NBER WORKING PAPER SERIES

LOCAL PROJECTIONS

Òscar Jordà
Alan M. Taylor

Working Paper 32822
http://www.nber.org/papers/w32822

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2024

We are grateful to Regis Barnichon, Colin Cameron, James Cloyne, Olivier Coibion, Yuriy Gorodnichenko, Amaze Lusompa, Christian Matthes, Valerie Ramey, Sanjay Singh, and Takuya Ura for useful comments and suggestions. Research by many of the authors that we discuss in the article, and others that we unfortunately will have inevitably missed, played an important role in clarifying local projections and advancing them into mainstream empirical research. Steven Durlauf and David Romer helped guide this article to fruition, along with several anonymous referees, and for their help and advice we are very grateful. The views expressed herein do not necessarily represent the views of any of the institutions of the Federal Reserve System. All errors are our own. The online repository of the STATA code that replicates all the examples in the article is available at: https://github.com/ojorda/JEL-Code. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Local Projections
Òscar Jordà and Alan M. Taylor
NBER Working Paper No. 32822
August 2024
JEL No. C01,C14,C22,C26,C32,C54

ABSTRACT

A central question in applied research is to estimate the effect of an exogenous intervention or shock on an outcome. The intervention can affect the outcome and controls on impact and over time. Moreover, there can be subsequent feedback between outcomes, controls and the intervention. Many of these interactions can be untangled using local projections. This method's simplicity makes it a convenient and versatile tool in the empiricist's kit, one that is generalizable to complex settings. This article reviews the state-of-the art for the practitioner, discusses best practices and possible extensions of local projections methods, along with their limitations.

Òscar Jordà
Economic Research, MS 1130
Federal Reserve Bank of San Francisco
San Francisco, CA 94105
and University of California, Davis and also CEPR
oscar.jorda@gmail.com

Alan M. Taylor
Columbia University
1313 International Affairs
420 West 118th Street
New York, NY 10027
and CEPR
and also NBER
amt2314@columbia.edu

1. Introduction

In the last 20 years, an increasingly convenient and widely-used way to estimate how an exogenous policy intervention or shock will affect an outcome over time—an impulse response—is with local projections (Jordà, 2005). Given the broad adoption and extensive development of this approach, a survey now seems very timely. Local projections (or LPs) are a sequence of regressions where the outcome, dated at increasingly distant horizons, is regressed on the intervention (directly, if randomly assigned; or perhaps instrumented, if not), conditional on a set of controls that include lags of both the outcome and the intervention, as well as other exogenous or predetermined variables.

At a basic level, LPs and vector autoregressions (VARs) aim to characterize the dynamic covariance structure of a system of variables. Perhaps not surprisingly, the impulse response estimates are asymptotically equivalent (though not in small samples) for the two methods under relatively general conditions when the data are generated by a VAR (Jordà, 2005; Plagborg-Møller and Wolf, 2021). They are also equivalent to the estimator of the moving average representation of the VAR proposed by Chang and Sakata (2007).

Given this equivalence with VARs, why are LPs necessary? The extensive literature on VARs, which we do not explicitly review here, has shown the many advantages of VARs as a forecasting tool, and as a simple way to obtain impulse responses. The solutions of many models in macroeconomics result in a system of difference (or differential equations, as the case may be), which can be well approximated with a VAR. In turn, impulse response functions can be conveniently estimated and policy experiments conducted. Depending on the setting, inference can also be less efficient with LPs. So there are good reasons to use VARs in certain applications.

However, over time LPs have gradually been seen to have several key advantages over VARs, some of which deserve to be highlighted up front. First, LPs rely on single-equation methods, which can be advantageous when specifying the full system is inconvenient due to data limitations or model complexity. Second, as a single-equation method, LPs can be useful in situations where there are nonlinearities or state-dependence though proper care must be exercised in their interpretation, as we will discuss. Third, LPs make estimation and inference convenient for many important objects of study, such as cumulative responses and multipliers. Fourth, LPs provide an encompassing framework for panel data and difference-in-difference, staggered event studies with heterogeneous treatment effects, though here the researcher once again faces a bias-variance trade-off.

LPs are of interest in their own right, where the connection to VARs is an advantageous feature but not necessarily an end in itself. In general experimental settings, the goal is to approximate the (conditional) mean difference between outcomes when an intervention is administered versus when it is counterfactually withheld. LPs can be seen as a semi-parametric method that imposes relatively mild assumptions on the data and on the shape of the response. In fact, Rambachan and Shephard (2019a,b) provide formal conditions using the potential outcomes paradigm that take this viewpoint even further in a fully nonparametric direction. The cost is that LPs will be less efficient than models that impose more structure, such as VARs.

In this respect, LPs provide a natural nexus between empirical macroeconomics, on the one hand, and the policy evaluation literature in applied microeconomics on the other. The questions are fundamentally the same, an exploration of policy counterfactuals and their effects. As a result, this link between applied problems in macro and microeconomics opens up many interesting and fertile opportunities for synergistic improvement in both areas (see, e.g., Dube, Girardi, Jordà, and Taylor, 2023, for an application to difference-in-differences estimation).

The goal of this survey is to overview the statistical properties of LPs, beginning with emphasis on estimation, inference, and small-sample properties as they have been developed in the literature. We start from the basics of how to set up and estimate LPs, and then discuss bias, multipliers, and smoothing, before moving on to inference. Next, we examine topics such as instrumental variables and other methods of identification, and impulse response decompositions. We provide a brief discussion on impulse response matching estimators of general models before showcasing state-dependence and nonlinear extensions, and the latest developments and applications to panel data. The breadth of material by necessity limits the rigor we can bring to each topic and the extent of the literature we can cover, though we refer the reader to the original sources for details.[1]

2. Intuition and basics

We begin with a simple but fairly typical dynamic setting to introduce the main ideas. Let y_t denote an outcome variable of interest. Let the controls $\boldsymbol{x}_t$ denote a vector of exogenous or pre-determined variables, including lags of the outcome and of the policy intervention, which we denote as s_t. Think of the policy intervention as an exogenous shock, such as a natural disaster; or a structural shock, such as a surprise interest rate hike; or a treatment—as when in a panel, some states raise the minimum wage. Finally, let $\boldsymbol{z}_t$ denote a vector of instruments for s_t, if these are available.[2]

We are interested in characterizing how an intervention today affects the average outcome at some time in the future relative to a baseline of no-intervention. Formally, we define an impulse response as

$$\mathcal{R}_{s\to y}(h,\delta) \equiv E[y_{t+h}|s_t = s_0+\delta;\boldsymbol{x}_t] - E[y_{t+h}|s_t = s_0;\boldsymbol{x}_t]\,; \quad h=0,1,\dots,H\,, \tag{1}$$

where δ is the size of the intervention, or *dose*. A common scale choice is to normalize $\delta = 1$ in some units—e.g., a 1 percentage point shock to the interest rate, a 1% of GDP fiscal shock, or a 1 s.d. perturbation. When the unit dose is obvious we may omit δ from the notation, and write $\mathcal{R}_{s\to y}(h,1) \equiv \mathcal{R}_{s\to y}(h)$. Further, the subscript $s \to y$ indicates that the intervention s affects the outcome y. More simply, we may write $\mathcal{R}_{sy}(h)$; if the context is clear, we may just omit the subscript.

