Machine learning time series regressions with an application to nowcasting*

Andrii Babii† Eric Ghysels‡ Jonas Striaukas§

June 1, 2020

Abstract

This paper introduces structured machine learning regressions for high-dimensional time series data potentially sampled at different frequencies. The sparse-group LASSO estimator can take advantage of such time series data structures and outperforms the unstructured LASSO. We establish oracle inequalities for the sparse-group LASSO estimator within a framework that allows for the mixing processes and recognizes that the financial and the macroeconomic data may have heavier than exponential tails. An empirical application to nowcasting US GDP growth indicates that the estimator performs favorably compared to other alternatives and that the text data can be a useful addition to more traditional numerical data.

Keywords: high-dimensional time series, text data, mixed frequency data, sparse-group LASSO, Fuk-Nagaev inequality, tau-dependent processes.

*We thank participants at the Financial Econometrics Conference at the TSE Toulouse, the JRC Big Data and Forecasting Conference, the Big Data and Machine Learning in Econometrics, Finance, and Statistics Conference at the University of Chicago, the Nontraditional Data, Machine Learning, and Natural Language Processing in Macroeconomics Conference at the Board of Governors and the AI Innovations Forum organized by SAS and the Kenan-Flagler Business School as well as Jianqing Fan, Michele Lenza and Dacheng Xiu for comments. All remaining errors are ours.

†Department of Economics, University of North Carolina–Chapel Hill - Gardner Hall, CB 3305 Chapel Hill, NC 27599-3305. Email: babii.andrii@gmail.com

‡Department of Economics and Kenan-Flagler Business School, University of North Carolina–Chapel Hill. Email: eghysels@unc.edu.

§LIDAM UC Louvain and FRS–FNRS Research Fellow. Email: jonas.striaukas@gmail.com.

1 Introduction

The statistical imprecision inherent in the quarterly gross domestic product (GDP) estimates, together with the fact that even the first estimate is available with a delay of nearly a month, pose a significant challenge to policymakers and other observers with an interest in monitoring the state of the economy in real time.[1] A term originated in meteorology, nowcasting pertains to the prediction of the present and very near future. Nowcasting is intrinsically a mixed frequency data problem as the object of interest is a low-frequency data series - observed say quarterly like GDP - whereas real-time information - daily, weekly or monthly - during the quarter can be used to assess and potentially continuously update the state of the low-frequency series, or put differently, *nowcast* the series of interest. Traditional methods being used for nowcasting rely on dynamic factor models that treat the underlying low frequency series of interest as a latent process with high frequency data noisy observations. These models are naturally cast in a state-space form, and inference can be performed using standard Kalman filtering techniques.[2]

So far, nowcasting has mostly relied on so-called standard macroeconomic data releases. Perhaps the most prominent among these releases in the US is the Bureau of Labor Statistics Employment Situation report, which is issued on the first Friday of every month. This report includes data on payroll employment, unemployment, earnings, and many other aspects of the labor market. The nature of business cycles, in which most sectors of the economy tend to move together, implies that good news for the labor market – or for manufacturing, construction, retail trade, and so on – usually reflects good news for the economy as a whole. The Employment report releases are followed closely not just by economists, but by market participants, people in business, and the media. Besides labor market data, nowcasting models typically also rely on construction spending, (non-)manufacturing report, price indices data, etc, which we will call traditional macroeconomic data. One prominent example is produced by the Federal Reserve Bank of New York, using a dynamic factor model with thirty-seven predictors of different frequencies.[3]

[1] See e.g. Ghysels, Horan, and Moench (2018) for a recent discussion of macroeconomic data revision and publication delays.

[2] See Bańbura, Giannone, Modugno, and Reichlin (2013) for a recent survey.

[3] See Bok, Caratelli, Giannone, Sbordone, and Tambalotti (2018) for more details. The Federal Reserve Bank of New York Staff Nowcast framework is run and updated daily at 10 a.m. whenever new data releases are issued, and updates to the nowcast are published weekly every Friday at 11:15 a.m. on the New York Feds public website https://www.newyorkfed.org/research/policy/nowcast.

Thirty-seven predictors of traditional macroeconomic series may be viewed as small in comparison to the non-traditional series possibly available and useful. We quickly would reach numerical complexities involved with estimating high-dimensional state space models – making the approach computationally prohibitively complex, slow, or simply infeasible. Macroeconomists increasingly rely on non-standard data such as textual analysis via machine learning. This means potentially hundreds of series. For example, a textual analysis data set based on *Wall Street Journal* articles that has been recently made available features a taxonomy of 180 topics.[4] Which topics are relevant? How should they be selected? Thorsrud (2020) constructs a daily business cycle index based on quarterly GDP growth and textual information contained in a daily business newspaper, using a time-varying dynamic factor model where dynamic sparsity is enforced upon the factor loadings using a latent threshold mechanism. His work shows the feasibility of using variations of the traditional state space setting. Yet, the challenges grow when we start thinking about also adding potentially large-dimensional traditional data sets as well as non-traditional data such as for example payment systems information or GPS tracking data.[5]

We study nowcasting low-frequency series – focusing on the key example of US GDP growth – in a data-rich environment, where our data not only includes conventional high-frequency series but also non-standard data generated by textual analysis. The latter type of data is shown to be statistically significant using HAC-based inference based on Babii, Ghysels, and Striaukas (2020). Our nowcasts are superior to those posted by the Federal Reserve Bank of New York which involves proprietary information whereas our models exclusively rely on public domain data, and most importantly involve high dimensional non-conventional data sources. To deal with such massive non-traditional datasets we need to rely on a different approach, one involving machine learning methods dealing with data sampled at different frequencies.[6] We adopt a MIDAS (Mixed

[4]See Bybee, Kelly, Manela, and Xiu (2020) and the website http://structureofnews.com/.

[5]Studies for Canada (Galbraith and Tkacz (2018)), Denmark (Carlsen and Storgaard (2010)), India (Raju and Balakrishnan (2019)), Italy (Aprigliano, Ardizzi, and Monteforte (2019)), Portugal (Duarte, Rodrigues, and Rua (2017)), and the United States (Barnett, Chauvet, Leiva-Leon, and Su (2016)) find that payment transactions can help with nowcasting and with forecasting GDP and private consumption in the short term. Other related applications are Moriwaki (2019) nowcasting unemployment rates with smartphone GPS data, among others.

[6]Relatively little is known about handling high-dimensional mixed frequency data. Among the exceptions is Andreou, Gagliardini, Ghysels, and Rubin (2019) who study principal component analysis with large dimensional panels and focus on mixed frequency data.

Data Sampling) regression-based approach which is more amenable to large dimensional data environments. Our general framework includes standard (same frequency) time series regressions as well.

Several novel contributions are required to achieve our goal. First, we argue that the high-dimensional time series regressions involve certain data structures that once taken into account should improve the performance of unrestricted estimators in small samples. These structures are represented by groups covering lagged dependent variables and groups of lags for a single (high-frequency) covariate. To that end, we leverage on the sparse-group LASSO (sg-LASSO) regularization that accommodates conveniently such structures.[7] The attractive feature of the sg-LASSO estimator is that it allows us to combine effectively the approximately sparse and dense signals; see e.g., Carrasco and Rossi (2016) for a comprehensive treatment of ill-posed dense time series regressions.

We recognize that the economic and financial time series data are frequently heavy-tailed, while the bulk of the machine learning methods assumes i.i.d. data and/or exponential tails for covariates and regression errors; see Belloni, Chernozhukov, Chetverikov, Hansen, and Kato (2018) for a comprehensive review of high-dimensional regressions with i.i.d. data. There have been several recent attempts to expand the asymptotic theory to settings involving time series dependent data, mostly for the LASSO estimator. For instance, Kock and Callot (2015) establish oracle inequalities for the VAR with i.i.d. errors; Wong, Li, and Tewari (2019) consider β-mixing series with exponential tails; Wu and Wu (2016), Han and Tsay (2017), and Chernozhukov, Härdle, Huang, and Wang (2019) allow for polynomial tails under the functional dependence measure of Wu (2005).

Despite these efforts, there is no complete estimation theory for high-dimensional time series regressions under the assumptions comparable to the classical GMM and QML estimators. To the best of our knowledge, the high-dimensional *mixing processes* with *polynomial tails* have not been treated in the relevant literature. This paper fills this gap in the literature relying on the Fuk-Nagaev inequality for τ-dependent processes[8] recently obtained in Babii, Ghysels, and Striaukas

[7]The sparse-group LASSO was introduced by Simon, Friedman, Hastie, and Tibshirani (2013). The idea to apply group structures to time series covariates is novel. In contrast to *group LASSO*, the sparse-group LASSO promotes sparsity *between* and *within* groups (i.e. lags of time series covariates).

[8]τ-dependence coefficients are introduced in Dedecker and Prieur (2004) and Dedecker and Prieur (2005) as

(2020) and establishes the non-asymptotic and asymptotic estimation and prediction properties of the sg-LASSO estimator with dependent data. To the best of our knowledge, these results are new and our paper is the first to introduce τ-dependent processes in the context of the LASSO estimator. The Fuk-Nagaev inequality, cf., Fuk and Nagaev (1971), describes the concentration of sums of random variables with a mixture of the sub-Gaussian and the polynomial tails. It provides sharp estimates of tail probabilities unlike Markov's bound in conjunction with the MarcinkiewiczZygmund or Rosenthal's moment inequalities. Our results cover the LASSO and the group LASSO as special cases and, to the best of our knowledge, such treatment of other multi-penalty regularized estimators, e.g., the elastic net, is not currently available even in the i.i.d. case.

The rest of the paper is organized as follows. Section 2 presents the generic time series regression setting used in the paper. Section 3 characterizes non-asymptotic estimation and prediction accuracy of the sg-LASSO estimator for τ-dependent processes with polynomial tails. We report on a Monte Carlo study in Section 4 which provides further insights about the validity of our theoretical analysis in small sample settings typically encountered in empirical applications. Section 5 covers the empirical application. Conclusions appear in Section 6.

Notation: For a random variable $X \in \mathbf{R}$, let $\|X\|_q = (\mathbb{E}|X|^q)^{1/q}, q \geq 1$ be its L_q norm. For $p \in \mathbf{N}$, put $[p] = \{1, 2, \ldots, p\}$. For a vector $\Delta \in \mathbf{R}^p$ and a subset $J \subset [p]$, let Δ_J be a vector in $\mathbf{R}^p$ with the same coordinates as Δ on J and zero coordinates on J^c. Let $\mathcal{G} = \{G_g : \ g \geq 1\}$ be a partition of $[p]$ defining the group structure. For a vector $\beta \in \mathbf{R}^p$, the sparse-group structure is described by a pair $(S_0, \mathcal{G}_0)$, where $S_0 = \{j \in [p] : \ \beta_j \neq 0\}$ is the support of β and $\mathcal{G}_0 = \{G \in \mathcal{G} : \ \beta_G \neq 0\}$ is its group support. For $b \in \mathbf{R}^p$, its $\ell_q, q \geq 1$ norm is denoted $|b|_q = \left(\sum_{j\geq 1}^p |b_j|^q\right)^{1/q}, q < \infty$ and $|b|_\infty = \max_{1\leq j\leq p} |b_j|$. For $\mathbf{u}, \mathbf{v} \in \mathbf{R}^T$, the empirical inner product is defined as $\langle \mathbf{u}, \mathbf{v}\rangle_T = \frac{1}{T}\sum_{t=1}^T u_t v_t$ with the induced empirical norm $\|.\|_T^2 = \langle ., .\rangle_T = |.|_2^2/T$. For a symmetric $p \times p$ matrix A, let $\text{vech}(A) \in \mathbf{R}^{p(p+1)/2}$ be its vectorization consisting of the lower triangular and the diagonal part. For $a, b \in \mathbf{R}$, we put $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Lastly, we write $a_n \lesssim b_n$ if there exists a (sufficiently large) absolute constant C such that $a_n \leq C b_n$ for all $n \geq 1$ and $a_n \sim b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

weaker than mixing coefficients. Therefore, our results cover mixing processes, which is not the case for the physical dependence measures.

2 Time series regressions and sparse-group LASSO

Let $(y_t)_{t\in[T]}$ be the target series measured at discrete time points $t \in [T]$.[9] Predictions of y_t can involve its lags as well as a large set of covariates and lags thereof. In the interest of generality, but more importantly because of the empirical relevance we allow the covariates to be sampled at higher frequencies - with same frequency being a special case. More specifically, let there be K covariates $\{x_{t-j/m,k}, j \in [m], t \in [T], k \in [K]\}$ possibly measured at some higher frequency with m observations every t and consider the following regression model

$$\phi(L)y_t = \sum_{k=1}^{K} \psi(L^{1/m};\beta_k)x_{t,k} + u_t, \qquad t \in [T], \tag{1}$$

where $\phi(L) = I - \rho_1 L - \rho_2 L^2 - \cdots - \rho_J L^J$ is the low-frequency lag polynomial and $\psi(L^{1/m};\beta_k)x_{t,k} = 1/m \sum_{j=1}^{m} \beta_{j,k} x_{t-j/m,k}$ is the high-frequency lag polynomial. For $m = 1$, we have a standard autoregressive distributed lag (ARDL) model, which is the workhorse regression model of the time series econometrics literature.[10]

The ARDL-MIDAS model (using the terminology of Andreou, Ghysels, and Kourtellos (2013)) features $J + 1 + m \times K$ parameters. In the big data setting with a large number of covariates sampled at high-frequency, the total number of parameters may be large compared to the effective sample size or even exceed it. This leads to poor estimation and out-of-sample prediction accuracy in finite samples. For instance, with $m = 3$ (quarterly/monthly setting) and 35 covariates at 4 lagged quarters, we need to estimate $m \times K = 420$ parameters. At the same time, say post-WWII quarterly GDP growth series has less than 300 observations.

