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

Andrii Babii[†] Eric Ghysels[‡] Jonas Striaukas[§]

January 26, 2022

Abstract

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.

Keywords: Granger causality, high-dimensional time series, fat tails, inference, HAC estimator, sparse-group LASSO, Fuk-Nagaev inequality.

*We thank Markus Pelger, our discussant at the SoFiE Machine Learning Virtual Conference, as well as participants at the Financial Econometrics Conference at the TSE Toulouse, the JRC Big Data and Forecasting Conference, the Big Data and Machine Learning in Econometrics, Finance, and Statistics Conference at the University of Chicago, the Nontraditional Data, Machine Learning, and Natural Language Processing in Macroeconomics Conference at the Board of Governors, the AI Innovations Forum organized by SAS and the Kenan Institute, the 12th World Congress of the Econometric Society, the 2021 Econometric Society Winter Meetings and seminar participants at Vanderbilt University and University of Connecticut for helpful comments.

[†]Department of Economics, University of North Carolina–Chapel Hill - Gardner Hall, CB 3305 Chapel Hill, NC 27599-3305. Email: babii.andrii@gmail.com

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

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

1 Introduction

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 over 800,000 daily *Wall Street Journal* news articles during 1984-2017 that have been shown by Babii, Ghysels, and Striaukas (2021) 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. This paper aims 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 work to the existing literature on Granger causality with high-dimensional data. Various dimensionality reduction schemes 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 (2021), 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 the post-double selection approach

[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.

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 (2021), 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 (2021).

The proper inference for time-series data relies on the heteroskedasticity and autocorrelation consistent (HAC) estimation of the long-run variance; see Eicker (1963), Huber (1967), White (1980), Gallant (1987), Newey and West (1987), and 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

[2]The distributed lag literature can be traced back to Fisher (1925); see also Almon (1965), Sims (1971), and Shiller (1973), as well as the 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).

[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).

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 considered in Granger (1969). 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 model and, at the same time, it contains the class of α-mixing processes as a special case.

[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.

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 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, we establish a suitable version of the Fuk-Nagaev inequality for τ-mixing processes. We report on a Monte Carlo study in section 4 which provides further insights about the validity of our theoretical analysis in finite sample settings typically encountered in empirical applications. Section 5 covers an empirical application examining the Granger causal relations between the VIX and financial news. Conclusions appear in section 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 a positive integer p, we put $[p] = \{1, 2, \ldots, p\}$. For $G \subset [p]$ and $\Delta \in \mathbf{R}^p$, we also use $\Delta_G = \{\Delta_j : j \in G\}$ to denote the sub-vector of Δ indexed by G. 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 $q \geq 1$, the 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 $p\times p$ matrices A and B, let $\langle A, B\rangle = \mathrm{trace}(A^\top B)$ be the inner product and let $\|A\| = \sqrt{\langle ., .\rangle}$ be the corresponding Frobenius norm. We will also use the following matrix norm $\|A\|_\infty = \max_{j\in[p]} |A_j|_1$. For $a, b \in \mathbf{R}$, we put $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Lastly, we write $a_n \lesssim b_n$ if there exists a (sufficiently large) absolute constant C such that $a_n \leq C b_n$ for all $n \geq 1$ and $a_n \sim b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

2 HAC-based inference for sg-LASSO

This section covers the HAC-based inference for the sparse-group LASSO (sg-LASSO) estimator. In the first subsection, we introduce the high-dimensional linear projection model and the sg-LASSO estimator. Next, we introduce the τ-mixing and other relevant assumptions. The following three subsections cover the debiased central limit theorem and the HAC estimator based on the sg-LASSO residuals. Lastly, we discuss the practical implementation of the high-dimensional Wald-type test for Granger causality.

2.1 High-dimensional time series regressions and sg-LASSO

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 $\{x_{t,j} : j \geq 1\}$ may include the intercept, time-varying 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 (2021).

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 (2021), 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

$$\min_{b \in \mathbf{R}^p} \|\mathbf{y} - \mathbf{X}b\|_T^2 + 2\lambda\Omega(b), \tag{1}$$

with the penalty function

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

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.

2.2 τ-mixing and other assumptions

We measure the persistence of a 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 $\mathrm{Lip}_1 = \{f : \mathbf{R}^l \to \mathbf{R} : \ |f(x) - f(y)| \leq |x - y|_1\}$ 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

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

τ-mixing coefficient is defined as

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

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 A.1.1 for the comparison of this coefficient to the mixingale and the α-mixing coefficients as well as Dedecker and Prieur (2004, 2005) and Dedecker, Doukhan, Lang, Rafael, Louhichi, and Prieur (2007) for more details. The following assumptions impose tail and moment conditions.

