LINEAR REGRESSION, PRIORS, AND MODEL SELECTION

September 08, 2019

Jon Fintzi

Biostatistics Research Branch
National Institute of Allergy and Infectious Diseases
National Institutes of Health

Dead and Company are Playing in Hampton, VA on 11/8

Nothing like Bayesian modeling to make you feel alive! Last time, we talked about:

- Bayesian inference *always* starts with a model for the **joint distribution** of θ and y:.

$$\pi(\theta, y) = f(y|\theta)\pi(\theta) = \pi(\theta|y)m(y).$$

- **Bayes rule** yields the **posterior distribution**

$$\pi(\theta|y) = \frac{f(y,\theta)}{m(y)} = \frac{f(y|\theta)\pi(\theta)}{m(y)} \propto Likelihood \times Prior.$$

.

- All of the information used in the *update* to our prior is encoded in the **likelihood**,

$$L(\mathbf{y}|\theta) = \prod_{i=1}^{N} f(y_i|y_{1,\dots,i-1}\theta).$$

- Re-Analysis of PREVAIL II data with non-conjugate priors:
  - Prior distributions for geometric mean log odds of 28 day mortality and ratio of the odds of death in 28 days for ZMapp vs. oSOC.
  - Posteriors are distributions of these *parameters* given the data, updated under binomial likelihood.
  - First look at `Stan`.

Lectures 3-4 of Statistical Rethinking

Briefly:

- Probability statements that describe key aspects of the data generating mechanism.
- Language for modeling:

$$
\begin{aligned}
y_i &\sim \mathrm{N}(\mu_i, \sigma^2), \\
\mu_i &= \beta_0 + \beta_1 x_{i,1} + \beta_p x_{i,p}, \\
\beta &\sim \mathrm{N}(0, 10^2), \\
\sigma &\sim \mathrm{Exponential}(1), \\
x_i &\sim \mathrm{N}(0, 1).
\end{aligned}
$$

- Here, the prior is full of lines, so too is the posterior is full of lines.
- Nothing special about lines, e.g., could use polynomials, splines, etc. Nothing special about linearity either.
- Bayesian framework takes a generative model with lots of uncertainty as input, learns which configurations of parameters are consonant with the data, and returns a generative model with (hopefully) less uncertainty.

Common Hangups

Probably the two most common hangups are

1. How to choose priors? What happens if the priors are “wrong”?
2. Small world model, big world data.

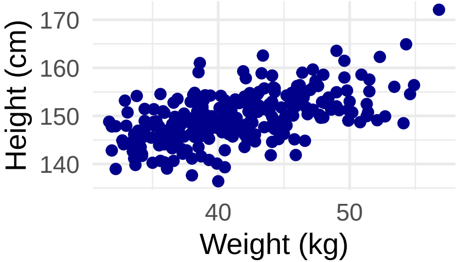
Today, we’ll talk about

1. How the shape or a prior affects the posterior.
2. Some principles for prior selection.
3. Good practices for workflow as they relate to model building and validation.

Example - Linear Regression for Height vs. Weight

(Similar to Statistical Rethinking, lecture 3) Suppose we want to understand how height predicts weight in a sample of 241 adult women.

- Simulate: $h_i = 150 + 0.55w_i + \epsilon_i,\ \epsilon_i \sim N(0, 4.5^2)$.
- Straightforward, probably won't matter much what we do.



Example - Linear Regression for Height vs. Weight

Model with "non-informative" priors[1]:

$$\begin{aligned} h_i &\sim Normal(\mu_i, \sigma^2) \\ \mu_i &= \alpha + \beta(w_i - \bar{w}) \\ \alpha &\sim Normal(160, 160^2) \\ \beta &\sim Normal(0, 100^2) \\ \sigma &\sim Half - Cauchy(10) \end{aligned}$$

I'll fit this using the `brms` package, which generates `Stan` code, compiles, and fits the model. Take a look at the raw Rmarkdown file if you're interested in seeing the code.

[1] Stupid priors.

Example - Linear Regression for Height vs. Weight

Prior predictive draws are clearly absurd, but it's a really simple model with a fair amount of data so let's just pretend to not care.

