Inference in High-Dimensional Linear Projections: Multi-Horizon Granger Causality and Network Connectedness*

[Click here for the latest version]

Eugène Dettaa[†] and Endong Wang[‡]

November 6, 2024

Abstract

This paper presents a Wald test for multi-horizon Granger causality within a high dimensional sparse Vector Autoregression (VAR) framework. The null hypothesis focuses on the causal coefficients of interest in a local projection (LP) at a given horizon. Nevertheless, the post-double-selection method on LP may not be applicable in this context, as a sparse VAR model does not necessarily imply a sparse LP for horizon $h > 1$. To validate the proposed test, we develop two types of de-biased estimators for the causal coefficients of interest, both relying on first-step machine learning estimators of the VAR slope parameters. The first estimator is derived from the Least Squares method, while the second is obtained through a two-stage approach that offers potential efficiency gains. We further derive heteroskedasticity- and autocorrelation-consistent (HAC) inference for each estimator. Additionally, we propose a robust inference method for the two-stage estimator, eliminating the need to correct for serial correlation in the projection residuals. Monte Carlo simulations show that the two-stage estimator with robust inference outperforms the Least Squares method in terms of the Wald test size, particularly for longer projection horizons. We apply our methodology to analyze the interconnectedness of policy-related economic uncertainty among a large set of countries in both the short and long run. Specifically, we construct a causal network to visualize how economic uncertainty spreads across countries over time. Our empirical findings reveal, among other insights, that in the short run (1 and 3 months), the U.S. influences China, while in the long run (9 and 12 months), China influences the U.S. Identifying these connections can help anticipate a country's potential vulnerabilities and propose proactive solutions to mitigate the transmission of economic uncertainty.

Keywords. Multi-horizon Granger causality, High dimensional VAR, Machine learning, De-biased estimator, Robust inference, Network connectedness, Policy-related economic uncertainty.

JEL Classification: C13, C26, C32, C36, C55, F44.

*We are extremely grateful to Marine Carrasco, Benoit Perron, Mathieu Marcoux, Jean-Marie Dufour, and Victoria Zinde-Walsh for their helpful discussions and guidance. We also thank René Garcia, Prosper Dovonon, and participants in the CIREQ Econometrics Conference in Honor of Eric Ghysels, the 2024 NBER-NSF Time Series Conference, and the Dagenais Econometrics Seminars for their constructive comments and discussions.

[†]Email: eugene.delacroix.dettaa.mboudjiho@umontreal.ca; Ph.D. candidate at the University of Montreal.

[‡]Email: endong.wang@mail.mcgill.ca; Ph.D. candidate at McGill University.

Contents

1. Introduction

Granger causality test is widely used in economics and finance to analyze the interconnectedness between time series in a multivariate system. The concept of Granger causality at a single horizon was initially introduced by Granger (1969) and later extended by Dufour and Renault (1998) to multiple horizons, enabling the exploration of interconnectedness between variables over extended time periods.[1] This paper develops a simple and user-friendly method for testing multi-horizon Granger causality in a high-dimensional (HD) system, where the number of time series is relatively large compared to the time series length. Several applications involving high-dimensionality are relevant in economics and finance. These include: (i) exploring spillovers and contagion among policy-related Economic Uncertainty Indices (see Baker et al., 2016) at the country level, (ii) evaluating the spillover effects of U.S. monetary policy on developing countries, and (iii) investigating volatility transmission in stock return prices.

Multi-horizon Granger causality test is typically conducted under the assumption that the underlying process follows a Vector Autoregressive (VAR) model. The null hypothesis of the test includes the parameters in a multi-horizon linear projection (LP) model, which projects future outcomes (up to a specified horizon) on current information (see Dufour et al., 2006 and Dufour and Wang, 2024). In macroeconomics, the linear projection is commonly referred to as Local Projection (Jordà, 2005), particularly when estimating impulse responses. However, high dimensionality render the standard Least Squares approach inappropriate since the covariance matrix of the explanatory variables could be singular. A widely used solution is the post-double-selection LASSO (pds-LASSO) method, which operates under the assumption of sparsity.[2] For instance, Hecq, Margaritella, and Smeekes (2023) apply pds-LASSO in the spirit of Belloni et al. (2014b), assuming sparsity in the underlying VAR process, to test (horizon one) Granger causality. However, extending this method to test multi-horizon Granger causality might not be feasible. Indeed, the LP is a nonlinear transformation of the underlying VAR process and it implies that a sparse VAR does not necessarily lead to a sparse LP for horizons $h > 1$. Directly imposing the assumption of sparsity on LP for all horizons $h > 1$ can be overly restrictive. Therefore, assuming sparsity only in the underlying VAR model is essential for testing multi-horizon Granger causality.

In this paper, we contribute to the literature by introducing two de-biased estimation methods with statistical inference for multi-horizon Granger-causal coefficients within a sparse high-dimensional VAR framework. Our approach enhances the application of multi-horizon Granger causality tests in high-dimensional datasets. Specifically, our contribution is fourfold.

First, we propose de-biased Least Squares (LS) estimators for multi-horizon Granger-causal coefficients, which are a finite subset of parameters in the Local Projection (LP)

[1] For instance, see Lütkepohl (1993), Dufour and Renault (1998), Dufour and Taamouti (2010), Diebold and Yılmaz (2014), Salamaliki and Venetis (2019), among others.

[2] Another approach to handle high dimensionality is principal component analysis (PCA), which assumes that only a few common factors drive the high-dimensional controls. Examples include factor VAR models, as discussed in Bernanke et al. (2005) and Stock and Watson (2016). In this paper, we focus on a sparse high-dimensional model without the common factor assumption.

equation. These estimators assume sparsity only in the underlying data-generating process (VAR model), rather than in the LP model itself. Our research highlights a crucial yet often overlooked fact: within a sparse VAR framework, when the projection horizon exceeds one, the LP equation may not exhibit sparsity, as LP coefficients are highly nonlinear transformations of the VAR matrix coefficients. We derive the asymptotic Gaussian distribution for the de-biased LS estimates and provide Heteroskedasticity- and Autocorrelation-Consistent (HAC) standard errors to account for serial correlation in the projection residuals.

Second, we extend the two-stage estimator for multi-horizon Granger-causal coefficients, originally proposed in Dufour and Wang (2024) for low-dimensional frameworks, to the high-dimensional VAR context. The two-stage estimators offer two primary advantages over the LS estimators: (1) they are generally more efficient when the horizon exceeds one, and (2) they could provide robust inference, eliminating the need to correct for serial correlation in the projection residuals (see Montiel Olea and Plagborg-Møller, 2021 and Dufour and Wang, 2024). We derive an asymptotic Gaussian distribution for these estimators with HAC standard error estimators under weak regularity conditions. Moreover, under additional conditions on the VAR disturbances, we propose Heteroskedasticity-Consistent (HC) standard errors. These HC standard errors eliminate the reliance on HAC estimators, addressing issues such as over-rejection of confidence intervals in small samples and challenges with bandwidth and kernel function selection (see Lazarus et al., 2018 and Lazarus et al., 2021), as well as the computational inefficiency of bootstrap methods in high-dimensional settings.

Third, we derive de-biased multi-horizon Granger-causal coefficient estimators using the de-sparsification technique proposed by van de Geer et al. (2014), as applied to structural impulse response estimates in Adamek et al. (2023). Instead of directly applying LASSO or post-double-selection LASSO to the LP, we first estimate the regularized VAR slope coefficients using methods such as LASSO and its variants (e.g., adaptive LASSO, elastic net). We then compute the multi-horizon Granger-causal coefficients using explicit formulas from two distinct estimation methods. To address the bias introduced by high-dimensional control variables, we de-bias these estimates, ensuring valid Gaussian inference in high-dimensional settings. Our de-biasing procedures can be interpreted in terms of Neyman orthogonalization (see, e.g., Chernozhukov et al., 2018), allowing to mitigate the impact of a potential regularization bias in the first-step estimation of VAR coefficients on the second-step estimators of the causal coefficients of interest.

We assess the performance of the Wald test based on both de-biased estimators and various variance estimators. We use the size of the Wald test as a measure of performance. Our results reveal that the two-stage approach with heteroskedastic-consistent (HC) standard errors outperforms the two-stage or least-squares approaches with HAC-type standard errors, particularly for large projection horizons. Indeed, as the projection horizon increases, while HC robust inference provides good size, sizes for HAC-type inference worsen. This size distortion arises because HAC-type variance estimators tend to become imprecise as the projection horizon increases due to high dimensionality. Moreover, our procedures outperform the post-double-selection procedure with HAC inference for all horizons. Additionally, we show in simulations that the size of the test converges to the nominal level for all inference procedures.

Finally, we apply the multi-horizon Granger causality test to study economic uncertainty interconnectedness among a large set of countries and construct a causal network to observe how uncertainty spreads across these countries. Our sample consists of 20 series of country-level monthly economic uncertainty indices collected from January 2003 to February 2024 (see Baker et al., 2016 for the construction of this index). Our objective is to visualize the strength of connectedness through Granger causality across multiple horizons. We implement pair-wise Granger causality tests at different horizons while controlling each time for the remaining countries in the sample. We then construct a heatmap, based on the significance levels of the test statistics, to illustrate interconnectedness in country-level economic uncertainty indices. Our empirical results show, among other insights, that in terms of economic uncertainties, the U.S. Granger-causes China in the short run (1 and 3 months), while China exerts influence over the U.S. in the long run (9 and 12 months). Our intuition for this result is that: (i) the U.S. has a dominant role in global economic policy, causing immediate spillovers to China, and trade dependency may amplify short-run transmission from the U.S. to China; (ii) China's growing influence on the global market, including raw materials and manufacturing, increasingly affects U.S. economic conditions over time, and potential long-term adjustments in trade and strategic U.S. sectors shift uncertainty from China to the U.S. in the long run. However, gaining more insights into the channels behind the interconnections we have identified requires deeper analysis of the types of transactions between countries in our sample.

Relevant Literature: Our study is related to the literature on regularized estimation in high-dimensional time series, drawing on work by Basu and Michailidis (2015), Medeiros and Mendes (2016), Wong et al. (2020), Masini et al. (2022), and Adamek et al. (2023). While these papers primarily focus on regularized estimation techniques, our research shifts the emphasis to de-biased estimation and inference for parameters in local projection (LP) equations, which are built upon these regularized estimates of VAR slope coefficients. The de-biasing technique we adopt is closely related to the debiased/desparsified methods in the literature, see Belloni et al. (2012), van de Geer et al. (2014), Chernozhukov et al. (2018), and Krampe et al. (2023), among others. To the best of our knowledge, however, we are the first to investigate multi-horizon Granger-causal coefficients within a high-dimensional VAR framework.

Our investigation into multi-horizon Granger causality in high-dimensional settings complements the growing literature on Granger causality at a single horizon in large datasets, such as the studies by Hecq et al. (2023) and Babii et al. (2024). While Adamek et al. (2024) examine debiased estimates for impulse responses in high-dimensional LP models, their focus remains on impulse responses, whereas our study specifically addresses multi-horizon Granger-causal coefficients. The distinction between Granger causality at a single horizon and at multiple horizons is conceptually grounded in the work of Dufour and Renault (1998).

Our debiased least squares (LS) estimators with HAC inference in high-dimensional LP models extend the low-dimensional estimation methods discussed by Jordà (2005) and Dufour et al. (2006). Additionally, our heteroskedasticity-robust inference for two-stage debiased estimates builds upon the literature on robust inference in LP models, including Montiel Olea and Plagborg-Møller (2021), Breitung and Brüggemann (2023), Xu and Guo (2024), and Dufour and Wang (2024). However, these studies focus exclusively on low-dimensional frameworks. To our knowledge, we are the first to propose heteroskedasticity-robust inference in a high-dimensional LP model.

This paper tackles high-dimensionality by employing the sparsity assumption and regularized estimation. An alternative common approach involves assuming common factors and applying principal component analysis (PCA), as in the classical Factor VAR (FAVAR) framework developed by Bernanke et al. (2005) and Stock and Watson (2016). Recently, Miao, Phillips, and Su (2023) incorporated latent factors into a sparse high-dimensional VAR model, though their algorithm is notably complex due to the simultaneous estimation of high-dimensional coefficient matrices and the common component matrix.

Our research on causal connectedness visualizes the significance levels of Wald test statistics for multi-horizon Granger causality. This approach relates to the work on network connectedness by Diebold and Yılmaz (2014), which accounts for connectedness using generalized variance decompositions by Koop, Pesaran, and Potter (1996) and Pesaran and Shin (1998). Multi-horizon Granger causality reveals the specific information that a given variable contributes to the forecast of a target outcome at various horizons.

Outlines: This paper is structured as follows. Section 2 outlines the econometric framework. In Section 3, we review a range of regularized estimators in high-dimensional models. Section 4 introduces a de-biased Least Squares estimation method. Section 5 presents a de-biased two-stage estimation method. We derive asymptotic Gaussian inference for both estimators, as well as robust inference for the de-biased two-stage estimators, in Section 6. The results of Monte Carlo simulations are presented in Section 7. Section 8 provides an empirical application of our methods and visualizes the connectedness of country-level economic uncertainties. Finally, Section 9 concludes the paper. Proofs of the results are collected in the Appendix.

Notations: The following notations are used throughout the paper. $C > 1$ will denote a generic constant of n that may be different in different uses. Let $r, s \in \mathbb{N}$. $\tilde{e}_{rj}$, $j = 1, \dots, r$ denote the r-dimensional unit vectors, where $\tilde{e}_{rj}$ contains 1 at the j$^{\text{th}}$ position and 0 elsewhere. For any vector $x \in \mathbb{R}^r$, $\|x\|_1 := \sum_{j=1}^{r} \left|x_j\right|$ denotes its l_1 norm, and $\|x\|_2^2 := \sum_{j=1}^{p} \left|x_j\right|^2$ is the squared l_2 norm. Furthermore, for a $r \times s$ matrix $B = \left(b_{i,j}\right)_{i=1,\dots,r, j=1,\dots,s}$, $\|B\|_1 := \max_{1\le j\le s} \sum_{i=1}^{r} \left|b_{i,j}\right| = \max_{1\le j\le s} \left\|B\tilde{e}_{sj}\right\|_1$ is the maximum absolute column sum norm, $\|B\|_\infty := \max_{1\le i\le r} \sum_{j=1}^{s} \left|b_{i,j}\right| = \max_{1\le i\le r} \left\|\tilde{e}'_{ri}B\right\|_1$ is the maximum absolute raw sum norm, and $\|B\|_{\max} := \max_{1\le i\le r, 1\le j\le s} \left|b_{i,j}\right|$ is the maximum norm. Also, denote the largest absolute eigenvalue of a square matrix B by $\rho(B)$ and let $\|B\|_2^2 := \rho\left(BB'\right)$ denote the spectral norm. The r-dimensional identity matrix is denoted by I_r and for two matrices B_1 and B_2, their Kronecker product is denoted by $B_1 \otimes B_2$. For any symmetric and positive semi-definite matrix B, $\lambda_{\min}(B)$ and $\lambda_{\max}(B)$ denote its minimum and maximum eigenvalues, respectively.

2. Framework

Consider a high-dimensional (HD) d-variate process $\left\{w_t\right\}_{t=1}^{n}$ generated by a VAR(p) process:

$$w_t = A_1 w_{t-1} + A_2 w_{t-2} + \cdots + A_p w_{t-p} + u_t, \tag{2.1}$$

where u_t is a serially uncorrelated random process with zero mean and non-singular covariance Σ_u, such that $u_t \sim (0, \Sigma_u)$, $\lambda_{\min}(\Sigma_u) > 0$. The order p is assumed to be finite, and the number of series d grows with the sample size n. To facilitate the discussion, the high-dimensional vector w_t is partitioned as $w_t = (y_t, x_t, q_t')'$, where q_t is a high-dimensional vector of control variables, and x_t and y_t are two scalar variables.

Following typical literature on time series (Lütkepohl (2005), Kilian and Lütkepohl (2017)), the VAR model can be written in a compact form

$$w_t = J\mathbf{A}W_{t-1} + u_t, \tag{2.2}$$

where J is a $d \times dp$ selection matrix, $J = [I_d, 0, \cdots, 0]$, and $W_{t-1} = [w_{t-1}', w_{t-2}', \cdots, w_{t-p}']'$, and $\mathbf{A}$ is the companion matrix,

$$\mathbf{A} = \begin{bmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I & & & & \\ & I & & & \\ & & \ddots & & \\ 0 & & & I & 0 \end{bmatrix}. \tag{2.3}$$

Granger causality is widely used in time series and economics, see Granger (1969), Geweke (1984). Moreover, multi-horizon Granger causality is an extended definition which has been used to better understand the dynamic causality for a multivariate system, see Lütkepohl (1993), Dufour and Renault (1998). Without loss of generality, we will focus on Granger causality from x_t to y_t over h periods. It is stated that x does not Granger-cause y at horizon h if the following equation holds[3]

$$\mathrm{P}_L(y_{t+h} \mid W_t) = \mathrm{P}_L(y_{t+h} \mid W_{-x,t}), \tag{2.4}$$

where $W_{-x,t}$ denotes the vector W_t excluding $(x_t, x_{t-1}, \cdots, x_{t-p+1})$.

To test the above equality, we could investigate the row equation of y_t in the linear projection model,

$$w_{t+h} = J\mathbf{A}^h W_t + u_t^{(h)} \tag{2.5}$$

where $u_t^{(h)} = \sum_{i=0}^{h-1} J\mathbf{A}^i J' u_{t+h-i}$, and in particular $J\mathbf{A}^i J'$ is the reduced-form impulse response; see Dufour and Renault (1998), Kilian and Lütkepohl (2017), Lusompa (2023), among others. Note that this equation has long been used as a general forecasting model in economics and finance. Moreover, Jordà (2005) uses it to estimate the reduced-form impulse response and obtains the structural impulse response by post-multiplying it with the structural matrix estimator of Θ_0.

We investigate the equation of y_{t+h} in the multivariate equation in (2.5) and write it in a generic way,

$$y_{t+h} = \beta_h' W_t + e_{t,h}, \tag{2.6}$$

[3]Where $P_L(X|Y)$ denotes the linear projection of X onto Y for given random vectors X and Y.

where β_h' is the row line of $J\mathbf{A}^h$ corresponding to variable y_t, and $e_{t,h}$ is the corresponding element in $u_t^{(h)}$.

Without loss of generality, we partition the set of regressors as $W_t = \{W_{1,t}, W_{2,t}\}$, where $W_{1,t} = R_1 W_t$ represents the low-dimensional vector of regressors of interest, and $W_{2,t} = R_2 W_t$ represents the high-dimensional vector of control variables, containing the remaining variables. The local projection equation (2.6) can be rewritten as

$$y_{t+h} = \beta_{1,h}' W_{1,t} + \beta_{2,h}' W_{2,t} + e_{t,h}, \tag{2.7}$$

where $\beta_{1,h}$ and $\beta_{2,h}$ are the corresponding coefficients for $W_{1,t}$ and $W_{2,t}$, respectively. In the exercise of multi-horizon Granger causality test, $W_{1,t} = (x_t, x_{t-1}, \cdots, x_{t-p+1})'$ and $W_{2,t}$ contains the lagged values of y and the control variable q.[4] Thus, the null hypothesis of Granger non-causality (2.4) is stated as

$$\mathscr{H}_0 : \beta_{1,h} = 0. \tag{2.8}$$

Testing $\mathscr{H}_0$ involves estimating and making inferences about $\beta_{1,h}$. Estimating $\beta_{1,h}$ is a challenging exercise because Equation (2.7) includes the high-dimensional nuisance parameter $\beta_{2,h}$. For example, if $d = 20$ and $p = 4$, as in our empirical application, then $\beta_{1,h}$ is a 4×1 vector, while $\beta_{2,h}$ consists of 76 nuisance parameters, which is a large vector if the sample size is around $n = 120$, as is often the case. Estimation in this high-dimensional setting is often feasible by assuming the sparsity of the underlying VAR model given by Equation (2.1), meaning that only a small number of coefficients in the VAR representation are non-zero [5]. Even under this assumption, the sparsity of the local projection equation (2.7) is not always guaranteed.

Indeed, if the causality from x to y at horizon one is of interest, i.e., $h = 1$, the post-double selection method could be employed to produce de-biased estimates, see Hecq et al. (2023). This is because (2.6) is essentially a single equation in the VAR system. Therefore, the sparsity assumption imposed on the VAR system implies that the high-dimensional coefficient $\beta_{2,h}$ is sparse for $h = 1$. However, for causality tests at horizons larger than one, $h > 1$, it might not be feasible to directly apply the post-double selection method to (2.7) to obtain de-biased estimates of $\beta_{1,h}$. This is because the sparsity assumption on VAR matrix slope coefficients does not necessarily imply the sparsity of the local projection coefficient β_h for all $h > 1$. Specifically, β_h is a highly non-linear transformation of the VAR matrix coefficient $\mathbf{A}$. We propose two approaches for estimation and inference on $\beta_{1,h}$ in the sparse high-dimensional VAR model (2.1) under a potentially non-sparse local projection equation (2.7). Since consistent estimation of the VAR matrix coefficient is a primary step in our procedures, we review methodologies for regularized estimation of $\mathbf{A}$ in the next section before presenting our methods.

[4]See multi-horizon Granger causality test in low dimensional setup in Dufour et al. (2006).

[5]This sparsity assumption is often supported by the belief that in a high-dimensional time series system, a given variable will be associated with only a small number of other variables in the system. In Section 6, we will present the form of sparsity we will rely on in the theoretical derivations. The sparsity assumption is typically incorporated into the estimation procedure via l_1-penalization methods, such as lasso and its variants (adaptive lasso, elastic net, etc.). Note that this sparsity assumption is imposed on the underlying VAR equation and not on the local projection equations.

3. Review of regularized estimation on high dimensional VAR

The curse of dimensionality in high-dimensional time series frameworks is widely recognized. For instance, in a d-variate VAR(p) model, estimating pd^2 parameters poses a formidable task, particularly as the number of parameters grows significantly with d. Even with extensive data, such as 20 years of daily observations for the S&P 100 index, the number of parameters $\left(\propto 100^2\right)$ remains large compared to the sample size (roughly 5000 observations). To keep the model complexity tractable, a sparsity assumption is often made (see Assumption 2(i)). Consistent estimation of the VAR matrix coefficients is then possible via l_1-type penalization. In this section, we briefly review the methodologies of l_1-regularized estimation of the VAR model under the sparsity assumption.

The Least Absolute Shrinkage and Selection Operator (LASSO), proposed by Tibshirani (1996), is one of the most popular l_1-regularized methods used in high-dimensional time series. Its variant, adaptive LASSO (adaLASSO), was introduced by Zou (2006) to overcome the limitations of LASSO. In fact, in addition to providing a sparse solution like LASSO, adaptive LASSO enjoys the oracle property, meaning that it has the same asymptotic distribution as OLS, conditional on knowing the regressors that should be included in the model. The adaLASSO method involves estimating $A_1, \ldots, A_p$ through row-wise regression on d equations of the VAR model:

$$\hat{A}^{(re)}_{j\bullet,1:p} = \text{argmin} \frac{1}{n-p} \sum_{t=p+1}^{n} \left\| w_t - \sum_{i=1}^{p} A_{j\bullet,i} w_{t-i} \right\|_2^2 + \lambda \sum_{i=1}^{p} \|A_{j\bullet,i}\Pi_i\|_1, \tag{3.1}$$

for $j = 1, 2, \cdots, d$, where $A_{j\bullet,1:p}$ denotes the j-th row of slope coefficient matrices $A_{1:p} = [A_1, A_2, \cdots, A_p]$ and Π_i is a diagonal matrix specifying penalty loadings $\Pi_i = \text{diag}[\pi_{ik}]_{k=1,2,\cdots,d}$. If $\pi_{ik} = 1$ for all k, (3.1) reduces to a LASSO estimation equation. For instance, Belloni et al. (2012) consider a diagonal matrix representing a data-dependent penalty loadings for the self-normalization of the first-order conditions in the Lasso problem. Practically, determining these data-dependent penalty loadings involves two steps. First, the 'first step coefficients' are obtained by applying Lasso with a specific information criterion, such as the Bayesian Information Criterion (BIC). Then, the data-dependent penalty loading for each coefficient is computed using the formula $|\text{'first step coefficients'} + (n-h)^{-1/2}|^{-\tau}$, where $\tau = 1$. This indicates that the data-dependent penalty loading is inversely related to the first step Lasso coefficient. In addition, the penalty parameter λ in adaptive Lasso is also determined through a model selection process using a specific information criterion.

Besides LASSO and adaLASSO, the elastic net (ElNet), proposed by Zou and Hastie (2005), provides a way to combine the strengths of LASSO and ridge regression. While the l_1 part of the ElNet method performs variable selection, its l_2 part stabilizes the solution. ElNet is particularly well-suited to cases where there is a strong correlation among regressors. Moreover, Hecq et al. (2023) mention the possibility of using ElNet, which allows the penalty function to be strictly convex. As a result, ElNet can select highly correlated variables as a group, while LASSO only selects one of these variables. Similar observations are supported by the simulation results in the appendix of Wilms et al. (2021).

There is a large strand of the literature on the derivation of theoretical properties of l_1-penalized least squares estimates of VAR models; see, e.g., Basu and Michailidis (2015); Davis et al. (2016); Han and Liu (2013); Song and Bickel (2011); Wu and Wu (2014), among others. For instance, Basu and Michailidis (2015) demonstrates the possibility of consistent estimation under high-dimensional scaling through l_1-regularization for a broad class of stable time series processes, subject to sparsity constraints. As the aim of this paper is not to investigate the properties of l_1-penalized estimators of the VAR matrix coefficients, we assume that we have a consistent estimator $\hat{\mathbf{A}}^{(re)}$ of the matrix $\mathbf{A}$ (in the sense of Assumption 2(iii)), regardless of whether LASSO or one of its variants (adaLASSO or ElNet) is used.

4. De-biased Least Square estimation

In this subsection, we consider the de-biased LS approach to identify and estimate the parameter of interest, $\beta_{1,h}$, in (2.6).

4.1. Lease Squares identification

Suppose weak exogeneity condition holds for u_t and the contemporaneous covariance matrix Σ_u is of full rank, then

$$\beta_h = \mathbb{E}[W_t W_t']^{-1}\mathbb{E}[W_t y_{t+h}], \tag{4.1}$$

where the covariance matrix $\mathbb{E}[W_t W_t']$ can be written as a function of VAR slope coefficient matrices and the contemporaneous covariance matrix Σ_u, as presented in Krampe et al. (2023) and Lütkepohl (2005):

$$\mathbb{E}[W_t W_t'] = \sum_{j=0}^{\infty} \mathbf{A}^j J' \Sigma_u J (\mathbf{A}')^j = \mathrm{vec}_{dp}^{-1}\Big((I_{d^2p^2} - \mathbf{A}\otimes\mathbf{A})^{-1}\mathrm{vec}(J'\Sigma_u J)\Big) \tag{4.2}$$

As shown in Section 2.1 of Lütkepohl (2005), the stability of VAR system, that is, the largest eigenvalue of matrix $\mathbf{A}$ is bounded from unit disk, ensures the invertibility of matrix $(I_{d^2p^2} - \mathbf{A}\otimes\mathbf{A})$. Moreover, the full rankness of covariance matrix Σ_u ensures that the covariance matrix $\mathbb{E}[W_t W_t']$ is positive definite. This result follows from the fact that u_t (the residual of the linear projection of w_t onto the past information set) has a non-singular covariance matrix.

By the Frisch–Waugh–Lovell theorem, the parameter of interest, $\beta_{1,h}$, can be expressed as

$$\beta_{1,h} = \mathbb{E}[W_{1,t}^{\perp} W_{1,t}^{'}]^{-1}\mathbb{E}[W_{1,t}^{\perp} y_{t+h}], \tag{4.3}$$

where

$$W_{1,t}^{\perp} := W_{1,t} - \mathrm{P}_L(W_{1,t} \mid W_{2,t}) = W_{1,t} - \mathbb{E}[W_{1,t}W_{2,t}']\Big(\mathbb{E}[W_{2,t}W_{2,t}']\Big)^{-1} W_{2,t}. \tag{4.4}$$

In the low-dimensional case, practitioners will take the LS projection residual as an estimator of $W_{1,t}^{\perp}$ in line with (4.4) and thereby conduct estimation on $\beta_{1,h}$ through sample covariance. Precisely, an estimator of $W_{1,t}^{\perp}$ is obtained in a low-dimensional setting by replacing population means in Equation (4.4) with their sample counterparts. However, in the high-dimensional setup, $W_{2,t}$ is a high-dimensional control variable, making the standard LS projection potentially infeasible. In fact, the sample counterpart of $\mathbb{E}[W_{2,t}W_{2,t}']$ can be singular in this case. The Least Squares estimation approach we propose below still uses the identification equation (4.3) but relies on an alternative way to estimate the rotated regressor $W_{1,t}^{\perp}$.

