Lecture 2: Unsupervised methods

Julian Ashwin

London Business School

Heidelberg, 2022

Road Map

1. What are unsupervised methods?
2. Embeddings
3. Topic models
4. Application to central bank communication

Unsupervised

Supervised versus unsupervised distinction:

1. Supervised methods aim to use some input to predict an output (e.g. regression or classification)
2. Unsupervised methods aim to summarise and identify patterns in an unlabelled dataset (e.g. principle components analysis).

Embeddings

Context

Bengio et al (2000, 2003, 2006) developed several "Neural probabilistic language models" which aim to use past words to predict the next word so generated embeddings as a by-product.

Mikolov et al (2013) introduced Word2Vec, a toolkit which allowed much faster training of vector space models of language.

Vector space models of text go back as far as the 1960s however.

Basics

You shall know a word by the company it keeps!

John Firth, 1962

Learn distributed representations of the vocabulary that capture its co-occurence statistics.

- Individual words are represented as real-valued vectors is some pre-defined space.
- Similar words to have similar representations (in that vector space), while in a bag of words model, different words have equally different representations.
- Can be pre-trained on a readily available large unannotated corpus and then used on smaller labelled datasets.
- Word2vec, GloVe and BERT include commonly used pre-trained embeddings.

Continuous Bag of Words Model

Objective of the CBOW is to predict a word given it's context.

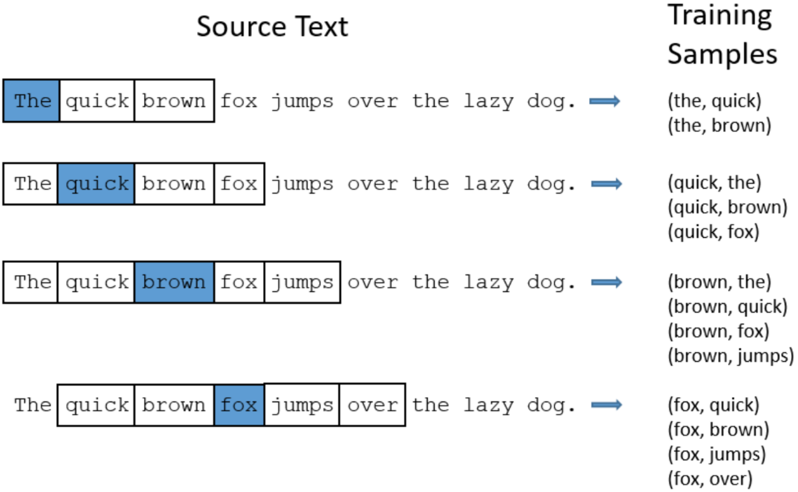
$$\sum_{t=1}^{T} \log p(w_t | w_{c_t})$$

where w_{c_t} are the words in the training context or "window".

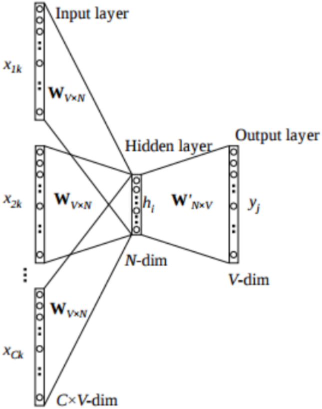
- Input is a (set of) one-hot vectors representing the context words
- Output is a softmax layer which is used to sum the probabilities obtained in the output layer to 1
- Two representation vectors for each term in the vocabulary: an "embedding" vector and a "context" vector

The conditional probability depends on the interaction between a word's embedding vector and the context vectors of the surrounding words.

Continuous Bag of Words Model



Continuous Bag of Words Model



- N here is the dimension for the vector representation of words.
- The weights between the hidden layer and the output layer is the word vector representation.

Continuous Bag-of-Words Model

Objective function

Unlike a standard MLP neural network, where the objective function is a mean squared error between the target and the output vector. The objective function of the CBOW model is a log-likelihood where

$$p(w_v|w_{c_t}) = \frac{\exp\left(\rho_v^T \alpha_{c_t}\right)}{\sum_{v=1}^{V} \exp\left(\rho_v^T \alpha_{c_t}\right)}$$

The "embedding vector" ρ is the hidden-output weights and the "context vectors" α are the input-hidden weights.

Further innovations

- Much more than I can cover, or than I know about...
- Bidirectional Encoder Representations from Transformers (BERT) is bi-directional so looks at words before and after.
- RoBERTa: A Robustly Optimized BERT Pretraining Approach, provides a version of BERT more robust to hyperparameter choice etc...
- Some approaches add an attention mechanism - different weight to co-occuring words
- Can add supervised layer to use embeddings in a prediction task
- Progress is very fast in this area, even BERT (2018) is seen as quite out of date now.
- State of the art often not very transparent.

Worked example

In the 2_unsupervised_methods.R script we estimate a simple embedding model on the FOMC minutes.

Useful to show that there are differences/similarities in language along certain dimensions, e.g. Acemoglu et al. (2022).

If we want to study the changing meaning of a word over time, the embeddings themselves can be quite interesting, e.g. has the semantic meaning of “nationalism” changed?

Can also use embeddings in a non-text context, e.g. items in a shopping cart Rudolph et al. (2017).

What is a topic model?

There are several forms of probabilistic topic model, but broadly they are:

statistical methods that analyse the words of the original texts to discover the themes that run through

Blei (2012)

Seminal and best-known is the Latent Dirichlet Allocation model developed by Blei et al. (2003).

- Assigns each word to one of K topics
- Thus each document is a distribution over topics
- Each topic is a distribution over words

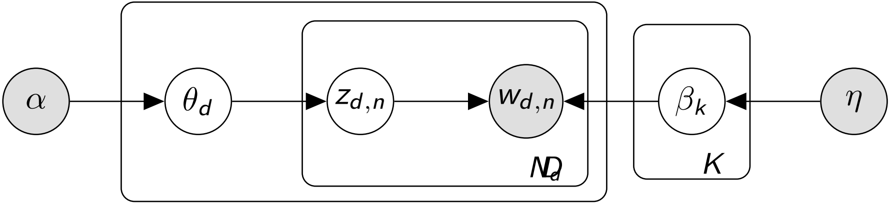
LDA Generative model

Back

1. For each of K topics, draw $\beta_k \sim \mathrm{Dir}(\eta)$
2. For each of D documents, draw $\theta_d \sim \mathrm{Dir}(\alpha)$
3. For each word n in document d:
   1. Draw topic assignment $z_{d,n}$ from $\mathrm{Mult}(\theta_d)$
   2. Draw $w_{d,n}$ from $\mathrm{Mult}(\beta_{z_{d,n}})$

Estimated at paragraph/article level by Gibbs sampling from multinomial posterior for z Griffiths and Steyvers (2004).

