Panel Data Nowcasting in a Data-Rich Environment: The Case of Price-Earnings Ratios*

Andrii Babii† Ryan T. Ball‡ Eric Ghysels§ Jonas Striaukas¶

August 24, 2022

Abstract

This paper uses structured machine learning regressions for nowcasting with panel data consisting of series sampled at different frequencies. Motivated by the problem of predicting corporate earnings for a large cross-section of firms with macroeconomic, financial, and news time series sampled at different frequencies, we focus on the sparse-group LASSO regularization which can take advantage of the mixed frequency time series panel data structures. Our empirical results show the superior performance of our machine learning panel data regression models over analysts' predictions, forecast combinations, firm-specific time series regression models, and standard machine learning methods.

Keywords: Corporate earnings, nowcasting, high-dimensional panels, mixed frequency data, textual news data, sparse-group LASSO.

---

*We benefited from comments by Rudy De Winne, Geert D'Haene, Max Farrell, Christian Hafner, Peter Reinhard Hansen, Dacheng Xiu, and participants at the 2021 SoFiE UC San Diego conference, 26th International Panel Data Conference, Data Science and Machine Learning workshop at the University of Amsterdam, the 2022 IAAE Conference, King's College, London, and the 2022 Viennaâ€"Copenhagen Conference on Financial Econometrics. This work was in part completed when Jonas Striaukas was a Research Fellow at Fonds de la Recherche Scientifique â€" FNRS.

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

‡Stephen M. Ross School of Business, University of Michigan, 701 Tappan Street, Ann Arbor, MI 48109. Email: rtball@umich.edu.

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

¶Department of Finance, Copenhagen Business School, Frederiksberg, Denmark. Email: jonas.striaukas@gmail.com.

1 Introduction

Nowcasting is intrinsically a mixed frequency data problem as the object of interest is a low-frequency data series — observed say quarterly — 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 which 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 techniques (in particular the Kalman filter, see Bańbura, Giannone, Modugno, and Reichlin (2013) for a recent survey).

Things get more complicated when we are operating in a data-rich environment *and* we have many target variables. Put differently, we are no longer interested in nowcasting a single key series such as the GDP growth where we could devote a lot of resources to that particular series. A good example is corporate earnings nowcasting for a large cross-section of corporate firms. The fundamental value of equity shares is determined by the discounted value of future payoffs. Every quarter investors get a glimpse of firms' potential payoffs with the release of corporate earnings reports. In a data-rich environment, stock analysts have many indicators regarding future earnings that are available much more frequently. Ball and Ghysels (2018) took a first stab at automating the process using MIDAS regressions. Since their original work, much progress has been made on machine learning (ML) regularized mixed frequency regression models.

In the context of earnings, we are potentially dealing with a large set of individual firms for which there are many predictors. From a practical point of view, this is clearly beyond the realm of nowcasting using state space models. In the current paper, we significantly expand the tools of nowcasting in a data-rich environment by exploiting panel data structures. Panel data regression models are well suited for the firm-level data analysis as both the time series and cross-sectional dimensions can be exploited. In such models, time-invariant firm-specific effects are typically used to capture cross-sectional heterogeneity in the data. This is combined with regularized regression machine learning methods which are becoming increasingly popular in economics and finance as a flexible way to model predictive relationships via variable selection. We focus on the panel data regressions in a high-dimensional data setting where the number of covariates could be large and potentially exceed the available sample size. This may happen when the number of firm-specific characteristics, such as textual analysis news data or firm-level stock returns, is large, and/or the number

of aggregates, such as market returns, macro data, etc., is large.

To the best of our knowledge, it is an open question of how to implement nowcasting in such a data-rich environment of high-dimensional mixed-frequency panels. For instance, Khalaf, Kichian, Saunders, and Voia (2021) consider low-dimensional dynamic mixed frequency panel data models but do not deal with high-dimensional data situations in the context of nowcasting or forecasting. Similarly, Fosten and Greenaway-McGrevy (2019) consider nowcasting with a mixed-frequency VAR panel data model, but not in the context of a high-dimensional data-rich environment that we are interested in here. Babii, Ball, Ghysels, and Striaukas (2022) introduce the sparse-group LASSO (sg-LASSO) regularization machine learning methods for heavy-tailed dependent panel data regressions potentially sampled at different time series frequencies. They derive oracle inequalities for the pooled and fixed effects models, the debiased inference for pooled regression, and consider an application to the Granger causality testing. In this paper, we explore how to use their framework for nowcasting large panels of low-frequency time series.

We focus on nowcasting current quarter firm-specific price-earnings ratios (henceforth P/E ratios). This means we focus on evaluating model-based within-quarter predictions for very short horizons. It is widely acknowledged that P/E ratios are a good indicator of the future performance of a company and, therefore, are used by analysts and investment professionals to base their decisions on which stocks to pick for their investment portfolios. Typically investors rely on consensus forecasts of earnings made by a pool of analysts. We, therefore, choose such consensus forecasts as the benchmark for our proposed machine learning methods. Ball and Ghysels (2018) and Carabias (2018) documented that analysts tend to focus on their firm/industry when making earnings predictions while not fully taking into account the macroeconomic events affecting their firm/industry. Babii, Ball, Ghysels, and Striaukas (2022) tested formally in a high-dimensional data setting the hypothesis that systematic and predictable errors occur in analyst forecasts and confirmed empirically that they *leave money on the table*. The analysis in the current paper is therefore an logical extension of this prior work. In addition, we also compare our proposed new methods with the MIDAS regression forecast combination approach used by Ball and Ghysels (2018) as well as a simple random walk model.

Our high-frequency regressors include traditional macro and financial series as well as non-standard series generated by textual analysis of financial news. We consider structured pooled and fixed effects sg-LASSO panel data regressions with mixed frequency data (sg-LASSO MIDAS). By "structured" we mean that the ML procedure is set up such that it recognizes the time series and panel structure of the

data. This is in contrast to standard ML rooted in a tradition of i.i.d. covariates. As a representative example of standard ML, we will consider the elastic net type of estimators.

The fixed effects estimator yields sparser models compared to pooled regressions with revenue growth and the first lag of the dependent variable selected throughout the out-of-sample period. The BAA minus AAA bond yield spread, firm-level volatility, and news textual analysis aggregate event sentiment index are also selected very frequently. Our results show the superior performance of sg-LASSO MIDAS compared to (a) analysts' (consensus) predictions, (b) forecast combination methods, and (c) firm-specific time series regression models. Besides, the sg-LASSO MIDAS regressions perform better than standard machine learning panel data regressions with elastic net regularization.

The paper is organized as follows. Section 2 introduces the models and estimators. A simulation study reporting the finite sample nowcasting performance of our proposed methods appears in Section 3. The results of our empirical application analyzing price-earnings ratios for a panel of individual firms are reported in Section 4. Section 5 concludes. All technical details and detailed data descriptions appear in the Appendix and the Online Appendix.

