STAFF REPORTS

NO. 1090
MARCH 2024

REVISED
SEPTEMBER 2024

Micro Responses to Macro Shocks

Martín Almuzara | Víctor Sancibrián

FEDERAL RESERVE BANK *of* NEW YORK

Micro Responses to Macro Shocks
Martín Almuzara and Víctor Sancibrián
Federal Reserve Bank of New York Staff Reports, no. 1090
March 2024; revised September 2024
https://doi.org/10.59576/sr.1090

Abstract

We study panel data regression models when the shocks of interest are aggregate and possibly small relative to idiosyncratic noise. This speaks to a large empirical literature that targets impulse responses via panel local projections. We show how to interpret the estimated coefficients when units have heterogeneous responses and how to obtain valid standard errors and confidence intervals. A simple recipe leads to robust inference: including lags as controls and then clustering at the time level. This strategy is valid under general error dynamics and uniformly over the degree of signal-to-noise of macro shocks.

JEL classification: C32, C33, C38, C51
Key words: panel data, local projections, impulse responses, aggregate shocks, inference, signal-to-noise, heterogeneity

---

Almuzara: Federal Reserve Bank of New York (email: martin.almuzara@ny.frb.org). Sancibrián: CEMFI (email: victor.sancibrian@cemfi.edu.es). The authors greatly benefited from discussions with Manuel Arellano, Dmitry Arkhangelsky, Daniel Lewis, Mikkel Plagborg-Møller, and Enrique Sentana. They also thank Stéphane Bonhomme, Richard Crump, Òscar Jordà, Geert Mesters, and seminar participants at Boston University, CEMFI, Erasmus University Rotterdam, Georgetown University, Princeton University, Universidad Autónoma de Madrid, University of Pennsylvania, and conference participants at the EABCN-UPF conference, the Greater New York Econometrics Colloquium, the NBER Summer Institute, and the SED 2024 Annual Meeting for valuable comments. Babur Kocaoglu provided excellent research assistance. Sancibrián acknowledges financial support from Grant PRE2022-000906 funded by MCIN/AEI/10.13039/501100011033 and by "ESF +," the María de Maeztu Unit of Excellence CEMFI MDM-2016-0684, funded by MCIN/AEI/10.13039/501100011033, Fundación Ramón Areces, and CEMFI.

This paper presents preliminary findings and is being distributed to economists and other interested readers solely to stimulate discussion and elicit comments. The views expressed in this paper are those of the author(s) and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System. Any errors or omissions are the responsibility of the author(s).

To view the authors' disclosure statements, visit
https://www.newyorkfed.org/research/staff_reports/sr1090.html.

1 Introduction

Applied macroeconomists are increasingly interested in empirical estimates of the transmission of aggregate uncertainty to individual outcomes, often in the form of impulse responses.

A popular approach is to formulate estimating equations of the form

$$Y_{i,t+h} = \beta(h) s_i X_t + \text{controls} + v_{h,it}, \tag{1}$$

where Y_{it} is a *micro outcome* for unit i ($i = 1, \ldots, N$) at time t ($t = 1, \ldots, T$), such as household income or firm sales, and X_t an observed *macro shock* of interest, such as a monetary policy or oil supply shock. Shocks are often interacted with unit-level covariates s_i to document heterogeneity in transmission along observables. Estimates $\hat{\beta}(h)$ of the response at horizon h are then obtained via least squares; a panel local projections version of Jordà (2005).

Despite its routine application, little is known about the statistical properties of $\hat{\beta}(h)$. The way standard errors are computed in the empirical literature illustrates it well: in our own survey of almost 50 recent papers, around half compute two-way clustered standard errors, one-third cluster within units only, and many others resort to Driscoll and Kraay (1998). This reflects the vastly different ways in which researchers perceive the nature of shocks, the role of each dimension of the panel for precision, and the importance of aggregate variation in the data.

In this paper, we provide the first treatment of estimation and inference for this problem. We show how to interpret $\hat{\beta}(h)$ when impulse-response heterogeneity is unrestricted and propose standard errors and confidence intervals that are easy to compute and robust to the signal-to-noise of macro shocks in the microdata. As a result, a very simple recipe for inference emerges: clustering standard errors at the time level and ex-ante including sufficient lags as controls. We refer to this strategy as *time-clustered lag-augmented heteroskedasticity-robust (t-LAHR) inference*.

We establish our results in a comprehensive setup that features observed and unobserved macro and micro shocks, cross-sectional heterogeneity in responses, general forms of serial dependence in outcomes, and unrestricted signal-to-noise. We first show that $\hat{\beta}(h)$ recovers the slope coefficient of a population linear projection of unit-specific impulse responses on characteristics s_i, thereby formalizing what practitioners have in mind when including interactions in equation (1). If $s_i = 1$, the estimand boils down to the average response in the population. Notably, because we place no restrictions on the underlying impulse-response heterogeneity or s_i, our characterization of the estimand is in effect nonparametric.[1]

Signal-to-noise. The degree of signal-to-noise of macro shocks in the microdata is a crucial parameter of the problem. Here, it is *common* shocks to all units that drives identification, and how sizable they are relative to micro shocks determines both the strength of identifying variation and the extent of unaccounted-for spatial dependence.[2] The notion of different signal regimes also reflects the scope of empirical work, which takes interest in atomistic and granular agents, administrative and narrow datasets, unit-specific and aggregate regression controls, etc.

Hence, one of our key contributions is to introduce a novel asymptotic framework where the signal value of aggregates may be arbitrarily low (or high) in the limit. We achieve this by indexing the relative standard deviation of macro to micro shocks by a parameter κ that can drift with the sample size. This device allows for a range of data generating processes (DGPs) in which estimation uncertainty is dominated by micro-level terms, a combination of micro and macro errors, or aggregate components only.[3] On the contrary, standard asymptotic plans where κ is fixed only capture the latter and ignore idiosyncratic shocks, potentially leading to poor approximations in small samples. It is clear then that the nature of estimation error depends on κ and the question is whether inference procedures are robust to

[1]We discuss extensions to (exogenous) time-varying characteristics s_{it} in Section 3 (Remark 7).

[2]It is immediate that if $s_i = 1$ in equation (1), including time fixed effects causes collinearity. If s_i varies over units, for time indicators to remove all additional aggregate variation one would need the untenable assumption that *only* impulse responses to X_t at horizon h are heterogeneous. In our exposition, we always allow for time indicators as controls when s_i displays cross-sectional variation.

[3]Our approach also resonates with the renewed interest on the potential for unit-level shocks to explain aggregate fluctuations, as in Gabaix (2011) and subsequent literature. Our device to obtain non-negligible micro errors is closer to Jovanovic (1987) in that we rely on scaling micro variation up rather than on fat-tailed distributions. However, we conjecture that similar inference results can be obtained in the latter under appropriate regularity conditions.

different macro signal regimes. Our main result is that *t-LAHR inference is uniformly valid* over κ, in other words, t-LAHR confidence intervals have correct asymptotic coverage for the (nonparametric) local projection estimand uniformly over κ.

Inference. The key assumption in our framework is the availability of an observed macro shock X_t. Our notion of shocks is that of mean independent innovations with respect to both its own lags and leads and other shocks, in line with the time series literature on local projections inference (Stock and Watson, 2018; Montiel Olea and Plagborg-Møller, 2021). We first focus on the case where the shock of interest is observed — an assumption prevalent in most empirical applications — and then consider settings where the shock of interest is recoverable (spanned by X_t and its lags) or contaminated with measurement error but a proxy is available (as in local projection-instrumental variable estimators; LP-IV for short).[4]

The *macro* and *shock* nature of X_t delivers the following observation which serves as a guiding principle throughout the paper: panel local projections with macro shocks are equivalent to *synthetic* time series local projections with an appropriately aggregated dependent variable. This is true even if shocks interact with covariates s_i and if unit and time effects are included. Therefore, aggregating the microdata by collapsing the cross-sectional dimension of the panel and treating it as a time series delivers valid inference for any κ.[5] This is precisely what t-LAHR inference does, since time-level clustering in the panel problem and heteroskedasticity-robust inference in the synthetic time series problem are essentially equivalent.

The *macro* and *shock* nature of X_t also clarifies the role of lag augmentation. In a panel local projection that controls for p lags of $s_i X_t$, the regression scores (the product of shocks and residuals) are *nearly* uncorrelated even if residuals are not.

[4]Examples of popular identification methods include narrative approaches (as in Crouzet and Mehrotra, 2020, for monetary policy shocks), high-frequency identification (as in Känzig, 2021, for oil supply shocks) or a combination of Cholesky/structural VAR restrictions (as in Drechsel, 2023, for firm investment shocks). See Ramey (2016, Section 2.3) for a review of identification methods in macroeconometrics.

[5]This synthetic time series representation is also illustrative of the fact that the concentration rate of the estimation error is at most $T^{-1/2}$, even in situations where $N \gg T$. This suggests caution regarding the conventional wisdom in many empirical applications that a larger cross-sectional dimension somehow compensates for a shorter time series.

Specifically, they are a moving average of order h where the first p autocovariances are zero and the remaining ones are independent of κ. This has two major implications. First, it confers a double layer of simplicity to inference: up to horizon $h \leq p$, there is no need for unit-level clustering or heteroskedasticity and autocorrelation robust (HAR) approaches to deal with serial dependence. Second, it explains why t-LAHR inference might have only small coverage distortions even for horizons exceeding p: these distortions depend on the size of the autocorrelation coefficients of the score, which are small in low-signal environments. In fact, if the DGP is well approximated by a low-order vector autoregression (VAR), we prove t-LAHR inference is uniformly valid over both κ and $h \propto T$, a result reminiscent of those in Montiel Olea and Plagborg-Møller (2021) for time series local projections.

We complement our theoretical results with simulations for realistic designs and sample sizes, allowing for moderately long horizons and substantial persistence in micro shocks. We study the performance of a battery of approaches, including an alternative to t-LAHR that substitutes lag augmentation with HAR inference, and incorporating small-sample refinements (Müller, 2004; Imbens and Kolesár, 2016; Lazarus, Lewis, Stock, and Watson, 2018). We find that t-LAHR inference shows remarkable performance relative to all other competitors, particularly in low-signal environments, in near non-stationary scenarios, and over moderate horizons even if we do not impose a VAR on outcomes.[6] In practice, we recommend to supplement t-LAHR inference (controlling for a reasonable number of lags of both outcome and shock variables) with the refinement proposed by Imbens and Kolesár (2016).

Empirical survey and illustration. We reviewed a large body of empirical work that precedes this paper. The typical application uses administrative data for firms, tracks units at the quarterly or annual frequency for a limited number of periods, and estimates impulse responses to monetary policy shocks via local projections.

[6] It is known that ad-hoc parameter choices and small-sample biases in sample autocovariances contribute to the subpar relative performance of HAR estimators (Herbst and Johansenn, 2023).

Most applications include interactions of the form $s_i X_t$ and both unit and time fixed effects, but vary widely in the number and nature of additional controls.[7]

In otherwise comparable empirical designs, we document large dispersion in the way practitioners compute standard errors: among 47 different papers, 24 compute two-way clustered standard errors (within units and time), 15 cluster within units only, 7 use Driscoll and Kraay (1998) and 1 clusters within time only.

These choices reflect very different views on the dominant sources of statistical uncertainty, from ruling out serial dependence to ruling out spatial dependence; from a suggestion that both unit-level and aggregate dynamics need to be accounted for to an explicit stance that either of the two dominates. Often these choices are made with little discussion or citing previous work as justification.[8] Our framework allows us to revisit them. First, off-the-shelf autocorrelation consistent methods such as Driscoll and Kraay (1998) leave information on the table (the autocovariance function of the regression scores is known), which comes at a cost in small samples. Second, validity in the case where standard errors do not explicitly adjust for serial dependence (as in two-way clustering) boils down to whether a reasonable number of lags was included in estimation. Third, clustering within units is superfluous, even in low-signal regimes where the size of unit-level dynamics is comparable to that of aggregates. Fourth, clustering only within units breaks down even in the face of small amounts of spatial dependence induced by aggregate shocks (that is, moderate-signal environments). In fact, we offer a way to reinterpret these

[7] We reproduce the full list in Supplemental Appendix D which includes 47 empirical papers that run panel regressions with macro shocks. A few focus on the case $h = 0$ only, but the vast majority compute impulse responses over several horizons. The economic content of X_t is very diverse, including fiscal policy shocks, investment shocks, TFP and innovation shocks, carbon pricing shocks, temperature shocks, etc. In these applications, the cross-sectional dimension is usually orders of magnitude larger than the effective time-series dimension. In our review we leave out empirical work with very small cross-sections where entities are meaningful and a unit-by-unit treatment is feasible. Nonetheless, when these units are pooled together, as in Fukui, Nakamura, and Steinsson (2023), our results still apply.

[8] The availability of a large cross-sectional dimension and the interaction of shocks with covariates s_i are also often argued as sources of large gains in statistical precision, also reflecting an implicit stance on the presence of macro shocks. We elaborate on the (im)plausibility of these notions in Remark 5.

confidence intervals as providing valid inference for an estimand indexed to the actual realizations of aggregate shocks during the sample period.

Finally, we illustrate our methods in an empirical exercise inspired by a booming literature that investigates the role of financial frictions and firm heterogeneity in the transmission of monetary policy.[9] The exercise highlights the importance of the choice of inference method, and the value of the synthetic time series representation as a way to gain intuition about the source of identifying variation.

Related literature. Our paper contributes to various strands of the literature.

First, it relates to the time series literature on inference for local projections (Hansen and Hodrick, 1980; Jordà, 2005; Stock and Watson, 2018; Montiel Olea and Plagborg-Møller, 2021; Lusompa, 2023; Xu, 2023; Montiel Olea, Plagborg-Møller, Qian, and Wolf, 2024). Relative to this literature we are (to our knowledge) the first to deal with the panel data case with aggregate shocks.[10]

In a time series finite-order VAR setup, Montiel Olea and Plagborg-Møller (2021) show the uniform validity of heteroskedasticity-robust inference on lag-augmented local projections over the persistence in the data and horizon h. They also postulate mean independent innovations, the same type of assumption we impose on X_t. Our Proposition 2 (and, more generally, Section 3.3) can be interpreted as the panel version of their results. Nonetheless, our focus is on uniformity with respect to the macro signal-to-noise κ, which has no obvious counterpart in the time series setup, and we derive most of our results without assuming a VAR model.

[9] For instance, Crouzet and Mehrotra (2020), Ottonello and Winberry (2020), Anderson and Cesa-Bianchi (2024) and Jeenas and Lagos (2024) target impulse responses of firm investment to monetary policy shocks interacted with external covariates s_i such as firm size, default risk or stock turnover.

[10] Our results on limited serial dependence in regression scores relate to the earlier multi-step forecast literature (Hansen and Hodrick, 1980), which relied on including infinitely many lags to ensure that the forecast errors have a MA(h) representation. In the local projection context, Jordà (2005) arrived at a similar result under a finite-order VAR model while Lusompa (2023) provided a recent reformulation. Instead, we exploit the orthogonality properties of macro shocks to show that the *scores* have MA(h) dynamics. The distinction is reminiscent of the difference between design-based and model-based/conditional unconfoundedness assumptions.

Second, we contribute to the literature on estimation and inference with aggregate shocks. Using stylized models, Hahn, Kuersteiner, and Mazzocco (2020) bring attention to the drastic consequences of drawing inferences from short panels with aggregate uncertainty. Although our focus is on thought experiments where macro shocks are a key source of identification, we can connect to their results by reinterpreting confidence intervals that exploit independence across units as valid for an approach to inference that conditions on the path of aggregate shocks.

Recent additions to this literature study regional-exposure designs where the researcher has access to low-rank instruments of the form $s_i X_t$ (s_i are region-specific exposures to aggregate conditions) and so the reduced-form equation looks like (1) for $h = 0$. Arkhangelsky and Korovkin (2023) argue that exogenous variation comes from the time series shock X_t and focus on threats to instrument validity, whereas Majerovitz and Sastry (2023) consider either s_i or X_t as sources of identification and suggest that inference needs to take spatial dependence into account in the latter case. Our paper extends these ideas by giving formal inference results that cover dynamic responses and different macro signal environments.

Third, our paper relates to the cross-sectional dependence literature that studies models where the scores feature varying degrees of spatial dependence (Driscoll and Kraay, 1998; Andrews, 2005; Pesaran, 2006; Gonçalves, 2011; Pakel, 2019). Our framework falls in the polar case where the shock of interest only varies over time, precluding solutions based on partialling out the common component from the regressors, as in Pesaran (2006). Moreover, our uniformity result (which translates into robustness to the degree of spatial dependence) is new to the literature.

Outline. Section 2 provides an overview of our results in a simple static model, illustrating the role of aggregate shocks and their signal relative to micro shocks. Section 3 presents our main inference result in a general, heterogeneous dynamic model. Section 4 discusses a comprehensive simulation study and Section 5 the empirical illustration. Proofs can be found in Appendix A with additional details in the Supplemental Appendix. A Matlab code repository is available online.[11]

[11] https://github.com/TinchoAlmuzara/PanelLocalProjections.

2 Simple model

We illustrate the main points of the paper in a simple, static regression model with homogeneous responses. We keep the exposition simple and omit technical details with the goal of building insights. The more general setup is studied in Section 3.

Model assumptions. We observe a micro outcome Y_{it} and a macro shock X_t for units $i = 1, \ldots, N$ and over periods $t = 1, \ldots, T$. They are related by

$$\begin{aligned} Y_{it} &= \beta_0 X_t + v_{it}, \\ v_{it} &= Z_t + \kappa u_{it}, \end{aligned} \tag{2}$$

where v_{it} is an error term including both aggregate and idiosyncratic unobservables, denoted Z_t and u_{it}, respectively. Here κ regulates their relative importance in the micro data, as explained below. The goal is to estimate and do inference on β_0.

This simple model is a stylized representation of an empirical setting where we are interested in the transmission of aggregate uncertainty to individual outcomes; the effect of X_t on Y_{it}. Examples of the former include changes in interest rates, tax regulations or oil prices, which might leave a mark on household consumption, worker's labor income or firm sales. In fact, one could entertain any combination of macro variables and micro outcomes in these examples. When interest centers around one aggregate variable — captured by X_t — it would be hard to ex-ante rule out the presence of any others — embedded in Z_t. This basic premise is at the core of the our results.

We now make two sets of assumptions, later generalized in Section 3 to allow for observable and unobservable heterogeneity, and more flexible dynamics.

Assumption S1 (Stationarity and iidness in the simple model).

(i) $\{X_t, Z_t, \{u_{it}\}_{i=1}^N\}_{t=1}^T$ *is stationary.*

(ii) $\{\{u_{it}\}_{i=1}^N\}_{t=-\infty}^{\infty}$ *are i.i.d. over i conditional on* $\{X_t, Z_t\}_{t=1}^T$.

Assumption S1(i) implies Y_{it} is stationary too. Assumption S1(ii) simply assigns the role of inducing cross-sectional dependence in the error term v_{it} to Z_t.[12]

Assumption S2 (Shocks and independence in the simple model).

(i) $E\left[X_t \middle| \{X_\tau\}_{\tau\neq t}, \left\{Z_\tau, \{u_{i\tau}\}_{i=1}^N\right\}_{\tau=1}^T\right] = 0.$

(ii) $E\left[Z_t \middle| \{Z_\tau\}_{\tau\neq t}, \left\{X_\tau, \{u_{i\tau}\}_{i=1}^N\right\}_{\tau=1}^T\right] = 0.$

(iii) $E\left[u_{it} \middle| \{u_{i\tau}\}_{\tau\neq t}, \{X_\tau, Z_\tau\}_{\tau=1}^T\right] = 0.$

Assumption S2 implies X_t, Z_t and u_{it} are mutually unpredictable and serially uncorrelated. Assumption S2(i) is ultimately an identification condition, whereas S2(ii) and S2(iii) are made for symmetry. Indeed, mutual unpredictability of macro shocks lies at the core of macroeconometrics and is typically necessary to give structural interpretation to impulse-response calculations (see, for instance, Ramey, 2016; Stock and Watson, 2016; Plagborg-Møller and Wolf, 2021).[13] Assumption S2(i) is an empirically realistic starting point, since in the majority of applications X_t is the (perhaps noisy) measurement of a shock.

Remark 1 (Relaxing Assumption S2). In practice, we might only observe a proxy shock X_t^*, which may be contaminated with measurement error or possess some residual autocorrelation structure, say $X_t^* = \sum_{\ell=1}^k \alpha_\ell X_{t-\ell}^* + X_t$ for known $k < \infty$. These cases can be handled by treating X_t^* as an instrument — a panel version of the LP-IV estimator (Stock and Watson, 2018, Section 1.3), which we study in Section 3.4 — or by including lags of X_t^* as controls, see also Section 3.3.

[12]Both assumptions can be relaxed; we briefly discuss departures from S1(i) in Section 3 and 4. Allowing for weak spatial dependence in u_{it} in place of S1(ii) is also possible with minor modifications.

[13]Mean independence assumptions with respect to past and future innovations are a slight strengthening of the more standard martingale difference assumptions, and are convenient in representations where both leads and lags of the variable might enter the model, cf. Montiel Olea and Plagborg-Møller (2021, Assumption 1) in a similar context of local projection inference. This still allows for dynamics on the second- or higher-order moments given the paths of other shocks. It permits that, say, monetary, fiscal or oil supply shocks (X_t, Z_t) increase the variance of household-level income (Y_{it}) via higher order dynamics in u_{it}.

Estimation and inference. A natural estimator of β_0 is pooled least squares,

$$\hat{\beta} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} X_t Y_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T} X_t^2} = \frac{\sum_{t=1}^{T} X_t \left(N^{-1}\sum_{i=1}^{N} Y_{it}\right)}{\sum_{t=1}^{T} X_t^2},$$

which is also a panel local projection (LP) estimator at horizon $h = 0$ and the estimator in a time series regression involving the synthetic outcome $\hat{Y}_t = N^{-1}\sum_{i=1}^{N} Y_{it}$ and X_t. The double nature of $\hat{\beta}$ as panel and time series estimator arises naturally in the presence of macro shocks, as we further demonstrate in Section 3.

Denote the residual by $\hat{\xi}_{it} = Y_{it} - \hat{\beta} X_t$. A key takeaway from our paper is that a reliable approach to inference uses the time-level cluster heteroskedasticity-robust standard error $\hat{\sigma}$, given by $\hat{\sigma}^2 = \hat{V}/T\hat{J}^2$ where $\hat{J} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} X_t^2 = T^{-1}\sum_{t=1}^{T} X_t^2$ is the least squares denominator and

$$\hat{V} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N} X_t \hat{\xi}_{it}\right)^2.$$

Another sign of the duality between panel regressions with aggregate shocks and time series regression is that $\hat{\sigma}$ is also the usual Eicker–Huber–White standard error computed using the synthetic time series residuals $\hat{\xi}_t = N^{-1}\sum_{i=1}^{N} \hat{\xi}_{it}$.

As mentioned in the Introduction, two popular inferential choices in applications are based on one-way (unit-level) cluster and two-way (unit- and time-level) cluster standard errors, $\hat{\sigma}_{1W}$ and $\hat{\sigma}_{2W}$, given by $\hat{\sigma}_{1W}^2 = \hat{V}_{1W}/T\hat{J}^2$ and $\hat{\sigma}_{2W}^2 = \hat{V}_{2W}/T\hat{J}^2$ where

$$\hat{V}_{1W} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} X_t \hat{\xi}_{it}\right)^2, \quad \hat{V}_{2W} = \hat{V} + \hat{V}_{1W} - \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} X_t^2 \hat{\xi}_{it}^2.$$

These standard errors reflect different concerns about the nature of estimation error or, more precisely, the correlation of the regression score $X_t v_{it}$ over units and time.

Substituting (2), the estimation error decomposes as

$$\hat{\beta} - \beta_0 = \underbrace{\frac{\sum_{t=1}^{T} X_t Z_t}{\sum_{t=1}^{T} X_t^2}}_{O_p(1)} + \frac{\kappa}{\sqrt{N}} \underbrace{\left(\frac{1}{\sqrt{N}} \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} X_t u_{it}}{\sum_{t=1}^{T} X_t^2}\right)}_{O_p(1)}, \tag{3}$$

i.e., as the sum of macro and micro components. The former induces cross-sectional correlation while the latter is uncorrelated across units and both have limited serial dependence — for $t \neq \tau$, $E\left[X_t Z_t \cdot X_\tau Z_\tau\right] = E\left[X_t u_{it} \cdot X_\tau u_{i\tau}\right] = 0$ by Assumption S2(i) and iterated expectations.[14] This is a direct consequence of X_t being a shock.

The intuition for why $\hat{\sigma}$ gives valid inference is the following. If the macro term is not asymptotically small, $X_t v_{it}$ displays correlation over i but not over t, the type of situation for which $\hat{\sigma}$ is designed. If, on the other hand, the micro term dominates, $X_t v_{it}$ is uncorrelated over both i and t. Yet $\hat{\sigma}$ still works: while it does not impose that the cross-sectional covariances of $X_t v_{it}$ are zero, it will correctly estimate them to be zero. One may wish to switch to a non-clustered heteroskedasticity-robust standard error in that case, but we show both analytically (Proposition 1) and in simulations (Section 4) that there is no loss in simply using $\hat{\sigma}$.

Clearly, correlation over t at the unit-level is never a concern; that is why unit-level clustering either fails or is not needed. In fact, $\hat{\sigma}_{1W}$ is asymptotically equivalent to the non-clustered standard error, and the same holds for $\hat{\sigma}_{2W}$ and $\hat{\sigma}$.

Macro-micro signal-to-noise ratio. Which term dominates the decomposition (3) will depend upon $\kappa/\sqrt{N}$. We now provide another interpretation of this quantity. Consider the average outcome $\hat{Y}_t = N^{-1}\sum_{i=1}^{N} Y_{it}$ and, for the sake of illustration, suppose $\text{Var}(Z_t) = \text{Var}(u_{it}) = 1$. By Assumptions S1 and S2, the proportion of the variance of $\hat{Y}_t$ explained by the unobserved macro error can be measured as

$$\bar{R}^2(\kappa) = 1 - \frac{\text{Var}\left(\hat{Y}_t \middle| X_t, Z_t\right)}{\text{Var}\left(\hat{Y}_t \middle| X_t\right)} = \frac{1}{1 + \kappa^2/N}, \tag{4}$$

that is, the signal-to-noise ratio is $O(N/\kappa^2)$. It increases with N since cross-sectional averaging reduces the variance from idiosyncratic errors, but decreases with $|\kappa|$.

