14D010 Text Mining for Social Sciences

Project: Analyzing ECB Press Statements

Hans-Peter Höllwirth

June 19, 2017

Abstract

This report describes text analysis conducted on official press conference statements by the President of the European Central Bank (ECB). In particular, I use different Latent Dirichlet Allocation models on the bag-of-word representation of the statement text to find to what extend language and focus of the ECB changed over time. I observe that for small topic numbers change in language can be mainly attributed to presidencies. For larger topic numbers, however, topics can be directly associated to major events and policy directions in the history of the ECB.

1 Data

The European Central Bank (ECB) manages the monetary policy in the Eurozone of the European Union (EU). Since its establishment in June 1998, its acting President who heads the executive board, governing council and general council of the ECB holds regular press conferences in which the President provides an update on the ECB's monetary policy.

Press conferences are usually scheduled monthly (12 meetings per year). Since 2015, however, press conferences are held roughly every 45 days (thus, 8 conferences per year). This gives a total of **222 press conferences** since June 1998 at the time of working on

this project. Each press conference starts with an introductory statement by the President, followed by a Q&A session between press representatives and the President. I decided to filter the press questions and only keep the President's responses. Thus, I solely focus on the words used by the president and disregard any other sources. The main purpose of this filtering is to ensure I do not consider political topics the press might have asked the president about but who was unwilling to comment on it.

1.1 Web Scraping

The transcript of each regular press conference from 1998 to 2017 is published on the ECB's webpage under the link https://www.ecb.europa.eu/press/pressconf/. The webpage directs to to the press conference list for each year.

The web scraping procedure iterates over the list of press statements for every year from 1998 to 2017. From each year page, the procedure extracts the date of speech (stored in <dt> tags) and the link to the English transcript (stored in tags of class 'doc-title'). For every press conference link, I consider the full transcript (encapsulated in tag <article>) but filter questions (as explained before), the head of every answer (which starts by the name of the president, eg. 'Draghi: '), and remove the first paragraph from the transcript which only contains links or metadata. Finally, I concatenate all remaining paragraphs to one transcript string.

1.2 Metadata

I annotate each of the 222 press conferences with the date of speech and the name of the ECB's president at that time. Since its establishment in June 1998, the ECB has been headed by 3 presidents, shown in *Table 1*.

Took office	Left office	Name	State
01.06.1998	31.10.2003	Wim Duisenberg	Netherlands
01.11.2003	31.10.2011	Jean-Claude Trichet	France
01.11.2011	incumbent	Mario Draghi	Italy

Table 1: List of ECB presidents

The annotated transcripts are then stored in a single comma-separated file: combined.csv. The size of the file is 5.8MB.

2 The Question

I use the European Central Bank press conference transcripts as a proxy to study whether and how the focus of the ECB shifted over time. In particular, I am interested to see whether

major economic events in the past 20 years are reflected in the content of the speeches. Such events include the introduction of the Euro currency in 2002, the international financial crisis from 2007, and the European debt crisis from late 2009. Once established, one could add new press conferences later and test whether they cover topics from the recent past, or whether they constitute new policies.

3 Data Pre-processing

In a first step, I pre-process the raw text data obtained from web scraping. The text pre-processing involves several sequential steps:

1. **Tokenize**: First, I transform the strings of speeches into arrays of tokens.
2. **Case-folding**: Next, I convert all tokens into lower-case representation.
3. **Alphanumeric**: I then remove all non-alphanumeric characters from every token, using regular expressions.
4. **Stopwords**: Next, I remove all English stopwords from the token list, using the NLTK corpus English stopword list.
5. **Stemming**: Finally, I replace the tokens with their stem, using the Porter stemmer.

