Deep Latent Variable Models for Unstructured Data

Germain Gauthier, Philine Widmer, Elliott Ash

Summer Institute in Computational Social Science (SICSS)

July 2025

Latent Variables: New Data, New Challenges

- Many latent variable models in the social sciences (e.g., ideal point models based on voting records, latent factors in SVAR models from macro time-series).
- Unstructured data is plentiful and offers new research opportunities (e.g., texts, images, audio, and video recordings).
- BUT unstructured data often means:
  - Large number of observations (***large n***)
    → Common estimation algorithms are intractable!
  - High dimensionality (***large p***)
    → We need embeddings!
  - Multiple modalities (**e.g., text *and* images *and* videos**)
    → We need to accommodate joint representations!

Latent Variables: New Data, New Challenges

- Many latent variable models in the social sciences (e.g., ideal point models based on voting records, latent factors in SVAR models from macro time-series).
- Unstructured data is plentiful and offers new research opportunities (e.g., texts, images, audio, and video recordings).
- BUT unstructured data often means:
  - Large number of observations (***large n***)
    → Common estimation algorithms are intractable!
  - High dimensionality (***large p***)
    → We need embeddings!
  - Multiple modalities (**e.g., text *and* images *and* videos**)
    → We need to accommodate joint representations!

Latent Variables: New Data, New Challenges

- Many latent variable models in the social sciences (e.g., ideal point models based on voting records, latent factors in SVAR models from macro time-series).
- Unstructured data is plentiful and offers new research opportunities (e.g., texts, images, audio, and video recordings).
- BUT unstructured data often means:
  - Large number of observations (***large n***)
    → Common estimation algorithms are intractable!
  - High dimensionality (***large p***)
    → We need embeddings!
  - Multiple modalities (**e.g., text *and* images *and* videos**)
    → We need to accommodate joint representations!

Latent Variables: New Data, New Challenges

- Many latent variable models in the social sciences (e.g., ideal point models based on voting records, latent factors in SVAR models from macro time-series).
- Unstructured data is plentiful and offers new research opportunities (e.g., texts, images, audio, and video recordings).
- BUT unstructured data often means:
  - Large number of observations (***large n***)

    → Common estimation algorithms are intractable!
  - High dimensionality (***large p***)

    → We need embeddings!
  - Multiple modalities (**e.g., text *and* images *and* videos**)

    → We need to accommodate joint representations!

Latent Variables: New Data, New Challenges

- Many latent variable models in the social sciences (e.g., ideal point models based on voting records, latent factors in SVAR models from macro time-series).
- Unstructured data is plentiful and offers new research opportunities (e.g., texts, images, audio, and video recordings).
- BUT unstructured data often means:
  - Large number of observations (***large n***)
    → Common estimation algorithms are intractable!
  - High dimensionality (***large p***)
    → We need embeddings!
  - Multiple modalities (**e.g., text *and* images *and* videos**)
    → We need to accommodate joint representations!

This Paper

- **Deep latent variable models**
  - Fairly general class of latent variable models that nests many special cases
  - Extended regularized autoencoder to approximate the marginal likelihood
  - Fast, scalable, multimodal, allows for covariates and auxiliary outcomes
  - Good finite-sample performance in Monte Carlo simulations
- **Some applications to topic modeling**
  - We introduce the Generalized Topic Model (GTM).
  - We unpack speech polarization in the US Congress.
  - We nowcast macroeconomic outcomes with business news.
  - We analyze written posts and shared images of US politicians on Facebook.

This Paper

- **Deep latent variable models**
  - Fairly general class of latent variable models that nests many special cases
  - Extended regularized autoencoder to approximate the marginal likelihood
  - Fast, scalable, multimodal, allows for covariates and auxiliary outcomes
  - Good finite-sample performance in Monte Carlo simulations
- **Some applications to topic modeling**
  - We introduce the Generalized Topic Model (GTM).
  - We unpack speech polarization in the US Congress.
  - We nowcast macroeconomic outcomes with business news.
  - We analyze written posts and shared images of US politicians on Facebook.

Table of Contents

The Structural Model

- Notations:
  - Z is a vector of latent variables.
  - W_m is a vector of response variables for each modality m.
  - Y is a vector of auxiliary outcomes.
  - P_Z is a prior distribution (modeled as a GLM).
  - X^p, X^c, and X^s are covariates that influence the latent, response or outcome variables.
- We assume that

$$Z \sim P_{Z|X^p} \quad \text{and} \quad W_m = G_m(Z, X^c) \quad \text{and} \quad Y = F(Z, X^s).$$

- This is a fairly general formulation, as P_Z, G_1,...,G_m, and F are left unspecified (we will parametrize them later for estimation).

Table of Contents

Estimation: Overview

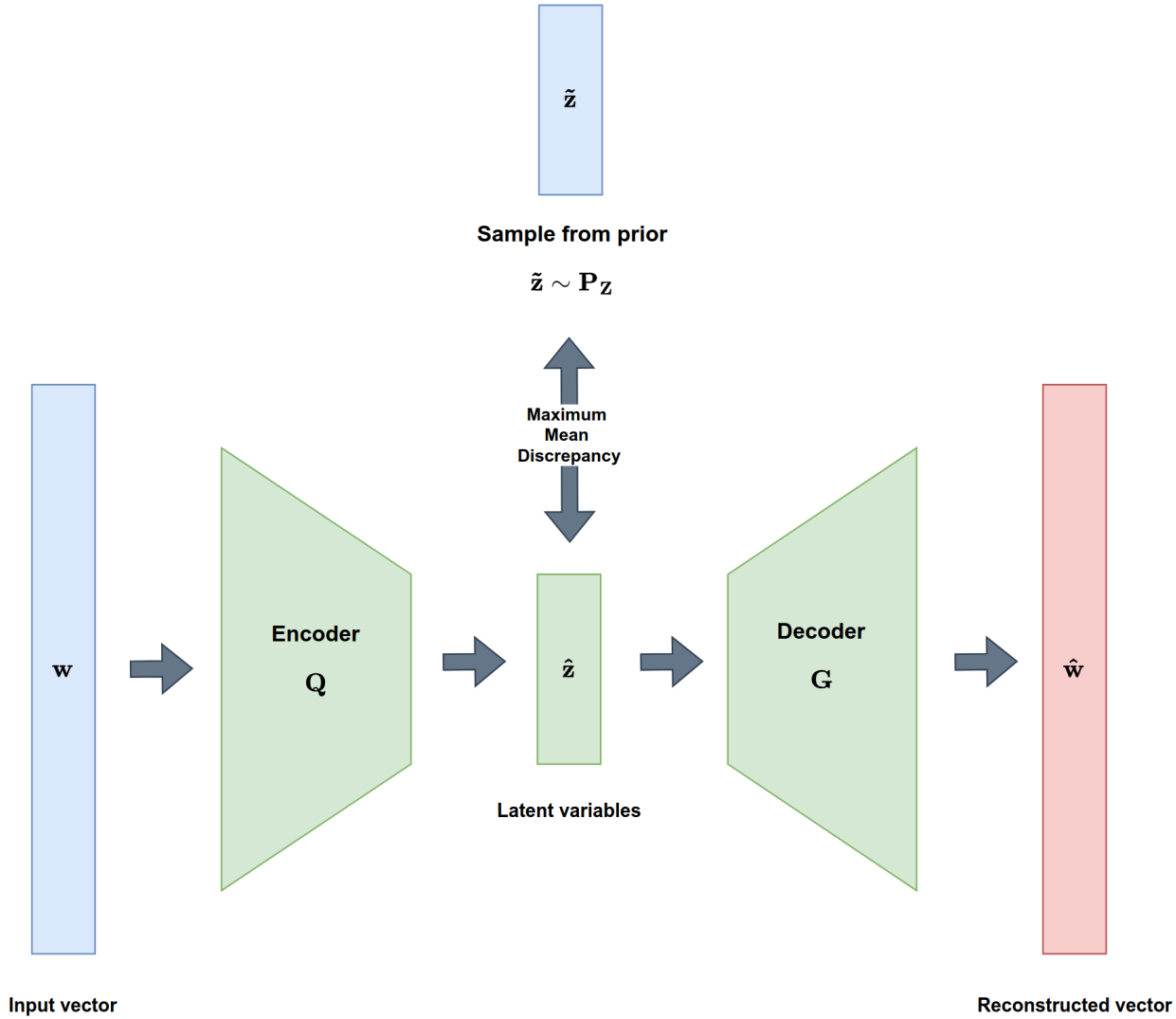
For response variables $\{w_m\} \in \{1, ..., M\}$, outcome y, latent variables z and covariates x^p, x^c, and x^s:

$$P(w_1, ..., w_m, y \mid x^p, x^c, x^s) = \int_{\mathcal{Z}} P(z \mid x^p) P(y \mid z, x^s) \prod_{m=1}^{M} P(w_m \mid z, x^c) dz.$$

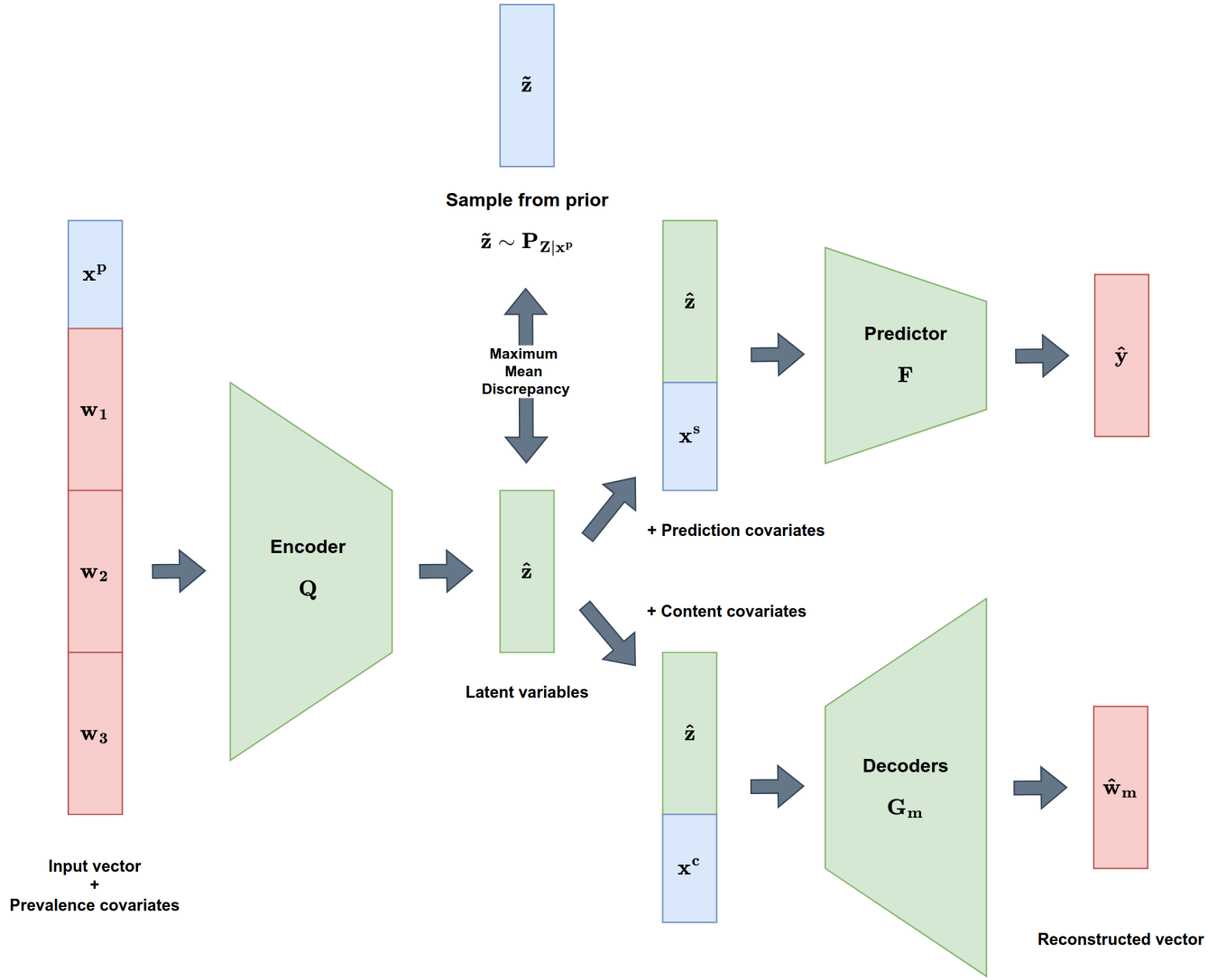
→ **Problem:** The marginal likelihood is *intractable*.

→ **Solution:** Approximate $P(w_1, ..., w_m \mid x^p, x^c)$ as a regularized autoencoder:

- Encoder $Q(w_1, ..., w_m, x^p) \approx P(z \mid w_1, ..., w_m, x^p)$
- Decoders $G_m(z, x^c) \approx P(w_m \mid z, x^c)$
- Predictor $F(z, x^s) \approx P(y \mid z, x^s)$
- We nudge $Q(w_1, ..., w_m, x^p)$ to remain close to the prior distribution.