Assumption 2.1 (Data)**.** *The process $(u_t,x_t)_{t\in\mathbf{Z}}$ is stationary 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 $(u_tx_{t,j})_{t\in\mathbf{Z}}$ 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 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 (2021) for verification of these conditions for a heavy-tailed autoregressive model with covariates. Next, we require that the covariance matrix of covariates is invertible.

Assumption 2.2 (Covariance)**.** *There exists a universal constant $\gamma>0$ such that the smallest eigenvalue of $\Sigma=\mathbb{E}[x_tx_t^\top]$ is bounded away from zero by γ.*

Assumption 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 equation (7).

Assumption 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 2.1.

In practice we recommend selecting 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 $\sqrt{s_\alpha} = \alpha\sqrt{|S_0|} + (1-\alpha)\sqrt{|\mathcal{G}_0|}$, and the sample size T.

Assumption 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 and $\mu = ((b+1)r-2)/(r+2(b-1))$.*

The effective sparsity constant $\sqrt{s_\alpha}$ 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 OA.1 in the supplementary material.

2.3 Nodewise LASSO

The LASSO estimator $\hat{\beta}$ is biased and may have a complicated distribution. Nonetheless, the distribution of subvectors of debiased LASSO estimator $\hat{\beta} + B$ of fixed size can be approximated with the Gaussian distribution. The bias-correction term $B = \hat{\Theta}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\beta})/T$ depends on the estimator of the precision matrix $\Theta = \Sigma^{-1}$, where $\Sigma = \mathbb{E}[XX^\top]$ is a covariance matrix of $X \in \mathbf{R}^p$. In this section, we discuss the nodewise LASSO estimator of Θ following Meinshausen and Bühlmann (2006) and van de Geer, Bühlmann, Ritov, and Dezeure (2014). 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}

$$X_j = X_{-j}^\top \gamma_j + v_j, \qquad \mathbb{E}[X_{-j}v_j] = 0, \tag{2}$$

and $\sigma_j^2 = \Sigma_{j,j} - \Sigma_{j,-j}\gamma_j = \mathbb{E}[v_j^2]$ is the variance of the projection error.[7] 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 $\hat{\Theta} = \hat{B}^{-1}\hat{C}$ with

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

[7] 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.

Let $v_{t,j}$ be the projection error in j^{th} nodewise LASSO regression. 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 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 variances of* $(u_t^2)_{t\in\mathbf{Z}}$ *and* $(v_{t,j}^2)_{t\in\mathbf{Z}}$ *exist 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 an 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 the nonparametric literature.

It is also worth mentioning that Assumption 2.5 does not require that β or Θ are sparse, i.e., we can have $s = O(p)$, in which case a larger T may be required. It is also known that one can use the alternative weak sparsity conditions defined as a small value of the ℓ_q norm of β with $q \in (0, 1)$; see van de Geer (2016), Section 2.10. Lastly, for simplicity of presentation we state all conditions in terms of the largest sparsity constant $s = s_\alpha \vee \max_{j\in G} S_j$. One can easily relax this at the cost of heavier notation.

2.4 Debiased CLT

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 2.1. *Suppose that Assumptions 2.1, 2.2, 2.3, 2.4, and 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 $\Xi_G = \lim_{T\to\infty} \operatorname{Var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T u_t \Theta_G X_t\right)$.

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

Theorem 2.1 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, and Kato (2013) for a comprehensive review of related coupling results in the i.i.d. case.

Remark 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 Theorem 2.1 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.*

2.5 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

$$\hat{\Xi}_G \triangleq \sum_{|k|<T} K\left(\frac{k}{M_T}\right) \hat{\Gamma}_k, \tag{3}$$

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 $K(0) = 1$ 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

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

It is worth recalling that the Parzen and the Quadratic spectral kernels are high-order kernels that are 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 2.1 simplifies to

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

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 2.2. *Suppose that Assumptions 2.1, 2.2, 2.3, 2.4, 2.5 are satisfied for the sg-LASSO regression and for each nodewise LASSO regression $j \in G$. Suppose also that Assumptions A.1.1 in the Appendix, and OA.1.1 in the Online 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*

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

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 2.2 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. 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. We deduce from Theorem 2.2 that the HAC estimator is consistent whenever $M_T\left(\frac{sp^{1/\kappa}}{T^{1-1/\kappa}} \vee s\sqrt{\frac{\log p}{T}} + \frac{s^2p^{2/\kappa}}{T^{2-3/\kappa}} + \frac{s^3p^{5/\kappa}}{T^{4-5/\kappa}}\right) \to 0$ as $M_T \to \infty$ and $T \to \infty$.