Graphical model



Sampling vs Variational Inference

Topic models are generally estimated through either MCMC sampling or variational inferences

Sampling (e.g. Gibbs, Metropolis Hastings, HMC) is fairly common in economics:

- Can't analytically characterise the posterior distribution
- Sample from it in a way that, eventually, approximates the true posterior arbitraily well

Variational Inference common in NLP, but rare in economics

- Make simplifying assumptions such that the variational posterior can be found as a result of an optimisation.
- No guarantee of accuracy
- Much faster - sampling is unfeasible for large models and large datasets.

Collapsed Gibbs sampling

Griffiths and Steyvers (2004) identified a way to collapse the LDA model so that only the topic assignments z need to be sampled!

Gibbs sampling involves drawing from a posterior for x_i while holding x_{-i} fixed.

Gibbs sampling algorithm requires multinomial sampling from a distribution defined by

$$\Pr[z_{d,n} = k | Z_{-(d,n)}, W, \alpha, \eta]$$

Gibbs sampling

$$\Pr[z_{d,n} = k | Z_{-(d,n)}, W, \alpha, \eta]$$

break this down into a probability of topic assignment and token assignment parts

$$\frac{\Pr[z_{d,n} = k, Z_{-(d,n)}|\alpha]}{\Pr[Z_{-(d,n)}|\alpha]} \times \Pr[z_{d,n} = k | Z_{-(d,n)}, W, \alpha, \eta] \propto \frac{\Pr[W|z_{d,n} = k, Z_{-(d}}{\Pr[W_{-(d,n)}|Z_{-(d,n)}}$$

Given the Dirichlet assumptions on the priors this can be simplified down into

$$\Pr[z_{d,n} = k | Z_{-(d,n)}, W, \alpha, \eta] \propto (s_{d,k,-n} + \alpha) \frac{m_{k,v,-(d,n)} + \eta}{\sum_v m_{k,v,-(d,n)} + V\eta}$$

The η and α are priors, V is observed in the data and the m and s terms are either being chosen or based on a previous iteration.

Algorithm

The Gibbs sampling algorithm:

- Starts with a randomly allocated topic assignment
- For each token sequentially, draw topic assignment from the multinomial distribution defined above
- Discard initial iterations as burn in and then apply thinning interval to make the draws approximately *iid*.

Back out θ and β parameters from z:

$$\hat{\theta}_{d,k} = \frac{s_{d,k} + \alpha}{\sum_k (s_{d,k} + \alpha)}$$

$$\hat{\beta}_{k,v} = \frac{m_{k,v} + \eta}{\sum_v (m_{k,v} + \eta)}$$

Worked example

In the 2_unsupervised_methods.R script we estimate a 20 topic LDA model on the FOMC minutes.

Motivation

Central banks face two choices in their communication

1. What should they talk about
2. What should they say about it

A lot of literature on the second point, e.g. forward guidance, but central banks communicate about many different dimensions of the economy and have limited communication capacity.

This paper: quantifies the *focus* of central bank communication, and shows that it is greater where there is more uncertainty.

Key Results

- Central bank communication focuses more on aspects of the economy around which:
  - There is greater uncertainty in the private sector
  - The central bank has received a more extreme signal
- The focus of central bank communication leads and potentially impacts that of the media
- The focus of Federal Reserve communication impacts the focus of other central banks.

Related Literature

- Central Bank Communication
  - Information Channel: Jarociński and Karadi (2020); Cieslak and Schrimpf (2019); Miranda-Agrippino and Ricco (2021)
  - Co-movement in communication and influence on media: Armelius et al. (2020); Binder (2017); Munday and Brookes (2021)
- Text as Data
  - Several approaches to extract tone from central bank communication Apel and Grimaldi (2012); Gonzalez et al. (2021)
  - This paper: concentrates on focus, not tone.

Multidimensional Uncertainty

- Central bank wishes to communicate information about N different dimensions of the economy.
- Communication capacity is limited.
- Devote focus to where it is most useful

Information Structure

Central Bank

We assume N such state variables that evolve exogenously.

$$X_{i,t} = \mu_1 + \rho_1 X_{i,t-1} + \epsilon_{i,t} \quad \text{where} \quad \epsilon_{i,t} \sim \mathcal{N}(0, \sigma_{\epsilon,i}) \quad \text{and} \quad i \in \{1, ..., N\}$$

CB observes a *private* signal, $s_{i,t}$, for each shocks.

$$s_{i,t} = \epsilon_{i,t} + \nu_{i,t} \quad \text{where} \quad \nu_{i,t} \sim \mathcal{N}(0, \sigma_{\nu,i})$$

CB's conditional distribution of the structural shocks is therefore

$$\epsilon_{i,t} | s_{i,t} \sim \mathcal{N}\left(\frac{\sigma^2_{\epsilon,i}}{\sigma^2_{\epsilon,i} + \sigma^2_{\nu,i}} s_{i,t}, \quad \sigma^2_{\epsilon,i}\left(1 - \frac{\sigma^2_{\epsilon,i}}{\sigma^2_{\epsilon,i} + \sigma^2_{\nu,i}}\right)\right) \tag{1}$$

Information structure

Private sector

CB devote a fraction of fixed communication to each variable $1 \leq a_{i,t} \geq 0$, creating a public version of it's signal.

$$\hat{s}_{i,t} = s_{i,t} + \eta_{i,t} \quad \text{where} \quad \eta_{i,t} \sim \mathcal{N}(0, (1 - a_{i,t})^2)$$

Private sector has M agents who receive private signals

$$q_{m,i,t} = \epsilon_{i,t} + \zeta_{m,i,t} \quad \text{where} \quad \zeta_{m,i,t} \sim \mathcal{N}\left(0, \sigma_{\zeta,i}\right)$$

Private sector's expectation of $X_{i,t}$ is given by

$$\begin{aligned} \mathbb{E}_{m,t}[X_{i,t}|\hat{s}_{i,t}] &= \mu_i + \rho_i X_{i,t-1} + \mathbb{E}_t[\epsilon_{i,t}|q_{m,i,t}, \hat{s}_{i,t}] \\ &= \lambda_{q,i,t} q_{i,t} + \lambda_{\hat{s},i,t} \hat{s}_{i,t} \end{aligned} \tag{2}$$

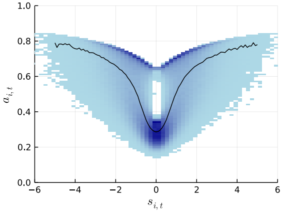
Central Bank problem

Minimise private sector error variance.

