Daily news sentiment and monthly surveys: A mixed–frequency dynamic factor model for nowcasting consumer confidence

Andres Algaba[a,b,*], Samuel Borms[a], Kris Boudt[a,b,c], Brecht Verbeken[a]

[a] *Faculty of Social Sciences and Solvay Business School, Vrije Universiteit Brussel, Belgium*
[b] *Department of Economics, Ghent University, Belgium*
[c] *School of Business and Economics, Vrije Universiteit Amsterdam, the Netherlands*

Abstract

Policymakers, firms, and investors closely monitor traditional survey–based consumer confidence indicators and treat it as an important piece of economic information. To obtain a daily nowcast of monthly consumer confidence, we introduce a latent factor model for the vector of monthly survey–based consumer confidence and daily sentiment embedded in economic media news articles. The proposed mixed–frequency dynamic factor model uses a Toeplitz correlation matrix to account for the serial correlation in the high–frequency sentiment measurement errors. We find significant accuracy gains in nowcasting survey–based Belgian consumer confidence with economic media news sentiment.

Keywords: dynamic factor model, mixed–frequency, nowcasting, sentometrics, state space, Toeplitz matrix.

JEL classification: C32, C51, C53, C55.

*We thank the Editor and two anonymous referees, as well as David Ardia, Raïsa Basselier, Keven Bluteau, Nabil Bouamara, Leopoldo Catania, Selien De Schryder, Eric Ghysels, Koen Inghelbrecht, Hande Karabiyik, Siem Jan Koopman, Geert Langenus, Geoffrey Minne, Juan Rubio Ramírez, Peter Reusens, James Thewissen, Steven Vanduffel, Jeroen Van Pelt, Marjan Wauters, and Raf Wouters for stimulating discussions and feedback on earlier drafts of this work. We further thank seminar participants at Ghent University, Vrije Universiteit Brussel, and the National Bank of Belgium, as well as participants at the 2019 CFE conference in London and the 2020 SoFiE summer school in Chicago. We are grateful to the Belgian News Agency (Belga) for providing us with their media news archive. Part of this research was conducted while Andres Algaba was a visiting researcher at the National Bank of Belgium. This project benefited from financial support from the National Bank of Belgium, the Swiss National Science Foundation (`https://www.snf.ch`, grant #17928), and Innoviris. Any remaining errors or shortcomings are those of the authors.

*Corresponding author: Andres Algaba (andres.algaba@vub.be). Pleinlaan 2, 1050 Brussel, Belgium.

"Americans reading the paper, listening to the news every single day, and all you hear is things are getting worse and worse. And that has a psychological effect on consumer confidence. That's what consumer confidence is."

– Howard Schultz (former Chairman and CEO of Starbucks Coffee Corporation)

1. Introduction

The confidence of consumers towards the future state of the economy guides their decision–making and ultimately impacts consumption, production, investment, and other relevant macroeconomic outcomes. It is traditionally measured through a national survey in which the respondent's outlook on personal and general economic developments is questioned (see e.g., Ludvigson, 2004). This kind of surveys are conducted over multiple days and thus give an aggregated view on the sentiment within a past period. This implies that the subsequent indicators are published at a low frequency and with a substantial release lag. It seems self–evident that their accuracy and timeliness can be improved by augmenting the low–frequency survey information with the daily sentiment embedded in media news articles. However, such a data augmentation approach requires a flexible model that can accommodate for the lack of a precise high–frequency timestamp of the low–frequency indicator, the high variability in the sentiment data, and the arbitrary pattern of days with missing sentiment information.

Our solution to this problem consists of modelling the high–frequency daily sentiment variables and the low–frequency survey–based indicator jointly as a monthly vector driven by a common latent consumer confidence factor. To account for the serial correlation of the measurement errors of economic media news sentiment, we provide an extension to the Toeplitz correlation matrix (see e.g., Mukherjee and Maiti, 1988). This extension allows for AR(1) dynamics in the autocorrelation of the high–frequency measurement errors, and puts a bound on the correlation between the high– and low–frequency measurement errors to ensure positive definiteness of the resulting correlation matrix. The combined use of survey data and economic media news sentiment leads to a more timely and frequent estimation of the latent factor, and nowcasts of survey–based consumer confidence.

The proposed mixed–frequency Dynamic Factor Model (DFM) complements the current literature on the use of a DFM for nowcasting economic variables in a mixed–frequency setting. Aruoba et al. (2009) show the usefulness of a DFM approach by blending low– and high–frequency economic data into a latent coincident index that tracks real business conditions at high observation frequency. Bańbura and Modugno (2014) and Hindrayanto et al. (2016) find that a mixed–frequency DFM with monthly and quarterly indicators is effective for nowcasting the quarterly euro area GDP growth rate. For an application with textual data, we refer to Thorsrud (2020) who decomposes daily newspaper data into sentiment–adjusted news topic variables, and subsequently uses those with quarterly GDP growth in a factor model with dynamic sparsity to construct a daily business cycle index.

We show the practical usefulness of the proposed framework for nowcasting survey–based Belgian consumer confidence. The daily economic media news sentiment variables are constructed using the media archive of the national Belgian News Agency (Belga). This archive contains around 40 million media news articles in Dutch and French over the period November 2001 until April 2020. We apply keyword filters to only select media news articles that are related to consumer confidence (see e.g., Baker et al., 2016). To extract the sentiment from the economic media news articles, we use a lexicon that we obtain via annotation of Belga news articles from January 2005 until December 2011. In an out–of–sample exercise from January 2012 until April 2020, we find significant nowcasting accuracy gains by using economic media news sentiment in combination with the extended Toeplitz correlation matrix. The recent COVID–19 pandemic serves as an interesting illustration to show the usefulness of different specifications of the proposed mixed–frequency model. We find that in crisis periods when economic indicators can be subject to sudden and rapid changes, the estimation of the latent factor may benefit from a larger sensitivity to on the high–frequency information.

The remainder of this paper is organized as follows. In Section 2, we introduce the mixed–frequency DFM with the Toeplitz correlation matrix and show how it can be used to construct a latent consumer confidence coincident index, and to nowcast survey–based

consumer confidence. In Section 3, we present an empirical application for consumer confidence and find that the proposed model implemented using economic media news sentiment is useful for nowcasting survey–based consumer confidence. Section 4 concludes.

2. Estimating real–time consumer confidence

In this section, we present our framework for nowcasting a latent low–frequency factor driving the observations of high– and low–frequency variables. In our application, this is latent consumer confidence with as observables monthly survey–based consumer confidence and daily economic media news sentiment. Other possible applications for our approach include, for example, nowcasting quarterly GDP growth. Central banks publish their flash estimate towards the end of the quarter. Economic sentiment data can then be used to nowcast the GDP growth data, as in Barbaglia et al. (2021). Another application is the monitoring of government popularity. Low–frequency survey results can be complemented by the sentiment about the government in newspapers. A further direction for research is to apply the proposed model with other high–frequency time series based on mobility data, electricity data or bank transaction data.

2.1. Notation

Our variable of interest is monthly (latent) consumer confidence, which we denote by α_t for month $t = 1, 2, \ldots, T$. It represents the average consumer confidence over the month. Let y_t be an observable proxy variable for α_t. The observations of y_t are often an estimate of consumer confidence measured via a survey over (all, or a part, of) the days i in each month t, with $i = 1, 2, 3, \ldots, d$. Note that d can be time–varying, i.e., d_t, but for simplicity of notation we will use d throughout this paper. We also have a high–frequency proxy based on daily economic media news sentiment. Denote these by $m_{t,i}$ for each day i in month t. We then stack all observables for a given month in the

$n \times 1$ monthly observation vector $\boldsymbol{y}_t$ as follows:[1]

$$\boldsymbol{y}_t = \left[m_{t,1}, m_{t,2}, ..., m_{t,d}, y_t\right]'. \tag{1}$$

All variables are assumed to be covariance–stationary, and standardized with mean zero and unit variance. A suitable model for $\boldsymbol{y}_t$ needs to account for the commonality in the proxies, the differene in precision of the proxies, and the serial correlation in the measurement errors of $m_{t,i}$. The order of the variables in $\boldsymbol{y}_t$ matters to account for the order dependence of the Cholesky decomposition which is performed during the estimation of our model. In our application, we construct pseudo–months ending on the last day of the survey. Consistent with that calender definition, we recommend to put the survey–based variable last since it reflects the average consumer confidence over the period. This order also replicates most closely the real–time news-–flow, with media news articles coming in daily and survey–based consumer confidence being released at the end of each month.

