Thorsrud, Leif Anders

Working Paper

Words Are the New Numbers: A Newsy Coincident Index of Business Cycles

Working Paper, No. 21/2016

Provided in Cooperation with:
Norges Bank, Oslo

Suggested Citation: Thorsrud, Leif Anders (2016) : Words Are the New Numbers: A Newsy Coincident Index of Business Cycles, Working Paper, No. 21/2016, ISBN 978-82-7553-953-1, Norges Bank, Oslo,
https://hdl.handle.net/11250/2495606

This Version is available at:
https://hdl.handle.net/10419/210110

WORKING PAPER

Words are the new numbers: A newsy coincident index of business cycles

NORGES BANK RESEARCH

21 | 2016

LEIF ANDERS THORSRUD

NORGES BANK

Working papers fra Norges Bank, fra 1992/1 til 2009/2 kan bestilles over e-post:
FacilityServices@norges-bank.no

Fra 1999 og senere er publikasjonene tilgjengelige på www.norges-bank.no

Working papers inneholder forskningsarbeider og utredninger som vanligvis ikke har fått sin endelige form. Hensikten er blant annet at forfatteren kan motta kommentarer fra kolleger og andre interesserte. Synspunkter og konklusjoner i arbeidene står for forfatternes regning.

Working papers from Norges Bank, from 1992/1 to 2009/2 can be ordered by e-mail
FacilityServices@norges-bank.no

Working papers from 1999 onwards are available on www.norges-bank.no

Norges Bank's working papers present research projects and reports (not usually in their final form) and are intended inter alia to enable the author to benefit from the comments of colleagues and other interested parties. Views and conclusions expressed in working papers are the responsibility of the authors alone.

ISSN 1502-819-0 (online)
ISBN 978-82-7553-953-1 (online)

Words are the new numbers: A newsy coincident index of business cycles*

Leif Anders Thorsrud†

December 21, 2016

Abstract

I construct a daily business cycle index based on quarterly GDP and textual information contained in a daily business newspaper. The newspaper data are decomposed into time series representing newspaper topics using a Latent Dirichlet Allocation model. The business cycle index is estimated using the newspaper topics and a time-varying Dynamic Factor Model where dynamic sparsity is enforced upon the factor loadings using a latent threshold mechanism. The resulting index is shown to be not only more timely but also more accurate than commonly used alternative business cycle indicators. Moreover, the derived index provides the index user with broad based high frequent information about the type of news that drive or reflect economic fluctuations.

JEL-codes: C11, C32, E32

Keywords: Business cycles, Dynamic Factor Model, Latent Dirichlet Allocation (LDA)

*This Working Paper should not be reported as representing the views of Norges Bank. The views expressed are those of the authors and do not necessarily reflect those of Norges Bank. I thank Hilde C. Bjørnland, Fabio Canova, Pia Glæserud, Juan F. Rubio-Ramírez, Maximilian Rohrer, and Christian Schumacher for valuable comments. Vegard Larsen provided helpful technical assistance for which I am grateful. Comments from participants at the Joint Research Workshop of Norges Bank and Deutsche Bundesbank and the CAMP Workshop on Commodities, business cycles and monetary policy also helped improve the paper. This work is part of the research activities at the Centre for Applied Macro and Petroleum economics (CAMP) at the BI Norwegian Business School.

†Norges Bank and Centre for Applied Macro and Petroleum economics, BI Norwegian Business School. Email: leif.a.thorsrud@bi.no

1 Introduction

Policy makers and forecasters need to assess the state of the economy in real time to devise appropriate policy responses and condition on an updated information set. However, in real time, our main measure of economic activity, GDP growth, is not observed as it is compiled on a quarterly frequency and published with a considerable lag, usually up to at least one month. To mediate these caveats, various more timely indicators (like financial and labor market data) are monitored closely, and coincident indexes constructed.[1]

However, these common approaches face at least two drawbacks. First, the relationships between the timely indicators typically monitored, e.g., financial market data, and GDP growth are inherently unstable (see, e.g., Stock and Watson (2003)). Second, due to limited availability of high frequency data, the type of data from which coincident indexes often are constructed is constrained. As a result, changes in any coincident index constructed from such series do generally not give the index user broad information about what's leading to the changes in the index. For example, changes in financial returns can be observed daily and are commonly believed to be due to new information about future fundamentals, but the changes themselves do not reveal what this new information is. For policy makers in particular, as reflected in the broad coverage of various financial and macroeconomic data in monetary policy reports and national budgets, understanding why an index changes might be as important as the movement itself. Related to this, the indicators often used are typically obtained from structured databases and professional data providers. In contrast, the agents in the economy likely use a plethora of high-frequency information to guide their actions and thereby shape aggregate economic fluctuations. It is not a brave claim to assert that this information is highly unstructured and does not come (directly) from professional data providers, but more likely reflect information shared, generated, or filtered through a large range of channels, including media.

In this paper, I propose a new coincident index of business cycles aimed at addressing the drawbacks discussed above. In the tradition of Mariano and Murasawa (2003) and Aruoba et al. (2009), I estimate a latent daily coincident index using a Bayesian time-varying Dynamic Factor Model (DFM) mixing observed daily and quarterly data. To this, I make two contributions. First, the daily data set comes from a novel usage of textual information contained in a daily business newspaper, represented as topic frequencies across time. Thus, words are the new numbers, and the name: A newsy coincident index of

[1] Stock and Watson (1988) and Stock and Watson (1989) provide early examples of studies constructing coincident indexes using single frequency variables and latent factors, while Mariano and Murasawa (2003) extent this line of research to a mixed frequency environment using monthly and quarterly data. Later contributions mixing even higher frequency data, e.g., daily, with quarterly observations are given by, e.g., Evans (2005) and Aruoba et al. (2009).

business cycles (*NCI*). In turn, this innovation allows for decomposing the changes in the latent daily business cycle index into the (time-varying) news components it constitutes, and therefore also say something more broadly about why (in terms of news topics) the index changes at particular points in time. My hypothesis is simple: To the extent that the newspaper provides a relevant description of the economy, the more intensive a given topic is represented in the newspaper at a given point in time, the more likely it is that this topic represents something of importance for the economy's current and future needs and developments. Instead of relying on a limited set of conventional high frequency indicators to measure changes in business cycle conditions, I use a primary source for new broad based information directly - the newspaper.[2]

Second, building on the Latent Threshold Model (LTM) idea introduced by Nakajima and West (2013), and applied in a factor model setting in Zhou et al. (2014), the DFM is specified using an explicit threshold mechanism for the time-varying factor loadings. This enforces sparsity on the system, but also explicitly takes into account that the relationship between the latent daily business cycle index and the indicators used to derive it might be unstable (irrespective of whether newspaper data or more standard high frequent data is used to derive the index).

My main results show that both innovations listed above are important. I demonstrate, using Receiver Operating Characteristic (ROC) curves, that compared to more traditional business cycle indicators and coincident indexes, the *NCI* provides a more timely and trustworthy signal about the state of the economy. This gain is achieved through the combined usage of newspaper data and allowing for time-variation in the factor loadings. Moreover, the *NCI* contains important leading information, suggesting that the *NCI* would be a highly useful indicator for turning point predictions and nowcasting. Decomposing the *NCI* into the individual news topic contributions it constitutes reveals that on average, across different business cycle phases, news topics related to monetary and fiscal policy, the stock market and credit, and industry specific sectors seem to provide the most important information about business cycle conditions. Finally, the sign and timing of their individual contributions map well with the historical narrative we have about recent business cycle phases.

In using newspaper data the approach taken here shares many features with a growing number of studies using textual information to predict and explain economic outcomes, but extends this line of research it into the realm of coincident index construction. For

[2]Economic theory suggests that news might be important for explaining economic fluctuations because it contains new fundamental information about the future (see, e.g., Beaudry and Portier (2014)). Alternatively, as in, e.g., Angeletos and La'O (2013), news is interpreted as some sort of propagation channel for sentiment. Results reported in Larsen and Thorsrud (2015) indicate that information in the newspaper, represented as topic frequencies, contain new fundamental information about the future.

example, Tetlock (2007) classifies textual information using negative and positive word counts, and links the derived time series to developments in the financial market; Baker et al. (2013) construct an uncertainty index based on the occurrence of words in newspapers associated with uncertainty and link it to policy-related economic uncertainty; Choi and Varian (2012) use Google Trends and search for specific categories to construct predictors for present developments in a wide range of economic variables.[3]

In this paper, textual information is utilized using a Latent Dirichlet Allocation (LDA) model. The LDA model statistically categorizes the corpus, i.e., the whole collection of words and articles, into topics that best reflect the corpus's word dependencies. A vast information set consisting of words and articles can thereby be summarized in a much smaller set of topics facilitating interpretation and usage in a time-series context.[4] Compared with existing textual approaches, the LDA approach offers several advantages. In terms of word counting, which words are positive and which negative obviously relates to an outcome. A topic does not. A topic has content in its own right. Moreover, the LDA is an automated machine learning algorithm, so (subjectively) choosing the words or specific categories to search for is not needed. Instead, the LDA automatically delivers topics that best describe the whole corpus. This permits us to examine if textual information in the newspaper is representative for economic fluctuations, and if so, identify the type of new information (in terms of topics) that might drive or reflect economic fluctuations. In Larsen and Thorsrud (2015), it is shown that individual news topics extracted using a LDA model adds marginal predictive power for a large range of economic aggregates at a quarterly frequency. Here I build on this knowledge and use similar topics to construct the daily *NCI*.

The perhaps most closely related paper to this is Balke et al. (2015). They use customized text analytics to decompose the Beige Book, a written description of economic conditions in each of the twelve districts banks of the Federal Reserve System in the U.S., into time series and construct a coincident index for the U.S. business cycle. They find that this textual data source contains information about current economic activity not contained in quantitative data. Their results are encouraging and complement

[3] Bloom (2014) provides a summary of the literature which constructs aggregate uncertainty indexes based on (among other things) counting pre-specified words in newspapers. See Tetlock (2014) for a short overview of the usage of textual data in the finance literature. In macroeconomics, there is a growing literature utilizing textual data to examine the effects of central bank's communication (see, e.g., Apel and Blix Grimaldi (2012) and the references therein).

[4] Blei et al. (2003) introduced the LDA as a natural language processing tool. Since then the methodology has been heavily applied in the machine learning literature and for textual analysis. Surprisingly, in economics, it has hardly been applied. See, e.g., Hansen et al. (2014), Hansen and McMahon (2015), and Larsen and Thorsrud (2015) for exceptions.

my findings. However, the Beige Book is published at an irregular frequency, and not all countries have Beige Book-type information. In contrast, most countries have publicly available newspapers published (potentially) daily.[5] Finally, as alluded to above, in contrast to existing studies using textual data, with the news topic approach one can decompose the daily changes in the coincident index into news topic contributions.

The rest of this paper is organized as follows. Section 2 describes the newspaper data, the topic model, and the estimated news topics. The mixed frequency and time-varying DFM is described in Section 3. Results are presented in Section 4. Section 5 concludes.

2 Data

The raw data used in this analysis consists of a long sample of the entire newspaper corpus for a daily business newspaper and quarterly GDP growth for Norway. I focus on Norway because it is a small and open economy and thereby representative of many western countries, and because small economies, like Norway, typically have only one or two business newspapers, making the choice of corpus less complicated. Here, I simply choose the corpus associated with the largest and most read business newspaper, *Dagens Næringsliv* (DN), noting that DN is also the fourth largest newspaper in Norway irrespective of subject matter. DN was founded in 1889, and has a right-wing and neo-liberal political stance. Importantly, however, the methodology for extracting news from newspaper data, and analyze whether or not it is informative about business cycle developments, is general and dependent neither on the country nor newspaper used for the empirical application.

To make the textual data applicable for time series analysis, the data is first decomposed into time series of news topics using a Latent Dirichlet Allocation (LDA) model. In general, topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them and the themes' connection to one another. Although topic models are well known, and have been massively applied, in the machine learning literature, their usage in the field of economics has been rare.

Blei (2012) provides a nice layman's introduction to topic modeling. The newspaper corpus and the LDA specification in this paper is similar to that described in Larsen and Thorsrud (2015). Still, as the usage of textual data and the application of a LDA model are relatively new in economics, I provide a summary of the computations below. I then

[5] In relation to this, the U.S. is in many aspects a special case when it comes to quantitatively available economic data, simply because there is so much available at a wide variety of frequencies. For most other countries, this is not the case. The usage of daily newspaper data can potentially mitigate such missing information.

examine the mapping between the estimated news topics and GDP growth using simple principal components analysis, before presenting the proposed time-varying and mixed frequency Dynamic Factor Model (DFM) in the subsequent section.

2.1 The news corpus, the LDA and topics

The DN news corpus is extracted from Retriever's "Atekst" database, and covers all articles published in DN from May 2, 1988, to December 29, 2014. In total this amounts to $N^a = 459745$ articles, well above one billion words, more than a million unique tokens, and a sample of $T^d = 9741$ days. This massive amount of data makes statistical computations challenging, but as is customary in this branch of the literature, some steps are taken to clean and reduce the raw dataset before estimation. A description of how this is done is given in Appendix C. I note here that around 250 000 unique tokens are kept after the filtering procedure.

The "cleaned", but still unstructured, DN corpus is decomposed into news topics using a Latent Dirichlet Allocation (LDA) model. The LDA model is an unsupervised topic model introduced by Blei et al. (2003) that clusters words into topics, which are distributions over words, while at the same time classifying articles as mixtures of topics.[6] By unsupervised learning algorithm we mean an algorithm that can learn/discover an underlying structure in the data without the algorithm being given any labeled samples to learn from. The term "latent" is used, because the words, which are the observed data, are intended to communicate a latent structure, namely the meaning of the article. The term "Dirichlet" is used because the topic mixture is drawn from a conjugate Dirichlet prior.

Figure 1 illustrates the LDA model graphically. The outer box, or plate, represents the whole corpus as M distinct documents (articles). $N = \sum_{m=1}^{M} N_m$ is the total number of words in all documents, and K is the total number of latent topics. Letting bold-font variables denote the vector version of the variables, the distribution of topics for a document is given by $\boldsymbol{\theta}_m$, while the distribution of words for each topic is determined by $\boldsymbol{\varphi}_k$. Both $\boldsymbol{\theta}_m$ and $\boldsymbol{\varphi}_k$ are assumed to have conjugate Dirichlet distributions with (hyper) parameter (vectors) α and β, respectively. Each document consists of a repeated choice of topics $Z_{m,n}$ and words $W_{m,n}$, drawn from the Multinomial distribution using $\boldsymbol{\theta}_m$ and $\boldsymbol{\varphi}_k$. The circle associated with $W_{m,n}$ is gray colored, indicating that these are the only observable variables in the model.

At an intuitive level, the best way to understand the LDA model is likely to make a thought experiment of how the articles in the newspaper (the corpus) were generated.

[6] This latter point is important, because it distinguishes the LDA model from other often used text classifying algorithms where each article is assumed to be described by only one single topic.

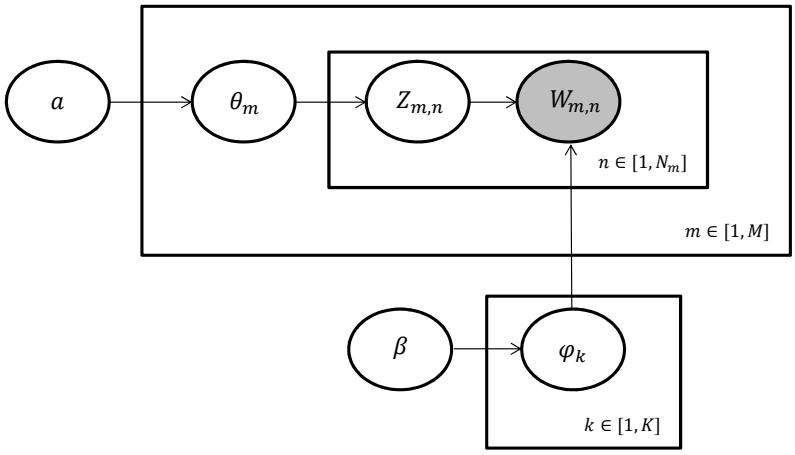

Figure 1. The LDA model visualized using plate notation.

1. Pick the overall theme of articles by randomly giving them a distribution over topics, i.e.: Choose $\boldsymbol{\theta}_m \sim \text{Dir}(\alpha)$, where $m \in \{1, \ldots, M\}$.
2. Pick the word distribution for each topic by giving them a distribution over words, i.e.: Choose $\boldsymbol{\varphi}_k \sim \text{Dir}(\beta)$, where $k \in \{1, \ldots, K\}$.
3. For each of the word positions m, n, where $n \in \{1, \ldots, N_m\}$, and $m \in \{1, \ldots, M\}$
   3.1. From the topic distribution chosen in 1., randomly pick one topic, i.e.: Choose a topic $Z_{m,n} \sim \text{Multinomial}(\boldsymbol{\theta}_m)$.
   3.2. Given that topic, randomly choose a word from this topic, i.e.: Choose a word $W_{m,n} \sim \text{Multinomial}(\boldsymbol{\varphi}_{z_{m,n}})$.

More formally, the total probability of a document, i.e., the joint distribution of all known and hidden variables given the hyper-parameters, is:

$$P(\boldsymbol{W}_m, \boldsymbol{Z}_m, \boldsymbol{\theta}_m, \boldsymbol{\Phi}; \alpha, \beta) = \overbrace{\underbrace{\prod_{n=1}^{N_m} P(W_{m,n}|\boldsymbol{\varphi}_{z_{m,n}}) P(Z_{m,n}|\boldsymbol{\theta}_m)}_{\text{word plate}} \cdot P(\boldsymbol{\theta}_m; \alpha)}^{\text{document plate (1 document)}} \cdot \underbrace{P(\boldsymbol{\Phi}; \beta)}_{\text{topic plate}} \quad (1)$$

where $\boldsymbol{\Phi} = \{\boldsymbol{\varphi}_k\}_{k=1}^K$ is a $(K \times V)$ matrix, and V is the size of the vocabulary. The two first factors in (1) correspond to the word plate in Figure 1, the three first factors to the document plate, and the last factor to the topic plate. Different solution algorithms exist for solving the LDA model. I follow Griffiths and Steyvers (2004), and do not treat $\boldsymbol{\theta}_m$ and $\boldsymbol{\varphi}_k$ as parameters to be estimated, but instead integrate them out of (1). Considering the corpus as a whole, this results in an expression for $P(\boldsymbol{W}, \boldsymbol{Z}; \alpha, \beta) = P(\boldsymbol{Z}|\boldsymbol{W}; \alpha, \beta) P(\boldsymbol{W}; \alpha, \beta)$ which can be solved using Gibbs simulations. Estimates of $\boldsymbol{\theta}_m$ and $\boldsymbol{\varphi}_k$ can subsequently be obtained from the posterior distribution. Further technical details, and a short description of estimation and prior specification, are described in Appendix D.