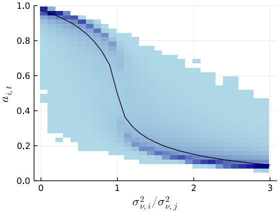
$$
\begin{gathered}
\min_{\mathbf{a}_t} L(\mathbf{a}_t, \mathbf{s}_t) = \mathbb{E}\left[\sum_i \sum_m \left(\mathbb{E}[X_{i,t}|q_{m,i,t}, \hat{s}_{i,t}] - X_{i,t}\right)^2 |s_{i,t}\right] \\
\text{s.t.} \\
\sum_i a_{i,t} = 1, \\
a_{i,t} \geq 0, \quad \text{for} \quad i \in \{1, ..., N\}
\end{gathered} \tag{3}
$$

Central Bank Focus

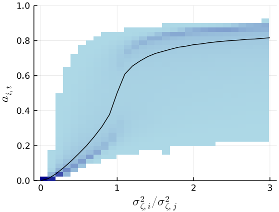
(a) Extreme $s_{i,t}$ increases $a_{i,t}$



(b) Volatile $\nu_{i,t}$ decreases $a_{i,t}$



(c) Volatile $\zeta_{i,t}$ increases $a_{i,t}$



Data

Quantifying the focus of communication and uncertainty over different dimensions of the economy.

1. Uncertainty:
   - Survey of Professional Forecasters give private sector expectations
   - Tealbook gives Fed signals and expectations
2. Focus
   - FOMC minutes
   - FOMC speeches
   - NYT news articles

Data

Uncertainty: series covered by Tealbook and SPF

SPF code	Tealbook code	Description	Key terms
NGDP	gNGDP	Nominal GDP growth	economi, growth
RGDP	gRGDP	Real GDP growth	economi, growth
CPI	gPCPI	CPI inflation	price, inflat
UNEMP	UNEMP	Unemployment Rate	job, employ
EMP	-	Nonfarm Employment	job, emp
CPROF	-	Corporate Profits (after tax)	corpor, profit
INDPROD	gIP	Industrial Production Index	industri, manufac
HOUSING	HSTART	Housing starts	hous, home
RRESINV	gRRES	Residential Investment	hous, home
RNRESIN	gBF	Nonresidential Investment	invest, capit
RCONSUM	gRPCE	Personal Consumption Expenditure	spend. consum
RFEDGOV	gRGOVF	Federal Government Expenditure	tax, budget
RSLGOV	gRGOVSL	State Government Expenditure	tax, budget

Uncertainty

Three measures

1. Dispersion in SPF nowcasts $disp^{\text{SPF}}_{k,t}$ acts as a proxy for the noisiness of the private sectors private signals (σ^2_ζ).
2. Tealbook update is gap between forecasts at $t-1$ and nowcast at t.

$$s^{Fed}_{k,t} = |\mathbb{E}^{GB}_t x_{k,t} - \mathbb{E}^{GB}_{t-1} x_{k,t}|$$

3. Tealbook-SPF error difference as the difference between the absolute nowcast errors, which acts as a proxy for the accuracy of the Fed's signal.

$$\nu^{Fed}_{k,t} = |\mathbb{E}^{GB}_t x_{k,t} - x_{k,t}| - |\mathbb{E}^{SPF}_t x_{k,t} - x_{k,t}|$$

Data

Text

1. Published FOMC minutes from Jan 1993 to December 2017
2. Published speeches by FOMC member from June 1996 to December 2017
3. Articles are taken from *New York Times* (NYT) between January 1993 to December 2017, tagged as "economic news"

Corpus	Total documents	Total paragraphs	Total words
FOMC minutes	235	9,133	1,163,211
FOMC speeches	1,289	40,433	4,012,218
New York Times	72,763	NA	30,655,062

Standard cleaning carried out on all documents. ▸ Cleaning

Quantifying focus of communication

Estimate a LDA models over combined FOMC and NYT corpora.

- Probabilistic model that assigns each word to one of these 30 topics.
- For each topic, we have a distribution over words in the vocabulary (what is the topic about).
- For each document, we have a distribution over topics (what is the document about)

▸ LDA generative model

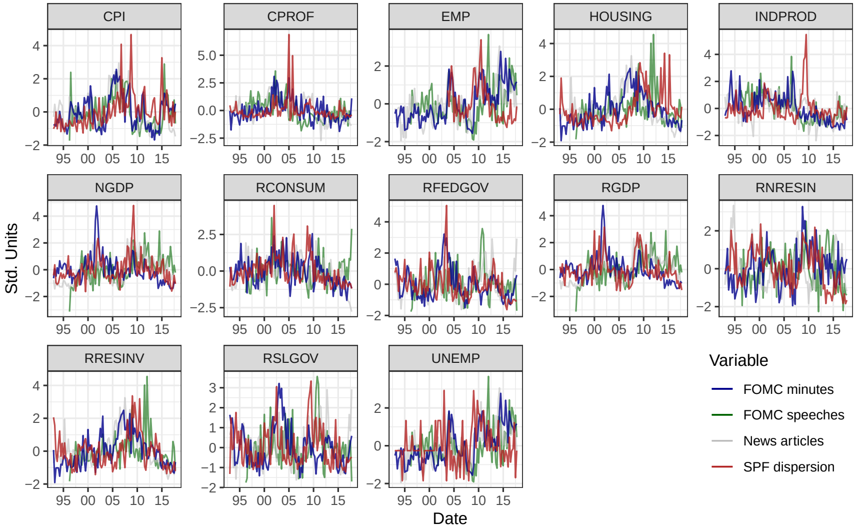
Key difference over sentiment/hawkishness measures that are often used, is that we quantify the *focus* of communication rather than tone.

Match topics to economic variables through prevalence of manually selected key terms.