2 High-dimensional mixed frequency panel data models

The target variables are a panel of low-frequency series (quarterly in our application) denoted $y_{i,t}$ for $t \in [T]$ (sample size T quarters) and a cross-section $i \in [N]$ (in our application a selection of individual firms).[1] Let K be the total number of time-varying predictors $\{x_{i,t-(j-1)/m,k} : i \in [N], t \in [T], j \in [m], k \in [K]\}$ possibly measured at some higher frequency with m observations for every low-frequency period $t \in [T]$ and every entity $i \in [N]$. Consider the following mixed frequency panel data regression

$$y_{i,t+h} = \alpha_i + \sum_{k=1}^{K} \psi(L^{1/m}; \beta_k) x_{i,t,k} + u_{i,t},$$

[1] For a positive integer p, we put $[p] = \{1, \ldots, p\}$.

where $h \geq 0$ is the prediction horizon, α_i is the entity-specific intercept, and

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

is a high-frequency lag polynomial with $\beta_k = (\beta_{1,k}, \ldots, \beta_{m,k})^\top \in \mathbf{R}^m$. More generally, the frequency can also be specific to the predictor $k \in [K]$, in which case we would have m_k instead of m observations for covariate k. In addition, we can also absorb the (low-frequency) lags of $y_{i,t}$ in covariates. A large number of predictors K with a potentially large number of high-frequency measurements m can be a rich source of predictive information, yet at the same time, estimating $N + m \times K$ parameters is costly and may reduce the predictive performance in small samples.

We follow the MIDAS ML literature, see Babii, Ghysels, and Striaukas (2021, 2022), and estimate a weight function ω parameterized by $\beta_k \in \mathbf{R}^L$ with $L < m$

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

where

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

and $(w_l)_{l \geq 0}$ is a collection of L approximating functions, called the *dictionary*. An example of a dictionary is the set of orthogonal Legendre polynomials on $[0,1]$ (see for instance Babii, Ghysels, and Striaukas (2022) for further details). The linear in parameters dictionaries map the MIDAS regression to a standard linear regression framework. In particular, define $\mathbf{x}_i = (X_{i,1}W, \ldots, X_{i,K}W)$, where for each $k \in [K]$, $X_{i,k} = (x_{i,t-(j-1)/m,k})_{t \in [T], j \in [m]}$ is a $T \times m$ matrix of predictors and $W = (w_l((j-1)/m)/m)_{j \in [m], 0 \leq l \leq L-1}$ is an $m \times L$ matrix corresponding to the dictionary $(w_l)_{l \geq 0}$. In addition, let $\mathbf{y}_i = (y_{i,1+h}, \ldots, y_{i,T+h})^\top$ and $\mathbf{u}_i = (u_{i,1}, \ldots, u_{i,T})^\top$. Then the regression equation after stacking time series observations for each $i \in [N]$ is

$$\mathbf{y}_i = \iota \alpha_i + \mathbf{x}_i \beta + \mathbf{u}_i,$$

where $\iota \in \mathbf{R}^T$ is the all-ones vector and $\beta \in \mathbf{R}^{LK}$ is a vector of slope coefficients. Lastly, put $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_N^\top)^\top$, $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_N^\top)^\top$, and $\mathbf{u} = (\mathbf{u}_1^\top, \ldots, \mathbf{u}_N^\top)^\top$. Then the regression equation after stacking all cross-sectional observations is

$$\mathbf{y} = B\alpha + \mathbf{X}\beta + \mathbf{u},$$

where $B = I_N \otimes \iota$, I_N is $N \times N$ identity matrix, and $\otimes$ is the Kronecker product.

Given that the number of potential predictors K can be large, additional regularization can improve the predictive performance in small samples. To that end, we take advantage of the sg-LASSO regularization that was shown to be attractive for individual time series ML regressions in Babii, Ghysels, and Striaukas (2022). The fixed effects sg-LASSO estimator $\hat{\rho} = (\hat{\alpha}^\top, \hat{\beta}^\top)^\top$ solves

$$\min_{(a,b)\in\mathbf{R}^{N+p}} \|\mathbf{y} - Ba - \mathbf{X}b\|_{NT}^2 + 2\lambda\Omega(b), \tag{2}$$

where Ω is the sg-LASSO regularizing functional. It is worth stressing that the design matrix $\mathbf{X}$ does not include the intercept and that we do not penalize the fixed effects which are typically not sparse. In addition, $\|.\|_{NT}^2 = |.|^2/(NT)$ is the empirical norm and

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

is a regularizing functional. It is a linear combination of the ℓ_1 LASSO and $\ell_{2,1}$ group LASSO norms. Note that for a group structure $\mathcal{G}$ described as a partition of $[p] = \{1, 2, \ldots, p\}$, the group LASSO norm is computed as $\|b\|_{2,1} = \sum_{G\in\mathcal{G}} |b_G|_2$, while $|.|_q$ denotes the usual ℓ_q norm. The group LASSO penalty encourages sparsity between groups whereas the ℓ_1 LASSO norm promotes sparsity within groups and allows us to learn the shape of the MIDAS weights from the data. The parameter $\gamma \in [0, 1]$ determines the relative weights of the ℓ_1 (sparsity) and the $\ell_{2,1}$ (group sparsity) norms, while the amount of regularization is controlled by the regularization parameter $\lambda \geq 0$.

In the Introduction, we called our approach structured ML because the group structure allows us to embed the time series structure of the data. More specifically, these structures are represented by groups covering lagged dependent variables and groups of lags for a single (high-frequency) covariate. Throughout the paper, we assume that groups have fixed size, and the group structure is known by the econometrician. Both are reasonable assumptions to make in the context of our empirical application.

For pooled regressions, we assume that all entities share the same intercept parameter $\alpha_1 = \cdots = \alpha_N = \alpha$. The pooled sg-LASSO estimator $\hat{\rho} = (\hat{\alpha}, \hat{\beta}^\top)^\top$ solves

$$\min_{r=(a,b)\in\mathbf{R}^{1+p}} \|\mathbf{y} - a\iota - \mathbf{X}b\|_{NT}^2 + 2\lambda\Omega(r). \tag{3}$$

Pooled regressions are attractive since the effective sample size NT can be huge, yet the heterogeneity of individual time series may be lost. If the underlying series have

a substantial heterogeneity over $i \in [N]$, then taking this into account might reduce the projection error and improve the predictive accuracy.

Babii, Ball, Ghysels, and Striaukas (2022) provide the theoretical analysis of predictive performance of regularized panel data regressions with the sg-LASSO regularization, including the standard LASSO and the group LASSO regularizations as well as generic high-dimensional panels not involved mixed frequency data regularized with the sg-LASSO penalty function as special cases. Finally, Babii, Ball, Ghysels, and Striaukas (2022) also develop the debiased inferential methods and Granger causality tests for pooled panel data regressions.

3 Monte Carlo experiments

It is not clear that the aforementioned theory is of practical use in the context of nowcasting using modestly sized samples of data. For this reason, we investigate in this section the finite sample nowcasting performance of the machine learning methods covered so far. We consider the standard (unstructured) elastic net with UMIDAS (see Foroni, Marcellino, and Schumacher (2015)) and sg-LASSO with MIDAS. Both methods require selecting two tuning parameters λ and γ. In the case of sg-LASSO, γ is the relative weight of LASSO and group LASSO penalties while in the case of the elastic net γ interpolates between LASSO and ridge. In both cases we report results on a grid $\gamma \in \{0, 0.2, \dots, 1\}$.

