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

Abstract

Policymakers, firms, and investors closely monitor traditional survey–based consumer confidence indicators and treat it as an important piece of economic information. We propose 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 framework uses a novel covariance matrix specification. Model estimation and real–time filtering of the latent consumer confidence index are computationally simple. In an empirical application concerning Belgian consumer confidence, we document the economically significant accuracy gains obtained by including daily news sentiment in the dynamic factor model for nowcasting consumer confidence.

Keywords: dynamic factor model, mixed–frequency, nowcasting, sentiment index, sentometrics, state space.

JEL classification: C32, C51, C53, C55.

"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 several days or weeks 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 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 indices and the low–frequency survey–based indicator jointly as a monthly vector driven by a common latent consumer confidence factor. To account for the intra–monthly serial correlation of the measurement errors of high–frequency economic media news sentiment, we provide a non–trivial 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. Furthermore, by imposing a sensible structure on the system matrices, we avoid the curse of dimensionality and allow for a standard Maximum Likelihood estimation and exact filtering via the Kalman filter (see e.g., Durbin and Koopman, 2012). The combined use of survey data and economic media news sentiment leads to a more timely and frequent estimation of the latent state, and imputation of the missing high–frequency observations of the low–frequency observables.

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.[1] 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 the Belgian consumer confidence index. The high–frequency economic media news sentiment variables are computed 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 media news articles, we use a lexicon that we obtain via annotation of relevant articles. We find that the daily average economic media news sentiment is useful for nowcasting survey–based consumer confidence, and for constructing a latent coincident consumer confidence index. The recent COVID–19 pandemic serves as an interesting illustration to show the usefulness of our mixed–frequency model. Our real–

[1] Diebold (2020) writes that "the workhorse nowcasting approaches involve dynamic factor models". An alternative strand of nowcasting models is the family of MIxed DAta Sampling (MIDAS) models, as in Andreou et al. (2013) and Lehrer et al. (2019). The two approaches coexist and have their respective (dis)advantages. The DFM approach is in our setup more suitable given the irregular pattern in missing economic media sentiment observations and the objective to estimate current latent consumer confidence (modelled as a latent factor).

time index correctly indicates a steep drawdown in survey—based consumer confidence.

The remainder of this paper is organized as follows. In Section 2, we introduce our mixed–frequency DFM and show how it can be used to construct a real–time consumer confidence index. The main novelty is a covariance model for the mixed-frequency vector. We investigate the impact of its calibration on the forecasting accuracy in Section 3. In Section 4, we present an empirical application for consumer confidence in Belgium and find that the proposed model implemented using economic media news sentiment is useful for nowcasting survey–based consumer confidence. Section 5 concludes.

2. Constructing a real–time consumer confidence index

In this section, we present our framework for estimating (latent) consumer confidence based on high–frequency economic media news sentiment variables and a low–frequency survey–based proxy of consumer confidence. We first introduce the notation. Next, we present our mixed–frequency DFM and describe the estimation and filtering method. Finally, we discuss some dynamic properties of the model predictions.

2.1. Notation

Our variable of interest is monthly consumer confidence, which we denote by α_t for month $t = 1, 2, \ldots, T$. It represents the average consumer confidence over the month.[2] 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

[2] We take the viewpoint of a public institution that needs to publish a single value for the consumer confidence over a period. As in the high–frequency literature on integrated variance estimation, this reference value over a period can be considered as a normalized integrated quantity (see e.g., Kristensen, 2010). In the application to real-time filtering, we we will be estimating daily nowcasts of the integrated consumer confidence over the month. In case a daily estimate of the "spot" value of consumer confidence is the parameter of interest, we refer the reader to Aruoba et al. (2009).

vector $\boldsymbol{y}_t$ as follows:

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

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 complexity of the model needs to be balanced against the requirement of computational convenience for filtering consumer confidence in real time.

2.2. A mixed-frequency covariance model

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:

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

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

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

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 specification of the correlation model for $\boldsymbol{y}_t$ is our main theoretical contribution. The correlation matrix $\boldsymbol{R}$ is obtained by assuming that the cross–correlations between the measurement errors of the low–frequency variable and the measurement errors of all the high–frequency variables are equal to r_1. For the high–frequency measurement errors, we assume an AR(1) process where 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):

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

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:*

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

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:*

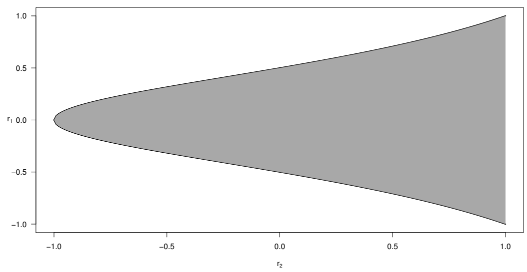
$$r_1 \in \left(-\sqrt{\frac{1 + r_2}{(n-1) - (n-3)r_2}}, \sqrt{\frac{1 + r_2}{(n-1) - (n-3)r_2}} \right).$$

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.

