Topic Models Estimators Are Statistically Inconsistent and Substantively Misleading: Evidence and Solutions*

Danny Ebanks†

Sara Kangaslahti‡

March 13, 2025

Abstract

We demonstrate that the most commonly used topic model estimators are statistically inconsistent and yield systematically misleading substantive conclusions. Specifically, (1) Latent Dirichlet Allocation models using variational inference diverges from the truth in the large-sample limit and is statistically inconsistent, and (2) this inconsistency is degenerate, which we define as learning noise as if it were meaningful structure. As a consequence, studies that rely on such models produce misleading descriptions of fundamental quantities of interest in American politics, such as legislative polarization and party-leader alignment. We then provide a consistent, spectral alternative that scales to large corpora and is easily applied to a wide variety of social science applications using text. We provide new evidence from 14 million Congressional Floor speeches which reveals that polarization and party-leader alignment, which appear intertwined when measured with roll-call votes, actually diverge substantially over the last 120 years, challenging conventional theories of legislative power and underscoring that modern statistical architectures fundamentally shape our substantive understanding of American politics, renewing the promise of text-based methods.

*Latest version available at https://dannyebanks.com/files/topic_consist.pdf. Our thanks to Mike Alvarez, Jonathan N. Katz, Gary King, Dominic Skinnion, Aleksandra Conevska, Matt Estes, Ransi Weerasooriya, Andrew Strasberg, Brandon Stewart, Justin Grimmer, Adam Breuer, Scott Abramson, and audiences at PolMeth 2024 for many helpful comments.

†Postdoctoral Fellow, Institute for Quantitative Social Science, Harvard University. Email: debanks@g.harvard.edu; Website: www.DannyEbanks.com.

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

1 Introduction

Topic models are powerful tools for labeling, summarizing, and interpreting large text corpora. They have gained wide adoption in political science for extracting themes in legislative speeches, legal rulings, media articles, and social media posts. Yet, as these models have moved into more central roles, e.g., measuring ideological polarization or estimating issue salience, we show these methods suffer from a particularly pernicious form of statistical inconsistency, resulting in incorrect conclusions about substantive issues in American politics that are of pressing interest to political scientists. Text-based evidence reveals that prevailing theories of legislative power based on roll-call votes alone are incomplete; while voting behavior implies polarization and party-leader alignment move in tandem, we provide evidence from 120 years of congressional floor speech data demonstrating that these core features of American legislative democracy have been diverging for nearly the entire period from 1899 to 2017. We find polarization is dramatically increasing during this time, while party-leader alignment is in a state of continual decline for both the Democratic and Republican parties. This new evidence underscores how modern statistical architecture fundamentally shapes substantive conclusions about American legislative democracy, renewing the promise of text data as a compelling source of evidence in political science research.

Through an empirical application to nearly 14 million U.S. Congressional floor speeches, we show that the most popular topic model methods are (1) inconsistent, and (2) misrepresent fundamental descriptive measures of American legislative politics over the last 120 years, particularly with respect to political polarization and party-leader alignment. Standard Latent Dirichlet Allocation (LDA) approaches imply that legislative polarization is in decline and that party leaders are more aligned than ever with rank-and-file members. By contrast, a consistent estimation method based on spectral decomposition, which we introduce, reveals the exact opposite trends.

Latent Dirichlet Allocation estimation via variational inference is an extremely popular text method used across the social sciences, computer science, machine learning, and political science. Accruing more than $56,882$ citations, LDA is employed in four distinct ways: (1) *measurement*, where researchers rely on topic proportions as key variables in subsequent analyses; (2) *exploratory investigation*, in which topic labels serve as a rough summary of large corpora; (3) *human-assisted document coding*, where a topic model merely aids manual labeling; and (4) *descriptive analyses* that focus on broad trends, rather than precise estimation. For uses (2)–(4), the standard practice is to deploy post-hoc validation (manually labeling or verifying topics), so the potential inconsistency of popular LDA estimators is not at issue. However, we show that when LDA is used for (1) *measurement*, the most common estimation approach, variational inference, yields results that are statistically inconsistent in a *degenerate* fashion, which we define in this paper as statistical inconsistency leading to systematically misleading parameter estimates. Such statistical inconsistency is not detected by post-hoc labeling. Although we are the first to characterize this particularly dangerous form of statistical inconsistency, we describe and discuss how variational inference suffers other types of inconsistency and instabilities in Section 7.

A systematic review of peer-reviewed publications in political science reveals that measurement is one of the leading ways that LDA is used in the discipline, with researchers frequently treating topic proportions as key variables in subsequent analyses. Indeed, from 2000 to 2022, topic models were the single most popular machine-learning technique in the field (de Slegte, Van Droogenbroeck, et al., 2024). Extending the review through 2025, we identify a total of 102 published papers that use topic modeling across the entire period, 93 percent of which report no attempts at downstream validation, and 84 percent of which rely on provably inconsistent estimators when reporting quantities of interest. Such divergence between standard LDA methods and consistent alternatives poses a serious risk for studies that adopt LDA as a measurement tool. In particular, no amount of post-hoc topic labeling can fix the resulting *degenerate inconsistency*; the estimation procedure itself must yield valid recovery. Accordingly, this paper provides (1) theoretical justification for why standard variational methods fail, and (2) a large-scale, *consistent* alternative based on tensor-based spectral decomposition that resolves the problem.

In Section 2, we introduce a novel taxonomy of types of statistical inconsistency and explain why the nature of LDA's inconsistency is especially harmful for substantive claims in political science applications. Section 3 lays out the LDA data-generating process. Section 4 then offers a mathematical proof that variational inference fails for LDA, and subsequently provides theoretical and simulation-based evidence of the problem. In a previous technical paper (Kangaslahti, Ebanks, et al., 2025), we introduce a spectral approach to topic modelling that provides consistency guarantees; Section 5 of this paper contributes a brief intuitive explanation of this approach. In (Section 6) we illustrate the dramatic differences in how our approach describes America political history, demonstrating starkly different conclusions on polarization and party-leader alignment when consistent methods are employed, in contrast with previous inconsistent methods. We then show that text-based measures build on existing empirical methods to study these questions (such as D.W. Nominate) and provide new evidence revealing polarization and party-leader alignment have moved in exactly the opposite direction over the last 120 years, confounding predictions from long-standing theories of legislative power. These results renew the promise of using statistically consistent, text-based methods to gain new insights into core questions about the nature of American legislative democracy and beyond. We close by discussing known limitations and broader implications for text-based measurement in political science (Sections 7 and 8).

2 When Inconsistency Implicates Substantive Findings

Statistical inconsistency is the failure of an estimator to converge to the true parameter value as the sample size tends to infinity. We offer here a classification of different types of inconsistency. We then show that the form of inconsistency plaguing topic models is a particularly pernicious type, which we call *degenerate* inconsistency, as it yields systematically misleading results. Below, we propose a taxonomy of inconsistency and then discuss why, when treating topic proportions as genuine measures of political phenomena, consistency is critical for substantive interpretation.

We define, interpret and provide examples for a taxonomy of three categories of inconsistency:

1. **Benign Inconsistency.**

   Interpretation. For this type of inconsistent estimator, the estimator carries a bias (i.e., offset from the true value), yet this offset remains small enough that substantive inferences are not drastically distorted. In other words, as the sample grows, the estimator does *not* hone in on the truth precisely, but it also does not "drift away" or explode in magnitude. This kind of stable, slight bias can be tolerable in many practical settings.

   Definition. Formally, let $\theta^* \in \Theta$ be the true parameter (or parameter vector) of interest, and let $\{\hat{\theta}_n\}$ be an estimator sequence based on n observations. We say that $\{\hat{\theta}_n\}$ exhibits *benign inconsistency* if it does *not* converge to θ^* but remains within a fixed, bounded neighborhood of θ^* as $n \to \infty$. Concretely,

   $$\limsup_{n\to\infty} \|\hat{\theta}_n - \theta^*\| \;=\; c < \infty,$$

   where $c > 0$ is a constant that does not vanish, is not too large, but also does not grow with n.

   Example (Additive Offset). Suppose the true parameter is $\theta^* = 0$. An estimator $\hat{\theta}_n$ such that

   $$\hat{\theta}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i \;+\; 0.05$$

never converges exactly to 0 (because of the fixed $+0.05$), so it is inconsistent in the strict sense. However, the estimator is *benignly inconsistent* because the bias of 0.05 does not grow with n; it remains a small, bounded discrepancy. Empirical results based on $\hat{\theta}_n$ may still be quite accurate for large n, albeit slightly shifted.

2. **Useful Inconsistency.**

   Interpretation. This scenario arises if the parameter space is high-dimensional, or the data-generating process drifts over time, but the quantity we truly care about (captured by a function of the parameters, g) is still recovered. In forecasting or control problems, for instance, the exact parameter may be less relevant than stable predictions of future observations[1].

   We next define a scenario in which $\{\hat{\theta}_n\}$ does not converge to θ^* but remains effective for certain predictive tasks.

   Definition. Let $g : \Theta \to \mathcal{Y}$ be a functional (e.g., a prediction rule). If

   $$\hat{\theta}_n \text{ does not converge to } \theta^*, \quad \text{but} \quad g(\hat{\theta}_n) \xrightarrow{\text{p}} g(\theta^*),$$

   then $\{\hat{\theta}_n\}$ is *usefully inconsistent*. It fails in a strict, parameter-estimation sense, yet achieves accurate asymptotic performance with respect to the predictive function g. In other words, even though $\hat{\theta}_n$ itself is inconsistent, its *predictions* or *implied decisions* may converge to the true ones.

   Example. Consider a simple linear model:

   Consider the same simple linear model:

   $$Y = \theta^* X + \varepsilon,$$

   where Y is the outcome, X the predictor, θ^* the true (but unknown) slope parameter, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is mean-zero Gaussian noise. Suppose we estimate θ by a sequence $\{\hat{\theta}_n\}$ that actually *diverges* (i.e. it does not stay near θ^*), but in such a way that the product $\hat{\theta}_n X_n$ remains near $\theta^* X_n$. For example, this can happen if $X_n \to 0$ as $n \to \infty$ while $\hat{\theta}_n \to \infty$ in such a manner that $\hat{\theta}_n X_n \to \theta^* X_n$.

   Now let

   $$g(\theta) \;=\; \Big[\, \theta X \;-\; z_{\alpha/2}\, \sigma,\;\; \theta X \;+\; z_{\alpha/2}\, \sigma \Big]$$

[1] In real-world applications, GPS location algorithms are a widely used technology that rely on this type of inconsistency.

denote the $(1-\alpha)$ predictive interval for Y, centered at θX with half-width $z_{\alpha/2}\,\sigma$. Although $\hat{\theta}_n$ may be diverging and never settles to θ^* itself, the *interval* $g(\hat{\theta}_n)$ converges to the "true" interval $g(\theta^*)$, so that its coverage probability and center/width are asymptotically correct for predicting Y. This is precisely a *useful inconsistency*: the estimator $\hat{\theta}_n$ fails to converge to the true parameter, yet its predictive intervals remain valid and converge to those that we would form if θ^* were known.

This type of behavior occurs most frequently in high-dimensional settings or machine learning, where we care about predictive accuracy rather than exact parameter recovery[2].

3. **Degenerate Inconsistency:**

   Interpretation. The *worst-case* form of inconsistency arises when $\{\hat{\theta}_n\}$ not only fails to converge to θ^* but actively diverges or systematically encodes noise as signal. Degenerate inconsistency is highly damaging because it may *reverse* the researchers conclusions or produce misleading patterns that appear coherent but are in fact just overfitted noise. When n increases, we expect better estimates under standard large-sample arguments, but this form of inconsistency flouts that intuition by growing worse over time.

   Definition. Formally, we say $\{\hat{\theta}_n\}$ is *degenerately inconsistent* if there exists an unbounded sequence $a_n \to \infty$ such that

$$\limsup \|\hat{\theta}_n - \theta^*\| \;=\; O_p(a_n), \quad \text{with } a_n \to \infty \text{ as } n \to \infty.$$

   The $O_p(a_n)$ notation means that the distance between estimator, $\hat{\theta}_n$ and the true value, θ^*, grows to infinity in the probability limit. In other words, the estimators distance from the truth will grow indefinitely as more observations are collected.

   Example (Unbounded Parameter Drift). Consider a scenario where $\hat{\theta}_n$ is obtained by an ill-posed optimization that inadvertently amplifies noise as n grows (e.g., by introducing spurious extra parameters or runaway scaling factors). The estimator

[2] We note that inconsistent methods can be useful, outside of the toy example provided above. For example, non-Convergent Kalman Filters are a popular dynamic state-based forecasting method, with applications from engineering to macroeconomic models. Consider a linear state-space model in which a Kalman filter aims to estimate a time-varying state θ_t. If the underlying state evolves unpredictably, the filter may *never converge* in distribution to any fixed θ^* (strictly speaking, it cannot as no θ^* exists). Yet, for each t, it can track the signal well enough that its predictive error remains bounded and small. So as an estimator of "the truth," it fails because there is no fixed truth to which to converge, but as a tool for real-time prediction, it is highly effective. Thus, it is *usefully inconsistent*: it does not converge in the parameter-estimation sense, yet it accomplishes its predictive task reliably.

might select $\hat{\theta}_n$ such that

$$\|\hat{\theta}_n\| \approx n^{\gamma}, \quad \gamma > 0,$$

so the distance $\|\hat{\theta}_n - \theta^*\|$ explodes with sample size. This renders the inference systematically misleading, as $\hat{\theta}_n$ is not merely biased by a fixed amount but becomes arbitrarily large or erratic, completely losing touch with θ^*.