The model is estimated using 7500×10 draws. The first 15000 draws of the sampler are disregarded, and only every 10th draw of the remaining simulations are recorded and used for inference. $K = 80$ topics are classified. Marginal likelihood comparisons across LDA models estimated using smaller numbers of topics (see Larsen and Thorsrud (2015)), indicate that 80 topics provide the best statistical decomposition of the DN corpus.

Now, the LDA estimation procedure does not give the topics any name or label. To do so, labels are subjectively given to each topic based on the most important words associated with each topic. As shown in Table 3 in Appendix A, which lists all the estimated topics together with the most important words associated with each topic, it is, in most cases, conceptually simple to classify them. I note, however, that the labeling plays no material role in the experiment, it just serves as a convenient way of referring to the different topics (instead of using, e.g., topic numbers or long lists of words). What is more interesting, however, is whether the LDA decomposition gives a meaningful and easily interpretable topic classification of the DN newspaper. As illustrated in Figure 2, it does: The topic decomposition reflects how DN structures its content, with distinct sections for particular themes, and that DN is a Norwegian newspaper writing about news of particular relevance for Norway. We observe, for example, separate topics for Norway's immediate Nordic neighbors (*Nordic countries*); largest trading partners (*EU* and *Europe*); and biggest and second biggest exports (*Oil production* and *Fishing*). A richer discussion about a similar decomposition is provided in Larsen and Thorsrud (2015).

2.2 News Topics as time series

Given knowledge of the topics (and their distributions), the topic decompositions are translated into time series. To do this, I proceed in three steps.

Step 1. For each day, the frequency with which each topic is represented in the newspaper that day is calculated. This is done by collapsing all the articles in the newspaper for a particular day into one document, and then computing, using the estimated word distribution for each topic, the topic frequencies for this newly formed document. See Appendix D.1 for details. By construction, across all topics, this number will sum to one for any given day. On average, across the whole sample, each topic will have a more or less equal probability of being represented in the newspaper. Across shorter time periods, i.e., days, the variation can be substantial. I define this as the D_0 data set, which will be a $T^d \times K$ matrix.[7]

[7]Since DN is not published on Sundays, but economic activity also takes place on Sundays, missing observations are filled by simple linear interpolation. Note also that the construction described in *Step 1* does not mean that only one topic is used as representative for a given day. For such an assumption mixture models other than the LDA would have been more appropriate.

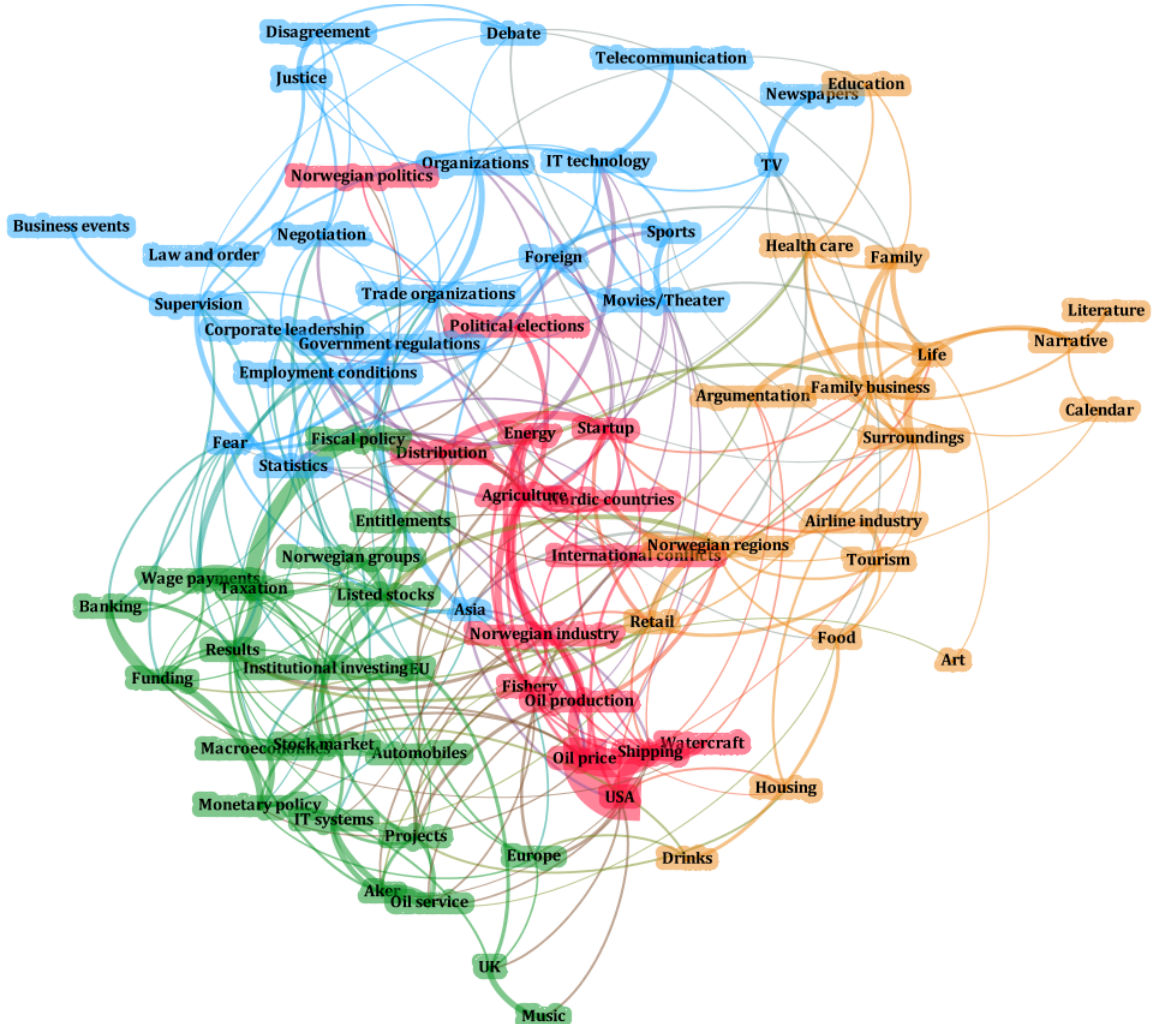

Figure 2. A network representation of the estimated news topics. The nodes in the graph represent the identified topics. All the edges represent words that are common to the topics they connect. The thickness of the edges represents the importance of the word that connect the topics, calculated as edge weight = 1/ (ranking of word in second topic + ranking of word in first topic). The topics with the same color are clustered together using a community detection algorithm called Louvain modularity. Topics for which labeling is *Unknown*, c.f. Table 3 in Appendix A, are removed from the graph for visual clarity.

Step 2. Since the time series objects constructed in *Step 1* will be intensity measures, i.e., reflecting how much DN writes about a given topic at a specific point in time, their tone is not identified. That is, whether the news is positive or negative. To mediate this, a sign-identified data set based on the number of positive relative to negative words in the text is constructed. In particular, for each day t, all N_t^a newspaper articles that day, and each news topic in D_0, the article that news topic k describes the best is found. Given knowledge of this topic article mapping, positive/negative words in the articles are identified using an external word list and simple word counts. The word list used here takes as a starting point the classification of positive/negative words defined by the *Harvard IV-4 Psychological Dictionary*. As this dictionary contains English words only, it must be translated into Norwegian. The translated set of words consists of 40 positive

and 39 negative Norwegian words, which is somewhat different from the *Harvard IV-4 Psychological Dictionary* both in terms of numbers and exact meaning.[8]

The count procedure delivers two statistics for each article, containing the number of positive and negative words. These statistics are then normalized such that each article observation reflects the fraction of positive and negative words, i.e.:

$$Pos_{t,n^a} = \frac{\#positivewords}{\#totalwords} \quad Neg_{t,n^a} = \frac{\#negativewords}{\#totalwords} \tag{2}$$

The overall mood of article n^a, for $n^a = 1, \ldots, N_t^a$ at day t, is defined as:

$$S_{t,n^a} = Pos_{t,n^a} - Neg_{t,n^a} \tag{3}$$

Using the S_{t,n^a} statistic and the topic article mapping described above, the sign of each topic in D_0 is adjusted accordingly as:

$$D_{t,1} = S_{t,n^a} D_{t,\tilde{k},0}$$

where $\tilde{k}$ reflects that article n^a is mapped to topic k.

Step 3. To remove daily noise from the topic time series in the D_1 data set, each topic time series is filtered using a 60 day (backward looking) moving average filter. As is common in factor model studies, see, e.g., Stock and Watson (2012), I also eliminate very low frequency variation, i.e., changes in the local mean, by removing a simple linear trend and standardize the data. For future reference I label this data set D_1^f.[9]

Figure 3 reports six of the topic time series, and illustrates how the different steps described above affect the data. The gray bars show the data as topic frequencies across time, i.e., as constructed in *Step 1* above. As is clearly visible in the graphs, these measures are very noisy. Applying the transformations described in *Step 2* and *Step 3* changes the intensity measures into sign identified measures and removes much of the most high frequent movements in the series. As seen from the figures, the differences between the D_0 and D_1^f measures are sometimes substantial, highlighting the influence of the sign identification procedure. The effect on the *Monetary Policy* topic is particular clear. From

[8]The translated word list can be obtained upon request. Counting the number of positive and negative words in a given text using the *Harvard IV-4 Psychological Dictionary* is a standard methodology in this branch of the literature (see, e.g., Tetlock et al. (2008)). In finance, Loughran and Mcdonald (2011) among others, show that word lists developed for other disciplines mis classify common words in financial texts, and suggest an alternative (English language) list. I leave it for future research to investigate if this also holds for macroeconomic applications and languages other than English.

[9]The estimated *NCI*, see Section 4, becomes more (less) noisy if a shorter (longer) window size is used to smooth the news topics (for similar prior specifications), but the overall cyclical pattern remains the same. I have also experimented using other word count and topic article mappings to construct the D_1 data set (in *Step 2*), observing that the methodology described above works best. Details about these alternative transformations may be obtained on request.

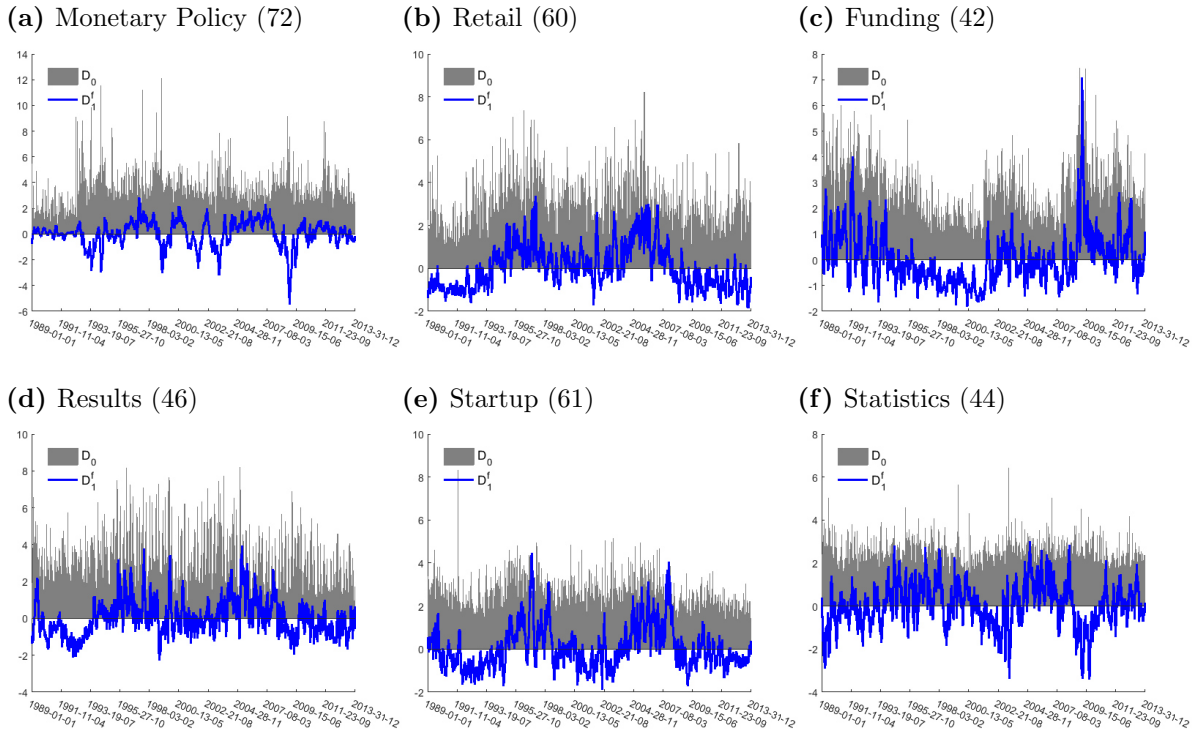

Figure 3. Individual news topics (topic numbers, confer Table 3 in Appendix A, in parenthesis). The grey bars and blue lines report topic time series from the D_0, and D_1^f data sets, respectively. See the text for details.

Figure 3 we also observe that topics covary, at least periodically. The maximum (minimum) correlation across all topics is 0.57 (-0.40) using the D_1^f data set. However, overall, the average absolute value of the correlation among the topics is just 0.1, suggesting that different topics are given different weight in the DN corpus across time.

2.3 GDP and news

Gross Domestic Product for mainland Norway, measured in constant 2012 prices (million NOK), is obtained from Statistics Norway (SSB).[10] The raw series is transformed to quarterly growth rates. Likewise to above, the local mean of the growth rates is removed using a linear time trend, and the series is standardized prior to estimation. In the rest of this paper the raw quarterly growth rates will be referred to as GDP, and the adjusted version, used for estimation, as GDP^a.

How do the news topics relate to GDP^a? To get a first pass impression I compute the first principal component of the sign identified data set, D_1^f, using either all 80 topics ($PCA1$), or only the 20 topics most correlated with linearly interpolated daily GDP^a ($PCA2$) (see Table 3 in Appendix A). These single common components explain only

[10] In Norway, using GDP excluding the petroleum sector is the commonly used measure of economic activity. I follow suit here because it facilitates the formal evaluation of the NCI in Section 4.

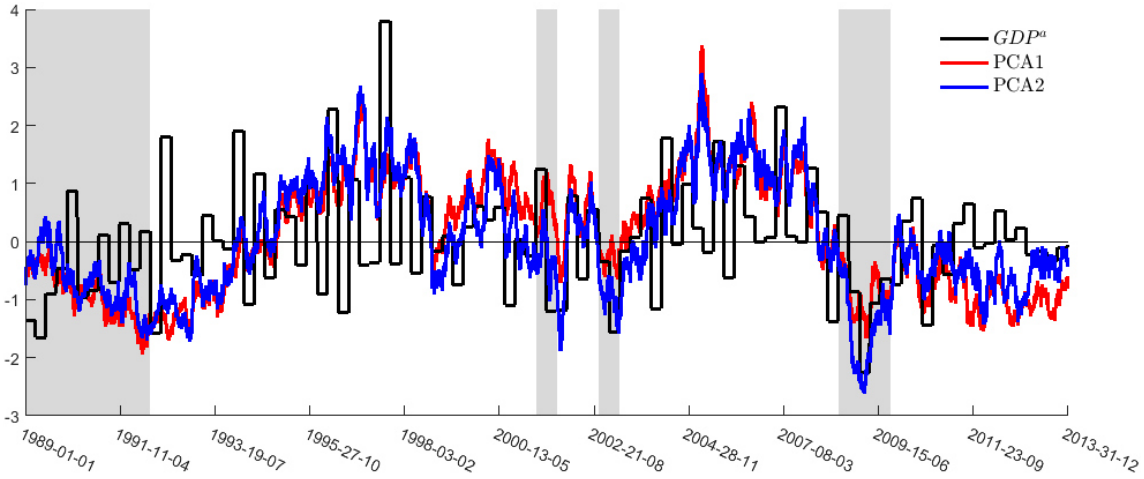

Figure 4. GDP^a is recorded at the end of each quarter, but reported on a daily basis in the graph using previous end-of-period values throughout the subsequent quarter. The red and blue lines are the first principal component estimate of the D_1^f data set using 80 and 20 topics, respectively. Recession periods, defined by a *MS-FMQ* model, see Section 4.1, are illustrated using gray color shading.

roughly 12 and 27 percent, respectively, of the overall variation in the data set, but seems to capture important business cycle fluctuations surprisingly well (see Figure 4). However, the factors derived from the simple PCA analysis do not seem to move in tandem with output during the early and later part of the sample. In addition, they are far from able to track the more high frequency movements in output. Having said this, it is still interesting that an unsupervised LDA and PCA decomposition of a business newspaper provides information about GDP^a in the manner reported here. It is not only a novel finding in itself, but also motivates the usage of a more supervised factor model using this type of data. I turn to this next.

3 The Dynamic Factor Model

To estimate a coincident index of business cycles utilizing the joint informational content in quarterly output growth and daily news topic series, I build on Mariano and Murasawa (2003) and Aruoba et al. (2009) and develop a mixed frequency time-varying Dynamic Factor Model (DFM).

Measured at the highest frequency among the set of mixed frequency observables, which is daily in this analysis, the DFM can be written as:

$$\boldsymbol{y}_t = \boldsymbol{z}_{0,t}\boldsymbol{a}_t + \cdots + \boldsymbol{z}_{s,t}\boldsymbol{a}_{t-s} + \boldsymbol{e}_t \tag{4a}$$

$$\boldsymbol{a}_t = \boldsymbol{F}_1\boldsymbol{a}_{t-1} + \cdots + \boldsymbol{F}_h\boldsymbol{a}_{t-h} + \boldsymbol{\omega}_t \tag{4b}$$

$$\boldsymbol{e}_t = \boldsymbol{P}_1\boldsymbol{e}_{t-1} + \cdots + \boldsymbol{P}_p\boldsymbol{e}_{t-p} + \boldsymbol{u}_t \tag{4c}$$

Equation (4a) is the observation equation of the system. $\boldsymbol{y}_t$ is a $N \times 1$ vector of observable

and unobservable variables assumed to be stationary with zero mean, decomposed as follows:

$$\boldsymbol{y}_t = \begin{pmatrix} \boldsymbol{y}^*_{1,t} \\ \boldsymbol{y}_{2,t} \end{pmatrix} \tag{5}$$

where $\boldsymbol{y}^*_{1,t}$ is a $N_q \times 1$ vector of unobserved daily output growth rates, mapping into quarterly output growth rates as explained below, and $\boldsymbol{y}_{2,t}$ is a $N_d \times 1$ vector of daily newspaper topic variables, described in Section 2.2. $N = N_q + N_d$, and $\boldsymbol{z}_{j,t}$ is a $N \times q$ matrix with dynamic factor loadings for $j = 0, 1, \cdots, s$, and s denotes the number of lags used for the dynamic factors $\boldsymbol{a}_t$. The dynamic factors, containing the daily business cycle index, follow a VAR(h) process given by the transition equation in (4b), where $\boldsymbol{\omega}_t \sim i.i.d.N(0, \boldsymbol{\Omega})$. Finally, equation (4c) describes the time series process for the $N \times 1$ vector of idiosyncratic errors $\boldsymbol{e}_t$. It is assumed that these evolve as independent AR(p) processes with $\boldsymbol{u}_t \sim i.i.d.N(0, \boldsymbol{U})$, and that $\boldsymbol{u}_t$ and $\boldsymbol{\omega}_t$ are independent. The model's only time-varying parameters are the factor loadings ($\boldsymbol{z}_{j,t}$), which are restricted to follow independent random walk processes.