Topics

Topic	Description	Top Terms	$\bar{\theta}_k^{mins}$	$\bar{\theta}_k^{speech}$	$\bar{\theta}_k^{news}$	Macro Variables
1	Healthcare	pay, cost, health, care, insur, save, year, plan, benefit, mani,	0.021	0.035	0.035	
2	Interest rates	fed, interest, feder, reserv, polici, economi, rais, term, point, meet	0.029	0.038	0.036	
3	Inflation	price, inflat, oil, increas, energi, month, rise, consum, measur, higher	0.055	0.032	0.024	CPI
4	China	china, unit, state, world, countri, global, american, chines, develop, asia	0.022	0.032	0.035	
5	Committe views	particip, member, risk, note, recent, continu, outlook, effect, growth, howev	0.084	0.050	0.019	
6	Education	school, incom, peopl, famili, univers, educ, work, student, american, studi	0.021	0.037	0.037	
7	Data I	quarter, growth, busi, pace, continu, fourth, inventori, remain, increas, spend	0.080	0.031	0.021	
8	Real estate	hous, home, real, year, mortgag, new, estat, apart, price, start	0.030	0.028	0.029	HOUSING,RRESINV
9	Expectations	expect, period, term, continu, remain, declin, project, chang, longer, forecast	0.077	0.035	0.017	
10	Bond markets	bond, secur, treasuri, yield, million, debt, issu, note, agenc, week	0.034	0.028	0.028	
11	Investment	fund, invest, investor, manag, money, stock, capit, financi, firm, asset	0.026	0.042	0.032	RNRESIN
12	Production	product, industri, manufactur, factori, high, car, produc, technolog, like, cost	0.029	0.034	0.030	INDPROD
13	Fiscal policy	tax, cut, budget, state, spend, year, billion, govern, plan, propos	0.022	0.028	0.038	RFEDGOV,RSLGOV
14	General I	one, make, way, get, now, like, even, think, say, just	0.020	0.047	0.049	
15	Policy decision	committe, polici, feder, inflat, monetari, condit, reserv, direct, fund, financi	0.092	0.050	0.018	
16	Infrastructure	citi, new, build, develop, york, plan, project, center, offic, million, area	0.021	0.031	0.041	
17	Politics	presid, administr, elect, vote, polit, hous, support, white, polici, issu	0.020	0.031	0.044	
18	Consumption	consum, sale, report, spend, month, retail, increas, data, expect, good	0.041	0.027	0.030	RCONSUM
19	General II	peopl, one, year, like, day, time, say, work, now, just	0.020	0.029	0.048	
20	Europe	european, bank, europ, euro, countri, central, union, govern, germani, german	0.022	0.031	0.036	
21	Corporations	compani, busi, million, billion, execut, share, profit, year, corpor, oper	0.020	0.028	0.041	CPROF
22	Finance	bank, loan, credit, financi, borrow, debt, mortgag, interest, lend, financ	0.033	0.056	0.033	
23	Japan	trade, dollar, japan, foreign, export, currenc, import, unit, american, japanes	0.030	0.029	0.030	
24	Foreign policy	govern, polit, nation, countri, minist, russia, war, offici, power, leader, militari	0.020	0.030	0.045	
25	Stock market	stock, percent, point, index, fell, rose, investor, share, gain, week	0.025	0.024	0.049	
26	Growth	economi, year, recess, last, still, now, even, time, fall, growth	0.024	0.032	0.040	NGDP,RGDP
27	Legal	state, new, offici, rule, law, deal, group, case, agenc, regul	0.022	0.048	0.042	
28	Data II	percent, year, last, month, sinc, averag, increas, first, three, annual	0.026	0.026	0.044	
29	Labour market	job, worker, unemploy, labor, employ, wage, work, month, benefit, week	0.032	0.029	0.030	EMP,UNEMP

Topics and SPF dispersion



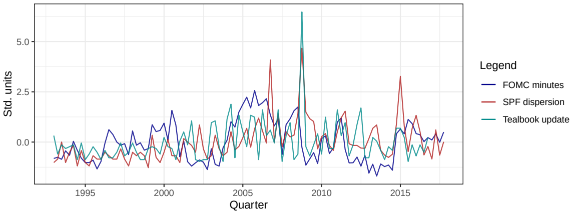
Correlations

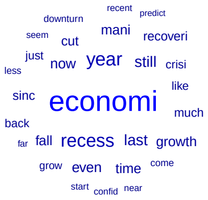
	FOMC minutes			FOMC speeches			NYT articles		
Variable	$disp^{SPF}_{k,t}$	$s^{Fed}_{k,t}$	$\nu^{Fed}_{k,t}$	$disp^{SPF}_{k,t}$	$s^{Fed}_{k,t}$	$\nu^{Fed}_{k,t}$	$disp^{SPF}_{k,t}$	$s^{Fed}_{k,t}$	$\nu^{Fed}_{k,t}$
CPI	0.288***	0.210**	-0.175*	0.031	0.057	-0.024	0.225**	0.196*	-0.215**
CPROF	0.417***			-0.098			0.140		
EMP	0.051			-0.201			0.197		
HOUSING	0.204**	0.317***	-0.201	0.105	0.076	0.105	0.099	0.227**	-0.091
INDPROD	0.101	0.043	-0.060	-0.121	-0.225**	0.007	0.176*	0.097	-0.157
NGDP	0.372***	0.203**	0.030	0.193*	-0.024	-0.016	0.451***	0.076	0.029
RCONSUM	0.489 ***	0.245 **	-0.173 *	0.036	-0.025	0.074	0.12	-0.02	0.177 *
RFEDGOV	0.353 ***	0.226 **	-0.053	0.004	0.185 *	-0.164	0.134	0.271 ***	-0.157
RGDP	0.356 ***	0.362 ***	-0.179 *	0.088	0.026	0.03	0.469 ***	0.285 ***	-0.157
RNRESIN	0.233 **	0.232 **	-0.123	0.034	-0.023	0.004	0.042	-0.082	-0.039
RRESINV	0.344 ***	0.163	-0.093	0.169	0.078	0.013	0.3 ***	0.029	-0.154
RSLGOV	0.175 *	0.121	0.04	0.078	0.151	0.085	0.179 *	0.154	-0.075
UNEMP	0.012	0.096	0.065	-0.058	0.035	0.166	0.21 **	0.074	0.009
Overall	0.267***	0.200***	-0.067**	0.024	0.029	0.024	0.211***	0.119***	-0.078**

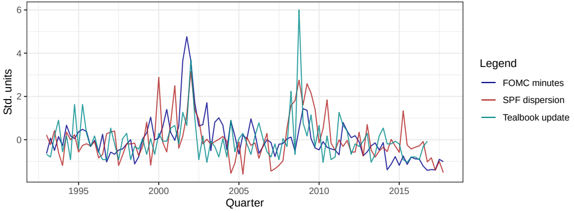
Note: *p<0.1; **p<0.05; ***p<0.01

Examples


import
gasolin
expect
core
product
remain
recent
month
increas
cost
declin
energi
measur
price
index
rise
oil
year
level
end
higher
consum
low
food
inflat
compens
produc
pressur
commod
earlier


5.0
2.5
0.0
Std. units
1995
2000
2005
2010
2015
Quarter
Legend
FOMC minutes
SPF dispersion
Tealbook update


recent
predict
downturn
mani
seem
cut
recoveri
just
now
year
still
crisi
less
like
sinc
economi
much
back
far
fall
recess
last
growth
grow
even
time
come
start
confid
near


6
4
2
0
Std. units
1995
2000
2005
2010
2015
Quarter
Legend
FOMC minutes
SPF dispersion
Tealbook update

Central bank focus and Uncertainty

Variables

- Focus of FOMC and NYT

$$\theta_{k,t}^{\text{mins}}, \quad \theta_{k,t}^{\text{speech}}, \quad \theta_{k,t}^{\text{NYT}}$$