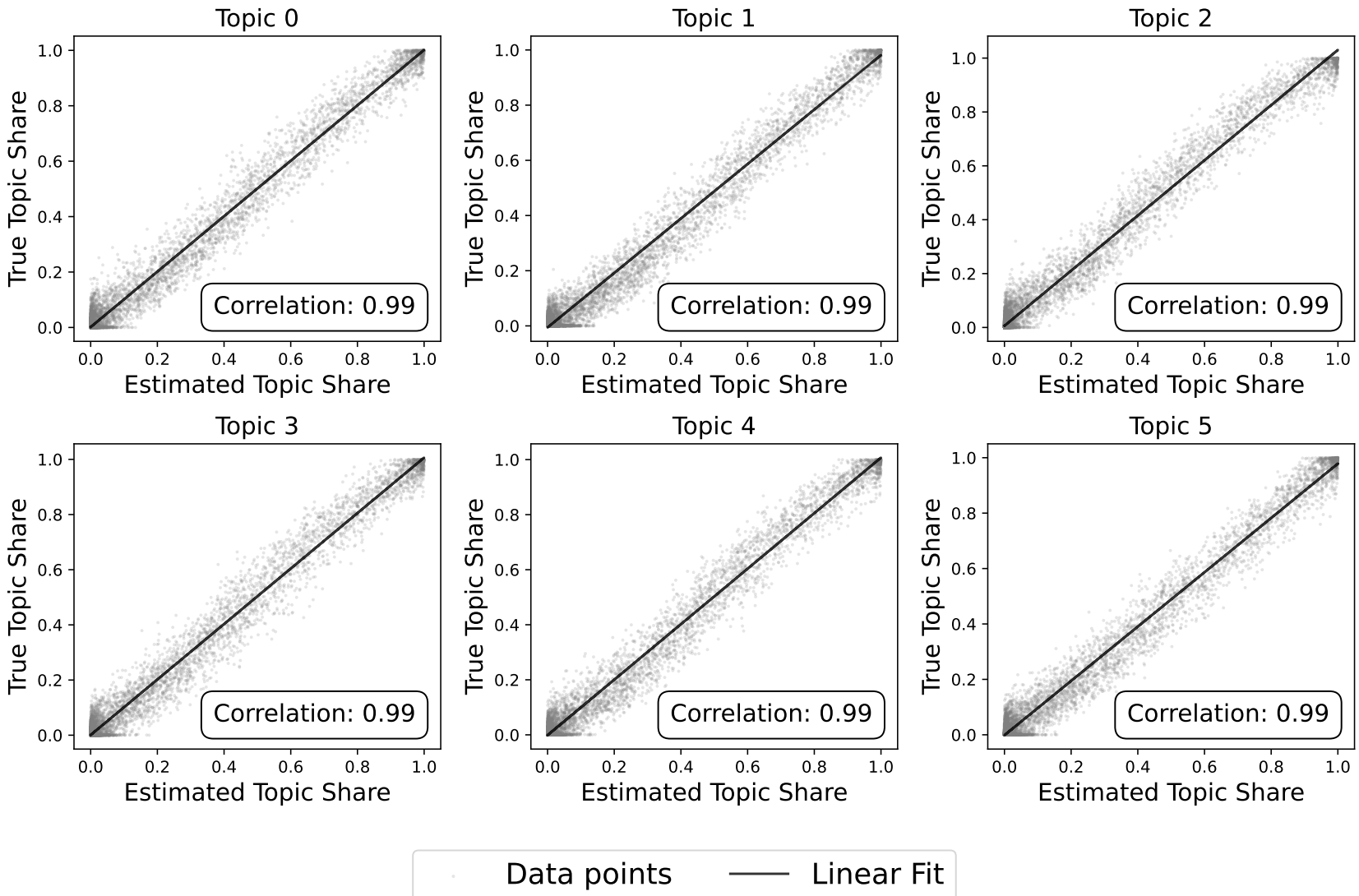


$$\textbf{TotalLoss} = \textbf{ReconstructionLoss}(w, \hat{w}) + \lambda \widehat{\text{MMD}}_k^2(\hat{z}, \tilde{z})$$

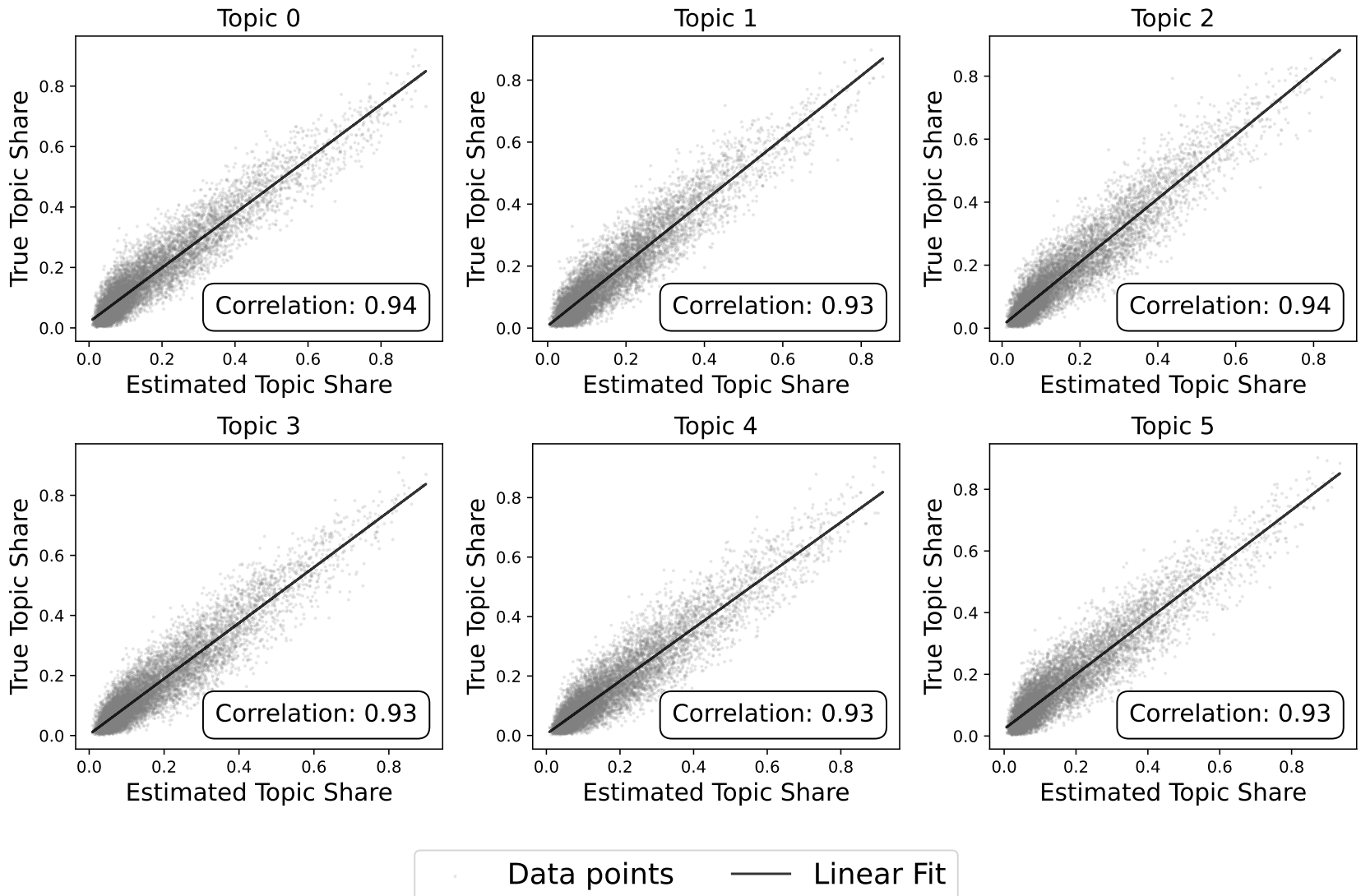


$$\text{TotalLoss} = \sum_{m=1}^{M} \text{RecLoss}(w_m, \hat{w}_m) + \lambda_0 \widehat{\text{MMD}}_k^2(\hat{z}, \tilde{z}) + \lambda_1 \text{PredLoss}(y, \hat{y})$$

Monte Carlo Simulations: Dirichlet Prior



Monte Carlo Simulations: Logistic Normal Prior



Main Advantages

- Very fast, modular, and scalable approach
- Incorporates covariates and outcomes

  → Do causal inference or supervised learning
- Learns embeddings and/or can take external embeddings as input

  → Learn complex patterns and build on top of pre-trained models
- Amortized inference

