STAFF
REPORTS

NO. 1090
MARCH 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
https://doi.org/10.59576/sr.1090

Abstract

We study estimation and inference in panel data regression models when the regressors of interest are macro shocks, which speaks to a large empirical literature that targets impulse responses via local projections. Our results hold under general dynamics and are uniformly valid over the degree of signal-to-noise of aggregate shocks. We show that the regression scores feature strong cross-sectional dependence and a known autocorrelation structure induced only by leads of the regressor. In general, including lags as controls and then clustering over the cross-section leads to simple, robust inference.

JEL classification: C32, C33, C38, C51
Key words: panel data, local projections, impulse responses, aggregate shocks, inference, 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 comments and discussion with Manuel Arellano, Dmitry Arkhangelsky, Mikkel Plagborg-Moller, and Enrique Sentana. They also thank Stephane Bonhomme, Oscar Jorda, Daniel Lewis, Geert Mesters, seminar participants at CEMFI, Erasmus University Rotterdam, Georgetown University, Princeton University, Universidad Autonoma de Madrid, conference participants at the EABCN-UPF conference, and the Greater New York Econometrics Colloquium for valuable comments and discussions. Víctor Sancibrián gratefully acknowledges funding from Fundacion Ramon Areces.

1 Introduction

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

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 monetary policy or oil price shocks, which satisfy certain exogeneity conditions. 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).[1]

Despite its routine application, a formal treatment of estimation and inference for these type of problems is lacking: we document below several conflicting views on the appropriate way to compute standard errors and confidence intervals, on the relative merits of each dimension of the panel for precision, and on the role of covariates s_i as sources of additional variation. Our paper fills this gap.

We study these issues in a comprehensive setup that explicitly acknowledges the micro-macro nature of the problem and features cross-sectional heterogeneity in responses and general forms of serial dependence in outcomes.

In a nutshell, we show that inference in regressions with macro shocks has two main features, namely strong spatial dependence and limited serial correlation. The former suggests that statistical uncertainty in estimates can only be dominated with more abundant time-series variation; the latter obviates the need to account for general patterns of correlation in unit-level residuals and holds true even if unit-level noise is pervasive. On a practical level, this suggests a simple approach based on clustering at the time level and accounting for the leftover autocorrelation terms by including lagged outcomes and shocks as controls.

[1] See also Jordà (2023) for an updated, accessible review of methods for local projections.

First, we show that in what looks like an otherwise standard panel data model, the precision at which β_h can be recovered is of order $T^{-1/2}$ even in situations where $N \gg T$, driven by the simultaneous and hardly avoidable presence of both observed and unobserved sources of aggregate variation. This is a manifestation of strong cross-sectional dependence — randomness that is common to all units in the panel — and inherent to local projections. We suggest caution regarding the conventional wisdom in these applications that a larger cross-sectional dimension necessarily compensates for a shorter time series.

Second, for a panel local projection at horizon h, we show that the regression scores have the serial dependence structure of a moving average process of order h, or MA(h), thanks to the unpredictable nature of X_t and in spite of richer and more complicated dynamics in outcomes. Importantly, the variance of the score depends on all omitted sources of micro and macro variation, whereas the autocovariances at different lags only depend on leads of the macro shock X_t up to horizon h.

These results hold true over data generating processes with different degrees of signal-to-noise of macro shocks, capturing empirically realistic environments where idiosyncratic noise might be pervasive. Standard asymptotic plans where $N, T \to \infty$ imply that unit-level dynamics are always small relative to aggregate ones and thus might not need to be accounted for. But neglecting the contribution of unit-level variation might lead to poor approximations in small samples. Because of that, we introduce an asymptotic framework where the signal value of aggregates remains low in the limit by allowing for a local-to-zero relative standard deviation of macro to micro shocks. Under this embedding, the limiting autocovariance function of the scores might feature additional micro-level terms, but crucially these do not show up at nonzero lags. Since the degree of signal-to-noise induces a discontinuity in the asymptotic distribution of panel local projection estimators, we prove that our approximations are uniformly valid over this parameter.

Third, we show that asymptotically valid inference can be achieved by first aggregating the residual regression scores over the cross-section, and then computing heteroskedasticity and autocorrelation robust (HAR) standard errors with up to h lags on the resulting synthetic time series. This approach to inference already contains a first layer of simplicity; clustering at the unit level in the micro

data is not necessary, and neither is a conservative HAR approach at moderate horizons. Additionally, when a vector autoregressive structure (VAR) is imposed and local projections are augmented with lagged outcomes and regressors as unit-by-unit controls, we show that the regression scores are serially uncorrelated, and heteroskedasticity-robust (HR) standard errors on the aggregated data deliver valid inference. This is reminiscent of the results by Montiel Olea and Plagborg-Møller (2021) in the time series context. The aggregation step over the cross-sectional dimension is critical: it amounts to clustering at the time level in the micro data. In fact, inference that exploits independence across units can be reinterpreted as inference that conditions on the realized path of aggregate shocks.

Fourth, we clarify the dual role of the cross-sectional variable s_i as descriptor of heterogeneity in the transmission of macro shocks and as source of additional precision. Our setup features individual heterogeneity in impulse responses whose conditional distribution given covariates we do not restrict. This allows us to formalize what practitioners have in mind when including interactions similar to those in (1). In particular, β_h describes the slope of a linear projection of heterogeneous responses onto s_i; letting $s_i = 1$ recovers the average response over the panel units, as expected. On the other hand, only under special conditions one can conceive s_i as devices to obtain large gains in statistical precision. Essentially, these need to satisfy exogeneity conditions akin to cross-sectional instruments: orthogonality with respect to heterogeneous exposures of other shocks is required. We also discuss extensions to state-dependent controls of the form s_{it}. These require allowing for time-varying impulse responses but entail no conceptual difference.

A key element in our framework is the availability of an observed macro shock X_t. We assume that X_t is mean independent of all other shocks, including its own lags and leads. This is rooted in empirical practice — similar definitions of shocks are routinely being made in the literature, and a certain form of unpredictability is nonetheless necessary for identification and consistent estimation of impulse responses. For instance, this is natural under the interpretation of shocks as unanticipated structural disturbances. Similar notions of unpredictability have been considered in the time series literature on local projections inference, see for instance Stock and Watson (2018) and Montiel Olea and Plagborg-Møller (2021).

These assumptions can be relaxed when the econometrician is willing to impose VAR dynamics on outcomes. Our results also cover endogeneity settings (local projection-instrumental variable estimators; LP-IV for short) where the structural innovations of interest are observed with noise but an instrument satisfying analogous conditions is available.

In practice, it is sensible to expect lag augmentation to approximate well the underlying covariance structure given a conservative lag length choice, which suggests clustering over the cross-sectional dimension as a reasonable and simple approach to inference. Moreover, as discussed above, any remaining components inducing serial correlation in the regression scores are likely to be relatively small, even more so when idiosyncratic noise is pervasive. We complement our theoretical results with simulations for realistic designs and sample sizes, including responses at moderately long horizons and a substantial degree of persistence in micro shocks. We study the performance of a battery of approaches to inference, incorporating finite-sample refinements proposed in the literature (Müller, 2004; Imbens and Kolesár, 2016; Lazarus, Lewis, Stock, and Watson, 2018). We find that HR inference on the aggregated scores displays a remarkable performance relative to HAR approaches, even if the latter uses more liberal lag choices and even if we do not impose a VAR structure on outcomes. Our general practical recommendation is to compute standard errors that cluster at the time level together with the refinement proposed by Imbens and Kolesár (2016) in local projections with a conservative number of lags as controls.

The state of empirical practice. We review a large body of empirical applications that precede our work. 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.[2]

[2]We have reviewed almost 40 recent empirical papers that fall within our framework: micro data, macro shocks, and local projections. The economic content of X_t is very diverse, including fiscal policy shocks, investment shocks, total factor productivity and innovation shocks, carbon pricing shocks, etc. Our list includes both published work and working papers (from 2019) and is available upon request.

In otherwise comparable empirical designs, we document large dispersion in the way practitioners compute standard errors: around 65% of applications default to the standard two-way clustering (Cameron, Gelbach, and Miller, 2011), 20% favor popular options in short panels such as clustering at the unit level (Liang and Zeger, 1986; Arellano, 1987) (or bootstrap alternatives, to a lesser extent), and around 15% opt for panel versions of conventional time-series tools (Driscoll and Kraay, 1998). Recent, representative examples are Crouzet and Mehrotra (2020), Ottonello and Winberry (2020) and Holm, Paul, and Tischbirek (2021), respectively.

Perhaps most importantly, these choices are often either not discussed at all or justified on the basis of alignment with common practice, citing previous similar empirical work. For instance, two-way clustering is usually simply reported as a way to account for general "autocorrelation within time and within units". We show that the former is crucial — and clarify why — and the latter is superfluous. Our results also demonstrate how information from the known autocovariance structure can be exploited and illustrate the potential pitfalls of off-the-shelf autocorrelation consistent methods such as Driscoll and Kraay (1998). Finally, we offer a way to reinterpret confidence intervals that ignore the macro nature of X_t, as those constructed by clustering at the unit level.

On top of this, in many cases the availability of a very large cross-sectional dimension is intuitively argued as a source of additional statistical precision, and so is the use of covariates s_i. In applications, these external variables are of intrinsic interest, rather than carefully constructed instruments.[3] We qualify these statements

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 relatively small cross-sectional dimensions (such as cross-country regressions), where entities are meaningful and a purely time-series treatment might be feasible. Nonetheless, when these units are pooled, as in Fukui, Nakamura, and Steinsson (2023), our results still apply.

[3]For instance, a booming literature studies the determinants of firm responsiveness to monetary policy, including differential responses by default risk (Ottonello and Winberry, 2020), firm size (Crouzet and Mehrotra, 2020) or predetermined measures of stock turnover (Jeenas and Lagos, forthcoming).

and provide conditions under which the latter is correct, in the sense of improved convergence rates.

Related literature. Our paper contributes to various strands of the econometrics literature, on top of the large empirical literature for which our results are relevant.

First, it relates to the time series literature on inference for local projections (Jordà, 2005; Stock and Watson, 2018; Montiel Olea and Plagborg-Møller, 2021; Lusompa, 2023; Xu, 2023). Jordà (2005) shows that the regression scores have MA(h) structure when the true model is a finite-order VAR. This also holds true in the empirically relevant case with observed shocks and general dynamics that we consider. Again under the finite-order VAR model, Montiel Olea and Plagborg-Møller (2021) show that lag augmented local projections induce serially uncorrelated scores and heteroskedasticity-robust inference suffices. Under this structure, we show a panel version of their results when local projections are augmented with unit-by-unit controls. In fact, the aggregate nature of X_t allows an intuitive, close parallel: heteroskedasticity-robust standard errors are valid too — when calculated on the synthetic (aggregated) time series of regression scores. Note that they require mean independent innovations, the same type of assumption we place on X_t.[4]

Second, we contribute to the literature on estimation and inference with aggregate shocks. Hahn, Kuersteiner, and Mazzocco (2020) bring attention to the drastic consequences of drawing inferences from short panels with aggregate uncertainty through several stylized economic models, and propose combining cross-sectional data with external time series data. Recent contributions have considered regional-exposure designs, which have a similar flavor to the representation in (1) for $h = 0$. These are situations where the regressor of interest varies over both dimensions of the panel but an interacted instrument of the form $s_i X_t$ (where s_i are region-specific exposures to changes in aggregate conditions) is a vailable. Arkhangelsky and Korovkin (2023) argue that in these setups the exogenous variation comes from time-series shocks X_t and focus on threats to instrument validity, whereas

[4]These assumptions are comparable, since in practice one could think of innovations as "observed" if the autoregressive structure is known (Montiel Olea and Plagborg-Møller, 2021, p. 1784).

Majerovitz and Sastry (2023) consider either s_i or X_t as sources of identification. They recognize the spatial autocorrelation induced by omitted aggregate shocks in the latter case and note that clustering at the time level is necessary. When X_t is i.i.d., they suggest the use and explore the validity of two-way clustering in simulations.[5] Our paper provides a framework and formal conditions under which this is the case. In fact, mean independence of X_t and clustering at the time level suffices for valid inference ($h = 0$), even in high noise environments.

Third, our paper relates to the literature on models with cross-sectional dependence, which often considers general setups where the scores feature varying degrees of spatial dependence (Driscoll and Kraay, 1998; Andrews, 2005; Pesaran, 2006; Gonçalves, 2011; Pakel, 2019). Our model falls in the polar case where the error term contains a source of aggregate variation and the regressor of interest only varies over time. This rules out solutions proposed in the literature based on partialling out the common component from the regressors, as in Pesaran (2006).

Outline. Section 2 provides a non-technical overview of our results in a simple model without dynamics, illustrating the role of aggregate shocks and their signal relative to micro shocks. Section 3 presents our formal results in the context of a general dynamic model and gives recommendations for empirical practice. Section 4 discusses the role of cross-sectional variation, and Section 5 presents a comprehensive set of simulations. Section 6 concludes. Proofs are relegated to Appendix C and the Supplemental Material (Almuzara and Sancibrián, 2024).

2 Overview of the results

We illustrate the main points of the paper in a simple, static regression model with homogeneous responses. We will keep the exposition simple and omit most technicalities, but all insights in this section extrapolate to the more general setup developed in Section 3.

[5] Their results that the size of the estimation error is of order $N^{-1/2}$ when X_t are i.i.d. shocks (and independent of unobservables) is driven by the restriction $N/T \to c$ where c is a finite contant.

We observe a micro outcome Y_{it} and a macro shock X_t for units $i = 1, \dots, N$ and over periods $t = 1, \dots, T$. 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 $\kappa \geq 0$ regulates their relative importance in the micro data, as explained below.

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 of these aggregate variables — captured by X_t — it would be hard to exante rule out the presence of any others — embedded in Z_t. This hardly escapable symmetry of the problem will be at the core of many results in what follows.

We now make two sets of assumptions related to the elements in (2), which are only slightly generalized to Assumptions 1 and 2 in Section 3.

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 here that 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.[6]

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

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

[6]Both of these assumptions could be relaxed; we discuss alternatives to S1(i) in Sections 3 and 5. Allowing for weak spatial dependence into u_{it} in place of S1(ii) is also possible with minor modifications.

