The Baby Name Model

Hanming Yang, Columbia University

December 15, 2023

Motivation & Data

- The Social Security Administration has been collecting data on the first names that are given to newborns since 1880. This data is available per year, on a state wide resolution. There is little to no research that has been done on the subject of baby names. This project seeks to use topic models to explore the question of **whether there is any significant correlation between the socioeconomic factors of a state and the newborn first names that are chosen within that state**. Do parents under the influence of different socioeconomic factors use different names? Do name frequencies have any predictive power over socioeconomic factors?

Why Topic Models?

- Relating this work to the existing topic models such as the LDA, each baby name can be thought of as word, and each state can be thought of as a document. Hence, given a certain year of name counts across states, it is possible to run LDA out of the box on this data set.
- This then allows us to find topics. Here I make the assumption that each topic represents the influence over names of a combination of socioeconomic factors, and will refer to each topic interchangeably as a "factor". This then allows me to define the interpretability of a factor by how well a small set of socioeconomic indicators (such as political belief, GDP, or demographic data) can predict the factor's proportions within each document.

Existing Models

- There are several issues when applying the LDA model to my data. In each state, the influence of all factors forms a simplex. This means that if a state is being influenced by a factor that indicates high GDP, it would limit the influence of other factors such as a state being very conservative or religious. However, if we were to assume that each topic were a socioeconomic factor, there should be states that can be heavily wealthy as well as conservative/liberal.
- As a direct consequence of the previous point, the distribution of a factor's proportion in each document is not particularly meaningful. Yet, socioeconomic data such a GDP is exactly a distribution of an attribute over each state. Hence, the generative model of the LDA conflicts with the interpretation of topics in this context.

The Baby Name Model (BNM)

- The new model's generative model is as follows.

  1. For each topic k in K:
     1.1 Draw from V dimensional Dirichlet to get the topic's distribution over words β_k
     1.2 Draw from N dimensional Dirichlet to get the topic's distribution over documents ϕ_k
  2. Draw from K dimensional Dirichlet to get global topic proportions θ
  3. For each word i out of the W words:
     3.1 Draw topic assignment z_i from θ
     3.2 Draw document assignment d_i from ϕ_{z_i}
     3.3 Draw word x_i from β_{z_i}

- This generative model removes the simplex constraint for document-topic proportions and instead places the constraint on a new topic-document distribution. This allows for each topic to have an interpretable topic-word distribution, as well as an interpretable topic-document distribution. While topics still need to cooperate in distributing every word in the vocabulary, the changes to the generative model does however provide less encouragement for each topic to be different. The model is now also more prone to differences in document length, requiring word frequency weighting methods such as the Term Frequency-Inverse Document Frequency (TF-IDF), or a simple normalization of document-word counts.
- This then has the following joint probability expression, with the graphical model provided as a reference as well.

$$p(\theta,\beta,\phi,z,d,x) = p(\theta;\gamma)\prod_{k=1}^{K} p(\beta_k;\eta)p(\phi_k;\alpha)\prod_{i=1}^{W} p(z_i|\theta)p(d_i|\phi,z_i)p(x_i|\beta,z_i)$$