In addition to evaluating the performance over the grid of γ tuning parameter values, we need to select the λ tuning parameter. To do so, we consider several approaches. First, we adapt the K-fold cross-validation to the panel data setting. To that end, we resample the data by blocks respecting the time-series dimension and creating folds based on cross-sectional units instead of the pooled sample. We use the 5-fold cross-validation both in the simulation experiments and the empirical application. We also consider the following three information criteria: BIC, AIC, and corrected AIC (AICc) of Hurvich and Tsai (1989). Assuming that $y_{i,t}|x_{i,t}$ are i.i.d. draws from $N(\alpha_i + x_{i,t}^\top \beta, \sigma^2)$, the log-likelihood of the sample is

$$\mathcal{L}(\alpha, \beta, \sigma^2) \propto -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - \alpha_i - x_{i,t}^\top \beta)^2.$$

Then, the BIC criterion is

$$\text{BIC} = \frac{\|\mathbf{y} - \hat{\mu} - \mathbf{X}\hat{\beta}\|_{NT}^2}{\hat{\sigma}^2} + \frac{\log(NT)}{NT} \times df,$$

where df denotes the degrees of freedom, $\hat{\sigma}^2$ is a consistent estimator of σ^2, $\hat{\mu} = \hat{\alpha}\iota$ for the pooled regression, and $\hat{\mu} = B\hat{\alpha}$ for fixed effects regression. The degrees of freedom are estimated as $\widehat{df} = |\hat{\beta}|_0 + 1$ for the pooled regression and $\widehat{df} = |\hat{\beta}|_0 + N$ for the fixed effects regression, where $|.|_0$ is the ℓ_0-norm defined as a number of non-zero coefficients; see Zou, Hastie, and Tibshirani (2007) for more details. The AIC is computed as

$$\text{AIC} = \frac{\|\mathbf{y} - \hat{\mu} - \mathbf{X}\hat{\beta}\|^2_{NT}}{\hat{\sigma}^2} + \frac{2}{NT} \times \widehat{df},$$

and the corrected Akaike information criteria is

$$\text{AICc} = \frac{\|\mathbf{y} - \hat{\mu} - \mathbf{X}\hat{\beta}\|^2_{NT}}{\hat{\sigma}^2} + \frac{2\widehat{df}}{NT - \widehat{df} - 1}.$$

The AICc is typically a better choice when p is large relative to the sample size. We report the results for each of the tuning parameter selection criteria for λ, along the grid choice for γ.

3.1 Simulation Design

To assess the predictive performance of pooled panel data models, we simulate the data from the following DGP:

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

where $i \in [N]$, $t \in [T]$, α is the common intercept, $1/m \sum_{j=1}^{m} \omega((j-1)/m; \beta_k)$ the weight function for k-th high-frequency covariate and the error term is either $u_{i,t} \sim_{i.i.d.} N(0,1)$ or $u_{i,t} \sim_{i.i.d.}$ student-$t(5)$. The DGP corresponds to the target variable of interest $y_{i,t}$ driven by two autoregressive lags augmented with high-frequency series, and therefore is a pooled MIDAS panel data model.

We set $\rho_1 = 0.4$, $\rho_2 = 0.01$, and the number of relevant high-frequency regressors $K = 6$. We are interested in a quarterly/monthly data mix, and use four quarters of data for the high-frequency regressors so that $m = 12$, which covers four low-frequency lags of each high-frequency regressor. The high-frequency regressors are generated as K i.i.d. realizations of the univariate autoregressive (AR) process $x_h = \rho x_{h-1} + \varepsilon_h$, where $\rho = 0.6$ and either $\varepsilon_h \sim_{i.i.d.} N(0,1)$ or $\varepsilon_h \sim_{i.i.d.}$ student-$t(5)$, where h denotes the high-frequency sampling. We rely on a commonly used weighting scheme in the MIDAS literature, namely $\omega(s; \beta_k)$ for $k = 1, 2, \ldots, 6$ are determined

by beta densities respectively equal to Beta$(1,3)$ for $k=1,4$, Beta$(2,3)$ for $k=2,5$, and Beta$(2,2)$ for $k=3,6$; see Ghysels, Sinko, and Valkanov (2007) or Ghysels and Qian (2019), for further details. The MIDAS regressions are estimated using Legendre polynomials of degree $L=3$. Lastly, we simulate the intercepts as $\alpha \sim$ Uniform$(-4,4)$.

We also consider DGPs featuring fixed effects. They are identical to the pooled MIDAS panel data model except for the common intercept α which is replaced by α_i and the individual fixed effects are simulated as $\alpha_i \sim_{\text{i.i.d}}$ Uniform$(-4,4)$ and are kept fixed throughout the experiment.

For the *Baseline scenario*, in the estimation procedure we add 24 noisy covariates which are generated in the same way as the relevant covariates, use 4 low-frequency lags and the error terms $u_{i,t}$ and ε_h are Gaussian. In the student-$t(5)$ scenario we replace the Gaussian error terms with a student-$t(5)$ distribution while in the *large dimensional* scenario we add 94 noisy covariates. For each scenario, we simulate $N=25$ i.i.d. time series of length $T=50$; next we increase the cross-sectional dimension to $N=75$ and time series to $T=100$.

Finally, the thought experiment in the simulation design is one where low-frequency data up to time $t-1$ and the first high-frequency observations during time t are available. The goal is to nowcast $y_{i,t}$ $\forall i \in [N]$, with lagged low-frequency data and real-time data $x_{i,t-(m-1)/m,k}$. The nowcaster of course does not know which of the covariates are relevant nor does she know the parameters of the prediction rule. We will call this scheme "one-step ahead" nowcasts. In the context of our application with monthly/quarterly data, this means that we are looking at the first monthly nowcast for quarter t and call it one-step ahead.

3.2 Simulation results

Table 1 covers the average mean squared forecast errors for one-step ahead nowcasts for the Baseline scenario. We report results for pooled panel data (left block) and fixed effects (right block) estimators.[2]

First, for all DGPs and both estimators, structured sg-LASSO-MIDAS performs better compared to unstructured elastic net UMIDAS. In the case of sg-LASSO-MIDAS the best performance is achieved for $\gamma \notin \{0,1\}$ for both pooled panel data and fixed effects cases, while $\gamma=0$, i.e. ridge regression, seems to dominate in the

[2]Results for the student-$t(5)$ and large dimensional DGPs are reported in the Online Appendix Tables OA.1-OA.2.

case of elastic net UMIDAS for both the pooled and fixed effects cases. For the student-$t(5)$ and large dimensional DGP, we observe a decrease in the performance for all methods. However, the decrease in the performance is larger for the student-$t(5)$ DGP, suggesting that heavy-tailed data may have a stronger impact on the performance of the estimators. Moreover, when comparing results for $\gamma = 1$, i.e., the LASSO case, we see a notable improvement in prediction quality when we apply MIDAS polynomials compared with UMIDAS across all scenarios.

For the pooled panel data case, increasing N from 25 to 75 seems to have a larger positive impact on the performance than an increase in the time-series dimension from $T = 50$ to $T = 100$. The difference appears to be larger for student-$t(5)$ and large dimensional DGPs and/or for the elastic net case. Turning to the fixed effects results, the differences seem to be even sharper, in particular for student-$t(5)$ and large dimensional DGPs.

When comparing results across the different model selection methods, i.e., cross-validation and the three information criteria, we find that almost always cross-validation leads to smaller prediction errors in both pooled and fixed effects panel data cases. Notably, the gains appear to be larger for the large N and T values. Comparing BIC, AIC, and AICc information criteria, the results appear to be similar for AIC and AICc across DGPs and different sample sizes, while BIC performance is slightly worse than AIC and AICc.