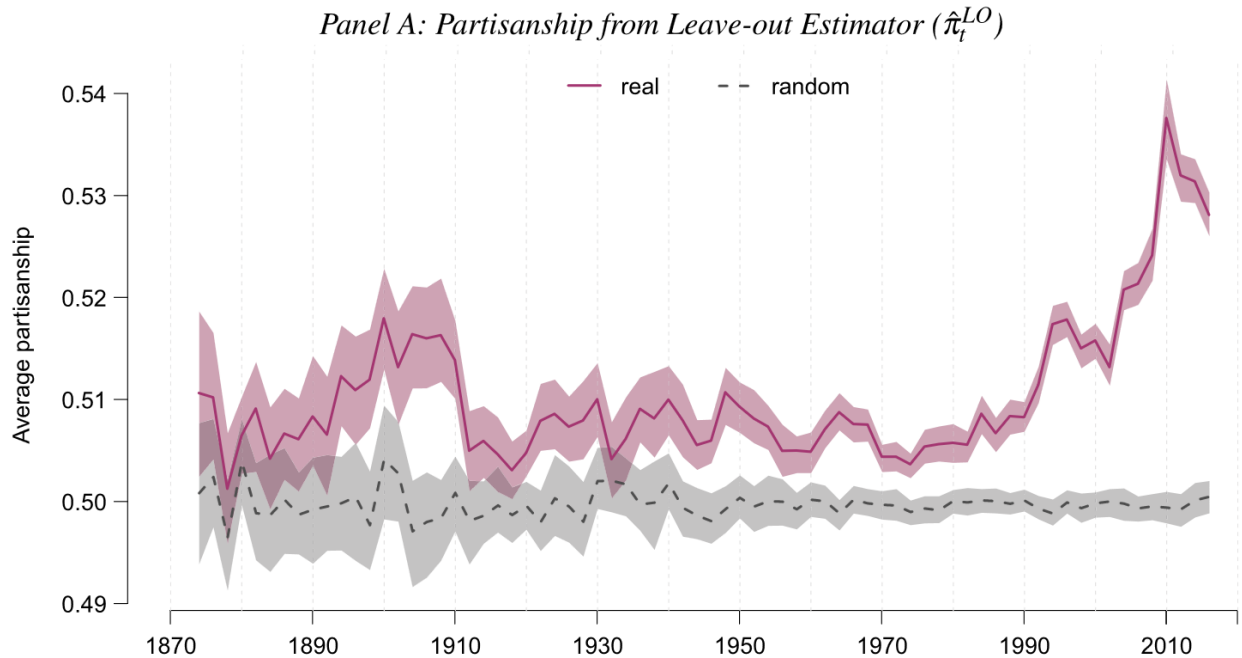
  → Train on a subsample of the data and scale out-of-sample in seconds
- Transfer learning capabilities

  → Train on texts and scale out-of-sample to images and videos in seconds

Table of Contents

Application 1: Speech Polarization in the US Congress

The increase in speech polarization in US politics is now a well-established stylized fact (Gentzkow, Shapiro, & Taddy, 2019).

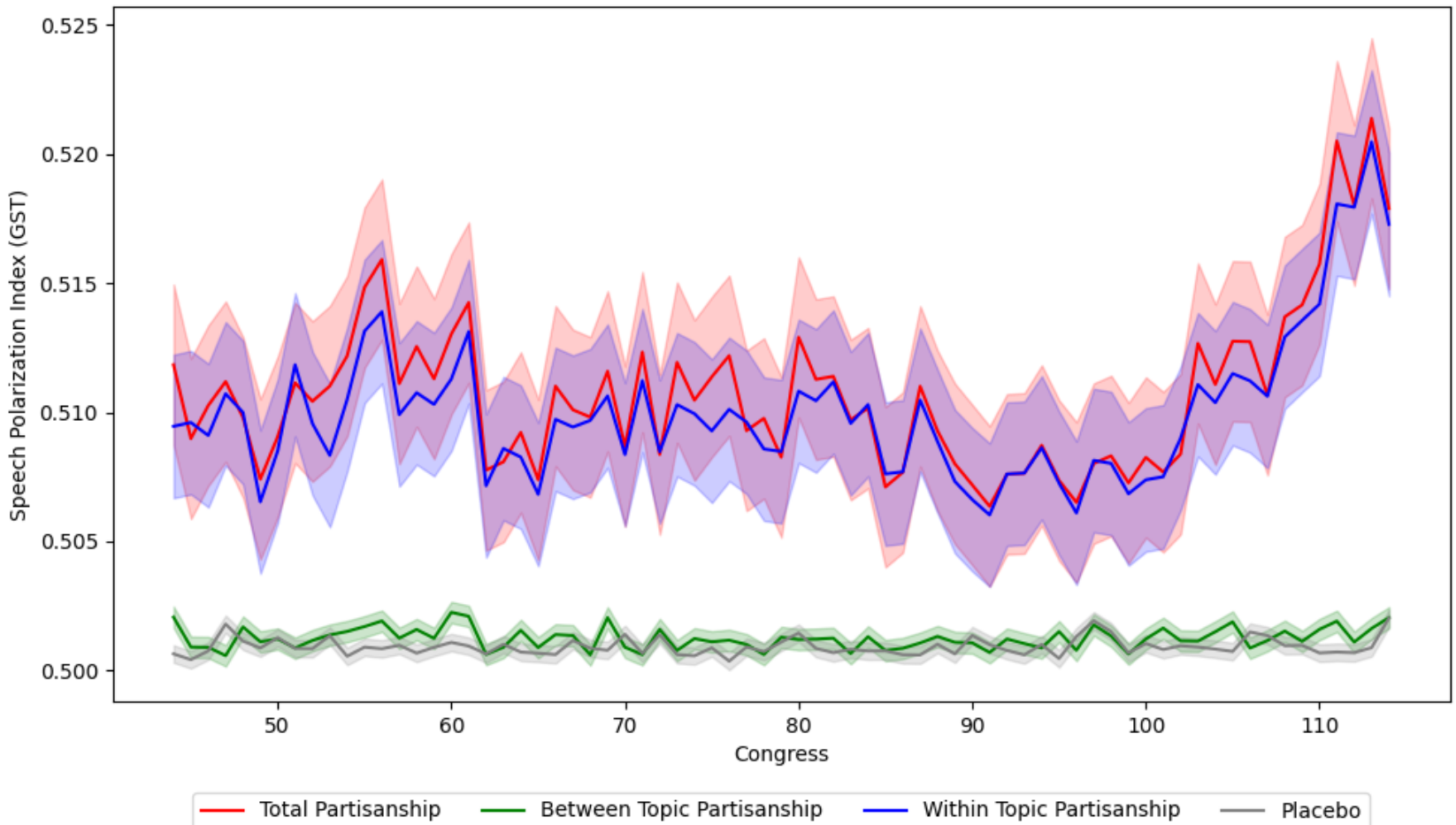


→ Is this increase driven by topics or slant?

Proposed (Unsupervised) Decomposition Methodology

- We estimate the GTM on the full *US Congressional Record* (1873 – 2016).
- We allow party × session covariates to affect topic choice *and* content.
- We use our model of speech to generate samples of partisan speeches.
- We simulate 1000 typical speeches for each party × session under four distinct scenarios:
  - **Total polarization (observed):** Republicans and Democrats differ in topic choice and topical content.
  - **Between-topic polarization (topics):** Republicans and Democrats differ in topic choice but not in topical content.
  - **Within-topic polarization (slant):** Republicans and Democrats differ in topical content but not in topic choice.
  - **Placebo:** Republicans and Democrats do not differ in either dimension.
- We compute a speech polarization index per session and scenario. ▸ Technical Details

Slant matters (much) more than topic choice.



