Tutorial07 Solutions

The Dirichlet Distribution and the Dirichlet-Categorical Distribution

The Dirichlet distribution is a distribution over probability distributions; that is, draws from a Dirichlet are a probability distribution. In the lecture on overhypotheses, the Dirichlet distribution was parameterised with two values: Dirichlet($\alpha\boldsymbol{\beta}$), where $\alpha > 0$ is a scalar and $\boldsymbol{\beta}$ is a probability distribution of size K: all $\beta_k > 0$ and $\sum_k \beta_k = 1$.

The two parameters play different roles: $\boldsymbol{\beta}$ is the **base distribution** and α is the **concentration parameter**, governing how much draws from Dirichlet($\alpha\boldsymbol{\beta}$) will diverge from the base

This exercise is meant to strengthen your intuitions about the role of the α and $\boldsymbol{\beta}$ hyperparameters in the context of Dirichlet priors with categorical likelihoods, by examining how the hyperparameters influence the prediction of the next draw.

Dirichlet priors with categorical likelihoods have a convenient closed form for the predictive posterior, integrating over all possible draws from the Dirichlet (θ in the lectures). With the Dirichlet($\alpha\boldsymbol{\beta}$) parameterisation, the predictive posterior is:

$$p(y = k|D, \alpha, \boldsymbol{\beta}) = \frac{\alpha\beta_k + N_k}{\sum_{k'} \alpha\beta_{k'} + N_{k'}}$$

where N_k refers to the number of items of category k in D.

Exercise: Consider a dataset consisting of ten marbles, two of which are black (marbles can only be black and white; so $K = 2; N = 10; N_{\textbf{black}} = 2$). Given this dataset, calculate the probability of the next marble being black, using the following hyperparameter settings (9 different settings; preferably vary α while keeping $\boldsymbol{\beta}$ stable):

- $\boldsymbol{\beta} = [0.5, 0.5], \boldsymbol{\beta} = [0.2, 0.8], \boldsymbol{\beta} = [0.8, 0.2]$
- $\alpha = 0.01, 1, 100$

You can calculate these by hand (preferably) or use R/Matlab or any language you prefer to calculate the probability.

***Solution*:**

- $\alpha = 0.01, b = [0.5, 0.5], \frac{0.01*0.5+2}{0.01*0.5+2+0.01*0.5+8} = 2.005/10.01 = 0.2003$
- $\alpha = 1, b = [0.5, 0.5], 2.5/11 = 0.2273$
- $\alpha = 100, b = [0.5, 0.5], 52/110 = 0.4727$
- $\alpha = 0.01, b = [0.2, 0.8], 2.002/10.01 = 0.2$
- $\alpha = 1, b = [0.2, 0.8], 2.2/11.0000 = 0.2$
- $\alpha = 100, b = [0.2, 0.8], 22/110.0000 = 0.2$
- $\alpha = 0.01, b = [0.8, 0.2], 2.008/10.0100 = 0.2006$
- $\alpha = 1, b = [0.8, 0.2], 2.8/11 = 0.2545$
- $\alpha = 100, b = [0.8, 0.2], 82/110 = 0.7455$

Questions:

- What is the effect of a small α? A large α? *Small α adds a tiny psuedocount that gets washed away by the data, so pred post is very close to empirical likelihood/data distribution; large α reweights the data distribution more severely, towards the base distribution $\boldsymbol{\beta}$*
- The empirical likelihood of black is 0.2. When and why do the predictive posteriors match from the empirical likelihood? *Pred posterior only matches the empirical likelihood when the base distribution is the same as the empirical likelihoood - pseudocounts behave the same as the real counts.*
- In this example, the posterior predictive probability always matches or increases the probability of black, as compared to the empirical likelihood. What kind of hyperparameters would result in the predictive probability of black being **lower** than the empirical likelihood? *A base distribution that puts more weight on white (more skewed than the data distribution); bigger α will skew this even further.*

Stan Tutorial

In this part of the tutorial you'll be exploring the marbles-in-bags model from Kemp, Perfors, and Tenenbaum (2007) using a package called Stan. You will need to have the Kemp paper in front of you as well (the course website has a link to this paper).

Stan is a useful library that allows you to define a Bayesian model using the kind of notation you saw in the lecture (and will see in the code). Given a model specification and data, it calculates the posterior distribution using a sampler. Note that unlike the word learning model, which calculated a MAP solution, here we will be dealing with the full posterior distribution. Stan performs inference by building and running a sampler generated for your model.