The value s_0 is a baseline level. In linear models, the baseline will not matter, as it will cancel out

[1] A full set of replication code accompanies the paper providing a template for users. It is available at: https://github.com/ojorda/JEL-Code.

[2] We will often use the term *intervention*, *shock*, or *treatment* interchangeably.

when taking the difference in expectations shown in Equation 1. In nonlinear models, this will not be the case: the baseline s_0 from which the intervention is evaluated can influence the effect of the intervention itself. Note that as $\delta \to 0$, the interpretation of $\mathcal{R}_{s\to y}(h,\delta)/\delta$ is that of a *derivative*. This is how impulse responses from VARs are often derived and interpreted and the same can be said with LPs, though this way of thinking is relatively new to applied macroeconomics and time series (exceptions include, e.g., Angrist and Kuersteiner, 2011; Angrist, Jordà, and Kuersteiner, 2016).

Baseline LP and LP-IV Each of the expectations in Equation 1 could be estimated with a flexible estimator. Here, however, we assume linearity to make the example easier to follow. Following Jordà (2005), the *local projection* or LP of y_{t+h} on s_t can be estimated with the following regressions:

$$y_{t+h} = \alpha_h + \beta_h s_t + \boldsymbol{\gamma}_h' \boldsymbol{x}_t + v_{t+h}\,; \quad h = 0, 1, \ldots, H\,; \tag{2}$$

with $\mathcal{R}_{s\to y}(h) = \beta_h$ by direct application of the definition in Equation 1. Here, the specific properties of the residual v_{t+h} and their effect on inference will depend on the data generating process considered. For now we assume that $E(s_t, v_{t+h}) = 0$, e.g., as when s_t is exogenous (i.e., determined at random). In this case, the LP is identified and might be estimated by OLS; we might say LP-OLS.

If s_t is not exogenously determined but we have $\boldsymbol{z}_t$ available as instruments for s_t, we can then estimate the LP using instrumental variable methods. This we will call LP-IV, introduced by Jordà, Schularick, and Taylor (2015), a technique that has quickly become a mainstay of applied macroeconomics research (the literature is now too vast to cite; see, e.g., Ramey, 2016, for a nice review). Deeper discussion of identification is something we defer and revisit in more detail in Section 8 where we will further discuss specific conditions that the instruments must meet.

In this section we explore some basic ideas. First, the assumption of linearity may appear restrictive. However, since a different regression is estimated for each horizon h, one can think of Equation 2 as a *semi-parametric* estimate of $\mathcal{R}_{s\to y}(h)$: a different regression model at each horizon approximates the conditional mean described in Equation 1, rather than specifying a model that characterizes the full dynamic evolution of $y_t, s_t, \boldsymbol{x}_t$, and $\boldsymbol{z}_t$ from which one can then derive $\mathcal{R}_{s\to y}(h)$ at all horizons. Second, for reasons that will become clear shortly, v_{t+h} is likely to be serially correlated up to h lags. Though this feature will not affect the consistency of our estimator for β_h, it affects how inference should be obtained and has implications for small samples that we shall also discuss in a moment. Third, although Equation 2 is presented as a setup for time series data, it is clear that it can be extended to panel data settings straightforwardly, as we discuss later. Finally, many useful nonlinear extensions are easier to implement because the setting in Equation 2 is a collection of single equations rather than a system.

On this last point, linearity imposes restrictions, whether in a VAR or in a LP, that are seldom appreciated. Under linearity, (*i*) interventions have symmetric effects, $\mathcal{R}_{s\to y}(h,\delta) = -\mathcal{R}_{s\to y}(h,-\delta)$. For example, this property implies that interest rate increases reduce inflation by as much as interest rate decreases boost inflation. Linearity also means that (*ii*) responses are independent from recent

history as embedded in the controls (i.e, independent of the state), $\mathcal{R}_{s\to y}(h,\delta|\boldsymbol{x}_t) = \mathcal{R}_{s\to y}(h,\delta)$. Thus, a rate hike in a recession, say, is expected to have the same effect as in an expansion. Finally, linearity means that (*iii*) responses are linearly proportional to the size of the intervention, $\mathcal{R}_{s\to y}(h,\delta) = \delta\mathcal{R}_{s\to y}(h) = \delta\beta_h$. Hence, doubling an interest rate hike is expected to double the reduction in inflation. These features are illustrated further in Section 12.

Other times, we may define Equation 1 for other moments of the data. For example, a practitioner may be interested in the probability of default on a debt at some point in the future if the interest rate were to increase today. Then Equation 1 could be redefined as follows,

$$\mathcal{R}_{s\to y}(h,\delta) \equiv P[y_{t+h}=1|s_t=s_0+\delta;\boldsymbol{x}_t] - P[y_{t+h}=1|s_t=s_0;\boldsymbol{x}_t]\,; \quad h=0,1,\dots,H\,, \tag{3}$$

and this LP form could be estimated with simple logit or probit models. Then, depending on functional form, the initial value s_0 could matter a lot. The effect on the default probability of a 1 percentage point increase in rates when rates are already high could be quite different than when rates are low. Of course, this point applies more generally when LPs are extended to nonlinear settings. We leave this and other extensions (e.g., to quantiles) for later sections of the paper.

LPs in relation to VARs Why LPs? The rationale for an LP might be simply that it estimates a moment in the data of possible interest: that is, without further assumptions on the underlying data generating process or DGP, the setup at Equation 2 allows one to estimate an economically interesting statistic and, therefore, no further justification for this regression would be needed.

However, Jordà (2005) and later more formally Plagborg-Møller and Wolf (2021) show that in large samples, the impulse responses from LPs and VARs of infinite order will be equivalent under relatively mild conditions. A simple example illustrates this equivalency and other important properties that help us better understand how to set up LPs. The lengthier details of various identification approaches will be covered fully in Section 8.

