,sigma)
}
}
```

WAIC using Stan + 'loo' package

```
library(loo)

## Extract log-likelihood
logLikelihood <- extract_log_lik(fit,'logLikelihood')

## Calculate WAIC
aWAIC <- waic(logLikelihood)

## Print answer
print(aWAIC)

## Computed from 12000 by 5000 log-likelihood matrix
##
##           Estimate   SE
## elpd_waic -10543.1 49.9
## p_waic          4.0  0.1
## waic        21086.3 99.7
```

"loo" also estimates leave-one-out-cross-validation errors using importance sampling

```
## Estimate LOO-CV
aLOO <- loo(logLikelihood)

## Print answer
print(aLOO)

## Computed from 12000 by 5000 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo -10543.1 49.9
## p_loo         4.0  0.1
## looic     21086.3 99.7
##
## All Pareto k estimates OK (k < 0.5)
```

Note: this is a preferred measure to WAIC, but be careful $\Longrightarrow$ if get Pareto k estimates warning do manual cross validation.

Important: use pairwise comparison to do model selection with WAIC or LOO-CV

```
## Extract log-likelihood for both models
logLikelihood_1 <- extract_log_lik(fit_1,'logLikelihood')
logLikelihood_2 <- extract_log_lik(fit_2,'logLikelihood')

## Estimate LOO-CV for eac
aLOO_1 <- loo(logLikelihood_1)
aLOO_2 <- loo(logLikelihood_2)

## Print answer
compare(aLOO_1,aLOO_2)

## elpd_diff        se
##      -0.7       0.7
```

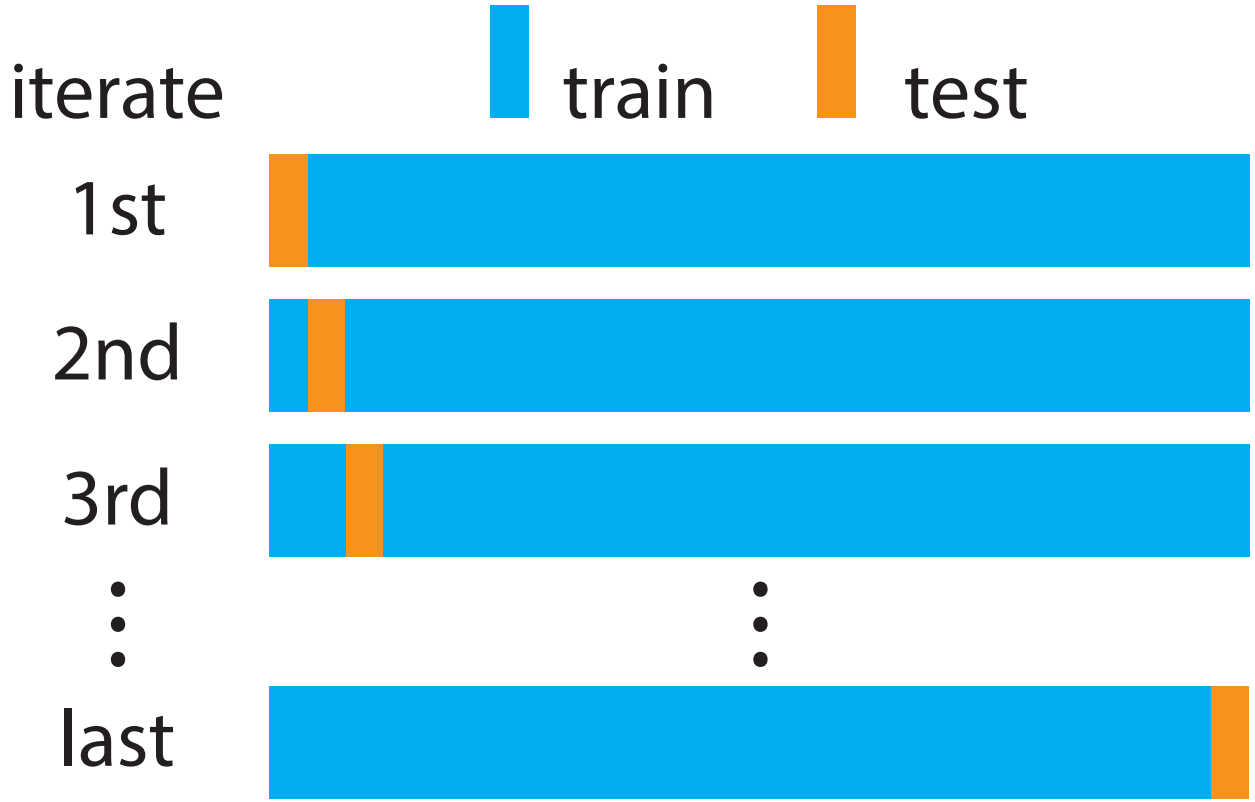
$\Longrightarrow$ pairwise comparison takes a proper account of uncertainty in fit of each model **relative** to the other.

Manual cross-validation

- The aforementioned methods **approximate** out-of-sample predictive capability by use of post-hoc corrections to account for overfitting.
- A better class of approximations is to partition the dataset and do proper **cross-validation.**

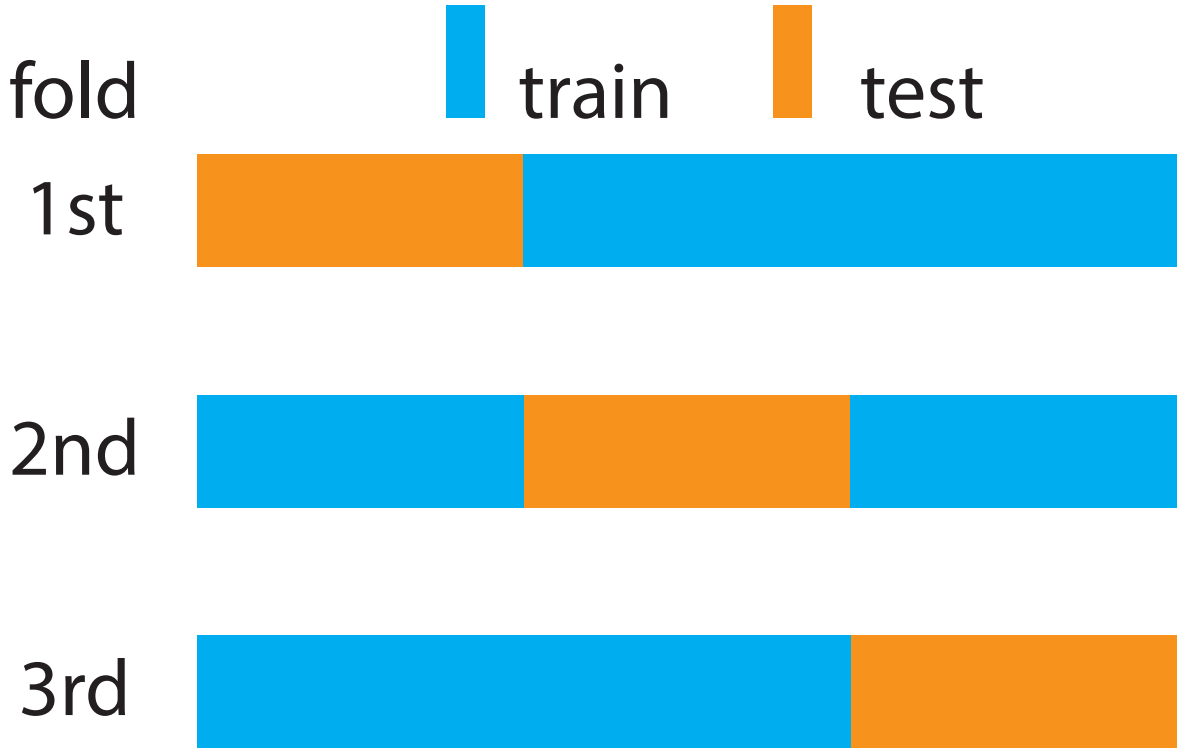
Question: how should we choose our training and test sets?

Manual cross-validation: leave-one-out cross-validation



$\Longrightarrow$ average log-likelihood across all posterior θ across all partitions. **Note:** no need for overfitting correction!

Manual cross-validation: k-fold cross-validation



$\Longrightarrow$ average log-likelihood across all posterior θ across all folds. **Note:** no need for overfitting correction!

Cross validation: issues for consideration

- Cross-validation method should reflect eventual use of model. For example, if predicting individual data points is most important use leave-one-out.
- Manual-refitting the model for each train/set can be costly $\Longrightarrow$ k-folds is less time-consuming than leave-one-out.
- Leave-one-out may be difficult for hierarchical models where data is naturally grouped $\Longrightarrow$ use k-folds where one fold = one group.
- Remember cross-validation is still an approximation to out-of-sample estimation $\Longrightarrow$ best to get new and independent data set!

Bayes factors

An alternative approach is to use Bayes factors to choose between models:

$$\frac{p(\text{model } 1|X)}{p(\text{model } 2|X)} = \frac{p(X|\text{model } 1)}{p(X|\text{model } 2)} \times \frac{p(\text{model } 1)}{p(\text{model } 2)} \tag{19}$$

where blue expression is known as the **Bayes factor.**

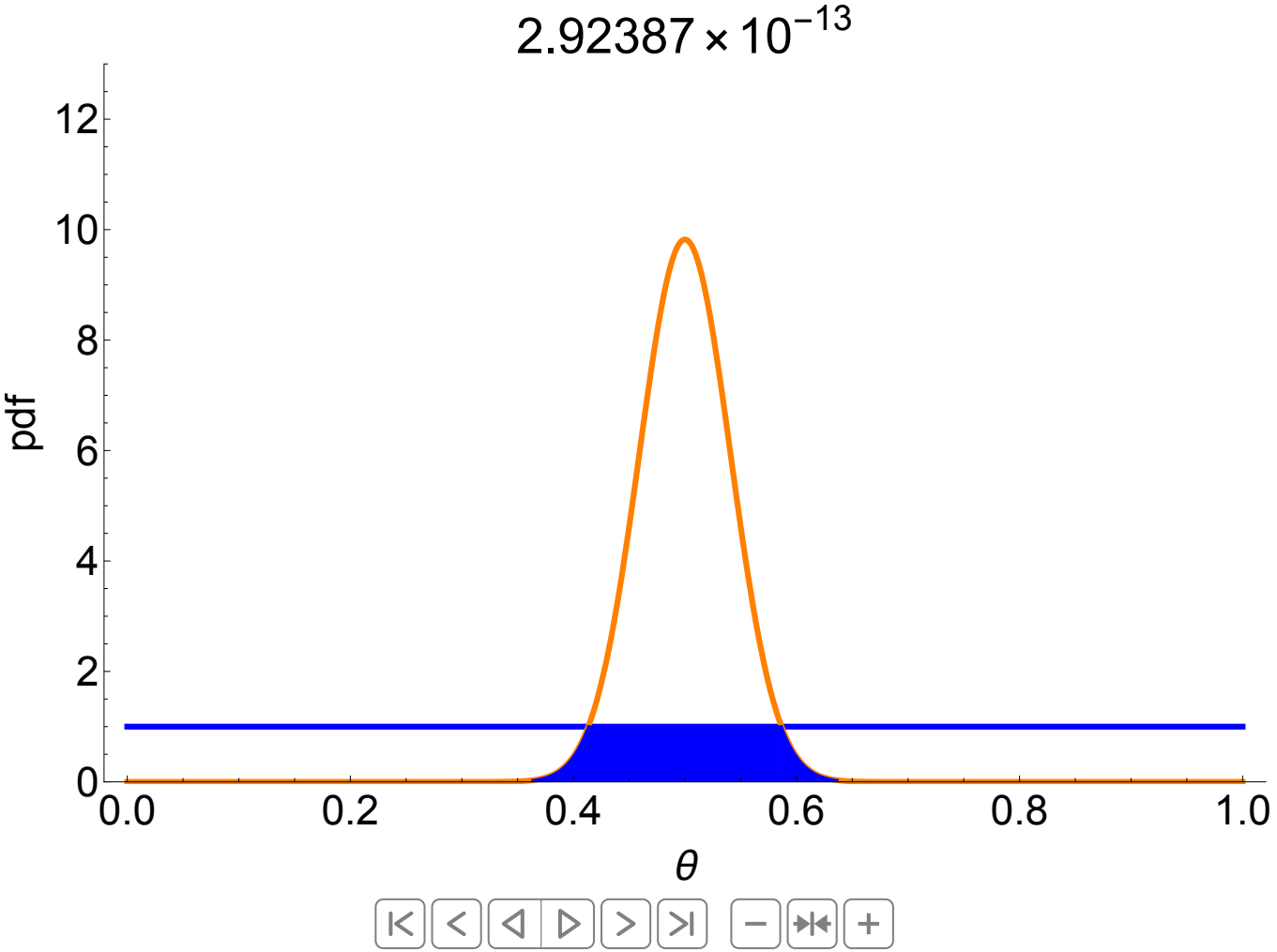
$\Longrightarrow$ requires:

- *a priori* specification of our preferences over models (not simple for models with different dimensions.)
- Calculate $p(X|\text{model})$ – known as the **marginal likelihood** for a model. Calculate by,

