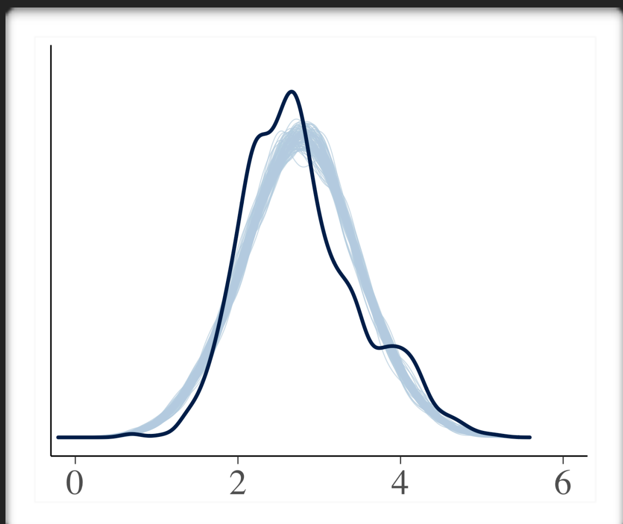
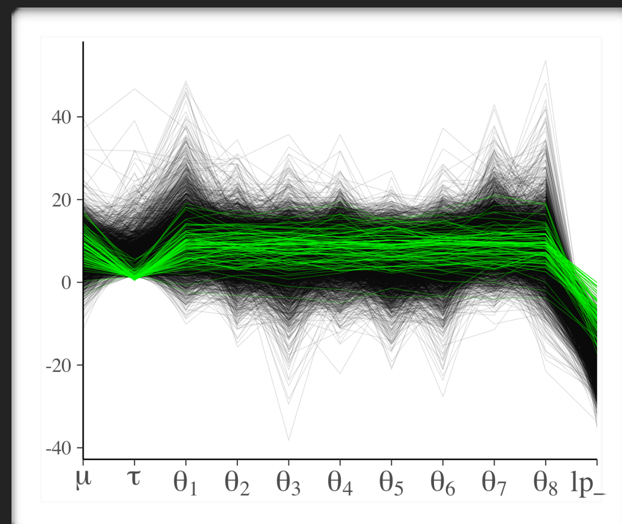
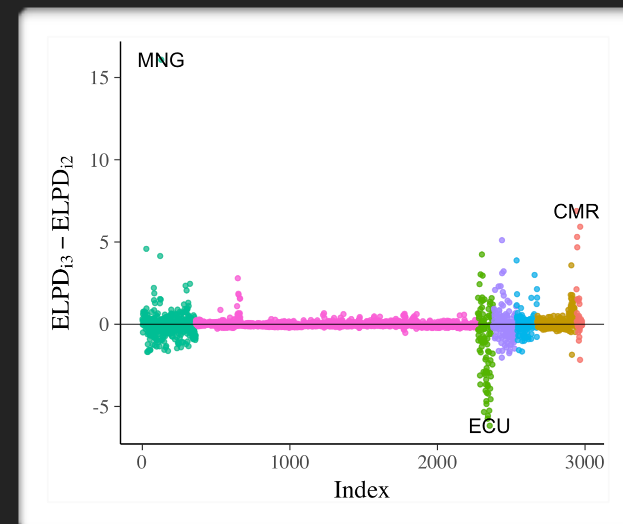
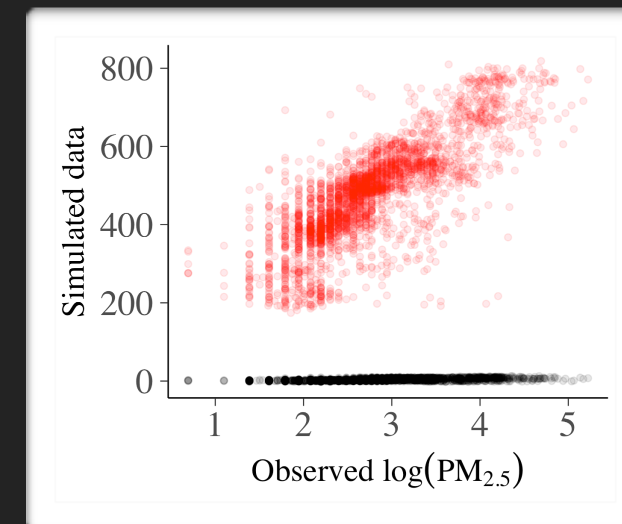
Visualization in Bayesian workflow









Jonah Gabry

Columbia University
Stan Development Team

Workflow
Bayesian data analysis

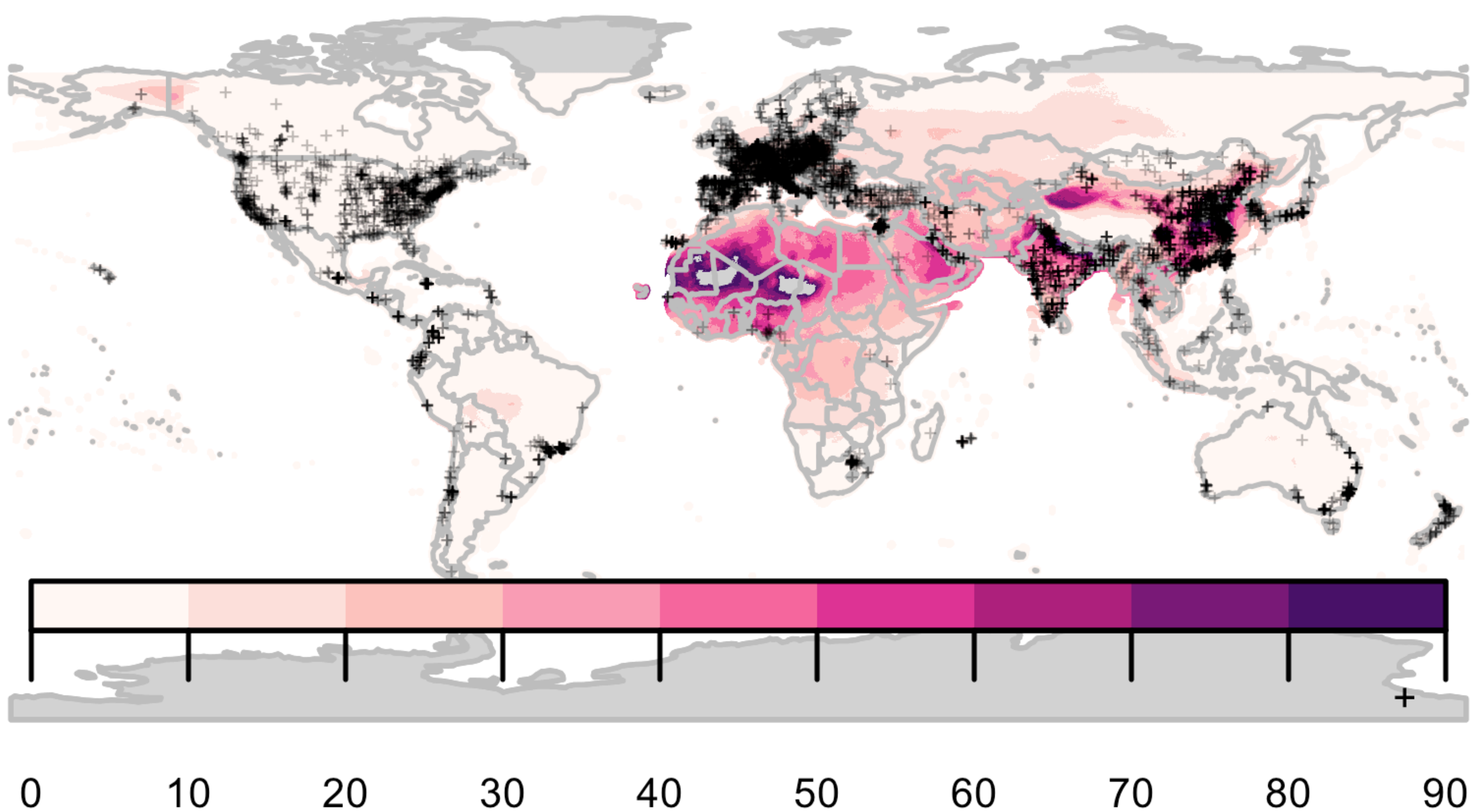
- Exploratory data analysis
- *Prior* predictive checking
- Model fitting and algorithm diagnostics
- *Posterior* predictive checking
- Model comparison (e.g., via cross-validation)

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2018).
Visualization in Bayesian workflow.
Forthcoming in *Journal of the Royal Statistical Society Series A*
arXiv preprint: arxiv.org/abs/1709.01449
Code: github.com/jgabry/bayes-vis-paper

Example

Goal Estimate global PM2.5 concentration

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



black points indicate ground monitor locations

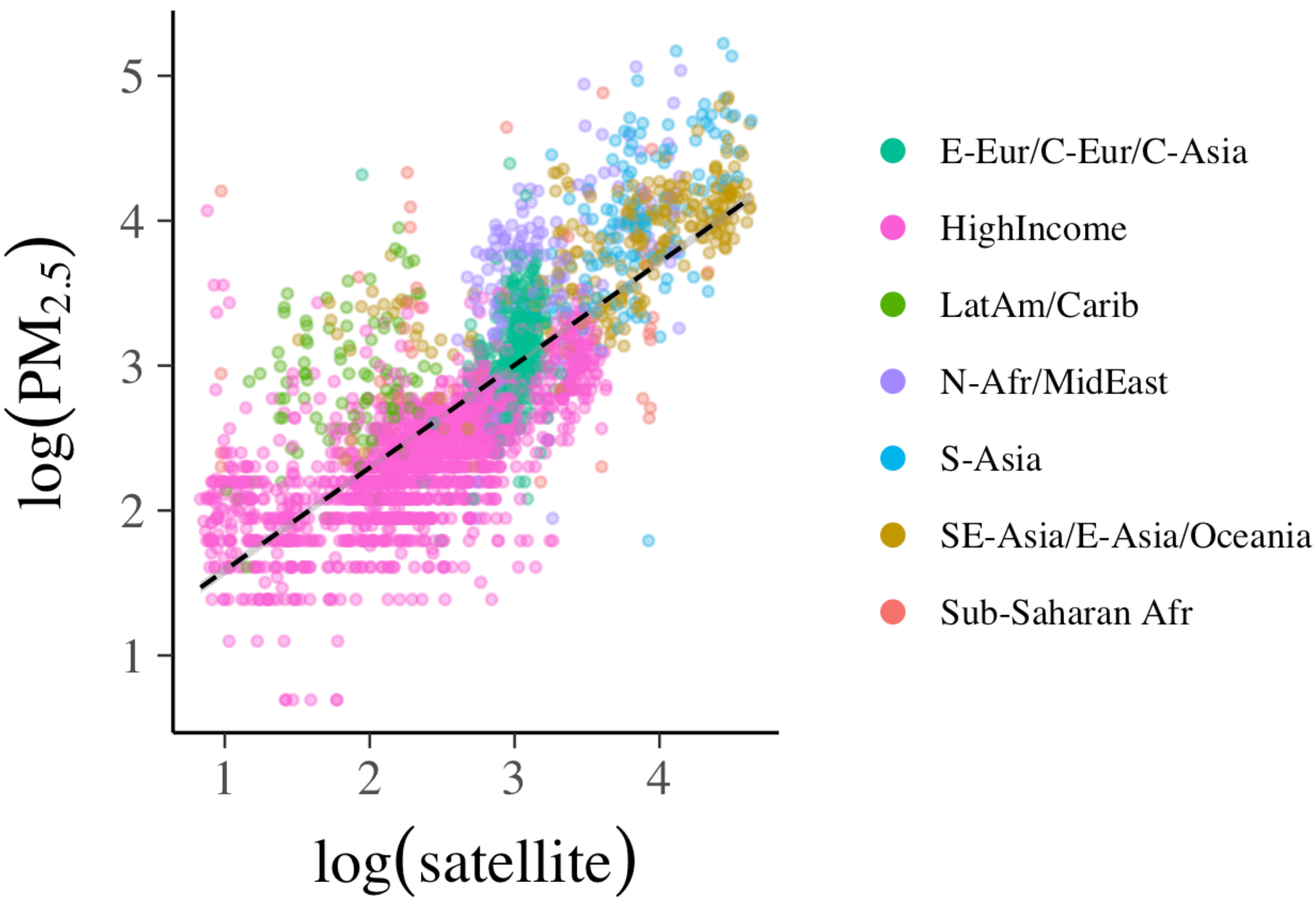
Satellite estimates of PM2.5 and ground monitor locations

