Nowcasting GDP using tone-adjusted time varying news topics: Evidence from the financial press[*]

Dorinth van Dijk[†] and Jasper de Winter[‡]

February 17, 2023

Abstract

We extract tone-adjusted, time-varying and hierarchically ordered topics from a large corpus of Dutch financial news and investigate whether these topics are useful for monitoring the business cycle and nowcasting GDP growth in the Netherlands. The financial newspaper articles span the period January 1985 up until January 2021. Our newspaper sentiment indicator has a high concordance with the business cycle. Further, we find newspaper sentiment increases the accuracy of our nowcast for GDP growth using a dynamic factor model, especially in periods of crisis. We conclude that our tone-adjusted newspaper topics contain valuable information not embodied in monthly indicators from statistical offices.

Keywords: Factor models, topic modeling, nowcasting

JEL classification: C8, C38, C55, E3.

[*] We gratefully acknowledge comments from Olaf Sleijpen, Peter van Els, Maurice Bun and participants of the 2021 Joint Statistical Meeting and research seminars at the European Commission Joint Research Centre, the MIT Center for Real Estate, Atradius Netherlands, the Dutch Foreign Bankers' Association, Ortec Finance, the AFM (Autoriteit Financiële Markten) and the Dutch Central Bank (De Nederlandsche Bank). We thank Eugenie van Wiechen, Edwin Fillekes and Mathijs Rotteveel of Het Financieele Dagblad and Olaf van der Wiel of De Nederlandsche Bank for useful discussions and their help to unlock the digital archive of Het Fiancieele Dagblad. The views expressed herein are those of the authors and do not necessarily reflect the position of De Nederlandsche Bank. All remaining errors are our own.

[†] Economics and Research Division, De Nederlandsche Bank – d.w.van.dijk@dnb.nl.

[‡] Corresponding author, Economics and Research Division, De Nederlandsche Bank – j.m.de.winter@dnb.nl.

1 Introduction

Most short-term forecasting models underestimated the depth of the Covid-19 crisis, partly because most monthly economic indicators are published with a considerable time lag. Practitioners and economists began to look for other, coronavirus related and new fast-moving indicators to improve short-term forecasting models (see e.g. De Smedt and Daelemans, 2021 and Barbaglia et al., 2022). The use of these big data sources for short-term forecasting is gaining popularity recently. One relatively new source of data is text, and especially texts in newspaper articles (see e.g. Gentzkow et al., 2019, Ardia et al., 2019, Thorsrud, 2020, Kalamara et al. (2022) and Barbaglia et al., 2022). Recent advances in machine-learning enable extraction of sentiment indicators and topics from these texts. A key question is whether this newspaper sentiment would have helped to produce a more accurate forecast during the Covid-19 crisis.

The signal and the gains in forecasting accuracy from using textual data is the main topic of this research. We explore two related issues: First, we explore if tone-adjusted news topics are good indicators for measuring the business cycle. Second, we investigate whether the tone-adjusted news topics add forecasting power to state-of-the art nowcasting models, focusing on short-term forecast of the quarterly growth rate of gross domestic product (GDP).

We use a novel newspaper corpus, spanning the period January 1st 1985 up until January 18th 2021, with articles published in the largest and only Dutch financial newspaper *Het Financieele Dagblad*. In total, we analyze around one million news articles. Given the relatively long time-period we are able to isolate periods of crisis (Global Financial Crisis and the Covid-19–crisis) and more tranquil times. The articles are written in Dutch, and we analyze them without translating them into English, keeping the nuances specific to the Dutch language. We construct a novel dictionary, stemming technique en sentiment list tailored to Dutch financial news articles.

There is a growing literature on Bayesian estimation of topic models and nowcasting. The most popular variant, Latent Dirichlet Allocation (LDA) was first introduced by Blei et al. (2003). The paper started a growing literature in the machine-learning field, inspiring a flurry of extensions to the base model. See, amongst others, Churchill and Singh (2022) and Chauhan and Shah (2021) for recent surveys. Up until recently however, its use in economics has been quite limited. With the increasing availability of large text and newspaper corpuses and the introduction of machine-learning methods economic applications are becoming more widespread. Hansen et al. (2018) was one of the first to use the LDA model, examining the effect of transparency on the deliberation of the Federal Open Market Committee (FOMC) of the Board of Governors of the Federal Reserve System. The FOMC transcripts were also used to extract financial stability concerns (Wischnewsky et al., 2021), amongst others. The LDA model was later applied to a corpus of financial newspaper articles by, amongst others, Thorsrud (2020) and Larsen and Thorsrud (2018). Recent contributions in the nowcasting literature combine sentiment and uncertainty extracted via pre-defined lists from newspaper texts (e.g. Shapiro et al., 2022 and Gentzkow et al., 2019) and incorporate these in nowcasting models (e.g. Aprigliano et al., 2022, Barbaglia et al., 2022, Bybee et al., 2020 and Rambaccussing and Kwiatkowski, 2020).

Our research adds to the existing literature in several dimensions. Firstly, we introduce a new variant of the plain-vanilla LDA model, the tone-adjusted time-varying layered topic model. Here, we subdivide news into time-varying hierarchical topics with their sentiment. Subsequently, we check if the resulting sentiment indicator is a good indicator for the stance of the Dutch business cycle. Secondly, we construct a novel Dutch economic dictionary specifically tailored to measure the sentiment of economic news regarding the Dutch economy. We enrich our approach with valence shifters that are specific to the Dutch language. Thirdly, we introduce a parsimonious Bayesian estimation scheme to easily estimate our topic model, using the posterior distributions of previous time slices and layers as priors for estimation of new time slices and deeper layers. Finally, we add to the literature on topic models and nowcasting, by including our novel tone-adjusted topics in a pseudo real-time out of sample forecast comparison between a state-of-the art nowcasting model with and without several variants of tone-adjusted newspaper topics.

The remainder of the paper is organized as follows. Section 2 describes the data sets we used in our analysis, i.e. the corpus of newspaper articles and a data set of macro-economic indicators. Section 3 describes the steps needed to construct our tone-adjusted time-varying layered topic model. Section 4 describes the details of our nowcasting exercise. Section 5 describes the main outcomes of our topic model. Section 6 presents the outcomes of the nowcasting exercise. Section 7 concludes.

2 Data

This section describes the data we used in the estimation of our tone-adjusted time-varying topic model and our nowcasting exercise, respectively. Section 2.1 describes the data set of newspaper-articles and the vocabulary used to estimate of the time-varying layered topic model. Section 2.2 describes the data set of monthly macro-economic indicators that we use in the nowcasting exercise, to determine the value added of using tone-adjusted newspaper topics.

2.1 Corpus of newspaper articles

Raw database

Our source of textual data is a full database of the only and largest financial newspaper of the Netherlands, *Het Financieele Dagblad* (FD). The database contains all published articles in the newspaper (both in print and online) for the period January 1st 1985 up until January 18th 2021. The raw database contains a total of $1,093,477$ articles. We have the complete text of each article, the article title, the publication URL, the publication date, the section the article was published in and one or more one-word tags describing the article content. Part of the articles database includes opinions by policy makers, plans by government and Parliament and news items on topics not directly related to the economy. We use these attributes to clean the database for non-relevant articles that have a direct relation to economic developments. The cleaning process is carried out in several steps, described in Section A.1 of the Appendix. After cleaning, we end up with $582,981$ articles, a reduction of approximately 47% compared to the raw database. Below, we describe the steps to prepare the text of the remaining articles in the database for use in our topic model. In short, we remove stopwords, stem all words and prune the total number of words to a vocabulary so it can be used in our topic model.[1]

Stopwords

Moreover, we delete very common words which add little value to understanding the meaning of an article, so-called stopwords, like "the", "a" and "and", count words, e.g. "duizend", "miljoen", months and days of the week. We compile a list of these stopwords by synthesizing the Dutch stopwords list in the R-package `snowballC` with stopwords that often appear in our corpus of articles. We do *not* delete stopwords that express sentiment, like "nothing" or " less" because these words are included in our sentiment list (see Section 3.4).

Collocations

An issue with analyzing texts is that when two (or more) words naturally belong together, this is not automatically understood by the topic model. In order to uncover the most relevant

[1] Our list of Dutch stopwords and transformation from conjugate verbs to their verbstem is publicly available and free to use, and can be downloaded here and here, respectively.

so-called *bi-grams* we analyzed all single words and bi-grams with a minimum frequency of $4,000$ in the newspaper. In most cases bi-grams relate to word-combinations that do not have any special meaning like “their customers”, “red car”. However, some bi- and tri-grams do have a special meaning, such as company names, like “Royal Dutch Shell”, “London stock exchange", “Thomas Cook”, “Standard & Poor′s” and two or more words that have a different meaning when combined, e.g.: “convertible stocks” , “financial markets” , “private equity” , “Statistics Netherlands” , “euro crisis”, “industrial production”, “current account”, “PMI index”, “interest rate” and “Central Europe”.

Stemming

Next, we transform all conjugate verbs and words to their stem. For example, the verbs “is”, “be”, “are” are all reduced to the stem “are”. This step reduces the number of unique words in the corpus, without loss of meaning of the words. Usually this is done with a mechanical so-called Porter stemmer (Porter, 1980). We experimented with this stemming technique, but concluded that it yields unsatisfactory results in Dutch. Therefore, we follow another –more labor-intensive– process. We proceed as follows. Firstly, we stem all Dutch verbs using the verb list in the web-mining **Python**-module `Pattern` and augment the list with verbs that are specific to financial news. Our lists of conjugate and stemmed verbs contains $20,058$ and $3,687$ words, respectively. Secondly, we *manually* stem all nouns with a total frequency of $2,000$ or higher. Furthermore, we check for synonyms, and replace words that are clearly synonyms, e.g. the Dutch language has two words for global: “globaal” and “mondiaal”. To keep our list of words parsimonious we combine these words into the word “mondiaal”.

Vocabulary

The cleaned corpus of newspaper articles contains $1,287,851$ unique word tokens, which makes statistical computation very challenging. Moreover, a topic-model with that many tokens risks being severely over-fitted, and does not generalize well to hold-out data. Following, amongst others, Thorsrud (2020) and Barbaglia et al. (2022) we take several steps to reduce the raw data set. First, we set a minimum and a maximum number of articles a token should appear in, based on the complete database of articles. We set the minimum number of documents a word appears in at least 0.1% of all newspapers, so words with a very low frequency are not included in the vocabulary. To clean the vocabulary of very common words, we only include words that occur in maximum 50% of the articles.[2] This so-called "pruning" decreased the number of unique word-tokens from $1,287,851$ to $9,613$, a reduction of > 99 percent. Second we check all words and exclude verbs, adjectives, count-words (million, thousand), and keep the nouns. The idea here is that we only "catch" the main topic of the sentence, and the noun is the most valuable word in this respect. We check for synonyms and transform all words to their singular form. After this second "pruning" step we have a final list of $2,135$ nouns in our vocabulary.

2.2 Data set of macro-economic indicators

In our nowcasting exercise as described in Section 4, we combine the extracted topics from the corpus of newspaper articles with a monthly data set of macro-economic indicators. The data set of macro-economic indicators consists of 70 monthly time series and quarterly GDP that were downloaded on February 1th 2021. The statistical monthly information set reflects the public knowledge at the beginning of the month, and covers a broad range of information readily available to economic agents. The indicators fall into five categories. The first category is hard, quantitative information on production and sales, such as industrial production in various sectors, retail trade turnover, household consumption, world trade and unemployment. The second category is soft, qualitative information on expectations derived from surveys among consumers and firms. The third category contains financial variables, both quantities (money stock and credit volume) and prices (interest rates and stock prices). These determine financing conditions for firms and consumers. Moreover, financial market prices partly reflect financial market expectations on output developments in the near future. The fourth category refers to input and output prices, i.e. headline consumer and producer prices, and world market commodity prices. The fifth category contains information on the development of soft and hard indicators for the main trading partners of the Netherlands, i.e. the United Kingdom, Germany, France, Italy, Spain and Belgium. These indicators are potentially important for a small open economy such as the Netherlands.

Section A.2 in the Appendix provides details on the sources and availability of the data series. The available monthly data are usually already adjusted for seasonality (and calendar effects). Where necessary, raw data series are seasonally adjusted using the US Census X12-method. All monthly series are made stationary by differencing or log-differencing (in the

[2] An automated variant of this approach would be to use the *term frequency–inverse document frequency*. We experimented with this automated approach, but in our case the more labor intensive manual approach delivered a more meaningful vocabulary.

case of trending data, such as industrial production, retail sales and monetary aggregates). All variables are standardized by subtracting the mean and dividing by the standard deviation. This normalization is necessary to avoid overweighting series with large variances in the determination of common factors in our nowcasting model (see Section 4). The data transformations are the same for all estimated models. In the nowcasting exercise, these 70 macroeconomic indicators are combined with the appropriately transformed tone-adjusted topics extracted from the tone-adjusted time-varying layered topic model described in Section 3. All tone-adjusted topics are differenced and normalized before inclusion in the nowcasting model.

3 Tone-adjusted time-varying layered topic model

The LDA model has emerged as a powerful tool to analyze document collections in an unsupervised fashion. We extend the plain-vanilla LDA (Blei et al., 2003) model by introducing time-variation in the topic-content and introduce hierarchy in the extracted topics. Section 3.1 lays out the groundwork, by describing the working of the plain-model underlying our tone-adjusted time-varying layered topic model. The following sections extend the base-model, i.e.: time-variation in Section 3.2, layering in Section 3.3 and tone-adjustment in Section 3.4. Section 3.5 describes the Bayesian inference algorithm used to infer the model parameters.

3.1 Latent Dirichlet Allocation: base model

The main idea of LDA is that each document is a part of a probability distribution over topics, and each topic is part of a probability distribution over words. The model generates automatic summaries of topics in terms of a discrete probability distribution over words for each topic, and further infers per-document discrete distributions over topics. Most importantly, LDA makes the implicit assumption that each word is generated from an underlying topic.