2.2. A mixed–frequency DFM with a Toeplitz correlation matrix

We propose a mixed–frequency DFM where the low– and high–frequency observables are all driven by a common low–frequency latent consumer confidence factor through the following state space representation relating the observable variables $\boldsymbol{y}_t$ to the unobserved state of consumer confidence α_t:

$$\boldsymbol{y}_t = \boldsymbol{\lambda}\alpha_t + \boldsymbol{\varepsilon}_t, \quad \text{with} \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{H}\right), \tag{2}$$

where the $n \times 1$ vector $\boldsymbol{\lambda}$ contains the n factor loadings of $\boldsymbol{y}_t$ on α_t. The measurement errors $\boldsymbol{\varepsilon}_t$ are assumed to be normally distributed with mean zero and a $n \times n$ covariance matrix:

$$\boldsymbol{H} = \boldsymbol{DRD}, \tag{3}$$

[1] In the standard mixed–frequency DFM framework the model would be defined at the highest possible frequency and the low–frequency variable would be interpreted as a high–frequency variable with missing data (see e.g., Bańbura and Modugno, 2014). We opt for a different setup as in our model low–frequency consumer confidence is considered to be a flow variable which is a real–time function of the high–frequency economic media news sentiment variables.

where $\boldsymbol{D}$ is an $n \times n$ diagonal matrix with the standard deviations on the diagonal, and $\boldsymbol{R}$ is the $n \times n$ correlation matrix.

The correlation matrix $\boldsymbol{R}$ in its full generality can have $n(n-1)/2$ parameters. We restrict it to two parameters in such a way that it can still accommodate the empirical fact that there is a positive autocorrelation in the errors of economic media news sentiment due to the presence of both news cycles and economic cycles. The parsimony is achieved by assuming that the high–frequency measurement errors follow an AR(1) process. It follows that the autocorrelation between the economic media news sentiment variables decreases exponentially with the absolute lag difference between the days. Note that while we allow the autocorrelation coefficient r_2 to be either positive or negative, we implicitly assume that daily economic media news sentiment is positively serially correlated, i.e., high (low) sentiment days are more likely to be followed by high (low) sentiment days. To formalize this AR(1) process in matrix form, we consider a Toeplitz correlation matrix which has the distinctive property that the elements only depend on the differences of the indices (see e.g., Mukherjee and Maiti, 1988):

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_2 & r_2^2 & \dots & r_2^{n-2} & r_1 \\ r_2 & 1 & r_2 & \ddots & \vdots & r_1 \\ r_2^2 & r_2 & \ddots & \ddots & r_2^2 & r_1 \\ \vdots & \ddots & \ddots & 1 & r_2 & \vdots \\ r_2^{n-2} & \dots & r_2^2 & r_2 & 1 & r_1 \\ r_1 & r_1 & r_1 & \dots & r_1 & 1 \end{bmatrix}. \tag{4}$$

To the best of our knowledge, the properties of this correlation matrix have not been studied elsewhere in the literature. The determinant of the correlation matrix $\boldsymbol{R}$ in Equation (4) is given in Lemma 1.

Lemma 1. *The determinant of the $n \times n$ matrix $\boldsymbol{R}$ is given by:*

$$\det(\boldsymbol{R}) = (1-r_2)^{(n-2)}(1+r_2)^{(n-3)}\left(1+nr_1^2(r_2-1)+(r_1^2+r_2-3r_1^2r_2)\right).$$

The proof is given in Appendix A. Note that the function is decreasing in n and that to ensure positive definiteness of $\boldsymbol{R}$, we thus need parameter restrictions for r_1 and r_2. We have the following corollary that gives the upper and lower bound for r_1 given $r_2 \in (-1, 1)$.

Corollary 1. *The $n \times n$ matrix $\boldsymbol{R}$ is a positive–definite correlation matrix if and only if $r_2 \in (-1, 1)$ and:*

$$r_1 \in \left(-\sqrt{\frac{1 + r_2}{(n-1) - (n-3)r_2}}, \sqrt{\frac{1 + r_2}{(n-1) - (n-3)r_2}} \right).$$

The proof is given in Appendix B. Note in Equation (3) that the positive definiteness of $\boldsymbol{H}$ is guaranteed when $\boldsymbol{R}$ is positive–definite as all the elements on the diagonal matrix $\boldsymbol{D}$ are positive.

Figure 1 shows an illustration of the upper and lower bound of r_1 given $n = 5,\ 10,\ 30$ and 50. The upper (lower) bound starts at 0 when $r_2 = -1$, and monotonically increases (decreases) non–linearly. Eventually the upper (lower) bound goes to 1 (-1) when $r_2 = 1$. In general, the bounds for r_1 are larger in absolute value for large values of r_2, and small values of n.[2] We refer the interested reader to Appendix C for a brief discussion of the impact of the mixed–frequency measurement errors covariance matrix on the prediction accuracy.

[2]In the implementation, we impose these bounds using parameter transformations, as in Koopman et al. (2018) and Buccheri et al. (2020). The transformed unconstrained parameters are r_1^* and r_2^* which can take any real value. The back–transformation is:

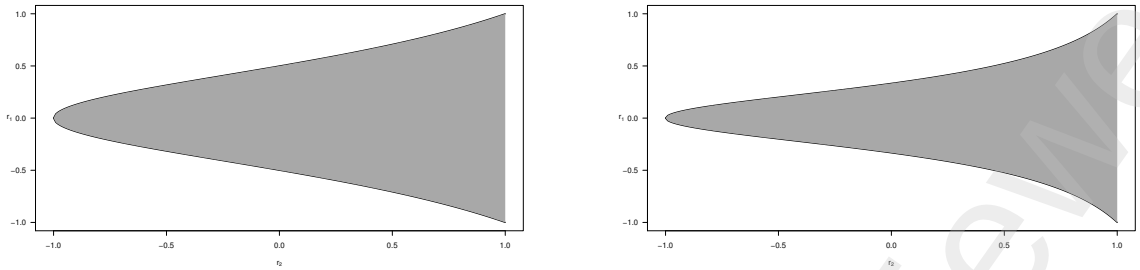
$$r_2 = \tanh(r_2^*), \text{ and } r_1 = \frac{1}{2}\left[(a+b) + (a-b)\tanh(r_1^*)\right],$$

where tanh denotes hyperbolic tangent, and a and b are the maximum and minimum allowed value for r_1, respectively. Following Corollary 1, this leads to the following formulation for r_1:

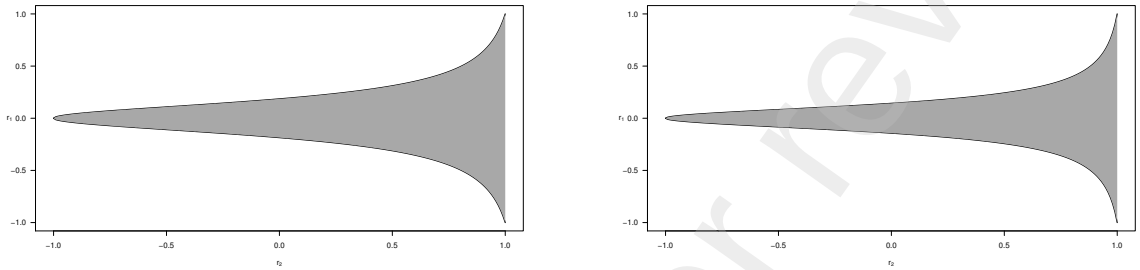
$$r_1 = \tanh(r_1^*)\sqrt{\frac{1 + r_2}{(n-1) - (n-3)r_2}}.$$

Figure 1: Upper and lower bounds of r_1 given $r_2 \in (-1, 1)$ for different values of n.

(a) $n = 5$. **(b)** $n = 10$.



(c) $n = 30$. **(d)** $n = 50$.



Note: The shaded area indicates the allowed parameter space for r_1 given $r_2 \in (-1, 1)$. The black lines are the upper and lower bounds.

2.3. Additional assumptions

We make the common assumption that the unobserved state of consumer confidence α_t follows an autoregressive process of order one with AR(1) coefficient ρ:

$$\alpha_t = \rho\alpha_{t-1} + \eta_t, \quad \text{with} \quad \eta_t \sim \mathcal{N}\left(0, \sigma_\eta^2\right), \tag{5}$$

where the innovation shocks η_t are normally distributed with mean zero and variance σ_η^2. We further assume that the error terms $\boldsymbol{\varepsilon}_t$ and η_t are uncorrelated with each other for identification purposes (see e.g., Harvey, 1989). The normality assumption is quite natural from two points of view. First, since the observables are an average across many observations, (approximate) normality follows from the central limit theorem. Second, the normality assumption leads to a more reactive filter than when a fat-tailed distributed is assumed (see e.g., Creal et al., 2013).

To implement this mixed–frequency DFM in practice, we need to account for the distinct features of textual data, such as the high variability in the sentiment data and the arbitrary pattern of days with missing sentiment information. To deal with these

features and to avoid the curse of dimensionality, we impose some structure on the factor loadings $\boldsymbol{\lambda}$ and the covariance matrix of the measurement errors $\boldsymbol{H}$.

For $\boldsymbol{\lambda}$, we restrict the factor loading of the low–frequency variable to be equal to one to identify the sign and size of α_t (see e.g., Bai and Wang, 2015). Further, we assume that daily economic media news sentiment is, on average, of equal importance across all days i of each month t, and set the d factor loadings of the high–frequency variables all equal to λ. This leads to the following structure for the $n \times 1$ vector $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda \iota_{n-1} \\ 1 \end{bmatrix}, \tag{6}$$

where ι_{n-1} is a $(n-1)$–dimensional vector of ones.