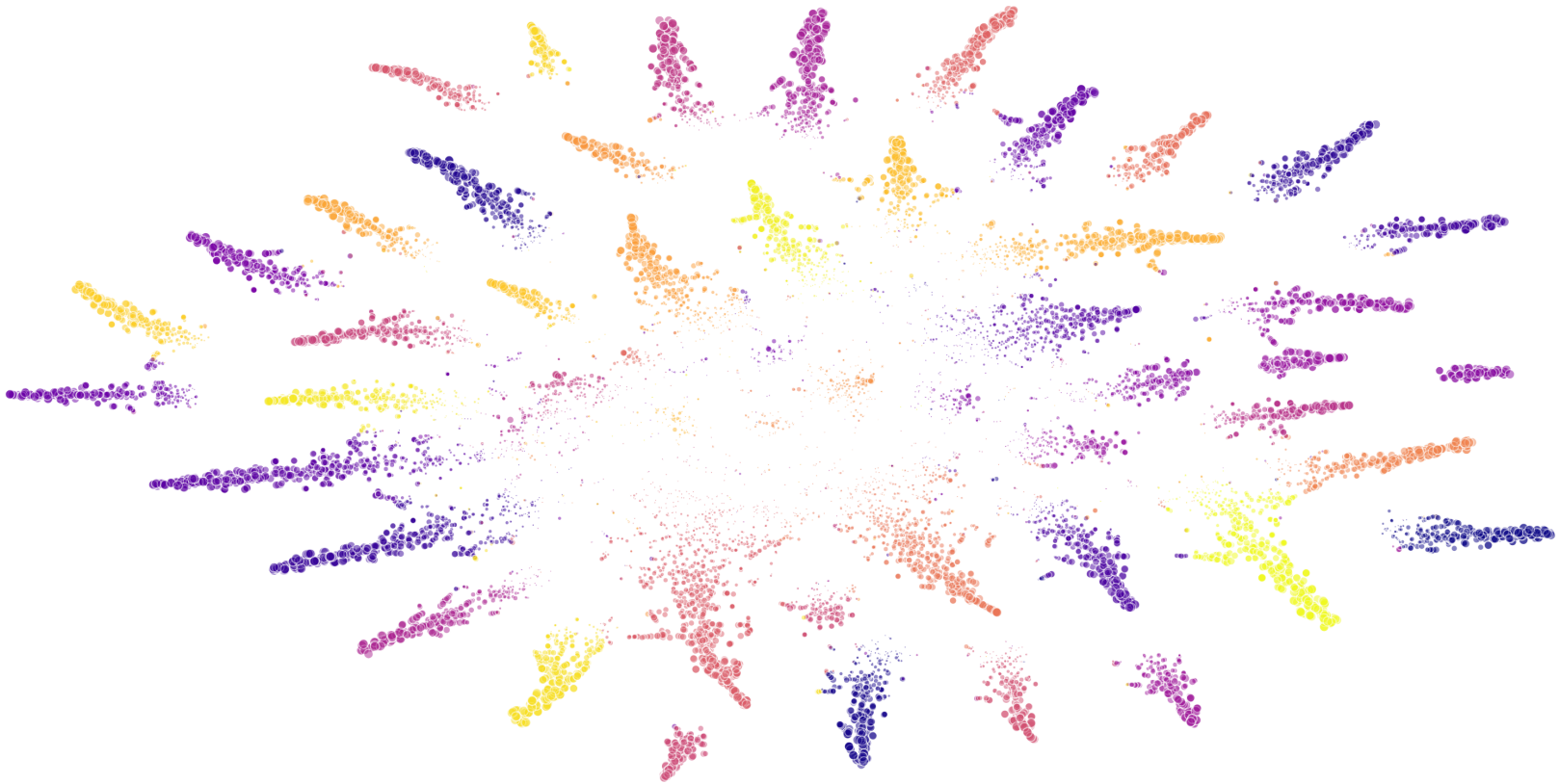
The Topics

Topic Label	Top Words
Welfare and Unemployment	food stamp, unemployment benefit, welfare reform, money borrow, poor people, unemployment
Military Weapons	missile, ballistic missile, ballistic, submarine, silo, cruiser
Education	pupil, school district, private school, vocational education, school, local school district
Transportation and Infrastructure	passenger, airport, highway trust fund, highway trust, rail, mail service
Human Rights	captive nation, polish people, genocide, human right, human right violation, anniversary independence
Health Insurance	private insurance, health insurance, health plan, insurance company, premium, hospital care
Budget and Appropriations	budget authority outlay, authority outlay, budget authority, amount recommend, budget estimate
Land Entitlement	vessel, act entitle act, certain land, act entitle, subdivision, entitle act
Condolences and Tribute	condolence, mourn, untimely death, eulogy, passing, sense humor
Water and Navigation	navigable, nuclear waste, dam, flood control, basin, navigation
Judiciary and Trials	jury trial, jury, trial jury, grand jury, district judge, defendant
Trade and Industry	build ship, freedom independence, coastwise, farmer buy, cut taxis, coal miner
Scientific Research	space station, space, shuttle, scientist engineer, basic research, library
Postal Services	cent pound cent, pound cent, cent pound, postal employee, pound, advertising
Pensions and Social Security	grant pension, pension, increase pension, widow, social security benefit, retirement system
Banking and Finance	commercial bank, bank, national bank, silver dollar, loan association, depositor
Banking Regulation	hold company, national bank, banking, bank bank, bank, loan association
Foreign Aid	foreign aid, military assistance, military aid, military budget, economic aid, aid
Agriculture and Farming	wheat, bushel, feed grain, cent bushel, percent parity, bale
Elections and Voting	elector, ballot, vacancy, suffrage, legislative power, fill vacancy
Tariffs and Public Service	protective tariff, law degree, career public, pastor, treasurer, farm organization
Taxation	sale tax, property tax, tax income, property taxes, tax return, local taxes
Trade and Textile	trade agreement, textile, fair trade, free trade, textile industry, export
Public Buildings and Construction	circuit judge, cubic, public building, lieutenant, cadet, erection
Military Service and Veterans	active duty, veteran, disabled veteran, enlisted, reservist, enlisted man

The Topics (continued)

Topic Label	Top Words
Intelligence and Investigation	subpoena, grand jury, impeachment, investigation, intelligence agency, intelligence
Campaign Finance	spending limit, campaign finance, electoral, finance reform, candidate, campaign finance reform
Labor Market	wage rate, minimum wage, increase minimum wage, wage worker, cent hour, farm labor
Nuclear Treaties	nuclear test, nuclear arm, treaty, arm control, nuclear weapon, treaty treaty
Oil and Energy	domestic oil, barrel day, barrel oil, crude oil, oil, barrel
Public Health Risks	quarantine, ordnance, beverage, follow resolve, act entitle act, peace security
National Forests	national forest, wilderness, acre, wilderness area, grazing, timber
Legislative Process	unfinished business, divide control, general appropriation, unfinished, open hour, complete action
Energy Policy	change exist law, free election, national energy policy, testimony hearing, change exist, debatable
Capital Gains and Taxes	capital gain, gain tax, capital gain tax, tax rate, surtax, estate tax
Military Construction	military assistance, military construction, new project, increase pension, complete project, river harbor
Judicial and Highway Funding	district attorney, highway trust fund, highway trust, judgeship, parking, controller
Fisheries and Maritime Activities	fishery, coastwise, merchant marine, vessel, fishing, canal
Legal Claims	claimant, pay claim, claim claim, attorney fee, statute limitation, claim pay
Public Health	home health, disease, heart disease, mental health, health center, dental
Housing and Urban Development	affordable housing, public housing, urban renewal, housing unit, housing, voucher
Law Enforcement	law enforcement officer, enforcement officer, absent, pair, prisoner war, enlisted man
Forest Conservation	forest, rainfall, pest, timber, insect, acre
Natural Disasters	earthquake, evacuate, parking, evacuation, flooding, fire department
Crime and Justice	juvenile, crime, deport, firearm, violent crime, handgun
Intellectual Property and Commerce	copyright, patent, common carrier, interstate commerce, plaintiff, antitrust
War Negotiations	military aid, aggression, embargo, peace process, troop, peace
Sports and Social Issues	championship, colored, basketball, colored people, abortion, coach
Agricultural Commodities	butter, wool, pesticide, cent pound, cent pound cent, cent ad
Debt and Budget	debt limit, increase debt, public debt, balance budget, debt increase, interest debt