4.2. De-biased Least Squares estimator

Denote $\Sigma_W := \mathbb{E}[W_t W_t']$. Since $W_{1,t}$ and $W_{2,t}$ are sub-vector of W_t, we rewrite them in the form of $W_{1,t} = R_1 W_t$ and $W_{2,t} = R_2 W_t$, where R_1, R_2 are selection matrices, such that $R = [R_1', R_2']'$ and $RR' = I_{dp}$. Estimation of $W_{1,t}^{\perp}$ in high-dimensional setting is carried out using the following relation obtained from Equation (4.4) through block matrix inversion,

$$W_{1,t}^{\perp} = (R_1 \Sigma_W^{-1} R_1')^{-1} R_1 \Sigma_W^{-1} W_t. \tag{4.5}$$

Let $\hat{A}_{1:p}^{(re)}$ be the regularized estimators of $(A_1, \cdots, A_p)$ as defined in Equation (3.1). Compute the covariance matrix of u_t as $\hat{\Sigma}_u = \frac{1}{n-p}\sum_{t=p+1}^{n} \hat{u}_t \hat{u}_t'$ where $\hat{u}_t := w_t - \sum_{i=1}^{p} \hat{A}_i^{(re)} w_{t-i} = w_t - J\hat{\mathbf{A}} W_{t-1}$.

Step 1: We use the explicit formula (4.2) and compute $\hat{\Sigma}_W$, the estimate of Σ_W, by using $\hat{A}_{1:p}^{(re)}$ and $\hat{\Sigma}_u$.

Step 2: Following (4.5), we estimate the regressor $W_{1,t}^{\perp}$. Instead of estimating Σ_W through sample variance, we use the estimate $\hat{\Sigma}_W$ obtained from Step 1.

$$\hat{W}_{1,t}^{\perp} = (R_1 \hat{\Sigma}_W^{-1} R_1')^{-1} R_1 \hat{\Sigma}_W^{-1} W_t. \tag{4.6}$$

The equation can be readily checked by the block matrix inverse formula.

Step 3: Compute the LS estimate of $\beta_{1,h}$,

$$\hat{\beta}_{1,h}^{(LS)} = \left(\sum_t \hat{W}_{1,t}^{\perp} W_{1,t}'\right)^{-1}\left(\sum_t \hat{W}_{1,t}^{\perp} y_{t+h}\right). \tag{4.7}$$

Step 4: Compute the de-biased LS estimate of $\beta_{1,h}$,

$$\hat{\beta}_{1,h}^{(de-LS)} = \hat{\beta}_{1,h}^{(LS)} - \left(\sum_t \hat{W}_{1,t}^{\perp} W_{1,t}'\right)^{-1}\left(\sum_t \hat{W}_{1,t}^{\perp} W_{2,t}' \hat{\beta}_{2,h}\right) \tag{4.8}$$

where $\hat{\beta}_{2,h}$ is selected from $J\left(\hat{\mathbf{A}}^{(re)}\right)^h$. Eventually, we obtain the de-biased estimates $\hat{\beta}_{1,h}^{(de-LS)}$.

Remark 4.1.

(i) In the low-dimensional setting, an estimator of $W_{1,t}^{\perp}$ is obtained by replacing the population covariance and variance in Equation (4.4) with their sample counterparts. Algebraically equivalently, this involves replacing Σ_W in (4.5) with its sample counterpart. However, in the high-dimensional setting, where the sample counterpart of the high-dimensional covariance matrix Σ_W is singular, $\hat{W}_{1,t}^{\perp}$ is obtained as in (4.6), in which the sample covariance is estimated through the explicit formula presented in (4.2), with $\mathbb{A}$ replaced with its regularized estimator.

(ii) Equation (4.6) implicitly assumes that the sample covariance of Σ_W computed from (4.2) is non-singular. Implied by (4.2), the nonsingularity of the sample covariance matrix estimate entails that

$$\lambda'\hat{\Sigma}_W\lambda = \lambda'J'\hat{\Sigma}_u J\lambda + \sum_{j=1}^{\infty}\lambda'(\hat{\mathbf{A}}^{(re)})^j J'\hat{\Sigma}_u J(\hat{\mathbf{A}}^{(re)})'^j\lambda > 0, \tag{4.9}$$

for all $\|\lambda\| = 1, \lambda \in \mathbb{R}^{pd}$. It is easy to check that one sufficient condition that $\hat{\Sigma}_W$ is non-singular is the full rankness of the covariance matrix of the VAR residuals $\hat{\Sigma}_u$. Since the sample covariance $\hat{\Sigma}_u$ is computed as $\hat{\Sigma}_u = \frac{1}{n-p}\sum_{t=p+1}^{n}\hat{u}_t\hat{u}_t'$, then one necessary condition of the non-singularity of the sample covariance $\hat{\Sigma}_u$ is the dimension of the VAR is less than the sample size, $d < n$. Otherwise, $\hat{\Sigma}_u$ will be singular, and this could potentially result in the sample covariance $\hat{\Sigma}_W$ being singular, though not necessarily. This is because the full rankness of the matrix $\hat{\Sigma}_u$ is not a necessary condition for the full rankness of $\hat{\Sigma}_W$, which in turn depends on the values of the VAR companion matrix estimates $\hat{\mathbf{A}}^{(re)}$.

(iii) Notice that Step 4 is crucial to obtain de-biased estimate of $\beta_{1,h}$. It is due to the fact that

$$\begin{aligned}
\hat{\beta}_{1,h}^{(LS)} &= \left(\sum_t \hat{W}_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\sum_t \hat{W}_{1,t}^{\perp}y_{t+h}\right) \\
&= \left(\sum_t \hat{W}_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\sum_t \hat{W}_{1,t}^{\perp}(\beta_{1,h}'W_{1,t} + \beta_{2,h}'W_{2,t} + e_{t,h})\right) \\
&= \beta_{1,h} + \left(\sum_t \hat{W}_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\sum_t \hat{W}_{1,t}^{\perp}e_{t,h}\right) + \left(\sum_t \hat{W}_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\sum_t \hat{W}_{1,t}^{\perp}W_{2,t}'\beta_{2,h}\right).
\end{aligned}$$

The bias term emerges due to the high dimensionality. In standard time series literature, according to the Frisch-Waugh-Lovell (FWL) theorem, $\hat{W}_{1,t}^{\perp}$ is the residual of $W_{1,t}$ after partialling out the control variable $W_{2,t}$. This leads to the term $\sum_t \hat{W}_{1,t}^{\perp}W_{2,t}'$ being equal to zero because the projection residual is orthogonal to the projection space. However, in a high-dimensional setup, $\hat{W}_{1,t}^{\perp}$ is obtained through an explicit formula rather than the projection residual. This induces the high-dimensional bias if $\beta_{2,h} \neq 0$.

(iv) Our de-biased estimator $\hat{\beta}_{1,h}^{(de-LS)}$ can be seen as a Neyman orthogonalized version of $\hat{\beta}_{1,h}^{(LS)}$ (see Chernozhukov et al., 2018 for the definition of Neyman orthogonality).

This interpretation further justifies the importance of our de-biasing procedure in mitigating bias in $\hat{\beta}_{1,h}^{(LS)}$ due to potential contamination by regularization bias from the first-step machine learning estimation of the VAR matrix coefficient $\mathbf{A}$.

To clarify, first note that $\hat{\beta}_{1,h}^{(LS)}$ is the solution to the sample counterpart of the moment condition

$$\mathbb{E}\left[\varphi_t^{ls}\left(\beta_{1,h},\eta_0\right)\right]=0, \tag{4.10}$$

where

$$\varphi_t^{ls}\left(\beta_{1,h},\eta\right)=\left(W_{1,t}-\delta W_{2,t}\right)\left(y_{t+h}-W_{1,t}'\beta_{1,h}\right),$$

$\delta_0 := \mathbb{E}[W_{1,t}W_{2,t}']\left(\mathbb{E}[W_{2,t}W_{2,t}']\right)^{-1}$ is such that $W_{1,t}^{\perp}=W_{1,t}-\delta_0 W_{2,t}$, and $\eta_0=\text{vec}\left(\delta_0\right)$ is a high-dimensional $(d-1)p^2\times 1$ vector of nuisance parameters. The estimator of δ_0 is obtained by replacing population means with $\hat{\mathbb{E}}[W_{1,t}W_{2,t}']$ and $\hat{\mathbb{E}}[W_{2,t}W_{2,t}']$, which are sub-matrices of $\hat{\Sigma}_W$. This estimator can be seen as a machine learning estimator of δ_0 as $\hat{\Sigma}_W$ involves the regularized estimator $\hat{\mathbf{A}}^{(re)}$ of $\mathbf{A}$. However, the score function φ_t^{ls} is not Neyman orthogonal with respect to the high-dimensional nuisance parameter η. This implies that a noisy estimation of η_0 will introduce bias in $\hat{\beta}_{1,h}^{(LS)}$ (see, Chernozhukov et al., 2018).

In contrast, $\hat{\beta}_{1,h}^{(de-LS)}$ is the solution to the sample counterpart of the moment condition

$$\mathbb{E}\left[\psi_t^{dls}\left(\beta_{1,h},\eta_0\right)\right]=0, \tag{4.11}$$

where

$$\psi_t^{dls}\left(\beta_{1,h},\eta\right)=\left(W_{1,t}-\delta W_{2,t}\right)\left(y_{t+h}-W_{1,t}'\beta_{1,h}-W_{2,t}'\beta_{2,h}\right),$$

$\beta_{2,h}=R_2\left(\mathbf{A}^h\right)'J'e_y$, and $\eta_0=\left(\beta_{2,h}',\text{vec}\left(\delta_0\right)'\right)'$ is a high-dimensional $(d-1)(p+p^2)\times 1$ vector of nuisance parameters[6]. The score function ψ_t^{dls} is Neyman orthogonal with respect to the high-dimensional nuisance parameter η. As an implication, $\hat{\beta}_{1,h}^{(de-LS)}$ is less sensitive to noisy estimation of η_0.

(v) In our simulation, we have several findings about the high-dimensional bias: (1) the distribution of student-t test statistics of the debiased estimates has a well-shaped density similar to the standard Gaussian distribution; see Figure 1. (2) The distribution of student-t test statistics of the non-debiased estimates deviates noticeably from the Gaussian distribution. (3) The effect of the bias on the empirical level of Student's t-test statistics is most pronounced at shorter horizons and diminishes as the horizon lengthens; see Figure 2. This is because the value of $\beta_{2,h}$ declines exponentially to zero as the projection horizon increases under stationarity (the absolute value of the maximum eigenvalue of the VAR companion matrix is bounded by unity).

[6] e_y is a d-dimensional unit vector with 1 in the position of y_t in the vector w_t and 0 elsewhere.

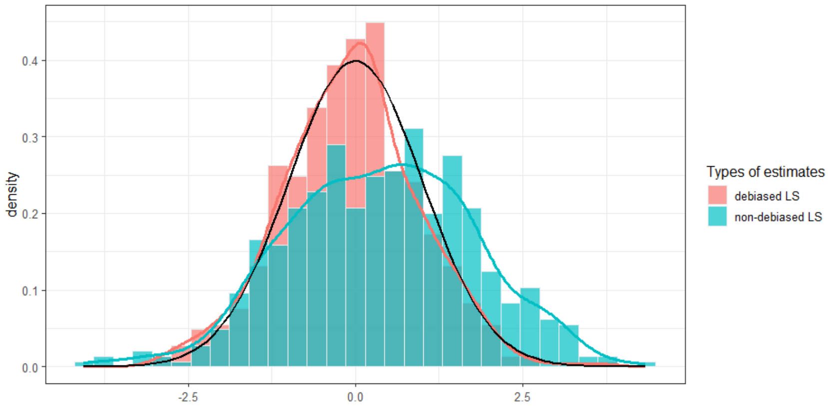

Fig. 1: The black curve represents the density of the standard Gaussian distribution, the red curve depicts the fitted density of the debiased least squares (LS) estimate, and the green curve shows the fitted density of the non-debiased LS estimate. The coefficient of interest is the coefficient of x_t on y_{t+1}. The sample size is $n = 240$, the dimension of the VAR is $d = 120$, and the number of simulations is 500. The data generating process (DGP) is a VAR(2). The values of the VAR coefficients and the covariance matrix are determined in the same manner as in Figure 3.

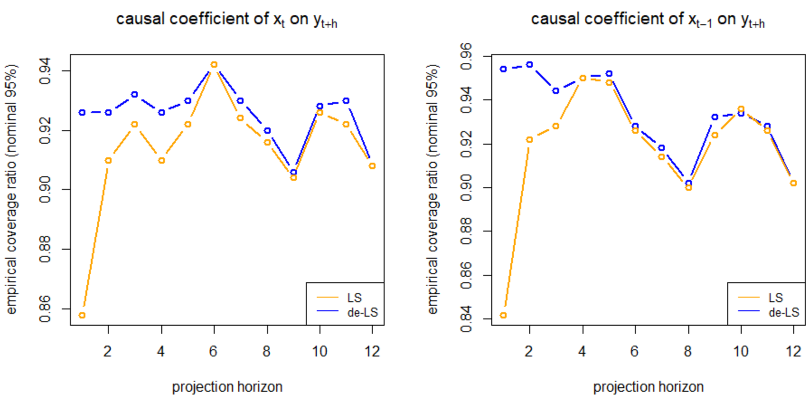

Fig. 2: Empirical coverage ratio for debiased and non-debiased LS for the coefficient of x_t to y_{t+1}. The sample size is $n = 120$, the dimension of the VAR is $d = 60$, and the number of simulations is 500. The data generating process (DGP) is a VAR(2). The values of the VAR coefficients and the covariance matrix are determined in the same manner as in Figure 3.

4.3. Asymptotic variance of de-biased LS estimators

To conduct a statistical test, it is crucial to derive the asymptotic variance and provide a consistent estimator. Obtaining the asymptotic variance requires disentangling the main term from the negligible term in the $\sqrt{n}$-normalized estimation error, $\sqrt{n}\left(\hat{\beta}_{1,h}^{(de-LS)} - \beta_{1,h}\right)$. Under Assumption 2 and certain restrictions on the growth rate of the number of series d with respect to the sample size n (see Condition 6.1 below), we show (see Lemma A.4 of the Appendix) that

$$\sqrt{n}\left(\hat{\beta}_{1,h}^{(de-LS)} - \beta_{1,h}\right) = \left(E\left[W_{1,t}^{\perp} W_{1,t}'\right]\right)^{-1}\left(n^{-1/2}\sum_{t=p}^{n-h} W_{1,t}^{\perp} e_{t,h}\right) + o_p(1), \tag{4.12}$$

so that the asymptotic variance of the de-biased LS estimator is given by[7]

$$\begin{aligned}\mathrm{AVar}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-LS)}\right) &= \lim_{n\to\infty}\left(E\left[W_{1,t}^{\perp}W_{1,t}'\right]\right)^{-1}\mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}W_{1,t}^{\perp}e_{t,h}\right)\left(E\left[W_{1,t}W_{1,t}^{\perp\prime}\right]\right)^{-1} \\ &= \lim_{n\to\infty}(R_1\Sigma_W^{-1}R_1')\Omega_{W_1,h}(R_1\Sigma_W^{-1}R_1'),\end{aligned} \tag{4.13}$$

where $\Omega_{W_1,h}$ is the long run variance of the regression score function,

$$\Omega_{W_1,h} := \lim_{n\to\infty}\mathrm{Var}\left(n^{-1/2}\sum_{t=p}^{n-h}W_{1,t}^{\perp}e_{t,h}\right) = \lim_{n\to\infty}\sum_{k=-\infty}^{\infty}\mathbb{E}[W_{1,t}^{\perp}W_{1,t+k}^{\perp\prime}e_{t,h}e_{t+k,h}].$$

Analogous to conventional time series literature, one possible consistent estimator of the asymptotic variance $\mathrm{AVar}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-LS)}\right)$ could be obtained by replacing each term within the expression on the right-hand side of the second equality of Equation (4.13) with its sample counterpart. Note that the long-run variance matrix $\Omega_{W_1,h}$ estimate may not be positive semidefinite if it is computed simply by summing up all lead-lag autocovariances at some truncated bandwidth. This motivates the use of HAC-type covariance matrix. Therefore, researchers could, for instance, choose Newey-West estimates for the sample regression score function $\hat{W}_{1,t}^{\perp}\hat{e}_{t,h}$, where $\hat{W}_{1,t}^{\perp}$ is given by Equation (4.6) and $\hat{e}_{t,h} = y_{t+h} - \hat{\beta}_h W_t$ and $\hat{\beta}_h$ is selected from $(\hat{\mathbf{A}}^{(re)})^h$. A consistent estimator of the asymptotic variance of the de-LS is then given by

$$\widehat{\mathrm{AVar}}^{(hac)}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-LS)}\right) = (R_1\hat{\Sigma}_W^{-1}R_1')\hat{\Omega}_{W_1,h}^{(hac)}(R_1\hat{\Sigma}_W^{-1}R_1'), \tag{4.14}$$

where $\hat{\Omega}_{W_1,h}^{(hac)}$ is some consistent HAC estimator of $\Omega_{W_1,h}$.

5. De-biased two-stage estimation

This subsection proposes a two-stage approach to identify β_h. The introduction of this second approach is motivated by potential drawbacks of the de-biased least squares estimation with HAC-type inference. In fact, as we will argue in the simulation section, the de-biased LS with HAC-type inference exhibits size distortion, especially for longer horizons. This is due to poor estimation of lead-lag autocovariances that appear in the long-run variance. Here, we extend the two-stage approach originally proposed by Dufour and Wang (2024) from a low-dimensional to a high-dimensional setting. This approach provides two potential gains. First, it offers a potential efficiency gain as it can be viewed as an instrumental variable approach. Second, it eliminates the need to correct for serial correlation in the variance estimation at the cost of a certain restriction on the VAR innovations, this restriction being satisfied for a wide range of innovation processes. We start by presenting the

[7] Note that Σ_W is a $dp \times dp$ matrix. Since we are in a high-dimensional setting, d is allowed to go to infinity in our asymptotic regime and therefore implicitly depends on the sample size n. These arguments justify the presence of the limit sign on the right-hand side of the second equality in Equation (4.13).

two-stage identification strategy.

5.1. Two-stage identification

If weak exogeneity condition holds for u_t and the covariance matrix Σ_u is of full rank, then

$$\mathrm{P}_L\left(y_{t+h} - \beta_h' W_t | U_t\right) = 0 \tag{5.1}$$

where $U_t = (u_t', u_{t-1}', \cdots, u_{t-p+1}')'$. Here, the variable U_t serves as instrumental variables to W_t. It yields an alternative moment-based identification method for the projection coefficients,

$$\beta_h = \mathbb{E}[U_t W_t']^{-1}\mathbb{E}[U_t y_{t+h}] \tag{5.2}$$

where $\mathbb{E}[U_t W_t']$ and $\mathbb{E}[U_t y_{t+h}]$ have explicit form containing the covariance matrix of the innovation process and the reduced-form impulse response functions,

$$\mathbb{E}[U_t W_t'] = (I_p \otimes \Sigma_u)\Psi(p), \tag{5.3}$$
$$\mathbb{E}[U_t y_{t+h}] = (I_p \otimes \Sigma_u)[\Psi_h', \Psi_{h+1}', \cdots, \Psi_{h+p-1}']' \nu_1, \tag{5.4}$$

and $\Psi(p)$ is a $p \times p$ block matrix whose ij-th block is a $d \times d$ matrix of Ψ_{i-j}' for $i \geq j$ and zero otherwise; and $\Psi_h = J\mathbf{A}^h J'$. It is easy to check that $\mathbb{E}[U_t W_t']$ is of full rank as long as Σ_u is non-singular. Analogous to the IV identification in static model, the full rankness of Σ_u implies there exists no under-identification, that is, the number of valid instruments is identical to the number of variables.

The parameter of interest $\beta_{1,h}$ is identified by applying the Frisch–Waugh–Lovell theorem,

$$\beta_{1,h} = \mathbb{E}[U_{1,t}^{\perp} W_{1,t}^{'}]^{-1}\mathbb{E}[U_{1,t}^{\perp} y_{t+h}], \tag{5.5}$$

where $U_{1,t}^{\perp} := U_{1,t} - \Gamma U_{2,t}$, $U_{1,t}$ is the residual corresponding to regressor of interest $W_{1,t}$, $U_{1,t} = R_1 U_t$, and $U_{2,t}$ is the residual corresponding to HD control variable $W_{2,t}$, $U_{2,t} = R_2 U_t$. Notice that $U_{1,t}^{\perp}$ is a linear transformation of $U_{1,t}$, $U_{2,t}$, such that the appropriation of being a valid instrument for $W_{1,t}$ implies that $U_{1,t}^{\perp}$ being orthogonal to the control variable $W_{2,t}$. Therefore, the parameter Γ is identified through the second moment,

$$\Gamma = \mathbb{E}[U_{1,t} W_{2,t}']\mathbb{E}[U_{2,t} W_{2,t}']^{-1} \tag{5.6}$$

which is derived from the moment condition $\mathrm{P}_L\left(U_{1,t} - \Gamma U_{2,t} \mid W_{2,t}\right) = 0$. For matrix algebra, we denote $\bar{\Gamma}_R = [I, -\Gamma][R_1', R_2']'$ and thereby we could rewrite $U_{1,t}^{\perp}$ as a rotated U_t,[8]

$$U_{1,t}^{\perp} = [I, -\Gamma][U_{1,t}', U_{2,t}']' = \bar{\Gamma}_R U_t = (R_1 \Sigma_{UW}^{-1} R_1')^{-1} R_1 \Sigma_{UW}^{-1} U_t. \tag{5.7}$$

5.2. De-biased two-stage estimator

We provide a de-biased two-stage estimator. Denote $\Sigma_{UW} := \mathbb{E}[U_t W_t']$.

Step 1: We use the explicit formula (5.3) to obtain an estimator of the matrix Σ_{UW}, denoted as $\hat{\Sigma}_{UW}$. The matrix Σ_{UW} consists of Ψ_h whose estimates are obtained through regularized slope coefficient estimates $\hat{A}_{1:p}^{(re)}$, such that $\hat{\Psi}_h = J\left(\hat{\mathbf{A}}^{(re)}\right)^h J'$.

Step 2: Estimate $U_{1,t}^{\perp}$ through (5.7):

$$\hat{U}_{1,t}^{\perp} = (R_1 \hat{\Sigma}_{UW}^{-1} R_1')^{-1} R_1 \hat{\Sigma}_{UW}^{-1} \hat{U}_t \tag{5.8}$$

where $\hat{\Sigma}_{UW}$ is from Step 1, and $\hat{U}_t = (\hat{u}_t', \hat{u}_{t-1}', \cdots, \hat{u}_{t-p+1}')'$, and $\hat{u}_t = w_t - \sum_i \hat{A}_i^{(re)} w_{t-i}$.

Step 3: Compute the two-stage estimate of $\beta_{1,h}$,

$$\hat{\beta}_{1,h}^{(2S)} = \left(\sum_t \hat{U}_{1,t}^{\perp} W_{1,t}'\right)^{-1} \left(\sum_t \hat{U}_{1,t}^{\perp} y_{t+h}\right). \tag{5.9}$$

Step 4: Compute the de-biased two-stage estimate of $\beta_{1,h}$,

$$\hat{\beta}_{1,h}^{(de-2S)} = \hat{\beta}_{1,h}^{(2S)} - \left(\sum_t \hat{U}_{1,t}^{\perp} W_{1,t}'\right)^{-1} \left(\sum_t \hat{U}_{1,t}^{\perp} W_{2,t}' \hat{\beta}_{2,h}\right) \tag{5.10}$$

where $\hat{\beta}_{2,h}$ is selected from $J\left(\hat{\mathbf{A}}^{(re)}\right)^h$.

Remark 5.1.

(i) The covariance matrix Σ_{UW} in finite samples can be readily computed through the sample covariance of $\hat{U}_t$ and W_t, where the $\hat{U}_t$ could be the stacked Least Squares VAR residuals. However, as illustrated in Dufour and Wang (2024), it is still recommended to estimate Σ_{UW} using the explicit formula. It is because the matrix Σ_{UW} has

[8] Here is the matrix algebra to support (5.7):

$$\begin{aligned}
U_{1,t}^{\perp} &= [I, -\Gamma] R U_t \\
&= \left([I,0](R\Sigma_{UW}R')^{-1}[I,0]'\right)^{-1} \left([I,0](R\Sigma_{UW}R')^{-1}\right) R U_t \\
&\quad \left(\text{Use block matrix inverse formula: } [I,-\Gamma] = \left([I,0](R\Sigma_{UW}R')^{-1}[I,0]'\right)^{-1} \left([I,0](R\Sigma_{UW}R')^{-1}\right)\right) \\
&= \left(R_1 \Sigma_{UW}^{-1} R_1'\right)^{-1} \left(R_1 \Sigma_{UW}^{-1} U_t\right). \\
&\quad \left(\text{Since } RR' = I, \text{ then } [I,0](R')^{-1} = R_1.\right)
\end{aligned}$$

a specific structure: it is a lower triangular matrix and, more precisely, a block Toeplitz matrix, meaning each diagonal from top-left to bottom-right (main and others) contains identical blocks. Although the sample covariance of $\hat{U}_t$ and W_t can produce consistent results, it often results in upper triangular part of the matrix being non-zero, and the block matrices are not identical on each diagonal.

(ii) Note that Step 4 is crucial for obtaining debiased two-stage estimates. The bias introduced by high dimensionality is analogous to that in least squares estimation. Therefore, it is essential to remove this bias.

$$\begin{aligned}\hat{\beta}_{1,h}^{(2S)} &= \left(\sum_t \hat{U}_{1,t}^{\perp} W_{1,t}'\right)^{-1}\left(\sum_t \hat{U}_{1,t}^{\perp} y_{t+h}\right)\\ &= \left(\sum_t \hat{U}_{1,t}^{\perp} W_{1,t}'\right)^{-1}\left(\sum_t \hat{U}_{1,t}^{\perp}(\beta_{1,h}' W_{1,t} + \beta_{2,h}' W_{2,t} + e_{t,h})\right)\\ &= \beta_{1,h} + \left(\sum_t \hat{U}_{1,t}^{\perp} W_{1,t}'\right)^{-1}\left(\sum_t \hat{U}_{1,t}^{\perp} e_{t,h}\right) + \left(\sum_t \hat{U}_{1,t}^{\perp} W_{1,t}'\right)^{-1}\left(\sum_t \hat{U}_{1,t}^{\perp} W_{2,t}'\beta_{2,h}\right).\end{aligned}$$

(iii) The de-biased two-stage estimator $\hat{\beta}_{1,h}^{(de-2S)}$ is the solution to the sample counterpart of the moment condition

$$\mathbb{E}\left[\psi_t^{d2s}\left(\beta_{1,h},\eta_0\right)\right] = 0, \tag{5.11}$$

where[9]

$$\psi_t^{d2s}\left(\beta_{1,h},\eta\right) = \left(R_1 - \Gamma R_2\right)\sum_{j=0}^{p-1} \tilde{e}_{p(j+1)} \otimes I_d\left(w_{t-j} - J\mathbf{A}W_{t-j-1}\right)\left(y_{t+h} - W_{1,t}'\beta_{1,h} - W_{2,t}'\beta_{2,h}\right),$$

$\Gamma_0 := E[U_{1,t}W_{2,t}']\left(E[U_{2,t}W_{2,t}']\right)^{-1}$, $\beta_{2,h} = R_2\left(\mathbf{A}^h\right)'J'e_y$, and $\eta_0 = \left(\beta_{2,h}', \mathrm{vec}\left(\Gamma_0\right)', \mathrm{vec}\left(\mathbf{A}\right)'\right)'$ is a high-dimensional $[(d-1)(p+p^2)+d^2p^2]\times 1$ vector of nuisance parameters. Note that the score function ψ_t^{d2s} is Neyman orthogonal with respect to the nuisance parameter η.

(iv) One crucial aspect in demonstrating the asymptotic distribution for de-biased two-stage estimators is the negligibility of the bias caused by using the estimated residual $\hat{u}_t$ as instruments. This has been proven in the low-dimensional case by Dufour and Wang (2024), showing that the estimation bias for the VAR residual $(\hat{u}_t - u_t)$ does not affect the two-stage estimator (5.9) at the $\sqrt{n}$-level asymptotically. In the subsequent section, we will elaborate on how the estimation bias of the VAR residual has asymptotically negligible effect on two-stage estimates in a high-dimensional framework.

