Master Thesis in Data Science

Barcelona Graduate School of Economics

Tracking the Economy Using FOMC Speech Transcripts

Laura Battaglia, Maria Salunina

Supervisor: Omiros Papaspiliopoulos

June 2020

To my old self, for finding the courage

To Katya, for being my mentor and greatest support

To each other, for making diversity our core strength

Abstract

In this study, we propose an approach for the extraction of a low-dimensional signal from a collection of text documents ordered over time. The proposed framework foresees the application of Latent Dirichlet Allocation (LDA) for obtaining a meaningful representation of documents as a mixture over a set of topics. Such representations can then be modeled via a Dynamic Linear Model (DLM) as noisy realisations of a limited number of latent factors that evolve with time. We apply this approach to Federal Open Market Committee (FOMC) speech transcripts for the period of Greenspan presidency. We are able to extract a latent factor that fairly resembles the Economic Policy Uncertainty Index for United States. This study serves as exploratory research for the investigation into how unstructured text data can be incorporated into economic modeling. In particular, our findings point at the fact that a meaningful state-of-the-world signal can be extracted from expert's language, and pave the way for further exploration into the building of macroeconomic forecasting models, and in general into the usage of variation in language for learning about latent economic conditions.

Keywords: Signal extraction; Topic model; Dynamic linear model; FOMC.

Acknowledgement

We would like to express our sincere gratitude to our supervisor, Omiros Papaspiliopoulos, for the extensive guidance and encouragement he provided not only during this research process but also throughout the whole year. Without his support and professional expertise in probabilistic inference, this project would not have been possible.

We are as well indebted to Stephen E. Hansen for both providing data for this research and sharing with us his profound knowledge in the application of unstructured data in economics and monetary policy. This work would not have advanced this far without his precious input and feedback.

Contents

Chapter 1

Introduction

This chapter provides a general background to our research question and explains the primary motivations behind our study. It as well includes a content outline and the main conventions used in the paper.

1.1 Background and Motivation

According to the International Data Corporation (IDC), while in 2013 the size of the "digital universe" amounted to 4.4 Zettabytes (ZB, 1 ZB = trillion gigabytes) (IDC 2014), it rocketed to 33 ZB already by 2018, and it is predicted to grow further up to 175 ZB by 2025 (Reinsel et al. 2018). Most of this information consists of unstructured data that is not organised in a pre-defined manner and is thus challenging to process and analyse. Yet, unstructured data is an incredibly rich source of information that can be relevant in infinitely many applications.

Most of this data consists of text, and this also explains why the field of text analytics is steadily gaining broader general interest and market share over time[1]. In the past few years, also in social sciences several research papers focused on different applications of text mining techniques to address economic problems, including the analysis of Central Bank communication, the estimation of a variety of macroeconomic variables, and the measuring of policy uncertainty and of the political slant of media content (see Gentzkow et al. (2019) for a recent literature review). However,

[1] According to the report of Global Market Insights, "Text Analytics Market size surpassed USD 4 billion in 2018 and is anticipated to grow at over 18% CAGR from 2019 to 2026" (Wadhwani & Kasnale 2019).

empirical work in social sciences still mainly relies on numeric data, leaving most of the potential of text information untapped.

In particular, our main motivation for this analysis lies in exploring ways for extracting meaningful information from text corpora that are evolving with time. Specifically, we will propose an approach for extracting a low-dimensional signal tracking the evolution of topic usage in a collection of ordered text documents. Exploiting the public availability of US Federal Reserve's Federal Open Market Committee (FOMC) transcripts, we will implement our framework to extract a low-dimensional representation of FOMC monetary policy deliberations, and explore whether this signal can be put in relation with macroeconomic variables of interest. This analysis could pave the way for further enriching macroeconomic models and improving macroeconomic forecasts. Indeed, there is extensive literature analysing FOMC transcripts (e.g. Hansen et al. (2018), Woolley & Gardner (2017), Acosta (2015) Schonhardt-Bailey (2013)), but the main focus is generally on transparency of central banking decisions and on how this might affect policy makers' deliberations. Moreover, as far as we are aware, combinations of topic and dynamic linear modelling techniques were so far not investigated.

1.2 Thesis Outline

The following chapters are organised as follows. Chapter 2 provides an overview of the theoretical framework behind our analysis. In Chapter 3, we introduce our approach for the extraction of a low-dimensional signal from ordered text documents and present results of its application to FOMC speech transcripts. We also discuss limitations and possible extensions. Finally, in Chapter 4 we provide concluding remarks and point at potential future explorations.

1.3 Conventions

In this paper, we will use the following notation:

- bold lowercase letters denote vectors, e.g. $\boldsymbol{y}$;

- bold uppercase letters denote matrices, e.g. $\boldsymbol{A}$;
- scalars can be represented by both upper and lowercase letters, but never in bold, e.g. k, N;
- a colon denotes a collection of random variables, e.g. $y_{1:t} = (y_1, y_2, \ldots, y_t)$.

Chapter 2

Preliminaries

This chapter provides the background theory to the proposed framework. In particular, it includes an overview of Latent Dirichlet Allocation and Multivariate Dynamic Linear Models.

2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model which represents documents as random mixtures over latent topics, where each latent topic is characterised by a distribution over unique terms in a vocabulary (Blei et al. 2003). The general idea of topic modeling is thus to define ’topics’ as specific term distributions and to decompose each document into the shares devoted to each topic. Table 2.1 summarizes the notation used for the description of the LDA approach.

The assumed generative process for LDA can be expressed as follows:

Symbol	Description
K	number of topics
V	number of unique terms in the vocabulary
D	number of documents
N_d	number of words in document d
$\boldsymbol{\theta}_d$	topic proportions specific to document d
$\boldsymbol{\beta}_k$	word proportions specific to topic k
$z_{d,n}$	identity of the topic of the n-th word in document d
$w_{d,n}$	identity of the n-th word in document d
α, η	parameters of the prior Dirichlet distributions

Table 2.1: LDA notations

1. For each topic $k = 1, \dots, K$

   (a) Draw word proportions $\boldsymbol{\beta}_k \sim \text{Dirichlet}(\eta)$

2. For each document $d = 1, \dots, D$

   (a) Draw topic proportions $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha)$

   (b) For each word $n = 1, \dots, N_d$

      i. Draw a topic assignment $z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$

      ii. Draw a word $w_{d,n} \sim \text{Multinomial}(\boldsymbol{\beta}_{z_{d,n}})$

where Dirichlet(·) and Multinomial(·) represent Dirichlet and Multinomial distributions respectively. Parameters of the Multinomial distributions, i.e. $\boldsymbol{\theta}_d$ and $\boldsymbol{\beta}_k$, are drawn from the conjugate prior Dirichlet distributions, which allows for efficient calculations of the likelihood function. A graphical representation of the LDA generative process, which illustrates dependencies between parameters and variables, is shown in Figure 2.1.

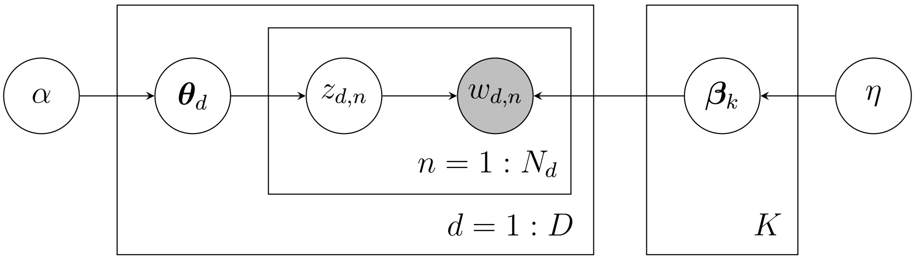


Figure 2.1: Graphical model representation of LDA (Blei 2012)[1]

Given this generative process for LDA, the joint distribution of the latent and observed variables can be written as follows:

$$p(\boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D}, z_{1:D}, w_{1:D}) = \prod_{k=1}^{K} p(\boldsymbol{\beta}_k) \prod_{d=1}^{D} p(\boldsymbol{\theta}_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\boldsymbol{\theta}_d) p(w_{d,n}|\boldsymbol{\beta}_{1:K}, z_{d,n}) \right).$$

To infer the topic structure, we would like to estimate the posterior distribution given the observed set of documents. Applying the Bayes theorem, we get an expression

[1] Shaded nodes stand for observed variables, rectangles denote replication.

for the corresponding posterior:

$$p(\boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D}, z_{1:D} | w_{1:D}) = \frac{p(\boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}.$$

Since the number of possible topic structures is exponentially large, marginal probability of the observations, $p(w_{1:D})$, which is theoretically computed as a sum of the joint distributions over every possible topic structure, is intractable to estimate in practice. Hence, variational or sampling-based algorithms are used to efficiently approximate the described posterior distribution (Blei 2012).

While sampling-based methods attempt approximating the posterior with empirical distribution estimated on collected samples, variational algorithms assume a parametrized family of distributions over the hidden structure and try to identify the member of the family which is closest to the posterior. In our analysis, we will use the first approach and in particular apply the collapsed Gibbs sampling algorithm for topic modelling of Griffiths & Steyvers (2004)[2] to infer the hidden topic structure of FOMC transcripts.

2.2 Multivariate Dynamic Linear Models

2.2.1 Definition of Multivariate DLMs

The first Bayesian approach to forecasting based on a dynamic linear model (DLM) was introduced in 1976 (Harrison & Stevens 1976) and was later developed by West and Harrison (West & Harrison 1997).