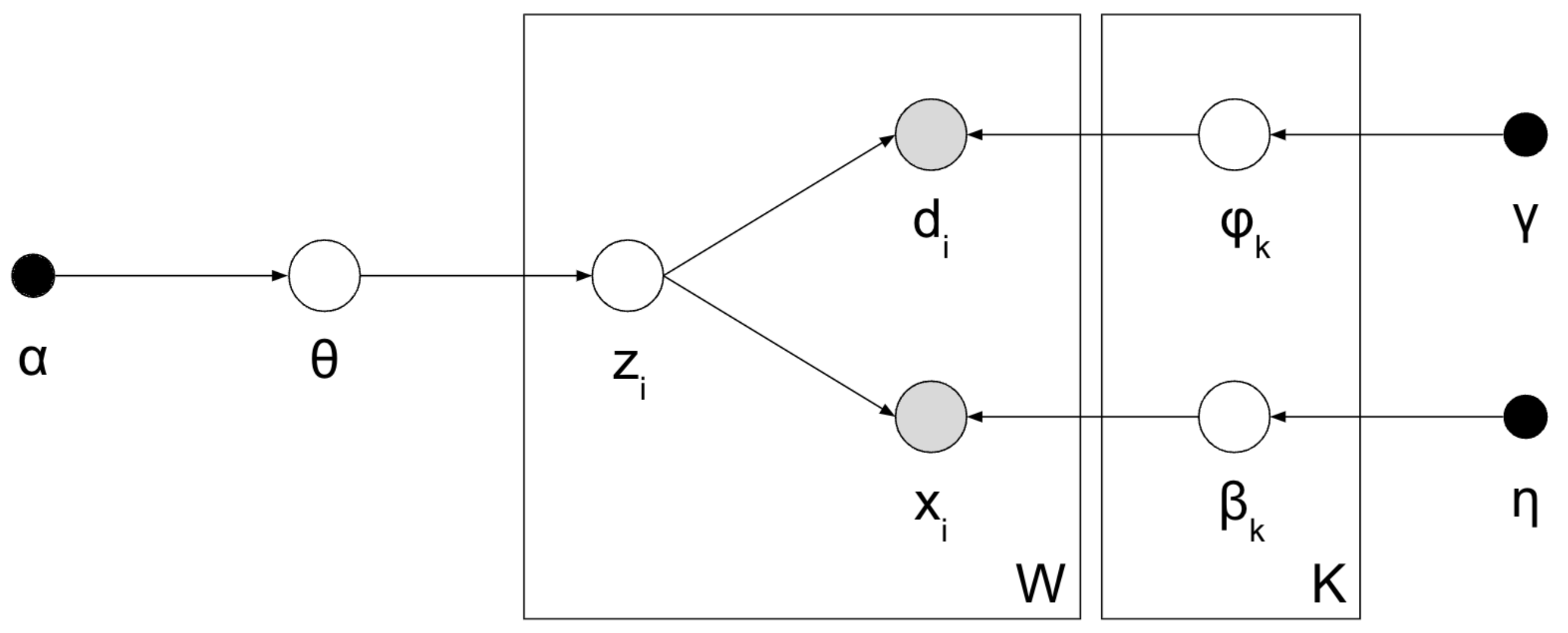

Figure: The BNM Graphical Model

- The Gibbs sampler is the simplest approach to optimizing the joint probability. θ, β, and ϕ obey the Dirichlet-categorical conjugacy,and can therefore be sampled from their respective Dirichlet posteriors. z's categorical posterior can also be easily constructed through multiplying topic proportion, topic-word compatibility, and topic-document compatibility.

Staring at the Joint Probability Expression

- We now play the game of "staring at the join probability expression". Since the product over the words dominate the expression, let us consider how the three terms within the product interact. For $p(\theta|\gamma)$ to be increase, θ needs to be come more sparse. This would mean that more names would be assigned to the same topic, which would mean that ϕ and β would jointly have to become more uniform to accommodate for the generality of the few topics that have high proportion. This in turn decreases the value of $p(d_i|\phi,z_i)p(x_i|\beta,z_i)$. If ϕ and β were to become more sparse, θ would then have to become more uniform.
- Now if we condition on a constant θ, if $p(d_i|\phi,z_i)$ increases by having a sparse ϕ, $p(x_i|\beta,z_i)$ would in turn decrease as concentrating on fewer documents would likely force the topic-word distribution to cover a diverse vocabulary. The same holds for when β becomes more sparse.
- While this three-way competition between θ, ϕ, and β does allow for more flexibility in terms of hyper-parameter tuning, it does mean that the wrong hyper-parameters can cause divergent training loss. In practice, the baby name data set worked well when θ's prior was larger than that of ϕ and β. This allowed for the model to focus its representation power over the topics while keeping topic proportions more uniform.
- The following figures illustrate how this joint probability distribution is expected to behave upon optimization.