We show in Section H that variational inference for LDA exhibits this degenerate form of inconsistency. Instead of refining topic estimates as more documents are ingested, the estimator is increasingly likely to merge or split topics incorrectly, in ways that are masked by the models approximate posterior. The very intuition that more data should improve the reliability of the model estimates is turned on its head.

3 Definition the Latent Dirichlet Allocation Data-Generation Process

We now summarize the data generation process for Latent Dirichlet Allocation (LDA), which we fully detail in Appendix A.2. We introduce a minimal running example for the purposes of explanation. Suppose for the sake of this example that a document contains only three words, “trade", “treaty", and “tariff", out of a possible total of V words that exist across all possible documents. Now, we have collected N documents in total, indexed by $n = 1, ..., N$. According to LDA, the document is composed of words, or tokens, which we will collect in a vector f (thus f will tell us how many times “trade", “treaty", and “tariff" appear in the document). This vector is one row, of length W, the total number of words in the document (in our case, 3, and a subset of a possible total of V words that appear across the entire set of documents). Each entry in this vector represents the number of times a word appears in that document. Suppose for the sake of illustration that each word appears one time. Then, the LDA generative story for this document is one in which (1) we select a topic for the document according to a multinomial probability distribution over K latent topics with topic label z_n. Suppose we have two topics, “Trade Relations" and “Foreign Affairs". We find this multinomial distribution itself by drawing it from a Dirichlet distribution with mixing parameter α and (2) conditional on each topic, each word in the document corresponds to a multinomial distribution over V possible words, again itself a distribution drawn from a Dirichlet distribution with mixing parameter β. Each document then has two features:

1. *topic mixture*: a vector of length K which states the probability a document belongs to each topic.

2. *word probability matrix*: a matrix of size $V \times K$ which tells us the probability each word appears in a topic.

Given our data, we might find that the document is 75 percent likely to be a “Trade Relations" topic and 25 percent likely to be a “"Foreign Affairs" topic. The model assumes no grammatical structure and each document is assumed a collection of words. While not an accurate model of language, LDA's simplifying assumptions allow for researchers to measure precise amounts of each of multiple, unknown topics in a each document, as we have done in our simple running example. This approach remains popular in political science applications, particularly cases where documents do not necessarily belong to a single topic, where the researcher does not have strong priors on what topics are in the data, and when theoretical frameworks are not especially informative about the underlying nature of the data. We now summarize the data generation process for LDA.

We have a single document n containing W_n words (out of a total possible V words), represented by $f_{n,1}, \ldots, f_{n,W_n}$, the collection of all words in the document. Latent variables:

- $h_n \in \Delta^{K-1}$: the topic-mixing vector (drawn from a Dirichlet(α)),
- $z_n = (z_{n,1}, \ldots, z_{n,W_n})$: topic assignments for each word,
- $\mu_k \in \Delta^{V-1}$: the word distribution for topic k, each drawn from $\mathrm{Dirichlet}(\beta)$.

Given these primitives, the data are generated according the following procedure:

1. Each document n draws a *topic-mixing vector* $h_n \sim \mathrm{Dirichlet}(\alpha)$.
2. Each topic k draws a *word-distribution vector* $\mu_k \sim \mathrm{Dirichlet}(\beta)$.
3. To form document i, each word position is assigned to a topic z_i according to h_i, and then an observed token f_n is drawn from μ_{z_i}.

The researcher observes only the words and attempts to back out the topic distributions $\{\mu_k\}$ and the document-specific topic weights $\{h_i\}$. The likelihood function for each document has the following form, with explicit dependence between the joint probability of observing a word in a given topic and a given topic given a document.

Complete-Data Likelihood (Document n)

The complete-data likelihood for all observed words f_n *and* latent variables $(z_n, h_n, \{\mu_k\})$ is:

$$
\begin{aligned}
&p\big(f_n,\ z_n,\ h_n,\ \{\mu_k\}_{k=1}^K \mid \alpha, \beta\big) \\
&= \underbrace{p(h_n \mid \alpha)}_{\text{document's topic mixture}} \times \underbrace{\prod_{k=1}^{K} p(\mu_k \mid \beta)}_{\text{topic word distributions}} \times \underbrace{\prod_{j=1}^{W_n} \Big[p\big(z_{n,j} \mid h_n\big)\ p\big(f_{n,j} \mid \mu_{z_{n,j}}\big)\Big]}_{\text{topic assignment and word generation}} \\
&= \left[\frac{\Gamma\big(\sum_{k=1}^K \alpha_k\big)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^{K} h_{n,k}^{\alpha_k - 1}\right] \times \prod_{k=1}^{K} \left[\frac{\Gamma\big(\sum_{v=1}^V \beta_v\big)}{\prod_{v=1}^V \Gamma(\beta_v)} \prod_{v=1}^{V} \mu_{k,v}^{\beta_v - 1}\right] \\
&\quad \underbrace{\times \prod_{j=1}^{W_n} \Big[\prod_{k=1}^{K} h_{n,k}^{\mathbf{1}\{z_{n,j}=k\}} \times \prod_{v=1}^{V} \mu_{z_{n,j},v}^{\mathbf{1}\{f_{n,j}=v\}}\Big]}_{\text{Intractable interdependency between } \mu \text{ and } h}. \qquad (3)
\end{aligned}
$$

Estimation is rendered challenging by the intractable normalizing constant in the joint posterior, which is typically bypassed using *variational inference* (VI). Although faster than exact Bayesian or Gibbs sampling methods in high dimensions, VI's simplifying assumptions *break* the dependence between μ and h in ways that cause more fundamental statistical pathologies, as we show below.

4 Evidence of Degenerate Inconsistency

We next present mathematical proofs and simulation evidence to underscore that standard LDA under variational inference is not merely inconsistent in a a benign way but *degenerately* inconsistent. That is, these approaches give substantively misleading results as the dataset includes more documents. In Appendix J, we provide an empirical demonstration of how changing contexts surrounding the political rhetoric trade drive this manifest as this inconsistency. Finally, in Section 6, we show how degenerate inconsistency leads to misleading substantive findings in the study of legislative politics using Congressional floor speeches over a 120 year period.

4.1 Statistical Intuition for Degenerate Inconsistency in Topic Models

The key insight for the inconsistency argument is that the the underlying topic-probability matrices for words and documents might be permuted in such a way that yields the same

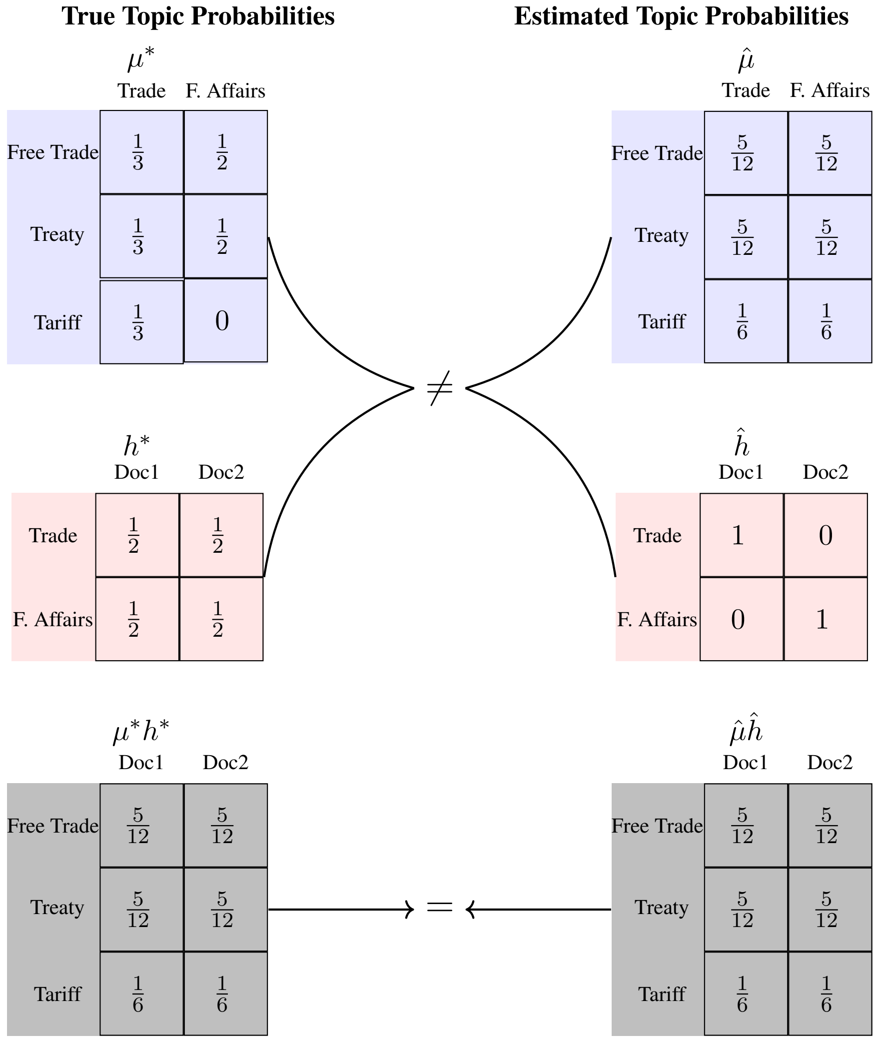


Figure 1: Numerical Example: Even though $\mu^* h^* = \hat{\mu}\hat{h}$, the factorized components differ. The rows of μ are words in the corpous: “Free Trade," “Treaty," “Tariff, and the columns are topics: “Trade" and “Foreign Affairs." The rows of h are topics, “Trade" and “Foreign Affairs." This mirrors illustrates the core source of the inconsistency: identical joint probabilities but permuted factorization in the underlying topic-word and document-topic probability matrices.

overall joint probabilities. We now demonstrate the intuition for this mathematical argument in Figure 1, by way of a numerical example. In blue, we write out the topic-word probability matrix, μ, which denotes the probability a word appears in a topic. Then in red, we write out the document-topic matrix, h, which denotes the probability a document belongs to a topic. Finally, in gray, we report the joint probability of observing a word and document at the same time, μh.

In Figure 1, we permute the underlying columns of the topic-word probability matrix, μ^*, and the rows of the document-topic probability matrix, h^*. We do so in such a way that μ^* and h^*, the ground-truth values, do not equal $\hat{\mu}$ and $\hat{h}$. Yet, when we multiply these underlying matrices, their products yield precisely the *same* document-word matrix $\hat{\mu}\hat{h}$ and $\mu^* h^*$. This fact holds true even though the latent true probabilities for the topics differ drastically. As we show below, for increasingly large corpora, such degenerate solutions become more likely. This leaves researchers in the undesirable scenario of having estimated systematically incorrect, but plausible, topic probabilities, as in our example here.

Figure 2 offers a conceptual depiction at a more general level. The left panel shows the true joint probability matrix for words and topics; the right panel shows an apparently identical joint distribution but formed by *swapped* or *duplicated* topics. Although the final likelihood is identical, the interpretation of those latent topics is completely changed and can systematically distort measurement. The figure follows a step-by-step process demonstrating how variational approximations introduce systematic topic misalignment, ultimately leading degenerately inconsistent topic distributions as dataset size increases.

The leftmost panel represents the Ground-Truth Joint Probability Matrix, denoted as $\mu\mathbf{h}$. This matrix correctly encodes the true relationship between topics and words, ensuring that document-topic distributions and word-topic distributions align as expected. However, the middle panel illustrates the variational approximation step, where the fundamental challenge arises. Instead of estimating the true posterior, variational inference factorizes dependencies between document-topic and word-topic probabilities, leading to a simplified form:

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

While this factorization enables computational feasibility, it introduces systematic estimation errors. The rightmost panel in the figure depicts the Estimated Joint Probability Matrix, which retains the same overall structure as the ground-truth matrix but with incorrectly reallocated topic proportions. Despite this misalignment, the estimated model produces the same joint probability outputs, effectively masking the underlying topic misallocation.

