Université de Montréal

Essays in High-Dimensional Econometrics

This version: 12 juin 2025
[Please click here for the latest version]

par
Eugène DETTAA

Département de sciences économiques
Faculté des arts et des sciences

Thèse présentée à la Faculté des arts et des sciences
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en sciences économiques

Juin, 2025

Université de Montréal
Faculté des arts et des sciences

Cette thèse intitulée:

Essays in High-Dimensional Econometrics

This version: 12 juin 2025
[Please click here for the latest version]

présentée par:

Eugène DETTAA

a été évaluée par un jury composé des personnes suivantes:

Alain-Philippe Fortin,	président-rapporteur
Marine Carrasco,	directrice de recherche
Benoit Perron,	membre du jury
Alain Guay,	examinateur externe

Thèse acceptée le: .....

À

ma mère Hermine,

mon épouse Sandra et mon fils Owen,

et à la mémoire de mon père Jonas, dont le souvenir m'accompagne chaque jour.

ACKNOWLEDGMENTS/REMERCIEMENTS

Je souhaite exprimer ma profonde gratitude à Marine Carrasco pour sa gentillesse, sa patience et sa grande disponibilité. Je me sens chanceux de l'avoir eue comme directrice de thèse. Ses encouragements, ses conseils et ses suggestions m'ont permis de traverser les périodes d'incertitude et m'ont aidé à surmonter les difficultés tout au long de mon doctorat. J'ai beaucoup appris de sa discipline, de sa rigueur et de son sens pointu du détail.

Je voudrais également remercier Benoit Perron, Mathieu Marcoux et Prosper Dovonon pour leurs conseils, commentaires et suggestions. Pour leur implication, leur dynamisme et leur promptitude durant mon année de marché, je tiens à remercier Marine Carrasco, Benoit Perron et Mathieu Marcoux, qui ont bien voulu accepter d'être mes référents. Mes remerciements s'adressent également aux professeurs et au personnel administratif du département de sciences économiques ainsi que du CIREQ. La rédaction de cette thèse n'aurait pas été possible sans le soutien financier de Marine Carrasco, du Département de sciences économiques, des Fonds de recherche du Québec et de la Fondation McConnell.

Je suis également reconnaissant envers mes amis et collègues doctorants, dont la présence et les échanges ont enrichi cette étape déterminante de mon parcours académique. Un merci particulier à mon collègue et ami Endong Wang pour la fructueuse collaboration, ainsi que pour les moments de partage de nos succès comme de nos échecs. Je remercie tous mes collègues membres du CIREQ et du Quebec Workshop in Economics, Statistics, and Finance (QESF) pour les échanges autour de diverses thématiques de recherche. Sans être exhaustif, je pense en particulier à Fansa, Féraud, Coulibaly, Sidi, Stéphane, Firmin, Juste, Kodjo, Régis, Moudachirou, Bruno, Jaures, Anicet, Alexandre et Arsène.

Finalement, je voudrais remercier du fond du cœur mon épouse Sandra, qui a été présente à mes côtés du début jusqu'à la fin de cette thèse. Merci pour le soutien psychologique et la stabilité émotionnelle que tu me procures. Merci pour toutes les fois où tu m'as encouragé à ne pas abandonner. Je remercie aussi mes frères et sœurs, Judicaëlle, Gaëlle, Stéphanie, Elsa, Elvira et Hermann, pour leur soutien multiforme.

RÉSUMÉ

Cette thèse s'appuie sur des techniques de régularisation issues de la littérature en apprentissage automatique pour développer de nouvelles approches méthodologiques répondant à certains défis de l'économétrie de grande dimension, en particulier dans des contextes où les méthodes traditionnelles s'avèrent moins fiables. Elle est organisée en trois chapitres.

Le premier chapitre, rédigé en collaboration avec Endong Wang, propose un test de Wald pour la causalité au sens de Granger à plusieurs horizons dans un cadre autorégressif vectoriel (VAR) parcimonieux incluant un grand nombre de séries temporelles. L'hypothèse nulle du test porte sur un sous-ensemble fini de coefficients causaux d'intérêt dans l'équation de projection locale (LP) en grande dimension à un horizon donné. Alors que l'inférence classique basée sur la méthode des moindres carrés ordinaires (MCO) dans l'équation LP souffre de la malédiction de la dimension dans ce contexte, la méthode de double sélection postérieure appliquée à l'équation LP peut ne pas être applicable, car un modèle VAR parcimonieux n'implique pas nécessairement une équation LP parcimonieuse pour un horizon $h > 1$. Pour pallier ces limites des approches classiques, nous développons deux types d'estimateurs débiaisés des coefficients causaux d'intérêt, reposant tous deux sur un estimateur de première étape des coefficients matriciels du modèle VAR obtenus par régularisation de type Lasso. Le premier estimateur est dérivé de la méthode des moindres carrés, tandis que le second est obtenu à l'aide d'une approche en deux étapes qui offre des gains potentiels d'efficacité au prix d'hypothèses plus restrictives. Nous proposons également une méthode d'inférence robuste à l'hétéroscédasticité et à l'autocorrélation (HAC) pour chaque estimateur. De plus, nous proposons une méthode d'inférence robuste pour l'estimateur en deux étapes, qui élimine la nécessité de corriger l'autocorrélation des résidus de projection. Les simulations Monte Carlo montrent que l'estimateur en deux étapes avec inférence robuste surpasse la méthode des moindres carrés en ce qui concerne le niveau du test de Wald, en particulier pour les horizons de prévision longs. Nous appliquons notre méthodologie pour analyser la transmission de l'incertitude économique liée aux politiques entre un grand nombre de pays,

à court et à long terme. Plus précisément, nous construisons un réseau causal afin de visualiser la propagation de l'incertitude économique entre les pays au fil du temps.

Le deuxième chapitre, rédigé en collaboration avec Endong Wang, introduit une équation de projection locale parcimonieuse en grande dimension pour analyser la transmission de la volatilité des rendements boursiers à travers plusieurs horizons temporels, tout en tenant compte d'un grand nombre d'actifs. L'équation LP proposée, motivée par le modèle autorégressif vectoriel hétérogène (HVAR) des volatilités réalisées, exploite les performances en termes de prévision des composantes de la volatilité mesurées à différents horizons temporels (quotidien, hebdomadaire et mensuel) et est ainsi appelée équation de *projection locale hétérogène (HLP)*. Pour surmonter le défi posé par l'inférence gaussienne asymptotique standard dans l'équation HLP incluant un grand nombre de variables, nous utilisons la méthode d'estimation et d'inférence des moindres carrés débiaisés (*de-biased least squares*, deLS) introduite au premier chapitre, permettant une inférence valide sans recourir à une hypothèse de parcimonie sur l'équation HLP. En appliquant ce cadre à des données haute fréquence sur 30 actions américaines entre 2010 et 2014, nous construisons des réseaux de causalité au sens de Granger à plusieurs horizons (quotidien, hebdomadaire, bimensuel et mensuel) et comparons les performances de l'inférence basée sur le deLS avec celles des méthodes classiques basées sur les MCO. Nos résultats indiquent que la méthode deLS produit des réseaux plus parcimonieux et plus stables, en particulier aux horizons plus longs, où les MCO ont tendance à surajuster et à détecter des connexions non fiables. Ces résultats sont robustes aux analyses par sous-échantillons et demeurent cohérents lorsque les corrélations réalisées sont ajoutées comme variables de contrôle supplémentaires, ce qui permet d'éviter davantage la détection de connexions fallacieuses entre les actifs tout en augmentant considérablement la dimension du modèle.

Le troisième chapitre, rédigé en collaboration avec Marine Carrasco, propose une nouvelle approche pour l'estimation basée sur les conditions de moments dans des contextes de grande dimension où de nombreux moments/instruments sont disponibles. En effet, plusieurs applications économiques impliquent l'utilisation d'un grand nombre d'instruments. Ce chapitre traite de la question de l'estimation effi-

cace dans de tels cadres. Bien que l'utilisation de nombreuses conditions de moments puisse améliorer l'efficacité des estimateurs de type GMM, elle peut également entraîner un biais substantiel en raison d'une forte suridentification. Nous considérons un cadre spécifique dans lequel le paramètre d'intérêt est défini par une unique restriction de moment conditionnelle, ce qui conduit à un grand nombre de conditions de moments. L'estimateur de référence que nous considérons est l'estimateur CUE (*continuous updating GMM estimator*), en raison de son biais relativement faible en présence de nombreuses conditions de moments. Nous introduisons une version régularisée de type Ridge de CUE (RRCUE) afin de résoudre le problème de singularité de la matrice de poids. Nous montrons que l'estimateur RRCUE est convergent, asymptotiquement normal, et atteint la borne d'efficacité semi-paramétrique dans un cadre asymptotique où la taille de l'échantillon et le nombre de conditions de moments tendent vers l'infini, et où le paramètre de régularisation tend vers zéro à une certaine vitesse. Nous proposons une méthode fondée sur la validation croisée pour choisir de manière optimale le paramètre de régularisation. Nous évaluons la performance de RRCUE à l'aide de simulations Monte Carlo. Nos résultats révèlent que la régularisation réduit le problème de dispersion du CUE et améliore l'efficacité, bien qu'elle introduise un biais qui demeure relativement faible avec un nombre modérément élevé de moments. Dans le cas particulier d'un modèle linéaire avec variables instrumentales, notre estimateur se montre compétitif par rapport à certains estimateurs de pointe dans la littérature. Nous appliquons notre méthodologie pour réestimer l'effet de la qualité des institutions et des politiques publiques (ce que l'on appelle *l'infrastructure sociale*) sur le PIB par tête. Nos résultats empiriques sont cohérents avec les résultats de simulation, fournissant des estimations plus précises.

Mots-clés : Causalité au sens de Granger à horizons multiples, VAR en grande dimension, régularisation, estimation dé-biaisée, VAR hétérogène en grande dimension, projection locale hétérogène, réseaux de volatilités réalisées, grand nombre de conditions de moments ou de variables instrumentales, estimateur CUE, estimation efficace.

ABSTRACT

This thesis leverages regularization techniques from the machine learning literature to develop new methodological approaches that address challenges in high-dimensional econometrics, particularly in contexts where traditional methods are less reliable. It is organized into three chapters.

The first chapter, co-authored with Endong Wang, presents a Wald test for multi-horizon Granger causality within a high-dimensional sparse Vector Autoregressive (VAR) framework. The null hypothesis focuses on a finite subset of causal coefficients of interest in the high-dimensional local projection (LP) equation at a given horizon. While OLS-based inference on LP coefficients suffers from the curse of dimensionality in this context, the post-double-selection method on LP may not be applicable, as a sparse VAR model does not necessarily imply a sparse LP for horizon $h > 1$. To address these limitations of standard approaches, we develop two types of de-biased estimators for the causal coefficients of interest, both relying on a first-step ℓ_1–regularized estimators of the VAR matrix coefficients. 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 at the cost of stronger assumptions. We further derive heteroskedasticity- and autocorrelation-consistent (HAC) inference for each estimator. In addition, we propose a robust inference method for the two-stage estimator that eliminates 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 size of the Wald test, particularly for longer projection horizons. We apply our methodology to analyze the interconnectedness of policy-related economic uncertainty across a large set of countries, both in the short and long run. Specifically, we construct a causal network to visualize how economic uncertainty spreads across countries over time.

The second chapter, co-authored with Endong Wang, introduces a parsimonious high-dimensional local projection (LP) framework to analyze volatility transmission

in stock returns across multiple horizons, while accommodating a large cross-section of assets. The proposed LP equation, motivated by the heterogeneous vector autoregressive (HVAR) model of realized volatilities, leverages the improved forecasting performance of volatility components realized over different time horizons—daily, weekly, and monthly—and is thus termed the *heterogeneous local projection (HLP)*. To address the challenge of standard asymptotic Gaussian inference in the high-dimensional HLP, we employ the de-biased least squares (deLS) estimation and inference procedure introduced in the first chapter, enabling valid inference without relying on a sparsity assumption in the HLP equation. Applying this framework to high-frequency data on 30 U.S. stocks from 2010 to 2014, we construct Granger-causal networks at multiple horizons—daily, weekly, biweekly, and monthly—and compare the performance of deLS-based inference with that of standard OLS-based methods. Our findings indicate that deLS produces sparser and more stable networks, particularly at longer horizons, where OLS tends to overfit and detect misleading connections. These results are robust to subsample analyses and remain consistent when realized correlations are included as additional controls, which helps further avoid the detection of spurious spillovers while substantially increasing the model's dimensionality.

The third chapter, co-authored with Marine Carrasco, provides a user-friendly approach for moment-based estimation in high-dimensional contexts where many moments/instruments are available. Several economic applications involve many instruments. This chapter addresses the issue of efficient estimation in such frameworks. Although using many moment conditions can improve the efficiency of GMM-type estimators, it may also result in substantial bias due to strong overidentification. We consider a specific setting where the parameter of interest is defined by a single conditional moment restriction, resulting in a large set of moment conditions. The benchmark estimator we consider is the continuous updating GMM estimator (CUE) due to its relatively low bias under many moment conditions. We introduce a Ridge-type regularized version of CUE (RRCUE) to address the singularity problem of the

weighting matrix under many moments. We show that the RRCUE estimator is consistent, asymptotically normal, and reaches the semiparametric efficiency bound under an asymptotic framework where both the sample size and the number of moment conditions go to infinity and the regularization parameter converges to zero at a certain rate. We propose a data-driven approach for selecting the optimal regularization parameter based on cross-validation criteria. We evaluate the performance of RRCUE through Monte Carlo simulations. Our results reveal that regularization reduces the dispersion problem of CUE and improves efficiency, although it introduces some bias that remains relatively low with a moderately large number of moments. In the specific linear instrumental variables framework, our estimator is shown to be competitive with some state-of-the-art estimators in the field. We apply our methodology to re-estimate the effect of the quality of institutions and government policies—the so-called *social infrastructure*—on output per worker. Our empirical results are consistent with simulation results, providing estimates with better precision.

Keywords: Multi-horizon Granger causality, high-dimensional VAR, regularization, de-biased estimation, high-dimensional heterogeneous VAR, heterogeneous local projection, networks of realized volatilities, many moment conditions or instrumental variables, continuous updating estimator, efficient estimation.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

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

1.1. Introduction

The 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 in-

⋆. This chapter is co-authored with Endong Wang. The authors are extremely grateful to Marine Carrasco, Benoit Perron, Mathieu Marcoux, Jean-Marie Dufour, and Victoria Zinde-Walsh for their helpful discussions and guidance. They 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.

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

formation (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) 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 coeffi-

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.

cients 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 de-biased/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. However, to the best of our knowledge, we are the first to investigate multi-horizon Granger causality testing within a high-dimensional VAR framework.

Our investigation of multi-horizon Granger causality in high-dimensional settings complements the growing literature on single-horizon Granger causality in large datasets, as illustrated by the works of Hecq et al. (2023) and Babii et al. (2024). While Adamek et al. (2024) examine de-biased 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 de-biased 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 de-biased 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). The multi-horizon Granger causality reveals the specific information that a given variable contributes to the forecast of a target outcome at various horizons.

This paper is structured as follows. Section 1.2 outlines the econometric framework. In Section 1.3, we review a range of regularized estimators in high-dimensional models. Section 1.4 introduces a de-biased Least Squares estimation method. Section 1.5 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 1.6. The results of Monte Carlo simulations are presented in Section 1.7. Section 1.8 provides an empirical application of our methods and visualizes the connectedness of country-level economic uncertainties. Finally,

Section 1.9 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 $r,s \in \mathbb{N}$. $\tilde{e}_{rj},\ j = 1,\ldots,r$ 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 $B = \left(b_{i,j}\right)_{i=1,\ldots,r,j=1,\ldots,s}$, $\|B\|_1 := \max_{1\le j\le s} \sum_{i=1}^{r} \left|b_{i,j}\right| = \max_{1\le j\le s} \left\|B\tilde{e}_{sj}\right\|_1$ 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.

1.2. Framework

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

$$w_t = A_1 w_{t-1} + A_2 w_{t-2} + \cdots + A_p w_{t-p} + u_t, \tag{1.1}$$

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 Lütke-

pohl (2017)), the VAR model can be written in a compact form

$$w_t = J\mathbf{A}W_{t-1} + u_t, \tag{1.2}$$

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,

$$\mathbf{A} = \begin{bmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I & & & & \\ & I & & & \\ & & \ddots & & \\ 0 & & & I & 0 \end{bmatrix}. \tag{1.3}$$

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]

$$\mathrm{P}_L(y_{t+h} \mid W_t) = \mathrm{P}_L(y_{t+h} \mid W_{-x,t}), \tag{1.4}$$

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,

$$w_{t+h} = J\mathbf{A}^h W_t + u_t^{(h)} \tag{1.5}$$

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

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

(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 (1.5) and write it in a generic way,

$$y_{t+h} = \beta_h' W_t + e_{t,h}, \tag{1.6}$$

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 (1.6) can be rewritten as

$$y_{t+h} = \beta_{1,h}' W_{1,t} + \beta_{2,h}' W_{2,t} + e_{t,h}, \tag{1.7}$$

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 (1.4) is stated as

$$\mathscr{H}_0 : \beta_{1,h} = 0. \tag{1.8}$$

Testing $\mathscr{H}_0$ involves estimating and making inferences about $\beta_{1,h}$. Estimating $\beta_{1,h}$ is a challenging exercise because Equation (1.7) includes the high-dimensional nuisance parameter $\beta_{2,h}$. For example, if $d = 20$ and $p = 4$, as in our empirical application,

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

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 (1.1), 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 (1.7) 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 (1.6) 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 (1.7) 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 (1.1) under a potentially non-sparse local projection equation (1.7). 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.

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 1.6, 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.

1.3. 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 1.2(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:

$$\hat{A}^{(re)}_{j\bullet,1:p} = \operatorname{argmin} \frac{1}{n-p} \sum_{t=p+1}^{n} \left\| w_t - \sum_{i=1}^{p} A_{j\bullet,i} w_{t-i} \right\|_2^2 + \lambda \sum_{i=1}^{p} \|A_{j\bullet,i}\Pi_i\|_1, \tag{1.9}$$

for $j = 1,2,\cdots,d$, where $A_{j\bullet,1:p}$ denotes the j-th row of slope coefficient matrices $A_{1:p} = [A_1,A_2,\cdots,A_p]$ 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, (1.9) reduces to a LASSO estimation equation. For instance, Belloni et al. (2012) considers a diagonal matrix representing a data-dependent penalty loadings for the self-normalization of the first-order con-

ditions 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) mentions 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 1.2(iii)), regardless of whether LASSO or one of its variants (adaLASSO or ElNet) is used.

1.4. De-biased least squares estimation

In this subsection, we consider the de-biased LS approach to identify and estimate the parameter of interest, $\beta_{1,h}$, in (1.6).

1.4.1. Least squares identification

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

$$\beta_h = \mathbb{E}[W_t W_t']^{-1}\mathbb{E}[W_t y_{t+h}], \tag{1.10}$$

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):

$$\mathbb{E}[W_t W_t'] = \sum_{j=0}^{\infty} \mathbf{A}^j J' \Sigma_u J (\mathbf{A}')^j = \mathrm{vec}_{dp}^{-1}\big((I_{d^2p^2} - \mathbf{A}\otimes\mathbf{A})^{-1}\mathrm{vec}(J'\Sigma_u J)\big) \tag{1.11}$$

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 the 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

$$\beta_{1,h} = \mathbb{E}[W_{1,t}^{\perp} W_{1,t}^{'}]^{-1}\mathbb{E}[W_{1,t}^{\perp} y_{t+h}], \tag{1.12}$$

where

$$W_{1,t}^{\perp} := W_{1,t} - \mathrm{P}_L(W_{1,t} \mid W_{2,t}) = W_{1,t} - \mathbb{E}[W_{1,t}W_{2,t}'] \Big(\mathbb{E}[W_{2,t}W_{2,t}'] \Big)^{-1} W_{2,t}. \quad (1.13)$$

In the low-dimensional case, practitioners will take the LS projection residual as an estimator of $W_{1,t}^{\perp}$ in line with (1.13) 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 (1.13) 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 (1.12) but relies on an alternative way to estimate the rotated regressor $W_{1,t}^{\perp}$.

1.4.2. 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 the high-dimensional setting is carried out using the following relation obtained from Equation (1.13) through block matrix inversion,

$$W_{1,t}^{\perp} = (R_1 \Sigma_W^{-1} R_1')^{-1} R_1 \Sigma_W^{-1} W_t. \quad (1.14)$$

Let $\hat{A}_{1:p}^{(re)}$ be the regularized estimators of $(A_1, \cdots, A_p)$ as defined in Equation (1.9). 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 (1.11) and compute $\hat{\Sigma}_W$, the estimate of Σ_W, by using $\hat{A}_{1:p}^{(re)}$ and $\hat{\Sigma}_u$.

Step 2: Following (1.14), we estimate the regressor $W_{1,t}^{\perp}$. Instead of estimat-

ing Σ_W through sample variance, we use the estimate $\hat{\Sigma}_W$ obtained from Step 1.

$$\hat{W}_{1,t}^{\perp} = (R_1 \hat{\Sigma}_W^{-1} R_1')^{-1} R_1 \hat{\Sigma}_W^{-1} W_t. \tag{1.15}$$

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

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

$$\hat{\beta}_{1,h}^{(LS)} = \left(\sum_t \hat{W}_{1,t}^{\perp} W_{1,t}' \right)^{-1} \left(\sum_t \hat{W}_{1,t}^{\perp} y_{t+h} \right). \tag{1.16}$$

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

$$\hat{\beta}_{1,h}^{(de-LS)} = \hat{\beta}_{1,h}^{(LS)} - \left(\sum_t \hat{W}_{1,t}^{\perp} W_{1,t}' \right)^{-1} \left(\sum_t \hat{W}_{1,t}^{\perp} W_{2,t}' \hat{\beta}_{2,h} \right) \tag{1.17}$$

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 1.1.

(i) In the low-dimensional setting, an estimator of $W_{1,t}^{\perp}$ is obtained by replacing the population covariance and variance in Equation (1.13) with their sample counterparts. Algebraically, this involves replacing Σ_W in (1.14) 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 (1.15), in which the sample covariance is estimated through the explicit formula presented in (1.11), with $\mathbb{A}$ replaced with its regularized estimator.

(ii) Equation (1.15) implicitly assumes that the sample covariance of Σ_W computed from (1.11) is non-singular. Implied by (1.11), the non-singularity of the sam-

ple covariance matrix estimate entails that

$$\lambda'\hat{\Sigma}_W\lambda = \lambda' J'\hat{\Sigma}_u J\lambda + \sum_{j=1}^{\infty}\lambda'(\hat{\mathbf{A}}^{(re)})^j J'\hat{\Sigma}_u J(\hat{\mathbf{A}}^{(re)})'^j\lambda > 0, \tag{1.18}$$

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

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

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 ver-

sion 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 debiasing 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

$$\mathbb{E}\left[\varphi_t^{ls}\left(\beta_{1,h}, \eta_0\right)\right] = 0, \tag{1.19}$$

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

$$\mathbb{E}\left[\psi_t^{dls}\left(\beta_{1,h}, \eta_0\right)\right] = 0, \tag{1.20}$$

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

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.

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 de-biased estimates has a well-shaped density similar to the standard Gaussian distribution; see Figure 1.1. (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 1.2. 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).

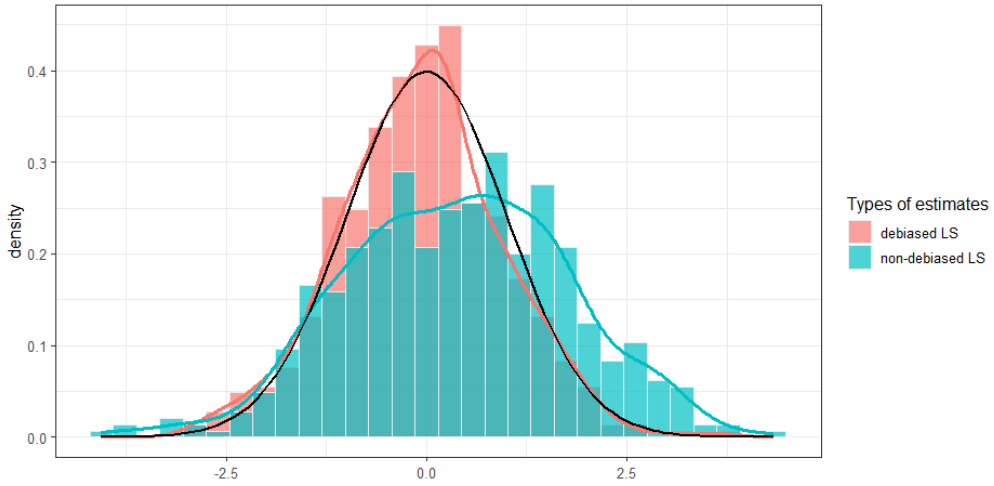

Figure 1.1 – The black curve represents the density of the standard Gaussian distribution, the red curve depicts the fitted density of the de-biased least squares (LS) estimate, and the green curve shows the fitted density of the non-debiased LS estimate. The coefficient of interest is the coefficient 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 1.3.

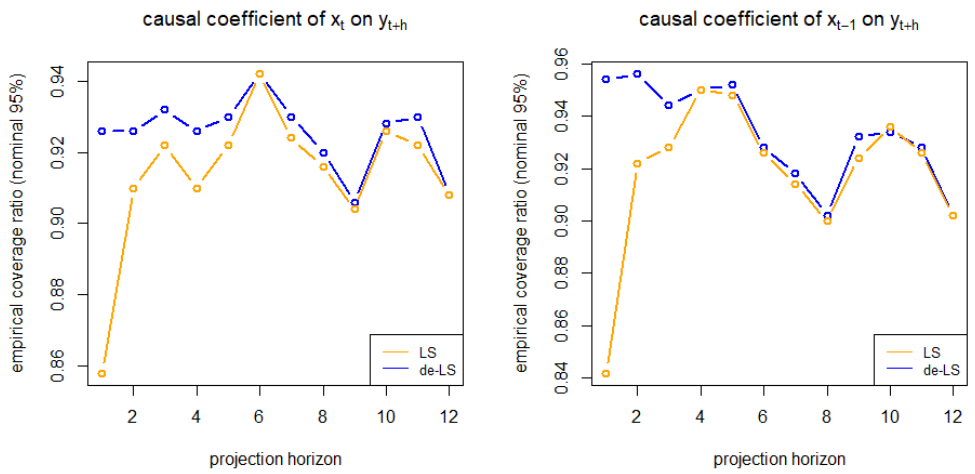

Figure 1.2 – Empirical coverage ratio for de-biased and non-debiased LS for the coefficient of x_t to y_{t+1}. The sample size is $n = 120$, the dimension of the VAR is $d = 60$, and the number of simulations is 500. The data generating process (DGP) is a VAR(2). The values of the VAR coefficients and the covariance matrix are determined in the same manner as in Figure 1.3.

1.4.3. Asymptotic variance of de-biased least squares estimator

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 1.2 and certain restrictions on the growth rate of the number of series d with respect to the sample size n (see Condition 1.1 below), we show (see Lemma 1.4 of the Appendix) that

$$\sqrt{n}\left(\hat{\beta}_{1,h}^{(de-LS)} - \beta_{1,h}\right) = \left(E\left[W_{1,t}^{\perp} W_{1,t}^{\prime}\right]\right)^{-1} \left(n^{-1/2} \sum_{t=p}^{n-h} W_{1,t}^{\perp} e_{t,h}\right) + o_p(1), \qquad (1.21)$$

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

$$\begin{aligned}\text{AVar}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-LS)}\right) &= \lim_{n\to\infty}\left(E\left[W_{1,t}^{\perp}W_{1,t}'\right]\right)^{-1}\text{Var}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}W_{1,t}^{\perp}e_{t,h}\right)\left(E\left[W_{1,t}W_{1,t}^{\perp\prime}\right]\right)^{-1} \\ &= \lim_{n\to\infty}(R_1\Sigma_W^{-1}R_1')\Omega_{W_1,h}(R_1\Sigma_W^{-1}R_1'),\end{aligned} \tag{1.22}$$

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

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

Analogous to conventional time series literature, one possible consistent estimator of the asymptotic variance $\text{AVar}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-LS)}\right)$ could be obtained by replacing each term within the expression on the right-hand side of the second equality of Equation (1.22) with its sample counterpart. Note that the long-run variance matrix $\Omega_{W_1,h}$ estimate may not be positive semi-definite 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 (1.15) 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

$$\widehat{\text{AVar}}^{(hac)}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-LS)}\right) = (R_1\hat{\Sigma}_W^{-1}R_1')\hat{\Omega}_{W_1,h}^{(hac)}(R_1\hat{\Sigma}_W^{-1}R_1'), \tag{1.23}$$

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

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 (1.22).

1.5. 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 two-stage identification strategy.

1.5.1. Two-stage identification

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

$$\mathrm{P}_L\left(y_{t+h} - \beta_h' W_t | U_t\right) = 0 \tag{1.24}$$

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

$$\beta_h = \mathbb{E}[U_t W_t']^{-1} \mathbb{E}[U_t y_{t+h}] \tag{1.25}$$

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,

$$\mathbb{E}[U_t W_t'] = (I_p \otimes \Sigma_u)\Psi(p), \tag{1.26}$$

$$\mathbb{E}[U_t y_{t+h}] = (I_p \otimes \Sigma_u)[\Psi_h', \Psi_{h+1}', \cdots, \Psi_{h+p-1}']'\nu_1, \tag{1.27}$$

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 the 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,

$$\beta_{1,h} = \mathbb{E}[U_{1,t}^{\perp} W_{1,t}^{'}]^{-1}\mathbb{E}[U_{1,t}^{\perp} y_{t+h}], \tag{1.28}$$

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,

$$\Gamma = \mathbb{E}[U_{1,t} W_{2,t}']\mathbb{E}[U_{2,t} W_{2,t}']^{-1} \tag{1.29}$$

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 ro-

tated U_t,[8]

$$U_{1,t}^{\perp} = [I, -\Gamma][U_{1,t}', U_{2,t}']' = \bar{\Gamma}_R U_t = (R_1 \Sigma_{UW}^{-1} R_1')^{-1} R_1 \Sigma_{UW}^{-1} U_t. \tag{1.30}$$

1.5.2. 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 (1.26) 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 (1.30):

$$\hat{U}_{1,t}^{\perp} = (R_1 \hat{\Sigma}_{UW}^{-1} R_1')^{-1} R_1 \hat{\Sigma}_{UW}^{-1} \hat{U}_t \tag{1.31}$$

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 $\beta_{1,h}$,

$$\hat{\beta}_{1,h}^{(2S)} = \left(\sum_t \hat{U}_{1,t}^{\perp} W_{1,t}'\right)^{-1} \left(\sum_t \hat{U}_{1,t}^{\perp} y_{t+h}\right). \tag{1.32}$$

8. Here is the matrix algebra to support (1.30):

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

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

$$\hat{\beta}_{1,h}^{(de-2S)} = \hat{\beta}_{1,h}^{(2S)} - \left(\sum_t \hat{U}_{1,t}^{\perp} W_{1,t}'\right)^{-1} \left(\sum_t \hat{U}_{1,t}^{\perp} W_{2,t}' \hat{\beta}_{2,h}\right) \tag{1.33}$$

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

Remark 1.2.

(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 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 the 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 de-biased 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.

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

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

$$\mathbb{E}\left[\psi_t^{d2s}\left(\beta_{1,h}, \eta_0\right)\right] = 0, \tag{1.34}$$

where [9]

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

$\Gamma_0:=E[U_{1,t}W_{2,t}']\left(E[U_{2,t}W_{2,t}']\right)^{-1}$, $\beta_{2,h}=R_2\left(\mathbf{A}^h\right)'J'e_y$, 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 (1.32) at the $\sqrt{n}$-level asymptotically. In the subsequent section, we will elaborate on how the estimation bias of the VAR residual has an asymptotically negligible effect on two-stage estimates in a high-dimensional framework.

1.5.3. Asymptotic variance of de-biased two-stage estimator

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,
$\sqrt{n}\left(\hat{\beta}_{1,h}^{(de-2S)}-\beta_{1,h}\right)$. Under Assumption 1.2 and certain restrictions on the growth rate of the number of time series d with respect to the sample size n, we show that

$$\sqrt{n}\left(\hat{\beta}_{1,h}^{(de-2S)}-\beta_{1,h}\right)=\left(\mathbb{E}\left[U_{1,t}^{\perp}W_{1,t}'\right]\right)^{-1}\left(n^{-1/2}\sum_{t=p}^{n-h}U_{1,t}^{\perp}e_{t,h}\right)+o_p(1), \tag{1.35}$$

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

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

$$\begin{aligned}\text{AVar}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-2S)}\right) &= \lim_{n\to\infty}\left(\mathbb{E}\left[U_{1,t}^{\perp}W_{1,t}'\right]\right)^{-1}\text{Var}\left(n^{-1/2}\sum_{t=p}^{n-h}U_{1,t}^{\perp}e_{t,h}\right)\left(\mathbb{E}\left[U_{1,t}^{\perp}W_{1,t}'\right]\right)^{'-1} \\ &= \lim_{n\to\infty}(R_1\Sigma_{UW}^{-1}R_1')\Omega_{U_1,h}(R_1\Sigma_{UW}^{'-1}R_1'),\end{aligned} \tag{1.36}$$

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

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

Analogous to conventional time series literature, one possible consistent estimate of the variance $\text{AVar}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-2S)}\right)$ could be obtained by replacing each term within the expression on the right-hand side of the second equality of Equation (1.36) with its sample counterpart. One advantage of the two-stage estimation method is to obviate the 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,

$$\widehat{\text{AVar}}^{(hac)}\left(\sqrt{n}\hat{\beta}_{1,h}^{(de-2S)}\right) = (R_1\hat{\Sigma}_{UW}^{-1}R_1')\hat{\Omega}_{U_1,h}^{(hac)}(R_1(\hat{\Sigma}_{UW}^{-1})'R_1'), \tag{1.37}$$

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 $\hat{e}_{t,h} = y_{t+h} - \hat{\beta}_h W_t$ and $\hat{\beta}_h = \nu_1'\left(\hat{\mathbf{A}}^{(re)}\right)^h$, and $\hat{U}_{1,t}^{\perp}$ comes from (1.31).

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 semi-definite 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 semi-definite.

By (1.30), $U_{1,t}^{\perp} = (R_1\Sigma_{UW}^{-1}R_1')^{-1}R_1\Sigma_{UW}^{-1}U_t$. Then, the long-run variance matrix $\Omega_{U_1,h}$ can be rewritten as

$$\Omega_{U_1,h} = (R_1\Sigma_{UW}^{-1}R_1')^{-1}R_1\Sigma_{UW}^{-1}\Omega_{U,h}\Sigma_{UW}^{\prime-1}R_1'(R_1\Sigma_{UW}^{-1}R_1')^{\prime-1} \tag{1.38}$$

by defining

$$\Omega_{U,h} := \left(\lim_{n\to\infty}\sum_{k=-\infty}^{\infty}\mathbb{E}[U_tU_{t+k}'e_{t,h}e_{t+k,h}]\right). \tag{1.39}$$

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 $U_te_{t,h}$,

$$U_te_{t,h} = (u_t', u_{t-1}', \cdots, u_{t-p+1}')'e_{t,h}. \tag{1.40}$$

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

$$s_t := (e_{t,h}, e_{t+1,h}, \cdots, e_{t+p-1,h})' \otimes u_t. \tag{1.41}$$

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:

$$\sum_{k=-\infty}^{\infty} \mathbb{E}[s_t s_{t+k}'] = \Omega_{U,h}. \tag{1.42}$$

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,

$$\begin{aligned}
\mathbb{E}[s_t s_\tau'] &= \mathbb{E}[\mathbb{E}[s_t s_\tau \mid u_{t+1}, u_{t+2}, \cdots]] \\
&= \mathbb{E}[(e_{t,h}, e_{t+1,h}, \cdots, e_{t+p-1,h}) \otimes \mathbb{E}[u_t \mid u_{t+1}, u_{t+2}, \cdots] s_\tau'] \\
&= 0 \\
&\quad (\text{since } \mathbb{E}[u_t \mid u_{t+1}, u_{t+2}, \cdots] = 0 \text{ by i.i.d. assumption})
\end{aligned} \tag{1.43}$$

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 1.1. *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 1.1(*i*) constrains u_t to be a martingale difference sequence, a common condition in the time series literature. Assumption 1.1(*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 1.1 for the serial uncorrelation of the process s_t.

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

$$\mathbb{E}[s_t s_\tau'] = \mathbb{E}[(e_{t,h}, e_{t+1,h}, \cdots, e_{t+p-1,h})(e_{\tau,h}, e_{\tau+1,h}, \cdots, e_{\tau+p-1,h})' \otimes u_t u_\tau']. \tag{1.44}$$

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

$$\mathbb{E}[s_t s_\tau'] = \mathbf{0} \quad \Leftrightarrow \quad \mathbb{E}[u_t u_\tau' \otimes e_{t+i,h} e_{\tau+j,h}] = \mathbf{0}, \tag{1.45}$$

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,

$$\begin{aligned}
&\mathbb{E}[u_t u_\tau' \otimes e_{t+i,h} e_{\tau+j,h}] \\
=&\mathbb{E}\left[u_t u_\tau' \otimes \left(\sum_{m=1}^{h} v_1' \Psi_{h-m} u_{t+i+m} \right) \left(\sum_{n=1}^{h} v_1' \Psi_{h-n} u_{\tau+j+n} \right) \right] \\
=&\mathbb{E}\left[u_t u_\tau' \otimes v_1' \left(\sum_{m=1}^{h} \sum_{n=1}^{h} \Psi_{h-m} u_{t+i+m} u_{\tau+j+n}' \Psi_{h-n}' \right) v_1 \right]
\end{aligned} \tag{1.46}$$

Without loss of generality, suppose $t < \tau$. If $t+i+m = \tau+j+n$, then Assumption 1.1 (ii) ensures that

$$\mathbb{E}\left[u_t u_\tau' \otimes (v_1' \Psi_{h-m} u_{t+i+m} u_{\tau+j+n}' \Psi_{h-n}' v_1) \right] = 0, \tag{1.47}$$

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

If $t+i+m < \tau+j+n$, then Assumption 1.1 (i) with Law of Iterated Expectation (LIE) ensures that

$$\mathbb{E}\left[u_t u_\tau' \otimes (v_1' \Psi_{h-m} u_{t+i+m} u_{\tau+j+n}' \Psi_{h-n}' v_1) \right] = 0, \tag{1.48}$$

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

If $t+i+m > \tau+j+n$, similarly, Assumption 1.1 (i) with Law of Iterated Expectation (LIE) ensures that

$$\mathbb{E}\left[u_t u_\tau' \otimes (v_1' \Psi_{h-m} u_{t+i+m} u_{\tau+j+n}' \Psi_{h-n}' v_1) \right] = 0, \tag{1.49}$$

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

Therefore, combining the results of (1.47)-(1.49), Assumption 1.1 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 (1.42), Assumption 1.1 yields the matrix $\Omega_{U,h}$ equals to the

covariance matrix of s_t,

$$\Omega_{U,h} = \text{Var}(s_t). \tag{1.50}$$

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

$$\hat{\Omega}^{(HC)}_{U_1,h} = (R_1\hat{\Sigma}^{-1}_{UW}R_1')^{-1}R_1\hat{\Sigma}^{-1}_{UW}\widehat{\text{Var}}(\hat{s}_t)\hat{\Sigma}^{\prime-1}_{UW}R_1'(R_1\hat{\Sigma}^{-1}_{UW}R_1')^{\prime-1} \tag{1.51}$$

where $\widehat{\text{Var}}(\hat{s}_t) = \frac{1}{n-h}\sum_t \hat{s}_t\hat{s}_t'$, $\hat{s}_t = (\hat{e}_{t,h}, \hat{e}_{t+1,h}, \cdots, \hat{e}_{t+p-1,h}) \otimes \hat{u}_t$, $\hat{u}_t = w_t - \hat{\Phi}^{(re)}_{1:p}W_{t-1}$, 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

$$\widehat{\text{AVar}}^{(HC)}(\sqrt{n}\hat{\beta}^{(de-2S)}_{1,h}) = R_1\hat{\Sigma}^{-1}_{UW}\widehat{\text{Var}}(\hat{s}_t)\hat{\Sigma}^{\prime-1}_{UW}R_1'. \tag{1.52}$$

1.6. 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 1.1 and 1.2 below). We then derive the asymptotic inference for the estimators. In particular, we propose a HAC standard error for both de-biased estimators. Additionally, we propose an HC standard error for the de-biased 2S estimator. Regarding the regularity conditions on the innovation process, the consistency of the HC standard error requires a slightly stronger assumption (see Assumption 1.1), which can be viewed as a cost for the convenience of avoiding the HAC standard error.

1.6.1. 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],

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

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., Adamek et al., 2023; Basu and Michailidis, 2015). 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 1.2 below. Assumption 1.2 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 1.2 is partially

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).

similar to Assumption 1 of Krampe et al. (2023) in deriving the consistency of the (Lasso) regularized estimator of the structural impulse response.

Assumption 1.2.

(i) ***Row-wise and Column-wise Approximate Sparsity:*** $\mathbf{A} \in \mathscr{U}(k_A, \mu)$ *and* $\mathbf{A}' \in \mathscr{U}(k_A, \mu)$ *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* $m \in \mathbb{N}$,

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

(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* $\|U\|_2 = 1 = \|V\|_2$,

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

(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* $\left\|\Sigma_W^{-1}\right\|_\infty = O(k_W)$ *and* $\left\|\Sigma_{UW}^{-1}\right\|_\infty = O(k_{UW})$, *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 1.2(*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 1.2(*ii*) specifies the standard stability condition of the VAR system. This assumption implies, in part, that the process $\{W_t, t \in \mathbb{Z}\}$ possesses a geometrically decaying functional dependence coefficient.