Exploratory Data Analysis

Building a network of models

Exploratory data analysis

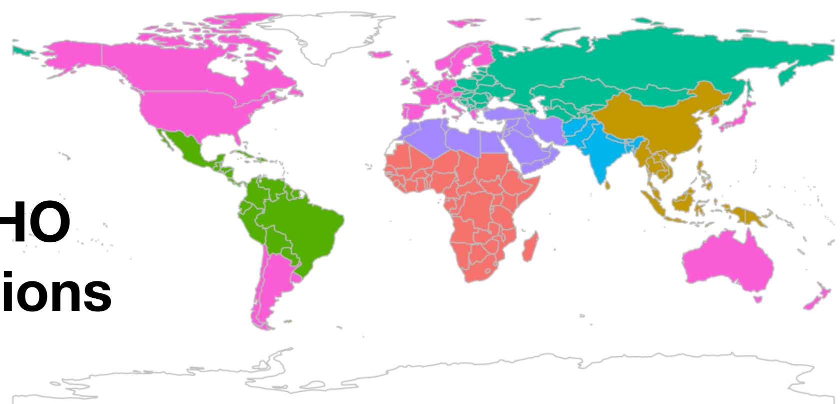
building a network of models

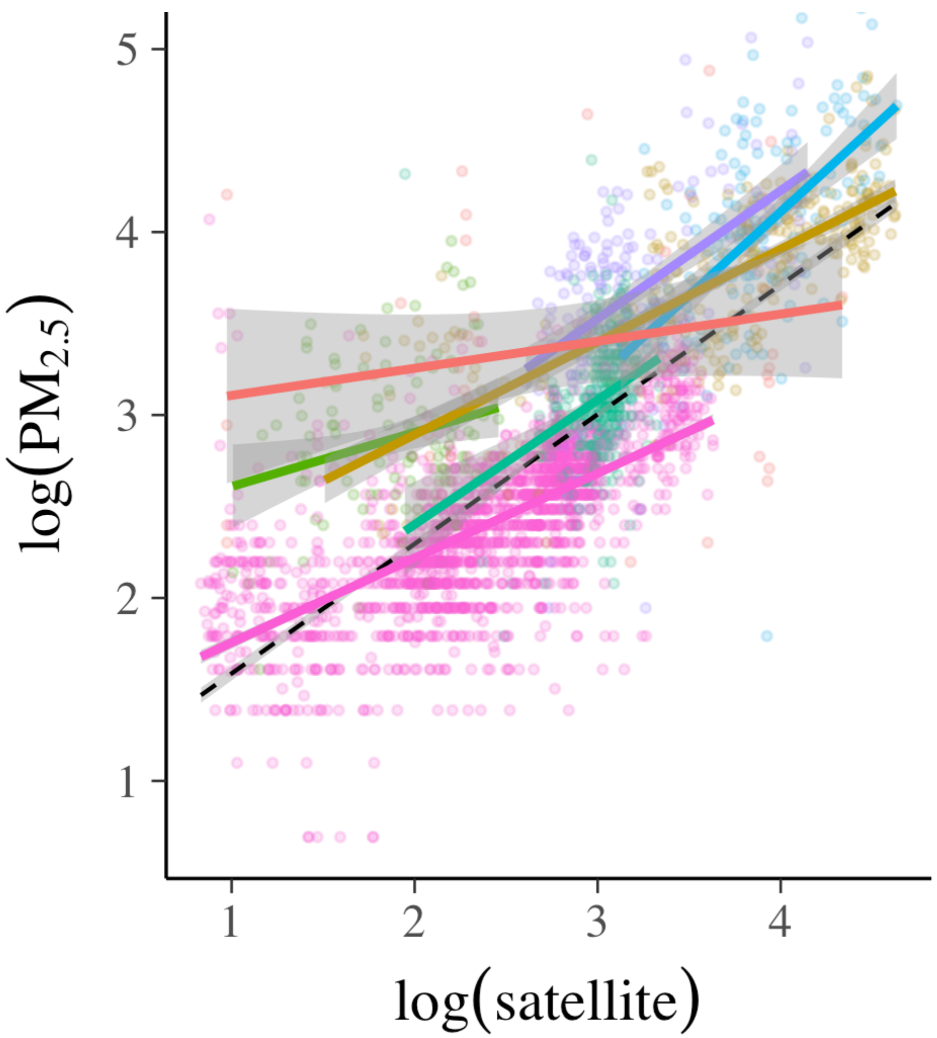


Exploratory data analysis

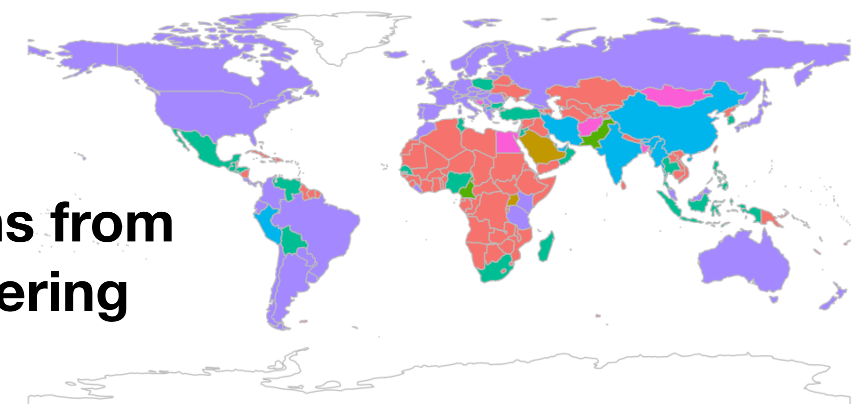
building a network of models

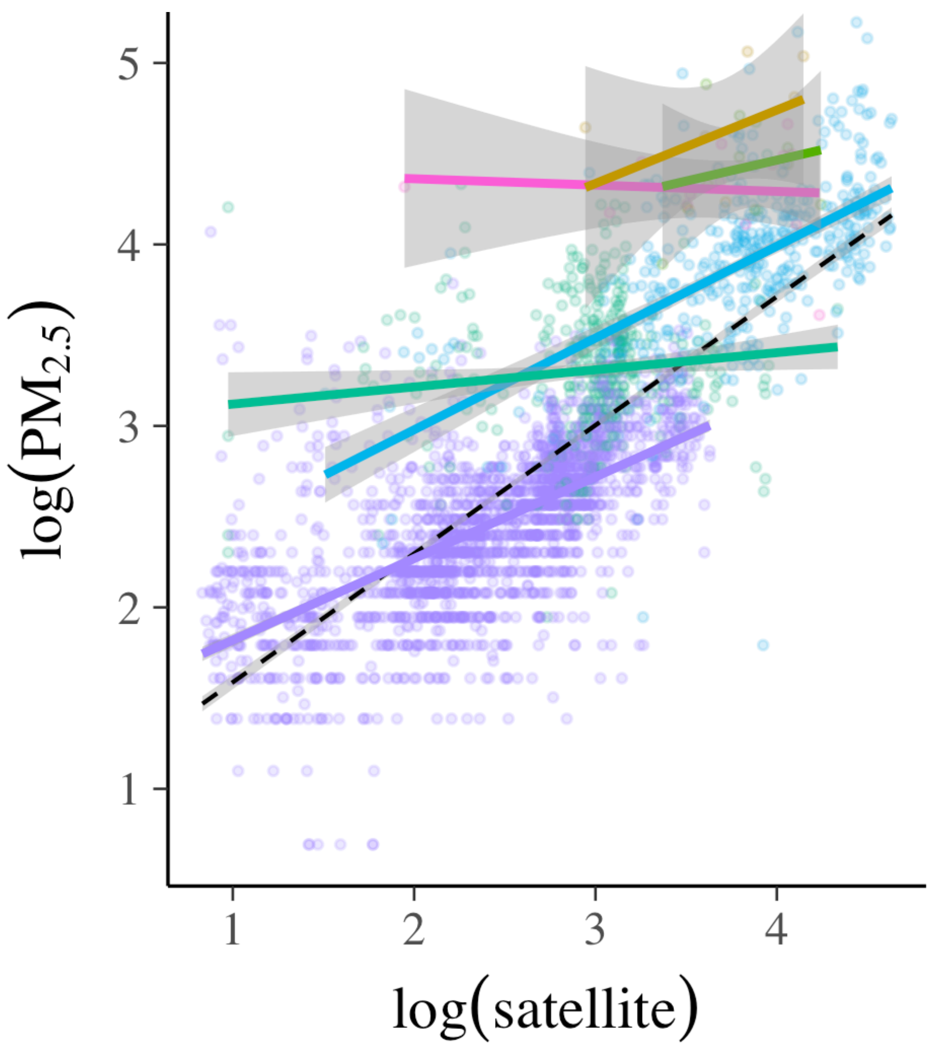
WHO Regions





Regions from clustering





Exploratory data analysis
building a network of models

For measurements $n = 1, \ldots, N$

and regions $j = 1, \ldots, J$

Model 1

$$\log\left(\mathrm{PM}_{2.5,nj}\right) \sim N(\alpha + \beta \log\left(\mathrm{sat}_{nj}\right), \sigma)$$

Exploratory data analysis
building a network of models

For measurements $n = 1, \ldots, N$

and regions $j = 1, \ldots, J$

Models 2 and 3

$$\log\left(\mathrm{PM}_{2.5,nj}\right) \sim N(\mu_{nj}, \sigma)$$

$$\mu_{nj} = \boxed{\alpha_0 + \alpha_j} + \boxed{(\beta_0 + \beta_j)} \log\left(\mathrm{sat}_{nj}\right)$$

$$\alpha_j \sim N(0, \tau_\alpha) \quad \beta_j \sim N(0, \tau_\beta)$$

Prior predictive checks

Fake data can be almost as valuable as real data

A Bayesian modeler commits to an a priori *joint distribution*

Likelihood x Prior

$$p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathbf{y}) p(\mathbf{y})$$

Posterior x
Marginal Likelihood

Data
(observed)

Parameters
(unobserved)

What is the problem with “vague” priors?

- If we use an *improper* prior, then we do not specify a joint model for our data and parameters
- More importantly, we do not specify a data generating mechanism ***p(y)***
- By construction, these priors do not regularize inferences, which is quite often a bad idea
- Proper but diffuse is better than improper but is still often problematic

Generative models

- If we disallow improper priors, then Bayesian modeling is generative

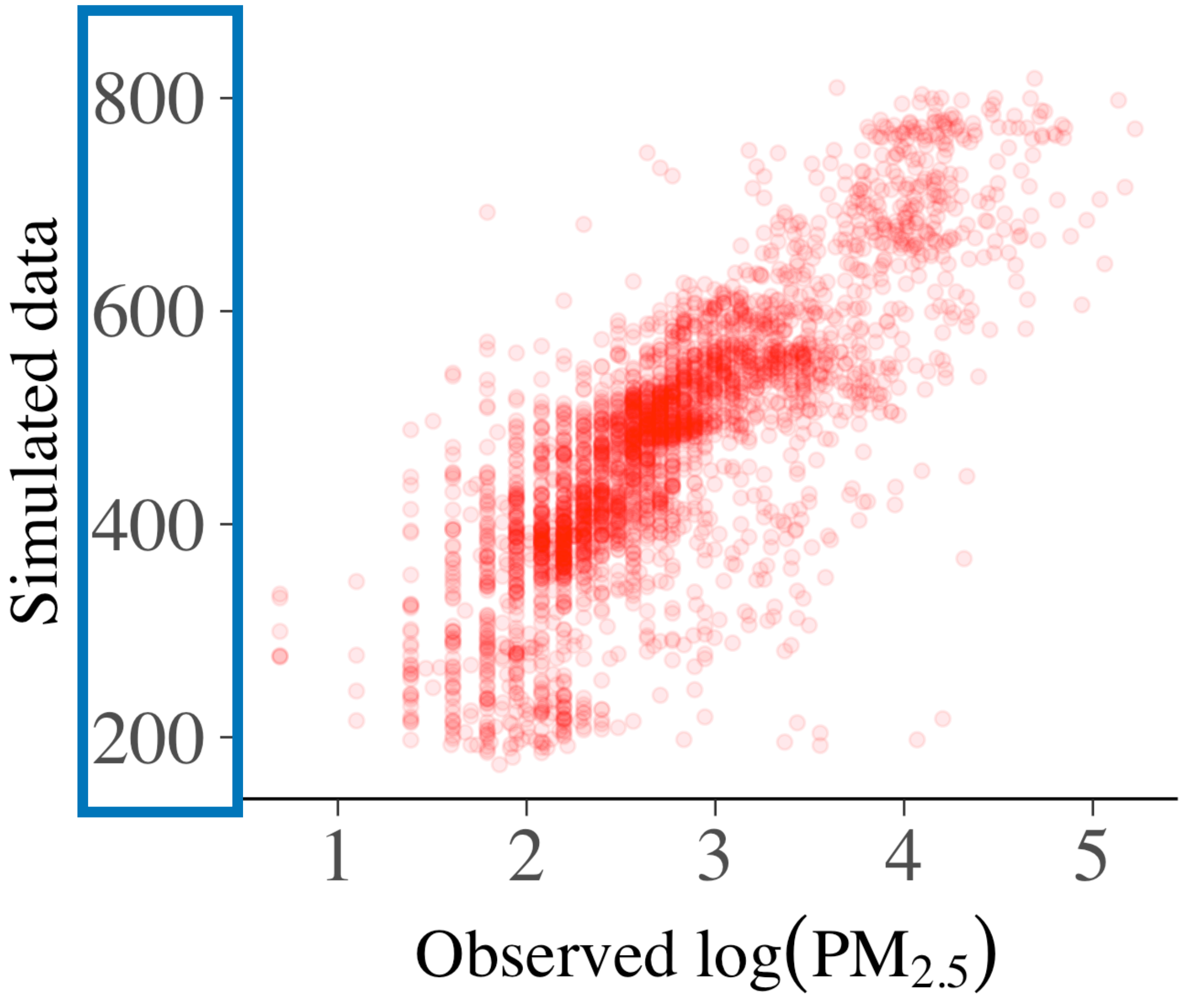
- In particular, we have a simple way to simulate from *p(y)*:

$$\begin{array}{c} \theta^\star \sim p(\theta) \\ \downarrow \\ y^\star \sim p(y|\theta^\star) \end{array} \quad \Longleftrightarrow \quad y^\star \sim p(y)$$

