Statistical Machine Learning for Large and Unstructured Data

Topic Models

Stephen Hansen
University College London

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

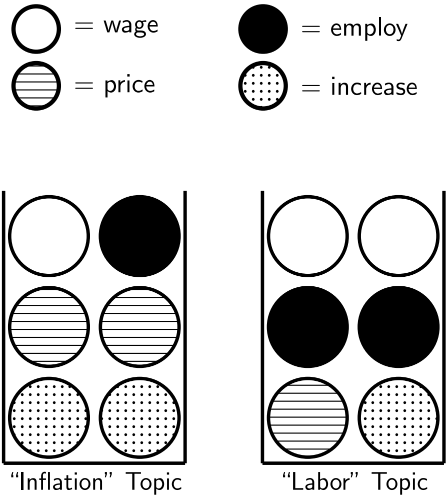
Our latent variable models begin with the idea of topics, which are groups of words that express a similar theme.

Imagine K separate term distributions $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$, each of which represent a topic.

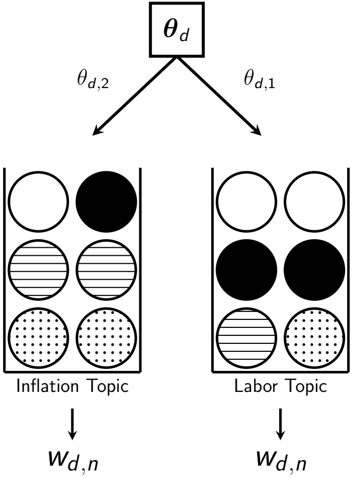
$\beta_{k,v}$ is the probability that term v appears in topic k.

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

Topics as Urns



Mixed-Membership Model for Document



Inference for Mixed-Membership Model

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

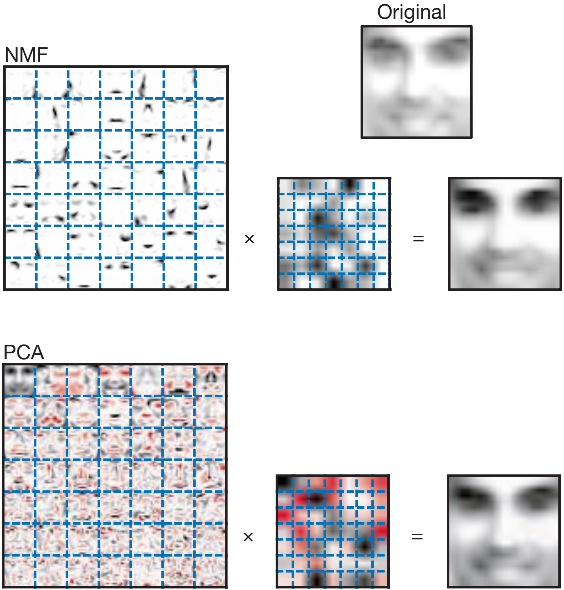
Likelihood function is $\prod_d \prod_v (\sum_k \theta_{d,k} \beta_{k,v})^{x_{d,v}}$.

Model known as probabilistic LSA [Hofmann, 1999].

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.

NMF for Image Data [Lee and Seung, 1999]



NMF is not Unique [Ke et al., 2021]

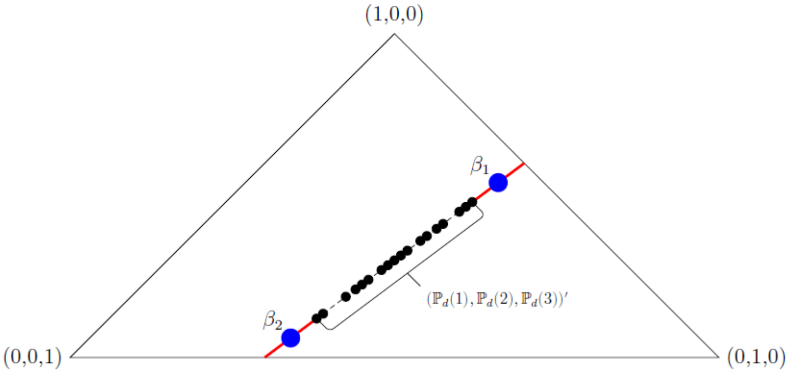


Figure 2: Lack of identification when $K = 2$, $V = 3$, and D is large. The small black circles are the document-specific term probabilities—the columns of P. The dotted line is the 2-simplex. The large blue circles represent one of the possible topic distributions B. The solid red line is the set of all possible topic distributions.