Assumption 1.2(*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 $\mathbf{A}$ (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 $k_A^{1.5} v_n^{(1-\nu)/2}$ tends to be large compared to $n^{(1-\nu)/2}$. The term $v_n := \nu(d, p, q, n)$, where ν is an increasing function of d. Its specific form depends on the regularization approach used for estimating $\mathbf{A}$, as well as on the number of finite moments q of the innovations u_t. As v_n lowers the convergence speed of $\hat{\mathbf{A}}^{(re)}$ to $\mathbf{A}$, it can be thought of as the cost of using regularization to estimate the high-dimensional object $\mathbf{A}$. 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 $v_n = \log(dp) + (ndp)^{2/q}$, and in particular, $v_n = \log(dp)$ 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 1.2(*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 (1.9), then a suitable candidate satisfying Assumption 1.2(*iii*) is the Thresholded Adaptive LASSO, denoted by $\hat{\mathbf{A}}^{(\text{thr})}$,

which is given by

$$\hat{A}_k^{(\text{thr})} = \text{THR}_\lambda(\hat{A}_k^{(re)}) := \left(\text{THR}_\lambda(\hat{A}_{ij,k}^{(re)})\right)_{i,j=1,2,\cdots,d}, \quad k = 1,2,\cdots,p. \tag{1.53}$$

where $\text{THR}_\lambda(z) = z(1-|\lambda/z|^\nu)_+$ 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 1.2(*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 1.2(*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 1.2(*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 $\phi < q(1-3\psi/(1-\mu))/4 - 1$. These conditions imply, in the case of exact sparsity ($\mu = 0$), that $\psi < (1-4/q)/3$ and $\phi < q(1-3\psi)/4 - 1$ 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 1.2(*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 1.2(*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 samples, 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 1.2(*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 1.2(*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 1.1 (**Consistency results**). *Let $\widehat{\Sigma}_W$ and $\widehat{\Sigma}_{UW}$ denote the regularized estimator of Σ_W and Σ_{UW}, respectively, using explicit formulas* (1.11) *and* (1.26). *Under Assumption 1.2, the following assertions are true:*

$$
\begin{aligned}
&(i) \quad \left\|\widehat{\Sigma}_u - \Sigma_u\right\|_\infty = O_p\left(d\left[k_A^3(\nu_n/n)^{1-\mu} + \sqrt{\tilde{\nu}_n/n}\right]\right);\\
&(ii) \quad \left\|\widehat{\Sigma}_W - \Sigma_W\right\|_\infty = O_p\left(dk_A^2\left\{k_A^{2.5}(\nu_n/n)^{(1-\mu)/2} + \sqrt{\tilde{\nu}_n/n}\right\}\right);\\
&(iii) \quad \left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_\infty = O_p\left(dk_A\left\{k_A^{1.5}(\nu_n/n)^{(1-\mu)/2} + \sqrt{\tilde{\nu}_n/n}\right\}\right).
\end{aligned}
$$

Remark 1.3. 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.

1.6.2. 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 (1.17). We show that the de-biased LS estimator $\hat{\beta}_{1,h}^{(de-LS)}$ is asymptotically normal and derive the consistency of its variance estimator defined by Equation (2.14) under certain restrictions on the growth rate of d relative to n (see Conditions 1.1 and 1.2). Importantly, we show under the conditional homoskedastic martingale difference sequence (m.d.s.) [11] assumption that the asymptotic variance defined by Equation (1.22) 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 1.1 and 1.2 implicitly restrict the growth rate of d relative 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 1.2(*iii*).

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

Assumption 1.3 (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 1.3 is a standard regularity condition on the time-dependence of the

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

VAR innovation process and observables. The strong mixing condition (Assumption 1.3(i)) and the boundedness of moments (Assumption 1.3(ii)) are important for asymptotic normality. Moreover, Assumption 1.3 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 1.3(ii) just requires the existence of more than 4 moments, as in Assumption 1.2(v). Furthermore, Assumption 1.3 is less restrictive than the i.i.d. innovation assumption imposed, 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 1.1. $\tilde{\nu}_n^{1/2} d k_A^2 k_W \left\{ k_A^{2.5} (\nu_n/n)^{(1-\mu)/2} + (\tilde{\nu}_n/n)^{1/2} \right\} = o(1)$.

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 (1.21). The following theorem establishes the asymptotic normality of the de-biased LS estimator.

Theorem 1.2 (**Asymptotic normality of the de-LS estimator**). *Under Assumptions 1.2 and 1.3, if the number of series d grows with n such that Condition 1.1 is satisfied, then for any vector $v \in \mathbb{R}^p$ such that $\|v\|_1 = 1$,*

$$\frac{\sqrt{n} v' (\hat{\beta}_{1,h}^{(de-LS)} - \beta_{1,h})}{s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v)} \xrightarrow{d} \mathcal{N}(0,1), \quad as \quad n \to \infty, \tag{1.54}$$

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

Condition 1.2. $k_W^2 \left\{ d k_A^2 \left[k_A^{2.5} (\nu_n/n)^{(1-\mu)/2} + (\tilde{\nu}_n/n)^{1/2} \right] + (\bar{\nu}_n/n)^{1/2} \right\} = o(1)$.

Condition 1.2 imposes an additional restriction to obtain consistency of the variance estimator for de-biased LS, as stated by the following theorem.

Theorem 1.3 (**Consistency of the variance estimator for de-LS**). *Under Assumptions 1.2 and 1.3, if the number of series d grows with n such that Conditions 1.1 and 1.2 are satisfied, then for any vector* $v \in \mathbb{R}^p$ *such that* $\|v\|_1 = 1$,

$$\left| \widehat{s.e.}_{\hat{\beta}_{1,h}^{(de-LS)}}^{(hac)}(v)^2 - s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v)^2 \right| \xrightarrow{p} 0, \tag{1.55}$$

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

Furthermore, the result (1.54) *in Theorem 1.2 still holds if* $s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v)$ *is replaced by its estimator,* $\widehat{s.e.}_{\hat{\beta}_{1,h}^{(de-LS)}}^{(hac)}(v)$.

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

$$s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v)^2 + o(1) = \sum_{j,l=0}^{h-1} e_y' \Psi_j \Sigma_u \Psi_l' e_y v' R_1 \Sigma_W^{-1} \Sigma_W (l-j) \Sigma_W^{-1} R_1' v, \tag{1.56}$$

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

Corollary 1.4 (**Limiting distribution of the Wald test statistic**). *If Assumptions 1.2 and 1.3 hold and the number of series d grows with n such that Conditions 1.1 and 1.2*

12. The notation $\Sigma_W(r)$ in (1.56) refers to the lag-r autocovariance matrix of W_t and has the closed-form representation:

$$\Sigma_W(r) := E[W_t W_{t-r}'] = \sum_{j=0}^{\infty} \mathbf{A}^{j+r} J' \Sigma_u J (\mathbf{A}')^j = \mathbf{A}^r \operatorname{vec}_{dp}^{-1}\left((I_{d^2p^2} - \mathbf{A} \otimes \mathbf{A})^{-1} \operatorname{vec}(J' \Sigma_u J) \right),$$

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

are satisfied, then under the null hypothesis $\mathscr{H}_0 : \beta_{1,h} = 0$*, the Wald test statistic*

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

1.6.3. 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 (1.33). We demonstrate that the de-biased 2S estimator $\hat{\beta}_{1,h}^{(de-2S)}$ is asymptotically normal and derive the consistency of the HAC variance estimator, defined by Equation (1.37), under certain restrictions on the growth rate of d relative to n (see Conditions 1.3 and 1.4). Furthermore, we establish the consistency of the HC variance estimator, defined by (1.52), under additional restrictions on the structure of the innovations u_t (see Assumptions 1.1 and 1.4). Note that Conditions 1.3 and 1.4 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 1.3. $\tilde{\nu}_n^{1/2} d k_A k_{\text{UW}} \left\{ k_A^{1.5} (\nu_n/n)^{(1-\mu)/2} + (\tilde{\nu}_n/n)^{1/2} \right\} = o(1)$.

Condition 1.3 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 (1.35). The following theorem demonstrates the asymptotic normality of the de-biased 2S estimator.

Theorem 1.5 (**Asymptotic normality of the de-2S estimator**)**.** *Under Assumptions 1.2 and 1.3, if the number of series* d *grows with* n *such that Condition 1.3 is satisfied, then for any vector* $v \in \mathbb{R}^p$ *such that* $\|v\|_1 = 1$*,*

$$\frac{\sqrt{n}v'(\hat{\beta}_{1,h}^{(de-2S)} - \beta_{1,h})}{s.e._{\hat{\beta}_{1,h}^{(de-2S)}}(v)} \xrightarrow{d} \mathcal{N}(0,1), \quad as \quad n \to \infty, \tag{1.57}$$

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 1.4. $k_W^2 \left\{ dk_A \left[k_A^{1.5} \left(v_n/n \right)^{(1-\mu)/2} + \sqrt{\tilde{v}_n/n} \right] \right\} = o(1)$.

Condition 1.4 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 1.2 and 1.3 are sufficient to ensure the consistency of the HAC-type estimator of the asymptotic variance of the de-biased LS, as defined by Equation (1.37), the consistency of the HC-type variance estimator, as defined by Equation (1.52), requires stronger moment restrictions on the innovation process, as stated by Assumption 1.4 below.

Assumption 1.4 (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 1.3(i).*

Assumption 1.4 is a stronger version of the regularity condition in Assumption 1.3(*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 (1.41).

Theorem 1.6 (**Consistency of the variance estimators for de-2S**)**.** *Under Assumptions 1.2 and 1.3, if the number of series* d *grows with* n *such that Conditions 1.3 and 1.4 are satisfied, then for any vector* $v \in \mathbb{R}^p$ *such that* $\|v\|_1 = 1$*,*

$$\left| \widehat{s.e.}^{(hac)}_{\hat{\beta}_{1,h}^{(de-2S)}}(v)^2 - s.e._{\hat{\beta}_{1,h}^{(de-2S)}}(v)^2 \right| \xrightarrow{p} 0, \tag{1.58}$$

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

Furthermore, the result (1.57) *in Theorem* 1.5 *still holds if* $s.e._{\hat{\beta}_{1,h}^{(de-2S)}}(v)$ *is replaced by its estimator,* $\widehat{s.e.}^{(hac)}_{\hat{\beta}_{1,h}^{(de-2S)}}(v)$.

If, in addition, the VAR *error term* u_t *satisfies Assumptions* 1.1 *and* 1.4*, then*

$$\left| \widehat{s.e.}^{(HC)}_{\hat{\beta}_{1,h}^{(de-2S)}}(v)^2 - s.e._{\hat{\beta}_{1,h}^{(de-2S)}}(v)^2 \right| \xrightarrow{p} 0, \tag{1.59}$$

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

Corollary 1.7 (Limiting distribution of the Wald test statistic). *If Assumptions* 1.2 *and* 1.3 *hold and the number of series* d *grows with* n *such that Conditions* 1.3 *and* 1.4 *are satisfied, then under the null hypothesis* $\mathscr{H}_0 : \beta_{1,h} = 0$*, the Wald test statistic*

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

1.7. 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:

$$(I - \Lambda_1 L)(I - \Lambda_2 L) w_t = u_t \tag{1.60}$$

where L is the lag operator, Λ_k is the root matrix.

Inspired by the literature on high-dimensional VAR, e.g., Miao et al. (2023), we consider two types of root matrices:

(i) DGP 1 (Tridiagonal root matrix): $\Lambda_{ij,k} = \rho^{|i-j|+1}$, with $\rho = 0.31$.

(ii) DGP 2 (Random root matrix): Λ_k is a random $d \times d$ matrix generated as follows. For each row i, the diagonal element $\Lambda_{ii,k}$ is set to a constant $\delta = 0.3$,

while four off-diagonal positions $j \neq i$ are selected uniformly at random without replacement from $\{1, \ldots, d\} \setminus \{i\}$. At these selected positions, the entries $\Lambda_{ij,k}$ are independently drawn from a continuous uniform distribution on the interval $[-0.2, 0.2]$.

Once the root matrices are determined, the VAR slope coefficients can be obtained as

$$\begin{aligned} A_1 &= \Lambda_1 + \Lambda_2, \\ A_2 &= -\Lambda_1 \Lambda_2, \end{aligned} \tag{1.61}$$

In addition, the error term u_t is assumed to follow a standard normal distribution [13], that is,

$$u_t \sim i.i.d. \mathcal{N}(0, \Sigma_u), \quad \text{with} \quad \Sigma_{u,ij} = 0.5^{|i-j|}.$$

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

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 horizons 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 [14] and 'bartlett' kernel function. We use '*HDeconometrics*' package and '*glmnet*' package in R program to implement adaptive LASSO. The penalty coefficient for adaptive LASSO is chosen by the Bayesian information criterion (BIC). Below, we present results only for tridiagonal root matrices (DGP 1). The results for random root matrices (DGP 2) are qualitatively similar, with only minor differences (see Appendix Figures 1.11, 1.12,

13. Simulation results for alternative specifications of the VAR innovations u_t—namely, a centered multivariate log-normal distribution and a Student-t distribution—are available upon request.

14. Results can be improved using automatic bandwidth selection (see, e.g., Andrews, 1991), but this improvement comes with a computational cost.

and 1.13).

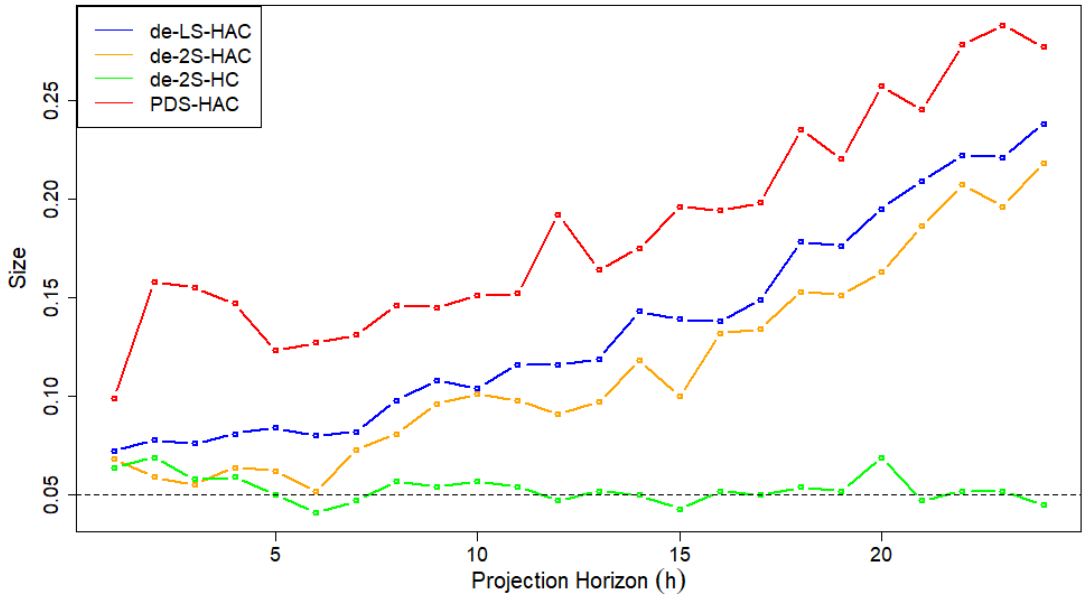

Figure 1.3 – Size of the Wald test at the 5% nominal level for different horizons. The red, blue, orange, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased two-stage with HAC standard errors, 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 $h = 1, \ldots, 24$. The number of replications is $1,000$.

Figure 1.3 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 infer-

ence 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 1.6 and 1.7 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.

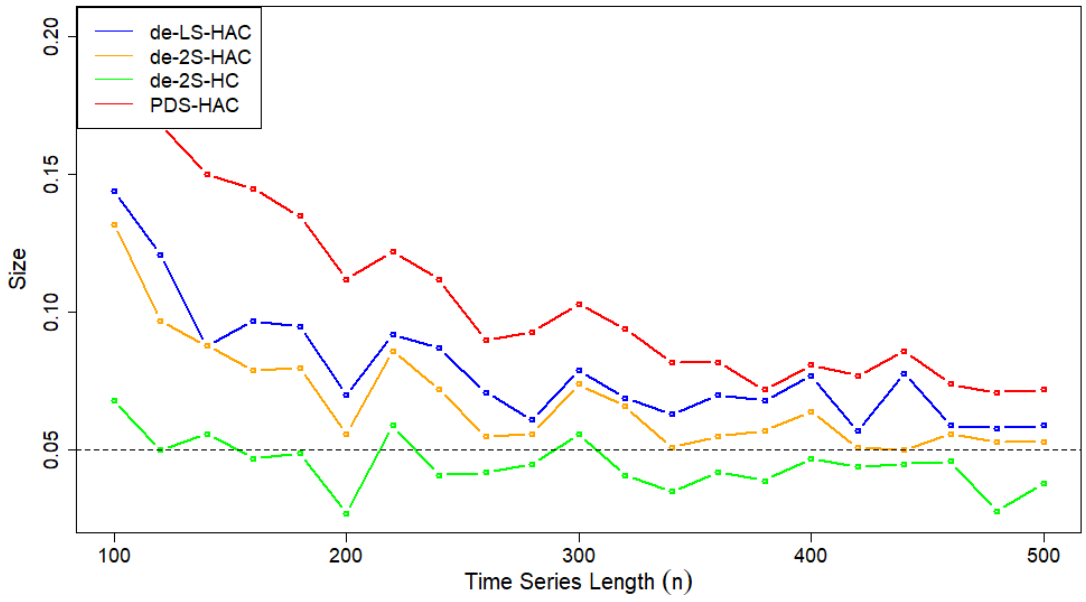

Figure 1.4 – Size of the Wald test at the 5% nominal level for different sample sizes and a given horizon ($h = 12$). The red, blue, orange, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased 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 1.4 compares the size of the Wald test across different ap-

proaches 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 1.8 to 1.10.

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.

1.8. 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 debiased estimator. The tests are performed at a 10% significance level, and critical values are taken from $\chi^2(df)$, with $df = 4$. In Figure 1.5, 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 1.5 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.

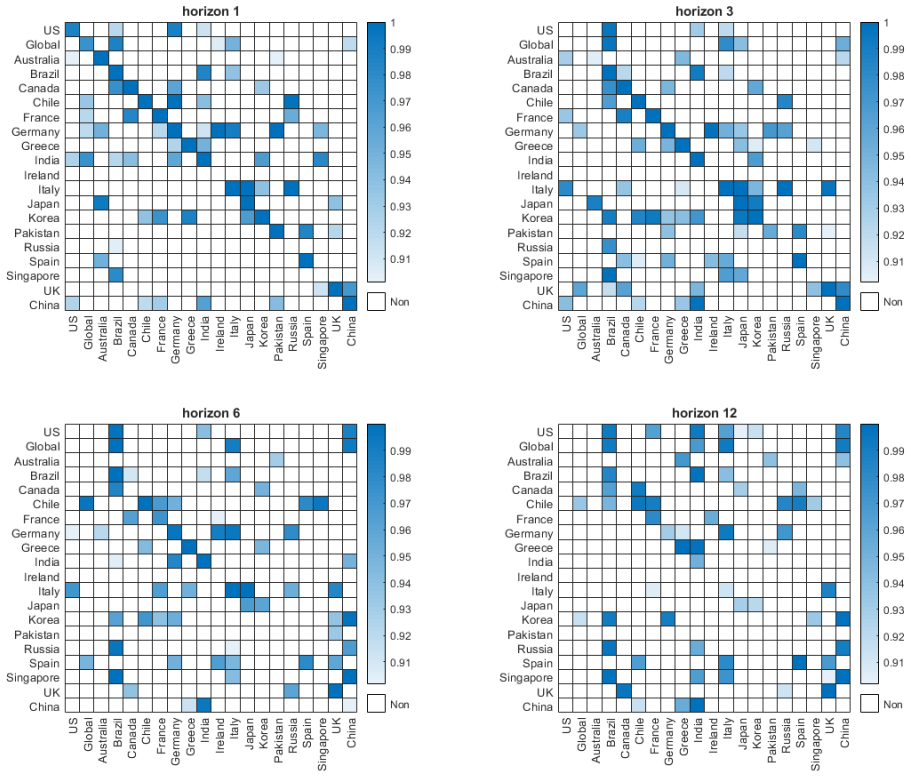

Figure 1.5 – Granger causality test at different horizons. For each cell, the darkness of the color is $1-p$-value and represents the strength of the Granger causality from the column variable to the row variable. Data: 2003:01 - 2024:02 time span, 254 number of observations, and 20 countries. VAR estimation method: adaptive LASSO on VAR(4). Causality test: two-stage estimation, Wald test with critical value from $\chi^2(df)$, $df = 4$.

1.9. 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) recognizes 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.

Bibliography

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.

D. W. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, pages 817–858, 1991.

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.

1.10. Appendix

1.10.1. 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:

$$\left\|B_1'\right\|_{\max} = \|B_1\|_{\max}, \quad \left\|B_1'\right\|_{\infty} = \|B_1\|_1, \quad \|B_1B_2\|_{\max} \leq \|B_1\|_{\infty}\|B_2\|_{\max}, \quad \|B_1x\|_l \leq \|B\|_l\|x\|_l$$
$$\|B_1B_2\|_l \leq \|B_1\|_l\|B_2\|_l \quad \text{for } l \in \{1,2,\infty\}, \quad \text{and} \quad \|UB_1V\|_2 = \|B_1\|_2 \quad \text{if } \|U\|_2 = \|V\|_2 = 1$$

Proof of Theorem 1.1. 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

$$I_3 := \left\|\frac{1}{n-p}\sum_{t=p+1}^{n} u_t u_t' - \Sigma_u\right\|_{\max} = O_p\left(\sqrt{\tilde{\nu}_n/n}\right) \quad \text{by Assumption 1.2}(iv),$$
$$I_1 = \left\|\frac{1}{n-p}\sum_{t=p+1}^{n} (\hat{u}_t - u_t)(\hat{u}_t - u_t)\right\|_{\max} \quad \text{and} \quad I_2 = \left\|\frac{1}{n-p}\sum_{t=p+1}^{n} (\hat{u}_t - u_t)u_t'\right\|_{\max}.$$

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 1.1 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 from Assumption 1.2*(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

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

and

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

Therefore,

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

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 1.2*(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 from 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 from 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 $\left\|\tilde{e}_{p(j+1)} \otimes I_d\right\|_\infty = 1$,

$$\left\|\hat{\Sigma}_{UW} - \Sigma_{UW}\right\|_\infty \leqslant \sum_{j=0}^{p-1}\left\|\hat{\Sigma}_u\hat{\Upsilon}_j' - \Sigma_u\Upsilon_j'\right\|_\infty. \tag{1.62}$$

Moreover, by Assumption 1.2*(v)*

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

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.(1.62) are of the same order for all j, then by T,

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

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

$$\left\|\hat{\Sigma}_W-\Sigma_W\right\|_\infty=\left\|\sum_{m=0}^{\infty}\hat{\Upsilon}_m\hat{\Sigma}_u\hat{\Upsilon}_m'-\sum_{m=0}^{\infty}\Upsilon_m\Sigma_u\Upsilon_m'\right\|_\infty\leq\gamma_1+\gamma_2+\gamma_3+\gamma_4+\gamma_5+\gamma_6+\gamma_7, \tag{1.63}$$

where,

$$\begin{aligned}\gamma_1 &:= \left\|\sum_{m=0}^{\infty}\left(\hat{\Upsilon}_m-\Upsilon_m\right)\left(\hat{\Sigma}_u-\Sigma_u\right)\left(\hat{\Upsilon}_m'-\Upsilon_m'\right)\right\|_\infty \\ &\leq \left\|\hat{\Sigma}_u-\Sigma_u\right\|_\infty\left(\sum_{m=0}^{\infty}\left\|\hat{\Upsilon}_m-\Upsilon_m\right\|_\infty\right)\left(\sum_{m=0}^{\infty}\left\|\hat{\Upsilon}_m'-\Upsilon_m'\right\|_\infty\right)=O_p\left(\tilde{a}_nk_A^7(\nu_n/n)^{1-\mu}\right), \\ \gamma_2 &:= \left\|\sum_{m=0}^{\infty}\left(\hat{\Upsilon}_m-\Upsilon_m\right)\left(\hat{\Sigma}_u-\Sigma_u\right)\Upsilon_m'\right\|_\infty \\ &\leqslant \left\|\hat{\Sigma}_u-\Sigma_u\right\|_\infty\left(\sum_{m=0}^{\infty}\left\|\hat{\Upsilon}_m-\Upsilon_m\right\|\right)\left(\sum_{m=0}^{\infty}\left\|\Upsilon_m'\right\|_\infty\right)=O_p\left(\tilde{a}_nk_A^{4.5}(\nu_n/n)^{(1-\mu)/2}\right), \\ \gamma_3 &:= \left\|\sum_{m=0}^{\infty}\left(\hat{\Upsilon}_m-\Upsilon_m\right)\Sigma_u\left(\hat{\Upsilon}_m'-\Upsilon_m'\right)\right\|_\infty=O_p\left(dk_A^7(\nu_n/n)^{1-\mu}\right), \\ \gamma_4 &:= \left\|\sum_{m=0}^{\infty}\left(\hat{\Upsilon}_m-\Upsilon_m\right)\Sigma_u\Upsilon_m'\right\|_\infty=O_p\left(dk_A^{4.5}(\nu_n/n)^{(1-\mu)/2}\right), \\ \gamma_5 &:= \left\|\sum_{m=0}^{\infty}\Upsilon_m\left(\hat{\Sigma}_u-\Sigma_u\right)\left(\hat{\Upsilon}_m'-\Upsilon_m'\right)\right\|_\infty=O_p\left(\tilde{a}_nk_A^{4.5}(\nu_n/n)^{(1-\mu)/2}\right),\end{aligned}$$

$$\gamma_6 := \left\| \sum_{m=0}^{\infty} \Upsilon_m \left(\hat{\Sigma}_u - \Sigma_u \right) \Upsilon_m' \right\|_\infty = O_p\left(\tilde{a}_n k_A^2 \right),$$
$$\gamma_7 := \left\| \sum_{m=0}^{\infty} \Upsilon_m \Sigma_u \left(\hat{\Upsilon}_m' - \Upsilon_m' \right) \right\|_\infty = O_p\left(d k_A^{4.5} \left(\nu_n / n \right)^{(1-\mu)/2} \right).$$

Plugging derived rates into (1.63) 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 1.1 (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 1.2(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 1.1. See Appendix A of Krampe et al. (2023). □

Lemma 1.2. *Under Assumptions 1.2(ii), and (iv)-(vi), it holds true that:*

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

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

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

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

Proof of Lemma 1.2. 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 1.2(ii)), the following VAR(∞) representation:

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

Note that the same stability condition ensures that the filter $\left\{\Upsilon_j, j=0,1,\ldots,\infty\right\}$ 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

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

and

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

Then, Lemma 1.1 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 1.2(v) implies that $\|\Sigma_W\|_{\max}=O(1)$ and the second result in part (a) follows from 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 1.2(vi)). The result follows from Lemma 1.1 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 Lemma1.1 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 Lemma1.1 due to the stability condition and the fact that $\|J\|_2 = 1$ and $\left\|\Sigma_W^{-1}\right\|_2 = O(1)$ (see Assumption 1.2(vi)). □

Lemma 1.3. *Let*

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

Under Assumptions 1.2(ii), and (iv)-(vi), it holds true that:

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

Proof of Lemma 1.3. First of all, it worth noting that part (b) of Lemma 1.2 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 $\|R_1\|_\infty = 1$, $\left\|\hat{\Sigma}_W^{-1}\right\|_\infty = O_p(k_W)$ (see Assumption 1.2(vi)) and $A^{-1}-B^{-1} = A^{-1}(B-A)B^{-1}$ that

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

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 the 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 1.2. It then follows from T and the 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 1.4. *If Assumption 1.2 is satisfied, then for any vector $\nu\in\mathbb{R}^p$ such that $\|\nu\|_1 = 1$,*

$$\begin{aligned}\sqrt{n}\nu'\left(\hat{\beta}_{1,h}^{(de-LS)}-\beta_{1,h}\right) &= \nu'\left(E\left[W_{1,t}^{\perp}W_{1,t}'\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}W_{1,t}^{\perp}e_{t,h}\right) \\ &+ O_p\left(\tilde{\nu}_n/\sqrt{n}+\left\|\hat{\Sigma}_W-\Sigma_W\right\|_\infty k_W\sqrt{\tilde{\nu}_n}+\left\|\hat{\beta}_{2,h}-\beta_{2,h}\right\|_\infty\left(\sqrt{\tilde{\nu}_n}+\left\|\hat{\Sigma}_W-\Sigma_W\right\|_\infty k_W\sqrt{\tilde{\nu}_n}\right)\right)\end{aligned} \tag{1.64}$$

Proof of Lemma 1.4. By the definition of the de-LS estimator,

$$\begin{aligned}\sqrt{n}\nu'\left(\hat{\beta}_{1,h}^{(de-LS)}-\beta_{1,h}\right) &= \nu'\left(\frac{1}{n}\sum_{t=p}^{n-h}\hat{W}_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\hat{W}_{1,t}^{\perp}e_{t,h}\right)\\ &\quad+\nu'\left(\frac{1}{n}\sum_{t=p}^{n-h}\hat{W}_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\hat{W}_{1,t}^{\perp}W_{2,t}'\left(\beta_{2,h}-\hat{\beta}_{2,h}\right)\right)\\ &= \nu'\left(E\left[W_{1,t}^{\perp}W_{1,t}'\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}W_{1,t}^{\perp}e_{t,h}\right)+\Lambda_0+\Lambda_1+\Lambda_2,\end{aligned} \tag{1.65}$$

where,

$$\Lambda_0 := \nu'\left\{\left(\frac{1}{n}\sum_{t=p}^{n-h}W_{1,t}^{\perp}W_{1,t}'\right)^{-1}-\left(E\left[W_{1,t}^{\perp}W_{1,t}'\right]\right)^{-1}\right\}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}W_{1,t}^{\perp}e_{t,h}\right) \tag{1.66}$$

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

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

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

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

where $\widehat{DN}$ and DN are defined as in the statement of Lemma 1.3 and

$$\begin{aligned}
\Lambda_{11} :=& \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(\widehat{DN}^{-1} - DN^{-1} \right) R_1 \Sigma_W^{-1} W_t e_{t,h} \\
\Lambda_{12} :=& \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' DN^{-1} R_1 \left(\hat{\Sigma}_W^{-1} - \Sigma_W^{-1} \right) W_t e_{t,h} \\
\Lambda_{13} :=& \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(\widehat{DN}^{-1} - DN^{-1} \right) R_1 \left(\hat{\Sigma}_W^{-1} - \Sigma_W^{-1} \right) W_t e_{t,h}.
\end{aligned} \tag{1.67}$$

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

$$\begin{aligned}
\Lambda_{21} :=& \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' R_1 \Sigma_W^{-1} W_t W_t' R_2' \\
\Lambda_{22} :=& \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' R_1 \left(\hat{\Sigma}_W^{-1} - \Sigma_W^{-1} \right) W_t W_t' R_2' \\
\Lambda_{23} :=& \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(\widehat{DN}^{-1} - I_p \right) R_1 \Sigma_W^{-1} W_t W_t' R_2' \\
\Lambda_{24} :=& \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(\widehat{DN}^{-1} - I_p \right) R_1 \left(\hat{\Sigma}_W^{-1} - \Sigma_W^{-1} \right) W_t W_t' R_2'.
\end{aligned} \tag{1.68}$$

It remains to show that the terms in (1.66), (1.67) and (1.68) are of the specified orders so that the result follows. By Lemma 1.2, Lemma 1.3 and the fact that

$\left\|DN^{-1}\right\|_{\max} = O_p(1)$ and $\left\|\hat{\Sigma}_W^{-1}\right\|_{\infty} = O_p(k_W)$, we have

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

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

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

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

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

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

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

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

By substituting the derived rates into Equation (1.65) and neglecting the higher-order terms, we obtain the result. □

Lemma 1.5. *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 1.2(ii), and (iv)-(vi), it holds true that:

$$(a)\left\|\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\Sigma_{UW}^{-1}U_t e_{th}\right\|_{max}=O_p\left(\sqrt{\tilde{\nu}_n}\right)\quad and\left\|\frac{1}{n}\sum_{t=p}^{n-h}\Sigma_{UW}^{-1}U_t W_t'-I_{dp}\right\|_{max}=O_p\left(\sqrt{\tilde{\nu}_n/n}\right);$$
$$(b)\left\|\widehat{CN}^{-1}-CN^{-1}\right\|_{max}=O_p\left(\|\hat{\mathbf{A}}-\mathbf{A}\|_\infty k_{UW}+\left\|\hat{\Sigma}_{UW}-\Sigma_{UW}\right\|_\infty k_{UW}\right);$$
$$(c)\left\|\widehat{CN}^{-1}-I_p\right\|_{max}=O_p\left(\sqrt{\hat{\nu}_n/n}+\|\hat{\mathbf{A}}-\mathbf{A}\|_\infty k_{UW}+\left\|\hat{\Sigma}_{UW}-\Sigma_{UW}\right\|_\infty k_{UW}\right).$$

Proof of Lemma 1.5. First, recall that the VAR(∞) representation of the VAR model, under the stability condition, allows us to write $W_t=\sum_{k=0}^{\infty}\Upsilon_k u_{t-k}$. Also, let $\tilde{e}_{pj}$, $j=1,\ldots,p$ denote the p-dimensional unit vectors, where $\tilde{e}_{pj}$ contains 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 1.2(vi)). The result follows from Lemma 1.1 applied to the filters $\left\{\Phi_j^{(k)},j=0,1,\ldots\right\},k=1,2$. The first result in part (a) follows from the same Lemma 1.1 applied to the

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

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 1.1 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 from $\|\Sigma_W(j)\|_{\max} = O(1)$ 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 from T, $\left\|\hat{\Sigma}_{UW}^{-1}\right\|_\infty = O_p(1)$, $\left\|\Sigma_{UW}^{-1}\right\|_\infty = O_p(1)$, and $\|R_1\|_\infty = 1$ that

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

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

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

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 $R_1R_1' = I_p$.. It then follows from 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 1.6. *If Assumption 1.2 is satisfied, then for any vector $v \in \mathbb{R}^p$ such that $\|v\|_1 =$*

1,

$$\sqrt{n}v'\left(\hat{\beta}_{1,h}^{(de-2S)}-\beta_{1,h}\right)=v'\left(E\left[U_{1,t}^{\perp}W_{1,t}'\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}U_{1,t}^{\perp}e_{t,h}\right)$$
$$+O_p\left(\tilde{\nu}_n/\sqrt{n}+\|\hat{\mathbf{A}}-\mathbf{A}\|_\infty k_{UW}\sqrt{\tilde{\nu}_n}+\left\|\hat{\Sigma}_{UW}-\Sigma_{UW}\right\|_\infty k_{UW}\sqrt{\tilde{\nu}_n}\right.$$
$$\left.+\left\|\hat{\beta}_{2,h}-\beta_{2,h}\right\|_\infty\left\{\sqrt{\tilde{\nu}_n}+\|\hat{\mathbf{A}}-\mathbf{A}\|_\infty k_{UW}\sqrt{\tilde{\nu}_n}+\left\|\hat{\Sigma}_{UW}-\Sigma_{UW}\right\|_\infty k_{UW}\sqrt{\tilde{\nu}_n}\right\}\right) \tag{1.69}$$

Proof of Lemma 1.6. By the definition of the de-2S estimator,

$$\begin{aligned}\sqrt{n}v'\left(\hat{\beta}_{1,h}^{(de-2S)}-\beta_{1,h}\right)&=v'\left(\frac{1}{n}\sum_{t=p}^{n-h}\hat{U}_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\hat{U}_{1,t}^{\perp}e_{t,h}\right)\\&\quad+v'\left(\frac{1}{n}\sum_{t=p}^{n-h}\hat{U}_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\hat{U}_{1,t}^{\perp}W_{2,t}'\left(\beta_{2,h}-\hat{\beta}_{2,h}\right)\right)\\&=v'\left(E\left[U_{1,t}^{\perp}W_{1,t}'\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}U_{1,t}^{\perp}e_{t,h}\right)+\tilde{\Lambda}_0+\tilde{\Lambda}_1+\tilde{\Lambda}_2,\end{aligned} \tag{1.70}$$

where,

$$\tilde{\Lambda}_0:=v'\left\{\left(\frac{1}{n}\sum_{t=p}^{n-h}U_{1,t}^{\perp}W_{1,t}'\right)^{-1}-\left(E\left[U_{1,t}^{\perp}W_{1,t}'\right]\right)^{-1}\right\}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}U_{1,t}^{\perp}e_{t,h}\right) \tag{1.71}$$
$$\tilde{\Lambda}_1:=v'\left(\frac{1}{n}\sum_{t=p}^{n-h}\hat{U}_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\hat{U}_{1,t}^{\perp}e_{t,h}\right)-v'\left(\frac{1}{n}\sum_{t=p}^{n-h}U_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}U_{1,t}^{\perp}e_{t,h}\right)$$
$$\tilde{\Lambda}_2:=v'\left(\frac{1}{n}\sum_{t=p}^{n-h}\hat{U}_{1,t}^{\perp}W_{1,t}'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\hat{U}_{1,t}^{\perp}W_{2,t}'\left(\beta_{2,h}-\hat{\beta}_{2,h}\right)\right).$$

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 1.5 and

$$\begin{aligned}
\tilde{\Lambda}_{11} :=& \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(\widehat{CN}^{-1} - CN^{-1} \right) R_1 \left(\hat{\Sigma}_{UW}^{-1} - \Sigma_{UW}^{-1} \right) \hat{U}_t e_{t,h} \\
\tilde{\Lambda}_{12} :=& \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(\widehat{CN}^{-1} - CN^{-1} \right) R_1 \Sigma_{UW}^{-1} \hat{U}_t e_{t,h} \\
\tilde{\Lambda}_{13} :=& \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' CN^{-1} R_1 \left(\hat{\Sigma}_{UW}^{-1} - \Sigma_{UW}^{-1} \right) \hat{U}_t e_{t,h} \\
\tilde{\Lambda}_{14} :=& \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' CN^{-1} R_1 \Sigma_{UW}^{-1} (\hat{U}_t - U_t) e_{t,h}
\end{aligned} \tag{1.72}$$

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

$$\begin{aligned}
\tilde{\Lambda}_{21} :=& \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' R_1 \Sigma_{UW}^{-1} \hat{U}_t W_t' R_2' \\
\tilde{\Lambda}_{22} :=& \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' R_1 \left(\hat{\Sigma}_{UW}^{-1} - \Sigma_{UW}^{-1} \right) \hat{U}_t W_t' R_2' \\
\tilde{\Lambda}_{23} :=& \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(\widehat{CN}^{-1} - I_p \right) R_1 \Sigma_{UW}^{-1} \hat{U}_t W_t' R_2' \\
\tilde{\Lambda}_{24} :=& \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} v' \left(\widehat{CN}^{-1} - I_p \right) R_1 \left(\hat{\Sigma}_{UW}^{-1} - \Sigma_{UW}^{-1} \right) \hat{U}_t W_t' R_2'.
\end{aligned} \tag{1.73}$$

It remains to show that the terms in (1.71), (1.72) and (1.73) 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 $\tilde{\Lambda}_0=\sum_{t=p}^{n-h}v'\left(CN^{-1}I_p\right)R_1\Sigma_{uW}^{-1}U_te_{t,h}/\sqrt{n}$. It then follows from Lemma 1.5 that

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

Lemma 1.1 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 from 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 1.5, T and the fact that $\left\|\hat{\Sigma}_{UW}^{-1}\right\|_{\infty}=O_p\left(k_{UW}\right)$, we have

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

$$
\begin{aligned}
\left|\tilde{\Lambda}_{12}\right| &\leqslant p\|v\|_1 \left\|\widehat{CN}^{-1} - CN^{-1}\right\|_{\max} \left\|\Sigma_{UW}^{-1}\right\|_{\infty} \|R_1\|_{\infty} \left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{n-h} \left(\hat{U}_t - U_t\right) e_{t,h} \right\|_{\max} \\
&\quad + p\|v\|_1 \left\|\widehat{CN}^{-1} - CN^{-1}\right\|_{\max} \|R_1\|_{\infty} \left\| \frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} \Sigma_{UW}^{-1} U_t e_{t,h} \right\|_{\max} \\
&= O_p\left(\left\|\widehat{CN}^{-1} - CN^{-1}\right\|_{\max} \sqrt{\tilde{\nu}_n} \left(1 + k_{UW} \left\|\hat{\mathbf{A}} - \mathbf{A}\right\|_{\infty}\right) \right)
\end{aligned}
$$

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

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

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 1.5 and the fact that $R_1 R_2' = 0_{p \times d(p-1)}$. It fololows by T that

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

By substituting the derived rates into Equation (1.70) and neglecting the higher-order terms, we obtain the result. □

Lemma 1.7. *If Assumptions 1.2 and 1.3 are satisfied, then for any vector $v \in \mathbb{R}^p$ such that $\|v\|_1 = 1$, it holds that*

$$
\frac{1}{\sqrt{n}} \sum_{t=p}^{n-h} \frac{v' R_1 \Sigma_W^{-1} W_t e_{t,h}}{s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v)} \xrightarrow{d} \mathcal{N}(0,1), \tag{1.74}
$$

and

$$\frac{1}{\sqrt{n}}\sum_{t=p}^{n-h}\frac{v'R_1\Sigma_{UW}^{-1}U_t e_{t,h}}{s.e._{\hat{\beta}_{1,h}^{(de-2S)}}(v)} \xrightarrow{d} \mathcal{N}(0,1). \tag{1.75}$$

Proof of Lemma 1.7. 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 [15],

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

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

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

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

(iii) $\bar{\sigma}_n^2 := \text{Var}\left(n^{1/2}\sum_{t=p}^{n-h} z_{nt}\right) > \delta > 0$ 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, $\left\|R_1'\tilde{v}\right\|_2^2 = \tilde{v}'R_1R_1'\tilde{v}' = \tilde{v}'\tilde{v} = 1$. Also, we have $\|R_1\|_2 = 1$ by $R_1R_1' = I_p$. 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 from $\|v\|_2 \geq p\|v\|_1 = p$ and Assumption 1.2(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 from 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 1.2(vii) that $1/s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v) = O(1)$.

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

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 $\tilde{v}_{1,j} = v_{1,j} / \left\|v_{1,j}\right\|_2$ 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 1.2(ii) and 1.3(ii), and given that $\|J\|_2 = \|R_1\|_2 = 1$, $\|v\|_2 \leq 1$, and $\left\|\Sigma_W^{-1}\right\|_2 \leqslant C$, it follows, for $r > 2$ as defined in Assumption 1.3(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 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:

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

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 1.3(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 (1.22), 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,

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

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 (1.74) follows from the conclusion of Theorem 5.20 (White, 1999).

To prove the second result (1.75), 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

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

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 1.3(i).

Furthermore, $\tilde{\sigma}_n^2 := \text{Var}\left(n^{-1/2}\sum_{t=p}^{n-h}\tilde{z}_{nt}\right) \to 1, \quad \text{as } n \to \infty$. 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 (1.75) follows from the conclusion of Theorem 5.20 (White, 1999) applied to the double array of scalars $\{\tilde{z}_{nt}\}$. □

Proof of Theorem 1.2. Recall that $1/s.e._{\hat{\beta}_{1,h}^{(de-LS)}}(v) = O(1)$, as justified in the proof of Lemma 1.7 above. Using this result and Lemma 1.4, we obtain, under Condition 1.1, that

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

Additionally, according to the first result of Lemma 1.7 (see Equation (1.74)),

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

giving the result.

□

Proof of Theorem 1.3. 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,

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

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

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

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

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

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

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

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

Therefore,

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

under Condition 1.2.

Note that this result and the fact that $1/s.e._{\hat{\beta}^{(de-LS)}_{1,h}}(v) = O(1)$ (see the proof of Theorem 1.2) imply $\widehat{s.e.}^{(hac)}_{\hat{\beta}^{(de-LS)}_{1,h}}(v)^2 \Big/ s.e._{\hat{\beta}^{(de-LS)}_{1,h}}(v)^2 \xrightarrow{p} 1$. The second result then follows from Theorem 1.2 and Slutsky's theorem.

It remains to show that the asymptotic variance of the de-biased LS has a simplified representation of the form (1.56) if the contemporaneous error term u_t is a conditional m.d.s. First recall that using the expression of $W^{\perp}_{1,t}$, this asymptotic variance can be rewritten as

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

where

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

with

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

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, $1 \le h-j \le h < k < k+1 \le k+h-l$ and by the law of iterated expectations (LIE) and the m.d.s. assumption,

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

Case 2: $k < -h$. In this case, $k < k+h-l \leq k+h < 0 < 1 \leq h-j$. It follows from 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: $-h \leq k \leq h$. We consider three sub-cases.

