Université Catholique de Louvain

Louvain School of Management
Doctoral Program in Economics and Management

Estimation and Inference for High Dimensional Time Series Data Models

---

Jonas Striaukas

Composition of the jury:

Committee:
Prof. Andrii Babii (University of North Carolina – Chapel Hill)
Prof. Rudy De Winne (Université catholique de Louvain) – Supervisor
Prof. Geert Dhaene (KU Leuven)
Prof. Eric Ghysels (University of North Carolina – Chapel Hill) – Supervisor

President of the jury:
Prof. Christian Hafner (Université catholique de Louvain)

Louvain-la-Neuve, February, 2022

To my grandma Sofija, parents and wife

"If you torture the data long enough, it will confess."

— Ronald Coase *(1991 Nobel Prize in Economics)*

Abstract

This doctoral thesis consists of three research articles on the general topic of high dimensional mixed frequency data models. The papers are preceded by an introductory chapter. Each chapter is devoted to analyze and propose specific methods tailored for the specific data structures which are motivated by the empirical application(s) consider in the chapter.

Chapter 2 introduces structured machine learning regressions for high-dimensional time series data potentially sampled at different frequencies. The sparse-group LASSO estimator can take advantage of such time series data structures and outperforms the unstructured LASSO. We establish oracle inequalities for the sparse-group LASSO estimator within a framework that allows for the mixing processes and recognizes that the financial and the macroeconomic data may have heavier than exponential tails. An empirical application to nowcasting US GDP growth indicates that the estimator performs favorably compared to other alternatives and that text data can be a useful addition to more traditional numerical data. Our methodology is implemented in the R package *midasml*, available from CRAN.

In Chapter 3, we study Granger causality testing for high-dimensional time series using regularized regressions. To perform proper inference, we rely on heteroskedasticity and autocorrelation consistent (HAC) estimation of the asymptotic variance and develop the inferential theory in the high-dimensional setting. To recognize the time series data structures we focus on the sparse-group LASSO estimator, which includes the LASSO and the group LASSO as special cases. We establish the debiased central limit theorem for low dimensional groups of regression coefficients and study the HAC estimator of the long-run variance based on the sparse-group LASSO residuals. This leads to valid time series inference for individual regression coefficients as well as groups, including Granger causality tests. The treatment relies on a new Fuk-Nagaev inequality for a class of τ-mixing processes with heavier than Gaussian tails, which is of independent interest. In an empirical application, we study the Granger causal relationship between the VIX and financial news.

Chapter 4 extends the structured machine learning regressions for prediction and inference to panel data consisting of series sampled at different frequencies. Motivated by the empirical problem of predicting corporate earnings for a large cross-section of firms with macroeconomic, financial, and news time series sampled at different frequencies, we focus

on the sparse-group LASSO regularization. This type of regularization can take advantage of the mixed frequency time series panel data structures and we find that it empirically outperforms the unstructured machine learning methods. We obtain oracle inequalities for the pooled and fixed effects sparse-group LASSO panel data estimators recognizing that financial and economic data exhibit heavier than Gaussian tails. To that end, we leverage on a novel Fuk-Nagaev concentration inequality for panel data consisting of heavy-tailed τ-mixing processes which may be of independent interest in other high-dimensional panel data settings. Lastly, we provide a valid inference method based on HAC estimator and the debiased LASSO framework for long panels. In two empirical applications, we consider nowcasting large pool of firm-level P/E ratios and studying which factors Granger cause the earnings prediction errors made by the analysts. In the former, we show the usefulness of structured machine learning techniques compared to the unstructured LASSO, and show that ML-based methods are indeed affected by the tail behavior of the data. In the latter application, we show that macro information is largely missed by the analysts when forming predictions; a result that was previously documented in the literature based on individual regression methods.

Acknowledgments

Firstly, I would like to sincerely thank my thesis advisor Eric Ghysels. During these four years of the thesis, I always had great support and research advice, resulting in three thesis chapters we co-authored. I have always been very interested in each of your ideas and am very honored to have been able to work with you. Thank you for inviting me to Chapel Hill twice for short visits during my thesis and many other great opportunities. I am also very grateful to Andrii Babii for his excellent advice and rigor to guide me through my research. Our team has done excellent research; I hope to collaborate with both of you in the years to come.

Besides, I would like to warmly thank the rest of my thesis committee, Geert Dhaene and Rudy De Winne, for their insightful comments and professional advice, which helped widen my research and enrich this thesis. I particularly thank Rudy for his great all-around support during my thesis years.

During various conferences, seminars, summer schools and visits, I had the chance to meet remarkable researchers who were kind enough to chat with me and whose work inspired me. I would like to quote, among others: Matias Cattaneo, Domenico Giannone, Christian Hafner, Christian Hansen, Juan-Pablo Ortega, Eugen Pircalabelu, Rainer von Sachs, Martin Spindler.

I warmly thank my collaborators on different projects and other people that helped me throughout the thesis period and before: Ryan Ball, Daniel Buncic, Leonardo Iania, and Matthias Weber. I especially thank Daniel, my master thesis advisor, for his valuable comments on my early PhD thesis work and his guidance during my master studies. I also thank Matthias, who kick-started my interest in high-dimensional statistics.

During my studies or seminars, I had the pleasure of meeting many PhD students with whom I had many fruitful discussions: Angelo, Cheikh, Cyrille, Francesco, Foti, Jean-Charles, Paolo, Pavel, Sofonias, Taiki, ...

I also owe a big thanks to the whole kiaulytės-crew: Gabrielius, Gytis, Jonas G., Jonas K., Julius, Justinas, Martynas, Tomas, Tumas, Vilius and Žilvinas.

Some administrative and IT related burdens were quickly reduced thanks to the efficiency of Alain, Catherine, Jennifer, Nancy, Raphaël and Sandrine. I especially thank Catherine for helping me to efficiently go through administrative hurdles.

I am grateful to Fonds de la Recherche Scientifique – FNRS for supporting my thesis through PDR grant and the Aspirant Research Fellowship grant FC21388. The fund also supported my two visits at Chapel Hill which I am grateful for.

I close with a big thank you to my family. My parents Eglė and Gintaras, my sister Akvilė – a big thank you for your constant support and help. I also thank my wife's family for constant encouragement. Last but definitely not least, I thank Akvilė, my life partner, for her incredible support and patience, for cheering me up when I needed the most, for helping me and giving advice on my thesis and beyond.

Contents

Contents **8**

1 Introduction and summary **11**
1.1 Introduction 11
1.2 High-dimensional regression 11
1.3 Time series data and LASSO 13
1.3.1 Monte Carlo evidence 13
1.4 Structured high-dimensional time series regressions 15
1.5 Conclusion 18

2 Machine Learning Time Series Regressions with an Application to Nowcasting **19**
2.1 Introduction 19
2.2 High-dimensional mixed frequency regressions 24
2.3 High-dimensional time series regressions 27
2.3.1 High-dimensional regressions and τ-mixing 27
2.3.2 Estimation and prediction properties 29
2.4 Monte Carlo experiments 33
2.4.1 Simulation Design 34
2.4.2 Simulation results 35
2.5 Nowcasting US GDP with macro, financial and textual news data 36
2.6 Conclusion 43
A2.1 Dictionaries . . . . . Appx. - 46
A2.2 Proofs of main results . . . . . Appx. - 47
A2.3 ARDL-MIDAS: moments and τ-mixing coefficients . . . . . Appx. - 52
A2.4 Monte Carlo Simulations . . . . . Appx. - 54
A2.5 Detailed description of data and models . . . . . Appx. - 60
A2.5.1 Additional results . . . . . Appx. - 64

3 High-Dimensional Granger Causality Tests with an Application to VIX and News **65**
3.1 Introduction . . . 65
3.2 HAC-based inference for sg-LASSO . . . 69
3.2.1 Debiased central limit theorem . . . 69
3.2.2 Nodewise LASSO . . . 74
3.2.3 HAC estimator . . . 75
3.2.4 High-dimensional Granger causality tests . . . 76
3.3 Fuk-Nagaev inequality . . . 77
3.4 Monte Carlo experiments . . . 78
3.5 Testing Granger causality for VIX and financial news . . . 81
3.5.1 Main results . . . 82
3.5.1.1 Granger causality of news topics . . . 83
3.5.1.2 Bi-directional Granger causality . . . 84
3.5.1.3 Granger causal clusters of news topics . . . 84
3.6 Conclusion . . . 85
A3.1 Proofs . . . Appx. - 86
A3.2 Data . . . Appx. - 105

4 Machine Learning Panel Data Regressions with Heavy-tailed Dependent Data: Theory and Applications **109**
4.1 Introduction . . . 109
4.2 High-dimensional (mixed frequency) panels . . . 112
4.3 Oracle inequalities . . . 115
4.3.1 τ-mixing . . . 116
4.3.2 Pooled regression . . . 116
4.3.3 Fixed effects . . . 120
4.4 Debiased inference . . . 122
4.5 Monte Carlo experiments . . . 124
4.5.1 Simulation design (nowcasting) . . . 125
4.5.2 Simulation results (nowcasting) . . . 126
4.5.3 Simulation design (Granger causality) . . . 126
4.5.4 Simulation results (Granger causality) . . . 129
4.6 Empirical Applications . . . 130
4.6.1 Nowcasting P/E ratios . . . 131
4.6.1.1 Data description . . . 132
4.6.1.2 Tuning parameters . . . 133
4.6.1.3 Models and main results . . . 134
4.6.1.4 Significance test of nowcasts . . . 139

4.6.2 Do analysts leave money on the table? 142
4.6.2.1 Granger causality tests 143
4.7 Conclusions 147
A4.1 Concentration and moment inequalities . . . . . Appx. - 149
A4.2 Large N and T central limit theorem . . . . . Appx. - 153
A4.3 Proofs . . . . . Appx. - 155
A4.4 Additional empirical results – nowcasting application . . . . Appx. - 165
A4.5 Additional empirical results – Granger causality application Appx. - 168
A4.6 Data description . . . . . Appx. - 169
A4.6.1 Firm-level data . . . . . Appx. - 169
A4.6.1.1 Firm sample selection . . . . . Appx. - 169
A4.6.1.2 Firm-specific text data . . . . . Appx. - 169

Bibliography **173**

Chapter 1

Introduction and summary

1.1 Introduction

Over the past decade or so, machine learning – or statistical learning – techniques have been increasingly used in econometrics literature covering both theory and empirical research. Thus far, most of the methods have dealt with independent and identically distributed (i.i.d.) and sub-Gaussian data-generating processes, with little attention paid to the time series data typically encountered in economics and finance. This thesis develops new methods for high-dimensional time series and panel data in several contexts and applies those techniques in various novel applications, showing the utility of such methods.

In this introductory chapter, I review the state-of-art methods for regularized regressions. Next, I provide a brief simulation study to show that once the data generating process (DGP) deviates from i.i.d. and Gaussian-like data assumptions, standard results on performance guarantees may no longer hold. I then introduce the remaining chapters of the thesis, provide a glimpse of ideas put forward, and offer an overview of the results obtained in the subsequent chapters.

1.2 High-dimensional regression

To introduce the regularized regressions, I consider a linear model

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

where $\mathbf{y} \in \mathbf{R}^T$ is the response variable, $\mathbf{X} \in \mathbf{R}^{T \times p}$ is the covariate matrix, $\beta \in \mathbf{R}^p$ is the slope coefficient vector and $\mathbf{u}$ are the residuals. In the low-dimensional case, $p < T$, the model can be estimated by applying an ordinary least squares (OLS) estimator. However, the case of $p \gg T$ is

more relevant for modern data sets with a much larger number of potential predictors relative to the sample size. In such cases, we typically apply regularization by adding a penalty function in the minimization problem. More concretely, we consider estimators of the form that solve the following minimization problem:

$$\min_{\beta \in \mathbf{R}^p} \|\mathbf{y} - \mathbf{X}\beta\|_T^2 + h(\beta, \lambda),$$

where $\|.\|_T^2$ is the empirical norm, i.e. $\|u\|_T^2 = \langle u, u \rangle / T$, and $h(\beta, \lambda)$ is a penalty function. A natural choice for the penalty function is a function which counts the number of non-zero coordinates in the coefficient vector, i.e. ℓ_0 zero "norm":

$$h(\beta, \lambda) = \lambda |\beta|_0 = \lambda \sum_{i \in [p]} \mathbb{1}\{\beta_i \neq 0\}.$$

The minimization problem with this penalty function is called the *best subset selection*. However, the minimization problem is non-convex and NP-hard, which precludes the possibility that any algorithm will successfully find the optimal solution in a linear time.[1]

A convex alternative is the so-called LASSO regression, first studied by Tibshirani (1996). In LASSO regression, the ℓ_1 norm is considered instead of ℓ_0, i.e.:

$$h(\beta, \lambda) = \lambda |\beta|_1 = \lambda \sum_{i \in [p]} |\beta_i|.$$

For large enough λ values, the penalty function induces sparsity in the high-dimensional β coefficient vector, which leads to the variable selection property of the LASSO estimator. For any $\lambda > 0$, the estimator also shrinks the whole coefficient vector towards zero. It is worth noting that even though recent advances in optimization provide techniques to find approximate solutions for the best subset selection problem, it seems that there is no gain in performance compared to the computationally attractive LASSO; see e.g., Hastie, Tibshirani, and Tibshirani (2019).

There are additional alternatives for the penalty functions to achieve regularization, including, for example, the ℓ_2-norm, referred to as Tikhonov regularization or ridge regression; the elastic net of Zou and Hastie (2005), which is a convex combination of LASSO and ridge; and the group LASSO of

[1] Recent advances in integer programming allow us to find an approximate solution to such a minimization problem.

Yuan and Lin (2006), which contains a penalty for selection and estimation in regressions with grouped variables. In this thesis, we study the structured regularization called sparse-group LASSO; see Simon, Friedman, Hastie, and Tibshirani (2013), which turns out to be useful for time series; see Chapter 2 and 3; or panel data applications, for which, see Chapter 4. In these cases, we show that sparse-group LASSO is particularly relevant in mixed-frequency data applications as one can easily trace the structure between covariates due to the temporal dependence of the lags of the specific variable and use this information to more accurately estimate regression coefficients.

1.3 Time series data and LASSO

To date, the literature on LASSO-type econometric methods has typically assumed i.i.d. and sub-Gaussian data with exponential tails. In this section, I provide some preliminary Monte-Carlo simulation-based evidence that LASSO estimator performance depends on the time series properties of the data.

1.3.1 Monte Carlo evidence

The purpose of the following Monte Carlo study is to show that the performance of the LASSO-type regression also depends on the time series properties of the data – dependence and heaviness of the tails. Thus, I conduct the following Monte Carlo experiment. The data generating process is the following linear regression model:

$$\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{u},$$

where $\beta_0 = (5, 4, 3, 2, 1, \underbrace{0, \ldots, 0}_{p-5})$ is the true target parameter vector.

I consider two scenarios for $\mathbf{X}$ and $\mathbf{u}$ and simulate the model for a grid of $p \in \{10, 20, 40, \ldots, 400\}$ values and

1. (dependence) for each column of $\mathbf{X}$ denoted as X^j, $j \in [p]$ is simulated as a Gaussian AR(1) process: $X^j_{t+1} = \rho X^j_t + \epsilon$, where $t \in [T]$ and $\epsilon \sim_{i.i.d.} N(0, 1)$. The error term $\mathbf{u}$ is simulated as an i.i.d. Gaussian process. The persistence parameter ρ is set to one of the values in the grid $\rho \in \{0.0, 0.05, \ldots, 0.95\}$;

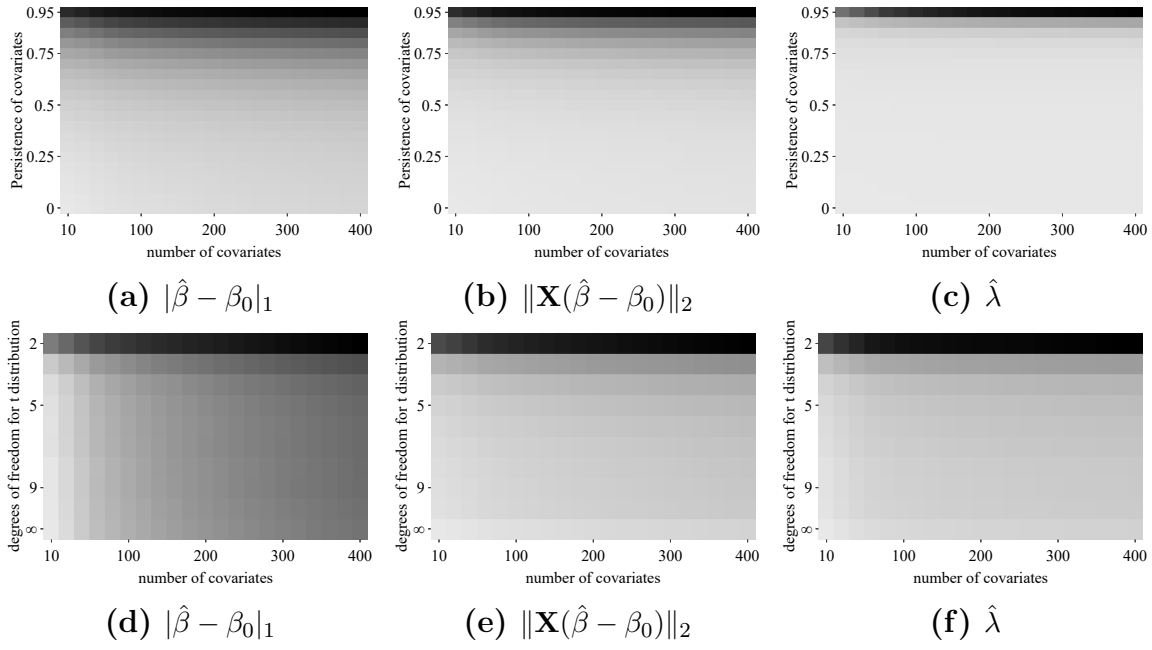
2. (heavy tails) the design matrix $\mathbf{X}$ and the error term $\mathbf{u}$ are drawn from $\sim_{i.i.d.}$ student-$t(\nu)$. The degrees of freedom parameter ν is set to one of the values in the grid $\nu \in \{\infty, 10, 9, \ldots, 2\}$.

I simulate two separate data sets, $\mathbf{u}^k, \mathbf{X}^k, \mathbf{y}^k, k \in \{1, 2\}$, fixing the sample size to $T = 100$. The first sample, $k = 1$, is used to train the model on a grid of λ parameter values.[2] The second sample, $k = 2$, is used to optimize the λ value choice. The results are reported in a heatmap form; see 1.1, where I plot the estimation error $|\hat{\beta} - \beta_0|_1$, the prediction error $\|\mathbf{X}(\hat{\beta} - \beta_0)\|_2$, and the optimal tuning parameter $\hat{\lambda}$ on a two-dimensional grid that reflects changes in the dimensionality of the regression problem p and the strength of the DGP properties. Estimation and prediction error is computed using the training sample $k = 1$ data using the best $\hat{\lambda}$, which is based on the $k = 2$ sample. Each entry in the heatmap reflects the magnitude of the value computed for the specific metric averaged over 5000 simulations. In each heatmap, x-axis reflects the growing number of covariates used in the model, while y-axis reflects the change of either the persistence parameter ρ (first row - dependence scenario) or the degrees of freedom ν (second row - heavy-tailed data scenario).

Discussion: The results are reported in Figure 1.1. First, it is clear that the dependence or the heaviness of tails in $\mathbf{X}$ affect the performance of the LASSO estimator. When looking at the first row (dependent data scenario), the results indicate that the worst case for the estimation error appears to be when p and ρ are large (Subfigure a). It is also clear that for a fixed p, the results deteriorate once ρ increases. For prediction,the results appear to be more homogeneous across the p choice – the worst case is $\rho = 0.95$ (Subfigure b). The λ parameter seems to increase with the ρ parameter (Subfigure c).

The results for the heavy-tailed data scenario seem to be slightly different, albeit with a similar pattern. As in the dependent data scenario, the estimation (Subfigure d) and prediction (Subfigure e) performance of LASSO deteriorates in the extreme cases when $\nu \to 2$, LASSO tuning parameter λ increases as $\nu \to 2$ (Subfigure f). The performance of LASSO deteriorates once the dimensionality increases.

[2] In this simulation study, I use `glmnet` implementation of the LASSO estimation procedure, which relies on a cyclical coordinate descent algorithm. The λ parameter sequence is constructed in a data-driven way, i.e. the maximum value of λ, $\lambda_{\max}$, is taken as $|\mathbf{X}^\top \mathbf{y}|_\infty$, which ensures that all β coefficients are zero for the $\lambda_{\max}$ tuning parameter. The remaining tuning parameters are set on an equidistant decreasing grid in log space, where $\lambda_{\min} = 0.0001\lambda_{\max}$.



(a) $|\hat{\beta} - \beta_0|_1$ **(b)** $\|\mathbf{X}(\hat{\beta} - \beta_0)\|_2$ **(c)** $\hat{\lambda}$

(d) $|\hat{\beta} - \beta_0|_1$ **(e)** $\|\mathbf{X}(\hat{\beta} - \beta_0)\|_2$ **(f)** $\hat{\lambda}$

Figure 1.1: Simulation results – Each column represents different quantity: estimation error (Subfigure a and d), prediction error (Subfigure b and e) and λ tuning parameter estimate (Subfigure c and f). Each row represents different simulation scenarios: row pertaining to Subfigure (a), (b), and (c) shows the results for the dependence scenarios, while the remaining plots show the results for the heavy-tailed data scenarios.

1.4 Structured high-dimensional time series regressions

The main idea exploited in this thesis to estimate high-dimensional time series and panel data regressions is to apply structured regularization over the lags of each covariate, thereby introducing additional structure and regularization of the regression function.

To put the idea into context, suppose we are interested in modeling a low-frequency variable $\mathbf{y} = (y_t)_{t\in[T]}$ as a function of some high-frequency variable lags. Suppose only one low-frequency time period lag of a high-frequency variable is included in such a model and suppose the sampling frequency ratio is m, i.e. the high-frequency lags are $\{x_{t-(j-1)/m}, j \in [m], t \in [T]\}$. Such a regression model (excluding the intercept) is:

$$y_t = \sum_{j=1}^{m} b_j x_{t-(j-1)/m} + u_t, \quad t \in [T], \tag{1.1}$$

where b_j are the regression coefficients associated with each high-frequency lag. In the literature, this type of model is called a Mixed Data Sampling (MIDAS) regression. If the model parameters are left unconstrained as in

(1.1), the model can be estimated with OLS and is typically referred to as an unconstrained MIDAS model (UMIDAS). In most cases, however, UMIDAS is not practical, or even feasible, as one needs to estimate a large number of parameters, which grows with the sampling ratio m. For instance, if we were to regress a quarterly variable on daily ($m = 66$), we need $T > 66$ to be able to estimate the single-variate model. In practice, quarterly time series data are not long; hence, such an approach is not particularly appealing. To make the regression model (1.1) more suitable for practical purposes, the MIDAS literature suggests various ways of parameterizing the lag polynomial $\{b_j, j \in [m]\}$ using certain weight functions to alleviate the dimensionality problem; see e.g., Ghysels, Sinko, and Valkanov (2007).[3] Such parametrizations typically lead to more accurate prediction, estimation and inference in various settings.

Note that in case m is large, we may apply the LASSO estimator without restricting the lag polynomial, i.e. estimate the LASSO-UMIDAS model. However, as argued in Ghysels, Kvedaras, and Zemlys-Balevičius (2020), sparsity imposed on the lag polynomial may not be appealing since typically many individually small and comparable in size coefficients are non-zero, which LASSO would screen out. This would lead to a biased prediction. At the same time, it is well known that the LASSO estimator cannot handle highly correlated predictors well, a feature that the UMIDAS model possesses due to the temporal dependence of the high-frequency variable. The structure of the ridge regression is not appealing either in MIDAS regression settings due to its inability to recover smooth decay functional shapes of the lag polynomial, a pattern which is typically encountered in MIDAS settings (Ghysels, Kvedaras, and Zemlys-Balevičius, 2020).

In a high-dimensional setting, an additional dimensionality problem occurs when we include a large set of high-frequency covariates in the regression model. Denoting the total number of covariates K, our covariate set becomes $\{x_{t-(j-1)/m,k}, j \in [m], t \in [T], k \in [K]\}$. In such a case, the model becomes:

$$y_t = \sum_{k=1}^{K} \sum_{j=1}^{m} b_{k,j} x_{t-(j-1)/m,k} + u_t, \quad t \in [T]. \tag{1.2}$$

As noted previously, applying LASSO together with UMIDAS scheme is not appealing even if $K = 1$, and this approach becomes even more

[3]Examples of weight functions applied in MIDAS literature are Almon polynomials, exponential polynomails, Beta density; see Ghysels, Sinko, and Valkanov (2007) for more detail.

stringent for large K cases. The key idea exploited in this thesis applies *structured* regularization and MIDAS polynomials instead of (blindly) applying regularization together with the UMIDAS scheme. In particular, we apply Legendre polynomials to model MIDAS weight functions; see Chapter 2 for more detail on Legendre polynomials, and exploit the structure of the regression model (1.2) by applying group structure over the lags of each covariate. For example, for $K = 2$ the model is:

$$y_t = \underbrace{\sum_{j=1}^{m} b_{1,j} x_{t-(j-1)/m,1}}_{\text{group 1}} + \underbrace{\sum_{j=1}^{m} b_{2,j} x_{t-(j-1)/m,2}}_{\text{group 2}} + u_t, \quad t \in [T]. \tag{1.3}$$

Hence, we may apply group structure on the first and second covariate lags and encode this information into the penalty function through a group LASSO. The general approach to modeling high-dimensional MIDAS models is to apply a sparse-group LASSO (sg-LASSO) estimator, which strikes a good balance in learning the shape of MIDAS weights and the important covariates in the regression model; Chapter 2 provides more detail on the sg-LASSO estimator.

Overview of thesis chapters: Chapter 2 introduces structured machine learning regressions for high-dimensional time series data potentially sampled at different frequencies. Oracle inequalities are established for the sg-LASSO estimator within a framework that allows for the mixing processes and recognizes that the financial data and the macroeconomic data may have heavier than exponential tails. An empirical application to nowcasting US GDP growth indicates that the estimator performs favorably compared to other alternatives and that text data can be a useful addition to more traditional numerical data.

In Chapter 3, Granger causality testing is studied for high-dimensional time series using regularized regressions. To perform proper inference, the method relies on heteroskedasticity and autocorrelation consistent (HAC) estimation of the asymptotic variance. The inferential theory in the high-dimensional setting is developed. The debiased central limit theorem is established for low-dimensional groups of regression coefficients and the HAC estimator of the long-run variance based on the sg-LASSO residuals is studied. This leads to valid time series inference for individual regression coefficients as well as groups, including Granger causality tests. The treatment relies on a new Fuk-Nagaev inequality for a class of τ-mixing processes with heavier than Gaussian tails, which is of independent interest.

In an empirical application, a study of the Granger causal relationship between the VIX and financial news is provided.

Chapter 4 extends the structured machine learning regressions for prediction and inference to panel data consisting of series sampled at different frequencies. We obtain oracle inequalities for the pooled and fixed effects sg-LASSO panel data estimators, recognizing that financial and economic data exhibit heavier than Gaussian tails. A new Fuk-Nagaev concentration inequality for panel data consisting of heavy-tailed τ-mixing processes is proven, which may be of independent interest in other high-dimensional panel data settings. Lastly, we provide a valid inference method based on the HAC estimator and the debiased sg-LASSO framework for long panels. In two empirical applications, we consider nowcasting a large pool of firm-level P/E ratios and studying which factors in the Granger analyses cause prediction errors of earnings made by analysts. In the former, we show the usefulness of structured machine learning techniques compared to the unstructured LASSO, and show that ML-based methods are indeed affected by the tail behavior of the data. In the latter application, we show that macro information is largely missed by analysts when forming predictions, a result that has been documented in the literature based on individual regression methods.

1.5 Conclusion

In this introductory chapter, I discussed high-dimensional regularized regression models applied for time series data regression models. In particular, I discuss the LASSO estimator, which is arguably the most widely applied estimator for such models. I provide a brief Monte-Carlo study and show evidence of the performance of LASSO-type methods under different DGP scenarios. The results demonstrate that LASSO performance, both in terms of estimation and prediction, depends on the properties of the time series data.

This thesis is devoted to studying high-dimensional time series and panel data regression models, and provides both theoretical and practical tools to analyze such data. A key distinctive feature of this thesis is the recognition that time series data typically encountered in applications are sampled at a mixed frequency; hence, the proposed methods are general enough to handle both single-frequency and more generally mixed-frequency data models.

Chapter 2

Machine Learning Time Series Regressions with an Application to Nowcasting

with Andrii Babii and Eric Ghysels

2.1 Introduction

The statistical imprecision of quarterly gross domestic product (GDP) estimates, along with the fact that the first estimate is available with a delay of nearly a month, pose a significant challenge to policy makers, market participants, and other observers with an interest in monitoring the state of the economy in real time; see, e.g., Ghysels, Horan, and Moench (2018) for a recent discussion of macroeconomic data revisions and publication delays. A term originated in meteorology, nowcasting pertains to the prediction of the present and very near future. Nowcasting is intrinsically a mixed frequency data problem as the object of interest is a low-frequency data series (e.g., quarterly GDP), whereas the real-time information (e.g., daily, weekly, or monthly) can be used to update the state, or to put it differently, to *nowcast* the low-frequency series of interest. Traditional methods used for nowcasting rely on dynamic factor models that treat the underlying low frequency series of interest as a latent process with high frequency data noisy observations. These models are naturally cast in a state-space form and inference can be performed using likelihood-based methods and Kalman filtering techniques; see Bańbura, Giannone, Modugno, and Reichlin (2013) for a survey.

So far, nowcasting has mostly relied on the so-called standard macroeconomic data releases, one of the most prominent examples being the Employment Situation report released on the first Friday of every month by the US Bureau of Labor Statistics. This report includes the data on the nonfarm payroll employment, average hourly earnings, and other summary

statistics of the labor market activity. Since most sectors of the economy move together over the business cycle, good news for the labor market is usually good news for the aggregate economy. In addition to the labor market data, the nowcasting models typically also rely on construction spending, (non-)manufacturing report, retail trade, price indices, etc., which we will call the traditional macroeconomic data. One prominent example of nowcast is produced by the Federal Reserve Bank of New York relying on a dynamic factor model with thirty-six predictors of different frequencies; see Bok, Caratelli, Giannone, Sbordone, and Tambalotti (2018) for more details.

Thirty-six predictors of traditional macroeconomic series may be viewed as a small number compared to hundreds of other potentially available and useful nontraditional series. For instance, macroeconomists increasingly rely on nonstandard data such as textual analysis via machine learning, which means potentially hundreds of series. A textual analysis data set based on *Wall Street Journal* articles that has been recently made available features a taxonomy of 180 topics; see Bybee, Kelly, Manela, and Xiu (2020). Which topics are relevant? How should they be selected? Thorsrud (2020) constructs a daily business cycle index based on quarterly GDP growth and textual information contained in the daily business newspapers relying on a dynamic factor model where time-varying sparsity is enforced upon the factor loadings using a latent threshold mechanism. His work shows the feasibility of traditional state space setting, yet the challenges grow when we also start thinking about adding other potentially high-dimensional data sets, such as payment systems information or GPS tracking data. Studies for Canada (Galbraith and Tkacz (2018)), Denmark (Carlsen and Storgaard (2010)), India (Raju and Balakrishnan (2019)), Italy (Aprigliano, Ardizzi, and Monteforte (2019)), Norway (Aastveit, Fastbø, Granziera, Paulsen, and Torstensen (2020)), Portugal (Duarte, Rodrigues, and Rua (2017)), and the United States (Barnett, Chauvet, Leiva-Leon, and Su (2016)) find that payment transactions can help to nowcast and to forecast GDP and private consumption in the short term; see also Moriwaki (2019) for nowcasting unemployment rates with smartphone GPS data, among others. We could quickly reach numerical complexities involved with estimating high-dimensional state space models, making the dynamic factor model approach potentially computationally prohibitively complex and slow, although some alternatives to the Kalman filter exist for the large data environments; see e.g., Chan and Jeliazkov (2009) and Delle Monache and Petrella (2019). In this paper, we study nowcasting a low-frequency

series – focusing on the key example of US GDP growth – in a data-rich environment, where our data not only includes conventional high-frequency series but also nonstandard data generated by textual analysis of financial press articles. Several novel contributions are required to achieve our goal. The contributions of our paper are both theoretical and practical. Regarding the former: (a) we propose a new structured approach to high-dimensional regularized time regression problems, (b) we establish a complete estimation and prediction theory for high-dimensional time series regressions under assumptions comparable to the classical GMM and QML estimators, and (c) we establish nonasymptotic and asymptotic estimation and prediction properties of our regularized time series regression approach. Regarding the practical contributions we document superior nowcasting performance with respect to the state-of-the-art state space model approach to nowcasting implemented by the Federal Reserve Bank of New York. In the remainder of this Introduction we devote a paragraph to each of these contributions, starting with the theoretical ones.

First, we argue that the high-dimensional mixed frequency time series regressions involve certain data structures that once taken into account should improve the performance of unrestricted estimators in small samples. These structures are represented by groups covering lagged dependent variables and groups of lags for a single (high-frequency) covariate. To that end, we leverage on the sparse-group LASSO (sg-LASSO) regularization that accommodates conveniently such structures; see Simon, Friedman, Hastie, and Tibshirani (2013). The attractive feature of the sg-LASSO estimator is that it allows us to combine effectively the approximately sparse and dense signals; see e.g., Carrasco and Rossi (2016) for a comprehensive treatment of high-dimensional dense time series regressions as well as Mogliani and Simoni (2021) for a complementary to ours Bayesian view of penalized MIDAS regressions.

Second, we recognize that the economic and financial time series data are persistent and often heavy-tailed, while the bulk of the machine learning methods assumes i.i.d. data and/or exponential tails for covariates and regression errors; see Belloni, Chernozhukov, Chetverikov, Hansen, and Kato (2020) for a comprehensive review of high-dimensional econometrics with i.i.d. data. There have been several recent attempts to expand the asymptotic theory to settings involving time series dependent data, mostly for the LASSO estimator. For instance, Kock and Callot (2015) and Uematsu and Tanaka (2019) establish oracle inequalities for regressions with i.i.d. errors with sub-Gaussian tails; Wong, Li, and Tewari (2020)

consider β-mixing series with exponential tails; Wu and Wu (2016), Han and Tsay (2017), and Chernozhukov, Härdle, Huang, and Wang (2021) establish oracle inequalities for causal Bernoulli shifts with independent innovations and polynomial tails under the functional dependence measure of Wu (2005); see also Medeiros and Mendes (2016) and Medeiros and Mendes (2017) for results on the adaptive LASSO based on the triplex tail inequality for mixingales of Jiang (2009). Despite these efforts, there is no complete estimation and prediction theory for high-dimensional time series regressions under the assumptions comparable to the classical GMM and QML estimators. For instance, the best currently available results are too restrictive for the MIDAS projection model, which is typically an example of a causal Bernoulli shift with *dependent innovations*. Moreover, the *mixing processes* with *polynomial tails* that are especially relevant for the financial and macroeconomic time series have not been properly treated due to the fact that the sharp Fuk-Nagaev inequality was not available in the relevant literature until recently. The Fuk-Nagaev inequality, see Fuk and Nagaev (1971), describes the concentration of sums of random variables with a mixture of the sub-Gaussian and the polynomial tails. It provides sharp estimates of tail probabilities unlike Markov's bound in conjunction with the Marcinkiewicz-Zygmund or Rosenthal's moment inequalities.

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

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

The rest of the paper is organized as follows. Section 2.2 presents the setting of (potentially mixed frequency) high-dimensional time series regressions. Section 2.3 characterizes nonasymptotic estimation and prediction accuracy of the sg-LASSO estimator for τ-mixing processes with polynomial tails. We report on a Monte Carlo study in Section 2.4 which provides further insights regarding the validity of our theoretical analysis in small sample settings typically encountered in empirical applications. Section 2.5 covers the empirical application. Conclusions appear in Section 2.6.

Notation: For a random variable $X \in \mathbf{R}$, let $\|X\|_q = (\mathbb{E}|X|^q)^{1/q}$ be its L_q norm with $q \geq 1$. For $p \in \mathbf{N}$, put $[p] = \{1, 2, \ldots, p\}$. For a vector $\Delta \in \mathbf{R}^p$ and a subset $J \subset [p]$, let Δ_J be a vector in $\mathbf{R}^p$ with the same coordinates as Δ on J and zero coordinates on J^c. Let $\mathcal{G}$ be a partition of $[p]$ defining the group structure, which is assumed to be known to the econometrician. For a vector $\beta \in \mathbf{R}^p$, the sparse-group structure is described by a pair $(S_0, \mathcal{G}_0)$, where $S_0 = \{j \in [p] : \beta_j \neq 0\}$ and $\mathcal{G}_0 = \{G \in \mathcal{G} : \beta_G \neq 0\}$ are the support and respectively the group support of β. We also use $|S|$ to denote the cardinality of arbitrary set S. For $b \in \mathbf{R}^p$, its ℓ_q norm is denoted as $|b|_q = \left(\sum_{j\in[p]} |b_j|^q\right)^{1/q}$ for $q \in [1, \infty)$ and $|b|_\infty = \max_{j\in[p]} |b_j|$ for $q = \infty$. For $\mathbf{u}, \mathbf{v} \in \mathbf{R}^T$, the empirical inner product is defined as $\langle \mathbf{u}, \mathbf{v} \rangle_T = T^{-1} \sum_{t=1}^T u_t v_t$ with the induced empirical norm $\|.\|_T^2 = \langle ., . \rangle_T = |.|_2^2/T$. For a symmetric $p \times p$ matrix A, let $\text{vech}(A) \in \mathbf{R}^{p(p+1)/2}$ be its vectorization consisting of the lower triangular and the diagonal elements. For $a, b \in \mathbf{R}$, we put $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Lastly, we write $a_n \lesssim b_n$ if there

exists a (sufficiently large) absolute constant C such that $a_n \leq Cb_n$ for all $n \geq 1$ and $a_n \sim b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

2.2 High-dimensional mixed frequency regressions

Let $\{y_t : t \in [T]\}$ be the target low frequency series observed at integer time points $t \in [T]$. Predictions of y_t can involve its lags as well as a large set of covariates and lags thereof. In the interest of generality, but more importantly because of the empirical relevance we allow the covariates to be sampled at higher frequencies - with same frequency being a special case. More specifically, let there be K covariates $\{x_{t-(j-1)/m,k}, j \in [m], t \in [T], k \in [K]\}$ possibly measured at some higher frequency with $m \geq 1$ observations for every t and consider the following regression model

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

where $\phi(L) = I - \rho_1 L - \rho_2 L^2 - \cdots - \rho_J L^J$ is a low-frequency lag polynomial and $\psi(L^{1/m}; \beta_k)x_{t,k} = 1/m \sum_{j=1}^{m} \beta_{j,k} x_{t-(j-1)/m,k}$ is a high-frequency lag polynomial. For $m = 1$, we have a standard autoregressive distributed lag (ARDL) model, which is the workhorse regression model of the time series econometrics literature. Note that the polynomial $\psi(L^{1/m}; \beta_k)x_{t,k}$ involves the same m number of high-frequency lags for each covariate $k \in [K]$, which is done for the sake of simplicity and can easily be relaxed; see Section 2.5.

The ARDL-MIDAS model (using the terminology of Andreou, Ghysels, and Kourtellos (2013)) features $J + 1 + m \times K$ parameters. In the big data setting with a large number of covariates sampled at high-frequency, the total number of parameters may be large compared to the effective sample size or even exceed it. This leads to poor estimation and out-of-sample prediction accuracy in finite samples. For instance, with $m = 3$ (quarterly/monthly setting) and 35 covariates at 4 lagged quarters, we need to estimate $m \times K = 420$ parameters. At the same time, say the post-WWII quarterly GDP growth series has less than 300 observations.

The LASSO estimator, see Tibshirani (1996), offers an appealing convex relaxation of a difficult nonconvex best subset selection problem. It allows increasing the precision of predictions via the selection of sparse and parsimonious models. In this paper, we focus on the structured sparsity

with additional dimensionality reductions that aim to improve upon the unstructured LASSO estimator in the time series setting.

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

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

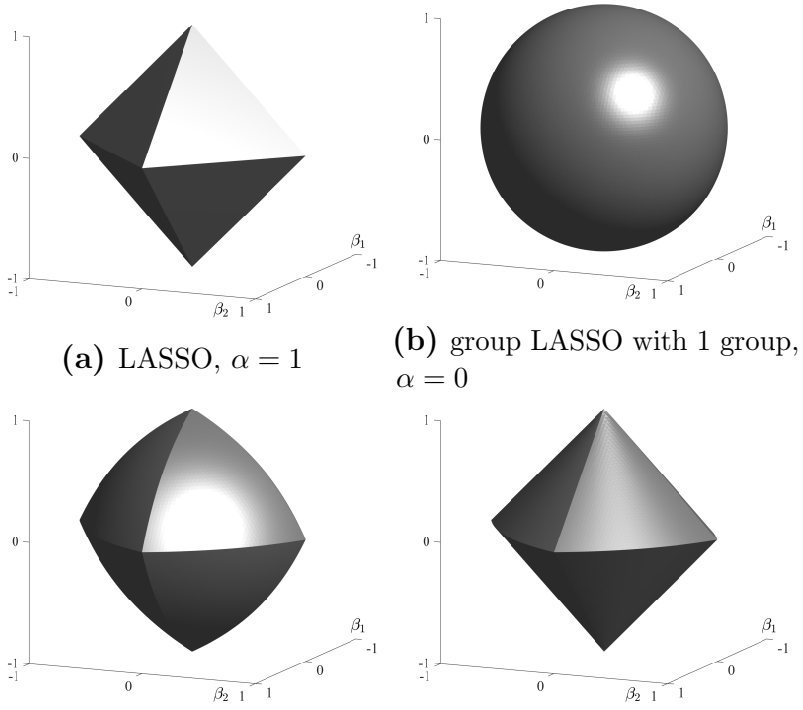
where β_k is L-dimensional vector of coefficients with $L \leq m$ and $\omega : [0,1] \times \mathbf{R}^L \to \mathbf{R}$ is some weight function. Second, we approximate the weight function as

$$\omega(u;\beta_k) \approx \sum_{l=1}^{L}\beta_{k,l}w_l(u), \qquad u \in [0,1], \tag{2.1}$$

where $\{w_l : \ l = 1,\dots,L\}$ is a collection of functions, called the *dictionary*. The simplest example of the dictionary consists of algebraic power polynomials, also known as Almon (1965) polynomials in the time series regression analysis literature. More generally, the dictionary may consist of arbitrary approximating functions, including the classical orthogonal bases of $L_2[0,1]$; see Appendix Section A.2.1 for more examples. Using orthogonal polynomials typically reduces the multicollinearity and leads to better finite sample performance. It is worth mentioning that the specification with dictionaries deviates from the standard MIDAS regressions and leads to a computationally attractive convex optimization problem, cf. Marsilli (2014a).