The LASSO estimator, see Tibshirani (1996), offers an appealing convex relaxation of a difficult non-convex best subset selection problem. By construction, it produces sparse parsimonious models zeroing-out a large number of the estimated parameters. The model selection is not free and comes at a price that can be high in the low signal-to-noise environment with heavy-tailed dependent data. In this paper, we focus on the structured sparsity with additional dimensionality reductions that aim to improve upon the unstructured LASSO estimator.

[9] For a natural number N, we denote $[N] = \{1, 2, \ldots, N\}$.

[10] Note that the polynomial $\psi(L^{1/m};\beta_k)x_{t,k}$ only involves m lags, which is done for the sake of simplicity and without the loss of generality. In addition, regression (1) involves high-frequency lags $t - j/m$. In some applications lags $t - 1 - j/m$ might be more suitable, or a combination of both, again without the loss of generality.

First, we parameterize the high-frequency lag polynomial following the MIDAS regression or the distributed lag econometric literatures (see Ghysels, Santa-Clara, and Valkanov (2006)) as

$$\psi(L^{1/m};\beta_k)x_{t,k} = \frac{1}{m}\sum_{j=1}^{m}\omega(j/m;\beta_k)x_{t-j/m,k}, \tag{2}$$

where $\dim(\beta_k) = L < m$. The weight function $\omega : [0,1] \times \mathbf{R}^L \to \mathbf{R}$ is approximated as

$$\omega(t;\beta_k) \approx \sum_{l=1}^{L}\beta_{k,l}w_l(t), \qquad t \in [0,1], \tag{3}$$

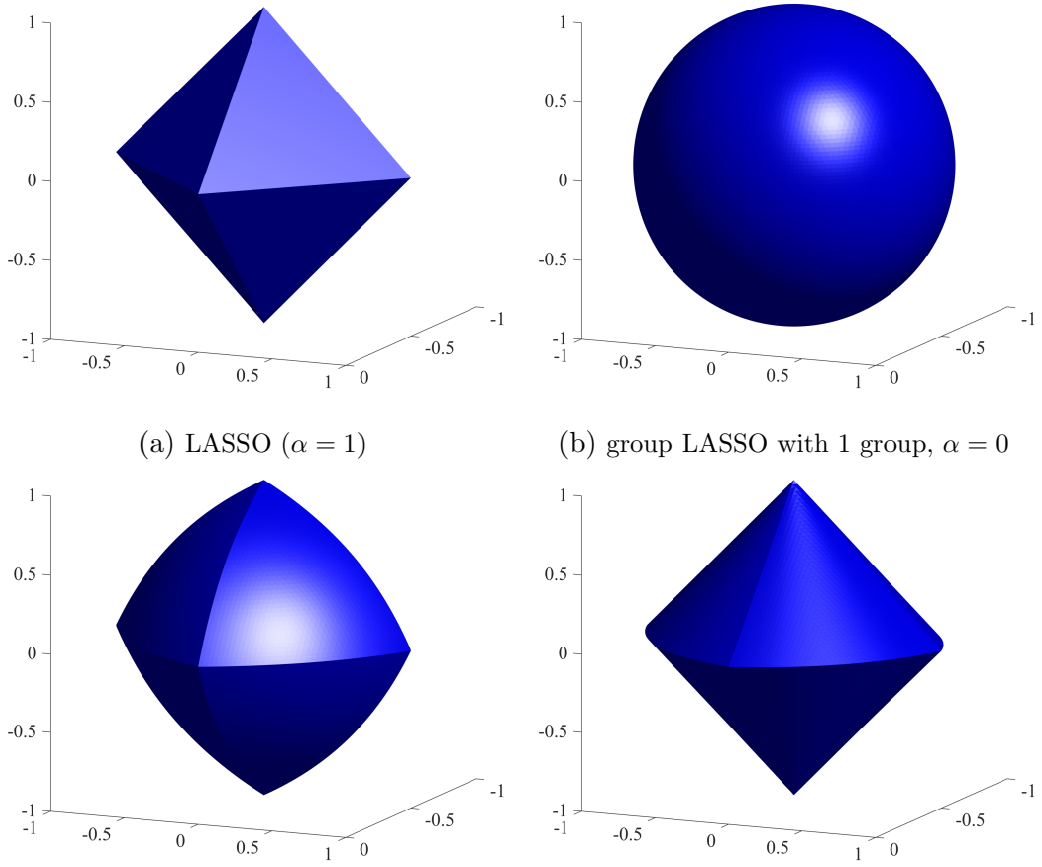
where $\{w_l : \ l = 1,\dots,L\}$ is a collection of functions, called the *dictionary*. The simplest example of the dictionary consists of algebraic power polynomials, also known as Almon (1965) polynomials in the time series analysis. More generally, the dictionary may consist of arbitrary approximating functions, including classical orthonormal bases.[11]

The size of the dictionary L and the number of covariates K can still be large and the *approximate sparsity* is a key assumption imposed throughout the paper. With the approximate sparsity, we recognize that assuming that most of the estimated coefficients are zero is overly restrictive and that the approximation error should be taken into account. For instance, the weight function may have an infinite series expansion, nonetheless, most can be captured by a relatively small number of orthogonal basis functions. Similarly, there can be a large number of economically relevant predictors, nonetheless, it might be sufficient to select only a smaller number of the most relevant ones to achieve good out-of-sample forecasting performance. Both model selection goals can be achieved with the LASSO estimator. However, the LASSO does not recognize that covariates at different (high-frequency) lags are temporally related.

In the baseline model, all high-frequency lags (or approximating functions once we parametrize the lag polynomial) of a single covariate constitute a group. We can also assemble all lag dependent variables into a group. Other group structures could be considered, for instance combining various covariates into a single group, but we will work with the simplest group setting of the aforementioned baseline model. The sparse-group LASSO (sg-LASSO), see Simon, Friedman, Hastie,

[11] See appendix section A.1 for more examples. Using orthogonal polynomials typically reduces the multicollinearity and leads to better finite sample performance. The specification in (2) deviates from the standard MIDAS polynomial specification and results in a linear regression model - a subtle but key innovation as it maps MIDAS regressions in the standard regression framework.

and Tibshirani (2013), allows us to incorporate such structure into the estimation procedure. In contrast to the group LASSO, see Yuan and Lin (2006), the sg-LASSO promotes sparsity *between* and *within* groups, and allows us to capture the predictive information from each group, such as approximating functions from the dictionary or specific covariates from each group.[12]



(a) LASSO ($\alpha = 1$)

(b) group LASSO with 1 group, $\alpha = 0$

(c) sg-LASSO with 1 group, $\alpha = 0.5$

(d) sg-LASSO with 2 groups, $\alpha = 0.5$

Figure 1: Geometry of $\{b \in \mathbf{R}^2 : \Omega(b) \leq 1\}$ for different groupings and values of α.

To describe the estimation procedure, let $\mathbf{y} = (y_1, \ldots, y_T)^\top$, be a vector of dependent variable and let $\mathbf{X} = (\iota, \mathbf{y}_1, \ldots, \mathbf{y}_J, Z_1 W, \ldots, Z_K W)$, be a design matrix, where $\iota = (1, 1, \ldots, 1)^\top$ is a vector of ones, $\mathbf{y}_j = (y_{1-j}, \ldots, y_{T-j})^\top$, $Z_k = (x_{k,t-j/m})_{t\in[T],j\in[m]}$ is a $T \times m$ matrix of the covariate

[12]Selecting the most important elements from the dictionary to approximate the MIDAS weights is superior to selecting, e.g., the polynomial of a fixed degree, see DeVore (1998) for the comparison between the linear and nonlinear approximation. It should also be noted that Marsilli (2014), and Uematsu and Tanaka (2019) are recent examples extending the MIDAS regression setting to a penalized regression setting. None of these existing papers provide an asymptotic theory supporting the proposed methods.

$k \in [K]$, and $W = \left(\frac{1}{m} w_l\left(j/m\right)\right)_{j\in[m],l\in[L]}$ is an $m \times L$ matrix of weights. In addition, put $\beta = (\beta_0^\top, \beta_1^\top, \ldots, \beta_K^\top)^\top$, where $\beta_0 = (\rho_0, \rho_1, \ldots, \rho_J)^\top$ is a vector of parameters pertaining to the group of autoregressive coefficients and $\beta_k \in \mathbf{R}^L$ denotes parameters of the high-frequency lag polynomial pertaining to the covariate $k \geq 1$. Then, the sparse-group LASSO estimator, denoted $\hat{\beta}$, solves the penalized least-squares problem

$$\min_{b\in\mathbf{R}^p} \|\mathbf{y} - \mathbf{X}b\|_T^2 + 2\lambda\Omega(b) \tag{4}$$

with a penalty function that interpolates between the ℓ_1 LASSO penalty and the ℓ_2 group LASSO penalty

$$\Omega(b) = \alpha|b|_1 + (1-\alpha)\|b\|_{2,1},$$

where $\|b\|_{2,1} = \sum_{G\in\mathcal{G}} |b_G|_2$ is the group LASSO norm.

The amount of penalization is controlled by the regularization parameter $\lambda > 0$ while $\alpha \in [0,1]$ is a weight parameter that determines the relative importance of the sparsity and the group structure. Setting $\alpha = 1$, we obtain the LASSO estimator while setting $\alpha = 0$, leads to the group LASSO estimator.[13] In practice, groups are defined by a particular problem, while α can be fixed or selected in a data-driven way. Figure 1 illustrates the geometry of Ω for different groupings and different values of α. The estimator can be computed efficiently using an appropriate coordinate descent algorithm, cf., Simon, Friedman, Hastie, and Tibshirani (2013).

3 Oracle inequalities and convergence rates

3.1 Dynamic regressions

We focus on the generic dynamic linear regression model that nests the ARDL-MIDAS regression as a special case

$$y_t = \mathbb{E}[y_t|\mathcal{F}_t] + u_t, \qquad \mathbb{E}[u_t|\mathcal{F}_t] = 0,$$

where $(y_t)_{t\in\mathbf{Z}}$ is a real-valued stochastic process and $(\mathcal{F}_t)_{t\in\mathbf{Z}}$ is a filtration. The filtration reflects the information set available at a particular point of time and is generated by a large number of covariates, lags of covariates, as well as lags of the dependent variable. We approximate the

[13] Note that with a single group, the penalty resembles the elastic net penalty with the only difference that we have $|.|_2$ instead of $|.|_2^2$, so that the sg-LASSO may achieve similar to the elastic net regularization goals.

conditional mean with its best linear approximation with respect to the L_2 norm, denoted $X_t^\top \beta$, where $(X_t)_{t\in\mathbf{Z}}$ is a stochastic process in $\mathbf{R}^p$ that may include some covariates, lags of covariates up to a certain order, as well as lags of the dependent variable.[14]

Using the setting of equation (4), in the vector notation, we write

$$\mathbf{y} = \mathbf{m} + \mathbf{u},$$

where $\mathbf{y} = (y_1, \dots, y_T)^\top$, $\mathbf{m} = (\mathbb{E}[y_1|\mathcal{F}_1], \dots, \mathbb{E}[y_T|\mathcal{F}_T])^\top$, and $\mathbf{u} = \mathbf{y} - \mathbf{m}$. The best linear approximation is denoted $\mathbf{X}\beta$, where $\mathbf{X}$ is a $T \times p$ design matrix and $\beta \in \mathbf{R}^p$ is a vector of unknown parameters.

We measure the time series dependence with τ-dependence coefficients. For a σ-algebra $\mathcal{M}$ and a random vector $\xi \in \mathbf{R}^l$, the τ coefficient is defined as

$$\tau(\mathcal{M}, \xi) = \sup_{f\in\Lambda(\mathbf{R}^l)} \int_{\mathbf{R}} \|F_{f(\xi)|\mathcal{M}}(t) - F_{f(\xi)}(t)\|_1 \mathrm{d}t,$$

where $\Lambda(\mathbf{R}^l) = \{f : \mathbf{R}^l \to \mathbf{R} : \ |f(x) - f(y)| \le |x - y|_2\}$ is a set of 1-Lipschitz functions, $F_{f(\xi)}$ is the CDF of $f(\xi)$, and $F_{f(\xi)|\mathcal{M}}$ is the CDF of $f(\xi)$ conditionally on $\mathcal{M}$. Let $(\xi_t)_{t\in\mathbf{Z}}$ be a stochastic process and let $\mathcal{M}_t = \sigma(\xi_t, \xi_{t-1}, \dots)$ be its natural filtration. The τ-dependence coefficient is defined as

$$\tau_k = \sup_{j\ge 1} \max_{1\le l\le j} \frac{1}{l} \sup_{t+k\le t_1<\dots<t_l} \tau(\mathcal{M}_t, (\xi_{t_1}, \dots, \xi_{t_l})), \qquad k \ge 0.$$

The process is called τ-dependent if its τ-dependence coefficients tend to zero. The τ-dependence coefficients were introduced in Dedecker and Prieur (2004) as dependence measures weaker than mixing;[15] see also Dedecker and Prieur (2005), Lemma 1 for an equivalent variational characterization.

3.2 Non-asymptotic bounds and convergence rates

In this section, we introduce main assumptions and study estimation and prediction properties of the sg-LASSO estimator. We restrict the class of stochastic processes below.

[14] Since dynamic time series regressions are fundamentally misspecified, our theoretical treatment allows for the misspecification. Note that in the correctly specified case the regression score is a martingale difference sequence which simplifies significantly the probabilistic treatment.

