

Machine Learning for Textual and Unstructured Data

Lecture 2:

Stephen Hansen
University College London



FUNDACIÓN
RAMÓN ARECES



IESE

Business School
University of Navarra

Center for
International
Finance

Introduction

The document-term matrix is the foundation of much of text analysis in economics.

One important issue that the bag-of-words model ignores is the strong dependence structure among words.

In this lecture, we address ways of reducing the dimensionality of the document-term matrix while preserving the relevant heterogeneity across documents.

Focus on topic models which are factor models for discrete data.

Two Core NLP Problems

The problem of *synonymy* is that several different words can be share similar meanings. Cosine similarity between following documents?

| school | university | college | teacher | professor |
|--------|------------|---------|---------|-----------|
| 0 | 5 | 5 | 0 | 2 |
| school | university | college | teacher | professor |
| 10 | 0 | 0 | 4 | 0 |

The problem of *polysemy* is that the same word can have multiple meanings. Cosine similarity between following documents?

| tank | seal | frog | animal | navy | war |
|------|------|------|--------|------|-----|
| 10 | 10 | 3 | 2 | 0 | 0 |
| tank | seal | frog | animal | navy | war |
| 10 | 10 | 0 | 0 | 4 | 3 |

Latent Semantic Analysis

One of the first NLP models for finding low-dimensional structure in a corpus is Latent Semantic Analysis [Deerwester et al., 1990].

A linear algebra approach that applies a singular value decomposition to document-term matrix.

Closely related to classical principal components analysis.

Provides many foundational ideas that later models extend and refine.

Applications

Concept detection: [Boukus and Rosenberg, 2006] apply LSA to central bank communication documents, relate document representations to market responses.

Distance between documents:

1. [Iaria et al., 2018] apply LSA to scientific documents to measure overlap in research agendas across countries.
2. [Ter Ellen et al., 2021] apply LSA to financial newspapers to derive narrative monetary policy shock.

Statistical Models of Dimensionality Reduction

LSA has statistical foundations, but is not itself a statistical model.

Advantages of statistical models:

1. Make clear the statistical foundations for dimensionality reduction, allows for well-defined inference procedures.
2. Easier to interpret the latent components onto which data is projected.
3. Relatively straightforward to extend to incorporate additional dependencies of interest.

Disadvantage: require more elaborate inference algorithms.

Latent Dirichlet Allocation

Topics


Our latent variable models begin with the idea of topics, which are groups of words that express a similar theme.


Imagine K separate term distributions β_1, \dots, β_K , each of which represent a topic.


$\beta_{k,v}$ is the probability that term v appears in topic k .


Note that topic membership is not exclusive: same term can appear in multiple topics, with differing probabilities.

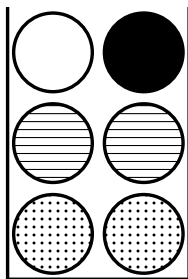
Topics as Urns

 = wage

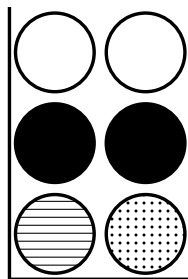
 = price

 = employ

 = increase

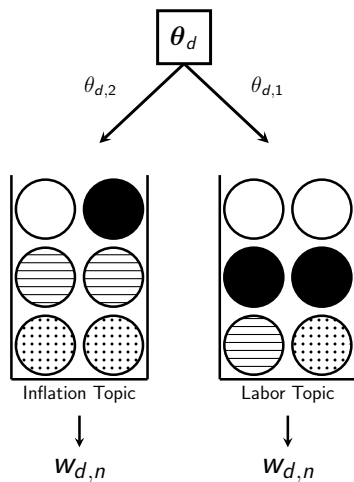


"Inflation" Topic



"Labor" Topic

Mixed-Membership Model for Document



Inference for Mixed-Membership Model

Under mixed-membership model, $\mathbf{x}_d \sim \text{Multinomial}(\sum_k \theta_{d,k} \boldsymbol{\beta}_k, N_d)$.

Likelihood function is $\prod_d \prod_v (\sum_k \theta_{d,k} \beta_{k,v})^{x_{d,v}}$.

