Machine Learning Panel Data Regressions with an Application to Nowcasting Price-Earnings Ratios

Andrii Babii[*] Ryan T. Ball[†] Eric Ghysels[‡] Jonas Striaukas[§]

March 22, 2021

Abstract

This paper introduces structured machine learning regressions for nowcasting and forecasting with panel data consisting of series sampled at different frequencies. Motivated by the empirical 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. This type of regularization can take advantage of the mixed frequency time series panel data structures and we find that it empirically outperforms the unstructured machine learning methods. We obtain oracle inequalities for the pooled and fixed effects sparse-group LASSO panel data estimators recognizing that financial and economic data can have fat tails. To that end, we leverage on a new Fuk-Nagaev concentration inequality for panel data consisting of heavy-tailed τ-mixing processes. Our empirical results show the superior performance of introduced machine learning panel data regressions over analysts' predictions, forecast combinations, firm-specific time series regression models, and unstructured elastic net.

Keywords: corporate earnings, nowcasting, high-dimensional panels, mixed frequency data, textual news data, sparse-group LASSO, fat tails, tau-mixing, Fuk-Nagaev inequalities.

[*] 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.

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

1 Introduction

The fundamental value of equity shares is determined by the discounted value of future payoffs. Every quarter investors get a glimpse of a firms' potential payoffs with the release of corporate earnings reports. In a data-rich environment, stock analysts have many indicators regarding future cash flows 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 regularized mixed frequency regression 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 time series and cross-section dimensions can be properly modeled. In such models, time-invariant firm-specific effects are typically modeled in a flexible way which allows capturing heterogeneity in the data. At the same time, machine learning methods are becoming increasingly popular in economics and finance as a flexible way to model relationships between the response and covariates.

In the present paper, we analyze panel data regressions in a high-dimensional setting where the number of time-varying covariates can be very large and potentially exceed the 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 how to implement nowcasting in such 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 not in the context of nowcasting or forecasting. On the other hand, Fosten and Greenaway-McGrevy (2019) consider nowcasting with a mixed-frequency VAR panel data model, but not in the context of high-dimensional data-rich environment that we are interested in the present paper.

It this paper, we aim to fill this gap in the literature and focus on nowcasting and forecasting with high-dimensional mixed frequency panel data. In contrast to the previous literature, we allow for the number of predictors to be large compared to the effective sample size. To that end, we leverage on the structured sparsity approach with the sparse-group LASSO (sg-LASSO) regularization and aggregation of mixed frequency lags with dictionaries. The advantages of this approach for nowcasting individual time series data sampled at mixed frequencies have been reported recently in Babii, Ghysels, and Striaukas (2021b) who focus on nowcasting the US GDP growth in a data-rich environment. However, Babii, Ghysels, and Striaukas (2021b) do not address the problem of nowcasting with panel data. In this paper, we first show how to leverage on the sparse group regularization in the mixed frequency panel data setting. Second, we study the benefits of cross-sectional dimension for prediction with panel data potentially consisting of fat-tailed series. Lastly, we apply the developed methodology to the problem of nowcasting price earnings ratios in a data-rich environment.

Our paper also relates to the literature on high-dimensional panel data models and/or the (group) LASSO regularization; see Harding and Lamarche (2019), Chiang, Rodrigue, and Sasaki (2019), Chernozhukov, Hausman, and Newey (2019), Belloni, Chen, Padilla, et al. (2019), Belloni, Chernozhukov, Hansen, and Kozbur (2016), Lu and Su (2016), Kock (2016), Su, Shi, and Phillips (2016), Lu and Su (2016), Farrell (2015), Kock (2013), Lamarche (2010), Koenker (2004) among many others. However, to the best of our knowledge, the existing literature relates mostly to the microeconometric problems and does not address properly the nowcasting/forecasting problems and does not address comprehensively: 1) the advantages of long panels; 2) the performance of regularized panel data estimators with potentially heavy-tailed covariates and regression errors; 3) the sparse-group LASSO regularization used for individual time series regressions in Babii, Ghysels, and Striaukas (2021b), which was initially motivated by a genetic application in Simon, Friedman, Hastie, and Tibshirani (2013).

We recognize that the economic and financial time series data are often persistent with fat tails. To that end, we leverage on new Fuk-Nagaev concentration inequality for long panels; cf. Babii, Ghysels, and Striaukas (2021a) for individual time series regressions. Using this inequality, we obtain oracle inequalities for the sparse-group LASSO that shed new light on how the predictive performance of pooled and fixed effect estimators scales with N (cross-section) and T (time series), which is especially relevant for modern panel data applications, where both N and T can be large; see Fernández-Val and Weidner (2018). Importantly, our theory covers the LASSO and the group-LASSO estimators as special cases of sparse group LASSO and we show how tails and persistence of the data affect the predictive performance. Lastly, we also illustrate the finite sample advantages of the sparse-group LASSO regularization with mixed-frequency panel data.

An empirical application to nowcasting firm-specific price/earnings ratios (henceforth P/E ratio) is provided. We focus on the current quarter nowcasts, hence 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 particular company and therefore used by analysts and investment professionals to base their decisions on which stocks to pick for their investment portfolios. A typical value investor relies on consensus forecasts of earnings made by a pool of analysts. Hence, we naturally benchmark our proposed machine learning methods against such predictions. Besides, we compare our methods with a forecast combination approach used by Ball and Ghysels (2018) and a simple random walk (RW).

Our high-frequency regressors include traditional macro and financial series as well as non-standard series generated by the textual analysis. We consider structured pooled and fixed effects sg-LASSO panel data regressions with mixed frequency data (sg-LASSO-MIDAS). The fixed effects estimator yields sparser models compared to pooled regressions with the Revenue growth and the first lag of the dependent variable are selected throughout the out-of-sample period. BAA less 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 over analysts' predictions, forecast combination method, and firm-specific time series regression models. Besides, the sg-LASSO-MIDAS regressions perform better than unstructured panel data regressions with the elastic net regularization.

Regarding the textual news data, it is worth emphasizing that the time series of news data is sparse since for many days are without firms-specific news and we impute zero values. The nice property of our mixed frequency data treatment with dictionaries, imputing zeros also implies that non-zero entries get weights with a decaying pattern for distant past values in comparison to the most recent daily news data. As a result, our ML approach is particularly useful to model news data which is sparse in nature.

The paper is organized as follows. Section 2 introduces the models and estimators. Oracle inequalities for sparse group LASSO panel data regressions appear in Section 3. 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 description appears in the Appendix.

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 a given group structure $\mathcal{G}$, the $\ell_{2,1}$ group norm of $b \in \mathbf{R}^p$ is defined as $\|b\|_{2,1} = \sum_{G\in\mathcal{G}} |b_G|_2$. For $\mathbf{u}, \mathbf{v} \in \mathbf{R}^J$, the empirical inner product is defined as $\langle \mathbf{u}, \mathbf{v}\rangle_J = J^{-1}\sum_{j=1}^J u_j v_j$ with the induced empirical norm $\|.\|_J^2 = \langle ., .\rangle_J = |.|_2^2/J$. For a symmetric $p \times p$ matrix A, let $\operatorname{vech}(A) \in \mathbf{R}^{p(p+1)/2}$ 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) panels

Motivated by our empirical application, we allow the high-dimensional set of predictors to be sampled at a higher frequency than the target variable. 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

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

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 $\beta_k = (\beta_{1,k}, \dots, \beta_{m,k})^\top \in \mathbf{R}^m$. 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. In addition, we can also absorb the (low-frequency) lags of $y_{i,t}$ in covariates. When $m = 1$, we retain the standard panel data regression model

$$y_{i,t+h} = \alpha_i + \sum_{k=1}^{K} \beta_k x_{i,t,k} + u_{i,t},$$

while $m > 1$ signifies that the high-frequency lags of $x_{i,t,k}$ are also included. The large number of predictors K with 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 statistically costly and may reduce the predictive performance in small samples.

To reduce the proliferation of lag parameters, we follow the MIDAS literature; see Ghysels, Santa-Clara, and Valkanov (2006), Ghysels, Sinko, and Valkanov (2006), and Babii, Ghysels, and Striaukas (2021a,b). Instead of estimating m individual slopes of high-frequency covariate $k \in [K]$ in equation (1), we estimate a weight function ω parametrized by $\beta_k \in \mathbf{R}^L$ with $L < m$

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

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 approximating functions, called the *dictionary*. An example of the dictionary is the orthogonal Legendre polynomials[1] on $[0,1]$ that can be computed via the Rodrigues' formula $w_l(s) = \frac{1}{l!}\frac{\mathrm{d}^l}{\mathrm{d}s^l}(s^2-s)^l$. For instance, the first five elements are

$$\begin{aligned}
w_0(s) &= 1\\
w_1(s) &= 2s-1\\
w_2(s) &= 6s^2-6s+1\\
w_3(s) &= 20s^3-30s^2+12s-1\\
w_4(s) &= 70s^4-140s^3+90s^2-20s+1.
\end{aligned}$$

More generally, we can use Gegenbauer polynomials, trigonometric polynomials, or wavelets. The orthogonal polynomials usually have better numerical properties than their popular non-orthogonal counterpart, cf. Almon (1965). The attractive feature of linear in parameters dictionaries is that we can map the MIDAS regression to the linear regression framework that can be solved via the convex optimization. To that end, put $\mathbf{x}_i = (X_{i,1}W,\dots,X_{i,K}W)$, where for each $k \in [K]$, $X_{i,k} = (x_{i,t-(j-1)/m,k})_{t\in[T],j\in[m]}$ is $T\times m$ 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 $(w_l)_{l\geq 0}$. Put also

$$\mathbf{y}_i = (y_{i,1+h},\dots,y_{i,T+h})^\top \qquad \text{and} \qquad \mathbf{u}_i = (u_{i,1},\dots,u_{i,T})^\top.$$

[1] The Legendre polynomials have the universal approximation property and can approximate any continuous weight function uniformly on $[0,1]$.

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 slopes. Lastly, put $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_N^\top)^\top$, $\mathbf{X} = (\mathbf{x}_1^\top, \dots, \mathbf{x}_N^\top)^\top$, and $\mathbf{u} = (\mathbf{u}_1^\top, \dots, \mathbf{u}_N^\top)^\top$. Then the regression equation after stacking all cross-sectional observations is

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

where $B = I_N \otimes \iota$ and $\alpha = (\alpha_1, \dots, \alpha_N)$.

The MIDAS approach allows us reducing effectively the dimensionality pertaining to the high-frequency lags. While assuming that the individual lag coefficients in equation (1) are approximately sparse is *highly* restrictive, the approximate sparsity of slopes of the dictionary elements $(w_l)_{l\geq 0}$ is plausible. For instance, if $w_0(s) = 1$ with $\beta_{0,k} \neq 0$ and $\beta_{l,k} = 0, \forall l \geq 1$, we recover the averaging of high-frequency lags of covariate k as a special case. More generally, the weight ω may be a decreasing function over lags and we may want to learn its shape from the data maximizing the predictive performance.