[15] Andrews (1984) constructs a simple example of an AR(1) process which is not mixing. Roughly speaking, τ-dependent processes are somewhere between mixingales and mixing processes and can accommodate such counterexamples; see Dedecker and Prieur (2005) and Dedecker and Doukhan (2003) for a more detailed comparison.

Assumption 3.1 (Data). *$(y_t, X_t)_{t\in\mathbf{Z}}$ is a stationary process such that (i) $\max_{j\in[p]} \|u_t X_{t,j}\|_q = O(1)$ for some $q > 2$; (ii) $\max_{j,k\in[p]} \|X_{t,j}X_{t,k}\|_{\tilde{q}} = O(1)$ for some $\tilde{q} > 2$; (iii) $(u_t X_t)_{t\in\mathbf{Z}}$ is a vector of τ-dependent processes with $\tau_k \leq ck^{-a}$ for some $a > (q-1)/(q-2)$; (iv) $(X_t X_t^\top)_{t\in\mathbf{Z}}$ is a matrix of τ-dependent processes with $\tilde{\tau}_k \leq \tilde{c}k^{-\tilde{a}}$ for some $\tilde{a} > (\tilde{q}-1)/(\tilde{q}-2)$.*

Assumption 3.1 imposes very mild moment and weak dependence restrictions on the data-generating process. It is worth mentioning that the stationarity is not essential and can be relaxed at the costs of introducing heavier notation. We require only $q > 2$ finite moments for stochastic processes and do not require sub-Gaussianity which could be particularly restrictive for financial and economic data. Note also that we do not require that the dependence fades away exponentially fast.

We also need an appropriate restricted eigenvalue condition. For the support S_0 and the group support $\mathcal{G}_0$ of β, put[16]

$$\Omega_0(b) \triangleq \alpha|b_{S_0}|_1 + (1-\alpha)\sum_{G\in\mathcal{G}_0} |b_G|_2 \quad \text{and} \quad \Omega_1(b) \triangleq \alpha|b_{S_0^c}|_1 + (1-\alpha)\sum_{G\in\mathcal{G}_0^c} |b_G|_2$$

and consider the following cone $\mathcal{C}(c_0) = \{\Delta \in \mathbf{R}^p : \ \Omega_1(\Delta) \leq c_0\Omega_0(\Delta)\}$ for some $c_0 > 0$.

Assumption 3.2 (Restricted eigenvalue). *There exists a universal constant $\gamma > 0$ such that $|\Sigma^{1/2}\Delta|_2 \geq \gamma\sqrt{\sum_{G\in\mathcal{G}_0} |\Delta_G|_2^2}$ for all $\Delta \in \mathcal{C}(c_0)$, where $c_0 = \frac{c+1}{c-1}$ for some $c > 1$.*

Assumption 3.2 is a population counterpart to the frequently used restricted eigenvalue or compatibility condition imposed on the sample covariance matrix. It is trivially satisfied whenever the smallest eigenvalue of the population covariance matrix $\Sigma = \mathbb{E}[\mathbf{X}^\top\mathbf{X}/T]$ is bounded away from zero.[17] In econometric literature, the one-to-one property of Σ is also known as the completeness condition. Interestingly, sparse vectors can be identified and accurately estimated even when the covariance matrix is singular. The only requirement is that Σ is well-behaved on the cone $\mathcal{C}(c_0)$; see Babii and Florens (2020) for a related discussion in the context of ill-posed econometric models. The regularization parameter is determined by the Fuk-Nagaev concentration inequality, appearing in the Appendix equation (A.1).

[16]Note that the sg-LASSO penalty is not decomposable with respect to the support or the group support of β and the results from the theory of decomposable regularizers are not directly applicable, see, e.g., Negahban, Ravikumar, Wainwright, and Yu (2012).

[17]We call loosely Σ the covariance matrix as it coincides with the covariance matrix of a zero-mean random vector $X \in \mathbf{R}^p$. Likewise, we refer to $\hat{\Sigma} = \mathbf{X}^\top\mathbf{X}/T$ as the sample covariance matrix.

Assumption 3.3 (Regularization parameter)**.** *The regularization parameter satisfies*

$$\lambda \sim \left(\frac{p}{\delta T^{\kappa-1}}\right)^{1/\kappa} \vee \sqrt{\frac{\log(8p/\delta)}{T}},$$

for some $\delta \in (0,1)$ *and* $\kappa = \frac{(a+1)q-1}{a+q-1}$*, where* a, q *are as in Assumption 3.1.*

The regularization parameter depends on the temporal dependence, measured by a and heaviness of tails, measured by q. This dependence is reflected in the *dependence-tails exponent* κ. Under maintained assumptions, we obtain the following bound on the estimation and prediction accuracy of the sg-LASSO estimator, see Appendix A.2 for the proof.

Theorem 3.1. *Suppose that Assumptions 3.1, 3.2, and 3.3 are satisfied. Then there exists* $A_1, A_2 > 0$ *such that with probability at least* $1 - \delta - A_1 \frac{s_\alpha^{\tilde{\kappa}} p^2}{T^{\tilde{\kappa}-1}} - 2p(p+1)e^{-A_2 T/s_\alpha^2}$

$$\|\mathbf{X}(\hat{\beta} - \beta)\|_T^2 \lesssim s_\alpha \lambda^2 + \|\mathbf{m} - \mathbf{X}\beta\|_T^2$$

and

$$\Omega(\hat{\beta} - \beta) \lesssim s_\alpha \lambda + \lambda^{-1}\|\mathbf{m} - \mathbf{X}\beta\|_T^2 + s_\alpha^{1/2}\|\mathbf{m} - \mathbf{X}\beta\|_T,$$

where $s_\alpha^{1/2} = \alpha\sqrt{|S_0|} + (1-\alpha)\sqrt{|\mathcal{G}_0|}$ *and* $\tilde{\kappa} = \frac{(\tilde{a}+1)\tilde{q}-1}{\tilde{a}+\tilde{q}-1}$.

In the special case of the LASSO estimator, $\alpha = 1$, we obtain the counterpart to the result of Belloni, Chen, Chernozhukov, and Hansen (2012) for the LASSO with i.i.d. data that takes into account the approximation error. For another extreme $\alpha = 0$, we obtain non-asymptotic bounds for the group LASSO reflecting the approximation error. We call the constant s_α the *effective sparsity*. The effective sparsity is a linear combination of sparsity and group sparsity constants with weights defined by the penalty function.

An immediate consequence of the bounds stated in Theorem 3.1 is the asymptotic guarantee for the sg-LASSO estimator presented in the following corollary which we state under the assumption that the approximation error is negligible and the dimension/sparsity increase at a certain rate.

Assumption 3.4. *Suppose that (i)* $\|\mathbf{m} - \mathbf{X}\beta\|_T^2 = O_P(s_\alpha \lambda^2)$*; (ii)* $\frac{s_\alpha^{\tilde{\kappa}} p^2}{T^{\tilde{\kappa}-1}} \to 0$ *and* $p^2 e^{-A_2 T/s_\alpha^2} \to 0$.

Corollary 3.1. *Suppose that Assumptions 3.1, 3.2, 3.3, and 3.4 are satisfied. Then*

$$\|\mathbf{X}(\hat{\beta} - \beta)\|_T^2 = O_P\left(\frac{s_\alpha p^{2/\kappa}}{T^{2-2/\kappa}} \vee \frac{s_\alpha \log p}{T}\right).$$

and

$$\Omega(\hat{\beta} - \beta) = O_P\left(\frac{s_\alpha p^{1/\kappa}}{T^{1-1/\kappa}} \vee s_\alpha \sqrt{\frac{\log p}{T}}\right).$$

If the effective sparsity constant is fixed, then $p = o(T^{\kappa-1})$ is a sufficient condition for the prediction error and the Ω-norm error to converge to zero, whenever $\tilde{\kappa} \geq 2\kappa - 1$. In this case Assumption 3.4 (ii) is vacuous. Convergence rates reflect a trade-off between tails, dependence, and the number of covariates. The number of covariates p can increase at a faster rate than the sample size, provided that $\kappa > 2$, which is not the case for the classical OLS, ridge regression, and PCR estimators that require $p/T \to 0$.

Remark 3.1. *Since the ℓ_1-norm is equivalent to the Ω-norm whenever groups have fixed size ℓ_1-norm convergence rate is the same.*

Remark 3.2. *For a fixed sparsity constant, in the special case of the LASSO estimator with independent data, Caner and Kock (2018) obtain the convergence rate of order $O_P\left(\frac{p^{1/q}}{T^{1/2}}\right)$. Since $\kappa \to q$ as $a \to \infty$, we recover the $O_P\left(\frac{p^{1/q}}{T^{1-1/q}} \vee \sqrt{\frac{\log p}{T}}\right)$ convergence rate that one would obtain in the case of independent data applying directly the Fuk and Nagaev (1971), Corollary 4, whence we conclude that the dependence on q is optimal. Furthermore, increasing q, the polynomial term can be made arbitrarily small compared to the sub-Gaussian term. Therefore, the Fuk-Nagaev inequality provides a more accurate description of the performance of the LASSO estimator for the financial and the economic time series data that are often believed to have heavier than sub-Gaussian tails.*[18]

4 Monte Carlo experiments

In this section, we aim to assess the out-of-sample predictive performance (forecasting and nowcasting), and the MIDAS weights recovery of the sg-LASSO with dictionaries. We benchmark the performance of our novel sg-LASSO setup against two alternatives: (a) unstructured, meaning standard, LASSO with MIDAS and (b) unstructured LASSO with unrestricted lag polynomial. The former allows us to assess the benefits of exploiting group structures, whereas the latter focuses on the advantages of using dictionaries in a high dimensional setting.

[18] Recall that the sub-Gaussianity requires that moments of all order $q \geq 1$ exist. There exist alternative approaches to high-dimensional regressions with i.i.d. data exhibiting heavier than Gaussian tails based on using the loss function different from the MSE; see Fan, Li, and Wang (2017) for the estimator based on the Huber loss and Lecué and Lerasle (2019) for the median-of-means estimator. The investigation of such alternative approaches is beyond the scope of the present paper.

4.1 Simulation Design

To assess the predictive performance and the MIDAS weight recovery, we simulate the data from the following DGP:

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \sum_{k=1}^{K} \frac{1}{m} \sum_{j=1}^{m} \omega(j/m; \beta_k) x_{t-j/m,k} + u_t$$

where $u_t \sim_{i.i.d.} N(0, \sigma_u^2)$ and the DGP for covariates $\{x_{k,t-j/m} : k = 1, \ldots, K\}$ is specified below. This corresponds to a target of interest y_t driven by two autoregressive lags augmented with high frequency series, hence, the DGP is an ARDL-MIDAS model. We set $\sigma_u^2 = 1$, $\rho_1 = 0.3$, $\rho_2 = 0.01$, and take the number of relevant high frequency regressors $K = 3$. In some scenarios we also decrease the signal-to-noise ratio setting $\sigma_u^2 = 5$. We are interested in quarterly/monthly data, and use four quarters of data for the high frequency regressors so that $m = 12$. We rely on a commonly used weighting scheme in the MIDAS literature, namely $\omega(s; \beta_k)$ for $k = 1$, 2 and 3 are determined by beta densities respectively equal to $\text{Beta}(1, 3), \text{Beta}(2, 3)$, and $\text{Beta}(2, 2)$; see Ghysels, Sinko, and Valkanov (2007) or Ghysels and Qian (2019), for further details.

The high frequency regressors are generated as either one of the following:

1. K i.i.d. realizations of the univariate autoregressive (AR) process $x_h = \rho\, x_{h-1} + \varepsilon_h$, where $\rho = 0.2$ and either $\varepsilon_h \sim_{i.i.d.} N(0, \sigma_\varepsilon^2)$, $\sigma_\varepsilon^2 = 5$, or $\varepsilon_h \sim_{i.i.d.}$ student-$t(5)$, where h denotes the high-frequency sampling.

2. Multivariate vector autoregressive (VAR) process $X_h = \Phi\, X_{h-1} + \varepsilon_h$, where $\varepsilon_h \sim_{i.i.d.} N(0, I_K)$.

The latter creates contemporaneously correlated high frequency regressors.[19] In the estimation procedure, we add 7 noisy covariates which are generated in the same way as the relevant covariates and use 5 low-frequency lags. The empirical models use a dictionary which consists of Legendre polynomials up to degree $L = 10$ shifted to $[0, 1]$ interval with $\omega(s; \beta_k)$ defined in equation (3). The sample size is $T \in \{50, 100, 200\}$. Throughout the experiment, we use 5000 simulation replications and 10-fold cross-validation to select the tuning parameter.

[19] In the AR case, we initiate the two processes from $x_0 \sim N\left(0, \frac{\sigma^2}{1-\rho^2}\right)$, $y_0 \sim N\left(0, \frac{\sigma^2(1-\rho_2)}{(1+\rho_2)((1-\rho_2)^2-\rho_1^2)}\right)$. In the VAR case, we use the same initial value for (y_t) and initiate $X_0 \sim N(0, I_K)$. For all cases, the first 200 observations are treated as burn-in.

We assess the performance of different methods by varying assumptions on error terms of the high-frequency process ε_h, considering multivariate high-frequency process, changing the degree of Legendre polynomials L, increasing the noise level of the low-frequency process σ_u, using only half of the high-frequency lags in predictive regressions, and adding a larger number of noisy covariates. In the case of VAR high-frequency process, we set Φ to be block-diagonal with the first 5×5 block having entries 0.15 and the remaining 5×5 block(s) having entries 0.075.

