Journal of Econometrics 237 (2023) 105315

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Machine learning panel data regressions with heavy-tailed dependent data: Theory and application

Andrii Babii [a], Ryan T. Ball [b], Eric Ghysels [c,*], Jonas Striaukas [d]

[a] University of North Carolina at Chapel Hill - Gardner Hall, CB 3305 Chapel Hill, NC 27599-3305, United States of America
[b] Stephen M. Ross School of Business, University of Michigan, 701 Tappan Street, Ann Arbor, MI 48109, United States of America
[c] Department of Economics and Kenan-Flagler Business School, University of North Carolina–Chapel Hill, United States of America
[d] FRS-FNRS Research Fellow, LIDAM UC Louvain, Belgium

ARTICLE INFO

Article history:
Received 9 August 2020
Received in revised form 19 June 2022
Accepted 10 July 2022
Available online 26 July 2022

Keywords:
High-dimensional panels
Large N and T panels
Mixed-frequency data
Sparse-group LASSO
Fat tails

ABSTRACT

The paper introduces structured machine learning regressions for heavy-tailed dependent panel data potentially 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 improve the quality of the estimates. 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.

1. Introduction

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. We leverage on the structured sparsity approach using sparse-group LASSO (sg-LASSO) regularization for time series data with dictionaries. The advantages of this approach for individual time series data, potentially sampled at mixed frequencies, have been recently reported in Babii et al. (2022b), who focus on nowcasting the US GDP growth in a data-rich environment. In this paper, we first show how to leverage on the sparse group regularization in a 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 are 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 provide an illustrative empirical example involving systematically predictable errors in analysts with individual firm earnings forecasts.

Our paper relates to the literature on high-dimensional panel data models and the LASSO regularization; see Harding and Lamarche (2019), Chiang et al. (2019), Chernozhukov et al. (2019), Belloni et al. (2019, 2016), Lu and Su (2016), Kock (2016), Su et al. (2016), Farrell (2015), Kock (2013), Lamarche (2010), Koenker (2004), among others.

It is worth emphasizing what sets our paper apart from the existing literature. An applied researcher facing high-dimensional panel data has several modeling decisions to make. Should (s)he simply run single regularized regression models, or take advantage of the cross-sectional dimension on the panel? If the latter is the case, how should one proceed? Pool the regressions or consider a fixed effect panel model to capture cross-sectional heterogeneity?

[*] Corresponding author.
E-mail addresses: andrii@email.unc.edu (A. Babii), rtball@umich.edu (R.T. Ball), eghysels@unc.edu (E. Ghysels), jonas.striaukas@uclouvain.be (J. Striaukas).

https://doi.org/10.1016/j.jeconom.2022.07.001
0304-4076/© 2022 Elsevier B.V. All rights reserved.

Note that there are a lot of subtleties involved. Take for example the decision to run individual regularized time series regressions versus pooled or fixed effect panel regressions. In a high-dimensional setting, it is not simply a matter of replacing T by NT to account for the panel data environment. In the fixed effect panel model we pay price for estimating N fixed effects which plays a role similar to the effective dimension of covariates. In a pooled panel model, we do not incur this penalty, but is pooling warranted? What about simply ignoring the panel structure? If we add to the model choice also the fact that the data are not sampled at the same frequency — a situation often encountered in many high-dimensional prediction problems — then we also need to make a choice on how to approach parameter proliferation and regularization in terms of the MIDAS parameterization (e.g. UMIDAS as in Foroni et al. (2015) or constrained polynomials as in Babii et al. (2022b)). Moreover, 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) the debiased HAC inference for regularized panel data, and (4) the sg-LASSO regularization of Simon et al. (2013) in a panel data setting.

In this paper we provide the theoretical underpinnings for all the aforementioned issues an applied researcher faces, report on Monte Carlo simulations to gauge the finite sample behavior of the various estimation methods and finally include an empirical application which provides new insights on potential inefficiencies and biases in earnings forecasting by analysts.

Our theory recognizes that the economic and financial time series data are often persistent with fat tails. To that end, we develop new Fuk–Nagaev and Rosenthal's inequalities and the central limit theorem for long τ-mixing panel.[1] Using new results, we obtain the 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). Additionally, we present the powerful debiased inferential methods for the regularized long panel data. The supporting theoretical results imply that the standard errors and the length of confidence intervals scale at the $O(1/\sqrt{NT})$ rate instead of the $O(1/\sqrt{N})$ rate that we typically encounter for fixed T approximations or the $O(1/\sqrt{T})$ rate for the individual time series regressions considered in Babii et al. (2022b).

In this way we also contribute to the modern panel data literature, 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. In contrast to this literature, we focus on the high-dimensional panels and our theory covers the regularized regression including the LASSO, the group-LASSO, and the sparse-group LASSO estimators.

The theory is extensively supported by the Monte Carlo experiments, where we focus on the debiased inference in the high-dimensional panel data and find that the individual regressions feature worse performance compared to the pooled regressions, hence, showing the usefulness of pooling the data. The results of these experiments also indicate that the structured regularization leads to better Granger causality tests in small samples.

In our empirical application we revisit a topic raised by Ball and Ghysels (2018) and Carabias (2018), but not resolved via formal inference in a high-dimensional setting. Namely, their empirical findings suggest that analysts tend to focus on their firm/industry when making earnings predictions while not fully taking into account the macroeconomic events affecting their firm/industry. More broadly, Ball and Ghysels (2018) argue that analysts do not fully exploit information embedded in high-dimensional data and therefore *leave money on the table*. Thanks to the theoretical contributions in the current paper we can formally test that hypothesis in a data-rich environment. Note that, as Ball and Ghysels (2018) point out, it is important to take into account the mixed frequency nature of the data flow, which is why the machine learning panel regression methods presented in the paper apply to mixed frequency data. We use 26 predictors, including traditional macro and financial series as well as non-standard series generated by textual analysis of financial news. Using such a rich set of covariates, we test whether analyst's consensus earnings prediction errors are systematically related to either one of the aforementioned variables.

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. Monte Carlo simulations are reported in Section 5. The results of our empirical application 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 $[p] = \{1, 2, \dots, p\}$. 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 $\mathrm{vech}(A) \in \mathbf{R}^{p(p+1)/2}$ be its vectorization consisting

[1] It is worth mentioning that the direct application of the theory developed in Babii et al. (2022a,b) for single time series regression models leads to inferior results for heavy-tailed panel data.

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 $a_n \leq Cb_n$ 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. We can also absorb the (low-frequency) lags of $y_{i,t}$ in the set of covariates. When $m = 1$, we retain the standard (dynamic if we include lagged dependent variables) 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.