We will study estimation and inference in sequences of data generating processes (DGPs) where κ is allowed to grow as $T, N \to \infty$. This leads, in essence, to three regimes. If $\kappa/\sqrt{N} = o(1)$, (such as if κ is fixed), $\bar{R}^2(\kappa) \to 1$ and macro shocks are the only source of aggregate variation; we call this the asymptotically high-signal case.

[14]The lack of serial correlation would remain true even if Z_t and u_{it} were serially correlated.

If $\kappa \propto \sqrt{N}$, $\bar{R}^2(\kappa)$ is bounded away from 0 and 1 in the limit and both macro and micro shocks matter for aggregate fluctuations; this is the asymptotically moderate-signal case. Finally, if $\kappa/\sqrt{N}$ diverges, $\bar{R}^2(\kappa) \to 0$, macro shocks are imperceptible and we are in the asymptotically low-signal case.[15]

The intuitive notion of κ-regimes has a natural counterpart in our asymptotic approximations, in that there is a close relation between the contribution of macro shocks to $\hat{Y}_t$ and the nature of estimation error for β_0, as illustrated by (2) and (4). In particular, the macro term dominates in the high-signal case, the micro term dominates in the low-signal case, and they are roughly balanced in the moderate-signal case. Moreover, it is not always possible to consistently detect what κ-regime applies. It is important then to derive inference procedures that are robust in the sense of uniform validity with respect to κ.[16]

Uniformity over κ. From the decomposition in (3), letting $N, T \to \infty$ and under regularity conditions specified in Section 3,

$$\sigma_0(\kappa)^{-1}\sqrt{T}\left(\hat{\beta} - \beta_0\right) \xrightarrow{\text{d}} N(0,1),$$

where

$$\left\{E\left[X_t^2\right]\right\}^2 \times \sigma_0(\kappa)^2 = \begin{cases} E\left[X_t^2 Z_t^2\right], & \text{if } \kappa/\sqrt{N} \to 0, \\ E\left[X_t^2\left(Z_t^2 + \bar{\kappa}^2 u_{it}^2\right)\right], & \text{if } \kappa/\sqrt{N} \to \bar{\kappa}, \\ \left(\kappa^2/N\right) E\left[X_t^2 u_{it}^2\right], & \text{if } \kappa/\sqrt{N} \to \infty, \end{cases}$$

This shows two things. First, the rate of concentration of the estimation error $\hat{\beta} - \beta_0$ is either $\sqrt{T}$ in the high- and moderate-signal cases or $\sqrt{NT}/\kappa$ (i.e., slower than $\sqrt{T}$

[15]Of course, letting κ grow with the sample size should not be taken literally — it is simply a device to ensure our approximations suitably interpolate between high and low signal-noise environments. This type of embeddings are common in econometrics; an example which also has a low-signal interpretation is weak IV (Staiger and Stock, 1997).

[16]We will consider inference procedures that are invariant to rescaling. It follows that all of our results can be equivalently obtained in an embedding that scales down the macro component of the model in (2) by κ^{-1}. Put differently, what matters is the relative size of macro and micro components.

and possibly even zero, thus making $\hat{\beta}$ inconsistent) in the low-signal case. Second, the asymptotic distribution of $\hat{\beta}$ changes discontinuously across κ-regimes.

Despite the discontinuity, our main result is that the $(1-\alpha)$ confidence interval $\hat{C}_\alpha = \left[\hat{\beta} \pm z_{1-\alpha/2}\hat{\sigma}\right]$, where z_q is the q-quantile of the standard normal distribution, has correct coverage for β_0 *uniformly* over κ,

$$\lim_{T,N\to\infty} \sup_{\kappa} \left| P_\kappa\left(\beta_0 \in \hat{C}_\alpha\right) - (1-\alpha)\right| = 0. \tag{5}$$

where P_κ denotes probabilities for a DGP with a given κ. This is much stronger than pointwise validity, as it implies that the quality of the asymptotic approximation to the coverage probability of $\hat{C}_\alpha$ is itself robust to the κ-regime. Statement (5) also means that if sample information about macro shocks is extremely scarce and $\hat{\beta}$ is inconsistent, the length of $\hat{C}_\alpha$ adjusts as needed to reflect the weak macro signal.

One might wonder how much the static nature of (2) limits these results. The rest of the paper will show that they extrapolate to a substantially more general and empirically realistic framework with rich forms of dynamics and heterogeneity.

Remark 2 (Inference conditional on aggregate shocks). Ignoring the unobservable macro component in (3) when doing inference is equivalent to conditioning on its realization. In that situation, $\hat{\sigma}_{1W}$ is a valid standard error for responses defined by moment restrictions that condition on the realized path of aggregate shocks during the sample period.[17] In general, this induces an internal/external validity trade-off whereby practitioners may be able to pin down certain parameters very precisely but these might lack generalizability to other contexts.

3 General case

In this section, we establish estimation and inference results for impulse responses to aggregate shocks in a general setup featuring observed and unobserved, macro and micro shocks, and unrestricted heterogeneity of individual responses.

[17] A proof and additional details are available upon request. As a practical example, we think of the responses of micro outcomes to monetary and fiscal policies during the COVID-19 pandemic.

We introduce the setup in Section 3.1 and state the main results in Section 3.2. We treat the important case of finite-order VAR DGPs in Section 3.3 and local projections with instrumental variables (LP-IV) in Section 3.4. Proofs are developed in Appendix A with technical lemmas in Supplemental Appendix B.

3.1 Setup

The researcher observes an outcome Y_{it}, an aggregate shock X_t and characteristics s_i for units $i = 1, \dots, N$ and over periods $t = 1, \dots, T$. Everything is scalar but it is straightforward to extend the results to the multivariate case. We assume

$$Y_{it} = \mu_i + \sum_{\ell=0}^{\infty} \beta_{i\ell} X_{t-\ell} + v_{it}, \tag{6}$$

$$v_{it} = \sum_{\ell=0}^{\infty} \gamma_{i\ell} Z_{t-\ell} + \kappa \sum_{\ell=0}^{\infty} \delta_{i\ell} u_{i,t-\ell}, \tag{7}$$

where Z_t and u_{it} are unobserved serially uncorrelated aggregate and idiosyncratic errors. We denote $\beta_i = \{\beta_{i\ell}\}_{\ell=0}^{\infty}$, $\gamma_i = \{\gamma_{i\ell}\}_{\ell=0}^{\infty}$, $\delta_i = \{\delta_{i\ell}\}_{\ell=0}^{\infty}$ and $\theta_i = \{\mu_i, \beta_i, \gamma_i, \delta_i\}$. These are draws from a cross-sectional distribution and below we specify conditions so that the infinite sums in (6)-(7) are well defined with probability one.

Here, θ_i traces out cross-sectionally heterogeneous responses to both aggregate and idiosyncratic shocks, and access to external variables s_i allows the researcher to study their transmission along unit-level observables. Our premise is that there is usually more heterogeneity in θ_i than can be explained by s_i alone and our goal is to characterize estimation and inference in that context.

As in Section 2, we consider a range of DGPs indexed by κ to cover different signal-to-noise environments. We also make the following assumptions:

Assumption 1 (Stationarity and iidness).

(i) $\{X_t, Z_t, \{u_{it}\}_{i=1}^{N}\}_{t=-\infty}^{\infty}$ *is stationary conditional on* $\{\theta_i, s_i\}_{i=1}^{N}$.

(ii) $\{\theta_i, s_i, \{u_{it}\}_{t=-\infty}^{\infty}\}_{i=1}^{N}$ *is i.i.d. over i conditional on* $\{X_t, Z_t\}_{t=-\infty}^{\infty}$.

Assumption 2 (Shocks and mean independence).

(i) $E\left[X_t \middle| \{X_\tau\}_{\tau \neq t}, \left\{Z_\tau, \{u_{i\tau}\}_{i=1}^{N}\right\}_{\tau=-\infty}^{\infty}, \{\theta_i, s_i\}_{i=1}^{N}\right] = 0.$

(ii) $E\left[Z_t \middle| \{Z_\tau\}_{\tau\neq t}, \left\{X_\tau, \{u_{i\tau}\}_{i=1}^N\right\}_{\tau=-\infty}^{\infty}, \{\theta_i, s_i\}_{i=1}^N\right] = 0.$

(iii) $E\left[u_{it} \middle| \{u_{i\tau}\}_{\tau\neq t}, \{X_\tau, Z_\tau\}_{\tau=-\infty}^{\infty}, \theta_i, s_i\right] = 0.$

Assumptions 1 and 2 generalize S1 and S2 to accommodate the presence of both unobserved heterogeneity and external covariates. Assumption 2 requires them to be strictly exogenous with respect to shocks. Importantly, the joint distribution of (θ_i, s_i) is left unrestricted, and so is that of $\{\theta_i, s_i\}_{i=1}^N$ conditional on $\{X_t\}_{t=-\infty}^{\infty}$, as in pure fixed effects approaches. For a discussion of all other components, we refer the reader to Section 2. Again, the crucial assumption is 2(i) on the availability of an observed macro shock satisfying certain orthogonality conditions. We consider alternatives to it in the form of mismeasurement with an instrument in Section 3.4.

3.1.1 Estimator and inference procedure

We now introduce the panel LP estimator and inference procedure. We denote by $W_{it} \in \mathbb{R}^d$ the vector of controls (d may change with the sample size). If W_{it} contains no time fixed effects, let $\hat{s}_i = s_i$ — this accommodates the case $s_i = 1$. Otherwise, let $\hat{s}_i = s_i - N^{-1}\sum_{j=1}^N s_j$ and note that if time fixed effects are included, local projections on $s_i X_t$ and $\hat{s}_i X_t$ produce numerically the same estimate $\hat{\beta}(h)$ below. In addition to unit and possibly time dummies, we consider below cases in which W_{it} contains lags of $s_i X_t$ or Y_{it} and we assume that W_{it} is observed for $t = 1, \ldots, T$.[18]

The fitted equation for the panel LP estimator $\hat{\beta}(h)$ is

$$Y_{i,t+h} = \hat{\beta}(h)\hat{s}_i X_t + \hat{\eta}(h)' W_{it} + \hat{\xi}_{it}(h),$$

where the residual $\hat{\xi}_{it}(h)$ is orthogonal to $\hat{s}_i X_t$ and W_{it}. To characterize $\hat{\beta}(h)$ we use Frisch–Waugh–Lovell. Consider the auxiliary regression of $\hat{s}_i X_t$ on W_{it},

$$\hat{s}_i X_t = \hat{\pi}(h)' W_{it} + \hat{x}_{it}(h), \tag{8}$$

[18]Since Y_{it}, X_t and s_i could be multivariate, this is without loss of generality. For example, a panel LP of Y_{it} on $s_i X_t$ controlling for X_t and lags of Y_{it} and another micro control $\tilde{Y}_{it}$ is covered by redefining Y_{it} to $(Y_{it}, \tilde{Y}_{it})$ and s_i to $(1, s_i)$. Also, note that if W_{it} includes lags of shocks or outcomes we assume we observe $s_i X_t$ or Y_{it} for $t < 1$.

where the residual $\hat{x}_{it}(h)$ is orthogonal to W_{it}. Then, an explicit formula for $\hat{\beta}(h)$ is

$$\hat{\beta}(h) = \frac{\sum_{t=1}^{T-h} \sum_{i=1}^{N} \hat{x}_{it}(h) Y_{i,t+h}}{\sum_{t=1}^{T-h} \sum_{i=1}^{N} \hat{x}_{it}(h)^2}. \tag{9}$$

The time-clustered heteroskedasticity-robust standard error is

$$\hat{\sigma}(h) = \sqrt{\frac{\hat{V}(h)}{(T-h)\hat{J}(h)^2}}, \tag{10}$$

with

$$\hat{J}(h) = \frac{1}{N(T-h)} \sum_{t=1}^{T-h} \sum_{i=1}^{N} \hat{x}_{it}(h)^2, \quad \hat{V}(h) = \frac{1}{(T-h)} \sum_{t=1}^{T-h} \left(\frac{1}{N} \sum_{i=1}^{N} \hat{x}_{it}(h) \hat{\xi}_{it}(h) \right)^2. \tag{11}$$

Finally, the $(1-\alpha)$ confidence interval is

$$\hat{C}_\alpha(h) = \left[\hat{\beta}(h) \pm z_{1-\alpha/2} \hat{\sigma}(h) \right], \tag{12}$$

where z_q is the q-quantile of the standard normal distribution.

3.1.2 Additional assumptions

To establish our uniform asymptotic approximations, we need the following:

Assumption 3 (Regularity conditions).

(i) There is a positive finite constant M_8 such that, almost surely,

$$E\left[X_t^8 \middle| \{\theta_i, s_i\}_{i=1}^N\right] \leq M_8, \quad E\left[Z_t^8 \middle| \{\theta_i, s_i\}_{i=1}^N\right] \leq M_8, \quad E\left[u_{it}^8 \middle| \theta_i, s_i\right] \leq M_8.$$

(ii) There is a positive finite constant $\underline{M}$ such that, almost surely,

$$E\left[X_t^2 \middle| \{X_\tau\}_{\tau \neq t}, \{\theta_i, s_i\}_{i=1}^N\right] \geq \underline{M}, \quad E\left[Z_t^2 \middle| \{X_\tau\}, \{\theta_i, s_i\}_{i=1}^N\right] \geq \underline{M}, \quad E\left[u_{it}^2 \middle| \{X_\tau\}, \theta_i, s_i\right] \geq \underline{M}.$$

(iii) The conditional cumulants up to fourth-order of $\operatorname{vec}\{(X_t, Z_t, u_{it})(X_t, Z_t, u_{it})'\}$ *given $\{\theta_i, s_i\}_{i=1}^N$ are almost surely absolutely summable.*

(iv) There are positive finite constants C_ℓ such that $C = \sum_{\ell=0}^\infty C_\ell < \infty$ and, almost surely,

$$|\beta_{i\ell}| \leq C_\ell, \quad |\gamma_{i\ell}| \leq C_\ell, \quad |\delta_{i\ell}| \leq C_\ell, \quad |s_i| < C.$$

(v) There is a positive finite constant $\underline{C}$ such that, almost surely,

$$\sum_{\ell=0}^{\infty}\left(N^{-1}\sum_{i=1}^{N}\hat{s}_i\beta_{i\ell}\right)^2 \geq \underline{C}, \quad \sum_{\ell=0}^{\infty}\left(N^{-1}\sum_{i=1}^{N}\hat{s}_i\gamma_{i\ell}\right)^2 \geq \underline{C}, \quad N^{-1}\sum_{\ell=0}^{\infty}\sum_{i=1}^{N}\hat{s}_i^2\delta_{i\ell}^2 \geq \underline{C}.$$

Our model interprets θ_i as unit-specific parameters and $\{X_t, Z_t, u_{it}\}$ as sources of uncertainty. This calls for making time series assumptions on the uncertainty given parameters (parts (i), (ii) and (iii)) while requiring that parameters ensure sufficient regularity for all units in the cross-sectional population (parts (iv) and (v)).

Parts (i), (ii) and (iii) are standard in the time series context (see, for instance, Assumption 2 in Montiel Olea and Plagborg-Møller (2021)). They put limits on the tails of the distributions of shocks, as well as the predictability and dependence of their second moments. Part (iv), on the other hand, guarantees that infinite moving averages, such as $\sum_{\ell=0}^{\infty}\beta_{i\ell}X_{t-\ell}$, are well defined for all units. Absolute summability rules out unit roots but still allows for rich persistence patterns — such as those from stationary ARMA and other short-memory processes.[19]

Lastly, part (v) requires that $N^{-1}\sum_{i=1}^{N}\hat{s}_i\sum_{\ell=0}^{\infty}\beta_{i\ell}X_{t-\ell}$, $N^{-1}\sum_{i=1}^{N}\hat{s}_i\sum_{\ell=0}^{\infty}\gamma_{i\ell}Z_{t-\ell}$ and $N^{-1/2}\sum_{i=1}^{N}\hat{s}_i\sum_{\ell=0}^{\infty}\delta_{i\ell}u_{i,t-\ell}$ display non-zero variability conditional on $\{\theta_i, s_i\}_{i=1}^{N}$. It is mostly a technical condition to prevent trivial cases in which the regression score has zero variance. Nevertheless, it is compatible with, say, a non-negligible fraction of units having zero exposure to macro or micro shocks. It also places no restriction on the relative importance of macro versus micro shocks which is governed by κ.

3.2 Main result

The main contribution of the paper is to characterize the large-sample properties of $\hat{\beta}(h)$, $\hat{\sigma}(h)$ and $\hat{C}_\alpha(h)$. In the asymptotic plan, we take $T, N \to \infty$ and we are interested in uniform approximations with respect to κ. The key result is Proposition 1 which states that $\hat{C}_\alpha(h)$ delivers uniformly valid inference for the coefficient in a regression of β_{ih} on $\hat{s}_i$ if enough lags of $\hat{s}_iX_t$ are used as controls.

[19] We conjecture, however, that many of our results remain valid at moderate horizons in the presence of near unit roots and our simulation evidence supports this claim. See Section 3.3 for further discussion.

We describe first the estimand and then the uniform inference result. We use P_κ to indicate probabilities under a DGP associated to a given value of κ and we omit the subindex from objects whose probabilities (or expectations) do not depend on κ (such as those in Assumptions 2 and 3).

Estimand. If s_i is not a constant and time fixed effects are included, the population object targeted by the panel LP is

$$\beta(h) = \frac{\text{Cov}(s_i, \beta_{ih})}{\text{Var}(s_i)}. \tag{13}$$

In other words, panel LPs estimate the slope in a population linear projection of β_{ih} on characteristics s_i including an intercept. Similarly, if $s_i = 1$, the estimand becomes the mean impulse response $\beta(h) = E[\beta_{ih}]$. Note that omitting either X_t or time dummies as controls in a panel LP has the effect of forcing the regression of β_{ih} on s_i through the origin, leading to the estimand $\beta(h) = (E[s_i^2])^{-1} E[s_i \beta_{ih}]$. In order to obtain a rich summary of the heterogeneity in β_{ih}, therefore, the researcher will typically need to explore different choices of s_i or allow s_i to be a vector.[20]

Under the conditions of Proposition 1, $\hat{\beta}(h) = \beta(h) + o_{P_\kappa}(1)$ for any DGP sequence P_κ such that $\kappa/\sqrt{TN} = o(1)$: that is, if the panel LP estimator converges, it is to $\beta(h)$.

This clarifies the sense in which panel LPs can be interpreted when the underlying population of interest features unrestricted heterogeneity in responses to shocks, as in (6). Precisely because we place virtually no restriction on the joint distribution of (θ_i, s_i), the characterization of the estimand is of a nonparametric nature.

[20] For example, the best linear approximation $E^*[\beta_{ih}|s_i] = E[\beta_{ih}] + (\text{Cov}(s_i, \beta_{ih})/\text{Var}(s_i))(s_i - E[s_i])$ requires both estimands or, alternatively, the interaction of X_t with $(1, s_i)$ rather than s_i alone (omitting time effects). If s_i is multivariate, a confidence region constructed on the basis of a time-clustered heteroskedasticity-robust variance estimate enjoys the same uniform validity property of Proposition 1. We illustrate this in our empirical calculations in Section 5.

Uniformly valid inference. Let p be the number of lags of $\hat{s}_i X_t$ included in the controls W_{it}. Both p and h are fixed as $T, N \to \infty$ while $T/N \to 0$.[21] Our main result is that $\hat{C}_\alpha(h)$ has correct coverage for $\beta(h)$ uniformly over κ so long as $h \leq p$:

Proposition 1. *Under Assumptions 1, 2 and 3, for $h \leq p$,*

$$\lim_{T,N\to\infty} \sup_{\kappa} \left| P_\kappa \left(\beta(h) \in \hat{C}_\alpha(h) \right) - (1-\alpha) \right| = 0. \tag{14}$$

Proof. See Appendix A. □

Proposition 1 states that valid inference results from clustering standard errors at the time level, which accounts for cross-sectional dependence induced by omitted aggregate shocks, and from ex-ante including lags of $\hat{s}_i X_t$ as controls, which renders the regression scores uncorrelated. We refer to this strategy as time-clustered lag-augmented heteroskedasticity-robust (t-LAHR) inference. As in Section 2 and as explained below, it is closely linked to inference in time series LPs.

Despite the general error dynamics in (6)–(7), the regression score $\sum_{i=1}^{N} X_t \hat{s}_i \xi_{it}(h, \kappa)$, with $\xi_{it}(h, \kappa)$ the population counterpart to $\hat{\xi}_{it}(h)$ defined in (19), has limited serial correlation. It is an MA(h) process with the first p autocovariances set to zero. Thus, it becomes uncorrelated when $p \geq h$ which is why t-LAHR works. Besides, when $p < h$, the autocovariances stem only from leftover leads of X_t and not from the unobserved macro error Z_t or micro error u_{it}. In fact, they will tend to be small compared to the variance of the score in low-signal (large κ) DGPs or if $\beta_{i\ell}$ decays quickly. We therefore expect t-LAHR inference to have small coverage distortions even for $p < h$; we provide affirmative evidence via simulations in Section 4.

A striking implication of Proposition 1 is that t-LAHR inference remains valid even in the low-signal setting $\kappa/\sqrt{N} \to \infty$ where there is scarcity of information about aggregate shocks in the sample and $\hat{\beta}(h)$ is inconsistent. The uniformity over DGPs with different macro-micro signal-noise obviates the need to take a stand on the κ-regime, which is important because κ is not always consistently estimable.

[21] We regard $T/N \to 0$ as a mild requirement for the empirical applications of reference. It follows from the proof of Proposition 1 that if T/N is not asymptotically negligible (as if taking N as fixed), (14) holds with $\beta(h)$ replaced by the *finite-population* estimand $\tilde{\beta}(h) = (\sum_{i=1}^{N} \hat{s}_i^2)^{-1} \sum_{i=1}^{N} \hat{s}_i \beta_{ih}$.

In contrast, inference based on unit-level clustering of the regression score is not uniformly valid as it tends to severely undercover $\beta(h)$ in high- and moderate-signal regimes. Similarly to Section 2, provided lags of $\hat{s}_i X_t$ are included, unit-level clustering is asymptotically equivalent to not clustering at all, whereas two-way clustering is equivalent to time-level clustering. That is, unit-level clustering is neither necessary nor sufficient for valid inference — yet another implication of X_t being a shock that has no counterpart in a more generic time series setup.

Remark 3 (Proof steps). To establish (14), we decompose the problem into showing (A) asymptotic normality of the score, (B) consistency of the standard error, and (C) negligibility of some remainder terms. We obtain uniformity via the drifting parameter sequence approach (see Andrews, Cheng, and Guggenberger (2020)).

In (A), although the regression score is serially uncorrelated, it contains leads and lags of macro and micro errors. This makes the reverse martingale technique of Montiel Olea and Plagborg-Møller (2021) inapplicable. Instead, using a similar insight to that of Xu (2021), we produce a martingale approximation by rearranging the score so that the leads at time t become the lags at a time in the future of t. See Lemma 1 in Supplemental Appendix B for the details.

In (B) and (C), we rely on direct calculation of uniform bounds. The presence of heterogeneity poses a challenge with no parallel in the time series case. Because of Assumption 3, we can derive many of the bounds by first conditioning on $\{\theta_i, s_i\}_{i=1}^N$, exploiting the connection between conditional and unconditional convergence.

Remark 4 (Synthetic time series). A useful device to interpret panel LPs is the following representation. The residual $\hat{x}_{it}(h)$ in (8) can be written as $\hat{x}_{it}(h) = \hat{s}_i \hat{X}_t(h)$, where $\hat{X}_t(h)$ is the residual from regressing X_t on $X_{t-1}, \dots, X_{t-p}$ and an intercept (on $T - h$ observations).[22]

Then, the panel LP estimator in (9) can be written as

$$\hat{\beta}(h) = \frac{\sum_{t=1}^{T-h} \sum_{i=1}^{N} \hat{s}_i \hat{X}_t(h) Y_{i,t+h}}{\sum_{t=1}^{T-h} \sum_{i=1}^{N} \hat{s}_i \hat{X}_t(h)^2} = \frac{\sum_{t=1}^{T-h} \hat{X}_t(h) \hat{Y}_{t+h}}{\sum_{t=1}^{T-h} \hat{X}_t(h)^2},$$

[22]To see this, note that $\hat{x}_{it}(h)$ is $\hat{s}_i X_t$ minus a linear combination of $\hat{s}_i X_{t-1}, \dots, \hat{s}_i X_{t-p}$ and unit and possibly time indicators which is orthogonal to all of the latter. When W_{it} includes additional controls, the synthetic time series representation is asymptotically but not numerically equivalent.

i.e., the time series LP estimator that regresses cross-sectional regression coefficients $\hat{Y}_{t+h} = (\sum_{i=1}^N \hat{s}_i^2)^{-1} \sum_{i=1}^N \hat{s}_i Y_{i,t+h}$ on X_t controlling for $X_{t-1}, \ldots, X_{t-p}$ and an intercept. The standard error $\hat{\sigma}(h)$ in (10) is also the Eicker–Huber–White standard error calculated on the time series LP residuals $\hat{\xi}_t(h) = (\sum_{i=1}^N \hat{s}_i^2)^{-1} \sum_{i=1}^N \hat{s}_i \hat{\xi}_{it}(h)$. Hence, t-LAHR inference for panel LPs and lag-augmented heteroskedasticity-robust inference for time series LPs are intimately related.