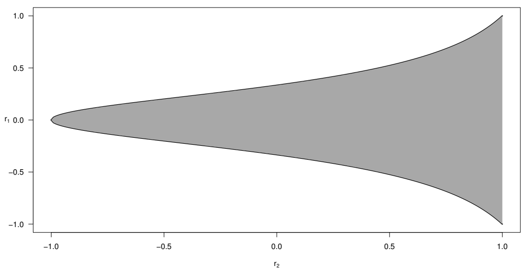
Figure 1 shows an illustration of the upper and lower bound of r_1 given $n = 5,\ 10,\ 30$ and 50. The upper (lower) bound starts at 0 when $r_2 = -1$, and monotonically increases (decreases) non–linearly. Eventually the upper (lower) bound goes to 1 (-1) when $r_2 = 1$. In general, the bounds for r_1 are larger in absolute value for large values of r_2, and small values of n.[3]

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

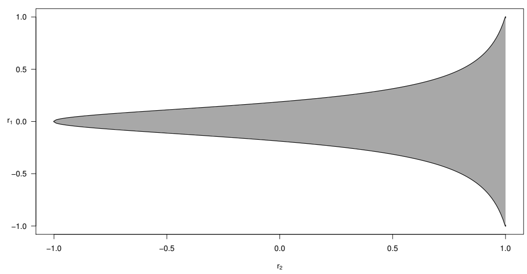
(a) $n = 5$.



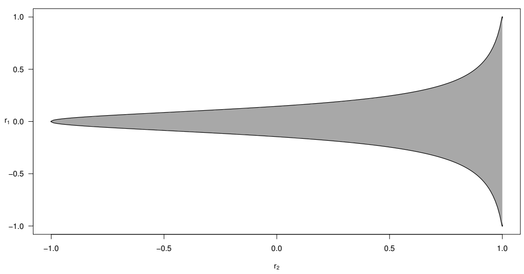
(b) $n = 10$.



(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.

[3]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:

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

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:

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

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 while keeping estimation of the parameters feasible. To avoid the curse of dimensionality, we limit the number of parameters by imposing some structure on the system matrices, i.e., 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 $\boldsymbol{\lambda}$

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

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}.[4] This leads to the following structure for $\boldsymbol{D}$:

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

where $diag\{\cdot\}$ creates a diagonal matrix.

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 $p_{t|t} = \text{Var}[\alpha_t|\boldsymbol{y}_t]$, 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.

[4]The flexibility of our approach allows for extensions and generalizations, e.g., the choices for the factor loadings and the variance of the high–frequency variables can be adapted to account for calendar effects. Moreover, for our empirical application to consumer confidence in Belgium in Section 4, we have checked whether the imposed structure on economic media news sentiment is consistent with the properties of the data by testing for equal averages and variances among all high–frequency variables.

Our approach can be used either to create a latent coincident index in its standard setting, or can be optimized for nowcasting the low–frequency observable by setting the variance of the low–frequency measurement errors ($\sigma^2_{\varepsilon_1}$) and the cross–correlations between the measurement errors of the low– and high–frequency variables (r_1) to zero. In the latter approach, one assumes that the low–frequency variable is observed without any measurement errors. In this paper we will refer to the real–time estimates in the standard setting as the latent coincident index and to the real–time estimates without measurement errors for the low–frequency variable as the real–time nowcasting index.

2.5. Real–time filtering at a daily frequency

Our approach allows for daily updates of the latent factor and the low–frequency observable 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 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.[5] Note that this approach also deals with the time–varying number of days in each month t (i.e., d_t).

3. Impact of mixed-frequency error covariance on prediction accuracy

Our main methodological contribution is the covariance matrix specification $\boldsymbol{H}$ for the mixed-frequency vector y_t. In this section we study the sensitivty of the prediction accuracy of the Kalman filter to the values of the variance and (auto)correlation parameters in

[5] Since we allow for correlations between the measurement errors, 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.

$\boldsymbol{H}$. As the dependency is highly non–linear, we use here two numeric studies to document the sensitivity. The first analysis involves the calculation of the partial derivatives of the conditional variance of a_t, namely $p_{t|t}$. The second analysis studies the sensitivity of the mean squared forecast error in a simulation setup.

3.1. Numeric evaluation of partial deriviatives of $p_{t|t}$

We are first interested in analyzing how $p_{t|t}$ is affected by the covariance matrix of the measurement errors. In Appendix C, we show that the gradient of $p_{t|t}$ with respect to the covariance matrix of the measurement errors $\boldsymbol{H}$ is given by:

$$\frac{\partial p_{t|t}}{\partial \boldsymbol{H}} = \boldsymbol{\lambda}^\intercal \left(p_{t|t-1} \left(\boldsymbol{H}^{-1} - \frac{p_{t|t-1}\boldsymbol{H}^{-1}\boldsymbol{\lambda}\boldsymbol{\lambda}^\intercal \boldsymbol{H}^{-1}}{1 + p_{t|t-1}\boldsymbol{\lambda}^\intercal \boldsymbol{H}^{-1}\boldsymbol{\lambda}} \right) \right)^2 \boldsymbol{\lambda}. \tag{9}$$

