Topic Models are Statistically Inconsistent, More Biased with More Data, and Substantively Misleading: Evidence, Solutions, and Applications*

Danny Ebanks† Sara Kangaslahti ‡

September 8, 2024

Updated Regularly: Latest Version Here.

Abstract

We demonstrate that conventional topic models (1) are statistically inconsistent, (2) are more biased as we put incrementally more effort into collecting additional data, and (3) lead to misleading substantive results, which we correct with our consistent estimator for topic models. We conclude that 36,000 papers, including 10,000 in social and political science, must be re-estimated with consistent and unbiased topic model methods or discarded altogether. These papers include those relying on the original Latent Dirichlet Allocation method, structural topic models, and every other topic model method that relies on variational inference.

We illustrate substantive bias by demonstrating that standard topic models mislead one to conclude that ideological polarization is decreasing when, in fact, it is increasing. We then show that using these same methods misleads researchers by suggesting party-leader agreement is increasing when it is, in fact, decreasing. Finally, we show the promise of unbiased and consistent text methods by illustrating results that would not be otherwise possible.
Words: 6300

*Our thanks to Mike Alvarez, Jonathan N. Katz, Gary King, Dominic Skinnion, and audiences at PolMeth for many helpful comments.

†Postdoctoral Fellow, Institute for Quantitative Social Science, Harvard University; DannyEbanks.com, DEbanks@g.harvard.edu.

‡First-Year Graduate Student, Computer Science, Harvard University;

1 Introduction

Political scientists have long operated under the notion that as they collect more and more data, their efforts to collect each additional piece of data will earn them more accurate findings. The ideal that researchers learn more about the world as they collect more data constitutes one of the fundamental tenets of applied political methodology. That is, our statistical estimators should be statistically consistent. In this paper, we show substantive problems when this principle is violated and demonstrate for the first time under realistic conditions, both theoretically and empirically, that topic models are asymptotically inconsistent when estimated via variational inference. The intuition for why we have this result is simple: as we collect incrementally more data, there are an increasing number of opportunities for topics to be incorrectly combined or split apart. We further show that such topic models are not only inconsistent but that the social scientists' natural intuitions about data collection are precisely the opposite of reality when using topic models. The marginal spent effort as data are collected results in increasingly biased substantive results. The stunning implication of this paper is that topic models favored by researchers would perform better if researchers collected less data, not more. Whether endeavoring to classify open-ended responses in a qualitative study or downloading terabytes of text data from social media, researchers should have confidence that their statistical methods provide increasingly correct answers for the additional effort to collect more data.

In the best case, nearly 10,000 papers in applied political and social science and 36,000 papers that use the method would need to be re-estimated using a topic model method with appropriate large-sample accuracy guarantees or discarded entirely.[1] To illustrate how the existing method misleads researchers, we show that when a consistent and unbiased method is applied, we obtain substantially different and more plausible results than biased approaches. For this purpose, we use our correct approach, which we fully de-

[1] We calculated these numbers by looking at the most popular topic model methods that use variational inference. In total, we find about 60,000 papers cite to methods that use variational inference. We then randomly sampled 100 papers that cite LDA, Correlated Topic models, Dynamic Topic models, Structural Topic Models, and other methods that use variational inference methods. We found that 60 percent of the sample were papers that used inconsistent methods for applied purposes or as inputs for other methods. Of the total sample, 16 percent concerned the political and social sciences.

scribe in Kangaslahti et al., 2023. By using a consistent and unbiased approach such as ours, we ultimately find polarization has fluctuated over time, with several instances of depolarization over the last 120 years. In contrast, biased measures suggest that polarization is an emergent feature of the 21st century. Further, we find that party-leader agreement has *decreased* since 1891, whereas inconsistent approaches misleadingly suggest it has increased. Finally, we show that our speech-based approaches provide better and more plausible estimates than popular approaches to measuring party-leader agreement and polarization derived from roll-call votes, renewing the promise of these text-based methods. We show in this paper that unsupervised text methods can yet provide new insights into essential quantities of interest for political scientists.

We formally analyze the large sample properties of topic models as detailed in Blei, Ng, and Jordan (2003b). Here, we define consistency in the traditional sense: as more observations for a dataset are collected, the statistical estimates of the model should approach the actual value. We show that topic models are inconsistent under this most natural definition – when we add more and more observations.[2]

Up to now, variational inference methods have been greatly helpful to computational social science and studies of political science text, especially because they were computationally convenient and easy to implement. This estimation method allowed for a generation of scholars to study the latent structure of text for the first time and demonstrate important findings in political science (Grimmer, 2011; Grimmer and Stewart, 2013; Grimmer, Roberts, and Stewart, 2022). With the new theoretical insights presented here, combined with increasingly powerful computational technology, we demonstrate how we can build upon this rich methodological tradition and improve applied empirical work in text.

We establish that topic models estimated via variational inference fail to converge to the truth *precisely* in the cases where researchers most need it: when there are a large number of documents. In these same cases, researchers may not be able to sample from the data to improve inference. Topics of critical importance for a substantive question are

[2]We build on the theoretical framework from Nakajima et al. (2014), which considers many interesting and important cases, but not the one of most interest to political scientists and social scientists writ large. We extend it to these natural natural cases under relaxed assumptions.

often highly concentrated in a small number of documents. In this case, sampling might not capture the important small-sample quantity of direct interest for the researcher. This result implicates popular approaches such as the original topic model method, structural topic models, and any model estimated with variational inference.

We demonstrate that the effects of using a consistent and unbiased method significantly change our understanding of even the most basic text analysis. In section 7, we discuss common theoretical approaches, which, while theoretically elegant, apply only under the most unrealistic conditions or define consistency in nontraditional ways to achieve a desirable result.

We then show that this inconsistency is not merely a trivial theoretical artifact of interest to statisticians. The substantive bias is significant and unmistakable. To demonstrate the effects of the statistical inconsistency on our substantive understanding of essential quantities of interest, we show how applying a consistent and unbiased method – in this case, a tensor decomposition method that we developed in Kangaslahti et al. (2023) – substantially alters our understanding of how ideological polarization party-leader agreement evolved since the late 1800s.[3]

Specifically, we use a consistent and unbiased topic model method to study two essential quantities for studying legislatures: polarization and party-leader agreement. By using a consistent and unbiased topic model framework (see Kangaslahti et al. (2023)), we provide the first statistically consistent estimates of the topic-word and document-topic probabilities jointly for a large-scale political science text dataset of Congressional floor speeches. We show that existing topic model methods produce substantive bias because they cannot consistently identify topics in the text, without bias. The inconsistent methods inappropriately combine, divide, or ignore topics in the text; worse yet, we cannot know what went wrong unless compared against a consistent and unbiased model's estimates. In the case of Congressional speeches, popular topic models overestimate party-leader agreement and underestimate ideological polarization. Then, we demonstrate potential

[3]In Kangaslahti et al. (2023) and Yao et al. (2018), we showed that traditional variational inference methods fail in practice. However, here we contribute by formalizing the theoretical best-case scenarios for VI and establishing that they will never work when data are voluminous and unstructured, precisely when they should be more helpful.

sources of substantive bias in the case of historical congressional speeches. In this case, one topic, trade, is discussed using changing contexts over 120 years. We show in this case that popular topic models fail to recognize which words belong to a topic related to trade because the words used to discuss it change over time.

