Machine Learning in Econometrics

Introduction to Text Analysis

Prof. Daniel Wilhelm
Chair of Statistics and Econometrics
Department of Statistics
LMU Munich

SS 2023

Trends in Data Types

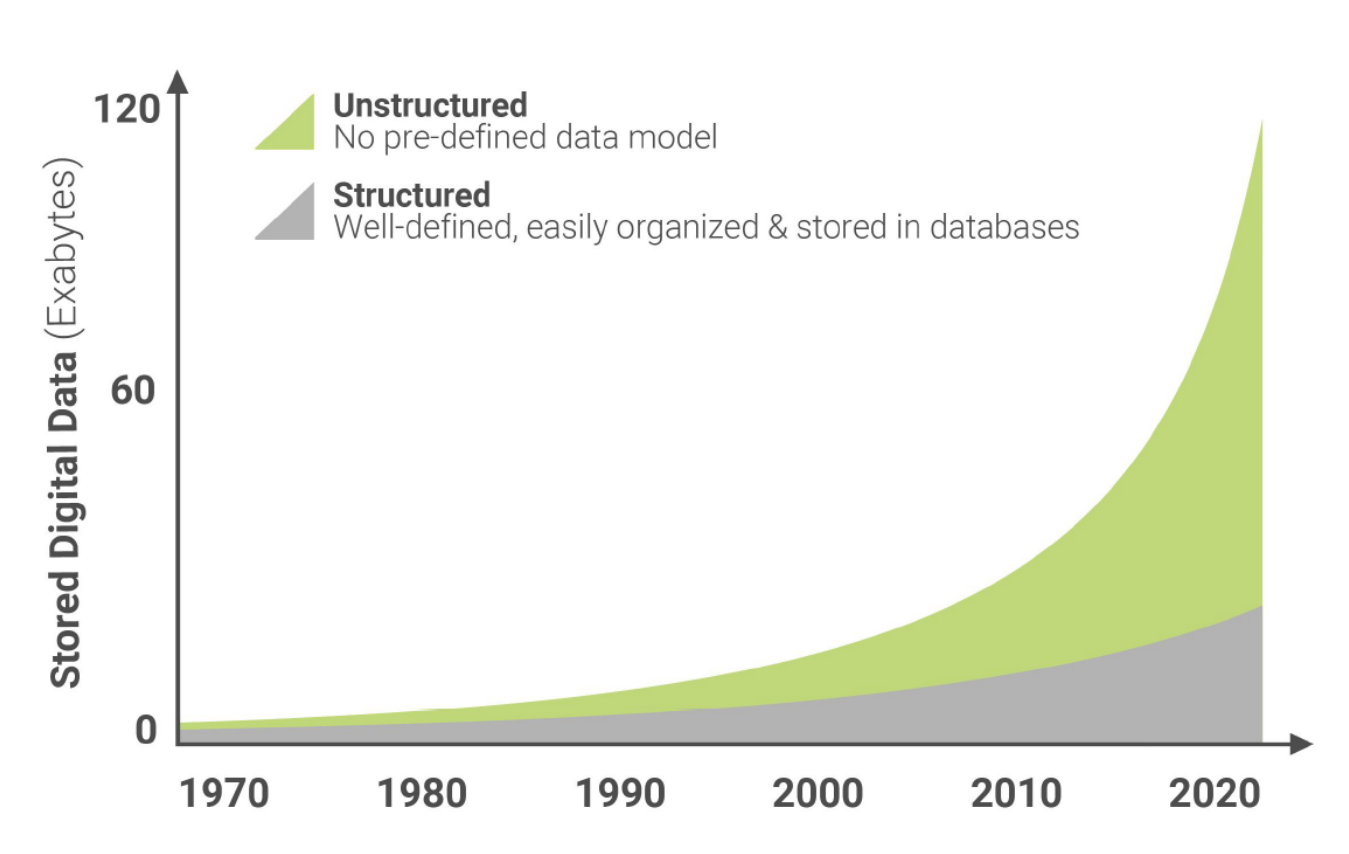


Figure 1: Source: Harvard Business Review

Unstructured Data

- unstructured data is not organized in traditional relational databases
- **key challenge:** extracting relevant information and separating it from irrelevant information
- **Examples:**
  - digital text
  - scanned historical texts
  - satellite images
  - audio recordings

Happenstance Data

Administrative data is collected

- for subsequent statistical analysis
- with clear definition of what is supposed to be measured
- with consistent, representative sampling frame
- data access arrangements exist

Happenstance data

- arises as byproduct of everyday activities by agents
- not collected with consistent, representative sampling frame
- data access often difficult

Examples of Types of Data

	administrative	**happenstance**
structured	traditional economic data	credit card transactions
unstructured	company filings	tweets

We will focus on text data . . .

Text Data Is Very Useful in Economics

Examples:

- **text from financial news, social media and company filings** is used
  - to predict asset price movements
  - study the causal impact of new information
- **text from news** is used
  - to forecast variation in inflation and unemployment
  - to estimate the causal effect of policy uncertainty
- **text from news and social media** is used
  - to study the drivers and effects of political slant
- **text from advertisements and product reviews** is used
  - to study drivers of consumers decision making
- **text from politicians' speeches** is used
  - to study the dynamics of political agendas and debate

Text Data Is Inherently High-Dimensional

Suppose a text

- is w words long
- words drawn from dictionary of p possible words

There are p^w unique texts!

- e.g., twitter messages ($w = 30$) using the 1,000 most common words in English: $1,000^{30} = 10^{90}$
- roughly same order of magnitude as the number of atoms in the universe

Workflow For Text Analysis

Typical Workflow

1. Harvest the data
2. Tidy the data
3. Analyze the data

Harvesting Text Data

Text data can be obtained from many different sources:

- **web-scraping**
  - e.g., news sites
  - R package `rvest`
- **APIs**
  - e.g., Twitter
  - R package `rtweet`
- **OCR**
  - e.g., scanned historical texts
  - R package `tesseract`
  - useful ressources on Melissa Dells's website: `https://dell-research-harvard.github.io`

Examples: Harvesting Text Data

(In-class)

Text Data Is Complex

- Text often obtained as **corpus** of several **documents**, e.g.,
  - collection of speeches
  - collection of sentences
- Text is a string of characters.
- Some characters may be from the Latin alphabet ('a', 'b', . . . , 'A', 'B', . . .), but there may also be:
  1. Decorated Latin letters (e.g. ö)
  2. Non-Latin alphabetic characters (e.g. Chinese and Arabic)
  3. Punctuation (e.g. '!')
  4. White spaces, tabs, newlines
  5. Numbers
  6. Non-alphanumeric characters (e.g. '@')

How can we obtain an informative, quantitative representation of these character strings?

Tidying Text Data

Typical steps for **pre-processing** text:

1. break up corpus into documents
    - e.g., into emails, speeches, sentences

2. feature selection (dimension reduction):
    - remove unwanted characters (e.g., punctuation)
    - remove stopwords (e.g., articles, prepositions)
    - represent words by their stem (e.g., stems of "duties" and "duty" are "duti")

3. create document term matrix
    - matrix containing counts of words occurring for each document

Example of Document Term Matrix

- Corpus containing three documents:
  1. 'stephen is nice'
  2. 'john is also nice'
  3. 'george is mean'
- set of unique terms: $\{\text{stephen}, \text{is}, \text{nice}, \text{john}, \text{also}, \text{george}, \text{mean}\}$
- document term matrix:

$$C := \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

contains

- a row for each document
- a column for each of the unique terms in the corpus
- element C_{ij} indicates how often term j occurs in document i

Key features of document term matrix:

- high-dimensionality
- sparsity

Examples: Tidying Text Data

(In-class)

Bag of Words Versus N-Grams

In previous example, **term** was defined as a **word**

$\rightarrow$ document term matrix shows occurence of each word

$\rightarrow$ "bag of words representation"

Sometimes useful to define **term** as an **"n-gram"**:

- term is a string of n words
- e.g., 3-grams like "head of state"
- accounts for dependence in occurence of words

Idea of Word Embeddings

- representing text as word counts is crude
- ignores similarity of words
  - e.g., "king" and "monarch" very similar
  - bag of word representations treat these as separate words

Idea of Word Embeddings:

Capture meaning of words by measuring similarity between them!