Maximum likelihood solution closely related to the problem of finding a *non-negative matrix factorization* of the form $\mathbf{X}' \approx \Theta B$
[Ding et al., 2006]

1. Rows of \mathbf{X}' are \mathbf{x}_d / N_d .
2. Θ is $D \times K$ row-stochastic matrix.
3. B is $K \times V$ row-stochastic matrix.

[Ke et al., 2021] point out that many such matrix factorizations exist, so MLE estimates are not unique.

Latent Dirichlet Allocation

[Blei et al., 2003] adds Dirichlet prior distributions to the multinomial probability vectors:

1. $\theta_d \sim \text{Dirichlet}(\alpha)$.
2. $\beta_k \sim \text{Dirichlet}(\eta)$.

Symmetric priors for simplicity, can be relaxed as in original paper.

Bayesian approach can be motivated in terms of regularization, and also to overcome weak identification.

LDA is the most popular probabilistic topic model for text, also influential in other domains (e.g. population genetics).

Essentially a Bayesian factor model for discrete data.

Example statement: Yellen, March 2006, #51

Raw Data → Remove Stop Words → Stemming → Multi-word tokens = Bag of Words

We have noticed a change in the relationship between the core CPI and the chained core CPI, which suggested to us that maybe something is going on relating to substitution bias at the upper level of the index. You focused on the nonmarket component of the PCE, and I wondered if something unusual might be happening with the core CPI relative to other measures.

Example statement: Yellen, March 2006, #51

Raw Data → Remove Stop Words → Stemming → Multi-word tokens =
Bag of Words

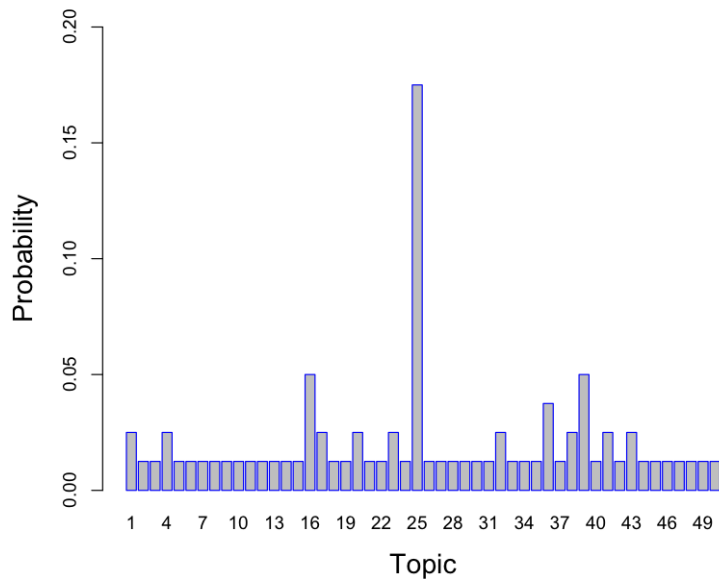
noticed change relationship between core CPI
chained core CPI suggested maybe something
going relating substitution bias upper level index
focused nonmarket component PCE wondered
something unusual happening core CPI relative
measures

Example statement: Yellen, March 2006, #51

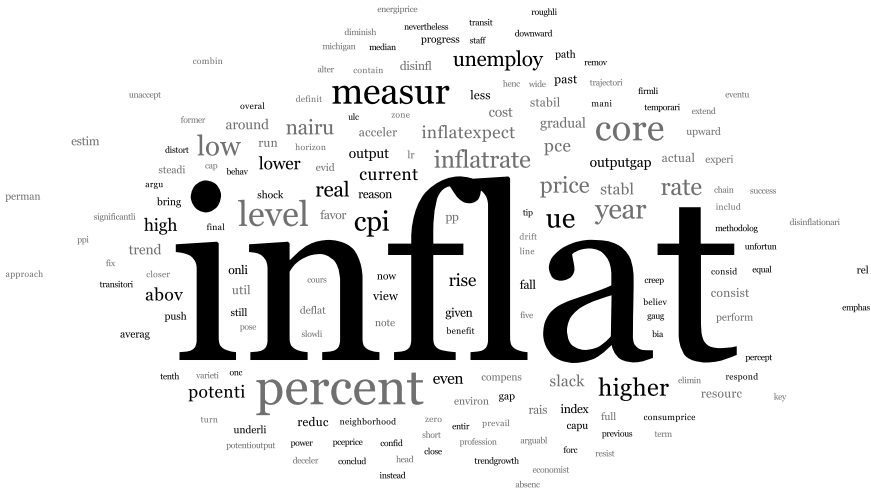
Raw Data → Remove Stop Words → Stemming → Multi-word tokens =
Bag of Words

notic chang relationship between core CPI
chain core CPI suggest mayb someth
go relat substitut bia upper level index
focus nonmarket compon PCE wonder
someth unusu happen core CPI rel
measur