Multivariate DLMs represent a particular class of state-space models and can be

[2] The following packages provide efficient implementations of the collapsed Gibbs sampling method: R package **lda** http://CRAN.R-project.org/package=lda (Chang 2015), Python package **topicmodels** https://github.com/sekhansen/text-mining-tutorial (Hansen et al. 2018).

defined by the following set of equations:

$$\begin{array}{lll} \text{Observation equation:} & \boldsymbol{y}_t = \boldsymbol{F}_t\boldsymbol{\theta}_t + \boldsymbol{v}_t, & \boldsymbol{v}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_t) \\ \text{State equation:} & \boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{w}_t, & \boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{W}_t), \end{array}$$

where (i) $\boldsymbol{y_t}$ is an observable k-dimensional time series of observations at time t, for $t = 1, \ldots, T$, (ii) $\boldsymbol{\theta}_t$ is a p-dimensional unobserved state vector, (iii) $\boldsymbol{F_t}$ is a known regression $k \times p$ matrix, (iv) $\boldsymbol{G}_t$ is a $p \times p$ state transition matrix, (v) $\boldsymbol{v}_t$ is a zero-mean k-dimensional vector of the observation equation residuals and (vi) $\boldsymbol{w}_t$ is a zero-mean p-dimensional vector representing evolution noise. The sequences $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$ are assumed to be independent and mutually independent, and independent of $\boldsymbol{\theta}_0$ (West & Harrison 1997).

While the variance and other structural parameters in DLMs can be estimated by numerical optimization or by Markov Chain Monte Carlo (MCMC) methods, evaluation of the states, assuming a known vector of parameters, can be efficiently performed using standard recursive Kalman formulas[3] (Laine 2020).

2.2.2 Recursive Kalman Formulas

Below we provide necessary formulas for Kalman filtering that allow us to estimate the conditional distributions of the DLM states, given observable time series and an assumed vector of parameters.

Kalman filtering can be described as a two-step process recursively repeated at each time stamp t, that provides distributions of the states at each time t given the observations up to the current time. In the first prediction step, the algorithm produces an estimation for the prior distribution of the one-step-ahead states. In the second update step, the Kalman filter estimates the posterior distribution of these states, taking into account the information about observed measurements.

[3] R package **dlm** `https://CRAN.R-project.org/package=dlm` (Petris 2010) focuses on Bayesian analysis of DLMs, providing high flexibility in defining user's models and offering methods for estimating both parameters and states of the DLM.

For the estimation of the prior $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{y}_{1:t-1}, \boldsymbol{F}_t, \boldsymbol{G}_t, \boldsymbol{V}_t, \boldsymbol{W}_t) = \mathcal{N}(\widehat{\boldsymbol{\theta}}_t, \widehat{\boldsymbol{C}}_t)$ at the first step, the mean and covariance matrices for $\boldsymbol{\theta}_t$ and $\boldsymbol{y}_t$ are calculated as follows:

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_t &= \boldsymbol{G}_t\overline{\boldsymbol{\theta}}_{t-1} && \text{prior mean for } \boldsymbol{\theta}_t, \\ \widehat{\boldsymbol{C}}_t &= \boldsymbol{G}_t\overline{\boldsymbol{C}}_{t-1}\boldsymbol{G}_t^T + \boldsymbol{W}_t && \text{prior covariance for } \boldsymbol{\theta}_t, \\ \widehat{\boldsymbol{C}}_{y,t} &= \boldsymbol{F}_t\widehat{\boldsymbol{C}}_t\boldsymbol{F}_t^T + \boldsymbol{V}_t && \text{covariance for predicting } \boldsymbol{y}_t. \end{aligned}$$

Next, we estimate the posterior distribution $\mathcal{N}(\overline{\boldsymbol{\theta}}_t, \overline{\boldsymbol{C}}_t)$, using the Kalman gain matrix $\boldsymbol{K}_t$ as follows:

$$\begin{aligned} \boldsymbol{K}_t &= \widehat{\boldsymbol{C}}_t\boldsymbol{F}_t^T\widehat{\boldsymbol{C}}_{y,t}^{-1} && \text{Kalman gain,} \\ \boldsymbol{r}_t &= \boldsymbol{y}_t - \boldsymbol{F}_t\widehat{\boldsymbol{\theta}}_t && \text{prediction residual,} \\ \overline{\boldsymbol{\theta}}_t &= \widehat{\boldsymbol{\theta}}_t + \boldsymbol{K}_t\boldsymbol{r}_t && \text{posterior mean for } \boldsymbol{\theta}_t, \\ \overline{\boldsymbol{C}}_t &= \widehat{\boldsymbol{C}}_t - \boldsymbol{K}_t\boldsymbol{F}_t\widehat{\boldsymbol{C}}_t && \text{posterior covariance for } \boldsymbol{\theta}_t. \end{aligned}$$

These calculations are repeated for every time t, and the values of $\overline{\boldsymbol{\theta}}_t$ and $\overline{\boldsymbol{C}}_t$ are stored for consecutive iterations. For the first iteration we assume that the initial distribution of $\boldsymbol{\theta}_0$ at $t = 0$ is available. A notable feature of the linear Gaussian case is that the formulas above are exact and easily implemented in computer as long as the model state dimension or the number of observations at one time is not too large (Laine 2020).

Chapter 3

Proposed Approach and Experiments

3.1 Framework Overview

In our framework, we assume that evolution of topic proportions in a collection of ordered documents is driven by an underlying low-dimensional signal, which can be modeled by a dynamic linear model and represents the object of interest. We also assume that the term-distribution within topics remains stable over the time period. For the extraction of this signal, we suggest the following sequential approach:

1. **Estimate a representation of topic usage over time via LDA.**

   We hereby assume that each document corresponds to a unique timestamp. Therefore, in this framework, time-level is equivalent to document-level, and we denote topic proportions at time t as vector $\boldsymbol{\theta}_t$. At this first stage, we can estimate the vectors $\boldsymbol{\beta}_k$ of term-distributions over topics and the vectors $\boldsymbol{\theta}_t$ of topic proportions at each point in time by fitting a classic LDA model as described in Paragraph 2.1 e.g. via Gibbs Sampling method (Griffiths & Steyvers 2004). At this stage, we are thus still making no account of the time-varying element of $\boldsymbol{\theta}_t$. We obtain as well estimates for the latent variable $\boldsymbol{z}_{t,n}$ representing the allocation of each word in the corpus to a topic. We can thus retrieve two representations of the evolution of topic usage over time, i.e. the topic proportions $\boldsymbol{\theta}_t$, as well as the word-counts per topic $\boldsymbol{c}_t$ representing how many words were allocated to each topic at each point in time.

   It should be noted that, if data has a richer granularity than the time-level, e.g.

if we have a set of individual documents within each time stamp, there is the need to aggregate the topic shares of interest from the document level ($\boldsymbol{\theta}_{d,t}$) or the word-counts per topic ($\boldsymbol{c}_{d,t}$) to the time level ($\boldsymbol{\theta}_t$, $\boldsymbol{c}_t$). There is no obvious way to address this problem. A sensible strategy could be that of pursuing a two-steps approach, where (i) first, word-distributions over topics $\boldsymbol{\beta}_k$ are estimated via an LDA at the document-time level, (ii) then, raw documents are aggregated to the time level and (iii) finally, topic-time proportions $\boldsymbol{\theta}_t$ or word-counts per topic $\boldsymbol{c}_t$ are re-estimated on the aggregated documents by keeping the original $\boldsymbol{\beta}_k$ as fixed. We find this approach more suitable than the alternative of aggregating documents to time-level first and then estimating LDA directly at the time level: indeed, it allows us to retrieve topics at a more granular level - where it is more likely that each document is more or less centered around one topic - and then retrieve a representation of topic usage at the time-level of interest.

2. **Estimate the latent factor(s) driving topic usage via DLM**

We will then specify an appropriate dynamic linear model for the chosen representation of topic usage over time, that is aimed at obtaining a lower dimensional representation of topic usage into a limited number of factors. Depending on the data and on the specific assumptions made, an appropriate number of factors should be chosen and components such as trends or seasonality should be taken into account.

As mentioned above, the variances and other structural parameters in DLMs can be estimated by numerical optimization or by Markov Chain Monte Carlo (MCMC) methods, while evaluation of the states, assuming a known vector of parameters, can be efficiently performed using standard recursive Kalman formulas. In this way, we can retrieve an estimate for the underlying latent factor(s) of interest.

3.2 Analysis on FOMC Transcripts

In this section, we apply the proposed framework on FOMC transcripts and put the extracted lower-dimensional signal in relation with specific macroeconomic variables.

3.2.1 Data Description

In order to examine the suggested framework, we use Federal Open Market Committee (FOMC) transcripts from the period 08-1986 to 01-2006 of Greenspan presidency. The FOMC is a committee consisting of the Governors of the Fed's Board and the presidents of five Federal Reserve Banks, that defines monetary policy for the Federal Reserve System by setting a target for the federal funds rate[1].

The FOMC holds eight meetings per year. In these meetings, two main topics are at the center of the discussion: economic situation (FOMC1) and monetary policy strategy (FOMC2). In our work, we decided to exclusively focus on FOMC1, being interested in investigating whether the estimated latent factors driving topics of this section can be reasonably put in relation with some macroeconomic variable describing aspects of the economy state. We also filter out staff statements, as they mostly represent only series of questions to the FOMC members.

The data under analysis thus consists into transcripted statements of FOMC1 members for a total of 148 meetings.

