Applied Probabilistic Programming & Bayesian Machine Learning

§1-4: Identifiability (Part 1)

MIT IAP 2017 — Max W. Shen

At this point, you should be fairly comfortable with Stan code, enabling discussion on a more abstract level.

The **three steps of Bayesian data analysis** are:

1. Write a model
2. Obtain posteriors by running inference
3. Check the model, and repeat from step 1 if needed

Thus far, we have loosely covered topics 1 and 2 for basic Bayesian models using Stan. Topic 3 becomes crucial when working with real data, the domain where all models are wrong but only some are useful. This crucial step will be the focus of the next lecture.

The remainder of the course will be spent on expanding the capabilities of Bayesian modeling (topic 1) by pushing Stan to its limits (topic 2), and addressing subtleties that arise during modeling. We briefly discuss two topics within this domain here.

1 Identifiability

In this folder, you should find model_1-4.py, model_1-4.R, and model_1-4.stan. All the code has been written for you - take a moment to read through the Stan code and the python or R code to understand the model it's describing.

Exercise 2

Write down the Stan model on paper, using statistical language (the squiggly tilde's).

Run the model.

This model may take longer to run than previous models (so if needed, go ahead and decrease the number of iterations). You'll get some output like this:

```
        mean se_mean     sd   2.5%    25%    50%    75%  97.5%  n_eff   Rhat
mu_a  228.26  1115.5 1932.1  -2794  -1355 162.06 1298.4 4187.7    3.0   2.39
mu_b  -228.6  1115.5 1932.1  -4188  -1298 -162.5 1354.7 2793.9    3.0   2.39
sigma   1.11  8.6e-3    0.2    0.8   0.97   1.08   1.22   1.56  520.0   1.01
lp__  -11.42    0.03   1.03 -14.28 -11.81  -11.1 -10.69 -10.42 1295.0    1.0
```

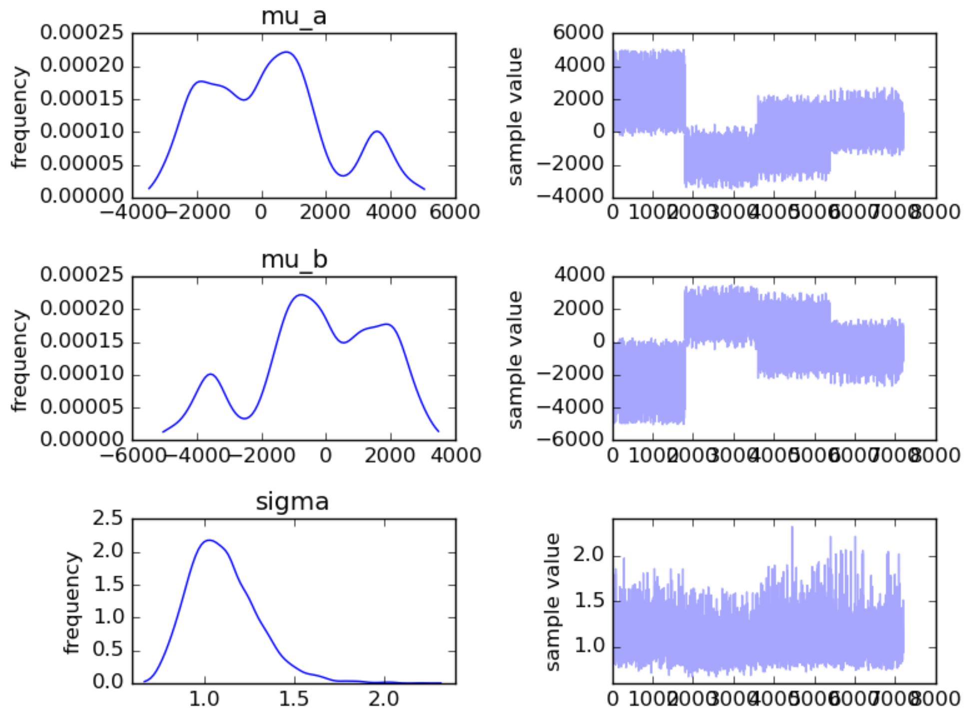

Figure 1: Unconstrained mu's

The first thing to notice is that n_eff and Rhat are very poor for *mu_a* and *mu_b*, suggesting incomplete convergence. We've told the model to find $mu\_a + mu\_b = 0$ which is roughly what we see, but there are infinite values that *mu_a* and *mu_b* can take on!

The right-side of the plotted fit, denoting the values explored by the chains during inference, shows that its exploration hasn't stabilized (as expected, since we have under-specified the model). Compare with the plot for sigma, which has nearly converged with Rhat = 1.01.

