Final Project Report

The Baby Names Project

Hanming Yang (hy2781)

December 21, 2023

Abstract

Baby name data, despite rarely being the subject of academic research, offers the potential to be an unique perspective on socioeconomic trends. This paper introduces the Baby Name Model (BNM), a novel topic model that seeks to provide insight into baby name data from a socioeconomic perspective. With each name being a word and each state being a document, the BNM treats each topic as a unique factor of socioeconomic influence. By replacing the per-document topic proportions with a global topic proportion and introducing an additional distribution over documents for each topic, our approach enables topics to independently influence documents conditioned on the global topic proportion. Upon analyzing the topics generated by applying the BNM to the 2020 data set, we identified distinct topics, each correlating to different socioeconomic factors. We provide an implementation of the BNM at `https://github.com/Albertyangyhm/baby_name_model`.

1 Introduction

The Social Security Administration has been collecting data on the first names that are given to newborns since 1880 at SSA (2020). This data is available per year, on a state wide resolution. As there is little to no research that has been done on the data set, 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? Given the extensive availability of census and economic data for 2020, this report will primarily focus on that year.

Relating this work to the existing topic models, each baby name can be thought of as a word, and each state can be thought of as a document. Hence, having selected a certain year, it is possible to run Latent Dirichlet Allocation (LDA), introduced in Blei (2003), out of the box on this data set. This then allows us to define how a topic should be conceptualized. An interpretable topic is characterized as one whose prevalence in each document closely correlates with the impact of a specific socioeconomic factor on each state. We will refer to each topic interchangeably as a "factor". Factors may encompass aspects like political beliefs, gross domestic product (GDP), or demographic information.

2 Existing Models

There are several issues when applying LDA to baby name 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. However, a states wealth should not restrain a state's political orientation. As a consequence of this simplex constraint, a state that is immensely affluent and conservative may exhibit equal proportions of a factor representing high GDP when compared to states that are moderately affluent, but politically neutral.

This results in the distribution of a factor's presence in each document to not be immediately interpretable. Yet, socioeconomic data such a GDP is exactly a distribution over each state. Hence, the generative model of the LDA conflicts with the interpretation of topics in this context. Given these issues with the LDA, related models such as the Dynamic Topic Model (DTM) and the Supervised Latent Dirichlet Allocation (sLDA) also share these challenges under the current context.

3 The Baby Name Model (BNM)

To circumvent the issues that existing models face when modeling baby name data, we introduce the Baby Name Model (BNM). The BNM removes the per-document topic proportion simplex constraint by replacing per document proportions with a single global topic proportion, and providing each topic with an additional distribution over documents. This allows for each topic to have an interpretable topic-word distribution, as well as an interpretable topic-document distribution.

3.1 Generative Model

The new model's generative model is as follows.

1. For each topic k in K:
   (a) Draw from V dimensional Dirichlet to get the topic's distribution over words β_k
   (b) 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:
   (a) Draw topic assignment z_i from θ
   (b) Draw document assignment d_i from ϕ_{z_i}
   (c) Draw word x_i from β_{z_i}

The introduction of an additional topic-document distribution raises the consideration that topics might allocate greater weight to documents with longer lengths, implying that socioeconomic factors would appear more prominently in more populous states. We argue that this characteristic serves to enhance interpretability. First, consider that this bias is uniformly applied across all topics, ensuring equitable competition. Moreover, it is logical for socioeconomic factors to be more pronounced in states with larger populations. Take New York as an example: with its vast population and diverse demographics, it should naturally exhibit a multitude of factors to a significant degree. This model's structure, therefore, reflects a realistic representation of how different socioeconomic elements manifest in populous regions.

This generative model has the following joint probability expression, with the graphical model provided as a reference as well.

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

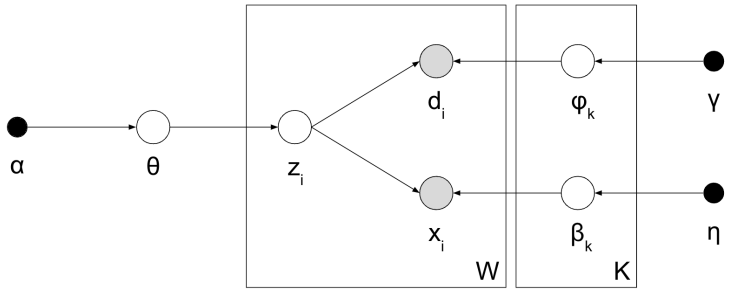

Figure 1: The BNM Graphical Model

3.2 Analyzing the Log Posterior

We now play the game of "staring at the log posterior" invented during Blei (2023b). Since the log posterior is dominated by the summation over words, let us consider how the three terms within the summation interact.