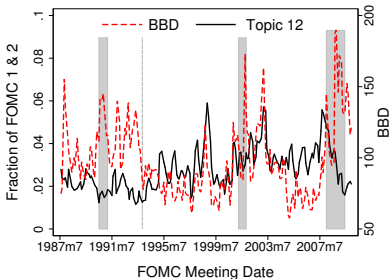
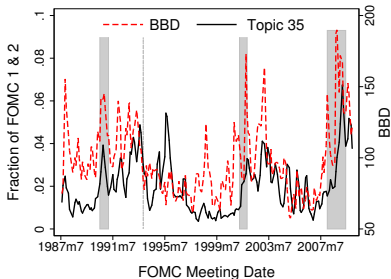
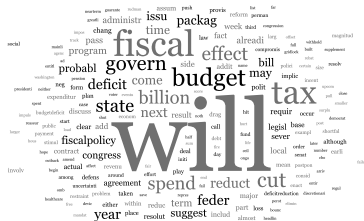
Distribution of Attention



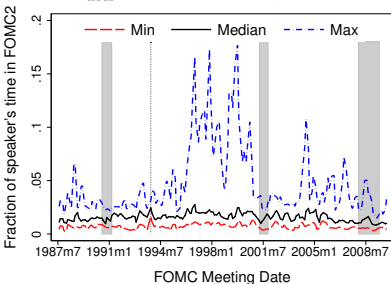
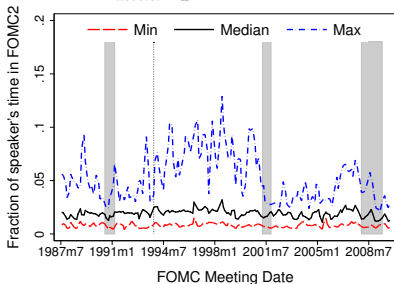
Topic 25



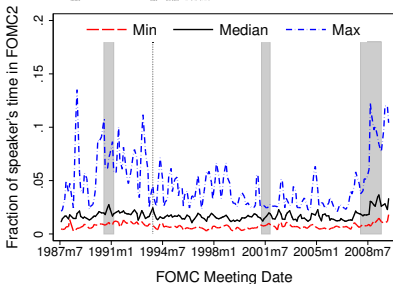
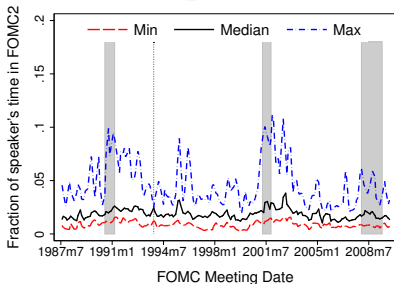
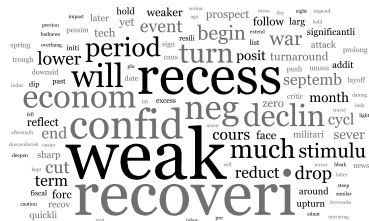
External Validation—BBD



Pro-Cyclical Topics



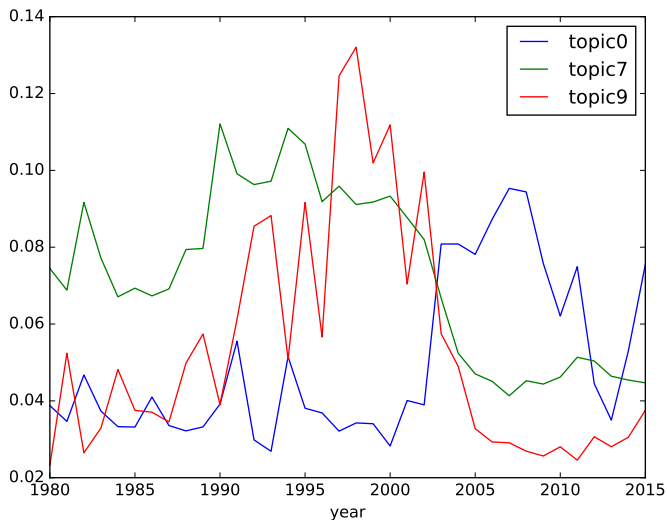
Counter-Cyclical Topics



Topics on NYT Data (Iraq, Iran, Syria from mid-1980s)

