FEDERAL RESERVE BANK OF SAN FRANCISCO

WORKING PAPER SERIES

Inference for Local Projections

Atsushi Inoue
Vanderbilt University

Òscar Jordà
Federal Reserve Bank of San Francisco
University of California, Davis
CEPR

Guido M. Kuersteiner
University of Maryland

August 2024

Working Paper 2024-29

https://doi.org/10.24148/wp2024-29

Suggested citation:

Inoue, Atsushi, Òscar Jordà, and Guido M. Kuersteiner. 2024. "Inference for Local Projections." Federal Reserve Bank of San Francisco Working Paper 2024-29. https://doi.org/10.24148/wp2024-29

The views in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Federal Reserve Bank of San Francisco or the Board of Governors of the Federal Reserve System.

Inference for Local Projections*

Atsushi Inoue† Òscar Jordà‡ Guido M. Kuersteiner§

13th August 2024

Abstract

Inference for impulse responses estimated with local projections presents interesting challenges and opportunities. Analysts typically want to assess the precision of individual estimates, explore the dynamic evolution of the response over particular regions, and generally determine whether the impulse generates a response that is any different from the null of no effect. Each of these goals requires a different approach to inference. In this article, we provide an overview of results that have appeared in the literature in the past 20 years along with some new procedures that we introduce here.

JEL classification codes: C11, C12, C22, C32, C44, E17.

Keywords: local projections, impulse response, instrumental variables, confidence bands, simultaneous bands, significance bands, wild block bootstrap.

*We thank the Royal Economic Society for the invitation to participate in the 2024 meetings in Belfast. Comments and suggestions from Raffaella Giacomini, Jaap Abbring, and conference participants helped us improve the paper. The views expressed in this paper are the sole responsibility of the authors and do not necessarily reflect the views of the Federal Reserve Bank of San Francisco or the Federal Reserve System.

†Vanderbilt University (atsushi.inoue@vanderbilt.edu).

‡Federal Reserve Bank of San Francisco; and Department of Economics, University of California, Davis (oscar.jorda@sf.frb.org; ojorda@ucdavis.edu) and CEPR.

§Department of Economics, University of Maryland (gkuerste@umd.edu).

1. Introduction

Impulse responses are often used to characterize dynamic systems. When the analyst wants to remain agnostic about the data generating process (DGP), impulse responses are often estimated using the method of local projections proposed by Jordà (2005) or LPs for short. LPs consist of projecting future outcome variables on current and past information on the intervention (or impulse), the outcome and other exogenous variables. In other words, they are simple regressions with a particular dynamic error structure.

This article discusses how to conduct inference for LPs with an emphasis on general principles. Relative to vector autoregressions (VARs), LPs generally have lower bias but higher variance though in infinite samples, when the lag structure grows with the sample, both methods generate the same impulse response and similar parameter estimation uncertainty (see, e.g. Plagborg-Møller & Wolf, 2021; Xu, 2023). We organize our discussion around three main topics: (1) point-wise inference; (2) simultaneous inference; and (3) significance.

Point-wise inference, by far the most popular, refers to the uncertainty with which individual parameters of the impulse response are estimated. It is presented graphically by means of error bands. Simultaneous inference refers to the uncertainty about sets of impulse response coefficients. It can be presented graphically as error bounds that accommodate families of hypotheses. Because these bounds are generally wider than error bands, practitioners do not generally show them. Both point-wise and simultaneous inference are based on the Wald principle.

In this paper, we introduce a new concept: *significance bands*. These bands refer to the null hypothesis that the impulse does not generate a response, or in the parlance of experiments, that the policy intervention has no effect. The Lagrange-Multiplier (LM) principle considerably simplifies the construction of significance bands, which are a useful complement to the practice of presenting error bands since many researchers are often concerned about the absence of a response to an impulse.

We set the stage with a brief introduction to LPs to highlight where the properties of the estimated response come from and how they affect inference. Next, we discuss point-wise inference. We highlight several recent developments in the literature, including feasible generalized least-squares methods (LP-FGLS), lag-augmentation (LP-LA), and with a brief mention of Bayesian inference (BLP).

Thinking about simultaneous inference requires system estimation of LPs. Thus we begin by framing such system estimation within the class of generalized method of moments (GMM) estimators. This allows us to cover instrumental variable estimation. With this scaffolding in place, we review two important papers in this arena, Jordà (2009) and Montiel-Olea & Plagborg-Møller (2019).

The review of the literature up to this stage sets us up to introduce the novel concept of significance bands. We will show that the LM principle provides a much simpler approach to think

about the absence of a response and to present the information in a convenient graphical way. In addition, we also introduce a straightforward wild block-bootstrap procedure that is simple to implement.

By now, there is a sprawling literature that extends LPs into a variety of new areas. Though we cannot cover all these new developments, we will briefly touch on methods to smooth LPs in the context of producing more efficient responses. We will also discuss applications with panel data and the inferential challenges that these data introduce before concluding.

2. Local projections: an introduction

There are many situations in the study of dynamic systems where the analyst is interested in the following statistic:

$$\mathcal{R}_{sy|x}(h, s_0, \delta) \equiv \mathbb{E}(y_{t+h}|s_t = s_0 + \delta; \boldsymbol{x}_t) - \mathbb{E}(y_{t+h}|s_t = s_0, \boldsymbol{x}_t) \quad \text{for } h = 0, 1, \ldots, H$$

where s_t is the variable that generates the impulse from $s_t = s_0$ to $s_t = s_0 + \delta$ for some initial value s_0. Denote y_t as the outcome variable of interest, whose response to an impulse in s_t we want to characterize. Finally, the vector of variables $\boldsymbol{x}_t$ contains lags of the outcome, lags of the impulse, and lags of any other exogenous or predetermined variables, including the constant and deterministic time trends.

We will refer to s_t interchangeably as the *treatment*, *intervention*, or *shock*. At this point, we assume that the shift from s_0 to $s_0 + \delta$ is exogenously determined. In practice, identification assumptions and methods will be required. However, since the emphasis here is on inference, we will often prefer to keep things simple. The conditional mean function, $\mathbb{E}(\cdot|\cdot)$ can in principle be nonlinear, a case considered by Angrist & Kuersteiner (2011) and Angrist *et al.* (2016). We will mostly focus on linear settings, but it is important to maintain the distinction for now.

We use the notation $\mathcal{R}_{s\to y|\boldsymbol{x}}(h, s_0, \delta)$ instead of $\mathcal{R}_{sy|\boldsymbol{x}}(h, s_0, \delta)$ when additional identifying assumptions, such as exogeneity of s_t justify a causal interpretation of the difference in conditional means. The notation $\mathcal{R}_{s\to y|\boldsymbol{x}}(h, s_0, \delta)$ is meant to convey the direction of the intervention s onto the outcome y, conditional on $\boldsymbol{x}$. When it is clear, we will simplify the notation to $\mathcal{R}_{sy}(h)$ to denote the response of y to an impulse in s, h periods in the future.

For example, when $s_t \in \{0, 1\}$ is randomly assigned, the conditional expectation $E(y_{t+h}|s_t, \boldsymbol{x}_t)$ does not depend on $\boldsymbol{x}_t$ and only takes two values. Hence, a natural estimate of the response is simply:

$$\mathcal{R}_{s\to y|\boldsymbol{x}}(h, s_0, \delta) = \frac{\sum_{t=h}^{T} y_{t+h} s_t}{\sum_{t=h}^{T} s_t} - \frac{\sum_{t=h}^{T} y_{t+h}(1 - s_t)}{\sum_{t=h}^{T}(1 - s_t)}; \quad h = 0, 1, \ldots, H.$$

For $h = 0$, one can think of this difference in means as a measure of the average treatment effect in a randomized controlled trial or RCT. The same statistic can be estimated in regression form:

$$y_{t+h} = \mu_h + \beta_h s_t + v_{t+h}, \tag{1}$$

where $\mu_h = \mathbb{E}(y_{t+h}|s_t = 0)$ and $\beta_h = \mathbb{E}(y_{t+h}|s_t = 1) - \mathbb{E}(y_{t+h}|s_t = 0) = \mathcal{R}_{sy}(h)$.

Equation 1 can be easily generalized to allow for $s_t \in \mathbb{R}$ and to include a vector of controls $\boldsymbol{x}_t$ which, even if s_t is randomly assigned, will improve efficiency. Using a linear framework, we arrive at the typical expression of a local projection:

$$y_{t+h} = \mu_h + \beta_h s_t + \boldsymbol{\gamma}_h \boldsymbol{x}_t + v_{t+h}. \tag{2}$$

If s_t is not purely randomly assigned but depends on $\boldsymbol{x}_t$, then including $\boldsymbol{x}_t$ may reduce bias while the effect on the standard errors is ambiguous. The reason is that $\boldsymbol{x}_t$ could be correlated with s_t, but have no direct effect on y_{t+h}. In that case, including $\boldsymbol{x}_t$ in the regression reduces the variance of s_t without reducing the variance of the regression residual. However, if one is willing to assume that the assignment of s_t is as good as random only given $\boldsymbol{x}_t$, then one can appeal to the conditional independence assumptions carefully spelled out in, e.g., Angrist & Kuersteiner (2011) and Angrist *et al.* (2016). In this scenario, $\boldsymbol{x}_t$ needs to be included in the regression.

If s_t is not randomly assigned even conditional on $\boldsymbol{x}_t$, but there is an instrument or instruments for it, call them $\boldsymbol{z}_t$, then clearly Equation 2 can be estimated using instrumental variables and we will be more specific later on about the relevance and exogeneity assumptions needed in this case.

Another obvious extension is to allow for the relationship between y_{t+h} and the right-hand side variables of this expression to be possibly nonlinear, in which case the values of s_0 and $\boldsymbol{x}_t$ will be determinative of the actual response estimated. Hence, consider the following general additive conditional mean specification:

$$y_{t+h} = \mu_h(s_t, \boldsymbol{x}_t; s_0, \delta; \boldsymbol{\theta}_h) + v_{t+h}; \quad h = 0, 1, \ldots, H, \tag{3}$$

where s_0 is the initial state for s_t, δ is the impulse from that initial state, and $\boldsymbol{\theta}$ is the parameter vector. All other components have been previously defined.

Based on Equation 3, then:

$$\mathcal{R}_{sy}(s_0, \delta, \boldsymbol{x}_t, h) = \mu_h(s_t = s_0 + \delta, \boldsymbol{x}_t; \boldsymbol{\theta}_h) - \mu_h(s_t = s_0, \boldsymbol{x}_t; \boldsymbol{\theta}_h); \quad h = 0, 1, \ldots, H,$$

where the notation makes explicit the dependence of the response on s_0, δ, and $\boldsymbol{x}_t$. For example, consider a simple nonlinear LP (though linear in parameters):

$$y_{t+h} = \alpha_h + \beta_h s_t + \gamma_h x_t + \phi_h s_t^2 x_t + v_{t+h} \tag{4}$$

then clearly $\mu_h(s_t = s_0 + \delta, x_t; \boldsymbol{\theta}_h) = \alpha_h + \beta_h(s_0 + \delta) + \gamma_h x_t + \phi_h(s_0 + \delta)^2 x_t$ whereas $\mu_h(s_t = s_0, x_t; \boldsymbol{\theta}_h) = \alpha_h + \beta_h s_0 + \gamma_h x_t + \phi s_0^2 x_t$ and hence $\mathcal{R}_{sy}(s_0, \delta, x_t, h) = \beta_h \delta + \phi_h(\delta^2 + 2\delta s_0)x_t$, which will depend on s_0, the initial value of the intervention variable, δ the size of the intervention, and x_t, the value of the current state variable. This example highlights the importance of being careful in how nonlinear responses are calculated as it has direct bearing on how inference needs to be constructed.

