JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2021, VOL. 00, NO. 0, 1–13
https://doi.org/10.1080/07350015.2021.1899933

Machine Learning Time Series Regressions With an Application to Nowcasting

Andrii Babii[a], Eric Ghysels[b,c], and Jonas Striaukas[d]

[a]Department of Economics, University of North Carolina–Chapel Hill, Chapel Hill, NC; [b]Department of Economics and Kenan-Flagler Business School, University of North Carolina–Chapel Hill; [c]CEPR, London, UK; [d]LIDAM UC Louvain and Research Fellow of the Fonds de la Recherche Scientifique—FNRS, Brussels, Belgium

ABSTRACT

This article 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 text data can be a useful addition to more traditional numerical data. Our methodology is implemented in the R package midasml, available from CRAN.

ARTICLE HISTORY
Received May 2020
Accepted February 2021

KEYWORDS
Fat tails; High-dimensional time series; Mixed-frequency data; Sparse-group LASSO; Tau-mixing; Textual news data

1. Introduction

The statistical imprecision of quarterly gross domestic product (GDP) estimates, along with the fact that the first estimate is available with a delay of nearly a month, pose a significant challenge to policy makers, market participants, and other observers with an interest in monitoring the state of the economy in real time; see, for example, Ghysels, Horan, and Moench (2018) for a recent discussion of macroeconomic data revisions and publication delays. 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 (e.g., quarterly GDP), whereas the real-time information (e.g., daily, weekly, or monthly) can be used to update the state, or to put it differently, to *nowcast* the low-frequency series of interest. Traditional methods 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 likelihood-based methods and Kalman filtering techniques; see Bańbura et al. (2013) for a survey.

So far, nowcasting has mostly relied on the so-called standard macroeconomic data releases, one of the most prominent examples being the Employment Situation report released on the first Friday of every month by the U.S. Bureau of Labor Statistics. This report includes the data on the nonfarm payroll employment, average hourly earnings, and other summary statistics of the labor market activity. Since most sectors of the economy move together over the business cycle, good news for the labor market is usually good news for the aggregate economy. In addition to the labor market data, the nowcasting models typically also rely on construction spending, (non-)manufacturing report, retail trade, price indices, etc., which we will call the traditional macroeconomic data. One prominent example of nowcast is produced by the Federal Reserve Bank of New York relying on a dynamic factor model with 36 predictors of different frequencies; see Bok et al. (2018) for more details.

Thirty-six predictors of traditional macroeconomic series may be viewed as a small number compared to hundreds of other potentially available and useful nontraditional series. For instance, macroeconomists increasingly rely on nonstandard data such as textual analysis via machine learning, which means potentially hundreds of series. A textual analysis dataset based on *Wall Street Journal* articles that has been recently made available features a taxonomy of 180 topics; see Bybee et al. (2020). 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 the daily business newspapers relying on a dynamic factor model where time-varying sparsity is enforced upon the factor loadings using a latent threshold mechanism. His work shows the feasibility of traditional state space setting, yet the challenges grow when we also start thinking about adding other potentially high-dimensional datasets, such as payment systems information or GPS tracking data. Studies for Canada (Galbraith and Tkacz (2018)), Denmark (Carlsen and Storgaard 2010), India (Raju and Balakrishnan 2019), Italy (Aprigliano, Ardizzi, and Monteforte 2019), Norway (Aastveit et al. 2020), Portugal (Duarte, Rodrigues, and Rua 2017), and the United States (Barnett et al. 2016) find that payment transactions can help to nowcast and to forecast GDP and private consumption in the short term; see also Moriwaki (2019) for nowcasting

CONTACT Eric Ghysels eghysels@gmail.com Department of Economics and Kenan-Flagler Business School, University of North Carolina–Chapel Hill and CEPR.

unemployment rates with smartphone GPS data, among others. We could quickly reach numerical complexities involved with estimating high-dimensional state space models, making the dynamic factor model approach potentially computationally prohibitively complex and slow, although some alternatives to the Kalman filter exist for the large data environments; see, for example, Chan and Jeliazkov (2009) and Delle Monache and Petrella (2019). In this article, we study nowcasting a 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 nonstandard data generated by textual analysis of financial press articles. Several novel contributions are required to achieve our goal. The contributions of our article are both theoretical and practical. Regarding the former: (a) we propose a new structured approach to high-dimensional regularized time regression problems, (b) we establish a complete estimation and prediction theory for high-dimensional time series regressions under assumptions comparable to the classical GMM and QML estimators, and (c) we establish nonasymptotic and asymptotic estimation and prediction properties of our regularized time series regression approach. Regarding the practical contributions we document superior nowcasting performance with respect to the state-of-the-art state space model approach to nowcasting implemented by the Federal Reserve Bank of New York. In the remainder of this Introduction we devote a paragraph to each of these contributions, starting with the theoretical ones.

First, we argue that the high-dimensional mixed frequency 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; see Simon et al. (2013). The attractive feature of the sg-LASSO estimator is that it allows us to combine effectively the approximately sparse and dense signals; see, for example, Carrasco and Rossi (2016) for a comprehensive treatment of high-dimensional dense time series regressions as well as Mogliani and Simoni (2020) for a complementary to ours Bayesian view of penalized MIDAS regressions.

Second, we recognize that the economic and financial time series data are persistent and often heavy-tailed, while the bulk of the machine learning methods assumes iid data and/or exponential tails for covariates and regression errors; see Belloni et al. (2020) for a comprehensive review of high-dimensional econometrics with iid 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) and Uematsu and Tanaka (2019) established oracle inequalities for regressions with iid errors with sub-Gaussian tails; Wong, Li, and Tewari (2020) consider β-mixing series with exponential tails; Wu and Wu (2016), Han and Tsay (2017), and Chernozhukov et al. (2020) established oracle inequalities for causal Bernoulli shifts with independent innovations and polynomial tails under the functional dependence measure of Wu (2005); see also Medeiros and Mendes (2016) and Medeiros and Mendes (2017) for results on the adaptive LASSO based on the triplex tail inequality for mixingales of Jiang (2009). Despite these efforts, there is no complete estimation and prediction theory for high-dimensional time series regressions under the assumptions comparable to the classical GMM and QML estimators. For instance, the best currently available results are too restrictive for the MIDAS projection model, which is typically an example of a causal Bernoulli shift with *dependent innovations*. Moreover, the *mixing processes* with *polynomial tails* that are especially relevant for the financial and macroeconomic time series have not been properly treated due to the fact that the sharp Fuk-Nagaev inequality was not available in the relevant literature until recently. The Fuk-Nagaev inequality, see 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 Marcinkiewicz-Zygmund or Rosenthal's moment inequalities.

Third, our article fills these gaps in the literature relying on the Fuk-Nagaev inequality for τ-mixing processes of Babii, Ghysels, and Striaukas (2020) and establishes the nonasymptotic and asymptotic estimation and prediction properties of the sg-LASSO projections under weak tail conditions and potential misspecification. The class of τ-mixing processes is fairly rich covering he α-mixing processes, causal linear processes with infinitely many lags of β-mixing processes, and nonlinear Markov processes; see Dedecker and Prieur (2004, 2005) for more details, as well as Carrasco and Chen (2002) and Francq and Zakoian (2019) for mixing properties of various processes encountered in time series econometrics. We show that the sparse-group LASSO estimator works when the data have fat tails. In particular, our weak tail conditions require at least $4 + \epsilon$ finite moments for covariates, while the number of finite moments for the error process can be as low as $2 + \nu$, provided that covariates have sufficiently light tails. From the theoretical point of view, we impose *approximate sparsity*, relaxing the assumption of exact sparsity of the projection coefficients and allowing for other forms of misspecification (see Giannone, Lenza, and Primiceri 2018 for further discussion on the topic of sparsity). Finally, we cover the LASSO and the group LASSO as special cases.

We find that our nowcasts are either superior to or at par with those posted by the Federal Reserve Bank of New York (henceforth NY Fed). This is the case when (a) we compare our approach with the NY Fed using the same data, or (b) when we compare our approach using an expanded high-dimensional dataset. The former is a comparison of methods, whereas the latter pertains to the value of the additional (nonstandard) big data. To deal with such massive nontraditional datasets, instead of using the likelihood-based dynamic factor models, we rely on a different approach that involves machine learning methods based on the regularized empirical risk minimization principle and data sampled at different frequencies. We adopt the MIDAS (mixed data sampling) projection approach which is more amenable to high-dimensional data environments. Our general framework also includes the standard same frequency time series regressions.