Install Stan by running `install.packages("rstan")` in your R-console - this might take a while and will display warnings (that you can ignore). You might also need to install the graphing library `ggplot2`.

0. Getting started with Stan

Open the model file in `model.stan` and read it. The stan syntax will be new to you, but you should be able to see how the stan model specification corresponds to the model defined in the Appendix of Kemp, Perfors, and Tenenbaum (2007) (the first model, for Figure 1a).

The model specification is made up of different program blocks:

- The *data* block, where the required input variables are declared. These include:
  - dimensionalities/sizes (B, M and V);
  - the fixed (hyper)parameter ω (omega)
  - the data, in `items`
- The *parameters* block, where the model parameters are declared; these are the parameters that we will be inferring, via sampling.
- The *model* block. Here the relationships between the parameters and the data are defined, by specifying a generative story for the data.
- The *generated quantities* block. We may want to look at certain quantities that are functions of the parameters and data, but not explicitly in the model. Here we're calculating the distribution over `theta_new`, the distribution of colors in a bag from which we've seen only one marble.

Note that there are extra higher layers of parameters in the code (which actually matches the model we saw in lecture, not the paper).

Question: What is the difference between the code and the Kemp paper's model description?

Solution

The model in Kemp, Perfors, and Tenenbaum (2007) is:

$$\alpha \sim Exponential(\lambda)\beta \sim Dirichlet(1)\theta_i \sim Dirichlet(\alpha\beta)y^i|n^i \sim Multinomial(\theta_i).$$

The code provided has two extra hyperparameters, $\lambda \sim Exponential(\omega), \mu \sim Exponential(\omega)$

1. Homogenous bags

Exercise: Look at Figure 3a) i and ii in the Kemp paper; read the caption and make sure you understand why the distributions look the way they do. Write down your estimate/prediction of the mean of each of the distributions (for alpha, beta, a white-bag theta and a black-bag theta), so you can compare this to the model at the next step.

We provide you the code to generate the marble bags in `data.R`. First we'll be running the model with the data from the single color per bag setting and look at the output that's produced.

```
data.single <- generate_data.onecolor()
table(data.single$item)
```

```
##
##   1   2
## 200 200
```

Now we run the model specified in `model.stan` by providing it the data set. This will take a while, as Stan runs 4 chains of MCMC for 2000 iterations.

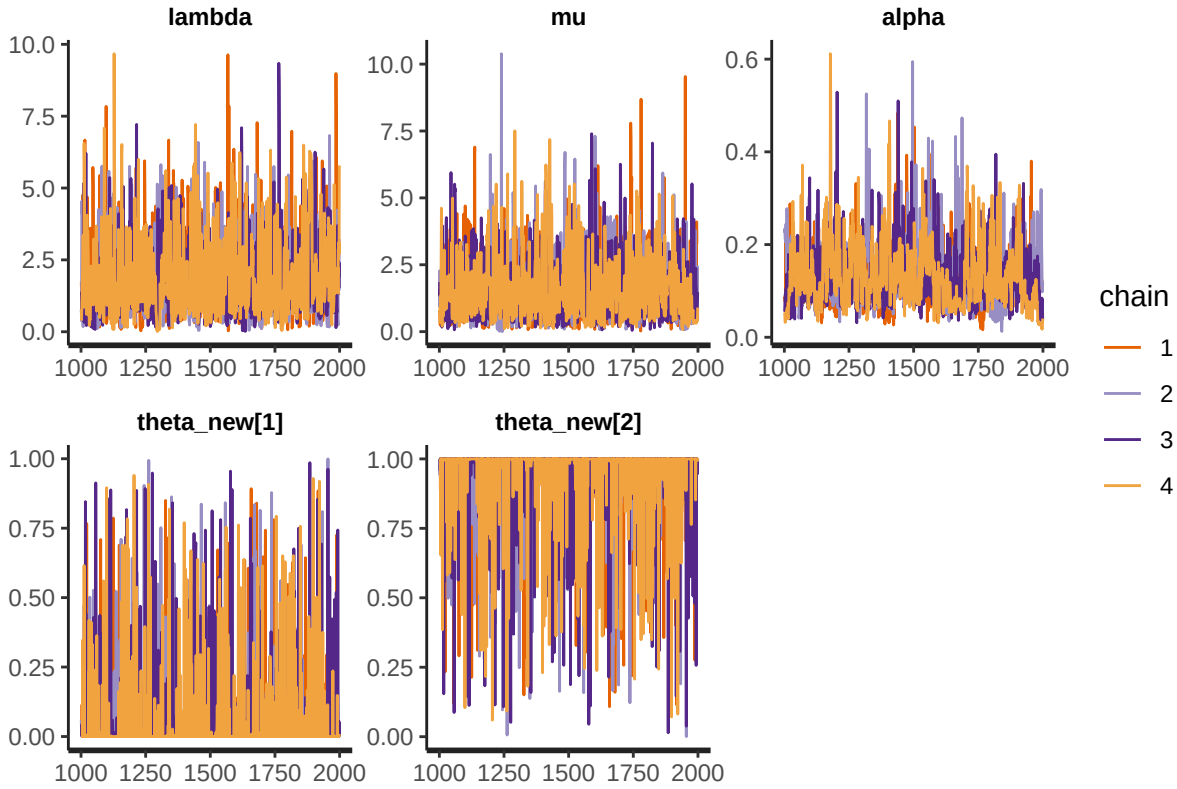
```
fitsingle <- stan(file = 'model.stan', data = data.single, verbose = FALSE,
            iter = 2000, chains = 4 , control=list(adapt_delta=0.95))
```

```
## Warning: There were 1519 divergent transitions after warmup. Increasing adapt_delta above 0.95 may h
## http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
```

```
## Warning: Examine the pairs() plot to diagnose sampling problems
```