Assuming instruments are available and they meet relevance and exogeneity conditions that we will make more precise below, then Equation 3 can be estimated by the Generalized Method of Moments (GMM) based on the moment conditions:

$$\mathbb{E}\left[(y_{t+h} - \mu_h(s_t, \boldsymbol{x}_t; \boldsymbol{\theta}_h))\, \boldsymbol{z}_t\right] = 0; \quad h = 0, 1, \ldots, H. \tag{5}$$

Furthermore, as we shall explain below, Equation 5 for each h can be stacked and estimated jointly as a system. Stacking as a system will prove useful when estimates of the covariance matrix of responses across all horizons is required for simultaneous inference or when it is required for nonlinear models such as the one in Equation 4.

2.1. Properties of the residual

Inference based on Equation 1 obviously depends on the properties of the residual, v_{t+h}. We present the main ideas of what differentiates local projections from traditional regression results in as simple a setup as possible. Hence suppose the data are generated by the following covariance-stationary AR(1) model. Using the companion form, the AR(1) encompasses more general AR(p) models and of course, in vector form, general VAR(p) models so the example, while simple, is helpful in thinking of more complex settings:

$$y_t = m + \rho y_{t-1} + u_t; \qquad |\rho| < 1; \tag{6}$$

where u_t is a zero-mean white noise process with $\mathbb{E}u_t^2 = \sigma_u^2$. All of these assumptions can be relaxed and made much more general but they suffice to make the point.

The goal is to calculate the response of y_{t+h} to a shock u_t. Iterating forward h periods into the future on the previous expression, we arrive at a common expression for a local projection, along the lines of that in Equation 2 where the shock of interest is now u_t instead of some other exogenous variable. Hence we have:

$$y_{t+h} = \mu_h + \beta_h y_t + v_{t+h}, \tag{7}$$

where $\mu_h = m(1 + \rho + \ldots + \rho^{h-1})$; $\beta_h = \rho^h$, and $v_{t+h} = u_{t+h} + \rho u_{t+h-1} + \ldots + \rho^{h-1} u_{t+1}$. Given our assumptions, Equation 7 can be estimated by ordinary least-squares (OLS) and thus, a consistent estimate of $\mathcal{R}_{uy}(h)$ is $\hat{\beta}_h$.

However, notice that the residuals have an MA(h) structure that affects the computation of the standard error $\hat{\beta}_h$ since:

$$\hat{\beta}_h = \beta_h + \frac{\frac{1}{T-h}\sum_{t=2}^{T-h} y_t v_{t+h}}{\frac{1}{T-h}\sum_{t=2}^{T-h} y_t^2}$$

from where

$$(T-h)^{1/2}(\hat{\beta}_h - \beta_h) = \frac{\frac{1}{(T-h)^{1/2}}\sum_{t=2}^{T-h} y_t v_{t+h}}{\frac{1}{T-h}\sum_{t=2}^{T-h} y_t^2}.$$

Given our assumptions on u_t it is easy to see that $\frac{1}{T-h}\sum_{t=2}^{T-h} y_t^2 \xrightarrow{p} \frac{1}{(1-\rho^2)}\sigma_u^2$, whereas the numerator will converge in distribution to a Normal random variable whose variance, ω^2 is:

$$\omega^2 = Var\left(\frac{1}{(T-h)^{1/2}}\sum_{t=2}^{T-h} y_t v_{t+h}\right) \approx \sum_{j=-\infty}^{\infty} \mathbb{E}(y_t v_{t+h} y_{t-j} v_{t+h-j}),$$

where $v_{t+h} = u_{t+h} + \rho u_{t+h-1} + \ldots + \rho^{h-1} u_{t+1}$ and $y_t = u_t + \rho u_{t-1} + \ldots$. Putting the pieces back together we arrive at:

$$(T-h)^{1/2}\frac{\sigma_u^2}{\omega(1-\rho^2)}(\hat{\beta}_h - \beta_h) \xrightarrow{d} N(0,1). \tag{8}$$

These derivations show that approximating ω in finite samples will require a heteroscedasticity and autocorrelation consistent estimator. Jordà (2005) originally proposed using the Newey-West estimator as a simple solution. Since then, several developments that we discuss next have provided more elegant solutions.

3. LP Inference: The issues

In thinking about inference, it will be important to clearly outline the objective of the inferential procedure to design the best approach. In this regard at least three obvious objectives come to mind:

1. **Pointwise inference:** How should one assess the precision of individual estimated response coefficients and the value that they attain?

2. **Simultaneous inference:** How should one assess subsets of response coefficients or the trajectory of the response as a whole?

3. **Significance:** What is the best way to test the null that the intervention has no effect on the outcome?

Layered on top of these three possible objectives, the analyst should consider the properties of alternative procedures available. Plagborg-Møller & Wolf (2021) and Xu (2023) show that VARs

and LPs are asymptotically equivalent when the data are generated by a VAR(∞) as long as the lag length is allowed to grow to infinity with the sample size. In finite samples and under the same VAR(∞) assumption, Jordà *et al.* (2024) show that, whereas the truncation lag used to estimate the VAR, say p, ensures consistency of the first p lags of the VAR, it does not ensure consistency of the impulse response beyond the p^{th} horizon whereas this is not the case for LPs: LPs remains consistent. Moreover, recent work by Plagborg-Møller *et al.* (2024) shows that LPs are even robust to misspecification of the truncation lag.

Further, the bias-efficiency trade-off between VARs and LPs, determine the probability coverage properties of each approach. Going back to Plagborg-Møller *et al.* (2024), these authors show that LP inference is robust to relatively large amounts of misspecification even when compared to VARs that are only mildly misspecified.

How did they arrive at this conclusion? Plagborg-Møller *et al.* (2024) use a setting where they assume that the data are generated by a VAR where the residuals follow a moving-average structure that vanishes as the sample grows. In large samples, the VAR is clearly correctly specified. However, VARs undercover even when misspecification is so small, that it would be difficult or impossible to detect in small samples. On the other hand, they show that LPs have the correct coverage even when misspecification is so large that it can be easily detected in a finite sample.

Finally, a common concern is not just to derive procedures that are valid in large samples, but to construct methods that account for small sample properties of the data that may be far from the large sample ideal. These will usually include simulation-based inference, such as the bootstrap, as we shall discuss. Bayesian inference is also possible. Although LPs are not a generative model with which to construct the likelihood, Tanaka (2020) and Ferreira *et al.* (2023), for example, provide Bayesian estimation methods of LPs. LPs can also be used to shrink large dimensional Bayesian VARs toward the responses generated with LPs as Miranda-Agrippino & Ricco (2017) show.

4. Pointwise Inference

Section 2 and Equation 8 more specifically, provide intuition for why, generally speaking, the residuals of the LP will have a moving average structure and how this might affect inference as a result. Jordà (2005) proposed using a heteroscedasticity and autocorrelation consistent (HAC) estimator such as Newey & West (1987). Much of the literature appears to follow a similar strategy, even when it comes to panel data, where the Driscoll-Kraay (Driscoll & Kraay, 1998) covariance estimator is used instead.[1] However, whether the Driscoll-Kraay estimator is the right approach depends on the dimensions of the panel. Implicitly, the assumption is that $T >> N$, where T is the time dimension of the panel and N is the cross-section dimension. We reserve a more thorough discussion of panel data inference to a later section. Instead, we begin by discussing the recent LP inferential procedures introduced for time series data in the literature.

[1]The Driscoll-Kraay estimator can be seen as the extension to panel data of the Newey-West estimator.

4.1. LP-FGLS

Lusompa (2023) proposes a parametric feasible GLS procedure that directly accounts for the specific MA structure of the residuals. This LP-FGLS procedure can be best explained with the simple AR(1) example in Equation 6. The idea is to use the residuals from the first local projection, $\hat{u}_t$ as follows.

LP-GLS algorithm

- For $h = 1$ estimate:

$$y_t = m_1 + \beta_1 y_{t-1} + u_t \quad \rightarrow \quad \hat{\beta}_1,\ \hat{u}_t$$

- For $h = 2$, construct $\tilde{y}_{t+1} = y_{t+1} - \hat{\beta}_1 \hat{u}_t$ and hence estimate:

$$\tilde{y}_{t+1} = m_2 + \beta_2 y_t + v_{t+1} \quad \rightarrow \quad \hat{\beta}_2$$

- For $h = 3$, construct $\tilde{y}_{t+2} = y_{t+2} - \hat{\beta}_1 \hat{u}_{t+1} - \hat{\beta}_2 \hat{u}_t$ and hence estimate:

$$\tilde{y}_{t+2} = m_3 + \beta_3 y_t + v_{t+2} \quad \rightarrow \quad \hat{\beta}_3$$

- For $h > 3$ sequentially generate $\tilde{y}_{t+h} = y_{t+h} - \hat{\beta}_1 \hat{u}_{t+h-1} - \hat{\beta}_2 \hat{u}_{t+h-2} - \ldots - \hat{\beta}_h \hat{u}_t$ and regress:

$$\tilde{y}_{t+h} = m_h + \beta_{h+1} y_t + v_{t+h}$$

The residuals $\hat{v}_{t+h}$ will be approximately white noise and hence heteroscedasticity robust inference is sufficient.

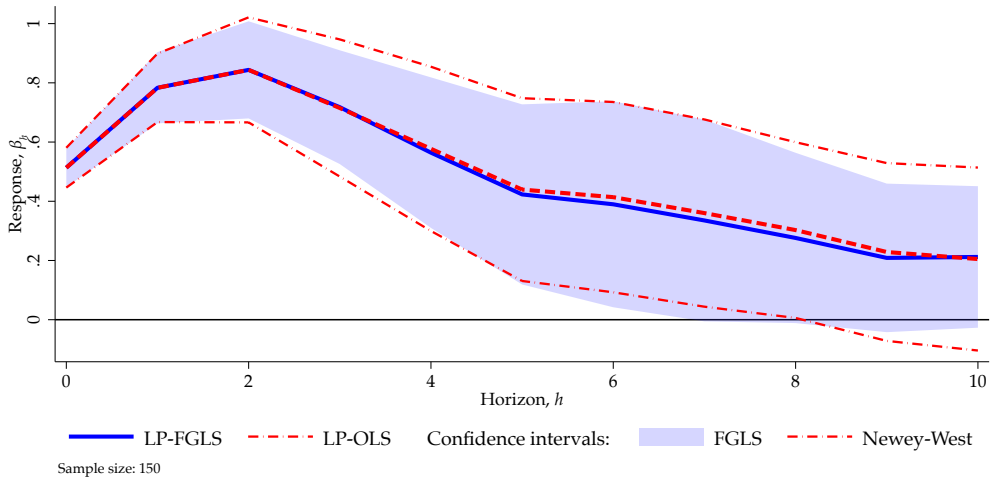
In a related paper, Breitung & Brüggemann (2023) propose a similar approach that consists of using the transformation $\tilde{y}_{t+h} = y_{t+h} - \hat{u}_{t+h-1}$ and use $\{\hat{u}_t, \ldots, \hat{u}_{t+h-2}\}$ as additional regressors. Breitung & Brüggemann (2023) show that this correction is as efficient as responses estimated with the correct vector autoregression in finite samples. Lusompa (2023) also provides bootstrap versions of the LP-FGLS procedure, which have good efficiency gains relative to standard procedures.