4.2 Theoretical Evidence

We prove in Section B that LDA estimated via variational inferenc is inconsistent and then we prove in Section H that this inconsistency is degenerate. Here, we provide a short proof sketch to provide intuition for the core of the two arguments. Recall that degenerate inconsistency implies there exists a sequence $a_n \to \infty$ such that

$$\|\hat{\mu}_n - \mu^*\| \; = \; O_p(N),$$

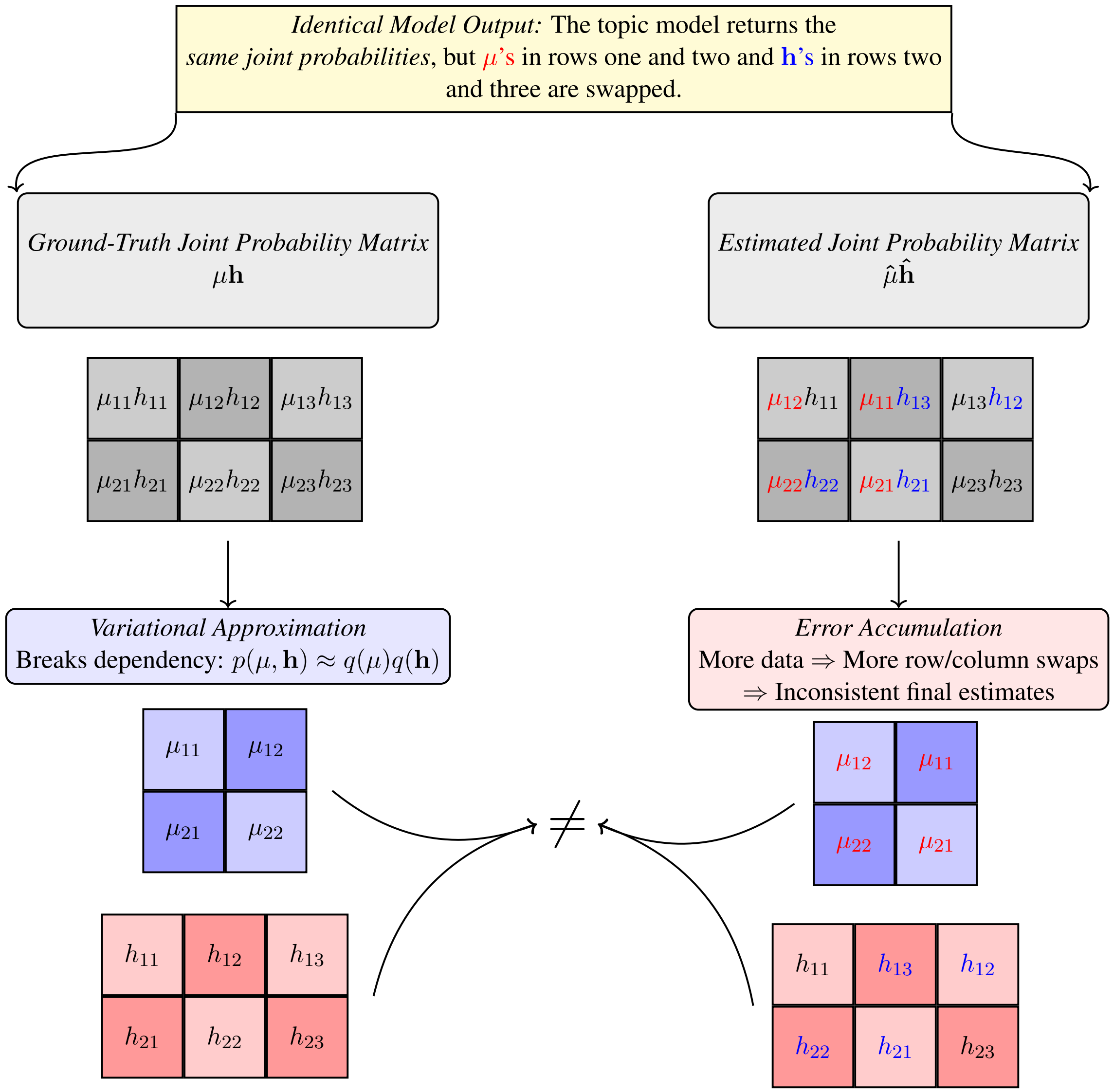


Figure 2: *Illustration of Variational Inference Error in LDA.* Even if the estimated joint probability matrix P_{est} matches the true P_{true}, the underlying decomposition into topic-word (μ) and document-topic (h) matrices may be *misaligned*.

where μ^* is the true topic-word matrix, and $\hat{\mu}_n$ is the estimator from variational inference (VI) on an LDA model with n documents. We outline why VI-based LDA satisfies this criterion:

1. *Revisiting the VI the factorization.* Variational inference for LDA factorizes the posterior in equation 3 into

$$q(\mu, h, z \mid \gamma, \phi) = \underbrace{q(z)}_{\text{topic assignments}} \underbrace{\prod_{i=1}^{n} q(\mu \mid \phi_i)\, q(h \mid \gamma)}_{\text{breaks } (\mu, h) \text{ dependence}},$$

and minimizes the KL divergence between $p(\mu, h, z \mid f_i)$ and $q(\mu, h, z)$.

2. *Duplicate-column argument.* As shown in many LDA consistency results (e.g., Nakajima, Sato, et al. 2014), one can *duplicate* a column of the topic matrix μ, divide those two columns by $1/2$, and still yield *identical* joint probabilities for the document-word matrix μh. Concretely:

$$\left(\mu^*\right) \;=\; \left[\mu_{:,1},\; \mu_{:,2},\; \ldots,\; \mu_{:,K}\right] \quad \mapsto \quad \left[\tfrac{1}{2}\mu_{:,1},\; \tfrac{1}{2}\mu_{:,1},\; \mu_{:,2}, \ldots, \mu_{:,K}\right]$$

   yields the same product μh. This "column-splitting" phenomenon illustrates how *multiple spurious solutions* might arise in the variational solution.

3. *Increasing dimension of spurious solutions.* As $n \to \infty$, the probability that VI converges to one of these spurious (incorrect) solutions grows, because more documents create more opportunities for minor errors in topic assignments that lead to duplicated or merged columns. Hence, the estimator $\hat{\mu}_n$ can systematically stray from μ^*.

4. *Formalizing the divergence.* If the measure of mismatch $\|\hat{\mu}_n - \mu^*\|$ is evaluated under an L_1 or L_2 norm, it can *increase* with n. Indeed, as new documents enter, VI is more likely to split or merge columns, compounding the misallocation of words to topics and giving $\hat{\mu}_n$ a distance from μ^* that grows unbounded. Formally,

$$\|\hat{\mu}_n - \mu^*\| \;=\; O_p(I\,W + V\,N),$$

   where $\|\cdot\|$ shows that for the word-topic probability matrix μ, the distance between the population parameter and the estimate grows linearly in the number of documents, N.

5. *Conclusion: Degenerate inconsistency.* By definition, an estimator that *fails to converge* and indeed *diverges* (and grows unboundedly in the limit) is degenerately inconsistent. Thus, VI-based LDA satisfies the third, most severe category of inconsistency described in Section 2, in which errors *accumulate* with sample size and may systematically invert substantive conclusions.

Hence, LDA with variational inference does more than simply fail to converge, it introduces opportunities for topics to be duplicated, merged, or permuted with probability increasing in n, thereby driving the parameter estimates arbitrarily far away from μ^*, in an unconstrained fashion. This behavior meets the formal criterion for *degenerate inconsistency*.

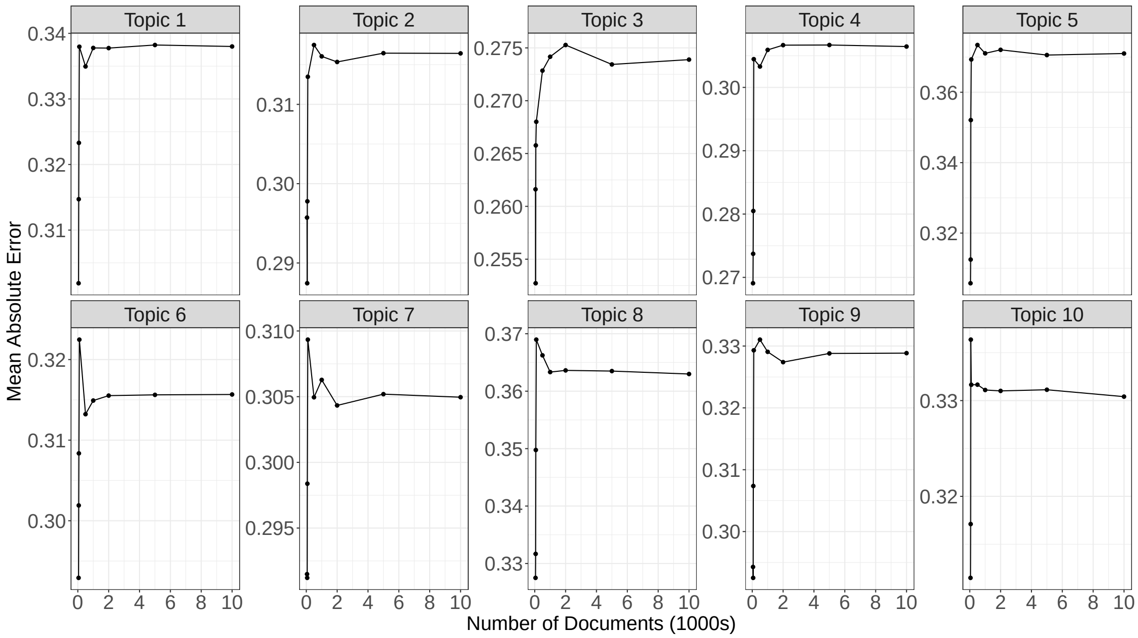


Figure 3: Mean Absolute Error (MAE) Between True (μ^*) and Estimated ($\hat{\mu}$) Word Probabilities: This figure presents the MAE for the top 15 words across ten topics that were used to generate a synthetic dataset. We generated 10,000 synthetic documents from the LDA data-generating process described earlier, and then conducted 100 simulation runs where subsets of these documents, ranging from 50 to the full 10,000, were sampled. For each subset size, we estimated the topic model via variational inference and compared the posterior probabilities of the top words in each topic to their known true probabilities. The plotted values represent the average MAE across all simulations for each subset size.

4.3 Simulation Evidence

Applied researchers commonly expect that as the volume of data increases, topic model estimates should become more accurate. However, with standard Latent Dirichlet Allocation (LDA) using variational inference, we encounter precisely the opposite scenario: beyond a certain dataset size, model errors not only persist but can worsen significantly.

To grasp why this happens, imagine the task of categorizing politically charged words like *free trade*, *treaty* or *tariff*. Depending on context, these terms could reasonably belong to distinct topics such as "international trade," "foreign policy," or even "sources of domestic political conflict." With smaller datasets, researchers typically resolve these ambiguities by direct manual interpretation. As datasets grow, manual interpretation becomes impractical, and automated models must resolve these ambiguities without direct oversight.

Our simulation demonstrates this numerically. We created 10,000 documents from a known mixture of fifteen distinct topics. For increasingly large subsets of this corpus, we repeatedly estimated LDA models via variational inference. Each simulation yields an estimated probability of words belonging to specific topics. We then compared these estimates to the known "ground truth" probabilities.

Surprisingly, rather than improving accuracy, we found that estimates produced by variational inference became increasingly error-prone with more data (Figure 3). Errors increase because larger datasets introduce more possibilities for topic conflation or duplication, what we term a "catastrophic merge" or "split." Thus, instead of benefiting from more data, variational inference methods experience degenerative inconsistencies, providing increasingly distorted views of the underlying topic structure. In essence, more data ironically introduces more opportunities for systematic errors in standard LDA.

We also find that the probability of an exact match between the estimated model and the ground truth *decreases* as N increases. For a large corpus, the best-case chance of correct topic identification can become vanishingly small under the standard LDA approach, regardless of the true number of topics.

5 A Consistent Spectral Method for Large-Scale Topic Modeling

To remedy these problems, we adopt a *tensor-based spectral decomposition* for LDA, as developed by (Anandkumar, Foster, et al., 2013; Anandkumar, Ge, et al., 2014) and adapted and implemented for large scale data, which we developed in (Kangaslahti, Ebanks, et al., 2025). The solution itself is highly technical, and so we provide an intuitive explanation for spectral approaches that are targeted for applied researchers in political science. The core insight is that we can attain consistent estimates for LDA by first estimating the topic-word probabilities, μ. Through method of moments and eigenvalue decomposition, we can obtain consistent estimates for the topic-word probabilities, μ. With consistent, accurate estimates of μ in hand, we can then estimate the document-topic probabilities, h. By estimating μ and h separately, the spectral approach sidesteps the dependence of μ and h in the estimation step, avoiding the core source of inconsistency for LDA model estimates obtained via variational inference.

The spectral approach has three desirable properties. First, it ensure a unique topic-word probability result, $\hat{\mu}$.[3] Second, the approach has *closed-form consistency guarantees*, meaning that with enough data N, the estimated $\hat{\mu}$ converges to μ^*. And third, the spectral approach is computationally tractable even for corpora with millions of documents, by performing calculations in a streaming and batched fashion.[4]

[3] We achieve an accurate eigenvalue decomposition of the word-topic matrix μ^* by exploiting orthogonality constraints.