Here, we will consider an illustration just using a VAR(1) with uncorrelated errors. Assume that, expressed in *differences*, a random $k\times 1$ vector $\Delta\boldsymbol{w}_t$ follows a first-order stationary VAR(1) process

$$\Delta\boldsymbol{w}_t = \boldsymbol{\Phi}\,\Delta\boldsymbol{w}_{t-1} + \boldsymbol{u}_t\,; \qquad \boldsymbol{u}_t \sim D(0,\Omega_u)\,; \quad |\lambda_l(\boldsymbol{\Phi})|<1 \quad \text{for} \quad l=1,\dots,k\,, \tag{4}$$

where the constant and any other deterministic terms (such as time trends) are omitted for convenience, but without loss of generality they could have been easily included, and $\mathcal{D}(0,\Omega_u)$ denotes a generic density with mean 0 and variance Ω_u. The notation $\lambda_l(\boldsymbol{\Phi})$ refers to one element of the set of eigenvalues (spectrum) of the matrix $\boldsymbol{\Phi}$; our assumption of stationarity means that the eigenvalues are inside the unit circle. We will now assume here that the residuals $\boldsymbol{u}_t$ are a white noise process with diagonal covariance matrix Ω_u; of course, this will generally not be the case in practice.

A first order process may seem restrictive, but in reality a wide class of time series processes have state-space representations where the states evolve as a first-order process as described by Equation 4.

Examples include the more general VAR(p) as well as the less common VARMA(p, q) models, and countless others (see, e.g., Harvey, 1991; Hamilton, 1994a,b, for these and other examples).

If we propagate forward the process in Equation 4 by recursive substitution, we obtain

$$\Delta \boldsymbol{w}_{t+h} = \boldsymbol{\Phi}^{h+1} \Delta \boldsymbol{w}_{t-1} + \boldsymbol{u}_{t+h} + \boldsymbol{\Phi}\, \boldsymbol{u}_{t+h-1} + \ldots + \boldsymbol{\Phi}^{h}\, \boldsymbol{u}_{t}\,; \quad h = 1, \ldots, H\,, \tag{5}$$

and if we allow $H \to \infty$, given our assumption $|\lambda_l(\boldsymbol{\Phi})| < 1$ for $l = 1, \ldots, k$, and we obtain the well-known Wold representation

$$\Delta \boldsymbol{w}_t = \boldsymbol{u}_t + \boldsymbol{\Phi}\boldsymbol{u}_{t-1} + \boldsymbol{\Phi}^2\boldsymbol{u}_{t-2} + \ldots\,,$$

since $||\boldsymbol{\Phi}^{\infty}|| \to 0$.[3] This expression makes clear how a shock propagates through the system since

$$\frac{\partial \Delta \boldsymbol{w}_{t+h}}{\partial \boldsymbol{u}_t} = \boldsymbol{\Phi}^h \quad \Longrightarrow \quad \frac{\partial \Delta w_{i,t+h}}{\partial u_{jt}} = \boldsymbol{e}_i \Phi^h \boldsymbol{e}_j' \equiv \phi_{ij}^{(h)}\,; \quad h = 0, 1, \ldots, H\,, \tag{6}$$

where $\boldsymbol{e}_l$ is the l^{th} row of the identity matrix of order k for $l = i, j$, and simply selects the appropriate entries of the coefficients in $\boldsymbol{\Phi}^h$. Here, $\boldsymbol{\Phi}^h$ is the matrix $\boldsymbol{\Phi}$ raised to the power h, and we define its ijth entry to be $\phi_{ij}^{(h)}$. Thus the impulse response can be expressed as,

$$\mathcal{R}_{j \to i}(h, \delta) = \boldsymbol{e}_i\, \boldsymbol{\Phi}^h\, \boldsymbol{e}_j'\, \delta = \phi_{ij}^{(h)}\, \delta\,, \quad h = 0, 1, \ldots, H\,. \tag{7}$$

Here the notation $\mathcal{R}_{j \to i}(h, \delta)$, or later simply $\mathcal{R}_{ji}(h, \delta)$, uses the index j to denote the shock variable, and i to denote the response variable.

Turning to the representation of the system in *levels*, note that from the Wold representation

$$\boldsymbol{w}_t = \boldsymbol{u}_t + (\mathbf{I} + \boldsymbol{\Phi})\boldsymbol{u}_{t-1} + (\mathbf{I} + \boldsymbol{\Phi} + \boldsymbol{\Phi}^2)\boldsymbol{u}_{t-2} + \ldots\,, \tag{8}$$

or, in other words,

$$\frac{\partial \boldsymbol{w}_{t+h}}{\partial \boldsymbol{u}_t} = \mathbf{I} + \boldsymbol{\Phi} + \ldots + \boldsymbol{\Phi}^h\,, \tag{9}$$

so that the impulse response in levels is just the cumulative of the responses in differences. This observation can also be seen by realizing that $\boldsymbol{w}_{t+h} - \boldsymbol{w}_{t-1} = \Delta \boldsymbol{w}_{t+h} + \ldots + \Delta \boldsymbol{w}_t$. Using similar notation to Equation 7, we may denote the cumulative response as

$$\mathcal{R}^c_{j \to i}(h, \delta) = \boldsymbol{e}_i\, (\mathbf{I} + \boldsymbol{\Phi} + \ldots + \boldsymbol{\Phi}^h)\, \boldsymbol{e}_j'\, \delta = (1 + \phi_{ij}^{(1)} + \ldots + \phi_{ij}^{(h)})\, \delta\,, \quad h = 0, 1, \ldots, H\,, \tag{10}$$

where we now use the superscript c to denote that we are calculating the *cumulative* response and hence, $\mathcal{R}^c_{j \to i}(h, \delta) = \sum_{l=1}^{h} \mathcal{R}_{j \to i}(l, \delta)$.

[3] The notation $||\boldsymbol{A}|| = [\text{Tr}(\boldsymbol{A}'\boldsymbol{A})]^{1/2}$ refers to the Frobenius norm where $Tr(B)$ is the trace of the square matrix B, that is the sum of the elements in the main diagonal.

Comments and caveats Several observations deserve comment. First, estimation of $\mathcal{R}_{j\to i}(h,\delta)$ or $\mathcal{R}^c_{j\to i}(h,\delta)$ appears straightforward—it only requires the estimation of the VAR and a simple transformation of the estimated matrix of parameters, $\mathbf{\Phi}$, to obtain impulse response estimates. Second, the previous discussion, however, also suggests that $\mathcal{R}_{j\to i}(h,\delta)$ can be directly estimated from a univariate regression of the *first difference* $\Delta w_{i,t+h}$ on $\Delta w_{j,t}$, or if $\mathcal{R}^c_{j\to i}(h,\delta)$ is desired, a regression of the *long difference* $\Delta_h w_{i,t+h} \equiv w_{i,t+h} - w_{i,t-1}$ on $\Delta w_{j,t}$.[4] Either of these can clearly be seen as special cases of the local projection presented in Equation 2.