$$p(X|\text{model}) = \int \underbrace{p(X|\theta, \text{model})}_{\text{likelihood}} \times \underbrace{p(\theta|\text{model})}_{\text{prior}} \, \mathrm{d}\theta \tag{20}$$

Problem with Bayes factors 1: marginal likelihood's sensitivity to priors

Suppose $X \sim \text{binomial}(10, \theta)$ likelihood, and $\theta \sim \text{beta}(a, a)$ prior. **Question:** how does the marginal likelihood vary as $a \uparrow$?



Problem with Bayes factors 2: difficulty estimating marginal likelihood

For most problems $\theta = \theta_1, ..., \theta_p$,

$$p(X|\text{model}) = \int ... \int p(X|\theta_1, ..., \theta_p, \text{model}) \times \tag{21}$$

$$p(\theta_1, ..., \theta_p|\text{model}) \, \mathrm{d}\theta_1 ... \mathrm{d}\theta_p \tag{22}$$

$\Longrightarrow$ integral too difficult to compute in practice. But can approximate using Monte Carlo integration,

$$p(X|\text{model}) \approx \frac{1}{S} \sum_{s=1}^{S} p(X|\theta_{1s}, ..., \theta_{ps}, \text{model}) \tag{23}$$

where $\theta_{1s}, ..., \theta_{ps} \sim p(\theta_1, ..., \theta_p|\text{model})$, i.e. priors.

Problem with Bayes factors 2: difficulty estimating marginal likelihood

$$p(X|\text{model}) \approx \frac{1}{S}\sum_{s=1}^{S} p(X|\theta_{1s}, ..., \theta_{ps}, \text{model}) \tag{24}$$

However simple Monte Carlo integration is very slow to converge, particularly when there is a mismatch between the position of the prior and likelihood. Solutions exist, for example:

- Bayesian Monte Carlo.
- Importance Sampling Annealing.
- Adiabatic Monte Carlo.

But still early days in the research.

Bayes factors: summary

- Require us to specify priors over model choice and calculate marginal likelihood.
- Marginal likelihood is highly sensitive to priors, even for nuisance parameters.
- Marginal likelihood hard to calculate exactly (multi-dimensional integral), and so approximate using sampling.
- $\Longrightarrow$ simple Monte Carlo is slow to converge.

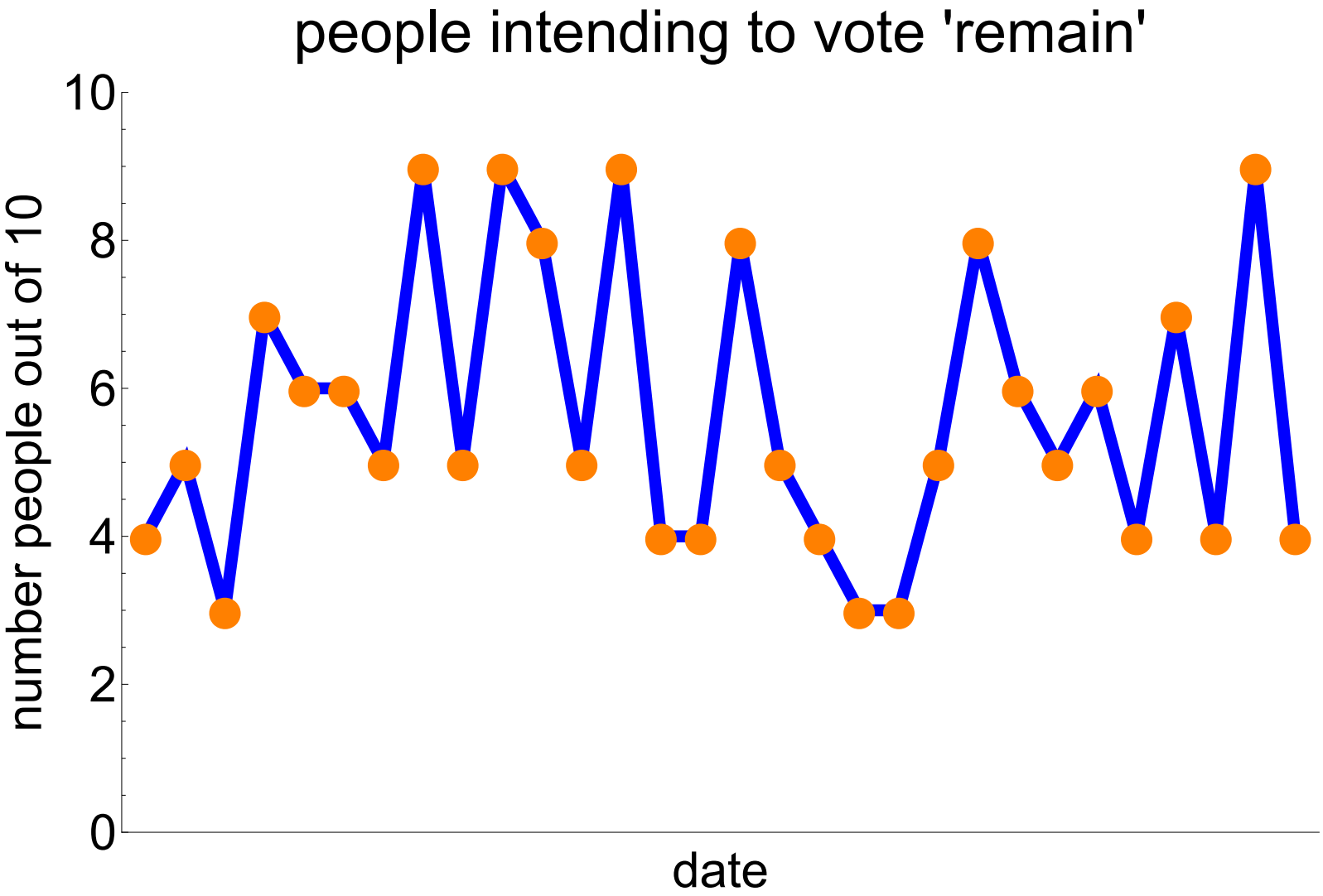
Model selection: summary

- Model selection inevitable part of scientific process $\Longrightarrow$ need a method for choosing between rival models.
- Posterior predictive checks useful in model building process but use predictive performance to select between hypotheses.
- Out-of-sample predictive performance is our aim.
- Two methods:
  - Post-hoc correction: Approximate LOO-CV using "loo" R package, or failing that use WAIC.
  - (Better) cross-validation: leave-one-out or k-folds.
- Bayes factors can be useful for model selection although need to be careful!

Example: EU referendum

- Imagine a dystopian future where the UK decides to hold a referendum on its membership of the EU.
- Have data from 20 polls on EU membership carried out over the past month (fake data.)
- Polls conducted by a range of different agencies.
- The sample size of each poll is 10.
- Ultimately want to use model to forecast the result of the EU referendum.

EU referendum: data



EU referendum: complete pooling model

Suppose that the we assume that the data across all polls are:

- Independent.
- Identically-distributed.

Sample size is fixed and data are discrete $\implies$ binomial likelihood:

$$Pr(X = X_i|\theta) = \theta^{X_i}(1-\theta)^{10-X_i} \tag{25}$$

where X_i is the number of people voting 'remain' in poll i, and θ is the probability that a randomly-chosen person will vote 'remain'.

Important: we are assuming that θ is the same across all polls.

EU referendum: complete pooling model in Stan