The rest of the article is organized as follows. Section 2 presents the setting of (potentially mixed frequency) high-dimensional time series regressions. Section 3 characterizes nonasymptotic estimation and prediction accuracy of the sg-LASSO estimator for τ-mixing processes with polynomial tails. We report on a Monte Carlo study in Section 4 which provides further insights regarding 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 $\|X\|_q = (\mathbb{E}|X|^q)^{1/q}$ be its L_q norm with $q \geq 1$. 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 $\mathcal{G}$ be a partition of $[p]$ defining the group structure, which is assumed to be known to the econometrician. For a vector $\beta \in \mathbf{R}^p$, the sparse-group structure is described by a pair $(S_0, \mathcal{G}_0)$, where $S_0 = \{j \in [p] : \beta_j \neq 0\}$ and $\mathcal{G}_0 = \{G \in \mathcal{G} : \beta_G \neq 0\}$ are the support and, respectively, the group support of β. We also use $|S|$ to denote the cardinality of arbitrary set S. For $b \in \mathbf{R}^p$, its ℓ_q norm is denoted as $|b|_q = \left(\sum_{j\in[p]} |b_j|^q\right)^{1/q}$ for $q \in [1, \infty)$ and $|b|_\infty = \max_{j\in[p]} |b_j|$ for $q = \infty$. For $\mathbf{u}, \mathbf{v} \in \mathbf{R}^T$, the empirical inner product is defined as $\langle \mathbf{u}, \mathbf{v} \rangle_T = T^{-1} \sum_{t=1}^T u_t v_t$ 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 elements. 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 $b_n \lesssim a_n$.

2. High-Dimensional Mixed-Frequency Regressions

Let $\{y_t : t \in [T]\}$ be the target low-frequency series observed at integer time points $t \in [T]$. 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 $\{x_{t-(j-1)/m,k}, j \in [m], t \in [T], k \in [K]\}$ possibly measured at some higher frequency with $m \geq 1$ observations for every t and consider the following regression model:

$$\phi(L) y_t = \rho_0 + \sum_{k=1}^K \psi(L^{1/m}; \beta_k) x_{t,k} + u_t, \qquad t \in [T],$$

where $\phi(L) = I - \rho_1 L - \rho_2 L^2 - \cdots - \rho_J L^J$ is a low-frequency lag polynomial and $\psi(L^{1/m}; \beta_k) x_{t,k} = 1/m \sum_{j=1}^m \beta_{j,k} x_{t-(j-1)/m,k}$ is a 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. Note that the polynomial $\psi(L^{1/m}; \beta_k) x_{t,k}$ involves the same m number of high-frequency lags for each covariate $k \in [K]$, which is done for the sake of simplicity and can easily be relaxed; see Section 5.

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 the post-WWII quarterly GDP growth series has less than 300 observations.

The LASSO estimator, see Tibshirani (1996), offered an appealing convex relaxation of a difficult nonconvex best subset selection problem. It allows increasing the precision of predictions via the selection of sparse and parsimonious models. In this article, we focus on the structured sparsity with additional dimensionality reductions that aim to improve upon the unstructured LASSO estimator in the time series setting.

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

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

where β_k is L-dimensional vector of coefficients with $L \leq m$ and $\omega : [0, 1] \times \mathbf{R}^L \to \mathbf{R}$ is some weight function. Second, we approximate the weight function as

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

where $\{w_l : l = 1, \ldots, L\}$ 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 regression analysis literature. More generally, the dictionary may consist of arbitrary approximating functions, including the classical orthogonal bases of $L_2[0, 1]$; see online appendix Section A.1 for more examples. Using orthogonal polynomials typically reduces the multicollinearity and leads to better finite sample performance. It is worth mentioning that the specification with dictionaries deviates from the standard MIDAS regressions and leads to a computationally attractive convex optimization problem, cf. Marsilli (2014).

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 article. 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 parameterize 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) 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.

To describe the estimation procedure, let $\mathbf{y} = (y_1, \dots, y_T)^\top$, be a vector of dependent variable and let $\mathbf{X} = (\iota, \mathbf{y}_1, \dots, \mathbf{y}_J, Z_1 W, \dots, Z_K W)$, be a design matrix, where $\iota = (1, 1, \dots, 1)^\top$ is a vector of ones, $\mathbf{y}_j = (y_{1-j}, \dots, y_{T-j})^\top$, $Z_k = (x_{k,t-(j-1)/m})_{t\in[T],j\in[m]}$ is a $T \times m$ matrix of the covariate $k \in [K]$, and $W = \left(w_l\left((j-1)/m\right)/m\right)_{j\in[m],l\in[L]}$ is an $m \times L$ matrix of weights. In addition, put $\beta = (\beta_0^\top, \beta_1^\top, \dots, \beta_K^\top)^\top$, where $\beta_0 = (\rho_0, \rho_1, \dots, \rho_J)^\top$ is a vector of parameters pertaining to the group consisting of the intercept and the 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

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

with a penalty function that interpolates between the ℓ_1 LASSO penalty and the 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 and $\mathcal{G}$ is a group structure (partition of $[p]$) specified by the econometrician. Note that estimator in Equation (2) is defined as a solution to the convex optimization problem and can be computed efficiently, for example, using an appropriate coordinate descent algorithm; see Simon et al. (2013).

The amount of penalization in Equation (2) 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, which is reminiscent of the elastic net. In Figure 1 we illustrate the geometry of the penalty function for different groupings and different values of α covering (a) LASSO with $\alpha = 1$, (b) group LASSO with one group, $\alpha = 0$, and two sg-LASSO cases (c) one group and (d) two groups both with $\alpha = 0.5$. In practice, groups are defined by a particular problem and are specified by the econometrician, while α can be fixed or selected jointly with λ in a data-driven way such as using the cross-validation.

3. High-Dimensional Time Series Regressions

3.1. High-Dimensional Regressions and τ-Mixing

We focus on a generic high-dimensional linear projection model with a countable number of regressors

$$y_t = \sum_{j=0}^{\infty} x_{t,j}\beta_j + u_t, \qquad \mathbb{E}[u_t x_{t,j}] = 0, \quad \forall j \geq 1, \qquad t \in \mathbf{Z}, \tag{3}$$

where $x_{t,0} = 1$ and $m_t \triangleq \sum_{j=0}^{\infty} x_{t,j}\beta_j$ is a well-defined random variable. In particular, to ensure that y_t is a well-defined economic quantity, we need $\beta_j \downarrow 0$ sufficiently fast, which is a form of the *approximate sparsity* condition, see Belloni et al. (2020). This setting nests the high-dimensional ARDL-MIDAS projections described in the previous section and more generally may allow for other high-dimensional time series models. In practice, given a (large) number of covariates, lags thereof, as well as lags of the dependent variable, denoted $x_t \in \mathbf{R}^p$, we would approximate m_t with $x_t^\top\beta \triangleq \sum_{j=0}^{p} x_{t,j}\beta_j$, where $p < \infty$ and the regression coefficient $\beta \in \mathbf{R}^p$ could be sparse. Importantly, our settings allows for the approximate sparsity as well as other forms of misspecification and the main result of the following section allows for $m_t \neq x_t^\top\beta$.

Using the setting of Equation (2), for a sample $(y_t, x_t)_{t=1}^T$, write

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

where $\mathbf{y} = (y_1, \dots, y_T)^\top$, $\mathbf{m} = (m_1, \dots, m_T)^\top$, and $\mathbf{u} = (u_1, \dots, u_T)^\top$. The approximation to $\mathbf{m}$ is denoted $\mathbf{X}\beta$, where $\mathbf{X} = (x_1, \dots, x_T)^\top$ is a $T \times p$ matrix of covariates and $\beta = (\beta_1, \dots, \beta_p)^\top$ is a vector of unknown regression coefficients.