2.6 High-dimensional Granger causality tests

We would like to test whether a series $(w_t)_{t\in\mathbf{Z}}$ Granger causes another series $(y_t)_{t\in\mathbf{Z}}$ at some horizon h. Consider the following projection equation

$$y_{t+h} = c + \sum_{j\geq 1} z_{t,j}\gamma_j + \mathbf{w}_{t-1}^\top \alpha + u_t, \tag{4}$$

where $(z_{t,j})_{j\geq 1}$ is a high-dimensional set of controls representing "all the information in the universe available at time t"; see Granger (1969), Section 4; $\alpha = (\alpha_1, \dots, \alpha_K)^\top$ and $\mathbf{w}_{t-1} = (w_{t-1}, \dots, w_{t-K})^\top$ are lags of $(w_t)_{t\in\mathbf{Z}}$. This high-dimensional set of controls may contain low- and/or high-frequency lags of $(y_t)_{t\in\mathbf{Z}}$ or some other series.

The null hypothesis that $(w_t)_{t\in\mathbf{Z}}$ does not Granger cause $(y_t)_{t\in\mathbf{Z}}$ and the corresponding alternative hypothesis are

$$H_0: \ \alpha = 0 \qquad \text{vs.} \qquad H_1: \ \alpha \neq 0.$$

It is worth mentioning our framework is based on the weakest notion of the Granger causality, as in Granger (1969) corresponding to the marginal improvement in time series projections due to the information contained in w_t. A stronger notion of Granger non-causality appears when projections are replaced by conditional means, so that the conditional mean y_t given w_t and all other available information does not depend on w_t. Yet, even stronger version of Granger non-causality pertains to the full conditional independence; see Florens and Mouchart (1982).

The Granger causality test is nested by our more general HAC-based inference framework. To that end, consider

$$y_{t+h} = \sum_{j\in G} \beta_j x_{t,j} + \sum_{j\in G^c} \beta_j x_{t,j} + u_t,$$

where we define $x_t^\top \beta_G = \mathbf{w}_{t-1}^\top \alpha$ and $\sum_{j\in G^c} \beta_j x_{t,j} = c + \sum_{j\geq 1} z_{t,j}\gamma_j$, and $G \subset [p]$ is a relevant group of projection coefficients.

Therefore, testing the Granger causality amounts to

$$H_0: \ \beta_G = 0 \qquad \text{vs.} \qquad H_1: \ \beta_G \neq 0.$$

Theorems 2.1 and 2.2 imply that, under H_0, the asymptotic distribution of the Wald statistics is as follows

$$W_T \triangleq T\left[(\hat{\beta}_G + B_G - \beta_G)\right]^\top \hat{\Xi}_G^{-1} \left[(\hat{\beta}_G + B_G - \beta_G)\right] \xrightarrow{d} \chi^2_{|G|}, \tag{5}$$

provided that consistency conditions for HAC estimator are satisfied. Note that $\hat{\Xi}_G$ is a positive definite matrix for commonly used kernel functions; see Newey and West (1987) and Andrews (1991). The Wald test rejects when $W_T > q_{1-\alpha}$, where $q_{1-\alpha}$ is the quantile of order $1-\alpha$ of $\chi^2_{|G|}$.

The practical implementation of the Granger causality test is as follows:

1. Use sg-LASSO to estimate $\alpha \in \mathbf{R}^K$ in equation (4).

2. Compute the HAC estimator using the sg-LASSO residuals as in equation (2).

3. Compute the bias-corrected Wald statistics for $H_0: \alpha = 0$ in equation (5). Note that the bias correction term requires fitting K nodewise LASSO regressions described in section 2.3.

4. Reject the Granger non-causality if $W_T > q_{1-\alpha}$ and do not reject otherwise.

To tune the sg-LASSO estimator and the nodewise LASSO regressions we consider the following time series cross-validation. For some $l \in \mathbf{N}$ and each $t = 1, \ldots, T$

1. If $t > l+1$ and $t < T-l$, use observations $I_{t,l} = \{1, \dots, t-l-1, t+l+1, \dots, T\}$ to estimate the sg-LASSO parameter, denoted $\hat{\beta}_{-t,l}(\lambda, \alpha)$. For $t = 1, \dots, l+1$, use $I_{t,l} = \{t+l+1, \dots, T\}$ as the training sample. Similarly, for $t = T-l, \dots, T$, use $I_{t,l} = \{1, \dots, T-l-1\}$ as the training sample.

