STCS 6701: Foundations of Graphical Models: Homework 2

Maxime TCHIBOZO (MT3390)

Tuesday December 1, 2020

1 Coordinate Ascent Variational Inference for LDA

Latent Dirichlet Allocation assumes a Mixed-Membership generative process to summarize text data using distributions of topics over a vocabulary of words. We implement a Variational Inference (VI) algorithm to approximate both topic distributions and topic proportions for the Associated Press dataset, which contains 2246 documents with a vocabulary of 10474 unique words. Unless otherwise specified, we use the entire dataset for our analysis.

Assessing model convergence

CAVI parameters for the variational family q are initialized randomly according to gamma distributions. We run the same VI algorithm 5 times with random initializations, and find that the variational objective function (ELBO) converges for each run (Fig. 1). Since VI is an approximate inference method, each ELBO converges to a different value, indicating that the parameters for q converge to local optima.

Checking coherence of LDA topics

After checking convergence, we evaluate if the local optima reached by CAVI are informative for LDA summarization. We check both the coherence of the topic distributions over words (i.e do the LDA topics group words which are actually similar together?) and coherence of document distributions over topics (i.e do the documents actually contain the topics LDA assigned them to?).

A classic method for LDA topic behavior assessment is to look at the words with highest probability for each topic. Fig. 2 shows the top words for each topic for each run of Fig. 1. These topics seem to clearly segregate between politics (domestic), economics, politics (foreign), justice and law & order.

We also suggest new methods to quantitatively compare how well our method captures meaning using word2vec embeddings (which we will consider to be a ground truth for word similarity). We find that the LDA topics naturally cluster words with similar word2vec embeddings together (Fig. 3, Fig. 4). To confirm the coherence of the topic assignments, we find that a KMeans clustering using only topic proportions yields similar results to a KMeans clustering using average word2vec embeddings for each document (Fig. 5). This word2vec embedding approach also allows us to **automate** the model checking process.

Model performance on held-out data

We divide a subset of the data into an (x_in, x_out) in-sample and out-of-sample split (OOS). We estimate topic distributions over words using the in-sample split and use these topics to perform a MAP approximation of the topic proportions θ^{OOS} and topic assignments z on the out-of-sample set. Using the MAP estimate approxation, we compute the log-likelihood of observing the out-of-sample documents for different values of K. For computational purposes, we restrict the in-sample dataset to 561 documents, and the out-of-sample dataset to 10 documents. The log-likelihood is plotted for K varying from 2 to 50 in increments of 3. Fig. 6 shows the optimal K with this method is 2. For future analyses, we could increase the size of the out-of-sample dataset and the range of the K values.

CAVI sensitivity to hyperparameters

Our CAVI algorithm relies upon assumptions on our prior belief of ϕ (MFVI approximation concentration parameter for the word distribution over topics β_k) and α (concentration parameter for the document distribution over topics θ_d). Specifically, Fig. 1 was obtained for $K = 5$ and Dirichlet prior concentration parameters $\phi = 0.1$ and $\alpha = 10$.

To verify that our method converges independently of both hyper-priors (ϕ, α) and hyper-parameters (K), we plot the ELBO as a function of iteration for 9 different prior parameter configurations ($\phi \in \{0.1, 1, 10\}$ and $\alpha \in \{0.1, 1, 10\}$) with K=10 on a subset of the data (first 33.3% of documents). Fig. 7 shows that our CAVI method converges for all configurations in the grid, although these methods converge to different local optima, have different speeds of convergence, and may be subject to more or less abrupt increases in the ELBO.

Challenges faced (ordered by decreasing order of inconvenience):

- **Working with continous and discrete variables simultaneously.**

LDA uses both discrete (z) and continuous (θ) latent variables. This means optimization problems over both variables become NP-hard. Notably, I was unable to optimize over all combinations of parameters z for the MAP estimation of the in-of-sample topic parameters ϕ_k to be used on held-out data. A brute force evaluation of all possible z_i configurations led to an exponential number of values of z_{OOS} to evaluate, so I resorted to a probabilistic approach, approximating the out-of-sample MAP parameters θ, z with parameters that yielded the best posterior log-likelihood over 10,000 randomly selected configurations of z. The issues this raises are twofold: as K increases, since the out-of-sample dataset is of fixed size, the curse of dimensionality makes it more likely that some clusters be assigned no points, and we can no longer estimate the in-sample $\theta_d^{\hat{MAP}}$ parameters. Secondly, this stochastic method only gives a local optima of (z^{MAP}, θ^{MAP}).