Latent Dirichlet Allocation

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

1. $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha)$.
2. $\boldsymbol{\beta}_k \sim \text{Dirichlet}(\eta)$.

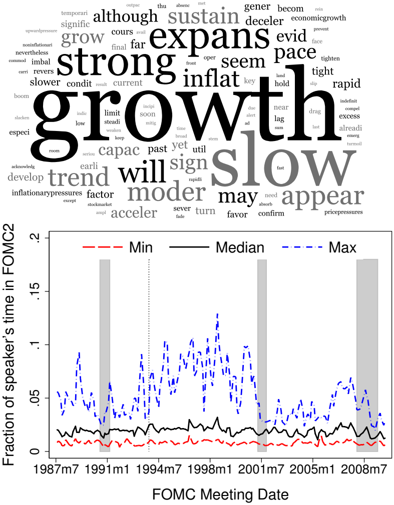
Symmetric priors for simplicity, can be relaxed as in original paper.

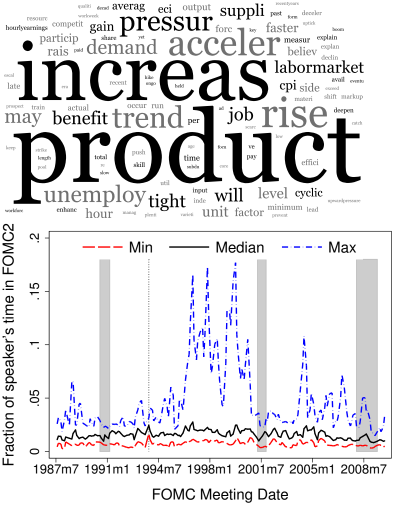
Bayesian approach can be motivated in terms of regularization, and also to overcome weak identification.

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

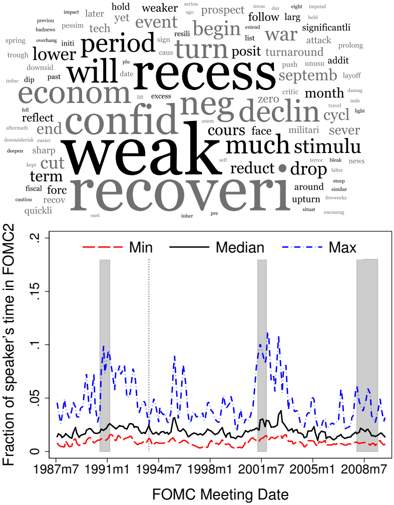
Essentially a Bayesian factor model for discrete data.

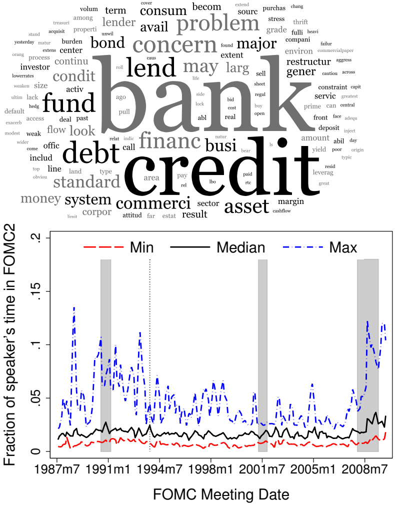
Pro-Cyclical Topics





Counter-Cyclical Topics

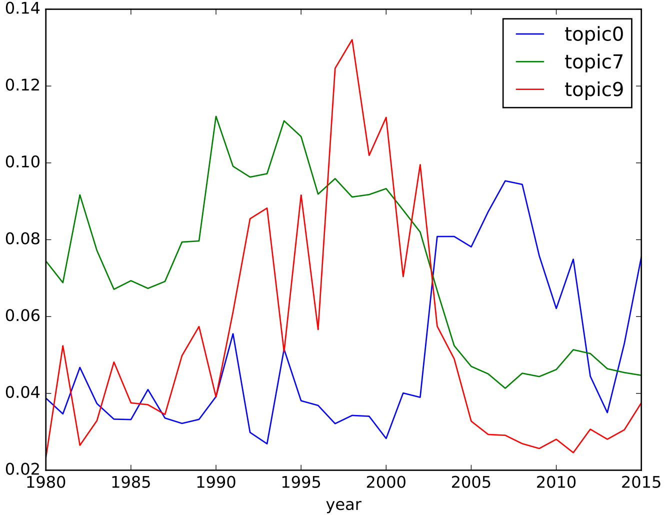




Topics on NYT Data (Iraq, Iran, Syria from mid-1980s)

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.

