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

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

July 8, 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, large N and T asymptotics.

*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 (ML) 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 a mixed frequency panel data setting. Second, we study the benefits of using the cross-sectional dimension for prediction with panel data paying particular attention to the issues of fat-tailed series which is relevant for the application involving financial time series. Third, we develop the debiased heteroskedasticity autocorrelation consistent (HAC) inference for regularized panel data regressions. 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 others. However, to the best of our knowledge, the existing literature relates mostly to the microeconometric 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) debiased HAC inference for regularized panel data, and (4) the sg-LASSO regularization of Simon, Friedman, Hastie, and Tibshirani (2013) in a panel data setting.

We recognize that the economic and financial time series data are often persistent with fat tails. To that end, we introduce a new Fuk-Nagaev concentration inequality for long panels. Using this inequality, we obtain oracle inequalities for the sg-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 (2016), Hansen (2007), Alvarez and Arellano (2003), Hahn and Kuersteiner (2002), and Phillips and Moon (1999), among others. Importantly, our theory covers the LASSO and the group-LASSO estimators as special cases of sg-LASSO.

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 company and therefore used by analysts and investment professionals to base their decisions on which stocks to pick for their investment portfolios. Typically investors rely on consensus forecasts of earnings made by a pool of analysts. We therefore choose such consensus forecasts as the benchmark for our proposed machine learning

methods. In addition, we also compare our proposed new methods with the MIDAS regression forecast combination approach used by Ball and Ghysels (2018) and a simple random walk model.

Our high-frequency regressors include traditional macro and financial series as well as non-standard series generated by textual analysis of financial news. We consider structured pooled and fixed effects sg-LASSO panel data regressions with mixed frequency data (sg-LASSO-MIDAS). The fixed effects estimator yields sparser models compared to pooled regressions with revenue growth and the first lag of the dependent variable selected throughout the out-of-sample period. The BAA minus AAA bond yield spread, firm-level volatility, and news textual analysis aggregate event Sentiment index are also selected very frequently. Our results show the superior performance of sg-LASSO-MIDAS 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 as many days are without firms-specific news and we impute zero values. The nice property of our mixed frequency data treatment with dictionaries is that 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 sg-LASSO panel data regressions appear in Section 3. Section 4 develops the debiased HAC inference for regularized panel data regressions. We report on a Monte Carlo study in Section 5 which provides insights about the validity of the asymptotic analysis in finite sample settings typically encountered in empirical applications. The results of our empirical application analyzing price earnings ratios for a panel of individual firms are reported in Section 6. Section 7 concludes. All technical details and detailed data descriptions appear in the Appendix and the Online 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 a set S. For $b \in \mathbf{R}^p$, its ℓ_q norm is denoted as $|b|_q = (\sum_{j\in[p]} |b_j|^q)^{1/q}$ if $q \in [1, \infty)$ and $|b|_\infty = \max_{j\in[p]} |b_j|$ if $q = \infty$. For a 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 $\text{vech}(A) \in \mathbf{R}^{p(p+1)/2}$ be its vectorization consisting of the lower triangular and the diagonal elements. Let A_G be a sub-matrix consisting of rows of A corresponding to indices in $G \subset [p]$. If $G = \{j\}$ for some $j \in [p]$, then we simply write $A_G = A_j$. Let $\|A\|_\infty = \max_{j\in[p]} |A_j|$ be the matrix norm. 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 $\beta_k = (\beta_{1,k}, \ldots, \beta_{m,k})^\top \in \mathbf{R}^m$. More generally, the frequency can also be specific to the predictor $k \in [K]$, in which case we would have m_k instead of m. 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 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), with some abuse of notation, we estimate a weight function ω parameterized by $\beta_k \in \mathbf{R}^L$ with $L < m$

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

where

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

and $(w_l)_{l \geq 0}$ is a collection of L approximating functions, called the *dictionary*. An example of a dictionary is the set of orthogonal Legendre polynomials on $[0,1]$ that can be computed via the Rodrigues' formula $w_l(s) = \frac{1}{l!} \frac{\mathrm{d}^l}{\mathrm{d}s^l}(s^2 - s)^l$.[1] 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, such as the Almon (1965) lag structure. 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 a convex

[1] The Legendre polynomials have the universal approximation property and can approximate any continuous function uniformly on $[0,1]$. At the same time they can generate a rich family of MIDAS weights with a relatively small number of parameters which is attractive in time series applications where the signal-to-noise ratio is often low.

optimization. To that end, define $\mathbf{x}_i = (X_{i,1}W, \ldots, X_{i,K}W)$, where for each $k \in [K]$, $X_{i,k} = (x_{i,t-(j-1)/m,k})_{t\in[T],j\in[m]}$ is a $T \times m$ matrix of predictors and $W = (w_l((j-1)/m)/m)_{j\in[m],0\leq l\leq L-1}$ is an $m \times L$ matrix corresponding to the dictionary $(w_l)_{l\geq 0}$. In addition, let $\mathbf{y}_i = (y_{i,1+h}, \ldots, y_{i,T+h})^\top$ and $\mathbf{u}_i = (u_{i,1}, \ldots, u_{i,T})^\top$. Then the regression equation after stacking time series observations for each $i \in [N]$ is

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

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

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

where $B = I_N \otimes \iota$, $\alpha = (\alpha_1, \ldots, \alpha_N)$, and $\otimes$ is the Kronecker product.

The MIDAS approach allows us to effectively reduce 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.[2]

Given that the number of potential predictors K can be large, additional regularization can improve the predictive performance in small samples. To that end, we take advantage of the sg-LASSO regularization that was shown to be attractive for individual time series ML regressions in Babii, Ghysels, and Striaukas (2021b). The fixed effects panel data estimator with sparse-group regularization solves

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

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

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

is a regularizing functional, which is a linear combination of LASSO and group LASSO penalties. The parameter $\gamma \in [0,1]$ determines the relative weights of the

[2]See Ball and Easton (2013) and Ball and Gallo (2018) for further discussion on interpreting the shape of MIDAS polynomials in accounting data applications considered in our empirical application.

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 $[p] = \{1, 2, \ldots, p\}$, 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.[3] 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

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 panel data projection 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.

[3] See Babii (2021) for a continuous-time mixed-frequency regression where the group size is allowed to increase with the sample size under the in-fill asymptotics.

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.

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 $\mathbf{R}$.[4] 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

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

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,[5] 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 sg-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 $\alpha_1 = \cdots = \alpha_N = \alpha$. The pooled sg-LASSO estimator $\hat{\rho} = (\hat{\alpha}, \hat{\beta}^\top)^\top$ solves

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

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

[5] The class of α-mixing processes is too restrictive for the predictive linear projection model with covariates and autoregressive lags, see also Babii, Ghysels, and Striaukas (2021b), Proposition A.3.1.

Define (a) $z_{i,t} = (1, x_{i,t}^\top)^\top$, where $x_{i,t} \in \mathbf{R}^p$ is a vector of predictors, (b) $u_i = (u_{i,1}, \ldots, u_{i,T})$ and (c) $x_i = (x_{i,1}^\top, \ldots, x_{i,T}^\top)^\top$ for $i \in [N]$. The following assumption imposes mild restrictions on the data.

Assumption 3.1 (Data)**.** $\{(u_i, x_i^\top)^\top : i \in \mathbf{N}\}$ *are independent vectors in* $\mathbf{R}^{(p+1)} \times \mathbf{R}^T$ *such that (i)* $\max_{i\in[N],t\in[T],j\in[p+1]} \|u_{i,t} z_{i,t,j}\|_q = O(1)$ *for some* $q > 2$*; (ii) the* τ*-mixing coefficients of* $(u_{i,t} z_{i,t})_{t\in\mathbf{Z}}$ *satisfy* $\max_{i\in[N],j\in[p+1]} \tau_{k-1}^{(i,j)} = O(k^{-a}), \forall k \geq 1$ *with* $a > (q-1)/(q-2)$*; (iii)* $\max_{i\in[N],t\in[T],j,k\in[p+1]} \|z_{i,t,j} z_{i,t,k}\|_{\tilde{q}} = O(1)$ *for some* $\tilde{q} > 2$*; (iv) the* τ*-mixing coefficients of* $\mathrm{vech}((z_{i,t} z_{i,t}^\top))_{t\in\mathbf{Z}}$ *satisfy* $\max_{i\in[N],j\in[(p+1)(p+2)/2]} \tilde{\tau}_{k-1}^{(i,j)} \leq \tilde{c} k^{-\tilde{a}}, \forall k \geq 1$ *with* $\tilde{c} > 0$ *and* $\tilde{a} > (\tilde{q}-1)/(\tilde{q}-2)$*.*

Note that we do not impose stationarity over $t \in \mathbf{Z}$ and 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 slowly – at a polynomial rate as measured by the τ-mixing coefficients.

Next, we assume that the $(1+p) \times (1+p)$ matrix $\Sigma_{N,T} = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \mathbb{E}[z_{i,t} z_{i,t}^\top]$ exists and is non-singular uniformly over N, T, p:

Assumption 3.2 (Covariance matrix)**.** *The smallest eigenvalue of* $\Sigma_{N,T}$ *is uniformly bounded away from zero by some universal constant* $\gamma_{\min} > 0$*.*

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 $\Sigma_{N,T}$; see Babii, Ghysels, and Striaukas (2021b). In addition, we can allow for $\gamma_{\min} \downarrow 0$ as $N, T, p \uparrow \infty$, in which case $\gamma_{\min}^{-1}$ would slow down the convergence rates in oracle inequalities and could be interpreted as a measure of ill-posedness; see also 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 the tail parameter q and the persistence parameter 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 non-trivial 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 that 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 $\gamma = 0$. Define $r_{N,T}^{\text{pooled}} = s^{\tilde{\kappa}} p^2/(NT)^{\tilde{\kappa}-1} + p^2 \exp(-cNT/s^2)$.

