Lecture 4: Causal NLP

Julian Ashwin

London Business School

Heidelberg, 2022

Road Map

- Key challenge: feature selection
- Double Machine Learning
- Bayesian Topic Regression
- Application: Financial news media and volatility

Causal inference with text

- Causal inference using observational text data is becoming increasingly popular in many research areas
  - Text can be confounder, proxy for confounder or treatment
  - We focus on text as (proxy for) confounder

- Key challenge is the representation and selection of text features
  - We need to capture dependencies between text confounders, treatment and outcomes correctly - otherwise risk of biased estimates
  - Can use unsupervised feature extraction, but these are unlikely to be optimal
  - Can use supervised feature extraction, but then need to be very careful with interpretation.

Conceptual framework

- **Data**:
  - y_i is the outcome
  - t_i is the treatment
  - W_i and C_i are text and numerical confounders
- **Goal**:
  - Identify the effect of t of y, controlling for W and C
  - To do this, we need to extract features Z from from W
- **Unsupervised**: use an unsupervised feature extraction approach
  - Won't lead to bias
  - But might not extract relevant features
- **Supervised**: extract Z that helps predict y
  - Extract more relevant features
  - Leads to bias if Z, C and t are correlated
- **Joint estimation**: extract Z in a supervised fashion taking into account dependencies between W, y, t, and C.

Frisch-Waugh-Lovell Theorem

The FWL theorem [Frisch and Waugh, 1933, Lovell, 1963] tells us that:

> *Any predictor's regression coefficient in a multivariate model is equivalent to the regression coefficient estimated from a bivariate model in which the residualised outcome is regressed on the residualised component of the predictor; where the residuals are taken from models regressing the outcome and the predictor on all other predictors in the multivariate regression (separately).*

Frisch-Waugh-Lovell Theorem

More formally, let

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$$

then we can residualise both y and x_1 on x_2

$$y_i = \beta_2^y x_{2,i} + \epsilon_i^y$$

$$x_{1,i} = \beta_2^{x_1} x_{2,i} + \epsilon_i^{x_1}$$

Then β_1 is the same in the joint model as in

$$\epsilon_i^y = \beta_1 \epsilon_i^{x_1} + \epsilon_i$$

Double Machine Learning

Just an extension of FWL

A model that is linear in the treatment x, but general in control variables Z.

$$y_i = \beta x_i + g_y(Z, \theta_y) + \epsilon_{x,i}$$

$$x_i = g_x(Z, \theta_x) + \epsilon_{x,i}$$

Analogously to FWL, we can thus:

1. Predict x based on Z using whatever supervised model we like
2. Predict y based on Z using whatever supervised model we like
3. Run a linear regression of the residuals from (2) on the residuals from (1) to get an estimate for β.

[Veitch et al., 2020] provide an application of word embeddings to this logic.

Conceptual outline of Bayesian Topic Regression

Want to estimate:

- $y = \mathbb{E}[y|Z, C, t] + \epsilon = t\omega_t + Z\omega_Z + C\omega_C + \epsilon$
- Cannot estimate equation in separate steps when Z and C are correlation with both treatment and outcome as (causal) regression coefficients ω would get biased (Frisch-Waugh-Lovell theorem)

Bayesian Topic Regression model:

- **Jointly estimate** the text representation Z and regression coefficients ω in fully Bayesian model taking into account dependencies between y_i, W_i, t_i, and C_i.
- Given focus on causal interpretation, use **Gibbs sampling** implementation as it provides statistical guarantees of asymptotically exact samples of the target density while (neural) mean-field variational inference does not

Related Literature ▸ Back

1. **Causal Inference with Text**
   - Increasing focus in NLP literature [Keith et al., 2020]
   - [Veitch et al., 2020]: restricted to binary treatment effects and only text as confounder.
2. **Topic Modelling**
   - canonical LDA model: [Blei et al., 2003], see [Vayansky and Kumar, 2020] for review of applications
   - supervised topic models [Blei and McAuliffe, 2008, Chen et al., 2015].
3. **Joint estimation**
   - Omitted Variable Bias/Frisch-Waugh-Lovell theorem [Frisch and Waugh, 1933, Lovell, 1963]
   - SCHOLAR [Card et al., 2018]: a neural classification model that incorporates both numerical and text features. We implement a regression extension (rSCHOLAR) as a benchmark.

Variables

- Corpus of D documents, vocab of V terms and K topics
- $w_{d,n}$, each document $d \in \{1, ..., D\}$ made up of N_d words from the vocabulary of V unique terms.
- $z_{d,n}$ the topic assignment (one of K) for word $w_{d,n}$. This is a latent variable that needs to be estimated.
- x_d, each document associated with some numerical metadata/covariates
- y_d, each document associated with an outcome that (potentially) depends on $\bar{z}_d$ and x_d:

$$y = A\omega + \epsilon$$

where A is a design matrix based on Z, X and optionally interaction terms, and $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$.

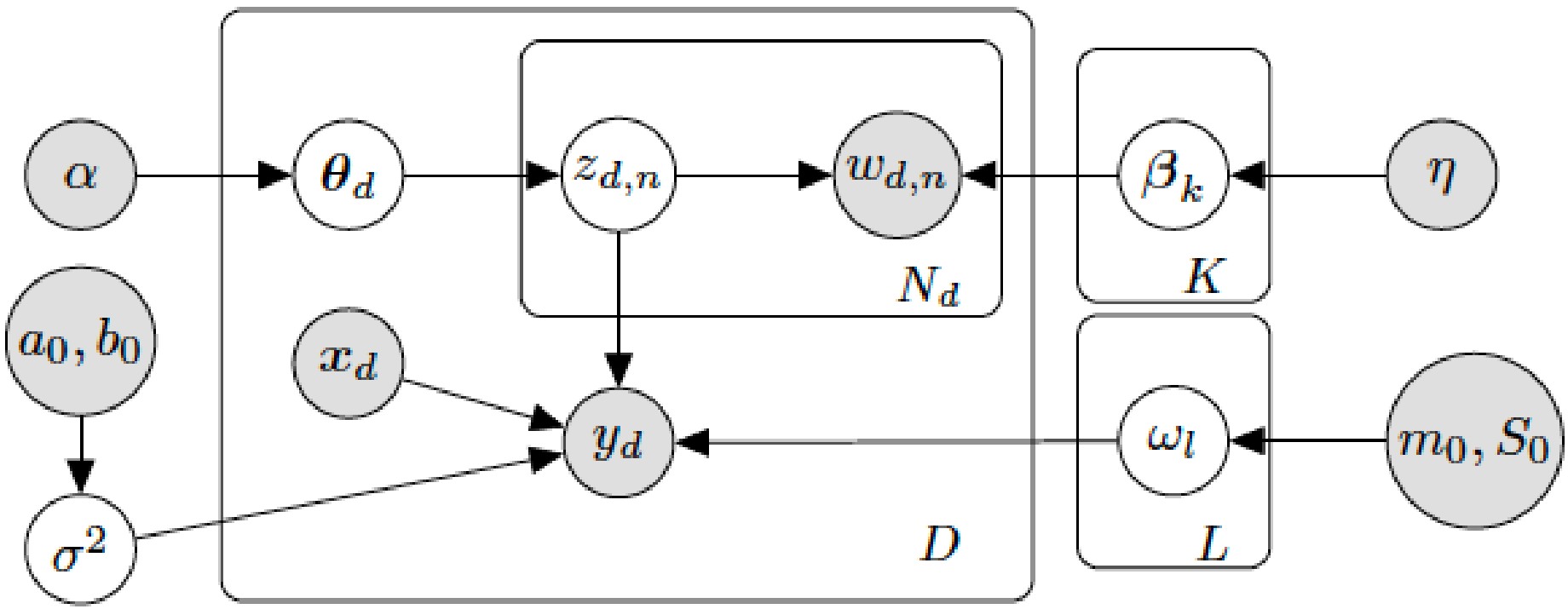
Generative model

1. $\omega \sim \mathcal{N}(\omega | m_0, S_0)$ and $\sigma^2 \sim \text{Inv-Gamma}(\sigma^2 | a_0, b_0)$
2. **for** $k = 1, \ldots, K$:
   1. $\beta_k \sim \text{Dir}(\eta)$
3. **for** $d = 1, \ldots, D$:
   1. $\theta_d \sim \text{Dir}(\alpha)$
   2. **for** $n = 1, \ldots, N_d$:
      1. topic assignment $z_{d,n} \sim \mathit{Mult}(\theta_d)$
      2. term $w_{d,n} \sim \mathit{Mult}(\beta_{z_{d,n}})$
4. $y \sim \mathcal{N}\left(A\omega, \sigma^2 I\right)$, where A is a design matrix based on Z and X.