$$\sum_{i=1}^{W} (\log \theta_{z_i} + \log \phi_{z_i,d_i} + \log \beta_{z_i,x_i})$$

For $\log \theta_{z_i}$ to increase, θ needs to be come more sparse. This would mean that more names would be assigned to the same factor, 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 $\log \phi_{z_i,d_i} + \log \beta_{z_i,x_i}$. Equally, 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 $\log \phi_{z_i,d_i}$ increases by having a sparse ϕ, $\log \beta_{z_i,x_i}$ 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. Figure 6 and 7 in Appendix A illustrate the 3-way competition described above.

While this three-way competition between θ, ϕ, and β does allow for more flexibility in terms of hyperparameter tuning, it does make fine-tuning more complex. On the baby name data set, the BNM seemed to offer the most interpretability when θ's prior was slightly larger than that of ϕ and β. This allowed for the model to focus its representation power over the topics.

4 The BNM Gibbs Sampler

4.1 Complete Conditionals

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. While their complete conditionals are included in Appendix B, their Dirichlet distributions can be collapsed and represented as counts during implementation (much like the LDA Gibbs Sampler). The categorical posterior of z encapsulates the core mechanism of the BNM, whereby words are selected to optimize topic proportion, topic-word compatibility, and topic-document compatibility.

$$p(z_i = k | \boldsymbol{z_{-i}}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{d}, \boldsymbol{x}) \propto \frac{n_k + \alpha}{W + K\alpha} \cdot \frac{n_{kd} + \gamma}{n_k + N\gamma} \cdot \frac{n_{kv} + \eta}{n_k + V\eta}$$

Where:

- z_{-i}: All topic assignments excluding the i-th word.
- n_k: The total number of words assigned to topic k, not including the current instance of the word.
- n_{kd}: The number of times words in document d are assigned to topic k , not including the current instance of the word.
- n_{kv}: The number of times word v is assigned to topic k, across all documents, not including the current instance of the word.

4.2 Hyperparameter Tuning

The BNM generative model fails to generalize to new documents since each topic maintains a simplex over the documents in the training set. While this is suitable for the BNM, as documents represents each of the 50 states, it does mean that there is no obvious method to detect over-fitting through a validation set. In practice, the employment of a patience mechanism that halts training once in-sample likelihood plateaus provided consistently interpretable results. However, the randomness of the gibbs sampler did mean that the early stoppage would occur at different likelihoods depending on the initialization. These factors as a whole prevented quantifiable fine tuning, forcing hyperparameter selection to be done on the average training likelihood across several runs.

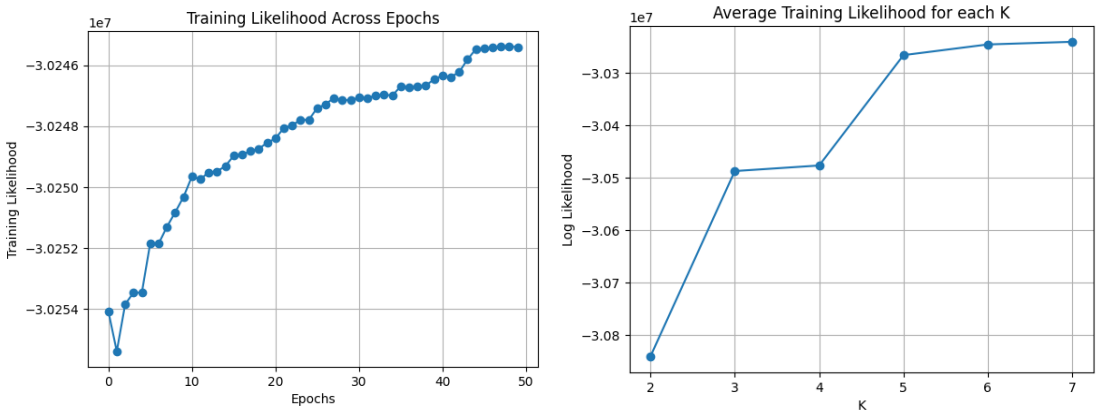

Figure 2: Training was halted when patience ran out. While $K = 5$ was chosen after comparing average training likelihood across 5 runs of the BNM.