Prior predictive checking:
fake data is almost as useful as real data

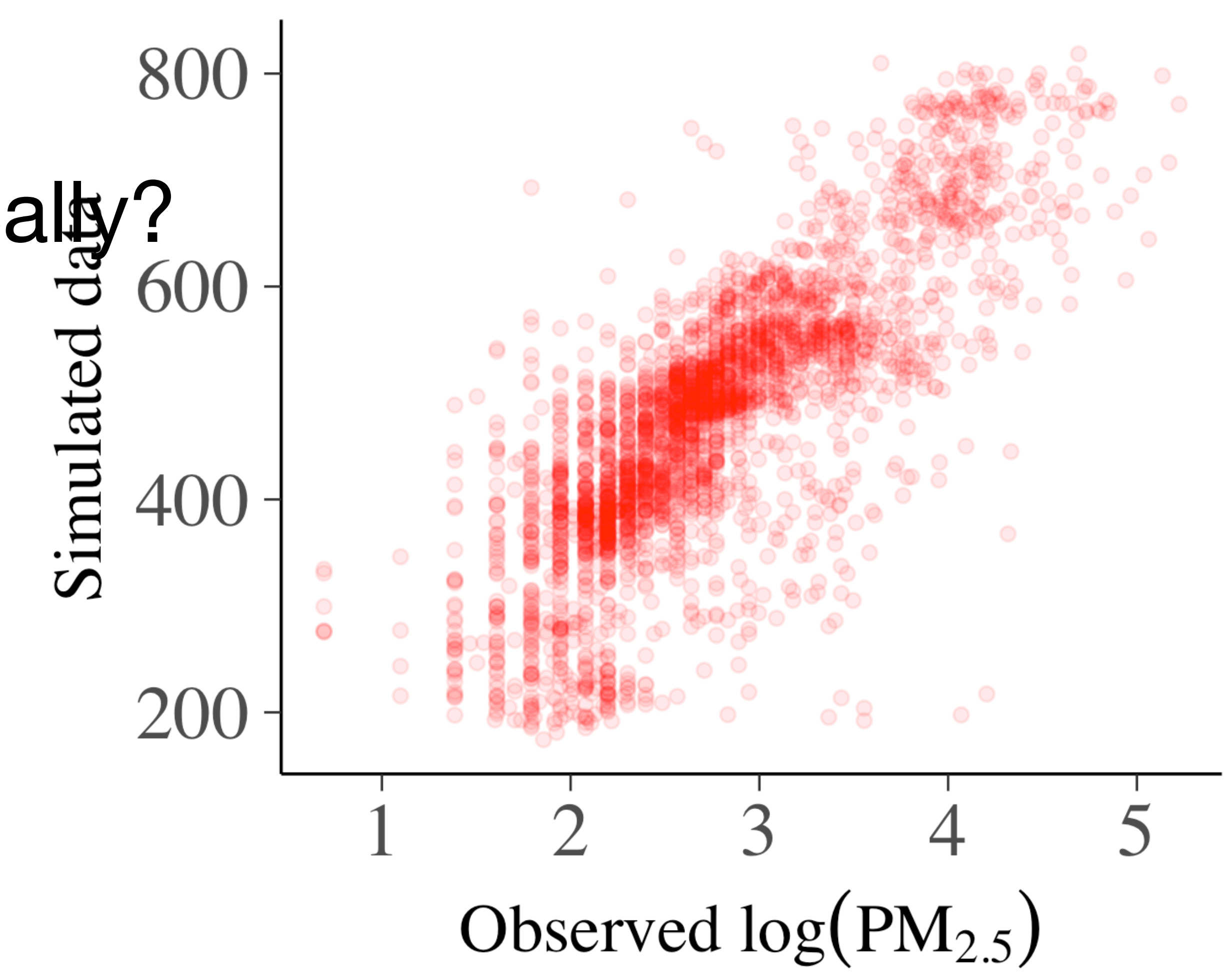
What do vague/non-informative priors imply about the data our model can generate?

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



Prior predictive checking:
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…

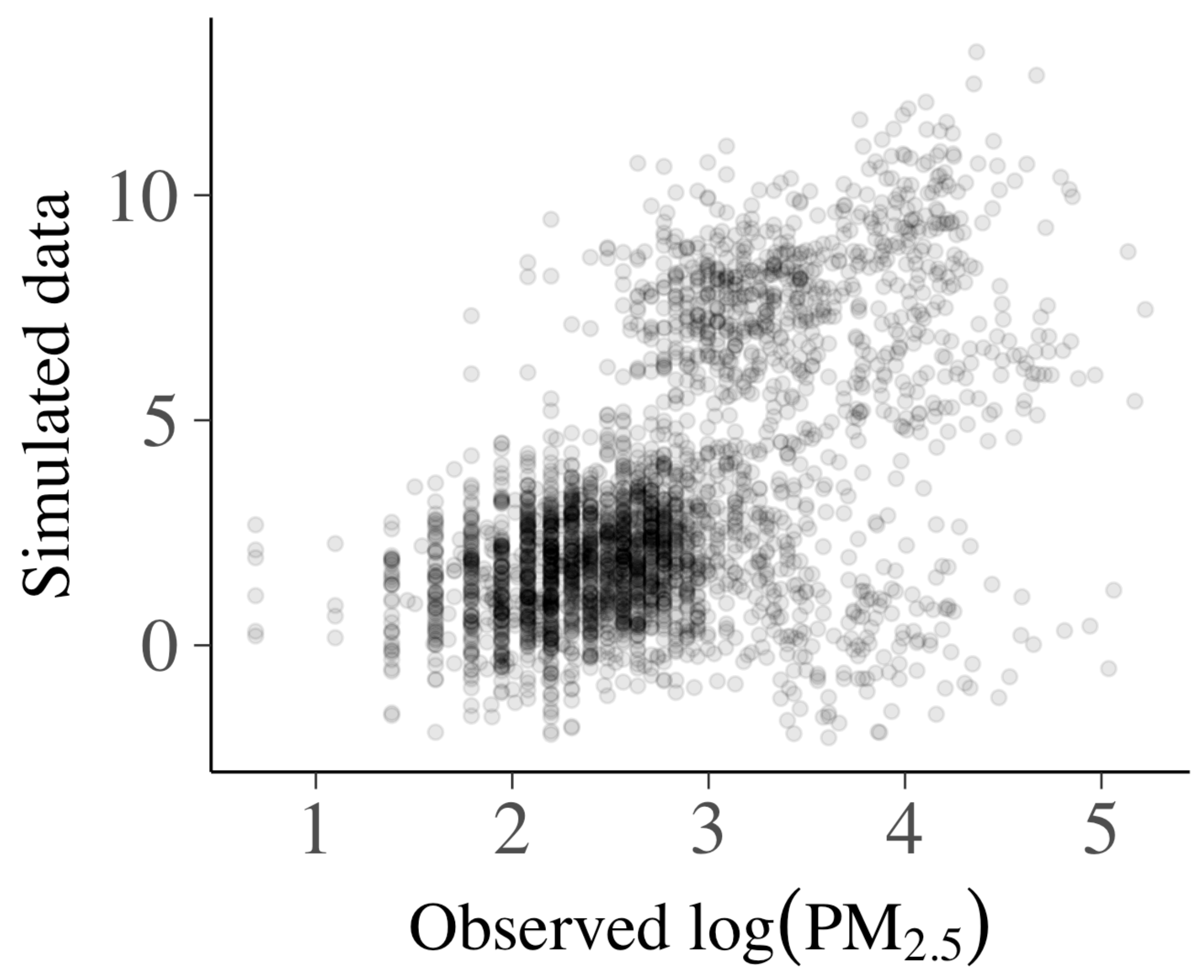


Prior predictive checking:

fake data is almost as useful as real data

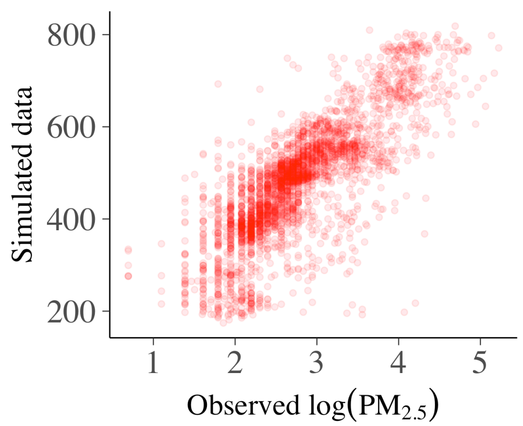
What are better priors for the global intercept and slope and the hierarchical scale parameters?

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

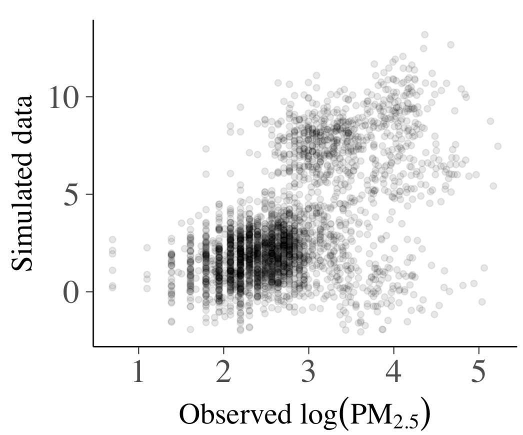


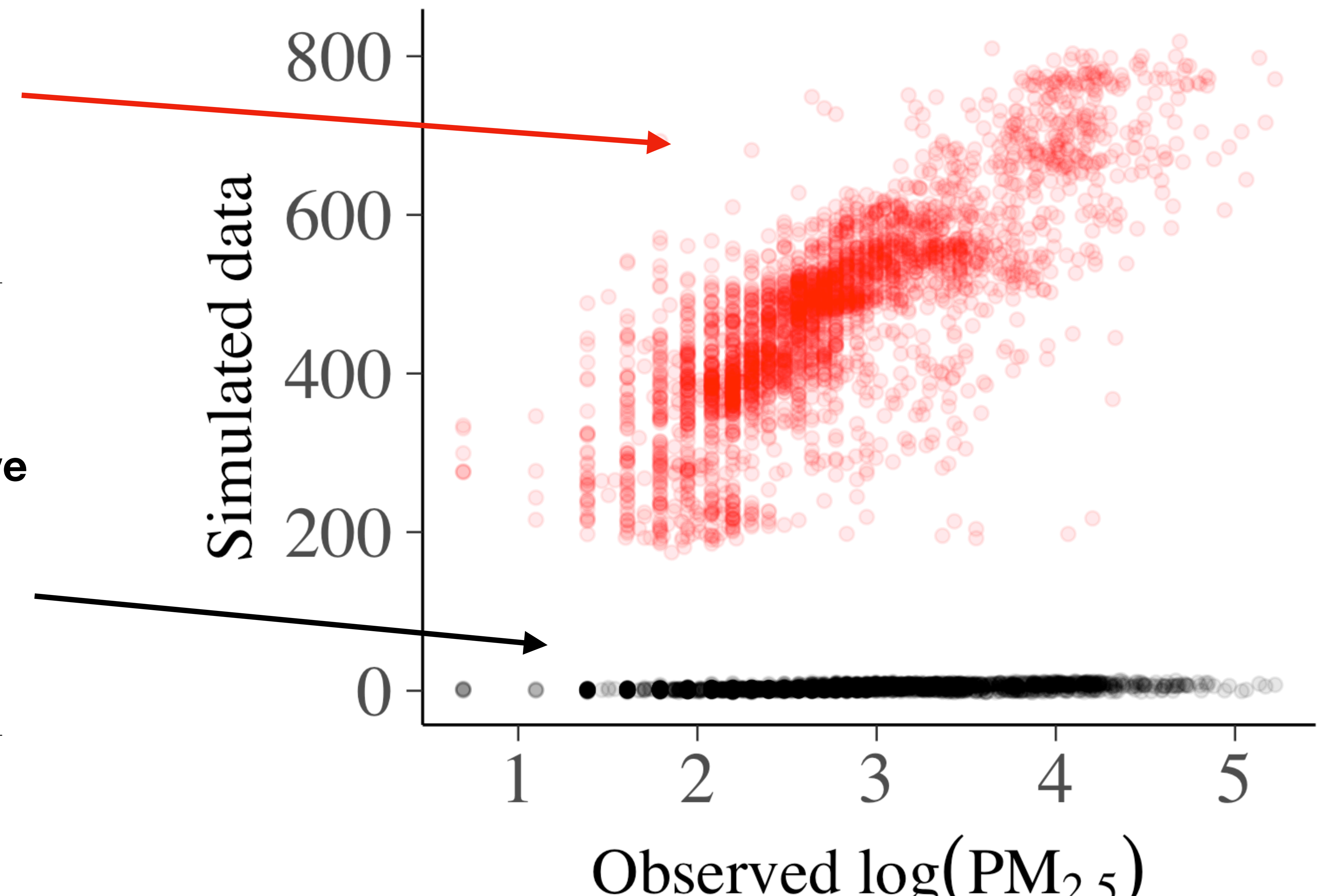
Prior predictive checking:
fake data is almost as useful as real data