5.3. Asymptotic variance of de-biased two-stage estimators

Analogous to the derivation of the asymptotic variance for de-biased Least Square estimators, obtaining the asymptotic variance of de-biased two-stage estimators requires disentangling the main term from the negligible term in the $\sqrt{n}$-normalized estimation error,

[9] $\tilde{e}_{pj}$, $j = 1,\ldots,p$ denote the d-dimensional unit vectors, where $\tilde{e}_{pj}$ containts 1 at the j[th] position and 0 elsewhere.

$\sqrt{n}\left(\hat{\beta}_{1,h}^{(de-2S)}-\beta_{1,h}\right)$. Under Assumption 2 and certain restrictions on the growth rate of the number of time series d with respect to the sample size n, we show that

$$\sqrt{n}\left(\hat{\beta}_{1,h}^{(de-2S)}-\beta_{1,h}\right)=\left(\mathbb{E}\left[U_{1,t}^{\perp}W_{1,t}'\right]\right)^{-1}\left(n^{-1/2}\sum_{t=p}^{n-h}U_{1,t}^{\perp}e_{t,h}\right)+o_p(1), \tag{5.12}$$

so that the asymptotic variance of the de-biased two-stage estimator is given by

$$\begin{aligned}\mathrm{AVar}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-2S)}\right)&=\lim_{n\to\infty}\left(\mathbb{E}\left[U_{1,t}^{\perp}W_{1,t}'\right]\right)^{-1}\mathrm{Var}\left(n^{-1/2}\sum_{t=p}^{n-h}U_{1,t}^{\perp}e_{t,h}\right)\left(\mathbb{E}\left[U_{1,t}^{\perp}W_{1,t}'\right]\right)^{\prime -1}\\&=\lim_{n\to\infty}(R_1\Sigma_{UW}^{-1}R_1')\Omega_{U_1,h}(R_1\Sigma_{UW}^{\prime -1}R_1'),\end{aligned} \tag{5.13}$$

where $\Omega_{U_1,h}$ is the long run variance of the regress score function,

$$\Omega_{U_1,h}:=\lim_{n\to\infty}\mathrm{Var}(n^{-1/2}\sum_{t}U_{1,t}^{\perp}e_{t,h})=\lim_{n\to\infty}\sum_{k=-\infty}^{\infty}\mathbb{E}[U_{1,t}^{\perp}U_{1,t+k}^{\perp\prime}e_{t,h}e_{t+k,h}].$$

Analogous to conventional time series literature, one possible consistent estimate of the variance $\mathrm{AVar}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-2S)}\right)$ could be obtained by replacing each term within the expression on the right-hand side of the second equality of Equation (5.13) with its sample counterpart. One advantage of two-stage estimation method is to obviate HAC-type covariance matrix. We present a heteroskedastic robust method to compute the covariance matrix following the general HAC-type estimates with a slightly stronger assumption on the innovation process.

HAC/HAR covariance estimates

Practically, researchers can obtain a consistent estimate of the asymptotic variance by applying some HAC type covariance matrix estimates,

$$\widehat{\mathrm{AVar}}^{(hac)}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-2S)}\right)=(R_1\hat{\Sigma}_{UW}^{-1}R_1')\hat{\Omega}_{U_1,h}^{(hac)}(R_1(\hat{\Sigma}_{UW}^{-1})'R_1'), \tag{5.14}$$

where $\hat{\Omega}_{U_1,h}^{(hac)}$ is some consistent HAC estimator of $\Omega_{U_1,h}$, e.g., Newey-West estimate. Since the projection error $e_{t,h}$ is unobservable, it is replaced by a consistent estimate $\hat{e}_{t,h}$, where $\hat{e}_{t,h}=y_{t+h}-\hat{\beta}_hW_t$ and $\hat{\beta}_h=v_1'\left(\hat{\mathbf{A}}^{(re)}\right)^h$, and $\hat{U}_{1,t}^{\perp}$ comes from (5.8).

HC/HR covariance estimates

HAC-type covariance matrix estimates often perform poorly in small samples, particularly regarding the empirical size of statistical tests in linear projection models at horizon h. The practical application of these estimates is further complicated by the need to choose an appropriate bandwidth and kernel function. Consequently, it is worthwhile to explore alternative methods for covariance matrix estimation that rely solely on heteroskedasticity-robust estimation techniques.

Replacing HAC-type covariance matrix estimation with a heteroskedasticity-robust method requires adherence to specific conditions. The motivation for HAC estimation is to ensure a positive semidefinite covariance matrix, which is not necessarily achieved by simply summing all lead-lag autocovariances. Therefore, an alternative method must be found to transform the regression scores into a serially uncorrelated process, thereby equating the long-run variance to the variance. This transformation ensures that the sample variance matrix is inherently positive semidefinite.

By (5.7), $U_{1,t}^{\perp} = (R_1 \Sigma_{UW}^{-1} R_1')^{-1} R_1 \Sigma_{UW}^{-1} U_t$. Then, the long run variance matrix $\Omega_{U_1,h}$ can be rewritten as

$$\Omega_{U_1,h} = (R_1 \Sigma_{UW}^{-1} R_1')^{-1} R_1 \Sigma_{UW}^{-1} \Omega_{U,h} \Sigma_{UW}^{'-1} R_1' (R_1 \Sigma_{UW}^{-1} R_1')^{'-1} \tag{5.15}$$

by defining

$$\Omega_{U,h} := \left(\lim_{n\to\infty} \sum_{k=-\infty}^{\infty} \mathbb{E}[U_t U_{t+k}' e_{t,h} e_{t+k,h}] \right). \tag{5.16}$$

The critical part of estimating $\Omega_{U_1,h}$ is obtaining the sample estimator of $\Omega_{U,h}$, which is the long run variance of the sequence $U_t e_{t,h}$,

$$U_t e_{t,h} = (u_t', u_{t-1}', \cdots, u_{t-p+1}')' e_{t,h}. \tag{5.17}$$

Following the method proposed in Dufour and Wang (2024), we consider an alternative sequence, denoted as

$$s_t := (e_{t,h}, e_{t+1,h}, \cdots, e_{t+p-1,h})' \otimes u_t. \tag{5.18}$$

where the sequence s_t is constructed by replacing each component $u_{t-i} e_{t,h}$ (for $i = 0, 1, \ldots, p-1$) in the vector $U_t e_{t,h}$ with its corresponding i-period lead, $u_{t+i} e_{t+i,h}$. For example, the first component, $u_t e_{t,h}$, remains unchanged; the second component, $u_{t-1} e_{t,h}$, is replaced by $u_t e_{t+1,h}$; and this pattern continues for the remaining terms

Since s_t and $U_t e_{t,h}$ encapsulate the same underlying terms across different time periods, their long-run variance matrices are equivalent. As a result, the long-run variance of both $U_t e_{t,h}$ and s_t yields identical matrices:

$$\sum_{k=-\infty}^{\infty} \mathbb{E}[s_t s_{t+k}'] = \Omega_{U,h}. \tag{5.19}$$

To avoid the need for correcting serial correlation in the projection error and to provide robust statistical inference, we provide regularity conditions that guarantee the process s_t is serially uncorrelated. Thereby the long run variance of s_t is identical to its covariance matrix.

Since u_t is the current shock and $(e_{t,h}, e_{t+1,h}, \cdots, e_{t+p-1,h})$ contains future shocks only. If certain conditions are met, for instance, u_t is i.i.d., then we could derive that the process s_t is serially uncorrelated. The serial uncorrelation for the i.i.d. case can be easily verified by the Law of Iterated Expectations,

$$\begin{aligned}\mathbb{E}[s_t s_\tau'] &= \mathbb{E}[\mathbb{E}[s_t s_\tau \mid u_{t+1}, u_{t+2}, \cdots]] \\ &= \mathbb{E}[(e_{t,h}, e_{t+1,h}, \cdots, e_{t+p-1,h}) \otimes \mathbb{E}[u_t \mid u_{t+1}, u_{t+2}, \cdots] s_\tau'] \\ &= 0 \\ &\text{(since } \mathbb{E}[u_t \mid u_{t+1}, u_{t+2}, \cdots] = 0 \text{ by i.i.d. assumption)}\end{aligned} \tag{5.20}$$

for all $t < \tau$. However, the i.i.d. assumption may be too restrictive, even though it is widely seen in high dimensional time series. We consider a weaker and more general condition.

Assumption 1.
For all $t \geq 1$, let

- *(i) (m.d.s. assumption)* $\mathbb{E}[u_t \mid \{u_s\}_{s<t}] = 0$*, almost surely.*
- *(ii) (some fourth moment assumption)* $\mathbb{E}[(u_t u_\tau') \otimes (u_{\tau+k} u_{\tau+k}')] = \mathbf{0}$*,* $\forall \tau > t, k > 0$*.*

Assumption 1(i) constrains u_t to be a martingale difference sequence, a common condition in the time series literature. Assumption 1(ii) imposes a condition on a specific fourth moment of the disturbances, which is crucial for ensuring the serial uncorrelation of s_t. This condition is met by a wide array of disturbance processes. For instance, it holds when (1) u_t is independent and identically distributed, (2) u_t is mean-independent, (3) u_t follows an ARCH(1) process with Gaussian errors, or (4) u_t is a conditionally homoskedastic process. However, this condition may not hold if u_t follows an ARCH(1) process with skewed errors. In such cases, researchers should employ HAC covariance matrix estimates rather than HC estimates.

We briefly illustrate the sufficiency of Assumption 1 for the serial uncorrelation of the process s_t.

First, we write explicitly the autocovariance of the process s_t by expanding its definition,

$$\mathbb{E}[s_t s_\tau'] = \mathbb{E}[(e_{t,h}, e_{t+1,h}, \cdots, e_{t+p-1,h})(e_{\tau,h}, e_{\tau+1,h}, \cdots, e_{\tau+p-1,h})' \otimes u_t u_\tau']. \tag{5.21}$$

Then, it is easy to see that an equivalent condition for this expectation to be equal to zero is

$$\mathbb{E}[s_t s_\tau'] = \mathbf{0} \quad \Longleftrightarrow \quad \mathbb{E}[u_t u_\tau' \otimes e_{t+i,h} e_{\tau+j,h}] = \mathbf{0}, \tag{5.22}$$

for all $i, j = 0, 1, \cdots, p-1$.

Recall that $e_{t,h} = \sum_{m=1}^{h} v_1' \Psi_{h-m} u_{t+m}$. Then,

$$
\begin{aligned}
&\mathbb{E}[u_t u_\tau' \otimes e_{t+i,h} e_{\tau+j,h}] \\
=&\mathbb{E}\left[u_t u_\tau' \otimes \left(\sum_{m=1}^{h} v_1' \Psi_{h-m} u_{t+i+m} \right) \left(\sum_{n=1}^{h} v_1' \Psi_{h-n} u_{\tau+j+n} \right) \right] \\
=&\mathbb{E}\left[u_t u_\tau' \otimes v_1' \left(\sum_{m=1}^{h} \sum_{n=1}^{h} \Psi_{h-m} u_{t+i+m} u_{\tau+j+n}' \Psi_{h-n}' \right) v_1 \right]
\end{aligned}
\tag{5.23}
$$

Without loss of generality, suppose $t < \tau$. If $t + i + m = \tau + j + n$, then Assumption 1 (ii) ensures that

$$\mathbb{E}\left[u_t u_\tau' \otimes (v_1' \Psi_{h-m} u_{t+i+m} u_{\tau+j+n}' \Psi_{h-n}' v_1) \right] = 0, \tag{5.24}$$

since $t + i + m = \tau + j + n > \tau > t$.

If $t + i + m < \tau + j + n$, then Assumption 1 (i) with Law of Iterated Expectation (LIE) ensures that

$$\mathbb{E}\left[u_t u_\tau' \otimes (v_1' \Psi_{h-m} u_{t+i+m} u_{\tau+j+n}' \Psi_{h-n}' v_1) \right] = 0, \tag{5.25}$$

since $\tau + j + n > \max(t + i + m, \tau, t)$.

If $t + i + m > \tau + j + n$, similarly, Assumption 1 (i) with Law of Iterated Expectation (LIE) ensures that

$$\mathbb{E}\left[u_t u_\tau' \otimes (v_1' \Psi_{h-m} u_{t+i+m} u_{\tau+j+n}' \Psi_{h-n}' v_1) \right] = 0, \tag{5.26}$$

since $t + i + m > \tau + j + n > \tau > t$.

Therefore, combining the results of (5.24)-(5.26), Assumption 1 ensures that s_t is serially uncorrelated. Due to the equality between the long run variance of s_t and the matrix $\Omega_{U,h}$ by (5.19), Assumption 1 yields the matrix $\Omega_{U,h}$ equals to the covariance matrix of s_t,

$$\Omega_{U,h} = \mathrm{Var}(s_t). \tag{5.27}$$

Thus, it entails this is an alternative method to estimate the Long run variance $\Omega_{U,h}$ through the sample variance of s_t:

$$\hat{\Omega}_{U_1,h}^{(HC)} = (R_1 \hat{\Sigma}_{UW}^{-1} R_1')^{-1} R_1 \hat{\Sigma}_{UW}^{-1} \widehat{\mathrm{Var}}(\hat{s}_t) \hat{\Sigma}_{UW}'^{-1} R_1' (R_1 \hat{\Sigma}_{UW}^{-1} R_1')'^{-1} \tag{5.28}$$

where $\widehat{\mathrm{Var}}(\hat{s}_t) = \frac{1}{n-h} \sum_t \hat{s}_t \hat{s}_t'$, $\hat{s}_t = (\hat{e}_{t,h}, \hat{e}_{t+1,h}, \cdots, \hat{e}_{t+p-1,h}) \otimes \hat{u}_t$, $\hat{u}_t = w_t - \hat{\Phi}_{1:p}^{(re)} W_{t-1}$, and $\hat{e}_{t,h}$ is obtained from VAR residuals, as in the HAC-type variance estimation. The heteroskedasticity-robust covariance matrix for two-stage estimates can be computed as

$$\widehat{\mathrm{AVar}}^{(HC)}(\sqrt{n} \hat{\beta}_{1,h}^{(de-2S)}) = R_1 \hat{\Sigma}_{UW}^{-1} \widehat{\mathrm{Var}}(\hat{s}_t) \hat{\Sigma}_{UW}'^{-1} R_1'. \tag{5.29}$$

6. Asymptotic properties of estimators

This section is devoted to the derivation of the asymptotic properties of both estimators. First, we derive the rate of some auxiliary terms that are needed to establish asymptotic normality in the sequel. We show that each of the two estimators is asymptotically normal under certain regularity conditions and restrictions on the growth rate of the number of series d with respect to the sample size n (see Conditions 6.1 and 6.2 below). We then derive the asymptotic inference for the estimators. In particular, we propose a HAC standard error for both de-biased estimator. Additionally, we propose an HC standard error for the de-biased 2S estimaor. Regarding the regularity conditions on the innovation process, the consistency of the HC standard error requires a slightly stronger assumption (see Assumption 1), which can be viewed as a cost for the convenience of avoiding the HAC standard error.

6.1. Preliminary results

This section aims to provide the preliminary consistency results required for establishing the asymptotic normality of the de-biased LS estimate $\hat{\beta}_{1,h}^{(de-LS)}$ and the de-biased 2S estimate $\hat{\beta}_{1,h}^{(de-2S)}$. Before presenting those results, we first state the assumptions needed. A sparsity assumption is needed to establish the consistency of Lasso-type regularized estimators, $\hat{\mathbf{A}}_j^{(re)}, j = 1, \ldots, p$, for VAR slope coefficients $\mathbf{A}_j, j = 1, 2, \ldots, d$, with the corresponding stacked form $\hat{\mathbf{A}}^{(re)}$. We consider the Krampe et al. (2023) adaptation of the Bickel and Levina (2008)'s concept of approximately sparse matrices defined by the following class, $\mathscr{U}(k, \mu)$, of row-wise approximately sparse matrices[10],

$$\mathscr{U}(k,\mu) = \left\{ B = \left(b_{ij}\right)_{i=1,\ldots,r, j=1,\ldots,s} \in \mathbb{R}^{r\times s} : \max_{1\le i\le r} \sum_{j=1}^{s} \left|b_{ij}\right|^{\mu} \le k, \|B\|_2 \le C < \infty \right\}.$$

This class includes the standard exact sparsity class for the special choice of $\mu = 0$, if we adopt the convention that $\sum_{j=1}^{s} \left|b_{ij}\right|^{\mu}$ counts the number of nonzero coefficients in the i^{th} row of the matrix B for $\mu = 0$. Approximate sparsity is considered by allowing to choose μ in a flexible way within the interval $[0, 1)$.

Various papers have investigated the theoretical properties of l_1-regularized estimators in sparse high-dimensional time series models, including stochastic regressions and transition matrix estimation in VAR models (see, e.g., Basu and Michailidis, 2015; Adamek et al., 2023). For this reason, we assume, under the approximate sparsity assumption, the consistency of the regularized estimator $\widehat{\mathbf{A}}^{(re)}$ as specified by part (*iii*) of Assumption 2 below. Assumption 2 collects all the regularity conditions required to obtain the consistency of covariance estimators $\hat{\Sigma}_u$, $\hat{\Sigma}_W$, and $\hat{\Sigma}_{UW}$. Sufficient conditions to obtain asymptotic normality and consistency of variance estimators of both de-biased estimators are given as well. Assumption 2 is partially similar to Assumption 1 of Krampe et al. (2023) in deriving the consistency of the (Lasso) regularized estimator of the structural impulse response.

[10]Note that in this definition, k potentially depends on the dimensions r and s of the matrix B. It measures the degree of row-wise approximate sparsity. The lower k is, the sparser the matrix B is (row-wise).

Assumption 2.

(i) ***Row-wise and Column-wise Approximate Sparsity:*** $\mathbf{A} \in \mathcal{U}\left(k_A, \mu\right)$ *and* $\mathbf{A}' \in \mathcal{U}\left(k_A, \mu\right)$ *for some* $\mu \in [0,1)$ *and* $k_A > 0$.

(ii) ***Stability Conditions:*** *There exists* $\varphi \in (0,1)$ *such that* $\rho(\mathbf{A}) \leq \varphi$ *and for any* $m \in \mathbb{N}$,

$$\|\mathbf{A}^m\|_2 = O\left(\varphi^m\right) \text{ and } \|\mathbf{A}^m\|_l = O\left(k_A \varphi^m\right) \quad \text{for} \quad l \in \{1, \infty\}.$$

(iii) ***Convergence Rate of the Lasso-type Regularized Estimator:*** $\widehat{\mathbf{A}}^{(re)}$ *satisfies,*

$$\left\|\widehat{\mathbf{A}}^{(re)} - \mathbf{A}\right\|_l = O_p\left(k_A^{1.5}\left(\frac{\nu_n}{n}\right)^{(1-\mu)/2}\right) \quad \text{for} \quad l \in \{1, \infty\}.$$

(iv) ***Convergence Rate of the Sample Covariance of Innovations:*** *The sample covariance* $\sum_{t=1}^n u_t u_t' / n$ *satisfies for all* $U, V \in \mathbb{R}^{d\times d}$ *with* $\|U\|_2 = 1 = \|V\|_2$,

$$\left\|\frac{1}{n}\sum_{t=1}^n U\left(u_t u_t' - \Sigma_u\right) V\right\|_{\max} = O_p\left(\sqrt{\bar{\nu}_n / n}\right).$$

(v) ***Moment Restrictions:*** *For all* $j = 1, \ldots, d$, *it holds true that* $\mathbb{E}|\tilde{e}_{jd}' u_t|^q \leq C < \infty$ *for some* $q > 4$, *where* e_{jd}, $j = 1, \ldots, d$, *are* d*-dimensional unit vectors.*

(vi) ***Stability of the Inverse of Covariance Matrices*** Σ_W ***and*** Σ_{UW}***:*** *There exist two sample-dependent functions* $k_W := k_W(n)$ *and* $k_{UW} := k_{UW}(n)$ *such that* $\left\|\Sigma_W^{-1}\right\|_\infty = O\left(k_W\right)$ *and* $\left\|\Sigma_{UW}^{-1}\right\|_\infty = O\left(k_{UW}\right)$, *with* $1/k_W = o(1)$ *and* $1/k_{UW} = o(1)$. *Also,* $\frac{1}{C} \leq \left\|\Sigma_W^{-1}\right\|_2 \leq C$, *and* $\frac{1}{C} \leq \left\|\Sigma_{UW}^{-1}\right\|_2 \leq C$.

(vii) ***Convergence Rate of HAC Estimators and Boundedness of Eigenvalues:*** *There exist a certain function* $\bar{\nu}_n := \bar{\nu}(d, p, q, n)$ *such that* $\left\|\widehat{\Omega}_{W_1,h}^{(hac)} - \Omega_{W_1,h}\right\|_{\max} = O_p\left(\sqrt{\bar{\nu}_n/n}\right)$ *and* $\left\|\widehat{\Omega}_{U_1,h}^{(hac)} - \Omega_{U_1,h}\right\|_{\max} = O_p\left(\sqrt{\bar{\nu}_n/n}\right)$. *Also,* $\frac{1}{C} \leq \lambda_{\min}\left(\Omega_{W_1,h}\right) \leq \lambda_{\max}\left(\Omega_{W_1,h}\right) \leq C$, *and* $\frac{1}{C} \leq \lambda_{\min}\left(\Omega_{U_1,h}\right) \leq \lambda_{\max}\left(\Omega_{U_1,h}\right) \leq C$.

Assumption 2(*i*) imposes both row-wise and column-wise approximate sparsity on the VAR matrix coefficient $\mathbf{A}$. Since $\mathbf{A}$ is a dp-dimensional square matrix, k_A depends on d and thus implicitly depends on the sample size n through d in a high-dimensional context. k_A captures the degree of row-wise sparsity of the matrices $\mathbf{A}$ and $\mathbf{A}'$. A sparse $\mathbf{A}$ will be associated with a low k_A. We expect k_A to be larger than 1 and to increase with d. Assumption 2(*ii*) specifies the standard stability condition of the VAR system. This assumption implies, in part, that the process $\left\{W_t, t \in \mathbb{Z}\right\}$ possesses a geometrically decaying functional dependence coefficient.

Assumption 2(*iii*) provides the rate for estimating the VAR slope coefficients. It assumes consistency of the regularized estimates $\hat{\mathbf{A}}^{(re)}$ and $\left(\hat{\mathbf{A}}^{(re)}\right)'$ in the sense of the maximum absolute raw sum norm. The rate of convergence is formulated in a flexible way, allowing for estimating the VAR slope coefficients using alternative lasso-type approaches, such as adaptive lasso. This rate holds under the sparsity assumption, and the convergence will be faster for a sparser matrix **A** (i.e., for lower k_A). It is evident that lasso-type regularized estimates may fail to be consistent or may converge very slowly if sparsity is wrongly assumed, such that $k_A^{1.5}\nu_n^{(1-\nu)/2}$ tends to be large compared to $n^{(1-\nu)/2}$. The term $\nu_n := \nu(d,p,q,n)$, where ν is an increasing function of d. Its specific form depends on the regularization approach used for estimating **A**, as well as on the number of finite moments q of the innovations u_t. As ν_n lowers the convergence speed of $\hat{\mathbf{A}}^{(re)}$ to **A**, it can be thought of as the cost of using regularization to estimate the high-dimensional object **A**. As emphasized by Krampe et al. (2023), if the innovation process $\{u_t, t \in \mathbb{Z}\}$ has only q moments, then the desired rate is $\nu_n = \log(dp) + (ndp)^{2/q}$, and in particular, $\nu_n = \log(dp)$ in the case of sub-Gaussian innovations. Also, note that if the naive regularized estimator $\hat{\mathbf{A}}^{(re)}$ does not converge at the rate specified in Assumption 2(*iii*), thresholding can be used to obtain an estimator with the desired rate (see Cai and Liu, 2011; Rothman et al., 2009). Thus, if $\hat{\mathbf{A}}^{(re)}$ is defined as in Equation (3.1), then a suitable candidate satisfying Assumption 2(*iii*) is the Thresholded Adaptive LASSO, denoted by $\hat{\mathbf{A}}^{(\mathrm{thr})}$, which is given by

$$\hat{A}_k^{(\mathrm{thr})} = \mathrm{THR}_\lambda(\hat{A}_k^{(re)}) := \left(\mathrm{THR}_\lambda(\hat{A}_{ij,k}^{(re)})\right)_{i,j=1,2,\cdots,d}, \quad k = 1,2,\cdots,p. \tag{6.1}$$

where $\mathrm{THR}_\lambda(z) = z\left(1 - |\lambda/z|^\nu\right)_+$ with $\nu \geq 1$. Soft thresholding ($\nu = 1$) and hard thresholding ($\nu = \infty$) represent boundary cases of this function (see Krampe and Paparoditis, 2021 and Section 4.1 in Krampe et al., 2023).

Assumption 2(*iv*) outlines the requirement for entry-wise consistency of the sample covariance of the innovations. This assumption comes directly from the number of finite moments (see Assumption 2(*v*)) and does not require any sparsity assumption on the contemporaneous covariance matrix of innovations, Σ_u. Additionally, $\tilde{\nu}_n := \tilde{\nu}(d,q,n)$ represents the cost associated with increasing dimensionality. Specifically, if only q moments of the innovations are finite, then the desired rate is $\tilde{\nu}_n = \log(d) + (nd)^{4/q}$ and for sub-Gaussian innovations, we have $\tilde{\nu}_n = \log(d)$.

Note that Assumptions 2(*iii*) and (*iv*) implicitly impose restrictions on the rate at which d grows to infinity relative to the sample size n. To illustrate, assume that d and k_A scale as $d = O(n^\phi)$ and $k_A = O(n^\psi)$, where $\phi > 0$ and $\psi > 0$. Simple calculations indicate that $\tilde{\nu}_n/n \to 0$ implies $\phi < q/4 - 1$ if $q > 4$. Similarly, $k_A^{3/(1-\mu)}(\nu_n/n) \to 0$ implies that $\psi < (1-\mu)(1-4/q)/3$ and $\phi < q\left(1 - 3\psi/(1-\mu)\right)/4 - 1$. These conditions imply, in the case of exact sparsity ($\mu = 0$), that $\psi < (1-4/q)/3$ and $\phi < q\left(1-3\psi\right)/4 - 1$ if $q > 4$. The restriction on the growth rate of d will be less stringent if the innovations have moments of higher orders (i.e., if q is large). For example, if $q = 8$ (i.e., $\psi < 1/6$) and $\psi = 1/7$, then the restriction on the growth rate of d with respect to n is $\phi < 1/7$. Similarly, if $q = 16$ (i.e., $\psi < 1/4$) and $\psi = 1/5$, then the restriction $\phi < 3/5$, is less stringent.

Assumption 2(*vi*) states additional conditions for deriving the convergence rate of the inverse of the covariance matrix estimators $\hat{\Sigma}_W$ and $\hat{\Sigma}_{UW}$. k_U and k_{UW} can be seen as the costs of inverting dp-square matrices Σ_W and Σ_{UW} when allowing for increasing dimensionality. Additionally, Assumption 2(*vi*) implies that there exists a constant C such that $C\lambda_{\min}\left(\Sigma_W\right) \geq 1/k_W = o(1)$ and $C\lambda_{\min}\left(\Sigma_{UW}\right) \geq 1/k_{UW} = o(1)$. Thus, the matrices Σ_W and Σ_{UW} are non-singular in finite sample, but their inverses might be slightly unstable in an asymptotic regime where d goes to infinity with n. The degree of stability of these inverses depends on how fast the functions k_U and k_{UW} go to infinity with the sample size. As we expect k_U and k_{UW} to increase very slowly with the sample size, we will end up with matrices Σ_W and Σ_{UW} that are relatively non-singular so that their inverses exist for large n.

Finally, Assumption 2(vii) specifies the convergence rate of the HAC estimators $\hat{\Omega}^{hac}_{W_1,h}$ and $\hat{\Omega}^{hac}_{U_1,h}$. $\bar{\nu}_n := \bar{\nu}(d,p,q,n)$ can be thought of as the cost of allowing the dimension d to increase and using regularization to obtain the estimator $\hat{\mathbf{A}}^{(re)}$ that enters the computation of $\hat{\Omega}^{hac}_{W_1,h}$ and $\hat{\Omega}^{hac}_{U_1,h}$. This assumption is useful to show the consistency of the variance estimator of the de-biased estimators. Additionally, Assumption 2(vii) imposes some restrictions on the structure of the long-run variances $\Omega_{W_1,h}$ and $\Omega_{U_1,h}$. In particular, it requires the eigenvalues of both matrices to be bounded above and below, away from zero, by a constant.

Under these assumptions, we have the following consistency results for the covariance matrix estimators $\hat{\Sigma}_u$, $\hat{\Sigma}_W$, and $\hat{\Sigma}_{UW}$.