As an illustration, Figure 1 compares estimates of an impulse response using the usual Newey-West correction with estimates based on this LP-GLS procedure. The responses are estimated using simulated data from a simple bivariate model given by:

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} 0.7 & 0.4 \\ 0.4 & 0.7 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} u_t^y \\ u_t^x \end{pmatrix}; \qquad u_t^y = e_t^y + e_t^x, \quad u_t^x = e_t^x; \quad e_t^y, e_t^x \overset{iid}{\sim} N(0,1), \tag{9}$$

with a sample of 150 observations (after disregarding 1,000 initial observations). Figure 1 shows the response of x to a shock to u^x and illustrates that both methods generate almost identical response

Figure 1: *Comparing Newey-West versus FGLS error bands*



Notes: Data generated from a bivariate VAR(1). The simulated sample size is 150 observations after disregarding 1,000 initialization observations. LP-OLS response (dashed line) with Newey-West (dot dashed lines) versus LP-FGLS (solid line) with FGLS (shaded region) error bands at 95% confidence level. See text.

estimates and very similar error bands with a slight efficiency edge for LP-FGLS. This is expected since the DGP fits quite well with the theoretical background.

4.2. Lag-augmentation of LPs

Montiel Olea & Plagborg-Møller (2021) introduce the idea of using lag-augmentation to conduct robust inference with LPs. They show that this procedure is uniformly valid over both stationary and non-stationary data and over a wide range of response horizons. Going back to the stylized model AR(1) in equation Equation 6 with $m = 0$, assume u_t is strictly stationary and further assume $E(u_t|\{u_s\}_{s\neq t}) = 0$ *almost surely*. We make these assumptions to follow the setup in Montiel Olea & Plagborg-Møller (2021). Using similar notation to that in their paper, let $\beta(\rho, h)$ denote the LP parameter used to estimate the impulse response ρ^h, that is

$$y_{t+h} = \beta(\rho, h)\, y_t + \xi_t(\rho, h); \qquad \xi_t(\rho, h) \equiv \sum_{l=1}^{h} \rho^{h-l}\, u_{t+l}\,. \tag{10}$$

Next, Montiel Olea & Plagborg-Møller (2021) suggest adding y_{t-1} as an additional regressor to Equation 10. The purpose of this *lag augmentation* is to make the effective regressor of interest stationary even if the data y_t has a unit root. Montiel Olea & Plagborg-Møller (2021) show that, rearranging, the lag-augmented local projection can be written as

$$y_{t+h} = \beta(\rho, h)\, u_t + \beta(\rho, h+1) y_{t-1} + \xi_t(\rho, h)\,. \tag{11}$$

Although u_t is stationary and therefore would sidestep distortions to the normal distribution caused by near-to-unity asymptotics, it is not directly observed. However, due to the linear relationship between y_t and u_t, the feasible local projection onto (y_t, y_{t-1}) provides an estimate of $\beta(\rho, h)$ precisely equal to the one that would be obtained from the projection onto (u_t, y_{t-1}).

Thus, the actual regression to be estimated is:

$$y_{t+h} = \beta(h)y_t + \beta(h+1)y_{t-1} + \xi_t(h) \quad \rightarrow \quad \hat{\beta}(h), \hat{\xi}_t(h)\,. \tag{12}$$

Lag-augmentation has two benefits. As Montiel Olea & Plagborg-Møller (2021) show, the distribution of $\hat{\beta}(h)$ of this feasible lag-augmented local projection is uniformly normal in $\rho \in [-1, 1]$ using similar arguments to those used of lag-augmentation in AR inference (see, e.g., Sims *et al.*, 1990; Toda & Yamamoto, 1995; Dolado & Lütkepohl, 1996; Inoue & Kilian, 2002, 2020). The second benefit is that it simplifies the computation of standard errors (though now the convergence rate will be $T^{1/2}$ instead of T).

In particular, it is sufficient to use a heteroskedasticity-robust routine to estimate standard errors for $\hat{\beta}(h)$, like the usual White correction (in STATA, `reg` with the option `robust` or even better, `hc3`). How can we magically dispense with the moving average structure of the residuals evident in Equation 10? From Equation 11, note that u_t was assumed to be uncorrelated with past and future values of itself, and therefore the regression *score* $\xi_t(\rho, h)u_t$ is serially uncorrelated. To see this, note that the standard error formula in the ideal regression of Equation 11 would be

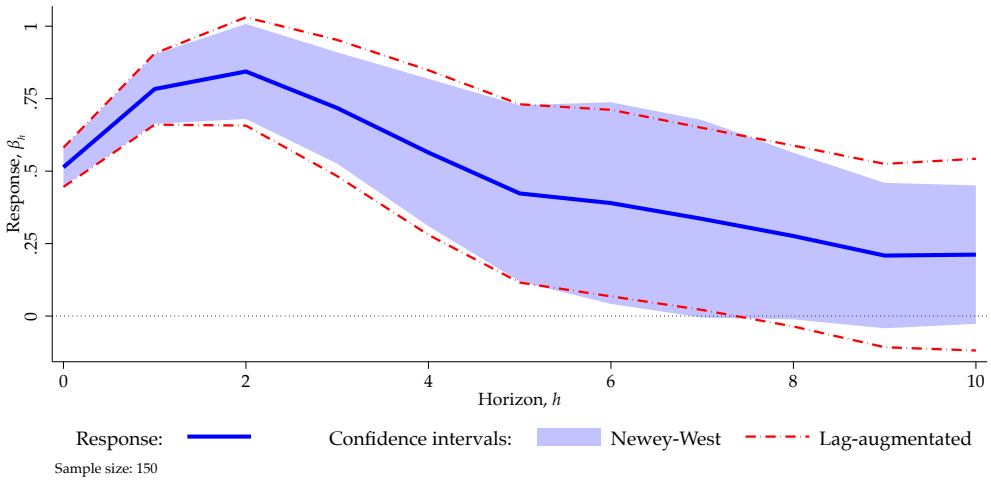
$$\hat{\mathbf{s}}_h = \frac{(\sum_{t=1}^{T-h} \hat{\xi}_t(\rho, h)^2 \hat{u}_t^2)^{1/2}}{\sum_{t=1}^{T-h} \hat{u}_t^2}\,.$$

But by similar linearity arguments used to justify the feasible augmented local projection, it can be calculated directly from Equation 12 using White corrected standard errors as indicated.

Several remarks deserve mention. First, Montiel Olea & Plagborg-Møller (2021) show that lag-augmented LP inference is relatively robust to persistent data and provides appropriate coverage even at relatively long horizons (as long as $h_T/T \to 0$). Second, lag-augmentation is shown to work more generally when the DGP is assumed to be a VAR(p) or a vector error correction model (VECM), though we are not aware that similar results have been derived for panel data in settings where the time dimension is larger than the cross section dimension, i.e., $T \gg N$. Of course, when $N \gg T$, asymptotic results are driven by the cross-sectional dimension of the panel and then the asymptotic distribution is normal even when the data are persistent. Third, lag-augmentation can also be applied to identified LPs (Plagborg-Møller & Wolf, 2021; Montiel Olea & Plagborg-Møller, 2021).

As an illustration of how Newey-West and lag-augmented error bands compare, we revert back to data simulated from the process shown in Equation 9 but this time use LP-OLS estimates with error bands calculated using Newey-West vs lag-augmentation. This is shown in Figure 2. The figure shows that both methods generate similar error bands. In fact, several experiments (not

Figure 2: *Comparing Newey-West versus lag-augmented error bands*



Notes: Data generated from a bivariate VAR(1). The simulated sample size is 150 observations after disregarding 1,000 initialization observations. Response shown with Newey-West (shaded region) versus lag-augmented (dot-dashed line) error bands at 95% confidence level. See text.

reported here) suggest that, for stationary data, the coverage is very similar between methods. Lag-augmented bands tend to be somewhat more conservative the more persistent the data.

Finally, Montiel Olea & Plagborg-Møller (2021) provide bootstrap procedures that we briefly sketch here though the reader should go to the original source for details. Suppose that one wants to provide inference for an impulse response estimated with lag-augmented LPs for which one can also obtain the standard error as described earlier (i.e., using White corrected standard errors). Montiel Olea & Plagborg-Møller (2021) then suggest estimating the corresponding VAR(p).[2] This VAR will serve two purposes. One is to construct the equivalent response to that estimated with LPs, whose difference is then used to construct the t-ratio using the LP standard error. The second is to generate bootstrap replicates of the data using a parametric wild bootstrap (see, e.g., Gonçalves & Kilian, 2004) based on the VAR(p). Using these bootstrap replicates, then one estimates the lag-augmented LP responses and their standard errors. These are the ingredients necessary to then construct a percentile-t confidence interval as usual.

5. Joint inference

Error bands, as they are overwhelmingly presented in the literature, are the result of inverting the t-statistics of individual hypotheses, as we have seen. When the interest turns to assessing the overall trajectory of the impulse response, the question becomes one of simultaneous inference. Testing joint hypotheses requires the covariance matrix of the response coefficients. Thus, we begin this section by providing a system Generalized Method of Moments (GMM) setup that will allow us

[2]One can also bias-adjust the VAR coefficients using the correction by Pope (1990).

to obtain this covariance matrix. Moreover, it will also allow us to consider instrumental variable estimation methods.

Although joint hypotheses tests are straightforward to construct (under rather general assumptions), presenting the evidence graphically, as we did with error bands, is not. After introducing the GMM estimator for LPs, we discuss two approaches to achieve this, one based on Scheffé's S-method, and another based on a sup-t test. Both approaches generate more conservative error *bounds* designed to accommodate hypotheses tests on subsets of response coefficients. In practice, these methods have found less echo in the literature precisely because the bounds are wider. However, some of the insights from these procedures are valuable.

5.1. LP-GMM

GMM provides a convenient framework to estimate LPs jointly, as we advanced when discussing Equation 5. Let $\boldsymbol{y}_t(H) = (y_t, \ldots, y_{t+H})'$ be an $(H+1) \times 1$ vector that collects the outcome variable observed at increasingly distant horizons into the future. Let $S_t = I_{H+1} \otimes s_t$ where I_{H+1} is the identity matrix of order $H+1$ and s_t is the intervention variable. We collect the error terms in $\boldsymbol{v}_t(H) = (v_t, \ldots, v_{t+h})'$. The response coefficients are $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_H)'$. Exogenous and predetermined variables collected in $\boldsymbol{x}_t$ could be easily included by defining $X_t = I_{H+1} \otimes \boldsymbol{x}_t$. However, for simplicity, we will mostly set them aside for the presentation. In linear settings, we could simply invoke the Frisch-Waugh-Lovell theorem and orthogonalize the outcome and the intervention with respect to the controls before proceeding. Finally, suppose an $1 \times l$ vector $\boldsymbol{z}_t$ of instrumental variables is available with $l \geq 1$. Hence we construct $Z_t = I_{H+1} \otimes \boldsymbol{z}_t$. If no instruments are available, but s_t is exogenously determined (perhaps conditional on controls), then one can simply set $Z_t = S_t$.

Using these definitions, the population moment conditions of the system of local projections can be expressed as:

$$E[Z_t'(\boldsymbol{y}_t(H) - S_t\boldsymbol{\beta})] = 0.$$

Thus, the corresponding finite sample GMM problem can be written as:

$$\min_{\beta} \left[\frac{1}{N}\sum_{t=t_0}^{T^*} Z_t'(\boldsymbol{y}_t(H) - S_t\boldsymbol{\beta})\right]' \hat{\Lambda}^{-1} \left[\frac{1}{N}\sum_{t=t_0}^{T^*} Z_t'(\boldsymbol{y}_t(H) - S_t\boldsymbol{\beta})\right],$$

with $N = T^* - t_0$, where t_0 denotes the first observation available after accounting for possible lags in the control set and $T^* = T - H - 1$. One could choose to set $\hat{\Lambda} = I_{l\times(H+1)}$. The estimator, referred to as the equally weighted estimator, yields consistent estimates of $\boldsymbol{\beta}$, but the covariance matrix is not valid for inference, as is well-known.