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

$$\tau(\mathcal{M}, \xi) = \left\| \sup_{f\in \mathrm{Lip}_1} \left|\mathbb{E}(f(\xi)|\mathcal{M}) - \mathbb{E}(f(\xi))\right| \right\|_1,$$

where $\mathrm{Lip}_1 = \left\{f : \mathbf{R}^l \to \mathbf{R} : \ |f(x) - f(y)| \leq |x - y|_1\right\}$ is a set of 1-Lipschitz functions. 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 canonical filtration. The τ-*mixing coefficient* of $(\xi_t)_{t\in\mathbf{Z}}$ is defined as

$$\tau_k = \sup_{j\geq 1} \frac{1}{j} \sup_{t+k\leq t_1<\dots<t_j} \tau(\mathcal{M}_t, (\xi_{t_1}, \dots, \xi_{t_j})), \qquad k \geq 0.$$

If $\tau_k \downarrow 0$ as $k \to \infty$, then the process $(\xi_t)_{t\in\mathbf{Z}}$ is called τ-*mixing*. The τ-mixing coefficients were introduced in Dedecker and Prieur (2004) as dependence measures weaker than mixing. Note that the commonly used α- and β-mixing conditions are too restrictive for the linear projection model with an ARDL-MIDAS process. Indeed, a causal linear process with dependent innovations is not necessary α-mixing; see also Andrews (1984) for an example of AR(1) process which is not α-mixing. Roughly speaking, τ-mixing processes are somewhere between mixingales and α-mixing processes and can accommodate such counterexamples. At the same time, sharp Fuk-Nagaev inequalities are available for τ-mixing processes which to the best of our knowledge is not the case for the mixingales or near-epoch dependent processes; see Babii, Ghysels, and Striaukas (2020).

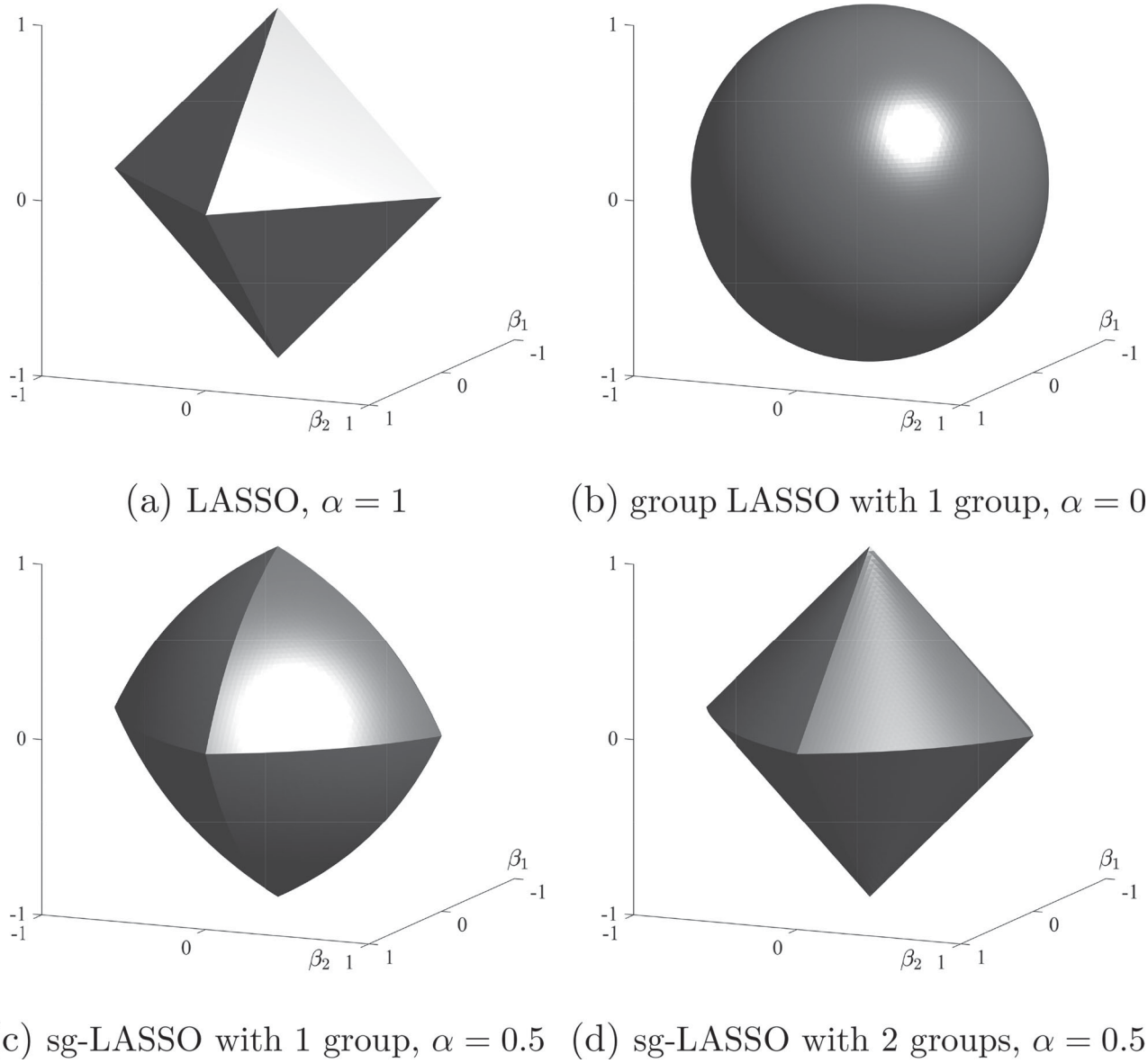

Figure 1. The figure shows the geometry of the constrained set, $\{b \in \mathbf{R}^2 : \Omega(b) \leq 1\}$, corresponding to the sparse-group LASSO penalty function for several groupings and values of α.

Dedecker and Prieur (2004, 2005) discussed how to verify the τ-mixing property for causal Bernoulli shifts with dependent innovations and nonlinear Markov processes. It is also worth comparing the τ-mixing coefficient to other weak dependence coefficients. Suppose that $(\xi_t)_{t\in\mathbf{Z}}$ is a real-valued stationary process and let $\gamma_k = \|\mathbb{E}(\xi_k|\mathcal{M}_0) - \mathbb{E}(\xi_k)\|_1$ be its L_1 mixingale coefficient. Then we clearly have $\gamma_k \leq \tau_k$ and it is known that

$$\begin{aligned}|\mathrm{cov}(\xi_0, \xi_k)| &\leq \int_0^{\gamma_k} Q \circ G(u)\mathrm{d}u \\ &\leq \int_0^{\tau_k} Q \circ G(u)\mathrm{d}u \\ &\leq \tau_k^{\frac{q-2}{q-1}} \|\xi_0\|_q^{q/(q-1)},\end{aligned}$$

where Q is the generalized inverse of $x \mapsto \Pr(|\xi_0| > x)$ and G is the generalized inverse of $x \mapsto \int_0^x Q(u)\mathrm{d}u$; see Babii, Ghysels, and Striaukas (2020), Lemma A.1.1. Therefore, the τ-mixing coefficient provides a sharp control of autocovariances similarly to the L_1 mixingale coefficients, which in turn can be used to ensure that the long-run variance of $(\xi_t)_{t\in\mathbf{Z}}$ exists. The τ-mixing coefficient is also bounded by the α-mixing coefficient, denoted α_k, as follows:

$$\tau_k \leq 2\int_0^{2\alpha_k} Q(u)\mathrm{d}u \leq 2\|\xi_0\|_q(2\alpha_k)^{1/r},$$

where the first inequality follows by Dedecker and Prieur (2004), Lemma 7 and the second by Hölder's inequality with $q, r \geq 1$ such that $q^{-1} + r^{-1} = 1$. It is worth mentioning that the mixing properties for various time series models in econometrics, including GARCH, stochastic volatility, or autoregressive conditional duration are well-known; see, for example, Carrasco and Chen (2002), Francq and Zakoian (2019), Babii, Chen, and Ghysels (2019); see also Dedecker et al. (2007) for more examples and a comprehensive comparison of various weak dependence coefficients.

3.2. Estimation and Prediction Properties

In this section, we introduce the main assumptions for the high-dimensional time series regressions and study the estimation and prediction properties of the sg-LASSO estimator covering the LASSO and the group LASSO estimators as special cases. The following assumption imposes some mild restrictions on the stochastic processes in the high-dimensional regression equation (3).

Assumption 3.1 (Data). For every $j,k \in [p]$, the processes $(u_t x_{t,j},)_{t\in\mathbf{Z}}$ and $(x_{t,j}x_{t,k})_{t\in\mathbf{Z}}$ are stationary such that (i) $\|u_0\|_q < \infty$ and $\max_{j\in[p]}\|x_{0,j}\|_r = O(1)$ for some constants $q > 2r/(r-2)$ and $r > 4$; (ii) the τ-mixing coefficients are $\tau_k \le ck^{-a}$ and, respectively, $\tilde{\tau}_k \le ck^{-b}$ for all $k \ge 0$ and some $c > 0$, $a > (\varsigma-1)/(\varsigma-2)$, $b > (r-2)/(r-4)$, and $\varsigma = qr/(q+r)$.