4.3 Topic-Word and Topic-Document Interpretability

In traditional LDA, stop words are omitted because they do not contribute to discerning the specific topic of a given document. Similarly, in baby name data, there exists a subset of names that are prevalently used across all states (see Figure 8). However, unlike LDA, BNM does not exclude these commonly occurring names. This is because, in the context of baby names, the varying frequencies of popular names across different states yields valuable insights. Rather than eliminating common names, we can achieve topic-word interpretability by finding the names that exhibit the greatest positive deviation in their topic-specific weighting compared to their average frequency across other topics. This approach allows BNM to effectively highlight names that are uniquely characteristic of individual topics.

Various attempts were made to quantify topics-document interpretability. The most promising method first sets a threshold to determine if a topic is significant for a certain state. States meeting this threshold are considered to have the topic 'activated.' The model then applies a similar threshold to various socioeconomic factors. The factor with the closest match to the distribution of states for a given topic, measured by the Jaccard distance, is identified as related to that topic. While this method worked on certain runs of the BNM, effective threshhold values deviated arbitrarily across each topic on other runs, resulting in inconsistent quantification of interpretability. Attempts were also made to use regression to relate topics and socioeconomic factors. However, this approach performed poorly. The regression models had virtually no predictive power over

the validation set. Including too many socioeconomic factors also resulted in overfitting, particularly given that there are only 50 data points. However, in the context of LDA, topic recognition is not quantitatively measured either. For example, if a topic relates to biology, we deem it interpretable if it includes words like 'DNA', 'cells', and 'animals'. Referring to Figure 3, 4, & 5, we argue that our results show sufficient resemblance to the distribution of specific socioeconomic factors to be considered interpretable.

5 Results

Several topics of the BNM model are consistently recognizable. The following is the result of a 5 topic BNM model run (the 2 remaining topics are shown in Appendix C). Further work is required to develop a reliable method for held-out validation of the model. On the other hand, while the Jaccard distance-based approach offers a promising way to quantify topic interpretability, improvements are needed in fine-tuning the threshold that determines a state's inclusion in the set of states activated by each factor.

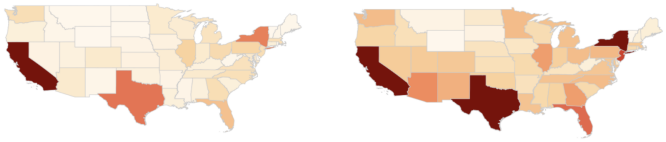

Figure 3: GDP distribution (left), and Topic 3 from BNM Model result (right). Daniel (M), Luke (M), Zoe (F), Aurora (F), Emily (F) were the most distinct names in the topic.

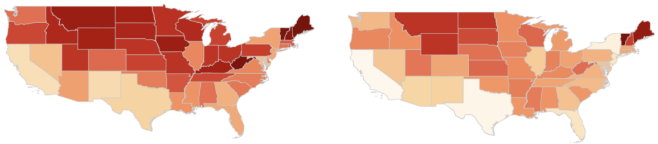

Figure 4: Caucasian population distribution (left), and Topic 1 from BNM Model result (right). Wyatt (M), Olivia (F), Harper (F), Elijah (M), Oliver (M) were the most distinct names in the topic.

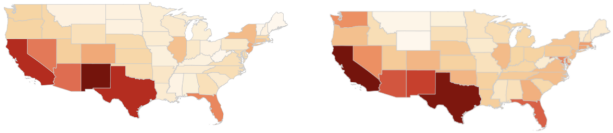

Figure 5: Hispanic and Latino population distribution (left), and Topic 4 from BNM Model result (right). Mateo (M), Sophia (F), Sebastian (M), Santiago (M), and Camila (F) were the most distinct names in the topic.

References

David M. Blei. *Probabilistic Models*. Columbia University, 2023.

David M. Blei. *Lectures in Probabilistic Models and Machine Learning*, 2023. Course conducted at Columbia University.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3:993–1022, 2003.

United States Census Bureau. *2020 Census*, 2020. Online: `https://www.census.gov/programs-surveys/decennial-census/decade/2020/2020-census-results.html`. Accessed: Fall 2023.

Bureau of Economic Analysis. *Gross Domestic Product*, 2020. Online: `https://www.bea.gov/data/gdp/gross-domestic-product`. Accessed: Fall 2023.

Social Security Administration. *Popular Baby Names*, 2020. Online: `https://www.ssa.gov/oact/babynames/limits.html`. Accessed: Fall 2023.