Word Embeddings

Main idea:

- each term is assigned a (low-dimensional) vector
- length of vector normalized to 1
- angle between vectors measures similarity of terms

Techniques for word embeddings:

- need to estimate the low-dimensional representation of a dictionary
- based on ML techniques
- popular embedding techniques:
  - Word2Vec
  - GloVe

Example: Word Embedding

- corpus represented by unique terms {king, queen, prince, man, woman, child}
- document term matrix

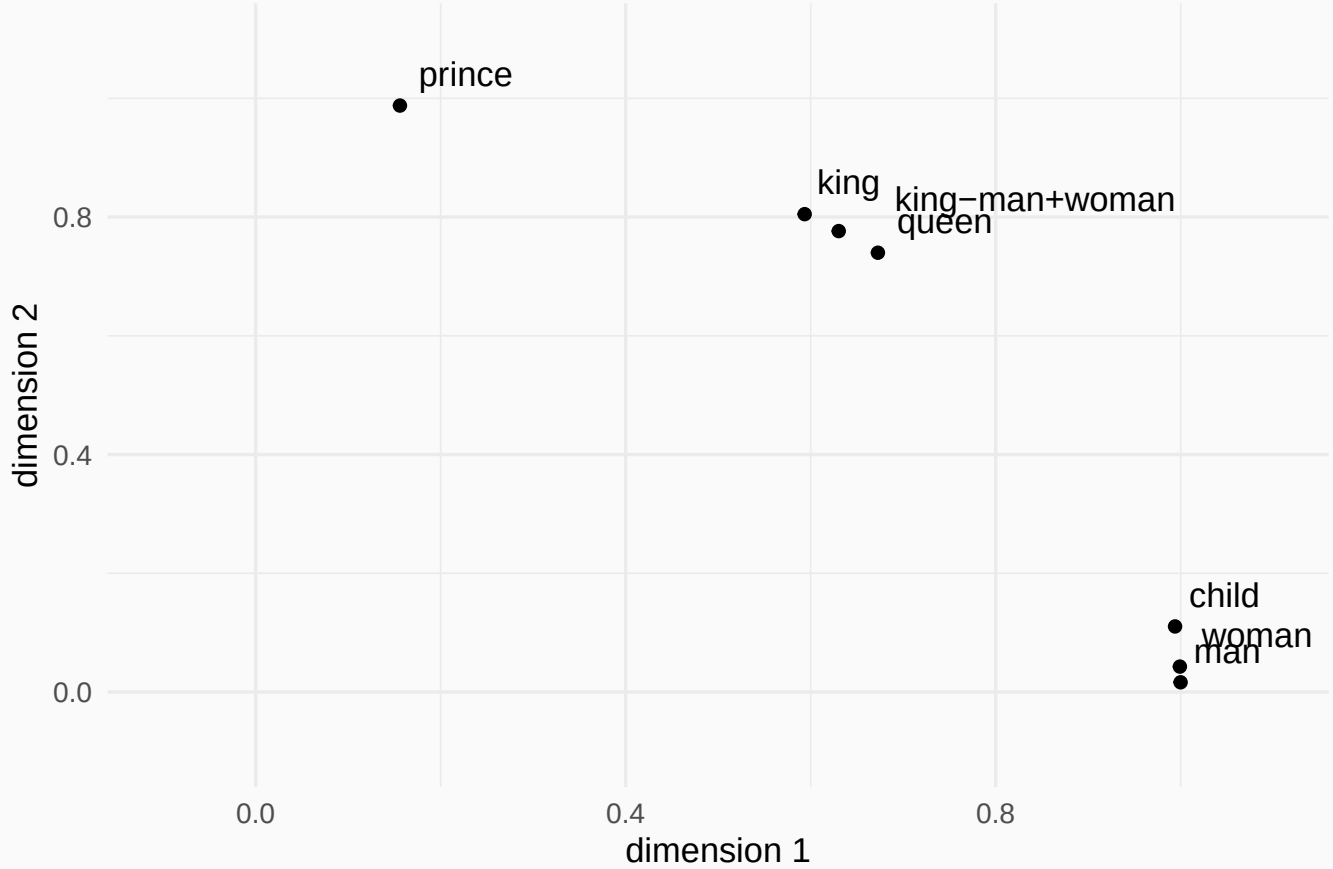
$$C := \begin{pmatrix} 10 & 4 & 1 & 0 & 0 & 9 \\ 2 & 1 & 5 & 1 & 2 & 0 \\ 3 & 2 & 0 & 0 & 3 & 1 \\ & & \vdots & & & \end{pmatrix}$$

- naive word embedding would be to assign each word the corresponding column of counts in C, i.e.
  - "king" is assigned the vector $(10, 2, 3, \ldots)$
  - high-dimensional
  - closeness of vectors only encode frequency of words occuring together (not necessarily similarity in "meaning")
- more meaningful word embedding might look like:

dimension	king	queen	prince	man	woman	child
1	0.80	0.74	0.99	0.02	0.04	0.11
2	0.59	0.67	0.16	1.00	1.00	0.99

  - low-dimensional
  - measures similarity of words in two dimensions

Example: Word Embeddings (Cont'd)



- "king" is close to "queen", but not close to "child"
- "king - man + woman" is closer to "queen"

Analyzing Text

1. dictionary-based methods
2. text regression
3. generative models

Dictionary-Based Methods

Goal

- C_i: row i of document term matrix (document i's count vector)
- Y_i: outcome of interest, e.g.
  - sentiment of document i
  - topic of document i

Goal:

compute sentiment Y_i using document count vector C_i

Dictionary-Based Methods

Idea:

Compute

$$Y_i = f(C_i)$$

for a pre-specified (i.e., known) function $f(\cdot)$.

Example 1 (sentiment analysis)

- dictionary that assigns to each word the label "positive" or "negative"
- $Y_i = f(C_i)$ counts how many words in C_i have the label "positive"
- dictionary is given, i.e. $f(\cdot)$ is known
  $\rightarrow$ no statistical analysis needed to compute Y_i

Example

(In-class)

Example: Sentiment of Wall Street Journal Column "Abreast of the Market"

Tetlock (2007):

- **text source:** articles from the Wall Street Journal column "Abreast of the Market" (1984 - 1999)
  - covers developments on financial markets the previous day
  - expert opinions and commentary
- dictionary "General Inquirer" lists words together with indicators for whether they fit into each of 77 categories
  - categories from the Harvard psychosocial dictionary, e.g., "positive", "optimistic"
- Y_i is a 77-dimensional vector with the k-th element showing the count of words in document i that belong to category i
- **principal component analysis**:
  - from Y_i extract the first principal component
  - highly correlated with negative words
  - factor interpreted as measure of "bad news" (or "pessimism factor")

How is the pessimism factor related to stock market returns?