It is worth mentioning that the stationarity condition is not essential and can be relaxed to the existence of the limiting variance of partial sums at costs of heavier notations and proofs. Condition (i) requires that covariates have at least 4 finite moments, while the number of moments required for the error process can be as low as $2+\epsilon$, depending on the integrability of covariates. Therefore, (i) may allow for heavy-tailed distributions commonly encountered in financial and economic time series, for example, asset returns and volatilities. Given the integrability in (i), (ii) requires that the τ-mixing coefficients decrease to zero sufficiently fast; see online appendix, Section A.3 for moments and τ-mixing coefficients of ARDL-MIDAS. It is known that the β-mixing coefficients decrease geometrically fast, for example, for geometrically ergodic Markov chains, in which case (ii) holds for every $a,b > 0$. Therefore, (ii) allows for relatively persistent processes.

For the support S_0 and the group support $\mathcal{G}_0$ of β, put

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

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

For some $c_0 > 0$, define $\mathcal{C}(c_0) \triangleq \{\Delta \in \mathbf{R}^p : \Omega_1(\Delta) \le c_0\Omega_0(\Delta)\}$. The following assumption generalizes the restricted eigenvalue condition of Bickel, Ritov, and Tsybakov (2009) to the sg-LASSO estimator and is imposed on the population covariance matrix $\Sigma = \mathbb{E}[\mathbf{X}^\top\mathbf{X}/T]$.

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

Recall that if Σ is a positive definite matrix, then for all $\Delta \in \mathbf{R}^p$, we have $\Delta^\top\Sigma\Delta \ge \gamma|\Delta|_2^2$, where γ is the smallest eigenvalue of Σ. Therefore, in this case Assumption 3.2 is trivially satisfied because $|\Delta|_2^2 \ge \sum_{G\in\mathcal{G}_0}|\Delta_G|_2^2$. The positive definiteness of Σ is also known as a completeness condition and Assumption 3.2 can be understood as its weak version; see Babii and Florens (2020) and references therein. It is worth emphasizing that $\gamma > 0$ in Assumption 3.2 is a universal constant independent of p, which is the case, for example, when Σ is a Toeplitz matrix or a spiked identity matrix. Alternatively, we could allow for $\gamma \downarrow 0$ as $p \to \infty$, in which case the term γ^{-1} would appear in our nonasymptotic bounds slowing down the speed of convergence, and we may interpret γ as a measure of ill-posedness in the spirit of econometrics literature on ill-posed inverse problems; see Carrasco, Florens, and Renault (2007).

The value of the regularization parameter is determined by the Fuk-Nagaev concentration inequality, appearing in the online appendix, see Theorem A.1.

Assumption 3.3 (Regularization). For some $\delta \in (0,1)$

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

where $\kappa = ((a+1)\varsigma - 1)/(a+\varsigma-1)$ and a,ς are as in Assumption 3.1.

The regularization parameter in Assumption 3.3 is determined by the persistence of the data, quantified by a, and the tails, quantified by $\varsigma = qr/(q+r)$. This dependence is reflected in the *dependence-tails exponent* κ. The following result describes the nonasymptotic prediction and estimation bounds for the sg-LASSO estimator, see Online Appendix, Section A.2 for the proof.

Theorem 3.1. Suppose that Assumptions 3.1, 3.2, and 3.3 are satisfied. Then with probability at least $1-\delta-O(p^2(T^{1-\mu}s_\alpha^\mu + \exp(-cT/s_\alpha^2)))$

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

and

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

for some $c > 0$, where $\sqrt{s_\alpha} = \alpha\sqrt{|S_0|} + (1-\alpha)\sqrt{|\mathcal{G}_0|}$ and $\mu = ((b+1)r-2)/(r+2(b-1))$.

Theorem 3.1 provides nonasymptotic guarantees for the estimation and prediction with the sg-LASSO estimator reflecting potential misspecification. In the special case of the LASSO estimator ($\alpha = 1$), we obtain the counterpart to the result of Belloni et al. (2012) for the LASSO estimator with iid data taking into account that we may have $m_t \ne x_t^\top\beta$. At another extreme, when $\alpha = 0$, we obtain the nonasymptotic bounds for the group LASSO allowing for misspecification which to the best of our knowledge are new, cf. Negahban et al. (2012) and van de Geer (2016). We call s_α the *effective sparsity constant*. This constant reflects the benefits of the sparse-group structure for the sg-LASSO estimator that cannot be deduced from the results currently available for the LASSO or the group LASSO.

Remark 3.1. Since the ℓ_1-norm is equivalent to the Ω-norm whenever groups have fixed size, we deduce from Theorem 3.1 that

$$|\hat{\beta}-\beta|_1 \lesssim s_\alpha\lambda + \lambda^{-1}\|\mathbf{m}-\mathbf{X}\beta\|_T^2 + \sqrt{s_\alpha}\|\mathbf{m}-\mathbf{X}\beta\|_T.$$

Next, we consider the asymptotic regime, in which the misspecification error vanishes when the sample size increases as described in the following assumption.

Assumption 3.4. (i) $\|\mathbf{m} - \mathbf{X}\beta\|_T^2 = O_P\left(s_\alpha \lambda^2\right)$; and (ii) $p^2 T^{1-\mu} s_\alpha^\mu \to 0$ and $p^2 \exp(-cT/s_\alpha^2) \to 0$.

The following corollary is an immediate consequence of Theorem 3.1.

Corollary 3.1. Suppose that Assumptions 3.1–3.4 hold. 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

$$|\hat{\beta} - \beta|_1 = O_P\left(\frac{s_\alpha p^{1/\kappa}}{T^{1-1/\kappa}} \vee s_\alpha \sqrt{\frac{\log p}{T}}\right).$$

If the effective sparsity constant s_α is fixed, then $p = o(T^{\kappa-1})$ is a sufficient condition for the prediction and estimation errors to vanish, whenever $\mu \geq 2\kappa - 1$. In this case Assumption 3.4 (ii) is vacuous. More generally, s_α is allowed to increase slowly with the sample size. Convergence rates in Corollary 3.1 quantify the effect of tails and persistence of the data on the prediction and estimation accuracies of the sg-LASSO estimator. In particular, lighter tails and less persistence allow us to handle a larger number of covariates p compared to the sample size T. In particular p can increase faster than T, provided that $\kappa > 2$.

Remark 3.2. In the special case of the LASSO estimator with iid data, Corollary 4 of Fuk and Nagaev (1971) leads to the convergence rate of order $O_P\left(\frac{p^{1/\varsigma}}{T^{1-1/\varsigma}} \vee \sqrt{\frac{\log p}{T}}\right)$. If the τ-mixing coefficients decrease geometrically fast (e.g., stationary AR(p)), then $\kappa \approx \varsigma$ for a sufficiently large value of the dependence exponent a, in which case the convergence rates in Corollary 3.1 are close to the iid case. In this sense these rates depend sharply on the tails exponent ς, and we can conclude that for geometrically decreasing τ-mixing coefficients, the persistence of the data should not affect the convergence rates of the LASSO.

Remark 3.3. In the special case of the LASSO estimator, if $(u_t)_{t\in\mathbf{Z}}$ and $(x_t)_{t\in\mathbf{Z}}$ are causal Bernoulli shifts with independent innovations and at least $q = r \geq 8$ finite moments, one can deduce from Chernozhukov et al. (2020), Lemma 5.1 and Corollary 5.1, the convergence rate of order $O_P\left(\frac{(p\omega_T)^{1/\varsigma}}{T^{1-1/\varsigma}} \vee \sqrt{\frac{\log p}{T}}\right)$, where $\omega_T = 1$ (weakly dependent case) or $\omega_T = T^{\varsigma/2-1-a\varsigma} \uparrow \infty$ (strongly dependent case), provided that the physical dependence coefficients are of size $O(k^{-a})$. Note that for causal Bernoulli shifts with independent innovations, the physical dependence coefficients are not directly comparable to τ-mixing coefficients; see Dedecker et al. (2007), Remark 3.1 on p.32.

4. Monte Carlo Experiments

We assess via simulations 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 the 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.