Non-informative



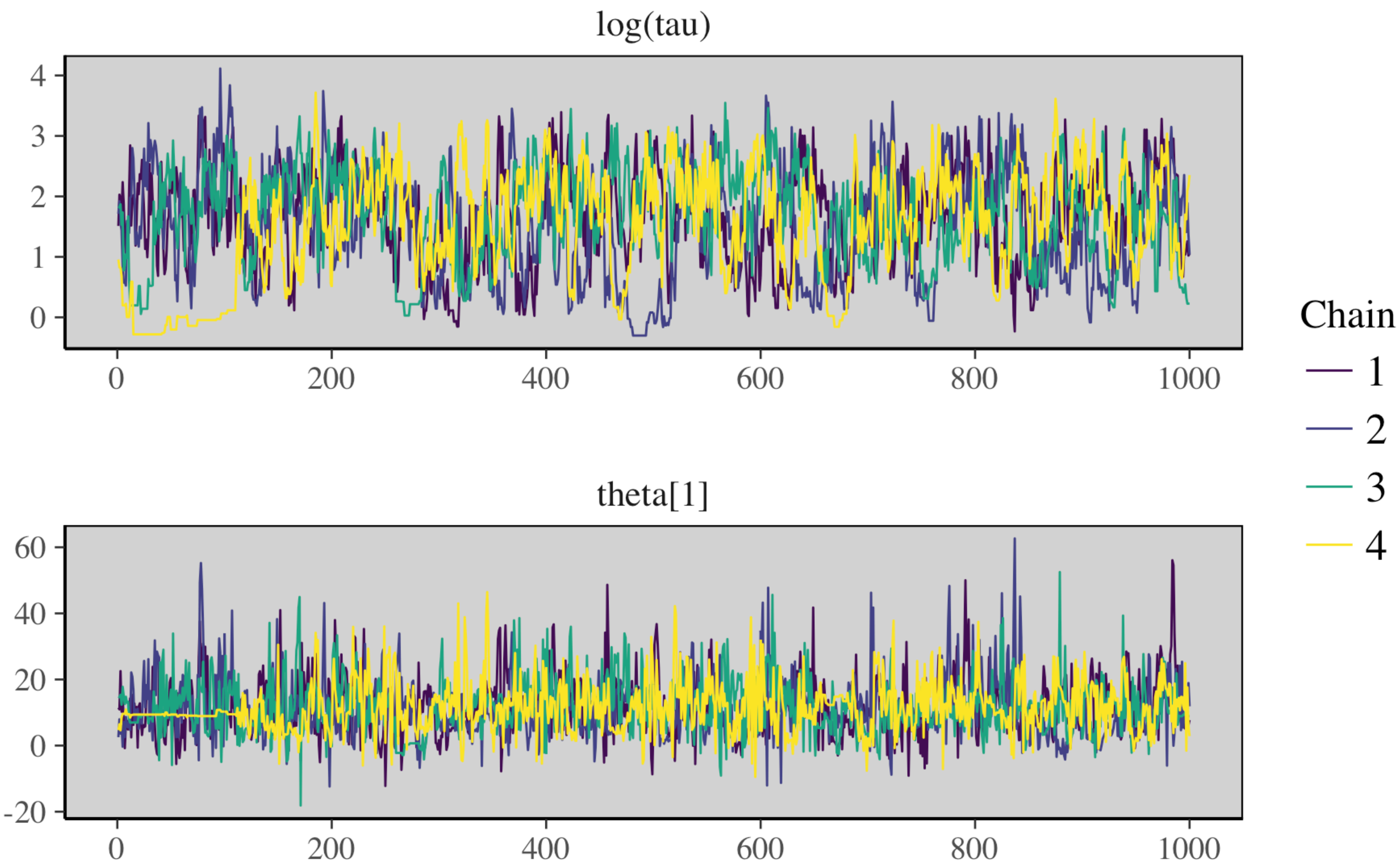
Weakly informative



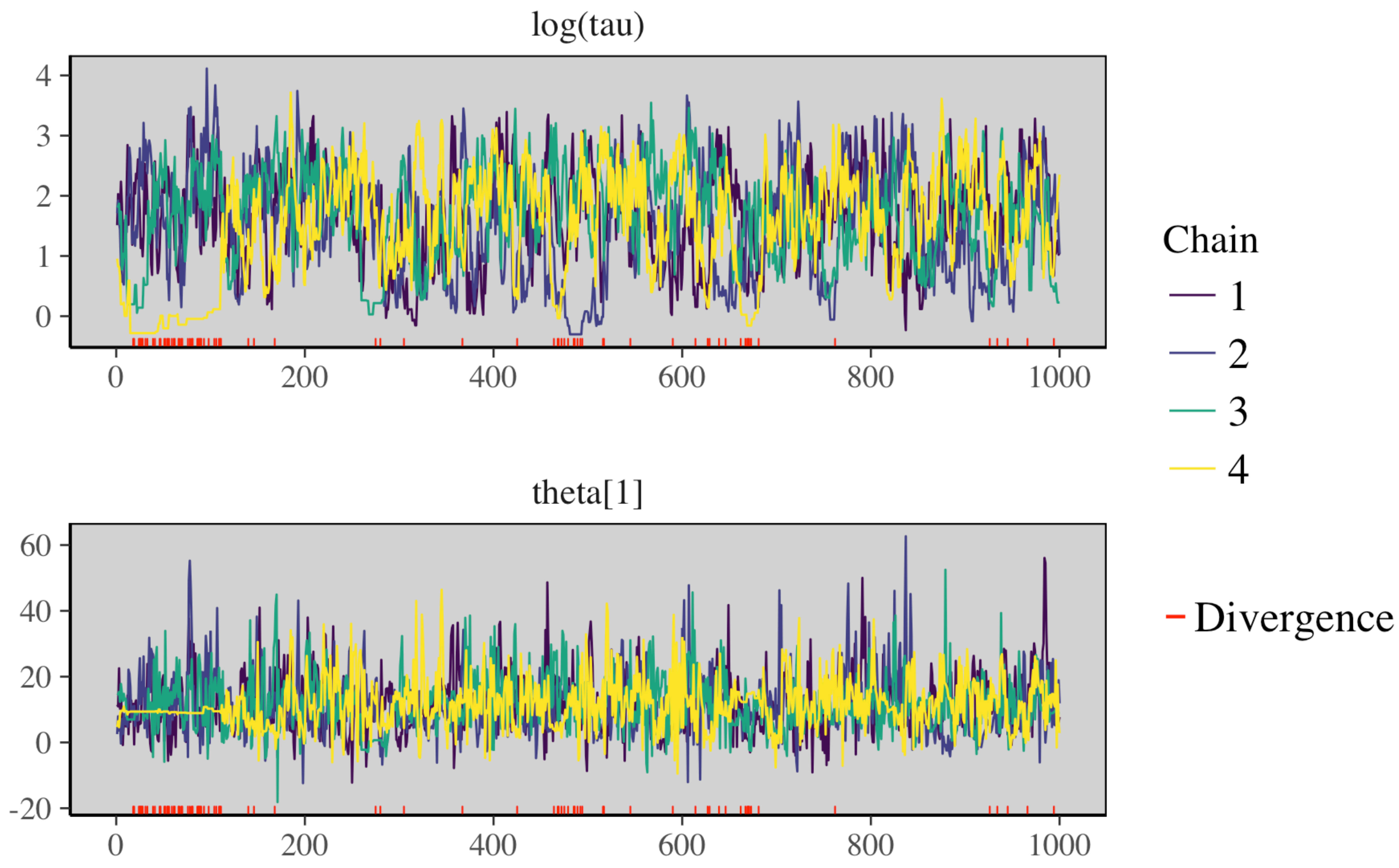


MCMC diagnostics

Beyond trace plots



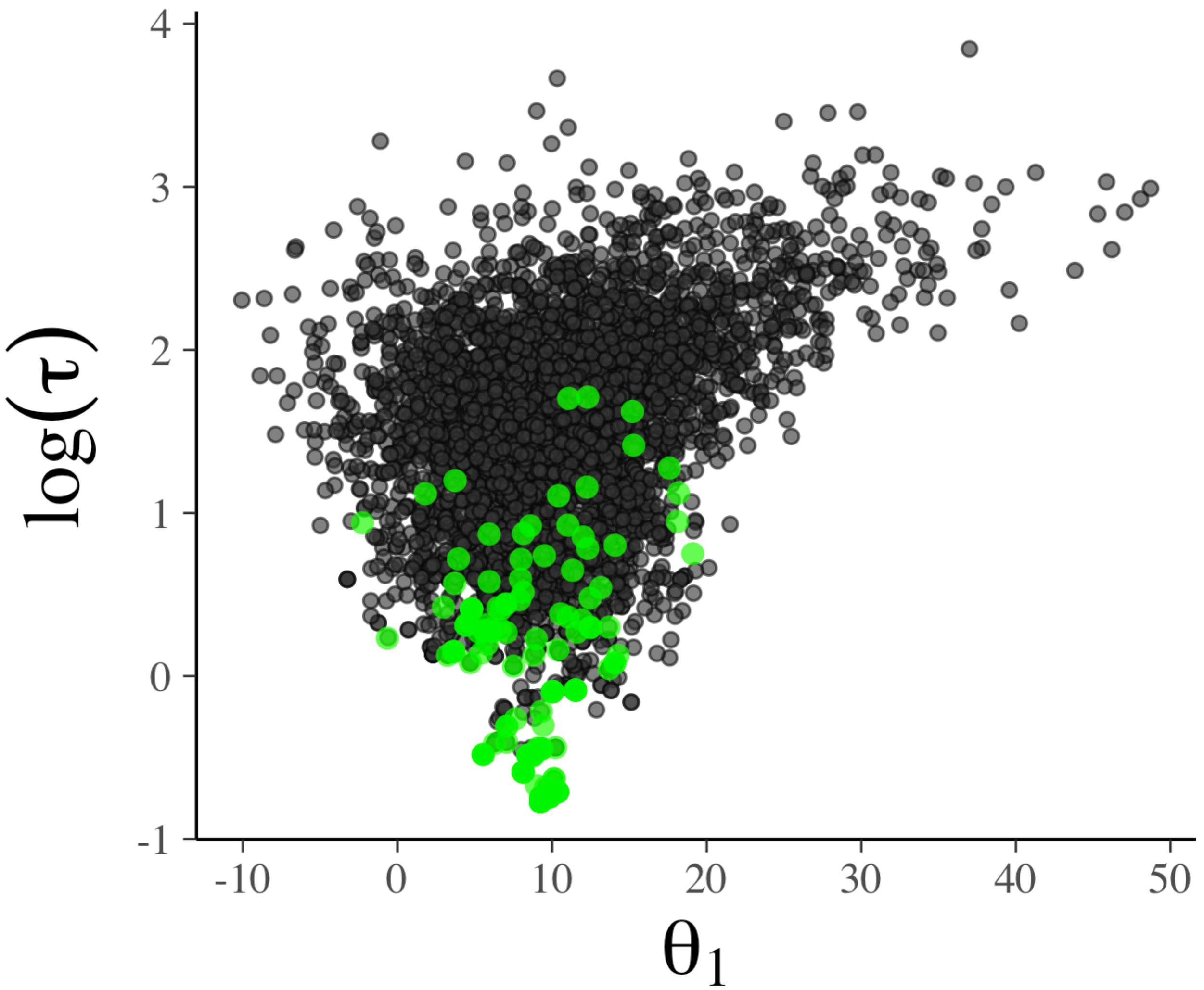
log(tau)
4
3
2
1
0
0
200
400
600
800
1000
theta[1]
60
40
20
0
-20
0
200
400
600
800
1000
Chain
1
2
3
4