Model Selection

Traditional way of selecting K: out-of-sample goodness-of-fit, information criteria.

[Chang et al., 2009] propose an objective way of determining whether topics are interpretable.

Two tests:

1. *Word intrusion*. Form set of words consisting of top five words from topic k + word with low probability in topic k. Ask subjects to identify inserted word.
2. *Topic intrusion*. Show subjects a snippet of a document + top three topics associated to it + randomly drawn other topic. Ask to identify inserted topic.

Estimate LDA and other topic models on NYT and Wikipedia articles for $K = 50, 100, 150$.

Results

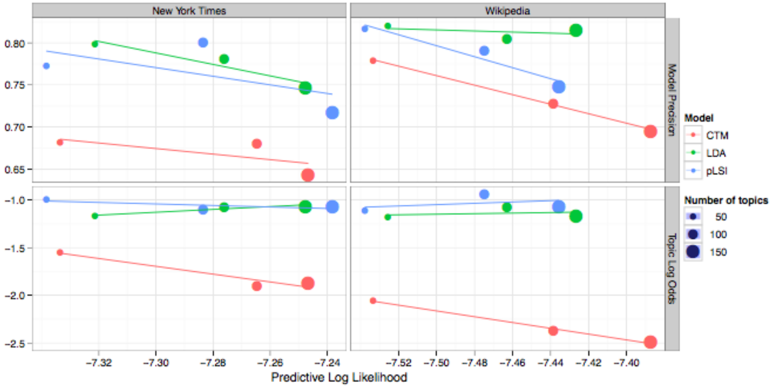


Figure 5: A scatter plot of model precision (top row) and topic log odds (bottom row) vs. predictive log likelihood. Each point is colored by model and sized according to the number of topics used to fit the model. Each model is accompanied by a regression line. Increasing likelihood does not increase the agreement between human subjects and the model for either task (as shown by the downward-sloping regression lines).

Takeaway

Topics seem objectively interpretable in many contexts.

Tradeoff between goodness-of-fit and interpretablility, which is generally more important in social science.

[Newman et al., 2010] propose a method based on mutual pointwise information between top words in topics as computed via co-occurrence in Wikipedia.

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.

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.

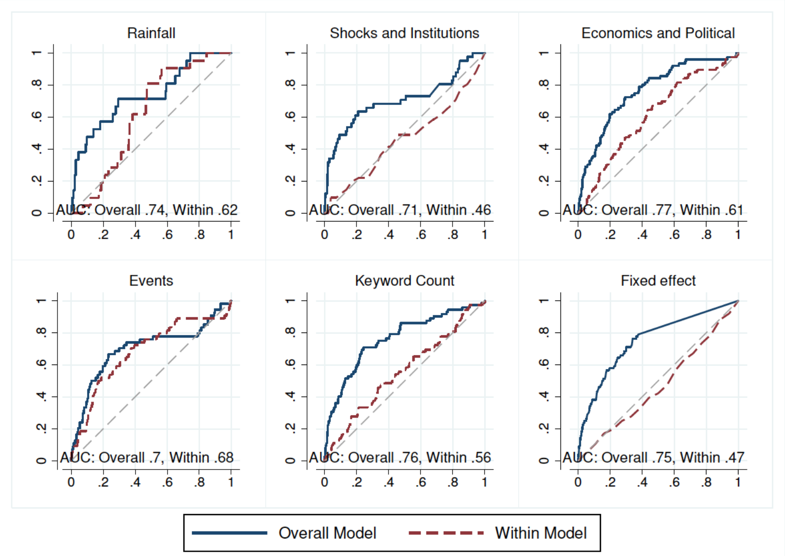
The existing literature fits models of the form $y_{i,t} = \alpha_i + \mathbf{x}_{i,t-1}^T \boldsymbol{\beta} + \varepsilon_{i,t-1}$ where $\mathbf{x}_{it}$ include variables like institutions; income shocks; etc. Sample is $t = 1, \ldots, T$.

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

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

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

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.