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 leverage on the sparse-group LASSO regularization that was shown to be attractive for individual time series regressions in Babii, Ghysels, and Striaukas (2021b). The fixed effects panel data estimator with sparse-group regularization solves

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

where $\|.\|_{NT}^2 = |.|^2/(NT)$ is the empirical norm and Ω is a regularizing functional,

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

which is a linear combination of LASSO and group LASSO penalties. 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 $\lambda \geq 0$. Recall also that for a group structure $\mathcal{G}$ described as a partition of $[p] = \{1, 2, \dots, p\}$, the group LASSO norm is computed as $\|b\|_{2,1} = \sum_{G\in\mathcal{G}} |b_G|_2$. The group structure is assumed to be known to the econometrician, which in our setting corresponds to time series lags of covariates. More generally, we may also combine covariates of a similar nature in groups. Throughout the paper we assume that groups have fixed size, which is well-justified in our empirical applications.[2] Therefore, the selection of covariates is performed by the group LASSO penalty, which encourages sparsity between groups. In addition, the ℓ_1 LASSO norm promotes sparsity within groups and allows us to learn the shape of the MIDAS weights from the data.

It is worth mentioning that the linear in parameters approximation to the MIDAS weight function leads to the convex optimization parameter problem in equation (2) that can be solved efficiently, e.g., via the proximal gradient descent algorithm, or its block-coordinate descent versions. In contrast, a popular beta weighting scheme, leads to a nonlinear non-convex optimization problem that becomes challenging to solve in high-dimensions; cf. Marsilli (2014) and Khalaf, Kichian, Saunders, and Voia (2021).

3 Oracle inequalities for panel data

In this section, we 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 special cases. It is worth stressing that the analysis of this section is not tied to the mixed-frequency data setting and applies to the generic high-dimensional panel data regularized with the sg-LASSO penalty function. Importantly, we focus on panels consisting of potentially persistent τ-mixing time series with polynomial tails. Consider a generic linear projection panel data model with a countable number of predictors

$$y_{i,t+h} = \alpha_i + \sum_{j=1}^{\infty} \beta_j x_{i,t,j} + u_{i,t}, \qquad \mathbb{E}[u_{i,t}x_{i,t,j}] = 0, \quad \forall j \geq 1.$$

This model subsumes the mixed-frequency data regressions as a special case, in which case covariates are obtained, e.g., from the aggregation with Legendre polynomials. The covariates may also include the time-varying covariates common for all entities (macroeconomic factors), lags of $y_{i,t}$, the intercept, as well as additional lags of a baseline covariate.

[2] See Babii (2020) for a continuous-time mixed-frequency regression where the group size is allowed to increase with the sample size under the in-fill asymptotics.

3.1 τ-mixing

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

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

where $\mathrm{Lip}_1 = \{f : \mathbf{R}^l \to \mathbf{R} : \ |f(x) - f(y)| \leq |x - y|_1\}$ is a set of 1-Lipschitz functions from $\mathbf{R}^l$ to $\mathbf{R}$.[3] For a stochastic process $(\xi_t)_{t \in \mathbf{Z}}$ with a natural filtration generated by its past $\mathcal{M}_t = \sigma(\xi_t, \xi_{t-1}, \dots)$, the τ-mixing coefficients are defined as

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

where the supremum is taken over all $t, t_1, \dots, t_j \in \mathbf{Z}$. If $\tau_k \downarrow 0$, as $k \uparrow \infty$ then the process is called τ-mixing. The class of τ-mixing processes can be placed somewhere between the α-mixing processes and mixingales – the τ-mixing condition is less restrictive than the α-mixing condition,[4] yet at the same time, there exists a convenient for us coupling result for τ-mixing processes, which is not the case for the mixingales or near-epoch dependent processes; see Dedecker and Doukhan (2003) and Dedecker and Prieur (2004, 2005) for more details. This allows us to obtain concentration inequalities and performance guarantees for the sparse-group LASSO estimator; see Appendix B for more details.

3.2 Pooled regression

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

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

Put $z_{i,t} = (1, x_{i,t}^\top)^\top$, where $x_{i,t} \in \mathbf{R}^p$ is a vector of predictors. The following assumption imposes mild restrictions on the data.

[3] See Dedecker and Prieur (2004) and Dedecker and Prieur (2005) for equivalent definitions.

[4] The class of α-mixing processes is too restrictive for the predictive linear projection model with covariates and autoregressive lags.

Assumption 3.1 (Data)**.** *(i) for each $t \in \mathbf{Z}$, $\{(u_{i,t}, z_{i,t}) : i \geq 1\}$ are i.i.d. and for each $i \geq 1$, $\{(u_{i,t}, z_{i,t}) : \ t \in \mathbf{Z}\}$ is a stationary process; (ii) $\max_{j\in[p+1]} \|u_{i,t}z_{i,t,j}\|_q = O(1)$ for some $q > 2$; (iii) each process in $(u_{i,t}z_{i,t})_{t\in\mathbf{Z}}$ has τ-mixing coefficients $\tau_k \leq ck^{-a}, \forall k \geq 1$ with universal constants $c > 0$ and $a > (q-1)/(q-2)$; (iv) $\max_{j,k\in[p+1]} \|z_{i,t,j}z_{i,t,k}\|_{\tilde{q}} = O(1)$ for some $\tilde{q} > 2$; (v) each process in $(z_{i,t}z_{i,t}^\top)_{t\in\mathbf{Z}}$ has τ-mixing coefficients $\tilde{\tau}_k \leq \tilde{c}k^{-\tilde{a}}, \forall k \geq 1$ with $\tilde{c} > 0$ and $\tilde{a} > (\tilde{q}-1)/(\tilde{q}-2)$.*

It is worth mentioning that the stationarity hypothesis can be relaxed at costs of introducing heavier notation. We require that only $2+\epsilon$ moments exist with $\epsilon > 0$, which is a realistic assumption in our empirical application and more generally for datasets encountered in time series and financial econometrics applications. Note also that the time series dependence is assumed to fade away relatively slow – at a polynomial rate as measured by the τ-mixing coefficients.

Next, we assume that the matrix $\Sigma = \mathbb{E}[z_{i,t}z_{i,t}^\top]$ is non-singular.

Assumption 3.2 (Covariance matrix)**.** *The smallest eigenvalue of Σ is uniformly bounded away from zero by some universal constant $\gamma > 0$.*

Assumption 3.2 is satisfied for the spiked identity and Topelitz covariance structures. It can be interpreted as a completeness condition, see Babii and Florens (2020), and can also be relaxed to the restricted eigenvalue condition imposed on the population covariance matrix Σ; see Babii, Ghysels, and Striaukas (2021b). In addition, we can allow for $\gamma \downarrow 0$ as $p \uparrow \infty$, in which case γ^{-1} would slow down the convergence rates in oracle inequalities and could be interpreted as a measure of ill-posedness; see Carrasco, Florens, and Renault (2007).

Lastly, we assume that the regularization parameter λ scales appropriately with the number of covariates p, the length of the panel T, the size of the cross-section N, and a certain exponent κ that depends on tail parameter q and persistence of the data a. The precise order of the regularization parameter is described by the Fuk-Nagaev inequality for long panels appearing in Appendix, see Theorem A.1.

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

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

where $\kappa = ((a+1)q-1)/(a+q-1)$ and a, q are as in Assumptions 3.1.

Our first result is the oracle inequality for the pooled sg-LASSO estimator described in equation (3). The result allows for misspecified regressions with a nontrivial approximation error in the sense that we consider more generally

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

where $\mathbf{m} \in \mathbf{R}^{NT}$ is approximated with $\mathbf{Z}\rho$, $\mathbf{Z} = (\iota, \mathbf{X})$, $\iota \in \mathbf{R}^{NT}$ is all-ones vector, and $\rho = (\alpha, \beta^\top)^\top$. The approximation error $\mathbf{m} - \mathbf{Z}\rho$ might come from the fact the MIDAS weight function may not have the exact expansion in terms of the specified dictionary or from the fact that some of the relevant predictors are not included in the regression equation. To state the result, let $S_0 = \{j \in [p] : \beta_j \neq 0\}$ be the support of β and let $\mathcal{G}_0 = \{G \in \mathcal{G} : \beta_G \neq 0\}$ be the group support of β. Consider the *effective sparsity* of the sparse-group structure, defined as $s^{1/2} = \gamma\sqrt{|S_0|} + (1-\gamma)\sqrt{|\mathcal{G}_0|}$. Note that s is proportional to the sparsity $|S_0|$, when $\gamma = 1$ and to the group sparsity $|\mathcal{G}_0|$ when $\gamma = 0$. Put $r_{N,T}^{\text{pooled}} = s^{\tilde{\kappa}}p^2/(NT)^{\tilde{\kappa}-1} + p^2\exp(-cNT/s^2)$.

Theorem 3.1. *Suppose that Assumptions 3.1, 3.2, and 3.3 are satisfied. Then with probability at least* $1 - \delta - O(r_{N,T}^{\text{pooled}})$

$$\|\mathbf{Z}(\hat{\rho} - \rho)\|_{NT}^2 \lesssim s\lambda^2 + \|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2$$

and

$$|\hat{\rho} - \rho|_1 \lesssim s\lambda + \lambda^{-1}\|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2 + s^{1/2}\|\mathbf{m} - \mathbf{Z}\rho\|_{NT},$$

for some $c > 0$ *and* $\tilde{\kappa} = ((\tilde{a}+1)\tilde{q} - 1)/(\tilde{a} + \tilde{q} - 1)$.

The proof of this result can be found in the Appendix. Theorem 3.1 describes the non-asymptotic oracle inequalities for the prediction and the estimation accuracy in the environment where the number of regressors p is allowed to scale with the effective sample size NT. Importantly, the result is stated under the weak tail and persistence conditions in Assumption 3.1. Parameters κ and $\tilde{\kappa}$ are the dependence-tails exponents for stochastic processes driving the regression score and the covariance matrix respectively. Theorem 3.1 shows that the prediction and the estimation accuracy of pooled panel data regressions improves when the sparse-group structure is taken into account. Indeed, for the LASSO regression, the effective sparsity reduces to $s^{1/2} = \sqrt{|S_0|}$, which is larger than $\gamma\sqrt{|S_0|} + (1-\gamma)\sqrt{|\mathcal{G}_0|}$ in the case of sparse-group LASSO.

Next, we consider the convergence rates of the prediction and estimation errors. The following assumption considers a simplified setting, where the approximation error vanishes sufficiently fast, and the total number of regressors scales appropriately with the effective sample size NT.

Assumption 3.4. *(i)* $\|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2 = O_P(s\lambda^2)$*; and (ii)* $s^{\tilde{\kappa}}p^2(NT)^{1-\tilde{\kappa}} \to 0$ *and* $p^2\exp(-cNT/s^2) \to 0$.

Note that Assumption 3.4 allows for 1) $N \to \infty$ while T is fixed; 2) $T \to \infty$ while N is fixed; and 3) both $N \to \infty$ and $T \to \infty$ without restricting the relative growth of the two. The following result describes the prediction and the estimation convergence rates in the asymptotic environment outlined in Assumption 3.4 and is an immediate consequence of Theorem 3.1.