The optimal weighting matrix, correcting for heteroscedasticity and autocorrelation with, for example, a Barlett correction, is:

$$\hat{\Lambda} = \hat{\Gamma}_0 + \sum_{j=1}^{J} K(j)(\hat{\Gamma}_j + \hat{\Gamma}_j'); \quad K(j) = \left[1 - \frac{j}{J+1}\right]; \quad \hat{\Gamma}_j = \frac{1}{N}\sum_{t_0}^{N} Z_t' \hat{\boldsymbol{v}}_t(H) \hat{\boldsymbol{v}}_{t-j}'(H) Z_{t-j}.$$

where $\tilde{\boldsymbol{v}}_t(H)$ refers to the residuals based on the equally weighted estimator, which we know to be consistent.

Stock & Watson (2018) and Plagborg-Møller & Wolf (2022) provide appropriate relevance and exogeneity conditions (under relatively general assumptions for the underlying DGP) for the analysis of LPs. In our model, these can be stated as:

- ***Relevance:*** $E(\boldsymbol{z}_t' s_t | \boldsymbol{x}_t) \neq \mathbf{0}$
- ***Exogenenity:*** $E(Z_t' \boldsymbol{v}_t(H) | X_t) = \mathbf{0}$

where we explicitly include the controls in the conditioning set to make clearer their influence (despite having proceeded by projecting their influence away in a first stage or simply when they are subsumed in the vector of instruments, $\boldsymbol{z}_t$). Our exogeneity condition is satisfied under the contemporaneous and lead-lag exogeneity conditions of Stock & Watson (2018).

Based on the relevance and exogeneity assumptions just stated and standard results from the algebra of GMM, estimates of the impulse response can be obtained from:

$$\hat{\mathcal{R}}_{sy} = \hat{\boldsymbol{\beta}} = \left(\frac{1}{N}\sum_{t_0}^{T^*} S_t' Z_t \hat{\Lambda}^{-1} Z_t' S_t\right)^{-1} \left(\frac{1}{N}\sum_{t_0}^{T^*} S_t' Z_t \hat{\Lambda}^{-1} Z_t' \boldsymbol{y}_t(H)\right),$$

which will be consistent and asymptotically normally distributed with approximate covariance matrix given by:

$$\hat{\Omega}_\beta = \left[\left(\frac{1}{N}\sum_{t_0}^{T^*} S_t' Z_t\right) \hat{\Lambda}^{-1} \left(\frac{1}{N}\sum_{t_0}^{T^*} Z_t' S_t\right)\right]^{-1}. \tag{13}$$

Joint estimation of the impulse response using this GMM set-up is useful as it provides an estimate of the covariance matrix, $\hat{\Omega}_\beta$, a key ingredient to construct any joint hypothesis test of interest. It will also be an important ingredient in the construction of error bands that accommodate simultaneous inference, as the next section briefly explains.

5.2. Simultaneous inference

Analysts interested in testing features of the impulse response involving more than one parameter can set up the appropriate hypothesis test as usual using the F-test, for example, and report the results of the test in a table. However, how should one represent simultaneous inference

graphically if one were interested in representing bounds with appropriate probability coverage that accommodate a variety of hypothesis tests an analyst may conduct? Impulse response coefficients are correlated. Thus, the usual practice of inverting the t-ratio to display error bands will not provide the correct probability coverage. The correct approach requires that we construct error bands that account for the simultaneous nature of the family of hypotheses of interest.

This problem was highlighted by Jordà (2009). His solution relied on Scheffé's multiple comparison approximation or S-method (Scheffé, 1953). Asymptotically, the sum of the squares of the t-ratios of the LP is approximately χ^2_H distributed. Thus, the critical value with which to construct simultaneous error bands is $(c^2_\alpha/H)^{1/2}$ where c^2_α refers to the critical value of a χ^2_H. The advantage of this method is that it does not require simulation methods to obtain the critical value.

More recently, Montiel-Olea & Plagborg-Møller (2019) proposed a more efficient approximation based on the sup-t procedure. Although this method requires simulation methods, it is shown to produce the narrowest bands for a given probability coverage. In particular, let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_H)$, and assume $\boldsymbol{\beta} \xrightarrow{d} N(0, \Omega_\beta)$. The goal is to find c_α such that the error bands defined by:

$$\hat{\mathcal{B}}(c_\alpha) = [\hat{\beta}_0 - c_\alpha \hat{\Omega}_{0,0}, \hat{\beta}_0 + c_\alpha \hat{\Omega}_{0,0}] \times \ldots \times [\hat{\beta}_H - c_\alpha \hat{\Omega}_{H,H}, \hat{\beta}_H + c_\alpha \hat{\Omega}_{H,H}]$$

such that:

$$P(\boldsymbol{\beta} \in \hat{\mathcal{B}}(c)) \geq 1 - \alpha$$

Montiel-Olea & Plagborg-Møller (2019) show that:

$$P(\boldsymbol{\beta} \in \hat{\mathcal{B}}(c)) \to P\left(\max_{h=0,1,\ldots,H} \left| \Omega^{-1/2}_{h,h} V_h \right| \leq c_\alpha \right); \quad \boldsymbol{V} = (V_0, V_1, \ldots, V_H)' \sim N(0, \Omega_\beta).$$

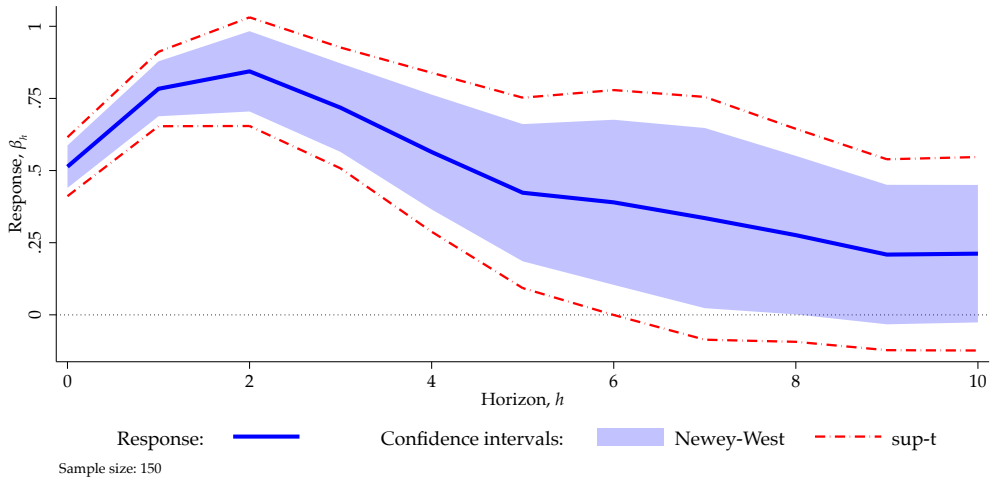
Although the distribution of $\max_{h=0,1,\ldots,H} \left| \Omega^{-1}_{h,h} V_h \right|$ is unknown, it is relatively easy to obtain any desired quantile of this distribution. Thus, error bands for simultaneous inference can be constructed according to the following algorithm:

Plug-in sup-t algorithm

- **Step 1:** Draw M i.i.d. normal vectors $\hat{V}^{(m)} \sim N_H(\mathbf{0}_H, \hat{\Omega}_\beta)$, $m = 1, \ldots, M$.
- **Step 2:** Define $\hat{q}_{1-\alpha}$, the empirical $1-\alpha$ quantile of $\max_{h=0,1,\ldots,H} \left| \Omega^{-1/2}_{h,h} V_h \right|$ across $m = 1, \ldots, M$.
- **Step 3:** Construct the error bands for each h as: $[\hat{\beta}_h - \hat{q}_{1-\alpha} \hat{\Omega}_{h,h}, \hat{\beta}_h + \hat{q}_{1-\alpha} \hat{\Omega}_{h,h}]$; $h = 0, 1, \ldots, H$

Montiel-Olea & Plagborg-Møller (2019) further provide bootstrap and Bayesian versions of this procedure. The interested reader should consult their paper for more details.

Figure 3: *Comparing Newey-West versus sup-t error bands*



Notes: Data generated from a bivariate VAR(1). The simulated sample size is 150 observations after disregarding 1,000 initialization observations. Response shown with Newey-West (shaded region) versus sup-t (dashed-dotted lines) error bands at 95% confidence level. See text.

To get a sense of how much wider the sup-t bands are, we revisit the example of Figure 1 and Figure 2. We simulate data from the same model and then construct 95% sup-t bands using joint estimation of all responses simultaneously to obtain the resulting covariance matrix. In Figure 3, the sup-t bands are shown as dashed blue lines. Relative to the usual Newey-West bands, shown as a shaded region in red, it is easy to see that they are wider but not by a very large amount.

6. Significance

In a randomized controlled trial (RCT), a common hypothesis of interest is to test the null that the treatment is ineffective. Similarly, an impulse response can be thought of as the response to a treatment observed over time and therefore, a common hypothesis of interest will be to assess whether the response is different from zero, just as in the RCT example. It turns out that under the null, the distribution of the appropriate hypothesis test can simplify the construction of *significance* bands considerably.

To understand the basic issues, we use a simple example where the controls are set aside to keep the notation simple. For example, we may presume that the outcome y_{t+h}, the intervention, s_t, and the instrument z_t have been previously orthogonalized with respect to the controls $\boldsymbol{x}_t$ based on the Frisch-Waugh-Lovell theorem. Thus, consider the following local projections estimated by instrumental variables regression:

$$y_{t+h} = s_t\beta_h + v_{t+h} \quad \text{for } h = 0, 1, \ldots, H-1; \quad t = 1, \ldots, T$$

We assume that z_t, meets the conditions for relevance and exogeneity, as well as an exclusion restriction. Then we postulate that:

- ***Relevance:*** $E(s_t\, z_t) \neq 0$.
- ***Exogeneity:*** $E(v_{t+h}\, z_t) = 0 \;\; \forall h = 0, 1, \ldots, H-1$.

Depending on the setting, z_t may include s_t itself, such as when s_t is an observable shock, and then the discussion returns to a more traditional OLS setting. Or if s_t is, conditional on x_t, sequentially exogenous. This would be the case in a recursive identification scheme. We further assume that y_t, s_t, and z_t are covariance stationary. This assumption is not necessary to ensure consistency of the local projection, but will make deriving our inferential procedures and the presentation in this section straightforward.