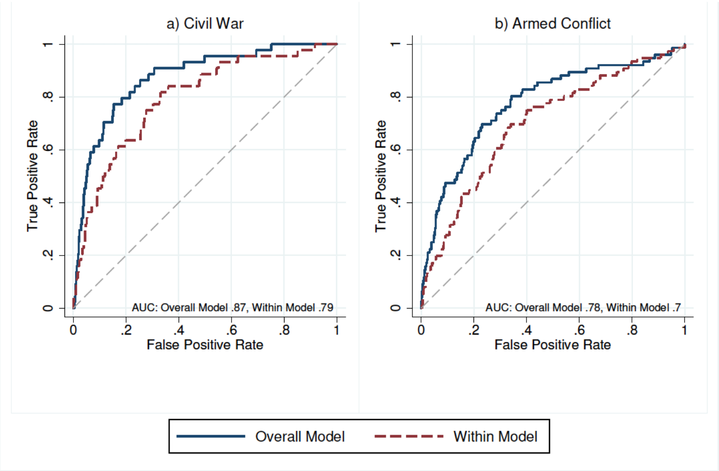
Moreover, the forecasting performance of the within model is better for predicting conflict in countries where conflict has not occurred recently.

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 $\mathfrak{L}_j = \{1, \dots, L_j\}$.

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

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

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

Modeling Individual Heterogeneity

Suppose we observe N separate survey respondents.

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

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

Executive Time Use Project

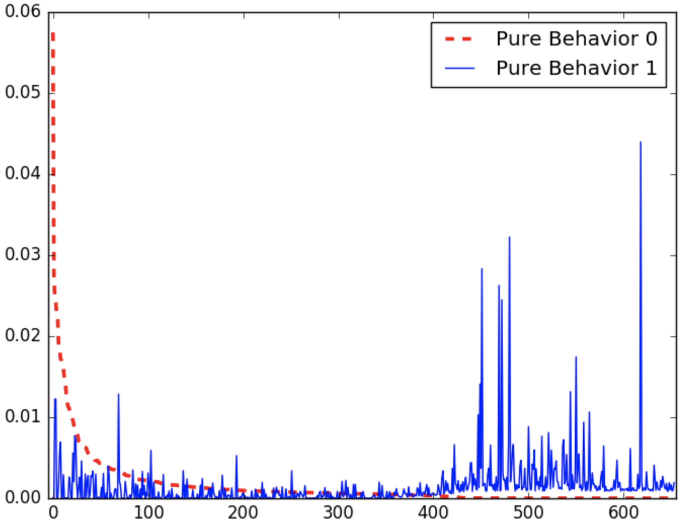
Data on each 15-minute block of time for one week of 1,114 CEOs' time classified according to

1. type (e.g. meeting, public event, etc.)
2. duration (15m, 30m, etc.)
3. planning (planned or unplanned)
4. number of participants (one, more than one)
5. functions of participants, divided between employees of the firms or "insiders" (finance, marketing, etc.) and "outsiders" (clients, banks, etc.).

There are 4,253 unique combinations of these five features in the data.

One can summarize the data with a 1114×4253 matrix where the (i, j)th element is the number of 15-minute time blocks that CEO i spends in activities with a particular combination of features j.

Pure Behaviors are Sharply Distinct



Differences across Pure Behaviors

Feature	X times less likely in Behavior 1	Feature	X times more likely in Behavior 1
Plant Visits	0.11	Communications	1.9
Just Outsiders	0.5	Outsiders + Insiders	1.9
Production	0.5	C-suite	34
Suppliers	0.3	Multifunction	1.5

Managers vs. Leaders

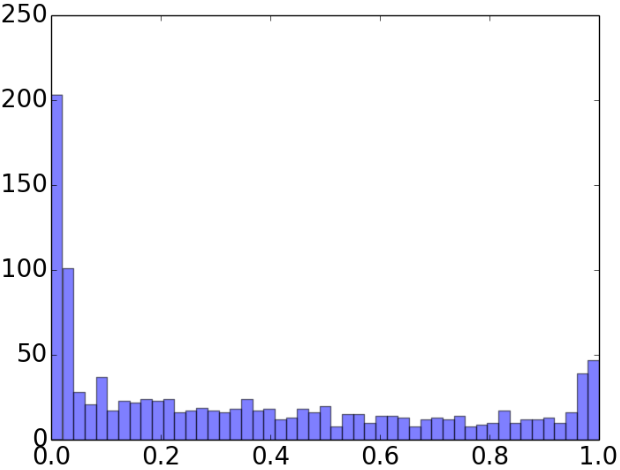
Kotter (1999) emphasizes a behavioral distinction between "management" and "leadership".

Management involves monitoring and implementing tasks, i.e. "setting up systems to ensure that plans are implemented precisely and efficiently."