For the covariance matrix $\boldsymbol{H}$, we have already described the parsimonious specification of the correlation matrix $\boldsymbol{R}$ in Section 2.2. We further assume that daily economic media news sentiment exhibits, on average, the same volatility across all days i of each month t. Therefore, we set the d standard deviations of the high–frequency variables all equal to σ_{ε_2}. This leads to the following structure for $\boldsymbol{D}$:

$$\boldsymbol{D} = diag\{\sigma_{\varepsilon_2} \iota_{n-1}, \sigma_{\varepsilon_1}\}, \tag{7}$$

where σ_{ε_1} denotes the standard deviation of survey–based consumer confidence, and $diag\{\cdot\}$ creates a diagonal matrix.

In this section we have imposed strong restrictions on $\boldsymbol{\lambda}$ and $\boldsymbol{H}$. They can be relaxed to take into account that there is a day–of–the–month effect in the exposure of economic media news sentiment to the latent factor α_t, or when some days tend to be associated with a higher measurement error variance. We recommend the analysis of this as a direction for further research.

2.4. Estimation

We use the Kalman filter to compute filtered estimates of the conditional mean and variance of latent consumer confidence α_t given $\boldsymbol{y}_t$, i.e., $a_{t|t} = \mathbb{E}[\alpha_t|\boldsymbol{y}_t]$ and $p_{t|t} = \mathrm{Var}[\alpha_t|\boldsymbol{y}_t]$,

and the one–step ahead forecasts, i.e., $a_{t+1|t} = \mathbb{E}[\alpha_{t+1}|\boldsymbol{y}_t]$ and $p_{t+1|t} = \text{Var}[\alpha_{t+1}|\boldsymbol{y}_t]$. The Kalman filter equations are given by:

$$\begin{aligned}
&\boldsymbol{v}_t = \boldsymbol{y}_t - \boldsymbol{\lambda} a_{t|t-1}, && \boldsymbol{F}_t = \boldsymbol{\lambda} p_{t|t-1} \boldsymbol{\lambda}^\intercal + \boldsymbol{H}, \\
&\boldsymbol{K}_t = p_{t|t-1} \boldsymbol{\lambda}^\intercal \boldsymbol{F}_t^{-1}, && \\
&a_{t|t} = a_{t|t-1} + \boldsymbol{K}_t \boldsymbol{v}_t, && p_{t|t} = p_{t|t-1} \left(1 - \boldsymbol{K}_t \boldsymbol{\lambda}\right), \\
&a_{t+1|t} = \rho a_{t|t}, && p_{t+1|t} = \rho^2 p_{t|t} + \sigma_\eta^2,
\end{aligned} \tag{8}$$

where $\boldsymbol{v}_t$ denotes an $n \times 1$ vector with the forecast errors of $\boldsymbol{y}_t$, $\boldsymbol{F}_t$ is the $n \times n$ covariance matrix of the forecast errors, and $\boldsymbol{K}_t$ is referred to as the $1 \times n$ Kalman gain vector.

The model parameters can be estimated by a Maximimum Likelihood procedure. As the error terms are assumed to be normally distributed, we obtain the Gaussian log-likelihood function via the forecast error decomposition. The loglikelihood can be easily computed by a routine application of the Kalman filter (see e.g., Durbin and Koopman, 2012). In our case, the initial conditions are unknown, and a diffuse initialization procedure is required. Therefore, we opt for an exact initialization with diffuse priors where an exact initial Kalman filter is derived as in Koopman and Durbin (2003). The effect of the initial conditions vanishes rapidly and the filter then reduces to a standard Kalman filter.

2.5. Updating estimates at a daily frequency

The variable of interest can be either the latent factor α_t or the low–frequency variable y_t. The former case corresponds to constructing a latent consumer confidence coincident index, while the latter consists of nowcasting survey–based consumer confidence. In the nowcasting case, the model simplifies since the variance of the low–frequency measurement errors ($\sigma^2_{\varepsilon_1}$) is set equal to zero.[3]

Our approach allows for daily updates of the latent consumer confidence coincident index and the nowcast of survey–based consumer confidence as we add the observations $m_{t,i}$ to the observation vector in real time, and y_t at the end of each month t (at the earliest if we assume there is no release lag). Even if the daily economic media news

[3] As shown in Corollary 1, the covariance matrix of the measurement errors will always be positive–definite in this case.

sentiment variables did not exhibit arbitrary patterns of missing data, we would still need to account for many missing values as most of the time we filter with partial information for the month t (the problem of the so–called "jagged" or "ragged" edge). To handle filtering with partial data, we apply a sequential processing approach that allows for a time–varying length n of the observation vector $\boldsymbol{y}_t$ (Koopman and Durbin, 2000). In the sequential processing approach, the elements of the observation vector $\boldsymbol{y}_t$ are brought into the analysis one at a time, thus in effect converting the multivariate time series into a univariate time series. We first diagonalize the covariance matrix of the measurement errors $\boldsymbol{H}$ via the Cholesky decomposition. We then transform the observation vector $\boldsymbol{y}_t$ accordingly such that the measurement errors are uncorrelated and the multivariate state space model can be treated as a univariate time series. Note that this approach also deals with the time–varying number of days in each month t (i.e., d_t).

3. Application to consumer confidence in Belgium

In this section, we perform an out–of–sample empirical application for Belgium over the period November 2001 until April 2020. First, we present monthly survey–based consumer confidence as measured by the National Bank of Belgium (NBB) which is currently the most prominent proxy of latent consumer confidence in Belgium. Next, we present the daily economic media news sentiment variables which are constructed from a rich media news archive that we obtain from the Belgian News Agency (Belga). Finally, we perform an out–of–sample nowcasting exercise on Belgian survey–based consumer confidence and illustrate the additional insights of monitoring the latent consumer confidence coincident index during the COVID–19 pandemic.

3.1. Survey–based consumer confidence

The National Bank of Belgium (NBB) measures consumer confidence in Belgium via a monthly survey. A stratified sampling technique is used to draw 1850 people each month on the basis of the public telephone directory. The survey is conducted in the first two weeks, and the results are published in the third week, of each month. Since November 2001, the questionnaire consists of the following four questions that assess

the twelve month forward–looking expectations around general economic developments, employment, savings and the financial situation of households:

- "How do you expect the general economic situation in Belgium to develop over the next twelve months?"
- "What do you think will happen to unemployment in Belgium over the next twelve months?"
- "How do you expect the financial position of your household to change over the next twelve months?"
- "Do you think that you will be able to put any money by, i.e., save, over the next twelve months?"

Respondents can choose between five possible answers on each question. Let PP_t stand for the percentage of respondents answering "much better" (or "total certainty"), P_t for "better", MM_t for "much worse" and M_t for "worse", then $Balance_t$ can be stated as follows:

$$Balance_t = (PP_t + 0.5P_t) - (MM_t + 0.5M_t)\,. \tag{9}$$

Monthly survey–based consumer confidence (y_t) is defined as the arithmetical average of the seasonally adjusted $Balance_t$ for the four questions over the period November 2001 until April 2020. Note that the fifth possible answer, which is "neutral", is not directly used in the computation of the consumer confidence indicator.

3.2. Economic media news sentiment

The use of economic media news sentiment as a proxy for consumer confidence is supported by the media dependency theory (Ball-Rokeach and DeFleur, 1976). This theory states that by reporting on current events, the media makes information about the (future) state of the economy more available to consumers and thereby influences their perception. We define economic media news sentiment as the polarity and strength of the sentiment that the media expresses about certain (economic) subjects and actors. It can

be measured via textual sentiment analysis which is a branch of the broad field of Natural Language Processing (NLP).

Belgium has three official languages, namely Dutch, French and German, of which the latter is the least prevalent primary language, spoken natively by less than 1% of the population. Therefore, we focus on the around 40 million media news articles in Dutch and French over the period November 2001 until April 2020 from the Belga archive. Besides text, the news articles are also tagged with relevant metadata, such as the publication date and news source. Since not all the articles are related to consumer confidence, we use some criteria to select a corpus which is only a subset of this text universe. First, we only select the twelve most popular newspapers in both Dutch and French which have been in the archive since November 2001.[4] This selection reduces the number of articles to 21 million. Next, we apply some keyword filters similar in spirit to the creation of the Economic Policy Uncertainty (EPU) index by Baker et al. (2016).[5] The keyword filters consist of four layers which ensure that we only select articles that are related to: 1) economic subjects, and 2) consumer confidence, and 3) Belgium, and 4) we apply a last filter to reduce the number of false positives.[6] The final corpus size is 234,000 news articles.

For each of the news articles in our final corpus, we compute the sentiment by using a lexicon approach which is a standard practice in sentiment analysis (see e.g., Algaba et al., 2020a). Let w_{j_a} be the polarity of a word j_a in a news article a with a total number of J_a words that convey a polarity, and v_{j_a} be a preceding valence shifter which may adjust

[4]For Dutch these are seven newspapers, namely "Het Laatste Nieuws", "Het Nieuwsblad", "De Standaard", "De Morgen", "De Tijd", "Het Belang van Limburg" and "De Gazet van Antwerpen". For French these are five newspapers, namely "Le Soir", "La Dernière Heure", "L'Avenir", "L'Echo" and "La Libre Belgique". The overweighting of Flemish versus French newspapers is consistent with the higher number of Dutch speaking people in Belgium.

