Large Language Models and Machine Learning for Unstructured Data

Lecture 4: Inference in Downstream Regression

Stephen Hansen
University College London

FUNDACIÓN RAMÓN ARECES

IESE Business School University of Navarra Center for International Finance

Text-as-Data Setup in Economics

1. z: latent variable of economic interest, e.g. policy uncertainty
2. x: text data, e.g. newspapers
3. y: outcome data, e.g. aggregate output

Ideal approach is to model y as a function of z.

Typical approach is to (i) use x to create proxy measure z'; (ii) model y as a function of z'.

Two Questions

1. How sensitive is downstream inference to choice of upstream model?

2. How is downstream inference affected by separation of (i) and (ii)?

Choosing Among Algorithms

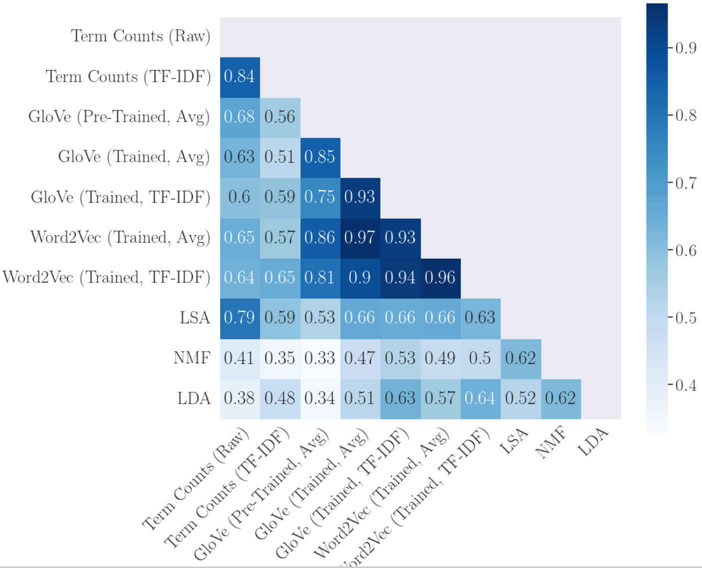
Multiple algorithms for document similarity: bag-of-words, word2vec, BERT, etc.

No clear metric of how to judge which is best and human labeling is hard.

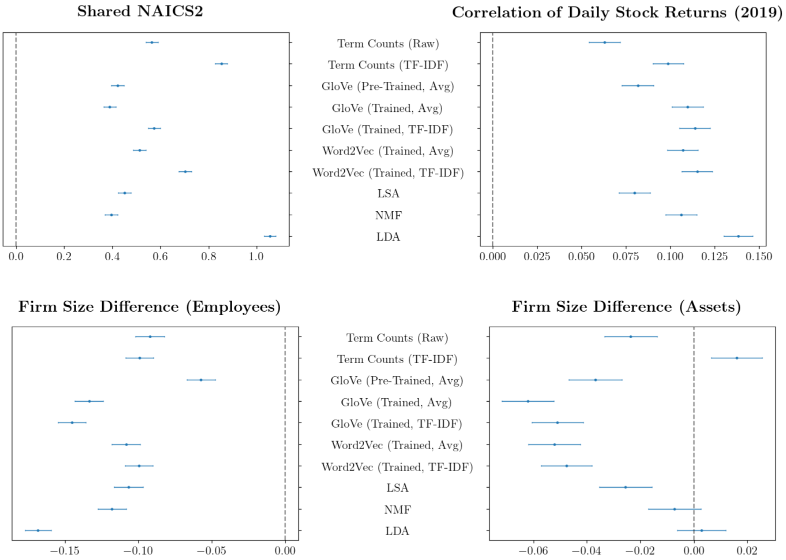
We compute document similarity in the context of 10-K risk factors using randomly sampled pairs from the universe of 2019 filing firms.

Keep data constant, and vary the algorithm used for similarity comparison.