The first thing to check is whether the model is working as expected; whether the inference procedure has converged to a good estimate of the posterior distribution. Since we ran 4 chains of MCMC, we can check how consistent these chains are - one common diagnostic is called the Gelman-Rubin diagnostic, or $\hat{R}$. Print out the full summary: `summary(fitsingle)` and check that the `Rhat` values are smaller than 1.1.

```
stan_trace(fitsingle, pars = c("lambda", "mu", "alpha", "theta_new"))
```




```
s <- summary(fitsingle)
kable(s$summary,  digits=2) #reports the merged chains
```

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
lambda	1.83	0.04	1.31	0.18	0.87	1.51	2.49	5.14	1203.69	1.00
mu	1.49	0.03	1.11	0.17	0.69	1.24	2.01	4.20	1605.10	1.00
alpha	0.13	0.00	0.07	0.04	0.09	0.12	0.17	0.29	305.05	1.01
beta[1]	0.45	0.01	0.13	0.22	0.36	0.45	0.54	0.71	388.86	1.02
beta[2]	0.55	0.01	0.13	0.29	0.46	0.55	0.64	0.78	388.86	1.02
theta[1,1]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	2506.40	1.00
theta[1,2]	1.00	0.00	0.01	0.99	1.00	1.00	1.00	1.00	2506.40	1.00
theta[2,1]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	3495.93	1.00
theta[2,2]	1.00	0.00	0.01	0.98	1.00	1.00	1.00	1.00	3495.93	1.00
theta[3,1]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	3319.26	1.00
theta[3,2]	1.00	0.00	0.01	0.98	1.00	1.00	1.00	1.00	3319.26	1.00
theta[4,1]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	3289.33	1.00
theta[4,2]	1.00	0.00	0.01	0.98	1.00	1.00	1.00	1.00	3289.33	1.00
theta[5,1]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	3068.21	1.00
theta[5,2]	1.00	0.00	0.01	0.98	1.00	1.00	1.00	1.00	3068.21	1.00
theta[6,1]	1.00	0.00	0.01	0.98	1.00	1.00	1.00	1.00	2013.19	1.00
theta[6,2]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	2013.19	1.00
theta[7,1]	1.00	0.00	0.01	0.98	1.00	1.00	1.00	1.00	3848.96	1.00
theta[7,2]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	3848.96	1.00
theta[8,1]	1.00	0.00	0.01	0.98	1.00	1.00	1.00	1.00	3036.87	1.00
theta[8,2]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	3036.87	1.00
theta[9,1]	1.00	0.00	0.01	0.98	1.00	1.00	1.00	1.00	3576.19	1.00
theta[9,2]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	3576.19	1.00
theta[10,1]	1.00	0.00	0.01	0.98	1.00	1.00	1.00	1.00	2851.13	1.00
theta[10,2]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	2851.13	1.00