Apart from the usage of newspaper data, the DFM described above is fairly standard. Similar specifications have been applied in recent work by Lopes and Carvalho (2007), Del Negro and Otrok (2008), Ellis et al. (2014), and Bjørnland and Thorsrud (2015). Some of these studies also include stochastic volatility in the DFM. In a mixed frequency setting for example, Marcellino et al. (2013) estimate a DFM (using monthly and quarterly data) without time-varying parameters, but with stochastic volatility. I abstract from this property here to focus on the innovations introduced in this paper.

Two extensions are applied here: First, sparsity is enforced on the system through the time-varying factor loadings using a latent threshold mechanism. Second, since the variables in the $\boldsymbol{y}_t$ vector are observed at different frequency intervals, cumulator variables are used to ensure consistency in the aggregation from higher to lower frequencies and make estimation feasible. Below I elaborate on these two extensions. A full description of the model, and its extensions, is given in Appendix E.[11]

[11] It follows from the above discussion that there is a conceptually close resemblance between the LDA model described in Section 2.1, and factor models commonly used in economics. In both instances, some set of observed variables are assumed to be determined by a (predefined) number of common latent variables. As such, one could envision a model where the observables, words and output growth in terms of this analysis, and their relationship to latent factors where estimated jointly within one model. I am, however, not aware of existing models in the literature that combine time series with textual data in this manner. Incorporating the mixed frequency and latent threshold dynamics into such model would complicate the problem further. Thus, as the first investigation of this sort, I opt for the simpler two-step approach in this analysis.

3.1 Enforcing sparsity and identification

Following the Latent Threshold Model (LTM) idea introduced by Nakajima and West (2013), and applied in a DFM setting in Zhou et al. (2014), sparsity is enforced on the system through the time-varying factor loadings using a latent threshold. For example, for one particular element in $\boldsymbol{z}_{0,t}$, $z_{i,0,t}$, the LTM structure can be written as:

$$z_{i,0,t} = z^*_{i,0,t}\varsigma_{i,0,t} \quad \varsigma_{i,0,t} = I(|z^*_{i,0,t}| \geq d_{i,0}) \tag{6}$$

where

$$z^*_{i,0,t} = z^*_{i,0,t-1} + w_{i,0,t} \tag{7}$$

with $w_{i,0,t} \sim i.i.d.N(0, \sigma^2_{i,0,w})$. In (6) $\varsigma_{i,0,t}$ is a zero one variable, whose value depends on the indicator function $I(|z^*_{i,0,t}| \geq d_{i,0})$. If $|z^*_{i,0,t}|$ is above the the threshold value $d_{i,0}$, then $\varsigma_{i,0,t} = 1$, otherwise $\varsigma_{i,0,t} = 0$.

In general, the LTM framework is a useful strategy for models where the researcher wants to introduce dynamic sparsity. For example, as shown in Zhou et al. (2014), allowing for such mechanism uniformly improves out-of-sample predictions in a portfolio analysis due to the parsimony it induces. Here, the LTM concept serves two purposes. First, if estimating the factor loadings without allowing for time variation, the researcher might conclude that a given topic has no relationship with $\boldsymbol{a}_t$, i.e., that $\boldsymbol{z}_{i,0:s} = 0$, simply because, on average, periods with a positive $\boldsymbol{z}_{i,0:s,t}$ cancels with periods with a negative $\boldsymbol{z}_{i,0:s,t}$. By using the time-varying parameter formulation above, this pitfall is avoided. Second, it is not very likely that one particular topic is equally important throughout the estimation sample. A topic might be very informative in some periods, but not in others. The threshold mechanism potentially captures such cases in a consistent and transparent way, safeguards against over-fitting, and controls for the fact that the relationship between the indicators and output growth might be unstable, confer the discussion in Section 1.[12]

As is common for all factor models, the factors and factor loadings in (4) are not identified without restrictions. To separately identify the factors and the loadings, the following identification restrictions on $\boldsymbol{z}_{0,t}$ in (4a) are enforced:

$$\boldsymbol{z}_{0,t} = \begin{bmatrix} \tilde{\boldsymbol{z}}_{0,t} \\ \hat{\boldsymbol{z}}_{0,t} \end{bmatrix}, \quad \text{for } t = 0, 1, \ldots, T \tag{8}$$

Here, $\tilde{\boldsymbol{z}}_{0,t}$ is a $q \times q$ identity matrix for all t, and $\hat{\boldsymbol{z}}_{0,t}$ is left unrestricted. Bai and Ng (2013) and Bai and Wang (2012) show that these restrictions uniquely identify the dynamic factors and the loadings, but leave the VAR(h) dynamics for the factors completely unrestricted.

[12] The same arguments naturally applies when constructing coincident indexes using more conventional indicators (like financial and labor market data).

3.2 Introducing mixed frequency variables

Due to the mixed frequency property of the data, the y_t vector in equation (4a) contains both observable and unobservable variables. Thus, the model as formulated in (4) can not be estimated. However, following Harvey (1990), and since $\boldsymbol{y}^*_{1,t}$ is a flow measure, the model can be reformulated such that observed quarterly series are treated as daily observations with missing observations. To this end, the $\boldsymbol{y}_t$ vector is decomposed as in equation (5). Assuming further that the quarterly variables, e.g., output growth defined in Section 2.3, are observed at the last day of each quarter, we can define:

$$\tilde{y}_{1,t} = \begin{cases} \sum_{j=0}^{m} y^*_{1,t-j} & \text{if } \tilde{y}_{1,t} \text{ is observed} \\ NA & \text{otherwise} \end{cases} \tag{9}$$

where $\tilde{y}_{1,t}$ is treated as the intra-period sum of the corresponding daily values, and m denotes the number of days since the last observation period. Because quarters have uneven number of days, $\tilde{y}_{1,t}$ is observed on an irregular basis. Accordingly, m will vary depending on which quarter and year we are in. This variation is however known, and easily incorporated into the model structure.

Given (9), temporal aggregation can be handled by introducing a cumulator variable of the form:

$$C_{1,t} = \beta_{1,t} C_{1,t-1} + y^*_{1,t} \tag{10}$$

where $\beta_{1,t}$ is an indicator variable defined as:

$$\beta_{1,t} = \begin{cases} 0 & \text{if } t \text{ is the first day of the period} \\ 1 & \text{otherwise} \end{cases} \tag{11}$$

and $y^*_{1,t}$ maps to the latent factor, $\boldsymbol{a}_t$, from equation (4b). Thus, $\tilde{y}_{1,t} = C_{1,t}$ whenever $\tilde{y}_{t,1}$ is observed, and treated as a missing observation in all other periods. Because of the usage of the cumulator variable in (10), one additional state variable is introduced to the system. Importantly, however, the system will now be possible to estimate using standard filtering techniques handling missing observations. Details are given in Appendix E.

Some remarks are in order. First, although mappings between mixed frequency variables have been applied extensively in both mixed frequency VARs and factor models, see Foroni and Marcellino (2013) for an overview, the cumulator approach has been exploited less regularly. For the purpose of this analysis it offers a clear advantage because it expands the number of state variables in the system only marginally. In contrast, using the mixed frequency approaches in, e.g., Mariano and Murasawa (2003) and Aruoba et al. (2009), would have expanded the number of state variables in the model by over 180 and 90, respectively. Such large number of states pose significant challenges for estimation,

making it almost infeasible in a Bayesian context.[13] Second, introducing (flow) variables of other frequencies than daily and quarterly into the system is not difficult. For each new frequency one simple constructs one new cumulator variable, specific for that frequency, and augment the system accordingly.

3.3 Model specification and estimation

In the model specification used to produce the main results one latent daily coincident index is identified. This latent daily coincident index is assumed to follow an AR(10) process, thus, $q = 1$ and $h = 10$. I do not allow for lags of the dynamic factors in the observation equation (4a) of the system, i.e., $s = 0$. Conceptually it would have been straightforward to use higher values for s for the N_d rows in (4a) associated with the observable daily observations. However, for the N_q rows associated with the quarterly variables, setting $s > 0$ would conflict with the temporal aggregation described in Section 3.2. For all the N elements in $\boldsymbol{e}_t$ (see equation 4c), the AR(p) dynamics are restricted to one lag, i.e., $p = 1$. To avoid end point issues due to data revisions with the latest vintage of output, I restrict the estimation sample to the period 1989-01-01 to 2013-31-12. Finally, based on simple correlation statistics between the news topic time series and output growth I truncate the D_1^f data set to include only the 20 most correlated (in absolute value) topics, see Table 3 in Appendix A. This latter adjustment is done to ease the computational burden, but, as seen from Figure 4, unsupervised PCA estimates of the topic time series result in almost identical factor estimates irrespective of whether 20 or 80 topics are used, suggesting that 20 topics are enough.[14]

The time-varying DFM is estimated by decomposing the problem of drawing from the joint posterior of the parameters of interest into a set of much simpler ones using Gibbs simulations. The Gibbs simulation employed here, together with the prior specifications, are described in greater detail in Appendix E. The results reported in this paper are all based on 9000 iterations of the Gibbs sampler. The first 6000 are discarded and only every sixth of the remaining are used for inference.[15]

[13] For example, Aruoba et al. (2009) employ Maximum Likelihood estimation, and note that one evaluation of the likelihood takes roughly 20 seconds. As Bayesian estimation using MCMC (see Section 3.3) requires a large number of iterations, the problem quickly becomes infeasible in terms of computation time.

[14] Still, the truncation is admittedly somewhat arbitrary. Noting that comparable coincident index models already proposed in the literature also resort to some type of variable selection prior to estimation, I leave it for future research to devise potentially more optimal methods to truncate the topics data set.

[15] As shown in Appendix E.7, and in Appendix E.8 for a simulation experiment, the convergence statistics seem satisfactory.

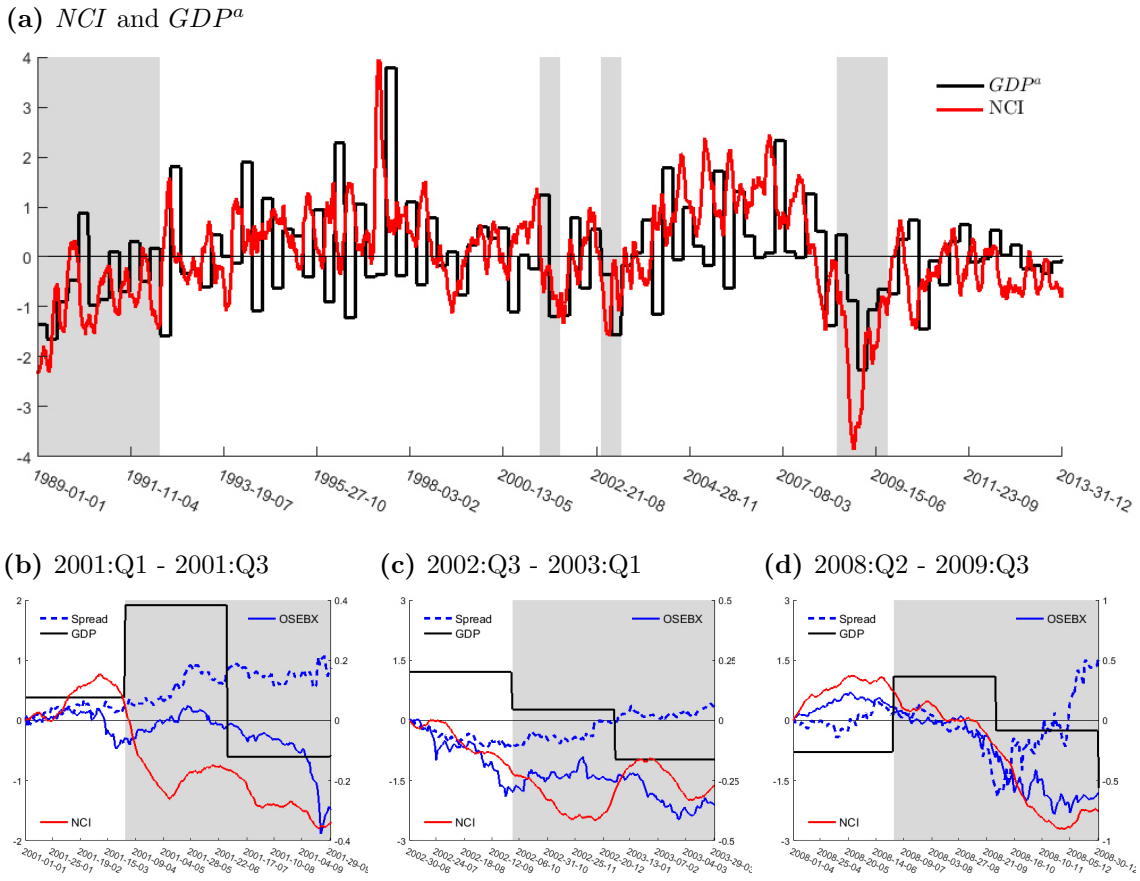

Figure 5. GDP^a is recorded at the end of each quarter, but reported on a daily basis in the graphs using previous end-of-period values throughout the subsequent quarter. *NCI* is the standardized measure of the daily business cycle index. Recession periods, defined by a *MS-FMQ* model (see Section 4.1), are illustrated using gray color shading. Figures 5b to 5d focus on three specific periods where output is illustrated using *GDP*. The indicators are normalized to zero on the first day of the first quarter displayed. *OSEBX* is the cumulative return over the period, and *Spread* is the difference between the 10 year and 3 month money market interest rate.

4 A newsy coincident index of the business cycle

Figure 5 reports the estimated *NCI*. As clearly seen in the upper part of the figure, the index tracks the general economic fluctuations closely. Compared to the simple PCA estimates reported in Figure 4, the *NCI* seems to provide a better fit: It captures the low growth period in the early 1990s, the boom and subsequent bust around the turn of the century, and finally the high growth period leading up to the Great Recession. Note, however, that in Norway, the downturn in the economy following the Norwegian banking crisis in the late 1980s was just as severe as the downturn following the global financial crisis in 2008.

An (informal) example of the importance of having timely information about the state of the economy is given in Figures 5b to 5d. They show the benefits of the *NCI* relative to using two timely and often-used indicators: the stock index (*OSEBX*) and

yield spreads (*Spreads*) (see, e.g., Estrella and Mishkin (1998)) around three important turning points in the Norwegian economy the last decades. For example, as seen in Figure 5d, between the second and third quarter of 2008 output growth declined considerably. During the month of August 2008, and in particular following Lehman Brothers collapse on September 15, 2008, the stock index, the yield spread, and the *NCI* plummet. Since the actual number for GDP growth in the third quarter of 2008 was not known before late 2008, both leading indicators and the *NCI* would have been useful for picking up the change in economic conditions prior to what we now know turned out to be a recession in this example. However, Figure 5d, and in particular Figures 5b and 5c, also show the problem with relying on the indicators alone: Their relationship with output growth is unstable. During the recession period in the early 2000s for example, see Figure 5b, the spread did not signal any downturn at all. Likewise, for this period the changes in the stock index did not turn significantly negative before almost one quarter after the recession started. In contrast, for all three recession periods reported in Figure 5, the *NCI* provides a more or less timely signal of the downturn.

4.1 Business cycles and index evaluation

Making a formal evaluation of the *NCI* is challenging. By construction, the quarterly sum of the daily *NCI* will equal the observed quarterly growth rates in GDP^a (plus a measurement error, c.f. Section 3.2), while daily business cycle conditions, on the other hand, are not observed. Alternatively, in the tradition of Burns and Mitchell (1946), and later work by, e.g., Bry and Boschan (1971) and Hamilton (1989), to mention just two of many, aggregate economic activity can be categorized as phases of expansions and contractions, and one can assess the index's ability to classify such phases. This is the route I take here.