Based on this simple set up, the instrumental variable estimator for β_h can be written as:

$$\sqrt{T-h}(\hat{\beta}_h - \beta_h) = \frac{(T-h)^{-1/2} \sum_1^n z_t\, y_{t+h}}{(T-h)^{-1} \sum_1^n z_t\, s_t} \quad \text{for } h = 0, 1, \ldots, H-1, \tag{14}$$

where we note that we will evaluate the statistic under the null $H_0 : \beta_h = 0$. Under standard regularity conditions, and the instrumental variable assumptions for local projections, it is easy to see that:

$$\frac{1}{T-h} \sum_1^n z_t\, s_t \xrightarrow{p} E(z_t\, s_t) \equiv \gamma_{zs}$$

Next, consider the numerator in Equation 14 evaluated at the null $H_0 : \beta_h = 0$:

$$\frac{1}{(T-h)^{1/2}} \sum_1^n z_t\, y_{t+h} \xrightarrow{d} N(0, s^2_{\eta,h})$$

where $s^2_{\eta,h}$ is given by

$$s^2_{\eta,h} = \lim_{T\to\infty} Var\left(\frac{1}{(T-h)^{1/2}} \sum_1^n z_t\, y_{t+h}\right) = \sum_{j=-\infty}^{\infty} E(z_t\, v_{t+h}\, z_{t-j}\, v_{t+h-j})$$

and where the RHS is the typical HAC type variance formula. The equality follows from the null hypothesis that $\beta_h = 0$ for $h = 0, 1, \ldots, H-1$. It then follows that the limiting distribution of $\hat{\beta}_h$ is given by

$$\sqrt{T-h}(\hat{\beta}_h - 0) \xrightarrow{d} N(0, \sigma^2_h); \quad \sigma^2_h = \frac{s^2_{\eta,h}}{\gamma^2_{zs}}; \quad \forall h \tag{15}$$

From Equation 15 it is easy to derive a $1-\alpha$ percent band around the zero null so that:

$$P\left[\zeta_{\alpha/2}\frac{\sigma_h}{\sqrt{T-h}} < \hat{\beta}_h < \zeta_{(1-\alpha/2)}\frac{\sigma_h}{\sqrt{T-h}}\right] = 1-\alpha$$

where $\zeta_{\alpha/2}$ is the critical value of a standard normal variable at $\alpha/2$ and for a standard normal, $\zeta_{1-\alpha/2} = -\zeta_{\alpha/2}$, as is well known.

To construct feasible confidence intervals we need to replace σ_h with an estimate. The LM principle requires that σ_h be estimated using the conventional formula for HAC robust standard errors for the just identified two-stage least squares estimator, but evaluated at $\beta_h = 0$. This is accomplished by estimating $s^2_{\eta,h}$ with the long-run variance of $\eta_{t,h} = z_t y_{t+h}$. The estimator for σ^2_h then is based on

$$\sigma^2_h = \frac{s^2_{\eta,h}}{\gamma^2_{sz}}. \tag{16}$$

When plotting a significance band of an impulse response up to H periods, we are essentially conducting a joint hypothesis test. Intuitively, the more horizons considered, the more likely it is to spuriously reject the null when the null is true in a finite sample. A simple way to address this issue is with a Bonferroni adjustment as proposed in Dunn (1961) so that the significance bands for each $\hat{\beta}_h$ become:

$$\left[\zeta_{\alpha/2(H+1)}\frac{\sigma_h}{\sqrt{T-h}},\ \zeta_{1-\alpha/2(H+1)}\frac{\sigma_h}{\sqrt{T-h}}\right].$$

The joint probability that the estimated impulse response lies within the confidence band is given by:

$$P\left(\bigcap_{h=0}^{H}\left\{\zeta_{\alpha/2(H+1)}\frac{\sigma_h}{\sqrt{T-h}} < \hat{\beta}_h < \zeta_{(1-\alpha/2(H+1))}\frac{\sigma_h}{\sqrt{T-h}}\right\}\right) \geq 1-\alpha$$

where the inequality holds in large samples and when the null hypothesis of a zero response is true. Similarly, the test of the joint hypothesis that all response coefficients are zero rejects when:

$$\hat{\beta}_h \notin \left[\zeta_{\alpha/2(H+1)}\frac{\sigma_h}{\sqrt{T-h}},\ \zeta_{1-\alpha/2(H+1)}\frac{\sigma_h}{\sqrt{T-h}}\right]$$

for at least one h. By the same argument, it follows that the size of such a test is not more than α in large samples.

Under stronger assumptions such as independence between z_t and v_s for all t and s, or homoskedasticity of v_{t+h} such that $E(v_{t+h-j}v_{t+h}|z_t, z_{t-j}) = E(v_{t+h-j}v_{t+h})$ a further simplification of

the expression for $s^2_{\eta,h}$ is possible. We obtain

$$s^2_\eta = \sum_{j=-\infty}^{\infty} E(z_t\, v_{t+h}\, z_{t-j}\, v_{t+h-j}) = \sum_{j=-\infty}^{\infty} E(z_t\, z_{t-j}) E(v_{t+h}\, v_{t+h-j}) \tag{17}$$
$$= \sum_{j=-\infty}^{\infty} \gamma_{z,j}\, \gamma_{y,j}$$

where the equality follows from the independence between z_t and v_{t+h}. We define $\gamma_{z,j}$ and $\gamma_{y,j}$ as the j^{th} autocovariances of z and y respectively. Importantly, note that ω is no longer a function of the horizon h under these additional restrictions. The implication is that under the additional restrictions of homoskedasticity or independence, the significance bands will be constant as a function of the horizon h.

Under these stronger conditions we can write Equation 15 under the null hypothesis as:

$$\sqrt{T-h}(\hat{\beta}_h - 0) \xrightarrow{d} N(0, \sigma^2); \quad \sigma^2 = \frac{\sum_{j=-\infty}^{\infty} \gamma_{z,j}\gamma_{y,j}}{\gamma^2_{zs}} = \frac{s^2_\eta}{\gamma^2_{zs}}; \quad \forall h \tag{18}$$

A simple example provides further intuition and a connection to well-known results. In the special case where $z = s$, and y and s are serially uncorrelated, Equation 17 simplifies even further to:

$$\sigma^2 = \frac{\gamma_{y,0}}{\gamma_{s,0}}$$

Thus, when $y = s = z$ and y is a white noise and hence $\gamma_{y,0} = \gamma_{s,0}$ so that $\sigma^2 = 1$, the local projection estimator is simply an estimator of the autocorrelation function. Hence, applying the same derivations as in Equation 18, it is easy to see that one recovers the well known[3] bands for the autocorrelogram of y. Specifically, focus on $h = 1$ in the special case that y is a white noise but one estimates an AR(1) model:

$$\sqrt{n}(\hat{\rho} - 0) \xrightarrow{d} N(0, 1). \tag{19}$$

This is the well known case where the 95% asymptotic significance bands in a correlogram are calculated as $\pm 1.96 \times 1/\sqrt{n}$ and provides a nice window into our proposed procedures. Importantly, notice that the bands do not depend on the horizon (in fact, they also do not depend on the variance in this special case). Whenever an autocorrelation coefficient exceeds the band, the interpretation is that said coefficient can be deemed to be different from zero. This, of course, means that the hypothesis that the impulse/treatment has no effect on the outcome can be rejected.

[3] Not Barlett corrected.

6.1. Practical implementation

Constructing significance bands in practice based on the results from the previous section is straightforward and can be implemented using standard statistical software. We provide a STATA example to illustrate this point and that corresponds to the figures displayed in the paper. The basic steps can be summarized as follows:

Significance bands using asymptotic approximations

1. Calculate the sample average of the product $s_t\, z_t$. Call this $\hat{\gamma}_{sz}$.
2. Construct the auxiliary variable $\eta_{t,h} = y_{t+h}\, z_t$ and regress $\eta_{t,h}$ on a constant. The Newey-West estimate of the standard error of the intercept coefficient is an estimate of $s_{\hat{\eta},h}$.
3. An estimate of $\sigma/\sqrt{T-h}$, call it $\hat{s}_{\beta_h}$, is therefore:

$$\hat{s}_{\beta_h} = \frac{\hat{s}_{\hat{\eta},h}}{\hat{\gamma}_{sz}}$$

4. Construct the significance bands as:

$$\left[\zeta_{\alpha/2(H+1)}\hat{s}_{\beta_h},\ \zeta_{1-\alpha/2(H+1)}\hat{s}_{\beta_h}\right]$$

A bootstrap procedure is equally easy to construct. Note that we do not take a position on the data generating process (DGP). Therefore, we apply the bootstrap directly to step 2 of the previous construction of the significance band. Because of the time series dependence and the possible existence of heteroscedasticity, we will use a wild-block bootstrap (see, e.g. Gonçalves & Kilian, 2007). The STATA implementation only requires a few lines of code. Thus, the entire procedure can be described as follows:

Significance bands using the Wild-Block Bootstrap

1. Calculate the sample average of $s_t\, z_t$. Call this $\hat{\gamma}_{sz}$.
2. Construct the auxiliary variable $\eta_{t,h} = y_{t+h}\, z_t$ and regress η_{t+h} on a constant. The Wild Block bootstrap estimate of the standard error of the intercept coefficient is an estimate of $s_{\hat{\eta},h}$.
3. An estimate of $\sigma/\sqrt{T-h}$, call it $\hat{s}^b_{\beta_h}$, is therefore:

$$\hat{s}^b_{\beta_h} = \frac{\hat{s}^b_{\hat{\eta},h}}{\hat{\gamma}_{sz}}$$

4. Construct the significance bands as:

$$\left[\zeta_{\alpha/2(H+1)}\hat{s}^b_{\beta_h},\ \zeta_{1-\alpha/2(H+1)}\hat{s}^b_{\beta_h}\right]$$

6.2. Application: the response of shelter inflation to monetary policy

We showcase these procedures with a simple application to shelter inflation. As the COVID-19 pandemic was winding down, inflation measured by the personal consumption expenditures (PCE) index, excluding food and energy (core PCE inflation), peaked at around 5.5% on March 2022. In response, the Federal Reserve raised the federal funds rate and subsequently inflation declined to 2.6% by June 2024 (the last data point as of the writing of this paper). However, inflation has been slow to travel back to the 2% target, in large part, because shelter inflation (which mainly measures rents that tenants face and owners de facto pay themselves) has been slow to decline.

Hence, we evaluate how responsive shelter inflation is to interest rates. We use the series of monetary shocks recently provided by Bauer & Swanson (2023). These shocks are obtained from high frequency financial data after removing information effects. The sample available is January 1988 to December 2019 so as to avoid polluting our estimates with the COVID-19 pandemic. We use as controls 12 lags of shelter Personal Consumption Expenditures (PCE) inflation, the unemployment rate (to control for aggregate conditions in the economy), and lags of the federal funds rate. The left-hand side variable is the long difference of 100 times the log of shelter PCE price index (PCEPI), that is $100(y_{t+h} - y_{t-1})$, where y_t is the log of shelter PCEPI. This means that the response reflects the cumulative percentage change in the level of shelter prices up to h periods since the shock.

Figure 4 displays estimates of this cumulative response along with Newey-West confidence bands (for 1 and 2 standard deviations in width) along with significance bands estimated with both analytic and bootstrap methods. Based on the conventional error bands displayed in the figure, one might be tempted to conclude that shelter inflation does not respond to monetary shocks. The error bands contain zero throughout the 48 periods displayed.

However, note that the response of shelter PCEPI is almost uniformly negative each month over the 4 years displayed. Indeed, an F-test of the null that the response coefficients are jointly zero is estrenuosly rejected (the p-value is 6.18e-78). And in fact, the significance bands displayed show that, except for about two and one half years after the shock, the response is clearly different from zero. Analytic bands are narrower than bootstrap bands for about 30 periods, after which they are slightly wider.