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 sg-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 vanishes sufficiently fast 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 sg-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 two practical consequences: (1) one may want to exclude (or suitably transform) 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 a 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. Therefore, 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

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

We will assume that the smallest eigenvalue of $\Sigma_{N,T}$ is uniformly bounded away from zero by some constant. Note that if $x_{i,t} \sim N(0, I_p)$, then $\Sigma_{N,T} = I_{N+p}$ 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 $r^{\text{fe}}_{N,T} = p(s \vee N)^{\tilde{\kappa}} T^{1-\tilde{\kappa}}(N^{1-\tilde{\kappa}/2} + pN^{1-\tilde{\kappa}}) + p(p \vee N)e^{-cNT/(s \vee N)^2}$ 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. Then with probability at least $1 - \delta - O(r^{\text{fe}}_{N,T})$*

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

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) $\|\mathbf{m} - \mathbf{Z}\rho\|^2_{NT} = O_P((s \vee N)\lambda^2)$; (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*

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

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 Debiased inference

In this section, we develop the debiased inferential methods for pooled panel data regressions. For a vector $\rho \in \mathbf{R}^{p+1}$, we use $\rho_G \in \mathbf{R}^{|G|}$ to denote the subvector of elements of $\rho \in \mathbf{R}^{p+1}$ indexed by $G \subset [p+1]$. Let $B = \hat{\Theta}\mathbf{Z}^\top(\mathbf{y} - \mathbf{Z}\hat{\rho})/NT$ denote the bias-correction for the sg-LASSO estimator, where $\hat{\Theta}$ is the nodewise LASSO estimator of the precision matrix $\Theta = \Sigma^{-1}$, where $\Sigma = \mathbb{E}[z_{i,t}z_{i,t}^\top]$. For pooled panel data, this estimator can be obtained as follows:

1. For each $j \in [p+1]$, let $\hat{\mu}_j = (\hat{\mu}_{j,1}, \ldots, \hat{\mu}_{j,p})^\top$ be a solution to
$$\min_{\mu \in \mathbf{R}^p} \|\mathbf{Z}_j - \mathbf{Z}_{-j}\mu\|_{NT}^2 + 2\lambda_j|\mu|_1,$$
where $\mathbf{Z}_j$ is $NT \times 1$ vector of stacked observations $\{z_{i,t,j} \in \mathbf{R} : i \in [N], t \in [T]\}$ and $\mathbf{Z}_{-j}$ is the $NT \times p$ matrix of stacked observations $\{(z_{i,t,k})_{k \neq j} \in \mathbf{R}^p : i \in [N], t \in [T]\}$. Put
$$\hat{\sigma}_j^2 = \|\mathbf{Z}_j - \mathbf{Z}_{-j}\hat{\mu}_j\|_{NT}^2 + \lambda_j|\hat{\mu}_j|,$$

2. Compute $\hat{\Theta} = \hat{B}^{-1}\hat{C}$, where $\hat{B} = \mathrm{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_{p+1}^2)$, and
$$\hat{C} = \begin{pmatrix} 1 & -\hat{\mu}_{1,1} & \cdots & -\hat{\mu}_{1,p} \\ -\hat{\mu}_{2,1} & 1 & \cdots & -\hat{\mu}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\mu}_{p,1} & \cdots & -\hat{\mu}_{p,p} & 1 \end{pmatrix}.$$

Let $v_{i,t,j} = z_{i,t,j} - \sum_{k \neq j} \mu_{j,k} z_{i,t,k}$ be the regression error for j^{th} nodewise LASSO regression. Let s_j be the number of non-zero elements in j^{th} row of precision matrix Θ_j, and put $S = \max_{j \in G} s_j$, and $s^* = s \vee S$.

The following assumption describes an additional set of conditions for the debiased central limit theorem.

Assumption 4.1. *(i)* $\sup_z \mathbb{E}[u_{i,t}^2 | z_{i,t} = z] = O(1)$*; (ii)* $\|\Theta_G\|_\infty = O(1)$ *for* $G \subset [p+1]$ *of fixed size; (iii) the long run variance of* $(u_{i,t}^2)_{t\in\mathbf{Z}}$ *and* $(v_{i,t,j}^2)_{t\in\mathbf{Z}}$ *exists for every* $j \in G$*; (iv)* $s^{*2} \log^2 p/T \to 0$ *and* $p/\sqrt{T^{\kappa-2}\log^\kappa p} \to 0$*; (v)* $\|\mathbf{m} - \mathbf{Z}\rho\|_{NT} = o_P(1/\sqrt{NT})$*; (vi) for every* $j, l \in [p]$ *and* $k \geq 0$*, the* τ*-mixing coefficients of* $(u_{i,t}u_{i,t+k}x_{i,t,j}x_{i,t+k,l})_{t\in\mathbf{Z}}$ *are* $\check{\tau}_t \leq ct^{-d}$ *for some universal constants* $c > 0$ *and* $d > 1$*; (vi) for each* i*,* $\{(u_{i,t}, z_{i,t}^\top)^\top : t \in \mathbf{Z}\}$ *is a stationary process that is also i.i.d. over* i*, Assumption 3.1 holds with* $a > (q-1)/(q-2) \vee (q\delta+1)/(q-2-\delta)$ *with* $q > 2+\delta$ *and* $\delta > 0$*.*

Assumption 4.1 (i) requires that the conditional variance of the regression error is bounded. Condition (ii) requires that the rows of the precision matrix have bounded ℓ_1 norm and is a plausible assumption in the high-dimensional setting, where the inverse covariance matrix is often sparse. Condition (iii) is a mild restriction needed for the consistency of the sample variance of regression errors. The rate conditions in (iv) is similar to the condition used in Babii, Ghysels, and Striaukas (2021a). Lastly, condition (v) is trivially satisfied when the projection coefficients are sparse and, more generally, it requires that the misspecification error vanishes asymptotically sufficiently fast.

The following result describes a large-sample approximation to the distribution of the debiased sg-LASSO estimator with serially correlated heavy-tailed errors.

Theorem 4.1. *Suppose that Assumptions 3.1, 3.2, 3.3, 3.4, and 4.1 are satisfied for the sg-LASSO regression and for each nodewise LASSO regression* $j \in G$*. Then*

$$\sqrt{NT}(\hat{\rho}_G + B_G - \rho_G) \xrightarrow{d} N(0, \Xi_G)$$

with the long-run variance $\Xi_G = \lim_{T\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T u_{i,t}\Theta_G z_{i,t}\right)$*.*

Theorem 4.1 applies to panel data consisting of non-Gaussian, heavy-tailed, and persistent time series under the large N and T large sample approximation. In contrast to the fixed T approximations, Theorem 4.1 leads to more precise inference, e.g., the standard errors and the length of confidence intervals would scale at $O(1/\sqrt{NT})$ rate instead of $O(1/\sqrt{N})$ that we typically encounter for fixed T approximations.

To estimate Ξ_G, we can use the following pooled HAC estimator

$$\hat{\Xi}_G = \frac{1}{N}\sum_{i=1}^N \sum_{|k|<T} K\left(\frac{k}{M_T}\right)\hat{\Gamma}_{k,i},$$

where $\hat{\Gamma}_{k,i} = \hat{\Theta}_G\left(\frac{1}{T}\sum_{t=1}^{T-k}\hat{u}_{i,t}\hat{u}_{i,t+k}x_{i,t}x_{i,t+k}^\top\right)\hat{\Theta}_G^\top$, $\hat{u}_{i,t}$ is the sg-LASSO residual, and $\hat{\Gamma}_{-k,i} = \hat{\Gamma}_{k,i}^\top$. The kernel function $K : \mathbf{R} \to [-1,1]$ with $K(0) = 1$ is puts less weight

on more distant noisy covariances, while $M_T \uparrow \infty$ is a bandwidth (or lag truncation) parameter; see Babii, Ghysels, and Striaukas (2021a) for more details as well as formal results on the validity of HAC-based inference using sg-LASSO residuals.

5 Monte Carlo experiments

In this section, we investigate the finite sample nowcasting performance of machine learning methods applied to large dimensional panel data. We consider the unstructured elastic net with UMIDAS and sg-LASSO with MIDAS. Both methods require selecting two tuning parameters λ and γ. In the case of sg-LASSO, γ is the relative weight of LASSO and group LASSO penalties while in the case of the elastic net γ interpolates between LASSO and ridge. We compute the optimal λ using BIC and we report results on a grid $\{0, 0.2, \dots, 1\}$.

5.1 Simulation Design

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

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

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

We set $\rho_1 = 0.4$, $\rho_2 = 0.01$, and take the number of relevant high frequency regressors $K = 6$. We are interested in quarterly/monthly data, and use four quarters of data for the high frequency regressors so that $m = 12$, which covers four low frequency lags of each high frequency regressor. The high frequency regressors are generated as K i.i.d. realizations of the univariate autoregressive (AR) process $x_h = \rho x_{h-1} + \varepsilon_h$, where $\rho = 0.6$ and either $\varepsilon_h \sim_{i.i.d.} N(0,1)$ or $\varepsilon_h \sim_{i.i.d.}$ student-$t(5)$, where h denotes the high-frequency sampling. We rely on a commonly used weighting scheme in the MIDAS literature, namely $\omega(s; \beta_k)$ for $k = 1, 2, \dots, 6$ are determined by beta densities respectively equal to Beta$(1,3)$ for $k = 1, 4$, Beta$(2,3)$ for $k = 2, 5$, and Beta$(2,2)$ for $k = 3, 6$; see Ghysels, Sinko, and Valkanov (2007) or Ghysels and Qian

(2019), for further details. The MIDAS regressions are estimated using Legendre polynomials of degree $L = 3$. Lastly, we draw the intercepts $\alpha \sim \text{Uniform}(-4, 4)$.

We also consider DGPs featuring fixed effects. They are identical to pooled MIDAS panel data model except for the common intercept α which replaced by

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

The individual fixed effects are simulated as $\alpha_i \sim_{\text{i.i.d}} \text{Uniform}(-4, 4)$ and are kept fixed throughout the experiment.

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

5.2 Simulation results

Table 1 covers the average mean squared forecast errors for one-step-ahead nowcasts. We report results for pooled panel data (left block) and fixed effects (right block) estimators. First, for all DGPs and both estimators, structured sg-LASSO-MIDAS performs better compared to unstructured elastic net. In the case of sg-LASSO-MIDAS the best performance is achieved for $\gamma \notin \{0, 1\}$ for both pooled panel data and fixed effects cases, while $\gamma = 0$, i.e. ridge regression, seems to dominate in the case of elastic net for both the pooled and fixed effects cases. For the student-$t(5)$ and large dimensional DGP, we observe a decrease in the performance for all methods. However, the decrease in the performance is larger for the student-$t(5)$ DGP, suggesting that heavy-tailed data may have a stronger impact on the performance of the estimators.