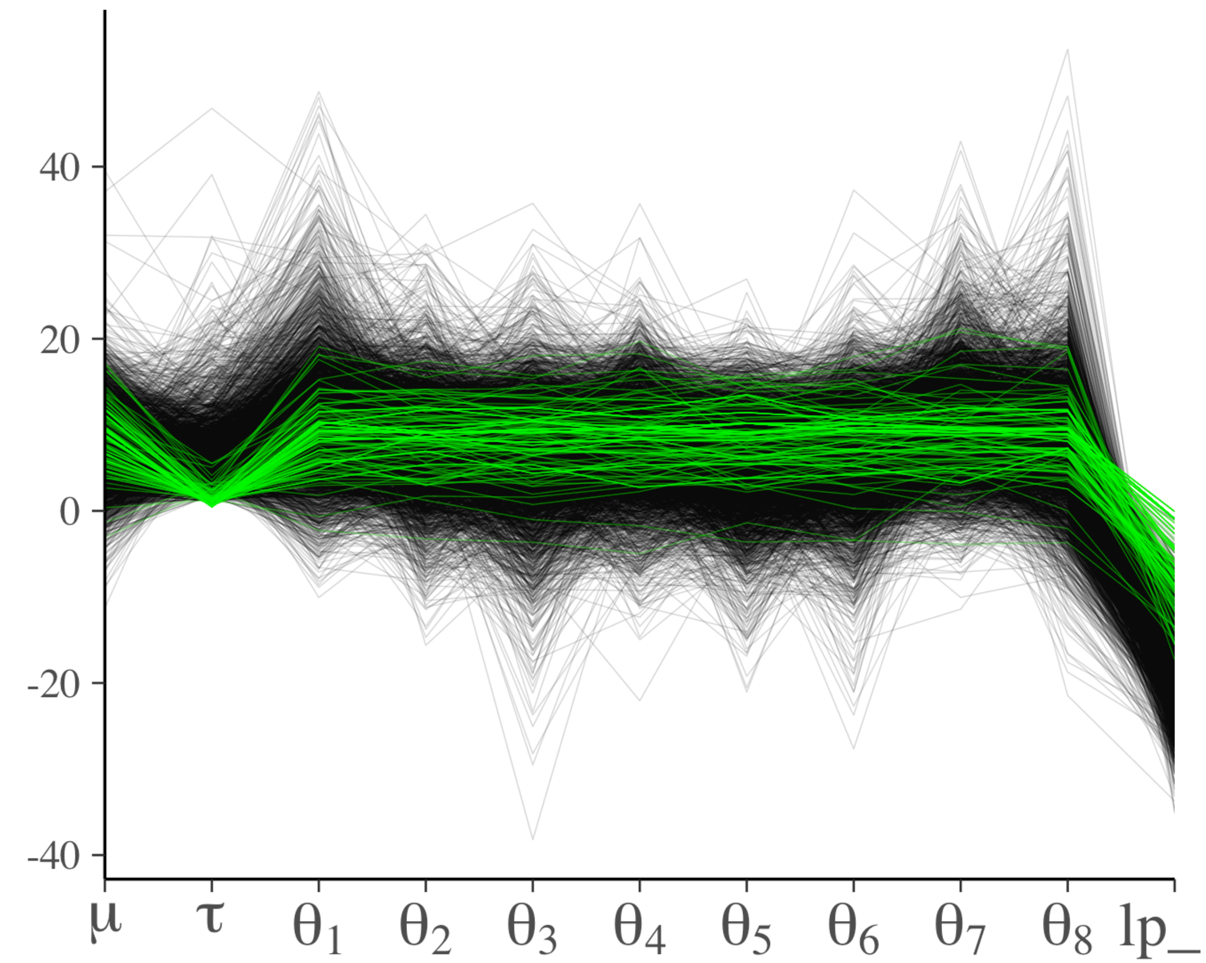


log(tau)
theta[1]
Chain
1
2
3
4
Divergence

MCMC diagnostics

beyond trace plots



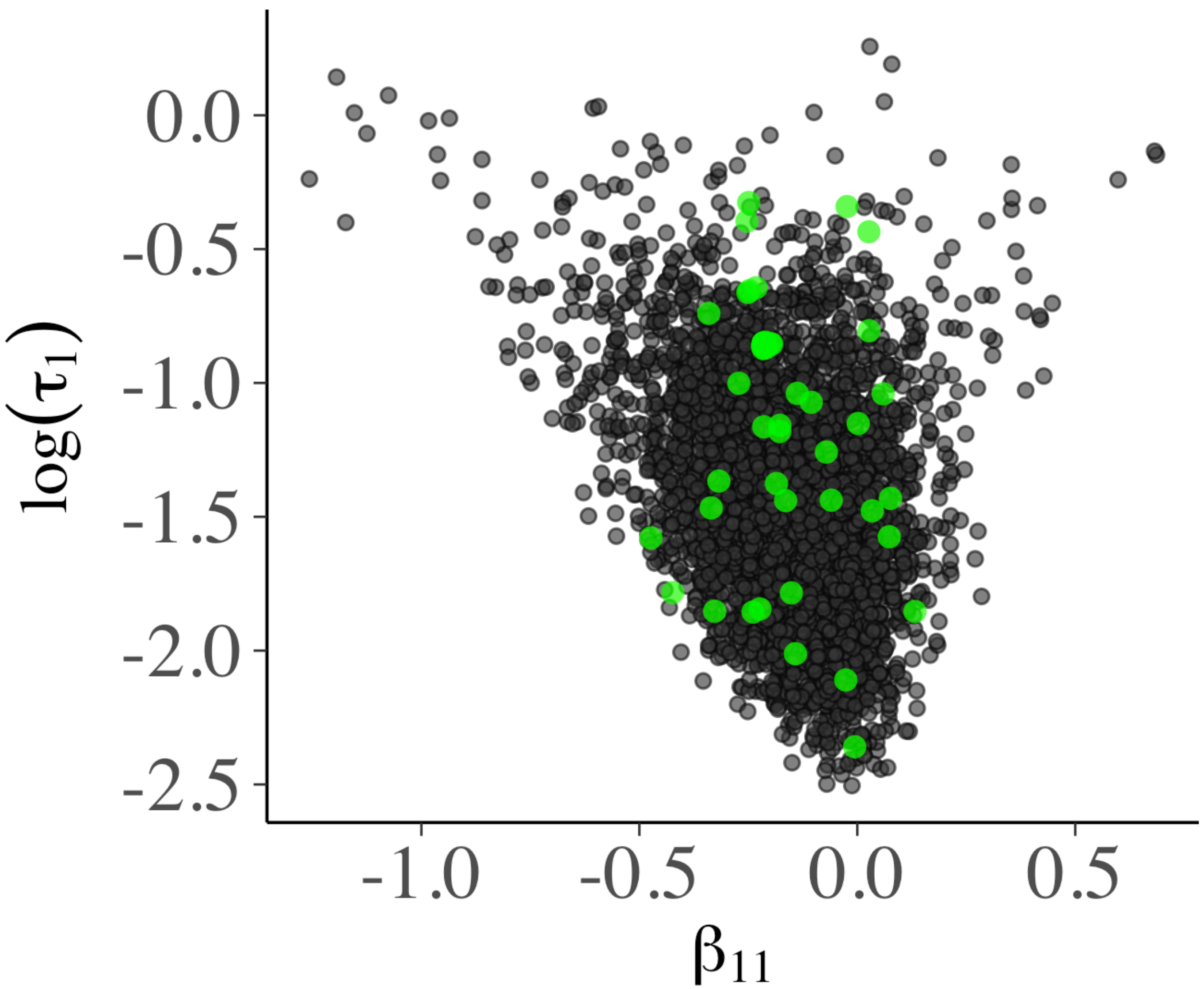


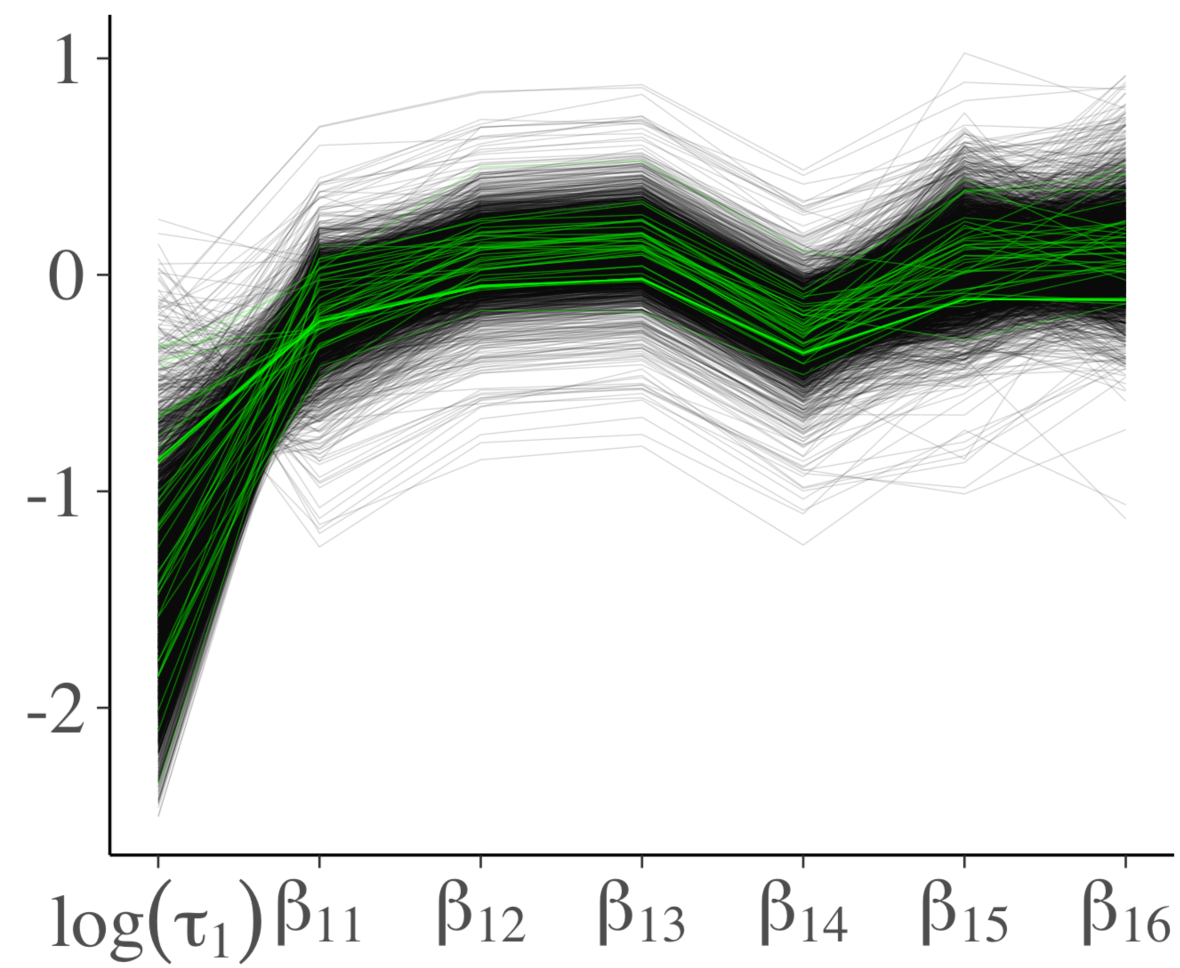
Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2018).
Visualization in Bayesian workflow.
Journal of the Royal Statistical Society Series A, accepted for publication.
arxiv.org/abs/1709.01449 | github.com/jgabry/bayes-vis-paper

Betancourt, M. (2017).
A conceptual introduction to Hamiltonian Monte Carlo.
arXiv preprint:
arxiv.org/abs/1701.02434