Visualizing the Topics



Topical Content Over Time

Content Associated with Covariates: Distinctive Words Over Time

Year of Congress	Top Words
Intercept	country, people, great, man, law, give
1917	conscription, prosecution war, win war, war measure, profiteer, present war
1919	profiteer, league, covenant, armistice, federal control, packer
1921	profiteer, armistice, prewar, bloc, tile, criticize
1939	conscription, peacetime, national income, totalitarian, setup, airplane
1941	war effort, nondefense, defense industry, sabotage, defense effort, rubber
1943	war effort, postwar, parity price, win war, directive, print record part
1945	decontrol, atomic bomb, postwar, war effort, atomic energy, full employment
1947	atomic energy, atomic bomb, atomic, free enterprise, decontrol, ideology
1973	energy crisis, mass transit, legal service, energy problem, impact statement, environmental
1977	energy problem, oversight hearing, economic stimulus, cruise missile, rate inflation, move opposition
1979	energy problem, oversight hearing, windfall profit, food stamp, inflation rate, auto industry
1981	regulatory reform, block grant, social security system, oversight hearing, auto industry, deferral
1993	health care reform, care reform, unfunded mandate, space station, deficit reduction, stimulus package
2001	war terrorism, homeland security, terrorist attack, economic stimulus, stimulus package, weapon mass destruction
2003	homeland security, war terror, war terrorism, terrorist attack, weapon mass destruction, terrorist
2005	war terror, homeland security, web site, border security, war terrorism, transparency
2007	transparency, war terror, homeland security, ensure, work bipartisan, urge reserve balance
2009	transparency, ensure, web site, stakeholder, job creation, health care reform
2011	transparency, job creation, job growth, infrastructure, ensure, bipartisan agreement

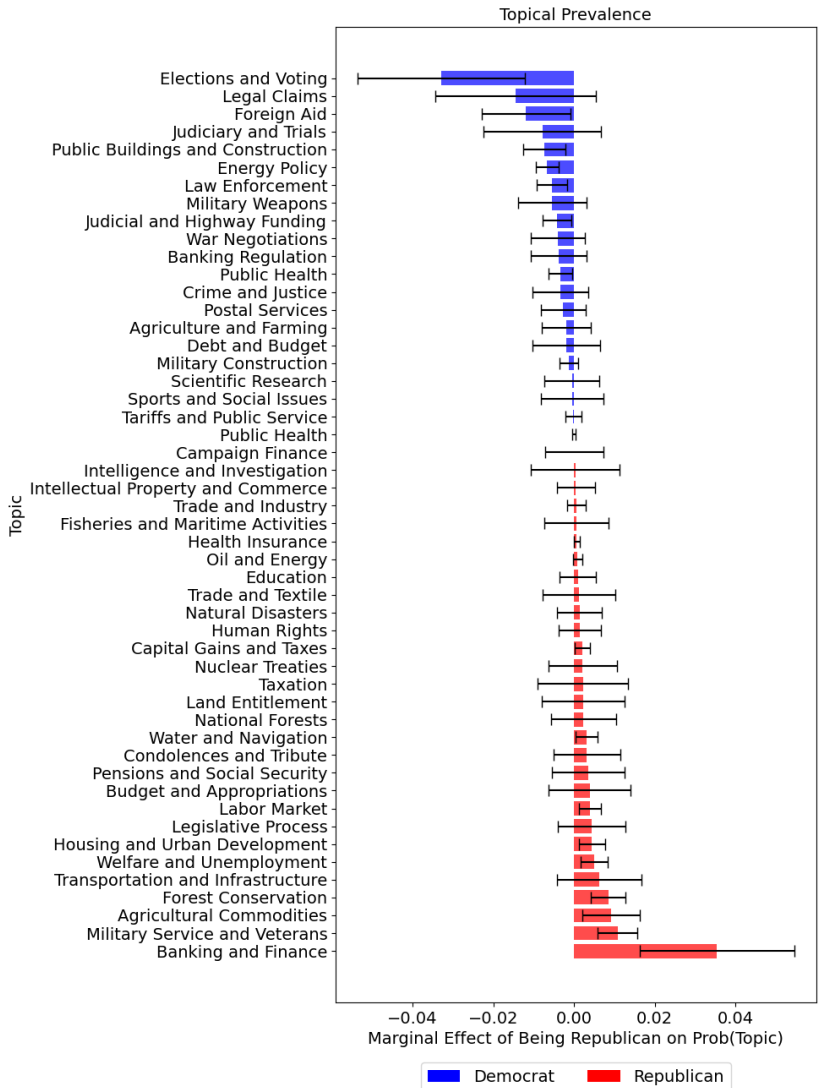
Learnt Word Embeddings

Most Similar Phrases in Embeddings Space for:			
taxation	**immigration**	**war**	**health**
levy tax	immigration reform	military	health care
burden taxation	immigration law	troop	medical
taxes impose	immigrant come	soldier	care
tax income	illegal immigration	war war	disease
tax power	deport	country	cancer
tax cent	immigrant	peace	drug
amount revenue	undocumented	man	hospital
taxes levy	deportation	nation	child
wage	**election**	**regulation**	**fiscal**
minimum wage	election hold	law	budget
wage increase	ballot	regulate	appropriation
increase minimum wage	election election	regulatory	fund
increase minimum	election law	federal regulation	fiscal fiscal
raise minimum	day election	propose regulation	authorization
wage worker	candidate	new regulation	increase
labor	general election	law regulation	expenditure
wage hour	voter	company	fiscal budget

Learnt Word Embeddings (continued)

Most Similar Phrases in Embeddings Space for:			
debt	**deficit**	**campaign**	**insurance**
national debt	deficit reduction	candidate	insurance company
public debt	budget deficit	campaign contribution	private insurance
debt debt	reduce deficit	campaign finance	insurance industry
pay debt	deficit problem	election	insurance coverage
debt pay	cut deficit	presidential campaign	insurance premium
debt limit	deficit deficit	election campaign	insurer
increase debt	federal deficit	political campaign	premium
pay interest	increase deficit	editorial	insured
recession	**growth**	**welfare**	**crime**
recession	economic growth	welfare reform	fight crime
unemployment benefit	growth rate	welfare system	violent crime
unemployed worker	growth economy	people welfare	commit crime
economic stimulus	rate growth	welfare people	crime commit
people unemployed	economic	welfare country	gun control
jobless	job creation	dependency	criminal justice
unemployment percent	economy grow	child care	crime rate
unemployment insurance	capital	public welfare	murder