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

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

Assumption S2 implies that X_t, Z_t and u_{it} are mutually unpredictable and serially uncorrelated. Assumption S2(i) is ultimately an identification assumption, and S2(ii) and S2(iii) are symmetric assumptions on unobservables.[7] Indeed, mutual unpredictability of macro shocks lies at the core of macroeconometrics and is typically necessary in order to give structural interpretation to impulse-response calculations (see, for instance, Ramey, 2016; Stock and Watson, 2016; Plagborg-Møller and Wolf, 2021).[8] 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. For example, a popular approach is to try to isolate the surprise component of monetary policy by measuring the change in asset prices in a tight window around policy announcements; see Ramey (2016, Section 2.3) for a review of this and many other identification methods.[9]

Remark 1. (Relaxing Assumption S2(i).) In practice, we might only observe a proxy shock that is contaminated with measurement error, but an instrument that satisfies an analogous condition to Assumption S2(i) and is correlated with X_t is available. Our results extend easily to the local projections instrumental variable (LP-IV) case (Stock and Watson, 2018, Section 1.3), as discussed in Section 3.4. In other instances,

[7] Since Z_t and u_{it} are unobserved, orthogonality between them is not strictly necessary, but is invoked to simplify the exposition and isolate "micro" and "macro" error terms.

[8] 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 similar context for inference on local projections.

This still permits dynamics on the second- or higher-order moments given the paths of other shocks. We permit that, for example, monetary, fiscal or oil supply shocks increase the variance of, say, household-level income via higher order dynamics in u_{it}.

[9] Examples of relevant applications include Crouzet and Mehrotra (2020, monetary policy shocks identified via narrative approaches), Känzig (2021, oil supply shocks via high-frequency identification) and Drechsel (2023, investment shocks identified via Cholesky/structural VAR restrictions), among many others.

a mismeasured shock $\widetilde{X}_t$ might display some residual autocorrelation structure, say

$$\widetilde{X}_t = \tilde{B}_1 \widetilde{X}_{t-1} + \cdots + \tilde{B}_p \widetilde{X}_{t-p} + X_t,$$

then including $(\widetilde{X}_{t-1}, \ldots, \widetilde{X}_{t-p})$ as controls will lead to analogous results.[10]

Remark 2. (An empirical illustration.) To fix ideas, consider Holm et al. (2021), who use a large administrative dataset to study the transmission of monetary policy in Norway. There, Y_{it} denotes income or consumption measured at the household level while X_t are monetary policy shocks to the key policy rate of Norges Bank, identified as in Romer and Romer (2004).

The period of analysis extends from 1996 to 2015, a time in which other aggregate shocks have likely occurred and contributed to variation in household income and spending. For example, the price of oil — Norway's main export — experienced large fluctuations over the same period, likely affecting economic activity through exports and exchange rate channels. In addition, fiscal stimulus measures were enacted in 2009 in response to the Global Financial Crisis. In model (2), they are captured by Z_t. Assumption S2(i) requires the monetary policy shock series to be orthogonal to all these other macro shocks.

On the signal value of aggregate variation. The setup in (2) describes individual outcomes as shaped by both idiosyncratic circumstances and changes in economy-wide conditions. Since interest is in responses to the latter, this bears the question of their relative importance, the signal-to-noise ratio of macro shocks.

We allow here for low signal-to-noise environments. This intends to capture settings with disaggregated data where idiosyncratic noise becomes more and more prevalent as these macro shocks trickle down through the economy, and become small relative to the wide range of idiosyncratic shocks that dictate the fate of individuals on a daily basis — such as health shocks, job losses or lucky streaks. Consider the in-sample average outcome, $\bar{Y}_t = N^{-1} \sum_{i=1}^{N} Y_{it}$. It is reasonable to expect a sizeable role for aggregates relative to micro shocks in explaining changes

[10] The situation where one is interested in the effects of persistent "shocks" themselves is analyzed by Alloza, Gonzalo, and Sanz (2023).

in $\bar{Y}_t$ since the latter tend to get averaged out; the opposite might be the case for Y_{it} where idiosyncratic noise might be pervasive. We now formalize this idea.

Back to (2), let $\text{Var}(X_t) = \text{Var}(Z_t) = \text{Var}(u_{it}) = 1$ and $\beta_0 = 1$ for the sake of illustration. Then, under assumptions S1 and S2, the proportion of the variance of $\bar{Y}_t$ explained by $\{X_t, Z_t\}$ is given by

$$\bar{R}^2_\kappa = 1 - \frac{\text{Var}_\kappa\left(\bar{Y}_t \middle| X_t, Z_t\right)}{\text{Var}_\kappa\left(\bar{Y}_t\right)} = \frac{1}{1 + \kappa^2/(2N)}, \tag{3}$$

so that κ regulates the signal value of aggregate variation. Intuitively, if κ is large, then micro shocks are ubiquitous at the unit level (low-signal environment) and do not get fully washed away; if κ is small averaging essentially gets rid of all micro-level variation.[11]

We consider a range of data generating processes where κ is such that $0 \leq \kappa \leq \bar{\kappa}\sqrt{N}$, for $N \to \infty$ and a positive, finite constant $\bar{\kappa}$. This allows for (asymptotically) non-trivial signal-to-noise: $R^2_\kappa \to 1/(1 + \bar{\kappa}^2/2)$, which is strictly between zero and one. Of course, indexing κ to the sample size should not be taken literally — it is simply a device to ensure that our approximations to the sampling distribution of $\hat{\beta}$ are able to capture the essence of low signal-to-noise regimes.[12] That is, if κ is constant, then $R^2_{\kappa,i} \to 1$ as $N \to \infty$, and approximations based on this result would attribute all randomness to aggregate shocks.

The practical usefulness of this device will become clear below. Essentially, under standard asymptotics where $N, T \to \infty$, micro shocks become negligible in the sampling distribution of $\hat{\beta}$. This has strong implications for uncertainty quantification, and is at odds with folk wisdom in our empirical applications —

[11] One can also consider the unit-level counterpart to (3):

$$R^2_{\kappa,i} = 1 - \frac{\text{Var}_\kappa\left(Y_{it} \middle| X_t, Z_t\right)}{\text{Var}_\kappa\left(Y_{it}\right)} = \frac{1}{1 + \kappa^2/2},$$

which is small if κ is large, corresponding to a low signal regime in which the macro data have weak explanatory power over Y_{it}.

[12] This type of embeddings are relatively common in the econometric literature and often referred to as asymptotics with local-to-zero parameters. An example which also has a low-signal flavor to it is the weak instrumental variables literature (Staiger and Stock, 1997).

as suggested by the outsized role that accounting for unit-level dynamics plays in many of these. Note that we do not need to take a stand on κ: we simply allow for scenarios with low, moderate and high signal-to-noise and seek robustness of our inferential procedures over these.[13,14] In Section 3, we will show that the concept of uniformity precisely captures this idea.

Estimation and inference. A natural estimator for the effect β_0 is the pooled least squares estimator,

$$\hat{\beta} = \frac{1}{T}\sum_{t=1}^{T} \omega_t \left(\frac{1}{N}\sum_{i=1}^{N} Y_{it} \right), \quad \omega_t = \frac{X_t}{T^{-1}\sum_{\tau=1}^{T} X_\tau^2}, \tag{4}$$

which is the default choice for the overwhelming majority of empirical applications; we incorporate controls — including unit-level fixed effects — in Section 3. In Section 3 we also introduce unobservable heterogeneity and characterize $\hat{\beta}$ explicitly as averages over unit-level responses.

It is useful to write $\hat{\beta}$ as

$$\hat{\beta} = \beta_0 + \frac{1}{T}\sum_{t=1}^{T} \omega_t Z_t + \frac{\kappa}{\sqrt{N}}\frac{1}{T}\sum_{t=1}^{T} \omega_t \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} u_{it} \right), \tag{5}$$

where the second term is a purely time-series component whereas the third term involves both dimensions of the panel. This micro-macro decomposition holds in the more general case studied in Section 3, and has major implications for inference.

Under Assumptions S1 and S2 and taking limits as $N, T \to \infty$ (and under some regularity conditions to be specified in Section 3), it is easy to show that $\hat{\beta}$ is

[13] Environments with relatively high signal value of macro shocks seem likely too; consider for instance bank-level credit data, which is likely highly responsive to aggregate macroeconomic and financial dynamics.

[14] Results are invariant to rescaling the macro component of the model in (2) by κ^{-1} instead; what matters is the relative signal of macro and micro shocks.

consistent for β_0. Additionally, we can bound the estimation error as

$$\hat{\beta} - \beta_0 = \underbrace{\frac{1}{T}\sum_{t=1}^{T} \omega_t Z_t}_{=O_p\left(T^{-1/2}\right)} + \underbrace{\frac{\kappa}{\sqrt{N}}\frac{1}{T}\sum_{t=1}^{T} \omega_t \left(\frac{1}{N}\sum_{i=1}^{N} u_{it}\right)}_{=O_p\left(\kappa(NT)^{-1/2}\right)}, \tag{6}$$

which shows that cross-sectional averaging does not contribute to averaging out macro shocks Z_t at all, while it helps partially dominate the contribution of micro shocks u_{it} to the estimation error. Indeed, in low-noise environments the latter is of standard stochastic order $(NT)^{-1/2}$; in pervasive noise environments it is of similar size as the macro variation.

In any case, it follows that the precision at which we can recover β_0 depends crucially on the amount of time-series variation in the data, and is of order $T^{-1/2}$ even in situations where $N \gg T$. An intuitive way to see this is to notice that $\hat{\beta}$ is numerically equal to the least-squares estimator using the synthetic time series $\bar{Y}_t = N^{-1}\sum_{i=1}^{N} Y_{it}$ and X_t.

Remark 3. (Dynamics and local projections.) The least squares estimator in (4) can be interpreted as a local projection estimator for $h = 0$, and these results carry over to horizons $h > 0$. As such, $\sqrt{T}$ convergence rates are to be expected in most empirical applications with macro shocks. In fact, once we allow for dynamic effects of X_t over time, these results arise organically: letting $Z_t = X_{t-1}$ in (2) does not change any of the reasoning so far.[15]

In this simple model, it follows that

$$\sqrt{T}\left(\hat{\beta} - \beta_0\right) \xrightarrow{\text{d}} N(0, V_0), \quad V_0 = \frac{\mathbb{E}\left[X_t^2 Z_t^2\right] + \left(\kappa^2/N\right)\mathbb{E}\left[X_t^2 u_{it}^2\right]}{\left(\mathbb{E}\left[X_t^2\right]\right)^2},$$

and the asymptotic variance V_0 depends on a common component involving observed and unobserved macro shocks that do not vanish as $N \to \infty$. When these shocks have a low explanatory power for individual outcomes, V_0 also incorporates

[15] Note however that under Assumption S2, β is also unbiased for β_0, but this is no longer the case with dynamics in X_t, which induce a (standard) bias of order T^{-1}. Contrary to the standard panel data case, the bias is here of smaller order than the standard errors and will not distort inference procedures.

an additional term involving micro shocks. Note that κ induces a discontinuity in the asymptotic distribution in the sense that the asymptotic variance of $\hat{\beta}$ may or may not include the micro component. Not allowing for low-signal regimes might greatly distort the quality of these approximations in finite samples.

For the construction of confidence intervals, we define the regression residual as $\hat{v}_{it} = Y_{it} - \hat{\beta} X_t$ and its cross-sectional average as $\hat{v}_t = N^{-1} \sum_{i=1}^{N} \hat{v}_{it}$. In Section 3, we will show that the following is a consistent estimator of V_0:

$$\hat{V}_0 = \frac{1}{T} \sum_{t=1}^{T} \omega_t^2 \hat{v}_t^2. \tag{7}$$

Two remarks are in order. First, this holds for any κ such that $0 \leq \kappa \leq \bar{\kappa} \sqrt{N}$, which will be behind the uniformity result in Section 3; second, valid inference simply requires clustering at the time level to account for the common aggregate component. As such, local projections inference with macro shocks is both robust to high noise regimes and very simple to implement. In fact, it amounts to computing the heteroskedasticity-robust variance formula on the synthetic time series $\{X_t, \hat{v}_t\}_{t=1}^{T}$.

One might wonder how much the static nature of (2) limits these results. The rest of the paper shows that one can accommodate a much richer framework with very little change.

Remark 4. (Inference conditional on aggregate shocks.) Ignoring the macro component — as in existing empirical work that exploits independence across units — is equivalent to conditioning on the realizations of aggregate shocks. Appendix A illustrates that this implies targeting a different object than β_0; one that is defined by conditional orthogonality restrictions. In general, this induces an internal/external validity trade-off whereby practitioners might be able to pin down responses very precisely and with fewer assumptions on the nature of aggregate variation — such as ergodicity — but these parameters 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 permanent unobserved heterogeneity of individual impulse-response functions (IRFs).

We introduce the setup in 3.1 and estimators and estimands in Section 3.2. We state the main results on inference in Section 3.3. In addition, we specialize our results to the relevant case of finite-order vector autoregressive (VAR) dynamics and provide recommendations for empirical practice. Extensions to local projections with instrumental variables (LP-IV) are discussed in Section 3.4. We postpone a discussion of the role of cross-sectional characteristics to Section 4.

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 vector case. We assume

$$Y_{it} = \mu_i + \beta_i(L)X_t + v_{it}, \tag{8}$$

$$v_{it} = \gamma_i(L)Z_t + \kappa\delta_i(L)u_{it}, \tag{9}$$

where $\beta_i(L) = \sum_{\ell=0}^{\infty} \beta_{i\ell}L^\ell$, $\gamma_i(L) = \sum_{\ell=0}^{\infty} \gamma_{i\ell}L^\ell$ and $\delta_i(L) = \sum_{\ell=0}^{\infty} \delta_{i\ell}L^\ell$ are polynomials in the lag operator L, and Z_t and u_{it} are unobserved serially uncorrelated aggregate and idiosyncratic errors. We adopt the notation $\beta_i = \{\beta_{i\ell}\}_{\ell=0}^{\infty}$, $\gamma_i = \{\gamma_{i\ell}\}_{\ell=0}^{\infty}$, $\delta_i = \{\delta_{i\ell}\}_{\ell=0}^{\infty}$ and $\theta_i = \{\mu_i, \beta_i, \gamma_i, \delta_i\}$. We later specify regularity conditions on these coefficients so that $\beta_i(L)X_t$, $\gamma_i(L)Z_t$ and $\delta_I(L)u_{it}$ are well defined with probability one. In this setup, θ_i traces out cross-sectionally heterogeneous IRFs to aggregate and idiosyncratic shocks. The availability of external variables s_i might help study their transmission along unit-level observable covariates.