Asymptotic results to derive inferential procedures for impulse responses estimated with a VAR are well developed. The closed-form asymptotic expressions rely on the delta-method, and simulation methods, such as the bootstrap, or even Bayesian Markov Chain Monte Carlo (MCMC) methods are readily available in most econometrics software packages. However, though it is very rarely acknowledged, even stationary VARs suffer from small sample biases as noted by Nicholls and Pope (1988) and Pope (1990). These biases are inversely related to the sample size T, and they are often ignored in applied work. However, in relevant small samples and with high persistence processes, the biases can be considerable, as we discuss below in comparison to LPs.[5]

Similarly, asymptotic-based inference for LPs is easy to derive since LP coefficient estimates are themselves the impulse response coefficients, although corrections for serial correlation are needed and will result in bigger standard errors. We return to these issues in more detail in Section 6. Further, LPs can also suffer from small sample issues, though these can be greatly remedied for many situations of interest, as we shall see, and this remains an area of ongoing research.

We conclude by noting that, to our knowledge, there is no well-established method to select how many lags to include in the LP. Selection criteria are invalid when residuals are autocorrelated, as in LPs for $h > 1$. However, since the LP in the first horizon $h = 1$ is equivalent to the corresponding equation in a VAR, a natural approach is to use information criteria to determine the lag-length as usual for that case, and then use the same lag-length at subsequent horizons. One caveat is that some inferential procedures call for lag-augmentation (that is, adding one more lag than needed), something that we discuss further in Section 6.

All that said, local projections do tend to be more forgiving when the lag length is not correctly chosen. Jordà, Singh, and Taylor (2024) show that in infinite order processes, LPs have lower bias than VARs at horizons greater than the optimal truncation lag-length. The reason is that in a local projection, the impulse response coefficient is directly estimated. Small misspecification errors do not compound, as they do when estimating impulse responses with a VAR. Plagborg-Møller, Montiel-Olea, Qian, and Wolf (2024) further show that, while VAR confidence intervals substantially undercover with misspecification so small that it is difficult to detect statistically, LPs enjoy a "doubly robust" property of having lower bias and providing correct coverage even with misspecification that can be detected with probability approaching 1.

[4] From here on we abuse the notation, and thus the *long difference* $\Delta_h w_{i,t+h}$ will mean $w_{i,t+h} - w_{i,t-1}$ rather than $w_{i,t+h} - w_{i,t}$ so as to keep notation to a minimum.

[5] Formally, this bias is order $O(T^{-1})$.

3. Specification choice: Levels versus long differences

The practitioner has several choices of LP specification to estimate impulse responses, as seen above. In the original formulation of local projections by Jordà (2005) the specification was set up in levels. But just because the system can be set up in levels, y_{t+h}, does not mean that it should be estimated that way. Indeed, over the subsequent years in our own applied work we have generally turned to the long difference specification, $y_{t+h} - y_{t-1}$, as our preferred tool. Note that going forward, we will now use the even simpler notation $\mathcal{R}_{sy}(h)$ rather than $\mathcal{R}_{s\to y}(h)$.

Stationary case Consider a simple but informative example. Suppose the DGP for the outcome y_t is given by $y_t = \alpha + s_t + \rho y_{t-1} + \epsilon_t$, the assumptions of the previous section hold, and $0 < \rho < 1$. The treatment, s_t, and noise, ϵ_t, are assumed i.i.d. standard normal (for simplicity). One could complicate matters in a variety of ways that we refrain from exploring here to focus on the intuition.

In this example the true impulse response for a unit shock is clearly $\mathcal{R}_{sy}(h) = \rho^h$. Consider two possible LP specifications that are often used in estimation of the impulse response,

$$\text{Level:} \qquad y_{t+h} = \alpha_h^L + \beta_h^L s_t + \gamma_h^L y_{t-1} + u_{t+h}^L\,; \tag{11}$$

$$\text{Long difference:} \qquad y_{t+h} - y_{t-1} = \alpha_h^{LD} + \beta_h^{LD} s_t + \gamma_h^{LD}(y_{t-1} - y_{t-2}) + u_{t+h}^{LD}\,. \tag{12}$$

In the first case, the *levels specification*, we regress the level of the outcome at horizon h (i.e., y_{t+h}) on its lag (y_{t-1}) and the treatment (s_t). In the second case, the *long difference specification*, we regress the long difference in the outcome at horizon h, $\Delta_h y_{t+h} \equiv y_{t+h} - y_{t-1}$, on the lagged first difference, $\Delta y_{t-1} = y_{t-1} - y_{t-2}$, and the treatment, s_t.[6] Hence the *levels* impulse response estimate is $\hat{\mathcal{R}}_{sy}^L(h) = \hat{\beta}_h^L$ whereas the *long-differences* estimate is $\hat{\mathcal{R}}_{sy}^{LD}(h) = \hat{\beta}_h^{LD}$.

Asymptotically, both of these specifications are equivalent and would recover the same impulse response $\mathcal{R}_{sy}(h) = \rho^h$ as in the true model. However, in small samples, the problem of bias with autocorrelation can be severe, as first identified by Orcutt (1948), Marriott and Pope (1954) and Kendall (1954). These earlier results were then expanded to express the small sample biases in VAR models by Nicholls and Pope (1988) and Pope (1990). This issue has been explored for LPs in recent papers by Piger and Stockwell (2023) and Herbst and Johannsen (2024).