[4] Computer scientists call a method that automatically updates with new data a method that is fully

5.1 Theory Underpinning the Spectral Approach

The spectral approach achieves accuracy by estimating the word-topic matrix μ and document topic matrix, h, separately. Let μ^* be the true topic-word matrix with K topics, each topic being a distribution over V words. If μ^* has no collinear columns (i.e. K linearly independent columns), there exist theoretical results showing that one can invert a system of second- and third-moment tensors (essentially, co-occurrence and tri-occurrence of words) to recover μ^* exactly (up to a simple permutation of topics). In formal terms, Anandkumar, Foster, et al., 2013 show that

$$\|\hat{\mu} - \mu^*\|_2 = O_p\Big(\frac{1}{\sqrt{N}}\Big),$$

so as $N \to \infty$, the difference goes to zero in probability, establishing consistency. With consistent estimates of μ in hand, we no longer have to worry about the interdependence between μ and h (we have already estimated μ). So, we can then estimate the the document-topic probability matrix h with variational inference, and standard consistency results will hold in that case. Although this approach seems incongruous with the preceding arguments against variational inference, recall that the reason variational inference does not provide consistent results is because of numerical pathologies introduced by estimating the topic word matrix μ and topic-document matrix h jointly. Spectral methods resolve this problem by first consistently estimating μ alone. Then, we can consistently estimate the implied topic-document matrix, h, with variational inference.

5.2 Proven Consistency at Scale

Spectral methods cleverly decompose the data into their principal components, which are then used to estimate the topic-word probabilities using a method-of-moments approach.[5] That is, these approaches match the empirical mean, variance, covariance, and skew to the theoretical mean, variance, covariance, and skew implied by the LDA data generation process. We calculate these theoretical values from the principal components of the data. By using a method-of-moments approach, spectral methods ensure that the parameter solution is uniquely identified. Furthermore, recent implementations (Kangaslahti, Ebanks, et al., 2025) batch or stream large corpora so that the memory usage remains modest, enabling *practical* runs on large datasets. Crucially, as N grows, the method's estimates *improve* rather than degrade.

online, meaning we can stream data in for real-time estimates of the results. Our spectral algorithm offers a fully online version (Kangaslahti, Ebanks, et al., 2025).

[5] Moments are various measurements of the shape of a statistical distribution, such as the mean, variance, covariance, skew, and kurtosis, among others.

While variational inference can be faster in small data regimes, the spectral method's advantage in consistency becomes vital for large-scale corpora. If the goal is truly to measure or infer stable latent topics, spectral decomposition offers an approach that is provably robust as data accumulate, avoiding the degeneracies that plague variational inference-based LDA.

6 How Topic Modeling Estimators Shape Our Understanding of Legislative Politics

We now show how these statistical issues manifest in practice by comparing standard LDA (variational inference) with a consistent spectral method on a large corpus of U.S. Congressional floor speeches. Comparing the two approaches reveals a dramatic shift in how we understand the last 120 years of American legislative party politics. We examine two key quantities of interest in American politics: (1) trends in polarization, and (2) levels of alignment between party leaders and rank-and-file members. To study these quantities, political scientists have relied on roll-call-votes-based measures, which are favored because they have a long historical record. We show that speech-based measures provide a significant departure from established theories of congressional party leadership, which traditionally anchor their expectations in ideological coherence and legislative institutional contexts. We fully explore these theoretical expectations in Appendix I. Ultimately, by employing consistent text-based methods, we illuminate previously obscured divergences between polarization and party-leader alignment, reshaping our theoretical understanding of legislative power dynamics. Our findings renew the promise of text-based methods in political science, providing a clearer picture of party leadership dynamics in the American legislature over the last 120 years.

6.1 Analyzing 14 Million Congressional Speeches

We use the dataset of Congressional floor speeches compiled by Gentzkow, Shapiro, and Taddy (2018), which is comprised of 13,818,250 floor speeches in the U.S. Congress (both House and Senate) from the 43rd through 114th Congress (1873 -2017). After standard text cleaning and allowing for bigrams, we fit two models to the entire dataset. First, we fit our preferred estimator *Spectral LDA* (`TensorLy-LDA`), which has proven asymptotic consistency (Anandkumar, Foster, et al., 2013; Kangaslahti, Ebanks, et al., 2025). We compare it against a standard, inconsistent estimator that uses *Variational LDA* (`Gensim-LDAMulticore`), a widely used implementation of the original Blei, Ng, and Jordan, 2003b method with parallelization. After tuning both models, we construct two measures of polarization and party-leader alignment.[6]

6.2 Defining Polarization and Party-Leader Alignment Using Text

To demonstrate the promise of our consistent topical modelling approach, we intuitively describe, justify, and formally define three quantities of interest. We will use the probability outputs of our topic model to measure which topics legislators are discussing on the Floor of the U.S. House. We we designate this measure as the legislators' floor speech topic profiles. Taking these floor speech topic profiles, we will construct measures of inter-party polarization and intra-party party-leader alignment.

1. **Legislators' Floor Speech Topic Profiles**

   Intuition: Imagine each legislator as maintaining a *Floor Speech Profile*, showing how their annual discourse (in period t) is proportionally allocated across K latent topics. Concretely, legislator i has a set of proportions $\left(\theta_{i,1,t}, \ldots, \theta_{i,K,t}\right)$ that sum to 1, indicating the distribution of their remarks among those topics.

   Theoretical Motivation: Defining a legislators topical mixture as their floor speech profile aligns with previous empirical work on issue attention and agenda-setting, wherein each official's political messaging profile spans multiple policy areas, allowing for topic heterogeneity within the same speech (Barbera, Casas, et al., 2019; de Slegte, Van Droogenbroeck, et al., 2024).

   Statistical Motivation: Topic models produce average topical probabilities for each legislator. These probabilities form a distribution over K categories, naturally summing to 1. This gives us a natural measure of how legislators engage with topics on the Floor of the U.S. House over a 120 year period. As floor speeches are long and generally contain more than one topic, a topic model approach better measures the topical mix of speech than a single-issue labelling approach.

[6] We set the number of topics to 30 for both methods, guided by standard coherence metrics (Röder, Both, and Hinneburg, 2015). Because of the extremely large sample size (≈ 14 million speeches), the likelihood of correct convergence under the standard LDA method is theoretically minuscule (Section 4). Below, we see the practical consequences. 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, 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.

Formal Definition. Let there be K latent topics. For each legislator i in period t, define legistor i's floor speech topic profile in year t as

$$\mathbf{x}_{i,t} \;=\; \big(\theta_{i,1,t},\;\ldots,\;\theta_{i,K,t}\big),$$

where each $\theta_{i,k,t}$ is the *average topical probability* that legislator i devotes to topic k during year t. By construction, $\sum_{k=1}^{K} \theta_{i,k,t} = 1$.

2. **Messaging Polarization**

   Intuition: We use these floor speech topic profiles to see how differently Republicans and Democrats engage in speech on the floor of the U.S. House. When cross-party profiles are highly correlated, polarization is low; when they diverge, polarization is high.

   Theoretical Motivation: Polarization reflects how far apart two groups are along important policy or rhetorical dimensions. Though roll-call votes are an important measure of this distance between the parties, votes are strategically constrained by party leadership decisions (e.g. which bills come to a vote) and can mask potential rhetorical disagreements. In contrast, analyzing floor speech topic profiles offers a more heterogeneous snapshot of how legislators publicly frame and prioritize issues.

   Statistical Motivation: Pearson's correlation coefficient captures how similarly two Floor Speech Profiles are distributed. Taking $1 - \mathrm{corr}$ yields a natural measure of dissimilarity.

   Formal Definition: Let $\mathbf{x}_i$ be the Floor Speech Profile for Republican legislator $i \in R$, and $\mathbf{x}_j$ be the Floor Speech Profile for Democratic legislator $j \in D$. We define polarization in year t as:

$$P_t \;=\; 1 \;-\; \frac{1}{|R| \times |D|} \sum_{i \in R} \sum_{j \in D} \mathrm{corr}\big(\mathbf{x}_i,\ \mathbf{x}_j\big).$$

   where $\mathrm{corr}\big(\mathbf{x}_i, \mathbf{x}_j\big)$ denotes the usual Pearson correlation between legislator i's floor speech topic profile $\mathbf{x}_i$ and legislator j's floor speech topic profile, $\mathbf{x}_j$. We subtract 1 so that lower cross-party correlation yields higher P_t.

3. **Party-Leader Alignment**

   Intuition: Within a single party, we compare each legislators Floor Speech Profile to those of the party leaders, who we define as the elected Leader, Whip, and Speaker (if applicable). A high correlation suggests leaders and party members discuss similar topics on the Floor, signaling strong alignment with leadership.

   Theoretical Motivation: This measure of party-leader alignment captures how closely a partys rank-and-file members align with their leaders rhetorical priorities. In contrast to roll-call votes, where a "no" vote might come from a hardline faction demanding a more extreme bill or a centrist faction rejecting a bill as too extreme, a speech-based measure allows for more heterogeneity on why members deviate (as it is inherently a higher-dimensional way to measure party-leader alignment).

By examining how their floor remarks compare to the leaders chosen topics and frames, we can better identify intra-party cohesion or fracture than by looking at votes alone.

Statistical Motivation: Correlations are scale-invariant and easily interpreted, allowing direct comparison of similarity of legislators' floor speech topic profiles.

Formal Definition: Let ℓ_p be the set of leaders in a party p. Denote a single party leader's floor speech topic profile as $\mathbf{x}_{\ell_p}$. Then:

$$A_{p,t} \;=\; \frac{1}{|p \setminus \{\ell_p\}|} \sum_{i \in p \setminus \{\ell_p\}} \operatorname{corr}\big(\mathbf{x}_i,\; \mathbf{x}_{\ell_p}\big),$$

which is the average correlation between each legislators floor speech topic profile and each party leaders', $\mathbf{x}_{\ell_p}$, in year t. The notation $i \in p \setminus \{\ell_p\}$ denotes the set of party members who are not in formal leadership.

6.3 Consistent Topic Models Show Party-Leader Alignment and Polarization are Diverging

In Figure 4, we report floor speech polarization as defined above. In the solid line, we show polarization when measured with a consistent topic model method, our spectral approach. In contrast, the dashed grey line shows implied polarization using the degenerately inconsistent variational inference approach. The first notable difference is the dramatic divergence in the two measures after 1930. Second, the inconsistent method implies polarization has declined since 1930, achieving nearly 100 year lows in late 2000s. Such a finding would contradict the overwhelming consensus of the literature across both quantitative and qualitative domains (Abramowitz, 2010; Hetherington, 2009; Mccarty, Poole, and Rosenthal, 2006; Paisley, 2016). Our consistent topic model methods imply very different trends in polarization compared with the degenerately inconsistent estimates. Namely, polarization is high in the late 1800s, the 1950s, and has increased from the sustained low that lasted from the 30-year period between 1970 through 2001, dramatically increasing through 2016.

We now examine party-leader alignment in within each party's messaging on the House floor. We show that our consistent, spectral topic model methods suggest the opposite dynamics for party-leader alignment than variational inference based models would imply. As formal party leadership began in 1899, our analysis of party-leader alignment begins in that year.

