SOMETIMES THE SNOW COMES DOWN IN JUNE SOMETIMES THE SUN GOES 'ROUND THE MOON

Daniel Simpson
Department of Statistical Sciences
University of Toronto

ESTHER WILLIAMS IN THE HAROLD HOLT MEMORIAL SWIMMING POOL

Some thoughts on complexity

Daniel Simpson
University of Toronto
@dan_p_simpson

HAROLD HOLT, THE MUSICAL (DRAMATIS PERSONAE)

HAROLD HOLT, THE MUSICAL (DRAMATIS PERSONAE)

- Harold Holt (17th Prime Minister of Australia)
  - Our metaphor for statisticians

HAROLD HOLT, THE MUSICAL (DRAMATIS PERSONAE)

- Harold Holt (17th Prime Minister of Australia)
  - Our metaphor for statisticians

HAROLD HOLT, THE MUSICAL (DRAMATIS PERSONAE)

HAROLD HOLT, THE MUSICAL (DRAMATIS PERSONAE)

- Harold Holt (17th Prime Minister of Australia)
  - Our metaphor for statisticians

HAROLD HOLT, THE MUSICAL (DRAMATIS PERSONAE)

- Harold Holt (17th Prime Minister of Australia)
  - Our metaphor for statisticians
- Harold Holt Memorial Swimming Pool (A swimming pool)
  - Our metaphor for statistical outputs

HAROLD HOLT, THE MUSICAL (DRAMATIS PERSONAE)

- Harold Holt (17th Prime Minister of Australia)
  - Our metaphor for statisticians
- Harold Holt Memorial Swimming Pool (A swimming pool)
  - Our metaphor for statistical outputs
- The Bass Strait (A large body of water)
  - Our metaphor for statistics

HAROLD HOLT, THE MUSICAL (DRAMATIS PERSONAE)

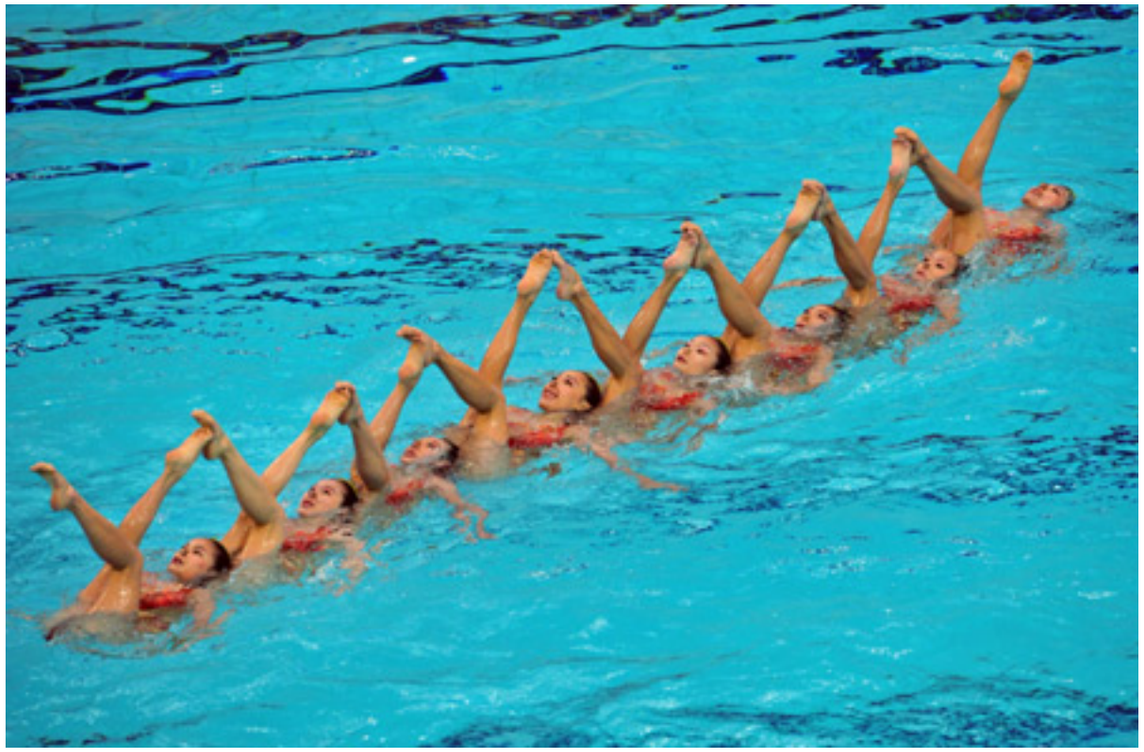
- Harold Holt (17th Prime Minister of Australia)
  - Our metaphor for statisticians
- Harold Holt Memorial Swimming Pool (A swimming pool)
  - Our metaphor for statistical outputs
- The Bass Strait (A large body of water)
  - Our metaphor for statistics
- Esther Williams (Esther Williams)
  - Our metaphor for prominent Bayesian inference libraries

GENERALLY SPEAKING, THINGS HAVE GONE ABOUT AS FAR AS THEY CAN POSSIBLY GO, WHEN THINGS HAVE GOTTEN ABOUT AS BAD AS THEY CAN REASONABLY GET.

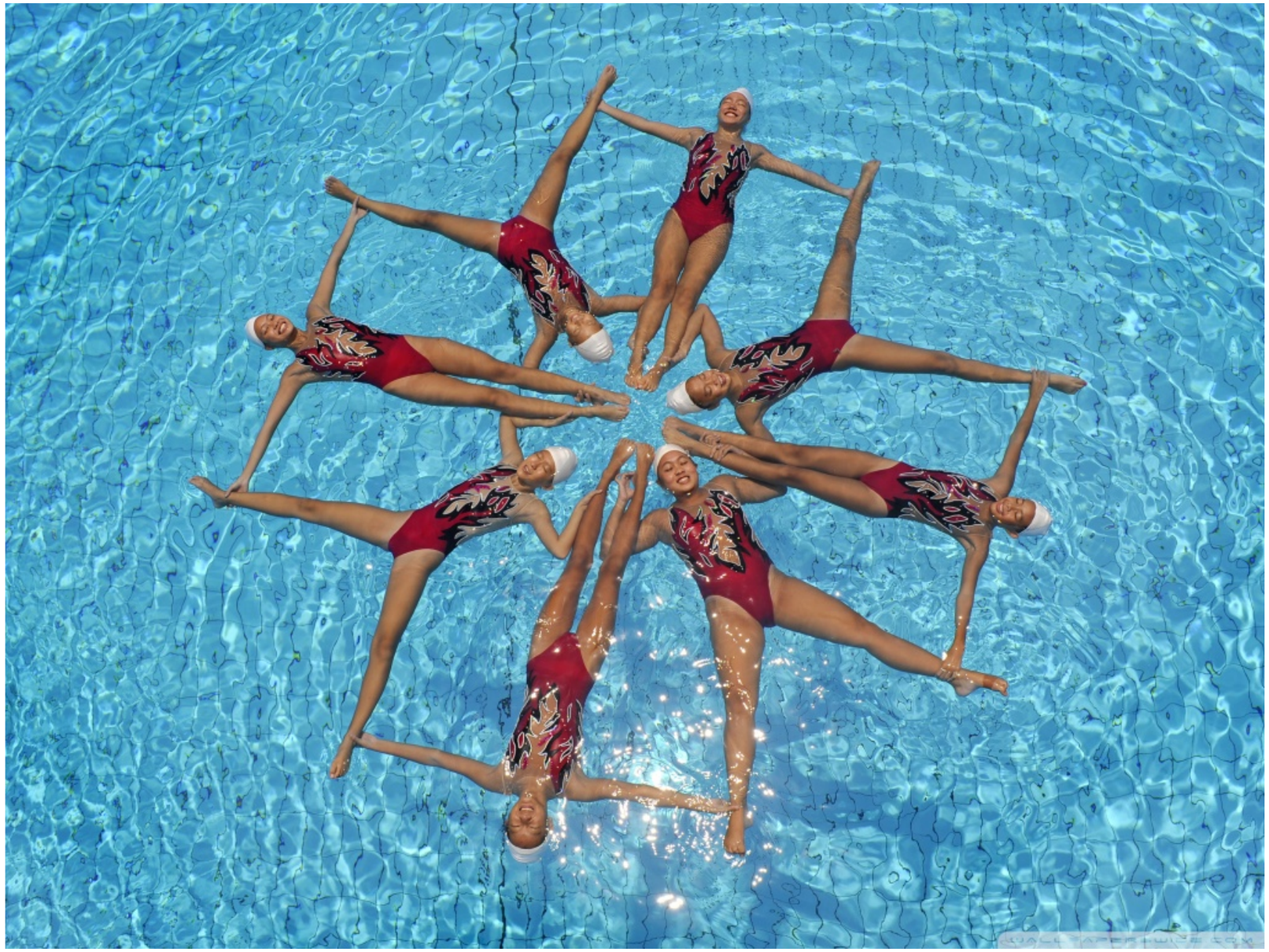
(Tom Stoppard)

WE USED TO JUST ESTIMATE MEANS OF GAUSSIANS

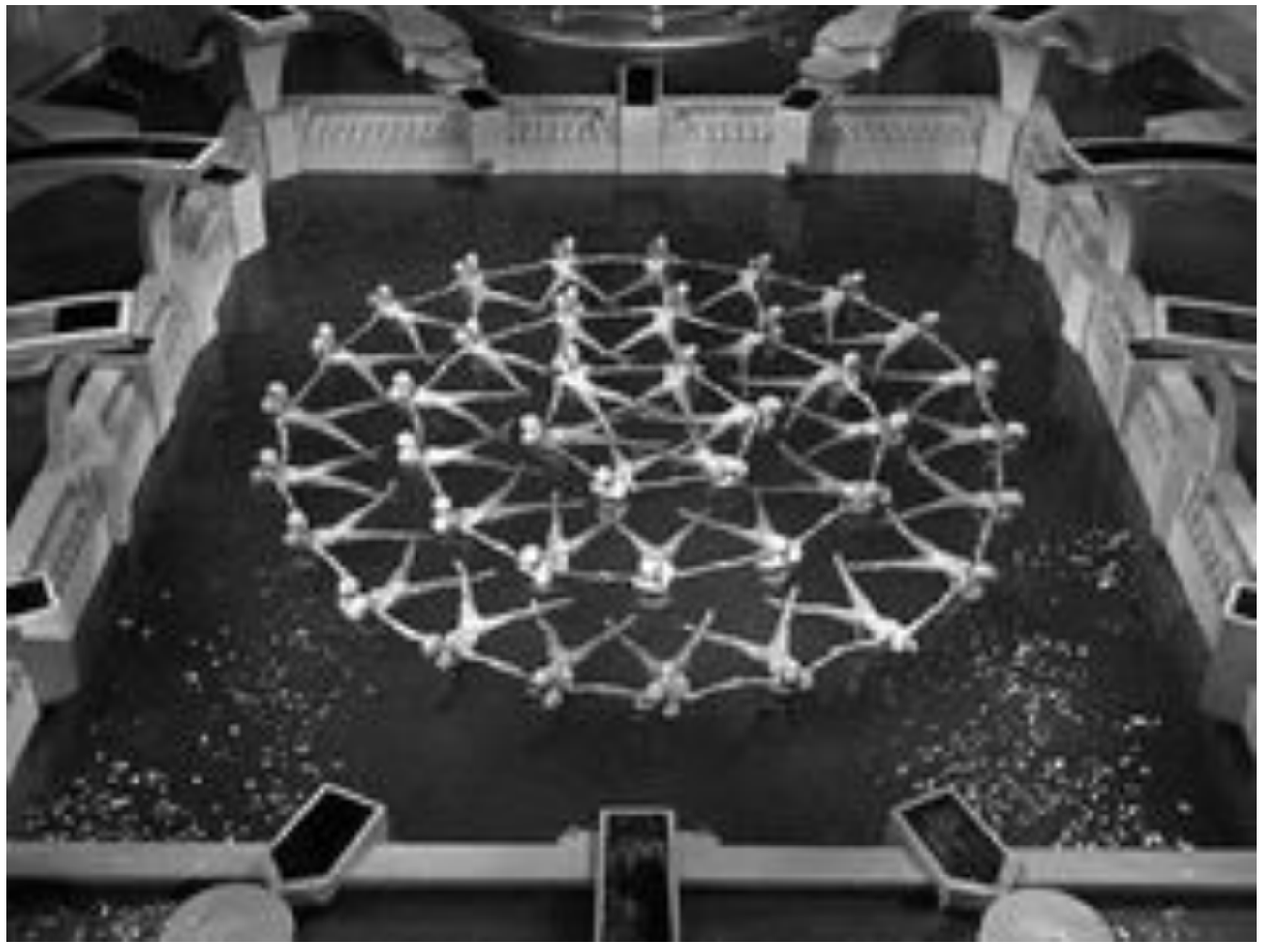
THEN THE MCMC REVOLUTION CHANGED EVERYTHING



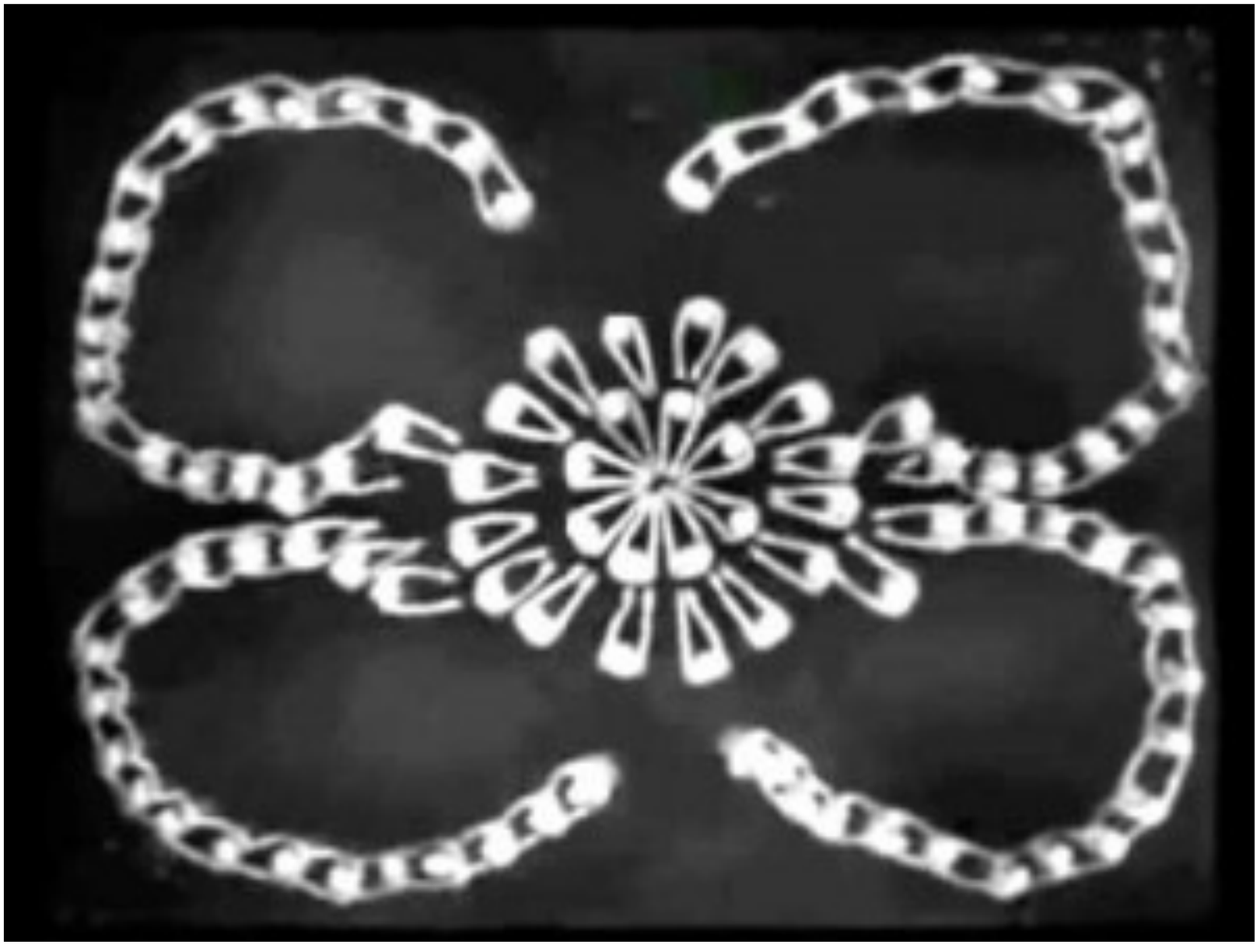
BUGS CAME ALONG AND REDEFINED THE POSSIBLE



METHODS LIKE INLA HELPED US SCALE UP



BUT THEN STAN CAME ALONG



WE ARE TIED DOWN TO A LANGUAGE THAT MAKES UP IN OBSCURITY WHAT IT LACKS IN STYLE

(Tom Stoppard)

A PARTIAL ORDER OF MASSIVE ASSUMPTIONS

Data gathering

Asymptotic regime

Model evaluation criteria

Likelihood

Prior

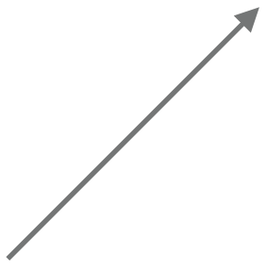
Computation

A PARTIAL ORDER OF MASSIVE ASSUMPTIONS

Data gathering

Asymptotic regime

Model evaluation criteria



Richard!

Likelihood

Prior

Computation

A PARTIAL ORDER OF MASSIVE ASSUMPTIONS

Data gathering

Asymptotic regime

Model evaluation criteria

Richard!

Likelihood

Prior

Maggie!

Computation

A PARTIAL ORDER OF MASSIVE ASSUMPTIONS

Data gathering

Asymptotic regime

Model evaluation criteria

Richard!

Likelihood

Prior

Maggie!

Computation

Sarah!

A PARTIAL ORDER OF MASSIVE ASSUMPTIONS

Data gathering

Asymptotic regime

Model evaluation criteria

Richard!