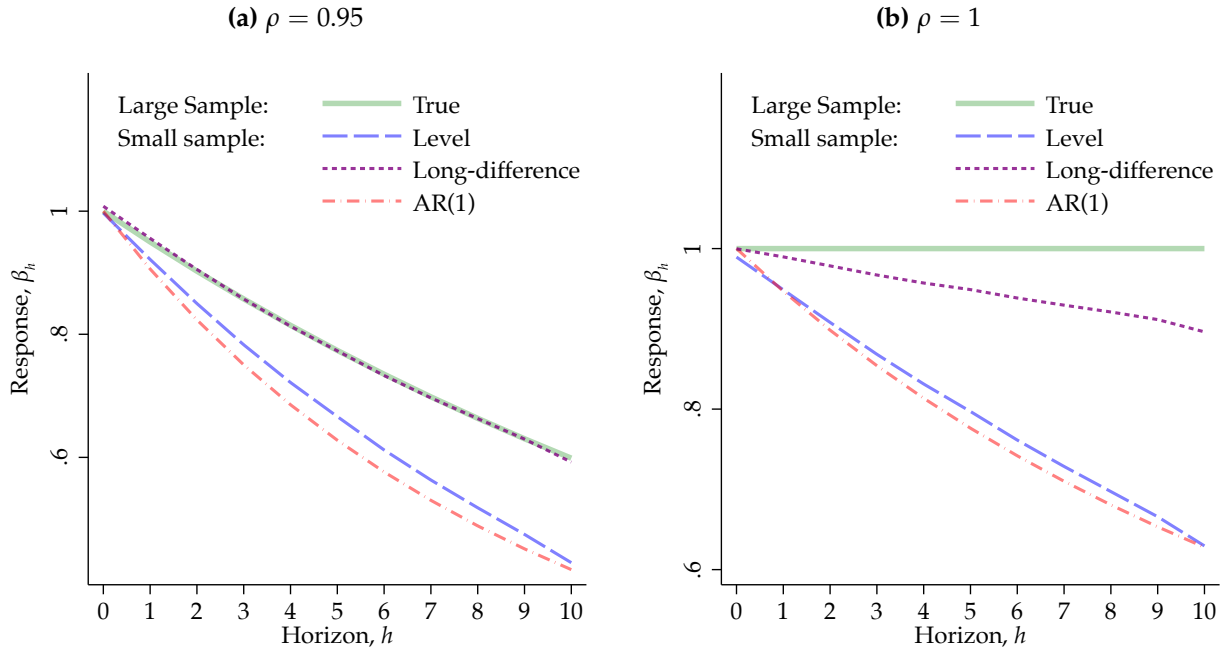
In particular, Piger and Stockwell (2023) explore the "pure" long-difference specifications above which include only the lagged difference, and not the lagged level, as a control. The results are striking. The small-sample bias[7] of order $O(T^{-1})$ discussed in Herbst and Johannsen (2024) is largely suppressed when $|\rho| < 1$. (It is even substantially reduced when $|\rho| = 1$.)

As an illustration, panel (a) of Figure 1 shows the bias reduction when estimating an AR(1) model, $y_t = \rho y_{t-1} + \epsilon_t$, where the DGP is $y_t = s_t + \rho y_{t-1} + \epsilon_t$ with $\rho = 0.95$ for a sample of $T = 100$

[6] As a reminder, note that the notation $\Delta_h y_{t+h}$ refers to $y_{t+h} - y_{t-1}$ and not $y_{t+h} - y_{t-1}$. Note also that we assume $\Delta s_t = s_t = 0, 1$, i.e., treatment does not happen in two consecutive periods.

[7] The shorthand $O(T^{-1})$ can be interpreted as $bias/T \to 0$ as $T \to \infty$.

Figure 1: *Small-sample biases for LPs estimated in level and long-difference forms*



Notes: The DGP is $y_t = s_t + \rho y_{t-1} + \epsilon_t$ with s_t and e_t standard normals, $\rho \in \{0.95, 1\}$, and a small sample size $T = 100$. Results from 10,000 Monte Carlo replications. See text.

observations. Thus it should be clear that $\mathcal{R}_{sy}(h) = \rho^h$ and hence the AR(1) would be correctly specified. Since Marriott and Pope (1954) we have known about the small-sample downward bias in autoregressive models as is evident also in our simulation, specially as the horizon increases. The LP estimated in levels also exhibits a similar bias. However, the long-difference estimate works to effectively eliminate the bias at all horizons.

What is the intuition for the source of the bias? When considering least-squares based estimators of the parameter of interest using time series data, the bias formula (see, e.g., Stuart and Ord, 2010) can be approximated with the Taylor series expansion, as in Marriott and Pope (1954),

$$E(\hat{\beta}_h) = \beta_h + E\left(\frac{N_t}{D_t}\right) \approx \underbrace{\frac{E(N_t)}{E(D_t)}}_{\substack{\text{usual} \\ \text{approximation}}} \underbrace{- \frac{cov(N_t, D_t)}{E(D_t)^2} + \frac{Var(N_t)E(N_t)}{E(D_t)^3}}_{\text{higher order asymptotic terms}} + O(T^{-3/2}), \quad (13)$$

where $N_t = \frac{1}{T-h}\sum_{t=1}^{T-h} u_{t+h}\, \boldsymbol{x}_t$ refers to the numerator, and $D_t = \frac{1}{T-h}\sum_{t=1}^{T-h} \boldsymbol{x}_t\, \boldsymbol{x}_t'$ refers to the denominator of typical least squares algebra with $\boldsymbol{x}_t = (s_t, y_{t-1})$ for the levels case and $\boldsymbol{x}_t = (s_t, \Delta y_{t-1})$ for the long-difference case. In time series, even if $E(N_t) = 0$, as is typically the case, $cov(N_t, D_t)$ need not be exactly 0 in small samples (though $cov(N_t, D_t) \to 0$ as $T \to \infty$). Long-differencing essentially works to suppress this covariance term, at least here for ρ close to 1. Understanding these patterns of small-sample bias reduction is a goal for further research.

Non-stationary case As is often the case, analysis of the non-stationary case is more complicated, even for the example of the simple model above with $\rho = 1$ shown in panel (b) of Figure 1 when $T = 100$. In the case of the AR(1) estimator it is well known that the estimate $\hat{\rho}_{AR(1)}$ suffers from an $O(T^{-1})$ downward bias, and various approximations have been presented in the literature, with some extensions to higher orders (White, 1957; Evans and Savin, 1981). Thus, when the impulse response at horizon h is then computed via compounding as $\hat{\rho}^h_{AR(1)}$, this well-known bias is propagated forwards and magnified at all horizons.

This problem of bias is clearly seen in the simulations in Figure 1, panel (b). We can also see that it equally contaminates the AR(1) estimate and the LP estimated using the levels specification. The long difference LP specification does not completely eliminate this bias, although it does attenuate it considerably, as noted by Piger and Stockwell (2023).

Large-sample asymptotics Of course, as $T \to \infty$ for fixed h, both levels and long-difference estimates are consistent; they converge to the true impulse response, and the distribution of the estimates is asymptotically normal. But problems might arise when h increases in proportion to sample size T, as $T \to \infty$. For example, in the non-stationary case, Phillips (1998) shows that the distribution of the impulse response coefficients in an autoregressive model is no longer normal, and the estimate is inconsistent. Pesavento and Rossi (2006, 2007) and Mikusheva (2012) develop methods to calculate confidence sets for impulse responses for this case. Though similar issues likely pervade responses estimated with LPs, including possibly also the stationary case, we are not aware of results yet in the literature that comprehensively deal with all of these issues. What about panel data? In a recent paper, Mei, Sheng, and Shi (2023) show that incidental parameter biases (Nickell, 1981) crop up when the dimensions of the panel $N, T \to \infty$ as $N/T \to c$ for $c \in (0, \infty)$. The solution that they propose in the paper is to use the split panel jacknife estimator of Dhaene and Jochmans (2016) and Chudik, Pesaran, and Yang (2018). Denote $\hat{\beta}_h$ the full sample panel estimate with fixed-effects and $\hat{\beta}^a_h$ and $\hat{\beta}^b_h$ estimates based on splitting the sample along the time series dimension into two halves. Then the bias corrected estimate of the impulse response, $\tilde{\beta}_h$, is simply $\tilde{\beta}_h = 2\hat{\beta}_h - \frac{1}{2}(\hat{\beta}^a_h + \hat{\beta}^b_h)$.