- **Exploding values for $\log(\Gamma(\sum ...))$ expressions where the sum is over the entire vocabulary V or all documents in D.**

A naïve implementation led the Γ function to explode to infinite values, rendering the log of this infinite value impossible to evaluate. It took time to solve this problem (using the scipy gammaln function).

- **Optimizing the code for faster runs without affecting its exactness.**

While we can usually optimize code using matrix multiplications or vectorizations instead of for loops, in many cases I found myself unable to optimize the code without affecting the order in which variables where updated, which would have led to an incorrect algorithm.

- **Memory and time requirements of these algorithms (saving all the parameters requires ~ 1GB of memory).**

2 Final Project: Aspirational Abstract

By design, Machine Learning models are trained to maximize predictive performance on a given dataset. Models can achieve accurate predictions through incorrect associations, if the data at hand is not adequately explored, interpreted and curated. Notably, machines trained on datasets containing racial, gender or disability bias may learn to internalize this bias, and perpetuate unwanted stereotypes when making predictions.

We use data from flagged social media comments to quantify the propensity for words or expressions to convey different types of bias and toxicity. Our algorithms build on statistical methods such as Supervised LDA and Structural Topic Models to identify bias and mitigate its pervasiveness in documents. We then compare different debiasing strategies to remove ambiguous connotations in text using an estimated treatment effect framework.