Likelihood

Prior

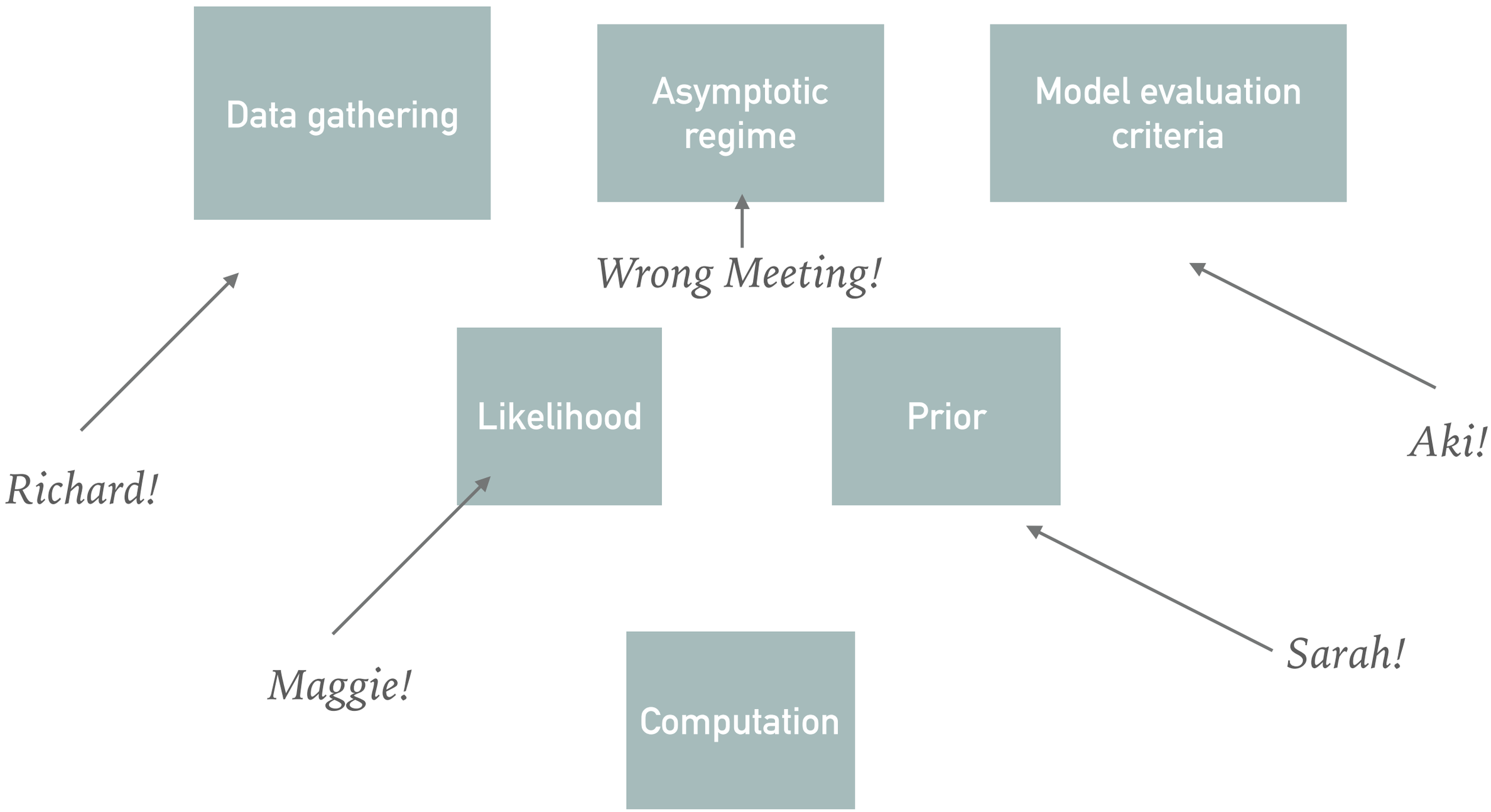
Aki!

Maggie!

Computation

Sarah!

A PARTIAL ORDER OF MASSIVE ASSUMPTIONS



A PARTIAL ORDER OF MASSIVE ASSUMPTIONS

Data gathering

Asymptotic regime

Model evaluation criteria

Richard!

Likelihood

Prior

Aki!

Maggie!

Computation

Sarah!

THE GREAT LIE OF STATISTICS

- Once the models get complex, we don't really know much about how they work.
- We can sometimes say some things about how things will work "eventually"
- But even that is limited to either essentially useless qualitative statements or very simple models

THE GREAT LIE OF STATISTICS

- Once the models get complex, we don't really know much about how they work.
- We can sometimes say some things about how things will work "eventually"
- But even that is limited to either essentially useless qualitative statements or very simple models

THEOREM 2.1. *Suppose that for a sequence* ε_n *with* $\varepsilon_n \to 0$ *and* $n\varepsilon_n^2 \to \infty$, *a constant* $C > 0$ *and sets* $\mathscr{P}_n \subset \mathscr{P}$, *we have*

$$\log D(\varepsilon_n, \mathscr{P}_n, d) \leq n\varepsilon_n^2, \tag{2.2}$$

$$\Pi_n(\mathscr{P} \setminus \mathscr{P}_n) \leq \exp\bigl(-n\varepsilon_n^2(C+4)\bigr), \tag{2.3}$$

$$\Pi_n\Bigl(P\colon -P_0\Bigl(\log\frac{p}{p_0}\Bigr) \leq \varepsilon_n^2,\ P_0\Bigl(\log\frac{p}{p_0}\Bigr)^2 \leq \varepsilon_n^2\Bigr) \geq \exp(-n\varepsilon_n^2 C). \tag{2.4}$$

Then for sufficiently large M, *we have that* $\Pi_n(P\colon d(P, P_0) \geq M\varepsilon_n | X_1, \ldots, X_n) \to 0$ *in* P_0^n*-probability.*

Ghosal, Ghosh, and van der Vaart
Convergence rates of posterior distributions (2000)

GOD IS PRESENT IN THE SWEEPING GESTURES,

BUT THE DEVIL IS IN THE DETAILS

- To do Bayesian statistics is to have long practical experience of pre-asympototic behaviour
- This was especially true with BUGS and JAGS, but is also true with Stan
- Because MCMC methods always converge asymptotically

Data gathering

Asymptotic regime

Model evaluation criteria

Likelihood

Prior

Computation

ETERNITY IS A TERRIBLE THOUGHT. I MEAN, WHERE'S IT GOING TO END?

(Tom Stoppard)

(YOU DRIVE ME) CRAZY

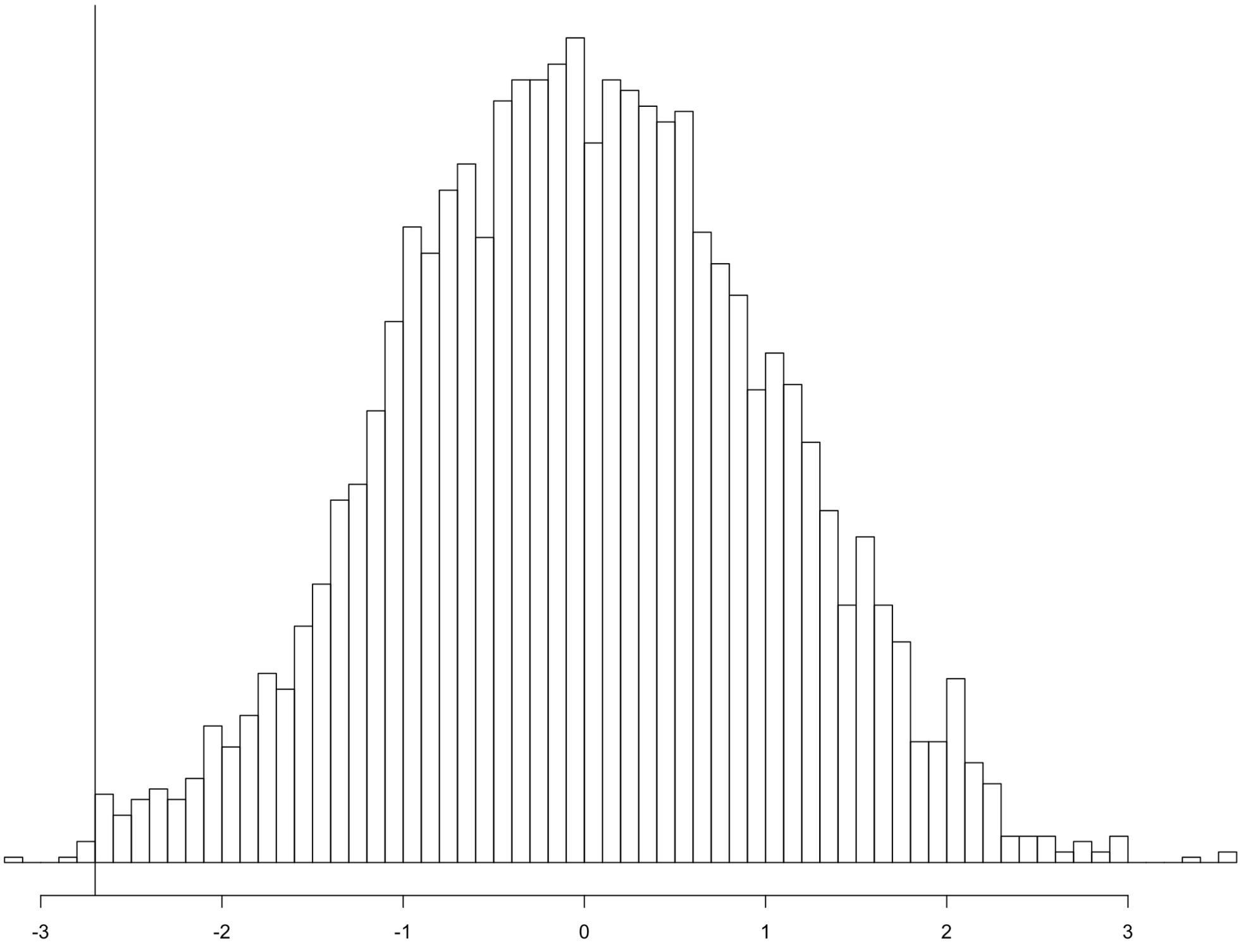
- If we know one thing about MCMC it's that it usually hasn't converged
- Stan implements a dynamic Hamiltonian Monte Carlo album wiht multinomial sampling of dynamic length trajectories, generalized termination criterion, and improved adaptation of the Euclidean metric.
- It is ok.
- Asymptotically, it probably works.
- But I want more.

HOW CAN WE TELL IF AN ALGORITHM ACTUALLY WORKS?

➤ Idea: Run the algorithm on simulated data.

1. Pick a parameter value θ_0
2. Generate data from $p(\mathbf{y} \mid \boldsymbol{\theta}_0)$
3. Fit model to data
4. Compare the posterior to the known true value

OKAY! IS THIS RIGHT?

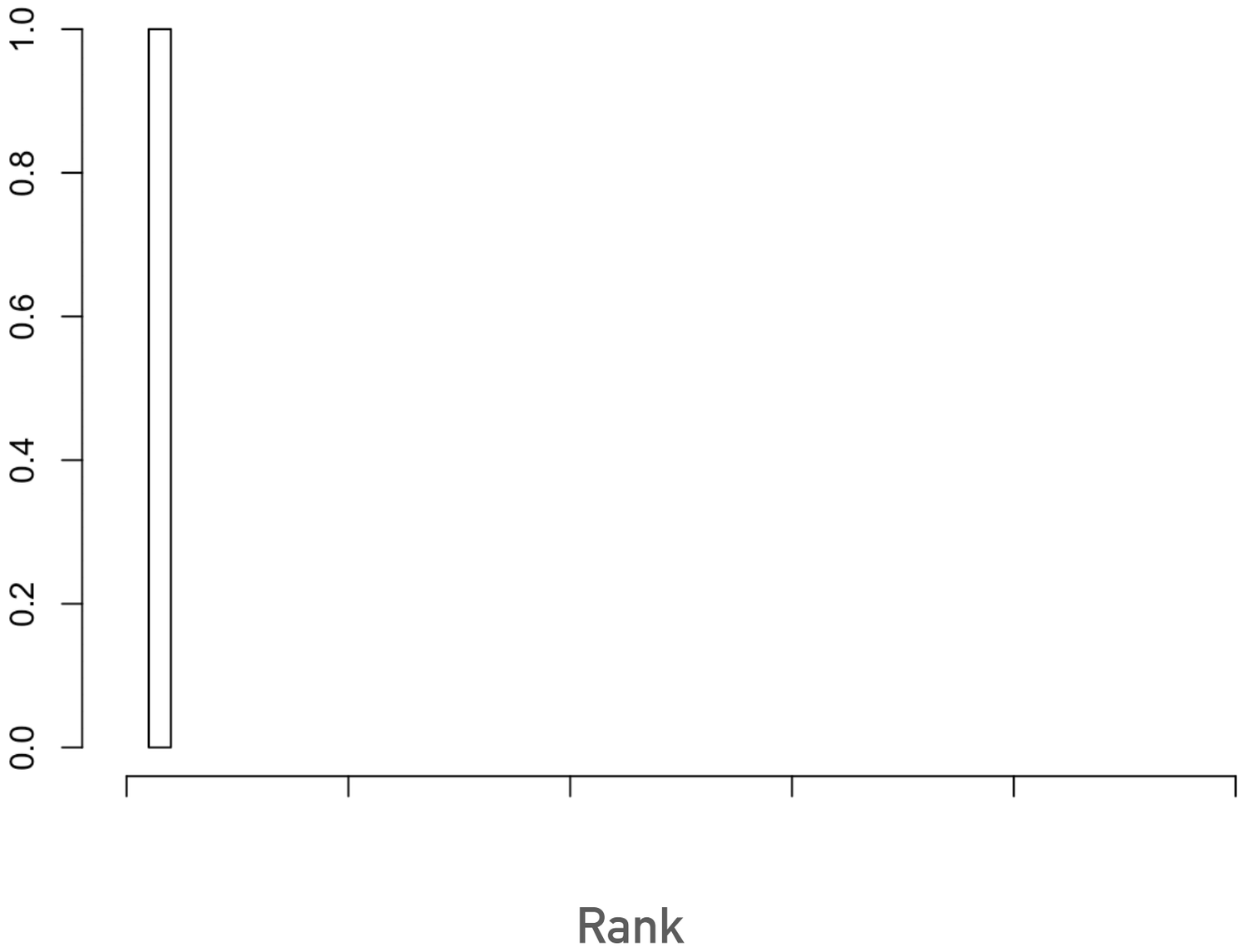


HOW DO YOU TELL IF IT FITS?

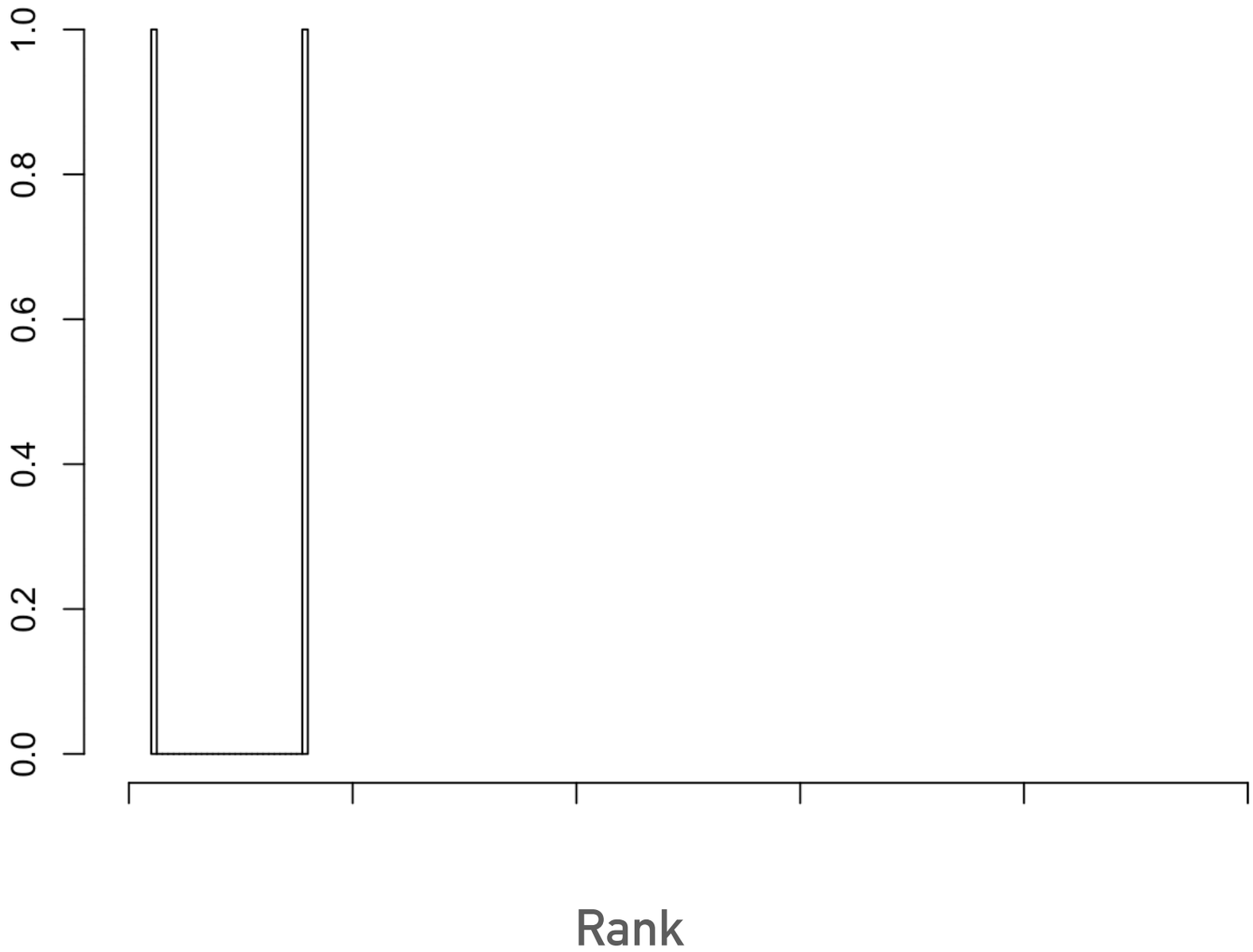
- We have a true value
- We have a bag of (approximately independent) posterior samples
- We can just look at where the true value lies in the bag of samples
- We look at the **rank** of the true value within the sample
- What happens when we do it a lot of times?