The size of the dictionary L and the number of covariates K can still be large and the *approximate sparsity* is a key assumption imposed throughout the paper. With the approximate sparsity, we recognize that assuming that most of the estimated coefficients are zero is overly restrictive and that the approximation error should be taken into account. For instance, the weight function may have an infinite series expansion, nonetheless, most can be captured by a relatively small number of orthogonal basis functions. Similarly, there can be a large number of economically relevant predictors, nonetheless, it might be sufficient to select only a smaller number of the most relevant ones to achieve good out-of-sample forecasting performance. Both model selection goals can be achieved with the LASSO estimator. However, the LASSO does not recognize that covariates at different (high-frequency) lags are temporally related.

In the baseline model, all high-frequency lags (or approximating functions once we parameterize the lag polynomial) of a single covariate constitute a group. We can also assemble all lag dependent variables into a group. Other group structures could be considered, for instance combining various covariates into a single group, but we will work with the simplest group setting of the aforementioned baseline model. The sparse-group LASSO (sg-LASSO) allows us to incorporate such structure into the estimation procedure. In contrast to the group LASSO, see Yuan and Lin (2006), the sg-LASSO promotes sparsity *between* and *within* groups, and allows us to capture the predictive information from each group, such as approximating functions from the dictionary or specific covariates from each group.



(a) LASSO, $\alpha = 1$

(b) group LASSO with 1 group, $\alpha = 0$

(c) sg-LASSO with 1 group, $\alpha = 0.5$

(d) sg-LASSO with 2 groups, $\alpha = 0.5$

Figure 2.1: The figure shows the geometry of the constrained set, $\{b \in \mathbf{R}^2 : \Omega(b) \leq 1\}$, corresponding to the sparse-group LASSO penalty function for several groupings and values of α.

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

consisting of the intercept and the autoregressive coefficients, and $\beta_k \in \mathbf{R}^L$ denotes parameters of the high-frequency lag polynomial pertaining to the covariate $k \geq 1$. Then, the sparse-group LASSO estimator, denoted $\hat{\beta}$, solves the penalized least-squares problem

$$\min_{b \in \mathbf{R}^p} \|\mathbf{y} - \rho_0 - \mathbf{X}b\|_T^2 + 2\lambda\Omega(b) \tag{2.2}$$

with a penalty function that interpolates between the ℓ_1 LASSO penalty and the group LASSO penalty

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

where ρ_0 is the intercept which is not penalized, $\|b\|_{2,1} = \sum_{G \in \mathcal{G}} |b_G|_2$ is the group LASSO norm and $\mathcal{G}$ is a group structure (partition of $[p]$) specified by the econometrician. Note that estimator in equation ((2.2)) is defined as a solution to the convex optimization problem and can be computed efficiently, e.g., using an appropriate coordinate descent algorithm; see Simon, Friedman, Hastie, and Tibshirani (2013).

The amount of penalization in equation ((2.2)) is controlled by the regularization parameter $\lambda > 0$ while $\alpha \in [0, 1]$ is a weight parameter that determines the relative importance of the sparsity and the group structure. Setting $\alpha = 1$, we obtain the LASSO estimator while setting $\alpha = 0$, leads to the group LASSO estimator, which is reminiscent of the elastic net. In figure 2.1 we illustrate the geometry of the penalty function for different groupings and different values of α covering (a) LASSO with $\alpha = 1$, (b) group LASSO with one group, $\alpha = 0$, and two sg-LASSO cases (c) one group and (d) two groups both with $\alpha = 0.5$. In practice, groups are defined by a particular problem and are specified by the econometrician, while α can be fixed or selected jointly with λ in a data-driven way such as using the cross-validation.

2.3 High-dimensional time series regressions

2.3.1 High-dimensional regressions and τ-mixing

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

$$y_t = \sum_{j=0}^{\infty} x_{t,j}\beta_j + u_t, \qquad \mathbb{E}[u_t x_{t,j}] = 0, \quad \forall j \geq 1, \qquad t \in \mathbf{Z}, \tag{2.3}$$

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

Using the setting of equation (2.2), for a sample $(y_t, x_t)_{t=1}^T$, write

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

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

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

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

where $\mathrm{Lip}_1 = \left\{ f : \mathbf{R}^l \to \mathbf{R} : \ |f(x) - f(y)| \leq |x - y|_1 \right\}$ is a set of 1-Lipschitz functions. Let $(\xi_t)_{t \in \mathbf{Z}}$ be a stochastic process and let $\mathcal{M}_t = \sigma(\xi_t, \xi_{t-1}, \dots)$ be its canonical filtration. The *τ-mixing coefficient* of $(\xi_t)_{t \in \mathbf{Z}}$ is defined as

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

If $\tau_k \downarrow 0$ as $k \to \infty$, then the process $(\xi_t)_{t \in \mathbf{Z}}$ is called *τ-mixing*. The τ-mixing coefficients were introduced in Dedecker and Prieur (2004) as dependence measures weaker than mixing. Note that the commonly used α- and β-mixing conditions are too restrictive for the linear projection model with an ARDL-MIDAS process. Indeed, a causal linear process with dependent innovations is not necessary α-mixing; see also Andrews (1984) for an example of AR(1) process which is not α-mixing. Roughly

speaking, τ-mixing processes are somewhere between mixingales and α-mixing processes and can accommodate such counterexamples. At the same time, sharp Fuk-Nagaev inequalities are available for τ-mixing processes which to the best of our knowledge is not the case for the mixingales or near-epoch dependent processes; see Babii, Ghysels, and Striaukas (2020a).

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

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

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

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

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

2.3.2 Estimation and prediction properties

In this section, we introduce the main assumptions for the high-dimensional time series regressions and study the estimation and prediction properties of the sg-LASSO estimator covering the LASSO and the group LASSO estimators as special cases. The following assumption imposes some mild

restrictions on the stochastic processes in the high-dimensional regression equation (2.3).

Assumption 2.3.1 (Data)**.** *For every $j, k \in [p]$, the processes $(u_t x_{t,j},)_{t\in\mathbf{Z}}$ and $(x_{t,j}x_{t,k})_{t\in\mathbf{Z}}$ are stationary such that (i) $\|u_0\|_q < \infty$ and $\max_{j\in[p]} \|x_{0,j}\|_r = O(1)$ for some constants $q > 2r/(r-2)$ and $r > 4$; (ii) the τ-mixing coefficients are $\tau_k \leq ck^{-a}$ and respectively $\tilde{\tau}_k \leq ck^{-b}$ for all $k \geq 0$ and some $c > 0$, $a > (\varsigma - 1)/(\varsigma - 2)$, $b > (r-2)/(r-4)$, and $\varsigma = qr/(q+r)$.*

It is worth mentioning that the stationarity condition is not essential and can be relaxed to the existence of the limiting variance of partial sums at costs of heavier notations and proofs. Condition (i) requires that covariates have at least 4 finite moments, while the number of moments required for the error process can be as low as $2 + \epsilon$, depending on the integrability of covariates. Therefore, (i) may allow for heavy-tailed distributions commonly encountered in financial and economic time series, e.g., asset returns and volatilities. Given the integrability in (i), (ii) requires that the τ-mixing coefficients decrease to zero sufficiently fast; see Appendix, Section A.2.3 for moments and τ-mixing coefficients of ARDL-MIDAS. It is known that the β-mixing coefficients decrease geometrically fast, e.g., for geometrically ergodic Markov chains, in which case (ii) holds for every $a, b > 0$. Therefore, (ii) allows for relatively persistent processes.

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

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

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

Assumption 2.3.2 (Restricted eigenvalue)**.** *There exists a universal constant $\gamma > 0$ such that $\Delta^\top\Sigma\Delta \geq \gamma\sum_{G\in\mathcal{G}_0} |\Delta_G|_2^2$ for all $\Delta \in \mathcal{C}(c_0)$, where $c_0 = (c+1)/(c-1)$ for some $c > 1$.*

Recall that if Σ is a positive definite matrix, then for all $\Delta \in \mathbf{R}^p$, we have $\Delta^\top\Sigma\Delta \geq \gamma|\Delta|_2^2$, where γ is the smallest eigenvalue of Σ. Therefore, in this

case Assumption 2.3.2 is trivially satisfied because $|\Delta|_2^2 \geq \sum_{G\in\mathcal{G}_0} |\Delta_G|_2^2$. The positive definiteness of Σ is also known as a completeness condition and Assumption 2.3.2 can be understood as its weak version; see Babii and Florens (2020) and references therein. It is worth emphasizing that $\gamma > 0$ in Assumption 2.3.2 is a universal constant independent of p, which is the case, e.g., when Σ is a Toeplitz matrix or a spiked identity matrix. Alternatively, we could allow for $\gamma \downarrow 0$ as $p \to \infty$, in which case the term γ^{-1} would appear in our nonasymptotic bounds slowing down the speed of convergence, and we may interpret γ as a measure of ill-posedness in the spirit of econometrics literature on ill-posed inverse problems; see Carrasco, Florens, and Renault (2007).

The value of the regularization parameter is determined by the Fuk-Nagaev concentration inequality, appearing in the Appendix, see Theorem A.2.1.

Assumption 2.3.3 (Regularization)**.** *For some* $\delta \in (0,1)$

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

where $\kappa = ((a+1)\varsigma - 1)/(a+\varsigma-1)$ *and* a, ς *are as in Assumption 2.3.1.*

The regularization parameter in Assumption 2.3.3 is determined by the persistence of the data, quantified by a, and the tails, quantified by $\varsigma = qr/(q+r)$. This dependence is reflected in the *dependence-tails exponent* κ. The following result describes the nonasymptotic prediction and estimation bounds for the sg-LASSO estimator, see Appendix, Section A.2.2 for the proof.

Theorem 2.1. *Suppose that Assumptions 2.3.1, 2.3.2, and 2.3.3 are satisfied. Then with probability at least* $1-\delta-O(p^2(T^{1-\mu}s_\alpha^\mu + \exp(-cT/s_\alpha^2)))$

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

and

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

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

Theorem 2.1 provides nonasymptotic guarantees for the estimation and prediction with the sg-LASSO estimator reflecting potential misspecification.

In the special case of the LASSO estimator ($\alpha = 1$), we obtain the counterpart to the result of Belloni, Chen, Chernozhukov, and Hansen (2012) for the LASSO estimator with i.i.d. data taking into account that we may have $m_t \neq x_t^\top \beta$. At another extreme, when $\alpha = 0$, we obtain the nonasymptotic bounds for the group LASSO allowing for misspecification which to the best of our knowledge are new, cf. Negahban, Ravikumar, Wainwright, and Yu (2012) and van de Geer (2016). We call s_α the *effective sparsity constant*. This constant reflects the benefits of the sparse-group structure for the sg-LASSO estimator that can not be deduced from the results currently available for the LASSO or the group LASSO.

Remark 2.3.1. *Since the ℓ_1-norm is equivalent to the Ω-norm whenever groups have fixed size, we deduce from Theorem 2.1 that*

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

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

Assumption 2.3.4. *(i) $\|\mathbf{m} - \mathbf{X}\beta\|_T^2 = O_P\left(s_\alpha \lambda^2\right)$; and (ii) $p^2 T^{1-\mu} s_\alpha^\mu \to 0$ and $p^2 \exp(-cT/s_\alpha^2) \to 0$.*

The following corollary is an immediate consequence of Theorem 2.1.

Corollary 2.3.1. *Suppose that Assumptions 2.3.1, 2.3.2, 2.3.3, and 2.3.4 hold. Then*

$$\|\mathbf{X}(\hat{\beta} - \beta)\|_T^2 = O_P\left(\frac{s_\alpha p^{2/\kappa}}{T^{2-2/\kappa}} \vee \frac{s_\alpha \log p}{T}\right)$$

and

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

If the effective sparsity constant s_α is fixed, then $p = o(T^{\kappa-1})$ is a sufficient condition for the prediction and estimation errors to vanish, whenever $\mu \geq 2\kappa - 1$. In this case Assumption 2.3.4 (ii) is vacuous. More generally, s_α is allowed to increase slowly with the sample size. Convergence rates in Corollary 2.3.1 quantify the effect of tails and persistence of the data on the prediction and estimation accuracies of the sg-LASSO estimator. In particular, lighter tails and less persistence allow us to handle a larger number of covariates p compared to the sample size T. In particular p can increase faster than T, provided that $\kappa > 2$.

Remark 2.3.2. *In the special case of the LASSO estimator with i.i.d. data, Corollary 4 of Fuk and Nagaev (1971) leads to the convergence rate of order* $O_P\left(\frac{p^{1/\varsigma}}{T^{1-1/\varsigma}} \vee \sqrt{\frac{\log p}{T}}\right)$. *If the* τ*-mixing coefficients decrease geometrically fast (e.g., stationary AR(p)), then* $\kappa \approx \varsigma$ *for a sufficiently large value of the dependence exponent a, in which case the convergence rates in Corollary 2.3.1 are close to the i.i.d. case. In this sense these rates depend sharply on the tails exponent* ς*, and we can conclude that for geometrically decreasing* τ*-mixing coefficients, the persistence of the data should not affect the convergence rates of the LASSO.*

Remark 2.3.3. *In the special case of the LASSO estimator, if* $(u_t)_{t\in\mathbf{Z}}$ *and* $(x_t)_{t\in\mathbf{Z}}$ *are causal Bernoulli shifts with independent innovations and at least* $q = r \geq 8$ *finite moments, one can deduce from Chernozhukov, Härdle, Huang, and Wang (2021), Lemma 5.1 and Corollary 5.1, the convergence rate of order* $O_P\left(\frac{(p\omega_T)^{1/\varsigma}}{T^{1-1/\varsigma}} \vee \sqrt{\frac{\log p}{T}}\right)$, *where* $\omega_T = 1$ *(weakly dependent case) or* $\omega_T = T^{\varsigma/2-1-a\varsigma} \uparrow \infty$ *(strongly dependent case), provided that the physical dependence coefficients are of size* $O(k^{-a})$. *Note that for causal Bernoulli shifts with independent innovations, the physical dependence coefficients are not directly comparable to* τ*-mixing coefficients; see Dedecker, Doukhan, Lang, Rafael, Louhichi, and Prieur (2007), Remark 3.1 on p.32.*

2.4 Monte Carlo experiments

We assess via simulations the out-of-sample predictive performance (forecasting and nowcasting), and the MIDAS weights recovery of the sg-LASSO with dictionaries. We benchmark the performance of our novel sg-LASSO setup against two alternatives: (a) unstructured, meaning standard, LASSO with MIDAS, and (b) unstructured LASSO with the unrestricted lag polynomial. The former allows us to assess the benefits of exploiting group structures, whereas the latter focuses on the advantages of using dictionaries in a high-dimensional setting.

2.4.1 Simulation Design

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

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

where $u_t \sim_{i.i.d.} N(0, \sigma_u^2)$ and the DGP for covariates $\{x_{k,t-(j-1)/m} : j \in [m], k \in [K]\}$ is specified below. This corresponds to a target of interest y_t driven by two autoregressive lags augmented with high frequency series, hence, the DGP is an ARDL-MIDAS model. We set $\sigma_u^2 = 1$, $\rho_1 = 0.3$, $\rho_2 = 0.01$, and take the number of relevant high frequency regressors $K = 3$. In some scenarios we also decrease the signal-to-noise ratio by setting $\sigma_u^2 = 5$. We are interested in quarterly/monthly data, and use four quarters of data for the high frequency regressors so that $m = 12$. We rely on a commonly used weighting scheme in the MIDAS literature, namely $\omega(s; \beta_k)$ for $k = 1$, 2 and 3 are determined by beta densities respectively equal to $\text{Beta}(1,3), \text{Beta}(2,3)$, and $\text{Beta}(2,2)$; see Ghysels, Sinko, and Valkanov (2007) or Ghysels and Qian (2019), for further details.

The high frequency regressors are generated as either one of the following:

1. K i.i.d. realizations of the univariate autoregressive (AR) process $x_h = \rho x_{h-1} + \varepsilon_h$, where $\rho = 0.2$ or $\rho = 0.7$ and either $\varepsilon_h \sim_{i.i.d.} N(0, \sigma_\varepsilon^2)$, $\sigma_\varepsilon^2 = 1$, or $\varepsilon_h \sim_{i.i.d.}$ student-$t(5)$, where h denotes the high-frequency sampling.
2. Multivariate vector autoregressive (VAR) process $X_h = \Phi X_{h-1} + \varepsilon_h$, where $\varepsilon_h \sim_{i.i.d.} N(0, I_K)$ and Φ is a block diagonal matrix described below.

For the AR simulation design, we initiate the processes as

$$x_0 \sim N\left(0, \sigma^2/(1-\rho^2)\right)$$

and

$$y_0 \sim N\left(0, \sigma^2(1-\rho_2)/((1+\rho_2)((1-\rho_2)^2 - \rho_1^2))\right).$$

For the VAR, the initial value of (y_t) is the same, while $X_0 \sim N(0, I_K)$. In all cases, the first 200 observations are treated as burn-in. In the estimation procedure, we add 7 noisy covariates which are generated in the same way as the relevant covariates and use 5 low-frequency lags. The empirical models use a dictionary which consists of Legendre polynomials up to degree $L = 10$ shifted to the $[0, 1]$ interval with the MIDAS weight function approximated

as in equation (2.1). The sample size is $T \in \{50, 100, 200\}$, and for all the experiments we use 5000 simulation replications.

We assess the performance of different methods by modifying the assumptions on the error terms of the high-frequency process ε_h, considering multivariate high-frequency processes, changing the degree of Legendre polynomials L, increasing the noise level of the low-frequency process σ_u^2, using only half of the high-frequency lags in predictive regressions, and adding a larger number of noisy covariates. In the case of VAR high-frequency process, we set Φ to be block-diagonal with the first 5×5 block having entries 0.15 and the remaining 5×5 block(s) having entries 0.075.

We estimate three different LASSO-type regression models. In the first model, we keep the weighting function unconstrained, and therefore we estimate 12 coefficients per high-frequency covariate using the unstructured LASSO estimator. We denote this model LASSO-U-MIDAS (inspired by the U-MIDAS of Foroni, Marcellino, and Schumacher (2015a)). In the second model we use MIDAS weights together with the unstructured LASSO estimator; we call this model LASSO-MIDAS. In this case, we estimate $L + 1$ number of coefficients per high-frequency covariate. The third model applies the sg-LASSO estimator together with MIDAS weights. Groups are defined as in Section 2.2; each low-frequency lag and high-frequency covariate is a group, therefore, we have $K + 5$ groups. We select the value of tuning parameters λ and α using the 5-fold cross-validation, defining folds as adjacent blocks over the time dimension to take into account the time series dependence. This model is denoted sg-LASSO-MIDAS.

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

2.4.2 Simulation results

Detailed results are reported in the Appendix. Tables A.2.1–A.2.2, cover the average mean squared forecast errors for one-step-ahead forecasts and nowcasts. The sg-LASSO with MIDAS weighting (sg-LASSO-MIDAS) outperforms all other methods in all simulation scenarios. Importantly,

both sg-LASSO-MIDAS and unstructured LASSO-MIDAS with nonlinear weight function approximations perform much better than all other methods when the sample size is small ($T = 50$). In this case, sg-LASSO-MIDAS yields the largest improvements over alternatives, in particular, with a large number of noisy covariates (bottom-right block). These findings are robust to increases in the persistence parameter of covariates ρ from 0.2 to 0.7. The LASSO without MIDAS weighting has typically large forecast errors. Comparing across simulation scenarios, all methods seem to perform worse with heavy-tailed or persistent covariates. In these cases, however, the impact on the sg-LASSO-MIDAS method is lesser compared to the other methods. This simulation evidence supports our theoretical results and findings in the empirical application. Lastly, forecasts using flow-aggregated covariates seem to perform better than other simple aggregation methods in all simulation scenarios, but significantly worse than the sg-LASSO-MIDAS.

In Table A.2.3–A.2.4 we report additional results for the estimation accuracy of the weight functions. In figure A.2.1–A.2.3, we plot the estimated weight functions from several methods. The results indicate that the LASSO without MIDAS weighting can not accurately recover the weights in small samples and/or low signal-to-noise ratio scenarios. Using Legendre polynomials improves the performance substantially and the sg-LASSO seems to improve even more over the unstructured LASSO.

2.5 Nowcasting US GDP with macro, financial and textual news data

We nowcast US GDP with macroeconomic, financial, and textual news data. Details regarding the data sources appear in the Appendix Section A.2.5. Regarding the macro data, we rely on 34 series used in the Federal Reserve Bank of New York nowcast model, discarding two series ("PPI: Final demand" and "Merchant wholesalers: Inventories") due to very short samples; see Bok, Caratelli, Giannone, Sbordone, and Tambalotti (2018) for more details regarding this data.

For all macro data, we use real-time vintages, which effectively means that we take all macro series with a delay as well real-time data releases. For example, if we nowcast the first quarter of GDP one month before the quarter ends, we use data up to the end of February, and therefore all macro series with a delay of one month that enter the model are available up to the end of January. As we use data real-time data releases, the January

observation in this case is also the first release of a particular series. We use Legendre polynomials of degree three for all macro covariates to aggregate twelve lags of monthly macro data. In particular, let $x_{t+(h+1-j)/m,k}$ be k^{th} covariate at quarter t with $m = 3$ months per quarter and $h = 2 - 1 = 1$ months into the quarter (2 months into the quarter minus 1 month due to publication delay), where $j = 1, 2, \ldots, 12$ is the monthly lag. We then collect all lags in a vector

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

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

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

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

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

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

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

Following Section 2.2, the high-frequency lag polynomial is defined as

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

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

We benchmark our predictions against the simple AR(1) model, which is considered to be a reasonable starting point for short-term GDP growth predictions. We focus on predictions of our method, sg-LASSO-MIDAS, with and without financial data combined with series based on the textual analysis. One natural comparison is with the publicly available Federal Reserve Bank of New York, denoted NY Fed, model implied nowcasts. We

Table 2.1: Nowcast comparisons for models with macro data only – Nowcast horizons are 2- and 1-month ahead, as well as the end of the quarter. Column *Rel-RMSE* reports root mean squared forecasts error relative to the AR(1) model. Column *DM-stat-1* reports Diebold and Mariano (1995) test statistic of all models relative to NY Fed nowcasts, while column *DM-stat-2* reports the Diebold Mariano test statistic relative to sg-LASSO-MIDAS model. The last row reports the p-value of the average Superior Predictive Ability (aSPA) test, see Quaedvlieg (2019), over the three horizons of sg-LASSO-MIDAS model compared to the NY Fed nowcasts. Out-of-sample period: 2002 Q1 to 2017 Q2.

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

adopt the following strategy. First, we focus on the same series that are used to calculate the NY Fed nowcasts. The purpose here is to compare

models since the data inputs are the same. This means that we compare the performance of dynamic factor models (NY Fed) with that of machine learning regularized regression methods (sg-LASSO-MIDAS). Next, we expand the data set to see whether additional financial and textual news series can improve the nowcast performance.

In Table 2.1, we report results based on real-time macro data used for the NY Fed model, see Bok, Caratelli, Giannone, Sbordone, and Tambalotti (2018). The results show that the sg-LASSO-MIDAS performs much better than the NY Fed nowcasts at the longer, i.e. 2-month, horizon. Our method significantly beats the benchmark AR(1) model for all the horizons, and the accuracy of the nowcasts improve with the horizon. Our end-of-quarter and 1-month horizon nowcasts are similar to the NY Fed ones, with the sg-LASSO-MIDAS being slightly better numerically but not statistically. We also report the average Superior Predictive Ability test of Quaedvlieg (2019) over all three horizons and the result reveals that the improvement of the sg-LASSO-MIDAS model versus the NY Fed nowcasts is significant at the 5% significance level. Lastly, we report results that do not discard two series ("PPI: Final demand" and "Merchant wholesalers: Inventories") due to short samples in the Appendix Section A.2.5.1. The results are very similar and do not change our conclusions.

The comparison in Table 2.1 does not fully exploit the potential of our methods, as it is easy to expand the data series beyond the small number used by the NY Fed nowcasting model. In Table 2.2 we report results with additional sets of covariates which are financial series, advocated by Andreou, Ghysels, and Kourtellos (2013), and textual analysis of news. In total, the models select from 118 series – 34 macro, 8 financial, and 76 news attention series. For the moment we focus only on the first three columns of the table. At the longer horizon of 2 months, the method seems to produce slightly worse nowcasts compared to the results reported in Table 2.1 using only macro data. However, we find significant improvements in prediction quality for the shorter 1-month and end-of-quarter horizons. In particular, a significant increase in accuracy relative to NY Fed nowcasts appears at the 1-month horizon. We report again the average Superior Predictive Ability test of Quaedvlieg (2019) over all three horizons with the same result that the improvement of sg-LASSO-MIDAS versus the NY Fed nowcasts is significant at the 5% significance level. Lastly, we report results for several alternatives, namely, PCA-OLS, ridge, LASSO, and Elastic Net, using the unrestricted MIDAS scheme. Our approach produces more accurate nowcasts compared to these alternatives.

Table 2.2: Nowcast comparison table – Nowcast horizons are 2- and 1-month ahead, as well as the end of the quarter. Column *Rel-RMSE* reports root mean squared forecasts error relative to the AR(1) model. Column *DM-stat-1* reports Diebold and Mariano (1995) test statistic of all models relative to the NY FED nowcast, while column *DM-stat-2* reports the Diebold Mariano test statistic relative to the sg-LASSO model. Columns *DM-stat-3* and *DM-stat-4* report the Diebold Mariano test statistic for the same models, but excludes the recession period. For the 1-month horizon, the last row *SPF (median)* reports test statistics for the same models comparing with the SPF median nowcasts. The last row reports the p-value of the average Superior Predictive Ability (aSPA) test, see Quaedvlieg (2019), over the three horizons of sg-LASSO-MIDAS model compared to the NY Fed nowcasts, including (left) and excluding (right) financial crisis period. Out-of-sample period: 2002 Q1 to 2017 Q2.

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

The inclusion of financial series is not common in traditional nowcasting models, see e.g. Bok, Caratelli, Giannone, Sbordone, and Tambalotti (2018), on the grounds that though timely, financial data is noisy hence do not contribute to the accuracy of the nowcasts. One may wonder how our model performs excluding these series. Therefore, we run our nowcasting regressions using only macro and news attention series, excluding financial data; results are reported in the Appendix Section A.2.5.1. Notably, results are slightly worse compared with the results that include financial data, supporting our initial choice. Similarly, Andreou, Ghysels, and Kourtellos (2013) find that financial data is helpful in GDP nowcasting applications.

Table 2.2 also features an entry called SPF (median), where we report results for the median survey of professional nowcasts for the 1-month horizon, and analyze how the model-based nowcasts compare with the predictions using the publicly available Survey of Professional Forecasters maintained by the Federal Reserve Bank of Philadelphia. We find that the sg-LASSO-MIDAS model-based nowcasts are similar to the SPF-implied nowcasts. We also find that the NY Fed nowcasts are significantly worse than the SPF.

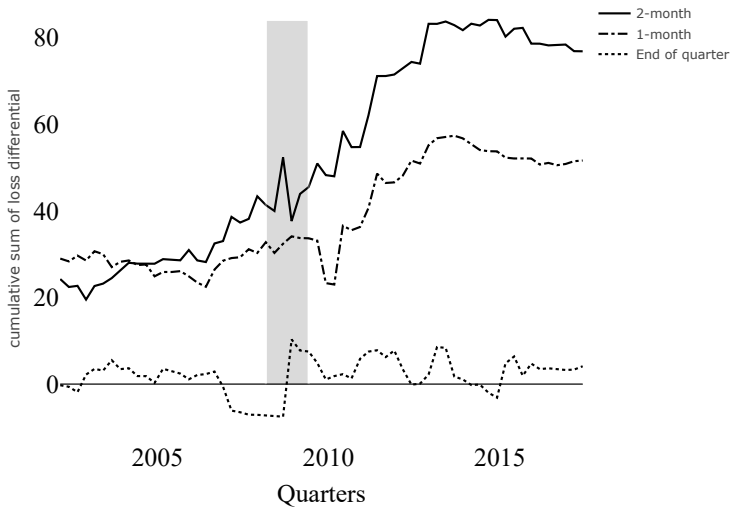


Figure 2.2: Cumulative sum of loss differentials of sg-LASSO-MIDAS model nowcasts including financial and textual data compared with the New York Fed model for three nowcasting horizons: solid black line cumsfe for the 2-months horizon, dash-dotted black line - cumsfe for the 1-month horizon, and dotted line for the end-of-quarter nowcasts. The gray shaded area corresponds to the NBER recession period.

In figure 2.2 we plot the cumulative sum of squared forecast error (CUMSFE) loss differential of sg-LASSO-MIDAS versus NY Fed nowcasts

for the three horizons. The CUMSFE is computed as

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

for model $M1$ versus $M2$. A positive value of $\text{CUMSFE}_{t,t+k}$ means that the model $M1$ has larger squared forecast errors compared to model $M2$ up to $t+k$, and negative values imply the opposite. In our case, $M1$ is the New York Fed prediction error, while $M2$ is the sg-LASSO-MIDAS model. We observe persistent gains for the 2- and 1-month horizons throughout the out-of-sample period. When comparing the sg-LASSO-MIDAS results with additional financial and textual news series versus those based on macro data only, we see a notable improvement at the 1-month horizon and a more modest one at the end-of-quarter horizons. In figuree 2.3, we plot the average CUMSFE for the 1-month and end-of-quarter horizons and observe that the largest gains of additional financial and textual news data are achieved during the financial crisis.

The result in figure 2.3 prompts the question whether our results are mostly driven by this unusual period in our out-of-sample data. To assess this, we turn our attention again to the last two columns of Table 2.2 reporting Diebold and Mariano (1995) test statistics which exclude the financial crisis period. Compared to the tests previously discussed, we find that the results largely remain the same, but some alternatives seem to slightly improve (e.g. LASSO or Elastic Net). Note that this also implies that our method performs better during periods with heavy-tailed observations, such as the financial crisis. It should also be noted that overall there is a slight deterioration of the average Superior Predictive Ability test over all three horizons when we remove the financial crisis.

In figure 2.4, we plot the fraction of selected covariates by the sg-LASSO-MIDAS model when we use the macro, financial, and textual analysis data. For each reference quarter, we compute the ratio of each group of variables relative to the total number of covariates. In each subfigure, we plot the three different horizons. For all horizons, the macro series are selected more often than financial and/or textual data. The number of selected series increases with the horizon, however, the pattern of denser macro series and sparser financial and textual series is visible for all three horizons. The results are in line with the literature – macro series tend to co-move, hence we see a denser pattern in the selection of such series, see e.g. Bok, Caratelli, Giannone, Sbordone, and Tambalotti (2018). On the other hand,

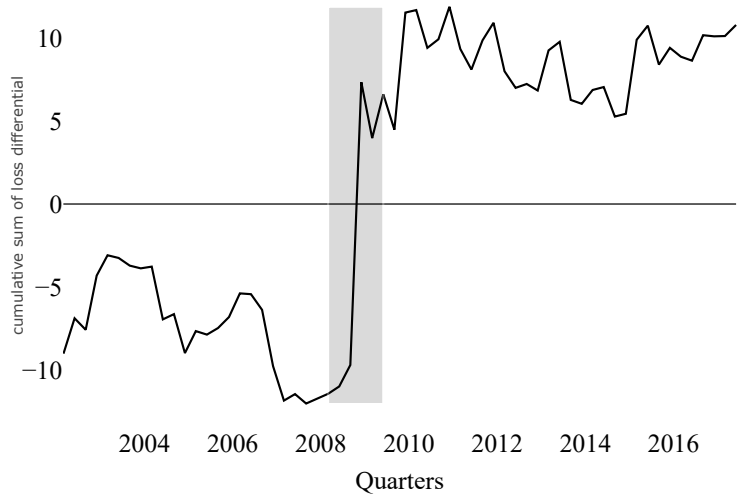


Figure 2.3: Cumulative sum of loss differentials (CUMSFE) of sg-LASSO-MIDAS nowcasts when we include vs. when we exclude the additional financial and textual news data, averaged over 1-month and the end-of-quarter horizons. The gray shaded area corresponds to the NBER recession period.

the alternative textual analysis data appear to be very sparse, yet still important for nowcasting accuracy, see also Thorsrud (2020).

2.6 Conclusion

This paper offers a new perspective on the high-dimensional time series regressions with data sampled at the same or mixed frequencies and contributes more broadly to the rapidly growing literature on the estimation, inference, forecasting, and nowcasting with regularized machine learning methods. The first contribution of the paper is to introduce the sparse-group LASSO estimator for high-dimensional time series regressions. An attractive feature of the estimator is that it recognizes time series data structures and allows us to perform the hierarchical model selection within and between groups. The classical LASSO and the group LASSO are covered as special cases.

To recognize that the economic and financial time series have typically heavier than Gaussian tails, we use a new Fuk-Nagaev concentration inequality, from Babii, Ghysels, and Striaukas (2020a), valid for a large class of τ-mixing processes, including α-mixing processes commonly used in econometrics. Building on this inequality, we establish the nonasymptotic and asymptotic properties of the sparse-group LASSO estimator.

Our empirical application provides new perspectives on applying machine

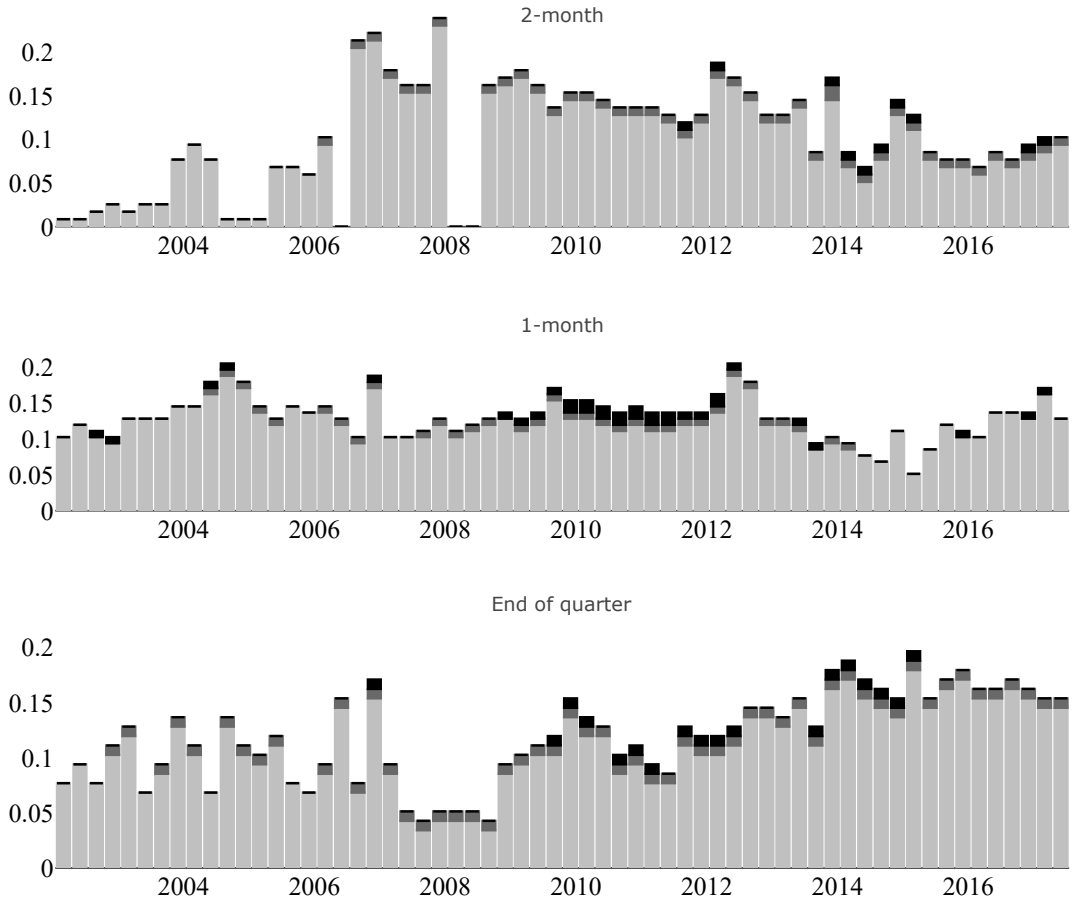


Figure 2.4: The fraction of selected covariates attributed to macro (light gray), financial (dark gray), and textual (black) data for three monthly horizons.

learning methods to real-time forecasting, nowcasting, and monitoring with time series data, including unconventional data, sampled at different frequencies. To that end, we introduce a new class of MIDAS regressions with dictionaries linear in the parameters and based on orthogonal polynomials with lag selection performed by the sg-LASSO estimator. We find that the sg-LASSO outperforms the unstructured LASSO in small samples and conclude that incorporating specific data structures should be helpful in various applications.

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

significant improvements in prediction quality for the shorter 1-month and end-of-quarter horizons.

APPENDIX

A2.1 Dictionaries

In this section, we review the choice of dictionaries for the MIDAS weight function. It is possible to construct dictionaries using arbitrary sets of functions, including a mix of algebraic polynomials, trigonometric polynomials, B-splines, Haar basis, or wavelets. In this paper, we mostly focus on dictionaries generated by orthogonalized algebraic polynomials, though it might be interesting to tailor the dictionary for each particular application. The attractiveness of algebraic polynomials comes from their ability to generate a variety of shapes with a relatively low number of parameters, which is especially desirable in low signal-to-noise environments. The general family of appropriate orthogonal algebraic polynomials is given by Jacobi polynomials that nest Legendre, Gegenbauer, and Chebychev's polynomials as a special case.

Example A2.1.1 (Jacobi polynomials)**.** *Applying the Gram-Schmidt orthogonalization to the power polynomials* $\{1, x, x^2, x^3, \dots\}$ *with respect to the measure*

$$\mathrm{d}\mu(x) = (1-x)^\alpha (1+x)^\beta \mathrm{d}x, \qquad \alpha, \beta > -1,$$

on $[-1, 1]$*, we obtain Jacobi polynomials. In practice Jacobi polynomials can be computed through the well-known tree-term recurrence relation for* $n = 1, 2, \dots$

$$P_{n+1}^{(\alpha,\beta)}(x) = axP_n^{(\alpha,\beta)}(x) + bP_n^{(\alpha,\beta)}(x) - cP_{n-1}^{(\alpha,\beta)}(x)$$

with $a = (2n+\alpha+\beta+1)(2n+\alpha+\beta+2)/2(n+1)(n+\alpha+\beta+1)$*,* $b = (2n+\alpha+\beta+1)(\alpha^2-\beta^2)/2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)$*, and* $c = (\alpha+n)(\beta+n)(2n+\alpha+\beta+2)/(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)$*. To obtain the orthogonal basis on* $[0, 1]$*, we shift Jacobi polynomials with affine bijection* $x \mapsto 2x - 1$*.*

For $\alpha = \beta$*, we obtain Gegenbauer polynomials, for* $\alpha = \beta = 0$*, we obtain Legendre polynomials, while for* $\alpha = \beta = -1/2$ *or* $\alpha = \beta = 1/2$*, we obtain Chebychev's polynomials of two kinds.*

In the mixed frequency setting, non-orthogonalized polynomials, $\{1, x, x^2, x^3, \dots\}$, are also called Almon polynomials. It is preferable to use orthogonal polynomials in practice due to reduced multicollinearity and better numerical

properties. At the same time, orthogonal polynomials are available in Matlab, R, Python, and Julia packages. Legendre polynomials is our default recommendation, while other choices of α and β are preferable if we want to accommodate MIDAS weights with other integrability/tail properties.

We noted in the main body of the paper that the specification in equation (2) deviates from the standard MIDAS polynomial specification as it results in a linear regression model - a subtle but key innovation as it maps MIDAS regressions in the standard regression framework. Moreover, casting the MIDAS regressions in a linear regression framework renders the optimization problem convex, something only achieved by Siliverstovs (2017) using the U-MIDAS of Foroni, Marcellino, and Schumacher (2015b) which does not recognize the mixed frequency data structure, unlike our sg-LASSO.

A2.2 Proofs of main results

Lemma A2.2.1. *Consider* $\|.\| = \alpha|.|_1 + (1-\alpha)|.|_2$, *where* $|.|_q$ *is* ℓ_q *norm on* $\mathbf{R}^p$. *Then the dual norm of* $\|.\|$, *denoted* $\|.\|^*$, *satisfies*

$$\|z\|^* \leq \alpha|z|_1^* + (1-\alpha)|z|_2^*, \qquad \forall z \in \mathbf{R}^p,$$

where $|.|_1^*$ *is the dual norm of* $|.|_1$ *and* $|.|_2^*$ *is the dual norm of* $|.|_2$.