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

	Pooled panel data						Fixed effects					
$\gamma =$	0	0.2	0.4	0.6	0.8	1	0	0.2	0.4	0.6	0.8	1
						Panel A. Baseline scenario						
						N = 25, T = 50						
Elnet	1.647	1.687	1.703	1.733	1.744	1.750	1.828	2.157	2.462	2.456	2.507	2.510
sg-LASSO-MIDAS	1.556	1.374	1.365	1.378	1.398	1.428	1.674	1.588	1.528	1.540	1.566	1.668
						N = 75, T = 50						
Elnet	1.290	1.311	1.318	1.322	1.323	1.325	1.464	1.575	1.693	1.718	1.720	1.720
sg-LASSO-MIDAS	1.211	1.210	1.211	1.212	1.218	1.256	1.257	1.230	1.262	1.262	1.304	1.398
						N = 25, T = 100						
Elnet	1.345	1.360	1.378	1.391	1.402	1.409	1.512	1.768	1.889	1.921	1.930	1.939
sg-LASSO-MIDAS	1.225	1.225	1.230	1.258	1.274	1.322	1.463	1.342	1.315	1.313	1.360	1.421
						Panel B. Student-$t(5)$						
						N = 25, T = 50						
Elnet	1.846	1.989	2.061	2.066	2.073	2.075	2.197	2.445	2.669	2.699	2.713	2.725
sg-LASSO-MIDAS	1.926	1.554	1.545	1.554	1.575	1.635	1.980	1.951	1.924	1.945	1.998	1.991
						N = 75, T = 50						
Elnet	1.425	1.444	1.466	1.475	1.484	1.491	1.634	1.721	1.818	1.868	1.886	1.894
sg-LASSO-MIDAS	1.333	1.324	1.339	1.340	1.340	1.360	1.424	1.396	1.395	1.391	1.393	1.530
						N = 25, T = 100						
Elnet	1.592	1.592	1.601	1.638	1.658	1.670	1.834	1.890	1.982	1.989	1.990	1.998
sg-LASSO-MIDAS	1.415	1.392	1.385	1.404	1.411	1.476	1.630	1.591	1.581	1.561	1.591	1.668
						Panel C. Large dimensional ($p = 100$)						
						N = 25, T = 50						
Elnet	1.992	1.664	1.720	1.735	1.740	1.746	2.351	2.347	2.034	2.132	2.166	2.192
sg-LASSO-MIDAS	1.757	1.413	1.387	1.399	1.424	1.484	2.131	1.951	1.601	1.710	1.826	1.893
						N = 75, T = 50						
Elnet	1.406	1.278	1.285	1.289	1.291	1.292	1.523	1.579	1.681	1.705	1.712	1.717
sg-LASSO-MIDAS	1.224	1.217	1.217	1.217	1.224	1.278	1.326	1.245	1.272	1.276	1.327	1.399
						N = 25, T = 100						
Elnet	1.405	1.393	1.401	1.412	1.421	1.429	1.789	1.601	1.727	1.756	1.773	1.776
sg-LASSO-MIDAS	1.299	1.277	1.277	1.292	1.310	1.342	1.549	1.408	1.386	1.378	1.427	1.481

Table 1: The table reports simulation results for nowcasting accuracy for pooled and fixed effects estimators. Panel A. reports results for the baseline DGP, Panel B. for student-$t(5)$ DGP and Panel C. for large dimensional DGP with 100 time-varying covariates. We vary the cross-sectional dimension $N \in \{25, 75\}$ and time series dimension $T \in \{50, 100\}$.

6 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 textual analysis of financial news. We apply pooled and individual 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. Moreover, our analysis includes results for sg-LASSO-MIDAS panel data models which include the median consensus analysts predictions. Adding the concensus forecast as a regressor allows us to address besides the question of ML versus analysts also the topic of a combined ML/analyst nowcasts – a theme explored by Ball and Ghysels (2018). Our analysis includes formal significance testing of predictors in the augmented sg-LASSO-MIDAS panel data model which allows us to determine whether analysts take all relevant information to them fully into account.

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

The remainder of the section is structured as follows. We start with a short review of the data, with more detailed descriptions and tables appearing in Online Appendix Section OA.2, followed by a summary of the methods and empirical results.

6.1 Data description

The full sample consists of observations between the 1^{st} of January, 2000 and the 30^{th} of June, 2017. Due to the lagged dependent variables in the models, our effective sample starts the third fiscal quarter of 2000. We use the first 25 observations for the initial sample, and use the remaining 42 observations for evaluating the

out-of-sample forecasts, which we obtain by using an expanding window forecasting scheme. We collect data from CRSP and I/B/E/S to compute the quarterly P/E ratios and firm-specific financial covariates; RavenPack is used to compute daily firm-level textual-analysis-based data; real-time monthly macroeconomic series are from the FRED-MD dataset, see McCracken and Ng (2016) for more details; FRED is used to compute daily financial markets data and, lastly, monthly news attention series extracted from the *Wall Street Journal* articles is retrieved from Bybee, Kelly, Manela, and Xiu (2019).[6] Online Appendix Section OA.2 provides a detailed description of the data sources.[7]

P/E ratio and analysts' forecasts sample construction. Our target variable is the P/E ratio for each firm. To compute it, we use CRSP stock price data and I/B/E/S earnings data. Earnings data are subject to release delays of 1 to 2 months depending on the firm and quarter. Therefore, to reflect the real-time information flow, we 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. The consensus forecast of the P/E ratio is computed using the same end of quarter price data which is divided by the earnings consensus forecast value. The consensus is computed by taking all individual prediction values up to the end of the quarter and aggregating those values by taking either the mean or the median. To compute the target variable, we adjust for publication lags and use prices of the publication date instead of the end of fiscal quarter prices. More precisely, suppose we predict the P/E ratio for the first quarter. Earnings are typically published with 1 to 2 months delay; say for a particular firm the data is published on the $25th$ of April. In this case, we record the stock price for the firm on $25th$ of April, and divide it by the earnings announced on that date.

6.2 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

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

[7] In particular, firm-level variables, including P/E ratios, are described in Online Appendix Table OA.5, and the other predictor variables in Online Appendix Table OA.6. The list of all firms we consider in our analysis appears in Online Appendix Table OA.7.

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) of Hurvich and Tsai (1989). Assuming that $y_{i,t}|x_{i,t}$ are i.i.d. draws from $N(\alpha_i + x_{i,t}^\top \beta, \sigma^2)$, the log-likelihood of the sample is

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

Then, the BIC criterion is

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

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

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

Lastly, the corrected Akaike information criteria 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 is typically a better choice when p is large relatively to the sample size. We report results for each of these four choices of the tuning parameters.

6.3 Models and main results

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

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

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

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

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

and compute predictions as

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

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

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

Second, we consider predictions of P/E implied by analysts earnings nowcasts using the information up to time $t+1$, i.e.

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

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

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

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

The main results for pooled panel data and fixed effects sg-LASSO-MIDAS regressions are reported in Table 2, while additional results for longer horizon predictions, unstructured LASSO estimators and the forecast combination approach appear in Online Appendix Tables OA.1-OA.3.

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

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

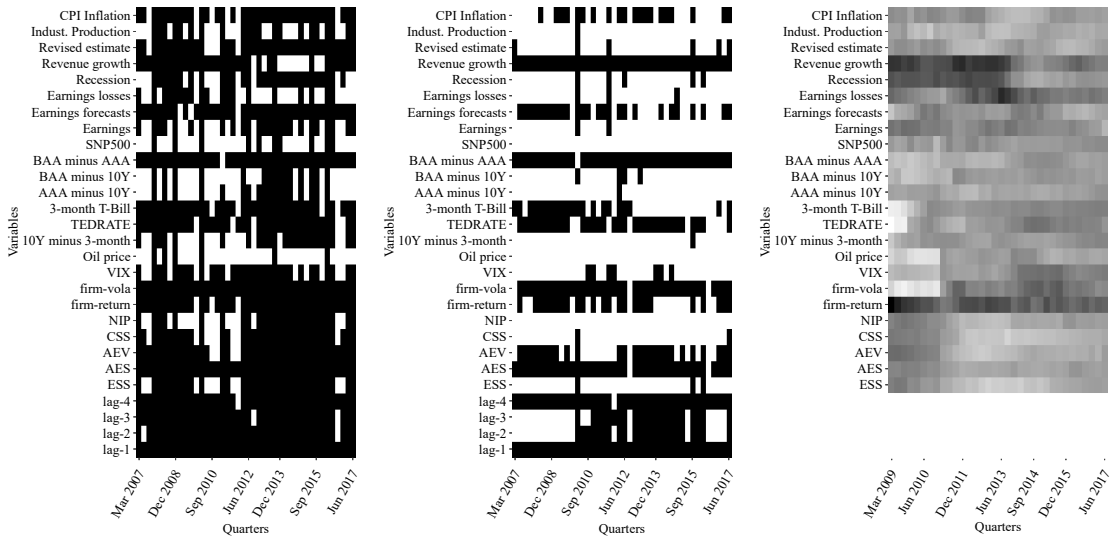
The first entries to Table 2 show that analysts-based predictions, both median and mean, have much larger mean squared forecast errors (MSEs) compared to model-based predictions. This is also the case for the RW predictions. The sharp increase in quality of model- versus analyst-based predictions indicates the usefulness of machine learning methods to nowcast P/E ratios, see Tables 2 Panel A1-D1 and OA.2. A better performance is achieved for almost all machine learning methods - single firm or panel data regressions - and all tuning parameter choices.[8] Table 2 also reports results for panel data ML methods augmented with median consensus analysts forecasts (see Panels A2-D2). Notably, it shows that the augmented models further improve upon ML-only models.

Turning to the comparison of model-based predictions, we see from the results in Table 2 that sg-LASSO-MIDAS panel data models improve the quality of predictions in comparison to individual sg-LASSO-MIDAS models irrespective of the γ weight or the tuning parameter choice. This indicates that panel data structures are relevant for nowcasting P/E ratios.[9] Among the panel data models, we observe that fixed effects regressions improve over the pooled regressions in most cases except when cross-validation is used, namely compare Panels A1 with Panel B1-D in Tables 2 and OA.3. The pooled model tuned by cross-validation seems to yield the best overall performance. In general, one can expect that cross-validation improves prediction performance over different tuning methods as it is directly linked to empirical risk minimization. In the case of fixed effects, however, we may lose the predictive gain due to the smaller samples with each fold used in estimating the model. Lastly, the best results per tuning parameter block seem to be achieved when $\gamma \notin \{0, 1\}$, indicating that both sparsity within the group and at the group level matters for prediction performance.