WSJ Column "Abreast of the Market"

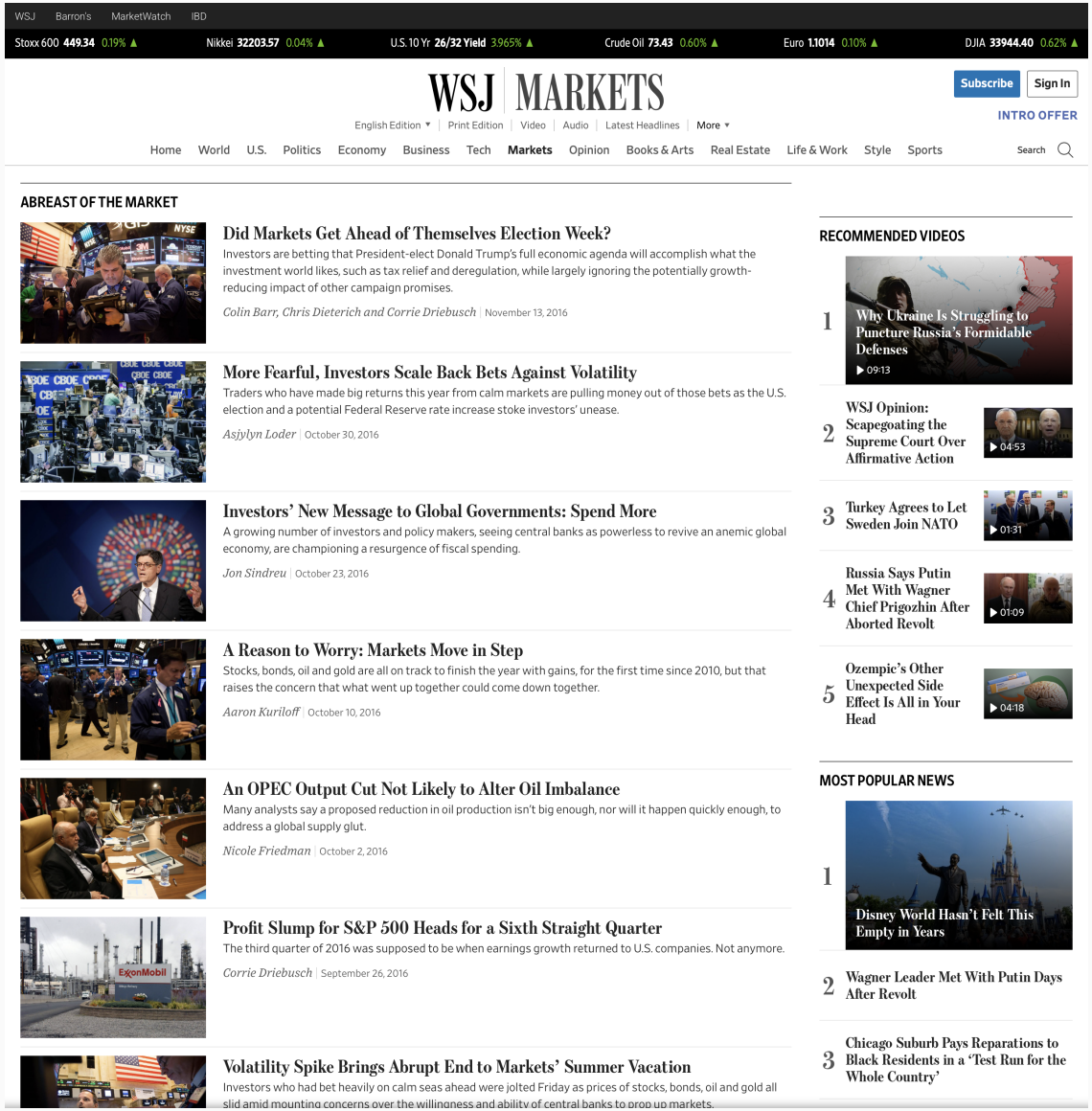


Figure 2: Source: `https://www.wsj.com/news/types/abreast-of-the-market`

Effect of Bad News on Stock Returns

	Regressand: Dow Jones Returns		
News Measure	Pessimism	Negative	Weak
$BdNws_{t-1}$	***−8.1***	***−4.4***	***−6.0***
$BdNws_{t-2}$	0.4	3.6	2.0
$BdNws_{t-3}$	0.5	−2.4	−1.2
$BdNws_{t-4}$	**4.7**	***4.4***	***6.3***
$BdNws_{t-5}$	1.2	2.9	**3.6**
$\chi^2(5)$ [*Joint*]	***20.0***	***20.8***	***26.5***
p-value	0.001	0.001	0.000
Sum of 2 to 5	**6.8**	***9.5***	***10.7***
$\chi^2(1)$ [*Reversal*]	**4.05**	***8.35***	***10.1***
p-value	0.044	0.004	0.002

Figure 3: Source: Tetlock (2007). Effect of "bad news" factor on one-day ahead forecasts of Dow Jones Industrial Average returns. Columns show different definitions of the "bad news" factor.

Conclusions:

1. "bad news" factor **significantly decreases** one-day ahead forecasts of stock returns (8.1 basis points)
2. negative effects **reversed in days 2-5** (coefficients of lags add up to 6.8 basis points)

→ "bad news" have **temporary negative effect**, but don't seem to contain information about fundamental value of stocks

Finance-Specific Dictionary

Loughran and McDonald (2011):

- argue that "General Inquirer" not suitable for financial applications
- create their own finance-specific dictionary of positive/negative terms
- show improved predictive power using new dictionary

Example: Policy Uncertainty

Baker, Bloom, and Davis (2016):

- develop **index of economic policy uncertainty** (EPU) using news articles
- news articles from 10 leading U.S. newspapers (185 - 2015)
- create their own dictionary assigning words to three categories (economy, policy, uncertainty)
- Y_i is a vector with k-th element containing the count of news articles from newspaper k in country-month i containing at least one word in each of the three categories
- apply some normalizations to Y_i
- **EPU index** for country-month i is then the average of elements in Y_i
- validate EPU index by human audit of 12,000 articles

How do firms and the economy as a whole respond to increased EPU?

EPU Index for the U.S.