Remark 5 (s_i and precision). This representation is also useful to illuminate the fact that estimation error is of order $T^{-1/2}$ in environments with $\kappa \propto \sqrt{N}$, despite what otherwise looks like a standard panel regression with potentially very rich micro data. We can give interpretable conditions under which variation in s_i affords faster convergence rates. These are akin to s_i being a cross-sectional instrument: we require s_i to correlate with β_{ih} — that is, be relevant for heterogeneity in transmission of X_t at horizon h — but to be orthogonal to all other exposures to aggregate shocks, $(\{\beta_{i\ell}\}_{\ell \neq h}, \gamma_i)$. These conditions seem particularly hard to meet: for each horizon h, a source of variation that is orthogonal to responses at all other horizons is required. (Assumption 3(v) rules this out in our formulation.) In some sense, this reveals an intrinsic trade-off between documenting interesting transmission mechanisms and finding valid instruments for precision.

Remark 6 (t-HAR). In principle, time-clustered HAR inference is a valid alternative to t-LAHR. An analogue to Proposition 1 can be established for a confidence interval that replaces $\hat{V}(h)$ in (11) with the Hansen and Hodrick (1980) variance estimator $\hat{V}(h) + 2\sum_{\ell=p+1}^h \tilde{V}_\ell(h)$ where

$$\tilde{V}_\ell(h) = \frac{1}{(T-h)} \sum_{t=\ell+1}^{T-h} \left(\frac{1}{N} \sum_{i=1}^N \hat{x}_{it}(h) \hat{\xi}_{it}(h) \right) \left(\frac{1}{N} \sum_{i=1}^N \hat{x}_{i,t-\ell}(h) \hat{\xi}_{i,t-\ell}(h) \right),$$

This boils down to $\hat{V}(h)$ for $p \geq h$. Unlike $\hat{V}(h)$, this alternative variance estimator is not guaranteed to be positive semidefinite. Also, t-LAHR inference is simpler to implement and refine, remains tractable over moderate horizons under VAR DGPs (Section 3.3), and performs better in small samples (Section 4).

Remark 7 (State-dependence). In some applications, interest is in the differential pass-through of shocks to responses along an observable (time-varying) state, denoted now s_{it}. Formalizing this requires extending (6)–(7) to allow for time-varying impulse responses:

$$Y_{it} = \mu_i + \sum_{\ell=0}^{\infty} \beta_{it\ell} X_{t-\ell} + v_{it}, \quad v_{it} = \sum_{\ell=0}^{\infty} \gamma_{it\ell} Z_{t-\ell} + \kappa \sum_{\ell=0}^{\infty} \delta_{it\ell} u_{i,t-\ell}.$$

Letting $\hat{s}_{it} = s_{it} - N^{-1} \sum_{j=1}^{N} s_{jt}$, the corresponding panel LP estimator on $\hat{s}_{it} X_t$ retains its interpretation as the slope coefficient of the linear projection $E^* \left[\beta_{ith} \middle| s_{it}\right]$ as long as s_{it} and impulse responses are exogenous with respect to X_t. Although a more detailed exploration is beyond the scope of our paper, the treatment of s_{it} is analogous to that of s_i, and all the results above carry over with little modification. We revisit this in simulations in Section 4 and in our empirical illustration in Section 5.[23]

3.3 Panel VAR model

It is not uncommon in applications that the researcher is interested in responses at an horizon h which is a non-negligible fraction of T. Proposition 1 guarantees exact coverage for short horizons depending on the number of lags of the outcome and shock used as controls. There is, however, one important class of DGPs for which our uniformity result extends to $h \propto T$: the VAR class.

We now assume a panel VAR(p) model (with $p < \infty$):

$$Y_{it} = m_i + \sum_{\ell=1}^{p} A_\ell Y_{i,t-\ell} + \sum_{\ell=0}^{p} B_{i\ell} X_{t-\ell} + C_{i0} Z_t + \kappa D_{i0} u_{it}. \tag{15}$$

If $\sum_{\ell=1}^{p} A_\ell < 1$, as implied by Assumption 3(iv), we can recover the unit-specific parameters $\mu_i, \{\beta_{i\ell}\}, \{\gamma_{i\ell}\}, \{\delta_{i\ell}\}$ from $m_i, \{A_\ell\}, \{B_{i\ell}\}, C_{i0}, D_{i0}$ by inverting the lag polynomial $A(L) = 1 - \sum_{\ell=1}^{p} A_\ell L^\ell$. That is, VAR model (15) is a special case of (6)–(7).

[23] Rambachan and Shephard (2021, Section 3.4) offer a nonparametric characterization of local projection estimands when states are endogenous in a time-series potential outcomes framework; see also Gonçalves, Herrera, Kilian, and Pesavento (forthcoming) for the case where $s_t = \mathbb{1}\{X_t > c\}$.

Assuming that p is known and that W_{it} contains p lags of Y_{it} and $s_i X_t$, the t-LAHR confidence interval $\hat{C}_\alpha(h)$ defined in (12) has uniform validity even for moderately long horizons h exceeding p:

Proposition 2. *Under Assumptions 1, 2 and 3, for some positive constant $\phi < 1$,*

$$\lim_{T,N\to\infty} \sup_{0\le h\le \phi T} \sup_{\kappa} \left| P_\kappa \left(\beta(h) \in \hat{C}_\alpha(h) \right) - (1-\alpha) \right| = 0. \tag{16}$$

Proof. See Appendix A. □

The intuition and proof for Proposition 2 mirror that of Proposition 1. Under VAR model (15) the regression score $\sum_{i=1}^N X_t \hat{s}_i \xi_{it}(h,\kappa)$, with $\xi_{it}(h,\kappa)$ now defined in (20), is serially uncorrelated not just for $h \le p$ but for any h. The basic consequence is that if a low-order VAR model is a reasonable approximation, the t-LAHR inference approach that relies on controlling for a small number of lags of the outcome and shock is robust over long horizons and regardless of the amount of micro noise.[24]

Remark 8 (LP inference when the shock is not observable). Proposition 2 can be read as the panel data counterpart to the result in Montiel Olea and Plagborg-Møller (2021) under stationarity when the shock is directly observable. That parallel implies that if X_t is unavailable but instead we observe $X_t^* = \sum_{\ell=1}^{p-1} \alpha_\ell X_{t-\ell}^* + X_t$ and we run a local projection of $Y_{i,t+h}$ on $s_i X_t^*$ including p lags of Y_{it} and $s_i X_t^*$ in the control vector W_{it}, t-LAHR inference is uniformly valid over h and κ.[25]

Remark 9 (Heterogeneity in VAR coefficients). Model (15) assumes homogeneous coefficients $\{A_\ell\}$. This is common in the microeconometric literature on panel VARs (Arellano, 2003, Chapter 6) but it is not necessary for (16). For example, we can

[24]The results in Montiel Olea et al. (2024) suggest that for a *fixed* horizon h, t-LAHR inference would also remain valid if the VAR model (15) were contaminated by moving averages of Z_t and u_{it} in a $T^{-1/4}$-neighborhood of zero — that is, if the VAR model holds only approximately. The simulation evidence in Section 4 based on DGPs which are not VARs is consistent with this idea.

[25]Lag-augmentation means including at least one more lag than the autoregressive order of X_t^* which is $p-1$. The connection with Montiel Olea and Plagborg-Møller (2021) also suggests that $\hat{C}_\alpha(h)$ is uniformly valid over the VAR parameter space (including unit roots) if a certain condition on uniform non-singularity of the least squares denominator matrix (Assumption 3 in their paper) holds.

establish Proposition 2 in a moderate heterogeneity environment that replaces A_ℓ with $A_{i\ell}$ where $\sup_{1\le i\le N} |A_{i\ell} - A_\ell| = O_p\left(T^{-1/2}\right)$. Proposition 2 can also be established (under slightly different regularity conditions) if we allow for heterogeneity in $\{A_\ell\}$ but we include p *unit-specific* lags of Y_{it} as controls in W_{it}.

3.4 Panel LP-IV and proxy shocks

The most common implementation of panel LPs in empirical work treats the shock of interest as observed. Nevertheless, it is sometimes more realistic to assume there is measurement error in the shock elicitation process. This creates an endogeneity problem that can be dealt with by using the shock measures as instruments for the actual underlying shock (Ramey, 2016; Stock and Watson, 2018).

The researcher observes the outcome Y_{it} and characteristics s_i, but instead of the actual shock X_t she observes an endogenous aggregate state variable $\tilde{X}_t$ and a proxy shock X_t^*. In addition to (6)–(7), we assume

$$\tilde{X}_t = \sum_{\ell=0}^{\infty} b_\ell X_{t-\ell} + \sum_{\ell=0}^{\infty} c_\ell Z_{t-\ell}, \tag{17}$$

$$X_t^* = a_0 X_t + \nu_t, \tag{18}$$

where ν_t is measurement error. We normalize $b_0 = 1$ to fix the scale of the estimand as only relative impulse responses are identified.[26] We also adopt the following:

Assumption 4 (LP-IV).

(i) $a_0 \neq 0$.

(ii) Assumptions 1, 2 and 3 hold with Z_t replaced by (Z_t, ν_t).

(iii) For the same constants C_ℓ and $\underline{C}$ of Assumption 3,

$$|b_\ell| \le C_\ell, \quad |c_\ell| \le C_\ell, \quad \sum_{\ell=0}^{\infty} b_\ell^2 \ge \underline{C}, \quad \sum_{\ell=0}^{\infty} c_\ell^2 \ge \underline{C}.$$

[26]It is straightforward to include intercepts in both (17) and (18). Additionally, as in Section 3.3, we can derive uniformity results with respect to the horizon h by assuming a VAR model in (6), (7) and (17).

Assumption 4(i) is needed for instrument relevance, and we restrict our attention to the strong instrument case where we keep a_0 fixed as $N, T \to \infty$. On the other hand, Assumption 4(ii) implies that ν_t is orthogonal to $\{X_\tau, Z_\tau\}$. This embodies the key lead-lag exogeneity condition requiring X_t^* to be contemporaneously correlated only with X_t, a well-known condition in the time series LP-IV context.[27] Finally, Assumption 4(iii) imposes regularity on the endogenous variable $\tilde{X}_t$.

LP-IV estimation and inference. LP-IV regresses $Y_{i,t+h}$ on $\tilde{\boldsymbol{X}}_t = (\tilde{X}_t, \tilde{X}_{t-1}, \ldots, \tilde{X}_{t-p})'$ using $\boldsymbol{X}_t^* = (X_t^*, X_{t-1}^*, \ldots, X_{t-p}^*)'$ as instruments (both interacted with s_i), controlling for unit and time effects (W_{it} denotes controls). The residualized instrument is

$$\hat{\boldsymbol{x}}_{it}(h) = \hat{s}_i \boldsymbol{X}_t^* - \hat{\boldsymbol{\pi}}(h)' W_{it} = \hat{s}_i \hat{\boldsymbol{X}}_t^*(h),$$

where $\hat{\boldsymbol{X}}_t^*(h) = \boldsymbol{X}_t^* - (T-h)^{-1} \sum_{t=1}^{T-h} \boldsymbol{X}_t^*$. The panel LP-IV estimator $\hat{\boldsymbol{\beta}}^{\text{IV}}(h)$ is then

$$\hat{\boldsymbol{\beta}}^{\text{IV}}(h) = \left(\sum_{t=1}^{T-h} \sum_{i=1}^{N} \hat{\boldsymbol{x}}_{it}(h) \hat{s}_i \tilde{\boldsymbol{X}}_t' \right)^{-1} \sum_{t=1}^{T-h} \sum_{i=1}^{N} \hat{\boldsymbol{x}}_{it}(h) Y_{i,t+h} = \left(\sum_{t=1}^{T-h} \hat{\boldsymbol{X}}_t^*(h) \tilde{\boldsymbol{X}}_t' \right)^{-1} \sum_{t=1}^{T-h} \hat{\boldsymbol{X}}_t^*(h) \hat{Y}_{i,t+h},$$

where $\hat{Y}_{i,t+h}$ is the synthetic outcome defined in Remark 4. Put another way, panel LP-IV admits a synthetic time series LP-IV representation.

The only entry of $\hat{\boldsymbol{\beta}}^{\text{IV}}(h)$ that has interpretation as an estimate of a relative impulse response is $\hat{\beta}_0^{\text{IV}}(h) = e_1' \hat{\boldsymbol{\beta}}^{\text{IV}}(h)$ where e_1 is the first column of I_{p+1}. The remaining entries are necessary for t-LAHR inference to be valid. Given residuals

$$\hat{\xi}_{it}^{\text{IV}}(h) = Y_{i,t+h} - \hat{s}_i \tilde{\boldsymbol{X}}_t' \hat{\boldsymbol{\beta}}^{\text{IV}}(h) - \hat{\eta}^{\text{IV}}(h)' W_{it},$$

we define

$$\hat{\boldsymbol{J}}^{\text{IV}}(h) = \frac{1}{N(T-h)} \sum_{t=1}^{T-h} \sum_{i=1}^{N} \hat{\boldsymbol{x}}_{it}(h) \hat{s}_i \tilde{\boldsymbol{X}}_t', \quad \hat{\boldsymbol{V}}^{\text{IV}}(h) = \frac{1}{(T-h)} \sum_{t=1}^{T-h} \left(\frac{1}{N} \sum_{i=1}^{N} \hat{\boldsymbol{x}}_{it}(h) \hat{\xi}_{it}^{\text{IV}}(h) \right)^2.$$

[27] See, for instance, Stock and Watson (2018, p. 924) and Plagborg-Møller and Wolf (2021, p. 970). The setup can be extended to allow ν_t to be serially correlated and to the case where X_t^* is valid only after conditioning on a set of controls.

The time-clustered heteroskedasticity-robust standard error for $\hat{\beta}_0^{\text{IV}}(h)$ is

$$\hat{\sigma}_0^{\text{IV}}(h) = \left[\frac{1}{(T-h)} \cdot \left(e_1' \hat{\boldsymbol{J}}^{\text{IV}}(h)^{-1}\right) \hat{\boldsymbol{V}}^{\text{IV}}(h) \left(e_1' \hat{\boldsymbol{J}}^{\text{IV}}(h)^{-1}\right)'\right]^{1/2}$$

and the $(1-\alpha)$ confidence interval, $\hat{C}_\alpha^{\text{IV}}(h) = \left[\hat{\beta}_0^{\text{IV}}(h) \pm z_{1-\alpha/2}\hat{\sigma}_0^{\text{IV}}(h)\right]$. Then:

Proposition 3. *Under Assumption 4, for $h \leq p$,*

$$\lim_{T,N\to\infty} \sup_{\kappa} \left| P_\kappa \left(\beta(h) \in \hat{C}_\alpha^{IV}(h)\right) - (1-\alpha)\right| = 0.$$

Proof. See Appendix A. □

Remark 10 (Absence of first-stage heterogeneity). The LP-IV estimand coincides (under the normalization $b_0 = 1$) with the LP estimand (13) despite the presence of heterogeneity. This is far from obvious: under treatment effect heterogeneity, IV estimands are generally (weighted averages of) local average treatment effects (Angrist and Imbens, 1995; Angrist, Imbens, and Graddy, 2000). It is the aggregate-only nature of the first-stage model that underlies this result. This is yet another illustration of the unique setting that we study in this paper.

4 Simulation study

We ran a comprehensive simulation study to verify the finite-sample robustness of the inference procedures analyzed in Section 3. Here we provide a summary and defer additional detail and results to Supplemental Appendix C.

Designs. Our study relies on two different DGPs. The first is the general setup (6)–(7) supplemented with (17)–(18) to cover the endogenous case. We begin by simulating shocks $\{X_t, Z_t, \nu_t, \{u_{it}\}_{i=1}^N\}$ as mutually and serially independent $N(0,1)$ random variables, and by drawing $\{\theta_i, s_i\}_{i=1}^N$ independently across units. To ensure correlation between observed and unobserved heterogeneity we use a technique described in Supplemental Appendix C. We calibrate the distribution of $\{\beta_{i\ell}, \gamma_{i\ell}, \delta_{i\ell}\}$ and the value of $\{b_\ell, c_\ell\}$ to produce realistic degrees of shock persistence.

Given these elements, we generate the inputs for panel LP and LP-IV procedures, namely Y_{it}, X_t, s_i, $\tilde{X}_t$, X_t^*. We also simulate the time-varying covariate $s_{it} = s_i + \zeta_{it}$

(where ζ_{it} is such that s_{it} remains strictly exogenous) to compare panel LPs on $s_i X_t$ and $s_{it} X_t$ — this illustrates the point we made in Remark 7.

The second DGP is the VAR model (15). Again we generate shocks as i.i.d. $N(0,1)$ and we simulate the heterogeneity as detailed in Supplemental Appendix C. When calibrating the VAR parameters $\{A_\ell\}$ we allow the largest AR root to be $1 - c/T$ (we use $c = 5$) to capture the essence of a near non-stationary environment.[28]

The results below are based on $n_{\text{MC}} = 5,000$ Monte Carlo samples. Motivated by our survey of the empirical literature, we look at designs with $T = 30$ and $T = 100$. We set $N = 1,000$ (although we also considered experiments with larger N) and we let κ take values consistent with $\bar{R}^2(\kappa) \in \{0.99, 0.66, 0.33\}$ as defined in (4). As a reference, $\bar{R}^2(\kappa) = 0.66$ corresponds to the one-third of aggregate fluctuations explained by micro shocks suggested by Gabaix (2011) for GDP growth, which we take as moderate signal-to-noise.

Inference procedures. We compare t-LAHR inference with one-way (1W), two-way (2W), and Driscoll-Kraay (DK98) inferences. These are implemented without lag augmentation, as is common practice. For illustrative purposes, we also include t-HR (the non-lag-augmented counterpart to t-LAHR) and t-HAR alternatives.

For t-LAHR inference we use the simple lag selection rule $p = \min\{h, (T-h)^{1/3}\}$ (except in the VAR DGP where p is known) and we apply the finite-sample refinement advocated by Imbens and Kolesár (2016). The lag selection rule is motivated by Xu (2023, Section 3.3) for fixed h and provides fairly generous lag augmentation. For t-HAR inference we use the equally-weighted cosine approach (Müller, 2004) with the choice of tuning parameter recommended in Lazarus et al. (2018).

Results. In Figure 1, we report pointwise coverage rates for horizons $0 \le h \le 0.25T$ with $T = 100$. These correspond to 90% confidence intervals for panel LP and LP-IV using s_i to interact the aggregate shock. Panels (a)-to-(c) display LP while (d)-to-(f) display LP-IV in the general DGP; panels (g)-to-(i) display LP in the VAR DGP.

[28] We also considered experiments where (a) in the first DGP shocks are conditionally heteroskedastic, and (b) in the VAR DGP we have unit-specific VAR parameters $\{A_{i\ell}\}$. We did not find any major difference with what we report here.

Figure 1 suggests four takeaways. First, t-LAHR performs best in all scenarios, with coverage close to the nominal rate even in low-signal cases and for horizons h well beyond p. Its mean absolute coverage distortion never exceeds 2%, whereas it is between 4% and 7% for the second best option (t-HAR) under high signal.

Second, estimating the long-run variance of the score (instead of lag augmenting) can be challenging with small T. This is particularly true for DK98 which relies on Newey–West. Interestingly, these approaches do better in low-signal DGPs where, as mentioned before, there is less to gain from doing HAC.

Third, one-way clustering is very sensitive to $\bar{R}^2(\kappa)$, suffering severe distortions in intermediate- and high-$\bar{R}^2(\kappa)$ cases. What is more, it is outperformed by t-LAHR even if micro shocks explain the majority of aggregate variation. This is consistent with the view that 1W guards against the wrong type of correlation in the score.

Finally, two-way clustering is usually close to t-HR, its non-i-clustered version; another indication that there is no clear advantage in clustering by units. In fact, in certain occasions (mainly low-signal and near non-stationary designs), 2W gives worse inferences than t-HR or 1W alone. This is possibly due to the non-standard behavior of variance estimators when there are micro (near) unit roots.

Identical takeaways emerge in experiments where we substitute s_i with either 1 or s_{it} (Supplemental Appendix C), and with a sample size $T = 30$ (Figure 2).

In sum, the small-sample evidence reinforces many of our theoretical results. It shows that the large-sample approximations of Section 3 provide reliable guidance for understanding estimation and inference with aggregate shocks. Furthermore, it illustrates the practical relevance of achieving uniformity with respect to κ, and it delivers a clear methodological prescription: t-LAHR inference.

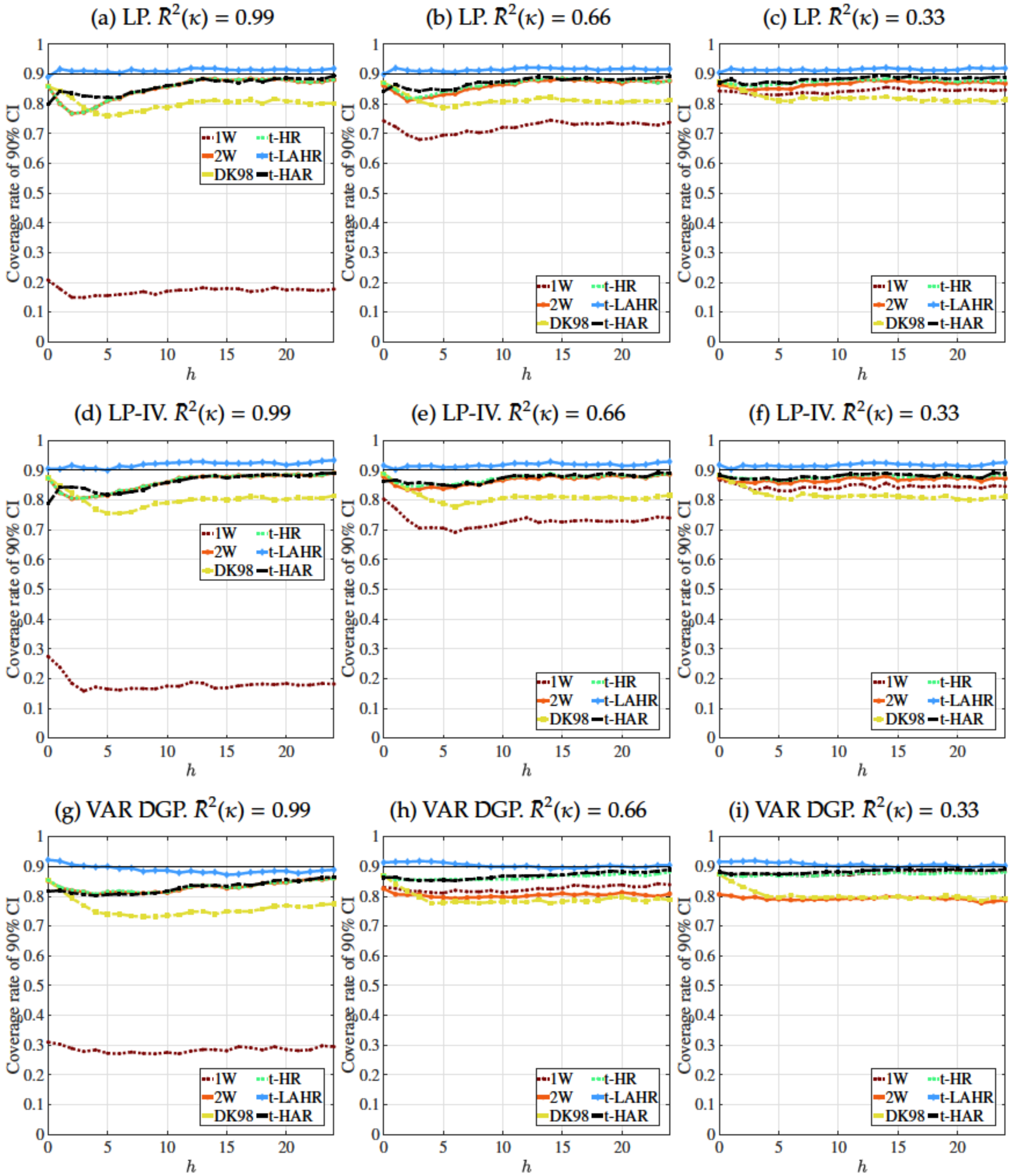

FIGURE 1. Coverage rates of 90% confidence intervals for $T = 100$.

Note: 1W refers to one-way (unit-level) clustering, 2W to two-way clustering, DK98 to Driscoll–Kraay, and t-HR/t-LAHR/t-HAR to the time-level clustering approaches discussed in the text.