3 Appendix: Tables, Figures for Problem 1

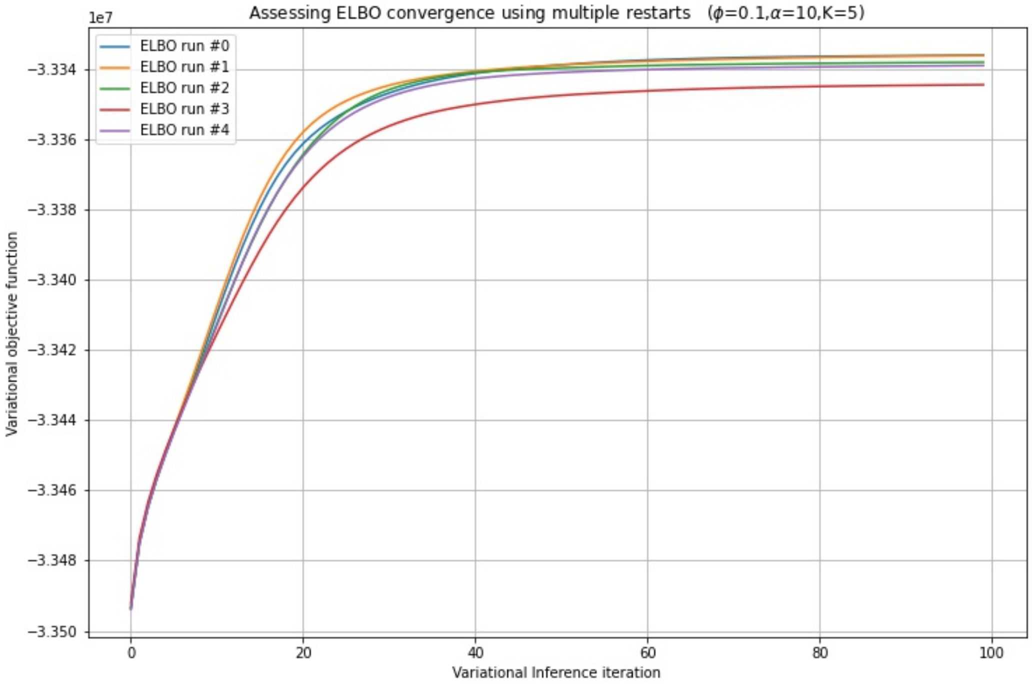
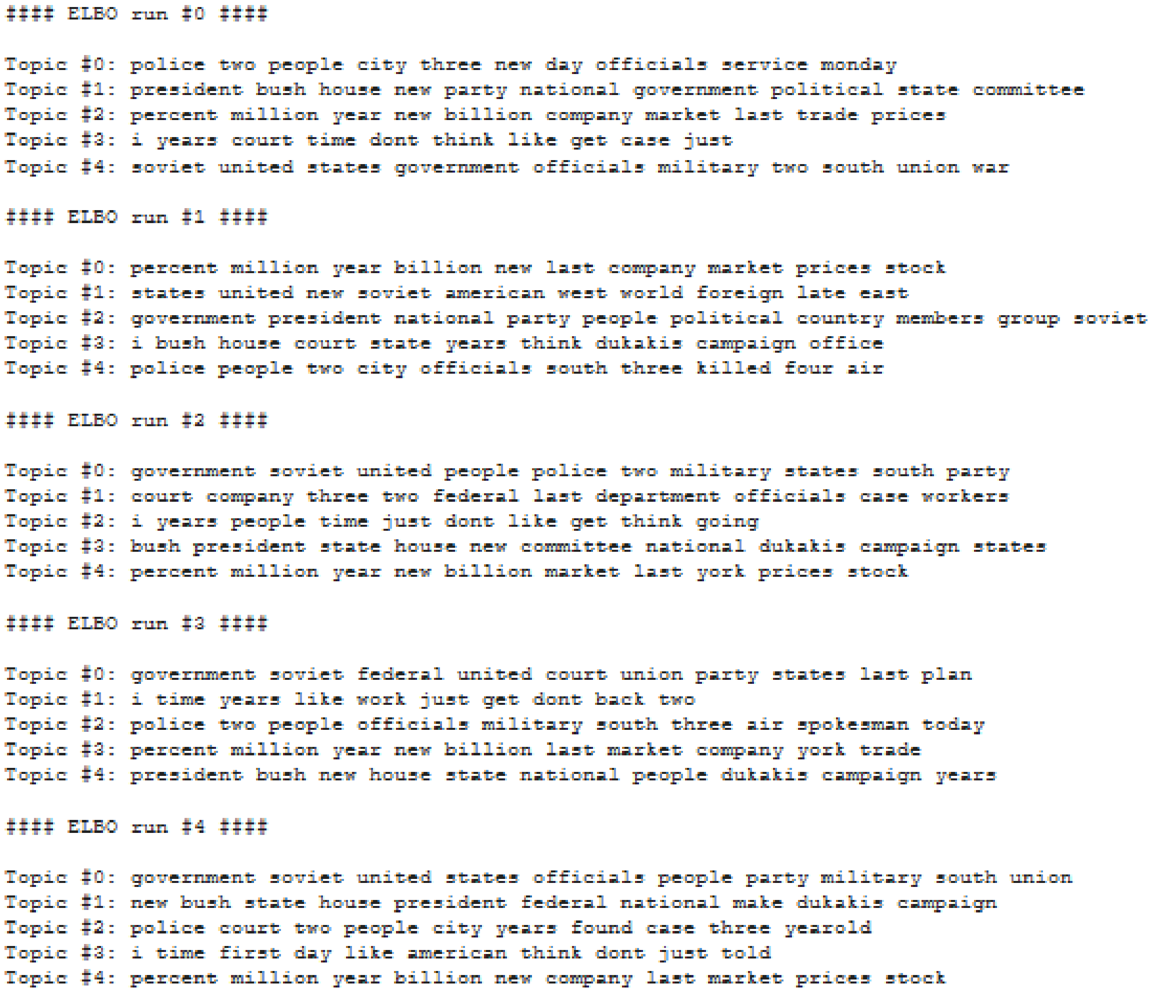