Following Travis and Jordà (2011), I use Receiver Operating Characteristic (ROC) curves and the area under the curve (AUROC) statistic to score the *NCI*'s ability to classify the state of the economy.[16] Here, I do so along four dimensions: How well it

[16] In economics, Travis and Jordà (2011) introduced the ROC methodology to classify economic activity into recessions and expansions. An ideal binary classifier would always indicate a recession when a recession actually occurs (true positive), while never indicate a recession when it does not occur (false positive). In Figure 6a, for example, such a classifier would be depicted by a point in the upper left corner. A model not performing any better than random guessing would end up at the 45 degree line. Thus, using the ROC one can easily compare the trade-offs (cost/benefit) one faces when using different models or indicators for classification. The AUROC is an often used summary statistic within the ROC framework. By definition the AUROC can not exceed 1, perfect classification, or be lower than 0.5. I compute the AUROC score non-parametrically using the algorithm described in Travis and Jordà (2011), and refer to their work for an overview of the ROC technicalities and advantages in terms of scoring business cycle

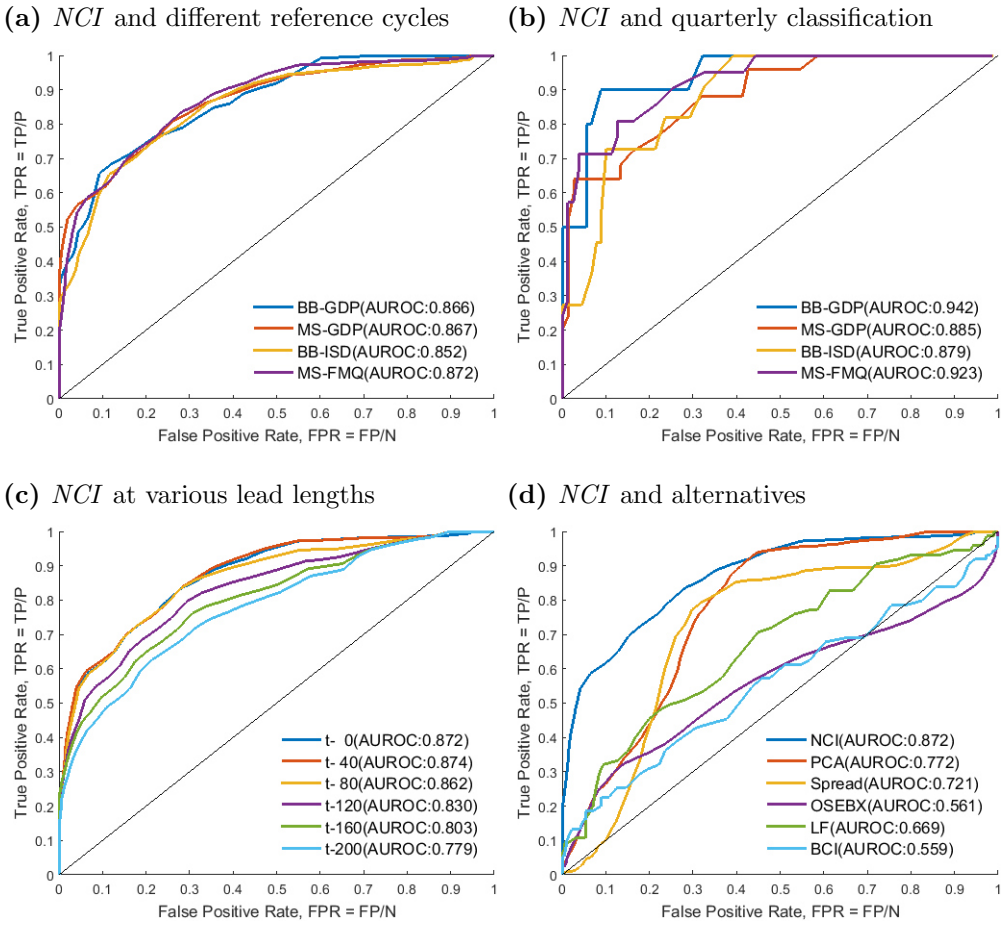

Figure 6. Receiver Operating Characteristics curves (ROC). Figure 6a reports the *NCI*'s ability of classifying business cycle phases across four different business cycle chronologies. In Figures 6b to 6d the *MS-FMQ* chronology is used as the reference cycle. Figure 6b reports the results when classification is scored at a quarterly frequency. Figure 6c reports the results when the *NCI* is lagged $p = \{0, 40, \ldots, 200\}$ days. Figure 6d compares the performance of the daily *NCI* against a set of daily and monthly alternatives. For the monthly indicators, *LFS* and *BCI*, daily numbers are obtained using previous end-of-period values throughout the subsequent month.

categorizes business cycles using different reference cycles; how well it categorizes business cycles at a different level of time aggregation; how well it categorizes business cycles at different lags; and finally, how well it categorizes business cycles compared to other often used and observable alternatives. See Section 4.4 for evaluations of the *NCI* relative to other estimated coincident indexes.

Figure 6a assesses the *NCI*'s classification ability against four different business cycle chronologies, developed by Aastveit et al. (2016) for the Norwegian economy.[17] Each

chronologies. Still, as the true underlying state of the economy is never observed, even retrospectively, and since the categorization of economic activity does not follow any universally agreed upon law, there will be an inherent uncertainty also with this type of evaluation. An evaluation of a different sort, but perhaps more hefty, can be obtained by running a real-time out-of-sample forecasting experiment.

[17] In contrast to in, e.g., the U.S., which has an official business cycle dating committee (NBER), no such

chronology is constructed using different methodologies to extract the unobserved phases: uni- and multivariate Bry-Boschan approaches (*BB-GDP* and *BB-ISD*), a univariate Markow-switching model (*MS-GDP*), and a Markov-Switching factor model (*MS-FMQ*). Aastveit et al. (2016) provide a description of these approaches and the data used. The resulting quarterly classifications, and additional model details, are summarized in Table 2 in Appendix A.[18] As seen from Figure 6a, irrespective of which reference cycle that is used to define the Norwegian business cycle, the *NCI* yields a true positive rate of roughly 80 percent, at the cost of only 25 percent false positives. The AUROC measures are also between 0.85 and 0.87 in all four cases, signaling very good classification. While these results are strong, but not perfect, it should be remembered that the *NCI* might provide an estimate of the economy's phases that is closer to the unknown truth than any of the other reference cycles I use to evaluate it. Moreover, the classification models typically used are at the quarterly (or monthly) frequency, while the *NCI* allows for daily classification. Aggregating the *NCI* to a quarterly time series, by simply computing the mean growth rate for each quarter, we observe that the index's classification ability becomes even better, see Figure 6b. When using the *MS-FMQ* as the reference cycle, for example, an AUROC of 0.92 is achieved at the quarterly frequency against 0.87 at the daily frequency. Compared with the results reported for quarterly Norwegian data in Aastveit et al. (2016), and U.S. data in Travis and Jordà (2011), this score is very competitive.[19]

The results reported in Figure 5 indicated that the *NCI* had leading properties. This is confirmed more formally in Figure 6c. Lagging the *NCI* 40 days yields a higher AUROC score than actually using the *NCI* as a contemporaneous classifier for the business cycle. The performance of the *NCI* does not really start to drop before it is lagged almost one quarter (80 days), suggesting that the *NCI* would be a highly useful indicator for turning point predictions and nowcasting.

Traditionally, coincident indexes are constructed using a number of observable daily and monthly variables. In Figure 6d, the classification properties of some of these variables (see Appendix A for data descriptions) are compared to the *NCI*. The best performing observable indicator in terms of ROC curve scoring is the daily *Spread* followed by the monthly labor force survey (*LFS*). Using stock returns or the business confidence indicator

institution or formal dating exists for Norway.

[18] Daily classifications are obtained by assuming that the economy remains in the same phase on each day within the quarterly classification periods.

[19] Using the reference cycle generated by the *MS-FMQ* model for Norwegian data, Aastveit et al. (2016) show that the *BB-GDP* model gets an AUROC of 0.93. Using U.S. data, and comparing various leading indicators and coincident indexes, Travis and Jordà (2011) show that the best performing coincident index is the one developed by Aruoba et al. (2009). This index receives an AUROC of 0.96 when the NBER business cycle chronology is used as a reference cycle.

(*OSEBX* and *BCI*) are almost no better than random guessing in terms of classifying the business cycle, confirming the impression from Figure 5. It is noteworthy that the PCA estimated news index (see Section 2.3) performs better than any of the other alternatives. At the cost of 40 percent false positive rates, it can give almost 100 percent true positive rates. Still, the AUROC score for the PCA estimated news index is well below the *NCI*'s.

In sum, the results presented above suggest that the *NCI* adds value. Although other alternatives also provide information that is competitive relative to the *NCI*, these alternatives are not necessarily available on a daily frequency and they do not provide the users of such information any broader rational in terms of why the indicators fall or rise. As shown in the next section, the *NCI* does.

4.2 News and index decompositions

Figure 7 illustrates how changes in the *NCI* can be decomposed into the contributions from the individual news topics, and thereby address what type of new information underlies changes in business cycle conditions.[20] To economize on space, I only report nine of the topics contributing to the *NCI* estimate. The 11 remaining topics are reported in Figure 9 in Appendix B. Three distinct results stand out.

First, the topics listed in Figure 7 do, for the most part, reflect topics one would expect to be important for business cycles in general, and for business cycles in Norway in particular. Examples of the former are the *Monetary policy*, *Fiscal policy*, *Wage payments/Bonuses*, *Stock market*, *Funding*, and *Retail* topics, while the *Oil production* and *Oil service* topics are examples of the latter.[21] The remaining topics (see Figure 9 in Appendix B) are typically related to general business cycle sensitive sectors (reflected by, e.g., *Airline industry* and *Automobiles* topics) and technological developments (reflected by, e.g., *IT-technology* and *Startup* topics). Still, although most topics are easily interpretable and provide information about what is important for the current state of the economy, some topics either have labels that are less informative, or reflect surprising categories. An example is the *Life* topic, reported in Figure 9. That said, such exotic or less informative named topics, are the exception rather than the rule. It is also the case that a given newspaper article contains many topics at the same time. To the extent that different topics, meaningful or not from an economic point of view, stand close to each

[20] Technically, these results are constructed using the Kalman Filter iterations and decomposing the state evolution at each updating step into news contributions (see Appendix E.5). The decompositions reported in Figure 7 are based on running the Kalman Filter using the posterior median estimates of the hyperparameters and the time-varying factor loadings (at each time t).

[21] Norway is a major petroleum exporter, and close to 50 percent of its export revenues are linked to oil and gas. See Bjørnland and Thorsrud (2015), and the references therein, for a more detailed analysis of the strong linkages between the oil sector and the rest of the mainland economy.

other in the decomposition of the corpus (see Figure 2) they might covary and therefore both add value in terms of explaining the current state of the economy.

Second, while some topics seem to be important almost every period throughout the sample, other topics only contribute significantly at certain time periods. The *Oil service* topic provides an example: Almost throughout the whole sample, until the mid 2000s, its contribution is close to zero. After 2004, however, its contribution becomes highly positive. Similar observations can also be confirmed for the *Stock market* topic, in particular. The extended periods of zero contribution are partly due to the threshold mechanism used when estimating the time-varying factor loadings. I return to this discussion in Section 4.3.

Third, the timing of when specific topics become important, either positively or negatively, resonates well with what we now know about the economic developments the last two decades. Without dredging too deep into the historical narrative of the Norwegian business cycle, I give three examples: It is by now well recognized that the extraordinary boom in the Norwegian economy during the 2000s was highly oil-driven. The large positive contributions from the two oil topics, *Oil service* and *Oil production*, reflect this.[22] It is also well known that Norwegian (cost) competitiveness has declined considerably during the two last decades. According to the National Accounts statistics, annual wages and salaries increased considerably during especially two periods: the mid-1990s and the mid-late 2000s. Both patterns are clearly visible in the graph showing how media coverage of the *Wage payments/Bonuses* topic contributes to the index fluctuations. Finally, we see from the bottom graph in Figure 7 that the *Funding* topic, a newspaper topic focused on words associated with credit and loans, contributed especially negatively during the Great Recession period. Again, this resonates well with the historical narrative, given what we today know about the Great Recession episode.

Some might find it tempting to interpret the news topics, and their contribution to the *NCI*, as some type of causal relationship between news and economic fluctuations. Technically, within the current framework, this is not a valid interpretation because the decompositions reported in Figure 7 are based on predictive properties. Instead, the newspaper topics should simply be interpreted as a broad panel of different high frequent economic indicators, informative about the current state of the economy. Still, there is a large literature emphasizing the role of news as an important channel for generating business cycles, see, e.g., Beaudry and Portier (2014) for an overview. In particular, in Larsen and Thorsrud (2015) it is shown, using the same raw text data as here, that

[22] During the 1980s and 1990s, value added in the oil service sector hardly grew in real terms (according to the National Accounts statistics for Norway). From the early 2000s until today, the sector has grown by over 300 percent.

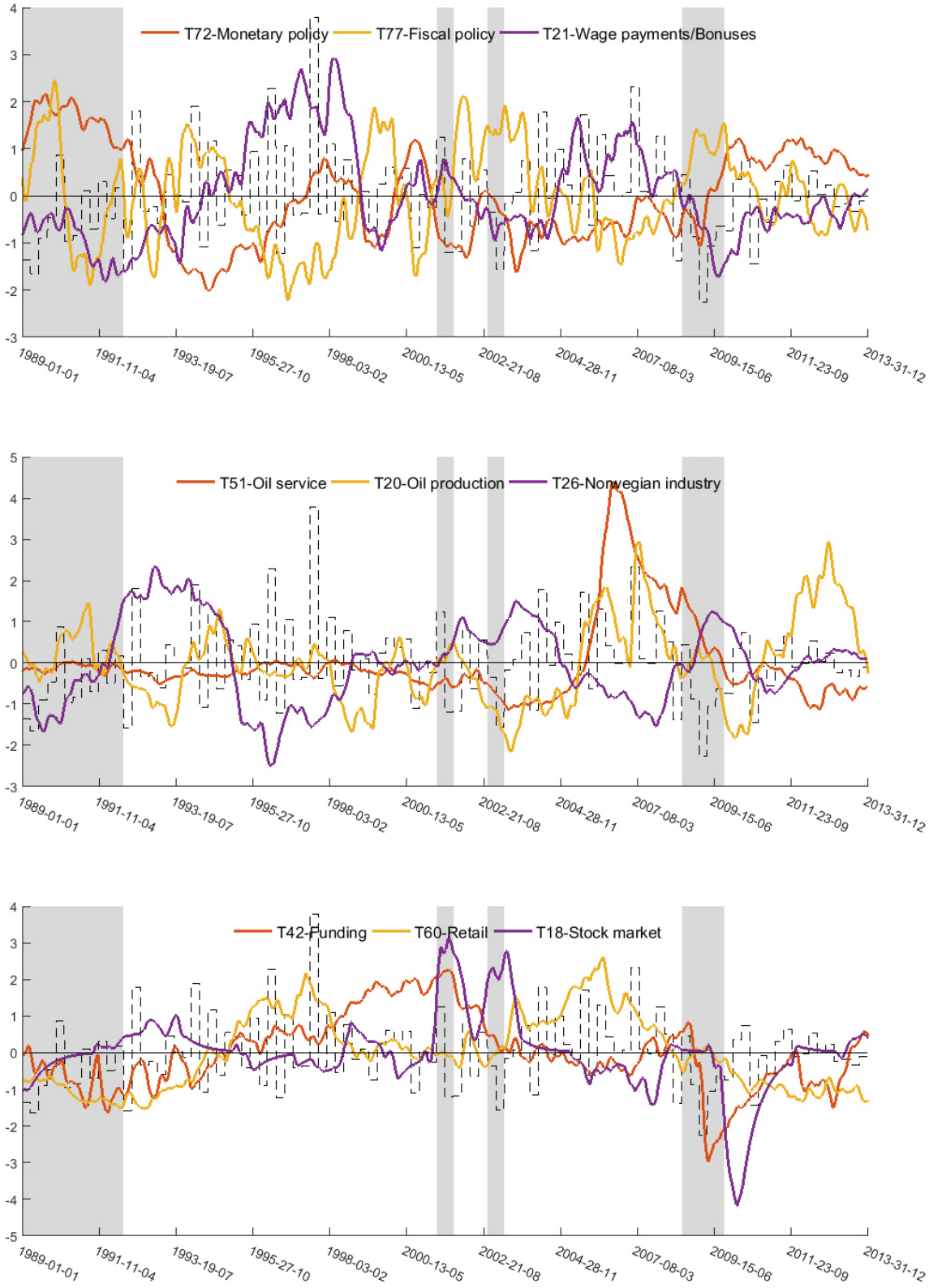

Figure 7. News topics and their (median) contribution to *NCI* estimates across time. The news topic contributions are standardized and illustrated using different colors. GDP^a, graphed using a dotted black line, is recorded at the end of each quarter, but reported on a daily basis using previous end-of-period values throughout the subsequent quarter. Recession periods, defined by a *MS-FMQ* model, see Section 4.1, are illustrated using gray shading.

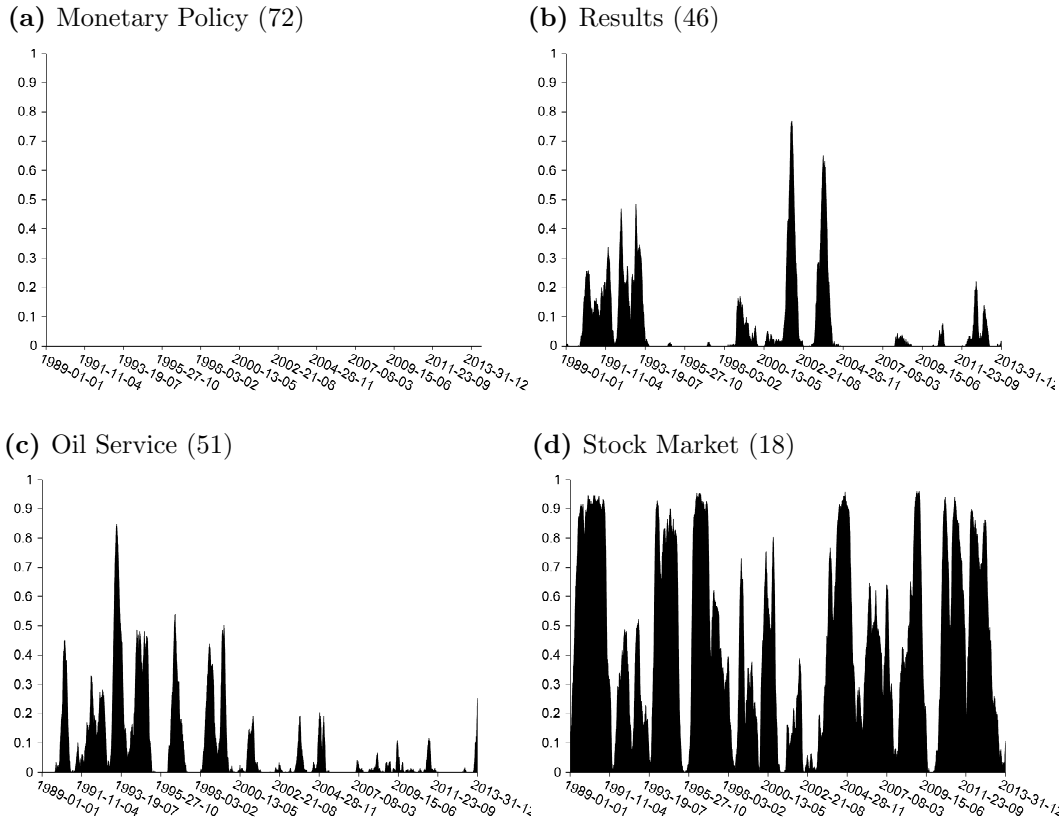

Figure 8. Topics and threshold probabilities across time (topic numbers, confer Table 3 in Appendix A, in parentheses). Each graph reports the posterior probability that the factor loading associated with topic i is 0.

unexpected innovations to a quarterly news index cause persistent fluctuations in both productivity, consumption, and output. While these responses are well in line with the predictions given by the news driven business cycle view, they stand in stark contrast to those one would obtain if the informational content of the news topics were associated with some type of sentiment, see, e.g., Beaudry et al. (2011) and Angeletos and La'O (2013). It is plausible that items in the newspaper generate a self-fulfilling feedback loop where the mood of the news changes economic activity, thus validating the original sentiment.

4.3 Threshold probabilities

An important aspect of the modeling strategy described in Section 3 is the time-varying factor loadings, and the sparsity enforced through the latent threshold mechanism. The effect of the threshold mechanism could be seen (partly) through how the news topic contributions varied across time, confer Figure 7. Following Nakajima and West (2013) and Zhou et al. (2014), a more direct illustration is offered in Figure 8.