2. Use the left-out observations to test the model

$$CV(\lambda, \alpha) = \frac{1}{T} \sum_{t=1}^{T} (Y_t - X_t^\top \hat{\beta}_{-t,l}(\lambda, \alpha))^2$$

3. Minimize $CV(\lambda, \alpha)$ with respect to (λ, α).

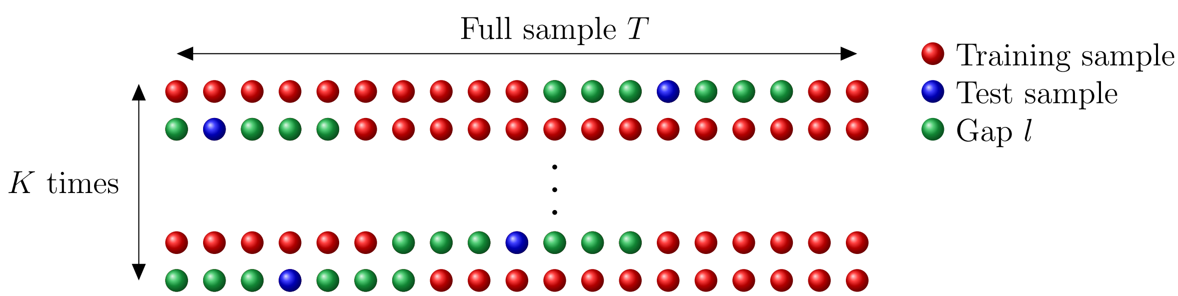

Figure 1: Time series cross-validation scheme with $l = 3$

Note that for $l = 0$ the procedure coincides with the usual leave-one-out cross-validation. For $l \geq 1$, there is a gap of l observations separating the test and the training samples, which aims to reduce the dependence between the two. To reduce the computational burden, in the spirit of the K-fold cross-validation, we can draw randomly a sub-sample $I \subset [T]$ of size K and minimize

$$CV_K(\lambda, \alpha) = \frac{1}{K} \sum_{t \in I} (Y_t - X_t^\top \hat{\beta}_{-t}(\lambda, \alpha))^2$$

instead. Figure 1 illustrates this procedure with gap $l = 3$ observations. We note that the training observations (red) are separated from a single test observation (blue) by a gap of 3 left-out observations (green) on each side. It is worth mentioning that similar procedures have been used to select the bandwidth parameter for nonparametric regression estimators with dependent data; see Chu and Marron (1991) and references therein. A variation of the "sample splitting with a gap" procedure for Tikhonov regularization is also considered in Babii (2021).

Lastly, for the HAC estimator, we propose the following "rule of thumb" to select the bandwidth parameter

$$M_T = \begin{cases} 1.3\left(\frac{T}{\log p}\right)^{\frac{1}{1+\varsigma}}, & \text{sub-Gaussian data} \\ 1.3\left(\frac{T^{2-2/q}}{p^{2/q}}\right)^{\frac{1}{1+\varsigma}}, & \text{heavy-tailed data,} \end{cases} \tag{6}$$

where ς is the order of the kernel ($\varsigma = 1$ for the Bartlett kernel and $\varsigma = 2$ for the Quadratic spectral and Parzen kernels), $q > 2$ is the number of finite moments in the data, T is the sample size, and p is the number of regressors. Note that for the Bartlett kernel in the sub-Gaussian case, we have $M_T = 1.3(T/\log p)^{1/2}$, which is similar to the "rule of thumb" $M_T = 1.3T^{1/2}$ proposed in Lazarus, Lewis, Stock, and Watson (2018). We also recommend using the quadratic spectral kernel which has certain optimality properties for inference; see Sun and Yang (2020).

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 $\tau_k^{(j)} \leq ck^{-a}$ 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[8] $\kappa = ((a+1)q-1)/(a+q-1)$, $B_T^2 = \max_{j\in[p]}\sum_{t=1}^{T}\sum_{k=1}^{T}|\mathrm{Cov}(\xi_{t,j},\xi_{k,j})|$.

[8]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. B_T here denotes the variance of partial sums and not the bias correction term of the LASSO estimator.

The inequality describes the mixture of the polynomial and Gaussian tails for the maximum of high-dimensional sums. In the limiting case of i.i.d. data, as $a \to \infty$, the dependence-tails exponent $\kappa \to q$, 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 (see 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 exponential tails.

For stationary processes, by Lemma A.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 $\delta \in (0,1)$

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\xi_t\right|_\infty \leq C\left(\frac{p}{\delta T^{\kappa-1}}\right)^{1/\kappa} \vee \sqrt{\frac{\log(8p/\delta)}{T}}\right) \geq 1-\delta, \tag{7}$$

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

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 report results for three different scenarios.