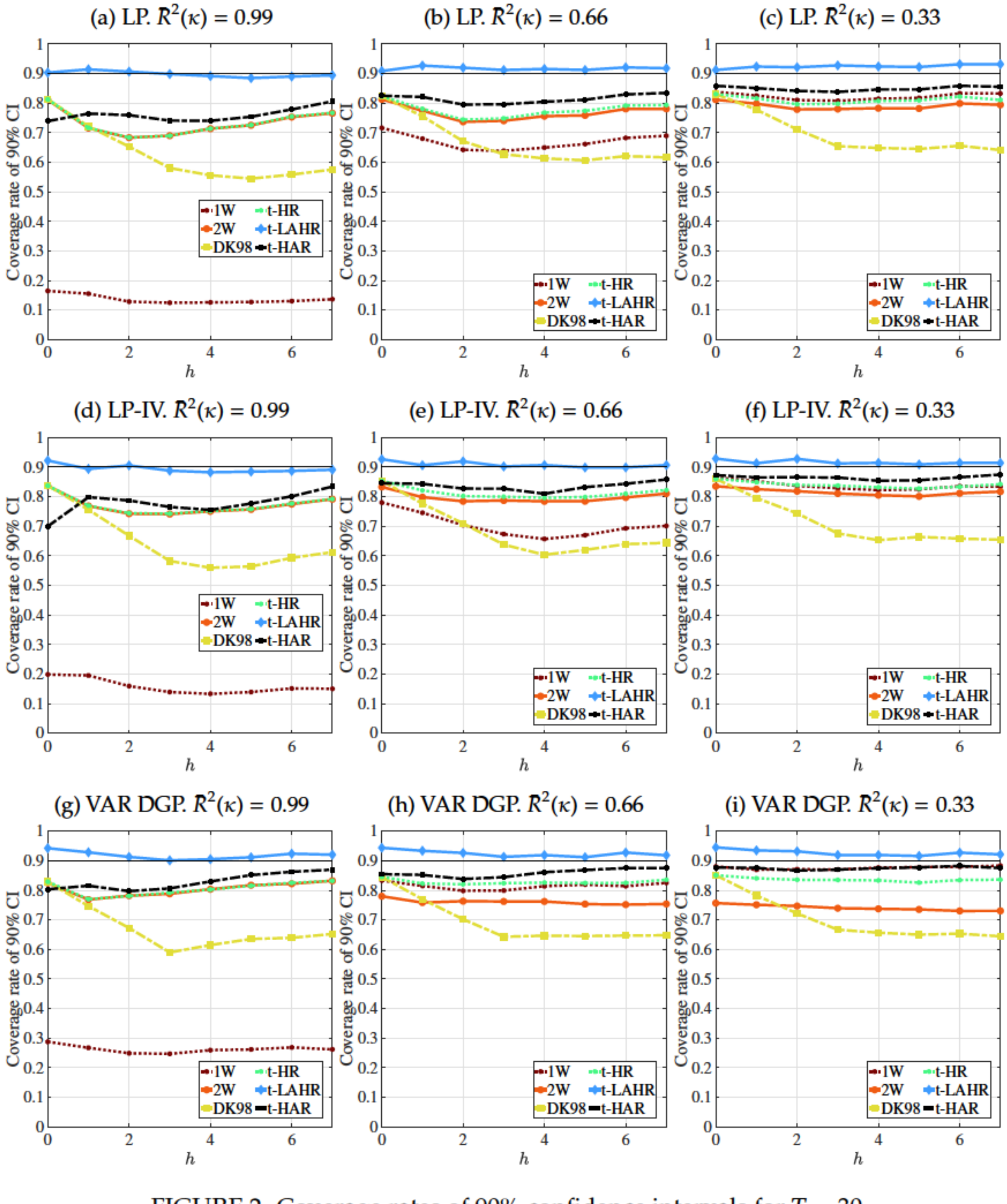

FIGURE 2. Coverage rates of 90% confidence intervals for $T = 30$.

Note: 1W refers to one-way (unit-level) clustering, 2W to two-way clustering, DK98 to Driscoll–Kraay, and *t*-HR/*t*-LAHR/*t*-HAR to the time-level clustering approaches discussed in the text.

5 Empirical illustration

We now discuss an empirical exercise that demonstrates the applicability of our methods in a setup featuring time-varying s_{it} and unbalanced panels, and compares our practical recommendation to popular alternatives. The exercise is motivated by the burgeoning literature on the role played by firm heterogeneity and financial frictions in the propagation of monetary policy.

Data and background. Quantifying firm-level responses to exogenous changes in policy is a key empirical goal as it is informative on the mechanisms through which monetary policy operates. For instance, Crouzet and Mehrotra (2020) focus on the role of firm size for investment response heterogeneity, finding larger (albeit noisy) responses for smaller firms; Ottonello and Winberry (2020) instead focus on default risk, finding larger responses for less risky companies.

For our empirical analysis, we construct a dataset similar to the latter based on Compustat and high-frequency identified monetary policy shocks (Gurkaynak, Sack, and Swanson, 2005; Gorodnichenko and Weber, 2016). This results in an unbalanced panel for the period 1990Q1–2010Q4 with observations on firm-level investment, size, and leverage.[29] In total, there are $T = 80$ quarters and $N = 4,187$ individual companies which, net of missing data, amount to 235,233 observations.

We consider regressions of cumulative investment changes $Y_{i,t+h} = \log(k_{i,t+h}/k_{i,t-1})$ (k_{it} being the capital stock) on policy shocks X_t interacted with s_{it}, a vector containing size, leverage, and their product. From Section 3, we know that under unrestricted heterogeneity the population counterpart is the linear projection of firm-level impulse responses on s_{it}. Thus, including size and leverage together (as well as their interaction) in s_{it} is a way to enrich the linear approximation.

[29]We use the paper's replication code to build the data and we verify that we can replicate the original results, with minor numerical differences that can be attributed to revisions in input data. Firm size is measured by the value of total assets held by a company while leverage is its debt-to-assets ratio. We have also tried the distance-to-default measure in Ottonello and Winberry (2020) with qualitatively similar results.

Synthetic time series representation. A fundamental insight of our paper is that the synthetic time series form of the microdata is a sufficient statistic for the panel LP; a low dimensional representation of a highly complex, unbalanced dataset.[30]

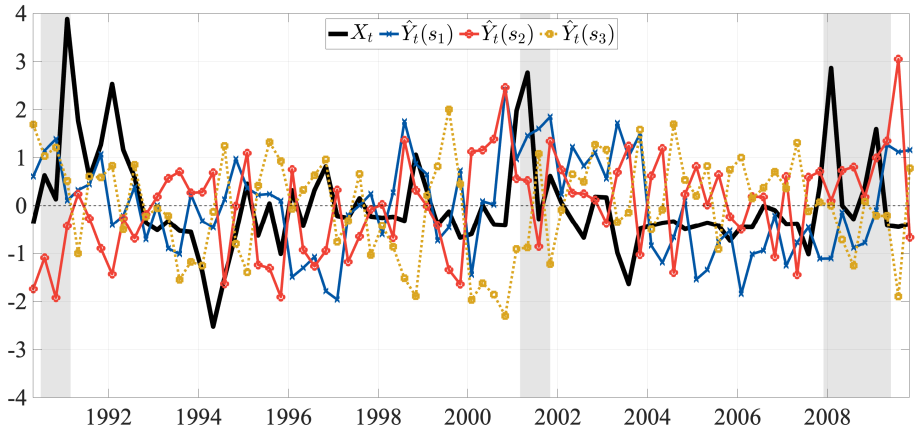

FIGURE 3. Synthetic time series representations.

Note: Grey areas are NBER-dated recessions. s_1 is size, s_2 is leverage and s_3 is the interaction. X_t and $\hat{Y}_t$ are standardized to zero mean and unit variance; $X_t > 0$ indicates a surprise cut in the Fed Funds rate.

Figure 3 displays it for the three components of s_{it}. It is clear that movements in synthetic outcomes concurrent with surprise cuts in policy rates, mostly around recessions, are the main source of identification. There is also substantial variation in synthetic outcomes unrelated to X_t, indicating the presence of omitted aggregate or non-negligible idiosyncratic shocks — the central premises of our paper.

Estimation and inference method comparison. Figure 4 reports point estimates and 90% confidence intervals for the coefficient on each entry of $s_{it}X_t$ at different

[30] Remark 4 generalizes as follows. Let $d_{it} = 0$ indicate a missing observation with $d_{it} = 1$ otherwise. Abstracting from controls, the panel local projection estimator with a time-varying s_{it} is

$$\hat{\beta}(h) = \frac{\sum_{t=1}^{T-h} \sum_{i=1}^{N} d_{it} s_{it} X_t Y_{i,t+h}}{\sum_{t=1}^{T-h} \sum_{i=1}^{N} d_{it} s_{it}^2 X_t^2} = \frac{\sum_{t=1}^{T-h} \omega_t X_t \hat{Y}_{t+h}}{\sum_{t=1}^{T-h} \omega_t X_t^2},$$

where $\omega_t = \sum_{i=1}^{N} d_{it} s_{it}^2$ and $\hat{Y}_{t+h} = (\sum_{i=1}^{N} s_{it}^2)^{-1} \sum_{i=1}^{N} s_{it} Y_{i,t+h}$. This is a weighted least squares regression of slope coefficients $\hat{Y}_{t+h}$ on X_t. Note that if $s_{it} = 1$ the weights boil down to the number of non-missing observations $\omega_t = \sum_{i=1}^{N} d_{it}$, as intuition suggests. Our theory applies with data missing at random.

horizons.[31] According to the t-LAHR intervals, the evidence favors the hypothesis that larger and less indebted firms respond less to monetary policy shocks, with the size effect more persistent and not much interaction between the two.

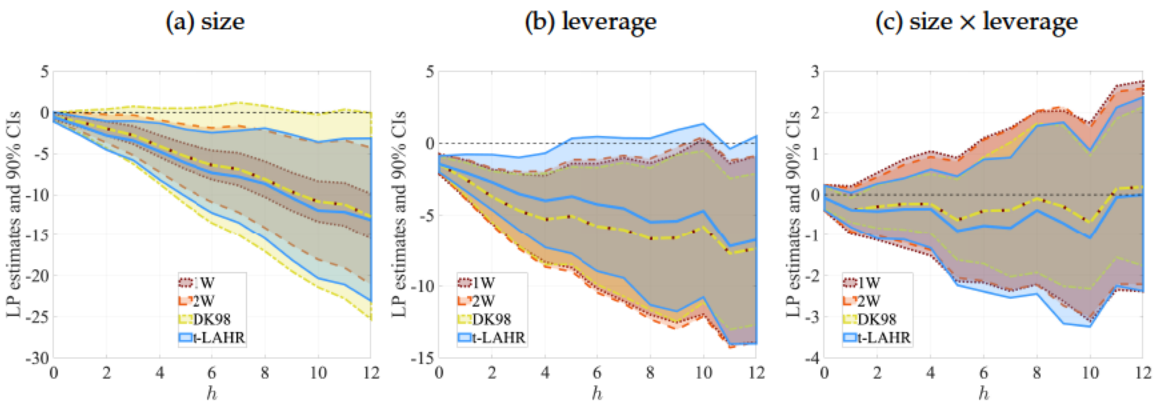

FIGURE 4. Point estimates and 90% confidence intervals.

Note: The procedures are one-way unit-level clustering (1W, dotted line), two-way clustering (2W, dashed line), Driscoll-Kraay (DK98, dash dotted line), and t-LAHR (solid line).

From an applied point of view, the main message is that popular methods can deviate significantly from the (asymptotically robust) t-LAHR method. For example, one-way clustering produces intervals that are too short (panel (a)) and too wide (panel (c)).[32] Two-way clustering is close to t-LAHR but produces different conclusions in panel (b) and may be unreliable in high-persistence, low-signal setups. Finally, Driscoll-Kraay intervals can be misleading with $T = 80$. In fact, they lead to exactly the opposite conclusions about the role of size and leverage.

[31] The panel local projections include as controls unit and time effects, lagged firm-level sales growth, and both lagged GDP growth and lagged unemployment interacted with s_{it}. For t-LAHR inference we include p lags of Y_{it} and $s_{it}X_t$. We limit $p = \min\{h, 2\}$ to discipline the number of regressors in view of the dimension of s_{it}. One-way, two-way and Driscoll-Kraay are implemented without lag-augmentation.

[32] This can happen even in the same exercise because the estimation errors of different coefficients load differently on the macro and micro components of the regression score. Figure 4 suggests the size coefficient is driven by the macro component and the other coefficients by the micro component.

6 Conclusion

The use of micro data to answer macro questions offers an exciting avenue to study how agents respond to economy-wide policies. Possibilities include a better understanding of the transmission of shocks and the nature of heterogeneity.

Challenges are ubiquitous too. We propose a disciplined approach to uncertainty quantification when both aggregate and idiosyncratic shocks coexist and interest is in parameters identified solely by macro shocks. One such scenario is the estimation of impulse responses to macro shocks when rich micro data and a measurement of the shock of interest are available. Despite the complex environment, inference is simple and robust: it involves lag augmentation and clustering at the time level, and is valid regardless of the relative signal of macro shocks in the microdata.

Our basic framework generalizes beyond the empirical applications we have focused on. Other, related literatures where identification comes from randomness in group level shocks include regional-exposure and shift-share designs. In fact, impulse responses are sometimes an object of interest too — see, for instance, the literature on cross-sectional fiscal multipliers (Chodorow-Reich, 2019).

We also leave some interesting dimensions for future research. Quantifying signal-to-noise (perhaps a lower bound) seems relevant in settings where uniform inference is not possible; we expect that these issues become more salient as macroeconomists embrace the use of microdata to sharpen identification (Nakamura and Steinsson, 2018). On a different note, strong persistence of micro-level shocks are likely a feature of many datasets, and these are only captured in an indirect sense by our signal-to-noise device. Formalizing the idea of (possibly heterogeneous) non-stationarities along these lines seems promising and full of empirical content. Finally, extensions to simultaneous inference over impulse response horizons could be made building on Jordà (2009) and Montiel Olea and Plagborg-Møller (2019).

A Proofs

Proposition 1

Let $\tilde{\beta}(h) = \left(\sum_{i=1}^N \hat{s}_i^2\right)^{-1} \sum_{i=1}^N \hat{s}_i \beta_{ih}$ be the coefficient in the (infeasible) regression of β_{ih} on $\hat{s}_i$ — the finite-population counterpart to $\beta(h)$. Also, define

$$\xi_{it}(h,\kappa) = \sum_{\ell=0}^{\infty} \left(\iota_\ell(h)\beta_{i\ell}X_{t+h-\ell} + \gamma_{i\ell}Z_{t+h-\ell} + \kappa\delta_{i\ell}u_{i,t+h-\ell}\right), \tag{19}$$

$$\xi_t(h,\kappa) = \frac{1}{N}\sum_{i=1}^N \hat{s}_i \xi_{it}(h,\kappa) = \sum_{\ell=0}^{\infty}\left(\iota_\ell(h)\bar{\beta}_\ell X_{t+h-\ell} + \bar{\gamma}_\ell Z_{t+h-\ell} + \frac{\kappa}{N}\sum_{i=1}^N \hat{s}_i\delta_{i\ell}u_{i,t+h-\ell}\right)$$

where $\iota_\ell(h) = 1 - \mathbb{1}\{h \le \ell \le h+p\}$, $\bar{\beta}_\ell = N^{-1}\sum_{i=1}^N \hat{s}_i\beta_{i\ell}$ and $\bar{\gamma}_\ell = N^{-1}\sum_{i=1}^N \hat{s}_i\gamma_{i\ell}$. Finally, let $V(h,\kappa) = \mathrm{Var}_\kappa\left(X_t\xi_t(h,\kappa)\middle|\{\theta_i, s_i\}_{i=1}^N\right)$.

Proof of Propositions 1. Let $\sum_{i,t}$ denote summation over $1 \le t \le T-h$ and $1 \le i \le N$. For any $\psi \in \mathbb{R}^d$,

$$\begin{aligned}
\left(\sum_{i,t}\hat{x}_{it}(h)^2\right)\left(\hat{\beta}(h) - \tilde{\beta}(h)\right) &= \sum_{i,t}\hat{x}_{it}(h)\left(Y_{i,t+h} - \tilde{\beta}(h)\hat{s}_i X_t - \psi' W_{it}\right) \\
&= \sum_{i,t}\hat{s}_i X_t\left(Y_{i,t+h} - \beta_{ih}X_t - \psi' W_{it}\right) \\
&\quad - \sum_{i,t}(\hat{\pi}(h)' W_{it})\left(Y_{i,t+h} - \tilde{\beta}(h)\hat{s}_i X_t - \psi' W_{it}\right).
\end{aligned}$$

The first line uses $\sum_{i,t}\hat{x}_{it}(h)^2 = \sum_{i,t}\hat{x}_{it}(h)\hat{s}_i X_t$ and $\sum_{i,t}\hat{x}_{it}(h) W_{it} = 0_{d\times 1}$ (to introduce ψ). The second line uses $\hat{x}_{it}(h) = \hat{s}_i X_t - \hat{\pi}(h)' W_{it}$ and $\sum_{i,t}\hat{s}_i X_t(\tilde{\beta}(h)\hat{s}_i X_t - \beta_{ih}X_t) = 0$.

We can choose ψ so that

$$\sum_{i,t}\hat{s}_i X_t\left(Y_{i,t+h} - \beta_{ih}X_t - \psi' W_{it}\right) = \sum_{i,t}\hat{s}_i X_t \xi_{it}(h,\kappa) = N\sum_{t=1}^{T-h} X_t \xi_t(h,\kappa).$$

Here, W_{it} consists of p lags of $\hat{s}_i X_t$, unit indicators, and (possibly) time indicators (so that $d = p + N + T$). To choose ψ, we set the coefficient on $\hat{s}_i X_{t-\ell}$ to $\tilde{\beta}(h+\ell) = \left(\sum_{i=1}^N \hat{s}_i^2\right)^{-1}\sum_{i=1}^N \hat{s}_i\beta_{i,h+\ell}$, the coefficient on the unit-i indicator to μ_i, and the coefficients on time indicators to zero. Moreover, $\hat{\pi}(h)'W_{it} = \hat{s}_i(X_t - \hat{X}_t(h))$ with $\hat{X}_t(h)$ the residual from a regression of X_t on $X_{t-1}, \ldots, X_{t-p}$ and an intercept. Then,

$$\sum_{i,t} (\hat{\pi}(h)'W_{it}) \left(Y_{i,t+h} - \tilde{\beta}(h)\hat{s}_i X_t - \psi' W_{it} \right) = \sum_{i,t} (\hat{\pi}(h)'W_{it}) \xi_{it}(h, \kappa).$$

It follows that the standardized estimation error can be written as

$$\begin{aligned} \frac{\hat{\beta}(h) - \tilde{\beta}(h)}{\hat{\sigma}(h)} &= \frac{\sum_{t=1}^{T-h} \sum_{i=1}^{N} \hat{x}_{it}(h)(Y_{i,t+h} - \tilde{\beta}(h)\hat{x}_{it}(h))}{N\sqrt{(T-h)\hat{V}(h)}} \\ &= \sqrt{\frac{V(h,\kappa)}{\hat{V}(h)}} \times \left(\frac{\sum_{t=1}^{T-h} X_t \xi_t(h,\kappa)}{\sqrt{(T-h)V(h,\kappa)}} + R_T(h,\kappa) \right) \end{aligned}$$

where the remainder term is

$$R_T(h,\kappa) = -\frac{\sum_{t=1}^{T-h} \sum_{i=1}^{N} (\hat{\pi}(h)'W_{it})\, \xi_{it}(h,\kappa)}{N\sqrt{(T-h)V(h,\kappa)}}.$$

To establish our uniform approximation we exploit drifting parameter sequences (see Andrews et al. (2020) for formal results connecting the two). For simplicity we index everything to T, including $N = N_T$. We show that for any $\{\kappa_T\}$, as $T \to \infty$,

(A) $\{(T-h)V(h,\kappa_T)\}^{-1/2} \sum_{t=1}^{T-h} X_t \xi_t(h,\kappa_T) \xrightarrow[P_{\kappa_T}]{\text{d}} N(0,1)$,

(B) $\hat{V}(h)/V(h,\kappa_T) \xrightarrow[P_{\kappa_T}]{\text{p}} 1$,

(C) $R_T(h,\kappa_T) \xrightarrow[P_{\kappa_T}]{\text{p}} 0$.

Hence, for any such $\{\kappa_T\}$,

$$\frac{\hat{\beta}(h) - \tilde{\beta}(h)}{\hat{\sigma}(h)} \xrightarrow[P_{\kappa_T}]{\text{d}} N(0,1).$$

We establish (A), (B) and (C) in Lemmas 1, 2 and 3 in Supplemental Appendix B. Now, Assumptions 1(ii) and 3(iv) imply $\tilde{\beta}(h) - \beta(h) = O_{P_{\kappa_T}}\left(N^{-1/2}\right)$ whereas Lemma 2 implies $\min\{1, \kappa_T^{-1}\}\hat{\sigma}(h) = O_{P_{\kappa_T}}\left((T-h)^{-1/2}\right)$. Since $T/N \to 0$,

$$\frac{(\hat{\beta}(h) - \beta(h))}{\hat{\sigma}(h)} = \frac{(\hat{\beta}(h) - \tilde{\beta}(h))}{\hat{\sigma}(h)} + o_{P_{\kappa_T}}(1)$$

and the result follows. $\square$

Proposition 2

Define

$$\xi_{it}(h,\kappa) = \sum_{\ell=0}^{h} \left(\iota_\ell(h)\beta_{i\ell}X_{t+h-\ell} + \gamma_{i\ell}Z_{t+h-\ell} + \kappa\delta_{i\ell}u_{i,t+h-\ell} \right), \tag{20}$$

$$\xi_t(h,\kappa) = \frac{1}{N}\sum_{i=1}^{N} \hat{s}_i \xi_{it}(h,\kappa) = \sum_{\ell=0}^{h} \left(\iota_\ell(h)\bar{\beta}_\ell X_{t+h-\ell} + \bar{\gamma}_\ell Z_{t+h-\ell} + \frac{\kappa}{N}\sum_{i=1}^{N} \hat{s}_i \delta_{i\ell} u_{i,t+h-\ell} \right),$$

and, as before, let $V(h,\kappa) = \mathrm{Var}_\kappa \left(X_t \xi_t(h,\kappa) \middle| \{\theta_i, s_i\}_{i=1}^N \right)$. By recursive substitution,

$$Y_{i,t+h} = m_i(h) + \sum_{\ell=1}^{p} (A_\ell(h) Y_{i,t-\ell} + B_{i\ell}(h) X_{t-\ell}) + \beta_{ih} X_t + \xi_{it}(h,\kappa),$$

for some $m_i(h), \{A_\ell(h)\}, \{B_{i\ell}(h)\}$ that depend on the VAR parameters $m_i, \{A_\ell\}, \{B_{i\ell}\}$.

Proof of Proposition 2. We follow exactly the same steps as for Proposition 1. The control vector W_{it} includes p lags of Y_{it} and $\hat{s}_i X_t$ in addition to unit and time effects. In the step where we choose ψ, we set the coefficient on $Y_{i,t-\ell}$ to $A_\ell(h)$, the coefficient on $\hat{s}_i X_{t-\ell}$ to $\tilde{B}_\ell(h) = \left(\sum_{i=1}^N \hat{s}_i^2\right)^{-1} \sum_{i=1}^N \hat{s}_i B_{i\ell}(h)$, the coefficient on the unit-i indicator to $m_i(h)$, and the coefficients on time indicators to zero.

The standardized estimation error can then be written as

$$\frac{\hat{\beta}(h) - \tilde{\beta}(h)}{\hat{\sigma}(h)} = \sqrt{\frac{V(h,\kappa)}{\hat{V}(h)}} \times \left(\frac{\sum_{t=1}^{T-h} X_t \xi_t(h,\kappa)}{\sqrt{(T-h)V(h,\kappa)}} + R_T(h,\kappa) \right)$$

where the remainder term is now

$$R_T(h,\kappa) = -\frac{\sum_{t=1}^{T-h}\sum_{i=1}^{N} (\hat{\pi}(h)' W_{it}) \left[(\beta_{ih} - \tilde{\beta}(h)\hat{s}_i) X_t + \sum_{\ell=1}^{p} (B_{i\ell}(h) - \tilde{B}_\ell(h)\hat{s}_i) X_{t-\ell} + \xi_{it}(h,\kappa) \right]}{N\sqrt{(T-h)V(h,\kappa)}}.$$

Let $\phi < 1$. In contrast to Proposition 1, instead of a single drifting parameter we now have two. We show that for any $\{h_T, \kappa_T\}$ such that $h_T \leq \phi T$,

(A) $\{(T-h_T)V(h_T,\kappa_T)\}^{-1/2} \sum_{t=1}^{T-h_T} X_t \xi_t(h_T,\kappa_T) \xrightarrow[P_{\kappa_T}]{d} N(0,1)$,

(B) $\hat{V}(h_T)/V(h_T,\kappa_T) \xrightarrow[P_{\kappa_T}]{\text{p}} 1$,

(C) $R_T(h_T,\kappa_T) \xrightarrow[P_{\kappa_T}]{\text{p}} 0$.

We prove (A), (B) and (C) in Lemmas 8, 9 and 10 in Supplemental Appendix B. The rest of the argument is identical to that of Proposition 1. □

Proposition 3

Using (17), substitute $\tilde{X}_t, \tilde{X}_{t-1}, \ldots, \tilde{X}_{t-p}$ in succession into (6)–(7) to obtain

$$Y_{i,t+h} = \mu_i + \beta_{ih}\tilde{X}_t + \sum_{\ell=1}^{p} \tilde{\eta}_{i\ell}\tilde{X}_{t-\ell} + \xi_{it}(h,\kappa),$$

$$\xi_{it}(h,\kappa) = \sum_{\ell=0}^{\infty} \left(\iota_\ell(h)\tilde{\beta}_{i\ell}X_{t+h-\ell} + \tilde{\gamma}_{i\ell}Z_{t+h-\ell} + \kappa\delta_{i\ell}u_{i,t+h-\ell}\right),$$

for some coefficients $\{\tilde{\eta}_{i\ell}\}$, $\{\tilde{\beta}_{i\ell}\}$, $\{\tilde{\gamma}_{i\ell}\}$ that depend on $\{\beta_{i\ell}\}$, $\{\gamma_{i\ell}\}$, $\{b_\ell\}$, $\{c_\ell\}$ and satisfy the bound conditions in Assumption 3 for a suitable choice of C_ℓ and $\underline{C}$. Also define $\tilde{\boldsymbol{\beta}}(h) = \left(\sum_{i=1}^N \hat{s}_i^2\right)^{-1} \sum_{i=1}^N \hat{s}_i \boldsymbol{\beta}_{ih}$ with $\boldsymbol{\beta}_{ih} = (\beta_{ih}, \tilde{\eta}_{i1}, \ldots, \tilde{\eta}_{ip})'$, $\xi_t(h,\kappa) = N^{-1}\sum_{i=1}^N \hat{s}_i\xi_{it}(h,\kappa)$ and $\boldsymbol{V}(h,\kappa) = \mathrm{Var}_\kappa\left(\boldsymbol{X}_t^*\xi_t(h,\kappa)\middle|\{\theta_i, s_i\}_{i=1}^N\right)$.