Proof. Clearly, $\|.\|$ is a norm. By the convexity of $x \mapsto x^{-1}$ on $(0, \infty)$

$$\begin{aligned}\|z\|^* &= \sup_{b\neq 0} \frac{|\langle z, b\rangle|}{\|b\|} \leq \sup_{b\neq 0} \left\{ \alpha \frac{|\langle z, b\rangle|}{|b|_1} + (1-\alpha)\frac{|\langle z, b\rangle|}{|b|_2} \right\} \\ &\leq \alpha \sup_{b\neq 0} \frac{|\langle z, b\rangle|}{|b|_1} + (1-\alpha) \sup_{b\neq 0} \frac{|\langle z, b\rangle|}{|b|_2} \\ &= \alpha|z|_1^* + (1-\alpha)|z|_2^*.\end{aligned}$$

□

Proof of Theorem 3.1. By Hölder's inequality for every $\varsigma > 0$

$$\max_{j\in[p]} \|u_0 x_{0,j}\|_\varsigma \leq \|u_0\|_{\varsigma q_1} \max_{j\in[p]} \|x_{0,j}\|_{\varsigma q_2}$$

with $q_1^{-1} + q_2^{-1} = 1$ and $q_1, q_2 \geq 1$. Therefore, under Assumption 3.1 (i), $\max_{j\in[p]} \|u_0 x_{0,j}\|_\varsigma = O(1)$ with $\varsigma = qr/(q+r)$. Recall also that $\mathbb{E}[u_t x_{t,j}] =$

$0, \forall j \in [p]$, see equation (3), which in conjunction with Assumption 3.1 (ii) verifies conditions of Theorem A2.1 and shows that there exists $C > 0$ such that for every $\delta \in (0,1)$

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T} u_t X_t\right|_\infty \leq C\left(\frac{p}{\delta T^{\kappa-1}}\right)^{1/\kappa} \vee \sqrt{\frac{\log(8p/\delta)}{T}}\right) \geq 1-\delta. \tag{A2.1}$$

Let $G^* = \max_{G\in\mathcal{G}} |G|$ be the size of the largest group in $\mathcal{G}$. Note that the sg-LASSO penalty Ω is a norm. By Lemma A2.2.1, its dual norm satisfies

$$\begin{aligned}
\Omega^*(\mathbf{X}^\top \mathbf{u}/T) &\leq \alpha|\mathbf{X}^\top \mathbf{u}/T|_\infty + (1-\alpha)\max_{G\in\mathcal{G}} |(\mathbf{X}^\top \mathbf{u})_G/T|_2 \\
&\leq (\alpha + (1-\alpha)\sqrt{G^*})|\mathbf{X}^\top \mathbf{u}/T|_\infty \\
&\leq (\alpha + (1-\alpha)\sqrt{G^*})C\left(\frac{p}{\delta T^{\kappa-1}}\right)^{1/\kappa} \vee \sqrt{\frac{\log(8p/\delta)}{T}} \\
&\leq \lambda/c,
\end{aligned} \tag{A2.2}$$

where the first inequality follows since $|z|_1^* = |z|_\infty$ and $\left(\sum_{G\in\mathcal{G}} |z_G|_2\right)^* = \max_{G\in\mathcal{G}} |z_G|_2$, the second by elementary computations, the third by equation ((A2.1)) with probability at least $1-\delta$ for every $\delta \in (0,1)$, and the last from the definition of λ in Assumption 3.3, where $c > 1$ is as in Assumption 3.2. By Fermat's rule, the sg-LASSO satisfies

$$\mathbf{X}^\top(\mathbf{X}\hat{\beta} - \mathbf{y})/T + \lambda z^* = 0$$

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

$$\langle \mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\beta}), \beta - \hat{\beta}\rangle_T = \lambda\langle z^*, \beta - \hat{\beta}\rangle \leq \lambda\left\{\Omega(\beta) - \Omega(\hat{\beta})\right\},$$

where the inequality follows from the definition of the subdifferential. Using $\mathbf{y} = \mathbf{m} + \mathbf{u}$ and rearranging this inequality

$$\begin{aligned}
\|\mathbf{X}(\hat{\beta} - \beta)\|_T^2 - \lambda\left\{\Omega(\beta) - \Omega(\hat{\beta})\right\} &\leq \langle \mathbf{X}^\top \mathbf{u}, \hat{\beta} - \beta\rangle_T + \langle \mathbf{X}^\top(\mathbf{m} - \mathbf{X}\beta), \hat{\beta} - \beta\rangle_T \\
&\leq \Omega^*\left(\mathbf{X}^\top \mathbf{u}/T\right)\Omega(\hat{\beta} - \beta) + \|\mathbf{X}(\hat{\beta} - \beta)\|_T \|\mathbf{m} - \mathbf{X}\beta\|_T \\
&\leq c^{-1}\lambda\Omega(\hat{\beta} - \beta) + \|\mathbf{X}(\hat{\beta} - \beta)\|_T \|\mathbf{m} - \mathbf{X}\beta\|_T.
\end{aligned}$$

where the second line follows by the dual norm inequality and the last by $\Omega^*(\mathbf{X}^\top \mathbf{u}/T) \leq \lambda/c$ as shown in equation ((A2.2)). Therefore,

$$\begin{aligned}
\|\mathbf{X}\Delta\|_T^2 &\leq c^{-1}\lambda\Omega(\Delta) + \|\mathbf{X}\Delta\|_T \|\mathbf{m} - \mathbf{X}\beta\|_T + \lambda\left\{\Omega(\beta) - \Omega(\hat{\beta})\right\} \\
&\leq (c^{-1} + 1)\lambda\Omega(\Delta) + \|\mathbf{X}\Delta\|_T \|\mathbf{m} - \mathbf{X}\beta\|_T
\end{aligned} \tag{A2.3}$$

with $\Delta = \hat{\beta} - \beta$. Note that the sg-LASSO penalty can be decomposed as a sum of two seminorms $\Omega(b) = \Omega_0(b) + \Omega_1(b),\ \forall b \in \mathbf{R}^p$ with

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

Note also that $\Omega_1(\beta) = 0$ and $\Omega_1(\hat{\beta}) = \Omega_1(\Delta)$. Then by the triangle inequality

$$\Omega(\beta) - \Omega(\hat{\beta}) \leq \Omega_0(\Delta) - \Omega_1(\Delta). \tag{A2.4}$$

If $\|\mathbf{m} - \mathbf{X}\beta\|_T \leq 2^{-1}\|\mathbf{X}\Delta\|_T$, then it follows from the first inequality in equation ((A2.3)) and equation ((A2.4)) that

$$\|\mathbf{X}\Delta\|_T^2 \leq 2c^{-1}\lambda\Omega(\Delta) + 2\lambda\left\{\Omega_0(\Delta) - \Omega_1(\Delta)\right\}.$$

Since the left side of this equation is positive, this shows that $\Omega_1(\Delta) \leq c_0\Omega_0(\Delta)$ with $c_0 = (c+1)/(c-1)$, and whence $\Delta \in \mathcal{C}(c_0)$, cf., Assumption 3.2. Then

$$\begin{aligned}
\Omega(\Delta) &\leq (1+c_0)\Omega_0(\Delta) \\
&\leq (1+c_0)\left(\alpha\sqrt{|S_0|}|\Delta_{S_0}|_2 + (1-\alpha)\sqrt{|\mathcal{G}_0|}\sqrt{\sum_{G\in\mathcal{G}_0}|\Delta_G|_2^2}\right) \\
&\leq (1+c_0)\sqrt{s_\alpha}\sqrt{\sum_{G\in\mathcal{G}_0}|\Delta_G|_2^2} \\
&\leq (1+c_0)\sqrt{s_\alpha/\gamma\Delta^\top\Sigma\Delta},
\end{aligned} \tag{A2.5}$$

where we use the Jensen's inequality, Assumption 3.2, and the definition of $\sqrt{s_\alpha}$. Next, note that

$$\begin{aligned}
\Delta^\top\Sigma\Delta &= \|\mathbf{X}\Delta\|_T^2 + \Delta^\top(\Sigma - \hat{\Sigma})\Delta \\
&\leq 2(c^{-1}+1)\lambda\Omega(\Delta) + \Omega(\Delta)\Omega^*\left((\hat{\Sigma} - \Sigma)\Delta\right) \\
&\leq 2(c^{-1}+1)\lambda\Omega(\Delta) + \Omega^2(\Delta)G^*|\mathrm{vech}(\hat{\Sigma} - \Sigma)|_\infty,
\end{aligned} \tag{A2.6}$$

where the first inequality follows from equation ((A2.3)) and the dual norm inequality and the second by Lemma A2.2.1 and elementary computations

$$\begin{aligned}
\Omega^*\left((\hat{\Sigma} - \Sigma)\Delta\right) &\leq \alpha|(\hat{\Sigma} - \Sigma)\Delta|_\infty + (1-\alpha)\max_{G\in\mathcal{G}}\left|[(\hat{\Sigma} - \Sigma)\Delta]_G\right|_2 \\
&\leq \alpha|\Delta|_1|\mathrm{vech}(\hat{\Sigma} - \Sigma)|_\infty + (1-\alpha)\sqrt{G^*}|\mathrm{vech}(\hat{\Sigma} - \Sigma)|_\infty|\Delta|_1 \\
&\leq G^*|\mathrm{vech}(\hat{\Sigma} - \Sigma)|_\infty\Omega(\Delta).
\end{aligned}$$

Combining the inequalities obtained in equations (A2.5) and (A2.6)

$$\begin{aligned}\Omega(\Delta) &\leq (1+c_0)^2\gamma^{-1}s_\alpha\left\{2(c^{-1}+1)\lambda + G^*|\text{vech}(\hat{\Sigma}-\Sigma)|_\infty\Omega(\Delta)\right\} \\ &\leq 2(1+c_0)^2\gamma^{-1}s_\alpha(c^{-1}+1)\lambda + (1-A^{-1})\Omega(\Delta),\end{aligned} \tag{A2.7}$$

where the second line holds on the event $E \triangleq \{|\text{vech}(\hat{\Sigma}-\Sigma)|_\infty \leq \gamma/2G^*s_\alpha(1+2c_0)^2\}$ with $1-A^{-1} = (1+c_0)^2/2(1+2c_0)^2 < 1$. Therefore, inequalities in equation ((A2.3) and (A2.7)) yield

$$\begin{aligned}\Omega(\Delta) &\leq \frac{2A}{\gamma}(1+c_0)^2(c^{-1}+1)s_\alpha\lambda \\ \|\mathbf{X}\Delta\|_T^2 &\leq \frac{4A}{\gamma}(1+c_0)^2(c^{-1}+1)^2s_\alpha\lambda^2.\end{aligned}$$

On the other hand, if $\|\mathbf{m}-\mathbf{X}\beta\|_T > 2^{-1}\|\mathbf{X}\Delta\|_T$, then

$$\|\mathbf{X}\Delta\|_T^2 \leq 4\|\mathbf{m}-\mathbf{X}\beta\|_T^2.$$

Therefore, on the event E we always have

$$\|\mathbf{X}\Delta\|_T^2 \leq C_1s_\alpha\lambda^2 + 4\|\mathbf{m}-\mathbf{X}\beta\|_T^2 \tag{A2.8}$$

with $C_1 = 4A\gamma^{-1}(1+c_0)^2(c^{-1}+1)^2$. This proves the first claim of Theorem 3.1 if we show that $\Pr(E^c) \leq 2p(p+1)(c_1T^{1-\mu}s_\alpha^\mu + \exp(-c_2T/s_\alpha^2)$. To that end, by the Cauchy-Schwartz inequality under Assumptions 3.1 (i)

$$\max_{1\leq j\leq k\leq p}\|x_{0,j}x_{0,k}\|_{r/2} \leq \max_{j\in[p]}\|x_{0,j}\|_r^2 = O(1).$$

This in conjunction with Assumption 3.1 (ii) verifies assumptions of Babii, Ghysels, and Striaukas (2020a), Theorem 3.1 and shows that

$$\begin{aligned}\Pr(E^c) &= \Pr\left(\left|\frac{1}{T}\sum_{t=1}^T x_tx_t^\top - \mathbb{E}[x_tx_t^\top]\right|_\infty > \frac{\gamma}{2G^*s_\alpha(1+2c_0)^2}\right) \\ &\leq c_1T^{1-\mu}s_\alpha^\mu p(p+1) + 2p(p+1)\exp\left(-\frac{c_2T^2}{s_\alpha^2B_T^2}\right)\end{aligned}$$

for some $c_1, c_2 > 0$ and $B_T^2 = \max_{j,k\in[p]}\sum_{t=1}^T\sum_{l=1}^T|\text{Cov}(x_{t,j}x_{t,k}, x_{l,j}x_{l,k})|$. Lastly, under Assumption 3.1, by Babii, Ghysels, and Striaukas (2020a), Lemma A.1.2 $B_T^2 = O(T)$.

To prove the second claim of Theorem 3.1, suppose first that $\Delta \in \mathcal{C}(2c_0)$. Then on the event E

$$\begin{aligned}
\Omega^2(\Delta) &= (\Omega_0(\Delta) + \Omega_1(\Delta))^2 \\
&\leq (1+2c_0)^2 \Omega_0^2(\Delta) \\
&\leq (1+2c_0)^2 \Delta^\top \Sigma \Delta s_\alpha/\gamma \\
&= (1+2c_0)^2 \left\{ \|\mathbf{X}\Delta\|_T^2 + \Delta^\top(\Sigma - \hat{\Sigma})\Delta \right\} s_\alpha/\gamma \\
&\leq (1+2c_0)^2 \left\{ C_1 s_\alpha \lambda^2 + 4\|\mathbf{m} - \mathbf{X}\beta\|_T^2 + \Omega^2(\Delta) G^* |\mathrm{vech}(\hat{\Sigma} - \Sigma)|_\infty \right\} s_\alpha/\gamma \\
&\leq (1+2c_0)^2 \left\{ C_1 s_\alpha \lambda^2 + 4\|\mathbf{m} - \mathbf{X}\beta\|_T^2 \right\} s_\alpha/\gamma + \frac{1}{2}\Omega^2(\Delta),
\end{aligned}$$

where we use the inequality in equations ((A2.5), (A2.6), and (A2.8)). Therefore,

$$\Omega^2(\Delta) \leq 2(1+2c_0)^2 \left\{ C_1 s_\alpha \lambda^2 + 4\|\mathbf{m} - \mathbf{X}\beta\|_T^2 \right\} s_\alpha/\gamma. \tag{A2.9}$$

On the other hand, if $\Delta \notin \mathcal{C}(2c_0)$, then $\Delta \notin \mathcal{C}(c_0)$, which as we have already shown implies $\|\mathbf{m} - \mathbf{X}\beta\|_T > 2^{-1}\|\mathbf{X}\Delta\|_T$. In conjunction with equations ((A2.3) and (A2.4)), this shows that

$$0 \leq \lambda c^{-1} \Omega(\Delta) + 2\|\mathbf{m} - \mathbf{X}\beta\|_T^2 + \lambda \left\{ \Omega_0(\Delta) - \Omega_1(\Delta) \right\},$$

and whence

$$\begin{aligned}
\Omega_1(\Delta) &\leq c_0 \Omega_0(\Delta) + \frac{2c}{\lambda(c-1)} \|\mathbf{m} - \mathbf{X}\beta\|_T^2 \\
&\leq \frac{1}{2}\Omega_1(\Delta) + \frac{2c}{\lambda(c-1)} \|\mathbf{m} - \mathbf{X}\beta\|_T^2.
\end{aligned}$$

This shows that

$$\begin{aligned}
\Omega(\Delta) &\leq (1 + (2c_0)^{-1}) \Omega_1(\Delta) \\
&\leq (1 + (2c_0)^{-1}) \frac{4c}{\lambda(c-1)} \|\mathbf{m} - \mathbf{X}\beta\|_T^2.
\end{aligned}$$

Combining this with the inequality in equation ((A2.9)), we obtain the second claim of Theorem 3.1. □

The following result is proven in Babii, Ghysels, and Striaukas (2020a), see their Theorem 3.1.

Theorem A2.1. *Let $(\xi_t)_{t \in \mathbf{Z}}$ be a centered stationary stochastic process in $\mathbf{R}^p$ such that (i) for some $\varsigma > 2$, $\max_{j \in [p]} \|\xi_{0,j}\|_\varsigma = O(1)$; (ii) for every*

$j \in [p]$, τ-mixing coefficients of $\xi_{t,j}$ satisfy $\tau_k^{(j)} \leq ck^{-a}$ for some constants $c > 0$ and $a > (\varsigma - 1)/(\varsigma - 2)$. Then there exists $C > 0$ such that for every $\delta \in (0, 1)$

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

with $\kappa = ((a + 1)\varsigma - 1)/(a + \varsigma - 1)$.

A2.3 ARDL-MIDAS: moments and τ-mixing coefficients

The ARDL-MIDAS process $(y_t)_{t\in\mathbf{Z}}$ is defined as

$$\phi(L)y_t = \xi_t,$$

where $\phi(L) = I - \rho_1 L - \rho_2 L^2 - \cdots - \rho_J L^J$ is a lag polynomial and $\xi_t = \sum_{j=0}^{p} x_{t,j}\gamma_j + u_t$. The process $(y_t)_{t\in\mathbf{Z}}$ is τ-mixing and has finite moments of order $q > 1$ as illustrated below.

Assumption A2.3.1. *Suppose that $(\xi_t)_{t\in\mathbf{Z}}$ is a stationary process such that (i) $\|\xi_t\|_q < \infty$ for some $q > 1$; (ii) the β-mixing coefficients satisfy $\beta_k \leq Ca^k$ for some $a \in (0, 1)$ and $C > 0$; and (iii) $\phi(z) \neq 0$ for all $z \in \mathbf{C}$ such that $|z| \leq 1$.*

Note that by Davydov (1973), (ii) holds if $(\xi_t)_{t\in\mathbf{Z}}$ is a geometrically ergodic Markov process and that (iii) rules out the unit root process.

Proposition A2.3.1. *Under Assumption A2.3.1, the ARDL-MIDAS process has moments of order $q > 1$ and τ-mixing coefficients $\tau_k \leq C(a^{bk} + c^k)$ for some $c \in (0, 1)$, $C > 0$, and $b = 1 - 1/q$.*

Proof. Under (iii) we can invert the autoregressive lag polynomial and obtain

$$y_t = \sum_{j=0}^{\infty}\psi_j\xi_{t-j}$$

for some $(\psi_j)_{j=0}^{\infty} \in \ell_1$. Note that $(y_t)_{t\in\mathbf{Z}}$ has dependent innovations. Clearly, $(y_t)_{t\in\mathbf{Z}}$ is stationary provided that $(\xi_t)_{t\in\mathbf{Z}}$ is stationary, which is the case by the virtue of Assumption A2.3.1. Next, since

$$\|y_t\|_q \leq \sum_{j=0}^{\infty}|\psi_j|\|\xi_0\|_q$$

and $\|\xi_0\|_q < \infty$ under (i), we verify that $\|y_t\|_q < \infty$. Let $(\xi'_t)_{t\in\mathbf{Z}}$ be a stationary process distributed as $(\xi_t)_{t\in\mathbf{Z}}$ and independent of $(\xi_t)_{t\leq 0}$. Then by Dedecker and Prieur (2005), Example 1, the τ-mixing coefficients of $(y_t)_{t\in\mathbf{Z}}$ satisfy

$$\begin{aligned}\tau_k &\leq \|\xi_0 - \xi'_0\|_q \sum_{j\geq k} |\psi_j| + 2\sum_{j=0}^{k-1} |\psi_j| \int_0^{\beta_{k-j}} Q_{\xi_0}(u)\mathrm{d}u \\ &\leq 2\|\xi_0\|_q \sum_{j\geq k} |\psi_j| + 2\|\xi_0\|_q \sum_{j=0}^{k-1} |\psi_j| \beta_{k-j}^{1-1/q},\end{aligned}$$

where $(\beta_k)_{k\geq 1}$ are β-mixing coefficients of $(\xi_t)_{t\in\mathbf{Z}}$ and the second line follows by Hölder's inequality. Brockwell and Davis (1991), p.85 shows that there exist $c \in (0,1)$ and $K > 0$ such that $|\psi_j| \leq Kc^j$. Therefore,

$$\sum_{j\geq k} |\psi_j| = O(c^k)$$

and under (ii)

$$\sum_{j=0}^{k-1} |\psi_j| \beta_{k-j}^{1-1/q} \leq CK \sum_{j=0}^{k-1} c^j a^{(k-j)(q-1)/q} \leq \begin{cases} CK\frac{a^{k(q-1)/q} - c^k}{1 - ca^{(1-q)/q}} & \text{if } c \neq a^{(q-1)/q}, \\ CKka^{k(q-1)/q} & \text{otherwise.} \end{cases}$$

This proves the second statement. □

A2.4 Monte Carlo Simulations

Table A2.1: Forecasting accuracy results. – See Table A2.2

	FLOW	STOCK	MIDDLE	LASSO-U	LASSO-M	SGL-M	FLOW	STOCK	MIDDLE	LASSO-U	LASSO-M	SGL-M
T			Baseline scenario						$\varepsilon_h \sim_{i.i.d.}$ student-$t(5)$			
50	1.920	2.086	2.145	1.848	1.731	1.537	2.081	2.427	2.702	2.399	2.038	1.702
	0.039	0.042	0.043	0.038	0.036	0.031	0.042	0.053	0.062	0.056	0.050	0.041
100	1.423	1.670	1.791	1.670	1.517	1.320	1.532	1.933	2.152	1.831	1.523	1.315
	0.029	0.033	0.036	0.034	0.031	0.027	0.030	0.039	0.044	0.037	0.031	0.027
200	1.292	1.502	1.645	1.407	1.268	1.170	1.410	1.741	2.017	1.493	1.278	1.194
	0.026	0.030	0.033	0.028	0.026	0.024	0.029	0.035	0.043	0.031	0.026	0.024
			High-frequency process: VAR(1)						Legendre degree $L = 5$			
50	1.869	2.645	2.863	2.192	1.712	1.431	1.920	2.086	2.145	1.848	1.741	1.598
	0.039	0.053	0.057	0.047	0.036	0.030	0.039	0.042	0.043	0.038	0.035	0.032
100	1.474	2.071	2.312	1.622	1.373	1.247	1.423	1.670	1.791	1.670	1.553	1.368
	0.030	0.042	0.048	0.033	0.028	0.026	0.029	0.033	0.036	0.034	0.032	0.028
200	1.335	1.919	2.080	1.369	1.239	1.216	1.292	1.502	1.645	1.407	1.298	1.187
	0.026	0.039	0.042	0.029	0.025	0.025	0.026	0.030	0.033	0.028	0.026	0.024
			Legendre degree $L = 10$						Low frequency noise level σ_u^2=5			
50	1.920	2.086	2.145	1.848	1.778	1.661	8.927	9.048	9.020	7.714	7.308	6.929
	0.039	0.042	0.043	0.038	0.037	0.034	0.182	0.184	0.181	0.155	0.149	0.140
100	1.423	1.670	1.791	1.670	1.617	1.446	6.643	7.300	7.536	7.510	6.953	6.305
	0.029	0.033	0.036	0.034	0.033	0.029	0.135	0.144	0.153	0.154	0.144	0.128
200	1.292	1.502	1.645	1.407	1.344	1.225	6.008	6.580	6.902	6.809	6.270	5.703
	0.026	0.030	0.033	0.028	0.027	0.025	0.123	0.131	0.137	0.137	0.127	0.115
			Half high-frequency lags						Number of covariates $p = 50$			
50	2.256	2.117	2.505	1.885	1.816	1.623				1.902	1.766	1.621
	0.047	0.044	0.050	0.038	0.037	0.033				0.038	0.035	0.032
100	1.655	1.685	2.079	1.679	1.595	1.370	3.593	3.277	3.318	1.754	1.599	1.403
	0.033	0.033	0.041	0.034	0.032	0.027	0.075	0.068	0.068	0.035	0.032	0.028
200	1.528	1.539	2.005	1.365	1.355	1.202	1.863	1.933	2.019	1.524	1.364	1.189
	0.031	0.030	0.040	0.027	0.027	0.024	0.038	0.039	0.039	0.030	0.027	0.024
			Baseline scenario, $\rho = 0.7$						Number of covariates $p = 50$, $\rho = 0.7$			
50	2.411	3.019	3.471	2.786	2.298	1.720				4.588	3.604	2.145
	0.051	0.059	0.069	0.061	0.051	0.036				0.093	0.077	0.044
100	1.717	2.423	2.943	1.710	1.501	1.331	5.351	5.030	4.854	2.275	1.910	1.424
	0.034	0.048	0.058	0.035	0.031	0.027	0.111	0.102	0.099	0.048	0.040	0.029
200	1.564	2.135	2.657	1.340	1.269	1.222	2.384	2.826	3.290	1.499	1.385	1.217
	0.032	0.043	0.052	0.027	0.026	0.025	0.048	0.056	0.065	0.030	0.028	0.024

Table A2.2: Nowcasting accuracy results. The table reports simulation results for nowcasting accuracy. The baseline DGP (upper-left block) is with the low-frequency noise level $\sigma_u^2 = 1$, the degree of Legendre polynomial $L = 3$, and Gaussian high-frequency noise. All remaining blocks report results for deviations from the baseline DGP. In the upper-right block, the noise term of high-frequency covariates is student-$t(5)$. Each block reports results for LASSO-U-MIDAS (LASSO-U), LASSO-MIDAS (LASSO-M), and sg-LASSO-MIDAS (SGL-M) (the last three columns). We also report results for aggregated predictive regressions with flow aggregation (FLOW), stock aggregation (STOCK), and taking the middle value (MIDDLE). We vary the sample size T from 50 to 200. Each entry in the odd row is the average mean squared forecast error, while each even row is the simulation standard error.

	FLOW	STOCK	MIDDLE	LASSO-U	LASSO-M	SGL-M	FLOW	STOCK	MIDDLE	LASSO-U	LASSO-M	SGL-M
T	Baseline scenario						$\varepsilon_h \sim_{i.i.d.}$ student-$t(5)$					
50	1.987	2.113	2.184	1.870	1.753	1.606	2.257	2.391	2.649	2.422	2.113	1.801
	0.043	0.042	0.043	0.038	0.036	0.032	0.046	0.054	0.057	0.052	0.046	0.038
100	1.446	1.632	1.769	1.667	1.541	1.345	1.659	1.889	2.139	1.903	1.678	1.462
	0.029	0.032	0.034	0.033	0.031	0.026	0.033	0.038	0.043	0.038	0.033	0.029
200	1.318	1.482	1.609	1.448	1.328	1.220	1.505	1.728	1.971	1.608	1.411	1.297
	0.026	0.029	0.032	0.029	0.026	0.024	0.030	0.035	0.041	0.033	0.028	0.026
	High-frequency process: VAR(1)						Legendre degree $L = 5$					
50	2.086	2.418	2.856	2.254	1.817	1.503	1.987	2.113	2.184	1.870	1.767	1.635
	0.044	0.050	0.057	0.049	0.039	0.031	0.043	0.042	0.043	0.038	0.036	0.033
100	1.642	1.935	2.365	1.690	1.459	1.328	1.446	1.632	1.769	1.667	1.564	1.389
	0.033	0.039	0.048	0.035	0.030	0.028	0.029	0.032	0.034	0.033	0.031	0.027
200	1.475	1.771	2.247	1.442	1.312	1.268	1.318	1.482	1.609	1.448	1.351	1.230
	0.029	0.036	0.046	0.029	0.027	0.026	0.026	0.029	0.032	0.029	0.027	0.024
	Legendre degree $L = 10$						Low frequency noise level σ_u^2=5					
50	1.987	2.113	2.184	1.870	1.799	1.698	9.121	9.208	9.167	7.700	7.397	7.087
	0.043	0.042	0.043	0.038	0.037	0.034	0.193	0.184	0.181	0.155	0.150	0.143
100	1.446	1.632	1.769	1.667	1.606	1.454	6.646	7.149	7.433	7.454	6.911	6.222
	0.029	0.032	0.034	0.033	0.032	0.029	0.135	0.141	0.144	0.149	0.138	0.123
200	1.318	1.482	1.609	1.448	1.400	1.267	6.052	6.482	6.777	6.835	6.345	5.780
	0.026	0.029	0.032	0.029	0.028	0.025	0.122	0.127	0.134	0.137	0.127	0.114
	Half high-frequency lags						Number of covariates $p = 50$					
50	2.378	2.164	2.540	1.875	1.827	1.723				1.912	1.767	1.611
	0.049	0.044	0.049	0.038	0.037	0.035				0.039	0.035	0.033
100	1.765	1.692	2.184	1.810	1.703	1.479	3.703	3.162	3.179	1.762	1.622	1.441
	0.035	0.033	0.042	0.036	0.033	0.029	0.076	0.064	0.068	0.035	0.032	0.028
200	1.605	1.520	1.976	1.544	1.495	1.324	1.912	1.871	2.017	1.546	1.428	1.260
	0.031	0.029	0.039	0.031	0.029	0.026	0.038	0.037	0.040	0.032	0.029	0.026
	Baseline scenario, $\rho = 0.7$						Number of covariates $p = 50$, $\rho = 0.7$					
50	2.606	2.872	3.618	2.927	2.599	1.884				4.606	3.816	2.242
	0.055	0.058	0.073	0.063	0.054	0.039				0.096	0.083	0.046
100	1.837	2.154	3.020	1.783	1.596	1.412	5.154	4.373	4.764	2.373	2.161	1.520
	0.037	0.043	0.059	0.037	0.032	0.028	0.102	0.089	0.100	0.051	0.046	0.030
200	1.661	1.919	2.753	1.389	1.341	1.287	2.622	2.555	3.364	1.563	1.500	1.315
	0.033	0.038	0.056	0.027	0.027	0.026	0.052	0.051	0.067	0.032	0.031	0.027

Table A2.3: Shape of weights estimation accuracy I. The table reports results for shape of weights estimation accuracy for the first four DGPs of Tables A2.1-A2.2 using LASSO-U, LASSO-M and SGL-M estimators for the weight functions Beta$(1,3)$, Beta$(2,3)$, and Beta$(2,2)$ with sample size $T = 50$, 100 and 200. Entries in odd rows are the average mean integrated squared error and in even rows the simulation standard error.

	LASSO-U	LASSO-M	SGL-M	LASSO-U	LASSO-M	SGL-M	LASSO-U	LASSO-M	SGL-M
	T=50			T=100			T=200		
				Baseline scenario					
Beta$(1,3)$	2.028	1.867	1.312	2.005	1.518	0.733	1.947	0.809	0.388
	0.001	0.009	0.015	0.001	0.011	0.010	0.002	0.009	0.005
Beta$(2,3)$	1.248	1.192	0.988	1.241	1.042	0.662	1.219	0.710	0.418
	0.001	0.006	0.011	0.001	0.006	0.008	0.001	0.006	0.005
Beta$(2,2)$	1.093	1.035	0.870	1.088	0.890	0.573	1.073	0.559	0.330
	0.001	0.005	0.009	0.001	0.006	0.007	0.001	0.005	0.004
				$\varepsilon_h \sim_{i.i.d.}$ student-$t(5)$					
Beta$(1,3)$	2.015	1.671	1.023	1.964	1.027	0.465	1.892	0.434	0.248
	0.001	0.011	0.014	0.002	0.011	0.007	0.001	0.005	0.004
Beta$(2,3)$	1.242	1.107	0.816	1.223	0.807	0.462	1.191	0.479	0.297
	0.001	0.007	0.010	0.001	0.007	0.006	0.001	0.005	0.004
Beta$(2,2)$	1.088	0.959	0.740	1.075	0.664	0.403	1.051	0.348	0.221
	0.001	0.006	0.009	0.001	0.006	0.006	0.001	0.004	0.003
				high-frequency process: VAR(1)					
Beta$(1,3)$	1.944	1.353	0.960	1.909	0.905	0.657	1.871	0.562	0.485
	0.003	0.014	0.014	0.002	0.010	0.009	0.002	0.006	0.006
Beta$(2,3)$	1.186	0.917	0.821	1.166	0.662	0.594	1.147	0.508	0.490
	0.002	0.012	0.013	0.002	0.009	0.008	0.001	0.006	0.005
Beta$(2,2)$	1.045	0.778	0.754	1.032	0.550	0.540	1.019	0.412	0.422
	0.002	0.011	0.012	0.001	0.008	0.008	0.001	0.005	0.005
				Legendre degree $L = 5$					
Beta$(1,3)$	2.028	1.907	1.487	2.005	1.619	0.909	1.947	0.915	0.436
	0.001	0.009	0.016	0.001	0.010	0.012	0.002	0.009	0.006
Beta$(2,3)$	1.248	1.211	1.090	1.241	1.091	0.783	1.219	0.772	0.462
	0.001	0.005	0.012	0.001	0.006	0.009	0.001	0.006	0.005
Beta$(2,2)$	1.093	1.055	0.962	1.088	0.938	0.672	1.073	0.619	0.356
	0.001	0.005	0.010	0.001	0.005	0.008	0.001	0.005	0.005
				Baseline scenario, $\rho = 0.7$					
Beta$(1,3)$	1.901	1.035	0.526	1.839	0.388	0.243	1.805	0.196	0.166
	0.003	0.012	0.009	0.003	0.005	0.004	0.002	0.002	0.002
Beta$(2,3)$	1.174	0.742	0.492	1.139	0.428	0.301	1.117	0.310	0.252
	0.002	0.009	0.008	0.002	0.005	0.004	0.002	0.003	0.003
Beta$(2,2)$	1.031	0.594	0.396	1.002	0.291	0.212	0.983	0.190	0.153
	0.002	0.007	0.006	0.002	0.003	0.003	0.002	0.002	0.002

Table A2.4: Shape of weights estimation accuracy II. – See Table A2.3

	LASSO-U	LASSO-M	SGL-M	LASSO-U	LASSO-M	SGL-M	LASSO-U	LASSO-M	SGL-M
	T=50			T=100			T=200		
				Legendre degree $L = 10$					
Beta$(1,3)$	2.028	1.962	1.685	2.005	1.769	1.150	1.947	1.078	0.528
	0.001	0.008	0.016	0.001	0.010	0.013	0.002	0.011	0.007
Beta$(2,3)$	1.248	1.247	1.247	1.241	1.168	0.960	1.219	0.869	0.522
	0.001	0.004	0.012	0.001	0.005	0.010	0.001	0.006	0.006
Beta$(2,2)$	1.093	1.086	1.091	1.088	1.011	0.823	1.073	0.710	0.398
	0.001	0.004	0.011	0.001	0.005	0.009	0.001	0.006	0.005
				low frequency noise level σ_u^2=5					
Beta$(1,3)$	2.038	1.941	1.588	2.025	1.816	1.109	1.983	1.436	0.563
	0.001	0.009	0.019	0.001	0.009	0.014	0.002	0.010	0.009
Beta$(2,3)$	1.252	1.215	1.144	1.246	1.160	0.878	1.230	0.996	0.529
	0.001	0.006	0.015	0.001	0.005	0.010	0.001	0.006	0.007
Beta$(2,2)$	1.096	1.065	1.022	1.092	1.007	0.773	1.080	0.845	0.460
	0.001	0.006	0.013	0.001	0.005	0.009	0.001	0.005	0.007
				Half high-frequency lags					
Beta$(1,3)$	2.028	1.826	1.219	1.990	1.504	0.825	1.924	0.964	0.611
	0.001	0.009	0.012	0.001	0.010	0.008	0.001	0.007	0.004
Beta$(2,3)$	1.252	1.206	1.072	1.243	1.133	0.925	1.224	0.968	0.779
	0.000	0.004	0.008	0.001	0.004	0.006	0.001	0.005	0.005
Beta$(2,2)$	1.096	1.060	0.991	1.090	1.007	0.878	1.076	0.890	0.783
	0.000	0.004	0.008	0.000	0.004	0.006	0.000	0.004	0.004
				Number of covariates $p = 50$					
Beta$(1,3)$	2.044	1.998	1.586	2.032	1.867	1.061	1.999	1.285	0.512
	0.000	0.004	0.012	0.001	0.007	0.011	0.001	0.009	0.006
Beta$(2,3)$	1.255	1.238	1.099	1.252	1.191	0.875	1.243	0.963	0.533
	0.000	0.002	0.007	0.000	0.004	0.007	0.001	0.005	0.005
Beta$(2,2)$	1.099	1.083	0.979	1.097	1.036	0.782	1.091	0.804	0.467
	0.000	0.002	0.007	0.000	0.003	0.006	0.000	0.005	0.005
				Number of covariates $p = 50$, $\rho = 0.7$					
Beta$(1,3)$	1.996	1.726	0.878	1.902	0.839	0.334	1.835	0.314	0.188
	0.002	0.010	0.011	0.002	0.009	0.005	0.002	0.003	0.002
Beta$(2,3)$	1.229	1.071	0.692	1.180	0.648	0.344	1.138	0.411	0.248
	0.001	0.006	0.008	0.002	0.006	0.004	0.002	0.003	0.003
Beta$(2,2)$	1.078	0.925	0.610	1.040	0.495	0.276	1.003	0.272	0.167
	0.001	0.005	0.007	0.001	0.005	0.004	0.001	0.002	0.002

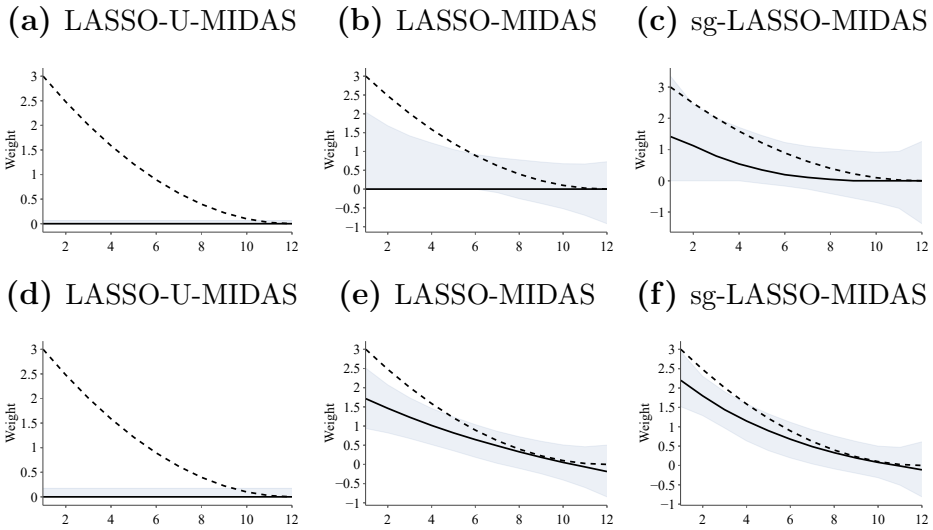


Figure A2.1: The figure shows the fitted Beta(1,3) weights. We plot the estimated weights for the LASSO-U-MIDAS, LASSO-MIDAS, and sg-LASSO-MIDAS estimators for the baseline DGP scenario. The first row plots weights for the sample size $T = 50$, the second row plots weights for the sample size $T = 200$. The black solid line is the median estimate of the weights function, the black dashed line is the population weight function, and the gray area is the 90% confidence interval.

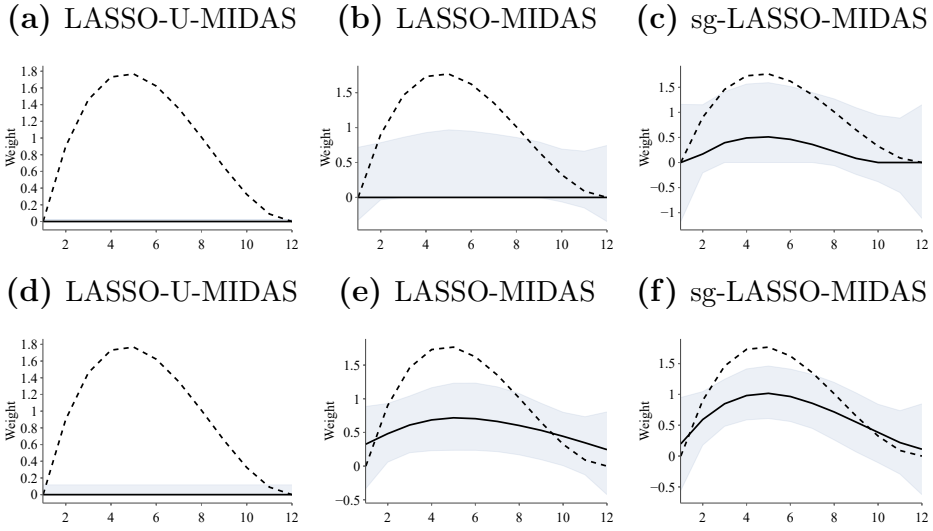


Figure A2.2: The figure shows the fitted Beta(2,3) weights. We plot the estimated weights for the LASSO-U-MIDAS, LASSO-MIDAS, and sg-LASSO-MIDAS estimators for the baseline DGP scenario. The first row plots weights for the sample size $T = 50$, the second row plots weights for the sample size $T = 200$. The black solid line is the median estimate of the weights function, the black dashed line is the population weight function, and the gray area is the 90% confidence interval.