4 Nowcasting price-earnings ratios

Ball and Ghysels (2018), Carabias (2018) and Babii, Ball, Ghysels, and Striaukas (2022) documented that analysts make systematic and predictable errors. In this section, we therefore consider nowcasting the P/E ratios of 210 US firms using a set of predictors that are sampled at mixed frequencies. We use 24 predictors, including traditional macro and financial series as well as non-traditional series from textual analysis of financial news. We apply pooled and individual fixed effects sg-LASSO MIDAS panel data models and compare them with several benchmark models, which include a random walk (RW) model, an elastic net UMIDAS model, and analysts' consensus forecasts.

We also compute predictions using individual-firm high-dimensional time series regressions and provide results for several choices of the tuning parameter. Moreover, our analysis includes results for sg-LASSO MIDAS panel data models which include the median consensus analysts' predictions. Adding the consensus forecast as a regressor allows us to address besides the question of ML versus analysts also the topic

		Pooled panel data						Fixed effects					
	N/T	$\gamma = 0$	0.2	0.4	0.6	0.8	1	$\gamma = 0$	0.2	0.4	0.6	0.8	1
							Cross-validation						
	25/50	1.213	1.208	1.206	1.206	1.213	1.228	1.229	1.226	1.223	1.222	1.222	1.227
sg-LASSO	75/50	1.163	1.161	1.160	1.160	1.163	1.171	1.171	1.169	1.169	1.168	1.169	1.169
	25/100	1.181	1.182	1.181	1.179	1.173	1.175	1.170	1.169	1.168	1.167	1.168	1.171
	25/50	1.237	1.232	1.232	1.232	1.234	1.234	1.267	1.268	1.268	1.269	1.269	1.269
elnet-U	75/50	1.178	1.168	1.168	1.168	1.168	1.168	1.188	1.188	1.189	1.190	1.195	1.196
	25/100	1.185	1.180	1.180	1.180	1.180	1.181	1.187	1.186	1.186	1.186	1.186	1.186
							BIC						
	25/50	1.556	1.374	1.365	1.378	1.398	1.428	1.674	1.588	1.528	1.540	1.566	1.668
sg-LASSO	75/50	1.211	1.210	1.211	1.212	1.218	1.256	1.257	1.230	1.262	1.262	1.304	1.398
	25/100	1.225	1.225	1.230	1.258	1.274	1.322	1.463	1.342	1.315	1.313	1.360	1.421
	25/50	1.647	1.687	1.703	1.733	1.744	1.750	1.828	2.157	2.462	2.456	2.507	2.510
elnet-U	75/50	1.290	1.311	1.318	1.322	1.323	1.325	1.464	1.575	1.693	1.718	1.720	1.720
	25/100	1.345	1.360	1.378	1.391	1.402	1.409	1.512	1.768	1.889	1.921	1.930	1.939
							AIC						
	25/50	1.294	1.289	1.286	1.288	1.296	1.317	1.350	1.337	1.334	1.335	1.342	1.382
sg-LASSO	75/50	1.189	1.185	1.183	1.181	1.175	1.172	1.211	1.208	1.206	1.205	1.212	1.232
	25/100	1.211	1.206	1.204	1.205	1.212	1.227	1.242	1.232	1.229	1.225	1.224	1.260
	25/50	1.382	1.382	1.382	1.382	1.384	1.384	1.420	1.490	1.510	1.521	1.528	1.532
elnet-U	75/50	1.211	1.199	1.198	1.199	1.199	1.199	1.237	1.271	1.266	1.265	1.264	1.264
	25/100	1.257	1.271	1.267	1.266	1.265	1.265	1.268	1.288	1.292	1.294	1.294	1.295
							AICc						
	25/50	1.296	1.290	1.287	1.289	1.298	1.321	1.358	1.342	1.340	1.341	1.351	1.393
sg-LASSO	75/50	1.189	1.185	1.183	1.182	1.175	1.173	1.214	1.208	1.207	1.206	1.212	1.234
	25/100	1.212	1.207	1.204	1.205	1.213	1.228	1.244	1.233	1.230	1.227	1.226	1.264
	25/50	1.382	1.391	1.389	1.390	1.391	1.392	1.420	1.518	1.540	1.550	1.559	1.562
elnet-U	75/50	1.211	1.199	1.200	1.200	1.200	1.200	1.237	1.272	1.268	1.267	1.266	1.266
	25/100	1.257	1.272	1.267	1.266	1.266	1.265	1.268	1.293	1.297	1.300	1.300	1.302

Table 1: The table reports simulation results for nowcasting accuracy for pooled and fixed effects estimators for the baseline DGP for the sg-LASSO-MIDAS (rows sg-LASSO) and elastic net UMIDAS (rows elnet-U). We vary the cross-sectional dimension $N \in \{25, 75\}$ and time series dimension $T \in \{50, 100\}$. We report results for 5-fold cross-validation, BIC, AIC, AICc information criteria λ tuning parameter calculation methods and for a grid of $\gamma \in \{0, 0.2, \ldots, 1\}$ tuning parameter.

of a combined ML/analyst nowcasts — a theme explored by Ball and Ghysels (2018). Our analysis includes formal significance testing of predictors in the augmented sg-LASSO MIDAS panel data model, which allows us to determine whether analysts use all relevant information that is available to them.

Lastly, we provide results for the low-dimensional single-firm MIDAS regressions using forecast combination techniques used by Andreou, Ghysels, and Kourtellos (2013) and Ball and Ghysels (2018). The latter is particularly relevant as it also deals with nowcasting price-earnings ratios. The forecast combination methods consist of estimating ARDL-MIDAS regressions for each of the high-frequency covariates separately. In our case this leads to 24 predictions, corresponding to the number of predictors. Then a combination scheme, typically of the discounted mean squared error type, produces a single nowcast with time-varying combination weights. One could call this a pre-machine learning large dimensional approach and it is interesting to assess how it compares with the regularized MIDAS panel regression machine learning approach introduced in the current paper.

The remainder of the section is structured as follows. We start with a short review of the data, with more detailed descriptions and tables appearing in the Online Appendix Section OA.3, followed by a summary of the methods and empirical results.

4.1 Data description

The full sample consists of observations between the 1^{st} of January, 2000 and the 30^{th} of June, 2017. Due to the lagged dependent variables in the models, our effective sample starts in the third fiscal quarter of 2000. We use the first 25 observations for the initial sample, and use the remaining 42 observations for evaluating the out-of-sample forecasts, which we obtain by using an expanding window forecasting scheme. We collect data from CRSP and I/B/E/S to compute the quarterly P/E ratios and firm-specific financial covariates; RavenPack is used to compute daily firm-level textual-analysis-based data; real-time monthly macroeconomic series are from the FRED-MD dataset, see McCracken and Ng (2016) for more details; FRED is used to compute daily financial markets data and, lastly, monthly news attention series extracted from the *Wall Street Journal* articles are retrieved from Bybee, Kelly, Manela, and Xiu (2021).[3] Online Appendix Section OA.3 provides a detailed description of the data sources.[4]

[3]The dataset is publicly available at http://www.structureofnews.com/.

