News Topics and Economic Fluctuations

Vegard Høghaug Larsen and Leif Anders Thorsrud

Norges Bank Conference on Big Data, Machine learning and the Macroeconomy
October 2–3, 2017

Introduction

The Unstructured Big Data Project

Using text as data:

A number is a fact, but the media in which it is presented/discussed/opinionated adds to the information

Rational (in)attention theory:

- Media an important channel for information diffusion

The fact and the media are both part of the (agent's) information set when forming expectations $\Rightarrow$ outcomes

But, putting text into models is somewhat new (to economists) - until recently...

Newspaper topics and the economy

Projects:

1. **News as a driver of the business cycle:**
   - The Value of News for Economic Developments, Larsen and Thorsrud (2015)
2. **A daily news based business cycle indicator:**
   - A newsy coincident index of business cycles, Thorsrud (2016a, 2016b)
3. **Uncertainty and the economy:**
   - Components of Uncertainty, Larsen (2017)
4. **Predicting daily stock market fluctuations using news:**
   - Asset returns, news topics, and media effects, Larsen and Thorsrud (2017)

Estimating news topics

Data: Newspaper articles

Data: Printed articles in Dagens Næringsliv, Norways biggest business newspaper and the forth largest irrespective of theme.

- Source: Retriever's "Atekst" database
- The data spans 1988–2017
- Close to 500 000 articles
- ≈ 2 GB of raw text

Data cleaning and preprocessing:

- Stopword removal (*and*, *but*, *where*)
- Stemming (*production* → *produc*)
- Term freq. - inverse document freq.

Machine Learning: Latent Dirichlet Allocation

- The newspaper is decomposed according to the topics it writes about using a topic model called Latent Dirichlet Allocation (LDA) introduced by Blei, Jordan, and Ng (JMLR, 2003).

- LDA as a generative model, where one article is created as follows:
  1. Pick the overall theme of an article by randomly giving it a distribution over topics
  2. For each word in the article
     i) From the topic distribution chosen in 1., randomly pick one topic
     ii) Given that topic, randomly choose a word from this topic

- Statistical techniques can be used to invert this process, inferring the set of topics that were responsible for generating a collection of articles, see e.g. Griffiths and Steyvers (PNAS, 2004)

- Application by economists: Hansen, McMahon and Prat (2014)

More on LDA

Estimating the topic model

- The LDA takes a set of articles as input and return two sets of distributions:
  - One set of distributions over words, one distribution for each topic j, given by φ_j
  - One set of distributions over topics, one distribution for each article i, given by θ_i
- The researcher must select the number of topics prior to estimation:

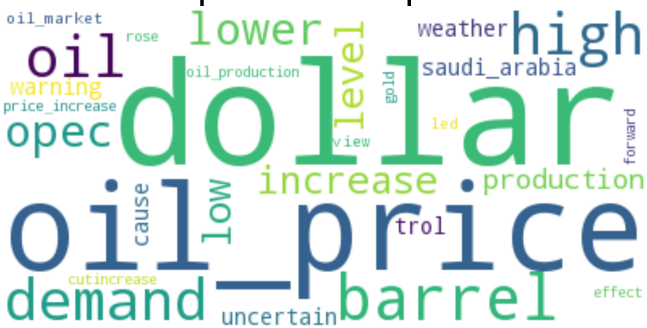
  We use 80 topics
- There is a trade off between interpretable topics and how well the topics are at explaining the whole newspaper, see Chang et al. (NIPS, 2009).
- Estimation is done using MCMC