Figure 8 exemplifies how the posterior probability of a binding threshold varies across time for factor loadings associated with four of the topics. It is particularly noteworthy

how the threshold probability for the *Stock market* topic varies substantially; perhaps capturing the conventional wisdom that the stock market predicts more recessions than what we actually observe in the data. Note, however, that during the slow growth periods in the beginning of the 2000s, when the world economy was hit by the bust of the dot-com bubble, the threshold probabilities came down considerably. A similar pattern is observed around the financial crisis, although the likelihood of a zero loading is only close to zero somewhat late in the crisis. In contrast, for the *Monetary Policy* topic, the threshold probability is essentially zero throughout the sample, indicating that this topic is an important part of the daily business cycle index in all periods. For both the *Oil Service* and *Results* topics the threshold probabilities show a more varied pattern. Still, on average there is a greater probability of a binding threshold for the factor loadings associated with the *Oil Service* topic in the early part of the sample compared to the later periods. For the *Results* topic, the mid 1990s and 2000s stand out as particularly informative, with long periods of (close to) zero threshold probabilities.

When investigating the threshold probabilities for the other factor loadings (not shown) I observe that they in most cases do not bind. For the researcher or index user wanting to enforce a larger degree of sparsity onto the system, a tighter prior for the threshold parameter likely needs to be imposed (confer Appendix E.2). However, in unreported experiments, and in this context, I find that imposing such a prior reduces the index's classification power.

4.4 Extensions and comparisons

In this section I do three experiments: First, I assess the importance of the LTM mechanism by estimating the DFM using the same information set as above, but without allowing for time-varying parameters. Second, I asses the importance of using a larger set of mixed frequency information by also including variables observed on a monthly frequency in the model.[23] Finally, to assess how well the *NCI* compares to other coincident indexes estimated without utilizing the daily newspaper topics, I compare its performance against three more standard alternatives.

The results from these experiments are summarized in Table 1. Across 40 different evaluations, non of the alternative specifications improve upon the benchmark *NCI*. On average, the benchmark *NCI* receives an AUROC score which is 27 (25) percent higher than the alternatives when evaluated against four different daily (quarterly) reference

[23]By not including more variables of lower frequency than daily in the benchmark model, the *NCI* model formulation departs from what is commonly used. For example, Mariano and Murasawa (2003) mix a small set of monthly variables with quarterly output growth to construct a coincident index, while Aruoba et al. (2009) mix both daily, weekly, monthly, and quarterly information to do the same.

Table 1. ROC comparison across models. Each entry in the table reports the AUROC score of the benchmark NCI model relative to five alternatives across different reference cycles (confer Section 4.1). The numbers in parentheses report the relative scores when the models are evaluated at a quarterly frequency. A value higher than one indicates that the NCI model receives a higher AUROC score. The alternative mixed frequency DFM models are: an NCI model estimated without allowing for time-varying factor loadings (*NCI-F*); an NCI model estimated with an augmented data set including monthly labor market data (*NCI-DM*); a coincident index constructed without newspaper data, but with monthly labor market and confidence data, and daily spreads and returns data (*CI-DM*); two coincident indexes constructed without newspaper data, but with daily spreads and returns data (*CI-D* and *CI-FD*). All alternative models, except *NCI-F* and *CI-FD*, are estimated allowing for time-varying factor loadings. A description of the data used is given in Appendix A.

	NCI-F	**NCI-DM**	**CI-DM**	**CI-D**	**CI-FD**
BB-GDP	1.22 (1.22)	1.36 (1.24)	1.36 (1.24)	1.17 (1.04)	1.38 (1.45)
MS-GDP	1.26 (1.34)	1.35 (1.28)	1.35 (1.28)	1.22 (1.16)	1.17 (1.20)
BB-ISD	1.11 (1.10)	1.50 (1.56)	1.51 (1.56)	1.16 (1.04)	1.09 (1.07)
MS-FMQ	1.23 (1.30)	1.29 (1.26)	1.30 (1.27)	1.25 (1.16)	1.20 (1.24)
Average	1.20 (1.24)	1.37 (1.33)	1.38 (1.34)	1.20 (1.10)	1.21 (1.24)

cycles. Focusing on the effect of including the time-varying parameter formulation we see that the benchmark model performs up to 26 (34) percent better than the alternative *NCI-F* when the daily (quarterly) *MS-GDP* reference cycle is used. Across all the reference cycles, the results strongly suggest that allowing for time-varying factor loadings increases classification precision. Moving to the second question raised above, including monthly information into the model does not to increase the model's classification abilities. Irrespective of which frequency and reference cycle the benchmark index is evaluated against, the *NCI* is on average 37 (33) percent better than the alternative *NCI-DM*. Given the results presented in Figure 6, where the monthly variables themselves did not actually obtain a very high AUROC, this is perhaps not very surprising. Finally, when mixed frequency time-varying DFMs are estimated using conventional daily and/or monthly variables only, the *CI-D*, *CI-DM*, and *CI-FD* models, the benchmark model is clearly better for three out of the four reference cycles used. Interestingly, comparing the results for the *NCI-F* model against the *CI-FD* model, we observe that they perform almost identical. Thus, it is the combined effect of allowing for the LTM mechanisms for the factor loadings and the usage of newspaper data which makes the *NCI* outperform the other alternatives. Obviously, one can not rule out that a more careful selection of other high frequent conventional variables might improve the relative score of the alternative models estimated here. On the other hand, a more careful selection of news topics might also improve the score of the *NCI*, and neither of the alternatives, *CI-D*, *CI-DM*, and *CI-FD*, offer the benefits in terms of news decomposition illustrated in Section 4.2.

5 Conclusion

In this paper I develop a time-varying mixed frequency Dynamic Factor Model where dynamic sparsity is enforced on the factor loadings using a latent threshold mechanism, and show how textual information contained in a business newspaper can be utilized to construct a daily coincident index of business cycles. Both contributions, the usage of newspaper data and the latent threshold mechanism, add value. The constructed index has almost perfect classification abilities, and outperforms many commonly used alternatives. In contrast to existing approaches, the usage of newspaper data also gives the index user broad based information about what is leading to the changes in the daily index. That is, when decomposing the coincident index into news topic contributions I show that news topics related to monetary and fiscal policy, the stock market and credit, and industry-specific sectors seem to provide the most important information about daily business cycle conditions. Moreover, the sign and timing of their individual contributions map well onto the historical narrative we have about recent business cycle swings.

The (macro)economic literature utilizing textual information and alternative data sources is fast growing, but still in its early stages. Although highly data and computationally intensive, the results presented here are encouraging and motivates further research. Going forward, an assessment of the predictive power of the proposed daily coincident index, and testing the methodology across different countries and media types, are natural extensions.

References

Aastveit, K. A., A. S. Jore, and F. Ravazzolo (2016). Identification and real-time forecasting of Norwegian business cycles. *International Journal of Forecasting 32*(2), 283 – 292.

Angeletos, G.-M. and J. La'O (2013). Sentiments. *Econometrica 81*(2), 739–779.

Apel, M. and M. Blix Grimaldi (2012). The information content of central bank minutes. Working Paper Series 261, Sveriges Riksbank (Central Bank of Sweden).

Aruoba, S. B., F. X. Diebold, and C. Scotti (2009). Real-Time Measurement of Business Conditions. *Journal of Business & Economic Statistics 27*(4), 417–427.

Bai, J. and S. Ng (2013). Principal components estimation and identification of static factors. *Journal of Econometrics 176*(1), 18 – 29.

Bai, J. and P. Wang (2012). Identification and estimation of dynamic factor models. MPRA Paper 38434, University Library of Munich, Germany.

Baker, S. R., N. Bloom, and S. J. Davis (2013). Measuring economic policy uncertainty. *Chicago Booth research paper* (13-02).

Balke, N. S., M. Fulmer, and R. Zhang (2015). Incorporating the Beige Book into a Quantitative Index of Economic Activity. Mimeo, Southern Methodist University.

Beaudry, P., D. Nam, and J. Wang (2011). Do Mood Swings Drive Business Cycles and is it Rational? NBER Working Papers 17651, National Bureau of Economic Research, Inc.

Beaudry, P. and F. Portier (2014). News-Driven Business Cycles: Insights and Challenges. *Journal of Economic Literature 52*(4), 993–1074.

Bjørnland, H. C. and L. A. Thorsrud (2015). Commodity prices and fiscal policy design: Procyclical despite a rule. Working Papers 0033, Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School.

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

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res. 3*, 993–1022.

Bloom, N. (2014). Fluctuations in Uncertainty. *Journal of Economic Perspectives 28*(2), 153–76.

Bry, G. and C. Boschan (1971). *Cyclical Analysis of Time Series: Selected Procedures and Computer Programs*. Number 71-1 in NBER Books. National Bureau of Economic Research, Inc.

Burns, A. F. and W. C. Mitchell (1946). *Measuring Business Cycles*. Number 46-1 in NBER Books. National Bureau of Economic Research, Inc.

Carter, C. K. and R. Kohn (1994). On Gibbs Sampling for State Space Models. *Biometrika 81*(3), 541–553.

Choi, H. and H. Varian (2012). Predicting the present with Google trends. *Economic Record 88*(s1), 2–9.

Del Negro, M. and C. Otrok (2008). Dynamic factor models with time-varying parameters: measuring changes in international business cycles. Staff Reports 326, Federal Reserve Bank of New York.

Ellis, C., H. Mumtaz, and P. Zabczyk (2014). What Lies Beneath? A Time-varying FAVAR Model for the UK Transmission Mechanism. *Economic Journal 0*(576), 668–699.

Estrella, A. and F. S. Mishkin (1998). Predicting US recessions: Financial variables as leading indicators. *Review of Economics and Statistics 80*(1), 45–61.

Evans, M. D. D. (2005). Where Are We Now? Real-Time Estimates of the Macroeconomy. *International Journal of Central Banking 1*(2).

Foroni, C. and M. Marcellino (2013). A survey of econometric methods for mixed-frequency data. Working Paper 2013/06, Norges Bank.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *In BAYESIAN STATISTICS*.

Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America 101*(Suppl 1), 5228–5235.

Hamilton, J. D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica 57*(2), 357–84.

Hansen, S. and M. McMahon (2015). Shocking Language: Understanding the macroeconomic effects of central bank communication. Discussion Papers 1537, Centre for Macroeconomics (CFM).

Hansen, S., M. McMahon, and A. Prat (2014). Transparency and Deliberation within the FOMC: A Computational Linguistics Approach. CEP Discussion Papers 1276, Centre for Economic Performance, LSE.

Harvey, A. C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge Books. Cambridge University Press.

Heinrich, G. (2009). Parameter estimation for text analysis. Technical report, Fraunhofer IGD.

Larsen, V. H. and L. A. Thorsrud (2015). The Value of News. Working Papers 0034, Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School.

Lopes, H. F. and C. M. Carvalho (2007). Factor Stochastic Volatility with Time Varying Loadings and Markov Switching Regimes. *Journal of Statistical Planning and Inference* (137), 3082–3091.

Loughran, T. and B. Mcdonald (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance 66*(1), 35–65.

Marcellino, M., M. Porqueddu, and F. Venditti (2013). Short-term GDP forecasting with a mixed frequency dynamic factor model with stochastic volatility. Temi di discussione (Economic working papers) 896, Bank of Italy, Economic Research and International Relations Area.

Mariano, R. S. and Y. Murasawa (2003). A new coincident index of business cycles based on monthly and quarterly series. *Journal of Applied Econometrics 18*(4), 427–443.

Nakajima, J. and M. West (2013). Bayesian Analysis of Latent Threshold Dynamic Models. *Journal of Business & Economic Statistics 31*(2), 151–164.

Raftery, A. E. and S. Lewis (1992). How many iterations in the Gibbs sampler. In *In BAYESIAN STATISTICS*.

Stock, J. and M. Watson (2012). Disentangling the channels of the 2007-2009 recession. *Brookings Papers on Economic Activity Spring 2012*, 81–135.

Stock, J. H. and M. W. Watson (1988). A Probability Model of The Coincident Economic Indicators. NBER Working Papers 2772, National Bureau of Economic Research, Inc.

Stock, J. H. and M. W. Watson (1989). New indexes of coincident and leading economic indicators. In *NBER Macroeconomics Annual 1989, Volume 4*, NBER Chapters, pp. 351–409. National Bureau of Economic Research, Inc.

Stock, J. H. and M. W. Watson (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature 41*(3), 788–829.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance 62*(3), 1139–1168.

Tetlock, P. C. (2014). Information Transmission in Finance. *Annual Review of Financial Economics 6*(1), 365–384.

Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy (2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. *Journal of Finance 63*(3), 1437–1467.

Travis, J. B. and s. Jordà (2011). Evaluating the Classification of Economic Activity into Recessions and Expansions. *American Economic Journal: Macroeconomics 3*(2), 246–77.

Zhou, X., J. Nakajima, and M. West (2014). Bayesian forecasting and portfolio decisions using dynamic dependent sparse factor models. *International Journal of Forecasting 30*, 963–980.

Appendices

Appendix A Data, reference cycles and topics

The newspaper corpus and data used for gross domestic product (GDP) are described in Sections 2.1 and 2.3, respectively. Different classifications of the business cycle into phases of expansions and recessions are listed in Table 2. A summary of all the estimated news topics, and the most important words associated with each topic, are reported in Table 3. The remaining data used in this analysis are obtained from Reuters Datastream, and are as follows: *Spread* is constructed as the difference between the 10-year benchmark rate and the interbank three month offered rate. *OSEBX* is (log) daily returns computed using the Oslo Stock Exchange Benchmark index. Both the *Spread* and *OSEBX* variables are recorded on a daily frequency. Missing observations, during weekends, are filled using simple linear interpolation. Like for the daily newspaper topic time series, prior to estimation I smooth the series using a 60-day (backward-looking) moving average filter, and standardize the resulting variables. *BCI* is the seasonally adjusted industrial confidence indicator for the manufacturing sector in Norway, and *LFS* is the seasonally adjusted labor force. Both variables are recorded on a monthly frequency, and transformed to (log) monthly growth rates. The series are smoothed using a two-month (backward-looking) moving average filter and standardized prior to estimation.

Table 2. Reference cycles 1986 to 2014 (as estimated in Aastveit et al. (2016)). The different chronologies build on: a Bry-Boschan approach using GDP growth (*BB-GDP*); a univariate Markow-switching model using GDP growth (*MS-GDP*); the Bry-Boschan approach applied to a coincident index based on inverse standard deviation weighting (*BB-ISD*); and a multivariate Markow-swithing model (*MS-FMQ*). For both the *BB-ISD* and *MS-FMQ* models six quarterly variables are included: the Brent Blend oil price, employment in mainland Norway, household consumption, private real investment in mainland Norway, exports of traditional goods and GDP for mainland Norway. See Aastveit et al. (2016), and the references therein, for more formal model, data, and estimation descriptions.

		BB-GDP	MS-GDP	BB-ISD	MS-FMQ
1986 - 1989	Peak	1987:Q2	1986:Q2	1987:Q4	1987:Q2
	Trough	1989:Q3		1989:Q1	
1990 - 1994	Peak			1991:Q1	
	Trough		1991:Q4	1991:Q4	1991:Q4
1995 - 2001	Peak	2001:Q1	2001:Q1	2001:Q1	2001:Q1
	Trough	2001:Q3	2001:Q3	2001:Q3	2001:Q3
2002 - 2003	Peak		2002:Q2		2002:Q3
	Trough		2002:Q4		2003:Q1
2004 - 2010	Peak	2008:Q2	2007:Q4	2007:Q4	2008:Q2
	Trough	2008:Q3	2010:Q1	2009:Q1	2009:Q3

Table 3. Estimated topics and labeling. The topics are labeled based on the meaning of the most important words (see the text for details). The "Corr" column reports the topics' correlation (using the D_1^f data set, see Section 2.2) with linearly interpolated daily GDP^a (the correlation rank is reported in parenthesis). The words are translated from Norwegian to English using Google Translate.

Topic	Label	Corr	First words
Topic 0	Calender	0.03(69)	january, march, october, september, november, february
Topic 1	Family business	0.18(19)	family, foundation, name, dad, son, fortune, brothers
Topic 2	Institutional investing	0.10(39)	fund, investments, investor, return, risk, capital
Topic 3	Justice	0.04(65)	lawyer, judge, appeal, damages, claim, supreme court
Topic 4	Surroundings	0.18(18)	city, water, meter, man, mountain, old, outside, nature
Topic 5	Housing	0.14(29)	housing, property, properties, apartment, square meter
Topic 6	Movies/Theater	0.08(50)	movie, cinema, series, game, producer, prize, audience
Topic 7	Argumentation	0.11(34)	word, besides, interesting, i.e., in fact, sure, otherwise
Topic 8	Unknown	0.09(42)	road, top, easy, hard, lift, faith, outside, struggle,fast
Topic 9	Agriculture	0.03(68)	industry, support, farmers, export, production, agriculture
Topic 10	Automobiles	0.18(17)	car, model, engine, drive, volvo, ford, møller, toyota
Topic 11	USA	0.09(47)	new york, dollar, wall street, president, usa, obama, bush
Topic 12	Banking	0.00(80)	dnb nor, savings bank, loss, brokerage firm, kreditkassen
Topic 13	Corporate leadership	0.05(59)	position, chairman, ceo, president, elected, board member
Topic 14	Negotiation	0.04(61)	solution, negotiation, agreement, alternative, part, process
Topic 15	Newspapers	0.22(9)	newspaper, media, schibsted, dagbladet, journalist, vg
Topic 16	Health care	0.00(77)	hospital, doctor, health, patient, treatment, medication
Topic 17	IT systems	0.17(24)	it, system, data, defense, siem, contract, tanberg, deliver
Topic 18	Stock market	0.23(8)	stock exchange, fell, increased, quote, stock market

Continued on next page

Table 3 – continued from previous page

Topic	Label	Corr	First words
Topic 19	Macroeconomics	0.07(53)	economy, budget, low, unemployment, high, increase
Topic 20	Oil production	0.18(20)	statoil, oil, field, gas, oil company, hydro, shelf, stavanger
Topic 21	Wage payments	0.26(7)	income, circa, cost, earn, yearly, cover, payed, salary
Topic 22	Norwegian regions	0.17(23)	trondheim, llc, north, stavanger, tromsø, local, municipality
Topic 23	Family	0.04(64)	woman, child, people, young, man, parents, home, family
Topic 24	Taxation	0.03(71)	tax, charge, revenue, proposal, remove, wealth tax, scheme
Topic 25	EU	0.04(62)	eu, eea, commission, european, brussel, membership, no
Topic 26	Norwegian industry	0.20(13)	hydro, forest, factory, production, elkem, industry, produce
Topic 27	Unknown	0.07(54)	man, he, friend, smile, clock, evening, head, never, office
Topic 28	Norwegian groups	0.09(45)	orkla, storebrand, merger, bid, shareholder, acquisitions
Topic 29	UK	0.06(57)	british, london, great britain, the, of, pound, england
Topic 30	Narrative	0.03(72)	took, did, later, never, gave, stand, happened, him, began
Topic 31	Shipping	0.10(36)	ship, shipping, dollar, shipowner, wilhelmsen, fleet, proud
Topic 32	Projects	0.10(38)	project, nsb, development, fornebu, entrepreneurship
Topic 33	Oil price	0.11(32)	dollar, oil price, barrel, oil, demand, level, opec, high
Topic 34	Sports	0.00(78)	olympics, club, football, match, play, lillehammer, sponsor
Topic 35	Organizations	0.10(41)	leader, create, organization, challenge, contribute, expertise
Topic 36	Drinks	0.13(30)	wine, italy, taste, drinks, italian, fresh, fruit, beer, bottle
Topic 37	Nordic countries	0.04(63)	swedish, sweden, danish, denmark, nordic, stockholm
Topic 38	Airline industry	0.21(12)	sas, fly, airline,norwegian, braathens, airport, travel
Topic 39	Entitlements	0.02(73)	municipality, public, private, sector, pension, scheme