As explained in section 2, we consider a range of data generating processes for which $0 \leq \kappa \leq \bar{\kappa}\sqrt{N}$ 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) $\{\theta_i, s_i, \{u_{it}\}_{t=-\infty}^{\infty}\}_{i=1}^N$ *are i.i.d. over i conditional on* $\{X_t, Z_t\}_{t=-\infty}^{\infty}$.

Assumption 2 (Shocks and mean independence).

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

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

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

Assumptions 1 and 2 are slight generalizations of S1 and S2, respectively, to accommodate the presence of unobserved heterogeneity and external covariates. Assumption 2 requires them to be strictly exogenous with respect to shocks.[16] Fixed-effect endogeneity, however, is permitted in the sense that the distribution of $\{\theta_i\}_{i=1}^N$ conditional on $\{X_t\}_{t=-\infty}^{\infty}$ is not restricted, as in pure fixed effects approaches. For a discussion of all other components, we refer the reader to Section 2. Again, the key assumption is 2(i) on the availability of an observed macro shock satisfying certain orthogonality requirements. We discuss alternatives to this assumption in the form of mismeasurement with an instrument in Section 3.4.

Remark 5. (The role of κ, revisited.) As in Section 2, κ controls the signal-to-noise of aggregate shocks in the micro data. To see this, let

$$\bar{Y}_t = \frac{1}{N}\sum_{i=1}^{N} Y_{it} = \bar{\mu} + \bar{\beta}(L)X_t + \bar{\gamma}(L)Z_t + \frac{\kappa}{\sqrt{N}}\tilde{U}_t,$$

where $\bar{\beta}(L) = N^{-1}\sum_{i=1}^N \beta_i(L)$ (similarly for $\bar{\gamma}(L)$ and $\bar{\mu}$) and $\tilde{U}_t = N^{-1/2}\sum_{i=1}^N \delta_i(L)u_{it}$. Then we can again compute the macro-level proportion of the variance of $\bar{Y}_t$ explained by by $\{X_\tau, Z_\tau\}$, which leads to a representation that is analogous to (3):

$$\bar{R}^2 = 1 - \frac{\mathrm{Var}_\kappa\left(\bar{Y}_t \middle| \{\theta_i\}_{i=1}^N, \{X_\tau, Z_\tau\}_{\tau=-\infty}^{\infty}\right)}{\mathrm{Var}_\kappa\left(\bar{Y}_t \middle| \{\theta_i\}_{i=1}^N\right)}$$

[16] This is essentially an identification condition for small T, see Arellano and Bonhomme (2012, Section 2.2) in the context of random coefficient models in short panels.

$$= 1 - \frac{(\kappa^2/N)\text{Var}_\kappa\left(\tilde{U}_t \middle| \{\theta_i\}_{i=1}^N, \{X_\tau, Z_\tau\}_{\tau=-\infty}^{\infty}\right)}{\text{Var}_\kappa\left(\bar{\beta}(L)X_t + \bar{\gamma}(L)Z_t \middle| \{\theta_i\}_{i=1}^N\right) + (\kappa^2/N)\text{Var}_\kappa\left(\tilde{U}_t \middle| \{\theta_i\}_{i=1}^N\right)} = 1 - O\left(\frac{\kappa^2}{N}\right).$$

and similarly at the unit level, where $R_i^2 = 1 - O\left(\kappa^2\right)$. When κ is proportional to $\sqrt{N}$, $\bar{R}^2$ remains asymptotically non-trivial (below one) and $R_i^2 \to 0$ as $N \to \infty$.

3.2 Estimators and estimands

We consider panel local projection (LP) estimators that target different features of the IRF of Y_{it} with respect to X_t, for a fixed horizon $h \geq 0$.

Popular choices in applied work include either X_t or $s_i X_t$ as regressors of interest ("pooled" and "interacted" specifications) and a battery of controls, such as unit or time fixed effects and lags of the outcome variable and shocks. Let W_{it} denote a vector of controls, to be specified below. The coefficient of the local projection on $s_i X_t$ including W_{it} is

$$\hat{\beta}_h = \frac{\sum_{i=1}^N \sum_{t=1}^{T-h} (s_i X_t - \hat{\Pi}' W_{it}) Y_{i,t+h}}{\sum_{i=1}^N \sum_{t=1}^{T-h} (s_i X_t - \hat{\Pi}' W_{it})^2}, \tag{10}$$

where $\hat{\Pi} = \left(\sum_{i,t} W_{it} W_{it}'\right)^{-1} \sum_{i,t} W_{it} s_i X_t$. For simplicity of exposition, we focus on the interacted local projection estimator in which W_{it} includes unit and time fixed effects, treating pooled local projections as a particular case where $s_i = 1$ and W_{it} only includes unit fixed effects. We discuss inclusion of additional controls below.[17] The estimator then has the representation

$$\hat{\beta}_h = \frac{1}{(T-h)} \sum_{t=1}^{T-h} \omega_t \left(\frac{1}{N} \sum_{i=1}^N \pi_i Y_{i,t+h} \right), \tag{11}$$

[17] At this time, we simply stress that under assumptions 1 and 2 the role of controls is reducing estimation noise, and thus all results in this section hold with a more general vector of controls.

with weights $N\pi_i = (s_i - \bar{s}) / \sum_{j=1}^{N}(s_j - \bar{s})^2$ and $(T-h)\omega_t = (X_t - \bar{X}) / \sum_{\tau=1}^{T-h}(X_\tau - \bar{X})^2$ where $\bar{s} = N^{-1}\sum_{i=1}^{N} s_i$ and $\bar{X} = (T-h)^{-1}\sum_{t=1}^{T-h} X_t$.[18] Pooled local projections can be recovered by setting $\pi_i = 1$.

Substituting (8) into (11) and using compact notation,

$$\hat{\beta}_h = \bar{\beta}_h + \frac{1}{T-h}\sum_{t=1}^{T-h} \omega_t \xi_{ht}, \tag{12}$$

where $\bar{\beta}_h = N^{-1}\sum_{i=1}^{N} \pi_i \beta_{ih}$ and $\xi_{ht} = \tilde{X}_{(-h),t+h} + \tilde{Z}_{t+h} + \frac{\kappa}{\sqrt{N}}\tilde{U}_{t+h}$ with

$$\begin{aligned}
\tilde{X}_{(-h),t+h} &= \sum_{\ell=0}^{\infty} \mathbb{1}\{\ell \neq h\}\left(\frac{1}{N}\sum_{i=1}^{N}\pi_i\beta_{i\ell}\right) X_{t+h-\ell} = \sum_{\ell=0}^{\infty}\mathbb{1}\{\ell\neq h\}\bar{\beta}_\ell X_{t+h-\ell}, \\
\tilde{Z}_{t+h} &= \sum_{\ell=0}^{\infty}\left(\frac{1}{N}\sum_{i=1}^{N}\pi_i\gamma_{i\ell}\right) Z_{t+h-\ell} = \sum_{\ell=0}^{\infty}\bar{\gamma}_\ell Z_{t+h-\ell}, \\
\tilde{U}_{t+h} &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{\ell=0}^{\infty}\pi_i\delta_{i\ell}u_{i,t+h-\ell},
\end{aligned} \tag{13}$$

with obvious definitions. Both the estimator and the estimation error are intuitive generalizations of their simple example counterparts in Section 2. In particular, $\hat{\beta}_h$ is numerically equal to the result of a time-series regression involving the aggregate (weighted) outcome $N^{-1}\sum_{i=1}^{N}\pi_i Y_{i,t+h}$, and the first two grand terms in ξ_{ht} contain common shocks to all units that can only be dominated with time series variation (unless special conditions hold).

Estimands. It follows that interacted and pooled local projections recover

$$\beta_h = \frac{\mathrm{Cov}(s_i, \beta_{ih})}{\mathrm{Var}\,(s_i)} \tag{14}$$

[18] This follows since under two-way fixed effects, we can write

$$s_i X_t - \hat{\Pi}' W_{it} = s_i X_t - \bar{s}X_t - s_i\bar{X} + \bar{s}\bar{X} = (s_i - \bar{s})(X_t - \bar{X}).$$

Interestingly, the weighted average $N^{-1}\sum_{i=1}^{N}\pi_i Y_{i,t+h} = \sum_{i=1}^{N}(s_i-\bar{s})Y_{i,t+h} / \sum_{i=1}^{N}(s_i-\bar{s})^2$ is the coefficient on s_i of a cross-sectional regression of $Y_{i,t+h}$ on s_i including an intercept.

and $\beta_h = \mathbb{E}[\beta_{ih}]$, respectively. This clarifies the sense in which pooled and interacted local projection specifications can be interpreted. The key ingredient is the presence of heterogeneous responses to shocks; this allows us to formalize s_i as a device to explore differential responsiveness to shocks along an observable gradient.

In general, both pooled and interacted local projections are needed for a meaningful description of heterogeneity. For example, if interest lies in the best linear approximation to β_{ih}, this involves both interacted and pooled LP estimands,[19]

$$\mathbb{E}^*[\beta_{ih}|s_i] = \mathbb{E}[\beta_{ih}] + \frac{\text{Cov}(s_i, \beta_{ih})}{\text{Var}(s_i)}(s_i - \mathbb{E}[s_i]).$$

3.3 Inference

We now turn to the main results of the paper. We first characterize the properties of the estimation error in (12) under model (8)–(9). We next study inference under our general setup (Proposition 2) and under a special assumption of autoregressive dynamics (Corollary 1). Based on these, we derive recommendations for empirical work that we verify in the simulation study of Section 5.

Let $\mathbb{P}_\kappa$ denote probabilities under a data generating process for a given value of κ; we omit κ from probabilities that do not depend on it (such as those that only involve X_t).

Proposition 1. *Under assumptions 1 and 2, the autocovariance function of the score $X_t\xi_{ht}$ in (12) is given by*

$$\begin{aligned}\text{Var}_\kappa(X_t\xi_{ht}) = \mathbb{E}\left[\sum_{\ell=0}^{\infty} \mathbb{1}\{\ell \neq h\} X_t^2 X_{t+h-\ell}^2 \bar{\beta}_\ell^2\right] + \mathbb{E}\left[\sum_{\ell=0}^{\infty} X_t^2 Z_{t+h-\ell}^2 \bar{\gamma}_\ell^2\right] \\ + \frac{\kappa^2}{N}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=0}^{\infty} X_t^2 \pi_i^2 \delta_{i,\ell}^2 u_{i,t+h-\ell}^2\right]\end{aligned} \tag{15}$$

[19] Recall that the inclusion of time effects absorbs the mean IRF. Importantly, omitting either X_t or time dummies when running interacted local projections has similar implications as forcing the regression of β_{ih} on s_i through the origin.

and, for $1 \leq |j| \leq h$,

$$\mathrm{Cov}_\kappa\Big(X_t\xi_{ht}, X_{t-j}\xi_{h,t-j}\Big) = \mathbb{E}\left[X_t X_{t-j}\tilde{X}_{(-h),t+h}\tilde{X}_{(-h),t+h-j}\right] = \mathbb{E}\left[X_t^2 X_{t-j}^2\bar{\beta}_{h+j}\bar{\beta}_{h-j}\right]. \tag{16}$$

For $|j| > h$, *we have* $\mathrm{Cov}_\kappa\Big(X_t\xi_{ht}, X_{t-j}\xi_{h,t-j}\Big) = 0$.

Proof. It follows from repeated application of 2(i) and iterated expectations.[20] □

Put another way, $X_t\xi_{ht}$ has the serial dependence structure of an MA(h) process. Moreover, the orthogonality properties of X_t imply that these dynamics are only induced by the presence of leftover leads between $t+1$ and $t+h$, independently of other unobserved aggregate shocks or micro noise.

Proposition 1 suggests the need for standard errors robust to heteroskedasticity and autocorrelation up to order h. Consider the unit-level residuals

$$\hat{v}_{h,it} = \widetilde{Y}_{i,t+h} - \hat{\beta}_h(s_i - \bar{s})(X_t - \bar{X}),$$

where $\widetilde{Y}_{i,t+h} = Y_{i,t+h} - \bar{Y}_i - \bar{Y}_{t+h} + \bar{Y}$ with $\bar{Y}_i = (T-h)^{-1}\sum_{t=1}^{T-h} Y_{i,t+h}$, $\bar{Y}_{t+h} = N^{-1}\sum_{i=1}^{N} Y_{i,t+h}$ and $\bar{Y} = N^{-1}(T-h)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T-h} Y_{i,t+h}$. Similar to Section 2, we first average over the cross-section,

$$\hat{\xi}_{ht} = \frac{1}{N}\sum_{i=1}^{N}\pi_i\hat{v}_{h,it} = \frac{1}{N}\sum_{i=1}^{N}\pi_i\widetilde{Y}_{i,t+h} - \hat{\beta}_h(X_t - \bar{X}), \tag{17}$$

which yields a synthetic time series of projection coefficients of unit-level residuals on (demeaned) s_i. Pooled local projections with unit fixed effects are recovered by setting $\pi_i = 1$ and $\widetilde{Y}_{i,t+h} = Y_{i,t+h} - \bar{Y}_i$. Define[21]

$$\hat{V}_j = \frac{1}{T-h}\sum_{t=j+1}^{T-h}(X_t - \bar{X})(X_{t-j} - \bar{X})\hat{\xi}_{ht}\hat{\xi}_{h,t-j},$$

[20] To establish the second equality in (16), note the typical element in $X_t X_{t-j}\tilde{X}_{(-h),t+h}\tilde{X}_{(-h),t+h-j}$ contains a cross-product involving X_t at potentially four different periods, among which at most only two might coincide. The only nonzero term is thus that where the four indices form two pairs.