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, Andrew Gelman (2018)
Validating Bayesian Inference Algorithms with Simulation-Based Calibration
arXiv preprint: https://arxiv.org/abs/1804.06788

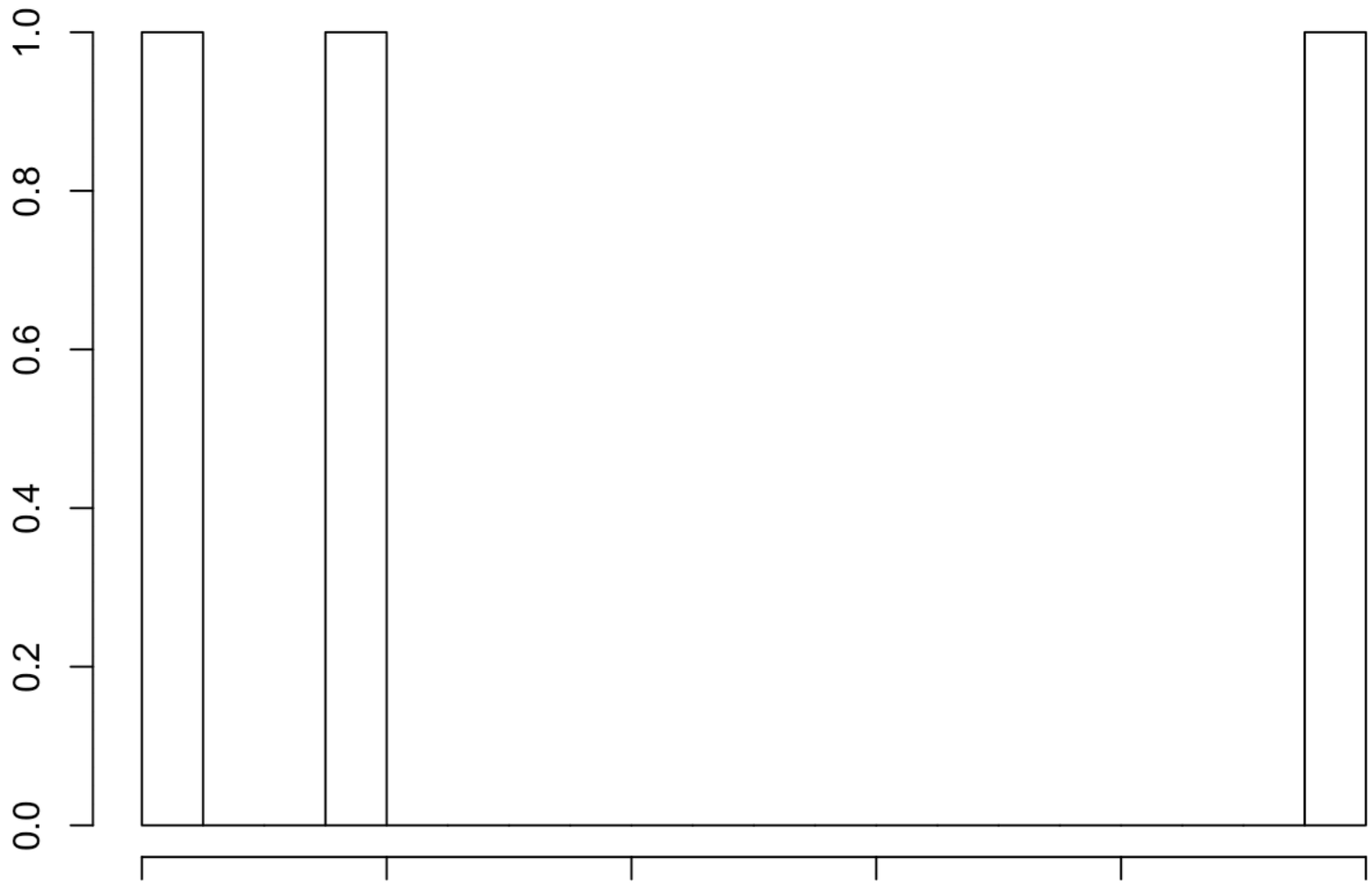
SINGLE RECOVERY



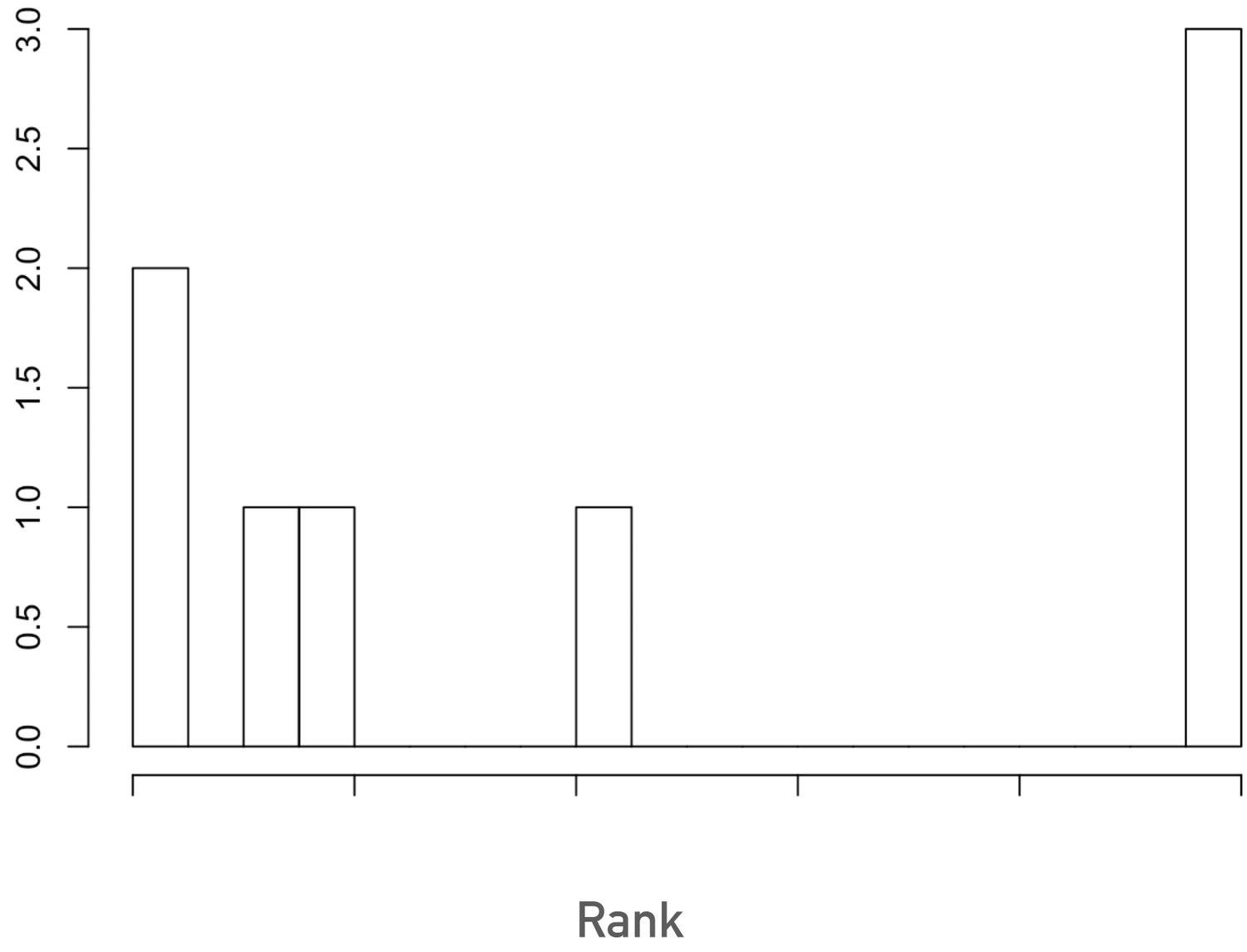
MULTIPLE RECOVERY



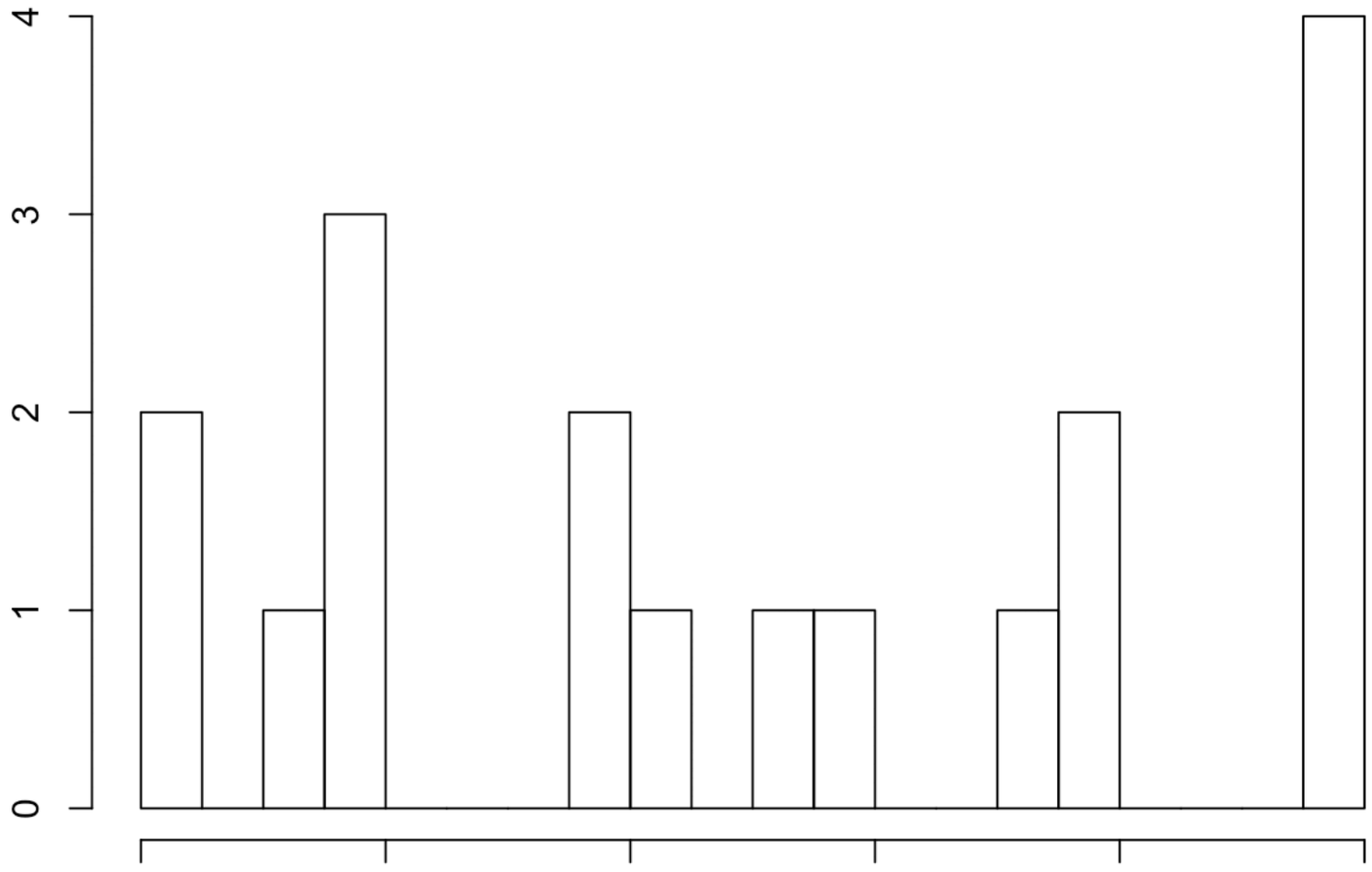
MULTIPLE RECOVERY



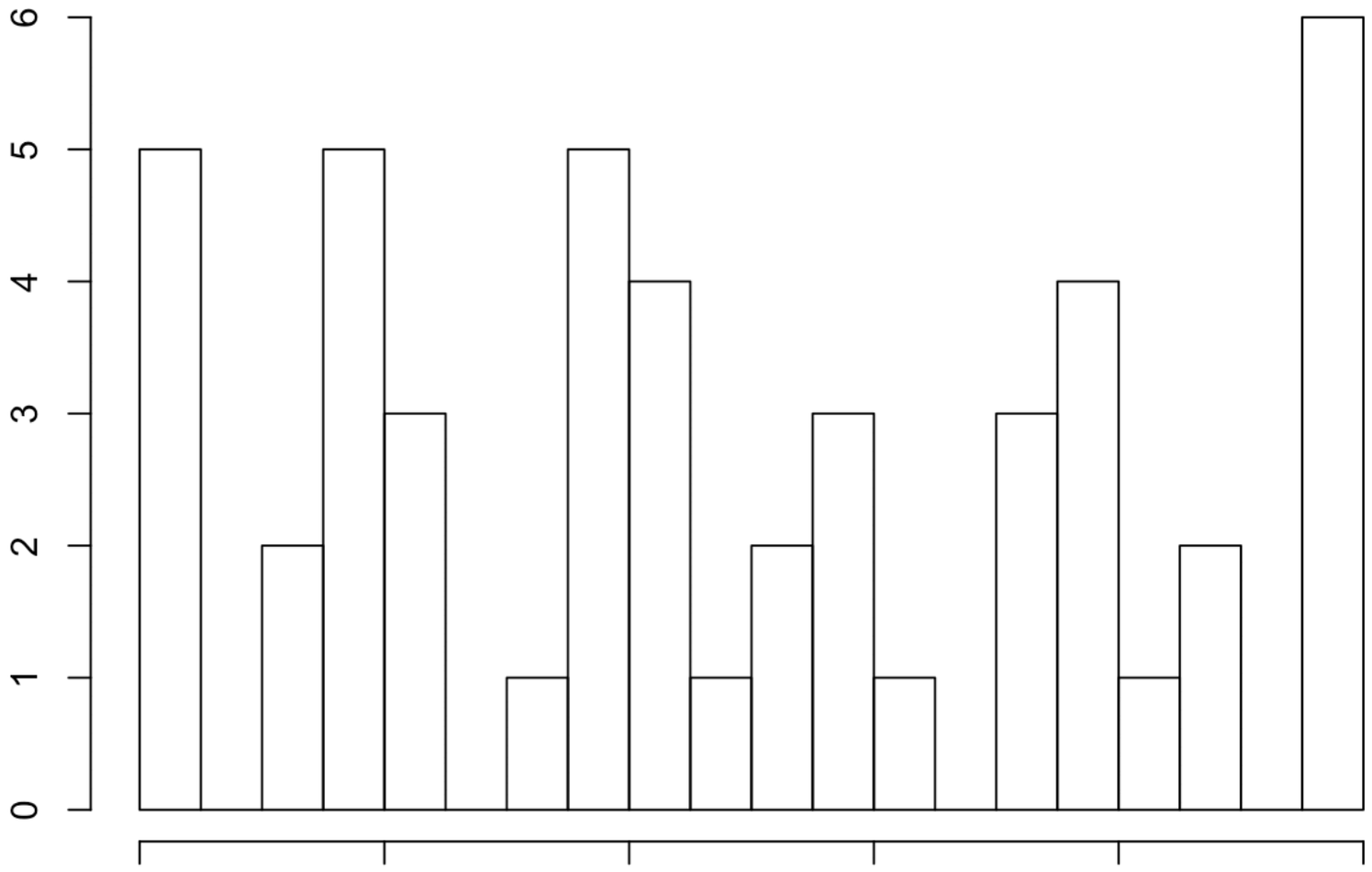
MULTIPLE RECOVERY



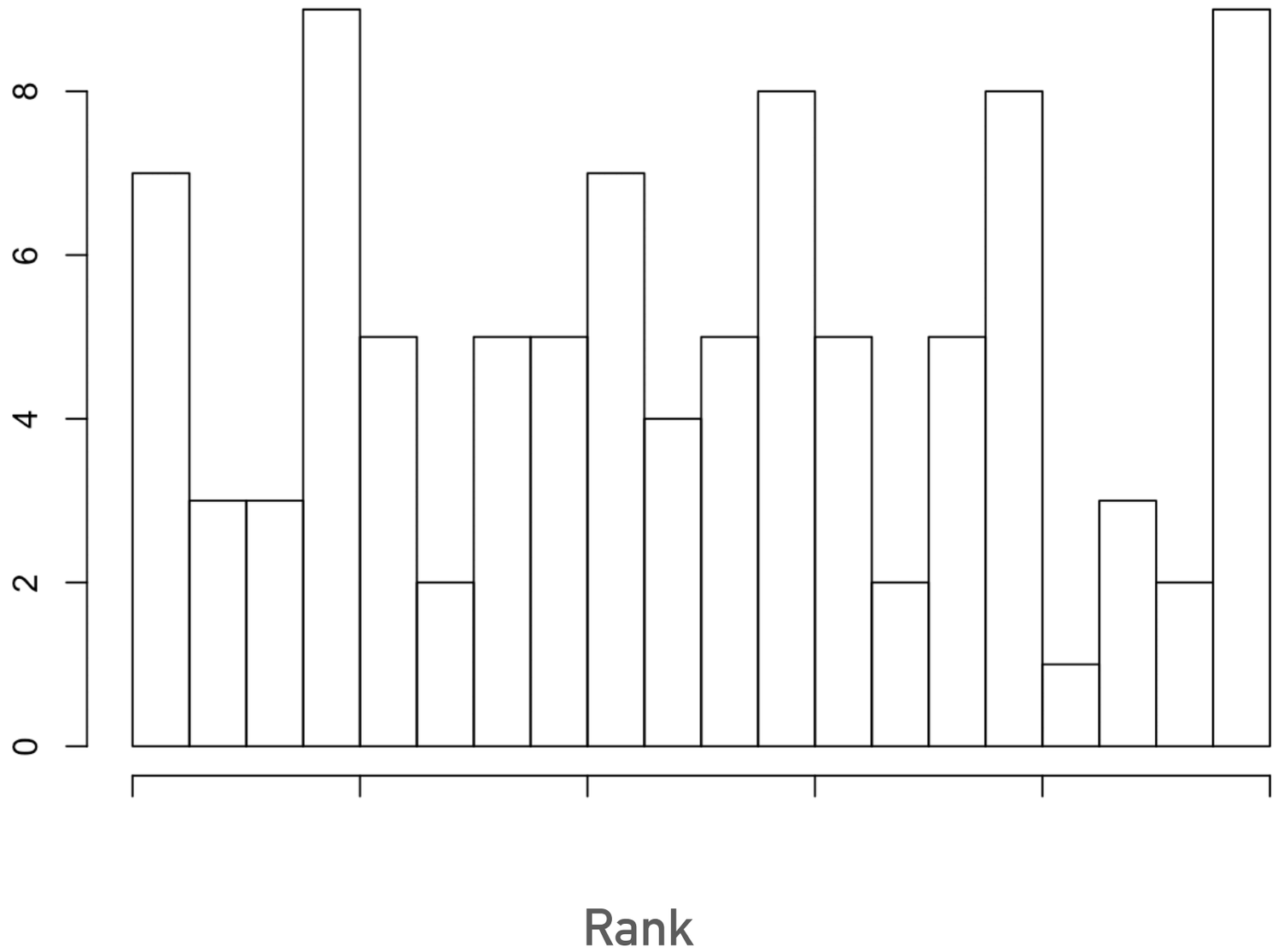
MULTIPLE RECOVERY



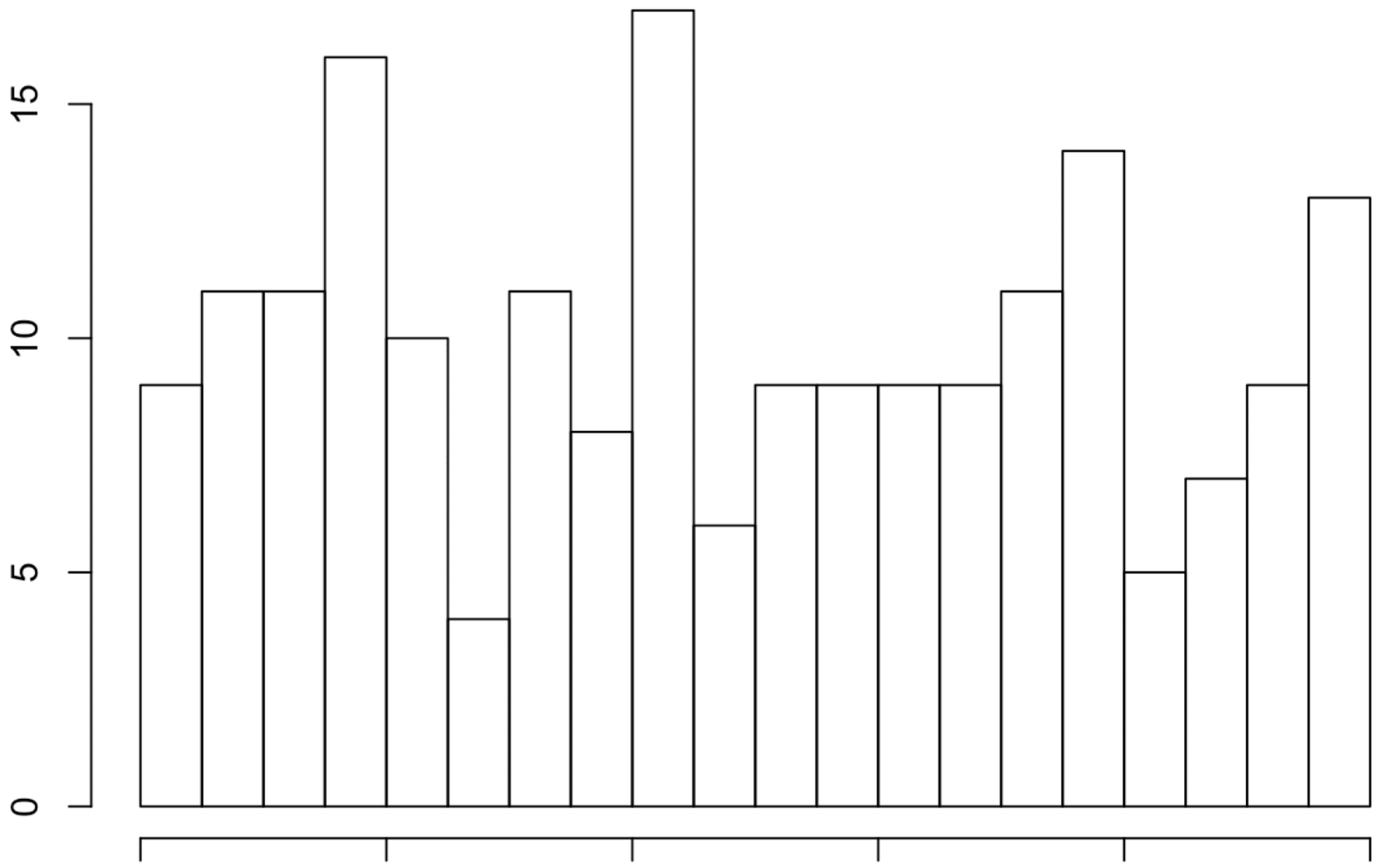
MULTIPLE RECOVERY



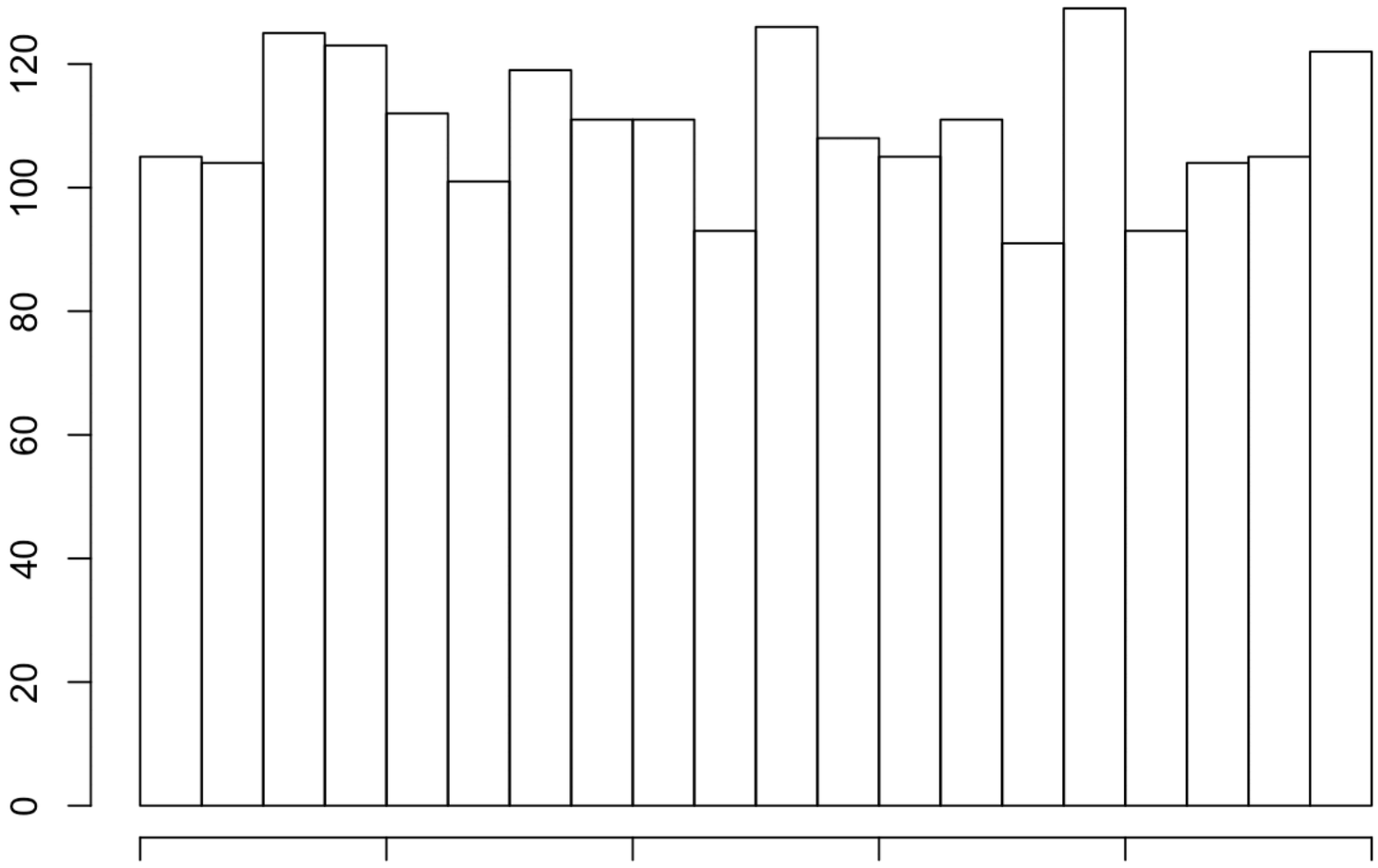
MULTIPLE RECOVERY



MULTIPLE RECOVERY



MULTIPLE RECOVERY

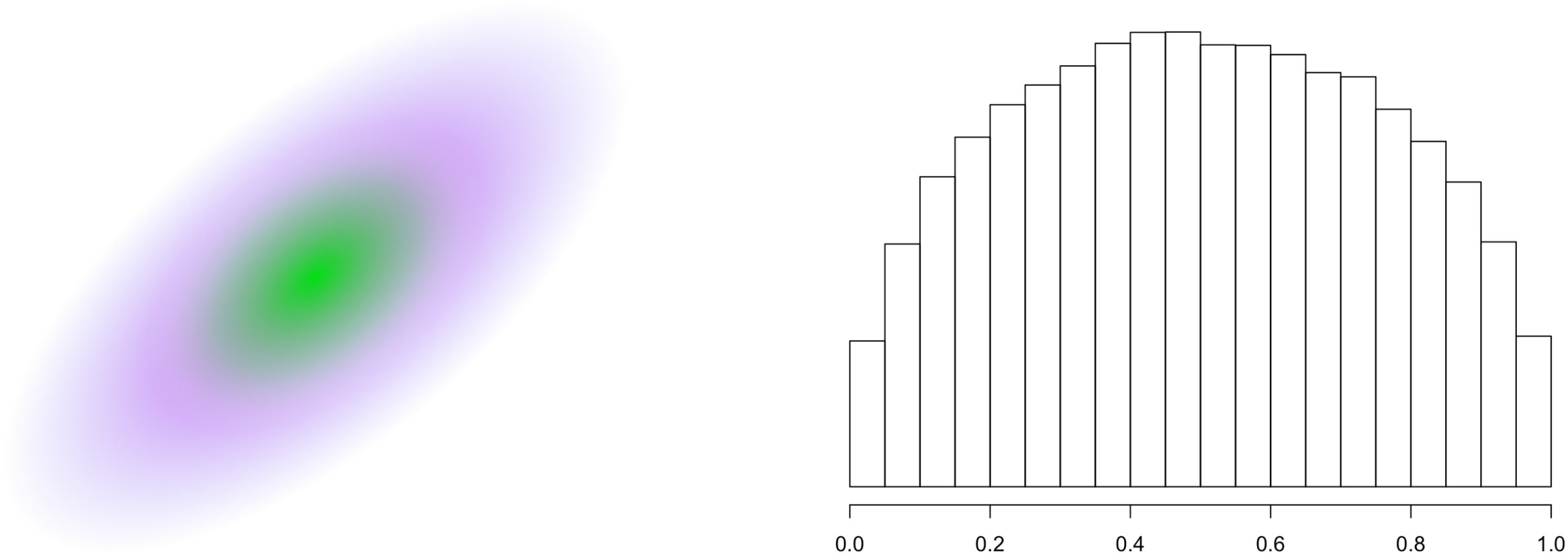


x

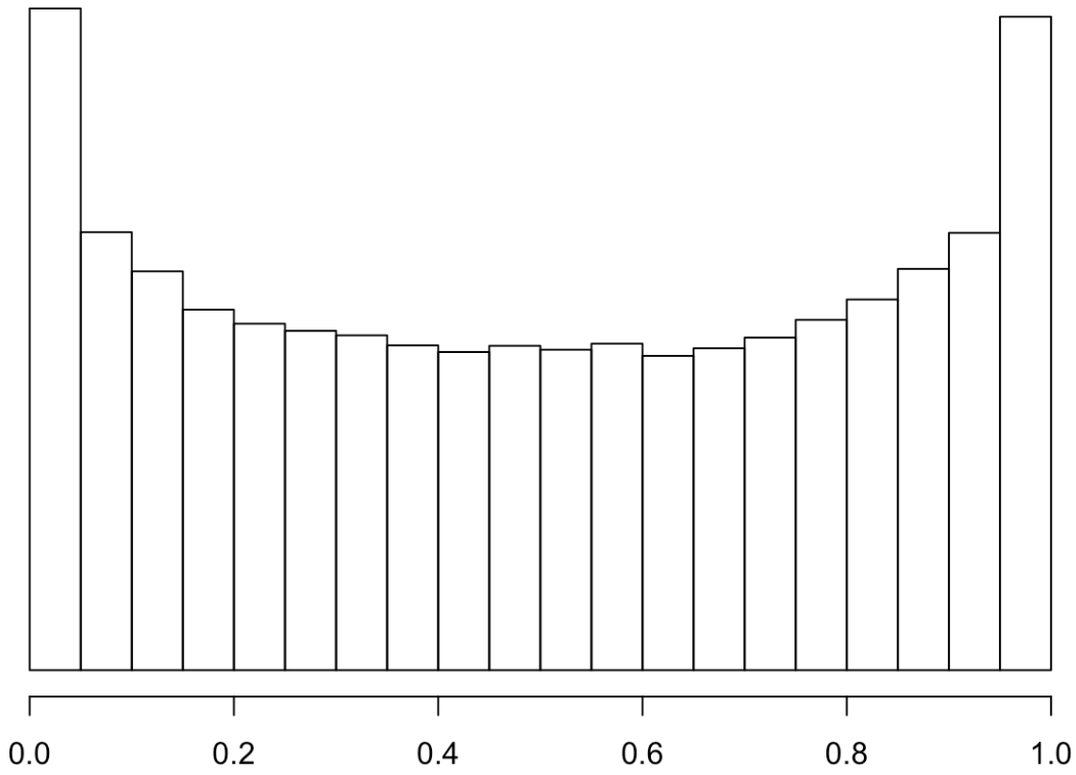
CORRECT

COMPUTED

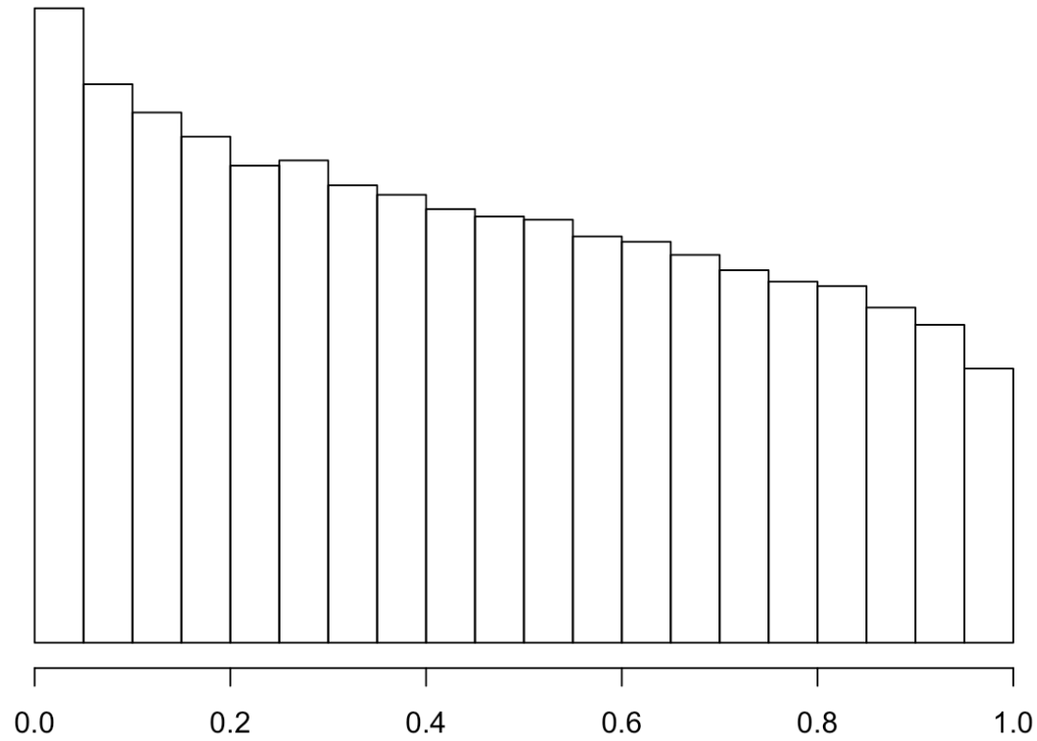
COMPUTED POSTERIOR IS TOO DIFFUSE



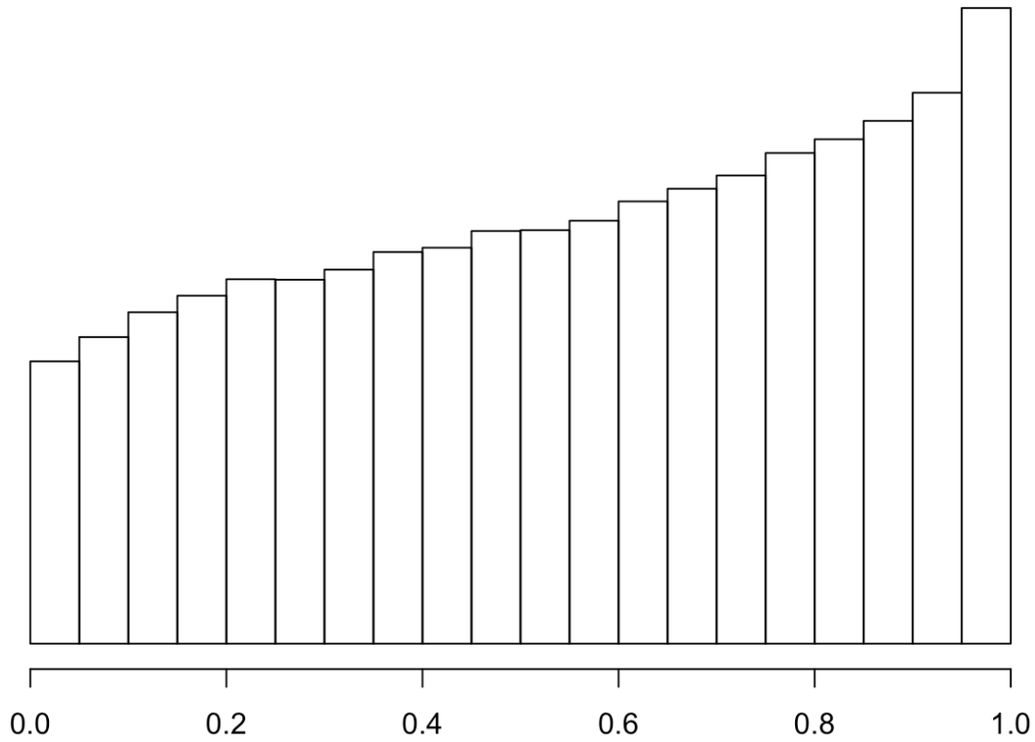
COMPUTED POSTERIOR IS TOO NARROW



COMPUTED POSTERIOR IS BIASED TOWARDS LARGER VALUES



COMPUTED POSTERIOR IS BIASED TOWARDS SMALLER VALUES



BUT THAT CAN BE QUITE EXPENSIVE

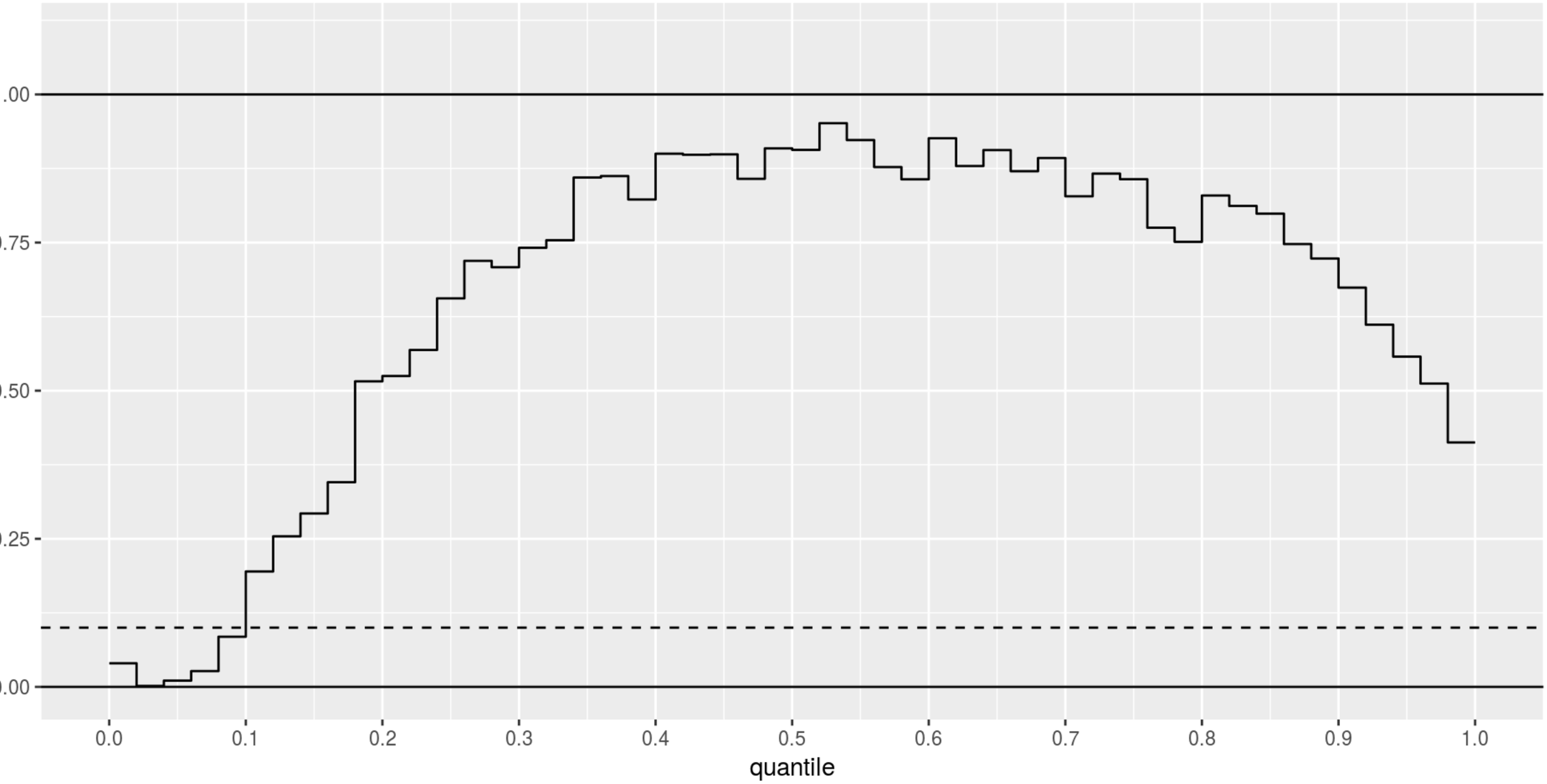
- We call this method **Simulation-Based Calibration (SBC)**
- **We** can run this on a cluster and it's not too bad, but it is a problem
- What about other methods of assessing the behaviour of a Markov Chain?
- Well there isn't much. Famously Gelman and Rubin suggested looking at the R-hat diagnostic
- But it's a little bit more complicated…

A QUICK PREVIEW

- R-hat is a little bit awkward:
  - If the marginal posterior has heavy tails, it is not sensitive to much