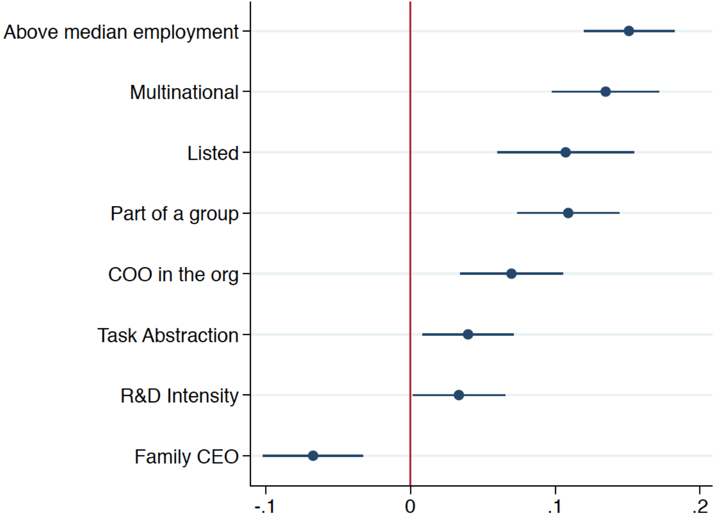
Leadership aims primarily at the creation of organizational alignment, and involves significant investments in interpersonal communication.

> *The knowledge worker makes much greater time demands than the manual worker on his superiors as well as on his associates...One has to sit down with a knowledge worker and think through with him what should be done and why, before when knowing whether he is doing a satisfactory job or not (Drucker 1967).*

Estimated Behavior Indices



Correlates of Behavior Index



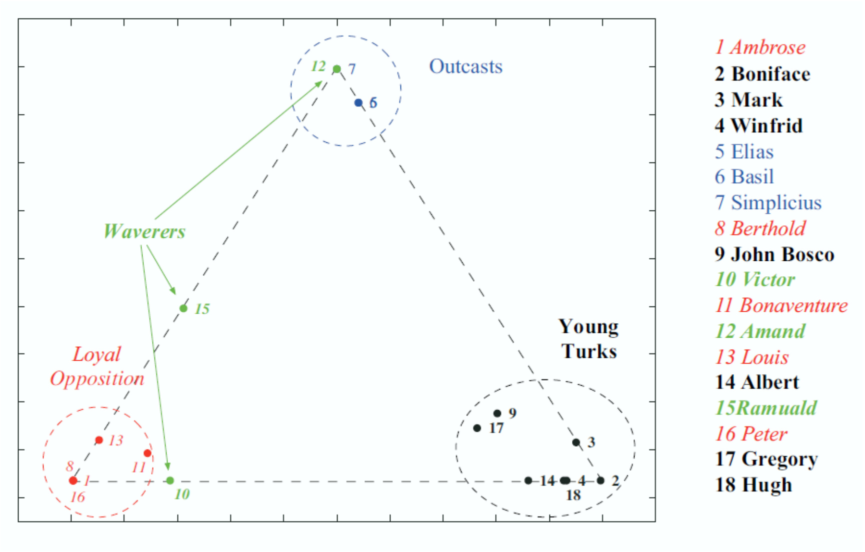
Network Data

Mixed-Membership Stochastic Blockmodel [Airoldi et al., 2008]

DGP for graph consisting of nodes $\mathcal{N}$ and adjacency matrix Y.

- For each node $p \in \mathcal{N}$:
  - Draw a K dimensional mixed membership vector $\vec{\pi}_p \sim \text{Dirichlet}\,(\,\vec{\alpha}\,)$.
- For each pair of nodes $(p,q) \in \mathcal{N} \times \mathcal{N}$:
  - Draw membership indicator for the initiator, $\vec{z}_{p \to q} \sim \text{Multinomial}\,(\,\vec{\pi}_p\,)$.
  - Draw membership indicator for the receiver, $\vec{z}_{q \to p} \sim \text{Multinomial}\,(\,\vec{\pi}_q\,)$.
  - Sample the value of their interaction, $Y(p,q) \sim \text{Bernoulli}\,(\,\vec{z}_{p \to q}^{\top} B\, \vec{z}_{p \leftarrow q}\,)$.

Example Output [Airoldi et al., 2008]



Incorporating Additional Structure

Relaxing Full Exchangeability

Vanilla topic models assume all documents are built from the same data generating process.

Generally not consistent with interesting behavioral hypotheses in social science.

Motivates need to build additional structure into the DGP for documents.

One common strategy replaces Dirichlet with logistic normal prior.

Automatic inference makes inference in these model feasible for applied researchers [Sacher et al., 2021].

Dynamic Topic Model

Suppose all documents carry a time stamp t.