[4]In particular, firm-level variables, including P/E ratios, are described in Online Appendix Table OA.7, and the other predictor variables in Online Appendix Table OA.8. The list of all firms we

P/E ratio and analysts' forecasts sample construction. Our target variable is the P/E ratio for each firm. To compute it, we use CRSP stock price data and I/B/E/S earnings data. Earnings data are subject to release delays of 1 to 2 months depending on the firm and quarter. Therefore, to reflect the real-time information flow, we compute the target variable using stock prices that are available in real-time. We also take into account that different firms have different fiscal quarters, which also affects the real-time information flow.

For example, suppose for a particular firm the fiscal quarters are at the end of the third month in a quarter, i.e. end of March, June, September, and December. The consensus forecast of the P/E ratio is computed using the same end-of-quarter price data which is divided by the earnings consensus forecast value. The consensus is computed by taking all individual prediction values up to the end of the quarter and aggregating those values by taking either the mean or the median. To compute the target variable, we adjust for publication lags and use prices of the publication date instead of the end of fiscal quarter prices. More precisely, suppose we predict the P/E ratio for the first quarter. Earnings are typically published with 1 to 2 months delay; say for a particular firm the data is published on the $25th$ of April. In this case, we record the stock price for the firm on $25th$ of April, and divide it by the earnings announced on that date.

4.2 Models and main results

To compute forecasts, we estimate several regression models. First, we estimate the individual sg-LASSO MIDAS regressions for each firm $i = 1, \ldots, N$, which in Table 2 we refer to as *Individual*,

$$\mathbf{y}_i = \iota\alpha_i + \mathbf{x}_i\beta_i + \mathbf{u}_i,$$

where the firm-specific predictions are computed as $\hat{y}_{i,t+1} = \hat{\alpha}_i + x_{i,t+1}^{\top}\hat{\beta}_i$. As noted in Section 2, $\mathbf{x}_i$ contains lags of the low-frequency target variable and high-frequency covariates to which we apply Legendre polynomials of degree $L = 3$.

Next, we estimate the following pooled and fixed effects sg-LASSO MIDAS panel data models

$$\begin{aligned} \mathbf{y} &= \alpha\iota + \mathbf{X}\beta + \mathbf{u} \quad \text{Pooled} \\ \mathbf{y} &= B\alpha + \mathbf{X}\beta + \mathbf{u} \quad \text{Fixed Effects} \end{aligned}$$

consider in our analysis appears in Online Appendix Table OA.9.

and compute predictions as

$$\begin{aligned} \hat{y}_{i,t+1} &= \hat{\alpha} + x_{i,t+1}^{\top}\hat{\beta} \quad \text{Pooled} \\ \hat{y}_{i,t+1} &= \hat{\alpha}_i + x_{i,t+1}^{\top}\hat{\beta} \quad \text{Fixed Effects.} \end{aligned}$$

We benchmark firm-specific and panel data regression-based nowcasts against two simple alternatives. First, we compute forecasts for the RW model as

$$\hat{y}_{i,t+1} = y_{i,t}.$$

Second, we consider predictions of P/E implied by analysts' earnings nowcasts using the information up to time $t+1$, i.e.

$$\hat{y}_{i,t+1} = \bar{y}_{i,t+1},$$

where $\bar{y}$ indicates that the forecasted P/E ratio is based on consensus earnings forecasts made at the end of the $t+1$ quarter, and the stock price is also taken at the end of $t+1$. Recall that the actual earnings are only available two months after the end-of-quarter $t+1$ as explained earlier in the section.

To measure the forecasting performance, we compute the mean squared forecast errors (MSE) for each method. Let $\bar{\mathbf{y}}_i = (y_{i,T_{is}+1}, \ldots, y_{i,T_{os}})^{\top}$ represent the out-of-sample realized P/E ratio values, where T_{is} and T_{os} denote the last in-sample observation for the first prediction and the last out-of-sample observation respectively, and let $\hat{\mathbf{y}}_i = (\hat{y}_{i,t_{is}+1}, \ldots, \hat{y}_{i,t_{os}})$ collect the out-of-sample forecasts. Then, the mean squared forecast errors are computed as

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{T - T_{is} + 1}(\bar{\mathbf{y}}_i - \hat{\mathbf{y}}_i)^{\top}(\bar{\mathbf{y}}_i - \hat{\mathbf{y}}_i).$$

The main results for pooled panel data and fixed effects sg-LASSO MIDAS regressions are reported in Table 2, while additional results for longer horizon predictions, unstructured LASSO estimators and the forecast combination approach appear in Online Appendix Tables OA.3-OA.5.

The first entries to Table 2 show that analysts-based predictions, both median and mean, have much larger mean squared forecast errors (MSEs) compared to model-based predictions. This is also the case for the RW predictions. The sharp increase in quality of model- versus analyst-based predictions indicates the usefulness of machine learning methods to nowcast P/E ratios, see Tables 2 Panel A1-D1 and OA.4. A better performance is achieved for almost all machine learning methods – single

RW	MSE An.-mean	MSE An.-median		MIDAS ML					
2.331	2.339	2.088	$\gamma =$	0	0.2	0.4	0.6	0.8	1
						sg-LASSO MIDAS			
						Panel A1. Cross-validation			
			Individual	1.545	1.551	1.567	1.594	1.614	1.606
			Pooled	1.459	1.456	**1.455**	1.456	1.455	1.459
			Fixed Effects	1.500	1.489	1.487	1.501	1.480	1.489
						Panel B1. BIC			
			Individual	1.657	1.634	1.609	1.543	1.561	1.610
			Pooled	1.482	1.498	1.491	1.495	1.493	1.483
			Fixed Effects	1.515	1.496	**1.472**	1.512	1.483	1.476
						Panel C1. AIC			
			Individual	1.622	1.589	1.560	1.603	1.674	1.688
			Pooled	1.494	1.492	1.488	1.487	1.490	1.492
			Fixed Effects	1.504	1.487	1.486	1.504	**1.479**	1.489
						Panel D1. AICc			
			Individual	2.025	2.122	2.272	2.490	2.923	3.255
			Pooled	1.494	1.484	1.488	1.487	1.490	1.492
			Fixed Effects	1.491	1.488	1.486	1.504	**1.479**	1.489
						sg-LASSO MIDAS augmented with An.-median			
						Panel A2. Cross-validation			
			Individual	1.528	1.542	1.552	1.552	1.537	1.534
			Pooled	1.422	1.419	1.417	1.418	1.420	1.425
			Fixed Effects	1.385	1.385	**1.358**	1.364	1.370	1.362
						Panel B2. BIC			
			Individual	1.638	1.610	1.584	1.566	1.506	1.508
			Pooled	1.453	1.425	1.398	1.425	1.453	1.447
			Fixed Effects	1.400	1.400	**1.372**	1.379	1.384	1.379
						Panel C2. AIC			
			Individual	1.618	1.580	1.565	1.577	1.621	1.610
			Pooled	1.453	1.453	1.482	1.483	1.486	1.488
			Fixed Effects	1.434	1.434	**1.405**	1.412	1.418	1.407
						Panel D2. AICc			
			Individual	1.618	1.580	1.565	1.577	1.621	1.610
			Pooled	1.453	1.453	1.482	1.483	1.486	1.488
			Fixed Effects	1.434	1.434	**1.405**	1.412	1.418	1.407

Table 2: Prediction results – The table reports the average over firms MSEs of out-of-sample predictions. The nowcasting horizon is the current quarter, i.e. we predict the P/E ratio using information up to the end of the current fiscal quarter. Block in Panel A1-D1 correspond to ML-only forecast errors while in Panel A2-D2 to ML models augmented with median consensus nowcasts. Each Panel A1-D1 and A2-D2 block represents different ways of calculating the tuning parameter λ. Bold entries are the best results in a block.

firm or panel data regressions – and all tuning parameter choices.[5] Table 2 also reports results for panel data ML methods augmented with median consensus analysts forecasts (see Panels A2-D2). Notably, it shows that the augmented models further improve upon ML-only models.