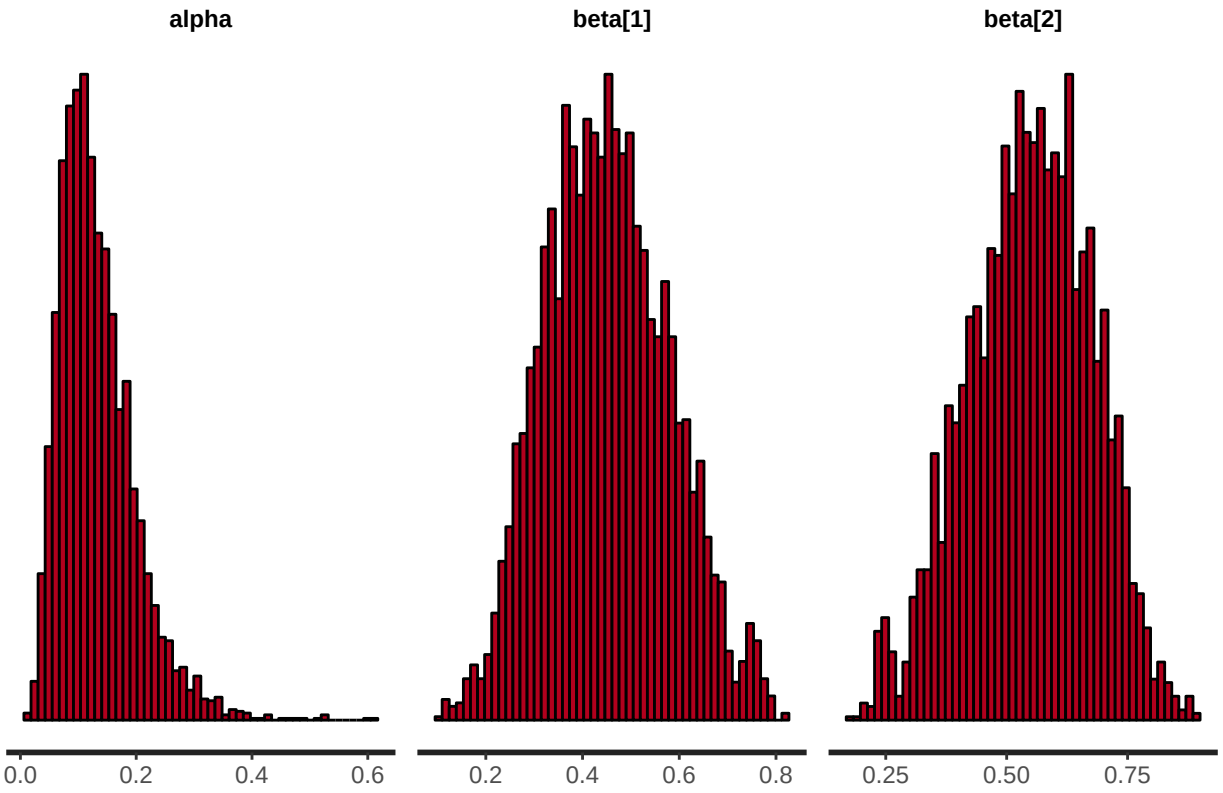
	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
theta_new[1]	0.05	0.00	0.15	0.00	0.00	0.00	0.01	0.57	3743.19	1.00
theta_new[2]	0.95	0.00	0.15	0.43	0.99	1.00	1.00	1.00	3743.19	1.00
lp__	-54.10	0.33	4.36	-63.69	-56.78	-53.60	-51.03	-46.82	172.92	1.02

In addition to the `Rhat` values the output also gives you the mean posterior estimates and distributions over the parameters in the model and the generated quantities.

Questions:

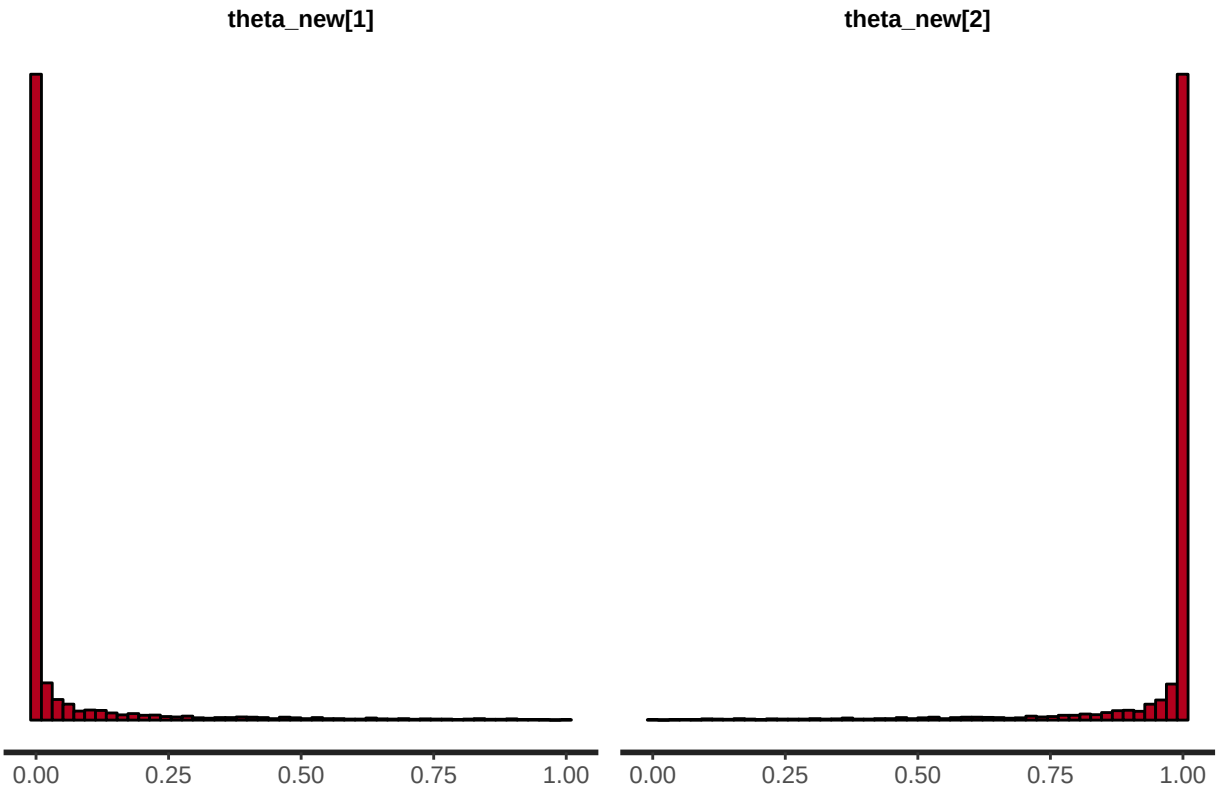
- Do these means correspond to what you predicted from the paper?
- What about the shape/spread of the distribution? You can look at the percentiles, and/or use `stan_hist(fit,pars=c("alpha", "beta"), bins=50)` to examine the sampled values, which correspond to the posterior. (I.e. once it's converged, the sampler draws from the posterior, so you would expect many samples from high-probability areas and few from low-probability areas).

```
stan_hist(fitsingle, pars = c("alpha", "beta"), bins=50)
```



- Check `theta_new`, the prediction about the distribution of a new bag of which you've seen only one marble. What does it say about the probable color of the next marble?

```
stan_hist(fitsingle, pars = c("theta_new"), bins=50)
```



2. Heterogeneous bags

Exercise: Do the same (go through all steps and questions from (1)) for the mixed bags scenario in Figure 3a(iii). Note that this means inputting different data – use `generate_data.mixed_proportion` in `data.R`.

```
data.mixed <- generate_data.mixed_proportion()
```

```
## [1] 5 1 0
## [1] 5 1 1
## [1] 5 1 2
## [1] 5 1 3
## [1] 5 1 4
## [1] 5 2 5
## [1] 5 2 6
## [1] 5 2 7
## [1] 5 2 8
## [1] 5 2 9
## [1]  5  3 10
## [1]  5  3 11
## [1]  5  3 12
## [1]  5  3 13
## [1]  5  3 14
## [1]  5  4 15
## [1]  5  4 16
## [1]  5  4 17
## [1]  5  4 18
```

```
## [1]  5  4 19
```

```
fitmixed <- stan(file = 'model.stan', data =data.mixed , verbose = FALSE,
            iter = 2000, chains = 4, control=list(adapt_delta=0.95))
```

```
## Warning: There were 5 divergent transitions after warmup. Increasing adapt_delta above 0.95 may help
## http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
```

```
## Warning: Examine the pairs() plot to diagnose sampling problems
```