We compare this against a popular approach – using roll-call votes– to study polarization and party-leader agreement. We provide qualitative evidence that our estimates are more plausible than these traditional measures. We compare these measures to vote-based measures for several reasons. First, they are the standard tools for measuring party-leader agreement and polarization in the U.S. Congress. Second, vote counts are easy to measure: they are either yes, no, or present. This paper shows that speech-based measures offer a different descriptive picture of how party-leader agreement and polarization have changed over the last 120 years. We find that there are lower levels of party leader agreement and more historical instances of polarization and de-polarization when using speech-based measures instead of roll-call votes-based measures.

We find speech-based measures provide more plausible qualitative descriptions of polarization and party-leader agreement than roll-call votes. Congressional votes are equilibrium outcomes, in that party leaders only bring votes to the floor where they know they will win the vote. It is generally bad legislative politics to call votes that divide a leader's party. Conversely, members of Congress are generally at liberty to speak their minds in their floor speeches – often to the chagrin of party leaders. This bifurcation in speech between party leaders and their co-partisans is especially noticeable in the modern Congress, where Republican and Democratic Speakers of the House during the 2010s and 2020s had infamously little control over their co-partisans' message discipline. We ultimately find that polarization has fluctuated far more frequently over the last century, and party-leader agreement is much lower than roll-call votes-based measures suggest.

2 Topic Models are Statistically Inconsistent

First, we establish that topic models do not abide by statistical consistency when estimated via variational inference. By statistical consistency, we mean that the parameters

of the topic models converge in probability to the actual population value as more documents are collected to augment the dataset. After establishing statistical inconsistency, we will show that topic models diverge from the truth as more documents are collected, confounding political scientists' notions of how data collection should improve our statistical estimates. Now, we describe the assumed data generation process that defines topic models, where the underlying parameters that describe this model are assumed to have ground-truth values. Then, we show formally, that topic models will never approach the truth. We then show that, in fact, as more data is collected, topic models grow less likely to converge to the truth.

The underlying assumption of statistical consistency is that a ground truth value exists for the parameters we wish to estimate. In Table 1, we summarize the notation used throughout the paper, which we adopt from Kangaslahti et al. (2023) for notational consistency and transparency.

2.1 Data Generation Process for Topic Models

We offer here a brief overview of the topic model framework as popularized in Blei, Ng, and Jordan (2003b). The goal of this model is to estimate underlying topics in text data. Unsupervised methods, such as topic modeling, are instrumental in political science as much of our text data is large-scale, and precise data generation processes for text are often undertheorized. Consider the congressional speeches we study in this paper. We might hypothesize that party is a crucial covariate for predicting speech; however, during the Civil Rights era in the 1950s and 1960s, learning which party affiliation of the interlocutor would reveal surprisingly little insight about speech in Congress. Southern and Northern Democrats often sounded little alike on civil rights, and the Republicans were often criticizing the New Deal. A more unstructured approach is helpful here.

The most popular topic model has an assumed Latent Dirichlet Allocation (LDA) structure (Structural Topic Models (STM), which are popular in political science and are similar to the underlying topic model. However, this method incorporates covariates (Roberts et al., 2014). We will study Blei's more popular method, but many results extend to other topic models estimated via variational inference (VI), of which STM is a

particular case. The model setup for LDA is simple:

We have a corpus of N documents comprised of some combinations of V total number of words in the vocabulary, all possible words that could appear in a document. These documents will contain W words each (which could include up to W duplicates of the same word). We capture the words contained by each document in the vector $\mathbf{f_i}$. The researcher determines that there are K hidden topics in the collection of documents. We then have the model,

$$\mathbb{E}\big[\, f_i|\mathbf{h}\,\big] = \boldsymbol{\mu}\mathbf{h}$$

where $\mathbf{h}$ is vector of multinomial distributions which describes the probability of seeing a topic given a document. Then $\boldsymbol{\mu}$ is a vector of multinomial distributions of the probability of seeing a word given a topic. Finally, $\mathbf{z}$ are topic labels to be uncovered by the LDA method.

In this setup, we have the following variables drawn in the following way:

$$\mathbf{h} \sim \text{Dirichlet}(\alpha)$$

$$\boldsymbol{\mu} \sim \text{Dirichlet}(\beta)$$

$$\mathbf{z} \sim \text{Multinomial}(\mathbf{h})$$

The hyperparameters α and β describe the amount of mixing in the documents. When values are closer to 0, the model collapses to one where each document belongs to a single topic. When values approach infinity, the documents become entirely mixed among all topics. The appropriate choice for mixing parameters will depend on the specific domain of the text data on which the topic model is estimated – social media posts on Twitter are likely to be single-topic documents so that researchers may favor a smaller mixing parameter. In contrast, Congressional speeches will generally comprise more than one topic, necessitating increasing the mixing parameter.

Given this setup, we write the LDA posterior distribution as

$$p(\boldsymbol{\mu}, \mathbf{h}, \mathbf{z}|\mathbf{f_i}, \alpha, \beta) = \frac{p(\boldsymbol{\mu}, \mathbf{h}, \mathbf{f_i}, \mathbf{z}|\alpha, \beta)}{p(\mathbf{f_i}|\alpha, \beta)} \tag{1}$$

As is common with these types of high-dimensional models, the normalizing constant $p(\mathbf{f_i}|\alpha, \beta)$ is intractable (Blei, Ng, and Jordan, 2003b):

$$p(\mathbf{f_i}|\alpha, \beta) = \frac{\Gamma(\alpha_k)}{\prod_k \Gamma(\alpha_k)} \int \left(\prod_K \mu_i^{\alpha_i - 1} \right) \prod_{k=1}^{K} \sum_{i=1}^{V} \prod_{i=1}^{N} (\mu_k h_{k,i})^{f_i^n} d\mu \tag{2}$$

where the identification problem lies in the term $\mu_k h_{k,i}$, which cannot be tractably estimated, and so an approximation technique is used (Blei, Ng, and Jordan, 2003b). This approximation technique is called variational inference, which we now explain intuitively.

The inconsistency arises due to underlying problems in the estimation routine for topic models. Because topic models are computationally taxing and the objective functions contain intractable constants, researchers rely on an approximation method called Variational Inference to estimate the topics (Blei, Ng, and Jordan, 2003b; Grimmer, 2011). This method is convenient because the intractable denominator, Equation 1, cancels out. The goal is to reduce the distance between the actual and approximated numerators. Unfortunately, this approximation introduces instabilities of its own.

2.2 Evidence of statistical inconsistency

To understand this inconsistency intuitively, imagine sorting all the words in a text dataset into topics. Researchers would try to find words that occur together in recognizable patterns. Then, our researchers came across two words that could be categorized among a diverse array of topics: say, apple and blackberry. Should these words go to the “phone" topic, the “fruit" topic, or even the “pie" topic? A good topic model should be able to determine which topic is appropriate based on the data with as little help from the researcher as possible. With a small dataset, this seems an easy enough task. We can read all the text, figure out the context of our dataset, create plausible heuristics for categories, and hire some research assistants to label the data. However, as researchers collect more and

more data, they can no longer read all the documents. Due to constraints on time and attention, they necessarily miss the context of the text they wish to analyze. The underlying data grow increasingly higher-dimensional, so discerning patterns become increasingly taxing.

Remarkably, our best topic models suffer the same problem as our human researchers! They cannot distinguish between "phone," "fruit," or "pie." Yet good topic models are supposed to provide good descriptions of patterns of text in the data that are too large, too complex, and too expensive to label by hand. The problem is that popular methods simply cannot determine where these words belong. Too many topics are mathematically plausible.

