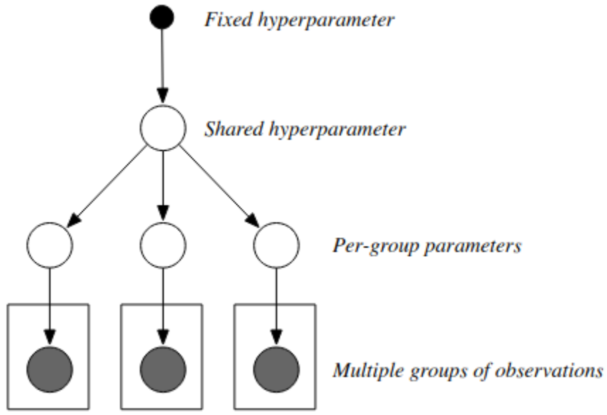
Hierarchical modeling.md

Hierarchical modeling

In Bayesian modeling it is easier and flexibile to implement a hierarchical model.

The classical hierarchical model looks like this:



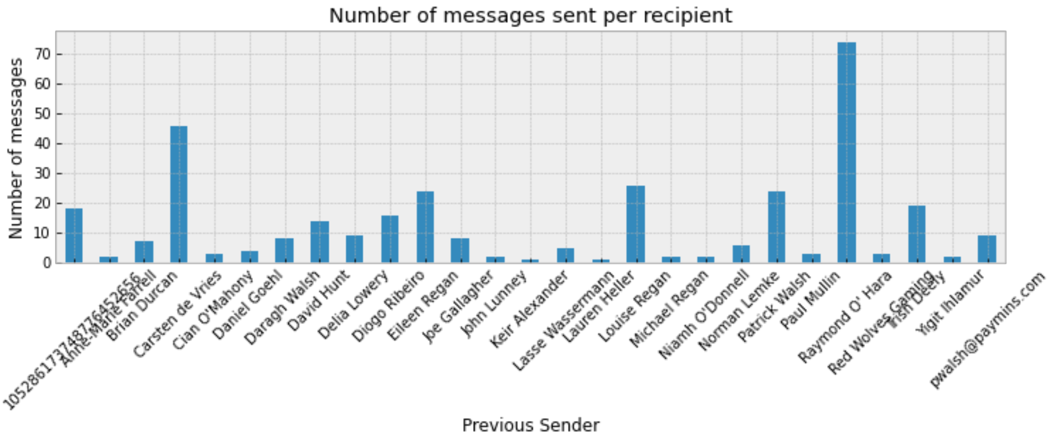
Contents

- Model pooling (separate models)
- Partial pooling (hierarchical models)
- Shrinkage effect of partial pooling