```
data {
  int<lower=1> K; // number of polls
  int<lower=0> X[K]; // numbers voting 'remain'
  int<lower=1> N; // sample size
}
parameters {
  real<lower=0,upper=1> theta;
}
model {
  for (i in 1:K){
    X[i] ~ binomial(N,theta);
  }
}
```

Implicitly $\Longrightarrow$ that we are using a uniform prior on (0,1) for theta.

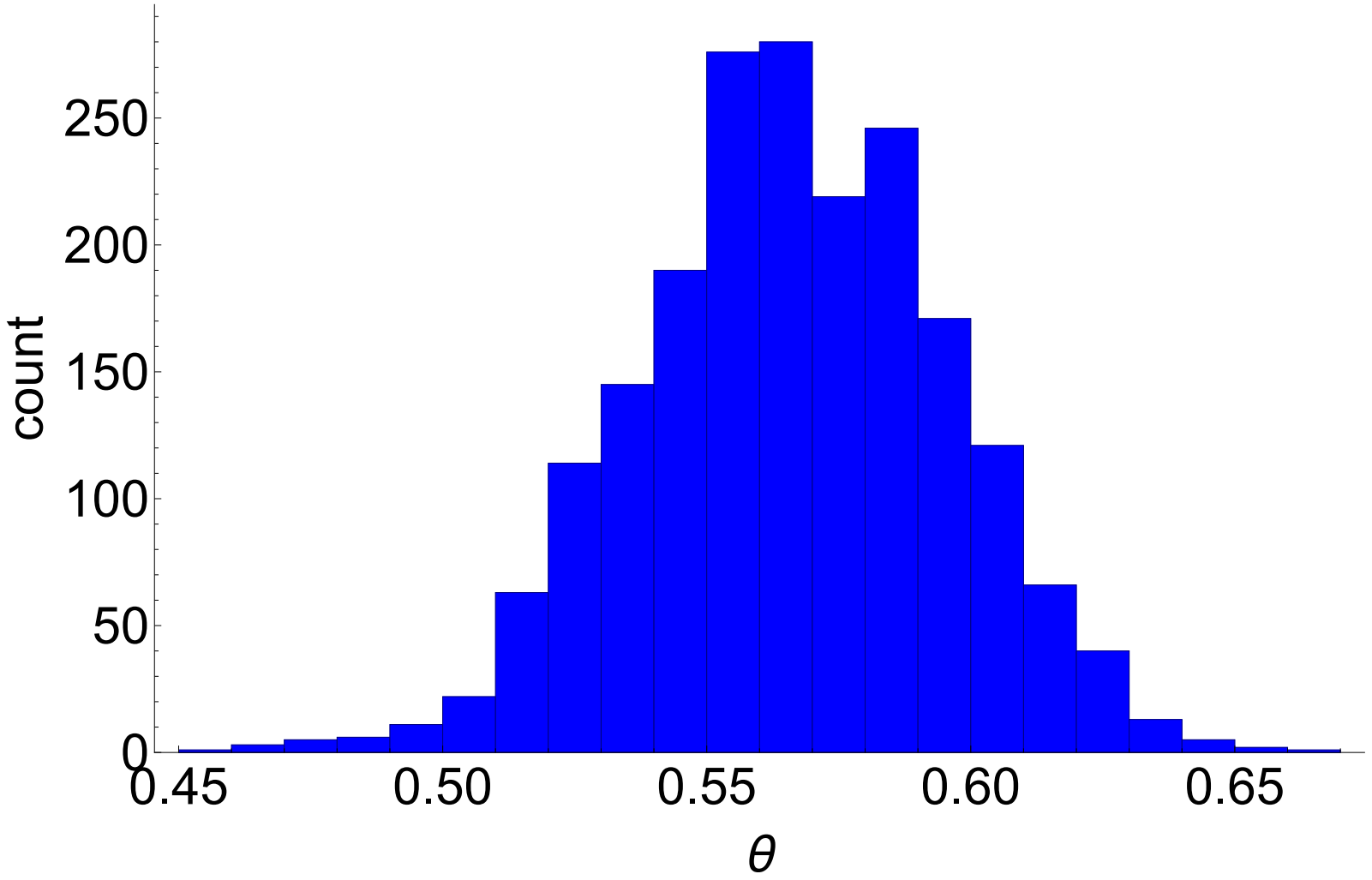
EU referendum: complete pooling model results