- If $k = l-j$, then $k < k+1 \leq k+h-l = h-j$ and $h-j \geq 1$ 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 $k < k+1 \leq k+h-l < h-j$ and $h-j \geq 1$. It follows from 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 $1 \leq h-j < k+h-l$ and $k < k+1 \leq k+h-l$. 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 (1.56). □

Proof of Theorem 1.5. 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 $\|R_1\|_2 = 1$ 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 from $\|v\|_2 \geq p\|v\|_1 = p$ and Assumption 1.2(*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 from 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 1.2(*vii*) that $1/s.e._{\hat{\beta}_{1,h}}^{(de-2S)}(v) = O(1)$. With this result and Lemma 1.6, we obtain, under Condition 1.3,

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

In addition, according to the second result of Lemma 1.7 (see Equation (1.75)),

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

giving the result. □

Lemma 1.8. *Let W_t follows (1.1), $\|\mathbf{A}\|_2 \leq \varphi \in [0,1)$, and Assumption 1.2(vii), Assumption 1.3(i), and Assumption 1.4 hold. Then, for any $v \in \mathbb{R}^{dp\times 1}$, with $\|v\|_1 = 1$.*

$$\lim_{n\to\infty} n^{-1}\sum_{t=p}^{n-h}(v's_t)^2/v'\Omega_{U,h}v = 1. \tag{1.76}$$

Proof of Lemma 1.8. The convergence of (1.76) 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 1.2 (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 1.3 (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 $\|\lambda\| = 1$,

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

By Assumption 1.4 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 1.6. Arguments similar to those presented in the first part of the proof of Theorem 1.3 can be invoked to obtain

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

under Condition 1.4.

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 1.2) 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 from Theorem 1.2 and Slutsky's theorem.

Consistency of the HC variance estimator follows from Lemma 1.8.

□

1.10.2. Additional simulation results

1.10.2.1. Tridiagonal root matrix (DGP 1)

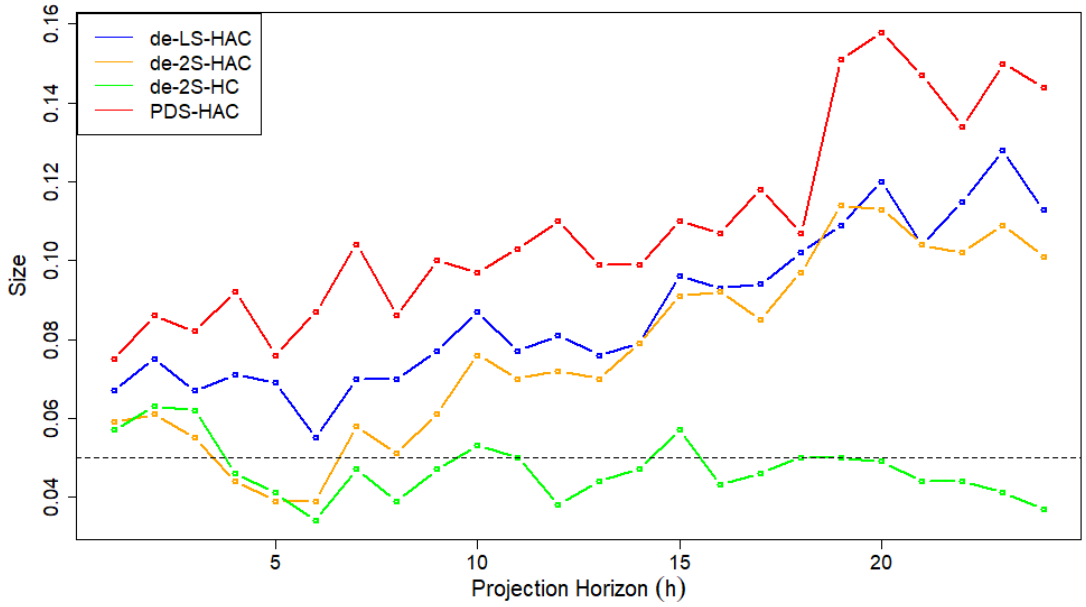

Figure 1.6 – Size of the Wald test at the 5% nominal level for different horizons. The red, blue, orange, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased two-stage with HAC standard errors, 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 $h = 1, \ldots, 24$. The number of replications is $1,000$.

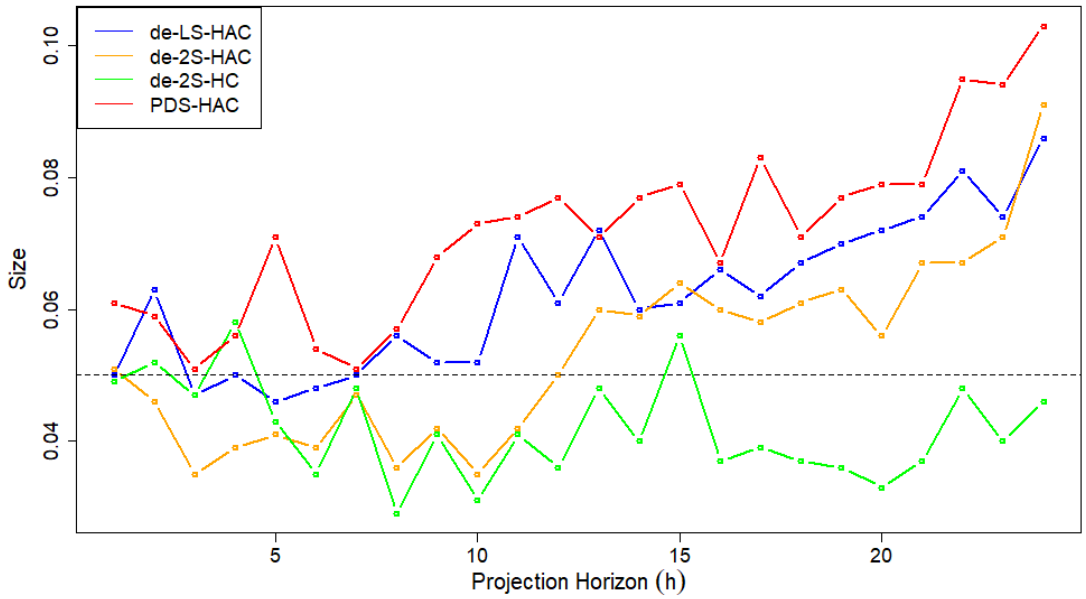

Figure 1.7 – Size of the Wald test at the 5% nominal level for different horizons. The red, blue, orange, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased two-stage with HAC standard errors, 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 $h = 1, \ldots, 24$. The number of replications is $1,000$.

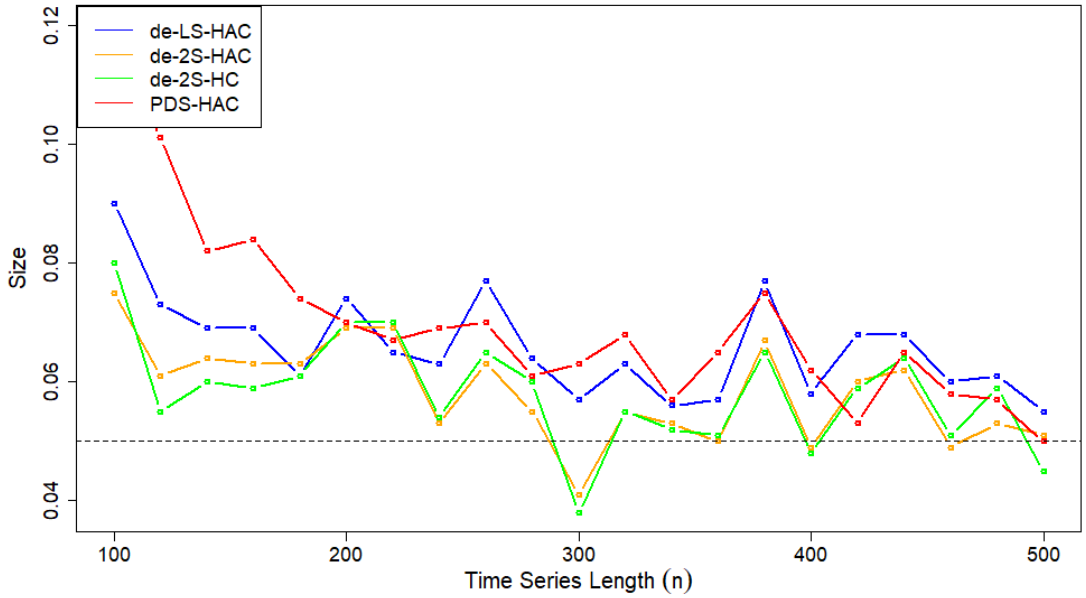

Figure 1.8 – Size of the Wald test at the 5% nominal level for different sample sizes and a given horizon ($h = 1$). The red, blue, orange, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased 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 $1,000$.

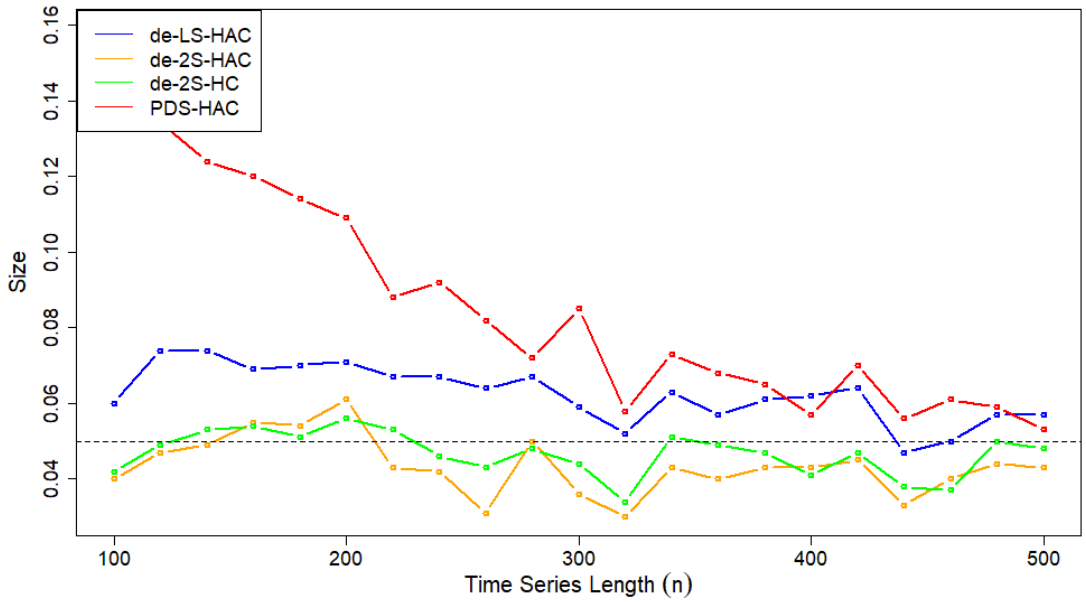

Figure 1.9 – Size of the Wald test at the 5% nominal level for different sample sizes and a given horizon ($h = 4$). The red, blue, orange, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased 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 $1,000$.

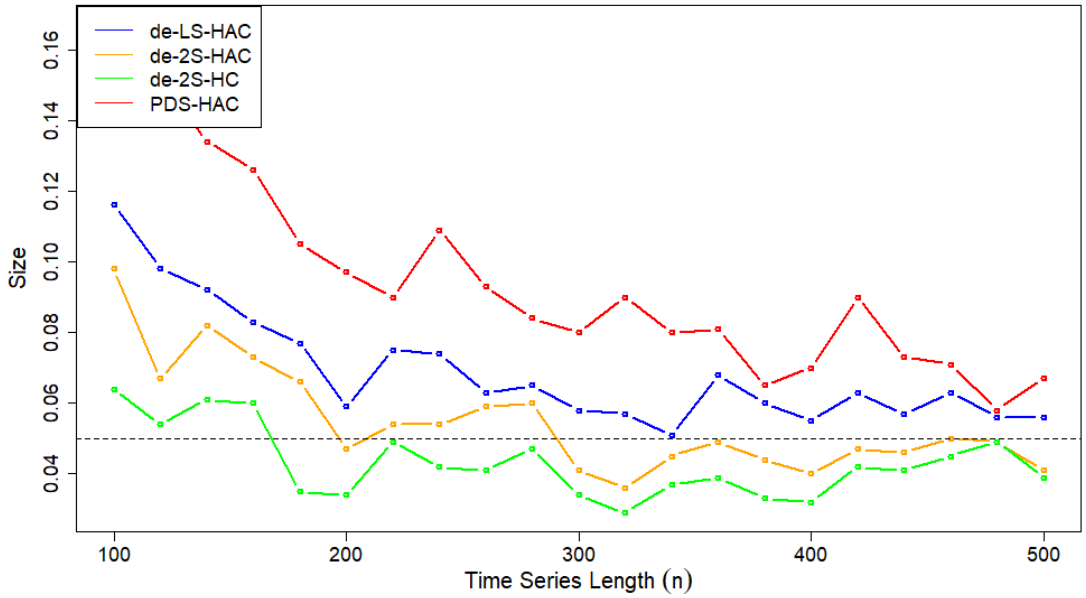

Figure 1.10 – Size of the Wald test at the 5% nominal level for different sample sizes and a given horizon ($h = 8$). The red, blue, orange, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased 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 $1,000$.

1.10.2.2. Random root matrix (DGP 2)

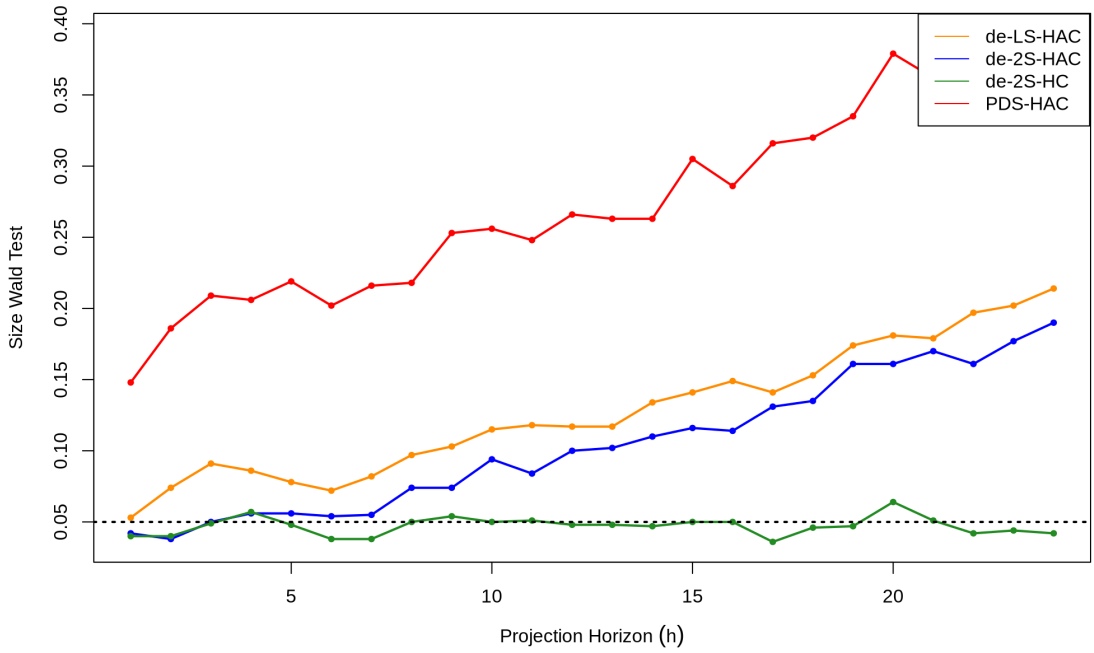

Figure 1.11 – Size of the Wald test at the 5% nominal level for different horizons. The red, orange, blue, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased two-stage with HAC standard errors, 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 $h = 1, \ldots, 24$. The number of replications is $1,000$.

,

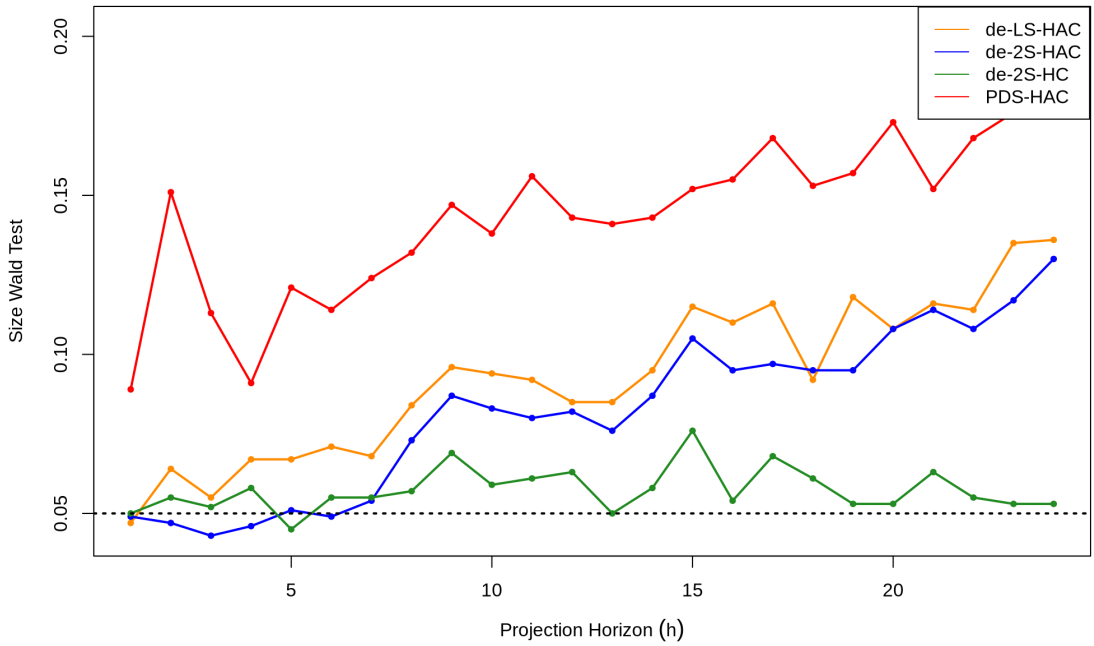

Figure 1.12 – Size of the Wald test at the 5% nominal level for different horizons. The red, orange, blue, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased two-stage with HAC standard errors, 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 $h = 1, \ldots, 24$. The number of replications is $1,000$.

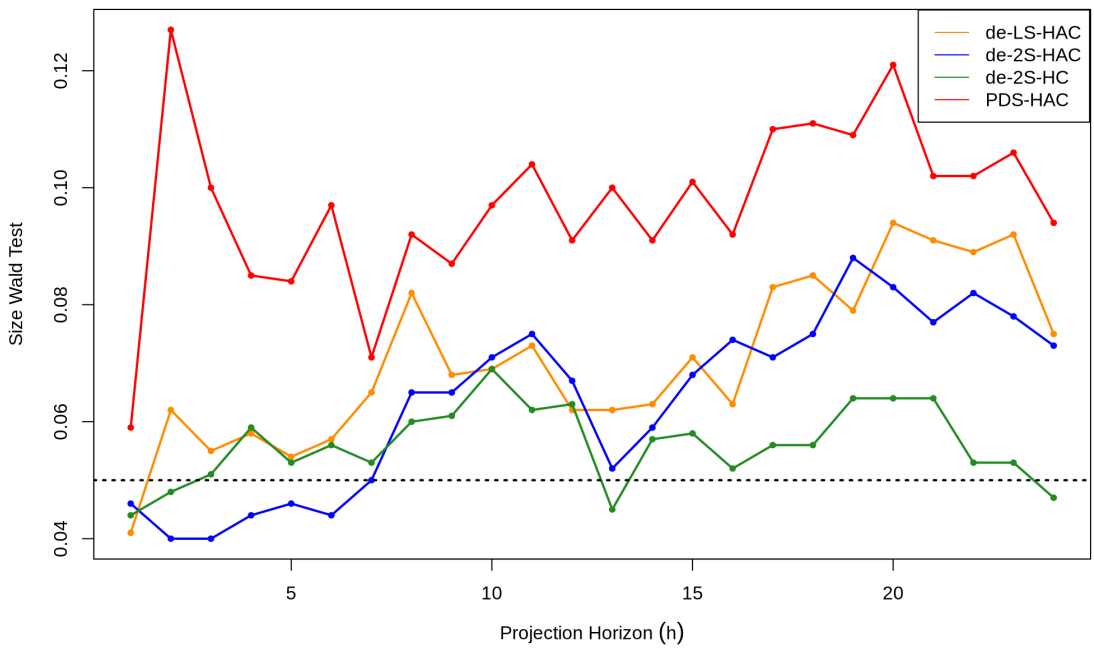

Figure 1.13 – Size of the Wald test at the 5% nominal level for different horizons. The red, orange, blue, and green curves correspond to the post-double selection procedure with HAC standard errors, the de-biased least squares with HAC standard errors, the de-biased two-stage with HAC standard errors, 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 $h = 1, \ldots, 24$. The number of replications is $1,000$.

CHAPTER 2

Volatility Transmission in Stock Returns: A High-Dimensional Heterogeneous Local Projection Framework*

2.1. Introduction

Understanding how volatility propagates from one asset to another is crucial for investors aiming to manage risk and construct more robust portfolios. In particular, capturing the transmission of volatility shocks across different time horizons can enhance investors' understanding of asset interdependencies, help anticipate risks associated with price fluctuations, and enable proactive adjustments to investment strategies—especially through the deployment of effective hedging mechanisms. This paper aims to construct networks of volatility spillovers in stock return prices within a data-rich environment, where many potentially interconnected assets are available.

In today's financial markets, investors have access to data on a wide range of assets, which are often interconnected in complex ways. To model the joint dynamics of realized volatilities, we employ the Heterogeneous Vector Autoregressive (HVAR) model of realized volatilities (see, e.g., Cubadda et al., 2019; Hecq et al., 2023). This HVAR model extends the Heterogeneous Autoregressive model of Realized Volatility (HAR-RV), introduced by Corsi (2009), from a univariate to a multivariate framework. The HAR-RV model is motivated by the fundamental idea that financial markets have a heterogeneous structure, comprising participants with different trading horizons (e.g., daily, weekly, and monthly). Then, the HAR-RV model has a simple autoregressive structure in the realized variance/volatility, but with the feature of considering volatilities realized over different time horizons. As argued by Corsi

⋆. This chapter is co-authored with Endong Wang. The authors are extremely grateful to Marine Carrasco for her helpful discussions. We also thank Luca Margaritella for providing us with the 10-minute realized variance and realized correlation data.

(2009), the HAR-RV model exhibits remarkably good forecasting performance and successfully achieves the purpose of modeling the long-memory behavior of volatility in a very simple and parsimonious way.

From an econometric perspective, the HVAR model of realized volatilities offers a parsimonious representation as a restricted vector autoregressive VAR(22) model, featuring economically meaningful constraints. In fact, instead of including 22 lags of realized volatilities as in the standard VAR(22) model, the HVAR model considers only three main volatility components realized over daily, weekly, and monthly time horizons and obtained by aggregation. Each component is associated with a specific segment of the financial market, leading to a '*heterogeneous*' model that accounts for heterogeneity in agents' trading behaviors. For example, the first component can be associated with short-term traders with daily or higher trading frequency, the second with medium-term investors who rebalance their positions weekly, and the third with agents who adjust their positions monthly or less frequently. The inclusion of these aggregated components leads to a parsimonious formulation that involves the estimation of only a limited number of parameters, thereby improving computational efficiency while preserving strong out-of-sample forecasting performance (see, e.g., Corsi, 2009; Cubadda et al., 2019).

As our ultimate goal is to analyze volatility transmission in stock returns, we rely on the multi-horizon Granger causality test, which essentially focuses on improving forecasting performance. As is well known in standard macroeconometrics, implementing this test requires the derivation and estimation of a local projection equation. Starting from the HVAR model, seen as a restricted VAR(22), our initial attempt resulted in a local projection equation that no longer preserved the heterogeneous property of the underlying HVAR model. This loss of heterogeneity stems from the nonlinear transformation of the VAR companion matrix. To preserve the benefits of heterogeneity—such as improved forecasting performance—we propose a heterogeneous local projection (HLP) framework. The HLP equation is derived from the standard 22-lag local projection equation by projecting past realized volatilities onto

the space spanned by daily-, weekly-, and monthly-aggregated realized volatilities. The resulting local projection equation includes only these latter variables and thus preserves the heterogeneous market structure.

In addition to considering the multivariate extension of the HAR-RV model, we allow for the inclusion of a large number of financial assets, leading to a high-dimensional HVAR setting. Compared to the low-dimensional framework, this high-dimensional representation has the advantage of enabling the identification of clear causal relationships between assets. As argued by Hecq et al. (2023), bivariate analyses often fail to uncover reliable connections between assets. In particular, spurious spillovers from stock k to stock l may arise if both assets are Granger-caused by a third stock that is omitted from the HVAR system. Including a large number of assets in the system thus helps to avoid detecting such spurious spillovers.

Including many assets in the HVAR model implies conditioning on a large number of variables in the heterogeneous local projection equation, which leads to a clear high-dimensionality problem that renders many standard statistical inference techniques invalid. Hecq et al. (2023) introduced a post-double selection (PDS) procedure that performs well in detecting single-horizon spillovers under a sparsity assumption on the underlying HVAR model. However, due to the potential lack of sparsity in our heterogeneous local projection, the PDS approach may fail to identify reliable connections between assets. This paper relies on the methodology developed by Dettaa and Wang (2024) to analyze volatility transmission between stocks over longer horizons (e.g., 1 week, 2 weeks, and 1 month). The advantage of the procedure proposed by Dettaa and Wang (2024) lies in its robustness to non-sparse local projections.

Our sample consists of 30 U.S. stocks. For each asset, we observe a time series of 10-minute daily realized variances, with data collected from March 2008 to May 2014. Since the 2008 global financial crisis may have caused substantial instability in the data (e.g., non-stationary), we discard the 2008–2009 sub-sample from our analysis, resulting in a 2010–2014 sample period totaling $n = 1{,}584$ trading days.

We refer to the 2010–2014 sample period as the *full sample*.

We begin by examining volatility transmission in stock return prices across different horizons (e.g., 1 day, 1 week, 2 weeks, and 1 month) using a heterogeneous local projection equation containing 90 variables (3 volatility components $\times$ 30 assets). For each horizon, we perform Granger causality tests for each pair of assets and use the identified connections to construct a network of realized volatilities. Our Granger causality tests rely on the Wald statistic, using Dettaa and Wang (2024)'s de-biased least squares(deLS) estimator adapted to this high-dimensional framework. Given the large sample size (n = 1,584) relative to the number of regressors, OLS-based inference on the heterogeneous local projection coefficients is a feasible option. For comparison purposes, we also construct a network of realized volatilities based on the standard least squares estimator. Our results suggest that the deLS estimator produces more parsimonious and stable networks, particularly at longer horizons, where OLS may overfit due to the curse of dimensionality and the absence of regularization.

In addition to the full sample results, we also consider a shorter subsample, namely 2013–2014, by discarding the 2010–2012 period due to potential instability caused by the U.S. debt-ceiling crisis of 2011 (Baker et al., 2019). In fact, the results of Hecq et al. (2023) for a single horizon suggest that a shorter time span may increase the likelihood of identifying more stable relationships between assets. The resulting subsample contains n = 488 observations. In this case, the Wald test based on OLS will fail to identify reliable spillover effects. In fact, OLS is still feasible but tends to identify a very large number of connections between assets across the different projection horizons we consider. This result is probably justified by the fact that the Wald test based on OLS will over-reject the null hypothesis of Granger non-causality in this high-dimensional setting. However, our de-biased least squares-based test still identifies a number of connections comparable to the full-sample case but with more sparse networks.

To further avoid the identification of spurious connections due to omitted variables, we enrich our model by allowing for spillovers from realized correlations to

variances. Although we remain mostly interested in contagion between the thirty realized volatilities, we add the 435 ($= 30 \times 29/2$) realized correlations between all these assets as control variables. The resulting HVAR model contains 465 ($= 30 + 435$) variables, with the corresponding HLP equation containing 1,395 ($= 465 \times 3$) variables. In this framework, OLS breaks down due to strong high-dimensionality even in the full sample. However, de-biased least squares still perform well in terms of identifying reliable connections. These results highlight the disparity in network density between the OLS and deLS estimators. Including realized correlations increases dimensionality and amplifies OLS overfitting, whereas the deLS estimator, with regularization and de-biasing, yields a more parsimonious volatility transmission structure.

This paper contributes to an important strand of the empirical finance literature that investigates network connectedness in the volatilities of financial asset prices. In this literature, network analysis generally relies either on the construction of connectedness measures (see, e.g., Demirer et al., 2018; Diebold and Yılmaz, 2014) or on Granger causality tests (see, e.g., Hecq et al., 2023). For instance, Diebold and Yılmaz (2014) proposed several connectedness measures based on variance decompositions and applied them to track the daily time-varying connectedness of major U.S. financial institutions' stock return volatilities, with a focus on the 2007–2008 financial crisis. Along similar lines, but using Granger causality analysis, spillovers and contagion among stock returns have been studied in networks by Barigozzi and Brownlees (2019), Lin and Michailidis (2017), Corsi et al. (2018), and Vỳrost et al. (2015).

Recent advances in regularization for estimating VAR-type models in high-dimensional settings have provided the literature with reliable tools for analyzing network connectedness in such contexts. For example, Miao et al. (2023) proposed a regularized estimator tailored to high-dimensional vector autoregressions (VARs) augmented with common factors, allowing for strong cross-sectional dependence, and applied their methodology to explore dynamic connectedness in financial asset price volatil-

ities. Hecq et al. (2023) developed a valid post-selection test for Granger causality in high-dimensional VARs and used their method to construct networks of realized volatilities, based on a sample of U.S. financial stocks modeled as a vector heterogeneous VAR. However, their analysis is restricted to identifying spillover effects at a single horizon. This paper extends the application of Hecq et al. (2023) by investigating spillovers in realized variances across multiple horizons, using the multi-horizon Granger causality test developed by Dettaa and Wang (2024).

The rest of the paper is organized as follows: Section 2.2 introduces the heterogeneous VAR model of realized volatilities and derives the heterogeneous local projection equation. Section 2.3 presents our estimation and testing procedure, inspired by the methodology of Dettaa and Wang (2024). Section 2.4 presents the data and conducts multi-horizon Granger causality tests to construct volatility spillover networks across different horizons. Section 2.5 concludes, and additional results are provided in the Appendix.

2.2. The local projection representation of heterogeneous VAR

2.2.1. Framework of heterogeneous VAR

To analyze volatility transmission in stock returns, we employ the multivariate version of the heterogeneous autoregressive (HAR) model of Corsi (2009) to model the joint dynamics of realized volatilities (see also Cubadda et al., 2019; Hecq et al., 2023).

If y_t is a d-dimensional vector of demean log-realized variances/volatilities associated with d financial assets observed at a daily frequency, the heterogeneous vector autoregressive (HVAR) specification is given by the following equation:

$$y_t = c + A^{(1)} y_{t-1} + A^{(2)} y_{t-1}^{(\text{week})} + A^{(3)} y_{t-1}^{(\text{month})} + u_t \tag{2.1}$$

where $y_t^{(\text{week})} = \frac{1}{5}\sum_{j=0}^{4} y_{t-j}$ and $y_t^{(\text{month})} = \frac{1}{22}\sum_{j=0}^{21} y_{t-j}$ are the vectors containing the average volatility over the last 5 (week) and 22 (month) days. u_t is a serially uncorrelated innovation process with zero mean.

The HVAR model (2.1) means that current log-realized volatilities are driven not only by their most recent daily values but also by medium- and long-term average components. Specifically, the model combines the influence of short-term dynamics through the lagged daily value y_{t-1}, medium-term dynamics via the weekly average $y_{t-1}^{(\text{week})}$, and long-term trends captured by the monthly average $y_{t-1}^{(\text{month})}$. This structure allows the model to differentiate between the behaviors of short-term traders, who react to daily fluctuations, medium-term investors, who adjust their positions on a weekly basis, and long-term agents, whose decisions are based on monthly trends. Also, $A^{(1)}$, $A^{(2)}$, and $A^{(3)}$ are non-random coefficient matrices.

Assume that we have a sample of n observations $(y_1, y_2, \ldots, y_n)$ drawn from the data-generating process given by (2.1). The number of series d is assumed to be *large*, leading to a high-dimensional framework. Our ultimate goal is to conduct pairwise multi-horizon Granger causality tests to examine volatility transmission between assets.

In standard macroeconometrics, the Granger causality test generally relies on the local projection (LP) equation. To obtain the LP equation, note that the HVAR model (2.1) has a restricted VAR(22) representation. [1] This restricted VAR(22) model has the following companion form, which is used in the next subsection to derive the LP

1. In fact, Equation (2.1) is equivalent to

$$y_t = \Phi_1 y_{t-1} + \cdots + \Phi_{22} y_{t-22} + u_t,$$

where

$$\Phi_i = \begin{cases} A^{(1)} + \frac{1}{5}A^{(2)} + \frac{1}{22}A^{(3)} & \text{for } i = 1 \\ \frac{1}{5}A^{(2)} + \frac{1}{22}A^{(3)} & \text{for } i = 2, \cdots, 5 \\ \frac{1}{22}A^{(3)} & \text{for } i = 6, \cdots, 22 \end{cases}.$$

equation:

$$Y_t = \mu + \mathbf{A}Y_{t-1} + U_t, \tag{2.2}$$

where $Y_{t-1} = [y'_{t-1}, y'_{t-2}, \cdots, y'_{t-22}]'$, $J := [I_d, 0, \cdots, 0]$ is a $d \times 22d$ selection matrix, $\mu = J'c$, $U_t = J'u_t$. The matrix coefficient $\mathbf{A}$ is a $22d \times 22d$ companion matrix defined by

$$\mathbf{A} = [\mathbf{A}'_1, \mathbf{A}'_2]' \tag{2.3}$$

such that $\mathbf{A}_1 = [A^{(1)}, A^{(2)}, A^{(3)}]G$ and $\mathbf{A}_2 = [I_{21d}, 0_{21d \times d}]$. G is the $3d \times 22d$ selection matrix defined by

$$G = \begin{bmatrix} I_d & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ \frac{1}{5}I_d & \frac{1}{5}I_d & \frac{1}{5}I_d & \frac{1}{5}I_d & \frac{1}{5}I_d & 0 & \cdots & 0 \\ \frac{1}{22}I_d & \frac{1}{22}I_d & \frac{1}{22}I_d & \frac{1}{22}I_d & \frac{1}{22}I_d & \frac{1}{22}I_d & \cdots & \frac{1}{22}I_d \end{bmatrix}.$$

The Heterogeneous Autoregressive (HAR) model represents a parsimonious yet robust approach to modeling the persistent dynamics of financial market volatility. This time-series framework was specifically developed to capture two key empirical features of volatility processes: long-memory dependence and multi-scale persistence. The model's theoretical foundation is based on the heterogeneous market hypothesis, which postulates that the aggregation of market participants operating across distinct time horizons generates the characteristic volatility persistence observed in financial markets.

The HAR model's econometric specification incorporates this heterogeneity through a strategically designed autoregressive structure that aggregates volatility components across daily, weekly, and monthly time horizons. From an econometric perspective, the HAR model offers significant practical benefits. Its parsimonious specification requires the estimation of only a limited number of parameters, enhancing computational efficiency while maintaining strong out-of-sample forecasting performance. This combination of theoretical coherence and empirical effectiveness has

led to the model's widespread adoption in financial econometrics.

2.2.2. Heterogeneous local projection and Granger causality

From a technical perspective, the heterogeneous VAR constitutes a parsimonious time series specification. We derived its general VAR representation above, which is particularly relevant for the analysis of multi-horizon predictability. This is a crucial consideration in financial econometrics, given the varying holding periods of financial instruments.

Iterating on the companion representation (2.2) and focusing on the outcome variable $y_{l,t+h}$ yields the following standard LP equation with 22 lags of y_t:

$$y_{l,t+h} = \mu_l^{(h)} + a_l^{(h)'} Y_t + u_{l,t}^{(h)}, \tag{2.4}$$

where $a_l^{(h)'} := J_l \mathbf{A}^h$ is the l^{th} row of the matrix $J\mathbf{A}^h$, $\mu_l^{(h)}$ is the l^{th} component of $\mu^{(h)} := \sum_{i=0}^{h-1} J\mathbf{A}^i J'c$, and $u_{l,t}^{(h)}$ the l^{th} component of $u_t^{(h)} := \sum_{i=0}^{h-1} J\mathbf{A}^i J' u_{t+h-i}$. $J_l = [0, \cdots, 1, \ldots, 0]$ is a $1 \times 22d$ selection matrix where 1 appears in the l^{th} position.

The standard local projection equation (2.4) facilitates the examination of predictability across multiple horizons, thereby providing a more comprehensive and nuanced understanding of connectedness relationships. However, this equation loses the long-memory dependence and multi-scale persistence characteristics that are central to heterogeneous VAR frameworks. This limitation arises because the local projection specification treats each covariate independently, rather than focusing on explicit weekly or monthly horizon-specific covariates.

To address this methodological gap, we propose an enhanced framework: a heterogeneous local projection model incorporating scaled independent variables. This extension aims to preserve the advantages of traditional local projections while reintroducing the crucial heterogeneous market dynamics captured in VAR specifications.

Note that the ultimate goal is to project $y_{l,t}$ onto y_t, $y_t^{(\text{week})}$, and $y_t^{(\text{month})}$ to reflect

the heterogeneous representation of the HVAR model (2.1). However, the standard local projection equation (2.4) derived from the restricted VAR(22) representation (2.2) does not inherit the specific structure of the HVAR model. Indeed, the nonlinear transformation of the restricted VAR(22) matrix coefficients yields a local projection equation corresponding to the projection of $y_{l,t+h}$ onto Y_t, rather y_t, $y_t^{(\text{week})}$, and $y_t^{(\text{month})}$, as needed. To obtain the desired representation, we go from Equation (2.4) and project $a_l^{(h)\prime} Y_t$ onto $w_t := \left[y_t', y_t^{(\text{week})\prime}, y_t^{(\text{month})\prime}\right]'$.

Let β_h denote the coefficient of the projection of $y_{l,t+h}$ onto w_t. Then, we have

$$\beta_h = \mathbb{E}[w_t w_t']^{-1}\mathbb{E}[w_t y_{l,t+h}]. \tag{2.5}$$

Also, it is straightforward to see that $w_t = GY_t$, where G is the selection matrix defined in the previous subsection. Since the projection error $u_{l,t}^{(h)}$ contains only future shocks beyond period t, we have the following explicit formula for β_h:

$$\beta_h = \mathbb{E}[w_t w_t']^{-1}\mathbb{E}[w_t Y_t' a_l^{(h)}] = \left(G\Sigma_Y G'\right)^{-1} G\Sigma_Y a_l^{(h)},$$

where $\Sigma_Y := \mathbb{E}\left[Y_t Y_t'\right]$

Adding and subtracting $\beta_h' w_t$ to the right-hand side of the standard local projection equation (2.4) leads to the following parsimonious equation, which we refer to as the *heterogeneous local projection* (HLP),

$$y_{l,t+h} = \mu_l^{(h)} + \beta_h' w_t + e_{l,t}^{(h)}. \tag{2.6}$$

The error term, $e_{l,t}^{(h)} = (a_l^{(h)\prime} Y_t - \beta_h' w_t) + u_{l,t}^{(h)}$, comprises two components. The first term, $a_l^{(h)\prime} Y_t - \beta_h' w_t$, arises because the HLP projects $y_{l,t+h}$ onto the subspace spanned by w_t, rather than the full information set defined by Y_t. The second component, $u_{l,t}^{(h)}$, captures the future shock that is orthogonal to all information available up to period t. Note that both residuals are uncorrelated with w_t as long as the underlying HVAR error term u_t is serially uncorrelated. Consequently, the projection coefficient,

β_h, can be identified by least squares, and estimation and inference can be conducted using the methodology proposed by Dettaa and Wang (2024). [2]

Recall that w_t is a $3d$-dimensional vector. Assume we are interested in Granger causality from $y_{k,t}$ to $y_{l,t}$. Let R_1 be the $3 \times 3d$ selection matrix defined by

$$R_{1,ij} = \begin{cases} 1 & \text{if } (i = 1 \text{ and } j = k) \\ 1 & \text{if } (i = 2 \text{ ad } j = d + k) \\ 1 & \text{if } (i = 2 \text{ and } j = 2d + k \\ 0 & \text{otherwise} \end{cases},$$

so that $w_{1,t} := \left(y_{k,t}, y_{k,t}^{(\text{week})}, y_{k,t}^{(\text{month})}\right)' = R_1 w_t$. If R_2 is the $3(d-1) \times 3d$ selection matrix obtained by stacking the $3d$-dimensional standard basis vectors for all positions except k, $d+k$, and $2d+k$. Then, Equation (2.6) can be rewritten as

$$y_{l,t+h} = \beta_{1,h}' w_{1,t} + \beta_{2,h}' w_{2,t} + e_{l,t}^{(h)}, \tag{2.7}$$

where $\beta_{1,h} := R_1 \beta_h = \left(\beta_{1,h}^{(1)}, \beta_{1,h}^{(2)}, \beta_{1,h}^{(3)}\right)'$, $\beta_{2,h} := \left(\mu_l^{(h)}, \beta_h' R_2'\right)'$, $w_{2,t} := \left(1, w_t' R_2'\right)'$. Note that $\beta_{1,h}^{(1)}$, $\beta_{1,h}^{(2)}$, and $\beta_{1,h}^{(3)}$ are, respectively, coefficients of $y_{k,t}$, $y_{k,t}^{(\text{week})}$, and $y_{k,t}^{(\text{month})}$ in Equation (2.6). To test for the null hypothesis of no Granger causality from $y_{k,t}$ to $y_{l,t}$ against the alternative of Granger causality, we test

$$\mathscr{H}_0 : \beta_{1,h}^{(1)} = \beta_{1,h}^{(2)} = \beta_{1,h}^{(3)} = 0 \quad \text{vs.} \quad \mathscr{H}_1 : \beta_{1,h}^{(1)}, \beta_{1,h}^{(2)}, \beta_{1,h}^{(3)} \neq 0. \tag{2.8}$$

For a given horizon h, we perform this test for every (l,k)-pair to obtain a full $d \times d$ network of spillover effects.

2. However, the standard two-stage procedure used in Dettaa and Wang (2024) fails to identify the heterogeneous local projection coefficient β_h due to the presence of the additive term $a_l^{(h)'} Y_t - \beta_h' w_t$ in the HLP error term, $e_{l,t}^{(h)}$. In fact, following the two-stage identification strategy of Dettaa and Wang (2024), the natural innovations-based vector, which is a candidate to serve as an instrumental variable for w_t, is defined by $v_t := G\Lambda_t$, with $\Lambda_t := (u_t', u_{t-1}', \cdots, u_{21}')'$. Since $a_l^{(h)'} Y_t - \beta_h' w_t$ may depend on past shocks, v_t is not a valid instrument for w_t in the heterogeneous local projection equation (2.6), as $\mathbb{E}\left[v_t e_{l,t}^{(h)}\right] \neq 0$.