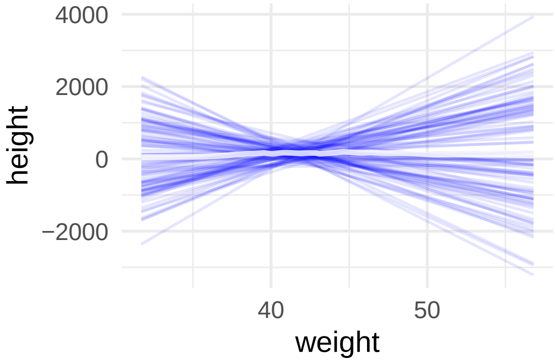


Figure 1: Prior distribution

Example - Linear Regression for Height vs. Weight

But, the posterior seems to shake out OK.

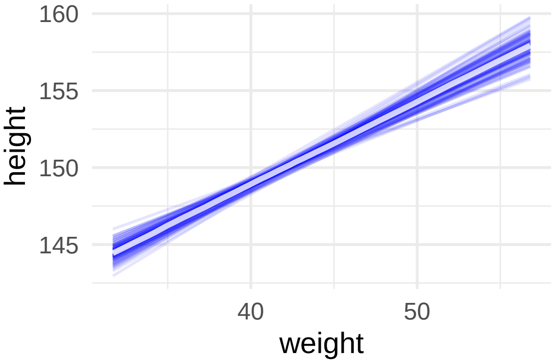


Figure 2: Posterior distribution of regression lines.

Example - Linear Regression for Height vs. Weight

Posterior distributions of model parameters seem reasonable and credible intervals contain the true values.

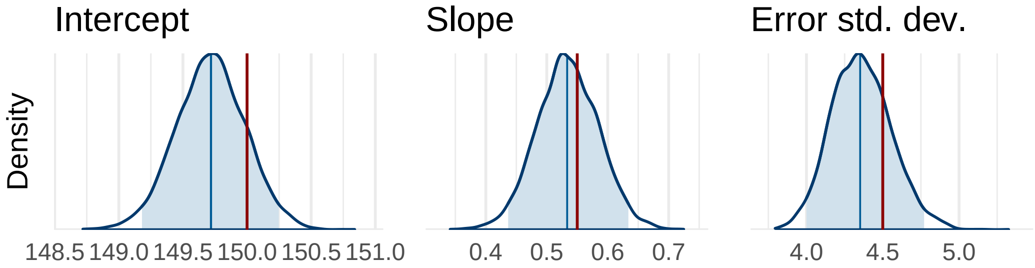


Figure 3: Parameter posteriors and 95 percent credible intervals.

Example - Linear Regression for Height vs. Weight

Observed data are contained within the posterior predictive distributions.

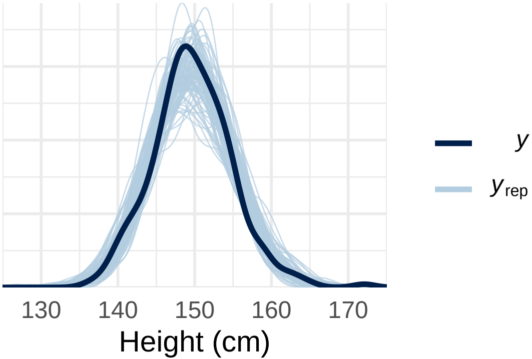


Figure 4: Density plots of observed heights and draws from the posterior predictive distribution.

Example - Linear Regression for Height vs. Weight

Leave-one-out posterior predictive distributions (`loo` package, more here) indicate good predictive performance.

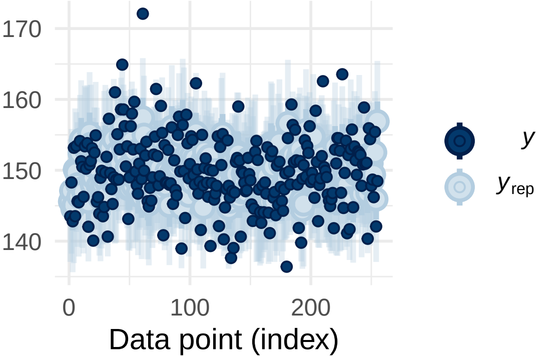


Figure 5: Observed heights and leave-one-out 95 percent posterior predictive intervals.

We Escaped the Zombies but They Ate the Dog

In this case, we had a lot of data to estimate a fairly strong signal.