```
print(fit,probs = c(0.25, 0.5, 0.75))

## Inference for Stan model: lec6_euBinomialHomgeneousSimple.
## 4 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500..
##
##          mean se_mean   sd     25%     50%     75% n_eff Rhat
## theta    0.57    0.00 0.03    0.55    0.57    0.59   656 1.01
## lp__  -206.59    0.02 0.65 -206.75 -206.32 -206.17   774 1.00
```

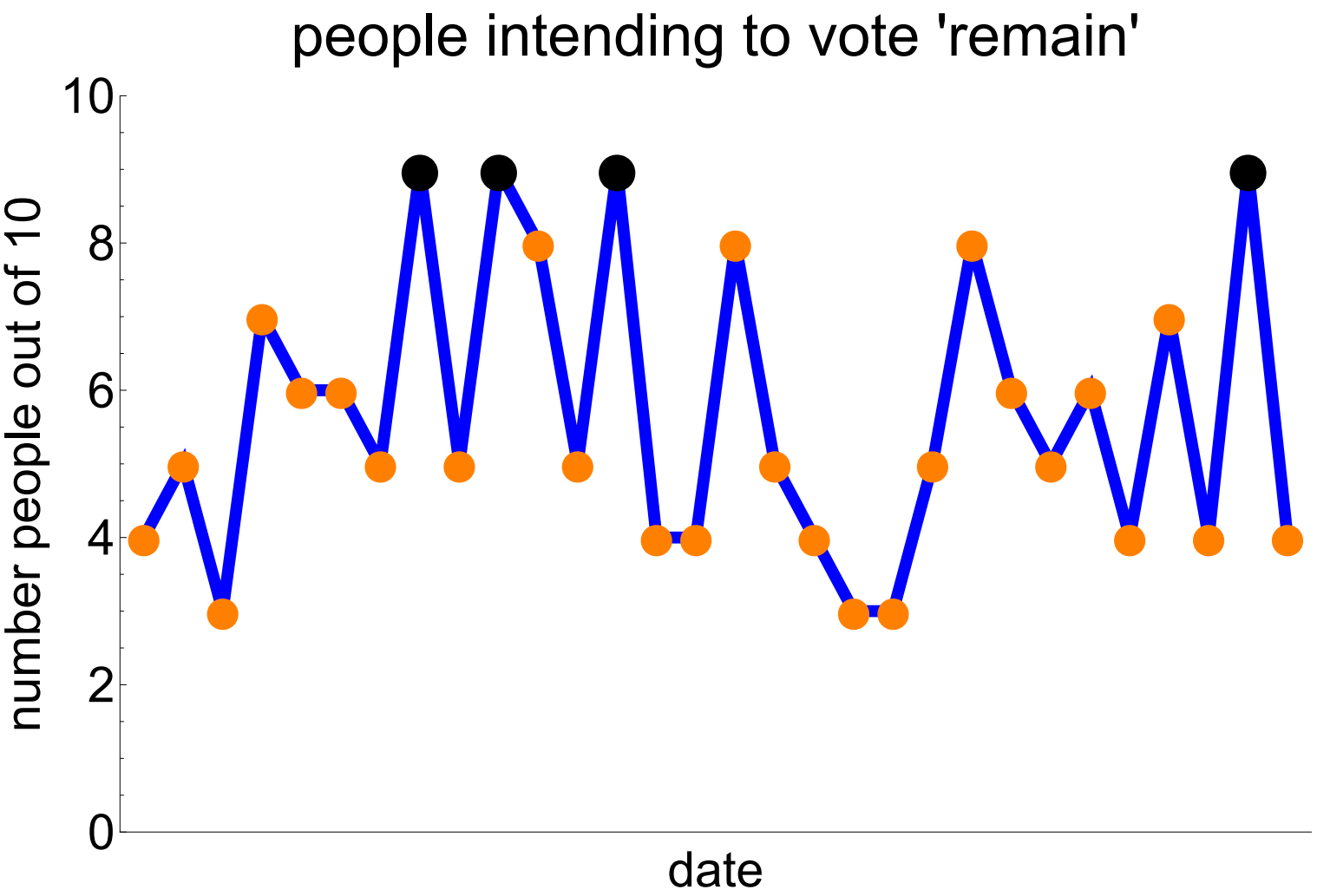
$\Longrightarrow$ all looks ok.

EU referendum: complete pooling model results



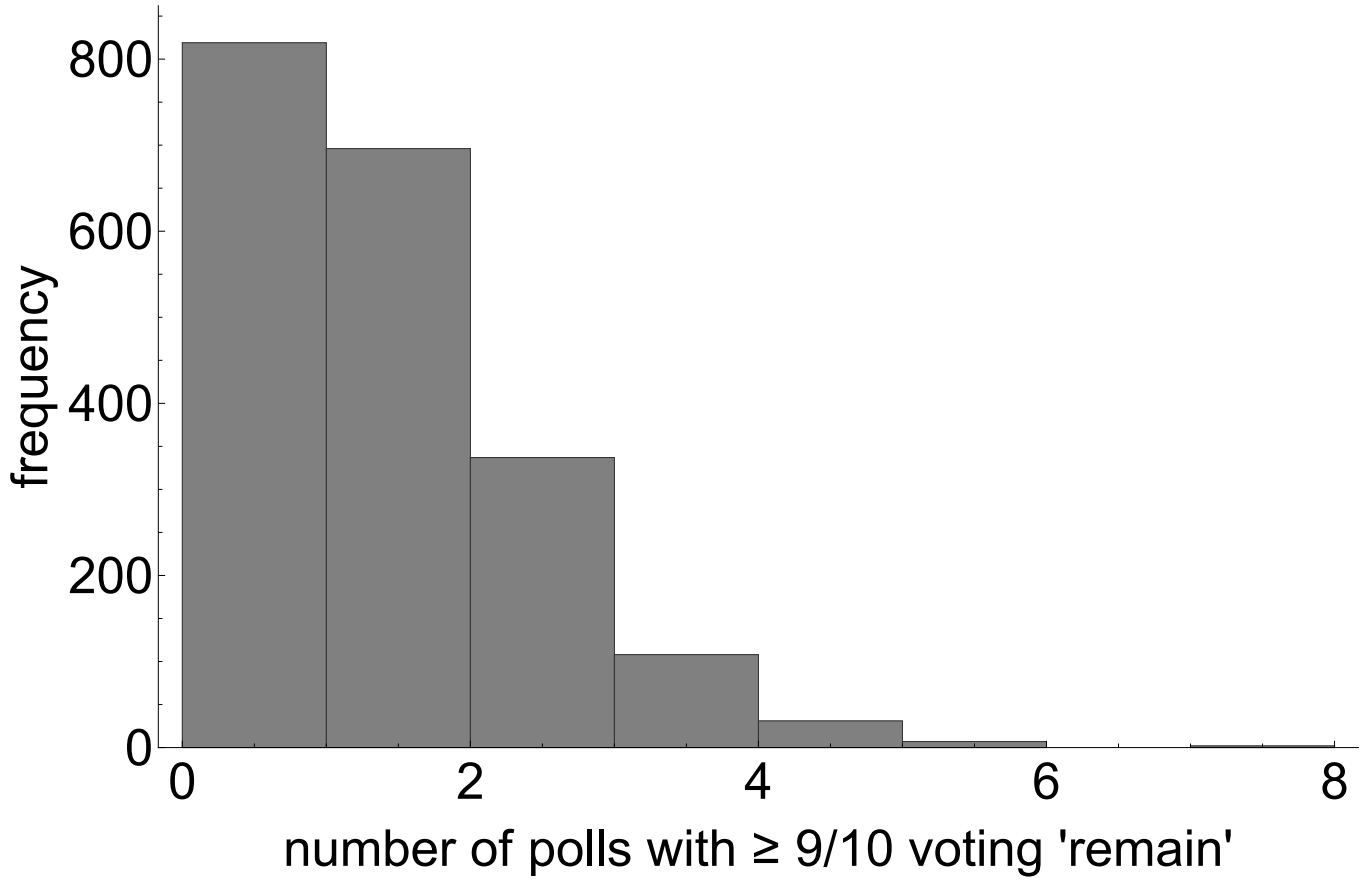
EU referendum: complete pooling model posterior predictive checks

Count number of times that 9 or more people vote 'remain', and find 4/30 cases.



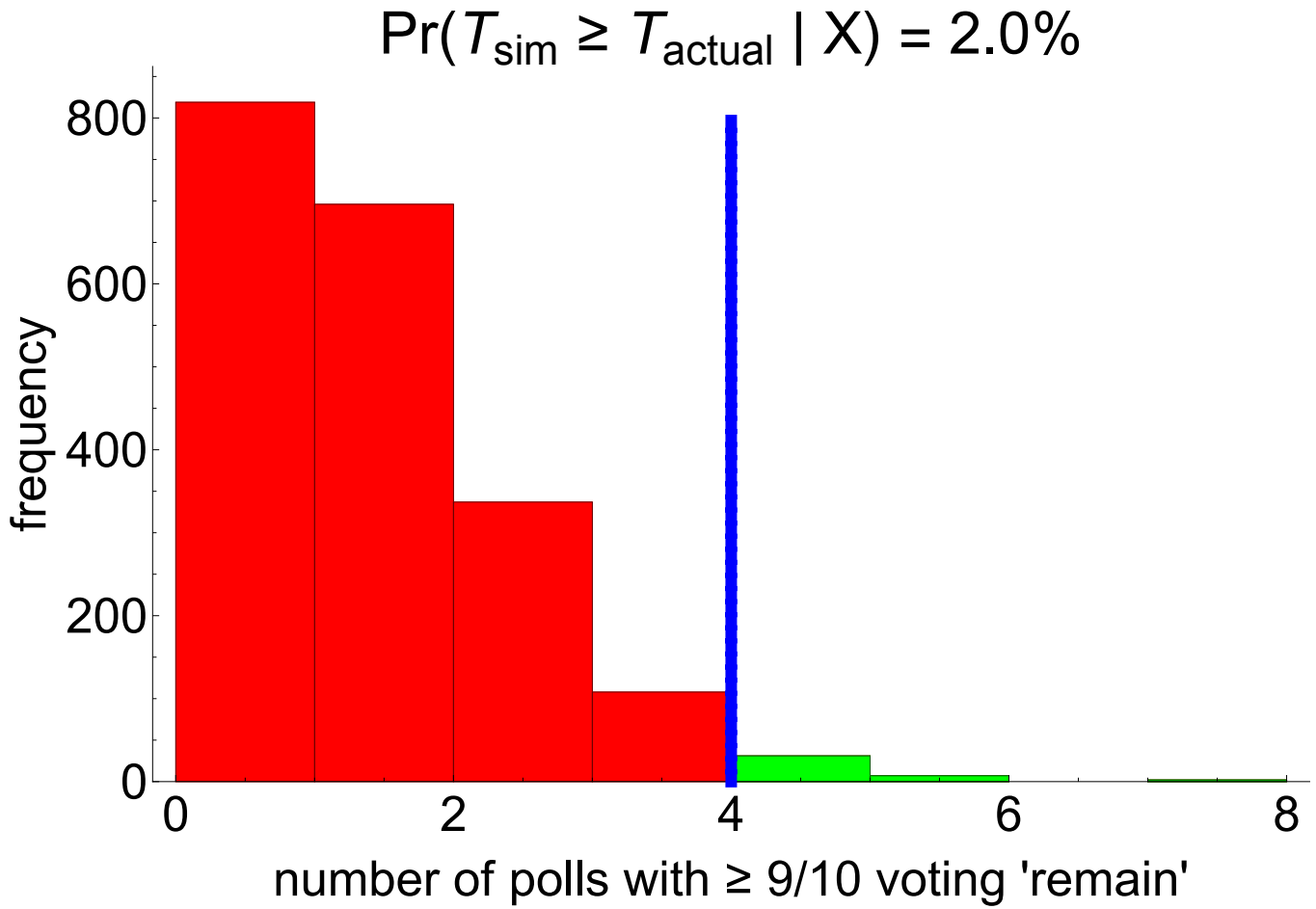
EU referendum: complete pooling model posterior predictive checks

Repeat for 2000 simulated data series; for each simulated dataset counting the number of $X_i \geq 9$.



EU referendum: complete pooling model posterior predictive checks

Low probability of replicating this aspect of real data.



EU referendum: complete pooling model summary

- Assumed a model where the probability of a polled individual intending to vote 'remain' is **identical** across all polls.
- However polls were conducted over a range of time by a range of agencies, each with their own methodology $\implies$ sampling method, exact interview process etc vary.
- Therefore assuming a common θ across polls is too strong an assumption.
- $\implies$ model will **understate** true uncertainty.

$\implies$ try the opposite extreme where we allow a different θ_i for each poll.

EU referendum: heterogeneous model

For each poll i we assume a binomial likelihood:

$$Pr(X = X_i|\theta_i) = \theta_i^{X_i}(1-\theta_i)^{10-X_i} \tag{26}$$

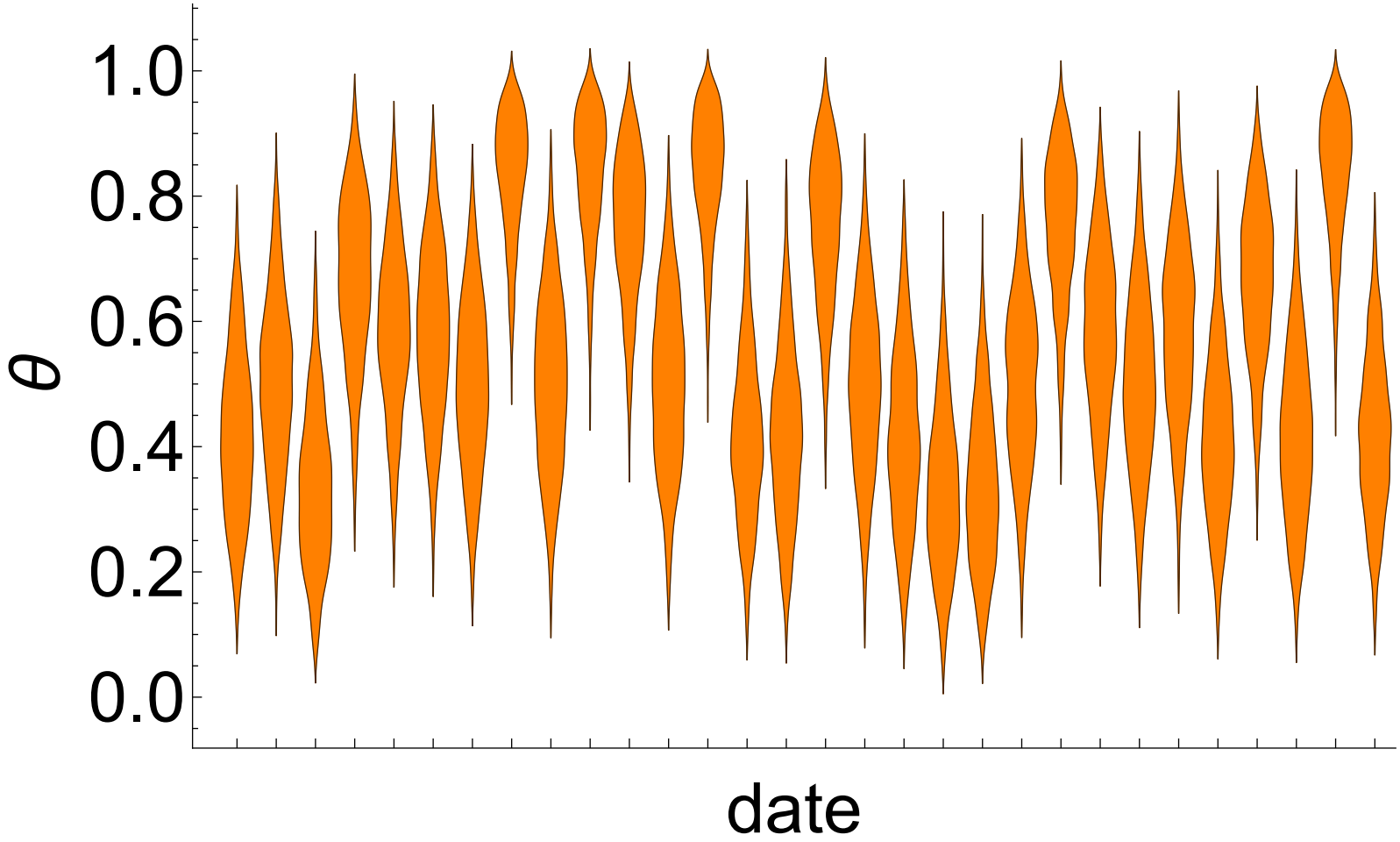
where θ_i is the probability of a interviewee indicating they intend to vote 'remain' in the referendum.
By allowing θ_i to vary across polls $\implies$ **different** data generating processes for each poll!

EU referendum: heterogeneous model