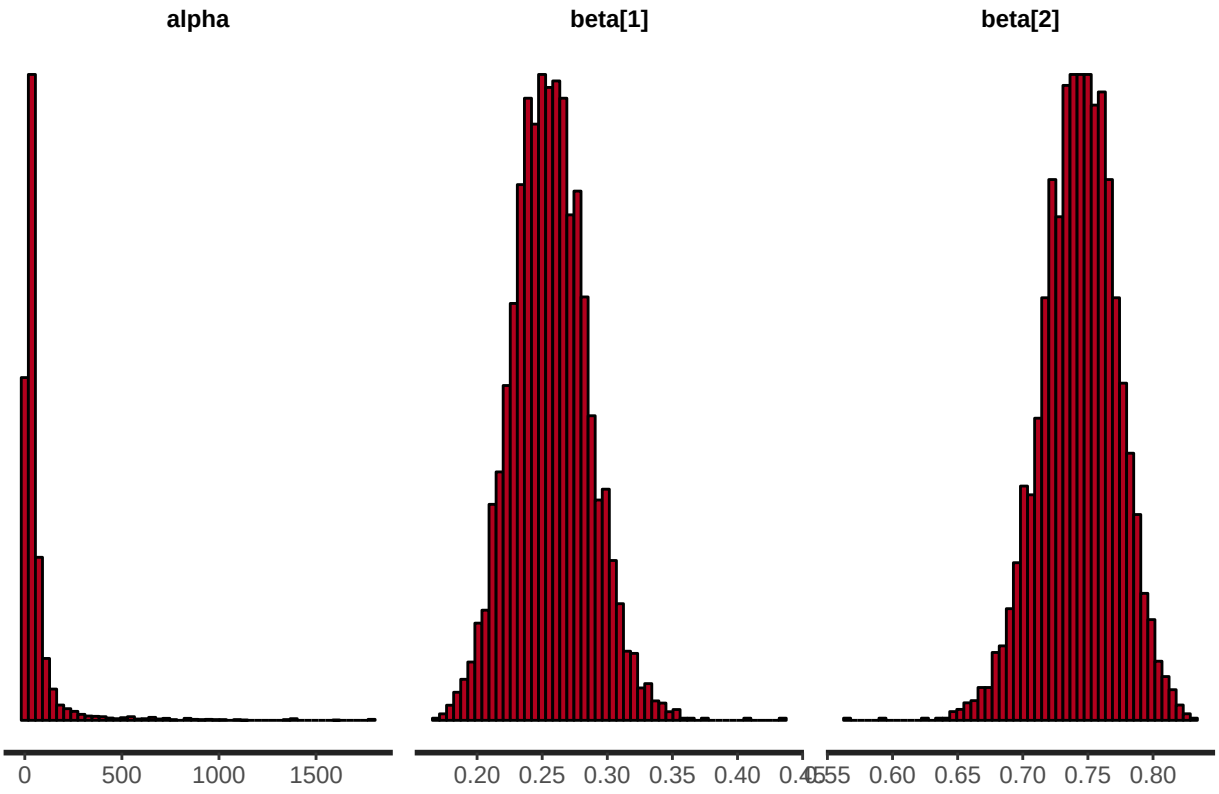
```
smixed <- summary(fitmixed)

kable(smixed$summary,  digits=2)
```