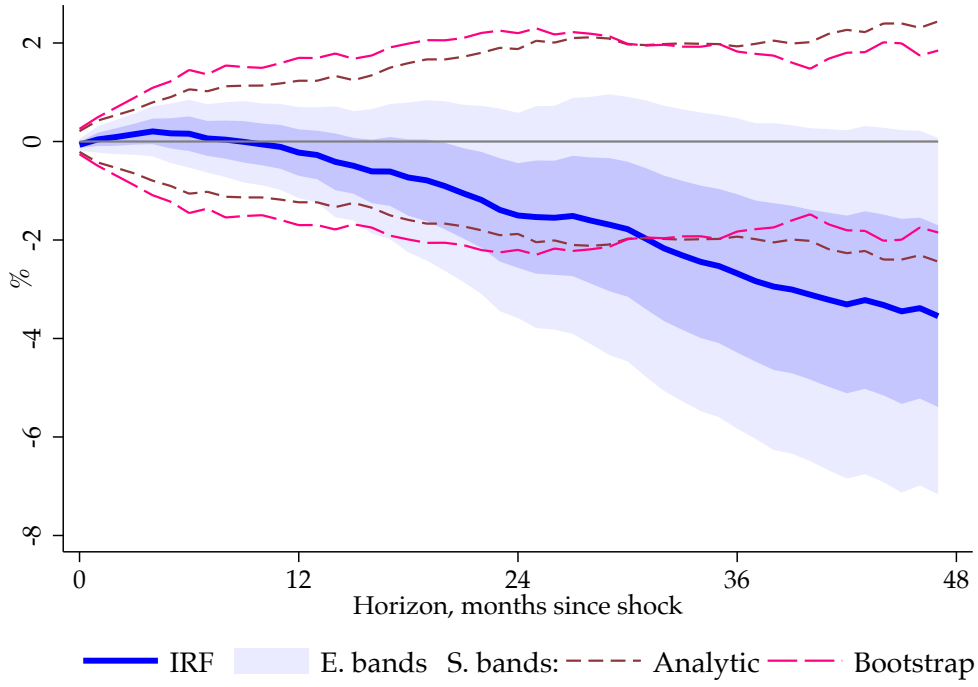
6.3. Monte Carlo evidence

This section presents a couple of simple experiments in graphical form to assess the calculation of significance bands using both the asymptotic approximation and the wild-block bootstrap procedures discussed in the previous section. The data are generated as follows:

$$\begin{cases} y_t = \beta s_t + 0.75 y_{t-1} + u_{yt} \\ s_t = 0.5 s_{t-1} - 0.25 y_{t-1} + z_t + u_{st} \\ z_t = u_{zt} \end{cases} \qquad u_{yt},\, u_{st},\, u_{zt} \sim N(0,1); \quad \beta \in \{0, 0.25, 0.50, 0.75\}$$

Figure 4: *Response of PCE shelter inflation in percent to a Bauer-Swanson monetary shock*

Significance bands displayed



Notes: cumulative response of shelter PCEPI to a Bauer-Swanson monetary shock (Bauer & Swanson, 2023). The specification includes 12 lags of shelter PCE inflation, the unemployment rate, and the federal funds rate. The sample is January 1998 to December 2019. Traditional, point-wise one and two standard error bands displayed as shaded regions using Newey-West standard errors along with analytic (in maroon, short-dash) and bootstrap (in pink long-dash), Bonferroni adjusted significance bands. See text.

This simple system encapsulates several features. First, the treatment variable, s_t, affects the outcome, y_t, contemporaneously. The outcome is itself serially correlated with a coefficient 0.75. The idea is to have internal propagation dynamics. Next, the intervention responds to feedback from the value of the outcome in the previous period, but also has some internal propagation dynamics. In addition, movements in the intervention are caused by the exogenous variable z_t, which will act as our instrumental variable. Finally, the coefficient β, which captures the effect of the treatment on the outcome, has values between 0 and 0.75. When $\beta = 0$ we have the null model with which to assess the size of the test. Increasing the value of β allows us to assess the power of the significance bands.

We generate samples of 100, and 500 observations with 500 burn-in observations that are discarded to avoid initialization problems. For each sample size and for the different values of β we generate 1,000 Monte Carlo replications. The implementation of the wild block-bootstrap is based on 1,000 bootstrap replications as well. For the Newey-West step as well as for the block size in the bootstrap, we use 8 lags. Figure 5 displays the results for sample sizes of 100 and 500 observations.

The figure summarizes quite a bit of information. The shaded bands around the mean estimate of the impulse response showcase the 25^{th} and the 975^{th} largest values for each coefficient estimate in the Monte Carlo simulation. The dashed lines correspond to the significance bands. Both Newey-West and the bootstrap procedures (using 8 lags) generate nearly indistinguishable values so the differences cannot be seen with the naked eye. For each Monte Carlo exercise, we construct rejection rates for each type of band constructed. The rate is calculated as the share of replications where one or more impulse response coefficients exceed the significance bands.

Several results deserve comment. First, consider the size of the test. We have chosen a rather conservative strategy with a window of size 8 both for Newey-West and for the block-size in the implementation of the bootstrap. As a result, with a small sample of 100 observations, the size is about 10% instead of the nominal 5%, though with 500 observations the size is close to 4%.

However, even with this conservative choice, the power of the test is respectable with a sample size of 100, improving from about 25% when $\beta = 0.25$ to about 95% when $\beta = 0.75$. These numbers jump with 500 observations with about 95% for $\beta = 0.25$ and 100% even for $\beta = 0.5$.

7. Other LP settings

Work on extensions to LPs is ongoing. Two settings in particular, have direct implications for how we think about LP inference. LPs can be seen as a semiparametric estimate of the true impulse response and the reason why they have lower bias. Thus, some authors have proposed methods to either smooth the LPs via some low dimensional series approximation (e.g. Barnichon & Brownlees, 2019; Barnichon & Matthes, 2018), or by pairing LPs with Bayesian methods to shrink high-dimensional VARs toward the LP response (e.g. Tanaka, 2020; Ferreira *et al.*, 2023).

The other setting that is popular in empirical research is panel data. Because LPs are a single-equation method, they lend themselves well to panel data settings. However, although there is an extensive literature using LPs with panels and there is an extensive literature on inference in panel data settings, little is known about LP inference in panel data settings. In the next sections, we give a brief overview on these less traditional settings.

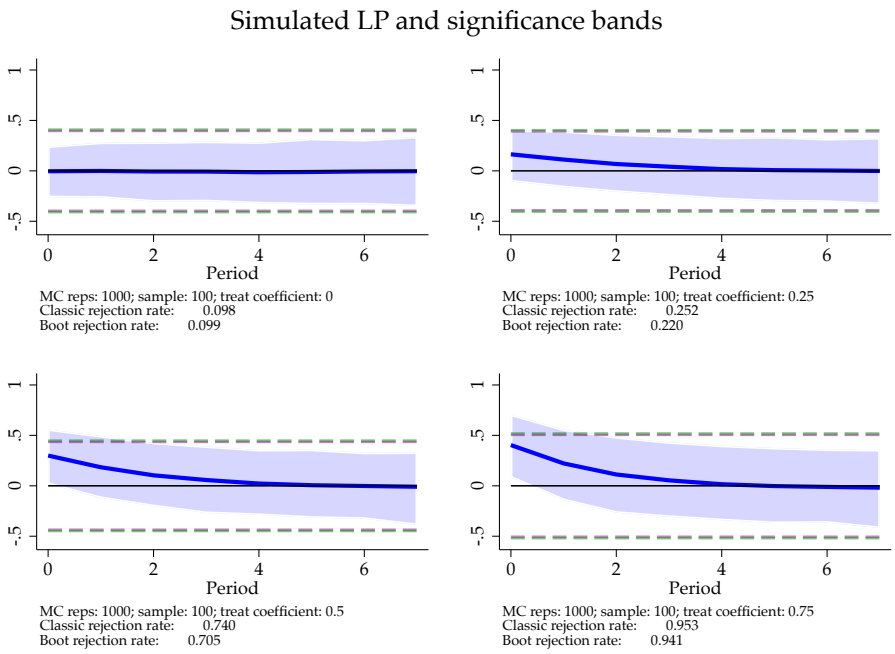
7.1. Smoothing local projections

Smoothing can be used to reduce the uncertainty about the impulse response. Consider the GMM moment condition that we introduced in subsection 5.1, repeated here for convenience:

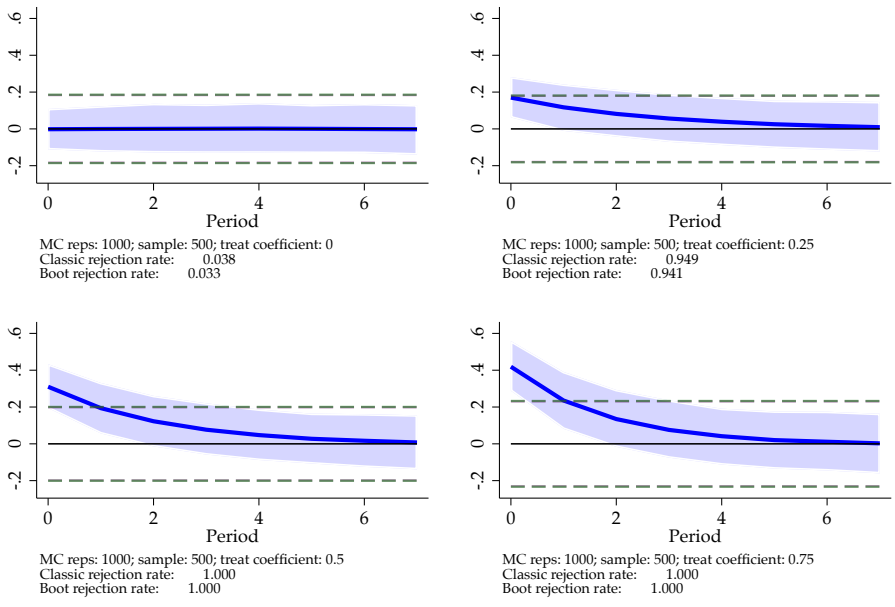
$$E[Z_t'(\boldsymbol{y}_t(H) - S_t\boldsymbol{\beta})] = 0$$

Figure 5: *Significance Bands: Monte Carlo exercise*

(a) *Sample size: 100*



(b) *Sample size: 500*



Notes: Monte Carlo exercise. Sample size = 100 and 500 observations (500 burn-in replications). Significance bands constructed using asymptotic approximations (Classic) and the Wild Block bootstrap (Boot). The rejection rate refers to the share of replications where one or more coefficients exceed the significance bands constructed with each procedure. Treat coefficient refers to the coefficient β described in the text. Significance bands constructed at 95% confidence level. Thus, when Treat coefficient = 0, the rejection rate should be 0.05, otherwise, it should be 1. See text.

where $\boldsymbol{\beta}$ is a $(H+1) \times 1$ dimensional vector with covariance matrix Ω_β, which can be estimated as shown in Equation 13. Smoothing can be thought of as replacing $\boldsymbol{\beta}$ with a lower-dimensional function, say $\phi(h, \boldsymbol{\theta})$ with $dim(\boldsymbol{\theta}) << dim(\boldsymbol{\beta})$. In Barnichon & Brownlees (2019), the authors propose using the B-spline method of Eilers & Marx (1996), whereas Barnichon & Matthes (2018) propose using Gaussian basis functions.

In general, assuming that one can obtain LP estimates that are asymptotically normal (such as when estimating LPs using the GMM setup in subsection 5.1) so that $\hat{\boldsymbol{\beta}} \xrightarrow{d} N(\boldsymbol{\beta}, \Omega_\beta)$ and $\hat{\Omega}_\beta \xrightarrow{p} \Omega_\beta$, then one can setup the minimum distance problem:

$$\min_\theta Q(\boldsymbol{\theta}) = \min_\theta (\hat{\boldsymbol{\beta}} - \phi(h; \boldsymbol{\theta}))' \hat{\Omega}_\beta^{-1} (\hat{\boldsymbol{\beta}} - \phi(h; \boldsymbol{\theta})) \tag{20}$$

to estimate $\hat{\phi}(h, \hat{\boldsymbol{\theta}})$ efficiently. Moreover, under standard regularity conditions, the quality of the approximation can be judged with an overidentification restrictions test since $\hat{Q}(\hat{\boldsymbol{\theta}}) \to \chi^2_q$ with $q = dim(\boldsymbol{\beta}) - dim(\boldsymbol{\theta})$. The variance of $\boldsymbol{\theta}$ is $\hat{\Omega}_\theta = (\Phi_0' \hat{\Omega}_\beta^{-1} \Phi_0)^{-1}$ where $\Phi_0 = \partial \phi(h, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ evaluated at the true value $\boldsymbol{\theta}_0$. An example of how smoothing can make response estimates more efficient is Li *et al.* (2024).