Bernstein-von Mises (BvM) theorem: under some conditions, the posterior will look asymptotically like the sampling distribution of a maximum likelihood estimator, i.e., multivariate normal with mean at the true population parameter, $\boldsymbol{\theta}_0$, and covariance matrix $\boldsymbol{\Sigma} = \frac{1}{n} I(\boldsymbol{\theta}_0)^{-1}$.

- See Section 2.2.5 here for a technical discussion.
- Related paper by Charlie Geyer on “no-n” asymptotics of MLEs, available here.
- Gelman 2017 (see refs) on prior selection.
- From the Wiki, quoting A. W. F. Edwards, “It is sometimes said…that the choice of prior distribution is unimportant in practice…when there are moderate amounts of data. The less said about this ‘defence’ the better.”

We Escaped the Zombies but They Ate the Dog

Dan Simpson summarizes the problem in an epic rant, Asymptotically we're all dead:

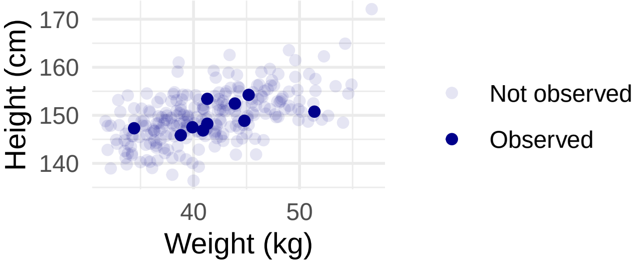
- There are some important assumptions needed for BvM:
    1. The MLE is consistent for the true population parameter.
    2. The model has a fixed, finite number of parameters.
    3. The true parameter θ_0 lies on the interior of the parameter space.
    4. The prior must be non-zero in a neighborhood around θ_0.
    5. The log-likelihood must be smooth.
- Incredibly difficult to apply BvM in practice.
    - Need independent replications of the same experiment, not enough to just have a lot of data.
    - Assumptions unlikely to hold in settings where we'd want to use penalized estimators or when we have an infinite dimensional parameter.
    - Most datasets are not instantaneous snapshots of a stationary process.

Moral of the story: The zombies could have bitten Fido and you'd never know until it's too late.

Poorly Informed Regression

Suppose we only had height-weight measurements for 5 women instead of on 241. How will the prior affect our inferences?

- Obviously, we expect a ton of uncertainty in the posterior. Might seem like a silly example, how far can we really get with only ten measurements?
- The things that go wrong in this setting are exactly what can go wrong without our realizing it in more complex settings.
- We are going through this exercise to understand the failure modes of different priors when the priors are not properly calibrated to the scale of the data.
- When the model fails, how does it fail?



Poorly Informed Regression: Analysis with "Non-informative" Priors

The posterior contains associations that don't make sense.

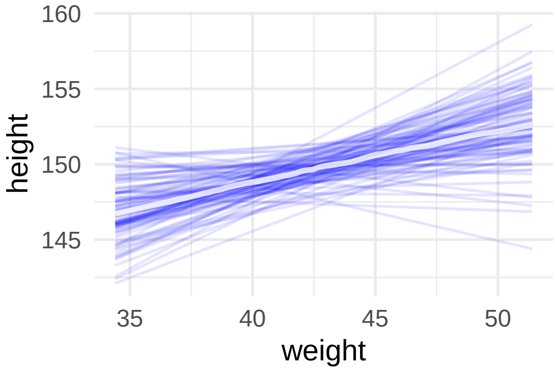


Figure 6: Posterior distribution of regression lines.

Poorly Informed Regression: Analysis with “Non-informative” Priors

Leave-one-out posterior predictive distributions (`loo` package) indicate poor predictive performance.

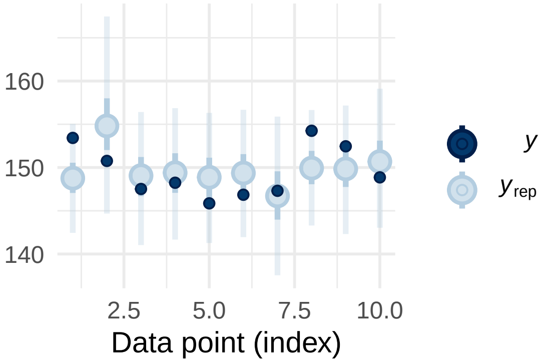


Figure 7: Observed heights and leave-one-out 95 percent posterior predictive intervals.