Pearson Correlation Between Similarity Scores Across Pairs



Downstream Regression Results



Inference for Regression with Variables Generated from Unstructured Data

Laura Battaglia
Oxford

Tim Christensen
UCL

Stephen Hansen
UCL, IFS, and CEPR

Szymon Sacher
Stanford

June 19, 2024

Stylized Model

The stylized model is based on Baker, Bloom and Davis (2016).

We are interested in effect of *policy uncertainty*, θ_i, on *investment*, Y_i. Assume that the relationship is linear:

$$Y_i = \gamma_0 + \gamma_1 \theta_i + \varepsilon_i. \quad (1)$$

Stylized Model

The stylized model is based on Baker, Bloom and Davis (2016).

We are interested in effect of *policy uncertainty*, θ_i, on *investment*, Y_i. Assume that the relationship is linear:

$$Y_i = \gamma_0 + \gamma_1 \theta_i + \varepsilon_i. \tag{1}$$

The policy uncertainty is unobserved, but we have access to text of the articles from 10 major newspapers.

BBD count the number of articles, X_i that contain certain words, and use X_i / C_i as a proxy for θ_i, where C_i is the total number of articles.

Stylized Model II

The measure X_i/C_i is related to the concept of *policy uncertainty* but is it the same?

- What if a different set of words is used?
- What if different newspapers are used?

Stylized Model II

The measure X_i/C_i is related to the concept of *policy uncertainty* but is it the same?

- What if a different set of words is used?
- What if different newspapers are used?

We propose that the following specification:

$$X_i \sim \text{Binomial}(C_i, \theta_i) \tag{2}$$

Note that:

- θ_i is the true policy uncertainty; $\hat{\theta}_i = X_i/C_i$ is the MLE of θ_i

Two-step Estimation

The usual strategy to estimate, γ_1, the effect of policy uncertainty on investment is to:

1. Estimate θ_i using $\hat{\theta}_i = X_i / C_i$.
2. Estimate γ_1 using $\hat{\theta}_i$ as a proxy for θ_i with OLS.

Two-step Estimation

The usual strategy to estimate, γ_1, the effect of policy uncertainty on investment is to:

1. Estimate θ_i using $\hat{\theta}_i = X_i / C_i$.
2. Estimate γ_1 using $\hat{\theta}_i$ as a proxy for θ_i with OLS.

This approach overlooks the fact that $\hat{\theta}_i$ is a noisy proxy for θ_i, which may lead to:

- *Attenuation bias* in the estimate of γ_1.
- *Incorrect standard errors* of γ_1.

Asymptotic Behavior I

Under standard assumptions on $(Y_i, X_i, C_i, \varepsilon_i)$, it's easy to show that as $n \to \infty$:

$$\hat{\gamma}_1 \to_p \gamma_1 \frac{\mathrm{Cov}(\theta_i, \hat{\theta}_i)}{\mathrm{Var}(\hat{\theta}_i)} = \gamma_1 \frac{\mathrm{Var}(\theta_i)}{\mathrm{Var}(\theta_i) + \mathbb{E}\left[C_i^{-1}\right] \mathbb{E}\left[\theta_i(1-\theta_i)\right]}$$

Then, if the number of articles is large, so $E[C_i^{-1}]$ is small, we have

$$\mathrm{plim}(\hat{\gamma}_1) \approx \gamma_1 - \mathbb{E}\left[\frac{1}{C_i}\right] \frac{\mathbb{E}\left[\theta_i(1-\theta_i)\right]}{\mathrm{Var}(\theta_i)} \gamma_1$$

Evidently, the estimate is biased and the bias is decreasing in the amount of unstructured data per observation, C_i.