For more mathematical intuition, imagine taking one column of the word-topic matrix $\boldsymbol{\mu}$, duplicating it, and dividing both the old and new columns by $\frac{1}{2}$. When we multiply it with document-topic matrix $\mathbf{h}$, we have the same result for the document-word matrix $\boldsymbol{\mu}\mathbf{h}$ as before.[4]

The reason this occurs is due to how topic models are estimated. The estimation technique involves minimizing the distance between the log-likelihood of the true posterior and the log-likelihood of a tractable, variational distribution. This approximating, variational distribution breaks the dependence between μ and h:

$$q(\mu, \mathbf{h}, \mathbf{z}|\gamma, \phi) = q(\mathbf{z}) \prod_{i=1}^{N} q(\boldsymbol{\mu}|\phi_n) q(\mathbf{h}|\gamma)$$

We then minimize the distance between the “true" and approximating distribution. We measure this distance using the Kullbach-Liebler Divergence:

$$D_{\text{KL}}(p \parallel q) = \sum_{x \in \mathcal{X}} p(\boldsymbol{\mu}, \mathbf{h}, \mathbf{f_i}, \mathbf{z}|\alpha, \beta) \ \log \left(\frac{p(\boldsymbol{\mu}, \mathbf{h}, \mathbf{f_i}, \mathbf{z}|\alpha, \beta)}{q(\mu, \mathbf{h}, \mathbf{z}|\gamma, \phi)} \right) \tag{3}$$

where ϕ and γ are called variational parameters. Said in more technical terms, because the normalizing constant in Equation 2.1 is intractable, practitioners have utilized variational inference as a convenient alternative to estimating Latent Dirichlet Allocation (LDA) Blei, Ng, and Jordan (2003b). In the Appendix, we show under general conditions, we show

[4]We note that Nakajima et al., 2014 were the first to suggest this intuition

the KL-Divergence in Equation 3 will not converge, and is thus inconsistent. We illustrate the underlying logic for this theoretical result in the next section.

3 Why Inconsistency and Bias Increases as More Data are Collected

We now show that even in the best case, Topic Models are not merely inconsistent – that they do not converge to the true population value – they in fact grow less and less likely to produce the correct results as we collect more and more data. Even worse, we show that the estimates grow more biased in a simulated setting as we sample more and more data. In fact, topic model estimates have less absolute error when we sample 50 documents from a 10,000 document corpus than when we estimate the model on the entire set of simulated 10,000 documents. These two properties of topic models, increasing rates of noncovergence and bias, produces substantial biases in the reported results. This result has staggering implications for political scientists in the empirical work on text data. Political scientists hope that, in the best case, they will better approximate the truth in their estimates as they endeavor to collect more data. However, with this popular method, used by tens of thousands of researchers, the more effort researchers put into data collection, the more biased the answer produced by the model. Researchers toil to collect these large datasets: digitizing historical archives, grappling with corporate APIs, and negotiating with lawyers to find terms to use social media data. Nevertheless, we find their reward for these efforts is increasingly biased estimates.

The intuition for why we have this result is simple: as we collect incrementally more data, there are an increasing number of opportunities for topics to be incorrectly combined or split. Like our researcher trying to categorize words in the last section, our model has more opportunities to encode mistakes as we collect more documents to supplement our dataset. In Figure 1, we report the theoretical best convergence rates under the lowest possible error rates. We mark seven papers on the graph from top political science journals that use variational inference to estimate their topic model outputs. Notice how usual intuitions about laws-of-large numbers fail here. As datasets grow large, convergence

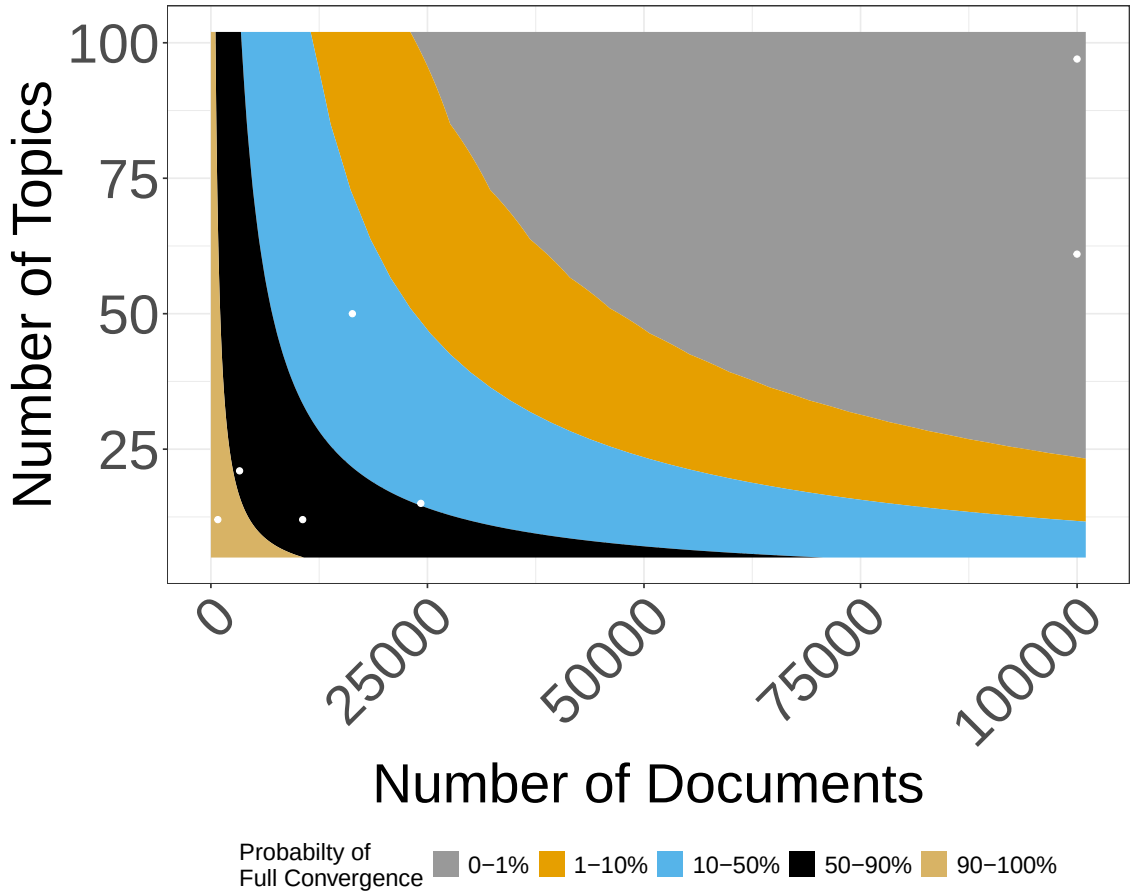


Figure 1: Best Case Convergence: We plot seven papers using variational Inference from APSR, JOP, AJPS, PNAS, and PA, as well as Blei, Ng, and Jordan (2003b)'s original application. Note that the two points at the 100,000 mark actually have millions of documents and would be orders of magnitude away from the other documents, so we truncate the graph for the ease of illustration. The probabilities of convergence for these results are actually far lower than depicted.

begins to fail. The reason is that as the dataset has more and more documents, the number of opportunities for topical parameters to "switch" also grows. At some point, by adding more documents, the researcher guarantees the model will make a mistake at least once, leading to asymptotic non-convergence and an inconsistent and unbiased result.

Figure 1 shows the theoretical best convergence rates for topic models estimated under variational inference. We calculate the probability that the entire corpus converges to derive these convergence rates. We assume the most conservative case: there is only one duplicated column, and then we divide that column (and its duplicate) by exactly one-half. The resultant matrix multiplication results in the same document-term matrix,

but the underlying probabilities are incorrect. One could imagine far more catastrophic scenarios (perhaps all the columns are divided by K and add a column multiplied by $K * N$). In that case, rates of non-convergence would be far worse, approaching 100 percent even for datasets with only 1 document.