1. Draw $\boldsymbol{\theta}_d$ independently for $d = 1, \ldots, D$ from Dirichlet(α).
2. Generate parameter vector $\tilde{\boldsymbol{\beta}}_{t,k} \in \mathbb{R}^V$ where $\tilde{\boldsymbol{\beta}}_{t,k} \sim \mathcal{N}(\tilde{\boldsymbol{\beta}}_{t-1,k}, \sigma^2 I_V)$.
3. Form topics according to $\beta_{t,k,v} = \frac{\exp(\tilde{\beta}_{t,k,v})}{\sum_v \exp(\tilde{\beta}_{t,k,v})}$.
4. $\mathbf{x}_d \sim$ Multinomial$(\sum_k \theta_{d,k} \boldsymbol{\beta}_{t(d),k}, N_d)$

Example with Journal Abstracts

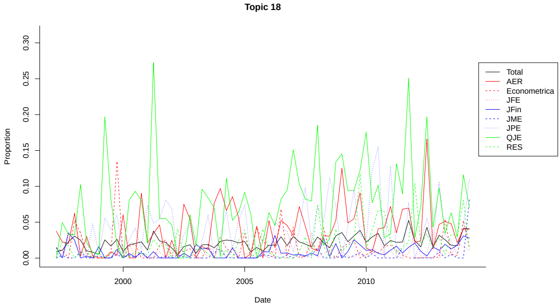
The following results are for a 20-topic version of DTM I on the abstracts of eight top economics journals from 1997:II to 2014:II (thanks to Julian Ashwin for collecting data and estimating):

- The Quarterly Journal of Economics
- Journal of Political Economy
- American Economic Review
- Econometrica
- Journal of Financial Economics
- Journal of Finance
- Review of Economic Studies
- Journal of Monetary Economics

Example with Journal Abstracts

Education Economics topic

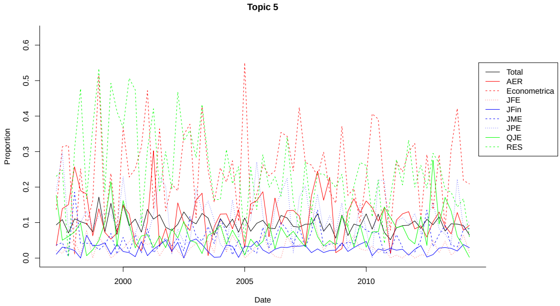
$$\begin{pmatrix} \text{school} \\ \text{educ} \\ \text{network} \\ \text{program} \\ \text{student} \\ \text{famili} \\ \text{score} \\ \text{year} \\ \text{black} \\ \text{class} \end{pmatrix} \rightarrow \begin{pmatrix} \text{school} \\ \text{educ} \\ \text{network} \\ \text{famili} \\ \text{student} \\ \text{program} \\ \text{colleg} \\ \text{class} \\ \text{group} \\ \text{year} \end{pmatrix} \rightarrow \begin{pmatrix} \text{school} \\ \text{educ} \\ \text{famili} \\ \text{group} \\ \text{student} \\ \text{network} \\ \text{program} \\ \text{black} \\ \text{score} \\ \text{white} \end{pmatrix} \rightarrow \begin{pmatrix} \text{school} \\ \text{student} \\ \text{network} \\ \text{group} \\ \text{famili} \\ \text{educ} \\ \text{social} \\ \text{peer} \\ \text{score} \\ \text{year} \end{pmatrix} \rightarrow \begin{pmatrix} \text{school} \\ \text{group} \\ \text{network} \\ \text{student} \\ \text{peer} \\ \text{social} \\ \text{educ} \\ \text{year} \\ \text{score} \\ \text{famili} \end{pmatrix}$$