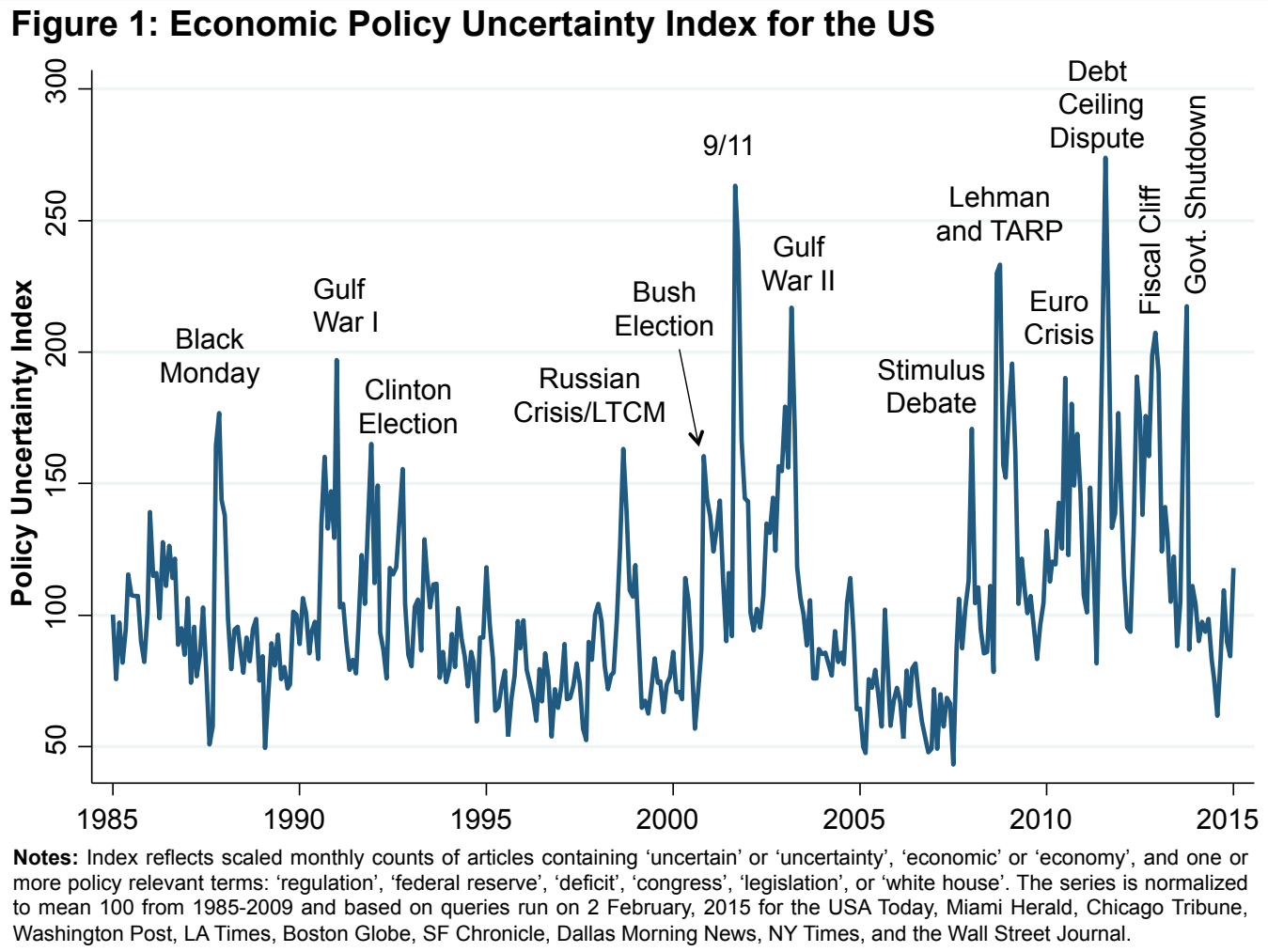


Figure 4: Source: Baker, Bloom, and Davis (2016).

→ `https://www.policyuncertainty.com` provides data and more indices

Effects of EPU on Economic Activity

Baker, Bloom, and Davis (2016)'s Findings

1. **firm-level regressions:** higher EPU leads to
   - reduced employment
   - reduced investment
   - larger asset price volatility

2. **country-level VAR regressions:** higher EPU correlated with
   - lower investment
   - lower employment
   - lower production

Text Regression

Predicting Attributes From Term Counts

Dictionary-Based Methods:

- compute attribute Y_i (e.g. sentiment, EPU) from term counts C_i using dictionary
- no statistical inference
- attribute Y_i can be directly computed from counts C_i and dictionary

Text Regression:

- predict attribute Y_i (e.g. sentiment, EPU) from term counts C_i using regression
- mapping from counts to attributes not known
- statistical inference
- attribute Y_i observed in training sample, but not observed in test sample

Text Regression Setup

- observed, scaler **outcome** Y_i for document i
- observed vector of **term counts** C_i for document i
  - $C_i := (C_{i1}, C_{i2}, \ldots)'$, where C_{ij} is the count of term j in document i
- p-dimensional vector of **transformations** X_i of C_i
  - e.g. could be equal to term counts C_i
  - e.g. could be word embedding derived from C_i
  - assume X_i contains a constant

Goal:

Want to predict Y_i with X_i!

Approaches:

Use ML methods for

1. linear regression (e.g., LASSO)
2. nonlinear regression (e.g., trees, forests, neural nets)
3. logistic regression

Logistic Regression

Suppose the **outcome is binary** ($Y_i \in \{0, 1\}$) and

$$P(Y_i = 1 | X_i = x) = \frac{e^{x'\beta}}{1 + e^{x'\beta}}$$

Then, maximizing the (conditional) log-likelihood is equivalent to

$$\min_{b \in \mathbb{R}^p} -\mathbb{E}_n \left[Y_i X_i' b - \log \left(1 + e^{X_i' b} \right) \right].$$

When X_i is high-dimensional, then minimize the **penalized logistic regression** objective

$$\min_{b \in \mathbb{R}^p} -\mathbb{E}_n \left[Y_i X_i' b - \log \left(1 + e^{X_i' b} \right) \right] + \frac{\lambda}{n} \sum_{k=1}^{p} |b_k|^q.$$

for some q (e.g., $q = 1$ for the LASSO).

Useful alternative to LASSO when outcome is binary!

Example

(In-class)

Example: Hedonic Pricing

Bajari, Cen, Chernozhukov, Manukonda, Vijaykumar, Wang, Huerta, Li, Leng, Monokroussos, and Wan (2023):

- Amazon data on apparel products (2013-2018):
  - price, description, image
- creating **predictors of price**:
  - large language model BERT creates word embeddings for each word in the description text
  - W_i is average of word embeddings of all words in product i's description
  - I_i is a similar embedding of product i's image using ResNet50
  - predictor $X_i := (W_i', I_i')'$ contains both embeddings
- Y_i is price of product i

Goals:

1. Estimate **hedonic price function** h in

$$Y_i = h(X_i) + \varepsilon_i, \qquad E[\varepsilon_i | X_i] = 0.$$

2. create hedonic **price index**

Figure 5: Source: Bajari, Cen, Chernozhukov, Manukonda, Vijaykumar, Wang, Huerta, Li, Leng, Monokroussos, and Wan (2023).

Hedonic Price Model

Hedonic regression:

$$Y_i = h(X_i) + \varepsilon_i, \qquad E[\varepsilon_i | X_i] = 0.$$

where

- Y_i is price of product i
- X_i is a vector of product i's characteristics

Interpretation:

price is combination of prices of characteristics

Construction of price index using hedonic price function h:

- $h(x)$ predicts price of bundle of characteristics x
- choose a bundle x of characteristics that represents a typical "**consumption basket**"
- hedonic price function $h(x)$ predicts price of that consumption basket

$\rightarrow$ price index

Bajari et al. (2023)

- estimate h with various ML methods on a training sample
- compare methods in terms of R^2 on a test sample