| Topic | Top Terms |
|-------|---|
| 0 | american.forc.militari.troop.command.iraqi.gener.armi.iraq.offic |
| 2 | shiit.mr.govern.sunni.polit.parti.leader.iraqi.elect.minist |
| 3 | iranian.attack.air.iraqi.gulf.report.today.missil.forc.fire |
| 4 | iran.iranian.islam.ayatollah.presid.leader.teheran.govern.polit.revolut |
| 6 | iran.nuclear.iranian.program.sanction.negoti.enrich.agenc.uranium.deal |
| 7 | iraq.iraqi.hussein.baghdad.war.saddam.kuwait.nation.today.countri |
| 8 | govern.compani.bank.state.money.work.million.billion.project.contract |
| 9 | weapon.intellig.report.use.inspector.chemic.nation.site.program.offici |
| 10 | syria.israel.syrian.arab.isra.mr.lebanon.assad.saudi.presid |
| 11 | oil.percent.year.price.countri.export.million.econom.day.trade |
| 13 | kill.american.attack.baghdad.bomb.iraqi.polic.offici.al.insurg |
| 14 | unit.nation.council.secur.mr.resolut.diplomat.meet.foreign.franc |
| 16 | mr.report.prison.releas.charg.case.court.arrest.accus.investig |
| 18 | govern.syria.group.kurdish.syrian.turkey.forc.opposit.border.rebel |

Distribution of Topics in Iraq Articles



Posterior Inference

Exact posterior inference in LDA is intractable so one must use approximation methods.

The conjugacy of the Dirichlet to the multinomial makes deriving a Gibbs sampler relatively straightforward [Griffiths and Steyvers, 2004].

Original paper instead use variational inference algorithm. Search within a simplified set of candidate posterior distributions for the element closest to the true posterior.

Modern automatic inference methods substantially simplify posterior inference.

Code for Hamiltonian Monte Carlo for LDA

```
1 def lda(X, K, alpha, eta):
2     # X: document-word matrix of BoWs
3     # K: number of topics
4     # alpha: Dirichlet hyperparameter for topic prevalence
5     # eta: Dirichlet hyperparameter for topic concentration
6
7     D, V = jnp.shape(X)
8     N = X.sum(axis = 1)
9
10    # document-topic distributions
11    with plate("docs", D):
12        theta = sample("theta", dist.Dirichlet(alpha*jnp.ones([K])))
13
14    # topic-word distributions
15    with plate("topics", K):
16        beta = sample("beta", dist.Dirichlet((eta) * jnp.ones([V])))
17
18    # likelihood
19    distMultinomial = dist.Multinomial(total_count = N,
20        probs = jnp.matmul(theta, beta))
21    with plate("hist", D):
22        sample("obs", distMultinomial, obs = X)
```


Applications

Topic Models in Empirical Economics

Economics and finance papers that use topic models typically follow a two-step approach:

1. LDA generates measures upstream.
2. Output is plugged into downstream econometric models.

For example, [Hansen et al., 2018] uses the similarity of FOMC members' topic coverage to proxy herding and studies its evolution in a DiD model.

Text and metadata hardly ever modeled jointly, although in principle they can (and should?) be [Blei and Lafferty, 2006], [Roberts et al., 2014], [Sacher et al., 2021].

Here we illustrate application in conflict forecasting, see also [Larsen and Thorsrud, 2019], [Thorsrud, 2020], [Bybee et al., 2021].

Predicting Conflict

[Mueller and Rauh, 2018] use media articles to predict conflict.

Corpus consists of 700,000 articles from 1975-2015 and covering 185 countries. Source: Economist, NYT, WP; accessed via LexisNexis.

Conflict indicator derived from Uppsala Conflict Data Program is 1 if > 25 battle-related deaths in the country in year t .