▸ Graphical model

Graphical model

Back



Gibbs EM algorithm

- Can collapse out θ and β, following [Griffiths and Steyvers, 2004].
- Posterior distribution is then:

$$p(Z, \omega, \sigma^2 | W, X, y) = p(Z | W, X, y, \omega, \sigma^2) p(\omega, \sigma^2 | Z, X, y). \quad (1)$$

- E-step: approximate $p(Z|W, X, y, \omega, \sigma^2)$ by Gibbs sampling.
- M-step: given $\mathbb{E}(Z|\cdot)$, we have can analytical express the Normal-Inverse-Gamma distribution for $p(\omega, \sigma^2|Z, X, y)$.

E-step

▸ Back

Probability of $w_{d,n}$ being assigned to a given topic k, given assignments of all other words $(Z_{-(d,n)})$

$$p(z_{d,n} = k | Z_{-(d,n)}, W, X, y, \omega, \sigma^2) \propto$$
$$p(z_{d,n} = k | Z_{-(d,n)}, W) p(y_d | z_{d,n} = k, Z_{-(d,n)}, x_d, \omega, \sigma^2).$$

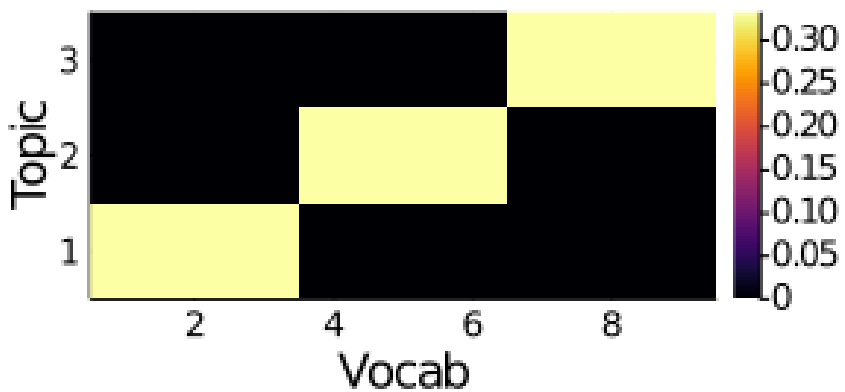
Sampling distribution for $z_{d,n}$ is multinomial distribution, defines the probability for each k that $z_{d,n}$ is assigned to that topic k:

$$p(z_{d,n} = k | z_{-(d,n)}, W, X, y, \alpha, \eta, \omega, \sigma^2) \propto$$
$$(s_{d,k,-n} + \alpha) \frac{m_{k,v,-(d,n)} + \eta}{\sum_v m_{k,v,-(d,n)} + V\eta} \exp\left\{ \frac{1}{2\sigma^2} \left(\frac{2\tilde{\omega}_{z,d,k}}{N_d} \left(y_d - \omega_x^\intercal x_{2,d} - \frac{\tilde{\omega}_{z,d}^\intercal}{N_d} s_{d,-n} \right) - \left(\frac{\tilde{\omega}_{z,d,k}}{N_d} \right)^2 \right) \right\}.$$

In the paper: extensions to multiple documents per observation and interaction terms...

Synthetic data

- Illustrate importance of joint approach for (causally) interpretable parameter estimation
- 10,000 documents of 50 words each and vocabulary of 9, following an LDA generate process



- Outcome depends on text and a numerical covariate: $y = x - \bar{z}_1 + \epsilon$, where x is the frequency of word 1

Benchmarks

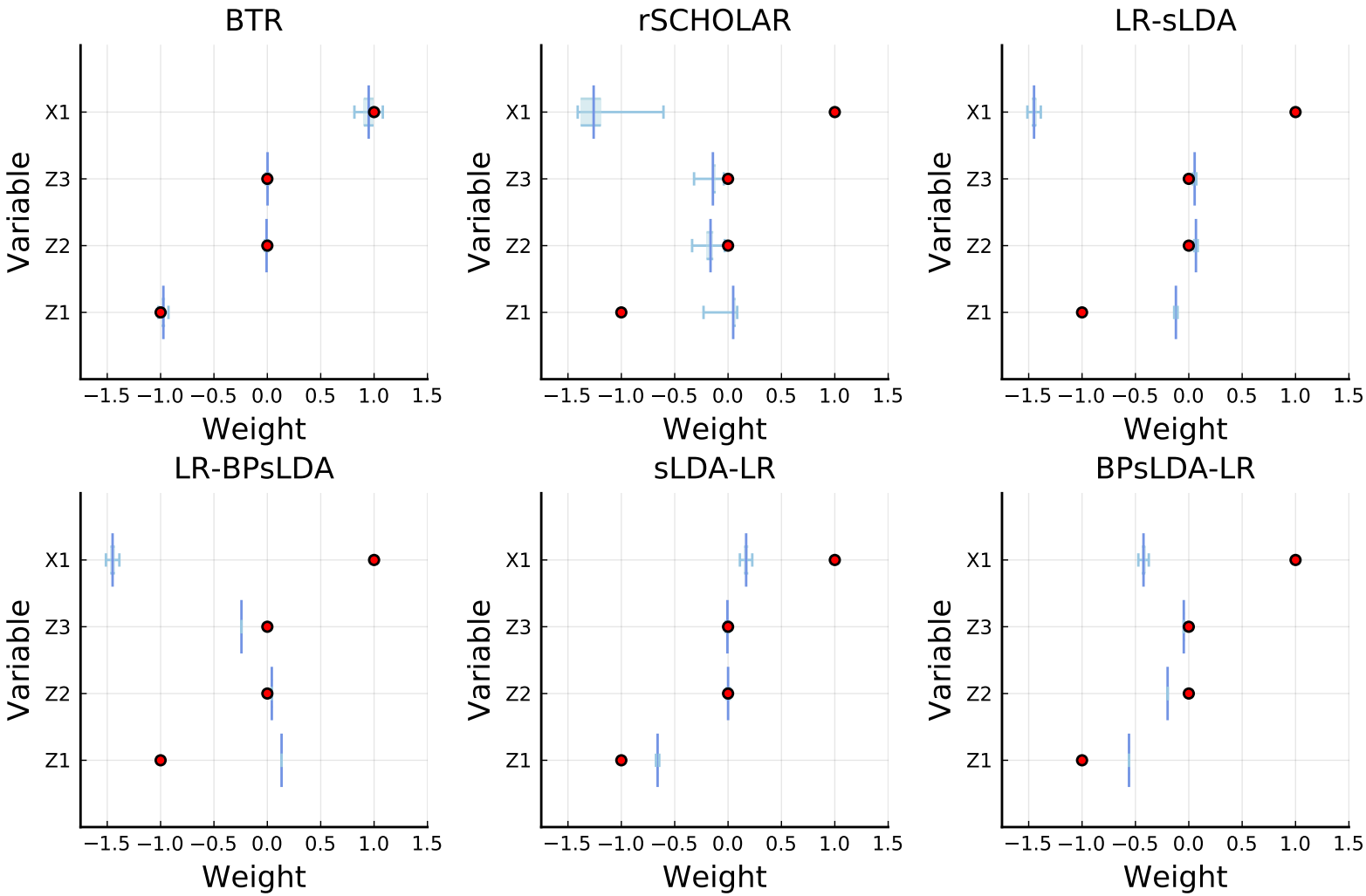
Joint estimation approaches:

1. **BTR:** our Bayesian model, estimated via Gibbs sampling
2. **rSCHOLAR**: the regression extension of SCHOLAR, estimated via neural VI

2-stage approaches:

1. **LR-sLDA:** first linearly regress y on x, then use the residual of that regression as the response in an sLDA model, estimated via Gibbs sampling
2. **sLDA-LR:** First sLDA, then linear regression
3. **BPsLDA-LR:** replacing sLDA with BPsLDA, which is sLDA estimated via the backpropagation approach of [Chen et al., 2015]
4. **LR-BPsLDA:** again replacing sLDA with BPsLDA

Results - synthetic data



Two examples

Real text (hotel/restaurant reviews), but synthetically generated outcome, **so that true data generating process (DGP) is known**

Text data is confounder, numerical data is treatment

1. **Yelp data**:
   - **Synthetic treatment variable**, so can control the degree of correlation between treatment and confounder
   - Binary treatment variable
2. **Booking.com data**
   - **Real treatment variable**, correlated with confounding text
   - Continuous treatment variable

Yelp

Do Americans give lower scores, conditional on their review text?