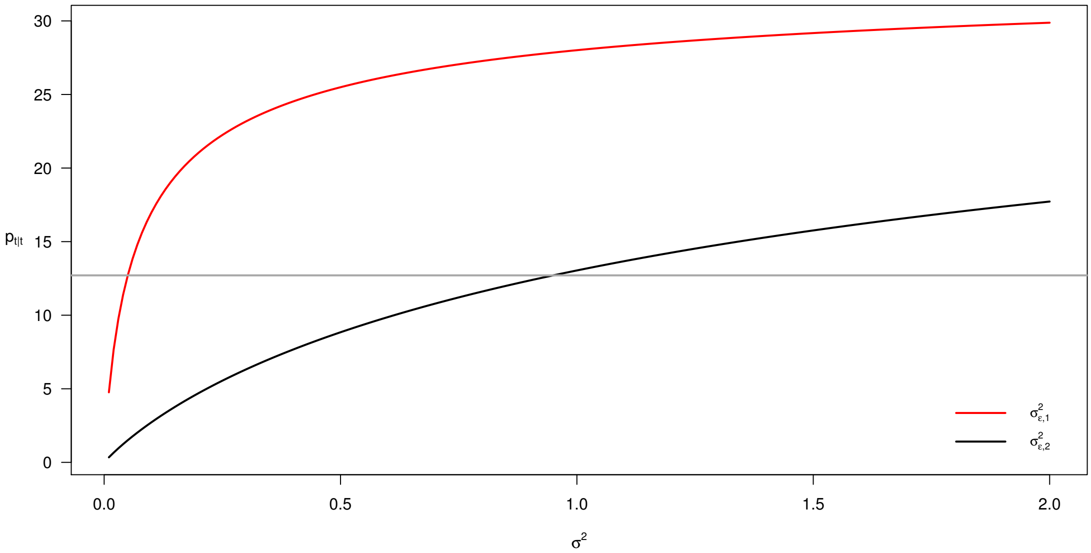
We plot the gradient in Figure 2 for $\sigma^2_{\varepsilon_1} = 0.05$, $\sigma^2_{\varepsilon_2} = 0.95$, $r_1 = -0.10$, and $r_2 = 0.20$ .These values correspond to the full–sample estimates of the parameters in the empirical application to consumer confidence in Belgium in Section 4. In the remainder of the paper, we use these values as default parameters in the illustrations, unless indicated otherwise. 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 0.15), and $n = 32$.

The upper (a) panel in Figure 2 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.

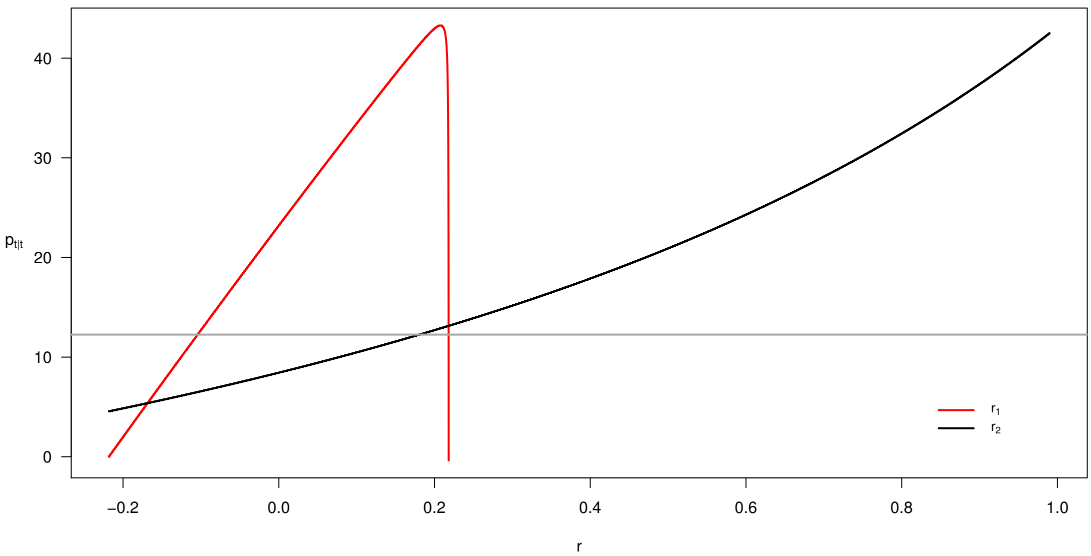
The lower (b) panel in Figure 2 shows the marginal sensitivity of $p_{t|t}$ ($\times 1000$) on the vertical axis for changes in r_1 (in red) and r_2 (in black) along the horizontal axis. For r_1,

Figure 2: 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.

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.

3.2. Impact on RMSE

We now study the impact using a Monte Carlo simulation study calibrated to our empirical setting of estimating consumer confidence using a mixed-frequency vector of monthly consumer surveys and daily news sentiment. Following the state dynamics specified in Equation (5), we first generate a monthly time series of latent consumer confidence α_t. We then generate a monthly survey–based consumer confidence indicator y_t with measurement errors as specified in Equation (2). We also create d high–frequency economic media news sentiment variables for each month t as specified in Equation (2). Each series consists of 250 months with a fixed number of 30 days per month ($d = 30$). We keep 200 months in–sample and simulate 200 series, resulting in 300,000 out–of–sample days in total (50 out–of–sample months times 30 days per month times 200 simulated series). We re–estimate the model at the end of each month t, and provide real–time estimates with the mixed–frequency models at each day i.

Table 1 shows the sensitivity of the RMSE of the the latent coincident index estimate to the variance and correlation parameters. As a benchmark, we also include the RMSE of the following AR(1) model which only uses the low–frequency survey–based consumer confidence observations to obtain one–step ahead forecasts of α_t ($a_{t|t-1}$):

$$y_t = \alpha_t + \varepsilon_t, \quad \text{with} \quad \varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right), \tag{10}$$

where the the measurement errors ε_t are normally distributed with mean zero and variance σ_ε^2 and the state dynamics of latent consumer confidence are given by Equation (5).

Panel A of Table 1 shows the impact of increasing $\sigma_{\varepsilon_1}^2$ on the RMSE of the low–frequency model and of the mixed–frequency model. We see that the performance of the low–frequency model deteriorates rapidly, while the performance of the latent coincident index remains quite stable due to the high–frequency information.