3.1 Data-driven Stopwords

In a last step, I remove data-driven stopwords. For this task I form a vocabulary list and then compute the tf-idf score for every vocabulary term. From these scores, I finally compute the topic stopword list, using a threshold value of 1.0. All terms with a tf-idf score below this threshold are put on that list. This results in the following topic stopwords to be removed from the transcript representations:

```
   ['activ', 'alreadi', 'also', 'analysi', 'annual', 'area', 'assess', 'bank',
'base', 'basi', 'chang', 'close', 'come', 'concern', 'condit', 'confid',
'confirm', 'continu', 'contribut', 'could', 'council', 'countri', 'cours',
'current', 'data', 'decid', 'decis', 'develop', 'dispos', 'ecb', 'econom',
'economi', 'effect', 'euro', 'european', 'expect', 'explain', 'financi',
'first', 'fiscal', 'follow', 'futur', 'gentlemen', 'govern', 'growth', 'hicp',
'high', 'howev', 'import', 'includ', 'increas', 'indic', 'inflat', 'inform',
'interest', 'ladi', 'last', 'let', 'level', 'like', 'line', 'look', 'made',
'maintain', 'make', 'market', 'medium', 'meet', 'monetari', 'month', 'much',
'need', 'one', 'order', 'outcom', 'outlook', 'particular', 'past', 'period',
'point', 'polici', 'posit', 'present', 'press', 'price', 'privat', 'public',
'question', 'rate', 'real', 'recent', 'reflect', 'reform', 'regard', 'relat',
'remain', 'report', 'risk', 'said', 'say', 'second', 'sector', 'see', 'sinc',
```

```
'stabil', 'start', 'still', 'structur', 'support', 'take', 'term', 'think',
'time', 'today', 'turn', 'two', 'vicepresid', 'well', 'would', 'year']
```

I end up with a cleaned token representation of the press conference speeches, i.e. each speech is represented by a list of tokens.

4 Content Analysis

This section describes the particular methods I used for the content analysis of the (pre-processed) ECB press conference speeches. As explained earlier, I intended to use the transcripts as a proxy to explain ECB's policy shifts over time and how those changes were related to major economic events. For a start, I used plain **Latent Dirichlet Allocation (LDA)**. The key idea behind my method choice is that I assume that at each press conference the president discusses of a range of different policies/topics, each of them characterized by different key words. This makes LDA the perfect method choice since it gives an estimate of the latent structure of the press conferences and allows for topic mixtures in each press conference.

Based on my early results obtained with the plain LDA method, I then used the **Structural Topic Model (STM)** for the later parts of the content analysis. The structural topic model allowed me to quantify the effect of presidents on the topic distribution observed with low topic numbers.

4.1 Latent Dirichlet Allocation (LDA)

My choice of implementation was Pyhton's *lda* package which implements LDA using collapsed Gibbs sampling. I start by constructing the document-term matrix from the cleaned token representation of the press conference speeches obtained from the pre-processing step. The obtained document-term matrix serves as input to the LDA implementation. Next, I train the model with $K = 3$ topics over 1000 sampling iterations. I tested the method for several prior hyper-parameter combinations and ended up using alpha=0.1 (Dirichlet parameter for distribution over topics) and eta=0.1 (Dirichlet parameter for distribution over words).

The topic distribution plot in *Figure 1* reveals two clear structures. First, each topic seems to be strongly connected to a particular president. Topic 2 dominates the first 6 years and abruptly gets replaced by topic 1 as the dominant topic at the end of 2004 which marks the transition from Wim Duisenberg to Jean-Claude Trichet. Another rather abrupt switch - this time from topic 1 to topic 3 - happens at the transition from Jean-Claude Trichet to Mario Draghi at the end of 2011. This observation suggests that the topic distribution is strongly associated to individual speech patterns, but at least to some part might also