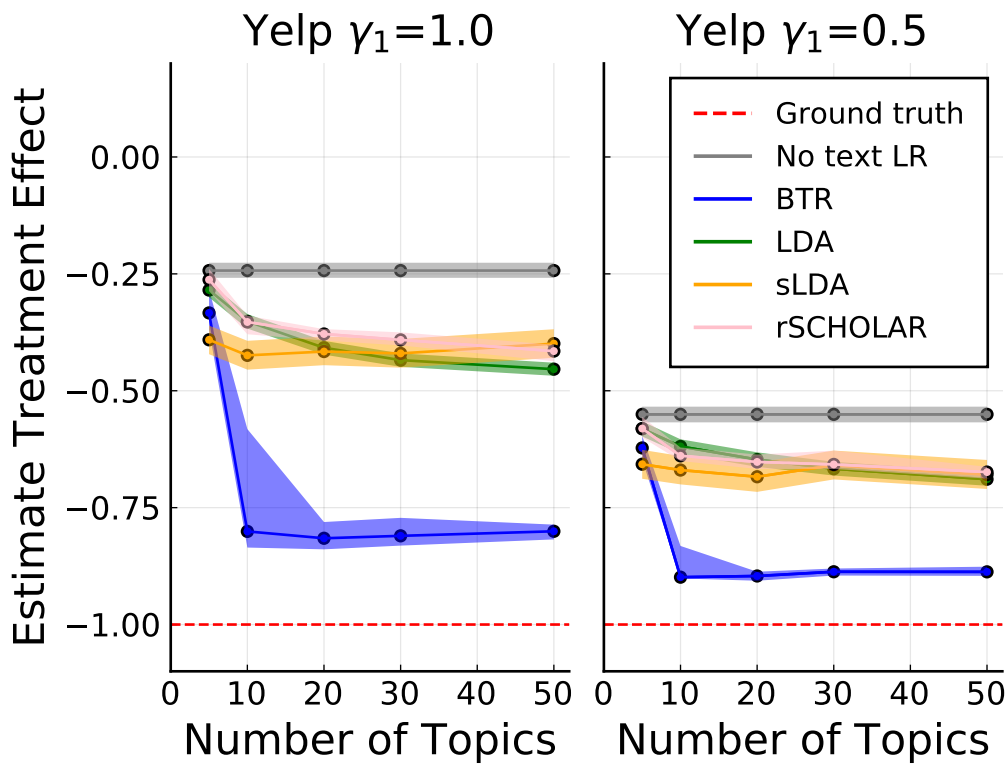
Synthetic treatment:

$$\Pr(US_i = 1) = \frac{\exp(\gamma_1 sent_i)}{1 + \exp(\gamma_1 sent_i)}$$

Yelp DGP:

$$y_{\text{yelp},i} = -\text{US}_i + \text{stars\_av\_b}_i + sent_i.$$

γ_1 controls the correlation between text and treatment.



Booking

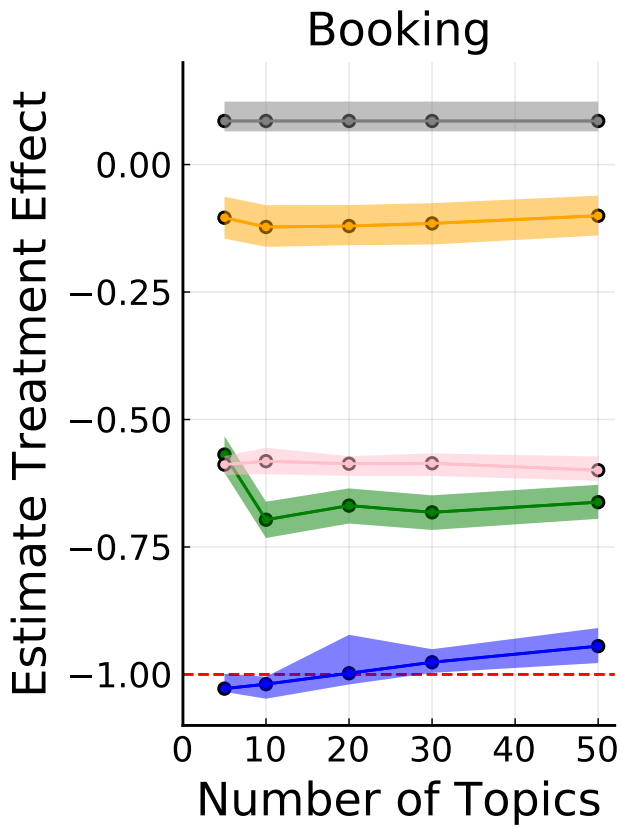
Do better hotels get lower scores, controlling for customer reviews?

Real treatment (historical average score for hotel)

Booking DGP:

$$y_{\text{booking},i} = -\text{hotel\_av}_i + 5\text{prop\_pos}_i + \epsilon_i, \quad (2)$$

where prop_pos is the proportion of positive words in a review



Experiment setup

- Joint supervised estimation approach (text and numerical) also improves prediction performance
- For both datasets - Yelp and Booking - we predict customer ratings (target) for reviews (text features) and business/customer metadata (numerical features)
  - **Yelp**: *historic avg. rating by user*, *historic avg. rating of business*, and a *Harvard Inquirer* sentiment score.
  - **Booking**: *historic avg hotel score*, *total neg word counts in customer review*, *total pos word counts in customer review*, *total num of reviews by customer*, *total num of reviews of hotel*
- For both datasets, we randomly sample $50,000$ observations and select 75% in Yelp, 80% in Booking for training
- Report mean and standard deviations from 20 runs

Benchmarks

- **OLS:** OLS regression on numerical metadata only
- **LR+sLDA:** first linear regression, then sLDA
- **LR+BPsLDA:** replacing sLDA with BPsLDA
- **BTR:** our Bayesian model.
- **rSCHOLAR:**the regression extension of SCHOLAR
- **LR+rSCHOLAR:** first linear regression, then rSCHOLAR
- **LDA+LR:** Gibbs sampled LDA, then regression
- **GSM+LR:** neural VI based topic models, then regression
- **LR+aRNN:** first linear regression, then a bidirectional Recurrent Neural Network with attention [Bahdanau et al., 2015]
- **LR+TAM:** first linear regression, then a bidirectional RNN using topic vector to enhance attention [Wang and Yang, 2020]

Yelp & Booking - prediction results (empirical data)

pR^2

Dataset	Booking		Yelp	
K	50	100	50	100
	pR^2 (higher is better)			
OLS	0.315		0.451	
LR+aRNN	0.479 (0.007)		0.582 (0.008)	
LR+TAM	0.479 (0.014)	0.487 (0.014)	0.585 (0.012)	0.587 (0.008)
LDA+LR	0.426 (0.003)	0.437 (0.002)	0.586 (0.006)	0.606 (0.007)
GSM+LR	0.386 (0.004)	0.395 (0.005)	0.495 (0.004)	0.517 (0.007)
LR+sLDA	0.432 (0.002)	0.438 (0.004)	0.571 (0.002)	0.574 (0.001)
LR+BPsLDA	0.419 (0.009)	0.455 (0.001)	0.603 (0.002)	0.609 (0.001)
LR+rSCHOLAR	0.469 (0.002)	0.465 (0.002)	0.550 (0.034)	0.557 (0.027)
rSCHOLAR	**0.494** (0.004)	**0.489** (0.003)	0.571 (0.01)	0.581 (0.009)
BTR	0.454 (0.003)	0.46 (0.002)	**0.630** (0.001)	**0.633** (0.001)

▸ Interpretability

Results

Perplexity

Table: Mean pR^2 and perplexity, standard deviation in brackets. 20 runs per model. Best model **bold**.