In Figure 1, we plot the sparsity patterns of the selected covariates for the two best-performing methods: (a) pooled sg-LASSO regressions, tuned using cross-validation with $\gamma = 0.4$, and (b) fixed effects sg-LASSO model with BIC tuning parameter and the same γ parameter. We also plot the forecast combination weights which are averaged over firms. The plots in Figure 1 reveal that the fixed effects estimator yields sparser models compared to pooled regressions, and the sparsity pattern is clearer. In the fixed-effects case, the revenue growth and the first lag of the dependent variable are selected throughout the out-of-sample period. BAA mi-

[8]Similar findings for one-quarter ahead predictions are reported in Table OA.1. The unstructured panel data methods and the forecast combination approach also yield more accurate forecasts, see Online Appendix Table OA.2-OA.3. The latter confirms the findings of Ball and Ghysels (2018).

[9]We also report similar findings for unstructured estimators (see Table OA.2) and one quarter ahead forecasts (see Table OA.1).



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

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

(c) Average forecast combination weights.

Figure 1: Sparsity patterns and forecast combination weights.

nus AAA bond yield spread, firm-level volatility, and the aggregate event sentiment index are also selected quite frequently. Similarly, these variables are selected in the pooled regression, but the pattern is less apparent. The forecast combination weights seem to yield similar, yet more dispersed patterns.[10] 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 coefficients.

In addition it is worth noting that the textual news data analytics also appear in the models according the results displayed in Figure 1. These are the ESS, AES, AEV, CSS and NEP regressors described in detail in Appendix Section OA.2. 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 as many days are without firms-specific news. For such days,

[10]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.

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.

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

6.4 Significance test of nowcasts

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

When testing the full sample of pooled nowcasts, the gain in prediction accuracy is not significant even though the MSEs are much lower for the panel data sg-LASSO regressions relative to the consensus forecasts. The result may not be surprising, however, as some firms have a large number of outliers. We report three additional columns where we pool the prediction based on the relative performance of machine learning methods versus analysts. First, we pool all errors for firms where sg-LASSO-MIDAS and elastic net outperform the analysts' median consensus forecasts, i.e. has 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.[11]

The results reveal heterogeneous performance for sg-LASSO-MIDAS and elastic net panel data regressions. First, for the pool of firms where both structured sg-LASSO-MIDAS and unstructured elastic net outperform the analysts, the gains over the analysts predictions are significant for both machine learning techniques.

[11]We do not report results for the pool of firms for which the elastic net outperforms analysts and the sg-LASSO-MIDAS does not, since there is only one such firm in the case of fixed effects regressions, while in the case of pooled regressions there are no such firms.

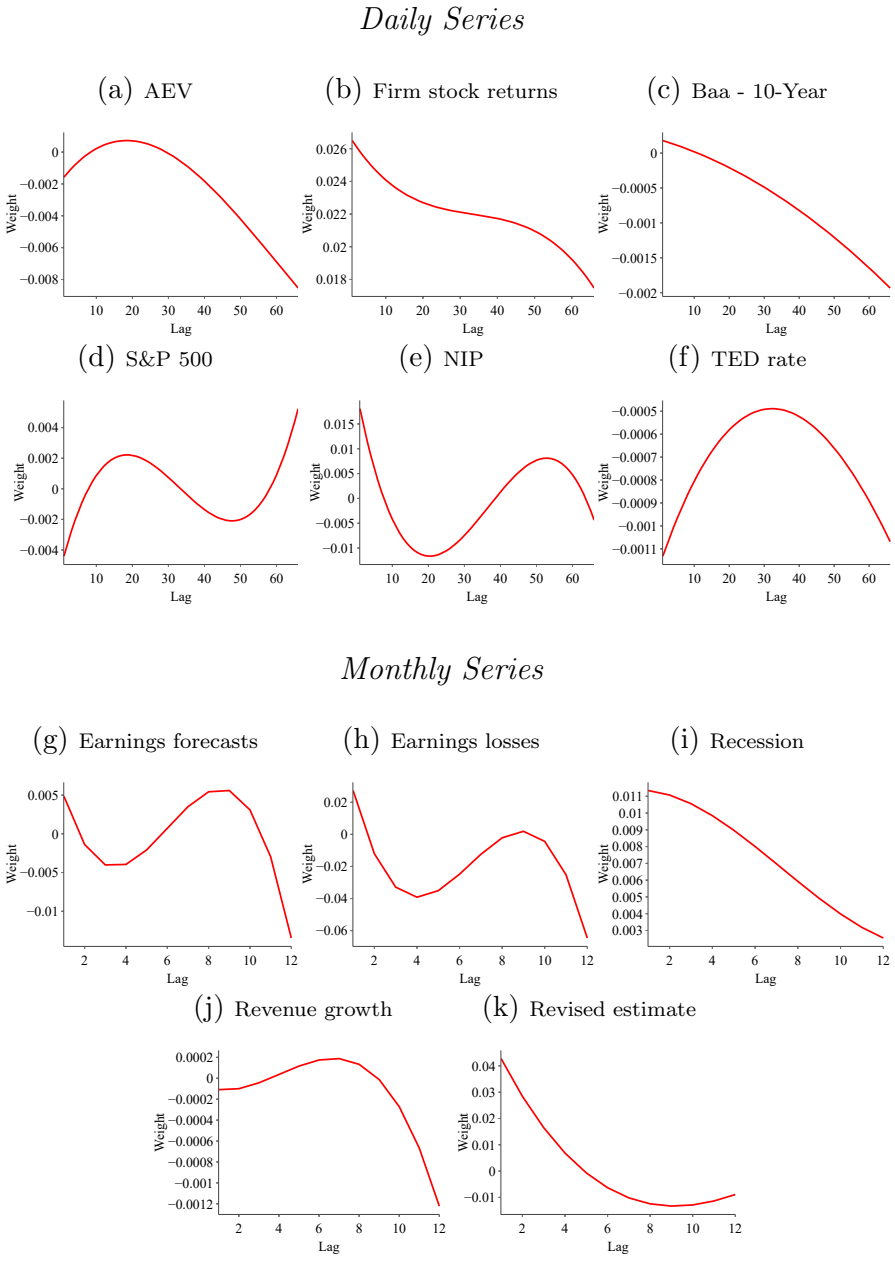


Figure 2: Weighting schemes for various covariates

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.

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

Table 3: Forecasting performance significance – The table reports the Diebold and Mariano (1995) test statistic for pooled nowcasts comparing machine learning panel data regressions with analysts' implied median consensus forecasts. We compare panel models that have the smallest forecast error per tuning parameter block in Table 2 (sg-LASSO-MIDAS) and Table OA.2 (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*.

Large differences in prediction accuracy for different pools of P/E ratios may relate to the heavy-tailedness of regression errors.[12] In Table 4, we report the maximum likelihood estimates of the degrees of freedom parameter of a student-t distribution for the in-sample residuals pooled as in Table 3.[13] The smaller values indicate that the tails are heavier, while the larger values correspond to lighter, closer to Gaussian, tails. In line with our theory, the results show that LASSO-type regressions yield much more accurate predictions when the residuals are less heavy-tailed. Interestingly, for the pool of firms where analysts' predictions are more accurate than both

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

[13]We follow a parametric approach since the time series are relatively short, see however also Online Appendix, Table OA.4 for the nonparametric tail index estimates.

machine learning methods (column *none*), tails of the residuals appear to be the heaviest.

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

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

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

6.5 Significance tests for predictors

The out-of-sample nowcasting analysis showed that combined machine learning and analysts nowcasts improve over ML-only and analysts-only predictions. We now assess the relative importance of each covariate in a combined panel data regression model by using HAC-based inference discussed in Section 4.

To analyze the significance, we estimate the pooled sg-LASSO-MIDAS panel data model on a full sample of the data. We compute the tuning parameter λ using 5-fold

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

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

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

The results are reported in Table 5. First, the median analyst forecasts, the third autoregressive lag, and the BAA minus 10Y yield spread are the only significant variables at 1% apart from five news attention series: Earnings, Earnings losses, Recession, Revenue growth and Revised estimate. The recession news attention series is an important predictor for US GDP nowcasting application, see Babii, Ghysels, and Striaukas (2021b), while the other four series are directly linked to earnings. Aggregate Event Sentiment (AES) and Treasury yield spread (10Y minus 3M) are additional significant variables at the 5% significance level. AES captures firm-level sentiment while 10Y minus 3M is thought to predict future recessions. Overall, the significant variables are either linked with economic conditions or to earnings themselves.

7 Conclusions

This paper introduces a new class of high-dimensional panel data regression models with dictionaries and sg-LASSO regularization. This type of regularization is an especially attractive choice for predictive panel data regressions, where the low- and/or the high-frequency lags define a clear group structure. The estimator nests the LASSO and the group LASSO estimators as special cases. Our theoretical treatment allows for heavy-tailed data frequently encountered in financial time series. To that end, we obtain a new panel data concentration inequality of the Fuk-Nagaev type for τ-mixing processes as well as oracle inequalities and debiased HAC inference for the panel data sg-LASSO estimator.

Our empirical analysis sheds light on the advantage of the regularized panel data regressions for nowcasting corporate earnings. We focus on nowcasting the P/E ratio of 210 US firms and find that the regularized panel data regressions outperform several benchmarks, including analysts consensus forecasts. Furthermore, we find that the regularized machine learning regressions outperform forecast combinations and that the panel data models improve upon single time series regressions for individual firms. In addition, our analysis also shows that ML methods augmented with consensus analysts predictions further improve the quality of P/E ratio nowcasts. Using the theory of HAC-based inference for pooled panel data models developed in our paper, we show the that textual analysis data, which are related to macro and earnings data, and interest rate spreads are highly significance predictors which go beyond median consensus and autoregressive lags.

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

References

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

Alvarez, J., and M. Arellano (2003): "The time series and cross-section asymptotics of dynamic panel data estimators," *Econometrica*, 71(4), 1121–1159.

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.

Apostol, T. M. (1974): *Mathematical analysis*. Pearson.

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

Babii, A. (2021): "High-dimensional mixed-frequency IV regression," *Journal of Business and Economic Statistics (forthcoming)*.

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

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 and Economic Statistics (forthcoming)*.

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

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

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

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

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.

Billingsley, P. (1995): *Probability and Measure*. John Wiley & Sons.

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," in *Handbook of Econometrics - Volume 6B*, ed. by J. J. Heckman, and E. E. Leamer, pp. 5633–5751. Elsevier.

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 (2016): "Individual and time effects in nonlinear panel models with large N, T," *Journal of Econometrics*, 192(1), 291–312.