  - It is not sensitive to one chain having smaller variance than the others
- One way to think about this is that R-hat tells us about the efficiency of the bulk of the chain (it's basically measuring if the centre is in the same place)
- But local behaviour is important

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, Andrew Gelman (2018)
Validating Bayesian Inference Algorithms with Simulation-Based Calibration
arXiv preprint: https://arxiv.org/abs/1804.06788

A COMMON CASE (LOCAL EFFICIENCY OF THE CHAIN)



Local relative efficiency for the group-level standard deviation Centred parameterization for 8 schools. (Figure by Aki Vehtari)

YOU OUGHTA KNOW

- There is a lot more research to be done on understanding local efficiency of MCMC methods
- But when it comes down to it, we still need to understand our models
- And asymptotics can come back to hurt us there too

Data gathering

Asymptotic regime

Model evaluation criteria

Prior

Computation

EVERYTHING HAS TO BE TAKEN ON TRUST; TRUTH IS ONLY THAT WHICH IS TAKEN TO BE TRUE.

(Tom Stoppard)

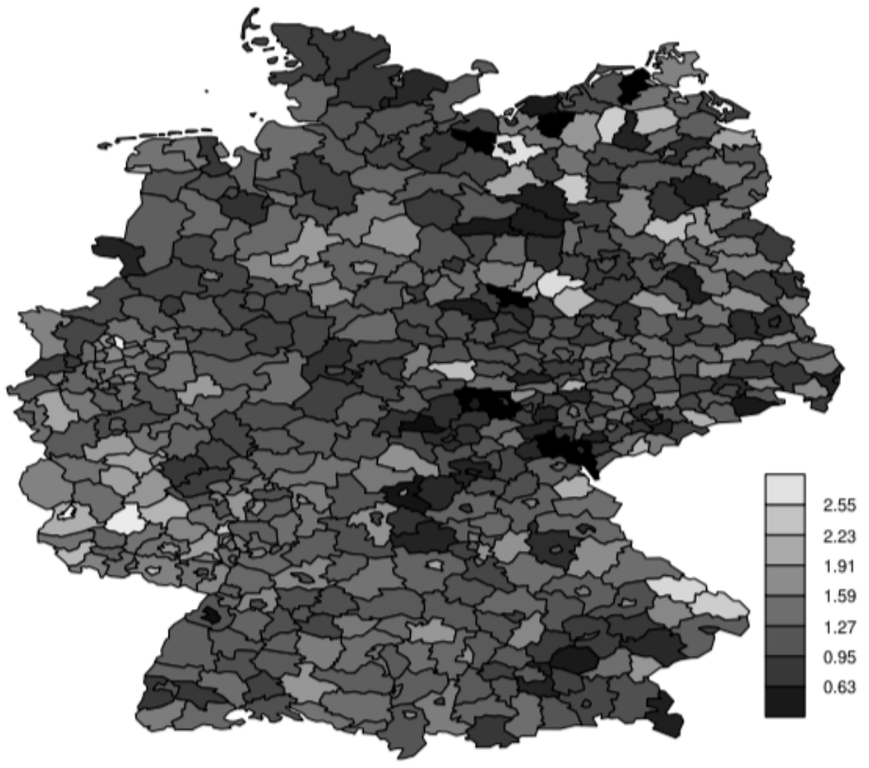
GAUSSIAN PROCESSES

- A Gaussian process is a mulivariate Gaussian $\mathbf{f} \sim N(\mathbf{0}, \mathbf{\Sigma})$
- Here $\Sigma_{ij} = c(x_i, x_j)$ is defined using a covariance function c(.,.) which usually depends on some hyperparameters
- A common example of a covariance function is

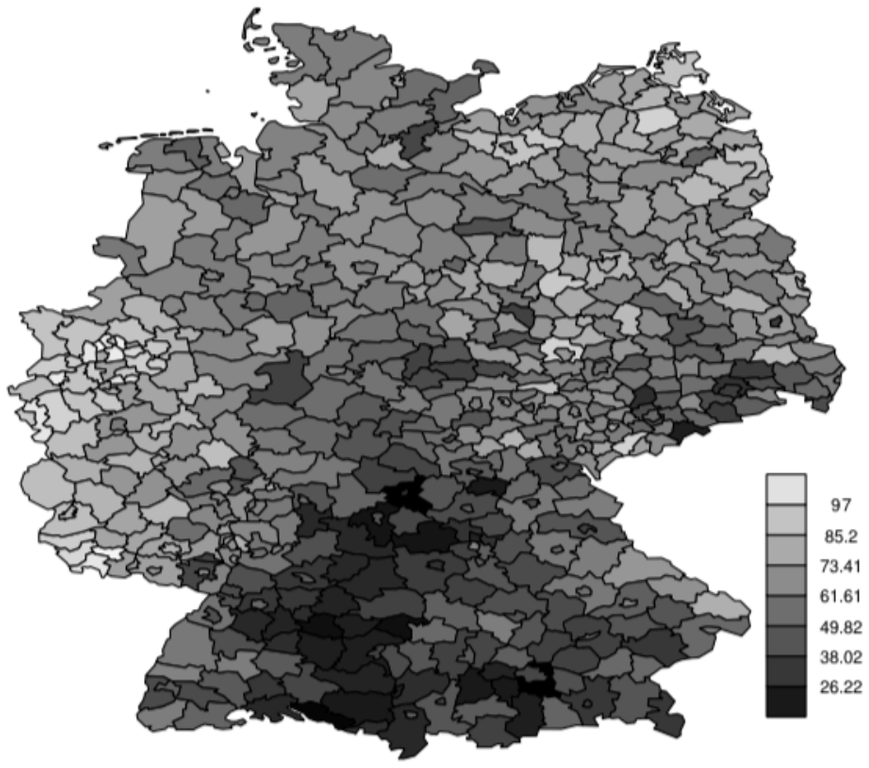
$$c(x_1, x_2) = \alpha^2 \exp\left(-\frac{\|x_1 - x_2\|^2}{2\ell^2}\right)$$