Continued on next page

Table 3 – continued from previous page

Topic	Label	Corr	First words
Topic 40	Employment conditions	0.08(51)	cut, workplace, measures, salary, labor, working, employ
Topic 41	Norwegian politics	0.05(60)	høyere, party, ap, labor party, stoltenberg, parlament, frp
Topic 42	Funding	0.31(3)	loan, competition, creditor, loss, bankruptcy, leverage
Topic 43	Literature	0.01(76)	book, books, read, publisher, read, author, novel, wrote
Topic 44	Statistics	0.27(6)	count, increase, investigate, share, average, decrease
Topic 45	Watercraft	0.01(75)	ship, boat, harbor, strait, shipowner, on board, color
Topic 46	Results	0.31(4)	quarter, surplus, deficit, tax, group, operating profit, third
Topic 47	TV	0.12(31)	tv, nrk, channel, radio, digital, program, media
Topic 48	International conflicts	0.10(40)	war, africa, irak, south, un, army, conflict, troops, attack
Topic 49	Political elections	0.02(74)	election, party, power, politics, vote, politician, support
Topic 50	Music	0.09(46)	the, music, record, of, in, artist, and, play, cd, band, song
Topic 51	Oil service	0.19(14)	rig, dollar, contract, option, offshore, drilling, seadrill
Topic 52	Tourism	0.21(11)	hotel, rom, travel, visit, stordalen, tourist, guest
Topic 53	Unknown	0.16(26)	no, ting, think, good, always, pretty, actually, never
Topic 54	Aker	0.11(35)	aker, kværner, røkke, contract, shipyard, maritime
Topic 55	Fishery	0.16(27)	fish, salmon, seafood, norway, tons, nourishment, marine
Topic 56	Europe	0.08(49)	german, russia, germany, russian, west, east, french, france
Topic 57	Law and order	0.06(56)	police, finance guards, aiming, illegal, investigation
Topic 58	Business events	0.00(79)	week, financial, previous, friday, wednesday, tdn, monday
Topic 59	Supervision	0.10(37)	report, information, financial supervision, enlightenment
Topic 60	Retail	0.31(2)	shop, brand, steen, rema, reitan, as, group, ica, coop

Continued on next page

Table 3 – continued from previous page

Topic	Label	Corr	First words
Topic 61	Startup	0.28(5)	bet, cooperation, establish, product, party, group
Topic 62	Food	0.19(15)	food, restaurant, salt, nok, pepper, eat, table, waiter
Topic 63	Listed stocks	0.11(33)	shareholder, issue, investor, holding, stock exchange listing
Topic 64	Asia	0.09(43)	china, asia, chinese, india, hong kong, south, authorities
Topic 65	Art	0.09(44)	picture, art, exhibition, gallery, artist, museum, munch
Topic 66	Disagreement	0.08(52)	criticism, express, asserting, fault, react, should, alleging
Topic 67	Debate	0.15(28)	degree, debate, context, unequal, actually, analysis
Topic 68	Life	0.18(21)	man, history, dead, him, one, live, church, words, strokes
Topic 69	Distribution	0.18(22)	customer, post, product, offers, service, industry, firm
Topic 70	Telecommunication	0.08(48)	telenor, mobile, netcom, hermansen, telia, nokia, ericsson
Topic 71	IT technology	0.21(10)	internet, net, pc, microsoft, technology, services, apple
Topic 72	Monetary policy	0.33(1)	interest rate, central bank, euro, german, inflation, point
Topic 73	Education	0.04(66)	school, university, student, research, professor, education
Topic 74	Government regulations	0.03(70)	rules, authorities, competition, regulations, bans
Topic 75	Trade organizations	0.16(25)	lo, nho, members, forbund, strike, organization, payroll
Topic 76	Fear	0.04(67)	fear, emergency, hit, severe, financial crisis, scared
Topic 77	Fiscal policy	0.19(16)	suggestions, parliamentary, ministry, selection, minister
Topic 78	Energy	0.05(58)	energy, emissions, statkraft, industry, environment
Topic 79	Foreign	0.07(55)	foreign, abroad, japan, japanese, immigration, games

Appendix B Additional results

Figure 9. News topics and their (median) contribution to *NCI* estimates across time. The news topic contributions are standardized and illustrated using different colors. GDP^a, graphed using a black dotted line, is recorded at the end of each quarter, but reported on a daily basis using previous end-of-period values throughout the subsequent quarter. Recession periods, defined by a *MS-FMQ* model (see Section 4.1) are illustrated using grey shading.

Appendix C Filtering the news corpus

To clean the raw textual data set, a stop-word list is first employed. This is a list of common words one does not expect to have any information relating to the subject of an article. Examples of such words are *the*, *is*, *are*, and *this*. The most common Norwegian surnames and given names are also removed. In total, the stop-word list together with the list of common surnames and given names removed roughly 1800 unique tokens from the corpus. Next, an algorithm known as stemming is run. The objective of this algorithm is to reduce all words to their respective word stems. A word stem is the part of a word that is common to all of its inflections. An example is the word *effective* whose stem is *effect*. Finally, a measure called *tf-idf*, which stands for term frequency - inverse document frequency, is calculated. This measures how important all the words in the complete corpus are in explaining single articles. The more often a word occurs in an article, the higher the *tf-idf* score of that word. On the other hand, if the word is common to all articles, meaning the word has a high frequency in the whole corpus, the lower that word's *tf-idf* score will be. Around 250 000 of the stems with the highest *tf-idf* score are kept, and used as the final corpus.

Appendix D LDA estimation and specification

The LDA model was developed in Blei et al. (2003). Here the estimation algorithm described in Griffiths and Steyvers (2004) is implemented. First, recall that the corpus consists of M distinct documents. $N = \sum_{m=1}^{M} N_m$ is the total number of words in all documents, K is the total number of latent topics, and V is the size of the vocabulary. Each document consists of a repeated choice of topics $Z_{m,n}$ and words $W_{m,n}$. Let t be a term in V, and denote $P(t|z = k)$, the mixture component, one for each topic, by $\boldsymbol{\Phi} = \{\boldsymbol{\varphi}_k\}_{k=1}^{K}$. Finally, let $P(z|d = m)$ define the topic mixture proportion for document m, with one proportion for each document $\Theta = \{\boldsymbol{\theta}_m\}_{m=1}^{M}$. The goal of the algorithm is then to approximate the distribution:

$$P(\boldsymbol{Z}|\boldsymbol{W}; \alpha, \beta) = \frac{P(\boldsymbol{W}, \boldsymbol{Z}; \alpha, \beta)}{P(\boldsymbol{W}; \alpha, \beta)} \tag{12}$$

using Gibbs simulations, where α and β are the (hyper) parameters controlling the prior conjugate Dirichlet distributions for $\boldsymbol{\theta}_m$ and $\boldsymbol{\varphi}_k$, respectively. A very good explanation for how this method works is found in Heinrich (2009). The description below provides a brief summary only.

With the above definitions, the total probability of the model can be written as:

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}; \alpha, \beta) = \prod_{k=1}^{K} P(\boldsymbol{\varphi}_i; \beta) \prod_{m=1}^{M} P(\boldsymbol{\theta}_m; \alpha) \prod_{t=1}^{N} P(z_{m,t}|\boldsymbol{\theta}_m) P(w_{m,t}|\boldsymbol{\varphi}_{z_{m,t}}) \tag{13}$$

Integrating out the parameters φ and θ:

$$\begin{aligned} P(\boldsymbol{Z}, \boldsymbol{W}; \alpha, \beta) &= \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Phi}} P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}; \alpha, \beta)\, d\boldsymbol{\Phi}\, d\boldsymbol{\Theta} \\ &= \int_{\boldsymbol{\Phi}} \prod_{k=1}^{K} P(\boldsymbol{\varphi}_k; \beta) \prod_{m=1}^{M} \prod_{t=1}^{N} P(w_{m,t} | \boldsymbol{\varphi}_{z_{m,t}})\, d\boldsymbol{\Phi} \int_{\boldsymbol{\Theta}} \prod_{m=1}^{M} P(\boldsymbol{\theta}_m; \alpha) \prod_{t=1}^{N} P(z_{m,t} | \boldsymbol{\theta}_m)\, d\boldsymbol{\Theta} \end{aligned} \tag{14}$$

In (14), the terms inside the first integral do not include a θ term, and the terms inside the second integral do not include a φ term. Accordingly, the two terms can be solved separately. Exploiting the properties of the conjugate Dirichlet distribution it can be shown that:

$$\int_{\boldsymbol{\Theta}} \prod_{m=1}^{M} P(\boldsymbol{\theta}_m; \alpha) \prod_{t=1}^{N} P(z_{m,t} | \boldsymbol{\theta}_m)\, d\boldsymbol{\Theta} = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma(n_m^{(k)} + \alpha_k)}{\Gamma\left(\sum_{k=1}^{K} n_m^{(k)} + \alpha_k\right)} \tag{15}$$

and

$$\int_{\boldsymbol{\Phi}} \prod_{k=1}^{K} P(\boldsymbol{\varphi}_k; \beta) \prod_{m=1}^{M} \prod_{t=1}^{N} P(w_{m,t} | \boldsymbol{\varphi}_{z_{m,t}})\, d\boldsymbol{\Phi} = \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{t=1}^{V} \beta_t\right)}{\prod_{t=1}^{V} \Gamma(\beta_t)} \frac{\prod_{t=1}^{V} \Gamma(n_k^{(t)} + \beta_t)}{\Gamma\left(\sum_{t=1}^{V} n_k^{(t)} + \beta_t\right)} \tag{16}$$

where $n_m^{(k)}$ denotes the number of word tokens in the m^{th} document assigned to the k^{th} topic, and $n_k^{(t)}$ is the number of times the t^{th} term in the vocabulary has been assigned to the k^{th} topic.

Since $P(\boldsymbol{W}; \alpha, \beta)$, in (12), is invariable for any of Z, the conditional distribution $P(\boldsymbol{Z}|\boldsymbol{W}; \alpha, \beta)$ can be derived from $P(\boldsymbol{W}, \boldsymbol{Z}; \alpha, \beta)$ directly using Gibbs simulation and the conditional probability:

$$P(Z_{(m,n)} \mid \boldsymbol{Z}_{-(m,n)}, \boldsymbol{W}; \alpha, \beta) = \frac{P(Z_{(m,n)}, \boldsymbol{Z}_{-(m,n)}, \boldsymbol{W}; \alpha, \beta)}{P(\boldsymbol{Z}_{-(m,n)}, \boldsymbol{W}; \alpha, \beta)} \tag{17}$$

where $Z_{(m,n)}$ denotes the hidden variable of the n^{th} word token in the m^{th} document, and $Z_{-(m,n)}$ denotes all Zs but $Z_{(m,n)}$. Denoting the index of a word token by $i = (m, n)$, and using the expressions in (15) and (16), cancellation of terms (and some extra manipulations exploiting the properties of the gamma function) yields:

$$P(Z_i = k \mid \boldsymbol{Z}_{-(i)}, \boldsymbol{W}; \alpha, \beta) \propto (n_{m,-i}^{(k)} + \alpha_k) \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,-i}^{(t)} + \beta_t} \tag{18}$$

where the counts $n_{\cdot,-i}^{(\cdot)}$ indicate that token i is excluded from the corresponding document or topic. Thus, sampling topic indexes using equation (18) for each word in a document and across documents until convergence allows us to approximate the posterior distribution given by (12). As noted in Heinrich (2009), the procedure itself uses only five larger data structures; the count variables $n_m^{(k)}$ and $n_k^{(t)}$, which have dimension $M \times K$

and $K \times V$, respectively, their row sums n_m and n_k, as well as the state variable $z_{m,n}$ with dimension W.

With one simulated sample of the posterior distribution for $P(\boldsymbol{Z}|\boldsymbol{W}; \alpha, \beta)$, φ and θ can be estimated from:

$$\hat{\varphi}_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} n_k^{(t)} + \beta_r} \tag{19}$$

and

$$\hat{\theta}_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k} \tag{20}$$

In the analysis of the main paper, the average of the estimated $\hat{\theta}$ and $\hat{\varphi}$ from the 10 last samples of the stored Gibbs simulations are used to construct the daily news topic frequencies.[24] In un-reported experiments, the topic extraction results reported in Section 2.1 do not change much when choosing other samples for inference, for example using the last sample only.

Before estimation, three parameters need to be pre-defined: the number of topics and the two parameter vectors of the Dirichlet priors, α and β. Here, symmetric Dirichlet priors, with α and β each having a single value, are used. In turn, these are defined as a function of the number of topics and unique words:

$$\alpha = \frac{50}{K}, \quad \text{and} \quad \beta = \frac{200}{N}$$

The choice of K is discussed in Section 2.1. In general, lower (higher) values for α and β will result in more (less) decisive topic associations. The values for the Dirichlet hyper-parameters also reflect a clear compromise between having few topics per document and having few words per topic. In essence, the prior specification used here is the same as the one advocated by Griffiths and Steyvers (2004).

D.1 Estimating daily topic frequencies

Using the posterior estimates from the LDA model, the frequency with which each topic is represented in the newspaper for a specific day is computed. This is done by first collapsing all the articles in the newspaper for one specific day into one document. Following Heinrich (2009) and Hansen et al. (2014), a procedure for querying documents outside the set on which the LDA is estimated is then implemented. In short, this corresponds to using the same Gibbs simulations as described above, but with the difference that the sampler is run with the estimated parameters $\boldsymbol{\Phi} = \{\boldsymbol{\varphi}_k\}_{k=1}^{K}$ and hyper-parameter α held fixed.

[24] Because of lack of identifiability, the estimates of $\hat{\theta}$ and $\hat{\varphi}$ can not be combined across samples for an analysis that relies on the content of specific topics. However, statistics insensitive to permutation of the underlying topics can be computed by aggregating across samples (see Griffiths and Steyvers (2004)).

Denote by $\tilde{W}$ the vector of words in the newly formed document. Topic assignments, $\tilde{Z}$, for this document can then be estimated by first initializing the algorithm by randomly assigning topics to words and then perform a number of Gibbs iterations using:

$$P(\tilde{Z}_i = k \mid \tilde{\boldsymbol{Z}}_{-(i)}, \tilde{\boldsymbol{W}}; \alpha, \beta) \propto (n^{(k)}_{\tilde{m},-i} + \alpha_k)\hat{\varphi}_{k,t} \tag{21}$$

Since $\hat{\varphi}_{k,t}$ do not need to be estimated when sampling from (21), fewer iterations are needed to form the topic assignment index for the new document than when learning both the topic and word distributions. Here 2000 iterations are performed, and only the average of every 10th draw is used for the final inference. After sampling, the topic distribution can be estimated as before:

$$\hat{\tilde{\theta}}_{\tilde{m},k} = \frac{n^{(k)}_{\tilde{m}} + \alpha_k}{\sum_{k=1}^{K} n^{(k)}_{\tilde{m}} + \alpha_k} \tag{22}$$

Appendix E The Dynamic Factor model and estimation

For estimation, the Dynamic Factor Model described in Section 3 is rewritten to incorporate the latent threshold mechanism for the time-varying factor loadings and the mixed frequency variables. For notational simplicity, I assume in the following that $s = 0$ (as in the benchmark model), $h = 1$, and $p = 1$. Moreover, I describe a model structure which includes both one quarterly and monthly variable, i.e., $N^q = 1$ and $N^m = 1$, in addition to a $N^d \times 1$ vector of daily observables. Accordingly, following Harvey (1990), the system used for estimation can be written in matrix form as:

$$\boldsymbol{y}_t = \boldsymbol{Z}_t \boldsymbol{a}_t + \boldsymbol{e}_t \tag{23a}$$

$$\boldsymbol{a}_t = \boldsymbol{F}_t \boldsymbol{a}_{t-1} + \boldsymbol{R}\boldsymbol{\omega}_t \tag{23b}$$

$$\boldsymbol{e}_t = \boldsymbol{P}\boldsymbol{e}_{t-1} + \boldsymbol{u}_t \tag{23c}$$

where

$$\boldsymbol{y}_t = \begin{bmatrix} y^q_{1,t} \\ y^m_{1,t} \\ y^d_{1,t} \\ y^d_{2,t} \\ \vdots \\ y^d_{N^d,t} \end{bmatrix} \quad \boldsymbol{Z}_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & z_{1,t} \\ 0 & 0 & z_{2,t} \\ \vdots & \vdots & \vdots \\ 0 & 0 & z_{N^d,t} \end{bmatrix} \quad \boldsymbol{a}_t = \begin{bmatrix} C_{t,q} \\ C_{t,m} \\ a_{t,d} \end{bmatrix} \quad \boldsymbol{e}_t = \begin{bmatrix} 0 \\ 0 \\ e_{1,t} \\ e_{2,t} \\ \vdots \\ e_{N^d,t} \end{bmatrix} \quad \boldsymbol{F}_t = \begin{bmatrix} \beta_{t,q} & 0 & z_q\Phi \\ 0 & \beta_{t,m} & z_m\Phi \\ 0 & 0 & \Phi \end{bmatrix}$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0 & z_q \\ 0 & 1 & z_m \\ 0 & 0 & 1 \end{bmatrix} \quad \boldsymbol{\omega}_t = \begin{bmatrix} \omega_{t,q} \\ \omega_{t,m} \\ \omega_{t,d} \end{bmatrix} \quad \boldsymbol{P} = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \Phi_1 & 0 & 0 \\ \vdots & \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \Phi_{N^d} \end{bmatrix} \quad \boldsymbol{u}_t = \begin{bmatrix} 0 \\ 0 \\ u_{1,t} \\ u_{2,t} \\ \vdots \\ u_{N^d,t} \end{bmatrix}$$

Here, q, m, or d superscripts denote that the variable is observed on a quarterly, monthly or daily frequency, respectively, while q, m, or d subscripts denote that the parameter or variable is associated with quarterly, monthly or daily variables, respectively. $C_{t,q}$ and $C_{t,m}$ are the quarterly and monthly cumulator variables, and $a_{t,d}$ is the daily coincident index. $\beta_{t,q}$ and $\beta_{t,m}$ are indicator variables, associated with quarterly and monthly variables, respectively.