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

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

——— (2007): "MIDAS regressions: Further results and new directions," *Econometric Reviews*, 26(1), 53–90.

Hahn, J., and G. Kuersteiner (2002): "Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large," *Econometrica*, 70(4), 1639–1657.

Hansen, C. B. (2007): "Asymptotic properties of a robust variance matrix estimator for panel data when T is large," *Journal of Econometrics*, 141(2), 597–620.

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

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.

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.

Phillips, P. C. B., and H. R. Moon (1999): "Linear regression limit theory for nonstationary panel data," *Econometrica*, 67(5), 1057–1111.

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. B. 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. By Fermat's rule, the pooled sg-LASSO satisfies

$$\mathbf{Z}^\top(\mathbf{Z}\hat{\rho} - \mathbf{y})/NT + \lambda z^* = 0_{p+1}$$

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

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

where the last line follows from the definition of the subdifferential. Since $\mathbf{y} = \mathbf{m} + \mathbf{u}$, the inequality can be rewritten as

$$\begin{aligned}\|\mathbf{Z}(\hat{\rho} - \rho)\|_{NT}^2 - \lambda\left\{\Omega(\rho) - \Omega(\hat{\rho})\right\} &\leq \langle \mathbf{Z}^\top(\mathbf{Z}\rho - \mathbf{y}), \rho - \hat{\rho}\rangle_{NT} \\ &= \langle \mathbf{Z}^\top\mathbf{u}, \hat{\rho} - \rho\rangle_{NT} + \langle \mathbf{m} - \mathbf{Z}\rho, \mathbf{Z}(\hat{\rho} - \rho)\rangle_{NT}.\end{aligned}$$

By the dual norm inequality $\langle \mathbf{Z}^\top\mathbf{u}, \hat{\rho} - \rho\rangle_{NT} \leq \Omega^*(\mathbf{Z}^\top\mathbf{u}/NT)\Omega(\hat{\rho} - \rho)$, where Ω^* is the dual norm of Ω. Then by Babii, Ghysels, and Striaukas (2021b), Lemma A.2.1

$$\begin{aligned}\Omega^*(\mathbf{Z}^\top\mathbf{u}/NT) &\leq \gamma|\mathbf{Z}^\top\mathbf{u}/NT|_\infty + (1-\gamma)\max_{G\in\mathcal{G}}|\mathbf{Z}_G^\top\mathbf{u}/NT|_2 \\ &\leq \max_{G\in\mathcal{G}}\sqrt{|G|}|\mathbf{Z}^\top\mathbf{u}/NT|_\infty \\ &\leq \lambda/c,\end{aligned}$$

where the last line follows from Theorem A.1 with probability at least $1-\delta$ and Assumption 3.3 for some $c > 1$. Therefore,

$$\|\mathbf{Z}\Delta\|_{NT}^2 - \lambda\left\{\Omega(\rho) - \Omega(\hat{\rho})\right\} \leq \frac{\lambda}{c}\Omega(\Delta) + \|\mathbf{m} - \mathbf{Z}\rho\|_{NT}\|\mathbf{Z}\Delta\|_{NT} \text{ with } \Delta = \hat{\rho} - \rho. \quad \text{(A.1)}$$

Note that the sg-LASSO penalty function can be decomposed as a sum of two semi-norms $\Omega(r) = \Omega_0(r) + \Omega_1(r), \forall r \in \mathbf{R}^{1+p}$ with

$$\Omega_0(r) = \gamma|r_{S_0}|_1 + (1-\gamma)\sum_{G\in\mathcal{G}_0}|r_G|_2 \quad \text{and} \quad \Omega_1(r) = \gamma|r_{S_0^c}|_1 + (1-\gamma)\sum_{G\in\mathcal{G}_0^c}|r_G|_2.$$

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

$$\begin{aligned}\Omega(\rho) - \Omega(\hat{\rho}) &= \Omega_0(\rho) - \Omega_0(\hat{\rho}) - \Omega_1(\hat{\rho}) \\ &\leq \Omega_0(\hat{\rho} - \rho) - \Omega_1(\hat{\rho} - \rho) = \Omega_0(\Delta) - \Omega_1(\Delta).\end{aligned} \quad \text{(A.2)}$$

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

$$\begin{aligned}\|\mathbf{Z}\Delta\|_{NT}^2 &\leq 2\frac{\lambda}{c}\Omega(\Delta) + 2\lambda\left\{\Omega_0(\Delta) - \Omega_1(\Delta)\right\} \\ &= 2\frac{\lambda}{c}\left\{\Omega_1(\Delta) + \Omega_0(\Delta)\right\} + 2\lambda\left\{\Omega_0(\Delta) - \Omega_1(\Delta)\right\}\end{aligned}$$

Since the left side of this equation is greater or equal to zero, this shows that

$$\Omega_1(\Delta) \leq \frac{c+1}{c-1}\Omega_0(\Delta). \tag{A.3}$$

Put $\Sigma_{N,T} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}[z_{i,t}z_{i,t}^\top]$. Therefore,

$$\begin{aligned}\Omega(\Delta) &\leq \frac{2c}{c-1}\Omega_0(\Delta) \leq \frac{2c}{c-1}\sqrt{s|\Delta|_2^2} \leq \frac{2c}{c-1}\sqrt{\frac{s}{\gamma_{\min}}|\Sigma_{N,T}^{1/2}\Delta|_2^2} \\ &= \frac{2c}{c-1}\sqrt{\frac{s}{\gamma_{\min}}\left\{\|\mathbf{Z}\Delta\|_{NT}^2 + \Delta^\top(\hat{\Sigma} - \Sigma_{N,T})\Delta\right\}} \\ &\leq \frac{2c}{c-1}\sqrt{\frac{s}{\gamma_{\min}}\left\{\|\mathbf{Z}\Delta\|_{NT}^2 + \Omega(\Delta)\Omega^*((\hat{\Sigma} - \Sigma_{N,T})\Delta)\right\}} \\ &\leq \frac{2c}{c-1}\sqrt{\frac{s}{\gamma_{\min}}\left\{2(1+c^{-1})\lambda\Omega(\Delta) + \Omega^2(\Delta)G^*|\mathrm{vech}(\hat{\Sigma} - \Sigma_{N,T})|_\infty\right\}},\end{aligned}$$

where we set $G^* = \max_{G\in\mathcal{G}}\sqrt{|G|}$ and use Hölder's inequality, inequalities in equations (A.1) and (A.3), Assumption 3.2, $\hat{\Sigma} = \mathbf{Z}^\top\mathbf{Z}/NT$, and Babii, Ghysels, and Striaukas (2021b), Lemma A.2.1. This shows that with probability at least $1-\delta$

$$\Omega(\Delta) \leq \frac{4c^2 s}{(c-1)^2\gamma_{\min}}\left\{2(1+c^{-1})\lambda + \Omega(\Delta)G^*|\mathrm{vech}(\hat{\Sigma} - \Sigma_{N,T})|_\infty\right\}. \tag{A.4}$$

Consider the following event $E = \{|\mathrm{vech}(\hat{\Sigma} - \Sigma_{N,T})|_\infty < (2c^*G^*s)^{-1}\}$ with $c^* = (3c+1)^2/(\gamma_{\min}(c-1)^2)$, and note that under Assumption 3.1 by Theorem A.1

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

for some $c > 0$. On the event E, the inequality in equation (A.4) implies $\Omega(\Delta) \lesssim s\lambda$, and whence from the equation (A.1) by the triangle inequality

$$\|\mathbf{Z}\Delta\|_{NT}^2 \leq 2(1+c^{-1})\lambda\Omega(\Delta) \lesssim s\lambda^2.$$

Therefore, we obtain the statement of the theorem as long as $\|\mathbf{m} - \mathbf{Z}\rho\|_{NT} \leq \frac{1}{2}\|\mathbf{Z}\Delta\|_{NT}$. Suppose now that $\|\mathbf{m} - \mathbf{Z}\rho\|_{NT} > \frac{1}{2}\|\mathbf{Z}\Delta\|_{NT}$. Then

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

Therefore, the first statement of the theorem always holds with probability at least $1 - \delta - O(r_{N,T}^{\text{pooled}})$

$$\|\mathbf{Z}\Delta\|_{NT}^2 \lesssim s\lambda^2 + \|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2.$$

For the second statement, suppose first that

$$\Omega_1(\Delta) \leq 2\frac{c+1}{c-1}\Omega_0(\Delta). \tag{A.5}$$

Then by the same arguments as before, on the event E, we have

$$\begin{aligned}
\Omega(\Delta) &\leq \left(1 + 2\frac{c+1}{c-1}\right)\Omega_0(\Delta) \\
&\leq \frac{3c+1}{c-1}\sqrt{\frac{s}{\gamma_{\min}}\left\{\|\mathbf{Z}\Delta\|_{NT}^2 + \frac{1}{2c^* s}\Omega^2(\Delta)\right\}} \\
&= \sqrt{\frac{(3c+1)^2}{(c-1)^2\gamma_{\min}}s\|\mathbf{Z}\Delta\|_{NT}^2 + \frac{1}{2}\Omega^2(\Delta)}
\end{aligned}$$

or simply

$$\Omega(\Delta) \leq \sqrt{2}\frac{(3c+1)}{(c-1)}\sqrt{\frac{s}{\gamma_{\min}}}\|\mathbf{Z}\Delta\|_{NT} \lesssim s\lambda + \sqrt{s}\|\mathbf{m} - \mathbf{Z}\rho\|_{NT},$$

where we use the first statement of the theorem. On the other hand, if the inequality in equation (A.5) does not hold, then the inequality in equation (A.3) also does not hold, which implies that

$$\|\mathbf{m} - \mathbf{Z}\rho\|_{NT} > \frac{1}{2}\|\mathbf{Z}\Delta\|_{NT}.$$

Then since $\|\mathbf{Z}\Delta\|_{NT} \geq 0$ from (A.1) we obtain

$$\begin{aligned}
0 &\leq \frac{1}{c}\Omega(\Delta) + \Omega(\rho) - \Omega(\hat{\rho}) + \frac{2}{\lambda}\|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2 \\
&\leq \frac{1}{c}\Omega(\Delta) + \Omega_0(\Delta) - \Omega_1(\Delta) + \frac{2}{\lambda}\|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2,
\end{aligned}$$

where we use equation (A.2). Since $\Omega(\Delta) = \Omega_1(\Delta) + \Omega_0(\Delta)$

$$\begin{aligned}
\Omega_1(\Delta) &\leq \frac{c+1}{c-1}\Omega_0(\Delta) + \frac{2c}{\lambda(c-1)}\|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2 \\
&\leq \frac{1}{2}\Omega_1(\Delta) + \frac{2c}{\lambda(c-1)}\|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2,
\end{aligned}$$

where we use the fact that the inequality in equation (A.5) does not hold. Therefore,

$$\Omega_1(\Delta) \leq \frac{4c}{\lambda(c-1)} \|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2,$$

which shows that

$$\Omega(\Delta) \lesssim \Omega_1(\Delta) \leq \frac{4c}{\lambda(c-1)} \|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2.$$