Nevertheless, as Figure 1 makes clear, convergence rates are poor even under the best-case scenario we illustrate here. We collected the five most cited and recent papers from the APSR, the JOP, PA, and the AJPS using Blei's method. We note these papers with white dots on the plot. Notice again that papers that are most likely to converge are those with the fewest number of documents. When one might find topic models are most needed, they perform worse: when data are prohibitively expensive to hand label.

Nevertheless, it is not necessarily resource-intensive to label the small corpora where such unstructured data when hand-labeling the text is so eminently feasible. It has been shown that the best practice for topic models is most helpful for data discovery and exploration on datasets that are large, expensive to label, and where the underlying data generation process is undertheorized (Grimmer and King, 2011; Grimmer, Roberts, and Stewart, 2022). Unsupervised methods, such as topic models, are most valuable for discovery in prohibitively large datasets. So, notably, the most popular methods fail precisely in the cases we most need them: when we have extensive data. Alternatively, in medium-size data, the rates of convergence are less encouraging. At best, researchers can expect a 50 percent chance that their results are consistent and unbiased under the most favorable conditions with as few as 25,000 documents and 12 topics. A coin flip's chance of getting consistent and unbiased answers - in the best case - is a risky bet in research!

We further investigate the disappointing behavior of topic models by demonstrating that as we collect more documents from a population, that the mean absolute error increases. across ten topics when using an LDA model with a corpus of 10,000 documents. We make several assumption for the best case scenario. First we assume that the number of topics is known to the researcher. We assume no measurement error in collecting the text data, the vocabulary is fully known, and that there is finite number of documents – 10,000 – in the full population. Even under these extraordinarily generous conditions,

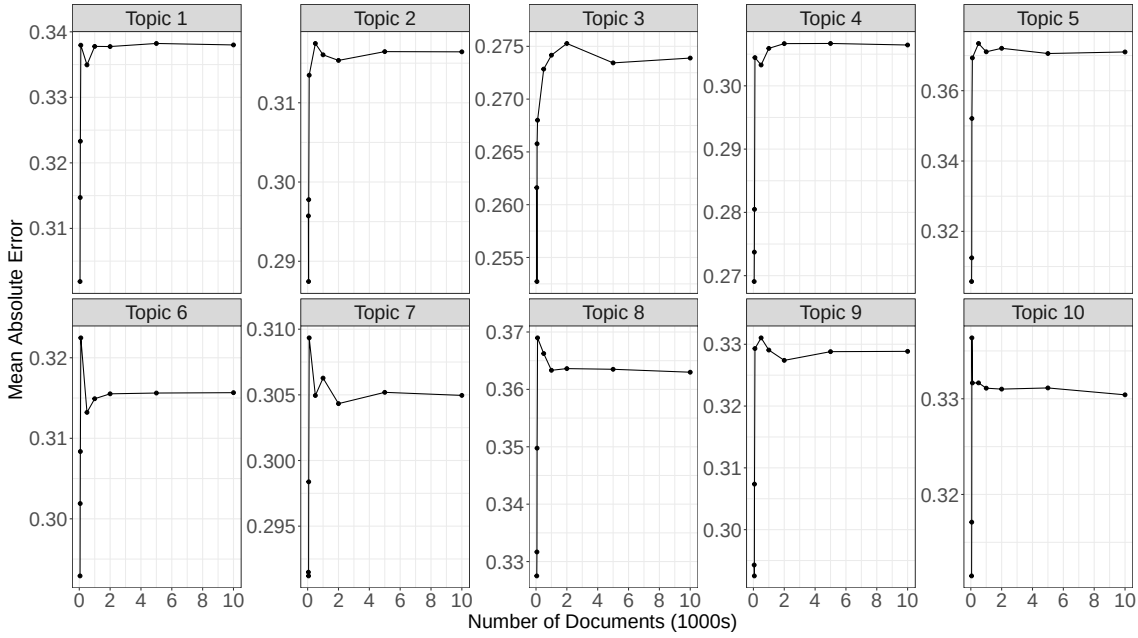


Figure 2: Mean Absolute Error: We simulated 10,000 documents drawn from the LDA data generation process described in the previous section. We plot the mean absolute error for the top 15 words for each of the ten topics. We simulate LDA 100 times for 50, 75,..., 5000 documents. We then acquire estimates on the full population of 10,000 documents. We compare posterior word probabilities for the top words in each topic to their ground truth values and report the mean absolute error across all documents in the sample. We report the average of this value across all 100 simulations.

we show that more data does not reduce bias, which we measure with mean anbsolute error. As detailed in Figure 2, the results from our simulations reveal a troubling trend: as we increase our sample size, the mean absolute error for identifying each topic increases precipitously as we collect more documents, eventually levelling off as we include the entire populaiton of documents. Specifically, we simulate LDA estimates 100 times, progressively sampling up to the complete set of 10,000 documents. The data indicates that, contrary to traditional assumptions about the benefits of larger datasets, the bias in mean squared error worsens with the addition of documents.

This counterintuitive phenomenon can be attributed to the number of new parameters introduced by each additional document. With every new entry, the likelihood of topic parameters being inaccurately estimated or misclassified grows, leading to a compounding effect of errors. The implications of this finding are significant for the field of political science: researchers, pursuing the ideal of more substantial datasets to refine their anal-

yses, may unknowingly exacerbate bias and reduce the reliability of their findings. As we can see, efforts to enhance data collection may ultimately yield diminishing returns, reinforcing the critical need for alternative methodologies that prioritize consistent and unbiased estimations over sheer volume. In this context, we emphasize the urgency for a re-evaluation of the methodologies currently in use, as clinging to the belief that more data equates to better models proves dangerously misleading.

All told, Figure 2 presents a clear warning: relying on topic models in their current form can lead to increasingly biased results as researchers collect more documents for ther datasets. The challenges posed by extensive datasets render traditional assumptions about convergence and accuracy obsolete, highlighting the pressing need for a paradigm shift in how political scientists approach text analysis methodologies.

4 A Correct Topic Model Method that is Accurate for Large Data

To demonstrate the bias with an example, we employ a topic model with established accuracy guarantees (Anandkumar, Foster, et al., 2013). Our preferred method builds on this theoretical foundation and sidesteps the challenges associated with variational inference by eliminating the need for approximation. Notably, by implementing a batching approach, the method also scales effectively to large datasets (Kangaslahti et al., 2023), unlike many accurate methods that may take months or years to yield results – a significant obstacle for data-driven researchers. This computational intractability is particularly problematic because the practicality of accurate methods diminishes when faced with vast data. Our approach leverages the finding that topic models can be framed as a method of moments problem, allowing for estimating population moments through straightforward algebraic transformations of the principal components of the word frequency, word co-occurrence, and word tri-occurrence matrices (Anandkumar, Foster, et al., 2013). By avoiding approximations inherent in variational inference, we mitigate the identification issues discussed in earlier sections that complicate approximation methods.

5 Illustration of the Bias through an Application to Congressional Speeches

By using a consistent and unbiased approach, we ultimately find that speech-based measures show party-leader agreement has *weakened* since 1891. In contrast, biased measures show an increase and speech-based measures of polarization have implied several instances of de-polarization over the last 120 years, whereas biased measures suggest a decline.