Monthly Out-of-Sample R^2

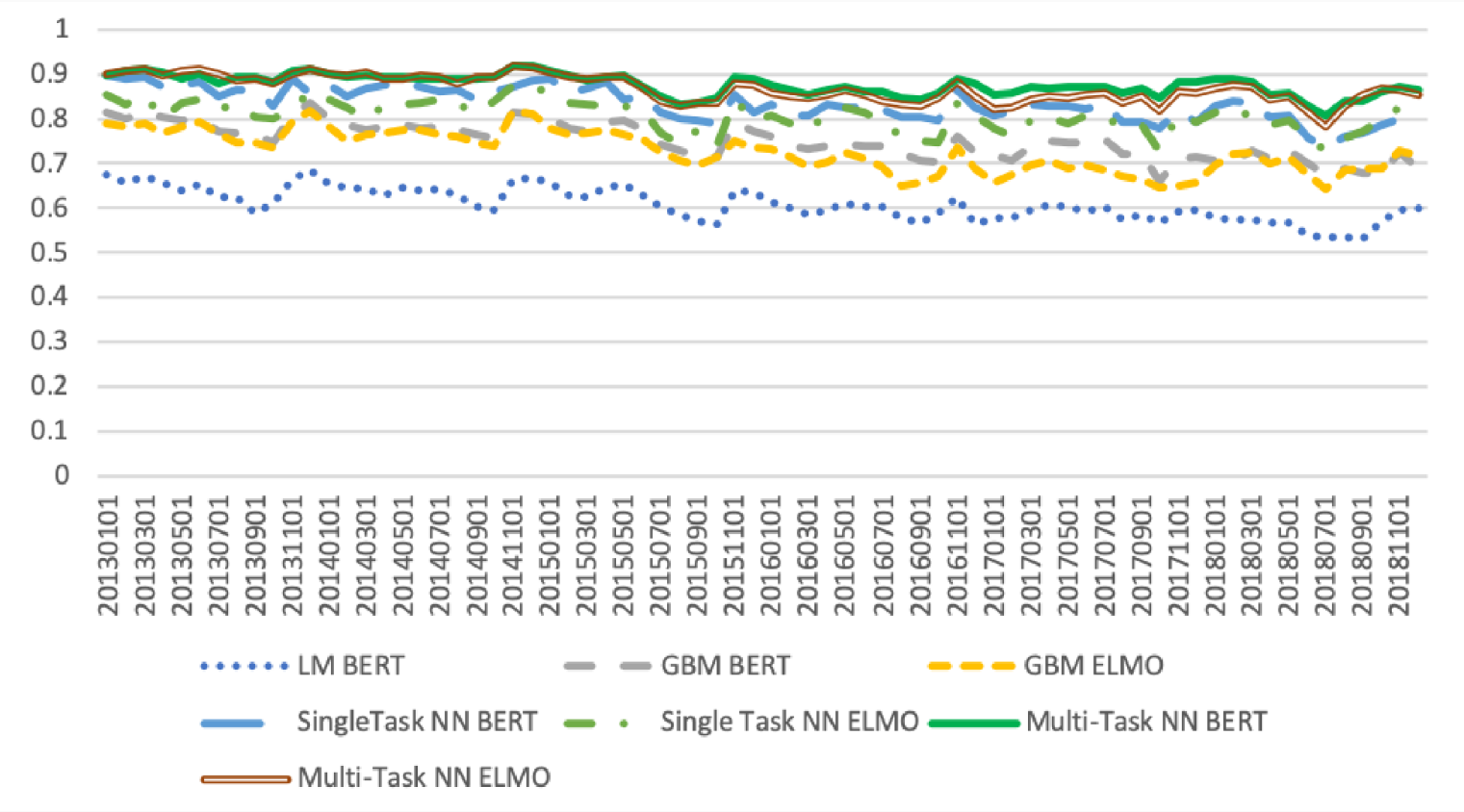


Figure 6: Source: Bajari, Cen, Chernozhukov, Manukonda, Vijaykumar, Wang, Huerta, Li, Leng, Monokroussos, and Wan (2023).

Conclusion:

Multi-Task neural net outperforms all other models

Average Annual Rate of Inflation in Apparel

Apparel Indices	*Change in Price Index*
Fisher Hedonic, Yearly Chaining (FY)	-0.77%
Fisher Hedonic, Monthly Chaining (FM)	-5.32%
Fisher Hedonic, Geometric Mean ($\sqrt{FY \cdot FM}$)	-3.16%
Fisher Matched, Monthly Chaining (FI)	-3.46%
Jevons Posted Price Index, Daily Chained (JPI)	-2.82%
Adobe Digital Price Index, Monthly Chained (DPI)	-1.30%
U.S. Urban Apparel (BLS)	-0.04%

Figure 7: Source: Bajari, Cen, Chernozhukov, Manukonda, Vijaykumar, Wang, Huerta, Li, Leng, Monokroussos, and Wan (2023).

Conclusion

Hedonic price index implies more decline than the consumer price index from the Bureau of Labor Statistics (BLS).

Generative Models

Text Regression Versus Generative Models

Text regression:

- treat text counts C_i as given predictors of an attribute Y_i (e.g., sentiment)
- learn about $E[Y_i|C_i]$ or, more generally, the distribution of $Y_i|C_i$

Generative models:

- model the generation of text through a model of the distribution of $C_i|Y_i$
- text is generated in different ways depending on the attribute Y_i
- e.g., Y_i indicates one of a finite number of topics or a political party

Generative models explicitly model how certain terms occur together!

Multinomial Distribution for Term Counts

Multinomial distribution for term counts:

$$C_i \sim MN(Q_i, m_i),$$

where

- $C_i := (C_{i1}, \ldots, C_{ip})'$ where
  - C_{ij} is count of how often term j occurs in document i
- choose m_i terms from dictionary of p terms
  - i.e., m_i is the total number of terms in document i: $m_i = \sum_{j=1}^{p} C_{ij}$
- $Q_i := (Q_{i1}, \ldots, Q_{ip})'$ where
  - Q_{ij} is the probability of choosing term j in document i
  - probabilities may be different in different documents
  - often modelled as function of some attribute(s) Y_i

$$Q_i = f(Y_i)$$

Multinomial distribution for term counts,

$$C_i \sim MN(Q_i, m_i),$$

combined with a **factor structure** for probability vector,

$$Q_i = Y_{i1}\theta_1 + \ldots + Y_{iK}\theta_K.$$

"Topic Model" interpretation:

- K topics
- θ_k is p-dimensional vector of probabilities for topic k
  $\rightarrow$ indicates which words in topic k are likely to occur!
- Y_{ik} is a scalar weight indicating importance of topic k for document i
  - weights satisfy $Y_{ik} \geq 0$ and $\sum_{k=1}^{K} Y_{ik} = 1$
  - **binary weights** ($Y_{ik} \in \{0, 1\}$) $\rightarrow$ indicates the **(single) topic** of document i
  - **real weights** $Y_{ik} \in [0, 1] \rightarrow$ document i is **mixture of topics**

Unsupervised Topic Model - Goal

Multinomial distribution for term counts:

$$C_i \sim MN(Q_i, m_i)$$

combined with a **factor structure** for probability vector

$$Q_i = Y_{i1}\theta_1 + \ldots + Y_{iK}\theta_K$$

Goal:

Want to infer $Y_{i1}, \ldots, Y_{iK}$ and $\theta_1, \ldots, \theta_K$ using observations on C_i and m_i.

Unsupervised Topic Model - Inference

Inference in topic models often **Bayesian**:

- place priors on $Y_{i1}, \ldots, Y_{iK}$ and $\theta_1, \ldots, \theta_K$
  - popular choice: conjugate Dirichlet priors
- Hofmann (1999) and Taddy (2012):
  - direct computation of posterior
- Blei, Ng, and Jordan (2003) ("Latent Dirichlet Allocation (LDA)"):
  - variational Bayes approximation of the posterior
  - many extensions, e.g. to unknown number of topics (Teh, Jordan, Beal, and Blei (2006)), time-varying topics (Blei and Lafferty (2006)), topics driven by sentiment (Blei and McAuliffe (2010))

Example

(In-class)

Example: CEO Behavior and Firm Performance

Bandiera, Prat, Hansen, and Sadun (2020):

- survey data on 1,114 CEOs in manufacturing firms from Brazil, France, Germany, India, the UK, and the U.S.
- record all daily activities by CEOs for a week
  - 654 types of activities
  - in 15min time blocks
- apply topic model (LDA):
  - each CEO i corresponds to a "document"
  - each CEO chooses activity (corresponding to a "term") for each 15min time block
  - each CEO described as mixture of K types (corresponding to K topics)
  - pick $K = 2$ types

Goal:

1. recover latent types of CEOs
2. study impact of CEO type on firm performance

Some Summary Statistics

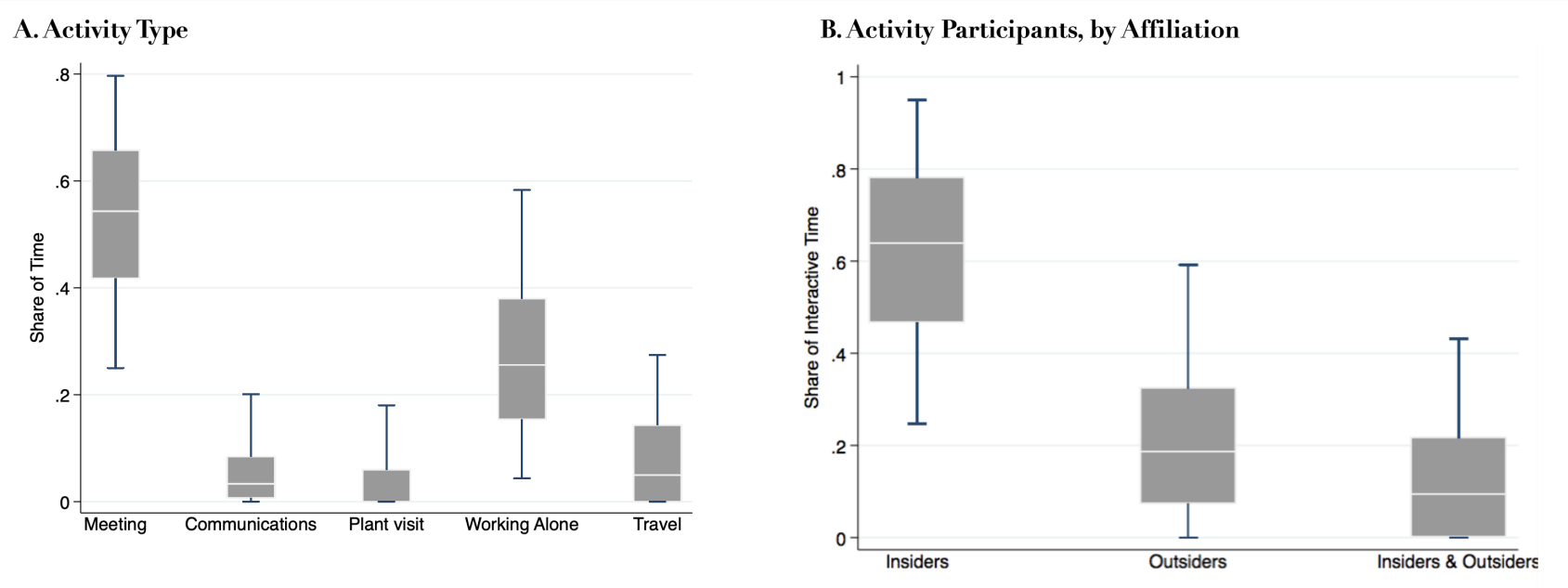


Figure 8: Source: Bandiera, Prat, Hansen, and Sadun (2020).

CEO Behavior Types

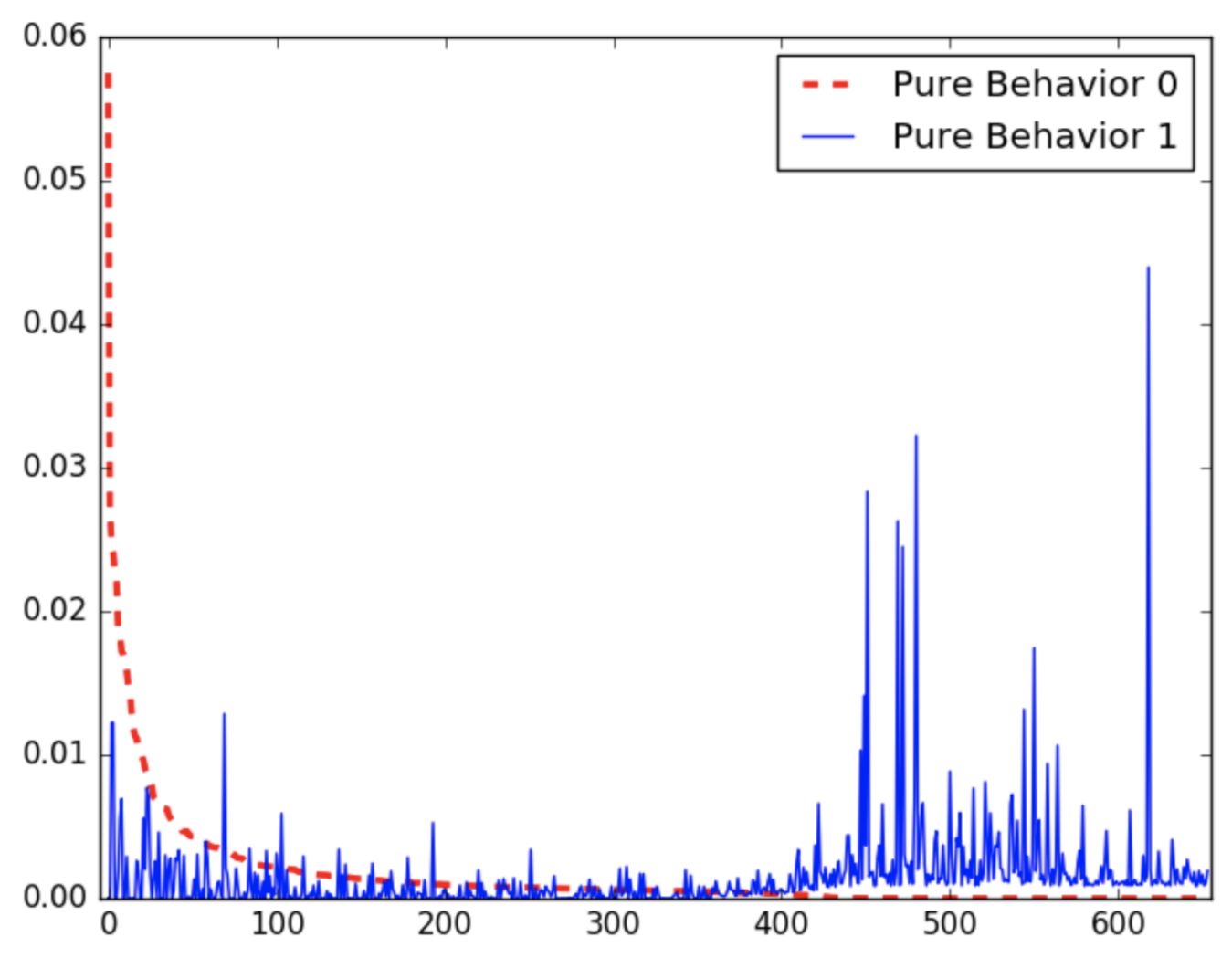


Figure 9: Probabilities for each of the 654 of activities occurring in the two types (corresponding to θ_1 and θ_2 for two "topics"). Source: Bandiera, Prat, Hansen, and Sadun (2020).