Examples of the word distributions: φ_j

Topic 19: Macroeconomics

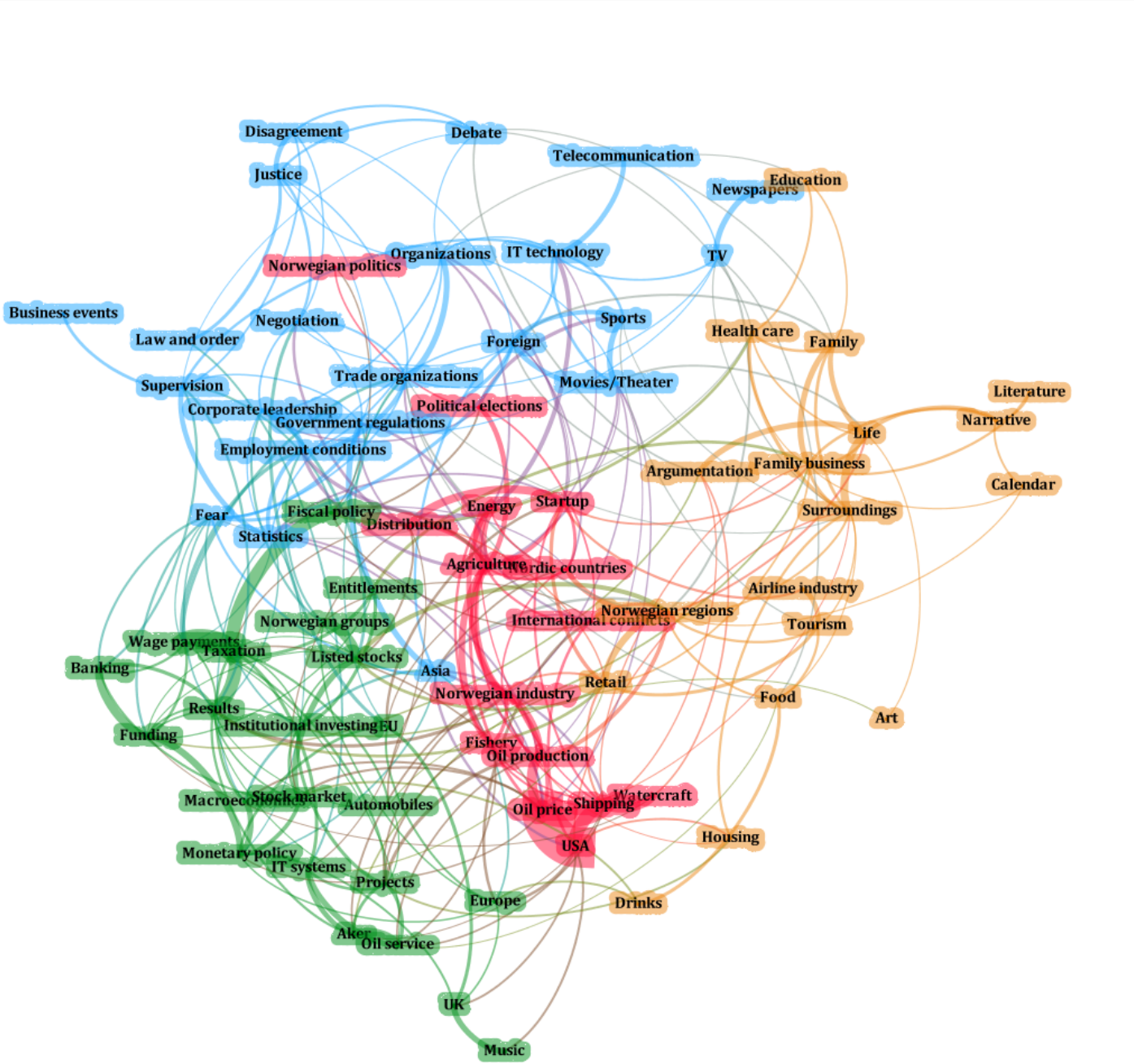
Topic 72: Monetary policy

Topic 33: Oil price

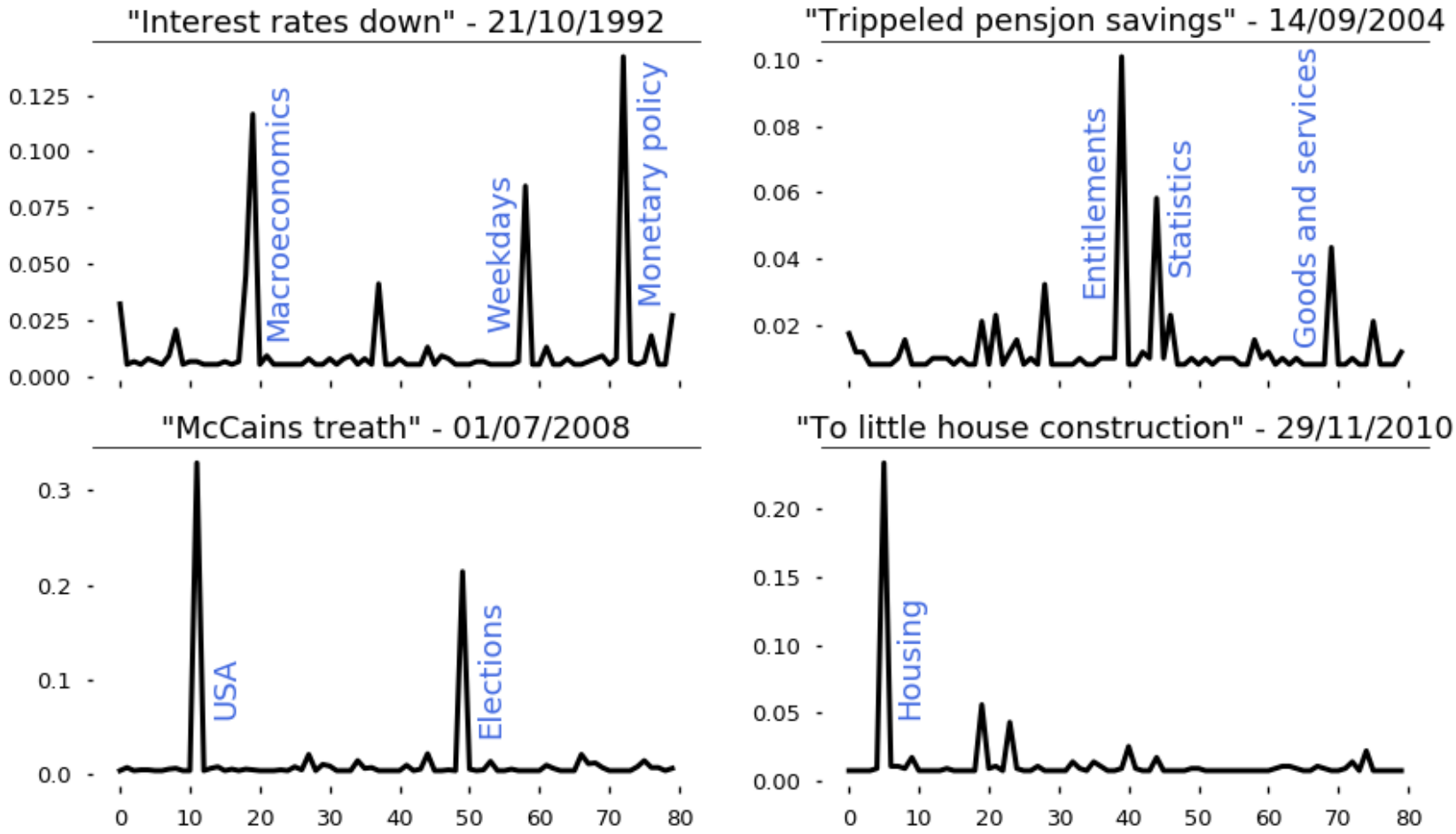


Topic 31: Shipping

Newspaper topics as a topic net



Examples of the topic distributions: θ_i

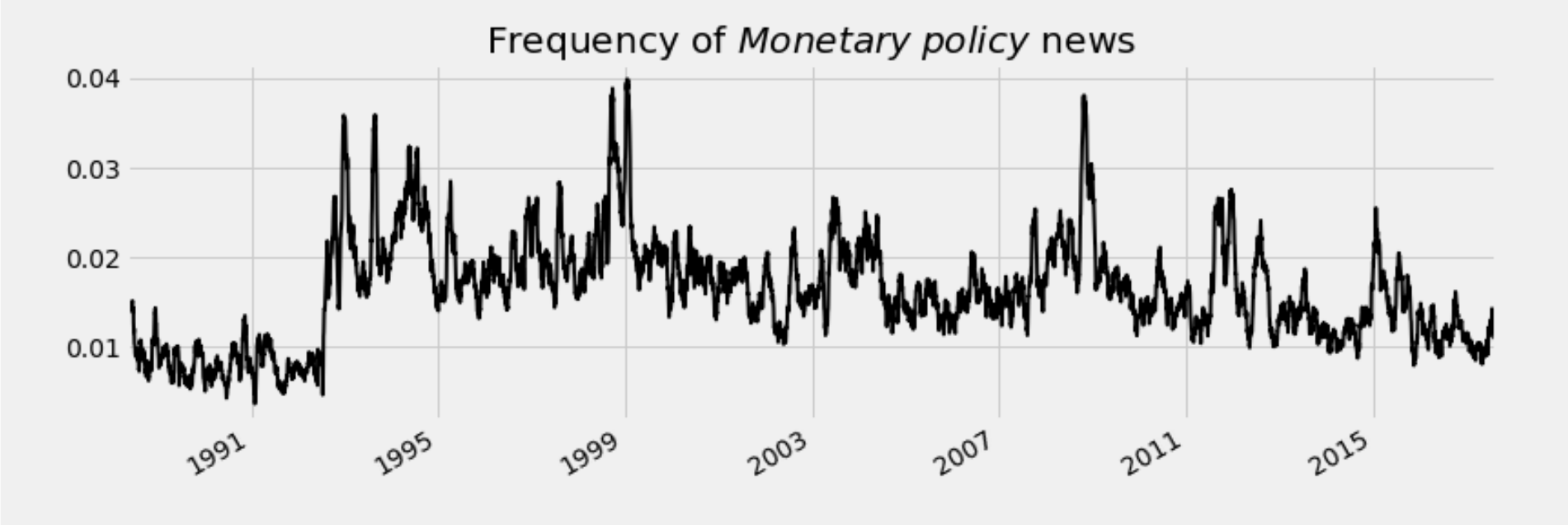


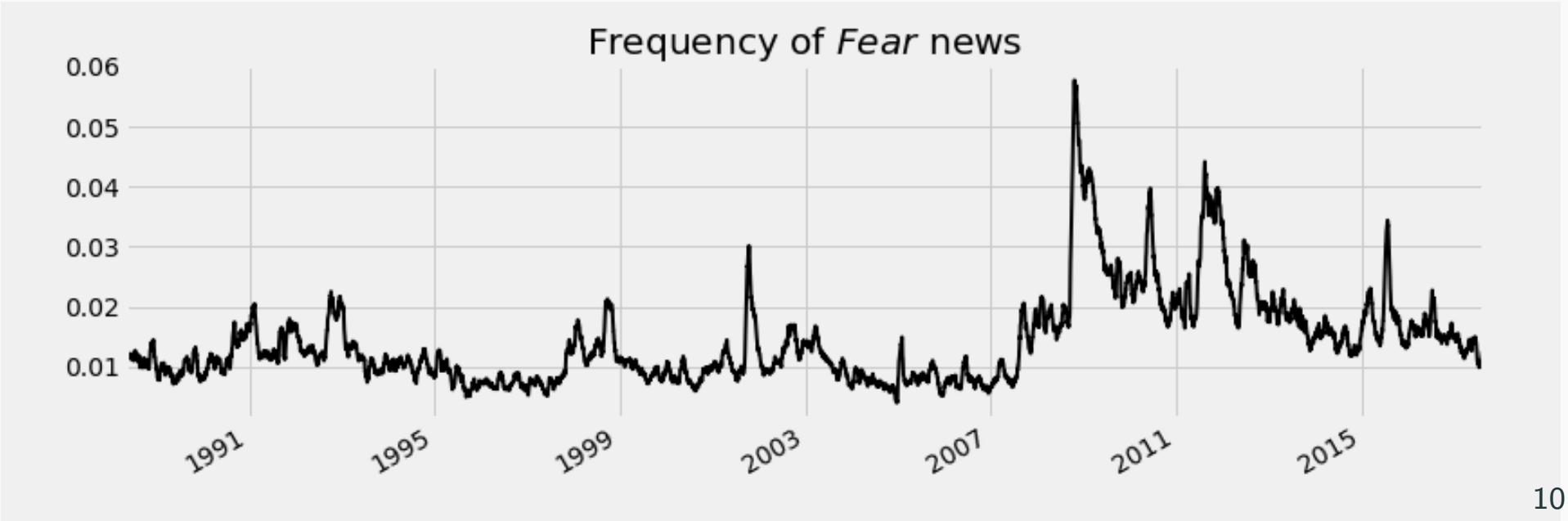