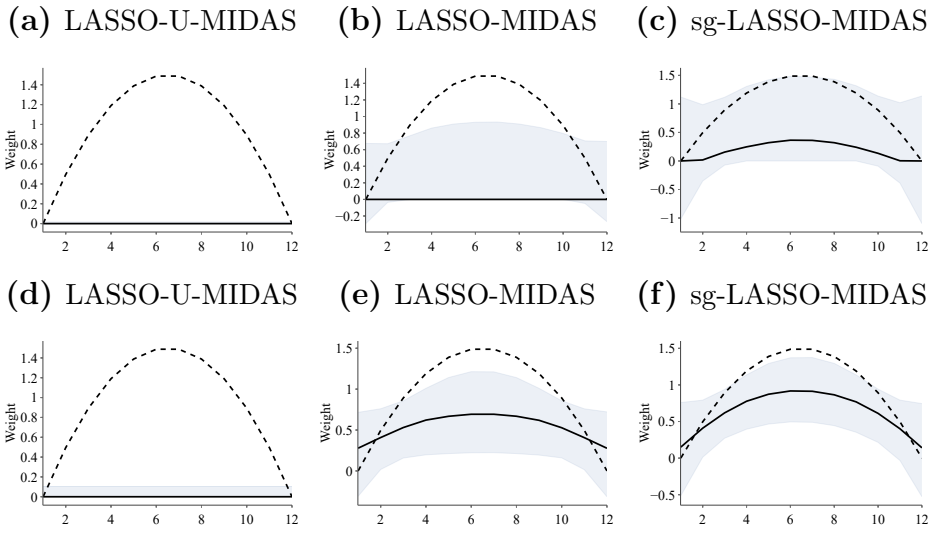


Figure A2.3: The figure shows the fitted Beta(2,2) weights. We plot the estimated weights for the LASSO-U-MIDAS, LASSO-MIDAS, and sg-LASSO-MIDAS estimators for the baseline DGP scenario. The first row plots weights for the sample size $T = 50$, the second row plots weights for the sample size $T = 200$. The black solid line is the median estimate of the weights function, the black dashed line is the population weight function, and the gray area is the 90% confidence interval.

A2.5 Detailed description of data and models

The standard macro variables are collected from *Haver Analytics* and *ALFRED* databases. ALFRED is a public data source for real-time data made available by the Federal Serve Bank of St. Louis; see the full list of the series with further details in Table A2.5. For series that are collected from the Haver Analytics database, we use *as reported* data, that is the first release is used for each data point. For the data that we collect from ALFRED, full data vintages are used. All the data is real-time, hence publication delays for each series are taken into consideration and we align each series accordingly. We use twelve monthly and four quarterly lags for each monthly and quarterly series respectively and apply Legendre aggregation with polynomial degree set to three. The groups are defined as lags of each series.

On top of macro data, we add eight financial series which are collected from *FRED* database; the full list of the series appears in Table A2.6. These series are available in real time, hence no publication delays are needed in this case. We use three monthly lags and apply Legendre aggregation with polynomial degree set to two. As for macro, we group all lags of each series.

Lastly, we add textual analysis covariates from www.structureofnews.com. The data is real time, i.e., topic models are estimated for each day and the monthly series are obtained by aggregating daily data; see Bybee, Kelly, Manela, and Xiu (2020) for further details on the data construction. We use categories of series are potentially closely tied with economic activity, which are *Banks*, *Economic Growth*, *Financial Markets*, *Government*, *Industry*, *International Affairs*, *Labor/income*, and *Oil & Mining*. In total, we add 76 news attention series; the full list is available in Table A2.7. Three lags are used and Legendre aggregation of degree two is applied to each series. In this case, we group variables based on categories.

To make the comparison with the NY Fed nowcasts as close as possible, we use 15 years (60 quarters) of the data and use rolling window estimation. The first nowcast is for the 2002 Q1 (first quarter that NY Fed publishes its historic nowcasts) and the effective sample size starts at 1988 Q1 (taking 15 years of data accounting for lags). We calculate predictions until the sample is exhausted, which is 2017 Q2, the last date for which news attention data is available. Real GDP growth rate data vintages are taken from *ALFRED* database. Some macro series start later than 1988 Q1, in which case we

impute zero values. Lastly, we use four lags of real GDP growth rate in all models.

Alternative estimators We implemented the following alternative machine learning nowcasting methods. The first method is the PCA factor-augmented autoregression, where we estimate the first principal component of the data panel and use it together with four autoregressive lags. We denote this model PCA-OLS. We then consider three alternative penalty functions for the same linear model: ridge, LASSO, and Elastic Net. For these methods, we leave high-frequency lags unrestricted, and thus we call these methods the unrestricted MIDAS (U-MIDAS). As for the sg-LASSO-MIDAS model, we tune one- and two-dimensional regularization parameters via 5-fold cross-validation.

Table A2.5: Data description table (macro data)– The *Series* column gives a time-series name, which is given in the second column *Source*. The column *Units* denotes the data transformation applied to a time-series.

	Series	Source	Units
1	ADP nonfarm private payroll employment	Haver	Level change (thousands)
2	Building permits	ALFRED	Level change (thousands)
3	Capacity utilization	ALFRED	Ppt. change
4	Civilian unemployment rate	ALFRED	Ppt. change
5	CPI-U: all items	ALFRED	MoM % change
6	CPI-U: all items less food and energy	ALFRED	MoM % change
7	Empire State Mfg. survey: general business conditions	Haver	Index
8	Exports: goods and services	Haver	MoM % change
9	Export price index	Haver	MoM % change
10	Housing starts	ALFRED	MoM % change
11	Imports: goods and services	Haver	MoM % change
12	Import price index	Haver	MoM % change
13	Industrial production index	ALFRED	MoM % change
14	Inventories: Total business	ALFRED	MoM % change
15	ISM mfg.: PMI composite index	Haver	Index
16	ISM mfg.: Prices index	Haver	Index
17	ISM mfg.: Employment index	Haver	Index
18	ISM nonmanufacturing: NMI composite index	Haver	Index
19	JOLTS: Job openings: total	Haver	Level change (thousands)
20	Manufacturers new orders: durable goods	ALFRED	MoM % change
21	Manufacturing payrolls	Haver	Level change (thousands)
22	Manufacturers shipments: durable goods	Haver	MoM % change
23	Manufacturers inventories: durable goods	Haver	MoM % change
24	Manufacturers' unfilled orders: total manufacturing	Haver	MoM % change
25	New single family houses sold	ALFRED	MoM % change
26	Nonfarm business sector: unit labor cost	ALFRED	QoQ % change (annual rate)
27	PCE less food and energy: chain price index	ALFRED	MoM % change
28	PCE: chain price index	ALFRED	MoM % change
29	Philly Fed Mfg. business outlook: current activity	Haver	Index
30	Retail sales and food services	ALFRED	MoM % change
31	Real personal consumption expenditures	ALFRED	MoM % change
32	Real gross domestic income	Haver	QoQ % change (annual rate)
33	Real disposable personal income	ALFRED	MoM % change
34	Value of construction put in place	Haver	MoM % change

Table A2.6: Data description table (financial and uncertainty series) – The *Series* column gives a time-series name, which is given in the second column *Source*. The column *Units* denotes the data transformation applied to a time-series.

	Series	Source	Units
1	BAA less AAA corporate bond spread	FRED	Level
2	BAA less 10-year bond spread	FRED	Level
3	S&500	FRED	Log-returns %
4	TED spread	FRED	Level
5	10-year less 3-month bond spread	FRED	Level
6	VIX	FRED	Level
7	Economic policy uncertainty index (EPUI)	FRED	Index
8	Equity market-related economic uncertainty index (EMEUI)	FRED	Index

	Group	Series
1	Banks	Bank loans
2	Banks	Credit ratings
3	Banks	Financial crisis
4	Banks	Mortgages
5	Banks	Nonperforming loans
6	Banks	Savings & loans
7	Economic Growth	Economic growth
8	Economic Growth	European sovereign debt
9	Economic Growth	Federal Reserve
10	Economic Growth	Macroeconomic data
11	Economic Growth	Optimism
12	Economic Growth	Product prices
13	Economic Growth	Recession
14	Economic Growth	Record high
15	Financial Markets	Bear/bull market
16	Financial Markets	Bond yields
17	Financial Markets	Commodities
18	Financial Markets	Currencies/metals
19	Financial Markets	Exchanges/composites
20	Financial Markets	International exchanges
21	Financial Markets	IPOs
22	Financial Markets	Options/VIX
23	Financial Markets	Share payouts
24	Financial Markets	Short sales
25	Financial Markets	Small caps
26	Financial Markets	Trading activity
27	Financial Markets	Treasury bonds
28	Government	Environment
29	Government	National security
30	Government	Political contributions
31	Government	Private/public sector
32	Government	Regulation
33	Government	Safety administrations
34	Government	State politics
35	Government	Utilities
36	Government	Watchdogs
37	Industry	Cable
38	Industry	Casinos
39	Industry	Chemicals/paper
40	Industry	Competition
41	Industry	Couriers
42	Industry	Credit cards
43	Industry	Fast food
44	Industry	Foods/consumer goods
45	Industry	Insurance
46	Industry	Luxury/beverages
47	Industry	Revenue growth
48	Industry	Small business
49	Industry	Soft drinks
50	Industry	Subsidiaries
51	Industry	Tobacco
52	Industry	Venture capital
53	International Affairs	Canada/South Africa
54	International Affairs	China
55	International Affairs	France/Italy
56	International Affairs	Germany
57	International Affairs	Japan
58	International Affairs	Latin America
59	International Affairs	Russia
60	International Affairs	Southeast Asia
61	International Affairs	Trade agreements
62	International Affairs	UK
63	Labor/income	Executive pay
64	Labor/income	Fees
65	Labor/income	Government budgets
66	Labor/income	Health insurance
67	Labor/income	Job cuts

68	Labor/income	Pensions
69	Labor/income	Taxes
70	Labor/income	Unions
71	Oil & Mining	Agriculture
72	Oil & Mining	Machinery
73	Oil & Mining	Mining
74	Oil & Mining	Oil drilling
75	Oil & Mining	Oil market
76	Oil & Mining	Steel

Table A2.7: Data description table (textual data) – The *Group* column is a group name of individual textual analysis series which appear in the column *Series*. Data is taken in levels.

A2.5.1 Additional results

Table A2.8: Nowcast comparisons for models with macro data including series with short samples – Nowcast horizons are 2- and 1-month ahead, as well as the end of the quarter. Column *Rel-RMSE* reports root mean squared forecasts error relative to the AR(1) model. Column *DM-stat-1* reports Diebold and Mariano (1995) test statistic of all models relative to NY Fed nowcasts, while column *DM-stat-2* reports the Diebold Mariano test statistic relative to sg-LASSO-MIDAS model. Out-of-sample period: 2002 Q1 to 2017 Q2.

	Rel-RMSE	DM-stat-1	DM-stat-2
		2-month horizon	
sg-LASSO-MIDAS	0.737	-2.500	
NY Fed	0.946		2.500
		1-month horizon	
sg-LASSO-MIDAS	0.726	-0.804	
NY Fed	0.805		0.804
		End-of-quarter	
sg-LASSO-MIDAS	0.704	-0.048	
NY Fed	0.708		0.048

Table A2.9: Nowcast comparison table (excluding financial data in Table A2.6) – Nowcast horizons are 2- and 1-month ahead, as well as the end of the quarter. Column *Rel-RMSE* reports root mean squared forecasts error relative to the AR(1) model. Column *DM-stat-1* reports Diebold and Mariano (1995) test statistic of all models relative to the NY FED nowcast. Out-of-sample period: 2002 Q1 to 2017 Q2.

	Rel-RMSE	DM-stat-1	DM-stat-2
		2-month horizon	
sg-LASSO-MIDAS	0.794	-1.780	
NY Fed	0.946		1.780
		1-month horizon	
sg-LASSO-MIDAS	0.693	-1.161	
NY Fed	0.805		1.161
		End-of-quarter	
sg-LASSO-MIDAS	0.691	-0.221	
NY Fed	0.707		0.221

Chapter 3

High-Dimensional Granger Causality Tests with an Application to VIX and News

with Andrii Babii and Eric Ghysels

3.1 Introduction

Modern time series analysis is increasingly using high-dimensional datasets, typically available at different frequencies. Conventional time series are often supplemented with non-traditional data, such as the high-dimensional data coming from the natural language processing. For instance, Bybee, Kelly, Manela, and Xiu (2020) extract 180 topic attention series from the over 800,000 daily *Wall Street Journal* news articles during 1984-2017 that have shown by Babii, Ghysels, and Striaukas (2020b) to be a useful supplement to more traditional macroeconomic and financial datasets for nowcasting US GDP growth.

In his seminal paper, Clive Granger defined causality in terms of high-dimensional time series data. His formal definition, see (Granger, 1969, Definition 1), considered all the information accumulated in the universe up to time $t-1$ (a process he called U_t) and examined predictability using U_t with and without a specific series of interest Y_t. It is still an open question how to implement Granger's test in a high-dimensional time series setting. It is the purpose of this paper to do this via regularized regressions using HAC-based inference. In a sense, we are trying to implement Granger's original idea of causality.[1]

It is worth relating our to the existing literature on Granger causality with high-dimensional data. Various dimensionality reduction schemes

[1]There exists an extensive literature on causal inference with machine learning methods within the *static* Neyman-Rubin's potential outcomes framework; see Athey and Imbens (2019) for the excellent review and further references.

have been considered. For example, Box and Tiao (1977) used canonical correlation analysis, Peña and Box (1987) and Stock and Watson (2002) proposed factor models and principle component analysis. Koop (2013) analyzed large dimensional Bayesian VAR models. More closely related to our paper are Yuan and Lin (2006), Simon, Friedman, Hastie, and Tibshirani (2013), Skripnikov and Michailidis (2019), Nicholson, Wilms, Bien, and Matteson (2020), and Babii, Ghysels, and Striaukas (2020b) who look at structured sparsity approaches without doing inference. Granger causality with sparsity and inference also appeared in a number of papers. Wilms, Gelper, and Croux (2016) use bootstrap but ignore post-selection issues, while Hecq, Margaritella, and Smeekes (2019) extend post-double selection approach of Belloni, Chernozhukov, and Hansen (2014) to Granger causality testing in linear sparse high-dimensional VAR. Finally, Ghysels, Hill, and Motegi (2020) propose a Granger causality test based on a seemingly overlooked, but simple, dimension reduction technique. The procedure involves multiple parsimonious regression models where key regressors are split across simple regressions. Each parsimonious regression model has one key regressor and other regressors not associated with the null hypothesis. The test is based on the maximum of the squared parameters of the key regressors.

Following Babii, Ghysels, and Striaukas (2020b), we focus on the structured sparsity approach based on the sparse-group LASSO (sg-LASSO) regularization for the high-dimensional time series analysis. The sg-LASSO allows capturing the group structures present in high-dimensional time series regressions where a single covariate with its lags constitutes a group. Alternatively, we can also combine covariates of similar nature in groups. An attractive feature of this estimator is that it encompasses the LASSO and the group LASSO as special cases, hence, it allows improving upon the unstructured LASSO in the high-dimensional time-series setting. At the same time, the sg-LASSO can learn the distribution of time series lags in a data-driven way solving elegantly the model selection problem that dates back to Fisher (1937).[2] In particular, the group structure can also accommodate data sampled at different frequencies as discussed in detail by Babii, Ghysels, and Striaukas (2020b).

The proper inference for time-series data relies on the heteroskedasticity

[2]The distributed lag literature can be traced back to Fisher (1925); see also Almon (1965), Sims (1971), and Shiller (1973), as well as more recent mixed frequency data sampling (MIDAS) approach in Ghysels, Santa-Clara, and Valkanov (2006), Ghysels, Sinko, and Valkanov (2007), and Andreou, Ghysels, and Kourtellos (2013).

and autocorrelation consistent (HAC) estimation of the long-run variance; see Eicker (1963), Huber (1967), White (1980), Gallant (1987), Newey and West (1987), Andrews (1991), among others.[3] Despite the increasing popularity of the LASSO in finance and more generally in time series empirical research, to the best of our knowledge, the validity of HAC-based inference for LASSO has not been established in the relevant literature.[4] The HAC-based inference is robust to the model misspecification and leads to the valid Granger causality tests even when the fitted regression function has only projection interpretation which is the case for the projection-based definition of the Granger causality. Developing the asymptotic theory for the linear projection model with autoregressive lags and covariates, however, is challenging because the underlying processes are typically *not* β-mixing.[5]

In this paper, we obtain the debiased central limit theorem with explicit bias correction for the sg-LASSO estimator and time series data, which extends van de Geer, Bühlmann, Ritov, and Dezeure (2014) and to the best of our knowledge is new. Next, we establish the formal statistical properties of the HAC estimator based on the sg-LASSO residuals in the high-dimensional environment when the number of covariates can increase faster than the sample size. The convergence rate of the HAC estimator can be affected by the tails and the persistence of the data, which is a new phenomenon compared to low-dimensional regressions. For the practical implementation, this implies that the optimal choice of the bandwidth parameter for the HAC estimator should scale appropriately with the number of covariates, the tails, and the persistence of the data. These results allow us to perform inference for groups of coefficients, including the (mixed-frequency) Granger causality tests.

Our asymptotic theory applies to the heavy-tailed time series data, which is often observed in financial and economic applications. To that end, we establish a new Fuk-Nagaev inequality, see Fuk and Nagaev (1971), for τ-*mixing* processes with *polynomial tails*. The class of τ-mixing processes is flexible enough for developing the asymptotic theory for the linear projection

[3]For stationary time series, the HAC estimation of the long-run variance is the same problem as the estimation of the value of the spectral density at zero which itself has even longer history dating back to the smoothed periodogram estimators; see Daniell (1946), Bartlett (1948), and Parzen (1957).

[4]See Chernozhukov, Härdle, Huang, and Wang (2021) for LASSO inference and causal Bernoulli shifts with independent innovations and Feng, Giglio, and Xiu (2020) for an asset pricing application; see also Belloni, Chernozhukov, and Hansen (2014) and van de Geer, Bühlmann, Ritov, and Dezeure (2014) for i.i.d. data; and Chiang and Sasaki (2019) for exchangeable arrays.

[5]More generally, it is known that the linear transformations based on infinitely many lags do not preserve the α- or β-mixing property.

model and, at the same time, it contains the class of α-mixing processes as a special case.

The paper is organized as follows. We start with the large sample approximation to the distribution of the sg-LASSO estimator (and as a consequence of the LASSO and the group LASSO) with τ-mixing data in section 3.2. Next, we consider the HAC estimator of the asymptotic long-run variance based on the sg-LASSO residuals and study the inference for groups of regression coefficients. In section 3.3, we establish a suitable version of the Fuk-Nagaev inequality for τ-mixing processes. We report on a Monte Carlo study in section 3.4 which provides further insights about the validity of our theoretical analysis in finite sample settings typically encountered in empirical applications. Section 3.5 covers an empirical application examining the Granger causal relations between the VIX and financial news. Conclusions appear in section 3.6. Proofs and supplementary results appear in the appendix and the supplementary material.

Notation: For a random variable $X \in \mathbf{R}$ and $q \geq 1$, let $\|X\|_q = (\mathbb{E}|X|^q)^{1/q}$ be its L_q norm. For $p \in \mathbf{N}$, put $[p] = \{1, 2, \ldots, p\}$. For a vector $\Delta \in \mathbf{R}^p$ and a subset $J \subset [p]$, let Δ_J be a vector in $\mathbf{R}^p$ with the same coordinates as Δ on J and zero coordinates on J^c. Let $\mathcal{G} = \{G_g : g \geq 1\}$ be a partition of $[p]$ defining groups. For a vector of regression coefficients $\beta \in \mathbf{R}^p$, the sparse-group structure is described by a pair $(S_0, \mathcal{G}_0)$, where $S_0 = \{j \in [p] : \beta_j \neq 0\}$ is the support of β and $\mathcal{G}_0 = \{G \in \mathcal{G} : \beta_G \neq 0\}$ is its group support. For $b \in \mathbf{R}^p$ and $q \geq 1$, its ℓ_q norm is denoted $|b|_q = \left(\sum_{j\in[p]} |b_j|^q\right)^{1/q}$ if $q < \infty$ and $|b|_\infty = \max_{j\in[p]} |b_j|$ if $q = \infty$. For $\mathbf{u}, \mathbf{v} \in \mathbf{R}^T$, the empirical inner product is defined as $\langle \mathbf{u}, \mathbf{v} \rangle_T = \frac{1}{T}\sum_{t=1}^T u_t v_t$ with the induced empirical norm $\|.\|_T^2 = \langle ., . \rangle_T = |.|_2^2/T$. For a symmetric $p \times p$ matrix A, let $\mathrm{vech}(A) \in \mathbf{R}^{p(p+1)/2}$ be its vectorization consisting of the lower triangular and the diagonal part. 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 put $A_G = A_j$. For a $p \times p$ matrix A, let $\|A\|_\infty = \max_{j\in[p]} |A_j|_1$ be its matrix norm. For $a, b \in \mathbf{R}$, we put $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Lastly, we write $a_n \lesssim b_n$ if there exists a (sufficiently large) absolute constant C such that $a_n \leq C b_n$ for all $n \geq 1$ and $a_n \sim b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

3.2 HAC-based inference for sg-LASSO

In this section, we cover the large sample approximation to the distribution of the sg-LASSO (LASSO and group LASSO) estimator for τ-mixing processes. Next, we consider the HAC estimator of the asymptotic long-run variance based on the sg-LASSO residuals and consider the Granger causality tests. In the first subsection, we cover the debiased central limit theorem. The next subsection covers the HAC estimator and the final subsection pertains to the Granger causality tests.

3.2.1 Debiased central limit theorem

Consider a generic linear projection model

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

where $(y_t)_{t\in\mathbf{Z}}$ is a real-valued stochastic process and predictors may include the intercept, some covariates, (mixed-frequency) lags of covariates up to a certain order, as well as lags of the dependent variable. For a sample of size T, in the vector notation, we write

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

where $\mathbf{y} = (y_1, \ldots, y_T)^\top$, $\mathbf{m} = (m_1, \ldots, m_T)^\top$ with $m_t = \sum_{j=1}^{\infty} \beta_j x_{t,j}$, and $\mathbf{u} = (u_1, \ldots, u_T)^\top$. We approximate m_t with $x_t^\top \beta = \sum_{j=1}^{p} \beta_j x_{t,j}$ and put $\mathbf{X}\beta$, where $\mathbf{X}$ is $T \times p$ design matrix and $\beta \in \mathbf{R}^p$ is the unknown projection parameter. This approximation can be constructed from lagged values of y_t, some covariates, as well as lagged values of covariates measured at a higher frequency, in which case, we obtain the autoregressive distributed lag mixed frequency data sampling model (ARDL-MIDAS) described as

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

where $\phi(L) = I - \rho_1 L - \rho_2 L^2 - \cdots - \rho_J L^J$ is a low frequency lag polynomial and the MIDAS part $\psi(L^{1/m}; \beta_k) x_{t,k} = \frac{1}{m} \sum_{j=1}^{m} \beta_{k,j} x_{t-(j-1)/m,k}$ is a high-frequency lag polynomial; see Andreou, Ghysels, and Kourtellos (2013) and Babii, Ghysels, and Striaukas (2020b). Note that when $m = 1$ we have all data sampled at the same frequency and recover the standard autoregressive

distributed lag (ARDL) model. The ARDL-MIDAS regression has a group structure where a single group is defined as all lags of $x_{t,k}$ or all lags of y_t and following Babii, Ghysels, and Striaukas (2020b), we focus on the sparse-group LASSO (sg-LASSO) regularized estimator.[6] The leading example here is the MIDAS regression involving the projection of future low frequency series onto its own lags and lags of high frequency data aggregated via some dictionary, e.g., the set of Legendre polynomials. The setup also covers what is sometimes called the reverse MIDAS, see Foroni, Guérin, and Marcellino (2018) and mixed frequency VAR, see Ghysels (2016), involving the projection of high frequency data onto its own (high frequency) lags and low frequency data. Such regressions, which appear in the empirical application of the paper, simply amount to a different group structure.

The sg-LASSO, denoted $\hat{\beta}$, solves the regularized least-squares problem

$$\min_{b\in\mathbf{R}^p} \|\mathbf{y}-\mathbf{X}b\|_T^2 + 2\lambda\Omega(b) \tag{3.1}$$

with the regularization functional

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

where $|b|_1 = \sum_{j=1}^p |b_j|$ is the ℓ_1 norm corresponding to the LASSO penalty, $\|b\|_{2,1} = \sum_{G\in\mathcal{G}} |b_G|_2$ is the group LASSO penalty, and the group structure $\mathcal{G}$ is a partition of $[p] = \{1, 2, \dots, p\}$ specified by the econometrician.

We measure the persistence of the series 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 $\mathrm{Lip}_1 = \{f : \mathbf{R}^l \to \mathbf{R} : \ |f(x)-f(y)| \le |x-y|_1\}$ is a set of 1-Lipschitz functions. Let $(\xi_t)_{t\in\mathbf{Z}}$ be a stochastic process and let $\mathcal{M}_t = \sigma(\xi_t, \xi_{t-1}, \dots)$ be its natural filtration. The τ-mixing coefficient is defined as

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

where the supremum is taken over t and $(t_1, \dots, t_j)$. The process is called τ-mixing if $\tau_k \downarrow 0$ as $k \uparrow \infty$; see Lemma A3.1.1 for the comparison of this coefficient to the mixingale and the α-mixing coefficients. The following assumptions impose tail and moment conditions on the series of interest.

[6] The sg-LASSO estimator allows selecting groups and important group members at the same time.

Assumption 3.2.1 (Data). *The processes $(u_t, x_t)_{t\in\mathbf{Z}}$ is stationary for every $p \geq 1$ and such that (i) $\|u_t\|_q < \infty$ and $\max_{j\in[p]} \|x_{t,j}\|_r = O(1)$ for some $q > 2r/(r-2)$ and $r > 4$; (ii) for every $j, l \in [p]$, the τ-mixing coefficients of $(u_t x_{t,j})_{t\in\mathbf{Z}}$ and $(x_{t,j}x_{t,l})$ are $\tau_k \leq ck^{-a}$ and $\tilde{\tau}_k \leq ck^{-b}$ for all $k \geq 0$ and some universal constants $c > 0$, $a > (\varsigma-1)/(\varsigma-2)$, $b > (r-2)/(r-4)$, and $\varsigma = qr/(q+r)$.*

Assumption 3.2.1 can be relaxed to non-stationary data with stable variances of partial sums at the cost of heavier notation. It allows for heavy-tailed and persistent data. For instance, it requires that either both covariates and the error process have at least $4+\epsilon$ finite moments, or that the error process has at least $2+\epsilon$ finite moments, whenever covariates are sufficiently integrable. It is also known that the τ-mixing coefficients decline exponentially fast for geometrically ergodic Markov chains, including the stationary AR(1) process, so condition (ii) allows for relatively persistent data; see also Babii, Ghysels, and Striaukas (2020b) for verification of these conditions in a toy heavy-tailed autoregressive model with covariates. Next, we require that the covariance matrix of covariates is invertible.

Assumption 3.2.2 (Covariance). *There exists a universal constant $\gamma > 0$ such that the smallest eigenvalue of $\Sigma = \mathbb{E}[x_t x_t^\top]$ is bounded away from zero by γ.*

Assumption 3.2.2 ensures that the precision matrix $\Theta = \Sigma^{-1}$ exists and rules out perfect multicollinearity. It also requires that the smallest eigenvalue of Σ is bounded away from zero by γ independently of the dimension p which is the case, e.g., for the spiked identity and the Toeplitz covariance structures. Strictly, speaking this condition can be relaxed to $\gamma \downarrow 0$ as $p \uparrow \infty$ at the cost of slower convergence rates and more involved conditions on rates, in which case γ can be interpreted as a measure of ill-posedness; see Carrasco, Florens, and Renault (2007). The next assumption describes the rate of the regularization parameter, which is governed by the Fuk-Nagaev inequality; see Theorem 3.1 and Eq. (3.4).

Assumption 3.2.3 (Regularization). *For some $\delta \in (0,1)$*

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

where $\kappa = ((a+1)\varsigma - 1)/(a+\varsigma-1)$, where a, ς are as in Assumption 3.2.1.

In practice we recommend to select the tuning parameter in a data-driven way. It is beyond the scope of the present paper to study properties of estimators with data-driven tuning parameters; see Chetverikov, Liao, and Chernozhukov (2021) for this type of analysis with i.i.d. data. Lastly, we impose the following condition on the misspecification error, the number of covariates p, the sparsity constant s_α, and the sample size T.

Assumption 3.2.4. *(i)* $\|\mathbf{m} - \mathbf{X}\beta\|_T^2 = O_P\left(s_\alpha \lambda^2\right)$*; (ii)* $s_\alpha^\mu p^2 T^{1-\mu} \to 0$ *and* $p^2 \exp(-cT/s_\alpha^2) \to 0$ *as* $T \to \infty$*, where* s_α *is the effective sparsity of* β *(defined below) and* $\mu = ((b+1)r - 2)/(r + 2(b-1))$*.*

The effective sparsity constant $\sqrt{s_\alpha} = \alpha\sqrt{|S_0|} + (1-\alpha)\sqrt{|\mathcal{G}_0|}$ is a linear combination of the sparsity $|S_0|$ (number of non-zero coefficients) and the group sparsity $|\mathcal{G}_0|$ (number of active groups). It reflects the finite sample advantages of imposing the sparse-group structure as $|\mathcal{G}_0|$ can be significantly smaller than $|S_0|$ that appears in the theory of the standard LASSO estimator. Throughout the paper we assume that the groups have fixed size, which is well-justified in time-series applications of interest.

The four assumptions listed above are needed for the prediction and estimation consistency of the sg-LASSO estimator; see Theorem A3.1 in the supplementary material. Next, let $v_{t,j}$ be the regression error in j^{th} nodewise LASSO regression; see the following subsection for more details. Put also $s = s_\alpha \vee S$, $S = \max_{j \in G} S_j$, where S_j is the number of non-zero coefficients in the j^{th} row of Θ. The following assumption describes an additional set of sufficient conditions for the debiased central limit theorem.

Assumption 3.2.5. *(i)* $\sup_x \mathbb{E}[u_t^2 | x_t = x] = O(1)$*; (ii)* $\|\Theta_G\|_\infty = O(1)$ *for some* $G \subset [p]$ *of fixed size; (iii) the long run variance of* $(u_t^2)_{t \in \mathbf{Z}}$ *and* $(v_{t,j}^2)_{t \in \mathbf{Z}}$ *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)* $\|\mathbf{m} - \mathbf{X}\beta\|_T = o_P(T^{-1/2})$*; (vi) for every* $j, l \in [p]$ *and* $k \geq 0$*, the* τ*-mixing coefficients of* $(u_t u_{t+k} x_{t,j} x_{t+k,l})_{t \in \mathbf{Z}}$ *are* $\check{\tau}_t \leq ct^{-d}$ *for some universal constants* $c > 0$ *and* $d > 1$*.*

Assumption (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, e.g., in the Gaussian graphical model. Condition (iii) is a mild restriction needed for the consistency of the sample variance of regression errors. The rate imposed on the sparsity constant, $s^2 \log^2 p / T \to 0$, is also used in van de Geer, Bühlmann, Ritov, and Dezeure (2014) who assume that the regression

errors are Gaussian, see their Corollary 2.1. On the other hand, the rate condition on the dimension $p/\sqrt{T^{\kappa-2}\log^{\kappa} p} \to 0$, is additional condition needed in our setting when regression errors are not Gaussian and may only have a certain number of finite moments. 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. Conditions of this type are standard in nonparametric literature.

Let $B = \hat{\Theta}\mathbf{X}^{\top}(\mathbf{y} - \mathbf{X}\hat{\beta})/T$ denote the bias-correction for the sg-LASSO estimator, where $\hat{\Theta}$ is the nodewise LASSO estimator of the precision matrix Θ; see the following subsection for more details. The following result describes a large-sample approximation to the distribution of the debiased sg-LASSO estimator with serially correlated non-Gaussian regression errors.

Theorem 3.1. *Suppose that Assumptions 3.2.1, 3.2.2, 3.2.3, 3.2.4, and 3.2.5 are satisfied for the sg-LASSO regression and for each nodewise LASSO regression $j \in G$. Then*

$$\sqrt{T}(\hat{\beta}_G + B_G - \beta_G) \xrightarrow{d} N(0, \Xi_G)$$

with the long-run variance[7] $\Xi_G = \lim_{T\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_t \Theta_G X_t\right)$.

It is worth mentioning that since the group G has fixed size and the rows of Θ have finite ℓ_1 norm, the long-run variance Ξ_G exists under the maintained assumptions; see Proposition A3.1.1 in the Appendix for a precise statement of this result.

Theorem 3.1 extends van de Geer, Bühlmann, Ritov, and Dezeure (2014) to non-Gaussian, heavy-tailed and persistent time series data and describes the long run asymptotic variance for the low-dimensional group of regression coefficients estimated with the sg-LASSO. One could also consider Gaussian approximations for groups of increasing size, which requires an appropriate high-dimensional Gaussian approximation result for τ-mixing processes and is left for future research; see Chernozhukov, Chetverikov, Kato, et al. (2013) for a comprehensive review of related coupling results in the i.i.d. case.

Remark 3.2.1. *It is worth mentioning that the debiasing with explicit bias correction addresses the post-model selection issues, see Leeb and Pötscher (2005), and it is fairly straightforward to show that the convergence in*

[7] With slight abuse of notation we use $\beta_G \in \mathbf{R}^{|G|}$ to denote the subvector of elements of $\beta \in \mathbf{R}^p$ indexed by G.

Theorem 3.1 holds uniformly over the set of sparse vectors; see also van de Geer, Bühlmann, Ritov, and Dezeure (2014), Corollary 2.1 and the remark following that corollary.

3.2.2 Nodewise LASSO

The bias-correction term B and the expression of the long-run variance in Theorem 3.1 depend on the appropriate estimator of the precision matrix $\Theta = \Sigma^{-1}$. Following Meinshausen and Bühlmann (2006) and van de Geer, Bühlmann, Ritov, and Dezeure (2014), we focus on the nodewise LASSO estimator of Θ. The estimator is based on the observation that the covariance matrix of the partitioned vector $X = (X_j, X_{-j}^\top)^\top \in \mathbf{R} \times \mathbf{R}^{p-1}$ can be written as

$$\Sigma = \mathbb{E}[XX^\top] = \begin{pmatrix} \Sigma_{j,j} & \Sigma_{j,-j} \\ \Sigma_{-j,j} & \Sigma_{-j,-j} \end{pmatrix},$$

where $\Sigma_{j,j} = \mathbb{E}[X_j^2]$ and all other elements similarly defined. By the partitioned inverse formula, the 1^{st} row of the precision matrix $\Theta = \Sigma^{-1}$ is

$$\Theta_j = \sigma_j^{-2} \begin{pmatrix} 1 & -\gamma_j^\top \end{pmatrix},$$

where $\gamma_j = \Sigma_{-j,-j}^{-1} \Sigma_{-j,j}$ is the projection coefficient in the regression of X_j on X_{-j}

$$X_j = X_{-j}^\top \gamma_j + v_j, \qquad \mathbb{E}[X_{-j} v_j] = 0, \tag{3.2}$$

and $\sigma_j^2 = \Sigma_{j,j} - \Sigma_{j,-j}\gamma_j = \mathbb{E}[v_j^2]$ is the variance of the projection error.[8] This suggests estimating the 1^{st} row of the precision matrix as $\hat{\Theta}_j = \hat{\sigma}_j^{-2} \begin{pmatrix} 1 & -\hat{\gamma}_j^\top \end{pmatrix}$ with $\hat{\gamma}_j$ solving

$$\min_{\gamma \in \mathbf{R}^{p-1}} \|\mathbf{X}_j - \mathbf{X}_{-j}\gamma\|_T^2 + 2\lambda_j |\gamma|_1$$

and

$$\hat{\sigma}_j^2 = \|\mathbf{X}_j - \mathbf{X}_{-j}\hat{\gamma}_j\|_T^2 + \lambda_j |\hat{\gamma}_j|,$$

where $\mathbf{X}_j \in \mathbf{R}^T$ is the column vector of observations of $x_j \in \mathbf{R}$ and $\mathbf{X}_{-j}$ is the $T \times (p-1)$ matrix of observations of $x_{-j} \in \mathbf{R}^{p-1}$. In the matrix notation, the nodewise LASSO estimator of Θ can be written then as $\hat{\Theta} = \hat{B}^{-1}\hat{C}$

[8] To ensure that the projection coefficient is well defined and does not change with the dimension of the model p, we can consider the limiting linear projection model and take into account the approximation error.

with

$$\hat{C} = \begin{pmatrix} 1 & -\hat{\gamma}_{1,1} & \cdots & -\hat{\gamma}_{1,p-1} \\ -\hat{\gamma}_{2,1} & 1 & \cdots & -\hat{\gamma}_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p-1,1} & \cdots & -\hat{\gamma}_{p-1,p-1} & 1 \end{pmatrix} \quad \text{and} \quad \hat{B} = \operatorname{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2).$$

3.2.3 HAC estimator

Next, we focus on the HAC estimator based on sg-LASSO residuals, covering the LASSO and the group LASSO as special cases. For a group $G \subset [p]$ of a fixed size, the HAC estimator of the long-run variance is

$$\hat{\Xi}_G = \sum_{|k|<T} K\left(\frac{k}{M_T}\right) \hat{\Gamma}_k, \tag{3.3}$$

where $\hat{\Gamma}_k = \hat{\Theta}_G \left(\frac{1}{T}\sum_{t=1}^{T-k} \hat{u}_t \hat{u}_{t+k} x_t x_{t+k}^\top\right) \hat{\Theta}_G^\top$, $\hat{u}_t$ is the sg-LASSO residual, and $\hat{\Gamma}_{-k} = \hat{\Gamma}_k^\top$. The kernel function $K : \mathbf{R} \to [-1, 1]$ with $K(0) = 1$ is puts less weight on more distant noisy covariances, while $M_T \uparrow \infty$ is a bandwidth (or lag truncation) parameter, see Parzen (1957), Newey and West (1987), and Andrews (1991). Several choices of the kernel function are possible, for example, the Parzen kernel is

$$K_{PR}(x) = \begin{cases} 1 - 6x^2 + 6|x|^3 & \text{for } 0 \le |x| \le 1/2, \\ 2(1 - |x|)^3 & \text{for } 1/2 \le |x| \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

It is worth recalling that the Parzen and the Quadratic spectral kernels are high-order kernels that superior to the Bartlett kernel, cf. Newey and West (1987); see appendix for more details on the choice of the kernel.

Note that under stationarity, the long-run variance in Theorem 3.1 simplifies to

$$\Xi_G = \sum_{k \in \mathbf{Z}} \Gamma_k,$$

where $\Gamma_k = \Theta_G \mathbb{E}[u_t x_t u_{t+k} x_{t+k}^\top] \Theta_G^\top$ and $\Gamma_{-k} = \Gamma_k^\top$. The following result characterizes the convergence rate of the HAC estimator pertaining to a group of regression coefficients $G \subset [p]$ based on the sg-LASSO residuals.

Theorem 3.2. *Suppose that Assumptions 3.2.1, 3.2.2, 3.2.3, 3.2.4, 3.2.5 are satisfied for the sg-LASSO regression and for each nodewise LASSO*

regression $j \in G$. Suppose also that Assumptions A3.1.1, and A3.1.2 in the Appendix are satisfied for $V_t = (u_t v_{t,j}/\sigma_j^2)_{j\in G}$, $\kappa \geq \tilde{q}$ and that $s^\kappa p T^{1-4\kappa/5} \to 0$ as $M_T \to \infty$ and $T \to \infty$. Then

$$\|\hat{\Xi}_G - \Xi_G\| = O_P\left(M_T\left(\frac{sp^{1/\kappa}}{T^{1-1/\kappa}} \vee s\sqrt{\frac{\log p}{T}} + \frac{s^2p^{2/\kappa}}{T^{2-3/\kappa}} + \frac{s^3p^{5/\kappa}}{T^{4-5/\kappa}}\right) + M_T^{-\varsigma} + T^{-(\varsigma\wedge 1)}\right).$$

The first term in the inner parentheses is of the same order as the estimation error of the maximum between the estimation errors of the sg-LASSO and the nodewise LASSO. Theorem 3.2 suggests that the optimal choice of the bandwidth parameter should scale appropriately with the number of covariates p, the sparsity constant s, and the dependence-tails exponent κ.[9] This contrasts sharply with the HAC theory for regressions without regularization developed in Andrews (1991), see also Li and Liao (2020), and allows for faster convergence rates of the HAC estimator.

3.2.4 High-dimensional Granger causality tests

Consider a linear projection model

$$y_{t+h} = \sum_{j\in G}\beta_j x_{t,j} + \sum_{j\in G^c}\beta_j x_{t,j} + u_t, \qquad \mathbb{E}[u_t x_{t,j}] = 0, \qquad \forall j \geq 1,$$

where $h \geq 0$ is the horizon, $G \subset [p]$ is a group of regression coefficients of interest, $x_t = \{x_{t,j} : j \in G\}$ represents the series for which we wish to test the Granger causality, and $\{x_{t,j} : j \in G^c\}$ represents all the remaining information available at time t. For instance, x_t may contain L low-frequency lags of some series $(z_t)_{t\in\mathbf{Z}}$, in which case $x_t = (z_t, z_{t-1}, z_{t-2}, \ldots, z_{t-L})^\top$. Alternatively, it may contain low and/or high-frequency lags of $(z_t)_{t\in\mathbf{Z}}$ aggregated with dictionaries, e.g., Legendre polynomials as in Babii, Ghysels, and Striaukas (2020b). In both cases the dimensionality of x_t small. On the other hand, the set of controls representing all the information available at time t is high-dimensional. The Granger causality test corresponds to the following hypotheses

$$H_0 : R\beta_G = 0 \qquad \text{against} \qquad H_1 : R\beta_G \neq 0,$$

where $\beta_G = \{\beta_j : j \in G\}$ and R is $r \times |G|$ matrix of linear restrictions imposed on β_G.

[9] A comprehensive study of the optimal bandwidth choice based on higher-order asymptotic expansions is beyond the scope of this paper and is left for future research, see, e.g., Lazarus, Lewis, Stock, and Watson (2018) for the recent literature review and practical recommendations in the low-dimensional case.

It is worth mentioning our framework is based on the weakest notion of the Granger causality corresponding to the marginal improvement in time series projections due to the information contained in x_t. A stronger notion of Granger non-causality appears when projections are replaced by conditional means, so that the conditional mean y_t given x_t and all other available information does not depend on x_t. Yet, even stronger version of Granger non-causality pertains to the full conditional independence; see Florens and Mouchart (1982).