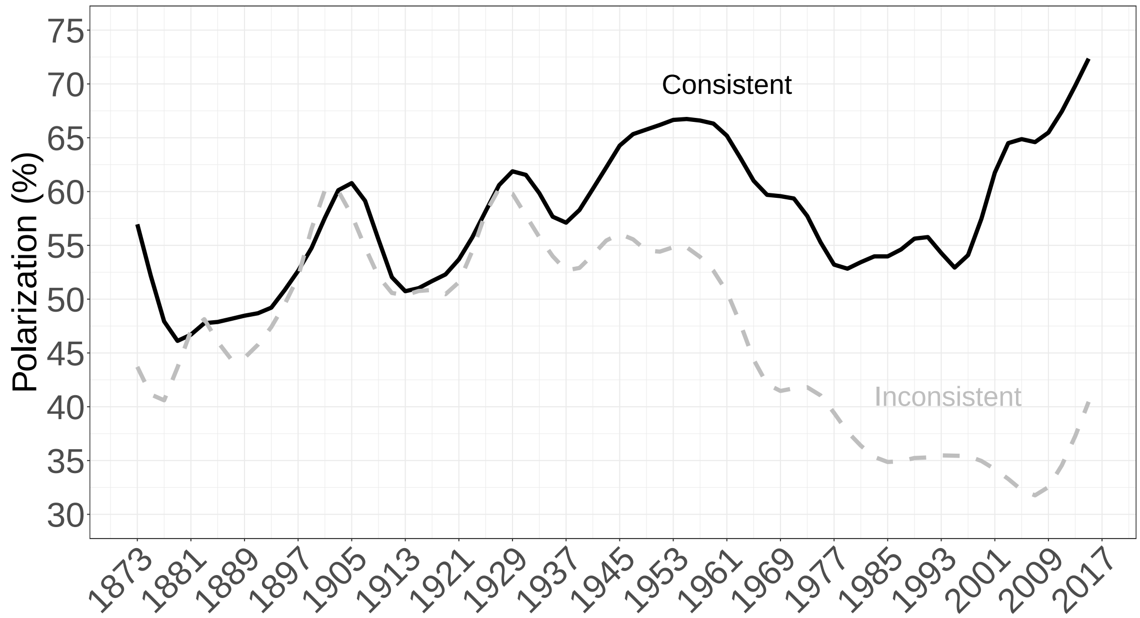
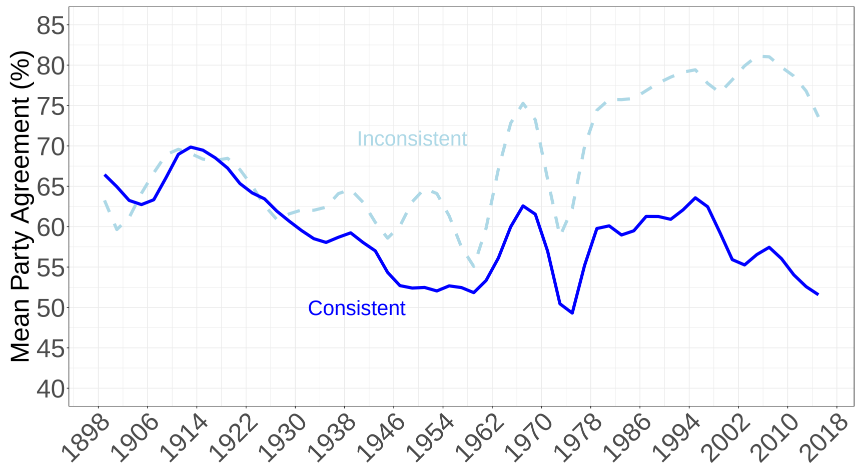


Figure 4: Speech Polarization: We calculate partisan polarization in speeches by measuring how correlated the two parties are across all topics in a given year.

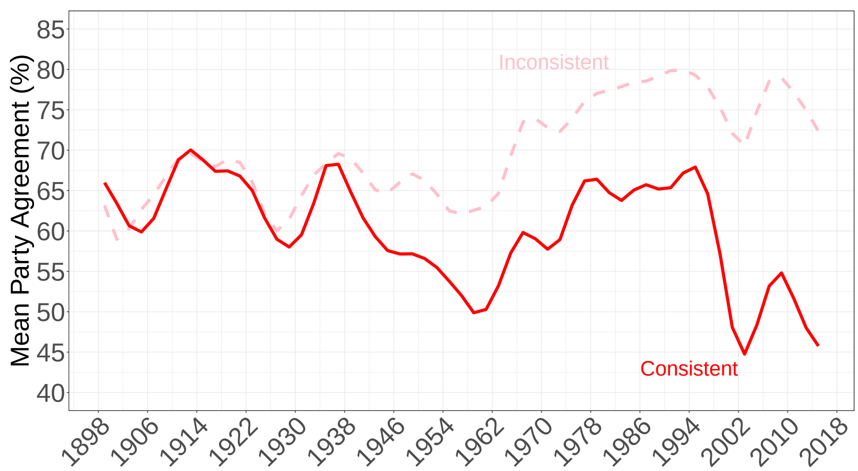
These results reveal a striking reversal in what we thought we knew about intra-party cohesion between party leaders and members over the 120 year period from 1899 through 2017. Figures 5a and 5b show that, when measured using a *consistent* topic model, both parties experience a *decline* in rhetorical alignment between leaders and rank-and-file members–whereas the standard variational (VI) approach would have us believe party-leader alignment keeps rising. For Democrats, alignment decreases steadily from the mid-1990s through 2017, mirroring previous historical lows in the 1940s and again in the post-Watergate 1970s.

The shift among Republicans is especially dramatic right after 1994–a watershed year when Newt Gingrichs leadership ushered in the "Contract with America" and a new GOP House majority. Under traditional assumptions, one might expect heightened leader-member cohesion in this era of tightly coordinated messaging. Instead, the consistent speech-based measure indicates a pronounced drop in rhetorical unity, suggesting a more multifaceted Republican discourse than voting behavior alone would imply.

In short, the consistent model shows us a party system whose members increasingly diverge from leaders in what they choose to emphasize on the floor, even though conventional methods (including VI-based approaches) might falsely suggest a party growing more cohesive over time. This fresh view of party-leader alignment underscores the power of speech data to uncover nuanced dynamics of intra-party politics that roll-call-based measure of party-leader alignment otherwise miss. By constructing quantities of interest from estimators based on floor speech data, political science researchers gain new



(a) Democratic Party-Leader Alignment



(b) Republican Party-Leader Alignment

Figure 5: Speech-Based Measures of Party-Leader Alignment: We measure the average correlation of topical compositions of members of Congress and their party leaders. We report the average correlation across all topics for all speeches for each party in a given year.

insights by directly measuring the content of each legislators messaging on the floor of the U.S. House rather than just an up-or-down vote.

That the drop in alignment occurs as party leadership seemingly tightens its grip underscores a powerful irony: while leaders may hold greater procedural control over the legislative agenda, their rank-and-file colleagues are increasingly carving out distinct rhetorical niches on the House floor. This divergence, revealed only by looking under the hood of speech-based measures, challenges conventional wisdom and implies that party discipline might be more fractured at the messaging level than at the final vote level.

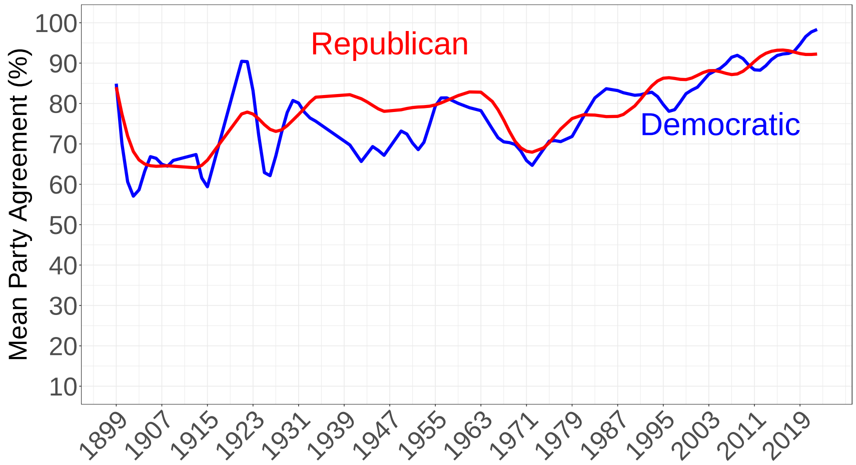
6.4 Renewing the Promise of Text Based Measures in Political Science

Text-based measures initially captivated political scientists with their promise to reveal deeper, more nuanced insights into American political dynamics. Unlike traditional vote-based methods, textual data offers a high-resolution window into legislators' ideological positions, free from the strategic distortions inherent in roll-call voting records. This original promise, however, has remained largely unfulfilled as conventional statistical methods struggled to capture these complexities at scale. With new consistent spectral methods, we now finally deliver on that original promise, uncovering previously hidden patterns of political polarization and party-leader alignment.

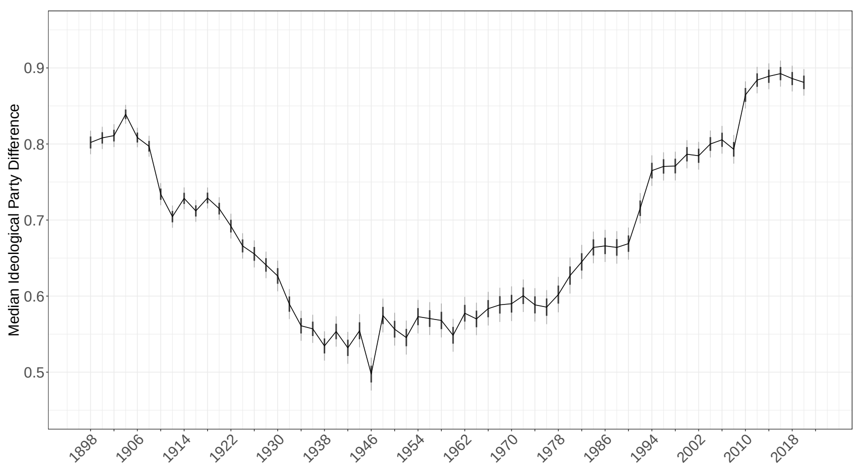
Vote-based measures, although historically abundant and conveniently quantifiable, systematically under-measure the amount of heterogeneity in the behavior of U.S. legislators. Roll-call votes frequently project an illusion of high party unity, with recorded alignment between members and their leaders typically ranging from 60 to 95 percent. At first glance, this might suggest stable, coherent leadership. But this portrayal is deeply misleading. The strategic behavior of party leaders explains why: leaders consciously sidestep divisive votes, deliberately avoiding issues that might expose fractures within their parties. Consequently, the very strategic decisions that drive the appearance of unity on recorded votes obscure authentic internal disagreements.

Similarly, vote-based ideological polarization measures like DW-Nominate depict genuine polarization in voting behavior. Augmenting these empirical findings, our consistent, text-based approach capitalizes on the expressive freedom inherent in congressional speeches to further explain the dynamics of polarization over time. Party leaders exercise far fewer veto powers over speech, making floor statements less susceptible to the strategic decisions that leads to highly constrained voting behavior. Consequently, legislators' speeches serve as an unfiltered reflection of genuine ideological divides and internal party dynamics. Using these richer textual data, our consistent topic-modeling approach uncovers substantially different, and more historically credible, patterns. Contrary to vote-based evidence, we document a sustained decline in party-leader alignment and capture a richer history of fluctuating polarization.

Now, we show and describe trends in vote-based measures favored by the literature. Figure 6a shows party-leader alignment from 1899 to 2017. We measure this by calculating the percentage of bills each year where party members voted in the same direction as their party leaders. The alignment levels are generally high, ranging from 60 to 95 percent in a given year. There are two notable periods where alignment is low: prior to the Progressive Era, when Republicans and Democrats contended with their respec-



(a) Party-Leader Alignment on Votes



(b) Ideological Polarization (DW-Nominate)

Figure 6: Votes-Based Measures

tive 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. This roll-call based evidence reflects real dynamics in how parties diverged in their pursuit of their legislative agendas, whereby the parties are further apart than ever by 2017. However, a text-based approach demonstrates floor speech polarization did not entirely dissipate as was the case for floor vote polarization. In fact, floor speach polarization experienced several peaks in 1900 and during the post-war period until the passage of the Civil Rights Act. Floor speech polarization does not increase again until 1994.

By using a consistent topic model method, we demonstrate that relationships between party leadership and their co-partisans was less cohesive in their floor speech messaging. Given floor votes are strategically constrained, the unsurprising outcome of this logic is that party leaders try 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.[7]

The same underlying issues mean that text-based approaches can provide additional insights into the study of polarization, building on the foundations established by roll-vall vote based methods, like DW-Nominate. Indeed, while other measures based on campaign finance data (Bonica, 2014) and Congressional Twitter activity (Barberá, 2015) are used to study legislative polarization, these approaches do not have a historical record of congressional floor speeches. Importantly, speech-based approaches can give historical insights over 120 years of American history and provide coherent historical descriptions of how polarization and party-leader alignment have evolved. 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.

Our text-based approach breaks through the strategic distortions that have long plagued roll-call data. By tapping the expressive range of legislative speech, we recover new evidence of 120 years intra-party tensions and ideological fragmentation. Rather than smoothing away complexity, text-based approaches reveal the churn and contestation that shape how political partys exercise power and organize over time. Far from merely challenging standard theories of Congress, these tools renew the original promise of text analysis in political science by providing new evidence pinpointing where and when parties cohere, fracture, and realign in ways that roll-call votes alone cannot capture.

7 Known Conceptual and Theoretical Issues

7.1 When Downstream Validation Is Not Enough

As Grimmer and Stewart (2013), Ying, Montgomery, and Stewart (2022), Grimmer, Roberts, and Stewart (2022), and others convincingly demonstrate, downstream validation practices are necessary for robustness tests for text labeling applications, and such approaches are critical for establishing measurement validity. Accordingly, one compelling rebuttal to concerns about topic model statistical inconsistency is that researchers routinely validate their models by inspecting topic labels, applying human judgment, or using topic distributions in downstream tasks. We note two problems, however. First, topic modelling is the most popular machine learning method for applied researchers (de Slegte,

[7] 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.

Van Droogenbroeck, et al., 2024), yet a systematic content analysis of all 102 papers that use topic modelling methods from 2000 to 2025, reveals that only 3 percent of these articles perform any rigorous downstream validation (including reporting an inter-coder reliability or validating against known or expert-derived labels). More disconcertingly, the articles that did perform downstream validation were exclusively papers advocating for the downstream validation approach in the first place. Second, of the remaining 97 percent of papers that used topic modelling, the vast majority, 84 percent, use inconsistent topic model probability estimates and document composition estimates, which cannot be validated by human hand-coding or by expert judgment.

