Sparse plus dense MIDAS regressions and nowcasting during the COVID pandemic

Jad Beyhum
Department of Economics, KU Leuven, Belgium

Jonas Striaukas
Department of Finance, Copenhagen Business School, Denmark.

December 3, 2023

Abstract

The common practice for GDP nowcasting in a data-rich environment is to employ either sparse regression using LASSO-type regularization or a dense approach based on factor models or ridge regression, which differ in the way they extract information from high-dimensional datasets. This paper aims to investigate whether sparse plus dense mixed frequency regression methods can improve the nowcasts of the US GDP growth. We propose two novel MIDAS regressions and show that these novel sparse plus dense methods greatly improve the accuracy of nowcasts during the COVID pandemic compared to either only sparse or only dense approaches. Using monthly macro and weekly financial series, we further show that the improvement is particularly sharp when the dense component is restricted to be macro, while the sparse signal stems from both macro and financial series.

Keywords: factor models, high-dimensional data, mixed-frequency data, nowcasting, sparse plus dense

1 Introduction

Nowcasting is particularly challenging during large economic *nonstandard* shocks such as the COVID pandemic. Standard nowcasting methods may fail because they are not well-equipped to model the new dynamics prevailing during these shocks. To properly model such shocks, a good practice is to rely on many variables for prediction since the shock may be less nonstandard in a larger information set. Increasing data, however, adds to the curse of dimensionality problem now well-known in nowcasting (Babii et al., 2022). Traditional forecasting methods in high-dimensional settings can be broadly categorized into two classes (Giannone et al., 2021). On the one hand, sparse approaches such as the LASSO (Tibshirani, 1996; Bickel et al., 2009) rely on the assumption that only very few covariates matter for predicting certain outcome variables. Such approaches leverage information in idiosyncratic shocks to forecast the variable of interest. On the other hand, dense methods form predictions based on common shocks. These techniques include the ridge estimator (Hsu et al., 2012) and factor-augmented regression (Stock and Watson, 2002; Brownlees et al., 2023). The two types of methods differ in how the information is extracted from high-dimensional datasets. Recently, a new class of sparse plus dense techniques has emerged (Chernozhukov et al., 2017; Fan et al., 2023a,b), which nest the two original classes and use both idiosyncratic and common shocks to form predictions. In this paper, we adapt sparse plus dense approaches to the specific challenges of nowcasting and show that they can dramatically improve nowcasts of the US GDP during the COVID pandemic.

The aforementioned dimension reduction techniques – sparse, dense, or sparse plus dense – need to be tailored to effectively handle mixed-frequency data. To address the mixed frequency data problem, two commonly employed approaches utilize either factor models or regression-based methods. The first approach applies dynamic factor models, or in short DFMs (Giannone et al., 2008). In DFMs, latent factors are extracted from a large panel of variables, which are then used to predict low-frequency series such as GDP growth. The DFM is formulated in a state-space form, and the low-frequency series are projected onto the latent factors to generate predictions. For a comprehensive review of the literature, we refer to Bok et al. (2018); Doz and Fuleky (2020); Ruiz et al. (2022). Another

strategy lies in mixed-frequency data sampling regression models, also known as MIDAS, see Marcellino and Schumacher (2010), Andreou et al. (2010, 2013); Babii et al. (2022). Unlike DFMs, MIDAS regressions do not model low-frequency series as latent processes. Instead, higher-frequency data is directly projected onto low-frequency series. To overcome the issue of parameter proliferation, MIDAS models employ a weighting scheme based on a low-dimensional parameter structure. It is worth noting that under certain data-generating processes, both DFM and MIDAS approaches are equivalent, see Bai et al. (2013) for detailed discussion.

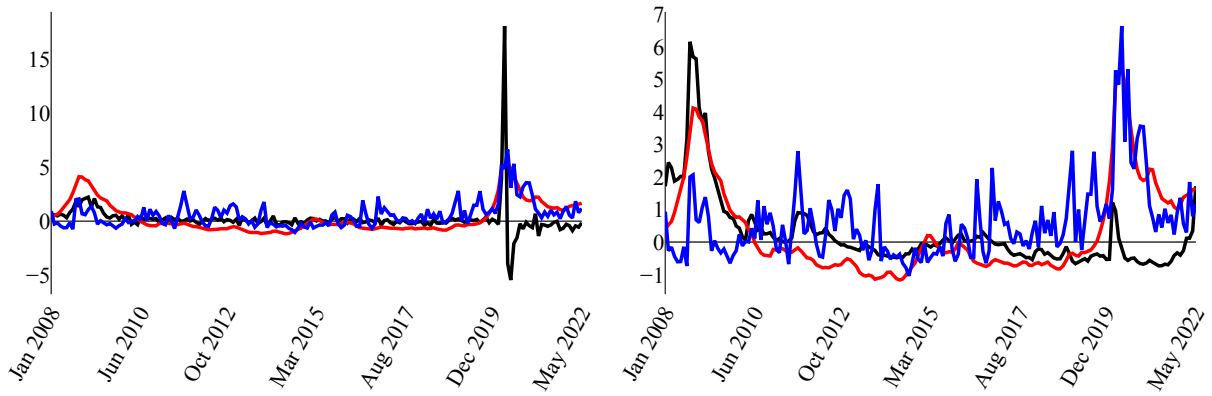
In this paper, we present a novel class of high-dimensional MIDAS regression models that depart from the standard assumptions of coefficients being either sparse or dense. Instead, our proposed models allow for sparse plus dense patterns, introducing a methodological innovation to the nowcasting literature. Traditionally, nowcasting approaches have leaned towards either dense methods (such as DFM) or sparse techniques (like LASSO-type estimators and MIDAS). In contrast, our approach embraces the flexibility of sparse plus dense patterns. The methods we propose enjoy the interpretability of high-dimensional regression and factor regression approaches. They are also MIDAS adaptations of estimators for which statistical guarantees have been derived in the literature, see, for instance, Chernozhukov et al. (2017), Fan et al. (2023a), Fan et al. (2023b).

The empirical contribution of our paper to the nowcasting literature predominantly centers on the COVID pandemic period, well-acknowledged as an exceptionally challenging time for generating accurate nowcasts (Ferrara and Sheng, 2022; Foroni et al., 2022). The best subset of novel models within our consideration shows significantly better performance during the COVID period than sparse-only or dense-only methods. Their performance remains competitive with state-of-the-art approaches in the pre-COVID period. Our findings underscore the effectiveness of methods that incorporate a sparse plus dense structure on the parameter estimates.

These results provide useful insights on the nature of the COVID shock. It is driven by both idiosyncratic shocks (hence, the importance of the sparse component) and common shocks (hence, the value of the dense component). Moreover, we find that the nowcasts are the most accurate when the method uses a sparse plus dense pattern for macroeconomic

predictors but only a sparse structure for financial regressors. This suggests that COVID included a large common shock on the macroeconomy (along with some idiosyncratic shocks on specific macro variables) but mostly affected the financial system through idiosyncratic shocks.

Uncertainty measures and COVID



(a) Macro factor.

(b) Financial factor.

Figure 1: The figure illustrates the standardized macro factor (the black line in Figure 1a) alongside the financial factor (the black line in Figure 1b). Additionally, the same two measures of economic uncertainty are presented in each Subfigure: the red line corresponds to the macroeconomic uncertainty, as defined by Jurado et al. (2015), while the blue line signifies the economic policy uncertainty, following the framework proposed by Baker et al. (2016)

. Note that the scales of the two Subfigures differ.

As a further illustration of the nature of the COVID shock, we plot two factors representing the first principal components of the macro (1a) and financial (1b) datasets used in our nowcasting models (see Figure 1). Additionally, we include measures of economic uncertainty from Jurado et al. (2015) and Baker et al. (2016). In accordance with our predictive findings, we see that the macro factor exhibits a significantly larger variation during the COVID pandemic compared to the financial crisis of 2008. Conversely, the financial factor displays a greater variation during the 2008 financial crisis than during the COVID

pandemic. Furthermore, both measures of economic uncertainty demonstrate a more pronounced variation during the COVID period in comparison to the financial crisis. The macro factor therefore seems to measure better economic uncertainty during the pandemic than the financial one.

These observations emphasize the dynamic nature of economic factors and uncertainties across different crisis contexts, providing valuable insights into the unique challenges and characteristics associated with each period. Consequently, it becomes imperative to approach macro and financial data with a nuanced perspective and design models that are well-equipped to effectively capture and respond to diverse economic shocks.

The paper is structured as follows. In Section 2, we present our novel dense, sparse, and sparse plus dense nowcasting methods. The data utilized for generating our results is outlined in Section 3. Our primary empirical findings, along with a discussion and an extension of the main analysis to alternative methods and data choices, are reported in Section 4. The paper concludes with a summary and discussion in Section 5.