For the Granger causality test, we set $R = I_{|G|}$, but more generally, we might be interested in testing other linear restrictions implied by the economic theory. Assuming that R is a full row rank matrix, consider the debiased Wald statistics

$$W_T = T\left[R(\hat{\beta}_G + B_G - \beta_G)\right]^\top \left(R\hat{\Xi}_G R^\top\right)^+ \left[R(\hat{\beta}_G + B_G - \beta_G)\right],$$

where A^+ is the generalized inverse of A. It follows from Theorems 3.1 and 3.2 that under H_0, $W_T \xrightarrow{d} \chi^2_r$. The Wald test rejects when $W_T > q_{1-\alpha}$, where $q_{1-\alpha}$ is the quantile of order $1-\alpha$ of χ^2_r. More generally, the linear restrictions can be extended to the nonlinear restrictions by the usual Delta method argument.

For testing hypotheses on the increasing set of regression coefficients, it might be preferable to use the non-pivotal sup-norm based statistics, see Ghysels, Hill, and Motegi (2020), due to the remarkable performance in the high-dimensional setting; see Chernozhukov, Chetverikov, Kato, et al. (2013) for high-dimensional Gaussian approximations with i.i.d. data.

3.3 Fuk-Nagaev inequality

In this section, we describe a suitable for us version of the Fuk-Nagaev concentration inequality for the maximum of high-dimensional sums. The inequality allows for the data with polynomial tails and τ-mixing coefficients decreasing at a polynomial rate. The following result does not require that the series is stationary.

Theorem 3.1. *Let $(\xi_t)_{t\in\mathbf{Z}}$ be a centered stochastic process in $\mathbf{R}^p$ such that (i) for some $q > 2$, $\max_{j\in[p],t\in[T]} \|\xi_{t,j}\|_q = O(1)$; (ii) for every $j \in [p]$, τ-mixing coefficients of $\xi_{t,j}$ satisfy $\tau_k^{(j)} \le ck^{-a}$ for some universal constants*

$a, c > 0$. Then there exist $c_1, c_2 > 0$ such that for every $u > 0$

$$\Pr\left(\left|\sum_{t=1}^{T}\xi_t\right|_\infty > u\right) \leq c_1 pTu^{-\kappa} + 4p\exp\left(-\frac{c_2u^2}{B_T^2}\right),$$

where[10] $\kappa = ((a+1)q-1)/(a+q-1)$, $B_T^2 = \max_{j\in[p]}\sum_{t=1}^{T}\sum_{k=1}^{T}|\mathrm{Cov}(\xi_{t,j},\xi_{k,j})|$.

The inequality describes the mixture of the polynomial and Gaussian tails for the maximum of high-dimensional sums. In the limiting case of the i.i.d. data, as $a \to \infty$, the dependence-tails exponent κ approaches q and we recover the inequality for the independent data stated in Fuk and Nagaev (1971), Corollary 4 for $p = 1$. In this sense, the inequality in Theorem 3.1 is sharp. It is well-known that the Fuk-Nagaev inequality delivers sharper estimates of tail probabilities in contrast to Markov's bound in conjunction with Rosenthal's moment inequality, cf. Nagaev (1998). The proof relies on the blocking technique, see Bosq (1993), and the coupling inequality for τ-mixing sequences, see Dedecker and Prieur (2004), Lemma 5. In contrast to previous results, e.g., Dedecker and Prieur (2004), Theorem 2, the inequality reflects the mixture of the polynomial and the exponential tails.

For stationary processes, by Lemma A3.1.2 in the appendix, $B_T^2 = O(T)$ as long as $a > (q-1)/(q-2)$, whence we obtain from Theorem 3.1 that for every $\delta \in (0,1)$

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\xi_t\right|_\infty \leq C\left(\frac{p}{\delta T^{\kappa-1}}\right)^{1/\kappa} \vee \sqrt{\frac{\log(8p/\delta)}{T}}\right) \geq 1-\delta, \tag{3.4}$$

where $C > 0$ is some finite universal constant.

3.4 Monte Carlo experiments

In this section, we aim to assess the debiased HAC-based inferences for the low-dimensional parameter in a high dimensional data setting. To that end, we draw covariates $\{x_{t,j}, j \in [p]\}$ independently from the AR(1) process

$$x_{t,j} = \rho x_{t-1,j} + \epsilon_{t,j}.$$

[10] It is worth mentioning that the notation in this section is specific to generic stochastic processes and is independent from the rest of the paper. Thus B_T here denotes the variance of partial sums and not the bias correction term of the LASSO estimator.

The regression error follows the AR(1) process

$$u_t = \rho u_{t-1} + \nu_t,$$

where errors are $\epsilon, \nu \sim_{i.i.d.} N(0,1)$. The vector of population regression coefficients β has the first five non-zero entries which are drawn from Uniform$(0,4)$ and all remaining entries are zero. The sample size is $T \in \{100, 1000\}$ and the number of covariates is $p \in \{10, 200\}$. We set the persistence parameter $\rho = 0.6$ and focus on the LASSO estimator to estimate coefficients $\hat{\beta}$. Throughout the experiment, we choose the LASSO tuning parameters using the 10-fold cross-validation, defining folds as adjacent over time blocks.

We report the average coverage (av. cov) and the average length of confidence intervals for the nominal coverage of 0.95 and on a grid of values of the bandwidth parameter $M_T \in \{5, 10, \ldots, 40\}$, using the Parzen kernel. We estimate the long run covariance matrix $\hat{\Xi}$ using the LASSO residuals, denoted $\hat{u}_t$. We also use the nodewise LASSO regressions to estimate the precision matrix Θ. The first step is to compute scores $\hat{V}_t = \hat{u}_t x_t$, where $\hat{u}_t = y_t - x_t^\top \hat{\beta}$, and $\hat{\beta}$ is the LASSO estimator. Then we compute the high-dimensional HAC estimator using the formuala in equation ((3.3)). We compute the pivotal statistics for each MC experiment $i \in [N]$ and each coefficient $j \in [p]$ as $\text{pivot}_j^{(i)} \triangleq (\hat{\beta}_j^{(i)} + B_j^{(i)} - \beta)/\sqrt{\hat{\Xi}_{j,j}^{(i)}/T}$, where $B_j^{(i)} = \hat{\Theta}_j^{(i)} \mathbf{X}^{\top(i)} \hat{\mathbf{u}}^{(i)}/T$, and $\hat{\mathbf{u}}^{(i)} = \mathbf{y}^{(i)} - \mathbf{X}^{(i)} \hat{\beta}^{(i)}$. Then we compute the empirical coverage as

$$\text{av.cov}_j = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\text{pivot}_j^i \in [-1.96, 1.96]\}$$

and the average confidence interval length as $\text{length}_j = \frac{1}{N} \sum_{i=1}^{N} 2 \times 1.96 \times \sqrt{\hat{\Xi}_{j,j}^{(i)}/T}$. The number of Monte Carlo experiments is set to N= 5000.

We report average results over the active and inactive sets of the vector of coefficients. Table 4.2 shows results for the small sample size ($T = 100$) and the large sample size ($T = 1000$). We find that the value of the bandwidth parameter M_T should be smaller when the number of regressors p is larger. For $T = 100$ (small sample size), for the active set of coefficients, the best coverage is achieved when the bandwidth parameter is set at 10 when $p = 10$ and at 5 when $p = 200$. Results for the inactive set and $T = 1000$ are similar. We also see that the increase in p relative to T leads to worse performance. Furthermore, the coverage improves when

Table 3.1: HAC-based inference simulation results – The table reports average coverage (first four columns) and average length of confidence intervals (last four columns) for active and inactive sets of β and for $T = 100$ and $T = 1000$. We report results for a set of bandwidth parameter M_T values. The data is generated from Gaussian distribution.

	Average coverage (av. cov)				Confidence interval length			
	Active set of β		Inactive set of β		Active set of β		Inactive set of β	
$M_T \backslash p$	10	200	10	200	10	200	10	200
				T=100				
5	0.830	0.755	0.827	0.747	0.286	0.382	0.277	0.368
10	0.834	0.750	0.835	0.746	0.305	0.401	0.291	0.376
15	0.824	0.753	0.832	0.745	0.314	0.411	0.294	0.376
20	0.821	0.735	0.826	0.743	0.320	0.420	0.296	0.374
25	0.816	0.739	0.820	0.741	0.325	0.427	0.297	0.372
30	0.807	0.738	0.814	0.740	0.329	0.434	0.297	0.370
35	0.806	0.733	0.807	0.738	0.333	0.441	0.297	0.368
40	0.801	0.737	0.802	0.735	0.337	0.447	0.297	0.366
				T=1000				
5	0.913	0.848	0.913	0.879	0.081	0.067	0.081	0.067
10	0.932	0.865	0.932	0.894	0.087	0.070	0.087	0.070
15	0.935	0.866	0.936	0.894	0.088	0.070	0.088	0.070
20	0.936	0.868	0.936	0.897	0.089	0.071	0.088	0.071
25	0.936	0.867	0.936	0.895	0.089	0.071	0.089	0.071
30	0.937	0.866	0.937	0.895	0.089	0.071	0.089	0.070
35	0.936	0.866	0.936	0.895	0.089	0.071	0.089	0.070
40	0.934	0.865	0.934	0.894	0.089	0.071	0.088	0.070

the bandwidth increases with the sample size. Lastly, as the sample size increases, the average coverage approaches the nominal level of 0.95 and the confidence interval shrinks in size. Overall, the simulation results confirm our theoretical findings.

3.5 Testing Granger causality for VIX and financial news

The CBOE Volatility Index, known as the VIX, is a popular measure of market-based expectation of future volatility and is often referred to as the "fear index". The VIX index quotes the expected annualized change in the S&P 500 index over the following 30 days, as computed from options-based theory and current options-market data.

There is a large literature studying the theoretical and empirical properties of the VIX and it is impossible to cite only a few papers to do justice to all the outstanding research output on the topic. Focusing on Granger causal patterns, there are several studies pertaining to causality between the VIX and VIX futures. For example, Bollen, O'Neill, and Whaley (2017) suggest that the VIX futures lagged the VIX in the first few years after its introduction, and show an increasing dominance of VIX futures over time. Along similar lines, Shu and Zhang (2012) study price-discovery between VIX futures the spot VIX index and find evidence of a bi-directional causal pattern.

We study the causal relationship between financial news and the VIX. There is also substantial literature on the impact of news releases on financial markets (e.g., Andersen, Bollerslev, Diebold, and Vega (2003)). Traditionally, such analysis looks at news releases and studies the behavior of asset prices pre- and post-release. News is usually quantified numerically via the surprise component measured as the difference between an expectation prior to the release and the announcement. In the age of machine learning, the characterization of news has been expanded into the textual analysis of news coverage. To paraphrase the title of Gentzkow, Kelly, and Taddy (2019), the text is treated as data. It is in this spirit that we conduct our high-dimensional Granger causality analysis between the VIX and news.

We use a data set from Bybee, Kelly, Manela, and Xiu (2020) which contains 180 news attention monthly series, all of which potentially Granger cause future US equity market volatility.[11] We estimate the following time series regression model

[11] We downloaded daily VIX data from St. Louis Fed FRED database and took the end-of-month values. The FRED mnemonic for the VIX is VIXCLS. Table with the full list of series appears in Appendix A3.1.

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

where y_{t+1} is the value of the VIX at the end of month $t+1$, $\psi(L^{1/m}; \beta) y_t$ is a MIDAS polynomial of 22 daily VIX lags where the first lag is the last day of the month t, and $x_{t,k}$ is the k-th news attention series. Note that we only take one lag for the news attention series to simplify the model (and also an empirically justified simplification). The MIDAS polynomial of daily lags of the VIX involves Legendre polynomials of degree 3. Note that the specification is what is sometimes called a reverse MIDAS regression as mentioned earlier in the paper. Prior to estimating the regression model, we time demean the response and covariates such that the intercept is zero. We further standardize all covariates to have a unit standard deviation. The daily VIX lags are standardized before the aggregation.

We apply the sg-LASSO estimator to estimate the slope coefficients and nodewise LASSO regressions to estimate the precision matrix. To fully exploit the group sparsity of sg-LASSO, we group all high-frequency lags of daily VIX, see Babii, Ghysels, and Striaukas (2020b) for further details on such grouping. The news attention series are monthly and we are interested in whether the most recent news Granger causes the VIX, hence we don't apply the group structure along the time dimension. Instead, we group news attention series based on a broader theme that is available for each series, see Bybee, Kelly, Manela, and Xiu (2020) for further details. Namely, the data set contains 24 broader topics which group each of the 180 news attention series.

3.5.1 Main results

We report the p-values for a range of M_T values for series that appear to be significant at the 1% or 5% significance level for all $M_T \in \{20, 40, 60\}$ values and for two kernel functions, namely Parzen and Quadratic Spectral. The sample starts January 1990 January and ends June 2017, determined by the availability of the textual analysis data. Both sg-LASSO and LASSO tuning parameters are selected via 10-fold cross-validation, defining folds as adjacent blocks over the time dimension to take into account the time series nature of the data. Similarly, we tune nodewise LASSO regressions for the precision matrix estimation.

Table 3.2: VIX Granger causality results. We report p-values of series that are significant at 1% and 5% significance level for a range of M_T values and both kernel functions.

Variable \ M_T	20	40	60	20	40	60
			sg-LASSO			
		Parzen			Quadratic Spectral	
			1% significance			
Daily VIX lags	0.000	0.000	0.001	0.000	0.001	0.002
Financial crisis	0.005	0.002	0.001	0.002	0.001	0.000
			5% significance			
Aerospace/defense	0.014	0.014	0.017	0.012	0.018	0.027
Recession	0.011	0.008	0.009	0.008	0.008	0.013
			LASSO			
		Parzen			Quadratic Spectral	
			1% significance			
Daily VIX lags	0.000	0.000	0.000	0.000	0.000	0.000
Financial crisis	0.001	0.000	0.000	0.000	0.000	0.000
Recession	0.003	0.002	0.002	0.002	0.002	0.004
Marketing	0.001	0.001	0.000	0.001	0.000	0.000
			5% significance			
Aerospace/defense	0.007	0.006	0.004	0.006	0.004	0.002
NY politics	0.012	0.015	0.013	0.016	0.012	0.008
Acquired investment banks	0.018	0.006	0.002	0.007	0.001	0.000

3.5.1.1 Granger causality of news topics

The results appear in Table 3.2 which contains two main row blocks reporting results for the structured sg-LASSO and unstructured LASSO estimators, and two-column blocks, reporting results for two kernel functions. Irrespective of the initial estimator and kernel function, the lagged daily VIX and the Financial crisis news series are highly significant at 1% significance level. Comparing results for the initial estimator, in the case of LASSO we see more significant predictors than for the sg-LASSO case, while the subset of significant covariates using sg-LASSO is a subset of the LASSO significant predictors. Many more series are selected by the initial LASSO estimator compared to the sg-LASSO, see Table A3.1. This suggests that relevant group structures are important, and may help in recovering salient relationships in the data.

Table 3.3: Bi-directional Granger causality results. We report p-values for a range of M_T values and both kernel functions.

Variable $\setminus M_T$	20	40	60	20	40	60
			sg-LASSO			
		Parzen			Quadratic Spectral	
Daily VIX lags	0.050	0.071	0.091	0.060	0.086	0.129

Table 3.4: Group Granger causality results. We report p-values for a range of M_T values and both kernel functions.

Variable $\setminus M_T$	20	40	60	20	40	60
			sg-LASSO			
		Parzen			Quadratic Spectral	
Banks	0.032	0.024	0.008	0.023	0.001	0.000

3.5.1.2 Bi-directional Granger causality

We also test whether the daily VIX Granger causes Financial crisis news series. For this we run the following MIDAS regression model

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

where $x_{t+1,j}$ is the Financial crisis news series. We test whether daily VIX Granger causes future values of Financial crisis news series. Note that we only need to estimate the initial initial coefficient vector, since the precision matrix remains the same. The results appear in Table 3.3. They show a rather weak predictability of future news series by daily VIX suggesting a unidirectional Granger causality pattern.

3.5.1.3 Granger causal clusters of news topics

The news attention series are classified into 24 broader meta topics that group the individual news series according to a common theme. We test which group of individual news series Granger causes future VIX values. The results are reported in Table 3.4. They show that the group *Banks* is significant at 5% significance level. This group consists of news series pertaining to news about Mortgages, Bank loans, Credit ratings, Nonperforming loans, Savings & loans, and the Financial crisis.

3.6 Conclusion

This paper develops valid inferential methods for high-dimensional time series regressions estimated with the sparse-group LASSO (sg-LASSO) estimator that encompasses the LASSO and the group LASSO as special cases. We derive the debiased central limit theorem with the explicit bias correction for the sg-LASSO with serially correlated regression errors. Furthermore, we also study HAC estimators of the long-run variance for low dimensional groups of regression coefficients and characterize how the optimal bandwidth parameter should scale with the sample size, the temporal dependence, as well as tails of the data. These results lead to the valid t- and Wald tests for the low-dimensional subset of parameters, such as Granger causality tests. Our treatment relies on a new suitable variation of the Fuk-Nagaev inequality for τ-mixing processes which allows us to handle the time series data with polynomial tails. An interesting avenue for future research is to study more carefully the problem of the optimal data-driven bandwidth choice based on higher-order asymptotic expansions, see, e.g., Sun, Phillips, and Jin (2008) for steps in this direction in low dimensional settings.

In an empirical application we use a high-dimensional news attention series to study causal patterns between the VIX, sometimes called the fear index, and financial news. We find that almost exclusively the topic of financial crisis exhibits unidirectional Granger causality for the VIX.

APPENDIX

A3.1 Proofs

Proof of Theorem 3.1. By Fermat's rule, the sg-LASSO satisfies

$$\mathbf{X}^\top(\mathbf{X}\hat{\beta} - \mathbf{y})/T + \lambda z^* = 0$$

for some $z^* \in \partial\Omega(\hat{\beta})$, where $\partial\Omega(\hat{\beta})$ is the sub-differential of $b \mapsto \Omega(b)$ at $\hat{\beta}$. Rearranging this expression and multiplying by $\hat{\Theta}$

$$\hat{\beta} - \beta + \hat{\Theta}\lambda z^* = \hat{\Theta}\mathbf{X}^\top\mathbf{u}/T + (I - \hat{\Theta}\hat{\Sigma})(\hat{\beta} - \beta) + \hat{\Theta}\mathbf{X}^\top(\mathbf{m} - \mathbf{X}\beta)/T,$$

where we use $\mathbf{y} = \mathbf{m} + \mathbf{u}$. Plugging in λz^* and multiplying by $\sqrt{T}$

$$\begin{aligned}\sqrt{T}(\hat{\beta} - \beta + B) &= \hat{\Theta}\mathbf{X}^\top\mathbf{u}/\sqrt{T} + \sqrt{T}(I - \hat{\Theta}\hat{\Sigma})(\hat{\beta} - \beta) \\ &\quad + \hat{\Theta}\mathbf{X}^\top(\mathbf{m} - \mathbf{X}\beta)/\sqrt{T} \\ &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_t\Theta x_t \\ &\quad + \frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_t(\hat{\Theta} - \Theta)X_t + \sqrt{T}(I - \hat{\Theta}\hat{\Sigma})(\hat{\beta} - \beta) \\ &\quad + \hat{\Theta}\mathbf{X}^\top(\mathbf{m} - \mathbf{X}\beta)/\sqrt{T}.\end{aligned}$$

Next, we look at coefficients corresponding to $G \subset [p]$

$$\begin{aligned}\sqrt{T}(\hat{\beta}_G - \beta_G + B_G) &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_t\Theta_G x_t + \frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_t(\hat{\Theta}_G - \Theta_G)x_t \\ &\quad + \sqrt{T}(I - \hat{\Theta}\hat{\Sigma})_G(\hat{\beta} - \beta) + \hat{\Theta}_G\mathbf{X}^\top(\mathbf{m} - \mathbf{X}\beta)/\sqrt{T} \\ &\triangleq I_T + II_T + III_T + IV_T.\end{aligned}$$

We will show that $I_T \xrightarrow{d} N(0, \Xi_G)$ by the triangular array CLT, see Neumann (2013), Theorem 2.1. To that end, by the Crámer-Wold theorem, it is sufficient to show that $z^\top I_T \xrightarrow{d} z^\top N(0, \Xi_G)$ for every $z \in \mathbf{R}^{|G|}$. Note that under Assumptions 3.2.1 and 3.2.5 (i)-(ii)

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left|\frac{z^\top\xi_t}{\sqrt{T}}\right|^2 &= \mathbb{E}|u_t z^\top\Theta_G x_t|^2 \\ &\leq Cz^\top\Theta_G\Sigma\Theta_G^\top z \\ &= O(1).\end{aligned}$$

Therefore, since $q > 2r/(r-2)$, we have $\varsigma > 2$, and for every $\epsilon > 0$

$$\sum_{t=1}^{T} \mathbb{E}\left[\left|\frac{z^\top \xi_t}{\sqrt{T}}\right|^2 \mathbf{1}\left\{\left|z^\top \xi_t\right| > \epsilon\sqrt{T}\right\}\right] \leq \frac{\mathbb{E}\left|z^\top \xi_t\right|^\varsigma}{(\epsilon\sqrt{T})^{\varsigma-2}} = o(1).$$

Next, under Assumptions 3.2.1 and 3.2.5 (i)-(ii), the long run variance

$$\lim_{T\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} z^\top \xi_t\right) = z^\top \Xi_G z$$

exists by Proposition A3.1.1.

Next, put $\mathcal{M} = \sigma(\xi_0, \xi_{-1}, \xi_{-2}, \dots)$, $Y = g(z^\top \xi_{t_1-t_h}/\sqrt{T}, \dots, z^\top \xi_0/\sqrt{T}) z^\top \xi_0$, and $X = z^\top \xi_r$ for some $t_h \geq 0$. Note that X and $|XY|$ are integrable and that Y is $\mathcal{M}$-measurable. Therefore, for every measurable function $g : \mathbf{R}^h \to \mathbf{R}$ with $\sup_x |g(x)| \leq 1$, by Dedecker and Doukhan (2003), Proposition 1, for all $h \in \mathbf{N}$ and all indices $1 \leq t_1 < t_2 < \cdots < t_h < t_h + r \leq t_h + s \leq T$

$$\begin{aligned}
&\left|\mathrm{Cov}\left(g(z^\top \xi_{t_1}/\sqrt{T}, \dots, z^\top \xi_{t_h}/\sqrt{T}) z^\top \xi_{t_h}/\sqrt{T}, z^\top \xi_{t_h+r}/\sqrt{T}\right)\right| \\
&= \frac{1}{T}\left|\mathrm{Cov}\left(Y, X\right)\right| \\
&\leq \frac{1}{T}\int_0^{\gamma(\mathcal{M}, z^\top \xi_r)} Q_Y \circ G_{z^\top \xi_r}(u)\mathrm{d}u \\
&\leq \frac{1}{T}\int_0^{\gamma(\mathcal{M}, z^\top \xi_r)} Q_{z^\top \xi_0} \circ G_{z^\top \xi_r}(u)\mathrm{d}u \\
&\leq \frac{1}{T}\|\mathbb{E}(z^\top \xi_r|\mathcal{M}) - \mathbb{E}(z^\top \xi_r)\|_1^{\frac{\varsigma-2}{\varsigma-1}} \|z^\top \xi_0\|_\varsigma^{\varsigma/(\varsigma-1)} \\
&\leq \frac{1}{T}|\Theta_G^\top z|_1^{\frac{\varsigma-2}{\varsigma-1}} \tau_r^{\frac{\varsigma-2}{\varsigma-1}} \|z^\top \xi_0\|_\varsigma^{\varsigma/(\varsigma-1)} \lesssim r^{-a\frac{\varsigma-2}{\varsigma-1}}
\end{aligned}$$

where the second line follows by stationarity and $\sup_x |g(x)| \leq 1$, the fourth by Hölder's inequality and the change of variables

$$\int Q_{z^\top \xi_0}^{\varsigma-1} \circ G_{z^\top \xi_r}(u)\mathrm{d}u = \int_0^1 Q_{z^\top \xi_0}^{\varsigma}(u)\mathrm{d}u = \|z^\top \xi_0\|_\varsigma^\varsigma,$$

and the last by Lemma A3.1.1 and Assumptions 3.2.1 (ii) and 3.2.5 (ii).

Similarly,

$$\begin{aligned}
&\left|\mathrm{Cov}\left(g(z^\top\xi_{t_1}/\sqrt{T},\ldots,z^\top\xi_{t_h}/\sqrt{T}),z^\top\xi_{t_h+r}/\sqrt{T}z^\top\xi_{t_h+s}/\sqrt{T}\right)\right| \\
&= \frac{1}{T}\left|\mathrm{Cov}\left(g(z^\top\xi_{t_1-t_h}/\sqrt{T},\ldots,z^\top\xi_0/\sqrt{T}),z^\top\xi_r z^\top\xi_s\right)\right| \\
&\leq \frac{1}{T}\int_0^{\gamma(\mathcal{M},z^\top\xi_r z^\top\xi_s)} Q_g\circ G_{z^\top\xi_r z^\top\xi_s}(u)\mathrm{d}u \\
&\leq \frac{1}{T}\left\|\mathbb{E}(z^\top\xi_r z^\top\xi_s|\mathcal{M})-\mathbb{E}(z^\top\xi_r z^\top\xi_s)\right\|_1 \\
&\leq \frac{1}{T}|\Theta_G^\top z|_1^2\check{\tau}_r \lesssim r^{-d}.
\end{aligned}$$

Since the sequence $(r^{-a(\varsigma-2)/(\varsigma-1)\wedge d})_{r\in\mathbf{N}}$ is summable under Assumption 3.2.1 (ii), all conditions of Neumann (2013), Theorem 2.1, are verified, whence $z^\top I_T \xrightarrow{d} z^\top N(0,\Xi_G)$ for every $z\in\mathbf{R}^{|G|}$.

Next,

$$\begin{aligned}
|II_T|_\infty &= \left|(\hat{\Theta}-\Theta)_G\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T u_t x_t\right)\right|_\infty \\
&\leq \|\hat{\Theta}_G-\Theta_G\|_\infty\left|\frac{1}{\sqrt{T}}\sum_{t=1}^T u_t x_t\right|_\infty \\
&= O_P\left(\frac{Sp^{1/\kappa}}{T^{1-1/\kappa}}\vee S\sqrt{\frac{\log p}{T}}\right)O_P\left(\frac{p^{1/\kappa}}{T^{1/2-1/\kappa}}+\sqrt{\log p}\right) \\
&= o_P(1),
\end{aligned}$$

where the second line follows by $|Ax|_\infty\leq\|A\|_\infty|x|_\infty$, the third line by Proposition A3.1.3 and the inequality in Eq. (3.4) under Assumption 3.2.1, and the last under Assumption 3.2.5 (iv). Likewise, using $|Ax|_\infty\leq\max_{j,k}|A_{j,k}|_\infty|x|_1$, by Proposition A3.1.3 and Theorem A3.1

$$\begin{aligned}
|III_T|_\infty &= \sqrt{T}|(I-\hat{\Theta}\hat{\Sigma})_G(\hat{\beta}-\beta)|_\infty \\
&\leq \sqrt{T}\max_{j\in G}|(I-\hat{\Theta}\hat{\Sigma})_j|_\infty|\hat{\beta}-\beta|_1 \\
&= O_P\left(\frac{p^{1/\kappa}}{T^{1/2-1/\kappa}}\vee\sqrt{\log p}\right)O_P\left(\frac{s_\alpha p^{1/\kappa}}{T^{1-1/\kappa}}\vee s_\alpha\sqrt{\frac{\log p}{T}}\right) \\
&= o_P(1)
\end{aligned}$$

under Assumption 3.2.5 (iv). Lastly, by the Cauchy-Schwartz inequality, under Assumption 3.2.5 (v)

$$\begin{aligned}|IV_T|_\infty &\leq \max_{j\in G} |\mathbf{X}\hat{\Theta}_j^\top|_2 \|\mathbf{m} - \mathbf{X}\beta\|_T \\ &= \max_{j\in G} \sqrt{\hat{\Theta}_j \hat{\Sigma} \hat{\Theta}_j^\top}\, o_P(1) \\ &= o_P(1),\end{aligned}$$

where the last line follows since $\hat{\Theta}_j$ are consistent for Θ_j in the ℓ_1 norm while $\hat{\Sigma}$ is consistent for Σ in the entrywise maximum norm under the maintained assumptions. □

Next, we focus on the HAC estimator based on LASSO residuals. Note that by construction of the precision matrix $\hat{\Theta}$, its j^{th} row is $\hat{\Theta}_j x_t = \hat{v}_{t,j}/\hat{\sigma}_j^2$, where $\hat{v}_{t,j}$ is the regression residual from the j^{th} nodewise LASSO regression and $\hat{\sigma}_j^2$ is the corresponding estimator of the variance of the regression error. Therefore, the HAC estimator based on the LASSO residuals in Eq. (3.3) can be written as

$$\hat{\Xi}_G = \sum_{|k|<T} K\left(\frac{k}{M_T}\right) \hat{\Gamma}_k,$$

where $\hat{\Gamma}_k$ has generic (j,h)-entry $\frac{1}{T}\sum_{t=1}^{T-k} \hat{u}_t \hat{u}_{t+k} \hat{v}_{t,j} \hat{v}_{t+k,h} \hat{\sigma}_j^{-2} \hat{\sigma}_h^{-2}$.

Similarly, we define

$$\tilde{\Xi}_G = \sum_{|k|<T} K\left(\frac{k}{M_T}\right) \tilde{\Gamma}_k,$$

where $\tilde{\Gamma}_k$ has generic (j,h)-entry $\frac{1}{T}\sum_{t=1}^{T-k} u_t u_{t+k} v_{t,j} v_{t+k,h} \sigma_j^{-2} \sigma_h^{-2}$ and note that the long-run variance Ξ_G has generic (j,h)-entry $\mathbb{E}[u_t u_{t+k} v_{t,j} v_{t+k,h}] \sigma_j^{-2} \sigma_h^{-2}$.

Assumption A3.1.1. *Suppose that uniformly over $k \in \mathbf{Z}$ and $j,h \in G$ (i) $\mathbb{E}|u_0 u_k v_{0,j} v_{k,h}| < \infty$; (ii) $\mathbb{E}|v_{0,j} u_k v_{k,h}|^2 < \infty$, $\mathbb{E}|u_0 u_k v_{k,h}|^2 < \infty$, $\mathbb{E}|u_0 v_{0,j} u_k|^2 < \infty$, and $\mathbb{E}|u_0 v_{0,j} v_{k,h}|^2 < \infty$; (iii) $\mathbb{E}|u_0|^{2q} < \infty$ and $\mathbb{E}|v_{0,j}|^{2q} < \infty$ for some $q \geq 1$.*

Proof of Theorem 3.2. By Proposition A3.1.4 with $V_t = (u_t v_{t,j}/\sigma_j^2)_{j\in G}$

$$\|\hat{\Xi}_G - \Xi_G\| \leq \|\hat{\Xi}_G - \tilde{\Xi}_G\| + O_P\left(\sqrt{\frac{M_T}{T}} + M_T^{-\varsigma} + T^{-(\varsigma\wedge 1)}\right). \tag{A3.1}$$

Next,

$$\begin{aligned}\|\hat{\Xi}_G - \tilde{\Xi}_G\| &\leq \sum_{|k|<T} \left|K\left(\frac{k}{M_T}\right)\right| \|\hat{\Gamma}_k - \tilde{\Gamma}_k\| \\ &\leq |G| \sum_{|k|<T} \left|K\left(\frac{k}{M_T}\right)\right| \max_{j,h\in G} \left|\frac{1}{\hat{\sigma}_j^2\hat{\sigma}_h^2 T}\sum_{t=1}^{T-k} \hat{u}_t\hat{u}_{t+k}\hat{v}_{t,j}\hat{v}_{t+k,h} - \frac{1}{\sigma_j^2\sigma_h^2 T}\sum_{t=1}^{T-k} u_t u_{t+k} v_{t,j} v_{t+k,h}\right| \\ &\leq |G| \sum_{|k|<T} \left|K\left(\frac{k}{M_T}\right)\right| \max_{j,h\in G} \frac{1}{\hat{\sigma}_j^2\hat{\sigma}_h^2} \left|\frac{1}{T}\sum_{t=1}^{T-k} \hat{u}_t\hat{u}_{t+k}\hat{v}_{t,j}\hat{v}_{t+k,h} - \frac{1}{T}\sum_{t=1}^{T-k} u_t u_{t+k} v_{t,j} v_{t+k,h}\right| \\ &\quad + |G| \max_{j,h\in G}\left|\frac{1}{\hat{\sigma}_j^2\hat{\sigma}_h^2} - \frac{1}{\sigma_j^2\sigma_h^2}\right| \sum_{|k|<T}\left|K\left(\frac{k}{M_T}\right)\right| \left|\frac{1}{T}\sum_{t=1}^{T-k} u_t u_{t+k} v_{t,j} v_{t+k,h}\right| \\ &\triangleq S_T^a + S_T^b.\end{aligned}$$

By Proposition A3.1.2, since $s_\alpha^2 \log p/T \to 0$ and $s_\alpha^\kappa p/T^{4\kappa/5-1} \to 0$, under stated assumptions, we obtain $\max_{j\in G}|\hat{\sigma}_j^2 - \sigma_j^2| = o_P(1)$, and whence $\max_{j\in G}\hat{\sigma}_j^{-2} = O_P(1)$. Using $\hat{a}\hat{b} - ab = (\hat{a}-a)b + a(\hat{b}-b) + (\hat{a}-a)(\hat{b}-b)$, by Proposition A3.1.2

$$S_T^b = O_P\left(\frac{s_\alpha p^{1/\kappa}}{T^{1-1/\kappa}} \vee s_\alpha\sqrt{\frac{\log p}{T}}\right) \sum_{|k|<T}\left|K\left(\frac{k}{M_T}\right)\right| \max_{j,h\in G}\left|\frac{1}{T}\sum_{t=1}^{T-k} u_t u_{t+k} v_{t,j} v_{t+k,h}\right|$$

Under Assumptions A3.1.1 and (i) A3.1.2 (i)

$$\begin{aligned}\mathbb{E}\left[\sum_{|k|<T}\left|K\left(\frac{k}{M_T}\right)\right| \max_{j,h\in G}\left|\frac{1}{T}\sum_{t=1}^{T-k} u_t u_{t+k} v_{t,j} v_{t+k,h}\right|\right] &\leq O(M_T)\sup_{k\in\mathbf{Z}}\sum_{j,h\in G}\mathbb{E}\left|\frac{1}{T}\sum_{t=1}^{T-k} u_t u_{t+k} v_{t,j} v_{t+k,h}\right| \\ &\leq O(M_T)|G|^2 \sup_{k\in\mathbf{Z}}\max_{j,h\in G}\mathbb{E}|u_t u_{t+k} v_{t,j} v_{t+k,h}| \\ &= O(M_T),\end{aligned}$$

and whence $S_T^b = O_P\left(M_T\left(\frac{s_\alpha p^{1/\kappa}}{T^{1-1/\kappa}} \vee s_\alpha\sqrt{\frac{\log p}{T}}\right)\right)$.

Next, we evaluate uniformly over $|k| < T$

$$\begin{aligned}&\left|\frac{1}{T}\sum_{t=1}^{T-k} \hat{u}_t\hat{u}_{t+k}\hat{v}_{t,j}\hat{v}_{t+k,h} - \frac{1}{T}\sum_{t=1}^{T-k} u_t u_{t+k} v_{t,j} v_{t+k,h}\right| \\ &\leq \left|\frac{1}{T}\sum_{t=1}^{T-k}(\hat{u}_t\hat{v}_{t,j} - u_t v_{t,j})u_{t+k} v_{t+k,h}\right| + \left|\frac{1}{T}\sum_{t=1}^{T-k} u_t v_{t,j}(\hat{u}_{t+k}\hat{v}_{t+k,h} - u_{t+k} v_{t+k,h})\right| \\ &\quad + \left|\frac{1}{T}\sum_{t=1}^{T-k}(\hat{u}_t\hat{v}_{t,j} - u_t v_{t,j})(\hat{u}_{t+k}\hat{v}_{t+k,h} - u_{t+k} v_{t+k,h})\right| \triangleq I_T + II_T + III_T.\end{aligned}$$

We bound the first term as

$$I_T \leq \left|\frac{1}{T}\sum_{t=1}^{T-k}(\hat{u}_t - u_t)v_{t,j}u_{t+k}v_{t+k,h}\right| + \left|\frac{1}{T}\sum_{t=1}^{T-k}u_t(\hat{v}_{t,j} - v_{t,j})u_{t+k}v_{t+k,h}\right| + \left|\frac{1}{T}\sum_{t=1}^{T-k}(\hat{u}_t - u_t)(\hat{v}_{t,j} - v_{t,j})u_{t+k}v_{t+k,h}\right| \triangleq I_T^a + I_T^b + I_T^c.$$

By the Cauchy-Schwartz inequality, under Assumptions of Theorem A3.1 for the sg-LASSO and Assumption A3.1.1 (ii)

$$\begin{aligned} I_T^a &= \left|\frac{1}{T}\sum_{t=1}^{T-k}\left(x_t^\top(\beta - \hat{\beta}) + m_t - x_t^\top\beta\right)v_{t,j}u_{t+k}v_{t+k,h}\right| \\ &\leq (\|\mathbf{X}(\hat{\beta} - \beta)\|_T + \|\mathbf{m} - \mathbf{X}\beta\|_T)\sqrt{\frac{1}{T}\sum_{t=1}^{T-k}v_{t,j}^2u_{t+k}^2v_{t+k,h}^2} \\ &= O_P\left(\frac{s_\alpha p^{1/\kappa}}{T^{1-1/\kappa}} \vee \sqrt{\frac{s_\alpha \log p}{T}}\right). \end{aligned}$$

Similarly, under Assumptions of Theorem A3.1 for the nodewise LASSO and Assumption A3.1.1 (ii)

$$\begin{aligned} I_T^b &\leq \left(\|\mathbf{X}_{-j}(\hat{\gamma}_j - \gamma_j)\|_T + o_P(T^{-1/2})\right)\sqrt{\frac{1}{T}\sum_{t=1}^{T-k}u_t^2u_{t+k}^2v_{t+k,h}^2} \\ &= O_P\left(\frac{S_j p^{1/\kappa}}{T^{1-1/\kappa}} \vee \sqrt{\frac{S_j \log p}{T}}\right). \end{aligned}$$

Note that for arbitrary $(\xi_t)_{t\in\mathbf{Z}}$ and $q \geq 1$, by Jensen's inequality

$$\mathbb{E}\left[\max_{t\in[T]}|\xi_t|\right] \leq \left(\mathbb{E}\left[\max_{t\in[T]}|\xi_t|^q\right]\right)^{1/q} \leq \left(\mathbb{E}\left[\sum_{t=1}^{T}|\xi_t|^q\right]\right)^{1/q} = T^{1/q}\left(\mathbb{E}|\xi_t|^q\right)^{1/q}.$$

Then by the Cauchy-Schwartz inequality under Assumption A3.1.1 (iii) and Theorem A3.1

$$\begin{aligned} I_T^c &\leq (\|\mathbf{X}(\hat{\beta} - \beta)\|_T + o_P(T^{-1/2}))(\|\mathbf{X}_{-j}(\hat{\gamma}_j - \gamma_j)\|_T + o_P(T^{-1/2}))\max_{t\in[T]}|u_tv_{t,h}| \\ &= O_P\left(\frac{s^2p^{2/\kappa}}{T^{2-3/\kappa}} \vee \frac{s\log p}{T^{1-1/\kappa}}\right), \end{aligned}$$

where we use the fact that $\kappa \leq q$. Therefore, under maintained assumptions

$$I_T = O_P\left(\frac{sp^{1/\kappa}}{T^{1-1/\kappa}} \vee \sqrt{\frac{s\log p}{T}} + \frac{s^2p^{2/\kappa}}{T^{2-3/\kappa}} \vee \frac{s\log p}{T^{1-1/\kappa}}\right)$$

and by symmetry

$$II_T = O_P\left(\frac{sp^{1/\kappa}}{T^{1-1/\kappa}} \vee \sqrt{\frac{s\log p}{T}} + \frac{s^2p^{2/\kappa}}{T^{2-3/\kappa}} \vee \frac{s\log p}{T^{1-1/\kappa}}\right).$$

Lastly, by the Cauchy-Schwartz inequality

$$\begin{aligned} III_T &\leq \sqrt{\frac{1}{T}\sum_{t=1}^{T-k}(\hat{u}_t\hat{v}_{t,j} - u_tv_{t,j})^2 \frac{1}{T}\sum_{t=1}^{T-k}(\hat{u}_{t+k}\hat{v}_{t+k,h} - u_{t+k}v_{t+k,h})^2} \\ &\leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_t\hat{v}_{t,j} - u_tv_{t,j})^2 \frac{1}{T}\sum_{t=1}^{T}(\hat{u}_t\hat{v}_{t,h} - u_tv_{t,h})^2}. \end{aligned}$$

For each $j \in G$

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}(\hat{u}_t\hat{v}_{t,j} - u_tv_{t,j})^2 &\leq \frac{3}{T}\sum_{t=1}^{T}|\hat{u}_t - u_t|^2v_{t,j}^2 + \frac{3}{T}\sum_{t=1}^{T}|\hat{v}_{t,j} - v_{t,j}|^2u_t^2 \\ &\quad + \frac{3}{T}\sum_{t=1}^{T}|\hat{u}_t - u_t|^2|\hat{v}_{t,j} - v_{t,j}|^2 \\ &\triangleq III_T^a + III_T^b + III_T^c. \end{aligned}$$