3.2.2 Data Preprocessing

Prior to estimation, the raw statement text needs to be preprocessed in several steps. These include the usual removal of stopwords and stemming or lemmatization, plus additional ad hoc preprocessing that one might deem appropriate given the problem at hand. Thankfully, for our analysis we were given access to topic allocations $z_{d,n}$ for each word in our text corpus, estimated via fitting a classic LDA

[1] See https://www.newyorkfed.org/aboutthefed/fedpoint/fed48.html for more details.

with a given set of hyperparameters[2] on already preprocessed statement-level data for the period of Greenspan governance (i.e. 08-1986 to 01-2006). This data was kindly provided by Stephen E. Hansen and was produced as part of his and his co-authors' analysis on how transparency of central banking policy-making impacts policy maker's deliberations (Hansen et al. 2018). The main advantage of making use of this data is that of accessing text that was specifically preprocessed to reduce the vocabulary to a set of terms that are most likely to reveal the underlying content of interest, thus facilitating the estimation of more semantically meaningful topics. This ad-hoc preprocessing thus also included the identification of bi-grams or tri-grams that have a specific meaning in our context, via tabulating frequencies of specific part-of-speech patterns and retaining those word sequences that have relatively high frequency in the corpus. Given that the quality of the data is of very high relevance to our analysis, we deemed this approach to be the most appropriate. On the other hand, it limited us to the choice of this specific time-span as well as of specific hyperparameters for LDA estimation - and in particular of $K = 40$ topics as the number of topics used for LDA estimation. As a future exploration, it would be interesting to relax this assumption and explore a wider dataset and a different range of hyperparameters.

3.2.3 Term-distribution over Topics

As said, as a first step we use LDA to retrieve from our text corpus a suitable representation of topics usage at each point in time. The main advantage of starting from estimated LDA word-topic allocations $z_{d,n}$ from text at the statement level, rather than already aggregated to the meeting (time-stamp) level, is that word-distributions over topics $\boldsymbol{\beta}_k$ are retrieved from a granular corpus - where it is more likely that each document (statement) is more or less centered around one topic. Retrieving term-distributions from text already aggregated to the meeting (time-stamp) level would instead result in making use of documents that include a wide range of different speakers going over potentially quite different subjects.

[2] Specifically, number of topics equal to $K = 40$, prior on $\boldsymbol{\theta}_{d,t}$ topic proportions equal to $\alpha = 50/K$, prior on $\boldsymbol{\beta}_k$ word proportions equal to $\eta = 0.025$.

Results for term-distributions over topics are shown in Figure 3.1. For each topic k, we show the 10 terms that are associated with the highest values $\beta_{v,k}$ ($v = 1, ...V$ total unique terms in vocabulary) in descending order. Darker shades on terms indicate higher probabilities. As expected, given the low value of η (hyperparameter responsible for the $\boldsymbol{\beta}_k$ prior), topics have a limited number of words with relatively high probability and a much larger number of words with relatively low probability. If, on the one hand, the use of a high number of topics ($K = 40$) results in some topics not being of particular relevance to our aims (e.g. topics which are generically including pleasantries), on the other hand, we see that most of the topics form quite natural distinguished groupings of words which are relatively easy to interpret. This feature allows us to associate natural labels to each of them, which are shown in the left-most column of the figure and will be used for our subsequent analysis[3].

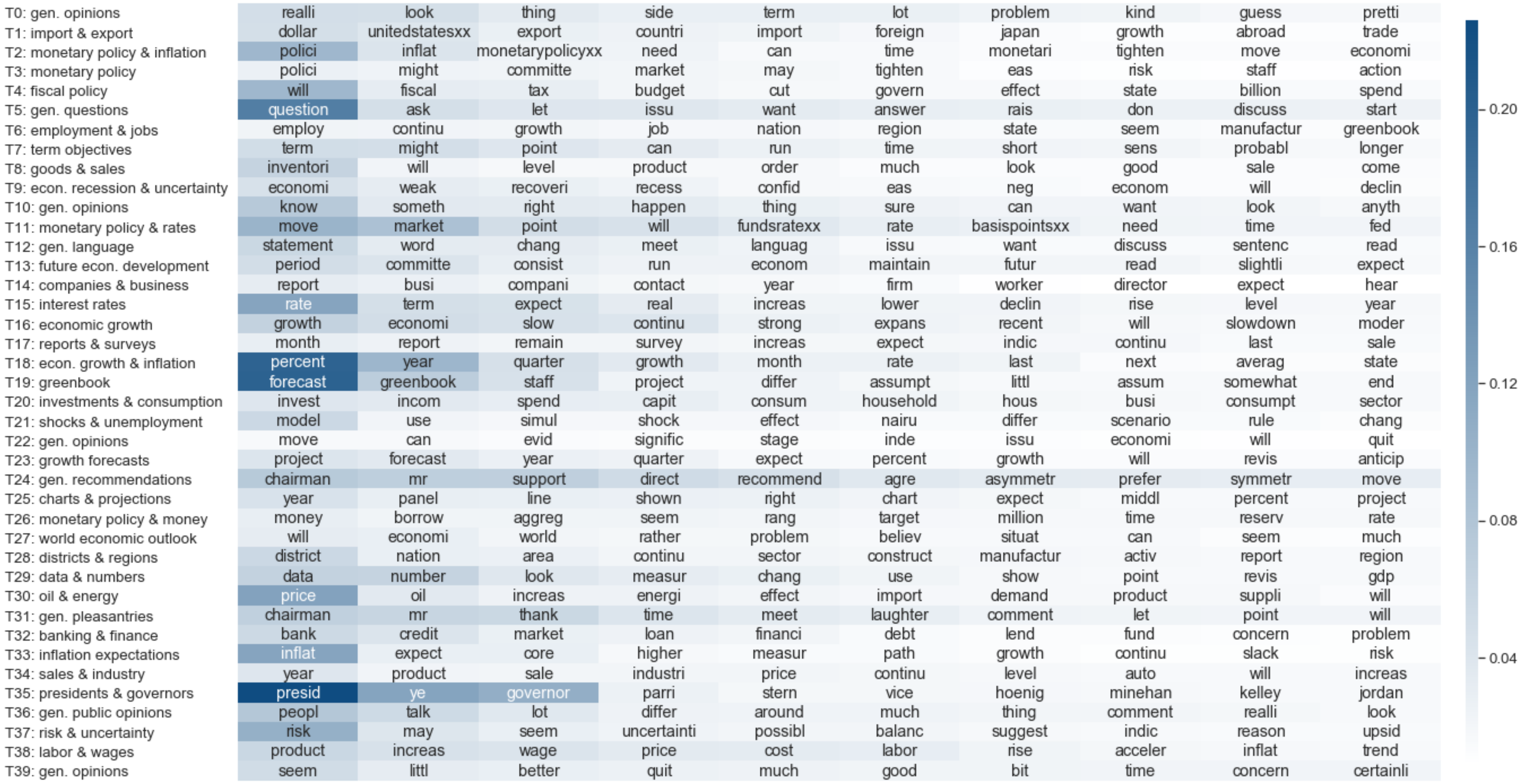
In particular, we can identify some topics related to the state of the economy from different perspectives (e.g. economic growth, economic recession & uncertainty), while some others related to labor and employment (e.g. employment & jobs, shocks unemployment, labor & wages) or monetary policy (monetary policy & inflation, monetary policy, monetary policy & rates).

3.2.4 Representation of Topic Usage

If, on the one hand, retrieving topic term-distributions from granular text allows for a convenient identification of fairly distinguishable topics, on the other hand, this choice implies the need for devising a sensible strategy for aggregating our measure for topic usage to the time-stamp level of interest.

To this end, we are pursuing two different strategies that will then shape two streams of further analysis. A first strategy is that of using the word-topic allocations $z_{d,n}$ to compute word counts per topic at the time-stamp level. One thus obtains 40 time series, each of length 148, representing how the number of words per each topic changed over time. While this approach gives a good insight into how the use of

[3] An important caveat here is that these interpretations are subjective to our judgement and are outside of the statistical model.