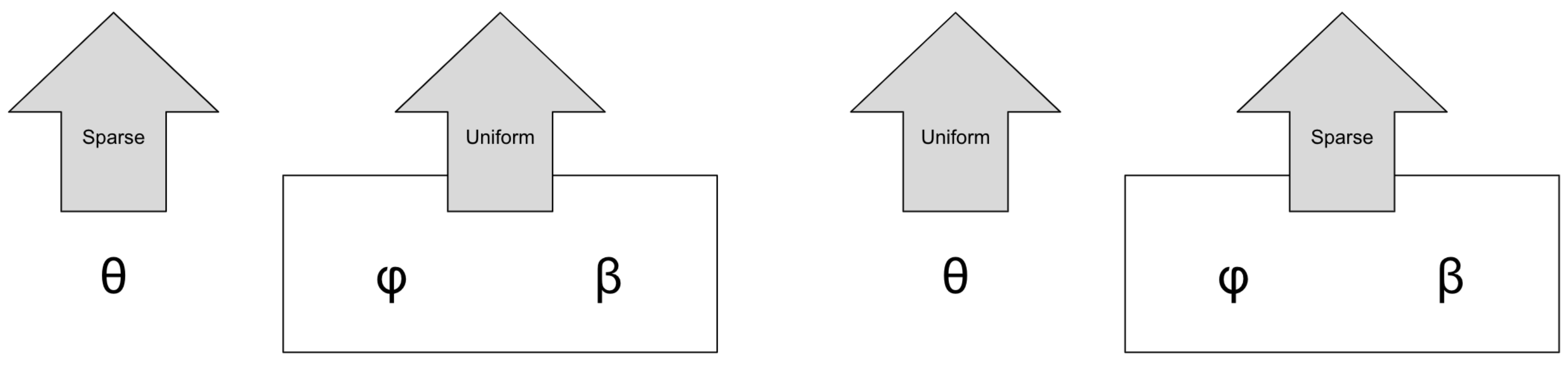

Figure: As θ becomes more sparse, ϕ and β together must become more uniform and vise versa.

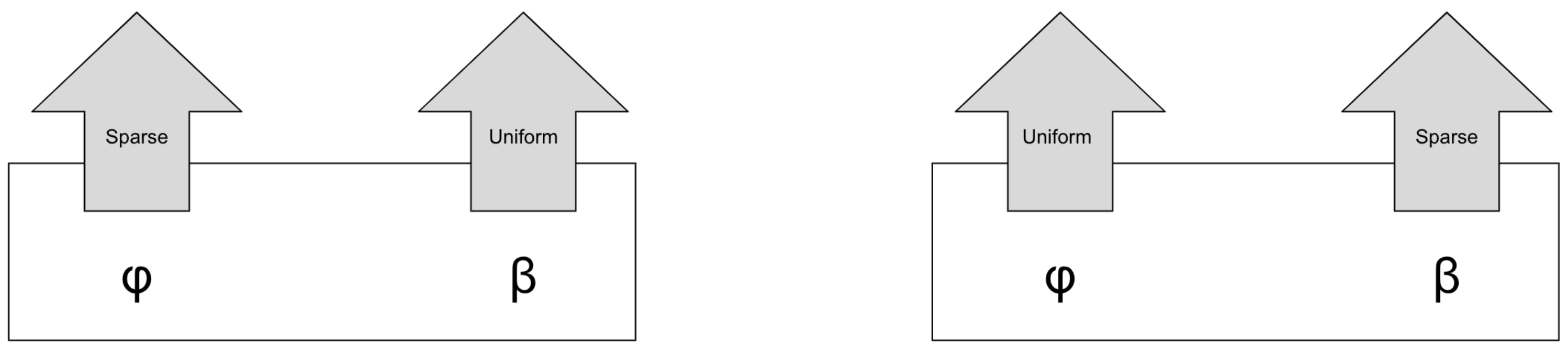

Figure: Conditioned on a constant θ value, as ϕ becomes more sparse, β must become more uniform and vise versa.

Results

- Several topics of the BNM model are consistently recognizable. For each topic, we can also yield interpretable results from topic-word distributions by finding the names that have the largest positive difference between the current topic's weighting of the name, and the average topic-word rating of the name across other topics. The following is the result of a 5 topic BNM model run.

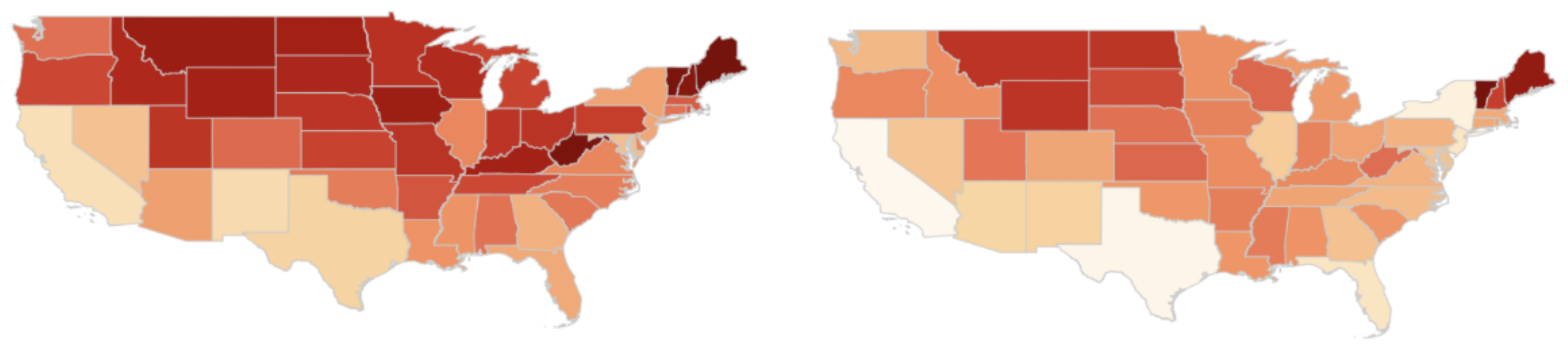

Figure: Caucasian population distribution (left), and Topic 1 from BNM Model result (right). Wyatt (M), Olivia (F), Harper (F), Elijah (M), Oliver (M) were the most distinct names in the topic.