Differences across Pure Behaviors

Feature	X times less likely in Type 1	Feature	X times more likely in Type 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

Interpretation of CEO types:

1. **Management (Pure Behavior 0)**:
   - focus on monitoring and implementing tasks
   - mainly operational activities
2. **Leadership (Pure Behavior 1)**:
   - focus on creation of organizational alignment
   - mainly activities related to interpersonal communication

CEO Behavior Are Mixtures of the Two Types

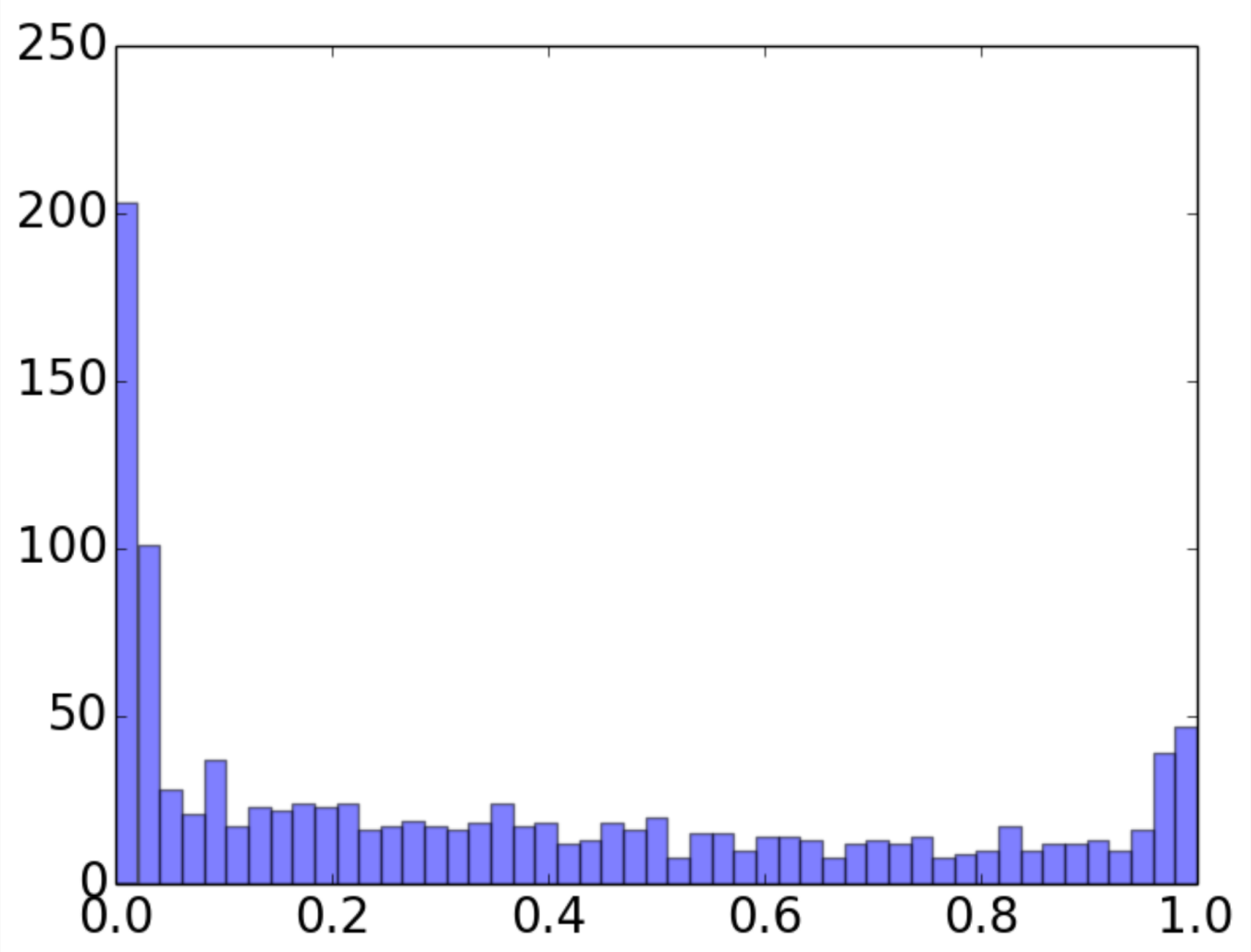


Figure 10: Each CEO i has a mixture weight Y_{i1} and $Y_{i2} = 1 - Y_{i1}$ on the two types. The graph shows the histogram of the predicted Y_{i2} across CEOs. Source: Bandiera, Prat, Hansen, and Sadun (2020).

CEO Behavior and Firm Performance

TABLE 3
CEO Behavior and Firm Performance

	Dependent Variable: log(sales)					Dependent Variable: Profits/ Employee
	(1)	(2)	(3)	(4)	(5)	(6)
CEO behavior index	.343***	.227**	.322***	.641**	.506**	10.029***
	(.108)	(.111)	(.121)	(.278)	(.236)	(3.456)
Log (employment)	(.040)	(.066)	(.099)	(.152)	(.090)	(.734)
	.889***	.555***	.346***	.339**	.784***	−.284
Log(capital)		.387***	.188***	.194*		
		(.042)	(.056)	(.098)		
Log(materials)			.447***	.421***		
			(.073)	(.109)		
Management					.179**	
					(.072)	
Observations (firms)	920	618	448	243	156	386
Observations used to compute means	2,202	1,519	1,054	604	383	1,028
Sample	All	With k	With k and m	With k and m, listed	With management score	With profits, listed

Figure 11: Y_{i2} = "CEO behavior index". Large values of Y_{i2} correspond to CEOs that are more like leaders rather than managers. Source: Bandiera, Prat, Hansen, and Sadun (2020).

Conclusions:

Firms with CEOs that are more like **leaders perform better** in terms of sales and profits!

Supervised Topic Model - Goal

Multinomial distribution for term counts:

$$C_i \sim MN(Q_i, m_i)$$

combined with a a model for how probability vector depends on attribute(s)

$$Q_i = f(Y_i)$$

Now, Y_i is observed!

Multinomial Logistic Regression - Goal

Multinomial logistic regression consists of a multinomial distribution for term counts,

$$C_i \sim MN(Q_i, m_i),$$

combined with a logistic model for how the probability vector depends on attribute(s),

$$Q_{ij} = \frac{e^{X_i' \beta_j}}{1 + e^{X_i' \beta_j}},$$

where $X_i := (1, Y_{i1}, \ldots, Y_{iK})'$ and β_j is a $(K+1)$-dim. vector of coefficients.

Goal:

Want to infer $\beta_1, \ldots, \beta_p$ using observations on $Y_{i1}, \ldots, Y_{iK}$, C_i and m_i.

Multinomial Logistic Regression - Estimation and Inference

Taddy (2013):

- proposes **multinomial inverse regression** for estimation of the multinomial logistic regression
- reduce estimation problem to standard logistic regression with
  - high-dimensional response (C_i)
  - low-dimensional predictors (Y_i)

Taddy (2015):

- proposes computationally attractive version of the above

$\rightarrow$ "distributed multinomial regression"

Example: Media Slant

Groseclose and Milyo (2005):

- media outlets have important role in political process
- potential power to sway both public opinion and policy

Goal:

How and why do media outlets slant the information they present?

Approach

Training sample:

- full text of U.S. congressional speeches from 1993-2002
- for each politician i, count how often each of **50 think tanks** are cited
  $\rightarrow$ term counts C_i
- to each politician i, assign **ADA score** (position on left-right political spectrum)
  $\rightarrow$ Y_i

Test sample:

- full text of U.S. news reports by media outlets from 1990-2004
- for each media outlet i, count how often each of **50 think tanks** are cited
  $\rightarrow$ term counts C_i

Approach:

1. **estimate multinomial logistic regression** model on training sample
2. **predict ADA score** of media outlets on the test sample