Topical Choice and Partisanship

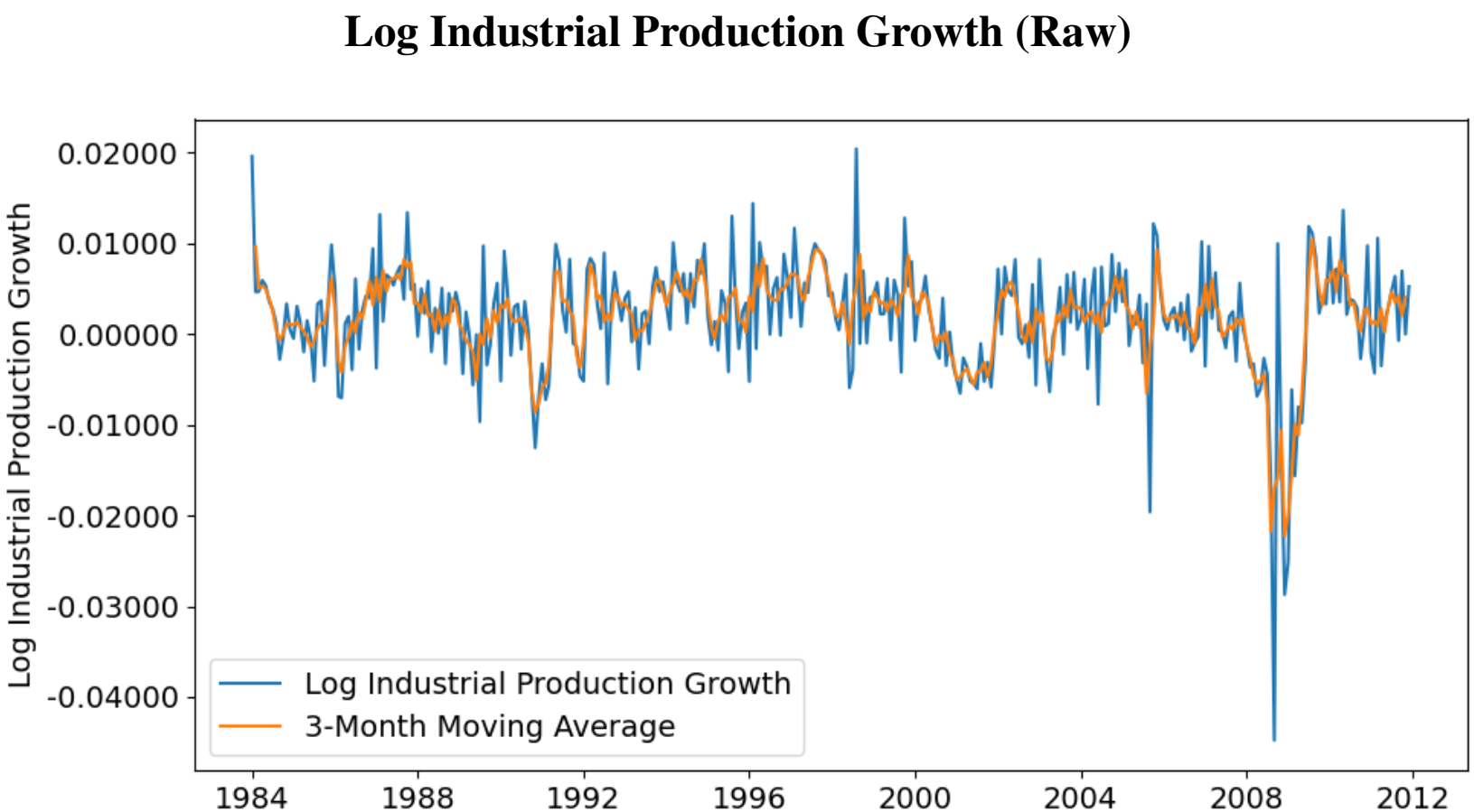


Topical Content and Partisanship

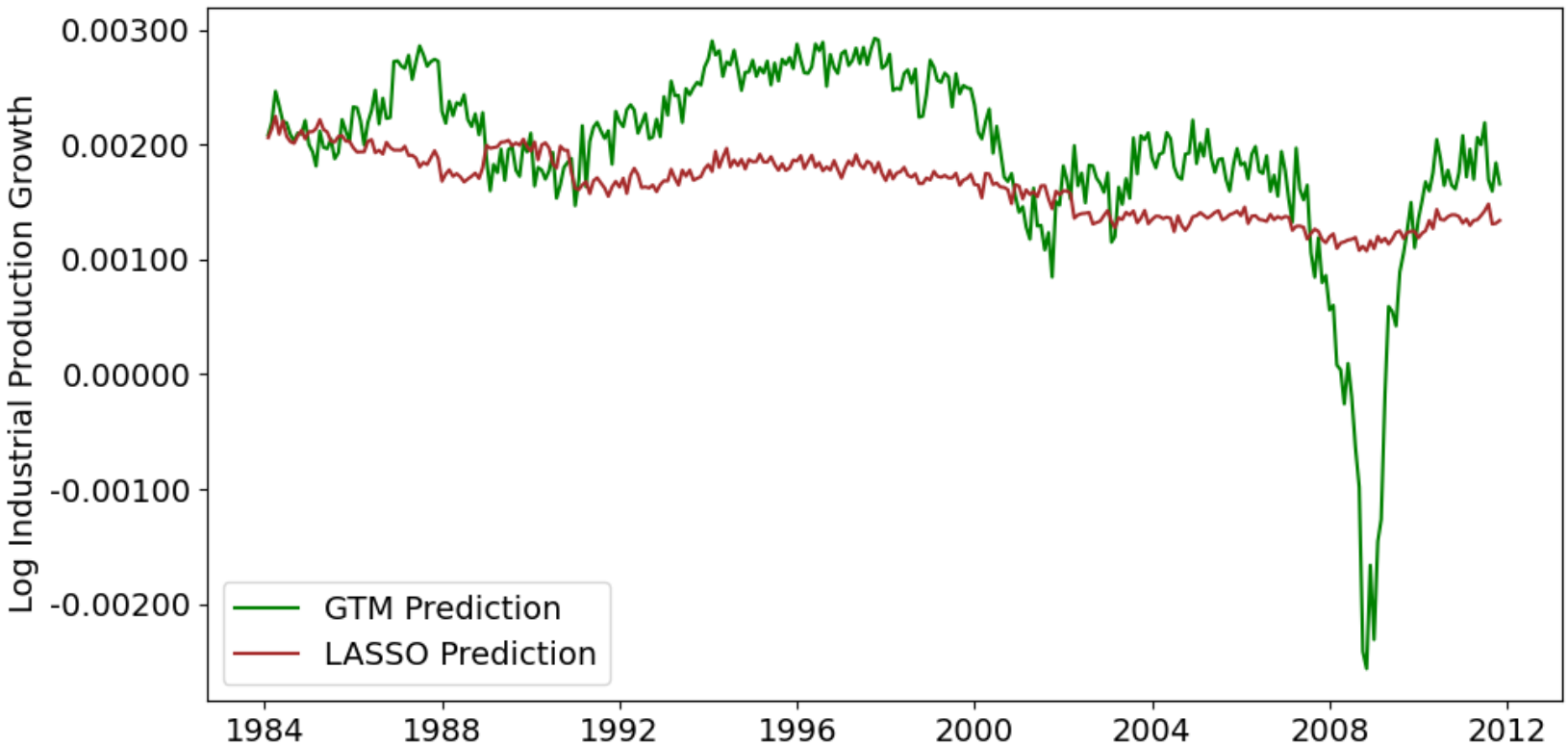
Topic-Covariate Interactions: Party-Specific Topic Language

