Fixing a finite mixture model with RStan

Jon Zelner

May 26, 2016

1 Introduction

This is meant as a short tutorial on how to fit a Stan model in a way that is fully reproducible using, R, Stan, Docker, Knitr, and Pandoc, and finally Make to orchestrate everything.

But along the way, this will also be a quick tutorial on how to fit a finite mixture model using Stan. In this case, we'll be simulating data from a 2-component Gaussian mixture, fitting a model written in Stan to the simulated data, plotting the results, and compiling it all into a PDF.

To generate new results, just go into the file `data/parameters.csv` and change a parameter value. When you type `make pdf` at the command line, the simulation model will run again with the new parameter values, parameters for the Stan model will be re-estimated, and figures (and this PDF) will be automatically regenerated.

2 Methods

First, a quick overview of the 2-component finite mixture model we'll be simulating from and fitting. The model has five parameters, μ_1 and μ_2 are the component means, and σ_1 and σ_2 are the standard deviations of components 1 and 2 respectively. Finally, p is the mixture probability, in this case the proportion of samples drawn from component 1.

The following table lists the parameters set in `data/parameters.csv`:

Table 1: Parameter values

parameter	value
m1	2.0
m2	10.0
sd1	1.0
sd2	3.0
p	0.5

To simulate from this model, we first draw a Bernoulli random variable with success probability p to select the component we'll be drawing from, i.e. $z_i \sim Bernoulli(p)$. So, $z_i = 1$, we'll sample x_i from $Normal(\mu_1, \sigma_1)$, otherwise $Normal(\mu_2, \sigma_2)$.

And here's what 10000 draws from this distribution looks like:

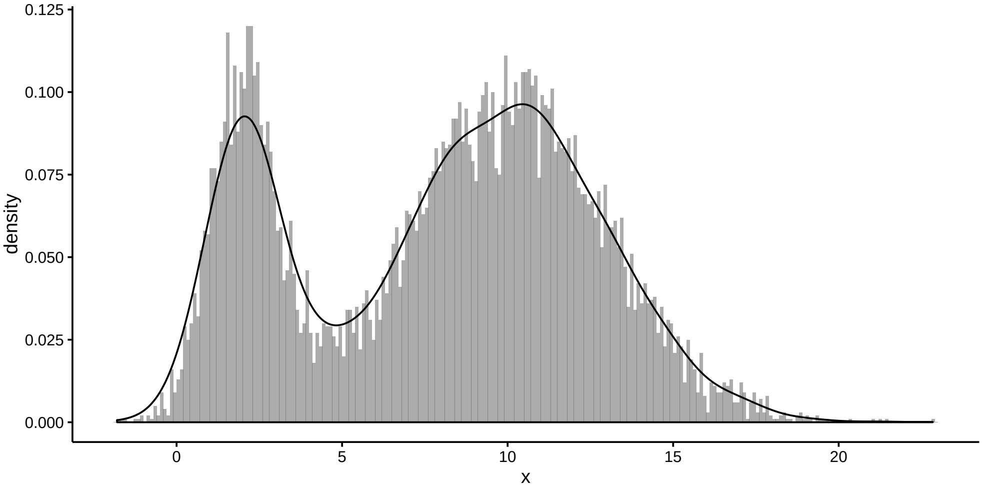

Figure 1: **Input data.**

To estimate the parameters of this model, we're going to use the marginalization approach to latent variables outlined in the Stan user's guide. We can think of each sample as being associated with a latent variable z_i indicates the mixture component it was sampled from. Stan allows us to easily marginalize out this variable using the full mixture likelihood. If we define $g(.|\mu, \sigma)$ to be the PDF of a Gaussian distribution, we can then calculate the likelihood of observing a sample x_i given the model parameters θ as follows:

$$Pr(x_i|\theta) = p \cdot g(x_i|\mu_1, \sigma_1) + (1 - p) \cdot g(x_i|\mu_2, \sigma_2) \tag{1}$$

3 Results

Figure 2 shows the posterior distributions of the component means as compared to the input values (indicated by dashed vertical lines):