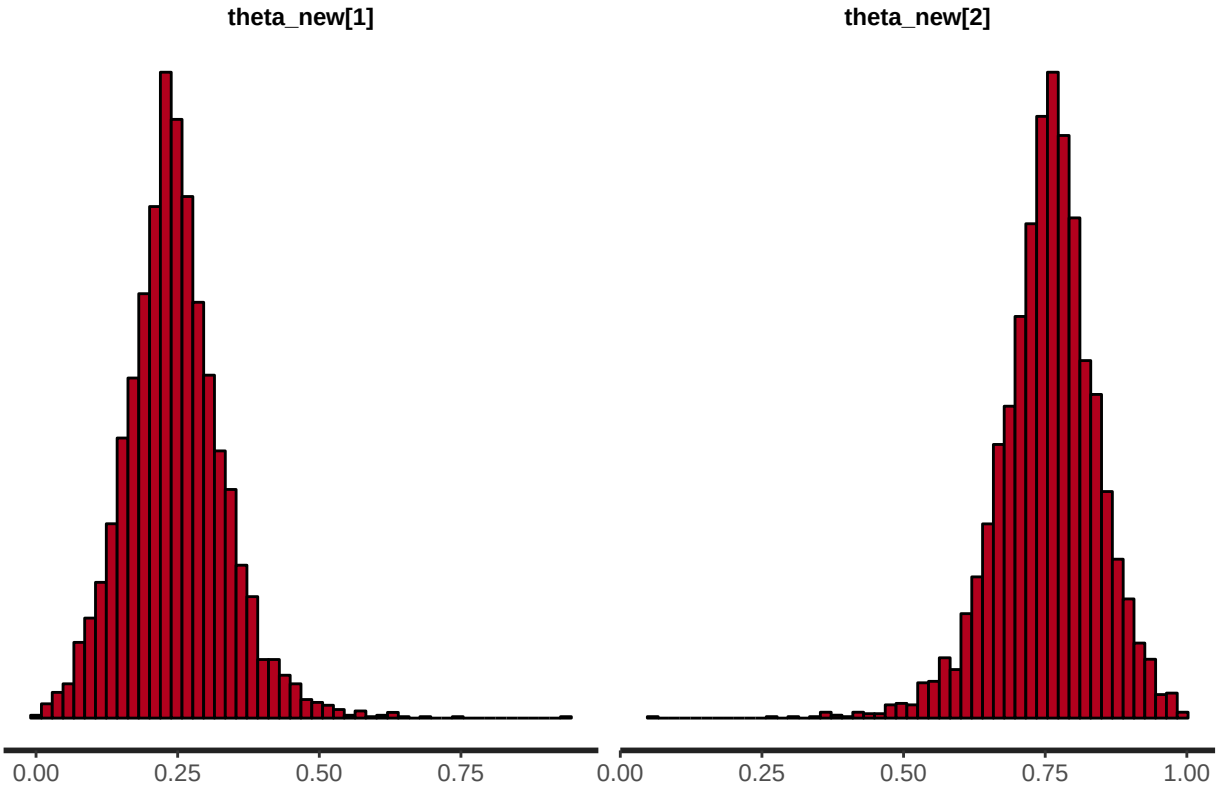
	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
lambda	0.08	0.00	0.09	0.00	0.02	0.05	0.10	0.32	817.59	1.00
mu	1.25	0.01	0.95	0.14	0.57	1.01	1.65	3.80	4653.32	1.00
alpha	63.86	11.75	131.15	7.44	18.02	29.94	55.66	392.90	124.67	1.05
beta[1]	0.26	0.00	0.03	0.20	0.24	0.26	0.28	0.32	1604.71	1.00
beta[2]	0.74	0.00	0.03	0.68	0.72	0.74	0.76	0.80	1604.71	1.00
theta[1,1]	0.19	0.00	0.06	0.07	0.15	0.19	0.24	0.31	1286.34	1.00
theta[1,2]	0.81	0.00	0.06	0.69	0.76	0.81	0.85	0.93	1286.34	1.00
theta[2,1]	0.19	0.00	0.06	0.07	0.15	0.19	0.24	0.31	1068.96	1.00
theta[2,2]	0.81	0.00	0.06	0.69	0.76	0.81	0.85	0.93	1068.96	1.00
theta[3,1]	0.19	0.00	0.06	0.07	0.15	0.19	0.24	0.31	1286.65	1.00
theta[3,2]	0.81	0.00	0.06	0.69	0.76	0.81	0.85	0.93	1286.65	1.00
theta[4,1]	0.19	0.00	0.06	0.08	0.15	0.19	0.24	0.31	1063.80	1.00
theta[4,2]	0.81	0.00	0.06	0.69	0.76	0.81	0.85	0.92	1063.80	1.00
theta[5,1]	0.19	0.00	0.06	0.07	0.15	0.20	0.24	0.32	1320.72	1.00
theta[5,2]	0.81	0.00	0.06	0.68	0.76	0.80	0.85	0.93	1320.72	1.00
theta[6,1]	0.23	0.00	0.06	0.12	0.19	0.23	0.27	0.37	4433.61	1.00
theta[6,2]	0.77	0.00	0.06	0.63	0.73	0.77	0.81	0.88	4433.61	1.00
theta[7,1]	0.23	0.00	0.06	0.11	0.19	0.23	0.27	0.36	3003.21	1.00
theta[7,2]	0.77	0.00	0.06	0.64	0.73	0.77	0.81	0.89	3003.21	1.00
theta[8,1]	0.23	0.00	0.06	0.11	0.19	0.23	0.27	0.37	3230.05	1.00
theta[8,2]	0.77	0.00	0.06	0.63	0.73	0.77	0.81	0.89	3230.05	1.00
theta[9,1]	0.23	0.00	0.06	0.12	0.19	0.23	0.27	0.36	3189.71	1.00
theta[9,2]	0.77	0.00	0.06	0.64	0.73	0.77	0.81	0.88	3189.71	1.00
theta[10,1]	0.23	0.00	0.06	0.11	0.19	0.23	0.27	0.37	3977.00	1.00
theta[10,2]	0.77	0.00	0.06	0.63	0.73	0.77	0.81	0.89	3977.00	1.00
theta[11,1]	0.27	0.00	0.06	0.16	0.23	0.27	0.31	0.42	3905.13	1.00
theta[11,2]	0.73	0.00	0.06	0.58	0.69	0.73	0.77	0.84	3905.13	1.00
theta[12,1]	0.27	0.00	0.06	0.16	0.23	0.27	0.31	0.41	3752.03	1.00
theta[12,2]	0.73	0.00	0.06	0.59	0.69	0.73	0.77	0.84	3752.03	1.00
theta[13,1]	0.27	0.00	0.07	0.16	0.23	0.27	0.31	0.42	3959.58	1.00
theta[13,2]	0.73	0.00	0.07	0.58	0.69	0.73	0.77	0.84	3959.58	1.00
theta[14,1]	0.27	0.00	0.06	0.16	0.23	0.27	0.31	0.41	4287.91	1.00
theta[14,2]	0.73	0.00	0.06	0.59	0.69	0.73	0.77	0.84	4287.91	1.00
theta[15,1]	0.27	0.00	0.07	0.16	0.23	0.27	0.31	0.42	4198.56	1.00
theta[15,2]	0.73	0.00	0.07	0.58	0.69	0.73	0.77	0.84	4198.56	1.00
theta[16,1]	0.31	0.00	0.07	0.19	0.26	0.30	0.35	0.47	1427.04	1.00
theta[16,2]	0.69	0.00	0.07	0.53	0.65	0.70	0.74	0.81	1427.04	1.00
theta[17,1]	0.31	0.00	0.07	0.19	0.26	0.30	0.36	0.48	1521.74	1.00
theta[17,2]	0.69	0.00	0.07	0.52	0.64	0.70	0.74	0.81	1521.74	1.00
theta[18,1]	0.31	0.00	0.07	0.20	0.26	0.30	0.36	0.48	1910.99	1.00