Therefore, with probability at least $1 - \delta - O(r_{N,T}^{\text{pooled}})$, we always have

$$\Omega(\Delta) \lesssim s\lambda + \sqrt{s}\|\mathbf{m} - \mathbf{Z}\rho\|_{NT} + \frac{1}{\lambda}\|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2.$$

The result follows from the equivalence between Ω and $|.|_1$ norms provided that groups have fixed size. □

Proof of Theorem 3.2. By Fermat's rule the solution to the fixed effects regression satisfies

$$\mathbf{Z}^\top(\mathbf{Z}\hat{\rho} - \mathbf{y})/NT + \lambda z^* = 0_{N+p}, \text{ for some } z^* = \begin{pmatrix} 0_N \\ z_b^* \end{pmatrix},$$

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

$$\begin{aligned}\|\mathbf{Z}(\hat{\rho} - \rho)\|_{NT}^2 - \lambda\left\{\Omega(\beta) - \Omega(\hat{\beta})\right\} \leq& \langle \mathbf{Z}^\top\mathbf{u}, \hat{\rho} - \rho\rangle_{NT} + \langle \mathbf{Z}^\top(\mathbf{m} - \mathbf{Z}\rho), \hat{\rho} - \rho\rangle_{NT} \\ \leq& \langle B^\top\mathbf{u}, \hat{\alpha} - \alpha\rangle_{NT} + \langle \mathbf{X}^\top\mathbf{u}, \hat{\beta} - \beta\rangle_{NT} \\ &+ \|\mathbf{m} - \mathbf{Z}\rho\|_{NT}\|\mathbf{Z}(\hat{\rho} - \rho)\|_{NT} \\ \leq& |B^\top\mathbf{u}/NT|_\infty|\hat{\alpha} - \alpha|_1 + \Omega^*(\mathbf{X}^\top\mathbf{u}/NT)\Omega(\hat{\beta} - \beta) \\ &+ \|\mathbf{m} - \mathbf{Z}\rho\|_{NT}\|\mathbf{Z}(\hat{\rho} - \rho)\|_{NT} \\ \leq& |B^\top\mathbf{u}/\sqrt{NT}|_\infty \vee \Omega^*(\mathbf{X}^\top\mathbf{u}/NT) \\ &\times \left\{|\hat{\alpha} - \alpha|_1/\sqrt{N} + \Omega(\hat{\beta} - \beta)\right\} \\ &+ \|\mathbf{m} - \mathbf{Z}\rho\|_{NT}\|\mathbf{Z}(\hat{\rho} - \rho)\|_{NT},\end{aligned} \tag{A.6}$$

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 $1-\delta/2$

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

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

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

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.6), this gives

$$\begin{aligned}\|\mathbf{Z}\Delta\|_{NT}^2 &\leq c^{-1}\lambda\left\{|\hat{\alpha}-\alpha|_1/\sqrt{N} + \Omega(\hat{\beta}-\beta)\right\} \\ &\quad + \|\mathbf{m}-\mathbf{Z}\rho\|_{NT}\|\mathbf{Z}\Delta\|_{NT} + \lambda\left\{\Omega(\beta)-\Omega(\hat{\beta})\right\} \\ &\leq (c^{-1}+1)\lambda\left\{|\hat{\alpha}-\alpha|_1/\sqrt{N} + \Omega(\hat{\beta}-\beta)\right\} + \|\mathbf{m}-\mathbf{Z}\rho\|_{NT}\|\mathbf{Z}\Delta\|_{NT}\end{aligned} \tag{A.7}$$

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 semi-norms $\Omega(b) = \Omega_0(b) + \Omega_1(b), \forall b \in \mathbf{R}^p$ with

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

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

$$\begin{aligned}\Omega(\beta)-\Omega(\hat{\beta}) &= \Omega_0(\beta)-\Omega_0(\hat{\beta})-\Omega_1(\hat{\beta}) \\ &\leq \Omega_0(\hat{\beta}-\beta)-\Omega_1(\hat{\beta}-\beta).\end{aligned} \tag{A.8}$$

Suppose that $\|\mathbf{m}-\mathbf{Z}\rho\|_{NT} \leq \frac{1}{2}\|\mathbf{Z}\Delta\|_{NT}$. Then from the first inequality in equation (A.7) and equation (A.2), we obtain

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

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

$$\Omega_1(\hat{\beta}-\beta) \leq \frac{c+1}{c-1}\Omega_0(\hat{\beta}-\beta) + (c-1)^{-1}|\hat{\alpha}-\alpha|_1/\sqrt{N}. \tag{A.9}$$

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

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

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.7) 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. □

Proof of Theorem 4.1. By Fermat's rule, the pooled sg-LASSO estimator in equation (3) satisfies

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

for some $z^* \in \partial\Omega(\hat{\rho})$. Rearranging this expression and multiplying by $\hat{\Theta}$

$$\hat{\rho} - \rho + \hat{\Theta}\lambda z^* = \hat{\Theta}\mathbf{Z}^\top\mathbf{u}/NT + (I - \hat{\Theta}\hat{\Sigma})(\hat{\rho} - \rho) + \hat{\Theta}\mathbf{Z}^\top(\mathbf{m} - \mathbf{Z}\rho)/NT,$$

where we use $\hat{\Sigma} = \mathbf{Z}^\top\mathbf{Z}/NT$ and $\mathbf{y} = \mathbf{m} + \mathbf{u}$. Plugging λz^* from the first-order conditions and multiplying by $\sqrt{NT}$

$$\sqrt{NT}(\hat{\rho} - \rho + B) = \hat{\Theta}\mathbf{Z}^\top\mathbf{u}/\sqrt{NT} + \sqrt{NT}(I - \hat{\Theta}\hat{\Sigma})(\hat{\rho} - \rho) + \hat{\Theta}\mathbf{Z}^\top(\mathbf{m} - \mathbf{Z}\rho)/\sqrt{NT}.$$

Then for a group of regression coefficients $G \subset [p+1]$, we have

$$\begin{aligned}\sqrt{NT}(\hat{\rho}_G - \rho_G + B_G) &= \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} u_{i,t}\Theta_G z_{i,t} + \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} u_{i,t}(\hat{\Theta}_G - \Theta_G)z_{i,t} \\ &\quad + \sqrt{NT}(I - \hat{\Theta}\hat{\Sigma})_G(\hat{\rho} - \rho) + \hat{\Theta}_G\mathbf{Z}^\top(\mathbf{m} - \mathbf{Z}\rho)/\sqrt{NT} \\ &\triangleq I_{N,T} + II_{N,T} + III_{N,T} + IV_{N,T}.\end{aligned}$$

We will show that by Theorem A.1, $I_{N,T} \xrightarrow{d} N(0, \Xi_G)$ as $N, T \to \infty$. To that end, by Minkowski's inequality under Assumptions 3.1 (i) and 4.1 (ii)

$$\begin{aligned}\max_{i\in[N],j\in G}\|u_{i,t}\Theta_j z_{i,t}\|_q &\leq \max_{i\in[N],j\in G}\sum_{k=1}^{p+1}\|u_{i,t}z_{i,t,k}\Theta_{j,k}\|_q \\ &\leq \|\Theta_G\|_\infty \max_{i\in[N],j\in G,k\in[p+1]}\|u_{i,t}z_{i,t,k}\|_q = O(1).\end{aligned}$$

Lastly, under Assumption 4.1 (i), for every $i, N \in \mathbf{N}$,

$$\begin{aligned}\lim_{T\to\infty}\mathrm{Var}(u_{i,t}\Theta_G z_{i,t}) &= \lim_{T\to\infty}\Theta_G\mathrm{Var}(u_{i,t}z_{i,t})\Theta_G^\top \\ &\lesssim \lim_{T\to\infty}\Theta_G\Sigma\Theta_G = (\Theta_G^\top)_G < \infty\end{aligned}$$

since groups have a fixed size. In conjunction with Assumption 3.1 (ii), this verifies conditions of Theorem A.1 and shows that $I_{N,T} \xrightarrow{d} N(0, \Xi_G)$.

Next,

$$|II_{N,T}| \leq \|\hat{\Theta}_G - \Theta_G\|_\infty \left| \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} u_{i,t} z_{i,t} \right|_\infty$$
$$= O_P\left(\frac{Sp^{1/\kappa}}{(NT)^{1-1/\kappa}} \vee S\sqrt{\frac{\log p}{NT}} \right) O_P\left(\frac{p^{1/\kappa}}{(NT)^{1/2-1/\kappa}} \vee \sqrt{\log p} \right) = o_P(1),$$

where we use Proposition OA.3.1 and Theorem A.1. Similarly by Proposition OA.3.1 and Corollary 3.1

$$|III_{N,T}| \leq \sqrt{NT} \max_{j \in G} |(I - \hat{\Theta}\hat{\Sigma})_j|_\infty |\hat{\rho} - \rho|_1$$
$$= O_P\left(\frac{p^{1/\kappa}}{(NT)^{1/2-1/\kappa}} \vee \sqrt{\log p} \right) O_P\left(\frac{sp^{1/\kappa}}{(NT)^{1-1/\kappa}} \vee s\sqrt{\frac{\log p}{NT}} \right) = o_P(1).$$

Lastly, by the Cauchy-Schwartz inequality

$$|IV_{N,T}|_\infty \leq \max_{j \in G} |\mathbf{Z}\hat{\Theta}_j^\top|_2 \|\mathbf{m} - \mathbf{Z}\rho\|_{NT} = \max_{j \in G} \sqrt{\hat{\Theta}_j^\top \hat{\Sigma} \hat{\Theta}_j} o_P(1)$$
$$\leq \|\hat{\Theta}_G\|_\infty \sqrt{|\text{vech}(\hat{\Sigma})|_\infty} o_P(1) = o_P(1),$$

where the second line follows under Assumption 4.1 (v), and the last by Proposition OA.3.1 and Theorem A.1 under maintained assumptions. □

B Concentration and moment inequalities

In this section we present a suitable for us Rosenthal's moment inequality for dependent data and a new Fuk-Nagaev concentration inequality for panel data reflecting the concentration jointly over N and T.