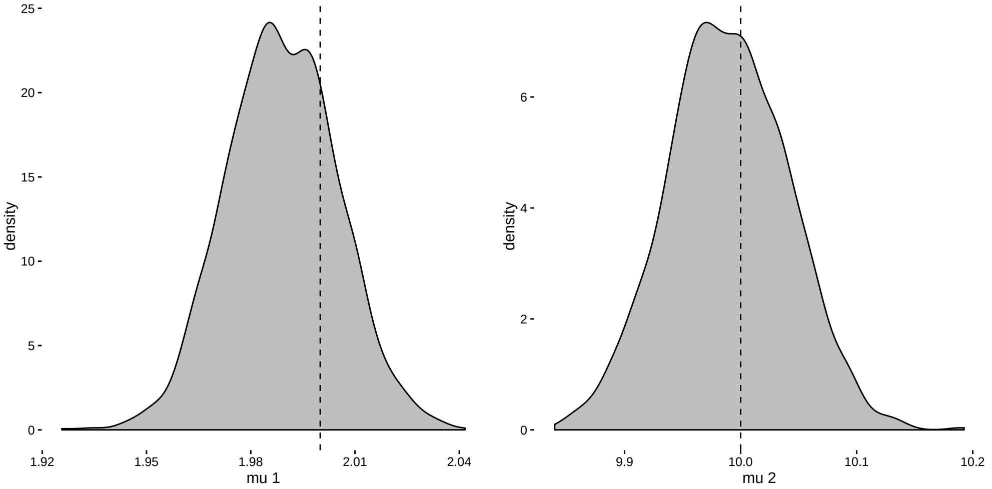

Figure 2: **Posterior distribution of component means**.

Figure 3 shows the posterior distributions of the component standard deviations as compared to the input values (indicated by a vertical dashed lines):

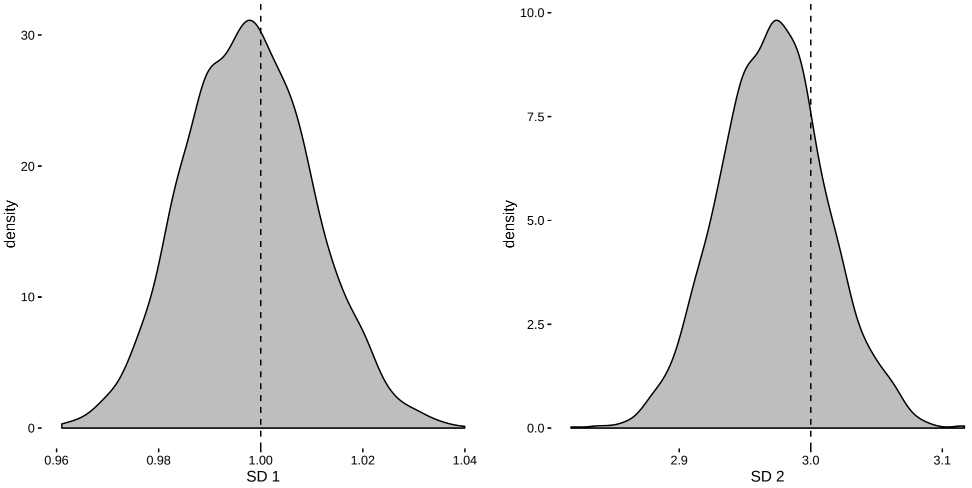

Figure 3: **Posterior distribution of component standard deviations**.

And Figure 4 shows the posterior distribution of the mixture probability as compared to its input value:

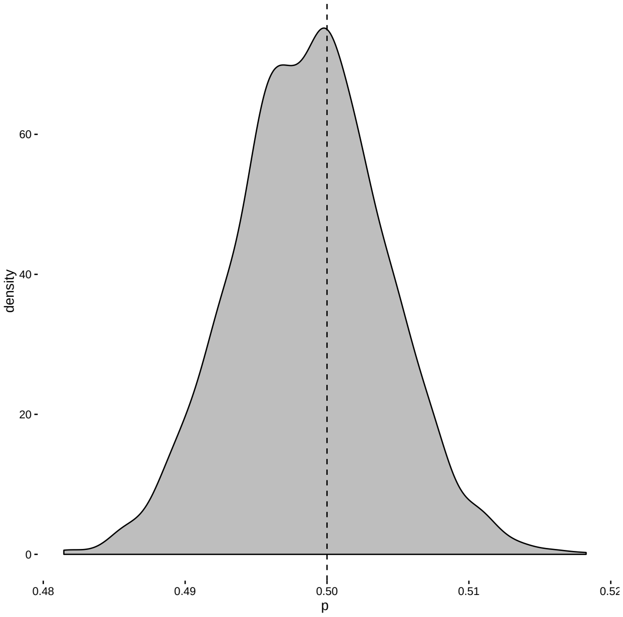

Figure 4: **Posterior distribution of mixture probability**.