Figure 1: Assessing model convergence: ELBO as a function of iteration



```
#### ELBO run #0 ####

Topic #0: police two people city three new day officials service monday
Topic #1: president bush house new party national government political state committee
Topic #2: percent million year new billion company market last trade prices
Topic #3: i years court time dont think like get case just
Topic #4: soviet united states government officials military two south union war

#### ELBO run #1 ####

Topic #0: percent million year billion new last company market prices stock
Topic #1: states united new soviet american west world foreign late east
Topic #2: government president national party people political country members group soviet
Topic #3: i bush house court state years think dukakis campaign office
Topic #4: police people two city officials south three killed four air

#### ELBO run #2 ####

Topic #0: government soviet united people police two military states south party
Topic #1: court company three two federal last department officials case workers
Topic #2: i years people time just dont like get think going
Topic #3: bush president state house new committee national dukakis campaign states
Topic #4: percent million year new billion market last york prices stock

#### ELBO run #3 ####

Topic #0: government soviet federal united court union party states last plan
Topic #1: i time years like work just get dont back two
Topic #2: police two people officials military south three air spokesman today
Topic #3: percent million year new billion last market company york trade
Topic #4: president bush new house state national people dukakis campaign years

#### ELBO run #4 ####

Topic #0: government soviet united states officials people party military south union
Topic #1: new bush state house president federal national make dukakis campaign
Topic #2: police court two people city years found case three yearold
Topic #3: i time first day like american think dont just told
Topic #4: percent million year billion new company last market prices stock
```

Figure 2: Assessing model behavior: Top Words for each Topic (K=5)

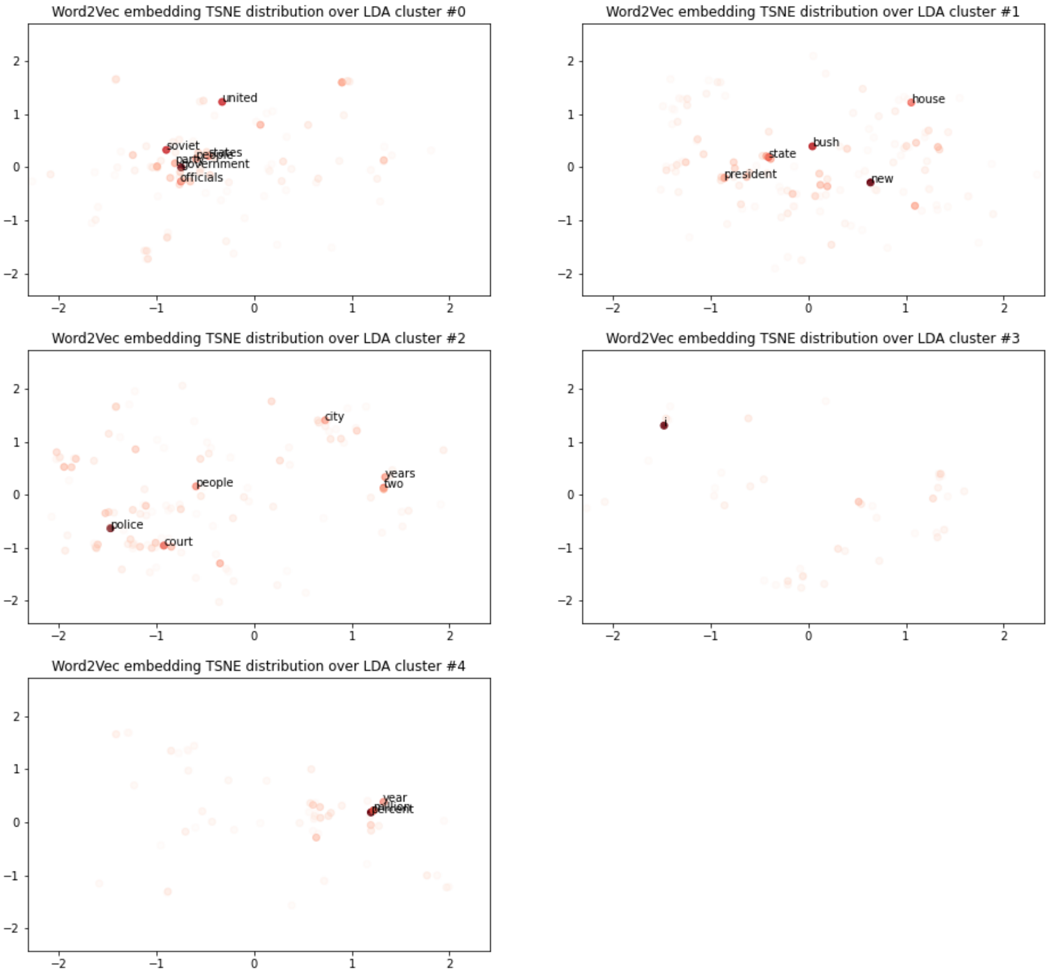


Figure 3: Distribution of topics over words in the word2vec embedding space (T-SNE dimensionality reduction). Red point intensity is equal to the Dirichlet parameter probability $(\phi_{k,v})$ for each word v probability in topic k