For the first scenario (denoted DGP 1), we draw independent AR(1) covariates $\{x_{t,j}, j \in [p]\}$

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

and the regression error follows the AR(1) process

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

where the errors $\epsilon_{t,j}(j \in [p]), \nu_t \sim_{i.i.d.} N(0,1)$. We set the persistence parameter $\rho = 0.6$. For the second heavy-tailed data scenario (DGP 2) the errors $\epsilon, \nu \sim_{i.i.d.}$ student-$t(5)$. Lastly, in the third scenario (DGP 3) the covariates follow a multivariate vector autoregressive (VAR) process

$$X_t = \Phi X_{t-1} + \epsilon_t$$

The sample size is $T \in \{100, 500\}$ and the number of covariates is $p \in \{10, 200\}$. For DGP 1 and DGP 2, we apply the LASSO estimator to estimate coefficients $\hat{\beta}$. For DGP 3, we use the block structure of the VAR to group covariates. Throughout the experiment, we choose all tuning parameters using the time series cross-validation described in section 2.6 with $K = 20$ and $l = 5$.

We report the empirical rejection frequency (ERF) for the Granger causality test based on the HAC estimator with two different kernel functions, namely Parzen and Quadratic spectral kernels. We set the bandwidth parameter according to the rules appearing in equation (6). For DGP 1 and DGP 3, we apply the sub-Gaussian rule, while for DGP 2 we use the heavy-tailed data rule with $q = 5$.

The vector of population regression coefficients β has the first five non-zero entries and all remaining entries are zero. To assess the performance we scale the coefficient vector by multiplying it with a constant $c \in \{0, 1/2, 1\}$, i.e., the coefficient for the relevant covariates is $c \times (1, 1, 1, 1, 1)$. For $c = 0$, the ERF shows the empirical size of the test for the nominal level of 5%, while $c \in \{1/2, 1\}$ the ERF shows the empirical power of the Granger causality test. For the larger scaling constant c, the alternatives are further away from the null hypothesis and therefore the Granger causality test is expected to perform better.

We estimate the long-run covariance matrix $\hat{\Xi}$ using the LASSO or sg-LASSO residuals, denoted $\hat{u}_t$. We use nodewise LASSO regressions to estimate the precision matrix Θ for which we also use time series cross-validation to compute LASSO tuning parameter. The first step is to compute the 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. Next, we compute the high-dimensional HAC estimator using the formula in equation (3), applying the relevant bandwidth parameter based on the DGP. 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 $\hat{\mathbf{u}}^{(i)} = \mathbf{y}^{(i)} - \mathbf{X}^{(i)} \hat{\beta}^{(i)}$. Finally, we compute the empirical rejection frequency rates by checking whether or not a variable is significant. The number of Monte Carlo experiments is set to $N = 2000$.

		Parzen		Quadratic spectral	
	$c, T \backslash p$	10	200	10	200
			Panel (a) – DGP 1		
$c = 0$	$T = 100$	0.083	0.097	0.075	0.080
	$T = 500$	0.060	0.077	0.058	0.069
$c = 0.5$	$T = 100$	0.765	0.701	0.825	0.719
	$T = 500$	0.904	0.889	0.915	0.863
$c = 1$	$T = 100$	1.0000	0.953	1.0000	0.944
	$T = 500$	1.0000	1.0000	1.0000	1.0000
			Panel (b) – DGP 2		
$c = 0$	$T = 100$	0.097	0.117	0.087	0.104
	$T = 500$	0.065	0.077	0.066	0.075
$c = 0.5$	$T = 100$	0.711	0.645	0.771	0.680
	$T = 500$	0.882	0.844	0.895	0.848
$c = 1$	$T = 100$	0.999	0.932	0.999	0.932
	$T = 500$	1.000	1.000	1.000	1.000
			Panel (c) – DGP 3		
$c = 0$	$T = 100$	0.093	0.115	0.082	0.091
	$T = 500$	0.064	0.080	0.058	0.071
$c = 0.5$	$T = 100$	0.711	0.689	0.691	0.666
	$T = 500$	0.893	0.869	0.908	0.856
$c = 1$	$T = 100$	1.000	0.910	1.000	0.897
	$T = 500$	1.000	1.000	1.000	1.000

Table 1: HAC-based inference simulation results — We report results for two kernel functions (Parzen and Quadratic spectral) with the bandwidth parameters selected using the rules-of-thumb appearing in equation (6). In addition, we report results for three data generating process: Panel (a) Gaussian data, Panel (b) student-$t(5)$ data, and Panel (c) Gaussian data with VAR model for covariates.