Panesl B, C and D are specific to the mixed-frequency model as they study the impact of increasingly larger parameter values for $\sigma_{\varepsilon_2}^2$, r_1 and r_2 on the RMSE of the latent coincident index. As these parameters have no effect on the performance of the low–frequency model, and the value of $\sigma_{\varepsilon_1}^2$ is fixed at 0.05, the RMSE of 0.2517 for the low–frequency model can serve as a benchmark. As expected, a larger $\sigma_{\varepsilon_2}^2$ leads to a larger RMSE, but even when $\sigma_{\varepsilon_2}^2$ is almost thirty times as large as $\sigma_{\varepsilon_1}^2$, the latent coincident index still outperforms the low–frequency model. We also see that even though $\sigma_{\varepsilon_1}^2$ has a relatively low value, high–frequency variables with small measurement errors still add substantial value to the performance.

We see that the lower the value of r_1 is, the more accurate the latent coincident index becomes in estimating the latent state compared to the low–frequency model since a higher diversification between the measurement errors (in terms of low and potentially negative correlations) improves the accuracy of the common factor extraction. However, note that when $r_1 = 0.20$, the RMSE is, on average, lower as the correlation matrix $\boldsymbol{R}$ is near its bound for positive definiteness. Lastly, the RMSE values indicate that for lower values of r_2, the latent coincident index performs better. This is intuitive as the more diversification there is between the errors, the more accurate the Kalman filter prediction becomes.

Bottomline, while the marginal sensitivities of $p_{t|t}$ to the variance and the autocorrelation are of the same order of magnitude, the total impact also depends on the magnitude

Table 1: Effect of the covariance matrix of the measurement errors on the estimation accuracy of the latent coincident index.

$\sigma^2_{\varepsilon_1}$	0.05	0.10	0.20	0.50	0.80
low–frequency	0.2517	0.2607	0.2853	0.3509	0.3945
mixed–frequency	0.1949	0.1963	0.1999	0.2066	0.2116
$\sigma^2_{\varepsilon_2}$	0.50	0.75	0.95	1.15	1.40
mixed–frequency	0.1446	0.1774	0.1949	0.2068	0.2180
r_1	-0.20	-0.10	0	0.10	0.20
mixed–frequency	0.1938	0.1949	0.1964	0.1968	0.1955
r_2	0	0.10	0.20	0.30	0.40
mixed–frequency	0.1812	0.1879	0.1949	0.2011	0.2084

Note: This table shows the RMSE of the low–frequency model and the latent coincident index with an increasingly larger parameter value for $\sigma^2_{\varepsilon_1}$, and the RMSE of the mixed–frequency model with increasingly larger parameter values for $\sigma^2_{\varepsilon_2}$, r_1 and r_2, respectively. The default parameter values are $\rho = 0.85$, $\sigma^2_\eta = 0.25$, $\lambda = 1$, $\sigma^2_{\varepsilon_1} = 0.05$, $\sigma^2_{\varepsilon_2} = 0.95$, $r_1 = -0.10$, and $r_2 = 0.20$, unless indicated otherwise.

of the variation of these parameters, which is limited for the autocorrelation parameters due to the positive definiteness constraint. As a result, we find that the autocorrelation has only a minor effect on the RMSE, while the variance of the measurement errors of the economic media news sentiment variables seems to have the largest impact. This emphasizes the importance of modelling economic media news sentiment and also suggests that for prediction purposes one should not be overly concerned about misspecification of the autocorrelation structure.

4. 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 the monthly survey–based consumer confidence indicator of the National Bank of Belgium (NBB) which is currently the most prominent proxy of latent consumer confidence in Belgium. Next, we discuss 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). Then, we evaluate the real–time estimates of both the latent coincident index and the real–time nowcasting index in an out–of–sample evaluation. Finally, the recent COVID–19 pandemic serves as an interesting illustration to show the usefulness of our mixed–frequency model.

4.1. Survey–based consumer confidence indicator

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:

$$Balance_t = (PP_t + 0.5P_t) - (MM_t + 0.5M_t)\,. \qquad (11)$$

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.

4.2. Economic media news sentiment

The use of economic media news sentiment as a proxy for latent 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.[6] 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).[7] 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

[6] 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.

[7] 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`.

last filter to reduce the number of false positives.[8] 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 the polarity of a word j_a. The sentiment per media news article s is then computed as:

$$s = \frac{1}{J_a} \sum_{j_a=1}^{J_a} v_{j_a} w_{j_a}. \quad (12)$$

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. 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.[9] Figure 3 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., 2020).[10]

To create the daily economic media news sentiment variables $m_{t,i}$, we aggregate the

[8] 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.

[9] 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.

[10] 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 (12), 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., $(-1 \times 1)/1$.