[5]Algaba et al. (2020b) use the same media news archive to construct an EPU index for Belgium. See also `http://policyuncertainty.com/belgium_monthly.html`.

[6]We remove all the articles which do not mention the word "economy" or variants thereof reducing the number of articles to 821,000. To ensure that the articles are specifically related to consumer confidence, we further reduce the selection by only selecting articles that contain certain keywords that are related to general economic developments, employement, savings and the financial situation of households. From the remaining 316,000 articles, we only keep the 258,000 articles that mention keywords that ensure that the article is related to Belgium. Finally, we remove articles from the corpus that are overwhelmingly associated with false positives, e.g., calendars, book and movie reviews, anniversaries, obituaries, etc.

the polarity of a word j_a. The sentiment per media news article s is then computed as:

$$s = \frac{1}{J_a} \sum_{j_a=1}^{J_a} v_{j_a} w_{j_a}. \tag{10}$$

We use a sentiment lexicon for Belgian economic news that we co–developed with the Belgian News Agency (Belga) based on the annotation of media news articles from their archive over the period January 2005 until December 2011. Twenty students were asked to read around 500 articles each, and to mark the most positive and negative words. The 500 most frequent positive and negative words in both Dutch and French were then used to compose the lexicons with a dichotomous (value -1 or 1) polarity.[7] Figure 2 shows a sample of the most frequent positive and negative words translated in English. Next to this lexicon, we also use valence shifters which are negators (value -1), amplifiers (value 1.8) and deamplifiers (value 0.2). We use the valence shifters from the sentometrics R package (Ardia et al., 2021).[8]

To create the daily economic media news sentiment variables $m_{t,i}$, we aggregate the resulting sentiment values by taking the daily average per newspaper and standardize them. We then average over all the values of the newspapers on a given day i in each month t. A missing value occurs if there are no economic news articles in any of the newspapers. However, if in some newspapers there are relevant news articles, the newspapers with no relevant news articles get a sentiment value of zero. When extending the observation vector $\boldsymbol{y}_t$ with the economic media news sentiment variables, we account for the fact that people are only surveyed in the first two weeks of each month by creating pseudo–months from the 15th of the previous month until the 14th of the surveyed month. We then relate

[7] Our target variable is survey–based consumer confidence. Given the limited time span and the high dimensionality of the potentially relevant words expressed in the newspapers every month, a supervised machine learning approach with our low–frequency target variable is not feasible. For a comparison between lexicon–based sentiment computation and supervised machine learning approaches on longer time spans and higher frequency data, we refer to Kalamara et al. (2020). The lexicons are available from the authors upon request.

[8] As an example, consider the sentence: "The National Bank of Belgium states that no positive effect can be expected from the recent regulations", where "no" is a valence shifter, namely a negator with a value of -1, and "positive" is a word with a polarity value of 1. Following Equation (10), the sentiment for this media news article is equal to -1, as we have one positive polarity word accompanied with one valence shifter, i.e., $(-1 \times 1)/1$.

Figure 2: The most frequent positive and negative words (translated in English) in the selected media news articles over the period November 2001 until April 2020.

Note: Green (red) indicates a positive (negative) word and the bigger the word, the more frequent it appears in the media news articles.

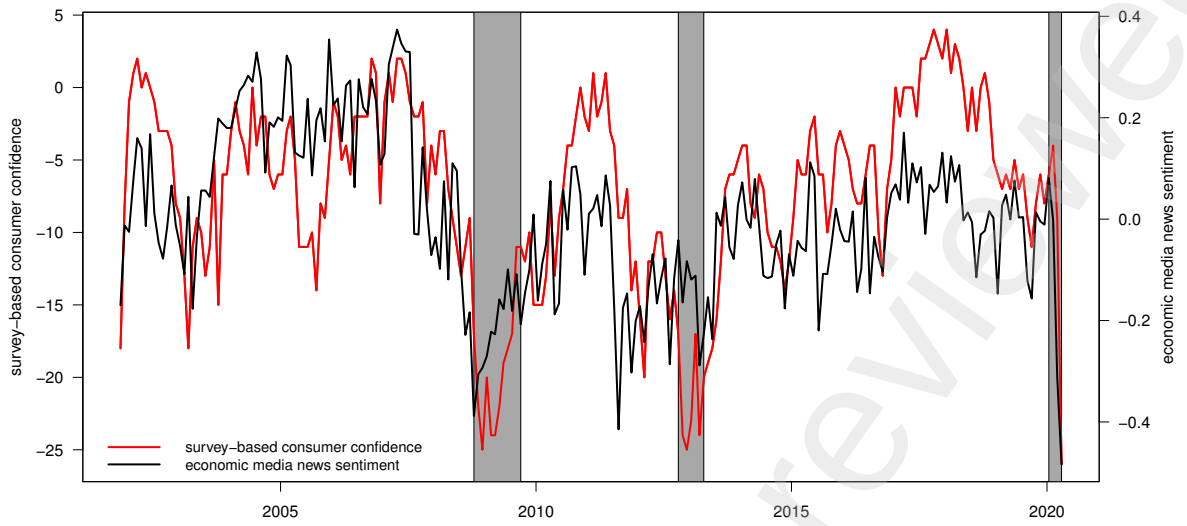
the high–frequency economic media news sentiment variables from the pseudo–months to the corresponding monthly survey–based consumer confidence.

Figure 3 shows the monthly average economic media news sentiment and the monthly survey–based consumer confidence. We see that there is a large degree of comovement between both time series with a contemporaneous correlation of 0.62. Note that both survey–based consumer confidence and economic media news sentiment experience their largest drawdown, and are at their lowest value, in April 2020 during the COVID–19 pandemic.

3.3. Out–of–sample evaluation

We perform an out–of–sample nowcasting exercise on Belgian survey–based consumer confidence. First, we present the competing nowcasting models. Then, we compare the nowcasting accuracy of survey–based consumer confidence of our proposed mixed–frequency DFM with the benchmark models. Finally, we provide anecdotal evidence of the added value of the latent consumer confidence coincident index for tracking consumer confidence during the turbulent period of the COVID–19 outbreak in February–April 2020.

Figure 3: Monthly economic media news sentiment and survey–based consumer confidence over the period November 2001 until April 2020.



Note: The red line indicates monthly survey–based consumer confidence, and the black line is the monthly average of daily sentiment values for the corresponding pseudo–months (right hand side). The shaded areas indicate recession periods defined as two consecutive quarters of negative economic growth as measured by Belgian Gross Domestic Product (GDP).

3.3.1. Nowcasting models

We compare the proposed mixed–frequency DFM with a single–frequency AR(1) model, a MIxed DAta Sampling (MIDAS) model, and with different specifications of the measurement errors covariance matrix in our mixed–frequency DFM.

MIDAS models are parsimonious regression specifications which use exponential lag polynomials for the coefficients. They are often used for nowcasting due to their ease of implementation and relatively good performance compared to other mixed–frequency approaches, such as mixed–frequency DFM and VAR models (see e.g., Foroni and Marcellino, 2014, and Kuzin et al., 2011). For an application of MIDAS with consumer confidence, we refer to Lehrer et al. (2019). In this paper we opt for a MIDAS specification with an AR(1) component which looks as follows:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 \sum_{i=0}^{d-1} w_i^{\theta} m_{t,d-i} + v_t, \tag{11}$$

where v_t is the error term, $m_{t,d-1}$ are the d high–frequency economic media news sentiment variables that have been observed during the corresponding nowcasting month, and w_i^{θ}

denote the weights.[9] For the parameterization of w_i^θ, we use an exponential Almon lag polynomial with positive weights which sum to one. More specifically, we consider the following functional form:

$$w_i^\theta = \frac{\exp(\theta_1 i + \theta_2 i^2)}{\sum_{i=1}^{d} \exp(\theta_1 i + \theta_2 i^2)}, \tag{12}$$

where only two parameters θ_1 and θ_2 have to be estimated. We refer to Ghysels et al. (2007) for a more detailed discussion. The AR(1) model is nested as a special case of the MIDAS implementation.

We use three implementations of the mixed–frequency DFM. In all three, we set the variance of the low–frequency measurement errors ($\sigma^2_{\varepsilon_1}$) equal to zero, as the goal is to nowcast survey–based consumer confidence. In the recommended implementation, we use the Toeplitz correlation matrix of Equation 4. Besides this specification, we also consider a diagonal covariance matrix and an unconstrained covariance matrix without any imposed structure. They cover the two extremes, where in the first case the (auto–)correlation is assumed to be equal to zero, and in the second case a more flexible (auto–)correlation is allowed. The proposed model with the Toeplitz correlation matrix strikes a parsimonious balance between the flexibility of the unconstrained covariance matrix and the simplicity of the diagonal matrix.

We re–estimate the AR(1) model and mixed–frequency DFMs at the end of each pseudo–month at the time that we obtain a new observation of survey–based consumer confidence (y_t) using an expanding estimation window. The MIDAS model is re–estimated each day at the time that we obtain a new observation of economic media news sentiment ($m_{t,d}$). We provide real–time nowcasts for each day i for each out–of–sample pseudo–month $t+1$. Due to its single–frequency nature, the AR(1) model will produce one–step ahead forecasts which remain constant during the entire out–of–sample month $t+1$.