We report the Monte Carlo simulation average ERF rates for each scaling constant $c \in \{0, 1/2, 1\}$ over the active set of regressors. Table 1 shows the results for sample sizes $T = 100$ and $T = 500$. The table has three panels, one for each DGP. We find that the empirical size approaches the nominal 5% level for all DGPs as the sample size increases. As expected, the empirical power of the Granger causality test also increases with the sample size. The best performance is for DGP 1, reported in Panel (a) of Table 1, whereas the performance deteriorates the most for the heavy-tailed DGP 2, as shown in Panel (b). Finally, the performance of the Granger causality tests for the DGP 3 looks like that of DGP 1. Hence, heavy-tailed data have more of an adverse effect on Granger causality testing than the multivariate structure of covariates. Interestingly, the performance deteriorates with the number of noisy regressors across all scenarios. The largest decrease in this case is for the DGP 2. Overall though, the simulation results corroborate our theoretical results.

5 Testing Granger causality for VIX and financial news

The CBOE Volatility Index, known as the VIX, is a popular measure of market-based expectations 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, computed using a combination of derivatives pricing theory and 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 only on Granger causal patterns, several studies pertain 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 pricediscovery between the VIX futures and the spot VIX index finding evidence of a bi-directional causal pattern.

We study the causal relationship between financial news and the VIX. There is also a substantial literature on the impact of news releases on financial markets (e.g., Andersen, Bollerslev, Diebold, and Vega (2003)). Traditionally, such analysis 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.[9] A table with the full list of news topics series appears in the Online Appendix Table OA.1. We estimate the following time series regression model

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

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 take only one lag for the news attention series to simplify the model (which is 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 and the daily VIX lags to have a unit standard deviation.

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 (2021), 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 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. As a robustness check, we also compute results including twelve lags of each news series. In this case, we apply Legendre polynomials of degree 3 and group all lags pertaining to each news series.

[9]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.

5.1 Main results

We report the p-values for series that appear to be significant at least at the 5% significance level for a range of bandwidth values and for two kernel functions, namely Parzen and Quadratic Spectral. For the bandwidth parameter we follow the rules-of-thumb appearing in equation (6), with M_T^{sG} for sub-Gaussian data and M_T^{ht} for heavy-tailed data with $q = 5$. The sample starts in January 1990 and ends in June 2017, determined by the availability of the textual analysis data. We also report results when we truncate the sample at December 2007. The sg-LASSO tuning parameters are selected via the time series cross-validation; see Section 4. Similarly, we tune the nodewise LASSO regressions for the precision matrix estimation using the same method. The results appear in Table 2. The table contains three panels with results for the structured sg-LASSO estimator using the full data sample in Panel (A), the sample prior to the global financial crisis in Panel (B) and a different lag structure for the news series in Panel (C).

5.1.1 Granger causality of news topics

The lagged daily VIX and the *Financial crisis* news series are highly significant at the 1% significance level according to the results in Table 2 Panel (A) for the full sample and Panel (B) for the pre-Great Depression sample, irrespective of the choice of kernel function. In the model with twelve lags for each news series, the covariate *Financial crisis* remains significant at the 5% significance level (Table 2 Panel (C)). The *Recession* news series appears to be significant in Panel (A)), but it is not significant if we exclude the data starting from January 2008 onward (see Panel (B)) or if we use the model with twelve lags for each news series reported in Panel (C). Therefore, the significance of the *Recession* news series seems to be affected by the sample choice or model specification, while the *Financial crisis* news series remains significant irrespective of these choices.

Variable $\backslash M_T$	M_T^{sG}	M_T^{ht}	M_T^{sG}	M_T^{ht}
	Parzen		QS	
	Panel (A) – Full sample			
Daily VIX lags	0.000	0.001	0.000	0.003
Financial crisis	0.001	0.001	0.001	0.001
Recession	0.001	0.001	0.001	0.002
	Panel (B) – up to 2008			
Daily VIX lags	0.000	0.001	0.000	0.012
Financial crisis	0.010	0.005	0.008	0.003
Recession	0.147	0.122	0.145	0.075
	Panel (C) – 12 lags			
Daily VIX lags	0.008	0.027	0.013	0.067
Financial crisis	0.011	0.037	0.033	0.023
Recession	0.255	0.434	0.367	0.518

Table 2: VIX Granger causality results. We report p-values of series that are significant at least at the 5% significance level with two kernel functions, Parzen and Quadratic spectral (QS). Following the rules-of-thumb appearing in equation (6) we use M_T for sub-Gaussian data, denoted M_T^{sG}, and for heavy-tailed data with $q = 5$, denoted M_T^{ht}.

Variable \ M_T	M_T^{sG}	M_T^{ht}	M_T^{sG}	M_T^{ht}
	Parzen		QS	
		Full sample		
Daily VIX lags	0.108	0.105	0.091	0.154

Table 3: Bi-directional Granger causality results. We report p-values for a range of M_T values and two kernel functions, Parzen and Quadratic spectral (QS). We use M_T for sub-Gaussian data, denoted M_T^{sG}, and for heavy-tailed data with $q = 5$, denoted M_T^{ht}.