2.1. Dealing with mixed frequency panel data

To reduce the proliferation of lag parameters, we follow the MIDAS literature; see Ghysels et al. (2006a,b), and Babii et al. (2022a,b). Instead of estimating m individual slopes of high-frequency covariate $k \in [K]$ in Eq. (1), with some abuse of notation, we estimate a weight function ω parametrized by $\beta_k \in \mathbf{R}^L$ with $L < m$

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

where

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

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

[2] 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.

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 optimization. To that end, define $\mathbf{x}_i = (X_{i,1}W, \dots, X_{i,K}W)$, where for each $k \in [K]$, $X_{i,k} = (x_{i,t-(j-1)/m,k})_{t\in[T],j\in[m]}$ is 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 $\mathbf{y}_i = (y_{i,1+h}, \dots, y_{i,T+h})^\top$ and $\mathbf{u}_i = (u_{i,1}, \dots, u_{i,T})^\top$. 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 $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_N^\top)^\top$, $\mathbf{X} = (\mathbf{x}_1^\top, \dots, \mathbf{x}_N^\top)^\top$, and $\mathbf{u} = (\mathbf{u}_1^\top, \dots, \mathbf{u}_N^\top)^\top$. Then the regression equation after stacking all cross-sectional observations is

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

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

The MIDAS approach allows us to effectively reduce the dimensionality pertaining to the high-frequency lags. Alternatively, we may apply what is known as the UMIDAS scheme, see e.g., Foroni et al. (2015), and directly estimate the coefficients associated with each high-frequency covariate lags separately (see Eq. (8) in Section 5 for example). Such a strategy, which as Foroni et al. (2015) argue works in single regressions when the ratio high to low-frequency sampling is small, may not be appealing in high-dimensional cases, as the estimation and prediction performance deteriorates due to the potentially large number of coefficients; see Babii et al. (2022b) for further discussion.

While the theory we present is general, we have the specific empirical application of our paper in mind with a cross-section of firms for which we want to analyze corporate earnings predictions. Some firms belong to the same industry, or are of similar size, yet all are affected by the state of the macroeconomy as well. This calls for a panel regression setting where we exploit both the cross-sectional dimension and the time series dynamics.

Assuming that the individual lag coefficients in Eq. (1) are approximately sparse is *highly* restrictive. Instead, the approximate sparsity of slopes of the dictionary elements $(w_l)_{l\geq 0}$ is more plausible for many applications of interest to economists and will be maintained here as well.[3] 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.[4]

2.2. *Fixed effect high-dimensional panel models*

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 shown to be attractive for individual time series ML regressions in Babii et al. (2022b). 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, i.e. $|b|_1$, and group LASSO penalties with for a group structure $\mathcal{G}$ described as a partition of $[p] = \{1, 2, \dots, p\}$, the group LASSO norm is computed as $\|b\|_{2,1} = \sum_{G\in\mathcal{G}} |b_G|_2$. The parameter $\gamma \in [0, 1]$ determines the relative weights of the ℓ_1 (sparsity) and the $\ell_{2,1}$ (group sparsity) norms, while the amount of regularization is controlled by the regularization parameter $\lambda \geq 0$. The group structure is assumed to be known to the econometrician, which in our setting corresponds to time series lags of covariates. For example, consider the earlier defined $T \times LK$ matrix of covariates $\mathbf{x}_i = (X_{i,1}W, \dots, X_{i,K}W)$, where each $X_{i,1}W$ is a $T \times L$ matrix corresponding to the aggregated time series lags of predictor $k \in [K]$. Ignoring the fixed effects, the time series lags define a group structure with K groups of size L described as $\mathcal{G} = \{\{1, \dots, L\}, \{L+1, \dots, 2L\}, \dots, \{p-L+1, \dots, p\}\}$.

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.[5] 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 Eq. (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 weight leads to a nonlinear non-convex optimization problem that becomes challenging to solve in high-dimensions; cf. Marsilli (2014) and Khalaf et al. (2021).

[3] For a formal definition of approximate sparsity, see Babii et al. (2022b).

[4] 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.

[5] 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.

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 via a MIDAS weighting scheme with Legendre polynomials. The covariates may also include the time-varying covariates common for all entities (macroeconomic factors), lags of $y_{i,t}$, the intercept, as well as additional lags of a baseline covariate.

3.1. Dealing with temporal dependence: τ-mixing

Many empirical applications, such as the prediction of corporate earnings considered later in the paper, involve persistent time series data both as dependent variables and covariates. To that end, we follow Babii et al. (2022b) and 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 $\mathbf{R}$.[6] For a stochastic process $(\xi_t)_{t \in \mathbf{Z}}$ with a natural filtration generated by its past $\mathcal{M}_t = \sigma(\xi_t, \xi_{t-1}, \ldots)$, the τ-mixing coefficients are defined as

$$\tau_k = \sup_{j \geq 1} \frac{1}{j} \sup_{t+k \leq t_1 < \cdots < t_j} \tau(\mathcal{M}_t, (\xi_{t_1}, \ldots, \xi_{t_j})), \qquad k \geq 0 \tag{3}$$

where the supremum is taken over all $t, t_1, \ldots, t_j \in \mathbf{Z}$. 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, 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.[7] This allows us to obtain concentration inequalities and performance guarantees for the sg-LASSO estimator; see Appendix B.

We would like to stress that the (non)asymptotic theory for the long panel data involves the concentration of double arrays

$$\sum_{i=1}^{N}\sum_{t=1}^{T} \xi_{i,t} = \sum_{i=1}^{N} \underbrace{\left(\sum_{t=1}^{T} \xi_{i,t}\right)}_{\equiv \xi_{i,T}} = \sum_{t=1}^{T} \underbrace{\left(\sum_{i=1}^{N} \xi_{i,t}\right)}_{\equiv \xi_{t,N}}$$

Instead of developing new results for such arrays, one could of course rely on the existing concentration inequalities for the i.i.d. or time-series data. However, this strategy does not lead to the optimal results for the heavy-tailed dependent data as it only reflects the concentration over one of the two dimensions — the i.i.d. inequality of Fuk and Nagaev (1971) applied to $\sum_{i=1}^{N} \xi_{i,T}$ does capture the benefits of the time dimension while the inequality of Babii et al. (2022a) applied to $\sum_{t=1}^{T} \xi_{t,N}$ does not capture the benefits of the cross-section. Our paper provides new Fuk–Nagaev and Rosenthal's inequalities for such double arrays consisting of τ-mixing processes as well as the central limit theorem when both the cross-section and the time dimension increase at the same time. Building on these results, we develop the inferential theory presented in subsequent sections.