For a random vector $\xi_{i,t} = (\xi_{i,t,1}, \ldots, \xi_{i,t,p}) \in \mathbf{R}^p$, let $\tau_k^{(i,j)}$ denote the τ-mixing coefficient of $\xi_{i,t,j}$. The following result describes a Fuk-Nagaev concentration inequality for panel data. It is worth mentioning that the inequality does not follow from Babii, Ghysels, and Striaukas (2021a) and is of independent interest for the high-dimensional panel data.[14]

Theorem A.1. *Let $\{\xi_{i,t} : i \in [N], t \in [T]\}$ be an array of centered random vectors in $\mathbf{R}^p$ such that $(\xi_{i,1}, \ldots, \xi_{i,T})$ are independent over i and (i) $\max_{i \in [N], t \in [T], j \in [p]} \|\xi_{i,t,j}\|_q = O(1)$*

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

for some $q > 2$; (ii) $\max_{i\in[N],j\in[p]} \tau_k^{(i,j)} = O(k^{-a})$ for some $a > (q-1)/(q-2)$. 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 universal constants $c_1, c_2 > 0$ 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 $i \in [N]$, split the partial sum into blocks with at most $J \in \mathbf{N}$ summands

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

where we set $V_{i,[[T/J]]+1} = 0$ if $[[T/J]]J = T$. Let $\{U_{i,t} : i \in [N], t \in [T]\}$ be i.i.d. random variables uniformly distributed on $(0,1)$ and independent of $\{\xi_{i,t} : i \in [N], t \in [T]\}$. Put $\mathcal{M}_{i,t} = \sigma(V_{i,1}, \ldots, V_{i,t-2})$ for every $t \geq 3$. For each $i \in [N]$, 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. $V_{i,t}^* \perp\!\!\!\perp (V_{i,1}, \ldots, V_{i,t-2})$.
3. $\|V_{i,t} - V_{i,t}^*\|_1 = \tau(\mathcal{M}_{i,t}, V_{i,t})$.

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

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

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

Decompose

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

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

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

where we use $V^*_{i,t} =_d V_{i,t}$ and $\sum_{i=1}^N \sum_{t\geq 1} \mathrm{Var}(V_{i,2t}) = O(T)$, which follows from Babii, Ghysels, and Striaukas (2021a), Lemma A.1.2 under assumptions (i) and (ii). Similarly,

$$\Pr(II > u/3) \leq c_1 u^{-q} \sum_{i=1}^{N} \sum_{t\geq 1} \mathbb{E}|V_{i,2t}|^q + 2\exp\left(-\frac{c_2 u^2}{NT}\right).$$

Finally, since $\mathcal{M}_{i,t}$ and $V_{i,t}$ are separated by $J+1$ lags of $\xi_{i,t}$, we have $\tau(\mathcal{M}_{i,t}, V_{i,t}) \leq J\tau_J^{(i,j)}(J+1)$. By Markov's inequality and property 3., this gives

$$\Pr(III > u/3) \leq \frac{3}{u} \sum_{i=1}^{N} \sum_{t=3}^{[[T/J]]+1} \|V_{i,t} - V^*_{i,t}\|_1 \leq \frac{3NT}{u} \max_{i\in[N]} \tau^{(i,1)}_{J+1}.$$

Combining all estimates together under (i)-(ii)

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

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

$$\Pr\left(\left|\sum_{i=1}^{N}\sum_{t=1}^{T} \xi_{i,t}\right| > u\right) \leq c_1 NT u^{-\kappa} + 4e^{-c_2 u^2/NT}$$

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 $\delta \in (0,1)$

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

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. It is worth mentioning that the concentration over N is also influence by the weak dependence, which probably can be relaxed with a sharper proof technique. However, for geometrically ergodic processes, e.g., for stationary $AR(p)$, we have $\kappa \approx q$, in which case the time series dependence does not influence the concentration at all.

Let $(\xi_t)_{t\in\mathbf{N}}$ be a real-valued stochastic process and let Q_t denote the generalized inverse of the tail function $x \mapsto \Pr(|\xi_t| \geq x)$. Let $\xi \in \mathbf{R}$ be a random variable corresponding to $(\xi_t)_{t\in\mathbf{Z}}$ such that $Q \geq \sup_{t\in\mathbf{N}} Q_t$, where Q is a generalized inverse of $x \mapsto \Pr(|\xi| \geq x)$. The following Rosenthal's moment inequality for τ-dependent sequences follows from Dedecker and Prieur (2004); see also Dedecker and Doukhan (2003).

Theorem A.2. *Let $(\xi_t)_{t\in\mathbf{N}}$ be a centered stochastic process such that (i) there exists $q > 2$ such that $\|\xi\|_q < \infty$, where $\xi \in \mathbf{R}$ corresponds to $(\xi_t)_{t\in\mathbf{N}}$; (ii) the τ-mixing coefficients are $\tau_{k-1} \leq ck^{-a}, \forall k \geq 1$ for some universal constants $c > 0$ and $a > (q(r-2)+1)/(q-r)$. Then for every $r \in [2, q)$*

$$\mathbb{E}\left|\sum_{t=1}^{T} \xi_t\right|^r \leq c_{q,r}\left(T^{r/2}\|\xi\|_q^{qr/2(q-1)} + T\|\xi\|_q^{q(r-1)/(q-1)}\right),$$

where the constant $c_{q,r}$ depend only on q and r.

Proof. Let G be the inverse of $x \mapsto \int_0^x Q(u)\mathrm{d}u$ and put $H(u) = \sum_{k=0}^{\infty} \mathbb{1}_{2u<\tau_k}$, where $(\tau_k)_{k\in\mathbf{N}}$ are τ-mixing coefficients of $(\xi_t)_{t\in\mathbf{N}}$. Note that for every $q \geq 1$,

$$\int_0^{\|\xi\|_1} |Q \circ G(u)|^{q-1}\mathrm{d}u = \int_0^1 Q^q(v)\mathrm{d}v = \|\xi\|_q^q.$$

Then by Hölder's inequality

$$\int_0^{\|\xi\|_1} |H(u)Q \circ G(u)|^{r-1}\mathrm{d}u \leq \left(\int_0^{\|\xi\|_1} H^{(q-1)(r-1)/(q-r)}(u)\mathrm{d}u\right)^{\frac{q-1}{q-r}} \|\xi\|_q^{q(r-1)/(q-1)}$$

Note also that for some constant $C_{q,r}$ that depends only on q and r we have

$$\begin{aligned}\int_0^{\|\xi\|_1} H^{(q-1)(r-1)/(q-r)}(u)\mathrm{d}u &\leq (1 \vee s_{q,r}) \int_0^{\|\xi\|_1} \sum_{k=0}^{\infty} (k+1)^{(q-1)(r-1)/(q-r)-1} \mathbb{1}_{2u<\tau_k} \mathrm{d}u \\ &\leq 0.5(1 \vee s_{q,r}) \sum_{k=0}^{\infty} (k+1)^{(q-1)(r-1)/(q-r)-1} \tau_k \\ &\leq 0.5c(1 \vee s_{q,r}) \sum_{k=1}^{\infty} k^{(q-1)(r-1)/(q-r)-1-a} \\ &\leq C_{q,r}\end{aligned}$$

where we use the fact that $H^s(u) = \sum_{k=0}^{\infty}((k+1)^s - k^s)\mathbb{1}_{2u<\tau_k}$, $(k+1)^s - k^s \leq (1 \vee s)(k+1)^{s-1}$ with $s = s_{q,r} = (q-1)(r-1)/(q-r)$, and the series converges since $a > (q(r-2)+1)/(q-r)$. Combining these estimates

$$\int_0^{\|\xi\|_1} |H(u)Q \circ G(u)|^{r-1}\mathrm{d}u \leq C_{q,r}^{\frac{q-1}{q-r}} \|\xi\|_q^{q(r-1)/(q-1)}. \tag{A.10}$$

By Dedecker and Prieur (2004), Corollary 1, for some constant $c_r > 0$ that depends only on r

$$\begin{aligned}\mathbb{E}\left|\sum_{t=1}^{T} \xi_t\right|^r &\leq c_r \left\{ \left(T \int_0^{\|\xi\|_1} H(u)Q \circ G(u)\mathrm{d}u \right)^{r/2} + T \int_0^{\|\xi\|_1} |H(u)Q \circ G(u)|^{r-1}\mathrm{d}u \right\} \\ &\leq c_r \left\{ T^{r/2} \left(C_{q,r}^{\frac{q-1}{q-2}} \|\xi\|_q^{q/(q-1)} \right)^{r/2} + T C_{q,r}^{\frac{q-1}{q-r}} \|\xi\|_q^{q(r-1)/(q-1)} \right\} \\ &\leq c_{q,r} \left(T^{r/2} \|\xi\|_q^{qr/2(q-1)} + T\|\xi\|_q^{q(r-1)/(q-1)} \right),\end{aligned}$$

where the second line follows by equation (A.10) and $c_{q,r} > 0$ depends only on q and r. □

C Large N and T central limit theorem

For a double sequence $\{a_{N,T} : N, T \in \mathbf{N}\}$, we use $\lim_{N,T\to\infty} a_{N,T}$ to denote the limit when $N, T \to \infty$ jointly and $\max_{N,T\in\mathbf{N}} a_{N,T} = \max\{a_{N,T} : N \in \mathbf{N}, T \in \mathbf{N}\}$. The following central limit theorem holds for panel data consisting of τ-mixing processes that may change over N and T.