Dataset	Booking		Yelp	
K	50	100	50	100
	Perplexity (lower is better)			
LR+TAM	521 (2)	522 (2)	1661 (7)	1655 (7)
LDA+LR	454 (1)	432 (1)	1306 (4)	1196 (2)
GSM+LR	**369** (8)	**348** (5)	1431 (34)	1387 (14)
LR+sLDA	436 (2)	411 (1)	1294 (5)	1174 (3)
LR+rSCHOLAR	441 (20)	458 (11)	1515 (34)	1516 (30)
rSCHOLAR	466 (19)	464 (9)	1491 (9)	1490 (9)
BTR	437 (1)	412 (1)	**1291 (5)**	**1165** (3)

Conclusions

- Bayesian Topic Regression incorporates both numerical and text data for modelling a response variable, jointly estimating all parameters.
- Designed to avoid potential bias: sound regression framework for statistical and causal inference
- Recovers the ground truth with lower bias than any other benchmark model on synthetic and semi-synthetic datasets.
- Yields superior prediction performance compared to 'two-stage' strategies, even competing with deep neural networks.
- BTR.jl *Julia* package

Summary of BTR

- **Bayesian Topic Regression (BTR):**
  - Allows for both text and numerical information to model an outcome variable
  - Allows estimation of discrete and continuous treatment effects, as well as additional numerical confounding factors next to text confounders
- **Causal inference**:
  - Synthetic and semi-synthetic datasets
  - Our joint approach recovers ground truth with lower bias than any benchmark model, when text and numerical features are correlated
- **Prediction**:
  - Two real-world customer review datasets
  - Joint and supervised learning strategy also yields superior prediction results compared to 2-stage strategies
  - BTR even competitive with more complex deep neural networks

Motivation

Open question: what is the role of the media in financial markets?

- In a textbook model of financial markets, public information is quickly and efficiently priced in, so little role for media.
- However, if investors have limited attention and information diffuses slowly media may play a crucial role.
- Media coverage can direct investors' attention towards certain firms, which can have an impact on the volatility of their stock price:
  - Decrease volatility - more efficient reaction
  - Increase volatility - less efficient reaction
- If there is an effect, what are its aggregate implications?

Key Results

- Media coverage of a firm in the print edition of the *Financial Times* **increases** stock price volatility.
- This is not explained by:
  1. Persistence
  2. Anticipated new information
  3. Content of the coverage
- There is spillover to related firms, but not the aggregate index.

Related Literature

- Causal effect of media coverage
  - **Increase** [Bushee et al., 2010], or **decrease** [Solomon et al., 2014] efficiency
  - **Case study** approach [Huberman and Regev, 2001, Carvalho et al., 2011]
  - **Exogenous variation** in coverage [Dougal et al., 2012, Peress, 2014, Larsen and Thorsrud, 2017, Engelberg and Parsons, 2011]
  - This paper: focus on volatility and temporal separation
- Attention and media coverage
  - Salience view of media coverage [Merton, 1987, Solomon et al., 2014]
  - Theory [Andrei and Hasler, 2015]
  - Empirical [Schmidt, 2013, Dimpfl and Jank, 2016, Goddard et al., 2015, Peress, 2008, Fan et al., 2020]
  - This paper: novel empirical evidence supporting the salience view
- New information and financial markets
  - Firm-specific news and equity price reactions [Hendershott et al., 2015, Alanyali et al., 2013, Jiao et al., 2020]
  - Aggregate effects [Cohen and Frazzini, 2008]
  - This paper: media coverage effect has measurable spillover effects.

Data

Novel dataset: *Financial Times* articles and daily stock price data

- Every article in the *Financial Times* print edition from May 1998 to December 2017.
- Daily London Stock Exchange trading data for every firm which appeared in the FTSE 100 Index from May 1998 to December 2017 (289 firms)
  - Realised volatility: high-low spread
  - Absolute intra-day return: open-close spread
  - Implied volatility: derived from option prices

Matching articles to firms

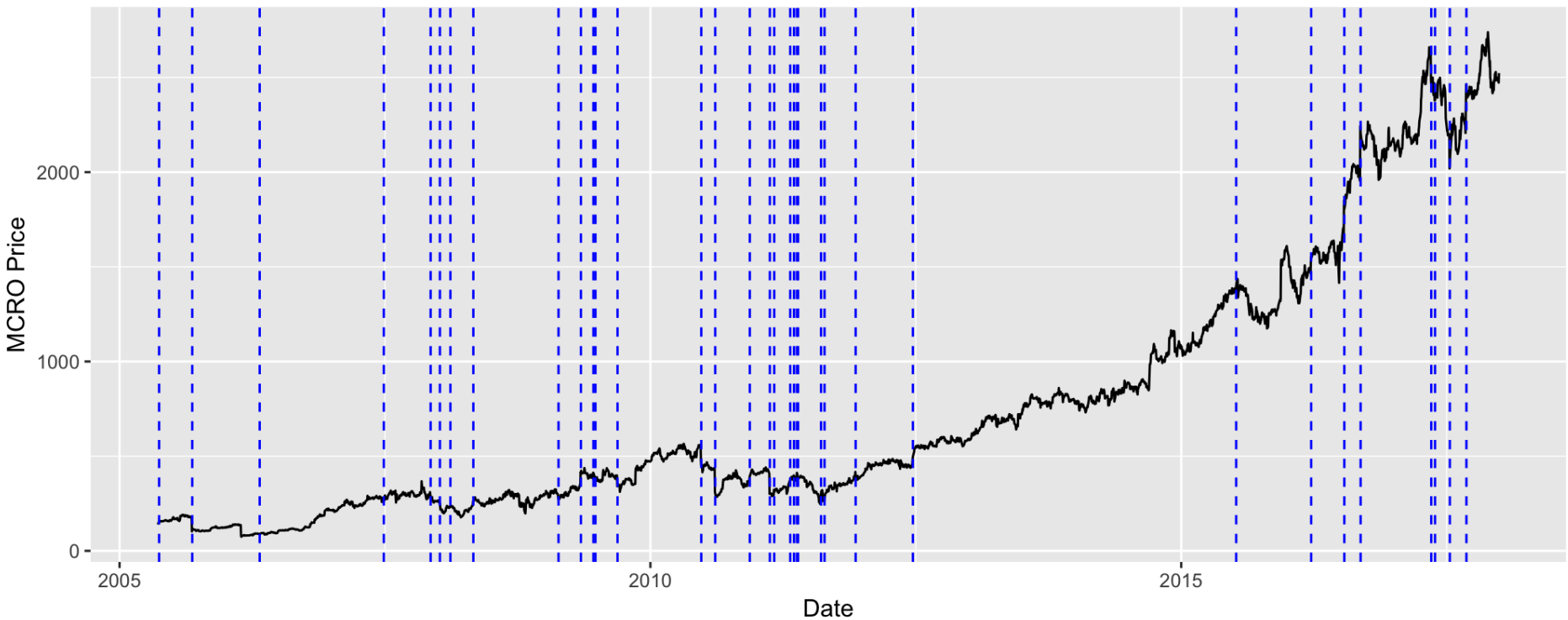
Match using both string-matching to headlines and a Named Entity Recognition algorithm, from Apache OpenNLP, on the full articles.

2003-01-09: Next confident of hitting its full-year target
Next yesterday dispersed the worst fears about its performance with news of ...

2001-01-24: Buoyant retail sales data defy bleak forecasts
... Surveys conducted in the first half of the month suggested sales were weak, verging on stagnant, and big retailers such as Dixons and Next said ...

	NER	
headline	0	1
0	765,401	56,644
1	21,225	17,027

Example



A Definition

Media Coverage vs new information

Media coverage and **new information** are distinct concepts, and this paper is primarily concerned with the former.

Holding the new information constant, greater media coverage may or may not have a causal effect.

Financial Times print newspaper well suited to investigate this, as it is unlikely to be investors' primary source of new information throughout the day.

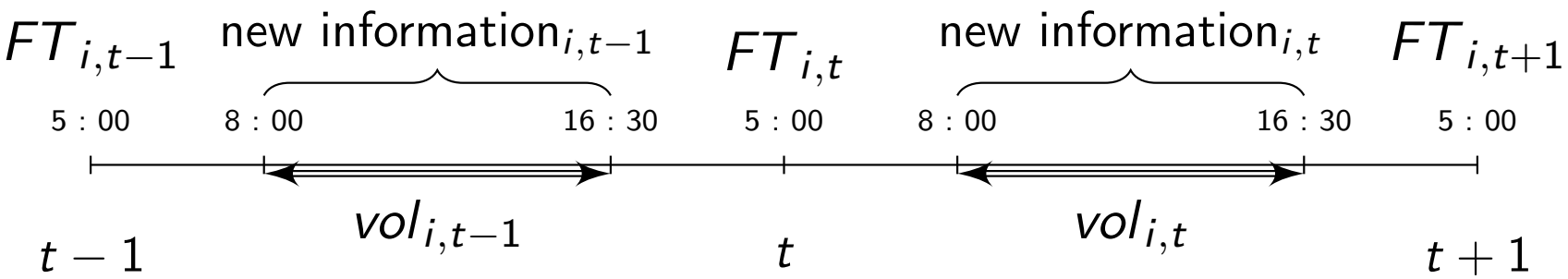
Media coverage effect

Summary

- An article in the FT increases the intra-day volatility of a firm's stock price by **around 10 basis points**.
- Exploit the timing structure of trading hours and newspaper publication to rule out reverse causality.
- Show that the effect is not driven by persistence.
- Show that the effect is not driven by new information or content of the coverage.