Since under Assumption A3.1.1 (iii), $\mathbb{E}|v_{t,j}|^{2q} < \infty$ and $\mathbb{E}|u_t|^{2q} < \infty$,

$$III_T^a \leq 3\max_{t\in[T]}|v_{t,j}|^2(\|\mathbf{X}(\hat{\beta} - \beta)\|_T^2 + o_P(T^{-1/2})) = O_P\left(\frac{s_\alpha p^{2/\kappa}}{T^{2-3/\kappa}} \vee \frac{s_\alpha \log p}{T^{1-1/\kappa}}\right)$$

and

$$III_T^b \leq 3\max_{t\in[T]}|u_t|^2(\|\mathbf{X}_{-j}(\hat{\gamma}_j - \gamma_j)\|_T^2 + o_P(T^{-1/2})) = O_P\left(\frac{S_j p^{2/\kappa}}{T^{2-3/\kappa}} \vee \frac{S_j \log p}{T^{1-1/\kappa}}\right).$$

For the last term, since under Assumption 3.2.1 (ii), $\sup_k \mathbb{E}|X_{t,k}|^{2\tilde{q}} < \infty$ and $\kappa \geq \tilde{q}$, by Theorem A3.1

$$\begin{aligned}
III_T^c &\leq 3(\|\mathbf{X}(\hat{\beta}-\beta)\|_T^2 + o_P(T^{-1/2})) \max_{t\in[T]} |X_{t,-j}^\top(\hat{\gamma}_j - \gamma_j) + m_t - X_t^\top \beta|^2 \\
&\leq O_P\left(\frac{s_\alpha p^{2/\kappa}}{T^{2-2/\kappa}} \vee \frac{s_\alpha \log p}{T}\right)\left(2\max_{t\in[T]}|X_t|_\infty^2 |\hat{\gamma}_j - \gamma_j|_1^2 + 2T\|\mathbf{m} - \mathbf{X}^\top\beta\|_T^2\right) \\
&= O_P\left(\left(\frac{s_\alpha p^{2/\kappa}}{T^{2-2/\kappa}} \vee \frac{s_\alpha \log p}{T}\right)\left(\frac{S^2 p^{2/\kappa}}{T^{2-2/\kappa}} \vee S^2\frac{\log p}{T}\right)(pT)^{1/\kappa}\right) \\
&= O_P\left(\frac{s^3 p^{5/\kappa}}{T^{4-5/\kappa}} + \frac{s^3 p^{3/\kappa}\log p}{T^{3-3/\kappa}} + \frac{s^3 p^{1/\kappa}\log^2 p}{T^{2-1/\kappa}}\right) \\
&= O_P\left(\frac{s^3 p^{5/\kappa}}{T^{4-5/\kappa}} + \frac{s^3 p^{3/\kappa}\log p}{T^{3-3/\kappa}}\right),
\end{aligned}$$

where we use the fact that $\kappa > 2$, $s = s_\alpha \vee S$, $s^\kappa p = o(T^{4\kappa/5-1})$, and $s^2 \log p/T \to 0$ as $T \to \infty$. Then for every $j \in G$

$$\frac{1}{T}\sum_{t=1}^T (\hat{u}_t\hat{v}_{t,j} - u_t v_{t,j})^2 = O_P\left(\frac{sp^{2/\kappa}}{T^{2-3/\kappa}} \vee \frac{s\log p}{T^{1-1/\kappa}} + \frac{s^3 p^{5/\kappa}}{T^{4-5/\kappa}} + \frac{s^3 p^{3/\kappa}\log p}{T^{3-3/\kappa}}\right),$$

and whence

$$III_T = O_P\left(\frac{sp^{2/\kappa}}{T^{2-3/\kappa}} \vee \frac{s\log p}{T^{1-1/\kappa}} + \frac{s^3 p^{5/\kappa}}{T^{4-5/\kappa}} + \frac{s^3 p^{3/\kappa}\log p}{T^{3-3/\kappa}}\right).$$

Therefore, since $\hat{\sigma}_j^2 \xrightarrow{P} \sigma_j^2$, we obtain

$$\begin{aligned}
S_T^a &= O_P\left(M_T\left(\frac{sp^{1/\kappa}}{T^{1-1/\kappa}} \vee \sqrt{\frac{s\log p}{T}} + \frac{s^2 p^{2/\kappa}}{T^{2-3/\kappa}} \vee \frac{s\log p}{T^{1-1/\kappa}} + \frac{s^3 p^{5/\kappa}}{T^{4-5/\kappa}} + \frac{s^3 p^{3/\kappa}\log p}{T^{3-3/\kappa}}\right)\right) \\
&= O_P\left(M_T\left(\frac{sp^{1/\kappa}}{T^{1-1/\kappa}} \vee \sqrt{\frac{s\log p}{T}} + \frac{s^2 p^{2/\kappa}}{T^{2-3/\kappa}} + \frac{s^3 p^{5/\kappa}}{T^{4-5/\kappa}}\right)\right),
\end{aligned}$$

where the last line follows since $s^\kappa p/T^{4\kappa/5-1} = o(1)$. Combining this estimate with previously obtained estimate for S_T^b

$$\|\hat{\Xi}_G - \tilde{\Xi}_G\| = O_P\left(M_T\left(\frac{sp^{1/\kappa}}{T^{1-1/\kappa}} \vee s\sqrt{\frac{\log p}{T}} + \frac{s^2 p^{2/\kappa}}{T^{2-3/\kappa}} + \frac{s^3 p^{5/\kappa}}{T^{4-5/\kappa}}\right)\right).$$

The result follows from combining this estimate with the estimate in equation ((A3.1)). □

Proof of Theorem 3.1. Suppose first that $p = 1$. For $a \in \mathbf{R}$, with some abuse of notation, let $[a]$ denote its integer part. We split partial sums into blocks $V_k = \xi_{(k-1)J+1} + \cdots + \xi_{kJ}, k = 1, 2, \ldots, [T/J]$ and $V_{[T/J]+1} = \xi_{[T/J]J+1} + \cdots + \xi_T$, where we set $V_{[T/J]+1} = 0$ if T/J is an integer. Let $\{U_t : \ t = 1, 2, \ldots, [T/J] + 1\}$ be i.i.d. random variables drawn from the uniform distribution on $(0, 1)$ independently of $\{V_t : \ t = 1, 2, \ldots, [T/J]+1\}$. Put $\mathcal{M}_t = \sigma(V_1, \ldots, V_{t-2})$ for every $t = 3, \ldots, [T/J] + 1$. Next, for $t = 1, 2$, set $V_t^* = V_t$, while for $t \geq 3$, by Dedecker and Prieur (2004), Lemma 5, there exist random variables $V_t^* =_d V_t$ such that:

1. V_t^* is $\sigma(V_1, \ldots, V_{t-2}) \vee \sigma(V_t) \vee \sigma(U_t)$-measurable;
2. $V_t^* \perp\!\!\!\perp (V_1, \ldots, V_{t-2})$;
3. $\|V_t - V_t^*\|_1 = \tau(\mathcal{M}_t, V_t)$.

It follows from properties 1. and 2. that $(V_{2t}^*)_{t\geq 1}$ and $(V_{2t-1}^*)_{t\geq 1}$ are sequences of independent random variables. Then

$$\begin{aligned}\left|\sum_{t=1}^{T} \xi_t\right| &\leq \left|\sum_{t\geq 1} V_{2t}^*\right| + \left|\sum_{t\geq 1} V_{2t-1}^*\right| + \sum_{t=3}^{[T/J]+1} |V_t - V_t^*| \\ &\triangleq I_T + II_T + III_T.\end{aligned}$$

By Fuk and Nagaev (1971), Corollary 4, there exist constants $c_q^{(j)}, j = 1, 2$ such that

$$\begin{aligned}\Pr(I_T \geq x) &\leq \frac{c_q^{(1)}}{x^q} \sum_{t\geq 1} \mathbb{E}|V_{2t}^*|^q + 2\exp\left(-\frac{c_q^{(2)} x^2}{\sum_{t\geq 1} \mathrm{Var}(V_{2t}^*)}\right) \\ &\leq \frac{c_q^{(1)}}{x^q} \sum_{t\geq 1} \mathbb{E}|V_{2t}|^q + 2\exp\left(-\frac{c_q^{(2)} x^2}{B_T^2}\right),\end{aligned}$$

where the second inequality follows since $\sum_{t\geq 1} \mathrm{Var}(V_{2t}^*) = \sum_{t\geq 1} \mathrm{Var}(V_{2t}) \leq B_T^2$. Similarly

$$\Pr(II_T \geq x) \leq \frac{c_q^{(1)}}{x^q} \sum_{t\geq 1} \mathbb{E}|V_{2t-1}|^q + 2\exp\left(-\frac{c_q^{(2)} x^2}{B_T^2}\right).$$

Lastly, by Markov's inequality and property 3.

$$\begin{aligned}\Pr\left(III_T \geq x\right) &\leq \frac{1}{x}\sum_{t=3}^{[T/J]+1}\tau(\mathcal{M}_t, V_t)\\ &\leq \frac{1}{x}\sum_{t=3}^{[T/J]+1}\tau(\mathcal{M}_t, (\xi_{(t-1)J+1}, \ldots, \xi_{tJ}))\\ &\leq \frac{1}{x}[T/J]\sup_{t+J+1\leq t_1<\cdots<t_J}\tau(\mathcal{M}_t, (\xi_{t_1}, \ldots, \xi_{t_J}))\\ &\leq \frac{T}{x}\tau_{J+1},\end{aligned}$$

where the second inequality follows since the sum is a 1-Lipschitz function with respect to $|.|_1$-norm and the third since $\mathcal{M}_t$ and $(\xi_{(t-1)J+1}, \ldots, \xi_{tJ})$ are separated by $J+1$ lags of $(\xi_t)_{t\in\mathbf{Z}}$.

Combining all the estimates together

$$\begin{aligned}\Pr\left(\left|\sum_{t=1}^{T}\xi_t\right| \geq 3x\right) &\leq \Pr(I_T \geq x) + \Pr(II_T \geq x) + \Pr(III_T \geq x)\\ &\leq \frac{c_q^{(1)}}{x^q}\sum_{t=1}^{[T/J]+1}\mathbb{E}|V_t|^q + 4\exp\left(-\frac{c_q^{(2)}x^2}{B_T^2}\right) + \frac{T}{x}\tau_{J+1}\\ &\leq \frac{c_q^{(1)}}{x^q}J^{q-1}\sum_{t=1}^{T}\|\xi_t\|_q^q + \frac{T}{x}c(J+1)^{-a} + 4\exp\left(-\frac{c_q^{(2)}x^2}{B_T^2}\right).\end{aligned}$$

To balance the first two terms, we shall set $J \sim x^{\frac{q-1}{q+a-1}}$, in which case we obtain the result under maintained assumptions. The result for $p > 1$ follows by the union bound. □

For a stationary process $(\xi_t)_{t\in\mathbf{Z}}$, let

$$\gamma_k = \|\mathbb{E}(\xi_k|\mathcal{M}_0) - \mathbb{E}(\xi_k)\|_1$$

be its L_1 mixingale coefficient with respect to the canonical filtration $\mathcal{M}_0 = \sigma(\xi_0, \xi_{-1}, \xi_{-2}, \ldots)$. Let α_k be the α-mixing coefficient and let Q be the quantile function of $|\xi_0|$. The following covariance inequality allows us controlling the autocovariances in terms of the τ-mixing coefficient as well as comparing the latter to the mixingale and the α-mixing coefficients.

Lemma A3.1.1. *Let $(\xi_t)_{t\in\mathbf{Z}}$ be a centered stationary stochastic process with $\|\xi_0\|_q < \infty$ for some $q > 2$. Then*

$$|\mathrm{Cov}(\xi_0, \xi_t)| \leq \gamma_t^{\frac{q-2}{q-1}} \|\xi_0\|_q^{q/(q-1)}$$

and

$$\gamma_t \leq \tau_t \leq 2\int_0^{2\alpha_t} Q(u)\mathrm{d}u.$$

Proof. Let G be the generalized inverse of $x \mapsto \int_0^x Q(u)\mathrm{d}u$. By Dedecker and Doukhan (2003), Proposition 1

$$\begin{aligned}
|\mathrm{Cov}(\xi_0, \xi_t)| &\leq \int_0^{\gamma_t} (Q \circ G)(u)\mathrm{d}u \\
&\leq \gamma_t^{\frac{q-2}{q-1}} \left(\int_0^{\|\xi_0\|_1} (Q \circ G)^{q-1}(u)\mathrm{d}u \right)^{1/(q-1)} \\
&= \gamma_t^{\frac{q-2}{q-1}} \|\xi_0\|_q^{q/(q-1)},
\end{aligned}$$

where the second line follows by Hölder's inequality and the last equality by the change of variables $\int_0^{\|\xi_0\|_1} (Q \circ G)^{q-1}(u)\mathrm{d}u = \int_0^1 Q^q(u)\mathrm{d}u = \mathbb{E}|\xi_0|^q$. The second statement follows from Dedecker and Doukhan (2003), Lemma 1 and Dedecker and Prieur (2004), Lemma 6. □

The following result shows that the variance of partial sums can be controlled provided that the τ-mixing coefficients decline sufficiently fast.

Lemma A3.1.2. *Let $(\xi_t)_{t\in\mathbf{Z}}$ be a centered stationary stochastic process such that $\|\xi_t\|_q < \infty$ for some $q > 2$ and $\tau_k = O(k^{-a})$ for some $a > \frac{q-1}{q-2}$. Then*

$$\sum_{t=1}^{T}\sum_{k=1}^{T} |\mathrm{Cov}(\xi_{t,j}, \xi_{k,j})| = O(T).$$

Proof. Under stationarity

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{k=1}^{T} |\mathrm{Cov}(\xi_{t,j}, \xi_{k,j})| &= T\mathrm{Var}(\xi_0) + 2\sum_{k=1}^{T-1} (T-k)\,\mathrm{Cov}(\xi_0, \xi_k) \\
&\leq T\mathrm{Var}(\xi_0) + 2T\|\xi_t\|_q^{q/(q-1)} \sum_{k=1}^{T-1} \tau_k^{\frac{q-2}{q-1}} \\
&= O(T),
\end{aligned}$$

where the second line follows by Proposition A3.1.1 and the last since the series $\sum_{k=1}^{\infty} k^{-a\frac{q-2}{q-1}}$ converges under the maintained assumptions. □

Lastly, we show that in the linear regression setting, the long-run variance for a group of projection coefficients $G \subset [p]$ of fixed size exists under mild conditions. Let $\xi_t = u_t \Theta_G x_t$, where Θ_G are rows of the precision matrix $\Theta = \Sigma^{-1}$ corresponding to indices in G.

Proposition A3.1.1. *Suppose that (i) $(u_t x_t)_{t\in\mathbf{Z}}$ is stationary for every $p \geq 1$; (ii) $\|u_t\|_q < \infty$ and $\max_{j\in[p]} \|x_{t,j}\|_r = O(1)$ for some $q > 2r/(r-2)$ and $r > 4$; (iii) for every $j \in [p]$, the τ-mixing coefficients of $(u_t x_{t,j})_{t\in\mathbf{Z}}$ are $\tau_k \leq ck^{-a}$ for all $k \geq 0$, where $c > 0$ and $a > (\varsigma - 1)/(\varsigma - 2)$, and $\varsigma = qr/(q+r)$ are some universal constants; (iv) $\|\Theta_G\|_\infty = O(1)$ and $\sup_x \mathbb{E}[|u_t|^2 | x_t = x] = O(1)$. Then for every $z \in \mathbf{R}^{|G|}$, the limit*

$$\lim_{T\to\infty} \operatorname{Var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} z^\top \xi_t\right)$$

exists.

Proof. Under assumption (i), by Hölder's inequality, $\max_{j\in[p]} \|u_t x_{t,j}\|_\varsigma = O(1)$ with $\varsigma = qr/(q+r)$, whence by the Minkowski inequality and assumption (iv)

$$\|z^\top \xi_t\|_\varsigma \leq \sum_{k\in G} \sum_{j\in[p]} |\Theta_{k,j}| \|u_t x_{t,j}\|_\varsigma \leq |G| \|\Theta_G\|_\infty = O(1). \tag{A3.2}$$

Since $\varsigma > 2$, this shows that $\operatorname{Var}(z^\top \xi_0)$ exists. Moreover,

$$\begin{aligned} \operatorname{Var}(z^\top \xi_0) &= z^\top \Theta_G \operatorname{Var}(u_0 x_0) \Theta_G^\top z \\ &= \sum_{j,k\in[p]} (z^\top \Theta_G)_j (z^\top \Theta_G)_k \mathbb{E}[u_0^2 x_{0,j} x_{0,k}], \end{aligned}$$

where the sum converges as $p \to \infty$ by the comparison test under assumption (iv) implying that $\lim_{T\to\infty} \operatorname{Var}(z^\top \xi_0)$ exists. Next, under assumption (i), for every $z \in \mathbf{R}^{|G|}$

$$\operatorname{Var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} z^\top \xi_t\right) = \operatorname{Var}(z^\top \xi_0) + 2 \sum_{k=1}^{T-1} \left(1 - \frac{k}{T}\right) \operatorname{Cov}(z^\top \xi_0, z^\top \xi_k).$$

By Lemma A3.1.1, we bound covariances by the mixingale coefficient for every $k \geq 1$

$$\begin{aligned}|\mathrm{Cov}(z^\top \xi_0, z^\top \xi_k)| &\lesssim \|z^\top \xi_0\|_\varsigma^{\varsigma/(\varsigma-1)} \|\mathbb{E}(z^\top \xi_k|\mathcal{M}_0) - \mathbb{E}(z^\top \xi_k)\|_1^{\frac{\varsigma-2}{\varsigma-1}} \\ &\lesssim |z^\top \Theta_G|_1^{\frac{\varsigma-2}{\varsigma-1}} \max_{j\in[p]} \|\mathbb{E}(u_k x_{k,j}|\mathcal{M}_0) - \mathbb{E}(u_k x_{k,j})\|_1^{\frac{\varsigma-2}{\varsigma-1}} \\ &\lesssim \tau_k^{\frac{\varsigma-2}{\varsigma-1}}\end{aligned}$$

where the first inequality follows by Lemma A3.1.1, the second by equation (A3.2), and the third by Lemma A3.1.1. Under assumption (iii), $\sum_{k=1}^\infty \tau_k^{(\varsigma-2)/(\varsigma-1)}$ converges, which implies that

$$\sum_{k=1}^\infty |\mathrm{Cov}(z^\top \xi_0, z^\top \xi_k)| < \infty$$

by the comparison test. Therefore, by Lebesgue's dominated convergence, this shows that the long run variance

$$\lim_{T\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^T z^\top \xi_t\right)$$

exists. □

We recall first the convergence rates for the sg-LASSO with weakly dependent data that will be needed throughout the paper from Babii, Ghysels, and Striaukas (2020b), Corollary 3.1.

Theorem A3.1. *Suppose that Assumptions 3.2.1, 3.2.2, 3.2.3, and 3.2.4 are satisfied. Then*

$$\|\mathbf{X}(\hat{\beta}-\beta)\|_T^2 = O_P\left(\frac{s_\alpha p^{2/\kappa}}{T^{2-2/\kappa}} \vee \frac{s_\alpha \log p}{T}\right).$$

and

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

Next, we consider the regularized estimator of the variance of the regression error

$$\hat{\sigma}^2 = \|\mathbf{y}-\mathbf{X}\hat{\beta}\|_T^2 + \lambda\Omega(\hat{\beta}),$$

where $\hat{\beta}$ is the sg-LASSO estimator. While the regularization is not needed to have a consistent variance estimator, the LASSO version of the regularized estimator ($\alpha = 1$) is needed to establish the CLT for the debiased sg-LASSO estimator. The following result describes the converges of this variance estimator to its population counterpart $\sigma^2 = \mathbb{E}\|\mathbf{u}\|_T^2$.

Proposition A3.1.2. *Suppose that Assumptions 3.2.1, 3.2.2, 3.2.3, and 3.2.4 are satisfied and that $(u_t^2)_{t\in\mathbf{Z}}$ has a finite long run variance. Then*

$$\hat{\sigma}^2 = \sigma^2 + O_P\left(\frac{s_\alpha p^{1/\kappa}}{T^{1-1/\kappa}} \vee s_\alpha\sqrt{\frac{\log p}{T}}\right)$$

provided that $\frac{s_\alpha p^{1/\kappa}}{T^{1-1/\kappa}} \vee s_\alpha\sqrt{\frac{\log p}{T}} = o(1)$.

Proof. We have

$$\begin{aligned}
|\hat{\sigma}^2-\sigma^2| &= \left|\|\mathbf{u}\|_T^2 + 2\langle \mathbf{u}, \mathbf{m}-\mathbf{X}\hat{\beta}\rangle_T - \|\mathbf{m}-\mathbf{X}\hat{\beta}\|_T^2 + \lambda\Omega(\hat{\beta}) - \sigma^2\right| \\
&\le |\sigma^2 - \|\mathbf{u}\|_T^2| + 2\|\mathbf{u}\|_T\|\mathbf{m}-\mathbf{X}\hat{\beta}\|_T + 2\|\mathbf{X}(\hat{\beta}-\beta)\|_T^2 + \\
&\quad + 2\|\mathbf{m}-\mathbf{X}\beta\|_T^2 + \lambda\Omega(\hat{\beta}) \\
&\triangleq I_T + II_T + III_T + IV_T + V_T.
\end{aligned}$$

By the Chebychev's inequality since the long-run variance exists, for every $\varepsilon > 0$

$$\Pr\left(\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(u_t^2 - \sigma^2)\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2}\sum_{t\in\mathbf{Z}}\operatorname{Cov}(u_0^2, u_t^2),$$

whence $I_T = O_P\left(\frac{1}{\sqrt{T}}\right)$. Therefore, by the triangle inequality and Theorem A3.1

$$\begin{aligned} II_T &= O_P(1)\|\mathbf{m} - \mathbf{X}\hat{\beta}\|_T \\ &\leq O_P(1)\left(\|\mathbf{m} - \mathbf{X}\beta\|_T + \|\mathbf{X}(\hat{\beta} - \beta)\|_T\right) \\ &= O_P\left(s_\alpha^{1/2}\lambda + \frac{s_\alpha^{1/2}p^{1/\kappa}}{T^{1-1/\kappa}} \vee \sqrt{\frac{s_\alpha \log p}{T}}\right). \end{aligned}$$

By Theorem A3.1 we also have

$$III_T + IV_T = O_P\left(\frac{s_\alpha p^{2/\kappa}}{T^{2-2/\kappa}} \vee \frac{s_\alpha \log p}{T} + s_\alpha\lambda^2\right).$$

Lastly, another application of Theorem A3.1 gives

$$\begin{aligned} V_T &= \lambda\Omega(\hat{\beta} - \beta) + \lambda\Omega(\beta) \\ &= O_P\left(\lambda\left(\frac{s_\alpha p^{1/\kappa}}{T^{1-1/\kappa}} \vee s_\alpha\sqrt{\frac{\log p}{T}}\right) + \lambda s_\alpha\right). \end{aligned}$$

The result follows from combining all estimates together. □

Next, we look at the estimator of the precision matrix. Consider nodewise LASSO regressions in equation ((3.2)) for each $j \in [p]$. Put $S = \max_{j\in G} S_j$, where S_j is the support of γ_j.

Proposition A3.1.3. *Suppose that Assumptions 3.2.1, 3.2.2, 3.2.3, and 3.2.4 are satisfied for each nodewise regression $j \in G$ and that $(v_{t,j}^2)_{t\in\mathbf{Z}}$ has a finite long-run variance for each $j \in G$. Then if $S^\kappa p T^{1-\kappa} \to 0$ and $S^2 \log p/T \to 0$*

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

and

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

Proof. By Theorem A3.1 and Proposition A3.1.2 with $\alpha = 1$ (corresponding to the LASSO estimator of γ_j and σ_j^2)

$$\begin{aligned}\|\hat{\Theta}_G - \Theta_G\|_\infty &= \max_{j\in G} |\hat{\Theta}_j - \Theta_j|_1 \\ &\leq \max_{j\in G} \left\{ |\hat{\gamma}_j|_1 \left|\hat{\sigma}_j^{-2} - \sigma_j^{-2}\right| + |\hat{\gamma}_j - \gamma_j|_1 |\sigma_j^{-2}| \right\} \\ &= O_P\left(\frac{Sp^{1/\kappa}}{T^{1-1/\kappa}} \vee S\sqrt{\frac{\log p}{T}} \right),\end{aligned}$$

where we use the fact that $|G|$ is fixed and that $\hat{\sigma}_j^2 \xrightarrow{p} \sigma_j^2$ under maintained assumptions.

Second, for each $j \in G$, by Fermat's rule,

$$\mathbf{X}_{-j}^\top (\mathbf{X}_j - \mathbf{X}_{-j}\hat{\gamma}_j)/T = \lambda_j z^*, \qquad z^* \in \partial |\hat{\gamma}_j|_1,$$

where $\hat{\gamma}_j^\top z^* = |\hat{\gamma}_j|_1$ and $|z^*|_\infty \leq 1$. Then

$$\begin{aligned}\mathbf{X}_j^\top (\mathbf{X}_j - \mathbf{X}_{-j}\hat{\gamma}_j)/T &= \|\mathbf{X}_j - \mathbf{X}_{-j}\hat{\gamma}_j\|_T^2 + \hat{\gamma}_j^\top \mathbf{X}_{-j}^\top (\mathbf{X}_j - \mathbf{X}_{-j}\hat{\gamma}_j)/T \\ &= \|\mathbf{X}_j - \mathbf{X}_{-j}\hat{\gamma}_j\|_T^2 + \lambda_j \hat{\gamma}_j^\top z^* = \hat{\sigma}_j^2,\end{aligned}$$

and whence

$$\begin{aligned}|(I - \hat{\Theta}\hat{\Sigma})_j|_\infty &= |I_j - (\mathbf{X}_j - \mathbf{X}_{-j}\hat{\gamma}_j)^\top \mathbf{X}/(T\hat{\sigma}_j^2)|_\infty \\ &= \max\left\{ |1 - \mathbf{X}_j^\top (\mathbf{X}_j - \mathbf{X}_{-j}\hat{\gamma}_j)/(T\hat{\sigma}_j^2)|, \right. \\ &\quad \left. |\mathbf{X}_{-j}^\top (\mathbf{X}_j - \mathbf{X}_{-j}\hat{\gamma}_j)/(T\hat{\sigma}_j^2)|_\infty \right\} \\ &= \lambda_j |z^*|_\infty / \hat{\sigma}_j^2 = O_P\left(\frac{p^{1/\kappa}}{T^{1-1/\kappa}} \vee \sqrt{\frac{\log p}{T}} \right),\end{aligned}$$

where the last line follows since $\hat{\sigma}_j^{-2} = O_P(1)$ and $|z^*|_\infty \leq 1$. The conclusion follows from the fact that $|G|$ is fixed. □

Next, we first derive the non-asymptotic Frobenius norm bound with explicit constants for a generic HAC estimator of the sample mean that holds uniformly over a class of distributions. We focus on the p-dimensional centered stochastic process $(V_t)_{t\in\mathbf{Z}}$ and put

$$\Xi = \sum_{k\in\mathbf{Z}} \Gamma_k \qquad \text{and} \qquad \tilde{\Xi} = \sum_{|k|<T} K\left(\frac{k}{M_T}\right) \tilde{\Gamma}_k,$$

where $\Gamma_k = \mathbb{E}[V_t V_{t+k}^\top]$ and $\tilde{\Gamma}_k = \frac{1}{T}\sum_{t=1}^{T-k} V_t V_{t+k}^\top$. Put also $\Gamma = (\Gamma_k)_{k\in\mathbf{Z}}$ and let $\langle .,. \rangle$ be the Frobenius inner product with corresponding Frobenius norm $\|.\|$. The following assumption describes the relevant class of distributions and kernel functions.

Assumption A3.1.2. *Suppose that (i) $K : \mathbf{R} \to [-1,1]$ is a Riemann integrable function such that $K(0) = 1$; (ii) there exists some $\varepsilon, \varsigma > 0$ such that $|K(0) - K(x)| \leq L|x|^\varsigma$ for all $|x| < \varepsilon$; (iii) $(V_t)_{t\in\mathbf{Z}}$ is fourth-order stationary; (iv) $\Gamma \in \mathcal{G}(\varsigma, D_1, D_2)$, where*

$$\mathcal{G}(\varsigma, D_1, D_2) = \left\{ \sum_{k\in\mathbf{Z}} |k|^\varsigma \|\Gamma_k\| \leq D_1, \quad \sup_{k\in\mathbf{Z}} \sum_{l\in\mathbf{Z}} \sum_{t\in\mathbf{Z}} \sum_{j,h\in[p]} |\mathrm{Cov}(V_{0,j}V_{k,h}, V_{t,j}V_{t+l,h})| \leq D_2 \right\}$$

for some $D_1, D_2 > 0$.

Condition (ii) describes the smoothness (or order) of the kernel in the neighborhood of zero. $\varsigma = 1$ for the Bartlett kernel and $\varsigma = 2$ for the Parzen, Tukey-Hanning, and Quadratic spectral kernels, see Andrews (1991). Since the bias of the HAC estimator is limited by the order of the kernel, it is typically not recommended to use the Bartlett kernel in practice. Higher-order kernels with $\varsigma > 2$ do not ensure the positive definiteness of the HAC estimator and require additional spectral regularization, see Politis (2011). Condition (iv) describes the class of autocovariances that vanish rapidly enough. Note that if (iv) holds for some $\bar{\varsigma}$, then it also holds for every $\varsigma < \bar{\varsigma}$ and that if (ii) holds for some $\tilde{\varsigma} > \varsigma$, then it also holds for $\tilde{\varsigma} = \varsigma$. The covariance condition in (iv) can be justified under more primitive moment and summability conditions imposed on L_1-mixingale/τ-mixing coefficients, see Proposition A3.1.1 and Andrews (1991), Lemma 1. The following result gives a nonasymptotic risk bound uniformly over the class $\mathcal{G}$ and corresponds to the asymptotic convergence rates for the spectral density evaluated at zero derived in Parzen (1957).

Proposition A3.1.4. *Suppose that Assumption A3.1.2 is satisfied. Then*

$$\sup_{\Gamma\in\mathcal{G}(\varsigma,D_1,D_2)} \mathbb{E}\|\tilde{\Xi} - \Xi\|^2 \leq C_1 \frac{M_T}{T} + C_2 M_T^{-2\varsigma} + C_3 T^{-2(\varsigma\wedge 1)},$$

where $C_1 = D_2\left(\int |K(u)|\mathrm{d}u + o(1)\right)$, $C_2 = 2\left(D_1 L + \frac{2D_1}{\varepsilon^\varsigma}\right)^2$, and $C_3 = 2D_1^2$.

Proof. By the triangle inequality, under Assumption A3.1.2 (i)

$$\begin{aligned}\|\mathbb{E}[\tilde{\Xi}] - \Xi\| &= \left\| \sum_{|k|<T} K\left(\frac{k}{M_T}\right)\frac{T-k}{T}\Gamma_k - \sum_{k\in\mathbf{Z}}\Gamma_k \right\| \\ &\leq \sum_{|k|<T}\left|K\left(\frac{k}{M_T}\right) - K(0)\right|\|\Gamma_k\| + \frac{1}{T}\sum_{|k|<T}|k|\|\Gamma_k\| + \sum_{|k|\geq T}\|\Gamma_k\| \\ &\triangleq I_T + II_T + III_T.\end{aligned}$$

For the first term, we obtain

$$\begin{aligned}I_T &= \sum_{|k|<\varepsilon M_T}\left|K(0) - K\left(\frac{k}{M_T}\right)\right|\|\Gamma_k\| + \sum_{\varepsilon M_T\leq|k|<T}\left|K\left(\frac{k}{M_T}\right) - K(0)\right|\|\Gamma_k\| \\ &\leq LM_T^{-\varsigma}\sum_{|k|<\varepsilon M_T}|k|^{\varsigma}\|\Gamma_k\| + 2\sum_{\varepsilon M_T\leq|k|<T}\|\Gamma_k\| \\ &\leq \frac{D_1 L}{M_T^{\varsigma}} + \frac{2}{\varepsilon^{\varsigma}M_T^{\varsigma}}\sum_{\varepsilon M_T\leq|k|<T}|k|^{\varsigma}\|\Gamma_k\| \\ &\leq \frac{D_1 L}{M_T^{\varsigma}} + \frac{2D_1}{\varepsilon^{\varsigma}M_T^{\varsigma}},\end{aligned}$$

where the second sum is defined to be zero if $T \leq \varepsilon M_T$, the second line follows under Assumption A3.1.2 (i)-(ii) and the last two under Assumption A3.1.2 (iii). Next, if $\varsigma \geq 1$,

$$\sum_{|k|<T}|k|\|\Gamma_k\| \leq \sum_{|k|<T}|k|^{\varsigma}\|\Gamma_k\|,$$

while if $\varsigma \in (0,1)$

$$\sum_{|k|<T}|k|\|\Gamma_k\| \leq T^{1-\varsigma}\sum_{|k|<T}|k|^{\varsigma}\|\Gamma_k\|.$$

Therefore, since $\sum_{|k|\geq T}\|\Gamma_k\| \leq T^{-\varsigma}\sum_{|k|\geq T}|k|^{\varsigma}\|\Gamma_k\|$, under Assumption A3.1.2 (iv)

$$\begin{aligned}II_T + III_T &\leq \begin{cases}\frac{D_1}{T} & \varsigma \geq 1 \\ \frac{D_1}{T^{\varsigma}} & \varsigma \in (0,1)\end{cases} \\ &= \frac{D_1}{T^{\varsigma\wedge 1}}.\end{aligned}$$

This shows that

$$\|\mathbb{E}[\tilde{\Xi}] - \Xi\| \leq \frac{D_1 L}{M_T^{\varsigma}} + \frac{2D_1}{\varepsilon^{\varsigma} M_T^{\varsigma}} + \frac{D_1}{T^{\varsigma \wedge 1}}. \tag{A3.3}$$

Next, under Assumption A3.1.2 (i)

$$\begin{aligned}\mathbb{E}\|\tilde{\Xi} - \mathbb{E}[\tilde{\Xi}]\|^2 &= \sum_{|k|<T} \sum_{|l|<T} K\left(\frac{k}{M_T}\right) K\left(\frac{l}{M_T}\right) \mathbb{E}\left\langle \tilde{\Gamma}_k - \mathbb{E}\tilde{\Gamma}_k, \tilde{\Gamma}_l - \mathbb{E}\tilde{\Gamma}_l \right\rangle \\ &\leq \sum_{|k|<T} \left| K\left(\frac{k}{M_T}\right) \right| \sup_{|k|<T} \sum_{|l|<T} \left| \mathbb{E}\left\langle \tilde{\Gamma}_k - \mathbb{E}\tilde{\Gamma}_k, \tilde{\Gamma}_l - \mathbb{E}\tilde{\Gamma}_l \right\rangle \right|,\end{aligned}$$

where under Assumptions A3.1.2 (iii)

$$\begin{aligned}T \left| \mathbb{E}\left\langle \tilde{\Gamma}_k - \mathbb{E}\tilde{\Gamma}_k, \tilde{\Gamma}_l - \mathbb{E}\tilde{\Gamma}_l \right\rangle \right| &\leq \frac{1}{T} \sum_{t=1}^{T-k} \sum_{r=1}^{T-l} \sum_{j,h\in[p]} |\mathrm{Cov}(V_{t,j} V_{t+k,h}, V_{r,j} V_{r+l,h})| \\ &\leq \sum_{t\in\mathbf{Z}} \sum_{j,h\in[p]} |\mathrm{Cov}(V_{0,j} V_{k,h}, V_{t,j} V_{t+l,h})|.\end{aligned}$$

Therefore, under Assumptions A3.1.2 (i), (iv)

$$\mathbb{E}\|\tilde{\Xi} - \mathbb{E}[\tilde{\Xi}]\|^2 \leq M_T \left(\int |K(u)| \mathrm{d}u + o(1) \right) \frac{D_2}{T}. \tag{A3.4}$$

The result follows from combining estimates in equations ((A3.3)) and ((A3.4)). □

A3.2 Data

	News topic	Meta topic	LASSO	sg-LASSO
1	Accounting	Asset Managers & I-Banks		
2	Acquired investment banks	Asset Managers & I-Banks	✓	✓
3	Activists	Activism/Language		
4	Aerospace/defense	Trans/Retail/Local Politics	✓	✓
5	Agreement reached	Negotiations		
6	Agriculture	Oil & Mining		
7	Airlines	Trans/Retail/Local Politics		
8	Announce plan	Activism/Language		
9	Arts	Social/Cultural		
10	Automotive	Trans/Retail/Local Politics		
11	Bank loans	Banks		
12	Bankruptcy	Buyouts & Bankruptcy		
13	Bear/bull market	Financial Markets	✓	✓
14	Biology/chemistry/physics	Science/Language		
15	Bond yields	Financial Markets		
16	Broadcasting	Entertainment	✓	✓
17	Buffett	Activism/Language		
18	Bush/Obama/Trump	Leaders		
19	C-suite	Management		
20	Cable	Industry		
21	California	Trans/Retail/Local Politics		
22	Canada/South Africa	International Affairs		
23	Casinos	Industry	✓	
24	Challenges	Challenges		
25	Changes	Challenges		
26	Chemicals/paper	Industry		
27	China	International Affairs		
28	Clintons	Leaders		
29	Committees	Negotiations		
30	Commodities	Financial Markets		
31	Company spokesperson	Negotiations		
32	Competition	Industry		
33	Computers	Technology		
34	Connecticut	Management		
35	Control stakes	Buyouts & Bankruptcy	✓	
36	Convertible/preferred	Buyouts & Bankruptcy		
37	Corporate governance	Buyouts & Bankruptcy	✓	✓
38	Corrections/amplifications	Activism/Language		
39	Couriers	Industry		
40	Courts	Courts	✓	
41	Credit cards	Industry		
42	Credit ratings	Banks	✓	✓
43	Cultural life	Social/Cultural		
44	Currencies/metals	Financial Markets		
45	Disease	Trans/Retail/Local Politics	✓	✓
46	Drexel	Buyouts & Bankruptcy	✓	
47	Earnings	Corporate Earnings		
48	Earnings forecasts	Corporate Earnings		
49	Earnings losses	Corporate Earnings	✓	
50	Economic growth	Economic Growth		
51	Economic ideology	Social/Cultural		
52	Elections	Leaders	✓	✓
53	Electronics	Technology	✓	✓

54	Environment	Government	✓	
55	European politics	Leaders	✓	
56	European sovereign debt	Economic Growth	✓	✓
57	Exchanges/composites	Financial Markets		
58	Executive pay	Labor/income		
59	Fast food	Industry	✓	
60	Federal Reserve	Economic Growth	✓	✓
61	Fees	Labor/income		
62	Financial crisis	Banks	✓	✓
63	Financial reports	Corporate Earnings		
64	Foods/consumer goods	Industry		
65	France/Italy	International Affairs		
66	Futures/indices	Activism/Language		
67	Gender issues	Social/Cultural		
68	Germany	International Affairs	✓	✓
69	Government budgets	Labor/income		
70	Health insurance	Labor/income		
71	Humor/language	Social/Cultural		
72	Immigration	Social/Cultural		
73	Indictments	Courts		
74	Insurance	Industry		
75	International exchanges	Financial Markets		
76	Internet	Technology		
77	Investment banking	Asset Managers & I-Banks		
78	IPOs	Financial Markets		
79	Iraq	Terrorism/Mideast	✓	
80	Japan	International Affairs	✓	
81	Job cuts	Labor/income		
82	Justice Department	Courts		
83	Key role	Challenges		
84	Latin America	International Affairs	✓	✓
85	Lawsuits	Courts		
86	Long/short term	Challenges		
87	Luxury/beverages	Industry		
88	M&A	Buyouts & Bankruptcy	✓	
89	Machinery	Oil & Mining		
90	Macroeconomic data	Economic Growth		
91	Major concerns	Activism/Language		
92	Management changes	Management	✓	
93	Marketing	Entertainment	✓	
94	Mexico	Activism/Language	✓	
95	Microchips	Technology		
96	Mid-level executives	Management		
97	Mid-size cities	Trans/Retail/Local Politics		
98	Middle east	Terrorism/Mideast	✓	
99	Mining	Oil & Mining		
100	Mobile devices	Technology		
101	Mortgages	Banks	✓	
102	Movie industry	Entertainment	✓	
103	Music industry	Entertainment		
104	Mutual funds	Asset Managers & I-Banks		
105	NASD	Asset Managers & I-Banks	✓	✓
106	National security	Government	✓	
107	Natural disasters	Trans/Retail/Local Politics	✓	✓
108	Negotiations	Negotiations		
109	News conference	Negotiations		
110	Nonperforming loans	Banks	✓	

111	Nuclear/North Korea	Terrorism/Mideast	✓	
112	NY politics	Trans/Retail/Local Politics	✓	
113	Oil drilling	Oil & Mining		
114	Oil market	Oil & Mining	✓	
115	Optimism	Economic Growth		
116	Options/VIX	Financial Markets	✓	✓
117	Pensions	Labor/income		
118	People familiar	Negotiations		
119	Pharma	Trans/Retail/Local Politics		
120	Phone companies	Technology		
121	Police/crime	Trans/Retail/Local Politics		
122	Political contributions	Government		
123	Positive sentiment	Social/Cultural		
124	Private equity/hedge funds	Asset Managers & I-Banks		
125	Private/public sector	Government		
126	Problems	Challenges		
127	Product prices	Economic Growth		
128	Profits	Corporate Earnings	✓	✓
129	Programs/initiatives	Science/Language		
130	Publishing	Entertainment	✓	
131	Rail/trucking/shipping	Trans/Retail/Local Politics		
132	Reagan	Leaders		
133	Real estate	Buyouts & Bankruptcy		
134	Recession	Economic Growth	✓	✓
135	Record high	Economic Growth		
136	Regulation	Government		
137	Rental properties	Trans/Retail/Local Politics	✓	
138	Research	Science/Language		
139	Restraint	Negotiations		
140	Retail	Trans/Retail/Local Politics		
141	Revenue growth	Industry		
142	Revised estimate	Corporate Earnings		
143	Russia	International Affairs	✓	
144	Safety administrations	Government		
145	Sales call	Social/Cultural		
146	Savings & loans	Banks	✓	✓
147	Scenario analysis	Science/Language		
148	Schools	Social/Cultural		
149	SEC	Buyouts & Bankruptcy		
150	Share payouts	Financial Markets		
151	Short sales	Financial Markets	✓	
152	Size	Science/Language		
153	Small business	Industry		
154	Small caps	Financial Markets	✓	✓
155	Small changes	Corporate Earnings		
156	Small possibility	Challenges		
157	Soft drinks	Industry		
158	Software	Technology	✓	
159	Southeast Asia	International Affairs		
160	Space program	Science/Language		
161	Spring/summer	Challenges		
162	State politics	Government		
163	Steel	Oil & Mining		
164	Subsidiaries	Industry		
165	Systems	Science/Language		
166	Takeovers	Buyouts & Bankruptcy	✓	
167	Taxes	Labor/income	✓	

168	Terrorism	Terrorism/Mideast		
169	Tobacco	Industry		
170	Trade agreements	International Affairs	✓	✓
171	Trading activity	Financial Markets		
172	Treasury bonds	Financial Markets		
173	UK	International Affairs		
174	Unions	Labor/income	✓	✓
175	US defense	Trans/Retail/Local Politics	✓	
176	US Senate	Leaders		
177	Utilities	Government	✓	
178	Venture capital	Industry		
179	Watchdogs	Government		
180	Wide range	Science/Language		

Table A3.1: News series – The column *News topic* are the news series topics, column *Meta topic* are meta topics/groups of news series. Columns *LASSO* and *sg-LASSO* reports whether the series was selected (✓) or not by the respective initial estimator.