5.1.2 Bi-directional Granger causality

We also test whether the daily VIX Granger causes the *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 the *Financial crisis* news series. Note that we only need to estimate the initial coefficient vector, since the precision matrix remains the same. The results appear in Table 3. They show a rather weak predictability of future news series by daily VIX, suggesting a unidirectional Granger causality pattern.

5.1.3 Granger causal clusters of news topics

The news attention series are classified into 24 broader meta topicsthat group the individual news series according to a common theme. See again Online Appendix Table OA.1 for details. We test which group of individual news series Granger causes future VIX values. The results are reported in Table 4.They show that the group *Banks* is significant at the 1% 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.*

Variable \M_T	M_T^{sG}	M_T^{ht}	M_T^{sG}	M_T^{ht}
	Parzen		QS	
Banks	0.002	0.005	0.001	0.001

Table 4: Group Granger causality results. We report p-values for a range of M_T values and two kernel functions, Parzen and Quadratic spectral (QS). We use M_T for sub-Gaussian data, denoted M_T^{sG}, and for heavy-tailed data with $q = 5$, denoted M_T^{ht}.

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, and 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.

References

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

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 and Economic Statistics*, 31(2), 240–251.

Andrews, D. W. (1991): "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica*, 59(3), 817–858.

Athey, S., and G. W. Imbens (2019): "Machine learning methods that economists should know about," *Annual Review of Economics*, 11, 685–725.

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

Babii, A., E. Ghysels, and J. Striaukas (2021): "Machine learning time series regressions with an application to nowcasting," *Journal of Business and Economic Statistics (forthcoming)*.

Bartlett, M. (1948): "Smoothing periodograms from time-series with continuous spectra," *Nature*, 161(4096), 686–687.

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.

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.

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.

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

Chernozhukov, V., D. Chetverikov, and K. Kato (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.

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., and Y. Sasaki (2019): "Lasso under multi-way clustering: Estimation and post-selection inference," *arXiv:1905.02107*.

Chu, C.-K., and J. S. Marron (1991): "Comparison of two bandwidth selectors with dependent errors," *Annals of Statistics*, 19(4), 1906–1918.

Daniell, P. (1946): "Discussion of paper by M.S. Bartlett," *Journal of the Royal Statistical Society Supplements*, 8, 88–90.

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.

EICKER, F. (1963): "Asymptotic normality and consistency of the least squares estimators for families of linear regressions," *Annals of Mathematical Statistics*, 34(2), 447–456.

FENG, G., S. GIGLIO, AND D. XIU (2020): "Taming the factor zoo: A test of new factors," *Journal of Finance*, 75(3), 1327–1370.

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.

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.

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., P. Santa-Clara, and R. Valkanov (2006): "Predicting volatility: Getting the most out of return data sampled at different frequencies," *Journal of Econometrics*, 131(1–2), 59–95.

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

Granger, C. W. (1969): "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, 37(3), 424–438.

Hecq, A., L. Margaritella, and S. Smeekes (2019): "Granger causality testing in high-dimensional VARs: A post-double-selection procedure," *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.

Koop, G. M. (2013): "Forecasting with medium and large Bayesian VARs," *Journal of Applied Econometrics*, 28(2), 177–203.

Lazarus, E., D. J. Lewis, J. H. Stock, and M. W. Watson (2018): "HAR inference: recommendations for practice," *Journal of Business and Economic Statistics*, 36(4), 541–559.

Leeb, H., and B. M. Pötscher (2005): "Model selection and inference: Facts and fiction," *Econometric Theory*, 21(1), 21–59.

Li, J., and Z. Liao (2020): "Uniform nonparametric inference for time series," 219(1), 38–51, Journal of Econometrics.

Meinshausen, N., and P. Bühlmann (2006): "High-dimensional graphs and variable selection with the LASSO," *Annals of Statistics*, 34(3), 1436–1462.

Nagaev, S. V. (1998): "Some refinements of probabilistic and moment inequalities," *Theory of Probability and Its Applications*, 42(4), 707–713.

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.

Politis, D. N. (2011): "Higher-order accurate, positive semidefinite estimation of large-sample covariance and spectral density matrices," *Econometric Theory*, 27(4), 703–744.

Shiller, R. J. (1973): "A distributed lag estimator derived from smoothness priors," *Econometrica*, 41(4), 775–788.

Shu, J., and J. E. Zhang (2012): "Causality in the VIX futures market," *Journal of Futures Markets*, 32(1), 24–46.