We estimate three different LASSO-type regression models. In the first model we keep the weighting function unconstrained, therefore, we estimate 12 coefficients per high-frequency covariate using the unstructured LASSO estimator. We denote this model LASSO-U-MIDAS (inspired by the U-MIDAS of Foroni, Marcellino, and Schumacher (2015a)). In the second model we use MIDAS weights together with unstructured LASSO estimator; we call this model LASSO-MIDAS. In this case, we estimate $L+1$ number of coefficients per high-frequency covariate. The third model applies sg-LASSO estimator together with MIDAS weights. Groups are defined as in Section 2; each low-frequency lag and high-frequency covariate is a group, therefore, we have K+5 number of groups. We set the relative weight α to 0.65. This model is denoted sg-LASSO-MIDAS.

For regressions with aggregated data, we consider: (a) Flow aggregation (FLOW): $x_{k,t}^A = \frac{1}{m}\sum_{j=1}^{m} w_k x_{k,t-j/m}$, (b) Stock aggregation (STOCK): $x_{k,t}^A = x_{k,t}$, and (c) Single high-frequency lag (MIDDLE): $x_{k,t-(m-1)/m}$. In these cases, the models are estimated using the OLS estimator.

4.2 Simulation results

Detailed results are reported in the Appendix. Tables A.1–A.2, cover the average mean squared forecast errors for one-steahead forecasts and nowcasts. The sg-LASSO with MIDAS weighting (sg-LASSO-MIDAS) outperforms all other methods in all simulation scenarios. Importantly, both sg-LASSO-MIDAS and unstructured LASSO-MIDAS with non-linear weight function approximation perform much better than all other methods in most of the scenarios when the sample size is small ($T = 50$). In this case, sg-LASSO-MIDAS yields the largest improvements over alternatives, in particular, with a large number of noisy covariates (bottom-right block). The LASSO without MIDAS weighting has typically large forecast errors. The method performs better when half of the high-frequency lags are included in the regression model. Lastly, forecasts using flow-aggregated covariates seem to perform better than other simple aggregation methods in all simulation scenarios,

but significantly worse than the sg-LASSO-MIDAS.

In Table A.3–A.4 we report additional results for the estimation accuracy of the weight functions. In Figure A.1–A.3, we plot the estimated weight functions from several methods. The results indicate that the LASSO without MIDAS weighting can not recover accurately weights in small samples and/or low signal-to-noise ratio. Using Legendre polynomials improves the performance substantially and the sg-LASSO seems to improve even more over the unstructured LASSO.

5 Nowcasting US GDP with textual news data

In this section we nowcast US GDP with macroeconomic, financial, and textual news data. The data used in our empirical analysis are described in Appendix Section A.4. For standard macro variables, we use a real-time FRED-MD monthly dataset. The data is available at the Federal Reserve Bank of St. Louis FRED database, see McCracken and Ng (2016) for more details on this dataset. For our main results, we use a subset of all available macro covariates which we list in Table A.5.[20] Next, we add data from the Survey of Professional Forecasters, namely, US GDP nowcasts and forecasts for several horizons, which we aggregate using Legendre polynomials. In addition, we augment predictive regression with news attention data based on textual analysis that has been recently made available by Bybee, Kelly, Manela, and Xiu (2020). Finally, we follow the literature on nowcasting real GDP and define our target variable to be the annualized growth rate. To measure forecast errors, we take 2019 February real GDP data vintage.

5.1 Models

Denote $x_{t,k}$ the k-th high-frequency covariate at time t. The general ARDL-MIDAS predictive regression is

$$\phi(L)y_{t+1} = \mu + \sum_{k=1}^{K} \psi(L^{1/m}; \beta_k)x_{t,k} + u_{t+1}, \qquad t = 1, \dots, T,$$

[20] Additional set of results for the full set of FRED-MD monthly covariates with detailed implementation description is available in Appendix Section A.5.

where $\phi(L)$ is the low-frequency lag polynomial, μ is the regression intercept and $\sum_{k=1}^{K} \psi(L^{1/m}; \beta_k) x_{tk}$ are high-frequency covariates. As discussed in Section 2, we parameterize the weight function as

$$\psi(L^{1/m}; \beta_k) x_{t,k} = \frac{1}{m} \sum_{j=1}^{m} \omega(j/m; \beta_k) x_{t+(h+1-j)/m,k},$$

where h indicates the number of leading months in the quarter t. For example, if $h = 2$, we shift high-frequency covariates two month in the quarter, and hence we nowcast the dependent variable one month ahead.

We benchmark our predictions with the simple random walk (RW) model, which is considered to be a reasonable benchmark for short-term GDP growth predictions. We focus on predictions of our method, sg-LASSO-MIDAS, with and without series based on textual analysis. One natural comparison is with Federal Reserve Bank of New York, denoted New York Fed, model implied nowcasts.[21]

5.2 Nowcasting results

Table 1 reports nowcasting results for US GDP growth rate real-time at one given instance, namely two months into a quarter (or put differently with one month left into the quarter). First, we observe from the table that the sg-LASSO-MIDAS model with standard macro information improves upon the New York Fed predictions in terms of smaller out-of-sample root mean squared errors - although the margin is slim reducing from .790 (ratio with respect to RW) to .761. Without text-based information, the improvement of sg-LASSO-MIDAS versus New York Fed nowcasts is therefore, not surprisingly, insignificant based on the Diebold and Mariano (1995) test statistic. We report similar findings using the full dataset of covariates in Appendix Section A.5, where we also compare alternative machine learning methods with the sg-LASSO-MIDAS method and find that the latter outperforms all other alternatives.

Turning to results using additional text-based covariates, we see a significant improvement in terms of the quality of out-of-sample predictions. Relative to New York Fed nowcasts, sg-LASSO-MIDAS with textual data decrease prediction errors by 19%. The gain is also large relative to the sg-LASSO-MIDAS model that does not condition on news attention information, albeit slightly smaller.The Diebold and Mariano (1995) test statistic reveals that the increase in prediction

[21] We downloaded the data from https://www.newyorkfed.org/research/policy/nowcast.

accuracy is significant at 10% significance level when compared with New York Fed predictions as well as sg-LASSO-MIDAS model without textual data.

	Rel-RMSE	DM-stat-1	DM-stat-2
RW	2.606	3.624	3.629
sg-LASSO-MIDAS (with textual data)	0.639	-1.687	
sg-LASSO-MIDAS (without)	0.761		1.687
NY Fed	0.790	0.490	1.727

Table 1: Nowcast real GDP comparison table – Forecast horizon is one month ahead. Column *Rel-RMSE* reports root mean squared forecasts error relative to the RW model. Column *DM-stat-1* reports the Diebold and Mariano (1995) test statistic for all models relative to sg-LASSO-MIDAS model without text-based information, while column *DM-stat-2* reports the Diebold Mariano test statistic relative to sg-LASSO-MIDAS model with text-based information. The out-of-sample period is 2002 Q1 to 2017 Q2.

In Figure 2, we plot a heat map of selected covariates through time for the sg-LASSO-MIDAS model which includes news attention data.[22] In addition to the heat map, which reveals sparsity patterns, we also plot the evolution of the number of selected covariates and the squared forecast errors across time. In general, the pattern is relatively sparse, with more covariates being selected after the Great Recession. Specifically, on average 14.93 covariates are selected before, while 19.63 after the crisis, and 17.58 for the entire out-of-sample exercise. Interestingly, after the crises the number of selected covariates is more stable - fourteen covariates are always selected. Three covariates are always selected throughout the out-of-sample period: Government budgets, All Employees: Financial Activities, and 3-Month AA Financial Commercial Paper Rate.[23]

In Figure 3, we plot the cumulative sum of loss differential (cumsfe), which is computed as

$$\text{cumsfe}_{t,t+k} = \sum_{q=t}^{t+k} e^2_{q,M1} - e^2_{q,M2} \tag{5}$$

for model $M1$ versus $M2$. Positive value of $\text{cumsfe}_{t,t+k}$ means model $M1$ has larger squared forecast errors compared with model $M2$ up to $t+k$ point, and negative value imply the opposite. In our case, $M1$ is the New York Fed prediction error, and $M2$ is the sg-LASSO-MIDAS model either with or without news attention series. The plot in Figure 3 reveals that predictions based on sg-LASSO-MIDAS with or without news attention data yield smaller squared errors compared with

[22]Figure A.5 in the Appendix is a similar plot for the full-sample results using only macro data.

[23]Note that without the news data, see Figure A.4, autoregressive lags are selected more often.

Figure 2: Sparsity pattern.

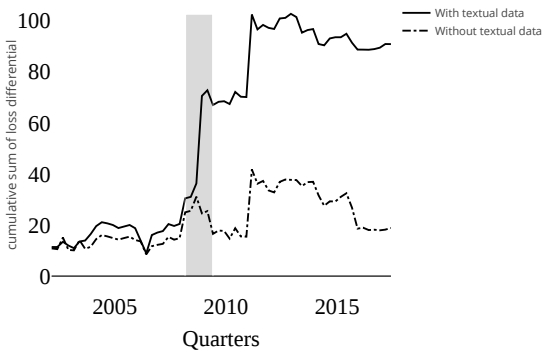


Figure 3: Cumulative sum of loss differential. Gray shaded area — NBER recession period.

New York Fed throughout the out-of-sample period. Interestingly, the largest gains are during the 2008-2009 recession period and at the beginning of 2011, which is around the period of the peak of European sovereign debt crises. The figure also shows marked improvements in prediction quality when using news attention series; notably, the largest gains are the two crises periods.[24]

5.3 Significance test

In this section, we test whether news attention series are significant predictors of real GDP growth rate using the inferential methods developed in Babii, Ghysels, and Striaukas (2020). We estimate the same sg-LASSO-MIDAS model using real-time macro data and news attention series. As in nowcasting application, we use 4 quarters of lagged data; the effective sample starts from 1985 February and ends in 2017 May and the sample size is 126 quarters. We select the tuning parameter by using 10-fold cross-validation and set $\alpha = 0.65$. For real-time macro data, we take the 2017 May FRED-MD vintage and use real GDP values as of May 30^{th}. To compute the precision matrix, we use nodewise LASSO regressions with a data-driven choice for the penalty parameter, see Babii, Ghysels, and Striaukas (2020) for more details. Since news attention series are high-frequency, we test the restriction that all coefficients associated with each series are jointly zero.

Table 2 reports values of Wald test for each series where we use the HAC estimator proposed by Babii, Ghysels, and Striaukas (2020) with Parzen kernel.[25] We report results for a grid of lag

[24] As an aside, using the *cumsfe* plots, we also show in the Appendix that the choice for $\alpha = 0.65$ is optimal, and yields the largest gains versus New York Fed nowcasts, see Figure A.6. The same figure also shows that favoring sparsity over group sparsity, i.e. $\alpha \in (0.5, 1]$, in general, improves predictions.

[25] In the Appendix Table A.8 we report the test statistic values.

truncation parameter M_T values. Results indicate that Government budgets and Oil market are highly significant predictors. Government budgets is significant at 1%, while Oil market at 5% significance level for all truncation parameter values. In the nowcasting application, the former was always selected throughout the out-of-sample period, while the latter was always selected after the crisis.

M_T	10	20	30
Commodities	0.533	0.514	0.554
Government budgets	0.002	0.009	0.008
Oil market	0.024	0.044	0.017
Recession	0.211	0.317	0.388
Savings & loans	0.754	0.685	0.655
Mortgages	0.750	0.604	0.553

Table 2: Significance test table – values for news attention series based on Wald test for a set of truncation parameter M_T values are reported. The number of lags in the HAC estimator correspond to M_T.

6 Conclusion

This paper offers a new perspective on the high-dimensional time series regressions with data sampled at the same or mixed frequencies and contributes more broadly to the rapidly growing literature on estimation, inference, forecasting, and nowcasting with regularized machine learning methods. The first contribution of the paper is to introduce the sparse-group LASSO estimator for high-dimensional time series regressions. An attractive feature of the estimator is that it recognizes time series data structures and allows us to perform the hierarchical model selection within and between groups. The classical LASSO and the group LASSO are covered as special cases.

To recognize that the economic and financial time series have typically heavier than Gaussian tails, we use a new Fuk-Nagaev concentration inequality, introduced in Babii, Ghysels, and Striaukas (2020), valid for a large class of τ-dependent processes, including mixing processes commonly used in econometrics. Building on this inequality, we establish non-asymptotic and asymptotic properties of the sparse-group LASSO estimator.

Our empirical application provides new perspectives on applying machine learning methods to real-time forecasting, nowcasting and monitoring using time series data, including non-conventional

data, sampled at different frequencies. To that end, we introduce a new class of MIDAS regressions with dictionaries linear in the parameters and based on orthogonal polynomials with lag selection using the sg-LASSO estimator. We find that the sg-LASSO estimator outperforms the unstructured LASSO in small samples and conclude that incorporating specific data structures should be helpful in various applications.

References

Almon, S. (1965): "The distributed lag between capital appropriations and expenditures," *Econometrica*, 33(1), 178–196.

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

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

Andrews, D. W. (1984): "Non-strong mixing autoregressive processes," *Journal of Applied Probability*, 21(4), 930–934.

Aprigliano, V., G. Ardizzi, and L. Monteforte (2019): "Using Payment System Data to Forecast Economic Activity," *International Journal of Central Banking*, 15(4), 55–80.

Babii, A., and J.-P. Florens (2020): "Is completeness necessary? Estimation in nonidentified linear models," arXiv preprint arXiv:1709.03473v3.

Babii, A., E. Ghysels, and J. Striaukas (2020): "Inference for high-dimensional regressions with heteroskedasticity and autocorrelation," arXiv preprint arXiv:1912.06307v2.

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

Barnett, W., M. Chauvet, D. Leiva-Leon, and L. Su (2016): "Nowcasting Nominal GDP with the Credit-Card Augmented Divisia Monetary Aggregates," .