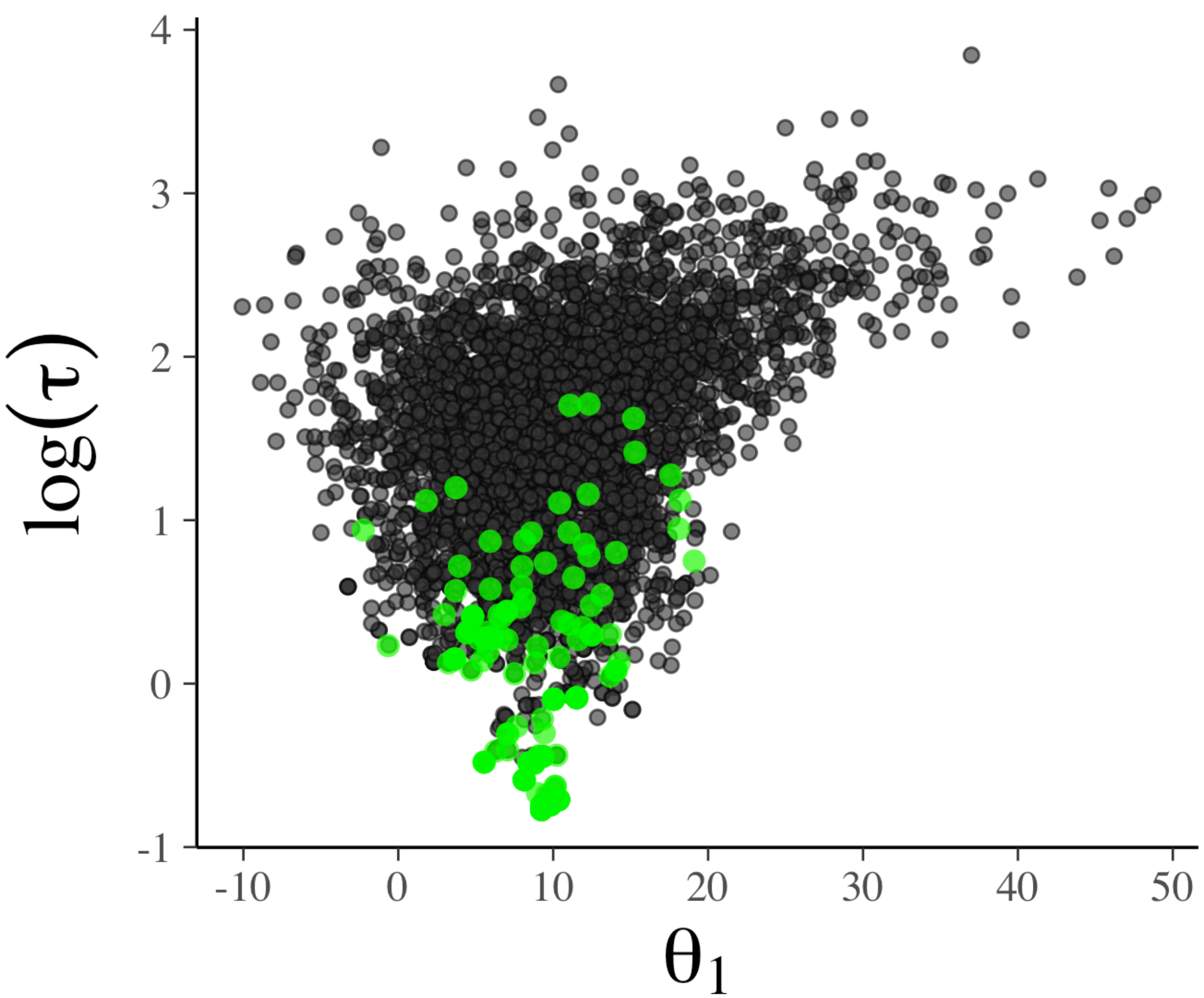
MCMC diagnostics

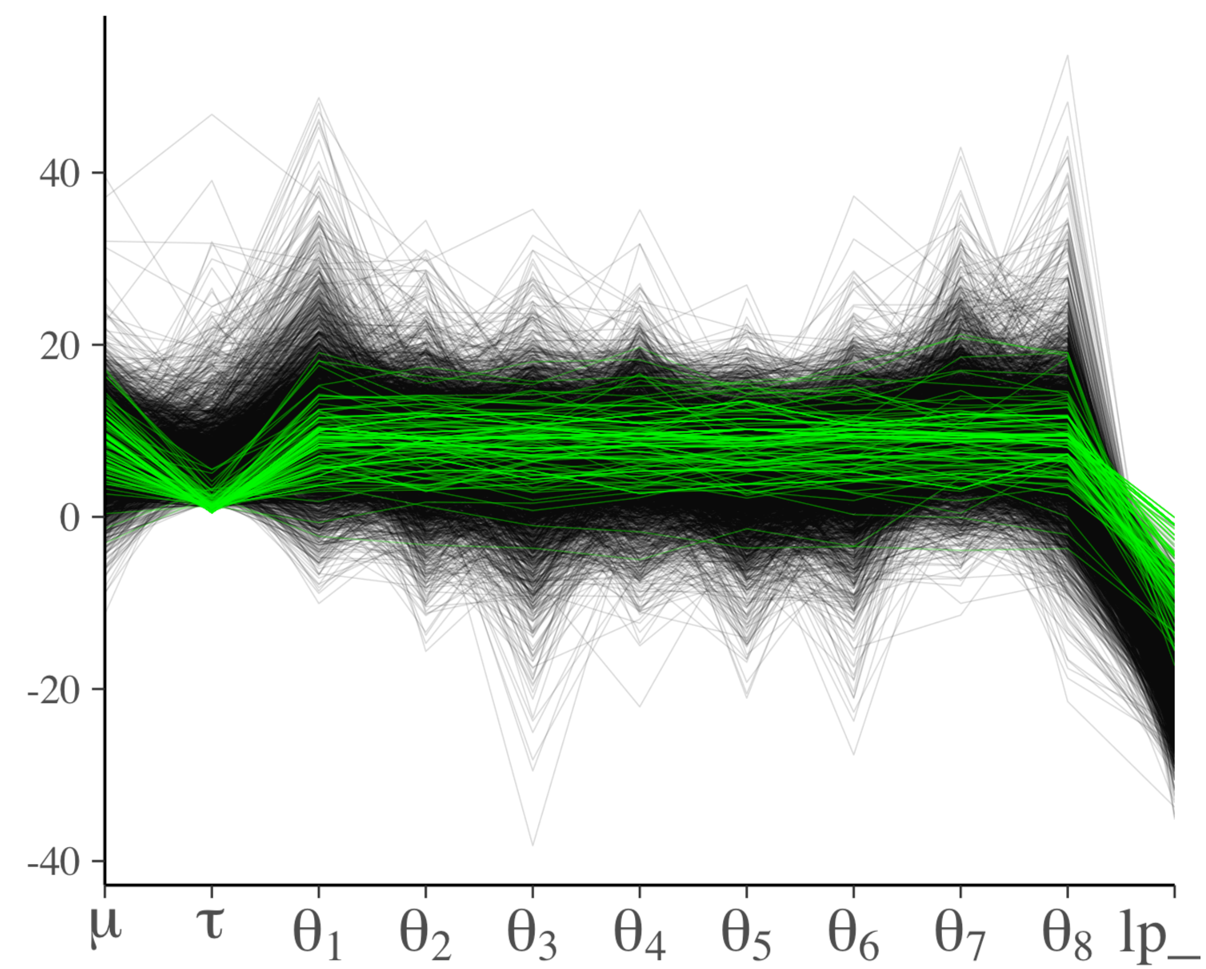
beyond trace plots



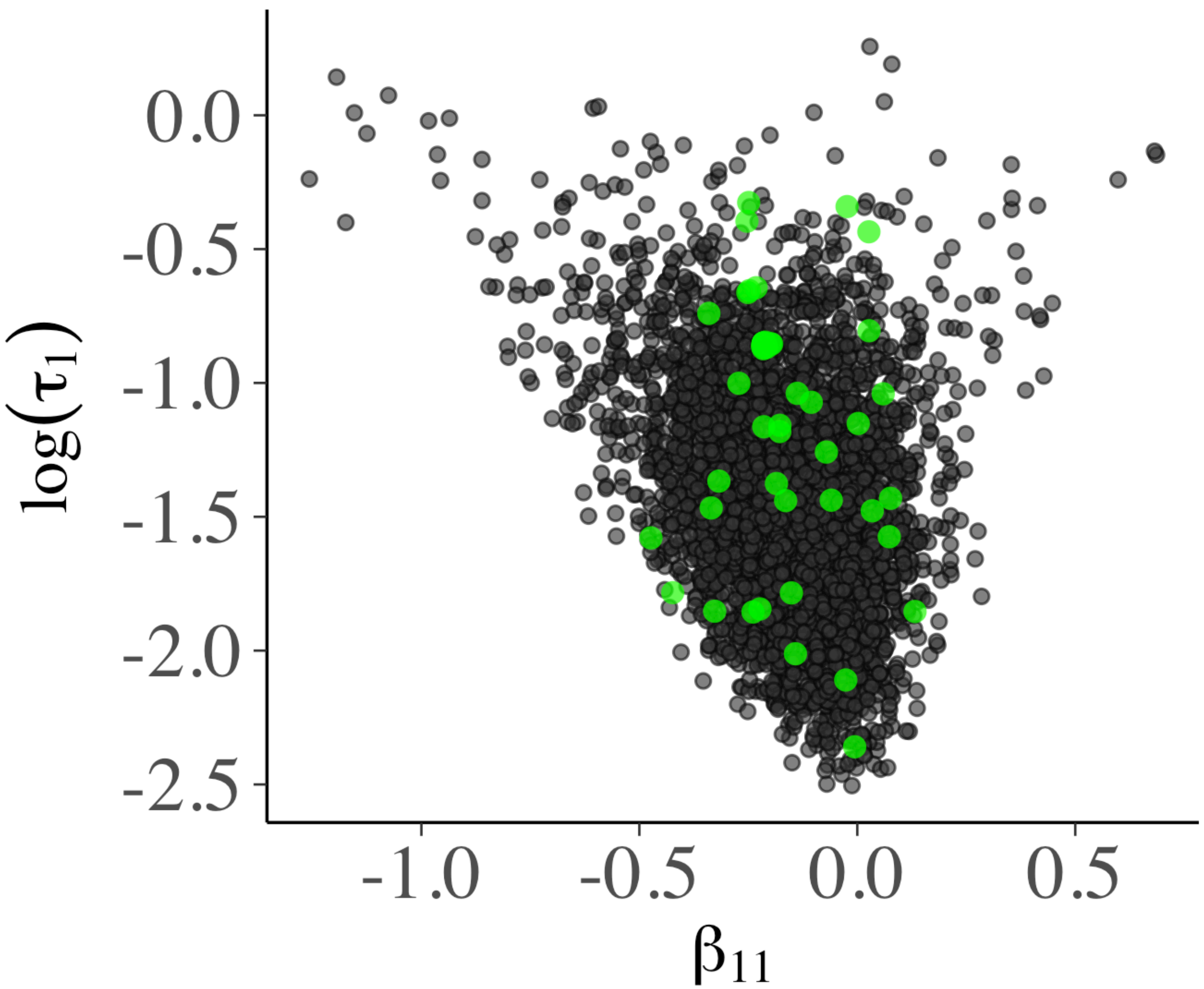


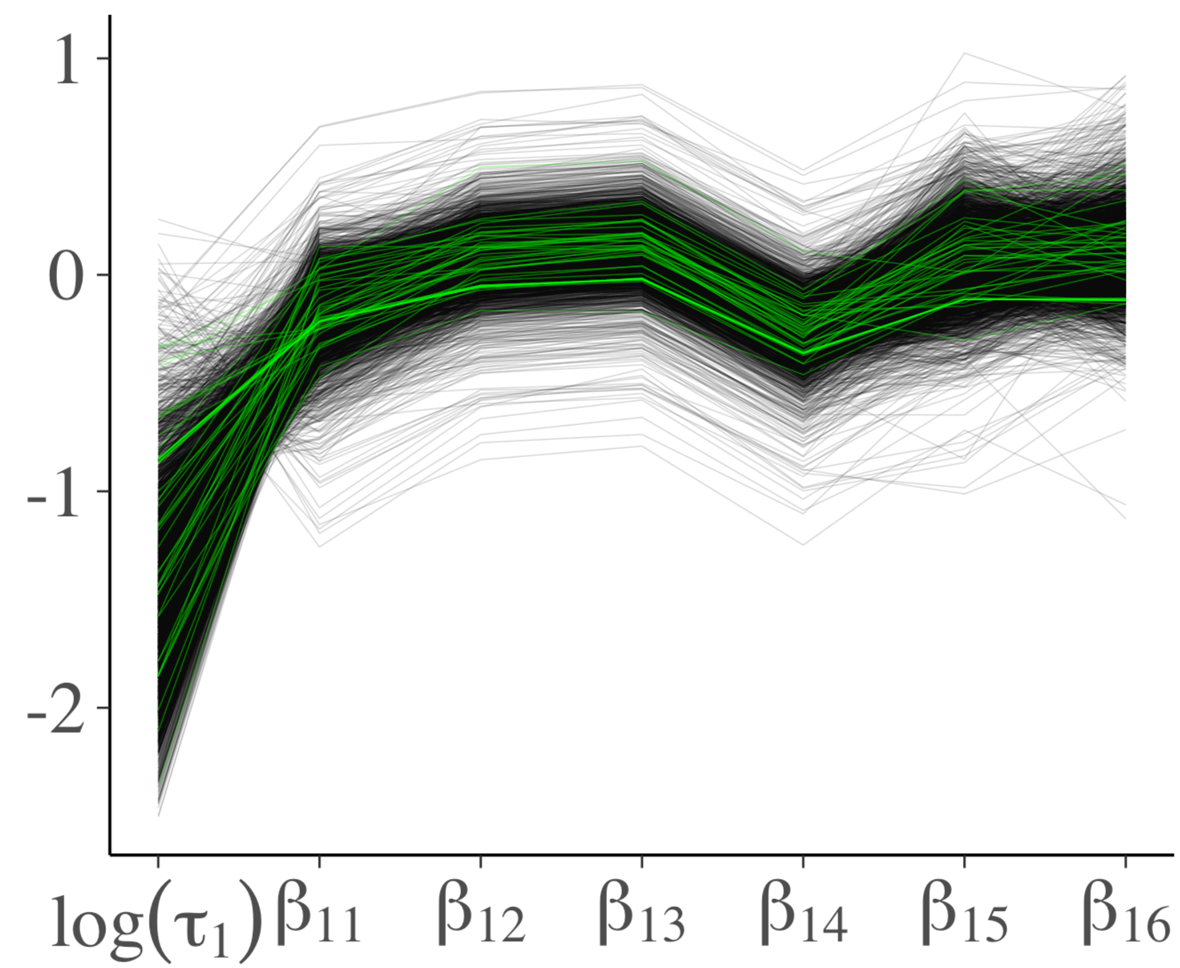
Pathological geometry





“False positives”





Posterior predictive checks

Visual model evaluation

Posterior predictive checking

visual model evaluation

The *posterior predictive distribution* is the average data generation process over the entire model

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)\, p(\theta|y)\, d\theta$$

Posterior predictive checking

visual model evaluation

- Misfitting and overfitting both manifest as tension between measurements and predictive distributions

- Graphical posterior predictive checks visually compare the observed data to the predictive distribution