5.1 Congressional Floor Speech Data and Pre-Processing

We study partisan alignment using $13,818,250$ congressional floor speeches collected and collated in (Gentzkow, Shapiro, and Taddy, 2018). We follow standard practices for pre-processing the text, removing common stop words, stemming the data, and allowing bigrams. After processing, we fit the model using all $13,818,250$ speeches. We then focus our analysis on the $4,388,931$ speeches from U.S. House of Representatives members. We now show that consistent and unbiased topic modeling methods correct for the substantive bias introduced by the statistical divergence of the existing topic modeling methods. We pick a technique with theoretical convergence guarantees (Anandkumar, Ge, et al., 2014). The underlying assumption here is mild – we need a topic-word matrix that is not co-linear. In Kangaslahti et al. (2023), we, alongside our co-authors, implement a scalable version of this model, which we apply to uncover the topics consistently and unbiasedly. We briefly summarize this method in the appendix. We note that there are other plausible methods for topic modeling with similar guarantees for large datasets or through Gibbs Sampling for smaller datasets (Breuer, 2024; Eshima, Imai, and Sasaki, 2024). Our method is open source and freely available through the Tensorly suite of ML algorithms[5].

We compare against the results derived from Gensim's LDAMulticore method for data at scale, which uses variational inference as popularized by Blei, Ng, and Jordan, 2003a. We optimize the number of topics based on an array of calculated coherence (Röder, Both,

[5]The software and accompanying webpage can be found here: https://tensorly.org/tlda/

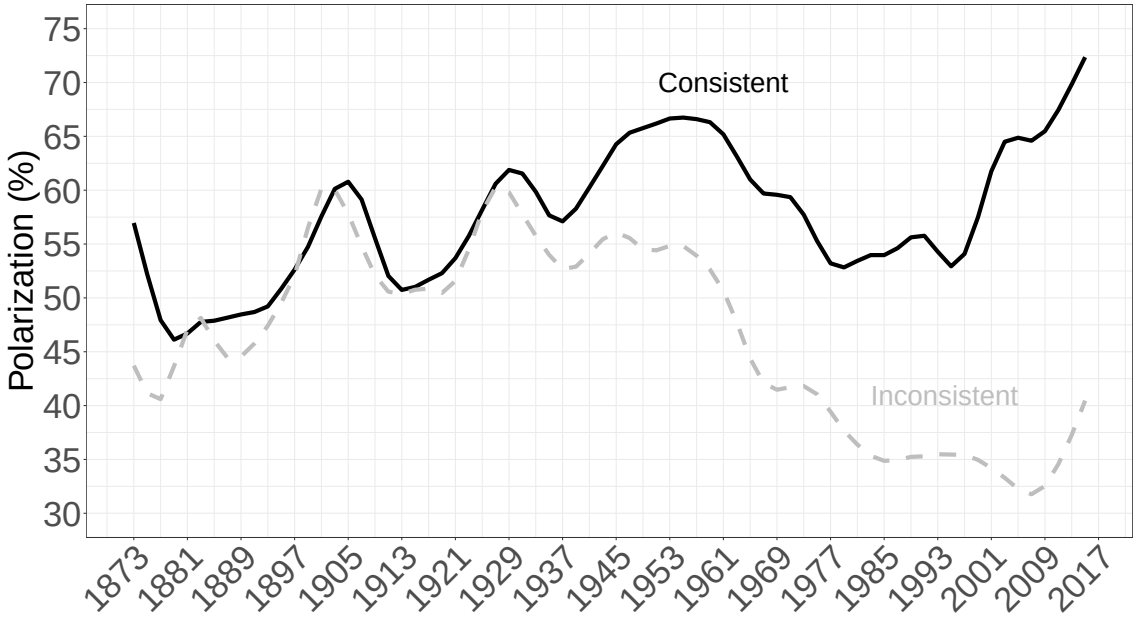


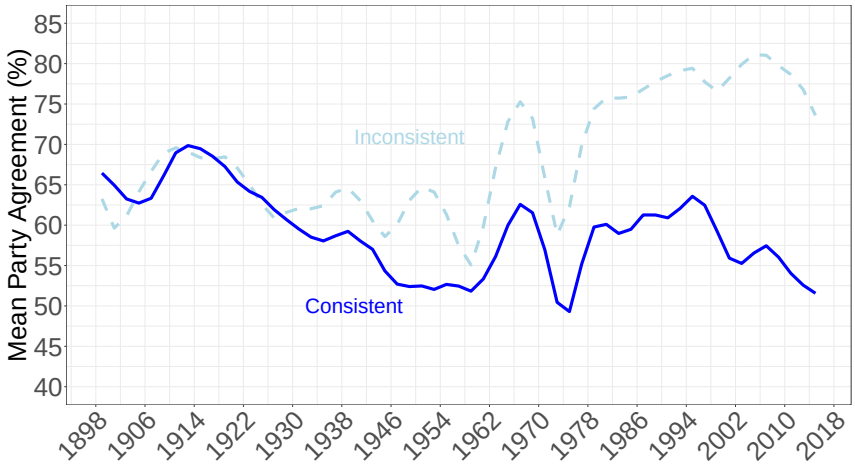
Figure 3: Speech Polarization

and Hinneburg, 2015). We find the optimal number of topics based on the availability of these coherence metrics, which is 30 for the TLDA method. We also optimized the LDAmulticore method and found that the same number of topics was optimal.

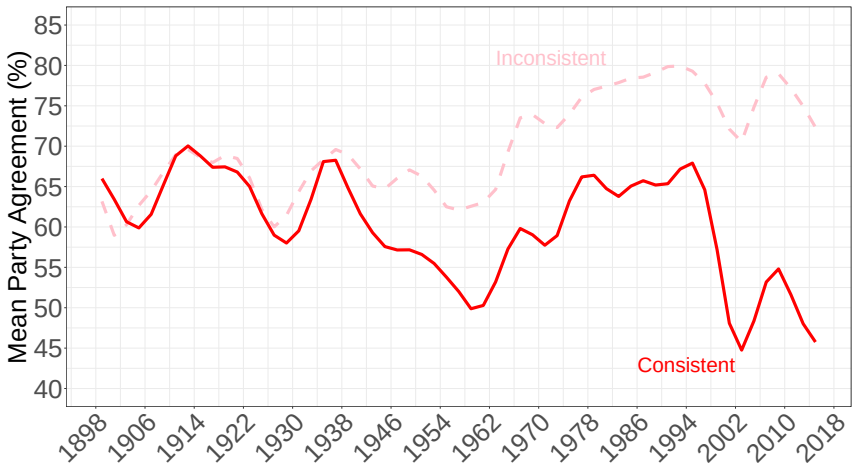
With 30 topics and 14 million documents, we note that our dataset is in the range where variational inference methods will have a less than 1-in-100,000 chance of consistent and unbiasedly converging to the actual underlying topical values.

5.2 Corrected Topic Models do not Underestimate a Century of Polarization or Overestimate a Century Party-Leader Agreement

In Figure 3, we show polarization as measured by the average inter-party pairwise correlations in topics discussed on the floor each year. In the solid line, we show polarization when measured with a consistent and unbiased topic model method. In contrast, the dashed grey line shows implied polarization using the popular Blei method. The first notable difference is the dramatic divergence after 1930. Second, the inconsistent and unbiased method implied polarization has declined over time. Our consistent and unbiased topic model methods imply very different trends in polarization compared with the biased estimates.



(a) Democratic Party-Leader Alignment



(b) Republican Party-Leader Alignment

Figure 4: Speech-Based Measures of Party-Leader Agreement

We now show here that consistent and unbiased topic model methods give speech-based evidence of very different patterns of party-leader agreement than popular topic models would imply. In Figures 4a and 4b, we show the average pairwise correlation between party leaders and party members within each party based on the topics they discuss in a given Congress. We show the implied level of party-leader agreement using a consistent and unbiased method in a solid line and then the same measure using the Blei-based estimates in a dashed line. Using a consistent and unbiased method, we find that party-leader agreement has decreased for both parties since the dawn of the modern leadership system in the late 1800s. In this figure, we find precisely the opposite pattern using the Blei estimator. The decline is especially acute for Republicans after 1994.