- SPF and Tealbook variables

$$disp_{k,t}^{\text{SPF}}, s_{k,t}^{GB}, \nu_{k,t}^{Fed}$$

- Estimate topic-quarter panel regressions (all variables are standardised to zero mean and unit variance).

$$\theta_{k,t}^{mins} = \alpha_k + \mu_t + \beta disp_{k,t}^{SPF} + \sum_{p=1}^{P} \rho_p \theta_{k,t-p}^{mins} + u_{k,t}$$

SPF dispersion

	Dependent variable:					
	$\theta^{mins}_{k,t}$		$\theta^{speech}_{k,t}$		$\theta^{news}_{k,t}$	
	(1)	(2)	(3)	(4)	(5)	(6)
$disp^{SPF}_{k,t}$	0.278***	0.120***	0.011	−0.021	0.204***	−0.021
	(0.050)	(0.033)	(0.027)	(0.029)	(0.038)	(0.023)
Dep variable lags		3		3		3
Topic fixed effects	✓	✓	✓	✓	✓	✓
Time fixed effects		✓		✓		✓
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.098	0.545	0.003	0.280	0.060	0.521
Adjusted R^2	0.086	0.498	−0.010	0.206	0.048	0.472
Residual Std. Error	0.986	0.731	0.991	0.879	0.991	0.738

Note: Driscoll-Kraay standard errors: *p<0.1; **p<0.05; ***p<0.01

Tealbook updates

	Dependent variable:					
	$\theta_{k,t}^{mins}$		$\theta_{k,t}^{speech}$		$\theta_{k,t}^{news}$	
	(1)	(2)	(3)	(4)	(5)	(6)
$s_{k,t}^{Fed}$	0.203***	0.086***	0.025	−0.006	0.117***	0.026
	(0.064)	(0.026)	(0.030)	(0.029)	(0.027)	(0.030)
Dep variable lags		3		3		3
Topic fixed effects	✓	✓	✓	✓	✓	✓
Time fixed effects		✓		✓		✓
Observations	858	858	858	858	858	858
R^2	0.050	0.548	0.003	0.275	0.032	0.509
Adjusted R^2	0.038	0.494	−0.009	0.189	0.019	0.450
Residual Std. Error	1.012	0.734	0.980	0.878	0.982	0.735

Note: Driscoll-Kraay standard errors: *p<0.1; **p<0.05; ***p<0.01

Tealbook-SPF error difference

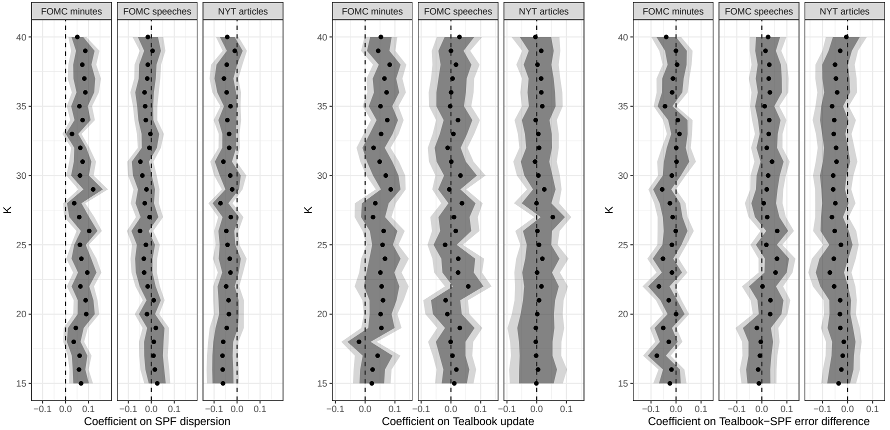
	Dependent variable:					
	$\theta^{mins}_{k,t}$		$\theta^{speech}_{k,t}$		$\theta^{news}_{k,t}$	
	(1)	(2)	(3)	(4)	(5)	(6)
$\nu^{Fed}_{k,t}$	-0.076^{*}	-0.055^{*}	0.015	0.020	-0.097^{***}	-0.057
	(0.041)	(0.028)	(0.035)	(0.034)	(0.030)	(0.037)
Dep variable lags		3		3		3
Topic fixed effects	✓	✓	✓	✓	✓	✓
Time fixed effects		✓		✓		✓
Observations	836	836	836	836	836	836
R^2	0.015	0.535	0.004	0.274	0.029	0.502
Adjusted R^2	0.002	0.480	-0.010	0.187	0.016	0.442
Residual Std. Error	1.028	0.742	0.982	0.881	0.980	0.737
Note:	Driscoll-Kraay standard errors: *p$<$0.1; **p$<$0.05; ***p$<$0.01					

Three measures

	Dependent variable:		
	$\theta^{mins}_{k,t}$		
	(1)	(2)	(3)
$disp^{SPF}_{k,t}$	0.217***	0.084**	0.064*
	(0.044)	(0.034)	(0.034)
$s^{Fed}_{k,t}$	0.127**	0.058*	0.059**
	(0.059)	(0.030)	(0.027)
$\nu^{Fed}_{k,t}$	0.027	−0.025	−0.006
	(0.039)	(0.031)	(0.028)
$\theta^{news}_{k,t}$			0.236***
			(0.039)
Dep variable lags		3	3
Topic fixed effects	✓	✓	✓
Time fixed effects		✓	✓
Observations	836	836	836
R^2	0.091	0.543	0.582
Adjusted R^2	0.077	0.487	0.531
Residual Std. Error	0.989	0.737	0.705

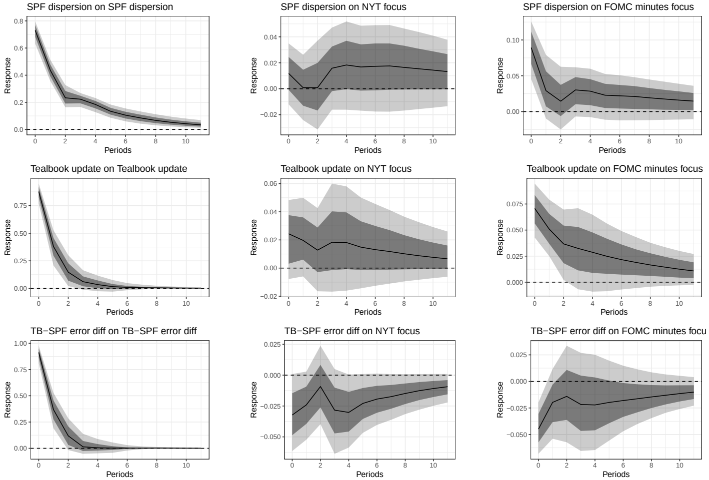
Note: DK standard errors: *p<0.1; **p<0.05; ***p<0.01