Theorem 6.1 (**Consistency results**).
Let $\widehat{\Sigma}_W$ and $\widehat{\Sigma}_{UW}$ denote the regularized estimator of Σ_W and Σ_{UW}, respectively, using explicit formulas (4.2) *and* (5.3). *Under Assumption 2, the following assertions are true:*

$$(i)\quad \left\|\widehat{\Sigma}_u - \Sigma_u\right\|_\infty = O_p\left(d\left[k_A^3\left(\nu_n/n\right)^{1-\mu} + \sqrt{\tilde{\nu}_n/n}\right]\right);$$

$$(ii)\quad \left\|\widehat{\Sigma}_W - \Sigma_W\right\|_\infty = O_p\left(dk_A^2\left\{k_A^{2.5}\left(\nu_n/n\right)^{(1-\mu)/2} + \sqrt{\tilde{\nu}_n/n}\right\}\right);$$

$$(iii)\quad \left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_\infty = O_p\left(dk_A\left\{k_A^{1.5}\left(\nu_n/n\right)^{(1-\mu)/2} + \sqrt{\tilde{\nu}_n/n}\right\}\right).$$

Remark 6.1. Note that the convergence rates of all those covariance estimators depend on both the convergence rate of the sample covariance of the innovations, $\sum_{t=1}^n u_t u_t'/n$, and the convergence rate of the regularized estimator $\hat{\mathbf{A}}^{(re)}$ of the VAR matrix coefficient $\mathbf{A}$. In all cases, the convergence speed also depends on the growth rate of the number of series d relative to the sample size n.

6.2. Asymptotic theory for de-biased LS estimator

In this section, we derive asymptotic normality and properties for statistical inference for the de-biased LS estimator defined by Equation (4.8). We show that the de-biased LS estimator $\hat{\beta}^{(de-LS)}_{1,h}$ is asymptotically normal and derive the consistency of its variance estimator defined by Equation (4.14) under certain restrictions on the growth rate of d relative to n (see Conditions 6.1 and 6.2). Importantly, we show under the conditional homoskedastic martingale difference sequence (m.d.s.)[11] assumption that the asymptotic variance defined by Equation (4.13) has a closed-form expression in terms of a truncated sum. Even in this particular case, we recommend using a kernel estimator, in the spirit of HAC estimation, to avoid situations where the covariance matrix estimator may be non-positive semi-definite due to truncation. Conditions 6.1 and 6.2 implicitly restrict the growth rate of d relative

[11]Note that the following two conditions should be satisfied for u_t to be a conditional homoskedastic m.d.s.: *(i)* $E\left[u_t \mid u_{t-1}, u_{t-2}, \cdots\right] = 0$ and *(ii)* $E\left[u_t u_t' \mid u_{t-1}, u_{t-2}, \cdots\right] = \Sigma_u$.

to n to ensure the asymptotic normality and consistency of the variance estimator for the de-biased LS. These conditions impose more stringent restrictions on d compared to those required for the consistency of the regularized estimator $\hat{\mathbf{A}}^{(re)}$ as stated in Assumption 2(*iii*).

In order to derive the asymptotic distributional theory, we impose additional standard regularity conditions on the innovation process and observables.

Assumption 3 (Regularity conditions).

(i) ***Strong mixing condition:*** *u_t and W_t are strong mixing (α-mixing) processes with a mixing size of $-r/(r-2)$, where $r > 2$.*

(ii) ***Boundedness of the moments of innovations:*** *For any $\lambda \in \mathbb{R}^{d\times 1}$ with $\|\lambda\|_2 = 1$, $\mathbb{E}|\lambda' u_t|^{2r+\delta} < c_0 < \infty$, for some constants c_0 and $\delta > 0$.*

Assumption 3 is a standard regularity condition on the time-dependence of the VAR innovation process and observables. The strong mixing condition (Assumption 3(*i*)) and the boundedness of moments (Assumption 3(*ii*)) are important for asymptotic normality. Moreover, Assumption 3 implies that there is a trade-off between the number of moments possessed by the innovation process and the memory of the series u_t and W_t. In fact, allowing for more dependence (i.e., for large r) will impose a strong restriction on the existing moments. In contrast, allowing for less dependence (i.e., for low r) will relax the restriction on the number of moments. For example, if the mixing coefficient $\alpha(k)$ exponentially decays with k (e.g., $\alpha(k) = c_1\rho^k$, for $0 < \rho < 1$ and c_1 a non-negative constant), then r can be set arbitrarily close to 2, so that Assumption 3(*ii*) just requires the existence of more than 4 moments, as in Assumption 2(*v*). Furthermore, Assumption 3 is less restrictive than the i.i.d innovation assumption impose, for example, by Krampe et al. (2023). It allows for a large range of time-dependent, although uncorrelated, innovation processes u_t, such as strongly mixing martingale difference sequences (e.g., ARCH and GARCH processes under certain restrictions).

Condition 6.1. $\tilde{\nu}_n^{1/2} d k_A^2 k_W \left\{ k_A^{2.5} \left(\nu_n/n\right)^{(1-\mu)/2} + \left(\tilde{\nu}_n/n\right)^{1/2} \right\} = o(1).$

This condition imposes an implicit restriction on the growth rate of d relative to n to ensure that the higher-order term in the $\sqrt{n}$-normalized estimation error, $\sqrt{n}\left(\hat{\beta}_{1,h}^{(de-LS)} - \beta_{1,h}\right)$, is effectively negligible. The asymptotic behavior of the de-biased LS estimator is then driven by the main term as specified in Equation (4.12). The following theorem establishes the asymptotic normality of the de-biased LS estimator.

Theorem 6.2 (Asymptotic normality of the de-LS estimator).
Under Assumptions 2 and 3, if the number of series d grows with n such that Condition 6.1 is satisfied, then for any vector $v \in \mathbb{R}^p$ such that $\|v\|_1 = 1$,

$$\frac{\sqrt{n} v'(\hat{\beta}_{1,h}^{(de-LS)} - \beta_{1,h})}{s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v)} \xrightarrow{d} \mathcal{N}(0,1), \quad as \quad n \to \infty, \tag{6.2}$$

where $s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v)^2 := v' \,\mathrm{AVar}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-LS)}\right) v.$

Condition 6.2. $k_W^2\left\{dk_A^2\left[k_A^{2.5}\left(\nu_n/n\right)^{(1-\mu)/2}+\left(\tilde{\nu}_n/n\right)^{1/2}\right]+\left(\bar{\nu}_n/n\right)^{1/2}\right\}=o(1).$

Condition 6.2 imposes an additional restriction to obtain consistency of the variance estimator for de-biased LS, as stated by the following theorem.

Theorem 6.3 (**Consistency of the variance estimator for de-LS**).
Under Assumptions 2 and 3, if the number of series d grows with n such that Conditions 6.1 and 6.2 are satisfied, then for any vector $v\in\mathbb{R}^p$ such that $\|v\|_1=1$,

$$\left|\widehat{s.e.}^{(hac)}_{\hat{\beta}^{(de-LS)}_{1,h}}(v)^2-s.e._{\hat{\beta}^{(de-LS)}_{1,h}}(v)^2\right|\xrightarrow{p}0, \tag{6.3}$$

where $\widehat{s.e.}^{(hac)}_{\hat{\beta}^{(de-LS)}_{1,h}}(v)^2:=v'\widehat{\text{AVar}}^{(hac)}\left(\sqrt{n}\hat{\beta}^{(de-LS)}_{1,h}\right)v.$

Furthermore, the result (6.2) in Theorem 6.2 still holds if $s.e._{\hat{\beta}^{(de-LS)}_{1,h}}(v)$ is replaced by its estimator, $\widehat{s.e.}^{(hac)}_{\hat{\beta}^{(de-LS)}_{1,h}}(v)$.

If in addition, the VAR *error term u_t is a conditional homoskedastic m.d.s, then $s.e._{\hat{\beta}^{(de-LS)}_{1,h}}(v)$ has a closed-form expression of the form*[12]*,*

$$s.e._{\hat{\beta}^{(de-LS)}_{1,h}}(v)^2+o(1)=\sum_{j,l=0}^{h-1}e_y'\Psi_j\Sigma_u\Psi_l'e_y v'R_1\Sigma_W^{-1}\Sigma_W(l-j)\Sigma_W^{-1}R_1'v, \tag{6.4}$$

and can be consistently estimated using $\hat{\Sigma}_u$, $\hat{\Psi}_j$, and $\hat{\Sigma}_W(j)$.

Corollary 6.4 (**Limiting distribution of the Wald test statistic**).
If Assumptions 2 and 3 hold and the number of series d grows with n such that Conditions 6.1 and 6.2 are satisfied, then under the null hypothesis $\mathscr{H}_0:\beta_{1,h}=0$, the Wald test statistic

$$W_n^{(de-LS)}:=n\hat{\beta}^{(de-LS)\prime}_{1,h}\left(\widehat{\text{AVar}}^{(hac)}\left(\sqrt{n}\hat{\beta}^{(de-LS)}_{1,h}\right)\right)^{-1}\hat{\beta}^{(de-LS)}_{1,h}\xrightarrow{d}\chi^2(p),\quad as\quad n\to\infty.$$

6.3. Asymptotic theory for de-biased 2S estimator

In this section, we derive the asymptotic normality and properties for statistical inference for the de-biased 2S estimator defined by Equation (5.10). We demonstrate that the de-biased 2S estimator $\hat{\beta}^{(de-2S)}_{1,h}$ is asymptotically normal and derive the consistency of the HAC variance estimator, defined by Equation (5.14), under certain restrictions on the growth rate of d relative to n (see Conditions 6.3 and 6.4). Furthermore, we establish the consistency of

[12]The notation $\Sigma_W(r)$ in (6.4) refers to the lag-r autocovariance matrix of W_t and has the closed-form representation:

$$\Sigma_W(r):=E[W_tW_{t-r}']=\sum_{j=0}^{\infty}\mathbf{A}^{j+r}J'\Sigma_uJ(\mathbf{A}')^j=\mathbf{A}^r\text{vec}_{dp}^{-1}\big((I_{d^2p^2}-\mathbf{A}\otimes\mathbf{A})^{-1}\text{vec}(J'\Sigma_uJ)\big),$$

for $r\geq 0$ and $\Sigma_W(r)=\Sigma_W(-r)'$ for $r<0$.

the HC variance estimator, defined by (5.29), under additional restrictions on the structure of the innovations u_t (see Assumptions 1 and 4). Note that Conditions 6.3 and 6.4 required for the asymptotic results of the de-biased 2S estimator are slightly less stringent than those required for deriving properties of the de-biased LS estimator.

Condition 6.3. $\tilde{\nu}_n^{1/2} d k_A k_{UW} \left\{ k_A^{1.5} \left(\nu_n/n\right)^{(1-\mu)/2} + \left(\tilde{\nu}_n/n\right)^{1/2} \right\} = o(1).$

Condition 6.3 sets a constraint on how d grows relative to n to ensure that the higher-order term in the $\sqrt{n}$-normalized estimation error, $\sqrt{n}\left(\hat{\beta}_{1,h}^{(de-2S)} - \beta_{1,h}\right)$, becomes negligible. Consequently, the asymptotic behavior of the de-biased 2S estimator is governed by the main term specified in Equation (5.12). The following theorem demonstrates the asymptotic normality of the de-biased 2S estimator.

Theorem 6.5 (**Asymptotic normality of the de-2S estimator**).
Under Assumptions 2 and 3, if the number of series d grows with n such that Condition 6.3 is satisfied, then for any vector $v \in \mathbb{R}^p$ such that $\|v\|_1 = 1$,

$$\frac{\sqrt{n} v'(\hat{\beta}_{1,h}^{(de-2S)} - \beta_{1,h})}{s.e._{\hat{\beta}_{1,h}^{(de-2S)}}(v)} \xrightarrow{d} \mathcal{N}(0,1), \quad as \quad n \to \infty, \tag{6.5}$$

where $s.e._{\hat{\beta}_{1,h}^{(de-2S)}}(v)^2 := v' \operatorname{AVar}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-2S)}\right) v$.

Condition 6.4. $k_W^2 \left\{ d k_A \left[k_A^{1.5} \left(\nu_n/n\right)^{(1-\mu)/2} + \sqrt{\tilde{\nu}_n/n} \right] \right\} = o(1).$

Condition 6.4 imposes an extra constraint on the growth rate of d necessary for the consistency of the variance estimators of de-biased 2S, as demonstrated by the following theorem.

Although Assumptions 2 and 3 are sufficient to ensure the consistency of the HAC-type estimator of the asymptotic variance of the de-biased LS, as defined by Equation (5.14), the consistency of the HC-type variance estimator, as defined by Equation (5.29), requires stronger moment restrictions on the innovation process, as stated by Assumption 4 below.

Assumption 4 (Regularity Conditions II).
For any $\lambda \in \mathbb{R}^{d\times 1}$ with $\|\lambda\|_2 = 1$, it holds that

$$\mathbb{E}\left|\lambda' u_t\right|^{4r+\delta} < c_0 < \infty,$$

for some constants c_0 and $\delta > 0$, where r is defined as in Assumption 3(i).

Assumption 4 is a stronger version of the regularity condition in Assumption 3(*ii*). It requires the boundedness of higher moments for the convergence of the sample covariance matrix of the process s_t as defined by Equation (5.18).

Theorem 6.6 (**Consistency of the variance estimators for de-2S**).
Under Assumptions 2 and 3, if the number of series d grows with n such that Conditions 6.3 and 6.4 are satisfied, then for any vector $v \in \mathbb{R}^p$ such that $\|v\|_1 = 1$,

$$\left| \widehat{s.e.}^{(hac)}_{\hat{\beta}_{1,h}^{(de-2S)}}(v)^2 - s.e._{\hat{\beta}_{1,h}^{(de-2S)}}(v)^2 \right| \xrightarrow{p} 0, \tag{6.6}$$

where $\widehat{s.e.}^{(hac)}_{\hat{\beta}^{(de-2S)}_{1,h}}(v)^2 := v'\widehat{\text{AVar}}^{(hac)}\left(\sqrt{n}\hat{\beta}^{(de-2S)}_{1,h}\right)v$.

Furthermore, the result (6.5) *in Theorem* 6.5 *still holds if* $s.e._{\hat{\beta}^{(de-2S)}_{1,h}}(v)$ *is replaced by its estimator,* $\widehat{s.e.}^{(hac)}_{\hat{\beta}^{(de-2S)}_{1,h}}(v)$.

If, in addition, the VAR *error term* u_t *satisfies Assumptions* 1 and 4, *then*

$$\left|\widehat{s.e.}^{(HC)}_{\hat{\beta}^{(de-2S)}_{1,h}}(v)^2 - s.e._{\hat{\beta}^{(de-2S)}_{1,h}}(v)^2\right| \xrightarrow{p} 0, \tag{6.7}$$

where $\widehat{s.e.}^{(HC)}_{\hat{\beta}^{(de-2S)}_{1,h}}(v)^2 := v'\widehat{\text{AVar}}^{(HC)}\left(\sqrt{n}\hat{\beta}^{(de-2S)}_{1,h}\right)v$.

Corollary 6.7 (Limiting distribution of the Wald test statistic).
If Assumptions 2 *and* 3 *hold and the number of series d grows with n such that Conditions* 6.3 *and* 6.4 *are satisfied, then under the null hypothesis* $\mathcal{H}_0 : \beta_{1,h} = 0$, *the Wald test statistic*

$$W_n^{(de-2S)} := n\hat{\beta}^{(de-2S)\prime}_{1,h}\left(\widehat{\text{AVar}}^{(hac)}\left(\sqrt{n}\hat{\beta}^{(de-2S)}_{1,h}\right)\right)^{-1}\hat{\beta}^{(de-2S)}_{1,h} \xrightarrow{d} \chi^2(p), \quad \text{as} \quad n \to \infty.$$

7. Monte Carlo simulations

This section reports the results of Monte Carlo experiment designed to evaluate the finite sample performance of the Wald test. We consider three cases of VAR(p), $p = 2$. The choice of a VAR(2) model is to accommodate more general empirical exercises. Since the DGP needs to be stationary, we generate the VAR slope coefficients by factorizing the VAR coefficient polynomial and determining the root matrices:

$$(I - \Lambda_1 L)(I - \Lambda_2 L)w_t = u_t \tag{7.1}$$

where L is the lag operator, Λ_k is the root matrix, and $u_t \sim i.i.d.N(0, \Sigma_u)$, $\Sigma_{ij,u} = 0.5^{|i-j|}$. Inspired by the literature on high dimensional VAR, e.g., Miao et al. (2023), we consider three types of root matrices:

(i) DGP 1 (Tridiagonal root matrix): $\Lambda_{ij,k} = 0.3^{|i-j|+1}$.

(ii) DGP 2 (Block-diagonal root matrix): Λ_k is a block diagonal matrix, where $\Lambda_k = \text{diag}[S_{i,k}]$ and $S_{i,k}$ is a square matrix of dimension 5. The diagonal entries of $S_{i,k}$ are fixed with 0.3. For each column of $S_{i,k}$, we randomly set two of those off-diagonal entries as −0.2.

(iii) DGP 3 (Random root matrix): We fix the diagonal entries of Λ_k to be 0.3. In each column of Λ_k, we randomly choose 3 out of $d-1$ entries and set them to be −0.2.

Once the root matrices are determined, we could obtain the VAR slope coefficients as

$$\begin{aligned} A_1 &= \Lambda_1 + \Lambda_2, \\ A_2 &= -\Lambda_1\Lambda_2, \end{aligned} \tag{7.2}$$

We fix the number of time series to $d = 60$. To accommodate the majority of macroeconomic datasets, we consider three cases of dimensionality: $n = 120$, $n = 240$, and $n = 480$, corresponding respectively to strong, moderate, and slight high-dimensionality.

We considered 1000 Monte Carlo replications. For each simulation, we implement the de-biased LS estimation, the de-biased two-stage estimation, and the post-double-selection LASSO estimation for horizon from one to twenty-four. The long run variance matrix of the regression score function is estimated by '*getLongRunVar*' command from '*cointReg*' package in R program with 'h' bandwidth and 'bartlett' kernel function. We use '*HDeconometrics*' package and '*glmnet*' package in R program to implement adaptive LASSO. The tuning parameter for the first step LASSO on VAR is chosen as $\lambda = \sqrt{\log d/n}$.

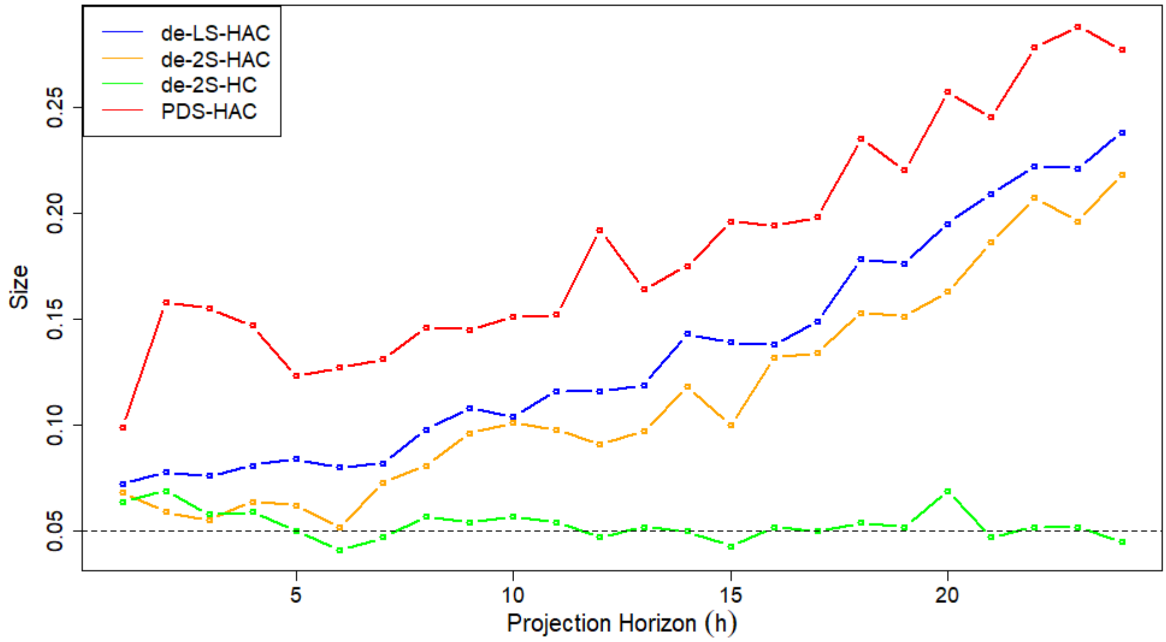

Fig. 3: Size of the Wald test at the 5% nominal level for different horizons. The red, blue, orange, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased two-stage with HAC standard errors, and the de-biased two-stage with HC standard errors, respectively. The number of time series is $d = 60$, and the sample size is $n = 120$. The horizon is $h = 0, 1, \ldots, 24$. The number of replications is 1,000.

Figure 3 provides a comparison of the performance of the Wald test in the case of strong high-dimensionality (with a sample size of $n = 120$ and the number of series $d = 60$) for different approaches used to perform the test: de-biased least squares with HAC standard errors, de-biased two-stage with HAC standard errors, de-biased two-stage with HC standard errors, and the post-double selection procedure with HAC standard errors. We use the size of the Wald test, approximated by the rejection frequency over the simulation replications, as a measure of performance. As can be seen, the two-stage approach with heteroskedastic-consistent (HC) robust standard errors outperforms the two-stage or least-squares approaches with HAC-type standard errors, particularly for large projection horizons. Indeed, as the projection horizon increases, HC inference provides good size, while sizes for HAC-type inference worsen. This size distortion arises because HAC-type variance estimators tend to become imprecise for higher horizons due to finite sample performance issues, as verified in low-dimensional local projection cases by Montiel Olea and Plagborg-Møller (2021) and Dufour and Wang (2024). However, this problem is exacerbated in our context by high dimensionality. Moreover, our procedures outperform the post-double-selection procedure with HAC inference for all horizons. Mitigating the degree of high dimensionality by increasing the sample size to $n = 240$ and $n = 480$ leads to similar results, although the size distortion attenuates and the discrepancy between curves reduces (see Figures A.1 and A.2 in the Appendix), denoting the convergence of all approaches toward the OLS benchmark for large samples. However, the de-biased two-stage method with HC standard errors tends to slightly under-reject the null hypothesis for larger samples and longer horizons.

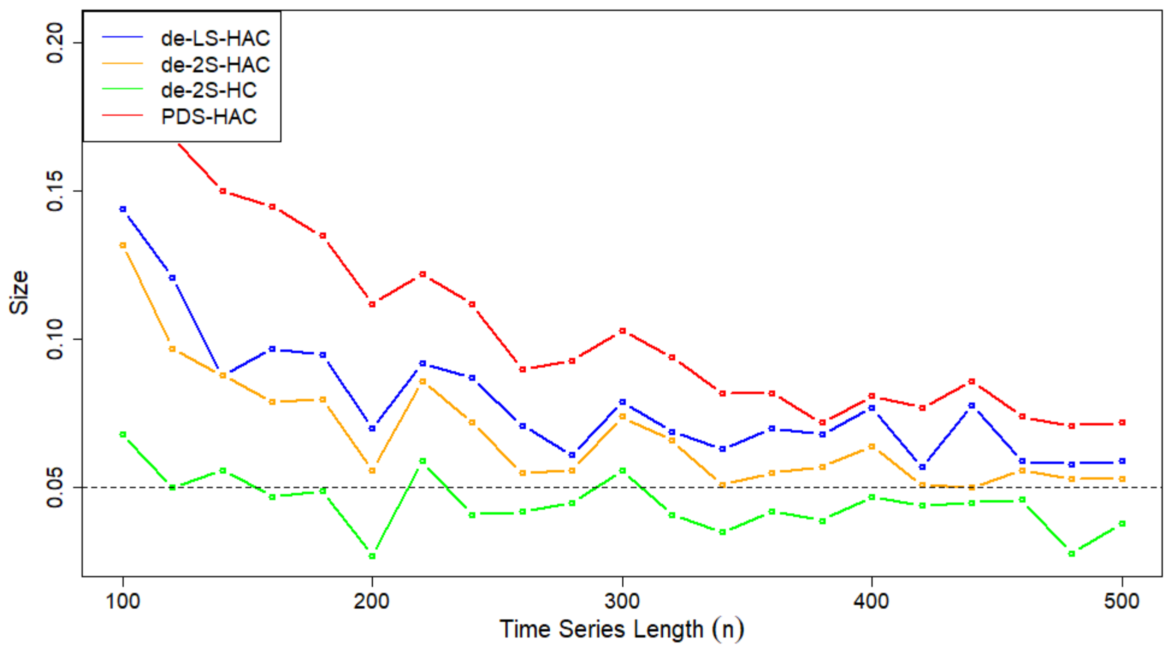

Fig. 4: Size of the Wald test at the 5% nominal level for different sample sizes and a given horizon ($h = 12$). The red, blue, orange, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased two-stage with HAC standard errors, and the de-biased two-stage with HC standard errors, respectively. The number of time series is $d = 60$. The number of replications is $1,000$.

Additionally, Figure 4 compares the size of the Wald test across different approaches for a fixed horizon ($h = 12$) and increasing sample sizes. It is obvious that the size of the Wald test converges to the nominal level for all inference procedures. Clearly, approaches based on the two-stage estimator provide better performance and better convergence rates to the nominal level. These results are consistent across different horizons, as shown by Figures A.3 to A.5.

In summary, the simulation results show that our approaches for testing the null hypothesis of Granger non-causality using the Wald test perform well. The de-biased two-stage estimator with heteroskedasticity-consistent (HC) standard errors undeniably provides the best performance in the simulation framework we considered, highlighting the advantage of debiasing the two-stage estimator and providing a robust variance estimator. This motivates our recommendation to use this approach in practice if the weak assumptions imposed on the VAR innovations to obtain the consistency of the HC variance estimator are likely to be satisfied. I.i.d. innovations and many martingale difference sequences obviously meet these conditions.

8. Empirical application: country-level economic policy uncertainty causal network

In this section, we apply our methodology to investigate the spillovers and contagions of economic uncertainty among a large set of countries and over time using multi-horizon Granger causality tests. We rely on the measure of policy-related economic uncertainty developed by Baker et al. (2016). The policy economic uncertainty index is constructed from three types of underlying components: (i) the first component quantifies newspaper coverage of policy-related economic uncertainty; (ii) the second component measures the level of uncertainty regarding the future path of the federal tax code; and (iii) the third component captures the level of uncertainty associated with macroeconomic variables. Data on uncertainty indices are collected from the Economic Policy Uncertainty website. Our sample consists of 20 series (20 countries) of country-level monthly indices collected from January 2003 to February 2024, totaling $n = 254$ observations.

We conduct pairwise Granger causality tests at different horizons. For a specific horizon h and for each pair of countries A and B, we check Granger causality from A to B conditional on countries other than A and B, and vice versa. We assume that our 'high-dimensional' system follows a VAR representation. We test the null hypothesis of Granger non-causality using a Wald test based on the two-stage de-biased estimator. The tests are performed at a 10% significance level, and critical values are taken from $\chi^2(df)$, with $df = 4$. In Figure 5, we represent the resulting causal graph at different horizons. For each horizon and for each cell, the darker the color, the stronger the causal relation from the corresponding column country to the row country. A white cell means that there is no causality from the column country to the row country.

Figure 5 reveals, among other things, that: (1) Almost all countries exhibit causality to themselves, either in the short run or in the long run; (2) There is causality from the US to China in the short run, but no causality in the long run. Conversely, there is no causality from China to the US in the short run, but there is strong causality in the long run; (3) There is no causality from the US to the UK, neither in the short run nor in the long run. The same result holds for causality from the UK to the US; (4) There is no causality from the US to Canada at any horizon we considered. Likewise, there is no causality from Canada to the US; (5) There is causality from Canada to France in the short run and mid-term. However, there is no causality from France to Canada, neither in the short run nor in the long run. It may seem surprising that there is no causal effect between the US and Canada, but this could be explained by the multi-dimensional nature of our sample. Indeed, it is possible that, conditional on all other countries considered besides Canada and the US, Canada does not provide relevant information to improve the prediction of uncertainty around the US economy at different horizons we consider, and vice versa. Of course, considering a model that includes only the indices of Canada and the US would obviously lead to a causal effect, which could be misleading due to the omission of other variables.