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

$$\min_{r=(a,b) \in \mathbf{R}^{1+p}} \|\mathbf{y} - a\iota - \mathbf{X}b\|_{NT}^2 + 2\lambda\Omega(r). \tag{4}$$

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

[7] The class of α-mixing processes is too restrictive for the predictive linear projection model with covariates and autoregressive lags; see also Babii et al. (2022b), 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) $u_i = (u_{i,1}, \dots, u_{i,T})$ and (c) $x_i = (x_{i,1}^\top, \dots, x_{i,T}^\top)^\top$ for $i \in [N]$. Let $\tau_k^{(i,j)}$ denote the τ-mixing coefficient of a stochastic process $(u_{i,t}z_{i,t,j})_{t\in\mathbf{Z}}$ with its natural filtration, where $i \in [N]$ and $j \in [p+1]$; see Eq. (3). Similarly, let $\tilde{\tau}_k^{(i,j)}$ denote the τ-mixing coefficient of jth coordinate of $(\mathrm{vech}(z_{i,t}z_{i,t}^\top))_{t\in\mathbf{Z}}$, where $i \in [N]$ and $j \in [(p+1)(p+2)/2]$. 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 Toeplitz 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 et al. (2022b). We can also 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 et al. (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 the Appendix; see Theorem B.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 Assumption 3.1.

Our first result is the oracle inequality for the pooled sg-LASSO estimator described in Eq. (4). 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.

The proof of Theorem 3.1 relies on the developments in Babii et al. (2022b) with some improvements. Notably, it relies on the new Fuk–Nagaev concentration inequality for long panel data, see Theorem B.1.

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 $O_P(\log p/NT)$ 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 the $p = o(T^{q-1})$ condition needed for time series regressions in Babii et al. (2022b). Moreover, since $q > 2$, $p = o((NT)^{q-1})$ can be significantly weaker than the $p = o(NT)$ 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 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 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

$$\begin{aligned} &\hat{\alpha} = (B^\top B)^{-1} B^\top (\mathbf{y} - \mathbf{X}\hat{\beta}) \\ &0 = \mathbf{X}^\top M_B(\mathbf{X}\hat{\beta} - \mathbf{y})/NT + \lambda z^* \end{aligned} \tag{5}$$

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

$$\hat{\Sigma}_{N,T} = \begin{pmatrix} \frac{1}{T}B^\top B & \frac{1}{\sqrt{NT}}B^\top \mathbf{X} \\ \frac{1}{\sqrt{NT}}\mathbf{X}^\top B & \frac{1}{NT}\mathbf{X}^\top \mathbf{X} \end{pmatrix} \quad \text{and} \quad \Sigma_{N,T} = \begin{pmatrix} I_N & \frac{1}{\sqrt{NT}}\mathbb{E}\left[B^\top \mathbf{X}\right] \\ \frac{1}{\sqrt{NT}}\mathbb{E}\left[\mathbf{X}^\top B\right] & \mathbb{E}[x_{i,t}x_{i,t}^\top] \end{pmatrix}. \tag{6}$$

In line with Assumption 3.2 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.

Next, we make an assumption about the order of the regularization parameter which is governed by the Fuk–Nagaev inequality for long panels; see Appendix, Theorem B.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 Assumption 3.1.

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

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

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

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

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

Then the following corollary is an immediate consequence of Theorem 3.2.

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

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

Corollary 3.2 allows for $s, p, N, T \to \infty$ at appropriate rates. However, we pay an additional price for estimating N fixed effects which plays a role similar to the effective dimension of covariates. An immediate practical implication is that to achieve accurate predictions with high-dimensional fixed effect regressions, the panel has to be sufficiently long 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 many empirical applications we are not only interested in building machine learning empirical panel data prediction models. We often also want to test hypotheses of interest as is the case in our empirical application where we want to identify the covariates which are significant in explaining prediction errors in a panel of analyst earnings predictions.

In this section, we therefore 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 $\hat{\mu}_j = (\hat{\mu}_{j,1}, \dots, \hat{\mu}_{j,p})^\top$ 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 $\hat{B} = \operatorname{diag}(\hat{\sigma}_1^2, \dots, \hat{\sigma}_{p+1}^2)$, and

$$\hat{C} = \begin{pmatrix} 1 & -\hat{\mu}_{1,1} & \dots & -\hat{\mu}_{1,p} \\ -\hat{\mu}_{2,1} & 1 & \dots & -\hat{\mu}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\mu}_{p,1} & \dots & -\hat{\mu}_{p,p} & 1 \end{pmatrix}.$$

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 jth nodewise LASSO regression. Let s_j be the number of non-zero elements in jth 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 $\check{\tau}_t \leq c t^{-d}$ 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) are similar to the condition used in Babii et al. (2022a). 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 $\Xi_G = \lim_{T \to \infty} \operatorname{Var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^T u_{i,t} \Theta_G z_{i,t}\right)$.

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 or $O(1/\sqrt{T})$ for individual time series regressions considered in Babii et al. (2022b). Note that this result is used in our empirical application to perform the Granger causality test based on the following Wald statistic

$$W_{NT} = NT(\hat{\rho}_G + B_G - \rho_G)\hat{\Xi}_G^{-1}(\hat{\rho}_G + B_G - \rho_G) \xrightarrow{d} \chi^2_{|G|}.$$

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 $K(0) = 1$ puts less weight on more distant noisy covariances, while $M_T \uparrow \infty$ is a bandwidth (or lag truncation) parameter; see Babii et al. (2022a) for more details as well as formal results on the validity of HAC-based inference using sg-LASSO residuals.

5. Monte Carlo simulations

In this section, we assess the finite sample performance of the Granger causality tests for high-dimensional pooled panel data MIDAS regressions. A first subsection describes the design, followed by a second reporting the findings.

5.1. Design

We simulate the data from the following DGP:

$$y_{i,t} = \alpha + \rho y_{i,t-1} + \sum_{k=1}^{K} \frac{1}{m} \sum_{j=1}^{m} \omega((j-1)/m; \beta_k) x_{i,t-(j-1)/m,k} + u_{i,t}, \tag{7}$$

where $i \in [N]$, $t \in [T]$, α is the common intercept, $\frac{1}{m}\sum_{j=1}^{m} \omega((j-1)/m; \beta_k)$ is the weight function for kth high-frequency covariate and the error term is $u_{i,t} \sim_{i.i.d.} N(0, 4)$. The DGP corresponds to the target variable of interest $y_{i,t}$ driven by one autoregressive lag augmented with high-frequency series. The DGP is therefore a pooled MIDAS panel data model.