2 Sparse plus dense models

We begin by presenting a general model that serves as the foundation for our methodological development. We explain how mixed-frequency data can be handled within this model in subsequent sections. Throughout the paper, we maintain a generic notation to ensure clarity and coherence. For an integer $L \in \mathbb{N}$, we let $[L] = \{1, \dots, L\}$ and $[L]_+ = \{0, \dots, L\}$.

2.1 The nowcasting challenge

Let $\{Y_t,\ t \in [T]\}$ be the low-frequency US GDP we want to predict. It is measured by quarter. Nowcasting relies on high-frequency covariates $\{\widetilde{X}^n_{k,t-(j-1)/m_n}, \in k \in [K_n], j \in [m_n], t \in [T], n \in [N]\}$ which are part of N panels of different frequencies. The number of covariates in panel n is K_n while that of dates is T. The predictors may be measured at some higher frequency with $m \geq 1$ observations for every t. Throughout the paper, we use a "$\sim$" to denote high-frequency variables. In our nowcasting application, we analyze two frequency panels: monthly and weekly, thus $N = 2$ in our case. The monthly panel

comprises $K_1 = 76$ macroeconomic variables. Regarding the weekly panel, we utilize $K_2 = 36$ financial predictor variables.

Nowcasting is characterized by two statistical challenges that researchers need to address. The first challenge involves effectively handling high-dimensional predictors. With a vast array of available high-frequency economic variables, it is crucial to carefully select or summarize the pertinent information to ensure accurate predictions. The process of variable selection plays a pivotal role in this regard. The second challenge arises from the mixed-frequency nature of the data. Simply regressing the low-frequency variable on numerous high-frequency lags can lead to an ultra-high-dimensional problem and parameter proliferation can bite even if estimators such as LASSO are used. This issue makes the model overly complex and potentially compromises its reliability.

Specific dimension reduction techniques are necessary to address both issues simultaneously. A popular dimension reduction approach for mixed-frequency data is the so-called MIDAS regression. Linear-in-parameters MIDAS specification is natural in the context of big data since it simplifies the computation (Babii et al., 2022), yet retains the flexibility of nonlinear weighting schemes. There are several ways to handle the high dimensionality of the predictors. Sparse methods are valid when only a few covariates matter for prediction. Dense methods instead rely on the assumption that most predictors are important and that their effect is in some sense uniform. In this paper, we put forward sparse plus dense MIDAS regression models, which prove to be useful in nowcasting during the COVID period.

2.2 General regression model

In our general regression model, we also incorporate real-valued factors $\{\widetilde{f}^n_{t-(j-1)/m,r}, n \in [N], t \in [T], j \in [m], r \in [R_n]\}$. The number of factors in panel n is R_n. We study the following mixed-frequency model:

$$\begin{aligned} Y_t = \rho_0 + \sum_{j=1}^{\mathrm{J}} \rho_j Y_{t-j} + \sum_{n=1}^{N}\sum_{k=1}^{K_n}\sum_{j=1-\ell_n}^{m_n\times q_n} \omega\left(\frac{(j-1+\ell_n)}{m_n}, \gamma^n_k\right) \widetilde{X}^n_{k,t-1-(j-1)/m_n} \\ + \sum_{n=1}^{N}\sum_{k=1}^{R_n}\sum_{j=1-\ell_n}^{m_n\times q_n} \omega\left(\frac{(j-1+\ell_n)}{m_n}, \gamma^n_k\right) \widetilde{f}^n_{r,t-1-(j-1)/m_n} + U_t, \end{aligned} \tag{1}$$

Here, q_n is the number of lags and U_t is a scalar error term. The quantity ℓ_n is the number of high-frequency leads for panel n, which we comment on in Section 3.1. For each variable in the macro panel, we consider $m_1 \times q_1 = 3$ monthly lags, accounting for publication delays and irregularities in data availability. It is worth noting that our approach can handle a higher number of lags by adjusting q_n. Regarding the weekly financial panel, we include $m_2 \times q_2 = 13$ weekly lags. The above model allows both regressors and factors to be useful predictors.

2.3 MIDAS weights

For each variable, macro, financial, or factor, we first apply MIDAS weights to map each high-frequency data to low-frequency periods. In practice, we approximate MIDAS weights using Legendre polynomials of degree $D = 3$ which are parsimonious and yield good performance, see Babii et al. (2022) for further details and the discussion on the choice of the polynomials. This allows us to address the parameter proliferation problem resulting from the mixture of frequencies in the data. Alternatively, one may opt for unrestricted MIDAS (UMIDAS) specification, see Foroni et al. (2015a) for low-dimensional models and Uematsu and Tanaka (2019) for LASSO-type implementations of the model. The disadvantage of the UMIDAS scheme is that it leads to highly over-parameterized regression model specification stemming from mixed frequency variables, which is particularly problematic when higher frequency covariates are considered such as weekly financial series. Therefore, for our main results, we apply flexible MIDAS weights but, later on, this paper also considers the UMIDAS scheme with LASSO as in Uematsu and Tanaka (2019) and its factor-augmented case.

Formally, the map is defined as

$$\begin{aligned} \omega(s;\beta_k^n) &= \sum_{d=0}^{D} \beta_{k,d}^n w_d(s), \quad \forall s \in [0,1] \\ \omega(s;\gamma_r^n) &= \sum_{d=0}^{D} \beta_{k,d}^n w_d(s), \quad \forall s \in [0,1] \end{aligned} \tag{2}$$

where $(w_d)_{d\geq 0}$ is the dictionary and D is the number of polynomials used to model high-frequency lag polynomials. In the MIDAS literature, equation (2) usually contains *approximate* equalities, reflecting that ω is well approximated by a low-dimensional polynomial.

Here, for simplicity, we write *exact* equalities.

Using (2), we can rewrite the regression model (1) as

$$Y_t = \rho_0 + \sum_{j=1}^{\mathrm{J}} \rho_j Y_{t-j} + \sum_{n=1}^{N} \sum_{k=1}^{K_n} \sum_{d=1}^{D} \beta_{k,d}^n X_{k,d,t} + \sum_{n=1}^{N} \sum_{k=1}^{R_n} \sum_{d=1}^{D} \gamma_{k,d}^n f_{r,d,t}^n + U_t, \tag{3}$$

where, for $n \in [N]$, $k \in [K_n]$, $d \in [D]_+$, $t \in [T]$

$$\begin{aligned} X_{k,d,t} &= \sum_{j=1-\ell_n}^{m_n \times q_n} w_d\left(\frac{(j-1+\ell_n)}{m_n}\right) \widetilde{X}_{k,t-1-(j-1)/m_n}. \\ f_{r,d,t}^n &= \sum_{j=1-\ell_n}^{m_n \times q_n} w_d\left(\frac{(j-1+\ell_n)}{m_n}\right) \widetilde{f}_{r,t-1-(j-1)/m_n}^n. \end{aligned} \tag{4}$$

are MIDAS-weighted variables. The rewritten model (2) is linear in parameters and parsimonious.

2.4 Sparse and dense models without factors

We study several approaches applied to nowcasting using the general MIDAS regression model 1. In this subsection, we describe the case where there are no factors, that is, we restrict $\gamma = 0$.

In our nowcasting application, the number of variables $K = \sum_{n=1}^{N} K_n$ is large relative to the sample size T. Therefore, using ordinary least squares would lead to over-fitting and poor out-of-sample performance. Instead, we rely on methodologies crafted for high-dimensional regression problems. The different possible techniques will depend on the assumptions made on the parameter β. A standard approach is to assume that β is sparse, i.e., that only a few coefficients are nonzero. Applying the ℓ_1-norm penalty on β coefficients leads to sparse and parsimonious parameter estimates, which is the rationale behind the popular LASSO estimator of Tibshirani (1996). Note that the sparsity assumption can be relaxed, e.g., β can assumed to be approximately sparse, which means that the coefficient vector β is closely approximated in ℓ_1 norm with a sparse vector (Belloni et al., 2020; Babii et al., 2022).