4.1. Simulation Design

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

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

where $u_t \sim_{\text{iid}} N(0, \sigma_u^2)$ and the DGP for covariates $\{x_{k,t-(j-1)/m} : j \in [m], k \in [K]\}$ 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 by 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 Beta(1, 3), Beta(2, 3), and 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 iid realizations of the univariate autoregressive (AR) process $x_h = \rho x_{h-1} + \varepsilon_h$, where $\rho = 0.2$ or $\rho = 0.7$ and either $\varepsilon_h \sim_{\text{iid}} N(0, \sigma_\varepsilon^2)$, $\sigma_\varepsilon^2 = 1$, or $\varepsilon_h \sim_{\text{iid}}$ student-$t(5)$, where h denotes the high-frequency sampling.
2. Multivariate vector autoregressive (VAR) process $X_h = \Phi X_{h-1} + \boldsymbol{\varepsilon}_h$, where $\varepsilon_h \sim_{\text{iid}} N(0, I_K)$ and Φ is a block diagonal matrix described below.

For the AR simulation design, we initiate the processes as $x_0 \sim N\left(0, \sigma^2/(1-\rho^2)\right)$ and $y_0 \sim N\left(0, \sigma^2(1-\rho_2)/((1+\rho_2)((1-\rho_2)^2 - \rho_1^2))\right)$. For the VAR, the initial value of (y_t) is the same, while $X_0 \sim N(0, I_K)$. In all cases, the first 200 observations are treated as burn-in. In the estimation procedure, we add seven noisy covariates which are generated in the same way as the relevant covariates and use five low-frequency lags. The empirical models use a dictionary which consists of Legendre polynomials up to degree $L = 10$ shifted to the $[0, 1]$ interval with the MIDAS weight function approximated as in Equation (1). The sample size is $T \in \{50, 100, 200\}$, and for all the experiments we use 5000 simulation replications.

We assess the performance of different methods by modifying the assumptions on the error terms of the high-frequency process $\boldsymbol{\varepsilon}_h$, considering multivariate high-frequency processes, changing the degree of Legendre polynomials L, increasing the noise level of the low-frequency process σ_u^2, 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, and 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 2015). In the second model we use MIDAS weights together with the 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 the 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 $K + 5$ groups. We select the value of tuning parameters λ and α using the 5-fold cross-validation, defining folds as adjacent blocks over the time dimension to take into account the time series dependence. This model is denoted sg-LASSO-MIDAS.

For regressions with aggregated data, we consider: (a) Flow aggregation (FLOW): $x^A_{k,t} = 1/m \sum_{j=1}^{m} x_{k,t-(j-1)/m}$, (b) Stock aggregation (STOCK): $x^A_{k,t} = x_{k,t}$, and (c) Middle high-frequency lag (MIDDLE): single middle value of the high-frequency lag with ties solved in favor of the most recent observation (i.e., we take a single 6th lag if $m = 12$). In these cases, the models are estimated using the OLS estimator, which is unfeasible when the number of covariates becomes equal to the sample size and we leave results blank in this case.

4.2. Simulation Results

Detailed results are reported in the online appendix of Tables A.1–A.2, cover the average mean squared forecast errors for one-step-ahead 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 nonlinear weight function approximations perform much better than all other methods 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). These findings are robust to increases in the persistence parameter of covariates ρ from 0.2 to 0.7. The LASSO without MIDAS weighting has typically large forecast errors. Comparing across simulation scenarios, all methods seem to perform worse with heavy-tailed or persistent covariates. In these cases, however, the impact on the sg-LASSO-MIDAS method is lesser compared to the other methods. This simulation evidence supports our theoretical results and findings in the empirical application. Finally, 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 cannot accurately recover the weights in small samples and/or low signal-to-noise ratio scenarios. Using Legendre polynomials improves the performance substantially and the sg-LASSO seems to improve even more over the unstructured LASSO.

5. Nowcasting U.S. GDP with Macro, Financial and Textual News Data

We nowcast U.S. GDP with macroeconomic, financial, and textual news data. Details regarding the data sources appear in the online appendix of Section A.5. Regarding the macro data, we rely on 34 series used in the Federal Reserve Bank of New York nowcast model, discarding two series ("PPI: Final demand" and "Merchant wholesalers: Inventories") due to very short samples; see Bok et al. (2018) for more details regarding this data.

For all macro data, we use real-time vintages, which effectively means that we take all macro series with a delay as well real-time data releases. For example, if we nowcast the first quarter of GDP one month before the quarter ends, we use data up to the end of February, and therefore all macro series with a delay of one month that enter the model are available up to the end of January. As we use data real-time data releases, the January observation in this case is also the first release of a particular series. We use Legendre polynomials of degree three for all macro covariates to aggregate 12 lags of monthly macro data. In particular, let $x_{t+(h+1-j)/m,k}$ be kth covariate at quarter t with $m = 3$ months per quarter and $h = 2 - 1 = 1$ months into the quarter (2 months into the quarter minus 1 month due to publication delay), where $j = 1, 2, \ldots, 12$ is the monthly lag. We then collect all lags in a vector

$$X_{t,k} = (x_{t+1/3,k}, x_{t+0/3,k}, \ldots, x_{t-10/3,k})^\top$$

and aggregate $X_{t,k}$ using a dictionary W consisting of Legendre polynomials, so that $X_{t,k}W$ defines as a single group for the sg-LASSO estimator.

In addition to macro and financial data, we also use the textual analysis data. We take 76 news attention series from Bybee et al. (2020) and use Legendre polynomials of degree two to aggregate three monthly lags of each news attention series. Note that the news attention series are used without a publication delay, that is, for the one-month horizon, we take the series up to the end of the second month. Moreover, the Bybee et al. (2020) news topic models involve rolling samples, avoiding look ahead biases when used in our nowcasts.

We compute the predictions using a rolling window scheme. The first nowcast is for 2002 Q1, for which we use fifteen years (60 quarters) of data, and the prediction is computed using 2002 January (2-month horizon) February (1-month), and March (end of the quarter) data. We calculate predictions until the sample is exhausted, which is 2017 Q2, the last date for which news attention data is available. As indicated above, we report results for the 2-month, 1-month, and the end-of-quarter horizons. Our target variable is the first release, that is, the advance estimate of real GDP growth. For each quarter and nowcast horizon, we tune sg-LASSO-MIDAS regularization parameters λ and α using 5-fold cross-validation, defining folds as adjacent blocks over the time dimension to take into account the time series nature of the data. Finally, we follow the literature on nowcasting real GDP and define our target variable to be the annualized growth rate.

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

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

where $\phi(L)$ is the low-frequency lag polynomial, μ is the regression intercept, and $\psi(L^{1/m};\beta_k)x_{tk}, k = 1,\ldots,K$ are lags of high-frequency covariates. Following Section 2, the high-frequency lag polynomial is defined as

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

where for the kth covariate, h_k indicates the number of leading months of available data in the quarter t, q_k is the number of quarters of covariate lags, and we approximate the weight function ω with the Legendre polynomial. For example, if $h_k = 1$ and $q_k = 4$, then we have 1 month of data into a quarter and use $q_k m = 12$ monthly lags for a covariate k.

We benchmark our predictions against the simple AR(1) model, which is considered to be a reasonable starting point for short-term GDP growth predictions. We focus on predictions of our method, sg-LASSO-MIDAS, with and without financial data combined with series based on the textual analysis. One natural comparison is with the publicly available Federal Reserve Bank of New York, denoted NY Fed, model implied nowcasts. We adopt the following strategy. First, we focus on the same series that are used to calculate the NY Fed nowcasts. The purpose here is to compare *models* since the data inputs are the same. This means that we compare the performance of dynamic factor models (NY Fed) with that of machine learning regularized regression methods (sg-LASSO-MIDAS). Next, we expand the dataset to see whether additional financial and textual news series can improve the nowcast performance.

In Table 1, we report results based on real-time macro data used for the NY Fed model, see Bok et al. (2018). The results show that the sg-LASSO-MIDAS performs much better than the NY Fed nowcasts at the longer, that is, 2-month, horizon. Our method significantly beats the benchmark AR(1) model for all the horizons, and the accuracy of the nowcasts improve with the horizon. Our end-of-quarter and 1-month horizon nowcasts are similar to the NY Fed ones, with the sg-LASSO-MIDAS being slightly better numerically but not statistically. We also report the average Superior Predictive Ability test of Quaedvlieg (2021) over all three horizons and the result reveals that the improvement of the sg-LASSO-MIDAS model versus the NY Fed nowcasts is significant at the 5% significance level. Finally, we report results that do not discard two series ("PPI: Final demand" and "Merchant wholesalers: Inventories") due to short samples in the online appendix of Section A.5.1. The results are very similar and do not change our conclusions.