T0: gen. opinions	realli	look	thing	side	term	lot	problem	kind	guess	pretti
T1: import & export	dollar	unitedstatesxx	export	countri	import	foreign	japan	growth	abroad	trade
T2: monetary policy & inflation	polici	inflat	monetarypolicyxx	need	can	time	monetari	tighten	move	economi
T3: monetary policy	polici	might	committe	market	may	tighten	eas	risk	staff	action
T4: fiscal policy	will	fiscal	tax	budget	cut	govern	effect	state	billion	spend
T5: gen. questions	question	ask	let	issu	want	answer	rais	don	discuss	start
T6: employment & jobs	employ	continu	growth	job	nation	region	state	seem	manufactur	greenbook
T7: term objectives	term	might	point	can	run	time	short	sens	probabl	longer
T8: goods & sales	inventori	will	level	product	order	much	look	good	sale	come
T9: econ. recession & uncertainty	economi	weak	recoveri	recess	confid	eas	neg	econom	will	declin
T10: gen. opinions	know	someth	right	happen	thing	sure	can	want	look	anyth
T11: monetary policy & rates	move	market	point	will	fundsratexx	rate	basispointsxx	need	time	fed
T12: gen. language	statement	word	chang	meet	languag	issu	want	discuss	sentenc	read
T13: future econ. development	period	committe	consist	run	econom	maintain	futur	read	slightli	expect
T14: companies & business	report	busi	compani	contact	year	firm	worker	director	expect	hear
T15: interest rates	rate	term	expect	real	increas	lower	declin	rise	level	year
T16: economic growth	growth	economi	slow	continu	strong	expans	recent	will	slowdown	moder
T17: reports & surveys	month	report	remain	survey	increas	expect	indic	continu	last	sale
T18: econ. growth & inflation	percent	year	quarter	growth	month	rate	last	next	averag	state
T19: greenbook	forecast	greenbook	staff	project	differ	assumpt	littl	assum	somewhat	end
T20: investments & consumption	invest	incom	spend	capit	consum	household	hous	busi	consumpt	sector
T21: shocks & unemployment	model	use	simul	shock	effect	nairu	differ	scenario	rule	chang
T22: gen. opinions	move	can	evid	signific	stage	inde	issu	economi	will	quit
T23: growth forecasts	project	forecast	year	quarter	expect	percent	growth	will	revis	anticip
T24: gen. recommendations	chairman	mr	support	direct	recommend	agre	asymmetr	prefer	symmetr	move
T25: charts & projections	year	panel	line	shown	right	chart	expect	middl	percent	project
T26: monetary policy & money	money	borrow	aggreg	seem	rang	target	million	time	reserv	rate
T27: world economic outlook	will	economi	world	rather	problem	believ	situat	can	seem	much
T28: districts & regions	district	nation	area	continu	sector	construct	manufactur	activ	report	region
T29: data & numbers	data	number	look	measur	chang	use	show	point	revis	gdp
T30: oil & energy	price	oil	increas	energi	effect	import	demand	product	suppli	will
T31: gen. pleasantries	chairman	mr	thank	time	meet	laughter	comment	let	point	will
T32: banking & finance	bank	credit	market	loan	financi	debt	lend	fund	concern	problem
T33: inflation expectations	inflat	expect	core	higher	measur	path	growth	continu	slack	risk
T34: sales & industry	year	product	sale	industri	price	continu	level	auto	will	increas
T35: presidents & governors	presid	ye	governor	parri	stern	vice	hoenig	minehan	kelley	jordan
T36: gen. public opinions	peopl	talk	lot	differ	around	much	thing	comment	realli	look
T37: risk & uncertainty	risk	may	seem	uncertainti	possibl	balanc	suggest	indic	reason	upsid
T38: labor & wages	product	increas	wage	price	cost	labor	rise	acceler	inflat	trend
T39: gen. opinions	seem	littl	better	quit	much	good	bit	time	concern	certainli

Figure 3.1: Estimated topics content: terms within topics ranked by probability

topics evolved with time, at the same time the evolution of word counts per se is an unbounded series that might be influenced by other factors or trends e.g. varying length of the meetings or of a number of speakers.

Another strategy is that of aggregating the documents to the time level and re-estimating time-topic distributions $\boldsymbol{\theta}_t$ at the time-stamp level while keeping the original $\boldsymbol{\beta}_k$ as fixed. This results in 40 time series of estimated use of topic probabilities over time.

Analysis and results under the two approaches are presented in the following two sections.

3.2.5 DLM for Topic-word Count Time Series

The aim of this part of the analysis is to retrieve a suitable lower dimensional representation of the estimated topic-word count time series that can well describe the main dynamics of use of topics across time. To this end, we want to specify a dynamic model where topic-word counts are regarded as a noisy representation of a limited number of factors.

Staying general, a way to specify such model could be the following:

$$\begin{aligned}
\boldsymbol{c}_t &= \text{Multinomial}(\boldsymbol{\theta}_t), && \\
\boldsymbol{\theta}_t &= \pi(\boldsymbol{u}_t), && \\
\boldsymbol{u}_t &= \boldsymbol{A}\boldsymbol{f}_t + \boldsymbol{\epsilon}_t, && \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{V}) \\
\boldsymbol{f}_t &= \boldsymbol{f}_{t-1} + \boldsymbol{\nu}_t, && \boldsymbol{\nu}_t \sim N(\mathbf{0}, \boldsymbol{W}),
\end{aligned}$$

where (i) $\boldsymbol{c}_t$ is the K-dimensional vector of word-counts per topic at time t as estimated in Section 3.2.4., (ii) $\boldsymbol{\theta}_t$ is the K-dimensional vector representing probabilities of each topic being used at time t, (iii) $\pi(\cdot)$ is a mapping of $\boldsymbol{u}_t$ real values to probabilities vectors $\boldsymbol{\theta}_t$, and (iv) $\boldsymbol{f}_t$ represents a vector of latent factors. In other words, we would assume counts to be generated from a Multinomial distribution with an associated vector of probabilities $\boldsymbol{\theta}_t$ representing the use of topics, whose real-mapped

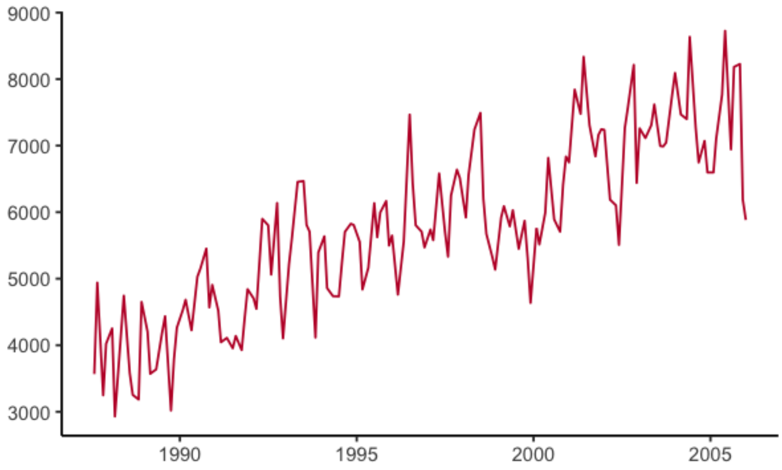


Figure 3.2: Evolution of the total number of words used in the FOMC1 section by non-staff speakers

values would then be modeled with a linear Gaussian dynamic factor model with a suitable number of factors. The matrices $\boldsymbol{V}$ and $\boldsymbol{W}$ should be appropriately modelled so that $\boldsymbol{V}$ well represents correlation across topic shares, while $\boldsymbol{W}$ models how far autocorrelated factors can evolve from their past values.

For estimation, we will make some assumptions so to simplify the model into a more tractable setting. In particular, we will model word-counts per topic with a linear Gaussian model with independent errors, thus disregarding correlations across topic time series.

As to the appropriate specification and number of factors, after an extensive exploration we have decided to fit our model to a topic-specific intercept, a linear trend and a latent factor which is modeled as a random walk. The main driver for our decision is the will to try and disentangle a "slow" and roughly stationary signal describing use of topics from the word-count dynamics related to the structure of the meetings per se. In this regard, we observe that in the period of interest there was a general steady increase in the total number of words per meeting over time (see Figure 3.2), likely due to an increase in the meeting length over time. Our attempt is thus to capture this trend with a linear deterministic trend and isolate a slow-moving factor that will then represent our signal.

The model under analysis thus collapses to the following state space model:

$$\begin{aligned} \boldsymbol{c}_t &= \boldsymbol{d} + \boldsymbol{\beta} t + \boldsymbol{l} f_t + \boldsymbol{\epsilon}_t & \qquad \boldsymbol{\epsilon}_t &\sim N(\boldsymbol{0}, \gamma_c \boldsymbol{I}) \\ f_t &= f_{t-1} + \nu_t & \qquad \nu_t &\sim N(0, \gamma_f) \end{aligned}$$

We fit the model via Kalman filtering with parameters obtained from Maximum Likelihood Estimation. To this end, we use the renowned *dlm* package in R[4]. The estimation requires initialisation values for the parameters and the mean and variance of the factor. Initial values are particularly important, since a bad initialisation could cause the algorithm to be trapped into some sub-optimal local maximum. As a sensible way to define initial values, we set initial $\boldsymbol{d}$, $\boldsymbol{\beta}$, $\boldsymbol{l}$ and γ_c to the estimated intercept, coefficients and average residual variance retrieved when regressing $\boldsymbol{c}_t$ on a trend and the unemployment rate time series for the period under scrutiny. Initial values for the factor mean and γ_f are set to the average unemployment rate over the period and the residual variance obtained when regressing unemployment on its lag. The choice of unemployment rate was driven by the will to initialise the factor around values resembling a slow-moving, quasi-stationary process such as the signal we wish to extract.

The algorithm converges to parameter values which are not too far off the initialization ones. Figure 3.3 shows the estimated factor against the initial unemployment values around which it was initialised. As we can see, the factor converged to values that are within a similar range to that of unemployment, but the evolution of the estimated latent factor is quite different.

In Figure 3.4 we show fitted values against word-count time series per each topic. In each plot, one can find the reference number and theme for the topic and the estimated loading for the latent factor at each topic. As it is reasonable, a linear trend plus a single factor does not manage alone to fit our data perfectly. However, such a low dimensional representation of the data already does a fairly good job

[4] R package **dlm** https://CRAN.R-project.org/package=dlm (Petris 2010) focuses on Bayesian analysis of DLMs, providing high flexibility in defining user's models and offering methods for estimating both parameters and states of the DLM.