In some cases, the sparsity pattern often exhibits a specific structure. For example, parameters can frequently be categorized into distinct groups, where it is common for all parameters within the same group to be either entirely zero or non-zero. To address

this group structure, a grouped ℓ_1-norm penalty can be applied, as introduced by (Yuan and Lin, 2006). In the context of nowcasting US GDP, Babii et al. (2022) argues that parameters associated with the same predictor variable, i.e., stemming from high-frequency lags, should be grouped. They advocate the use of the sparse-group LASSO (sg-LASSO) estimator to take advantage of structured sparsity patterns relevant to MIDAS regression models. Formally, let

$$B = \{(b_{k,d}^n)_{n\in[N],k\in[K_n],d\in[D]}\}$$

be the parameter space of coefficient vector β. The sg-LASSO estimator solves

$$\min_{(p,b)\in\mathbb{R}^{\mathrm{J}+1}\times B} \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - p_0 - \sum_{j=1}^{\mathrm{J}} p_j Y_{t-j} - \sum_{n=1}^{N}\sum_{k=1}^{K_n}\sum_{d=1}^{D} b_{k,d}^n X_{k,d,t}\right)^2 + \mu\Omega(b), \tag{5}$$

where $\mu > 0$ is the penalty level and $\Omega(b) = \alpha|b|_1 + (1-\alpha)\|b\|_{2,1}$ is a combination of the LASSO penalty $|b|_1 = \sum_{n=1}^{N}\sum_{k=1}^{K_n}\sum_{d=1}^{D}\left|b_{k,d}^n\right|$ and the group LASSO penalty $\|b\|_{2,1} = \sum_{n=1}^{N}\sum_{k=1}^{K_n}\sqrt{\left(\sum_{d=1}^{D}\left(b_{k,d}^n\right)^2\right)}$. The ℓ_1-penalty enforces sparsity at the level of each coefficient, while the group penalty accomplishes sparsity at the level of original predictors in each panel. We call this method sg-LASSO-MIDAS.

For certain data generating processes, the sparsity assumption may be too restrictive. In this case, we may use dense models, in which we estimate the parameters under the assumption that many are nonzero yet small. A standard estimator for dense linear regression is the ridge estimator (Hsu et al., 2012) which uses an ℓ_2 penalty. The latter has been adapted to mixed-frequency data by Babii (2022) and used in nowcasting by Ferrara and Simoni (2023). In our context, the ridge estimator solves

$$\min_{(p,b)\in\mathbb{R}^{\mathrm{J}+1}\times B} \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - p_0 - \sum_{j=1}^{\mathrm{J}} p_j Y_{t-j} - \sum_{n=1}^{N}\sum_{k=1}^{K_n}\sum_{d=1}^{D} b_{k,d}^n X_{k,d,t}\right)^2 + \mu|b|_2, \tag{6}$$

where $\mu > 0$ is the penalty level and $|b|_2^2 = \sum_{n=1}^{N}\sum_{k=1}^{K_n}\sum_{d=1}^{D}\left(b_{k,d}^n\right)^2$ is the ℓ_2-norm on B. We will refer to this method as Ridge-MIDAS.

The dense assumption may also not be justified. As noted by Chernozhukov et al. (2017), a way to relax further the assumptions, is to assume that $\beta = \zeta + \eta$, where ζ is sparse and η is dense. This is called a sparse plus dense model. Chernozhukov et al. (2017) proposes a LAVA estimator which penalizes ζ by the ℓ_1-norm and η by the ℓ_2-norm. As argued for the sg-LASSO-MIDAS method, in the context of nowcasting, it is beneficial to

rely on a structured sparsity pattern. Hence, to take advantage of the sparse plus dense pattern, we can define the sg-LAVA estimator, which is a natural counterpart of the LAVA estimator in our setting and solves

$$\min_{(p,z,h)\in\mathbb{R}^{\mathrm{J}+1}\times B^2} \frac{1}{T}\sum_{t=1}^{T}\Big(Y_t - p_0 - \sum_{j=1}^{\mathrm{J}} p_j Y_{t-j} - \sum_{n=1}^{N}\sum_{k=1}^{K_n}\sum_{d=1}^{D}(\zeta_{k,d}^n + \eta_{k,d}^n)X_{k,d,t}\Big)^2 + \mu_1\Omega(\zeta) + \mu_2|\eta|_2, \tag{7}$$

where $\mu_1, \mu_2 > 0$ are two penalty levels. We call sg-LAVA-MIDAS the method that applies the sg-LAVA estimator .

2.5 Factor-augmented regression models

In this section, we consider the case where γ is non-zero and hence we use factors in the regression models. To estimate the factors, we impose an approximate factor model on the regressors. That is, there exist real-valued factors loadings $\{\lambda_{k,r}^n, n \in [N], k \in [K_n], r \in [R]\}$ and error terms $\{\widetilde{E}_{k,t}^n, n \in [N], k \in [K_n], t \in [T]\}$ such that

$$\widetilde{X}_{k,t-(j-1)/m}^n = \sum_{r=1}^{R_n} \lambda_{k,r}^n \widetilde{f}_{t-(j-1)/m,r}^n + \widetilde{E}_{k,t-(j-1)/m}^n, \tag{8}$$

for all $n \in [N], t \in [T], j \in [m], k \in [K_n]$.

Equation (8) can be rewritten as

$$X_{k,d,t}^n = \sum_{r=1}^{R_n} \lambda_{k,r}^n f_{r,d,t}^n + E_{k,d,t}^n, \tag{9}$$

where

$$E_{k,d,t}^n = \sum_{j=1-\ell_n}^{m_n\times q_n} w_d\left(\frac{(j-1+\ell_n)}{m_n}\right)\widetilde{E}_{k,t-1-(j-1)/m_n}^n. \tag{10}$$

This allows for estimating the factors by principal components analysis on the variables $X_{k,d,t}^n$ and obtaining estimators $\{\hat{f}_{r,d,t}^n\}$. We also denote by $\widehat{R}_n$ the estimated number of factors in panel n, which can be estimated using e.g., the growth ratio estimator of Ahn and Horenstein (2013).

If $\beta = 0$, we can then regress the target variable on its lags and estimated factors, which corresponds to the classical principal component regression model. We let

$$C = \{(\gamma_{r,d}^n)_{n\in[N],r\in[R_n],d\in[D]}\}$$

be the parameter set for γ. Thus, we define the FAMIDAS method, which is equivalent to the factor-augmented regression of Stock and Watson (2002) in a single-frequency setting and the methodology of Andreou et al. (2013) in a mixed-frequency data context and solves

$$\min_{(p,c)\in\mathbb{R}^{\mathrm{J}+1}\times C} \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - p_0 - \sum_{j=1}^{\mathrm{J}} p_j Y_{t-j} - \sum_{n=1}^{N}\sum_{r=1}^{\widehat{R}_n}\sum_{d=1}^{D} c_{k,d}^{n}\hat{f}_{r,d,t}^{n}\right)^2. \tag{11}$$

This is a dense approach because (almost) all regressors matter to estimate the factors, which themselves are used to compute the predictions.

Instead, if $\beta \neq 0$, a natural approach to estimation is to assume that β follows a group-sparse structure as in the previous subsection and to apply the sg-LASSO estimator. We thus introduce a new sparse plus dense approach, which we call sg-LASSO-FAMIDAS and which solves

$$\begin{aligned}\min_{(p,c,b)\in\mathbb{R}^{\mathrm{J}+1}\times C\times B} \frac{1}{T}\sum_{t=1}^{T}\Big(Y_t - p_0 - \sum_{j=1}^{\mathrm{J}} p_j Y_{t-j} - \sum_{n=1}^{N}\sum_{k=1}^{K_n}\sum_{d=1}^{D} b_{k,d}^{n} X_{k,d,t} \\ - \sum_{n=1}^{N}\sum_{r=1}^{\widehat{R}_n}\sum_{d=1}^{D} c_{k,d}^{n}\hat{f}_{r,d,t}^{n}\Big)^2 + \mu\Omega(b),\end{aligned} \tag{12}$$

where μ is the penalty level.

3 Data

3.1 Data flow in nowcasting

Our primary objective is to conduct nowcasting for the current quarter's GDP growth and compute performance metrics across various monthly horizons. It is crucial to note that the release of US GDP data typically occurs one month after the quarter. Figure 2 provides a visual representation of the timeline involved in the nowcasting process. Consider the following scenario: we find ourselves at the end of the first month, aiming to nowcast the GDP for the first quarter. In pursuit of this goal, the econometrician trains the model using data spanning up to the first month of the fourth quarter of the preceding year. Moreover, there is a need to incorporate the data from the first month of the current quarter (Q1) to generate the prediction, or nowcast in our case. The parameter h represents the time remaining until the end of the target quarter. We will consider three nowcasting exercises:

forecasting the GDP of the current quarter two months before its conclusion, one month ahead, and nowcasting at the end of the quarter.

In our model, we use ℓ_n to denote the number of leads in high-frequency data used. It is frequency-indexed and varies based on the nowcasting horizon. For instance, in the context of the monthly panel of predictor variables, $\ell_1 \in (1, 2, 3)$, represents the number of leading months in monthly predictors. Conversely, in the weekly panel, $\ell_2 \in (4, 8, 13)$, which approximately corresponds to the number of leading weeks. The horizon index reflects the availability of high-frequency data within the quarter t. For example, when $\ell_1 = 1$ or $\ell_2 = 4$, we incorporate 1 leading month or 4 weeks of available high-frequency data within the quarter, resulting in a horizon of 2 months.