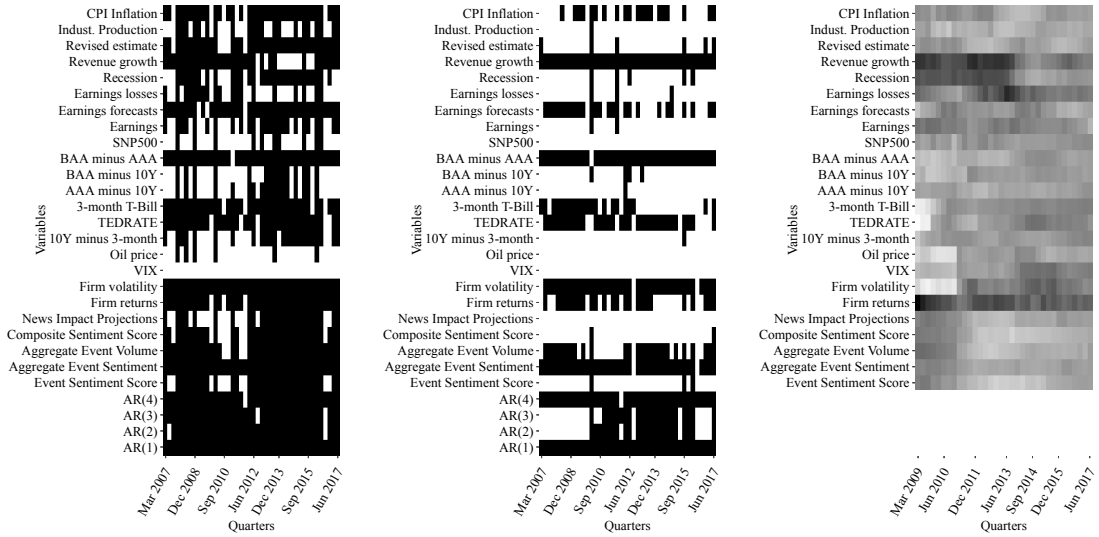
Turning to the comparison of model-based predictions, we see from the results in Table 2 that sg-LASSO MIDAS panel data models improve the quality of predictions in comparison to individual sg-LASSO MIDAS models irrespective of the γ weight or the tuning parameter choice. This indicates that panel data structures are relevant for nowcasting P/E ratios.[6] Among the panel data models, we observe that fixed effects regressions improve over the pooled regressions in most cases except when cross-validation is used, namely compare Panels A1 with Panel B1-D in Tables 2 and OA.5. The pooled model tuned by cross-validation seems to yield the best overall performance. In general, one can expect that cross-validation improves prediction performance over different tuning methods as it is directly linked to empirical risk minimization. In the case of fixed effects, however, we may lose the predictive gain due to the smaller samples with each fold used in estimating the model. Lastly, the best results per tuning parameter block seem to be achieved when $\gamma \notin \{0, 1\}$, indicating that both sparsity within the group and at the group level matters for prediction performance.

In Figure 1, we plot the sparsity patterns of the selected covariates for the two best-performing methods: (a) pooled sg-LASSO regressions, tuned using cross-validation with $\gamma = 0.4$, and (b) fixed effects sg-LASSO model with BIC tuning parameter and the same γ parameter. We also plot the forecast combination weights which are averaged over firms. The plots in Figure 1 reveal that the fixed effects estimator yields sparser models compared to pooled regressions, and the sparsity pattern is clearer. In the fixed-effects case, the revenue growth and the first lag of the dependent variable are selected throughout the out-of-sample period. BAA minus AAA bond yield spread, firm-level volatility, and the aggregate event sentiment index are also selected quite frequently. Similarly, these variables are selected in the pooled regression, but the pattern is less apparent. The forecast combination weights seem to yield similar, yet more dispersed patterns.[7] In this case, revenue growth and

[5]Similar findings for one-quarter ahead predictions are reported in Table OA.3. The unstructured panel data methods and the forecast combination approach also yield more accurate forecasts, see Online Appendix Table OA.4-OA.5. The latter confirms the findings of Ball and Ghysels (2018).

[6]We also report similar findings for unstructured estimators (see Table OA.4) and one quarter ahead forecasts (see Table OA.3).

[7]Note that forecast combination weights start in 2009 Q1 due to the first eight quarters being used as a pre-sample to estimate weights, see Ball and Ghysels (2018) for further details. Also, the



(a) Pooled sg-LASSO, $\gamma = 0.4$, cross-validation.

(b) Fixed effects sg-LASSO, $\gamma = 0.4$, BIC.

(c) Average forecast combination weights.

Figure 1: Sparsity patterns and forecast combination weights.

firm-level stock returns covariates obtain relatively larger weights compared to the rest of covariates, particularly for the first part of the out-of-sample period. Therefore, the gain of machine learning methods - both single-firm and panel data - can be associated with sparsity imposed on the regression coefficients.

In addition, it is worth noting that the textual news data analytics also appear in the models according to the results displayed in Figure 1. These are the Event Sentiment Score, Aggregate Event Sentiment, Aggregate Event Volume, Composite Sentiment Score, and News Impact Projections indices described in detail in Appendix Section OA.3. Among them, as already noted, the aggregate event sentiment index features most prominently in the sg-LASSO models. It is worth emphasizing that the time series of news data is sparse as many days are without firms-specific news. For such days, we impute zero values.[8] The nice property of our mixed frequency data treatment with dictionaries, imputing zeros also implies that non-zero

forecast combination weights figure does not contain autoregressive lags; all four lags are always included in all forecasting regressions.

[8]Note that all firm-level textual data series are centered around zero, where zero value means zero impact news, see Appendix Section OA.3. Therefore, imputing zeros is the same as assuming that no news on a given day implies zero impact news, which we believe is a reasonable assumption to make.

entries get weights with a decaying pattern for distant past values.[9]

Finally, Figure 2 shows the flexibility of our approach when dealing with high-dimensional MIDAS panel data models. First, we show that various shapes and forms of the weighting function can be estimated by applying Legendre polynomials over the high-frequency lags. For instance, the BAA minus 10-Year Treasury bond yield spread is estimated to have slowly decaying weights, while the TED rate covariate has a humped shape of the weights. Our approach provides a foundation for future research that focuses on the economic interpretations of the various MIDAS polynomial shapes (e.g., Ball (2013); Ball and Easton (2013); Ball and Gallo (2018)). Finally, our approach allows for the recovery of smooth lag functions for such series, even for daily textual news series that are sparse, see Figure 2 (a) and (e).

4.3 Significance test of nowcasts

To test for the superior forecast performance, we use the Diebold and Mariano (1995) test for the pool of P/E ratio nowcasts. We compare the median consensus forecasts versus panel data machine learning regressions with the smallest forecast error for pooled and fixed effects panel regressions and report the forecast accuracy test results in Table 3.