Table 1. Nowcast comparisons for models with macro data only – Nowcast horizons are 2- and 1-month ahead, as well as the end of the quarter.

	Rel-RMSE	DM-stat-1	DM-stat-2
		2-month horizon	
AR(1)	2.056	0.612	2.985
sg-LASSO-MIDAS	0.739	−2.481	
NY Fed	0.946		2.481
		1-month horizon	
AR(1)	2.056	2.025	2.556
sg-LASSO-MIDAS	0.725	−0.818	
NY Fed	0.805		0.818
		End-of-quarter	
AR(1)	2.056	2.992	3.000
sg-LASSO-MIDAS	0.701	−0.077	
NY Fed	0.708		0.077
		p-value of aSPA test	
			0.046

The comparison in Table 1 does not fully exploit the potential of our methods, as it is easy to expand the data series beyond the small number used by the NY Fed nowcasting model. In Table 2, we report results with additional sets of covariates which are financial series, advocated by Andreou, Ghysels, and Kourtellos (2013), and textual analysis of news. In total, the models select from 118 series—34 macro, 8 financial, and 76 news attention series. For the moment we focus only on the first three columns of the table. At the longer horizon of 2 months, the method seems to produce slightly worse nowcasts compared to the results reported in Table 1 using only macro data. However, we find significant improvements in prediction quality for the shorter 1-month and end-of-quarter horizons. In particular, a significant increase in accuracy relative to NY Fed nowcasts appears at the 1-month horizon. We report again the average Superior Predictive Ability test of Quaedvlieg (2021) over all three horizons with the same result that the improvement of sg-LASSO-MIDAS versus the NY Fed nowcasts is significant at the 5% significance level. Finally, we report results for several alternatives, namely, PCA-OLS, ridge, LASSO, and Elastic Net, using the unrestricted MIDAS scheme. Our approach produces more accurate nowcasts compared to these alternatives.

The inclusion of financial series is not common in traditional nowcasting models, see, for example, Bok et al. (2018), on the grounds that though timely, financial data are noisy, hence do not contribute to the accuracy of the nowcasts. One may wonder how our model performs excluding these series. Therefore, we run our nowcasting regressions using only macro and news attention series, excluding financial data; results are reported in the online appendix of Section A.5.1. Notably, results are slightly worse compared with the results that include financial data, supporting our initial choice. Similarly, Andreou, Ghysels, and Kourtellos (2013) find that financial data are helpful in GDP nowcasting applications.

Table 2 also features an entry called SPF (median), where we report results for the median survey of professional nowcasts for the 1-month horizon, and analyze how the model-based nowcasts compare with the predictions using the publicly available Survey of Professional Forecasters maintained by the Federal Reserve Bank of Philadelphia. We find that the sg-LASSO-MIDAS model-based nowcasts are similar to the SPF-implied nowcasts. We also find that the NY Fed nowcasts are significantly worse than the SPF.

In Figure 2 we plot the cumulative sum of squared forecast error (CUMSFE) loss differential of sg-LASSO-MIDAS versus

Table 2. Nowcast comparison table – Nowcast horizons are 2- and 1-month ahead, as well as the end of the quarter. Column *Rel-RMSE* reports root mean squared forecasts error relative to the AR(1) model. Column *DM-stat-1* reports Diebold and Mariano (1995) test statistic of all models relative to the NY FED nowcast, while column *DM-stat-2* reports the Diebold Mariano test statistic relative to the sg-LASSO model. Columns *DM-stat-3* and *DM-stat-4* report the Diebold Mariano test statistic for the same models, but excludes the recession period. For the 1-month horizon, the last row *SPF (median)* reports test statistics for the same models comparing with the SPF median nowcasts. The last row reports the p-value of the average Superior Predictive Ability (aSPA) test, see Quaedvlieg (2021), over the three horizons of sg-LASSO-MIDAS model compared to the NY Fed nowcasts, including (left) and excluding (right) financial crisis period. Out-of-sample period: 2002 Q1 to 2017 Q2.

	Rel-RMSE	DM-stat-1	DM-stat-2	DM-stat-3	DM-stat-4
			2-month horizon		
PCA-OLS	0.982	0.416	2.772	0.350	2.978
Ridge-U-MIDAS	0.918	−0.188	1.073	−1.593	0.281
LASSO-U-MIDAS	0.996	0.275	1.280	−1.983	−0.294
Elastic Net-U-MIDAS	0.907	−0.266	0.976	−1.725	0.042
sg-LASSO-MIDAS	0.779	−2.038		−2.349	
NY Fed	0.946		2.038		2.349
			1-month horizon		
PCA-OLS	1.028	2.296	3.668	2.010	3.399
Ridge-U-MIDAS	0.940	0.927	2.063	−0.184	1.979
LASSO-U-MIDAS	1.044	1.286	1.996	−0.397	1.498
Elastic Net-U-MIDAS	0.990	1.341	2.508	0.444	2.859
sg-LASSO-MIDAS	0.672	−1.426		−1.341	
NY Fed	0.805		1.426		1.341
SPF (median)	0.639	−2.317	−0.490	−1.743	0.282
			End-of-quarter		
PCA-OLS	0.988	3.414	3.400	3.113	3.155
Ridge-U-MIDAS	0.939	1.918	1.952	0.867	1.200
LASSO-U-MIDAS	1.014	1.790	1.773	0.276	0.517
Elastic Net-U-MIDAS	0.947	2.045	2.034	1.198	1.400
sg-LASSO-MIDAS	0.696	−0.156		−0.159	
NY Fed	0.707		0.156		0.159
			p-value of aSPA test		
			0.042		0.056

NOTES: Column *Rel-RMSE* reports root mean squared forecasts error relative to the AR(1) 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. The last row reports the *p*-value of the average Superior Predictive Ability (aSPA) test, see Quaedvlieg (2021), over the three horizons of sg-LASSO-MIDAS model compared to the NY Fed nowcasts. Out-of-sample period: 2002 Q1 to 2017 Q2.

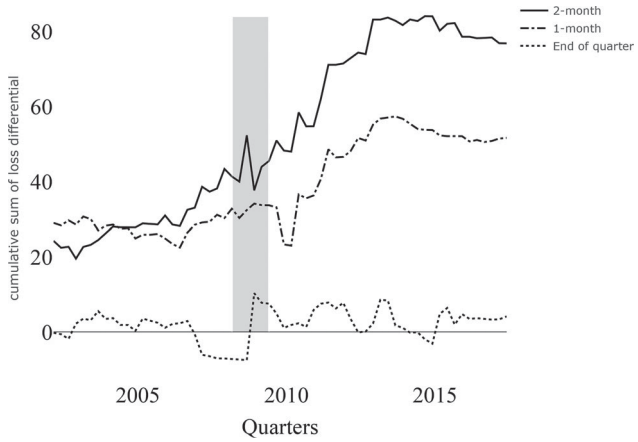

Figure 2. Cumulative sum of loss differentials of sg-LASSO-MIDAS model nowcasts including financial and textual data compared with the New York Fed model for three nowcasting horizons: solid black line cumsfe for the 2-months horizon, dash-dotted black line - cumsfe for the 1-month horizon, and dotted line for the end-of-quarter nowcasts. The gray shaded area corresponds to the NBER recession period.

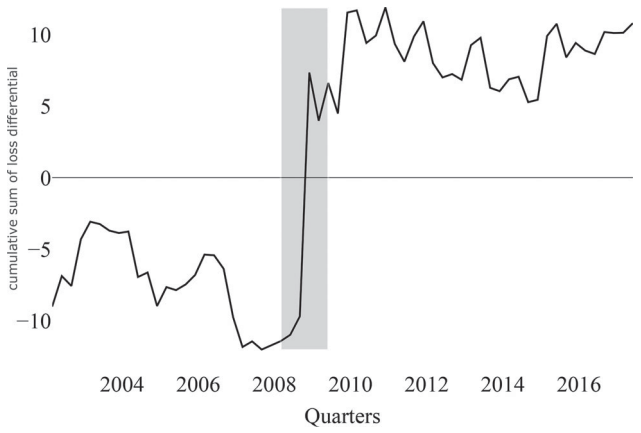

Figure 3. Cumulative sum of loss differentials (CUMSFE) of sg-LASSO-MIDAS nowcasts when we include vs. when we exclude the additional financial and textual news data, averaged over 1-month and the end-of-quarter horizons. The gray shaded area corresponds to the NBER recession period.