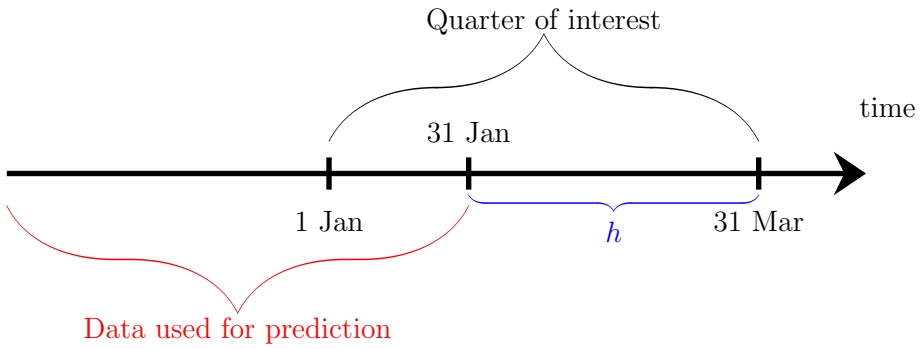


Figure 2: Timeline illustration of nowcasting.

We apply our methods to nowcast the initial release of real US GDP growth, often referred to as the "advance" release. Typically, this release becomes available towards the end of the first month of the subsequent quarter, except for the 2018 Q4 release. In this case, the first release was made available on February 28$^{\text{th}}$, 2019 (second month) due to a partial government shutdown from December 22, 2018, to January 25, 2019, lasting 35 days. Additional details regarding this event can be found at the Bureau of Economic Analysis. In our analysis, we utilized real-time GDP vintages sourced from the Archival Federal Reserve Economic Data St. Louis Fed (ALFRED) database. We have set the in-sample data period from 1984 Q1 to 2007 Q4. Our first nowcast is for 2008 Q1, and we use an expanding window approach to compute nowcasts for subsequent quarters.

Monthly macroeconomic panel. The monthly covariates and factors are based on the well-known FRED-MD real-time dataset, see McCracken and Ng (2016) for further details. We restrict the series such that all of them are available in all vintages used for out-of-sample forecasting. We also omit all financial series from the monthly panel. In total, we use 76 monthly series in our final monthly panel.

Weekly financial panel. For the weekly factor, we use weekly financial series which are also used to construct the National Financial Conditions Index (NFCI) by the Chicago Fed. The original NFCI is a latent factor comprised of weekly, monthly, and quarterly series and is computed at a weekly frequency. Instead of using the full list of series, we opt to use the majority of weekly series in the NFCI index without using monthly or quarterly series. In total, we collected 36 weekly financial series. Since financial data is accessible in real-time, there is no need to account for publication delays or irregularities for this set of predictor variables. However, some financial series may have shorter durations compared to macroeconomic and other longer financial series. To incorporate all series effectively, we employ matrix completion methods based on nuclear-norm regularization to impute missing data. This technique helps balance the financial series and is applied at the original frequency of the variables, see Section B of the Online Appendix for details.

We provide the full list of series for monthly macro and weekly financial data with additional details in Section A of the Online Appendix.

4 Results

4.1 Baseline results

Our candidate model set includes the AR(4) model, which we regard as the simplest and designate as our benchmark model. Consequently, our reported root mean squared forecast errors are presented relative to the AR(4). Notably, our conclusions remain consistent even when employing alternative benchmarks such as the random walk or AR(1). Our baseline results for sparse plus dense methods are based on computations assuming the monthly macro data has a sparse plus dense structure, while the weekly financial data exclusively comprises a sparse component. All other benchmarks utilize the full dataset by

assuming either a dense-only or sparse-only regression model. The nowcasts for all models are computed as detailed in Section 2. In addition, we choose all tuning parameters by cross-validation adjusting for time series dependence, see, e.g., Section 4 in Babii et al. (2022). We estimate the number of factors through the growth ratio estimator of Ahn and Horenstein (2013).

Our primary findings are summarized in Table 1, presenting results for three sub-samples: Panel A encompasses the entire dataset, spanning the out-of-sample evaluation period from 2008 Q1 to 2022 Q2 (Full sample); Panel B focuses on the sub-sample up to 2019 Q4, excluding the COVID-19 period (Up to COVID); and Panel C shows results for the sub-sample from 2020 Q1 to 2022 Q2, covering the COVID-19 pandemic and subsequent periods (COVID and afterward). The reported outcomes include root-mean-squared nowcast errors for six models: AR(4) benchmark, sparse (Panel A1-C1), dense (Panel A2-C2), and sparse plus dense (Panel A3-C3) methods. Notably, results for the autoregressive benchmark model are expressed in absolute terms, while the remaining RMSEs are presented relative to this benchmark. To assess the statistical significance, we employed the average superior predictive ability (aSPA) test proposed by Quaedvlieg (2021), enabling a comprehensive evaluation across three horizons. The results of the aSPA tests are detailed in Table 2.

First, our results reveal pronounced shifts in performance across the three out-of-sample periods and models. Notably, the benchmark model experiences a substantial increase in prediction errors during the COVID-19 period – an expected outcome given the autoregressive model's inherent lack of high-frequency information. The sparse model, specifically sg-LASSO-MIDAS, shows superior performance compared to other methods before the onset of the pandemic, with sg-LASSO-FAMIDAS being the exception (Table 1 Panel B). This improvement is statistically significant at a 10% confidence level (refer to Table 2 Panel B, column 3). Interestingly, the sparse model outperforms the sg-LAVA-MIDAS approach, which belongs to the sparse plus dense class of models, while not surpassing sg-LASSO-FAMIDAS, also from the same class of models. However, results change when considering the entire sample. Notably, both sparse plus dense models yield more accurate nowcasts compared to exclusively sparse or dense models. This improvement reaches

	2-month	1-month	EoQ
	Panel A. *Full sample*		
AR(4)	9.553	9.553	9.553
	Panel A1. *Sparse*		
sg-LASSO-MIDAS	0.572	0.435	0.559
	Panel A2. *Dense*		
Ridge-MIDAS	0.599	0.504	0.595
FAMIDAS	0.590	0.495	0.580
	Panel A3. *Sparse plus dense*		
LAVA-MIDAS	0.589	0.493	0.407
sg-LASSO-FAMIDAS	0.580	0.340	0.251
	Panel B. *Up to COVID*		
AR(4)	1.934	1.934	1.934
	Panel B1. *Sparse*		
sg-LASSO-MIDAS	0.827	0.837	0.733
	Panel B2. *Dense*		
Ridge-MIDAS	0.988	0.937	0.866
FAMIDAS	0.962	0.911	0.851
	Panel B3. *Sparse plus dense*		
LAVA-MIDAS	0.959	0.908	0.837
sg-LASSO-FAMIDAS	0.842	0.827	0.756
	Panel C. *COVID and afterwards*		
AR(4)	22.613	22.613	22.613
	Panel C1. *Sparse*		
sg-LASSO-MIDAS	0.561	0.414	0.552
	Panel C2. *Dense*		
Ridge-MIDAS	0.582	0.482	0.584
FAMIDAS	0.573	0.474	0.568
	Panel C3. *Sparse plus dense*		
sg-LAVA-MIDAS	0.572	0.472	0.383
sg-LASSO-FAMIDAS	0.569	0.309	0.213

Table 1: Nowcast comparisons — horizons are 2- and 1-month ahead, as well as the end of the quarter (EoQ). We report results for the full sample in Panel (A), and Panel (B) results excluding the COVID pandemic period, while Panel (C) reports results for the COVID pandemic period and beyond. The out-of-sample period starts from 2008 Q1 to 2022 Q2 Panel (A) and 2019 Q4 Panel (B) and starts from 2020 Q1 to 2022 Q2 Panel (C). The RMSEs are reported relative to the AR model.