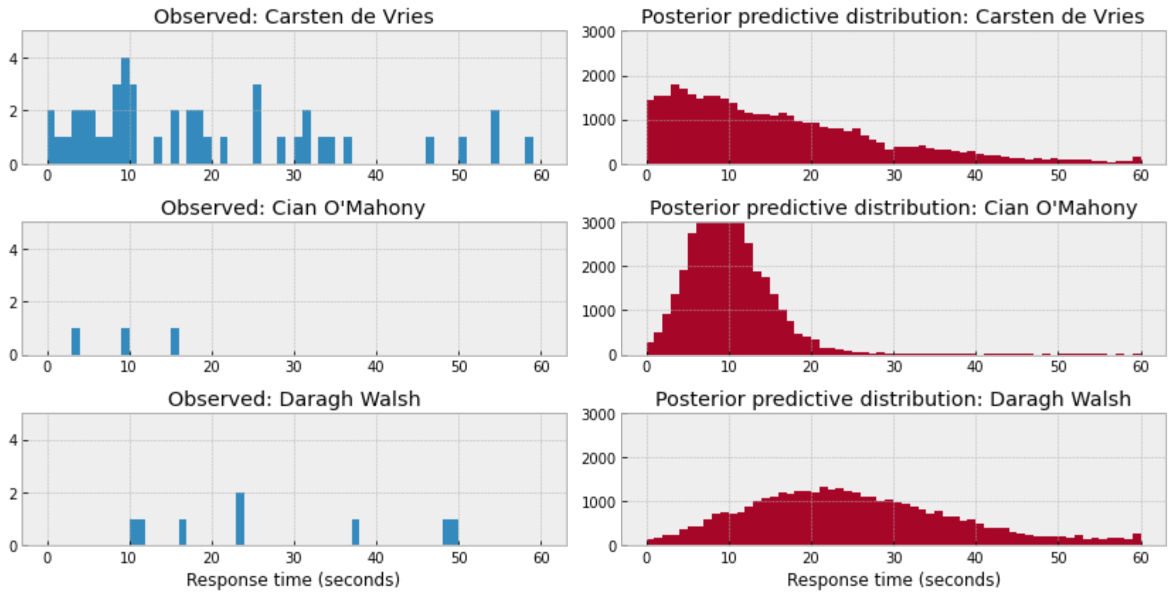
Model Pooling

A different way of modeling the response time for the hangout conversations. Intuition suggests that the tendency to reply quickly to a chat depends on who we are talking to. We might be more likely to respond quickly to the girlfriend than to a distant friend. As such, we can decide to model each conversation independently, estimating parameters $μ_i$ and $α_i$ for each conversation i.

One consideration we must make, is that some conversations have very few messages compared to others. As such, our estimates of response time for conversations with few messages will have a higher degree of uncertainty than conversations with a large number of messages. The below plot illustrates the discrepancy in sample size per conversation.

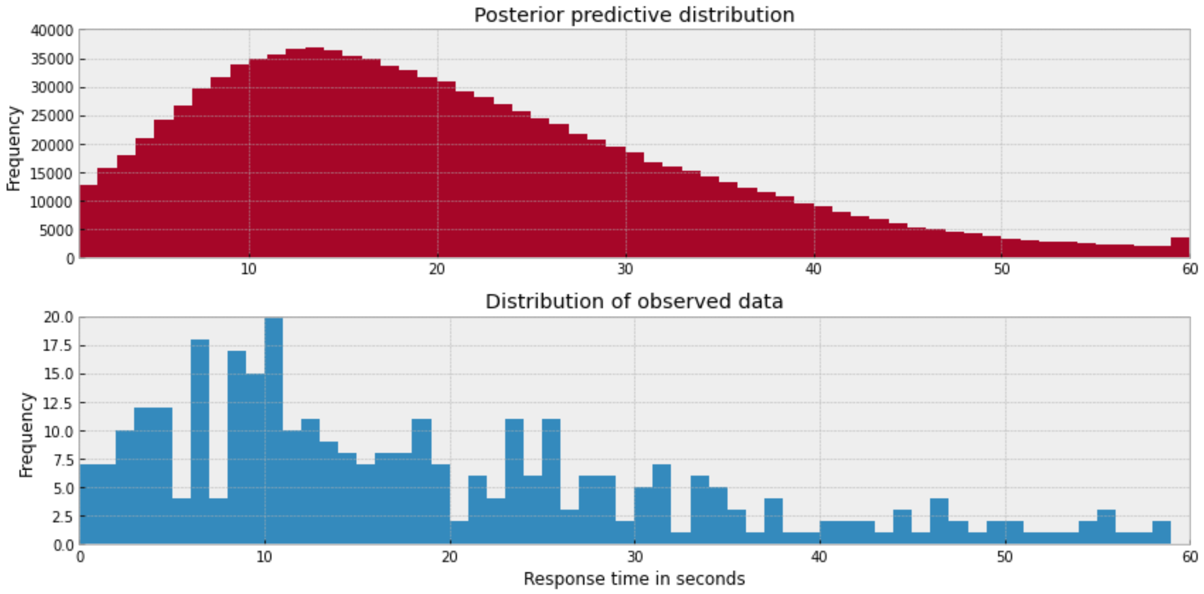


For each message j and each conversation i, we represent the model as: $y_{ji}~NegBinomial(μ_i,α_i)$ $μ_i=Uniform(0,100)$ $α_i=Uniform(0,100)$



The above plots show the observed data (left) and the posterior predictive distribution (right) for 3 example conversations we modeled. As you can see, the posterior predictive distribution can vary considerably across conversations. This could accurately reflect the characteristics of the conversation or it could be inaccurate due to small sample size.

If we combine the posterior predictive distributions across these models, we would expect this to resemble the distribution of the overall dataset observed. Let's perform the posterior predictive check.



Yes, the posterior predictive distribution resembles the distribution of the observed data. However, I'm concerned that some of the conversations have very little data and hence the estimates are likely to have high variance. One way to mitigate this risk to to share information across conversations - but still estimate μi for each conversation. We call this **partial pooling**.