We evaluate the accuracy gains of nowcasting survey–based consumer confidence (y_t) in terms of the Relative RMSE to compare the one–step ahead forecasts of the single–frequency AR(1) model with the daily nowcasts of the mixed–frequency models. More

[9] With the inclusion of the economic media news sentiment variables missing values are introduced which are not automatically handled by the MIDAS approach. Therefore, we impute the missing values with the real–time averages in the estimation of the MIDAS model.

formally, we define the Relative RMSE$_h$ at a daily forecasting horizon h as:

$$\text{Relative RMSE}_h = \frac{\sqrt{\frac{1}{S}\sum_{t=1}^{S}\left(\hat{y}_{t|t,h} - y_t\right)^2}}{\sqrt{\frac{1}{S}\sum_{t=1}^{S}\left(\hat{y}_{t|t-1} - y_t\right)^2}}, \tag{13}$$

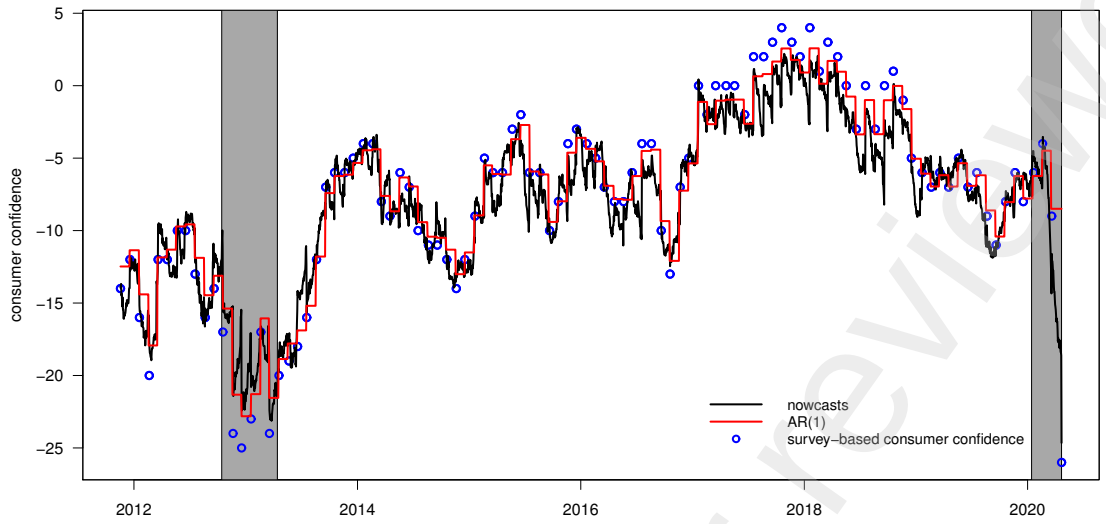
where S is the total number of forecast errors, $\hat{y}_{t|t,h}$ are the daily nowcasts of the mixed–frequency models computed at forecasting horizon h, and $\hat{y}_{t|t-1}$ are the corresponding one–step ahead forecasts of the AR(1) model. To test whether the difference in nowcasting accuracy is statistically significant, we compare the squared loss of our mixed–frequency DFM with Toeplitz correrlation matrix with the competing models jointly at all horizons using the average Superior Predictive Ability (aSPA) (Quaedvlieg, 2021). We use the null hypothesis that our mixed–frequency DFM with Toeplitz correrlation matrix does not outperform the competing models, and a block length of 3 and 999 bootstrap replications as in Quaedvlieg (2021).

3.3.2. Nowcasting survey–based consumer confidence

The first sample used to estimate the models consists of 121 observations from November 2001 until December 2011. The corresponding out–of–sample evaluation sample consists of 101 observations for the period of January 2012 until April 2020. Figure 4 shows the daily nowcasts of our mixed–frequency DFM with Toeplitz correlation matrix, the one–step ahead forecasts of the AR(1) model and survey–based consumer confidence as measured by the National Bank of Belgium. We see that there is substantial intra–monthly movement in the mixed–frequency nowcasts, while the forecasts of the AR(1) model are constant during an entire month t which results in a stepwise pattern.

In Table 1, we show the Relative RMSE for $h = 0, 1, 2, ..., 13$, and also the overall Relative RMSE which is computed by averaging over all the forecasting horizons. We see that, compared to the one–step ahead forecasts of the single–frequency AR(1) model, the addition of high–frequency economic media news sentiment adds value by obtaining more precise nowcasts with all the mixed–frequency models. The modelling of the serial correlation in the measurement errors of high–frequency economic media news sentiment via the Toeplitz matrix seems to also add value as its nowcasting accuracy performs better than

Figure 4: Daily nowcasts of our mixed–frequency DFM with Toeplitz correlation matrix, one–step ahead forecasts of the AR(1) model, and monthly survey–based consumer confidence as measured by the National Bank of Belgium over the period January 2012 until April 2020.



Note: The black line are the daily nowcasts of our mixed–frequency DFM with Toeplitz correlation matrix, the red line represents the one–step ahead forecasts of the AR(1) model, and the blue dots indicate survey–based consumer confidence as measured by the National Bank of Belgium. The shaded area indicates a recession period defined as two consecutive quarters of negative economic growth as measured by Belgian Gross Domestic Product (GDP).

all the competing models at any forecasting horizon. The outperformance compared to all the competing models, except for the diagonal specification, are statistically significant at the 5% significance level according to aSPA test.

3.4. Estimation of the latent coincent index: Application to the COVID–19 pandemic

In the previous section, the goal of the analysis is to nowcast survey–based consumer confidence (y_t). Policymakers may also be interested in monitoring the day–to–day evolution of latent consumer confidence (α_t). For stable months, the nowcast of survey–based consumer confidence and the latent coincident consumer confidence index are similar. However in turbulent months with substantial sentiment dynamics they can differ. The recent COVID–19 pandemic serves as an interesting illustration to show the usefulness of nowcasting a_t in addition to y_t.

To zoom in on the COVID–19 crisis, we consider the last three observations of survey–based consumer confidence which were published by the National Bank of Belgium on 19 February, 20 March, and 21 April, respectively. From 19 February 2020 until 21 April

Table 1: Relative RMSE over the period January 2012 until April 2020.

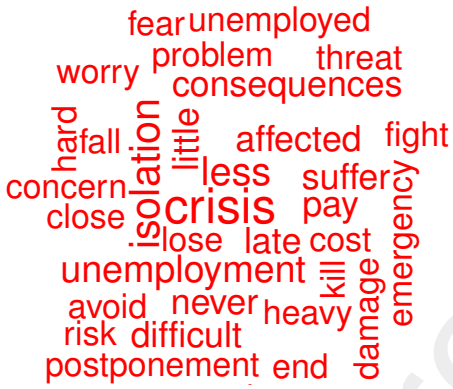
h	Relative RMSE (%)			
	Toeplitz	Unconstrained	Diagonal	MIDAS
0	86.29	96.40	90.92	88.58
1	87.64	96.10	91.15	89.49
2	85.41	92.58	88.69	88.33
3	85.76	92.30	88.77	88.66
4	85.17	89.82	87.86	88.84
5	85.39	90.23	87.77	88.63
6	85.72	90.34	87.57	90.06
7	86.02	90.93	87.31	90.13
8	84.97	91.41	85.94	89.51
9	85.93	91.49	86.56	90.15
10	85.69	91.79	86.05	90.19
11	85.16	91.42	85.66	89.50
12	85.91	91.63	86.59	90.59
13	87.81	92.44	88.19	91.41
Overall	85.91	92.06	87.79	89.58

Note: This table shows the Relative RMSE of the mixed–frequency models compared to the AR(1) model for forecasting horizons $h = 0, 1, 2, ..., 13$. The RMSE of the AR(1) model is 3.28.

2020, 90% of the selected media news articles contain at least one word related to the COVID–19 pandemic, i.e., coronavirus. Figure 5 shows the most frequent negative words appearing in the selected media news articles translated in English. These frequently appearing negative words, such as crisis, suffer, fear and kill, indicate that the negative sentiment corresponds well to what one would expect for the COVID–19 pandemic. We also see that economic related words such as unemployed are among the most frequently appearing negative words.

The first confirmed COVID–19 fatality in Belgium was reported on 11 March, after which the government decided that schools, restaurants and bars would need to shut down from 13 March onwards. On 17 March, the Belgian government decided on a so–called

Figure 5: The most frequent negative words (translated in English) in the selected media news articles over the period February 19 2020 until April 21 2020.