7.2. Panel data

Panel data offers more opportunities to explore data using LPs and more opportunities to conduct inference. As we anticipated in section 4, when the $T \to \infty$ with N fixed or when N grows at a slower rate than T, the asymptotic distribution of the LP estimates will be dominated by the time-dimension of the panel and in that case, the counterpart to Newey-West HAC standard errors is to use the Driscoll & Kraay (1998) covariance estimator.[4]

Cluster robust inference can be used in situations where $N \to \infty$ with T fixed. In such a setting, autocovariances are relatively efficiently estimated and there is no need to specify the residual autocorrelation structure. When T is relatively small, a recommended correction for heteroscedasticity is the wild bootstrap (see, e.g. Cameron *et al.*, 2008; Canay *et al.*, 2021; Roodman *et al.*, 2019)[5]. If N is relatively large, the asymptotic distribution will be dominated by the cross-sectional dimension of the panel. In that case, stationarity or lack thereof plays no role in computing standard errors.

Relatedly, in a recent paper, Mei *et al.* (2023) show that incidental parameter biases (Nickell, 1981) crop up when the dimensions of the panel $N, T \to \infty$ as $N/T \to c$, $c \in (0, \infty)$. To avoid this bias, these authors suggest using the split panel jacknife estimator of Dhaene & Jochmans (2016) and Chudik *et al.* (2018). Denote $\hat{\beta}_h$ the full sample estimate with fixed-effects and $\hat{\beta}_h^a$ and β_h^b estimates based on splitting the sample along the time series dimension into two halves, $T = T_a + T_b$. Then, the bias corrected estimate of the impulse response is $\tilde{\beta}_h = 2\hat{\beta}_h - 0.5(\hat{\beta}_h^a + \hat{\beta}_h^b)$. However, note that a

[4]The command `xtscc` in STATA implements this type of covariance matrix estimation.

[5]In STATA, this type of boostrap can be implemented with the user supplied command `boottest`.

recent paper by Hahn *et al.* (2024) shows that the split panel jackknife bias correction is generally higher order inefficient and may significantly increase estimator variance in finite samples compared to higher order efficient bias correction methods.

8. Conclusion

The error structure of local projections and the kind of hypotheses implicit in the way inference is obtained and communicated requires some care. Residual serial correlation in local projections can be dealt with and more recent developments show that it can be obtained rather easily using lag augmentation and heteroscedasticity robust methods. Bounds for simultaneous inference are also relatively easy to construct, though in most situations we expect that practitioners will simply report formal hypothesis tests.

The significance bands introduced in this review are simple to construct and can be easily displayed alongside the usual confidence bands. While confidence bands inform the reader about the estimation uncertainty of each coefficient, significance bands inform the reader about the significance of the impulse response itself.

Panel data settings present their own challenges and opportunities. The time series and cross-sectional dimensions of the panel play a critical role in choosing the best inferential procedures. Cluster robust inference generally offers an attractive approach, but cannot always be directly used. Inference in panel data settings is an ever growing field and new developments are constantly arriving to improve existing methods.

Applications and extensions of local projections continue to grow. In this review we are unable to cover every single scenario. However, we hope to have provided the reader with general principles that can then be tailored to each specific extension.

References

Angrist, Joshua D., & Kuersteiner, Guido M. 2011. Causal Effects of Monetary Shocks: Semiparametric Conditional Independence Tests with a Multinomial Propensity Score. *The Review of Economics and Statistics*, **93**(3), 725–747.

Angrist, Joshua D., Jordà, Òscar, & Kuersteiner, Guido M. 2016. Semiparametric Estimates of Monetary Policy Effects: String Theory Revisited. *Journal of Business and Economic Statistics*, **http://dx.doi.org/10.1080/07350015.2016.1204919**.

Barnichon, Regis, & Brownlees, Christian. 2019. Impulse Response Estimation by Smooth Local Projections. *Review of Economics and Statistics*, **101**(3), 522–530.

Barnichon, Regis, & Matthes, Christian. 2018. Functional Approximation of Impulse Responses. *Journal of Monetary Economics*, **99**, 41–55.

Bauer, Michael D, & Swanson, Eric T. 2023. A reassessment of monetary policy surprises and high-frequency identification. *NBER Macroeconomics Annual*, **37**(1), 87–155.

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

Cameron, A. Colin, Gelbach, Jonah B., & Miller, Douglas L. 2008. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, **90**(3), 414–427.

Canay, Ivan A., Santos, Andres, & Shaikh, Azeem M. 2021. The wild bootstrap with a "small" number of "large" clusters. *Review of Economics and Statistics*, **103**(2), 346–363.

Chudik, Alexander, Pesaran, M. Hashem, & Yang, Jui-Chung. 2018. Half-panel jackknife fixed-effects estimation of linear panels with weakly exogenous regressors. *Journal of Applied Econometrics*, **33**(6), 816–836.

Dhaene, Geert, & Jochmans, Koen. 2016. Bias-corrected estimation of panel vector autoregressions. *Economics Letters*, **145**, 98–103.

Dolado, Juan J., & Lütkepohl, Helmut. 1996. Making Wald tests work for cointegrated VAR systems. *Econometric Reviews*, **15**(4), 369–386.

Driscoll, John C., & Kraay, Aart C. 1998. Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data. *Review of Economics and Statistics*, **80**(4), 549–560.

Dunn, Olive Jean. 1961. Multiple comparisons among means. *Journal of the American statistical association*, **56**(293), 52–64.

Eilers, Paul H. C., & Marx, Brian D. 1996. Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**(2), 89–121.

Ferreira, Leonardo N., Miranda-Agrippino, Silvia, & Ricco, Giovanni. 2023. Bayesian Local Projections. *Review of Economics and Statistics*, 05, 1–45.

Gonçalves, Sílvia, & Kilian, Lutz. 2004. Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. *Journal of Econometrics*, **123**(1), 89–120.

Gonçalves, Sílvia, & Kilian, Lutz. 2007. Asymptotic and bootstrap inference for AR (∞) processes with conditional heteroskedasticity. *Econometric Reviews*, **26**(6), 609–641.

Hahn, Jinyong, Hughes, David W., Kuersteiner, Guido, & Newey, Whitney K. 2024. Efficient bias correction for cross section and panel data. *Quantitative Economics*, **15**, 783–816.

Inoue, Atsushi, & Kilian, Lutz. 2002. Bootstrapping autoregressive processes with possible unit roots. *Econometrica*, **70**(1), 377–391.

Inoue, Atsushi, & Kilian, Lutz. 2020. The uniform validity of impulse response inference in autoregressions. *Journal of Econometrics*, **215**(2), 450–472.

Jordà, Òscar. 2005. Estimation and Inference of Impulse Responses by Local Projections. *American Economic Review*, **95**(1), 161–182.

Jordà, Òscar. 2009. Simultaneous confidence regions for impulse responses. *Review of Economics and Statistics*, **91**(3), 629–647.

Jordà, Òscar, Singh, Sanjay R., & Taylor, Alan M. 2024. The long-run effects of monetary policy. *Review of Economics and Statistics*, **forthcoming**.

Li, Dake, Plagborg-Møller, Mikkel, & Wolf, Christian K. 2024. Local projections vs. vars: Lessons from thousands of dgps. *Journal of Econometrics,*, **forthcoming**.

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

Mei, Ziwei, Sheng, Liugang, & Shi, Zhentao. 2023. *Nickell Bias in Panel Local Projection: Financial Crises Are Worse Than You Think*. Unpublished. https://arxiv.org/pdf/2302.13455.pdf.

Miranda-Agrippino, Silvia, & Ricco, Giovanni. 2017. The Transmission of Monetary Policy Shocks. *Center for Macroeconomics working paper*, **DP**(11).

Montiel-Olea, José Luis, & Plagborg-Møller, Mikkel. 2019. Simultaneous Confidence Bands: Theory, Implementation, and an Application to SVARs. *Journal of Applied Econometrics*, **34**(1), 1–17.

Montiel Olea, José Luis, & Plagborg-Møller, Mikkel. 2021. Local projection inference is simpler and more robust than you think. *Econometrica*, **89**(4), 1789–1823.

Newey, Whitney K., & West, Kenneth D. 1987. A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, **55**(3), 703–708.

Nickell, Stephen. 1981. Biases in dynamic models with fixed effects. *Econometrica: Journal of the econometric society*, 1417–1426.

Plagborg-Møller, Mikkel, & Wolf, Christian K. 2021. Local projections and VARs estimate the same impulse responses. *Econometrica*, **89**(2), 955–980.

Plagborg-Møller, Mikkel, & Wolf, Christian K. 2022. Instrumental Variable Identification of Dynamic Variance Decompositions. *Journal of Political Economy*, **130**(8), 2164–2202.

Plagborg-Møller, Mikkel, Montiel-Olea, José Luis, Qian, Eric, & Wolf, Christian K. 2024. *Double Robustness of Local Projections and Some Unpleasant VARithmetic*. Tech. rept. 32495. NBER, http://www.nber.org/papers/w32495.

Pope, Alun Lloyd. 1990. Biases of estimators in multivariate non-Gaussian autoregressions. *Journal of Time Series Analysis*, **11**(3), 249–258.

Roodman, David, Nielsen, Morten Ørregaard, MacKinnon, James G., & Webb, Matthew D. 2019. Fast and wild: Bootstrap inference in Stata using boottest. *The Stata Journal*, **19**(1), 4–60.

Scheffé, Henry. 1953. A method for judging all contrasts in the analysis of variance*. *Biometrika*, **40**(1-2), 87–110.

Sims, Christopher A., Stock, James H., & Watson, Mark W. 1990. Inference in linear time series models with some unit roots. *Econometrica*, **58**(1), 113–144.

Stock, James H., & Watson, Mark W. 2018. Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments. *Economic Journal*, **128**(610), 917–948.

Tanaka, Masahiro. 2020. Bayesian inference of local projections with roughness penalty priors. *Computational Economics*, **55**(2), 629–651.

Toda, Hiro Y., & Yamamoto, Taku. 1995. Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics*, **66**(1-2), 225–250.

Xu, Ke-Li. 2023. *Local Projection Based Inference under General Conditions*. Unpublished. https://ssrn.com/abstract=4372388.

Appendix

In this appendix we provide some background on the construction of significance bands proposed in Section 6.