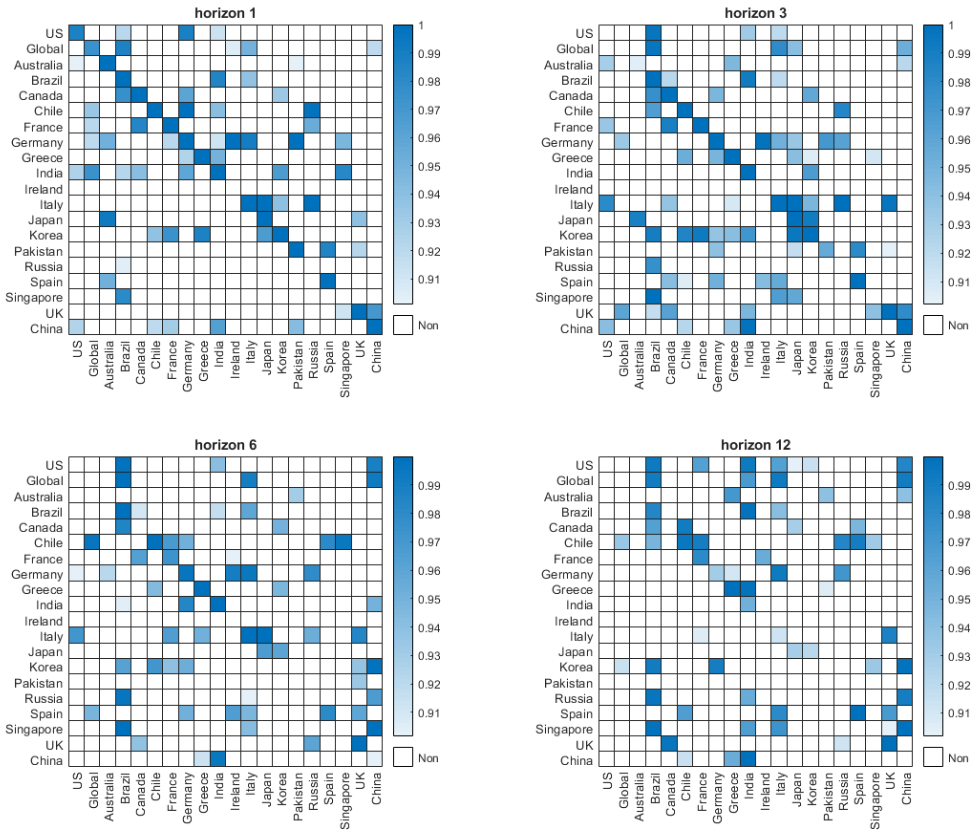

Fig. 5: Granger causality test at different horizons. For each cell, the darkness of the color is $1-p$-value and represents the strength of the Granger causality from the column variable to the row variable. Data: 2003:01 - 2024:02 time span, 254 number of observations, and 20 countries. VAR estimation method: adaptive LASSO on VAR(4). Causality test: two-stage estimation, Wald test with critical value from $\chi^2(df)$, $df = 4$.

9. Conclusion

In this paper, we investigate a Wald test for multi-horizon Granger causality within a high-dimensional sparse VAR framework. To define the Wald test statistics, we propose two types of de-biased estimation methods for the multi-horizon Granger-causal coefficients: the Least Squares method and a two-stage procedure, along with HAC standard error estimates. To ensure robust inference, we impose a specific regularity condition and derive HR/HC standard errors that do not require correcting for serial correlation in the projection residuals. Finally, we apply our methodology to analyze the spread of economic uncertainty at the country level and visualize causal connectedness based on the significance levels of the causality tests.

Our de-biased estimators address the econometric challenge posed by Local Projection (LP) equations for horizons $h > 1$, which may not be sparse even under a sparsity assumption on the underlying VAR process. From a practical perspective, our robust inference approach alleviates the relatively poor performance of HAC-based inference or the computational burden of bootstrap methods in high-dimensional settings. Our application underscores that high-dimensional multi-horizon Granger causality tests offer a more comprehensive understanding of the causal mechanisms within dynamic systems compared to single-horizon Granger tests, expanding the toolkit for practitioners conducting causality studies across multiple horizons.

Our study presents several avenues for future research. One compelling and econometrically challenging extension is to move beyond the linear structure of the HD-VAR framework and investigate Granger-causal coefficients and impulse responses in a nonlinear setting. This would result in Local Projection (LP) equations that become nonlinear transformations of the underlying nonlinear VAR model, thereby taking on a more complex form. A promising approach could involve adopting flexible functional approximations, such as nonparametric series estimators, as proposed by Belloni et al. (2014a). Similarly, Hecq et al. (2023) recognize the potential of incorporating nonlinear regressors, such as quadratic terms or Rectified Linear Units (ReLU), to enhance flexibility in high-dimensional VAR models. These advancements have significant applications in macroeconomics, particularly in the study of nonlinear (state-dependent) causal responses, as demonstrated by Gonçalves, Herrera, Kilian, and Pesavento, 2021 and 2024.

References

R. Adamek, S. Smeekes, and I. Wilms. Lasso inference for high-dimensional time series. *Journal of Econometrics*, 235(2):1114–1143, 2023.

R. Adamek, S. Smeekes, and I. Wilms. Local projection inference in high dimensions. *The Econometrics Journal*, page utae012, 2024.

A. Babii, E. Ghysels, and J. Striaukas. High-dimensional granger causality tests with an application to vix and news. *Journal of Financial Econometrics*, 22(3):605–635, 2024.

S. R. Baker, N. Bloom, and S. J. Davis. Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4):1593–1636, 2016.

S. Basu and G. Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, pages 1535–1567, 2015.

A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

A. Belloni, V. Chernozhukov, and C. Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50, 2014a.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650, 2014b.

B. S. Bernanke, J. Boivin, and P. Eliasz. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach. *The Quarterly journal of economics*, 120(1):387–422, 2005.

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577 – 2604, 2008. doi: 10.1214/08-AOS600. URL https://doi.org/10.1214/08-AOS600.

J. Breitung and R. Brüggemann. Projection estimators for structural impulse responses. *Oxford Bulletin of Economics and Statistics*, 85(6):1320–1340, 2023.

T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

R. A. Davis, P. Zang, and T. Zheng. Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096, 2016.

F. X. Diebold and K. Yılmaz. On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of econometrics*, 182(1):119–134, 2014.

J.-M. Dufour and E. Renault. Short run and long run causality in time series: theory. *Econometrica*, pages 1099–1125, 1998.

J.-M. Dufour and A. Taamouti. Short and long run causality measures: Theory and inference. *Journal of Econometrics*, 154(1):42–58, 2010.

J.-M. Dufour and E. Wang. Simple robust two-stage estimation and inference for generalized impulse responses and multi-horizon causality. *Technical Report, McGill University, Economic Department*, 2024.

J.-M. Dufour, D. Pelletier, and É. Renault. Short run and long run causality in time series: inference. *Journal of Econometrics*, 132(2):337–362, 2006.

J. Geweke. Inference and causality in economic time series models. *Handbook of econometrics*, 2:1101–1144, 1984.

S. Gonçalves, A. M. Herrera, L. Kilian, and E. Pesavento. Impulse response analysis for structural dynamic models with nonlinear regressors. *Journal of Econometrics*, 225(1): 107–130, 2021.

S. Gonçalves, A. M. Herrera, L. Kilian, and E. Pesavento. State-dependent local projections. *Journal of Econometrics*, page 105702, 2024.

C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

F. Han and H. Liu. Transition matrix estimation in high dimensional time series. In *International conference on machine learning*, pages 172–180. PMLR, 2013.

A. Hecq, L. Margaritella, and S. Smeekes. Granger causality testing in high-dimensional vars: a post-double-selection procedure. *Journal of Financial Econometrics*, 21(3):915–958, 2023.

Ò. Jordà. Estimation and inference of impulse responses by local projections. *American economic review*, 95(1):161–182, 2005.

L. Kilian and H. Lütkepohl. *Structural vector autoregressive analysis*. Cambridge University Press, 2017.

G. Koop, M. H. Pesaran, and S. M. Potter. Impulse response analysis in nonlinear multivariate models. *Journal of econometrics*, 74(1):119–147, 1996.

J. Krampe and E. Paparoditis. Sparsity concepts and estimation procedures for high-dimensional vector autoregressive models. *Journal of Time Series Analysis*, 42(5-6):554–579, 2021.

J. Krampe, E. Paparoditis, and C. Trenkler. Structural inference in sparse high-dimensional vector autoregressions. *Journal of Econometrics*, 234(1):276–300, 2023.

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

E. Lazarus, D. J. Lewis, and J. H. Stock. The size-power tradeoff in har inference. *Econometrica*, 89(5):2497–2516, 2021.

A. Lusompa. Local projections, autocorrelation, and efficiency. *Quantitative Economics*, 14 (4):1199–1220, 2023.

H. Lütkepohl. Testing for causation between two variables in higher-dimensional var models. In *Studies in applied econometrics*, pages 75–91. Springer, 1993.

H. Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

R. P. Masini, M. C. Medeiros, and E. F. Mendes. Regularized estimation of high-dimensional vector autoregressions with weakly dependent innovations. *Journal of Time Series Analysis*, 43(4):532–557, 2022.

M. C. Medeiros and E. F. Mendes. L1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics*, 191(1):255–271, 2016.

K. Miao, P. C. Phillips, and L. Su. High-dimensional vars with common factors. *Journal of Econometrics*, 233(1):155–183, 2023.

J. L. Montiel Olea and M. Plagborg-Møller. Local projection inference is simpler and more robust than you think. *Econometrica*, 89(4):1789–1823, 2021.

H. H. Pesaran and Y. Shin. Generalized impulse response analysis in linear multivariate models. *Economics letters*, 58(1):17–29, 1998.

A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

P. K. Salamaliki and I. A. Venetis. Transmission chains of economic uncertainty on macroeconomic activity: new empirical evidence. *Macroeconomic Dynamics*, 23(8):3355–3385, 2019.

S. Song and P. J. Bickel. Large vector auto regressions. *arXiv preprint arXiv:1106.3915*, 2011.

J. H. Stock and M. W. Watson. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of macroeconomics*, volume 2, pages 415–525. Elsevier, 2016.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3), 2014.

H. White. *Asymptotic theory for econometricians, revised edition*. Academic press, San Diego, Florida, 1999.

I. Wilms, S. Basu, J. Bien, and D. S. Matteson. Sparse identification and estimation of large-scale vector autoregressive moving averages. *Journal of the American Statistical Association*, pages 1–12, 2021.

K. C. Wong, Z. Li, and A. Tewari. Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, 48(2):1124–1142, 2020.

W. B. Wu and Y. N. Wu. High-dimensional linear models with dependent observations. *Preprint*, 2014.

K.-L. Xu and J. Guo. A new test for multiple predictive regression. *Journal of Financial Econometrics*, 22(1):119–156, 2024.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

A. Appendix

A.1. Proofs of results

This section collects the proofs of the theoretical results. For notation convenience, we will omit the '*re*' superscript in all regularized estimators. For example, $\hat{\mathbf{A}}$ will refer to $\hat{\mathbf{A}}^{(re)}$, the regularized estimator of the VAR matrix coefficient. Throughout this Appendix, C will denote a generic positive constant that may vary with different uses. We will also use the following abbreviations in what follows: T (triangle inequality), CS (Cauchy–Schwarz inequality), LIE (law of iterated expectations), and m.d.s. (martingale difference sequence). Moreover, we will apply the following matrix norm inequalities to any compatible matrices B_1, B_2, U, and V, and to any column vector x:

$$\left\|B_1'\right\|_{\max} = \left\|B_1\right\|_{\max}, \quad \left\|B_1'\right\|_{\infty} = \left\|B_1\right\|_{1}, \quad \left\|B_1 B_2\right\|_{\max} \leq \left\|B_1\right\|_{\infty}\left\|B_2\right\|_{\max}, \quad \left\|B_1 x\right\|_{l} \leq \|B\|_l \|x\|_l$$
$$\left\|B_1 B_2\right\|_{l} \leq \left\|B_1\right\|_{l}\left\|B_2\right\|_{l} \text{ for } l \in \{1,2,\infty\}, \text{ and } \left\|U B_1 V\right\|_{2} = \left\|B_1\right\|_{2} \text{ if } \|U\|_2 = \|V\|_2 = 1$$

Proof of Theorem 6.1. To obtain rate in part *(i)*, it worth notice that T implies $\left\|\hat{\Sigma}_u - \Sigma_u\right\|_{\max} \leq I_1 + 2I_2 + I_3$, where

$$I_3 := \left\|\frac{1}{n-p}\sum_{t=p+1}^{n} u_t u_t' - \Sigma_u\right\|_{\max} = O_p\left(\sqrt{\tilde{\nu}_n/n}\right) \quad \text{by Assumption 2}(iv),$$
$$I_1 = \left\|\frac{1}{n-p}\sum_{t=p+1}^{n} \left(\hat{u}_t - u_t\right)\left(\hat{u}_t - u_t\right)\right\|_{\max} \quad \text{and} \quad I_2 = \left\|\frac{1}{n-p}\sum_{t=p+1}^{n} \left(\hat{u}_t - u_t\right)u_t'\right\|_{\max}.$$

Note that $\hat{u}_t - u_t = J(\mathbf{A}-\hat{\mathbf{A}})W_{t-1}$ by $w_t = J\mathbf{A}W_{t-1} + u_t$ and $\hat{u}_t = w_t - J\hat{\mathbf{A}}W_{t-1}$. Also, by Lemma A.1 below applied to suitable filters, we have

$$\left\|\frac{1}{n-p}\sum_{t=p+1}^{n} W_{t-1}u_t'\right\|_{\max} = O_p\left(\sqrt{\tilde{\nu}_n/n}\right) \quad \text{and} \quad \left\|\frac{1}{n-p}\sum_{t=p+1}^{n} W_{t-1}W_{t-1}'\right\|_{\max} = O_p(1).$$

It follows by Assumption 2*(iii)*, by $\|AB\|_{\max} \leq \|A\|_\infty \|B\|_{\max}$ and $\|AB\|_\infty \leqslant \|A\|_\infty \|B\|_\infty$ for all compatible matrices A and B, and by $\|J\|_\infty = 1$ that

$$I_1 := \left\|J(\mathbf{A}-\hat{\mathbf{A}})\frac{1}{n-p}\sum_{t=p+1}^{n} W_{t-1}W_{t-1}'(\mathbf{A}-\hat{\mathbf{A}})'J'\right\|_{\max}$$
$$\leqslant \|J\|_\infty^2 \|\hat{\mathbf{A}}-\mathbf{A}\|_\infty^2 \left\|\frac{1}{n-p}\sum_{t=p+1}^{n} W_{t-1}W_{t-1}'\right\|_{\max} = O_p\left(\|\hat{\mathbf{A}}-\mathbf{A}\|_\infty^2\right)$$

and

$$I_2 := \left\| \frac{1}{n-p} \sum_{t=p+1}^{n} J(\hat{\mathbf{A}}-\mathbf{A}) W_{t-1} u_t' \right\|_{\max}$$
$$\leqslant \|\hat{\mathbf{A}}-\mathbf{A}\|_\infty \left\| \frac{1}{n-p} \sum_{t=p+1}^{n} W_{t-1} u_t' \right\|_{\max} = O_p\left(\|\hat{\mathbf{A}}-\mathbf{A}\|_\infty \sqrt{\tilde{\nu}_n/n} \right).$$

Therefore,

$$\left\|\hat{\Sigma}_u - \Sigma_u\right\|_{\max} = O_p\left(\|\hat{\mathbf{A}}-\mathbf{A}\|_\infty^2 + \sqrt{\tilde{\nu}_n/n}\right) = O_p\left(k_A^3\left(\nu_n/n\right)^{1-\mu} + \sqrt{\tilde{\nu}_n/n}\right).$$

The result (i) follows from the following equivalence inequality of norm matrix which hold fo any r-by-s matrix B: $\|B\|_\infty \leq d\|B\|_{\max}$.

The proofs of parts *(ii)* and *(iii)* require first deriving the orders of some auxiliary terms. First note that the stability condition, see Assumption 2*(ii)*, implies $\sum_{m=0}^{\infty} \|\mathbf{A}^m\|_\infty = O\left(k_A/(1-\varphi)\right) = O(k_A)$. Let $j \in \mathbb{N}$, $j \geq 1$, with j sample-independent. By the Binomial Theorem, $\hat{\mathbf{A}}^j = (\hat{\mathbf{A}}-\mathbf{A}+\mathbf{A})^j = \mathbf{A}^j + \sum_{i=0}^{j-1}\binom{j}{i}(\hat{\mathbf{A}}-\mathbf{A})^{j-i}\mathbf{A}^j$ Therefore, by $\|J\|_\infty = 1$ and T,

$$\left\|\hat{\Upsilon}_j - \Upsilon_j\right\|_\infty \leq \left\|\hat{\mathbf{A}}^j - \mathbf{A}^j\right\|_\infty \leq C\sum_{i=0}^{d-1}\|\hat{\mathbf{A}}-\mathbf{A}\|_\infty^{j-i}\left\|\mathbf{A}^j\right\|_\infty \leq O_p\left(\|\hat{\mathbf{A}}-\mathbf{A}\|_\infty\right)\sum_{i=0}^{\infty}\left\|\mathbf{A}^j\right\|_\infty$$
$$= O_p\left(k_A^{2.5}\left(\nu_n/n\right)^{(1-\mu)/2}\right)$$

Also, it is obvious that for all $m \in \mathbb{N}$, $m \geq 1$ $\hat{\mathbf{A}}^m - \mathbf{A}^m = (\hat{\mathbf{A}}-\mathbf{A})\left(\hat{\mathbf{A}}^{m-1} - \mathbf{A}^{m-1}\right) + (\mathbf{A}-\hat{\mathbf{A}})\mathbf{A}^{m-1}$. It follows by this recursive formula that $\hat{\mathbf{A}}^m - \mathbf{A}^m = \sum_{s=0}^{m-1}[(\hat{\mathbf{A}}-\mathbf{A})+\mathbf{A}]^s(\hat{\mathbf{A}}-\mathbf{A})\mathbf{A}^{m-1-s}$. I follows by T that,

$$\sum_{m=0}^{\infty}\left\|\hat{\Upsilon}_m - \Upsilon_m\right\|_\infty \leq \sum_{m=0}^{\infty}\left\|\hat{\mathbf{A}}^m - \mathbf{A}^m\right\|_\infty \leq \sum_{m=0}^{\infty}\sum_{s=0}^{m-1}\left\|\hat{\mathbf{A}}^s\right\|_\infty\|\hat{\mathbf{A}}-\mathbf{A}\|_\infty\left\|\mathbf{A}^{m-1-s}\right\|_\infty$$
$$\leq \|\hat{\mathbf{A}}-\mathbf{A}\|_\infty\sum_{m,s=0}^{\infty}\left\|\hat{\mathbf{A}}^s\right\|\|\mathbf{A}^m\|_\infty = O_p\left(k_A^2\|\hat{\mathbf{A}}-A\|_\infty\right) = O_p\left(k_A^{3.5}\left(\nu_n/n\right)^{(1-\mu)/2}\right).$$

Using similar argument yields $\sum_{m=0}^{\infty}\left\|(\mathbf{A}')^m\right\|_\infty = O(k_A)$, $\left\|\hat{\Upsilon}_j' - \Upsilon_j'\right\|_\infty = O_p\left(k_A^{2.5}\left(\nu_n/n\right)^{(1-\mu)/2}\right)$, and $\sum_{m=0}^{\infty}\left\|\hat{\Upsilon}_m' - \Upsilon_m'\right\|_\infty = O_p\left(k_A^{3.5}\left(\nu_n/n\right)^{(1-\mu)/2}\right)$.

Given the results above, the proof of part (iii) is straightforward. First, it is worth noting that $\Sigma_{UW} = \sum_{j=0}^{p-1}\left(\tilde{e}_{p(j+1)} \otimes I_d\right)\Sigma_u\Upsilon_j'$, so that by T and $\left\|\tilde{e}_{p(j+1)} \otimes I_d\right\|_\infty = 1$,

$$\left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_\infty \leqslant \sum_{j=0}^{p-1}\left\|\hat{\Sigma}_u\hat{\Upsilon}_j' - \Sigma_u\Upsilon_j'\right\|_\infty. \tag{A.1}$$

Moreover, by Assumption 2*(v)*

$$\left\|\Sigma_u\right\|_\infty = \max_{1\le i\le d}\sum_{j=1}^{d}\left|E\left[e_i' u_t u_t' e_j\right]\right| \le \max_{1\le i\le d}\sum_{j=1}^{d}\left\{E\left[\left|e_i' u_t\right|^2\right]E\left[\left|e_j' u_t\right|^2\right]\right\}^{1/2} \le Cd.$$

Let $\tilde{a}_n = d\left[k_A^3\left(\nu_n/n\right)^{1-\mu} + \sqrt{\tilde{\nu}_n/n}\right]$ so that $\left\|\hat{\Sigma}_u - \Sigma_u\right\|_\infty = O_p(\tilde{a}_n)$. Since p is finite and sample-independent, and the terms in the summation on the right-hand side of Eq.(A.1) are of the same order for all j, then by T,

$$\begin{aligned}
\left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_\infty &\leqslant \left\|\left(\hat{\Upsilon}_j' - \Upsilon_j'\right)\left(\hat{\Sigma}_u - \Sigma_u\right)\right\|_\infty + \left\|\left(\hat{\Upsilon}_j' - \Upsilon_j'\right)\Sigma_u\right\|_\infty + \left\|\Upsilon_j'\left(\hat{\Sigma}_u - \Sigma_u\right)\right\|_\infty \\
&= O_p\left(\tilde{a}_n\left\|\hat{\Upsilon}_j' - \Upsilon_j'\right\|_\infty + d\left\|\hat{\Upsilon}_j' - \Upsilon_j'\right\|_\infty + k_A\tilde{a}_n\right) \\
&= O_p\left(dk_A\left\{k_A^{1.5}\left(\nu_n/n\right)^{(1-\mu)/2} + k_A^3\left(\nu_n/n\right)^{1-\mu} + \sqrt{\tilde{\nu}_n/n}\right\}\right) \\
&= O_p\left(dk_A\left\{k_A^{1.5}\left(\nu_n/n\right)^{(1-\mu)/2} + \sqrt{\tilde{\nu}_n/n}\right\}\right).
\end{aligned}$$

Now let us consider the proof of part (*ii*). By arguments above and $\left\|\Sigma_u\right\|_\infty = O(1)$, it follows by T that,

$$\left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty = \left\|\sum_{m=0}^{\infty}\hat{\Upsilon}_m\hat{\Sigma}_u\hat{\Upsilon}_m' - \sum_{m=0}^{\infty}\Upsilon_m\Sigma_u\Upsilon_m'\right\|_\infty \le \gamma_1+\gamma_2+\gamma_3+\gamma_4+\gamma_5+\gamma_6+\gamma_7, \quad \text{(A.2)}$$

where,

$$\begin{aligned}
\gamma_1 &:= \left\|\sum_{m=0}^{\infty}\left(\hat{\Upsilon}_m - \Upsilon_m\right)\left(\hat{\Sigma}_u - \Sigma_u\right)\left(\hat{\Upsilon}_m' - \Upsilon_m'\right)\right\|_\infty \\
&\le \left\|\hat{\Sigma}_u - \Sigma_u\right\|_\infty\left(\sum_{m=0}^{\infty}\left\|\hat{\Upsilon}_m - \Upsilon_m\right\|_\infty\right)\left(\sum_{m=0}^{\infty}\left\|\hat{\Upsilon}_m' - \Upsilon_m'\right\|_\infty\right) = O_p\left(\tilde{a}_n k_A^7\left(\nu_n/n\right)^{1-\mu}\right), \\
\gamma_2 &:= \left\|\sum_{m=0}^{\infty}\left(\hat{\Upsilon}_m - \Upsilon_m\right)\left(\hat{\Sigma}_u - \Sigma_u\right)\Upsilon_m'\right\|_\infty \\
&\leqslant \left\|\hat{\Sigma}_u - \Sigma_u\right\|_\infty\left(\sum_{m=0}^{\infty}\left\|\hat{\Upsilon}_m - \Upsilon_m\right\|\right)\left(\sum_{m=0}^{\infty}\left\|\Upsilon_m'\right\|_\infty\right) = O_p\left(\tilde{a}_n k_A^{4.5}\left(\nu_n/n\right)^{(1-\mu)/2}\right), \\
\gamma_3 &:= \left\|\sum_{m=0}^{\infty}\left(\hat{\Upsilon}_m - \Upsilon_m\right)\Sigma_u\left(\hat{\Upsilon}_m' - \Upsilon_m'\right)\right\|_\infty = O_p\left(dk_A^7\left(\nu_n/n\right)^{1-\mu}\right), \\
\gamma_4 &:= \left\|\sum_{m=0}^{\infty}\left(\hat{\Upsilon}_m - \Upsilon_m\right)\Sigma_u\Upsilon_m'\right\|_\infty = O_p\left(dk_A^{4.5}\left(\nu_n/n\right)^{(1-\mu)/2}\right),
\end{aligned}$$

$$\gamma_5 := \left\| \sum_{m=0}^{\infty} \Upsilon_m \left(\hat{\Sigma}_u - \Sigma_u\right)\left(\hat{\Upsilon}_m' - \Upsilon_m'\right) \right\|_\infty = O_p\left(\tilde{a}_n k_A^{4.5} \left(\nu_n/n\right)^{(1-\mu)/2}\right),$$

$$\gamma_6 := \left\| \sum_{m=0}^{\infty} \Upsilon_m \left(\hat{\Sigma}_u - \Sigma_u\right)\Upsilon_m' \right\|_\infty = O_p\left(\tilde{a}_n k_A^2\right),$$

$$\gamma_7 := \left\| \sum_{m=0}^{\infty} \Upsilon_m \Sigma_u \left(\hat{\Upsilon}_m' - \Upsilon_m'\right) \right\|_\infty = O_p\left(d k_A^{4.5} \left(\nu_n/n\right)^{(1-\mu)/2}\right).$$

Plugging derived rates into (A.2) and dropping higher-order terms yiels,

$$\begin{aligned}
\left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty &= O_p\left(d k_A^{4.5}\left(\nu_n/n\right)^{(1-\mu)/2} + \tilde{a}_n k_A^2\right) \\
&= O_p\left(d k_A^2 \left\{k_A^{2.5}\left(\nu_n/n\right)^{(1-\mu)/2} + k_A^3\left(\nu_n/n\right)^{1-\mu} + \sqrt{\tilde{\nu}_n/n}\right\}\right) \\
&= O_p\left(d k_A^2 \left\{k_A^{2.5}\left(\nu_n/n\right)^{(1-\mu)/2} + \sqrt{\tilde{\nu}_n/n}\right\}\right),
\end{aligned}$$

giving the result for part *(ii)*.

□

Lemma A.1 (Lemma A.2 of Krampe et al. (2023)).
Let $\left\{\Phi_j^{(k)}, j = 0, 1, \ldots\right\}, k = 1, 2$, *be linear filters with* $\sum_{j=0}^{\infty} \left\|\Phi_j^{(k)}\right\|_2 = O(1), k = 1,\ 2$ *. Then under Assumption 2(iv)*

$$\left\| 1/\sqrt{n} \sum_{t=1}^{n} \sum_{j,k=0}^{\infty} \Phi_j^{(1)} \left(u_{t-j}u_{t-k}' - \mathbf{1}(j=k)\Sigma_u\right)\left(\Phi_k^{(2)}\right)' \right\|_{\max} = O\left(\sqrt{\tilde{\nu}_n}\right)$$

Proof of Lemma A.1. See Appendix A of Krampe et al. (2023). □

Lemma A.2.
Under Assumptions 2(ii), and (iv)-(vi), it holds true that:

(a) $\left\| \frac{1}{n} \sum_{t=p}^{n-h} W_t W_t' - \Sigma_W \right\|_{max} = O_p\left(\sqrt{\hat{\nu}_n/n}\right)$ *and* $\left\| \frac{1}{n} \sum_{t=p}^{n-h} W_t W_t' \right\|_{max} = O_p(1);$

(b) $\left\| \frac{1}{n} \sum_{t=p}^{n-h} \Sigma_W^{-1} W_t W_t' - I_{dp} \right\|_{\max} = O_p\left(\sqrt{\tilde{\nu}_n/n}\right);$

(c) $\left\| \frac{1}{n} \sum_{t=p}^{n-h} R_1 \Sigma_W^{-1} W_t W_t' R_1' - I_p \right\|_{\max} = O_p\left(\sqrt{\tilde{\nu}_n/n}\right)$ *and* $\left\| \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} R_1 \Sigma_W^{-1} W_t W_t' R_2' \right\|_{\max} = O_p\left(\sqrt{\tilde{\nu}_n}\right);$

(d) $\left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{n-h} W_{t-1} u_t' \right\|_{max} = O_p\left(\sqrt{\tilde{\nu}_n}\right)$ *and* $\left\| \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} \Sigma_W^{-1} W_t e_{t,h} \right\|_{\max} = O_p\left(\sqrt{\tilde{\nu}_n}\right).$

Proof of Lemma A.2. First note that the VAR(p) underlying equation has a companion representation in terms of VAR(1) of the form $W_t = \mathbf{A}W_{t-1} + J'u_t$. This implies, by the stability condition (see Assumption 2(ii)), the following VAR(∞) representation:

$$W_t = \sum_{j=0}^{\infty} \Upsilon_j u_{t-j} \text{ with } \Upsilon_j = \mathbf{A}^j J' \text{ for } j = 0, 1, \ldots, \infty.$$

Note that the same stability condition ensures that the filter $\{\Upsilon_j, j = 0, 1, \ldots, \infty\}$ satisfies the condition $\sum_{j=0}^{\infty} \left\|\Upsilon_j\right\|_2 = O(1/(1-\varphi)) = O(1)$. Also, the VAR($\infty$) representation implies that

$$W_t W_t' = \sum_{j,k=0}^{\infty} \Upsilon_j u_{t-j} u_{t-k}' \Upsilon_k',$$

and

$$\sum_{j,k=0}^{\infty} \mathbf{1}(j=k)\Upsilon_j \Sigma_u \Upsilon_k' = \sum_{j=0}^{\infty} \mathbf{A}^j J' \Sigma_u J \left(\mathbf{A}'\right)^j = \Sigma_W.$$

Then, Lemma A.1 applied to the filters $\Phi_j^{(1)} = \Phi_j^{(2)} = \Upsilon_j$, $j = 0, 1, \ldots, \infty$ leads to the first result in part (a) of the Lemma. Also, Assumption 2(v) implies that $\left\|\Sigma_W\right\|_{\max} = O(1)$ and the second result in part (a) follows by T.