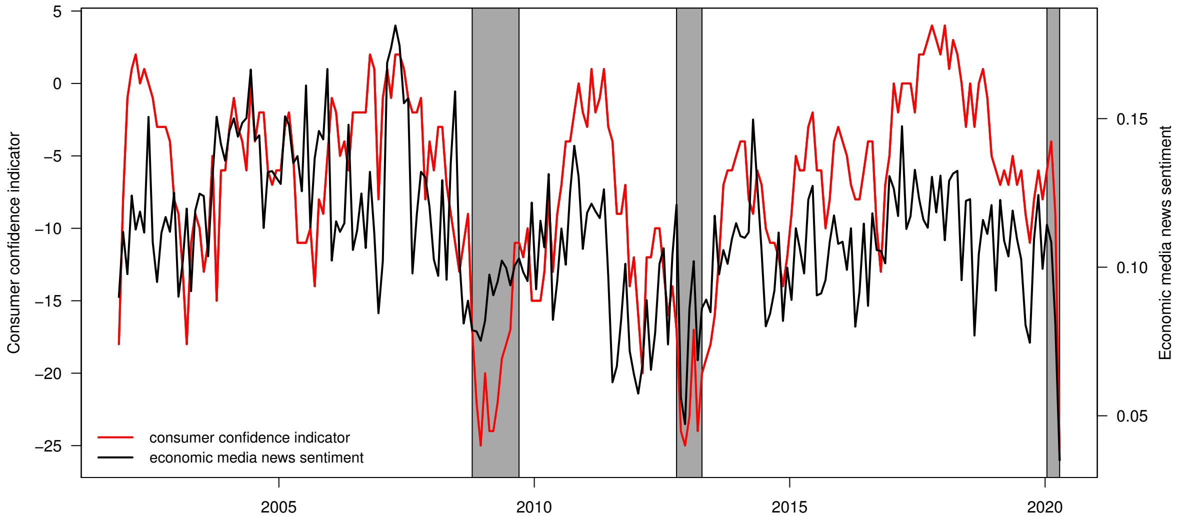
Figure 3: The most frequent positive and negative words (translated in English) in the selected media news articles over the period November 2001 until April 2020.

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

resulting sentiment values by taking the daily average per newspaper, and then by averaging over all 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 the high–frequency economic media news sentiment variables from the pseudo–months to the corresponding monthly survey–based consumer confidence indicator.

Figure 4 shows the monthly average economic media news sentiment and the monthly consumer confidence indicator. We see that there is a large degree of comovement between both time series with a contemporaneous correlation of 0.54. Note that both the consumer confidence indicator and economic media news sentiment experience their largest drawdown, and are at their lowest value, in April 2020 during the COVID–19 pandemic.

Figure 4: Monthly economic media news sentiment and survey–based consumer confidence indicator over the period November 2001 until April 2020.



Note: The red line indicates the monthly survey–based consumer confidence indicator, 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).

4.3. Out–of–sample evaluation

We evaluate the real–time estimates of both the real–time nowcasting index and the latent coincident index. First, we show how to construct the real–time consumer confidence indicators. Then, we assess the added value of the high–frequency economic media news sentiment in the real–time estimates of latent consumer confidence. Finally, we compare the nowcasting accuracy of observed survey–based consumer confidence by the mixed–frequency models and compare it with the performance of the one–step ahead forecasts of the low–frequency model.

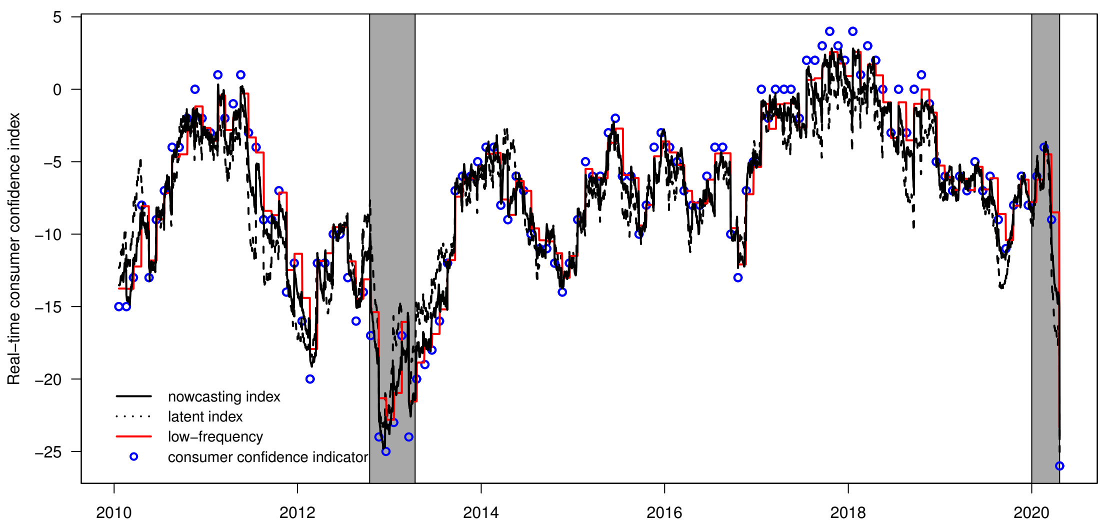
4.3.1. Construction of the latent coincident index and real–time nowcasting index

As a benchmark model, we use the low–frequency model, defined in Equation (10), to obtain one–step ahead forecasts of α_t and y_t (both forecasts are denoted by $a_{t|t-1}$). To obtain real–time filtered estimates of α_t and y_t (both estimates are denoted by $a_{t|t}$, as the factor loading for the low–frequency variable is set equal to one), we use the mixed–frequency model, both in its standard setting, i.e., the latent coincident index, and with the variance of the low–frequency measurement errors ($\sigma^2_{\varepsilon_1}$) and the cross–correlations between the measurement errors of the low– and high–frequency variables (r_1) set to zero, i.e., the real–time nowcasting index.