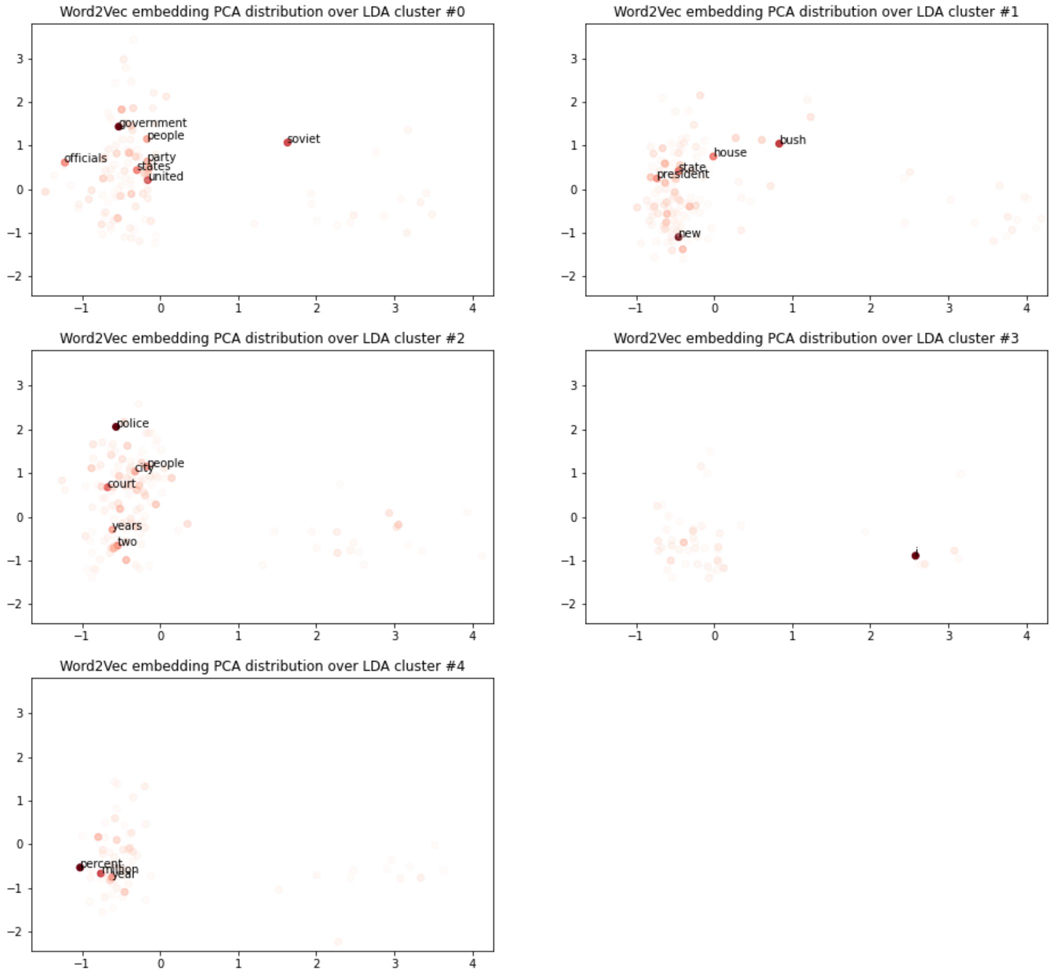


Figure 4: Distribution of topics over words in the word2vec embedding space (PCA dimensionality reduction). Red point intensity is equal to the Dirichlet parameter probability $(\phi_{k,v})$ for each word v probability in topic k