[21] Note that (18) is not guaranteed to be positive semidefinite, but it is straightforward to introduce a kernel that ensures so (as in Newey and West, 1987).

$$\hat{S}_h = \hat{V}_0 + 2\sum_{j=1}^{h} \hat{V}_j, \tag{18}$$

and

$$\hat{\sigma}_h = \frac{\sqrt{(T-h)\hat{S}_h}}{\sum_{t=1}^{T-h}(X_t - \bar{X})^2}. \tag{19}$$

Additionally, if Φ is the standard normal c.d.f., we can form a $(1-\alpha)$-CI for β_h as

$$\mathrm{CI}_\alpha = \left[\hat{\beta}_h \pm \hat{\sigma}_h \Phi^{-1}(\alpha/2)\right].$$

We derive asymptotic approximations to the sampling distribution of $\hat{\beta}_h$ and to the coverage probability of CI_α as $T-h \to \infty$ where $N = N_T \to \infty$ with $(T-h)/N_T \to 0$ and $\kappa = \kappa_T \to \infty$.[22] Regularity conditions are specified in assumptions 4 and 5 in Appendix B, and essentially require absolute summability of the lag polynomial coefficients in (8)–(9) and bounds on higher-order moments of shocks. The main result is that CI_α is asymptotically similar uniformly over $0 \le \kappa \le \bar{\kappa}\sqrt{N}$:

$$1-\alpha = \liminf_{T\to\infty}\left\{\inf_{0\le\kappa\le\bar{\kappa}\sqrt{N}} \mathbb{P}_\kappa\left[\beta_h \in \mathrm{CI}_\alpha\right]\right\} = \limsup_{T\to\infty}\left\{\sup_{0\le\kappa\le\bar{\kappa}\sqrt{N}} \mathbb{P}_\kappa\left[\beta_h \in \mathrm{CI}_\alpha\right]\right\}.$$

This is a corollary of the following:

Proposition 2. *Under assumptions 1, 2, 4 and 5, for $(T-h)/N \to 0$ and for all $\eta \in \mathbb{R}$,*

$$\lim_{T\to\infty}\left[\sup_{0\le\kappa\le\bar{\kappa}\sqrt{N}}\left|\mathbb{P}_\kappa\left[\frac{\hat{\beta}_h - \beta_h}{\hat{\sigma}_h} \le \eta\right] - \Phi(\eta)\right|\right] = 0.$$

Proof. See Appendix C. □

Proposition 2 states that approximations are valid uniformly over data generating processes with different degrees of signal-to-noise of macro shocks. Intuitively,

[22] That N grows faster than $T-h$ is a mild requirement given the state of empirical practice. The derivation is analogous in square panels, in which case some terms of order $N^{-1/2}$ might not be asymptotically negligible. In practice, applications where N is proportional to the effective time-series sample size, such as cross-country regressions, might entail an additional clustering step at the unit level.

even though the asymptotic distribution of the (rescaled) estimation error has a discontinuity around κ (see for instance (15)), this does not affect the calculation of confidence intervals. The t-statistic can then be shown to be asymptotically normal uniformly over κ.

VAR dynamics. In practice, it would seem natural to include lags of the outcome variable and regressor as controls in estimation in order to reduce estimation noise. It turns out that, in some important cases, this further simplifies inference.

We study this formally by imposing additional structure on (8)–(9) so that it conforms with a finite-order joint vector autoregressive model (VAR) on $(Y_{it}, X_t)'$. Under the shock assumption on X_t, we let the outcome be

$$Y_{it} = m_i + \sum_{\ell=1}^{p} A_{i\ell} Y_{i,t-\ell} + \sum_{\ell=0}^{p} B_{i\ell} X_{t-\ell} + C_{i0} Z_t + \kappa D_{i0} u_{it}, \tag{20}$$

which is a special case of (8)–(9) provided one can invert the lag polynomial $A_i(L) = 1 - \sum_{\ell=1}^{p} A_{i\ell} L^\ell$. If so, the mapping is $\mu_i = m_i / A_i(1)$ and (with $B_i(L) = \sum_{\ell=0}^{p} B_{i\ell} L^\ell$)

$$\begin{aligned}
\beta_i(L) &= (A_i(L))^{-1} B_i(L), \\
\gamma_i(L) &= (A_i(L))^{-1} C_{i0}, \\
\delta_i(L) &= (A_i(L))^{-1} D_{i0}.
\end{aligned}$$

In other words, the heterogeneous VAR model requires that the responses to unobserved macro and micro shocks be proportional to each other.

We now consider local projections augmented with p lags of Y_{it} and $s_i X_t$. That is, we will include as controls unit and time fixed effects (as in (10)) together with $Y_{i,t-1}, \ldots, Y_{i,t-p}, s_i X_{t-1}, \ldots, s_i X_{t-p}$. Importantly, the coefficients on lags of Y_{it} are unit-specific. To express the estimator, we use the Frisch-Waugh-Lovell logic. Define

$$\mathbf{Y}_{it} = \begin{pmatrix} \widetilde{Y}_{i,t-1} \\ \vdots \\ \widetilde{Y}_{i,t-p} \end{pmatrix}, \quad \mathbf{X}_t = \begin{pmatrix} X_{t-1} - \bar{X} \\ \vdots \\ X_{t-p} - \bar{X} \end{pmatrix},$$

where $\widetilde{Y}_{it} = Y_{it} - N^{-1}\sum_{i=1}^{N} Y_{it} - (T-h)^{-1}\sum_{t=1}^{T-h} Y_{it} + N^{-1}(T-h)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T-h} Y_{it}$ is the residual from a two-way regression of Y_{it}. The estimator becomes

$$\begin{aligned}\hat{\beta}_h^{(p)} &= \frac{\sum_{t=1}^{T-h}\sum_{i=1}^{N}(X_t - \bar{X})(s_i - \bar{s})\left[Y_{i,t+h} - \hat{\Pi}_{YY,i}\mathbf{Y}_{it} - \hat{\Pi}_{YX}\mathbf{X}_t(s_i - \bar{s})\right]}{\sum_{i=1}^{N}\sum_{t=1}^{T-h}(X_t - \bar{X})(s_i - \bar{s})\left[X_t s_i - \hat{\Pi}_{XY,i}\mathbf{Y}_{it} - \hat{\Pi}_{XX}\mathbf{X}_t(s_i - \bar{s})\right]} \\ &= \bar{\beta}_h + \frac{1}{T-h}\sum_{t=1}^{T-h}\omega_t\xi_{ht} + o_p\left(T^{-1/2}\right),\end{aligned} \tag{21}$$

where $\hat{\Pi}_{YY,i}, \hat{\Pi}_{YX}, \hat{\Pi}_{XY,i}, \hat{\Pi}_{XX}$ are the corresponding partial regression coefficients, $\bar{\beta}_h = N^{-1}\sum_{i=1}^{N}\pi_i\beta_{ih}$ and $\xi_{ht} = \tilde{X}_{(-h),t+h} + \tilde{Z}_{t+h} + \frac{\kappa}{\sqrt{N}}\tilde{U}_{t+h}$ with

$$\begin{aligned}\tilde{X}_{(-h),t+h} &= \sum_{\ell=0}^{h-1}\left(\frac{1}{N}\sum_{i=1}^{N}\pi_i\beta_{i\ell}\right)X_{t+h-\ell} = \sum_{\ell=0}^{h-1}\bar{\beta}_\ell X_{t+h-\ell}, \\ \tilde{Z}_{t+h} &= \sum_{\ell=0}^{h-1}\left(\frac{1}{N}\sum_{i=1}^{N}\pi_i\gamma_{i\ell}\right)Z_{t+h-\ell} = \sum_{\ell=0}^{h-1}\bar{\gamma}_\ell Z_{t+h-\ell}, \\ \tilde{U}_{t+h} &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{\ell=0}^{h-1}\pi_i\delta_{i\ell}u_{i,t+h-\ell}.\end{aligned}$$

Corollary 1 below shows that under the VAR model in (20), the panel regression scores implied by (21) are serially uncorrelated, a panel version of the result in Montiel Olea and Plagborg-Møller (2021).[23]

Corollary 1. *Under assumptions 1 and 2, the autocovariance function of the score $X_t\xi_{ht}$ in (21) is given by*

$$\mathrm{Var}_\kappa\left(X_t\xi_{ht}\right) = \mathbb{E}\left[\sum_{\ell=0}^{h-1}X_t^2X_{t+h-\ell}^2\bar{\beta}_\ell^2\right] + \mathbb{E}\left[\sum_{\ell=0}^{h-1}X_t^2Z_{t+h-\ell}^2\bar{\gamma}_\ell^2\right] + \frac{\kappa^2}{N}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=0}^{h-1}X_t^2\pi_i^2\delta_{i,\ell}^2u_{i,t+h-\ell}^2\right],$$

and, for $|j| > 0$, $\mathrm{Cov}_\kappa\left(X_t\xi_{ht}, X_{t-j}\xi_{h,t-j}\right) = 0$.

[23] Strictly speaking, this result holds with p — rather than $p+1$ lags. The lag-augmentation step in Montiel Olea and Plagborg-Møller (2021) is needed when X_t is unobserved. See the remark below. Still, we would often refer to (21) as a lag-augmented local projection.

This leads to an even simpler inference recipe. The average residuals in (17) are now given by

$$\hat{\xi}_{ht} = \frac{1}{N}\sum_{i=1}^{N} \pi_i \left(\widetilde{Y}_{i,t+h} - \hat{\Pi}_{YY,i}\mathbf{Y}_{it} - \hat{\Pi}_{YX}\mathbf{X}_t(s_i - \bar{s}) \right) - \hat{\beta}_h(X_t - \bar{X}).$$

Define

$$\hat{\sigma}_h^{(p)} = \frac{1}{T-h}\left(\sum_{t=1}^{T-h} \omega_t^2 \hat{\xi}_{ht}^2 \right)^{1/2}, \tag{22}$$

which are the HR standard errors corresponding to a simple linear regression of the synthetic outcome time-series $N^{-1}\sum_{i=1}^{N} \pi_i \left(\widetilde{Y}_{i,t+h} - \hat{\Pi}_{YY,i}\mathbf{Y}_{it} - \hat{\Pi}_{YX}\mathbf{X}_t(s_i - \bar{s}) \right)$ on X_t and a constant, and the time-level cluster robust standard errors in the micro data. The standard errors in (22) can then be used to construct confidence intervals for β_h that will be uniformly valid over the degree of signal-to-noise in the data.

Remark 6. (Relaxing assumption 2(i).) It is possible to extend the model to allow for more flexible dynamics in the impulse. Suppose we replace the shock X_t in (20) by $\widetilde{X}_t$ which satisfies

$$\widetilde{X}_t = \sum_{\ell=0}^{p} \widetilde{B}_\ell \widetilde{X}_{t-\ell} + X_t,$$

The vector $(Y_{it}, \tilde{X}_t)$ then follows a heterogeneous finite-order joint VAR model and the results of this section apply. Lag-augmentation (including p+1 lags) is important to obtain Corollary 1, as in Montiel Olea and Plagborg-Møller (2021).

Discussion and recommendations for empirical work. These results also provides intuitive guidance for the practical implementation of inference recipes.

First, it is reasonable to expect lag augmentation to approximate reasonably well the underlying covariance structure, provided a conservative lag choice is made, along the lines of Montiel Olea and Plagborg-Møller (2021) and recommendations in the VAR tradition (Kilian and Lütkepohl, 2017).

Second, it is worth stressing the residual nature of the terms inducing serial correlation in Proposition 1, even under general dynamics in (8)–(9), which only

gets magnified as we consider environments with more aggregate and idiosyncratic noise. This intuition applies to simple local projection estimators; lag augmentation is likely to further mitigate these concerns. On the other hand, off-the-shelf (panel) HAC estimators such as (18) are likely to display a poor finite-sample performance, in line with existing results in the time series literature (Lazarus et al., 2018; Herbst and Johansenn, 2023). The results in our simulations align with this discussion; see Section 5.3 for further elaboration.

Third, heteroskedasticity-robust (HR) inference on the synthetic data is simple and easiest to implement, not only relative to standard practice in applied work but also relative to the HAR standard errors in (19), only requiring a one-step clustering approach over the cross-sectional dimension. The relative simplicity of HR versus HAR approaches is maintained when we consider refined counterparts; namely the equally-weighted cosine approach of Müller (2004) recommended by Lazarus et al. (2018) and the proposal by Imbens and Kolesár (2016), respectively. Our simulations suggest that lag-augmented HR inference tends to perform best across a wide range of scenarios, including samples of just $T = 30$ periods and moderate horizons. HAR inference is a competitive alternative if coupled with the refinement and few controls are used.

Our general practical recommendation is to include a reasonable number of controls — unit-level lagged outcomes and lagged shocks — in local projections and then compute standard errors that simply cluster over the cross-sectional dimension. Our simulations also suggest that performance is significantly improved with the correction suggested by Imbens and Kolesár (2016).

3.4 Proxy shocks and instrumental variables

The pooled and interacted local projection estimators in (10), where X_t are assumed direct observations of some structural shock, are by far the most common implementation in empirical work. Even if this is the case, ultimate interest is in an aggregate, endogenous state variable $\widetilde{X}_t$ and not in X_t — in changes in the policy rate and not in shocks themselves.

In fact, it is more realistic to allow for some measurement error in the shock elicitation process, and treat these generated variables as (external) instruments for the actual underlying shock (Ramey, 2016; Stock and Watson, 2018). In order to accommodate these environments, consider the following design,

$$Y_{it} = \mu_i + \beta_i(L)\widetilde{X}_t + v_{it}, \tag{23}$$

$$v_{it} = \gamma_i(L)Z_t + \kappa\delta_i(L)u_{it}, \tag{24}$$

where $\widetilde{X}_t$ might be endogenous in the sense that

$$\widetilde{X}_t = \tilde{\beta}_0 X_t + \tilde{\gamma}(L)Z_t,$$

where $\tilde{\gamma}(L) = \sum_{\ell=0}^{\infty}\tilde{\gamma}_\ell L^\infty$ (absolutely summable), $\tilde{\beta}_0 \neq 0$ and X_t is unobserved but a valid instrument V_t satisfying Assumption 3 below is available.

Assumption 3 (LP-IV).