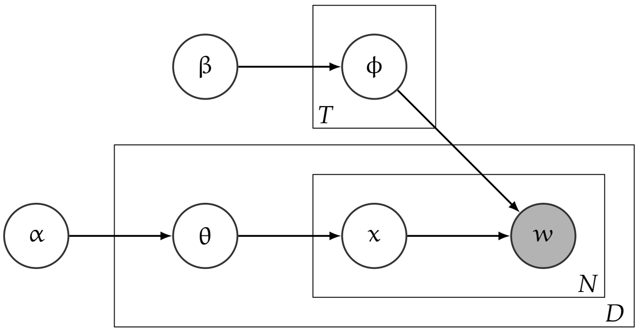
Define a document is a sequence of N words denoted by $\mathbf{d} = (w_{d1}, \ldots, w_{dn})$, where w_{dn} is the nth word in the sequence of document $\mathbf{d}$. A corpus is a collection of D documents denoted by $\mathcal{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_D\}$. Each document is composed of T topics. LDA assumes the following generative process for each document $\mathbf{d}$ in a corpus $\mathcal{D}$:

1. For each topic $t = 1, \ldots, T$,
   - Draw a distribution over words from a Dirichlet distribution with hyperparameter β, i.e: $\phi_t \sim \mathrm{Dir}(\beta)$.

2. For each document, **d**,
   - Draw a vector of topic proportions from a Dirichlet distribution with hyperparameter α, i.e: $\theta_d \sim \text{Dir}(\alpha)$.
   - For each word w_{dn}:
     (a) Draw a topic assignment x_{dn} for word $w_{dn} \sim \text{Mult}(\theta_d)$, $x_{dn} \in \{1, \ldots, T\}$;
     (b) Draw a word w_{dn} from the $\sim \text{Mult}(\phi_t)$, where t is the drawn topic assignment in the previous step.

The generative probabilistic process with repeated sampling described above, can be conveniently illustrated using plate notation. In this graphical notation, shown in Figure 1, shaded variables indicate observed variables, in our case the words (w). Latent, or unobserved, variables are unshaded. Arrows indicate conditional dependencies between variables, while plates (the boxes in Figure 1) refer to repetitions of sampling steps with the variables in the lower right corner referring to the number of samples. For example, the inner plate over topic x and w illustrates the repeated sampling of topics and words until N words have been generated for document **d**. The plate surrounding θ illustrates sampling over topics for each document **d** for a total of D documents. The plate surrounding ϕ_t illustrates the repeated sampling of word distributions for each topic until T topics have been generated. Hyperparameters α and β determine the shape of the Dirichlet distribution. All subscript for a variable on a plate carry-over to all variables on that plate. For example, w in Figure 1 equals w_{dn} because it is located within the plates D and N.

Figure 1: The graphical model for the topic model using plate notation



3.2 Extension 1: Time-variation

The first extension we propose to the base model above is time variation. In our "time-varying" topic model, we use the estimated topic-word distribution for a time slice as initialization for the estimation of the topic-word distribution in the next time slice. We do this to circumvent a flaw of regular topic models in real-time forecasting competitions. Usually, topic models are estimated over the *entire* sample period. Most of the papers in the recent literature are not using a time-varying LDA approach (see e.g. Larsen and Thorsrud, 2022, Larsen and Thorsrud, 2018, and Bybee et al., 2020). This process implicitly assumes that the documents are drawn from the same (time-invariant) set of topics. Hence, *not* taking into account time slices and estimating one topic model over the whole time period leads to misspecification of the topic model. In case of the word "virus", this would lead to overestimating its importance in the pre-coronavirus period and underestimating its importance during and after 2020. For our newspaper corpus, however, the order of the documents reflects an evolving set of topics, i.e the content of topics in 2022 will be different from the same topics in 1985.

Figure 2: The graphical model for the time-varying topic model using plate notation

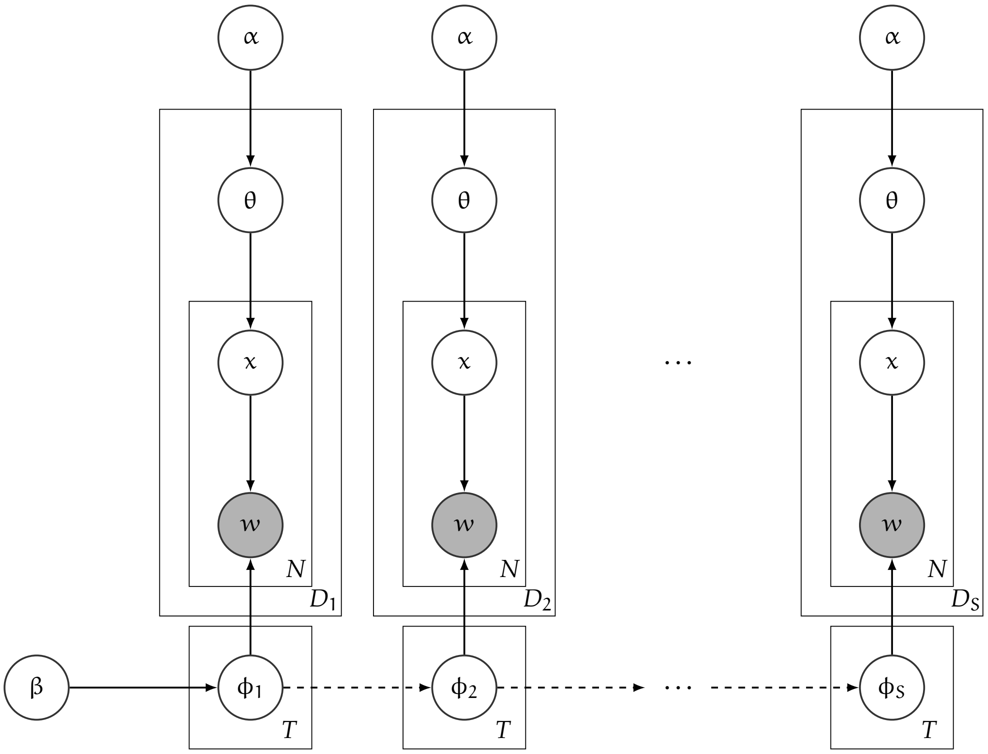


Figure 3 shows the repeated sampling process for the first time slice. In later time slices β will be replaced by the estimated ϕ in the previous time slice. We introduce the time slice subscript $s = 1 \ldots, S$, such that $D_1 \ldots, D_S$ constitute different time slices of documents. In our time-varying topic model, we divide the newspaper articles in *monthly rolling* time slices of 15 years. In our application we have newspaper articles over the period January 1st 1985 up until January 18th 2021, so our first time slice ($s = 1$) contains articles published in the period January 1st 1985 up until January 1st 2000. The second time slice consists of articles published in the period February 1st 1985 up until February 1st 2000, and so on until our last time slice

($s = S$), which contains articles published in the period January 1st 2006 up until January 1st 2021. $\phi_1 \ldots, \phi_S$ indicate the topic-word distribution in time slices $s = 1 \ldots, S$.

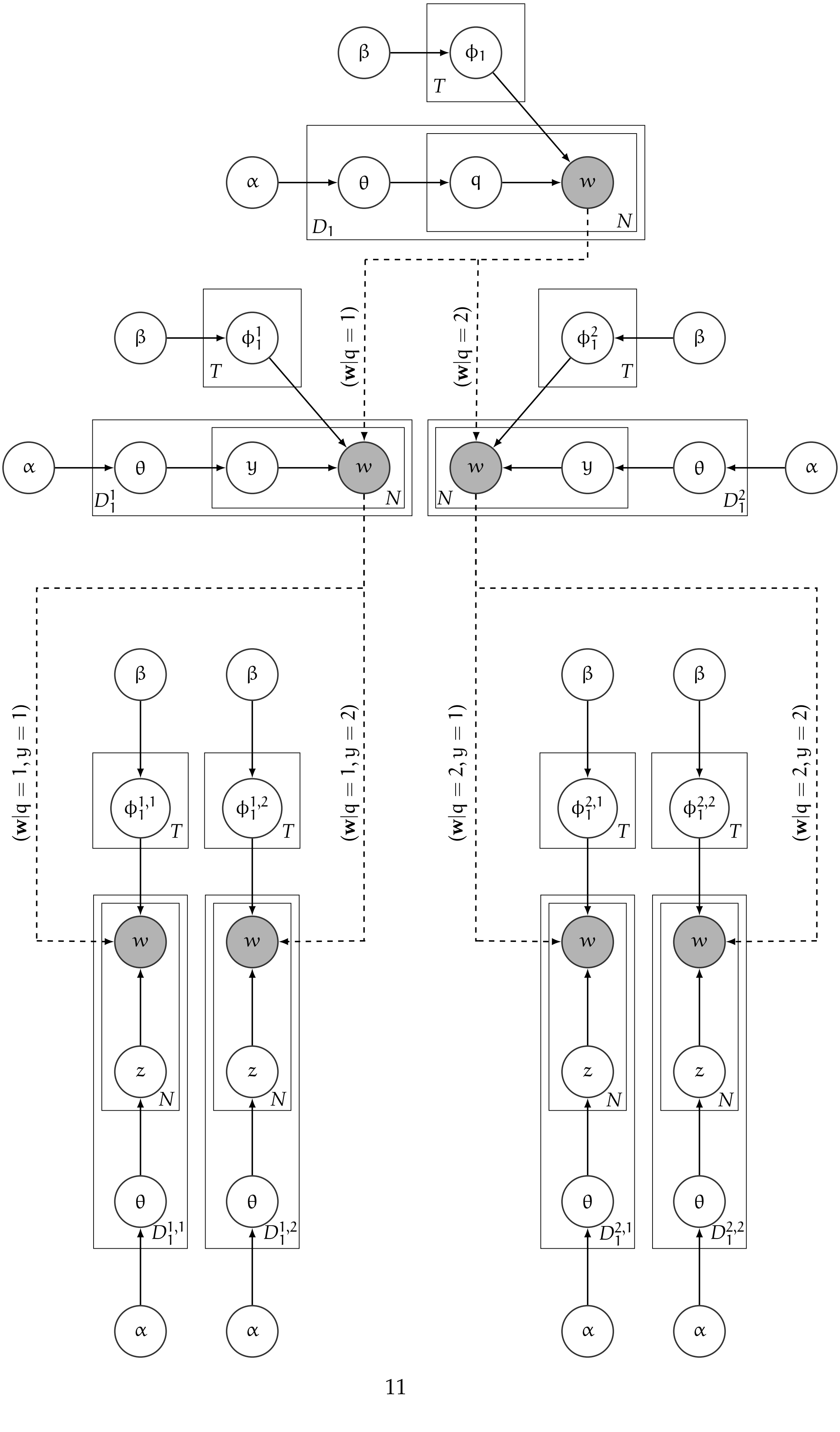
Figure 2 shows the generative probabilistic process with repeated sampling. dashed arrows indicate that the variable is used in a subsequent time slice. Note, that the time slice subscript for a variable on a plate carry-over to all variables on that plate as in Figure 1, For example, the variable w in the plate above ϕ_1 equals $w_{d_1 n}$ because it is located within the plates D_1, indicating the first time slice, and N, indicating that the repeated sampling process is done for each word in each document in time slice 1. The modeling of the dynamics in our topic model is related to the class of dynamic topic models (e.g. Blei and Lafferty, 2006). The main difference with dynamic topic models is that we do not make any explicit assumptions on the dynamics of the topic-word distributions and re-estimate the model each time slice, instead of one estimation with time-varying word distributions within topics. Our method therefore is more flexible, and can more easily accommodate shifts in the vocabulary. The way we model dynamic is most closely related to the rolling topic model Bittermann and Rieger (2022). The main difference is that we estimate overlapping time slices instead of non-overlapping adjacent time slices, and we explicitly use the estimates of the Gibbs sampler in time slice $t - 1$ as starting values in the Gibbs sampler in time slice t, which leads to more stable topics.

3.3 Extension 2: Hierarchy

Besides time variation, we also include three layers in our model. For exposition purposes, we first simplify our model to the plain vanilla LDA model. Next, we assume that we can further sub-divide the estimated topics in more detailed topics. The layers in the topic model are indicated as follows: $D^{\text{layer1...,current layer}}$. For example $D^{1,2}$ indicates all words in documents assigned to topic 1 in the first layer, and topic 2 in the second layer. To keep notation parsimonious, we refer to the topic assignments in layer 1, 2 and 3 as q, y and z, respectively.

Figure 3 shows a stylized example for a layered topic model with three layers, with two extracted topics in each layer. In the second layer we estimate separate LDA models based on a subset of words in articles assigned to topic 1 and 2 in the first layer. This is indicated by the dashed arrows with the labels $(\mathbf{w}|q = 1)$ and $(\mathbf{w}|q = 2)$. Let's call the first layer topics “economics” and ”politics” for illustrative purposes. Our layered topic model aims to further detail the topic into two separate topics, both within “economics” as well as “politics”.

Figure 3: Plate notation for a stylized layered topic model with three layers



Let's name these topics "indicators" $(\mathbf{w}|q = 1, y = 1)$ and "euro area" $(\mathbf{w}|q = 1, y = 2)$ for economics and "parliament" $(\mathbf{w}|q = 2, y = 1)$ and "social partners", $(\mathbf{w}|q = 2, y = 2)$ for politics. The third and final layer of the layered topic model splits these four topics into 8 extracted topics, i.e: $(\mathbf{w}|q = 1, y = 1, z = 1, 2)$, $(\mathbf{w}|q = 1, y = 2, z = 1, 2)$, $(\mathbf{w}|q = 2, y = 1, z = 1, 2)$ and $(\mathbf{w}|q = 2, y = 2, z = 1, 2)$. Note that the time slice subscript and the superscript for the layers for a variable on a plate carry-over to all variables on that plate as in Figure 1 and Figure 2. The layering we propose is different from the canonical hierarchical topic models of Griffiths et al. (2003). In that model the hierarchy stems from the correlation between topics. We explicitly enforce hierarchy and estimate the model in separate layers.

3.4 Extension 3: Tone-adjustment

Using a measure of sentiment, we can "tone-adjust" the topics in our model. This enables us to redistribute the headline sentiment of the newspaper on a certain day to the identified topics. Following Thorsrud (2020), we start from article-level sentiment and aggregate to eventually obtain headline sentiment. We proceed as follows. First, we calculate sentiment per article by taking the sum of the sentiment scores of the words and dividing by the number of words per article.[3] Next, we assign this sentiment to topics according to the estimated θ per article for all third-layer topics. From the topic proportions in the third layer of our model, we are able to construct tone-adjusted topics in the second and first layer. From the topic-adjusted sentiment in the first layer, we can construct headline sentiment. Note that this calculated headline sentiment is equivalent to calculating headline sentiment by using all articles without this aggregation procedure.