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

$$\|\mathbf{Z}(\hat{\rho}-\rho)\|_{NT}^2 = O_P\left(\frac{sp^{2/\kappa}}{(NT)^{2-2/\kappa}} \vee \frac{s\log p}{NT}\right)$$

and

$$|\hat{\rho}-\rho|_1 = O_P\left(\frac{sp^{1/\kappa}}{(NT)^{1-1/\kappa}} \vee s\sqrt{\frac{\log p}{NT}}\right).$$

Corollary 3.1 describes the prediction and the estimation accuracy of pooled sparse-group panel data regressions. It suggests that the predictive performance of the sparse-group LASSO (and consequently LASSO and group LASSO) regressions may deteriorate when regression errors and/or predictors are heavy-tailed or when the data are extremely persistent. However, for geometrically ergodic Markov processes, e.g., stationary AR(1) process, the τ-mixing coefficients decline geometrically fast, so that $\kappa \approx q$ and $\tilde{\kappa} \approx \tilde{q}$. In this case, the prediction accuracy scales approximately at the rate $O_P\left(\frac{p^{2/q}}{(NT)^{2-2/q}} \vee \frac{\log p}{NT}\right)$ and the predictive performance may be affected only by the tails constant q.

If additionally, the data are sub-Gaussian, then moments of all order $q \geq 2$ exist and for any particular effective sample size NT, the first term can be made arbitrarily small relatively to the second term. In this case we recover the $O_P\left(\frac{\log p}{NT}\right)$ rate typically obtained for sub-Gaussian data. On the other hand, if the polynomial tail dominates, then we need $p = o((NT)^{q-1})$ for the prediction and the estimation consistency provided that $\tilde{q} \geq 2q-1$ and the sparsity constant s is fixed. In this case, we have a *significantly weaker* requirement than $p = o(T^{q-1})$ needed for time series regressions in Babii, Ghysels, and Striaukas (2021b). Moreover, since $q > 2$, $p = o((NT)^{q-1})$ can be significantly weaker than $p = o(NT)$ condition typically needed for QMLE/GMM estimators without regularization.

Theorem 3.1 and Corollary 3.1 imply immediately two practical consequences: 1) one may want to exclude the heavy-tailed series from the high-dimensional predictive regressions based on the preliminary estimates of the tail index, e.g., using the Hill estimator; 2) if the individual heterogeneity can be ignored, then pooling panel data can improve significantly the predictive performance. In the latter case, one can also preliminary cluster similar series in groups, e.g., based on the unsupervised clustering algorithms, which may strike a good balance between the pooling benefits and heterogeneity.

3.3 Fixed effects

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 substantial heterogeneity over $i \in [N]$, then taking this into account might reduce the projection error and improve the predictive accuracy. At a very extreme side, the cross-sectional structure can be completely ignored and individual time-series regressions can be used for prediction. The fixed effects panel data regressions strike a good balance between the two extremes controlling for heterogeneity with entity-specific intercepts.

The fixed effects sg-LASSO estimator $\hat{\rho} = (\hat{\alpha}^\top, \hat{\beta}^\top)^\top$ solves

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

where $B = I_N \otimes \iota$, I_N is $N \times N$ identity matrix, $\iota \in \mathbf{R}^T$ is an all-ones vector, and 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, that are typically not sparse. By Fermat's rule, the first-order conditions are

$$\begin{aligned} \hat{\alpha} &= (B^\top B)^{-1}B^\top(\mathbf{y} - \mathbf{X}\hat{\beta}) \\ 0 &= \mathbf{X}^\top M_B(\mathbf{X}\hat{\beta} - \mathbf{y})/NT + \lambda z^* \end{aligned} \tag{4}$$

for some $z^* \in \partial\Omega(\hat{\beta})$, where $b \mapsto \partial\Omega(b)$ is the subdifferential of Ω and $M_B = I - B(B^\top B)^{-1}B^\top$ is the orthogonal projection matrix. It is easy to see from the first-order conditions that the estimator of $\hat{\beta}$ is equivalent to: 1) penalized GLS estimator for the first-differenced regression; 2) penalized OLS estimator for the regression written in the deviation from time means; and 3) penalized OLS estimator where the fixed effects are partialled-out. Thus, the equivalence between the three approaches is not affected by the penalization; cf. Arellano (2003) for low-dimensional panels.

With some abuse of notation, redefine

$$\hat{\Sigma} = \begin{pmatrix} \frac{1}{T}B^\top B & \frac{1}{\sqrt{NT}}B^\top\mathbf{X} \\ \frac{1}{\sqrt{NT}}\mathbf{X}^\top B & \frac{1}{NT}\mathbf{X}^\top\mathbf{X} \end{pmatrix} \qquad \text{and} \qquad \Sigma = \begin{pmatrix} I_N & \frac{1}{\sqrt{NT}}\mathbb{E}\left[B^\top\mathbf{X}\right] \\ \frac{1}{\sqrt{NT}}\mathbb{E}\left[\mathbf{X}^\top B\right] & \mathbb{E}[x_{i,t}x_{i,t}^\top] \end{pmatrix}. \tag{5}$$

We will assume that the smallest eigenvalue of Σ is uniformly bounded away from zero by some constant. Note that if $x_{i,t} \sim N(0, I_p)$, then $\Sigma = I_{N+p}$ and this assumption is trivially satisfied.

The order of the regularization parameter is governed by the Fuk-Nagaev inequality for long panels; see Appendix, Theorem A.1.

Assumption 3.5 (Regularization). *For some $\delta \in (0,1)$*

$$\lambda \sim \left(\frac{p \vee N^{\kappa/2}}{\delta(NT)^{\kappa-1}}\right)^{1/\kappa} \vee \sqrt{\frac{\log(p \vee N/\delta)}{NT}},$$

where $\kappa = ((a+1)q-1)/(a+q-1)$, and a, q are as in Assumptions 3.1.

Similarly to the pooled regressions, we state the oracle inequality allowing for the approximation error. For fixed effects regressions, with some abuse of notation we redefine $\mathbf{Z} = (B, \mathbf{X})$ and $\rho = (\alpha^\top, \beta^\top)^\top$. Put also $r_{N,T}^{\mathrm{fe}} = p(s \vee N)^{\tilde{\kappa}} T^{1-\tilde{\kappa}}(N^{1-\tilde{\kappa}/2} + pN^{1-\tilde{\kappa}}) + p(p \vee N)e^{-cNT/(s\vee N)^2}$ with $\tilde{\kappa} = ((\tilde{a}+1)\tilde{q}-1)/(\tilde{a}+\tilde{q}-1)$ and some $c > 0$.

Theorem 3.2. *Suppose that Assumptions 3.1, 3.2, and 3.5 are satisfied (with redefined Σ). Then with probability at least $1 - \delta - O(r_{N,T}^{\mathrm{fe}})$*

$$\|\mathbf{Z}(\hat{\rho} - \rho)\|_{NT}^2 \lesssim (s \vee N)\lambda^2 + \|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2.$$

Theorem 3.2 states a non-asymptotic oracle inequality for the prediction error in the fixed effects panel data regressions estimated with the sg-LASSO. To see clearly, how the prediction accuracy scales with the sample size, we make the following assumption.

Assumption 3.6. *Suppose that (i) $\|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2 = O_P((s \vee N)\lambda^2)$; (ii) $(p + N^{\tilde{\kappa}/2})p(s \vee N)^{\tilde{\kappa}} N^{1-\tilde{\kappa}} T^{1-\tilde{\kappa}} \to 0$ and $p(p \vee N)e^{-cNT/(s\vee N)^2} \to 0$.*

The following corollary is an immediate consequence of Theorem 3.2.

Corollary 3.2. *Suppose that Assumptions 3.1, 3.2, 3.5, and 3.6 are satisfied. Then*

$$\|\mathbf{Z}(\hat{\rho} - \rho)\|_{NT}^2 = O_P\left(\frac{(s \vee N)(p^{2/\kappa} \vee N)}{N^{1-2/\kappa}T^{2-2/\kappa}} \vee \frac{(s \vee N)\log(p \vee N)}{NT}\right).$$

Corollary 3.2 allows for $s, p, N, T \to \infty$ at appropriate rates. However, we pay additional price for estimating N fixed effects which plays a similar role to the effective dimension of covariates. An immediate practical implication is that in order to achieve accurate predictions with high-dimensional fixed effect regressions, the panel has to be sufficiently long in order to offset the estimation error of the individual fixed effects. Likewise, the tails and the persistence of the data may also reduce the prediction accuracy in small samples through κ, which is approximately equal to q for geometrically decaying τ-mixing coefficients.

4 Nowcasting price-earnings ratios

In this section, we 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-standard series generated by the textual analysis. We apply pooled and fixed effects sg-LASSO-MIDAS panel data models and compare them with several benchmarks such as random walk (RW), analysts consensus forecasts, and unstructured elastic net. We also compute predictions using individual-firm high-dimensional time series regressions and provide results for several choices of the tuning parameter. Lastly, we provide results for 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 regarding the analysis in the current paper as it also deals with nowcasting price earnings ratios. The forecast combination methods consist of estimating ADL-MIDAS regressions with 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 discounted mean squared error type, produces a single nowcast. One could call this a pre-machine learning large dimensional approach. It will, therefore, be interesting to assess how this approach compares to the regularized MIDAS panel regression machine learning approach introduced in the current paper.

We start with a short review of the data, with more detailed descriptions and tables available appearing in Appendix Section D, followed by a summary of the methods used and the empirical results obtained.

4.1 Data description

The full sample consists of observations between 1^{st} of January, 2000 and 30^{th} of June, 2017. Due to the lagged dependent variables in the models, our effective sample starts 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 the data from CRSP and I/B/E/S to compute 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 obtained from 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 is retrieved from Bybee, Kelly, Manela, and

Xiu (2019).[5] Appendix Section D provides a detailed description of the data sources. In particular, firm-level variables, including P/E ratios, are described in Appendix Table A.3, and the other predictor variables in Appendix Table A.4. The list of all firms we consider in our analysis appears in Appendix Table A.5.

P/E ratio and analysts' forecasts sample construction. Our target variable is the P/E ratio for each individual 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 separately compute the dependent variable, analysts' consensus forecasts, and the target variable using stock prices that were 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. Our dependent variable used in regression models is computed by taking the end of quarter prices and dividing them by the respective earnings value. 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 which we use to measure the prediction performance, 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 example for a particular firm the data is published on the $25th$ of April. In this case, we record the stock price for this particular firm on $25th$ of April, and divide it by the realized earnings value.

4.2 Tuning parameters

We consider several approaches to select the tuning parameter λ. 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 individual firms instead of the pooled sample. We use 5-fold cross-validation as the sample size of the dataset we consider in our empirical application is relatively small. We also consider the following three information criteria: BIC, AIC, and corrected AIC (AICc). Assuming that $y_{i,t}|x_{i,t}$ are i.i.d. draws from $N(\alpha_i + x_{i,t}^\top \beta, \sigma^2)$,

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

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

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

Lastly, the corrected Akaike information criteria, see Hurvich and Tsai (1989), is

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

The AICc might be a better choice when p is large compared to the sample size. We report results for each of these four choices of the tuning parameters.