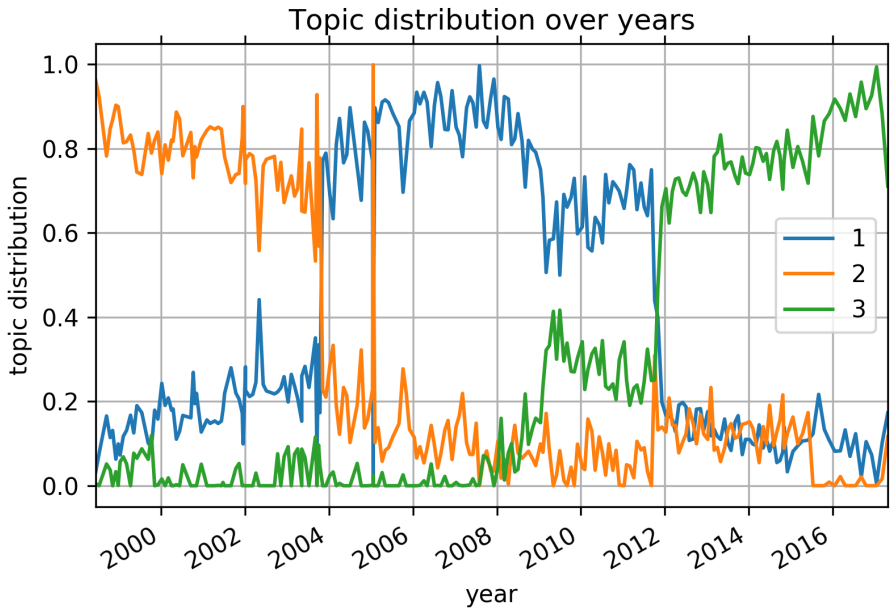


Figure 1: LDA topics distribution with $K = 3$

represent a shift of policy introduced by the new president.

The last idea is supported by the second structure we can observe in the topic distribution plot: While practically irrelevant in the first 10 years, topic 3 quickly becomes relevant in 2008 and more so at the start of 2009 with a share of 20-40% - almost 3 years before Mario Draghi took office. The time of its emergence coincides with two major (linked) events: In 2008 the **global financial crisis** started to hit the Eurozone which then in late 2009 triggered the start of the **European debt crisis**. One could interpret topic 3 as European debt crisis and claim that the focus on this crisis was intensified with the introduction of Mario Draghi. An alternative hypothesis is that topic 3 combines language used by Mario Draghi and financial/debt crisis language.

Figure 2 shows the word clouds for the 3 topics. Topics 1 and 2, associated with Jean-Claude Trichet and Wim Duisenberg's presidency, mostly contain rather general terms that one would expect to be touched on in any central bank policy discussion. Some prominent terms in topic 1 could be linked to the global financial crisis (eg. US, global). In constrast, more specific terms dominate topic 3. Clearly, this topic focuses on policies dealing with the aftermath of the global financial crisis and the European debt crisis.

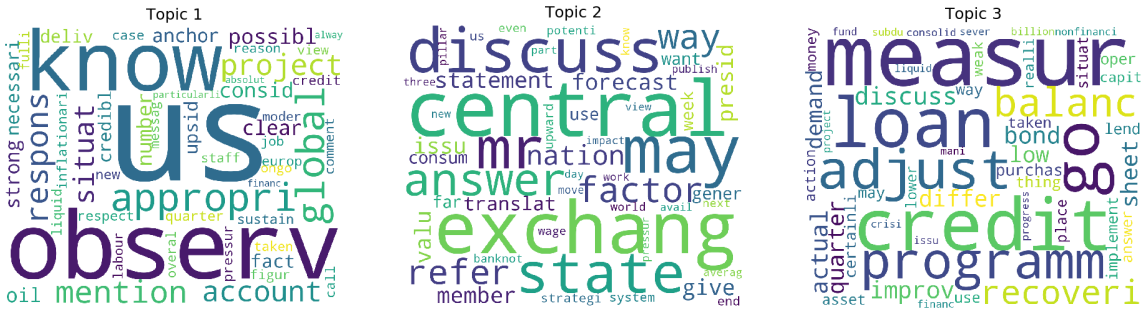


Figure 2: Word clouds for $K = 3$

To better understand the effect of individual speech patterns on topic grouping, I also ran the same LDA method for $K = 2$ and 4 topics. The resulting topic distribution plots are reported in *Figure 3*. When only allowing for 2 topics, topics 2 and 3 associated with the presidency of Wim Duisenberg and Mario Draghi from the 3-topics model are combined.

In the case of $K = 4$, we find that topic 3 from the 3-topics model (which has been associated to Mario Draghi's presidency) is now split into 2 topics: One topic (topic 3) emerges roughly at the start of the European debt crisis (but more pronounced than in the 3-topics model). The topic then loses relevance in the following years but suddenly re-emerges again in 2015 from which point on it dominates most of the speeches. The time of its re-emergence coincides with the launch of the ECB's **quantitative easing** program. Thus, topic 3 in the 4-topics model seems to combine coverage of the financial and European debt crisis as well as ECB's recovery policy. In contrast, the other topic (topic 2) related to Mario Draghi's presidency can be interpreted as Draghi's individual speech patterns and policy focus, which are not related to the crisis management captured by topic 3.