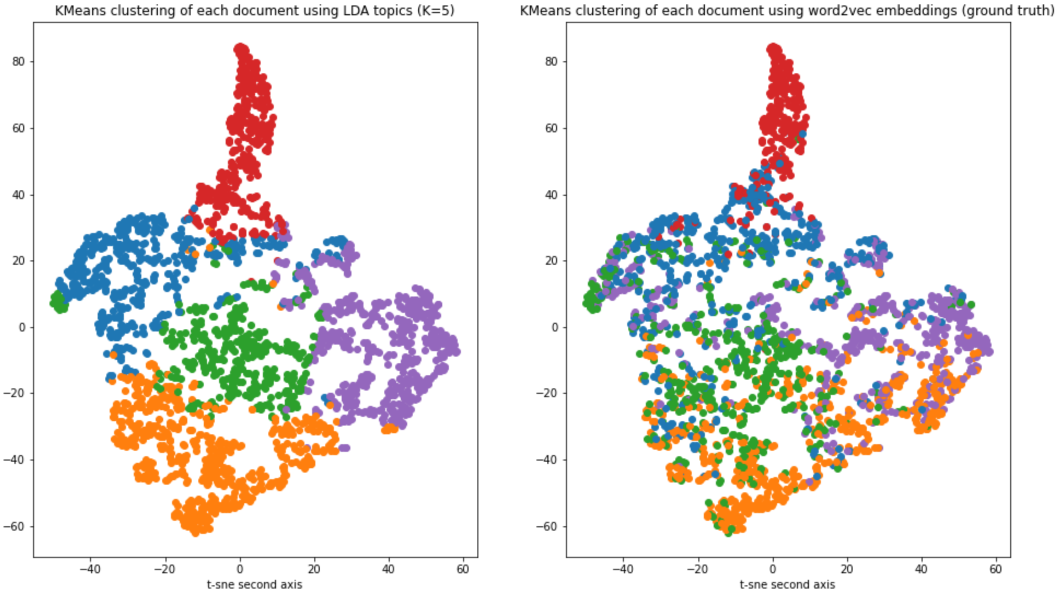


Figure 5: Comparison of KMeans clustering of all documents using their LDA 5-dimensional topic proportions (left ; K=5) and their 300-dimensional average word2vec embeddings (right)

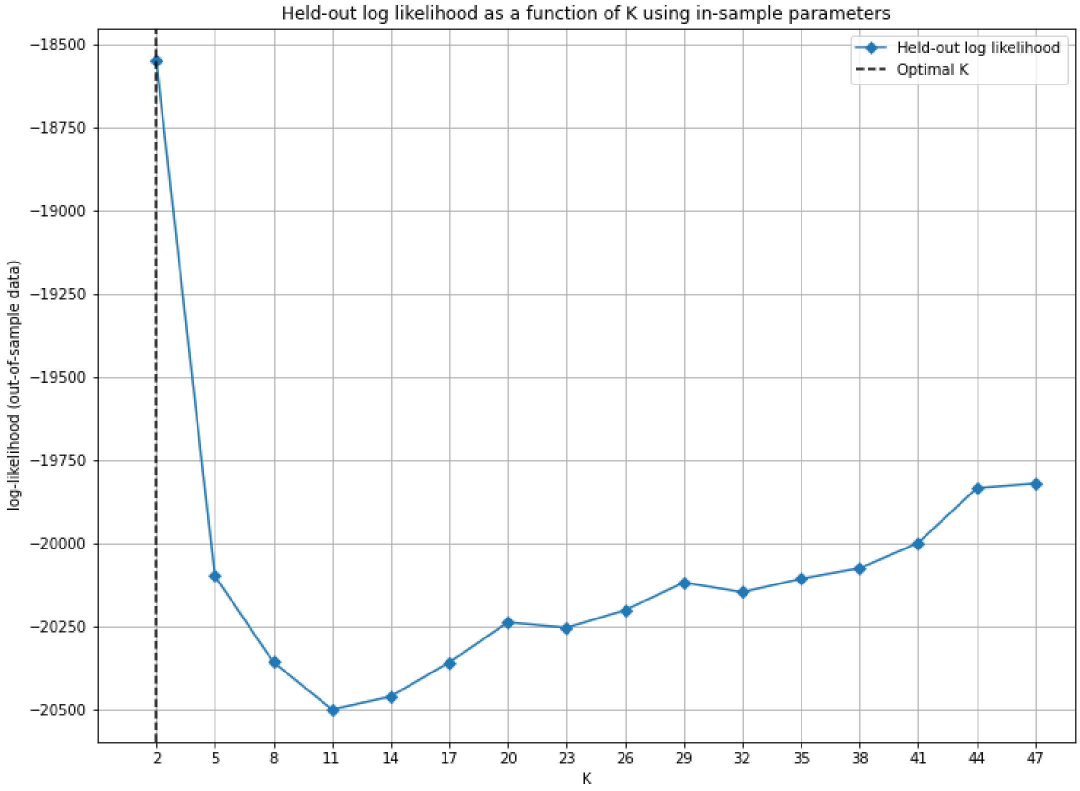


Figure 6: Held-out log likelihood using in-sample MAP estimated topic proportions ϕ_k^{in} for different values of k ($|x_{in}| = 561$ documents, $|x_{out}| = 10$ documents)