Proof of Proposition 3. Following similar steps to the derivation in Proposition 1, let $\sum_{i,t}$ denote summation over $1 \le t \le T-h$ and $1 \le i \le N$. For any ψ,

$$\left(\sum_{i,t} \hat{\boldsymbol{x}}_{it}(h)\hat{s}_i\tilde{\boldsymbol{X}}_t'\right)\left(\hat{\boldsymbol{\beta}}^{\mathrm{IV}}(h) - \tilde{\boldsymbol{\beta}}(h)\right) = \sum_{i,t} \hat{s}_i\boldsymbol{X}_t^*\left(Y_{i,t+h} - \tilde{\boldsymbol{X}}_t'\tilde{\boldsymbol{\beta}}_{ih} - \psi' W_{it}\right) - \left(\frac{\sum_{t=1}^{T-h}\boldsymbol{X}_t^*}{(T-h)}\right)\sum_{i,t}\hat{s}_i\left(Y_{i,t+h} - \hat{s}_i\tilde{\boldsymbol{X}}_t'\tilde{\boldsymbol{\beta}}(h) - \psi' W_{it}\right).$$

Note W_{it} includes unit and (possibly) time effects. To choose ψ, set the coefficient on the unit-i indicator to μ_i and the coefficients on time indicators to zero, so that

$$\sum_{i,t}\hat{s}_i\boldsymbol{X}_t^*\left(Y_{i,t+h} - \tilde{\boldsymbol{X}}_t'\tilde{\boldsymbol{\beta}}_{ih} - \psi' W_{it}\right) = N\sum_{t=1}^{T-h}\boldsymbol{X}_t^*\xi_t(h,\kappa),$$

$$\left(\frac{\sum_{t=1}^{T-h} \boldsymbol{X}_t^*}{(T-h)}\right)\sum_{i,t} \hat{s}_i \left(Y_{i,t+h} - \hat{s}_i \tilde{\boldsymbol{X}}_t' \tilde{\boldsymbol{\beta}}(h) - \psi' W_{it}\right) = \left(\frac{\sum_{t=1}^{T-h} \boldsymbol{X}_t^*}{(T-h)}\right)\sum_{t=1}^{T-h} \xi_t(h,\kappa).$$

Thus, the standardized estimation error can be written as

$$\frac{\hat{\beta}_0^{\mathrm{IV}}(h) - \tilde{\beta}(h)}{\hat{\sigma}_0^{\mathrm{IV}}(h)} = \sqrt{\frac{\left(e_1'\boldsymbol{J}^{-1}\right)\boldsymbol{V}(h,\kappa)\left(e_1'\boldsymbol{J}^{-1}\right)'}{\left(e_1'\hat{\boldsymbol{J}}^{\mathrm{IV}}(h)^{-1}\right)\hat{\boldsymbol{V}}^{\mathrm{IV}}(h)\left(e_1'\hat{\boldsymbol{J}}^{\mathrm{IV}}(h)^{-1}\right)'}} \times\left(\frac{\left(e_1'\hat{\boldsymbol{J}}^{\mathrm{IV}}(h)^{-1}\right)\sum_{t=1}^{T-h}\boldsymbol{X}_t^*\tilde{\xi}_t(h,\kappa)}{\sqrt{(T-h)\left(e_1'\boldsymbol{J}^{-1}\right)\boldsymbol{V}(h,\kappa)\left(e_1'\boldsymbol{J}^{-1}\right)'}} + R_T(h,\kappa)\right)$$

where $\boldsymbol{J} = (N^{-1}\sum_{i=1}^N \hat{s}_i^2)E\left[\boldsymbol{X}_t^*\tilde{\boldsymbol{X}}_t'\right]$ and the remainder term is

$$R_T(h,\kappa) = -\frac{\left\{(T-h)^{-1}\left(e_1'\hat{\boldsymbol{J}}^{\mathrm{IV}}(h)^{-1}\right)\sum_{t=1}^{T-h}\boldsymbol{X}_t^*\right\}\sum_{t=1}^{T-h}\xi_t(h,\kappa)}{\sqrt{(T-h)\left(e_1'\boldsymbol{J}^{-1}\right)\boldsymbol{V}(h,\kappa)\left(e_1'\boldsymbol{J}^{-1}\right)'}}.$$

As in Proposition 1, we show that for any $\{\kappa_T\}$ and $\boldsymbol{\lambda} \neq 0_{(p+1)\times 1}$

(A) $\{(T-h)\boldsymbol{\lambda}'\boldsymbol{V}(h,\kappa_T)\boldsymbol{\lambda}\}^{-1/2}\sum_{t=1}^{T-h}\boldsymbol{\lambda}'\boldsymbol{X}_t^*\xi_t(h,\kappa_T) \xrightarrow[P_{\kappa_T}]{\mathrm{d}} N(0,1)$,

(B) $\left(\boldsymbol{\lambda}'\hat{\boldsymbol{V}}^{\mathrm{IV}}(h)\boldsymbol{\lambda}\right)/(\boldsymbol{\lambda}'\boldsymbol{V}(h,\kappa_T)\boldsymbol{\lambda}) \xrightarrow[P_{\kappa_T}]{\mathrm{p}} 1$ and $\hat{\boldsymbol{J}}^{\mathrm{IV}}(h) \xrightarrow[P_{\kappa_T}]{\mathrm{p}} \boldsymbol{J}$,

(C) $R_T(h,\kappa_T) \xrightarrow[P_{\kappa_T}]{\mathrm{p}} 0$.

The technical steps for (A), (B), and (C) are stated in Lemmas 11, 12 and 13 in Supplemental Appendix B. The rest of the argument is as in Proposition 1. □

References

Anderson, G. and A. Cesa-Bianchi (2024): "Crossing the Credit Channel: Credit Spreads and Firm Heterogeneity." *American Economic Journal: Macroeconomics*, 16, 417–446.

Andrews, D., X. Cheng, and P. Guggenberger (2020): "Generic results for establishing the asymptotic size of confidence sets and tests," *Journal of Econometrics*, 218, 496–531.

Andrews, D. K. (2005): "Cross-section regression with common shocks." *Econometrica*, 73, 1551–1585.

Angrist, J. D. and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association*, 90, 431–442.

Angrist, J. D., G. W. Imbens, and K. Graddy (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish." *Review of Economic Studies*, 67, 499–527.

Arellano, M. (2003): *Panel Data Econometrics*, Oxford University Press.

Arkhangelsky, D. and V. Korovkin (2023): "On Policy Evaluation with Aggregate Time-Series Shocks." Working Paper arXiv:1905.13660.

Chodorow-Reich, G. (2019): "Geographic Cross-Sectional Fiscal Multipliers: What Have We Learned?" *American Economic Journal: Economic Policy*, 11, 1–34.

Crouzet, N. and N. R. Mehrotra (2020): "Small and Large Firms over the Business Cycle." *American Economic Review*, 110, 3549–3601.

Drechsel, T. (2023): "Earnings-Based Borrowing Constraints and Macroeconomic Fluctuations." *American Economic Journal: Macroeconomics*, 15, 1–34.

Driscoll, J. and A. Kraay (1998): "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data." *Review of Economics and Statistics*, 80, 549–560.

Fukui, M., E. Nakamura, and J. Steinsson (2023): "The Macroeconomic Consequences of Exchange Rate Depreciations." NBER Working Paper w31279.

Gabaix, X. (2011): "The Granular Origins of Aggregate Fluctuations." *Econometrica*, 79, 733–772.

Gonçalves, S. (2011): "The moving blocks bootstrap for panel linear regression models with individual fixed effects." *Econometric Theory*, 27, 1048–1082.

Gonçalves, S., A. M. Herrera, L. Kilian, and E. Pesavento (forthcoming): "State-dependent local projections." *Journal of Econometrics*.

Gorodnichenko, Y. and M. Weber (2016): "Are Sticky Prices Costly? Evidence from the Stock Market." *American Economic Journal: Macroeconomics*, 8, 160–199.

Gurkaynak, R. S., B. Sack, and E. T. Swanson (2005): "Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements." *American Economic Review*, 95, 425–437.

Hahn, J., G. Kuersteiner, and M. Mazzocco (2020): "Estimation with Aggregate Shocks." *Review of Economic Studies*, 87, 1365–1398.

Hansen, L. P. and R. Hodrick (1980): "Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis." *Journal of Political Economy*, 88, 829–853.

Herbst, E. P. and B. K. Johansenn (2023): "Bias in Local Projections." Working paper.

Imbens, G. W. and M. Kolesár (2016): "Robust Standard Errors in Small Samples: Some Practical Advice." *Review of Economics and Statistics*, 98, 701–712.

Jeenas, P. and R. Lagos (2024): "Q-Monetary Transmission." *Journal of Political Economy*, 132.

Jordà, O. (2009): "Simultaneous Confidence Regions for Impulse Responses." *The Review of Economics and Statistics*, 91, 629–647.

Jordà, O. (2005): "Estimation and inference of impulse responses by local projections." *American Economic Review*, 95, 161–182.

Jovanovic, B. (1987): "Micro Shocks and Aggregate Risk." *Quarterly Journal of Economics*, 102, 395–409.

Känzig, D. (2021): "The Macroeconomic Effects of Oil Supply News: Evidence from OPEC Announcements." *American Economic Review*, 111, 1092–1125.

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

Lusompa, A. (2023): "Local Projections, Autocorrelation, and Efficiency." *Quantitative Economics*, 14, 1199–1220.

Majerovitz, J. and K. A. Sastry (2023): "How Much Should We Trust Regional-Exposure Designs?." Working paper.

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

Montiel Olea, J. L. and M. Plagborg-Møller (2019): "Simultaneous Confidence Bands: Theory, Implementation, and an Application to SVARs." *Journal of Applied Econometrics*, 34, 1–17.

Montiel Olea, J. L., M. Plagborg-Møller, E. Qian, and C. K. Wolf (2024): "Double Robustness of Local Projections and Some Unpleasant VARithmetic." Working Paper NBER Working Paper 32495.

Müller, U. K. (2004): "A theory of robust long-run variance estimation." Working paper.

Nakamura, E. and J. Steinsson (2018): "Identification in Macroeconomics." *Journal of Economic Perspectives*, 32, 59–86.

Ottonello, P. and T. Winberry (2020): "Financial Heterogeneity and the Investment Channel of Monetary Policy." *Econometrica*, 88, 2473–2502.

Pakel, C. (2019): "Bias reduction in nonlinear and dynamic panels in the presence of cross-section dependence." *Journal of Econometrics*, 213, 459–492.

Pesaran, M. H. (2006): "Estimation and inference in large heterogeneous panels with a multifactor structure." *Econometrica*, 74, 967–1012.

Plagborg-Møller, M. and C. K. Wolf (2021): "Local projections and VARs estimate the same impulse responses." *Econometrica*, 89, 955–980.

Rambachan, A. and N. Shephard (2021): "When do common time series estimands have nonparametric causal meaning?" Working paper.

Ramey, V. (2016): "Macroeconomic shocks and their propagation." in *Handbook of Macroeconomics*, ed. by J. B. Taylor and H. Uhlig, Elsevier, vol. 2, chap. 2.

Staiger, D. and J. H. Stock (1997): "Instrumental variables regression with weak instruments." *Econometrica*, 65, 557–586.

Stock, J. H. and M. W. Watson (2016): "Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics." in *Handbook of Macroeconomics*, ed. by J. B. Taylor and H. Uhlig, Elsevier, vol. 2, chap. 8.

——— (2018): "Identification and estimation of dynamic causal effects in macroeconomics using external instruments." *Economic Journal*, 128, 917–948.

Xu, K.-L. (2023): "Local Projection Based Inference under General Conditions." Working paper.

Xu, R. (2021): "Potential outcomes and finite-population inference for M-estimators." *Econometrics Journal*, 24, 162–176.

Supplemental Material

Micro responses to macro shocks

Martín Almuzara* Víctor Sancibrián†

This version: August 2024

[Link to the paper]

B Additional proofs

We adopt the following notation in the proofs below. We use P_N, E_N, Var_N, Cov_N to denote probability, expectation, variance and covariance given $\{\theta_i, s_i\}_{i=1}^N$ (we insert a subindex κ or κ_T when necessary).

With a slight abuse of nomenclature we sometimes call Loève's inequality to the statement $|\sum_{i=1}^m X_i|^r \leq c_r \sum_{i=1}^m |X_i|^r$ (with $c_r = 1$ if $r \leq 1$ and $c_r = m^{r-1}$ otherwise) where $X_1, \ldots, X_m$ are random variables and not just to $E\left[|\sum_{i=1}^m X_i|^r\right] \leq c_r \sum_{i=1}^m E\left[|X_i|^r\right]$ (which is implied by the former). See Davidson (1994, Theorem 9.28).

Without loss of generality we assume $\kappa \geq 0$. We also define the scaling function $g(\kappa) = \max\{1, \kappa\}$ and note that $g(\kappa)/\kappa = g(\kappa^{-1})$. In Proposition 1

$$\frac{V(h,\kappa)}{g(\kappa^2/N)} = \frac{\sum_{\ell=0}^{\infty}\left\{\iota_\ell(h)\bar{\beta}_\ell^2 E_N\left[X_t^2 X_{t+h-\ell}^2\right] + \bar{\gamma}_\ell^2 E_N\left[X_t^2 Z_{t+h-\ell}^2\right]\right\}}{g(\kappa^2/N)} + \frac{\sum_{i=1}^N \sum_{\ell=0}^{\infty} \hat{s}_i^2 \delta_{i\ell}^2 E_N\left[X_t^2 u_{i,t+h-\ell}^2\right]}{N g(N/\kappa^2)}$$

is bounded below by $\underline{\mathrm{CM}}^2 > 0$ and above by $3C^4 M_4 < \infty$ for any κ (and h). The same applies to $V(h,\kappa)/g(\kappa^2/N)$ in Proposition 2. In Proposition 3, $\mathrm{tr}\{\boldsymbol{V}(h,\kappa)\}/g(\kappa^2/N)$ is bounded below by $(a_0^2+1)\underline{\mathrm{CM}}^2 > 0$ and above by $6(p+1)(a_0^2+1)C^4 M_4 < \infty$.

*Federal Reserve Bank of New York: martin.almuzara@ny.frb.org

†CEMFI: victor.sancibrian@cemfi.edu.es

Proposition 1

Parts (A), (B) and (C) of the proof of Proposition 1 in Appendix A are established in Lemmas 1, 2 and 3 below. Lemmas 4 and 5 provide auxiliary results for Lemma 1, while 6 and 7 do the same for 2. At all times, we make Assumptions 1, 2 and 3 and we fix h and $p \geq h$ as $T, N \to \infty$ (note we do not need $T/N \to 0$ here).

Lemma 1 (Asymptotic normality of the score).

$$\frac{\sum_{t=1}^{T-h} X_t \xi_t(h, \kappa_T)}{\sqrt{(T-h)V(h, \kappa_T)}} \xrightarrow[P_{\kappa_T}]{\text{d}} N(0, 1).$$

Proof. The argument relies on the martingale representation:

$$\sum_{t=1}^{T-h} \frac{X_t \xi_t(h, \kappa_T)}{\sqrt{(T-h)V(h, \kappa_T)}} = \sum_{t=1}^{T} \chi_{T,t}(h, \kappa_T)$$

where we have defined

$$\chi_{T,t}(h, \kappa) = \frac{X_t \Xi_{X,t}(h, \kappa) + Z_t \Xi_{Z,t}(h) + (\kappa_T/N) \sum_{i=1}^{N} u_{it} \Xi_{U,it}(h)}{\sqrt{(T-h)V(h, \kappa_T)}}$$

with

$$\begin{aligned}
\Xi_{X,t}(h, \kappa) &= \sum_{\ell=1}^{h} \mathbb{1}\{t - \ell \geq 1\} \bar{\beta}_{h-\ell} X_{t-\ell} + \sum_{\ell=p+1}^{\infty} \mathbb{1}\{t \leq T - h\} \bar{\beta}_{h+\ell} X_{t-\ell} \\
&\quad + \sum_{\ell=0}^{\infty} \mathbb{1}\{t \leq T - h\} \left[\bar{\gamma}_{h+\ell} Z_{t-\ell} + \frac{\kappa}{N} \sum_{i=1}^{N} \hat{s}_i \delta_{i,h+\ell} u_{i,t-\ell} \right], \\
\Xi_{Z,t}(h) &= \sum_{\ell=1}^{h} \mathbb{1}\{t - \ell \geq 1\} \bar{\gamma}_{h-\ell} X_{t-\ell}, \\
\Xi_{U,it}(h) &= \sum_{\ell=1}^{h} \mathbb{1}\{t - \ell \geq 1\} \hat{s}_i \delta_{i,h-\ell} X_{t-\ell}.
\end{aligned}$$

Under Assumption 2, it can be readily verified that $\{\chi_{T,t}(h, \kappa_T)\}_{t=1}^{T}$ is a martingale difference array adapted to the natural filtration $\{\mathcal{F}_{T,t}\}_{t=1}^{T}$,

$$\mathcal{F}_{T,t} = \sigma\left(\{X_\tau, Z_\tau, \{u_{i\tau}\}_{i=1}^{N}\}_{\tau \leq t}, \{\theta_i, s_i\}_{i=1}^{N}\right),$$

that is, $\chi_{T,t}(h,\kappa_T)$ is $\mathcal{F}_{T,t}$-measurable and $E_{\kappa_T}\left[\chi_{T,t}(h,\kappa_T)\middle|\mathcal{F}_{T,t-1}\right]=0$.

By construction, $\sum_{t=1}^{T} E_{\kappa_T}\left[\chi_{T,t}(h,\kappa_T)^2\right]=1$ and we can show (Lemmas 4 and 5)

$$\sum_{t=1}^{T}\chi_{T,t}(h,\kappa_T)^2 \xrightarrow[P_{\kappa_T}]{\text{p}} 1 \text{ and } \lim_{T\to\infty}\sum_{t=1}^{T} E_{\kappa_T}\left[\chi_{T,t}(h,\kappa_T)^4\right]=0.$$

By Davidson (1994, Theorems 23.11, 23.16 and 24.3), the Lemma follows. □

Lemma 2 (Consistency of the standard error).

$$\frac{\hat{V}(h)}{V(h,\kappa_T)} \xrightarrow[P_{\kappa_T}]{\text{p}} 1.$$

Proof. Since $V(h,\kappa_T)>0$ holds P_{κ_T}-a.s., it suffices to show that

$$\frac{\hat{V}(h)-V(h,\kappa_T)}{g(\kappa_T^2/N)} \xrightarrow[P_{\kappa_T}]{\text{p}} 0.$$

Write

$$\frac{\hat{V}(h)-V(h,\kappa_T)}{g(\kappa_T^2/N)}=D_{T,1}(h,\kappa_T)+D_{T,2}(h,\kappa_T),$$

where we have defined

$$D_{T,1}(h,\kappa_T)=\sum_{t=1}^{T-h}\frac{\left(X_t^2\xi_t(h,\kappa_T)^2-E_{\kappa_T}\left[X_t^2\xi_t(h,\kappa_T)^2\middle|\{\theta_i,s_i\}_{i=1}^N\right]\right)}{(T-h)g(\kappa_T^2/N)},$$

$$D_{T,2}(h,\kappa_T)=\sum_{t=1}^{T-h}\left[\frac{\left(N^{-1}\sum_{i=1}^N\hat{x}_{it}(h)\hat{\xi}_{it}(h)\right)^2-X_t^2\xi_t(h,\kappa_T)^2}{(T-h)g(\kappa_T^2/N)}\right].$$

Next, using $(x^2-y^2)=(x-y)(x+y)$ and the Cauchy-Schwarz inequality,

$$\left|D_{T,2}(h,\kappa_T)\right|\leq\sqrt{D_{T,2}^{-}(h,\kappa_T)}\sqrt{D_{T,2}^{+}(h,\kappa_T)},$$

with

$$D_{T,2}^{-}(h,\kappa_T)=\sum_{t=1}^{T-h}\frac{\left[\left(N^{-1}\sum_{i=1}^N\hat{x}_{it}(h)\hat{\xi}_{it}(h)\right)-X_t\xi_t(h,\kappa_T)\right]^2}{(T-h)g(\kappa_T^2/N)},$$

$$D^{+}_{T,2}(h,\kappa_T) = \sum_{t=1}^{T-h} \frac{\left[\left(N^{-1}\sum_{i=1}^{N}\hat{x}_{it}(h)\hat{\xi}_{it}(h)\right) + X_t\xi_t(h,\kappa_T)\right]^2}{(T-h)g(\kappa_T^2/N)}.$$

Adding and subtracting $X_t\xi_t(h,\kappa_T)$ within the squares in $D^{+}_{T,2}(h,\kappa_T)$ and applying Loève's inequality,

$$D^{+}_{T,2}(h,\kappa_T) \le 2D^{-}_{T,2}(h,\kappa_T) + 8|D_{T,1}(h,\kappa_T)| + \frac{8V(h,\kappa_T)}{g(\kappa_T^2/N)}.$$

We can show (Lemmas 6 and 7) that $D_{T,1}(h,\kappa_T) = o_{P_{\kappa_T}}(1)$ and $D^{-}_{T,2}(h,\kappa_T) = o_{P_{\kappa_T}}(1)$. Given that $V(h,\kappa_T)/g(\kappa_T^2/N)$ is bounded P_{κ_T}-a.s., $D^{+}_{T,2}(h,\kappa_T) = O_{P_{\kappa_T}}(1)$ which implies $D_{T,2}(h,\kappa_T) = o_{P_{\kappa_T}}(1)$ and the Lemma follows. □

Lemma 3 (Negligibility of the reminder).

$$R_T(h,\kappa_T) \xrightarrow[P_{\kappa_T}]{\text{p}} 0.$$

Proof. Let $\bar{x}_t(h) = (X_{t-1} - \bar{X}_1(h), \dots, X_{t-p} - \bar{X}_p(h))'$ where $\bar{X}_\ell(h) = (T-h)^{-1}\sum_{t=1}^{T-h} X_{t-\ell}$. Since either $\hat{s}_i$ was demeaned or time effects were not included as controls,

$$\hat{\pi}(h)'W_{it} = \hat{\pi}_{0,i}(h) + \sum_{\ell=1}^{p}\hat{\pi}_{X,\ell}(h)\hat{s}_i X_{t-\ell} = \hat{s}_i\left(\bar{X}_0(h) + \hat{\pi}_X(h)'\bar{x}_t(h)\right),$$

where $\{\hat{\pi}_{0,i}(h)\}$, $\pi_X(h) = (\hat{\pi}_{X,1}(h),\dots,\hat{\pi}_{X,p}(h))'$ are the coefficients from the regression of s_iX_t on unit fixed effects and p lags of $\hat{s}_iX_t$. Furthermore, it is readily seen that $\hat{\pi}_X(h)$ are also the coefficients in a regression of X_t on $\bar{x}_t(h)$,

$$\hat{\pi}_X(h) = \left[\sum_{t=1}^{T-h}\bar{x}_t(h)\bar{x}_t(h)'\right]^{-1}\sum_{t=1}^{T-h}\bar{x}_t(h)X_t.$$

Note that $E\left[X_{t-\ell}\right] = E\left[X_{t-\ell}X_t\right] = 0$ and that $\mathrm{Var}\left(\sum_{t=1}^{T-h}X_{t-\ell}\right), \mathrm{Var}\left(\sum_{t=1}^{T-h}X_{t-\ell}X_t\right)$ are bounded by a constant (M_2 and M_4, respectively) times $(T-h)$ under Assumptions 1, 2 and 3. Also note that $(T-h)^{-1}\sum_{t=1}^{T-h}\bar{x}_t(h)\bar{x}_t(h)' = E\left[X_t^2\right]\times I_p + o_{P_{\kappa_T}}(1)$. All of this is independent of κ_T. It follows that

$$\bar{X}_0(h) = O_{P_{\kappa_T}}\left((T-h)^{-1/2}\right), \quad \hat{\pi}_X(h) = O_{P_{\kappa_T}}\left((T-h)^{-1/2}\right).$$

Write

$$R_T(h, \kappa_T) = -\frac{\bar{X}_0(h) \sum_{t=1}^{T-h} \xi_t(h, \kappa_T)}{\sqrt{(T-h)V(h, \kappa_T)}} - \frac{\hat{\pi}_X(h)' \sum_{t=1}^{T-h} \bar{x}_t(h)\xi_t(h, \kappa_T)}{\sqrt{(T-h)V(h, \kappa_T)}}.$$

To obtain $R_T(h, \kappa_T) = o_{P_{\kappa_T}}(1)$, we show $\{(T-h)V(h, \kappa_T)\}^{-1/2} \sum_{t=1}^{T} \xi_t(h, \kappa_T) = O_{P_{\kappa_T}}(1)$ and $\{(T-h)V(h, \kappa_T)\}^{-1/2} \sum_{t=1}^{T} \bar{x}_t(h)\xi_t(h, \kappa_T) = O_{P_{\kappa_T}}(1)$. We do so by direct calculation.