We use a dictionary based technique to construct our measure of newspaper sentiment, following Tetlock (2007) and Loughran and McDonald (2011). The basis of our sentiment measure is the sentiment list introduced Loughran and McDonald. We translate this list to Dutch using two freely available online translation tools, i.e "Google Translate" and "DeepL" . We manually check the translations, and when there are more options to choose from, we choose the most appropriate translation for our use case, or added both terms when appropriate. Besides, we add words and collocations to the sentiment-list that are often used in the Dutch financial press, e.g. "begrotingstekort" (budget deficit), "laagconjunctuur" (economic

[3] We tested several variants of the weighing of our newspaper sentiment score, i.e. weighing by the number of sentiment words per article or weighing by both the number of total and sentiment words. Overall, the results are comparable tot our chosen measure, but our measure resulted in a better fit with the year-on-year growth of GDP; our measure for the stance of the business cycle. Results available upon request with the authors. See Algaba et al., 2020 for a comprehensive treatment on the measurement of sentiment.

downturn) as negative terms, and “begrotingsoverschot” (budget surplus) and “hoogconjunctuur” (economic boom) as positive terms. Moreover, we include word-combinations that reverse the meaning of sentiment words, so-called negations. For instance, “unemployment increased” would be scored as positive because increased is a positive word, but the meaning of increased in this particular case is clearly a negative sign for economic development. This is also true for negations such as “fiscal deficit increased”, and “no increase”. The total augmented Loughran and McDonald dictionary contains $1,672$ words and collocations.[4]

Combining the techniques described in the previous sections on time-variation, layering, and tone-adjustment constitute our tone-adjusted time-varying layered topic model, i.e. we extend the time-varying nature to all layers of the model. In short, the estimation of the successive slices equals the (median of) the estimated posterior topic-word distribution from the previous time slice. In our base model we estimate three layers, one more layer than the model described in the stylized example above and in Figure 3.

3.5 Bayesian inference

The topic model we propose at its core is still a conventional LDA. The key idea of our approach is to re-estimate this LDA model for each layer and time slice separately, using the outcome of the previous layer and time slice as initial value for the Gibbs sampler. The procedure for inference of our model-coefficients is as follows: For the first layer of the first time slice, running from January 1st 1985 up until January 1st 2000 we estimate $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ using the collapsed Gibbs sampling procedure, which is explained in more detail in Section B in the Appendix. We re-estimate the topic model each month. We construct newspaper time slices with a *monthly rolling* as described in Section 3.2. In total we have 253 time slices.[5] For the 2nd to last time slice of the first layer, we re-estimate the model using the posterior estimate of $\boldsymbol{\phi}$ of the previous slice of the month to initialize the count matrix in the Gibbs Sampler.

The aim of this “chain of count initializiatons” is to stabilize the topics, i.e. the extracted topics have a high probability of having the same ordering in each time slice and therefore a more stable topic-word distribution. To clarify this point, let’s assume for the sake of exposition that we extract two topics, and after $4,000$ Gibbs iterations the word “inflation” has a count of 500 in topic 1 and 10 in topic 2. If we were to use this estimate as our final estimate of

[4] Our list of Dutch sentiment words and their sentiment is publicly available and free to use, and can be downloaded here.

[5] We experimented with smaller time windows, i.e. 5 and 10 year but found the shorter time windows entailed a costs in terms of less interpretable and more volatile topics. The results of the estimations with these shorter time windows are available upon request with the authors.

the topic-word distribution in slice t and use the posterior estimate of ϕ of the first time slice to initialize the count matrix in the Gibbs sampling algorithm in time slice $t + 1$, the count of the word "inflation" in topic 2 needs to increase by a large amount in order to become more prominent in topic 2 than in topic 1. This can only be the case when the data provide a very strong signal that the word "inflation" has to shifted to topic 2. Our approach is therefore different from random initialization, where the word "inflation" is randomly assigned to topic 1 and 2 in the first iteration *not* taking into account the counts in the first time slice. The stability of the topics over time is verified by calculating the cosine distance of the topics, see e.g. Newman et al. (2010) and Aletras and Stevenson (2014). The minimum distance for the same topic numbers in t and $t + 1$ is always higher than 0.95. The distances with other topics numbers are always smaller. A number of 1 (0) indicates to topics are very similar (dissimilar). This indicates that the ordering of topics has not shifted from period to period.[6] It is crucial that the topics are stable terms of ordering, because we aim to construct time series per topic.

After estimation of θ and ϕ for each time slice in the first layer of the topic model, we assign every word w to a specific topic using the assignment in the median of the posterior distribution of ϕ. Next, we assign every word that is assigned to topic 1 to the first "branch" in the second layer (see Figure 3). We do the same for topics $2-4$ and assign the words belonging to these topics at an article level to "branch" $2 - 4$. For each branch, we again estimate plain vanilla LDA's, in our case four LDA's, one for the words of every topic in the first layer. Since the "branches" are time-varying too, we use random initialization of the Gibbs sampler for the first time slice and use the posterior estimate of ϕ of the first time slice to initialize the count matrix in the Gibbs sampler for the second time slice. For the third layer we make a further split in branches based on the 16 topic assignments in the second layer. Again, we estimate 4 topics per branch following the same procedure as described above. In the end we infer three topic layers, with 4 topics in the first layer, 16 topics in the second layer and 64 topics in the third layer. The first layer indicates the general topic of the article, the second layer gives some more details about the topic and the third layer provides the most granular information. Finally, we attach sentiment to all articles and topics by calculating the sentiment-score for each of the original article texts as described in Section 3.4.

[6] Results available upon request with the authors.

4 Nowcasting model and forecast design

In order to test whether the sentiment indicators extracted from the newspaper have any value for short-term forecasts of GDP growth, we compare nowcasting models with and without the extracted tone-adjusted topics. We use a dynamic factor model (DFM) –the workhorse forecasting model for many central banks and policymakers (see e.g. Bańbura and Rünstler, 2011, Jansen et al., 2016 and Jansen and de Winter, 2018)– without newspaper sentiment as baseline model. The model is described in more detail in Section 4.2 below. The next section describes the specificity's of our forecast design to determine the value added of our newspaper indicators.

4.1 Forecast design

Section C in the Appendix discusses the dynamic factor model and the data set of monthly indicators in more detail, here we provide the most important characteristics. In short, we estimate a DFM based on the specification in Bańbura et al. (2011), and use a data set of 70 monthly economic indicators, comprised of five distinct groups of data. These are production & sales, surveys, financial indicators, prices and indicators of the most important trading partners of the Netherlands. This data set is enriched with time series of the topics resulting from different version of the tone-adjusted time-varying layered topic model described in Section 3. We convert the daily estimated topic models by extracting the sentiment per topic per day and aggregating the daily data to a six-month moving average. The latter is necessary because the newspaper sentiment is extremely volatile at a daily basis.[7]

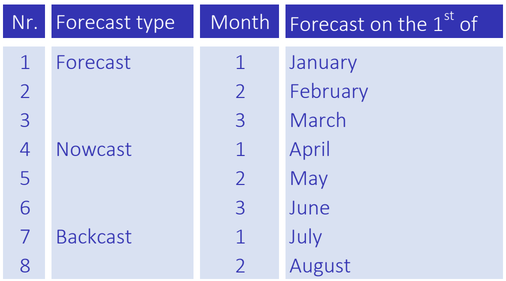
We construct a sequence of eight forecasts for GDP growth in a given quarter, obtained in consecutive months. Table I explains the timing of the forecasting exercise, taking the forecast of a second quarter as an example. We make the first forecast on the January 1st with the monthly data and the time series derived from our topic model that were available at the time. This forecast is called the one-quarter-ahead forecast in month one. We subsequently produce a monthly forecast for the next seven months, through August. The last forecast is made on August 1st, two weeks before the first release of GDP for the second quarter.[8] Following the conventional terminology, *forecasts* refer to (one-quarter) ahead forecasts, *nowcasts* refer to current quarter forecasts and *backcasts* refer to forecasts for the preceding quarter, as

[7] Our choice for a six-month moving average was motivated by its high concordance with the business cycle (turning points). The outcomes of the indicator with longer and shorter moving averages are available upon request with the authors.

[8] Statistics Netherlands publishes the first estimate of GDP growth approximately 45 days after the end of a quarter.

long as official GDP figures are not yet available. The forecast design entails the construction of six consecutive forecasts of the DFM for real GDP growth for each quarter in the period 2003Q3–2020Q3. The start of the estimation period is 1996M1, i.e. no monthly data before 1996M1 are used in the estimation of the models. All monthly indicators were downloaded on February 1st 2022. We start evaluating the forecasts errors of the model in 2003M9, and use an expanding window for the estimation of the model. This implies that final backcast for 2020Q3 on November 1st.

Table I: Timing of forecast exercise for the second quarter



Nr.	Forecast type	Month	Forecast on the 1st of
1	Forecast	1	January
2		2	February
3		3	March
4	Nowcast	1	April
5		2	May
6		3	June
7	Backcast	1	July
8		2	August

4.2 Dynamic factor model

In practice, taking advantage of auxiliary information for the forecasting of real GDP in the short-run poses several challenges. The first challenge is posed by the large size of the information set, also known as the "curse of dimensionality". There are countless potentially useful variables for forecasting GDP. The data sets used in the empirical literature vary greatly in size, and may include more than 300 variables. Moreover, the limited length of the time series involved makes over-parametrization a real issue. The second problem relates to the fact that the indicator variables are observed at a monthly frequency and GDP at a quarterly frequency. Moreover, there may be variation in publication lags due to different release dates. This is known as the "ragged edge" problem.

Dynamic factor models tackle the "curse of dimensionality" by summarizing the information of a potentially large data set in a limited number of factors. The dynamic behavior is specified as a vector-autoregressive process. Another key feature of the model is the use of the Kalman filter which efficiently handles the unbalanced character of the data set and is able to handle differences in frequency. The Kalman filter replaces any missing monthly indicator observations with optimal predictions and also generates estimates of unobserved monthly real GDP, subject to a temporal aggregation constraint for the quarterly observation. Jansen et al. (2016) found in their comparative multi-country study that the DFM had the highest forecasts accuracy on average, in particular for nowcasting and backcasting. Here, we employ the

DFM version proposed by Bańbura and Rünstler (2011), because of it's relatively good forecasting performance. See Hindrayanto et al. (2016) for a elaborate review of the forecasting performance of different dynamic factor model specifications. See Section C in the Appendix for more information on the model equations, the state space representation of the DFM, and modeling choices.

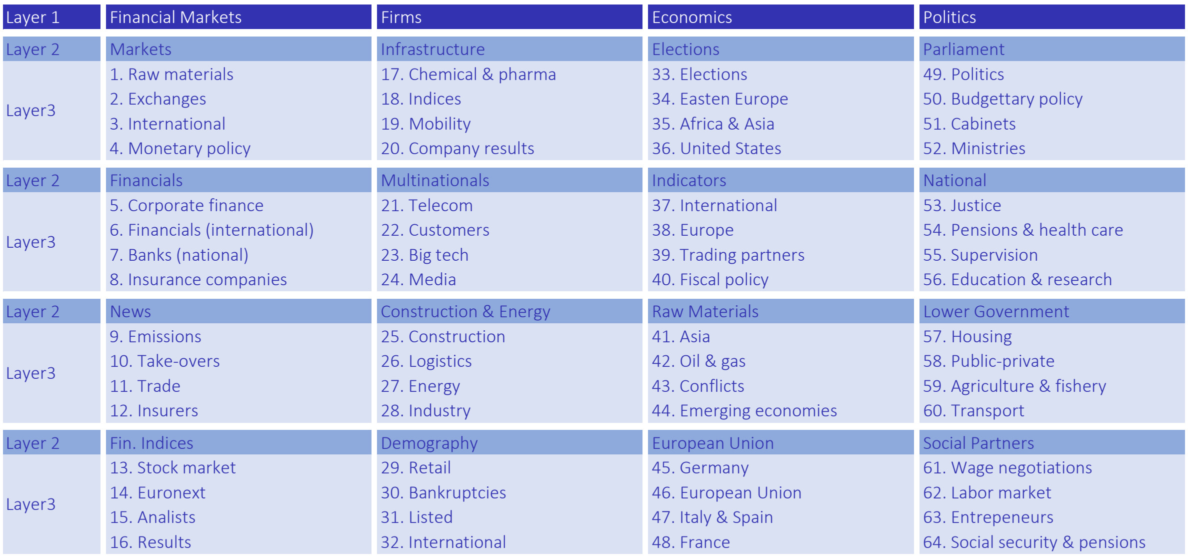
5 Outcome topic model

5.1 Topics and their interpretation

Our model is unsupervised in nature, but we ex-ante impose a layered structure of the model. The framework is flexible, and can be easily extended to more than 3 layers. Our choice of 64 topics symmetrically spread over three layers is based on two considerations. First, in our experience with more than 64 topics the topics become highly event specific, i.e., there are signs of over-fitting. Conversely, extracting substantially fewer than 64 topics results in too general topics. Unfortunately, there are no generally accepted quality metrics of quality that take into account the features in our model. Moreover, when applying layered and dynamic topic models, the type of topical layering, e.g. the number of layers is unknown in advance. However, we conducted a battery of statistical tests that are routinely used in plain vanilla topic models, to determine the number of topics. The results are presented in B.2 in the Appendix. The tests are not conclusive, but using 64 topics seems to add little to the explanatory power of the LDA. Our choice for 4 topics in the first layer was mainly based on the the organisational structure of Het Financieele Dagblad, i.e. we were informed that the editorial staff was spread over four groups each focusing on a theme that incidentally largely overlapped with the outcome of an LDA model with 4 topics. Overall, our choice of the number of topics and layers in our model is based on a mix of statistical criteria, interpretability and knowledge on the organisation of the editorial staff of the newspaper. We leave a further investigation of the optimal number of topics and layers or "depth" of the topic model to future research. As we will show in Section 4 the hierarchy in the model is especially helpful for getting extracting the inter-relatedness and hierarchy in topics.

Table II shows the extracted topics. In the first layer, we distinguish four topics, "financial markets", "firms", "economics", and "politics". To illustrate further layering, consider the first-layer topic "financial markets". This topic can be broken down in four second-layer topics, i.e.: "markets", "financials", "news", and "financial indices". These second-layer topics are each divided in four third-layer topics. For example, the "financials" topic is divided in "corporate finance", "financials (international)", "banks (national)", and "insurance companies".