The time-varying factor loadings are modeled as random walks following the Latent Threshold Model (LTM) idea introduced by Nakajima and West (2013). For example, for one particular element in the $\boldsymbol{Z}_t$ matrix, $z_{i,t}$, the LTM structure can be written as:

$$z_{i,t} = z^*_{i,t}\varsigma_{i,t} \quad \varsigma_{i,t} = I(|z^*_{i,t}| \geq d_i) \tag{24}$$

where

$$z^*_{i,t} = z^*_{i,t-1} + w_{i,t} \tag{25}$$

with $w_{i,t} \sim i.i.d.N(0, \sigma^2_{i,w})$. In (24) $\varsigma_{i,t}$ is a zero one variable, whose value depends on the indicator function $I(|z^*_{i,t}| \geq d_i)$. If $|z^*_{i,t}|$ is above the the threshold value d_i, then $\varsigma_{i,t} = 1$, otherwise $\varsigma_{i,t} = 0$.

The vectors of error terms, $\boldsymbol{v}_t$, $\boldsymbol{u}_t$, and $\boldsymbol{w}_t$ are independent:

$$\begin{bmatrix} \boldsymbol{\omega}_t \\ \boldsymbol{u}_t \\ \boldsymbol{w}_t \end{bmatrix} \sim i.i.d.N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Omega} & 0 & 0 \\ 0 & \boldsymbol{U} & 0 \\ 0 & 0 & \boldsymbol{W} \end{bmatrix} \right)$$

and both $\boldsymbol{\Omega}$, $\boldsymbol{U}$, and $\boldsymbol{W}$ are diagonal matrices:

$$\boldsymbol{\Omega} = \begin{bmatrix} \sigma^2_{\omega_q} & 0 & 0 \\ 0 & \sigma^2_{\omega_m} & 0 \\ 0 & 0 & \sigma^2_{\omega_d} \end{bmatrix} \quad \boldsymbol{U} = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \sigma^2_{1,u} & 0 & 0 \\ \vdots & \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{N^d,u} \end{bmatrix} \quad \boldsymbol{W} = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \sigma^2_{1,w} & 0 & 0 \\ \vdots & \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{N^d,w} \end{bmatrix}$$

I note here that by restricting the error matrix $\boldsymbol{\Omega}$ to be non-singular, the model specification basically assumes that the quarterly and monthly variables contain some measurement error relative to the latent daily business cycle factor. Accordingly, identification of

the latent factors, $C_{t,q}$, $C_{t,m}$, and $a_{t,d}$, is obtained by restricting the upper 3×3 block of the time-varying factor loadings matrix $\boldsymbol{Z}_t$ to be an identity matrix. Thus, $z_{1,t} = 1$ for all t, and $\sigma^2_{1,w} = 0$.

The model's hyper-parameters are $\boldsymbol{\Omega}$, $\boldsymbol{U}$, $\boldsymbol{W}$, $\boldsymbol{F}_t$, $\boldsymbol{P}$, and $\boldsymbol{d}$. Inside $\boldsymbol{F}_t$, the indicator variables $\beta_{t,q}$ and $\beta_{t,m}$ are time-varying, but their evolution is deterministic and need not be estimated. Thus, the only time-varying parameters to be estimated within the model are those in $\boldsymbol{Z}_t$, which together with $\boldsymbol{a}_t$, are the model's unobserved state variables.[25]

Estimation consists of sequentially drawing the model's unobserved state variables and hyper-parameters utilizing 4 blocks until convergence is achieved. In essence, each block involves exploiting the state space nature of the model using the Kalman Filter and the simulation smoother suggested by Carter and Kohn (1994), coupled with a Metropolis-Hastings step to simulate the time-varying loadings. Below, I describe each block in greater detail. For future reference and notational simplicity it will prove useful to define the following: $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T]'$, $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_T]'$, $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T]'$, $\boldsymbol{E} = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_T]'$, $\boldsymbol{F} = [\boldsymbol{F}_1, \ldots, \boldsymbol{F}_T]'$, and $\boldsymbol{Q} = \boldsymbol{R}\boldsymbol{\Omega}\boldsymbol{R}'$.

E.1 Block 1: $\boldsymbol{A}|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{E}, \boldsymbol{F}, \boldsymbol{P}, \boldsymbol{U}, \boldsymbol{Q}$

Equations (23a) and (23b) constitute a state space system we can use to draw the unobserved state $\boldsymbol{a}_t$ using the Carter and Kohn's multimove Gibbs sampling approach. However, to do so we need to make the errors in the observation equation conditionally i.i.d. Given knowledge of equation (23c), we can define $\boldsymbol{P}(L) = (\boldsymbol{I} - \boldsymbol{P}L)$ and pre-multiply equation (23a) by $\boldsymbol{P}(L)$ to obtain the system:

$$\tilde{\boldsymbol{y}}_t = \tilde{\boldsymbol{Z}}_t \boldsymbol{a}_t + \boldsymbol{u}_t \sim N(0, \boldsymbol{U}) \tag{26a}$$

$$\boldsymbol{a}_t = \boldsymbol{F}_t \boldsymbol{a}_{t-1} + \boldsymbol{R}\boldsymbol{\omega}_t \sim N(0, \boldsymbol{Q}) \tag{26b}$$

where $\tilde{\boldsymbol{y}}_t = \boldsymbol{P}(L)\boldsymbol{y}_t$ and $\tilde{\boldsymbol{Z}}_t = \boldsymbol{P}(L)\boldsymbol{Z}_t$.

Since all hyper-parameters and state variables, less $\boldsymbol{A}$, are known (or conditionally known), we can use the equations in (26) together with Carter and Kohn's multimove Gibbs sampling approach (see Appendix E.5) to sample $\boldsymbol{a}_t$ from:

$$\boldsymbol{a}_T | \cdots \sim N(\boldsymbol{a}_{T|T}, \boldsymbol{P}^a_{T|T}) \quad t = T \tag{27a}$$

$$\boldsymbol{a}_t | \cdots \sim N(\boldsymbol{a}_{t|t,a_{t+1}}, \boldsymbol{P}^a_{t|t,a_{t+1}}) \quad t = T-1, T-2, \cdots, 1 \tag{27b}$$

to get $\boldsymbol{A}$. Note here that the Kalman Filter can be run straightforwardly despite the fact that the $\tilde{\boldsymbol{y}}_t$ vector contains missing values (see Harvey (1990) for details).

[25] Note that, in principle, the z_q, and z_m hyper-parameters could have been made time-varying. However, I find that estimation of the model then becomes much more sensitive to the prior specification, and have therefore chosen to treat them as constant.

E.2 Block 2: $\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{W}, \boldsymbol{d}|\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{E}, \boldsymbol{P}$

Conditionally on $\boldsymbol{A}$, the errors in (23a) are independent across the N variables in $\boldsymbol{y}_t$. Moreover, we have assumed that the covariance matrix $\boldsymbol{W}$ associated with the time-varying factor loadings in equation (25) is diagonal. Consequently, one can draw $\boldsymbol{Z}$ one equation at a time. As above, we deal with the fact that the errors in the observation equation are not conditionally i.i.d. by applying the quasi differencing operator, $\boldsymbol{P}(L)$, to each equation. Thus, for each $i = 2, \ldots, N^d$, we obtain the following Gaussian system:

$$\tilde{y}^d_{i,t} = \tilde{a}_{t,d} z_{i,t} + u_{i,t} \tag{28a}$$

$$z_{i,t} = z^*_{i,t} \varsigma_{i,t} \quad \varsigma_{i,t} = I(|z^*_{i,t}| \geq d_i) \tag{28b}$$

$$z^*_{i,t} = z^*_{i,t-1} + w_{i,t} \tag{28c}$$

where $\tilde{y}^d_{i,t} = (I - \Phi_i L) y^d_{i,t}$, and $\tilde{a}_{t,d} = (I - \Phi_i L) a_{t,d}$.

To simulate from the conditional posterior of $z^*_{i,ti}$ and d_i in (28), the procedure outlined in Nakajima and West (2013) is followed. That is, conditional on all the data and hyper-parameters, we draw the conditional posterior of $z^*_{i,t}$ sequentially for $t = 1 : T$ using a Metropolis-Hastings (MH) sampler. As described in Nakajima and West (2013), the MH proposals come from a non-thresholded version of the model specific to each time t, as follows: Fixing $\varsigma_{i,t} = 1$, take proposal distribution $N(z^*_{i,t}|m_t, M_t)$ where:

$$M_t^{-1} = \sigma^{-2}_{i,u} \tilde{a}_{t,d} \tilde{a}_{t,d} + \sigma^{-2}_{i,w} (I + 1) \tag{29a}$$

$$m_t = M_t [\sigma^{-2}_{i,u} \tilde{a}_{t,d} \tilde{y}^d_{i,t} + \sigma^{-2}_{i,w} \{(z^*_{i,t-1} + z^*_{i,t+1}) + (I - 1) z^*_{i,0}\}] \tag{29b}$$

for $t = 2 : T - 1$. For $t = 1$ and $t = T$, a slight modification is needed. Details can be found in Nakajima and West (2013). The candidate is accepted with probability:

$$\alpha(z^*_{i,t}, z^{p*}_{i,t}) = min \left\{ 1, \frac{N(\tilde{y}^d_{i,t}|\tilde{a}_{t,d} z^p_{i,t}, \sigma^2_{i,u}) N(z^*_{i,t}|m_t, M_t)}{N(\tilde{y}^d_{i,t}|\tilde{a}_{t,d} z_{i,t}, \sigma^2_{i,u}) N(z^{p*}_{i,t}|m_t, M_t)} \right\} \tag{30}$$

where $z_{i,t} = z^*_{i,t} \varsigma_{i,t}$ is the current state, and $z^p_{i,t} = z^{p*}_{i,t} \varsigma^p_{i,t}$ is the candidate.

The independent latent thresholds in d_i can then be sampled conditional on the data and the hyper-parameters. For this, a direct MH algorithm is employed. Let $d_{i,-j} = d_{i,0:s} \backslash d_{i,j}$. A candidate is drawn from the current conditional prior, $d^p_{i,j} \sim U(0, |\beta_0| + K)$, where K is described below, and accepted with probability:

$$\alpha(d_{i,j}, d^p_{i,j}) = min \left\{ 1, \Pi^T_{t=1} \frac{N(\tilde{y}^d_{i,t}|\tilde{a}_{t,d} z^p_{i,t}, \sigma^2_{i,u})}{N(\tilde{y}^d_{i,t}|\tilde{a}_{t,d} z_{i,t}, \sigma^2_{i,u})} \right\} \tag{31}$$

where $z_{i,t}$ is the state based on the current thresholds $(d_{i,j}, d_{i,-j})$, and $z^p_{i,t}$ the candidate based on $(d^p_{i,j}, d_{i,-j})$.

Lastly, conditional on the data, the hyper parameters and the time-varying parameters, we can sample the elements of $\boldsymbol{U}$ and $\boldsymbol{W}$ using the inverse Gamma distributions. Let

letters denoted with an underscore reflect the prior, then:

$$\sigma_{i,u}^2|\cdots \sim IG(\bar{v}^u, \bar{\sigma}_{i,u}^2) \tag{32}$$

where $\bar{v}^u = T + \underline{T}^u$ and $\bar{\sigma}_{i,u}^2 = [\underline{\sigma}_{i,u}^2 \underline{T}^u + \sum_{t=1}^{T}(\tilde{y}_{i,t}^d - \tilde{a}_{t,d} z_{i,t})'(\tilde{y}_{i,t}^d - \tilde{a}_{t,d} z_{i,t})]/\bar{v}^u$, and:

$$\sigma_{i,w}^2|\cdots \sim IG(\bar{v}^w, \bar{\sigma}_{i,w}^2) \tag{33}$$

where $\bar{v}^w = T + \underline{T}^w$ and $\bar{\sigma}_{i,w}^2 = [\underline{\sigma}_{i,w}^2 \underline{T}^w + \sum_{t=1}^{T}(z_{i,t}^* - z_{i,t-1}^*)'(z_{i,t}^* - z_{i,t-1}^*)]/\bar{v}^w$. Note here that for $\sigma_{i,u}^2$, the simulations are done for $i = 1, \ldots, N^d$, while for $\sigma_{i,w}^2$ they are only done for $i = 2, \ldots, N^d$ because $z_{1,t} = 1$ for all t by restriction.

E.3 Block 3: $\boldsymbol{F}, \boldsymbol{\Omega}|\boldsymbol{A}$

Conditional on $\boldsymbol{A}$, the transition equation in (23b) is independent of the rest of the system. While the first and second equations of (23b) do depend on the estimates of Φ, the part of the transition equation associated with $a_{t,d}$ is independent of the rest of the components in (23b). Accordingly, Φ and $\sigma_{\omega_d}^2$, the element in the lower right corner of $\boldsymbol{\Omega}$, can first be simulated independently from the rest of the parameters in (23b), and then $\sigma_{\omega_m}^2$, $\sigma_{\omega_q}^2$, z_m, and z_q can be simulated conditionally on Φ and $\sigma_{\omega_d}^2$.

To simulate Φ and $\sigma_{\omega_d}^2$, we employ the independent Normal-Gamma prior. Accordingly, continuing with letting letters denoted with an underscore reflect the prior, the conditional posterior of Φ is:

$$\Phi|\cdots \sim N(\bar{\Phi}, \bar{V}^{\Phi})_{I[s(\Phi)]} \tag{34}$$

with

$$\bar{V}^{\Phi} = (\underline{V}^{\Phi^{-1}} + \sum_{t=1}^{T} a_{t-1,d}' \sigma_{\omega_d}^{-2} a_{t-1,d})^{-1} \tag{35}$$

$$\bar{\Phi} = \bar{V}^{\Phi}(\underline{V}^{\Phi^{-1}} \underline{\Phi} + \sum_{t=1}^{T} a_{t-1,d}' \sigma_{\omega_d}^{-2} a_{t,d}) \tag{36}$$

and $I[s(\Phi)]$ is an indicator function used to denote that the roots of Φ lie outside the unit circle. Further, the conditional posterior of $\sigma_{\omega_d}^2$ is:

$$\sigma_{\omega_d}^2|\cdots \sim IG(\bar{v}^{\omega_d}, \bar{\sigma}_{\omega_d}^2) \tag{37}$$

with $\bar{v}^{\omega_d} = T + \underline{T}^{\omega_d}$, and $\bar{\sigma}_{\omega_d}^2 = [\underline{\sigma}_{\omega_d}^2 \underline{T}^{\omega_d} + \sum_{t=1}^{T}(a_t - a_{t-1}\Phi)'(a_t - a_{t-1}\Phi)]/\bar{v}^{\omega_d}$.

Once Φ and $\sigma_{\omega_d}^2$ are drawn, we can construct, for $j = \{q, m\}$:

$$C_{t,j} - \beta_{t,j} C_{t-1,j} \equiv C_{t,j}^* = z_j a_{t,d} + \omega_{t,j} \tag{38}$$

and draw from the conditional posterior of z_j and $\sigma_{\omega_j}^2$. Using again the independent Normal-Gamma prior:

$$z_j|\cdots \sim N(\bar{z}_j, \bar{V}^{z_j}) \tag{39}$$

with

$$\bar{V}^{z_j} = (\underline{V}^{z_j^{-1}} + \sum_{t=1}^{T} a'_{t,d}\sigma_{\omega_j}^{-2}a_{t,d})^{-1} \tag{40}$$

$$\bar{z}_j = \bar{V}^{z_j}(\underline{V}^{z_j^{-1}}\underline{z}_j + \sum_{t=1}^{T} a'_{t,d}\sigma_{\omega_j}^{-2}C^*_{t,j}) \tag{41}$$

Finally, the conditional posterior of $\sigma^2_{\omega_j}$ is:

$$\sigma^2_{\omega_j}|\cdots \sim IG(\bar{v}^{\omega_j}, \bar{\sigma}^2_{\omega_j}) \tag{42}$$

with $\bar{v}^{\omega_j} = T + \underline{T}^{\omega_j}$, and $\bar{\sigma}^2_{\omega_j} = [\underline{\sigma}^2_{\omega_j}\underline{T}^{\omega_j} + \sum_{t=1}^{T}(C^*_{t,j} - a_{t,d}z_j)'(C^*_{t,j} - a_{t,d}z_j)]/\bar{v}^{\omega_j}$.

E.4 Block 4: $\boldsymbol{E}|\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{Z}$ and $\boldsymbol{P}|\boldsymbol{E}, \boldsymbol{U}$

For each observation of the N^d daily variables, we have that:

$$e_{i,t} = y_{i,t} - z_{i,t}a_{t,d} \tag{43}$$

Thus, conditional on $\boldsymbol{Y}$, $\boldsymbol{A}$ and $\boldsymbol{Z}$, $\boldsymbol{E}$ is observable. As above, since $\boldsymbol{E}$ is independent across the N^d equations, we can sample the elements of $\boldsymbol{P}$ in (23c) one equation at the time. As this is done in the same manner as in equations (34) to (36) of Block 3 (with the obvious change of notation), I do not repeat the computations here.

E.5 The Carter and Kohn algorithm and observation weights

Consider a generic state space system, written in companion form, and described by:

$$\boldsymbol{y}_t = \boldsymbol{Z}_t\boldsymbol{a}_t + \boldsymbol{u}_t \sim N(0, \boldsymbol{U}) \tag{44a}$$

$$\boldsymbol{a}_t = \boldsymbol{F}\boldsymbol{a}_{t-1} + \boldsymbol{R}\boldsymbol{\omega}_t \sim N(0, \boldsymbol{Q}) \tag{44b}$$

where we assume that the hyper-parameters $\theta = \{\boldsymbol{U}, \boldsymbol{F}, \boldsymbol{R}, \boldsymbol{Q}\}$, and $\boldsymbol{Z}_t$ are known, and we wish to estimate the latent state $\boldsymbol{a}_t$ for all $t = 1, \ldots, T$. To do so, we can apply Carter and Kohn's multimove Gibbs sampling approach (see Carter and Kohn (1994)).