Identification strategy

Timings are key



- FT print newspaper is published hours before the London Stock Exchange opens
- Any media coverage effect on volatility is not due to the media simply reporting on that day's volatility.
- $vol_{i,t} = \alpha_i + \mu_t + \beta mention_{i,t} + \sum_{p=1}^{P} \gamma_p vol_{i,t-p} + \text{controls} + \varepsilon_{i,t}$

Identification strategy

Initial results

	Dependent variable:			
	$vol_{i,t}$ (in percentage points)			
	(1)	(2)	(3)	(4)
$mention_{i,t}$	0.501^{***}	0.597^{***}	0.517^{***}	¡+¿0.121^{***}
	(0.020)	(0.020)	(0.017)	(0.016)
$vol_{i,t-1}$				0.390^{***}
				(0.001)
Constant	3.086^{***}			
	(0.003)			
Firm fixed effects		✓	✓	✓
Time fixed effects			✓	✓
Observations	862,974	862,974	862,974	849,061
R^2	0.001	0.123	0.330	0.433
Adjusted R^2	0.001	0.123	0.326	0.430
Residual Std. Error	2.616	2.451	2.149	1.973
Note:				*p<0.1; **p<0.05; ***p<0.01

Persistence

Volatility in financial markets is persistent.

I carry out three tests to show that the media coverage effect is not driven by reporting on past price movements.

1. Control for past price movements.
   - A single lag of volatility appears to sufficiently control for past price movements. ▸ Results
2. Forward looking have a greater effect. ▸ Results
   - Using forward-looking dictionary ▸ More detail
   - Articles which appear on Mondays ▸ Results
3. The effect of today's media coverage survives controlling for tomorrow's media coverage. ▸ Results

New Information

An Example

Both media coverage and stock price movements are driven by new information, if media anticipates that information will arrive.

RBS set to reveal it suffered £20bn loss last year (FT, 2009-01-19)
Royal Bank of Scotland is today expected to announce it suffered a £20bn ...

$$vol_{i,t} = 71\%, \qquad |\Delta p|_{i,t} = -66\%$$

RBS counts £28bn cost of past ambition (FT, 2009-01-20)
Stephen Hester, chief executive of Royal Bank of Scotland, yesterday resisted ...

New Information

I carry out three tests to show that the effect is not driven by the omission of underlying new information.

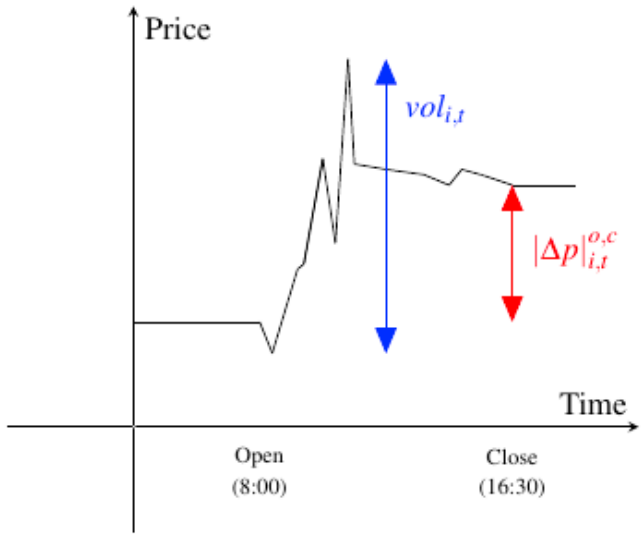
1. Show that sentiment of *FT* articles does not predict *subsequent* returns. ▸ Measuring sentiment ▸ Sentiment results

2. Use absolute intra-day and overnight returns, as the broadest possible control for realised new information that day.
   - Volatility effect greater than that explained by absolute return

3. Use an implied volatility measure derived from options on the underlying stocks, to control for market expectations of volatility
   - Demonstrate that the effect is not driven by the arrival of new information that is anticipated by the media

New Information

Volatility and returns ▸ Back



I do not explicitly control for information events, e.g. earnings reports [Peress, 2008], as the price changes constitute the broadest possible measure.

- Limits analysis to media coverage of a specific type of event
- Not all earnings report are equal - some are more anticipated than others

New Information

Implied Volatility

If the media can anticipate newsworthy events, so can investors who can then hedge against large movements through the options market.

- Calculated from the values of the nearest expiry month options, using the Black-Scholes formula
- Very good predictors of realised volatility: R^2 of 0.525
- Available for 152 of the 275 firms (362,833/863,641 observations)

New Information

Implied Volatility Results

	Dependent variable:				
	$vol_{i,t}$				
	(1)	(2)	(3)	(4)	(5)
$mention_{i,t}$	0.121***	0.088***	0.090***	0.084***	0.086***
	(0.016)	(0.012)	(0.012)	(0.013)	(0.013)
$\lvert\Delta p\rvert_{i,t}^{o,c}$		0.716***	0.708***	0.702***	0.700***
$\lvert\Delta p\rvert_{i,t}^{c,o}$			0.089***	0.124***	0.124***
$VI_{i,t-1}^{put}$				0.768***	0.214***
$VI_{i,t-1}^{call}$				0.211***	0.015
$VI_{i,t}^{put}$					0.882***
$VI_{i,t}^{call}$					0.244***
Lags & FE	✓	✓	✓	✓	✓
Observations	849,061	730,836	719,292	306,741	306,721
R^2	0.433	0.707	0.712	0.768	0.769

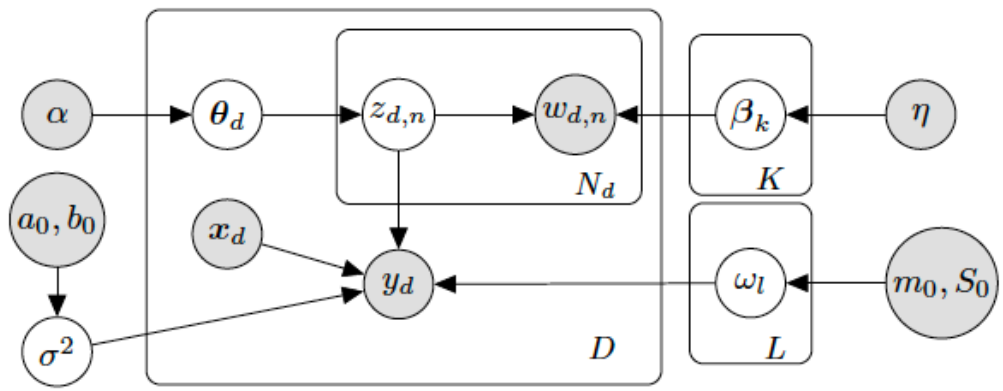
Article Content

- We can move beyond the somewhat ad hoc approach of dictionary/sentiment methods.
- Interested in volatility, not direction, so unclear which text features might be relevant.
- Bayesian Topic Regression [Ahrens et al., 2021] estimates a regression in which text acts as an explanatory variable, while jointly estimating a topic representation of the text.
  - Can include other controls - respect FWL theorem
  - Supervised learning - identify text features relevant to our context
  - Interpretable topic-based text features