- There are two parameters here:
  - α is the marginal variance of f_i
  - ℓ is the length scale

YES BUT WHY DO I CARE?



Incidence of larynx cancer



Smoking rates

How would we model risk?

A BASIC MODEL FOR ESTIMATING DISEASE COUNTS

$$Counts_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \log(Expected\ Counts) + u_i + v_i + f(x_i)$$

A BASIC MODEL FOR ESTIMATING DISEASE COUNTS

$$Counts_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \log(Expected\ Counts) + u_i + v_i + f(x_i)$$

Relative risk!

A BASIC MODEL FOR ESTIMATING DISEASE COUNTS

$$Counts_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \log(Expected\ Counts) + u_i + v_i + f(x_i)$$

Relative risk!

Overdispersion

A BASIC MODEL FOR ESTIMATING DISEASE COUNTS

$$Counts_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \log(Expected\ Counts) + u_i + v_i + f(x_i)$$

Relative risk!

Overdispersion

Spatial correlation

A BASIC MODEL FOR ESTIMATING DISEASE COUNTS

$$Counts_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \log(Expected\ Counts) + u_i + v_i + f(x_i)$$

Relative risk!

Overdispersion

Spatial correlation

Covariate GP

I don't know how to deal with all of this stuff together

LET'S IGNORE ALL OF THE INTERESTING BITS

- I'm a mathematician by training and disposition, so I'm going to ignore all of the interesting bits of that model and just focus on the maths bit!
- So let's just focus on GP regression
- Because I **know** somethings about GP regression
- Namely, I know that I can consistently estimate all of the parameters, so the MCMC should be fine.