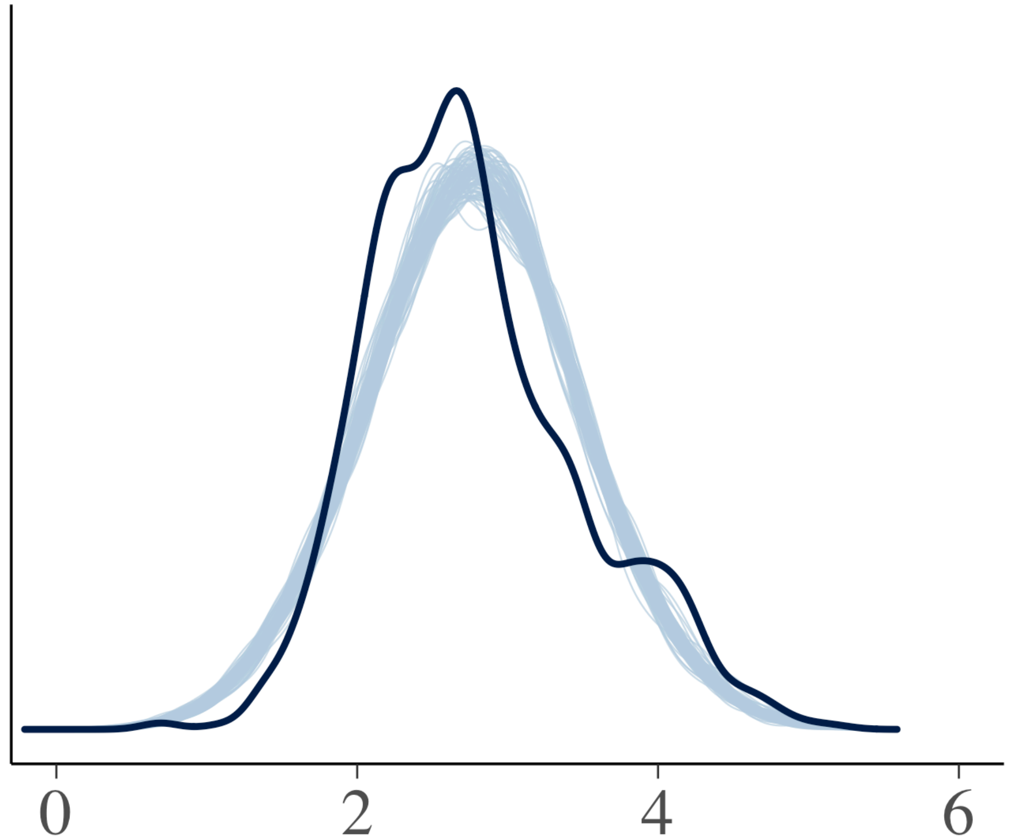
$$\begin{array}{c} \theta^\star \sim p(\theta|y) \\ \downarrow \\ \tilde{y} \sim p(y|\theta^\star) \end{array} \quad \Longleftrightarrow \quad \tilde{y} \sim p(\tilde{y}|y)$$

Posterior predictive checking

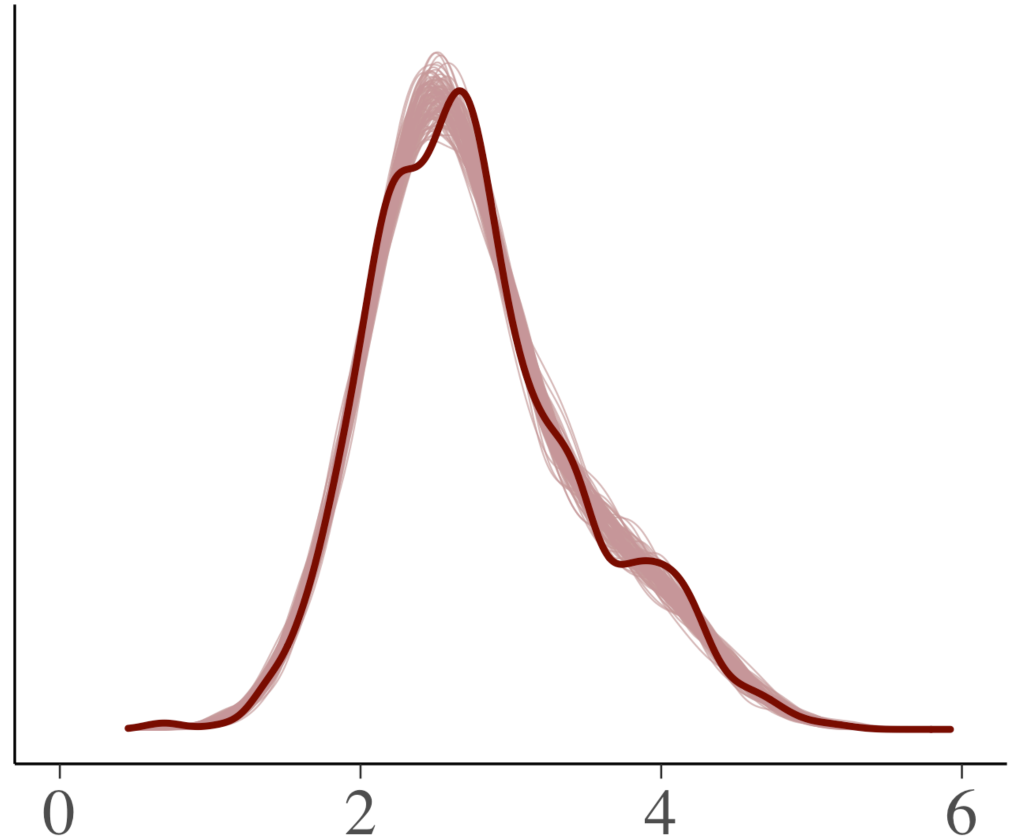
visual model evaluation

Observed data vs posterior predictive simulations

Model 1 (single level)



Model 3 (multilevel)

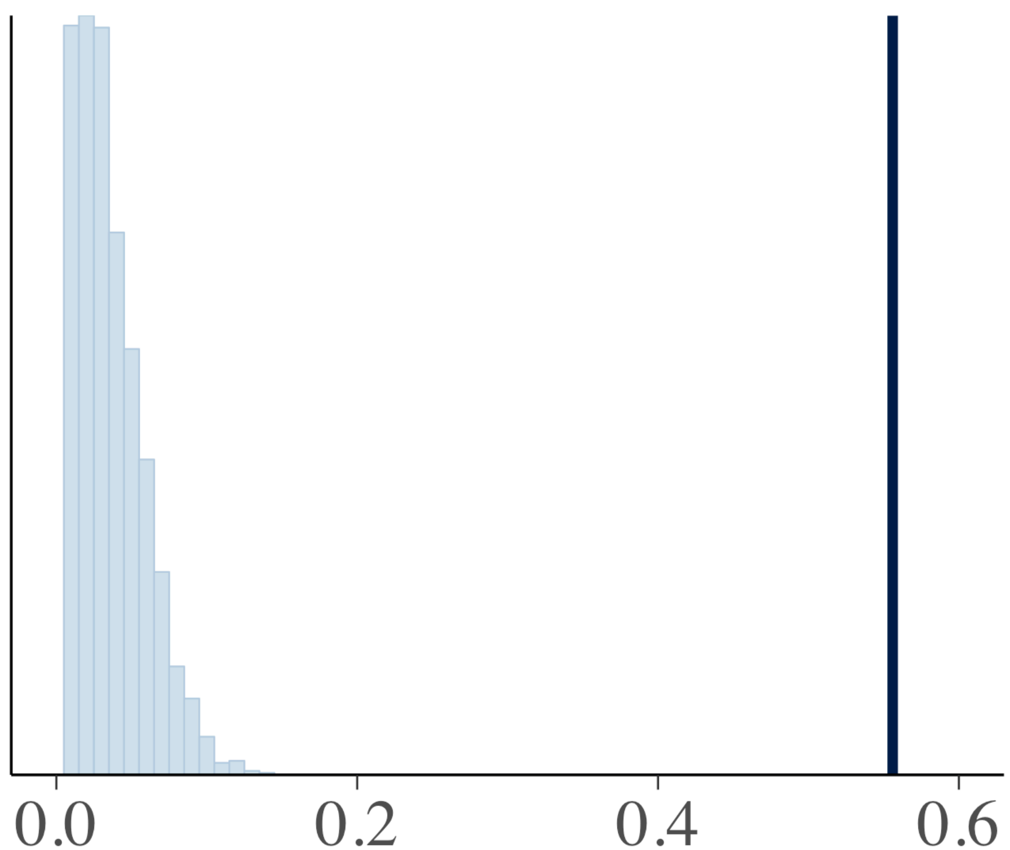


Posterior predictive checking

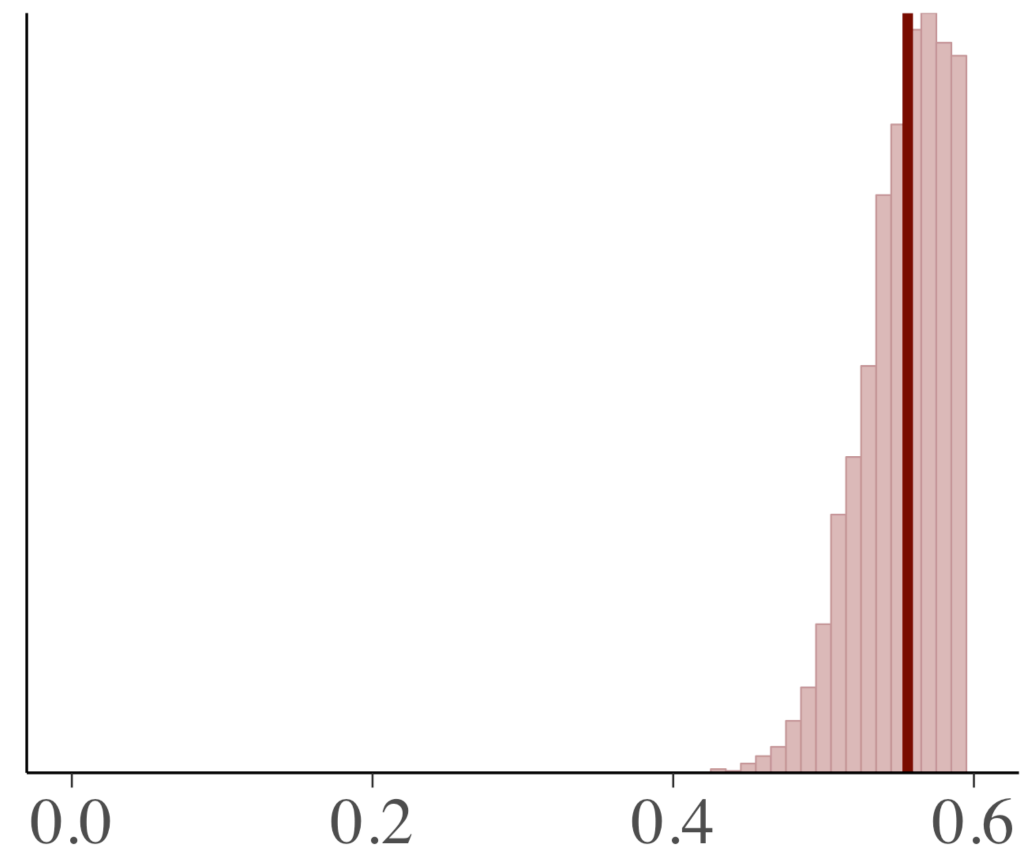
visual model evaluation

Observed statistics vs posterior predictive statistics

Model 1 (single level)



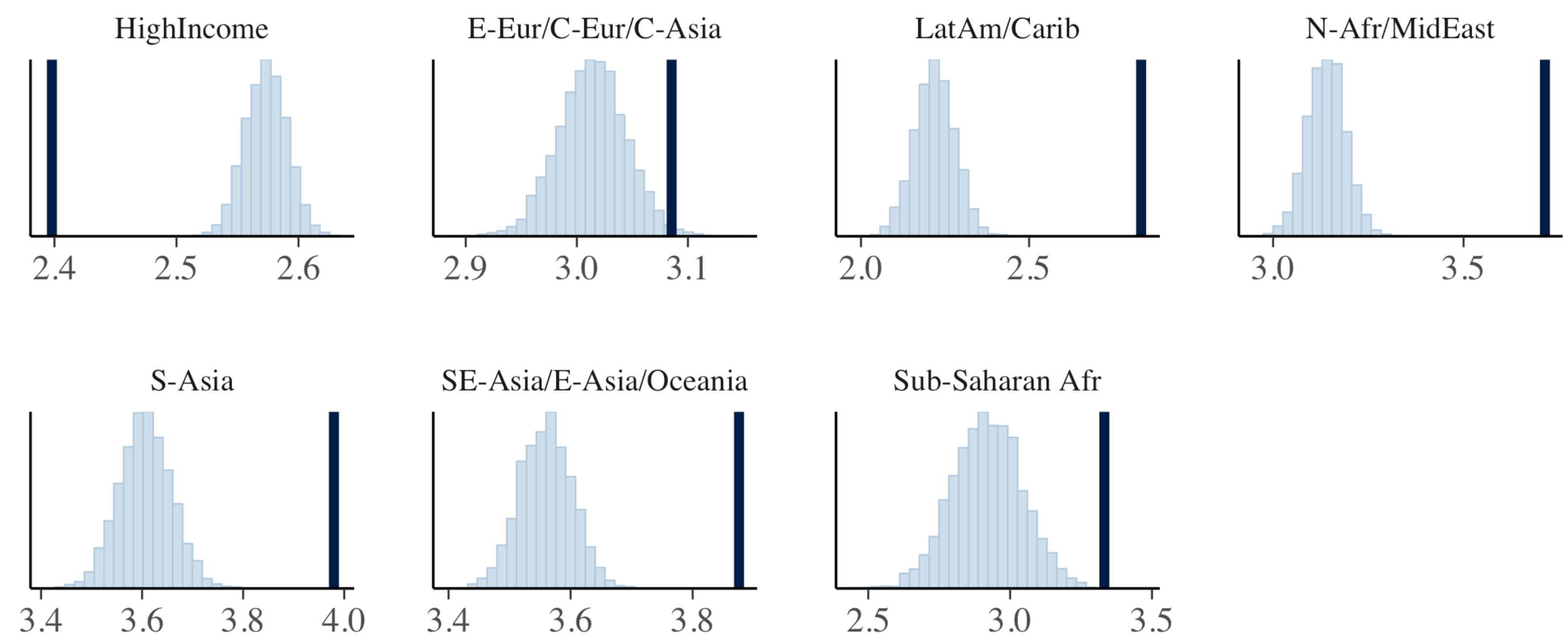
Model 3 (multilevel)