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

(ii) $\mathrm{Cov}(X_t, V_t) \neq 0.$

Assumption 3(i) requires V_t to satisfy an analogous condition to 2(i) and Assumption 3(ii) requires V_t to be informative about X_t.[24] This is a standard assumption in the literature on instrument relevance and (lead-lag) exogeneity,[25] and allows for V_t to be a noisy measurement of the underlying structural shock of interest.

Consider the reduced-form estimator

$$\hat{\beta}_h^{\mathrm{RF}} = \frac{\sum_{i=1}^N\sum_{t=1}^{T-h}(s_i V_t - \tilde{\Pi}' W_{it})Y_{i,t+h}}{\sum_{i=1}^N\sum_{t=1}^{T-h}(s_i V_t - \tilde{\Pi}' W_{it})^2}, \tag{25}$$

where $\bar{V}_t = (T-h)^{-1}\sum_{t=1}^{T-h} V_t$, W_{it} contain unit and time fixed effects and we have defined $\tilde{\Pi} = \left(\sum_{i,t} W_{it}W_{it}'\right)^{-1}\sum_{i,t} W_{it}s_i V_t$. Consider the first-stage estimator

$$\hat{\beta}_h^{\mathrm{FS}} = \frac{\sum_{t=1}^{T-h}(V_t - \bar{V})\widetilde{X}_t}{\sum_{t=1}^{T-h}(V_t - \bar{V})^2}, \tag{26}$$

[24] Trivially, the results in this section apply to the case $Z_t = X_t$ (almost surely) too.

[25] See, for instance, Stock and Watson (2018, p. 924) or Plagborg-Møller and Wolf (2021, p. 970). Extensions where Assumption 3(i) holds conditional on controls might also be entertained.

where for simplicity we have omitted any possible controls.[26] The LP-IV estimator is then simply $\hat{\beta}_h^{\text{IV}} = \hat{\beta}_h^{\text{RF}} / \hat{\beta}_h^{\text{FS}}$.

Corollary 2 shows that the LP-IV estimator recovers the interacted (or pooled) local projection estimand (14) and that an analogous characterization of inference follows by adapting Proposition 1. The result is immediate once we recognize that $\hat{\beta}_h^{\text{RF}}$ has the same representation as that of $\hat{\beta}_h$ in (11).

Corollary 2. *Under Assumptions IV1, IV2 and 3, as $N, T \to \infty$,*

$$\hat{\beta}_h^{IV} \xrightarrow{\text{p}} \beta_h,$$

where β_h is defined in (14). Additionally, the autocovariance function of the reduced-form regression score has the same properties as $X_t \xi_{ht}$ in Proposition 1.

Proof. See Supplemental Material SM.A. Assumptions IV1 and IV2 simply extend 1 and 2 to include V_t. □

Through the lens of model (23)–(24), the LP-IV estimand has a relative impulse response interpretation: it identifies average responses to a shock in X_t that raises $\widetilde{X}_t$ by one unit on impact.

Remark 7. (No first-stage heterogeneity.) That the LP-IV estimand correctly identifies (relative) impulse responses might seem a natural consequence of Assumption 3; see, for instance, Stock and Watson (2018). However, it is not so obvious: under treatment effect heterogeneity, IV estimands often identify (weighted averages of) local average treatment effects (Angrist and Imbens, 1995; Angrist, Imbens, and Graddy, 2000). Yet another manifestation of the micro-macro duality implies that this is not the case here, due to the aggregate-only nature of the first-stage model.

4 The role of cross-sectional variation

In this section, we revisit the previous results through the lens of different views on the nature of observable unit-level characteristics, denoted s_i so far.

[26] Note that there is no need to use $T-h$ observations in the first-stage. We keep them here for analytic simplicity.

Heterogeneity and state-dependence. A quick glance at the existing body of applications reveals the widespread use of both time-invariant s_i and time-varying s_{it} observables. A more meaningful distinction is that between heterogeneity and state-dependence.

In some cases, interest is in the differential sensitivity of responses along some strictly exogenous characteristic, and these are thought of as immutable during the panel period. This includes applications where s_{it} might be approximately time-invariant, as it is the case with states that very over very low frequencies. For instance, Crouzet and Mehrotra (2020) list reasons that might alleviate the concern of firm reclassification across size bins when studying differential responses by size to monetary policy shocks.[27]

In other instances, focus might shift to the differential (and possibly dynamic) pass-through of shocks to responses along an observable state, which might vary substantially over time and thus interest is on its level at the time of the shock, or right before. For example, Ottonello and Winberry (2020) explore how firms respond to monetary policy when they have higher or lower default risk than usual. We can formalize these choices by extending (8)–(9) to allow for time-varying impulse responses,

$$\begin{aligned} Y_{it} &= \mu_i + \beta_{it}(L)X_t + v_{it}, \\ v_{it} &= \gamma_{it}(L)Z_t + \kappa\delta_{it}(L)u_{it}, \end{aligned}$$

where $\beta_{it}(L) = \sum_{\ell=0}^{\infty} \beta_{it\ell} L^{\ell}$ and so on. The coefficient on $s_{i\,t} X_t$ of a local projection at horizon h with unit and time fixed effects retains its interpretation as the slope coefficient of the linear projection $\mathbb{E}^* \left[\beta_{i\,th} \middle| s_{i\,t} \right]$ as long as $s_{i\,t}$ and impulse responses are exogenous with respect to X_t. Although a more detailed exploration is beyond

[27] Similar arguments are made in Drechsel (2023), Singh, Suda, and Zervou (2023) and Caglio, Darst, and Kalemli-Özcan (2024), among others. When observables are subpopulation indicators such as wealth deciles or size bins, these concerns are greatly alleviated. Often s_{it} is set at its value at $t = 0$ on similar grounds.

the scope of this paper, the treatment of s_{it} is here analogous to that of s_i, and the results in Section 3.3 carry over without much modification.[28,29]

When are s_i instruments for precision? The availability of rich micro data offers the promise of both revealing new insights into the heterogeneous transmission channels of macro shocks and their precise identification and estimation.

In particular, it can be shown that s_i can be exploited for additional precision — in the sense of faster-than-$\sqrt{T-h}$ convergence rates — if the following two conditions are satisfied:

1. Instrument relevance: $\text{Cov}(s_i, \beta_{ih}) \neq 0$.
2. Instrument exogeneity:
   (i) $\text{Cov}(s_i, \beta_{i\ell}) \neq 0$ for all $\ell \geq 0$ and $\ell \neq h$, and
   (ii) $\text{Cov}(s_i, \gamma_{i\ell}) \neq 0$ for all $\ell \geq 0$.

Essentially, we require that s_i is informative about heterogeneity in transmission of X_t at horizon h, but orthogonal to all other exposures to aggregate shocks. Condition 2(i) seems particularly hard to meet: for a each horizon h, a source of variation that is orthogonal to responses at all other horizons is required. In some sense, this reveals an intrinsic trade-off between documenting interesting transmission mechanisms and finding valid instruments for precision.[30]

These conditions can be directly read off the components of the estimation error in (12)–(13). Loosely, we need that $\tilde{X}_{(-h),t+h}$ and $\tilde{Z}_{t+h}$ are of order $O_p\left(N^{-1/2}\right)$, which requires the limiting value of all $\bar{\beta}_\ell$ and $\bar{\beta}_\ell$ in (13) to be zero.[31]

---

[28] We also explore these setups in simulations in Section 5.

[29] 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 $s_t = \mathbb{1}\{X_t > c\}$.

[30] It should be stressed that these conditions need to hold on top of the more basic strict exogeneity (identification) condition in assumption 2, which is regularly discussed in empirical work.

[31] Note that it is natural to normalize $\kappa = 1$ in these scenarios, which are closer to standard panel data models with both unit and time variation. Intuitively, there is no micro-macro duality in the estimation

These insights are part of a more general message: strong spatial dependence induced by aggregate shocks and all its implications for inference are inherent to our setup unless one is willing to restrict the (arguably natural) symmetry conditions placed on the observed and unobserved macro shocks in (8)–(9); be it in the form of absence of the latter, availability of cross-sectional instruments or homogeneity in responses.

5 Simulations

We now explore the finite-sample properties of the estimators and inference procedures analyzed in section 3 by means of an extensive Monte Carlo simulation study. We consider three designs:

(i) A heterogeneous linear process model (HLP design) that allows for both unit-specific IRFs and flexible dynamics in observed and unobserved, macro and micro shocks as in our general setup in (8)–(9). We use this design to assess the properties of pooled and interacted LP estimators and the performance of simple lag-augmented HR inference versus HAR inference. We also illustrate the role of genuine cross-sectional variation in the interacted LP case, including the notions discussed in Section 4.

(ii) A heterogeneous vector autoregression model (HVAR design) that restricts outcome dynamics to be generated by a VAR while allowing for individual heterogeneity, as in model (20). With this design, we explore the role of nonstationarity (in the form of near unit roots) of both macro and micro shocks.

(iii) A heterogeneous local projection instrumental variable model (HLPIV design) as in (23)–(24) that generalizes the LP-IV setup to the panel case allowing for heterogeneous micro IRFs. We use this design to explore the behavior of inference procedures in the presence of endogeneity.

error anymore, since macro shocks "look like" micro shocks in the sense that they average out to zero along a particular subpopulation defined by s_i.

For each design, we simulate $n_{\text{MC}} = 5,000$ samples on which we implement the confidence intervals to be considered. We then measure the coverage rates of confidence intervals. The designs are indexed by sample sizes (N, T) and κ (the micro noise-to-macro signal index), which we choose to cover a wide range of empirically realistic settings.[32]

Outline. This section is organized as follows. We describe the DGPs in 5.1, we present the inference procedures in 5.2, and we summarize the results in 5.3.

5.1 Data generating processes

Our simulation study relies on three different DGPs.

HLP design. We simulate samples from model (8)–(9) under the following conditions:

- We draw the macro shocks X_t and Z_t as $N(0,1)$ i.i.d. over time and the micro shocks u_{it} as $N(0,1)$ i.i.d. over units and time.
- We draw $\mu_i \sim N(0,1)$ as i.i.d. over units, independent of shocks and coefficients.
- We draw a vector $\tilde{s}_i$ as i.i.d. over units according to

$$\tilde{s}_i = \begin{pmatrix} s_i \\ s_{\gamma,i} \\ s_{\delta,i} \end{pmatrix} \sim N\left(1_{3\times 1}, (1-\rho_s)I_3 + \rho_S 1_{3\times 3}\right).$$

  Here, s_i is the descriptor of heterogeneity observable to the researcher whereas the unobserved $s_{\gamma,i}, s_{\delta,i}$ introduce correlation between s_i and the responses to Z_t and u_{it} at different horizons.

[32] For clarity we begin by presenting results for a set of designs with scalar outcome Y_{it} and descriptor of heterogeneity s_i, and with i.i.d. normally distributed shocks. In the supplemental appendix, we provide results for multivariate Y_{it} and s_i, and for both non-normal and conditionally heteroskedastic shocks. The findings we discuss below appear robust to reasonable amounts of non-normality and heteroskedasticity, and generalize to larger dimensions of Y_{it} and s_i.

- For a large $\bar{L}$ and $\beta_i(L) = \sum_{\ell=0}^{\bar{L}} \beta_{i\ell} L^\ell$, $\gamma_i(L) = \sum_{\ell=0}^{\bar{L}} \gamma_{i\ell} L^\ell$, $\delta_i(L) = \sum_{\ell=0}^{\bar{L}} \delta_{i\ell} L^\ell$, we set

$$\beta_{i\ell} = \frac{s_i \tilde{\beta}_{i\ell}}{\sqrt{\sum_{l=0}^{\bar{L}} \tilde{\beta}_{il}^2}}, \quad \gamma_{i\ell} = \frac{s_{\gamma,i} \tilde{\gamma}_{i\ell}}{\sqrt{\sum_{l=0}^{\bar{L}} \tilde{\gamma}_{il}^2}}, \quad \delta_{i\ell} = \frac{s_{\delta,i} \tilde{\delta}_{i\ell}}{\sqrt{\sum_{l=0}^{\bar{L}} \tilde{\delta}_{il}^2}},$$

  where coefficients $\{\tilde{\beta}_{i\ell}, \tilde{\gamma}_{i\ell}, \tilde{\delta}_{i\ell}\}_{\ell=0}^{\bar{L}}$ are obtained by drawing the roots of ARMA polynomials from a set of Beta distributions, computing their implied MA(∞) representations, and truncating them at $\bar{L}$. See Appendix C for details.[33]

- We also generate time-varying heterogeneity $s_{it} = s_i + \zeta_{it}$ where $\zeta_{it} \sim N(0,1)$, i.i.d. over units and time, and independent of s_i and everything else.

- We calibrate $\bar{L} = 2T$ and $\rho_s = 0.5$. We also set the parameters of the distributions underlying $\{\tilde{\beta}_{i\ell}, \tilde{\gamma}_{i\ell}, \tilde{\delta}_{i\ell}\}_{\ell=0}^{\bar{L}}$ to generate rich cross-sectional variation in IRFs and responses to X_t, Z_t and (particularly) u_{it} that for a large fraction of the cross-sectional population last for several periods (see figure 1). We are interested in situations featuring lots of positive persistence of micro shocks as this seems the empirically relevant case.[34]

[33] We normalize the coefficient by $\sqrt{\sum_{l=0}^{\bar{L}} \tilde{\beta}_{i\ell}^2}$ to separate the roles of s_i (controlling the scale of the response to X_t) and $\{\tilde{\beta}_{i\ell}\}_{\ell=0}^{\bar{L}}$ (controlling the persistence of the response over time). Assuming X_t is white noise with unit variance conditional on $\{\beta_{i\ell}\}_{\ell=0}^{\bar{L}}$, the variance of $\beta_i(L)X_t$ is $\sum_{\ell=0}^{\bar{L}} \beta_{i\ell}^2 = s_i^2$ while the ratio of long-run variance to variance of $\beta_i(L)X_t$ (a measure of persistence) is

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

which does not depend on s_i.