First,

$$\begin{aligned}
E_{N,\kappa_T}\left[\left(\sum_{t=1}^{T-h} \xi_t(h, \kappa_T)\right)^2\right] &= E_N\left[\left(\sum_{t=1}^{T-h}\sum_{\ell=0}^{\infty} \iota_\ell(h)\bar{\beta}_\ell X_{t+h-\ell}\right)^2\right] + E_N\left[\left(\sum_{t=1}^{T-h}\sum_{\ell=0}^{\infty} \bar{\gamma}_\ell Z_{t+h-\ell}\right)^2\right] \\
&\quad + \frac{\kappa_T^2}{N^2} E_N\left[\left(\sum_{t=1}^{T-h}\sum_{i=1}^{N}\sum_{\ell=0}^{\infty} \hat{s}_i \delta_{i\ell} u_{i,t+h-\ell}\right)^2\right] \\
&\leq 2(T-h)\Bigg[\left(\sum_{\ell=0}^{\infty} \iota_\ell(h)|\bar{\beta}_\ell|\right)^2 E_N\left[X_t^2\right] + \left(\sum_{\ell=0}^{\infty} |\bar{\gamma}_\ell|\right)^2 E_N\left[Z_t^2\right] \\
&\quad + \frac{\kappa_T^2}{N^2}\sum_{i=1}^{N}\left(\sum_{\ell=0}^{\infty} |\hat{s}_i\delta_{i\ell}|\right)^2 E_N\left[u_{it}^2\right]\Bigg] \\
&\leq (T-h) \times 2(2 + \kappa_T^2/N)C^4 M_2,
\end{aligned}$$

where the last line uses Assumption 3(i)–(iv).[1] By iterated expectations and Chebyshev's inequality, for any $\varepsilon > 0$,

$$\begin{aligned}
P_{\kappa_T}\left(\left|\frac{\sum_{t=1}^{T} \xi_t(h, \kappa_T)}{\sqrt{(T-h)V(h, \kappa_T)}}\right| > \varepsilon\right) &= E_{\kappa_T}\left[P_{N,\kappa_T}\left(\left|\frac{\sum_{t=1}^{T} \xi_t(h, \kappa_T)}{\sqrt{(T-h)V(h, \kappa_T)}}\right| > \varepsilon\right)\right] \\
&\leq \frac{1}{\varepsilon^2} E_{\kappa_T}\left[\frac{2(2 + \kappa_T^2/N)C^4 M_2}{V(h, \kappa_T)}\right] \leq \frac{1}{\varepsilon^2}\frac{6C^4 M_2}{\underline{\text{CM}}^2} < \infty,
\end{aligned}$$

[1] We also used the fact that for any linear process $\omega_t = \sum_{\ell=0}^{\infty} \varphi_\ell \varepsilon_{t-\ell}$ where $\{\varphi_\ell\}$ are absolutely summable and $\{\varepsilon_t\}$ is white noise with $E[\varepsilon_t] = 0$ and $E\left[\varepsilon_t^2\right] = 1$,

$$E\left[\left(\sum_{t=1}^{T} \omega_t\right)^2\right] = \sum_{m=-(T-1)}^{T-1} (T - |m|) \sum_{\ell=0}^{\infty} \varphi_\ell \varphi_{\ell+|m|} \leq T \sum_{\ell=0}^{\infty} |\varphi_\ell| \sum_{m=-\infty}^{\infty} |\varphi_{\ell+|m|}| \leq 2T\left(\sum_{\ell=0}^{\infty} |\varphi_\ell|\right)^2.$$

where the bound on $(2+\kappa_T^2/N)/V(h,\kappa_T) = ((2+\kappa_T^2/N)/g(\kappa_T^2/N))\times(g(\kappa_T^2/N)/V(h,\kappa_T))$ uses $(2+\kappa)/g(\kappa) \le 3$ and $V(h,\kappa_T)/g(\kappa_T^2/N) \ge \underline{\mathrm{C}}\underline{\mathrm{M}}^2$.

Similarly for any $k = 1,\dots,p$,

$$E_{N,\kappa_T}\left[\left(\sum_{t=1}^{T-h} X_{t-k}\xi_t(h,\kappa_T)\right)^2\right] \le (T-h)\left[\sum_{\ell=0}^{\infty}\iota_\ell(h)\bar{\beta}_\ell^2 E_N\left[X_{t-k}^2X_{t+h-\ell}^2\right] + \sum_{\ell=0}^{\infty}\bar{\gamma}_\ell^2 E_N\left[X_{t-k}^2Z_{t+h-\ell}^2\right]\right.$$
$$\left. + \frac{\kappa_T^2}{N^2}\sum_{i=1}^{N}\sum_{\ell=0}^{\infty}\hat{s}_i^2\delta_{i\ell}^2 E_N\left[X_{t-k}^2u_{i,t+h-\ell}^2\right] + 2\sum_{\ell=1}^{h+k}\iota_{h+k-\ell}(h)\iota_{h+k+\ell}(h)|\bar{\beta}_{h+k-\ell}\bar{\beta}_{h+k+\ell}|E_N\left[X_{t-k}^2X_{t-k-\ell}^2\right]\right]$$
$$\le (T-h)\times(4+\kappa_T^2/N)C^4M_4,$$

where we used the autocovariances of $X_{t-k}\xi_t(h,\kappa_T)$ and Assumption 3(i)–(iv) again. By iterated expectations and Chebyshev, for any $\varepsilon > 0$,

$$P_{\kappa_T}\left(\left|\frac{\sum_{t=1}^T X_{t-r}\xi_t(h,\kappa_T)}{\sqrt{(T-h)V(h,\kappa_T)}}\right| > \varepsilon\right) \le \frac{1}{\varepsilon^2}E_{\kappa_T}\left[\frac{(4+\kappa_T^2/N)C^4M_4}{V(h,\kappa_T)}\right] \le \frac{1}{\varepsilon^2}\frac{5C^4M_4}{\underline{\mathrm{C}}\underline{\mathrm{M}}^2} < \infty.$$

Thus, $R_T(h,\kappa_T) = o_{P_{\kappa_T}}(1)$ and the Lemma follows. □

Lemma 4. *Under the conditions of Lemma 1,*

$$\sum_{t=1}^{T}\chi_{T,t}(h,\kappa_T)^2 \xrightarrow[P_{\kappa_T}]{\mathrm{p}} 1.$$

Proof. We show $\mathrm{Var}_{N,\kappa_T}\left(\sum_{t=1}^T \chi_{T,t}(h,\kappa_T)^2\right) \le \bar{V}/(T-h)$ for a constant $\bar{V}$ independent of κ_T. Since $E_{N,\kappa_T}\left[\sum_{t=1}^T \chi_{T,t}(h,\kappa_T)^2\right] = 1$, by iterated expectations and Chebyshev's inequality, for any $\varepsilon > 0$,

$$P_{\kappa_T}\left(\left|\sum_{t=1}^{T}\chi_{T,t}(h,\kappa_T)^2 - 1\right| > \varepsilon\right) = E_{\kappa_T}\left[P_{N,\kappa_T}\left(\left|\sum_{t=1}^{T}\chi_{T,t}(h,\kappa_T)^2 - 1\right| > \varepsilon\right)\right]$$
$$\le \frac{\bar{V}}{\varepsilon^2(T-h)} \to 0.$$

As argued at the beginning of the section, $V(h,\kappa)/g(\kappa^2/N)$ is bounded away from zero and infinity uniformly over κ. Thus, it suffices to show

$$\mathrm{Var}_{N,\kappa_T}\left(\sum_{t=1}^{T}\frac{V(h,\kappa_T)\chi_{T,t}(h,\kappa_T)^2}{g(\kappa_T^2/N)}\right) \le \frac{\bar{V}}{T-h},$$

P_{κ_T}-a.s., for some constant $\bar{V}$ independent of κ_T. We do this by a direct calculation. Define $\bar{\chi}_{T,t}(h,\kappa_T) = \chi_{T,t}(h,\kappa_T)\left\{(T-h)V(h,\kappa_T)/g(\kappa_T^2/N)\right\}^{1/2}$ so that

$$
\begin{aligned}
g\left(\frac{\kappa_T}{\sqrt{N}}\right)\bar{\chi}_{T,t}(h,\kappa_T) &= X_t \Xi_{X,t}(h,\kappa) + Z_t \Xi_{Z,t}(h) + \frac{\kappa_T}{N}\sum_{i=1}^{N} u_{it}\Xi_{U,it}(h) \\
&= \underbrace{\sum_{\ell=1}^{\infty} b_{t,\ell} X_t X_{t-\ell}}_{\equiv g(\kappa_T/\sqrt{N})\zeta_{1,t}} + \underbrace{\sum_{\ell=0}^{\infty} c_{t,\ell} X_t Z_{t-\ell}}_{\equiv g(\kappa_T/\sqrt{N})\zeta_{2,t}} + \underbrace{\sum_{\ell=1}^{h} \tilde{c}_{t,\ell} Z_t X_{t-\ell}}_{\equiv g(\kappa_T/\sqrt{N})\zeta_{3,t}} \\
&\quad + \underbrace{\frac{\kappa_T}{N}\sum_{i=1}^{N}\sum_{\ell=0}^{\infty} d_{it,\ell} X_t u_{i,t-\ell}}_{\equiv g(\kappa_T/\sqrt{N})\zeta_{4,t}} + \underbrace{\frac{\kappa_T}{N}\sum_{i=1}^{N}\sum_{\ell=1}^{h} \tilde{d}_{it,\ell} u_{it} X_{t-\ell}}_{\equiv g(\kappa_T/\sqrt{N})\zeta_{5,t}} \qquad \text{(B.1)}
\end{aligned}
$$

for some $\{b_{t,\ell}, c_{t,\ell}, \tilde{c}_{t,\ell}, \{d_{it,\ell}, \tilde{d}_{it,\ell}\}_{i=1}^N\}$ that depend on $\{\theta_i, s_i\}_{i=1}^N$ (and h). Note that the coefficients depend on t only via the indicator functions $\mathbb{1}\{t-\ell \le 1\}$ and $\mathbb{1}\{t \le T-h\}$. It will be convenient to define $\{b_\ell, c_\ell, \tilde{c}_\ell, \{d_{i,\ell}, \tilde{d}_{i,\ell}\}_{i=1}^N\}$ as the coefficients we would get by setting the indicators to one. This implies $|b_{t,\ell}| \le |b_\ell|$, $|c_{t,\ell}| \le |c_\ell|$, and so on. By Assumption 3(iv), $|b_\ell|, |c_\ell|, |\tilde{c}_\ell|, |d_{i\ell}|, |\tilde{d}_{i\ell}| \le \bar{C}_\ell$ almost surely for finite constants $\bar{C}_\ell$ such that $\bar{C} = \sum_{\ell=1}^{\infty} \bar{C}_\ell < \infty$ (in fact, we can take $\bar{C} \le C^2$ independent of h).

Consider the variance

$$
\mathrm{Var}_{N,\kappa_T}\left(\sum_{t=1}^{T} \frac{V(h,\kappa_T)\chi_{T,t}(h,\kappa_T)^2}{g(\kappa_T^2/N)}\right) = \frac{\sum_{t=1}^{T}\sum_{\tau=1}^{T}\Gamma_T(t,\tau)}{(T-h)^2}
$$

where (omitting the dependence on h, κ_T and $\{\theta_i, s_i\}_{i=1}^N$)

$$
\Gamma_T(t,\tau) = \mathrm{Cov}_{N,\kappa_T}\Big(\tilde{\chi}_{T,t}(h,\kappa_T)^2, \tilde{\chi}_{T,\tau}(h,\kappa_T)^2\Big).
$$

Expanding the square of $\tilde{\chi}_{T,t}(h,\kappa_T)$ and using the linearity of the covariance we can express $\Gamma_T(t,\tau)$ as the sum of covariances $\Gamma_{T,k_1k_2k_3k_4}(t,\tau) = \mathrm{Cov}_{N,\kappa_T}\Big(\zeta_{k_1,t}\zeta_{k_2,t}, \zeta_{k_3,\tau}\zeta_{k_4,\tau}\Big)$ where k_1, k_2, k_3, k_4 range over the five terms in (B.1). Moreover, if $k_1 = k_2$, $\Gamma_{T,k_1k_2k_3k_4}(t,\tau)$ can only be non-zero if $k_3 = k_4$, while if $k_1 \neq k_2$, only if either $k_1 = k_3$

and $k_2 = k_4$ or $k_1 = k_4$ and $k_2 = k_3$. Then, by the triangle inequality,

$$\begin{aligned}\left|\Gamma_T(t,\tau)\right| &= \left|\sum_{k_1=1}^{5}\sum_{k_2=1}^{5}\sum_{k_3=1}^{5}\sum_{k_4=1}^{5}\Gamma_{T,k_1k_2k_3k_4}(t,\tau)\right| \\ &\leq \sum_{k_1=1}^{5}\sum_{k_3=1}^{5}\left|\Gamma_{T,k_1k_1k_3k_3}(t,\tau)\right| + 2\sum_{k_1=1}^{5}\sum_{k_2=1}^{5}\left|\Gamma_{T,k_1k_2k_1k_2}(t,\tau)\right|. \qquad \text{(B.2)}\end{aligned}$$

We begin with $\sum_{t=1}^{T}\sum_{\tau=1}^{T}\Gamma_{T,k_1k_1k_3k_3}(t,\tau)$. Consider $k_1 = k_3 = 1$:

$$\begin{aligned}g\left(\frac{\kappa_T^4}{N^2}\right)\left|\Gamma_{T,1111}(t,\tau)\right| &= \left|\mathrm{Cov}_N\left(\left(\sum_{\ell=1}^{\infty}b_{t,\ell}X_tX_{t-\ell}\right)^2, \left(\sum_{\ell=1}^{\infty}b_{\tau,\ell}X_\tau X_{\tau-\ell}\right)^2\right)\right| \\ &= \left|\sum_{\ell_1=1}^{\infty}\sum_{\ell_2=1}^{\infty}\sum_{\ell_3=1}^{\infty}\sum_{\ell_4=1}^{\infty}b_{t,\ell_1}b_{t,\ell_2}b_{\tau,\ell_3}b_{\tau,\ell_4}\mathrm{Cov}_N\left(X_t^2X_{t-\ell_1}X_{t-\ell_2}, X_\tau^2X_{\tau-\ell_3}X_{\tau-\ell_4}\right)\right| \\ &\leq \sum_{\ell_1=1}^{\infty}\sum_{\ell_3=1}^{\infty}b_{\ell_1}^2b_{\ell_3}^2\left|\mathrm{Cov}_N\left(X_t^2X_{t-\ell_1}^2, X_\tau^2X_{\tau-\ell_3}^2\right)\right| \\ &\quad + 2\sum_{\ell_1=1}^{\infty}\sum_{\ell_2\neq\ell_1}|b_{\ell_1}b_{\ell_2}b_{\ell_1+\tau-t}b_{\ell_2+\tau-t}|\left|\mathrm{Cov}_N\left(X_t^2X_{t-\ell_1}X_{t-\ell_2}, X_\tau^2X_{t-\ell_1}X_{t-\ell_2}\right)\right|.\end{aligned}$$

The inequality uses the fact that by Assumption 2, $\mathrm{Cov}_N\left(X_t^2X_{t-\ell_1}X_{t-\ell_2}, X_\tau^2X_{\tau-\ell_3}X_{\tau-\ell_4}\right)$ can only be non-zero if $\ell_1 = \ell_2$ and $\ell_3 = \ell_4$ or, with $\ell_1 \neq \ell_2$, if either $\ell_3 = \ell_1 + \tau - t$ and $\ell_4 = \ell_2 + \tau - t$ or $\ell_3 = \ell_2 + \tau - t$ and $\ell_4 = \ell_1 + \tau - t$.[2] We also use $|b_{t,\ell}| \leq |b_\ell|$.

For the first double sum, now summing over t and τ,

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{\tau=1}^{T}\sum_{\ell_1=1}^{\infty}\sum_{\ell_3=1}^{\infty}b_{\ell_1}^2b_{\ell_3}^2\left|\mathrm{Cov}_N\left(X_t^2X_{t-\ell_1}^2, X_\tau^2X_{\tau-\ell_3}^2\right)\right| \\ &\qquad \leq 2T\sum_{m=0}^{T-1}\sum_{\ell_1=1}^{\infty}\sum_{\ell_3=1}^{\infty}\bar{C}_{\ell_1}^2\bar{C}^2\left|\mathrm{Cov}_N\left(X_t^2X_{t-\ell_1}^2, X_{t-m}^2X_{t-m-\ell_3}^2\right)\right| \\ &\qquad \leq 2T\bar{C}^2\sum_{\ell_1=1}^{\infty}\bar{C}_{\ell_1}^2\left(\sum_{j_1=-\infty}^{\infty}\sum_{j_2=-\infty}^{\infty}\left|\mathrm{Cov}_N\left(X_t^2X_{t-\ell_1}^2, X_{t-m}^2X_{t-m-\ell_3}^2\right)\right|\right)\end{aligned}$$

[2] This is similar to the proof of Montiel Olea and Plagborg-Møller (2021, Lemma A.6)

$$\leq 2T\bar{C}^2\bar{K}\sum_{\ell_1=1}^{\infty}\bar{C}_{\ell_1}^2 \leq 2T\bar{C}^4\bar{K}$$

for some constant $\bar{K}$ that can be shown to exist as by Assumption 3(iii) the fourth-order cumulants of X_t^2 conditional on $\{\theta_i, s_i\}_{i=1}^N$ are absolutely summable.

Turning to the second double sum, by Assumption 2, since $\ell_1 \neq \ell_2$,

$$\left|\mathrm{Cov}_N\Big(X_t^2 X_{t-\ell_1}X_{t-\ell_2}, X_\tau^2 X_{\tau-\ell_1}X_{\tau-\ell_2}\Big)\right| = \left|E_N\left[X_t^2X_\tau^2X_{t-\ell_1}^2X_{t-\ell_2}^2\right]\right| \leq E_N\left[X_t^8\right] \leq M_8,$$

where M_8 is the moment bound from Assumption 3(i). Then,

$$\begin{aligned}
2\sum_{t=1}^{T}\sum_{\tau=1}^{T}\sum_{\ell_1=1}^{\infty}\sum_{\ell_2\neq\ell_1}&|b_{\ell_1}b_{\ell_2}b_{\ell_1+\tau-t}b_{\ell_2+\tau-t}|\left|\mathrm{Cov}_N\Big(X_t^2X_{t-\ell_1}X_{t-\ell_2}, X_\tau^2X_{t-\ell_1}X_{t-\ell_2}\Big)\right| \\
&\leq 4TM_8\sum_{m=0}^{T-1}\sum_{\ell_1=1}^{\infty}\sum_{\ell_2\neq\ell_1}|b_{\ell_1}b_{\ell_2}b_{\ell_1+m}b_{\ell_2+m}| \\
&\leq 4TM_8\sum_{\ell_1=1}^{\infty}\sum_{\ell_2=1}^{\infty}|b_{\ell_1}||b_{\ell_2}|\left(\sum_{m=0}^{\infty}|b_{\ell_1+m}||b_{\ell_2+m}|\right) \\
&\leq 4TM_8\sum_{\ell_1=1}^{\infty}\sum_{\ell_2=1}^{\infty}|b_{\ell_1}||b_{\ell_2}|\left(\sum_{m_1=1}^{\infty}|b_{m_1}|^2\sum_{m_2=1}^{\infty}|b_{m_2}|^2\right)^{1/2} \\
&\leq 4T\bar{C}^4M_8,
\end{aligned}$$

where the second inequality increases the range of summation over ℓ_2 and m, the third uses Cauchy-Schwarz and the fourth follows from Assumption 3(iv).

Putting these calculations together and using $g(\kappa) \geq 1$,

$$\frac{\sum_{t=1}^T\sum_{\tau=1}^T\left|\Gamma_{T,1111}(t,\tau)\right|}{(T-h)^2} \leq \frac{T\times 2\bar{C}^4(\bar{K}+2M_8)}{g(\kappa_T^4/N^2)(T-h)^2} \leq \frac{2\bar{C}^4(\bar{K}+2M_8)}{(1-h/T)(T-h)}.$$

In fact, the same bound works for $\sum_{t=1}^T\sum_{\tau=1}^T\left|\Gamma_{T,k_1k_1k_3k_3}(t,\tau)\right|$ for any $k_1, k_3 \in \{1,2,3\}$.

Next consider $k_1 = k_3 = 4$:

$$g\left(\frac{\kappa_T^4}{N^2}\right)\frac{\left|\Gamma_{T,4444}(t,\tau)\right|}{(\kappa_T^4/N^4)} = \left|\mathrm{Cov}_N\left(\left(\sum_{i=1}^N\sum_{\ell=1}^{\infty}d_{it,\ell}X_tu_{i,t-\ell}\right)^2, \left(\sum_{i=1}^N\sum_{\ell=1}^{\infty}d_{i\tau,\ell}X_\tau u_{i,\tau-\ell}\right)^2\right)\right|$$

$$
\begin{aligned}
&= \Bigg| \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \sum_{i_3=1}^{N} \sum_{i_4=1}^{N} \sum_{\ell_1=1}^{\infty} \sum_{\ell_2=1}^{\infty} \sum_{\ell_3=1}^{\infty} \sum_{\ell_4=1}^{\infty} d_{i_1 t,\ell_1} d_{i_2 t,\ell_2} d_{i_3 \tau,\ell_3} d_{i_4 \tau,\ell_4} \\
&\qquad \times \mathrm{Cov}_N\Big(X_t^2 u_{i_1,t-\ell_1} u_{i_2,t-\ell_2}, X_\tau^2 u_{i_3,\tau-\ell_3} u_{i_4,\tau-\ell_4} \Big) \Bigg| \\
&\le \sum_{i_1=1}^{N} \sum_{i_3=1}^{N} \Bigg| \sum_{\ell_1=1}^{\infty} \sum_{\ell_2=1}^{\infty} \sum_{\ell_3=1}^{\infty} \sum_{\ell_4=1}^{\infty} d_{i_1 t,\ell_1} d_{i_1 t,\ell_2} d_{i_3 \tau,\ell_3} d_{i_3 \tau,\ell_4} \\
&\qquad \times \mathrm{Cov}_N\Big(X_t^2 u_{i_1,t-\ell_1} u_{i_1,t-\ell_2}, X_\tau^2 u_{i_3,\tau-\ell_3} u_{i_3,\tau-\ell_4} \Big) \Bigg| \\
&\quad + \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \Bigg| \sum_{\ell_1=1}^{\infty} \sum_{\ell_2=1}^{\infty} \sum_{\ell_3=1}^{\infty} \sum_{\ell_4=1}^{\infty} d_{i_1 t,\ell_1} d_{i_2 t,\ell_2} d_{i_1 \tau,\ell_3} d_{i_2 \tau,\ell_4} \\
&\qquad \times \mathrm{Cov}_N\Big(X_t^2 u_{i_1,t-\ell_1} u_{i_2,t-\ell_2}, X_\tau^2 u_{i_1,\tau-\ell_3} u_{i_2,\tau-\ell_4} \Big) \Bigg| \\
&\quad + \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \Bigg| \sum_{\ell_1=1}^{\infty} \sum_{\ell_2=1}^{\infty} \sum_{\ell_3=1}^{\infty} \sum_{\ell_4=1}^{\infty} d_{i_1 t,\ell_1} d_{i_2 t,\ell_2} d_{i_2 \tau,\ell_3} d_{i_1 \tau,\ell_4} \\
&\qquad \times \mathrm{Cov}_N\Big(X_t^2 u_{i_1,t-\ell_1} u_{i_2,t-\ell_2}, X_\tau^2 u_{i_2,\tau-\ell_3} u_{i_1,\tau-\ell_4} \Big) \Bigg| .
\end{aligned}
$$

The inequality uses the fact that $\mathrm{Cov}_N\Big(X_t^2 u_{i_1,t-\ell_1} u_{i_1,t-\ell_2}, X_\tau^2 u_{i_3,\tau-\ell_3} u_{i_3,\tau-\ell_4} \Big)$ can only be non-zero if $i_1 = i_2$ and $i_3 = i_4$, or $i_1 = i_3$ and $i_2 = i_4$, or $i_1 = i_4$ and $i_2 = i_3$.

Summing over t and τ and applying to each of the three summands on the right hand side the same steps as the case $k_1 = k_3 = 1$,

$$
\frac{\sum_{t=1}^{T} \sum_{\tau=1}^{T} \left| \Gamma_{T,4444}(t,\tau) \right|}{(T-h)^2} \le \frac{3N^2 \times \kappa_T^4 / N^4 \times 2\bar{C}^4 (\bar{K} + 2M_8)}{g(\kappa_T^4/N^2)(1-h/T)(T-h)} \le \frac{6\bar{C}^4(\bar{K} + 2M_8)}{(1-h/T)(T-h)}.
$$

Repeating the calculation for the remaining cases (and noting that this bound is three times larger than the one we computed for $k_1 = k_3 = 1$) we conclude that $6\bar{C}^4(\bar{K}+2M_8)/(1-h/T)(T-h)$ works for any $k_1, k_3 \in \{1,2,3,4,5\}$. By similar reasoning, the bound also works for $\sum_{t=1}^{T} \sum_{\tau=1}^{T} \Gamma_{T,k_1 k_2 k_1 k_2}(t,\tau)$ whenever $k_1 \neq k_2$. We then get

$$
\frac{\sum_{t=1}^{T} \sum_{\tau=1}^{T} \Gamma_T(t,\tau)}{(T-h)^2} \le \frac{\bar{V}}{(T-h)},
$$

where $\bar{V} = 75 \times 6\bar{C}^4(\bar{K} + 2M_8)/(1 - h/T)$ does not depend on κ_T (75 is the number of covariances in (B.2)). This establishes $\sum_{t=1}^{T} \chi_{T,t}(h, \kappa_T)^2 = 1 + o_{P_{\kappa_T}}(1)$. □

Lemma 5. *Under the conditions of Lemma 1,*

$$\lim_{T\to\infty} \sum_{t=1}^{T} E_{\kappa_T}\left[\chi_{T,t}^4\right] = 0.$$

Proof. Using the notation of Lemma 4 and Loève's inequality,

$$E_N\left[\bar{\chi}_{T,t}(h,\kappa_T)^4\right] \le 5^3 \sum_{k=1}^{5} E_N\left[\zeta_{k,t}^4\right]. \tag{B.3}$$

Each of the five terms in (B.3) is under Assumption 3(i)–(iv) bounded by a constant that does not depend on κ_T. For $k = 1$,

$$\begin{aligned}
g\left(\frac{\kappa_T^4}{N^2}\right) E_N\left[\zeta_{1,t}^4\right] &= E_N\left[\left(\sum_{\ell=1}^{\infty} b_{t,\ell} X_t X_{t-\ell}\right)^4\right] \\
&\le \sum_{\ell_1=1}^{\infty}\sum_{\ell_2=1}^{\infty}\sum_{\ell_3=1}^{\infty}\sum_{\ell_4=1}^{\infty} |b_{t,\ell_1} b_{t,\ell_2} b_{t,\ell_3} b_{t,\ell_4}| \left|E_N\left[X_t^4 X_{t-\ell_1} X_{t-\ell_2} X_{t-\ell_3} X_{t-\ell_4}\right]\right| \\
&\le M_8 \sum_{\ell_1=1}^{\infty}\sum_{\ell_2=1}^{\infty}\sum_{\ell_3=1}^{\infty}\sum_{\ell_4=1}^{\infty} |b_{\ell_1} b_{\ell_2} b_{\ell_3} b_{\ell_4}| \le M_8 \left(\sum_{\ell=1}^{\infty} |b_\ell|\right)^4 \le M_8 \bar{C}^4,
\end{aligned}$$

where $\bar{C}$ is the constant we defined in the first part. The same bound works for $k = 2$ and $k = 3$ in (B.3). For $k = 4$,

$$\begin{aligned}
g\left(\frac{\kappa_T^4}{N^2}\right) \frac{E_N\left[\zeta_{4,t}^4\right]}{(\kappa_T^4/N^4)} &= E_N\left[\left(\sum_{i=1}^{N}\sum_{\ell=1}^{\infty} d_{it,\ell} X_t u_{i,t-\ell}\right)^4\right] \\
&\le \sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\sum_{i_3=1}^{N}\sum_{i_4=1}^{N}\sum_{\ell_1=1}^{\infty}\sum_{\ell_2=1}^{\infty}\sum_{\ell_3=1}^{\infty}\sum_{\ell_4=1}^{\infty} |d_{i_1t,\ell_1} d_{i_2t,\ell_2} d_{i_3t,\ell_3} d_{i_4t,\ell_4}| \\
&\quad \times \left|E_N\left[X_t^4 u_{i_1,t-\ell_1} u_{i_2,t-\ell_2} u_{i_3,t-\ell_3} u_{i_4,t-\ell_4}\right]\right| \\
&\le 3 \sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\sum_{\ell_1=1}^{\infty}\sum_{\ell_2=1}^{\infty} |d_{i_1t,\ell_1}^2 d_{i_2t,\ell_2}^2| \left|E_N\left[X_t^4 u_{i_1,t-\ell_1}^2 u_{i_2,t-\ell_2}^2\right]\right|
\end{aligned}$$

$$\leq 3N^2 M_8 \sum_{\ell_1=1}^{\infty}\sum_{\ell_2=1}^{\infty} d_{\ell_1}^2 d_{\ell_2}^2 \leq 3N^2 M_8 \bar{C}^4,$$

where the second inequality uses that for $E_N\left[X_t^4 u_{i_1,t-\ell_1} u_{i_2,t-\ell_2} u_{i_3,t-\ell_3} u_{i_4,t-\ell_4}\right]$ to be non-zero we need $i_1 = i_2$ and $i_3 = i_4$, or $i_1 = i_3$ and $i_2 = i_4$, or $i_1 = i_4$ and $i_2 = i_3$ because of Assumptions 1(ii) and 2. The same bound applies to $k = 5$ in (B.3).