Predicted ADA Scores

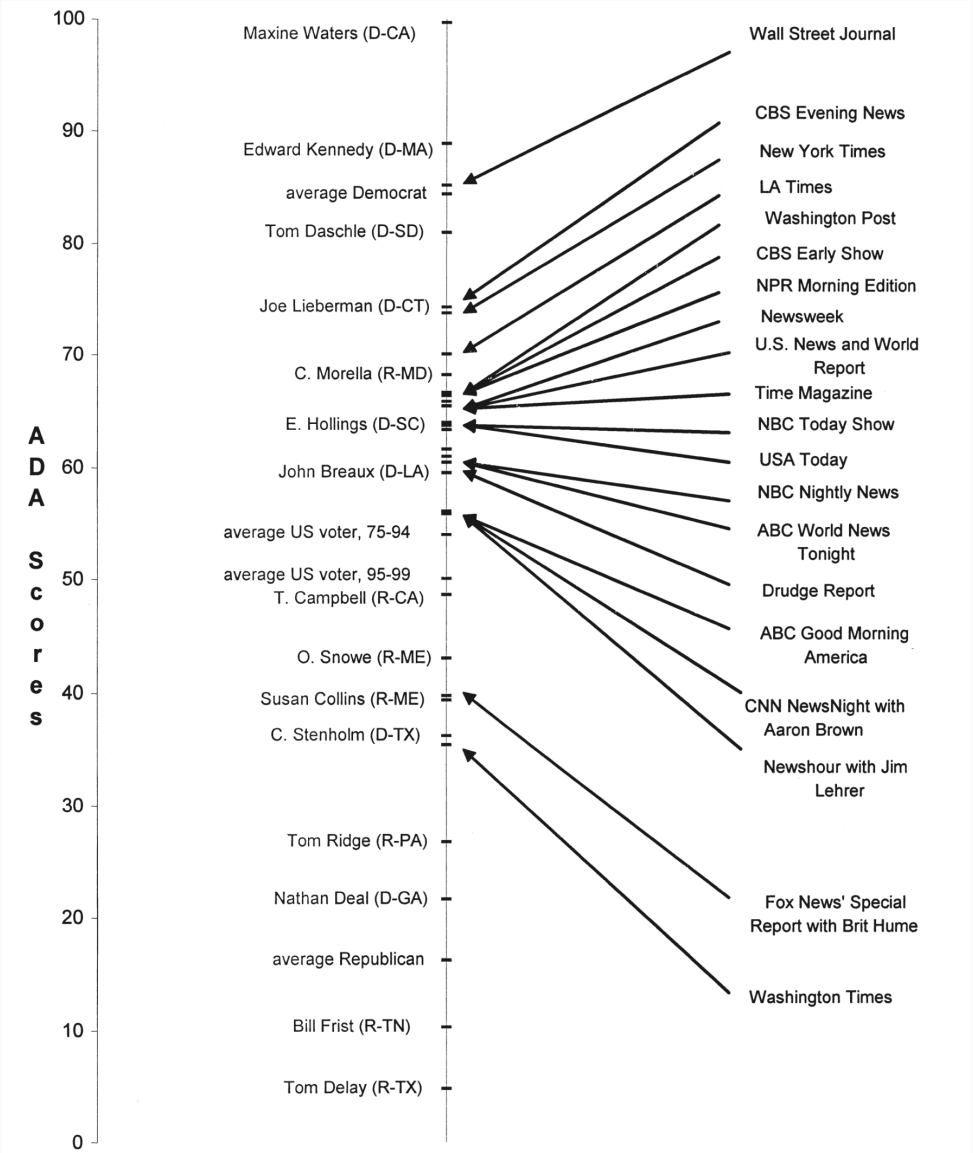


Figure 12: Predicted ADA scores for selected media outlets. Source: Groseclose and Milyo (2005).

Conclusions

Conclusions:

1. media outlets are **relatively centrist**: they are all
   - to the left of the average Republican
   - to the right of the average Democrat (with one exception)

2. ordering of media outlets **matches conventional wisdom**

3. majority of outlets fall to the left of the average US voter

   $\rightarrow$ overall **liberal bias** in the media

Example: What Drives Media Slant?

Gentzkow and Shapiro (2010):

- **predict ADA score of newspapers** ("slant index") similarly as in Groseclose and Milyo (2005), except:
  - for a much larger number of newspapers
  - using all term counts rather than only think tank citations
- estimate **model of newspaper demand**
  - $\rightarrow$ consumer's utility depends on closeness of newspapers' slant index to consumer's own ideology
  - $\rightarrow$ compute each newspaper's slant index if they were profit-maximizing

Findings:

1. profit-maximizing slant close to observed slant index
2. identity of newspapaer owner or party of local incumbent politician do not explain slant

Comparison of Approaches

Dictionary-based methods:

- useful when mapping from text C_i to attribute Y_i is known
- e.g., Baker, Bloom, and Davis (2016)'s definition of economic policy uncertainty

Text regression:

- mapping from text C_i to attribute Y_i not known
- want to predict attributes from the text
- e.g., hedonic pricing as in Bajari, Cen, Chernozhukov, Manukonda, Vijaykumar, Wang, Huerta, Li, Leng, Monokroussos, and Wan (2023)

Generative models:

- model how text is generated $\rightarrow$ full probability model of attributes Y_i and text C_i
- can learn interesting features from distributions of $Y_i|C_i$ and $C_i|Y_i$
- e.g., Groseclose and Milyo (2005) modeling $C_i|Y_i$, but then predicting $Y_i|C_i$

Readings

Recommended readings:

- Gentzkow, Kelly, and Taddy (2019)

Further readings:

- Gentzkow, Shapiro, and Taddy (2019): study partisanship in congressional speeches using Taddy (2015)'s methods
- Egami, Fong, Grimmer, Roberts, and Stewart (2022): causal inference with text as treatment/outcome

References

Bajari, P., Z. Cen, V. Chernozhukov, M. Manukonda, S. Vijaykumar, J. Wang, R. Huerta, J. Li, L. Leng, G. Monokroussos, and S. Wan (2023): "Hedonic Prices and Quality Adjusted Price Indices Powered By AI," Discussion paper.

Baker, S. R., N. Bloom, and S. J. Davis (2016): "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics*, 131(4), 1593–1636.

Bandiera, O., A. Prat, S. Hansen, and R. Sadun (2020): "CEO Behavior and Firm Performance," *Journal of Political Economy*, 128(4), 1325–1369.

Blei, D. M., and J. D. Lafferty (2006): "Dynamic Topic Models," in *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 113–120, New York, NY, USA. Association for Computing Machinery.

Blei, D. M., and J. D. McAuliffe (2010): "Supervised Topic Models," .

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003): "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, 3(null), 993–1022.

Egami, N., C. J. Fong, J. Grimmer, M. E. Roberts, and B. M. Stewart (2022): "How to make causal inferences using texts," *Science Advances*, 8(42), eabg2652.

Gentzkow, M., B. Kelly, and M. Taddy (2019): "Text as Data," *Journal of Economic Literature*, 57(3), 535–74.

References

Gentzkow, M., and J. M. Shapiro (2010): "What Drives Media Slant? Evidence From U.S. Daily Newspapers," *Econometrica*, 78(1), 35–71.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2019): "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech," *Econometrica*, 87(4), 1307–1340.

Groseclose, T., and J. Milyo (2005): "A Measure of Media Bias," *The Quarterly Journal of Economics*, 120(4), 1191–1237.

Hofmann, T. (1999): "Probabilistic Latent Semantic Indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pp. 50–57, New York, NY, USA. Association for Computing Machinery.

Loughran, T., and B. McDonald (2011): "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance*, 66(1), 35–65.

Taddy, M. (2012): "On Estimation and Selection for Topic Models," in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, ed. by N. D. Lawrence, and M. Girolami, vol. 22 of *Proceedings of Machine Learning Research*, pp. 1184–1193, La Palma, Canary Islands. PMLR.

——— (2013): "Multinomial Inverse Regression for Text Analysis," *Journal of the American Statistical Association*, 108(503), 755–770.

References

——— (2015): "Distributed multinomial regression," *The Annals of Applied Statistics*, 9(3), 1394 – 1414.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006): "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101(476), 1566–1581.

Tetlock, P. C. (2007): "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," *The Journal of Finance*, 62(3), 1139–1168.