[34] As a measure of persistence, we report the ratio of long- to short-run standard deviations of $\beta_i(L)X_t$, $\gamma_i(L)Z_t$ and $\delta_i(L)u_{it}$ given $\{\beta_{i\ell}, \gamma_{i\ell}, \delta_{i\ell}\}_{\ell=0}^{\infty}$. An AR(1) process with AR coefficient ρ has a long- to short-run ratio of $\sqrt{(1+\rho)/(1-\rho)}$. The mean ratio for $\beta_i(L)X_t$ (around 3) is comparable to that of an AR(1) process with $\rho = 0.8$. The mean ratio for $\gamma_i(L)Z_t$ is comparable to $\rho = 0.77$ while that for $\delta_i(L)u_{it}$ to $\rho = 0.94$. The dynamics produced by our model are however richer than that of an AR process and cannot be fully captured by a finite number of lags of Y_{it} or X_t — although they can be well approximated. A DGP with such exact VAR dynamics is discussed below.

HVAR design. We simulate samples from the heterogeneous VAR model in (20).[35] We specify:

- We draw macro shocks X_t and Z_t, micro shocks u_{it} and observable s_i (together with unobserved $s_{\gamma,i}, s_{\delta,i}$) as in the HLP model.
- We draw $m_i \sim N(0,1)$ independent of the shocks and coefficients and i.i.d. over units, and we simulate the coefficients by drawing the roots of polynomials $A_i(L)$ and $B_i(L)$ from Beta distributions. Then,

$$B_{i\ell} = \frac{s_i \tilde{B}_{i\ell}}{\sqrt{\sum_{l=0}^{p} \tilde{B}_{il}^2}}, \quad C_{i0} = s_{\gamma,i}, \quad D_{i0} = s_{\delta,i}.$$

- We calibrate $\bar{L} = 2T$ and $\rho_s = 0.5$. We also set the mean of the largest root of $A_i(L)$ to $1 - 2/T$.[36] We illustrate the IRFs in figure 1, which shows substantially more persistence in responses to shocks compared to the HLP design.

HLPIV design. We simulate samples from the heterogeneous linear model (23)–(23) with

$$\widetilde{X}_t = \tilde{\beta}_0 X_t + \tilde{\gamma}(L) Z_t + \tilde{\delta}(L) \tilde{U}_t,$$

where $\tilde{\gamma}(L) = \sum_{\ell=0}^{\bar{L}} \tilde{\gamma}_\ell L^\ell$ and $\tilde{\delta}(L) = \sum_{\ell=0}^{\bar{L}} \tilde{\delta}_\ell L^\ell$ for the same truncation lag $\bar{L}$ used for $\beta_i(L)$, $\gamma_i(L)$ and $\delta_i(L)$.

- We draw macro shocks X_t, Z_t and errors $\tilde{U}_t$ as $N(0,1)$ i.i.d. over time, and micro shocks u_{it} as $N(0,1)$ i.i.d. over units and time. We draw $\mu_i, \{\beta_{i\ell}, \gamma_{i\ell}, \delta_{i\ell}\}_{\ell=0}^{\infty}$ as in the HLP design together with the observable characteristics s_i.
- We generate $\tilde{\gamma}(L)$ and $\tilde{\delta}(L)$ by setting the roots of AR and MA polynomials and proceeding as before: we obtain the implied MA(∞) representations by long

[35] Note that we truncate the infinite-order lag polynomials in the HLP model while we use the recursive nature to simulate the path of outcomes in the HVAR model. This is relevant when trying to generate data with high persistence as we do below.

[36] Indexing the mean of the largest AR root to T is in the spirit of the local-to-unity analysis of time series models with high persistence.

division, we truncate them at $\bar{L}$ and we normalize the transfer functions so that $\tilde{\gamma}(L)Z_t$ and $\tilde{\delta}(L)\tilde{U}_t$ have unit variance. We set $\tilde{\beta}_0 = 1$.

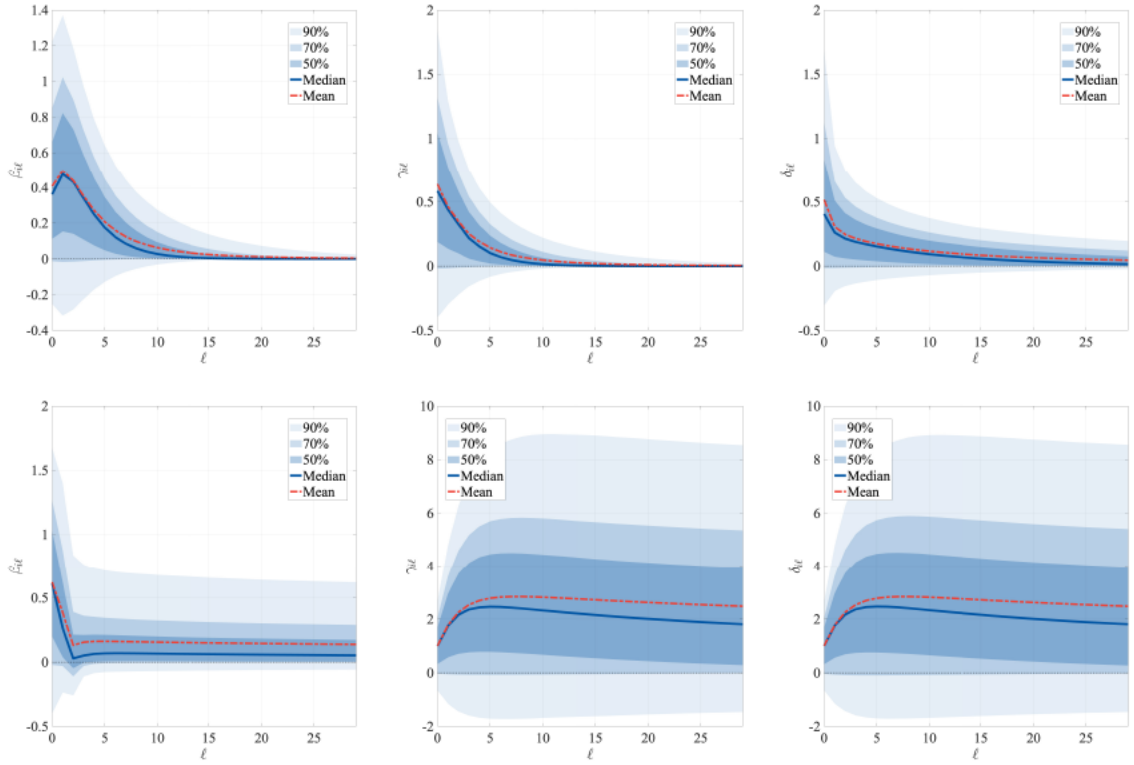

FIGURE 1. Heterogeneity of IRFs in the HLP, HVAR and HLPIV designs.

Note: Upper panels present the distributions (pointwise in ℓ) of $\beta_{i\ell}$, $\gamma_{i\ell}$ and $\delta_{i\ell}$ estimated from 100,000 draws from HLP and HLPIV DGPs. Lower panels present the distributions for the HVAR model.

Values of T, N and κ. We choose $T = 30$ and $T = 100$, which are typical values in the empirical literature for datasets with annual and quarterly observations, respectively. We take $N = 1,000$ and use $\kappa = 5, 10, 30$. Expressed in terms of macro and micro R^2 of the contribution of macro shocks to the variance of $\bar{Y}_t$ and Y_{it}, respectively, $\kappa = 5$ implies a macro R^2 of 0.99 and a micro R^2 of 0.07. Similarly,

$\kappa = 10$ implies macro R^2 of 0.95 and micro R^2 of 0.02 while $\kappa = 30$ implies macro R^2 of 0.69 and micro R^2 of 0.002.[37]

5.2 Estimators and confidence intervals

For $i = 1, \ldots, N$ and $t = 1, \ldots, T$, the researcher observes Y_{it}, X_t and s_i. She also observes s_{it} in the HLP model and $\tilde{X}_t$ in the HLPIV model.

Estimators. The estimators implemented in each sample are the following:

(i) Pooled LP estimator obtained by regressing $Y_{i,t+h}$ on X_t and unit fixed effects.

(ii) Interacted LP estimator obtained by regressing $Y_{i,t+h}$ on $s_i X_t$ and both unit and time fixed effects.

(iii) Micro interacted LP estimator obtained by regressing $Y_{i,t+h}$ on $s_{it} X_t$ and both unit and time fixed effects (only in HLP model).

(iv) Pooled LP-IV estimator obtained by regressing $Y_{i,t+h}$ on $\tilde{X}_t$ and unit fixed effects using X_t as an instrument of $\tilde{X}_t$ (only in HLPIV model).

(v) Interacted LP estimator obtained by regressing $Y_{i,t+h}$ on $s_i \tilde{X}_t$ and both unit and time fixed effects using $s_i X_t$ as an instrument of $s_i \tilde{X}_t$ (only in HLPIV model).

Inference procedures. The procedures we will compare include one-way unit-level clustering (Arellano (1987)), HR inference (Eicker (1967), Huber (1967), White (1985)) applied to the synthetic time series, two-way clustering, HAR inference (Newey and West (1987)) applied to the aggregated data (i.e., Driscoll and Kraay (1998)), and HAR inference combined with unit-level clustering (Thompson (2011)). HR inference is implemented with lag augmentation while HAR inference is done without lag augmentation.

[37] We also produced results for $N = 10,000$ with similar findings. As in our asymptotic analysis, provided N is large relative to T, the key quantity appears to be $\kappa/\sqrt{N}$, which controls the proportion of macro signal to micro noise in the data.

We also include simple finite-sample refinements of confidence intervals in our simulations, motivated by the relatively small time-series sample size in applications. For HR inference we follow Imbens and Kolesár (2016) and apply the HC2 correction of MacKinnon and White (1985) to the Eicker-Huber-White standard error. For HAR inference we adopt the equally-weighted cosine (EWC) estimator of the long-run variance (Müller (2004)) recommended by Lazarus et al. (2018). Both standard errors are combined with Student-t critical values with easy-to-determine degrees of freedom and are readily available in standard econometric software.

The confidence intervals implemented are the following:

(1W) One-way clustering standard errors.

(HAR) HAR standard errors (Newey-West with $0.75(T-h)^{1/3}$ autocovariances) on the synthetic time series (i.e., Driscoll and Kraay (1998)). No lags.

(HAR$_h$) HAR standard errors that exploit the MA(h) dynamics of the score (i.e., Newey-West with h autocovariances). No lags.

(HAR$_+$) HAR standard errors (EWC with $0.4(T-h)^{2/3}$ cosine functions) on the synthetic time series with Student-t critical values (i.e., Lazarus et al. (2018)). No lags.

(2W-HAR) HAR + unit-level clustered standard errors with $0.75(T-h)^{1/3}$ autocovariances (i.e., Thompson (2011)). No lags.

(HR$_0$) HR standard errors (Eicker-Huber-White). No lags.

(HR) HR standard errors (Eicker-Huber-White). Lag augmentation with $p = T^{1/3}$ lags of Y_{it} and X_t (or $s_i X_t$ or $s_{it} X_t$).

(HR$_h$) HR standard errors (Eicker-Huber-White). Lag augmentation including h lags of X_t (or $s_i X_t$ or $s_{it} X_t$).

(HR$_+$) HR standard errors (Eicker-Huber-White + HC2 correction) on the synthetic time series with Student-t critical values (i.e., Imbens and Kolesár (2016)). Lag augmentation including $p = T^{1/3}$ lags of Y_{it} and X_t (or $s_i X_t$ or $s_{it} X_t$).

(2W-HR) Two-way clustering standard errors. Lag augmentation with $p = T^{1/3}$ lags of Y_{it} and X_t (or $s_i X_t$ or $s_{it} X_t$).

5.3 Results

We report coverage rates for the HLP design in figures 2 ($T = 30$) and 3 (for $T = 100$). Results for the HVAR and HLPIV designs are displayed in figures 4 and 5

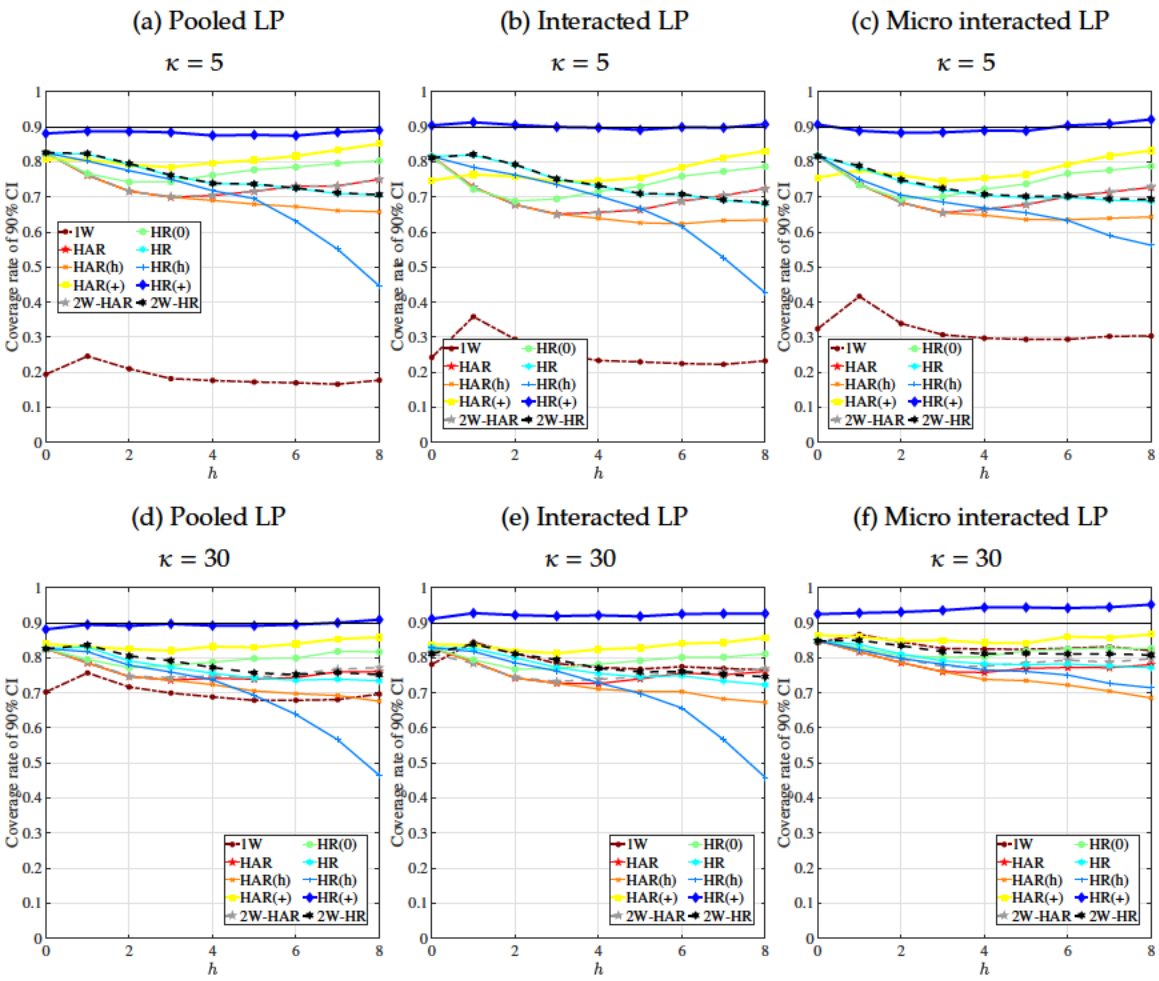

FIGURE 2. Coverage rates of 90% CIs. HLP design: $T = 30$.