```
data {
  int<lower=1> K; // number of polls
  int<lower=0> X[K]; // numbers voting 'remain'
  int<lower=1> N; // sample size
}
parameters {
  real<lower=0,upper=1> theta[K]; // now array
}
model {
  for (i in 1:K){
   X[i] ~ binomial(N,theta[i]);// select element
  }
}
```

$\Longrightarrow$ only two changes to homogeneous model necessary.

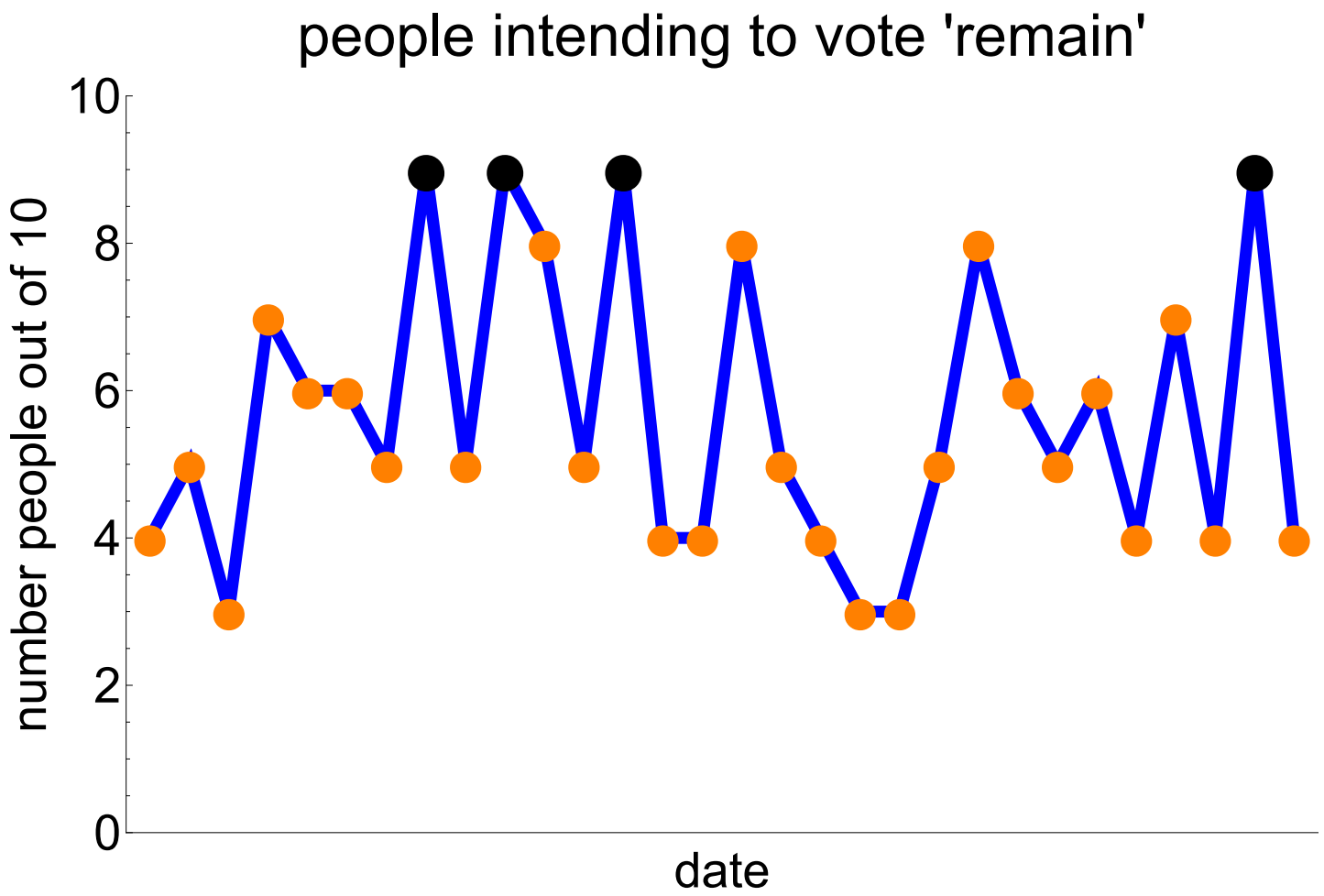
Heterogeneous model results



$\Longrightarrow$ large uncertainty associated with each θ_i.

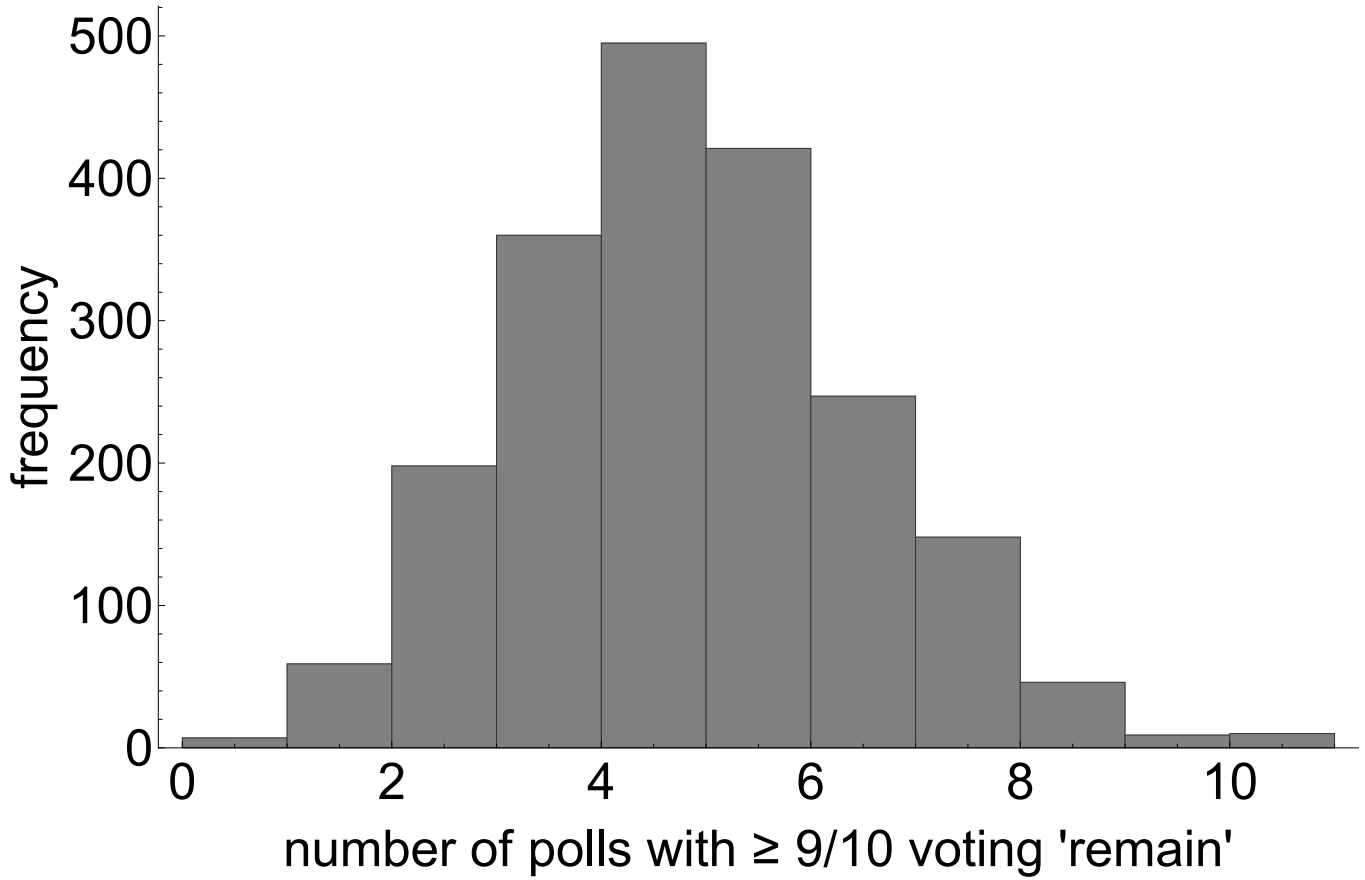
Heterogeneous model posterior predictive checks

Compare again occurrence of 9+/10 'remain' voters with real data; where 4/30.



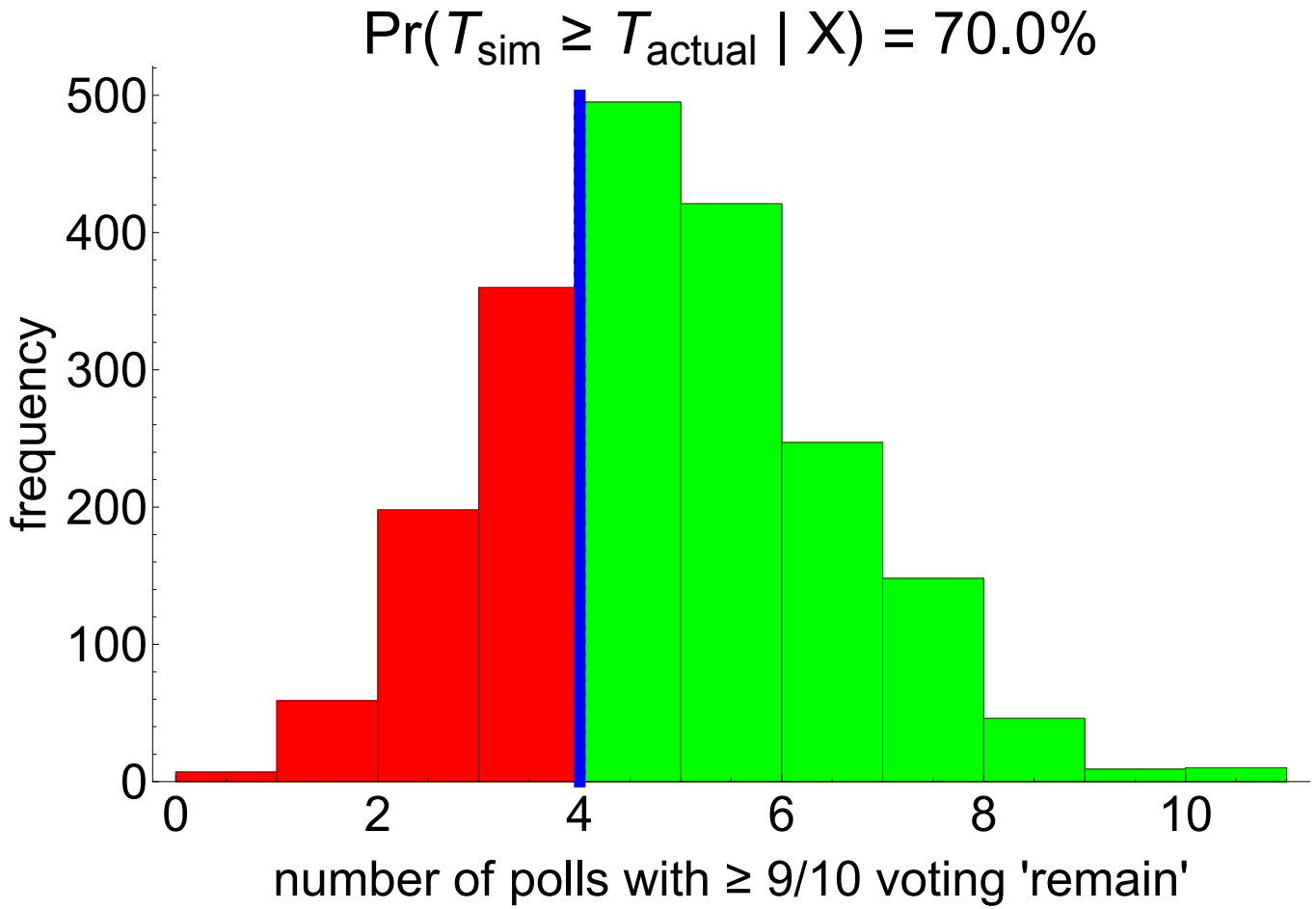
Heterogeneous model posterior predictive checks

For 2000 posterior predictive samples.



Heterogeneous model posterior predictive checks

$\Longrightarrow$ better than complete pooling model.



Heterogeneous model problems

- Heterogeneous model is a better fit to the data than fully-pooled $\Longleftarrow$ is overfitting the data?
- However not clear what we should now forecast for the polls? Should we use:
  - Estimate from one poll.
  - Average across point estimates for all polls.
- Further how should we quantify our uncertainty?

Heterogeneous model: summary

- Same binomial likelihood as before but allowed θ to vary across polls.
- $\implies$ obtained separate estimates of θ across each poll.
- Found considerable variability and uncertainty in estimates of θ_i.
- Heterogeneous θ model better captured the extremes seen in the data $\implies$ fully-pooled model was too strong.
- However key question: **what do we forecast will be the result of the EU referendum, and with what uncertainty?**

Introducing a hierarchical model

- In fully-pooled model case we assumed that data from all the polls was the same; i.e. θ was constant.
- However there are differences between polling methodologies, and time when polls were taken $\implies$ data generated by different processes.