The first sample used to estimate the models consists of 99 observations from November 2001 until December 2009. The corresponding out–of–sample evaluation sample consists of 123 observations for the period of January 2010 until April 2020. By dividing the data up like this, we respect the findings of Hansen and Timmermann (2012) that the forecast evaluation period should be a relatively large proportion of the available data, while preserving enough data to keep estimation of all the models feasible. We re–estimate the models each month at the time that we obtain a new observation of the survey–based consumer confidence indicator (y_t) using an expanding estimation window, and provide real–time estimates at each day i for each out–of–sample month $t+1$.

Figure 5 shows the daily real–time estimates of the mixed–frequency models, the one–step ahead forecasts of the low–frequency model and the survey–based consumer confidence indicator as measured by the National Bank of Belgium. We see that there is

Figure 5: Daily real–time estimates of the mixed–frequency models, one–step ahead forecasts of the low–frequency model, and the monthly survey–based consumer confidence indicator as measured by the National Bank of Belgium over the period January 2010 until April 2020.



Note: The (dotted) black line(s) are the real–time estimates of the mixed–frequency models, the red line represents the one–step ahead forecasts of the low–frequency model, and the blue dots indicate the survey–based consumer confidence indicator observations. The shaded area indicates a recession period defined as two consecutive quarters of negative economic growth as measured by Belgian Gross Domestic Product (GDP).

substantial intra–monthly movement in the mixed–frequency estimates, and that the latent coincidence index produces more volatile estimates than the real–time nowcasting index. This is the result of the restriction of setting the variance of the measurement errors of the survey–based consumer confidence indicator equal to zero, and we therefore also expect that the real–time nowcasting index is more suitable for nowcasting y_t. Finally, note that the forecasts of the low–frequency model are constant during an entire month t which results in a stepwise pattern.

4.3.2. Added value of high–frequency sentiment in estimating latent consumer confidence

We assess the added value of incorporating the high–frequency sentiment in estimating latent consumer confidence by comparing the real–time estimates of the conditional state variance $p_{t|t}$ obtained under the mixed–frequency models with the one–step ahead forecasts of the state variance $p_{t|t-1}$ obtained under the low–frequency model. We therefore compute the Variance Reduction Ratio (VRR) which compares the average $p_{t|t}$ with the average $p_{t|t-1}$. We define the VRR_h at a daily forecasting horizon h as:

$$\text{VRR}_h = \frac{\frac{1}{T}\sum_{t=1}^{T} p_{t|t,h}}{\frac{1}{T}\sum_{t=1}^{T} p_{t|t-1}}, \tag{13}$$

where h is equal to the number of days before the end of the pseudo–month so that the 14th of each month corresponds to $h = 0$, $p_{t|t,h}$ are the real–time estimates of the conditional state variance computed at forecasting horizon h, and T is the total number of out–of–sample months.

In Table 2, we show the VRR for $h = 0, 1, 2, ..., 13$, and the overall VRR which is computed by averaging over all the forecasting horizons (not only until $h = 13$). The VRR of the latent coincident index ranges in between 69.89% and 81.14% which is substantially lower than the VRR of the real–time nowcasting index that ranges in between 89.70% and 96.57%. As expected, this suggests that the restriction of setting the variance of the low–frequency measurement errors ($\sigma^2_{\varepsilon_1}$) and the cross–correlations between the measurement errors of the low– and high–frequency variables (r_1) to zero does not allow this model to fully exploit the high–frequency information. However, both mixed–frequency approaches

Table 2: The VRR of the mixed–frequency models over the period January 2010 until April 2020.

h	VRR (%)	
	Nowcasting index	Latent index
0	89.70	69.89
1	90.20	70.64
2	90.70	71.42
3	91.22	72.22
4	91.70	72.96
5	92.22	73.80
6	92.75	74.66
7	93.28	75.54
8	93.82	76.44
9	94.38	77.37
10	94.92	78.30
11	95.49	79.29
12	96.01	80.19
13	96.57	81.14
Overall	94.64	79.52

show that adding high–frequency sentiment allows for a better latent factor extraction. Finally, note that as h becomes smaller the latent coincident index gets, on average, rapidly less volatile.

4.3.3. Nowcasting accuracy

We evaluate the accuracy gains of estimating the low–frequency survey–based indicator y_t in terms of the Relative RMSE to compare the one–step ahead forecasts of the low–frequency model with the real–time estimates of the mixed–frequency models. More formally, we define the Relative RMSE$_h$ at a daily forecasting horizon h as:

$$\text{Relative RMSE}_h = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(a_{t|t,h} - y_t\right)^2}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(a_{t|t-1} - y_t\right)^2}}, \tag{14}$$

where $a_{t|t,h}$ are the real–time estimates of the mixed–frequency models computed at forecasting horizon h, and $a_{t|t-1}$ are the corresponding one–step ahead forecasts of the low–frequency model. To test whether the difference is statistically significant, we perform a pairwise Diebold–Mariano test on the squared errors with a Null hypothesis of equal, or worse, performance with the low–frequency model (Diebold and Mariano, 2002).

We also compute the Mean Directional Accuracy (MDA) of the mixed–frequency (and low–frequency) models to examine whether the estimates correctly indicate in which direction the survey–based consumer confidence indicator is moving. More formally, we define the MDA_h at a daily forecasting horizon h as:

$$\text{MDA}_h = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\left((a_{t|t,h} - y_{t-1})(y_t - y_{t-1}) > 0\right), \tag{15}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. To test whether the difference is statistically significant, we perform a pairwise χ^2–test with a Null hypothesis of equal, or worse, performance with the low–frequency model.