We set $\rho = 0.15$ and take the first high-frequency regressor, $k = 1$, as relevant, i.e. the first regressor Granger causes the response variable. We are interested in quarterly/monthly data, and use four quarters of data for the high-frequency regressors so that $m = 12$. The high-frequency regressors are generated as K i.i.d. realizations of univariate autoregressive (AR) processes $x_h = \rho x_{h-1} + \varepsilon_h$, where $\rho = 0.7$ and $\varepsilon_h \sim_{i.i.d.} N(0, 1)$, where h denotes the high-frequency sampling. For the DGP we rely on a commonly used weighting scheme in the MIDAS literature, namely the weights $\omega(s; \beta_k)$ for the only relevant high-frequency regressor $k = 1$ determined by the beta density, Beta(3, 3); see Ghysels et al. (2007) or Ghysels and Qian (2019), for further details. The empirical estimation involves MIDAS regressions with Legendre polynomials of degree $L = 3$. Lastly, we draw the intercepts $\alpha \sim$ Uniform(−4, 4). Throughout the experiment, we fix the sample sizes to $T = 50$ and $N = 30$.

We compare the empirical size and power of the Granger causality test under different structures placed on the regression models.

First, we compare sg-LASSO-MIDAS with LASSO-UMIDAS pooled panel data models. The former exploits the group structure of covariates by applying the sg-LASSO penalty function and a flexible way to model lags for each covariate using the MIDAS weight functions parametrized by low-dimensional coefficients. The latter pertains to the unstructured LASSO estimator together with the UMIDAS scheme. Introduced by Foroni et al. (2015), UMIDAS consists of estimating a regression coefficient for each high-frequency lag separately, and therefore the weight function for each covariate is

$$\sum_{j=1}^{m} \omega((j-1)/m; \beta_k) x_{i,t-(j-1)/m,k} = \sum_{j=1}^{m} b_{j,k} x_{i,t-(j-1)/m,k} \tag{8}$$

where $b_{j,k}$ is a regression coefficient associated with each high-frequency lag. We estimate regression coefficients by applying the standard unstructured LASSO estimator; hence we call the model LASSO-UMIDAS.

Second, we compare the pooled panel with individual time series regressions, for sg-LASSO-MIDAS and LASSO-UMIDAS, where the former exploits the benefits of the panel structure and the latter does not. In this case, we take the first sample $i = 1$ to compute empirical size and power of the Granger test for the individual regression models. Babii et al. (2022a) propose tests of Granger causality in univariate regularized regressions and high-dimensional data.

5.2. Simulation results

In Table 1, we report the empirical rejection frequency (ERF) for the Granger causality test based on the HAC estimator with two different kernel functions, Parzen and Quadratic spectral, and two different estimation strategies, sg-LASSO-MIDAS and LASSO-UMIDAS. We test whether the first high-frequency covariate Granger causes the low-frequency series, which corresponds to the DGP potential causal pattern. We report results for a set of bandwidth parameters, denoted M_T = 10, 20 and 30. The reported results are based on 2000 Monte Carlo replications.

To assess the performance we scale the Beta density function by multiplying it with a constant $a \in \{0, 1/5, 1/4, 1/3\}$, i.e. the weight function for the relevant covariate is:

$$a\frac{1}{m}\sum_{j=1}^{m} \omega((j-1)/m; \beta_k)$$

For $a = 0$, the ERF shows the empirical size of the test for the nominal level of 5%, while $a \in \{1/5, 1/4, 1/3\}$ the ERF shows the empirical power of the Granger causality test. For the larger scaling constant a, the alternatives are separated further away from the null hypothesis and the Granger causality test is expected to perform better.

The results reported in Table 1 show that the Granger causality test based on the sg-LASSO-MIDAS has empirical size close to the nominal level of 5%. In contrast, the LASSO-UMIDAS leads to undersized Granger causality tests with size distortions around 0.01. The Granger causality test based on the sg-LASSO-MIDAS has also better empirical power against each of the alternative hypotheses $a \in \{1/5, 1/4, 1/3\}$. Additionally, it approaches 1 much faster as opposed to the LASSO-UMIDAS.

The results for individual regressions reveal worse performance compared to pooled panel data regressions, hence showing the usefulness of pooling the data. The empirical size shows considerable size distortions of around 0.05. Tests for individual regressions have worse power compared to the pooled panel data cases. Nonetheless, similar to the pooled

Table 1
HAC-based inference simulation results — We report results for a set of bandwidth parameters, denoted M_T, and two kernel functions.

	Pooled panel							
	Parzen kernel				*Quadratic spectral kernel*			
$M_T\backslash a$	0	1/5	1/4	1/3	0	1/5	1/4	1/3
	sg-LASSO-MIDAS							
10	0.051	0.835	0.959	0.999	0.056	0.841	0.963	0.998
20	0.049	0.822	0.954	0.999	0.047	0.828	0.957	0.998
30	0.046	0.803	0.953	0.999	0.047	0.823	0.956	0.998
	LASSO-UMIDAS							
10	0.039	0.551	0.788	0.978	0.042	0.549	0.797	0.979
20	0.030	0.514	0.762	0.970	0.033	0.535	0.780	0.977
30	0.021	0.494	0.735	0.964	0.025	0.514	0.758	0.972
	Individual regressions							
	Parzen kernel				*Quadratic spectral kernel*			
$M_T\backslash a$	0	1/5	1/4	1/3	0	1/5	1/4	1/3
	sg-LASSO-MIDAS							
10	0.090	0.356	0.406	0.548	0.094	0.349	0.356	0.486
20	0.097	0.345	0.406	0.548	0.094	0.350	0.360	0.492
30	0.092	0.345	0.403	0.547	0.093	0.356	0.379	0.524
	LASSO UMIDAS							
10	0.110	0.201	0.228	0.362	0.107	0.210	0.236	0.378
20	0.111	0.240	0.272	0.406	0.108	0.212	0.206	0.388
30	0.107	0.245	0.370	0.494	0.105	0.204	0.206	0.386

panel data cases, the sg-LASSO-MIDAS estimation method seems to have better empirical power when comparing to LASSO-UMIDAS.

Overall, the results of the Monte Carlo experiments indicate that the structured regularization leads to better Granger causality tests in small samples.

6. Do analysts leave money on the table?

In this section we revisit a topic raised by Ball and Ghysels (2018) and Carabias (2018). Their empirical findings suggest that analysts tend to focus on their firm/industry when making earnings predictions while not fully taking into account the impact of macroeconomic events. While their findings were suggestive, there was no formal testing in a data-rich environment. The theory established in the previous sections allows us to do so.