Asymptotic Behavior II

Consider the *drifting sequence* where $\sqrt{n} \times \mathbb{E}\left[\frac{1}{C_i}\right] \to \kappa$

Proposition 1

$$\sqrt{n}(\hat{\gamma}_1 - \gamma_1) \to_d N\left(-\kappa\,\gamma_1 \frac{\mathbb{E}[\theta_i(1-\theta_i)]}{\mathrm{Var}(\theta_i)}, \frac{\mathbb{E}[\varepsilon_i^2(\theta_i - \mathbb{E}[\theta_i])^2]}{\mathrm{Var}(\theta_i)^2}\right).$$

- κ governs relative importance of sampling error vs measurement error
- If $\kappa = 0$, can ignore measurement error asymptotically
- If $\kappa > 0$, bias present of order $\mathbb{E}[C_i^{-1}]$
- In either case, the standard errors are correct

Empirical relevance of κ

Many common empirical settings have $\kappa >> 0$:

Minimum Data Set (Nursing Homes) (Einav et al, 2022; Olenski & Sacher, 2024):

- $n \approx 24$ million patients, $C_i = 107$ health measures, $\kappa \approx 221,000$

Empirical relevance of κ

Many common empirical settings have $\kappa >> 0$:

Minimum Data Set (Nursing Homes) (Einav et al, 2022; Olenski & Sacher, 2024):

- $n \approx 24$ million patients, $C_i = 107$ health measures, $\kappa \approx 221,000$

Lightcast data:

- $n = 45$ million job postings, inverse post length $E[C_i^{-1}] \approx 0.003$, $\kappa \approx 20$

Empirical relevance of κ

Many common empirical settings have $\kappa >> 0$:

Minimum Data Set (Nursing Homes) (Einav et al, 2022; Olenski & Sacher, 2024):

- $n \approx 24$ million patients, $C_i = 107$ health measures, $\kappa \approx 221,000$

Lightcast data:

- $n = 45$ million job postings, inverse post length $E[C_i^{-1}] \approx 0.003$, $\kappa \approx 20$

Nielsen Homescan:

- $n = 40{,}000$ households, 53 purchases per household in average category, $\kappa > 3.77$

Empirical relevance of κ

Many common empirical settings have $\kappa >> 0$:

Minimum Data Set (Nursing Homes) (Einav et al, 2022; Olenski & Sacher, 2024):

- $n \approx 24$ million patients, $C_i = 107$ health measures, $\kappa \approx 221,000$

Lightcast data:

- $n = 45$ million job postings, inverse post length $E[C_i^{-1}] \approx 0.003$, $\kappa \approx 20$

Nielsen Homescan:

- $n = 40{,}000$ households, 53 purchases per household in average category, $\kappa > 3.77$

10K Business Descriptions (Hoberg & Phillips, 2016):

- $n = 5{,}000$ firms, $E(C_i^{-1}) \approx 0.0064$, $\kappa \approx 0.45$

Stylized Model

Supervised Topic Models with Covariates

Simulation Study

Replication: CEO Behavior and Firm Performance

Theoretical Model and Results

Model:

- Outcome Y_i depends on K-dimensional latent variables $\boldsymbol{\theta}_i$ and covariates $\mathbf{q}_i$:

$$Y_i = \boldsymbol{\gamma}^T \boldsymbol{\theta}_i + \boldsymbol{\alpha}^T \mathbf{q}_i + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | \boldsymbol{\theta}_i, \mathbf{q}_i] = 0$$

- V-dimensional unstructured data $\mathbf{x}_i$ generated by latent $\boldsymbol{\theta}_i$:

$$\mathbf{x}_i | (C_i, \boldsymbol{\theta}_i) \sim \text{Multinomial}(C_i, \mathbf{B}^T \boldsymbol{\theta}_i)$$

  where C_i is total number of features, $\mathbf{B}$ is $K \times V$ matrix of topic weights.

- Two-step: (i) Estimate $\hat{\boldsymbol{\theta}}_i$ from $\mathbf{x}_i$, (ii) Regress Y_i on $\hat{\boldsymbol{\theta}}_i$ and $\mathbf{q}_i$.