News as a driver of the business cycle

Key hypothesis:

The more a newspaper writes about a topic the more likely it is that this topic reflects something of importance for the economys current and future needs and developments.

Topic frequencies for Monetary policy and Fear news





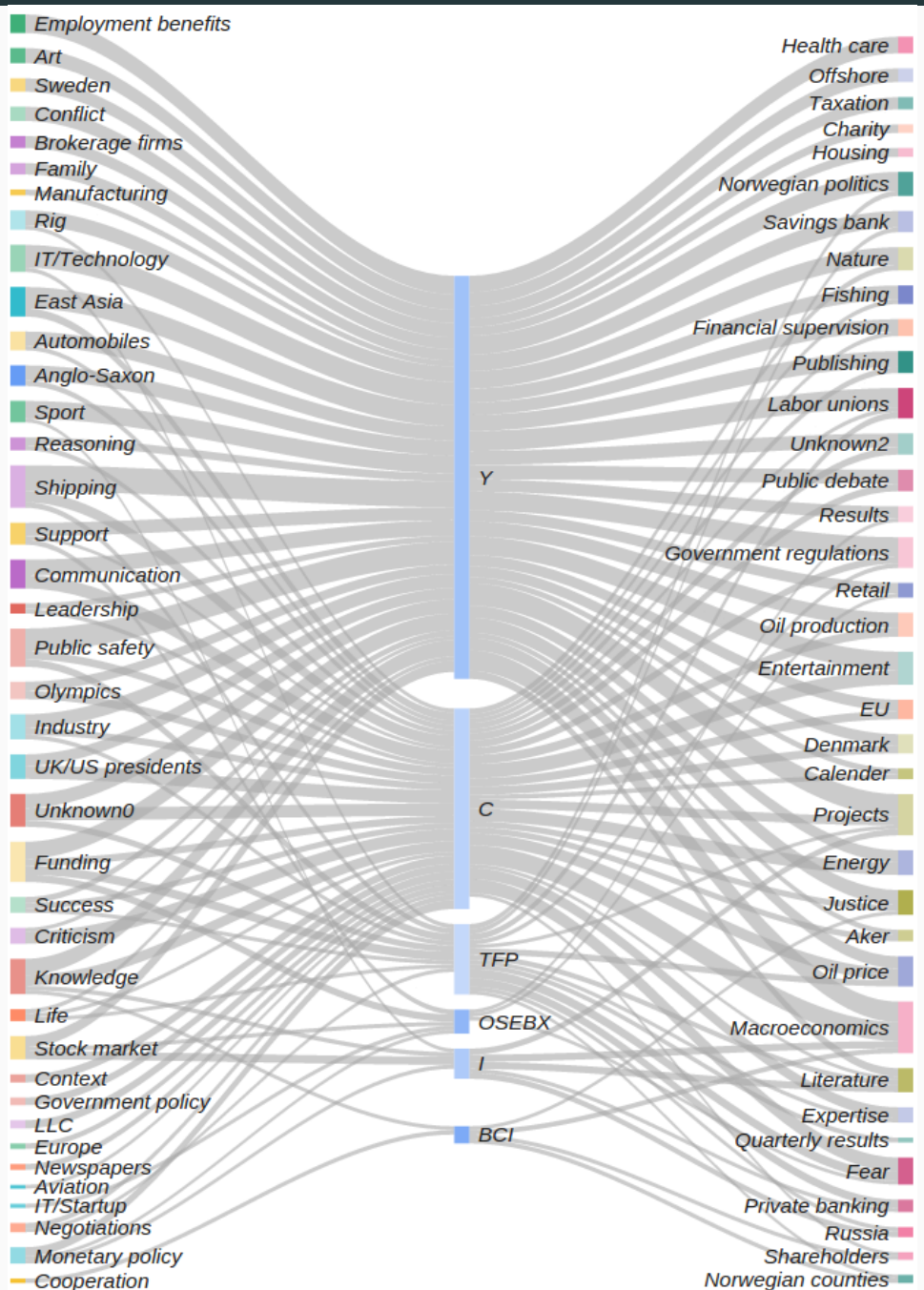
Topics adding marginal predictive power

Posterior odds ratios:

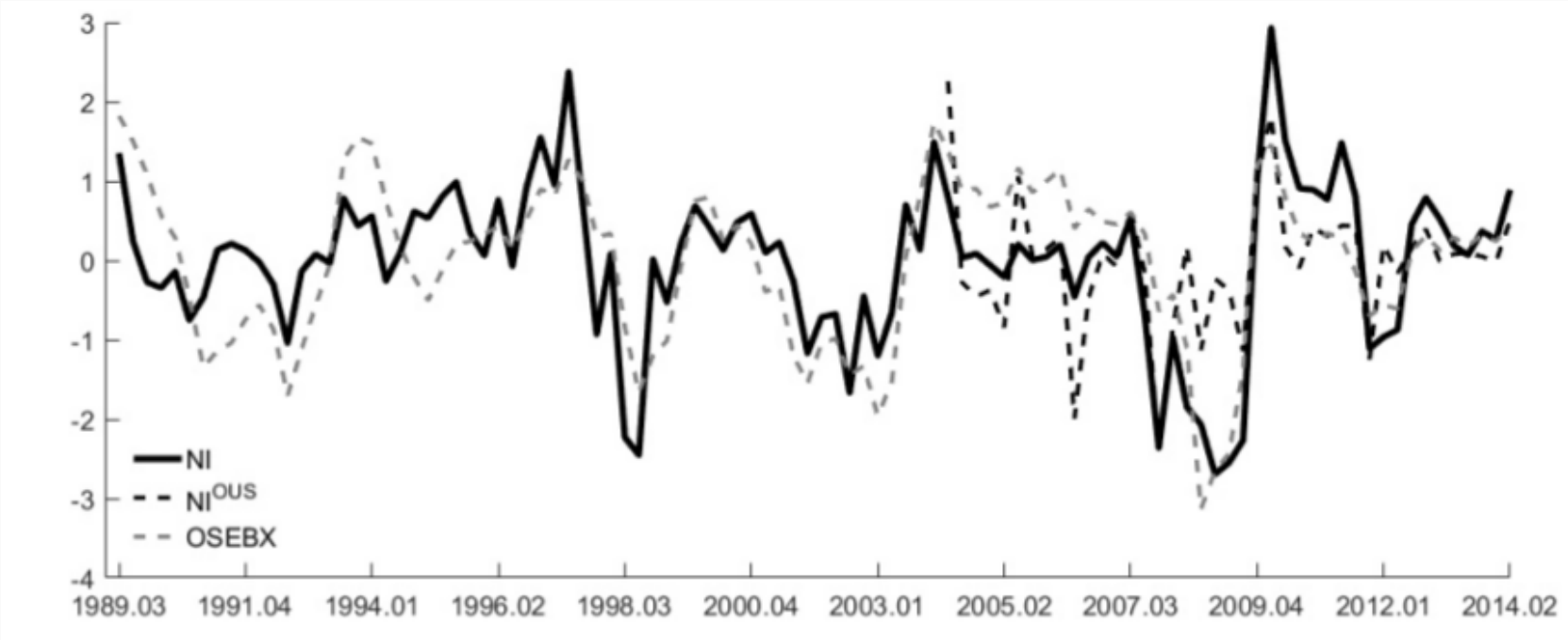
$$PO_{jt} = \frac{p(y|M_i)}{p(y|M_j)}$$

Plot report topics where:

$$2 \ln PO_{jt} > 2$$



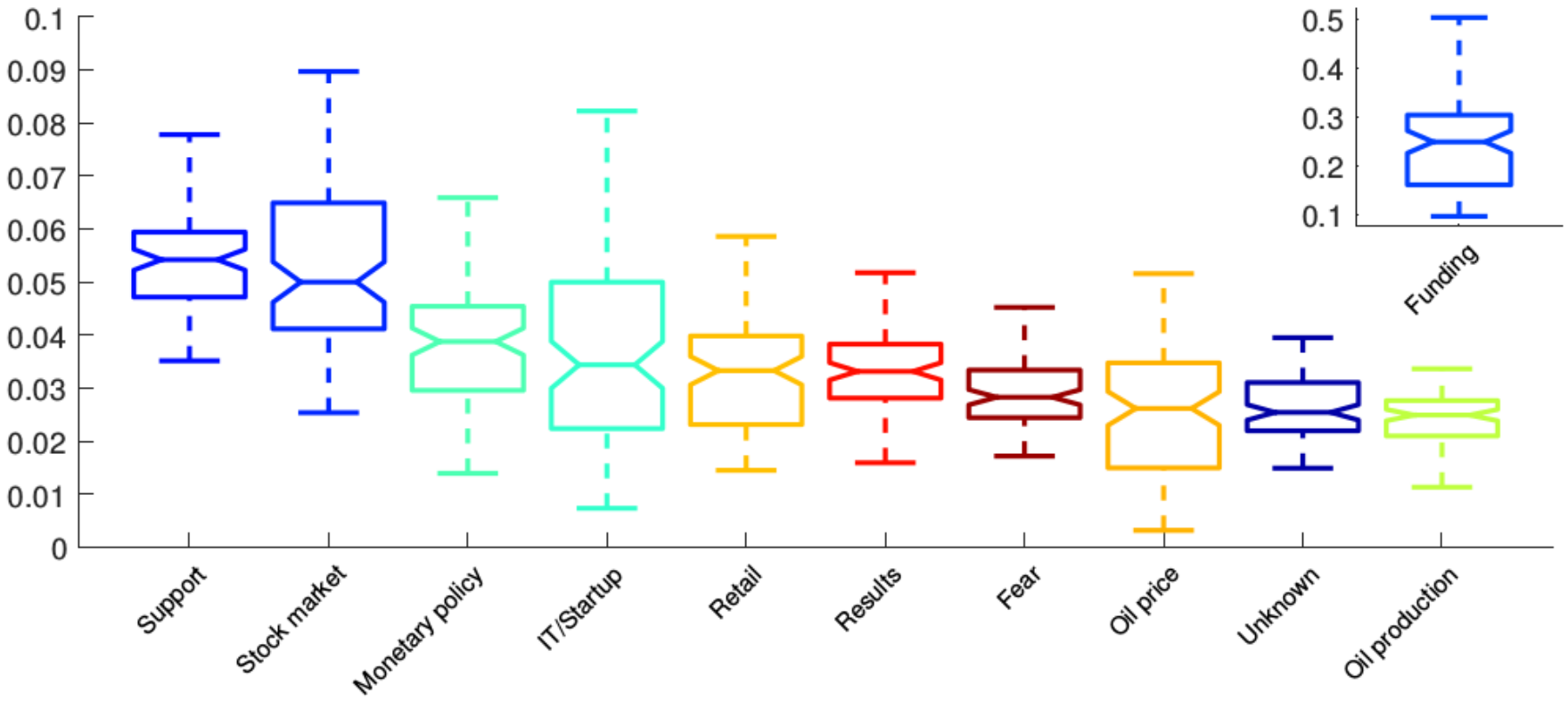
The News Index:



$$ni_t = \sum_{i=1}^{T} \omega_i b_{i,t} n_{i,t-1}, \quad \text{where} \quad \omega_t = \frac{p(y|M_i)}{\sum_j p(y|M_j)}$$

Latent Threshold Model | IRFs for news shocks

What constitutes a news shock?



A daily news based business cycle indicator

What is the state of the economy today?