NY Fed nowcasts for the three horizons. The CUMSFE is computed as follows:

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

for model $M1$ versus $M2$. A positive value of $\text{CUMSFE}_{t,t+k}$ means that the model $M1$ has larger squared forecast errors compared to model $M2$ up to $t+k$, and negative values imply the opposite. In our case, $M1$ is the New York Fed prediction error, while $M2$ is the sg-LASSO-MIDAS model. We observe persistent gains for the 2- and 1-month horizons throughout the out-of-sample period. When comparing the sg-LASSO-MIDAS results with additional financial and textual news series versus

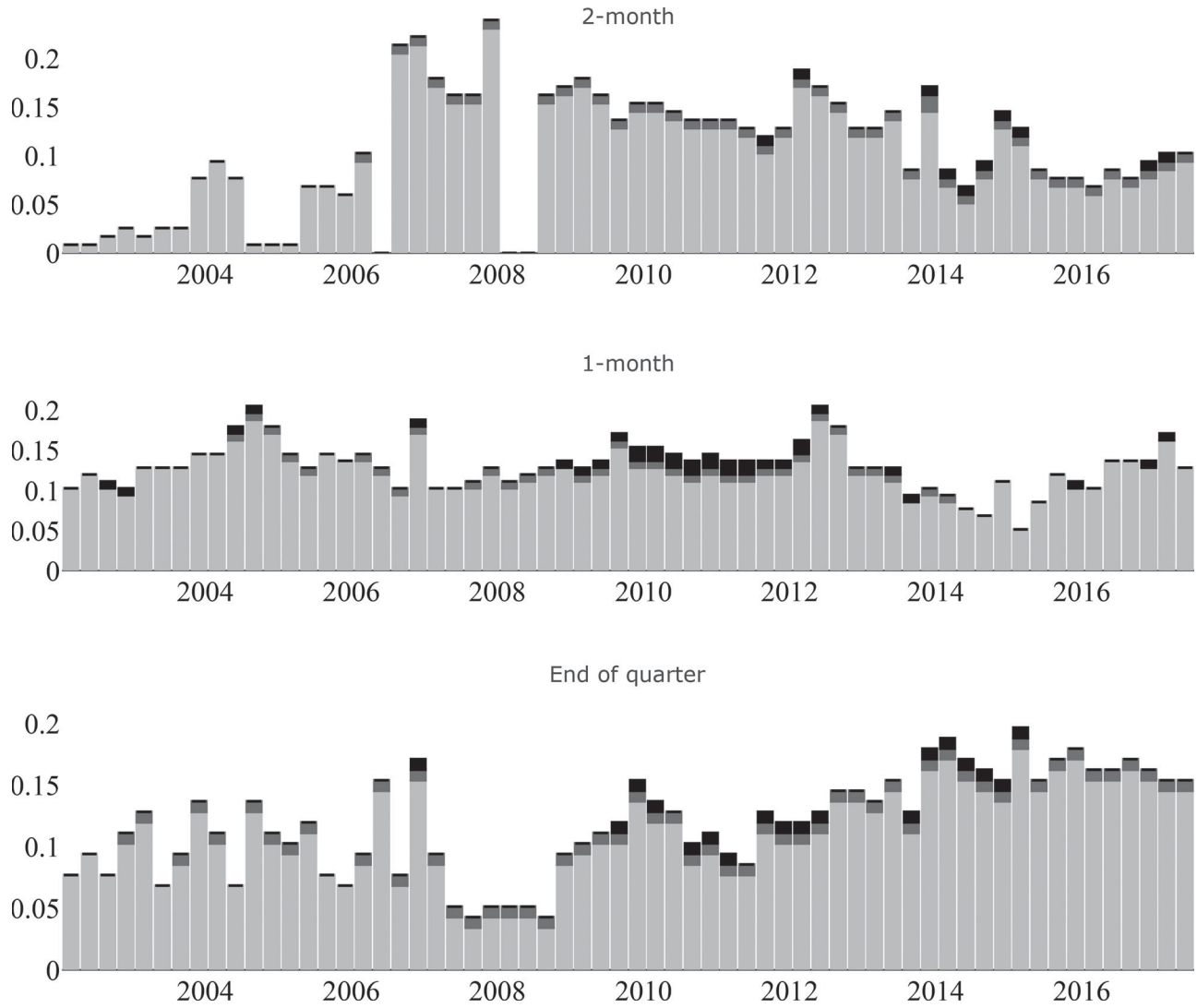

Figure 4. The fraction of selected covariates attributed to macro (light gray), financial (dark gray), and textual (black) data for three monthly horizons.

those based on macro data only, we see a notable improvement at the 1-month horizon and a more modest one at the end-of-quarter horizons. In Figure 3, we plot the average CUMSFE for the 1-month and end-of-quarter horizons and observe that the largest gains of additional financial and textual news data are achieved during the financial crisis.

The result in Figure 3 prompts the question whether our results are mostly driven by this unusual period in our out-of-sample data. To assess this, we turn our attention again to the last two columns of Table 2 reporting Diebold and Mariano (1995) test statistics which exclude the financial crisis period. Compared to the tests previously discussed, we find that the results largely remain the same, but some alternatives seem to slightly improve (e.g., LASSO or Elastic Net). Note that this also implies that our method performs better during periods with heavy-tailed observations, such as the financial crisis. It should also be noted that overall there is a slight deterioration of the average Superior Predictive Ability test over all three horizons when we remove the financial crisis.

In Figure 4, we plot the fraction of selected covariates by the sg-LASSO-MIDAS model when we use the macro, financial, and textual analysis data. For each reference quarter, we compute the ratio of each group of variables relative to the total number of covariates. In each subfigure, we plot the three different horizons. For all horizons, the macro series are selected more often than financial and/or textual data. The number of selected series increases with the horizon, however, the pattern of denser macro series and sparser financial and textual series is visible for all three horizons. The results are in line with the literature—macro series tend to co-move, hence we see a denser pattern in the selection of such series, see, for example, Bok et al. (2018). On the other hand, the alternative textual analysis data appear to be very sparse, yet still important for nowcasting accuracy, see also Thorsrud (2020).

6. Conclusion

This article 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 the estimation, inference, forecasting, and nowcasting with regularized machine learning methods. The first contribution of the article 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, from Babii, Ghysels, and Striaukas (2020), valid for a large class of τ-mixing processes, including α-mixing processes commonly used in econometrics. Building on this inequality, we establish the nonasymptotic 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 with time series data, including unconventional 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 performed by the sg-LASSO estimator. We find that the sg-LASSO outperforms the unstructured LASSO in small samples and conclude that incorporating specific data structures should be helpful in various applications.

Our empirical results also show that the sg-LASSO-MIDAS using only macro data performs statistically better than NY Fed nowcasts at 2-month horizons and overall for the 1- and 2-month and end-of-quarter horizons. This is a comparison involving the same data and, therefore, pertains to models. This implies that machine learning models are, for this particular case, better than the state space dynamic factor models. When we add the financial data and the textual news data, a total of 118 series, we find significant improvements in prediction quality for the shorter 1-month and end-of-quarter horizons.

Acknowledgments

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, the AI Innovations Forum organized by SAS and the Kenan Institute, the 12th World Congress of the Econometric Society, and seminar participants at the Vanderbilt University, as well as Harold Chiang, Jianqing Fan, Jonathan Hill, Michele Lenza, and Dacheng Xiu for comments. We are also grateful to the referees and the editor whose comments helped us to improve our article significantly. All remaining errors are ours.