	1	2	3	4
Panel A. *Full sample*				
2	**0.026**			
3	**0.016**	**0.030**		
4	**0.018**	**0.022**	**0.024**	
5	**0.020**	**0.034**	**0.020**	0.984
Panel B. *Up to COVID*				
2	**0.020**			
3	0.560	0.976		
4	**0.010**	0.152	**0.030**	
5	**0.028**	0.380	**0.070**	0.834

Table 2: Nowcast comparisons — We report the p-values of the average superior predictive ability bootstrap tests (aSPA) of Quaedvlieg (2021) over all three horizons comparing 1 (sg-LASSO-FAMIDAS), 2 (sg-LAVA-MIDAS), 3 (sg-LASSO-MIDAS), 4 (Ridge-MIDAS), and 5 (FAMIDAS). We test the null hypothesis that the average out-of-sample loss over the three horizons is smaller for the models in the column versus in the row. Bold entries indicate 10% significance.

statistical significance at a 5% confidence level (refer to Table 2 Panel A, columns 1-2) for both sg-LAVA-MIDAS and sg-LASSO-FAMIDAS methods. Furthermore, the sparse model outperforms dense models even when considering the full sample (see Table 2 Panel A, column 3). In contrast, there is no significant difference between the two dense models under consideration (refer to Table 2 Panel A-B, column 4). We should note that it is not clear if such tests of equal forecasting accuracy are valid in such a context with penalized estimators and estimated factors, so that these inference results should be considered with caution.

In Figure 3, we plot the square-root cumulative sum of squared forecast error (CUMSUM) for the sparse or dense models (Figure 3a and 3c) and for the sparse plus dense models (Figure 3b and 3d). With these graphs, we try to visualize the differences in the performance of these models throughout the out-of-sample period. We use forecast errors

for the end-of-quarter horizon. The CUMSUM is computed as

$$\text{CUMSUM}_{t,t+k} = \sqrt{\sum_{q=t}^{t+k} \hat{\epsilon}_{q,j}^2},$$

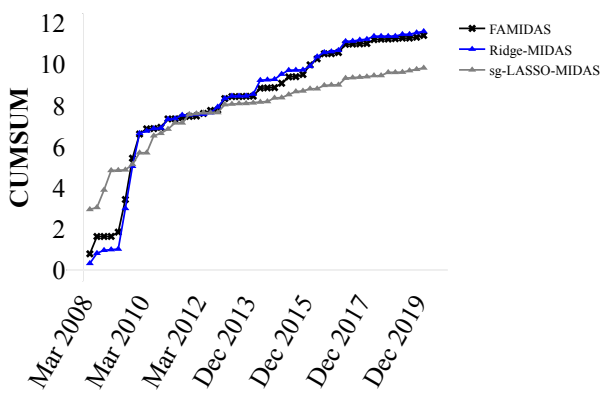
where $\hat{\epsilon}_{q,j}$, $j \in \{$sg-LASSO-MIDAS, Ridge-MIDAS, FAMIDAS$\}$ for either sparse or dense models and $\hat{\epsilon}_{q,j}$, $j \in \{$sg-LAVA-MIDAS, sg-LASSO-FAMIDAS$\}$ for sparse plus dense models, are the out-of-sample nowcast errors. We plot the CUMSUM for the full sample period, corresponding to Panel A in Table 1, and up to the COVID pandemic which corresponds to Panel B in the same Table.

The plots highlight the contrast between sparse or dense models versus the sparse plus dense models, especially during the onset of the COVID outbreak. Before the pandemic, the sparse model, sg-LASSO-MIDAS, appears to yield the most accurate results (Figure 3a). However, the performance of the sparse plus dense sg-LASSO-FAMIDAS model is similar. This implies that factor augmentation may not have a large impact on the performance during stable periods even though we need to estimate additional unpenalized regression coefficients as well as factors themselves. As the COVID period unfolds, both sparse and dense models experience a marked decline in prediction quality, whereas the sparse plus dense methods exhibit a notable resilience in handling the substantial shock introduced by the pandemic.

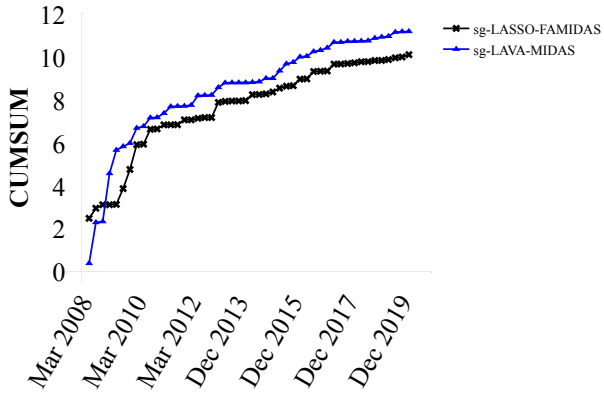
Notably, both sg-LASSO-FAMIDAS and sg-LAVA-MIDAS are less influenced by this shock. Nevertheless, the latter displays a more pronounced impact compared to the former and, overall, produces less accurate nowcasts throughout the out-of-sample evaluation period. Further discussion of potential reasons for these differences is provided in Section 4.2.

4.2 On model structure

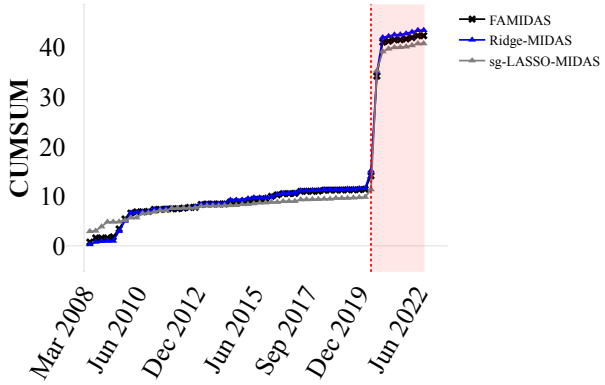
Our out-of-sample nowcasting results indicate that applying a sparse plus dense structure on a macro panel of predictors improves the overall performance when compared to sparse or dense alternatives. In the current section, we further explore the role of the sparse and dense components.



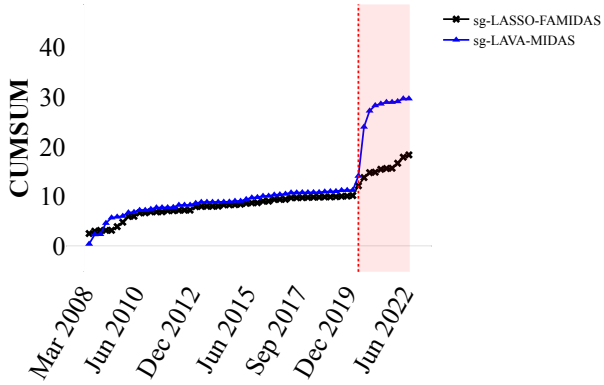
(a) Sparse or dense.



(b) Sparse plus dense.



(c) Sparse or dense.



(d) Sparse plus dense.

Figure 3: The figure illustrates the CUMSUM for the sparse or dense models (Figure 3a-3c) and for the sparse and dense models (Figure 3b-3d). Figures 3a-3b plot CUMSUM using the pre-COVID pandemic period while 3c-3d plot the whole out-of-sample period.

4.2.1 Model specification tests

First, we test the presence of the sparse and the dense parts of the model. Specifically, applying the tuning-free bootstrap test of Beyhum and Striaukas (2023), we test the null hypothesis that the regression coefficient of the sparse part of the model is equal to zero. Our findings reveal that for the entire (real-time) sample using the end of the quarter horizon, the sparse component is significant with a p-value of 0.002. When applying the test, from the beginning of the period to just before the COVID pandemic, we obtain a p-value of 0.037. This implies that idiosyncratic shocks are useful to explain the GDP. Furthermore, we test the null hypothesis that the coefficients of the factors are zero. To do so, we use the debiased LASSO with the Gaussian multiplier bootstrap procedure, as outlined in Dezeure et al. (2017), to examine the significance of the PCA factors as predictors. The initial parameter estimates are based on a LASSO regression where we do not penalize the coefficients of the factors. We obtain a p-value of 0.003 when using the full sample, confirming that PCA factors are significant predictor variables. Upon applying the testing procedure to the pre-COVID data, we observe a p-value of 0.668. Consequently, the factors are not significant for the period preceding the pandemic. These significance testing results align with our out-of-sample nowcasting results. Specifically, the sparse model demonstrates superior accuracy in nowcasts before the onset of COVID. In contrast, the sparse plus dense models show substantial improvement when considering the entire sample, including the COVID period.

4.2.2 Factor-augmented regression versus LAVA regularization

Although the sparse plus dense structure proves resilient to the COVID shock in providing good nowcasting results, a nuanced distinction in performance emerges between the two methods considered. Specifically, the LAVA estimator-based procedure yields less accurate nowcasts when compared to the factor-augmented sparse regression. An intuitive explanation for this difference lies in the factor model's greater flexibility in capturing dense shocks, as opposed to the potentially more intricate penalized regression approach employed by the LAVA estimator. Furthermore, the LAVA estimator necessitates the selection of an additional tuning parameter, introducing the possibility of increased noise in the parameter

estimates.