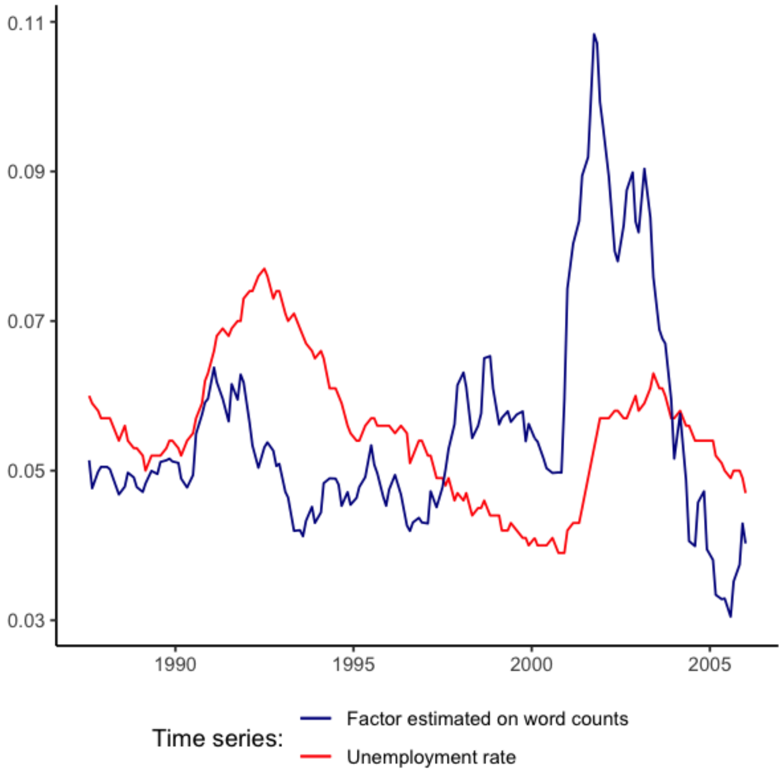


Figure 3.3: Factor estimated on the topic-word count time series versus unemployment rate

with a number of topics time series. In particular, we can see that while a few topics are mainly only fitted with the linear trend (e.g. T3 - monetary policy, T5 - general questions, T17 - reports & surveys), there is a subgroup of topics for which the factor is majorly contributing and providing a better fit to the respective word counts (e.g. T33 - inflation expectations, T13 - future economic development, T9 - economic recession & uncertainty, T14 - companies & business). It thus seems that the factor is trying to find a common low-dimensional signal that could fairly fit the word counts for a subgroup of specific topics, rather than trying to fit all the topics to the same extent.

However, we again point out that word counts per topic not only capture the evolution of use of topics over time, but also exogenous elements such as the increasing meeting length over time that we mentioned above. Since our aim is to find a low dimensional representation of use of topics, this implies that the model is trying to fit to features of our data that are not at the core of our interest. Moreover, such

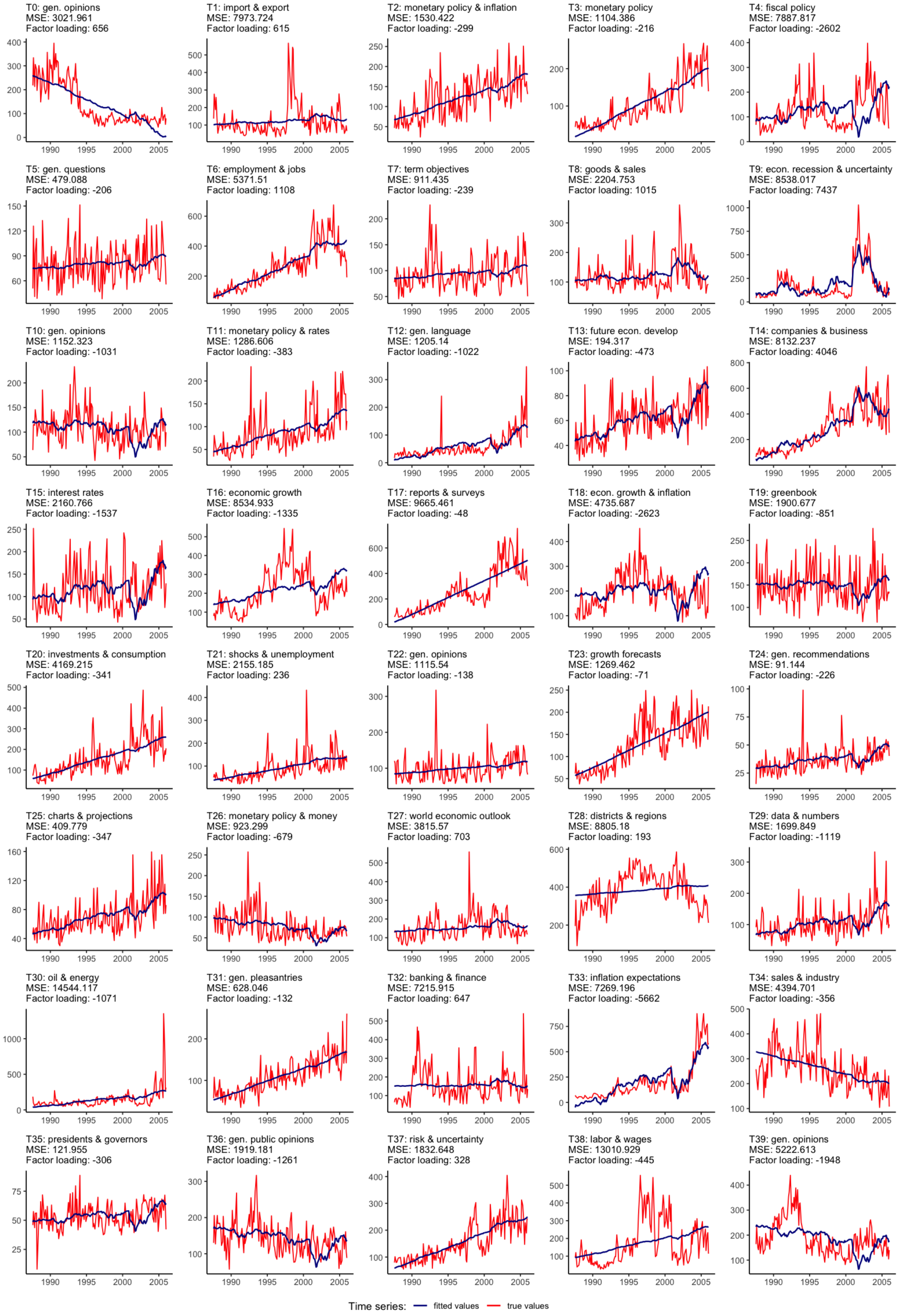


Figure 3.4: Fitted versus true values for topic-word counts time series

features, including the upward trend in meeting length, could be masking trends in use of topics that could be of our interest. We thus decide to compare results to a similar model fitted on time-topic probabilities $\boldsymbol{\theta}_t$ estimated as described in Paragraph 3.2.4.

3.2.6 DLM for Topic Proportions

In this section, we directly model the time-topic distributions $\boldsymbol{\theta}_t$ estimated as in Paragraph 3.2.4. When investigating the evolution of the $\boldsymbol{\theta}_t$ series, we observed that values for some topics show as well signs of a linear trend. We thus choose a specification aligned to that of word-counts. We first map $\boldsymbol{\theta}_t$ values to real-valued $\boldsymbol{u}_t$ vectors via inverse soft-max transformation[5]. Vector $\boldsymbol{u}_t$ will then represent our input to the following dynamic linear model:

$$\begin{aligned} \boldsymbol{u_t} &= \boldsymbol{d} + \boldsymbol{\beta} t + \boldsymbol{l} f_t + \boldsymbol{\epsilon}_t & \boldsymbol{\epsilon}_t &\sim N(\boldsymbol{0}, \gamma_c I) \\ f_t &= f_{t-1} + \nu_t & \nu_t &\sim N(0, \gamma_f) \end{aligned}$$

We are thus fitting our model to a topic-specific intercept, a linear trend and a latent factor which is modeled as a random walk. Initialisation values are chosen with an analogous strategy to that for word-counts, but making use of estimated $\boldsymbol{u}_t$ as a dependent variable.

Figure 3.5 shows the estimated factor against the initial unemployment values around which it was initialised. Again, the factor converged to values that are within a similar range to that of unemployment, but its evolution is not fully aligned to unemployment. However, we can see that the extracted signal is not too far off from that extracted from word counts (Figure 3.6).

[5] $u_{k,t} = \log(\theta_{k,t}) - (\sum_k \log(\theta_{k,t}))/K$

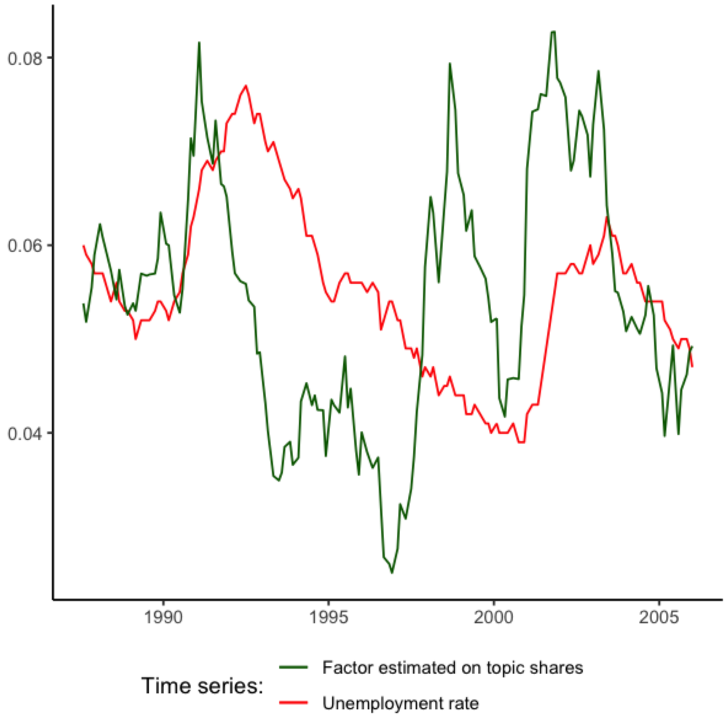


Figure 3.5: Factor estimated on the topic proportions time series versus unemployment rate