For part (b), consider the following filters:

$$\Phi_j^{(1)} = \Sigma_W^{-1}\Upsilon_j \quad \text{and} \quad \Phi_j^{(2)} = \Upsilon_j \quad \text{for} \quad j = 0, 1, \ldots, \infty.$$

It is obvious that $\sum_{i=0}^{\infty} \left\|\Phi_j^{(1)}\right\|_2 = O(1) = \sum_{k=0}^{\infty} \left\|\Phi_k^{(2)}\right\|_2$, where the second equality follows from the stability condition as mentioned above and the first equality follows from the same condition and the fact that $\left\|\Sigma_W^{-1}\right\|_2 = O(1)$ (see Assumption 2(vi)). The result follows from Lemma A.1 applied to the filters $\left\{\Phi_j^{(k)}, j = 0, 1, \ldots\right\}, k = 1, 2$.

Given part (b), the results in part (c) are straightforward. They follows from the fact that $R_1 R_1' = I_p$ and $R_1 R_2' = O_{p\times(d-1)p}$.

Finally, the first result in part (d) follows from LemmaA.1 applied to the filters defined by

$$\begin{cases} \Phi_0^{(1)} = 0 \quad \text{and} \quad \Phi_j^{(1)} = \Upsilon_{j-1} \text{ for } j \geqslant 1 \\ \Phi_0^{(2)} = I_p \quad \text{and} \quad \Phi_k^{(2)} = 0 \text{ for } k \geqslant 1 \end{cases},$$

and the second result is obtained by applying the same lemma to the following filters:

$$\Phi_j^{(1)} = \begin{cases} 0 & \text{if} \quad j < h \\ \Sigma_W^{-1}\Upsilon_j & \text{if} \quad j \geqslant h \end{cases} \quad \text{and} \quad \Phi_k^{(2)} = \begin{cases} e_y' J \mathbf{A}^k J' & \text{if} \quad k < h \\ 0 & \text{if} \quad k \geqslant h, \end{cases}$$

where e_y is the d-dimensional unit vector such that $e_{t,h} = \left(u_t^{(h)}\right)' e_y$, meaning that e_y contains 1 at the position of y_t in the vector w_t.

Note that all these filters satisfy the condition required for applying LemmaA.1 due to the stability condition and the fact that $\|J\|_2 = 1$ and $\left\|\Sigma_W^{-1}\right\|_2 = O(1)$ (see Assumption 2(vi)).

□

Lemma A.3.
Let

$$\widehat{DN} := \frac{1}{n}\sum_{t=p}^{n-h} R_1 \hat{\Sigma}_W^{-1} W_t W_t' R_1' \quad and \quad DN := \frac{1}{n}\sum_{t=p}^{n-h} R_1 \Sigma_W^{-1} W_t W_t' R_1'.$$

Under Assumptions 2(ii), and (iv)-(vi), it holds true that:

$(a)\ \left\|\widehat{DN}^{-1} - DN^{-1}\right\|_{max} = O_p\left(k_W \left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty\right);$

$(b)\ \left\|\widehat{DN}^{-1} - I_p\right\|_{max} = O_p\left(\sqrt{\tilde{\nu}_n/n} + k_W \left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty\right).$

Proof of Lemma A.3. First of all, it worth noting that part (b) of Lemma A.2 implies, by T, that

$$\left\|\frac{1}{n}\sum_{t=p}^{n-h}\Sigma_W^{-1}W_tW_t'\right\|_{\max} \leq \left\|\frac{1}{n}\sum_{t=1}^{n-h}\Sigma_W^{-1}W_tW_t' - I_{dp}\right\|_{\max} + \left\|I_{dp}\right\|_{\max} = O_p\left(\sqrt{\tilde{\nu}_n/n}\right) + 1 = O_p(1).$$

It then follows from $\left\|R_1\right\|_\infty = 1$, $\left\|\hat{\Sigma}_W^{-1}\right\|_\infty = O_p(k_W)$ (see Assumption 2(vi)) and $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$ that

$$\begin{aligned}\|\widehat{DN} - DN\|_{\max} &= \left\|\frac{1}{n}\sum_{t=p}^{n-h} R_1\hat{\Sigma}_W^{-1}\left(\Sigma_W - \hat{\Sigma}_W\right)\Sigma_W^{-1}W_tW_t'R_1\right\|_{\max} \\ &\leq \left\|R_1\right\|_\infty^2 \left\|\hat{\Sigma}_W^{-1}\right\|_\infty \left\|\Sigma_W - \hat{\Sigma}_W\right\|_\infty \left\|\frac{1}{n}\sum_{t=p}^{n-h}\Sigma_W^{-1}W_tW_t'\right\|_{\max} = O_p\left(k_W\left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty\right).\end{aligned}$$

Also, it is easy to verify that $\|DN^{-1}\|_1 = O_p(1)$ and $\|\widehat{DN}^{-1}\|_\infty = O_p(1)$, so

$$\begin{aligned}\left\|\widehat{DN}^{-1} - DN^{-1}\right\|_{\max} &= \left\|\widehat{DN}^{-1}(DN - \widehat{DN})DN^{-1}\right\|_{\max} \\ &\leq \left\|\widehat{DN}^{-1}\right\|_\infty \|DN - \widehat{DN}\|_{\max} \left\|DN^{-1}\right\|_1 = O_p\left(k_W\left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty\right),\end{aligned}$$

giving result in part (a).

To obtain the result in part (b), first note that $\left\|DN - I_p\right\|_{\max} = O_p\left(\sqrt{\tilde{\nu}_n/n}\right)$ by part (c) of Lemma A.2. It then follows by T and result we have just derived in part (a) that $\left\|\widehat{DN} - I_p\right\|_{\max} = O_p\left(\sqrt{\tilde{\nu}_n/n} + k_W\left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty\right)$. The result is obtained by noting that $\|\widehat{DN}^{-1}\|_\infty = O_p(1)$ and $\widehat{DN}^{-1} - I_p = \widehat{DN}^{-1}\left(I_p - \widehat{DN}\right)$. □

Lemma A.4.

If Assumption 2 is satisfied, then for any vector $v \in \mathbb{R}^p$ such that $\|v\|_1 = 1$,

$$\sqrt{n}v'\left(\hat{\beta}_{1,h}^{(de-LS)} - \beta_{1,h}\right) = v'\left(E\left[W_{1,t}^{\perp}W_{1,t}'\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} W_{1,t}^{\perp}e_{t,h}\right)$$
$$+ O_p\left(\tilde{\nu}_n/\sqrt{n} + \left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty k_W\sqrt{\tilde{\nu}_n} + \left\|\hat{\beta}_{2,h} - \beta_{2,h}\right\|_\infty\left(\sqrt{\tilde{\nu}_n} + \left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty k_W\sqrt{\tilde{\nu}_n}\right)\right) \tag{A.3}$$

Proof of Lemma A.4. By the definition of the de-LS estimator,

$$\begin{aligned}\sqrt{n}v'\left(\hat{\beta}_{1,h}^{(de-LS)} - \beta_{1,h}\right) &= v'\left(\frac{1}{n}\sum_{t=p}^{n-h}\hat{W}_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\hat{W}_{1,t}^{\perp}e_{t,h}\right) \\ &\quad + v'\left(\frac{1}{n}\sum_{t=p}^{n-h}\hat{W}_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\hat{W}_{1,t}^{\perp}W_{2,t}'\left(\beta_{2,h} - \hat{\beta}_{2,h}\right)\right) \\ &= v'\left(E\left[W_{1,t}^{\perp}W_{1,t}'\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}W_{1,t}^{\perp}e_{t,h}\right) + \Lambda_0 + \Lambda_1 + \Lambda_2,\end{aligned} \tag{A.4}$$

where,

$$\Lambda_0 := v'\left\{\left(\frac{1}{n}\sum_{t=p}^{n-h}W_{1,t}^{\perp}W_{1,t}'\right)^{-1} - \left(E\left[W_{1,t}^{\perp}W_{1,t}'\right]\right)^{-1}\right\}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}W_{1,t}^{\perp}e_{t,h}\right) \tag{A.5}$$

$$\Lambda_1 := v'\left(\frac{1}{n}\sum_{t=p}^{n-h}\hat{W}_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\hat{W}_{1,t}^{\perp}e_{t,h}\right) - v'\left(\frac{1}{n}\sum_{t=p}^{n-h}W_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}W_{1,t}^{\perp}e_{t,h}\right)$$

$$\Lambda_2 := v'\left(\frac{1}{n}\sum_{t=p}^{n-h}\hat{W}_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\hat{W}_{1,t}^{\perp}W_{2,t}'\left(\beta_{2,h} - \hat{\beta}_{2,h}\right)\right).$$

Using the fact that

$$W_{1,t}^{\perp} = \left(R_1\Sigma_W^{-1}R_1'\right)^{-1}R_1\Sigma_W^{-1}W_t \quad \text{and} \quad \hat{W}_{1,t}^{\perp} = \left(R_1\hat{\Sigma}_W^{-1}R_1'\right)^{-1}R_1\hat{\Sigma}_W^{-1}W_t,$$

Λ_1 can be rewritten as

$$\Lambda_1 = \frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}v'\left(\widehat{DN}^{-1}R_1\hat{\Sigma}_W^{-1} - DN^{-1}R_1\Sigma_W\right)W_t e_{t,h} = \Lambda_{11} + \Lambda_{12} + \Lambda_{13},$$

where $\widehat{DN}$ and DN are defined as in the statement of Lemma A.3 and

$$\Lambda_{11} := \frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}v'\left(\widehat{DN}^{-1} - DN^{-1}\right)R_1\Sigma_W^{-1}W_t e_{t,h}$$

$$\Lambda_{12} := \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' DN^{-1} R_1 \left(\hat{\Sigma}_W^{-1} - \Sigma_W^{-1}\right) W_t e_{t,h} \tag{A.6}$$

$$\Lambda_{13} := \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(\widehat{DN}^{-1} - DN^{-1}\right) R_1 \left(\hat{\Sigma}_W^{-1} - \Sigma_W^{-1}\right) W_t e_{t,h}.$$

Also,

$$\Lambda_2 = \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \widehat{DN}^{-1} R_1 \hat{\Sigma}_W^{-1} W_t W_t' R_2' \left(\beta_{2,h} - \hat{\beta}_{2,h}\right) = \left(\Lambda_{21} + \Lambda_{22} + \Lambda_{23} + \Lambda_{24}\right)\left(\beta_{2,h} - \hat{\beta}_{2,h}\right),$$

where

$$\Lambda_{21} := \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' R_1 \Sigma_W^{-1} W_t W_t' R_2'$$

$$\Lambda_{22} := \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' R_1 \left(\hat{\Sigma}_W^{-1} - \Sigma_W^{-1}\right) W_t W_t' R_2'$$

$$\Lambda_{23} := \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(\widehat{DN}^{-1} - I_p\right) R_1 \Sigma_W^{-1} W_t W_t' R_2' \tag{A.7}$$

$$\Lambda_{24} := \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(\widehat{DN}^{-1} - I_p\right) R_1 \left(\hat{\Sigma}_W^{-1} - \Sigma_W^{-1}\right) W_t W_t' R_2'.$$

It remains to show that the terms in (A.5), (A.6) and (A.7) are of the specified orders so that the result follows. By Lemma A.2, Lemma A.3 and the fact that $\left\|DN^{-1}\right\|_{\max} = O_p(1)$ and $\left\|\hat{\Sigma}_W^{-1}\right\|_{\infty} = O_p\left(k_W\right)$, we have

$$\begin{aligned}
\left|\Lambda_0\right| &= \left|\frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(DN^{-1} - I_p\right) R_1 \Sigma_W^{-1} W_t e_{t,h}\right| \\
&\leq \|v\|_1 \left\|DN^{-1} - I_p\right\|_{\max} \left\|\frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} R_1 \Sigma_W^{-1} W_t e_{t,h}\right\|_{\max} = O_p\left(\tilde{v}_n / \sqrt{n}\right);
\end{aligned}$$

$$\begin{aligned}
&\left|\Lambda_{11}\right| \leq \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} \sum_{r,j=1}^{p} \left|v_r\right| \left|e_r' \left(\widehat{DN}^{-1} - DN^{-1}\right) e_j\right| \left|e_j' R_1 \Sigma_W^{-1} W_t e_{t,h}\right| \\
&\leqslant p\|v\|_1 \left\|\widehat{DN}^{-1} - DN^{-1}\right\|_{\max} \left\|\frac{1}{\sqrt{n}} \sum_{t=P}^{n-h} R_1 \Sigma_W^{-1} W_t e_{t,h}\right\|_{\max} = O_p\left(\left\|\hat{\Sigma}_W - \Sigma_W\right\|_{\infty} k_W \sqrt{\tilde{v}_n}\right);
\end{aligned}$$

$$
\begin{aligned}
\left|\Lambda_{12}\right| &\leq \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} \sum_{r,j=1}^{p} |\nu_r| \left|e_r' DN^{-1} e_j\right| \left|e_j' R_1 \left(\hat{\Sigma}_W^{-1} - \Sigma_W^{-1}\right) W_t e_{t,h}\right| \\
&\leq p\|\nu\|_1 \left\|DN^{-1}\right\|_{\max} \left\|\hat{\Sigma}_W^{-1}\right\|_\infty \left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty \left\| \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} \Sigma_W^{-1} W_t e_{t,h} \right\|_{\max} \\
&= O_p\left(\left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty k_W \sqrt{\tilde{\nu}_n}\right);
\end{aligned}
$$

$$
\begin{aligned}
\left|\Lambda_{13}\right| &\leq p\|\nu\|_1 \left\|\widehat{DN}^{-1} - DN^{-1}\right\|_{\max} \left\|\hat{\Sigma}_W^{-1}\right\|_\infty \left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty \left\| \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} \Sigma_W^{-1} W_t e_{t,h} \right\|_{\max} \\
&= O_p\left(\left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty^2 k_W^2 \sqrt{\tilde{\nu}_n}\right);
\end{aligned}
$$

$$
\begin{aligned}
\left|\Lambda_{21}\left(\beta_{2,h} - \hat{\beta}_{2,h}\right)\right| &\leqslant \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} \sum_{r=1}^{p} |\nu|_r \left|e_r' R_1 \Sigma_W^{-1} W_t W_t' R_2' \left(\beta_{2,h} - \hat{\beta}_{2,h}\right)\right| \\
&\leqslant \|\nu\|_1 \left\|\beta_{2,h} - \hat{\beta}_{2,h}\right\|_\infty \left\| \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} R_1 \Sigma_W^{-1} W_t W_t' R_2' \right\|_{\max} \\
&= O_p\left(\left\|\hat{\beta}_{2,h} - \beta_{2,h}\right\|_\infty \sqrt{\tilde{\nu}_n}\right);
\end{aligned}
$$

$$
\begin{aligned}
\left|\Lambda_{22}\right| &\leq \|\nu\|_1 \left\|\hat{\Sigma}_W^{-1}\right\|_\infty \left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty \left\| \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} R_1 \Sigma_W^{-1} W_t W_t' R_2' \right\|_{\max} \\
&= O_p\left(\left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty k_W \sqrt{\tilde{\nu}_n}\right);
\end{aligned}
$$

$$
\begin{aligned}
&\left\|\Lambda_{23}\right\| \leq \|\nu\|_1 \left\|\widehat{DN}^{-1} - I_p\right\|_{\max} \left\| \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} R_1 \Sigma_W^{-1} W_t W_t' R_2' \right\|_{\max} \\
&= O_p\left(\tilde{\nu}_n / \sqrt{n} + \left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty k_W \sqrt{\tilde{\nu}_n}\right);
\end{aligned}
$$

$$
\begin{aligned}
\left|\Lambda_{24}\right| &\leq \|\nu\|_1 \left\|\hat{\Sigma}_W^{-1}\right\|_\infty \left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty \left\|\widehat{DN}^{-1} - \Sigma_p\right\|_{\max} \left\| \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} R_2 \Sigma_W^{-1} W_t W_t' R_2' \right\|_{\max} \\
&= O_p\left(\left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty k_W \left(\tilde{\nu}_n / \sqrt{n} + \left\|\hat{\Sigma}_W - \Sigma_W\right\|_\infty k_W \sqrt{\tilde{\nu}_n}\right)\right).
\end{aligned}
$$

By substituting the derived rates into Equation (A.4) and neglecting the higher-order terms, we obtain the result. $\square$

Lemma A.5.
Let

$$\widehat{CN} := \frac{1}{n}\sum_{t=p}^{n-h} R_1 \hat{\Sigma}_{UW}^{-1} \hat{U}_t W_t' R_1' \quad \textit{and} \quad CN := \frac{1}{n}\sum_{t=p}^{n-h} R_1 \Sigma_{UW}^{-1} U_t W_t' R_1'.$$

Under Assumptions 2(ii), and (iv)-(vi), it holds true that:

$$(a) \left\| \frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} \Sigma_{UW}^{-1} U_t e_{th} \right\|_{max} = O_p\left(\sqrt{\tilde{\nu}_n}\right) \quad \textit{and} \left\| \frac{1}{n}\sum_{t=p}^{n-h} \Sigma_{UW}^{-1} U_t W_t' - I_{dp} \right\|_{max} = O_p\left(\sqrt{\tilde{\nu}_n / n}\right);$$

$$(b) \left\| \widehat{CN}^{-1} - CN^{-1} \right\|_{max} = O_p\left(\|\hat{\mathbf{A}} - \mathbf{A}\|_\infty k_{UW} + \left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_\infty k_{UW} \right);$$

$$(c) \left\| \widehat{CN}^{-1} - I_p \right\|_{max} = O_p\left(\sqrt{\hat{\nu}_n / n} + \|\hat{\mathbf{A}} - \mathbf{A}\|_\infty k_{UW} + \left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_\infty k_{UW} \right).$$

Proof of Lemma A.5. First recall that the VAR(∞) representation of the VAR model, under the stability condition, allows to write $W_t = \sum_{k=0}^{\infty} \Upsilon_j u_{t-k}$. Also, let $\tilde{e}_{pj}$, $j = 1, \dots, p$ denote the d-dimensional unit vectors, where $\tilde{e}_{pj}$ containts 1 at the j$^{\text{th}}$ position and 0 elsewhere. Then,

$$U_t := \left(u_t', u_{t-1}', \cdots, u_{t-p+1}'\right)' = \sum_{j=0}^{p-1} \left(\tilde{e}_{p(j+1)} \otimes I_d\right) u_{t-j},$$

and therefore

$$\Sigma_{UW} := E\left[U_t W_t'\right] = \sum_{j=0}^{p-1} \left(\tilde{e}_{p(j+1)} \otimes I_d\right) \Sigma_u \Upsilon_j'.$$

To derive the second result in part (a), consider the following filters:

$$\Phi_j^{(1)} = \begin{cases} \Sigma_{UW}^{-1}\left(\tilde{e}_{p(j+1)} \otimes I_d\right) & \text{if} \quad j < p \\ 0 & \text{otherwise} \end{cases} \quad \text{and } \Phi_k^{(2)} = \Upsilon_k$$

It is obvious that $\sum_{i=0}^{\infty} \left\|\Phi_j^{(1)}\right\|_2 = O(1) = \sum_{k=0}^{\infty} \left\|\Phi_k^{(2)}\right\|_2$, where the second equality follows from the stability condition and the first equality follows from the fact that $\left\|\Sigma_{UW}^{-1}\right\|_2 = O(1)$ (see Assumption 2(vi)). The result follows from Lemma A.1 applied to the filters $\left\{\Phi_j^{(k)}, j = 0, 1, \dots\right\}, k = 1, 2$. The first result in part (a) follows from the same Lemma A.1 applied to the

$$\Phi_j^{(1)} = \begin{cases} 0 & \text{if } j < h \text{ and } j \geqslant p + h - 1 \\ \Sigma_{UW}^{-1}\left(\tilde{e}_{p(j-h+1)} \otimes I_d\right) & \text{otherwise} \end{cases} \quad \text{and } \Phi_k^{(2)} = \begin{cases} e_y' J \mathbf{A}^k J' & \text{if } k < h \\ 0 & \text{otherwise} \end{cases}$$

and given the fact that $\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} \Sigma_{UW}^{-1} U_t e_{t,h} = \frac{1}{\sqrt{n}}\sum_{t=p+h}^{n} \Sigma_{UW}^{-1} U_{t-h} e_{t-h,h}$, where

$$U_{t-h} = \sum_{j=h}^{p+h-1} \left(\tilde{e}_{p(j-h+1)} \otimes I_d\right) u_{t-j} \quad \text{and} \quad e_{t-h,h} = \sum_{k=0}^{h-1} u_{t-k}' J \left(\mathbf{A}'\right)^k J' e_y.$$

For part (b), first noting that

$$\begin{aligned}\left\|\widehat{CN}-CN\right\|_{\max} &= \left\|\frac{1}{n}\sum_{t=p}^{n-h} R_1\left(\hat{\Sigma}_{UW}^{-1}\hat{U}_t-\Sigma_{UW}^{-1}U_t\right)W_t'R_1'\right\|_{\max} \\ &\leqslant \left\|\frac{1}{n}\sum_{t=p}^{n-h} R_1\left(\hat{\Sigma}_{W}^{-1}-\Sigma_{UW}^{-1}\right)\left(\hat{U}_t-U_t\right)W_t'R_1'\right\|_{\max} \\ &+ \left\|\frac{1}{n}\sum_{t} R_1\left(\hat{\Sigma}_{UW}^{-1}-\Sigma_{UW}^{-1}\right)U_tW_t'R_1'\right\|_{\max} \\ &+ \left\|\frac{1}{n}\sum_{t} R_1\Sigma_{UW}^{-1}\left(\hat{U}_t-U_t\right)W_t'R_1'\right\|_{\max} = \tilde{I}_1+\tilde{I}_2+\tilde{I}_3.\end{aligned}$$

Also, applying again Lemma A.1 to suitable filters give

$$\left\|\frac{1}{n}\sum_{t} W_{t+j}W_t'-\Sigma_W(j)\right\|_{\max} = O_p\left(\sqrt{\tilde{\nu}_n/n}\right) \quad \text{where} \quad \Sigma_W(j):=E\left[W_{t+j}W_t'\right].$$

It then follows by $\left\|\Sigma_W(j)\right\|_{\max}=O(1)$ that $\left\|\sum_t W_{t+j}W_t'/n\right\|_{\max}=O_p(1)$. This result is implies by T, $\|J\|_\infty=1$, $\left\|\tilde{e}_{p(j+1)}\otimes I_d\right\|_\infty=1$, and

$$\hat{U}_t-U_t=\sum_{j=0}^{p-1}\left(\tilde{e}_{p(j+1)}\otimes I_d\right)\left(\hat{u}_{t-j}-u_{t-j}\right)=\sum_{j=0}^{p-1}\left(\tilde{e}_{p(j+1)}\otimes I_d\right)J(\mathbf{A}-\hat{\mathbf{A}})W_{t-j-1}$$

that

$$\begin{aligned}\left\|\frac{1}{n}\sum_{t=p}^{n-h}\left(\hat{U}_t-U_t\right)W_t'\right\|_{\max} &\leqslant \sum_{j=0}^{p-1}\left\|\tilde{e}_{p(j+1)}\otimes I_d\right\|_\infty \|J\|_\infty\|\hat{\mathbf{A}}-\mathbf{A}\|_\infty \left\|\frac{1}{n}\sum_{t=p}^{n-h}W_{t-j-1}W_t'\right\|_{\max} \\ &= O_p\left(\|\hat{\mathbf{A}}-\mathbf{A}\|_\infty\right).\end{aligned}$$

In addition, it worth noting that the second assertion in part (a) and T impliy that

$$\left\|\frac{1}{n}\sum_{t=p}^{n-h}\Sigma_{UW}^{-1}U_tW_t'\right\|_{\max} \leq \left\|\frac{1}{n}\sum_{t=1}^{n-h}\Sigma_{UW}^{-1}U_tW_t'-I_{dp}\right\|_{\max}+\left\|I_{dp}\right\|_{\max}=O_p\left(\sqrt{\tilde{\nu}_n/n}\right)+1=O_p(1).$$

It then follows by T, $\left\|\hat{\Sigma}_{UW}^{-1}\right\|_\infty=O_p(1)$, $\left\|\Sigma_{UW}^{-1}\right\|_\infty=O_p(1)$, and $\left\|R_1\right\|_\infty=1$ that

$$\tilde{I}_1 \leq \left\|R_1\right\|_\infty^2\left\|\hat{\Sigma}_{UW}^{-1}-\Sigma_{UW}^{-1}\right\|_\infty\left\|\frac{1}{n}\sum_{t=p}^{n-h}\left(\hat{U}_t-U_t\right)W_t'\right\|_{\max}=O_p\left(\left\|\hat{\Sigma}_{UW}-\Sigma_{UW}\right\|_\infty\|\hat{\mathbf{A}}-\mathbf{A}\|_\infty k_{UW}^2\right)$$

$$\tilde{I}_2 \leqslant \left\|R_1\right\|_\infty^2 \left\|\hat{\Sigma}_{UW}^{-1}\right\|_\infty \left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_\infty \left\|\frac{1}{n}\sum_{t=p}^{n-h} \Sigma_{UW}^{-1} U_t W_t'\right\|_{\max} = O_p\left(\left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_\infty k_{UW}\right)$$

$$\tilde{I}_3 \leqslant \left\|R_1\right\|_\infty^2 \left\|\Sigma_{UW}^{-1}\right\|_\infty \left\|\frac{1}{n}\sum_{t=p}^{n-h} \left(\hat{U}_t - U_t\right) W_t'\right\|_{\max} = O_p\left(\|\hat{\mathbf{A}} - \mathbf{A}\|_\infty k_{UW}\right).$$

By plugging in the derived rates and dropping the higher-order terms, we obtain

$$\left\|\widehat{CN} - CN\right\|_{\max} = O_p\left(\|\hat{\mathbf{A}} - \mathbf{A}\|_\infty k_{UW} + \left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_\infty k_{UW}\right).$$

The assertion in part (b) follows from $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, $\|CN^{-1}\|_1 = O_p(1)$, and $\|\widehat{CN}^{-1}\|_\infty = O_p(1)$.

Given parts (a) and (b), assertion in part (c) is straightforward. In fact, note that $\left\|CN - I_p\right\|_{\max} = O_p\left(\sqrt{\tilde{\nu}_n/n}\right)$ by the second assertion in part (a) and the fact that $R_1 R_1' = I_p$.. It then follows by T and result we have just derived in part (b) that $\left\|\widehat{CN} - I_p\right\|_{\max} = O_p\left(\sqrt{\tilde{\nu}_n/n} + \|\hat{\mathbf{A}} - \mathbf{A}\|_\infty k_{UW} + \left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_\infty k_{UW}\right)$. The result in part (c) is obtained by noting that $\|\widehat{CN}^{-1}\|_\infty = O_p(1)$ and $\widehat{CN}^{-1} - I_p = \widehat{CN}^{-1}\left(I_p - \widehat{CN}\right)$.