2.3. Estimation and test procedure

Our objective is to conduct inference on the causal coefficient of interest $\beta_{1,h}$, while accounting for the high-dimensional set of control variables $w_{2,t}$ (note that the HLP equation contains $3d$ regressors, which can be very large if d itself is large). Thus, estimation and inference based on standard OLS may fail to identify reliable causal relationships. To address this issue, we rely on the de-biased least squares (de-LS) identification, estimation, and inference procedure proposed by Dettaa and Wang (2024), which is robust to high-dimensional local projections with potentially non-sparse coefficients. For inference purposes, we use the asymptotic theory developed in the same paper. To test the null hypothesis of Granger non-causality, we proceed as follows:

Step 1: Lasso estimation of the HVAR model

Let $A_{1:3} := \left[A^{(1)}, A^{(2)}, A^{(3)}\right]$. The HVAR model (2.1) can be rewritten as

$$y_t = c + A_{1:3} w_{t-1} + u_t,$$

where $w_t := \left[y_t', y_t^{(\text{week})\prime}, y_t^{(\text{month})\prime}\right]'$. Under appropriate sparsity assumption, consistent estimator of $A_{1:3}$ can be obtained using l_1-type penalization. We will rely on the Zou (2006)'s adaptive lasso (AdaLasso) method which involves estimating $A_{1:3}$ through row-wise penalized regression as follows:

$$\left(\hat{c}_j, \hat{A}^{(re)}_{j\cdot,\,1:3}\right) = \underset{c_j,\,\beta=(\beta_1,\ldots,\beta_{3d})'\in\mathbb{R}^{3d}}{\arg\min} \frac{1}{n-22}\sum_{t=23}^{n}\left(y_{j,t} - c_j - \beta' w_{t-1}\right)^2 + \lambda\sum_{s=1}^{3d}\pi_s\left|\beta_s\right|, \quad (2.9)$$

where $j = 1, \cdots, d$, $\pi_s := \left|n^{-1/2} + \bar{\beta}_s\right|^{-\tau}$ represents different weights[3] used to penalize each variable, and $\bar{\beta} = \left(\bar{\beta}_s,\ s = 1, 2, \ldots, 3d\right)$ are the Lasso estimators of

3. As Lasso is used as a first-step model, some estimates $\bar{\beta}_s$ will be equal to zero. Garcia et al. (2017) suggest adding $n^{-1/2}$ to avoid infinite weights.

$A_{j\cdot,1:3}$ obtained from

$$\bar{\beta} = \underset{\beta \in \mathbb{R}^{3d}}{\operatorname{argmin}} \frac{1}{n-22} \sum_{t=23}^{n} \left(y_{j,t} - c_j - \beta' w_{t-1} \right)^2 + \lambda \|\beta\|_1.$$

This data-driven method for choosing weights has been used in several papers in the literature and has shown good forecasting performance, particularly for time series data (see, e.g., Garcia et al., 2017; Krampe et al., 2023; Medeiros and Vasconcelos, 2016). The parameter $\tau > 0$ determines the degree of emphasis placed on differences in the weights. Although τ can be selected using cross-validation or information criteria, a common choice is $\tau = 1$. Finally, λ is the regularization parameter. We use the Bayesian Information Criterion (BIC), adapted to Lasso-type penalization problems (see Zou et al., 2007), to select the optimal regularization parameter. To implement the adaptive Lasso with BIC-based tuning parameter selection, we use an estimator from the R *HDeconometrics* package (Garcia et al., 2017), which internally relies on the *glmnet* package (Friedman and Schwartz, 2008).

With the regularized estimator $\hat{A}_{1:3}^{(re)}$ in hand, we use the explicit formulas in (2.3) to obtain a regularized estimator of the companion matrix $\hat{\mathbf{A}}^{(re)}$.

Step 2: Covariance matrices estimation

We apply the explicit formula (2.10) to obtain an estimator of the covariance matrix of Y_t. Indeed, the covariance matrix of Y_t is determined by the companion matrix and the covariance matrix of the disturbance term through the following explicit relationship:

$$\Sigma_Y := \mathbb{E}[Y_t Y_t'] = \sum_{j=0}^{\infty} \mathbf{A}^j J' \Sigma_u J (\mathbf{A}')^j, \tag{2.10}$$

and a straightforward estimator would be

$$\hat{\Sigma}_Y^{(re)} = \sum_{j=0}^{N} (\hat{\mathbf{A}}^{(re)})^j J' \hat{\Sigma}_u J (\hat{\mathbf{A}}^{(re)'})^j$$

where $\hat{\Sigma}_u = \frac{1}{n-22}\sum_{t=23}^{n}\hat{u}_t\hat{u}_t'$, $\hat{u}_t = y_t - \hat{c} - \hat{A}_{1:3}^{(re)}w_{t-1}$, and setting N as some large number (e.g., 50).

The estimator of the covariance matrix of $w_t = GY_t$ is straightforward,

$$\hat{\Sigma}_w = G\hat{\Sigma}_Y^{(re)}G', \tag{2.11}$$

and its precision matrix $\hat{\Theta} = \hat{\Sigma}_w^{-1}$.

Step 3: Estimation of the rotated regressors

Let $w_{1,t}^{\perp} := w_{1,t} - P_L\left(w_{1,t} \mid w_{2,t}\right)$, where $P_L\left(w_{1,t} \mid w_{2,t}\right)$ denotes the linear projection of the vector of regressors of interest, $w_{1,t}$, onto the high-dimensional set of control variables, $w_{2,t}$. Then, using simple block matrix inversion, we obtain $w_{1,t}^{\perp} = (R_1\Theta R_1')^{-1}R_1\Theta w_t$, where $\Theta := \Sigma_w^{-1}$ is the precision matrix of w_t. A straightforward estimator of $w_{1,t}^{\perp}$ can be constructed using the estimator of Θ defined in the previous step.

Step 4: De-biased least squares estimation (de-LS)

The least squares estimator of $\beta_{1,h}$ is given by: [4]

$$\hat{\beta}_{1,h}^{(ls)} = \left(\sum_{t=22}^{n-h}\hat{w}_{1,t}^{\perp}w_{1,t}'\right)^{-1}\sum_{t=22}^{n-h}\hat{w}_{1,t}^{\perp}y_{l,t+h}, \tag{2.12}$$

where $\hat{w}_{1,t}^{\perp} = (R_1\hat{\Theta}R_1')^{-1}R_1\hat{\Theta}w_t$.

Dettaa and Wang (2024) pointed out a potential regularization bias arising from the first-step machine learning estimation of $A_{1:3}$, which can contaminate the (naive) least squares estimator defined above, and proposed a de-biased estimator (de-LS)

4. This least squares estimator can be interpreted as a '*regularized*' estimator of $\beta_{1,h}$, since the estimation of $w_{1,t}^{\perp}$ relies on the precision matrix estimator $\hat{\Theta}$, obtained from the first-step regularized estimator of the HVAR coefficient matrix, $\hat{A}_{1:3}^{(re)}$.

with improved bias properties in finite samples. This de-LS estimator is defined as:

$$\hat{\beta}_{1,h}^{(de-ls)} = \hat{\beta}_{1,h}^{(ls)} - \left(\sum_{t=22}^{n-h} \hat{w}_{1,t}^{\perp} w_{1,t}' \right)^{-1} \sum_{t=22}^{n-h} \hat{w}_{1,t}^{\perp} w_{2,t}' \hat{\beta}_{2,h}^{(re)}, \tag{2.13}$$

where, $\hat{\beta}_{2,h}^{(re)} := \left(\hat{\mu}_l^{(h)}, \hat{\beta}_h^{(re)\prime} R_2'\right)'$ is a first step regularized estimator of $\beta_{2,h}$ with $\hat{\mu}^{(h)} = \sum_{i=0}^{h-1} J \left(\hat{\mathbf{A}}^{(re)}\right)^i J' \hat{c}$, $\hat{\beta}_h^{(re)} = \left(G \hat{\Sigma}_Y^{(re)} G'\right)^{-1} G \hat{\Sigma}_Y^{(re)} (\hat{a}_l^{(h)})^{(re)}$, and $(\hat{a}_l^{(h)})^{(re)} := J_l \left(\hat{\mathbf{A}}^{(re)}\right)^h$.

Step 5: Asymptotic variance estimation

We rely on the Heteroskedasticity and Autocorrelation Consistent (HAC) variance estimator proposed by Dettaa and Wang (2024), defined as

$$\widehat{\text{AVar}}^{(hac)} \left(\sqrt{n} \hat{\beta}_{1,h}^{(de-ls)} \right) = (R_1 \hat{\Theta} R_1') \hat{\Omega}_{w_1,h}^{(hac)} (R_1 \hat{\Theta} R_1'), \tag{2.14}$$

where $\hat{\Omega}_{w_1,h}^{(hac)}$ is a consistent HAC estimator (e.g., the Newey–West estimator) of the long-run variance of the sample regression score function $\hat{w}_{1,t}^{\perp} \hat{e}_{l,t}^{(h)}$, with $\hat{e}_{l,t}^{(h)} = y_{l,t+h} - \hat{\mu}_l^{(h)} - \hat{\beta}_h^{(re)'} w_t$. Practically, we use the R function *getBandwidthAnd* for automatic bandwidth selection in the spirit of Andrews (1991). The Newey and West (1994) long-run variance estimator is then obtained using the R function *getLongRunVar* with the Bartlett kernel.

Step 6: Compute the Wald test statistic

The Wald test statistic for testing the null $\mathscr{H}_0 : \beta_{1,h} = 0$ against the bilateral alternative is given by

$$W_n^{(de-ls)} := n \hat{\beta}_{1,h}^{(de-ls)'} \left(\widehat{\text{AVar}}^{(hac)} \left(\sqrt{n} \hat{\beta}_{1,h}^{(de-ls)} \right) \right)^{-1} \hat{\beta}_{1,h}^{(de-ls)}.$$

Step 7: Decision rule

The asymptotic theory developed in Dettaa and Wang (2024) establishes that, under the null hypothesis $\mathscr{H}_0 : \beta_{1,h} = 0$, we have $W_n^{(de-ls)} \xrightarrow{d} \chi^2(3)$ as $n \to \infty$, under

certain regularity conditions. We will reject the null and conclude in favor of Granger causality at horizon h from asset k to asset l at the α significance level if $W_n^{(de\text{-}ls)} > \chi^2_{1-\alpha}(3)$, where $\chi^2_{1-\alpha}(3)$ denotes the $(1-\alpha)$-quantile of the $\chi^2(3)$ distribution.

2.4. Data and estimation results

2.4.1. Data

Our dataset consists of thirty (30) time series of daily realized variances observed for thirty U.S. assets, as presented in Table 2.1.[5] More precisely, the data consist of 10-minute daily realized variances defined, for a given stock, by

$$RV10_t \equiv \sum_{j=1}^{M} r_{j,t}^2, \quad r_{j,t} = \ln P_{j,t} - \ln P_{j-1,t}, \tag{2.15}$$

where $j = 1, \ldots, M$ denotes intraday 10-minute stock prices $P_{j,t}$. Realized variances are computed from 10-minute returns, as this frequency minimizes microstructure noise in the original high-frequency return data (see Hecq et al., 2023; McAleer and Medeiros, 2008). The initial sample consists of data collected from March 2008 until May 2014 for a total of 2,236 trading days.

Figure 2.1 presents the series of realized variances (top panel) and the series of log-realized variances along with the corresponding density function (bottom panel) for Apple Inc. (AAPL), one of the assets in our sample. The series of realized variances appears to be non-stationary and exhibits pronounced right skewness. This pattern is observed for nearly all assets in the sample. To improve finite sample properties, we apply a logarithmic transformation to reduce the right-skewness of the volatilities[6]; the transformed series tends to be more stationary (see the right panel of Figure

5. We thank Luca Margaritella for providing us with the 10-minute realized variance data used in Hecq et al. (2023).

6. For example, the log transformation reduces the skewness of AAPL's realized variance from 8.88 to 0.76.

Table 2.1 – List of the 30 U.S. stocks used

No.	Symbol	Issue name	No.	Symbol	Issue name
1	AAPL	APPLE INC	16	KO	COCA-COLA CO
2	AXP	AMERICAN EXPRESS CO	17	MCD	MCDONALD'S CORP
3	BA	BOEING CO	18	MMM	3M
4	CAT	CATERPILLAR	19	MRK	MERCK & CO
5	CSCO	CISCO SYSTEMS	20	MSFT	MICROSOFT CORPORATION
6	CVX	CHEVRON CORP	21	NKE	NIKE INC
7	DD	DOW CHEMICAL COMPANY	22	PFE	PFIZER INC
8	DIS	WALT DISNEY CO	23	PG	PROCTER & GAMBLE CO
9	GE	GENERAL ELEC	24	TRV	TRAVELERS COMPANIES INC
10	GS	GOLDMAN SACHS GROUP INC	25	UNH	UNITEDHEALTH GROUP INC
11	HD	HOME DEPOT INC	26	UTX	UNITED TECHNOLOGIES CORPORATION
12	IBM	INTL BUS MACHINE	27	V	VISA INC
13	INTC	INTEL CORP	28	VZ	VERIZON COMMUNICATIONS INC
14	JNJ	JOHNSON & JOHNSON	29	WMT	WALMART INC
15	JPM	JPMORGAN CHASE & CO	30	XOM	EXXON MOBIL CORPORATION

2.1). However, the log realized variances show a decreasing trend, probably due to the 2008 global financial crisis. For this reason, we discard the 2008–2009 subsample from our analysis, resulting in a 2010–2014 sample period totaling $n = 1{,}584$ trading days. We refer to the 2010–2014 sample period as the *full sample*.

2.4.2. Network of realized variances

In this section, we present and discuss the results of our volatility network analysis. We begin with the full sample of $n = 1{,}584$ trading days and evaluate four projection horizons: 1 day, 5 days (1 week), 10 days (2 weeks), and 22 days (1 month). For each horizon, we conduct pairwise Granger causality tests at the 1% significance level. The first-step estimation of the HVAR model employs Adaptive Lasso with the Bayesian Information Criterion (BIC) guiding the selection of the regularization parameter λ. We use the p-values from Wald tests to assess whether one asset Granger-causes another at a given horizon.

Our analysis does not adjust the p-values for multiple tests, as the objective is exploratory to uncover patterns of pairwise linkages rather than to produce a definitive set of statistically significant connections. Although multiple testing adjustments may

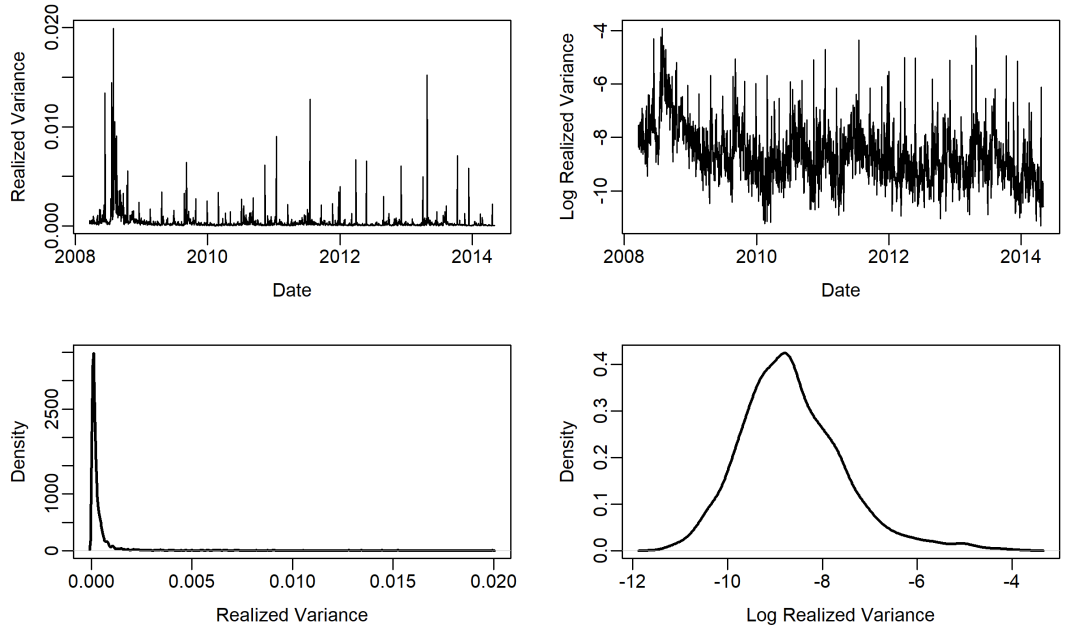

Figure 2.1 – AAPL: Realized and log realized variances – Time series and densities.

influence the number of reported associations, they do not materially alter the qualitative nature of the results. Implementing such size corrections would mainly alter the number of detected associations without changing the overall interpretation or qualitative conclusions. Since our goal is to understand general patterns rather than make strict inferential claims, we consider multiple testing adjustments unnecessary for this study, though they could be incorporated in future work if needed.

Figure 2.2 displays the estimated volatility transmission networks for the entire sample during the 2010–2014 period on the four projection horizons. Each directed edge represents a statistically significant spillover from one asset to another at a given horizon. A key observation is that the number of detected connections generally declines as the horizon increases, indicating that volatility spillovers are more pronounced over shorter horizons and tend to dissipate over time.

To quantify the differences between estimators, we compare the Wald tests based on deLS and OLS estimation over the same period. At the daily horizon ($h = 1$), the deLS estimator identifies 71 connections, while OLS detects 62 (see Appendix

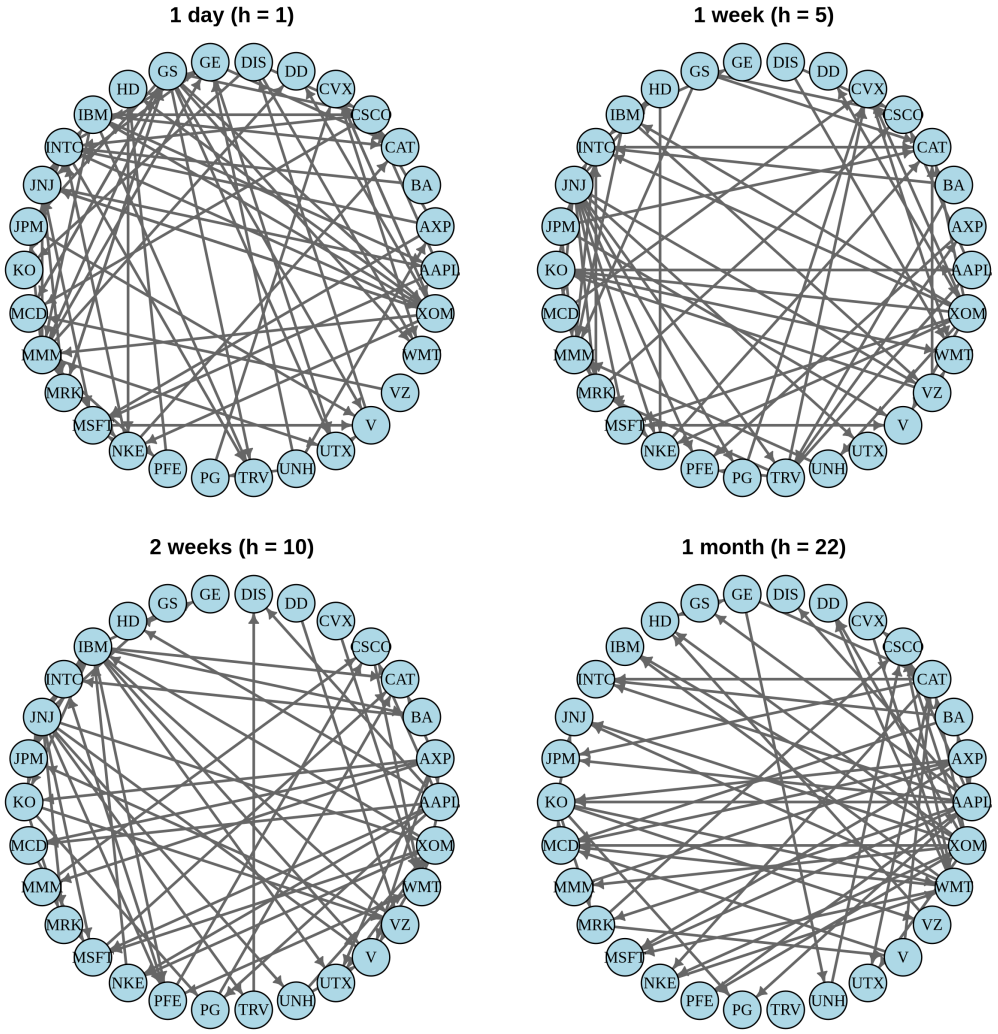

Figure 2.2 – Volatility networks for the 2010–2014 sample period estimated with de-biased least squares for different projection horizons. The Wald test based on de-LS detects networks consisting of 71, 61, 62, and 58 connections for projection horizons of 1 day, 1 week, 2 weeks, and 1 month, respectively.

Figure 2.5), with 35 connections in common. At the weekly horizon ($h = 5$), the two methods yield 61 (deLS) and 52 (OLS) connections, respectively, with 20 overlapping connections. For the biweekly horizon ($h = 10$), deLS identifies 62 connections and OLS 64, with 28 shared links. Finally, at the monthly horizon ($h = 22$), the deLS test detects 58 connections, whereas the OLS identifies 114, with 33 in common. These comparisons suggest that the deLS estimator produces more parsimonious and stable networks, particularly at longer horizons, where OLS may overfit due to limited signal-to-noise ratios and the absence of regularization.

To assess the stability of the network structure, we repeat the analysis over a shorter time period. Specifically, we consider the 2013–2014 subsample, which comprises 488 trading days. This more limited window introduces a higher dimensionality relative to the sample size (note that the HVAR model of realized volatilities with 30 assets leads to an HLP equation containing $30 \times 3 = 90$ variables) and helps mitigate concerns about structural instability. We deliberately exclude major market disruptions such as the 2008 global financial crisis and the 2011 U.S. debt-ceiling crisis (Baker et al., 2019), allowing for a cleaner evaluation of volatility transmission mechanisms during more stable conditions. Figure 2.3 depicts the resulting networks at each projection horizon. For the projection horizon $h = 1$, the Wald test based on deLS identifies 34 significant connections in the volatility transmission network, while the OLS-based test detects 127 connections (see Appendix Figure 2.6), with 17 connections in common. At horizon $h = 5$, the deLS test reveals 17 connections compared to 102 from OLS, sharing 9 common links. For $h = 10$, deLS identifies 15 connections, while OLS yields 170, with 13 overlapping connections. Finally, at the longest horizon $h = 22$, the deLS test detects 11 connections, while OLS produces 198, with only 6 connections in common.

These findings highlight key differences between the deLS and OLS estimators in finite-sample, high-dimensional contexts. OLS-based Wald tests, when applied without regularization, are prone to overfitting and inflated false positive rates, particularly when the number of predictors is large relative to the sample size. In contrast,

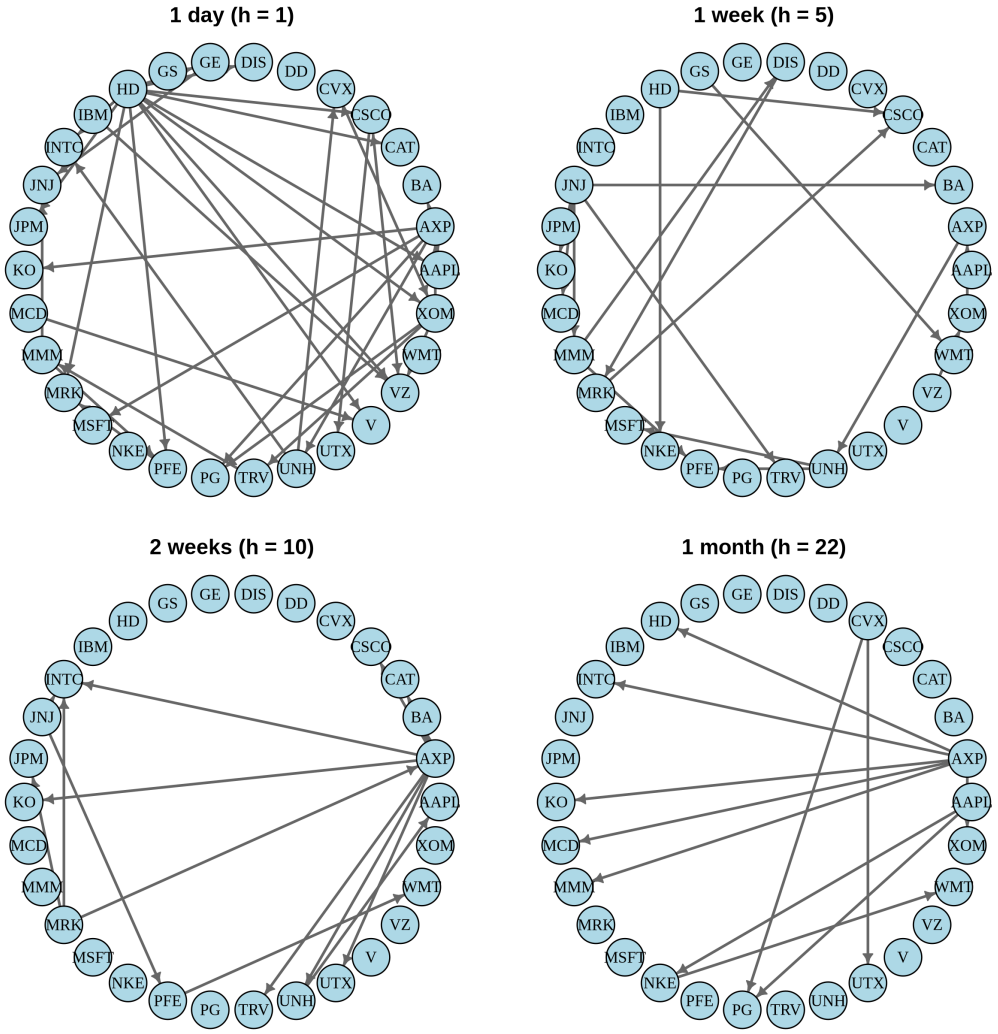

Figure 2.3 – Volatility networks for the 2013–2014 sample period estimated with de-biased least squares for different projection horizons. The Wald test based on de-LS detects networks consisting of 34, 17, 15, and 11 connections for projection horizons of 1 day, 1 week, 2 weeks, and 1 month, respectively.

the deLS estimator incorporates a de-biasing correction following variable selection via regularization, thereby mitigating shrinkage bias and improving finite-sample performance. As a result, deLS tends to yield sparser, more stable networks that are less sensitive to noise. These differences become especially salient at longer horizons, where the signal-to-noise ratio diminishes and the benefits of regularization-based correction are more pronounced.

2.4.3. Network of realized variances and covariances

In this subsection, we extend the analysis to investigate potential spillover effects from realized correlations to realized variances. Although the previous subsection was high-dimensional only when restricted to a shorter sample period, the inclusion of realized correlations on the order of $30^2 = 900$ enables the full-sample analysis to be inherently high-dimensional. It is plausible that changes in correlations influence volatility dynamics, and omitting these variables may result in biased inference due to omitted variable bias. By incorporating correlations as additional predictors within the HVAR framework, we are able to assess how the estimated volatility transmission network evolves when conditional dependencies on correlation dynamics are accounted for.

Let y_{1t} denote the 30-dimensional vector of log-realized variances and y_{2t} the .-dimensional vector of Fisher z-transformed realized correlations.[7] Define the joint process as $y_t = (y_{1t}', y_{2t}')'$. Following Hecq et al. (2023), we assume that y_t follows a high-dimensional vector autoregressive (HVAR) process of dimension 465.

We apply Wald tests based on deLS and OLS estimation to this extended set-

7. The arctanh transformation (inverse hyperbolic tangent) is applied to map correlations—bounded between -1 and 1—onto the real line. This transformation ensures the transformed variables are unbounded, making them suitable for linear modeling and multivariate analysis that assume unrestricted support. The arctanh function is defined as:

$$\operatorname{arctanh}(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right).$$

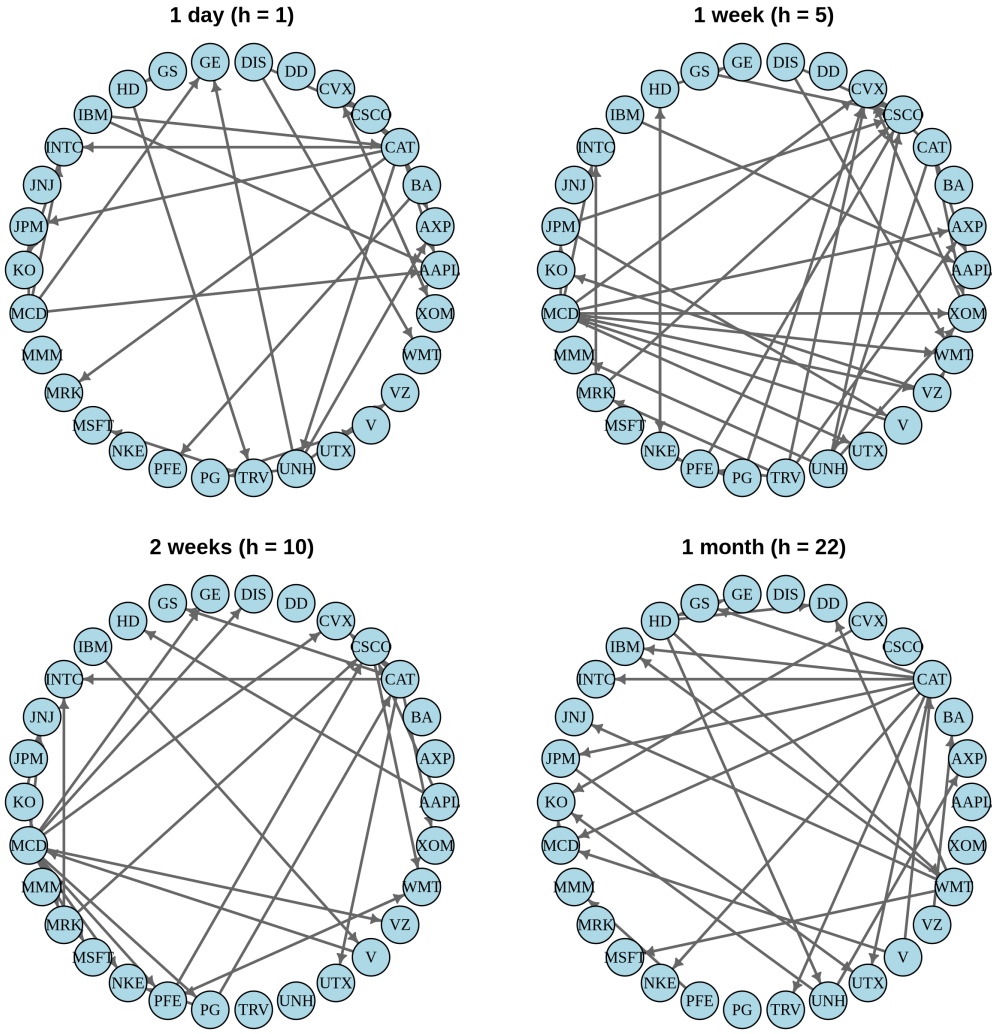

Figure 2.4 – Volatility networks after controlling for realized correlations, obtained for the 2010–2014 sample period and estimated with de-biased least squares for different projection horizons. The Wald test based on de-LS detects networks consisting of 28, 38, 28, and 26 connections for projection horizons of 1 day, 1 week, 2 weeks, and 1 month, respectively.

ting using the 2010–2014 subsample and evaluate the structure of the network in four horizons. At the daily horizon ($h = 1$), deLS identifies 28 significant volatility spillovers, while OLS detects 593 (see Appendix Figure 2.7), with 18 connections in common. At the weekly horizon ($h = 5$), the deLS test yields 38 connections, compared to 615 from OLS, with 29 overlapping. At the biweekly horizon ($h = 10$), deLS identifies 28 connections and OLS 591, with 22 shared links. Finally, at the monthly horizon ($h = 22$), deLS detects 26 connections, while OLS identifies 637, with 17 connections in common.

These results again illustrate the disparity in network density produced by the two estimators. The inclusion of realized correlations magnifies the high-dimensionality of the system, which appears to exacerbate the overfitting tendency of OLS-based inference. In contrast, the deLS estimator, through its regularization and de-biasing mechanism, maintains a more parsimonious representation of the volatility transmission structure. This contrast reinforces the value of regularization-based approaches in high-dimensional predictive settings, particularly when incorporating complex interdependencies among volatility and correlation measures.

2.5. Conclusion

This paper introduces a high-dimensional heterogeneous local projection framework to study volatility transmission across financial assets at multiple horizons. By extending the HAR-RV model into a multivariate local projection setting, we develop a flexible approach that captures horizon-specific dynamics while accommodating a large cross-section of predictors. To address the challenges of statistical inference in high-dimensional systems, we apply the de-biased least squares estimator introduced by Dettaa and Wang (2024), which remains valid without assuming sparsity in the projection coefficients.

To enable reliable inference, we employ the de-biased least squares estimator, which corrects for the bias introduced by regularization and high dimensionality,

allows for valid Wald-type testing without requiring sparsity on multi-horizon. Applying this framework to high-frequency data on U.S. equities, we estimate horizon-specific volatility transmission networks and document clear patterns of dynamic connectedness. Incorporating realized correlations into the system further enriches the structure of the networks and highlights the flexibility of the method in handling additional sources of information under high dimensionality.

In general, this study provides a robust and flexible approach to modeling and inferring volatility spillovers in high-dimensional systems. It contributes to the growing literature on financial network connectedness by offering a multi-horizon perspective and a tractable inferential framework suited to modern financial data. Extensions to time-varying structures, nonlinear dynamics, or cross-market interactions present promising directions for future work.

Bibliography

D. W. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, pages 817–858, 1991.

S. R. Baker, N. Bloom, S. J. Davis, and K. J. Kost. Policy news and stock market volatility. Technical report, National Bureau of Economic Research, 2019.

M. Barigozzi and C. Brownlees. Nets: Network estimation for time series. *Journal of Applied Econometrics*, 34(3):347–364, 2019.

F. Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009.

F. Corsi, F. Lillo, D. Pirino, and L. Trapin. Measuring the propagation of financial distress with granger-causality tail risk networks. *Journal of Financial Stability*, 38:18–36, 2018.

G. Cubadda, A. Hecq, and A. Riccardo. Forecasting realized volatility measures with multivariate and univariate models: The case of the us banking sector. In *Financial Mathematics, Volatility and Covariance Modelling*, pages 286–307. Routledge, 2019.

M. Demirer, F. X. Diebold, L. Liu, and K. Yilmaz. Estimating global bank network connectedness. *Journal of Applied Econometrics*, 33(1):1–15, 2018.

E. Dettaa and E. Wang. Inference in high-dimensional linear projections: Multi-horizon granger causality and network connectedness. *arXiv preprint arXiv:2410.04330*, 2024.

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.

M. Friedman and A. J. Schwartz. *A monetary history of the United States, 1867-1960*, volume 9. Princeton University Press, 2008.

M. G. Garcia, M. C. Medeiros, and G. F. Vasconcelos. Real-time inflation forecasting with high-dimensional models: The case of brazil. *International Journal of Forecasting*, 33(3):679–693, 2017.

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.

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

J. Lin and G. Michailidis. Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models. *Journal of Machine Learning Research*, 18(117):1–49, 2017.

M. McAleer and M. C. Medeiros. Realized volatility: A review. *Econometric reviews*, 27(1-3):10–45, 2008.

M. C. Medeiros and G. F. Vasconcelos. Forecasting macroeconomic variables in data-rich environments. *Economics Letters*, 138:50–52, 2016.

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

W. K. Newey and K. D. West. Automatic lag selection in covariance matrix estimation. *The review of economic studies*, 61(4):631–653, 1994.

T. Vỳrost, Š. Lyócsa, and E. Baumöhl. Granger causality stock market networks: Temporal proximity and preferential attachment. *Physica A: Statistical Mechanics and its Applications*, 427:262–276, 2015.

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

H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173 – 2192, 2007. doi: 10.1214/009053607000000127. URL https://doi.org/10.1214/009053607000000127.

2.6. Appendix

2.6.1. Additional empirical results: OLS-based networks

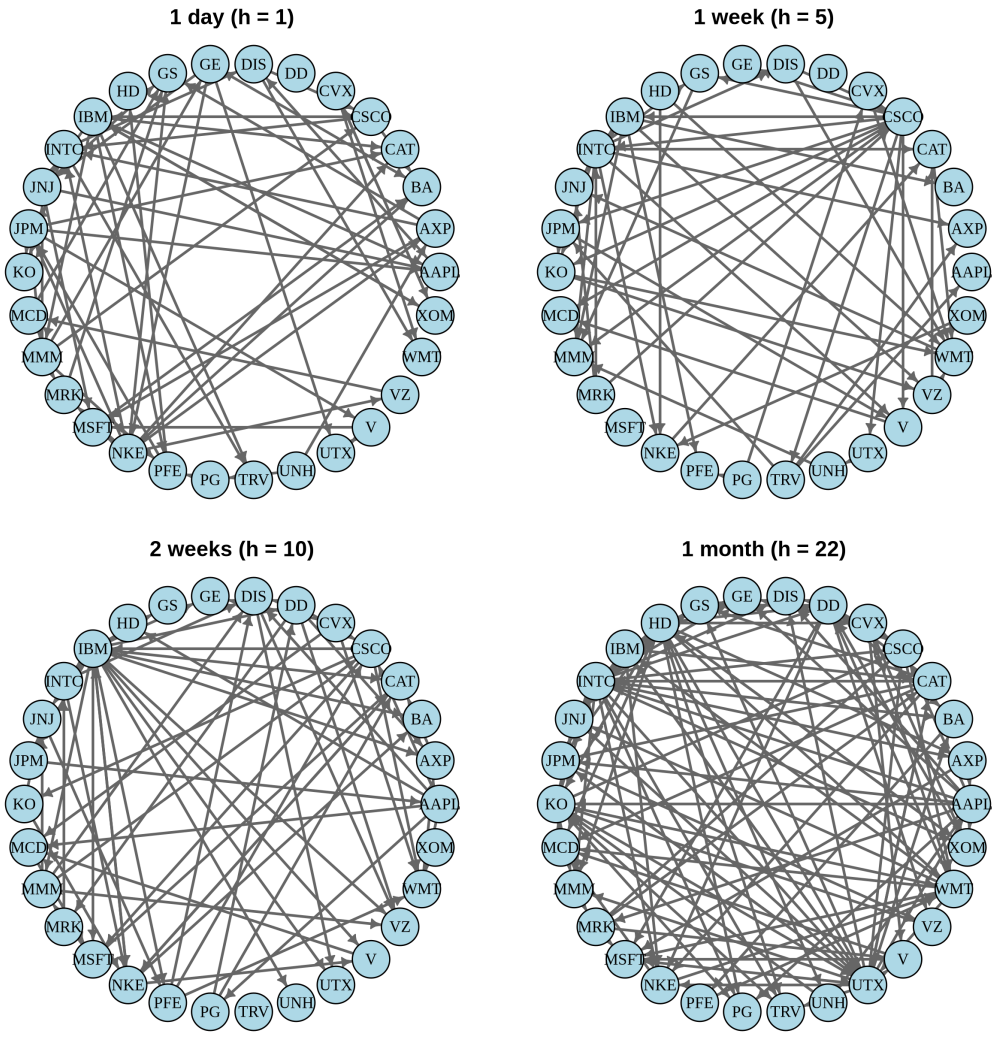

Figure 2.5 – Volatility networks for the 2010–2014 sample period estimated with least squares for different projection horizons. The Wald test based on OLS detects networks consisting of 62, 52, 64, and 114 connections for projection horizons of 1 day, 1 week, 2 weeks, and 1 month, respectively.

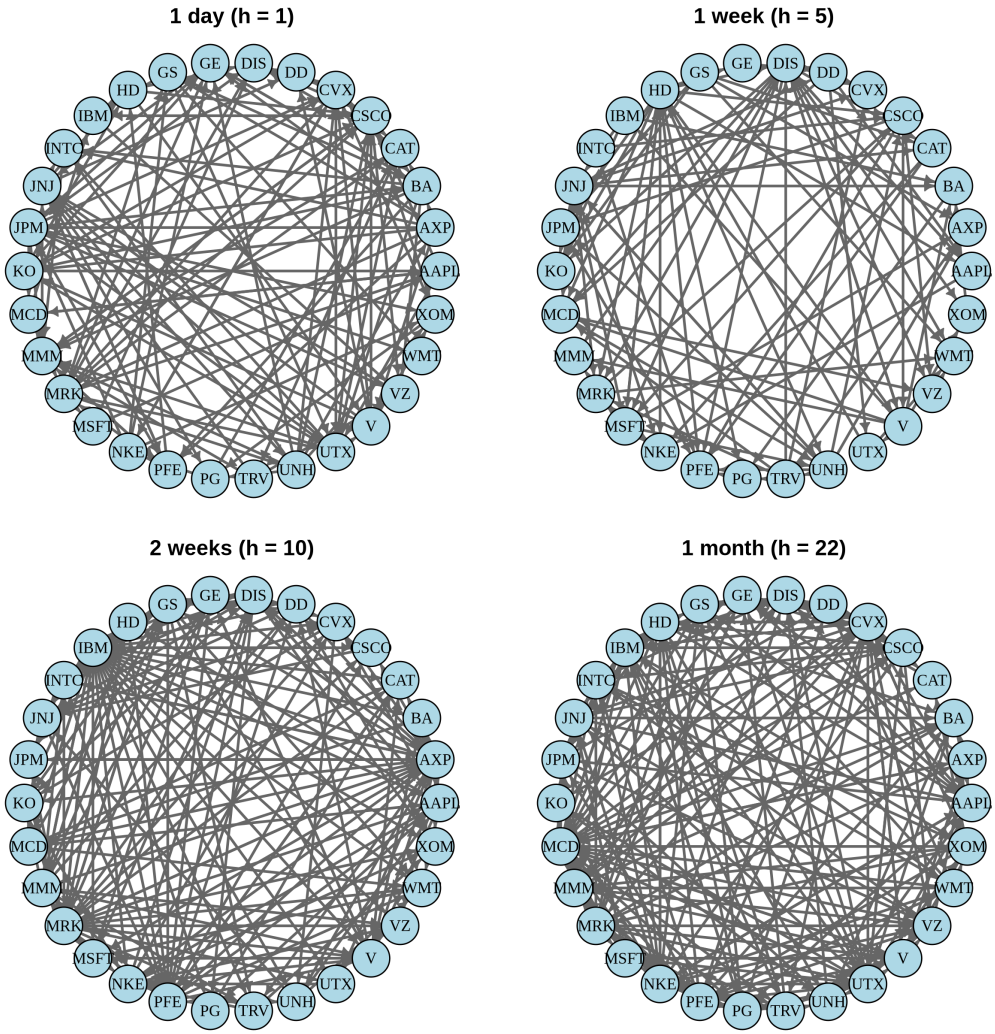

Figure 2.6 – Volatility networks for 2013–2014 sample period estimated with least squares for different projection horizons. The Wald test based on OLS detects networks consisting of 127, 102, 170, and 192 connections for projection horizons of 1 day, 1 week, 2 weeks, and 1 month, respectively.