Appropriate Uses of Downstream Validation:

- *Exploratory Research*: If the goal is to summarize a corpus or identify broad patterns, human labeling and post-hoc validation can be sufficient.
- *Human-Assisted Coding*: When topic models are used to assist human coders in labeling documents, their statistical properties matter less because final coding decisions rest with researchers.

However, downstream validation alone cannot salvage a model that is asymptotically misleading if the analyst is using those estimated topic distributions as *variables* in subsequent analyses. Suppose a researcher treats the estimated topic proportions for each legislator as measures of ideology or legislative attention and then regress on them. If the estimator has systematically mislabeled or merged topics, no amount of localized, post-hoc inspection can fix that. Indeed, researchers might simply confirm that the mislabeled topics have plausible top words, inadvertently lending credibility to a misleading measure.

This is especially problematic in the realm of *measurement inferences*, where topic proportions are used as *probabilistic estimates* of an underlying quantity of interest (e.g., ideological alignment, rhetorical framing). If the estimation routine is fundamentally inconsistent, the measured patterns can be a function of spurious artifacts that become more pronounced as the corpus grows.

7.2 Known Statistical Problems with Variational Inference

Given the widespread use of topic models, applied researchers might assume that the variational inference method such models rely upon has reliable statistical guarantees. As it turns out, 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, Sato, 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, Sato, 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, Vehtari, 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 illustrate 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 errors introduced by degenerately inconsistent topic model estimators, we use our correct topic model method with theoretical accuracy guarantees at scale, implemented in a convenient Python software (Kangaslahti, Ebanks, et al., 2025).

References

Abramowitz, Alan I. (2010). *The Disappearing Center: Engaged Citizens, Polarization, and American Democracy*. New Haven, CT: Yale University Press.

Aldrich, John and David Rohde (Jan. 1998). *Measuring Conditional Party Government*. Working Paper. Duke University. URL: https://www.researchgate.net/publication/251842516%5C_Measuring%5C_Conditional%5C_Party%5C_Government.

— (2001). "The logic of conditional party government". In: *Congress Reconsidered*. Ed. by LC Dodd and BI Oppenheimer. Washington, DC: CQ Press, pp. 269–92.

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.

Barbera, Pablo, Andreau Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost, and Joshua A. Tucker (2019). "Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data". In: *American Political Science Review* 113.4, pp. 883–901. DOI: 10.1017/S0003055419000352.

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.

Bendix, William (2016). "Bypassing Congressional Committees: Parties, Panel Rosters, and Deliberative Processes". In: *Legislative Studies Quarterly* 41.3, pp. 687–714. DOI: https://doi.org/10.1111/lsq.12125. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lsq.12125. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/lsq.12125.

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*.

Curry, James M. (2015). *Legislating in the Dark: Information and Power in the House of Representatives*. Chicago: The University of Chicago Press.

de Slegte, Jef, Filip Van Droogenbroeck, Bram Spruyt, Sam Verboven, and Vincent Ginis (July 2024). "The Use of Machine Learning Methods in Political Science: An In-Depth Literature Review". English. In: *Political Studies Review*. Publisher Copyright: l' The Author(s) 2024. ISSN: 1478-9299. DOI: 10.1177/14789299241265084.

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.
Gamm, Gerald and Steven Smith (2020). "The dynamics of party government in Congress". In: *Congress Reconsidered, 11th Edition*. Ed. by LC Dodd and BI Oppenheimer. Washington, DC: CQ Press. Chap. 7, pp. 197–224.
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, Margaret E. Roberts, and Brandon M. Stewart (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Presss.
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.
Harbridge, Laurel (2015). *Is Bipartisanship Dead? Policy Agreement and Agenda-Setting in the House of Representatives*. New York: Cambridge University Press.
Hetherington, Marc (Apr. 2009). "Review Article: Putting Polarization in Perspective". In: *British Journal of Political Science* 39, pp. 413–448. DOI: 10.1017/S0007123408000501.
Howard, Nicholas O. and Mark E. Owens (2020). "Circumventing Legislative Committees: The US Senate". In: *Legislative Studies Quarterly* 45.3, pp. 495–526. DOI: https://doi.org/10.1111/lsq.12269. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lsq.12269. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/lsq.12269.
Kangaslahti, Sara, Danny Ebanks, Jean Kossaifi, R. Michael Alvarez, and Anima Anandkumar (2025). "TensorLy-LDA: Analyzing Social Media Conversations at Scale with Online Tensor LDA". In: *Conditionally Accepted at Political Analysis*.
Koger, Gregory and Matthew J. Lebo (2020). *Strategic Party Government: Why Winning Trumps Ideology*. University of Chicago Press.
Mccarty, Nolan, Keith Poole, and Howard Rosenthal (Jan. 2006). *Polarized America: The Dance of Ideology and Unequal Riches*.
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.

Paisley, Laura (Nov. 2016). *Political Polarization at Its Worst Since the Civil War: Data Scientists Try to Explain the U.S. Government's Shifting Ideologies Over the Past Four Decades*. https://today.usc.edu/political-polarization-at-its-worst-since-the-civil-war-2/. Accessed: 2023-03-12.

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.

Tiefer, Charles (2016). *The Polarized Congress: The Post-Traditional Procedure of Its Current Struggles*. Lanham, MD: University Press of America.

Wallner, James I. (2013). *The Death of Deliberation: Partisanship and Polarization in the United States Senate*. New York: Lexington Books.

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.

Ying, Luwei, Jacob M. Montgomery, and Brandon M. Stewart (2022). "Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures". In: *Political Analysis* 30.4, pp. 570–589. DOI: 10.1017/pan.2021.33.

A Notation

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

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	$\mathbb{N}$
$\boldsymbol{\mu}$	$\mathbb{E}\left[f_i \mid \mathbf{h}\right] = \boldsymbol{\mu}\mathbf{h}$	$\mathbb{R}^V$
$\mathbf{f}_i$	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 $\mathbf{x}_i$	$\mathbb{R}^{n_b \times D}$
$\mathbf{\Phi}$	learned factors of the decomposition	$\mathbb{R}^{D \times K}$

A.1 Defining Statistical Consistency

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, Ebanks, et al. (2025) for notational consistency and transparency. We also provide the following technical definitions for the limit notation used throughout this paper. This notation is common in formal statistical and computer science settings, but less so in political science. We provide it here

Convergence in Probability If a sequence, X_n, is $o_p(a_n)$ then we say it converges in probability. That is, for all $\varepsilon > 0$,

$$\lim_{n \to \infty} \Pr\left[\frac{X_n}{a_n} \geq \varepsilon\right] = 0$$

Stochastic Bound If a sequence, X_n, is $O_p(a_n)$ then we say it is stochastically bounded. That is, for all $\varepsilon > 0$, there exist at least one $\delta > 0$ and $N > 0$ such that for $n \geq N$

$$\Pr\left[\frac{X_n}{a_n} > \delta\right] < \varepsilon$$

We then make use of the following notations to describe when a sequence is convergent in probability.:

1. A sequence that is $o_p(1)$ converges in probability.

2. If a sequence is $o_p(1)$ then it must be $O_p(1)$. That is, if it is convergent in probability, then it must be stochastically bounded.

3. The converse is false. That is, $O_p(1)$ does not imply $o_p(1)$. In English, a sequence can be stochastically bounded and not be convergent in probability.

4. If a sequence, X_n, is $O_p(a_n)$ and $a_n \to 0$, then X_n is $o_p(1)$, and thus convergences in probability. This is easily seen because we can apply the squeeze theorem to X_n by squeezing it between 0 and δa_n. Since δ is fixed and $a_n \to 0$, then $X_n \to 0$.

5. Suppose we have a sequence, X_n, that is $O_p(N)$. X_n is stochastically unbounded, as it grows linear in N. Thus it cannot be $o_p(1)$, which implies it diverges. This is easily seen as it is the contrapositive of (2).

We then define consistency and inconsistency for topic models in the following way: *Consistent.* Suppose we have a true parameter vector $\boldsymbol{\mu}^*$ and an estimate of that parameter vector, $\hat{\boldsymbol{\mu}}$. Then if $||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*|| = o(1)$, we say it is asymptotically consistent under the $l1-$norm. If $||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*||_2 = o(1)$, we say it is asymptotically consistent under the $l2$ norm. *Not Consistent.* Suppose we have a true parameter vector $\boldsymbol{\mu}^*$ and an estimate of that parameter vector, $\hat{\boldsymbol{\mu}}$. Then if $||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*|| = O(a_n)$ and $a_n \to \infty$, then we say it is *not* asymptotically consistent under the $l1$ norm. Further, if $||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*||_2 = O(a_n)$ and $a_n \to \infty$, then we say it is *not* asymptotically consistent under the $l2$ norm.

A.2 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, which we describe here. We note that other methods which rely on Variational Inference (VI) techniques, such as Structural Topic Models (STM), are exhibit the same statistical pathology we describe in this paper. Although STM methods include covariates, these covariates do not directly resolve the underlying problems introduced by VI, as STM estimates the analogous joint topic-document probabilities and word-topic probabilities as we describe below (Roberts, Stewart, et al., 2014)[8]. Under VI, these probabilities will still

[8] To see this, note that the posterior of STM and LDA both contain the multiplied topic-topic and topic-word matrices, which is the source of statistical pathologies for all LDA-like models. The inclusion of covariates provides more structure, but the underlying problem is not fully mitigated, as there is an implicit interdependence between topic-document and topic-word probabilities that implicates all LDA-like methods.

suffer from the same statistical identification problems as LDA. We study Gensim's LDA method as it is more analytically tractable, but these results extend to other topic models estimated via VI, of which STM is a more general 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 following *quantity of interest*,

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

where $\mathbf{h}$ is vector of multinomial distributions which describes the probability of seeing a topic given a document. Under topic models, these are mixtures that generated the documents in our data. 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

$$p(\boldsymbol{\mu}, \mathbf{h}, \mathbf{z}|\mathbf{f_i}, \alpha, \beta) = \frac{p(\boldsymbol{\mu}, \mathbf{h}, \mathbf{f_i}, \mathbf{z}|\alpha, \beta)}{p(\mathbf{f_i}|\alpha, \beta)} \tag{2}$$

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):

$$p(\mathbf{f_i}|\alpha, \beta) = \frac{\Gamma(\alpha_k)}{\prod_k \Gamma(\alpha_k)} \int \left(\prod_K \mu_i^{\alpha_i - 1} \right) \prod_{k=1}^{K} \sum_{i=1}^{V} \prod_{i=1}^{N} (\mu_k h_{k,i})^{f_i^n} \, d\mu \tag{3}$$

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 2, cancels out. The goal is to reduce the distance between the actual and approximated numerators. Unfortunately, this approximation introduces instabilities of its own.

A.3 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.[9]

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:

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

$$D_{\mathrm{KL}}(p \parallel q) = \sum_{x \in \mathcal{X}} p(\boldsymbol{\mu}, \mathbf{h}, \mathbf{f_i}, \mathbf{z} | \alpha, \beta) \, \log \left(\frac{p(\boldsymbol{\mu}, \mathbf{h}, \mathbf{f_i}, \mathbf{z} | \alpha, \beta)}{q(\mu, \mathbf{h}, \mathbf{z} | \gamma, \phi)} \right) \tag{4}$$

where ϕ and γ are called variational parameters. Said in more technical terms, because the normalizing constant in Equation A.2 is intractable, practitioners have utilized variational inference as a convenient alternative to estimating Latent Dirichlet Allocation (LDA) Blei, Ng, and Jordan (2003b).

We have the following result for estimating LDA via variational inference:

Theorem 1. *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 $O_p(N)$ under both the $l1$ and $l2$ norms.*

B Main Proof of Statistical Inconsistency

We can expand equation 4 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 $\boldsymbol{\mu}$

$$\begin{aligned}
\mathbf{z} &= \frac{\exp\left(\Psi(\gamma_{i,k}) + \sum_{v=1}^{V} f_{v,i} \left(\Psi(\phi_{v,k}) - \Psi\left(\sum_{k'=1}^{K} \phi_{v',k}\right)\right)\right)}{\sum_{k'=1}^{K} \exp\left(\Psi(\gamma_{i,k'}) + \sum_{v=1}^{V} f_{v,i} \left(\Psi(\phi_{v,k'}) - \Psi\left(\sum_{v'=1}^{V} \phi_{v',k'}\right)\right)\right)} && (5) \\
\boldsymbol{\gamma} &= \alpha + \sum_{v=1}^{V} \mathbf{z} && (6) \\
\boldsymbol{\phi} &= \beta + \sum_{i=1}^{N} \sum_{v=1}^{V} \mathbf{f_i} \mathbf{z} && (7)
\end{aligned}$$

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

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

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, Sato, et al., 2014 below.

Theorem 2. *From Nakajima, Sato, 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.*

We recapitulate their argument below, and then prove our extension.

First, Nakajima, Sato, et al., 2014 prove two lemmas the bound the limiting behavior of D_{KL}.