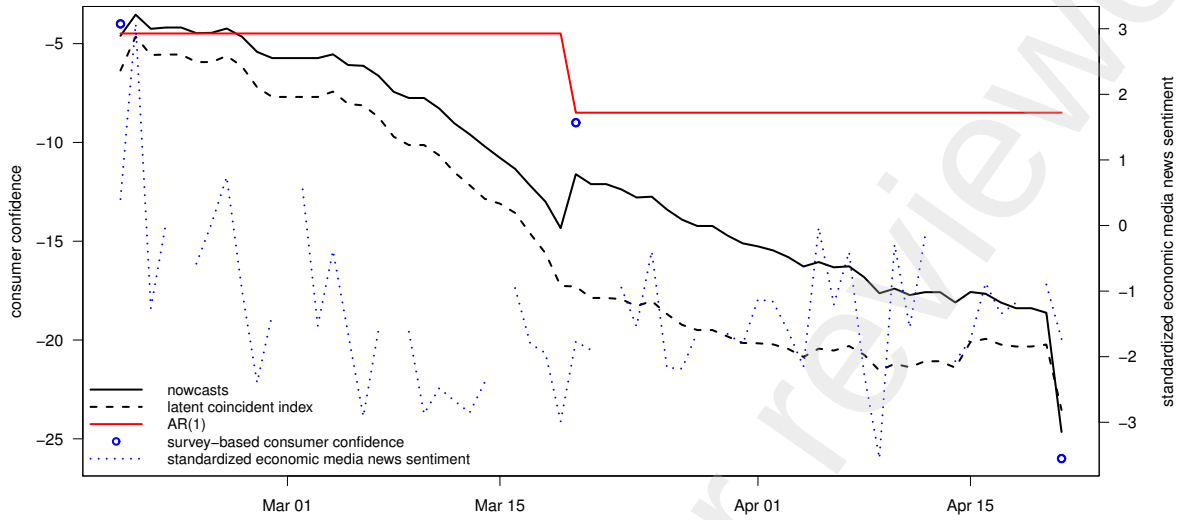
"lockdown light" from 18 march onwards. Some important events thus happened after, or at the end of, the survey period for the consumer confidence indicator of March. In their press release about consumer confidence on 20 March, the National Bank of Belgium explicitly acknowledges this shortcoming of monthly surveys[10]

> "The consumer confidence indicator is the averaged sentiment measured during a survey period of two successive weeks within a month, which runs this month from 2 to 16 March. It therefore does not yet reflect the full impact of the measures adopted by the government to combat the coronavirus. At the end of the survey period, the confidence indicator deteriorated sharply, to such a point that, in the three last days, consumer confidence reached a level close to the historical low (−28)."

The numbers discussed by the National Bank of Belgium are shown in Figure 6, where the blue dots indicate the monthly survey–based consumer confidence as measured by the National Bank of Belgium. The (dotted) black line(s) are the daily nowcasts and latent coincident index of our mixed–frequency DFM with Toeplitz correlation matrix, the red line represents the one–step ahead forecasts of the AR(1) model, and the dotted blue line are the standardized economic media news sentiment observations. We see that during the first half of March, the mixed–frequency DFM specifications correctly assess

[10] See `https://www.nbb.be/doc/dq/e/dq3/histo/pee2003.pdf`.

Figure 6: Daily nowcasts and latent coincident index of our mixed–frequency DFM with Toeplitz correlation matrix, one–step ahead forecasts of the AR(1) model, monthly survey–based consumer confidence, and standardized economic media news sentiment over the period 19 February 2020 until 21 April 2020.



Note: The (dotted) black line(s) are the daily nowcasts and latent coincident index of our mixed–frequency DFM with Toeplitz correlation matrix, the red line represents the one–step ahead forecasts of the AR(1) model, the blue dots indicate the monthly survey–based consumer confidence as measured by the National Bank of Belgium, and the dotted blue line are the standardized economic media news sentiment observations (right hand side).

that consumer confidence is going down. However, the moment that the survey–based consumer confidence for March is published, the DFM specification where the variance of survey–based consumer confidence measurement errors ($\sigma^2_{\varepsilon_1}$) are set equal to zero goes up again while the latent coincident consumer confidence index remains going down. We see that in crisis periods when economic indicators can be subject to sudden and rapid changes, the estimation of the latent factor is characterized by a higher reactivity to the high–frequency information in the economic media news sentiment.

4. Conclusion

Policymakers, firms, and investors closely monitor traditional survey–based consumer confidence indicators and treat it as an important piece of economic information. This kind of surveys are conducted over multiple days and thus give an aggregated view on the sentiment within a past period. This implies that the subsequent indicators are published at a low frequency and with a substantial release lag. To obtain a daily nowcast of

monthly consumer confidence, we introduce a latent factor model for the vector of monthly survey–based consumer confidence and daily sentiment embedded in economic media news articles. The proposed mixed–frequency dynamic factor model uses a Toeplitz correlation matrix to account for the serial correlation of the media news sentiment measurement errors. This allows for AR(1) dynamics in the autocorrelation of the high–frequency measurement errors, and puts a bound on the correlation between the high– and low–frequency measurement errors to ensure positive definiteness of the resulting correlation matrix. In an out–of–sample test of nowcasting survey–based consumer confidence in Belgium from January 2012 until April 2020, we find significant nowcasting accuracy gains by using economic media news sentiment.

References

Algaba, A., Ardia, D., Bluteau, K., Borms, S., Boudt, K., 2020a. Econometrics meets sentiment: An overview of methodology and applications. Journal of Economic Surveys 34, 512–547.

Algaba, A., Borms, S., Boudt, K., Van Pelt, J., 2020b. The Economic Policy Uncertainty index for Flanders, Wallonia and Belgium. BFW digitaal / RBF numérique 2020/6.

Ardia, D., Bluteau, K., Borms, S., Boudt, K., 2021. The R package **sentometrics** to compute, aggregate and predict with textual sentiment. Journal of Statistical Software 99, 1–40.

Aruoba, S., Diebold, F., Scotti, C., 2009. Real-time measurement of business conditions. Journal of Business & Economic Statistics 27, 417–427.

Bai, J., Wang, P., 2015. Identification and Bayesian estimation of dynamic factor models. Journal of Business & Economic Statistics 33, 221–240.

Baker, S.R., Bloom, N., Davis, S.J., 2016. Measuring economic policy uncertainty. The Quarterly Journal of Economics 131, 1593–1636.

Ball-Rokeach, S.J., DeFleur, M.L., 1976. A dependency model of mass–media effects. Communication research 3, 3–21.

Bańbura, M., Modugno, M., 2014. Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. Journal of Applied Econometrics 29, 133–160.

Barbaglia, L., Consoli, S., Manzan, S., 2021. Forecasting GDP in Europe with textual data. Working paper.

Bartlett, M., 1951. An inverse matrix adjustment arising in discriminant analysis. Annals of Mathematical Statistics 22, 107–111.

Buccheri, G., Bormetti, G., Corsi, F., Lillo, F., 2020. A score-driven conditional correlation model

for noisy and asynchronous data: An application to high-frequency covariance dynamics. Journal of Business & Economic Statistics, forthcoming.

Creal, D., Koopman, S.J., Lucas, A., 2013. Generalized autoregressive score models with applications. Journal of Applied Econometrics 28, 777–795.

Durbin, J., Koopman, S.J., 2012. Time series analysis by state space methods. Second ed., Oxford university press, New York.

Foroni, C., Marcellino, M., 2014. A comparison of mixed frequency approaches for nowcasting euro area macroeconomic aggregates. International Journal of Forecasting 30, 554–568.

Ghysels, E., Sinko, A., Valkanov, R., 2007. MIDAS regressions: Further results and new directions. Econometric Reviews 26, 53–90.

Harvey, A.C., 1989. Forecasting, structural time series models and the Kalman filter. Cambridge university press.

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

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., Kapadia, S., 2020. Making text count: Economic forecasting using newspaper text. Bank of England Working Paper No. 865.

Koopman, S., Lit, R., Lucas, A., Opschoor, A., 2018. Dynamic discrete copula models for high-frequency stock price changes. Journal of Applied Econometrics 33, 966–985.

Koopman, S.J., Durbin, J., 2000. Fast filtering and smoothing for multivariate state space models. Journal of Time Series Analysis 21, 281–296.

Koopman, S.J., Durbin, J., 2003. Filtering and smoothing of state vector for diffuse state–space models. Journal of Time Series Analysis 24, 85–98.

Kuzin, V., Marcellino, M., Schumacher, C., 2011. MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area. International Journal of Forecasting 27, 529–542.

Lehrer, S., Xie, T., Zeng, T., 2019. Does high-frequency social media data improve forecasts of low-frequency consumer confidence measures? Journal of Financial Econometrics, forthcoming.

Ludvigson, S.C., 2004. Consumer confidence and consumer spending. Journal of Economic Perspectives 18, 29–50.

Mukherjee, B.N., Maiti, S.S., 1988. On some properties of positive definite Toeplitz matrices and their possible applications. Linear algebra and its applications 102, 211–240.

Quaedvlieg, R., 2021. Multi–horizon forecast comparison. Journal of Business & Economic Statistics 39, 40–53.