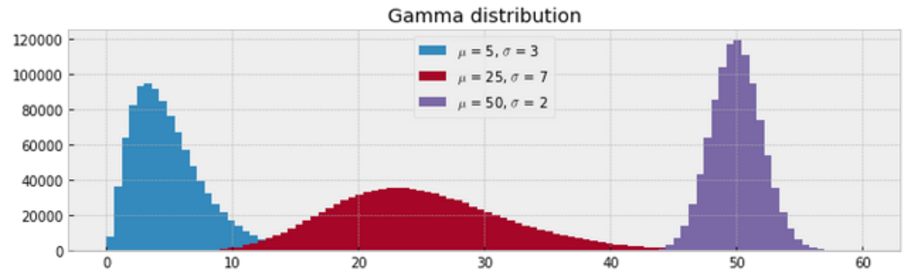
Partial pooling

Just like in the pooled model, a partially pooled model has paramater values estimated for each conversation i. However, parameters are connected together via hyperparameters. This reflects our belief that `response_time`'s per conversation have similarities with one another via natural tendancy to respond quickly or slowly.

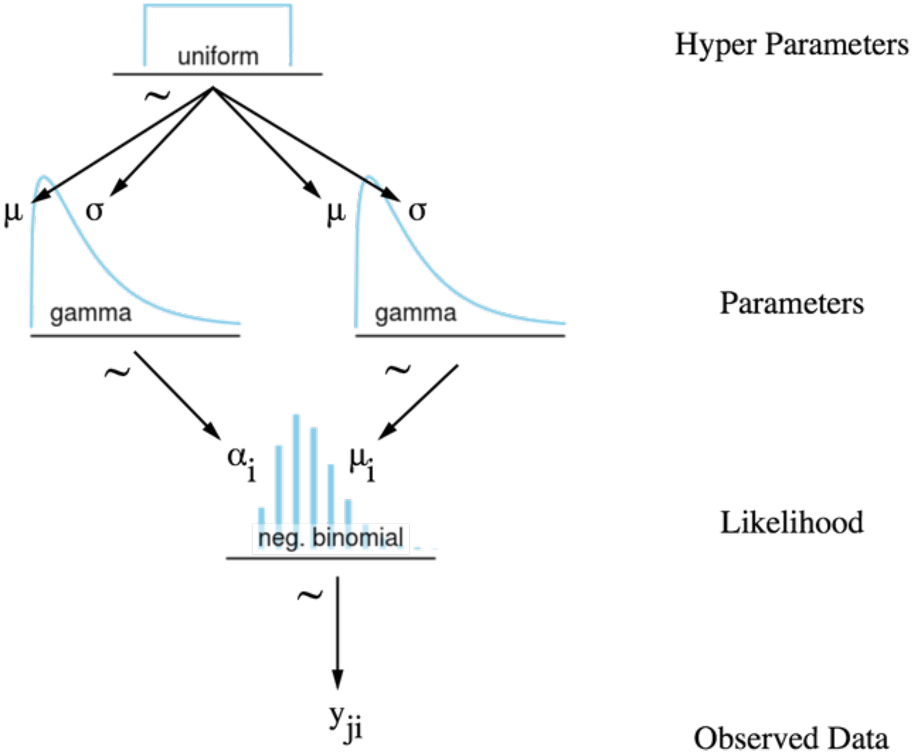
$y_{ji} \sim NegBinomial(\mu_i, \alpha_i)$