When testing the full sample of pooled nowcasts, the gain in prediction accuracy is not significant even though the MSEs are much lower for the panel data sg-LASSO regressions relative to the consensus forecasts. The result may not be surprising, however, as some firms have a large number of outliers. We report three additional columns where we pool the prediction based on the relative performance of machine learning methods versus analysts. First, we pool all errors for firms where sg-LASSO MIDAS and elastic net outperform the analysts' median consensus forecasts, i.e. has a smaller average prediction error. Second, we pool the errors where sg-LASSO MIDAS outperforms the analysts, but the elastic net does not. Lastly, we pool prediction errors where none of the methods outperforms analysts.[10] Â

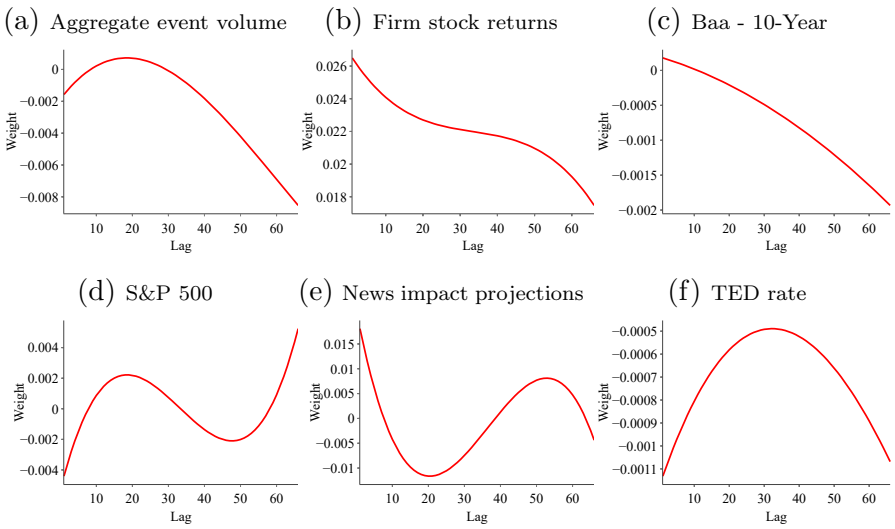
[9]Note that the imputed zeros will cancel any weights applied to the high-frequency data, such that the MIDAS polynomial will only pick the non-zero entries with the proper weight corresponding to the time lag. More specifically, recall we estimate a weight function ω parametrized by $\beta_k \in \mathbf{R}^L$ with $L < m$

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

We can see from the above equation that with imputed zeros only the non-zero entries will appear in the prediction formula with the weight associated with the time lag.

[10]We do not report results for the pool of firms for which the elastic net outperforms analysts

Daily Series



Monthly Series

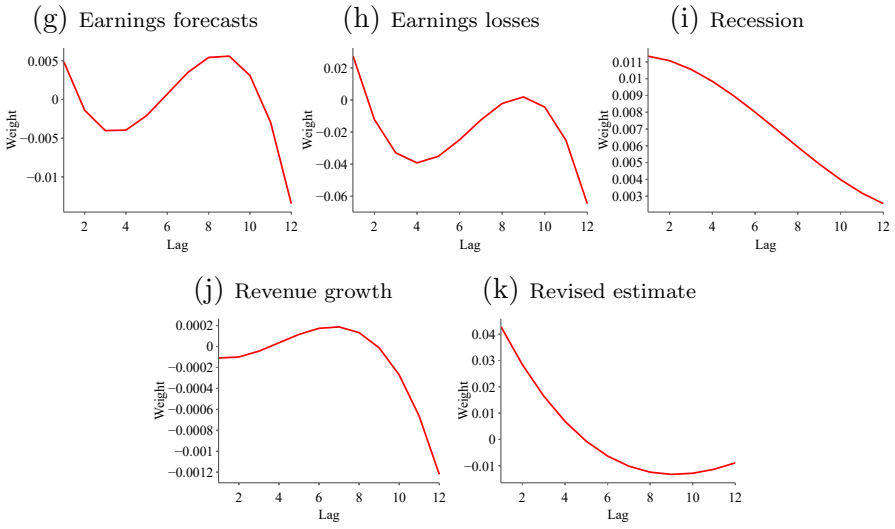


Figure 2: Weighting schemes for various covariates

The results reveal heterogeneous performance for sg-LASSO MIDAS and elastic net panel data regressions. First, for the pool of firms where both structured sg-LASSO MIDAS and unstructured elastic net outperform the analysts, the gains over the analysts' predictions are significant for both machine learning techniques. Second, for the firms where both methods yield less accurate forecasts compared to the analysts, the loss in prediction accuracy is also significant. Lastly, the portion of firms sg-LASSO outperforms analysts while elastic net does not yield significantly higher quality predictions for sg-LASSO and significantly worse for the elastic net.Â

	Full sample	sg-LASSO & elnet	sg-LASSO	none
		sg-LASSO		
Pooled	0.694	2.328	1.924	-2.738
Fixed Effects	0.672	2.319	1.681	-2.555
		Elastic net		
Pooled	0.656	2.299	-3.112	-2.698
Fixed Effects	0.656	2.314	-2.244	-2.571
		Number of firms		
Pooled	210	63	12	135
Fixed Effects	210	66	8	134

Table 3: Forecasting performance significance – The table reports the Diebold and Mariano (1995) test statistic for pooled nowcasts comparing machine learning panel data regressions with analysts' implied median consensus forecasts. We compare panel models that have the smallest forecast error per tuning parameter block in Table 2 (sg-LASSO-MIDAS) and Table OA.4 (elastic net or elastic net UMIDAS) for pooled and fixed effects regressions respectively. We report test statistics for a) all firms in column *Full sample*, b) pooled firms where both sg-LASSO and elastic net outperform analysts in column *sg-LASSO & elnet*, c) pooled firms where sg-LASSO outperforms analysts but elastic net does not in column *sg-LASSO* , and d) where none of the machine learning methods outperforms analysts' forecasts in column *none*.Â

Large differences in prediction accuracy for different pools of P/E ratios may relate to the heavy-tailedness of regression errors.[11] In Table 4, we report the maximum likelihood estimates of the degrees of freedom parameter of a student-t distribution for the in-sample residuals pooled as in Table 3.[12] The smaller values indicate that

and the sg-LASSO MIDAS does not, since there is only one such firm in the case of fixed effects regressions, while in the case of pooled regressions there are no such firms.

[11]Our theory applies to the tail behavior of covariates as well as regression errors. However, some of the covariates do not feature cross-sectional variation, which is why we focus only on the errors.

[12]We follow a parametric approach since the time series are relatively short, see however also

the tails are heavier, while the larger values correspond to lighter, closer to Gaussian, tails. In line with our theory, the results show that LASSO-type regressions yield much more accurate predictions when the residuals are less heavy-tailed. Interestingly, for the pool of firms where analysts' predictions are more accurate than both machine learning methods (column *none*), the tails of the residuals appear to be the heaviest.