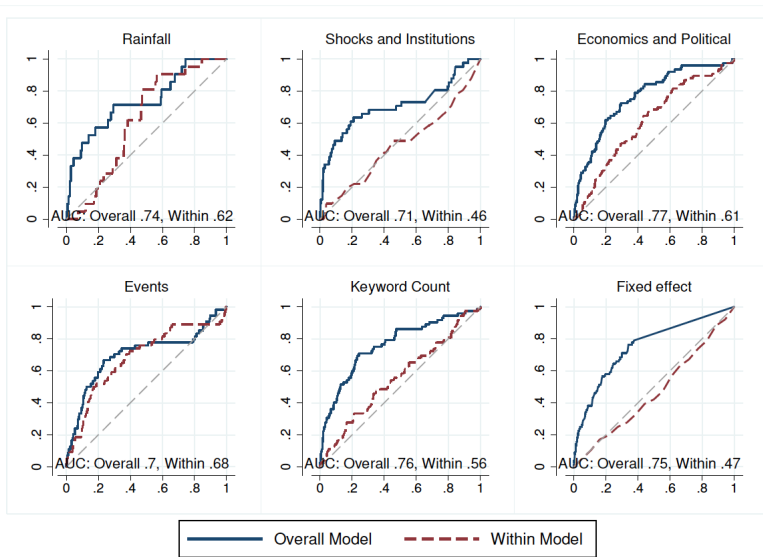
The existing literature fits models of the form $y_{i,t} = \alpha_i + \mathbf{x}_{i,t-1}^T \boldsymbol{\beta} + \varepsilon_{i,t-1}$ where \mathbf{x}_{it} include variables like institutions; income shocks; etc. Sample is $t = 1, \dots, T$.

Fitted values used to form estimate $\hat{y}_{T+1} = \hat{\alpha}_i + \mathbf{x}_{i,t}^T \hat{\boldsymbol{\beta}}$.

Paper points out that nearly all of the predictive power in such models comes from the country fixed effects.

ROC Curves for Standard Models

(b) Armed Conflict



Topics as Covariates

As an alternative forecasting model, the paper estimates models

$$y_{i,t} = \alpha_i + \boldsymbol{\theta}_{i,t-1}^T \boldsymbol{\beta} + \varepsilon_{i,t} \quad (1)$$

$$y_{i,t} = \alpha + \boldsymbol{\theta}_{i,t-1}^T \boldsymbol{\beta} + \epsilon_{i,t} \quad (2)$$

The 'within' model produces nearly as good forecasts as the 'overall' model.

The lesson is that there is substantial within-country variation in (English-language) media coverage correlated with the onset of conflict.

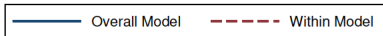
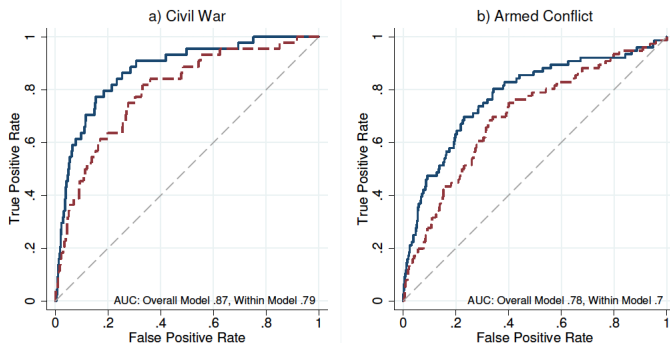
Moreover, the forecasting performance of the within model is better for predicting conflict in countries where conflict has not occurred recently.

When $T=2010$, Yemen is one of the countries predicted to be most likely to enter conflict in 2011 according to (2) but not according to (1)

(2) also puts much higher probability on onset of conflict in Syria and Libya in 2011.

ROC Curves for Media Models

Figure 7: ROC Curves for Onset (Only Non-Conflict Topics)



Survey Data

Overview

Survey data is arguably neither fundamentally unstructured nor happenstance (e.g. Survey of Professional Forecasters).

Often summarized in terms of headline numbers or averages, which ignores potentially rich underlying heterogeneity and important elements of the data structure.

Many surveys generate categorical data if they are structured as a sequence of multiple choice questions.

The latent variable models we introduced for text are also useful for capturing unobserved heterogeneity in such data.

Why Latent Variable Models?

The motivation for recovering low-dimensional structure in text is that there are fewer semantic dimensions than vocabulary terms.