$$y_i = f(x_i) + \epsilon_i$$

iid Gaussian

EVERYTHING IS PROBABLY NOT GOING TO BE OK

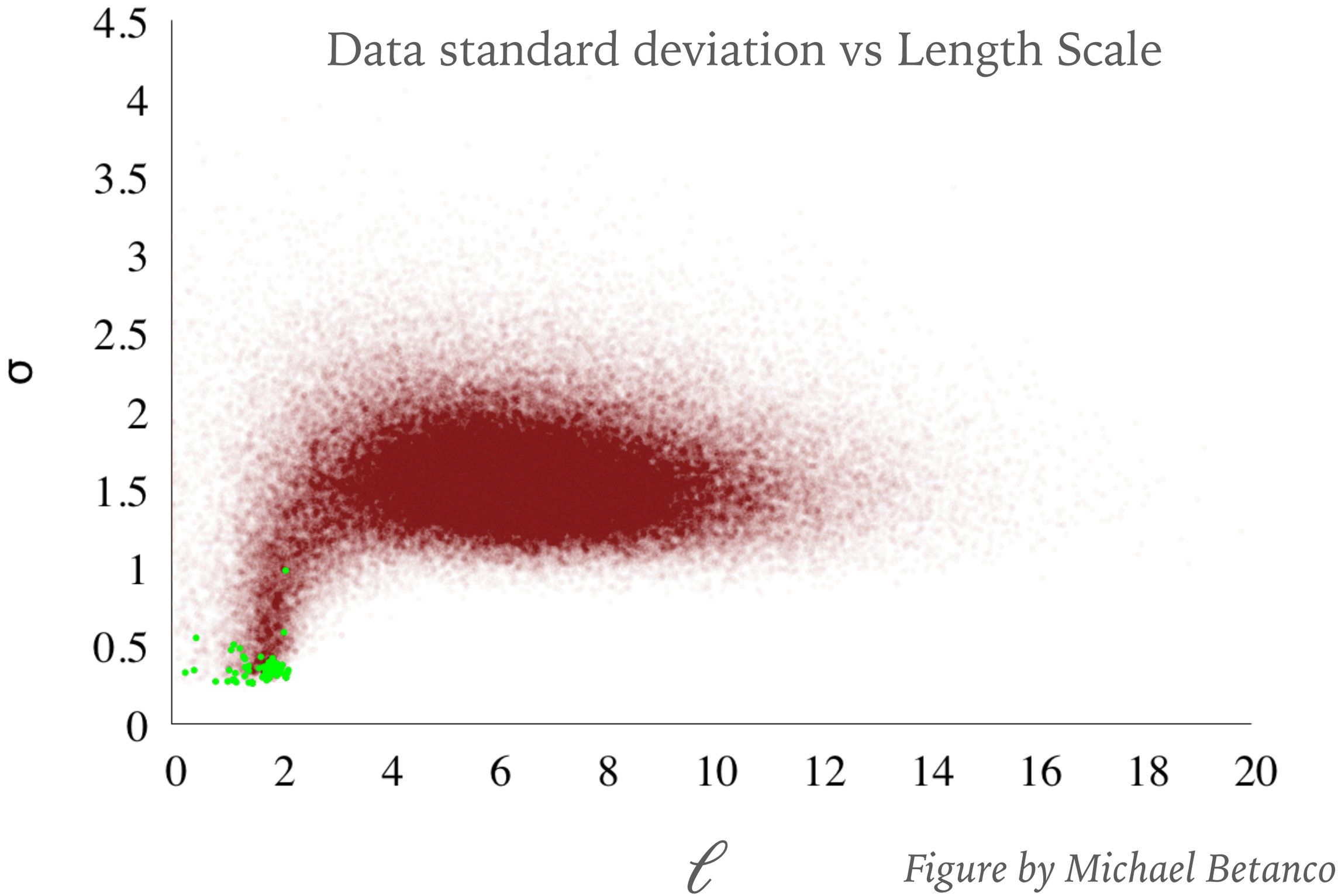


Figure by Michael Betancourt

EVERYTHING IS PROBABLY NOT GOING TO BE OK

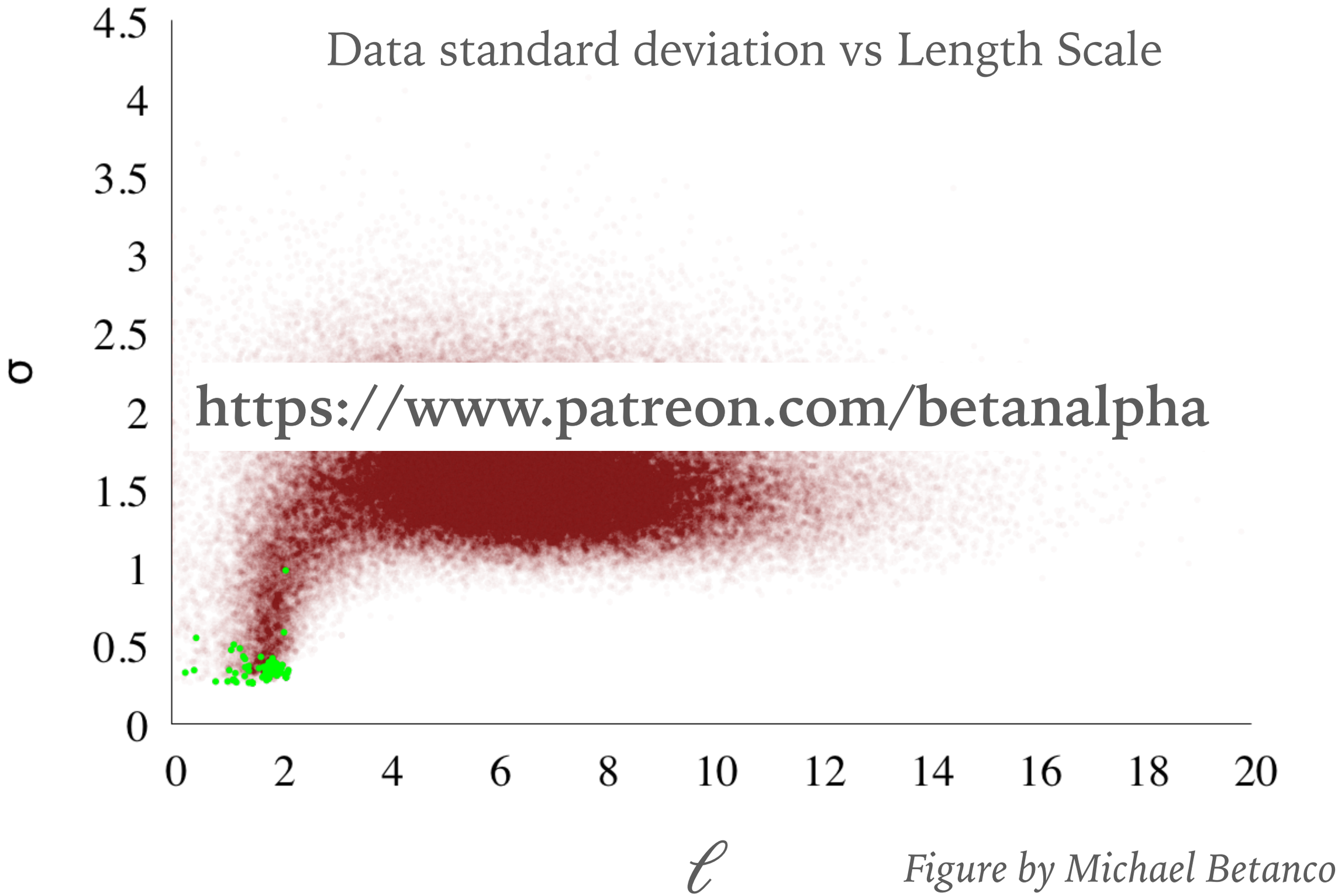


Figure by Michael Betancourt

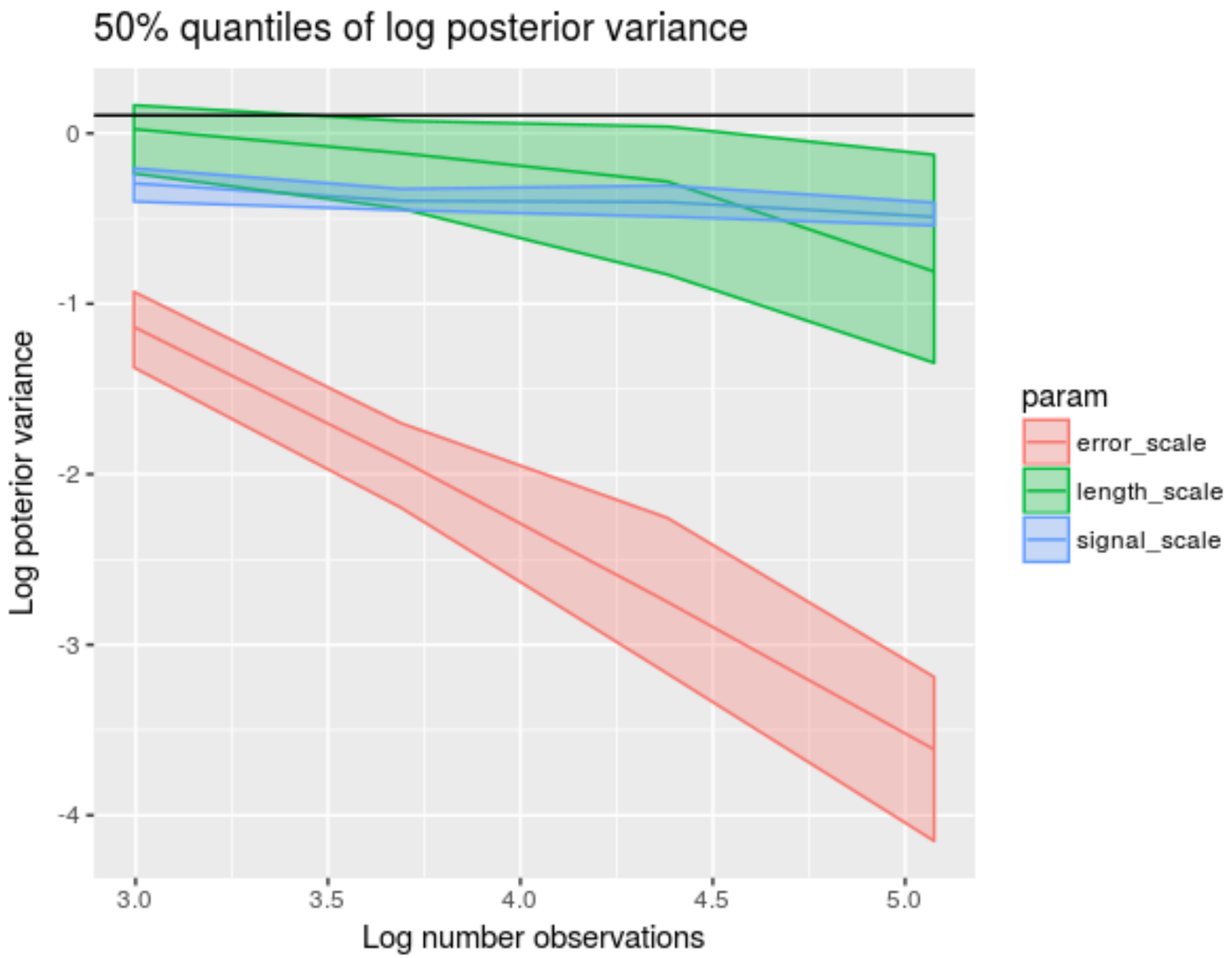
AND I SAID HEYYYYY YEAH YEAH YEAH

- The posterior for the parameters have **strong** correlations
- This is even true for the largest GPs we can do in Stan
- But we **know** that asymptotically this isn't true…
- We just never see the good behaviour :(

AND IT ONLY GETS WORSE

- GP regression is fairly simple as a model. There are only three parameters
- And there is a lot of data
- So they should all be reasonably well constrained by the data

MORE THAN JUST BOOK LEARNING



Plot: Rob Trangucci

ME AGAINST THE MUSIC

- The problem with all of this is that even the simplest case behaves counter to our expectations.
- But we routinely want to use GPs as **components** of real models
- How can we tell if it works?
- A lot more work needs to be done on prior specification and model checking for GPs

Data gathering

Asymptotic regime

Model evaluation criteria

Likelihood

Computation

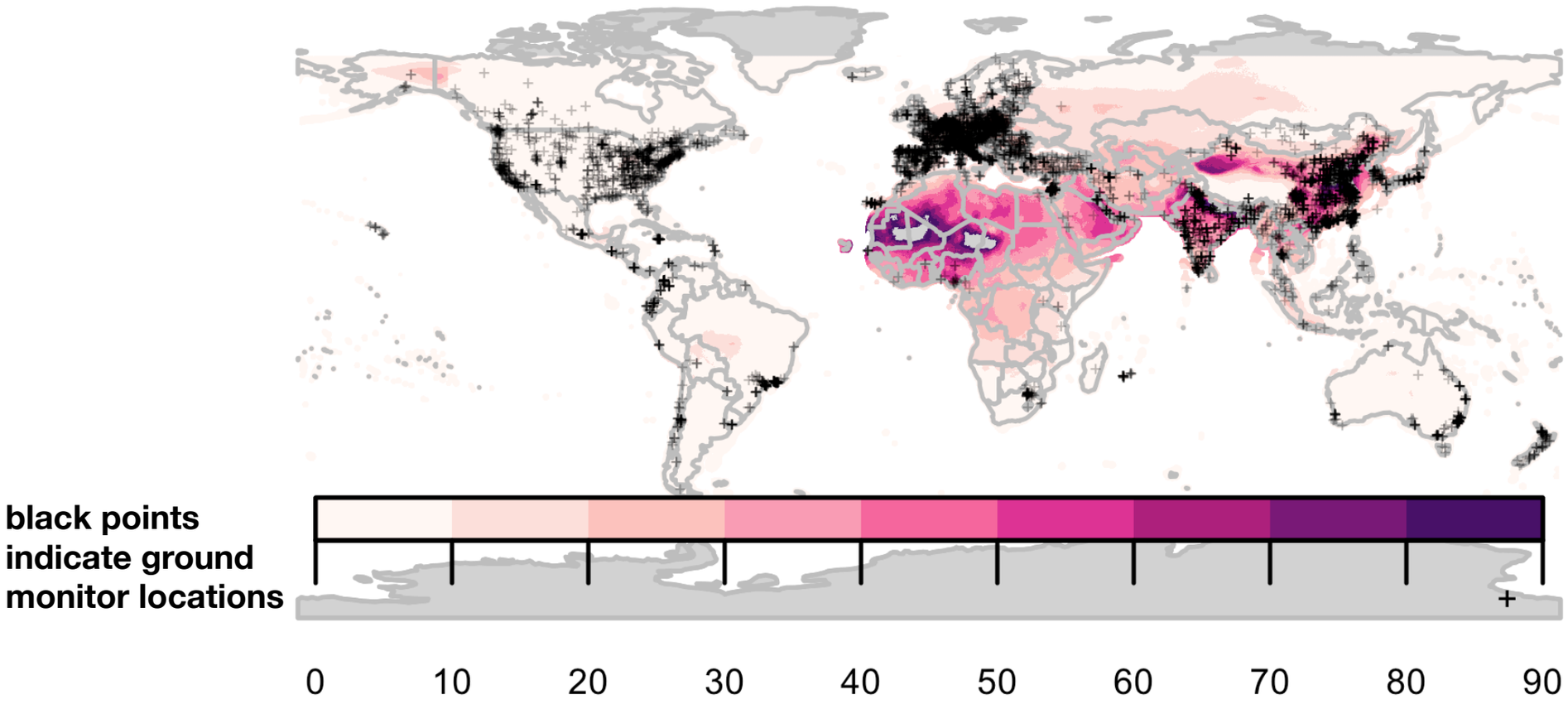
THE COLOURS RED, BLUE AND GREEN ARE REAL. THE COLOUR YELLOW IS A MYSTICAL EXPERIENCE SHARED BY EVERYBODY.

(Tom Stoppard)

WHEN KYLIE SAID "BREATHE" THIS WASN'T WHAT SHE WANTED

Goal Estimate global PM2.5 concentration

Problem Most data from noisy satellite measurements (ground monitor network provides sparse, heterogeneous coverage)



Satellite estimates of PM2.5 and ground monitor locations

ARIANISM WAS A HERESY FOR A REASON

- Many are taught that the likelihood is the fundamental building block of a Bayesian model and the prior is a secondary object
- This is a very limiting view.
- In reality, we build a **joint distribution** for the data and the likelihood
- People who don't do this (like people who use reference priors) are making some heavy assumptions
- (and, in this analogy, are heretics but don't worry so much about that)

Gelman, A., Simpson, D., and Betancourt, M. (2017).
The prior can often only be understood in the context of the likelihood.
arXiv preprint: arxiv.org/abs/1708.07487

THE MAJESTY OF GENERATIVE MODELS

- If we disallow improper priors, then Bayesian modelling is generative.
- In particular, we have a simple way to simulate from $p(y)$:
  - Simulate $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$
  - Simulate $\mathbf{y}^* \sim p(\mathbf{y} \mid \boldsymbol{\theta}^*)$
  - (Repeat for each sample)

PRIOR PREDICTIVE CHECKING

What do vague/non-informative priors imply about the data our model can generate?

PRIOR PREDICTIVE CHECKING

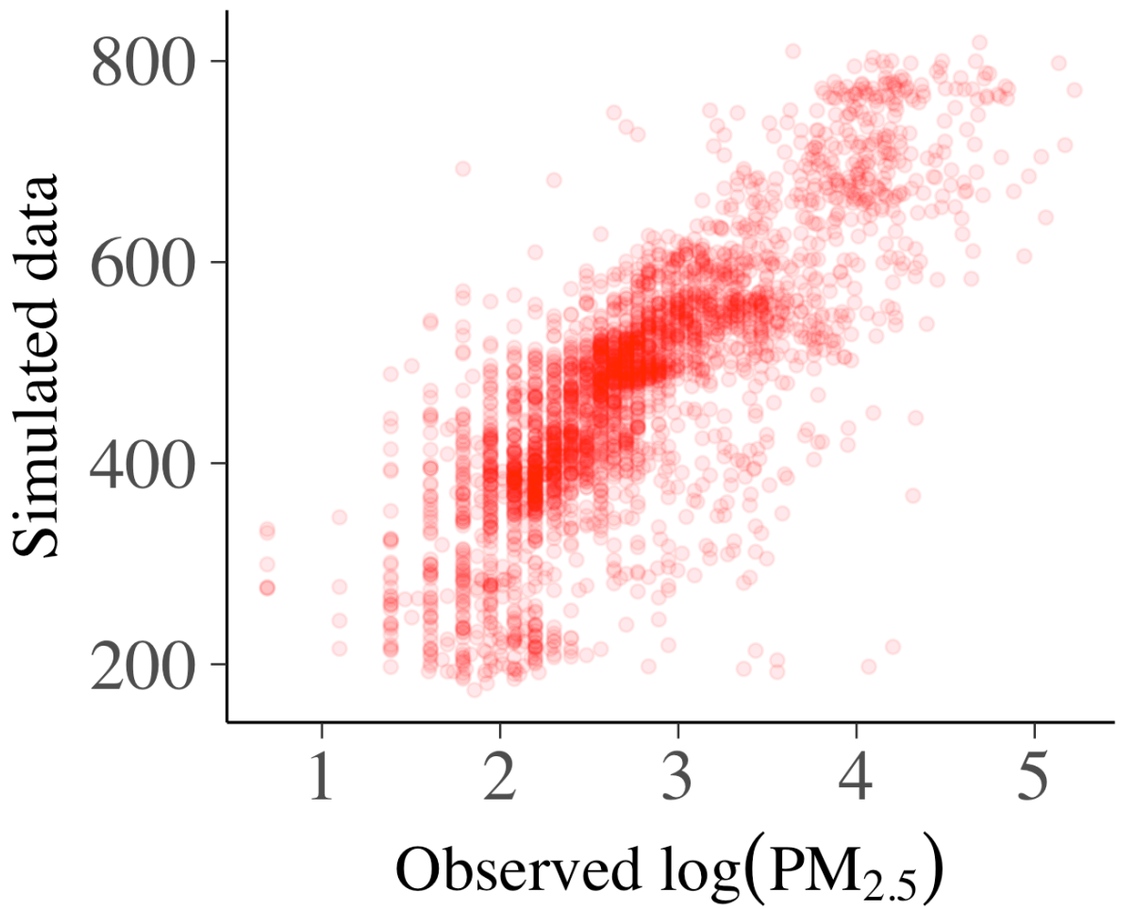
What do vague/non-informative priors imply about the data our model can generate?

$$\alpha_0 \sim N(0, 100)$$
$$\beta_0 \sim N(0, 100)$$
$$\tau_\alpha^2 \sim \text{InvGamma}(1, 100)$$
$$\tau_\beta^2 \sim \text{InvGamma}(1, 100)$$

PRIOR PREDICTIVE CHECKING

What do vague/non-informative priors imply about the data our model can generate?