Topic Label	Type	Top Words
Welfare and Unemployment	Baseline	food stamp, unemployment benefit, welfare reform, money borrow, poor people, unemployment
	Republicans	money, man, pay, people, pension, come
	Democrats	unemployment, poor people, poverty, hungry, food stamp, unemployed
Elections and Voting	Baseline	elector, ballot, vacancy, suffrage, legislative power, fill vacancy
	Republicans	precinct, election, contestant, elector, ballot, seat
	Democrats	civil right, racial, equal protection, segregation, equal protection law, filibuster
Military Service and Veterans	Baseline	active duty, veteran, disabled veteran, enlisted, reservist, enlisted man
	Republicans	officer, man, soldier, enlistment, regiment, enlisted man
	Democrats	veteran, disabled veteran, personnel, care veteran, veteran family, veteran benefit
Campaign Finance	Baseline	spending limit, campaign finance, electoral, finance reform, candidate, campaign finance reform
	Republicans	election, candidate, party, money, postage, expenditure
	Democrats	campaign finance, spending limit, finance reform, campaign finance reform, campaign contribution
Labor Market	Baseline	wage rate, minimum wage, increase minimum wage, wage worker, cent hour, farm labor
	Republicans	laborer, labor, man, railroad, workman, railway
	Democrats	minimum wage, wage worker, worker, increase minimum wage, cent hour, wage
Crime and Justice	Baseline	juvenile, crime, deport, firearm, violent crime, handgun
	Republicans	officer, liquor, alien, consular, gun, intoxicate
	Democrats	juvenile, law enforcement, drug, enforcement, crime, criminal justice

Application 2: Nowcasting Macroeconomic Indicators



Predictions Based on Topics from the *Wall Street Journal*



Application 3: US Politicians on Facebook

- We collect the universe of posts written and images shared by US Congress representatives on Facebook between 2010 and 2020.
- We train the GTM on texts only using embeddings that handle texts and images (LongCLIP) as input, and try to reconstruct the bag of words as output.
- The model can then predict topics out-of-sample for texts and images.
- Some open questions:
  - Are texts or images more polarized?
  - How do image and text topics correlate within a post?
  - Are some topics more likely to be addressed with images?

The 20 Topics

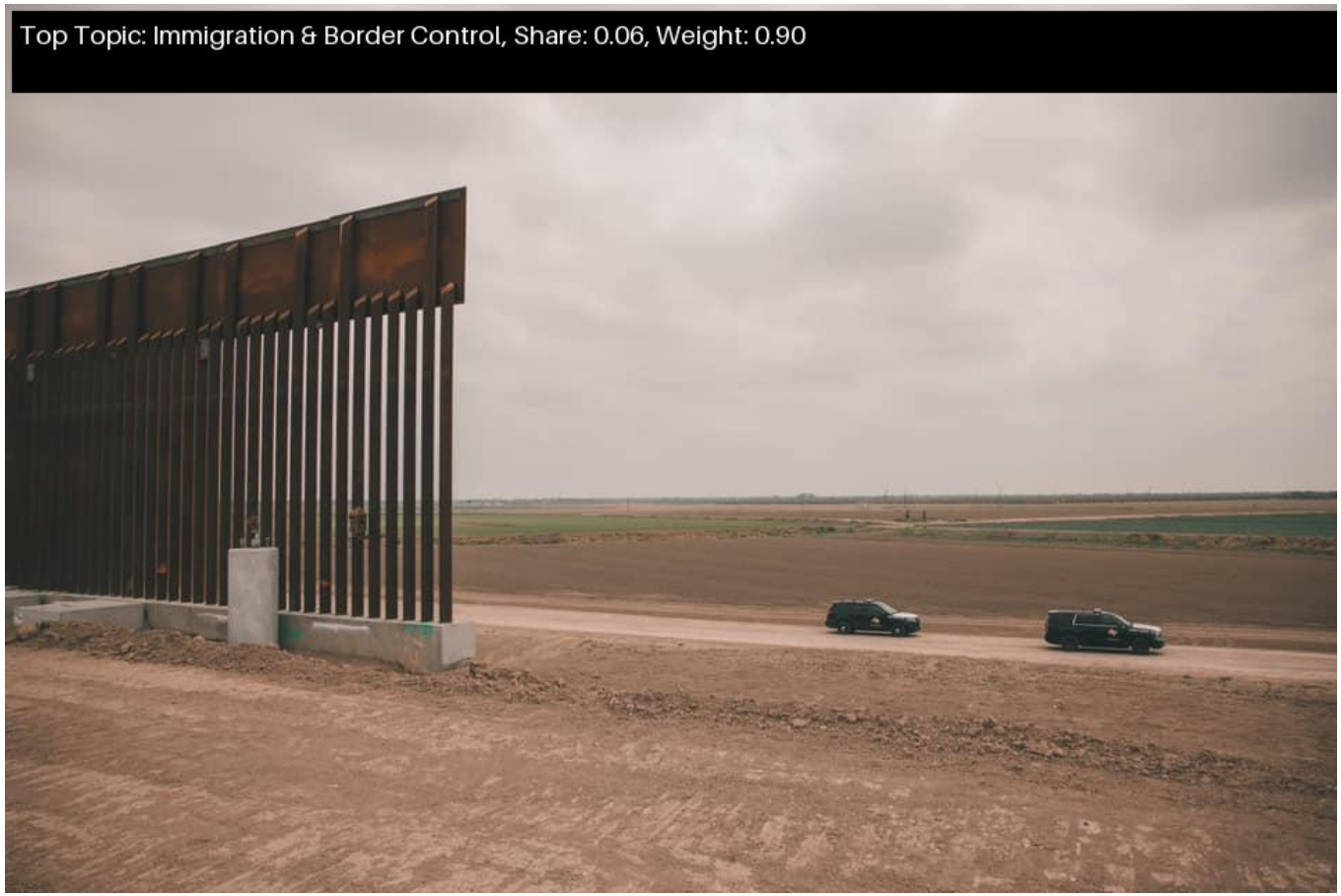
Label	**Representative Words**
Sports and Competitions	game, win, team, football, fun
Holidays and Awards	wish, family, love, present, hope
Voting System	vote, election, democracy, ballot, state
Military Honors	brave, medal, hero, military, present
Gun Violence	gun, violence, shooting, action, border
COVID-19 Health Services	vaccine, test, health, covid, site
Natural Disasters	storm, fire, damage, evacuation, area
Clean Energy & Infrastructure	energy, infrastructure, clean, water, project
Campaign Volunteering & Outreach	volunteer, campaign, update, constituent, folk
Tragedy & Mourning	lose, prayer, love, fight, attack
Women's Rights	woman, equality, vote, fight, legacy
Voting Encouragement	vote, early, polling, ballot, location
Taxes	tax, cost, insurance, bill, plan
Business Openings	facility, manufacturing, opening, company, economy
Public Health	health, funding, flood, fire, recovery
Military Sacrifice & Remembrance	sacrifice, hero, brave, forget, woman
Economy	economy, tax, growth, cost, american
Farming & Small Business	farmer, product, company, grow, industry
Schools	teacher, school, education, class, young
Immigration & Border Control	border, immigrant, illegal, crisis, southern

Some examples: Military Honors

Some examples: Business Openings

Some examples: Schools

Some examples: Immigration and Border Controls



Concluding Remarks

We are preparing an open-source Python package `generalized_topic_models` to support future applications

...

and we look forward to your feedback.

Thanks for listening!

Deep Latent Variable Models for Unstructured Data

Germain Gauthier, Philine Widmer, Elliott Ash

Summer Institute in Computational Social Science (SICSS)

July 2025