Table II: Names of topics in the three different layers of the topic model.



Layer 1	Financial Markets	Firms	Economics	Politics
Layer 2	Markets	Infrastructure	Elections	Parliament
Layer3	1. Raw materials	17. Chemical & pharma	33. Elections	49. Politics
	2. Exchanges	18. Indices	34. Easten Europe	50. Budgettary policy
	3. International	19. Mobility	35. Africa & Asia	51. Cabinets
	4. Monetary policy	20. Company results	36. United States	52. Ministries
Layer 2	Financials	Multinationals	Indicators	National
Layer3	5. Corporate finance	21. Telecom	37. International	53. Justice
	6. Financials (international)	22. Customers	38. Europe	54. Pensions & health care
	7. Banks (national)	23. Big tech	39. Trading partners	55. Supervision
	8. Insurance companies	24. Media	40. Fiscal policy	56. Education & research
Layer 2	News	Construction & Energy	Raw Materials	Lower Government
Layer3	9. Emissions	25. Construction	41. Asia	57. Housing
	10. Take-overs	26. Logistics	42. Oil & gas	58. Public-private
	11. Trade	27. Energy	43. Conflicts	59. Agriculture & fishery
	12. Insurers	28. Industry	44. Emerging economies	60. Transport
Layer 2	Fin. Indices	Demography	European Union	Social Partners
Layer3	13. Stock market	29. Retail	45. Germany	61. Wage negotiations
	14. Euronext	30. Bankruptcies	46. European Union	62. Labor market
	15. Analists	31. Listed	47. Italy & Spain	63. Entrepeneurs
	16. Results	32. International	48. France	64. Social security & pensions

In LDA the topics need to be labeled by the researcher. To interpret the topics we use the *relevance* of a term, a measure introduced in Sievert and Shirley (2014). The relevance weights the most probable term in a topic, using the topic-word distribution $\boldsymbol{\phi}$ and the *lift*. The lift is defined as the ratio of a term's probability within a topic to its marginal probability across the corpus (Taddy, 2012). We set the weight for $\boldsymbol{\phi}$ to 0.4 and the weight for the lift at $(1 - 0.6)$. In our experience this setting results in easily interpretable topics. See Sievert and Shirley (2014) for a more elaborate treatment on relevance, lift and choice of weights in the relevance measure.

Note that we use a single vocabulary for all time slices, which is a mix of important words in different time periods. The word "Covid" understandably was omnipresent in the newspaper in 2020, but had a very low frequency in the years before. In our setup this does not cause problems in inferring the model coefficients, because words that do not appear in a time slice will get a negligible weight. This can be easily seen by looking at the collapsed Gibbs sampling algorithm, i.e. the weight when the word does not appear will be equal to the smoothing parameter β, divided by the sum of the total number of words in each topic and the total number of words in all document. Also note that the collapsed Gibbs sampling algorithm "automatically" picks up the increased use of the word "Covid" because the algorithm is a counting algorithm at it's core.

Figures 4a to 4d show the top 20 words per topic with the highest relevance in the first time slice (panel A) and last time time slice (panel B), ordered from high to low relevance. The light blue bar shows the overall term frequency, the dark blue bar is the estimated within topic term frequency. Figures 4 shows some interesting outcomes. First, the topics can be relatively easily labeled based on the relevance rankings. For instance, the financial markets topic (Figure 4a) is identified by the high relevance of words such as “stock”, “bank”, “stock exchange” and “stock price”. The same holds for the firms (Figure 4b), economics and politics topics. Second, there is a rather large shift between the first and last time slice in terms of the relevance of words within topics. This can be seen from the red and green arrows and the green circles indicating a decrease, increase and new entry in the top 20 ranking, respectively. For instance, within the financial markets topic, the terms “ECB” and “Euro” rose in importance. This seems logical as their prominence in the media surged after the introduction of the euro in 2002. Within the firms topic, “corona” understandably increased greatly in relevance, climbing from position 749 to position 8. Besides, the relevance of “car manufacturers” increased, possibly caused by the large innovations (electric car) and some misappropriations in this sector. Within the economics topic, “Europe”, “China” and the “European Union” gained prominence, whilst “Germany”, “the United States”, and “Russia” lost a few positions in the ranking. This seems to indicate the increased importance of Europe for the Dutch economy and the economic rise of China as a global player, respectively. Within the politics topic , the opposite seems to have happened. There is a rather large increase in the relevance of the term Netherlands, signaling the larger focus on domestic political issues instead of international issues.

Figure 4: Top-20 words with highest *relevance*, first and last time slice.

(a) Financial markets

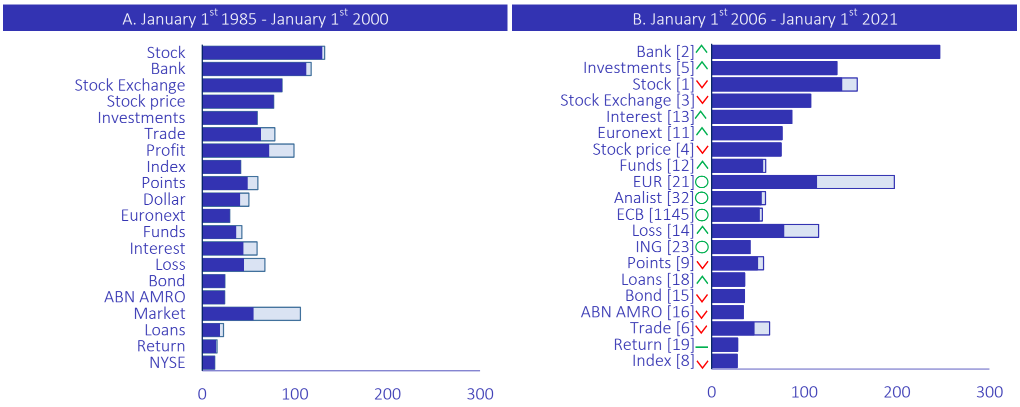
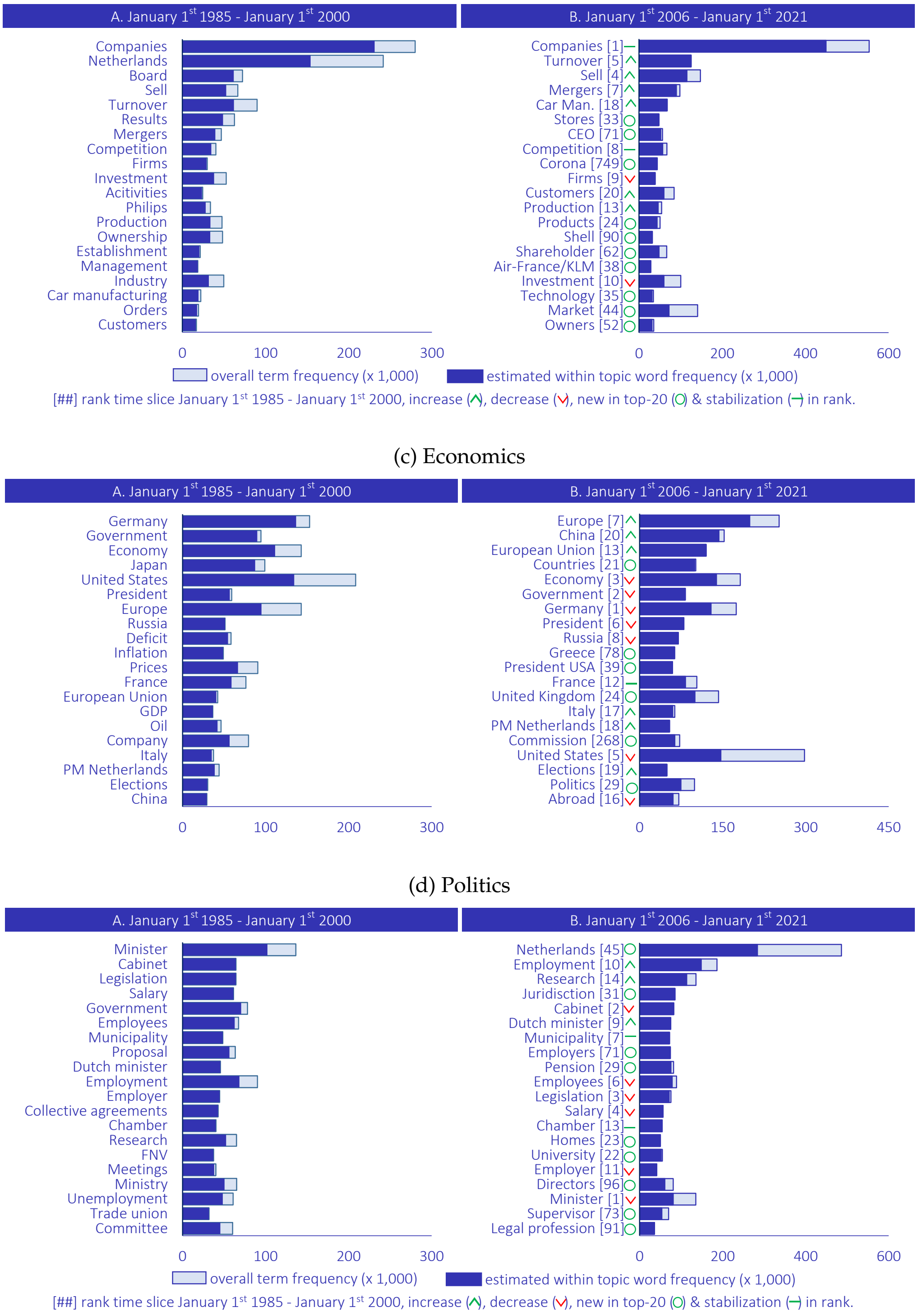


Figure 4: – continued from previous page



(b) Firms

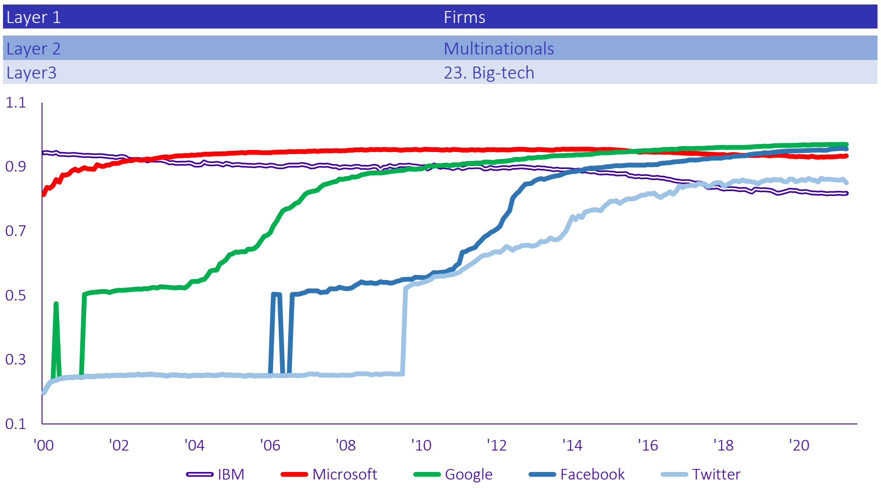
(c) Economics

(d) Politics

This is corroborated by the entry of the word “pension” in the top 20. The redesign of the Dutch pension system was a hotly debated topic over the last years. Interestingly, the words “jurisdiction” and “legal profession” entered the top 20, possibly signaling the increasing role of the judicial system influencing politics. Especially, surrounding refugees and climate issues. [9]

Figure 5 shows the time-varying character of our model from another angle, and plots –for illustration purposes– the *relevance* of five words within the "big tech" topic over time (topic 23 in Table II). We observe that "IBM", one of the largest IT hardware companies in the 1980s and 1990s, got less interest in the Dutch financial press. By contrast, the usage of the internet-firms "Google", "Facebook" and "Twitter", which were founded in 1998, 2004 and 2006 respectively, increased over time in line with their increased popularity. Overall, the examples in this section clearly indicate that time-variation in topic content and term relevance clearly add to the insights that can be gained from topic analysis.

Figure 5: Time variation in the percentile of the rank of the word based on the relevance score in Layer 3 topic 23 "Big tech", 1 = most relevant word, 0 = least relevant word.



5.2 Tone-adjusted topics

In this section we discuss the outcome of the final part of our tone-adjusted topic model, i.e. the tone adjustment of the topics. We "score" the sentiment of articles using dictionary techniques, as described in Section 3.4. Figure 6 displays a wordcloud with the frequency of the most common positive and negative words. As expected, words such as "increase", "large", "grow", and "profit" are common positive words. Words such as "decrease", "loss", "low", "problem" and "crisis" are frequently occurring negative words.

[9] We only show the most relevant words for the first-layer topics here to conserve space. Relevant words for other layers and their topics, are available upon request with the authors.

Figure 6: Frequency of most common positive (left) and negative (right) sentiment words, larger font size indicates a higher frequency. Words translated into English.

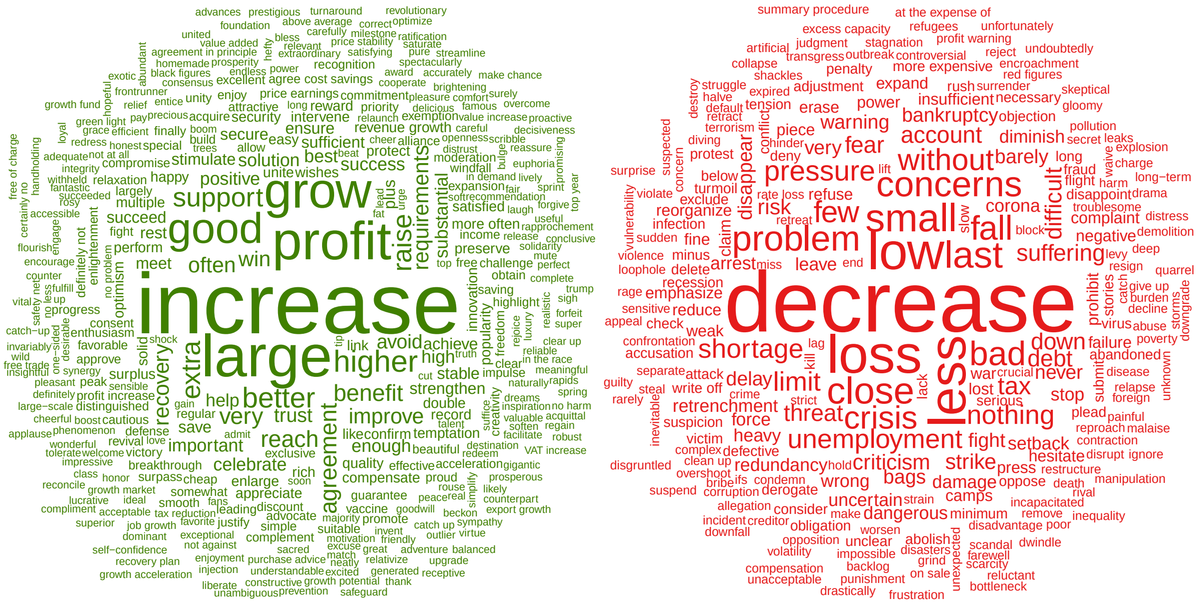


Figure 7: Normalized newspaper sentiment and normalized year-on-year GDP growth, shaded areas indicate OECD-recessions.