Chapter 4

Machine Learning Panel Data Regressions with Heavy-tailed Dependent Data: Theory and Applications

with Andrii Babii, Ryan Ball and Eric Ghysels

4.1 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, Ghysels, and Striaukas (2021b), 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 (group) LASSO regularization; see Harding and Lamarche (2019), Chiang, Rodrigue, and Sasaki (2019), Chernozhukov, Hausman, and Newey (2019), Belloni, Chen, Padilla, et al. (2019), Belloni, Chernozhukov, Hansen, and Kozbur (2016), Lu and Su (2016), Kock (2016), Su, Shi, and Phillips (2016), Farrell (2015), Kock (2013), Lamarche (2010), Koenker (2004),

among others. However, to the best of our knowledge, the existing literature relates mostly to the microeconometric problems and does not address comprehensively (1) the advantages of long panels; (2) the performance of regularized panel data estimators with potentially heavy-tailed covariates and regression errors, (3) the debiased HAC inference for regularized panel data, and (4) the sg-LASSO regularization of Simon, Friedman, Hastie, and Tibshirani (2013) in a panel data setting.

We recognize that the economic and financial time series data are often persistent with fat tails. To that end, we introduce a new Fuk-Nagaev concentration inequality for long panels. Using this inequality, we obtain oracle inequalities for the sg-LASSO that shed new light on how the predictive performance of pooled and fixed effect estimators scales with N (cross-section) and T (time series), which is especially relevant for modern panel data applications, where both N and T can be large; see Fernández-Val and Weidner (2016), Hansen (2007), Alvarez and Arellano (2003), Hahn and Kuersteiner (2002), and Phillips and Moon (1999), among others. Importantly, our theory covers the LASSO and the group-LASSO estimators as special cases of sg-LASSO.

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

In our second 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 are able to 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' consensus earnings prediction errors are systematically related to either one of the aforementioned variables.

The chapter is organized as follows. Section 4.2 introduces the models and estimators. Oracle inequalities for sg-LASSO panel data regressions appear in Section 4.3. Section 4.4 develops the debiased HAC inference for regularized panel data regressions. The results of our empirical applications are reported in Section 4.6. Section 4.7 concludes. All technical details and detailed data descriptions appear in the Appendix section.

Notation: For a random variable $X \in \mathbf{R}$, let $\|X\|_q = (\mathbb{E}|X|^q)^{1/q}$ be its L_q norm with $q \geq 1$. For $p \in \mathbf{N}$, put $[p] = \{1, 2, \ldots, p\}$. For a vector $\Delta \in \mathbf{R}^p$ and a subset $J \subset [p]$, let Δ_J be a vector in $\mathbf{R}^p$ with the same coordinates as Δ on J and zero coordinates on J^c. Let $\mathcal{G}$ be a partition of $[p]$ defining the group structure, which is assumed to be known to the econometrician. For a vector $\beta \in \mathbf{R}^p$, the sparse-group structure is described by a pair $(S_0, \mathcal{G}_0)$, where $S_0 = \{j \in [p] : \beta_j \neq 0\}$ and $\mathcal{G}_0 = \{G \in \mathcal{G} : \beta_G \neq 0\}$ are the support and respectively the group support of β.

We also use $|S|$ to denote the cardinality of a set S. For $b \in \mathbf{R}^p$, its ℓ_q norm is denoted as $|b|_q = (\sum_{j\in[p]} |b_j|^q)^{1/q}$ if $q \in [1, \infty)$ and $|b|_\infty = \max_{j\in[p]} |b_j|$ if $q = \infty$. For a group structure $\mathcal{G}$, the $\ell_{2,1}$ group norm of $b \in \mathbf{R}^p$ is defined as $\|b\|_{2,1} = \sum_{G\in\mathcal{G}} |b_G|_2$. For $\mathbf{u}, \mathbf{v} \in \mathbf{R}^J$, the empirical inner product is defined as $\langle \mathbf{u}, \mathbf{v} \rangle_J = J^{-1} \sum_{j=1}^J u_j v_j$ with the induced empirical norm $\|.\|_J^2 = \langle ., . \rangle_J = |.|_2^2/J$. For a symmetric $p \times p$ matrix A, let $\text{vech}(A) \in \mathbf{R}^{p(p+1)/2}$ be its vectorization consisting of the lower triangular and the diagonal elements. Let A_G be a sub-matrix consisting of rows of A corresponding to indices in $G \subset [p]$. If $G = \{j\}$ for some $j \in [p]$, then we simply write $A_G = A_j$. Let $\|A\|_\infty = \max_{j\in[p]} |A_j|$ be the matrix norm. For $a, b \in \mathbf{R}$, we put $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Lastly, we write $a_n \lesssim b_n$ if there exists a (sufficiently large) absolute constant C such that $a_n \leq C b_n$ for all $n \geq 1$ and $a_n \sim b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

4.2 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

$$\psi(L^{1/m}; \beta_k) x_{i,t,k} = \frac{1}{m} \sum_{j=1}^{m} \beta_{j,k} x_{i,t-(j-1)/m,k} \tag{4.1}$$

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 covariates. When $m = 1$, we retain the standard panel data regression model

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

while $m > 1$ signifies that the high-frequency lags of $x_{i,t,k}$ are also included. The large number of predictors K with potentially large number of high-frequency measurements m can be a rich source of predictive information, yet at the same time, estimating $N + m \times K$ parameters is costly and may reduce the predictive performance in small samples.

To reduce the proliferation of lag parameters, we follow the MIDAS literature; see Ghysels, Santa-Clara, and Valkanov (2006), Ghysels, Sinko, and Valkanov (2006), and Babii, Ghysels, and Striaukas (2021a,b). Instead of estimating m individual slopes of high-frequency covariate $k \in [K]$ in equation ((4.1)), with some abuse of notation, we estimate a weight function ω parameterized by $\beta_k \in \mathbf{R}^L$ with $L < m$

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

where

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

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

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

More generally, we can use Gegenbauer polynomials, trigonometric polynomials, or wavelets. The orthogonal polynomials usually have better numerical properties than their popular non-orthogonal counterpart, such as the Almon (1965) lag structure. The attractive feature of linear in parameters dictionaries is that we can map the MIDAS regression to the linear regression framework that can be solved via a convex optimization. To that end, define $\mathbf{x}_i=(X_{i,1}W,\ldots,X_{i,K}W)$, where for each $k\in[K]$,

$$X_{i,k}=(x_{i,t-(j-1)/m,k})_{t\in[T],j\in[m]}$$

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

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

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

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

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

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

The MIDAS approach allows us to effectively reduce the dimensionality pertaining to the high-frequency lags. Alternatively, we may apply what is known as the UMIDAS scheme, see e.g., Foroni, Marcellino, and Schumacher (2015a), and directly estimate the coefficients associated with each high-frequency covariate lags separately (see equation ((4.7)) in Section 4.5 for example). Such a strategy, which as Foroni, Marcellino, and Schumacher (2015a) 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, Ghysels, and Striaukas (2021b) for further discussion. Also, while assuming that the individual lag coefficients in equation (4.1) are approximately sparse is *highly* restrictive, the approximate sparsity of slopes of the dictionary elements $(w_l)_{l\geq 0}$ is plausible. For instance, if $w_0(s) = 1$ with $\beta_{0,k} \neq 0$ and $\beta_{l,k} = 0, \forall l \geq 1$, we recover the averaging of high-frequency lags of covariate k as a special case. More generally, the weight ω may be a decreasing function over lags and we may want to learn its shape from the data maximizing the predictive performance.[2]

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

$$\min_{(a,b)\in\mathbf{R}^{N+LK}} \|\mathbf{y} - Ba - \mathbf{X}b\|_{NT}^2 + 2\lambda\Omega(b), \tag{4.2}$$

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

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

is a regularizing functional, which is a linear combination of LASSO and group LASSO penalties. The parameter $\gamma \in [0,1]$ determines the relative weights of the ℓ_1 (sparsity) and the $\ell_{2,1}$ (group sparsity) norms, while the amount of regularization is controlled by the regularization parameter $\lambda \geq 0$. Recall also that for a group structure $\mathcal{G}$ described as a partition of $[p] =$

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

$\{1, 2, \ldots, p\}$, the group LASSO norm is computed as $\|b\|_{2,1} = \sum_{G \in \mathcal{G}} |b_G|_2$. The group structure is assumed to be known to the econometrician, which in our setting corresponds to time series lags of covariates. More generally, we may also combine covariates of a similar nature in groups. Throughout the paper we assume that groups have fixed size, which is well-justified in our empirical applications.[3] Therefore, the selection of covariates is performed by the group LASSO penalty, which encourages sparsity between groups. In addition, the ℓ_1 LASSO norm promotes sparsity within groups and allows us to learn the shape of the MIDAS weights from the data.

It is worth mentioning that the linear in parameters approximation to the MIDAS weight function leads to the convex optimization parameter problem in equation ((3.1)) that can be solved efficiently, e.g., via the proximal gradient descent algorithm, or its block-coordinate descent versions. In contrast, a popular beta weights leads to a nonlinear non-convex optimization problem that becomes challenging to solve in high-dimensions; cf. Marsilli (2014b) and Khalaf, Kichian, Saunders, and Voia (2021).

4.3 Oracle inequalities

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

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

This model subsumes the mixed-frequency data regressions as a special case, in which case covariates are obtained, e.g., from the aggregation with Legendre polynomials. 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] See Babii (2020) for a continuous-time mixed-frequency regression where the group size is allowed to increase with the sample size under the in-fill asymptotics.

4.3.1 τ-mixing

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

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

where $\text{Lip}_1 = \{f : \mathbf{R}^l \to \mathbf{R} : \ |f(x) - f(y)| \leq |x - y|_1\}$ is a set of 1-Lipschitz functions from $\mathbf{R}^l$ to $\mathbf{R}$.[4] For a stochastic process $(\xi_t)_{t \in \mathbf{Z}}$ with a natural filtration generated by its past $\mathcal{M}_t = \sigma(\xi_t, \xi_{t-1}, \dots)$, the τ-mixing coefficients are defined as

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

where the supremum is taken over all $t, t_1, \dots, t_j \in \mathbf{Z}$. If $\tau_k \downarrow 0$, as $k \uparrow \infty$ then the process is called τ-mixing. The class of τ-mixing processes can be placed somewhere between the α-mixing processes and mixingales — the τ-mixing condition is less restrictive than the α-mixing condition,[5] yet at the same time, there exists a convenient for us coupling result for τ-mixing processes, which is not the case for the mixingales or near-epoch dependent processes; see Dedecker and Doukhan (2003) and Dedecker and Prieur (2004, 2005) for more details. This allows us to obtain concentration inequalities and performance guarantees for the sg-LASSO estimator; see Appendix A4.1 for more details.

4.3.2 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

$$\min_{r=(a,b) \in \mathbf{R}^{1+p}} \|\mathbf{y} - a\iota - \mathbf{X}b\|_{NT}^2 + 2\lambda\Omega(b). \tag{4.3}$$

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 $i \in [N]$. The following assumption imposes mild restrictions on the data.

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

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

Assumption 4.3.1 (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 4.3.2 (Covariance matrix). *The smallest eigenvalue of* $\Sigma_{N,T}$ *is uniformly bounded away from zero by some universal constant* $\gamma_{\min} > 0$.

Assumption 4.3.2 is satisfied for the spiked identity and Topelitz covariance structures. It can be interpreted as a completeness condition, see Babii and Florens (2020), and can also be relaxed to the restricted eigenvalue condition imposed on the population covariance matrix $\Sigma_{N,T}$; see Babii, Ghysels, and Striaukas (2021b). 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, Florens, and Renault (2007).

Lastly, we assume that the regularization parameter λ scales appropriately with the number of covariates p, the length of the panel T, the size of the cross-section N, and a certain exponent κ that depends on the tail parameter q and the persistence parameter a. The precise order of the regularization parameter is described by the Fuk-Nagaev inequality for long panels appearing in the Appendix; see Theorem A4.1.

Assumption 4.3.3 (Regularization)**.** *For some* $\delta \in (0,1)$

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

where $\kappa = ((a+1)q-1)/(a+q-1)$ *and* a, q *are as in Assumptions 4.3.1.*

Our first result is the oracle inequality for the pooled sg-LASSO estimator described in equation (4.3). The result allows for misspecified regressions with a non-trivial approximation error in the sense that we consider more generally

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

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

Theorem 4.1. *Suppose that Assumptions 4.3.1, 4.3.2, and 4.3.3 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 4.1 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 4.3.1. Parameters κ and $\tilde{\kappa}$ are the dependence-tails exponents for stochastic processes driving the regression score and the covariance matrix respectively. Theorem 4.1 shows that the prediction and the estimation accuracy of pooled

panel data regressions improves when the sparse-group structure is taken into account. Indeed, for the LASSO regression, the effective sparsity reduces to $s^{1/2} = \sqrt{|S_0|}$, which is larger than $\gamma\sqrt{|S_0|} + (1-\gamma)\sqrt{|\mathcal{G}_0|}$ in the case of sg-LASSO.

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

Assumption 4.3.4. *(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 4.3.4 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 4.3.4 and is an immediate consequence of Theorem 4.1.

Corollary 4.3.1. *Suppose that Assumptions 4.3.1, 4.3.2, 4.3.3, and 4.3.4 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 4.3.1 describes the prediction and the estimation accuracy of pooled sparse-group panel data regressions. It suggests that the predictive performance of the sg-LASSO (and consequently LASSO and group LASSO) regressions may deteriorate when regression errors and/or predictors are heavy-tailed or when the data are extremely persistent. However, for geometrically ergodic Markov processes, e.g., stationary AR(1) process, the τ-mixing coefficients decline geometrically fast, so that $\kappa \approx q$ and $\tilde{\kappa} \approx \tilde{q}$. In this case, the prediction accuracy scales approximately at the rate $O_P\left(\frac{p^{2/q}}{(NT)^{2-2/q}} \vee \frac{\log p}{NT}\right)$ and the predictive performance may be affected only by the tails constant q.

If additionally, the data are sub-Gaussian, then moments of all order $q \geq 2$ exist and for any particular effective sample size NT, the first

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

Theorem 4.1 and Corollary 4.3.1 imply two practical consequences: (1) one may want to exclude (or suitably transform) the heavy-tailed series from the high-dimensional predictive regressions based on the preliminary estimates of the tail index, e.g., using the Hill estimator; (2) if the individual heterogeneity can be ignored, then pooling panel data can improve significantly the predictive performance. In the latter case, one can also preliminary cluster similar series in groups, e.g., based on the unsupervised clustering algorithms, which may strike a good balance between the pooling benefits and heterogeneity.

4.3.3 Fixed effects

Pooled regressions are attractive since the effective sample size NT can be huge, yet the heterogeneity of individual time series may be lost. If the underlying series have a substantial heterogeneity over $i \in [N]$, then taking this into account might reduce the projection error and improve the predictive accuracy. At a very extreme side, the cross-sectional structure can be completely ignored and individual time-series regressions can be used for prediction. The fixed effects panel data regressions strike a good balance between the two extremes controlling for heterogeneity with entity-specific intercepts.

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

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

where $B = I_N \otimes \iota$, I_N is $N \times N$ identity matrix, $\iota \in \mathbf{R}^T$ is an all-ones vector, and Ω is the sg-LASSO regularizing functional. It is worth stressing that the design matrix $\mathbf{X}$ does not include the intercept and that we do not penalize the fixed effects, that are typically not sparse. By Fermat's

rule, the first-order conditions are

$$\begin{aligned} \hat{\alpha} &= (B^\top B)^{-1} B^\top (\mathbf{y} - \mathbf{X}\hat{\beta}) \\ 0 &= \mathbf{X}^\top M_B (\mathbf{X}\hat{\beta} - \mathbf{y})/NT + \lambda z^* \end{aligned} \tag{4.4}$$

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

With some abuse of notation, redefine

$$\hat{\Sigma}_{N,T} = \begin{pmatrix} \frac{1}{T} B^\top B & \frac{1}{\sqrt{NT}} B^\top \mathbf{X} \\ \frac{1}{\sqrt{NT}} \mathbf{X}^\top B & \frac{1}{NT} \mathbf{X}^\top \mathbf{X} \end{pmatrix} \quad \text{and} \quad \Sigma_{N,T} = \begin{pmatrix} I_N & \frac{1}{\sqrt{NT}} \mathbb{E}\left[B^\top \mathbf{X}\right] \\ \frac{1}{\sqrt{NT}} \mathbb{E}\left[\mathbf{X}^\top B\right] & \mathbb{E}[x_{i,t} x_{i,t}^\top] \end{pmatrix}. \tag{4.5}$$

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

The order of the regularization parameter is governed by the Fuk-Nagaev inequality for long panels; see Appendix, Theorem A4.1.

Assumption 4.3.5 (Regularization)**.** *For some $\delta \in (0,1)$*

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

where $\kappa = ((a+1)q - 1)/(a + q - 1)$, and a, q are as in Assumptions 4.3.1.

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

Theorem 4.2. *Suppose that Assumptions 4.3.1, 4.3.2, and 4.3.5 are satisfied. Then with probability at least $1 - \delta - O(r^{\text{fe}}_{N,T})$*

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

Theorem 4.2 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 4.3.6. *Suppose that (i)* $\|\mathbf{m} - \mathbf{Z}\rho\|_{NT}^2 = O_P((s \vee N)\lambda^2)$*; (ii)* $(p + N^{\tilde{\kappa}/2})p(s \vee N)^{\tilde{\kappa}} N^{1-\tilde{\kappa}} T^{1-\tilde{\kappa}} \to 0$ *and* $p(p \vee N)e^{-cNT/(s \vee N)^2} \to 0$.

The following corollary is an immediate consequence of Theorem 4.2.

Corollary 4.3.2. *Suppose that Assumptions 4.3.1, 4.3.2, 4.3.5, and 4.3.6 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 4.3.2 allows for $s, p, N, T \to \infty$ at appropriate rates. However, we pay additional price for estimating N fixed effects which plays a similar role to the effective dimension of covariates. An immediate practical implication is that in order to achieve accurate predictions with high-dimensional fixed effect regressions, the panel has to be sufficiently long in order to offset the estimation error of the individual fixed effects. Likewise, the tails and the persistence of the data may also reduce the prediction accuracy in small samples through κ, which is approximately equal to q for geometrically decaying τ-mixing coefficients.

4.4 Debiased inference

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

1. For each $j \in [p+1]$, let $\hat{\mu}_j = (\hat{\mu}_{j,1}, \ldots, \hat{\mu}_{j,p})^\top$ be a solution to

$$\min_{\mu \in \mathbf{R}^p} \|\mathbf{Z}_j - \mathbf{Z}_{-j}\mu\|_{NT}^2 + 2\lambda_j|\mu|_1,$$

where $\mathbf{Z}_j$ is $NT \times 1$ vector of stacked observations $\{z_{i,t,j} \in \mathbf{R} : i \in [N], t \in [T]\}$ and $\mathbf{Z}_{-j}$ is the $NT \times p$ matrix of stacked observations

$\{(z_{i,t,k})_{k\neq j} \in \mathbf{R}^p : i \in [N], t \in [T]\}$. Put

$$\hat{\sigma}_j^2 = \|\mathbf{Z}_j - \mathbf{Z}_{-j}\hat{\mu}_j\|_{NT}^2 + \lambda_j|\hat{\mu}_j|,$$

2. Compute $\hat{\Theta} = \hat{B}^{-1}\hat{C}$, where $\hat{B} = \text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_{p+1}^2)$, and

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

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

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

Assumption 4.4.1. *(i)* $\sup_z \mathbb{E}[u_{i,t}^2 | z_{i,t} = z] = O(1)$*; (ii)* $\|\Theta_G\|_\infty = O(1)$ *for* $G \subset [p+1]$ *of fixed size; (iii) the long run variance of* $(u_{i,t}^2)_{t\in\mathbf{Z}}$ *and* $(v_{i,t,j}^2)_{t\in\mathbf{Z}}$ *exists for every* $j \in G$*; (iv)* $s^{*2}\log^2 p/T \to 0$ *and* $p/\sqrt{T^{\kappa-2}\log^\kappa p} \to 0$*; (v)* $\|\mathbf{m} - \mathbf{Z}\rho\|_{NT} = o_P(1/\sqrt{NT})$*; (vi) for every* $j, l \in [p]$ *and* $k \geq 0$*, the* τ*-mixing coefficients of* $(u_{i,t}u_{i,t+k}x_{i,t,j}x_{i,t+k,l})_{t\in\mathbf{Z}}$ *are* $\check{\tau}_t \leq ct^{-d}$ *for some universal constants* $c > 0$ *and* $d > 1$*; (vi) for each* i*,* $\{(u_{i,t}, z_{i,t}^\top)^\top : t \in \mathbf{Z}\}$ *is a stationary process that is also i.i.d. over* i*, Assumption 4.3.1 holds with* $a > (q-1)/(q-2) \vee (q\delta+1)/(q-2-\delta)$ *with* $q > 2+\delta$ *and* $\delta > 0$.

Assumption 4.4.1 (i) requires that the conditional variance of the regression error is bounded. Condition (ii) requires that the rows of the precision matrix have bounded ℓ_1 norm and is a plausible assumption in the high-dimensional setting, where the inverse covariance matrix is often sparse. Condition (iii) is a mild restriction needed for the consistency of the sample variance of regression errors. The rate conditions in (iv) is similar to the condition used in Babii, Ghysels, and Striaukas (2021a). Lastly, condition (v) is trivially satisfied when the projection coefficients are sparse and, more generally, it requires that the misspecification error vanishes asymptotically sufficiently fast.

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

Theorem 4.1. *Suppose that Assumptions 4.3.1, 4.3.2, 4.3.3, 4.3.4, and 4.4.1 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 $\Xi_G = \lim_{T\to\infty} \text{Var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} u_{i,t}\Theta_G z_{i,t}\right)$.

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

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

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

where $\hat{\Gamma}_{k,i} = \hat{\Theta}_G\left(\frac{1}{T}\sum_{t=1}^{T-k}\hat{u}_{i,t}\hat{u}_{i,t+k}x_{i,t}x_{i,t+k}^{\top}\right)\hat{\Theta}_G^{\top}$, $\hat{u}_{i,t}$ is the sg-LASSO residual, and $\hat{\Gamma}_{-k,i} = \hat{\Gamma}_{k,i}^{\top}$. The kernel function $K : \mathbf{R} \to [-1, 1]$ with $K(0) = 1$ is puts less weight on more distant noisy covariances, while $M_T \uparrow \infty$ is a bandwidth (or lag truncation) parameter; see Babii, Ghysels, and Striaukas (2021a) for more details as well as formal results on the validity of HAC-based inference using sg-LASSO residuals.

4.5 Monte Carlo experiments

This section presents Monte Carlo simulation results.

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

Second, 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.

4.5.1 Simulation design (nowcasting)

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

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

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

We set $\rho_1 = 0.4$, $\rho_2 = 0.01$, and take the number of relevant high frequency regressors $K = 6$. We are interested in quarterly/monthly data, and use four quarters of data for the high frequency regressors so that $m = 12$, which covers four low frequency lags of each high frequency regressor. The high frequency regressors are generated as K i.i.d. realizations of the univariate autoregressive (AR) process $x_h = \rho x_{h-1} + \varepsilon_h$, where $\rho = 0.6$ and either $\varepsilon_h \sim_{i.i.d.} N(0,1)$ or $\varepsilon_h \sim_{i.i.d.}$ student-$t(5)$, where h denotes the high-frequency sampling. We rely on a commonly used weighting scheme in the MIDAS literature, namely $\omega(s; \beta_k)$ for $k = 1, 2, \ldots, 6$ are determined by beta densities respectively equal to Beta$(1,3)$ for $k = 1, 4$, Beta$(2,3)$ for $k = 2, 5$, and Beta$(2,2)$ for $k = 3, 6$; see Ghysels, Sinko, and Valkanov (2007) or Ghysels and Qian (2019), for further details. The MIDAS regressions are estimated using Legendre polynomials of degree $L = 3$. Lastly, we draw the intercepts $\alpha \sim$ Uniform$(-4, 4)$.

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

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

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

For the *Baseline scenario*, in the estimation procedure we add 24 noisy covariates which are generated in the same way as the relevant covariates,

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

4.5.2 Simulation results (nowcasting)

Table 4.1 covers the average mean squared forecast errors for one-step-ahead nowcasts. We report results for pooled panel data (left block) and fixed effects (right block) estimators. First, for all DGPs and both estimators, structured sg-LASSO-MIDAS performs better compared to unstructured elastic net. In the case of sg-LASSO-MIDAS the best performance is achieved for $\gamma \notin \{0, 1\}$ for both pooled panel data and fixed effects cases, while $\gamma = 0$, i.e. ridge regression, seems to dominate in the case of elastic net for both the pooled and fixed effects cases. For the student-$t(5)$ and large dimensional DGP, we observe a decrease in the performance for all methods. However, the decrease in the performance is larger for the student-$t(5)$ DGP, suggesting that heavy-tailed data may have a stronger impact on the performance of the estimators.

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

4.5.3 Simulation design (Granger causality)

We simulate the data from the following DGP:

$$y_{i,t} = \alpha + \rho y_{i,t-1} + \sum_{k=1}^{K} \frac{1}{m} \sum_{j=1}^{m} \omega((j-1)/m; \beta_k) x_{i,t-(j-1)/m,k} + u_{i,t}, \quad (4.6)$$

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

Table 4.1: The table reports simulation results for nowcasting accuracy for pooled and fixed effects estimators. Panel A. reports results for the baseline DGP, Panel B. for student-$t(5)$ DGP and Panel C. for large dimensional DGP with 100 time-varying covariates. We vary the cross-sectional dimension $N \in \{25, 75\}$ and time series dimension $T \in \{50, 100\}$.

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

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, Sinko, and Valkanov (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, Marcellino, and Schumacher (2015a), UMIDAS consists of estimating a regression coefficient for each high-frequency lag separately, and therefore the weight function for each covariate is

$$\sum_{j=1}^{m} \omega((j-1)/m; \beta_k) x_{i,t-(j-1)/m,k} = \sum_{j=1}^{m} b_{j,k} x_{i,t-(j-1)/m,k} \tag{4.7}$$

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, Ghysels,

and Striaukas (2021a) propose tests of Granger causality in univariate regularized regressions and high-dimensional data.

Table 4.2: 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

4.5.4 Simulation results (Granger causality)

In Table 4.2, 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 4.2 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 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 and that pooling individual series improves the results even further.

4.6 Empirical Applications

The fundamental value of equity shares is determined by the discounted value of future payoffs. Every quarter investors get a glimpse of a firms' potential payoffs with the release of corporate earnings reports. In a data-rich environment, stock analysts have many indicators regarding future cash flows that are available much more frequently. Ball and Ghysels (2018) took a first stab at automating the process using MIDAS regressions. Since their original work, much progress has been made on machine learning regularized mixed frequency regression models. In the Section 4.6.1, we significantly expand the tools of nowcasting in a data-rich environment by exploiting panel data structures. Panel data regression models are well suited for the firm-level data analysis as both time series and cross-section dimensions can be properly modeled. In such models, time-invariant firm-specific effects are typically modeled in a flexible way which allows capturing heterogeneity

in the data. At the same time, machine learning methods are becoming increasingly popular in economics and finance as a flexible way to model relationships between the response and covariates.

In our second application, 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.

4.6.1 Nowcasting P/E ratios

In our first empirical application, we consider nowcasting the P/E ratios of 210 US firms using a set of predictors that are sampled at mixed frequencies. We use 24 predictors, including traditional macro and financial series as well as non-standard series generated by the textual analysis. We apply pooled and fixed effects sg-LASSO-MIDAS panel data models and compare them with several benchmarks such as random walk (RW), analysts consensus forecasts, and unstructured elastic net. We also compute predictions using individual-firm high-dimensional time series regressions and provide results for several choices of the tuning parameter. Lastly, we provide results for low-dimensional single-firm MIDAS regressions using forecast combination techniques used by Andreou, Ghysels, and Kourtellos (2013) and Ball and Ghysels (2018). The latter is particularly relevant regarding the analysis in the current paper as it also deals with nowcasting price earnings ratios. The forecast combination methods consist of estimating ARDL-MIDAS regressions with each of the high-frequency covariates separately. In our case this leads to 24 predictions, corresponding to the number of predictors. Then a combination scheme, typically discounted mean squared error type, produces a single nowcast. One could call this a pre-machine learning large dimensional approach. It will, therefore, be interesting to assess how this approach compares to the regularized MIDAS panel regression machine learning approach introduced in this chapter.

We consider nowcasting the P/E ratios of 210 US firms using a set of predictors that are sampled at mixed frequencies. We use 24 predictors, including traditional macro and financial series as well as non-standard series generated by textual analysis of financial news. We apply pooled and individual fixed effects sg-LASSO-MIDAS panel data models and

compare them with several benchmarks such as random walk (RW), analysts consensus forecasts, and unstructured elastic net.

We also compute predictions using individual-firm high-dimensional time series regressions and provide results for several choices of the tuning parameter. Moreover, our analysis includes results for sg-LASSO-MIDAS panel data models which include the median consensus analysts predictions. Adding the consensus forecast as a regressor allows us to address besides the question of ML versus analysts also the topic of a combined ML/analyst nowcasts – a theme explored by Ball and Ghysels (2018). Our analysis includes formal significance testing of predictors in the augmented sg-LASSO-MIDAS panel data model which allows us to determine whether analysts take all relevant information to them fully into account.

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

The remainder of the section is structured as follows. We start with a short review of the data, with more detailed descriptions and tables appearing in Appendix Section A4.6, followed by a summary of the methods and empirical results.

4.6.1.1 Data description

The full sample consists of observations between the 1^{st} of January, 2000 and the 30^{th} of June, 2017. Due to the lagged dependent variables in the models, our effective sample starts the third fiscal quarter of 2000. We use the first 25 observations for the initial sample, and use the remaining 42 observations for evaluating the out-of-sample forecasts, which we obtain by using an expanding window forecasting scheme. We collect data from CRSP and I/B/E/S to compute the quarterly P/E ratios and firm-specific financial covariates; RavenPack is used to compute daily firm-level textual-

analysis-based data; real-time monthly macroeconomic series are from the FRED-MD dataset, see McCracken and Ng (2016) for more details; FRED is used to compute daily financial markets data and, lastly, monthly news attention series extracted from the *Wall Street Journal* articles is retrieved from Bybee, Kelly, Manela, and Xiu (2020).[6] Appendix Section A4.6 provides a detailed description of the data sources.

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

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

4.6.1.2 Tuning parameters

We consider several approaches to select the tuning parameter λ. First, we adapt the k-fold cross-validation to the panel data setting. To that end, we resample the data by blocks respecting the time-series dimension and creating folds based on individual firms instead of the pooled sample. We use 5-fold cross-validation as the sample size of the dataset we consider in our empirical application is relatively small. We also consider the following three information criteria: BIC, AIC, and corrected AIC (AICc) of Hurvich and Tsai (1989). Assuming that $y_{i,t}|x_{i,t}$ are i.i.d. draws from $N(\alpha_i + x_{i,t}^\top \beta, \sigma^2)$,

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

the log-likelihood of the sample is

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

Then, the BIC criterion is

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

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

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

Lastly, the corrected Akaike information criteria is

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

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

4.6.1.3 Models and main results

To compute forecasts, we estimate several regression models. First, we estimate the individual sg-LASSO-MIDAS regressions for each firm $i = 1, \ldots, N$, which in Table 4.3 we refer to as *Individual*,

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

where the firm-specific predictions are computed as $\hat{y}_{i,t+1} = \hat{\alpha}_i + x_{i,t+1}^\top \hat{\beta}_i$. As noted in Section 4.2, $\mathbf{x}_i$ contains lags of the low-frequency target variable and high-frequency covariates to which we apply Legendre polynomials of degree $L = 3$.

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

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

and compute predictions as

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

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

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

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

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

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

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

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

The main results for pooled panel data and fixed effects sg-LASSO-MIDAS regressions are reported in Table 4.3, while additional results for longer horizon predictions, unstructured LASSO estimators and the forecast combination approach appear in Appendix Tables A4.1-A4.3.

Table 4.3: Prediction results – The table reports average over firms MSEs of out-of-sample predictions. The nowcasting horizon is the current quarter, i.e. we predict the P/E ratio using information up to the end of current fiscal quarter. Block in Panel A1-D1 correspond to ML-only forecast errors while in Panel A2-D2 to ML models augmented with median consensus nowcasts. Each Panel A1-D1 and A2-D2 block represents different ways of calculating the tuning parameter λ. Bold entries are the best results in a block.

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

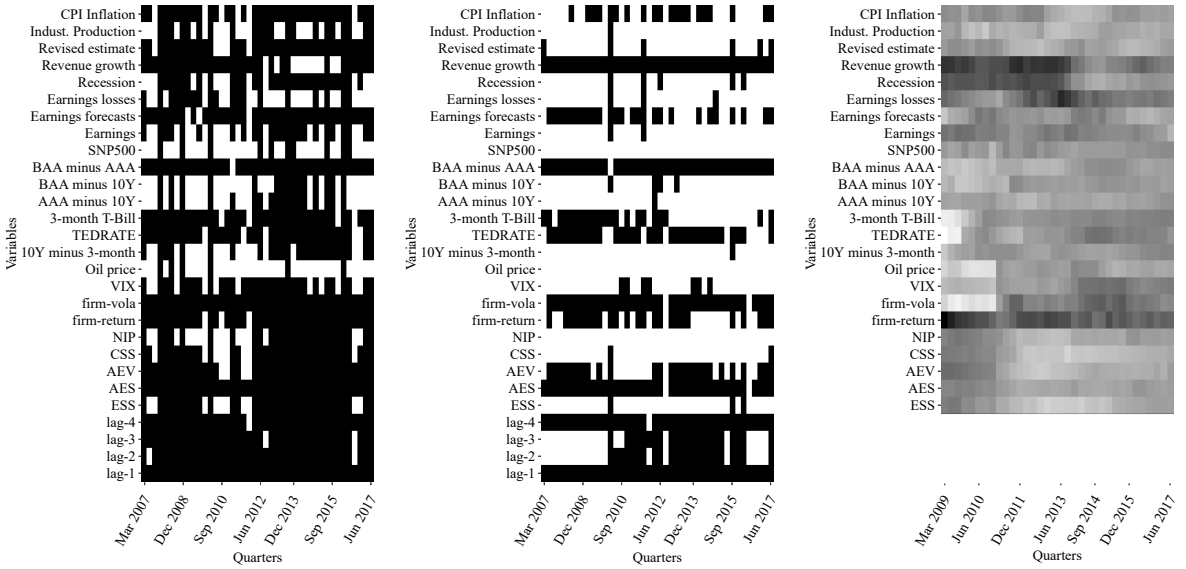
The first entries to Table 4.3 show that analysts-based predictions, both median and mean, have much larger mean squared forecast errors (MSEs) compared to model-based predictions. This is also the case for the RW predictions. The sharp increase in quality of model- versus analyst-based predictions indicates the usefulness of machine learning methods to nowcast P/E ratios, see Tables 4.3 Panel A1-D1 and A4.2. A better performance is achieved for almost all machine learning methods - single firm or panel data regressions - and all tuning parameter choices.[7] Table 4.3 also reports results for panel data ML methods augmented with median consensus analysts forecasts (see Panels A2-D2). Notably, it shows that the augmented models further improve upon ML-only models.

Turning to the comparison of model-based predictions, we see from the results in Table 4.3 that sg-LASSO-MIDAS panel data models improve the quality of predictions in comparison to individual sg-LASSO-MIDAS models irrespective of the γ weight or the tuning parameter choice. This indicates that panel data structures are relevant for nowcasting P/E ratios.[8] Among the panel data models, we observe that fixed effects regressions improve over the pooled regressions in most cases except when cross-validation is used, namely compare Panels A1 with Panel B1-D in Tables 4.3 and A4.3. The pooled model tuned by cross-validation seems to yield the best overall performance. In general, one can expect that cross-validation improves prediction performance over different tuning methods as it is directly linked to empirical risk minimization. In the case of fixed effects, however, we may lose the predictive gain due to the smaller samples with each fold used in estimating the model. Lastly, the best results per tuning parameter block seem to be achieved when $\gamma \notin \{0, 1\}$, indicating that both sparsity within the group and at the group level matters for prediction performance.

In Figure 4.1, we plot the sparsity patterns of the selected covariates for the two best-performing methods: (a) pooled sg-LASSO regressions, tuned using cross-validation with $\gamma = 0.4$, and (b) fixed effects sg-LASSO model with BIC tuning parameter and the same γ parameter. We also plot the forecast combination weights which are averaged over firms. The plots in Figure 4.1 reveal that the fixed effects estimator yields sparser models compared to pooled regressions, and the sparsity pattern is clearer. In the

[7]Similar findings for one-quarter ahead predictions are reported in Table A4.1. The unstructured panel data methods and the forecast combination approach also yield more accurate forecasts, see Appendix Table A4.2-A4.3. The latter confirms the findings of Ball and Ghysels (2018).

[8]We also report similar findings for unstructured estimators (see Table A4.2) and one quarter ahead forecasts (see Table A4.1).



(a) Pooled sg-LASSO, $\gamma = 0.4$, cross-validation. **(b)** Fixed effects sg-LASSO, $\gamma = 0.4$, BIC. **(c)** Average forecast combination weights.

Figure 4.1: Sparsity patterns and forecast combination weights.

fixed-effects case, the revenue growth and the first lag of the dependent variable are selected throughout the out-of-sample period. BAA minus AAA bond yield spread, firm-level volatility, and the aggregate event sentiment index are also selected quite frequently. Similarly, these variables are selected in the pooled regression, but the pattern is less apparent. The forecast combination weights seem to yield similar, yet more dispersed patterns.[9] In this case, revenue growth and firm-level stock returns covariates obtain relatively larger weights compared to the rest of covariates, particularly for the first part of the out-of-sample period. Therefore, the gain of machine learning methods - both single-firm and panel data - can be associated with sparsity imposed on the regression coefficients.

In addition it is worth noting that the textual news data analytics also appear in the models according the results displayed in Figure 4.1. These are the ESS, AES, AEV, CSS and NEP regressors described in detail in Appendix Section A4.6. Among them, as already noted, AES – the aggregate event sentiment index – features most prominently in the sg-LASSO models. It is worth emphasizing that the time series of news data is sparse as many days are without firms-specific news. For such

[9]Note that forecast combination weights start in 2009 Q1 due to the first eight quarters being used as a pre-sample to estimate weights, see Ball and Ghysels (2018) for further details. Also, the forecast combination weights figure does not contain autoregressive lags; all four lags are always included in all forecasting regressions.

days, we impute zero values. The nice property of our mixed frequency data treatment with dictionaries, imputing zeros also implies that non-zero entries get weights with a decaying pattern for distant past values.

Finally, Figure 4.2 shows the flexibility of our approach when dealing with high-dimensional MIDAS panel data models. First, we show that various shapes and forms of the weighting function can be estimated by applying Legendre polynomials over the high-frequency lags. For instance, the BAA minus 10-Year Treasury bond yield spread is estimated to have slowly decaying weights, while the TED rate covariate has a humped shape of the weights. Our approach provides a foundation for future research that focuses on the economic interpretations of the various MIDAS polynomial shapes (e.g., Ball (2013); Ball and Easton (2013); Ball and Gallo (2018)). Finally, our approach allows for the recovery of smooth lag functions for such series, even for daily textual news series that are sparse, see Figure 4.2 (a) and (e).

4.6.1.4 Significance test of nowcasts

To test for the superior forecast performance, we use the Diebold and Mariano (1995) test for the pool of P/E ratio nowcasts. We compare the median consensus forecasts versus panel data machine learning regressions with the smallest forecast error for pooled and fixed effects panel regressions and report the forecast accuracy test results in Table 4.4.