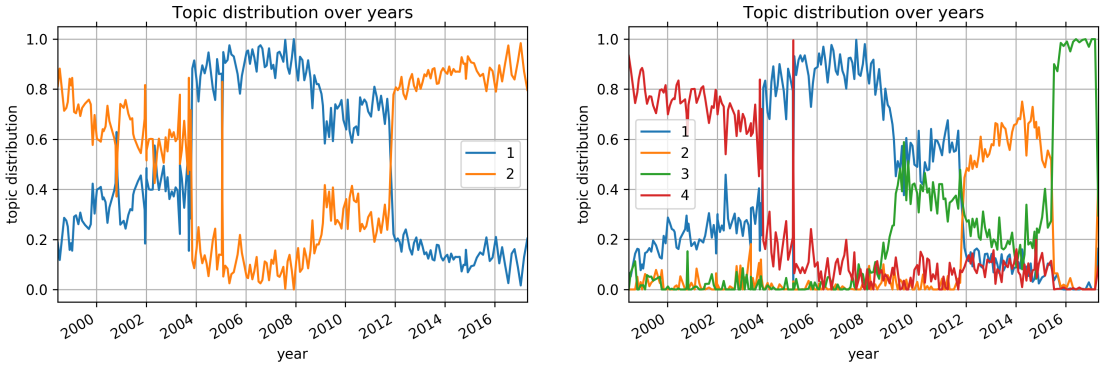


Figure 3: LDA topics distribution with $K = 2$ and $K = 4$

4.2 Structural Topic Model (STM)

The plain LDA results with few topics (K between 2 and 4) suggest that the president has a strong effect on the topic distribution in a speech. An extension to LDA - the Structural Topic Model - allows us to quantify this effect.

The *stm* package in R provides a powerful tool to conduct the structural topic analysis on the ECB press conference transcripts. The package takes the annotated transcripts from file *combined.csv* as input. It then provides a function called *textProcessor()* that performs all major text pre-processing steps that I described earlier combined: case-folding, removal of non-alphanumeric characters and language-specific stopwords, and stemming. In addition, the function also offers the option to remove additional stopwords. For this, I provide the list of data-driven stopwords which I listed earlier. A similar result could have been achieved with function *prepDocuments()* by using the *upper.tresh* option, however, I wanted to guarantee that exactly the same data-driven stopwords get removed. Thus, I use function *prepDocuments()* solely to obtain (the sparse representation of) the document-term matrix.

Next, I estimate the estimate the structural topic model with function *stm()* for $K = 3$ on the generated document-term matrix. I incorporate the president information in the form of one-hot-encoding (for reasons that are explained shortly) as contextual covariates in the prior distribution for the topic prevalence. The topic prevalence covariates formula looks as follows:

$$prevalence =\sim Duisenberg + Trichet + Draghi$$

Finally, the true workhorse function of the package - *estimateEffect()* - estimates the relationship between metadata and topics. In my case, I wish to quantify the effect of the president on the 3 topics from *Figure 1*. I run the effect estimation function on the constructed metadata for 10 simulation iterations and from this compute the median coefficients. The one-hot-encoding of the president metadata makes the result easier to interpret. The result is shown in *Table 2*.

	Intercept	Duisenberg	Trichet	Draghi
Topic 1	0.198	-0.197	**0.440**	-0.045
Topic 2	0.274	**0.540**	0.000	-0.267
Topic 3	0.278	-0.093	-0.189	**0.560**

Table 2: Effect of presidents on topics

The coefficients suggest that each topic is indeed connected to one particular president. In each case, a different president has a coefficient of around 0.5 while the other presidents are given negative coefficients which roughly cancel out the positive intercept value.

5 Results with More Topics

Having used mixed-membership models for small numbers of topics K, the question arises if we could get a more granulated description of ECB policies over time by increasing the number of topics. The earlier test with $K = 4$ has shown that additional topics might explain major phases in the Eurozone history that go beyond the association to a particular president.

I decided to adapt the structural topic model from before for this purpose which incorporates the president information as contextual covariates in the prior distribution for the topic prevalence. I started to explore the topic distribution of models with up to 25 different topics. I then manually decreased the topic number until the share of every topic exceeded 20% for at least 3 speeches. The highest number that met this condition was **10 topics**. *Figure 4* shows its topics distribution plot.