Robustness to K

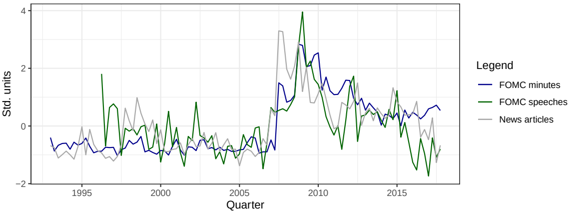


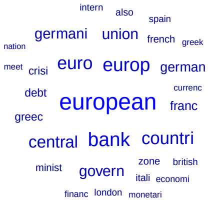
Panel VAR

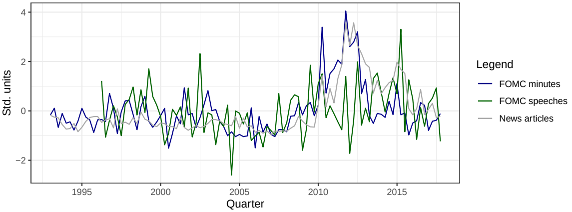
IRFs



Other topics







Publication of FOMC minutes impacts media focus

Two questions:

1. Do media articles the week prior to a meeting predict the minutes' focus?
2. Do minutes predict the content of media articles in the week following their publication?

Figure: Media focus around FOMC meeting

Window around minutes: $\Delta\theta^{news}_{m_p,k} = \theta^{news}_{m_p+w,k} - \theta^{news}_{m_p-w,k}$
Window around speeches: $\Delta\theta^{news}_{s,k} = \theta^{news}_{s+w,k} - \theta^{news}_{s-w,k}$

CBC and media focus event study

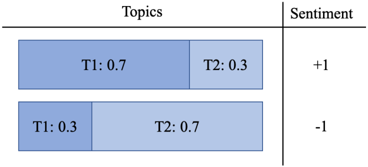
	Dependent variable:			
	$\Delta\theta^{news}_{m_p,k}$		$\Delta\theta^{news}_{s,k}$	
	(1)	(2)	(3)	(4)
$\theta^{mins}_{m,k}$	0.021**	0.026***		
	(0.008)	(0.010)		
$\theta^{speech}_{s,k}$			0.067***	0.062***
			(0.009)	(0.009)
Topic fixed effects	✓	✓	✓	✓
Time fixed effects	✓	✓	✓	✓
Topic-specific γ		✓		✓
Observations	5,742	5,742	31,784	31,784
R^2	0.369	0.402	0.540	0.546
Adjusted R^2	0.343	0.368	0.523	0.529
Residual Std. Error	0.008	0.008	0.018	0.018
Note:	DK standard errors: *p<0.1; **p<0.05; ***p<0.01			

Topic-specific tone

So the CB can influence what the media focuses on, but can they also influence the tone of coverage?

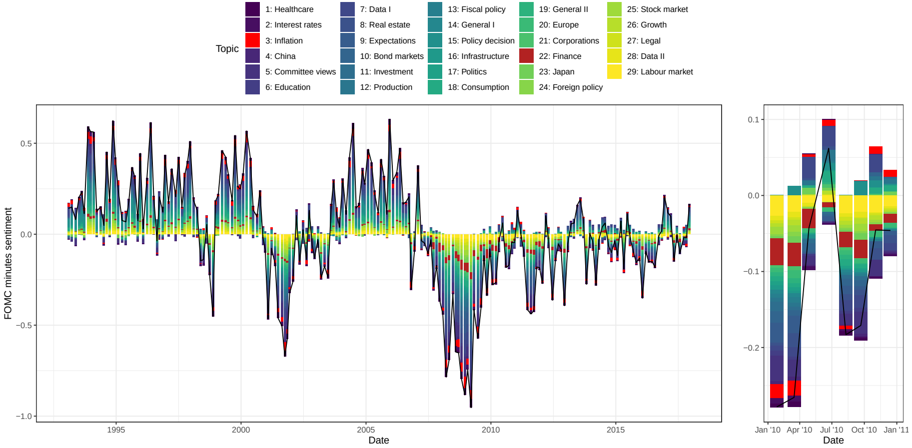
Use sentiment dictionaries from Loughran and McDonald (2011) to create a topic-specific tone metric.

$$\vartheta_{m,k}^{mins} = \theta_{m,k}^{mins} \times sent_{m}^{mins},$$



Overall sentiment is 0, $\vartheta_1^{mins} = 0.4$ and $\vartheta_2^{mins} = -0.4$

Sentiment decomposition

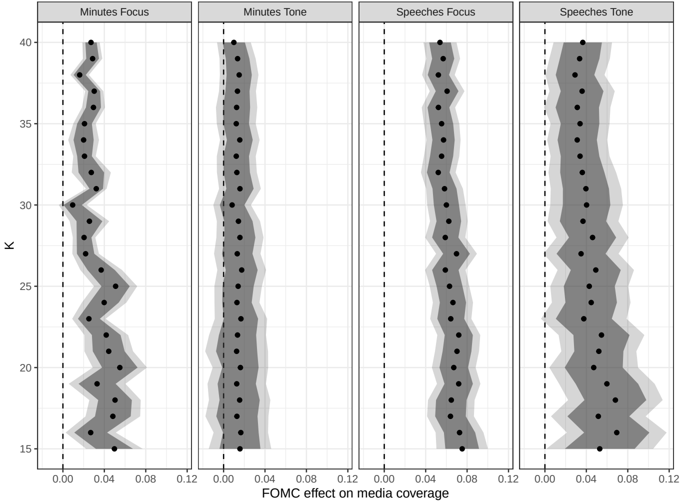


CB influence on tone

	Dependent variable:					
	$\Delta\vartheta^{news}_{m_p,k}$		$\Delta\vartheta^{news}_{s,k}$		$\Delta sent^{mins}_{m_p}$	$\Delta sent^{news}_{s}$
	(1)	(2)	(3)	(4)	(5)	(6)
$\vartheta^{mins}_{m,k}$	−0.009	0.014				
	(0.010)	(0.011)				
$\theta^{mins}_{m,k}$	−0.017**	−0.023***				
	(0.007)	(0.008)				
$\vartheta^{speech}_{s,k}$			0.036**	0.037**		
			(0.018)	(0.017)		
$\theta^{speech}_{s,k}$			−0.014*	−0.014*		
			(0.008)	(0.007)		
$sent^{mins}_{m}$					0.144**	
					(0.058)	
$sent^{speech}_{s}$						0.081**
						(0.032)
Topic fixed effects	✓	✓	✓	✓		
Time fixed effects	✓	✓	✓	✓		
Topic-specific γ		✓		✓		
Observations	5,742	5,742	31,784	31,784	198	1,096
R^2	0.619	0.633	0.698	0.700	0.276	0.431
Adjusted R^2	0.603	0.612	0.687	0.689	0.261	0.430
Residual Std. Error	0.008	0.007	0.018	0.018	0.184	0.416

Note: Driscoll-Kraay standard errors: *p<0.1; **p<0.05; ***p<0.01

Robustness to different K



Conclusion

- Examines which dimensions of the economy a CB chooses to devote its communication to.
- Model: where can the CB add the most value?
- Data:
  - Quantify focus and uncertainty in meaningful way
  - FOMC minutes place greater focus where model suggests is useful
  - FOMC speeches do not
- Can the CB convey information to the public?
  - CB communication can influence focus and tone of media coverage
  - Speeches have greater influence than minutes

Thank you for your attention!