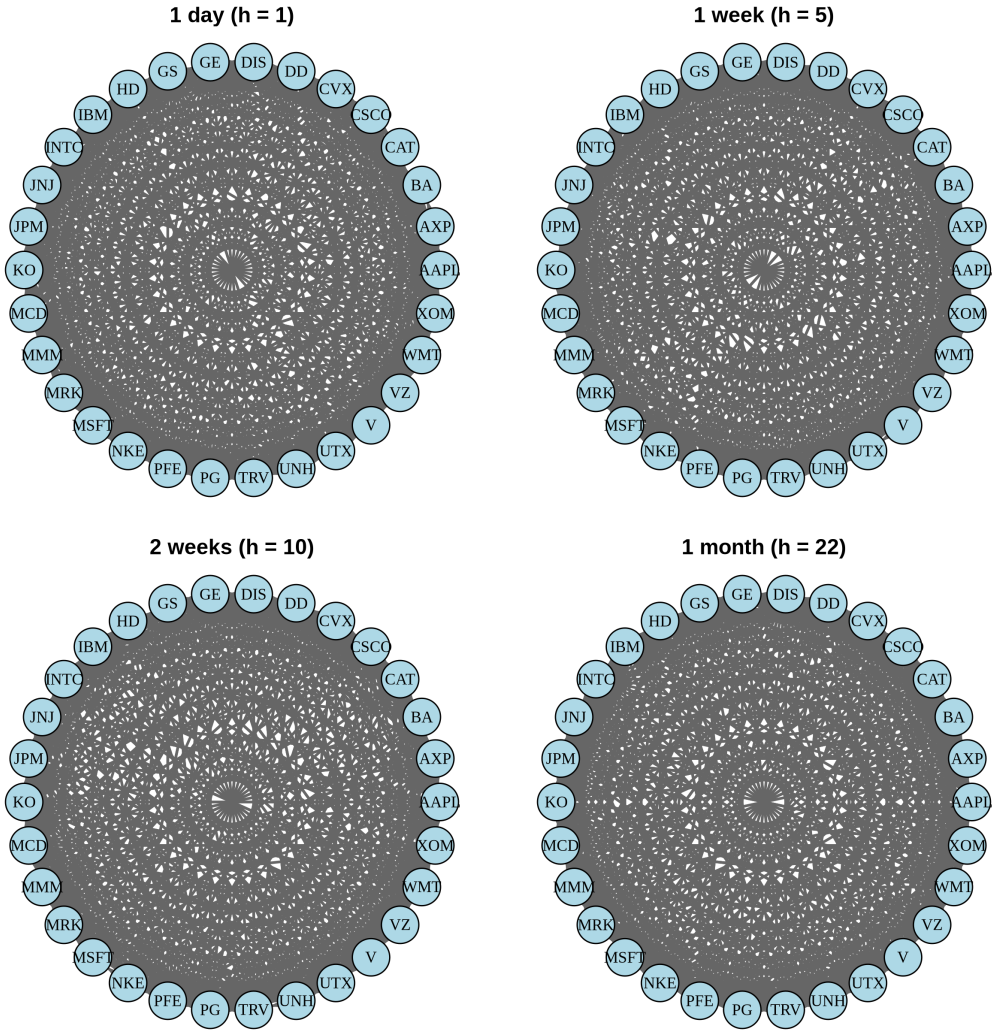

Figure 2.7 – Volatility networks after controlling for realized correlations, obtained for the 2010–2014 sample period and estimated with OLS for different projection horizons. The Wald test based on OLS detects networks consisting of 593, 615, 591, and 637 connections for projection horizons of 1 day, 1 week, 2 weeks, and 1 month, respectively.

CHAPTER 3

Ridge-Regularization for Moment-based Estimation in High-Dimensional Settings*

3.1. Introduction

The issue of efficient estimation of models involving many instruments/moments is an important part of the econometric literature. This paper considers the efficient estimation of a parameter of interest defined by a single conditional moment restriction. This framework is compatible with several applications in microeconomic data where a large set of valid instruments is available. Examples include an influential instrumental variable study that estimates the economic return to schooling. Angrist and Frandsen (2022) revisiting the Angrist and Krueger (1991)'s application, use up to $1,530$ instruments for schooling by interacting quarter of birth dummies, year of birth dummies, and state of birth dummies. Altonji et al. (2013) examine a joint model of earnings, employment, job changes, wage rates, and work hours over a career with a full specification of $2,429$ moments. Eaton et al. (2011) explore the sales of French manufacturing firms in 113 destination countries with $1,360$ moments. Han et al. (2005) investigate the cost efficiency of the Spanish saving banks in a time-varying coefficient model with 872 moments.

Efficient estimation of models with conditional moment restrictions poses a challenging problem. Indeed, many unconditional moment restrictions can be obtained from nonlinear transformations of an exogenous variable or by using interactions between various exogenous variables. Selecting a small number of moments may lead

⋆. This chapter is co-authored with Marine Carrasco. The authors thank the participants of the African Meeting of the Econometric Society (Abidjan, 2024), the 1st CIREQ Interdisciplinary PhD Student Conference on Big Data and Artificial Intelligence (Montreal, 2023), the 57th Annual Canadian Economics Association (CEA) Meetings (Winnipeg, 2023), and the 62nd Annual Congress of the Société canadienne de science économique (Quebec City, 2023). Carrasco thanks SSHRC for partial financial support.

to a loss of efficiency in finite dimensions since a single conditional moment restriction is equivalent to an infinite countable sequence of unconditional moments under certain conditions (see, Chamberlain, 1987; Donald et al., 2003). However, all the information in the conditional moment restriction will eventually be accounted for by allowing the number of unconditional moments to grow with the sample size, thus achieving asymptotic efficiency (Donald et al., 2003). This suggests that, rather than using the infinite set of moments available, one can use a reasonably large number of them (even larger than the sample size) to gain efficiency in applications.

GMM-type estimation with a large number of moments presents certain challenges. Firstly, while it can improve asymptotic efficiency, using an excessive number of moments can deteriorate the finite sample properties of the GMM estimator. Specifically, the standard two-step GMM estimator can exhibit significant bias and/or inaccuracy, even in applied work with a large number of observations (see, Hahn and Hausman, 2003; Hansen et al., 2008; Newey and Smith, 2004; Newey and Windmeijer, 2009). This trade-off between variance and bias is known in the literature as the "*many moments problem*". The benchmark estimator we consider is the Hansen et al. (1996)'s continuous updating estimator (CUE), due to its relatively low bias when dealing with a large number of moments. Secondly, the weighting matrix used in the CUE's objective function may become singular when dealing with a substantial number of moment conditions. Consequently, the CUE estimator may become infeasible or exhibit poor finite sample properties. Lastly, the CUE estimator is not guaranteed to have finite moments of any order, resulting in the undesirable property of significant dispersion in the estimates (see, Guggenberger, 2005; Hausman et al., 2011, 2012). This property is referred to in the literature as the "no-moments problem" of CUE.

We introduce ridge-type regularization in the weighting matrix to address the singularity problem. The resulting estimator is referred to as the ridge-regularized CUE (RRCUE). Unlike many conventional methods such as the standard CUE, our RRCUE estimator remains feasible even when the sample counterpart of the weighting matrix

is unstable or, worse, singular—in particular when the number of moment conditions exceeds the sample size. We demonstrate that the RRCUE estimator can be derived through L2 penalization of the generalized empirical likelihood (GEL) representation of CUE. We establish that RRCUE is consistent and asymptotically normal, subject to certain restrictions on the convergence rate of the regularization parameter. Our double asymptotic framework allows for both the sample size and the number of moment conditions to go to infinity. Furthermore, we show that the asymptotic variance of RRCUE achieves the Chamberlain (1987)'s semiparametric efficiency bound. To assess the benefits of regularization, we conduct a Monte Carlo simulation. Our findings demonstrate that regularization can help address the no-moments problem observed in CUE by reducing its dispersion. Additionally, regularization helps to improve the efficiency of the CUE estimator in finite samples, though it introduces some bias. However, this bias remains smaller than the GMM overidentification bias in nearly all the settings we consider.

This article contributes to the extensive literature on many instruments/moments. This literature can be divided into two parts. The first part, which dates back to Bekker (1994), focuses on an asymptotic framework where the number of moments, denoted by K, grows with the sample size, denoted by n, but remains relatively small (see, e.g., Chao and Swanson, 2005; Donald and Newey, 2001; Donald et al., 2003, 2009; Hansen et al., 2008; Hausman et al., 2012; Newey and Windmeijer, 2009, among others). Specifically, Hansen et al. (2008) derived asymptotics properties of the limited information maximum likelihood (LIML) and Fuller (1977) estimators under a variety of many instrument asymptotics, including the many instrument sequence of Bekker (1994) and the many weak instruments sequence of Chao and Swanson (2005) and Stock and Yogo (2005). Newey and Windmeijer (2009) derived properties of CUE under many weak moment conditions as those of Hansen et al. (2008). Donald et al. (2003) demonstrated that the generalized empirical likelihood (GEL) class, including CUE, offers consistent and asymptotically normal estimators for models with conditional moment restrictions, which attain the semiparametric

efficiency bound under a stringent condition on the growth rate of K relative to n. Donald and Newey (2001) and Donald et al. (2009) proposed a method to select the optimal number of moments by minimizing an approximate mean square error derived from a higher-order expansion.

This paper falls into the second strand of literature that does not require selecting moments. Papers in this strand allow for the sample size to grow faster than the number of moments (see, for example, Belloni et al., 2012; Carrasco, 2012; Carrasco and Tchuente, 2015; Shi, 2016, among others). In particular, in the linear instrumental variables framework, Belloni et al. (2012) recommended using LASSO in the first step to construct the optimal instrument when assuming the approximate sparsity of the first-stage equation. However, the empirical support for sparse models in macroeconomics, microeconomics, and finance is generally weak, as shown by Giannone et al. (2021). The methodology developed in this paper is therefore relevant, as it is robust to a *dense* first-stage equation. Carrasco (2012) proposed three regularized estimators to improve the small sample properties of the standard two-stage least squares (2SLS) estimator when dealing with a large number of instruments. Carrasco and Tchuente (2015) extended this regularization approach to the limited information maximum likelihood (LIML) estimator and demonstrated that the regularized LIML estimator has finite first moments provided that the sample size is sufficiently large. This paper can be seen as an extension of Carrasco and Tchuente (2015) from efficient estimation of linear homoskedastic models using LIML to potentially nonlinear heteroskedastic models using the continuous updating estimator (CUE). Other papers in the literature address specific issues such as heteroskedasticity and the no-moments problem. Hausman et al. (2012) addressed the problem of many instruments in heteroskedastic data and recommended using a jackknife version of the Fuller (1977) estimator in applications that align with this framework. Hansen and Kozbur (2014) proposed a ridge-regularized version of the jackknife instrumental variable estimator (JIVE) that is robust to heteroskedasticity in the presence of many instruments. Hausman et al. (2011) proposed a modification of the continuous updating estimator

(CUE) to address the no-moments problem and improve the finite sample properties of the standard CUE in time series settings with many weak moment conditions. Our regularization scheme, although simpler, is similar to theirs. Chang et al. (2015) proposed a penalized generalized empirical likelihood (GEL) estimator, but their setup is different, as they directly penalized a certain norm of the parameter of interest in the GEL form. The GEL representation of RRCUE obtained in this paper penalizes instead the l_2 norm of the auxiliary variable that appears in that representation. In a more recent work, Angrist and Frandsen (2022) investigated the performance of machine learning (ML) for instrument selection. They argued that the optimal way to leverage the ML toolkit is to combine it with a sample-splitting procedure. Also, standard LIML surprisingly performs well in their simulation framework, but there is no reason to believe that this will generally be the case, as their data generating process is very specific and motivated by the Angrist and Krueger (1991) empirical application.

The rest of the paper is structured as follows. Section 3.2 introduces the framework. Section 3.3 presents the Ridge regularized continuous updating estimator (RRCUE). In Section 3.4, we derive asymptotic properties of the RRCUE. In Section 3.5, we suggest a data-driven procedure for selecting the optimal regularization parameter using cross-validation. We conduct a large-scale Monte Carlo experiment to evaluate the benefits of regularization in Section 3.6. Section 3.7 applies our method to estimate the impact of institutions and government policies on productivity. Section 3.8 concludes our findings, while technical proofs and additional lemmas are provided in the Appendix.

3.2. The framework and moment restrictions

We consider an environment where there are many unrestricted moment conditions generated by a single conditional moment restriction (CMR) like Chamberlain (1987) and Donald et al. (2003, 2009). To describe this setting, let w denote a single

observation from an i.i.d. sequence $(w_1, w_2, \ldots)$, $\beta \in \mathscr{B}$ a p-dimensional parameter vector, and $\rho(w, \beta)$ a scalar that can be seen as a residual. z is a subvector acting as conditioning variables such that for a value β_0 of the parameters

$$E[\rho(w, \beta_0)|z] = 0, \tag{3.1}$$

where $E[\cdot|z]$ is the expectation taken with respect to the conditional distribution of w given z.

We rely on a GMM-type estimator to address the issue of efficient estimation of the parameter of interest β_0. We need a vector of unrestricted moment conditions for this purpose. It is well known that a conditional moment restriction as in Eq. (3.1) is equivalent to a countable number of unconditional moment restrictions under certain conditions like the one in Assumption 3.1(b) below, see e.g., Chamberlain (1987). Following Donald et al. (2003, DIN03 hereafter), our set of unconditional moment restrictions is based on splines or other approximating functions like power series. For each positive integer K, let $q^K(z) = (q_{1K}(z), \ldots, q_{KK}(z))'$ be a K-dimensional vector of instruments, also referred to as approximating functions [1]. Under Assumption 3.1(b) below DIN03 showed that the conditional moment restriction of (3.1) is equivalent to a sequence, indexed by $K \in \mathbb{N}$, of unconditional moment restrictions of the following form:

$$E[g(x, \beta_0)] = 0, \tag{3.2}$$

where $g(x, \beta) = q^K(z)\rho(w, \beta)$ and $x = (w', z')'$. The immediate consequence of this result is that an efficient estimation of β_0 under CMR of Eq. (3.1) can be obtained from the sequence of unconditional moments of Eq. (3.2) by letting K grow with the sample size n. Indeed, all the information in the CMR will be eventually accounted for by letting K go to infinity with n. For notational convenience, we omit the K superscript on $g(w, \beta)$, which indicates its dependence on the number K of

1. This terminology is widely used in the literature on efficient estimation under conditional moment restrictions and is motivated by the main objective of using the vector of instruments, $q^K(z)$, to linearly approximate a certain square-integrable function of z, typically the 'optimal instrument'.

approximating functions.

In the literature, there are several choices of the approximating functions, $q_{kK}(z)$, including power series, splines, and Fourier series. In this article, we will focus on the first two. They both have faster approximation rates for smoother functions, up to the order of the spline or the power series. Unlike power series, spline approximations are not severely affected by singularities (e.g., discontinuities) in the function being approximated. To describe $q^K(z)$ in detail, consider the simple case where z is a scalar. In this case, the vector of power series approximating functions is given by

$$q^K(z) = \left(1, z, z^2, \ldots, z^{K-1}\right)'. \tag{3.3}$$

For splines, let s be a positive scalar denoting the order of the spline. Let $t_1, \ldots, t_{K-s-1}$ denote knots and let $\xi(z) = z1_{\{z>0\}}$, where 1_A denotes the indicator function for the event A. Then a vector of spline approximating functions is given by

$$q^K(z) = \left(1, z, \ldots, z^s, \xi\left(z - t_1\right)^s, \ldots, \xi\left(z - t_{K-s-1}\right)^s\right)'. \tag{3.4}$$

The most common specification is $s = 3$. In practice, it is recommended to choose the knots t_j in the observed data range of z, see e.g. Donald et al. (2003). We impose the following conditions on the sequence $q^K(z)$ and the distribution of z. Let $\mathscr{Z}$ denote the support of z.

Assumption 3.1. *(a) For each K there is a constant $\zeta(K) \geq \sqrt{K}$ and a positive constant C such that: $E\left[q^K(z)'q^K(z)\right]$ is finite and $\sup_{z\in\mathscr{Z}} \left\|q^K(z)\right\| \leq C\zeta(K)$, (b) For any $a(z)$ with $E\left[a(z)^2\right] < \infty$ there are $K \times 1$ vectors γ_K such that as $K \longrightarrow \infty$, $E\left[\left\{a(z) - \tilde{q}^K(z)'\gamma_K\right\}^2\right] \longrightarrow 0$, where $\tilde{q}^K(z) = q^K(z)/\zeta(K)$, (c) For each K, the matrix $E\left[\tilde{q}^K(z)\tilde{q}^K(z)'\right]$ has only nonzero eigenvalues. For any function $\psi(w)$ belonging to the set $\left\{\rho(w,\beta) : \beta \in \mathscr{B}\right\} \cup \left\{\dfrac{\partial \rho\left(w, \beta_0\right)}{\partial \beta_k} : k = 1, \ldots, p\right\}$, its conditional expectation*

$b(z) := E[\psi(w) \mid z]$ is square-integrable, i.e., $E\left[b(z)^2\right] < \infty$, and[2]

$$\sum_{j=1}^{\infty} \frac{\left(E\left[b(z)\tilde{q}^K(z)\right]' \phi_j\right)^2}{\lambda_j} < \infty, \tag{3.5}$$

where $\left(\lambda_j, \phi_j,\ j = 1, 2, \ldots, K\right)$ are the eigenvalues and corresponding orthonormal eigenvectors of the $K \times K$ symmetric and positive semidefinite matrix $M := E\left[U(z)\tilde{q}^K(z)\tilde{q}^K(z)'\right]$. The function $U(z)$ is either equal to 1 or to $E\left[\rho\left(w, \beta_0\right)^2 \mid z\right]$, which is assumed to be bounded away from zero.

Assumption 3.1(a) is similar to a normalization of the approximating functions like that adopted by Newey (1997) and Donald et al. (2003, 2009). Assumption 3.1(b) is similar to Assumption 1 of DIN03. DIN03 argues that its specific role is to obtain estimators that achieve the Chamberlain (1987)'s semiparametric efficiency bound by ensuring that linear combinations of $q^K(z)$ can approximate certain square integrable function of z. The bound $\zeta(K)$ plays an important role in the asymptotic theory for GMM and the generalized empirical likelihood (GEL) class of estimators developed by DIN03. DIN03 showed under some mild conditions that Assumption 3.1(b) is sufficient to obtain the asymptotic efficiency of the continuous updating estimator (CUE), known as being an element of the GEL class (Newey and Smith, 2004), if the condition $\zeta(K)^2 K^2/n \longrightarrow 0$ holds. Such a condition on the growth rate of K restricts the number K of moment conditions that can be used in applications. We have shown that Assumption 3.1(b) is no longer sufficient to obtain the asymptotic efficiency of the regularized estimator that we will introduce in the next section. An additional condition given by Assumption 3.1(c) is required. On the one hand, Assumption 3.1(c) implies that eigenvalues of the matrix M are all non-zero for fixed

2. The sum in Eq. (3.5) is defined as

$$\lim_{K\to\infty} \sum_{j=1}^{K} \frac{\left(E\left[b(z)\tilde{q}^K(z)\right]' \phi_j\right)^2}{\lambda_j}.$$

K although they can converge to zero if K grows with the sample size [3]. On the other hand, condition (3.5) is similar to that used by Carrasco (2012) and Carrasco and Tchuente (2015). This condition is important to obtain the asymptotic efficiency of the regularized estimator. More precisely, as pointed out by Carrasco et al. (2007), this regularity condition will facilitate the calculation of the rate of convergence of the regularization bias.

We show, under Assumption 3.1 and some regularity conditions, that regularizing the second moment of $g(w,\beta)$ allows us to get rid of any constraint on the number of moments K, such as the one imposed by DIN03 to obtain asymptotic efficiency. Indeed, our convergence rates no longer depend on K but only on the sample size n and the regularization parameter α. For example we show that the regularized version of CUE is consistent under the asymptotic framework where K goes to infinity and α goes to zero as the sample size goes to infinity with the following restriction on the convergence rate of α relative to n: $\alpha^{-2}n^{-1/2+1/\gamma} \to 0$. Asymptotic normality of the regularized estimator requires a stronger restriction, that is, $\alpha^{-5/2}n^{-1/2+1/\gamma} \to 0$. The parameter $\gamma > 2$ is specified as in Assumption 3.2 below.

An explicit formula for $\zeta(K)$ is available for several cases. For example, it has been shown under mild conditions that $\zeta(K) = \sqrt{K}$ for splines and $\zeta(K) = K$ for power series, see e.g., Newey (1997) among others. Under Assumption 3.1(a), $E\left[q^K(z)'q^K(z)\right] = O(\zeta(K)^2)$ and therefore the second moment matrix of the approximating functions is a trace-class matrix [4] (even for large K) if they are normalized by the bound $\zeta(K)$. We show in this paper that this type of normalization is useful to obtain convergence rates that do not depend on the number K of moment conditions. DIN03 used an additional condition, that is $A := E\left[\tilde{q}^K(z)\tilde{q}^K(z)'\right]$ has a smallest eigenvalue bounded away from zero uniformly in K. This restriction limits the number of moments K that can be used in practice and prevents handling cases where

3. Which is less restrictive than the condition used by Donald et al. (2003), that is, $E\left[\tilde{q}^K(z)\tilde{q}^K(z)'\right]$ has smallest eigenvalue bounded away from zero uniformly in K.

4. A matrix is said to be trace-class matrix if its trace is a finite number. If a trace-class matrix is symmetric, then its maximum eigenvalue is bounded above.

the matrix A is nearly singular due to a large number of instruments. Regularization will allow us to avoid such a condition by allowing A to be nearly singular as K grows with the sample size [5].

3.3. Ridge-regularized version of CUE

One of the standard approaches to efficiently estimate the parameter of interest β_0 defined by the conditional moment restriction of Eq. (3.1) is to use GMM-type estimators based on unconditional moment restrictions as specified in Eq. (3.2). The role of the approximating functions in this specification, in terms of unconditional moment restrictions, is to make the CMR approximately satisfied in the sample. Asymptotic efficiency is then obtained by letting K grow with n at a certain rate (Donald et al., 2003). To gain efficiency in the finite sample, the use of many instruments may be necessary. As is well known, one of the costs of using many approximating functions is that the sample counterpart of the second moment matrix of the moment function, $g(w,\beta)$, might be ill-conditioned. Following Carrasco and Florens (2000), Carrasco (2012), and Carrasco and Tchuente (2015) among others, we use regularization to fix this problem.

The benchmark estimator we consider is the continuous updating GMM estimator (CUE) due to its relatively low bias with many moment conditions, see, e.g., Newey and Smith (2004). Before presenting the regularized CUE, we begin by recalling the standard CUE estimator. We introduce here some notations to ease the presentation of estimators. Let $q_i = q^K(z_i)/\zeta(K)$, $\rho_i(\beta) = \rho(w_i,\beta)$, $g_i(\beta) = q_i\rho_i(\beta)$, $\hat{g}(\beta) = n^{-1}\sum_{i=1}^n g_i(\beta)$, $\widehat{\Omega}(\beta) = n^{-1}\sum_{i=1}^n g_i(\beta)g_i(\beta)' = n^{-1}\sum_{i=1}^n \rho_i(\beta)^2 q_i q_i'$, and $\Omega(\beta) = E\left[\rho_i(\beta)^2 q_i q_i'\right]$. The Hansen et al. (1996)'s CUE uses $\widehat{\Omega}(\beta)^{-1}$ as weighting matrix without replacing β by a first step estimator as it is the case for the Hansen

5. The nearly singularity of A refers here to the cases where its minimum eigenvalue, although nonzero for a fixed K, approaches zero as K tends to infinity.

(1982)'s conventional two-step GMM estimator. It is defined by

$$\hat{\beta}_{CUE} = \arg\min_{\beta \in \mathscr{B}} \ \hat{g}(\beta)' \widehat{\Omega}(\beta)^{-1} \hat{g}(\beta). \tag{3.6}$$

In the sequel, we will refer to this estimator as the standard CUE. Note that the vector of approximating functions $q^K(z_i)$ is normalized by the bound $\zeta(K)$ prior to the definition of the CUE estimator. Although the CUE estimator is invariant to this normalization, it is not the case for its regularized version introduced below. However, this normalization allows us to search for the optimal regularization parameter around zero.

Note that the standard GMM estimator and the standard CUE are specialized for cases where $K < n$ and the inverse of $\widehat{\Omega}(\beta)$ is stable. They can be infeasible when the number of moment conditions is close to the sample size, restricting their use in empirical applications. Indeed, when K is larger than n or smaller than n but close to n, the weighting matrix used in the CUE's objective function is singular or nearly singular. To illustrate this fact, assume that the error term $\rho_i(\beta_0)$ is conditional homoskedastic with $E\left[\rho_i(\beta_0)^2 | z_i\right] = \sigma^2$. Then, by the law of iterated expectations, $\Omega := \Omega(\beta_0) = \sigma^2 E\left[q_i q_i'\right]$. If $q := [q_1, \ldots, q_n]'$, then the sample counterpart of Ω, $\widehat{\Omega} := \widehat{\sigma}^2 q'q/n$, is almost singular when $K > n$ or close to n. It is natural to think that in general, the matrix $\widehat{\Omega}(\beta)$ suffers from this singularity problem when $K > n$. Even when K is smaller than n but large, the naive inverse of $\widehat{\Omega}(\beta)$ will be unstable in the sense that a seemingly innocuous change in the sample moment function may induce a large variation of $\widehat{\Omega}(\beta)^{-1}\hat{g}(\beta)$. A matrix with such a property is said to be ill-conditioned.

A Monte Carlo experiment by Hausman et al. (2012) reveals that using a heteroskedasticity-consistent weighting matrix can worsen the finite sample performance of CUE with many moments. We suspect that the instability of the inverse of the weighting matrix plays an important role in the deterioration of the finite sample properties of CUE under many moment restrictions. This instability of the naive inverse of $\widehat{\Omega}(\beta)$ for a

large K is caused by the fact that its condition number is large. Since the condition number is defined as the ratio of the maximum eigenvalue (λ_{max}) and the minimum eigenvalue (λ_{min}), it is large if either λ_{max} is very large or λ_{min} is close to zero. Before regularizing $\widehat{\Omega}(\beta)$, we first normalize the approximating functions $q^K(z_i)$ by the upper bound $\zeta(K)$ so that it becomes a trace-class matrix. Thus, the only source of instability of the inverse of $\widehat{\Omega}(\beta)$ is the fact that its minimum eigenvalue may be close to zero when K is large compared to n.

We propose to stabilize the inverse of $\widehat{\Omega}(\beta)$ using Ridge [6] type regularization. This consists in replacing, in the CUE objective function, the naive inverse $\widehat{\Omega}(\beta)^{-1}$ by the regularized inverse $\left(\widehat{\Omega}(\beta)^{\alpha}\right)^{-1} = \left(\widehat{\Omega}(\beta) + \alpha I_K\right)^{-1}$, where $\alpha > 0$ is the regularization parameter and I_K is the $K \times K$ identity matrix. The resulting estimator that we refer to as the Ridge regularized CUE (RRCUE) depends on the regularization parameter α and is defined by

$$\hat{\beta} = arg\min_{\beta \in \mathscr{B}} \hat{g}(\beta)' \left(\widehat{\Omega}(\beta)^{\alpha}\right)^{-1} \hat{g}(\beta). \tag{3.7}$$

The main purpose of this paper is to study the properties of the RRCUE. Pre-normalization of $q^K(z_i)$ by $\zeta(K)$ plays an important role in deriving the theoretical properties of the RRCUE. It partly allows us to obtain convergence rates that do not depend on the number of moment conditions K, but only on the sample size n and the regularization parameter α. As a consequence, adequately controlling the convergence rate of the regularization parameter to zero is sufficient to obtain the consistency and asymptotic normality of the RRCUE without imposing any constraint on the growth rate of K with respect to n. In terms of finite-sample implementation, normalization also plays a crucial role, as it allows us to select the optimal regularization parameter within a grid around zero. Without normalization, RRCUE would still be feasible, but we would have no guidance on the appropriate range in which to search for the regularization parameter α. Note that, in practice, it is not necessary to know the exact form of $\zeta(K)$. However, for both types of approximating functions we

6. There are other regularization methods. For example, Carrasco (2012) considered three other regularization schemes: Tikhonov, Landweber–Fridman and Principal components.

consider—i.e., splines and power series—suitable choices are $\zeta(K) = \sqrt{K}$ for splines and $\zeta(K) = K$ for power series (see, e.g., Newey, 1997).

Newey and Smith (2004) showed that the standard CUE is part of a class of estimators introduced by Smith (1997, 2001) called generalized empirical likelihood (GEL) estimators. The GEL representation of CUE facilitates the theoretical derivation of asymptotic properties of CUE. To describe GEL, let $s(v)$ be a function of a scalar v that is concave on its domain, an open interval $\mathcal{V}$ containing zero with $s_0 = 0$ and $s_1 = s_2 = -1$ where $s_j(v) = \partial^j s(v)/\partial v^j$. Let $\widehat{\Lambda}(\beta) := \{\lambda : \lambda' g_i(\beta) \in \mathcal{V}, i = 1, \ldots, n\}$. The GEL estimator associated with the concave function s is the solution to a saddle point problem

$$\hat{\beta}_{\text{GEL}} = \arg\min_{\beta \in \mathcal{B}} \sup_{\lambda \in \widehat{\Lambda}(\beta)} n^{-1} \sum_{i=1}^{n} s\left(\lambda' g_i(\beta)\right). \tag{3.8}$$

Newey and Smith (2004) showed that $\hat{\beta}_{CUE} = \hat{\beta}_{\text{GEL}}$ if $s(v)$ is quadratic, e.g., if $s(v) = -v - v^2/2$. The following theorem establishes a similar result for the RRCUE.

Theorem 3.1. *If Assumption 3.1 (a) is satisfied, then for* $s(v) = -v - v^2/2$,

$$\hat{\beta} = \underset{\beta \in \mathcal{B}}{\text{argmin}} \sup_{\lambda \in \widehat{\Lambda}(\beta)} \hat{P}(\beta, \lambda) \text{ where } \hat{P}(\beta, \lambda) = n^{-1} \textstyle\sum_{i=1}^{n} s\left(\lambda' g_i(\beta)\right) - \dfrac{\alpha}{2} \lambda' \lambda.$$

This result shows that the regularized CUE can be obtained alternatively by penalizing the L^2 norm of λ in the GEL criterion. This is an important result that will be useful for deriving asymptotic properties of RRCUE.

Here we give first-order conditions for RRCUE which are useful for deriving first-order asymptotic properties. We need some notations. Let $\hat{\pi}_i$ $(i = 1, \cdots, n)$ denote the empirical probabilities associated with $\hat{\beta}$. They are defined by

$$\hat{\pi}_i = s_1\left(\hat{\lambda}' \hat{g}_i\right) \Big/ \sum_{j=1}^{n} s_1\left(\hat{\lambda}' \hat{g}_j\right) = \frac{1 + \hat{v}_i}{\sum_j \left(1 + \hat{v}_j\right)} \quad (i = 1, \cdots, n). \tag{3.9}$$

where $\hat{\lambda} = \arg\max_{\lambda \in \hat{\Lambda}\left(\hat{\beta}\right)} P\left(\hat{\beta}, \lambda\right)$, $\hat{g}_i = g_i\left(\hat{\beta}\right)$, and $\hat{v}_i = \hat{\lambda}' \hat{g}_i$. These empirical probabilities sum to one by construction and satisfy the sample moment condition $\sum_{i=1}^{n} \hat{\pi}_i \hat{g}_i = 0$ when the first order conditions for $\hat{\lambda}$ hold. For any function $a(w, \beta)$, these prob-

abilities can be used to form an efficient estimator $\sum_{i=1}^{n} \hat{\pi}_i a\left(w_i, \hat{\beta}\right)$ of $E\left[a\left(w, \beta_0\right)\right]$, as in Newey and Smith (2004). The following result gives first-order conditions for RRCUE.

Theorem 3.2. *If Assumption 3.1 (a) is satisfied, then the RRCUE first order conditions imply*

$$\left[\sum_{i=1}^{n} \hat{\pi}_i G_i\left(\hat{\beta}\right)\right]' \left[\hat{\Omega}\left(\hat{\beta}\right) + \alpha I\right]^{-1} \hat{g}\left(\hat{\beta}\right) = 0, \tag{3.10}$$

where $G_i(\beta) = \partial g_i(\beta)/\partial \beta'$.

3.4. Asymptotic properties of the RRCUE

In this section, we establish first-order asymptotic properties of the regularized estimator. We show that RRCUE is consistent and asymptotically normal, and achieves the Chamberlain (1987)'s semiparametric efficiency bound under some standard assumptions. We first give some regularity conditions for the consistency of RRCUE.

Assumption 3.2. *The data are i.i.d. and (a)* β_0 *is unique value of* β *in* $\mathscr{B}$ *satisfying* $E[\rho(w,\beta)|z] = 0$*; (b)* $\mathscr{B}$ *is compact; (c)* $E\left[\sup_{\beta\in\mathscr{B}} |\rho(w,\beta)|^2 \,|z\right]$ *is bounded and there is* $\delta_1(w)$ *and* $r > 0$ *such that for all* $\tilde{\beta}, \beta \in \mathscr{B}$, $|\rho(w,\tilde{\beta}) - \rho(w,\beta)| \leqslant \delta_1(w)\|\tilde{\beta} - \beta\|^r$ *and* $E\left[\delta_1(w)^2\right] < \infty$*; (d) there are* $\delta_2(w)$ *and a neighborhood* $\mathscr{N}$ *of* β_0 *such that* $E\left[\sup_{\beta\in\mathscr{N}} |\rho(w,\beta)|^4 \,|z\right]$ *is bounded and for all* $\beta \in \mathscr{N}$ $|\rho(w;\beta) - \rho\left(w,\beta_0\right)| \leq \delta_2(w)\|\beta - \beta_0\|$ *and* $E\left[\delta_2(w)^2|z\right]$ *is bounded; (e)* $\sigma(z)^2 := E\left[\rho\left(w,\beta_0\right)^2 |z\right]$ *is bounded away from zero; (f) there is* $\gamma > 2$ *with* $E\left[\sup_{\beta\in\mathscr{B}} |\rho(w,\beta)|^\gamma\right] < \infty$.

Assumption 3.2(a) is the minimal identification condition that β_0 is the unique value where the conditional moment restriction is satisfied. The stronger condition that there is a known K such that the unconditional moment restrictions $E\left[q^K(z)\rho(w,\beta)\right] = 0$ serves to identify β_0 is not required. As K grows with n, the weak condition in Assumption 3.2(a) is sufficient to identify β_0 as justified in Lemma 2.1 of DIN03. Assumption 3.2(b) is the usual compacity assumption. Assumption 3.2(c) imposes

a bounded second conditional moment and Lipschitz condition, which is useful to apply the uniform convergence result of Newey (1991). Assumption 3.2(d) plays an important role in obtaining a convergence rate for the sample second moment matrix $\hat{\Omega}\left(\hat{\beta}\right)$. Assumption 3.2(d) allows for the error term $\rho\left(w,\beta_0\right)$ to be conditionally heteroskedastic, so that $E\left[\rho\left(w,\beta_0\right)^2 \mid z\right]$ depends on z. Assumption 3.2(f) requires the existence of slightly higher moments than consistency for GMM, as in Hansen (1982).

The following theorem establishes the consistency of the RRCUE.

Theorem 3.3. *Let $\varepsilon > 0$ be such that $1/2 - 1/\gamma - \varepsilon > 0$, where $\gamma > 2$ is defined in Assumption 3.2(f). If Assumptions 3.1 and 3.2 are satisfied, $K \to \infty$, $\alpha \to 0$ and $\alpha n^{1/2-1/\gamma-\varepsilon} \to \infty$, then $\hat{\beta} \xrightarrow{p} \beta_0$.*

The restriction, $\alpha n^{1/2-1/\gamma-\varepsilon} \to \infty$, on the rate of convergence of the regularization parameter α is the counterpart of the restriction on the growth rate of K imposed by DIN03 to obtain consistency of the standard CUE, that is $\zeta(K)^2 K/n^{1-2/\gamma} \to 0$. This restriction on the rate of convergence of α is weaker when $\rho(z,\beta)$ has moments of higher orders. It implies in particular that α goes to zero slower than $1/\sqrt{n}$.

We need some additional conditions for asymptotic normality. Let $\rho_\beta(w,\beta) = \partial\rho(w,\beta)/\partial\beta'$, $D(z) = E\left[\rho_\beta\left(w,\beta_0\right)|z\right]$, and $\rho_{\beta\beta}(w,\beta) = \partial^2\rho(w,\beta)/\partial\beta\partial\beta'$.

Assumption 3.3. *(a) $\beta_0 \in int(\mathscr{B})$; (b) $\rho(w,\beta)$ is twice continuously differentiable in a neighborhood $\mathscr{N}$ of β_0; (c) $E\left[\sup_{\beta\in\mathscr{N}}\left\|\rho_\beta(w,\beta)\right\|^2|z\right]$ and $E\left[\left\|\rho_{\beta\beta}\left(w,\beta_0\right)\right\||z\right]$ are bounded; (d) $E\left[D(z)'D(z)\right]$ is nonsingular.*

These assumptions are quite standard regularity conditions used by DIN03. Parts (b) and (c) are standard smoothness conditions. Part (d) is the local identification condition that is essential for asymptotic normality.

We need to introduce some notions before stating the asymptotic normality result. Let $\hat{g}_i = g_i(\hat{\beta})$, $\widehat{G}_i = G_i(\hat{\beta})$, $\widehat{G} = \sum_{i=1}^n \hat{\pi}_i \widehat{G}_i$, $\widehat{\Omega} = \sum_{i=1}^n \hat{g}_i\hat{g}_i'/n$, and $\widehat{V} = \left(\hat{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\hat{G}\right)^{-1}$.

Theorem 3.4. *If Assumptions 3.1, 3.2 and 3.3 are satisfied, $K \to \infty$, $\alpha \to 0$, $\alpha^{3/2} n^{1/2-1/\gamma-\varepsilon} \to \infty$, and $\alpha^{5/2} n^{1/2} \to \infty$ then*

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} \mathcal{N}(0, V), \quad \hat{V} \xrightarrow{p} V, \quad V = \left(E\left[D(z)'\sigma(z)^{-2}D(z)\right]\right)^{-1}.$$

This result gives the restrictions on the convergence rate of α for the regularized CUE estimator to reach the semiparametric efficiency bound of Chamberlain (1987). These restrictions imply, in particular, that α goes to zero slower than $1/n^{1/5}$. This condition is the counterpart of the restriction on the growth rate of K imposed by DIN03 to obtain asymptotic efficiency of the standard CUE, that is, $\zeta(K)^2 K^2/n \to 0$. The main advantage is that our rate no longer depends on K, so more instruments can be used to gain efficiency in finite samples. Although the standard CUE and its regularized version are both asymptotically efficient under certain conditions, their small-sample properties can differ substantially, as shown in Monte Carlo simulations. Theorem 3.4 also provides an efficient estimator of the asymptotic variance of the regularized CUE. To the best of our knowledge, these asymptotic results are new and extend DIN03 to the asymptotic framework where no restriction is imposed on the growth rate of K relative to n.

Alternative variance estimator: more robust to many moments

Newey and Windmeijer (2009) argued that in the presence of many moments (potentially weak), the standard textbook variance estimator $\widehat{V}_0/n$, where $\widehat{V}_0 = \left(\widehat{G}'\widehat{\Omega}^{-1}\widehat{G}\right)^{-1}$, does not provide a good approximation to the finite sample distribution of the standard CUE estimator. They suggest instead approximating the finite sample variance of $\hat{\beta}_{CUE}$ by $\widetilde{V}_0/n$, where $\widetilde{V}_0 = \hat{H}^{-1}\hat{D}'\hat{\Omega}^{-1}\hat{D}\hat{H}^{-1}$. They argued that, in the presence of many moment conditions, the former underestimates the finite sample variance of $\hat{\beta}_{CUE}$, the latter then adjusts for the presence of many moments. In the same spirit, we suggest using the following estimator of the asymptotic variance of the regularized CUE estimator,

$$\tilde{V} = \hat{H}^{-1}\hat{D}'(\hat{\Omega} + \alpha I)^{-1}\hat{D}\hat{H}^{-1}, \tag{3.11}$$

where $\hat{H} = \left.\frac{\partial^2 \hat{Q}(\beta)}{\partial\beta\partial\beta'}\right|_{\beta=\hat{\beta}}$, $\hat{D} = \hat{D}(\hat{\beta})$, $\hat{\Omega} = \hat{\Omega}(\hat{\beta})$, with

$$\hat{Q}(\beta) = \hat{g}(\beta)'(\hat{\Omega}(\beta)+\alpha I)^{-1}\hat{g}(\beta)/2, \quad \hat{D}(\beta) = \sum_i \hat{\pi}_i(\beta)G_i(\beta) \quad \hat{\Omega}(\beta) = \frac{1}{n}\sum_i g_i(\beta)g_i(\beta)',$$

and $\hat{\pi}_i(\beta) = \dfrac{1 - \hat{g}(\beta)'(\hat{\Omega}(\beta) + \alpha I)^{-1}g_i(\beta)}{\sum_j \left(1 - \hat{g}(\beta)'(\hat{\Omega}(\beta) + \alpha I)^{-1}g_j(\beta)\right)}$.

We will not investigate consistency of $\tilde{V}$ in this paper, but we believe that strategies used to obtain consistency of $\hat{V}$ (see the proof of Theorem 3.4 in the Appendix) can be used to show consistency of $\tilde{V}$ under certain restrictions on the convergence rate of the regularization parameter. We see in simulation that a Wald test of $H_0 : \beta = \beta_0$ based on $\tilde{V}$ performs better (in terms of controlling size) in the presence of many moment conditions compared to the test constructed from $\hat{V}$, in almost all simulation frameworks we considered.

In the next section, we propose a data-driven method for choosing the regularization parameter in practice.

3.5. Data-driven selection of the regularization parameter

This section is devoted to the selection of the optimal regularization parameter. We propose to employ cross-validation to compute the distance of sample moments from zero, which is then used as a criterion to select the regularization parameter. This approach aims to choose, from the family of RRCUEs indexed by α, the estimator that best satisfies the sample counterpart of the moment condition (3.2). Indeed, if $\hat{\beta}$ is a '*good*' estimator of β_0, the sample moment function $\sum_{i=1}^n g(x_i, \beta_0)/n$ would be '*close*' to zero, in the sense of a certain norm, if β_0 were replaced by $\hat{\beta}$. Instead

of using the simple l_2 norm to measure the distance of the sample moment function from zero, we propose an alternative distance[7]. We use L-fold cross-validation to construct a '*suitable*' distance of the sample mean $\sum_{i=1}^{n} g\left(x_i, \hat{\beta}\right)/n$ from zero. Using this distance as criteria allows us to choose α such that the corresponding estimator best satisfies the moment condition of Eq.(3.2) even out-of-sample.