Example with Journal Abstracts

Game Theory topic

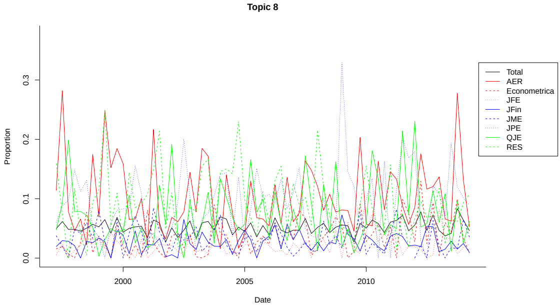
$$\begin{pmatrix} \text{equilibrium} \\ \text{agent} \\ \text{inform} \\ \text{contract} \\ \text{game} \\ \text{auction} \\ \text{optim} \\ \text{may} \\ \text{equilibria} \\ \text{effici} \end{pmatrix} \rightarrow \begin{pmatrix} \text{inform} \\ \text{equilibrium} \\ \text{agent} \\ \text{game} \\ \text{contract} \\ \text{optim} \\ \text{auction} \\ \text{equilibria} \\ \text{effici} \\ \text{player} \end{pmatrix} \rightarrow \begin{pmatrix} \text{agent} \\ \text{equilibrium} \\ \text{inform} \\ \text{game} \\ \text{optim} \\ \text{contract} \\ \text{auction} \\ \text{equilibria} \\ \text{effici} \\ \text{privat} \end{pmatrix} \rightarrow \begin{pmatrix} \text{inform} \\ \text{equilibrium} \\ \text{agent} \\ \text{game} \\ \text{contract} \\ \text{optim} \\ \text{effici} \\ \text{auction} \\ \text{set} \\ \text{player} \end{pmatrix} \rightarrow \begin{pmatrix} \text{inform} \\ \text{agent} \\ \text{equilibrium} \\ \text{game} \\ \text{privat} \\ \text{optim} \\ \text{contract} \\ \text{equilibria} \\ \text{player} \\ \text{effici} \end{pmatrix}$$



Example with Journal Abstracts

Industrial Organization topic

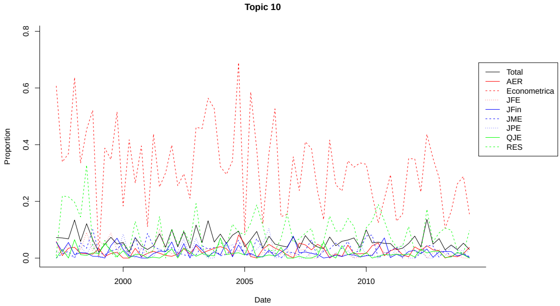
$$
\begin{pmatrix} \text{cost} \\ \text{product} \\ \text{competit} \\ \text{incent} \\ \text{effici} \\ \text{contract} \\ \text{demand} \\ \text{may} \\ \text{effort} \\ \text{welfar} \end{pmatrix}
\rightarrow
\begin{pmatrix} \text{product} \\ \text{cost} \\ \text{competit} \\ \text{good} \\ \text{industri} \\ \text{demand} \\ \text{profit} \\ \text{qualiti} \\ \text{effort} \\ \text{consum} \end{pmatrix}
\rightarrow
\begin{pmatrix} \text{product} \\ \text{cost} \\ \text{competit} \\ \text{consum} \\ \text{good} \\ \text{industri} \\ \text{demand} \\ \text{qualiti} \\ \text{entri} \\ \text{welfar} \end{pmatrix}
\rightarrow
\begin{pmatrix} \text{cost} \\ \text{product} \\ \text{competit} \\ \text{consum} \\ \text{industri} \\ \text{demand} \\ \text{entri} \\ \text{welfar} \\ \text{qualiti} \\ \text{good} \end{pmatrix}
\rightarrow
\begin{pmatrix} \text{product} \\ \text{cost} \\ \text{competit} \\ \text{good} \\ \text{demand} \\ \text{consum} \\ \text{entri} \\ \text{qualiti} \\ \text{export} \\ \text{industri} \end{pmatrix}
$$



Example with Journal Abstracts

Econometric Theory topic

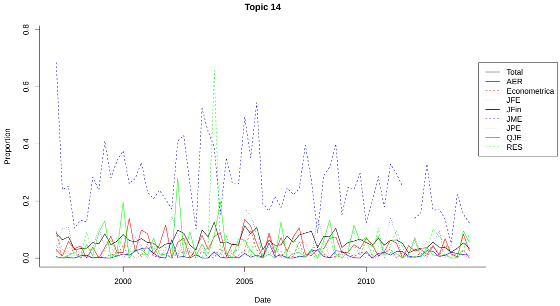
$$\begin{pmatrix} \text{estim} \\ \text{test} \\ \text{distribut} \\ \text{paramet} \\ \text{condit} \\ \text{asymptot} \\ \text{method} \\ \text{sampl} \\ \text{function} \\ \text{regress} \end{pmatrix} \rightarrow \begin{pmatrix} \text{estim} \\ \text{distribut} \\ \text{function} \\ \text{test} \\ \text{paramet} \\ \text{method} \\ \text{asymptot} \\ \text{condit} \\ \text{sampl} \\ \text{consist} \end{pmatrix} \rightarrow \begin{pmatrix} \text{estim} \\ \text{method} \\ \text{distribut} \\ \text{function} \\ \text{condit} \\ \text{test} \\ \text{paramet} \\ \text{asymptot} \\ \text{variabl} \\ \text{consist} \end{pmatrix} \rightarrow \begin{pmatrix} \text{estim} \\ \text{distribut} \\ \text{function} \\ \text{paramet} \\ \text{condit} \\ \text{method} \\ \text{identif} \\ \text{test} \\ \text{asymptot} \\ \text{variabl} \end{pmatrix} \rightarrow \begin{pmatrix} \text{estim} \\ \text{method} \\ \text{distribut} \\ \text{condit} \\ \text{test} \\ \text{paramet} \\ \text{function} \\ \text{identif} \\ \text{asymptot} \\ \text{propos} \end{pmatrix}$$