References I

Acemoglu, D., Mühlbach, N. S., and Scott, A. J. (2022). The rise of age-friendly jobs. *The Journal of the Economics of Ageing*, page 100416.

Apel, M. and Grimaldi, M. (2012). The information content of central bank minutes. *Riksbank Research Paper Series No. 92*.

Armelius, H., Bertsch, C., Hull, I., and Zhang, X. (2020). Spread the word: International spillovers from central bank communication. *Journal of International Money and Finance*, 103:102116.

Binder, C. (2017). Fed speak on main street: Central bank communication and household expectations. *Journal of Macroeconomics*, 52:238–251.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

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

Cieslak, A. and Schrimpf, A. (2019). Non-monetary news in central bank communication. *Journal of International Economics*.

Gonzalez, M., Tadle, R. C., et al. (2021). Monetary policy press releases: An international comparison. Technical report, Central Bank of Chile.

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

Jarociński, M. and Karadi, P. (2020). Deconstructing monetary policy surprises—the role of information shocks. *American Economic Journal: Macroeconomics*, 12(2):1–43.

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

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.

Miranda-Agrippino, S. and Ricco, G. (2021). The transmission of monetary policy shocks. *American Economic Journal: Macroeconomics*, 13(3):74–107.

Munday, T. and Brookes, J. (2021). Mark my words: the transmission of central bank communication to the general public via the print media.

Rudolph, M., Ruiz, F., Athey, S., and Blei, D. (2017). Structured embedding models for grouped data. In *Advances in neural information processing systems*, pages 251–261.

Text cleaning ▸ Back

- Split documents into paragraphs, remove preamble and administrative details.
- The documents are stripped of any additional white-spaces so that each term is separated by a single space.
- All numerical characters are removed, as is any punctuation.
- All characters are transformed to lower case and common stop-words are removed using the list provided by Lewis et al. (2004).
- The remaining terms are then stemmed using the Porter stemming algorithm, reducing each word to its root
- Terms fewer than three characters in length are removed.
- Remove all names of months and seasons as the obvious seasonality of these terms might generate spurious co-movement.
- Remove terms which appear in only one of the corpora

Text cleaning

Example

Raw text

Short term interest rates have registered small mixed changes since the day before the Committee meeting on November 12 1997 while bond yields have fallen somewhat. Share prices in U.S. equity markets recorded mixed changes over the period; equity markets in other countries notably in Asia have remained volatile. In foreign exchange markets the value of the dollar has risen over the intermeeting period in terms of both the trade weighted index of the other G 10 countries and the currencies of a number of Asian countries.

Clean text

short term interest rate regist small mix chang sinc day committe meet novemb bond yield fallen somewhat share price equiti market record mix chang period equiti market countri notabl asia remain volatil foreign exchang market valu dollar risen intermeet period term trade weight index countri currenc number asian countri

Clean text with corpus-specific and seasonal terms removed

short term interest rate regist small mix chang sinc day committe meet bond yield fallen somewhat share price equiti market record mix chang period equiti market countri notabl asia remain volatil foreign exchang market valu dollar risen period term trade weight index countri currenc number asian countri

Do central banks influence one another?

Summary

- Significant and robust co-movement of focus across FOMC, MPC and GC.
- FOMC communication Granger causes that of the MPC and GC.
- The focus of the most recently published communication has similar cross-central bank effects.
- Change in the publication policy of the FOMC's minutes can be used to show that they may have a causal influence on the MPC minutes.

Quantifying cross-CB focus

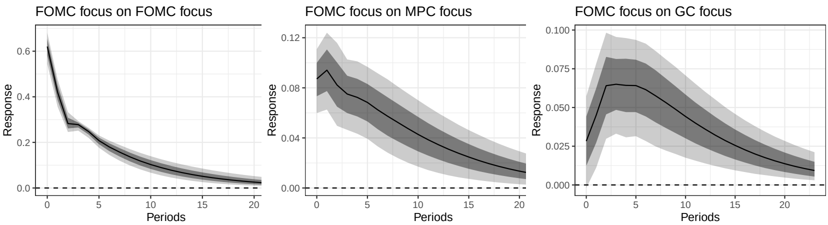
- Same process as for FOMC-NYT corpus
- Some CB-specific topics (e.g. 27 and 28)

Topic	Description	Top 5 words	$\bar{\theta}_k^{BoE}$	$\bar{\theta}_k^{ECB}$	$\bar{\theta}_k^{Fed}$
Topic 1	Economic data	fallen, sinc, risen, fall, average	0.0575	0.0126	0.0091
Topic 2	Growth expectations	seem, might, prospect, slowdown, recoveri	0.0641	0.0143	0.0131
Topic 3	Staff projections	project, forecast, report, staff, central	0.0348	0.0233	0.0240
Topic 4	International trade	trade, import, export, foreign, net	0.0310	0.0144	0.0428
Topic 5	Cost push factors	pressure, cost, product, wage, capac	0.0497	0.0212	0.0225
Topic 6	Inflation expectations	inflat, risk, target, committee, view	0.0641	0.0114	0.0180
Topic 7	Hypotheticals	might, possibl, earn, pay, one	0.0646	0.0106	0.0106
Topic 8	GDP data	quarter, first, second, gdp, estim	0.0394	0.0253	0.0335
Topic 9	Household consumption	consum, spend, household, consumpt, incom	0.0330	0.0112	0.0417
Topic 10	Credit conditions	credit, bank, loan, financi, lend	0.0360	0.0510	0.0220
Topic 11	Business investment	busi, invest, inventori, spend, capit	0.0212	0.0099	0.0680
Topic 12	Market expectations	particip, econom, note, improve, longer	0.0120	0.0124	0.0674
Topic 13	FOMC	comitte, feder, percent, consist, reserve	0.0073	0.0090	0.0737
Topic 14	Fiscal reforms	fiscal, countri, govern, reform, structur	0.0168	0.1310	0.0126
Topic 15	Core inflation	inflat, energi, oil, core, cpi	0.0326	0.0330	0.0431
Topic 16	Committee expectations	member, expans, prospect, factor, persist	0.0199	0.0179	0.0741
Topic 17	Output data	survey, data, output, manufactur, servic	0.0674	0.0130	0.0112
Topic 18	Interest rate	interest, point, short, basi, reduct	0.0459	0.0221	0.0178
Topic 19	Labour market	labour, employ, unemploy, measur, privat	0.0302	0.0173	0.0381
Topic 20	Policy committee	polici, member, committe, monetari, econom	0.0235	0.0150	0.0685
Topic 21	Bond market	period, yield, bond, spread, fund	0.0252	0.0115	0.0518
Topic 22	Policy decision	polici, financi, committe, decis, discuss	0.0322	0.0159	0.0248
Topic 23	Exchange rates	unit, state, sterl, dollar, exchang	0.0484	0.0102	0.0212
Topic 24	Industrial production	product, industri, moder, rose, manufactur	0.0097	0.0079	0.0805
Topic 25	Quantitative easing	bank, purchas, asset, committe, vote	0.0403	0.0111	0.0158
Topic 26	Housing market	hous, mortgag, home, sale, new	0.0237	0.0090	0.0463
Topic 27	ECB GC	govern, council, will, meet, ecb	0.0091	0.1040	0.0090
Topic 28	Eurozone	euro, area, econom, recoveri, global	0.0234	0.1121	0.0075
Topic 29	Monetary stability	monetari, medium, stabil, develop, econom	0.0121	0.1731	0.0113
Topic 30	Risk	risk, develop, uncertainti, downsid, global	0.0249	0.0694	0.0199