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 80(6), 2369–2429.

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2018): "High-dimensional econometrics and generalized GMM," arXiv preprint arXiv:1806.01888.

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

Bybee, L., B. T. Kelly, A. Manela, and D. Xiu (2020): "The structure of economic news," *National Bureau of Economic Research*, and http://structureofnews.com.

Caner, M., and A. B. Kock (2018): "Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso," *Journal of Econometrics*, 203(1), 143–168.

Carlsen, M., and P. E. Storgaard (2010): "Dankort payments as a timely indicator of retail sales in Denmark," .

Carrasco, M., and B. Rossi (2016): "In-sample inference and forecasting in misspecified factor models," *Journal of Business and Economic Statistics*, 34(3), 313–338.

Chernozhukov, V., W. K. Härdle, C. Huang, and W. Wang (2019): "Lasso-driven inference in time and space," *Annals of Statistics* (forthcoming).

Dedecker, J., and P. Doukhan (2003): "A new covariance inequality and applications," *Stochastic Processes and their Applications*, 106(1), 63–80.

Dedecker, J., and C. Prieur (2004): "Coupling for τ-dependent sequences and applications," *Journal of Theoretical Probability*, 17(4), 861–885.

——— (2005): "New dependence coefficients. Examples and applications to statistics," *Probability Theory and Related Fields*, 132(2), 203–236.

DeVore, R. A. (1998): "Nonlinear approximation," *Acta Numerica*, 7, 51–150.

Diebold, F. X., and R. S. Mariano (1995): "Comparing predictive accuracy," *Journal of Business and Economic Statistics*, 13(3), 253–263.

Duarte, C., P. M. Rodrigues, and A. Rua (2017): "A mixed frequency approach to the forecasting of private consumption with ATM/POS data," *International Journal of Forecasting*, 33(1), 61–75.

Fan, J., Q. Li, and Y. Wang (2017): "Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1), 247–265.

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

——— (2015b): "Unrestricted mixed data sampling (U-MIDAS): MIDAS regressions with unrestricted lag polynomials," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1), 57–82.

Fuk, D. K., and S. V. Nagaev (1971): "Probability inequalities for sums of independent random variables," *Theory of Probability and Its Applications*, 16(4), 643–660.

Galbraith, J. W., and G. Tkacz (2018): "Nowcasting with payments system data," *International Journal of Forecasting*, 34(2), 366–376.

Ghysels, E., C. Horan, and E. Moench (2018): "Forecasting through the Rearview Mirror: Data Revisions and Bond Return Predictability.," *Review of Financial Studies*, 31(2), 678–714.

Ghysels, E., and H. Qian (2019): "Estimating MIDAS regressions via OLS with polynomial parameter profiling," *Econometrics and Statistics*, 9, 1–16.

Ghysels, E., P. Santa-Clara, and R. Valkanov (2006): "Predicting volatility: getting the most out of return data sampled at different frequencies," *Journal of Econometrics*, 131, 59–95.

Ghysels, E., A. Sinko, and R. Valkanov (2007): "MIDAS regressions: Further results and new directions," *Econometric Reviews*, 26(1), 53–90.

Han, Y., and R. S. Tsay (2017): "High-dimensional Linear Regression for Dependent Observations with Application to Nowcasting," *arXiv preprint arXiv:1706.07899*.

Kock, A. B., and L. Callot (2015): "Oracle inequalities for high dimensional vector autoregressions," *Journal of Econometrics*, 186(2), 325–344.

Lecué, G., and M. Lerasle (2019): "Robust machine learning by median-of-means: theory and practice," *Annals of Statistics* (forthcoming).

Marsilli, C. (2014): "Variable Selection in Predictive MIDAS Models," Working papers 520, Banque de France.

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

Moriwaki, D. (2019): "Nowcasting Unemployment Rates with Smartphone GPS Data," in *International Workshop on Multiple-Aspect Analysis of Semantic Trajectories*, pp. 21–33. Springer.

Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu (2012): "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, 27(4), 538–557.

Raju, S., and M. Balakrishnan (2019): "Nowcasting economic activity in India using payment systems data," *Journal of Payments Strategy and Systems*, 13(1), 72–81.

Siliverstovs, B. (2017): "Short-term forecasting with mixed-frequency data: a MIDASSO approach," *Applied Economics*, 49(13), 1326–1343.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013): "A sparse-group LASSO," *Journal of Computational and Graphical Statistics*, 22(2), 231–245.

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

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

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

Wong, K. C., Z. Li, and A. Tewari (2019): "LASSO guarantees for β-mixing heavy tailed time series," *Annals of Statistics* (forthcoming).

Wu, W. B. (2005): "Nonlinear system theory: Another look at dependence," *Proceedings of the National Academy of Sciences*, 102(40), 14150–14154.

Wu, W.-B., and Y. N. Wu (2016): "Performance bounds for parameter estimates of high-dimensional linear models with correlated errors," *Electronic Journal of Statistics*, 10(1), 352–379.

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

Appendix

A.1 Dictionaries

In this section we review briefly the choice of dictionaries for the MIDAS weight function. It is possible to construct dictionaries using arbitrary sets of functions, including a mix of algebraic polynomials, trigonometric polynomials, B-splines, Haar basis, or wavelets. In this paper, we mostly focus on dictionaries generated by orthogonalized algebraic polynomials, though it might be interesting to tailor the dictionary for each particular application. The attractiveness of algebraic polynomials comes from their ability to generate a variety of shapes with a relatively low number of parameters, which is especially desirable in the low signal-to-noise environments. The general family of appropriate orthogonal algebraic polynomials is given by Jacobi polynomials that nest Legendre, Gegenbauer, and Chebychev's polynomials as a special case.

Example A.1.1 (Jacobi polynomials)**.** *Applying the Gram-Schmidt orthogonalization to* $\{1, x, x^2, x^3, \dots\}$ *with respect to the measure*

$$\mathrm{d}\mu(x) = (1-x)^{\alpha}(1+x)^{\beta}\mathrm{d}x, \qquad \alpha, \beta > -1,$$

on $[-1, 1]$*, we obtain Jacobi polynomials. In practice Jacobi polynomials can be computed through the well-known tree-term recurrence relation for* $n \geq 0$

$$P_{n+1}^{(\alpha,\beta)}(x) = axP_n^{(\alpha,\beta)}(x) + bP_n^{(\alpha,\beta)}(x) - cP_{n-1}^{(\alpha,\beta)}(x)$$

with $a = \frac{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)}{2(n+1)(n+\alpha+\beta+1)}$*,* $b = \frac{(2n+\alpha+\beta+1)(\alpha^2-\beta^2)}{2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)}$*,* $c = \frac{(\alpha+n)(\beta+n)(2n+\alpha+\beta+2)}{(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)}$*. To obtain the orthogonal basis on* $[0, 1]$*, we shift Jacobi polynomials with affine bijection* $x \mapsto 2x - 1$*.*

For $\alpha = \beta$*, we obtain Gegenbauer polynomials, for* $\alpha = \beta = 0$*, we obtain Legendre polynomials, while for* $\alpha = \beta = -1/2$ *or* $\alpha = \beta = 1/2$*, we obtain Chebychev's polynomials of two kinds.*

In the mixed frequency setting, non-orthogonalized polynomials, $\{1, x, x^2, x^3, \dots\}$, are also called Almon polynomials. It is preferable to use orthogonal polynomials in practice due to reduced multicolinearity and better numerical properties. At the same time, orthogonal polynomials are available in Matlab, R, Python, and Julia packages. Legendre polynomials is our default recommendation, while other choices of α and β are preferable if we want to accommodate heavier tails.

We noted in the main body of the paper that the specification in (2) deviates from the standard MIDAS polynomial specification as it results in a linear regression model - a subtle but key innovation as it maps MIDAS regressions in the standard regression framework. Moreover, casting the MIDAS regressions in a linear regression framework renders the optimization problem convex, something only achieved by Siliverstovs (2017) using the U-MIDAS of Foroni, Marcellino, and Schumacher (2015b) which does not recognize the mixed frequency data structure, unlike our sg-LASSO.

A.2 Proofs of main results

Lemma A.2.1. *Consider* $\|.\| = \alpha|.|_1 + (1-\alpha)|.|_2$, *where* $|.|_q$ *is* ℓ_q *norm on* $\mathbf{R}^p$. *Then the dual norm of* $\|.\|$, *denoted* $\|.\|^*$, *satisfies*

$$\|z\|^* \leq \alpha|z|_1^* + (1-\alpha)|z|_2^*, \qquad \forall z \in \mathbf{R}^p,$$

where $|.|_1^*$ *is the dual norm of* $|.|_1$ *and* $|.|_2^*$ *is the dual norm of* $|.|_2$.

Proof. Clearly, $\|.\|$ is a norm. By the convexity of $x \mapsto x^{-1}$ on $(0,\infty)$

$$\begin{aligned}
\|z\|^* &= \sup_{b\neq 0} \frac{|\langle z,b\rangle|}{\|b\|} \leq \sup_{b\neq 0}\left\{\alpha\frac{|\langle z,b\rangle|}{|b|_1} + (1-\alpha)\frac{|\langle z,b\rangle|}{|b|_2}\right\} \\
&\leq \alpha \sup_{b\neq 0}\frac{|\langle z,b\rangle|}{|b|_1} + (1-\alpha)\sup_{b\neq 0}\frac{|\langle z,b\rangle|}{|b|_2} \\
&= \alpha|z|_1^* + (1-\alpha)|z|_2^*.
\end{aligned}$$

□

Proof of Theorem 3.1. Note that the sg-LASSO penalty Ω is a norm. By Lemma A.2.1, its dual norm satisfies

$$\begin{aligned}
\Omega^*(\mathbf{X}^\top\mathbf{u}/T) &\leq \alpha|\mathbf{X}^\top\mathbf{u}/T|_\infty + (1-\alpha)\max_{G\in\mathcal{G}}|(\mathbf{X}^\top\mathbf{u})_G/T|_2 \\
&\lesssim |\mathbf{X}^\top\mathbf{u}/T|_\infty \\
&\lesssim \left(\frac{p}{\delta T^{\kappa-1}}\right)^{1/\kappa} \vee \sqrt{\frac{\log(8p/\delta)}{T}} \\
&\lesssim \lambda,
\end{aligned} \tag{A.1}$$

where the first inequality follows since $|z|_1^* = |z|_\infty$ and $\left(\sum_{G\in\mathcal{G}}|z_G|_2\right)^* = \max_{G\in\mathcal{G}}|z_G|_2$, the second by elementary computations, the third under Assumptions 3.1 (i) and (iii), by Theorem A.1 with

probability at least $1-\delta$, and the last from the definition of λ in Assumption 3.3. By Fermat's rule, the sg-LASSO satisfies

$$\mathbf{X}^\top(\mathbf{X}\hat{\beta}-\mathbf{y})/T+\lambda z^*=0$$

for some $z^*\in\partial\Omega(\hat{\beta})$, where $\partial\Omega(\hat{\beta})$ is the subdifferential of $b\mapsto\Omega(b)$ at $\hat{\beta}$. Taking the inner product with $\beta-\hat{\beta}$

$$\langle\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\hat{\beta}),\beta-\hat{\beta}\rangle_T=\lambda\langle z^*,\beta-\hat{\beta}\rangle\leq\lambda\left\{\Omega(\beta)-\Omega(\hat{\beta})\right\},$$

where the inequality follows from the definition of the subdifferential. Using $\mathbf{y}=\mathbf{m}+\mathbf{u}$ and rearranging this inequality

$$\begin{aligned}\|\mathbf{X}(\hat{\beta}-\beta)\|_T^2-\lambda\left\{\Omega(\beta)-\Omega(\hat{\beta})\right\}&\leq\langle\mathbf{X}^\top\mathbf{u},\hat{\beta}-\beta\rangle_T+\langle\mathbf{X}^\top(\mathbf{m}-\mathbf{X}\beta),\hat{\beta}-\beta\rangle_T\\&\leq\Omega^*\left(\mathbf{X}^\top\mathbf{u}/T\right)\Omega(\hat{\beta}-\beta)+\|\mathbf{X}(\hat{\beta}-\beta)\|_T\|\mathbf{m}-\mathbf{X}\beta\|_T\\&\leq c^{-1}\lambda\Omega(\hat{\beta}-\beta)+\|\mathbf{X}(\hat{\beta}-\beta)\|_T\|\mathbf{m}-\mathbf{X}\beta\|_T.\end{aligned}$$

where the second line follows by the dual norm inequality and the last by $\Omega^*(\mathbf{X}^\top\mathbf{u}/T)\leq c^{-1}\lambda$ for some $c>1$ as shown in Eq. A.1. Therefore,

$$\begin{aligned}\|\mathbf{X}\Delta\|_T^2&\leq c^{-1}\lambda\Omega(\Delta)+\|\mathbf{X}\Delta\|_T\|\mathbf{m}-\mathbf{X}\beta\|_T+\lambda\left\{\Omega(\beta)-\Omega(\hat{\beta})\right\}\\&\leq(c^{-1}+1)\lambda\Omega(\Delta)+\|\mathbf{X}\Delta\|_T\|\mathbf{m}-\mathbf{X}\beta\|_T\end{aligned}\tag{A.2}$$

with $\Delta=\hat{\beta}-\beta$. Note that the sg-LASSO penalty can be decomposed as a sum of two seminorms $\Omega(b)=\Omega_0(b)+\Omega_1(b),\ \forall b\in\mathbf{R}^p$ with

$$\Omega_0(b)=\alpha|b_{S_0}|_1+(1-\alpha)\sum_{G\in\mathcal{G}_0}|b_G|_2\quad\text{and}\quad\Omega_1(b)=\alpha|b_{S_0^c}|_1+(1-\alpha)\sum_{G\in\mathcal{G}_0^c}|b_G|_2.$$