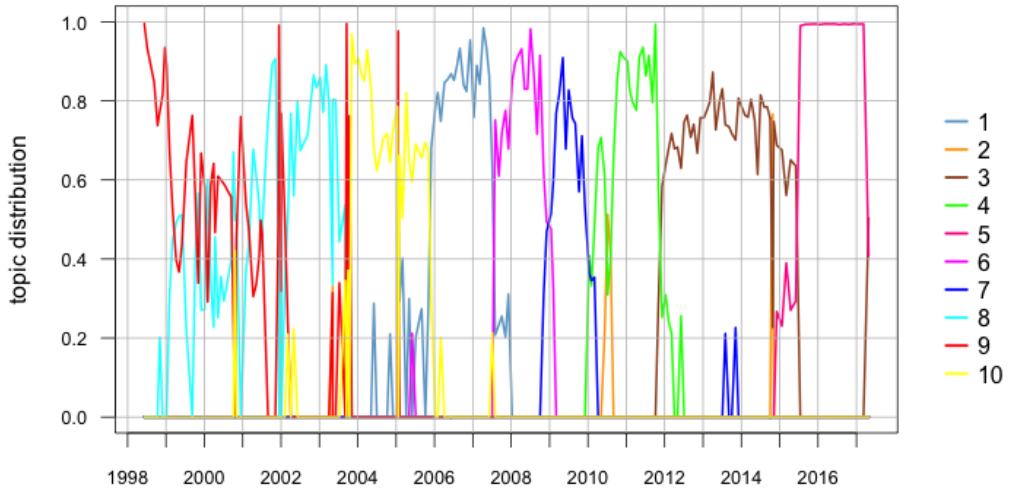


Figure 4: STM topics distribution with $K = 10$ (shares below 0.2 cut)

I then analyze the topics in a number of ways. First, I plot the topic shares over time to assess the period of relevance for each topic. All of those topic plots can be found in the appendix. Next, I compute word sets that describe each topic and put them into *word frequency tables*. The function *labelTopics()* outputs are variety of metrics:

- **HProb**: highest probability words
- **FREX**: words that are both frequent and exclusive
- **Lift**: words weighted by dividing their frequency by their frequency in other topics

- **Score**: words weighted by dividing their log frequency by their log frequency in other topics

I report the top 7 words for all 4 metrics. A sample word frequency table - in this case for topic 4 - is shown below in *Table 3*. The complete set of topic word frequency tables are added as an appendix.

Topic 4	1	2	3	4	5	6	7
HProb	respons	appropri	programm	anchor	necessari	consid	deliv
FREX	efsf	non-standard	ireland	tension	allot	greec	head
Lift	africa	dormant	double-dip	flaw	geithner	grade	hungari
Score	non-standard	efsf	doctrin	smp	recapitalis	transmiss	default

Table 3: Word frequency table for topic 4

The word sets turn out to be most powerful for identifying latent policies/events when considered together. Note, that the highest probability word set is in most cases the least informative metric because it mostly consists of rather general economic terms. This is even true after removing all terms that appear in more than $185/222 \approx 83\%$ of all speeches with function *prepDocuments()*, option *upper.thresh*. As a consequence, simple word frequency clouds also provide little information and are therefore omitted from this report.

5.1 Topics Interpretation

Next, I tried to give an ad-hoc interpretation to every topic, based mainly on the period in which a topic was relevant, and the word frequency tables (find both in the appendix). The result is shown in *Table 4*.

Topic	Interpretaton	Period	Key terms
1	Early Trichit presidency	2004-2008	-
2	Banking system stability	-	stress, test, billion, shortfal
3	European debt crisis 2	2012-1015	fragment, omt, esm
4	European debt crisis 1	2010-2013	efsf, ireland, greece, non-standard
5	Quantitative easing	2015-	recoveri, purchas, tltro
6	Global financial crisis 1	2007-2009	turbul, abcp
7	Global financial crisis 2	2009-2010	intensif, exit, non-standard, refinanc
8	Duisenberg presidency	1999-2003	-
9	Euro introduction	1998-2003	banknot, changeov, ecsb
10	Treaty of Lisbon	2002-2006	vigil, lisbon

Table 4: Topic interpretations

Looking at the topic interpretations, note the following things:

- Every period is clearly dominated by one particular topic. The share of the dominating topic often exceeds 80%. Most periods last for about 2 years before being replaced by a new topic.
- Only topics 5 (quantitative easing) and 2 (banking system stability) are discussed in the majority of press conferences - with shares below 10% in most cases. Out of those, banking system stability is the only topic that does not dominate any period in the ECB's history.
- 2 topics still seem to be assigned to presidencies (Duisenberg and early Trichit presidencies) with no apparent link to any particular event.
- Both arguably most challenging crisis in the history of the ECB - the global financial crisis (2007-2010) and the European debt crisis (2010-2015) - have been assigned 2 topics each. This could be interpreted as a policy shift mid-way through the crisis.