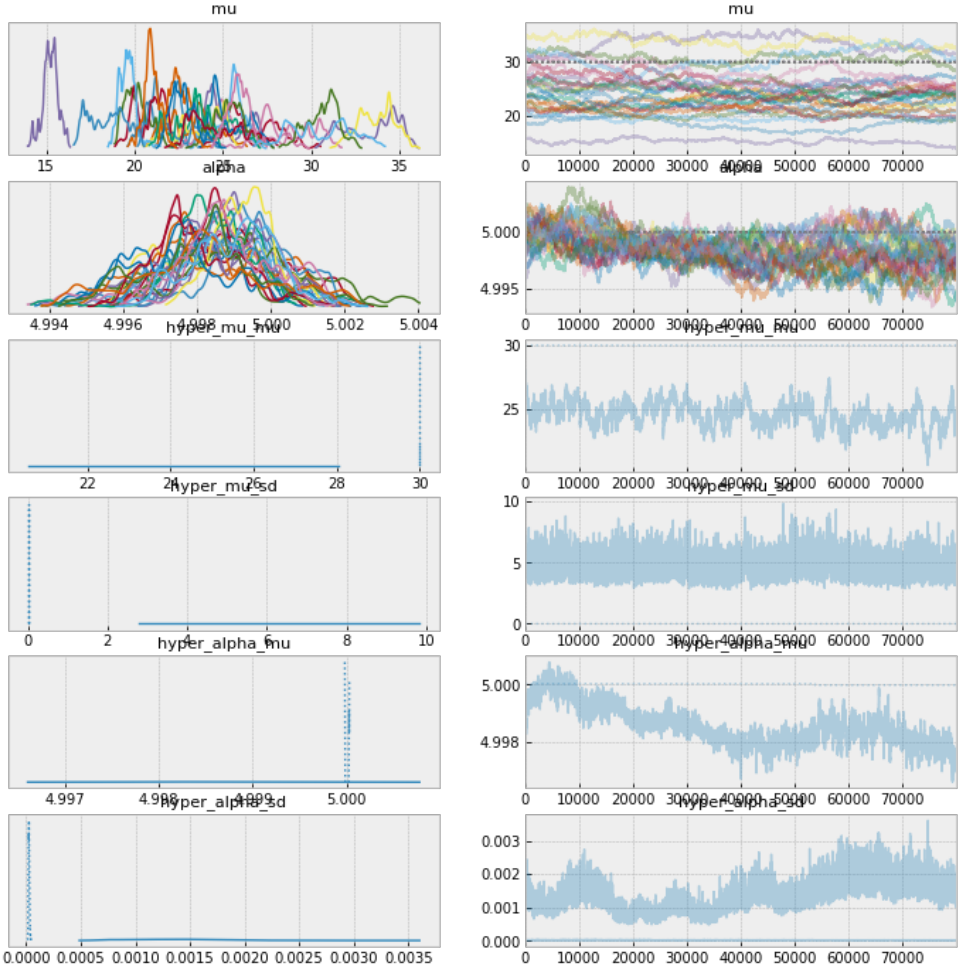
Following on from the above example, we will estimate parameter values (μ_i) and (α_i) for a Poisson distribution. Rather than using a uniform prior, we will use a Gamma distribution for both μ and σ. This will enable us to introduce more prior knowledge into the model as we have certain expectations as to what vales μ and σ will be.

First, let's have a look at the Gamma distribution. As you can see below, it is very flexible.



The partially pooled model can be formally described by: $y_{ji} \sim NegBinomial(\mu_i, \alpha_i)$ $\mu_i = Gamma(\mu_\mu, \sigma_\mu)$ $\alpha_i = Gamma(\mu_\alpha, \sigma_\alpha)$ $\mu_\mu = Uniform(0, 60)$ $\sigma_\mu = Uniform(0, 50)$ $\mu_\alpha = Uniform(0, 10)$ $\sigma_\alpha = Uniform(0, 50)$



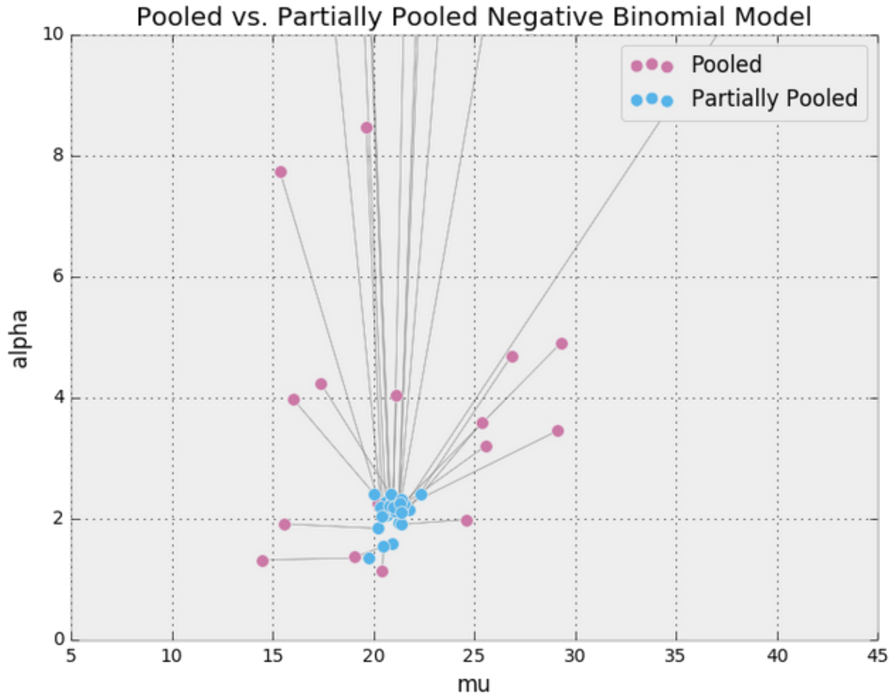


You can see for the estimates of μ and α that we have multiple plots - one for each conversation i. The difference between the pooled and the partially pooled model is that the parameters of the partially pooled model (μ_i and α_i) have a hyperparameter that is shared across all conversations i. This brings two benefits:

1. Information is shared across conversations, so for conversations that have limited sample size, they "borrow" knowledge from other conversations during estimation to help reduce the variance of the estimate
2. We get an estimate for each conversation and an overall estimate for all conversations.

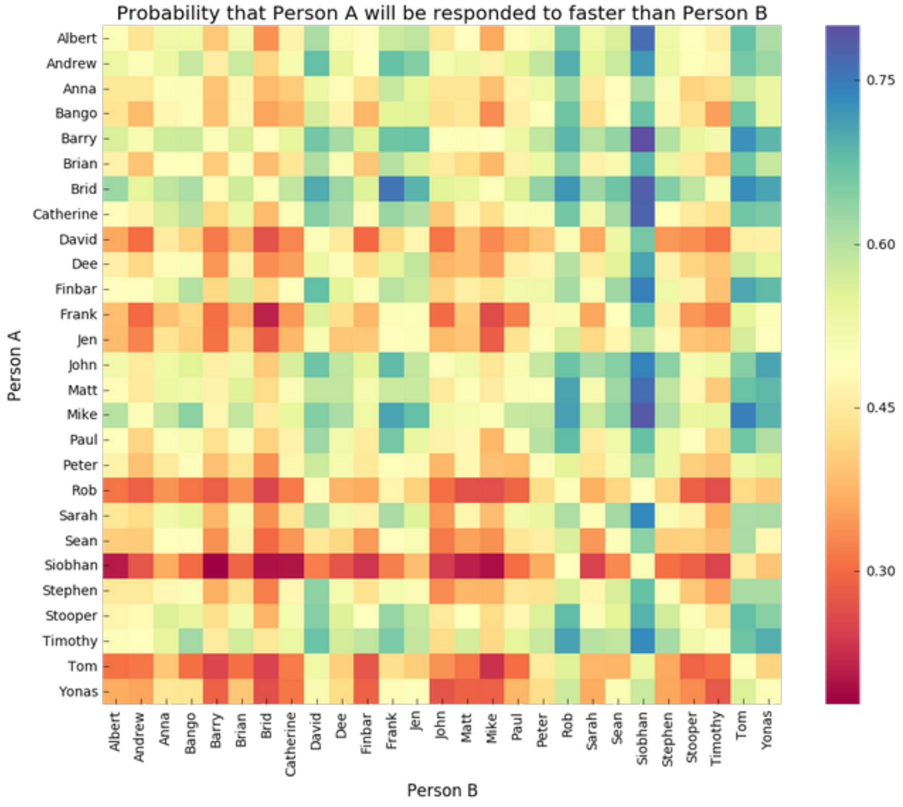
Shrinkage Effect

In statistics, shrinkage is the reduction in the effects of sampling variation. In regression analysis, a fitted relationship appears to perform less well on a new data set than on the data set used for fitting. In particular the value of the coefficient of determination 'shrinks'. This idea is complementary to overfitting and, separately, to the standard adjustment made in the coefficient of determination to compensate for the subjunctive effects of further sampling, like controlling for the potential of new explanatory terms improving the model by chance: that is, the adjustment formula itself provides "shrinkage." But the adjustment formula yields an artificial shrinkage.



Comparing all of the people's response times

```
f, ax = plt.subplots(figsize=(12, 9))
cmap = plt.get_cmap("Spectral")
sns.heatmap(ab_dist_df, square=True, cmap=cmap)
plt.title('Probability that Person A will be responded to faster than Person B')
plt.ylabel('Person A')
plt.xlabel('Person B')
```


Probability that Person A will be responded to faster than Person B
Person A
Person B
Albert
Andrew
Anna
Bango
Barry
Brian
Brid
Catherine
David
Dee
Finbar
Frank
Jen
John
Matt
Mike
Paul
Peter
Rob
Sarah
Sean
Siobhan
Stephen
Stooper
Timothy
Tom
Yonas