4.2.3 Further specifications of predictive regressions

In the application of the sg-LASSO-FAMIDAS method, we explore alternative structures by i) incorporating panel-specific factors, encompassing both macro and financial data factors; ii) implementing a group factor structure; and iii) using a different method for determining the optimal number of factors. We address these variations and present corresponding additional out-of-sample nowcasting results in Table 3.

Firstly, the inclusion of panel-specific factors improves nowcasts for the 1-month horizon, but this improvement is not observed at the end of the quarter (Table 3 Panel A). Notably, and in alignment with our primary findings, the overall results appear to be less favorable compared to the models that use macro-specific factors. Secondly, the estimation of both group and panel-specific factors does not seem to contribute significantly to the accuracy of GDP nowcasting (Table 3 Panel B). In this case, we follow Andreou et al. (2019) estimation and inference procedure to estimate the group and panel-specific factors. Both results appear to hold true particularly for the COVID and post-COVID sample periods. The distinct shocks experienced by macro and financial data series during the pandemic likely account for this outcome. Lastly, our findings are robust across various methods used to determine the number of factors (Table 3 Panel C). In this case, we use the same model structure as in Table 1, but replace the growth ratio estimator with the eigenvalue ratio estimator for the number of macro factors. This emphasizes the reliability and consistency of our results regardless of the chosen methodology.

4.2.4 Economic rationale

The above results suggest that the data follows a sparse pattern before the COVID pandemic, but that the COVID shock is a mix of idiosyncratic shocks and common shocks on the macroeconomic variables, while only consisting in idiosyncratic shocks on the financial market.

It is possible to give an economic rationale explaining these results. The stringent physical restrictions imposed during the pandemic exerted a profound impact on most of the real

	2-month	1-month	EoQ
	Panel (A) *Panel specific factors*		
Full sample	0.611	0.315	0.356
Up to COVID	0.958	0.883	0.818
COVID and afterwards	0.595	0.275	0.328
	Panel (B) *Group factors*		
Full sample	0.786	0.536	0.467
Up to COVID	0.943	0.808	0.815
COVID and afterwards	0.780	0.524	0.450
	Panel (C) *Eigenvalue ratio*		
Full sample	0.580	0.340	0.251
Up to COVID	0.842	0.827	0.756
COVID and afterwards	0.569	0.309	0.213

Table 3: Nowcast comparisons of different factors models — horizons are 2- and 1-month ahead, as well as the end of the quarter (EoQ). We report results for panel-specific factors in Panel (A), Panel (B) reports results based on group factors, while Panel (C) reports results for the factors based on the eigenvalue ratio estimator for the number of factors. The out-of-sample period starts from 2008 Q1 to 2022 Q2. The RMSEs are reported relative to the AR model.

economy, significantly affecting variables classified as macro, this is the common shocks. An in-depth examination of the events during the COVID period and their influence on key macroeconomic indicators is elaborated upon by Diebold (2020). In contrast, financial markets, while initially impacted by the COVID crisis in February and March 2020, swiftly rebounded. This rapid recovery led to a pronounced disconnect between financial markets and the real economy, a phenomenon that has garnered substantial attention from policy institutions, as evidenced by studies such as Igan et al. (2020). There was less of a general movement of stocks due to COVID, explaining the absence of the dense financial component.

4.3 Alternative methods and factors

In this section, we consider alternative methods and factors to evaluate the robustness of our results.

4.3.1 Methods

We demonstrate that our conclusions remain by exploring alternative methodologies. Specifically, we employ the UMIDAS scheme and the unstructured LASSO estimator, comparing both sparse and sparse plus dense models. Using the UMIDAS scheme in conjunction with a dimension reduction technique aligns with a widely used strategy in the literature on modeling mixed frequency data, as proposed by the studies Foroni et al. (2015b) and Uematsu and Tanaka (2019), among others. In this context, the sg-LASSO-FAMIDAS method transforms into LASSO-FAUMIDAS, while sg-LASSO-MIDAS becomes LASSO-UMIDAS. Results of both methods appear in Table 4.

Our analysis indicates that the UMIDAS model with an unstructured LASSO estimator consistently underperforms the MIDAS approach using the structured sg-LASSO estimator across the out-of-sample period. Notably, the sparse-only method improved over the sparse plus dense approach before the COVID-19 pandemic, but the subsequent shock affected both unstructured approaches and similarly structured approaches. That is, the macro factor augmentation proves beneficial in improving nowcast quality. Thus, the findings remain similar using alternative approaches.

	2-month	1-month	EoQ
	Panel A. *Full sample*		
LASSO-UMIDAS	0.567	0.395	0.433
LASSO-FAUMIDAS	0.569	0.403	0.407
	Panel B. *Up to COVID*		
LASSO-UMIDAS	0.835	0.748	0.851
LASSO-FAUMIDAS	0.921	0.862	0.858
	Panel C. *COVID and afterwards*		
LASSO-UMIDAS	0.556	0.377	0.410
LASSO-FAUMIDAS	0.582	0.376	0.382

Table 4: Nowcast comparisons — horizons are 2- and 1-month ahead, as well as the end of the quarter (EoQ). We report results for the full sample in Panel (A), and Panel (B) results excluding the COVID pandemic period, while Panel (C) reports results for the COVID pandemic period and beyond. The out-of-sample period starts from 2008 Q1 to 2022 Q2 Panel (A) and 2019 Q4 Panel (B) and starts from 2020 Q1 to 2022 Q2 Panel (C). The RMSEs are reported relative to the AR(4) model.

4.3.2 Alternative factors

	2-month	1-month	EoQ
	Panel A. *Full sample*		
ADS	0.553	0.379	0.258
CFNAI	0.325	0.446	0.334
NFCI	0.699	0.505	0.549
	Panel B. *Up to COVID*		
ADS	0.810	0.782	0.742
CFNAI	0.800	0.814	0.732
NFCI	0.895	0.932	0.924
	Panel C. *COVID and afterwards*		
ADS	0.542	0.357	0.222
CFNAI	0.295	0.427	0.311
NFCI	0.681	0.483	0.531

Table 5: Nowcast comparisons — horizons are 2- and 1-month ahead, as well as the end of the quarter (EoQ). We report results for the full sample in Panel (A), and Panel (B) results excluding the COVID pandemic period, while Panel (C) reports results for the COVID pandemic period and beyond. The out-of-sample period starts from 2008 Q1 to 2022 Q2 Panel (A) and 2019 Q4 Panel (B) and starts from 2020 Q1 to 2022 Q2 Panel (C). The RMSEs are reported relative to the AR(4) model.

Further to analyzing alternative approaches, we conduct a comparative assessment of the sg-LASSO-FAMIDAS method replacing the macro principal components factors with commonly used indicators in the literature. Specifically, we consider the Aruoba-Diebold-Scotti index (ADS, weekly) as detailed in Aruoba et al. (2009), the Chicago Fed National Activity Index (CFNAI) as outlined in Brave et al. (2019), and the National Financial Conditions Index (NFCI, Brave and Kelley, 2017, Amburgey and McCracken, 2023). For all three factors, we use vintages sourced from the Federal Reserve Banks, specifically from Philadelphia, Chicago, and St. Louis, respectively. We use the same MIDAS scheme and the number of lags for all factors as for the predictors, that is, ADS and NFCI are

treated as a weekly variable while CFNAI is monthly. Notably, for NFCI, we use vintages made available by Amburgey and McCracken (2023) due to its longer period, aligning more effectively with our empirical application. Further details regarding the exact data sources, including web links, are provided in the Online Appendix, while results are reported in Table 5.

We find that both ADS and CFNAI factors outperform the NFCI factor, yielding more accurate nowcasts across various subsamples. Notably, both ADS and CFNAI factors rely exclusively on macroeconomic indicators, while NFCI is rooted in financial series. This emphasizes the robustness of our primary conclusion, even when considering factors not originating from our main datasets. In the comparison between ADS and CFNAI, it becomes evident that ADS consistently enhances the nowcasts of the sg-LASSO-FAMIDAS method. One plausible explanation is that ADS, being a weekly series, contributes to smoother and less noisy MIDAS weight estimates. Additionally, the ADS series is based on eight predictors, including lagged GDP growth, whereas CFNAI is comprised of many variables, often exceeding a hundred. Consequently, the ADS indicator itself appears to be less prone to noise, offering a more accurate signal regarding the current state of the economy. This finding is in line with Diebold (2020), who also found the ADS index particularly useful for nowcasting during the pandemic.

5 Conclusion

This paper introduces an extension to the class of high-dimensional MIDAS regression models designed for handling sparse plus dense signals. We propose two innovative methods tailored for handling such patterns in the context of mixed-frequency data and apply these methods to nowcast US GDP growth using conventional monthly macro and weekly financial datasets. Our empirical findings reveal that tackling the COVID shock is particularly challenging, yet models incorporating a dense macro component demonstrate promising results. These findings stress the importance of adapting prediction approaches to the data structures, thereby improving the quality of predictions, as evident in our nowcasts.