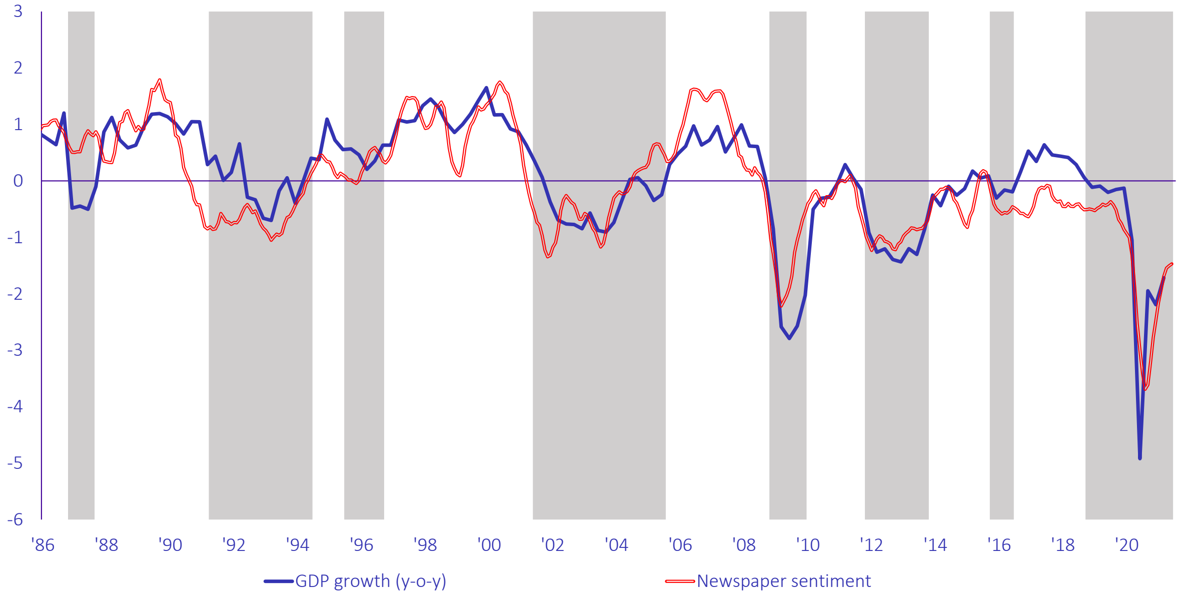
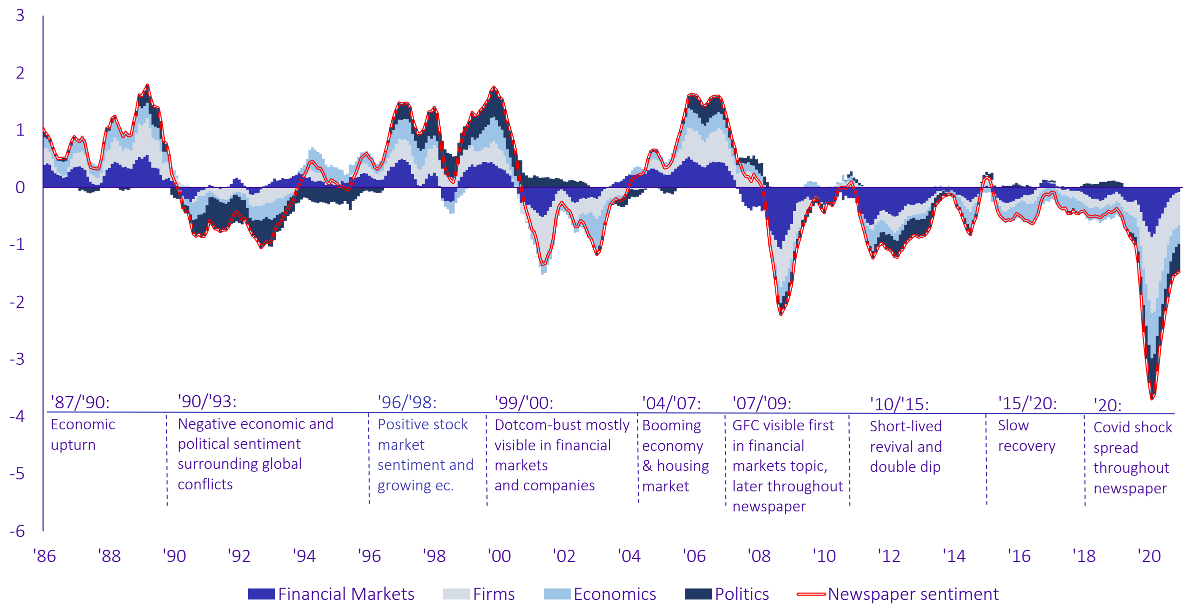


Figure 7 shows the average sentiment over the full sample. The daily and monthly sentiment series are very noisy, hence we calculate a moving average to see the underlying trend in newspaper sentiment. Here, we show the trailing 6-month moving average. Figure 7 also includes our measure of the business cycle, the year-on-year growth of GDP. The shaded areas indicate recessions, as defined by the reference turning points derived from the OECD composite leading indicator. The co-movement between the sentiment indicator and GDP growth is clearly visible. The correlation between the two series is high, i.e. 0.79. Additionally, sentiment turns more negative when the economy is in a downturn. This preliminary finding suggests that –without involved analysis– sentiment derived from financial news could be a valuable indicator for tracking the stance of the business cycle.

The allocation of headline sentiment to topics can give more insight into why sentiment is moving, as shown in Figure 8. We subdivided the headline newspaper sentiment in the topics we have distilled in the first layer, i.e. financial markets, firms, economics and politics. Some interesting patterns appear. For example, at the beginning of the financial crisis in 2008, sentiment declined solely in the topic “financial markets”. The declining sentiment expressed itself much later in the other topics, when the financial crisis deepened. This is in sharp contrast with the synchronous nature of the decline in tone-adjusted topic-sentiment at the start of the corona crisis in 2020.

Figure 8: Normalized newspaper sentiment by topic (first layer).



6 Outcome nowcasting exercise

6.1 Outcome complete sample

Table III presents data the forecast performances of the nowcasting model including and excluding the newspaper sentiment indicators for the complete sample period 2003Q3 − 2020Q3 (69 quarters). We measure forecast performances using the root mean square forecast error

(RMSFE). The RMSFE of the baseline nowcasting model without newspaper sentiment in the final row of the table ("no topics"). Our nowcasting exercise is in line with the well know result in the nowcasting literature, that the forecast error declines as more hard-based information becomes available throughout the quarter (Bańbura et al., 2013 and Giannone et al., 2008). The table further shows the *relative* forecast errors for three different versions of the DFM with newspaper sentiment, i.e: our base model with all three layers of our topic model (64 topics), the first two layers of our model (16 topics) and only the first layer (4 topics). The second row of Table III is "64 topics TVL (4x4x4)", indicating a DFM with the 70 monthly economic indicators of the baseline DFM augmented with 64 FD sentiment indicators, originating from a time-varying (TV) layered (L) topic model (TVL). Moreover, "(4x4x4)" indicates the layers of the model and the composition of the layers, i.e. 4 topics in the first layer (4), where each first layer topic has 4 topics in the second layer (4x4), and each second layer topic has 4 topics in the third layer (4x4x4).

Following, amongst others, Jansen and de Winter (2018) we present both a formal and informal measure to assess the observed differences in RMSFEs. We roughly assess the economic importance of the gain by looking at the percentage difference in RMSFE between two models. Bold faced entries indicate that the RMSFE of the DFM *with* FD sentiment is at least 5% lower than the DFM *without* FD sentiment. We conducted (one-sided) Diebold and Mariano (1995) (DM) tests as a formal test of statistical significance at the conventional levels (denoted by asterisks).[10] Non-starred, normal-type entries thus indicate models that are equal in terms of forecasting accuracy, both statistically and economically. We will follow the same two-way approach to statistical/economic significance in all tables that feature RMSFEs in this paper. The results in Table III indicate that including newspaper sentiment increases the accuracy of the DFM. This only holds for the short-run, when backcasting and nowcasting and only for the models with one or two layers.

Table III shows that the topic model with two layers and 16 topics is better than the other models when backcasting, according to both our formal and informal measure of significance. When nowcasting the advantage disappears, except for the nowcast in the third month. The decline in RMSFE ranges from 14% for the backcast in month 2 to 7% for the nowcast in month 3. The RMSFE of the DFM with one layer and 4 tone-adjusted topics is 6% better when

[10] The DM test broadly paints the same picture as the informal 5% improvement criterion, although the two do not always match. In some cases, large differences in accuracy are not statistically significant, whilst the reverse also happens. Jansen and de Winter (2018) note that he power of the DM test may be low due to the small number of observations. Moreover, the differences might signal that statistical significance and economic importance are different concepts.

Table III: Forecasting performance of DFM models with and without newspaper sentiment, RMSFE, 2003Q3–2020Q3.

	Backcast		Nowcast			Forecast		
	M2	M1	M3	M2	M1	M3	M2	M1
DFM with newspaper sentiment (relative RMSFE)								
64 topics TVL (4x4x4)	0.98	1.11	1.03	1.05	1.04	1.01	1.00	1.01
16 topics TVL (4x4)	**0.86***	**0.95***	**0.93***	1.00	1.01	1.00*	1.00	1.00
4 topics TV (4)	**0.94**	0.99	0.98*	0.99*	1.00**	1.00*	1.00	1.00
DFM without newspaper sentiment (absolute RMSFE)								
No topics	0.78	0.95	1.31	1.47	1.71	1.73	1.65	1.58

RMSFE Baseline denotes the out-of-sample root mean squared forecast error (RMSFE) of the baseline dynamic factor model (DFM) without newspaper sentiment. The DFM with newspaper sentiment variations indicate the relative RMSFEs of dynamic factors models that additionally include 64 tone-adjusted topics, 16 tone-adjusted topics or 4-tone-adjusted topics. Bold cells indicate the RMSFE is at least 5% better than the baseline. Starred entries (*, **, ***) indicate that the one-sided Diebold-Mariano test (alternative is more accurate than the baseline) is significant at the 10%, 5%, and 1% levels, respectively.

backcasting in the second month. According to our measure of economic importance this is a sizeable difference. However, the DM-test is not significant. Interestingly, the DM-test indicates significant differences for the RSMFE for the topic model with one layer and 4 tone-adjusted topics when nowcasting in months 1 to 3 and forecasting in month 3. However, the differences are very small from an economic standpoint (all average differences are $\leq$ 2%). Finally, we observe that the DFM with all 64 tone-adjusted topics actually performs worse (with the exception of the backcast in month 2) than the baseline DFM. This might reflect that the third-layer is too granular and noisy to add value in short-term forecasts.[11]. Our outcomes are in line with, amongst others, Barbaglia et al. (2022), Ellingsen et al. (2020) and Thorsrud (2016), who find significant improvements the forecasting accuracy in a nowcasting exercise when using both news sentiment and macroeconomic indicators.

6.2 Outcome periods of crisis

Table IV shows the results of our nowcasting exercise when we leave out periods with extreme GDP contractions (2009Q1, 2020Q2, and 2020Q3). As can be seen from comparing DFM without newspaper sentiment in Table IV with Table III, the RMSFE declines drastically. This coroborates previous results of, amongst others Jansen et al. (2016) who find that the FMSFE of nowcasting models increases during sudden declines in GDP.

[11] The result that the DFM that contains 64 topics performs worse than DFMs with fewer topics also hold when only comparing specifications that contain 6 factors.

Table IV: Forecasting performance of DFM models with and without newspaper sentiment without crisis period, RMSFE, 2003Q3–2020Q3.

	Backcast		Nowcast			Forecast		
	M2	M1	M3	M2	M1	M3	M2	M1
DFM with newspaper sentiment (relative RMSFE)								
64 topics TVL (4x4x4)	**0.93**	**0.94**	0.96	0.97	1.00	1.01	1.04	1.08
16 topics TVL (4x4)	1.01*	1.00*	1.00*	1.00	0.99	0.97*	0.99	0.97
4 topics TV (4)	0.98	0.99	0.99*	0.99*	0.99**	1.00*	1.00	1.00
DFM without newspaper sentiment (absolute RMSFE)								
No topics	0.53	0.54	0.55	0.56	0.57	0.57	0.59	0.59

RMSFE Baseline denotes the out-of-sample root mean squared forecast error (RMSFE) of the baseline dynamic factor model (DFM) without newspaper sentiment with very large GDP contractions (2009Q1, 2020Q2, and 2020Q3) removed. The DFM with newspaper sentiment variations indicate the relative RMSFEs of dynamic factors models that additionally include 64 tone-adjusted topics, 16 tone-adjusted topics or 4-tone-adjusted topics. Bold cells indicate the RMSFE is at least 5% better than the baseline. Starred entries (*, **, ***) indicate that the one-sided Diebold-Mariano test (alternative is more accurate than the baseline) is significant at the 10%, 5%, and 1% levels, respectively.