Poorly Informed Regression: Analysis with “Non-informative” Priors

Failure mode of diffuse priors in weak data settings:

- Posterior contains implausible values.
- Poor predictive performance.

Uncontroversial opinions:

- We could have ruled out negative associations and extreme associations by choosing coherent priors.
- We should still have a lot of uncertainty, no reason to pretend that we have strong information. It would be fine for the prior to be dominant in the posterior.
- Our uncertainty should, at least, be constrained to coherent ranges that could plausibly have produced the data.
- For more examples, see Gabry (2019).

Weakly informative priors

Basic idea: Introduce scale information, e.g., about order of magnitude or signs of parameters, in order to regularize inferences.

- Well defined units and meaningful parameterizations.
- Does not necessarily leverage full domain-specific knowledge.
- Requires an understanding of how the likelihood, prior, and data interact in the *joint* model.
- In our example, weakly informed prior on average height for a person of average weight and on slope s.t. encode a positive association and unlikely to observe someone more extreme than the shortest and tallest people in the world.
- Another example, `RStanArm` puts a weakly informative prior on transformed parameters in linear regression via a QR decomposition of the design matrix, see here.

Poorly Informed Regression: Analysis with Weakly Informative Priors

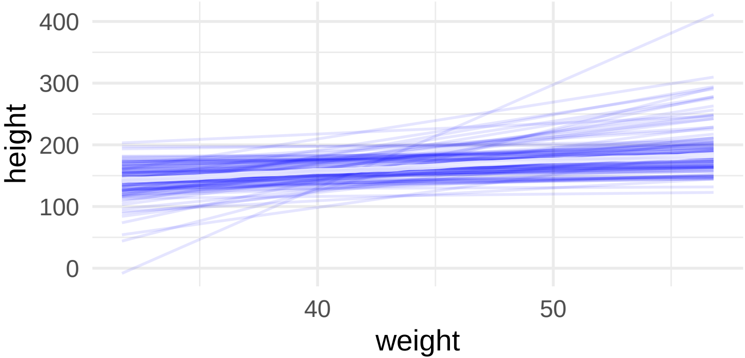
Model with weakly informative priors [2]:

$$
\begin{aligned}
h_i &\sim Normal(\mu_i, \sigma^2) \\
\mu_i &= \alpha + \beta(w_i - \bar{w}) \\
\alpha &\sim Normal(160, 20^2) \\
\beta &\sim LogNormal(0, 1) \\
\sigma &\sim HalfNormal(5)
\end{aligned}
$$

[2] Be a good Bayesian and look some stuff up: height and weight.

Poorly Informed Regression: Analysis with Weakly Informative Priors

Prior distribution:



Poorly Informed Regression: Analysis with Weakly Informative Priors

Much more sensible posterior.

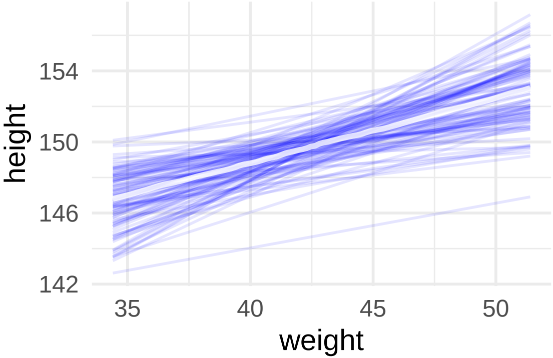


Figure 8: Posterior distribution of regression lines.

Poorly Informed Regression: Analysis with Weakly Informative Priors

Posterior distributions of model parameters seem reasonable, not crazy wide.

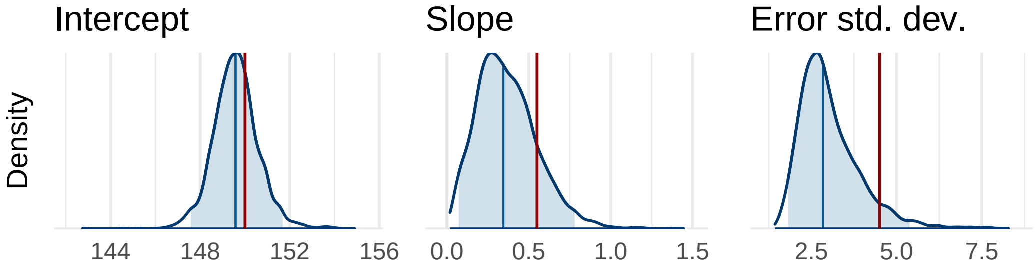


Figure 9: Parameter posteriors and 95 percent credible intervals.