The motivation in survey data is that there exist unobserved types in the population that generate correlation patterns across questions:

1. If pessimistic about the economy, more likely to believe 'stock market value lower next year' and 'business investment is falling'.
2. If socially conservative, more likely to believe 'abortion is wrong' and 'religion is important in public life'.
3. If a firm well managed, more likely to 'conduct performance reviews' and 'have inventory management system'.

Type-Specific Distributions

Suppose there are J survey questions in total.

Question j has L_j possible responses, encoded as $\mathcal{L}_j = \{1, \dots, L_j\}$.

Responses need not have ordinal interpretation nor be comparable across questions, but important that there be a discrete number.

Suppose there are K separate response profiles.

Let $\beta_{k,j} \in \Delta^{L_j-1}$ be the distribution over question j responses induced by type k , i.e. $\beta_{k,j,r}$ is the probability of observing the r th response to question j when type is k .

Important assumption is that responses are independent across question conditional on type.

Prior distribution on $\beta_{k,j} \sim \text{Dirichlet}(\eta)$.

Modeling Individual Heterogeneity

Suppose we observe N separate survey respondents.

Let $x_{i,j} \in \mathcal{X}_j$ be the response of individual i to question j .

Let $\theta_i \sim \text{Dirichlet}(\alpha)$ represent distribution of person i across latent types, where $\theta_{i,k}$ represents i 's association with type k .

$$x_{i,j} \sim \text{Multinomial}(\sum_k \theta_{i,k} \beta_{k,j}, 1)$$

Likelihood function is

$$\prod_i \prod_j \sum_k \theta_{i,k} \beta_{k,j, x_{ij}}$$

Inference issues same as in LDA.

Known as Bayesian Grade-of-Membership Model [Erosheva et al., 2007].

Application to Election Survey

[Gross and Manrique-Vallier, 2014] apply Bayesian GoM to the American National Election Study conducted on Election Day 1982.

19 separate questions regarding political beliefs and values related to *equal opportunity*, *economic individualism*, and *free enterprise*.

Responses coded as 'agree', 'can't decide', 'disagree'.

$K = 3$, but two types dominate responses roughly corresponding to the conservative-liberal distinction.

| j | Question | Level: $l = 1$ (Agree) | | $l = 3$ (Disagree) | |
|-----|---|------------------------|-------------|--------------------|-------------|
| | | $k = 1$ | $k = 2$ | $k = 1$ | $k = 2$ |
| 1 | <i>Equal treatment</i> | 0.61 (0.10) | 0.92 (0.05) | 0.37 (0.10) | 0.07 (0.05) |
| 2 | <i>Equality goal misguided</i> | 0.27 (0.05) | 0.14 (0.06) | 0.7 (0.05) | 0.83 (0.06) |
| 3 | <i>Equal opportunity society's responsibility</i> | 0.82 (0.05) | 0.89 (0.05) | 0.17 (0.05) | 0.10 (0.05) |
| 4 | <i>Natural inequality 1</i> | 0.87 (0.04) | 0.76 (0.07) | 0.12 (0.04) | 0.22 (0.07) |
| 5 | <i>Natural inequality 2</i> | 0.95 (0.02) | 0.85 (0.06) | 0.05 (0.02) | 0.14 (0.05) |
| 6 | <i>Democracy</i> | 0.86 (0.04) | 0.94 (0.04) | 0.14 (0.04) | 0.05 (0.04) |
| 7 | <i>Inequality big problem</i> | 0.30 (0.14) | 0.88 (0.07) | 0.69 (0.14) | 0.10 (0.07) |
| 8 | <i>Hard work optimism</i> | 0.97 (0.02) | 0.45 (0.17) | 0.02 (0.02) | 0.54 (0.17) |
| 9 | <i>Hard work realism</i> | 0.12 (0.06) | 0.47 (0.09) | 0.87 (0.06) | 0.52 (0.09) |
| 10 | <i>Individual responsibility for failure</i> | 0.77 (0.06) | 0.19 (0.12) | 0.22 (0.06) | 0.77 (0.12) |
| 11 | <i>Ambition pessimism</i> | 0.76 (0.05) | 0.88 (0.05) | 0.23 (0.05) | 0.11 (0.05) |
| 12 | <i>Hard work idealism</i> | 0.64 (0.06) | 0.21 (0.12) | 0.35 (0.06) | 0.78 (0.11) |
| 13 | <i>Effort pessimism</i> | 0.75 (0.07) | 0.95 (0.03) | 0.25 (0.07) | 0.04 (0.03) |
| 14 | <i>Less intervention is better</i> | 0.81 (0.05) | 0.42 (0.13) | 0.17 (0.05) | 0.55 (0.13) |
| 15 | <i>Intervention populism</i> | 0.62 (0.06) | 0.83 (0.06) | 0.36 (0.05) | 0.11 (0.06) |
| 16 | <i>Laissez-faire capitalism</i> | 0.36 (0.05) | 0.07 (0.07) | 0.63 (0.05) | 0.91 (0.07) |
| 17 | <i>Regulations not a threat to freedom</i> | 0.33 (0.05) | 0.49 (0.08) | 0.66 (0.05) | 0.49 (0.08) |
| 18 | <i>Intervention causes problems</i> | 0.94 (0.04) | 0.58 (0.13) | 0.05 (0.03) | 0.40 (0.13) |
| 19 | <i>Free enterprise not intrinsic feature of gov't</i> | 0.12 (0.07) | 0.41 (0.08) | 0.87 (0.07) | 0.58 (0.08) |