To elaborate, let $\{J_l, l = 1, \cdots, L\}$ denote a partition of the set of data indices $[n] := \{1, 2, \cdots, n\}$. For each $l = 1, \ldots, L$, let J_{-l} denote the set of all indices in $[n]$ except those in J_l, and let n_l denote the cardinality of J_l. Let $\hat{\beta}_{-l}$ denote the version of the regularized CUE estimator obtained using the part of the sample indexed by J_{-l}. The following algorithm describes the general procedure for choosing the optimal α using L-fold cross-validation.

Algorithm (L-fold CV approach for selecting the optimal α).

1. Consider a grid Δ_K of values of α.
2. For each $\alpha \in \Delta_K$ and for each $l = 1, \cdots, L$, compute $\hat{\beta}_{-l}^{\alpha}$ (resp. $\hat{\beta}_{l}^{\alpha}$) , the version of the RRCUE obtained using the part of the sample indexed by J_{-l} (resp. J_l). Let $\widetilde{\Omega}_l = diag\left(\widehat{\Omega}(\hat{\beta}_{l}^{\alpha})\right)$.
3. For each $\alpha \in \Delta_K$ and for each $l = 1, \cdots, L$, compute the distance $\mathscr{I}_{nl}(\alpha)$ (it measures how well $\hat{\beta}_{-l}^{\alpha}$ satisfies the moment condition $E\left[g(x_i, \beta_0)\right] = 0, i \in J_l$), defined by

$$\mathscr{I}_{nl}(\alpha) = \left(\frac{1}{n_l}\sum_{i\in J_l} g\left(x_i, \hat{\beta}_{-l}^{\alpha}\right)\right)' \widetilde{\Omega}_l^{-1} \left(\frac{1}{n_l}\sum_{i\in J_l} g\left(x_i, \hat{\beta}_{-l}^{\alpha}\right)\right) \tag{3.12}$$

4. The optimal α is obtained as

$$\hat{\alpha} = \underset{\alpha\in\Delta_K}{\arg\min}\left(\mathscr{I}_n(\alpha) := \sum_{l=1}^{L} \mathscr{I}_{nl}(\alpha)\right) \tag{3.13}$$

7. As the entire sample was used to obtain $\hat{\beta}$ by minimizing a quadratic function of the sample moment function $\sum_{i=1}^{n} g\left(x_i, \beta_0\right)/n$, the simple l_2 norm of $\sum_{i=1}^{n} g\left(x_i, \hat{\beta}\right)/n$ may provide a misleading measure of how well $\hat{\beta}$ satisfies the sample counterpart of the moment condition (3.2).

Remark 3.1.

— The dependence of the grid Δ_K in K is motivated by the fact that the regularization of the covariance matrix of $\hat{\Omega}(\beta)$ is preceded by the normalization of q_i by the upper bound, $\zeta(K)$, of the sup-norm of the approximating functions. Our simulation exercise suggests choosing the grid Δ_K to be inversely proportional to $\sqrt{K}$, so that the grid shrinks to zero as K increases. This consideration seems counterintuitive, as one might think that the regularization parameter would be higher for larger values of K. However, this intuition could be misleading here because the normalization performed prior to regularization tends to considerably reduce the eigenvalues of the matrix $\hat{\Omega}(\beta)$. As a consequence, large values of α will introduce substantial regularization bias. With this normalization, one expects the optimal choice of α to be a decreasing function of K. This is why we anticipated this by choosing a grid that narrows towards zero as K increases. Even when considering a grid that does not depend on K, we observed in simulations that the optimal choice of α, according to our procedure, decreases with K.

— The number of folds L has to be chosen. We see in simulations that our result is not very sensitive to a small number of folds (ranging from 2 to 10). In the empirical application, we choose $L = 5$.

3.6. Monte Carlo study

This section aims to examine the small sample properties of our RRCUE estimator in order to evaluate the gain of regularization. The baseline setup for our simulation is given by the following system,

$$\begin{cases} y_i = h(x_i, \beta_0) + e_i \\ x_i = f(z_i) + u_i \end{cases} \quad \text{with } \beta_0 = 0.1, \tag{3.14}$$

where the first equation is the main structural model and the second is the reduced-form equation for the right-hand side endogenous regressor. Both y_i and x_i are univariate, but the underlying vector of instruments z_i is potentially high-dimensional. Our parameter of interest is β_0. We generate data in a way that $E[e_i|z_i] = 0$, given Eq. (3.1) with $\rho(w_i, \beta_0) = y_i - h(x_i, \beta_0)$ and $w_i = (x_i, y_i)'$.

We consider different experiments. Each framework is designed to mimic a specific situation that can arise in application. The data-generating processes differ: (i) by the type of specification of the main equation (the function h is either linear or nonlinear with respect to β); (ii) by the nature of the structural disturbance (homoskedastic or heteroskedastic); (iii) by the number of relevant instruments that enter the first-stage equation (small number or large number); (iv) or by the way the explanatory power, captured by the concentration parameter, is distributed among instruments.

In experiments 1 and 2 below, we assume that the reduced-form error term $u_i \overset{i.i.d.}{\sim} N(0, \sigma_u^2)$ with $\sigma_u^2 = 1$. Following Hausman et al. (2012), we suppose that the structural disturbance e_i, which is allowed to be heteroskedastic, is given by

$$e_i = \rho u_i + \sqrt{\frac{1-\rho^2}{\phi^2 + \psi^4}}(\phi v_{1i} + \psi v_{2i}),$$

where $\rho = 0.3$, $\psi = 0.86$ and conditional on z_{i1} (where z_{ij} is the jth component of z_i), $v_{1i} \overset{i.i.d.}{\sim} N\left(0, z_{i1}^2\right)$ and $v_{2i} \overset{i.i.d.}{\sim} N(0, \psi^2)$ are independent of u_i. $\phi = 0$ or 1.38072 is chosen so that the R-squared for the regression of e^2 on the instruments [8], $\mathscr{R}^2_{e^2|z}$, is 0 or 0.2, corresponding to homoskedastic and heterosckedastic cases respectively.

We compare the performance of RRCUE to that of some state-of-the-art estimators in the literature, including the standard two-step GMM of Hansen (1982) and the standard CUE of Hansen et al. (1996). In the case of a linear homoskedastic structural equation, we consider four additional alternative estimators: 2SLS, LIML, the Tikhonov regularized 2SLS estimator (T2SLS) of Carrasco (2012), and the Tikhonov

8. $R^2_{e^2|z} = \text{var}\left\{\text{E}\left(e^2 \mid z\right)\right\} / \left[\text{var}\left\{\text{E}\left(e^2|z\right)\right\} + \text{E}\left\{\text{var}\left(e^2 \mid z\right)\right\}\right]$.

regularized LIML estimator (TLIML) of Carrasco and Tchuente (2015). In the case of linear heteroskedastic structural equation, we consider two more alternative estimators: Hausman et al. (2012)'s heteroskedasticity-robust version of Fuller (1977)'s estimator (HFUL) and heteroskedasticity-robust LIML (HLIM).

We set $n = 500$ observations, and the number of instrumental variables K is chosen from the set $\{2, 30, 50, 100\}$ for experiment 1 and from the set $\{15, 30, 50, 100\}$ for experiment 2. As our vector of instruments is almost a spline basis, we choose $\zeta(K) = \sqrt{K}$. With two specifications for the structural equation (linear or nonlinear), two specifications for the corresponding error term (homoskedastic or heteroskedastic), and four different choices for the number of instruments, there are a total of 16 specifications for experiment 1. Similarly, with two different sets of first-stage coefficients and four different choices for the number of instruments, there are a total of 8 specifications for experiment 2. Therefore, the two experiments total 24 specifications.

For each specification, we performed 10,000 Monte Carlo simulations. For each draw, we compute the optimal RRCUE estimator and the alternative comparison estimators. The computation of the optimal RRCUE requires a suitable choice of the grid Δ_K of values for α. We search for the optimal α in the following 100-point grid,

$$\Delta_K = \left\{ \frac{1}{\sqrt{K}} \times \left(0.0001 + (i-1) \times \frac{0.0499}{99} \right) \middle| \, i = 1, 2, \ldots, 100 \right\},$$

which is the set of 100 points uniformly distributed between 0.0001 and 0.05, where each element is multiplied by $1/\sqrt{K}$. This normalization allows the grid to approach zero as K increases.

After 10,000 replications, we calculate several performance measures for each estimator. We consider two measures of bias: the mean bias (Mean.bias) and the median bias (Med.bias); four measures of dispersion: the variance of estimates (Var), the median absolute deviation (MAD, defined as the median of the absolute value of the difference between simulated estimates and the median simulation estimate),

and the nine decile range (Ndr$(0.95-0.05)$, defined as the range between the 0.05 and 0.95 quantiles of the distribution of simulated estimates). We also compute the nominal 5% rejection frequency (0.05 rej.Freq) for the Wald test of $H_0 : \beta = \beta_0$. To compute the Wald statistic for CUE, we rely on the many-instruments robust standard error of Newey and Windmeijer (2009), and for RRCUE, we use its regularized counterpart presented in section 3.4.

3.6.1. Experiment 1: Small number of relevant instruments

We first consider a setup with only one relevant instrument. In particular, we suppose that $z_i \overset{i.i.d.}{\sim} N(0,1)$ and $f(z_i) = \pi z_i$, where π is a scalar chosen so that the concentration parameter $n\pi^2 = \mu^2 = 32$. Also, we consider the following set of instruments in the spirit of Hausman et al. (2012),

$$q^K(z_i) = \begin{cases} (1, z_i) & \text{if} \quad K = 2 \\ \left(1, z_i, z_i^2, z_i^3, z_i^4, z_i D_{i1}, \cdots, z_i D_{i,K-5}\right)' & \text{if} \quad K \in \{30, 50, 100\} \end{cases}$$

where $D_{ik} \in \{0,1\}, \Pr(D_{ik} = 1) = 1/2$. This instrument set consists of powers of z up to fourth power plus interactions with dummy variables. Note that only z affects the reduced-form. Since the exact specification of the reduced-form is unknown in practice, this framework will also help evaluate the effect of including more power series than necessary.

In this experiment, we consider two types of specifications for the structural equation:

(i) **Model 1a.** linear specification, $h(x_i, \beta) = \beta x_i$,

(ii) **Model 1b.** nonlinear specification, $h(x_i, \beta) = \exp(\beta x_i)$.

For each of these frameworks, we consider cases where the structural disturbance is either homoskedastic ($\phi = 0$) or heteroskedastic ($\phi = 1.38072$.)

Table 3.1 presents results for experiment 1 with linear structural equation (Model

1a). It offers a detailed comparison of the RRCUE estimator's performance against several competitors, including CUE, GMM, T2SLS, TLIML, HFUL, and HLIM, under both homoskedastic and heteroskedastic disturbances across varying numbers of instruments (K). The focus will be on bias (Mean.bias, Med.bias) and dispersion (Var, MAD, Ndr($0.95-0.05$)), particularly highlighting RRCUE's performance relative to CUE, GMM, and other estimators.

Panel A - Homoskedasticity: For small K ($K = 2$, only instruments that enter the reduced-form equation are used), regularization is not needed. In fact, the median bias for RRCUE is close to that of CUE and slightly better than GMM. In terms of dispersion, RRCUE and CUE report almost the same variance and MAD, while GMM shows slightly lower variance, though this comes at the cost of higher median bias. As K increases ($K = 30, 50, 100$), RRCUE outperforms GMM in terms of median bias, although regularization introduces a certain amount of bias compared to CUE. RRCUE remains comparable to competitors like T2SLS and TLIML in terms of median bias. In terms of dispersion, RRCUE exhibits substantially lower variance compared to CUE. This pattern persists as K increases, with RRCUE maintaining a much lower variance, MAD, and nine-decile range than CUE. Compared to T2SLS and TLIML, RRCUE consistently shows comparable performance in terms of bias and dispersion as the number of instruments grows.

Panel B - Heteroskedasticity: Under heteroskedastic disturbances, RRCUE continues to demonstrate strong performance relative to CUE, GMM, HFUL, and HLIM, with some exceptions. For $K = 2$, RRCUE's median bias is comparable to that of CUE, GMM, HFUL, and HLIM. Similarly, RRCUE is almost equivalent to standard competitors in terms of dispersion, denoting again the fact that regularization is not needed for small K. As K increases, the effect of regularization is more apparent. RRCUE keeps dispersion under control while maintaining a reasonable level of bias that remains smaller than the GMM over-identification bias. Against HFUL and HLIM, RRCUE shows clear advantages in terms of dispersion. For example, at $K = 30, 50, 100$, RRCUE substantially outperforms HFUL and HLIM in terms of dis-

persion (variance, MAD, and nine-decile range), while keeping the regularization bias at a manageable level.

Moreover, RRCUE's rejection frequency remains close to the 5% nominal level in both homoskedastic and heteroskedastic settings, although there is a slight under-rejection, indicating accurate hypothesis testing. In contrast, the rejection frequencies of CUE and GMM deteriorate as K increases, while HFUL and HLIM also exhibit slight deviations from the nominal level.

Table 3.2 presents results for experiment 1 with nonlinear structural equation (Model 1b). The results highlight that RRCUE generally performs well compared to CUE and GMM in both homoskedastic and heteroskedastic settings. In terms of bias, RRCUE shows low median bias across different numbers of moments (K), particularly it maintains comparable or slightly better performance than GMM. Regarding dispersion, RRCUE exhibits smaller variance than CUE, especially as K increases, while maintaining lower MAD compared to CUE, indicating tighter concentration around the median. Notably, RRCUE's Ndr (0.95-0.05) remains relatively stable across all K values, outperforming CUE and GMM in controlling over-dispersion for larger sets of moments ($K = 30, 50, 100$), particularly under heteroskedastic disturbances. Finally, in terms of rejection frequency at the 5% nominal level, RRCUE maintains competitive performance, consistently rejecting less often than GMM and CUE, particularly when K is large, which suggests better size control in finite samples.

Overall, RRCUE demonstrates a balance between bias and dispersion across different numbers of instruments and disturbance structures. Results in Tables 3.1 & 3.2 reveal that even if a small number of instruments enter the reduced-form equation, using power series and splines as additional instruments, together with regularization, can help improve the efficiency of the CUE while maintaining the regularization bias at a manageable level if the regularization parameter is chosen in a suitable manner.

Table 3.1 – Simulation results: Experiment 1 - Small number of relevant instruments and linear model (Model 1a)

		Panel A: Homoskedasticity							Panel B: Heteroskedasticity				
	Estimator	**RRCUE**	**CUE**	**GMM**	**2SLS**	**T2SLS**	**LIML**	**TLIML**	**RRCUE**	**CUE**	**GMM**	**HFUL**	**HLIM**
K=2	Mean.bias	-0.010	-0.010	0.000	0.000	0.000	-0.010	-0.010	-0.008	-0.009	0.010	0.013	0.003
	Med.bias	0.002	0.002	0.010	0.010	0.010	0.001	0.001	-0.001	-0.001	0.019	0.021	0.013
	Var	0.029	0.029	0.027	0.027	0.027	0.029	0.029	0.072	0.072	0.062	0.057	0.062
	MAD	0.106	0.106	0.103	0.102	0.102	0.106	0.106	0.170	0.171	0.158	0.153	0.158
	Ndr (0.95-0.05)	0.553	0.556	0.528	0.530	0.529	0.553	0.552	0.870	0.877	0.813	0.779	0.810
	0.05 rej.Freq	0.040	0.042	0.047	0.045	0.045	0.041	0.040	0.045	0.046	0.049	0.049	0.046
K=30	Mean.bias	0.048	1.211e+09	0.120	0.143	0.046	-0.014	-0.007	0.046	6.147e+08	0.122	-0.014	-0.048
	Med.bias	0.054	0.019	0.122	0.147	0.050	0.004	0.004	0.055	0.014	0.123	0.022	0.012
	Var	0.034	6.063e+23	0.017	0.012	0.021	0.838	0.033	0.061	2.078e+23	0.037	0.197	6.672
	MAD	0.112	0.186	0.084	0.074	0.095	0.147	0.112	0.148	0.226	0.121	0.205	0.215
	Ndr (0.95-0.05)	0.575	1.054	0.419	0.361	0.472	0.839	0.579	0.788	1.343	0.613	1.307	1.510
	0.05 rej.Freq	0.054	0.098	0.327	0.281	0.096	0.035	0.041	0.041	0.095	0.285	0.060	0.058
K=50	Mean.bias	0.074	2.870e+09	0.136	0.182	0.061	-5.927e+07	-0.009	0.078	2.651e+10	0.143	-0.005	-3.248
	Med.bias	0.080	0.044	0.136	0.182	0.064	0.003	0.005	0.087	0.043	0.142	0.023	0.013
	Var	0.031	3.464e+23	0.014	0.009	0.018	3.693e+20	0.033	0.051	1.421e+25	0.034	0.270	6.854e+04
	MAD	0.101	0.219	0.077	0.064	0.087	0.173	0.112	0.133	0.266	0.116	0.241	0.255
	Ndr (0.95-0.05)	0.525	1.315	0.387	0.311	0.443	1.048	0.582	0.708	1.664	0.593	1.593	2.047
	0.05 rej.Freq	0.052	0.145	0.463	0.495	0.106	0.037	0.038	0.041	0.145	0.424	0.065	0.062
K=100	Mean.bias	0.102	-1.806e+09	0.132	0.226	0.096	-2.19e+09	-0.013	0.114	1.016e+09	0.143	0.021	0.019
	Med.bias	0.110	0.088	0.131	0.227	0.096	0.008	0.002	0.123	0.093	0.139	0.040	0.027
	Var	0.030	1.919e+22	0.012	0.005	0.013	1.451e+23	0.057	0.055	1.054e+23	0.033	0.420	107.265
	MAD	0.093	0.239	0.072	0.050	0.075	0.225	0.115	0.120	0.282	0.114	0.303	0.329
	Ndr (0.95-0.05)	0.498	1.543	0.356	0.241	0.377	1.694	0.601	0.677	1.821	0.589	2.159	3.250
	0.05 rej.Freq	0.061	0.235	0.620	0.864	0.154	0.038	0.038	0.064	0.268	0.604	0.074	0.070

Note: Simulation results based on 10,000 replications with a sample size of $n = 500$. We report six (06) measures of performance: the mean bias (Mean.bias), the median bias (Med.bias), the variance of estimates (Var), the median absolute deviation (MAD), the nine decile range (Ndr$(0.95-0.05)$), and the nominal 5% rejection frequency (0.05 rej.Freq) for the Wald test of $H_0 : \beta = \beta_0$.

Table 3.2 – Simulation results: Experiment 1 - Small number of relevant instruments and nonlinear structural equation (Model 1b)

		Panel A: Homoskedasticity			**Panel B: Heteroskedasticity**		
		RRCUE	CUE	GMM	RRCUE	CUE	GMM
K=2	Mean.bias	-0.020	-0.020	-0.014	-0.038	-0.038	-0.033
	Med.bias	0.001	0.000	0.007	-0.011	-0.011	-0.004
	Var	0.022	0.022	0.021	0.033	0.033	0.032
	MAD	0.094	0.094	0.091	0.122	0.122	0.117
	Ndr (0.95-0.05)	0.485	0.486	0.476	0.587	0.587	0.585
	0.05 rej.Freq	0.128	0.132	0.119	0.205	0.209	0.176
K=30	Mean.bias	0.019	-0.026	0.086	-0.006	-0.041	0.064
	Med.bias	0.044	0.011	0.095	0.033	0.003	0.082
	Var	0.023	0.041	0.010	0.028	0.037	0.016
	MAD	0.092	0.138	0.062	0.101	0.131	0.071
	Ndr (0.95-0.05)	0.494	0.644	0.318	0.543	0.606	0.410
	0.05 rej.Freq	0.114	0.177	0.329	0.134	0.179	0.284
K=50	Mean.bias	0.045	-0.020	0.103	0.020	-0.045	0.091
	Med.bias	0.066	0.030	0.110	0.056	0.011	0.103
	Var	0.019	0.054	0.008	0.025	0.046	0.015
	MAD	0.081	0.153	0.057	0.090	0.144	0.071
	Ndr (0.95-0.05)	0.455	0.737	0.293	0.523	0.677	0.391
	0.05 rej.Freq	0.107	0.224	0.455	0.115	0.196	0.407
K=100	Mean.bias	0.072	-0.003	0.106	0.053	-0.043	0.108
	Med.bias	0.090	0.065	0.111	0.085	0.034	0.113
	Var	0.017	0.078	0.008	0.024	0.073	0.018
	MAD	0.071	0.163	0.057	0.076	0.157	0.080
	Ndr (0.95-0.05)	0.415	0.873	0.284	0.493	0.814	0.437
	0.05 rej.Freq	0.127	0.310	0.614	0.120	0.268	0.597

Note: Simulation results based on $10,000$ replications with a sample size of $n = 500$. We report six (06) measures of performance: the mean bias (Mean.bias), the median bias (Med.bias), the variance of estimates (Var), the median absolute deviation (MAD), the nine decile range (Ndr(0.95—0.05)), and the nominal 5% rejection frequency (0.05 rej.Freq) for the Wald test of $H_0 : \beta = \beta_0$.

3.6.2. Experiment 2: Large number of relevant instruments

Our second experiment design involves a large number of relevant instruments in a framework where both structural and reduced-form models are linear. In particular, we assume that $h(x_i, \beta) = \beta x_i$ and $f(z_i) = \pi' z_i$, where π is a high-dimensional vector of first-stage coefficients that satisfies $n\sigma_u^{-2}\pi'\Sigma_z\pi = \mu^2$, with the concentration parameter μ^2 measuring the strength of the instruments and $\Sigma_z = E\left[z_i z_i'\right]$. Following Hansen and Kozbur (2014) we consider Gaussian instruments that are correlated with one another. Under this Gaussian instrument design, all instruments are drawn with mean 0 and variance $\text{var}\left(z_{ij}\right) = \Sigma_{zjj} = 0.3$. Dependence between instruments is given by the Pearson correlation coefficient $\text{corr}\left(z_{ij}, z_{ik}\right) = 0.5^{|j-k|}$. As z_i is already a large vector, we consider it as the instrument set without adding power series; that is, $q^K(z_i) = z_i$. In this experiment, we focus on heteroskedastic structural disturbance and consider two different sets of first-stage coefficients in the spirit of Donald and Newey (2001) and Carrasco (2012):

(i) **Model 2a.** $\pi_l = d\left(1 - \dfrac{l}{K+1}\right)^4$, for $l = 1, \ldots, K$, where the constant $d := \sqrt{\dfrac{\sigma_u^2\mu^2}{n\tilde{\pi}'\Sigma_z\tilde{\pi}}}$, with $\tilde{\pi} = \pi/d$, is chosen so that $\mu^2 = 32$. This specification is relevant for applications in which some instruments are more relevant than others.

(ii) **Model 2b.** $\pi_l = d$, for $l = 1, \ldots, K$, where the constant $d := \sqrt{\dfrac{\sigma_u^2\mu^2}{n\iota_K'\Sigma_z\iota_K}}$, with ι_K a $K \times 1$ vector of ones, is chosen so that $\mu^2 = 32$. This framework is relevant for applications where the instruments are equally important.

Table 3.3 presents results for the case where a large number of instruments enter the reduced-form equations. In Panel A, which considers the case of instruments ranked in decreasing order (Model 2a), the performance of RRCUE in terms of dispersion is generally superior to that of its competitors, particularly CUE and GMM. With a median absolute deviation (MAD) of 0.103 and a variance of 0.027 at $K = 15$, RRCUE demonstrates lower dispersion compared to CUE, which exhibits erratic be-

havior with a MAD of 0.126 and an excessively high variance. Moreover, RRCUE maintains a favorable Ndr (0.95-0.05) of 0.529, indicating a balanced performance across different deciles. In terms of rejection frequency, RRCUE exhibits a consistent and controlled rate, significantly lower than that of GMM and closer to the nominal 5% level, further underscoring its robustness in maintaining type I error rates. In comparison to HFUL and HLIM, RRCUE exhibits comparable MAD and variance, indicating its competitiveness in terms of dispersion. The same pattern is observed when the number of instruments increases. We have similar results in Panel B, which evaluates the scenario with equally important instruments (Model 2b). Overall, RRCUE consistently demonstrates superior performance in terms of bias (compared to GMM) and dispersion (compared to CUE, HFUL, and HLIM). Moreover, the role of regularization seems to be much more important when there is a large number of relevant instruments that enter the reduced-form equation.

In summary, regularization allows solving the moment problem of CUE by reducing its dispersion. RRCUE can take advantage of a bunch of moments/instruments and gain efficiency while maintaining the regularization bias at a relatively low and reasonable level.

3.7. Empirical application: Institutions and growth

This section revisits the empirical work of Hall and Jones (1999), aiming to answer the famous question: *Why do some countries produce so much more output per worker than others?*. This question is primarily motivated by the simple fact that output per worker varies enormously across countries. Hall and Jones (1999) argue that the differences in capital accumulation, productivity, and therefore output per worker are driven by differences in institutions and government policies, which they call social infrastructure.

To quantify the effect of social infrastructure on per capita income, they treat social infrastructure as endogenous. Identification is then based on the idea that

social infrastructure is determined historically by location and other factors captured in part by language, all of which are exogenous. More precisely, they use 2SLS with four instruments for social infrastructure: the fraction of population speaking English at birth (EnL), the fraction of population speaking one of the five major European languages at birth (EuL), the distance from the equator (latitude, Lt), and Frankel and Romer (1999) geography-predicted trade intensity (FR). The linear IV regression model is given by

$$y_i = c + \delta S_i + \varepsilon_i \quad i = 1, 2 \ldots, n = 79,$$

where y_i is country i's log income per capita, S_i is country i's proxy for social infrastructure, c is a constant, and δ is the scalar parameter of interest.

The baseline $n \times 4$ matrix of instruments is given by $z = [EnL, EuL, Lt, FR]$. Recently, Dmitriev (2013) pointed out that these instruments are weak. To boost their identification strength, we increase the set of instruments from 4 to 18, as suggested by Carrasco and Tchuente (2016) [9].

Our enriched set of instruments is given by [10]

$$\begin{aligned} q(z) = \big[& z, z.^2, z.^3, z(:,1) \times z(:,2), z(:,1) \times z(:,3), z(:,1) \times z(:,4), \\ & z(:,2) \times z(:,3), z(:,2) \times z(:,4), z(:,3) \times z(:,4) \big], \end{aligned}$$

where all instruments are divided by their standard deviation prior to regularization. Our sample consists of $n = 79$ countries for which no data were imputed [11]. Results are collected in Table 3.4.

Table 3.4 presents estimates of the effect of social infrastructure on growth using various estimators for two different numbers of instruments: $K = 4$ (benchmark) and $K = 18$ (many instruments). Across the estimators, RRCUE demonstrates a notable

9. Carrasco and Tchuente (2016) argued that the use of many instruments allows the concentration parameter to increase from $\hat{\mu}_n^2 = 28.6$ for 4 instruments to $\hat{\mu}_n^2 = 51.48$ for 18 instruments, resulting in a moderately strong set of instruments.

10. $z.^k = \left[z_{ij}^k \right], z(:,j)$ is the j^{th} column of z, and $z(:,j) \times z(:,l)$ is a vector of interactions between columns j and l.

11. Data used are collected from https://web.stanford.edu/~chadj/HallJones400.asc.

Table 3.3 – Simulation results: Experiment 2 - Large number of relevant instruments and heteroskedastic structural disturbance

		Panel A: Instruments with decreasing relevance (Model 2a)					Panel B: Instruments equally important (Model 2b)				
	Estimator	**RRCUE**	**CUE**	**GMM**	**HFUL**	**HLIM**	**RRCUE**	**CUE**	**GMM**	**HFUL**	**HLIM**
K=15	Mean.bias	0.024	2.494e+10	0.102	0.003	-0.005	0.014	-1.23e+10	0.087	-0.002	-0.018
	Med.bias	0.030	0.003	0.106	0.014	0.005	0.023	0.001	0.091	0.010	0.001
	Var	0.027	4.577e+24	0.013	0.033	0.319	0.021	1.509e+24	0.011	0.029	0.224
	MAD	0.103	0.126	0.077	0.110	0.115	0.091	0.106	0.070	0.099	0.103
	Ndr (0.95-0.05)	0.529	0.685	0.375	0.585	0.621	0.470	0.579	0.350	0.528	0.564
	0.05 rej.Freq	0.042	0.049	0.215	0.052	0.049	0.038	0.043	0.192	0.050	0.046
K=30	Mean.bias	0.054	1.033e+11	0.144	-0.002	-0.022	0.044	-3.576e+10	0.135	-0.006	-0.071
	Med.bias	0.057	0.006	0.145	0.009	-0.000	0.047	0.005	0.136	0.011	0.001
	Var	0.022	2.760e+25	0.009	0.042	0.196	0.020	7.573e+24	0.008	0.039	10.418
	MAD	0.095	0.141	0.062	0.117	0.123	0.088	0.123	0.058	0.111	0.116
	Ndr (0.95-0.05)	0.481	0.800	0.307	0.643	0.692	0.456	0.724	0.291	0.618	0.674
	0.05 rej.Freq	0.046	0.056	0.449	0.052	0.048	0.043	0.048	0.426	0.054	0.050
K=50	Mean.bias	0.099	1.318e+10	0.174	-0.006	0.051	0.091	2.973e+10	0.168	-0.010	-0.095
	Med.bias	0.097	0.015	0.174	0.011	0.001	0.090	0.013	0.168	0.011	0.001
	Var	0.018	7.978e+25	0.006	0.065	137.613	0.015	5.630e+25	0.006	0.059	31.513
	MAD	0.079	0.162	0.051	0.131	0.138	0.075	0.151	0.050	0.126	0.133
	Ndr (0.95-0.05)	0.410	1.025	0.256	0.754	0.839	0.391	0.948	0.250	0.732	0.815
	0.05 rej.Freq	0.054	0.072	0.716	0.052	0.049	0.051	0.068	0.701	0.050	0.047
K=100	Mean.bias	0.160	-1.969e+10	0.207	-0.000	-0.531	0.157	5.229e+09	0.204	-0.001	-0.004
	Med.bias	0.161	0.073	0.207	0.020	0.008	0.155	0.065	0.205	0.018	0.006
	Var	0.010	5.630e+25	0.004	0.122	1.258e+03	0.009	1.749e+25	0.004	0.121	60.889
	MAD	0.064	0.199	0.041	0.167	0.177	0.063	0.193	0.041	0.167	0.177
	Ndr (0.95-0.05)	0.317	1.297	0.203	1.072	1.288	0.308	1.286	0.200	1.051	1.265
	0.05 rej.Freq	0.095	0.138	0.965	0.067	0.062	0.090	0.128	0.965	0.065	0.061

Note: Simulation results based on $10,000$ replications with a sample size of $n = 500$. We report six (06) measures of performance: the mean bias (Mean.bias), the median bias (Med.bias), the variance of estimates (Var), the median absolute deviation (MAD), the nine decile range (Ndr$(0.95-0.05)$), and the nominal 5% rejection frequency (0.05 rej.Freq) for the Wald test of $H_0 : \beta = \beta_0$.

balance between the size of the estimates and the precision, particularly in the case of $K = 18$. For $K = 4$, RRCUE produces an estimate of 5.704, comparable to LIML and HLIM, but with a standard error of 1.059, which is slightly higher than other methods and might indicate the fact that regularization is not needed in this low-dimensional case. However, as the number of instruments increases to $K = 18$, RRCUE achieves a precise estimate of 4.813 with a relatively low standard error of 0.639, maintaining stability in the presence of many instruments.

Compared to other estimators, RRCUE performs robustly with many instruments, showing superior precision compared to LIML, HLIM, and HFUL, which all experience substantial increases in standard errors. For instance, HLIM and HFUL report estimates of 6.828 and 6.561, respectively, but with much larger standard errors (1.610 and 1.497), indicating less precision, which is consistent with the simulation results. In contrast, RRCUE's regularized approach effectively controls for many instruments, offering both a reasonable estimate and greater reliability, making it a competitive choice for empirical analysis in such settings.

3.8. Conlusion

This paper introduces the Ridge-regularized Continuous Updating Estimator (RRCUE) to address the challenges posed by using a large number of instruments/ moments to improve efficiency in moment-based estimation in a framework where the parameter of interest is defined by a single conditional moment restriction. Through theoretical analysis and Monte Carlo simulations, we demonstrate that RRCUE offers a significant reduction in dispersion and improves efficiency compared to standard CUE. Despite introducing a small bias, the estimator remains robust in high-dimensional settings where the number of moments increases with the sample size, providing consistent and precise estimates. We show that RRCUE is competitive and sometimes outperforms state-of-the-art estimators like HLIM and HFUL of Hausman et al. (2012), specifically in the linear instrumental variable framework with het-

Table 3.4 – Estimates of the effect of social infrastructure on growth

	OLS	2SLS	LIML	GMM	CUE	HLIM	HFUL	T2SLS	TLIML	RRCUE
$K = 4$	3.074	5.412	5.938	5.367	5.728	5.982	5.812	5.628	5.958	5.704
	(0.296)	(0.777)	(1.012)	(0.695)	(1.017)	(0.936)	(0.874)	(0.807)	(0.948)	(1.059)
								$\alpha = 0.500$	$\alpha = 0.500$	$\alpha = 0.025$
$K = 18$	3.074	3.986	6.093	3.603	4.764	6.828	6.561	4.438	5.523	4.813
	(0.296)	(0.427)	(1.597)	(0.169)	(0.778)	(1.610)	(1.497)	(0.507)	(1.002)	(0.639)
								$\alpha = 0.020$	$\alpha = 0.010$	$\alpha = 0.012$

Note: The sample consists of $n = 79$ countries for which no data were imputed. We present results for both $K = 4$ instruments (benchmark) and for many instruments ($K = 18$), as well as for alternative estimators for comparison purposes. Standard errors are in parentheses. For CUE with 18 instruments, we report the many-instruments robust standard error of Newey and Windmeijer (2009), and for RRCUE, we report its regularized counterpart defined in Section 3.4.

eroskedasticity and many instruments. These findings make RRCUE a promising tool for empirical research, particularly in econometric applications where a large set of instruments/moments is available and there is either no rule to select a subset of them or they have almost the same explanatory power. Promising future research may explore the extension of our regularization scheme to the generalized empirical likelihood (GEL) class of estimators. This research path is particularly interesting as Newey and Smith (2004) justified that the empirical likelihood (EL) estimator enjoys good performance in terms of higher-order bias within the GEL class. Therefore, the EL estimator might be a good candidate to address overidentification bias in the many moments setting. Additionally, an extension to weakly dependent data would be possible by adapting our methodology to the three-step Euclidean empirical likelihood estimators (see, e.g., Guay and Pelgrin, 2016). Further promising future work includes the investigation of higher-order expansions to derive an approximate mean squared error for the optimal selection of the regularization parameter, as suggested by Carrasco (2012) and Carrasco and Tchuente (2015).

Bibliography

J. G. Altonji, A. A. Smith Jr, and I. Vidangos. Modeling earnings dynamics. *Econometrica*, 81(4):1395–1454, 2013.

J. D. Angrist and B. Frandsen. Machine labor. *Journal of Labor Economics*, 40(S1): S97–S140, 2022.

J. D. Angrist and A. B. Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

P. A. Bekker. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society*, pages 657–681, 1994.

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.

M. Carrasco. A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398, 2012.

M. Carrasco and J.-P. Florens. Generalization of gmm to a continuum of moment conditions. *Econometric Theory*, 16(6):797–834, 2000.

M. Carrasco and G. Tchuente. Regularized liml for many instruments. *Journal of Econometrics*, 186(2):427–442, 2015.

M. Carrasco and G. Tchuente. Efficient estimation with many weak instruments using regularization techniques. *Econometric Reviews*, 35(8-10):1609–1637, 2016.

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

G. Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of econometrics*, 34(3):305–334, 1987.

J. Chang, S. X. Chen, and X. Chen. High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics*, 185(1): 283–304, 2015.

J. C. Chao and N. R. Swanson. Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692, 2005.

A. Dmitriev. Institutions and growth: evidence from estimation methods robust to weak instruments. *Applied Economics*, 45(13):1625–1635, 2013.

S. G. Donald and W. K. Newey. Choosing the number of instruments. *Econometrica*, 69(5):1161–1191, 2001.

S. G. Donald, G. W. Imbens, and W. K. Newey. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, 117 (1):55–93, 2003.

S. G. Donald, G. W. Imbens, and W. K. Newey. Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics*, 152(1):28–36, 2009.

J. Eaton, S. Kortum, and F. Kramarz. An anatomy of international trade: Evidence from french firms. *Econometrica*, 79(5):1453–1498, 2011.

J. A. Frankel and D. H. Romer. Does trade cause growth? *American Economic Review*, 89(3):379–399, June 1999. doi: 10.1257/aer.89.3.379. URL https://www.aeaweb.org/articles?id=10.1257/aer.89.3.379.

W. A. Fuller. Some properties of a modification of the limited information estimator. *Econometrica: Journal of the Econometric Society*, pages 939–953, 1977.

D. Giannone, M. Lenza, and G. E. Primiceri. Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437, 2021.

A. Guay and F. Pelgrin. Using implied probabilities to improve the estimation of unconditional moment restrictions for weakly dependent data. *Econometric Reviews*, 35(3):344–372, 2016.

P. Guggenberger. Monte-carlo evidence suggesting a no moment problem of the continuous updating estimator. *Economics Bulletin*, 3(13):1–6, 2005.

J. Hahn and J. Hausman. Weak instruments: Diagnosis and cures in empirical econometrics. *American Economic Review*, 93(2):118–125, 2003.

R. E. Hall and C. I. Jones. Why do some countries produce so much more output per worker than others? *The quarterly journal of economics*, 114(1):83–116, 1999.

C. Han, L. Orea, and P. Schmidt. Estimation of a panel data model with parametric temporal variation in individual effects. *Journal of Econometrics*, 126(2):241–267, 2005.

C. Hansen and D. Kozbur. Instrumental variables estimation with many weak instruments using regularized jive. *Journal of Econometrics*, 182(2):290–308, 2014.

C. Hansen, J. Hausman, and W. Newey. Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 26(4):398–422, 2008.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, pages 1029–1054, 1982.

L. P. Hansen, J. Heaton, and A. Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, 1996.

J. Hausman, R. Lewis, K. Menzel, and W. Newey. Properties of the cue estimator and a modification with moments. *Journal of Econometrics*, 165(1):45–57, 2011.

J. A. Hausman, W. K. Newey, T. Woutersen, J. C. Chao, and N. R. Swanson. Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics*, 3(2):211–255, 2012.

W. K. Newey. Uniform convergence in probability and stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pages 1161–1167, 1991.

W. K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of econometrics*, 79(1):147–168, 1997.

W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

W. K. Newey and R. J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

W. K. Newey and F. Windmeijer. Generalized method of moments with many weak moment conditions. *Econometrica*, 77(3):687–719, 2009.

Z. Shi. Econometric estimation with high-dimensional moment equalities. *Journal of Econometrics*, 195(1):104–119, 2016.

J. Stock and M. Yogo. *Asymptotic distributions of instrumental variables statistics with many instruments*, volume 6. Chapter, 2005.

3.9. Appendix

Throughout the Appendix, C will denote a generic positive constant that may be different in different uses, and M, CS, and T the Markov, Cauchy-Schwarz, and triangle inequalities, respectively. Also, with probability approaching one will be abbreviated as w.p.a.1., p.d. and p.s.d. will be the abbreviations for positive definite and positive semidefinite matrix, respectively, and CLT will refer to the Lindeberg-Lévy central limit theorem.

Proof of Theorem 3.1. By $s(\nu)$ quadratic, a second-order Taylor expansion is exact,

giving

$$\begin{aligned}\widehat{P}(\beta,\lambda) &= \sum_{i=1}^{n} s\left(\lambda' g_i(\beta)\right)/n - \frac{\alpha}{2}\lambda'\lambda \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{s(0) + s'(0)\lambda' g_i(\beta) + \frac{s''(0)}{2}\left(\lambda' g_i(\beta)\right)^2\right\} - \frac{\alpha}{2}\lambda'\lambda \\ &= -\hat{g}(\beta)'\lambda - \frac{1}{2}\lambda'\hat{\Omega}(\beta)\lambda - \frac{\alpha}{2}\lambda'\lambda \\ &= -\hat{g}(\beta)'\lambda - \frac{1}{2}\lambda'\left(\hat{\Omega}(\beta) + \alpha I_K\right)\lambda.\end{aligned}$$

By concavity of $\widehat{P}(\beta,\lambda)$ in λ, any solution $\hat{\lambda}(\beta)$ to the FOCs $0 = \hat{g}(\beta) + (\widehat{\Omega}(\beta) + \alpha I)\lambda$ will maximize $\widehat{P}(\beta,\lambda)$ with respect to λ holding β fixed. That is $\hat{\lambda}(\beta) = -\left[\hat{\Omega}(\beta) + \alpha I\right]^{-1}\hat{g}(\beta)$ maximizes $\widehat{P}(\beta,\lambda)$ holding holding β fixed. Then the penalized GEL objective function is given by

$$\widehat{P}\left(\beta, \hat{\lambda}(\beta)\right) = \frac{1}{2}\hat{g}(\beta)'\left[\hat{\Omega}(\beta) + \alpha I_K\right]^{-1}\hat{g}(\beta).$$

Therefore, the penalized GEL objective function is a monotonic increasing transformation of the regularized CUE objective function, so that the result follows. □

Proof of Theorem 3.2. As justified in the proof of Theorem 3.1, $\hat{\lambda}(\beta) = -\left[\hat{\Omega}(\beta) + \alpha I\right]^{-1}\hat{g}(\beta)$ maximizes $\widehat{P}(\beta,\lambda)$ holding β fixed.

By the envelope theorem, the FOCs for the regularized CUE $\hat{\beta}$ are given by

$$\begin{aligned}
0 &= \left.\frac{\partial \widehat{P}(\beta, \lambda(\beta))}{\partial \beta}\right|_{\beta=\hat{\beta}} \\
&= \left.\frac{\partial \widehat{P}(\beta, \lambda)}{\partial \beta}\right|_{\lambda=\hat{\lambda}(\beta),\, \beta=\hat{\beta}} \\
&= n^{-1} \sum_{i=1}^{n} s_1\left(\lambda' g_i(\beta)\right) G_i(\beta)' \lambda \Big|_{\lambda=\hat{\lambda}(\beta),\, \beta=\hat{\beta}} \\
&= n^{-1} \sum_{i=1}^{n} s_1\left(\hat{v}_i\right) G_i\left(\hat{\beta}\right)' \hat{\lambda},
\end{aligned}$$

where $\hat{v}_i = \hat{\lambda}\hat{g}_i$. Multiplying by $-n\left(\sum_{i=1}^{n} s_1\left(\hat{v}_i\right)\right)^{-1}$ and using $\hat{\lambda} = -\left[\hat{\Omega}\left(\hat{\beta}\right) + \alpha I\right]^{-1} \hat{g}\left(\hat{\beta}\right)$ give the result. □

We now give some preliminary lemmas for the proof of Theorem 3.3 and of Theorem 3.4. The proof of Theorem 3.3 will be based on the following lemma, borrowed from DNI03, with suitable choices of functions $\hat{R}(\beta)$ and $R(\beta)$. The proof of this lemma can be found in the DIN03's Appendix.