Theoretical Model and Results

Model:

- Outcome Y_i depends on K-dimensional latent variables $\boldsymbol{\theta}_i$ and covariates $\mathbf{q}_i$:

$$Y_i = \boldsymbol{\gamma}^T \boldsymbol{\theta}_i + \boldsymbol{\alpha}^T \mathbf{q}_i + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | \boldsymbol{\theta}_i, \mathbf{q}_i] = 0$$

- V-dimensional unstructured data $\mathbf{x}_i$ generated by latent $\boldsymbol{\theta}_i$:

$$\mathbf{x}_i | (C_i, \boldsymbol{\theta}_i) \sim \text{Multinomial}(C_i, \mathbf{B}^T \boldsymbol{\theta}_i)$$

  where C_i is total number of features, $\mathbf{B}$ is $K \times V$ matrix of topic weights.

- Two-step: (i) Estimate $\hat{\boldsymbol{\theta}}_i$ from $\mathbf{x}_i$, (ii) Regress Y_i on $\hat{\boldsymbol{\theta}}_i$ and $\mathbf{q}_i$.

Theorem 1: (Fixed DGP)

- OLS of Y_i on $\hat{\boldsymbol{\theta}}_i$ is inconsistent for γ.
- Bias depends on average inverse features per observation $\mathbb{E}[C_i^{-1}]$.

Theoretical Model and Results

Model:

- Outcome Y_i depends on K-dimensional latent variables $\boldsymbol{\theta}_i$ and covariates $\mathbf{q}_i$:

$$Y_i = \boldsymbol{\gamma}^T \boldsymbol{\theta}_i + \boldsymbol{\alpha}^T \mathbf{q}_i + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | \boldsymbol{\theta}_i, \mathbf{q}_i] = 0$$

- V-dimensional unstructured data $\mathbf{x}_i$ generated by latent $\boldsymbol{\theta}_i$:

$$\mathbf{x}_i | (C_i, \boldsymbol{\theta}_i) \sim \text{Multinomial}(C_i, \mathbf{B}^T \boldsymbol{\theta}_i)$$

  where C_i is total number of features, $\mathbf{B}$ is $K \times V$ matrix of topic weights.

- Two-step: (i) Estimate $\hat{\boldsymbol{\theta}}_i$ from $\mathbf{x}_i$, (ii) Regress Y_i on $\hat{\boldsymbol{\theta}}_i$ and $\mathbf{q}_i$.

Theorem 2: (Drifting DGP)

- Consider sequence where $\sqrt{n} \times \mathbb{E}[C_i^{-1}] \to \kappa \geq 0$.
- If $\kappa = 0$, two-step is asymptotically valid.
- If $\kappa > 0$, two-step is consistent but biased and CIs have incorrect coverage.

Integrated Model

Supervised Topic Model with Covariates

Upstream Topic Model:

$$\boldsymbol{\theta}_i \sim \text{LogisticNormal}\left(\boldsymbol{\Phi}\mathbf{g}_i, \mathrm{I}_K\sigma_\theta^2\right)$$
$$\mathbf{x}_i \sim \text{Multinomial}\left(C_i, \mathbf{B}^T\boldsymbol{\theta}_i\right)$$

Downstream Regression Model:

$$Y_i \sim \text{Normal}\left(\boldsymbol{\gamma}^T\boldsymbol{\theta}_i + \boldsymbol{\alpha}^T\mathbf{q}_i, \sigma_Y^2\right)$$

- Observed: covariates $\mathbf{g}_i$ and $\mathbf{q}_i$; outcomes Y_i; counts $\mathbf{x}_i$; and total counts C_i
- Unobserved: topic proportions $\boldsymbol{\theta}_i$

Estimation with HMC using Probabilistic Programming

STMC defines joint likelihood $\ell(\mathbf{x}_i, Y_i|\boldsymbol{\theta}_i, C_i, \mathbf{g}i, \mathbf{q}i)$