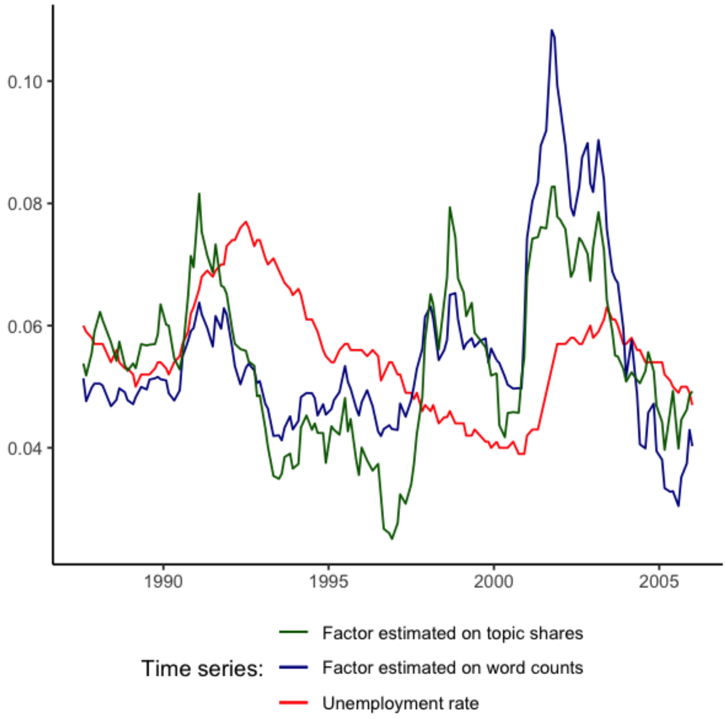


Figure 3.6: Comparison of latent factors estimated on different time series with unemployment rate

In Figure 3.7 we show fitted values against $\boldsymbol{u}_t$ time series per each topic. Looking at the $\boldsymbol{u}_t$ time series, we can see that now the meeting-length effect is no longer present, some topics (usually very general ones) that before showed to be trending upwards now are fairly steady (e.g. T24 - general recommendations, T25 - charts & projections, T31 - general pleasantries), pointing at the fact that the use of topics per se did not particularly increase over time for these topics. In turn, for other topics a linear upward trend can still be detected (e.g. T6 - employment & jobs, T17 - reports & surveys, T23 - growth forecasts), hinting that in these cases the upward trend was likely not only due to a general increase in meeting length, but also to an increase in the use of topic. However, we can still distinguish between topic series that are mainly fitted with the linear trend (e.g. T11 - monetary policy & rates, T23 - growth forecasts, T25 - charts & projections) and topic series for which the factor is majorly contributing and providing a better fit to the respective $\boldsymbol{u}_t$ series (e.g. T9 - economic recession & uncertainty, T33 - inflation expectations, T37 - risk & uncertainty, T18 - economic growth inflation). Interestingly, we see that a number of these overlap to those identified for word counts (e.g. T9 - economic recession & uncertainty, T33 - inflation expectations). It thus seems that both models roughly go in the same direction when trying to find a common low-dimensional signal for mildly trending topics.

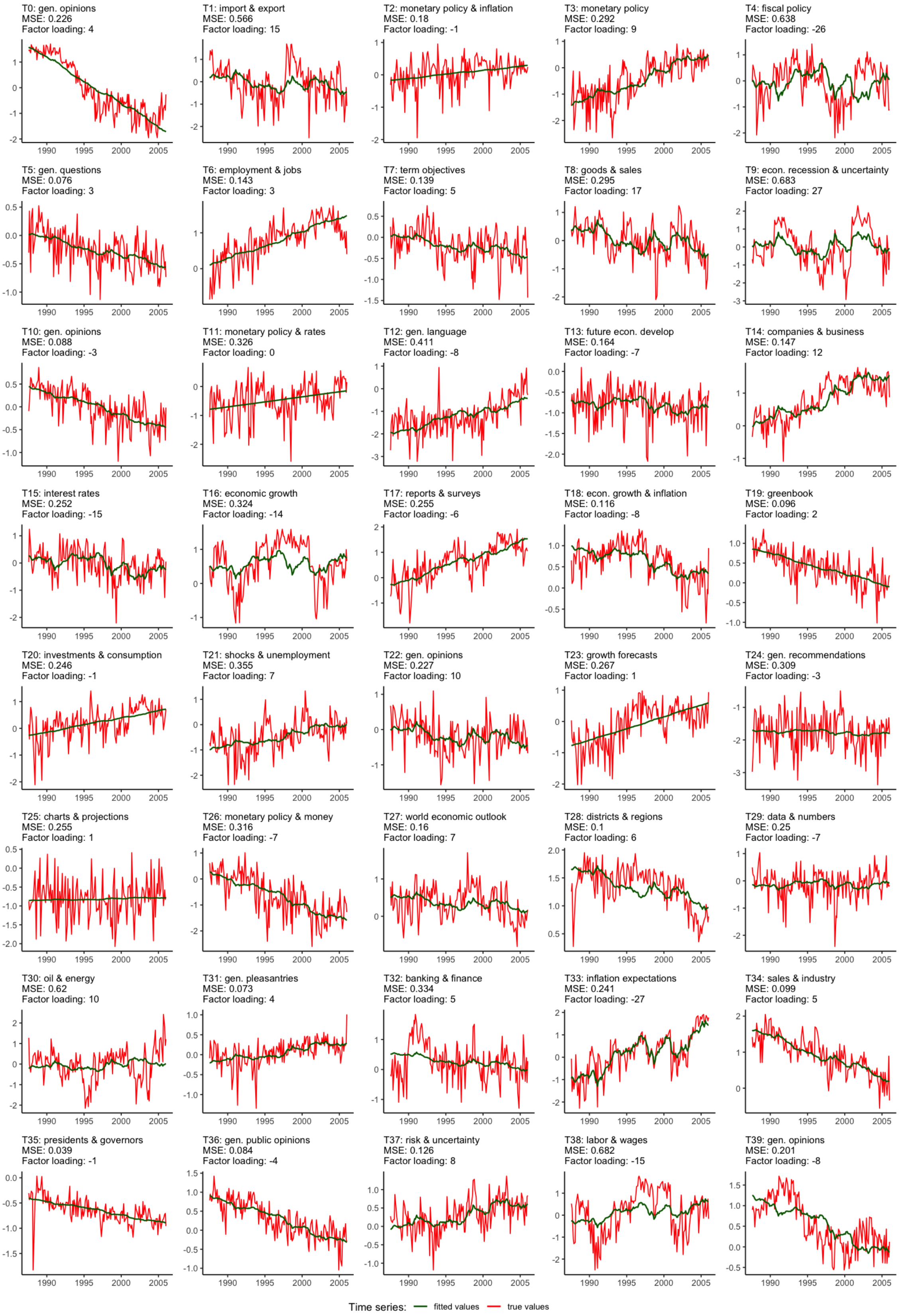


Figure 3.7: Fitted versus true values for topic proportions time series

3.2.7 Connection with Macroeconomic Uncertainty

Investigating the results produced by our models, we have seen how in both cases the latent factor focuses on a subset of specific topics and tries to retrieve a common low dimensional signal that can explain the variability in the observed time series. We are now interested in seeing whether this signal can be put in relation with some meaningful macroeconomic variable that might drive this common variability across topics.

In this regard, we observe that a few of the topics mainly fitted with the estimated latent factor are somewhat related to the concept of macroeconomic uncertainty and expectations. In particular, this is the case for T9 - economic recession & uncertainty, T13 - future economic development, T18 - economic growth & inflation, T29 - data & numbers, T33 - inflation expectations. It thus comes natural to compare the estimated factors with a measure for macroeconomic uncertainty. To this end, we use two measures approximating policy-related economic uncertainty: the US Economic Policy Uncertainty Index and the US News-Based Economic Policy Uncertainty Index (Baker et al. 2016). While the first index is calculated based on three main components: news coverage about policy-related economic uncertainty, tax code expiration data, and economic forecaster disagreement, the second one is based exclusively on information drawn from large newspapers[6]. None of these makes use of FOMC speech data. Comparisons for both models are shown in Figures 3.8 and 3.9. As we can see, in both cases the retrieved factor shows to be able to track the targeted variables quite faithfully.

We thus argue that there is an inherent state-of-the-world dynamic - that we here identify with macroeconomic uncertainty - that is driving FOMC discussions over a number of different but somewhat related topics such as economic recession and growth, inflation expectations, or future economic development. FOMC speeches can thus be used to extract a signal that resembles the one obtained when purely trying to measure uncertainty in the economy via a combination of numerical in-