Note also that $\Omega_1(\beta)=0$ and $\Omega_1(\hat{\beta})=\Omega_1(\Delta)$. Then by the triangle inequality

$$\Omega(\beta)-\Omega(\hat{\beta})\leq\Omega_0(\Delta)-\Omega_1(\Delta).\tag{A.3}$$

If $\|\mathbf{m}-\mathbf{X}\beta\|_T\leq\frac{1}{2}\|\mathbf{X}\Delta\|_T$, then it follows from the first inequality in Eq. A.2 and Eq. A.3 that

$$\|\mathbf{X}\Delta\|_T^2\leq 2c^{-1}\lambda\Omega(\Delta)+2\lambda\left\{\Omega_0(\Delta)-\Omega_1(\Delta)\right\}.$$

Since the left side of this equation is positive, this shows that $\Omega_1(\Delta)\leq c_0\Omega_0(\Delta)$ with $c_0=\frac{c+1}{c-1}$, and

whence $\Delta \in \mathcal{C}(c_0)$, cf., Assumption 3.2. Then

$$\begin{aligned}\Omega(\Delta) &\leq (1+c_0)\Omega_0(\Delta) \\ &\leq (1+c_0)\left(\alpha\sqrt{|S_0|}|\Delta_{S_0}|_2 + (1-\alpha)\sqrt{|\mathcal{G}_0|}\sqrt{\sum_{G\in\mathcal{G}_0}|\Delta_G|_2^2}\right) \\ &\leq (1+c_0)s_\alpha^{1/2}\sqrt{\sum_{G\in\mathcal{G}_0}|\Delta_G|_2^2} \\ &\leq (1+c_0)s_\alpha^{1/2}\gamma^{-1}|\Sigma^{1/2}\Delta|_2,\end{aligned} \tag{A.4}$$

where the second line follows by Jensen's inequality and the last under Assumption 3.2 with $s_\alpha^{1/2} = \alpha\sqrt{|S_0|} + (1-\alpha)\sqrt{|\mathcal{G}_0|}$. Next, put $\bar{G} = \max_{G\in\mathcal{G}}|G|$, where $|G|$ is the cardinality of the group $G \subset [p]$, and note that

$$\begin{aligned}|\Sigma^{1/2}\Delta|_2^2 &= \Delta^\top\Sigma\Delta \\ &= \|\mathbf{X}\Delta\|_T^2 + \Delta^\top(\Sigma - \hat{\Sigma})\Delta \\ &= 2(c^{-1}+1)\lambda\Omega(\Delta) + \Omega(\Delta)\Omega^*\left((\hat{\Sigma}-\Sigma)\Delta\right) \\ &\leq 2(c^{-1}+1)\lambda\Omega(\Delta) + \Omega^2(\Delta)\bar{G}|\mathrm{vech}(\hat{\Sigma}-\Sigma)|_\infty,\end{aligned}$$

where the third inequality follows by the inequality in Eq. A.2 and the dual norm inequality, and the fourth by Lemma A.2.1 and elementary computations

$$\begin{aligned}\Omega^*\left((\hat{\Sigma}-\Sigma)\Delta\right) &\leq \alpha|(\hat{\Sigma}-\Sigma)\Delta|_\infty + (1-\alpha)\max_{0\leq k\leq K}|[(\hat{\Sigma}-\Sigma)\Delta]_{G_k}|_2 \\ &\leq \alpha|\Delta|_1|\mathrm{vech}(\hat{\Sigma}-\Sigma)|_\infty + (1-\alpha)\bar{G}^{1/2}|\mathrm{vech}(\hat{\Sigma}-\Sigma)|_\infty|\Delta|_1 \\ &\leq \bar{G}|\mathrm{vech}(\hat{\Sigma}-\Sigma)|_\infty\Omega(\Delta).\end{aligned}$$

Combining these computations with the inequality in Eq. A.4

$$\begin{aligned}\Omega(\Delta) &\leq (1+c_0)^2\gamma^{-2}s_\alpha\left\{2(c^{-1}+1)\lambda + \bar{G}|\mathrm{vech}(\hat{\Sigma}-\Sigma)|_\infty\Omega(\Delta)\right\} \\ &\leq 2(1+c_0)^2\gamma^{-2}s_\alpha(c^{-1}+1)\lambda + (1-A^{-1})\Omega(\Delta),\end{aligned}$$

where the second line holds on the event $E \triangleq \left\{|\mathrm{vech}(\hat{\Sigma}-\Sigma)|_\infty \leq \frac{\gamma^2}{2\bar{G}s_\alpha(1+2c_0)^2}\right\}$ with $1 - A^{-1} = \frac{(1+c_0)^2}{2(1+2c_0)^2} < 1$. This observation in conjunction with the inequality in Eq. A.2 gives

$$\begin{aligned}\Omega(\Delta) &\leq 2A(1+c_0)^2 s_\alpha(c^{-1}+1)\lambda \\ \|\mathbf{X}\Delta\|_T^2 &\leq 4A(1+c_0)^2 s_\alpha(c^{-1}+1)^2\lambda^2.\end{aligned}$$

On the other hand, if $\|\mathbf{m}-\mathbf{X}\beta\|_T > \frac{1}{2}\|\mathbf{X}\Delta\|_T$, then

$$\|\mathbf{X}\Delta\|_T^2 \leq 4\|\mathbf{m}-\mathbf{X}\beta\|_T^2.$$

Therefore, on E, we always have

$$\|\mathbf{X}\Delta\|_T^2 \leq C_1 s_\alpha \lambda^2 + 4\|\mathbf{m}-\mathbf{X}\beta\|_T^2 \tag{A.5}$$

with $C_1 = 4A(1+c_0)^2(c^{-1}+1)^2$, which proves the first claim of Theorem 3.1.

For the second claim, suppose first that $\Delta \in \mathcal{C}(2c_0)$. Then on E

$$\begin{aligned}
\Omega^2(\Delta) &\leq (1+2c_0)^2 s_\alpha |\Sigma^{1/2}\Delta|_2^2 \\
&= (1+2c_0)^2 s_\alpha \left\{\|\mathbf{X}\Delta\|_T^2 + \Delta^\top(\Sigma-\hat{\Sigma})\Delta\right\} \\
&\leq (1+2c_0)^2 s_\alpha \left\{C_1 s_\alpha \lambda^2 + 4\|\mathbf{m}-\mathbf{X}\beta\|_T^2 + \Omega^2(\Delta)\bar{G}|\mathrm{vech}(\hat{\Sigma}-\Sigma)|_\infty\right\} \\
&\leq (1+2c_0)^2 s_\alpha \left\{C_1 s_\alpha \lambda^2 + 4\|\mathbf{m}-\mathbf{X}\beta\|_T^2\right\} + 2^{-1}\Omega^2(\Delta),
\end{aligned}$$

where the first inequality follows by computations similar to Eq. A.4 and the second inequality from Eq. A.5. Therefore,

$$\Omega^2(\Delta) \leq 2(1+2c_0)^2 s_\alpha \left\{C_1 s_\alpha \lambda^2 + 4\|\mathbf{m}-\mathbf{X}\beta\|_T^2\right\}. \tag{A.6}$$

On the other hand, if $\Delta \notin \mathcal{C}(2c_0)$, then $\Delta \notin \mathcal{C}(c_0)$, which as we have already shown implies $\|\mathbf{m}-\mathbf{X}\beta\|_T > \frac{1}{2}\|\mathbf{X}\Delta\|_T$. In conjunction with Eq. A.2 and Eq. A.3, this shows that

$$0 \leq \lambda c^{-1}\Omega(\Delta) + 2\|\mathbf{m}-\mathbf{X}\beta\|_T^2 + \lambda\left\{\Omega_0(\Delta) - \Omega_1(\Delta)\right\},$$

and whence

$$\begin{aligned}
\Omega_1(\Delta) &\leq c_0\Omega_0(\Delta) + \frac{2c}{\lambda(c-1)}\|\mathbf{m}-\mathbf{X}\beta\|_T^2 \\
&\leq 2^{-1}\Omega_1(\Delta) + \frac{2c}{\lambda(c-1)}\|\mathbf{m}-\mathbf{X}\beta\|_T^2.
\end{aligned}$$

This shows that $\Omega(\Delta) \leq (1+(2c_0)^{-1})\Omega_1(\Delta) \leq (1+(2c_0)^{-1})\frac{4c}{\lambda(c-1)}\|\mathbf{m}-\mathbf{X}\beta\|_T^2 = \lambda^{-1}\frac{2c(3c-1)}{c^2-1}\|\mathbf{m}-\mathbf{X}\beta\|_T^2$. Combining this with the inequality in Eq. A.6, we obtain the second claim of Theorem 3.1.

Lastly, under Assumptions 3.1 (ii) and (iv), by Theorem 3.1 in Babii, Ghysels, and Striaukas (2020)

$$\begin{aligned}
\Pr(E^c) &= \Pr\left(|\mathrm{vech}(\hat{\Sigma}-\Sigma)|_\infty > \frac{\gamma^2}{2\bar{G}s_\alpha(1+2c_0)^2}\right) \\
&\leq \frac{A_1 s_\alpha^{\tilde{\kappa}} p^2}{T^{\tilde{\kappa}-1}} + 2p(p+1)\exp\left(-\frac{c_2 T^2}{s_\alpha^2 B_T^2}\right)
\end{aligned}$$

for some universal constants A_1 and c_2 and $B_T^2 = \max_{j,k\in[p]} \sum_{t=1}^T \sum_{l=1}^T |\mathrm{Cov}(X_{t,j}X_{t,k}, X_{l,j}X_{l,k})|$. Lastly, under Assumptions 3.1 (ii) and (iv), by Babii, Ghysels, and Striaukas (2020), Lemma A.1.2 that $B_T^2 = O(T)$. □

The following result is proven in Babii, Ghysels, and Striaukas (2020), see their Theorem 3.1 and Eq. (4) following it.

Theorem A.1. *Let $(\xi_t)_{t\in\mathbf{Z}}$ be a centered stationary stochastic process in $\mathbf{R}^p$ such that (i) $\max_{j\in[p]} \|\xi_{0,j}\|_q = O(1)$ for some $q > 2$; (ii) for every $j \in [p]$, τ-dependence coefficients of $\xi_{t,j}$ satisfy $\tau_k^{(j)} \leq ck^{-a}$ for some universal constants $c > 0$ and $a > \frac{q-1}{q-2}$. Then there exists $C > 0$ such that for every $\delta \in (0,1)$*

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^T \xi_t\right|_\infty \leq C\left(\frac{p}{\delta T^{\kappa-1}}\right)^{1/\kappa} \vee \sqrt{\frac{\log(8p/\delta)}{T}}\right) \geq 1-\delta. \tag{A.7}$$

A.3 Monte Carlo Simulations

	FLOW	STOCK	MIDDLE	LASSO-U	LASSO-M	SGL-M	FLOW	STOCK	MIDDLE	LASSO-U	LASSO-M	SGL-M
T			Baseline scenario						$\varepsilon_h \sim_{i.i.d.}$ student-$t(5)$			
50	2.847	3.839	4.660	4.213	2.561	2.188	2.081	2.427	2.702	2.334	2.066	1.749
	0.059	0.077	0.090	0.087	0.054	0.044	0.042	0.053	0.062	0.056	0.051	0.041
100	2.110	2.912	3.814	2.244	1.579	1.473	1.504	1.900	2.155	1.761	1.535	1.343
	0.041	0.057	0.076	0.045	0.032	0.030	0.030	0.038	0.043	0.034	0.030	0.026
200	1.882	2.772	3.681	1.539	1.302	1.230	1.357	1.714	1.986	1.414	1.238	1.192
	0.037	0.056	0.072	0.031	0.026	0.025	0.027	0.035	0.040	0.029	0.025	0.024
			High-frequency process: VAR(1)						Legendre degree $L = 3$			
50	1.869	2.645	2.863	2.135	1.726	1.533	2.847	3.839	4.660	4.213	2.339	1.979
	0.039	0.053	0.057	0.046	0.036	0.032	0.059	0.077	0.090	0.087	0.050	0.041
100	1.453	2.073	2.245	1.575	1.373	1.284	2.110	2.912	3.814	2.244	1.503	1.386
	0.028	0.042	0.046	0.031	0.028	0.025	0.041	0.057	0.076	0.045	0.031	0.029
200	1.283	1.921	2.040	1.348	1.240	1.201	1.882	2.772	3.681	1.539	1.277	1.196
	0.026	0.038	0.041	0.026	0.024	0.023	0.037	0.056	0.072	0.031	0.025	0.024
			Legendre degree $L = 10$						Low frequency noise level σ_u^2=5			
50	2.847	3.839	4.660	4.213	2.983	2.583	9.598	10.429	10.726	9.799	8.732	7.785
	0.059	0.077	0.090	0.087	0.063	0.053	0.196	0.211	0.213	0.198	0.180	0.159
100	2.110	2.912	3.814	2.244	1.719	1.633	7.319	8.177	8.880	8.928	7.359	6.606
	0.041	0.057	0.076	0.045	0.035	0.032	0.147	0.163	0.176	0.179	0.147	0.135
200	1.882	2.772	3.681	1.539	1.348	1.300	6.489	7.699	8.381	7.275	6.391	5.919
	0.037	0.056	0.072	0.031	0.027	0.026	0.127	0.154	0.165	0.146	0.126	0.117
			Half high-frequency lags						Number of covariates $p = 50$			
50	2.750	2.730	3.562	2.455	2.344	1.905				5.189	3.610	2.658
	0.058	0.056	0.070	0.050	0.048	0.038				0.104	0.075	0.054
100	2.134	2.167	3.082	1.899	1.718	1.468	5.582	5.633	6.298	3.527	2.034	1.753
	0.043	0.043	0.061	0.038	0.034	0.030	0.117	0.113	0.126	0.075	0.042	0.036
200	1.833	1.971	2.808	1.400	1.356	1.225	2.679	3.573	4.399	1.867	1.413	1.319
	0.036	0.039	0.055	0.028	0.027	0.024	0.053	0.071	0.090	0.038	0.028	0.026