Finally, I use the *estimateEffect()* function to check to what extend the presidency had an effect of these topics. The result is shown in *Table 5*, with coefficients above 0.15 highlighted.

	Intercept	Duisenberg	Trichet	Draghi
Topic 1	0.061	-0.045	**0.169**	-0.061
Topic 2	0.028	0.002	0.029	-0.004
Topic 3	0.115	-0.115	-0.115	**0.346**
Topic 4	0.060	-0.057	0.100	0.029
Topic 5	0.093	-0.076	-0.069	**0.237**
Topic 6	0.054	-0.054	**0.160**	-0.053
Topic 7	0.046	-0.030	0.061	0.016
Topic 8	0.111	**0.294**	-0.100	-0.083
Topic 9	0.109	**0.293**	-0.082	-0.101
Topic 10	0.071	0.030	0.105	-0.069

Table 5: Effect of presidents on 10 topics

Clearly, the effect of presidencies on topics declined compared to the 3-topic model. The method links topics 8 and 9 (Euro introduction) to the Duisenberg presidency. A somewhat weak link is also established between topics 1 and 6 (global financial crisis - phase 1) and Jean-Claude Trichet. Interestingly, the strongest link is established between topic 3 (European debt crisis - phase 2) and Draghi's presidency.

6 Conclusion

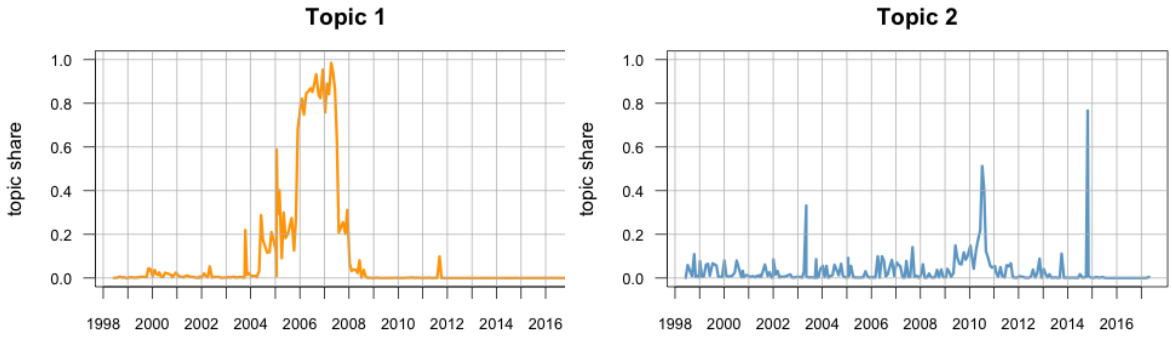
Topic distribution plots obtained with mixed-membership methods such as Latent Dirichlet Allocation (LDA) or Structural Topic Models (STM) show very well how ECB's policy and attention shifts over time. For small number of topics, the primary separation criterion is the president. However, for larger number of topics, the majority of topics can be linked to major events over the past 20 years that each dominated policy discussions for 2-5 years before (often swiftly) being replaced by a new agenda.

Topic interpretation has been done ad-hoc in this case. One could certainly improve the topic association by analyzing key terms and periods more carefully. The *stm* package in R provides a set of additional functions for this task that I did not exploit for the this content analysis (eg. *topicCorr()* to estimate topic relations or *findThoughts()* which outputs the most representative documents for a particular topic). Analysis could be further extended by introducing additional metadata such as stock index data or interest rates.

The project showed that ECB's policy focus is - to a surprisingly large extend - dominated by major economic events. Only 1 out of 10 topics - banking system stability - is not linked to a particular event or period in the history of the ECB.