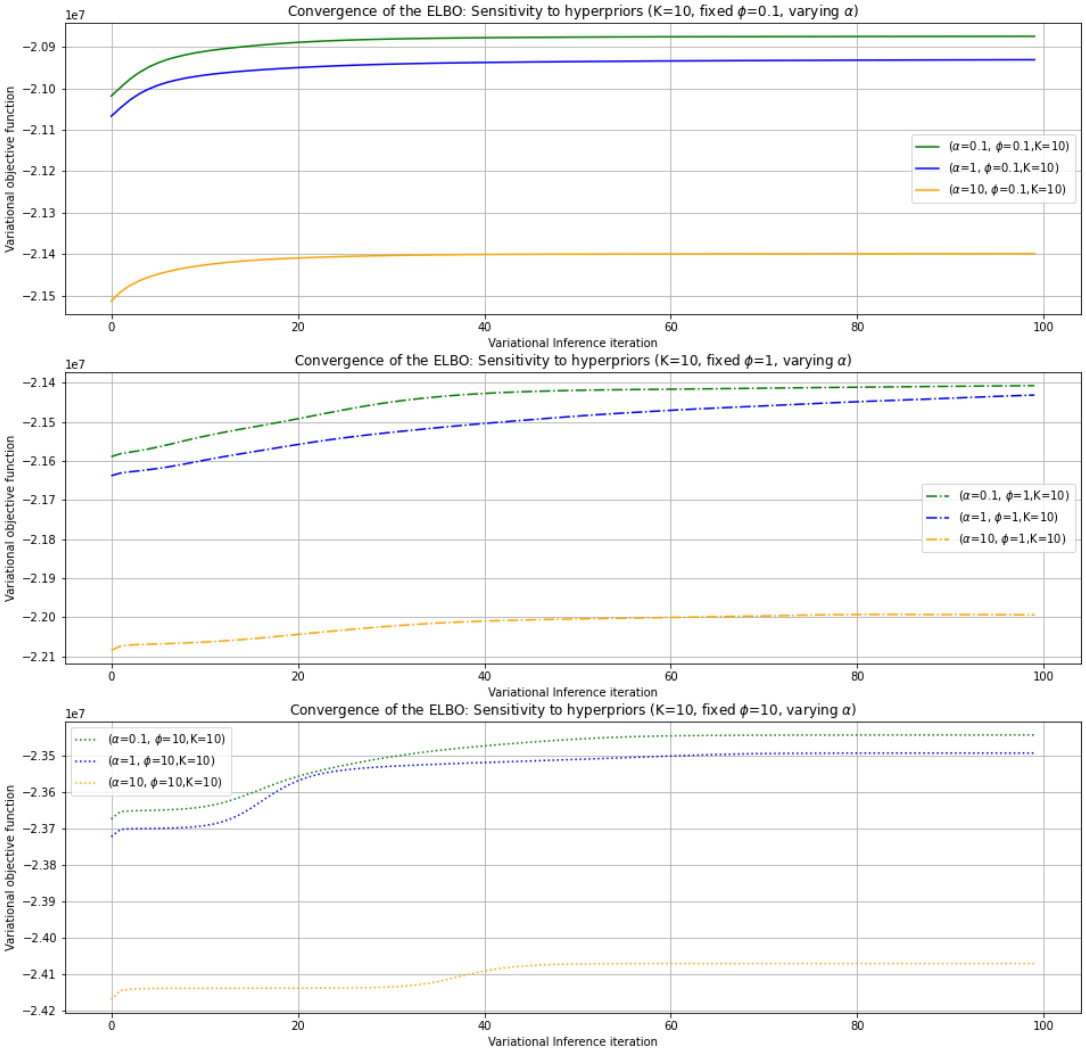


Figure 7: ELBO sensitivity to hyperparameters