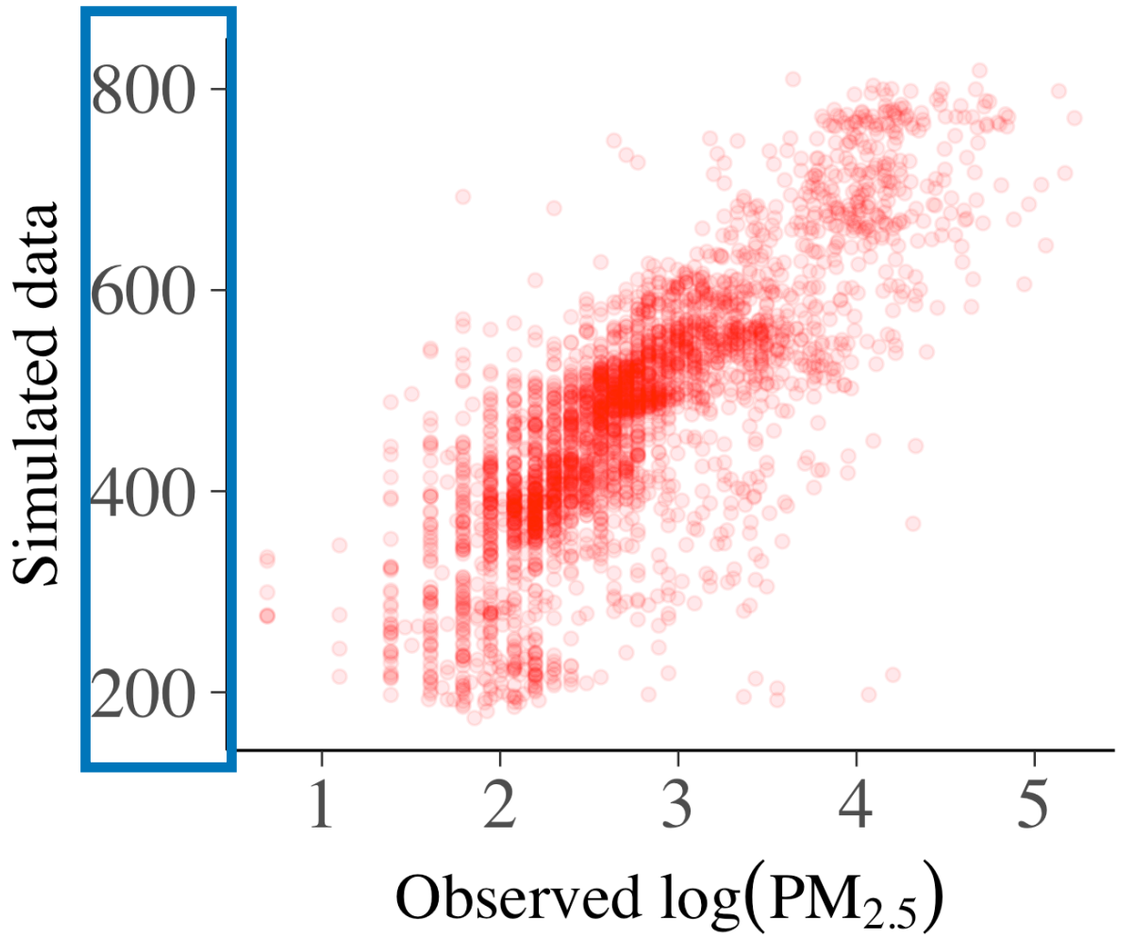
$$\alpha_0 \sim N(0, 100)$$
$$\beta_0 \sim N(0, 100)$$
$$\tau_\alpha^2 \sim \text{InvGamma}(1, 100)$$
$$\tau_\beta^2 \sim \text{InvGamma}(1, 100)$$



PRIOR PREDICTIVE CHECKING

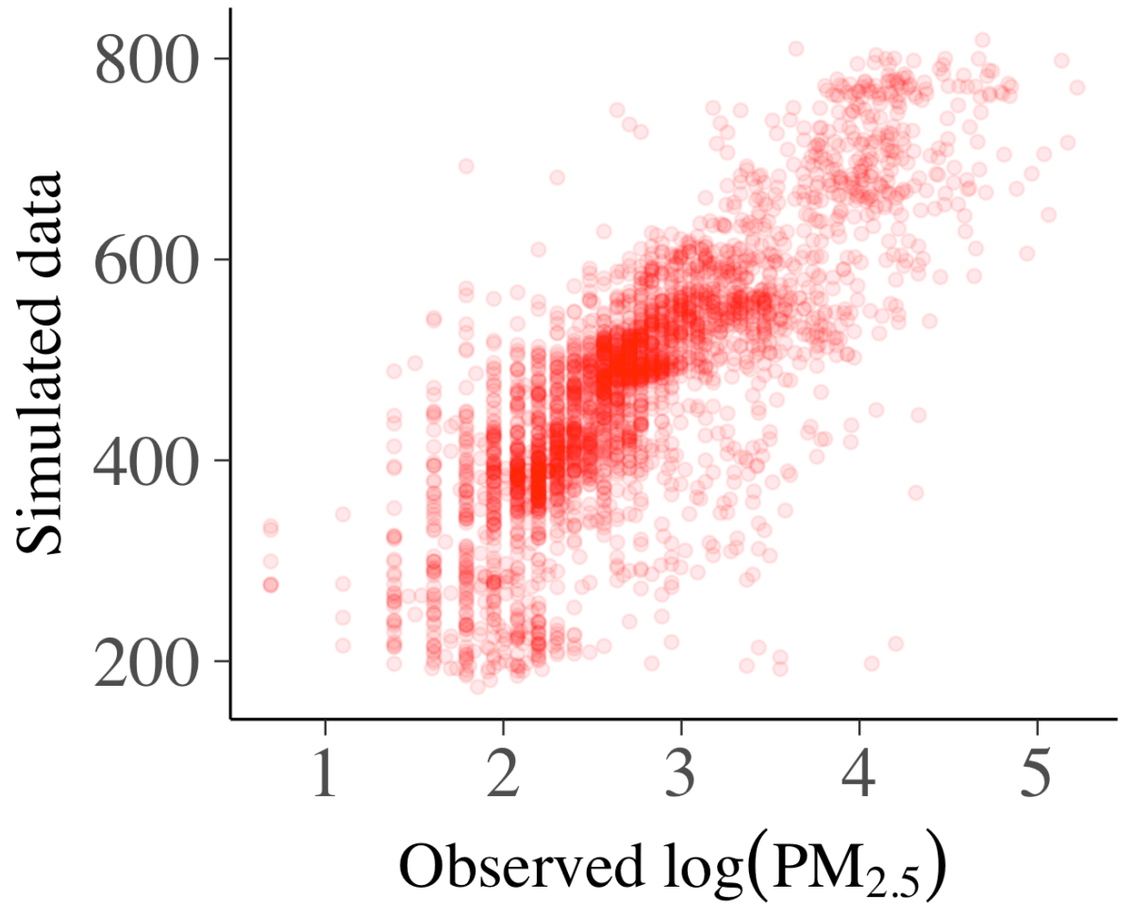
What do vague/non-informative priors imply about the data our model can generate?

$$\alpha_0 \sim N(0, 100)$$
$$\beta_0 \sim N(0, 100)$$
$$\tau_\alpha^2 \sim \text{InvGamma}(1, 100)$$
$$\tau_\beta^2 \sim \text{InvGamma}(1, 100)$$



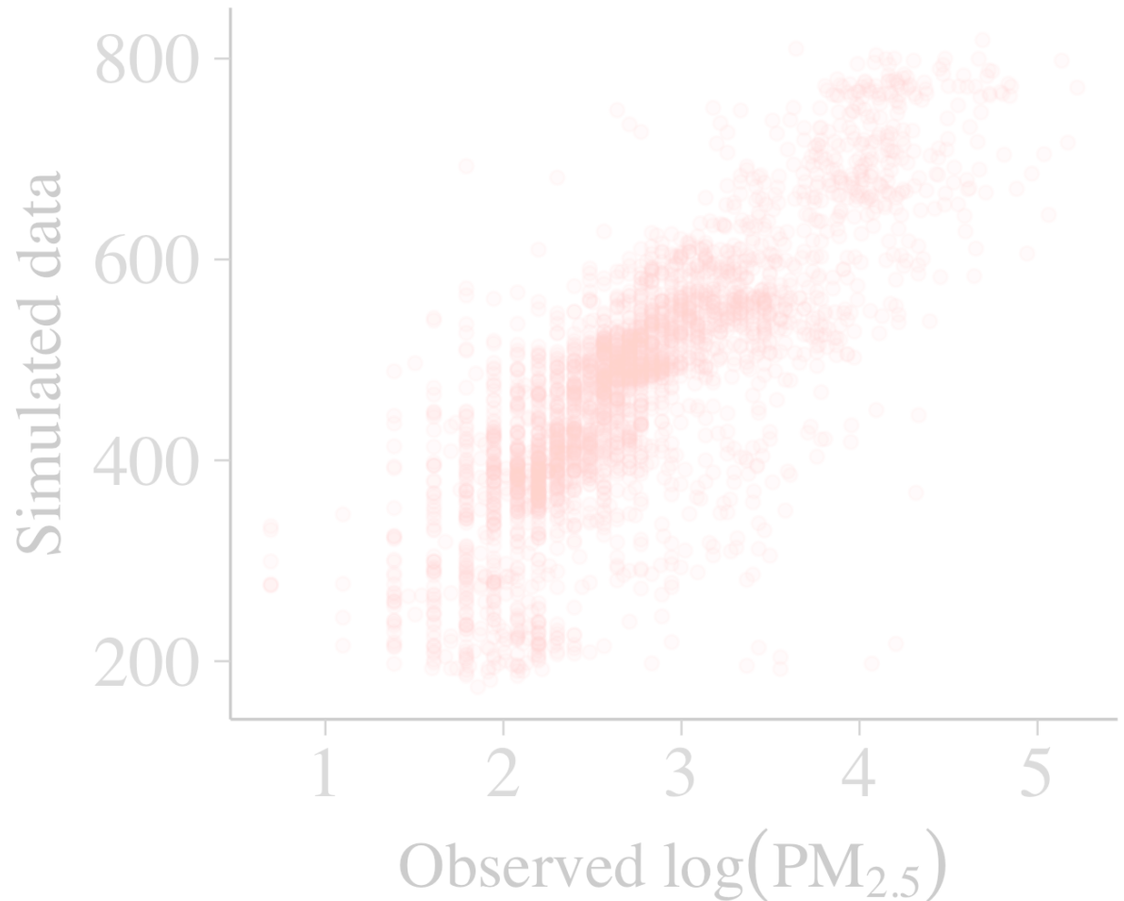
FAKE DATA IS ALMOST AS USEFUL AS REAL DATA

- The prior model is **two orders of magnitude** off the real data
- Two orders of magnitude **on the log scale!**



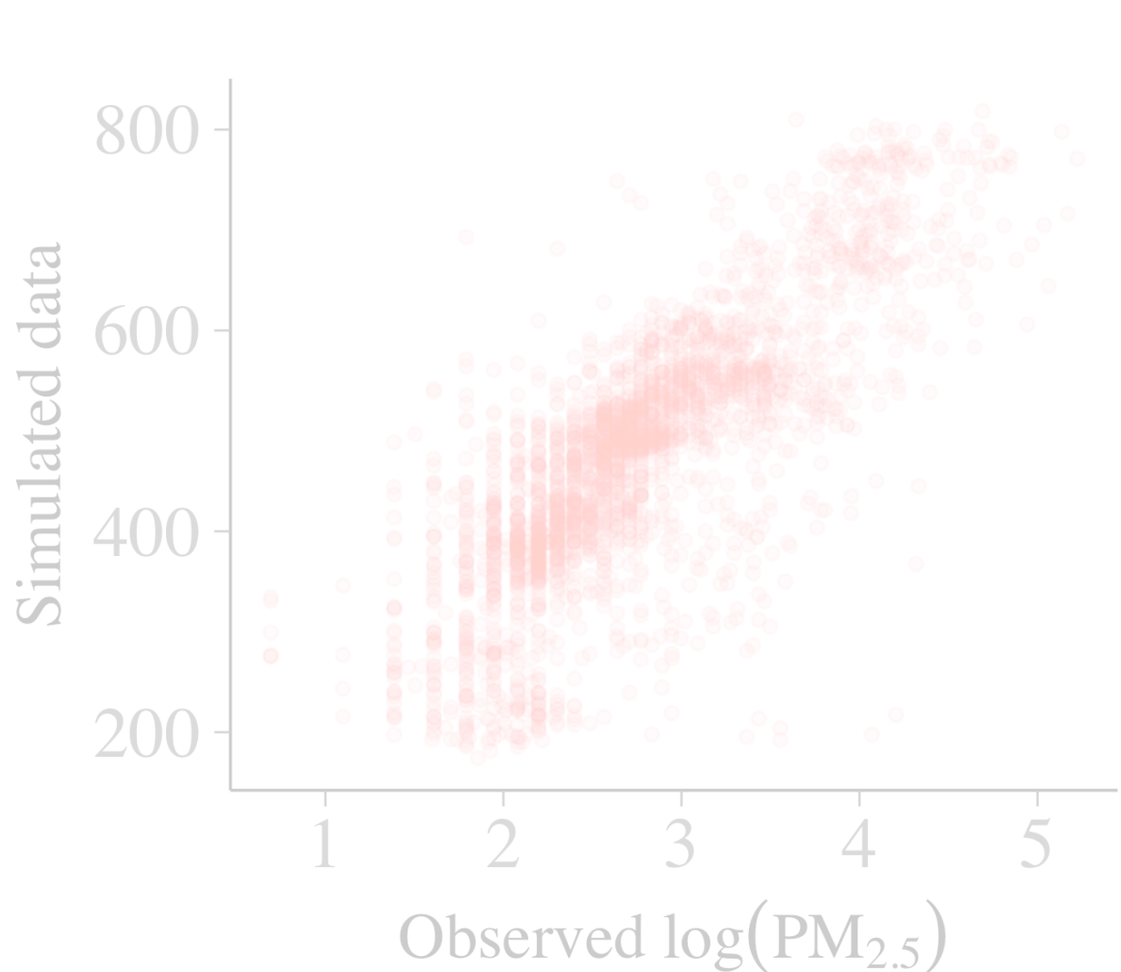
FAKE DATA IS ALMOST AS USEFUL AS REAL DATA

- The prior model is **two orders of magnitude** off the real data
- Two orders of magnitude **on the log scale!**



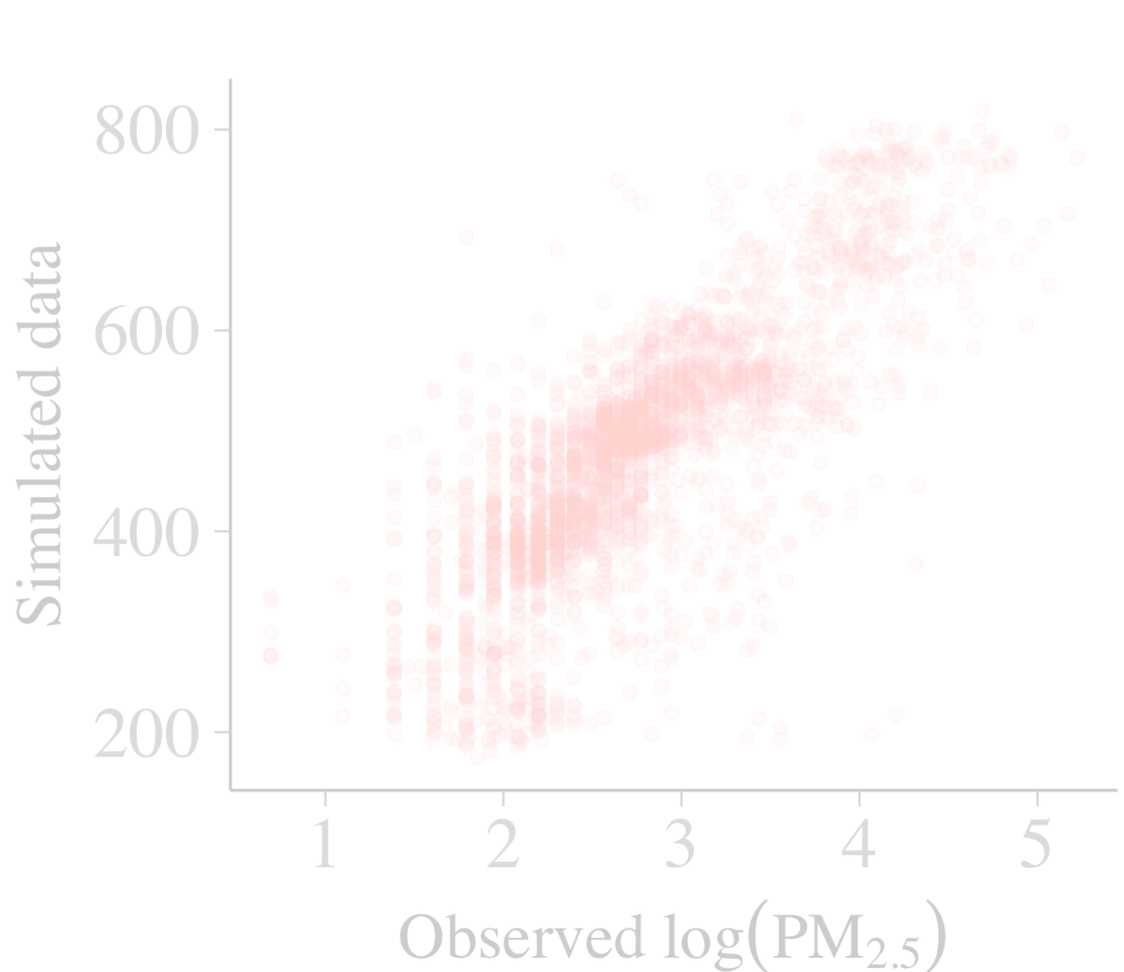
FAKE DATA IS ALMOST AS USEFUL AS REAL DATA

- The prior model is **two orders of magnitude** off the real data
- Two orders of magnitude **on the log scale!**
- What does this mean practically?



FAKE DATA IS ALMOST AS USEFUL AS REAL DATA

- The prior model is **two orders of magnitude** off the real data
- Two orders of magnitude **on the log scale!**
- What does this mean practically?
- The data will have to overcome the prior…



IT CAN GUIDE YOUR CHOICE OF PRIOR

What are better priors for the global intercept and slope and the hierarchical scale parameters?