Appendix

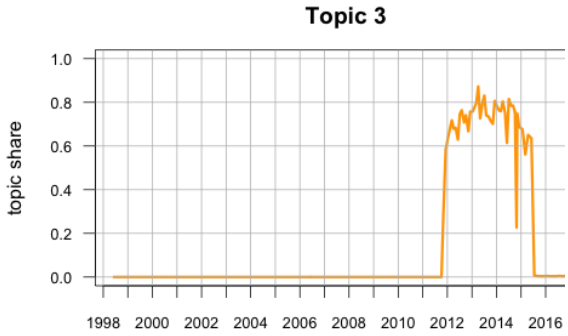
Topic 1	1	2	3	4	5	6	7
HProb	project	oil	observ	mention	dynam	upsid	figur
FREX	ampl	dynam	hous	upsid	baselin	materialis	counterpart
Lift	-appreci	aberr	ambigu	anonym	backward-look	barrier	bi-partisan
Score	plausibl	almunia	pertin	slovenia	pre-commit	romania	singapor

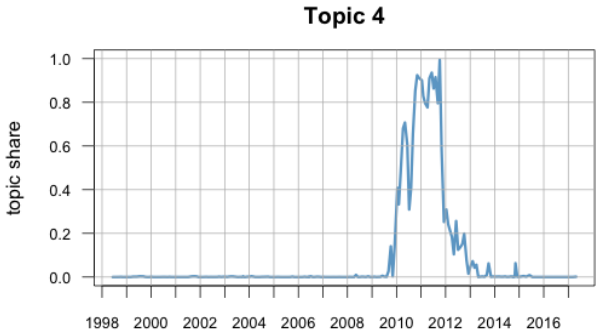


Topic 2	1	2	3	4	5	6	7
HProb	result	billion	capit	situat	exercis	stress	credit
FREX	exercis	test	shortfal	stress	billion	independ	result
Lift	file	beauti	quicker	ceb	doabl	red	chart
Score	test	shortfal	billion	slide	exposur	file	constâncio

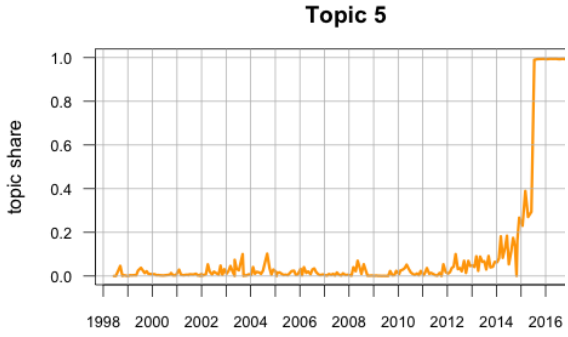
Topic 3	1	2	3	4	5	6	7
HProb	certain	credit	actual	differ	thing	answer	bond
FREX	omt	fragment	guidanc	basic	actual	sens	that
Lift	-go	-prolong	acquir	ad-hoc	anglo-irish	appl	artilleri
Score	omt	fragment	esm	guidanc	ssm	aqr	tltros

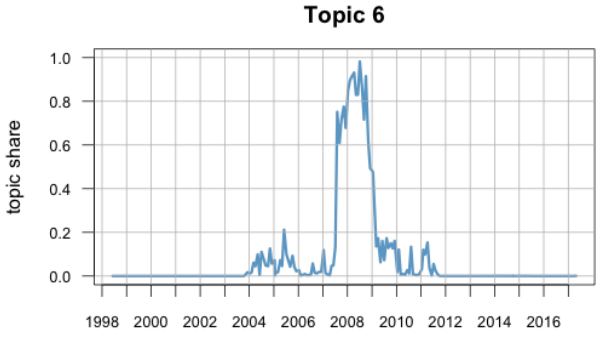
Topic 4	1	2	3	4	5	6	7
HProb	respons	appropri	programm	anchor	necessari	consid	deliv
FREX	efsf	non-standard	ireland	tension	allot	greec	head
Lift	africa	dormant	double-dip	flaw	geithner	grade	hungari
Score	non-standard	efsf	doctrin	smp	recapitalis	transmiss	default





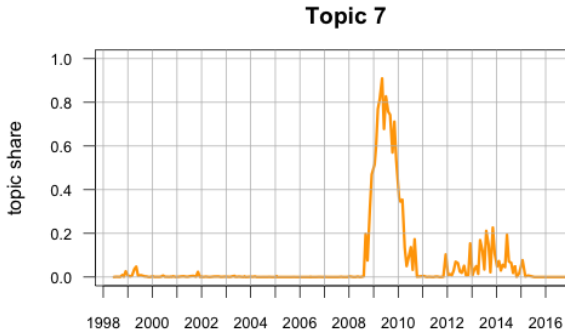
Topic 5	1	2	3	4	5	6	7
HProb	recoveri	purchas	loan	implement	project	oil	adjust
FREX	purchas	path	recoveri	march	boost	household	macroeconom
Lift	refuge	app	dombrovski	t-bill	cspp	incentivis	npl
Score	tltro	non-financi	securitis	growth-friend	recoveri	app	dombrovski