4.3 Models and main results

To compute forecasts, we estimate several regression models. First, we estimate firm-specific sg-LASSO-MIDAS regressions for firm $i=1,\dots,N$, which in Table 1 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 MIDAS weights for each of the high-frequency covariate. We then estimate the following pooled and fixed effects sg-LASSO-MIDAS panel data models

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

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 $t+1$ quarter, and the stock price is also taken at the end of $t+1$.

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 initial in-sample observation 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 from a specific method. 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 are reported in Table 1, while additional results for unstructured LASSO estimators and the forecast combination approach appear in Appendix Tables A.1-A.2. First, we document that analysts-based predictions have much larger mean squared forecast errors (MSEs) compared to model-based 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 1 and A.1. A better performance is achieved for almost all machine learning methods - single firm or panel data regressions - and all tuning parameter choices. Unstructured panel data methods and forecast combination approach also yield more accurate forecasts, see Appendix Table A.1-A.2. The latter confirms the findings of Ball and Ghysels (2018).

RW	MSE An.-mean	MSE An.-median		sg-LASSO					
2.331	2.339	2.088	$\gamma =$	0	0.2	0.4	0.6	0.8	1
			Panel A. 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 B. 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 C. 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 D. 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

Table 1: Prediction results – The table reports average over firms MSEs of out-of-sample predictions. The nowcasting horizon is the current month, i.e. we predict the P/E ratio using information up to the end of current fiscal quarter. Each Panel A-D block represents different ways of calculating the tuning parameter λ. Bold entries are the best results in a block.

Turning to the comparison of model-based predictions, we see from the results in Table 1 that sg-LASSO-MIDAS panel data models improve the quality of predictions over 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. We also report similar findings for unstructured estimators. Within the panel data framework, we observe that fixed effects improve over pooled regressions in most cases except when cross-validation is used; compare Table 1-A.2 Panel A with Table 1-A.2 Panel B-D. 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 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 Appendix Figure A.1, we plot the sparsity pattern 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 A.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 less AAA bond yield spread, firm-level volatility, and Aggregate Event Sentiment index are also selected very 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 a more blurred pattern.[6] In this case, Revenue growth and 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 coefficient vector.

It is also worth noting that the textual news data analytics also appear in the models according the results appear in Figure A.1. These are the ESS, AES, AEV, CSS and NEP regressors described in detail in Appendix Section D. Among them, as already noted, AES - 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 since for many days are without firms-specific news. For such days, we impute zero values. The nice property of our mixed frequency data treatment with dictionaries, imputing zeros also implies that non-zero entries get weights with a decaying pattern for distant past values in comparison to the most recent daily news data.

4.4 Significance test

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, see Table 1. We first report the forecast accuracy test results in Table 2.

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 outlier observations. We report three

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

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 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.[7]

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 yields significantly higher quality predictions for sg-LASSO and significantly worse for the elastic net.

Large differences in prediction accuracy for different pools of P/E ratios may relate to heavy-tailedness of regression errors for both machine learning methods. In Table 3, we report the maximum likelihood estimates of the shape parameter of generalized error distribution, see Nelson (1991), for the in-sample residuals pooled as in Table 2. The parameter less than two indicates the presence of heavy-tails in the residuals, while larger or equal to two suggests Gaussian-like or lighter tails. In line with our theory, 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*), tails of the residuals appear to be the heaviest.

5 Conclusions

This paper introduces a new class of high-dimensional panel data regression models with dictionaries and sparse-group LASSO regularization. This type of regularization is an especially attractive choice for the predictive panel data regressions, where the low- and/or the high-frequency lags define a clear group structure, and dictionaries are used to aggregate time series lags. The estimator nests the LASSO and the group LASSO estimators as special cases, as discussed in our theoretical analysis. Our theoretical treatment allows for the heavy-tailed data frequently encountered

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

	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 2: 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 1 (sg-LASSO-MIDAS) and Table A.1 (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*.

in time series and financial econometrics. To that end, we obtain a new panel data concentration inequality of the Fuk-Nagaev type for τ-mixing processes.

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 the analysts' predictions. Furthermore, we find that the regularized machine learning regressions outperform the forecast combinations and that the panel data approach improves upon the predictive time series regressions for individual firms.

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.

	Full sample	sg-LASSO & elnet	sg-LASSO	none
		sg-LASSO		
Pooled	1.167	1.506	1.398	1.010
Fixed Effects	1.226	1.607	1.237	1.077
		Elastic net		
Pooled	1.243	1.670	1.413	1.048
Fixed Effects	1.243	1.586	1.133	1.117
	GED parameter for the response variable			
	1.142	1.383	1.378	1.062
		Number of firms		
Pooled	210	63	12	135
Fixed Effects	210	66	8	134

Table 3: Heaviness of tails – The table reports the maximum likelihood estimate of the shape parameter of it Generalized Error Distribution of in-sample residuals. The results are reported for the models as in Table 2.

References

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

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.

Arellano, M. (2003): *Panel data econometrics*. Oxford University Press.

Babii, A. (2020): "High-dimensional mixed-frequency IV regression," *arXiv preprint arXiv:2003.13478*.

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

Babii, A., E. Ghysels, and J. Striaukas (2021a): "High-dimensional Granger causality tests with an application to VIX and news," *arXiv preprint arXiv:1912.06307*.

——— (2021b): "Machine learning time series regressions with an application to nowcasting," *Journal of Business & Economic Statistics (forthcoming)*.

Ball, R. T., and E. Ghysels (2018): "Automated earnings forecasts: beat analysts or combine and conquer?," *Management Science*, 64(10), 4936–4952.

Belloni, A., M. Chen, O. H. M. Padilla, et al. (2019): "High dimensional latent panel quantile regression with an application to asset pricing," *arXiv preprint arXiv:1912.02151*.

Belloni, A., V. Chernozhukov, C. Hansen, and D. Kozbur (2016): "Inference in high-dimensional panel models with an application to gun control," *Journal of Business and Economic Statistics*, 34(4), 590–605.

Bybee, L., B. T. Kelly, A. Manela, and D. Xiu (2019): "The structure of economic news," Available at SSRN 3446225.

Carrasco, M., J.-P. Florens, and E. Renault (2007): "Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization," *Handbook of Econometrics*, 6, 5633–5751.

Chernozhukov, V., J. A. Hausman, and W. K. Newey (2019): “Demand analysis with many prices,” National Bureau of Economic Research Discussion paper 26424.

Chiang, H. D., J. Rodrigue, and Y. Sasaki (2019): “Post-selection inference in three-dimensional panel data,” *arXiv preprint arXiv:1904.00211*.

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

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

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

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

Farrell, M. H. (2015): “Robust inference on average treatment effects with possibly more covariates than observations,” *Journal of Econometrics*, 189(1), 1–23.

Fernández-Val, I., and M. Weidner (2018): “Fixed effects estimation of large-T panel data models,” *Annual Review of Economics*, 10, 109–138.

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

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

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(1–2), 59–95.

Ghysels, E., A. Sinko, and R. Valkanov (2006): “MIDAS regressions: Further results and new directions,” *Econometric Reviews*, 26(1), 53–90.

Harding, M., and C. Lamarche (2019): “A panel quantile approach to attrition bias in Big Data: Evidence from a randomized experiment,” *Journal of Econometrics*, 211(1), 61–82.

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," 220(2), 589–605, *Journal of Econometrics* (forthcoming).

Kock, A. B. (2013): "Oracle efficient variable selection in random and fixed effects panel data models," *Econometric Theory*, 29(1), 115–152.

——— (2016): "Oracle inequalities, variable selection and uniform inference in high-dimensional correlated random effects panel data models," *Journal of Econometrics*, 195(1), 71–85.

Koenker, R. (2004): "Quantile regression for longitudinal data," *Journal of Multivariate Analysis*, 91(1), 74–89.

Kolanovic, M., and R. Krishnamachari (2017): "Big data and AI strategies: Machine learning and alternative data approach to investing," JP Morgan Global Quantitative & Derivatives Strategy Report.

Lamarche, C. (2010): "Robust penalized quantile regression estimation for panel data," *Journal of Econometrics*, 157(2), 396–408.

Lu, X., and L. Su (2016): "Shrinkage estimation of dynamic panel data models with interactive fixed effects," *Journal of Econometrics*, 190(1), 148–175.

Marsilli, C. (2014): "Variable selection in predictive MIDAS models," *Banque de France Working Paper*.

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.

Nelson, D. B. (1991): "Conditional heteroskedasticity in asset returns: A new approach," *Econometrica: Journal of the Econometric Societ*, pp. 347–370.

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

Su, L., Z. Shi, and P. C. Phillips (2016): "Identifying latent structures in panel data," *Econometrica*, 84(6), 2215–2264.

ZOU, H., T. HASTIE, AND R. TIBSHIRANI (2007): "On the degrees of freedom of the lasso," *Annals of Statistics*, 35(5), 2173–2192.

APPENDIX

A Proofs

Proof of Theorem 3.1. The proof is similar to the proof of Babii, Ghysels, and Striaukas (2021b), Theorem 3.1 and is omitted. The main difference in the proof is that instead of applying the Fuk-Nagaev inequality from Babii, Ghysels, and Striaukas (2021a), Theorem 3.1, we apply the concentration inequality from Theorem A.1 to

$$\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}u_{i,t}z_{i,t}\right|_{\infty} \qquad \text{and} \qquad \max_{j,k\in[p]}\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}z_{i,t,j}z_{i,t,k}-\Sigma_{j,k}\right|$$

under Assumptions 3.1, 3.2, and 3.3. □

Proof of Theorem 3.2. Put $r=(a^\top,b^\top)^\top$. Then we solve

$$\min_{r\in\mathbf{R}^{N+p}}\|\mathbf{y}-\mathbf{Z}r\|_{NT}^2+2\lambda\Omega(b).$$