Lemma 1, Nakajima, Sato, et al., 2014:

Proof

First, we note the limiting behavior of the Gamma and Digamma functions:

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

We have,

$$Q = -\sum_{i=1}^{N} V \sum_{v=1}^{V} f_{i,n}\log\left(\sum_{k=1}^{K}\frac{\exp\left(\Psi\left(\phi_{i,k}\right)\right)}{\exp\left(\Psi\left(\sum_{k'=1}^{K}\phi_{i,k'}\right)\right)}\frac{\exp\left(\Psi\left(\gamma_{v,k}\right)\right)}{\exp\left(\Psi\left(\sum_{v'=1}^{V}\gamma_{v',k}\right)\right)}\right)$$

And we obtain, applying the properties of the Digamma function ψ, from above:

$$\underline{Q} \leq Q \leq \bar{Q}$$

where,

$$\begin{aligned}\underline{Q} &= -\sum_{i=1}^{N} V \sum_{v=1}^{V} f_{i,n}\log\left(\sum_{k=1}^{K}\frac{\phi_{i,k}}{\sum_{k'=1}^{K}\Theta_{i,k'}}\frac{\gamma_{l,k}}{\sum_{v'=1}^{L}\gamma_{v',k}}\frac{2\exp\left(-\frac{1}{2\phi_{i,k}}\right)}{\exp\left(-\frac{1}{\sum_{k'=1}^{K}\gamma_{i,k'}}\right)}\frac{\exp\left(-\frac{1}{2\gamma_{v,k}}\right)}{\exp\left(-\frac{1}{\sum_{v'=1}^{V}\gamma_{v',k}}\right)}\right)\\ \bar{Q} &= -\sum_{i=1}^{N} V \sum_{v=1}^{V} f_{i,n}\log\left(\sum_{k=1}^{K}\frac{\phi_{i,k}}{\sum_{k'=1}^{K}\phi_{i,k'}}\frac{\gamma_{l,k}}{\sum_{v'=1}^{L}\gamma_{v',k}}\frac{\exp\left(-\frac{1}{\phi_{i,k}}\right)}{\exp\left(-\frac{1}{2\sum_{k'=1}^{K}\gamma_{i,k'}}\right)}\frac{\exp\left(-\frac{1}{\gamma_{v,k}}\right)}{\exp\left(-\frac{1}{2\sum_{v'=1}^{V}\gamma_{v',k}}\right)}\right)\end{aligned}$$

Lemma 3 from Nakajima, Sato, et al., 2014.

$$
\begin{aligned}
\underline{R} = \quad & -\sum_{m=1}^{M} \log\left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right) - \sum_{k=1}^{K} \log\left(\frac{\Gamma(V\eta)^V}{\Gamma(\eta)}\right) - \frac{M(H-1)+H(L-1)}{2}\log(2\pi) \\
& + \sum_{m=1}^{M} \left\{ \left(K\alpha - \tfrac{1}{2}\right) \log \sum_{k=1}^{K} \phi_{i,k} - \sum_{k=1}^{K} \left(\alpha - \tfrac{1}{2}\right) \log \phi_{i,k} \right\} \\
& + \sum_{k=1}^{K} \left\{ \left(V\eta - \tfrac{1}{2}\right) \log \sum_{l=1}^{V} \gamma_{v,k} - \sum_{l=1}^{V} \left(\eta - \tfrac{1}{2}\right) \log \gamma_{v,k} \right\} \\
& + \sum_{m=1}^{M} \left\{ -\sum_{k=1}^{K} \frac{1}{12\phi_{i,k}} - \sum_{k=1}^{K} (\phi_{i,k} - \alpha) \left(\frac{1}{\phi_{i,k}} - \frac{1}{2\sum_{h'=1}^{K} \phi_{m,k'}} \right) \right\} \\
& + \sum_{k=1}^{K} \left\{ -\sum_{l=1}^{V} \frac{1}{12\gamma_{v,k}} - \sum_{l=1}^{V} (\gamma_{v,k} - \eta) \left(\frac{1}{\gamma_{v,k}} - \frac{1}{2\sum_{v'=1}^{V} \gamma_{v',k}} \right) \right\} \\
\bar{R} = \quad & -\sum_{m=1}^{M} \log\left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right) - \sum_{k=1}^{K} \log\left(\frac{\Gamma(V\eta)^V}{\Gamma(\eta)}\right) - \frac{M(H-1)+H(L-1)}{2}\log(2\pi) \\
& + \sum_{m=1}^{M} \left\{ \left(K\alpha - \tfrac{1}{2}\right) \log \sum_{k=1}^{K} \phi_{i,k} - \sum_{k=1}^{K} \left(\alpha - \tfrac{1}{2}\right) \log \phi_{i,k} \right\} \\
& + \sum_{k=1}^{K} \left\{ \left(V\eta - \tfrac{1}{2}\right) \log \sum_{l=1}^{V} \gamma_{v,k} - \sum_{l=1}^{V} \left(\eta - \tfrac{1}{2}\right) \log \gamma_{v,k} \right\} \\
& + \sum_{m=1}^{M} \left\{ \frac{1}{12\sum_{k=1}^{K} \phi_{i,k}} - \sum_{k=1}^{K} (\phi_{i,k} - \alpha) \left(\frac{1}{2\phi_{i,k}} - \frac{1}{\sum_{h'=1}^{K} \phi_{m,k'}} \right) \right\} \\
& + \sum_{k=1}^{K} \left\{ \frac{1}{12\sum_{l=1}^{V} \gamma_{v,k}} - \sum_{l=1}^{V} (\gamma_{v,k} - \eta) \left(\frac{1}{2\gamma_{v,k}} - \frac{1}{\sum_{v'=1}^{V} \gamma_{v',k}} \right) \right\}
\end{aligned}
$$

which, applying the limiting properties of Γ and Ψ, we have

$$
\begin{aligned}
R \quad = \quad & \sum_{m=1}^{M} \left\{ \left(K\alpha - \frac{1}{2}\right) \log \sum_{k=1}^{K} \phi_{i,k} - \sum_{k=1}^{K} \left(\alpha - \frac{1}{2}\right) \log \phi_{i,k} \right\} \\
& + \sum_{k=1}^{K} \left\{ \left(V\eta - \frac{1}{2}\right) \log \sum_{l=1}^{V} \gamma_{v,k} - \sum_{l=1}^{V} \left(\eta - \frac{1}{2}\right) \log \gamma_{v,k} \right\} + O_p(H(M+L))
\end{aligned}
$$

We can now combine the terms from Lemma 1 and Lemma 2 to write out the asymptotic form for D_{KL}:

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

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

We show:

Theorem 3. *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.

First, we recapitulate the proof framework from Nakajima, Sato, et al., 2014 which we use to prove our more general inconsistency result.

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,

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

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, Sato, et al., 2014. From here, we get the following form for $D_{\mathrm{KL}}(p \parallel q)$,

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

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.

C Numerical Example: Identifiability and Variational Inference Error

This section provides a concrete numerical example illustrating how variational inference in topic models can misalign topic assignments while preserving joint probability estimates, leading to potential statistical inconsistencies.

C.1 Ground-Truth Model

Consider a topic model where documents are generated from a true topic distribution matrix μ and a word distribution matrix h for each topic. These matrices define the joint probability matrix as:

$$P_{\text{true}} = \mu h.$$

For a corpus with three topics and three words, we assume the following ground-truth values:

$$\mu_{\text{true}} = \begin{bmatrix} 0.4 & 0.1 & 0.0 \\ 0.0 & 0.3 & 0.2 \end{bmatrix}$$

$$h_{\text{true}} = \begin{bmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The resulting joint probability matrix is computed as:

$$P_{\text{true}} = \mu_{\text{true}} h_{\text{true}} = \begin{bmatrix} 0.4 & 0.1 & 0.2 \\ 0.0 & 0.3 & 0.2 \end{bmatrix}$$

C.2 Estimated Model Under Variational Approximation

Now, suppose that due to variational inference constraints, the estimated model reallocates topics while still returning the same joint probabilities. Specifically, the estimated parameters $\hat{\mu}$ and $\hat{h}$ are permuted versions of the true matrices:

$$\mu_{\text{est}} = \begin{bmatrix} 0.1 & 0.4 & 0.0 \\ 0.3 & 0.0 & 0.2 \end{bmatrix}$$

$$h_{\text{est}} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0.5 \\ 0 & 0 & 1 \end{bmatrix}$$

Multiplying these matrices yields:

$$P_{\text{est}} = \mu_{\text{est}} h_{\text{est}} = \begin{bmatrix} 0.4 & 0.1 & 0.2 \\ 0.0 & 0.3 & 0.2 \end{bmatrix}$$

which is identical to P_{true}.

C.3 Implications for Statistical Inference

This example highlights a key issue: although the estimated model produces the same joint probability matrix, the underlying topic allocations are incorrect. This issue arises because variational approximations break dependencies between μ and h, leading to statistical inconsistency as dataset size grows. The inability to correctly recover μ and h has downstream effects on topic interpretability and generalization.

This phenomenon explains why even well-performing topic models may misrepresent underlying document structures, they optimize for likelihood but fail to uniquely recover interpretable latent topics.

D Technical Details for Anankumar, 2013

To demonstrate the degenerate inconsistency with an example, we employ a topic model that uses spectral decomposition with established accuracy *and* consistency guarantees (**kangetal2023**; **anandkumar2013spectral**).

Here, Anandkumar, Foster, et al., 2013 establish two critical theoretical properties for spectral decomposition of their topic model estimator. Both results require a mild identification assumption:

Identification Assumption: μ^* is full rank. This assumption requires that the columns of the topic-word probability matrix are not collinear. We need this assumption because we need to calculate a pseudo-inverse for μ to recover α. In practice, this merely means that no two topics are identical or near identical.

The first theorem establishes that we recover the intended parameters. First, we do not recover spurious topics (no false positives) that do not belong to the set of true topics, μ^*. Second, with probability 1, we recover all the topics in μ^*. And finally, we recover the mixing parameter that governs the Dirichlet distribution that generates the documents.

Theorem 4. *Parameter Recovery (Theorem 4.3 in Anandkumar, Foster, et al., 2013) We have that:*

- *(No False Positives) For all $\theta \in \mathbb{R}^k$, spectral decomposition for LDA returns a subset of the columns of μ^*.*
- *(Topic Recovery) Suppose $\theta \in \mathbb{R}^k$ is a random vector uniformly sampled over the sphere $\mathcal{S}^{k-1}$. With probability 1, spectral decomposition returns all columns of μ^*.*
- *(Parameter Recovery) We have that:*

$$\alpha = \alpha_0 \left(\alpha_0 + 1\right) \left(\mu^T \mu\right)^{-1} \mu^T \frac{1}{(\alpha_0+1)\alpha + 0} \mu \tilde{\alpha} \mu^T \left(\left(\mu^T \mu\right)^{-1} \mu^T\right)^\top \overrightarrow{1}$$

where $\overrightarrow{1} \in \mathbb{R}^k$ is a vector of all ones and $\tilde{\alpha} = diag(\alpha)$.

Second, Anandkumar, Foster, et al., 2013 establish a sample complexity-bound.

Theorem 5. *Theorem 5.1 from Anandkumar, Foster, et al., 2013 (Sample Complexity for Topic Models). Set a $\delta \in (0,1)$. Let $p_{\min} = \min_i \frac{\alpha_i}{\alpha_0}$ and let $\sigma_k(\mu)$ $\mathrm{de-}$ note the smallest (non-zero) singular value of $\boldsymbol{\mu}$. Suppose that we obtain $N \geq \left(\frac{(\alpha_0+1)(6+6\sqrt{\ln(3/\delta)})}{p_{\min}\sigma_k(\mu)^2}\right)^2$ independent samples of f_i, that is the documents. With proba-*

bility greater than $1 - \delta$, the following holds: for $\theta \in \mathbb{R}^k$ sampled uniformly sampled over the sphere $\mathcal{S}^{k-1}$, with probability greater than $3/4$, spectral decompositions returns a set of word-topic probability vectors $\{\hat{\mu}_1, \hat{\mu}_2, \ldots \hat{\mu}_k\}$ such that there exists a permutation σ of $\{1, 2, \ldots k\}$ (a permutation of the columns) so that for all $i \in \{1, 2, \ldots k\}$

$$\left\|\mu_i^* - \hat{\mu}_{\sigma(i)}\right\|_2 \leq c \frac{(\alpha_0 + 1)^2 k^3}{p_{\min}^2 \sigma_k(\mu)^3} \left(\frac{1 + \sqrt{\ln(1/\delta)}}{\sqrt{N}} \right)$$

where c is a universal constant.