- $\implies$ we estimated a separate θ_i for each poll; i.e. assumed data from different polls was completely unrelated.

Introducing a hierarchical model

Question 1: do we really think that data from different polls is completely unrelated? **Answer 1:** no! After all the polls measure the same thing, at around the same point in time.

Question 2: do we really think that the polls are exactly the same? **Answer 2** no! We rejected this initially because of differences between polling agencies and time over which the polls were done.

Introducing a hierarchical model

Question: isn't there somewhere between the extremes of complete-separation and fully-pooled?

Answer:

Hierarchical model for EU referendum polls

As for heterogeneous model we allow θ to vary by poll:

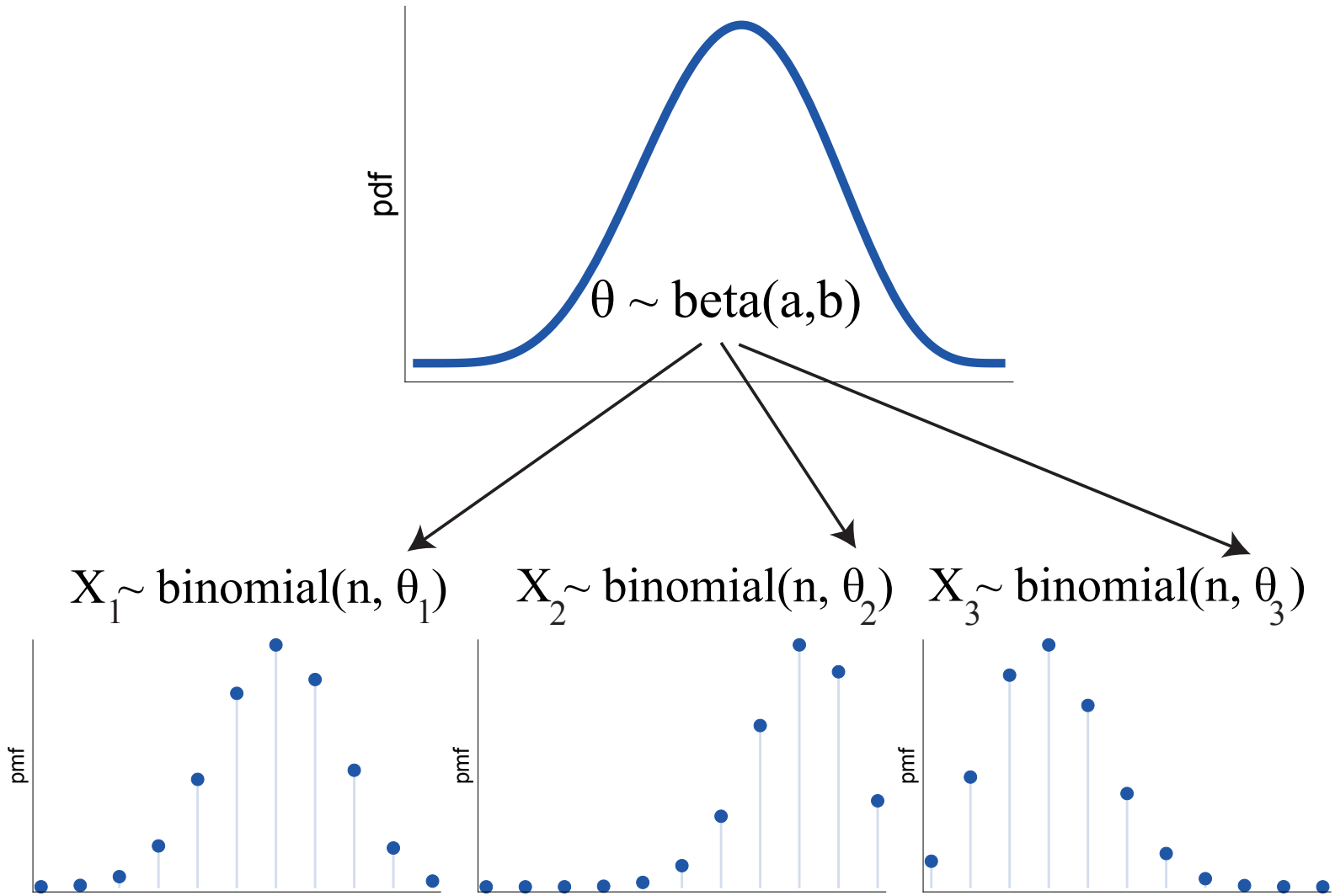
$$Pr(X = X_i|\theta_i) = {\theta_i}^{X_i}(1-\theta_i)^{10-X_i} \tag{27}$$

However now we assume that the θ_i are drawn from a common "population" distribution:

$$\theta_i \sim beta(a, b) \tag{28}$$

where a and b are parameters that define the population-level beta distribution.

Hierarchical model for EU referendum polls



Hierarchical model for EU referendum polls

$$\theta_i \sim \mathit{beta}(a, b) \tag{29}$$

(a, b) are parameters just like any other in Bayesian inference $\Longrightarrow$ assign them **priors**!
Actually easier to set priors for transformed parameters:

$$a = \alpha \times \kappa$$
$$b = (1 - \alpha) \times \kappa$$

where α represents the "population" chance of voting 'remain' and κ measures the concentration $\Longrightarrow$

$$\theta_i \sim \mathit{beta}(\alpha \times \kappa, (1 - \alpha) \times \kappa) \tag{30}$$

Hierarchical model for EU referendum polls

$$\theta_i \sim \textit{beta}(\alpha \times \kappa, (1-\alpha) \times \kappa) \tag{31}$$

Set independent priors:

$$p(\alpha, \kappa) = p(\alpha) \times p(\kappa) \tag{32}$$

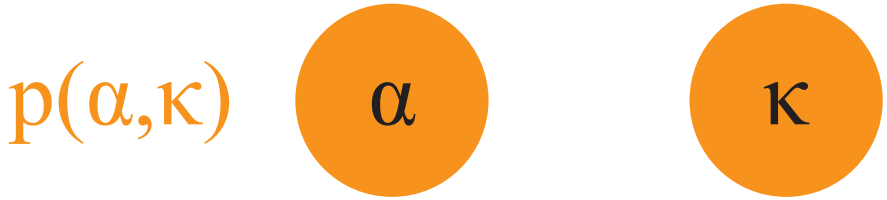
Question: what is the numerator of Bayes' rule for this problem?
Answer: the joint distribution of the data X and parameters; i.e. $p(X, \theta, \alpha, \kappa)$