Lemma 3.1. *(Lemma A1 of Donald et al. (2003)) Suppose that (i) $R(\beta)$ has a unique minimum at $\beta_0 \in \mathscr{B}$; (ii) B is compact; (iii) $R(\beta)$ is continuous; and (iv) $\sup_{\beta\in\mathscr{B}} |\hat{R}(\beta) - R(\beta)| \xrightarrow{p} 0$. Then for any $\tilde{\beta} \in \mathscr{B}$, if $\hat{R}(\tilde{\beta}) \xrightarrow{p} R(\beta_0)$ then $\tilde{\beta} \xrightarrow{p} \beta_0$.*

Lemma 3.2. *If Assumption 3.1(a) is satisfied and $\sigma(z)^2 := E\left[\rho(w, \beta_0)^2 | z\right]$ is bounded then $\|\hat{g}(\beta_0)\| = O_p(n^{-1/2})$*

Proof. Let $q_i = \tilde{q}^K(z_i) = q^K(z_i)/\zeta(K)$ and $\rho_i = \rho(w_i, \beta)$. By Assumption 3.1(a) $\sup_{\beta\in\mathscr{B}} \left\|\tilde{q}^K(z)\right\| \leq C$ so that $E\left[\|q_i\|^2\right] = O(1)$. It follows from i.i.d. data and the law of

iterated expectations that

$$\begin{aligned} E\left[\|\hat{g}(\beta_0)\|^2\right] &= E\left[\left\|\frac{1}{n}\sum_{i=1}^{n} q_i\rho_i\right\|^2\right] \\ &= E\left[\frac{1}{n^2}\sum_{i,j}(q_i\rho_i)'\left(q_j\rho_j\right)\right] \\ &= E\left[\rho_i^2\|q_i\|^2\right]/n \\ &= E\left[E\left[\rho_i^2|z_i\right]\|q_i\|^2\right]/n \\ &= E\left[\sigma_i^2\|q_i\|^2\right]/n \leq C/n. \end{aligned}$$

The conclusion then follows by M. □

Lemma 3.3. *If Assumption 3.1(a) is satisfied, $U_i := U(z_i)$ a nonnegative scalar function bounded away from zero, $P_i = q_i U_i^{1/2}$, $P = [P_1, \cdots, P_n]'$, $Q^\alpha = P\left[P'P/n + \alpha I\right]^{-1} P'/n$ with $\alpha > 0$, and*
$\left(\lambda_j, \phi_j : j = 1, 2, \ldots, K\right)$ the eigenvalues and orthonormal eigenvectors $E\left[U_i q_i q_i'\right]$ then

(i) $\operatorname{tr}(E[Q^\alpha]) = O(1/\alpha)$;

(ii) For all x and y, $x'Q^\alpha y \leq \|x\|\|y\|$ so that $\lambda_{\max}(Q^\alpha) \leq 1$;

(iii) $x'(I - Q^\alpha)^2 x \leq x'(I - Q^\alpha)x$ *for all x;*

(iv) If $\bar{x}$ is an n-dimensional vector such that $\|\bar{x}\|/\sqrt{n} = O_p(1)$, for each K there is a K-dimensional vector γ_K such that $\|\bar{x} - P\gamma_K\|/\sqrt{n} = o_p(1)$, and

$$\sum_{j=1}^{\infty} \frac{\left(E[\bar{x}_i P_i]'\phi_j\right)^2}{\lambda_j} < \infty,$$

then $\bar{x}'(I - Q^\alpha)\bar{x}/n \xrightarrow{p} 0$ as $n \to \infty$ and $\alpha \to 0$.

Proof. To prove (i) remark that $P'P/n + \alpha I \geqslant \alpha I$ so that $(P'P/n + \alpha I)^{-1} P'P \leqslant P'P/\alpha$ and therefore $\operatorname{tr}(Q^\alpha) \leqslant \operatorname{tr}(P'P)/(n\alpha)$. Also, note that $P'P = \sum_{i=1}^n U_i q_i q_i'$ so that $\operatorname{tr}(P'P)/n = n^{-1}\sum_{i=1}^n U_i \operatorname{tr}\left(q_i q_i'\right) \leqslant Cn^{-1}\sum_{i=1}^n \|q_i\|^2 \leq C$ by $U(z)$ bounded and Assumption 3.1(a).

To prove (ii) we make use of singular value decomposition (SVD). Recall that $T_n = P/\sqrt{n}$ is an $n \times K$ matrix. Let $T_n = \hat{\Psi} D \widehat{\Phi}$ denotes its SVD, where $\hat{\Psi}$ is an $n \times n$ orthogonal matrix, D is an $n \times K$ rectangular diagonal matrix with nonnegative real numbers on the diagonal, $\widehat{\Phi}$ is an $K \times K$ matrix. Let $\sqrt{\hat{\lambda}_i} = D_{ii}$ denote the diagonal entries of D known as singular values of P. The number of nonzero singular values is equal to the rank r of P. The columns of $\widehat{\Psi}$ and $\widehat{\Phi}$ form two sets of orthonormal bases $\hat{\psi}_1, \ldots, \hat{\psi}_n$ and $\hat{\phi}_1, \ldots, \hat{\phi}_K$. If singular values $\sqrt{\hat{\lambda}_i}$ are sorted in decreasing order such that $\sqrt{\hat{\lambda}_1} \geq \sqrt{\hat{\lambda}_2} \geq \cdots \geq \sqrt{\hat{\lambda}_r} > \sqrt{\hat{\lambda}_{r+1}} = 0$ then the SVD can be written as $T_n = \sum_{i=1}^{r} \sqrt{\hat{\lambda}_i} \hat{\psi}_i \hat{\phi}_i'$, where $r \leqslant \min(n, K)$. It follows that for all $j = 1, \ldots, K$, $T_n' T_n \hat{\phi}_j = \hat{\lambda}_j \hat{\phi}_j$ so that $\left(\hat{\lambda}_j, \hat{\phi}_j : j = 1, 2, \ldots, K\right)$ is the system of eigenvalues and orthonormal eigenvectors of $T_n' T_n$.

Also, note that $Q^\alpha = T_n \left[T_n' T_n + \alpha I\right]^{-1} T_n' = \sum_{j=1}^{n} \frac{\hat{\lambda}_j}{\hat{\lambda}_j + \alpha} \hat{\psi}_j \hat{\psi}_j'$ with $\hat{\lambda}_j = 0$ for all $j > r$. $(\hat{\psi}_1, \ldots, \hat{\psi}_n)$ being an orthonormal basis $Q^\alpha \hat{\psi}_j = \frac{\hat{\lambda}_j}{\hat{\lambda}_j + \alpha} \hat{\psi}_j$ for all $j = 1, \ldots, n$. It follows that eigenvalues of Q^α are $\frac{\hat{\lambda}_j}{\hat{\lambda}_j + \alpha}$ for $j = 1, \ldots, n$ and the associated eigenvectors are respectively $\hat{\psi}_j$, $j = 1, \ldots, n$. Therefore $\lambda_{\max}(Q^\alpha) \leqslant 1$ so that for all vectors x and y, $x' Q^\alpha y \leq \lambda_{\max}(Q^\alpha) \|x\| \|y\| \leq \|x\| \|y\|$.

To prove (iii) note that $I - Q^\alpha = \sum_{j=1}^{n} \frac{\alpha}{\hat{\lambda}_j + \alpha} \hat{\psi}_j \hat{\psi}_j'$ so that $\lambda_{\max}(I - Q^\alpha) \leq 1$. Also, for $\alpha > 0$, $I - Q^\alpha$ is a symmetric positive semidefinite matrix. Let $(I - Q^\alpha)^{1/2}$ be a symmetric square root of $I - Q^\alpha$. Then,

$$x' (I - Q^\alpha)^2 x = x' (I - Q^\alpha)^{1/2} (I - Q^\alpha) (I - Q^\alpha)^{1/2} x \leqslant \left\| (I - Q^\alpha)^{1/2} x \right\|^2 = x' (I - Q^\alpha) x,$$

giving the result.

It remains to prove (iv). Note that for $\alpha = 0$, Q^α coincide with $Q := P(P'P)^- P' = \sum_{j=1}^{r} \hat{\psi}_j \hat{\psi}_j'$, where $(\cdot)^-$ is the Moore-Penrose generalized inverse. By definition, $I - Q$ is idempotent and satisfies $QP = P$ so that

$$\bar{x}'(I-Q)\bar{x}/n = (\bar{x}-P\gamma_K)'(I-Q)(\bar{x}-P\gamma_K)/n \leqslant \frac{\|\bar{x}-P\gamma_K\|^2}{n} \xrightarrow{p} 0.$$

Also, the function $\lambda/(\alpha+\lambda)$ is increasing in λ and reaches its maximum for the maximal eigenvalue (which is bounded by $\mathrm{tr}(P'P)/n \leq C$) and therefore $\sup_\lambda \lambda/(\alpha+\lambda) \leq C$. Also, by the SVD, $\hat{\psi}_j = T_n\hat{\phi}_j/\sqrt{\hat{\lambda}_j} = \frac{1}{\sqrt{\hat{\lambda}_j}}\left[P_1'\hat{\phi}_j, \cdots, P_n'\hat{\phi}_j\right]'/\sqrt{n}$ so that $\left(\bar{x}'\hat{\psi}_j\right)^2 = \frac{n}{\hat{\lambda}_j}\left(\hat{E}\left[\bar{x}_i P_i\right]'\hat{\phi}_j\right)^2$ where $\hat{E}\left[\bar{x}_i P_i\right] = \sum_{i=1}^n \bar{x}_i P_i/n$. It follows from $Q-Q^\alpha = \sum_{j=1}^r \frac{\alpha}{\hat{\lambda}_j+\alpha}\hat{\psi}_j\hat{\psi}_j'$ that

$$\begin{aligned}
\bar{x}'(Q-Q^\alpha)\bar{x}/n &= \sum_{j=1}^r \frac{\alpha}{\alpha+\hat{\lambda}_j}\left(\bar{x}'\hat{\psi}_j\right)^2/n \\
&= \sum_{j=1}^r \frac{\alpha\hat{\lambda}_j}{\alpha+\hat{\lambda}_j}\frac{\left(\hat{E}\left[\bar{x}_i P_i\right]'\hat{\phi}_j\right)^2}{\hat{\lambda}_j} \\
&\leqslant \sup_\lambda\left(\frac{\alpha\lambda}{\alpha+\lambda}\right)\sum_{j=1}^r \frac{\left(\hat{E}\left[\bar{x}_i P_i\right]'\hat{\phi}_j\right)^2}{\hat{\lambda}_j} \\
&\leq C\alpha\sum_{j=1}^r \frac{\left(\hat{E}\left[\bar{x}_i P_i\right]'\hat{\phi}_j\right)^2}{\hat{\lambda}_j}. \qquad \text{(A.15)}
\end{aligned}$$

At the limit, the sum in (A.15) is finite by hypothesis in the statement of Lemma 3.3 so that $\bar{x}'(Q-Q^\alpha)\bar{x}/n = O_p(\alpha)$. It follows that $\bar{x}'(I-Q^\alpha)\bar{x}/n = \bar{x}'(I-Q)\bar{x}/n + \hat{x}'(Q-Q^\alpha)\bar{x}/n \xrightarrow{p} 0$ as $n\to\infty$ and $\alpha\to 0$, giving the conclusion in (iv). □

Lemma 3.4. *If Assumption 3.1 is satisfied, (i)* $\hat{\beta} \xrightarrow{p} \bar{\beta}$, *(ii)* $a_i(\beta) := a\left(w_{i,}, \beta\right)$ *and* $b_i(\beta) := b\left(w_i, \beta\right)$ *are scalar functions that are continuous at* $\bar{\beta}$ *w.p.1 and there is a neighborhood* $\mathcal{N}$ *of* $\bar{\beta}$ *such that* $E\left[\sup_{\beta\in\mathcal{N}}|a_i(\beta)|^2\right] < \infty$ *and* $E\left[\sup_{\beta\in\mathcal{N}}|b_i(\beta)|^2\right] < \infty$, $E\left[a_i(\bar{\beta})^2|z_i\right]$ *and* $E\left[b_i(\bar{\beta})^2|z_i\right]$ *are bounded; (iii)* $U_i := U(z_i)$ *is a nonnegative scalar function bounded away from zero; (iv)* $K\to\infty$, $\alpha\to 0$, *and* $n\alpha\to\infty$ *as*

$n \to \infty$, *then*

$$\widehat{\Lambda}^{\alpha} := \left(\frac{1}{n}\sum_{i=1}^{n} a_i(\hat{\beta})q_i\right)' \left(\frac{1}{n}\sum_{i=1}^{n} U_i q_i q_i' + \alpha I\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} b_i(\hat{\beta})q_i\right) \xrightarrow{p} \Lambda,$$

where $\Lambda := E\left[E\left[a_i(\bar{\beta})|z_i\right]U_i^{-1}E\left[b_i(\bar{\beta})|z_i\right]\right]$.

Proof. Let $P_i = q_i U_i^{1/2}$, $P = [P_1, \ldots, P_n]'$, $A_i(\beta) = U_i^{-1/2}a_i(\beta)$, $A(\beta) = [A_1(\beta), \cdots, A_n(\beta)]'$, $\hat{A} = A(\hat{\beta})$, $A = A(\bar{\beta})$, $B_i(\beta) = U_i^{-1/2}b_i(\beta)$, $B(\beta) = [B_1(\beta), \ldots, B_n(\beta)]'$, and $\hat{B} = B(\hat{\beta})$, and $B = B(\bar{\beta})$. Note that $\sum_{i=1}^{n} U_i q_i q_i' = P'P$, $\sum_{i=1}^{n} a_i(\hat{\beta})q_i' = \hat{A}'P$, and $\sum_{i=1}^{n} b_i(\hat{\beta})q_i = P\hat{B}$ so that for $Q^{\alpha} = P\left(P'P/n + \alpha I\right)P'/n$, $\hat{\Lambda}^{\alpha} = \hat{A}'P\left(P'P/n + \alpha I\right)^{-1}P'\hat{B}/n^2 = \hat{A}'Q^{\alpha}\hat{B}/n$.

Let $\Delta(w,\beta) = U(z)^{-1}[b(w,\beta) - b(w,\bar{\beta})]^2$. By hypothesis in the statement of Lemma 3.4, $\Delta(w,\beta)$ is continuous with respect β in a neighborhood $\mathcal{N}$ of $\bar{\beta}$. Moreover, $E\left[\sup_{\beta\in\mathcal{N}}|\Delta(w,\beta)|\right] \leq$
$CE\left[\sup_{\beta\in\mathcal{N}}|b(w,\beta)|^2\right] < \infty$. It follows by Lemma 4.3 of Newey and McFadden (1994) with $a(z,\theta)$ there equal to $\Delta(w,\beta)$ that $\|\hat{B} - B\|^2/n = \sum_{i=1}^{n}\Delta\left(\omega_i,\hat{\beta}\right)/n \xrightarrow{p}$ $E\left[\Delta\left(\omega_i,\bar{\beta}\right)\right] = 0$. Therefore by Lemma 3.3(ii)

$$\hat{T}_B^{\alpha} := (\hat{B} - B)'Q^{\alpha}(\hat{B} - B)/n \leqslant \lambda_{\max}(Q^{\alpha})\|\hat{B} - B\|^2/n \leq \|\hat{B} - B\|^2/n \xrightarrow{p} 0.$$

For $Z = [z_1, \cdots, z_n]'$, let $a_i = a_i(\bar{\beta})$, $\bar{a}_i = E\left[a_i|z_i\right]$, and note that $\bar{A} \overset{\text{def}}{=}$ $E[A|Z] = \left(U_1^{-1/2}\bar{a}_1, \cdots, U_n^{-1/2}\bar{a}_n\right)'$. Note that from i.i.d. observations,

$$E\left[(A - \bar{A})(A - \bar{A})'|Z\right] = \text{Diag}\left(U_1^{-1}V\left[a_1|z_1\right], \cdots, U_n^{-1}V\left[a_n|z_n\right]\right) \leq CI,$$

by $E\left[|a_i|^2|z_i\right]$ bounded and U_i bounded away from zero.

Let Σ be the Cholesky factor of the symmetric and positive definite matrix $(P'P + \alpha I)^{-1}$. Then $Q^{\alpha} = P\Sigma'\Sigma P$. By the law of iterated expectations and Lemma 3.3(i) it

follows for $\widetilde{T}_A^\alpha \overset{def}{=} (A-\bar{A})'Q^\alpha(A-\bar{A})/n$ that

$$\begin{aligned}
E\left[\widetilde{T}_A^\alpha\right] &= \operatorname{tr} E\left[E\left[\widetilde{T}_A^\alpha|Z\right]\right] \\
&= E\left[E\left[\operatorname{tr}\left((A-\bar{A})(A-\bar{A})'Q^\alpha\right)|Z\right]\right]/n \\
&= E\left[\operatorname{tr}\left(\Sigma P' E\left[(A-\bar{A})(A-\bar{A})'|Z\right]P\Sigma'\right)\right]/n \\
&\leq CE\left[\operatorname{tr}\left(\Sigma P' P\Sigma'\right)\right]/n \\
&\leq CE\left[\operatorname{tr}\left(P\Sigma'\Sigma P'\right)\right]/n \\
&\leq CE\left[\operatorname{tr}(Q^\alpha)\right]/n \leq C/(n\alpha) \to 0 \text{ as } n \to \infty.
\end{aligned}$$

It then follows by M that $\widetilde{T}_A^\alpha \xrightarrow{p} 0$. Also the same result holds for $\widetilde{T}_B^\alpha$.

By Assumption 3.1(b) there exists a $K\times 1$ vector, γ_K such that $E\left[\left\{U_i^{-1}\bar{a}_i - q_i'\gamma_K\right\}^2\right] \to 0$ as $K \to \infty$. Then by M

$$\begin{aligned}
\|\bar{A}-P\gamma_K\|^2/n &= \sum_{i=1}^n \left|U_i^{-1/2}\bar{a}_i - P_i'\gamma_K\right|^2/n \\
&= \sum_{i=1}^n U_i^{1/2}\left|U_i^{-1}\bar{a}_i - q_i'\gamma_K\right|^2/n \\
&\leq C\sum_{i=1}^n \left|U_i^{-1}\bar{a}_i - q_i'\gamma_K\right|^2/n \xrightarrow{p} 0 \text{ as } n,K \to \infty.
\end{aligned}$$

Also by M $\bar{A}'\bar{A}/n = O_p(1)$. By Assumption 3.1(c), hypothesis in Lemma 3.3 is satisfied for $\bar{x} = \bar{A}$ so that part (iv) of Lemma 3.3 gives $\bar{T}_A^\alpha \overset{def}{=} \bar{A}'(I-Q^\alpha)\bar{A}/n \xrightarrow{p} 0$. The analogous result holds for B replacing A.

Next, note that by CS

$$\begin{aligned}
T_A^\alpha &\overset{def}{=} (\hat{A}-\bar{A})'Q^\alpha(\hat{A}-\bar{A}) = (\hat{A}-A+A-\bar{A})'Q^\alpha(\hat{A}-A+A-\bar{A}) \\
&\leq \widehat{T}_A^\alpha + \tilde{T}_A^\alpha + 2\sqrt{\hat{T}_A^\alpha}\sqrt{\tilde{T}_A^\alpha} \xrightarrow{p} 0.
\end{aligned}$$

Then by CS and T,

$$|\hat{A}'Q^\alpha\hat{B}/n-\bar{A}'\bar{B}/n| =|(\hat{A}-\bar{A})'Q^\alpha(\hat{B}-\bar{B})+(\hat{A}-\bar{A})'Q^\alpha\bar{B}+\bar{A}'Q^\alpha(\hat{B}-\bar{B})-\bar{A}'(I-Q^\alpha)\bar{B}|/n$$
$$\leq\sqrt{T_A^\alpha}\sqrt{T_B^\alpha}+\sqrt{T_A^\alpha}\sqrt{\bar{B}'\bar{B}/n}+\sqrt{\bar{A}'\bar{A}/n}\sqrt{T_A^\alpha}+\sqrt{\bar{T}_A^\alpha}\sqrt{\bar{T}_B^\alpha}\xrightarrow{p}0.$$

Nothing that $\bar{A}'\bar{B}/n=\sum_{i=1}^n\bar{a}_iU_i^{-1}\bar{b}_i/n$ the conclusion follows by the standard law of large numbers. □

In the sequel, we will use the following notations

$$\hat{R}(\beta)=\hat{g}(\beta)'\widetilde{W}\hat{g}(\beta),\quad R(\beta)=E\left[(E\left[\rho(w,\beta)|z\right])^2\Big/E\left[\rho(w,\beta_0)^2|z\right]\right],\qquad\text{(A.16)}$$

where $\widetilde{W}^\alpha=\left(\widehat{A}+\alpha I\right)^{-1}$, $\hat{A}=\sum_{i=1}^n\rho_i^2q_iq_i'/n$, $\rho_i=\rho(w_i,\beta_0)$ and $q_i=q^K(z_i)/\zeta(K)$.

Lemma 3.5. *If Assumptions 3.1 and 3.2(a)-(c) are satisfied, $K\to\infty$, $\alpha\to0$, and $n\alpha\to\infty$ as $n\to\infty$, then $R(\beta)$ has a unique minimum at β_0, $R(\beta)$ is continuous on $\mathscr{B}$ and $\sup_{\beta\in\mathscr{B}}|\hat{R}(\beta)-R(\beta)|\xrightarrow{\mathrm{p}}0$.*

Proof. For any $\beta\neq\beta_0$ it follows from Assumption 3.2(a) that $E[\rho(w,\beta)|z]\neq0$ so that $R(\beta)=$
$E\left[(E\left[\rho(w,\beta)|z\right])^2\Big/E\left[\rho(w,\beta_0)^2|z\right]\right]>0=R(\beta_0)$ giving the first result.

To show the continuity of $R(\beta)$, note that by Assumptions 3.2(c)&3.2(e) , for all $\beta,\tilde{\beta}\in\mathscr{B}$

$$\begin{aligned}\left|R(\tilde{\beta})-R(\beta)\right|&\leqslant C\left|E\left[\left(E[\rho(w,\tilde{\beta})|z]\right)^2-(E[\rho(w,\beta)|z])^2\right]\right|\\&\leq CE\left[(E[\rho(w,\tilde{\beta})-\rho(w,\beta)|z])^2\right]\\&\leq CE\left[E\left[\left(\rho(w,\tilde{\beta})-\rho(w,\beta)\right)^2|z\right]\right]\\&\leqslant CE\left[\left(\rho(w,\tilde{\beta})-\rho(w,\beta)\right)^2\right]\\&\leqslant CE\left[\delta_1(w)^2\right]\|\tilde{\beta}-\beta\|^{2r}\\&\leq C\|\tilde{\beta}-\beta\|^{2r}.\end{aligned}$$

Therefore, $R(\beta)$ is continuous, being uniformly continuous.

To obtain the last conclusion, $\sup_{\beta\in B}|\hat{R}(\beta)-R(\beta)|\xrightarrow{p}0$, it suffices, by Corollary 2.2 of Newey (1991), to show that: (i) $\mathscr{B}$ is compact (it is the case by Assumption 3.2(b)); (ii) $\widehat{R}(\beta)\xrightarrow{p}R(\beta)$ for all $\beta\in\mathscr{B}$; and (iii) there is $\hat{D}=O_p(1)$ with $\left|\hat{R}(\tilde{\beta})-\hat{R}(\beta)\right|\leq\hat{D}\|\tilde{\beta}-\beta\|^r$ for all $\beta,\tilde{\beta}\in\mathscr{B}$.

To show (ii), apply Lemma 3.4 with $a(w,\beta)=b(w,\beta)=\rho(w,\beta)$ and $U(z)=1$ with β fixed. Hypotheses in the statement of Lemma 3.4 are satisfied by Assumption 3.2(c). The conclusion of Lemma 3.4 implies that $\widehat{R}(\beta)\xrightarrow{p}R(\beta)$ for all $\beta\in\mathscr{B}$ giving (ii).

To show (iii), let $\rho=(\rho_1,\ldots,\rho_n)'$ and $\tilde{\rho}=(\tilde{\rho}_1,\ldots,\tilde{\rho}_n)$ with $\rho_i=\rho(w_i,\beta)$ and $\tilde{\rho}_i=\rho\left(\omega_i,\tilde{\beta}\right)$. Also, note that $\hat{R}(\beta)=\rho'Q^\alpha\rho/n$ where Q^α is defined as in the statement of Lemma 3.3 for $U_i=1$. It follows by Lemma 3.3(ii) and by CS that

$$\begin{aligned}\left|\hat{R}(\tilde{\beta})-\hat{R}(\beta)\right|&=\left|\tilde{\rho}'Q^\alpha\tilde{\rho}-\rho'Q^\alpha\rho\right|/n\\&=\left|(\tilde{\rho}-\rho)'Q^\alpha\tilde{\rho}+\rho'Q^\alpha(\tilde{\rho}-\rho)\right|/n\\&\leq\lambda_{max}(Q^\alpha)\|\tilde{\rho}-\rho\|(\|\tilde{\rho}\|+\|\rho\|)/n\\&\leqslant\|\tilde{\rho}-\rho\|(\|\tilde{\rho}\|+\|\rho\|)/n.\end{aligned}$$

Note by Assumption 3.2(c) and M that $n^{-1/2}\|\tilde{\rho}-\rho\|=\left[\sum_{i=1}^n(\tilde{\rho}_i-\rho_i)^2/n\right]^{1/2}\leq\hat{D}_{\delta_1}\|\tilde{\beta}-\beta\|^r$, where $\hat{D}_{\delta_1}=\left[\sum_{i=1}^n\delta_1(w_i)^2/n\right]^{1/2}=O_p(1)$. Also, for any fixed $\bar{\beta}\in\mathscr{B}$, by Assumption 3.2(b),

$$\begin{aligned}\sup_{\beta\in\mathscr{B}}|\rho(w_i,\beta)|/\sqrt{n}&\leq\sup_{\beta\in\mathscr{B}}\left|\rho(w_i,\beta)-\rho\left(w_i,\bar{\beta}\right)\right|/\sqrt{n}+\bar{D}\\&\leq\hat{D}_{\delta_1}\sup_{\beta\in\mathscr{B}}\|\beta-\bar{\beta}\|^r+\bar{D}\leq C\hat{D}_{\delta_1}+\bar{D},\end{aligned}$$

where $\bar{D}=\left[\sum_{i=1}^n\rho\left(w_i,\bar{\beta}\right)^2/n\right]^{1/2}=O_p(1)$ by Assumption 3.2(c) and M. Therefore $\left|\hat{R}(\tilde{\beta})-\hat{R}(\beta)\right|\leq\hat{D}\|\tilde{\beta}-\beta\|^r$, where $\hat{D}=2\hat{D}_{\delta_1}\left(C\hat{D}_{\delta_1}+\bar{D}\right)=O_p(1)$, giving (iii). □

Lemma 3.6. *If Assumptions 3.1(a) and 3.2(f) are satisfied, for* $\Lambda_n=\{\lambda:\|\lambda\|\leq\delta_n\}$,

where δ_n is a sequence of nonnegative real numbers such that $\delta_n n^{1/\gamma} \to 0$ as n goes to infinity, then we have $\max_{\substack{\beta\in\mathscr{B},\,\lambda\in\Lambda_n\\ 1\le i\le n}} |\lambda' g_i(\beta)| \xrightarrow{p} 0$ and w.p.a.1, $\Lambda_n \subseteq \hat{\Lambda}(\beta)$ for all $\beta \in \mathscr{B}$.

Proof. Let $b_i = \sup_{\beta\in\mathscr{B}} |\rho(w_i,\beta)|$. By Assumption 3.2(f), $E\left[b_i^\gamma\right] < C$. If follows by M that $\max_{1\le i\le n} b_i = \left\{\max_i b_i^\gamma\right\}^{1/\gamma} \le \left\{\sum_{i=1}^n b_i^\gamma\right\}^{1/\gamma} = n^{1/\gamma}\left\{\sum_{i=1}^n b_i^\gamma/n\right\}^{1/\gamma} \le n^{1/\gamma} O_p\left(\left\{E\left[b_i^\gamma\right]\right\}^{1/\gamma}\right) = O_p\left(n^{1/\gamma}\right)$.

It then follows by Assumption 3.1(a) that

$$X_n := \max_{\substack{\beta\in\mathscr{B},\,\lambda\in\Lambda_n\\ 1\le i\le n}} \left|\lambda' g_i(\beta)\right| = \max_{\substack{\beta\in\mathscr{B},\,\lambda\in\Lambda_n\\ 1\le i\le n}} \left|\lambda' q_i \rho(w_i,\beta)\right| \le \delta_n \max_{1\le i\le n} b_i = \delta_n O_p\left(n^{1/\gamma}\right) \xrightarrow{p} 0,$$

giving the first conclusion.

Also, since $\mathscr{V}$ is a neighborhood of 0, by the first conclusion $X_n \in \mathscr{V}$ w.p.a.1. Equivalently, $|\lambda' g_i(\beta)| \in \mathscr{V}$ for all $\beta \in \mathscr{B}$, $\lambda \in \Lambda_n$, and $i = 1,\dots,n$. It follows that $\Lambda_n \subseteq \hat{\Lambda}(\beta)$ for all $\beta \in \mathscr{B}$, giving the second conclusion. □

Lemma 3.7. *If Assumptions 3.1(a) and 3.2(f) are satisfied, δ_n a sequence of nonnegative real numbers such that $\delta_n n^{1/\gamma} \to 0$, a_n a sequence of real numbers such that $\alpha\delta_n a_n \to \infty$ as n goes to infinity, $\tilde{\beta}$ an estimator of β_0 with $\left\|\hat{g}\left(\tilde{\beta}\right)\right\| = O_p\left(a_n^{-1}\right)$ then $\sup_{\lambda\in\hat{\Lambda}(\tilde{\beta})} \hat{P}(\tilde{\beta},\lambda) = O_p\left(\alpha^{-1}a_n^{-2}\right)$, $\tilde{\lambda} = \arg\min_{\lambda\in\hat{\Lambda}(\tilde{\beta})} \hat{P}(\tilde{\beta},\lambda)$ exists w.p.a.1 and $\left\|\tilde{\lambda}\right\| = O_p\left(\alpha^{-1}a_n^{-1}\right)$.*

Proof. Let $\tilde{g} = \hat{g}(\tilde{\beta})$ and $\tilde{\Omega} = \hat{\Omega}(\tilde{\beta})$. Also let Λ_n be as defined in Lemma 3.6. It is obvious that $\hat{P}(\beta,\lambda)$ is twice continuously differentiable on Λ_n (as it is a quadratic function of λ). Then by compacity of Λ_n, $\bar{\lambda} := \arg\max_{\lambda\in\Lambda_n} \hat{P}(\tilde{\beta},\lambda)$ exists. Furthermore, by $\lambda_{\min}\left(\tilde{\Omega}+\alpha I\right) \ge \alpha$, the following inequalities hold

$$0 = \hat{P}\left(\tilde{\beta},0\right) \le \hat{P}\left(\tilde{\beta},\bar{\lambda}\right) = -\bar{\lambda}'\tilde{g} - \frac{1}{2}\bar{\lambda}'\left(\tilde{\Omega}+\alpha I\right)\bar{\lambda} \le \left\|\bar{\lambda}\right\|\|\tilde{g}\| - \alpha C\left\|\bar{\lambda}\right\|^2. \quad \text{(A.17)}$$

Adding $\alpha C\left\|\bar{\lambda}\right\|^2$ from both sides and dividing by $C\left\|\bar{\lambda}\right\|$ we find that $\alpha\left\|\bar{\lambda}\right\| \leqslant C\|\tilde{g}\|$. Then by the hypothesis in the statement of Lemma 3.7, $\left\|\bar{\lambda}\right\| = O_p\left(\alpha^{-1}a_n^{-1}\right) =$

$\delta_n O_p\left(\alpha^{-1}\delta_n^{-1}a_n^{-1}\right) = \delta_n o_p(1)$. Therefore $\lim_{n\to\infty} P\left(\left\|\bar{\lambda}\right\| < \delta_n\right) = 1$ and then $\bar{\lambda} \in \text{int}(\Lambda_n)$ w.p.a.1. It follows that $\bar{\lambda} = \arg\max_{\lambda\in\Lambda_n} \hat{P}(\tilde{\beta},\lambda)$ satisfies the first order conditions, $\partial\hat{P}\left(\tilde{\beta},\lambda\right)/\partial\lambda\big|_{\lambda=\bar{\lambda}} = 0$. By Lemma 3.6 $\bar{\lambda} \in \Lambda_n \subseteq \hat{\Lambda}(\tilde{\beta})$ w.p.a.1. Then by concavity of $\hat{P}\left(\tilde{\beta},\lambda\right)$ with respect of λ, and convexity of $\hat{\Lambda}(\tilde{\beta})$ it follows that $\hat{P}\left(\tilde{\beta},\bar{\lambda}\right) = \max_{\lambda\in\hat{\Lambda}(\tilde{\beta})}\hat{P}(\tilde{\beta},\lambda)$, giving the second and the third conclusions with $\tilde{\lambda} = \bar{\lambda}$. The last inequality of Eq. (A.17) gives

$$\hat{P}\left(\tilde{\beta},\tilde{\lambda}\right) \leq \left\|\tilde{\lambda}\right\|\|\tilde{g}\| - \alpha C\|\tilde{\lambda}\|^2 \leqslant O_p\left(\frac{1}{\alpha a_n}\right)O_p\left(\frac{1}{a_n}\right) - \alpha O_p\left(\frac{1}{\alpha^2 a_n^2}\right) = O_p\left(\alpha^{-1}a_n^{-2}\right),$$

giving the first result. □

Lemma 3.8. *If Assumptions 3.1(a) and 3.2(f) are satisfied, $\alpha n^{1/2-1/\gamma-\varepsilon} \to \infty$ as $n \to \infty$, where $\varepsilon > 0$ is such that $1/2 - 1/\gamma - \varepsilon > 0$, then for any $\bar{\lambda} \in \hat{\Lambda}\left(\hat{\beta}\right)$ it is the case that w.p.a.1 $\hat{P}\left(\hat{\beta},\bar{\lambda}\right) \leqslant \sup_{\lambda\in\hat{\Lambda}\left(\hat{\beta}\right)}\hat{P}\left(\hat{\beta},\lambda\right) = O_p\left(\alpha^{-1}n^{-1}\right)$.*

Proof. The inequality is obvious by the fact that $\bar{\lambda} \in \hat{\Lambda}\left(\hat{\beta}\right)$. Let $a_n = n^{1/2}$, then $\|\hat{g}(\beta_0)\| = O_p\left(a_n^{-1}\right)$ by Lemma 3.2. For $\delta_n = n^{-1/\gamma-\varepsilon}$, $\delta_n n^{1/\gamma} = n^{-\varepsilon} \to 0$, and $\alpha\delta_n a_n = \alpha n^{1/2-1/\gamma-\varepsilon} \to \infty$, so that the hypotheses in the statement of Lemma 3.7 is satisfied for $\tilde{\beta} = \beta_0$. The conclusion of Lemma 3.7 gives $\sup_{\lambda\in\hat{\Lambda}(\beta_0)}\hat{P}(\beta_0,\lambda) = O_p\left(\alpha^{-1}n^{-1}\right)$. By Theorem 3.1,

$$\sup_{\lambda\in\hat{\Lambda}(\hat{\beta})}\hat{P}(\hat{\beta},\lambda) \leq \sup_{\lambda\in\hat{\Lambda}(\beta_0)}\hat{P}(\beta_0,\lambda) = O_p\left(\alpha^{-1}n^{-1}\right),$$

giving the second result. □

We need the following notations for the next result. Let $g_i = g_i(\beta_0)$, $\Omega = E\left[g_i g_i'\right]$, $\hat{\Omega}(\beta) = \sum_{i=1}^n g_i(\beta)g_i(\beta)'/n$, $\hat{\Omega} = \hat{\Omega}(\hat{\beta})$, $\tilde{\Omega} = \sum_{i=1}^n g_i g_i'/n$ and $\bar{\Omega} = n^{-1}\sum_{i=1}^n \sigma_i^2 q_i q_i'$, with $\sigma_i = \sigma(z_i)$ and $\sigma(z)^2 := \left[\rho(w,\beta_0)^2|z\right]$

Lemma 3.9. *If Assumptions 3.1(a) and 3.2(b)-(e) are satisfied then for any $\hat{\beta} \in \mathscr{B}$*

(i) If $\alpha \to 0$ as $n \to 0$ then for n large enough $\lambda_{\max}\left(\hat{\Omega} + \alpha I\right) \leq C$ w.p.a.1;

(ii) If $\hat{\beta} = \beta_0 + O_p(\tau_n)$ *with* $\tau_n \to 0$, *then*

$$\left\|\hat{\Omega} - \tilde{\Omega}\right\| = O_p(\tau_n), \quad \left\|\tilde{\Omega} - \bar{\Omega}\right\| = O_p\left(n^{-1/2}\right), \text{ and } \left\|\bar{\Omega} - \Omega\right\| = O_p\left(n^{-1/2}\right).$$

Moreover $\lambda_{\max}(\Omega) \leq C$ *and w.p.a.1* $\lambda_{\max}\left(\bar{\Omega}\right) \leq C$, *and* $\lambda_{\max}\left(\hat{\Omega}\right) \leq C$. *If in addition* $\left\|\tilde{\lambda}\right\| = O_p(\kappa_n)$ *then for* $\check{\Omega} = -\sum_{i=1}^n s_1\left(\tilde{\lambda}' g_i\right) g_i g_i'/n$, *we have* $\left\|\check{\Omega} - \bar{\Omega}\right\| = O_p\left(\kappa_n + \tau_n + n^{-1/2}\right)$.

Proof. *(i)* Let $\hat{\rho}_i = \rho\left(w_i, \hat{\beta}\right)$ and $\rho_i = \rho(w_i, \beta_0)$. For $b_i = \sup_{\beta \in \mathscr{B}} \|\rho(w_i, \beta)\|$ we have $\hat{\Omega} \leqslant \sum_{i=1}^n b_i^2 q_i q_i'/n$
$\overset{def}{=} \dot{\Omega}$. Also, by Assumptions 3.1(a) and 3.2(d),

$$\begin{aligned} E\left[\left\|\dot{\Omega} - E\left[\dot{\Omega}\right]\right\|^2\right] &= E\left[\left\|\sum_{i=1}^n b_i^2 q_i q_i'/n - E\left[b_i^2 q_i q_i'\right]\right\|^2\right] \\ &= \operatorname{tr} E\left[\left(b_i^2 q_i q_i' - E\left[b_i^2 q_i q_i'\right]\right)^2\right]/n \\ &\leq \operatorname{tr} E\left[b_i^4 \left\{q_i q_i'\right\}^2\right]/n \\ &\leqslant \operatorname{tr} E\left[E\left[b_i^4 | z_i\right]\left\{q_i q_i'\right\}^2\right]/n \\ &\leqslant \operatorname{tr} E\left[\|q_i\|^4\right]/n \leqslant C/n. \end{aligned}$$

It follows by M that $\left\|\dot{\Omega} - E\left[\dot{\Omega}\right]\right\| = O_p(n^{-1/2})$. Also, by Assumptions 3.1(a) and 3.2(c),

$$\begin{aligned} \lambda_{\max}\left(E\left[\dot{\Omega}\right]\right) &= \lambda_{\max}\left(E\left[b_i^2 q_i q_i'\right]\right) \\ &= \lambda_{\max}\left(E\left[E\left[b_i^2 | z_i\right] q_i q_i'\right]\right) \\ &\leqslant C \operatorname{tr} E\left[q_i q_i'\right] \\ &\leq C E\left[\|q_i\|^2\right] \leq C. \end{aligned}$$

It follows that $\left\|\lambda_{\max}\left(\dot{\Omega}\right) - \lambda_{\max}\left(E\left[\dot{\Omega}\right]\right)\right\| \leqslant \left\|\dot{\Omega} - E\left[\dot{\Omega}\right]\right\| \xrightarrow{p} 0$ and therefore $\lambda_{\max}\left(\dot{\Omega}\right) \leq C$ w.p.a.1. Moreover $\alpha \to 0$ as $n \to \infty$, then $\alpha \leqslant C$ for n large enough; giving the result (i) by $\hat{\Omega} \leq \dot{\Omega}$.

(ii) Also, by $\hat{\beta} \xrightarrow{p} \beta_0$, $\hat{\beta} \in \mathcal{N}$ w.p.a.1 so that by Assumption 3.2(d) $|\hat{\rho}_i - \rho_i| \leq \delta_i \left\|\hat{\beta} - \beta_0\right\|$ for all $i = 1, \ldots, n$ w.p.a.1, where $\delta_i = \delta_2(w_i)$. $M_i = \delta_i^2 + 2\delta_i \|\rho_i\|$ has $E[M_i|z_i]$ bounded by CS and Assumption 3.2(d) so that $E\left[M_i \|q_i\|^2\right] = E\left[E\left[M_i|z_i\right]\|q_i\|^2\right] \leq C$. It follows by Assumption 3.2(b), CS and M that

$$\begin{aligned}
\left\|\hat{\Omega} - \tilde{\Omega}\right\| &= \left\| n^{-1} \sum_{i=1}^{n} \left(\hat{\rho}_i^2 - \rho_i^2\right) q_i q_i' \right\| \\
&\leqslant n^{-1} \sum_{i=1}^{n} \left|\hat{\rho}_i^2 - \rho_i^2\right| \|q_i\|^2 \\
&\leqslant n^{-1} \sum_{i=1}^{n} \left[(\hat{\rho}_i - \rho_i)^2 + 2|\hat{\rho}_i - \rho_i||\rho_i|\right] \|q_i\|^2 \\
&\leqslant n^{-1} \sum_{i=1}^{n} \left[\delta_i^2 \|\hat{\beta} - \beta_0\|^2 + 2\delta_i |\rho_i| \left\|\hat{\beta} - \beta_0\right\|\right] \|q_i\|^2 \\
&\leqslant \left\|\hat{\beta} - \beta_0\right\| \sum_{i=1}^{n} M_i \|q_i\|^2 / n \\
&\leq O_p(\tau_n) O_p\left(E\left[M_i \|q_i\|^2\right]\right) = O_p(\tau_n),
\end{aligned}$$

giving the first result in (ii).

Note that by Assumptions 3.1(a) and 3.2(d),

$$\begin{aligned}
E\left[\left\|\tilde{\Omega} - \bar{\Omega}\right\|^2\right] &= E\left[\left\|\sum_{i=1}^{n} \left(\rho_i^2 - \sigma_i^2\right) q_i q_i' / n\right\|^2\right] \\
&= \operatorname{tr} E\left[\left(\rho_i^2 - \sigma_i^2\right)^2 \left\{q_i q_i'\right\}^2\right] / n \\
&\leq \operatorname{tr} E\left[\rho_i^4 \left\{q_i q_i'\right\}^2\right] / n \\
&\leqslant \operatorname{tr} E\left[E\left[\rho_i^4 | z_i\right] \{q_i q_i'\}^2\right] / n \\
&\leq C E\left[\|q_i\|^4\right] / n \leq C / n,
\end{aligned}$$

so that the second result in (ii) follows by M.