It is worth noting that there is a vibrant body of work that explores the utilization of novel types of data for nowcasting GDP. For example, Ferrara and Simoni (2022) exam-

ine the usefulness of Google data in improving GDP nowcasting. Interestingly, they also discover that the impact on performance varies between periods of economic stability and recession. Barbaglia et al. (2022); Ellingsen et al. (2022) explore the use of news-based data for forecasting purposes. Such data are available at a much higher frequency and there is preliminary evidence that they can help in the daunting task of nowcasting during the COVID pandemic (Ferrara and Sheng, 2022). Our research differs from these studies in that we focus on factor models and LASSO regression as alternative approaches to extracting information from the available data, rather than considering different types of data sources. As an avenue for further research, we believe that combining our novel methods and these innovative datasets may yield even more accurate results.

In addition, our focus is primarily on frequentist MIDAS regression to handle the mixed-frequency aspect of our data. However, it is important to note that alternative approaches, such as dynamic factor models (Bok et al., 2018; Doz and Fuleky, 2020; Ruiz et al., 2022) (for factors) or Bayesian MIDAS regression/mixed frequency VAR (Kuzin et al., 2011; Schorfheide and Song, 2015; Mogliani and Simoni, 2021) can also be employed in practice. By considering sparse plus dense representations within these frameworks, one may obtain different results and potentially uncover improved empirical findings. Exploring alternative methods may provide further insights and enrich the overall quality of the data analysis.

Acknowledgments

The authors thank Andrii Babii, Luca Barbaglia, Ferre De Graeve, Catherine Doz, Geert Dhaene, Domenico Giannone, Eric Ghysels, Peter Reusens, Boriss Siliverstovs, Wouter Van der Veken, Raf Wouters, seminar and conference participants at CBS, COMPSTAT 2023, CREST, Paris School of Economics and EcoSta 2023 for helpful comments. An earlier version of the paper was circulated under the title "Factor-augmented sparse MIDAS regression for nowcasting". Jad Beyhum undertook most of this work while employed by CREST, ENSAI (Rennes). Jad Beyhum gratefully acknowledges financial support from the Research Fund KU Leuven through the grant STG/23/014. Jonas Striaukas gratefully acknowledges the financial support of F.R.S.-FNRS PDR project Nr. PDR T.0044.22 and from the European Commission, MSCA-2022-PF Individual Fellowship, Project 101103508.

Supplementary material

The Online Appendix contains additional details on the data and the matrix completion algorithm used to impute missing data.

References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227.

Amburgey, A. J. and McCracken, M. W. (2023). On the real-time predictive content of financial condition indices for growth. *Journal of Applied Econometrics*, 38(2):137–163.

Andreou, E., Gagliardini, P., Ghysels, E., and Rubin, M. (2019). Inference in group factor models with an application to mixed-frequency data. *Econometrica*, 87(4):1267–1305.

Andreou, E., Ghysels, E., and Kourtellos, A. (2010). Regression models with mixed sampling frequencies. *Journal of Econometrics*, 158(2):246–261.

Andreou, E., Ghysels, E., and Kourtellos, A. (2013). Should macroeconomic forecasters use daily financial data and how? *Journal of Business & Economic Statistics*, 31(2):240–251.

Aruoba, S. B., Diebold, F. X., and Scotti, C. (2009). Real-time measurement of business conditions. *Journal of Business & Economic Statistics*, 27(4):417–427.

Babii, A. (2022). High-dimensional mixed-frequency IV regression. *Journal of Business & Economic Statistics*, 40(4):1470–1483.

Babii, A., Ghysels, E., and Striaukas, J. (2022). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, 40(3):1094–1106.

Bai, J., Ghysels, E., and Wright, J. H. (2013). State space models and midas regressions. *Econometric Reviews*, 32(7):779–813.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics*, 131(4):1593–1636.

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

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., and Kato, K. (2020). High-dimensional econometrics and generalized GMM. *Handbook of Econometrics (forthcoming)*.

Beyhum, J. and Striaukas, J. (2023). Tuning-free testing of factor regression against factor-augmented sparse alternatives. *arXiv preprint arXiv:2307.13364*.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705 – 1732.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., and Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. *Annual Review of Economics*, 10:615–643.

Brave, S. A., Butters, R. A., Kelley, D., et al. (2019). A new "big data" index of us economic activity. *Economic Perspectives, Federal Reserve Bank of Chicago*, 1.

Brave, S. A. and Kelley, D. (2017). Introducing the chicago fed's new adjusted national financial conditions index. *Chicago Fed Letter*, 386:2017.

Brownlees, C., Guomundsson, G. S., and Wang, Y. (2023). Performance of empirical risk minimization for principal component regression.

Chernozhukov, V., Hansen, C., and Liao, Y. (2017). A lava attack on the recovery of sums of dense and sparse signals.

Dezeure, R., Bühlmann, P., and Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *Test*, 26:685–719.

Diebold, F. X. (2020). Real-time real economic activity: Exiting the great recession and entering the pandemic recession. Working Paper 27482, National Bureau of Economic Research.

Doz, C. and Fuleky, P. (2020). Dynamic factor models. *Macroeconomic Forecasting in the Era of Big Data: Theory and Practice*, pages 27–64.

Ellingsen, J., Larsen, V. H., and Thorsrud, L. A. (2022). News media versus FRED-MD for macroeconomic forecasting. *Journal of Applied Econometrics*, 37(1):63–81.

Fan, J., Lou, Z., and Yu, M. (2023a). Are latent factor regression and sparse regression adequate? *Journal of the American Statistical Association*, (just-accepted):1–77.

Fan, J., Masini, R., and Medeiros, M. C. (2023b). Bridging factor and sparse models. *Annals of Statistics* (forthcoming).

Ferrara, L. and Sheng, X. S. (2022). Guest editorial: Economic forecasting in times of covid-19. *International Journal of Forecasting*, 38(2):527–528.

Ferrara, L. and Simoni, A. (2022). When are Google data useful to nowcast GDP? an approach via preselection and shrinkage. *Journal of Business & Economic Statistics*, pages 1–15.

Ferrara, L. and Simoni, A. (2023). When are google data useful to nowcast GDP? an approach via preselection and shrinkage. *Journal of Business & Economic Statistics*, 41(4):1188–1202.

Foroni, C., Marcellino, M., and Schumacher, C. (2015a). Unrestricted mixed data sampling (midas): Midas regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 178(1):57–82.

Foroni, C., Marcellino, M., and Schumacher, C. (2015b). Unrestricted mixed data sampling (midas): Midas regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 178(1):57–82.

Foroni, C., Marcellino, M., and Stevanovic, D. (2022). Forecasting the Covid-19 recession and recovery: Lessons from the financial crisis. *International Journal of Forecasting*, 38(2):596–612.

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437.

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

Hsu, D., Kakade, S. M., and Zhang, T. (2012). Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1. JMLR Workshop and Conference Proceedings.

Igan, D., Kirti, D., and Peria, S. M. (2020). The disconnect between financial markets and the real economy. *IMF Special Notes Series on COVID-19, August*, 26:2020.

Jurado, K., Ludvigson, S. C., and Ng, S. (2015). Measuring uncertainty. *American Economic Review*, 105(3):1177–1216.

Kuzin, V., Marcellino, M., and Schumacher, C. (2011). MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area. *International Journal of Forecasting*, 27(2):529–542.

Marcellino, M. and Schumacher, C. (2010). Factor midas for nowcasting and forecasting with ragged-edge data: A model comparison for german GDP. *Oxford Bulletin of Economics and Statistics*, 72(4):518–550.

McCracken, M. W. and Ng, S. (2016). FRED-MD: a monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Mogliani, M. and Simoni, A. (2021). Bayesian MIDAS penalized regressions: estimation, selection, and prediction. *Journal of Econometrics*, 222(1):833–860.

Quaedvlieg, R. (2021). Multi-horizon forecast comparison. *Journal of Business & Economic Statistics*, 39(1):40–53.

Ruiz, E., Poncela, P., et al. (2022). Factor extraction in dynamic factor models: Kalman filter versus principal components. *Foundations and Trends® in Econometrics*, 12(2):121–231.

Schorfheide, F. and Song, D. (2015). Real-time forecasting with a mixed-frequency VAR. *Journal of Business & Economic Statistics*, 33(3):366–380.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460):1167–1179.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Uematsu, Y. and Tanaka, S. (2019). High-dimensional macroeconomic forecasting and variable selection via penalized regression. *The Econometrics Journal*, 22(1):34–56.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67.