4. One-off treatments versus treatment plans: Multipliers

The previous section clarifies the statistical connection between impulse responses estimated with LPs and VARs. In this section instead, we discuss the economic connection and what it means for interpretation of impulse responses. To explain the main ideas, we use a simple model discussed in Alloza, Gonzalo, and Sanz (2019). Consider a setting where the data are generated by the following, highly stylized, structural process

$$\begin{pmatrix} 1 & -\beta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} y_t \\ s_t \end{pmatrix} = \begin{pmatrix} \phi_{yy} & \phi_{ys} \\ \phi_{sy} & \phi_{ss} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ s_{t-1} \end{pmatrix} + \begin{pmatrix} u^y_t \\ u^s_t \end{pmatrix}; \quad E(\boldsymbol{u}_t) = \mathbf{0}; \quad Cov(u^y_t, u^s_t) = 0. \tag{14}$$

Think of Equation 14 as a structural VAR, where s_t is the policy variable and y_t the outcome variable and hence β is the effect on impact of a shock to u_t^s. This simple process allows one to think of four cases of interest.

1. No propagation Suppose $\phi_{ij} = 0$ for any $i, j \in \{y, s\}$. In this case, the intervention $s_t = u_t^s$ is completely exogenous and there is no propagation of the shock. A regression of y_t on s_t would recover the effect of the intervention, $\mathcal{R}_{sy}(0) = \beta$ with $\mathcal{R}_{sy}(h) = 0$ for $h > 0$. In fact, if $s_t \in \{0, 1\}$, then β could be estimated as a simple difference in means between treated and control observations, just like in any *randomized controlled trial* or RCT.

2. Persistent interventions Suppose instead that $\phi_{ss} = \phi \neq 0$, but $\phi_{yy} = \phi_{ys} = \phi_{sy} = 0$. The intervention s_t is still exogenous, though an intervention today is followed by subsequent interventions. Think of it as a *treatment plan*. The impulse response is $\mathcal{R}_{sy}(h) = \beta\phi^h$. However, we may ask, what would be the outcome effect if there were no subsequent interventions? In this example, it is easy to see that the answer would be $\mathcal{R}_{sy}(h|s_{t+1} = \ldots = s_{t+h} = 0) = 0$ for $h > 1$, just as in the *no propagation* case.

Another way to think about this issue is to compute the ratio of the overall outcome and intervention responses, often referred to as the *multiplier* (see e.g., Mountford and Uhlig, 2009; Uhlig, 2010; Ramey, 2016; Ramey and Zubairy, 2018). Intuitively, the multiplier calculates something like an *average effect per intervention*. This turns out to be exact in our example

$$m(h) = \frac{\beta(1 + \phi + \phi^2 + \ldots + \phi^h)}{(1 + \phi + \phi^2 + \ldots + \phi^h)} = \beta = R_{sy}(0)\,. \tag{15}$$

Which should one report, $\mathcal{R}_{sy}(h)$, $\mathcal{R}_{sy}(h|s_{t+1} = \ldots = s_{t+h} = 0)$, or $m(h)$? There is no correct answer as each responds to a different question. $\mathcal{R}_{sy}(h)$ more closely resembles what we are likely to observe in practice following an intervention in s_t and is the typical response reported in macroeconomics. $\mathcal{R}_{sy}(h|s_{t+1} = \ldots = s_{t+h} = 0)$ more directly represents the thought experiment of a one-off intervention, more typical in micro studies of RCTs. In principle, the $\mathcal{R}_{sy}(h)$ can be obtained as the convolution of $\mathcal{R}_{sy}(h|s_{t+1} = \ldots = s_{t+h} = 0)$ and $\mathcal{R}_{ss}(h)$. The multiplier $m(h)$ offers a natural bridge between these two concepts. In this simple example $\mathcal{R}_{sy}(0) = m(h)$ for $h \geq 0$, but in general, this need not be the case.

3. Internal propagation Suppose $\phi_{yy} = \phi \neq 0$ but $\phi_{ys} = \phi_{sy} = \phi_{ss} = 0$. An intervention has an effect on impact and over subsequent periods so that $\mathcal{R}_{sy}(h) = \beta\phi^h$, the same response that we obtained previously! But the logic is different. Assignment is still random since $s_t = u_t^s$ and hence an LP would recover this response, which is the dynamic effect of a one-off intervention. Again, if $s_t \in \{0, 1\}$, one can think of this as a typical RCT and the response can be calculated by comparing the means of treated and control observations at different points in time. Note that the cumulative response is $\mathcal{R}_{sy}^c(h) = \beta(1 + \phi + \ldots + \phi^h)$ whereas $\mathcal{R}_{ss}^c(h) = 1$ since $\mathcal{R}_{ss}(0) = 1$ and is 0 for $h > 0$. Hence the ratio of cumulative responses or

multiplier is $m(h) = \beta(1-\phi^{h+1})/(1-\phi)$ if $|\phi| < 1$. This is a very different number than what we obtained in the *persistent interventions* case since the experiment is a one-off intervention but the process has internal propagation dynamics.

4. **The general case** Without restrictions, there can be feedback from outcomes to future interventions and vice versa. In this case, disentangling the effects of treatment plans and feedback is difficult without specifying a model. However, one way to get a sense of the granular effects of the intervention is to compute the multiplier. We now discuss how to estimate multipliers.

4.1. Multipliers

Consider, as an example, a fiscal policy evaluation exercise. An initial exogenous one dollar of government spending at time t may lead to more than one dollar of output on impact and over subsequent periods. Measured this way, the dollar payoff from such a fiscal intervention might seem large and thus desirable. However, the initial boost to spending is often followed by additional spending in subsequent periods. Thus, the cost-benefit calculus changes substantially when the comparison is of the overall increase in output relative to the overall increase in spending over a given period time. This is the way fiscal multipliers are calculated in, for example, Mountford and Uhlig (2009); Uhlig (2010); Ramey (2016); Ramey and Zubairy (2018).

Specifically, as in Equation 1, let y denote the outcome variable, and s denote the intervention or policy variable, then, using the same notation introduced earlier, the *cumulative multiplier* can be defined as

$$m(h) = \frac{\mathcal{R}^c_{sy}(h)}{\mathcal{R}^c_{ss}(h)}\,. \tag{16}$$

The two cumulative impulse responses could be estimated at each h and then plugged in here to estimate the ratio $m(h)$. However, calculating standard errors for a ratio of random variables is complicated as it requires stochastic approximations, similar to the approximation used in Equation 13. In addition, when $\mathcal{R}^c_{ss}(h) \to 0$, the estimate of the multiplier can become unstable, which can degrade the stochastic approximation and introduce bias in the standard error computation.