The third result follows by M and

$$\begin{aligned} E\left[\left\|\bar{\Omega}-\Omega\right\|^2\right] &= E\left[\left\|\sum_{i=1}^n \sigma_i^2 q_i q_i'/n - E\left[\rho_i^2 q_i q_i'\right]\right\|^2\right] \\ &= \operatorname{tr} E\left[\left(\sigma_i^2 q_i q_i' - E\left[\rho_i^2 q_i q_i'\right]\right)^2\right]/n \\ &\leqslant \operatorname{tr} E\left[\sigma_i^4 \left\{q_i q_i'\right\}^2\right]/n \\ &\leq \operatorname{tr} E\left[E\left[\rho_i^2|z_i\right]^2 \left\{q_i q_i'\right\}^2\right]/n \\ &\leq C E\left[\|q_i\|^4\right]/n \leq C/n. \end{aligned}$$

As in the proof of (i), $\lambda_{\max}(\Omega) \leq C$ by Assumptions 3.1(a) and 3.2(c). Similarly to the proof of (i), it follows from $|\lambda_{\max}(A) - \lambda_{\max}(B)| \leq \|A-B\|$ that w.p.a.1 $\lambda_{\max}\left(\bar{\Omega}\right) \leq C$ and $\lambda_{\max}\left(\hat{\Omega}\right) \leq C$.

It remains to show that $\left\|\breve{\Omega}-\bar{\Omega}\right\| = O_p\left(\kappa_n + \tau_n + n^{-1/2}\right)$. Given previous results, it will be sufficient to show that $\left\|\breve{\Omega}-\hat{\Omega}\right\| = O_p\left(\kappa_n\right)$. Note that $s_1(v) = -(1+v)$ so that by Assumptions 3.1(a) and 3.2(d), and CS,

$$\begin{aligned} \left\|\breve{\Omega}-\hat{\Omega}\right\| &= \left\|n^{-1}\sum_{i=1}^n \left(\lambda'\hat{g}_i\right) b_i^2 q_i q_i'\right\| \\ &\leqslant \left(\sum_{i=1}^n \left|\tilde{\lambda}'\hat{g}_i\right|^2/n\right)^{1/2} \left(\sum_{i=1}^n b_i^4 \|q_i\|^4/n\right)^{1/2} \\ &\leqslant \sqrt{\tilde{\lambda}'\hat{\Omega}\tilde{\lambda}} \sqrt{\sum_{i=1}^n b_i^4 \left\|q_i\right\|^4/n} \\ &\leq \lambda_{\max}\left(\hat{\Omega}\right)^{1/2} \|\tilde{\lambda}\| O_p\left(\left\{E\left[E\left[b_i^4|z_i\right]\|q_i\|^4\right]\right\}^{1/2}\right) \\ &\leq C\|\tilde{\lambda}\| O_p(1) = O_p\left(\kappa_n\right). \end{aligned}$$

□

Lemma 3.10. *If Assumptions 3.1 and 3.2 are satisfied, $\alpha n^{1/2-1/\gamma-\varepsilon} \to \infty$ as $n \to \infty$, where $\varepsilon > 0$ is such that $1/2 - 1/\gamma - \varepsilon > 0$, then $\left\|\hat{g}\left(\hat{\beta}\right)\right\| = O_p\left((n\alpha)^{-1/2}\right)$.*

Proof. Let $\hat{\Omega} = \hat{\Omega}(\hat{\beta})$, $\hat{g} = \hat{g}(\hat{\beta})$, and Λ_n be as defined in Lemma 3.6. Also, let $\delta_n = n^{-1/\gamma-\varepsilon}$ and $\bar{\lambda} = -\delta_n \hat{g}/\|\hat{g}\|$ so that $\bar{\lambda}'\hat{g} = -\delta_n \|\hat{g}\|$ and $\|\bar{\lambda}\| = \delta_n$. Then $\bar{\lambda} \in \Lambda_n$ and by Lemma 3.6 $\bar{\lambda} \in \Lambda_n \subseteq \hat{\Lambda}(\widehat{\beta})$ w.p.a.1. Also, by Lemma 3.9(i), $\lambda_{\max}\left(\hat{\Omega} + \alpha I\right) \leq C$ so that Lemma 3.8 applied to $\bar{\lambda}$ gives

$$O_p(1/(\alpha n)) = \hat{P}(\hat{\beta}, \bar{\lambda}) = -\bar{\lambda}'\hat{g} - \frac{1}{2}\bar{\lambda}'\left(\hat{\Omega} + \alpha I\right)\bar{\lambda} \geqslant \delta_n \|\hat{g}\| - C\delta_n^2,$$

or equivalently $\delta_n \|\hat{g}\| - C\delta_n^2 \leq O_p(1/(n\alpha))$. Adding $C\delta_n^2$ on both sides and dividing by δ_n gives

$$\|\hat{g}\| \leq O_p\left(1/\left(n\alpha\delta_n\right)\right) + C\delta_n = \frac{1}{\alpha n^{\frac{1}{2}-\frac{1}{\gamma}-\varepsilon} n^{\frac{1}{2}}} O_p(1) + C\delta_n = o(1)O_p(1) + C\delta_n = O_p\left(\delta_n\right).$$

Now consider any $\varepsilon_n \to 0$. Let $\tilde{\lambda} = -\varepsilon_n \hat{g}$. Then $\|\tilde{\lambda}\| = |\varepsilon_n| \|\hat{g}\| = o(1)O_p\left(\delta_n\right)$ so that w.p.a.1. $\tilde{\lambda} \in \Lambda_n \subseteq \hat{\Lambda}(\hat{\beta})$ by Lemma 3.6. For n large enough,

$$\hat{P}(\hat{\beta}, \tilde{\lambda}) \geqslant -\tilde{\lambda}'\hat{g} - C\|\tilde{\lambda}\|^2 = \left(\varepsilon_n - C\varepsilon_n^2\right)\|\hat{g}\|^2 \geqslant \|\hat{g}\|^2 \varepsilon_n/2.$$

It then follows by Lemma 3.8 $\|\hat{g}\|^2 \varepsilon_n = O_p(1/(n\alpha))$. Since ε_n is any sequence converging to zero, it follows that $\|\hat{g}\|^2 = O_p(1/(n\alpha))$ giving the result. □

Proof of Theorem 3.3. By $\hat{A} = n^{-1}\sum_{i=1}^n \rho_i^2 q_i q_i'$ being positive semidefinite, $\hat{A} + \alpha I \geqslant \alpha I$ so that $\lambda_{\min}(\hat{A} + \alpha I) \geqslant \alpha$ and therefore $\lambda_{\max}\left(\tilde{W}\right) \leqslant 1/\alpha$ for $\tilde{W} = (\hat{A} + \alpha I)^{-1}$. By CS and Lemma 3.10, $\hat{R}(\hat{\beta}) = \hat{g}\tilde{W}\hat{g} \leqslant \alpha^{-1}\|\hat{g}\|^2 = O_p\left(\alpha^{-2}n^{-1}\right) \xrightarrow{p} 0$ since $\alpha^2 n = \left(\alpha n^{\frac{1}{2}-\frac{1}{\gamma}-\varepsilon}\right)^2 n^{\frac{2}{\gamma}+2\varepsilon} \to \infty$ as $n \to \infty$. It follows that $\hat{R}(\hat{\beta}) \xrightarrow{p} 0 = R(\beta_0)$. By Lemma 3.5, hypotheses in the statement of Lemma 3.1 are satisfied. The conclusion then follows from Lemma 3.1. □

We introduce the following notation for the next result. Let $\hat{g} = \hat{g}(\hat{\beta})$ and $\bar{g} = \hat{g}(\beta_0)$, where $\hat{g}(\beta) := n^{-1}\sum_{i=1}^n g_i(\beta)$ and $g_i(\beta) := q_i\rho(w_i, \beta)$.

Lemma 3.11. *If Assumptions 3.1, 3.2, and 3.3 are satisfied, $\alpha n^{1/2-1/\gamma-\varepsilon} \to \infty$ as $n \to \infty$, where $\varepsilon > 0$ is such that $1/2 - 1/\gamma - \varepsilon > 0$, then $\hat{\beta} = \beta_0 + O_p\left((\alpha\sqrt{n})^{-1}\right)$.*

Proof. By an expansion $\hat{g} = \bar{g} + \dot{G}\left(\hat{\beta} - \beta_0\right)$ for $\dot{G} = n^{-1}\sum_{i=1}^{n} q_i \rho_\beta\left(w_i, \bar{\beta}\right)$, where $\bar{\beta}$ is on the line joining $\hat{\beta}$ and β_0. Therefore,

$$\begin{aligned}\hat{R}(\hat{\beta}) := \hat{g}'\tilde{W}\hat{g} &= \left(\bar{g} + \dot{G}\left(\hat{\beta} - \beta_0\right)\right)'\tilde{W}\left(\bar{g} + \dot{G}\left(\hat{\beta} + \hat{\beta}_0\right)\right) \\ &= \hat{R}(\beta_0) + 2\bar{g}'\tilde{W}\dot{G}\left(\hat{\beta} - \beta_0\right) + \hat{D}^2\end{aligned}$$

where $\hat{D} = \left[\left(\hat{\beta} - \beta_0\right)\dot{G}'\tilde{W}\dot{G}\left(\hat{\beta} - \beta_0\right)\right]^{1/2}$. Then for $\hat{F} = \left[\hat{R}(\hat{\beta}) + \hat{R}(\beta_0)\right]^{1/2}$, it follows from T, CS and $\hat{R}(\beta_0)^{1/2} \leq \hat{F}, \hat{D} \geqslant 0$ that

$$\begin{aligned}\hat{D}^2 &= \hat{R}(\hat{\beta}) - \hat{R}(\beta_0) - 2\bar{g}'\tilde{W}\dot{G}\left(\hat{\beta} - \beta_0\right) = \left|\hat{R}(\hat{\beta}) - \hat{R}(\beta_0) - 2\bar{g}'\tilde{W}\dot{G}\left(\hat{\beta} - \beta_0\right)\right| \\ &\leq \hat{R}(\hat{\beta}) + \hat{R}(\beta_0) + 2\left|\bar{g}'\tilde{W}\dot{G}\left(\hat{\beta} - \beta_0\right)\right| \leq \hat{R}(\hat{\beta}) + \hat{R}(\beta_0) + 2\hat{R}(\beta_0)^{1/2}\hat{D} \leq \hat{F}^2 + 2\hat{F}\hat{D}.\end{aligned}$$

Subtracting $2\hat{F}\hat{D}$ from both sides, adding $\hat{F}^2$ to both sides, and then taking square roots gives $|\hat{D} - \hat{F}| \leq \sqrt{2}\hat{F}$. Also, by T, $|\hat{D} - \hat{F}| \geqslant \hat{D} - \hat{F}$, so that $\hat{D} \leq (\sqrt{2}+1)\hat{F} = C\hat{F}$. By Lemma 3.2, $\|\bar{g}\| = O_p\left(n^{-1/2}\right)$ and by Lemma 3.10 $\|\hat{g}\| = O_p\left((n\alpha)^{-1/2}\right)$. Also, as in the proof of Theorem 3.3 $\lambda_{\max}\left(\tilde{W}\right) \leq 1/\alpha$ w.p.a.1 so that by T

$$\hat{F}^2 \leq \hat{R}(\hat{\beta}) + \hat{R}(\beta_0) \leq \frac{1}{\alpha}\left(\|\hat{g}\|^2 + \|\bar{g}\|^2\right) \leq \frac{1}{\alpha}\left(O_p(1/n) + O_p\left(1/(\alpha n)\right)\right) = O_p\left(1/\alpha^2 n\right).$$

Also, Lemma 3.4 applied to $\hat{\beta} = \bar{\beta}, \bar{\beta} = \beta_0, U(z) = 1$, $a(w,\beta) = \partial\rho(w,\beta)/\partial\beta_k$, and $b(w,\beta) = \partial\rho(w,\beta)/\partial\beta_l$ for $k, l = 1, \ldots, p$ gives $\left(\dot{G}'\tilde{W}\dot{G}\right)_{kl} \xrightarrow{p} E\left[(D(z)'D(z))_{kl}\right]$ for all k, l so that $\dot{G}'\tilde{W}\dot{G} \xrightarrow{p} E\left[D(z)'D(z)\right]$ which is non singular by Assumption 3.3(d). It then follows that $\lambda_{\min}\left(\dot{G}'\tilde{W}\dot{G}\right) \geqslant C$ w.p.a.1 and then $\hat{D}^2 = \left(\hat{\beta} - \beta_0\right)\dot{G}'\tilde{W}\dot{G}\left(\hat{\beta} - \beta_0\right) \geq C\left\|\hat{\beta} - \beta_0\right\|$. Therefore, $C\left\|\hat{\beta} - \beta_0\right\|^2 \leq \hat{D}^2 \leq C\hat{F}^2 = O_p\left(1/(n\alpha^2)\right)$, giving the result. □

Some useful notations are needed for the next result. Let $D_i = D(z_i)$, $\hat{G} = n^{-1}\sum_{i=1}^{n} q_i \rho_\beta\left(w_i, \hat{\beta}\right)$, $\bar{G} = n^{-1}\sum_{i=1}^{n} q_i D_i$, $G = E\left[q_i D_i\right]$, and $\tilde{G} = n^{-1}\sum_{i=1}^{n} q_i \rho_\beta\left(\omega_i, \beta_0\right)$.

Lemma 3.12. *If Assumptions 3.1, 3.2(b)-(e), and 3.3(b)-(c) are satisfied and* $\hat{\beta} = \beta_0 + O_p(\tau_n)$ *with* $\tau_n \to 0$, *then*

(i) $\left\|\hat{G} - \bar{G}\right\| = O_p\left(\tau_n + n^{-1/2}\right)$ *and* $\left\|\bar{G} - G\right\| = O_p\left(n^{-1/2}\right)$;

(ii) If in addition $\alpha \to 0$ *and* $\alpha n \to \infty$ *as* $n \to \infty$, *then* $\left\|\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left(\widehat{\Omega}-\bar{\Omega}\right)\right\| = O_p\left(\tau_n + n^{-1/2}\right)$;

(iii) For $\|\tilde{\lambda}\| = O_p(\kappa_n)$ *and* $\breve{G} = -n^{-1}\sum_{i=1}^n s_1\left(\tilde{\lambda}'\hat{g}_i\right)\partial g_i(\hat{\beta})/\partial\beta'$ *then* $\left\|\breve{G}-\hat{G}\right\| = O_p(\kappa_n)$ *and* $\left\|\breve{G}-\bar{G}\right\| = O_p\left(\kappa_n + \tau_n + n^{-1/2}\right)$.

Proof. Let $\rho_{\beta i} = \rho_\beta(w_i, \beta_0)$, then $E\left[\rho_{\beta i}\right] = E[D_i]$ by iterated expectation. Also, by Assumption 3.3(c)

$$\begin{aligned} E\left[\left\|\tilde{G}-\bar{G}\right\|^2\right] &= E\left[\left\|n^{-1}\sum_{i=1}^n q_i\left(\rho_{\beta_i}-D_i\right)\right\|^2\right] \\ &= n^{-1}\operatorname{tr} E\left[\left(\rho_{\beta_i}-D_i\right)'\left(\rho_{\beta i}-D_i\right)q_i'q_i\right] \\ &\le n^{-1}\operatorname{tr} E\left\{E\left[\rho_{\beta i}'\rho_{\beta i}|z_i\right]\|q_i\|^2\right\} \\ &\le n^{-1}CE\left[E\left[\left\|\rho_{\beta i}\right\|^2|z_i\right]\|q_i\|^2\right] \le C/n. \end{aligned}$$

It follows by M that $\left\|\tilde{G}-\bar{G}\right\| = O_p\left(n^{-1/2}\right)$.

Also, by the mean value theorem for vector-valued functions, $\left\|\rho_\beta(w,\beta)-\rho_\beta(w,\beta_0)\right\| \le \delta_3(w)\|\beta-\beta_0\|$, for all $\beta \in \mathcal{N}$, where $\delta_3(w) = \sup_{\beta\in\mathcal{N}}\left\|\rho_{\beta\beta}(w,\beta)\right\|$ with $E[\delta_3(w)]$ bounded by Assumption 3.3(c). For $\hat{\rho}_{\beta i} = \rho_\beta\left(w_i, \hat{\beta}\right)$, it follows from T, CS, and M that

$$\begin{aligned} \left\|\hat{G}-\tilde{G}\right\| &= \left\|n^{-1}\sum_{i=1}^n q_i\left(\hat{\rho}_{\beta i}-\rho_{\beta i}\right)\right\| \\ &\le n^{-1}\sum_{i=1}^n\left\|\hat{\rho}_{\beta i}-\rho_{\beta i}\right\|\|q_i\| \\ &\le C\left\|\hat{\beta}-\beta_0\right\|\sum_{i=1}^n\delta_3(w_i)/n = O_p(\tau_n)O_p(E[\delta_3(w_i)]) \le O_p(\tau_n). \end{aligned}$$

It then follows by T that $\left\|\hat{G}-\bar{G}\right\| \leqslant \left\|\hat{G}-\tilde{G}\right\| + \left\|\tilde{G}-\bar{G}\right\| = O_p\left(\tau_n + n^{-1/2}\right)$, giving the first result in (i).

Also, by the Jensen's Inequality and Assumption 3.3(c), $\|D_i\|^2 = \left\|E\left[\rho_{\beta i}|z_i\right]\right\|^2 \leq E\left[\left\|\rho_{\beta i}\right\|^2|z_i\right] \leqslant C$. The second conclusion in (i), then follows by M and

$$E\left[\left\|\bar{G}-G\right\|^2\right] = E\left[\left\|n^{-1}\sum_{i=1}^{n} q_i D_i - G\right\|^2\right] \leq n^{-1}E\left[\|D_i\|^2\|q_i\|^2\right] \leq C/n.$$

For the proof of (ii) let $D(z) = \left[D^1(z), \cdots, D^p(z)\right]$ where $D^k(z) = E\left[\partial\rho\left(w,\beta_0\right)/\partial\beta_k|z\right]$. Then by Assumptions 3.2(e) and 3.3(c), the hypothesis in the statement of Lemma 3.4 are satisfied for $a(w,\beta) = D^k(z), b(w,\beta) = D^l(z)$ and $U(z) = \sigma(z)^2$, for $k,l = 1,\cdots,p$. It follows by the conclusion of Lemma 3.4 that $\left(\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{G}\right)_{kl} \xrightarrow{p} E\left[\sigma(z)^{-2}\left(D(z)'D(z)\right)_{kl}\right]$ for all $k,l = 1,\ldots,p$ so that $\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{G} \xrightarrow{p} E\left[\sigma(z)^{-2}D(z)'D(z)\right]$ and hence $\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{G} = O_p(1)$. Let $H_i = \bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}q_i$. Then by Σ_i bounded away from zero and $\bar{\Omega} \leq \bar{\Omega}+\alpha I$,

$$\begin{aligned}\sum_{i=1}^{n}\|H_i\|^2/n &= \operatorname{tr}\left(\sum_{i=1}^{n}H_iH_i'/n\right) = \operatorname{tr}\left(\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}q_iq_i'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{G}\right)\\ &\leq \operatorname{tr}\left(\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{\Omega}\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{G}\right)\\ &\leq \operatorname{tr}\left(\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{G}\right) = O_p(1).\end{aligned}$$

Next, let $M_i = \delta_i^2 + 2\delta_i|\rho_i|$, where $\delta_i = \delta_2(w_i)$ and $\rho_i = \rho(w_i,\beta_0)$. Also, let $Z = (z_1,\cdots,z_n)$. It is well known that if $E\left[\hat{R}_n|Z\right] = O_p(\nu_n)$ for some ν_n then $\hat{R}_n = O_p(\nu_n)$. For $\hat{R}_n = \sum_{i=1}^{n}M_i\|H_i\|\|q_i\|/n$, it follows from CS and M that

$$\begin{aligned}E\left[\hat{R}_n|Z\right] &= \sum_{i=1}^{n}E\left[M_i|Z\right]\|H_i\|\,\|q_i\|/n \leq C\sum_{i=1}^{n}\|H_i\|\,\|q_i\|/n\\ &\leq C\left(\sum_{i=1}^{n}\|H_i\|^2/n\right)^{1/2}\left(\sum_{i=1}^{n}\|q_i\|^2/n\right)^{1/2} = O_p(1),\end{aligned}$$

so that $\hat{R}_n = O_p(1)$. Therefore by Assumption 3.2(d), CS and T

$$\begin{aligned}
\left\|\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left(\hat{\Omega}-\tilde{\Omega}\right)\right\| &= \left\|n^{-1}\sum_{i=1}^{n}\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left\{\left(\hat{\rho}_i^2-\rho_i^2\right)q_i q_i'\right\}\right\| \\
&= \left\|n^{-1}\sum_{i=1}^{n}\left(\hat{\rho}_i^2-\rho_i^2\right)H_i q_i'/\zeta(K)^2\right\| \\
&\leq n^{-1}\sum_{i=1}^{n}\left|\hat{\rho}_i^2-\rho_i^2\right|\|H_i\|\,\|q_i\| \\
&\leqslant \left\|\hat{\beta}-\beta_0\right\|\hat{R}_n = O_p(\tau_n).
\end{aligned}$$

Also by Assumptions 3.1(a) and 3.2(d)

$$\begin{aligned}
E\left[\left\|\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left(\tilde{\Omega}-\bar{\Omega}\right)\right\|^2\middle|Z\right] &= E\left[\left\|\sum_{i=1}^{n}\left(\rho_i^2-\sigma_i^2\right)H_i q_i'/n\right\|^2\middle|Z\right] \\
&= \frac{1}{n^2}\operatorname{tr}E\left[\sum_{i=1}^{n}\left(\rho_i^2-\sigma_i^2\right)^2 H_i q_i' q_i H_i\middle|Z\right] \\
&\leq \frac{1}{n^2}E\left[\sum_{i=1}^{n}E\left[\rho_i^4|z_i\right]\operatorname{tr}\left\{H_i\|q_i\|^2 H_i'\right\}\right] \\
&\leq \frac{C}{n}\sum_{i=1}^{n}\|H_i\|^2/n = O_p(1/n),
\end{aligned}$$

so that $\left\|\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left(\tilde{\Omega}-\bar{\Omega}\right)\right\| = O_p\left(n^{-1/2}\right)$. The conclusion in (ii) follows by T, that is,
$\left\|\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left(\hat{\Omega}-\bar{\Omega}\right)\right\| = O_p\left(\tau_n+n^{-1/2}\right)$.

To prove (iii), note that $s_1(v) = -(1+v)$ so that $\breve{G} = \sum_{i=1}^{n}\left(1+\tilde{\lambda}'\hat{g}_i\right)q_i\hat{\rho}_{\beta i}/n$,

where $\hat{\rho}_{\beta i} = \rho_\beta\left(w_i, \hat{\beta}\right)$. Let $b_i = \sup_{\beta\in\mathcal{N}} \left\|\rho_\beta\left(w_i,\beta\right)\right\|$, then by 3.3(c), T, CS, and M

$$\begin{aligned}
\left\|\check{G} - \hat{G}\right\| &= \left\| n^{-1}\sum_{i=1}^{n}\left(\tilde{\lambda}'\hat{g}_i\right) q_i\hat{\rho}_{\beta i}\right\| \\
&\leqslant \sum_{i=1}^{n}\left|\tilde{\lambda}'\hat{g}_i\right| |b_i| \left\|q_i\right\| /n \\
&\leqslant \sqrt{\sum_{i=1}^{n}\left|\tilde{\lambda}'\hat{g}_i\right|^2 /n}\sqrt{\sum_{i=1}^{n} b_i^2 \left\|q_i\right\|^2 /n} \\
&\leqslant \sqrt{\tilde{\lambda}'\hat{\Omega}\tilde{\lambda}}\sqrt{\sum_{i=1}^{n} b_i^2 \left\|q_i\right\|^2 /n} \\
&\leq C\|\tilde{\lambda}\| O_p\left(\left\{E\left[E\left[b_i^2|z_i\right]\|q_i\|^2\right]\right\}^{1/2}\right) = O_p\left(\kappa_n\right),
\end{aligned}$$

giving the first result in (iii). The second result in (iii) follows by T. □

Lemma 3.13. *If Assumption 3.1 is satisfied, ε_i, Y_i are random variables with $E\left[\varepsilon_i|z_i\right] = 0$, $E\left[\left\|\varepsilon_i\right\|^2|z_i\right] \leq C$, $E\left[\left\|Y_i\right\|^2|z_i\right] \leq C$, $U_i = U(z_i)$ is a nonnegative scalar function that is bounded away from zero, $K \to \infty, \alpha \to 0, \alpha\sqrt{n} \to \infty$ then,*

$$\left(\frac{1}{n}\sum_{i=1}^{n} q_i Y_i\right)'\left(\frac{1}{n}\sum_{i=1}^{n} U_i q_i q_i' + \alpha I\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_i\varepsilon_i\right) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} E\left[Y_i|z_i\right]' U_i^{-1}\varepsilon_i \xrightarrow{p} 0.$$

Proof. Let P and Q^α be as defined in the proof of Lemma 3.4, $A_i = U_i^{-1/2}Y_i$, $\bar{A}_i = E\left[A_i|z_i\right] = U_i^{-1/2}E\left[Y_i|z_i\right]$, $A = (A_1,\cdots,A_n)'$, $\bar{A} = \left(\bar{A}_1,\cdots,\bar{A}_n\right)'$, $B_i = U_i^{-1/2}\varepsilon_i$, and $B = (B_1,\ldots,B_n)'$. Then, similarly to the proof of Lemma 3.4

$$\begin{aligned}
&\left(\frac{1}{n}\sum_{i=1}^{n} q_i Y_i\right)'\left(\frac{1}{n}\sum_{i=1}^{n} U_i q_i q_i' + \alpha I\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_i\varepsilon_i\right) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} E\left[Y_i|z_i\right]' U_i^{-1}\varepsilon_i \\
&= A'Q^\alpha B/\sqrt{n} - \bar{A}'B/\sqrt{n} = (A-\bar{A})'Q^\alpha B/\sqrt{n} - \bar{A}'\left(I-Q^\alpha\right)B/\sqrt{n}.
\end{aligned}$$

It follows as in the proof of Lemma 3.4 that $(A-\bar{A})'Q^\alpha(A-\bar{A}) = O_p\left(\operatorname{tr}\left(Q^\alpha\right)\right) = O_p(1/\alpha)$. Also, by $\bar{B}_i := E\left[B_i|z_i\right] = U_i^{-1/2}E\left[\varepsilon_i|z_i\right] = 0$ for all i, $B'Q^\alpha B = (B-\bar{B})'Q^\alpha(B-\bar{B}) =$

$O_p(1/\alpha)$ as in the proof of Lemma 3.4. It follows by CS that

$$\left|(A-\bar{A})'Q^\alpha B/\sqrt{n}\right| \leq \sqrt{(A-\bar{A})'Q^\alpha(A-\bar{A})}\sqrt{B'Q^\alpha B}/\sqrt{n} = O_p(1/\alpha\sqrt{n}) \xrightarrow{p} 0$$

Also, as in Lemma 3.4, $E\left[\bar{A}'(I-Q)\bar{A}\right]/n \to 0$, so by the law of iterated expectations, and Lemma 3.3,

$$\begin{aligned} E\left[\left|\bar{A}'(I-Q^\alpha)B/\sqrt{n}\right|^2\right] &= E\left[\bar{A}'(I-Q^\alpha)E\left[BB' \mid Z\right](I-Q^\alpha)\bar{A}\right]/n \\ &= E\left[\bar{A}'(I-Q^\alpha)U_i^{-1}E\left[\varepsilon_i^2|z_i\right](I-Q^\alpha)\bar{A}\right]/n \\ &\leq CE\left[\bar{A}'(I-Q^\alpha)^2\bar{A}\right]/n \\ &\leq CE\left[\bar{A}'(I-Q^\alpha)\bar{A}\right]/n \to 0. \end{aligned}$$

The conclusion follows by M and T. □

Proof of Theorem 3.4. Let $a_n = (n\alpha)^{1/2}$, then $\|g(\hat{\beta})\| = O_p(a_n^{-1})$ by Lemma 3.10. For $\delta_n = n^{-1/\gamma-\varepsilon}$, $\delta_n n^{1/\gamma} \to 0$ and $\alpha\delta_n a_n = \alpha^{3/2}n^{1/2-1/\gamma-\varepsilon} \to \infty$ as $n \to \infty$. It follows that the hypotheses of Lemma 3.7 are satisfied with $\tilde{\beta} = \hat{\beta}$ so that $\hat{\lambda} = \arg\max_{\lambda\in\hat{\Lambda}(\hat{\beta})} \hat{P}(\hat{\beta},\lambda)$ exists w.p.a.1, and $\|\hat{\lambda}\| = O_p\left(\alpha^{-3/2}n^{-1/2}\right)$.. It follows that $\|\hat{\lambda}\|/\delta_n = O_p\left(\delta_n^{-1}\alpha^{-3/2}n^{-1/2}\right) \xrightarrow{p} 0$ by $\delta_n\alpha^{3/2}n^{1/2} = \alpha^{3/2}n^{1/2-1/\gamma-\varepsilon} \to \infty$ as $n \to \infty$. It follows that $\hat{\lambda} \in \Lambda_n$ w.p.a.1 so that $\max_{1\leq i\leq n}\left|\hat{\lambda}'\hat{g}_i\right|$
$\xrightarrow{p} 0$ by Lemma 3.6.

Also, from the consistency of $\hat{\beta}$ (Theorem 3.3), $\hat{\beta}$ will be an element of $int(\mathscr{B})$. It follows by Assumption 3.3(b) that w.p.a.1, $\hat{P}(\beta,\lambda)$ is twice continuously differentiable in a neighborhood of $(\hat{\beta},\hat{\lambda})$. Then by the first-order condition for $\hat{\lambda}$, $\partial\hat{P}(\hat{\beta},\lambda)/\partial\lambda\big|_{\lambda=\hat{\lambda}} = 0$. Also, by the implicit function theorem, for all β in a neighborhood of $\hat{\beta}$ there is $\hat{\lambda}(\beta)$ such that $\partial\hat{P}(\beta,\lambda)/\partial\lambda\big|_{\lambda=\hat{\lambda}(\beta)} = 0$ and $\hat{\lambda}(\beta)$ is continuously differentiable in β with $\hat{\lambda}(\hat{\beta}) = \hat{\lambda}$. By concavity of $\hat{P}(\beta,\lambda)$ it reaches it maximum at $\hat{\lambda}(\beta)$ holding β fixed. Then the first order conditions for $\hat{\beta}$ and the envelope theorem give $0 = \partial\hat{P}(\beta,\lambda(\beta))/\partial\beta\big|_{\beta=\hat{\beta}} = \partial\hat{P}(\hat{\beta},\hat{\lambda})/\partial\beta = \breve{G}'\hat{\lambda}$, with $\breve{G} = -n^{-1}\sum_{i=1}^n s_1\left(\hat{\lambda}'\hat{g}_i\right)g_i(\hat{\beta})/\partial\beta'$, where $s_1(v) = -(1+v)$. Also, by $\hat{P}(\hat{\beta},\lambda) =$

$-\hat{g}'\lambda - \frac{1}{2}\lambda'\left(\hat{\Omega}+\alpha I\right)\lambda$, FOC for $\hat{\lambda}$ implies $0 = -\hat{g} - \left(\hat{\Omega}+\alpha I\right)\hat{\lambda} = 0$ so that $\hat{\lambda} = -\left(\hat{\Omega}+\alpha I\right)^{-1}\hat{g}$. Plugging the equation for $\hat{\lambda}$ in the first order conditions for $\hat{\beta}$ gives $\check{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\hat{g} = 0$.

Expanding $\hat{g}$ around β_0 gives, for a mean value $\dot{\beta}$, $\dot{G} = n^{-1}\sum_{i=1}^{n} q_i \rho_\beta\left(w_i, \dot{\beta}\right)$, and $\bar{g} = \hat{g}\left(\beta_0\right)$

$$\check{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\dot{G}\left(\hat{\beta}-\beta_0\right) + \dot{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g} = 0. \tag{A.18}$$

Note by Assumptions 3.2(e) and 3.3(d) that $E\left[D(z)'\sigma(z)^{-2}D(z)\right] \geqslant CE\left[D(z)'D(z)\right]$ so that
$V = \left\{E\left[D(z)'\sigma(z)^{-2}D(z)\right]\right\}^{-1}$ exists. Now successively apply Lemma 3.4 with $\theta = \left(\beta', \lambda'\right)'$, $\bar{\theta} = \left(\beta_0', 0'\right)'$, $\hat{\theta} = \left(\hat{\beta}', \hat{\lambda}'\right)'$, $a\left(w_i, \theta\right) = s_1\left(\lambda' g_i(\beta)\right)\partial\rho\left(w, \beta\right)/\partial\beta_r$, $b\left(w_i, \theta\right) = \partial\rho\left(w_i, \beta\right)/\partial\beta_s$, and $U(z) = \sigma(z)^2 = E\left[\rho\left(w, \beta_0\right)^2 | z\right]$ for $r, s = 1, \ldots, p$, we obtain $\check{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\dot{G} \xrightarrow{p} V^{-1}$. Also, by Lemma 3.11 $\hat{\beta} = \beta_0 + O_p\left(\tau_n\right)$ with $\tau_n = \alpha^{-1}n^{-1/2}$ so that the conclusion of Lemma 3.9(ii) gives $\left\|\hat{\Omega}-\bar{\Omega}\right\| = O_p\left(\tau_n + n^{-1/2}\right) = O_p\left(\alpha^{-1}n^{-1/2}\right) = o_p(1)$ by $\alpha n^{1/2} \to \infty$. It follows that

$$\check{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\dot{G} = \check{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\dot{G} + o_p(1) \xrightarrow{p} V^{-1}. \tag{A.19}$$

Also as previously justified, $\|\hat{\lambda}\| = O_p\left(\kappa_n\right)$, where $\kappa_n = \alpha^{-3/2}n^{-1/2}$. By Lemma 3.12(ii)&(iii) applied to $\tilde{\lambda} = \hat{\lambda}$, we have $\left\|\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left(\hat{\Omega}-\bar{\Omega}\right)\right\| = O_p\left(\tau_n + n^{-1/2}\right) = O_p\left(\alpha^{-1}n^{-1/2}\right)$ and $\left\|\check{G}-\bar{G}\right\| = O_p\left(\kappa_n + \tau_n + n^{-1/2}\right) = O_p\left(\alpha^{-3/2}n^{-1/2}\right)$ so that $\hat{A} \overset{def}{=} \left\|\check{G}-\bar{G}\right\| + \left\|\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left(\hat{\Omega}-\bar{\Omega}\right)\right\| = O_p\left(\alpha^{-3/2}n^{-1/2}\right)$. By $\hat{\Omega}$ p.s.d. $\lambda_{\min}\left(\hat{\Omega}+\alpha I\right) \geq \alpha$ and therefore $\lambda_{\max}\left\{\left(\hat{\Omega}+\alpha I\right)^{-1}\right\} \leq 1/\alpha$. By Lemma 3.2 $\|\bar{g}\| = O_p\left(n^{-1/2}\right)$ so that by CS

$$\begin{aligned}\left\|\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g}\right\| &= \left\{\bar{g}'\left(\hat{\Omega}+\alpha I\right)^{-1}\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g}\right\}^{1/2} \\ &\leq \frac{1}{\alpha}\left\{\bar{g}'\bar{g}\right\}^{1/2} = \frac{1}{\alpha}\|\bar{g}\| = O_p\left(\alpha^{-1}n^{-1/2}\right).\end{aligned}$$

It follows that $\hat{A}\left\|\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g}\right\| = b_n O_p\left(n^{-1/2}\right)$ with $b_n = \alpha^{-5/2}n^{-1/2} = o(1)$ by the

hypothesis in Theorem 3.4, so that $\hat{A}\left\|\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g}\right\| = o_p(n^{-1/2})$. Then by T and CS

$$\begin{aligned}
&\left\|\check{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g} - \bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{g}\right\| \\
&= \left\|\left(\check{G}-\bar{G}+\bar{G}\right)'\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g} - \bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{g}\right\| \\
&\leq \left\|\left(\check{G}-\bar{G}\right)\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g}\right\| + \left\|\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left[\bar{\Omega}-\hat{\Omega}\right]\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g}\right\| \\
&\leqslant \hat{A}\left\|\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g}\right\| = o_p(1/\sqrt{n}).
\end{aligned}$$

It follows that $\bar{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g} = \bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{g} + o_p(1/\sqrt{n})$.

Furthermore, Lemma 3.13 applied to $Y_i' = D(z_i) \stackrel{def}{=} D_i$, $\varepsilon_i = \rho(w_i,\beta_0) \stackrel{def}{=} \rho_i$, and $U_i = \sigma_i^2 \stackrel{def}{=} \sigma(z_i)^2$, leads to $\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{g} - n^{-1}\sum_{i=1}^n D_i'\sigma_i^2\rho_i = o_p(1/\sqrt{n})$. Also, by the Lindbergh-Levy central limit theorem $\sum_{i=1}^n D_i'\sigma_i^{-2}\rho_i/\sqrt{n} \rightarrow \stackrel{d}{\rightarrow} N(0,\Lambda)$, where, by iterated expectation, $\Lambda \stackrel{def}{=} E\left[D_i'\sigma_i^{-2}\rho_i^2\sigma_i^{-2}D_i\right] = E\left[D_i'\sigma_i^{-2}D_i\right] = V^{-1}$. Therefore,

$$\begin{aligned}
\sqrt{n}\bar{G}'(\bar{\Omega}+\alpha I)^{-1}\bar{g} &= \sqrt{n}\left(\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{g} - \sum_{i=1}^n D_i'\sigma_i^{-2}\rho_i/n\right) + \sum_{i=1}^n D_i'\sigma_i^{-2}\rho_i/\sqrt{n} \\
&= o_p(1) + \sum_{i=1}^n D_i'\sigma_i^{-2}\rho_i/\sqrt{n} \stackrel{d}{\rightarrow} \mathcal{N}\left(0, V^{-1}\right),
\end{aligned}$$

and thus

$$\sqrt{n}\check{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g} = \sqrt{n}\bar{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\bar{g} + o_p(1) \stackrel{d}{\rightarrow} \mathcal{N}\left(0, V^{-1}\right) \tag{A.20}$$

By Eq. (A.18)

$$\sqrt{n}(\hat{\beta}-\beta_0) = -\left[\check{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\dot{G}\right]^{-1}\sqrt{n}\check{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\bar{g},$$

so that by Eq. (A.19), Eq. (A.20), the Slutzky's theorem, and the continuous mapping theorem, $\sqrt{n}(\hat{\beta}-\beta_0) \stackrel{d}{\rightarrow} \mathcal{N}(0,V)$, giving the first result.

We now establish the consistency of the variance estimator. First, applying Lemma 3.4 gives $\check{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\check{G}$

$\xrightarrow{p} V^{-1}$. Also, by $\bar{\Omega}$ p.s.d. $\lambda_{\max}\left[\left(\bar{\Omega}+\alpha I\right)^{-1}\right] \leqslant 1/\alpha$ so that for $\hat{B} = \left(\bar{\Omega}+\alpha I\right)^{-1}\check{G}$, it follows from CS that

$$\begin{aligned}\|\hat{B}\|^2 = \operatorname{tr}\left(\hat{B}'\hat{B}\right) &= \operatorname{tr}\left(\check{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\left(\bar{\Omega}+\alpha I\right)^{-1}\check{G}\right) \\ &\leq \frac{1}{\alpha}\operatorname{tr}\left(\check{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\check{G}\right) \leq O_p(1/\alpha).\end{aligned}$$

Also by Lemma 3.9 applied to $\kappa_n = \alpha^{-3/2}n^{-1/2}$ and $\tau_n = \alpha^{-1}n^{-1/2}$ we have $\left\|\check{\Omega}-\bar{\Omega}\right\| = O_p\left(\kappa_n+\tau_n+n^{-1/2}\right) = O_p\left(\alpha^{-3/2}n^{-1/2}\right)$ for $\check{\Omega} = -\sum_{i=1}^n s_1\left(\hat{\lambda}'\hat{g}_i\right)\hat{g}_i\hat{g}_i'/n$. By T, CS, $\lambda_{\max}\left\{(\check{\Omega}+\alpha I)^{-1}\right\} \leq 1/\alpha$, and $A^{-1}-B^{-1} = B^{-1}\left[(B-A)+(B-A)A^{-1}(B-A)\right]B^{-1}$,

$$\begin{aligned}\left\|\check{G}'\left(\check{\Omega}+\alpha I\right)^{-1}\check{G} - \check{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\check{G}\right\| &= \left\|\check{G}'\left[\left(\check{\Omega}+\alpha I\right)^{-1}-\left(\bar{\Omega}+\alpha I\right)^{-1}\right]\check{G}\right\| \\ &= \left\|\hat{B}'\left\{\bar{\Omega}-\check{\Omega}+\left(\bar{\Omega}-\check{\Omega}\right)\left(\check{\Omega}+\alpha I\right)^{-1}\left(\bar{\Omega}-\check{\Omega}\right)\right\}\hat{B}\right\| \\ &\leq \|\hat{B}\|^2\left(\left\|\bar{\Omega}-\check{\Omega}\right\| + \alpha^{-1}\left\|\bar{\Omega}-\check{\Omega}\right\|^2\right) \\ &\leq C\alpha^{-1}\left(O_p\left(\alpha^{-3/2}n^{-1/2}\right)+\alpha^{-1}O_p\left(\alpha^{-3}n^{-1}\right)\right) \\ &\leq O_p\left(\alpha^{-5/2}n^{-1/2}\right)+O_p\left(\left(\alpha^{-5/2}n^{-1/2}\right)^2\right) \xrightarrow{p} 0.\end{aligned}$$

It follows that $\check{G}'(\check{\Omega}+\alpha I)^{-1}\check{G} = \check{G}'\left(\bar{\Omega}+\alpha I\right)^{-1}\check{G} + o_p(1) \xrightarrow{p} V^{-1}$. Also, by $s_1(v) = -(1+v)$, we have $\left|1+\sum_{i=1}^n s_1\left(\hat{v}_i\right)/n\right| = \left|\sum_{i=1}^n \hat{v}_i/n\right| \leq \max_{1\leq i\leq n}\left|\hat{\lambda}'\hat{g}_i\right| \xrightarrow{p} 0$ so that $\sum_{i=1}^n s_1\left(\hat{v}_i\right)/n \xrightarrow{p} -1$. Note that $\check{\Omega} = -n^{-1}\sum_{i=1}^n s_1\left(\hat{v}_i\right)\hat{\Omega}$ and $\check{G} = -n^{-1}\sum_{i=1}^n s_1\left(\hat{v}_i\right)\hat{G}$ so that $\hat{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\hat{G} = -\left(n\Big/\sum_{i=1}^n s_1\left(\hat{v}_i\right)\right)\check{G}'\left(\check{\Omega}+\alpha I\right)^{-1}\check{G}$. It follows by continuous mapping that,

$$\begin{aligned}\hat{V}^{-1} &\overset{def}{=} \hat{G}'\left(\hat{\Omega}+\alpha I\right)^{-1}\hat{G} \\ &= -\check{G}'\left(\check{\Omega}+\alpha I\right)^{-1}\check{G}\left[\frac{1+n^{-1}\sum_{i=1}^n s_1\left(\hat{v}_i\right)}{n^{-1}\sum_{i=1}^n s_1\left(\hat{v}_i\right)}\right] + \check{G}'\left(\check{\Omega}+\alpha I\right)^{-1}\check{G} \xrightarrow{p} V^{-1},\end{aligned}$$

giving the consistency result for the variance estimator. □