[6] For more information see https://www.policyuncertainty.com/us_monthly.html

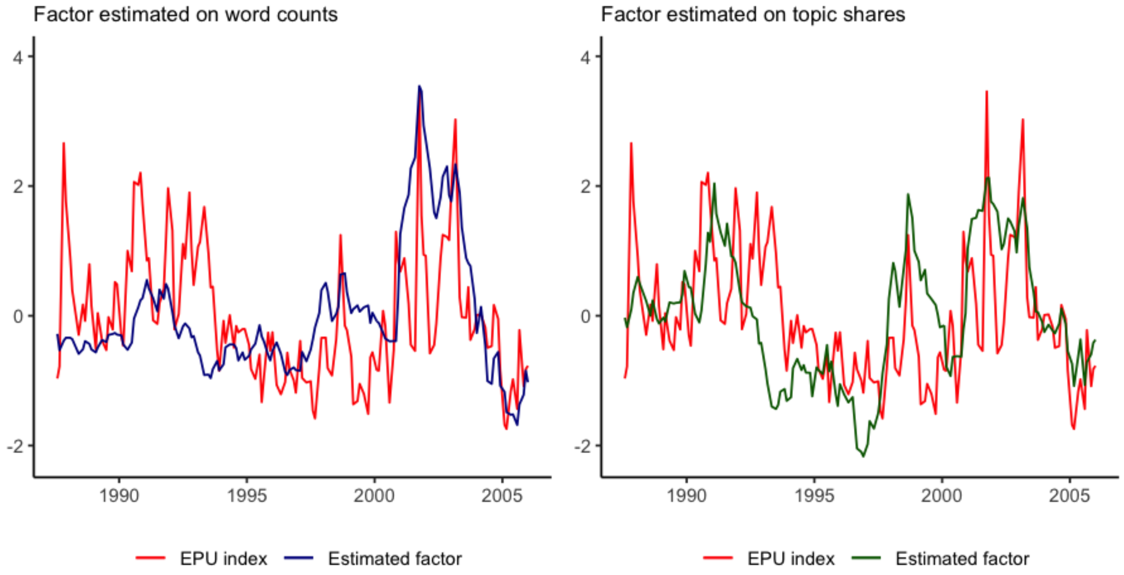


Figure 3.8: Comparison of latent factors estimated on different time series with the Economic Policy Uncertainty Index for United States[7]

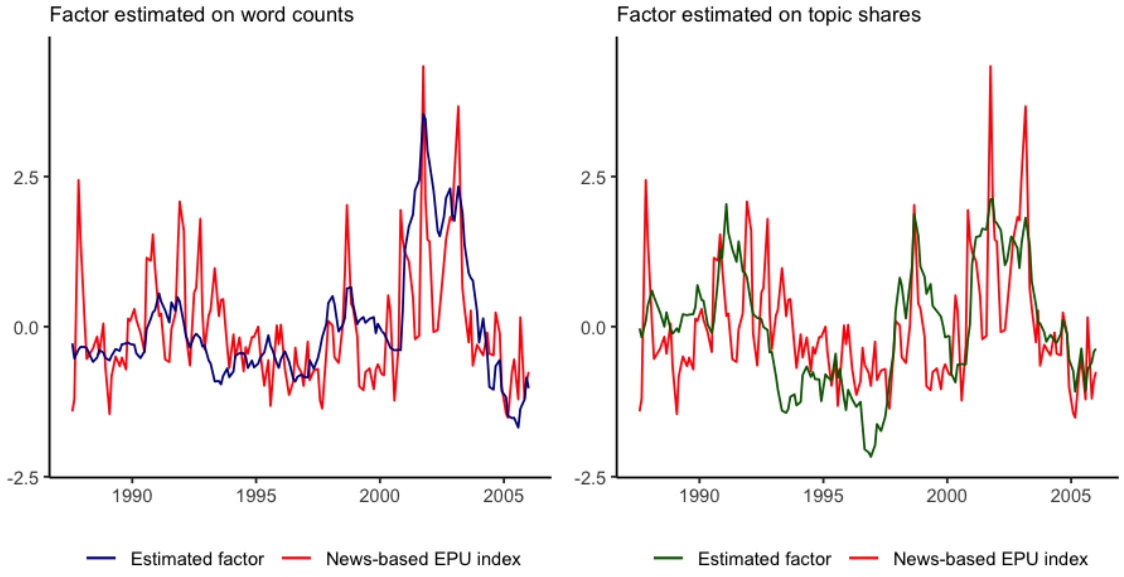


Figure 3.9: Comparison of latent factors estimated on different time series with the News-Based Economic Policy Uncertainty Index for United States[7]

[7] All time series are normalized.

dices and trends in news and reports. Building on this result, an interesting area for further exploration would thus be to investigate the ability of this signal to predict future uncertainty, and the opportunity to incorporate the latent factor into macroeconomic forecast or structural models.

3.3 Limitations and Extensions

Here we discuss the main limitations and potential extensions to our analysis.

First, we observe that when constructing the DLM for both our model specifications, we made the assumption that our observations follow a Multivariate Gaussian distribution with uncorrelated homoscedastic variances. Assuming uncorrelated time series with homoscedastic variance was preferred as a way to keep the estimation more manageable, but it could be relaxed. Moreover, we are treating word-counts in the first case and (mapped to reals) probabilities in the second case. A logical extension would thus be to modify the Gaussianity assumption and model observation equations with specifications which are more suitable to the type of data, for instance modeling counts with a Multinomial distribution or probabilities with a Logistic-Normal distribution. Else, one could model correlations across topic series so to better reflect these aspects.

Secondly, in our framework we apply a sequential approach where first we make use of LDA to retrieve an estimation for word-counts/topic shares, and then we use the LDA output as an input to our DLM for extracting a low-dimensional signal. This approach has the disadvantage that, at the time of word-counts/topic shares estimation, LDA does not take into account time dynamics anyhow and perceives the given data as a collection of documents exchangeable over time. Indeed, the two models, LDA and DLM, do not "communicate" between each other. This implies that during the evaluation of the time-evolving latent factor we do not transmit the retrieved information at every timestamp into the LDA, that thus does not dynamically update topic proportions accordingly. Treating the output as data also ignores the uncertainty in the estimate of the LDA low-dimensional space.

A suitable approach would be that of incorporating time dynamics into a holistic extension of LDA. In this regard, interesting extensions of LDA were proposed by Blei & Lafferty (2006) (Discrete Dynamic Topic Model, dDTM) and Wang et al. (2012) (Continuous Dynamic Topic Model, cDTM) in an attempt to relax the implicit assumption about exchangeability of documents in a collection (for more detail, see Appendix). A sophisticated approach explicitly integrating LDA with a state space model for topic proportions is that proposed by Glynn et al. (2019) under the name of Dynamic Linear Topic Model (DLTM). This model allows topic probabilities to exhibit a rich set of dynamic behaviors and incorporates document-specific covariates, such as author or publisher. The last aspect would also be of interest to our case, in that one could perform the analysis at the speaker-time level and foresee the inclusion of speaker-specific covariates. Besides, another important contribution made by authors of the DLTM is a development of a fully Bayesian posterior inference algorithm making use of a Gibbs sampler with Polya-Gamma data augmentation. Indeed, even keeping our sequential approach, different approaches might be considered for DLM estimation, including MCMC or variational inference methods.

Thirdly, one might be interest in tuning LDA hyperparameters, and in particular the number of topics K. As we chose to rely on preprocessed data obtained from the analyses of Hansen et al. (2018), this type of exploration fell out of the scope of this particular study. However, it would be an interesting extension for further investigation, and a general suggestion for approaching such types of analysis.

Chapter 4

Conclusions and Future Research

In this study, we suggest a sequential approach for the extraction of a low-dimensional signal from a collection of documents ordered over time. This approach foresees using Latent Dirichlet Allocation (LDA) for retrieving estimates for word-distributions over different topics and a representation of topics usage for a given set of documents. It then foresees modeling different representations of topics usage with a Dynamic Linear Model (DLM) that is able to capture the dynamic evolution of topics usage over time, and achieve a low-dimensional representation of the given time series into evolving latent factors that capture the driving dynamics in the data. We apply this framework to the US Fed's FOMC speech transcripts for the period 08-1986 to 01-2006. We retrieve estimates for a single latent factor, that seem to track fairly well a specific set of topics connected with risk, uncertainty, and expectations. Finally, we find a remarkable correspondence between this factor and the Economic Policy Uncertainty Indices for United States.

This exploratory research provides solid motivation for the investigation of the potential use of extracted low-dimensional signals from unstructured text data. For the specific case at hand, further research can be extended in several directions. First of all, one can consider exploring a more complex structure of a DLM, perhaps including more latent factors. This would imply not only an increase in the computational complexity but also the need for an in-depth analysis of signals interpretation in their combinations. Secondly, this approach can be improved by foreseeing a holistic and integrated approach between LDA and DLM. As it was discussed in the previous section, a DLTM is a sophisticated alternative, which resolves some limitations of our framework. Finally, this study can become a foundation for building macroeconomic forecasting models, or in general for using variation in language to estimate a model of learning about latent economic conditions.

List of Figures

List of Tables

Bibliography

Acosta, M. (2015), FOMC Responses to Calls for Transparency, Finance and economics discussion series, Board of Governors of the Federal Reserve System (U.S.).

Baker, S. R., Bloom, N. & Davis, S. J. (2016), 'Measuring Economic Policy Uncertainty', *The Quarterly Journal of Economics* **131**(4), 1593–1636.

Blei, D. & Lafferty, J. (2006), Dynamic topic models, *in* 'Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006', Vol. 148 of *ACM International Conference Proceeding Series*, pp. 113–120.

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

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

Chang, J. (2015), 'Collapsed gibbs sampling methods for topic models'.

Gentzkow, M., Kelly, B. & Taddy, M. (2019), 'Text as data', *Journal of Economic Literature* **57**(3), 535–74.

Glynn, C., Tokdar, S. T., Howard, B. & Banks, D. L. (2019), 'Bayesian analysis of dynamic linear topic models', *Bayesian Anal.* **14**(1), 53–80.

Griffiths, T. & Steyvers, M. (2004), 'Finding scientific topics', *Proceedings of the National Academy of Sciences of the United States of America* **101**(Suppl. 1), 5228–5235.

Hansen, S., McMahon, M. & Prat, A. (2018), 'Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach', *The Quarterly Journal of Economics* **133**(2), 801–870.

Harrison, P. J. & Stevens, C. F. (1976), 'Bayesian forecasting', *Journal of the Royal Statistical Society. Series B (Methodological)* **38**(3), 205–247.

IDC (2014), The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, Technical report, International Data Corporation.

Laine, M. (2020), *Introduction to Dynamic Linear Models for Time Series Analysis*, pp. 139–156.

