hqt part in final essay

Qitian (Jason) Hu

March 17, 2021

1 structural topic model

The raw topic model (raw TM) only adds topic and document as internal organization of the corpus, but in practice, we usually have a much larger range of metadata. Structural topic modeling (sTM [**?**]) allows us to add document-level metadata on which topical prevalence or topical content is of interest. The main idea is to use generalized linear models as priors and then condition on metadata of the document.

Based on three variants of LDA [**?**], correlated topic model (CTM, [**?**]), the Dirichlet-Multinomial Regression topic model (DMR, [**?**]) and the Sparse Additive Generative (SAGE, [**?**]) topic model.

The logistic normal prior for topics in the standard CTM is replaced by a logistic-normal linear model. The design matrix for the covariates X allows for flexible forms of the covariates that the prior could be conditioned on.

The distribution over words is replaced with a multinomial logistic function such that a token's distribution is jointly determined by topic, covariates, and topic-covariate interaction. We adopted the R implementation provided by the authors [**?**].

A more specific description of the generative process is described here.

1. Draw the document-level attention to each topic from a logistic-normal generalized linear model based on a vector of document covariates X_d.

$$\vec{\theta}_d \mid X_d\gamma, \Sigma \sim \text{Logistic Normal}\left(\mu = X_d\gamma, \Sigma\right)$$

   where X_d is a 1-by- p vector, γ is a p -by- $K-1$ matrix of coefficients and Σ is $K-1$ -by$K-1$ covariance matrix.

2. Given a document-level content covariate y_d, form the document-specific distribution over words representing each topic (k) using the baseline word distribution (m), the topic specific deviation $\kappa_k^{(\text{t})}$, the covariate group deviation $\kappa_{y_d}^{(\text{c})}$ and the interaction between the two $\kappa_{y_d,k}^{(\text{i})}\Big)$

$$\beta_{d,k} \propto \exp\left(m + \kappa_k^{(\text{t})} + \kappa_{y_d}^{(\text{c})} + \kappa_{y_d,k}^{(\text{i})}\right)$$

   m, and each $\kappa_k^{(\text{t})}, \kappa_{y_d}^{(\text{c})}$ and $\kappa_{y_d,k}^{(\text{i})}$ are V -length vectors containing one entry per word in the vocabulary. When no convent covariate is present β can be formed as $\beta_{d,k} \propto \exp\left(m + \kappa_k^{(\text{t})}\right)$ or simply point estimated (this latter behavior is the default).

3. For each word in the document, $(n \in 1, \ldots, N_d)$ - Draw word's topic assignment based on the document-specific distribution over topics.

$$z_{d,n} \mid \vec{\theta}_d \sim \text{Multinomial}\left(\vec{\theta}_d\right)$$

   - Conditional on the topic chosen, draw an observed word from that topic.

$$w_{d,n} \mid z_{d,n}, \beta_{d,k=z_{d,n}} \sim \text{Multinomial}\left(\beta_{d,k=z_{d,n}}\right)$$

2 comparison between the two methods

We choose to use People's Daily to compare the two topic models, because

1. The textual quality of the corpus is better
2. It is large and thick enough to demonstrate the difference
3. It has a long time frame so structural topic model might be effective

Although we used high-performance 128G-memory server, structural topic model could not be run probably, probably due to the limit of R language. Thus, we removed every word that appears in less than 10 documents or appear in more than half of all the documents. A brief comparison with the topics using the original corpus using raw topic model reveals that the quality of topics is not severely affected, and major historical themes are correctly identified.

We compare raw topic model and structural topic from 3 aspects. Firstly, we interpretatively examine the topics and check if important historical themes are captured. We discovered that both models accurately captured the themes of anti-imperialism, Chairman Mao and cultural revolution, socialism and communism, United Nations and diplomacy, Korea and Vietnam wars, market economy, firm and investment, etc. sTM tends to capture more detailed topics, like Israel and Palestine, India and Pakistan, and municipal constructions, while these topics were not captured by raw TM.