Takeaways. First, in all of the designs we have entertained, lag-augmented HR inference with modest lag length and the refinement of Imbens and Kolesár (2016) emerges as the best alternative. It attains coverage very close to the nominal rate even in samples of just $T = 30$ periods and for responses at horizons about a third of the time series sample size. HAR inference based on the EWC approach of Lazarus

et al. (2018) also performs well with small T when no lags are used.[38] In addition, a weaker aggregate signal (in the form of a larger κ) tilts the comparison in favor of HR inference as it decreases the relative importance of the autocovariance terms in the long-run variance of the score. This is validated by our simulations.

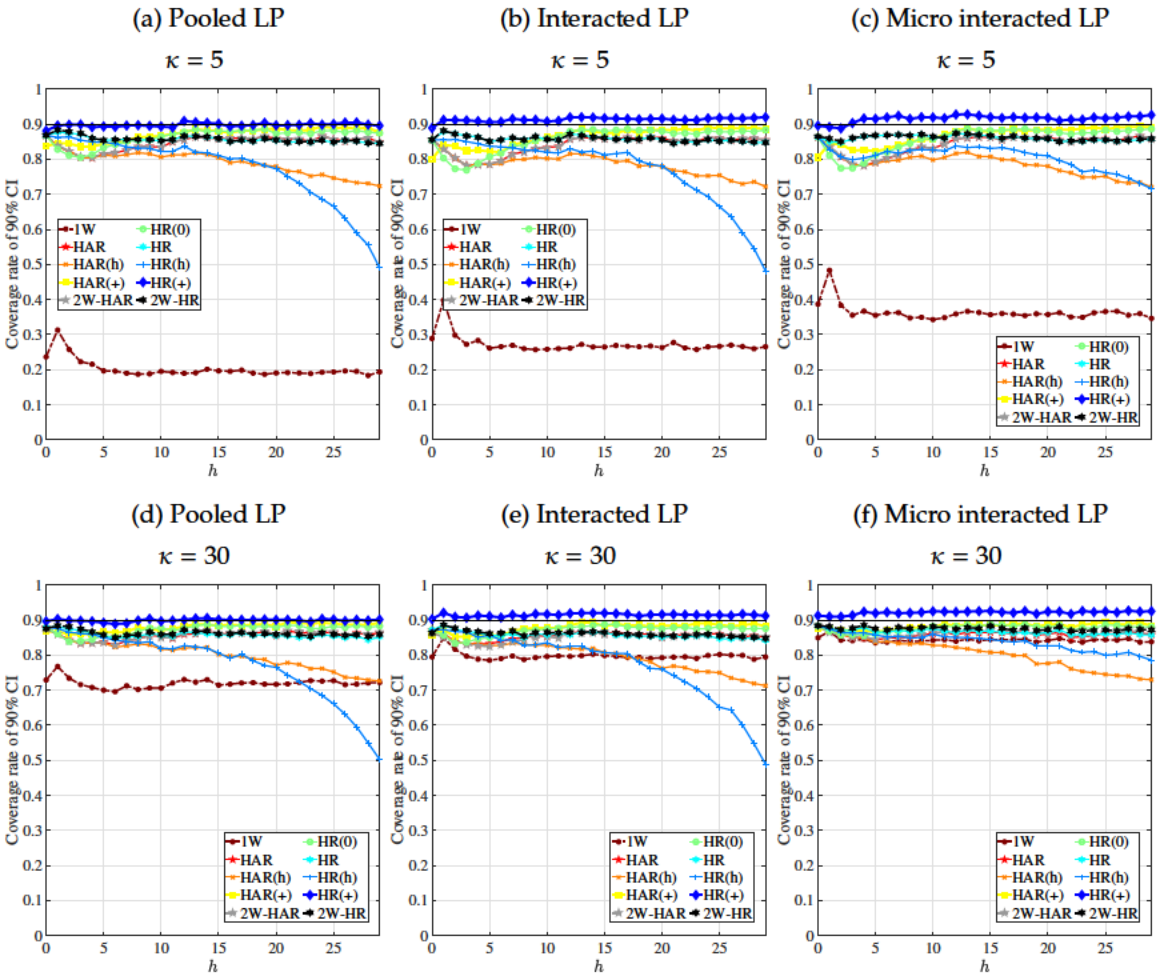

FIGURE 3. Coverage rates of 90% CIs. HLP design: $T = 100$.

Second, when small-sample refinements are not employed, lag-augmented HR and non-lag-augmented HAR approaches perform well for T large ($T = 100$ in our

[38] The performance of HAR procedures deteriorates when lags are used. That including lags and accounting for serial correlation in the score are substitutes from the point of view of obtaining correct inference aligns well with the theory of Section 3.

experiments) but suffer large coverage distortions for smaller T. Moreover, recipes that estimate the h autocovariance terms in the long-run variance of the score or control for h lags during estimation do not retain coverage at long horizons.[39]

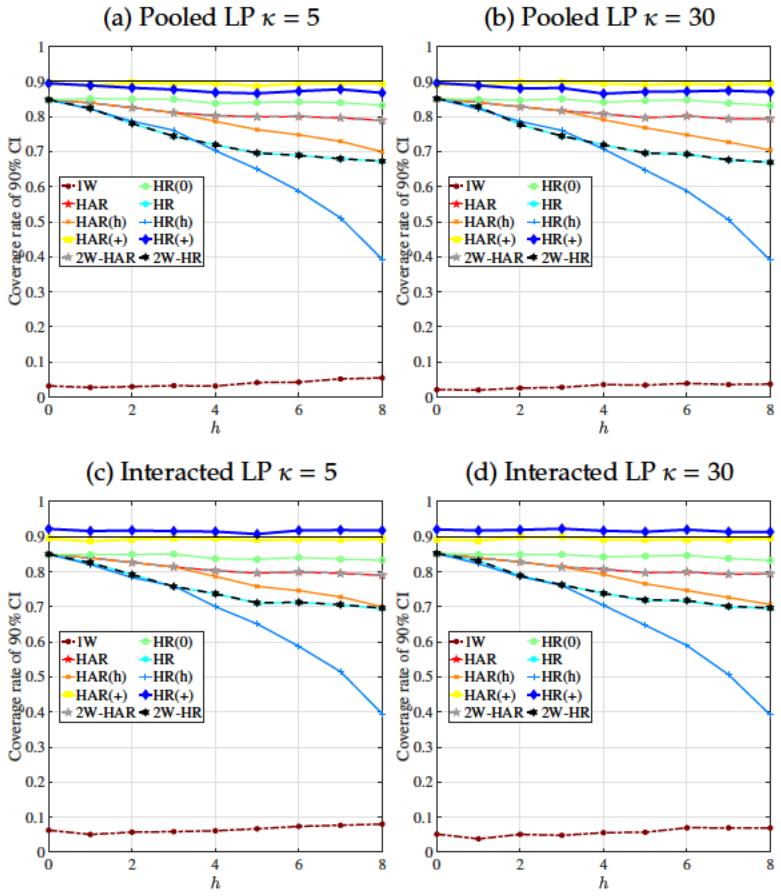

FIGURE 4. Coverage rates of 90% CIs. HVAR design: $T = 30$.

Third, clustering at the unit level either produces undercoverage (one-way clustering) or statistical decisions that are no different to what one would get from their non-clustered counterparts. For example, HR inference on the synthetic time series and two-way clustering are almost identical; the same applies to Driscoll and Kraay (1998) and Thompson (2011). This holds true regardless of the value of κ in

[39] This suggests that the requirement that h be sufficiently small relative to T in proposition 2 cannot be dispensed with when characterizing inference recipes that exploit the MA(h) serial dependence structure of the score.

line with the uniformity results of our theory. Overall, clustering at the unit level produces intervals that are too short or contribute terms that cancel out with the double-count terms in two-way approaches.

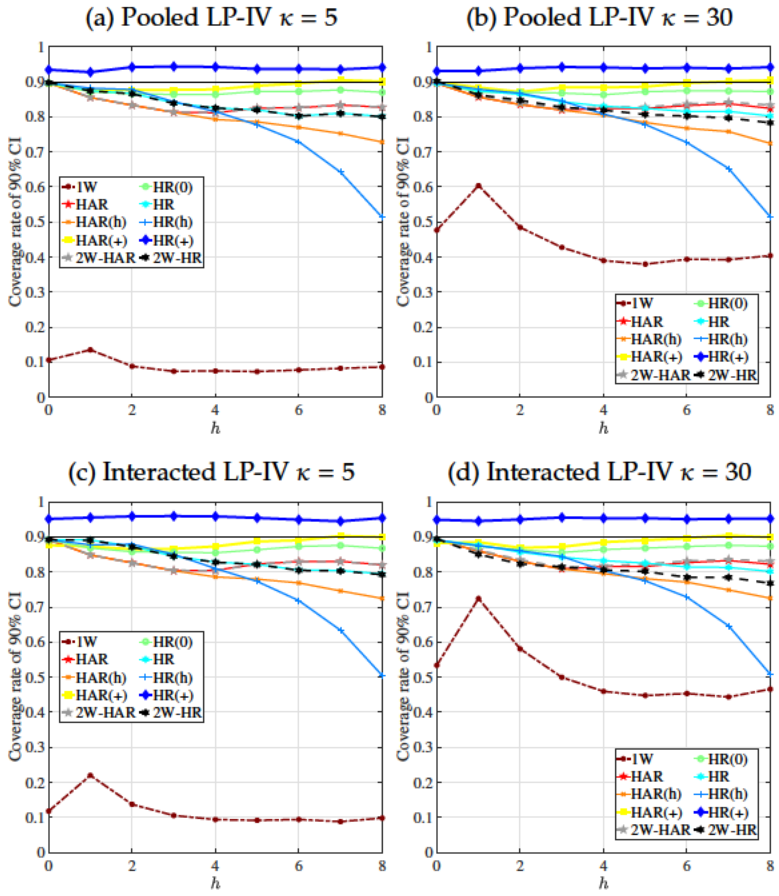

FIGURE 5. Coverage rates of 90% CIs. HLPIV design: $T = 30$.

6 Conclusion

The use of micro data to answer macro questions offers a n e xciting a venue to study how agents respond to economy-wide policies. Possibilities include a better understanding of the transmission of shocks and the extent of heterogeneity in responses.

Challenges are ubiquitous too. We propose a disciplined approach to uncertainty quantification in an environment where interest is in parameters identified mostly by aggregate variation and (pervasive) idiosyncratic and common shocks coexist. We then suggest simple yet robust inferential tools.

Simplicity by no means should suggest plain enumeration of options, however: different choices are informative about different notions of statistical uncertainty, target parameters and external validity of results. These features are particularly salient when combining micro and macro data (Nakamura and Steinsson, 2018). As such, we believe that applied practice would benefit from a more explicit discussion of these notions in a given particular problem. Our hope is that the tools offered in this paper can help such an effort.

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. 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 embedding. Formalizing the idea of (possibly heterogeneous) non-stationarities along these lines seems promising and full of empirical content. On a different note, extensions to simultaneous inference over impulse response horizons could be made building on Jordà (2009) and Montiel Olea and Plagborg-Møller (2019).

A Inference conditional on aggregate shocks

The micro-macro dichotomy has been a recurrent topic throughout this paper. Estimation of impulse responses with macro regressors needs to be explicit about the presence of statistical uncertainty that is common to all units in the panel which, in general, requires accounting for a strong form of spatial dependence in the data.

This has drastic consequences for uncertainty quantification approaches that exploit independence over units in the micro data, such as clustering at the unit level, which leads to confidence intervals that are too short and vastly undercover.

In this section, we provide a reinterpretation of such confidence intervals. We show that they provide valid inference for a different object — which can be interpreted as impulse responses that are indexed to the particular realizations of aggregate shocks that occurred during the historical period under consideration.

In other words, inference when the analyst behaves as-if there were no aggregate shocks is tantamount to inference conditional on the path of aggregate shocks. Conditional inference induces an internal/external validity trade-off; these impulse responses are tightly linked to the time window in which they were collected and thus lack generalizability to other environments.[40] At the same time, conditional inference requires substantially weaker conditions: it applies to short panels under fixed-T, large-N asymptotics and places virtually no assumptions on the nature of aggregate variation (ergodicity, stationarity...).

The focus on this section is conceptual: it highlights the importance of a structured approach to inference where an explicit estimand is a precursor of the inferential task, and exploits the convenient micro-macro duality of our framework to illustrate the implications of ignoring its second ingredient.

[40] The lack of generalizability of estimates obtained from micro data when the population is subject to unaccounted aggregate shocks has also been discussed by Rosenzweig and Udry (2020) and Deeb and de Chaisemartin (2022) in the context of RCTs or causal studies with individual-level treatments. Deeb and de Chaisemartin (2022) note that if units are subject to village-level shocks but one clusters at the level of randomization, one can still draw inference on a conditional average treatment effects estimand indexed by the realizations of these shocks.