In Table 3, we show the Relative RMSE and MDA for $h = 0, 1, 2, ..., 13$, and also the overall Relative RMSE and MDA which is computed by averaging over all the forecasting horizons (not only until $h = 13$). We see that overall the real–time nowcasting index performs significantly better than the low–frequency model in terms of both RMSE and MDA. At a forecasting horizon h of zero until twelve days, the outperformance compared to the low–frequency model is statistically significant at a 5% significance with a Relative RMSE which is in between 83.83% and 88.64%. When $h = 13$, the outperformance is statistically significant at a 10% significance level with a Relative RMSE of 90.09%. We see that the latent coincident index does not perform substantially better or worse than the low–frequency model in terms of Relative RMSE. This result is not surprising as the restriction of setting the variance of the low–frequency measurement errors ($\sigma^2_{\varepsilon_1}$) and the cross–correlations between the measurement errors of the low– and high–frequency variables (r_1) to zero makes the mixed–frequency model more suitable for nowcasting the low–frequency variable. Finally, note that as h becomes smaller the nowcasts get, on average, rapidly more accurate.

Table 3: Relative RMSE and MDA of the mixed–frequency models over the period January 2010 until April 2020.

h	Relative RMSE (%)		MDA (%)	
	Nowcasting index	Latent index	Nowcasting index	Latent index
0	83.83**	101.16	74.59***	70.49**
1	84.00**	100.56	75.41***	69.67**
2	83.38**	99.44	73.77***	70.49**
3	84.15**	99.88	74.59***	71.31**
4	84.15**	99.48	73.77***	72.13**
5	84.93**	99.85	71.31**	69.67**
6	85.05**	99.02	72.95***	68.85**
7	85.15**	98.62	73.77***	69.67**
8	85.23**	98.14	70.49***	70.49**
9	86.66**	99.70	70.49***	71.31**
10	87.18**	99.81	70.49***	69.67**
11	87.41**	99.61	72.13**	68.03*
12	88.64**	100.42	71.31**	68.03*
13	90.09*	102.51	70.49**	67.21*
Overall	89.10**	100.45	70.53***	68.18***

Note: The RMSE of the low–frequency model is 3.21 and its sign accuracy is 57.38%. We perform a pairwise Diebold–Mariano test on the squared errors with a Null hypothesis of equal, or worse, performance with the low–frequency model in terms of RMSE. To account for the autocorrelation, we use Heteroskedasticity and Autocorrelation Consistent (HAC) standard errors. We further perform a pairwise χ^2–test with a Null hypothesis of equal, or worse, performance with the low–frequency model in terms of MDA. The significance at the 10%, 5%, and 1% levels are denoted as *, **, and ***, respectively.

The MDA of the real–time nowcasting index ranges from 70.49% to 75.41% which confirms its good performance in terms of the Relative RMSE. The outperformance is also statistically significant at $h = 1, 2, \ldots, 11$ at a 1% significance level, except for $h = 5$. While the latent coincident index does not outperform in terms of Relative RMSE, it does in terms of MDA which ranges in between 69.67% to 72.13% and is statistically significant at a 5% level for $h = 1, 2, \ldots, 10$.

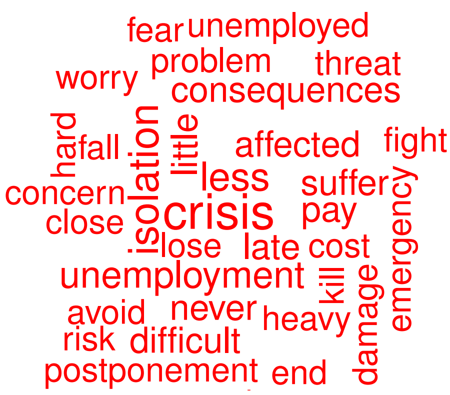
4.4. COVID–19 pandemic

In general, our model is most useful in crisis periods when economic indicators can be subject to sudden and rapid changes. As the augmentation of a low–frequency proxy with high–frequency sentiment information allows us to capture these changes more timely. In this regard, the recent COVID–19 pandemic serves as an interesting illustration to demonstrate the applicability of our mixed–frequency model.

To zoom in on the COVID–19 crisis, we consider the last three observations on the survey–based consumer confidence indicator 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 2020, 90% of the selected media news articles contain at least one word related to the COVID–19 pandemic, i.e., coronavirus. Figure 6 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

Figure 6: The most frequent negative words (translated in English) in the selected media news articles over the period February 19 2020 until April 21 2020.



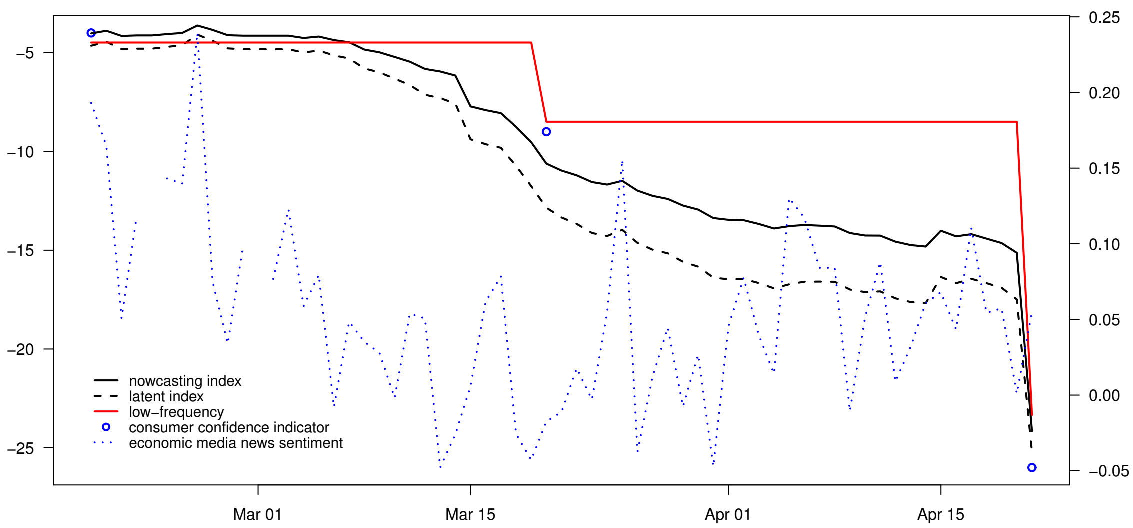
which the government decided that schools, restaurants and bars would need to shut down from 13 March onwards. On 17 March, the Belgian government decided on a so–called "lockdown light" from 18 march onwards. Some important events thus happened after, or at the end of, the survey period for the consumer 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[11]