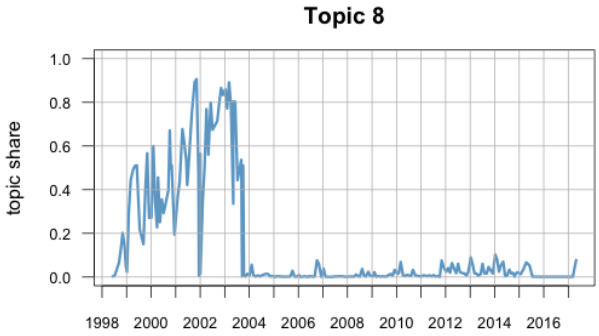




Topic 6	1	2	3	4	5	6	7
HProb	observ	number	money	credit	situat	upsid	certain
FREX	second-round	turbul	episod	atlant	fellow	alert	tension
Lift	abolish	abcp	across–board	agro-food	brutal	cartel	commodity-export
Score	needl	fellow	pre-commit	atlant	compass	hump	episod

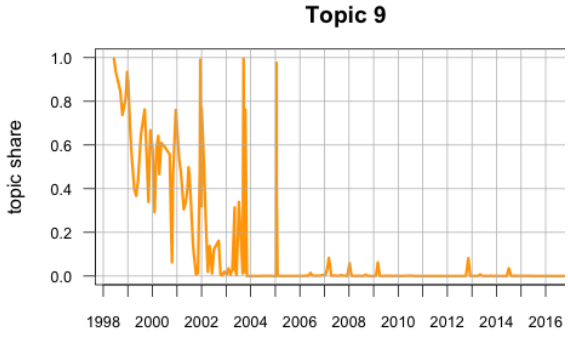
Topic 7	1	2	3	4	5	6	7
HProb	appropri	project	oper	loan	anchor	possibl	negat
FREX	intensif	refinanc	exit	deposit	negat	commod	non-standard
Lift	disinflationari	extrapol	feet	inelig	nationalist	plain	refi
Score	non-standard	intensif	exit	non-convent	recapitalis	non-financi	refinanc

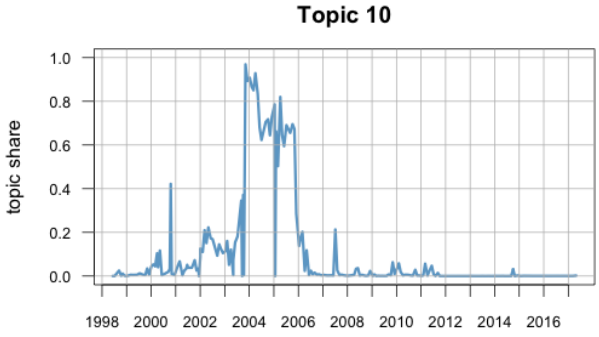




Topic 8	1	2	3	4	5	6
HProb	pillar	statement	uncertainti	answer	exchang	strategi
FREX	pillar	translat	statement	hope	consum	intervent
Lift	diseas	entrepreneurship	hämäläinen	knowledge-bas	lacklustr	outbreak
Score	duisenberg	forward-look	iss	co-ordin	unprocess	resumpt

Topic 9	1	2	3	4	5	6	7
HProb	banknot	exchang	issu	nation	system	valu	eurosystem
FREX	banknot	changeov	escb	coin	noyer	payment	cash
Lift	brochur	duck	esa	lame	lump-sum	maximis	mond
Score	banknot	changeov	co-oper	co-ordin	undervalu	duisenberg	escb





Topic 10	1	2	3	4	5	6	7
HProb	vigil	respons	oil	observ	mention	europ	situat
FREX	vigil	lisbon	diagnosi	excess	bodi	partner	save
Lift	algorithm	barroso	charli	cornerston	defect	kok	krone
Score	vigil	diagnosi	sticki	non-inflationari	lisbon	magnet	pragmat