Another question: how do we find this here? **Another answer:** exploit conditional independence of the problem!

Hierarchical model for EU referendum polls

Start with "population" level parameters.

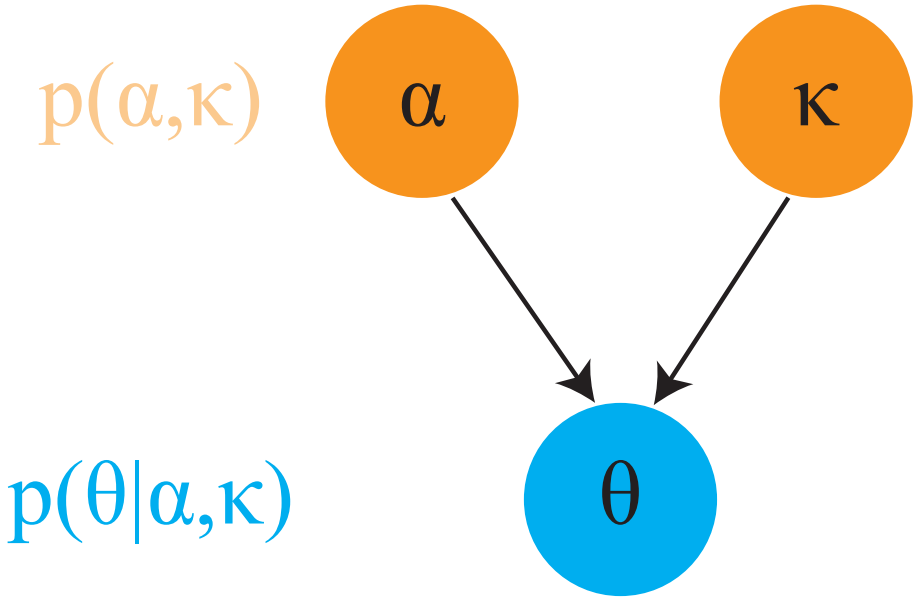
Hierarchical model for EU referendum polls

And determine their joint probability.



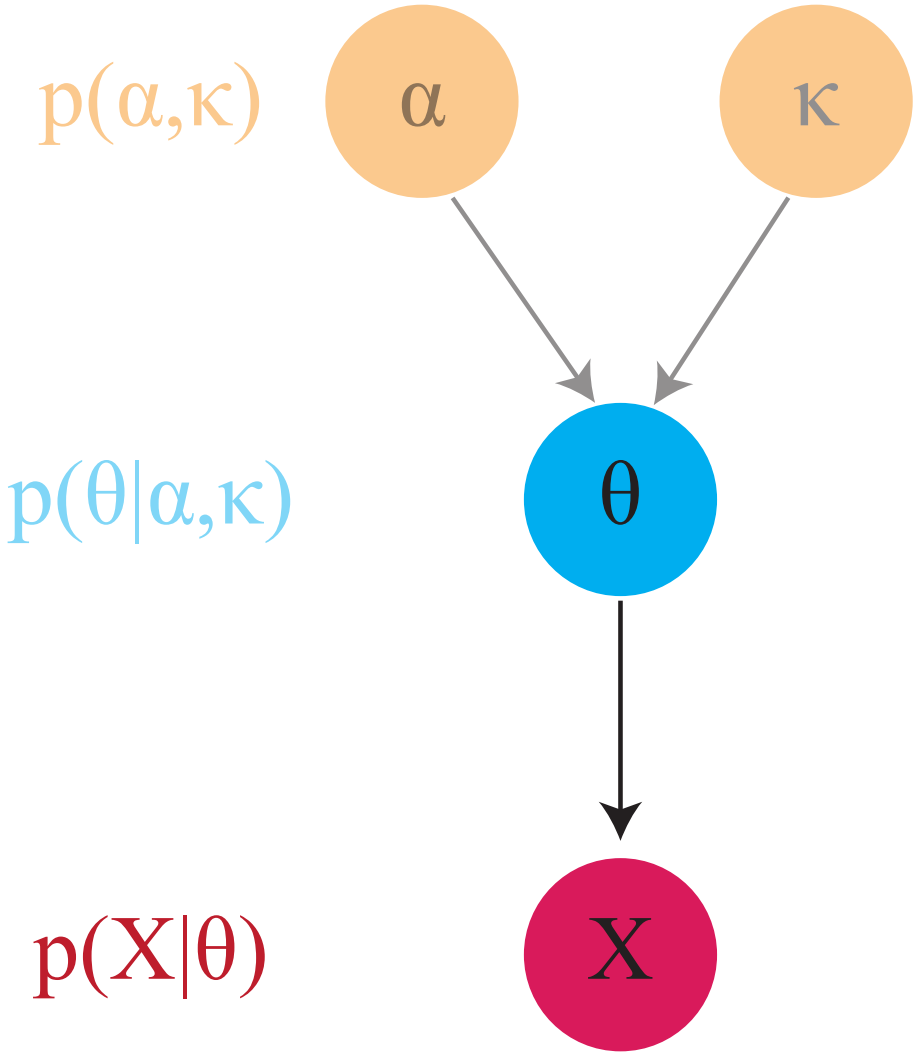
Hierarchical model for EU referendum polls

Find the probability of θ **conditional** on α and κ.



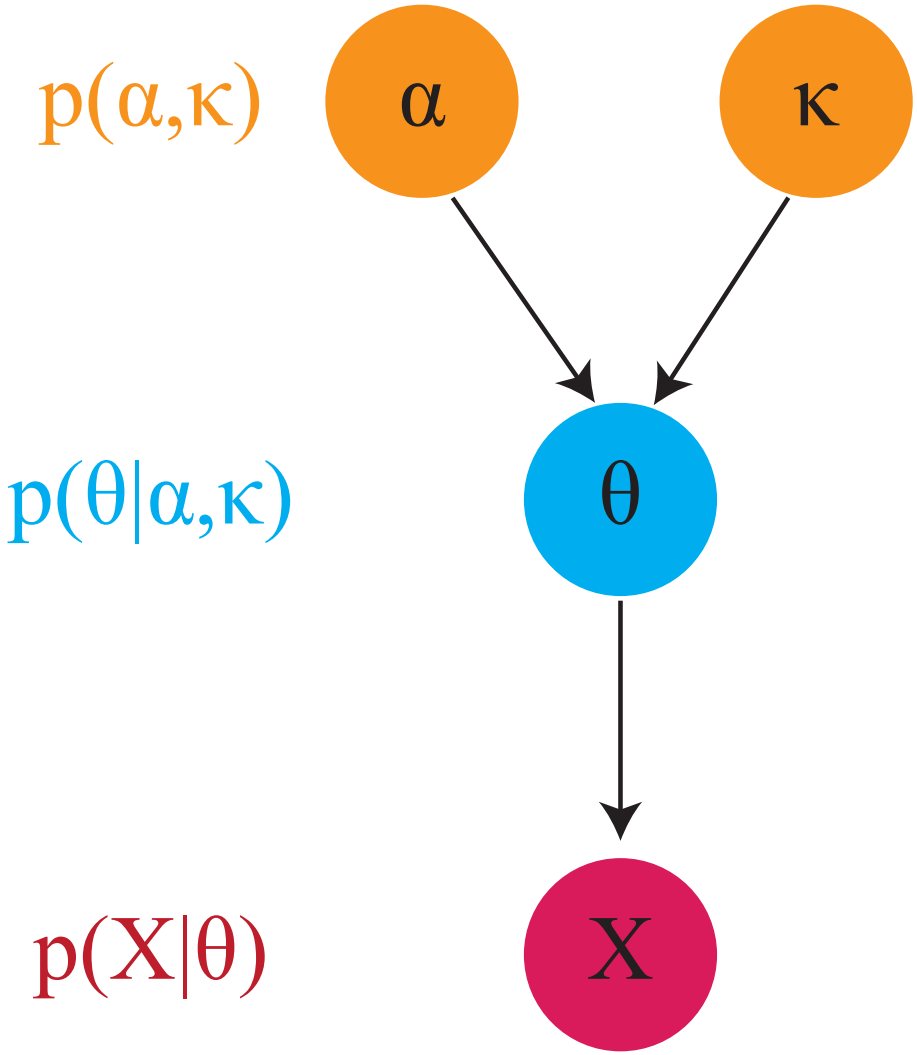
Hierarchical model for EU referendum polls

Find the probability of X **conditional** on θ .



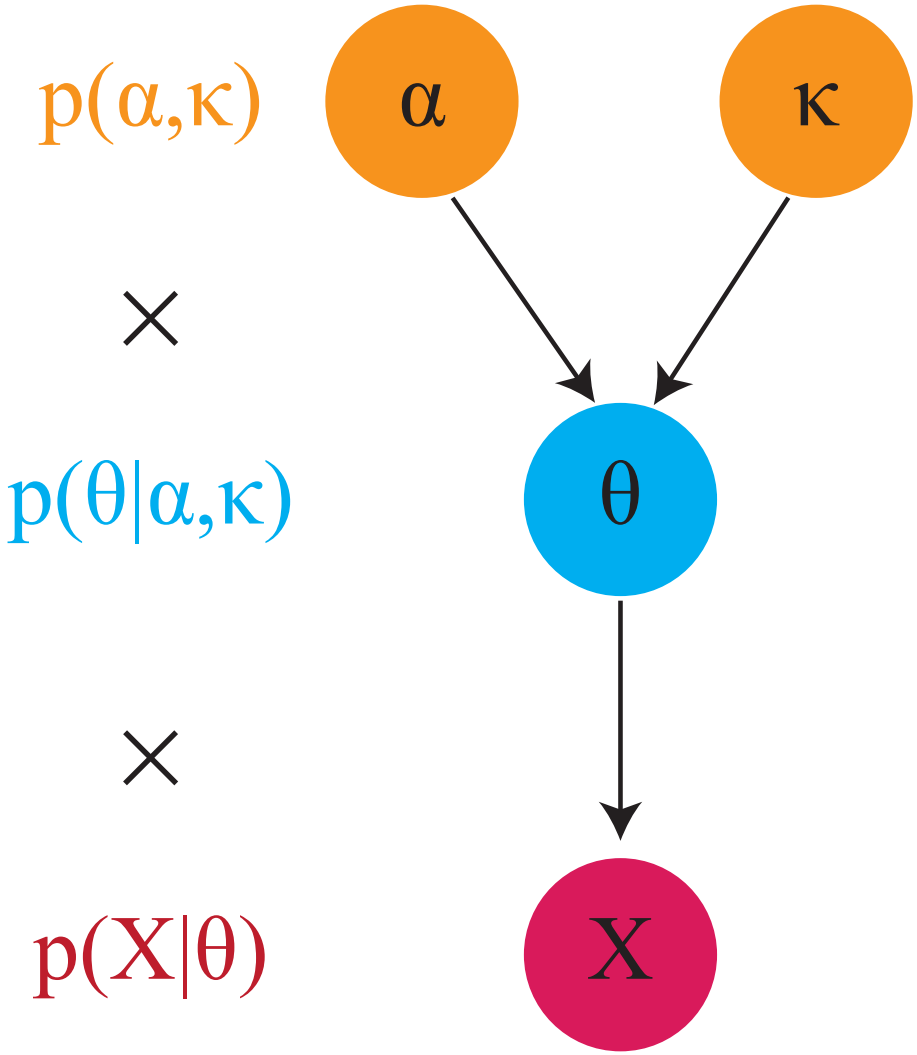
Hierarchical model for EU referendum polls

Finally to obtain the overall probability...



Hierarchical model for EU referendum polls

...multiply together all the terms.



Hierarchical model for EU referendum polls

So the posterior is found as:

$$p(\theta, \alpha, \kappa | X) \propto p(X, \theta, \alpha, \kappa)$$
$$= \underbrace{p(X|\theta)}_{\text{likelihood}} \times \underbrace{p(\theta|\alpha, \kappa)}_{\text{prior}} \times \underbrace{p(\alpha, \kappa)}_{\text{hyper-prior}}$$

- $p(X|\theta)$ is just the **likelihood**.
- $p(\theta|\alpha, \kappa)$ is the **prior** on θ.
- $p(\alpha, \kappa)$ is the **hyper-prior** on the **hyper-parameters** α and κ.
- However the word "hyper" is really just a fancy word we use to represent priors on "population" level parameters.
- In hierarchical models there is a blurring of the likelihood/prior boundary.

Hierarchical model for EU referendum polls: back to the problem

We set the following (hyper-)priors on α and κ:

$$\alpha \sim \textit{beta}(5, 5)$$
$$\kappa \sim \textit{pareto}(1, 0.3)$$

where:

- beta(5,5) has a mean of 0.5, and only has support for $0 \leq \alpha \leq 1$.
- pareto(1,0.3) is a distribution only with support for values of $\kappa \geq 1$.

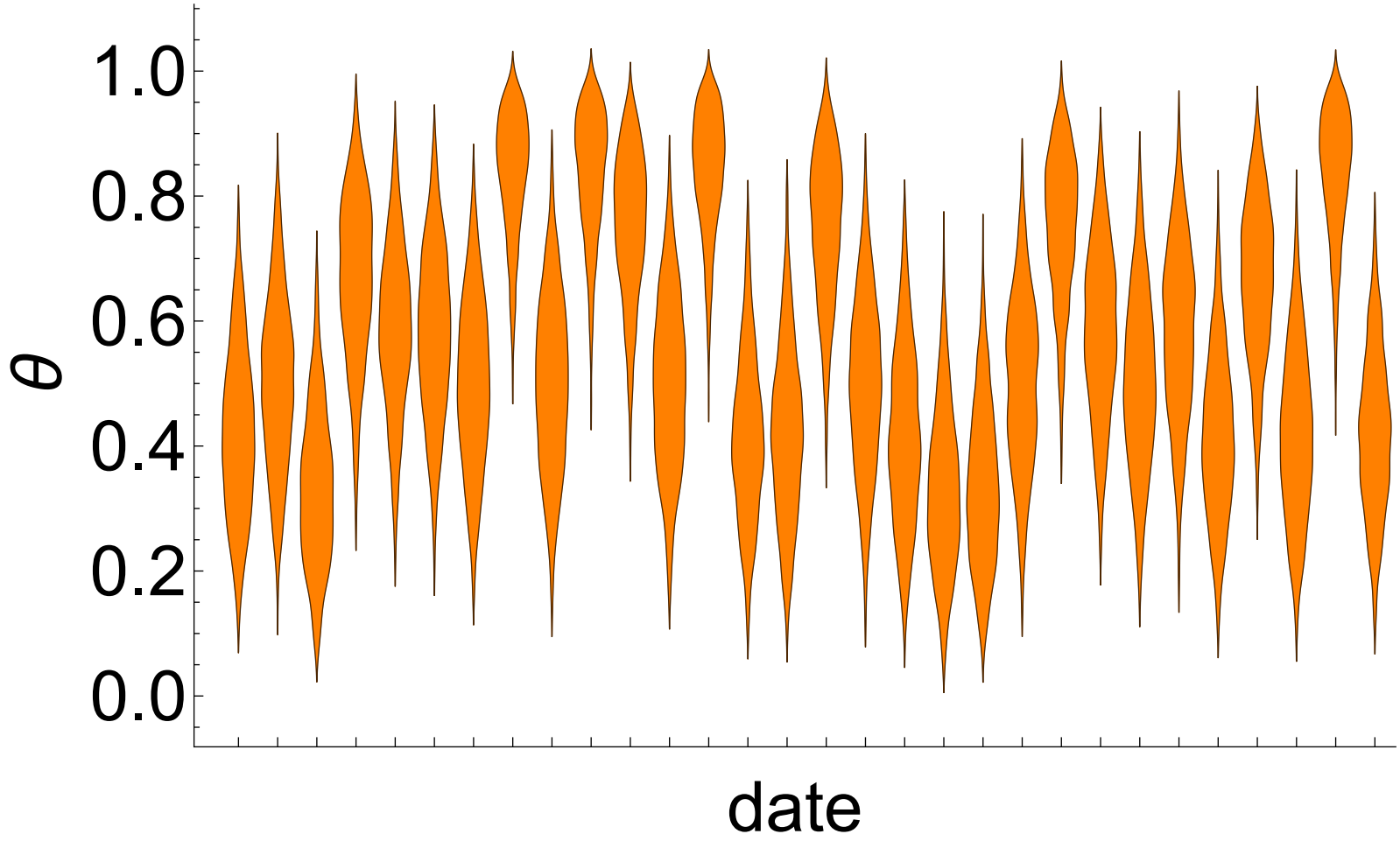
Hierarchical model for EU referendum polls: coding up model in Stan