By Fermat's rule the solution to this problem satisfies

$$\mathbf{Z}^\top(\mathbf{Z}\hat{\rho}-\mathbf{y})/NT+\lambda z^*=0_{N+p}$$

for some $z^*=\binom{0_N}{z_b^*}$, where 0_N is N-dimensional vector of zeros, $z_b^*\in\partial\Omega(\hat{\beta})$, $\hat{\rho}=(\hat{\alpha}^\top,\hat{\beta}^\top)^\top$, and $\partial\Omega(\hat{\beta})$ is the sub-differential of $b\mapsto\Omega(b)$ at $\hat{\beta}$. Taking the inner product with $\rho-\hat{\rho}$

$$\begin{aligned}\langle\mathbf{Z}^\top(\mathbf{y}-\mathbf{Z}\hat{\rho}),\rho-\hat{\rho}\rangle_{NT}&=\lambda\langle z^*,\rho-\hat{\rho}\rangle\\&=\lambda\langle z_b^*,\beta-\hat{\beta}\rangle\\&\leq\lambda\left\{\Omega(\beta)-\Omega(\hat{\beta})\right\},\end{aligned}$$

where the last line follows from the definition of the sub-differential. Rearranging this inequality and using $\mathbf{y}=\mathbf{m}+\mathbf{u}$

$$\begin{aligned}\|\mathbf{Z}(\hat{\rho}-\rho)\|_{NT}^2-\lambda\left\{\Omega(\beta)-\Omega(\hat{\beta})\right\} \leq & \langle\mathbf{Z}^\top\mathbf{u},\hat{\rho}-\rho\rangle_{NT}+\langle\mathbf{Z}^\top(\mathbf{m}-\mathbf{Z}\rho),\hat{\rho}-\rho\rangle_{NT} \\ = & \langle B^\top\mathbf{u},\hat{\alpha}-\alpha\rangle_{NT}+\langle\mathbf{X}^\top\mathbf{u},\hat{\beta}-\beta\rangle_{NT} \\ & \quad +\|\mathbf{m}-\mathbf{Z}\rho\|_{NT}\|\mathbf{Z}(\hat{\rho}-\rho)\|_{NT} \\ \leq & |B^\top\mathbf{u}/NT|_\infty|\hat{\alpha}-\alpha|_1+\Omega^*(\mathbf{X}^\top\mathbf{u}/NT)\Omega(\hat{\beta}-\beta) \\ & \quad +\|\mathbf{m}-\mathbf{Z}\rho\|_{NT}\|\mathbf{Z}(\hat{\rho}-\rho)\|_{NT} \\ \leq & |B^\top\mathbf{u}/\sqrt{N}T|_\infty\vee\Omega^*(\mathbf{X}^\top\mathbf{u}/NT)\times \\ & \times\left\{|\hat{\alpha}-\alpha|_1/\sqrt{N}+\Omega(\hat{\beta}-\beta)\right\} \\ & \quad +\|\mathbf{m}-\mathbf{Z}\rho\|_{NT}\|\mathbf{Z}(\hat{\rho}-\rho)\|_{NT},\end{aligned} \tag{A.1}$$

where the second line follows by the dual norm inequality and the Cauchy-Schwartz inequality, and Ω^* is the dual norm of Ω. By Babii, Ghysels, and Striaukas (2021b), Lemma A.2.1. and Theorem A.1 under Assumption 3.1, with probability at least $1-\delta/2$

$$\Omega^*(\mathbf{X}^\top\mathbf{u}/NT)\leq\max_{G\in\mathcal{G}}|G||\mathbf{X}^\top\mathbf{u}/NT|_\infty\lesssim\left(\frac{p}{\delta(NT)^{\kappa-1}}\right)^{1/\kappa}\vee\sqrt{\frac{\log(16p/\delta)}{NT}}.$$

Similarly, under Assumption 3.1 by Babii, Ghysels, and Striaukas (2021a), Theorem 3.1 with probability at least $1-\delta/2$

$$|B^\top\mathbf{u}/\sqrt{N}T|_\infty=\max_{i\in[N]}\left|\frac{1}{\sqrt{N}T}\sum_{t=1}^T u_{i,t}\right|\lesssim\left(\frac{N}{\delta N^{\kappa/2}T^{\kappa-1}}\right)^{1/\kappa}\vee\sqrt{\frac{\log(16N/\delta)}{NT}}.$$

Therefore, under Assumption 3.5 with probability at least $1-\delta$

$$|B^\top\mathbf{u}/NT|_\infty\vee\Omega^*(\mathbf{X}^\top\mathbf{u}/NT)\lesssim\left(\frac{(pN^{1-\kappa})\vee N^{1-\kappa/2}}{\delta T^{\kappa-1}}\right)^{1/\kappa}\vee\sqrt{\frac{\log(p\vee N/\delta)}{NT}}\lesssim\lambda.$$

In conjunction with the inequality in equation (A.1), this gives

$$\begin{aligned}\|\mathbf{Z}\Delta\|_{NT}^2 & \leq c^{-1}\lambda\left\{|\hat{\alpha}-\alpha|_1/\sqrt{N}+\Omega(\hat{\beta}-\beta)\right\}+\|\mathbf{m}-\mathbf{Z}\rho\|_{NT}\|\mathbf{Z}\Delta\|_{NT}+\lambda\left\{\Omega(\beta)-\Omega(\hat{\beta})\right\} \\ & \leq(c^{-1}+1)\lambda\left\{|\hat{\alpha}-\alpha|_1/\sqrt{N}+\Omega(\hat{\beta}-\beta)\right\}+\|\mathbf{m}-\mathbf{Z}\rho\|_{NT}\|\mathbf{Z}\Delta\|_{NT}\end{aligned} \tag{A.2}$$

for some $c > 1$ and $\Delta = \hat{\rho} - \rho$, where the second line follows by the triangle inequality. Note that the sg-LASSO penalty function can be decomposed as a sum of two seminorms $\Omega(b) = \Omega_0(b) + \Omega_1(b), \forall b \in \mathbf{R}^p$ with

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

Note also that $\Omega_1(\beta) = 0$ and $\Omega_1(\hat{\beta}) = \Omega_1(\hat{\beta} - \beta)$. Then

$$\begin{aligned} \Omega(\beta) - \Omega(\hat{\beta}) &= \Omega_0(\beta) - \Omega_0(\hat{\beta}) - \Omega_1(\hat{\beta}) \\ &\leq \Omega_0(\hat{\beta} - \beta) - \Omega_1(\hat{\beta} - \beta). \end{aligned} \tag{A.3}$$

Suppose that $\|\mathbf{m} - \mathbf{Z}\rho\|_{NT} \leq \frac{1}{2}\|\mathbf{Z}\Delta\|_{NT}$. Then it follows from the first inequality in equation (A.2) and equation (A.3) that

$$\|\mathbf{Z}\Delta\|_{NT}^2 \leq 2c^{-1}\lambda\left\{|\hat{\alpha} - \alpha|_1/\sqrt{N} + \Omega(\hat{\beta} - \beta)\right\} + 2\lambda\left\{\Omega_0(\hat{\beta} - \beta) - \Omega_1(\hat{\beta} - \beta)\right\}.$$

Since the left side of this equation is ≥ 0, this shows that

$$(1 - c^{-1})\Omega_1(\hat{\beta} - \beta) \leq (1 + c^{-1})\Omega_0(\hat{\beta} - \beta) + c^{-1}|\hat{\alpha} - \alpha|_1/\sqrt{N}$$

or equivalently

$$\Omega_1(\hat{\beta} - \beta) \leq \frac{c+1}{c-1}\Omega_0(\hat{\beta} - \beta) + (c-1)^{-1}|\hat{\alpha} - \alpha|_1/\sqrt{N}. \tag{A.4}$$

Put $\Delta_N = ((\hat{\alpha} - \alpha)^\top/\sqrt{N}, (\hat{\beta} - \beta)^\top)^\top$. Then under Assumption 3.2

$$\begin{aligned} |\Delta_N|_1 &\lesssim \Omega(\hat{\beta} - \beta) + |\hat{\alpha} - \alpha|_1/\sqrt{N} \\ &\leq \frac{2c}{c-1}\Omega_0(\hat{\beta} - \beta) + \frac{c}{c-1}|\hat{\alpha} - \alpha|_1/\sqrt{N} \\ &\lesssim |\hat{\alpha} - \alpha|_2 + \sqrt{s}|\hat{\beta} - \beta|_2 \\ &\leq \sqrt{s \vee N|\Delta_N|_2^2} \\ &\lesssim \sqrt{s \vee N|\Sigma^{1/2}\Delta_N|_2^2} \\ &= \sqrt{s \vee N\left\{\|\mathbf{Z}\Delta\|_{NT}^2 + \Delta_N^\top(\hat{\Sigma} - \Sigma)\Delta_N\right\}} \\ &\leq \sqrt{s \vee N\left\{\|\mathbf{Z}\Delta\|_{NT}^2 + |\Delta_N|_1^2|\mathrm{vech}(\hat{\Sigma} - \Sigma)|_\infty\right\}} \\ &\lesssim \sqrt{s \vee N\left\{\lambda|\Delta_N|_1 + |\Delta_N|_1^2|\mathrm{vech}(\hat{\Sigma} - \Sigma)|_\infty\right\}}. \end{aligned}$$

Consider the following event $E = \{|\mathrm{vech}(\hat{\Sigma} - \Sigma)|_\infty < 1/(2s \vee N)\}$. Under Assumption 3.1 by Theorem A.1 and Babii, Ghysels, and Striaukas (2021a), Theorem 3.1

$$\begin{aligned}
\Pr(E^c) \leq \Pr & \left(\max_{i\in[N], j\in[p]} \left| \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \{x_{i,t,j} - \mathbb{E}[x_{i,t,j}]\} \right| \geq \frac{1}{2s \vee N} \right) \\
& + \Pr \left(\max_{1\leq j \leq k \leq p} \left| \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{i,t,j} x_{i,t,k} - \mathbb{E}[x_{i,t,j} x_{i,t,k}] \right| \geq \frac{1}{2s \vee N} \right) \\
& \lesssim p(s \vee N)^{\tilde{\kappa}} T^{1-\tilde{\kappa}} (N^{1-\tilde{\kappa}/2} + pN^{1-\tilde{\kappa}}) + p(p \vee N) e^{-cNT/(s\vee N)^2}.
\end{aligned}$$

Therefore, on the event E

$$|\hat{\alpha} - \alpha|_1/\sqrt{N} + |\hat{\beta} - \beta|_1 = |\Delta_N|_1 \lesssim (s \vee N)\lambda,$$

and whence from equation (A.2) we obtain

$$\begin{aligned}
\|\mathbf{Z}\Delta\|_{NT}^2 & \lesssim \lambda \left\{ |\hat{\alpha} - \alpha|_1/\sqrt{N} + \Omega(\hat{\beta} - \beta) \right\} \\
& \lesssim \lambda |\Delta_N|_1 \\
& \leq (s \vee N)\lambda^2.
\end{aligned}$$

Suppose now that $\|\mathbf{m} - \mathbf{Z}\rho\|_{NT} > \frac{1}{2}\|\mathbf{Z}\Delta\|_{NT}$. Then, obviously,

$$\|\mathbf{Z}(\hat{\rho} - \rho)\|_{NT}^2 \leq 4\|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2.$$

Therefore, on the event E, we always have

$$\|\mathbf{Z}(\hat{\rho} - \rho)\|_{NT}^2 \lesssim (s \vee N)\lambda^2 + 4\|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2,$$

which proves the statement of the theorem. □

B Fuk-Nagaev inequality for panel data

In this section we obtain new Fuk-Nagaev concentration inequality for panel data reflecting the concentration jointly over N and T. It is worth stressing that the inequality does not follow directly from the Fuk-Nagaev inequality of Babii, Ghysels, and Striaukas (2021a) and is of independent interest for the high-dimensional panel data.[8]

[8] The direct application of the time series Fuk-Nagaev inequality of Babii, Ghysels, and Striaukas (2021a) leads to inferior concentration results for panel data.