References

Aastveit, K. A., Fastbø, T. M., Granziera, E., Paulsen, K. S., and Torstensen, K. N. (2020), "Nowcasting Norwegian Household Consumption with Debit Card Transaction Data," Discussion Paper, Norges Bank. [1]

Almon, S. (1965), "The Distributed Lag Between Capital Appropriations and Expenditures," *Econometrica*, 33, 178–196. [3]

Andreou, E., Ghysels, E., and Kourtellos, A. (2013), "Should Macroeconomic Forecasters Use Daily Financial Data and How?," *Journal of Business and Economic Statistics*, 31, 240–251. [3,9]

Andrews, D. W. (1984), "Non-strong Mixing Autoregressive Processes," *Journal of Applied Probability*, 21, 930–934. [4]

Aprigliano, V., Ardizzi, G., and Monteforte, L. (2019), "Using Payment System Data to Forecast Economic Activity," *International Journal of Central Banking*, 15, 55–80. [1]

Babii, A., Chen, X., and Ghysels, E. (2019), "Commercial and Residential Mortgage Defaults: Spatial Dependence With Frailty," *Journal of Econometrics*, 212, 47–77. [5]

Babii, A., and Florens, J.-P. (2020), "Is Completeness Necessary? Estimation in Nonidentified Linear Models," arXiv preprint arXiv:1709.03473v3. [6]

Babii, A., Ghysels, E., and Striaukas, J. (2020), "High-dimensional Granger Causality Tests With an Application to VIX and Newsn," arXiv preprint arXiv:1912.06307v2. [2,4,5,12]

Bańbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013), "Nowcasting 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. Amsterdam: Elsevier. [1]

Barnett, W., Chauvet, M., Leiva-Leon, D., and Su, L. (2016), "Nowcasting Nominal GDP with the Credit-Card Augmented Divisia Monetary Aggregates," Working Paper University of Kansas, Department of Economics. [1]

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012), "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, 80, 2369–2429. [6]

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., and Kato, K. (2020), "High-dimensional Econometrics and Generalized GMM," *Handbook of Econometrics* (forthcoming). [2,4]

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of LASSO and Dantzig Selector," *Annals of Statistics*, 37, 1705–1732. [6]

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. [1,8,9,11]

Bybee, L., Kelly, B. T., Manela, A., and Xiu, D. (2020), "The Structure of Economic News," *National Bureau of Economic Research*, and *http://structureofnews.com*. [1,8]

Carlsen, M., and Storgaard, P. E. (2010), "Dankort Payments as a Timely Indicator of Retail Sales in Denmark," Danmarks Nationalbank Working Papers. [1]

Carrasco, M., and Chen, X. (2002), "Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models," *Econometric Theory*, 18, 17–39. [2,5]

Carrasco, M., Florens, J.-P., and Renault, E. (2007), "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," *Handbook of Econometrics*, 6, 5633–5751. [6]

Carrasco, M., and Rossi, B. (2016), "In-sample Inference and Forecasting in Misspecified Factor Models," *Journal of Business and Economic Statistics*, 34, 313–338. [2]

Chan, J. C., and Jeliazkov, I. (2009), "Efficient Simulation and Integrated Likelihood Estimation in State Space Models," *International Journal of Mathematical Modelling and Numerical Optimisation*, 1, 101–120. [2]

Chernozhukov, V., Härdle, W. K., Huang, C., and Wang, W. (2020), "Lasso-driven Inference in Time and Space," *Annals of Statistics* (forthcoming). [2,7]

Dedecker, J., Doukhan, P., Lang, G., Rafael, L. R. J., Louhichi, S., and Prieur, C. (2007), *Weak Dependence: With Examples and Applications*, Berlin: Springer. [5,7]

Dedecker, J., and C. Prieur (2004), "Coupling for τ-dependent Sequences and Applications," *Journal of Theoretical Probability*, 17, 861–885. [2,4,5]

——— (2005), "New Dependence Coefficients. Examples and Applications to Statistics," *Probability Theory and Related Fields*, 132(2), 203–236. [2,5]

Delle Monache, D., and Petrella, I. (2019), "Efficient Matrix Approach for Classical Inference in State Space Models," *Economics Letters*, 181, 22–27. [2]

Diebold, F. X., and Mariano, R. S. (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13(3), 253–263. [10,11]

Duarte, C., Rodrigues, P. M., and Rua, A. (2017), "A Mixed Frequency Approach to the Forecasting of Private Consumption With ATM/POS Data," *International Journal of Forecasting*, 33, 61–75. [1]

Foroni, C., Marcellino, M., and Schumacher, C. (2015), "Unrestricted Mixed Data Sampling (U-MIDAS): MIDAS Regressions With

Unrestricted Lag Polynomials," *Journal of the Royal Statistical Society*, Series A, 178, 57–82. [8]

Francq, C., and Zakoian, J.-M. (2019), *GARCH Models: Structure, Statistical Inference and Financial Applications*, Wiley. [2,5]

Fuk, D. K., and Nagaev, S. V. (1971), "Probability Inequalities for Sums of Independent Random Variables," *Theory of Probability and Its Applications*, 16, 643–660. [2,7]

Galbraith, J. W., and Tkacz, G. (2018), "Nowcasting with Payments System Data," *International Journal of Forecasting*, 34(2), 366–376. [1]

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. [1]

Ghysels, E., and H. Qian (2019), "Estimating MIDAS Regressions Via OLS With Polynomial Parameter Profiling," *Econometrics and Statistics*, 9, 1–16. [7]

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. [3]

Ghysels, E., Sinko, A., and Valkanov, R. (2007), "MIDAS Regressions: Further Results and New Directions," *Econometric Reviews*, 26, 53–90. [7]

Giannone, D., Lenza, M., and Primiceri, G. E. (2018), "Economic Predictions With Big Data: The Illusion of Sparsity," Staff Reports 847, Federal Reserve Bank of New York. [2]

Han, Y., and Tsay, R. S. (2017), "High-Dimensional Linear Regression for Dependent Observations With Application to Nowcasting," arXiv preprint arXiv:1706.07899. [2]

Jiang, W. (2009), "On Uniform Deviations of General Empirical Risks with Unboundedness, Dependence, and High Dimensionality," *Journal of Machine Learning Research*, 10, 977–996. [2]

Kock, A. B., and Callot, L. (2015), "Oracle Inequalities for High Dimensional Vector Autoregressions," *Journal of Econometrics*, 186, 325–344. [2]

Marsilli, C. (2014), "Variable Selection in Predictive MIDAS Models," Working Papers 520, Banque de France. [3]

Medeiros, M. C., and Mendes, E. F. (2016), "ℓ_1-regularization of High-dimensional Time-series Models With Non-Gaussian and Heteroskedastic Errors," *Journal of Econometrics*, 191, 255–271. [2]

——— (2017), "Adaptive LASSO Estimation for ARDL Models With GARCH Innovations," *Econometric Reviews*, 36, 622–637. [2]

Mogliani, M., and Simoni, A. (2020), "Bayesian MIDAS Penalized Regressions: Estimation, Selection, and Prediction," *Journal of Econometrics* (forthcoming). [2]

Moriwaki, D. (2019), "Nowcasting Unemployment Rates With Smartphone GPS Data," in *International Workshop on Multiple-Aspect Analysis of Semantic Trajectories*, eds. K. Tserpes, C. Renso, and S. Matwin, pp. 21–33. Springer. [1]

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A Unified Framework for High-dimensional Analysis of M-estimators With Decomposable Regularizers," *Statistical Science*, 27, 538–557. [6]

Quaedvlieg, R. (2021), "Multi-horizon Forecast Comparison," *Journal of Business and Economic Statistics*, 39, 40–53. [9,10]

Raju, S., and Balakrishnan, M. (2019), "Nowcasting Economic Activity in India Using Payment Systems Data," *Journal of Payments Strategy and Systems*, 13, 72–81. [1]

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), "A Sparse-group LASSO," *Journal of Computational and Graphical Statistics*, 22, 231–245. [2,4]

Thorsrud, L. A. (2020), "Words Are the New Numbers: A Newsy Coincident Index of the Business Cycle," *Journal of Business and Economic Statistics*, 38, 393–409. [1,11]

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [3]

Uematsu, Y., and Tanaka, S. (2019), "High-dimensional Macroeconomic Forecasting and Variable Selection Via Penalized Regression," *Econometrics Journal*, 22, 34–56. [2]

van de Geer, S. (2016), *Estimation and Testing Under Sparsity: École d'Été de Probabilités de Saint-Flour XLV–2015* (Vol. 2159). Springer. [6]

Wong, K. C., Li, Z., and Tewari, A. (2020), "LASSO Guarantees for β-mixing Heavy Tailed Time Series," *Annals of Statistics*, 48, 1124–1142. [2]

Wu, W. B. (2005), "Nonlinear System Theory: Another Look at Dependence," *Proceedings of the National Academy of Sciences*, 102, 14150–14154. [2]

Wu, W.-B., and Wu, Y. N. (2016), "Performance Bounds for Parameter Estimates of High-dimensional Linear Models With Correlated Errors," *Electronic Journal of Statistics*, 10, 352–379. [2]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [4]