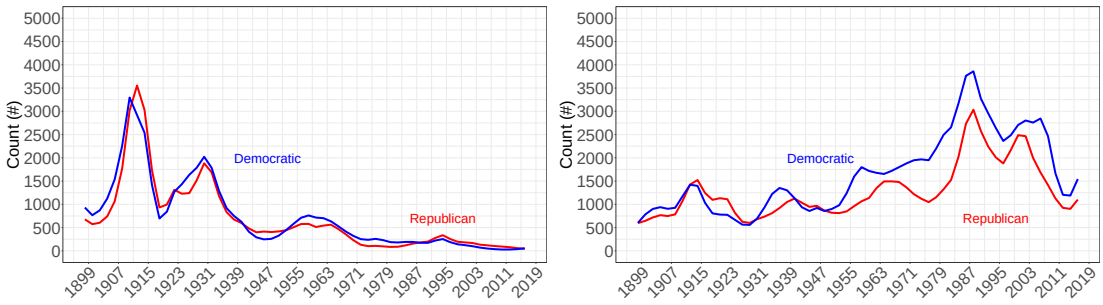
5.3 Substantive Bias Arises due to Changing Contexts

We rely on the method to provide a consistent and unbiased description of how text is related and organized into topics. We have formally and empirically demonstrated that inconsistent and unbiased topic model methods lead to substantive bias. Here, we illustrate how the bias arises with a specific example of topic identification, in this case, trade. With an unsupervised, consistent, and unbiased method, we cannot know *a priori* the topics in the text, nor which words belong in which topic. This bias is especially prevalent for extensive datasets that span hundreds of years and cover changing contexts such as those we have studied here. To fully illustrate the point, we show how word counts evolve for two potentially closely related words: tariff and trade. We show in Figure 5a how tariff was more frequently mentioned in the 1800s, whereas in Figure 5b, the word “trade" was more common after 1980. A consistent and unbiased topic model should realize these words will likely belong to the same topic despite the changing frequencies over time.

We can illustrate the source of the bias in inconsistent and unbiased methods by imagining five potential scenarios for a correct topic solution: (i) a trade topic that includes only the word “trade", (ii) one that includes only the word “tariff," (iii)one that includes both words, (iv)one that includes neither, or (v) the model uncovers no trade topic. Blei’s method settles on (ii), whereas a consistent and unbiased method resolves that (iii) is the correct topical description. Blei’s method decides that two trade topics exist, both dominated by the word "tariff," but neither of which features the word “trade." Given the long duration of the data under analysis, the changing contexts over time, and our domain-specific knowledge that trade has been a prominent topic of Congressional debate, these results are unsurprising.

6 Consistent Topic Models Provide More Plausible Estimates of Quantities of Interest

We now show that our measures offer a new description of the trends of two popular and critical trends in quantities of interest for political scientists. To study these quantities, political scientists have relied on roll-call-votes-based measures which is favored because



(a) Mentions of Tariff by Party

(b) Mentions of Trade by Party

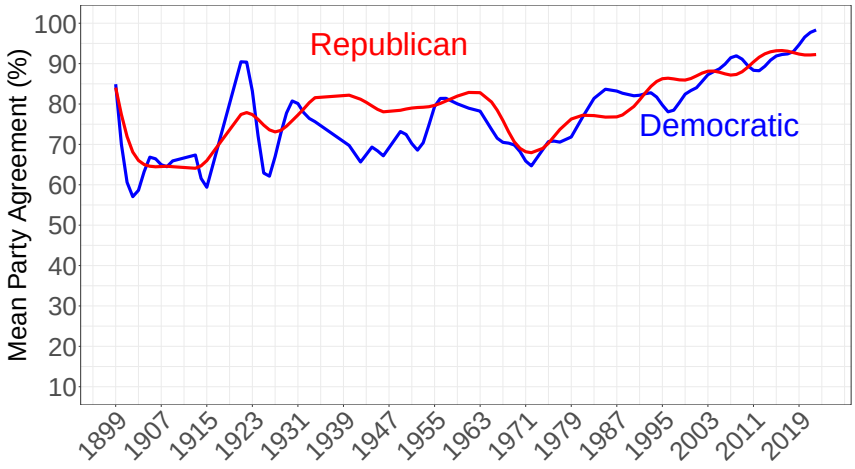
Figure 5: Trade-Related Word Frequencies over Time

they have long historical record. We will show that speech-based measures provide better estimates of these roll-call votes-derived quantities and address the fundamental shortcomings of these measures. We hope this renews some of the original promise that text data could provide more reliable descriptions of quantities of interest that political scientists endeavor to study.

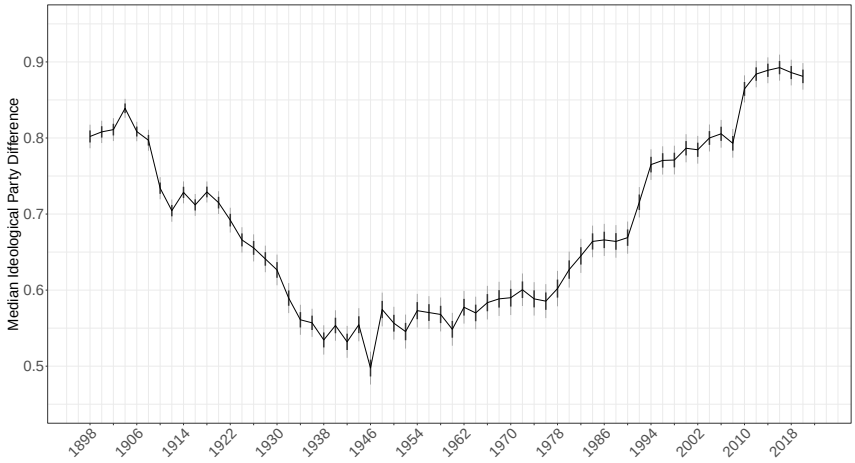
6.1 Existing Votes-Based Measures Show Implausibly Low Levels of Polarization and High Levels Party-Leader Agreement

First, we show and describe trends in vote-based measures favored by the literature. Figure 6a shows party-leader agreement from 1899 to 2015. We measure this by calculating the percentage of bills each year where party members voted in the same direction as their party leaders. The agreement levels are generally high, ranging from 60 to 95 percent in a given year. There are two notable periods where agreement is low: prior to the Progressive Era, when Republicans and Democrats contended with their respective progressive movements within their parties, and again during the Civil Rights era when Southern and Northern Democrats were at odds over ending segregation. Figure 6b shows the difference between the median Democrat and median Republican in the DW-Nominate scale. These measures show that ideological polarization was high in the 1800s and the 2000s but low for the rest of the period.

By using a consistent and unbiased method, we demonstrate that neither of these approaches fully captures the relationship between party leadership and their co-partisans. Roll-call votes are equilibrium outcomes. The unsurprising outcome of this strategic logic



(a) Party-Leader Agreement on Votes



(b) Ideological Polarization (DW-Nominate)

Figure 6: Votes-Based Measures

is that party leaders do their best to avoid bringing votes to the floor that divide their party. It is usually bad legislative politics to call votes that large numbers in their party oppose.[6]

Compared with the vote-based measure of party-leader agreement trends in Figure 6a, we observe starkly different trends in Figures 4a and 4b, where we plot a measure of party-leader agreement. Using a speech-based approach, we find that members of Congress are much less likely to agree with their leaders in speeches than on votes. This relatively lower level of party-leader agreement makes qualitative sense as members of Congress have much more latitude in speaking their minds on their particular topics of interest on

[6]There are exceptions, of course, where, especially in the modern Congress, GOP Speakers will bring must-pass legislation to the floor where a majority of the majority party opposes the legislation. Bringing divisive votes to the floor tends to result in the speedy end of the Speaker's leadership role.

the U.S. floor. Party leaders have fewer institutional veto points over speech, unlike floor votes, where party leaders can bring bills to the floor or let these bills die before being brought to the floor.