Secondly, to quantify the topic difference, we calculated the average number of characters of each word selected by the two topic models. Note that each Chinese word is composed by a number of characters, and usually the longer the word is, the more detailed and specific it is. The average word length for raw TM is 2.09 (std = 0.789), while that of sTM is 3.28 (std = 0.591). sTM tends to capture more detailed, specific words.

Third, we visualize the time range of documents in each topic. As the figure shows, while raw TM tends to capture topics that capture documents around 1976, sTM topics could vary in the mean year of related documents. Time range of Raw TM topics have standard deviation centered at 17 year, while sTM topics could be very specific to a period (low standard deviation) or vice versa.

To sum up, raw TM topics are more general and usually capture documents in a broader period. With metadata on the temporal change of prevalence of topics, sTM captures more specific topics of specific historical periods.

3 Added Exploration

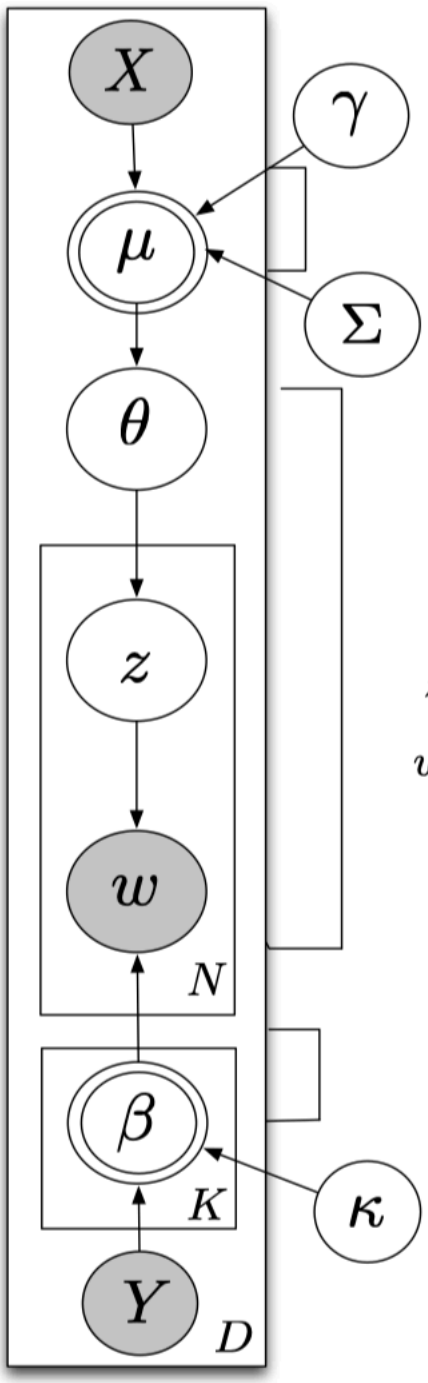
3.1 People's Daily Exploration

People's Daily is the largest newspaper group in China, and is the official newspaper of the Central Committee of the Chinese Communist Party. It was founded in 1948 and represents the official ideology and policy of the Party. The corpus we we has about 1.3 million articles, ranging from 1948 until 2003.

3.1.1 Confirmation of Historical Knowledge

4 further exploration

1. topic number selection
2. topic evaluation



Topic Prevalence:

$$\mu_{d,k} = X_d \gamma_k$$
$$\gamma_k \sim \mathcal{N}(0, \sigma_k^2)$$
$$\sigma_k^2 \sim \text{Gamma}(s^\gamma, r^\gamma)$$

Language Model:

$$\theta_d \sim \text{LogisticNormal}(\mu_d, \Sigma)$$
$$z_{d,n} \sim \text{Mult}(\theta_d)$$
$$w_{d,n} \sim \text{Mult}(\beta_d^{k=z_{d,n}})$$

Topical Content:

$$\beta_{d,v}^k \propto \exp(m_v + \kappa_v^{:,k} + \kappa_v^{y,\cdot} + \kappa_v^{y,k})$$
$$\kappa_v^{y,k} \sim \text{Laplace}(0, \tau_v^{y,k})$$
$$\tau_v^{y,k} \sim \text{Gamma}(s^\kappa, r^\kappa)$$

Figure 1: mechanism of structural topic model