Appendix A Supplementary Figures for the Log Posterior

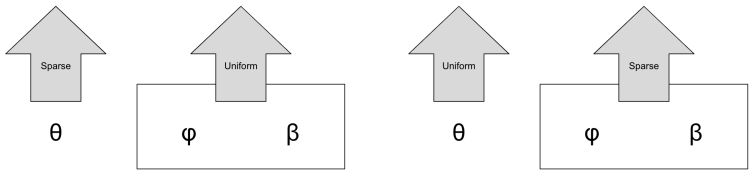

Figure 6: As θ becomes more sparse, ϕ and β together must become more uniform and vise versa.

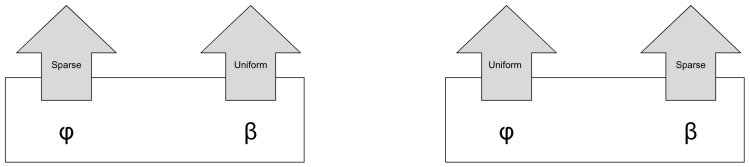

Figure 7: Conditioned on a constant θ value, as ϕ becomes more sparse, β must become more uniform and vise versa.

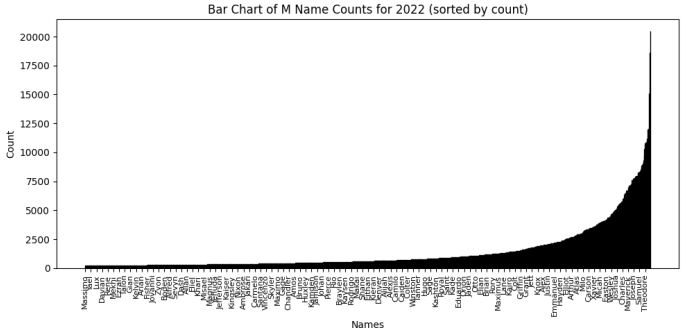

Figure 8: 2022 Distribution of Male Baby Name Counts

Appendix B Complete Conditionals

Results based on dirichlet-categorical conjugacy results in Chapter 2 of Blei (2023a).

$$p(\theta|z, \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{d}, \boldsymbol{x}) \sim \text{Dirichlet}(\alpha + n_k)$$
$$p(\beta_k|z, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\beta_{-k}}, \boldsymbol{d}, \boldsymbol{x}) \sim \text{Dirichlet}(\eta + n_{kv})$$
$$p(\phi_k|z, \boldsymbol{\theta}, \boldsymbol{\phi_{-k}}, \boldsymbol{\beta}, \boldsymbol{d}, \boldsymbol{x}) \sim \text{Dirichlet}(\gamma + n_{kd})$$

Appendix C Additional Results

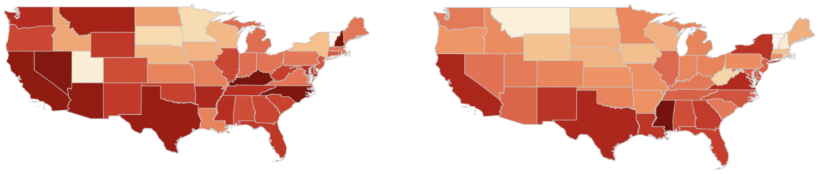

Figure 9: Topics 2 and 5 from the BNM run discussed in the results section. Topic 2 displays a relatively uniform distribution across states, potentially representing residual words not captured by other more distinct topics. Topic 5 shows a distribution that is the inverse of Topic 1's. While this characteristic renders Topic 5 interpretable, it fails to provide any new insights beyond what is already offered by Topic 1.

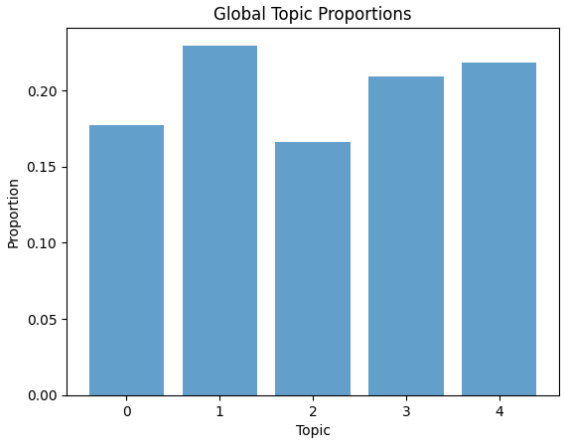

Figure 10: Global topic proportions of the BNM run discussed in the results section.