The same underlying issues lead to complications with studying polarization using roll-call votes, like with DW-Nominate. Although measures like using campaign finance (Bonica, 2014) or Twitter activity (Barberá, 2015) have been used to study polarization, these approaches do not have a historical record of congressional floor speeches.

Importantly, speech-based approaches can give historical insights over 120 years and provide more consistent and unbiased descriptions. Unlike roll-call votes-based estimates, we see many periods of re-polarization and de-polarization over time, as opposed to a U-shaped trend in Figure 6b.

7 Known Conceptual and Theoretical Issues

Given the widespread use of topic models, applied researchers might assume that the variational inference method such models rely upon has reliable statistical guarantees.

Existing literature has proven theorems establishing statistical consistency for topic models estimated via variational inference but under conditions unlikely to be encountered by applied researchers. While theoretically elegant, these results apply only under the most unrealistic of conditions. For example, we achieve consistency as the starting values being known *a priori* (Wang and Titterington, 2012), assuming a sparse posterior parameter space (Pati, Bhattacharya, and Yang, 2017), or by introducing additional hyperparameters to dampen the likelihood (Yang, Pati, and Bhattacharya, 2018). Still, others redefine consistency in ways that are not encountered in practice, such as allowing documents to grow to infinite length, all else fixed (Nakajima et al., 2014).

It is well-noted that variational inference methods for high-dimensional topic models lack accuracy and consistency guarantees (Anandkumar, Foster, et al., 2013) and suffer from various instabilities for posterior inference (Ghorbani, Javadi, and Montanari, 2019),. Statistics and computer science intuition suggest that desirable large-sample properties like Laws of Large Numbers and Central Limit Theorems come with ever-larger

data sets. In fact, topic models routinely use variational inference approximations with undesirable analytical properties and no such large-sample guarantees. Worse yet, this is true precisely in high-dimensional settings where political scientists are mainly likely to employ machine learning tools. At best, Nakajima et al. (2014) shows that consistency only holds if we fix the number of documents and the total number of unique words in the corpus, allowing the individual documents to grow asymptotically in document length. Allowing the number of documents and vocabulary size to grow in fixed ratios asymptotically breaks consistency. None of these theoretical approaches have large-sample consistency in realistic settings. The core of the problem arises from the intractable interdependencies first noted by Blei, Ng, and Jordan (2003b) and reiterated by others empirically (Ghorbani, Javadi, and Montanari, 2018; Yao et al., 2018).

8 Conclusion

Our results have several important implications. Our statistical results highlight the limitations of the existing popular methods for topic modeling. First, we show that popular topic model methods must be more convergent. Second, they are increasingly less likely to converge to the actual value as we collect incrementally more data. By using an example of floor speeches in the U.S. House of Representatives, we have illustrated that this non-convergence produces substantively misleading results, especially for large datasets. We show that partisan alignment between members and leaders has decreased over a century, whereas existing topic model methods would suggest the opposite. To address and correct the substantive biases introduced by inconsistent topic model estimators, we use our correct topic model method with theoretical accuracy guarantees at scale, implemented in a convenient Python software Kangaslahti et al., 2023.

References

Anandkumar, Animashree, Dean P. Foster, Daniel Hsu, Sham M. Kakade, and Yi-Kai Liu (2013). "A Spectral Algorithm for Latent Dirichlet Allocation". In: *arXiv preprint arxiv: 1204.6703*.

Anandkumar, Animashree, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky (2014). "Tensor Decompositions for Learning Latent Variable Models". In: *Journal of Machine Learning Research* 15.80, pp. 2773–2832. URL: http://jmlr.org/papers/v15/anandkumar14b.html.

Barberá, Pablo (2015). "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data". In: *Political Analysis* 23.1, pp. 76–91.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003a). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

— (Mar. 2003b). "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3.null, pp. 993–1022. ISSN: 1532-4435.

Bonica, Adam (2014). "Mapping the Ideological Marketplace". In: *American Journal of Political Science* 58 (2), pp. 367–386.

Breuer, Adam (2024). "Interpretable LDA Topic Models with Near-Optimal Posterior Probability". In: *Working Paper*.

Eshima, Shusei, Kosuke Imai, and Tomoya Sasaki (2024). "Keyword-Assisted Topic Models". In: *American Journal of Political Science* 68.2, pp. 730–750. DOI: https://doi.org/10.1111/ajps.12779. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12779. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12779.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy (2018). *Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts*. https://data.stanford.edu/congress_text. Accessed: 2023-07-16. Palo Alto, CA.

Ghorbani, Behrooz, Hamid Javadi, and Andrea Montanari (2018). *An Instability in Variational Inference for Topic Models*. arXiv: 1802.00568 [stat.ML].

— (June 2019). "An Instability in Variational Inference for Topic Models". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 2221–2231. URL: https://proceedings.mlr.press/v97/ghorbani19a.html.

Grimmer, Justin (2011). "An Introduction to Bayesian Inference via Variational Approximations". In: *Political Analysis* 19.1, pp. 32–47. DOI: 10.1093/pan/mpq027.

Grimmer, Justin and Gary King (2011). "General purpose computer-assisted clustering and conceptualization". In: *Proceedings of the National Academy of Sciences* 108.7, pp. 2643–2650. DOI: 10.1073/pnas.1018067108. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1018067108. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1018067108.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Grimmer, Justin and Brandon M. Stewart (2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts". In: *Political Analysis* 21.3, pp. 267–297. DOI: 10.1093/pan/mps028.

Kangaslahti, Sara, Daniel Ebanks, Jean Kossaifi, R. Michael Alvarez, and Anima Anandkumar (2023). "TensorLy-LDA: Analyzing Social Media Conversations at Scale with Online Tensor LDA". In: *Working Paper*.

Nakajima, Shinichi, Issei Sato, Masashi Sugiyama, Kazuho Watanabe, and Hiroko Kobayashi (2014). "Analysis of Variational Bayesian Latent Dirichlet Allocation: Weaker Sparsity Than MAP". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger. Vol. 27. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/5487315b1286f907165907aa8fc96619-Paper.pdf.

Pati, Debdeep, Anirban Bhattacharya, and Yun Yang (2017). *On Statistical Optimality of Variational Bayes*. arXiv: 1712.08983 [math.ST].

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand (2014). "Structural Topic Models for Open-Ended Survey Responses". In: *American Journal of Political Science* 58.4, pp. 1064–1082. DOI: 10.1111/ajps.12103. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12103. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12103.

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015). "Exploring the Space of Topic Coherence Measures". In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM '15. Shanghai, China: Association for Computing Machinery, pp. 399–408. ISBN: 9781450333177. DOI: 10.1145/2684822.2685324. URL: https://doi.org/10.1145/2684822.2685324.

Wang, Bo and D. Titterington (2012). *Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values*. arXiv: 1207.4159 [math.ST].

Yang, Yun, Debdeep Pati, and Anirban Bhattacharya (2018). *α-Variational Inference with Statistical Guarantees*. arXiv: 1710.03266 [math.ST].

Yao, Yuling, Aki Vehtari, Daniel Simpson, and Andrew Gelman (Oct. 2018). "Yes, but Did It Work?: Evaluating Variational Inference". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 5581–5590. URL: https://proceedings.mlr.press/v80/yao18a.html.