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
theta[18,2]	0.69	0.00	0.07	0.52	0.64	0.70	0.74	0.80	1910.99	1.00
theta[19,1]	0.31	0.00	0.07	0.19	0.26	0.31	0.35	0.48	1423.23	1.00
theta[19,2]	0.69	0.00	0.07	0.52	0.65	0.69	0.74	0.81	1423.23	1.00
theta[20,1]	0.31	0.00	0.07	0.19	0.26	0.31	0.36	0.47	1684.73	1.00
theta[20,2]	0.69	0.00	0.07	0.53	0.64	0.69	0.74	0.81	1684.73	1.00
theta_new[1]	0.25	0.00	0.09	0.08	0.19	0.24	0.29	0.44	3367.91	1.00
theta_new[2]	0.75	0.00	0.09	0.56	0.71	0.76	0.81	0.92	3367.91	1.00
lp___	-238.35	0.56	7.39	-251.38	-243.15	-238.98	-234.47	-220.10	171.25	1.03

```
stan_hist(fitmixed, pars = c("alpha", "beta"), bins=50)
```




```
stan_hist(fitmixed, pars = c("theta_new"), bins=50)
```



Question: Why is the shape/spread of the distribution of theta_new so different between Fig 3a(ii) and 3a(iii)?

Solution

Because the data is different! In 3a(ii) the bags are all single colour and (posterior) alpha is very small, so each theta (each bag) has a very skewed distribution that is far from beta (beta will capture the overall distribution over all bags/theta, which in this case is averaging over all-black and all-white bags. After you see a single black marble, you have a strong posterior belief that the rest of that bag is also going to be full of black marbles; this is what theta_new captures. In 3(a)iii the bags consist of different proportions of black/white marbles, which are all less extreme than in the single-color case; this leads to broader theta (probability of drawing a black marble is ~0.2, not $_{1/}0$). The hyperpameters (alpha, beta) capture this fact: log(alpha) is positive, so alpha > 1, so draws from Dir(alpha beta) will reflect more closely beta, the base distribution. (Beta again reflects the overall data distribution, where black marbles occur 25% of the time.) In the case of theta_new, after you see a black marble, you don't have as much information (compared to 3a(ii)) about what the colour of the next marble from that bag is going to be. If you think about the example, this should match your intuition: in the first scenario, you're in a context in which all the bags seem to consist of only one color of marble, and you'd expect the next bag to follow this pattern, whereas in the second scenario you know much less about what kind of color distribution you're likely to get in a new bag.

References

Kemp, Charles, Amy Perfors, and Joshua Tenenbaum. 2007. "Learning Overhypotheses with Hierarchical Bayesian Models." *Developmental Science* 10. Wiley Online Library.