Instead, the multiplier can be better calculated in one step from a particular specification of the local projection (as in Ramey, 2016; Ramey and Zubairy, 2018). Define $w^c_{t,h} = (w_t + \ldots + w_{t+h})$ for $w_t = y_t, s_t$ and note that cumulative impulse responses can be calculated from the following LPs,

$$\begin{aligned} y^c_{t,h} &= \beta^c_h s_t + v_{t+h}\,, \\ s^c_{t,h} &= \theta^c_h s_t + \eta_{t+h}\,; \quad h = 0, 1, \ldots, H\,. \end{aligned} \tag{17}$$

Suppose these LPs are estimated using a vector $\boldsymbol{z}_t$ as instruments for s_t (which includes the case where s_t itself is exogenous and thus an instrument for itself). As a start we will consider the just-identified case when there is only one instrument, a scalar z_t, before generalizing. Note that, without loss of generality, we have omitted other controls $\boldsymbol{x}_t$ for simplicity, but by appeal

to the Frisch-Waugh-Lovell theorem, the same derivations would follow.[8] Finally, assume that $E(v_{t+h}, z_t) = E(\eta_{t+h}, z_t) = 0$ for any h, as is expected of an instrumental variable. Later in Section 7 we discuss more precisely the conditions required of instrumental variables for LPs made by Stock and Watson (2018); Plagborg-Møller and Wolf (2022); and Rambachan and Shephard (2019b).

Taking the covariance of both sides of Equation 17 with z_t, it is easy to see that

$$\beta_h^c = \frac{\text{cov}(y_{t,h}^c, z_t)}{\text{cov}(s_t, z_t)}; \qquad \theta_h^c = \frac{\text{cov}(s_{t,h}^c, z_t)}{\text{cov}(s_t, z_t)}; \qquad m(h) = \frac{\beta_h^c}{\theta_h^c} = \frac{\text{cov}(y_{t,h}^c, z_t)}{\text{cov}(s_{t,h}^c, z_t)}. \tag{18}$$

However, by the same logic, the multiplier term of interest, $m(h)$, can be obtained directly from the local projection

$$y_{t,h}^c = m(h)\, s_{t,h}^c + \epsilon_{t+h} \tag{19}$$

estimated using z_t as an instrument for $s_{t,h}^c$ since by multiplying both sides by z_t and taking covariances we obtain

$$\text{cov}(y_{t,h}^c, z_t) = m(h)\, \text{cov}(s_{t,h}^c, z_t)\,,$$

where by assumption $\text{cov}(z_t,\, \epsilon_{t+h}) = 0$. Thus, the direct method in Equation 19 gives the same estimate of $m(h)$ as in Equation 18. Going further, after routine manipulations using the 2SLS estimator, the approach can be generalized to the over-identified case when $\boldsymbol{z}_t$ is a vector of dimension $r > 1$.[9]

In Figure 2 we present an example of this approach based on Jordà and Taylor (2016). The outcome of interest is log real GDP per capita taken from the annual panel Jordà, Schularick, and Taylor (2017) Macrohistory dataset (along with other control variables). The policy shock is a change in the cyclically-adjusted primary balance measured as a share of GDP. This shock is based on a narrative identification of fiscal consolidations for an OECD annual panel from 1978 to 2019 based on data constructed by Guajardo, Leigh, and Pescatori (2014) and updated to 2019. Both the cumulative responses and the multiplier are negative and relatively accurately estimated (based on point-wise basis confidence bands). Further, joint significant tests (to be discussed later) reject the zero null.

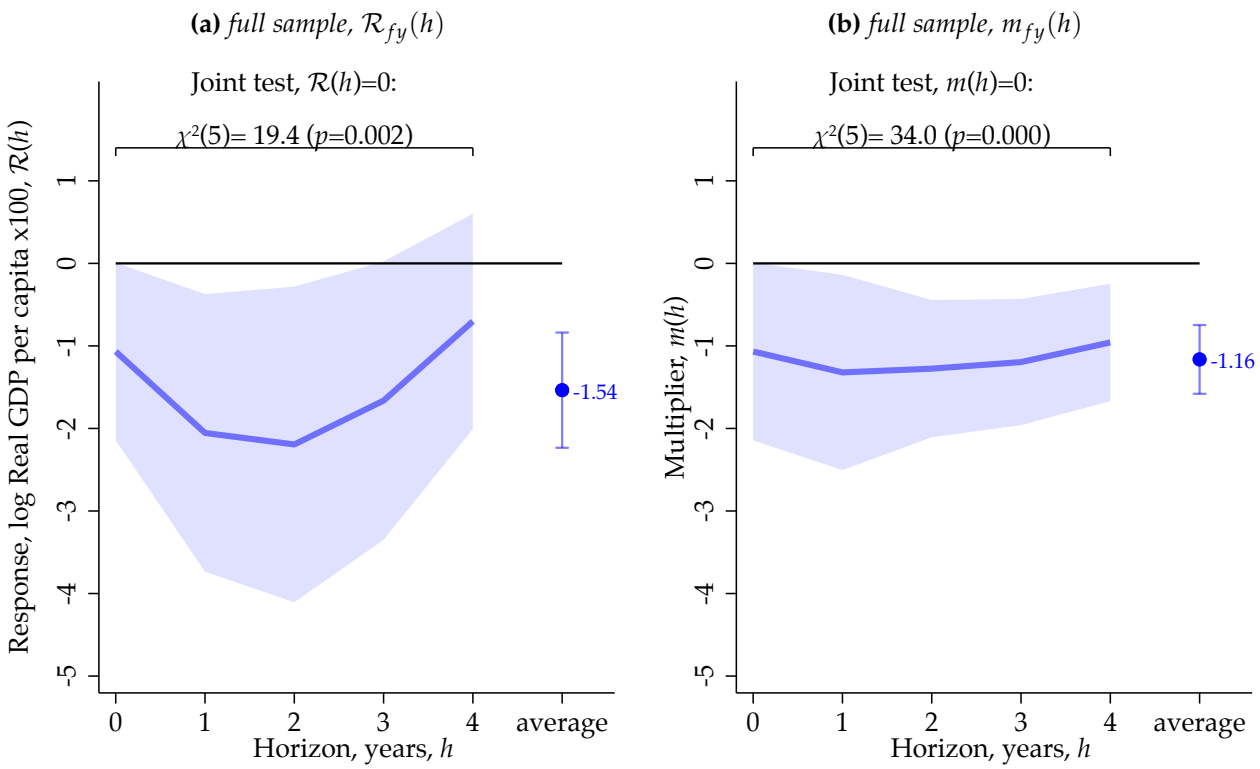
However, whereas the impulse response suggests that the output (real GDP) decline can be as large as 2% for a 1% of GDP fiscal consolidation (seen in years 1 and 2), the multiplier is much more stable, suggesting that there is a dollar for dollar effect: a consolidation that reduces the deficit by

[8] In linear models, one can remove the effect of the additional controls by running a preliminary regression of the outcome and the intervention on the controls. Then one can regress the residuals from these preliminary regressions on each other and obtain the same estimator as when the controls are included directly as right-hand side variables.