Lastly, we report the Diebold and Mariano (1995) test statistic comparing whether the sg-LASSO MIDAS model combined with the median consensus nowcasts (An.-median) outperforms the analysts-only nowcasts (see Table 2 Panels A2-D2). Note that median consensus predictions are always selected by the sg-LASSO MIDAS throughout the out-of-sample period while other covariates retain a selection pattern similar to that of sg-LASS-MIDAS regressions reported in Figure 1. We pick the best panel model specification and compute the statistic of out-of-sample residuals. The statistic is 1.327, suggesting that the combined model and analysts' predictions seem to outperform analysts when using a one-sided 10% level test.

	Full sample	sg-LASSO & elnet	sg-LASSO	none
		sg-LASSO		
Pooled	4.803	7.413	5.497	4.217
Fixed Effects	4.871	6.966	5.003	4.321
		Elastic net		
Pooled	4.926	7.588	5.762	4.341
Fixed Effects	5.332	7.422	5.479	4.741
		Regressands		
	5.627	7.031	5.303	5.228
		Number of firms		
Pooled	210	63	12	135
Fixed Effects	210	66	8	134

Table 4: The heaviness of tails – The table reports the maximum likelihood estimate of the degree of freedom of *student-t* distribution of in-sample residuals. The results are reported for the models as in Table 3.Â

Online Appendix, Table OA.6 for the nonparametric tail index estimates.

4.4 Significance tests for predictors

The out-of-sample nowcasting analysis showed that combined machine learning and analysts nowcasts improve over ML-only and analysts-only predictions. We now assess the relative importance of each covariate in a combined panel data regression model by using the HAC-based inference.

To analyze the significance, we estimate the pooled sg-LASSO MIDAS panel data model on a full sample of the data. We compute the tuning parameter λ using 5-fold cross-validation and set $\gamma = 0.4$, i.e. the best tuning parameter choice, see Table 2 Panels A2-D2. The precision matrix is estimated using nodewise LASSO regressions, which are tuned using 5-fold cross-validation. We use the HAC estimator and report the p-values for a range of M_T values for series that appear to be significant at 1% or 5% for all $M_T \in \{10, 20, 30\}$ values and for two kernel functions, namely Parzen and Quadratic Spectral.

The results are reported in Table 5. First, the median analyst forecasts, the third autoregressive lag, and the BAA minus 10Y yield spread are the only significant variables at 1% apart from five news attention series: Earnings, Earnings losses, Recession, Revenue growth and Revised estimate. The recession news attention series is an important predictor for US GDP nowcasting application, see Babii, Ghysels, and Striaukas (2022), while the other four series are directly linked to earnings. Aggregate Event Sentiment (AES) and Treasury yield spread (10Y minus 3M) are additional significant variables at the 5% significance levels. AES captures firm-level sentiment while 10Y minus 3M is thought to predict future recessions. Overall, the significant variables are either linked with economic conditions or to earnings themselves.

5 Conclusions

This paper introduces a new class of high-dimensional panel data nowcasting models with dictionaries and sg-LASSO regularization which is an attractive choice for the predictive panel data regressions, where the low- and/or the high-frequency lags define a clear group structure. The estimator covers the LASSO and the group LASSO estimators as special cases.

Our empirical analysis sheds light on the advantage of the regularized panel data regressions for nowcasting corporate earnings. We focus on nowcasting the P/E ratio of 210 US firms and find that the regularized panel data regressions outperform several benchmarks, including analysts consensus forecasts. Furthermore, we find

Variable $\backslash M_T$	10	20	30	10	20	30
	Quadratic Spectral			Parzen		
			1% significance			
An.median	0.118	0.189	0.129	0.112	0.115	0.100
AR(3)	0.149	0.224	0.195	0.117	0.220	0.301
BAA minus 10Y	0.154	0.193	0.208	0.132	0.170	0.206
Earnings	0.025	0.100	0.121	0.025	0.102	0.205
Earnings losses	0.123	0.104	0.105	0.092	0.162	0.341
Recession	0.103	0.198	0.305	0.165	0.197	0.209
Revenue growth	0.145	0.209	0.351	0.257	0.305	0.495
Revised estimate	0.097	0.102	0.158	0.102	0.256	0.145
			5% significance			
AES	2.288	3.211	3.791	1.990	2.524	3.128
10Y less 3M	0.998	1.559	2.119	0.846	1.127	1.484

Table 5: Significance testing results. We report p-values (in percent) of series that are significant at 1% and 5% significance level for a range of M_T values and both kernel functions.

that the regularized machine learning regressions outperform forecast combinations and that the panel data models improve upon single time series regressions for individual firms. In addition, our analysis also shows that ML methods augmented with consensus analysts' predictions further improve the quality of P/E ratio nowcasts. Using the HAC-based inference for pooled panel data models, we show that the textual analysis data, which are related to macro and earnings data, and interest rate spreads are highly significant predictors which go beyond median consensus and autoregressive lags.

While nowcasting earnings is a leading example of applying panel data MIDAS machine learning regressions, one can think of many other applications of interest in finance. Beyond earnings, analysts are also interested in sales, dividends, etc. Our analysis can also be useful for other areas of interest, such as regional and international panel data settings.

References

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

Babii, A., R. T. Ball, E. Ghysels, and J. Striaukas (2022): "Machine Learning Panel Data Regressions with Heavy-tailed Dependent Data: Theory and Application," *Journal of Econometrics*, (forthcoming).

Babii, A., E. Ghysels, and J. Striaukas (2021): "High-dimensional Granger causality tests with an application to VIX and news," *Journal of Financial Econometrics*, (forthcoming).

——— (2022): "Machine learning time series regressions with an application to nowcasting," *Journal of Business and Economic Statistics*, 40, 1094–1106.

Ball, R. T. (2013): "Does Anticipated Information Impose a Cost on Risk-Averse Investors? A Test of the Hirshleifer Effect," *Journal of Accounting Research*, 51(1), 31–66.

Ball, R. T., and P. Easton (2013): "Dissecting earnings recognition timeliness," *Journal of Accounting Research*, 51, 1099–1132.

Ball, R. T., and L. A. Gallo (2018): "A mixed data sampling approach to accounting research," Available at SSRN 3250445.

Ball, R. T., and E. Ghysels (2018): "Automated earnings forecasts: beat analysts or combine and conquer?," *Management Science*, 64, 4936–4952.

Bańbura, M., D. Giannone, M. Modugno, and L. Reichlin (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. Elsevier.

Bybee, L., B. T. Kelly, A. Manela, and D. Xiu (2021): "Business News and Business Cycles," Available at SSRN 3446225.

Carabias, J. M. (2018): "The real-time information content of macroeconomic news: implications for firm-level earnings expectations," *Review of Accounting Studies*, 23(1), 136–166.

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

Foroni, C., M. Marcellino, and C. Schumacher (2015): "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.

Fosten, J., and R. Greenaway-McGrevy (2019): "Panel data nowcasting," *Available at SSRN 3435691*.

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

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

Hurvich, C. M., and C.-L. Tsai (1989): "Regression and time series model selection in small samples," *Biometrika*, 76(2), 297–307.

Khalaf, L., M. Kichian, C. J. Saunders, and M. Voia (2021): "Dynamic panels with MIDAS covariates: Nonlinearity, estimation and fit," *Journal of Econometrics*, 220(2), 589–605.

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.

Zou, H., T. Hastie, and R. Tibshirani (2007): "On the "degrees of freedom" of the LASSO," *Annals of Statistics*, 35(5), 2173–2192.