More specifically, we consider the earnings of 210 US firms using a set of predictors sampled at mixed frequencies — quarterly, monthly and daily series. We use 26 predictors (and their lags), including traditional macro and financial series as well as non-standard series generated by textual analysis of financial news.

6.1. Data description

The full sample consists of observations between the 1st of January, 2000 and the 30th of June, 2017. Due to the lagged dependent variables in the models, our effective sample starts at the third fiscal quarter of 2000. We collected data from CRSP and I/B/E/S to compute quarterly earnings and firm-specific financial covariates; RavenPack was used to compute daily firm-level textual-analysis-based data; real-time monthly macroeconomic series are from the ALFRED; FRED is used to compute daily financial markets data and, lastly, monthly news attention series extracted from the *Wall Street Journal* articles were retrieved from Bybee et al. (2019).[8] Table 2 provides a list of the variables used in our analysis, whereas Online Appendix Section OA.1 covers a detailed description of the RavenPack data. Finally, the list of all firms we consider in our analysis appears in Online Appendix Table OA.1. Table 2 has six panels, namely three panels of firm-level series: A1 — describes earnings data, B1 — describes daily firm-level stock market data, and C1 — describes daily firm-level sentiment data series. The remaining three panels are: A2 — describes real-time monthly macro series, B2 — describes daily financial markets data, and C2 — describes monthly news attention series. In the models we include 365 daily lags, 12 monthly lags and 4 quarterly lags respectively.

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

Table 2
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 transformations applied to a time series, which are (1) not transformed, (2) $100[(x_t/x_{t-1})^4 - 1]$, (3) $\Delta \log(x_t)$, (4) $\Delta^2 \log(x_t)$. The block of firm-level series contains three panels: A1 — describes earnings data, B1 — describes daily firm-level stock market data, and C1 — describes daily firm-level sentiment data series. The block labeled "other series" also has three panels: A2 — describes real-time monthly macro series, B2 — describes daily financial markets data, and C2 — describes monthly news attention series. In the models we include 365 daily lags, 12 monthly lags and 4 quarterly lags respectively.

	id	Frequency	Source	T-code
Firm-level series				
Panel A1.				
–	Earnings	Quarterly	CRSP & I/B/E/S	1
–	Earnings consensus forecasts	Quarterly	CRSP & I/B/E/S	1
–	Other earnings/earnings forecast implied series	Quarterly	CRSP & I/B/E/S	1
Panel B1.				
1	Stock returns	Daily	CRSP	1
2	Realized variance measure	Daily	CRSP/computations	1
Panel C1.				
3	Event Sentiment Score (ESS)	Daily	RavenPack	1
4	Aggregate Event Sentiment (AES)	Daily	RavenPack	1
5	Aggregate Event Volume (AEV)	Daily	RavenPack	1
6	Composite Sentiment Score (CSS)	Daily	RavenPack	1
7	News Impact Projections (NIP)	Daily	RavenPack	1
Other series				
Panel A2.				
8	Industrial Production Index	Monthly	ALFRED	3
9	CPI inflation	Monthly	ALFRED	4
10	Unemployment rate	Monthly	ALFRED	1
11	Real GDP	Quarterly	ALFRED	2
Panel B2.				
12	Crude Oil Prices	Daily	FRED	4
13	S&P 500	Daily	CRSP	3
14	VIX Volatility Index	Daily	FRED	1
15	Moodys Aaa less 10-Year Treasury	Daily	FRED	1
16	Moodys Baa less 10-Year Treasury	Daily	FRED	1
17	Moodys Baa less Aaa (corporate yield spread)	Daily	FRED	1
18	10-Year Treasury minus 3-Month Treasury (term spread)	Daily	FRED	1
19	3-Month Treasury minus Effective Federal funds rate (short-term spread)	Daily	FRED	1
20	TED rate	Daily	FRED	1
Panel C2.				
21	Earnings	Monthly	Bybee et al. (2019)	1
22	Earnings forecasts	Monthly	Bybee et al. (2019)	1
23	Earnings losses	Monthly	Bybee et al. (2019)	1
24	Recession	Monthly	Bybee et al. (2019)	1
25	Revenue growth	Monthly	Bybee et al. (2019)	1
26	Revised estimate	Monthly	Bybee et al. (2019)	1

6.2. Granger causality tests

Whether analysts leave money on the table amounts to testing whether forecast errors in earnings can be predicted by current information variables. Hence, this amounts to performing something akin to the Granger causality test. In our empirical application we are dealing with a panel, and it is important to exploit the multivariate data structure to perform such tests.

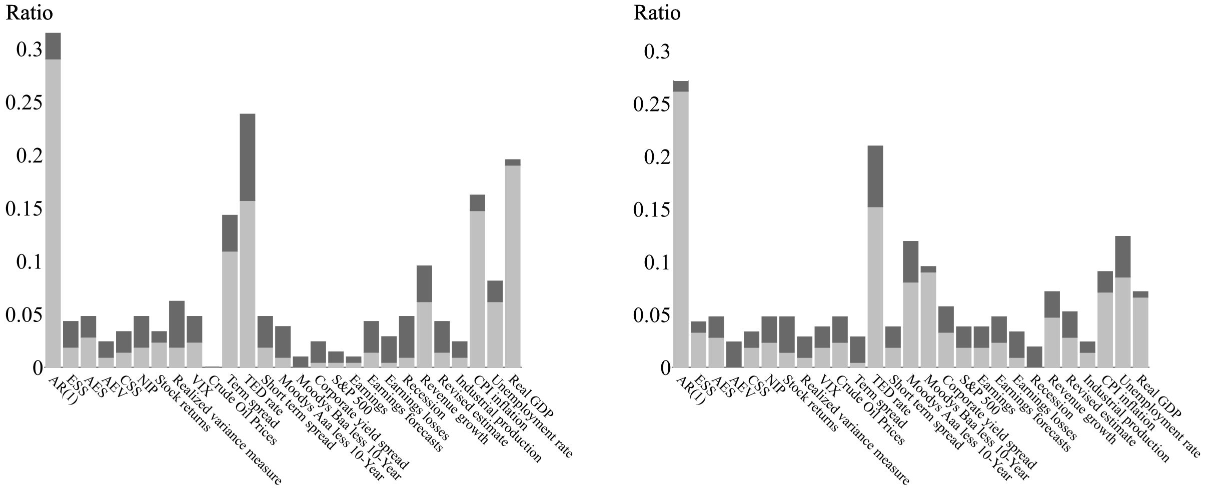
We analyze the difference between realized earnings and analysts' predictions, i.e., the response variable $y_{i,t+1}$ is computed by taking the difference between realized earnings, denoted $e_{i,t+1}$, and the median of analysts' predictions for the quarter $t+1$, denoted $f_{i,t+1|t}$,

$$y_{i,t+1} = e_{i,t+1} - f_{i,t+1|t}.$$

We then fit the following pooled panel data MIDAS model using sg-LASSO estimator:

$$y_{i,t+1} = \alpha + \rho y_{i,t} + \sum_{k=1}^{K} \psi(L^{1/m}; \beta_k) x_{i,t,k} + u_{i,t+1}. \tag{9}$$