A Notation

In this section, we summarize all the paper's notation for the reader's reference.

B Variational Inference and KL Divergence

We can expand equation 3 for $D_{\mathrm{KL}}(p \parallel q)$. We can then study this expanded term to understand the asymptotic properties of LDA as we collect additional documents.

By optimizing the KL-Divergence, we can find stationary points for ϕ and γ, writing them in terms of the critical parameters of interest: $\mathbf{h}$ and $\boldsymbol{\mu}$

Table 1: Table of Notations used in this paper.

Symbol	Meaning	Domain
K	Number of topics	$\mathbb{N}$
$\mathbf{h}$	Topic mixture	$\mathbb{R}^K$
$\mathbf{z}$	Topic label	$\mathbb{R}^K$
V	Vocabulary size	$\mathbb{N}$
W	Document size	$\mathbb{N}$
$\boldsymbol{\mu}$	$\mathbb{E}\left[f_i \middle\| \mathbf{h} \right] = \boldsymbol{\mu}\mathbf{h}$	$\mathbb{R}^V$
$\mathbf{f}_i$	Frequency vector for the i-th document	$\mathbb{R}^V$
$\tilde{\mathbf{x}}_i$	Centered frequency vector for the i-th document	$\mathbb{R}^V$
$\mathbf{x}_i$	Centered & whitened frequency vector	$\mathbb{R}^V$
N	Number of documents	$\mathbb{N}$
D	Whitening dimension size	$\mathbb{N}$
n_b	Number of documents in a mini-batch	$\mathbb{N}$
$\mathbf{X}$	centered, whitened matrix with columns $\mathbf{x}_i$	$\mathbb{R}^{n_b \times D}$
$\boldsymbol{\Phi}$	learned factors of the decomposition	$\mathbb{R}^{D \times K}$

$$\mathbf{z} = \frac{\exp\left(\Psi(\gamma_{i,k}) + \sum_{v=1}^{V} f_{v,i}\left(\Psi(\phi_{v,k}) - \Psi\left(\sum_{k'=1}^{K}\phi_{v',k}\right)\right)\right)}{\sum_{k'=1}^{K}\exp\left(\Psi(\gamma_{i,k'}) + \sum_{v=1}^{V} f_{v,i}\left(\Psi(\phi_{v,k'}) - \Psi\left(\sum_{v'=1}^{V}\phi_{v',k'}\right)\right)\right)} \tag{4}$$

$$\boldsymbol{\gamma} = \alpha + \sum_{v=1}^{V}\mathbf{z} \tag{5}$$

$$\boldsymbol{\phi} = \beta + \sum_{i=1}^{N}\sum_{v=1}^{V}\mathbf{f_i}\mathbf{z} \tag{6}$$

We can then expand the KL-Divergence term, following Nakajima et al., 2014.

$$\begin{aligned} D_{\mathrm{KL}}(p \parallel q) &= \sum_{i=1}^{N}\left(\log\frac{\Gamma\left(\sum_{k=1}^{K}\phi_{i,k}\right)}{\prod_{k=1}^{K}\Gamma\left(\phi_{i,k}\right)}\frac{\Gamma(\alpha)^K}{\Gamma(K\alpha)} + \sum_{k=1}^{K}\left(\phi_{i,k}-\alpha\right)\left(\Psi\left(\phi_{i,k}\right) - \Psi\left(\sum_{k'=1}^{K}\phi_{i,k'}\right)\right)\right) \\ &+ \sum_{k=1}^{K}\left(\log\frac{\Gamma\left(\sum_{v=1}^{V}\gamma_{v,k}\right)}{\prod_{v=1}^{V}\Gamma\left(\gamma_{v,k}\right)}\frac{\Gamma(\beta)^V}{\Gamma(V\beta)} + \sum_{v=1}^{V}\left(\gamma_{v,k}-\beta\right)\left(\Psi\left(\gamma_{v,k}\right) - \Psi\left(\sum_{l'=1}^{V}\gamma_{v',k}\right)\right)\right), \\ &- \sum_{i=1}^{N}V\sum_{v=1}^{V}f_{i,n}\log\left(\sum_{k=1}^{K}\frac{\exp\left(\Psi\left(\phi_{i,k}\right)\right)}{\exp\left(\Psi\left(\sum_{k'=1}^{K}\phi_{i,k'}\right)\right)}\frac{\exp\left(\Psi\left(\gamma_{l,k}\right)\right)}{\exp\left(\Psi\left(\sum_{v'=1}^{V}\gamma_{v',k}\right)\right)}\right) \end{aligned} \tag{7}$$

where Ψ is the digamma function. Taking advantage of the limiting properties of the digamma function, we can derive appropriate bounds and then calculate limiting values for KL-Divergence for Latent Dirichlet Allocation models. We summarize the results

from Nakajima et al., 2014 below.

Theorem 1. *From Nakajima et al., 2014: Define the KL-Divergence as above and let* $\boldsymbol{\mu}^*\mathbf{h}^*$ *be the true values of the topic-word probability matrix. Let* I *be the number of entries in* μh *which are not equal to* $\boldsymbol{\mu}^*\mathbf{h}^*$*. Then,*

1. *If* $N, W, V \to \infty$ *with* V *and* N *in a fixed ratio with* W*, then VI diverges with a magnitude* $O_p(W\log(W))$ *and* $o_p(log(W))$ *elements deviate in* I*.*

From this form, we extend Nakajima et al. (2014) to study the large sample properties by relaxing the strong assumptions imposed in their theorems.

We show:

Theorem 2. *Define the KL-Divergence as above and let* $\boldsymbol{\mu}^*\mathbf{h}^*$ *be the true values of the topic-word probability matrix. Let* I *be the number of entries in* μh *which are not equal to* $\boldsymbol{\mu}^*\mathbf{h}^*$*. If* $N \to \infty$ *for fixed V,W, then VI diverges* $O_p(N)$

Proof.

Taking advantage of the limiting properties of the digamma function, we can derive appropriate bounds and then calculate limiting values for KL-Divergence for Latent Dirichlet Allocation models.

We know the digamma function has the following bounds,

$$\begin{aligned}\left(y-\frac{1}{2}\right)\log y - y + \frac{1}{2}\log(2\pi) &\leq \log\Gamma(y) \leq \left(y-\frac{1}{2}\right)\log y - y + \frac{1}{2}\log(2\pi) + \frac{1}{12y} \\ \log y - \frac{1}{y} &\leq \Psi(y) \leq \log y - \frac{1}{2y}\end{aligned}$$

We can use these bounds to find limiting values for each part of $D_{\mathrm{KL}}(p \parallel q)$ and by applying Lemma 2 and Lemma 3 from Nakajima et al., 2014. From here, we get the following form for $D_{\mathrm{KL}}(p \parallel q)$,

$$\begin{aligned}D_{\mathrm{KL}}(p \parallel q) &= \left\{N\left(K\alpha - \frac{1}{2}\right) + K\left(V\beta - \frac{1}{2}\right) - \sum_{k=1}^{K}\left(N^{(k)}\left(\alpha - \frac{1}{2}\right) + V^{(k)}\left(\beta - \frac{1}{2}\right)\right)\right\}\log W \\ &\quad + (K - K^*)\left(V\beta - \frac{1}{2}\right)\log V + O_p(IW + VN)\end{aligned}$$

where $N^{(k)}$ are documents containing the k-th topic and $V^{(k)}$ are words drawn from the k-th topic, K^* are the true number of topics and K are the number of topics selected by the researcher. Then taking the limit with V, W, fixed, the result follows.