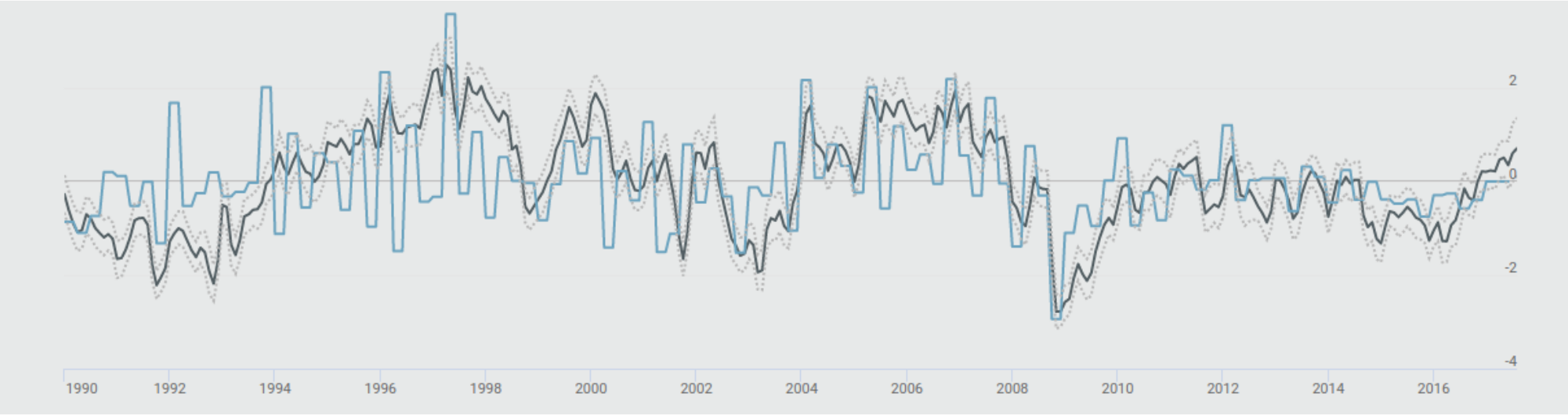
In real-time, our best known measure of economic activity, GDP growth, is not observed:

- It is registered on a quarterly frequency
- It is published with a considerable lag

One solution:

- A news-based coincident index
- Mixed-frequency time-varying Dynamic Factor Model
- Uses daily newspaper topics and quarterly GDP
- Enforces dynamic sparsity using a latent threshold mechanism

Financial News Index (FNI) for Norway



Published monthly at `www.retriever-info.com/fni/`

Decomposing the FNI

Uncertainty and the economy

Uncertainty term counts in topic specific news:

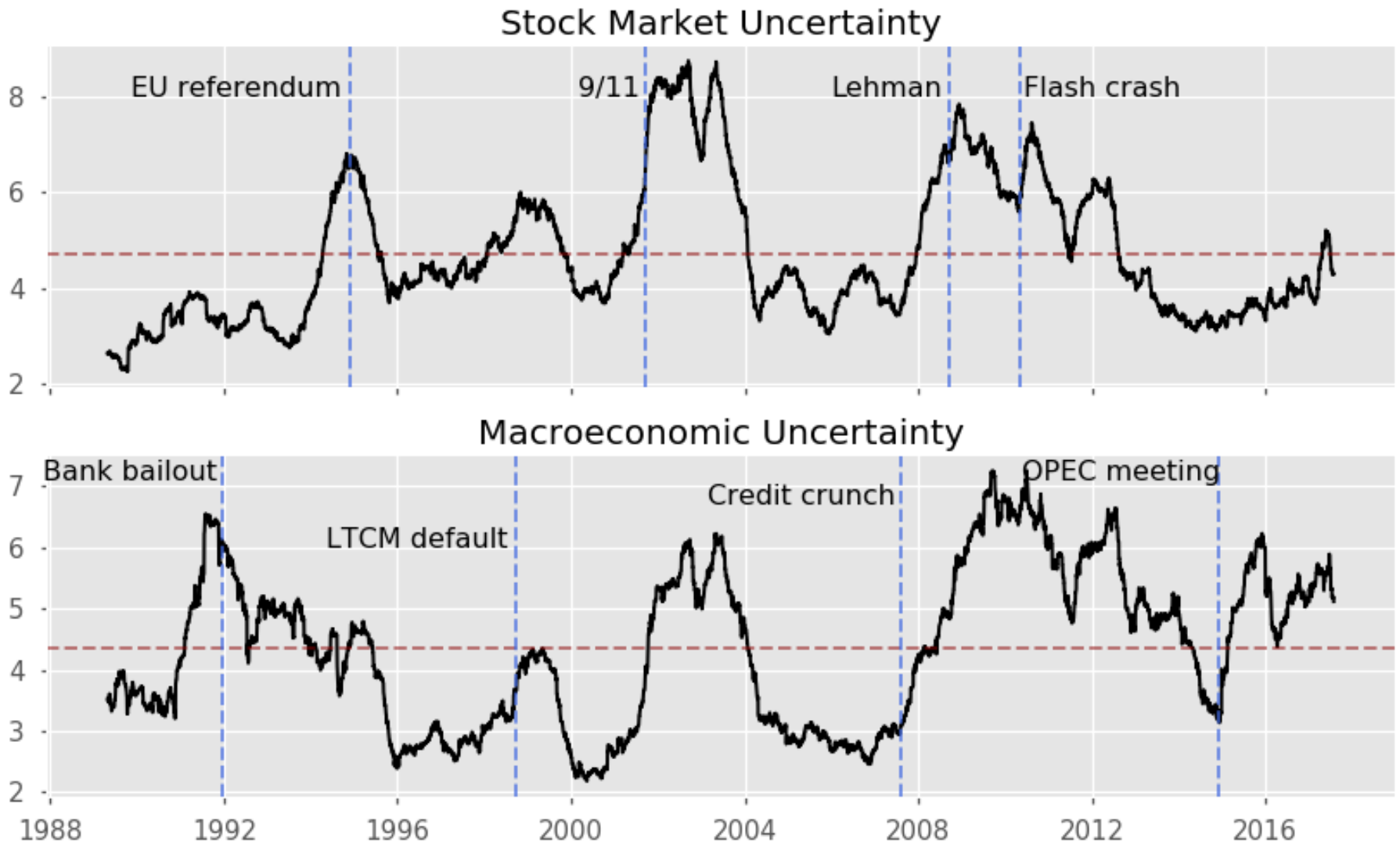
- Uncertainty is quantified on the article level by counting uncertainty terms.
- The words that are counted: *uncertain* and *uncertainty* (and also variations of these words)
- Let's define

$$\begin{aligned} \upsilon_i &\equiv \text{number of uncertainty terms in article } i \\ \omega_i &\equiv \text{number of total words in article } i \end{aligned}$$