The core reason we have poor convergence is because our model has a problem with **identifiability**. There are infinitely plausible values for mu_a and mu_b. All we need is:

$$\mu_a + \mu_b = 0 \tag{1}$$

Consider figure 2 - there is a "ridge" in the posterior distribution of equally likely values of mu_a and mu_b. This ridge is characteristic of non-identifiability.

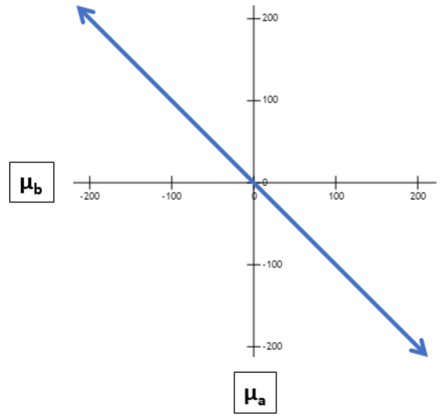

Figure 2: "Ridge" in the Posterior; Nonidentifiability

Since all values of mu_a and mu_b are equally likely, our true posterior distribution is completely flat from negative infinity to positive infinity, and we haven't given it nearly enough iterations to allow the inferred posterior to come close to this!

This is one danger of using a uniform prior, as we have implicitly used in model 1-4, the example above - it asserts that all regions of probability space should be explored equivalently.

Also note that identifiability issues can arise from periodicity. For example, the following model will also have identifiability issues:

```
model {
  y ~ N( sin(mu), sigma);
  mu ~ uniform(-500, 500);
}
```

Where uniform(a, b) sets a lower and upper bound, respectively. (See Stan reference for more details).

Identifiability issues arise frequently in mixture models, which will be discussed in a later section.

Models can be **weakly identifiable** if there exists strong correlation between parameters. The global optima may be unique for all parameters, but strong correlation between parameters will slow the speed of inference.

Addressing correlation will be discussed in a later section (Identifiability Part 2). In short, the discussed approach is to explicitly model the correlation in the model, for example by defining a multivariate Gaussian prior over parameters, and explicitly learning the covariance matrix.

2 Using Weakly-Informative Priors for Identifiability

In general, identifiability issues can be solved by specifying additional prior knowledge. One way to solve the identifiability issue in model_1.4 is to place a weakly informative prior over mu_a and mu_b, such as a normal(0, 100) prior.

Exercise 2

Solve the identifiability issue in model_1-4 by adding weakly informative priors.

This works for convergence because the parameters mu_a and mu_b now don't have literally infinite parameter space to explore, because the prior probability decays before infinity. The inference chains have been lightly constrained to explore a more sensible area of parameter space. However, if you change the mean of your prior, your posterior mean will also change!

Such a model, whose posteriors are dependent largely on the priors rather than the data, should not be trusted. An important part of model checking (the focus of the next lecture) is testing the sensitivity of models to assumptions baked into your choice of priors. A good model will be resilient to specific prior assumptions and instead draw most of its statistical strength from the data itself.

Inspecting the fit further emphasizes the weakness of this model:

```
        mean se_mean     sd   2.5%    25%    50%    75%  97.5%  n_eff   Rhat
mu_a    4.57    2.47  71.09 -138.9 -43.77   8.82  51.09 140.25  827.0    1.0
mu_b   -4.52    2.47  71.09 -140.2 -51.02  -8.69  43.87 138.81  827.0    1.0
sigma   1.06    0.01   0.19   0.78   0.92   1.04   1.17   1.51  334.0   1.01
lp__  -10.98    0.04   1.23 -14.32  -11.5 -10.65 -10.12  -9.58 1070.0   1.01
```

We now have successful convergence with Rhat = 1.0, but the posterior means of mu_a and mu_b at ± 4.57 are not centered at the global log likelihood optima, 0, as we would expect with perfect inference. Moreover, the standard deviation is huge at 71.09, a direct impact of our weakly informative prior. Finally, our n_eff is less than 10% of our total inference iterations, suggesting a poor model, to put it mildly.

3 In Practice...

In practice, it's better to rewrite the model to not have identifiability issues rather than rely on additional priors. We use weakly-informative priors here as a teaching tool, as an example to improve understanding on how weakly informative priors contribute to posterior inference.

In more complicated models, however, adding additional prior information is a powerful way to address identifiability when models cannot be rewritten in another way. This is discussed during mixture models in section 4.

4 Bibliography and Additional Resources

Identifiability is discussed in Chapter 8 of Statistical Rethinking.

The Stan reference is available here: http://mc-stan.org/documentation/