First, because the state space model given in equation (44) is linear and (conditionally) Gaussian, the distribution of $\boldsymbol{a}_t$ given $\boldsymbol{Y}$ and that of $\boldsymbol{a}_t$ given $\boldsymbol{a}_{t+1}$ and $\boldsymbol{Y}$ for $t = T - 1, \ldots, 1$ are also Gaussian:

$$\boldsymbol{a}_T|\boldsymbol{Y} \sim N(\boldsymbol{a}_{T|T}, \boldsymbol{P}_{T|T}), \quad t = T \tag{45a}$$

$$\boldsymbol{a}_t|\boldsymbol{Y}, \boldsymbol{a}_{t+1} \sim N(\boldsymbol{a}_{t|t,a_{t+1}}, \boldsymbol{P}_{t|t,a_{t+1}}), \quad t = T-1, T-2, \cdots, 1 \tag{45b}$$

where

$$\boldsymbol{a}_{T|T} = E(\boldsymbol{a}_T|\boldsymbol{Y}) \tag{46a}$$

$$\boldsymbol{P}_{T|T} = Cov(\boldsymbol{a}_T|\boldsymbol{Y}) \tag{46b}$$

$$\boldsymbol{a}_{t|t,a_{t+1}} = E(\boldsymbol{a}_t|\boldsymbol{Y}, \boldsymbol{a}_{t+1}) = E(\boldsymbol{a}_t|\boldsymbol{a}_{t|t}, \boldsymbol{a}_{t|t+1}) \tag{46c}$$

$$\boldsymbol{P}_{t|t,a_{t+1}} = Cov(\boldsymbol{a}_t|\boldsymbol{Y}, \boldsymbol{a}_{t+1}) = Cov(\boldsymbol{a}_t|\boldsymbol{a}_{t|t}, \boldsymbol{a}_{t|t+1}) \tag{46d}$$

Given $\boldsymbol{a}_{0|0}$ and $\boldsymbol{P}_{0|0}$, the unknown states $\boldsymbol{a}_{T|T}$ and $\boldsymbol{P}_{T|T}$ needed to draw from (45a) can be estimated from the (conditionally) Gaussian Kalman Filter as:

$$\boldsymbol{a}_{t|t-1} = \boldsymbol{F}\boldsymbol{a}_{t-1|t-1} \tag{47a}$$

$$\boldsymbol{P}_{t|t-1} = \boldsymbol{F}\boldsymbol{P}_{t-1|t-1}\boldsymbol{F}' + \boldsymbol{Q} \tag{47b}$$

$$\boldsymbol{K}_t = \boldsymbol{P}_{t|t-1}\boldsymbol{Z}_t'(\boldsymbol{Z}_t\boldsymbol{P}_{t|t-1}\boldsymbol{Z}_t' + \boldsymbol{U})^{-1} \tag{47c}$$

$$\boldsymbol{a}_{t|t} = \boldsymbol{a}_{t|t-1} + \boldsymbol{K}_t(\boldsymbol{y}_t - \boldsymbol{Z}_t\boldsymbol{a}_{t|t-1}) \tag{47d}$$

$$\boldsymbol{P}_{t|t} = \boldsymbol{P}_{t|t-1} - \boldsymbol{K}_t\boldsymbol{Z}_t\boldsymbol{P}_{t|t-1} \tag{47e}$$

That is, at $t = T$, equation 47d and 47e above, together with equation 45a, can be used to draw $\boldsymbol{a}_{T|T}$. Moreover, $\boldsymbol{a}_{t|t,a_{t+1}}$ for $t = T-1, T-2, \cdots, 1$ can also be simulated based on 45b, where $\boldsymbol{a}_{t|t,a_{t+1}}$ and $\boldsymbol{P}_{t|t,a_{t+1}}$ are generated from the following updating equations:

$$\boldsymbol{a}_{t|t,a_{t+1}} = \boldsymbol{a}_{t|t} + \boldsymbol{P}_{t|t}\boldsymbol{F}'(\boldsymbol{F}\boldsymbol{P}_{t|t}\boldsymbol{F}' + \boldsymbol{Q})^{-1}(\boldsymbol{a}_{t+1} - \boldsymbol{F}\boldsymbol{a}_{t|t}) \tag{48a}$$

$$\boldsymbol{P}_{t|t,a_{t+1}} = \boldsymbol{P}_{t|t} + \boldsymbol{P}_{t|t}\boldsymbol{F}'(\boldsymbol{F}\boldsymbol{P}_{t|t}\boldsymbol{F}' + \boldsymbol{Q})^{-1}\boldsymbol{F}\boldsymbol{P}_{t|t} \tag{48b}$$

When computing the news topic contributions in Figures 7 and 9, I decompose the state vector into a history of forecast error contributions. For simplicity, I use the notation introduced in Appendix E.5 to describe how this is done. At each time interval t, the forecast error in predicting $\boldsymbol{y}_t$ is given by $\boldsymbol{v}_t = \boldsymbol{y}_t - \boldsymbol{Z}_t\boldsymbol{a}_{t|t-1}$. In computing $\boldsymbol{a}_{t|t}$, equation (47d) above, the Kalman gain $\boldsymbol{K}_t$ is used to weight each forecast error when computing the updated state estimate. If the predictions of the ith observable at time t are perfect, $v_{i,t} = 0$ and this observation does not contribute to potential updates from $\boldsymbol{a}_{t|t-1}$ to $\boldsymbol{a}_{t|t}$. If the predictions of the ith observable at time t are not perfect, $v_{i,t} \neq 0$, the observation will influence the updated state estimate as long as it is given weight through the $\boldsymbol{K}_t$ matrix. As the updating equation in 47d has a recursive structure, the time evolution of $\boldsymbol{a}_{t|t}$ can easily be decomposed into a set of weighted forecast error contributions, resulting in the decompositions shown in Figures 7 and 9.

E.6 Prior specification

To implement the MCMC algorithm, and estimate the model, prior specifications for the initial state variables $\boldsymbol{a}_0$, $\boldsymbol{Z}_0$, and for the hyper-parameters $\boldsymbol{\Omega}$, $\boldsymbol{U}$, $\boldsymbol{W}$, $\boldsymbol{F}_t$, $\boldsymbol{P}$, and $\boldsymbol{d}$

are needed. The prior specifications used for the initial states take the following form: $\boldsymbol{a}_0 \sim N(0, I \cdot 100)$, and $\boldsymbol{Z}_0 \sim N(0, I \cdot 100)$. The priors for the hyper-parameters Φ and Φ, which are part of the $\boldsymbol{F}_t$ and $\boldsymbol{P}$ matrices, respectively, are set to:

$$\begin{aligned} \underline{\Phi} \sim & N(\hat{\Phi}_{OLS}, V(\hat{\Phi}_{OLS})) \\ \underline{\Phi}_i \sim & N(0, 0.5) \qquad\qquad\qquad\qquad \text{for } i = 1, \ldots, N^d \end{aligned}$$

where $\hat{\Phi}_{OLS}$ are the OLS estimates of an AR(h) using the first principal component of the daily news dataset as dependent variable. $V(\hat{\Phi}_{OLS})$ is a diagonal matrix where the non-zero entries are the variance terms associated with the $\hat{\Phi}_{OLS}$ elements. To draw z_q and z_m, which are part of the $\boldsymbol{F}_t$ matrix, I use: $\underline{z}_j \sim N(1, 1)$ for $j = \{q, m\}$.

The priors for the hyper-parameters $\boldsymbol{\Omega}$, $\boldsymbol{U}$, and $\boldsymbol{W}$, are all from the Inverse-Gamma distribution, where the first element in each prior distribution is the shape parameter, and the second the scale parameter: $\underline{\sigma}^2_{i,w} \sim IG(\underline{T}^w, \kappa^2_w)$ where $\underline{T}^w = 8000$ and $\kappa_w = 0.003$ for $i = 2, \ldots, N^d$; $\underline{\sigma}^2_{i,u} \sim IG(\underline{T}^u, \kappa^2_u)$ where $\underline{T}^u = 100$ and $\kappa_u = 0.1$ for all $i = 1, \ldots, N^d$; $\underline{\sigma}^2_{\omega_j} \sim IG(\underline{T}^{\omega_j}, \kappa^2_{\omega_j})$ where $\underline{T}^{\omega_j} = 1000$ for $j = \{q, m, d\}$, and $\kappa_{\omega_q} = 0.003$, and $\kappa_{\omega_m} = \kappa_{\omega_d} = 0.1$. In sum, as the full sample size $T = 8769$ observations, these priors are very informative for the variance terms associated with the time-varying factor loadings, but less so for the other parameters. Note, however, that the prior variance associated with the quarterly cumulator variable error term, $\sigma^2_{\omega_q}$, is assumed to be considerably lower than the other variance terms.

Finally, to draw the latent threshold, $\boldsymbol{d}$, using the algorithm described in Appendix E.2, the K parameter needs to be defined. K controls our prior belief concerning the marginal sparsity probability. For example, assuming that a time-varying parameter follows $B_t \sim N(0, v^2)$, and marginalizing over B_t, it can be shown that $Pr(|B_t| = 0) = 2\Phi(\frac{d}{v}) - 1$, where Φ is the standard normal CDF. Defining $K = \frac{d}{v}$ as the standardized scaling parameter with respect to the threshold, it can be seen that $K = 3$ implies a marginal sparsity probability exceeding 0.99. As described in Nakajima and West (2013), a neutral prior will support a range of sparsity values in order to allow the data to inform on relevant values, and they suggest that setting $K = 3$ is a reasonable choice.[26] However, in contrast to Nakajima and West (2013), where the time-varying parameters follows AR(1) dynamics, the time-varying factor loadings in (28) follows independent random walk processes. The random walk is non-stationary, and does not have a marginal distribution. For this reason I have experimented with estimating the model using different values for K, finding that higher values for K, coupled with the rather tight priors for the variance of the factor loadings, results in worse model performance (in terms of ROC and AUROC scoring). Accordingly,

[26] Note that when combined with the priors over the other hyper-parameters in the model, the implied marginal prior for each threshold will not be uniform (see Nakajima and West (2013) for details).

Table 4. Convergence statistics. The *AutoCorr* row reports the 10th-order sample autocorrelation of the draws, the *RNE* row reports the relative numerical efficiency measure, proposed by Geweke (1992), while the *IRL* row reports the i-statistic, proposed by Raftery and Lewis (1992). For each entry we report the mean value together with the minimum and maximum value obtained across all parameters in parentheses.

	Parameters					
Statistic	$\boldsymbol{U}$	$\boldsymbol{\Omega}$	$\boldsymbol{F}$	$\boldsymbol{P}$	$\boldsymbol{W}$	$\boldsymbol{d}$
AutoCorr	−0.0 (−0.1,0.1)	0.1 (0.1,0.1)	0.1 (−0.1,0.1)	−0.0 (−0.1,0.1)	0.0 (−0.0,0.1)	0.0 (−0.2,0.1)
RNE	1.1 (0.3,1.9)	0.3 (0.2,0.5)	0.3 (0.7,1.7)	1.0 (0.6,1.6)	0.7 (0.4,1.0)	0.9 (0.4,1.5)
IRL	1.9 (1.9,1.9)	1.4 (1.4,1.4)	1.4 (1.2,1.2)	1.2 (1.4,1.4)	1.0 (1.0,1.0)	1.0 (1.0,1.0)

$K = 0.05$, in the estimations conducted in this analysis.

E.7 Convergence of the Markov Chain Monte Carlo Algorithm

Table 4 summarizes the main convergence statistics used to check that the Gibbs sampler mixes well. In the first row of the table the mean, as well as the minimum and maximum, of the 10th-order sample autocorrelation of the posterior draws is reported. A low value indicates that the draws are close to independent. The second row of the table reports the relative numerical efficiency measure (RNE), proposed by Geweke (1992). The RNE measure provides an indication of the number of draws that would be required to produce the same numerical accuracy if the draws represented had been made from an i.i.d. sample drawn directly from the posterior distribution. An RNE value close to or below unity is regarded as satisfactory. Autocorrelation in the draws is controlled for by employing a 4 percent tapering of the spectral window used in the computation of the RNE. The last row, labelled IRL, reports the mean of the i-statistic. This statistic was proposed by Raftery and Lewis (1992). In essence, it measures the ratio of two other statistics: the total number of draws needed to achieve the desired accuracy for each parameter, and the number of draws that would be needed if the draws represented an i.i.d. chain, see Raftery and Lewis (1992) for details.[27] Values of IRL exceeding 5 indicate convergence problems with the sampler.

As can be seen from the results reported in Table 4, the sampler seems to have converged. That is, the mean autocorrelations are all very close to zero, and the minimum or maximum values obtained seldom exceed 0.1 in absolute value. Moreover, the mean RNE statistic does not exceed unity by a large margin for any of the parameters. Finally, the

[27] The parameters used for computing these diagnostics are as follows: quantile = 0.025; desired accuracy = 0.025; required probability of attaining the required accuracy = 0.95.

Table 5. Estimates and true DGP parameters. The numbers reported in parenthesis are standard deviations. See also Figure 10.

	$\boldsymbol{F}_1$	$\boldsymbol{\Omega}_{3,3}$	$\boldsymbol{U}_{3,3}$	$\boldsymbol{U}_{4,4}$	$\boldsymbol{U}_{5,5}$	$\boldsymbol{U}_{6,6}$
Estimated	0.99 (0.01)	0.48 (0.01)	1.29 (0.02)	1.05 (0.02)	0.99 (0.02)	1.00 (0.02)
True	0.99	0.5	1	1	1	1

IRL statistics are always well below 5. Additional convergence results can be obtained on request.

E.8 A simulation experiment

To control the estimation procedure, and verify the code, I run a simple simulation experiment. Artificial data is generated from a data generating process like the one described in Appendix E, with $T = 8769$ daily observations. $N_q = 1$, $N_m = 1$, and $N_d = 8$, such that $N = 10$. Quarterly and monthly observations are attributed across some generic year, quarters and months, such that the artificial sample contains roughly 100 and 300 observable quarterly and monthly observations, respectively.

Hyper-parameters used to simulate the data are set as follows: All diagonal elements in $\boldsymbol{W}$, $\boldsymbol{U}$, and $\boldsymbol{\Omega}$ are set to 0.001, 1, and 0.5, respectively. The threshold parameter $\boldsymbol{d}$ is set equal to 0.3 for all of the time-varying factor loadings. The autoregressive process for the law of motion for the latent daily factor, a_t, is specified with one lag and $\Phi = 0.99$. The autoregressive processes for the idiosyncratic errors are specified with $\Phi_i = 0$ for $i = 1, \ldots, N^d$. Finally, $z_j = 1$ for $j = \{q, m\}$, and the latent state variables in $\boldsymbol{A}_0$ and $\boldsymbol{Z}_0$ are initialized at zero. The prior specifications used for estimation are in expectation all set equal to the true values, but for neither specification is the degrees of freedom parameters set higher than 100.

Figure 10 reports the estimated latent daily factor alongside the simulated factor. As is clearly seen in the figure, they are very close to being perfectly correlated. In the figure four of the estimated time-varying factor loadings, together with their simulated counterparts, are also illustrated. Again, the estimated and simulated processes are very similar. As seen from the figures, the estimation procedure is also capable of identifying the true threshold value with a large degree of precision. Table 5 reports the posterior median and standard deviation of the parameter estimates for Φ_1, $\boldsymbol{\Omega}_{3,3}$, and $\boldsymbol{U}_{i,i}$ for $i = \{3, 4, 5, 6\}$. All estimates are precisely estimated and very close to their true values.

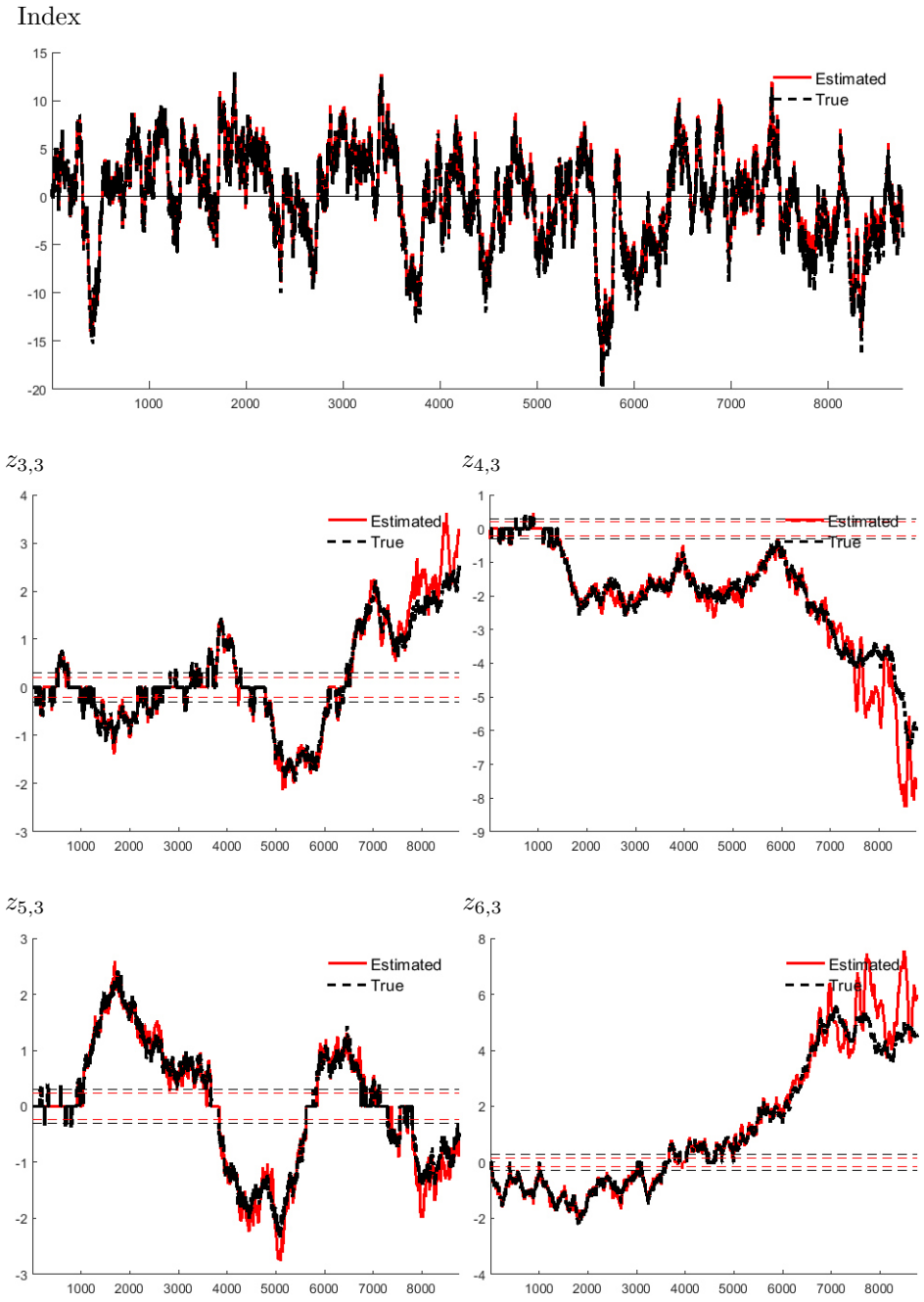

Figure 10. The upper graph reports the simulated daily index together with its estimate. The subsequent graphs report the simulated factor loadings for daily observations 3 to 6, together with the true thresholds (d) and the estimated loadings and thresholds.