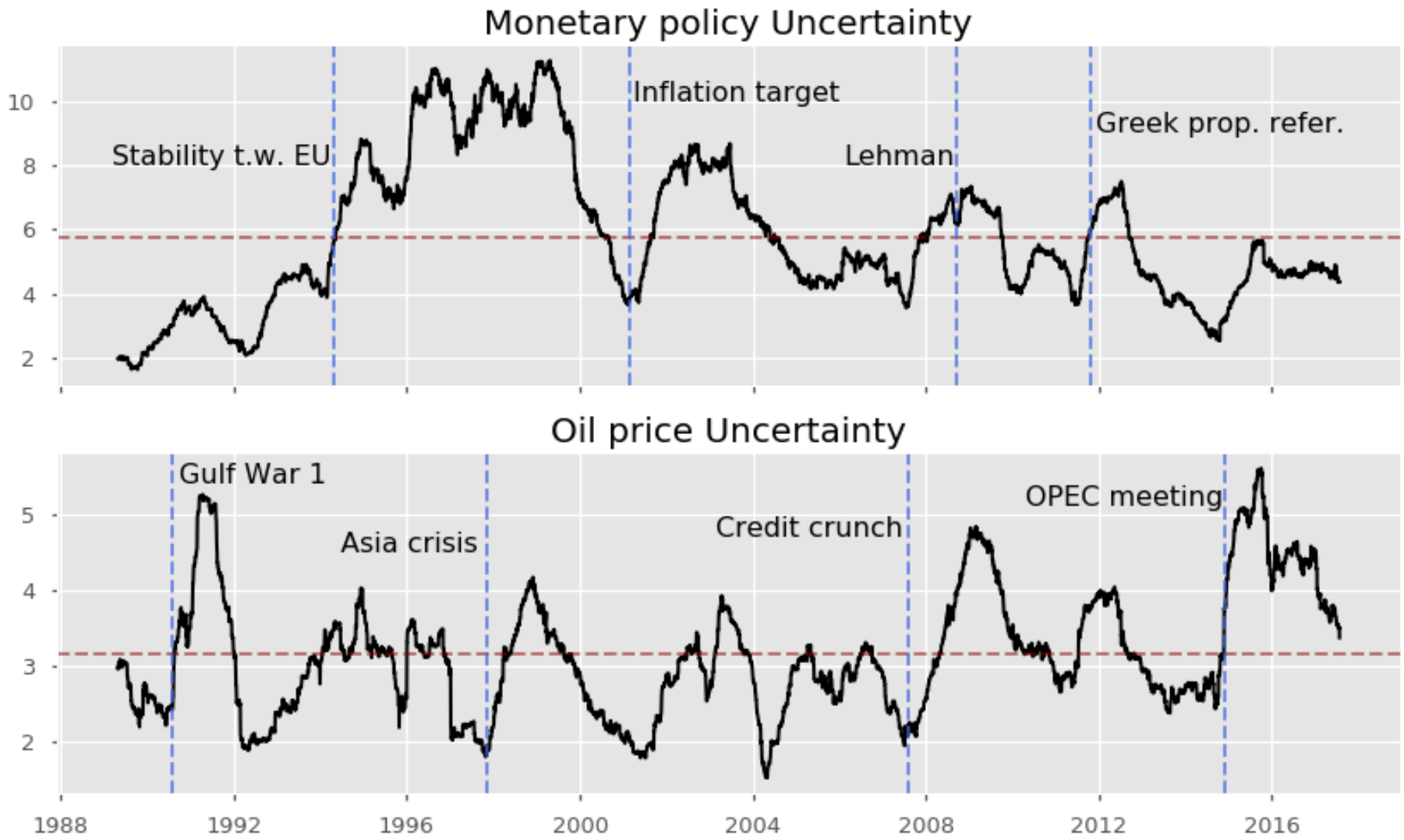
- The category specific uncertainty measures are calculated for all topics j:

$$\text{Topic } j \text{ uncertainty on day } t \;=\; \frac{\sum_{i \in \text{day } t} \upsilon_i \boldsymbol{\theta}_i(\text{topic } j)}{\sum_{i \in \text{day } t} \omega_i}$$