We test which factors Granger cause future errors of earnings forecasts made by the analysts. In the sg-LASSO, groups are defined as all lags of a single covariate k; Legendre polynomials up to degree three are applied to all weight functions $\psi(L^{1/m}; \beta_k)$. We use 5-fold cross-validation to tune both λ and γ, where we define folds as adjacent blocks over the



(a) sg-LASSO-MIDAS

(b) LASSO-UMIDAS

Fig. 1. Individual regression-based Granger causality tests. In Panel (a) we plot the ratios based on sg-LASSO estimator and MIDAS weighting scheme with Legendre polynomials, while in Panel (b) we plot for the ratios for the standard LASSO estimator with UMIDAS weighting scheme. The lighter-gray color shows the ratio for firms with high disagreement, while the dark-gray color shows the ratio for firms with low disagreement; see Table 4. All results are based on the 5% significance level.

time series dimension to take into account the time series dependence. Similarly, we estimate the precision matrix using nodewise LASSO regressions selecting the tuning parameter in a similar vein. The results are reported in Table 3.

In Panel (A) of Table 3 we find that the AR(1) lag is significant, leading us to conclude that the prediction errors made by the analysts are persistent. The autoregressive coefficient is significant throughout all specifications of the models, including in a simple pooled AR(1) model. In the latter case, the AR(1) coefficient is estimated to be 0.147.

Panel (B) of Table 3 reports that beyond the AR(1) we find that the highly significant covariates are TED rate, CPI inflation and real GDP growth. These results support previous findings that analysts tend to miss information associated with macroeconomic conditions — including real GDP growth and the TED spread, which is an indicator of measure credit risk. The latter is rather surprising, as it indicates that analysts tend to miss out on credit risk information at the macro level in their earnings forecasts. Lastly, the term spread (10-year less 3-month treasury yield), often viewed as a business cycle indicator, is also significant at the 10% level.

Finally, in Panel (C) of Table 3 we report results based on the unstructured LASSO applying UMIDAS for the lag polynomials of each covariate. The findings reveal similar results for the TED rate, but notably miss real GDP and CPI inflation as significant covariates.

In Table 4 we show results based on a different way of pooling analysts' prediction errors $y_{i,t+1}$. We split the data into two parts based on how large the average disagreement among analysts is. For each firm, we compute the forecast disagreement as the difference between 95% and 5% percentile of the empirical forecast distribution and take the average over the sample. We sort from high to low disagreement and split the sample of firms into two subsamples of equal size. The results show that macro variables which are significant for the full sample are also significant for the large disagreement subsample. On the other hand, little significance is reported for the low disagreement subsample. In this case, only the AR(1) lag and stock returns are significant at the 5% significance level.

Lastly, in Fig. 1 we plot the ratio of firms for which we find Granger causality based on individual regressions versus panel models. In Panel (a) we plot the ratios for sg-LASSO estimator using MIDAS weighting scheme while in Panel (b) we plot the ratios for the LASSO estimator with UMIDAS scheme. The plot shows ratios for each covariate representing the fraction with respect to sg-LASSO (Panel (a)) or LASSO with UMIDAS (Panel (b)) each covariate is significant by running individual regressions. For example, the AR(1) lag is significant for around 30% (0.3) of firms when running individual sg-LASSO-MIDAS regressions. Some covariates that are not significant in pooled panels are significant for some firms; therefore, we show results for all covariates, including those that are not significant in pooled panel cases. We also show how the ratios differ for low (dark-gray color) versus high disagreement (light-gray color) firms. They represent whether a specific firm we run an individual regression for is in the high-disagreement versus low-disagreement subsample. Interestingly, the largest ratios are for AR(1), TED rate, Real GDP, CPI inflation and term spread in the case of sg-LASSO-MIDAS. Moreover, the portion of firms in the high disagreement subsample seem to have the largest ratios. In the case of LASSO-UMIDAS, the ratios show a less clear pattern, with only the AR(1) and TED rate covariates significant for a larger number of firms.

Table 3
Significance testing results — We report p-values for the AR(1) in Panel (A) and for the sg-LASSO using the MIDAS scheme with Legendre polynomials in Panel (B) displaying series significant at the 5% or 10% significance level. We also report results for the standard LASSO estimator together with the UMIDAS scheme in Panel (C). The results are reported for a range of bandwidth parameters (M_T = 10, 20 and 30) and two kernel functions (Quadratic Spectral and Parzen).

Variable \M_T	10	20	30	10	20	30
	Quadratic Spectral			Parzen		
	Panel (A) — AR(1)					
AR(1)	0.001	0.000	0.000	0.002	0.001	0.001
	Panel (B) — sg-LASSO					
	Significant variables at 5% or less					
AR(1)	0.001	0.000	0.000	0.002	0.001	0.000
TED rate	0.001	0.001	0.000	0.003	0.001	0.001
CPI inflation	0.003	0.001	0.001	0.013	0.003	0.001
Real GDP	0.028	0.003	0.001	0.035	0.021	0.006
	Significant variables at 10% level					
Term spread	0.012	0.014	0.023	0.053	0.016	0.015
	Panel (C) — LASSO (significant for sg-LASSO)					
	Significant variables at 5% or less					
AR(1)	0.001	0.000	0.000	0.002	0.001	0.000
TED rate	0.000	0.000	0.000	0.000	0.000	0.000
CPI inflation	0.677	0.390	0.461	0.651	0.724	0.576
Real GDP	0.341	0.247	0.094	0.339	0.328	0.270
	Significant variables at 10% level					
Term spread	0.273	0.060	0.022	0.235	0.387	0.365
	LASSO (significant only for LASSO)					
	Significant variables at 5% or less					
AAA less 10 year	0.009	0.001	0.001	0.015	0.014	0.007
BAA less 10 year	0.000	0.000	0.000	0.000	0.000	0.000

6.3. *Out of sample analysis*

In addition to the Granger causality tests, we apply the sg-LASSO-MIDAS model to predict the earnings forecast errors and test whether machine learning can improve upon analysts' predictions. As we aim at predicting the earnings forecast errors, we test whether machine learning methods can generate accurate out-of-sample forecasts to correct analysts' forecasts and thereby improve the overall performance.