References I

Blei, D. M. and Lafferty, J. D. (2006).

Dynamic topic models.

In Proceedings of the 23rd International Conference on Machine Learning, ICML '06, pages 113–120, New York, NY, USA. Association for Computing Machinery.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).

Latent dirichlet allocation.

The Journal of Machine Learning Research, 3(null):993–1022.

Boukus, E. and Rosenberg, J. V. (2006).

The Information Content of FOMC Minutes.

Bybee, L., Kelly, B. T., Manela, A., and Xiu, D. (2021).

Business News and Business Cycles.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).

Indexing by latent semantic analysis.

Journal of the American Society for Information Science, 41(6):391–407.

References II

Ding, C., Li, T., and Peng, W. (2006).

Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence, chi-square statistic, and a hybrid method.

In Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1, AAAI'06, pages 342–347, Boston, Massachusetts. AAAI Press.

Erosheva, E. A., Fienberg, S. E., and Joutard, C. (2007).

Describing disability through individual-level mixture models for multivariate binary data.

The Annals of Applied Statistics, 1(2).

Griffiths, T. L. and Steyvers, M. (2004).

Finding scientific topics.

Proceedings of the National Academy of Sciences, 101(suppl 1):5228–5235.

Gross, J. H. and Manrique-Vallier, D. (2014).

A mixed membership approach to the assessment of political ideology from survey responses.

In Airolidi, E. M., Blei, D., Erosheva, E. A., and Fienberg, S. E., editors, Handbook of Mixed Membership Models and Its Applications. CRC Press.

References III

Hansen, S., McMahon, M., and Prat, A. (2018).

Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach.

[The Quarterly Journal of Economics](#), 133(2):801–870.

Iaria, A., Schwarz, C., and Waldinger, F. (2018).

Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science.

[The Quarterly Journal of Economics](#), 133(2):927–991.

Ke, S., Olea, J. L. M., and Nesbit, J. (2021).

Robust Machine Learning Algorithms for Text Analysis.

Unpublished manuscript.

Larsen, V. H. and Thorsrud, L. A. (2019).

The value of news for economic developments.

[Journal of Econometrics](#), 210(1):203–218.

Mueller, H. and Rauh, C. (2018).

Reading Between the Lines: Prediction of Political Violence Using Newspaper Text.

[American Political Science Review](#), 112(2):358–375.

References IV

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014).

Structural Topic Models for Open-Ended Survey Responses.

[American Journal of Political Science](#), 58(4):1064–1082.

Sacher, S., Battaglia, L., and Hansen, S. (2021).

Hamiltonian Monte Carlo for Regression with High-Dimensional Categorical Data.

[arXiv:2107.08112 \[econ, stat\]](#).

Ter Ellen, S., Larsen, V. H., and Thorsrud, L. A. (2021).

Narrative Monetary Policy Surprises and the Media.

[Journal of Money, Credit and Banking](#).

Thorsrud, L. A. (2020).

Words are the New Numbers: A Newsy Coincident Index of the Business Cycle.

[Journal of Business & Economic Statistics](#), 38(2):393–409.