Category/topic specific uncertainty

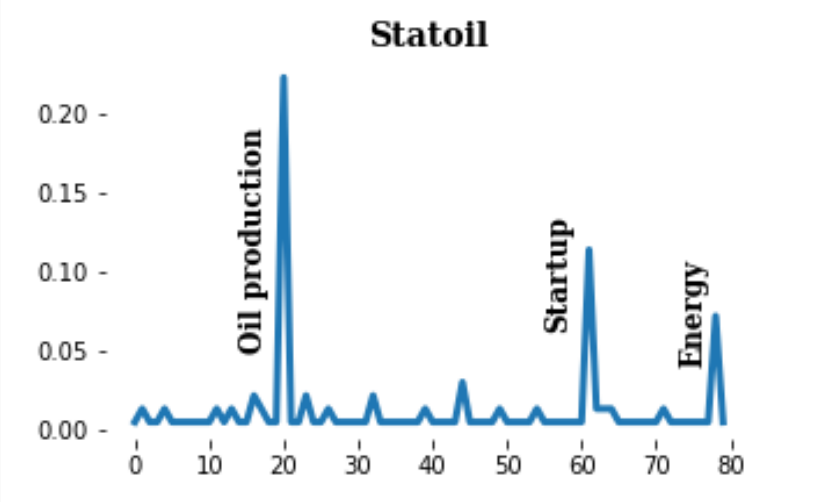


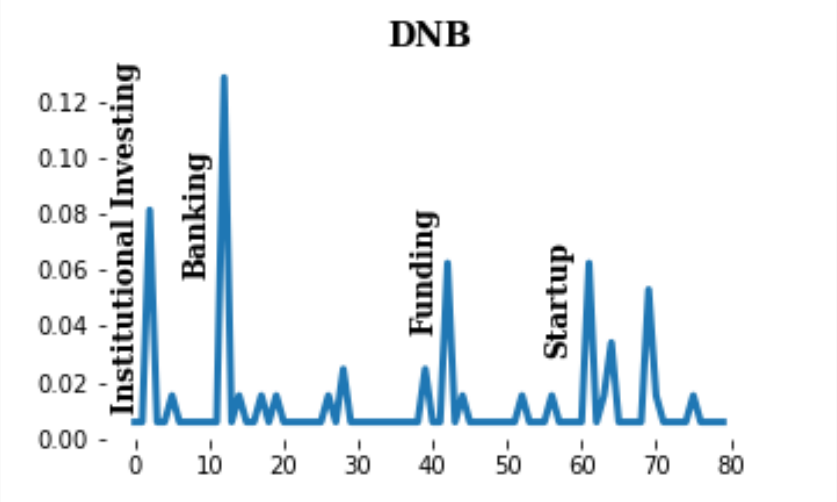
Category/topic specific uncertainty

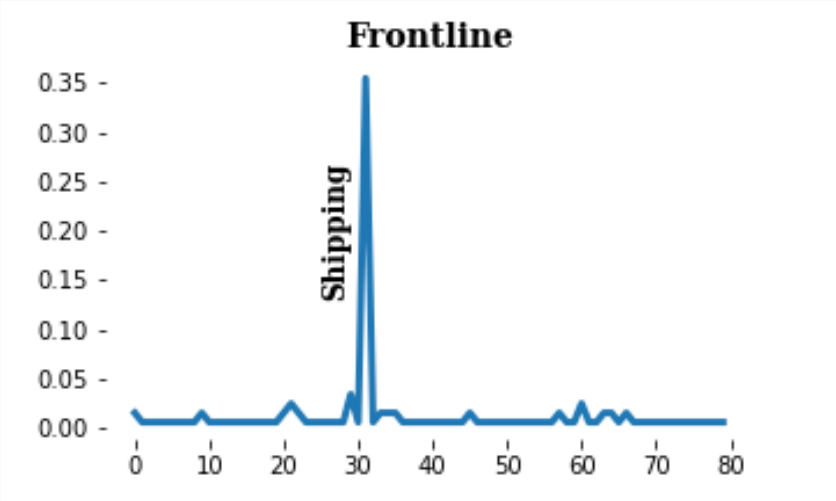


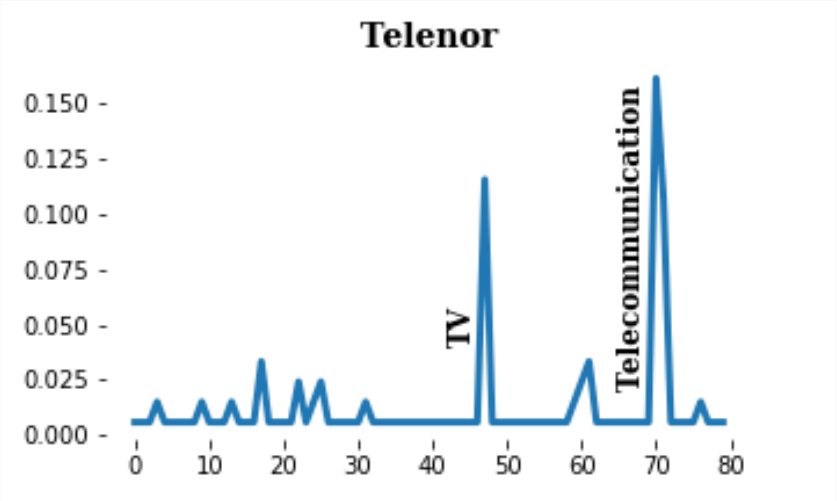
News and asset prices

Linking stocks to news topics:









News and returns

	c2o			c2c			
	I	**II**	**III**	**I**	**II**	**III**	**Std(X)**
$Topic_t$	0.0107***	0.0135***	0.0090**	0.0153***	0.0272***	0.0208***	0.017
	(0.0022)	(0.0031)	(0.0037)	(0.0031)	(0.0051)	(0.0062)	
R^{mi}_{t-1}		0.3453***	0.3410***		0.2816***	0.2774***	0.013
		(0.0177)	(0.0217)		(0.0285)	(0.0371)	
R^{mh}_{t-1}		-0.0120	-0.0180		-0.0012	-0.0023	0.016
		(0.0133)	(0.0158)		(0.0222)	(0.0275)	
R^{oil}_{t-1}		0.0139**	0.0067		0.0231*	0.0132	0.020
		(0.0067)	(0.0091)		(0.0120)	(0.0169)	
B/M_{t-1}		-0.0007***	-0.0007***		-0.0008***	-0.0007***	0.889
		(0.0001)	(0.0001)		(0.0002)	(0.0002)	
MV_{t-1}		0.0002*	0.0002*		-0.0003*	-0.0003*	1.802
		(0.0001)	(0.0001)		(0.0001)	(0.0002)	
$Turn_{t-1}$			0.0115***			0.0076***	0.047
			(0.0021)			(0.0022)	
R^2	0.0088	0.0298	0.0301	0.0054	0.0128	0.0128	
Obs.	540708	540708	391533	540708	540708	391533	
α_i	yes	yes	yes	yes	yes	yes	
δ_t	yes	no	no	yes	no	no	

Summing up

Conclusion

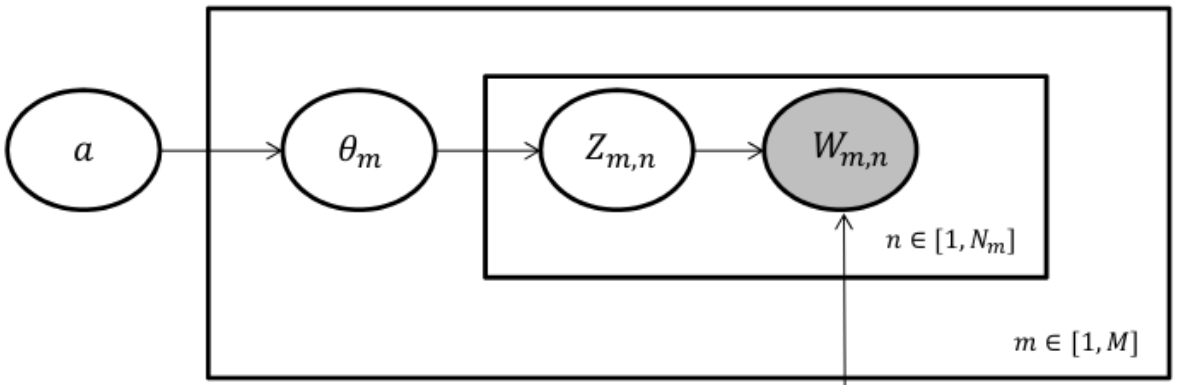
- Alternative data sources, e.g., newspaper data, provide valuable information
- We obtain coherent results across a variety of applications - at both daily and quarterly frequency - for both forecasting and structural analysis

More on LDA

Joint likelihood:

$$p(\varphi, \theta, \mathbf{z}, \mathbf{w}) = \left(\prod_{\mathbf{j}}^{\mathbf{K}} \mathbf{p}(\varphi_{\mathbf{j}}|\beta) \right) \left(\prod_{\mathbf{i}}^{\mathbf{M}} \mathbf{p}(\theta_{\mathbf{i}}|\alpha) \prod_{\mathbf{n}=\mathbf{1}}^{\mathbf{N}} (\mathbf{z}_{\mathbf{i},\mathbf{n}}|\theta_{\mathbf{i}}) \mathbf{p}(\mathbf{w}_{\mathbf{i},\mathbf{n}}|\varphi_{\mathbf{1:K}}, \mathbf{z}_{\mathbf{i},\mathbf{n}}) \right)$$

Graphical model:



Back

Latent Threshold Model:

We estimate both the AR(p) or ARX(p) specification using a Latent Threshold Model (LTM).