Theorem A.1. *Let $\{\xi_{i,t} : i \in \mathbf{N}, t \in \mathbf{Z}\}$ be an array of centered random vectors in $\mathbf{R}^p$ such that $\{(\xi_{i,1}, \dots, \xi_{i,T}) : i \in \mathbf{N}\}$ are i.i.d. for each $T \geq 1$ and for each $i \geq 1$, $(\xi_{i,t})_{t\in\mathbf{Z}}$ is a stationary stochastic process such that (i) $\max_{j\in[p]} \|\xi_{i,t,j}\|_q = O(1)$ for some $q > 2$; (ii) for every $j \in [p]$, τ-mixing coefficients of $(\xi_{i,t,j})_{t\in\mathbf{Z}}$ satisfy $\tau_k^{(j)} \leq ck^{-a}, \forall k \geq 1$ for some universal constants $c > 0$ and $a > (q-1)/(q-2)$. Then for every $u > 0$*

$$\Pr\left(\left| \sum_{i=1}^{N} \sum_{t=1}^{T} \xi_{i,t} \right|_\infty > u \right) \leq c_1 pNTu^{-\kappa} + 4pe^{-c_2u^2/NT}$$

for some $c_1, c_2 > 0$ and $\kappa = ((a+1)q-1)/(a+q-1)$.

Proof of Theorem A.1. Suppose first that $p = 1$. For $a \in \mathbf{R}$ with some abuse of notation, let $[a]$ denote its integer part. For each $i = 1, 2 \dots, N$, split partial sums into blocks with at most $J \in \mathbf{N}$ summands

$$\begin{aligned} V_{i,k} &= \xi_{i,(k-1)J+1} + \cdots + \xi_{i,kJ}, \qquad k = 1, 2, \dots, [T/J] \\ V_{i,[T/J]+1} &= \xi_{i,[T/J]J+1} + \cdots + \xi_{i,T}, \end{aligned}$$

where we set $V_{i,[T/J]+1} = 0$ if $[T/J]J = T$. Let $\{U_{i,t} : i, t \geq 1\}$ be i.i.d. random variables uniformly distributed on $[0, 1]$ and independent of $\{\xi_{i,t} : i, t \geq 1\}$. Put $\mathcal{M}_{i,t} = \sigma(V_{i,1}, \dots, V_{i,t-2})$ with $t \geq 3$. For each $i \geq 1$, if $t = 1, 2$, set $V_{i,t}^* = V_{i,t}$, while if $t \geq 3$, then by Dedecker and Prieur (2004), Lemma 5, there exist random variables $V_{i,t}^* =_d V_{i,t}$ such that

1. $V_{i,t}^*$ is $\mathcal{M}_{i,t} \vee \sigma(V_{i,t}) \vee \sigma(U_{i,t})$-measurable.
2. $V_{i,t}^*$ is independent of $\mathcal{M}_{i,t}$.
3. $\|V_{i,t} - V_{i,t}^*\|_1 = \tau(\mathcal{M}_{i,t}, V_{i,t})$.

Property 1. ensures that there exists a measurable function f_i such that

$$V_{i,t}^* = f_i(V_{i,t}, V_{i,t-2}, \dots, V_{i,1}, U_{i,t}).$$

Property 2. implies that $(V_{i,2t}^*)_{t\geq1}$ and $(V_{i,2t-1}^*)_{t\geq1}$ are sequences of independent random variables for every $i \geq 1$. Moreover, $\{V_{i,2t}^* : i \geq 1, t \geq 1\}$ and $\{V_{i,2t-1}^* : i \geq 1, t \geq 1\}$ are sequences of independent random variables since $\{\xi_{i,t} : \ t = 1, \dots, T\}$ are independent over $i = 1, \dots, N$.

Decompose

$$\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\xi_{i,t}\right| \leq \left|\sum_{i=1}^{N}\sum_{t\geq 1}V_{i,2t}^{*}\right| + \left|\sum_{i=1}^{N}\sum_{t\geq 1}V_{i,2t-1}^{*}\right| + \sum_{i=1}^{N}\sum_{t=3}^{[T/J]+1}\left|V_{i,t}-V_{i,t}^{*}\right|$$
$$\triangleq I + II + III.$$

By Fuk and Nagaev (1971), Corollary 4 there exist constants $c_1, c_2 > 0$ such that

$$\begin{aligned}\Pr(I > u/3) &\leq c_1 u^{-q} N \sum_{t\geq 1}\mathbb{E}|V_{i,2t}^{*}|^q + 2\exp\left(-\frac{c_2 u^2}{N\sum_{t\geq 1}\mathrm{Var}(V_{i,2t}^{*})}\right)\\ &\leq c_1 u^{-q} N \sum_{t\geq 1}\mathbb{E}|V_{i,2t}|^q + 2\exp\left(-\frac{c_2 u^2}{NT}\right),\end{aligned}$$

where we use $V_{i,t}^{*} =_d V_{i,t}$ and

$$\sum_{t\geq 1}\mathrm{Var}(V_{i,2t}) \leq \sum_{t\geq 1}\mathrm{Var}(V_{i,t}) = O(T),$$

which follows by Babii, Ghysels, and Striaukas (2021a), Lemma A.1.2. Similarly,

$$\Pr(II > u/3) \leq c_3 u^{-q} N \sum_{t\geq 1}\mathbb{E}|V_{i,2t}|^q + 2\exp\left(-\frac{c_4 u^2}{NT}\right)$$

for some constants $c_3, c_4 > 0$. Lastly, since $\mathcal{M}_{i,t}$ and $V_{i,t}$ are separated by $J+1$ lags of $\xi_{i,t}$, we have $\tau(\mathcal{M}_{i,t}, V_{i,t}) \leq J\tau_J(J+1)$. By Markov's inequality and property 3., this gives

$$\begin{aligned}\Pr(III > u/3) &\leq \frac{3N}{u}\sum_{t=3}^{[T/J]+1}\|V_{i,t}-V_{i,t}^{*}\|_1\\ &\leq \frac{3NT}{u}\tau_{J+1}.\end{aligned}$$

Combining all estimates together

$$\begin{aligned}\Pr\left(\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\xi_{i,t}\right| > u\right) &\leq \Pr(I > u/3) + \Pr(II > u/3) + \Pr(III > u/3)\\ &\leq c_1 u^{-q} N \sum_{t\geq 1}\|V_{i,t}\|_q^q + 4e^{-c_2 u^2/NT} + \frac{3NT}{u}\tau_{J+1}\\ &\leq c_1 u^{-q} J^{q-1} NT \|\xi_{i,t}\|_q^q + \frac{3NT}{u}(J+1)^{-a} + 4e^{-c_2 u^2/NT}\end{aligned}$$

for some constants $c_1, c_2 > 0$. To balance the first two terms, we shall choose the length of blocks $J \sim u^{\frac{q-1}{q+a-1}}$, in which case we get

$$\Pr\left(\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\xi_{i,t}\right| > u\right) \leq c_1 NTu^{-\kappa} + 4e^{-c_2u^2/NT}$$

for some $c_1, c_2 > 0$.

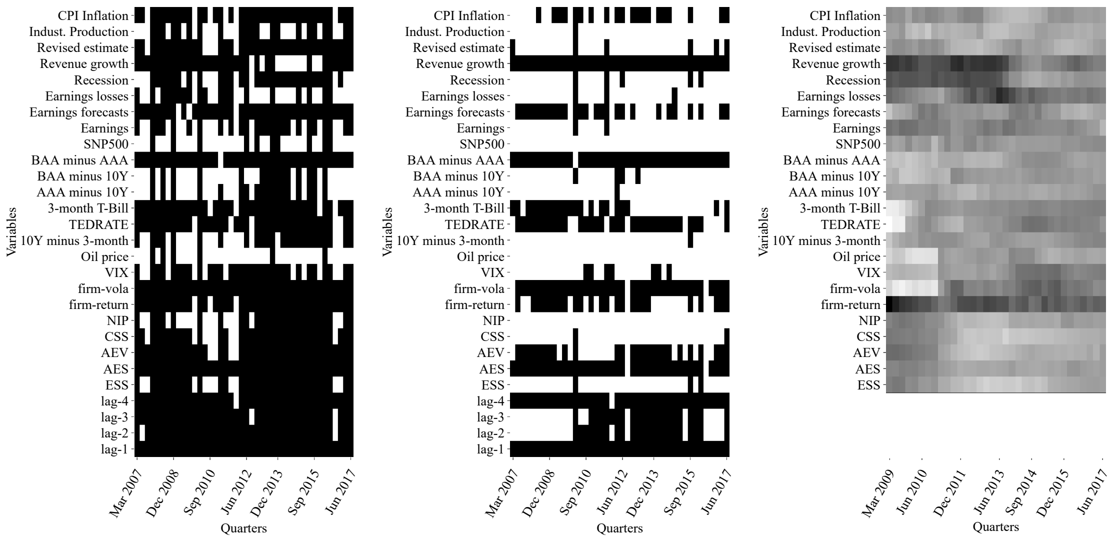
Finally, for $p > 1$, the result follows by the union bound. □

It follows from Theorem A.1 that there exists $C > 0$ such that for every $\delta \in (0, 1)$

$$\Pr\left(\left|\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\xi_{i,t}\right|_\infty \leq C\left(\frac{p}{\delta(NT)^{\kappa-1}}\right)^{1/\kappa} \vee \sqrt{\frac{\log(8p/\delta)}{NT}}\right) \geq 1 - \delta.$$

Note that the inequality reflects the concentration jointly over N and T and that tails and persistence play an important role through the mixing-tails exponent κ. The inequality is a key technical tool that allows us to handle panel data with heavier than Gaussian tails and non-negligible T and N.

C Additional empirical results



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

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

(c) Average forecast combination weights.

Figure A.1: Sparsity patterns and forecast combination weights.

RW	MSE An.-mean	MSE An.-median		sg-LASSO	elnet-U	elnet
2.331	2.339	2.088				
				Panel A. Cross-validation		
			Individual	1.545	1.606	1.606
			Pooled	**1.455**	1.489	1.499
			Fixed Effects	1.480	1.490	1.509
				Panel B. BIC		
			Individual	1.543	1.597	1.611
			Pooled	1.482	1.486	1.485
			Fixed Effects	**1.472**	1.489	1.489
				Panel C. AIC		
			Individual	1.560	1.640	1.652
			Pooled	1.487	1.491	1.494
			Fixed Effects	**1.479**	1.487	1.495
				Panel D. AICc		
			Individual	2.025	1.699	1.866
			Pooled	1.484	1.491	1.493
			Fixed Effects	**1.479**	1.487	1.495

Table A.1: Prediction results – The table reports average over firms MSEs of out-of-sample predictions. The nowcasting horizon is the current month, i.e. we predict the P/E ratio using information up to the end of current fiscal quarter. Each Panel A-D block represents different ways of calculating the tuning parameter λ. Bold entries are the best results in a block. We report the best elastic net MSEs over LASSO/ridge weight $[0, 0.2, 0.4, 0.6, 0.8, 1]$: elnet-U method is where high-frequency lags are unrestricted, elnet method is where we use only the first high-frequency lag for each covariate. We also report the best sg-LASSO specification for each tuning parameter method and each model specification, see Table 1.