```
parameters {
  real<lower=0, upper=1> alpha;
  real<lower=1> kappa;
  vector<lower=0, upper=1>[K] theta;
}
model {
  for (i in 1:K){
   Y[i] ~ binomial(N[i],theta[i]);// Likelihood
  }
  // prior
  theta ~ beta(alpha * kappa, (1 - alpha) * kappa);

  // hyper-priors
  kappa ~ pareto(1, 0.3);
  alpha ~ beta(5,5);
}
```

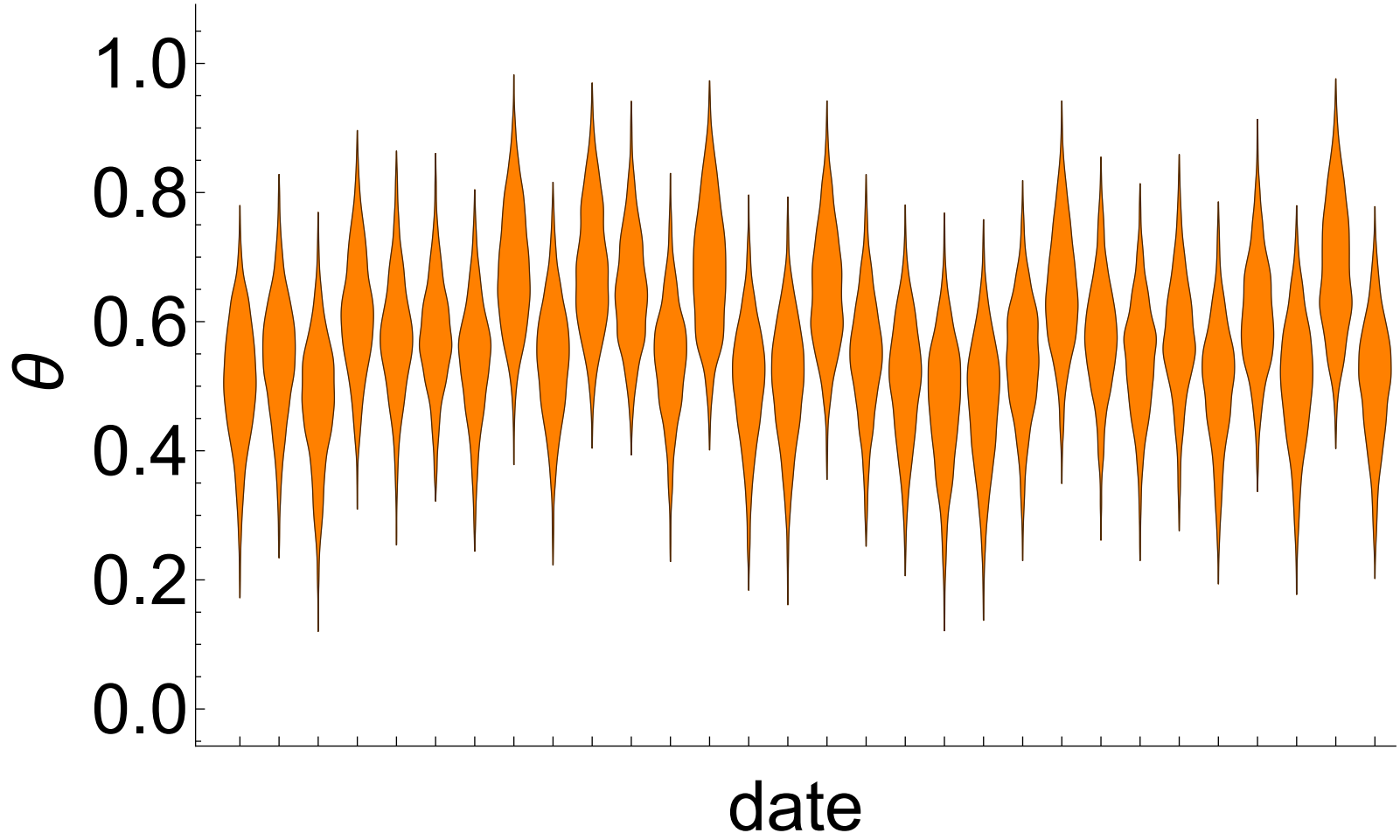
Hierarchical model for EU referendum polls: heterogeneous model estimates

Remember the posterior from the heterogeneous model?



Hierarchical model for EU referendum polls: hierarchical model estimates

Hierarchical model estimates.



Hierarchical model for EU referendum polls: hierarchical model estimates

Two effects evident:

- Shrinkage to **grand** mean.
- Shrinkage in variance.

Shrinkage to grand mean because hierarchical models lie on a spectrum between completely heterogeneous estimates and fully pooled.
Important: the data determines where on the spectrum we exactly end up! No choice in analysis.
Important: shrinkage helps reduce the variance of estimates $\implies$ outliers have less impact.

Also "partially-pooling" information across groups $\implies$ essentially a larger sample size and lower variance.

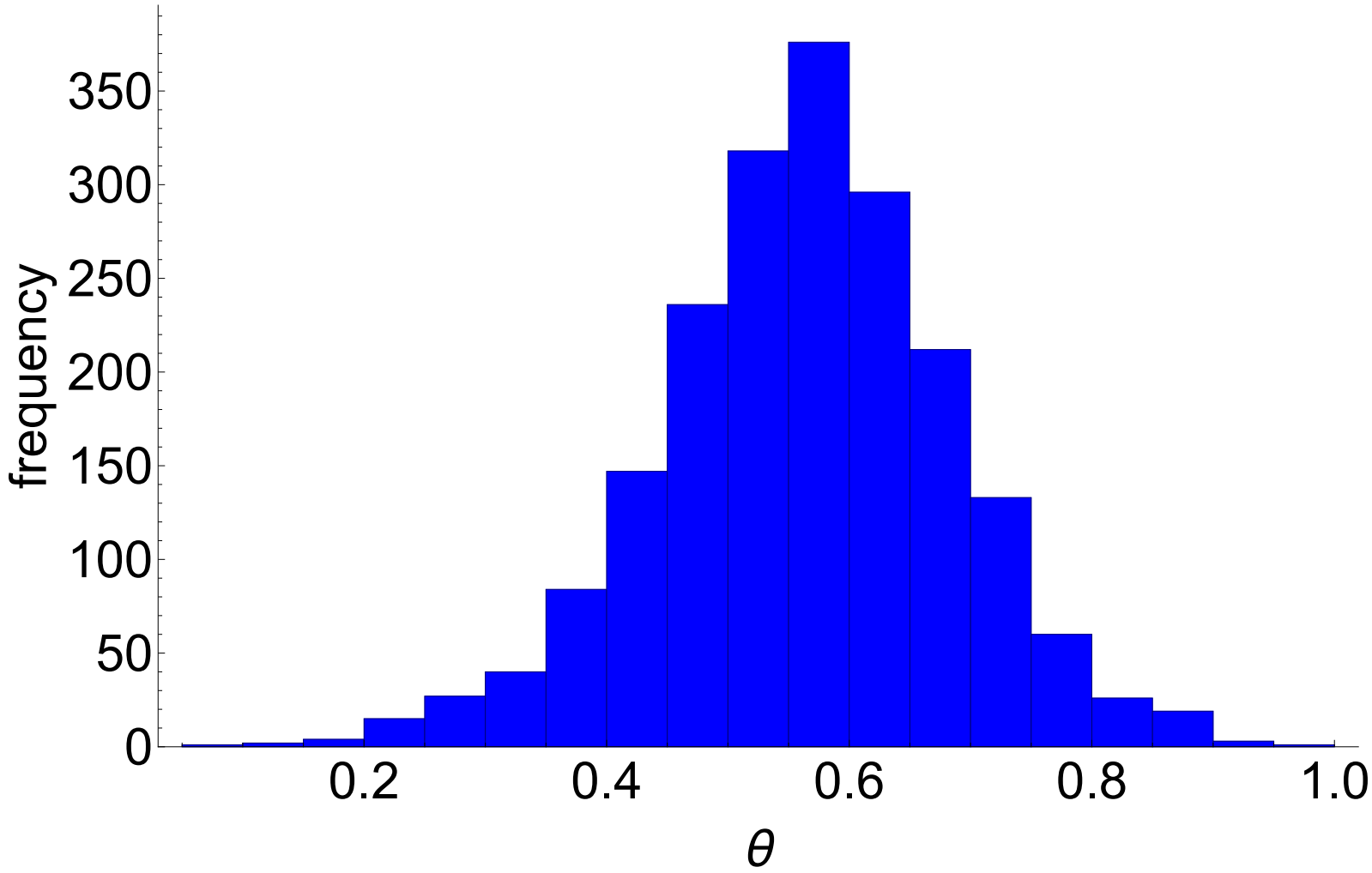
Hierarchical model: forecasting outcome of overall elections

Want to estimate the value of θ for the last poll: the referendum itself!
To do this do the following:

1. Sample (α, κ) from their posteriors.
2. Sample $\theta \sim \mathit{beta}(\alpha\kappa, (1-\alpha)\kappa)$.

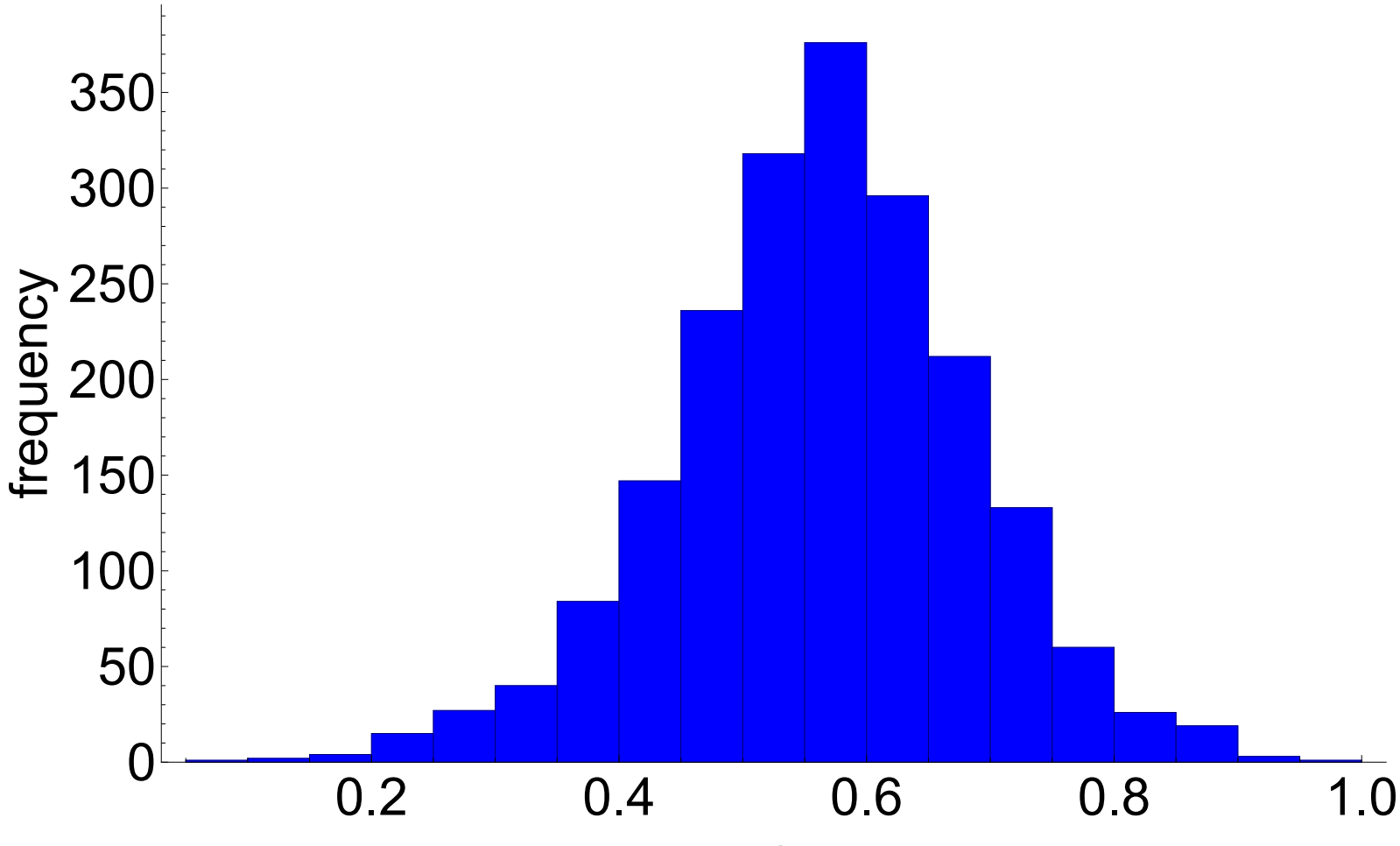
Hierarchical model: forecasting outcome of overall elections

Yields the posterior below (gulp):



Hierarchical model: forecasting outcome of overall elections

Conclusion: the result could go either way (I did a similar analysis for real poll data and it produced a posterior very much like the one below!)



Hierarchical models: summary

- Model with **same** parameters for all groups $\Longrightarrow$ data generating process is the **same**.
- Model with separately estimated group-level parameters $\Longrightarrow$ data generating process is completely **different**.
- Frequently neither of the aforementioned models are appropriate; i.e. we want some dependence between parameters but not 100%.
- $\Longrightarrow$ use hierarchical model where the data determines parameter dependence across groups.
- In hierarchical models $\Longrightarrow$ **shrinkage** of **group** means towards the **grand** mean.
- Also **shrinkage** of group variance $\Longleftarrow$ sample size$\uparrow$ by **partial-pooling** information across groups.

Lecture summary

- Stan carries out inference by a variant of HMC called NUTS $\implies$ very fast!
- Stan is (hopefully) a fairly intuitive language to learn.
- Makes MCMC sampling easier for a wide class of problems in Bayesian inference.
- ODEs can be fit in Stan although can be slow due to computational expense of solving for sensitivities.
- Non-hierarchical models have failings: understate uncertainty, or overfit.
- Hierarchical models $\implies$ "happy" medium between fully-pooled and completely heterogeneous models.

Not sure I underSTANd

Overfit.

Not sure I underSTANd

Underfit.

Not sure I underSTANd

Fit.