Ideally integrate out $\boldsymbol{\theta}_i$ and use for MLE of parameters $\boldsymbol{\delta} = (\mathbf{B}, \boldsymbol{\Phi}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \sigma_Y, \sigma\theta)$

- But integration high-dimensional with no closed form

Solution: Use Bayesian computation to implicitly integrate

- Specify prior on $\boldsymbol{\delta}$, treat $\boldsymbol{\theta}_i$ as latent parameters
- Sample from posterior of $(\boldsymbol{\delta}, (\boldsymbol{\theta}_i)_{i=1}^n)$ given data
- Posterior mean of $\boldsymbol{\delta}$ asymptotically equivalent to MLE

We use Hamiltonian Monte Carlo (HMC) implemented in NumPyro probabilistic programming language

- Only need to specify DGP as a probabilistic program and the prior
- HMC is efficient for high-dimensional parameters and easily utilizes GPUs

Code Sample

```
    #### Upstream Factor Model ###

    with plate("topics", self.K):
        beta = sample("beta", dist.Dirichlet(
            self.eta * jnp.ones(self.V - self.num_anchors_per_class)))

    phis = sample("phis", dist.Normal(0,2).expand([self.q, self.K-1]))

    with plate_stack("docs", sizes = [self.N, self.K - 1]):
        A = sample("A", dist.Normal(jnp.matmul(Q, phis) , self.alpha))

    # document-topic distributions
    theta = deterministic(
        "theta",
        softmax(jnp.hstack([A, jnp.zeros([self.D, 1])]), axis = -1)
        )

    distMultinomial = dist.Multinomial(
        total_count=C,
        probs = jnp.matmul(theta, beta)
        )
    with plate("hist", self.N):
        X_bows = sample("obs_x", distMultinomial, obs = X)

    #### Downstream Regression Model ###

    gammas = sample("gammas", dist.Normal(0, 10).expand([self.K-1]))
    zetas = sample("zetas", dist.Normal(0, 10).expand([self.z]))
    sigma = sample("sigma", dist.Gamma(1, 10))

    mean = jnp.matmul(theta[:,:(self.K-1)], gammas) + jnp.matmul(Z, zetas)

    with plate("y", self.N):
        Y = sample("obs_y",  dist.Normal(mean, sigma), obs = Y)
```

Stylized Model

Supervised Topic Models with Covariates

Simulation Study

Replication: CEO Behavior and Firm Performance

Setup

Simulation study:

- Simulate data from STMC with $K = 2$ topics
- 200 simulations per set with $n = 10,000$
- 100 "anchor words" per to ensure identification
- 3 sets of simulations varying $C_i \in \{25, 100, 200\}$ implying $\kappa \in \{4, 1, 0.5\}$
- Compare 1-step (STMC) vs 2-step on downstream γ_1 and upstream ϕ_1
- Also report 2-step using true θ_i as benchmark

Setup

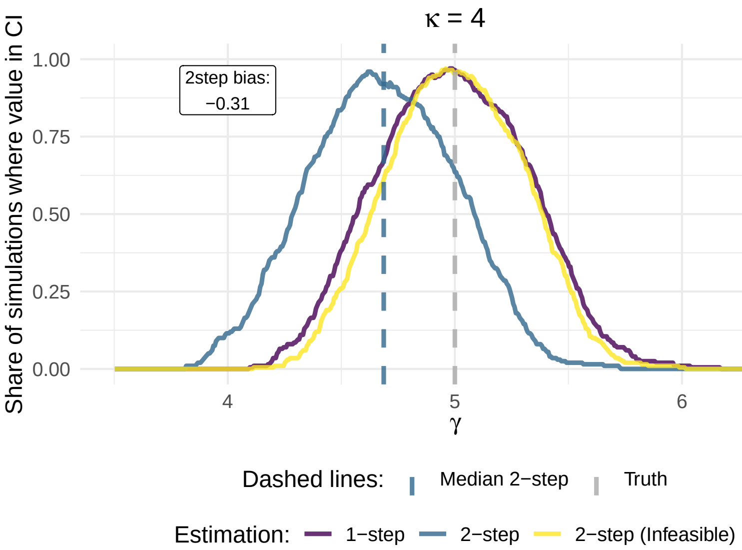
Simulation study:

- Simulate data from STMC with $K = 2$ topics
- 200 simulations per set with $n = 10,000$
- 100 "anchor words" per to ensure identification
- 3 sets of simulations varying $C_i \in \{25, 100, 200\}$ implying $\kappa \in \{4, 1, 0.5\}$
- Compare 1-step (STMC) vs 2-step on downstream γ_1 and upstream ϕ_1
- Also report 2-step using true θ_i as benchmark

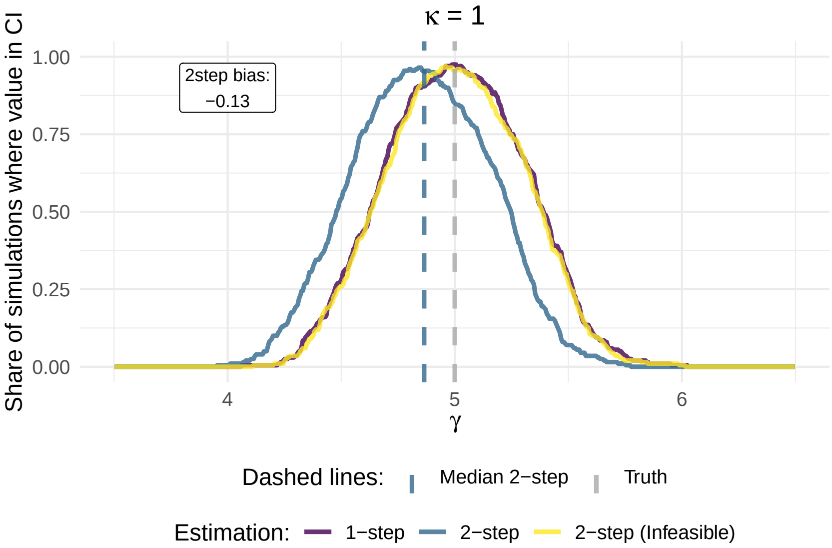
Theoretical predictions:

- Bias in 2-step which decreases when C_i increases (i.e. κ decreases)
- Width of CIs in 2-step is the same between 2-step and infeasible benchmark
- 1-step is unbiased, CIs have correct coverage but may be wider than 2-step

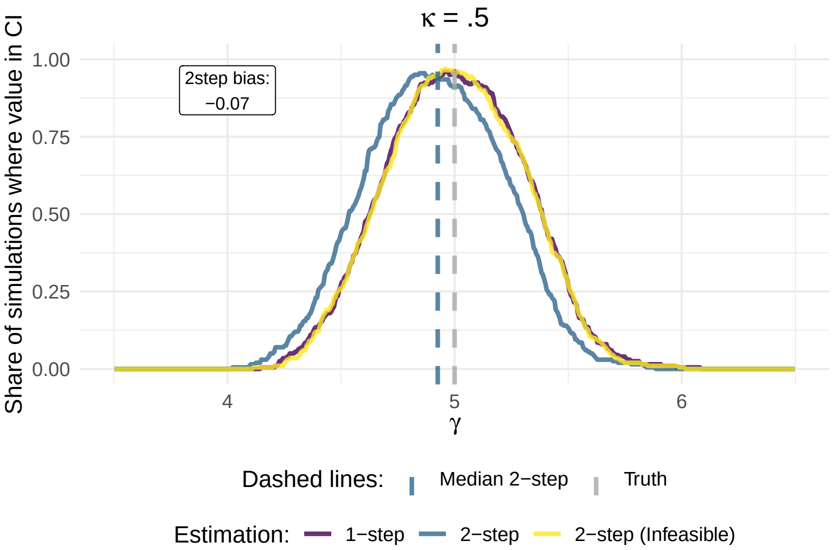
Downstream coefficient, γ



Downstream coefficient, γ



Downstream coefficient, γ



Stylized Model

Supervised Topic Models with Covariates

Simulation Study

Replication: CEO Behavior and Firm Performance

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 of feature counts.