Quarterly panel VAR

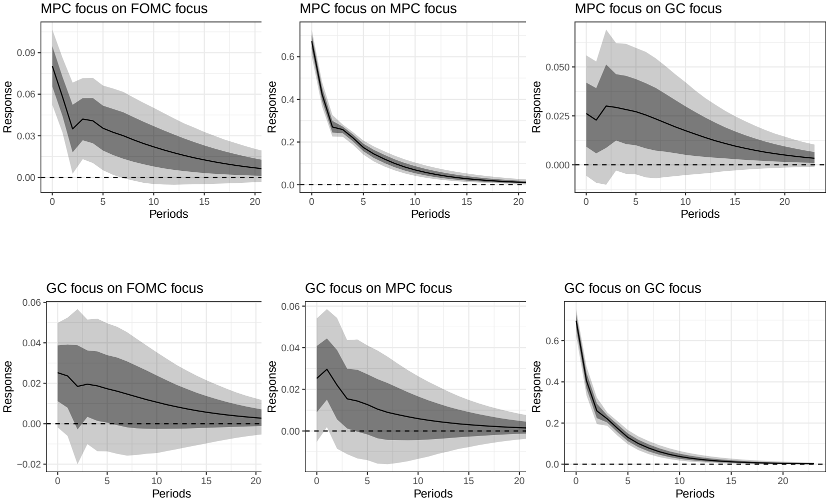
$$\theta_{k,t} = \alpha_k + \sum_{l=1}^{p} A_l \theta_{k,t-l} + \varepsilon_{k,t}$$

Figure: Generalised IRFs for shock to FOMC focus.



Note: The darker band represents the 70% confidence interval and the lighter the 95% confidence interval.

Other IRFs


MPC focus on FOMC focus
MPC focus on MPC focus
MPC focus on GC focus
GC focus on FOMC focus
GC focus on MPC focus
GC focus on GC focus
Response
Periods

Potentially influential publications

We use two alternative approaches to define “recently published” communication:

1. Most recent: other central banks’ most recently *published* piece of communication, prior to the meeting date.
2. 3 month window: an average of a central bank’s communication published in a rolling window of three months prior to a meeting.

Regression specification will be the same for both.

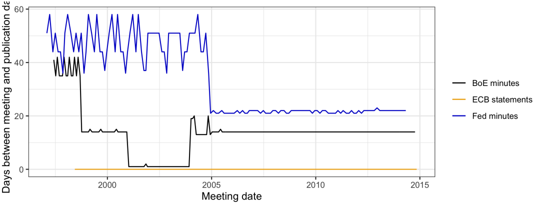
$$\theta^b_{b,m,k} = \alpha_{b,k} + \sum_{c \neq b} \gamma_{b,c} \theta^c_{b,m,k} + \sum_{p=1}^{P} \rho_{b,p} \theta^b_{b,m-p,k} + \text{controls} + \varepsilon_{b,m,k}$$

$\gamma_{\text{Fed,BoE}}$ indicates the effect of the BoE’s recently published communication on that of the Fed

Recently published

	Empirical strategy					
	3 month window			Most recent		
	(1)	(2)	(3)	(4)	(5)	(6)
$\gamma_{BoE,Fed}$	0.129***	0.053***	0.054***	0.108***	0.054***	0.052***
	(0.014)	(0.013)	(0.013)	(0.012)	(0.011)	(0.012)
$\gamma_{ECB,Fed}$	0.107***	0.068***	0.054***	0.822***	0.051***	0.039***
	(0.014)	(0.014)	(0.014)	(0.012)	(0.012)	(0.012)
$\gamma_{Fed,BoE}$	0.101***	0.042***	0.035*	0.068***	0.041***	0.035**
	(0.020)	(0.019)	(0.019)	(0.015)	(0.014)	(0.014)
$\gamma_{ECB,BoE}$	0.068***	0.030*	0.029**	0.030*	0.043***	0.037***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.018)	(0.018)
$\gamma_{Fed,ECB}$	0.075***	0.018	0.018	0.048***	0.015	0.013
	(0.019)	(0.018)	(0.018)	(0.014)	(0.013)	(0.013)
$\gamma_{BoE,ECB}$	0.047***	0.007	0.009	0.030***	0.006	0.005
	(0.015)	(0.014)	(0.015)	(0.012)	(0.012)	(0.012)
CB-specific lags	1	10	10	1	10	10
Macro controls			✓			✓
CB-topic FE	✓	✓	✓	✓	✓	✓
Observations	15,330	15,060	15,060	15,330	15,060	15,060
R^2	0.204	0.298	0.309	0.202	0.299	0.309
Residual Std. Error	0.885	0.823	0.819	0.886	0.822	0.819

Natural Experiment



Regressor			
$\theta^{Fed}_{BoE,t,k}$	$\theta^{Fed}_{ECB,t,k}$	$\theta^{Fed}_{BoE,t,k}\mathbb{I}_{\{t\geq 2005\}}$	$\theta^{Fed}_{ECB,t,k}\mathbb{I}_{\{t\geq 2005\}}$
0.020	0.050**	0.065**	0.001
(0.019)	(0.020)	(0.025)	(0.026)

Note: *p<0.1; **p<0.05; ***p<0.01