We split the full sample into two halves, estimate the sg-LASSO-MIDAS model as in Eq. (9) based on the first half of the time-series observations and generate the forecasts. We then reestimate the model parameters based on the expanding window scheme and use the real-time data updates for macroeconomic variables. We use the 5-fold cross-validation to compute the tuning parameters. The results appear in Table 5.

We find that machine learning can indeed predict and correct the forecast errors made by analysts. When using the full set of covariates, the sg-LASSO-MIDAS model reduces the earnings forecast errors by 3%. We find further reduction in earnings forecast errors by 4% when only using the covariates which we find to be significant, see Table 3 Panel (B). The gain in prediction quality is significant at the 5% level based on the Diebold and Mariano (1995) test in both cases. Overall, the results appear to indicate that model-based predictions can significantly improve earnings forecasts when combined with the analysts predictions.

7. Conclusions

This paper introduced 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, which allows us to establish oracle inequalities that are used subsequently to develop the debiased HAC inference for the panel data sg-LASSO estimator.

Using the theory of HAC-based inference for pooled panel data regressions developed in our paper, our empirical analysis revisits a topic raised by earlier literature that analysts tend to focus on firm and/or industry information when forming earnings forecasts, while not fully taking into account the macroeconomic data. Our results suggest that indeed analysts tend to miss on macro information, i.e., macro variables turn out to be significant in pooled panel regression models.

Table 4
Significance testing results — We report p-values for the AR(1) and for the sg-LASSO-MIDAS models, displaying series significant at the 5% or 10% significance level. The results are reported for a range of bandwidth parameters and two kernel functions. We pool the response based on large versus small disagreement, which we measure as the average (over time series) of the difference between 95% and 5% percentile of the empirical forecast distribution of the analysts.

Variable $\backslash M_T$	10	20	30	10	20	30
	Quadratic Spectral			Parzen		
	Large disagreement					
	Significant variables at 5% or less					
AR(1)	0.002	0.001	0.000	0.004	0.001	0.001
Term spread	0.029	0.023	0.016	0.085	0.036	0.026
TED rate	0.002	0.001	0.001	0.016	0.002	0.001
CPI inflation	0.016	0.009	0.007	0.040	0.018	0.011
	Significant variables at 10% level					
Real GDP	0.098	0.005	0.000	0.098	0.082	0.021
	Small disagreement					
	Significant variables at 5% or less					
AR(1)	0.000	0.000	0.000	0.000	0.000	0.000
Stock returns	0.008	0.004	0.003	0.015	0.008	0.006
	Significant variables at 10% level					
Unemployment rate	0.060	0.043	0.045	0.060	0.056	0.048

Table 5
Out of sample prediction results — We report one-quarter ahead real-time out-of-sample forecasting results. In the first row we report results based on a full set of covariates, while in the second row the results are based on covariates which we find to be significant, see Table 3 Panel (B). We report the mean-squared-error ratio (column (MSE ratio) of sg-LASSO-MIDAS model corrected prediction relative to the analyst-only prediction of earnings. The last column reports the (Diebold and Mariano, 1995) test statistic.

	MSE ratio	DM statistic
All covariates	0.968	1.986
Significant covariates	0.960	2.143

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 et al. (2022b), 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 B.1 with probability at least $1-\delta$ and Assumption 3.3 for some $c > 1$. Therefore,

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

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

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

Suppose that $\|\mathbf{m} - \mathbf{Z}\rho\|_{NT} \leq \frac{1}{2}\|\mathbf{Z}\Delta\|_{NT}$. Then it follows from Eqs. (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 than 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 Eqs. (A.1) and (A.3), Assumption 3.2, $\hat{\Sigma} = \mathbf{Z}^\top\mathbf{Z}/NT$, and Babii et al. (2022b), 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 B.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 Eq. (A.4) implies $\Omega(\Delta) \lesssim s\lambda$, and whence from Eq. (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

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

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 Eq. (A.5) does not hold, then the inequality in Eq. (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

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

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

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

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

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

which shows that

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

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

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

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

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

where the second line follows by the dual norm inequality and the Cauchy–Schwarz inequality, and Ω^* is the dual norm of Ω. By Babii et al. (2022b), Lemma A.2.1. and Theorem B.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 et al. (2022a), Theorem 3.1 with probability at least $1-\delta/2$

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

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

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

In conjunction with the inequality in Eq. (A.6), this gives

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

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 Eq. (A.7) and Eq. (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 B.1 and Babii et al. (2022a), Theorem 3.1

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

Therefore, on the event E

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

and whence from Eq. (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 Eq. (4) 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 C.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 C.1 and shows that $I_{N,T} \xrightarrow{d} N(0,\Xi_G)$.

Next,

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

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

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

Lastly, by the Cauchy–Schwarz inequality

$$\begin{aligned}|IV_{N,T}|_\infty &\leq \max_{j\in G}|\mathbf{Z}\hat\Theta_j^\top|_2\|\mathbf{m}-\mathbf{Z}\rho\|_{NT} = \max_{j\in G}\sqrt{\hat\Theta_j^\top\hat\Sigma\hat\Theta_j}o_P(1) \\ &\leq \|\hat\Theta_G\|_\infty\sqrt{|\mathrm{vech}(\hat\Sigma)|_\infty}o_P(1) = o_P(1),\end{aligned}$$

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

Proposition A.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 et al. (2022a), Propositions A.1.2 and A.1.3. □

Appendix 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 et al. (2022a) and is of independent interest for the high-dimensional panel data.[9]

[9] The direct application of the time series Fuk–Nagaev inequality of Babii et al. (2022a) leads to inferior concentration results for panel data.