$$T(y) = \mathrm{skew}(y)$$

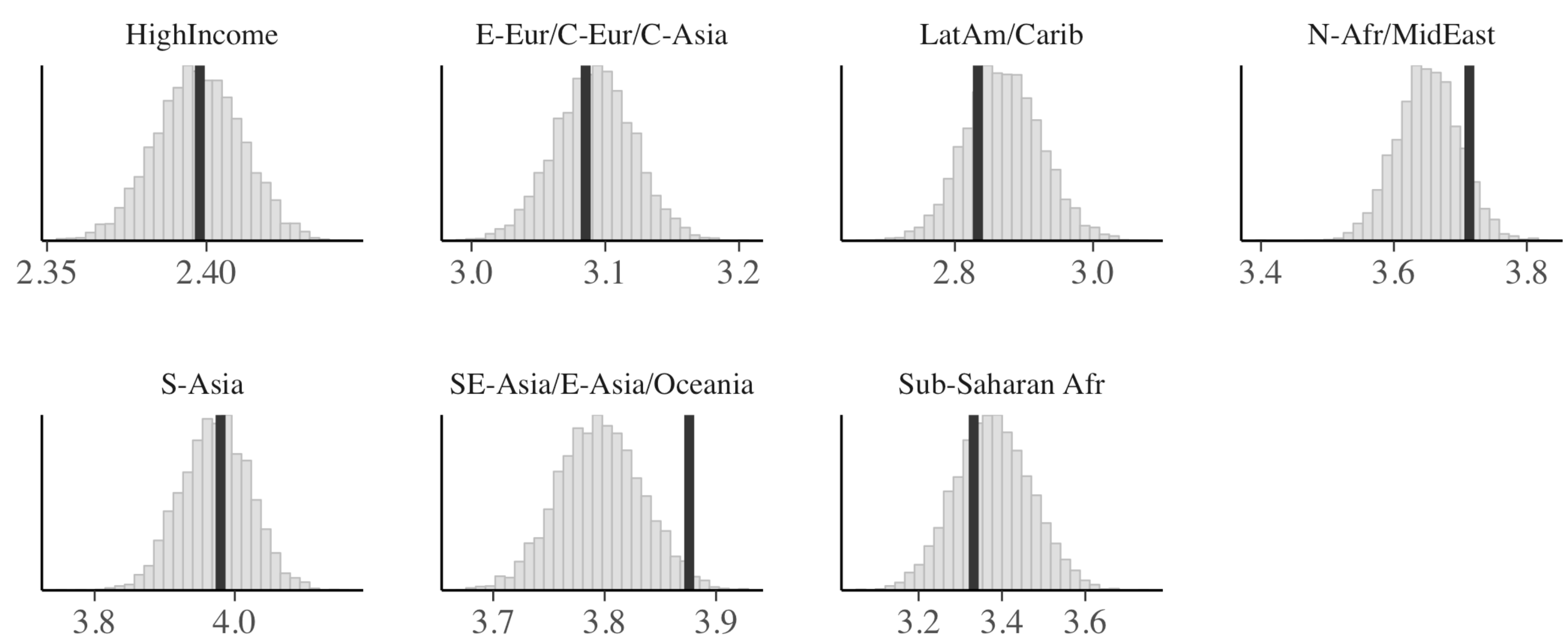
Posterior predictive checking:
visual model evaluation

Model 1 (single level)



$$T(y) = \mathrm{med}(y|\mathrm{region})$$

Model 2 (multilevel)



Model comparison

Pointwise predictive comparisons & LOO-CV

Model comparison

pointwise predictive comparisons & LOO-CV

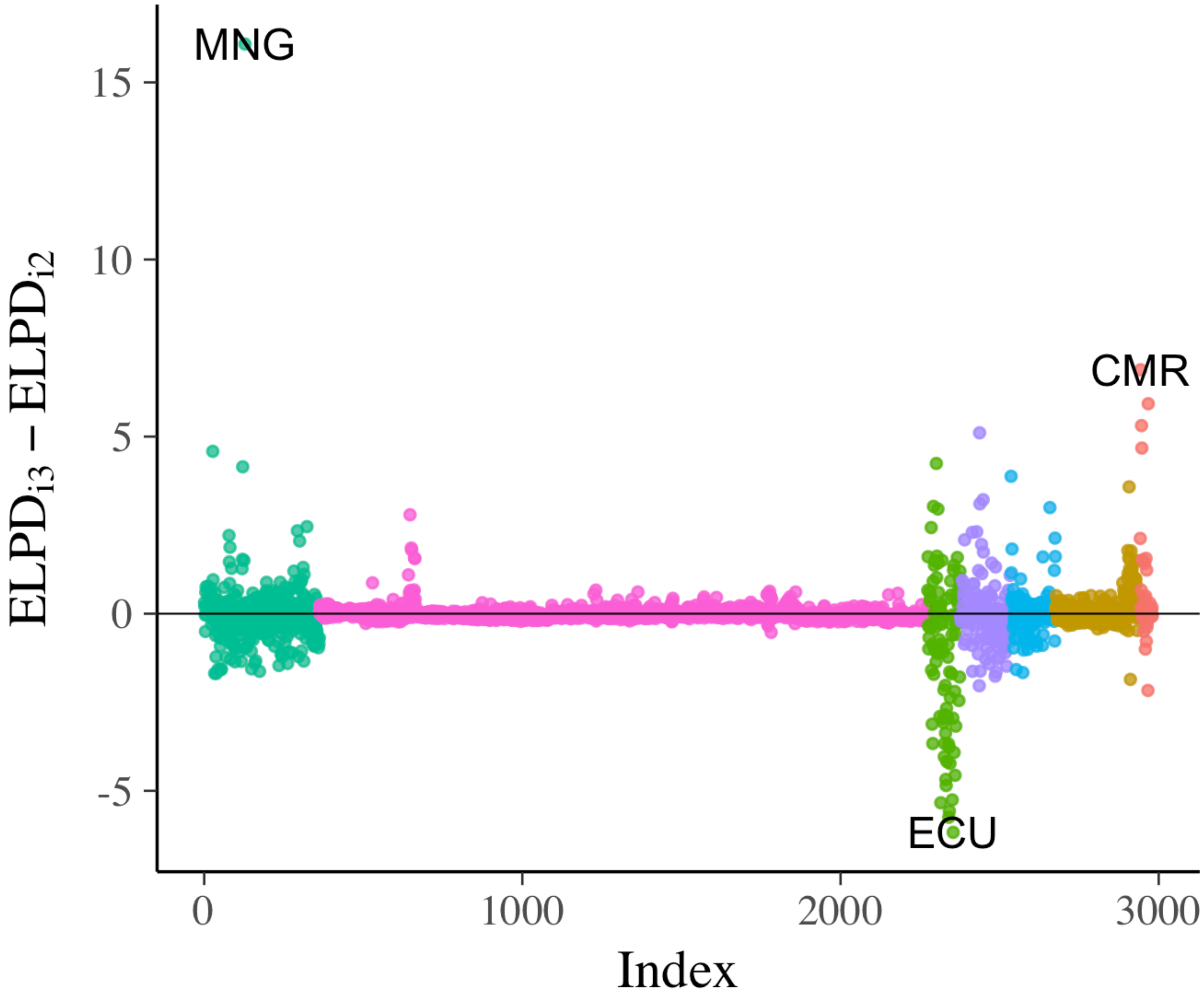
- Visual PPCs can also identify unusual/influential (outliers, high leverage) data points
- We like using cross-validated leave-one-out predictive distributions

$$p(y_i | y_{-i})$$

- Which model best predicts each of the data points that is left out?

Model comparison

pointwise predictive comparisons & LOO-CV



Model comparison
Efficient approximate LOO-CV

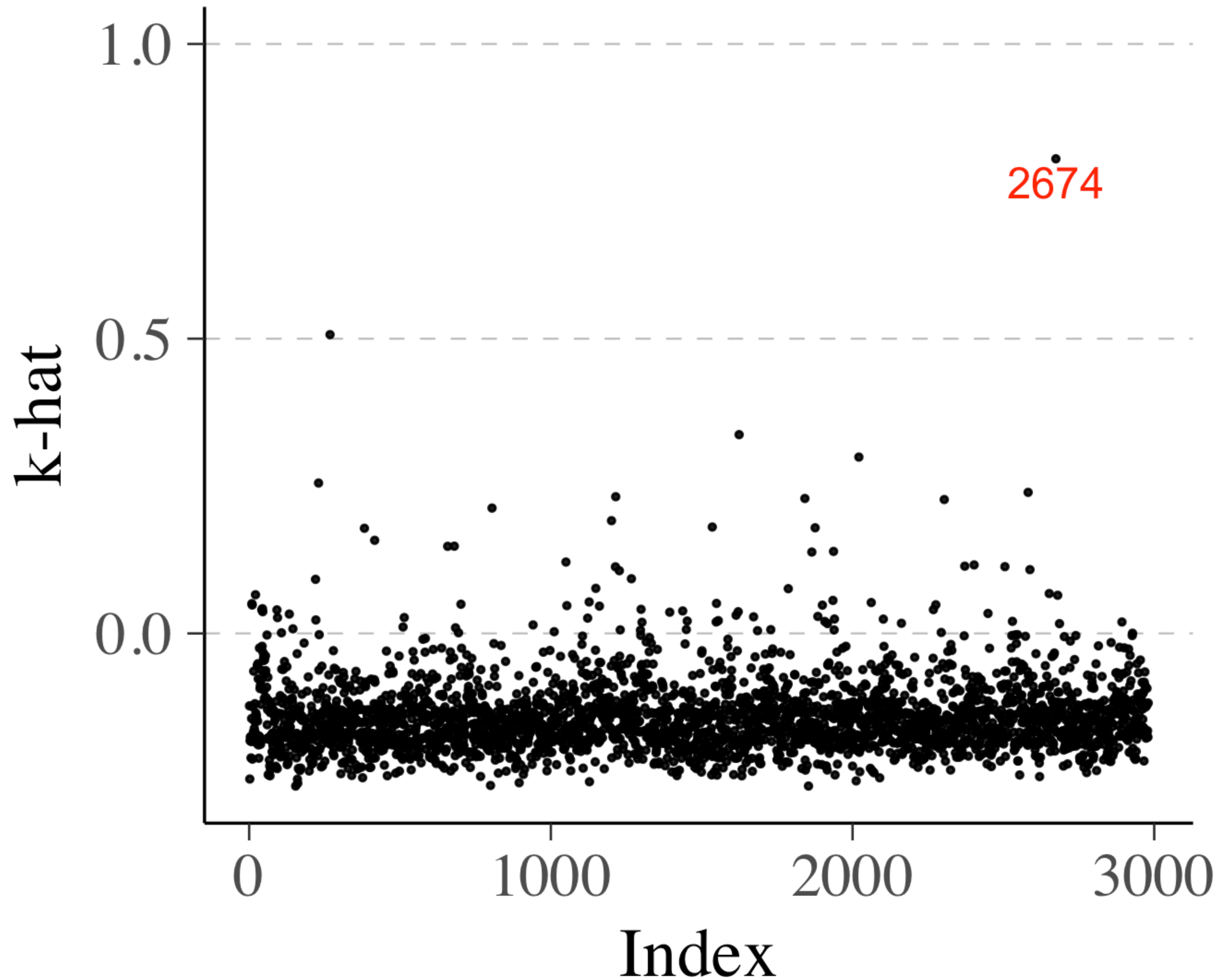
- How do we compute LOO-CV without fitting the model N times?
- Fit once, then use Pareto smoothed importance sampling (PSIS-LOO)
- Has finite variance property of truncated IS
- And less bias (replace largest weights with order stats of generalized Pareto)
- Assumes posterior not highly sensitive to leaving out single observations
- Asymptotically equivalent to WAIC
- Advantage: PSIS-LOO CV more robust + has diagnostics (check assumptions)

Vehtari, A., Gelman, A., and Gabry, J. (2017).
Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC.
Statistics and Computing. 27(5), 1413–1432.
doi: [10.1007/s11222-016-9696-4](https://doi.org/10.1007/s11222-016-9696-4)

Vehtari, A., Gelman, A., and Gabry, J. (2017).
Pareto smoothed importance sampling.
working paper
arXiv: arxiv.org/abs/1507.02646/

Diagnostics

Pareto shape parameter & influential observations



Thank You

bayesplot R package: **mc-stan.org/bayesplot**