Illustration in the simple example in Section 2. Recall the static setup in (2) with a two-way unobserved structure $v_{it} = Z_t + u_{it}$ and the pooled least squares estimator $\hat{\beta}$ in (4).[41] In Section 2, we studied the properties of $\hat{\beta}$ as an estimator of β_0. Letting $v_{it}(b) = Y_{it} - bX_t$ for each b, one can motivate the latter as the parameter that solves the following orthogonality condition:

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t v_{it}\left(\beta_0\right)\right] = 0, \tag{A.1}$$

which averages over both micro and macro sources of uncertainty. As in the time-series tradition, $\{X_t, Z_t\}_{t=1}^{T}$ are stochastic processes and $\hat{\beta}$ inherits randomness in aggregate conditions over repeated sampling. Alternatively, $\hat{\beta}$ can be seen as an estimator of β_0^{cond}, which is defined as the parameter that solves

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t v_{it}\left(\beta_0^{\text{cond}}\right)\middle|\{X_\tau, Z_\tau\}_{\tau=1}^{T}\right] = 0, \tag{A.2}$$

which takes $\{X_t, Z_t\}_{t=1}^{T}$ as fixed attributes of a specific historical episode, in the spirit of case studies or, more generally, of analyses that condition on aggregate variation.[42] In other words, the expectation in (A.1) averages over the hypothetical population distribution of the aggregate component; the expectation in (A.2) does so over its empirical distribution of aggregate shocks. Note that condition (A.2) implies

$$\beta_0^{\text{cond}} = \beta_0 + \frac{1}{T}\sum_{t=1}^{T} \omega_t Z_t, \tag{A.3}$$

[41] Note that here we are implicitly setting $\kappa = 1$. As it will become clear below, micro- and macro-level sources of uncertainty do not coexist anymore under this framework.

[42] The notion of conditioning on trends, business cycle variation or seasonal effects is natural to a long-standing panel data literature via the use of time dummies (see, for instance, Arellano, 2003, Section 5.2). More generally, it is implicit in many applications in short panels, including cross-sectional studies (see Andrews, 2005; Hahn et al., 2020, for reviews), where a stochastic modelling of the aggregate component is not possible. For instance, this is often the case in the literature on policy evaluation methods with few available periods, such as RCTs (Rosenzweig and Udry, 2020; Deeb and de Chaisemartin, 2022) and synthetic controls (Arkhangelsky and Hirshberg, 2023).

and the simple model thus helps illustrate how the estimand depends on the sample covariance between observed and unobserved aggregate shocks. Intuitively, the estimand is indexed to a particular macroeconomic regime; a sequence of realizations of the unobserved component which cannot be averaged out to zero in a given sample.[43]

In this case, only the micro term in the estimation error in (5), which varies over both dimensions of the panel, contributes to uncertainty in $\hat{\beta}$. Fixed-T inference is possible, and we can recover β_0^{cond} at a $N^{-1/2}$ rate. In the empirically prevalent case where $N \gg T$ with presumably limited time-series variation, $\hat{\beta}$ can be interpreted as a very precise estimate of β_0^{cond} and a noisy attempt to recover β_0. It can also be shown that

$$\sqrt{N}\left(\hat{\beta} - \beta_0^{\text{cond}}\right)\Big|\{X_\tau, Z_\tau\}_{\tau=1}^T \xrightarrow{\text{d}} N\left(0, V_0^{\text{cond}}\right), \quad V_0^{\text{cond}} = \frac{1}{T}\sum_{t=1}^T \omega_t^2 \mathbb{E}\left[u_{it}^2 \middle| \{X_\tau, Z_\tau\}_{\tau=1}^T\right]$$

regardless of the properties of $\{X_t, Z_t\}_{t=1}^T$. Given the residuals $\hat{v}_{it}$, the heteroskedasticity-robust variance on the micro data provides a consistent estimator of V_0^{cond}.

We show in the Supplemental Material that these insights extend to the more general setup considered in Section 3; the estimand is defined by conditional moment equations and fixed-T inference that is agnostic about aggregate dynamics is possible. Independence over units but unspecified patterns of serial dependence in unobservables now make clustering at the unit level a valid, convenient implementation.

Discussion. This approach to inference ties together the computation of standard errors and confidence intervals to an explicit discussion about the (micro/macro) sources of uncertainty in the data. In particular, it emphasizes the conditions needed to precisely recover the unconditional estimand β_0. Loosely, it boils down to

[43] The explicit form of the estimand in (A.3) is also reminiscent of the illustrative examples used by Hahn et al. (2020) to illustrate the consequences of ignoring aggregate shocks in estimation. In particular, setting $X_t = 1$ and $T = 1$ recovers their portfolio choice example (Section 3); in which case $\beta_0^{\text{cond}} = \beta_0 + Z_1$, where β_0 is the unconditional mean return of the risky asset.

whether the aggregate component provides sufficient "repetition" for approximate in-sample orthogonality between X_t and Z_t.

On a more speculative note, this might be an appealing approach to inference in some contexts. Robustness to aggregate dynamics is attractive when the notion of an unstable macroeconomic regime underlies the economic analysis (non-stationary or non-ergodic aggregates), as in the following example.

Example 1. (Case studies.) A limiting case of this lack of time series repetition are event studies, where $X_t = \mathbb{1}\{t = \tau\}$ for a particular time period τ. This is an obvious source of non-stationarities, and the possibility of time series averaging seems here far fetched. More generally, interest might be in the transmission of aggregate uncertainty to micro units during a recognizable historical episode, a notion natural to monetary policy applications, where regimes might be captured by breaks in the policy rule or other structural shocks. Coglianese, Olsson, and Patterson (2023) analyze one of such case studies, focusing on the worker and firm-level effects of a monetary quasi-experiment in Sweden in 2010, when the Riksbank raised interest rates substantially. Their monetary policy shock series spans a few pre- and post-2010 periods but nonetheless displays a single, large monetary policy shock at the time. Validity of conditional inference in short panels further suggests it might be a suitable choice here.

More generally, the notion of external validity inherent to β_0 and absent in β_0^{cond} is closely connected with the notion of time-invariant, "structural" unconditional responses. In more general setups, however, the issue of out-of-sample validity is not as obvious. For instance, if we rather hypothesize a sequence of parameters subject to structural breaks with no obvious connection between them, targetting conditional estimands seems more intuitive than some long-run average, which might not even be well-defined.

B Regularity conditions for limit theorems

We set separate conditions for the macro and micro components of the model. These involve restrictions on (i) the moments of X_t, Z_t and u_{it} and (ii) the temporal dependence induced by $(\beta_i, \gamma_i, \delta_i)$. Assumption 4 places (symmetric) conditions on existence of higher-order moments and on the persistence of macro shocks. Conditions on higher-order moments are required to establish uniformity results; the most stringent ones are needed to obtain bounds when proving consistency of the standard error in Proposition 2.[44] By iterated expectations, these also hold unconditionally. We also require sufficiently rapid decay almost surely on the persistence of macro shocks at the unit-level. As an example, the condition includes finite-order ARMA processes but rules out unit root processes. We impose analogous conditions for the micro component.

Assumption 4 (Regularity conditions for macro terms). *The following holds:*

(i) For some constants $\sigma_{X,\text{upp}}, C_{\beta,\ell}$ *such that* $C_\beta = \sum_{\ell=0}^{\infty} C_{\beta,\ell} < \infty$*, almost surely,*

$$\mathbb{E}\left[X_t^8 \middle| \{\theta_i\}_{i=1}^N\right] \leq \sigma_{X,\text{upp}}^8,$$
$$|\beta_{i\ell}| \leq C_{\beta,\ell}.$$

(ii) For some constants $\sigma_{Z,\text{upp}}, C_{\gamma,\ell}$ *such that* $C_\gamma = \sum_{\ell=0}^{\infty} C_{\gamma,\ell} < \infty$*, almost surely,*

$$\mathbb{E}\left[Z_t^8 \middle| \{\theta_i\}_{i=1}^N\right] \leq \sigma_{Z,\text{upp}}^8,$$
$$|\gamma_{i\ell}| \leq C_{\gamma,\ell}.$$

Assumption 5 (Regularity conditions for micro terms). *The following holds:*

(i) For some constants $\sigma_{U,\text{upp}}, C_{\delta,\ell}$ *such that* $C_\delta = \sum_{\ell=0}^{\infty} C_{\delta,\ell} < \infty$*, almost surely,*

$$\mathbb{E}\left[\|u_{it}\|^8 \middle| \theta_i\right] \leq \sigma_{U,\text{upp}}^8,$$
$$|\delta_{i\ell}| \leq C_{\delta,\ell}.$$

[44] Similar conditions are required in Montiel Olea and Plagborg-Møller (2021).

C Proof of Proposition 2

The (long-run) variance of the score is then given by

$$
\begin{aligned}
V_j &= \mathbb{E}_\kappa\left[X_t X_{t-j}\xi_{ht}\xi_{h,t-j}\right], \\
S_h &= V_0 + 2\sum_{j=1}^{h} V_j.
\end{aligned}
\tag{C.1}
$$

The proof requires various steps. The estimation error of $\hat{\beta}_h$ around $\bar{\beta}_h$ can be written as

$$
\begin{aligned}
\frac{\hat{\beta}_h - \bar{\beta}_h}{\hat{\sigma}_h} &= \frac{(T-h)^{-\frac{1}{2}}\sum_{t=1}^{T-h}(X_t - \bar{X})\,\xi_{ht}}{\sqrt{\hat{S}_h}} \\
&= \left[\frac{\sum_{t=1}^{T-h} X_t\xi_{ht}}{\sqrt{(T-h)S_h}} - \frac{\sqrt{T-h}\,\bar{X}\times(T-h)^{-1}\sum_{t=1}^{T-h}\xi_{ht}}{\sqrt{S_h}}\right]\times\sqrt{\frac{S_h}{\hat{S}_h}}.
\end{aligned}
$$

To establish uniformity, we use the connection with drifting parameter sequences highlighted by Andrews, Cheng, and Guggenberger (2020). For any sequence $\{\kappa_T\}$ such that $0 \le \kappa_T \le \bar{\kappa}\sqrt{N_T}$, we will show that

(A) $\{(T-h)S_h\}^{-\frac{1}{2}}\sum_{t=1}^{T-h} X_t\xi_{ht} \xrightarrow[\mathbb{P}_{\kappa_T}]{\mathrm{d}} N(0,1)$,

(B) $\hat{S}_h - S_h \xrightarrow[\mathbb{P}_{\kappa_T}]{\mathrm{p}} 0$,

(C) $(T-h)^{-1}\sum_{t=1}^{T-h}\xi_{ht} \xrightarrow[\mathbb{P}_{\kappa_T}]{\mathrm{p}} 0$ and $\sqrt{T-h}\,\bar{X} = O_{\mathbb{P}_{\kappa_T}}(1)$.

If (A), (B) and (C) hold, then for any sequence $\{\kappa_T\}$ satisfying $0 \le \kappa \le \bar{\kappa}\sqrt{N_T}$,

$$
\frac{\hat{\beta}_h - \bar{\beta}_h}{\hat{\sigma}_h} \xrightarrow[\mathbb{P}_{\kappa_T}]{\mathrm{d}} N(0,1),
$$

which is asymptotically equal to $(\hat{\beta}_h - \beta_h)/\hat{\sigma}_h$ under assumption 1(ii), $(T-h)/N_T \to 0$ and Slutsky's theorem. Proposition 2 follows from Andrews et al. (2020, Theorem 2.1(e)). We establish (A) in lemma 1, (B) in lemma 2 and (C) in lemma 3 in the Supplemental Material.

References

Alloza, M., J. Gonzalo, and C. Sanz (2023): "Dynamic Effects of Persistent Shocks." Working paper.

Almuzara, M. and V. Sancibrián (2024): "Micro responses to macro shocks." Federal reserve bank of new york staff report.

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. (1987): "Computing Robust Standard Errors for Within-Group Estimators." *Oxford Bulletin of Economics and Statistics*, 49, 431–434.

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

Arellano, M. and S. Bonhomme (2012): "Identifying distributional characteristics in random coefficients panel data models." *Review of Economic Studies*, 79, 987–1020.

Arkhangelsky, D. and D. Hirshberg (2023): "Large-Sample Properties of the Synthetic Control Method under Selection on Unobservables." Working Paper arxiv:2311.13575.

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

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.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011): "Robust Inference With Multiway Clustering." *Journal of Business and Economic Statistics*, 29, 238–249.

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

Coglianese, J., M. Olsson, and C. Patterson (2023): "Monetary Policy and the Labor Market: A Quasi-Experiment in Sweden," Working paper.

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

Deeb, A. and C. de Chaisemartin (2022): "Clustering and External Validity in Randomized Controlled Trials." Working Paper arxiv:1912.01052.

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.

Eicker, F. (1967): "Limit theorems for regressions with unequal and dependent errors." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 59–82.

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

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*.

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

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

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

Huber, P. J. (1967): "The behavior of maximum likelihood estimates under nonstandard conditions." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 221–233.

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 (forthcoming): "Q-Monetary Transmission." *Journal of Political Economy*.

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.

——— (2023): "Local Projections for Applied Economics." *Annual Review of Economics*, 15, 607–631.

Kilian, L. and H. Lütkepohl (2017): *Structural Vector Autoregressive Analysis*, Cambridge University Press.

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.

Liang, K. and S. Zeger (1986): "Longitudinal data analysis using generalized linear models." *Biometrika*, 73, 13–22.

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

MacKinnon, J. G. and H. White (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties." *Journal of Econometrics*, 29, 305–325.

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.

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.

Newey, W. K. and K. D. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, 55, 703–708.

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.

Romer, C. D. and D. H. Romer (2004): "A new measure of monetary shocks: derivation and implications," *American Economic Review*, 94, 1055–1084.

Rosenzweig, M. R. and C. Udry (2020): "External Validity in a Stochastic World: Evidence from Low-Income Countries." *Review of Economic Studies*, 87, 343–381.

Singh, A., J. Suda, and A. Zervou (2023): "Heterogeneity in labor market response to monetary policy: small versus large firms." Working paper.

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.

Thompson, S. (2011): "Simple formulas for standard errors that cluster by both firm and time." *Journal of Financial Economics*, 99, 1–10.

White, H. (1985): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica*, 48, 817–838.

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