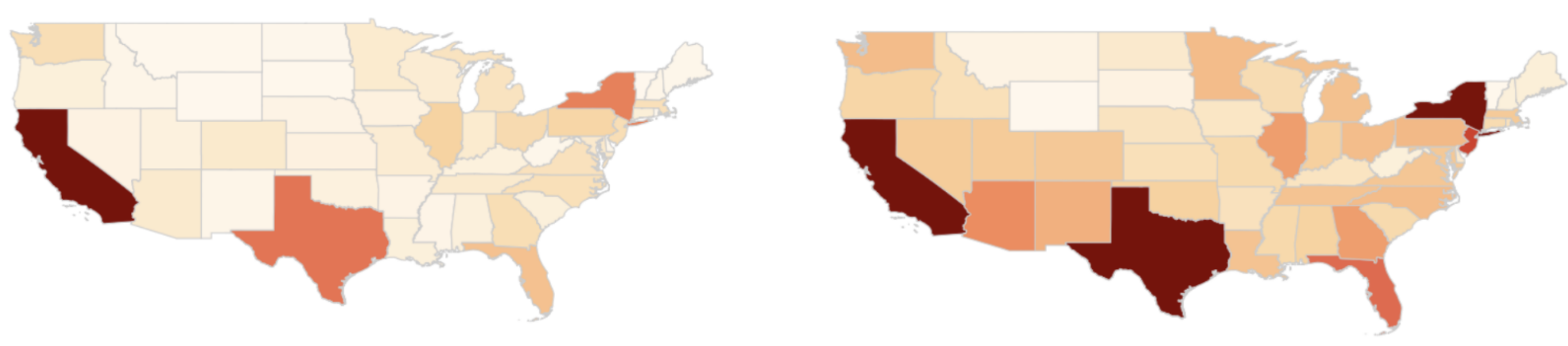

Figure: GDP distribution (left), and Topic 3 from BNM Model result (right). Daniel (M), Luke (M), Zoe (F), Aurora (F), Emily (F) were the most distinct names in the topic.

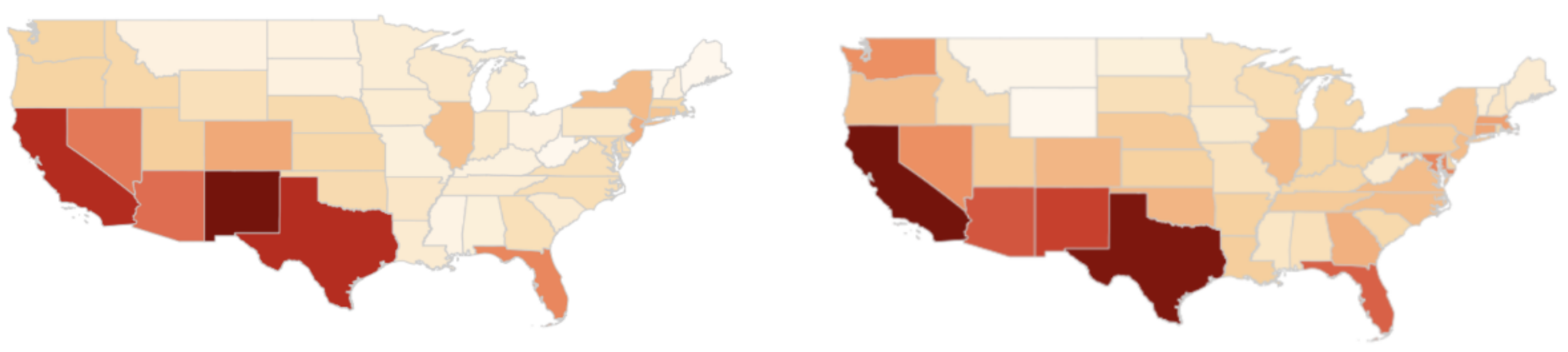

Figure: Hispanic and Latino population distribution (left), and Topic 4 from BNM Model result (right). Mateo (M), Sophia (F), Sebastian (M), Santiago (M), and Camila (F) were the most distinct names in the topic.