For example, the RMSFE for the DFM without newspaper sentiment drops from 1.22 in Table III to 0.55 in Table IV. We still observe added value of the FD sentiment indicators, although non of the differences is both formally and informally significant. For instance, the DFM including the 64 topics from the third layer of the tone-adjusted time-varying layered topic model is diverging by more than 5% from the DFM without the sentiment indicators but the difference is not significant according to the one-sided DM-test. Moreover, the DFM with 16 topics and 4 topics show significant differences depending on the back, now and forecasting horizon, but from an economic viewpoint these differences are small. Overall, these outcomes suggest that a large part of the added value of the sentiment indicators over the full sample can be attributed to their added value during times of crisis. This is an interesting result, as nowcasting during these times is notoriously difficult. Moreover, from a policy perspective, a good forecasting performance is especially important during periods of crisis. This finding corroborates earlier findings of, amongst others, Ashwin et al. (2021) and Kalamara et al. (2022).

6.3 Robustness tests

We have estimated several other nowcasting specifications with different versions of our topic model to assess the robustness of our findings. Table V shows the results of our nowcasting model when we only include nowcasting specifications with at least 4 dynamic and static factors instead of at least 3 in our baseline-outcomes in Table III. Qualitatively, our results are robust to this change: the forecasting errors are significantly smaller at short horizons and the results are economically meaningful. In the specification with 16 topics, the RMSFE for the two backcasts and the nowcast in month 3 is about $9 - 17\%$ lower.

Table V: Forecasting performance of DFM models with and without newspaper sentiment, extra specifications, RMSFE, 2003Q3–2020Q3.

	Backcast		Nowcast			Forecast		
	M2	M1	M3	M2	M1	M3	M2	M1
DFM with newspaper sentiment ≥ 4 factors (relative RMSFE)								
64 topics TVL (4x4x4)	0.98	1.09	1.01	1.06	1.05	1.01	1.01	1.02
16 topics TVL (4x4)	**0.83**	**0.92***	**0.91**	0.99	1.01	1.00	1.00	1.00
4 topics TV (4)	**0.93**	0.98	0.98*	0.99*	1.00**	1.00**	1.00	1.00
DFM without newspaper sentiment ≥ 4 factors (absolute RMSFE)								
No topics	0.76	0.92	1.30	1.46	1.69	1.72	1.65	1.58

RMSFE Baseline denotes the out-of-sample root mean squared forecast error (RMSFE) of the baseline dynamic factor model (DFM) without newspaper sentiment. The DFM with newspaper sentiment variations indicate the relative RMSFEs of dynamic factors models that additionally include tone-adjusted topics, TV = Time-varying, L = Layered, TVL = Time-varying layered. Bold cells indicate the RMSFE is at least 5% better than the baseline. Starred entries (*, **, ***) indicate that the one-sided Diebold-Mariano test (alternative is more accurate than the baseline) is significant at the 10%, 5%, and 1% levels, respectively.

Table VI: Comparing added value of time variation and time variation in forecasting performance of DFM models, RMSFE, 2003Q3–2020Q3.

	Backcast		Nowcast			Forecast		
	M2	M1	M3	M2	M1	M3	M2	M1
DFM with newspaper sentiment (relative RMSFE)								
16 topics TV (1x16)	1.00	0.99	0.97	0.98	0.98	1.00	1.00	1.00
16 topics L (4x4)	1.03	1.02	0.96	0.97	0.97	1.00	1.00	1.00
16 topics TVL (4x4)	1.02	0.98	**0.93**	**0.94**	0.96	1.00	1.00	1.00
Plain vanilla DFM with newspaper sentiment (absolute RMSFE)								
16 topics plain vanilla (1x16)	0.66	0.91	1.31	1.55	1.80	1.72	1.65	1.59

RMSFE Baseline denotes the out-of-sample root mean squared forecast error (RMSFE) of a Dynamic Factor Model (DFM) with newspaper sentiment in 16 topics based on a plain vanilla topic model. The DFM with newspaper sentiment variations indicate the relative RMSFEs of DFMs with 16 newspaper topics based on topic models that additionally include layering, time variation, or both. TV = Time-varying, L = Layered, TVL = Time-varying layered. Bold cells indicate the RMSFE is at least 5% better than the baseline. Differences in forecasting accuracy are not significant according to a one-sided Diebold-Mariano test at conventional levels.

Table VI shows the results of the nowcasting exercise with different types of topic models. The baseline is based on a DFM that includes 16 topics from a plain vanilla topic model without time variation or layering. We find that the DFMs that include 16 topics based on a model with layering, time variation, or both are slightly better. However, only the nowcasts in month 3 and month 2 from a model that includes both time variation and layering are economically significantly better: 7% and 6%, respectively. The Diebold-Mariano test, however, is not significant for any forecast. This reflects that layering or time-variation in the topic model does not add much value in the forecasting exercise. It does help in interpreting the topics over time (see Section 5.2).

7 Conclusions

This paper investigates two related questions. First, can newspaper sentiment can be used to assess the course of the business cycle? Second, Can newspaper sentiment be used to decrease the forecasting error of nowcasting models? To answer the first question we combine a dictionary based newspaper sentiment measure with a topic model that allows us to trace the sentiment by fine grained time-varying and hierarchically ordered topics. To the best of our knowledge, our tone-adjusted time-varying layered topic model is new in the literature. The second question analyzes the first question more formally by examining the added value of including newspaper sentiment indicators in a formal nowcasting horse-race between a state-of-the art nowcasting model using 70 monthly indicators versus a model using these indicators and 4 to 64, time-varying and hierarchically ordered, sentiment indicators derived from our topic model. Our findings can be summarized in five points.

First, the aggregate newspaper sentiment is a strong indicator of the business cycle. Sentiment correlates strongly with y-o-y GDP growth and sentiment turns negative when the economy is in a downturn. Second, combining sentiment with the topics from the topic model enables story-telling regarding sentiment movements. Moreover, the time-varying and layered character of the topics resulting from our time-varying layered topic model can provide policymakers and practitioners with valuable insights into the causes of sentiment swings. Third, sentiment indicators derived from newspaper articles contain valuable information not embodied in monthly indicators from statistical offices. The forecast accuracy of our DFM improves when including the tone-adjusted topics derived from the newspaper articles, especially when now- and backcasting. The added value of newspaper sentiment articles is particularly strong during periods of large declines in GDP. Fourth, time-variation and layering of the topic model add little to the forecasting power in the nowcasting exercise. Fifth, for nowcasting purposes, the amount of granularity of the topic model needs to be limited. We find that in our case at tone-adjusted time-varying topic model with 16 topics in the second layer and 4 topics in the first layer has the highest forecasting accuracy.

The results of our analysis may be useful to policymakers, financial analysts, and economic agents. We demonstrate that sentiment indicators derived from the financial press contain valuable additional information that can be extracted in real time. Our findings suggest that the number of topics and the number of layers in the topic model are important determinants of the forecasting power of their value added in a nowcasting model. In this research we investigated a maximum number of layers of three with an equal number of topics stemming from each topic in each layer, i.e. 64 topics in the three layered model. Developing selection criteria for determining the optimal number of layers and topics are interesting topic for future research. A deeper investigation of the real-time availability of the newspaper data could also be an interesting avenue to explore. In our current analysis we investigate the advantage of using newspaper data on one specific day in the month, i.e. the first day, using pseudo real-time vintages for the monthly economic indicators. Using real-time vintages of the indicators and could possibly better pinpoint when the newspaper data are most advantageous.

References

Aletras, N. and M. Stevenson (2014). Measuring the similarity between automatically generated topics. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pp. 22–27. link.

Algaba, A., D. Ardia, K. Bluteau, S. Borms, and K. Boudt (2020). Econometrics meets sentiment: an overview of methodology and applications. *Journal of Economic Surveys 34*(3), 512–547. link.

Aprigliano, V., S. Emiliozzi, G. Guaitoli, A. Luciani, J. Marcucci, and L. Monteforte (2022). The power of text-based indicators in forecasting Italian economic activity. *International Journal of Forecasting forthcoming*, 1–18. link.

Ardia, D., K. Bluteau, and K. Boudt (2019). Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. *International Journal of Forecasting 35*(4), 1370–1386. link.

Arun, R., V. Suresh, C. E. V. Madhavan, and M. N. N. Murthy (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In M. J. Zaki, J. X. Yu, B. Ravindran, and V. Pudi (Eds.), *Advances in knowledge discovery and data mining*, pp. 391–402. Springer Berlin Heidelberg. link.

Ashwin, J., E. Kalamara, and L. Saiz (2021). Nowcasting euro area GDP with news sentiment: a tale of two crise. Working Papers 2616, European Central Bank. link.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221. link.

Bańbura, M., D. Giannone, and L. Reichlin (2011). Nowcasting. In M. P. Clements and D. F. Hendry (Eds.), *Oxford Handbook on Economic Forecasting*, pp. 63–90. Oxford University Press. link.

Bańbura, M. and G. Rünstler (2011). A look into the factor model black box: Publication lags and the role of hard and soft data in forecasting GDP. *International Journal of Forecasting 27*(2), 333–346. link.

Barbaglia, L., S. Consolia, and S. Manzan (2022). Forecasting with economic news. *Journal of Business & Economic Statistics forthcoming*, 1–12. link.

Barbaglia, L., L. Frattarolo, L. Onorante, F. M. Pericoli, M. Ratto, and L. T. Pezzoli (2022). Testing big data in a big crisis: Nowcasting under Covid-19. *International Journal of Forecasting forthcoming*, 1–16. link.

Bańbura, M., D. Giannone, M. Modugno, and L. Reichlin (2013). Now-casting and the real-time data flow. In G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 2, Part A of *Handbook of Economic Forecasting*, pp. 195–237. Elsevier. link.

Bittermann, A. and J. Rieger (2022). Finding scientific topics in continuously growing text corpora. In *ACL2022: Proceedings of the Third Workshop on Scholarly Document Processing*, pp. 7–18. link.

Blei, D. M. and J. D. Lafferty (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 113–120. Association for Computing Machinery. link.

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

Bybee, L., B. T. Kelly, A. Manela, and D. Xiu (2020). The structure of economic news. NBER Working Papers 26648, National Bureau of Economic Research. link.

Cao, J., T. Xia, J. Lia, Y. Zhang, and ShengTanga (2009). A density-based method for adaptive LDA model selection. *Neurocomputing 72*(7), 1775–1781. link.

Chauhan, U. and A. Shah (2021). Topic modeling using latent dirichlet allocation: A survey. *ACM Computing Surveys 54*(7), 1–35. link.

Churchill, R. and L. Singh (2022). The evolution of topic modeling. *ACM Computing Surveys 54*(10), 1–35. link.

De Smedt, T. and W. Daelemans (2021). Mobility-based real-time economic monitoring amid the COVID-19 pandemic. *Nature: scientific reports 11*(13069), 1–15. link.

Deveaud, R., E. Sanjuan, and P. Bellot (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique 17*(1), 61–84. link.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 134–144. link.

Ellingsen, J., V. H. Larsen, and L. A. Throrsrud (2020). News media vs. FRED-MD for macroeconomic forecasting. Working Paper 14, Norges Bank. link.

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

Giannone, D., L. Reichlin, and D. Small (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics 55*(4), 665–676. link.

Griffiths, T., M. Jordan, J. Tenenbaum, and D. Blei (2003). Hierarchical topic models and the nested chinese restaurant process. In S. Thrun, L. Saul, and B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems*, Volume 16, pp. 17–24. MIT Press. link.

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

Hansen, S., M. McMahon, and A. Prat (2018). Transparency and deliberation within the

FOMC: A computational linguistics approach. *The Quarterly Journal of Economics 133*(2), 801–870. link.

Hindrayanto, I., S. J. Koopman, and J. M. de Winter (2016). Forecasting and nowcasting economic growth in the euro area using factor models. *International Journal of Forecasting 32*(4), 1284–1305. link.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal 42*(1), 177–196. link.

Jansen, W. J. and J. M. de Winter (2018). Combining model-based near-term GDP forecasts and judgmental forecasts: A real-time exercise for the G7 countries. *Oxford Bulletin of Economics and Statistics 80*(6), 1213–1242. link.

Jansen, W. J., X. Jin, and J. M. de Winter (2016). Forecasting and nowcasting real GDP: Comparing statistical models and subjective forecasts. *International Journal of Forecasting 32*(2), 411–436. link.

Kalamara, E., A. Turrell, C. Redl, G. Kapetanios, and S. Kapadia (2022). Making text count: Economic forecasting using newspaper text. *Journal of Applied Econometrics 37*(5), 896–919. link.

Kuzin, V., M. Marcellino, and C. Schumacher (2013). Pooling versus model selection for nowcasting GDP with many predictors: Empirical evidence for six industrialized countries. *Journal of Applied Econometrics 28*(3), 392–411. link.

Larsen, V. H. and L. A. Thorsrud (2018). Business cycle narratives. Working Papers 6, BI Norwegian Business School. link.

Larsen, V. H. and L. A. Thorsrud (2022). Asset returns, news topics, and media effects. *Scandinavian Journal of Economics 124*(3), 838–868. link.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? textual analysis, dictionaries, and 10Ks. *Journal of Finance 66*(1), 35–65. link.

Newman, D., J. H. Lau, K. Grieser, and T. Baldwin (2010). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pp. 100–108. link.

Porter, M. (1980). An algorithm for suffix stripping. *Program: Electronic library and information systems 14*(3), 130–137. link.

Rambaccussing, D. and A. Kwiatkowski (2020). Forecasting with news sentiment: Evidence with UK newspapers. *International Journal of Forecasting 36*(4), 1501–1506. link.

Resnik, P. and E. Hardisty (2009). Gibbs sampling for the uninitiated. mimeo, Institute for Advanced Computer Studies University of Maryland. link.

Shapiro, A. H., M. Sudhof, and D. J. Wilson (2022). Measuring news sentiment. *Journal of Econometrics 228*(2), 221–243. link.

Sievert, C. and K. Shirley (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70. link.

Steyvers, M. and T. Griffiths (2007). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch (Eds.), *Latent Semantic Analysis: A Road to Meaning*, pp. 424–440. Laurence Erlbaum. link.

Taddy, M. (2012). On estimation and selection for topic models. In *Artificial Intelligence and Statistics*, pp. 1184–1193. link.

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. link.

Thorsrud, L. A. (2016). Nowcasting using news topics: Big data versus big bank. Working Papers 6, Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School. link.

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics 38*(2), 393–409. link.

Wischnewsky, A., D.-J. Jansen, and M. Neuenkirch (2021). Financial stability and the Fed: Evidence from congressional hearings. *Economic Inquiry 59*, 1192–1214. link.

A Data

A.1 Cleaning the article database

This section, describes how we cleaned the raw-database in more detail. Cleaning reduced the size of our article database from 1, 093, 477 articles to articles to 582, 981 articles, a reduction of approximately 47%.