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

Appendix A. Proof of Lemma 1

We use mathematical induction to prove that the determinant of the $n \times n$ matrix $\boldsymbol{R}$, which we will further denote by $\boldsymbol{R}_n$, is given by:

$$\det(\boldsymbol{R}_n) = (1-r_2)^{(n-2)}(1+r_2)^{(n-3)}\left(1+nr_1^2(r_2-1)+(r_1^2+r_2-3r_1^2r_2)\right).$$

The case $n = 3$ is the first non–trivial one and an easy calculation shows that indeed $\det(\boldsymbol{R}_3) = (1-r_2)(1+r_2-2r_1^2)$, which settles the base case. Now, for the inductive step, suppose that the claim is true for $n = k$, so that:

$$\det(\boldsymbol{R}_k) = (1-r_2)^{(k-2)}(1+r_2)^{(k-3)}\left(1+kr_1^2(r_2-1)+(r_1^2+r_2-3r_1^2r_2)\right).$$

We will show that the claim holds for the case $n = k+1$ as well, which will settle the proof. Remark that $\boldsymbol{R}_k$ is nothing more than $\boldsymbol{R}_{k+1}$ without the first column and row. Subtracting r_2 times the second–to–first row from the first row of $\boldsymbol{R}_{k+1}$ yields:

$$\det(\boldsymbol{R}_{k+1}) = \begin{vmatrix} 1-r_2^2 & 0 & 0 & \dots & 0 & r_1(1-r_2) \\ r_2 & 1 & r_2 & \ddots & \vdots & r_1 \\ r_2^2 & r_2 & \ddots & \ddots & \vdots & r_1 \\ \vdots & \ddots & \ddots & \ddots & r_2 & \vdots \\ r_2^{k-1} & \dots & \dots & r_2 & 1 & r_1 \\ r_1 & r_1 & r_1 & \dots & r_1 & 1 \end{vmatrix}.$$

Expanding this determinant along the first row yields a sum of two terms, the first one being $(1-r_2)(1+r_2)\det(\boldsymbol{R}_k)$. The second term is given by $(-1)^{k+2}r_1(1-r_2)\det(\boldsymbol{T})$,

where the $k \times k$ matrix $\boldsymbol{T}$ is defined as:

$$\boldsymbol{T} = \begin{bmatrix} r_2^1 & 1 & r_2^1 & r_2^2 & \dots & r_2^{k-2} \\ r_2^2 & r_2^1 & 1 & \ddots & \ddots & r_2^{k-3} \\ r_2^3 & r_2^2 & r_2^1 & \ddots & \vdots & r_2^{k-4} \\ \vdots & \ddots & \ddots & r_2^1 & 1 & \vdots \\ r_2^{k-1} & \dots & \dots & r_2^2 & r_2^1 & 1 \\ r_1 & r_1 & r_1 & \dots & r_1 & r_1 \end{bmatrix}.$$

Now remark that $\boldsymbol{T}$ without the first column and the last row is a $(k-1)\times(k-1)$ Toeplitz matrix, which has determinant $(1-r_2^2)^{k-2}$ (see e.g., Mukherjee and Maiti, 1988). Subtracting r_2 times the second column from the first column before taking the determinant by expanding along the first column yields that:

$$\det(\boldsymbol{T}) = (-1)^{k+1} r_1 (1-r_2)(1-r_2^2)^{k-2}.$$

This implies that the second term is given by:

$$(-1)^{k+2} r_1 (1-r_2) \det(\boldsymbol{T}) = r_1^2 (r_2-1)(1-r_2)^{k-1}(1+r_2)^{k-2}.$$

Taking into account the other term, which was given by:

$$(1-r_2)(1+r_2)\det(\boldsymbol{R}_k) = (1-r_2)^{k-1}(1+r_2)^{k-2}\left(1 + k r_1^2 (r_2-1) + (r_1^2 + r_2 - 3r_1^2 r_2)\right),$$

and combining both terms, yields that:

$$\begin{aligned} \det(\boldsymbol{R}_{k+1}) &= (1-r_2)^{k-1}(1+r_2)^{k-2}\left(r_1^2(r_2-1)\right) \\ &\quad + (1-r_2)^{k-1}(1+r_2)^{k-2}\left(1 + k r_1^2 (r_2-1) + (r_1^2 + r_2 - 3r_1^2 r_2)\right) \\ &= (1-r_2)^{k-1}(1+r_2)^{k-2}\left[r_1^2(r_2-1) + 1 + k r_1^2 (r_2-1) + (r_1^2 + r_2 - 3r_1^2 r_2)\right] \\ &= (1-r_2)^{k-1}(1+r_2)^{k-2}\left[1 + (k+1) r_1^2 (r_2-1) + (r_1^2 + r_2 - 3r_1^2 r_2)\right], \end{aligned}$$

that is, the statement for $n = k + 1$ also holds true, establishing the inductive step and finishing the proof.

Appendix B. Proof of Corollary 1

By Sylvester's theorem, the $n \times n$ matrix $\boldsymbol{R}$ is positive–definite if and only if all upper left $k \times k$ corners of $\boldsymbol{R}$ have a positive determinant, with $2 \leq k \leq n$. From Lemma 1, it follows that:

$$\det(\boldsymbol{R}_k) = (1 - r_2)^{(k-2)}(1 + r_2)^{(k-3)} \left(1 + kr_1^2(r_2 - 1) + (r_1^2 + r_2 - 3r_1^2 r_2)\right).$$

Remark that $\boldsymbol{R}_k$ is nothing more than $\boldsymbol{R}_{k+1}$ without the last column and row. So it suffices to check for every k that $\det(\boldsymbol{R}_k) > 0$, but as this function is decreasing in k for $r_2 \in (-1, 1)$, it is sufficient that $\det(\boldsymbol{R}_n) > 0$. So we have to solve the following inequality:

$$\det(\boldsymbol{R}_n) = (1 - r_2)^{(n-2)}(1 + r_2)^{(n-3)} \left(1 + nr_1^2(r_2 - 1) + (r_1^2 + r_2 - 3r_1^2 r_2)\right) > 0.$$

As $r_2 \in (-1, 1)$, we can solve the condition as follows:

$$\begin{aligned}
& 1 + nr_1^2(r_2 - 1) + (r_1^2 + r_2 - 3r_1^2 r_2) > 0 \\
\iff & -1 - nr_1^2(r_2 - 1) - (r_1^2 + r_2 - 3r_1^2 r_2) < 0 \\
\iff & -1 - r_2 - r_1^2(1 - 3r_2 + nr_2 - n) < 0 \\
\iff & -r_1^2 \left(1 - n + (n - 3)r_2\right) < 1 + r_2 \\
\iff & r_1^2 \left((n - 1) - (n - 3)r_2\right) < 1 + r_2 \\
\iff & r_1^2 < \frac{1 + r_2}{(n - 1) - (n - 3)r_2} \\
\iff & r_1 \in \left(-\sqrt{\frac{1 + r_2}{(n - 1) - (n - 3)r_2}}, \sqrt{\frac{1 + r_2}{(n - 1) - (n - 3)r_2}}\right).
\end{aligned}$$

Appendix C. Impact of the mixed–frequency measurement errors covariance matrix on the prediction accuracy

In this Appendix, we study the sensitivty of the prediction accuracy of the Kalman filter to the values of the variance and (auto–)correlation parameters in $\boldsymbol{H}$. The filtered estimate $a_{t|t}$ obtained by performing the Kalman filter defined in Equation (8) minimizes the Mean Squared Error (MSE). From Lemma 2 in Durbin and Koopman (2012), it follows that its conditional variance $p_{t|t}$ is the lowest among all linear unbiased estimators. We are thus interested in analyzing how $p_{t|t}$ is affected by the covariance matrix of the measurement errors. Therefore, we derive the gradient of $p_{t|t}$ with respect to the covariance matrix of the measurement errors $\boldsymbol{H}$. From Equation (8), it follows that $p_{t|t}$ is given by:

$$p_{t|t} = p_{t|t-1}\left(1 - p_{t|t-1}\boldsymbol{\lambda}^{\intercal}(\boldsymbol{\lambda}p_{t|t-1}\boldsymbol{\lambda}^{\intercal} + \boldsymbol{H})^{-1}\boldsymbol{\lambda}\right). \tag{C.1}$$

We can rewrite it as follows:

$$p_{t|t} = p_{t|t-1}\left(1 - \boldsymbol{\lambda}^{\intercal}(\boldsymbol{\lambda}\boldsymbol{\lambda}^{\intercal} + p_{t|t-1}^{-1}\boldsymbol{H})^{-1}\boldsymbol{\lambda}\right). \tag{C.2}$$

It follows from the Sherman—Morrison formula (see e.g., Bartlett (1951)) that:

$$\begin{aligned}(\boldsymbol{\lambda}\boldsymbol{\lambda}^{\intercal} + p_{t|t-1}^{-1}\boldsymbol{H})^{-1} &= (p_{t|t-1}^{-1}\boldsymbol{H})^{-1} - \frac{(p_{t|t-1}^{-1}\boldsymbol{H})^{-1}\boldsymbol{\lambda}\boldsymbol{\lambda}^{\intercal}(p_{t|t-1}^{-1}\boldsymbol{H})^{-1}}{1 + \boldsymbol{\lambda}^{\intercal}(p_{t|t-1}^{-1}\boldsymbol{H})^{-1}\boldsymbol{\lambda}} \\ &= p_{t|t-1}\left(\boldsymbol{H}^{-1} - \frac{p_{t|t-1}\boldsymbol{H}^{-1}\boldsymbol{\lambda}\boldsymbol{\lambda}^{\intercal}\boldsymbol{H}^{-1}}{1 + p_{t|t-1}\boldsymbol{\lambda}^{\intercal}\boldsymbol{H}^{-1}\boldsymbol{\lambda}}\right).\end{aligned} \tag{C.3}$$