Poorly Informed Regression: Analysis with Weakly Informative Priors

Leave-one-out posterior predictive distributions are still wide, but perform better ($ELPD_{WI} - ELPD_{NI} \approx 0.4 \pm 0.2$).

```
## elpd_diff        se
##      -0.3       0.1
```

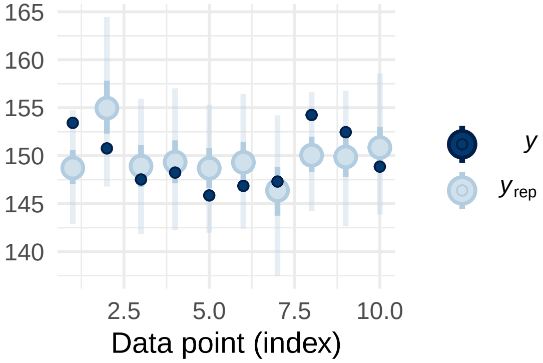


Figure 10: Observed heights and leave-one-out 95 percent posterior predictive intervals.

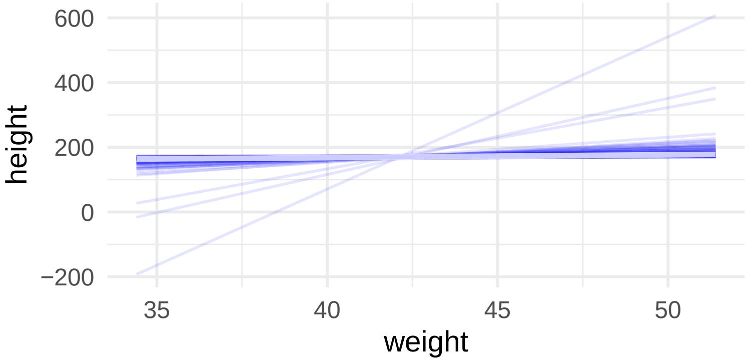
Poorly Informed Regression: Failure Mode of Light Tailed Priors

What if our prior is too tight and we get the location wrong?

$$\begin{aligned} h_i &\sim Normal(\mu_i, \sigma^2) \\ \mu_i &= \alpha + \beta(w_i - \bar{w}) \\ \alpha &\sim Normal(170, 2.5) \\ \beta &\sim LogNormal(0, 1.25) \\ \sigma &\sim HalfNormal(5) \end{aligned}$$

Poorly Informed Regression: Failure Mode of Light Tailed Priors

Prior distribution:



Poorly Informed Regression: Failure Mode of Light Tailed Priors

Nothing super glaring here?

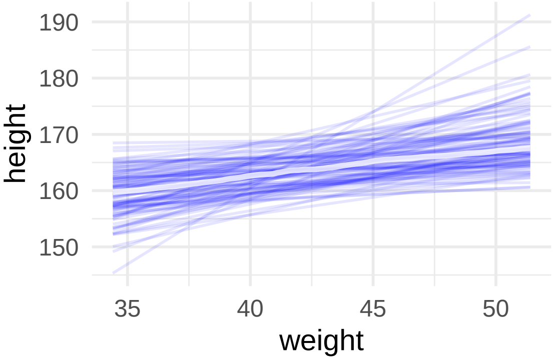


Figure 11: Posterior distribution of regression lines.

Poorly Informed Regression: Failure Mode of Light Tailed Priors

Ugh oh. Notice how the posterior doesn't contract relative to the prior.

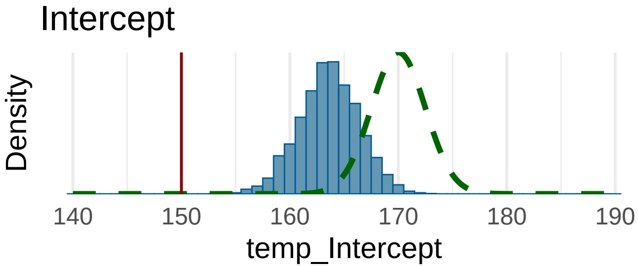


Figure 12: Parameter posteriors and 95 percent credible intervals.