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

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

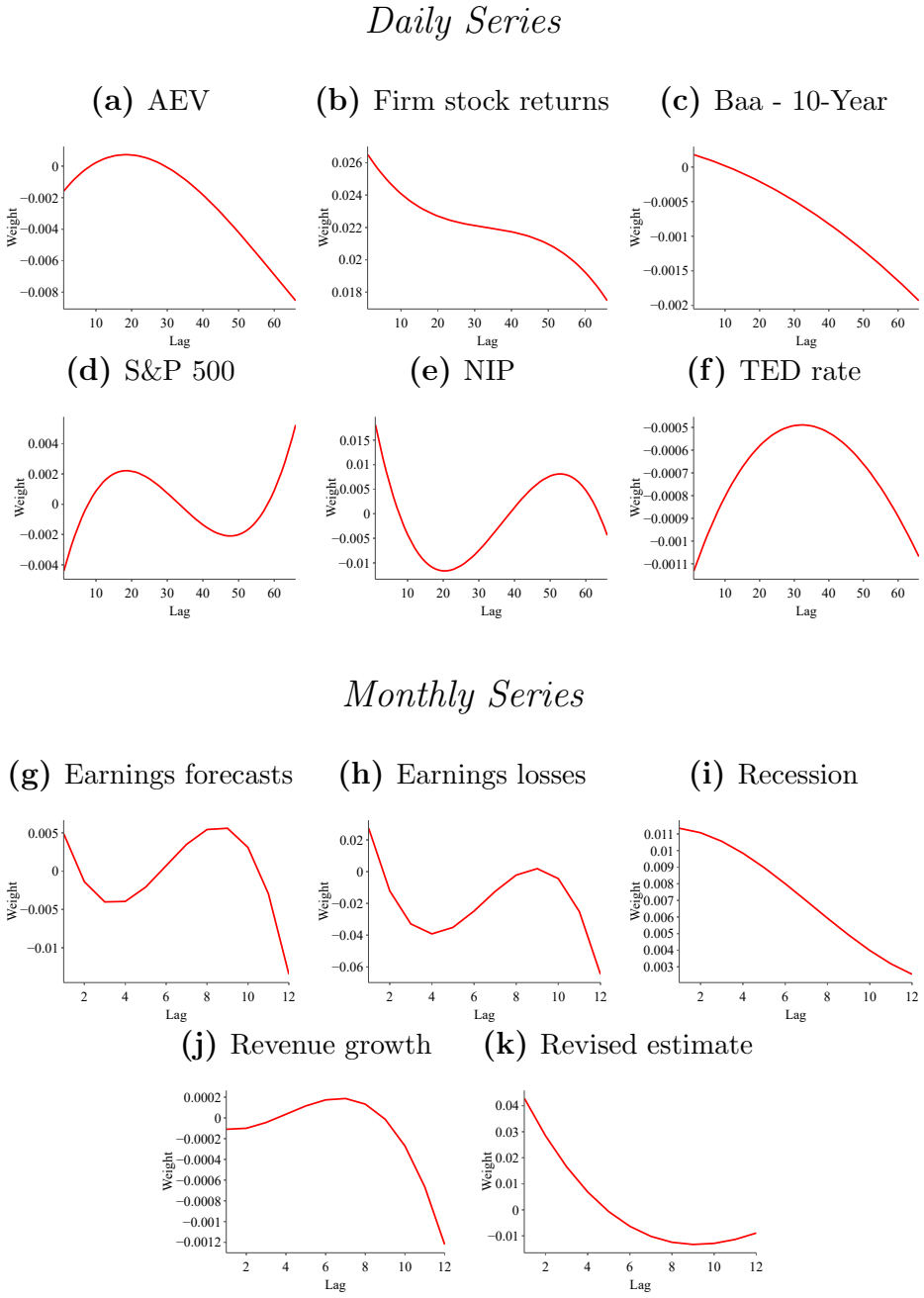


Figure 4.2: Weighting schemes for various covariates

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

Table 4.4: Forecasting performance significance – The table reports the Diebold and Mariano (1995) test statistic for pooled nowcasts comparing machine learning panel data regressions with analysts' implied median consensus forecasts. We compare panel models that have the smallest forecast error per tuning parameter block in Table 4.3 (sg-LASSO-MIDAS) and Table A4.2 (elastic net or elastic net UMIDAS) for pooled and fixed effects regressions respectively. We report test statistics for a) all firms in column *Full sample*, b) pooled firms where both sg-LASSO and elastic net outperform analysts in column *sg-LASSO & elnet*, c) pooled firms where sg-LASSO outperforms analysts but elastic net does not in column *sg-LASSO* , and d) where none of the machine learning methods outperforms analysts' forecasts in column *none*.

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

Large differences in prediction accuracy for different pools of P/E ratios may relate to the heavy-tailedness of regression errors.[11] In Table 4.5, we report the maximum likelihood estimates of the degrees of freedom parameter of a student-t distribution for the in-sample residuals pooled as in Table 4.4.[12] The smaller values indicate that the tails are heavier, while the larger values correspond to lighter, closer to Gaussian, tails. In line with our theory, the results show that LASSO-type regressions yield

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

[12] We follow a parametric approach since the time series are relatively short, see however also Appendix, Table A4.4 for the nonparametric tail index estimates.

much more accurate predictions when the residuals are less heavy-tailed. Interestingly, for the pool of firms where analysts' predictions are more accurate than both machine learning methods (column *none*), tails of the residuals appear to be the heaviest.

Lastly, we report the Diebold and Mariano (1995) test statistic comparing whether the sg-LASSO-MIDAS model combined with the median consensus nowcasts (An.-median) outperforms the analysts-only nowcasts (see Table 4.3 Panels A2-D2). Note that median consensus predictions are always selected by the sg-LASSO-MIDAS throughout out-of-sample period while other covariates retain a selection pattern similar to that of sg-LASS-MIDAS regressions reported in Figure 4.1. We pick the best panel model specification and compute the statistic of out-of-sample residuals. The statistic is 1.327, suggesting that combined model and analysts predictions seem to outperform analysts when using a one-sided 10% level test.

Table 4.5: Heaviness of tails – The table reports the maximum likelihood estimate of the degree of freedom of *student-t* distribution of in-sample residuals. The results are reported for the models as in Table 4.4.

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

4.6.2 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, similarly as in nowcasting application, we consider the earnings of 210 US firms using a set of predictors that are 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. We use similar data set as in nowcasting application, see Table A4.6, but include two additional predictors in unemployment rate and real GDP growth rate.

4.6.2.1 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:

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

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 10-fold cross-validation to tune both λ and γ, where we define folds as adjacent blocks over the 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 4.6.

In Panel (A) of Table 4.6 we find that the AR(1) lag is significant, leading us to conclude that the prediction errors made by the analysts

Table 4.6: 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 $\backslash M_T$	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

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 4.6 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 4.6 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.

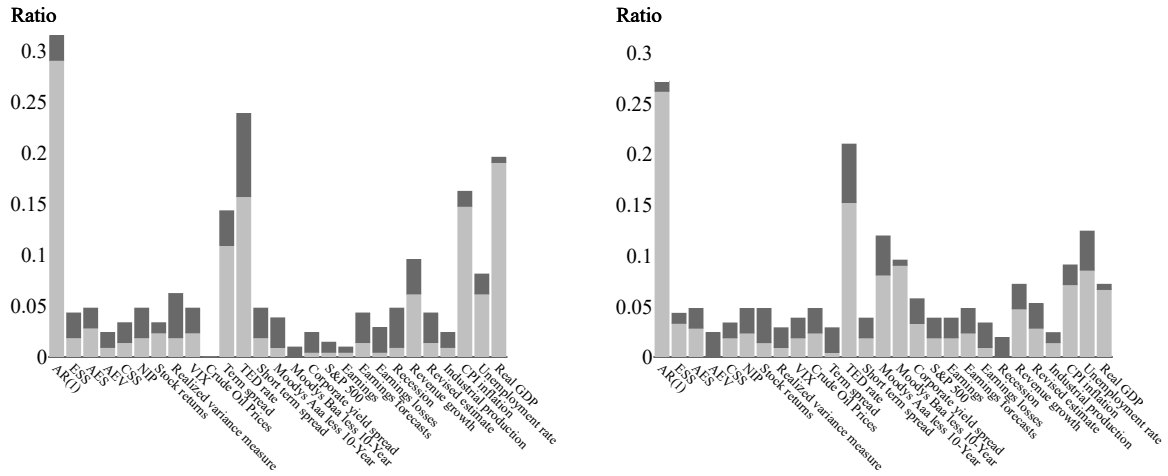
In Table 4.7 we show results based on a different way of pooling analysts' prediction errors $y_{i,t+1}$. We split the data into two parts based on how large the average disagreement among analysts is. For each firm, we compute the forecast disagreement as the difference between 95% and 5% percentile 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 Figure 4.3 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

Table 4.7: 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

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.



(a) sg-LASSO-MIDAS **(b)** LASSO-UMIDAS

Figure 4.3: 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 4.7. All results are based on the 5% significance level.

4.7 Conclusions

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

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

regressions outperform the forecast combinations and that the panel data approach improves upon the predictive time series regressions for individual firms.

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

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.

APPENDIX

A4.1 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 $\xi_{i,t} = (\xi_{i,t,1}, \dots, \xi_{i,t,p}) \in \mathbf{R}^p$, let $\tau_k^{(i,j)}$ denote the τ-mixing coefficient of $\xi_{i,t,j}$. The following result describes a Fuk-Nagaev concentration inequality for panel data. It is worth mentioning that the inequality does not follow from Babii, Ghysels, and Striaukas (2021a) and is of independent interest for the high-dimensional panel data.[13]

Theorem A4.1. *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 A4.1. 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.

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

2. $V_{i,t}^* \perp\!\!\!\perp (V_{i,1}, \ldots, V_{i,t-2})$.
3. $\|V_{i,t} - V_{i,t}^*\|_1 = \tau(\mathcal{M}_{i,t}, V_{i,t})$.

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

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

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

Decompose

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

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

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

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, Ghysels, and Striaukas (2021a), Lemma A.1.2 under assumptions (i) and (ii). Similarly,

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

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 3., this gives

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

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

$$\begin{aligned}\Pr\left(\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\xi_{i,t}\right|>u\right) &\leq \Pr(I>u/3)+\Pr(II>u/3)+\Pr(III>u/3)\\ &\leq c_1u^{-q}N\sum_{i=1}^{N}\sum_{t\geq 1}\|V_{i,t}\|_q^q+4e^{-c_2u^2/NT}+\\ &+\frac{3NT}{u}\max_{i\in[N]}\tau_{J+1}^{(i,1)}\\ &\leq c_1u^{-q}J^{q-1}NT+\frac{3NT}{u}(J+1)^{-a}+4e^{-c_2u^2/NT}\end{aligned}$$

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

$$\Pr\left(\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\xi_{i,t}\right|>u\right)\leq c_1NTu^{-\kappa}+4e^{-c_2u^2/NT}$$

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

It follows from Theorem A4.1 that there exists $C > 0$ such that for every $\delta \in (0, 1)$

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

Note that the inequality reflects the concentration jointly over N and T and that tails and persistence play an important role through the mixing-tails exponent κ. The inequality is a key technical tool that allows us to handle panel data with heavier than Gaussian tails and non-negligible T and N. It is worth mentioning that the concentration over N is also influence by the weak dependence, which probably can be relaxed with a sharper proof technique. However, for geometrically ergodic processes, e.g., for stationary $AR(p)$, we have $\kappa \approx q$, in which case the time series dependence does not influence the concentration at all.

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

Theorem A4.2. *Let $(\xi_t)_{t\in\mathbf{N}}$ be a centered stochastic process such that (i) there exists $q>2$ such that $\|\xi\|_q<\infty$, where $\xi\in\mathbf{R}$ corresponds to $(\xi_t)_{t\in\mathbf{N}}$; (ii) the τ-mixing coefficients are $\tau_{k-1}\leq ck^{-a},\forall k\geq 1$ for some universal constants $c>0$ and $a>(q(r-2)+1)/(q-r)$. Then for every $r\in[2,q)$*

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

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

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

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

Then by Hölder's inequality

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

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

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

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

$$\int_0^{\|\xi\|_1}|H(u)Q\circ G(u)|^{r-1}\mathrm{d}u\leq C_{q,r}^{\frac{q-1}{q-r}}\|\xi\|_q^{q(r-1)/(q-1)}. \tag{A4.1}$$

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

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

where the second line follows by equation ((A4.1)) and $c_{q,r} > 0$ depends only on q and r. □

A4.2 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 A4.1. *Let $\{\xi_{N,T,i,t} : i \in \mathbf{N}, t \in \mathbf{Z}\}$ be an array of centered random vectors in $\mathbf{R}^p$ such that for each N, T, and i, $\{\xi_{N,T,i,t} : t \in \mathbf{Z}\}$ is a stationary process in $\mathbf{R}^p$ and $\{(\xi_{N,T,i,1}, \ldots, \xi_{N,T,i,T}) : i \in \mathbf{N}\}$ are independent arrays in $\mathbf{R}^p \times \mathbf{R}^T$ satisfying (i) for some $q > 2$, $\max_{i\in[N],j\in[p]} \|\xi_{N,T,i,t,j}\|_q = O(1)$; (ii) for all N, T, i, j, the τ-mixing coefficients of $\{\xi_{N,T,i,t,j} : t \in \mathbf{Z}\}$ satisfy $\tau_{k-1} \leq ck^{-a}, \forall k \geq 1$ for some universal constants $c > 0$ and $a > (q-1)/(q-2) \vee (q\delta+1)/(q-2-\delta)$ with $q > 2+\delta$ and $\delta > 0$; (iii) for every $i, N \in \mathbf{N}$, $\lim_{T\to\infty} \mathrm{Var}(\xi_{N,T,i,t}) < \infty$. Then*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and whence under (ii)

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

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

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

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

By Theorem A4.2, for some $c_{q,\delta}$ that depends only on q and δ,

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

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

A4.3 Proofs

Proof of Theorem 4.1. 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 $\rho - \hat{\rho}$

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

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

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

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

Lemma A.2.1

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

where the last line follows from Theorem A4.1 with probability at least $1-\delta$ and Assumption 3.2.3 for some $c>1$. Therefore,

$$\|\mathbf{Z}\Delta\|_{NT}^2 - \lambda\{\Omega(\rho)-\Omega(\hat{\rho})\} \leq \frac{\lambda}{c}\Omega(\Delta) + \|\mathbf{m}-\mathbf{Z}\rho\|_{NT}\|\mathbf{Z}\Delta\|_{NT} \text{ with } \Delta = \hat{\rho}-\rho. \tag{A4.2}$$

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

$$\begin{aligned}\Omega(\rho) - \Omega(\hat{\rho}) &= \Omega_0(\rho) - \Omega_0(\hat{\rho}) - \Omega_1(\hat{\rho}) \\ &\leq \Omega_0(\hat{\rho}-\rho) - \Omega_1(\hat{\rho}-\rho) = \Omega_0(\Delta) - \Omega_1(\Delta).\end{aligned} \tag{A4.3}$$

Suppose that $\|\mathbf{m}-\mathbf{Z}\rho\|_{NT} \leq \frac{1}{2}\|\mathbf{Z}\Delta\|_{NT}$. Then it follows from equations (A4.2) and (A4.3) that

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

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

$$\Omega_1(\Delta) \leq \frac{c+1}{c-1}\Omega_0(\Delta). \tag{A4.4}$$

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,

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

where we set $G^* = \max_{G\in\mathcal{G}} \sqrt{|G|}$ and use Hölder's inequality, inequalities in equations ((A4.2)) and ((A4.4)), Assumption 3.2.2, $\hat{\Sigma} = \mathbf{Z}^\top\mathbf{Z}/NT$, and Babii, Ghysels, and Striaukas (2021b), Lemma A.2.1. This shows that with probability at least $1-\delta$

$$\Omega(\Delta) \leq \frac{4c^2 s}{(c-1)^2\gamma_{\min}} \left\{ 2(1+c^{-1})\lambda + \Omega(\Delta) G^* |\mathrm{vech}(\hat{\Sigma} - \Sigma_{N,T})|_\infty \right\}. \tag{A4.5}$$

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 3.2.1 by Theorem A4.1

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

for some $c > 0$. On the event E, the inequality in equation (A4.5) implies $\Omega(\Delta) \lesssim s\lambda$, and whence from the equation (A4.2) 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

$$\Omega_1(\Delta) \leq 2\frac{c+1}{c-1}\Omega_0(\Delta). \tag{A4.6}$$

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

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

or simply

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

where we use the first statement of the theorem. On the other hand, if the inequality in equation (A4.6) does not hold, then the inequality in equation (A4.4) 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 (A4.2) we obtain

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

where we use equation (A4.3). Since $\Omega(\Delta) = \Omega_1(\Delta) + \Omega_0(\Delta)$

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

where we use the fact that the inequality in equation (A4.6) does not hold. Therefore,

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

which shows that

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

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

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

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

Proof of Theorem 4.2. 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 $\rho-\hat{\rho}$

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

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

$$\begin{aligned}\|\mathbf{Z}(\hat{\rho}-\rho)\|_{NT}^2-\lambda\left\{\Omega(\beta)-\Omega(\hat{\beta})\right\}&\le\langle\mathbf{Z}^\top\mathbf{u},\hat{\rho}-\rho\rangle_{NT}+\langle\mathbf{Z}^\top(\mathbf{m}-\mathbf{Z}\rho),\hat{\rho}-\rho\rangle_{NT}\\&\le\langle B^\top\mathbf{u},\hat{\alpha}-\alpha\rangle_{NT}+\langle\mathbf{X}^\top\mathbf{u},\hat{\beta}-\beta\rangle_{NT}\\&\quad+\|\mathbf{m}-\mathbf{Z}\rho\|_{NT}\|\mathbf{Z}(\hat{\rho}-\rho)\|_{NT}\\&\le|B^\top\mathbf{u}/NT|_\infty|\hat{\alpha}-\alpha|_1+\Omega^*(\mathbf{X}^\top\mathbf{u}/NT)\Omega(\hat{\beta}-\beta)\\&\quad+\|\mathbf{m}-\mathbf{Z}\rho\|_{NT}\|\mathbf{Z}(\hat{\rho}-\rho)\|_{NT}\\&\le|B^\top\mathbf{u}/\sqrt{NT}|_\infty\vee\Omega^*(\mathbf{X}^\top\mathbf{u}/NT)\\&\quad\times\left\{|\hat{\alpha}-\alpha|_1/\sqrt{N}+\Omega(\hat{\beta}-\beta)\right\}\\&\quad+\|\mathbf{m}-\mathbf{Z}\rho\|_{NT}\|\mathbf{Z}(\hat{\rho}-\rho)\|_{NT},\end{aligned}\tag{A4.7}$$

where the second line follows by the dual norm inequality and the Cauchy-Schwartz inequality, and Ω^* is the dual norm of Ω. By Babii, Ghysels, and Striaukas (2021b), Lemma A.2.1. and Theorem A4.1 under Assumption 3.2.1, with probability at least $1-\delta/2$

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

Similarly, under Assumption 3.2.1 by Babii, Ghysels, and Striaukas (2021a), Theorem 3.1 with probability at least $1-\delta/2$

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

Therefore, under Assumption 4.3.5 with probability at least $1-\delta$

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

In conjunction with the inequality in equation ((A4.7)), this gives

$$\begin{aligned} \|\mathbf{Z}\Delta\|_{NT}^2 &\leq c^{-1}\lambda\left\{|\hat{\alpha}-\alpha|_1/\sqrt{N}+\Omega(\hat{\beta}-\beta)\right\} \\ &\quad + \|\mathbf{m}-\mathbf{Z}\rho\|_{NT}\|\mathbf{Z}\Delta\|_{NT} + \lambda\left\{\Omega(\beta)-\Omega(\hat{\beta})\right\} \\ &\leq (c^{-1}+1)\lambda\left\{|\hat{\alpha}-\alpha|_1/\sqrt{N}\Omega(\hat{\beta}-\beta)\right\} \\ &\quad + \|\mathbf{m}-\mathbf{Z}\rho\|_{NT}\|\mathbf{Z}\Delta\|_{NT} \end{aligned} \tag{A4.8}$$

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

$$\begin{aligned} \Omega(\beta)-\Omega(\hat{\beta}) &= \Omega_0(\beta)-\Omega_0(\hat{\beta})-\Omega_1(\hat{\beta}) \\ &\leq \Omega_0(\hat{\beta}-\beta)-\Omega_1(\hat{\beta}-\beta). \end{aligned} \tag{A4.9}$$

Suppose that $\|\mathbf{m}-\mathbf{Z}\rho\|_{NT} \leq \frac{1}{2}\|\mathbf{Z}\Delta\|_{NT}$. Then from the first inequality in equation ((A4.8)) and equation ((A4.3)), we obtain

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

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

$$\Omega_1(\hat{\beta}-\beta) \leq \frac{c+1}{c-1}\Omega_0(\hat{\beta}-\beta)+(c-1)^{-1}|\hat{\alpha}-\alpha|_1/\sqrt{N}. \tag{A4.10}$$

Put $\Delta_N = ((\hat{\alpha}-\alpha)^\top/\sqrt{N}, (\hat{\beta}-\beta)^\top)^\top$. Then under Assumption 3.2.2

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

Consider the following event $E = \{|\mathrm{vech}(\hat{\Sigma}-\Sigma)|_\infty < 1/(2s \vee N)\}$. Under Assumption 3.2.1 by Theorem A4.1 and Babii, Ghysels, and Striaukas (2021a), Theorem 3.1

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

Therefore, on the event E

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

and whence from equation ((A4.8)) 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 **??**. By Fermat's rule, the pooled sg-LASSO estimator in equation **??** 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 $\sqrt{NT}$

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

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

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

We will show that by Theorem A4.1, $I_{N,T} \xrightarrow{d} N(0,\Xi_G)$ as $N,T\to\infty$. To that end, by Minkowski's inequality under Assumptions 3.2.1 (i) and 3.2.5 (ii)

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

Lastly, under Assumption 3.2.5 (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 3.2.1 (ii), this verifies conditions of Theorem A4.1 and shows that $I_{N,T} \xrightarrow{d} N(0, \Xi_G)$.

Next,

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

where we use Proposition A4.3.1 and Theorem A4.1. Similarly by Proposition A4.3.1 and Corollary 4.3.1

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

Lastly, by the Cauchy-Schwartz 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 3.2.5 (v), and the last by Proposition A4.3.1 and Theorem A4.1 under maintained assumptions. □

Proposition A4.3.1. *Suppose that Assumptions 3.2.1, 3.2.2, 3.2.3, 3.2.4, and 3.2.5 are satisfied for each nodewise regression $j \in G$. Then if $S^\kappa p (NT)^{1-\kappa} \to 0$ and $S^2 \log p / NT \to 0$*

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

and

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

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

A4.4 Additional empirical results – nowcasting application

Table A4.1: Prediction results – The table reports average over firms MSEs of out-of-sample predictions. The nowcasting horizon is the next quarter, i.e. we predict the next quarter P/E ratio. Block in Panel A1-D1 correspond to ML-only forecast errors while in Panel A2-D2 to ML models augmented with median consensus nowcasts. Each Panel A1-D1 and A2-D2 block represents different ways of calculating the tuning parameter λ. Bold entries are the best results in a block.

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

Table A4.2: Prediction results – The table reports average over firms MSEs of out-of-sample predictions. The nowcasting horizon is the current month, i.e. we predict the P/E ratio using information up to the end of current fiscal quarter. Each Panel A-D block represents different ways of calculating the tuning parameter λ. Bold entries are the best results in a block. We report the best elastic net MSEs over LASSO/ridge weight $[0, 0.2, 0.4, 0.6, 0.8, 1]$: elnet-U method is where high-frequency lags are unrestricted, elnet method is where we use only the first high-frequency lag for each covariate. We also report the best sg-LASSO specification for each tuning parameter method and each model specification, see Table 4.3.

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

Table A4.3: Prediction results – The table reports average over firms MSEs of out-of-sample predictions for the same models as in Table 4.3 - discarding the first 8 quarters to compute for forecast combination weights - with additional result of prediction errors using forecast combination approach of Ball and Ghysels (2018), denoted as *F.Comb*. Hence the out-of-sample quarters start at 2009 Q1. The nowcasting horizon is the current month, i.e. we predict the P/E ratio using information up to the end of current fiscal quarter. Each Panel A-D block represents different ways of calculating the tuning parameter λ. Bold entries are the best results in a block.

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

Table A4.4: Heaviness of tails – The table reports the tail index of in-sample residuals. Tail index is computed using the Hill estimator. The results are reported for the models as in Table 4.4.

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

A4.5 Additional empirical results – Granger causality application

Table A4.5: Significance testing results – We report p-values for the AR(1) and for the sg-LASSO-MIDAS models, displaying series that are significant at 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 vs. 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 \ 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
Firm-level 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

A4.6 Data description

A4.6.1 Firm-level data

The full list of firm-level data is provided in Table A4.6. We also add two daily firm-specific stock market predictor variables: stock returns and a realized variance measure, which is defined as the rolling sample variance over the previous 60 days (i.e. 60-day historical volatility).

A4.6.1.1 Firm sample selection

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

A4.6.1.2 Firm-specific text data

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

We construct our news-based firm-level covariates by filtering only highly relevant news stories. More precisely, for each firm and each day, we filter out news that has the *Relevance Score* (REL) larger or equal to 75, as is suggested by the RavenPack News Analytics guide and used by practitioners, see for example Kolanovic and Krishnamachari (2017). REL is a score between 0 and 100 which indicates how strongly a news story is linked with a particular firm. A score of zero means that the entity is vaguely mentioned in the news story, while 100 means the opposite. A score of 75 is regarded as a significantly relevant news story. After applying the REL

filter, we apply a novelty of the news filter by using the *Event Novelty Score* (ENS); we keep data entries that have a score of 100. Like REL, ENS is a score between 0 and 100. It indicates the novelty of a news story within a 24-hour time window. A score of 100 means that a news story was not already covered by earlier announced news, while subsequently published news story score on a related event is discounted, and therefore its scores are less than 100. Therefore, with this filter, we consider only novel news stories. We focus on *five sentiment indices* that are available in both DJ and PR editions. They are:

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

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

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

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

News Impact Projections (NIP) represents the degree of impact a news flash has on the market over the following two-hour period. The algorithm

produces scores to accurately predict a relative volatility - defined as scaled volatility by the average of volatilities of large-cap firms used in the test set - of each stock price measured within two hours following the news. Tick data is used to train the algorithm and produce scores, which take values between 0 and 100, 50 representing zero impact news.

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

Table A4.6: Firm-level data description table – The *id* column gives mnemonics according to data source, which is given in the second column *Source*. The column *frequency* states the sampling frequency of the variable. The column *T-code* denotes the data transformation applied to a time-series, which are: (1) not transformed, (2) $100[(x_t/x_{t-1})^4 - 1]$, (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 - daily firm-level stock market data and C1 - 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 - monthly news attention series. In the models we include 365 daily lags, 12 monthly lags and 4 quarterly lags respectively. Series with (N) are not being used in nowcasting application.

	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 (N)	monthly	ALFRED	1
11	Real GDP (N)	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 EFFR	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

BIBLIOGRAPHY

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

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

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

Andersen, T. G., T. Bollerslev, F. X. Diebold, and C. Vega (2003): "Micro effects of macro announcements: Real-time price discovery in foreign exchange," *American Economic Review*, 93(1), 38–62.

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

Andrews, D. W. (1984): "Non-strong mixing autoregressive processes," *Journal of Applied Probability*, 21(4), 930–934.

——— (1991): "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica*, 59(3), 817–858.

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

Aprigliano, V., G. Ardizzi, and L. Monteforte (2019): "Using payment system data to forecast economic activity," *International Journal of Central Banking*, 15(4), 55–80.

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

Athey, S., and G. W. Imbens (2019): "Machine learning methods that economists should know about," *Annual Review of Economics*, 11, 685–725.

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

Babii, A., X. Chen, and E. Ghysels (2019): "Commercial and residential mortgage defaults: Spatial dependence with frailty," *Journal of Econometrics*, 212(1), 47–77.

Babii, A., and J.-P. Florens (2020): "Is completeness necessary? Estimation in nonidentified linear models," arXiv preprint arXiv:1709.03473v3.

Babii, A., E. Ghysels, and J. Striaukas (2020a): "High-dimensional Granger causality tests with an application to VIX and news," arXiv preprint arXiv:1912.06307v2.

——— (2020b): "Machine learning time series regressions with an application to nowcasting," arXiv preprint arXiv:2005.

——— (2021a): "High-dimensional Granger causality tests with an application to VIX and new," *arXiv preprint arXiv:1912.06307*.

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

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

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

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

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

Bańbura, M., D. Giannone, M. Modugno, and L. Reichlin (2013): "Now-casting and the real-time data flow," in *Handbook of Economic Forecasting, Volume 2 Part A*, ed. by G. Elliott, and A. Timmermann, pp. 195–237. Elsevier.

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

Bartlett, M. (1948): "Smoothing periodograms from time-series with continuous spectra," *Nature*, 161(4096), 686–687.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 80(6), 2369–2429.

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

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2020): "High-dimensional econometrics and generalized GMM," *Handbook of Econometrics (forthcoming)*.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on treatment effects after selection among high-dimensional controls," *Review of Economic Studies*, 81(2), 608–650.

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

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous analysis of LASSO and Dantzig selector," *Annals of Statistics*, 37(4), 1705–1732.

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

Bok, B., D. Caratelli, D. Giannone, A. M. Sbordone, and A. Tambalotti (2018): "Macroeconomic nowcasting and forecasting with big data," *Annual Review of Economics*, 10, 615–643.

Bollen, N. P., M. J. O'Neill, and R. E. Whaley (2017): "Tail wags dog: Intraday price discovery in VIX markets," *Journal of Futures Markets*, 37(5), 431–451.

Bosq, D. (1993): "Bernstein-type large deviations inequalities for partial sums of strong mixing processes," *Statistics*, 24(1), 59–70.

Box, G. E., and G. C. Tiao (1977): "A canonical analysis of multiple time series," *Biometrika*, 64(2), 355–365.

Brockwell, P. J., and R. Davis (1991): *Time series: theory and methods 2nd ed.* Springer-Verlag New York.

Bybee, L., B. T. Kelly, A. Manela, and D. Xiu (2020): "The structure of economic news," *National Bureau of Economic Research*, and http://structureofnews.com.

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

Carlsen, M., and P. E. Storgaard (2010): "Dankort payments as a timely indicator of retail sales in Denmark," Danmarks Nationalbank Working Papers.

Carrasco, M., and X. Chen (2002): "Mixing and moment properties of various GARCH and stochastic volatility models," *Econometric Theory*, pp. 17–39.

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

Carrasco, M., and B. Rossi (2016): "In-sample inference and forecasting in misspecified factor models," *Journal of Business & Economic Statistics*, 34(3), 313–338.

Chan, J. C., and I. Jeliazkov (2009): "Efficient simulation and integrated likelihood estimation in state space models," *International Journal of Mathematical Modelling and Numerical Optimisation*, 1(1-2), 101–120.

Chernozhukov, V., D. Chetverikov, K. Kato, et al. (2013): "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors," *Annals of Statistics*, 41(6), 2786–2819.

Chernozhukov, V., W. K. Härdle, C. Huang, and W. Wang (2021): "LASSO-driven inference in time and space," *Annals of Statistics*, 49(3), 1702–1735.

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

Chetverikov, D., Z. Liao, and V. Chernozhukov (2021): "On cross-validated lasso in high dimensions," *Annals of Statistics*, 49(3), 1300–1317.

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

Chiang, H. D., and Y. Sasaki (2019): "Lasso under multi-way clustering: estimation and post-selection inference," *arXiv preprint arXiv:1905.02107*.

Daniell, P. (1946): "Discussion of paper by M.S. Bartlett," *Journal of the Royal Statistical Society Supplements*, 8, 88–90.

Davydov, Y. A. (1973): "Mixing conditions for Markov chains (in Russian)," *Teoriya Veroyatnostei i ee Primeneniya*, 18(2), 321–338.

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

Dedecker, J., P. Doukhan, G. Lang, L. R. J. Rafael, S. Louhichi, and C. Prieur (2007): "Weak dependence," in *Weak dependence: With examples and applications*, pp. 9–20. Springer.

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

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

Delle Monache, D., and I. Petrella (2019): "Efficient matrix approach for classical inference in state space models," *Economics Letters*, 181, 22–27.

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

Duarte, C., P. M. Rodrigues, and A. Rua (2017): "A mixed frequency approach to the forecasting of private consumption with ATM/POS data," *International Journal of Forecasting*, 33(1), 61–75.

Eicker, F. (1963): "Asymptotic normality and consistency of the least squares estimators for families of linear regressions," *Annals of Mathematical Statistics*, pp. 447–456.

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

Feng, G., S. Giglio, and D. Xiu (2020): "Taming the factor zoo: A test of new factors," *Journal of Finance*, 75(3), 1327–1370.

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

Fisher, I. (1925): "Our unstable dollar and the so-called business cycle," *Journal of the American Statistical Association*, 20(150), 179–202.

——— (1937): "Note on a short-cut method for calculating distributed lags," *Bulletin de l'Institut international de Statistique*, 29(3), 323–328.

Florens, J.-P., and M. Mouchart (1982): "A note on noncausality," *Econometrica*, 50(3), 583–591.

Foroni, C., P. Guérin, and M. Marcellino (2018): "Using low frequency information for predicting high frequency variables," *International Journal of Forecasting*, 34(4), 774–787.

Foroni, C., M. Marcellino, and C. Schumacher (2015a): "Unrestricted mixed data sampling (U-MIDAS): MIDAS regressions with unrestricted lag polynomials," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1), 57–82.

——— (2015b): "Unrestricted mixed data sampling (U-MIDAS): MIDAS regressions with unrestricted lag polynomials," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1), 57–82.

Francq, C., and J.-M. Zakoian (2019): *GARCH models: structure, statistical inference and financial applications*. John Wiley & Sons.

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

Galbraith, J. W., and G. Tkacz (2018): "Nowcasting with payments system data," *International Journal of Forecasting*, 34(2), 366–376.

Gallant, A. R. (1987): *Nonlinear statistical models*. John Wiley, New York.

Gentzkow, M., B. Kelly, and M. Taddy (2019): "Text as data," *Journal of Economic Literature*, 57(3), 535–74.

Ghysels, E. (2016): "Macroeconomics and the reality of mixed frequency data," *Journal of Econometrics*, 193(2), 294–314.

Ghysels, E., J. B. Hill, and K. Motegi (2020): "Testing a large set of zero restrictions in regression models, with an application to mixed frequency Granger causality," *Journal of Econometrics*, 218(2), 633–654.

Ghysels, E., C. Horan, and E. Moench (2018): "Forecasting through the Rearview Mirror: Data Revisions and Bond Return Predictability.," *Review of Financial Studies*, 31(2), 678–714.

Ghysels, E., V. Kvedaras, and V. Zemlys-Balevičius (2020): "Mixed data sampling (MIDAS) regression models," in *Handbook of Statistics*, vol. 42, pp. 117–153. Elsevier.

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

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

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

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

Giannone, D., M. Lenza, and G. E. Primiceri (2018): "Economic predictions with big data: The illusion of sparsity," Staff Reports 847, Federal Reserve Bank of New York.

Granger, C. W. (1969): "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, 37, 424–438.

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

Han, Y., and R. S. Tsay (2017): "High-dimensional linear regression for dependent observations with application to nowcasting," *arXiv preprint arXiv:1706.07899*.

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

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

Hastie, T., R. Tibshirani, and R. Tibshirani (2019): "Best Subset, Forward Stepwise, or Lasso? Analysis and Recommendations Based on Extensive Comparisons," Stat Sci in press.

Hecq, A., L. Margaritella, and S. Smeekes (2019): "Granger causality testing in high-dimensional VARs: a post-double-selection procedure," arXiv preprint arXiv:1902.10991.

Huber, P. J. (1967): "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 221–233. University of California Press.

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

Jiang, W. (2009): "On Uniform Deviations of General Empirical Risks with Unboundedness, Dependence, and High Dimensionality.," *Journal of Machine Learning Research*, 10(4).

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

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

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

Kock, A. B., and L. Callot (2015): "Oracle inequalities for high dimensional vector autoregressions," *Journal of Econometrics*, 186(2), 325–344.

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

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

Koop, G. M. (2013): "Forecasting with medium and large Bayesian VARs," *Journal of Applied Econometrics*, 28(2), 177–203.

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

Lazarus, E., D. J. Lewis, J. H. Stock, and M. W. Watson (2018): "HAR Inference: Recommendations for Practice," *Journal of Business & Economic Statistics*, 36(4), 541–559.

Leeb, H., and B. M. Pötscher (2005): "Model selection and inference: Facts and fiction," *Econometric Theory*, pp. 21–59.

Li, J., and Z. Liao (2020): "Uniform nonparametric inference for time series," *Journal of Econometrics*, 219(1), 38–51.

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

Marsilli, C. (2014a): "Variable selection in predictive MIDAS models," Working papers 520, Banque de France.

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

McCracken, M. W., and S. Ng (2016): "FRED-MD: A monthly database for macroeconomic research," *Journal of Business & Economic Statistics*, 34(4), 574–589.

Medeiros, M. C., and E. F. Mendes (2016): "ℓ_1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors," *Journal of Econometrics*, 191(1), 255–271.

——— (2017): "Adaptive LASSO estimation for ARDL models with GARCH innovations," *Econometric Reviews*, 36(6-9), 622–637.

Meinshausen, N., and P. Bühlmann (2006): "High-dimensional graphs and variable selection with the LASSO," *Annals of Statistics*, 34(3), 1436–1462.

Mogliani, M., and A. Simoni (2021): "Bayesian MIDAS penalized regressions: estimation, selection, and prediction," *Journal of Econometrics*, 222(1), 833–860.

Moriwaki, D. (2019): "Nowcasting Unemployment Rates with Smartphone GPS Data," in *International Workshop on Multiple-Aspect Analysis of Semantic Trajectories*, pp. 21–33. Springer.

Nagaev, S. V. (1998): "Some refinements of probabilistic and moment inequalities," *Theory of Probability and Its Applications*, 42(4), 707–713.

Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu (2012): "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, 27(4), 538–557.

Neumann, M. H. (2013): "A central limit theorem for triangular arrays of weakly dependent random variables, with applications in statistics," *ESAIM: Probability and Statistics*, 17, 120–134.

Newey, W. K., and K. D. West (1987): "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, 55(3), 703–708.

Nicholson, W. B., I. Wilms, J. Bien, and D. S. Matteson (2020): "High dimensional forecasting via interpretable vector autoregression," *Journal of Machine Learning Research*, 21(166), 1–52.

Parzen, E. (1957): "On consistent estimates of the spectrum of a stationary time series," *Annals of Mathematical Statistics*, 28(2), 329–348.

Peña, D., and G. E. Box (1987): "Identifying a simplifying structure in time series," *Journal of the American statistical Association*, 82(399), 836–843.

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

Politis, D. N. (2011): "Higher-order accurate, positive semidefinite estimation of large-sample covariance and spectral density matrices," *Econometric Theory*, 27(4), 703–744.

Quaedvlieg, R. (2019): "Multi-horizon forecast comparison," *Journal of Business & Economic Statistics*, pp. 1–14.

Raju, S., and M. Balakrishnan (2019): "Nowcasting economic activity in India using payment systems data," *Journal of Payments Strategy and Systems*, 13(1), 72–81.

Shiller, R. J. (1973): "A distributed lag estimator derived from smoothness priors," *Econometrica*, 41, 775–788.

Shu, J., and J. E. Zhang (2012): "Causality in the VIX futures market," *Journal of Futures Markets*, 32(1), 24–46.

Siliverstovs, B. (2017): "Short-term forecasting with mixed-frequency data: a MIDASSO approach," *Applied Economics*, 49(13), 1326–1343.

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

Sims, C. A. (1971): "Discrete approximations to continuous time distributed lags in econometrics," *Econometrica*, 39, 545–563.

Skripnikov, A., and G. Michailidis (2019): "Joint estimation of multiple network Granger causal models," *Econometrics and Statistics*, 10, 120–133.

Stock, J. H., and M. W. Watson (2002): "Forecasting using principal components from a large number of predictors," *Journal of the American Statistical Association*, 97(460), 1167–1179.

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

Sun, Y., P. C. Phillips, and S. Jin (2008): "Optimal bandwidth selection in heteroskedasticity–autocorrelation robust testing," *Econometrica*, 76(1), 175–194.

Thorsrud, L. A. (2020): "Words are the new numbers: A newsy coincident index of the business cycle," *Journal of Business & Economic Statistics*, 38(2), 393–409.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B (Methodological)*, 58, 267–288.

Uematsu, Y., and S. Tanaka (2019): "High-dimensional macroeconomic forecasting and variable selection via penalized regression," *Econometrics Journal*, 22, 34–56.

van de Geer, S. (2016): *Estimation and testing under sparsity: École d'Été de Probabilités de Saint-Flour XLV–2015*, vol. 2159. Springer.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014): "On asymptotically optimal confidence regions and tests for high-dimensional models," *Annals of Statistics*, 42(3), 1166–1202.

White, H. (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, pp. 817–838.

Wilms, I., S. Gelper, and C. Croux (2016): "The predictive power of the business and bank sentiment of firms: A high-dimensional Granger Causality approach," *European Journal of Operational Research*, 254(1), 138–147.

Wong, K. C., Z. Li, and A. Tewari (2020): "Lasso guarantees for β-mixing heavy-tailed time series," *Annals of Statistics*, 48(2), 1124–1142.

Wu, W. B. (2005): "Nonlinear system theory: Another look at dependence," *Proceedings of the National Academy of Sciences*, 102(40), 14150–14154.

Wu, W.-B., and Y. N. Wu (2016): "Performance bounds for parameter estimates of high-dimensional linear models with correlated errors," *Electronic Journal of Statistics*, 10(1), 352–379.

Yuan, M., and Y. Lin (2006): "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49–67.

Zou, H., and T. Hastie (2005): "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, 67(2), 301–320.

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