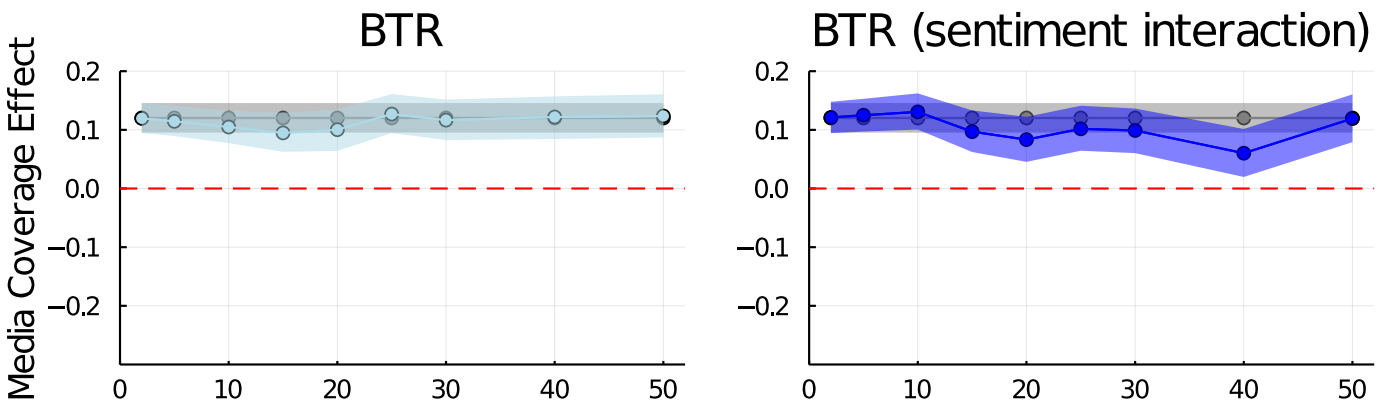
BTR details

$$y_{i,t} = \omega_x x_{i,t} + \omega_z \bar{z}_{i,t} + \epsilon_{i,t}$$

$$x_{i,t} = \begin{pmatrix} mention_{i,t} \\ sentiment_{i,t} \\ |\Delta p|^{o,c}_{i,t} \\ \overline{vol}_t \\ \overline{vol}_i \\ vol_{i,t-1} \\ vol_{i,t-2} \\ vol_{i,t-3} \\ vol_{i,t-4} \\ vol_{i,t-5} \\ VI^{put}_{i,t} \\ VI^{call}_{i,t} \end{pmatrix}'$$



BTR results



Note: Grey points and lines in each panel represent a linear regression on just the numerical covariates, as a reference point for the text models. The shaded bands represent 95% posterior credible intervals for the estimated media coverage effect.

▸ Example

Aggregate importance

- Media coverage has a causal effect on a firm's stock price, but is this of aggregate importance?
- Spreads to other firms linked by the production network.
- Does not affect index-level volatility.
- This provides evidence for a salience-based interpretation: investors' attention is drawn towards some firms, and consequently away from others.

Aggregate importance

Sector level measures

- Use NACE sector classifications to construct a sector level average media coverage measure.
- Use the input-output tables from UK ONS to weight sectors by their importance for one another. ▸ Network
- For each firm i in sector s, we then have a measure of media coverage in upstream (potential supplier) and downstream (potential customers) sectors, e.g.

$$downstream\_mentions_{i,t} = \sum_{\varsigma=1}^{S} m_{s,\varsigma}^{sell} \frac{1}{n_\varsigma} \sum_{j \in \varsigma} mention_{j,t}$$

Aggregate effects ▸ Placebo

Production network

	Dependent variable:			
	$vol_{i,t}$			
	(1)	(2)	(3)	(4)
$mention_{i,t}$	0.090***	0.083***	0.093***	0.085***
	(0.013)	(0.014)	(0.014)	(0.014)
$\lvert\Delta p\rvert_{i,t}^{o,c}$	0.723***	0.723***	0.723***	0.717***
	(0.001)	(0.001)	(0.001)	(0.001)
$downstream\_mentions_{i,t}$		0.161***	0.140**	0.168***
		(0.059)	(0.059)	(0.059)
$upstream\_mentions_{i,t}$		−0.002	−0.121	−0.134
		(0.113)	(0.121)	(0.121)
$sector\_mentions_{i,t}$			¡1¿0.125***	¡1¿0.059
			(0.046)	(0.046)
$sector\_vol_{i,t}$				¡1¿0.094***
				(0.002)
$vol_{i,t}$ lags	✓	✓	✓	✓
Implied vol	✓	✓	✓	✓
Firm & Time fixed effects	✓	✓	✓	✓
Observations	312,499	312,499	312,499	311,925
R^2	0.766	0.766	0.766	0.767
Note:				*p<0.1; **p<0.05; ***p<0.01

Aggregate effects

Index level

	Dependent variable:		
	vol_t^{FTSE}		
	(1)	(2)	(3)
$\overline{mention}_t$	3.893***	0.100	0.475
	(1.290)	(0.934)	(0.617)
$\lvert\Delta p\rvert_t^{FTSE}$			0.643***
			(0.010)
Constant	1.329***	0.127***	0.113***
	(0.029)	(0.027)	(0.018)
vol_t^{FTSE} lags		✓	✓
Observations	4,291	3,690	3,807
R^2	0.002	0.574	0.805
Note:		*p<0.1; **p<0.05; ***p<0.01	

Conclusion

- Robust and substantial link between financial news media coverage on stock price volatility.
- Not explained by persistence, new information or content of coverage.
- Consistent with a salience theory of media coverage.
- Spillovers to firms in linked sectors.

End of mini-course

Thanks to everyone for coming along!

You can contact me at julianashwin@gmail.com

References I

Ahrens, M., Ashwin, J., Calliess, J.-P., and Nguyen, V. (2021).
Bayesian topic regression for causal inference.
arXiv preprint arXiv:2109.05317.

Alanyali, M., Moat, H. S., and Preis, T. (2013).
Quantifying the relationship between financial news and the stock market.
Scientific reports, 3:3578.

Andrei, D. and Hasler, M. (2015).
Investor attention and stock market volatility.
Review of Financial Studies, 28(1).

Bahdanau, D., Cho, K., and Bengio, Y. (2015).
Neural machine translation by jointly learning to align and translate.
In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Blei, D. M. and McAuliffe, J. D. (2008).
Supervised topic models.
In *Advances in Neural Information Processing Systems*, pages 121–128.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent dirichlet allocation.
Journal of Machine Learning Research, 3(Jan):993–1022.

References II

Bushee, B. J., Core, J. E., Guay, W., and Hamm, S. J. (2010).
The role of the business press as an information intermediary.
Journal of Accounting Research, 48(1):1–19.

Card, D., Tan, C., and Smith, N. A. (2018).
Neural models for documents with metadata.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040. Association for Computational Linguistics.

Carvalho, C., Klagge, N., and Moench, E. (2011).
The persistent effects of a false news shock.
Journal of Empirical Finance, 18(4):597–615.

Chen, J., He, J., Shen, Y., Xiao, L., He, X., Gao, J., Song, X., and Deng, L. (2015).
End-to-end learning of lda by mirror-descent back propagation over a deep architecture.
arXiv preprint arXiv:1508.03398.

Cohen, L. and Frazzini, A. (2008).
Economic links and predictable returns.
The Journal of Finance, 63(4):1977–2011.

Dimpfl, T. and Jank, S. (2016).
Can internet search queries help to predict stock market volatility?
European Financial Management, 22(2):171–192.

References III

Dougal, C., Engelberg, J., Garcia, D., and Parsons, C. A. (2012).
Journalists and the stock market.
The Review of Financial Studies, 25(3):639–679.

Engelberg, J. E. and Parsons, C. A. (2011).
The causal impact of media in financial markets.
The Journal of Finance, 66(1):67–97.

Fan, R., Talavera, O., and Tran, V. (2020).
Social media bots and stock markets.
European Financial Management, 26(3):753–777.

Frisch, R. and Waugh, F. V. (1933).
Partial time regressions as compared with individual trends.
Econometrica: Journal of the Econometric Society, pages 387–401.

Goddard, J., Kita, A., and Wang, Q. (2015).
Investor attention and fx market volatility.
Journal of International Financial Markets, Institutions and Money, 38:79–96.

Griffiths, T. L. and Steyvers, M. (2004).
Finding scientific topics.
Proceedings of the National Academy of Sciences, 101(suppl 1):5228–5235.

References IV

Hendershott, T., Livdan, D., and Schürhoff, N. (2015).
Are institutions informed about news?
Journal of Financial Economics, 117(2):249–287.

Huberman, G. and Regev, T. (2001).
Contagious speculation and a cure for cancer: A nonevent that made stock prices soar.
The Journal of Finance, 56(1):387–396.

Jiao, P., Veiga, A., and Walther, A. (2020).
Social media, news media and the stock market.
Journal of Economic Behavior & Organization, 176:63–90.