□

Lemma A.6.
If Assumption 2 is satisfied, then for any vector $v \in \mathbb{R}^p$ such that $\|v\|_1 = 1$,

$$\begin{aligned}\sqrt{n}v'\left(\hat{\beta}_{1,h}^{(de-2S)} - \beta_{1,h}\right) = v'\left(E\left[U_{1,t}^\perp W_{1,t}'\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} U_{1,t}^\perp e_{t,h}\right) \\ + O_p\Big(\tilde{\nu}_n/\sqrt{n} + \|\hat{\mathbf{A}} - \mathbf{A}\|_\infty k_{UW}\sqrt{\tilde{\nu}_n} + \left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_\infty k_{UW}\sqrt{\tilde{\nu}_n} \\ + \left\|\hat{\beta}_{2,h} - \beta_{2,h}\right\|_\infty \left\{\sqrt{\tilde{\nu}_n} + \|\hat{\mathbf{A}} - \mathbf{A}\|_\infty k_{UW}\sqrt{\tilde{\nu}_n} + \left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_\infty k_{UW}\sqrt{\tilde{\nu}_n}\right\}\Big)\end{aligned} \tag{A.8}$$

Proof of Lemma A.6. By the definition of the de-2S estimator,

$$\begin{aligned}\sqrt{n}v'\left(\hat{\beta}_{1,h}^{(de-2S)} - \beta_{1,h}\right) &= v'\left(\frac{1}{n}\sum_{t=p}^{n-h}\hat{U}_{1,t}^\perp W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\hat{U}_{1,t}^\perp e_{t,h}\right) \\ &+ v'\left(\frac{1}{n}\sum_{t=p}^{n-h}\hat{U}_{1,t}^\perp W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\hat{U}_{1,t}^\perp W_{2,t}'\left(\beta_{2,h} - \hat{\beta}_{2,h}\right)\right) \\ &= v'\left(E\left[U_{1,t}^\perp W_{1,t}'\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} U_{1,t}^\perp e_{t,h}\right) + \tilde{\Lambda}_0 + \tilde{\Lambda}_1 + \tilde{\Lambda}_2,\end{aligned} \tag{A.9}$$

where,

$$\tilde{\Lambda}_0 := v'\left\{\left(\frac{1}{n}\sum_{t=p}^{n-h} U_{1,t}^{\perp} W_{1,t}'\right)^{-1} - \left(E\left[U_{1,t}^{\perp} W_{1,t}'\right]\right)^{-1}\right\}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} U_{1,t}^{\perp} e_{t,h}\right) \tag{A.10}$$

$$\tilde{\Lambda}_1 := v'\left(\frac{1}{n}\sum_{t=p}^{n-h} \hat{U}_{1,t}^{\perp} W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} \hat{U}_{1,t}^{\perp} e_{t,h}\right) - v'\left(\frac{1}{n}\sum_{t=p}^{n-h} U_{1,t}^{\perp} W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} U_{1,t}^{\perp} e_{t,h}\right)$$

$$\tilde{\Lambda}_2 := v'\left(\frac{1}{n}\sum_{t=p}^{n-h} \hat{U}_{1,t}^{\perp} W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} \hat{U}_{1,t}^{\perp} W_{2,t}'\left(\beta_{2,h} - \hat{\beta}_{2,h}\right)\right).$$

Using the fact that

$$U_{1,t}^{\perp} = \left(R_1 \Sigma_{UW}^{-1} R_1'\right)^{-1} R_1 \Sigma_{UW}^{-1} U_t \quad \text{and} \quad \hat{U}_{1,t}^{\perp} = \left(R_1 \hat{\Sigma}_{UW}^{-1} R_1'\right)^{-1} R_1 \hat{\Sigma}_{UW}^{-1} U_t,$$

$\tilde{\Lambda}_1$ can be rewritten as

$$\tilde{\Lambda}_1 = \frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} v'\left(\widehat{CN}^{-1} R_1 \hat{\Sigma}_{UW}^{-1} \hat{U}_t - CN^{-1} R_1 \Sigma_{UW}^{-1} U_t\right) e_{t,h} = \tilde{\Lambda}_{11} + \tilde{\Lambda}_{12} + \tilde{\Lambda}_{13},$$

where $\widehat{CN}$ and CN are defined as in the statement of Lemma A.5 and

$$\tilde{\Lambda}_{11} := \frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} v'\left(\widehat{CN}^{-1} - CN^{-1}\right) R_1\left(\hat{\Sigma}_{UW}^{-1} - \Sigma_{UW}^{-1}\right)\hat{U}_t e_{t,h}$$

$$\tilde{\Lambda}_{12} := \frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} v'\left(\widehat{CN}^{-1} - CN^{-1}\right) R_1 \Sigma_{UW}^{-1} \hat{U}_t e_{t,h} \tag{A.11}$$

$$\tilde{\Lambda}_{13} := \frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} v' CN^{-1} R_1\left(\hat{\Sigma}_{UW}^{-1} - \Sigma_{UW}^{-1}\right)\hat{U}_t e_{t,h}$$

$$\tilde{\Lambda}_{14} := \frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} v' CN^{-1} R_1 \Sigma_{UW}^{-1}(\hat{U}_t - U_t) e_{t,h}$$

Also,

$$\tilde{\Lambda}_2 = \frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} v'\widehat{CN}^{-1} R_1 \hat{\Sigma}_{UW}^{-1} \hat{U}_t W_t' R_2'\left(\beta_{2,h} - \hat{\beta}_{2,h}\right) = \left(\tilde{\Lambda}_{21} + \tilde{\Lambda}_{22} + \tilde{\Lambda}_{23} + \tilde{\Lambda}_{24}\right)\left(\beta_{2,h} - \hat{\beta}_{2,h}\right),$$

where

$$\tilde{\Lambda}_{21} := \frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} v' R_1 \Sigma_{UW}^{-1} \hat{U}_t W_t' R_2'$$

$$\tilde{\Lambda}_{22} := \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' R_1 \left(\hat{\Sigma}_{UW}^{-1} - \Sigma_{UW}^{-1}\right) \hat{U}_t W_t' R_2'$$

$$\tilde{\Lambda}_{23} := \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(\widehat{CN}^{-1} - I_p\right) R_1 \Sigma_{UW}^{-1} \hat{U}_t W_t' R_2' \tag{A.12}$$

$$\tilde{\Lambda}_{24} := \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(\widehat{CN}^{-1} - I_p\right) R_1 \left(\hat{\Sigma}_{UW}^{-1} - \Sigma_{UW}^{-1}\right) \hat{U}_t W_t' R_2'.$$

It remains to show that the terms in (A.10), (A.11) and (A.12) are of the specified orders so that the result follows.
First, note that $E\left[U_{1,t}^{\perp} W_{1,t}'\right] = \left(R_1 \Sigma_{UW}^{-1} R_1'\right)^{-1}$, so that $\tilde{\Lambda}_0 = \sum_{t=p}^{n-h} v' \left(CN^{-1} I_p\right) R_1 \Sigma_{uW}^{-1} U_t e_{t,h} / \sqrt{n}$. It then follows by Lemma A.5 that

$$\left|\widetilde{\Lambda}_0\right| \leq \|v\|_1 \left\|CN^{-1} - I_p\right\|_{\max} \left\| \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} R_1 \Sigma_{UW}^{-1} U_t e_{t,h} \right\|_{\max} = O_p\left(\tilde{\nu}_n / \sqrt{n}\right).$$

Lemma A.1 applied to suitable filters yields $\left\|\sum_{t=p}^{n-h} W_{t-j-1} e_{t,h} / \sqrt{n}\right\|_{\max} = O_p\left(\sqrt{\tilde{\nu}_n}\right)$ for $j = 0, 1, \cdots, p-1$. It then follows by T that

$$\begin{aligned}
\left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{n-h} \left(\hat{U}_t - U_t\right) e_{t,h} \right\|_{\max} &\leq \sum_{j=0}^{p-1} \left\| \tilde{e}_{p(j+1)} \otimes I_d \right\|_{\infty} \|J\|_{\infty} \|\hat{\mathbf{A}} - \mathbf{A}\|_{\infty} \left\| \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} W_{t-j-1} e_{t,h} \right\| \\
&= O_p\left(\|\hat{\mathbf{A}} - \mathbf{A}\|_{\infty} \sqrt{\tilde{\nu}_n}\right).
\end{aligned}$$

By Lemma A.5, T and the fact that $\left\|\hat{\Sigma}_{UW}^{-1}\right\|_{\infty} = O_p\left(k_{UW}\right)$, we have

$$\begin{aligned}
\left|\tilde{\Lambda}_{11}\right| \leq& \left| \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(\widehat{CN}^{-1} - CN^{-1}\right) R_1 \left(\hat{\Sigma}_{UW}^{-1} - \Sigma_{UW}^{-1}\right) (\hat{U}_t - U_t) e_{t,h} \right| \\
&+ \left| \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(\widehat{CN}^{-1} - CN^{-1}\right) R_1 \left(\hat{\Sigma}_{UW}^{-1} - \Sigma_{UW}^{-1}\right) U_t e_{t,h} \right| \\
&\leqslant p \|v\|_1 \left\|\widehat{CN}^{-1} - CN^{-1}\right\|_{\max} \left\|\hat{\Sigma}_{UW}^{-1} - \Sigma_{UW}^{-1}\right\|_{\infty} \left\|R_1\right\|_{\infty} \left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{n-h} \left(\hat{U}_t - U_t\right) e_{t,h} \right\|_{\max} \\
&+ p \|v\|_1 \left\|\widehat{CN}^{-1} - CN^{-1}\right\|_{\max} \left\|\hat{\Sigma}_{UW}^{-1}\right\|_{\infty} \left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_{\infty} \left\| \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} \Sigma_{UW}^{-1} U_t e_{t,h} \right\|_{\max} \\
&= O_p\left(\left\|\widehat{CN}^{-1} - CN^{-1}\right\|_{\max} \left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_{\infty} k_{UW} \sqrt{\tilde{\nu}_n} \left(1 + k_{UW} \left\|\hat{\mathbf{A}} - \mathbf{A}\right\|_{\infty}\right)\right)
\end{aligned}$$

$$\left|\tilde{\Lambda}_{12}\right| \leqslant p\|v\|_1 \left\|\widehat{CN}^{-1} - CN^{-1}\right\|_{\max} \left\|\Sigma_{UW}^{-1}\right\|_{\infty} \left\|R_1\right\|_{\infty} \left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{n-h}\left(\hat{U}_t - U_t\right)e_{t,h}\right\|_{\max}$$

$$+ p\|v\|_1 \left\|\widehat{CN}^{-1} - CN^{-1}\right\|_{\max} \left\|R_1\right\|_{\infty} \left\|\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\Sigma_{UW}^{-1}U_t e_{t,h}\right\|_{\max}$$

$$= O_p\left(\left\|\widehat{CN}^{-1} - CN^{-1}\right\|_{\max} \sqrt{\tilde{\nu}_n}\left(1 + k_{UW}\left\|\hat{\mathbf{A}} - \mathbf{A}\right\|_{\infty}\right)\right)$$

$$\left|\tilde{\Lambda}_{13}\right| = O_p\left(\left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_{\infty} k_{UW}\sqrt{\tilde{\nu}_n}\left(1 + k_{UW}\left\|\hat{\mathbf{A}} - \mathbf{A}\right\|_{\infty}\right)\right)$$

$$\left|\tilde{\Lambda}_{14}\right| = O_p\left(k_{UW}\left\|\hat{\mathbf{A}} - \mathbf{A}\right\|_{\infty}\sqrt{\tilde{\nu}_n}\right)$$

To obtain the order of $\tilde{\Lambda}_2$, it is worth noting that $\left\|\sum_{t=p}^{n-h} R_1 \Sigma_{UW}^{-1} U_t W_t R_2'\right\|_{\max} / \sqrt{n} = O_p\left(\sqrt{\tilde{\nu}_n}\right)$ by Lemma A.5 and the fact that $R_1 R_2' = 0_{p\times d(p-1)}$. It fololows by T that

$$\left|\tilde{\Lambda}_{21}\right| = O_p\left(\|\mathbf{A} - \hat{\mathbf{A}}\|_{\infty}\sqrt{\tilde{\nu}_n} + \sqrt{\tilde{\nu}_n}\right)$$
$$\left|\tilde{\Lambda}_{22}\right| = O_p\left(\left(1 + \|\mathbf{A} - \hat{\mathbf{A}}\|_{\infty}\right)\left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_{\infty} k_{UW}\sqrt{\tilde{\nu}_n}\right)$$
$$\left|\tilde{\Lambda}_{23}\right| = O_p\left(\left(1 + \|\mathbf{A} - \hat{\mathbf{A}}\|_{\infty}\right)\left\|\widehat{CN}^{-1} - I_p\right\|_{\max}\sqrt{\tilde{\nu}_n}\right)$$
$$\left|\tilde{\Lambda}_{24}\right| = O_p\left(\left(1 + \|\mathbf{A} - \hat{\mathbf{A}}\|_{\infty}\right)\left\|\widehat{CN}^{-1} - I_p\right\|_{\max}\left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_{\infty} k_{UW}\sqrt{\tilde{\nu}_n}\right).$$

By substituting the derived rates into Equation (A.9) and neglecting the higher-order terms, we obtain the result. □

Lemma A.7.
If Assumptions 2 and 3 are satisfied, then for any vector $v \in \mathbb{R}^p$ such that $\|v\|_1 = 1$, it holds that

$$\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\frac{v' R_1 \Sigma_W^{-1} W_t e_{t,h}}{s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v)} \xrightarrow{d} \mathcal{N}(0,1), \tag{A.13}$$

and

$$\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\frac{v' R_1 \Sigma_{UW}^{-1} U_t e_{t,h}}{s.e._{\hat{\beta}_{1,h}^{(de-2S)}}(v)} \xrightarrow{d} \mathcal{N}(0,1). \tag{A.14}$$

Proof of Lemma A.7. The proof of this lemma will rely on Theorem 5.20 (Wooldridge - White, p. 30) in White (1999). To prove the first result, consider, for any n and t, the following double array of scalars[13],

$$z_{nt} := \frac{v' R_1 \Sigma_W^{-1} W_t e_{t,h}}{s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v)}.$$

We need to justify that $\{z_{nt}\}$ satisfies the hypotheses in the statement of Theorem 5.20 (White, 1999), that is

(i) $E\left[\left|z_{nt}\right|^r\right] < \Delta < \infty$ for some $r \geqslant 2$ and all n, t;

(ii) $\left\{z_{nt}\right\}$ is mixing with mixing coefficient of size $-r/(r-2), r > 2$;

(iii) $\bar{\sigma}_n^2 := \mathrm{Var}\left(n^{1/2} \sum_{t=p}^{n-h} z_{nt}\right) > \delta > 0$ for all n sufficiently large.

The first step in proving *(ii)* is to justify that $1/s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v) = O(1)$. To do so, consider $\tilde{v} = v/\|v\|_2$ so that $\|\tilde{v}\|_2 = 1$. Then, $\left\|R_1' \tilde{v}\right\|_2^2 = \tilde{v}' R_1 R_1' \tilde{v}' = \tilde{v}'\tilde{v} = 1$. Also, we have $\left\|R_1\right\|_2 = 1$ by $R_1 R_1' = I_p$. By the unitary invariance property of the norm $\|\cdot\|_2$, we have $\left\|R_1 \Sigma_W^{-1} R_1' \tilde{v}\right\|_2 = \left\|\Sigma_W^{-1}\right\|_2$. Therefore, it follows by $\|v\|_2 \geq p\|v\|_1 = p$ and Assumption 2(*vii*) that

$$v'(R_1 \Sigma_W^{-1} R_1')\Omega_{W_1,h}(R_1 \Sigma_W^{-1} R_1')v \geqslant \lambda_{\min}\left(\Omega_{W_1,h}\right) \left\|R_1 \Sigma_W^{-1} R_1'\right\|_2^2 \geqslant \frac{1}{C} \left\|\Sigma_W^{-1}\right\|_2 \|v\|_2^2 = \frac{1}{C} \left\|\Sigma_W^{-1}\right\|_2^2$$

It follows by the fact that $s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v)^2 = \lim_{n\to\infty}\left(v'(R_1 \Sigma_W^{-1} R_1')\Omega_{W_1,h}(R_1 \Sigma_W^{-1} R_1')v\right)$ and Assumption 2(*vii*) that $1/s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v) = O(1)$.

Given this result, to show (i), it is sufficient to justify that

$$E\left|v' R_1 \Sigma_W^{-1} W_t e_{t,h}\right|^r < \Delta < \infty \quad \text{for some } r \geq 1.$$

Recall that the stability assumption implies the following VAR(∞) representation for W_t:

$$W_t = \sum_{j=0}^{\infty} \Upsilon_j u_{t-j} \text{ with } \Upsilon_j = \mathbf{A}^j J'.$$

Also, by definition,

$$e_{t,h} = e_y' u_t^{(h)} = \sum_{k=0}^{h-1} u_{t+h-k}' \Psi_k' e_y \text{ with } \Psi_k = J\mathbf{A}^k J'.$$

For $j = 0, 1, \cdots, \infty$ and $k = 0, 1, \cdots, h-1$, let $v_{1,j} = \Upsilon_j' \Sigma_W^{-1} R_1' v$ and $v_{2,k} = \Psi_k' e_y$. Also, let

[13]Note that z_{nt} implicitly depends on n through d, as Σ_W is a $dp \times dp$ matrix.

$\tilde{v}_{1,j} = v_{1,j} / \left\| v_{1,j} \right\|_2$ and $\tilde{v}_{2,k} = v_{2,k} / \left\| v_{2,k} \right\|_2$ so that $\left\| \tilde{v}_{1,i} \right\|_2 = 1 = \left\| \tilde{v}_{2,k} \right\|_2$. Then, it follows that,

$$v' R_1 \Sigma_W^{-1} W_t' e_{t,h} = \sum_{k=0}^{h-1} a_k, \text{ where } a_k = \left(v_{2,k}' u_{t+h-k} \right) \sum_{j=0}^{\infty} v_{1,j}' u_{t-j} \text{ for all } k = 0, 1, \cdots, h-1.$$

By Assumptions 2(*ii*) and 3(*ii*), and given that $\|J\|_2 = \left\| R_1 \right\|_2 = 1$, $\|v\|_2 \leq 1$, and $\left\| \Sigma_W^{-1} \right\|_2 \leqslant C$, it follows, for $r > 2$ as defined in Assumption 3(*ii*), that

$$E \left| v_{2,k}' u_{t+h-k} \right|^{2r} = \left\| v_{2,k} \right\|_2 E \left| \tilde{v}_{2,k}' u_{t+h-k} \right|^{2r} \leqslant c_0 \left\| e_y' J \mathbf{A}^k J' \right\|_2 \leqslant c_0 \left\| \mathbf{A}^k \right\|_2 \leqslant c_0 \varphi^k,$$

and

$$\begin{aligned} \sum_{j=0}^{\infty} \left(E \left| v_{1,j}' u_{t-j} \right|^{2r} \right)^{1/2r} &= \sum_{j=0}^{\infty} \left\| v_{1,j} \right\|_2 \left(E \left| \tilde{v}_{1,j}' u_{t-j} \right|^{2r} \right)^{1/2r} \\ &\leq c_0^{1/2r} \sum_{j=0}^{\infty} \left\| \Upsilon_j' \Sigma_W^{-1} R_1' v \right\|_2 \leq c_0^{1/2r} \sum_{j=0}^{\infty} \varphi^j = \frac{c_0^{1/2r}}{1-\varphi} < \infty. \end{aligned}$$

It then follows from the Minkowski's inequality that

$$E \left(\sum_{j=0}^{\infty} \left| v_{1,j}' u_{t-j} \right| \right)^{2r} \leqslant \left(\sum_{j=0}^{\infty} \left(E \left| v_{1,j}' u_{t-j} \right|^{2r} \right)^{1/2r} \right)^{2r} \leqslant \frac{c_0}{(1-\varphi)^{2r}}.$$

Then hypothesis *(ii)* is verified by T and CS as follows:

$$\begin{aligned} E \left| v' R_1 \Sigma_W^{-1} W_t e_{t,h} \right|^r &\leqslant E \left(\sum_{k=0}^{h-1} \left| a_k \right| \right)^r \leqslant 2^{(h-2)(r-1)} \sum_{k=0}^{h-1} E \left| a_k \right|^r \\ &\leqslant C \sum_{k=0}^{h-1} E \left| v_{2,k}' u_{t+h-k} \right|^r \left| \sum_{j=0}^{\infty} v_{1,j}' u_{t-j} \right|^r \\ &\leqslant C \sum_{k=0}^{h-1} \left\{ E \left| v_{2,k}' u_{t+h-k} \right|^{2r} E \left(\sum_{j=0}^{\infty} \left| v_{1,j}' u_{t-j} \right| \right)^{2r} \right\}^{1/2} \\ &\leqslant \frac{C}{(1-\varphi)^r} \sum_{k=0}^{h-1} \varphi^{k/2} = \frac{C \left(1 - \varphi^{h/2} \right)}{(1-\varphi)^r \left(1 - \varphi^{1/2} \right)} := \Delta < \infty \end{aligned}$$

To prove *(ii)*, first note that $e_{t,h}$ is strongly mixing of size $-r/(r-2)$, since it is a linear combination of h-periods u_t's, and h is some finite integer. Also W_t is mixing of size $-r/(r-2)$ by Assumption 3(*i*). Due to Proposition 3.50 in White (1999) (if two elements are strong mixing of size $-a$, then the product of two are also strong mixing of size $-a$), $W_t e_{t,h}$ is mixing of size $-r/(r-2)$. Therefore, z_{nt} is mixing of size $-r/(r-2)$ as a linear transformation of $W_t e_{t,h}$.

First, we check the term $U_{1,t}^{\perp}$ is a strong mixing process of size $-r/(r-2)$ for $r>2$. Due to Proposition 3.50 in White (1999) (if two elements are strong mixing of size $-a$, then the product of two are also strong mixing of size $-a$), it is easy to show that U_t is a strong

It remains to show *(iii)*. Using the expression for $W_{1,t}^{\perp}$, it is straightforward that the asymptotic variance of the de-biased LS, as defined by Equation (4.13), can be rewritten as

$$s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v)^2 = \lim_{n\to\infty}\left(v'R_1\Sigma_W^{-1}\Omega_{W,h}\Sigma_W^{-1}R_1 v\right).$$

Then,

$$\bar{\sigma}_n^2 := \mathrm{Var}\left(n^{-1/2}\sum_{t=p}^{n-h} z_{nt}\right) = \frac{v'R_1\Sigma_W^{-1}\,\mathrm{Var}\left(n^{-1/2}\sum_{t=p}^{n-h} W_t e_{t,h}\right)\Sigma_W^{-1}R_1' v}{s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v)^2} \to 1,$$

as $n\to\infty$. Therefore, for any arbitrarily small $\delta>0$ (e.g., $\delta<1/2$), we have $\bar{\sigma}_n^2>\delta>0$ for all n sufficiently large.

Given *(i)*, *(ii)*, and *(iii)*, the first result (A.13) follows by the conclusion of Theorem 5.20 (White, 1999).

To prove the second result (A.14), let

$$\tilde{z}_{nt} := \frac{v'R_1\Sigma_{UW}^{-1}U_t e_{t,h}}{s.e._{\hat{\beta}_{1,h}^{(de-2S)}}(v)}.$$

Similar to what was done above for de-biased LS, it is straightforward to check that $1/s.e._{\hat{\beta}_{1,h}^{(de-2S)}}(v) = O(1)$. Hence, z_{nt} has its r^{th} moment bounded as long as this is the case for $v'R_1\Sigma_{UW}^{-1}U_t e_{t,h}$. Recall that

$$U_t = \sum_{j=0}^{p-1}\left(\tilde{e}_{p(j+1)}\otimes I_d\right)u_{t-j},$$

so that if $v_{3,j} := \left(\tilde{e}_{p(j+1)}'\otimes I_d\right)\left(\Sigma_{UW}^{-1}\right)' R_1' v$, for $j=0,1,\cdots,p-1$, it holds that

$$v'R_1\Sigma_{UW}^{-1}U_t e_{t,h} = \sum_{k=0}^{h-1} b_k, \text{ where } b_k = \left(v_{2,k}' u_{t+h-k}\right)\sum_{j=0}^{p-1} v_{3,j}' u_{t-j} \text{ for all } k=0,1,\cdots,h-1.$$

Similar arguments as above lead to

$$E\left|v'R_1\Sigma_{UW}^{-1}U_t e_{t,h}\right| \le Cp^r\frac{1-\varphi^{h/2}}{1-\varphi^{1/2}} := \tilde{\Delta} < \infty,$$

so that hypothesis *(i)* is satisfied if z_{nt} is replaced by $\tilde{z}_{nt}$.

In addition, it is straightforward to show that U_t is a strong mixing process of size $-r/(r-2)$ by its definition, as U_t contains finite number of lagged u_t 's and u_t is a strong mixing process of size $-r/(r-2)$ by Assumption 3(*i*).

Furthermore, $\tilde{\sigma}_n^2 := \mathrm{Var}\left(n^{-1/2}\sum_{t=p}^{n-h}\tilde{z}_{nt}\right) \to 1$, as $n\to\infty$. Therefore, for any arbitrarily small $\tilde{\delta} > 0$ (e.g., $\tilde{\delta} < 1/2$), we have $\tilde{\sigma}_n^2 > \tilde{\delta} > 0$ for all n sufficiently large. The second result (A.14) follows by the conclusion of Theorem 5.20 (White, 1999) applied to the double array of scalars $\{\tilde{z}_{nt}\}$. □

Proof of Theorem 6.2. Recall that $1/s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(\nu) = O(1)$, as justified in the proof of Lemma A.7 above. Using this result and Lemma A.4, we obtain, under Condition 6.1, that

$$\frac{\sqrt{n}\nu'(\hat{\beta}_{1,h}^{(de-LS)} - \beta_{1,h})}{s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(\nu)} = \frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} \nu' R_1 \Sigma_W^{-1} W_t e_{t,h} \Big/ s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(\nu) + o_p(1).$$

Additionally, according to the first result of Lemma A.7 (see Equation (A.13)),

$$\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} \nu' R_1 \Sigma_W^{-1} W_t e_{t,h} \Big/ s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(\nu) \xrightarrow{d} \mathcal{N}(0,1),$$

giving the result.

□

Proof of Theorem 6.3. Recall that the variance estimator of the de-biased LS is given by

$$\widehat{AVar}^{(hac)}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-LS)}\right) = (R_1 \hat{\Sigma}_W^{-1} R_1')\hat{\Omega}_{W_1,h}^{(hac)}(R_1 \hat{\Sigma}_W^{-1} R_1'.$$

To simplify the proof, we drop the subscript and the superscript in $\hat{\Sigma}_W^{-1}$, Σ_W^{-1}, $\hat{\Omega}_{W_1,h}^{(hac)}$, and $\Omega_{W_1,h}$. Nothe by T that,

$$\begin{aligned}
&\left|\nu' R_1 \hat{\Sigma}^{-1} R_1' \hat{\Omega} R_1 \hat{\Sigma}^{-1} R_1' \nu - \nu' R_1 \Sigma^{-1} R_1' \Omega R_1 \Sigma^{-1} R_1' \nu\right| \\
&\quad \leq \left|\nu' R_1 \left(\hat{\Sigma}^{-1} - \Sigma^{-1}\right) R_1' \left(\hat{\Omega} - \Omega\right) R_1 \left(\hat{\Sigma}^{-1} - \Sigma^{-1}\right) R_1' \nu\right| \\
&\quad + 2\left|\nu' R_1 \left(\hat{\Sigma}^{-1} - \Sigma^{-1}\right) R_1' \left(\hat{\Omega} - \Omega\right) R_1 \Sigma^{-1} R_1' \nu\right| + \left|\nu' R_1 \left(\hat{\Sigma}^{-1} - \Sigma^{-1}\right) R_1' \Omega R_1 \left(\hat{\Sigma}^{-1} - \Sigma^{-1}\right) R_1' \nu\right| \\
&\quad + 2\left|\nu' R_1 \left(\hat{\Sigma}^{-1} - \Sigma^{-1}\right) R_1' \Omega R_1 \Sigma^{-1} R_1' \nu\right| + \left|\nu' R_1 \Sigma^{-1} R_1' \left(\hat{\Omega} - \Omega\right) R_1 \Sigma^{-1} R_1' \nu\right| \\
&\quad = S_1 + 2S_2 + S_3 + 2S_4 + S_5.
\end{aligned}$$

It remains to determine the orders of terms S_1 to S_5. First, note that by $\|\nu\|_1 = 1$, $\left\|R_1\right\|_\infty = 1$, and $\left\|\Sigma^{-1}\right\|_\infty = O(k_W)$

$$\begin{aligned}
S_1 &\leqslant \left|\sum_{j,r=1}^{p} \nu_r \nu_j e_r' R_1 \left(\hat{\Sigma}^{-1} - \Sigma^{-1}\right) R_1' \left(\hat{\Omega} - \Omega\right) R_1 \left(\hat{\Sigma}^{-1} - \Sigma^{-1}\right) R_1' e_j\right| \\
&\leqslant C \|\nu\|_1^2 \left\|R_1\right\|_\infty^2 \left\|\left(\hat{\Sigma}^{-1} - \Sigma^{-1}\right) R_1' \left(\hat{\Omega} - \Omega\right) R_1 \left(\hat{\Sigma}^{-1} - \Sigma^{-1}\right)\right\|_{\max} \\
&= O_p\left(\left\|\hat{\Sigma}^{-1} - \Sigma^{-1}\right\|_\infty^2 \|\hat{\Omega} - \Omega\|_{\max}\right);
\end{aligned}$$

$$S_2 \leq C \left\| \left(\hat{\Sigma}^{-1} - \Sigma^{-1}\right) R_1' \left(\hat{\Omega} - \Omega\right) R_1 \Sigma^{-1} \right\|_{\max} = O_p \left(\left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|_\infty \|\hat{\Omega} - \Omega\|_{\max} k_W \right);$$

$$S_5 \leq C \left\| \Sigma^{-1} R_1' \left(\hat{\Omega} - \Omega\right) R_1 \Sigma^{-1} \right\|_{\max} = O_p \left(\|\hat{\Omega} - \Omega\|_{\max} k_W^2 \right).$$

Also, it is well known that for any s.d.p. $r \times r$ matrix M and for all $x, y \in \mathbb{R}^r$, $\left|x'My\right| \leqslant \left(x'Mx\right)^{1/2} \left(y'My\right)^{1/2} \leqslant \lambda_{\max}(A)\|x\|_2\|y\|_2$. Applying this result to S_3 and S_4, it follows by $\lambda_{\max}(\Omega) \leq C$, $\|v\|_1 = 1$, $\left\|R_1\right\|_2 = 1$, and $\left\|\Sigma^{-1}\right\|_2 = O(1)$ that

$$S_3 \leqslant \lambda_{\max}(\Omega) \left\| R_1 \left(\hat{\Sigma}^{-1} - \Sigma^{-1}\right) R_1' v \right\|_2^2 \leq C \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|_\infty^2 \left\|R_1\right\|_\infty^2 \|v\|_2^2 = O_p \left(\left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|_\infty^2 \right) \quad ;$$

$$S_4 \leqslant \lambda_{\max}(\Omega) \left\| R_1 \left(\hat{\Sigma}^{-1} - \Sigma^{-1}\right) R_1' v \right\|_2 \left\| R_1 \Sigma^{-1} R_1' v \right\|_2 = O_p \left(\left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|_\infty \right).$$

Therefore,

$$\left| \widehat{s.e.}^{(hac)}_{\hat{\beta}_{1,h}^{(de-LS)}}(v)^2 - s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v)^2 \right| = O_p \left(\left\| \hat{\Sigma} - \Sigma \right\|_\infty k_W^2 + \left\| \hat{\Omega} - \Omega \right\|_{\max} k_W^2 \right) = o_p(1),$$

under Condition 6.2.