Table A.1: Forecasting accuracy results – See Table A.2

	FLOW	STOCK	MIDDLE	LASSO-U	LASSO-M	SGL-M	FLOW	STOCK	MIDDLE	LASSO-U	LASSO-M	SGL-M
T			Baseline scenario						$\varepsilon_h \sim_{i.i.d.}$ student-$t(5)$			
50	3.095	3.793	4.659	4.622	3.196	2.646	2.257	2.391	2.649	2.357	2.131	1.786
	0.067	0.078	0.094	0.094	0.064	0.055	0.046	0.054	0.057	0.050	0.047	0.038
100	2.393	2.948	3.860	2.805	2.113	1.888	1.598	1.840	2.068	1.824	1.653	1.433
	0.048	0.060	0.078	0.058	0.044	0.038	0.032	0.037	0.043	0.036	0.033	0.029
200	2.122	2.682	3.597	1.971	1.712	1.604	1.452	1.690	1.969	1.544	1.383	1.302
	0.042	0.055	0.072	0.039	0.034	0.032	0.030	0.035	0.041	0.032	0.028	0.026
			High-frequency process: VAR(1)						Legendre degree $L = 3$			
50	2.086	2.418	2.856	2.208	1.828	1.612	3.095	3.793	4.659	4.622	2.987	2.451
	0.044	0.050	0.057	0.049	0.039	0.033	0.067	0.078	0.094	0.094	0.061	0.050
100	1.571	1.906	2.341	1.671	1.430	1.329	2.393	2.948	3.860	2.805	2.020	1.796
	0.031	0.039	0.047	0.033	0.028	0.026	0.048	0.060	0.078	0.058	0.042	0.037
200	1.397	1.720	2.168	1.428	1.307	1.248	2.122	2.682	3.597	1.971	1.680	1.560
	0.028	0.034	0.043	0.028	0.026	0.024	0.042	0.055	0.072	0.039	0.033	0.031
			Legendre degree $L = 10$						Low frequency noise level σ_u^2=5			
50	3.095	3.793	4.659	4.622	3.528	2.948	9.934	10.566	10.921	9.819	9.037	8.091
	0.067	0.078	0.094	0.094	0.071	0.059	0.213	0.212	0.216	0.198	0.184	0.168
100	2.393	2.948	3.860	2.805	2.271	2.079	7.576	8.130	8.854	9.190	7.743	6.876
	0.048	0.060	0.078	0.058	0.047	0.042	0.150	0.166	0.180	0.188	0.160	0.141
200	2.122	2.682	3.597	1.971	1.777	1.693	6.830	7.580	8.351	7.648	6.820	6.258
	0.042	0.055	0.072	0.039	0.035	0.034	0.135	0.152	0.168	0.156	0.136	0.124
			Half high-frequency lags						Number of covariates $p = 50$			
50	3.014	2.773	3.638	2.455	2.509	2.201				5.222	3.919	3.002
	0.063	0.056	0.072	0.050	0.051	0.046				0.105	0.081	0.061
100	2.344	2.087	3.116	1.899	2.101	1.774	5.978	5.556	6.536	3.948	2.665	2.232
	0.046	0.041	0.063	0.038	0.043	0.036	0.121	0.112	0.132	0.083	0.053	0.044
200	2.119	1.985	2.988	1.400	1.761	1.590	2.974	3.422	4.412	2.355	1.938	1.725
	0.041	0.040	0.061	0.028	0.035	0.032	0.059	0.070	0.087	0.048	0.040	0.035

Table A.2: Nowcasting accuracy results

The table reports simulation results for nowcasting accuracy. We report eight different scenarios for the DGP: baseline scenario (upper-left block) DGP is with the low-frequency noise level $\sigma_u^2 = 1$ which we keep for all other scenarios except where we change it to $\sigma_u^2 = 5$, the degree of Legendre polynomial L=5, andGaussian high-frequency error term. All remaining blocks report results for different DGPs: e.g. in the upper-right block, we report results where the noise term of high-frequency covariates is i.i.d. student-$t(5)$. Each block reports results for LASSO-U-MIDAS (LASSO-U), LASSO-MIDAS (LASSO-M) and sg-LASSO-MIDAS (SGL-M) estimators (the last three columns). In addition, we report results for predictive regressions using aggregated data where we use different aggregation schemes: 1) flow aggregation (FLOW), stock aggregation (STOCK) and taking the middle value of high-frequency covariates (MIDDLE). We vary the sample size T from 50 to 200. Each entry in the odd row is the average mean squared forecast error, and each entry in the even row is the simulation standard error.

	LASSO-U	LASSO-M	SGL-M	LASSO-U	LASSO-M	SGL-M	LASSO-U	LASSO-M	SGL-M
	T=50			T=100			T=200		
				Baseline scenario					
Beta$(1,3)$	1.955	0.887	0.652	1.846	0.287	0.247	1.804	0.138	0.106
	0.002	0.012	0.010	0.002	0.004	0.004	0.001	0.002	0.002
Beta$(2,3)$	1.211	0.739	0.625	1.157	0.351	0.268	1.128	0.199	0.118
	0.001	0.008	0.008	0.001	0.004	0.004	0.001	0.002	0.002
Beta$(2,2)$	1.062	0.593	0.537	1.019	0.231	0.216	0.995	0.106	0.092
	0.001	0.007	0.007	0.001	0.003	0.003	0.001	0.001	0.001
				$\varepsilon_h \sim_{i.i.d.}$ student-$t(5)$					
Beta$(1,3)$	2.005	1.688	1.290	1.953	1.064	0.624	1.885	0.471	0.401
	0.002	0.012	0.014	0.002	0.011	0.009	0.002	0.005	0.005
Beta$(2,2)$	1.237	1.126	0.993	1.218	0.848	0.614	1.185	0.506	0.440
	0.001	0.007	0.010	0.001	0.007	0.007	0.001	0.005	0.004
Beta$(2,2)$	1.084	0.969	0.874	1.070	0.691	0.518	1.047	0.369	0.356
	0.001	0.006	0.008	0.001	0.006	0.006	0.001	0.004	0.004
				high-frequency process: VAR(1)					
Beta$(1,3)$	1.935	1.271	0.939	1.890	0.772	0.492	1.842	0.419	0.288
	0.003	0.016	0.015	0.002	0.010	0.008	0.002	0.005	0.004
Beta$(2,3)$	1.177	0.864	0.811	1.155	0.610	0.505	1.136	0.468	0.359
	0.002	0.011	0.012	0.002	0.008	0.008	0.001	0.005	0.005
Beta$(2,2)$	1.036	0.706	0.729	1.023	0.477	0.458	1.008	0.326	0.299
	0.002	0.009	0.011	0.002	0.007	0.007	0.001	0.004	0.004
				Legendre degree $L = 3$					
Beta$(1,3)$	1.955	0.727	0.484	1.846	0.248	0.178	1.804	0.123	0.081
	0.002	0.010	0.008	0.002	0.004	0.003	0.001	0.002	0.001
Beta$(2,3)$	1.211	0.642	0.491	1.157	0.313	0.201	1.128	0.181	0.094
	0.001	0.008	0.007	0.001	0.004	0.003	0.001	0.002	0.001
Beta$(2,2)$	1.062	0.508	0.414	1.019	0.200	0.156	0.995	0.094	0.069
	0.001	0.007	0.006	0.001	0.003	0.003	0.001	0.001	0.001

Table A.3: Shape of weights estimation accuracy I.

The table reports results for shape of weights estimation accuracy for the first four DGPs of Table A.1-A.2 using LASSO-U, LASSO-M and SGL-M estimators for the weight functions Beta$(1,3)$, Beta$(2,3)$ and Beta$(2,2)$ with sample size $T = 50$, 100 and 200. Entries in odd rows are the average point-wise mean squared error, and in even rows the simulation standard error.

	LASSO-U	LASSO-M	SGL-M	LASSO-U	LASSO-M	SGL-M	LASSO-U	LASSO-M	SGL-M
		T=50			T=100			T=200	
				Legendre degree $L = 10$					
Beta$(1, 3)$	1.955	1.155	0.952	1.846	0.378	0.386	1.804	0.163	0.179
	0.002	0.013	0.011	0.002	0.006	0.006	0.001	0.002	0.003
Beta$(2, 3)$	1.211	0.902	0.885	1.157	0.423	0.370	1.128	0.225	0.162
	0.001	0.008	0.009	0.001	0.005	0.005	0.001	0.002	0.002
Beta$(2, 2)$	1.062	0.747	0.775	1.019	0.293	0.314	0.995	0.126	0.135
	0.001	0.007	0.008	0.001	0.004	0.005	0.001	0.002	0.002
				low frequency noise level σ_1=5					
Beta$(1, 3)$	2.022	1.736	1.389	1.972	1.290	0.757	1.893	0.716	0.355
	0.001	0.012	0.017	0.002	0.011	0.011	0.002	0.008	0.006
Beta$(2, 3)$	1.244	1.132	1.060	1.220	0.929	0.700	1.186	0.657	0.385
	0.001	0.008	0.013	0.001	0.007	0.009	0.001	0.006	0.006
Beta$(2, 2)$	1.089	0.980	0.936	1.069	0.781	0.604	1.042	0.509	0.315
	0.001	0.007	0.012	0.001	0.007	0.008	0.001	0.005	0.005
				Half high-frequency lags					
Beta$(1, 3)$	1.997	1.509	1.083	1.925	0.882	0.686	1.878	0.571	0.535
	0.001	0.011	0.011	0.001	0.008	0.006	0.001	0.004	0.004
Beta$(2, 3)$	1.243	1.121	1.026	1.221	0.913	0.828	1.202	0.729	0.716
	0.001	0.006	0.007	0.001	0.005	0.006	0.001	0.004	0.004
Beta$(2, 2)$	1.090	0.998	0.955	1.074	0.838	0.813	1.059	0.719	0.740
	0.001	0.005	0.007	0.001	0.005	0.005	0.000	0.004	0.004
				Number of covariates $p = 50$					
Beta$(1, 3)$	2.031	1.563	1.038	1.931	0.620	0.401	1.841	0.223	0.174
	0.001	0.010	0.011	0.002	0.007	0.005	0.001	0.002	0.002
Beta$(2, 3)$	1.250	1.067	0.883	1.206	0.606	0.436	1.156	0.296	0.196
	0.001	0.006	0.007	0.001	0.006	0.005	0.001	0.003	0.002
Beta$(2, 2)$	1.095	0.923	0.782	1.062	0.461	0.360	1.023	0.178	0.153
	0.000	0.005	0.006	0.001	0.005	0.004	0.001	0.002	0.002

Table A.4: Shape of weights estimation accuracy II. – See Table A.3

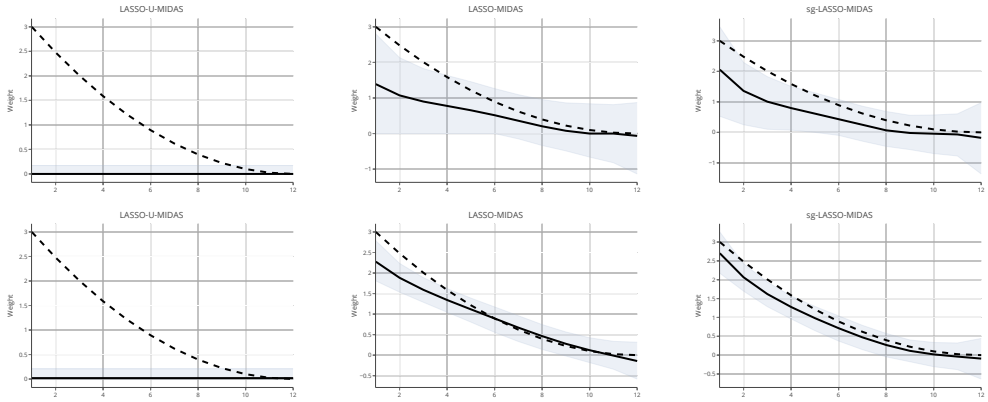


Figure A.1: The figure shows the fitted Beta(1,3) weights. We plot the estimated weights for the LASSO-U-MIDAS, LASSO-MIDAS and sg-LASSO-MIDAS estimators for the baseline DGP scenario. The first row plots weights for the sample size $T = 50$, the second row plot weights for the sample size $T = 200$. Black solid line is the median estimate of the weights function, black dashed line is the population weight function, and the grey area is the 90% confidence interval.