Context

Bandiera et al. (2020) studied the relationship between CEO behavior and firm performance.

- $n = 916$ CEOs; 1 week of data per CEO; 15-minute intervals; 654 unique types of activities
- Average number of activities per CEO $\bar{C}_i = 88.4$

Context

Bandiera et al. (2020) studied the relationship between CEO behavior and firm performance.

- $n = 916$ CEOs; 1 week of data per CEO; 15-minute intervals; 654 unique types of activities
- Average number of activities per CEO $\bar{C}_i = 88.4$

Data consists of:

- *Firm performance*: Y_i is the firm's net sales
- *CEO behavior*: $\boldsymbol{x}_i \in \mathbb{Z}^{654}$ is the number of times the CEO engaged in each activity
- *Covariates*: $\boldsymbol{g}_i$, $\boldsymbol{q}_i$ e.g. CEO education, firm's employment, etc.

Context

Bandiera et al. (2020) studied the relationship between CEO behavior and firm performance.

- $n = 916$ CEOs; 1 week of data per CEO; 15-minute intervals; 654 unique types of activities
- Average number of activities per CEO $\bar{C}_i = 88.4$

Data consists of:

- *Firm performance*: Y_i is the firm's net sales
- *CEO behavior*: $\boldsymbol{x}_i \in \mathbb{Z}^{654}$ is the number of times the CEO engaged in each activity
- *Covariates*: $\boldsymbol{g}_i$, $\boldsymbol{q}_i$ e.g. CEO education, firm's employment, etc.

The authors used LDA with $K = 2$ "topics" to summarize the CEO's behavior.

- The topic distributions $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are named *Pure Behaviors*.
- The CEO's share of topic 1, $\theta_{i,1}$, used as *CEO index*.

Our Approach

Used Supervised Topic Model with Covariates to jointly estimate the relationship between covariates ($\boldsymbol{g}_i$, $\boldsymbol{q}_i$), CEO behavior ($\boldsymbol{x}_i$), and firm performance (Y_i).

Our Approach

Used Supervised Topic Model with Covariates to jointly estimate the relationship between covariates ($\boldsymbol{g}_i$, $\boldsymbol{q}_i$), CEO behavior ($\boldsymbol{x}_i$), and firm performance (Y_i).

Also estimated the model in two steps keeping priors and hyperparameters and inference algorithms the same as in the joint model:

1. Estimate $\boldsymbol{B}$ and $\boldsymbol{\theta}_i$
2. Regression coefficient γ_i, $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$ using estimated $\hat{\theta}_i$ as a proxy for $\boldsymbol{\theta}_i$

Our Approach

Used Supervised Topic Model with Covariates to jointly estimate the relationship between covariates ($\boldsymbol{g}_i$, $\boldsymbol{q}_i$), CEO behavior ($\boldsymbol{x}_i$), and firm performance (Y_i).

Also estimated the model in two steps keeping priors and hyperparameters and inference algorithms the same as in the joint model:

1. Estimate $\boldsymbol{B}$ and $\boldsymbol{\theta}_i$
2. Regression coefficient γ_i, $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$ using estimated $\hat{\theta}_i$ as a proxy for $\boldsymbol{\theta}_i$

We also estimated both models on a 10% sample of activities:

- In the full data $\kappa = 0.44$
- In the sampled data $\kappa = 4.26$

Pure Behaviors

CEOs with high $\theta_{i,1}$ are dubbed *Leaders*; those with low $\theta_{i,1}$ are *Managers*.