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(3), 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.

Sun, Y., P. C. Phillips, and S. Jin (2008): "Optimal bandwidth selection in heteroskedasticity–autocorrelation robust testing," *Econometrica*, 76(1), 175–194.

Sun, Y., and J. Yang (2020): "Testing-optimal kernel choice in HAR inference," *Journal of Econometrics*, 219(1), 123–136.

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*, 48(4), 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.

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.

APPENDIX

A.1 Proofs

Proof of Theorem 2.1. 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 $\sqrt{T}$

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

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

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

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 2.1 and 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 2.1 and 2.5 (i)-(ii), the long-run variance

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

exists by Proposition A.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 $1 \leq t_1 < t_2 < \dots < t_h < t_h + r \leq t_h + s \leq T$

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

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 A.1.1 and Assumptions 2.1 (ii) and 2.5 (ii). Similarly,

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

Since the sequence $(r^{-a(\varsigma-2)/(\varsigma-1)\wedge d})_{r\in\mathbf{N}}$ is summable under Assumption 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 OA.1.2 and the inequality in equation (7) under Assumption 2.1, and the last under Assumption 2.5 (iv). Likewise, using $|Ax|_\infty\leq\max_{j,k}|A_{j,k}|_\infty|x|_1$, by Proposition OA.1.2 and Theorem OA.1 of the Online Appendix

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

under Assumption 2.5 (iv). Lastly, by the Cauchy-Schwartz inequality, under Assumption 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 equation (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 A.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 2.2. By Proposition OA.1.3 in the Online Appendix with $V_t = (u_t v_{t,j}/\sigma_j^2)_{j\in G}$

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

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 OA.1.1, 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 OA.1.1

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

Under Assumptions A.1.1 (i) and OA.1.1 (i) in the Online Appendix

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

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

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

By the Cauchy-Schwartz inequality, under Assumptions of Theorem OA.1 in the Online Appendix for the sg-LASSO and Assumption A.1.1 (ii)

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

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

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

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 A.1.1 (iii) and Theorem OA.1

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

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

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

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 A.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

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

For the last term, since under Assumption 2.1 (ii), $\sup_k \mathbb{E}|X_{t,k}|^{2\tilde{q}} < \infty$ and $\kappa \geq \tilde{q}$, by Theorem OA.1

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

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 $j \in G$

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

and whence

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

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

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

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

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

The result follows from combining this estimate with the estimate in equation (A.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

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

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

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

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 $(\xi_{(t-1)J+1}, \dots, \xi_{tJ})$ 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. $\square$

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 $\mathcal{M}_0 = \sigma(\xi_0, \xi_{-1}, \xi_{-2}, \dots)$. 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 A.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. $\square$

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

Lemma A.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 A.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 A.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*

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

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)

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

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

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

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

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

By Lemma A.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 A.1.1, the second by equation (A.2), and the third by Lemma A.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. □

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

Online Appendix

Andrii Babii Eric Ghysels Jonas Striaukas

Additional results and proofs: This file contains supplementary results with proofs.

OA.1 Additional theoretical results

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 (2021), Corollary 3.1.

Theorem OA.1. *Suppose that Assumptions 2.1, 2.2, 2.3, and 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 OA.1.1. *Suppose that Assumptions 2.1, 2.2, 2.3, and 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

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

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

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

whence $I_T = O_P\left(\frac{1}{\sqrt{T}}\right)$. Therefore, by the triangle inequality and Theorem OA.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 OA.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 OA.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 (2) for each $j \in [p]$. Put $S = \max_{j\in G} S_j$, where S_j is the support of γ_j.

Proposition OA.1.2. *Suppose that Assumptions 2.1, 2.2, 2.3, and 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 OA.1 and Proposition OA.1.1 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

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

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 $\Gamma = (\Gamma_k)_{k\in\mathbf{Z}}$. The following assumption describes the relevant class of distributions and kernel functions.

Assumption OA.1.1. *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 A.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 OA.1.3. *Suppose that Assumption OA.1.1 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 OA.1.1 (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

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

where the second sum is defined to be zero if $T \leq \varepsilon M_T$, the second line follows under Assumption OA.1.1 (i)-(ii) and the last two under Assumption OA.1.1 (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 OA.1.1 (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

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

Next, under Assumption OA.1.1 (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 OA.1.1 (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 OA.1.1 (i), (iv)

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

The result follows from combining estimates in equations (OA.1) and (OA.2). □

OA.2 News topics data details

	News topic	Meta topic		News topic	Meta topic
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 OA.1: News series — The column *News topic* is the news series topics, column *Meta topic* is meta topics/groups of news series.