We thus have

$$\left\|\mu_i^* - \hat{\mu}_{\sigma(i)}\right\|_2 = O\left(\frac{1}{\sqrt{N}}\right)$$

Then, because $\frac{1}{\sqrt{N}} \to 0$ as $N \to \infty$, we have,

$$\left\|\mu_i^* - \hat{\mu}_{\sigma(i)}\right\|_2 = o_p(1)$$

which implies we have a consistent estimator for the topic-word probability matrix under the $l2$-norm. To achieve consistency under the $l1$ norm, Anandkumar, Foster, et al., 2013 note in Remark 11 that we need one additional assumption: that a topic contains most of the probability mass on a finite number of words. In many use-cases in political science, where large vocabularies are common and topics are pinned down by 20 words (it is common to show the top 20 words i in a topic as a heursitic), this will be a reasonable assumption.

Now, with accurate estimates of μ^* and α_i in hand, *we can consistently estimate h^* via variational inference independently of μ.* The reason is because we are no longer estimating uh *jointly*, which is not identified because we are multiplying two sparse matrices together. We now have consistent and accurate estimates for all the troublesome parameters, and we can now apply standard methods.

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, Ebanks, et al., 2025), unlike many accurate methods that may take months or years to yield results – a significant obstacle for data-driven researchers.

We establish in that paper that our method is equivalent to the spectral method in Anandkumar, Foster, et al., 2013, endowing it with the same desirable statistical properties: parameter recovery and consistency under the $l2$ norm.

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.

Appendix: Additional Technical Details

E Selected Notation and Definitions

We briefly restate key notation. Let N be the number of documents, V the vocabulary size, K the number of topics. Each document i is associated with a topic mixture vector $h_i \in \mathbb{R}^K$, and each topic k with a word-distribution vector $\mu_k \in \mathbb{R}^V$. The quantity of interest is

$$\mathbb{E}[f_i \mid h_i] \;=\; \mu\, h_i,$$

where f_i is the vector of word counts in document i. For LDA, $h_i \sim \mathrm{Dirichlet}(\alpha)$ and $\mu \sim \mathrm{Dirichlet}(\beta)$ in a standard formulation. In practice, α and β may be known, estimated, or chosen via cross-validation.

E.1 Consistency vs. Inconsistency

An estimator $\hat{\mu}$ is consistent if $\|\hat{\mu} - \mu^*\| \to 0$ in probability, and inconsistent otherwise. For LDA with variational inference, we show that $\|\hat{\mu} - \mu^*\|$ can actually diverge with N, reflecting the degeneracies discussed in the main text. See **naka14** for related proofs under specific asymptotic regimes.

F Sketch of the Variational Inference Breakdown

The factorized variational distribution

$$q(\mu, h, z \mid \gamma, \phi) \;=\; q(z) \prod_i q(\mu \mid \phi_i)\, q(h \mid \gamma)$$

minimizes the Kullback–Leibler divergence $D_{\mathrm{KL}}(p\|q)$, but that objective has multiple global solutions in high dimensions due to permutations and duplications in μ. As N grows, these solutions become more numerous, leading to an increasing probability of settling on a misaligned $\hat{\mu}$.

G Spectral Method Guarantees

By contrast, the spectral method of Anandkumar, Foster, et al. (2013) recovers μ^* (up to permutation) by solving moment-based equations for co-occurrence and tri-occurrence of words. For large N, with high probability,

$$\|\hat{\mu} - \mu^*\|_2 = O\Big(\frac{1}{\sqrt{N}}\Big),$$

implying standard $1/\sqrt{N}$ consistency. Once μ^* is estimated, $\hat{h}_i$ can be recovered in a second step without the mutual degeneracy.

H Technical Proof of Degenerate Inconsistency for Variational LDA

Continuing the Proof to Show Degenerate Inconsistency

(Picking up precisely from where the above derivation of the asymptotic expression for $D_{\mathrm{KL}}(p\|q)$ leaves off, i.e., after obtaining)

$$D_{\mathrm{KL}}(p \parallel q) \;=\; \Big\{N\big(K\alpha-\tfrac{1}{2}\big) + K\big(V\beta-\tfrac{1}{2}\big) - \sum_{k=1}^{K}\Big[N^{(k)}\big(\alpha-\tfrac{1}{2}\big)+V^{(k)}\big(\beta-\tfrac{1}{2}\big)\Big]\Big\}\,\log W \,+\, (K-K^{*})\big(V\beta-\tfrac{1}{2}$$

We now show why this asymptotic form implies a degenerate *form of inconsistency under variational inference (VI).*

Step 1: Generalizing the $O_p(I\,W + V\,N)$ Term. A key piece of the expansion is the term $O_p(I\,W + V\,N)$, where:

- I is the number of entries in the matrix μh (topic-word *times* document-topic) for which the estimate diverges from the ground truth, $\mu^* h^*$,
- W is the (typical) document length, and
- V is the vocabulary size.

  Under naive consistency arguments (e.g., if the estimator were truly convergent), I would remain bounded as $N \to \infty$, or at least grow more slowly than N. However, as Nakajima, Sato, et al. (2014) and others have noted, *variational inference* in high-dimensional topic models can permit *spurious column duplication, merging, or re-scaling of topics*, increasing I *with* N. In other words, with more documents and more opportunities for small local improvements to the evidence lower bound (ELBO), the VI procedure can reassign words to partially duplicated topics (or incorrectly merge distinct topics) with only negligible costs in likelihood terms, thereby inflating the mismatch I.

  Step 2: Linking Growing I to Divergence from μ^*. If I increases sufficiently quickly with N, then the $O_p(I\,W + V\,N)$ contribution may dominate the entire expression as $N \to \infty$. This would drive the effective $D_{\mathrm{KL}}(p\|q)$ away from zero, indicating that the estimated parameters $(\hat{\mu}, \hat{h})$ *fail* to approximate the true posterior well as the sample grows. Concretely:

  $$I \;=\; I(N) \quad \text{may scale on the order of } N,$$

  so that $I\,W$ is effectively $O(N\,W)$ and can offset or overwhelm the main bracketed term that might otherwise shrink with better fits.

Step 3: Why This Implies *Degenerate Inconsistency*. Recall that an estimator $\hat{\mu}_N$ is *degenerately inconsistent* if its distance to the true parameter μ^* can grow unboundedly in probability as $N \to \infty$. Formally, there must exist a sequence $a_N \to \infty$ such that

$$\|\hat{\mu}_N - \mu^*\| \;=\; O_p\big(a_N\big).$$

Here, the mismatch $I(N)$ captures how many parameters or entries differ from their true values. If $I(N)$ is proportional to N (or grows faster), then $\|\hat{\mu}_N - \mu^*\|$ can blow up correspondingly. This growth is *not* constrained by the usual Dirichlet priors or the factorized nature of VI, because:

– *Variational factorization.* Breaking μ and h apart relaxes the global dependence that might otherwise penalize large misassignments.

– *Multiple local modes.* As N increases, the number of (near-)optimal local minima in the ELBO proliferates, some of which *duplicate or merge topic columns*. Consequently, the probability of landing in a solution *far* from μ^* *increases*.

Thus, the $O_p(I\,W + V\,N)$ term embodies precisely that phenomenon: if $I(N)$ scales with N, the distance can keep growing rather than converging.

Step 4: Concluding the Argument. Putting it all together:

1. We have expanded $D_{\mathrm{KL}}(p\|q)$ to show how its dominant terms split into a main bracket (involving N, $K\alpha - \frac{1}{2}$, $V\beta - \frac{1}{2}$, etc.) and a residual *mismatch* part, $O_p(I\,W + V\,N)$.
2. If the mismatch I remains $O(1)$ or grows very slowly, we might hope for partial consistency. But in reality, for large N, *variational inference* can systematically produce merges/splits that drive $I(N)$ in direct proportion to N.
3. Consequently, $\|\hat{\mu}_N - \mu^*\| \to \infty$ in probability (*or at least fails to shrink*), exactly matching the notion of *degenerate inconsistency*, the estimator does not just remain slightly biased, it *diverges*.

Therefore, even with V and W held fixed, once $N \to \infty$, the factorized approximation can allow unbounded overfitting in the topic columns (via duplication, merging, or other permutations), causing $\hat{\mu}_N$ to drift arbitrarily far from μ^*. This failure of $\hat{\mu}_N$ to *converge* under standard high-dimensional asymptotic assumptions confirms the degenerately inconsistent behavior of LDA with variational inference.

Summary of the Degenerate Inconsistency Result.

> *Because the $O_p(I\,W + V\,N)$ term in the asymptotic $D_{\mathrm{KL}}(p\|q)$ expansion can dominate (via merging or duplicating topics), the variational-inference estimator for LDA can systematically diverge as $N \to \infty$. Hence, in the sense of $\|\hat{\mu}_N - \mu^*\| \to \infty$ in probability, we obtain degenerate inconsistency.*

I Theories of Legislative Power

Although these theories offer diverse sets of predictions about the legislature, they all agree that increased polarization naturally enhances party-leader alignment, driven by the assumption that ideologically cohesive parties delegate greater authority to their leaders (Aldrich and Rohde, 2001). This notion aligns with the strategic party government theory articulated in Koger and Lebo (2020), who argue that polarization is actively chosen by parties to in the pursuit of electoral advantage. According to their framework, polarization should bolster the delegation of decision-making power to party leaders, thereby increasing alignment between leaders and their members.

However, our core findings fundamentally confound these conventional expectations. Using text-based methods that avoid the pitfalls associated with roll-call votes based measures, we find evidence for a substantial divergence between these polarization and party-leader alignment. This empirical reality directly challenges (Aldrich and Rohde, 2001) conditional government theory, which presupposes that internal ideological homogeneity and inter-party polarization inherently strengthen party leadership. Similarly, our results raise questions about the broad applicability of strategic party government, which fails to anticipate conditions under which polarization and leader-member alignment may substantially diverge.

Moreover, most research related to American legislative power relies on roll-call vote derived evidence, such as DW-Nominate scores to measure polarization (Aldrich and Rohde, 1998) and highlights top-down control by modern party leadership over legislative agendas (Gamm and Smith, 2020). Others emphasize the strong institutional powers available to contemporary party leaders, such as the ability to bypass committees (Bendix, 2016; Howard and Owens, 2020), directly negotiate policies Curry, 2015; Wallner, 2013, set legislative agendas (Harbridge, 2015), and control floor debates (Tiefer, 2016). Through the use of statistically consistent, text-based methods, we provide a more complete characterization than could be provided by roll-call votes alone. First, roll-call votes are highly constrained by party leaders' strategic decisions. And second, the two members may cast the same vote for very diametrically opposed reasons (such as a far-right and far-left member of congress voting against centrist legislation). Text-based evidence, such as we provide from 120 years of Congressional floor speeches, allows for new evidence based on the context of the rhetoric.

J Empirical Demonstration of Ill-Formed Topics Using Trade

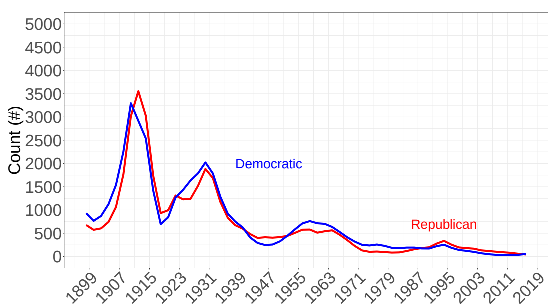
J.1 Empirical Evidence

We now show how inconsistency arises with a specific example of topic identification, in this case, trade. 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 topic modeling methods correct for the substantive errors introduced by the degenerate inconsistency plaguing existing topic modeling methods. We pick a technique with theoretical convergence guarantees (Anandkumar, Ge, et al., 2014), which we explain and provide detailed intuition in Section 5. The underlying assumption of this approach is mild – we need a topic-word matrix that is not co-linear. In Kangaslahti, Ebanks, et al. (2025), we, alongside our co-authors, implement a scalable version of this model, which we apply to uncover the topics consistently[10].

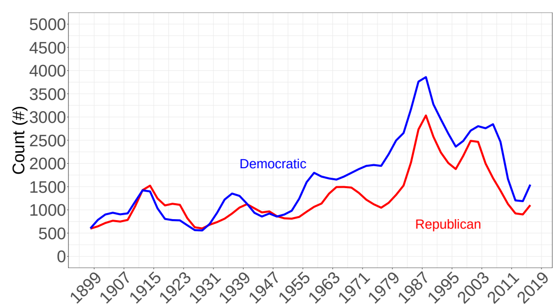
With an unsupervised, consistent method, we cannot know *a priori* the topics in the text, nor which words belong in which topic. In an applied setting, degenerate inconsistency is especially prevalent for datasets that span hundreds of years and cover changing contexts such as those we have study here. To fully illustrate the point, we show how word counts evolve for two potentially closely related words: tariff and trade. We show in Figure 7a how tariff was more frequently mentioned in the 1800s, whereas in Figure 7b, the word “trade" was more common after 1980. A consistent topic model will recognize that these words belong to the same topic despite the changing frequencies over time.

We can illustrate the source of the errors induced by inconsistent 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. A variational approach settles on (ii), whereas a consistent 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.

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



(a) Mentions of Tariff by Party



(b) Mentions of Trade by Party

Figure 7: Trade-Related Word Frequencies over Time