Combining Equation (C.2) and (C.3) leads to:

$$p_{t|t} = p_{t|t-1}\left(1 - p_{t|t-1}\boldsymbol{\lambda}^{\intercal}\left(\boldsymbol{H}^{-1} - \frac{p_{t|t-1}\boldsymbol{H}^{-1}\boldsymbol{\lambda}\boldsymbol{\lambda}^{\intercal}\boldsymbol{H}^{-1}}{1 + p_{t|t-1}\boldsymbol{\lambda}^{\intercal}\boldsymbol{H}^{-1}\boldsymbol{\lambda}}\right)\boldsymbol{\lambda}\right). \tag{C.4}$$

Taking the derivative with respect to the covariance matrix of the measurement errors $\boldsymbol{H}$ gives us:

$$\frac{\partial p_{t|t}}{\partial \boldsymbol{H}} = \boldsymbol{\lambda}^{\intercal}\left(p_{t|t-1}\left(\boldsymbol{H}^{-1} - \frac{p_{t|t-1}\boldsymbol{H}^{-1}\boldsymbol{\lambda}\boldsymbol{\lambda}^{\intercal}\boldsymbol{H}^{-1}}{1 + p_{t|t-1}\boldsymbol{\lambda}^{\intercal}\boldsymbol{H}^{-1}\boldsymbol{\lambda}}\right)\right)^{2}\boldsymbol{\lambda}. \tag{C.5}$$

We plot the gradient in Figure C.7 for $\sigma^2_{\varepsilon_1} = 0.05$, $\sigma^2_{\varepsilon_2} = 0.95$, $r_1 = -0.10$, and $r_2 = 0.20$. These values correspond to the full–sample estimates of the parameters in the empirical application to consumer confidence in Belgium in Section 3. We set $p_{t|t-1}$ and λ equal to one as these scaling parameters do not alter the findings (the estimated value for λ is 0.50), and $n = 32$.

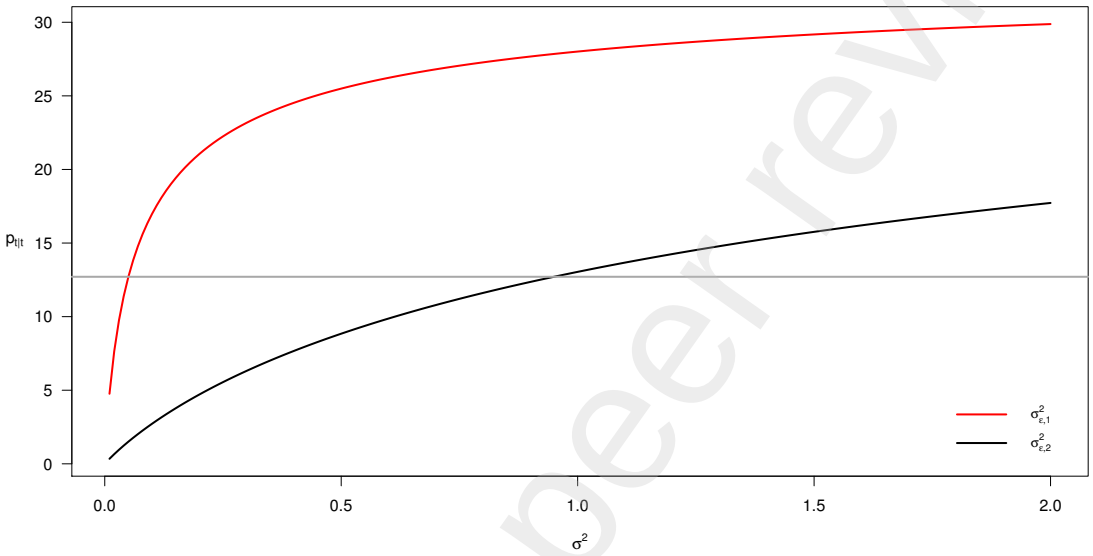
The upper (a) panel in Figure C.7 shows the marginal sensitivity of $p_{t|t}$ ($\times 1000$) on the vertical axis for changes in $\sigma^2_{\varepsilon_1}$ (in red) and $\sigma^2_{\varepsilon_2}$ (in black) along the horizontal axis. In our empirical setting with a relatively low variance for the measurement errors of the low–frequency variable compared to that of the high–frequency variables, we see that the performance of the model is very sensitive to (small) changes in $\sigma^2_{\varepsilon_1}$ from its default value 0.05. However, the marginal sensitivity of $p_{t|t}$ rapidly becomes smaller for changes in larger values of $\sigma^2_{\varepsilon_1}$. In contrast, the variance of the measurement errors of the high–frequency variables is less sensitive around its default value. This indicates the importance of the choice of the informative low–frequency variable, whereas the measurement accuracy of the high–frequency variables seems to be less important, which corresponds well to our empirical setting where we use a low–frequency survey–based indicator and daily economic media news sentiment to estimate latent consumer confidence. However, note that even when $\sigma^2_{\varepsilon_1}$ has a relatively low value, high–frequency variables with small measurement errors still adds value to the performance.

The lower (b) panel in Figure C.7 shows the marginal sensitivity of $p_{t|t}$ ($\times 1000$) on the vertical axis for changes in r_1 (in red) and r_2 (in black) along the horizontal axis. For r_1, we consider the values of (approximately) -0.218 until 0.218 as only these are allowed with $r_2 = 0.20$ and $n = 32$. For r_2, we consider the values of (approximately) -0.218 until 0.99. All these values are allowed with $r_1 = -0.10$. We see that a lower cross–correlation r_1 between the measurement errors of the low–frequency and high–frequency variables improves the model's performance. Intuitively, this means that a higher diversification between the measurement errors (in terms of low and potentially negative correlations) improves the accuracy of the common factor extraction. Note that at the bounds of the allowed values for r_1, i.e., at (approximately) -0.218 and 0.218, $p_{t|t}$ goes to zero. Further,

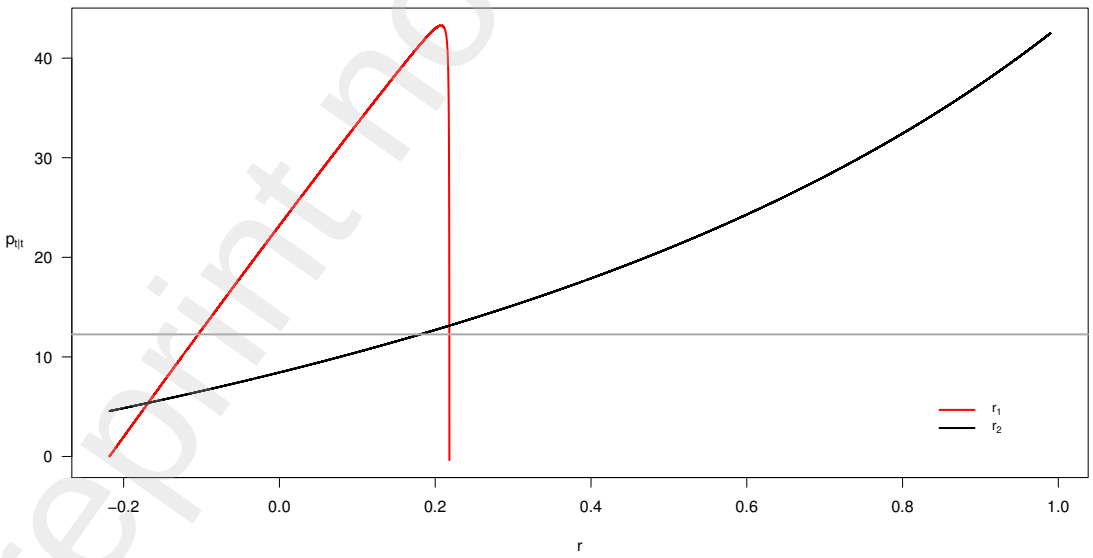
we see that a low autocorrelation r_2 in the measurement errors of the high–frequency variables also leads to a better performance. The intuition is the same as for r_1, the more diversification there is between the errors, the more accurate the Kalman filter prediction will be.

Figure C.7: Impact of the covariance matrix of the measurement errors on $p_{t|t}$.

(a) Marginal sensitivity of $p_{t|t}$ to $\sigma^2_{\varepsilon_1}$ and $\sigma^2_{\varepsilon_2}$.



(b) Marginal sensitivity of $p_{t|t}$ to r_1 and r_2.



Note: The upper (a) panel shows the marginal sensitivity of $p_{t|t}$ ($\times 1000$) to $\sigma^2_{\varepsilon_1}$ (in red) and $\sigma^2_{\varepsilon_2}$ (in black). The lower (b) panel shows the marginal sensitivity of $p_{t|t}$ ($\times 1000$) to r_1 (in red) and r_2 (in black). The default parameter values are $p_{t|t-1} = 1$, $\lambda = 1$, $\sigma^2_{\varepsilon_1} = 0.05$, $\sigma^2_{\varepsilon_2} = 0.95$, $r_1 = -0.10$, and $r_2 = 0.20$, unless indicated otherwise. The horizontal gray line indicates the value of $p_{t|t}$ when the default parameters are used.