First, we stripped the database of newspaper categories that were not relevant, e.g fashion, radio and television pages, letters from readers, profiles of entrepreneurs, personal finance, advertorials, photo-pages, newspaper service pages, announcement of events. This step reduces our database by 25, 928 articles to 1, 067, 549.

Second, we used the publication URLs of the articles to clean irrelevant articles from the database. The URLs are composed of several parts, where the last part indicates an abbreviated indication of the article title. After deleting the last part of the URLs there are 1, 046 unique URLs left. Each of these URL addresses contains information on the kind of article. We manually checked and re-categorized all 1, 046 unique URLs into categories as long as the category was equal or larger than 0.1% of all articles. Using this approach we were able to group 72% of all articles. For the articles that had an informative URL we grouped the articles in 11 categories, based on the tags that were given by the FD, i.e.: 1. company news, 2. economics & politics, 3. financial markets, 4. opinion-pages, 5. domestic news , 6. personal profiles, 7. foreign news, 8 human interest, 9. English pages and 10. short news, 11. archive. Category 9 might seem obscure for a Dutch financial newspaper, but until the mid 2000s the FD contained an English back-page with the most important news of that day in English as a service to non-native readers. We deleted all articles in categories 4, 6, 8, 9 and 11, reducing our database by 438, 550 articles to 628, 999 articles.

Third, we deleted all articles with a title that clearly signals the article is not relevant for our purpose, e.g. articles that contained summaries of closing and opening prices for stock exchanges, articles containing agendas for upcoming events and company press releases and overview summaries of newspaper content. This reduces the database by 18, 753 articles to 610, 246 articles.

Fourth, we deleted articles that appeared more than once in the database. This can occur because articles are adjusted later on, or are first published on the website and later in the printed newspaper. We decided to clear these articles, and keep the article when its first published (online). Deleting articles with the same title and the same publication date drops 23, 080 articles from the database. Dropping articles with the same title and one day publication difference drops $1,978$ articles from the database. A two day publication delay is relatively rare, but we deleted these articles as well, dropping another 538 articles. In total, this cleaning step reduces the number of article in the database by $25,596$ to $586,408$.

Fifth, we deleted all articles that have no content, but do have a title. These articles can be broadly classified in the following groups: headlines appearing on the front-page of the FD, with the actual article in the database with a different article identifier, titles of info-graphics, one-line articles about changes in stock-markets (e.g "AHOLD -5.4%"), or one liners indicating a recent release (e.g "industrial production: $+2,8\%$"). This step reduces ur database by 120 articles to $586,288$.

Sixth, we removed any reaming English articles. First, we identify articles that contain three of the most common English words, i.e. "the", "and" and "to" . Next, we check the language of these articles using the R-package `textcat` to verify the articles are English. This procedure deletes another $3,307$ articles from the database, and reduces the database to a total of $582,981$ articles.

Finally, we converted all words to lowercase letters, stripped HTML-codes from the text and removed all punctuation and numbers from the texts. We kept the dot (".") and the HTML-tag for sections "<\p>" to identify the number of sections and sentences in an article.

A.2 Data set of economic indicators

Table A.I provides the list of the monthly indicators that have been used for the estimation of the dynamic factor models. As mentioned in the main text, the data can be split-up into five groups: production & sales, surveys, financial indicators, prices and indicators of important trading partners; see the headings in the table. Furthermore, the table shows the start and end date for all variables in the data set. The main data source for our database are Statistics Netherlands, Eurostat and the Datawarehouse of the European Central Bank. A few series are downloaded via the subscription service Datastream. The world trade series from the World Trade Monitor, maintained by the Dutch CPB Netherlands Bureau for Economic Policy Analysis.

Table A.I: Description monthly database

Nr.	Variable name	Source	Start	End
Production & sales				
1	Industrial production: total	ECB	jan-65	nov-20
2	Industrial production: capital goods	ECB	jan-70	nov-20
3	Industrial production: durable consumption goods	ECB	jan-90	nov-20
4	Industrial production: non-durable consumption goods	ECB	jan-90	nov-20
5	Retail trade turnover	ECB	jan-94	dec-20
6	Household consumption, durable goods	CBS	jan-95	nov-20
7	Household consumption, other goods	CBS	jan-95	nov-20
8	Household consumption, services	CBS	jan-95	nov-20
9	Unemployment rate	CBS	jan-83	dec-20
10	Building permits	CBS	jan-95	nov-20
11	Personal car registrations	CBS	jan-65	jan-21
12	New commercial vehicles registration	ECB	jan-90	dec-20
13	Construction production	ECB	jan-85	nov-20
14	World Trade	CPB	jan-91	nov-20
15	Imports	CBS	jan-90	nov-20
16	Exports	CBS	jan-90	nov-20
17	Bankruptcies	CBS	jan-65	dec-20
Surveys				
18	Consumer confidence: headline	CBS	apr-86	jan-21
19	Consumer confidence: ec. situation < 12 months	CBS	apr-86	jan-21
20	Consumer confidence: ec. situation > 12 months	CBS	apr-86	jan-21
21	Consumer confidence: fin. situation < 12 months	CBS	apr-86	jan-21
22	Consumer confidence: financial situation > 12 months	CBS	apr-86	jan-21
23	Consumer confidence: major purchases, present	CBS	apr-86	jan-21
24	Consumer confidence: unemployment > 12 months	ES	jan-85	jan-21
25	Consumer confidence: major purchases > 12 months	ES	jan-85	jan-21
26	Industrial confidence: headline	CBS	jan-85	jan-21
27	Industrial confidence: production > months	CBS	jan-85	jan-21
28	Industrial confidence: assessment of order-book	CBS	jan-85	jan-21
29	Industrial confidence: assessment of stocks	CBS	jan-85	jan-21
30	Industrial confidence: employment > months	ES	jan-85	jan-21
31	PMI: new orders	DS	mrt-00	jan-21
32	Retail confidence: headline	ES	jan-86	jan-21
33	Retail confidence: sales < 3 months	ES	jan-86	jan-21
34	Retail confidence: volume of stock	ES	jan-86	jan-21
35	Retail confidence: sales > 3 months	ES	jan-86	jan-21
36	Retail confidence: employment > 3 months	ES	jan-86	jan-21
37	Service confidence: headline	ES	jan-96	jan-21
38	Service confidence: business situation < 3 months	ES	apr-93	jan-21
39	Service confidence: demand < 3 months	ES	apr-93	jan-21
40	Service confidence: demand > 3 months	ES	jan-96	jan-21
41	Construction confidence: headline	ES	jan-85	jan-21
42	Construction confidence: building activity < 3 months	ES	jan-85	jan-21
43	Construction confidence: % no limiting factors in prod.	ES	sep-85	jan-21
44	Construction confidence: evolution order books	ES	jan-85	jan-21
45	Construction confidence: employment > 3 months	ES	jan-85	jan-21
Financial				
46	Interest rate: loans on mortgages, 5-10 year	ECB	jan-80	nov-20
47	Amsterdam AEX-index	DS	jan-83	jan-21
48	Amsterdam Midkap-index	DS	jan-83	jan-21
49	Euro area: Dow Jones Euro Stoxx 50 index	ECB	jan-87	dec-20
50	Euro area: Dow Jones Euro Stoxx basic materials index	ECB	jan-87	dec-20
51	Euro area: Dow Jones Euro Stoxx technology index	ECB	jan-87	jan-21
52	Euro area: Dow Jones Euro Stoxx industrials Index	ECB	jan-87	dec-20
53	M1	ECB	jan-80	dec-20
54	Loans to the private sector	ECB	dec-82	dec-20
Prices				

continued on next page . . .

TableA.I – continued from previous page

Nr.	Variable name	Source	Start	End
55	Consumer-price index: headline	CBS	jan-65	dec-20
56	World market commodity prices: metals	HWWA	sep-78	jan-21
57	Producer price: industry (foreign market)	CBS	jan-81	dec-20
Trading partners				
58	United Kingdom: industrial production	ECB	jan-98	sep-20
59	Germany: industrial production (excl. construction)	ECB	jan-65	nov-20
60	Germany: IFO-indicator, expected business-situation	IFO	jan-91	jan-21
61	Germany: IFO-indicator, current business-situation	IFO	jan-91	jan-21
62	Germany: ZEW-indicator expected econ. situation	ZEW	dec-91	jan-21
63	Germany: ZEW-indicator current econ. situation	ZEW	dec-91	jan-21
64	France: industrial production (excl. construction)	ECB	jan-65	nov-20
65	Italy: industrial production (excl. construction)	ECB	jan-65	nov-20
66	Spain: industrial production (excl. construction)	ECB	jan-65	nov-20
67	Belgium: BNB-indicator	BNB	jan-85	jan-21
68	Belgium: consumer confidence	ES	jan-85	jan-21
69	Belgium: industrial production (excl. construction)	ECB	jan-65	nov-20
70	Belgium: retail trade	ECB	jan-70	dec-20

Source.: BNB: Banque Nationale de Belgique, CBS: Statistics Netherlands, CPB: CPB Netherlands Bureau for Economic Policy Analysis, DS: Datastream, ECB: European Central Bank, ES: Eurostat, HWWA: Hamburgisches Welt-Wirtschafts-Archive, IFO: IFO Institute, ZEW: Leibniz-Zentrum für Europäische Wirtschaftsforschung. **Start**: First month of the time-series **End**: Last month of the time series.

B Details on Bayesian inference algorithm

B.1 Inference of the posterior distribution

The full posterior distribution of the latent variables ϕ, θ and x, conditional on the observed corpus $\boldsymbol{w}$, and the priors α and β, can be inferred by using the definitions of conditional, marginal and joint distributions, as:

$$\Pr(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{x} | \boldsymbol{w}, \alpha, \beta) = \frac{\Pr(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{w} | \alpha, \beta)}{\Pr(\boldsymbol{w} | \alpha, \beta)} \tag{B.1}$$

The joint distribution in the numerator can be written as:

$$\Pr(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{w} | \alpha, \beta) = \underbrace{\prod_{t=1}^{T} \Pr(\phi_t | \beta)}_{\text{topic-word}} \prod_{d=1}^{D^1} \left[\Pr(\theta_d | \alpha) \prod_{n=1}^{N} \Pr(x_{dn} | \theta_d) \Pr(w_{dn} | \boldsymbol{\phi}, x_{dn}) \right] \tag{B.2}$$

The denominator, the *evidence* or *marginal likelihood*, can be obtained by marginalizing over the latent variables β, θ and x, i.e.:

$$\Pr(\boldsymbol{w} | \alpha, \beta) = \int \int \sum_{x} \left(\prod_{t=1}^{T} \Pr(\phi_t | \beta) \right) \left(\prod_{d=1}^{D^1} \Pr(\theta_d | \alpha) \prod_{n=1}^{N} \Pr(x_{dn} | \theta_d) \Pr(w_{dn} | \boldsymbol{\phi}, x_{dn}) \right) d\theta d\phi \tag{B.3}$$

The numerator in equation (B.2) can be computed easily, but the evidence in equation (B.3) is intractable to compute as the latent variables β and θ are not separable in summing over all the possible values of the latent topic structure, because there is an exponentially large number of possible topic structures, and this sum is intractable (see Blei et al., 2003 for a formal proof). Luckily, there is a variety of methods to approximate the inference, such as expectation-maximization (e.g. Hofmann, 2001), variational inference (e.g. Blei et al., 2003), or Gibbs sampling (e.g. Griffiths and Steyvers, 2004).

We used a variant of the Gibbs sampling algorithm for inference, i.e. the so-called *collapsed* Gibbs-sampling algorithm. The idea of collapsed Gibbs sampling, is that it allows us to sample from a distribution that asymptotically follows the full joint distribution $\Pr(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{w} | \alpha, \beta)$ without having to explicitly calculate any integrals. We present the main formulas of the collapsed Gibbs sampling algorithm without formal derivation, which are well described elsewhere (e.g. Resnik and Hardisty, 2009).

The collapsed Gibbs sampling procedure considers each word token in the text collection in turn, and estimates the probability of assigning the current word token to each topic, *con-*

ditioned on the topic assignments to all other word tokens. From this conditional distribution, a topic is sampled an stored as the new topic assignment for this word. We write this conditional distribution as $\Pr(x_i = j|\mathbf{x}_{-i}, w_i, d_i, .)$, where $x_i = j$ represents the topic assignment of token i to topic j, x_i refers to the topic assignments of all other word tokens, and "." refers to all other known or observed information such as all other word and document indices and hyperparameters. Griffiths and Steyvers (2004) and Steyvers and Griffiths (2007) show this can be rather easily calculated, by a counting rule, i.e.:

$$\Pr(x_i = j|\mathbf{x}_{-i}, w_i, d_i, .) \propto \frac{\mathbf{C}^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} \mathbf{C}^{WT}_{wj} + W\beta} \times \frac{\mathbf{C}^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} \mathbf{C}^{DT}_{w_i j} + T\alpha} \tag{B.4}$$

The first term on the right-side of the equal sign $\mathbf{C}^{WT}$ is the topic-word matrix and $\sum_{w=1}^{W} \mathbf{C}^{WT}_{wj}$ is the total number of tokens(words) in each topic. In the second term $\mathbf{C}^{DT}$ is the document-topic matrix and $\sum_{t=1}^{T} \mathbf{C}^{DT}_{w_i j}$ indicates the total number of tokens(words) in document i. α and β are the hyperparameters of the Dirichlet distributions of the document-topic and topic-word distribution, respectively. W is the total number of words in the set of documents and T is the number of topics. The first term is the probability of word w under topic j whereas the second term is the probability that topic j has under the current topic distribution for document d. The intuition is that once many tokens of a word have been assigned to topic j, it will increase the probability of assigning any particular token of that word to topic j (the first term). At the same time, if topic j has been used multiple times in a document, it will increase the probability that any word from that document will be assigned to topic j (the second term). Therefore, words are assigned to topics depending on how likely the word is for a topic, *as well as* how dominant a topic is in a document.