Example with Journal Abstracts

Monetary Policy topic

$$
\begin{pmatrix} \text{polici} \\ \text{monetari} \\ \text{rate} \\ \text{inflat} \\ \text{exchang} \\ \text{money} \\ \text{currenc} \\ \text{real} \\ \text{chang} \\ \text{nomin} \end{pmatrix}
\rightarrow
\begin{pmatrix} \text{polici} \\ \text{rate} \\ \text{monetari} \\ \text{inflat} \\ \text{exchang} \\ \text{interest} \\ \text{money} \\ \text{real} \\ \text{scienc} \\ \text{currenc} \end{pmatrix}
\rightarrow
\begin{pmatrix} \text{polici} \\ \text{monetari} \\ \text{inflat} \\ \text{rate} \\ \text{money} \\ \text{exchang} \\ \text{shock} \\ \text{rule} \\ \text{interest} \\ \text{real} \end{pmatrix}
\rightarrow
\begin{pmatrix} \text{polici} \\ \text{monetari} \\ \text{inflat} \\ \text{rate} \\ \text{shock} \\ \text{interest} \\ \text{real} \\ \text{nomin} \\ \text{optim} \\ \text{exchang} \end{pmatrix}
\rightarrow
\begin{pmatrix} \text{polici} \\ \text{rate} \\ \text{monetari} \\ \text{exchang} \\ \text{inflat} \\ \text{nomin} \\ \text{money} \\ \text{shock} \\ \text{interest} \\ \text{real} \end{pmatrix}
$$



Topic Prevalence Model [Roberts et al., 2014]

Suppose all documents have vector of associated covariates $\mathbf{g}_d$.

1. Draw $\boldsymbol{\beta}_k$ independently for $k = 1, \ldots, K$ from Dirichlet(η).
2. Generate parameter vector $\tilde{\boldsymbol{\theta}}_d \in \mathbb{R}^K$ from $\tilde{\theta}_{d,k} \sim \mathcal{N}(\mathbf{g}_d^T \gamma_k, \sigma^2)$.
3. $\theta_{d,k} = \frac{\exp(\tilde{\theta}_{d,k})}{\sum_v \exp(\tilde{\theta}_{d,k})}$.
4. $\mathbf{x}_d \sim \text{Multinomial}(\sum_k \theta_{d,k} \boldsymbol{\beta}_k, N_d)$

Topic model with regression structure in prior distribution on topic coverage in documents.

Conclusion

Topic models are Bayesian factor models for discrete data.

They explicitly model the document-level correlation structure among vocabulary terms.

This helps resolve the problem of synonymy in NLP.

Underlying data is still the document-term matrix which is an important limitation.

References I

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008).
Mixed Membership Stochastic Blockmodels.
Journal of Machine Learning Research, 9(65):1981–2014.

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.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., and Blei, D. (2009).
Reading Tea Leaves: How Humans Interpret Topic Models.
In Advances in Neural Information Processing Systems, volume 22. Curran Associates, Inc.

References II

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.

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.

References III

Gross, J. H. and Manrique-Vallier, D. (2014).
A mixed membership approach to the assessment of political ideology from survey responses.
In Airoldi, E. M., Blei, D., Erosheva, E. A., and Fienberg, S. E., editors, Handbook of Mixed Membership Models and Its Applications. CRC Press.

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.

Hofmann, T. (1999).
Probabilistic latent semantic analysis.
In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, UAI'99, pages 289–296, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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.

References IV

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.

Lee, D. D. and Seung, H. S. (1999).
Learning the parts of objects by non-negative matrix factorization.
Nature, 401(6755):788–791.

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.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010).
Automatic evaluation of topic coherence.
In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10, pages 100–108, USA. Association for Computational Linguistics.

References V

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.