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

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

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

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

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

$$\begin{aligned} \Pr(I > u/3) &\leq c_1 u^{-q}\sum_{i=1}^N\sum_{t\geq 1}\mathbb{E}|V_{i,2t}^*|^q + 2\exp\left(-\frac{c_2u^2}{\sum_{i=1}^N\sum_{t\geq 1}\mathrm{Var}(V_{i,2t}^*)}\right) \\ &\leq c_1 u^{-q}\sum_{i=1}^N\sum_{t\geq 1}\mathbb{E}|V_{i,2t}|^q + 2\exp\left(-\frac{c_2u^2}{NT}\right), \end{aligned}$$

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 et al. (2022a), Lemma A.1.2 under assumptions (i) and (ii). Similarly,

$$\Pr(II > u/3) \leq c_1 u^{-q}\sum_{i=1}^N\sum_{t\geq 1}\mathbb{E}|V_{i,2t}|^q + 2\exp\left(-\frac{c_2u^2}{NT}\right).$$

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 2, this gives

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

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

$$
\begin{aligned}
\Pr\left(\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\xi_{i,t}\right| > u\right) &\leq \Pr(I > u/3) + \Pr(II > u/3) + \Pr(III > u/3)\\
&\leq c_1 u^{-q} N \sum_{i=1}^{N}\sum_{t\geq 1}\|V_{i,t}\|_q^q + 4e^{-c_2u^2/NT} + \frac{3NT}{u}\max_{i\in[N]}\tau_{J+1}^{(i,1)}\\
&\leq c_1 u^{-q} J^{q-1} NT + \frac{3NT}{u}(J+1)^{-a} + 4e^{-c_2u^2/NT}
\end{aligned}
$$

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

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

for some $c_1, c_2 > 0$. Finally, for $p > 1$, the result follows by the union bound. □

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

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

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 influenced 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 B.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 $c_{q,r}$ depends 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

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

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

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

where the second line follows by Eq. (B.1) and $c_{q,r}>0$ depends only on q and r. □

Appendix 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 C.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 $\tau_{k-1}\leq ck^{-a}$, $\forall k\geq1$ 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*

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

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

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

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

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

see Apostol (1974), Theorem 8.39. By Babii et al. (2022a), Lemma A.1.1, for every $k \geq 1$

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

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 $i, N \in \mathbf{N}$,

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

and whence under (ii)

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

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 B.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). □

Appendix D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2022.07.001.

References

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

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

Apostol, T.M., 1974. Mathematical Analysis. Pearson, p. 512.

Arellano, M., 2003. Panel Data Econometrics. Oxford University Press.

Babii, A., 2021. High-dimensional mixed-frequency IV regression. J. Bus. Econ. Statist. (forthcoming).

Babii, A., Florens, J.-P., 2020. Is Completeness Necessary? Estimation in Nonidentified Linear Models. Mimeo-UNC Chapel Hill.

Babii, A., Ghysels, E., Striaukas, J., 2022a. High-dimensional Granger causality tests with an application to VIX and news. J. Financ. Econom. (forthcoming).

Babii, A., Ghysels, E., Striaukas, J., 2022b. Machine learning time series regressions with an application to nowcasting. J. Bus. Econ. Statist. 40 (3), 1094–1106.

Ball, R.T., Easton, P., 2013. Dissecting earnings recognition timeliness. J. Account. Res. 51 (5), 1099–1132.

Ball, R.T., Gallo, L.A., 2018. A mixed data sampling approach to accounting research. available at SSRN 3250445.

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

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

Belloni, A., Chernozhukov, V., Hansen, C., Kozbur, D., 2016. Inference in high-dimensional panel models with an application to gun control. J. Bus. Econom. Statist. 34 (4), 590–605.

Billingsley, P., 1995. Probability and Measure. John Wiley & Sons.

Bybee, L., Kelly, B.T., Manela, A., Xiu, D., 2019. The structure of economic news. available at SSRN 3446225.

Carabias, J.M., 2018. The real-time information content of macroeconomic news: implications for firm-level earnings expectations. Rev. Account. Stud. 23 (1), 136–166.

Carrasco, M., Florens, J.-P., Renault, E., 2007. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In: Heckman, J.J., Leamer, E.E. (Eds.), Handbook of Econometrics - Vol. 6B. Elsevier, pp. 5633–5751.

Chernozhukov, V., Hausman, J.A., Newey, W.K., 2019. Demand Analysis with Many Prices. National Bureau of Economic Research Discussion paper 26424.

Chiang, H.D., Rodrigue, J., Sasaki, Y., 2019. Post-selection inference in three-dimensional panel data. arXiv preprint arXiv:1904.00211.

Dedecker, J., Doukhan, P., 2003. A new covariance inequality and applications. Stochastic Process. Appl. 106 (1), 63–80.

Dedecker, J., Prieur, C., 2004. Coupling for τ-dependent sequences and applications. J. Theoret. Probab. 17 (4), 861–885.

Dedecker, J., Prieur, C., 2005. New dependence coefficients. Examples and applications to statistics. Probab. Theory Related Fields 132 (2), 203–236.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econom. Statist. 13 (3), 253–263.

Farrell, M.H., 2015. Robust inference on average treatment effects with possibly more covariates than observations. J. Econometrics 189 (1), 1–23.

Fernández-Val, I., Weidner, M., 2016. Individual and time effects in nonlinear panel models with large N, T. J. Econometrics 192 (1), 291–312.

Foroni, C., Marcellino, M., Schumacher, C., 2015. Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. J. Roy. Statist. Soc. Ser. A 178 (1), 57–82.

Fuk, D.K., Nagaev, S.V., 1971. Probability inequalities for sums of independent random variables. Theory Probab. Appl. 16 (4), 643–660.

Ghysels, E., Qian, H., 2019. Estimating MIDAS regressions via OLS with polynomial parameter profiling. Econom. Statist. 9, 1–16.

Ghysels, E., Santa-Clara, P., Valkanov, R., 2006a. Predicting volatility: getting the most out of return data sampled at different frequencies. J. Econometrics 131 (1–2), 59–95.

Ghysels, E., Sinko, A., Valkanov, R., 2006b. MIDAS regressions: further results and new directions. Econometric Rev. 26 (1), 53–90.

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

Hahn, J., Kuersteiner, G., 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. J. Econometrics 141 (2), 597–620.

Harding, M., Lamarche, C., 2019. A panel quantile approach to attrition bias in big data: Evidence from a randomized experiment. J. Econometrics 211 (1), 61–82.

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

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

Kock, A.B., 2016. Oracle inequalities, variable selection and uniform inference in high-dimensional correlated random effects panel data models. J. Econometrics 195 (1), 71–85.

Koenker, R., 2004. Quantile regression for longitudinal data. J. Multivariate Anal. 91 (1), 74–89.

Lamarche, C., 2010. Robust penalized quantile regression estimation for panel data. J. Econometrics 157 (2), 396–408.

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

Marsilli, C., 2014. Variable Selection in Predictive MIDAS Models. Banque de France Working Paper.

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

Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2013. A sparse-group LASSO. J. Comput. Graph. Statist. 22 (2), 231–245.

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