Both of our approaches yield similar topics.

Activity	1-step	2-step	Bandiera et al (2020)
Plant Visits	0.1	0.09	0.11
Suppliers	0.61	0.74	0.32
Production	0.38	0.33	0.46
Just Outsiders	0.74	1.21	0.58
Communication	1.44	1.23	1.49
Multi-Function	1.35	1.12	1.9
Insiders and Outsiders	1.8	1.83	1.9
C-suite	29.78	16.76	33.9

Note: The table shows the relative probability of different types of activities in each topic, $\frac{\beta_{1,j}}{\beta_{2,j}}$.

Effect of CEO Index on Firm Performance

	Dependent variable: Log(sales)			
	(1) 2-Step	(2) 1-Step	(3) 2-Step	(4) 1-Step
CEO Index	0.4	0.402	0.211	0.439
	(0.219, 0.572)	(0.240, 0.603)	(-0.028, 0.449)	(0.153, 0.711)
Log Employment	1.212	1.198	1.239	1.199
	(1.159, 1.268)	(1.154, 1.248)	(1.186, 1.29)	(1.148, 1.26)
Controls	X	X	X	X
Activities' Sample	Full	Full	10%	10%
κ statistic	0.44	0.44	4.26	4.26

- Coefficient on Log Employment ≈ 1.2 in all cases

Effect of CEO Index on Firm Performance

	Dependent variable: Log(sales)			
	(1) 2-Step	(2) 1-Step	(3) 2-Step	(4) 1-Step
CEO Index	0.4	0.402	0.211	0.439
	(0.219, 0.572)	(0.240, 0.603)	(-0.028, 0.449)	(0.153, 0.711)
Log Employment	1.212	1.198	1.239	1.199
	(1.159, 1.268)	(1.154, 1.248)	(1.186, 1.29)	(1.148, 1.26)
Controls	X	X	X	X
Activities' Sample	Full	Full	10%	10%
κ statistic	0.44	0.44	4.26	4.26

- Coefficient on Log Employment ≈ 1.2 in all cases
- In full data, coefficient on CEO index is ≈ 0.4 in both cases

Effect of CEO Index on Firm Performance

	Dependent variable: Log(sales)			
	(1) 2-Step	(2) 1-Step	(3) 2-Step	(4) 1-Step
CEO Index	0.4	0.402	0.211	0.439
	(0.219, 0.572)	(0.240, 0.603)	(-0.028, 0.449)	(0.153, 0.711)
Log Employment	1.212	1.198	1.239	1.199
	(1.159, 1.268)	(1.154, 1.248)	(1.186, 1.29)	(1.148, 1.26)
Controls	X	X	X	X
Activities' Sample	Full	Full	10%	10%
κ statistic	0.44	0.44	4.26	4.26

- Coefficient on Log Employment ≈ 1.2 in all cases
- In full data, coefficient on CEO index is ≈ 0.4 in both cases
- In 10% sample using 1-step, point estimate is similar, standard errors increase

Effect of CEO Index on Firm Performance

	Dependent variable: Log(sales)			
	(1) 2-Step	(2) 1-Step	(3) 2-Step	(4) 1-Step
CEO Index	0.4	0.402	0.211	0.439
	(0.219, 0.572)	(0.240, 0.603)	(-0.028, 0.449)	(0.153, 0.711)
Log Employment	1.212	1.198	1.239	1.199
	(1.159, 1.268)	(1.154, 1.248)	(1.186, 1.29)	(1.148, 1.26)
Controls	X	X	X	X
Activities' Sample	Full	Full	10%	10%
κ statistic	0.44	0.44	4.26	4.26

- Coefficient on Log Employment $\approx$ 1.2 in all cases
- In full data, coefficient on CEO index is $\approx$ 0.4 in both cases
- In 10% sample using 1-step, point estimate is similar, standard errors increase
- In 10% sample using 2-step, point estimate is halved and insignificant