The LM Statistic

We consider the just identified instrumental variables estimator discussed in the paper under the constraint that the null $H_0 : \beta_h = 0$ holds. The estimator $\hat{\beta}_h$ solves the problem

$$\underset{\beta_h}{\operatorname{argmin}} \frac{1}{2} \left(\sum_{t=1}^{T} z_t \left(y_{t+h} - s_t \beta_h \right) \right)^2 .$$

We can set up a Lagrangian to analyze the constrained problem $\beta_h = 0$ as

$$L\left(\beta_h\right) = \frac{1}{2} \left(\sum_{t=1}^{T} z_t \left(y_{t+h} - s_t \beta_h \right) \right)^2 + \lambda \beta_h .$$

The first order condition is

$$\frac{\partial L\left(\beta_h\right)}{\partial \beta} = - \sum_{t=1}^{T} z_t \left(y_{t+h} - s_t \beta_h \right) + \lambda = 0$$

such that the Lagrange multiplier, using $\beta = 0$, is

$$\hat{\lambda} = \sum_{t=1}^{T} z_t y_{t+h}.$$

The LM test now is based on the asymptotic χ_1^2 statistic

$$T_h^2 = \frac{(T-h)^{-1}\hat{\lambda}^2}{Var\left(T^{-1/2}\hat{\lambda}\right)} \rightarrow_d \chi_1^2 \text{ under } H_0.$$

Let $\hat{\omega}$ be an estimator of $Var\left(n^{-1/2}\hat{\lambda}\right)$. Then the test rejects $\beta_h = 0$ if

$$\hat{T}_h^2 = \frac{(T-h)^{-1}\hat{\lambda}^2}{\hat{\omega}} > c_{\chi_1^2,\alpha}$$

where $c_{\chi_1^2,\alpha}$ is the $1-\alpha$ quantile of the χ_1^2 distribution. Note that

$$\begin{aligned}\hat{T}_h^2 &= \frac{(T-h)^{-1}\left(\sum_{t=1}^{T} z_t y_{t+h}\right)^2}{\hat{\omega}} \\ &= \frac{(T-h)\left(\hat{\beta}_h\right)^2}{\hat{\omega}/\left(\frac{1}{T-h}\sum_1^T z_t\, s_t\right)^2}\end{aligned}$$

which shows that the test based on $\hat{T}_h^2$ is numerically identical to the test implied by inverting the confidence interval proposed below. The estimator $\hat{\omega}$ can be understood as the HAC estimator of

$$(T-h)^{-1}\sum_{t=1}^{T} z_t \tilde{u}_{t+h} = (T-h)^{-1}\sum_{t=1}^{T} z_t\left(y_{t+h} - s_t\beta_{h,0}\right) = (T-h)^{-1}\sum_{t=1}^{T} z_t y_{t+h}$$

where $(T-h)^{-1}\sum_{t=1}^{T} z_t y_{t+h}$ is the coefficient in an OLS regression of $z_t y_{t+h}$ on a constant. In other words the variance of $\hat{\beta}_h$ is computed by imposing the null when obtaining the model residual $\tilde{u}_{t+h} = y_{t+h}$. The long run variance (spectrum at zero frequency) of $z_t\tilde{u}_{t+h} = z_t y_{t+h}$ is $\sum_{j=-\infty}^{\infty} E(z_t\, y_{t+h}\, z_{t-j}\, y_{t+h-j})$ whether or not the null holds true in the DGP. The test statistic is evaluated at the null and is based on an estimator for $\sigma^2 = \frac{\tilde{\omega}}{\gamma_{zs}^2}$ where $\tilde{\omega} = \sum_{j=-\infty}^{\infty} E(z_t\, y_{t+h}\, z_{t-j}\, y_{t+h-j})$.

Bonferroni Inequality

The Bonferroni inequality in the context of our impulse response coefficients can be stated as

$$\begin{aligned}&P\left(\bigcup_{h=0}^{H-1}\left\{\hat{\beta}_h \notin \left[\zeta_{\alpha/2H}\frac{\sigma}{\sqrt{T-h}}, \zeta_{(1-\alpha/2H)}\frac{\sigma}{\sqrt{T-h}}\right]\right\}\right) \\ &\leq \sum_{h=0}^{H-1} P\left(\left\{\hat{\beta}_h \notin \left[\zeta_{\alpha/2H}\frac{\sigma}{\sqrt{T-h}}, \zeta_{(1-\alpha/2H)}\frac{\sigma}{\sqrt{T-h}}\right]\right\}\right) \\ &\approx \sum_{h=0}^{H-1}\frac{\alpha}{H} = \alpha.\end{aligned}$$

where the approximation in the last line refers to the fact that the individual confidence intervals have coverage $1-\alpha/H$ in large samples. Based on calculations in the next section one obtains the inequality (see also Dunn Equation 5), that the probability of all impulse response coefficients to be inside the confidence sets is

$$P\left(\bigcap_{h=0}^{H-1}\left\{\zeta_{\alpha/2H}\frac{\sigma}{\sqrt{T-h}}<\hat{\beta}_h<\zeta_{(1-\alpha/2H)}\frac{\sigma}{\sqrt{T-h}}\right\}\right)\geq 1-\alpha$$

and where $\left\{a<\hat{\beta}_h<b\right\}$ denotes the set of all samples where $\hat{\beta}_h\in[a,b]$.

Joint Test of Non-Zero Impulse Responses

We define a test statistic that rejects $H_0:\beta=0$ if at least one of the elements of $\beta=[\beta_0,...,\beta_{H-1}]$ is outside the confidence interval. Formally, we say that we reject H_0 if

$$T_n=\sum_{h=0}^{H-1}1\left\{\hat{\beta}_h\notin\left[\zeta_{\alpha/2H}\hat{s}^b_{\beta_h},\ \zeta_{1-\alpha/2H}\hat{s}^b_{\beta_h}\right]\right\}>0$$

where $1\{.\}$ is equal to 1 if the argument is true, and zero otherwise. In words, the expression counts the number of $\hat{\beta}_h$ that are not inside the confidence interval. This means that we reject the null if at least one of the estimates is outside the confidence band around zero. The probability of rejecting the null is

$$\begin{aligned}\Pr(T_n>0)&=\Pr\left(\bigcup_{h=0}^{H-1}\left\{\hat{\beta}_h\notin\left[\zeta_{\alpha/2H}\hat{s}^b_{\beta_h},\ \zeta_{1-\alpha/2H}\hat{s}^b_{\beta_h}\right]\right\}\right)\\&\leq\sum_{h=0}^{H-1}\Pr\left(\hat{\beta}_h\notin\left[\zeta_{\alpha/2H}\hat{s}^b_{\beta_h},\ \zeta_{1-\alpha/2H}\hat{s}^b_{\beta_h}\right]\right)\approx\sum_{h=0}^{H-1}\frac{\alpha}{H}=\alpha\end{aligned}$$

and where the inequality is Bonferroni's inequality.

Now we construct a confidence region in $\mathbb{R}^H$ that contains the estimated impulse response $\hat{\beta}$ in repeated samples with at least probability $1-\alpha$ if the null H_0 is true. Define the following sets

$$A_h=\left\{b=(b_0,...,b_{H-1})\in\mathbb{R}^H|b_h\in\left[\zeta_{\alpha/2H}\hat{s}^b_{\beta_h},\ \zeta_{1-\alpha/2H}\hat{s}^b_{\beta_h}\right]\right\}$$

where A_h is a strip in $\mathbb{R}^H$ that goes through the interval $\left[\zeta_{\alpha/2H}\hat{s}^b_{\beta_h},\ \zeta_{1-\alpha/2H}\hat{s}^b_{\beta_h}\right]$ on axis $h-1$. For example, for $H=2$, A_1 is the set of all values $b=(b_0,b_1)$ such that $b_1\in\left[\zeta_{\alpha/4}\hat{s}^b_{\beta_h},\ \zeta_{1-\alpha/4}\hat{s}^b_{\beta_1}\right]$and $b_0\in\mathbb{R}$ is unconstrained. Then,

$$\bigcap_{h=0}^{H-1}A_h=\left\{b\in\mathbb{R}^H|b_h\in\left[\zeta_{\alpha/2H}\hat{s}^b_{\beta_h},\ \zeta_{1-\alpha/2H}\hat{s}^b_{\beta_h}\right],h=0,...,H-1\right\}$$

is a rectangle in the H-dimensional space such that each of the coordinates of b are in the one dimensional confidence interval. By De Morgan's law $\left(\bigcap_{h=0}^{H-1}A_h\right)^c=\bigcup_{h=0}^{H-1}A_h^c$, where $(.)^c$ denotes

the complement. We then have

$$\begin{aligned}\Pr\left(\hat{\beta} \in \left(\bigcap_{h=0}^{H-1} A_h\right)\right) &= 1 - \Pr\left(\hat{\beta} \in \left(\bigcap_{h=0}^{H-1} A_h\right)^c\right) \\ &= 1 - \Pr\left(\bigcup_{h=0}^{H-1} \left\{\hat{\beta}_h \notin \left[\zeta_{\alpha/2H}\hat{s}^b_{\beta_h},\ \zeta_{1-\alpha/2H}\hat{s}^b_{\beta_h}\right]\right\}\right) \qquad (21) \\ &\geq 1 - \sum_{h=0}^{H-1} \Pr\left(\hat{\beta}_h \notin \left[\zeta_{\alpha/2H}\hat{s}^b_{\beta_h},\ \zeta_{1-\alpha/2H}\hat{s}^b_{\beta_h}\right]\right) = 1 - \alpha\end{aligned}$$

where the first equality uses that $\left(\bigcap_{h=0}^{H-1} A_h\right)$ and $\left(\bigcap_{h=0}^{H-1} A_h\right)^c$ are disjoint, and $\left(\bigcap_{h=0}^{H-1} A_h\right) \cup \left(\bigcap_{h=0}^{H-1} A_h\right)^c = \mathbb{R}^H$. The second equality uses

$$\left\{\hat{\beta} \in \left(\bigcap_{h=0}^{H-1} A_h\right)\right\} = \left(\bigcap_{h=0}^{H-1} \left\{\hat{\beta} \in A_h\right\}\right)$$

since $\hat{\beta}$ is in the hypercube $\left(\bigcap_{h=0}^{H-1} A_h\right)$ if all coordinates $\hat{\beta}_h$ are in the segments defining the hypercube. Here the $\{.\}$ brackets mean all outcomes for which $\hat{\beta}$ satisfies the condition. Since $\hat{\beta} \in A_h$ iff $\hat{\beta}_h \in \left[\zeta_{\alpha/2H}\hat{s}^b_{\beta_h},\ \zeta_{1-\alpha/2H}\hat{s}^b_{\beta_h}\right]$ it follows that $\left\{\hat{\beta} \in A_h\right\}^c = \left\{\hat{\beta}_h \notin \left[\zeta_{\alpha/2H}\hat{s}^b_{\beta_h},\ \zeta_{1-\alpha/2H}\hat{s}^b_{\beta_h}\right]\right\}$ Now apply de Morgan's law to the RHS. Finally, the inequality in the display above is the Bonferroni inequality.

Since we do not reject H_0 if $T_n = 0$ and $T_n = 0$ iff $\hat{\beta}_h \in \left[\zeta_{\alpha/2H}\hat{s}^b_{\beta_h},\ \zeta_{1-\alpha/2H}\hat{s}^b_{\beta_h}\right]$ for all $h = 0, ..., H-1$ we obtain

$$\begin{aligned}\Pr\left(T = 0\right) &= \Pr\left(\bigcap_{h=0}^{H-1} \left\{\hat{\beta}_h \in \left[\zeta_{\alpha/2H}\hat{s}^b_{\beta_h},\ \zeta_{1-\alpha/2H}\hat{s}^b_{\beta_h}\right]\right\}\right) \\ &= 1 - \Pr\left(\bigcup_{h=0}^{H-1} \left\{\hat{\beta}_h \notin \left[\zeta_{\alpha/2H}\hat{s}^b_{\beta_h},\ \zeta_{1-\alpha/2H}\hat{s}^b_{\beta_h}\right]\right\}\right) \\ &\geq 1 - \alpha\end{aligned}$$

where the second equality follows from the laws of total probability and De Morgan's law and the inequality follows from the bound in (21).