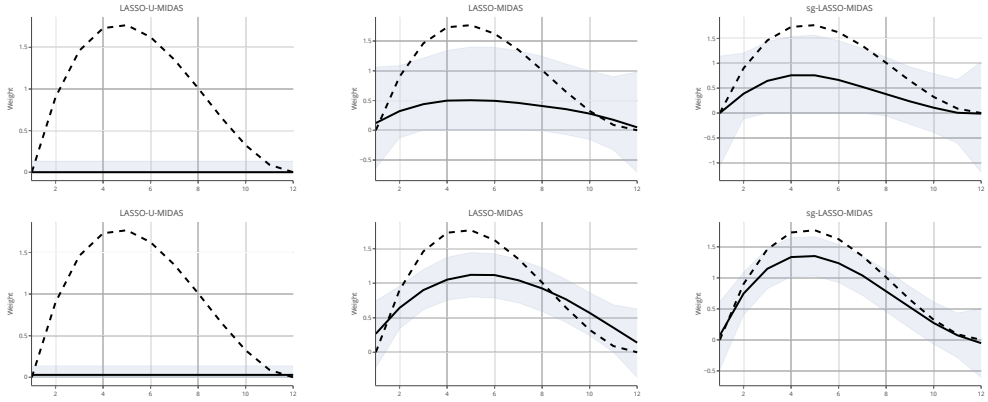


Figure A.2: The figure shows the fitted Beta(2,3) weights. We plot the estimated weights for the LASSO-U-MIDAS, LASSO-MIDAS and sg-LASSO-MIDAS estimators for the baseline DGP scenario. The first row plots weights for the sample size $T = 50$, the second row plot weights for the sample size $T = 200$. Black solid line is the median estimate of the weights function, black dashed line is the population weight function, and the grey area is the 90% confidence interval.

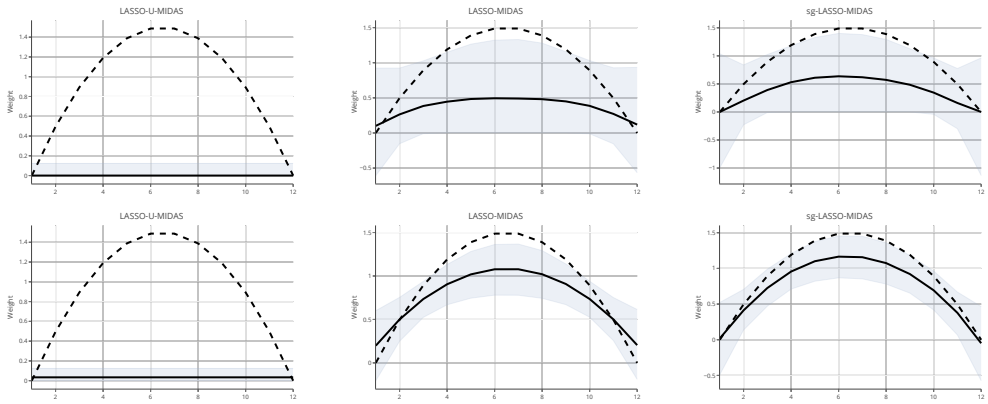


Figure A.3: The figure shows the fitted Beta(2,2) weights. We plot the estimated weights for the LASSO-U-MIDAS, LASSO-MIDAS and sg-LASSO-MIDAS estimators for the baseline DGP scenario. The first row plots weights for the sample size $T = 50$, the second row plot weights for the sample size $T = 200$. Black solid line is the median estimate of the weights function, black dashed line is the population weight function, and the grey area is the 90% confidence interval.

A.4 Detailed data description

To compute the main results reported in Table 1, we use thirty monthly macro series, eight quarterly survey covariates, and six news attention series which are aggregated using Legendre polynomials.[26] Thirty predictors are real-time macro series that are either directly taken from FRED-MD dataset, or are calculated by using FRED-MD series, denoted by FRED-MD and FRED-MD (calc.) respectively in the *Source* column of the data description table A.4 below. Note that for all monthly macro data, we use real-time vintages, which effectively means that we take all macro series with one month delay. For example, if we nowcast the first quarter of GDP one month ahead, we use data up to the end of February, and thus all macro series that enter the model areavailable up to the end of January. We then use Legendre polynomials of degree three for all covariates to aggregate twelve lags of monthly macro data. In particular, let $x_{t+(h+1-j)/m,k}$ be k-th $\in \{1,\dots,30\}$ covariate at quarter t, $j = 1,\dots,12$, $m = 3$, and $h = 2$, minus additional lag to account for the publication delay of macro series. Therefore, for macro series the first lag index is: $(h+1-j-1)/m = (2+1-1-1)/3 = 1/3$. We then collect all lags in X_{tk} vector, i.e.

$$X_{tk} = (x_{t+1/3,k}, x_{t+0/3,k}, \dots, x_{t-11/3,k})$$

and aggregate X_{tk} using dictionary W consisting of Legendre polynomials, $X_{tk}W$. In this case, $X_{tk}W$ is defined as a single group for the sg-LASSO estimator.

Furthermore, we use data from the Survey of Professional Forecasters (SPF) - nowcasts and forecasts - aggregated using Legendre polynomial of degree three. More precisely, denote $x^h_{mean,t}$ and $x^h_{median,t}$ the mean and median forecast at the horizon h. We collect all mean and median forecast horizons that do not have missing entries, $h = 0,1,2,3$, in the X_t vector

$$X_t = (x^0_{mean,t}, x^1_{mean,t}, x^2_{mean,t}, x^3_{mean,t}, x^0_{median,t}, x^1_{median,t}, x^2_{median,t}, x^3_{median,t}),$$

and aggregate this data using the same dictionary W, i.e. X_tW, and define X_tW as a single group for the sg-LASSO. Note that SPF data is quarterly; therefore, we aggregate it cross-sectionally rather than time series.

Lastly, we take six news attention series from http://structureofnews.com/, see Table A.5, and, as for macro series, use Legendre polynomials of degree three to aggregate twelve monthly lags of

[26] We transform all macro covariates using transformations suggested by McCracken and Ng (2016), see Table A.5. We then standardize all covariates before the aggregation step.

each attention news series. However, in this case, news attention series is used without a publication delay, that is, for the one-month horizon, we take the series up to the end of the second month.

We compute the predictions by using the expanding window scheme. The first nowcast is for the 2002 Q1, the effective sample size is from 1990 February to 2001 November, and the prediction is computed using 2002 February data. We calculate predictions until the sample is exhausted, which is 2017 Q2, the last date for which news attention data is available.

	id	Source	T-code
1	Commodities	Bybee, Kelly, Manela, and Xiu (2020)	1
2	Government budgets	Bybee, Kelly, Manela, and Xiu (2020)	1
3	Oil market	Bybee, Kelly, Manela, and Xiu (2020)	1
4	Recession	Bybee, Kelly, Manela, and Xiu (2020)	1
5	Savings & loans	Bybee, Kelly, Manela, and Xiu (2020)	1
6	Mortgages	Bybee, Kelly, Manela, and Xiu (2020)	1
7	IP: Business Equipment	FRED-MD	5
8	IP: Fuels	FRED-MD	5
9	IP: Manufacturing (SIC)	FRED-MD	5
10	IP: Durable Consumer Goods	FRED-MD	5
11	Civilians Unemployed - Less Than 5 Weeks	FRED-MD	5
12	All Employees: Financial Activities	FRED-MD	5
13	All Employees: Government	FRED-MD	5
14	Initial Claims	FRED-MD	5
15	All Employees: Total nonfarm	FRED-MD	5
16	All Employees: Service-Providing Industries	FRED-MD	5
17	All Employees: Mining and Logging: Mining	FRED-MD	5
18	Unemployment Rate	FRED-MD	2
19	All Employees: Manufacturing	FRED-MD	5
20	Housing Starts, Midwest	FRED-MD	4
21	Housing Starts, West	FRED-MD	4
22	Housing Starts: Total New Privately Owned	FRED-MD	4
23	Retail and Food Services Sales	FRED-MD	5
24	New Orders for Durable Goods	FRED-MD	5
25	MZM Money Stock	FRED-MD	6
26	Personal Cons. Expend.: Chain Index	FRED-MD	6
27	CPI: All Items	FRED-MD	6
28	S&P: Industrials	FRED-MD	5
29	3-Month AA Fin. Comm. Paper Rate	FRED-MD	2
30	Crude Oil	FRED-MD	6
31	5-Year Treasury	FRED-MD	2
32	3-Month Commercial Paper - FEDFUNDS	FRED-MD	1
33	Moodys Baa Corporate Bond - FEDFUNDS	FRED-MD	1
34	10-Year Treasury	FRED-MD	2
35	S&P 500	FRED-MD	5
36	Moodys Baa - Aaa Corporate Bond Spread	FRED-MD (calc.)	1
37	Survey of professional forecasters	Phil. Fed	1

Table A.5: Data description table – The *id* column gives mnemonics according to data source, which is given in the second column *Source*. The column *T-code* denotes the data transformation applied to a time-series, which are: (1) not transformed, (2) Δx_t, (3) $\Delta^2 x_t$, (4) $\log(x_t)$, (5) $\Delta \log(x_t)$, (6) $\Delta^2 \log(x_t)$.

A.5 Additional results for empirical application

A.5.1 Full-sample nowcasting results

As for the main results, we benchmark our predictions with the simple random walk (RW) model. We implemented the following alternative machine learning nowcasting methods. The first method is the PCA factor-augmented autoregression, where we estimate the first principal component of a monthly macro panel and use it together with four autoregressive lags. We denote this model PCA-OLS. We then consider three alternative penalty functions for the same linear model: Ridge, LASSO and Elastic Net. For these methods, we leave high-frequency lags unrestricted, and thus we call these methods the unrestricted MIDAS (U-MIDAS). Lastly, we use sg-LASSO estimator, where we also aggregate high-frequency lags using MIDAS weights. The weight function is approximated by using Legendre polynomials of degree three. For each method, we use four lags of GDP and twelve lags of each high-frequency covariates. The first prediction is for the 2002 Q1, and we use expanding window scheme up until 2017 Q2. In this case, we use larger samples to estimate all models, and thus the effective sample starts from 1960 February. For each quarter, we take predictors that do not having missing values. In addition, we compute corporate bond spread and discard NONBORRES series due to a possible break in this series, see Uematsu and Tanaka (2019). In total, the number of covariates (without taking into account lags) ranges from 94 to 114.

In Table A.6, we report out-of-sample nowcasting results for one-month horizon using real-time data vintages. We report root mean squared forecast error relative to the RW model (column Rel-RMSE) and the Diebold Mariano predictive accuracy test statistic (DM). In addition to models we implemented, we compare GDP growth nowcasts provided by the New York Fed (denoted NY Fed). The column DM-stat-1 reports Diebold Mariano test statistic where we compare NY Fed predictions with other methods, and the column DM-stat-2 compares sg-LASSO-MIDAS with alternative methods.

Using full-sample data, sg-LASSO-MIDAS model also gives smaller forecast errors when compared with NY Fed predictions, however, the gains are not statistically significant. Nonetheless, sg-LASSO-MIDAS model nowcasts give significantly smaller prediction errors compared with other alternative machine learning methods.

Figure A.4: Sparsity pattern for 50 most selected covariates.

	Rel-RMSE	DM-stat-1	DM-stat-2
RW	2.606	2.370	3.318
PCA-OLS	0.849	0.854	1.975
Ridge-U-MIDAS	0.838	0.763	1.974
LASSO-U-MIDAS	0.853	0.967	2.039
Elastic Net-U-MIDAS	0.833	0.699	1.888
sg-LASSO-MIDAS	0.750	-0.739	
NY Fed	0.790		0.739

Table A.6: Nowcast comparison table – Forecast horizon is one month ahead. Column *Rel-RMSE* reports root mean squared forecasts error relative to the RW model. Column *DM-stat-1* reports Diebold and Mariano (1995) test statistic of all models relative to NY Fed nowcasts, while column *DM-stat-2* reports the Diebold Mariano test statistic relative to sg-LASSO-MIDAS model. Out-of-sample period: 2002 Q1 to 2017 Q2.

Figure A.5: Cumulative sum of loss differential. Gray shaded are — NBER recession period.

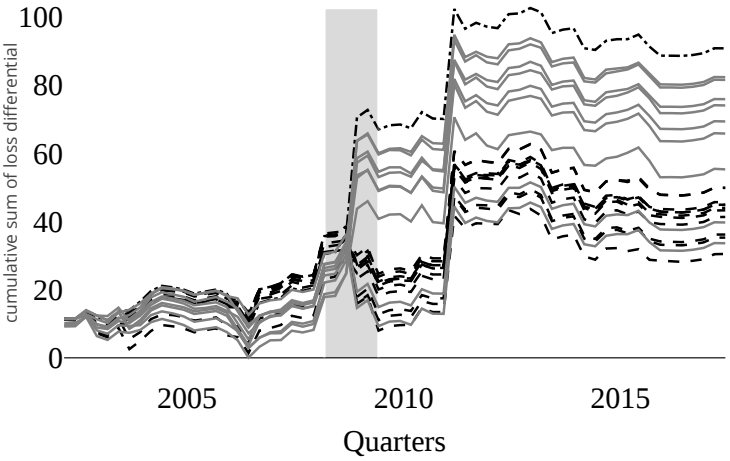


Figure A.6: Cumulative sum of loss differentials (cumsfe) of New York Fed nowcasts compared with sg-LASSO-MIDAS model with textual analysis based data for different $\alpha \in [0, 1]$ value. Dashed black lines are cumsfe's for $\alpha \in [0, 0.5]$, gray solid lines - $\alpha \in (0.5, 1]$ excluding 0.65, and dash-dotted black line - cumsfe for $\alpha = 0.65$. Gray shaded area — NBER recession period.

Table A.7: Significance test table

M_T	10	20	30
Commodities	3.152	3.271	3.024
Government budgets	16.461	13.512	13.680
Oil market	11.216	9.796	12.068
Recession	5.843	4.720	4.132
Savings & loans	1.900	2.275	2.440
Mortgages	1.923	2.728	3.031

Table A.8: Significance test table – Wald test statistic values for news attention series for a set of truncation parameter M_T values. The number of lags in the HAC estimator correspond to M_T. The critical value at 5% is 9.488.