Poorly Informed Regression: Failure Mode of Light Tailed Priors

Systematic bias in leave-one-out predictive densities is bad news bears!

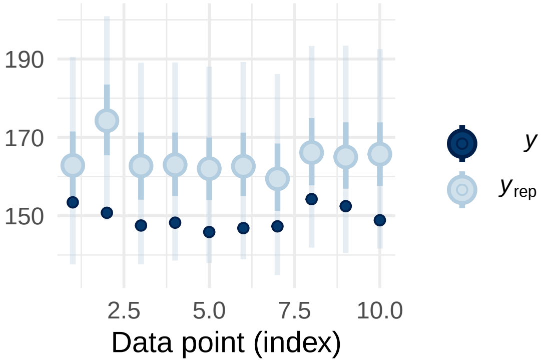


Figure 13: Observed heights and leave-one-out 95 percent posterior predictive intervals.

Poorly Informed Regression: Failure Mode of Heavy Tailed Priors

What if our prior is too diffuse and we get the location wrong?

$$
\begin{aligned}
h_i &\sim Normal(\mu_i, \sigma^2) \\
\mu_i &= \alpha + \beta(w_i - \bar{w}) \\
\alpha &\sim Cauchy(170, 2.5) \\
\beta &\sim LogNormal(0, 1.25) \\
\sigma &\sim HalfNormal(5)
\end{aligned}
$$

Poorly Informed Regression: Failure Mode of Heavy Tailed Priors

Posterior now contracts around the true value, but if we were to inspect more closely we'd see that it's also leaking mass out into the tails.

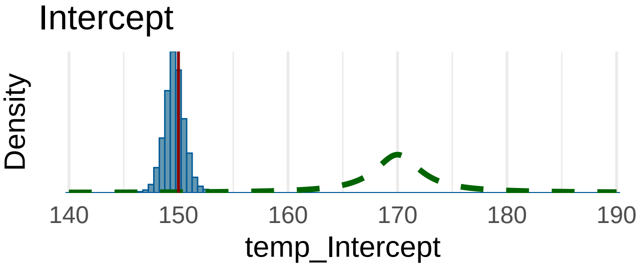


Figure 14: Parameter posteriors and 95 percent credible intervals.

Poorly Informed Regression: Failure Mode of Heavy Tailed Priors

Systematic bias in leave-one-out predictive densities is now gone. Huzzah!

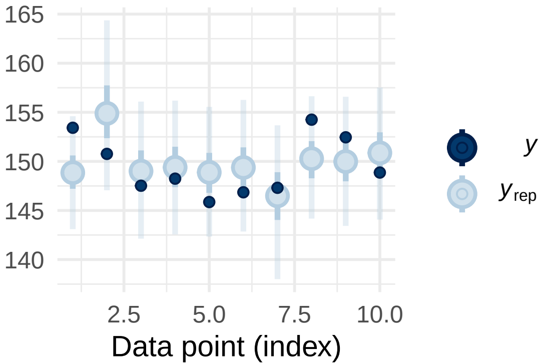


Figure 15: Observed heights and leave-one-out 95 percent posterior predictive intervals.

Recapping

Some key takeaways:

- We went through an iterative process to evaluate how different priors act on the posterior.
    1. Interrogate prior.
    2. Fit model.
    3. Criticize model.
    4. Wash-rinse-repeat.
- We could have done more at each step, but we only had an hour. Read about robustifying Bayesian workflow here.
- Incorporating information about *scales* can help regularize weakly informed models.
- But remember, our goal is to identify scales that are consistent with our prior beliefs, not exact values. - **Big warning:** Not necessarily a good idea to look at your data and pick the prior to look like it. We don't know as much as we think. Remember, it was hubris that killed the king.
- Not the only way to come up with priors, but thinking generatively about how parts of the model interact can help diagnose subtle issues that would otherwise have gone unnoticed.

References

M. Betancourt "How the Shape of a Weakly Informative Prior Affects Inferences." `https://betanalpha.github.io/assets/case_studies/weakly_informative_shapes.html` (2017).

J. Gabry, et al. "Visualization in Bayesian workflow." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182.2 (2019): 389-402.

A. Gelman, S. Simpson, and M. Betancourt. "The prior can often only be understood in the context of the likelihood." *Entropy* 19.10 (2017): 555.

C.J. Geyer. "Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity." *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*. Institute of Mathematical Statistics, 2013. 1-24.