Note that this result and the fact that $1/s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v) = O(1)$ (see the proof of Theorem 6.2) imply $\widehat{s.e.}^{(hac)}_{\hat{\beta}_{1,h}^{(de-LS)}}(v)^2 \Big/ s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v)^2 \xrightarrow{p} 1$. The second result then follows by Theorem 6.2 and Slutsky's theorem.

It remains to show that the asymptotic variance of the de-biased LS has a simplified representation of the form (6.4) if the contemporaneous error term u_t is a conditional m.d.s. First recall that using the expression of $W_{1,t}^{\perp}$, this asymptotic variance can be rewritten as

$$\text{AVar}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-LS)}\right) + o(1) = R_1 \Sigma_W^{-1} \Omega_{W,h} \Sigma_W^{-1} R_1',$$

where

$$\Omega_{W,h} + o(1) = \sum_{k=-\infty}^{\infty} \mathbb{E}[W_t W_{t+k}' e_{t,h} e_{t+k,h}] = \sum_{j,l=0}^{h-1} \sum_{k=-\infty}^{\infty} V_{jlk}(h),$$

with

$$V_{jlk}(h) := E[W_t W_{t+k}' e_{t,h} e_{t+k,h}] = E\left[e_y' \Psi_j u_{t+h-j} u_{t+k+h-l}' \Psi_l' e_y W_t W_{t+k}' \right],$$

and $W_t = \sum_{k=0}^{\infty} \Psi_s u_{t-s}$.

Let $j, l \in \{0, 1, \ldots, h-1\}$ and $k \in \mathbb{Z}$ fixed. Also, let $\mathscr{F}_t = \{u_t, u_{t-1}, \ldots\}$. In order to simply the expression of $V_{jlk}(h)$, we consider three cases.

Case 1: $k > h$. In this case, $1 \leq h - j \leq h < k < k + 1 \leq k + h - l$ and by the law of iterated expectations (LIE) and the m.d.s. assumption,

$$V_{jlk}(h) = E\left[e_y' \Psi_j u_{t+h-j} E\left[u_{t+k+h-l} \mid \mathscr{F}_{t+k+h-l-1} \right]' \Psi_l' e_y W_t W_{t+k}' \right] = 0.$$

Case 2: $k < -h$. In this case, $k < k+h-l \le k+h < 0 < 1 \le h-j$. It follows by the LIE and the m.d.s. assumption that

$$V_{jlk}(h) = E\left[e_y'\Psi_j E\left[u_{t+h-j} \mid \mathscr{F}_{t+h-j-1}\right] u_{t+k+h-l}'\Psi_l' e_y W_t W_{t+k}'\right] = 0.$$

Case 3: $-h \le k \le h$. We consider three subcases.

- If $k = l-j$, then $k < k+1 \le k+h-l = h-j$ and $h-j \ge 1$ so that by the LIE and the conditional homoskedasticity assumption

$$V_{jlk}(h) = E\left[e_y'\Psi_j E\left[u_{t+h-j}u_{t+h-j}' \mid \mathscr{F}_{t+h-j-1}\right]\Psi_l' e_y W_t W_{t+k}'\right] = e_y'\Psi_j\Sigma_u\Psi_l' e_y \Sigma_W(j-l).$$

- If $k < l-j$, then $k < k+1 \le k+h-l < h-j$ and $h-j \ge 1$. It follows by the LIE and the m.d.s. assumption that

$$V_{jlk}(h) = E\left[e_y'\Psi_j E\left[u_{t+h-j} \mid \mathscr{F}_{t+h-j-1}\right] u_{t+k+h-l}'\Psi_l' e_y W_t W_{t+k}'\right] = 0.$$

- If $k > l-j$, then $1 \le h-j < k+h-l$ and $k < k+1 \le k+h-l$. By the LIE and the m.d.s. assumption,

$$V_{jlk}(h) = E\left[e_y'\Psi_j u_{t+h-j} E\left[u_{t+k+h-l} \mid \mathscr{F}_{t+k+h-l-1}\right]'\Psi_l' e_y W_t W_{t+k}'\right] = 0.$$

It follows from all these calculations that

$$\Omega_{W,h} + o(1) = \sum_{j,l=0}^{h-1} e_y'\Psi_j\Sigma_u\Psi_l' e_y \Sigma_W(j-l),$$

leading to the result (6.4). □

Proof of Theorem 6.5. First of all, consider $\tilde{v} = v/\|v\|_2$ so that $\|\tilde{v}\|_2 = 1$. Then, $\left\|R_1'\tilde{v}\right\|_2^2 = \tilde{v}'R_1R_1'\tilde{v}' = \tilde{v}'\tilde{v} = 1$. Also, we have $\left\|R_1\right\|_2 = 1$ by $R_1R_1' = I_p$. By the unitary invariance property of the norm $\|\cdot\|_2$, we have $\left\|R_1\left(\Sigma_{UW}^{-1}\right)'R_1'\tilde{v}\right\|_2 = \left\|(\Sigma_{UW}^{-1})'\right\|_2 = \left\|\Sigma_{UW}^{-1}\right\|_2$. Therefore, it follows by $\|v\|_2 \ge p\|v\|_1 = p$ and Assumption 2(*vii*) that

$$\begin{aligned} v'\left(R_1\Sigma_{UW}^{-1}R_1'\right)\Omega_{U_1,h}\left(R_1\left(\Sigma_{UW}^{-1}\right)'R_1'\right)v &\geqslant \lambda_{\min}\left(\Omega_{U_1,h}\right)\left\|R_1\left(\Sigma_{UW}^{-1}\right)'R_1'\right\|_2^2 \\ &\geqslant \frac{1}{C}\left\|\Sigma_U^{-1}\right\|_2\|v\|_2^2 = \frac{1}{C}\left\|\Sigma_{UW}^{-1}\right\|_2^2 \end{aligned}$$

It follows by the fact that $s.e._{\beta_{1,h}^{de-2S}}(v)^2 = \lim_{n\to\infty}\left(v'\left(R_1\Sigma_{UW}^{-1}R_1'\right)\Omega_{U_1,h}\left(R_1\left(\Sigma_{UW}^{-1}\right)'R_1'\right)v\right)$ and Assumption 2(*vii*) that $1/s.e._{\beta_{1,h}}^{(de-2S)}(v) = O(1)$. With this result and Lemma A.6, we obtain, under Condition 6.3,

$$\frac{\sqrt{n}v'(\hat{\beta}_{1,h}^{(de-2S)} - \beta_{1,h})}{s.e._{\hat{\beta}_{1,h}^{(de-2S)}}(v)} = \frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} v'R_1\Sigma_{UW}^{-1}U_t e_{t,h}\Big/s.e._{\beta_{1,h}}^{(de-2S)}(v) + o_p(1).$$

In addition, according to the second result of Lemma A.7 (see Equation (A.14)),

$$\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h} v' R_1 \Sigma_{UW}^{-1} U_t e_{t,h} / s.e._{\beta_{1,h}}^{(de-2S)}(v) \xrightarrow{d} \mathcal{N}(0,1),$$

giving the result. □

Lemma A.8.
Let W_t follows (2.1), $\|\mathbf{A}\|_2 \le \varphi \in [0,1)$, and Assumption 2(vii), Assumption 3(i), and Assumption 4 hold. Then, for any $v \in \mathbb{R}^{dp\times 1}$, with $\|v\|_1 = 1$.

$$\lim_{n\to\infty} n^{-1}\sum_{t=p}^{n-h}(v's_t)^2/v'\Omega_{U,h}v = 1. \tag{A.15}$$

Proof of Lemma A.8. The convergence of (A.15) will be proved by verifying the conditions of Corollary 3.48 in White (1999), (1) the process $(v's_t)^2$ is strong mixing of size $-r/(1-r)$ for $r>1$, and (2) the $(r+\delta)$-th moment of the process $(v's_t)^2$ is finite for some $\delta>0$. In addition, Assumption 2 (vii) ensures the matrix $\Omega_{U,h}$ is non-singular.

First, we verify the first condition. Recall process $s_t := (e_{t,h}, e_{t+1,h}, \cdots, e_{t+p-1,h}) \otimes u_t$, and $e_{t+i,h}$ is a linear combination of u_{t+i+j} for $i = 0,1,\cdots,p-1$ and $j = 0,1,\cdots,h-1$. Since the horizon h is finite, and u_t is strong mixing (α-mixing) processes with mixing size $-r/(r-2)$, for $r>2$ (Assumption 3 (i)), then the process s_t is strong mixing (α-mixing) processes with mixing size $-r/(r-2)$. Moreover, since $-r/(r-2) < -r/(r-1)$ for $r>2$, then the process s_t and $(v's_t)^2$ are strong mixing (α-mixing) processes with mixing size $-r/(r-1)$.

Next, we verify the second condition. Without loss of generality, we show the following moment condition holds for all $\lambda \in \mathbb{R}^d$ and $\|\lambda\| = 1$,

$$\begin{aligned}
\mathbb{E}[\|(\lambda' u_t e_{t,h})^2\|^{r+\delta}] =& \mathbb{E}[\|\lambda' u_t \sum_{i=0}^{h-1} v_1' J\mathbf{A}^i J' u_{t+h-i}\|^{2r+2\delta}] \\
\le& \sum_{i=0}^{h-1}\mathbb{E}[\|\lambda' u_t u_{t+h-i}' J\mathbf{A}^i J' v_1\|^{2r+2\delta}] = \sum_{i=0}^{h-1}\mathbb{E}[\|\lambda' u_t u_{t+h-i}'\tilde{v}_1\|J\mathbf{A}^i J' v_1\|\|^{2r+2\delta}] \\
& \text{(denote } \tilde{v}_1 = J\mathbf{A}^i J' v_1/\|J\mathbf{A}^i J' v_1\|) \\
\le& \sum_{i=0}^{h-1}\mathbb{E}[\|\lambda' u_t u_{t+h-i}'\tilde{v}_1\|^{2r+2\delta}]\|J\mathbf{A}^i J' v_1\|^{2r+2\delta} \\
\le& \left(\mathbb{E}\|\lambda' u_t\|^{4r+4\delta}\mathbb{E}\|\tilde{v}_1' u_t\|^{4r+4\delta}\right)^{1/2}\sum_{i=0}^{h-1}\|J\mathbf{A}^i J' v_1\|^{2r+2\delta} \\
& \text{(Cauchy–Schwarz inequality)} \\
\le& \left(\mathbb{E}\|\lambda' u_t\|^{4r+4\delta}\mathbb{E}\|\tilde{v}_1' u_t\|^{4r+4\delta}\right)^{1/2}\sum_{i=0}^{\infty}\mathbb{E}\|\mathbf{A}^i\|_2^{2r+2\delta} \\
& \text{(apply } \|AB\| \le \|A\|\|B\|_2, \text{ and } \|v_1\| = 1, \|J\|_2 = 1)
\end{aligned}$$

$$\leq \left(\mathbb{E}\|\lambda' u_t\|^{4r+4\delta}\mathbb{E}\|\tilde{v}_1' u_t\|^{4r+4\delta}\right)^{1/2}\sum_{i=0}^{\infty}\varphi^{(2r+2\delta)i}$$
$$= \left(\mathbb{E}\|\lambda' u_t\|^{4r+4\delta}\mathbb{E}\|\tilde{v}_1' u_t\|^{4r+4\delta}\right)^{1/2}\frac{1}{1-\varphi^{(2r+2\delta)}}$$

By Assumption 4 and $\|\varphi\| < 1$, the above term is bounded by a constant. Thus, the moment condition on process $(v's_t)^2$ is verified. In turn, the convergence is proved. □

Proof of Theorem 6.6. Arguments similar to those presented in the first part of the proof of Theorem 6.3 can be invoked to obtain

$$\left|\widehat{s.e.}^{(hac)}_{\hat{\beta}^{(de-2S)}_{1,h}}(v)^2 - s.e._{\hat{\beta}^{(de-2S)}_{1,h}}(v)^2\right| = O_p\left(\left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_\infty k^2_{UW} + \left\|\hat{\Omega}^{(hac)}_{U_1,h} - \Omega_{U_1,h}\right\|_{\max} k^2_{UW}\right) = o_p(1),$$

under Condition 6.4.

Note that this result and the fact that $1/s.e._{\hat{\beta}^{(de-2S)}_{1,h}}(v) = O(1)$ (see the proof of Theorem 6.2) imply $\widehat{s.e.}^{(hac)}_{\hat{\beta}^{(de-LS)}_{1,h}}(v)^2 \Big/ s.e._{\hat{\beta}^{(de-2S)}_{1,h}}(v)^2 \xrightarrow{p} 1$. The second result then follows by Theorem 6.2 and Slutsky's theorem.
Consistency of the HC variance estimator follows from Lemma A.8.

□

A.2. Additional simulation results

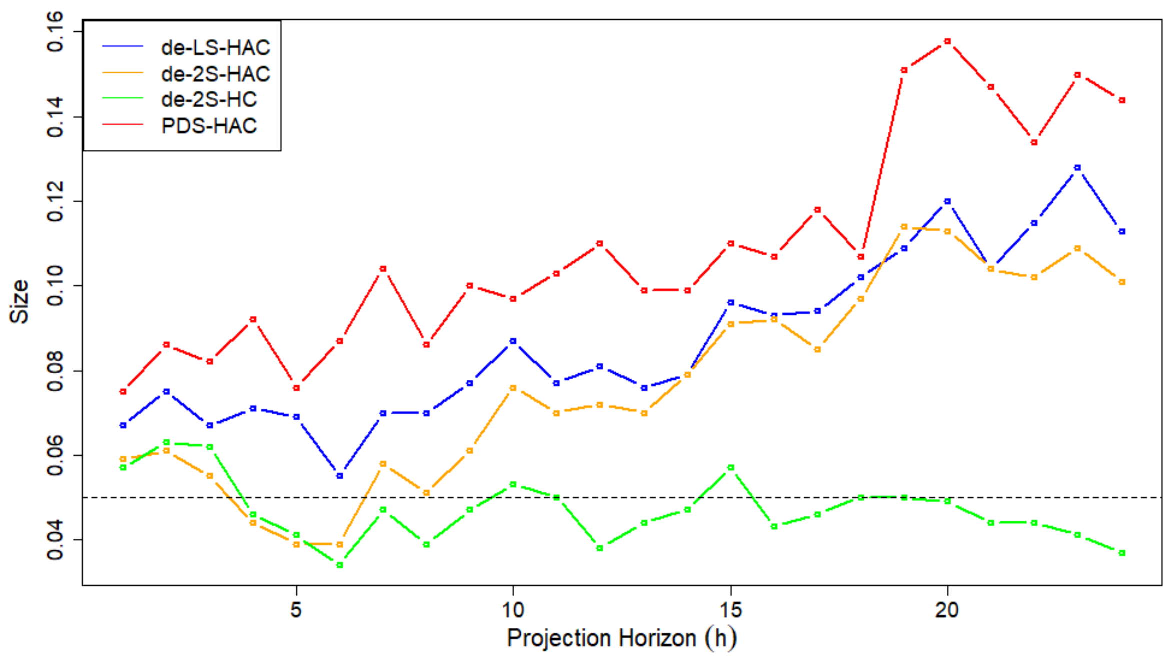

Fig. A.1: Size of the Wald test at the 5% nominal level for different horizons. The red, blue, orange, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased two-stage with HAC standard errors, and the de-biased two-stage with HC standard errors, respectively. The number of time series is $d = 60$, and the sample size is $n = 240$. The horizon is $h = 0, 1, \ldots, 24$. The number of replications is 1,000.

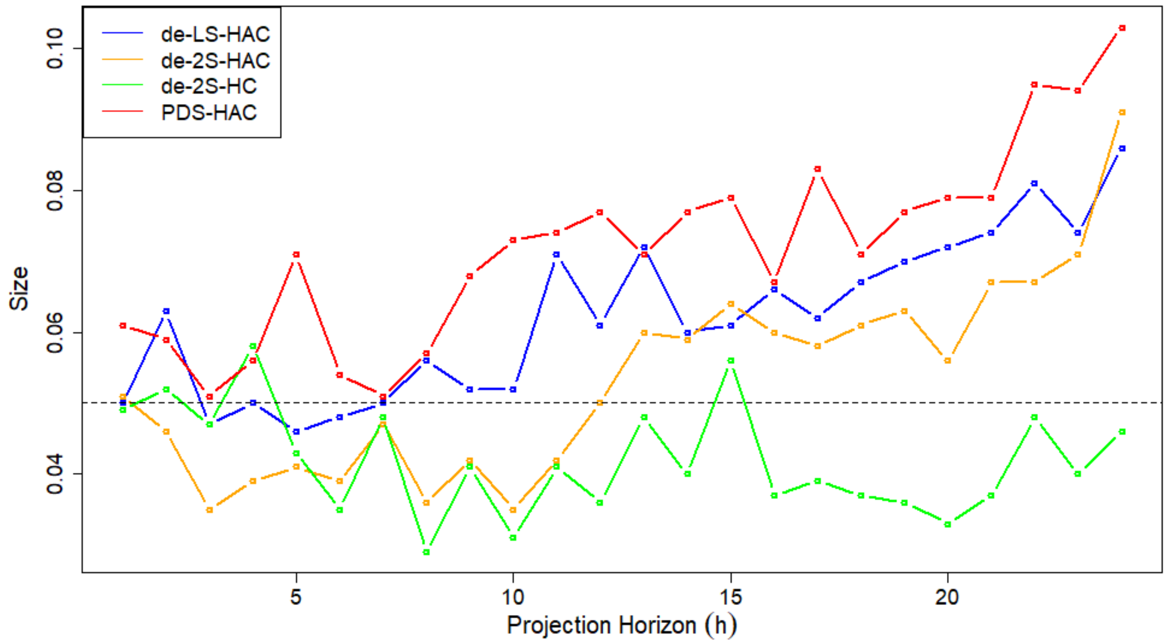

Fig. A.2: Size of the Wald test at the 5% nominal level for different horizons. The red, blue, orange, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased two-stage with HAC standard errors, and the de-biased two-stage with HC standard errors, respectively. The number of time series is $d = 60$, and the sample size is $n = 480$. The horizon is $h = 0, 1, \dots, 24$. The number of replications is $1,000$.

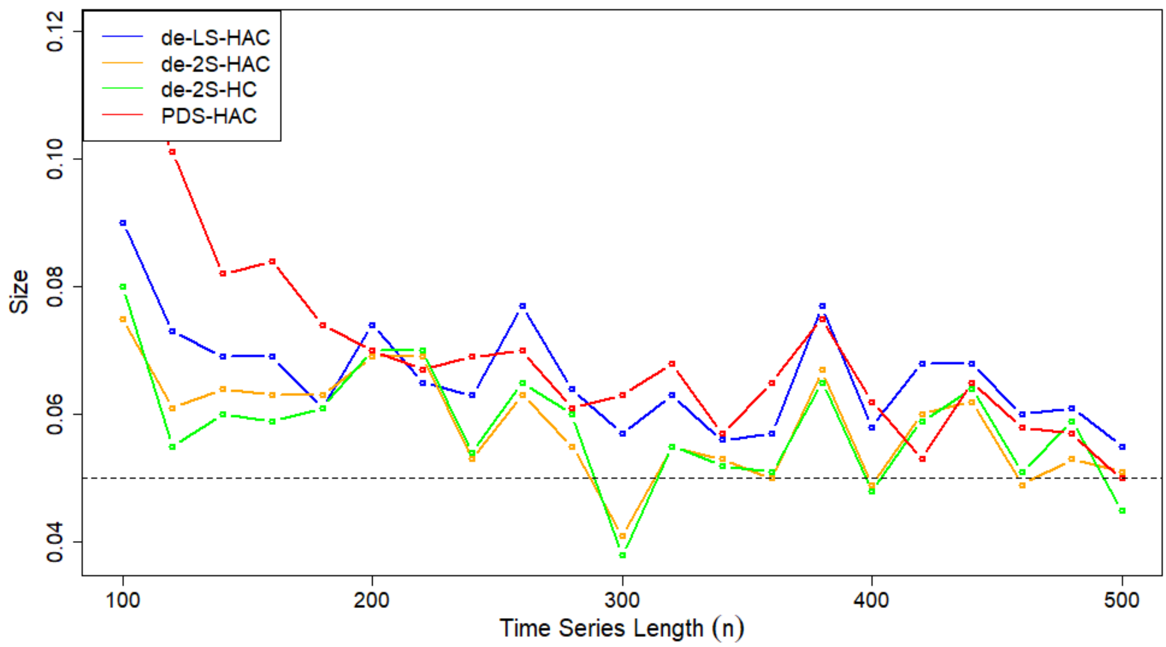

Fig. A.3: Size of the Wald test at the 5% nominal level for different sample sizes and a given horizon ($h = 1$). The red, blue, orange, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased two-stage with HAC standard errors, and the de-biased two-stage with HC standard errors, respectively. The number of time series is $d = 60$. The number of replications is $1,000$.

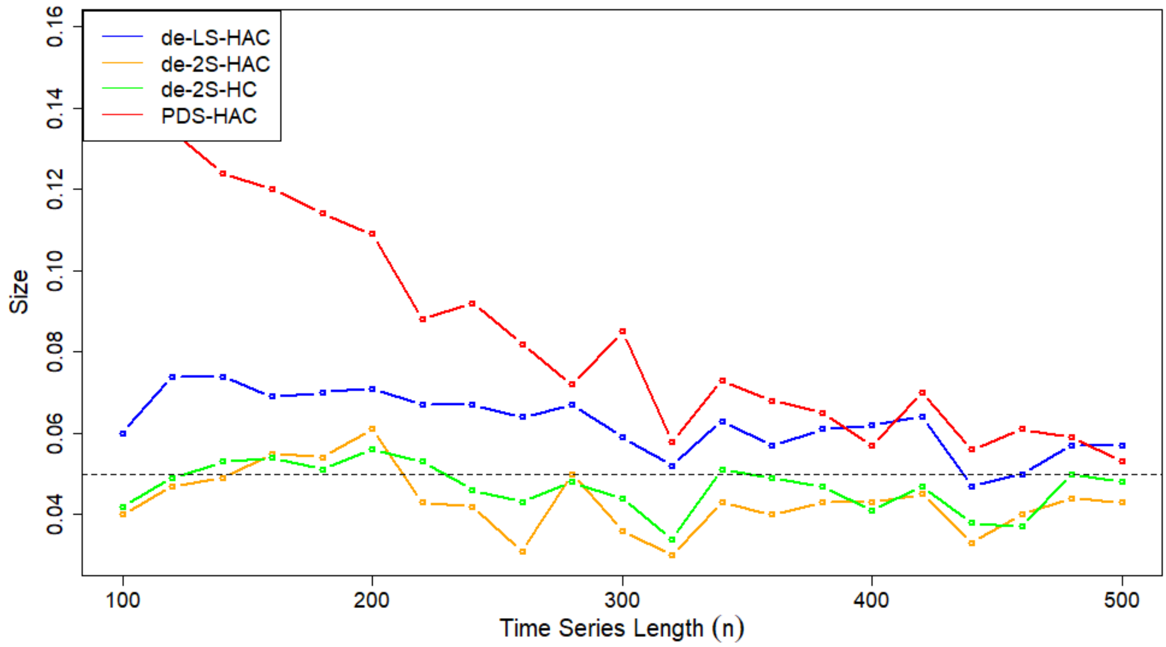

Fig. A.4: Size of the Wald test at the 5% nominal level for different sample sizes and a given horizon ($h = 4$). The red, blue, orange, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased two-stage with HAC standard errors, and the de-biased two-stage with HC standard errors, respectively. The number of time series is $d = 60$. The number of replications is $1,000$.

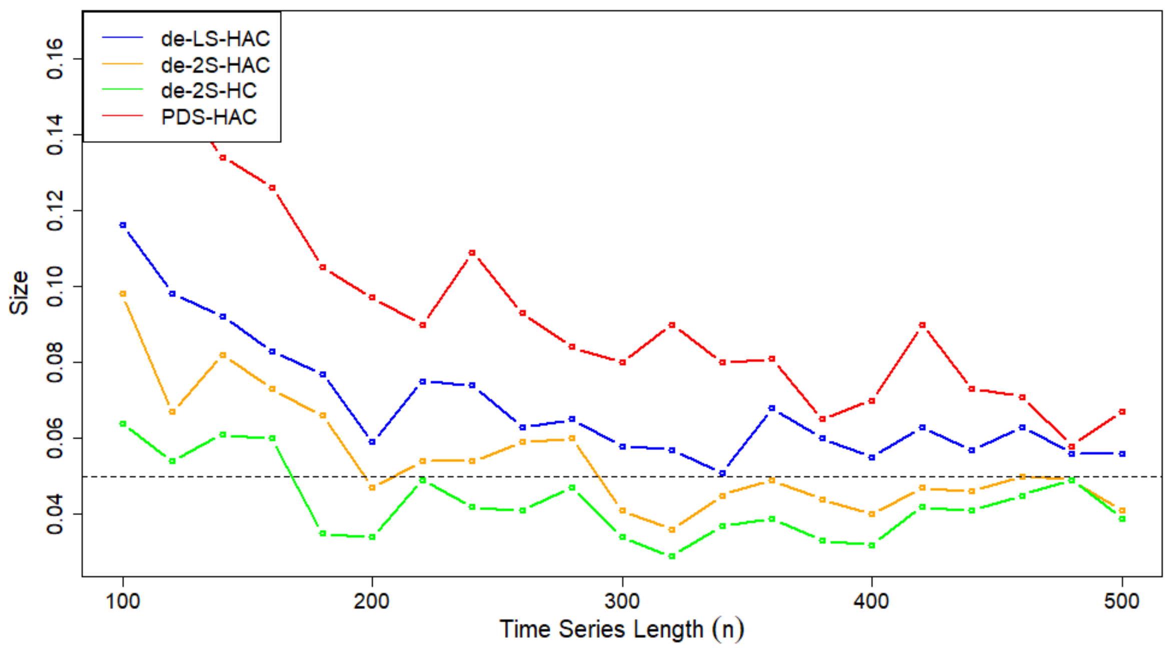

Fig. A.5: Size of the Wald test at the 5% nominal level for different sample sizes and a given horizon ($h = 8$). The red, blue, orange, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased two-stage with HAC standard errors, and the de-biased two-stage with HC standard errors, respectively. The number of time series is $d = 60$. The number of replications is $1,000$.