The LTM was introduced by Nakajima and West (JEBS, 2013), and can be written as follows:

$$y_t = x_{t-1}' b_t + u_t \qquad u_t \sim N(0, \sigma_u^2) \tag{1a}$$

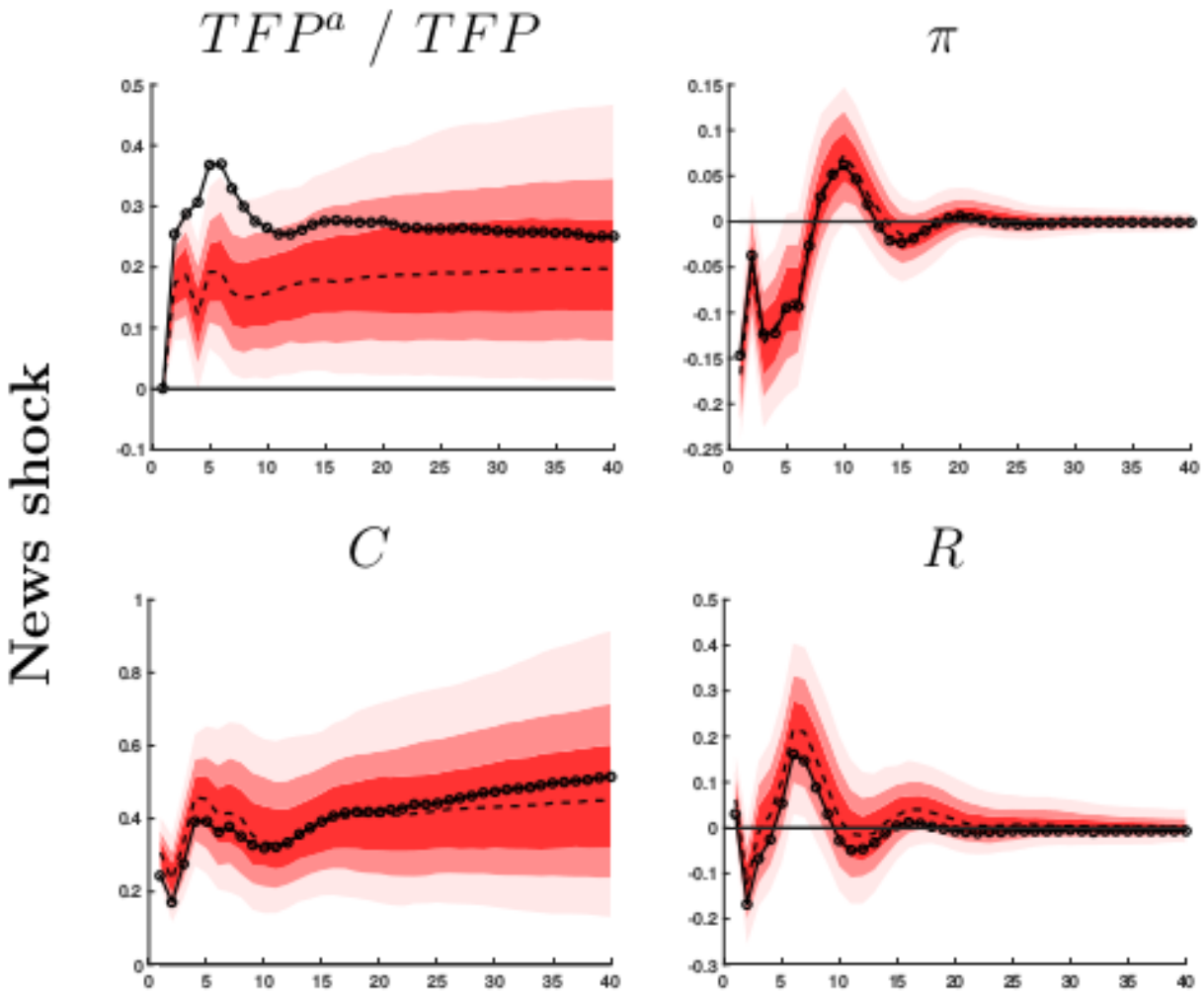
$$b_t = \beta_t \varsigma_t \qquad \varsigma_t = I(|\beta_t| \geq d) \tag{1b}$$

$$\beta_t = \beta_{t-1} + e_t \qquad e_t \sim N(0, \Sigma_e) \tag{1c}$$

Back

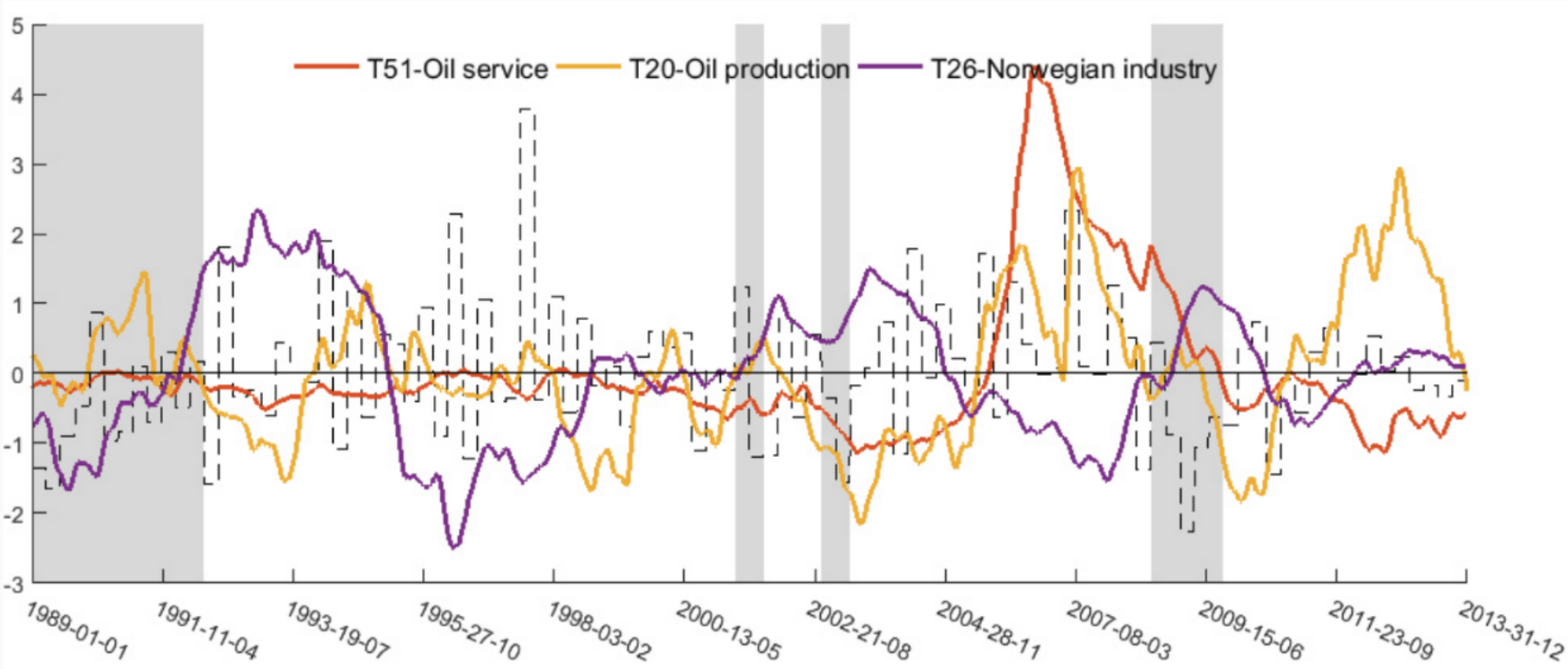
IRFs from a News shock

We estimate an S-VAR as in Beaudry and Portier (AER, 2006):



Back

Decomposing the FNI into topic contributions (some examples)



Back