Putting these bounds together,

$$\sum_{t=1}^{T} E_N\left[\chi_{T,t}(h,\kappa_T)^4\right] = \frac{\sum_{t=1}^{T} E_N\left[\bar{\chi}_{T,t}(h,\kappa_T)^4\right] g(\kappa_T^4/N^2)}{(T-h)^2 V(h,\kappa_T)^2} \leq \frac{9M_8\bar{C}^4 g(\kappa_T^4/N^2)}{(1-h/T)(T-h)V(h,\kappa_T^2)^2}.$$

Since $V(h,\kappa_T)^2/g(\kappa_T^4/N^2) \geq \underline{\mathrm{CM}}^2 > 0$, using iterated expectations we conclude that $\sum_{t=1}^{T} E_{\kappa_T}\left[\chi_{T,t}(h,\kappa_T)^4\right] = o(1)$ where the convergence is uniform over κ_T. ◻

Lemma 6. *Under the conditions of Lemma 2,*

$$\sum_{t=1}^{T-h} \frac{X_t^2 \xi_t(h,\kappa_T)^2 - E_{\kappa_T}\left[X_t^2 \xi_t(h,\kappa_T)^2 \middle| \{\theta_i, s_i\}_{i=1}^N\right]}{(T-h) g(\kappa_T^2/N)} \xrightarrow[P_{\kappa_T}]{\mathrm{p}} 0.$$

Proof. The proof is analogous to that of Lemma 4. We will show that for a constant $\bar{V}$ independent of κ_T, $\mathrm{Var}_{N,\kappa_T}\left(\sum_{t=1}^{T} X_t^2 \xi_t(h,\kappa_T)^2/g(\kappa_T^2/N)\right) \leq \bar{V}(T-h)$. By iterated expectations and Chebyshev's inequality it will follow that, for any $\varepsilon > 0$,

$$P_{\kappa_T}\left(\left|\sum_{t=1}^{T} \frac{X_t^2 \xi_t(h,\kappa_T)^2 - E_{N,\kappa_T}\left[X_t^2 \xi_t(h,\kappa_T)^2\right]}{(T-h) g(\kappa_T^2/N)}\right| > \varepsilon\right) \leq \frac{\bar{V}}{\varepsilon^2 (T-h)} \to 0.$$

We can write

$$\begin{aligned} X_t \xi_t(h,\kappa_T) &= \sum_{\ell=0}^{\infty} \iota_\ell(h) \bar{\beta}_\ell X_t X_{t+h-\ell} + \sum_{\ell=0}^{\infty} \bar{\gamma}_\ell X_t Z_{t+h-\ell} + \frac{\kappa_T}{N} \sum_{i=1}^{N}\sum_{\ell=0}^{\infty} \hat{s}_i \delta_{i\ell} X_t u_{i,t-\ell} \\ &= \underbrace{\sum_{\ell=0}^{\infty} b_\ell X_t X_{t+h-\ell}}_{\equiv g(\kappa_T/\sqrt{N})\zeta_{1,t}} + \underbrace{\sum_{\ell=0}^{\infty} c_\ell X_t Z_{t+h-\ell}}_{\equiv g(\kappa_T/\sqrt{N})\zeta_{2,t}} + \underbrace{\frac{\kappa_T}{N} \sum_{i=1}^{N}\sum_{\ell=0}^{\infty} d_{i\ell} X_t u_{i,t+h-\ell}}_{\equiv g(\kappa_T/\sqrt{N})\zeta_{3,t}}. \qquad \text{(B.4)} \end{aligned}$$

for some coefficients $\{b_\ell, c_\ell, \{d_{i\ell}\}_{i=1}^N\}$ that depend on $\{\theta_i, s_i\}_{i=1}^N$ (and h). By Assumption 3(iv), we have $|b_\ell|, |c_\ell|, |d_{i\ell}| \leq C_\ell$ almost surely for some positive finite constants C_ℓ

such that $C = \sum_{\ell=1}^{\infty} C_\ell < \infty$. Note that the coefficients, constants and variables $\zeta_{1,t}$, $\zeta_{2,t}$, $\zeta_{3,t}$ are different from the ones in the proof of Lemma 4.

Consider the variance

$$\mathrm{Var}_{N,\kappa_T}\left(\sum_{t=1}^{T} \frac{X_t^2 \xi_t(h,\kappa_T)^2}{g(\kappa_T^2/N)}\right) = \sum_{t=1}^{T-h}\sum_{\tau=1}^{T-h} \Gamma_T(t,\tau)$$

where (omitting the dependence on h, κ_T and $\{\theta_i, s_i\}_{i=1}^N$)

$$\Gamma_T(t,\tau) = \mathrm{Cov}_{N,\kappa_T}\left(\frac{X_t^2 \xi_t(h,\kappa_T)^2}{g(\kappa_T/\sqrt{N})}, \frac{X_\tau^2 \xi_\tau(h,\kappa_T)^2}{g(\kappa_T/\sqrt{N})}\right).$$

As in the proof of Lemma 4, we expand the square of $X_t^2 \xi_t(h,\kappa_T)^2$ to express $\Gamma_T(t,\tau)$ as the sum of covariances $\Gamma_{T,k_1k_2k_3k_4}(t,\tau) = \mathrm{Cov}_{N,\kappa_T}\left(\zeta_{k_1,t}\zeta_{k_2,t}, \zeta_{k_3,\tau}\zeta_{k_4,\tau}\right)$ where k_1, k_2, k_3, k_4 range over the three terms in (B.4). If $k_1 = k_2$, $\Gamma_{T,k_1k_2k_3k_4}(t,\tau)$ can only be non-zero if $k_3 = k_4$, while if $k_1 \neq k_2$, only if either $k_1 = k_3$ and $k_2 = k_4$ or $k_1 = k_4$ and $k_2 = k_3$. Then,

$$\begin{aligned}\left|\Gamma_T(t,\tau)\right| &= \sum_{k_1=1}^{3}\sum_{k_2=1}^{3}\sum_{k_3=1}^{3}\sum_{k_4=1}^{3} \left|\Gamma_{T,k_1k_2k_3k_4}(t,\tau)\right| \\ &= \sum_{k_1=1}^{3}\sum_{k_3=1}^{3} \left|\Gamma_{T,k_1k_1k_3k_3}(t,\tau)\right| + 2\sum_{k_1=1}^{3}\sum_{k_2=1}^{3} \left|\Gamma_{T,k_1k_2k_1k_2}(t,\tau)\right|. \end{aligned} \tag{B.5}$$

By calculations similar to that of Lemma 4, for any $k_1, k_2, k_3 \in \{1,2,3\}$,

$$\begin{aligned}\sum_{t=1}^{T-h}\sum_{\tau=1}^{T-h} \left|\Gamma_{T,k_1k_1k_3k_3}(t,\tau)\right| &\leq 6C^4(\bar{K} + 2M_8) \times (T-h), \\ \sum_{t=1}^{T-h}\sum_{\tau=1}^{T-h} \left|\Gamma_{T,k_1k_2k_1k_2}(t,\tau)\right| &\leq 6C^4(\bar{K} + 2M_8) \times (T-h).\end{aligned}$$

We therefore arrive at

$$\sum_{t=1}^{T}\sum_{\tau=1}^{T} \Gamma_T(t,\tau) \leq \bar{V}(T-h),$$

with $\bar{V} = 27 \times 6C^4(\bar{K} + 2M_8)$ independent of κ_T (27 is the number of terms in (B.5)). Hence, $\{(T-h)g(\kappa_T^2/N)\}^{-1} \sum_{t=1}^{T}\left(X_t^2\xi_t(h,\kappa_T)^2 - E_{N,\kappa_T}\left[X_t^2\xi_t(h,\kappa_T)^2\right]\right) = o_{P_{\kappa_T}}(1)$. □

Lemma 7. *Under the conditions of Lemma 2,*

$$\sum_{t=1}^{T-h} \frac{\left[\left(N^{-1}\sum_{i=1}^{N}\hat{x}_{it}(h)\hat{\xi}_{it}(h)\right) - X_t\xi_t(h,\kappa_T)\right]^2}{(T-h)g(\kappa_T^2/N)} \xrightarrow[P_{\kappa_T}]{\text{p}} 0.$$

Proof. We begin by writing

$$\hat{x}_{it}(h) = \hat{s}_i(X_t - \hat{X}_t(h)), \quad \hat{X}_t(h) = \bar{X}_0(h) + \hat{\pi}_X(h)'\bar{x}_t(h), \tag{B.6}$$

with $\bar{X}_0(h)$, $\hat{\pi}_X(h)$ and $\bar{x}_t(h) = (X_{t-1} - \bar{X}_1(h), \ldots, X_{t-p} - \bar{X}_p(h))'$ as in the proof of Lemma 3. As argued, $\bar{X}_0(h) = O_{P_{\kappa_T}}\left((T-h)^{-1/2}\right)$ and $\hat{\pi}_X(h) = O_{P_{\kappa_T}}\left((T-h)^{-1/2}\right)$.

Next, we write $\hat{\eta}(h)'W_{it} = \hat{\eta}_{0,i}(h) + \hat{\eta}_X(h)'\bar{x}_t(h)\hat{s}_i$ and $\eta_{X,ih} = (\beta_{i,h+1}, \ldots, \beta_{i,h+p})'$ so that

$$\hat{\xi}_{it}(h) - \xi_{it}(h,\kappa_T) = \left(\mu_i - \hat{\eta}_{0,i}(h) + \sum_{\ell=1}^{p}\beta_{i,h+\ell}\bar{X}_\ell(h)\right) + (\beta_{ih} - \hat{\beta}(h)\hat{s}_i)X_t + (\eta_{X,ih} - \hat{\eta}_X(h)\hat{s}_i)'\bar{x}_t(h)$$

and we note

$$\begin{aligned}
\begin{pmatrix}\hat{\beta}(h)\\ \hat{\eta}_X(h)\end{pmatrix} &= \left[\sum_{t=1}^{T-h}\begin{pmatrix}X_t - \bar{X}_0(h)\\ \bar{x}_t(h)\end{pmatrix}\begin{pmatrix}X_t - \bar{X}_0(h)\\ \bar{x}_t(h)\end{pmatrix}'\right]^{-1}\sum_{t=1}^{T-h}\begin{pmatrix}X_t - \bar{X}_0(h)\\ \bar{x}_t(h)\end{pmatrix}\hat{Y}_{t+h}\\
&= \begin{pmatrix}\tilde{\beta}(h)\\ \tilde{\eta}_X(h)\end{pmatrix} + \left[\sum_{t=1}^{T-h}\begin{pmatrix}X_t - \bar{X}_0(h)\\ \bar{x}_t(h)\end{pmatrix}\begin{pmatrix}X_t - \bar{X}_0(h)\\ \bar{x}_t(h)\end{pmatrix}'\right]^{-1}\sum_{t=1}^{T-h}\begin{pmatrix}X_t - \bar{X}_0(h)\\ \bar{x}_t(h)\end{pmatrix}\frac{\xi_t(h,\kappa_T)}{(N^{-1}\sum_{i=1}^{N}\hat{s}_i^2)}
\end{aligned}$$

where $\hat{Y}_{t+h} = (\sum_{i=1}^{N}\hat{s}_i^2)^{-1}\sum_{i=1}^{N}\hat{s}_iY_{i,t+h}$ and $\tilde{\eta}_X(h) = (\sum_{i=1}^{N}\hat{s}_i^2)^{-1}\sum_{i=1}^{N}\hat{s}_i\eta_{X,ih}$. Since the least squares denominator matrix when scaled by $(T-h)^{-1}$ converges to $E\left[X_t^2\right] \times I_{p+1}$ in probability uniformly over κ_T, the calculations in Lemma 3 imply that

$$\begin{aligned}
\frac{(N^{-1}\sum_{i=1}^{N}\hat{s}_i^2)(\hat{\beta}(h) - \tilde{\beta}(h))}{g(\kappa_T/\sqrt{N})} &= O_{P_{\kappa_T}}\left((T-h)^{-1/2}\right),\\
\frac{(N^{-1}\sum_{i=1}^{N}\hat{s}_i^2)(\hat{\eta}_X(h) - \tilde{\eta}_X(h))}{g(\kappa_T/\sqrt{N})} &= O_{P_{\kappa_T}}\left((T-h)^{-1/2}\right).
\end{aligned}$$

Because W_{it} includes unit effects, $\sum_{i=1}^{N}\hat{x}_{it}(h)(\hat{\eta}_{0,i}(h) - \mu_i + \sum_{\ell=1}^{p}\beta_{i,h+\ell}\bar{X}_\ell(h)) = 0$ and,

$$N^{-1}\sum_{i=1}^{N}\hat{x}_{it}(h)(\hat{\xi}_{it}(h) - \xi_{it}(h,\kappa_T)) = \left(N^{-1}\sum_{i=1}^{N}\hat{s}_i^2\right)(\tilde{\beta}(h) - \hat{\beta}(h))X_t(X_t - \hat{X}_t(h))$$

$$+\left(N^{-1}\sum_{i=1}^{N}\hat{s}_i^2\right)(\tilde{\eta}_X(h)-\hat{\eta}_X(h))'\bar{x}_t(h)(X_t-\hat{X}_t(h)). \tag{B.7}$$

To prove the Lemma, add and subtract $N^{-1}\sum_{i=1}^{N}\hat{x}_{it}(h)\xi_{it}(h,\kappa_T)$ within the squares and use Loève's inequality to obtain

$$\sum_{t=1}^{T-h}\frac{\left[\left(N^{-1}\sum_{i=1}^{N}\hat{x}_{it}(h)\hat{\xi}_{it}(h)\right)-X_t\xi_t(h,\kappa_T)\right]^2}{(T-h)g(\kappa_T^2/N)}\leq 2D_{T,2}^{\pi}(h,\kappa_T)+2D_{T,2}^{\eta}(h,\kappa_T),$$

where

$$D_{T,2}^{\pi}(h,\kappa_T)=\sum_{t=1}^{T-h}\frac{\left[N^{-1}\sum_{i=1}^{N}(\hat{s}_iX_t-\hat{x}_{it}(h))\xi_{it}(h,\kappa_T)\right]^2}{(T-h)g(\kappa_T^2/N)},$$

$$D_{T,2}^{\eta}(h,\kappa_T)=\sum_{t=1}^{T-h}\frac{\left[N^{-1}\sum_{i=1}^{N}\hat{x}_{it}(h)(\hat{\xi}_{it}(h)-\xi_{it}(h,\kappa_T))\right]^2}{(T-h)g(\kappa_T^2/N)}.$$

Inserting (B.6) into the first term and using Loève's inequality,

$$D_{T,2}^{\pi}(h,\kappa_T)\leq 2\left[\bar{X}_0(h)^2\frac{\sum_{t=1}^{T-h}\xi_t(h,\kappa_T)^2}{(T-h)g(\kappa_T^2/N)}+\|\hat{\pi}_X(h)\|^2\frac{\sum_{t=1}^{T-h}\|\bar{x}_t(h)\xi_t(h,\kappa_T)\|^2}{(T-h)g(\kappa_T^2/N)}\right],$$

where $\|\cdot\|$ is Euclidean norm. From calculations similar to those in Lemma 3,

$$\frac{\sum_{t=1}^{T-h}\xi_t(h,\kappa_T)^2}{(T-h)g(\kappa_T^2/N)}=O_{P_{\kappa_T}}(1)\ \text{and}\ \frac{\sum_{t=1}^{T-h}\|\bar{x}_t(h)\xi_t(h,\kappa_T)\|^2}{(T-h)g(\kappa_T^2/N)}=O_{P_{\kappa_T}}(1),$$

which allows us to conclude that $D_{T,2}^{\pi}(h,\kappa_T)=o_{P_{\kappa_T}}(1)$.

Inserting (B.7) into the second term and using Loève's inequality,

$$D_{T,2}^{\eta}(h,\kappa_T)\leq 2\left[\left(\frac{(N^{-1}\sum_{i=1}^{N}\hat{s}_i^2)(\tilde{\beta}(h)-\hat{\beta}(h))}{g(\kappa_T/\sqrt{N})}\right)^2\frac{\sum_{t=1}^{T-h}X_t^2(X_t-\hat{X}_t(h))^2}{T-h}\right.$$
$$\left.+\left\|\left(\frac{(N^{-1}\sum_{i=1}^{N}\hat{s}_i^2)(\tilde{\eta}_X(h)-\hat{\eta}_X(h))}{g(\kappa_T/\sqrt{N})}\right)\right\|^2\frac{\sum_{t=1}^{T-h}\|\bar{x}_t(h)(X_t-\hat{X}_t(h))\|^2}{T-h}\right].$$

Under Assumption 3(i), we can show that $(T-h)^{-1}\sum_{t=1}^{T-h}X_t^2(X_t-\hat{X}_t(h))^2=O_{P_{\kappa_T}}(1)$ and $(T-h)^{-1}\sum_{t=1}^{T-h}\|x_t(h)(X_t-\hat{X}_t(h))\|^2=O_{P_{\kappa_T}}(1)$. Thus, $D_{T,2}^{\eta}(h,\kappa_T)=o_{P_{\kappa_T}}(1)$. □

Proposition 2

Parts (A), (B) and (C) of the proof of Proposition 2 in Appendix A are established in Lemmas 8, 9 and 10 below. The argument closely resembles the proof of Proposition 1 and, therefore, in order to conserve space we only sketch the steps. Again, we adopt Assumptions 1, 2 and 3, we fix p and assume $h_T/T \leq \phi < 1$ as $T, N \to \infty$.

Lemma 8 (Asymptotic normality of the score).

$$\frac{\sum_{t=1}^{T-h_T} X_t \xi_t(h_T, \kappa_T)}{\sqrt{(T-h_T)V(h_T, \kappa_T)}} \xrightarrow[P_{\kappa_T}]{\text{d}} N(0,1).$$

Proof. The proof given for Lemma 1 goes through with the following adjustment: we can remove the terms $\bar{\beta}_\ell, \bar{\gamma}_\ell, \delta_{i\ell}$ from $\Xi_{X,t}(h,\kappa)$ whenever $\ell > h$. That is, we set

$$\Xi_{X,t}(h,\kappa) = \sum_{\ell=1}^{h} \mathbb{1}\{t-\ell \geq 1\}\bar{\beta}_{h-\ell} X_{t-\ell} + \mathbb{1}\{t \leq T-h\}\left[\bar{\gamma}_h Z_t + \frac{\kappa}{N}\sum_{i=1}^{N} \hat{s}_i \delta_{ih} u_{it}\right].$$

The calculations in Lemmas 4 and 5 apply with the same adjustment. In Lemma 4, $\bar{V} \leq 75 \times 6C^8(\bar{K} + 2M_8)/(1-\phi)$, which does not depend on κ_T or h_T. Similarly, in Lemma 5, $\sum_{t=1}^{T} E_N\left[\chi_{T,t}(h_T,\kappa_T)^4\right] \leq 9M_8C^8/(1-\phi)^2\underline{\text{CM}}^2T$, which tends to zero as $T \to \infty$ uniformly over κ_T and h_T. □

Lemma 9 (Consistency of the standard error).

$$\frac{\hat{V}(h_T)}{V(h_T,\kappa_T)} \xrightarrow[P_{\kappa_T}]{\text{p}} 1.$$

Proof. The proofs of Lemma 2 and auxiliary Lemma 6 go through without change. To establish the equivalent to Lemma 7 in this context, define $\bar{x}_t(h_T)$ as in its proof and let $\bar{y}_{it}(h_T) = (\hat{Y}_{i,t-1}(h_T), \ldots, \hat{Y}_{i,t-p}(h_T))$ with $\hat{Y}_{i,t-\ell}(h_T)$ the residual from regressing $g(\kappa_T)^{-1}Y_{i,t-\ell}$ on unit and time effects. We can write

$$\hat{\pi}(h_T)'W_{it} = \hat{s}_i\bar{X}_0(h_T) + \hat{s}_i\hat{\pi}_X(h_T)'\bar{x}_t(h_T) + \hat{\pi}_Y(h_T)'\bar{y}_{it}(h_T),$$
$$\hat{\eta}(h_T)'W_{it} = \hat{\eta}_{0,i}(h_T) + \hat{s}_i\hat{\eta}_X(h_T)'\bar{x}_t(h_T) + \hat{\eta}_Y(h_T)'\bar{y}_{it}(h_T).$$

Scaling $Y_{i,t-\ell}$ by $g(\kappa_T)^{-1}$ leaves the least square predictions $\hat{\pi}(h_T)'W_{it}$ and $\hat{\eta}(h_T)'W_{it}$ unchanged, but it helps bound them in probability uniformly over κ_T.

Calculations similar to those in Lemma 3 deliver

$$\begin{pmatrix} \bar{X}_0(h_T) \\ \hat{\pi}_X(h_T) \\ \hat{\pi}_Y(h_T) \end{pmatrix} = O_{P_{\kappa_T}}\left((T-h_T)^{-1/2}\right),$$

$$g\left(\frac{\kappa_T}{\sqrt{N}}\right)^{-1} \begin{pmatrix} (\hat{\beta}(h_T) - \tilde{\beta}(h_T)) \\ (\hat{\eta}_X(h_T) - \tilde{\eta}_X(h_T)) \\ (\hat{\eta}_Y(h_T) - \tilde{\eta}_Y(h_T)) \end{pmatrix} = O_{P_{\kappa_T}}\left((T-h_T)^{-1/2}\right),$$

where $\tilde{\eta}_X(h_T) = (\tilde{B}_1(h_T), \ldots, \tilde{B}_p(h_T))'$ and $\tilde{\eta}_Y(h_T) = g(\kappa_T)(A_1(h_T), \ldots, A_p(h_T))'$ with $A_\ell(h)$ and $\tilde{B}_\ell(h)$ as defined in the proof of Proposition 2 in Appendix A.

The rest of the proof follows the steps of Lemma 7. The convergence is uniform in both κ_T and h_T because $T - h_T \leq (1-\phi)T$ with $\phi < 1$. □

Lemma 10 (Negligibility of the remainder).

$$R_T(h_T, \kappa_T) \xrightarrow[P_{\kappa_T}]{\text{p}} 0.$$

Proof. We begin by defining $\bar{x}_t(h_T)$ and $\bar{y}_{it}(h)$ as in Lemma 9, by writing

$$\hat{\pi}(h_T)'W_{it} = \hat{s}_i \bar{X}_0(h_T) + \hat{s}_i \hat{\pi}_X(h_T)'\bar{x}_t(h_T) + \hat{\pi}_Y(h_T)'\bar{y}_{it}(h_T),$$

and by noting again that

$$\begin{pmatrix} \bar{X}_0(h_T) \\ \hat{\pi}_X(h_T) \\ \hat{\pi}_Y(h_T) \end{pmatrix} = O_{P_{\kappa_T}}\left((T-h_T)^{-1/2}\right).$$

Next, we write $r_{it}(h_T) = (\beta_{ih} - \tilde{\beta}(h)\hat{s}_i)X_t + \sum_{\ell=1}^{p}(B_{i\ell}(h) - \tilde{B}_\ell(h)\hat{s}_i)X_{t-\ell}$ and

$$R_T(h_T, \kappa_T) = -\frac{\bar{X}_0(h_T)\sum_{t=1}^{T-h_T}\xi_t(h_T, \kappa_T)}{\sqrt{(T-h_T)V(h_T, \kappa_T)}} - \frac{\hat{\pi}_X(h_T)'\sum_{t=1}^{T-h_T}\bar{x}_t(h_T)\xi_t(h_T, \kappa_T)}{\sqrt{(T-h_T)V(h_T, \kappa_T)}}$$
$$- \frac{\hat{\pi}_Y(h_T)'\sum_{i=1}^{N}\sum_{t=1}^{T-h_T}\bar{y}_{it}(h_T)(r_{it}(h_T) + \xi_{it}(h_T, \kappa_T))}{N\sqrt{(T-h_T)V(h_T, \kappa_T)}}$$

The rest of the argument mimics the proof of Lemma 3. □

Proposition 3

Parts (A), (B) and (C) of the proof of Proposition 3 in Appendix A are stated in Lemmas 11, 12 and 13 below. The proofs are virtually identical to their counterparts in Proposition 1 with some minor differences. Here we make Assumptions 4 and we hold h and $p \geq h$ fixed as $T, N \to \infty$.