RW	MSE An.-mean	MSE An.-median	F.Comb		sg-LASSO					
2.794	2.836	2.539	2.405	$\gamma =$	0	0.2	0.4	0.6	0.8	1
					Panel A. Cross-validation					
				Individual	1.808	1.817	1.836	1.864	1.889	1.884
				Pooled	1.692	1.689	**1.688**	**1.688**	**1.688**	1.689
				Fixed Effects	1.743	1.726	1.725	1.743	1.712	1.726
					Panel B. BIC					
				Individual	1.972	1.945	1.914	1.833	1.853	1.912
				Pooled	1.723	1.741	1.733	1.738	1.736	1.724
				Fixed Effects	1.760	1.734	**1.707**	1.756	1.717	1.710
					Panel C. AIC					
				Individual	1.929	1.889	1.853	1.903	1.989	2.003
				Pooled	1.737	1.735	1.729	1.728	1.732	1.734
				Fixed Effects	1.747	1.724	1.724	1.747	**1.712**	1.726
					Panel D. AICc					
				Individual	2.401	2.513	2.679	2.918	3.404	3.732
				Pooled	1.737	1.725	1.729	1.728	1.732	1.734
				Fixed Effects	1.732	1.725	1.724	1.747	**1.712**	1.726

Table A.2: Prediction results – The table reports average over firms MSEs of out-of-sample predictions for the same models as in Table 1 - discarding the first 8 quarters to compute for forecast combination weights - with additional result of prediction errors using forecast combination approach of Ball and Ghysels (2018), denoted as *F.Comb*. Hence the out-of-sample quarters start at 2009 Q1. The nowcasting horizon is the current month, i.e. we predict the P/E ratio using information up to the end of current fiscal quarter. Each Panel A-D block represents different ways of calculating the tuning parameter λ. Bold entries are the best results in a block.

D Data description

D.1 Firm-level data

The full list of firm-level data is provided in Table A.3. We also add two daily firm-specific stock market predictor variables: stock returns and a realized variance measure, which is defined as the rolling sample variance over the previous 60 days (i.e. 60-day historical volatility).

D.1.1 Firm sample selection

We select a sample of firms based on data availability. First, we remove all firms from I/B/E/S which have missing values in earnings time series. Next, we retain firms that we are able to match with CRSP dataset. Finally, we keep firms that we can match with the RavenPack dataset.

D.1.2 Firm-specific text data

We create a link table of RavenPack ID and PERMNO identifiers which enables us to merge I/B/E/S and CRSP data with firm-specific textual analysis generated data from RavenPack. The latter is a rich dataset that contains intra-daily news information about firms. There are several editions of the dataset; in our analysis, we use the Dow Jones (DJ) and Press Release (PR) editions. The former contains relevant information from Dow Jones Newswires, regional editions of the Wall Street Journal, Barron's and MarketWatch. The PR edition contains news data, obtained from various press releases and regulatory disclosures, on a daily basis from a variety of newswires and press release distribution networks, including exclusive content from PRNewswire, Canadian News Wire, Regulatory News Service, and others. The DJ edition sample starts at 1^{st} of January, 2000, and PR edition data starts at 17^{th} of January, 2004.

We construct our news-based firm-level covariates by filtering only highly relevant news stories. More precisely, for each firm and each day, we filter out news that has the *Relevance Score* (REL) larger or equal to 75, as is suggested by the RavenPack News Analytics guide and used by practitioners, see for example Kolanovic and Krishnamachari (2017). REL is a score between 0 and 100 which indicates how strongly a news story is linked with a particular firm. A score of zero means that the entity is vaguely mentioned in the news story, while 100 means the opposite. A score of 75 is regarded as a significantly relevant news story. After applying the REL filter, we apply a novelty of the news filter by using the *Event Novelty Score* (ENS); we keep data entries that have a score of 100. Like REL, ENS is a score between 0 and 100. It indicates the novelty of a news story within a 24-hour time window. A score of 100 means that a news story was not already covered by earlier announced news, while subsequently published news story score on a related event is discounted, and therefore its scores are less than 100. Therefore, with this filter, we consider only novel news stories. We focus on *five sentiment indices* that are available in both DJ and PR editions. They are:

Event Sentiment Score (ESS), for a given firm, represents the strength of the news measured using surveys of financial expert ratings for firm-specific events. The score value ranges between 0 and 100 - values above (below) 50 classify the news as being positive (negative), 50 being neutral.

Aggregate Event Sentiment (AES) represents the ratio of positive events reported on a firm compared to the total count of events measured over a rolling 91-day window in a particular news edition (DJ or PR). An event with ESS $>$ 50 is counted as a positive entry while ESS $<$ 50 as negative. Neutral news (ESS $=$ 50) and news that does not receive an ESS score does not enter into the AES computation. As ESS, the score values are between 0 and 100.

Aggregate Event Volume (AEV) represents the count of events for a firm over the last 91 days within a certain edition. As in AES case, news that receives a non-neutral ESS score is counted and therefore accumulates positive and negative news.

Composite Sentiment Score (CSS) represents the news sentiment of a given news story by combining various sentiment analysis techniques. The direction of the score is determined by looking at emotionally charged words and phrases and by matching stories typically rated by experts as having short-term positive or negative share price impact. The strength of the scores is determined by intra-day price reactions modeled empirically using tick data from approximately 100 large-cap stocks. As for ESS and AES, the score takes values between 0 and 100, 50 being the neutral.

News Impact Projections (NIP) represents the degree of impact a news flash has on the market over the following two-hour period. The algorithm produces scores to accurately predict a relative volatility - defined as scaled volatility by the average of volatilities of large-cap firms used in the test set - of each stock price measured within two hours following the news. Tick data is used to train the algorithm and produce scores, which take values between 0 and 100, 50 representing zero impact news.

For each firm and each day with firm-specific news, we compute the average value of the specific sentiment score. In this way, we aggregate across editions and groups, where the later is defined as a collection of related news. We then map the indices that take values between 0 and 100 onto $[-1, 1]$. Specifically, let $x_i \in \{\mathrm{ESS}, \mathrm{AES}, \mathrm{CSS}, \mathrm{NIP}\}$ be the average score value for a particular day and firm. We map $x_i \mapsto \bar{x}_i \in [-1, 1]$ by computing $\bar{x}_i = (x_i - 50)/50$.

	id	Frequency	Source	T-code
	Panel A.			
-	Price/Earnings ratio	quarterly	CRSP & I/B/E/S	1
-	Price/Earnings ratio consensus forecasts	quarterly	CRSP & I/B/E/S	1
	Panel B.			
1	Stock returns	daily	CRSP	1
2	Realized variance measure	daily	CRSP/computations	1
	Panel C.			
1	Event Sentiment Score (ESS)	daily	RavenPack	1
2	Aggregate Event Sentiment (AES)	daily	RavenPack	1
3	Aggregate Event Volume (AEV)	daily	RavenPack	1
4	Composite Sentiment Score (CSS)	daily	RavenPack	1
5	News Impact Projections (NIP)	daily	RavenPack	1

Table A.3: Firm-level data description table – The *id* column gives mnemonics according to data source, which is given in the second column *Source*. The column *frequency* states the sampling frequency of the variable. The column *T-code* denotes the data transformation applied to a time-series, which are: (1) not transformed, (2) Δx_t, (3) $\Delta^2 x_t$, (4) $\log(x_t)$, (5) Δ log (x_t), (6) Δ^2 log (x_t). Panel A. describes earnings data, panel B. describes quarterly firm-level accouting data, panel C. daily firm-level stock market data and panel D. daily firm-level sentiment data series.

	id	Frequency	Source	T-code
	Panel A.			
1	Industrial Production Index	monthly	FRED-MD	5
2	CPI Inflation	monthly	FRED-MD	6
	Panel B.			
1	Crude Oil Prices	daily	FRED	6
2	S&P 500	daily	CRSP	5
3	VIX Volatility Index	daily	FRED	1
4	Moodys Aaa - 10-Year Treasury	daily	FRED	1
5	Moodys Baa - 10-Year Treasury	daily	FRED	1
6	Moodys Baa - Aaa Corporate Bond	daily	FRED	1
7	10-Year Treasury - 3-Month Treasury	daily	FRED	1
8	3-Month Treasury - Effective Federal funds rate	daily	FRED	1
9	TED rate	daily	FRED	1
	Panel C.			
1	Earnings	monthly	Bybee, Kelly, Manela, and Xiu (2019)	1
2	Earnings forecasts	monthly	Bybee, Kelly, Manela, and Xiu (2019)	1
3	Earnings losses	monthly	Bybee, Kelly, Manela, and Xiu (2019)	1
4	Recession	monthly	Bybee, Kelly, Manela, and Xiu (2019)	1
5	Revenue growth	monthly	Bybee, Kelly, Manela, and Xiu (2019)	1
6	Revised estimate	monthly	Bybee, Kelly, Manela, and Xiu (2019)	1

Table A.4: Other predictor variables description table – The *id* column gives mnemonics according to data source, which is given in the second column *Source*. The column *frequency* states the sampling frequency of the variable. The column *T-code* denotes the data transformation applied to a time-series, which are: (1) not transformed, (2) Δx_t, (3) $\Delta^2 x_t$, (4) $\log(x_t)$, (5) Δ log (x_t), (6) Δ^2 log (x_t). Panel A. describes real-time monthly macro series, panel B. describes daily financial markets data and panel C. monthly news attention series.