Online Appendix

A Additional details on the data

Monthly macro data

	Description	Category	T-code
1	Real Personal Income	Output and income	5
2	Real personal income ex transfer receipts	Output and income	5
3	Industrial Production Index	Output and income	5
4	IP: Final Products and Nonindustrial Supplies	Output and income	5
5	IP: Final Products (Market Group)	Output and income	5
6	IP: Consumer Goods	Output and income	5
7	IP: Materials	Output and income	5
8	IP: Manufacturing (SIC)	Output and income	5
9	Capacity Utilization: Manufacturing	Output and income	5
10	Civilian Labor Force	Labour market	5
11	Civilian Employment	Labour market	2
12	Civilian Unemployment Rate	Labour market	5
13	Average Duration of Unemployment (Weeks)	Labour market	5
14	Civilians Unemployed - Less Than 5 Weeks	Labour market	2

15	Civilians Unemployed for 5-14 Weeks	Labour market	2
16	Civilians Unemployed - 15 Weeks & Over	Labour market	5
17	Civilians Unemployed for 15-26 Weeks	Labour market	5
18	Civilians Unemployed for 27 Weeks and Over	Labour market	5
19	Initial Claims	Labour market	5
20	All Employees: Total nonfarm	Labour market	5
21	All Employees: Goods-Producing Industries	Labour market	5
22	All Employees: Mining and Logging: Mining	Labour market	5
23	All Employees: Construction	Labour market	5
24	All Employees: Manufacturing	Labour market	5
25	All Employees: Durable goods	Labour market	5
26	All Employees: Nondurable goods	Labour market	5
27	All Employees: Service-Providing Industries	Labour market	5
28	All Employees: Wholesale Trade	Labour market	5
29	All Employees: Retail Trade	Labour market	5
30	All Employees: Financial Activities	Labour market	5
31	All Employees: Government	Labour market	5
32	Avg Weekly Hours : Goods-Producing	Labour market	5
33	Avg Weekly Overtime Hours : Manufacturing	Labour market	5
34	Avg Weekly Hours: Manufacturing	Labour market	1
35	Avg Hourly Earnings: Goods-Producing	Labour market	2
36	Avg Hourly Earnings: Construction	Labour market	1
37	Avg Hourly Earnings: Manufacturing	Labour market	4
38	Housing Starts: Total New Privately Owned	Housing	4
39	Housing Starts, Northeast	Housing	4
40	Housing Starts, Midwest	Housing	4
41	Housing Starts, South	Housing	4
42	Housing Starts, West	Housing	4
43	New Private Housing Permits, Northeast (SAAR)	Housing	4

44	New Private Housing Permits, Midwest (SAAR)	Housing	4
45	New Private Housing Permits, South (SAAR)	Housing	4
46	New Private Housing Permits, West (SAAR)	Housing	5
47	Real Manu. and Trade Industries Sales	Consumption	5
48	Retail and Food Services Sales	Consumption	5
49	New Orders for Durable Goods	Consumption	5
50	New Orders for Nondefense Capital Goods	Consumption	2
51	Unfilled Orders for Durable Goods	Consumption	6
52	Total Business Inventories	Consumption	6
53	Total Business: Inventories to Sales Ratio	Consumption	5
54	Consumer Sentiment Index	Consumption	6
55	M1 Money Stock	Money and credit	7
56	M2 Money Stock	Money and credit	6
57	Real M2 Money Stock	Money and credit	6
58	Total Reserves of Depository Institutions	Money and credit	6
59	Reserves Of Depository Institutions	Money and credit	2
60	Commercial and Industrial Loans	Money and credit	6
61	Real Estate Loans at All Commercial Banks	Money and credit	6
62	Total Nonrevolving Credit	Money and credit	6
63	Nonrevolving consumer credit to Personal Income	Money and credit	6
64	Consumer Motor Vehicle Loans Outstanding	Money and credit	6
65	Total Consumer Loans and Leases Outstanding	Money and credit	6
66	Securities in Bank Credit at All Commercial Banks	Money and credit	6
67	Crude Oil, spliced WTI and Cushing	Prices	6
68	PPI: Metals and metal products:	Prices	6
69	CPI : All Items	Prices	6
70	CPI : Apparel	Prices	6
71	CPI : Transportation	Prices	6
72	CPI : Medical Care	Prices	6

73	CPI : Commodities	Prices	2
74	CPI : Services	Prices	6
75	CPI : All Items Less Food	Prices	6
76	CPI : All items less medical care	Prices	6

Table 6: FRED MD monthly data subset. Definitions of t-codes are available in the primary data source.

Source: https://research.stlouisfed.org/econ/mccracken/fred-databases/

Weekly financial data

	Description	Category	T-code
1	1-mo. Nonfinancial commercial paper A2P2/AA credit spread	Credit	1
2	Moody's Baa corporate bond/10-yr Treasury yield spread	Credit	1
3	BofAML High Yield/Moody's Baa corporate bond yield spread	Credit	1
4	30-yr Jumbo/Conforming fixed rate mortgage spread	Credit	1
5	30-yr Conforming Mortgage/10-yr Treasury yield spread	Credit	1
6	10-yr Constant Maturity Treasury yield	Leverage	2
7	S&P 500 Financials/S&P 500 Price Index (Relative to 2-yr MA)	Leverage	5
8	S&P 500, S&P 500 mini, NASDAQ 100, NASDAQ mini Open Interest	Leverage	4
9	3-mo. Eurodollar, 10-yr/3-mo. swap, 2-yr and 10-yr Treasury Open Interest	Leverage	4
10	1-mo. Asset-backed/Financial commercial paper spread	Risk	1
11	BofAML Home Equity ABS/MBS yield spread	Risk	1
12	3-mo. Financial commercial paper/Treasury bill spread	Risk	1
13	Commercial Paper Outstanding	Risk	3
14	BofAML 3-5 yr AAA CMBS OAS spread	Risk	1
15	3-mo./1-wk AA Financial commercial paper spread	Risk	1
16	Treasury Repo Delivery Fails Rate	Risk	4
17	Agency Repo Delivery Failures Rate	Risk	4
18	Government Securities Repo Delivery Failures Rate	Risk	4
19	Agency MBS Repo Delivery Failures Rate	Risk	4

20	3-mo. Eurodollar spread (LIBID-Treasury)	Risk	1
21	On-the-run vs. Off-the-run 10-yr Treasury liquidity premium	Risk	1
22	Fed Funds/Overnight Treasury Repo rate spread	Risk	1
23	Fed Funds/Overnight Agency Repo rate spread	Risk	1
24	Fed Funds/Overnight MBS Repo rate spread	Risk	1
25	3-mo./1-wk Treasury Repo spread	Risk	1
26	10-yr/2-yr Treasury yield spread	Risk	1
27	2-yr/3-mo. Treasury yield spread	Risk	1
28	10-yr Interest Rate Swap/Treasury yield spread	Risk	1
29	2-yr Interest Rate Swap/Treasury yield spread	Risk	1
30	1-yr Interest Rate Swap/1-Year Treasury spread	Risk	1
31	3-mo. LIBOR/CME Term SOFR-Treasury spread	Risk	1
32	1-yr./1-mo. LIBOR/CME Term SOFR spread	Risk	1
33	Advanced Foreign Economies Trade-weighted US Dollar Value Index	Risk	3
34	CBOE Market Volatility Index VIX	Risk	1
35	1-mo. BofAML Option Volatility Estimate Index	Risk	1
36	3-mo. BofAML Swaption Volatility Estimate Index	Risk	1

Table 7: Financial weekly data set. Sources: Bloomberg & Haver Analytics. Definitions of t-codes are available on NFCI Chicago Fed website: https://www.chicagofed.org/research/data/nfci/current-data

Factors

Factor	Frequency	Source website
ADS	Weekly	https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/ads
CFNAI	Monthly	https://www.chicagofed.org/research/data/cfnai/historical-data
NFCI	Weekly	https://research.stlouisfed.org/econ/mccracken/fred-databases/

Table 8: Factors.

B Details on matrix completion

To implement matrix completion, we use the R package `softImpute` version 1.4—1 downloaded from CRAN. The algorithm fits a low-rank matrix approximation to a matrix with missing values via nuclear-norm regularization. We set the maximum number of rank, `max.rank`, to 6, which restricts the rank of the solution. Starting from λ_0, where λ is the regularization parameter for the nuclear norm minimization problem, we find λ so that the solution reached has rank slightly less than `rank.max`, as suggested in the package manual. λ_0 is the initial guess, which we set to a value computed by a function `lambda0` within the package. This function computes the smallest value for λ such that `softImpute` returns the zero solution.