Lemma 11 (Asymptotic normality of the score).

$$\frac{\sum_{t=1}^{T-h} \boldsymbol{\lambda}' \boldsymbol{X}_t^* \xi_t(h, \kappa_T)}{\sqrt{(T-h)\boldsymbol{\lambda}' \boldsymbol{V}(h, \kappa_T)\boldsymbol{\lambda}}} \xrightarrow[P_{\kappa_T}]{\text{d}} N(0,1).$$

Proof. The arguments given for Lemma 1 and auxiliary Lemmas 4 and 5 apply with the obvious change in notation. □

Lemma 12 (Consistency of the standard error and OLS denominator).

$$\frac{\boldsymbol{\lambda}' \hat{\boldsymbol{V}}^{\text{IV}}(h)\boldsymbol{\lambda}}{\boldsymbol{\lambda}' \boldsymbol{V}(h, \kappa_T)\boldsymbol{\lambda}} \xrightarrow[P_{\kappa_T}]{\text{p}} 1 \text{ and } \hat{\boldsymbol{J}}^{\text{IV}}(h) \xrightarrow[P_{\kappa_T}]{\text{p}} \boldsymbol{J}.$$

Proof. The first part follows from arguments analogous to those given for Lemma 2 and auxiliary Lemmas 6 and 7 (with obvious notational changes). For the second part, note $\text{Var}_{N,\kappa_T}\left(X_t^* \tilde{X}_t\right) \leq \bar{V}/(T-h)$ for some constant $\bar{V}$ independent of κ_T under Assumption 4(ii), so that $\left\|\hat{\boldsymbol{J}}^{\text{IV}}(h) - \boldsymbol{J}\right\| = o_{P_{\kappa_T}}(1)$ follows from iterated expectations and Chebyshev's inequality. □

Lemma 13 (Negligibility of the remainder).

$$R_T(h, \kappa_T) \xrightarrow[P_{\kappa_T}]{\text{p}} 0.$$

Proof. For any $\boldsymbol{\lambda} \neq 0_{(p+1)\times 1}$, by the same calculations as in Lemma 3,

$$\frac{\sum_{t=1}^{T-h} \boldsymbol{\lambda}' \boldsymbol{X}_t^*}{(T-h)} = O_{P_{\kappa_T}}\left((T-h)^{-1/2}\right) \text{ and } \frac{\sum_{t=1}^{T-h} \xi_t(h, \kappa_T)}{\sqrt{(T-h)\boldsymbol{\lambda}' \boldsymbol{V}(h, \kappa_T)\boldsymbol{\lambda}}} = O_{P_{\kappa_T}}(1)\,.$$

Since $\hat{\boldsymbol{J}}^{\text{IV}}(h) = \boldsymbol{J} + o_{P_{\kappa_T}}(1)$ by the second part of Lemma 12, the result follows. □

C Details of simulation study

Here we complement Section 4 with additional details. First, we describe how we simulate the heterogeneity. Second, we specify the calibration of our DGPs. Third and last, we present further simulation results.

Simulation of observable and unobservable heterogeneity. A primary feature is the correlation between s_i and $\{\beta_i, \gamma_i.\delta_i\}$.[3] We begin by drawing the vector

$$(s_i, s_{\gamma,i}, s_{\delta,i})' \sim N\left(1_{3\times1}, (1-\rho)I_3 + \rho 1_{3\times3}\right)$$

for some $\rho \neq 0$. Next, we set a very large $\bar{L}$ and compute

$$\beta_{i\ell} = s_i \breve{\beta}_{i\ell}, \quad \gamma_{i\ell} = s_{\gamma,i}\breve{\gamma}_{i\ell}, \quad \delta_{i\ell} = s_{\delta,i}\breve{\delta}_{i\ell},$$

where $\{\breve{\beta}_{i\ell}, \breve{\gamma}_{i\ell}, \breve{\delta}_{i\ell}\}_{\ell=0}^{\bar{L}}$ are obtained by (a) drawing the roots of ARMA polynomials from Beta distributions, (b) computing their MA(∞) representations, (c) truncating them at $\bar{L}$, and (d) normalizing them so that $\sum_{\ell=0}^{\bar{L}} \breve{\beta}_{i\ell}^2 = \sum_{\ell=0}^{\bar{L}} \breve{\gamma}_{i\ell}^2 = \sum_{\ell=0}^{\bar{L}} \breve{\delta}_{i\ell}^2 = 1$.[4]

To generate time-varying heterogeneity we set $s_{it} = s_i + \zeta_{it}$ with $\zeta_{it} \sim N(0,1)$, i.i.d. over units and time, and independent of s_i and everything else. This ensures s_{it} remains exogenous with respect to aggregate and idiosyncratic shocks.

Finally, in the VAR DGP, we set

$$B_{i\ell} = s_i \breve{B}_{i\ell}, \quad C_{i0} = s_{\gamma,i}, \quad D_{i0} = s_{\delta,i}.$$

where $\{\breve{B}_{i\ell}\}_{\ell=0}^{\bar{L}}$ are obtained in the same way as $\{\breve{\beta}_{i\ell}\}_{\ell=0}^{\bar{L}}$ above.

Our method does not satisfy Assumption 3(iv), although responses are bounded with sufficiently high probability that it does not seem to make a difference.

[3] Instead, μ_i (and m_i in the VAR setup) does not play a big role and we simply draw it as $N(0,1)$.

[4] The advantage of this representation is that it separates the scale and persistence. For example, if X_t is white noise with unit variance conditional on $\{\beta_{i\ell}\}_{\ell=0}^{\bar{L}}$, the variance of $\sum_{\ell=0}^{\bar{L}} \beta_{i\ell} X_{t-\ell}$ is $\sum_{\ell=0}^{\bar{L}} \beta_{i\ell}^2 = s_i^2$ while the ratio of long-run variance to variance of $\sum_{\ell=0}^{\bar{L}} \beta_{i\ell} X_{t-\ell}$ (a measure of persistence) is

$$\frac{\left(\sum_{\ell=0}^{\bar{L}} \beta_{i\ell}\right)^2}{\sum_{\ell=0}^{\bar{L}} \beta_{i\ell}^2} = \frac{\left(\sum_{\ell=0}^{\bar{L}} \breve{\beta}_{i\ell}\right)^2}{\sum_{\ell=0}^{\bar{L}} \breve{\beta}_{i\ell}^2},$$

which does not depend on s_i.

DGP calibration. In the general DGP, we set $\rho = 0.5$, and generate $\{\breve{\beta}_{i\ell}, \breve{\gamma}_{i\ell}, \breve{\delta}_{i\ell}\}_{\ell=0}^{L}$ from ARMA(4, 2) processes with expected roots $(0.7, 0.3, 0.2, 0.1)$ and $(0, 0)$ for $\breve{\beta}_{i\ell}$, $(0.7, 0.2, 0.1, -0.2)$ and $(0.2, -0.2)$ for $\breve{\gamma}_{i\ell}$, and $(0.9, 0.3, 0.1, 0.1)$ and $(0.5, 0.2)$ for $\breve{\delta}_{i\ell}$. We draw each root as $\text{Beta}(\bar{\lambda}\nu, (1-\bar{\lambda})\nu)$ where $\bar{\lambda}$ is the mean listed above and $\nu = 10$, and we truncate polynomials at $\bar{L} = 2T$ lags.

In the LP-IV case, we use a similar method for $\{b_\ell, c_\ell\}_{\ell=0}^{L}$. We obtain b_ℓ from an ARMA(1, 1) with roots 0.3 and -0.2, and c_ℓ from an ARMA(2, 2) with roots $(0.4, 0.2)$ and $(0.1, -0.1)$. We also set $a_0 = 10$ to be safely above standard weak IV thresholds.

Finally, for the VAR DGP, we draw $\{\breve{B}_{i\ell}\}_{\ell=0}^{p}$ from an MA(2) with roots $(0.8, -0.5)$ and $\nu = 10$, and we set $\{A_\ell\}_{\ell=1}^{p}$ to an AR(2) with roots $(1 - 5/T, 0.5)$.

The mean and quantiles of responses for each horizon can be seen in Figure C.1.

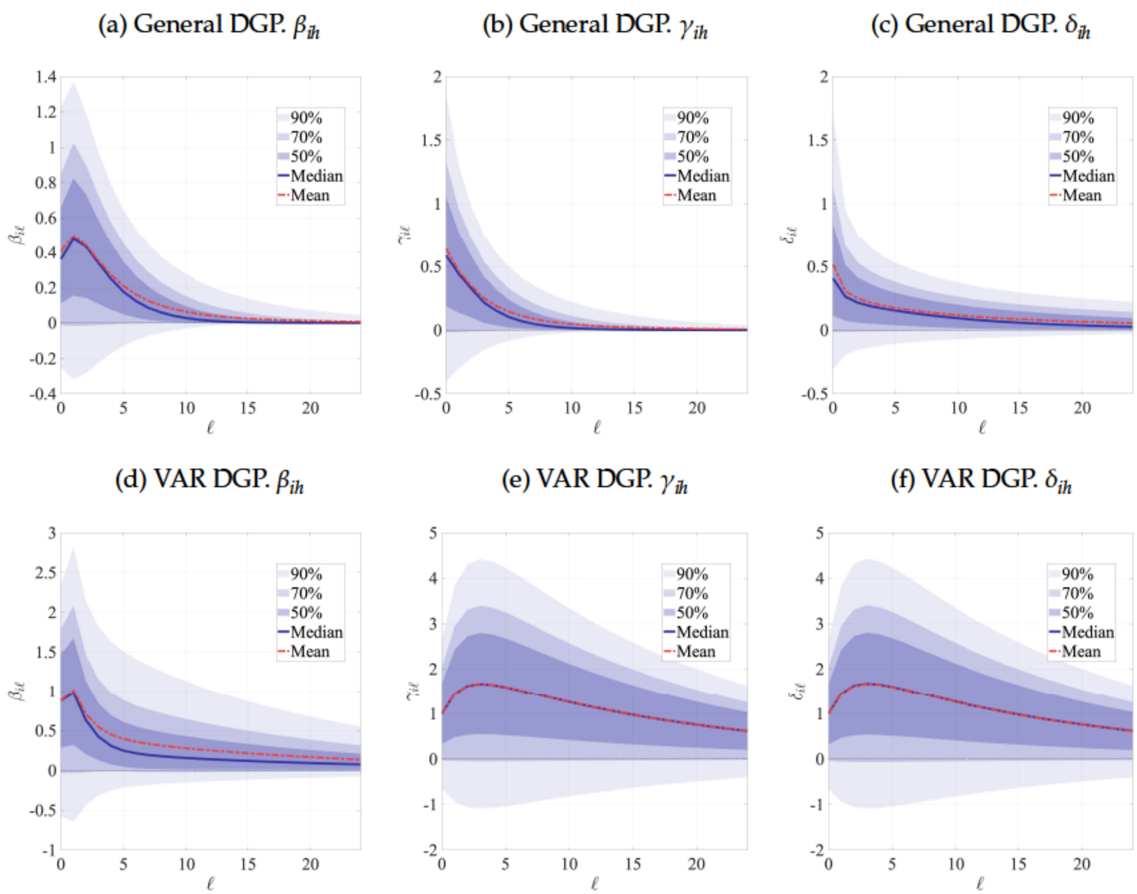

FIGURE C.1. Distributions of impulse responses across ℓ for general and VAR DGPs.

Additional results. Figure C.2 presents coverage rates of 90% confidence intervals in the general DGP with $T = 100$ for panel LPs on X_t (panels (a)-to-(c)) and on $s_{it}X_t$ (panels (d)-to-(f)).[5] As mentioned in the paper, the estimands are different: LPs on X_t recover the mean impulse response while LPs on $s_{it}X_t$ recover their projection on s_{it}. Yet, the observations we made about inference from Section 4 are unchanged. In particular, t-LAHR inference dominates all the alternatives in delivering correct coverage for the nonparametric panel local projection estimand.

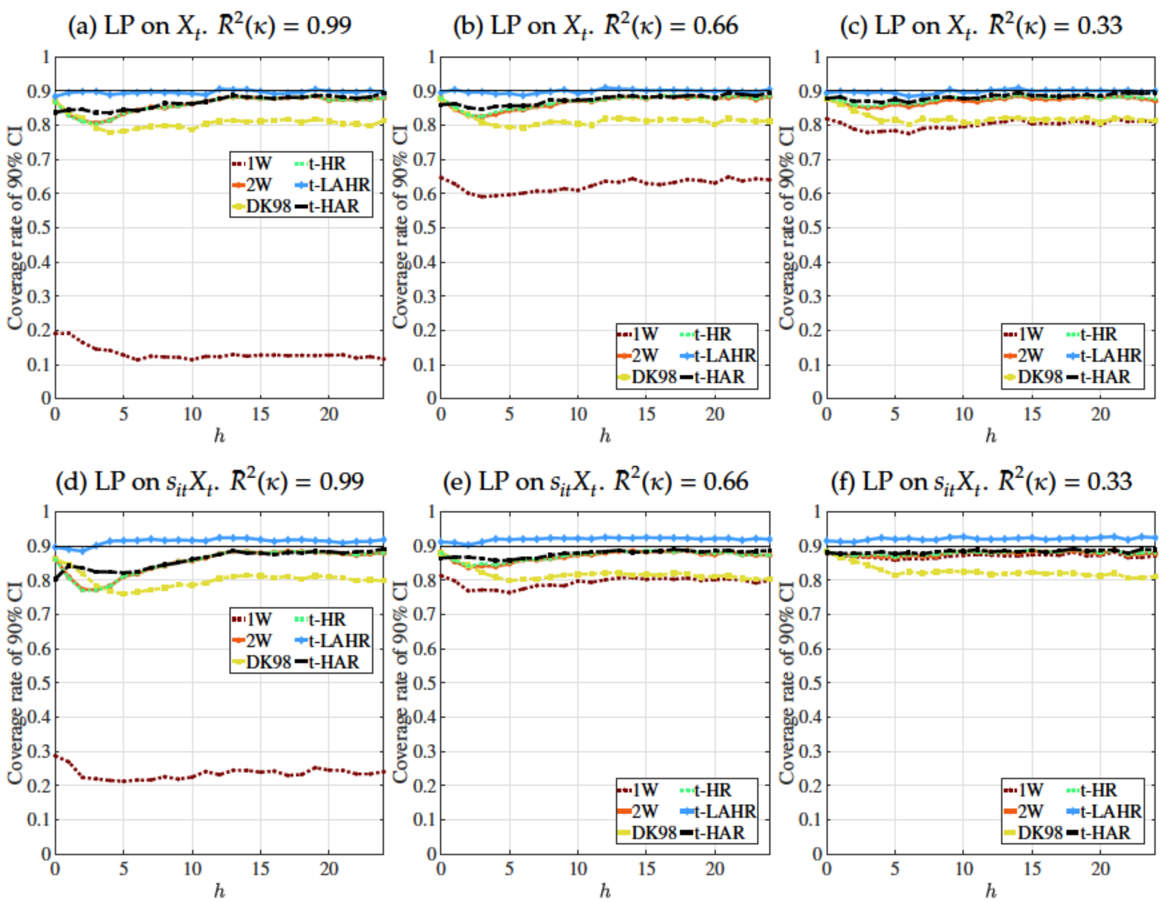

FIGURE C.2. Coverage rates of 90% confidence intervals for $T = 100$.

Note: 1W refers to one-way (unit-level) clustering, 2W to two-way clustering, DK98 to Driscoll–Kraay, and t-HR/t-LAHR/t-HAR to the time-level clustering approaches discussed in the text.

[5] For panel LPs on X_t time effects are excluded from the vector of controls. Otherwise, the estimation and inference procedures are the same as in Figure 1 in the paper.

D A survey of empirical applications

Below, we survey relevant empirical applications by the method used to calculate standard errors. The list reflects the recent surge in applications (with the oldest paper dated 2018) and includes both published work and working papers. We have aimed to make the list comprehensive, but it is possible that some might have been inadvertently omitted. When different methods were used, we favored the one used in the main specification and the one used in estimation of dynamic effects (non-zero horizons). We classified as one-way clustering (within units) applications that cluster at a higher level of aggregation than primary units; say, at the industry (or industry-time) level when units are firms. While allowing for sector-level shocks, these still rule out economy-wide spatial dependence. See the Introduction for additional details.

By method	
Two-way clustering (within units and time)	Ippolito, Ozdagli, and Perez-Orive (2018), Jeenas (2019), Ottonello and Winberry (2020), Amberg, Jansson, Klein, and Rogantini Picco (2022), Palazzo and Yamarthy (2022), Paz (2022), Bellifemine, Couturier, and Jamilov (2023), Cascaldi-Garcia, Vukotić, and Zubairy (2023), Drechsel (2023), Durante, Ferrando, and Vermeulen (2022), Duval, Furceri, Lee, and Tavares (2023), Ferreira, Ostry, and Rogers (2023), González, Nuño, Thaler, and Albrizio (2023), Lakdawala and Moreland (2023), Singh, Suda, and Zervou (2023), Thürwächter (2023), Zhou (2023), Anderson and Cesa-Bianchi (2024), Berthold, Cesa-Bianchi, Di Pace, and Haberis (2024), Caglio, Darst, and Kalemli-Özcan (2024), Camêlo (2024), Gulyas, Meier, and Ryzhenkov (2024), Paranhos (2024), Lakdawala and Moreland (forthcoming)
Clustering within units	Wu (2018), Ozdagli (2018), Crouzet and Mehrotra (2020), Singh, Suda, and Zervou (2022), Albrizio, González, and Khametshin (2023), Andersen, Johannesen, Jørgensen, and Peydró (2023), Camara and Ramirez Venegas (2023), Ghomi (2023), Indarte (2023), Bardóczy, Bornstein, Maggi, and Salgado (2024), Jeenas (2024), Jeenas and Lagos (2024), Lo Duca, Moccero, and Parlapiano (2024), Paranhos (2024), Ruzzier (2024)
Driscoll and Kraay (1998) standard errors	Holm, Paul, and Tischbirek (2021), Bahaj, Foulis, Pinter, and Surico (2022), Cloyne, Ferreira, Froemel, and Surico (2023), Fagereng, Gulbrandsen, Holm, and Natvik (2023), Gorea, Kryvtsov, and Kudlyak (2023), Bilal and Känzig (2024), Cao, Hegna, Holm, Juelsrud, König, and Riiser (2024)
Clustering within time	Gürkaynak, Karasoy-Can, and Lee (2022)

References

Albrizio, S., B. González, and D. Khametshin (2023): "A Tale of Two Margins: Monetary Policy and Capital Misallocation." Working paper.

Amberg, N., T. Jansson, M. Klein, and A. Rogantini Picco (2022): "Five Facts about the Distributional Income Effects of Monetary Policy Shocks." *AER: Insights*, 4, 289–304.

Andersen, A. L., N. Johannesen, M. Jørgensen, and J.-L. Peydró (2023): "Monetary Policy and Inequality." *Journal of Finance*, 78, 2945–2989.

Anderson, G. and A. Cesa-Bianchi (2024): "Crossing the Credit Channel: Credit Spreads and Firm Heterogeneity." *American Economic Journal: Macroeconomics*, 16, 417–446.

Bahaj, S., A. Foulis, G. Pinter, and P. Surico (2022): "Employment and the Residential Collateral Channel of Monetary Policy." *Journal of Monetary Economics*, 131, 26–44.

Bardóczy, B., G. Bornstein, C. Maggi, and S. Salgado (2024): "Monopsony Power and the Transmission of Monetary Policy." Working paper.

Bellifemine, M., A. Couturier, and R. Jamilov (2023): "The Regional Keynesian Cross." Working paper.

Berthold, B., A. Cesa-Bianchi, F. Di Pace, and A. Haberis (2024): "The Heterogeneous Effects of Carbon Pricing: Macro and Micro Evidence." Working paper.

Bilal, A. and D. R. Känzig (2024): "The Macroeconomic Impact of Climate Change: Global vs. Local Temperature." NBER Working Paper 32450.

Caglio, C. R., R. M. Darst, and c. Kalemli-Özcan (2024): "Collateral Heterogeneity and Monetary Policy Transmission: Evidence from Loans to SMEs and Large Firms." Working paper.

Camara, S. and S. Ramirez Venegas (2023): "The Transmission of US Monetary Policy Shocks: The Role of Investment & Financial Heterogeneity." Working paper.

Camêlo, F. (2024): "Capital Allocation and Firm Dynamics in Small Open Economies." Working paper.

Cao, J., T. Hegna, M. B. Holm, R. Juelsrud, T. König, and M. Riiser (2024): "The Investment Channel of Monetary Policy: Evidence from Norway." Working paper.

Cascaldi-Garcia, D., M. Vukotić, and S. Zubairy (2023): "Innovation During Challenging Times." Working paper.

Cloyne, J., C. Ferreira, M. Froemel, and P. Surico (2023): "Monetary Policy, Corporate Finance, and Investment." *Journal of the European Economic Association*, 21, 2586–2634.

Crouzet, N. and N. R. Mehrotra (2020): "Small and Large Firms over the Business Cycle." *American Economic Review*, 110, 3549–3601.

Davidson, J. (1994): *Stochastic Limit Theory: An Introduction for Econometricians*, Oxford University Press, 2nd edition ed.

Drechsel, T. (2023): "Earnings-Based Borrowing Constraints and Macroeconomic Fluctuations." *American Economic Journal: Macroeconomics*, 15, 1–34.

Driscoll, J. and A. Kraay (1998): "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data." *Review of Economics and Statistics*, 80, 549–560.

Durante, E., A. Ferrando, and P. Vermeulen (2022): "Monetary Policy, Investment and Firm Heterogeneity." *European Economic Review*, 148.

Duval, R., D. Furceri, R. Lee, and M. M. Tavares (2023): "Market Power, Financial Constraints, and Monetary Transmission." *AEA Papers and Proceedings*, 113, 99–104.

Fagereng, A., M. A. H. Gulbrandsen, M. B. Holm, and G. J. Natvik (2023): "Fisher Effects, Monetary Policy, and Household Indebtedness." Working paper.

Ferreira, T. R., D. A. Ostry, and J. Rogers (2023): "Firm Financial Conditions and the Transmission of Monetary Policy." Working paper.

Ghomi, M. (2023): "Monetary Policy, Economic Uncertainty, and Firms' R&D Expenditure." Working paper.

González, B., G. Nuño, D. Thaler, and S. Albrizio (2023): "Firm Heterogeneity, Capital Misallocation and Optimal Monetary Policy." Working paper.

Gorea, D., O. Kryvtsov, and M. Kudlyak (2023): "House Price Responses to Monetary Policy Surprises: Evidence from the U.S. Listings Data." Working paper, European Investment Bank, Bank of Canada, Federal Reserve Bank of San Francisco.

Gulyas, A., M. Meier, and M. Ryzhenkov (2024): "Labor Market Effects of Monetary Policy Across Workers and Firms." *European Economic Review*, 166.

Gürkaynak, R., H. G. Karasoy-Can, and S. S. Lee (2022): "Stock Market's Assessment of Monetary Policy Transmission: The Cash Flow Effect." *Journal of Finance*, 77, 2027–2073.

Holm, M. B., P. Paul, and A. Tischbirek (2021): "The Transmission of Monetary Policy under the Microscope." *Journal of Political Economy*, 129, 2861–2904.

Indarte, S. (2023): "Financial Crises and the Transmission of Monetary Policy to Consumer Credit Markets." *Review of Financial Studies*, 36, 4045–4087.

Ippolito, F., A. K. Ozdagli, and A. Perez-Orive (2018): "The Transmission of Monetary Policy through Bank Lending: The Floating Rate Channel." *Journal of Monetary Economics*, 95, 49–71.

Jeenas, P. (2019): "Monetary Policy Shocks, Financial Structure, and Firm Activity: A Panel Approach." Working paper.

——— (2024): "Firm Balance Sheet Liquidity, Monetary Policy Shocks, and Investment Dynamics." Working paper.

Jeenas, P. and R. Lagos (2024): "Q-Monetary Transmission." *Journal of Political Economy*, 132.

Lakdawala, A. and T. Moreland (2023): "The Effect of Monetary Policy on Firm-Level Uncertainty." *Economics Letters*, 232.

——— (forthcoming): "Firm-Level Uncertainty and the Transmission of Monetary Policy." *Review of Economics and Statistics*.

Lo Duca, M., D. Moccero, and F. Parlapiano (2024): "The Impact of Macroeconomic and Monetary Policy Shocks on Credit Risk in the Euro Area Corporate Sector." Working paper.

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

Ottonello, P. and T. Winberry (2020): "Financial Heterogeneity and the Investment Channel of Monetary Policy." *Econometrica*, 88, 2473–2502.

Ozdagli, A. K. (2018): "Financial Frictions and the Stock Price Reaction to Monetary Policy." *Review of Financial Studies*, 31, 3895–3936.

Palazzo, B. and R. Yamarthy (2022): "Credit Risk and the Transmission of Interest Rate Shocks." *Journal of Monetary Economics*, 130, 120–136.

Paranhos, L. (2024): "How Do Firms' Financial Conditions Influence the Transmission of Monetary Policy? A Non-parametric Perspective." Working paper.

Paz, P. (2022): "Bank Capitalization Heterogeneity and Monetary Policy." Working paper.

Ruzzier, G. (2024): "Specialized Banks and the Transmission of Monetary Policy: Evidence from the U.S. Syndicated Loan Market." Working paper.

Singh, A., J. Suda, and A. Zervou (2022): "Monetary Policy, Labor Market, and Sectoral Heterogeneity." *AEA Papers and Proceedings*, 112, 491–495.

——— (2023): "Heterogeneity in labor market response to monetary policy: small versus large firms." Working paper.

Thürwächter, C. (2023): "Firm Heterogeneity and Monetary Policy Transmission." Working paper.

Wu, W. (2018): "The Credit Channel at the Zero Lower Bound through the Lens of Equity Prices." *Journal of Money, Credit and Banking*, 50, 423–452.

Zhou, W. (2023): "The Firm Balance Sheet Channel of Uncertainty Shocks." Working paper.