> "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 7, where the blue dots indicate the monthly reported values from the survey–based consumer confidence indicator. The (dotted) black line(s) are the daily real–time estimates of the mixed–frequency models, the red line represents the one–step ahead forecasts of the low–frequency model, and the dotted blue line are the daily economic media news sentiment observations. We see that during the first half of March, the mixed–frequency models correctly assess that consumer confidence is going down from around 11 March onwards. However, the moment that the consumer confidence indicator for March is published, the mixed–frequency models indicate that on the date of the press release consumer confidence is already down again by two (real–time nowcasting index) to five (latent coincident index) points which indicate that the survey–based consumer confidence indicator is not an accurate estimate of consumer confidence at that date. Finally, note that, from 11 March until the beginning of April, daily economic media news sentiment observations are often approximately two standard deviations below their long–term average.

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

Figure 7: Daily one–step ahead forecasts of the low–frequency model, real–time estimates of the mixed–frequency models, the survey–based consumer confidence indicator, and economic media news sentiment over the period 19 February 2020 until 21 April 2020.



Note: The (dotted) black line(s) are the real–time estimates of the mixed–frequency models, the red line represents the one–step ahead forecasts of the low–frequency model, the blue dots indicate the survey–based consumer confidence indicator observations, and the dotted blue line are the economic media news sentiment observations (right hand side).

5. Conclusion

Policymakers, firms, and investors closely monitor traditional survey–based consumer confidence indicators and treat it as an important piece of economic information. To obtain early estimates of consumer confidence in real time, we augment the low–frequency survey–based consumer confidence indicator with the high–frequency sentiment embedded in economic media news articles. We take the viewpoint of a public institution that needs to publish a single value for the consumer confidence over a period. In this regard, we propose a novel covariance specification for the mixed–frequency vector consisting of a monthly survey–based indicator and daily economic media news sentiment values. The covariance specification is then used in a Dynamic Factor Model with a state space representation where the mixed-frequency vector is driven by a common latent consumer confidence factor. We apply the proposed framework to nocasting consumer confidence over the period November 2001 until April 2020. We find that adding daily news sentiment to the proposed dynamic factor model leads to a nowcasting accuracy gain of over ten percent.

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.

Andreou, E., Ghysels, E., Kourtellos, A., 2013. Should macroeconomic forecasters use daily financial data and how? Journal of Business & Economic Statistics 31, 240–251.

Ardia, D., Bluteau, K., Borms, S., Boudt, K., 2020. The R package **sentometrics** to compute, aggregate and predict with textual sentiment. Journal of Statistical Software, forthcoming.

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.

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.

Diebold, F., 2020. Real-Time Real Economic Activity: Exiting the Great Recession and Entering the Pandemic Recession. Working paper.

Diebold, F.X., Mariano, R.S., 2002. Comparing predictive accuracy. Journal of Business & Economic Statistics 20, 134–144.

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

Hansen, P., Timmermann, A., 2012. Choice of sample split in out–of–sample forecast evaluation. Working paper. European University Institute.

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.

Kristensen, D., 2010. Nonparametric filtering of the realized spot volatility: A kernel-based approach. Econometric Theory 26, 60–93.

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.

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:

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

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 suppose that:

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

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 last column and row. Subtracting r_2 times the second–to–last row from the last row of $\boldsymbol{R}_{k+1}$ yields:

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

Expanding this determinant along the last 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:

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

Now remark that $\boldsymbol{T}$ without the first row and the last column 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–to–last column from the last column before taking the determinant by expanding along the last 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:

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

and combining both terms, yields that:

$$\begin{aligned}\det(\boldsymbol{R}_{k+1}) &= (1-r_2)^{k-1}(1+r_2)^{k-2}\left(r_1^2(r_2-1)\right)\\ &+ (1-r_2)^{k-1}(1+r_2)^{k-2}\left(1+kr_1^2(r_2-1)+(r_1^2+r_2-3r_1^2r_2)\right)\\ &= (1-r_2)^{k-1}(1+r_2)^{k-2}\left[r_1^2(r_2-1)+1+kr_1^2(r_2-1)+(r_1^2+r_2-3r_1^2r_2)\right]\\ &= (1-r_2)^{k-1}(1+r_2)^{k-2}\left[1+(k+1)r_1^2(r_2-1)+(r_1^2+r_2-3r_1^2r_2)\right],\end{aligned}$$

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:

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

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:

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

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. Derivation of Equation (9)

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 Equation (9).