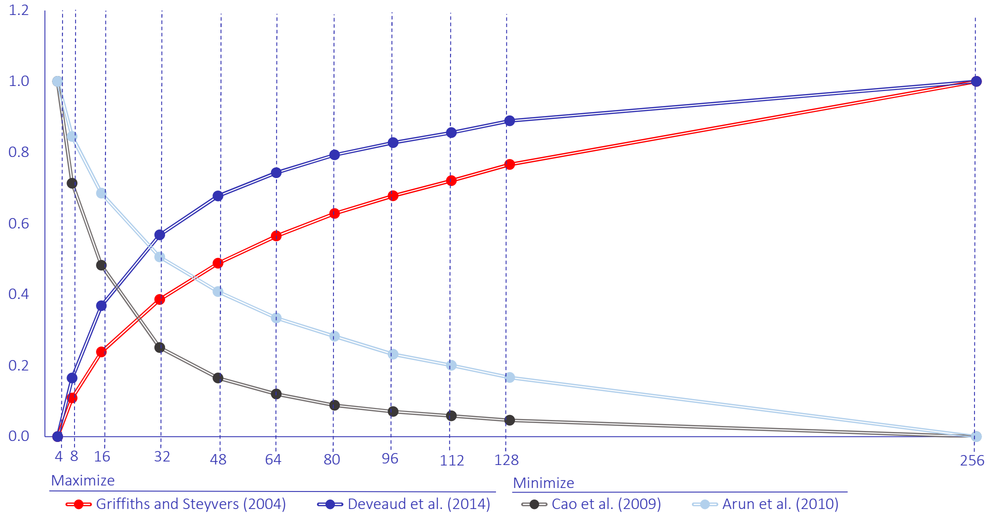
The Gibbs sampling algorithm starts by assigning each word token to a random topic in $[1, \ldots, T]$. For each word token, the count matrices $\mathbf{C}^{WT}$ and $\mathbf{C}^{DT}$ are first decremented by one for the entries that correspond to the current optic assignment. Then, a new topic is sampled from the distribution in equation (B.4) and the count matrices $\mathbf{C}^{WT}$ and $\mathbf{C}^{DT}$ are incremented with the new topic assignments. Each Gibbs sample consists of the set of topic assignment tot all $N \times M$ word tokens in the corpus, achieved by a single pass through all documents. During the initial stage of the sampling process, the "burnin period", the Gibbs samples have to be discarded because they are poor estimates of the posterior. After the burnin period, the successive Gibbs samples start to approximate the target distribution, i.e. the posterior distribution over topic assignments. At this point, to get a representative set of sampled from this distribution, a number of Gibbs samples are saved at regularly spaced intervals, to prevent correlations between samples (“skip sampling”). The sampling process is done sequentially and proceeds until the sampled values approximate the target distribution. We can then use the count-matrices $\mathbf{C}^{WT}$ and $\mathbf{C}^{DT}$ to approximate the estimated posterior topic-word matrices $\boldsymbol{\phi}$ and the estimated posterior topic matrix per document $\boldsymbol{\theta}$, respectively. Following Griffiths and Steyvers (2004) these matrices can be build from calculated probabilities per topic-word and document-topic combinations, defined as:

$$\hat{\phi}_{ij} = \frac{\mathbf{C}^{WT}_{ij} + \beta}{\sum_{k=1}^{W} \mathbf{C}^{WT}_{tj} + W\beta}, \qquad \hat{\theta}_{dj} = \frac{\mathbf{C}^{DT}_{dj} + \alpha}{\sum_{k=1}^{T} \mathbf{C}^{DT}_{dt} + T\alpha} \tag{B.5}$$

B.2 Setting of Bayesian inference model

Number of topics

Figure B.1: Selection of number of topics, indicators (re-scaled to 0 -1)



We conduct statistical tests to determine the optimal number of topics using the tests available in the **R**-package `ldatuning`, i.e. the test in Cao et al. (2009), Griffiths and Steyvers (2004), Arun et al. (2010) and Deveaud et al. (2014). For more information on the tests we refer to these papers. We tested for the optimal number of topics in the first time slice 1, running from January 1st 1985 to January 1st 2000. The results are presented in Figure B.1. All tests are re-scaled with a min-max transformation to lie between 0 and 1. Generally the tests indicate an optimal number of topics that is higher than 64, but the improvement becomes relatively small with more than 64 topics.

Iterations and hyperparameters

The Gibbs sampler also requires choices regarding the number of repeated samples (iterations) and the setting of the hyperparameters. In general, the choice of the hyperparameters α for the document-topic distribution (θ) and β for the topic-word distribution (ϕ) depends on the empirical application. With a higher α, documents are made up of more topics. Likewise, with a high β, topics are made up of most of the words in the corpus, and with a low β they consist of few words. Griffiths and Steyvers (2004) is an often used standard setting in the literature. in which α is to 50 divided by the number of topics ($50/T$) and β to 0.1. We set α to 0.1 and β to 0.01, after an informal grid-search, using several different hyperparameters in the range $0.01 - 0.1$. The chosen values for the hyperparameters resulted in the best interpretability of the topics.

We set the number of "burn-in" draws for the Gibbs sampling algorithm in the first time slice to $1,000$. After the "burn-in" period, we take another $2,000$ samples for the posterior distribution and save every 10^{th} iteration, resulting in 200 saved draws of the posterior distribution (skip-sampling). Next, we determine the draw with the highest posterior likelihood and take the posterior of the topic-word distribution and document-topic distribution as our estimate of ϕ and θ, respectively.

For the second until the last time slice we do a maximum of $1,000$ iterations, and save every 10^th iteration, resulting in a maximum of 100 saved draws of the posterior distribution. We stop drawing from the posterior distribution if the increase in the likelihood for a draw is less than $1e^{-9}$ with respect to the likelihood ten draws earlier. This stopping algorithm leads to less than $1,000$ iterations for all time slices > 1. Again, we determine the draw with the highest posterior likelihood and take the posterior of the topic-word distribution and document-topic distribution as our estimate of ϕ and θ.

C Dynamic factor model

This section describes the model equations of the nowcasting model in the main text, i.e. a dynamic factor model, the choice of the number of factors. Moreover, this section describes the state space representation of the dynamic factor model.

C.1 Model equations

We use the dynamic factor model specification in Bańbura et al. (2011). The main equation are as follows.

$$\mathbf{x}_m = \Lambda \mathbf{f}_m + \boldsymbol{\xi}_m, \qquad \boldsymbol{\xi}_m \sim N(0, \Sigma_\xi) \tag{C.1}$$

which relates the n monthly indicators $\mathbf{x}_m = (\mathbf{x}_{1,m}, \dots, \mathbf{x}_{n,m})'$ to r monthly static factors $\mathbf{f}_m = (\mathbf{f}_{1,m}, \dots, \mathbf{f}_{r,m})'$ via an $n \times r$ matrix of factor loadings Λ and an idiosyncratic component $\boldsymbol{\xi}_m = (\xi_{1,m}, \dots, \mathbf{xi}_{n,m})'$, where $r << n$. m is a monthly time index and the monthly indicators $\mathbf{x}_{i,m}$ are normalized three-month growth rates or differences. The DFM assumes that the idiosyncratic components are a multivariate white noise process, hence the covariance matrix Σ_ξ is diagonal. Furthermore, the DFM assumes that the factors follow a vector-autoregressive process of order p:

$$\mathbf{f}_m = \sum_{s=1}^{p} A_s \mathbf{f}_{m-s} + \zeta_m, \qquad \zeta_m \sim N(0, Q) \tag{C.2}$$

where A and Q are square $r \times r$ matrices. The final equation links the factors to mean-adjusted real GDP growth:

$$\mathbf{y}_m = \boldsymbol{\beta}' \mathbf{f}_m + \varepsilon_m, \qquad \varepsilon_m \sim N(0, \sigma_\varepsilon^2) \tag{C.3}$$

where $\mathbf{y}_m$ denotes the (unobserved) three-month growth rate of monthly real GDP. t is a quarterly time index. Quarterly real GDP growth in quarter t, $\mathbf{y}_t^Q$, is assigned to the third month of the quarter, i.e. month $3t$ on the monthly time scale. The relation between the quarterly GDP growth rate and quarter-on-quarter latent monthly GDP growth rates is given by

$$\mathbf{y}_t^Q = \frac{1}{3}(y_{3t} + y_{3t-1} + y_{3t-2}) \tag{C.4}$$

C.2 Number of factors

To estimate the model we need to specify the number of static common factors r, the number of dynamic factors q and the number of lags p in the factor VAR process. See Figure C.1 for a "scree plot" that indicates the explained variance for 1–20 factors. The plot clearly indicates that the model should include at least 3 factors. We refrain from choosing a particular combination of r, q and p to avoid potential misspecification and instability problems, see also Kuzin et al. (2013) and Jansen et al. (2016). Instead, we estimate a set of models for different combinations of r, q and p. We set the largest possible value of r and q at 5. Given the outcome of the scree test, we set the minimum number of static factors at 3. We set the maximum value of p at 2. This results in a total of 48 model forecasts.[12] The unweighted average of these forecasts form our baseline forecasts.[13] We take as our DFM forecast the (unweighted) average of the forecasts generated by all model specifications (108 in total). This strategy avoids any hindsight bias.

The dynamic factor model of Bańbura and Rünstler (2011) specified in equations (C.1)–(C.4) is estimated in three steps. In the first step we obtain the factors loadings Λ and initial estimates of the static factors $\widehat{f}_m$, applying a static principal components analysis to a balanced subsample of x_m.[14] In the second step we estimate the coefficient matrices A_s in equations (C.2) using $\widehat{f}_m$, and β in equation (C.3) by using a quarterly version of equation (C.3).[15] In the third step, we cast the model in state space form and use the Kalman filter and smoother to re-estimate the estimated factors ($\widehat{f}_m$) and monthly GDP growth.

We calculate forecasts of quarterly GDP growth by applying equation (C.3) to forecasts of monthly factors generated by equation (C.2), and then aggregate to quarterly values. The state-space setup of our dynamic factor model is outlined in the next section. See Bańbura and Rünstler (2011) for a more detailed treatment of the dynamic factor model and the estimation procedures.

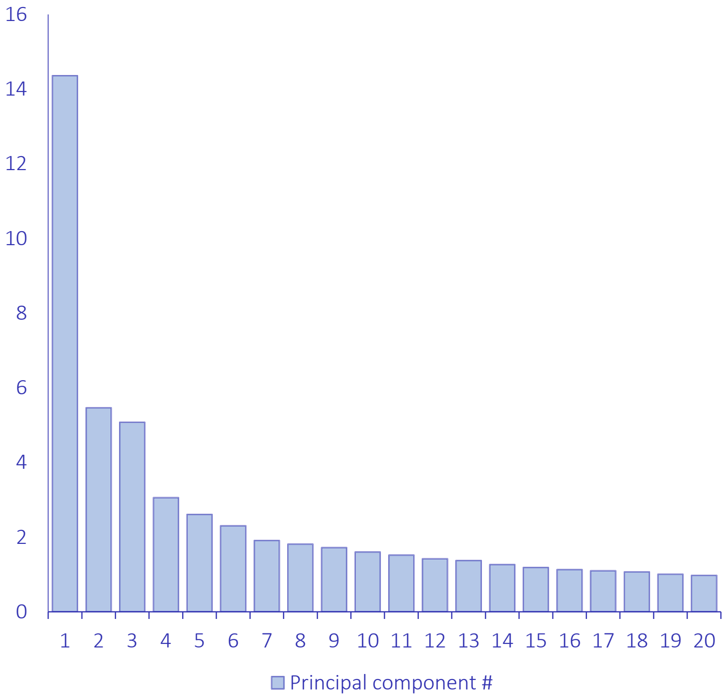
[12] Note that $q \leq r$.

[13] Jansen et al. (2016) found that the forecasting power of the dynamic factor model increases if r increases (until at least six), while it hardly changes if p increases. A different approach is to choose the number of factors on the basis of in-sample criteria, as described in Bai and Ng (2002). Bańbura and Rünstler (2011) and Jansen et al. (2016) report that these procedures tend to result in more volatile and less accurate forecasts.

[14] The balanced sub-sample is obtained by discarding the rows in x_m that contain missing observations due to publication delays.

[15] As in Bańbura and Rünstler (2011), β is estimated via the regression $y_t^Q = \beta f_t^Q + \varepsilon_t^Q$, where f_t^Q are three-month averages of sample estimates of f_m using the aggregation rule in equation (C.4). σ_ε^2 is estimated as the sample variance of ε_t^Q divided by 3.

Figure C.1: Percentage of variance explained for 1–20 factors.



C.3 State-space representation

The equations of the DFM, equations (C.1)–(C.4) in the main text, can be cast in state space form, as illustrated below for the case of $p = 1$. The aggregation rule is implemented in a recursive way in equation (C.6) by introducing a latent cumulator variable $\hat{\mathbf{y}}^Q{}_m = \Xi_m \hat{y}^Q_{m-1} + \frac{1}{3} y_m$, where $\Xi_m = 0$ for m corresponding to the first month of the quarter and $\Xi_m = 1$ otherwise (see Bańbura and Rünstler, 2011). The monthly state space representation is given by the following observation equation:

$$\begin{bmatrix} \mathbf{x}_m \\ \mathbf{y}^Q{}_t \end{bmatrix} = \begin{bmatrix} \Lambda & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{f}_m \\ \mathbf{y}_m \\ \hat{\mathbf{y}}^Q_m \end{bmatrix} + \begin{bmatrix} \xi_m \\ 0 \end{bmatrix} \tag{C.5}$$

and the transition equation:

$$\begin{bmatrix} I & 0 & 0 \\ -\beta' & 1 & 0 \\ 0 & -\frac{1}{3} & 1 \end{bmatrix} \begin{bmatrix} f_{m+1} \\ y_{m+1} \\ \hat{y}^Q_{m+1} \end{bmatrix} = \begin{bmatrix} A & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Xi_{m+1} \end{bmatrix} \begin{bmatrix} f_m \\ y_m \\ \hat{y}^Q_m \end{bmatrix} + \begin{bmatrix} \zeta_{m+1} \\ \varepsilon_m \\ 0 \end{bmatrix} \tag{C.6}$$

The application of the Kalman filter and smoother provides the minimum mean square linear estimates (MSLE) of the state vector $\boldsymbol{\alpha}_m = (f_m, y_m, \hat{y}^Q_m)$ and enables the forecasting of quarterly GDP growth y^Q_t and dealing efficiently with an unbalanced data set of missing observations at the beginning and at the end of the series by replacing the missing data with optimal predictions. Moreover, when compared with using principal components technique alone, the two-step estimator allows for dynamics of the common factors and cross-sectional heteroskedasticity of the idiosyncratic component.