Petris, G. (2010), 'An R package for dynamic linear models', *Journal of Statistical Software* **36**(12), 1–16.

Reinsel, D., Gantz, J. & Rydning, J. (2018), The Digitization of the World from Edge to Core, Technical report, International Data Corporation.

Schonhardt-Bailey, C. (2013), *Deliberating American monetary policy: a textual analysis*, MIT Press.

Wadhwani, P. & Kasnale, S. (2019), Text Analytics Market Research. Global Report 2019-2026, Technical report.

Wang, C., Blei, D. & Heckerman, D. (2012), 'Continuous time dynamic topic models'.

West, M. & Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models (2nd Ed.)*, Springer-Verlag, Berlin, Heidelberg.

Woolley, J. T. & Gardner, J. (2017), 'The effect of "sunshine" on policy deliberation: The case of the federal open market committee', *The Social Science Journal* **54**(1), 13–29.

Appendix A

Dynamic Topic Models

While LDA is an extremely powerful tool for inferring the topics' structure underlying a set of documents, one of its assumptions can be quite arguable for many practical applications. In fact, this topic model assumes the words of each document to be independently drawn from a mixture of Multinomials and does not consider the evolution of the topics or words use over time. In this section, we give a broad overview of two interesting extensions of LDA, which form the family of dynamic topic models (DTM), aimed at relaxing the implicit assumption about the exchangeability of documents in a collection.

A.1 Discrete-time DTM

A discrete-time dynamic topic model (dDTM) is a generative probabilistic model, developed to analyze the evolution of latent topics in a collection of documents over time (Blei & Lafferty 2006). It assumes that a collection of documents can be divided by time slice and that all of the K topics associated with slice t evolve from the topics associated with slice $t-1$. For sequential modeling of the words and topics proportions, dDTM uses random walk state-space models, evolving with a Gaussian noise. The generative process for slice t, chaining together topics and topic proportion distributions, can be written as follows:

1. For each topic $k = 1, \dots, K$

   (a) Draw topics $\boldsymbol{\beta}_{k,t} | \boldsymbol{\beta}_{k,t-1} \sim \mathcal{N}(\boldsymbol{\beta}_{k,t-1}, \sigma^2 \boldsymbol{I})$

2. Draw $\alpha_t | \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 \boldsymbol{I})$

3. For each document $d = 1, \ldots, D$

   (a) Draw $\boldsymbol{\theta}_{d,t} \sim \mathcal{N}(\alpha_t, a^2 \boldsymbol{I})$

   (b) For each word $n = 1, \ldots, N_d$

      i. Draw $z_{d,n,t} \sim \text{Multinomial}(\pi(\boldsymbol{\theta}_{d,t}))$

      ii. Draw $w_{d,n,t} \sim \text{Multinomial}(\pi(\boldsymbol{\beta}_{z_{d,n,t},t}))$

where π is a function mapping the multinomial natural parameters to the mean parameters, $\pi(\boldsymbol{\beta}_{k,t})_w = \frac{\exp(\boldsymbol{\beta}_{k,t,w})}{\sum_w \exp(\boldsymbol{\beta}_{k,t,w})}$ (Blei & Lafferty 2006). Graphical representation of the dDTM generative process is shown in Figure A.1.

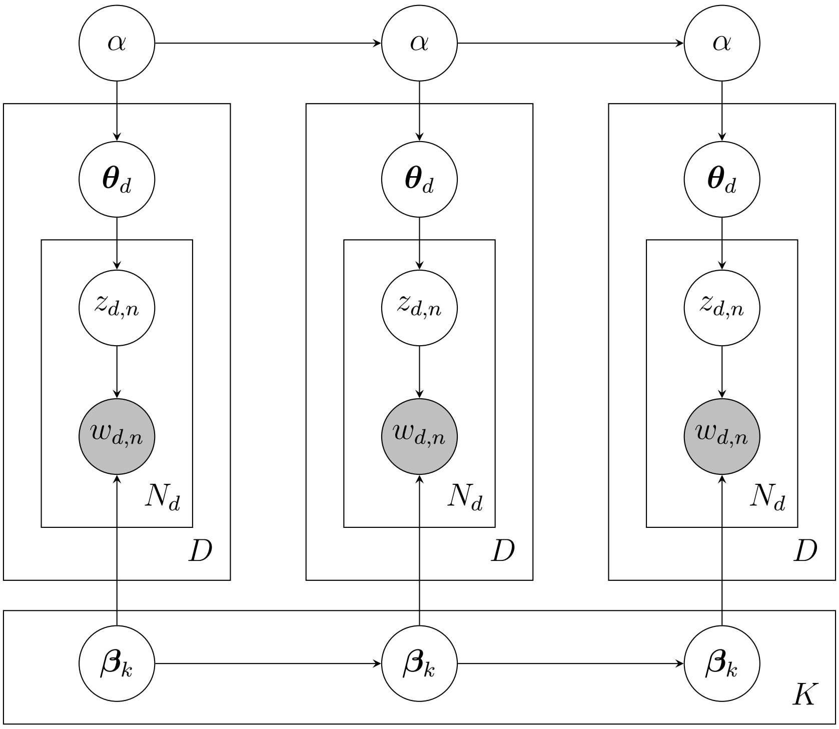


Figure A.1: Graphical model representation of a discrete-time dynamic topic model (Blei & Lafferty 2006)

A.2 Continuous DTM

A continuous dynamic topic model (cDTM) is a generative probabilistic model that uses Brownian motion to model the latent topics through a sequential collection of documents with arbitrary granularity (Wang et al. 2012). This model is a continuous version of the DTM which models the natural parameters with Brownian motion. This allows for more granular time discretization and eases the computation associated with finer time scales.

To define the generative process of cDTM, let us denote s_i and s_j as two arbitrary time stamps, and Δ_{s_i,s_j} – as the elapsed time between these time stamps. Then the data generative process can be introduced as follows:

1. For each topic $k = 1, \ldots, K$
   (a) Draw $\boldsymbol{\beta}_{k,0} \sim \mathcal{N}(\boldsymbol{m}, v_0\boldsymbol{I})$
2. For document d_t at time s_t $(t > 0)$
   (a) For each topic $k = 1, \ldots, K$
      i. From the Brownian motion model, draw $\boldsymbol{\beta}_{k,t}|\boldsymbol{\beta}_{k,t-1,s} \sim \mathcal{N}(\boldsymbol{\beta}_{k,t-1}, v\Delta_{s_t,s_{t-1}}\boldsymbol{I})$
   (b) Draw $\boldsymbol{\theta}_{d,t} \sim \text{Dirichlet}(\alpha)$
   (c) For each word $n = 1, \ldots, N_d$
      i. Draw $z_{d,n,t} \sim \text{Multinomial}(\boldsymbol{\theta}_{d,t})$
      ii. Draw $w_{d,n,t} \sim \text{Multinomial}(\pi(\boldsymbol{\beta}_{z_{d,n,t},t}))$

where π is a function mapping the multinomial natural parameters to the mean parameters, $\pi(\boldsymbol{\beta}_{k,t})_w = \frac{\exp{(\boldsymbol{\beta}_{k,t,w})}}{\sum_w \exp{(\boldsymbol{\beta}_{k,t,w})}}$ (Wang et al. 2012). Graphical representation of the cDTM generative process is illustrated in Figure A.2.

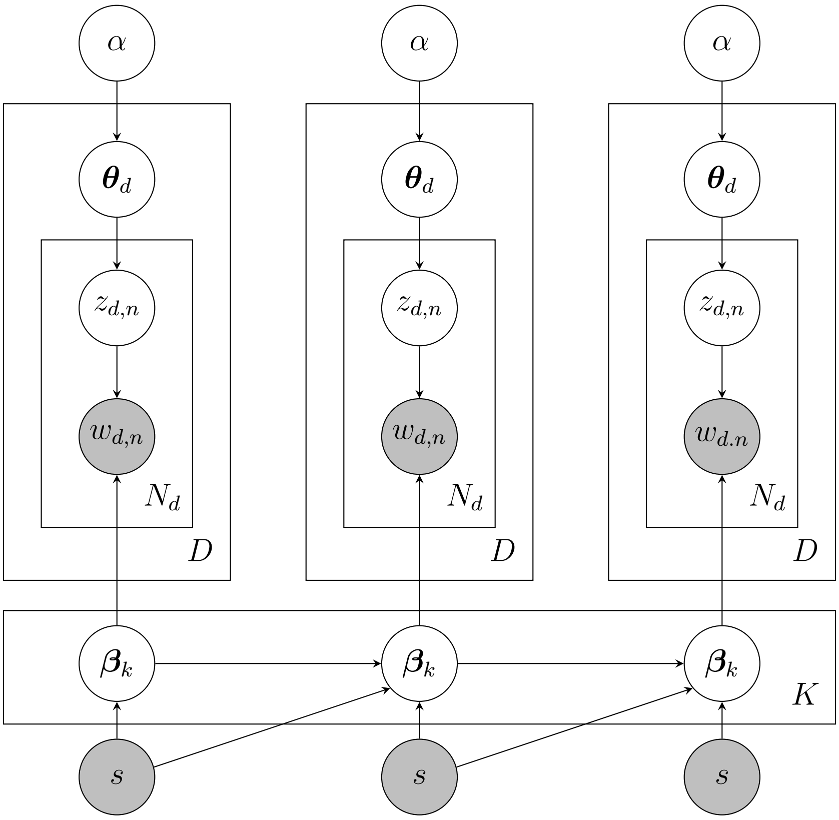


Figure A.2: Graphical model representation of a continuous dynamic topic model (Wang et al. 2012)