IT CAN GUIDE YOUR CHOICE OF PRIOR

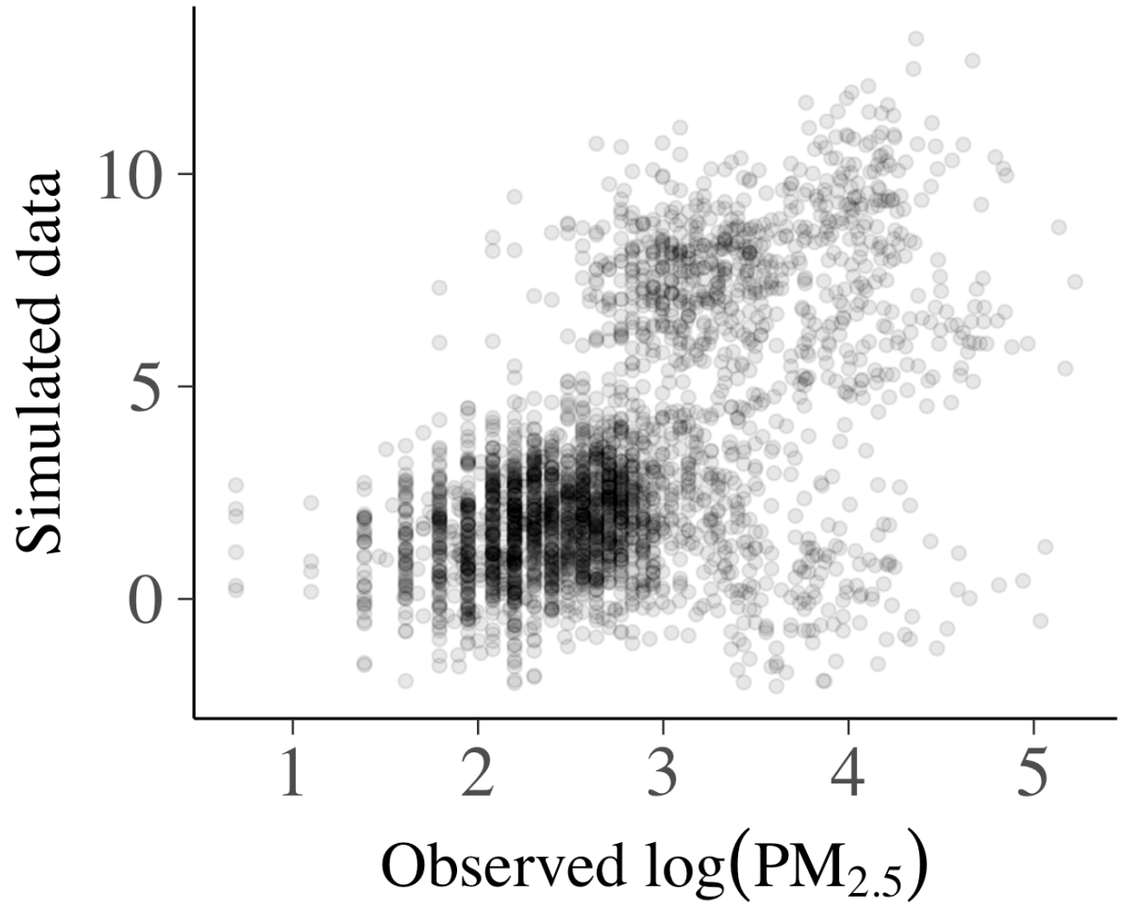
What are better priors for the global intercept and slope and the hierarchical scale parameters?

$$\alpha_0 \sim N(0, 1)$$
$$\beta_0 \sim N(1, 1)$$
$$\tau_\alpha \sim N_+(0, 1)$$
$$\tau_\beta \sim N_+(0, 1)$$

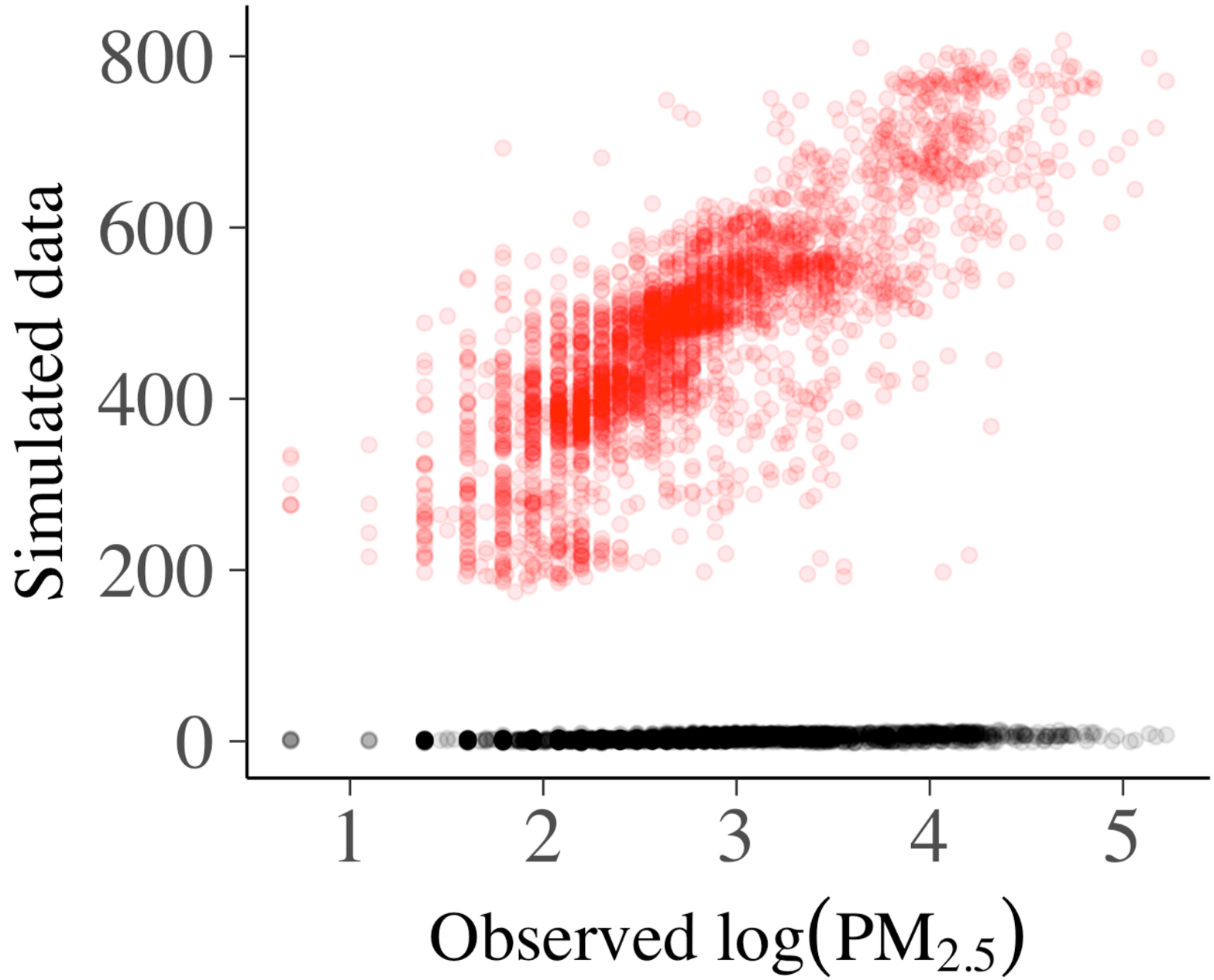
IT CAN GUIDE YOUR CHOICE OF PRIOR

What are better priors for the global intercept and slope and the hierarchical scale parameters?

$$\alpha_0 \sim N(0, 1)$$
$$\beta_0 \sim N(1, 1)$$
$$\tau_\alpha \sim N_+(0, 1)$$
$$\tau_\beta \sim N_+(0, 1)$$

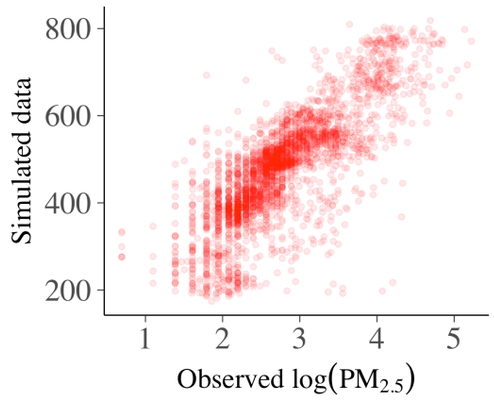


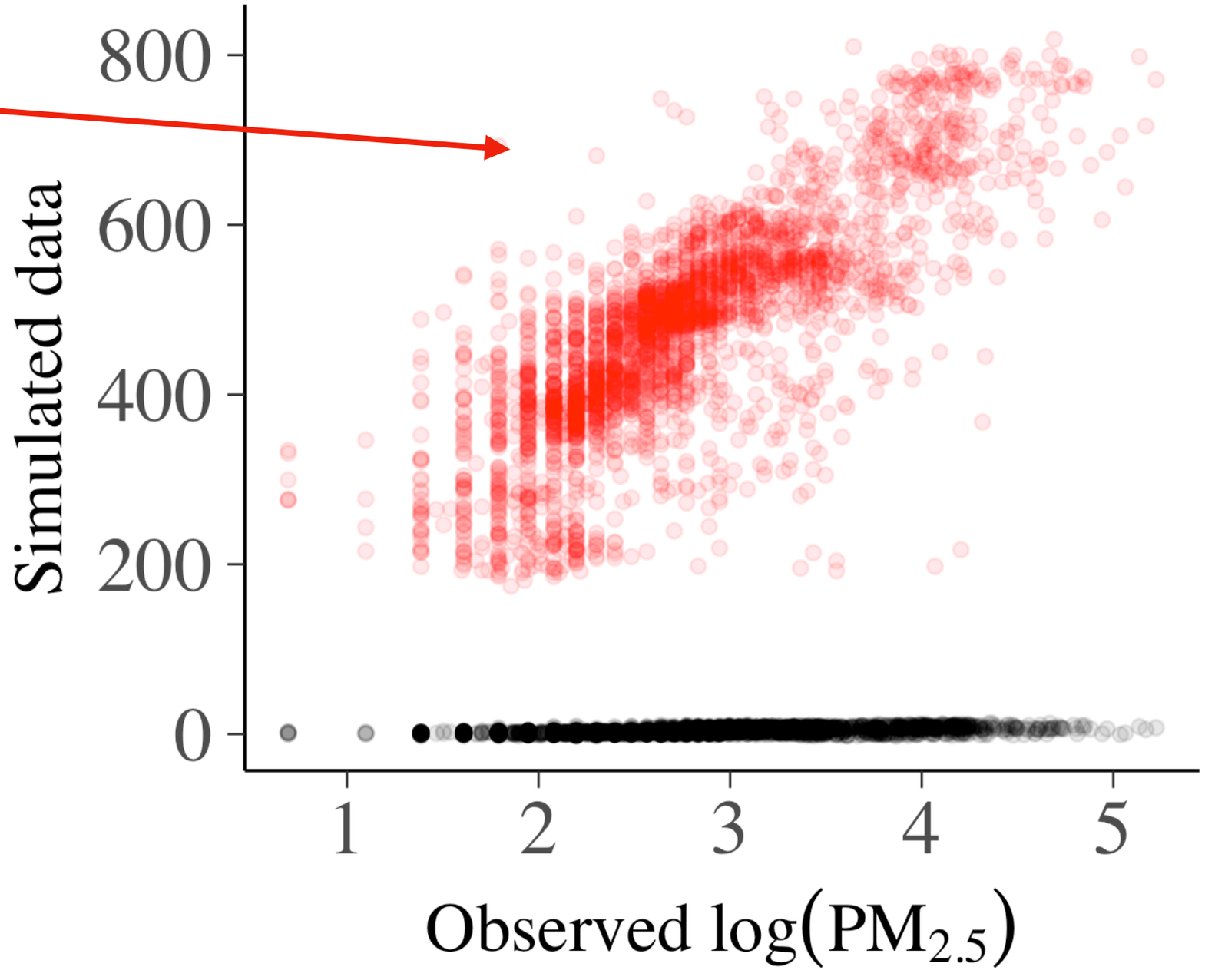
AND MAKE IT EASIER TO DEFEND YOUR MODELLING CHOICES



AND MAKE IT EASIER TO DEFEND YOUR MODELLING CHOICES

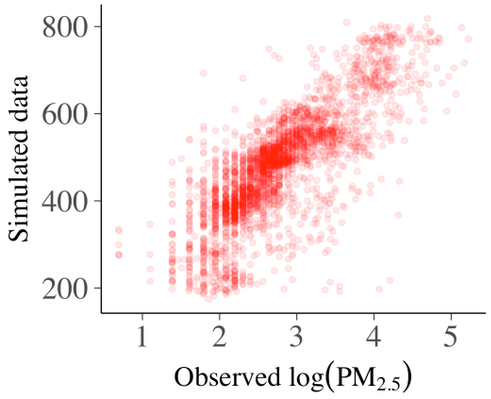
Non-informative



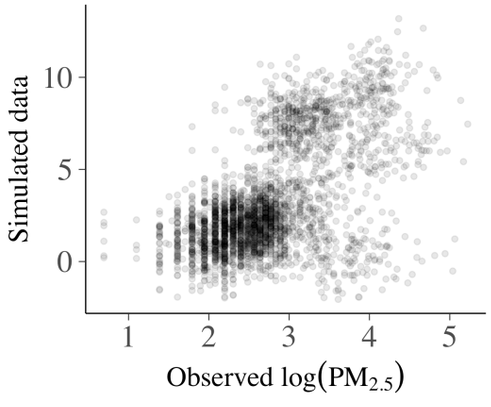


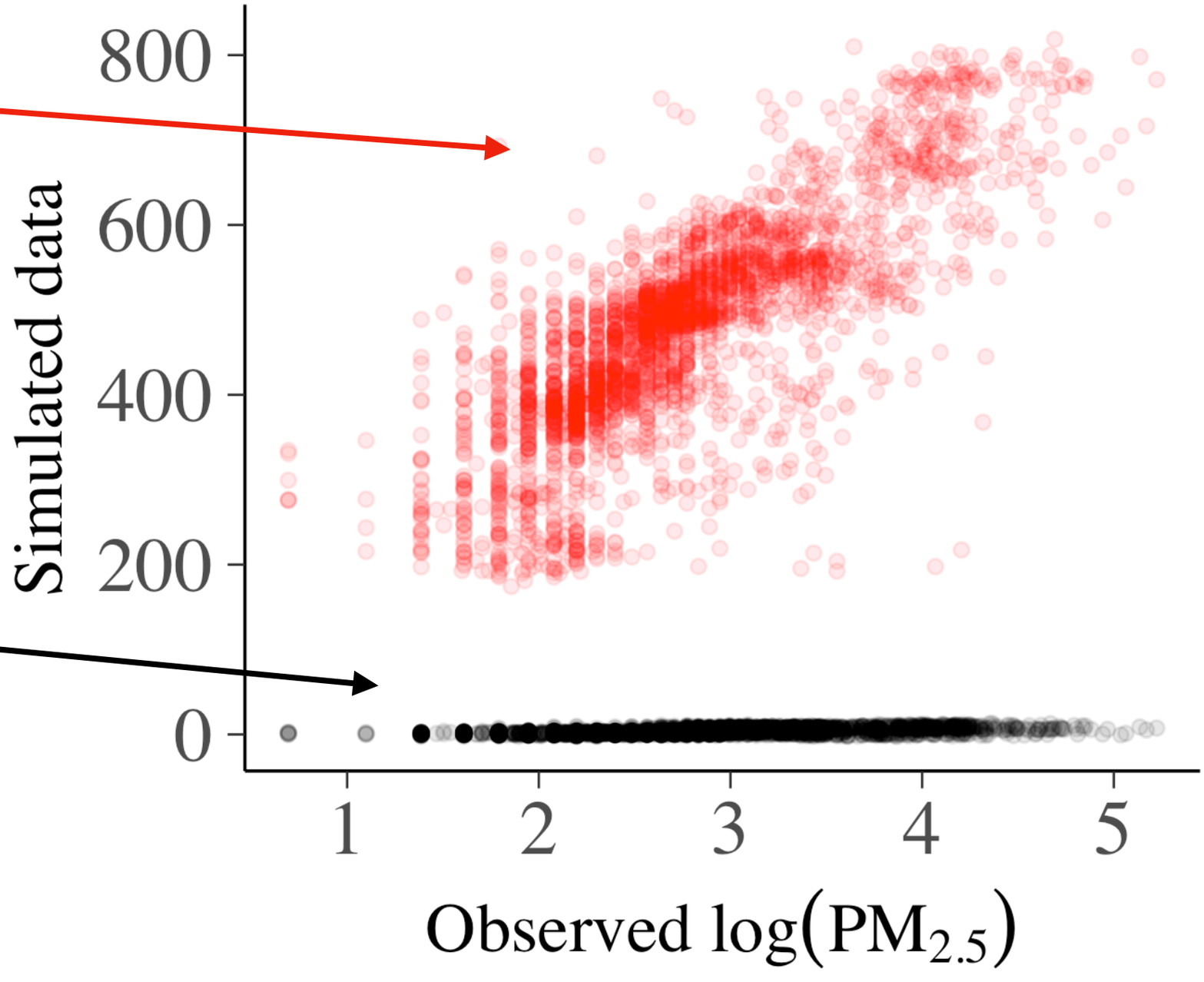
AND MAKE IT EASIER TO DEFEND YOUR MODELLING CHOICES

Non-informative



Weakly informative





Data gathering

Asymptotic regime

Model evaluation criteria

Likelihood

Prior

Computation

WHICH IS A KIND OF INTEGRITY, IF YOU LOOK ON EVERY EXIT BEING AN ENTRANCE SOMEWHERE ELSE

(Tom Stoppard)

STATISTICS IS HARD

- As tempting as it is, there is no way to avoid thinking of all of the aspects of the model simultaneously
- Think of the aspects of your data gathering, modelling, computation, and model evaluation as all being made of the same substance
- And right now, I'm not sure there are any good ways to keep track of anything at once

THERE WON'T BE TRUMPETS

- Sometimes there are loud warnings that things have gone badly:
  - Divergences
  - R-hat (kinda)
  - Simulation Based Calibration (expensive)
  - Prior predictive simulations (if you're clever)
  - Posterior predictive checks (watch your assumptions)
- But really, we need to build careful simulation studies (like in the GP case) and meaningful checks of the pre-observation joint distribution of the parameters and the data.

HAROLD HOLT'S HUBRIS

- Harold Holt went swimming in dangerous surf and drowned.
- No amount of synchronized swimming would not have saved him.
- So make sure you focus on the right things and stop just building memorial swimming pools.