Theorem A.1. *Let $\{\xi_{N,T,i,t} : i \in \mathbf{N}, t \in \mathbf{Z}\}$ be an array of centered random vectors in $\mathbf{R}^p$ such that for each N, T, and i, $\{\xi_{N,T,i,t} : t \in \mathbf{Z}\}$ is a stationary process in $\mathbf{R}^p$ and $\{(\xi_{N,T,i,1}, \ldots, \xi_{N,T,i,T}) : i \in \mathbf{N}\}$ are independent arrays in $\mathbf{R}^p \times \mathbf{R}^T$ satisfying (i) for some*

$q > 2$, $\max_{i\in[N],j\in[p]} \|\xi_{N,T,i,t,j}\|_q = O(1)$; *(ii) for all* N, T, i, j, *the* τ*-mixing coefficients of* $\{\xi_{N,T,i,t,j} : t \in \mathbf{Z}\}$ *satisfy* $\tau_{k-1} \leq ck^{-a}, \forall k \geq 1$ *for some universal constants* $c > 0$ *and* $a > (q-1)/(q-2) \vee (q\delta+1)/(q-2-\delta)$ *with* $q > 2+\delta$ *and* $\delta > 0$; *(iii) for every* $i, N \in \mathbf{N}$, $\lim_{T\to\infty} \mathrm{Var}(\xi_{N,T,i,t}) < \infty$. *Then*

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \xi_{N,T,i,t} \xrightarrow{\mathrm{d}} N(0, \Xi) \qquad as \qquad N, T \to \infty,$$

where $\Xi = \lim_{N,T\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathrm{Var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \xi_{N,T,i,t}\right)$ *is a finite matrix, assumed to be a positive definite.*

Proof. By the Cramér-Wold device, see Billingsley (1995), Theorem 29.4,

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \xi_{N,T,i,t} \xrightarrow{\mathrm{d}} N(0, \Xi) \qquad \text{as} \qquad N, T \to \infty$$

in $\mathbf{R}^p$ if and only if for every $z \in \mathbf{R}^p$, the following weak convergence holds in $\mathbf{R}$

$$z^\top \left(\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \xi_{N,T,i,t} \right) \xrightarrow{\mathrm{d}} N(0, z^\top \Xi z) \qquad \text{as} \qquad N, T \to \infty.$$

Note that under maintained assumptions, for each N, T and $z \in \mathbf{R}^p$,

$$z^\top \left(\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \xi_{N,T,i,t} \right) = \sum_{i=1}^{N} z^\top \left(\frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \xi_{N,T,i,t} \right)$$

is a sum of N independent zero-mean random variables. By independence and stationarity, the variance of this sum is

$$\begin{aligned} \sigma^2_{N,T,z} &\triangleq \frac{1}{N} \sum_{i=1}^{N} \mathrm{Var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} z^\top \xi_{N,T,i,t} \right) \\ &= \frac{1}{N} \sum_{i=1}^{N} \left\{ \mathrm{Var}(z^\top \xi_{N,T,i,t}) + 2 \sum_{k=1}^{T-1} \left(1 - \frac{k}{T}\right) \mathrm{Cov}(z^\top \xi_{N,T,i,0}, z^\top \xi_{N,T,i,k}) \right\}. \end{aligned}$$

If we show that the limit in the parentheses exists for every $i, N \in \mathbf{N}$, then the joint limit of $\sigma^2_{N,T,z}$ as $N, T \to \infty$ is the same as the sequential limit

$$\lim_{N\to\infty} \lim_{T\to\infty} \frac{1}{N} \sum_{i=1}^{N} \left\{ \mathrm{Var}(z^\top \xi_{N,T,i,t}) + 2 \sum_{k=1}^{T-1} \left(1 - \frac{k}{T}\right) \mathrm{Cov}(z^\top \xi_{N,T,i,0}, z^\top \xi_{N,T,i,k}) \right\};$$

see Apostol (1974), Theorem 8.39. By Babii, Ghysels, and Striaukas (2021a), Lemma A.1.1, for every $k \geq 1$

$$|\mathrm{Cov}(z^\top \xi_{N,T,i,0}, z^\top \xi_{N,T,i,k})| \leq \tau_k^{\frac{q-2}{q-1}} \|z^\top \xi_{N,T,i,0}\|_q^{q/(q-1)} = O(k^{-a}),$$

where the second inequality follows under (i)-(ii). Moreover, $\sum_{k=1}^\infty k^{-a} < \infty$ under (ii). Therefore, by Lebesgue's dominated convergence theorem, for every $i, N \in \mathbf{N}$,

$$\lim_{T\to\infty} \sum_{k=1}^{T-1} \left(1 - \frac{k}{T}\right) \mathrm{Cov}(z^\top \xi_{N,T,i,0}, z^\top \xi_{N,T,i,k}) < \infty,$$

and whence under (ii)

$$\lim_{N,T\to\infty} \sigma_{N,T}^2 = \lim_{N,T\to\infty} \frac{1}{N} \sum_{i=1}^N \mathrm{Var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^T z^\top \xi_{N,T,i,t}\right) = z^\top \Xi z < \infty.$$

The statement of the theorem follows by the central limit theorem for independent random variables, provided that the following Lyapunov condition holds

$$\lim_{N,T\to\infty} \frac{1}{(NT)^{1+\delta/2}} \sum_{i=1}^N \mathbb{E}\left|\sum_{t=1}^T z^\top \xi_{N,T,i,t}\right|^{2+\delta} = 0;$$

see Billingsley (1995), Theorem 27.3 and Phillips and Moon (1999), Theorem 2.

By Theorem A.2, for some $c_{q,\delta}$ that depends only on q and δ,

$$\mathbb{E}\left|\sum_{t=1}^T z^\top \xi_{N,T,i,t}\right|^{2+\delta} \leq c_{q,\delta} \left\{T^{1+\delta/2} \|z^\top \xi_{N,T,i,t}\|_q^{q(1+\delta/2)/(q-1)} + T \|z^\top \xi_{N,T,i,t}\|_q^{q(1+\delta)/(q-1)}\right\}.$$

Therefore, the Lyapunov condition holds under (i). □

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

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

July 8, 2021

*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.

OA.1 Additional empirical results

RW	MSE An.-mean	MSE An.-median		MIDAS ML					
2.831	2.762	2.614	$\gamma =$	0	0.2	0.4	0.6	0.8	1
						sg-LASSO-MIDAS			
						Panel A1. Cross-validation			
			Individual	1.816	1.813	1.805	1.866	1.810	1.812
			Pooled	1.756	**1.706**	1.724	1.741	1.784	1.937
			Fixed Effects	1.762	1.762	1.764	1.766	1.761	1.961
						Panel B1. BIC			
			Individual	1.940	1.906	1.957	1.982	1.950	1.935
			Pooled	1.808	1.796	1.794	1.798	1.811	1.889
			Fixed Effects	1.793	1.794	**1.789**	1.799	1.790	1.794
						Panel C1. AIC			
			Individual	1.971	1.953	1.971	1.937	1.981	1.934
			Pooled	1.785	1.785	1.793	1.794	1.794	1.792
			Fixed Effects	1.715	**1.706**	1.796	1.762	1.714	1.708
						Panel D1. AICc			
			Individual	2.047	2.154	2.278	2.452	2.659	2.862
			Pooled	1.785	1.785	1.793	1.794	1.794	1.792
			Fixed Effects	1.715	**1.706**	1.796	1.762	1.714	1.708
						sg-LASSO-MIDAS augmented with An.-median			
						Panel A2. Cross-validation			
			Individual	1.743	1.753	1.745	1.732	1.734	1.889
			Pooled	1.746	1.732	1.738	1.741	1.761	1.878
			Fixed Effects	1.723	**1.698**	1.702	1.725	1.764	1.867
						Panel B2. BIC			
			Individual	1.751	1.752	1.761	1.772	1.780	1.781
			Pooled	1.756	1.747	1.743	1.742	1.771	1.784
			Fixed Effects	1.749	**1.712**	1.721	1.735	1.761	1.835
						Panel C2. AIC			
			Individual	1.762	1.761	1.769	1.778	1.781	1.801
			Pooled	1.765	1.765	1.763	1.764	1.764	1.771
			Fixed Effects	1.755	**1.753**	1.760	1.757	1.757	1.789
						Panel D2. AICc			
			Individual	1.762	1.761	1.769	1.778	1.781	1.801
			Pooled	1.765	1.765	1.763	1.764	1.764	1.771
			Fixed Effects	1.755	**1.753**	1.760	1.757	1.757	1.789

Table OA.1: Prediction results – The table reports average over firms MSEs of out-of-sample predictions. The nowcasting horizon is the next quarter, i.e. we predict the next quarter P/E ratio. Block in Panel A1-D1 correspond to ML-only forecast errors while in Panel A2-D2 to ML models augmented with median consensus nowcasts. Each Panel A1-D1 and A2-D2 block represents different ways of calculating the tuning parameter λ. Bold entries are the best results in a block.

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 OA.2: 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 2.

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 OA.3: Prediction results – The table reports average over firms MSEs of out-of-sample predictions for the same models as in Table 2 - 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.

	Full sample	sg-LASSO & elnet	sg-LASSO	none
		sg-LASSO		
Pooled	4.827	7.004	5.370	5.001
Fixed Effects	4.261	7.777	5.648	5.256
		Elastic net		
Pooled	5.805	6.868	5.536	4.958
Fixed Effects	7.990	5.956	5.324	5.026
		Tail index for regressands		
	3.954	5.236	4.702	4.630
		Number of firms		
Pooled	210	63	12	135
Fixed Effects	210	66	8	134

Table OA.4: Heaviness of tails – The table reports the tail index of in-sample residuals. Tail index is computed using the Hill estimator. The results are reported for the models as in Table 3.

OA.2 Data description

OA.2.1 Firm-level data

The full list of firm-level data is provided in Table OA.5. 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).

OA.2.1.1 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.

OA.2.1.2 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 $x_i \in \{\text{ESS}, \text{AES}, \text{CSS}, \text{NIP}\}$ 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 OA.5: 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) $\Delta \log(x_t)$, (6) $\Delta^2 \log(x_t)$. 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 OA.6: 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) $\Delta \log(x_t)$, (6) $\Delta^2 \log(x_t)$. 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 OA.7: 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.

OA.3 Additional Technical Results

Proposition OA.3.1. *Suppose that Assumptions 3.1, 3.2, 3.3, 3.4, and 4.1 are satisfied for each nodewise regression $j \in G$. Then if $S^\kappa p(NT)^{1-\kappa} \to 0$ and $S^2 \log p/NT \to 0$*

$$\|\hat{\Theta}_G - \Theta_G\|_\infty = O_P\left(\frac{Sp^{1/\kappa}}{(NT)^{1-1/\kappa}} \vee S\sqrt{\frac{\log p}{NT}}\right)$$

and

$$\max_{j \in G} |(I - \hat{\Theta}\hat{\Sigma})_j|_\infty = O_P\left(\frac{p^{1/\kappa}}{(NT)^{1-1/\kappa}} \vee \sqrt{\frac{\log p}{NT}}\right).$$

Proof. The proof is similar to the proof of Babii, Ghysels, and Striaukas (2021), Propositions A.1.2 and A.1.3. □

References

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

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

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

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.