	Ticker	Firm name	PERMNO	RavenPack ID
1	MMM	3M	22592	03B8CF
2	ABT	Abbott labs	20482	520632
3	AUD	Automatic data processing	44644	66ECFD
4	ADTN	Adtran	80791	9E98F2
5	AEIS	Advanced energy industries	82547	1D943E
6	AMG	Affiliated managers group	85593	30E01D
7	AKST	A K steel holding	80303	41588B
8	ATI	Allegheny technologies	43123	D1173F
9	AB	AllianceBernstein holding l.p.	75278	CB138D
10	ALL	Allstate corp.	79323	E1C16B
11	AMZN	Amazon.com	84788	0157B1
12	AMD	Advanced micro devices	61241	69345C
13	DOX	Amdocs ltd.	86144	45D153
14	AMKR	Amkor technology	86047	5C8D61
15	APH	Amphenol corp.	84769	BB07E4
16	AAPL	Apple	14593	D8442A
17	ADM	Archer daniels midland	10516	2B7A40
18	ARNC	Arconic	24643	EC821B
19	ATTA	AT&T	66093	251988
20	AVY	Avery dennison corp.	44601	662682
21	BHI	Baker hughes	75034	940C3D
22	BAC	Bank of america corp.	59408	990AD0
23	BAX	Baxter international inc.	27887	1FAF22
24	BBT	BB&T corp.	71563	1A3E1B
25	BDX	Becton dickinson & co.	39642	873DB9
26	BBBY	Bed bath & beyond inc.	77659	9B71A7
27	BHE	Benchmark electronics inc.	76224	6CF43C
28	BA	Boeing co.	19561	55438C
29	BK	Bank of new york mellon corp.	49656	EF5BED
30	BWA	BorgWarner inc.	79545	1791E7
31	BP	BP plc	29890	2D469F
32	EAT	Brinker international inc.	23297	732449
33	BMY	Bristol-Myers squibb co.	19393	94637C
34	BRKS	Brooks automation inc.	81241	FC01C0
35	CA	CA technologies inc.	25778	76DE40
36	COG	Cabot oil & gas corp.	76082	388E00
37	CDN	Cadence design systems inc.	11403	CC6FF5
38	COF	Capital one financial corp.	81055	055018
39	CRR	Carbo ceramics inc.	83366	8B66CE
40	CSL	Carlisle cos.	27334	9548BB
41	CCL	Carnival corporation & plc	75154	067779
42	CERN	Cerner corp.	10909	9743E5
43	CHRW	C.H. robinson worldwide inc.	85459	C659EB
44	SCHW	Charles schwab corp.	75186	D33D8C
45	CHKP	Check point software technologies ltd.	83639	531EF1
46	CHV	Chevron corp.	14541	D54E62
47	CI	CIGNA corp.	64186	86A1B9
48	CTAS	Cintas corp.	23660	BFAEB4
49	CLX	Clorox co.	46578	719477
50	KO	Coca-Cola co.	11308	EEA6B3
51	CGNX	Cognex corp.	75654	709AED
52	COLM	Columbia sportswear co.	85863	5D0337
53	CMA	Comerica inc.	25081	8CF6DD
54	CRK	Comstock resources inc.	11644	4D72C8
55	CAG	ConAgra foods inc.	56274	FA40E2
56	STZ	Constellation brands inc.	69796	1D1B07
57	CVG	Convergys corp.	86305	914819

58	COST	Costco wholesale corp.	87055	B8EF97
59	CCI	Crown castle international corp.	86339	275300
60	DHR	Danaher corp.	49680	E124EB
61	DRI	Darden restaurants inc.	81655	9BBFA5
62	DVA	DaVita inc.	82307	EFD406
63	DO	Diamond offshore drilling inc.	82298	331BD2
64	D	Dominion resources inc.	64936	977A1E
65	DOV	Dover corp.	25953	636639
66	DOW	Dow chemical co.	20626	523A06
67	DHI	D.R. horton inc.	77661	06EF42
68	EMN	Eastman chemical co.	80080	D4070C
69	EBAY	eBay inc.	86356	972356
70	EOG	EOG resources inc.	75825	A43906
71	EL	Estee lauder cos. inc.	82642	14ED2B
72	ETH	Ethan allen interiors inc.	79037	65CF8E
73	ETFC	E*TRADE financial corp.	83862	28DEFA
74	XOM	Exxon mobil corp.	11850	E70531
75	FII	Federated investors inc.	86102	73C9E2
76	FDX	FedEx corp.	60628	6844D2
77	FITB	Fifth third bancorp	34746	8377DB
78	FISV	Fiserv inc.	10696	190B91
79	FLEX	Flex ltd.	80329	B4E00D
80	F	Ford motor co.	25785	A6213D
81	FWRD	Forward air corp.	79841	10943B
82	BEN	Franklin resources inc.	37584	5B6C11
83	GE	General electric co.	12060	1921DD
84	GIS	General mills inc.	17144	9CA619
85	GNTX	Gentex corp.	38659	CC339B
86	HAL	Halliburton Co.	23819	2B49F4
87	HLIT	Harmonic inc.	81621	DD9E41
88	HIG	Hartford financial services group inc.	82775	766047
89	HAS	Hasbro inc.	52978	AA98ED
90	HLX	Helix energy solutions group inc.	85168	6DD6BA
91	HP	Helmerich & payne inc.	32707	1DE526
92	HSY	Hershey co.	16600	9F03CF
93	HES	Hess corp.	28484	D0909F
94	HON	Honeywell international inc.	10145	FF6644
95	JBHT	J.B. Hunt transport services Inc.	42877	72DF04
96	HBAN	Huntington bancshares inc.	42906	C9E107
97	IBM	IBM corp.	12490	8D4486
98	IEX	IDEX corp.	75591	E8B21D
99	IR	Ingersoll-Rand plc	12431	5A6336
100	IDTI	Integrated device technology inc.	44506	8A957F
101	INTC	Intel corp.	59328	17EDA5
102	IP	International paper co.	21573	8E0E32
103	IIN	ITT corp.	12570	726EEA
104	JAKK	Jakks pacific inc.	83520	5363A2
105	JNJ	Johnson & johnson	22111	A6828A
106	JPM	JPMorgan chase & co.	47896	619882
107	K	Kellogg co.	26825	9AF3DC
108	KMB	Kimberly-Clark corp.	17750	3DE4D1
109	KNGT	Knight transportation inc.	80987	ED9576
110	LSTR	Landstar system inc.	78981	FD4E8D
111	LSCC	Lattice semiconductor corp.	75854	8303CD
112	LLY	Eli lilly & co.	50876	F30508
113	LFUS	Littelfuse inc.	77918	D06755
114	LNC	Lincoln national corp.	49015	5C7601
115	LMT	Lockheed martin corp.	21178	96F126

116	MTB	M&T bank corp.	35554	D1AE3B
117	MANH	Manhattan associates inc.	85992	031025
118	MAN	ManpowerGroup inc.	75285	C0200F
119	MAR	Marriott international inc.	85913	385DD4
120	MMC	Marsh & mcLennan cos.	45751	9B5968
121	MCD	McDonald's corp.	43449	954E30
122	MCK	McKesson corp.	81061	4A5C8D
123	MDU	MDU resources group inc.	23835	135B09
124	MRK	Merck & co. inc.	22752	1EBF8D
125	MTOR	Meritor inc	85349	00326E
126	MTG	MGIC investment corp.	76804	E28F22
127	MGM	MGM resorts international	11891	8E8E6E
128	MCHP	Microchip technology inc.	78987	CDFCC9
129	MU	Micron technology inc.	53613	49BBBC
130	MSFT	Microsoft corp.	10107	228D42
131	MOT	Motorola solutions inc.	22779	E49AA3
132	MSM	MSC industrial direct co.	82777	74E288
133	MUR	Murphy oil corp.	28345	949625
134	NBR	Nabors industries ltd.	29102	E4E3B7
135	NOI	National oilwell varco inc.	84032	5D02B7
136	NYT	New york times co.	47466	875F41
137	NFX	Newfield exploration co.	79915	9C1A1F
138	NEM	Newmont mining corp.	21207	911AB8
139	NKE	NIKE inc.	57665	D64C6D
140	NBL	Noble energy inc.	61815	704DAE
141	NOK	Nokia corp.	87128	C12ED9
142	NOC	Northrop grumman corp.	24766	FC1B7B
143	NTRS	Northern trust corp.	58246	3CCC90
144	NUE	NuCor corp.	34817	986AF6
145	ODEP	Office depot inc.	75573	B66928
146	ONB	Old national bancorp	12068	D8760C
147	OMC	Omnicom group inc.	30681	C8257F
148	OTEX	Open text corp.	82833	34E891
149	ORCL	Oracle corp.	10104	D6489C
150	ORBK	Orbotech ltd.	78527	290820
151	PCAR	Paccar inc.	60506	ACF77B
152	PRXL	Parexel international corp.	82607	EF8072
153	PH	Parker hannifin corp.	41355	6B5379
154	PTEN	Patterson-uti energy inc.	79857	57356F
155	PBCT	People's united financial inc.	12073	449A26
156	PEP	PepsiCo inc.	13856	013528
157	PFE	Pfizer inc.	21936	267718
158	PIR	Pier 1 imports inc.	51692	170A6F
159	PXD	Pioneer natural resources co.	75241	2920D5
160	PNCF	PNC financial services group inc.	60442	61B81B
161	POT	Potash corporation of saskatchewan inc.	75844	FFBF74
162	PPG	PPG industries inc.	22509	39FB23
163	PX	Praxair inc.	77768	285175
164	PG	Procter & gamble co.	18163	2E61CC
165	PTC	PTC inc.	75912	D437C3
166	PHM	PulteGroup inc.	54148	7D5FD6
167	QCOM	Qualcomm inc.	77178	CFF15D
168	DGX	Quest diagnostics inc.	84373	5F9CE3
169	RL	Ralph lauren corp.	85072	D69D42
170	RTN	Raytheon co.	24942	1981BF
171	RF	Regions financial corp.	35044	73C521
172	RCII	Rent-a-center inc.	81222	C4FBDC
173	RMD	ResMed inc.	81736	434F38

174	RHI	Robert half international inc.	52230	A4D173
175	RDC	Rowan cos. inc.	45495	3FFA00
176	RCL	Royal caribbean cruises ltd.	79145	751A74
177	RPM	RPM international inc.	65307	F5D059
178	RRD	RR R.R. donnelley & sons co.	38682	0BE0AE
179	SLB	Schlumberger ltd. n.v.	14277	164D72
180	SCTT	Scotts miracle-gro co.	77300	F3FCC3
181	SM	SM st. mary land & exploration co.	78170	6A3C35
182	SONC	Sonic corp.	76568	80D368
183	SO	Southern co.	18411	147C38
184	LUV	Southwest airlines co.	58683	E866D2
185	SWK	Stanley black & decker inc.	43350	CE1002
186	STT	State street corp.	72726	5BC2F4
187	TGNA	TEGNA inc.	47941	D6EAA3
188	TXN	Texas instruments inc.	15579	39BFF6
189	TMK	Torchmark corp.	62308	E90C84
190	TRV	The travelers companies inc.	59459	E206B0
191	TBI	TrueBlue inc.	83671	9D5D35
192	TUP	Tupperware brands corp.	83462	2B0AF4
193	TYC	Tyco international plc	45356	99333F
194	TSN	Tyson foods inc.	77730	AD1ACF
195	X	United states Steel corp.	76644	4E2D94
196	UNH	UnitedHealth group inc.	92655	205AD5
197	VIAV	Viavi solutions inc.	79879	E592F0
198	GWW	W.W. grainger inc.	52695	6EB9DA
199	WDR	Waddell & reed financial inc.	85931	2F24A5
200	WBA	Walgreens boots alliance inc.	19502	FACF19
201	DIS	Walt disney co.	26403	A18D3C
202	WAT	Waters corp.	82651	1F9D90
203	WBS	Webster financial corp.	10932	B5766D
204	WFC	Wells fargo & co.	38703	E8846E
205	WERN	Werner enterprises inc.	10397	D78BF1
206	WABC	Westamerica bancorp	82107	622037
207	WDC	Western digital corp.	66384	CE96E7
208	WHR	Whirlpool corp.	25419	BDD12C
209	WFM	Whole foods market inc.	77281	319E7D
210	XLNX	Xilinx inc.	76201	373E85

Table A.5: Final list of firms – The table contains the information about the full list of firms: tickers, firm names, CRSP PERMNO code and RavenPack ID. Tickers and firm names are taken as of June, 2017. PERMNO and RavenPack ID columns are used to match firms and firm news data.