Keith, K., Jensen, D., and O'Connor, B. (2020).
Text and causal inference: A review of using text to remove confounding from causal estimates.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, Online. Association for Computational Linguistics.

Larsen, V. and Thorsrud, L. (2017).
Asset returns, news topics, and media effects.
Technical report, Centre for Applied Macro-and Petroleum Economics (CAMP), BI Norwegian ....

Loughran, T. and McDonald, B. (2011).
When is a liability not a liability? textual analysis, dictionaries, and 10-ks.
The Journal of Finance, 66(1):35–65.

References V

Lovell, M. C. (1963).
Seasonal adjustment of economic time series and multiple regression analysis.
Journal of the American Statistical Association, 58(304):993–1010.

Merton, R. C. (1987).
A simple model of capital market equilibrium with incomplete information.
The journal of finance, 42(3):483–510.

Peress, J. (2008).
Media coverage and investors' attention to earnings announcements.
Available at SSRN 2723916.

Peress, J. (2014).
The media and the diffusion of information in financial markets: Evidence from newspaper strikes.
The Journal of Finance, 69(5):2007–2043.

Schmidt, D. (2013).
Investors' attention and stock covariation.
Technical report, Working paper, HEC Paris.

Solomon, D. H., Soltes, E., and Sosyura, D. (2014).
Winners in the spotlight: Media coverage of fund holdings as a driver of flows.
Journal of Financial Economics, 113(1):53–72.

References VI

Vayansky, I. and Kumar, S. A. (2020).
A review of topic modeling methods.
Information Systems, 94:101582.

Veitch, V., Sridhar, D., and Blei, D. (2020).
Adapting text embeddings for causal inference.
In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR.

Wang, X. and Yang, Y. (2020).
Neural topic model with attention for supervised learning.
In *International Conference on Artificial Intelligence and Statistics*, pages 1147–1156. PMLR.

Interpretability

Back

	Pos1	Pos2	Pos3	Neg3	Neg2	Neg1
Yelp BTR	love delici definit perfect tri super	definit ever toronto citi far amaz	best amaz everi friend free alway	custom ask said manag rude servic	never worst ever money bad terribl	disappoint tast bland dri better lack
BTR regr. weights	5.9	5.2	5.0	-8.4	-12.5	-14.5
Booking BTR	staff help friendli excel especi wonder	hotel wonder beauti love experi fabul	love great staff littl fab especi	old look carpet tire furnitur need	room small tini bathroom noisi far	poor posit servic bad never rude
BTR regr. weights	4.3	4.1	3.3	-6.1	-7.3	-14.2

Past price movements

Back

	Dependent variable:						
	$vol_{i,t}$						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$mention_{i,t}$	0.121***	0.160***	0.156***	0.137***	0.136***	0.123***	0.118***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.017)
$vol_{i,t}$ lags	1	10	10	10	10	10	20
$\lvert\Delta p\rvert_{i,t}^{o,c}$ lags			10	10	10	10	20
Volume lags				10	10	10	20
Turnover lags				10	10	10	20
$vol_{i,t}$ polynomials					✓	✓	✓
Other polynomials						✓	✓
Firm & Time FE		✓	✓	✓	✓	✓	✓
Observations	849,061	735,517	735,426	670,247	670,247	670,247	576,128
R^2	0.433	0.498	0.499	0.517	0.519	0.520	0.531
Adjusted R^2	0.430	0.495	0.495	0.513	0.516	0.517	0.527
Residual Std. Error	1.973	1.870	1.867	1.814	1.810	1.809	1.778

Note: Polynomials include $\frac{1}{2}$, 2, 3, $\frac{3}{2}$, 3 and 4 order terms

Forward looking Results

Back

	Dependent variable:				
			$vol_{i,t}$		
	(1)	(2)	(3)	(4)	(5)
$mention_{i,t}$	0.121***	0.173***	0.125***	0.135***	0.095***
	(0.016)	(0.030)	(0.029)	(0.017)	(0.016)
$future_{i,t}$		2.132**	6.029***		
$past_{i,t}$		−1.010*	−0.510		
$mention(Monday)_{i,t}$				0.213***	
$mention_{i,t+1}$					0.855***
Dictionary		LIWC	alt		
$vol_{i,t}$ lags	✓	✓	✓	✓	✓
Firm & Time FE	✓	✓	✓	✓	✓
Observations	849,016	735,517	735,517	735,517	723,882
R^2	0.433	0.498	0.498	0.498	0.501
Adjusted R^2	0.430	0.495	0.495	0.495	0.497
Residual Std. Error	1.973	1.870	1.870	1.870	1.867
Note:					*p<0.1; **p<0.05; ***p<0.01

Future focuse & LIWC dictionaries

Back

Use LIWC pastfocus and future/present focus to separate out articles which are backward-looking. This gives a percentage score for each past, present and future based on categories created to capture psychological states.

	NER		
Text	past	present	future
Logica said revenue in its telecoms business had jumped 69 per cent.	16.67	0.0	0.0
GlaxoSmithKline will consider selling part of its research operations if their discovery of new drugs does not accelerate.	0.0	16.0	4.0
Strong earnings growth is expected when oil group BP Amoco posts its first-quarter results tomorrow.	0.0	3.85	11.54

New Information

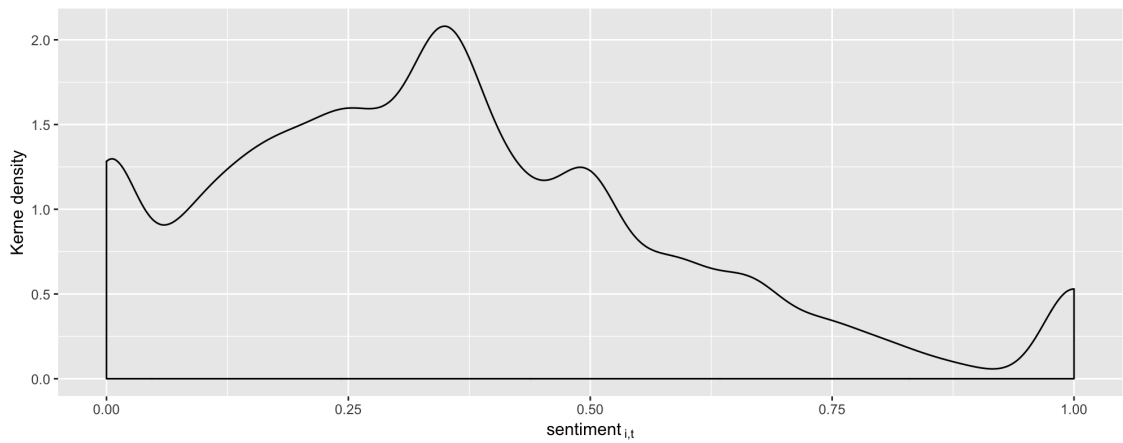
Back

Measuring sentiment

If the FT provides new information which is not yet priced in, then article sentiment should predict subsequent returns. Sentiment measure from [Loughran and McDonald, 2011]:

$$sentiment_{i,t} = \frac{positive_{i,t}}{positive_{i,t} + negative_{i,t}}$$

The sentiment measure is bounded by 0 and 1, with a mean around 0.35



New Information

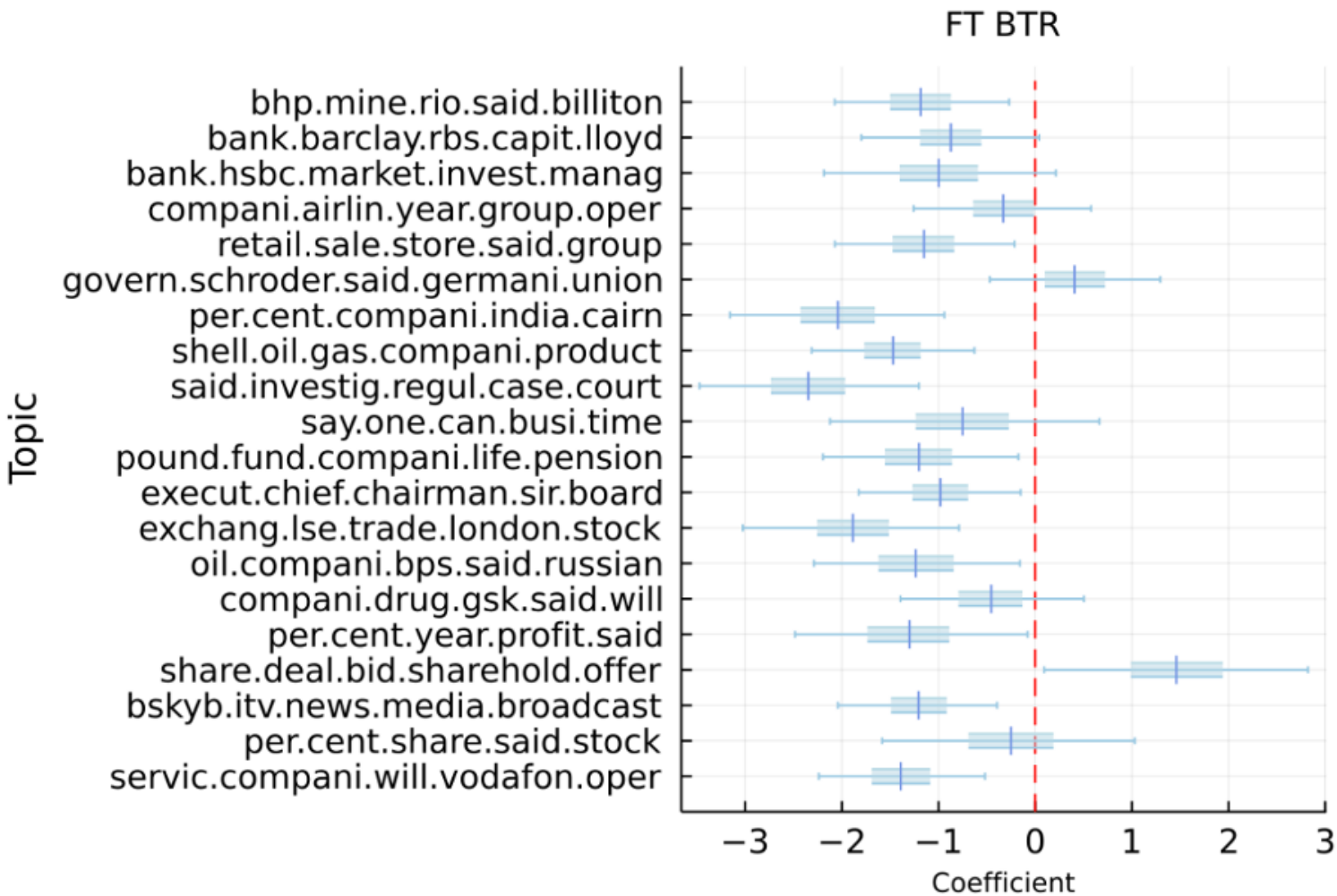
Back

Sentiment results

	Dependent variable:					
	$\Delta p^{o,c}_{i,t}$		$\Delta p^{c,o}_{i,t}$		$vol_{i,t}$	
	(1)	(2)	(3)	(4)	(5)	(6)
$mention_{i,t}$	0.015	0.023	0.035***	0.041***	0.161***	0.161***
	(0.019)	(0.019)	(0.013)	(0.013)	(0.016)	(0.016)
$sentiment_{i,t}$	0.115		0.101**		−0.277***	
	(0.073)		(0.050)		(0.064)	
$sentiment_{i,t+1}$		1.130***		0.866***		−0.137**
		(0.074)		(0.050)		(0.065)
$\Delta p^{o,c}_{i,t}$			−0.244***	−0.242***		
$\Delta p^{c,o}_{i,t}$	−0.531***	−0.528***				
Dependent lags	✓	✓	✓	✓	✓	✓
Firm & Time FE	✓	✓	✓	✓	✓	✓
Observations	719,337	707,836	719,337	707,836	735,517	723,882
R^2	0.387	0.385	0.578	0.576	0.498	0.499
Note:						*p<0.1; **p<0.05; ***p<0.01

BTR example

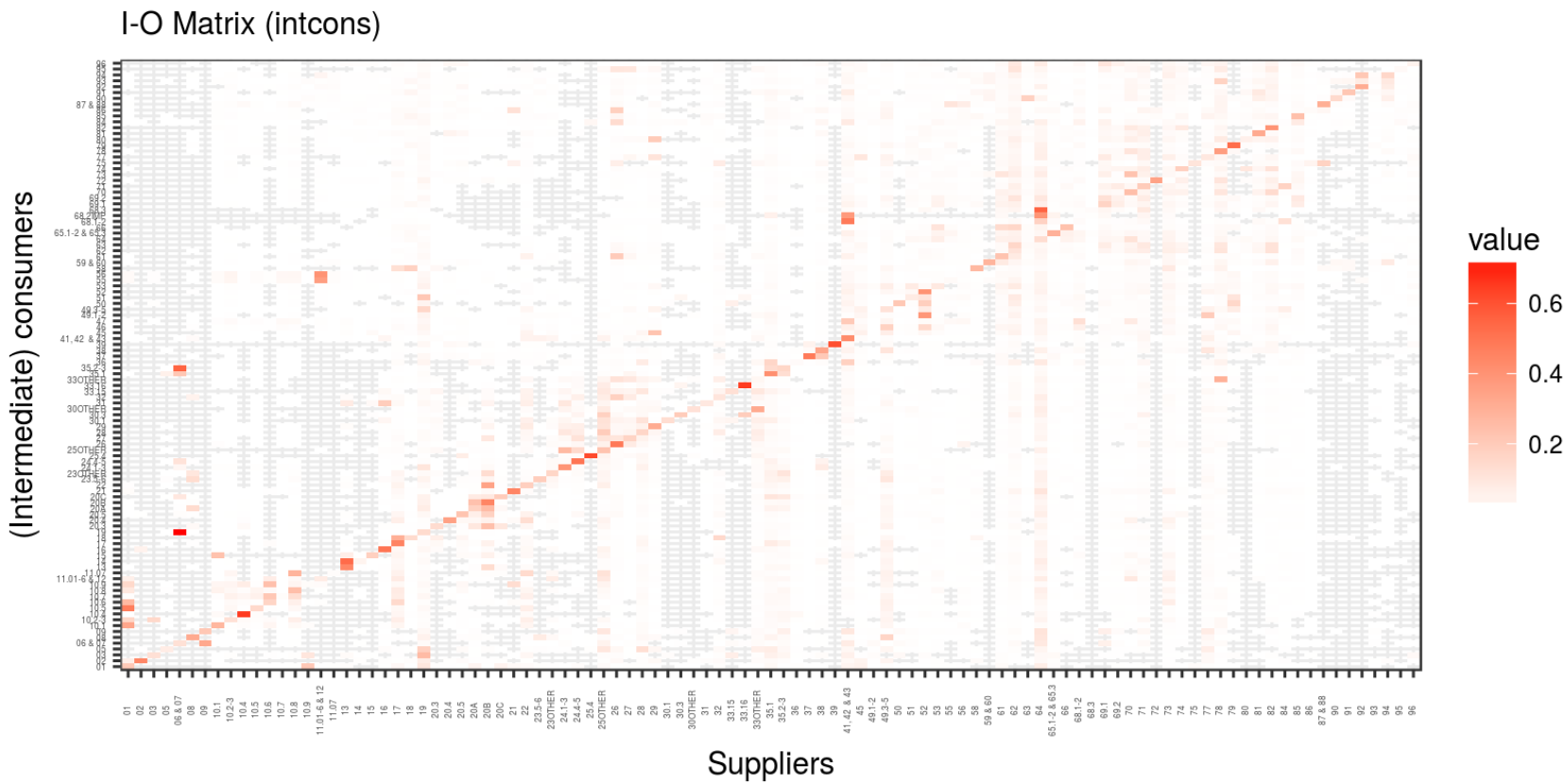
Back



Aggregate importance

Back

Production network



Aggregate effects

Back

Placebo

	Dependent variable:	
	$vol_{i,t}$	
	(1)	(2)
$mention_{i,t}$	0.161^{***}	0.087^{***}
	(0.016)	(0.013)
$\lvert\Delta p\rvert_{i,t}^{o,c}$		0.717^{***}
		(0.001)
$placebo\_mentions_{i,t}$	-0.093	-0.147
	(0.092)	(0.095)
$sector\_mentions_{i,t}$		0.067
		(0.042)
$sector\_vol_{i,t}$		0.094^{***}
		(0.002)
$vol_{i,t}$ lags	✓	✓
Implied vol	✓	✓
Firm & Time fixed effects	✓	✓
Observations	732,433	311,925
R^2	0.496	0.767
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	