[9] By appeal to the Frisch-Waugh-Lovell theorem, we again can project the controls onto the dependent variable, the intervention and the instruments in a first stage and thus write the three relevant 2SLS estimates as (see, e.g., Wooldridge, 2010, chap. 5):

$$\beta_h^c = \frac{\text{cov}(y_{t,h}^c, \Delta\widehat{s}_t)}{\text{cov}(\Delta s_t, \Delta\widehat{s}_t)}; \qquad \theta_h^c = \frac{\text{cov}(s_{t,h}^c, \Delta\widehat{s}_t)}{\text{cov}(\Delta s_t, \Delta\widehat{s}_t)}; \qquad m(h) = \frac{\beta_h^c}{\theta_h^c} = \frac{\text{cov}(y_{t,h}^c, s_{t,h}^c)}{\text{cov}(s_{t,h}^c, \widehat{s}_{t,h}^c)}. \tag{20}$$

Figure 2: *Cumulative fiscal impulse response $\mathcal{R}_{fy}(h)$ and multiplier, $m_{fy}(h)$*



Notes: Outcome y_{it} is log real GDP per capita from Jordà and Taylor (2016), and f denotes a fiscal shock, a treatment Δs_{it} is dCAPB from Guajardo, Leigh, and Pescatori (2014), updated to 2019, instrument z_{it} is GLP2 size of fiscal consolidation Guajardo, Leigh, and Pescatori (2014), updated to 2019. OECD sample, 1978–2019. Control variables are two lags of treatment, two lags of outcome, lag change in the public debt to GDP ratio, and lag of HP-filtered cyclical component of log real GDP per capita. 95% confidence bands are shown and the joint test (see later).

one percent of GDP, reduces output by the same amount.

Finally, the methods used to obtain the fiscal multiplier and presented in Equation 19 are, of course, applicable to other settings. As an example Alessandri, Jordà, and Venditti (2023) calculate financial multipliers from monetary tightenings that depend on the degree of financial market stress.

5. Smoothing

Local projections can be thought of as a semi-parametric method of estimating impulse responses. The advantage is that, as Jordà, Singh, and Taylor (2024) and Plagborg-Møller, Montiel-Olea, Qian, and Wolf (2024) show, they can reduce bias (sometimes considerably) relative to VARs in settings where the lag-length is misspecified, or in settings where the truncation lag (under the assumption that the data are generated by an infinite order process), is relatively short with respect to the impulse response horizon considered. Because LPs do not impose cross-horizon (smoothness) restrictions, as VARs do, this has the advantage of reducing bias at the cost of noisier looking

responses and possibly less precise estimates—the usual bias-efficiency trade-off resulting from imposing fewer restrictions (see, e.g., Li, Plagborg-Møller, and Wolf, 2024). That said and as we shall see, recent research suggests that these trade-offs can result in correct probability coverage when conducting formal inference (Xu, 2023; Plagborg-Møller, Montiel-Olea, Qian, and Wolf, 2024).

Smoothness can be easily restored to an LP response in a variety of ways, however, should one desire. If one views the choppy look of a raw LP response as a symptom of small-sample variation around a presumed true smooth response, a simple option is to use a rolling-window moving average over the response coefficients.[10] This is essentially what a nonparametric kernel estimator of the conditional mean does.

More generally, let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_H)'$ be the $(H+1) \times 1$ vector of LP response coefficients with covariance matrix $\hat{\Omega}_\beta$. We may conjecture that the true, unknown impulse response is described by a smooth function, say, $\beta_h = f(h) : \mathbb{R} \to \mathbb{R}$. In turn, one can approximate smooth functions in a variety of ways, from Taylor series expansions, to splines, to basis function models for supervised learning. These approximations will usually require regularization to prevent overfitting, which in turn involves choosing a tuning parameter, requiring a deeper discussion than this space permits.

Briefly, suppose that $f(h)$ can be approximated as $f(h) \approx \phi(h; \boldsymbol{\Theta}) = \sum_{j=1}^{J} c_j \phi_j(h; \boldsymbol{\theta}_j)$ with $\boldsymbol{\Theta} = (c_1, \boldsymbol{\theta}_1, \ldots, c_J, \boldsymbol{\theta}_J)'$ where $\phi_j(h; \boldsymbol{\theta}_j)$ is a basis function that depends on a small number of parameters, $\boldsymbol{\theta}_j$. We assume that the approximation can be made arbitrarily precise as $J \to \infty$. Ideally, we want to choose $\phi_j(h; \boldsymbol{\theta}_j)$ so that the shape of the response can be well approximated with as few basis functions as possible (sometimes as small as $J = 1$, as we will see). Hence, suppose $\beta_h \approx \phi(h; \boldsymbol{\Theta})$ with $\dim(\boldsymbol{\Theta}) \ll \dim(\boldsymbol{\beta})$. In such a scenario, given estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\Omega}_\beta$ such that $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$; $\hat{\Omega}_\beta \xrightarrow{p} \Omega_\beta$ for Ω_β a positive semidefinite matrix, then estimates of $\boldsymbol{\Theta}$ and Ω_Θ can be easily obtained by minimum distance as the solution to the problem:

$$\min_{\theta} Q(\boldsymbol{\Theta}) = \min_{\Theta} (\hat{\boldsymbol{\beta}} - \phi(\boldsymbol{\Theta}))' \hat{\Omega}_\beta^{-1} (\hat{\boldsymbol{\beta}} - \phi(\boldsymbol{\Theta})), \tag{21}$$

where $\hat{\boldsymbol{\Theta}} \xrightarrow{p} \boldsymbol{\Theta}$ and $\sqrt{T-H}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \xrightarrow{d} \mathcal{N}(0, \Omega_\Theta)$ with $\Omega_\Theta = (\Phi_0' \hat{\Omega}_\beta \Phi_0)^{-1}$ and where $\Phi_0 = \partial \phi(\boldsymbol{\Theta}) / \partial \boldsymbol{\Theta}'|_{\Theta_0}$. Moreover, when $k = \dim(\boldsymbol{\beta}) - \dim(\boldsymbol{\Theta}) > 0$, then

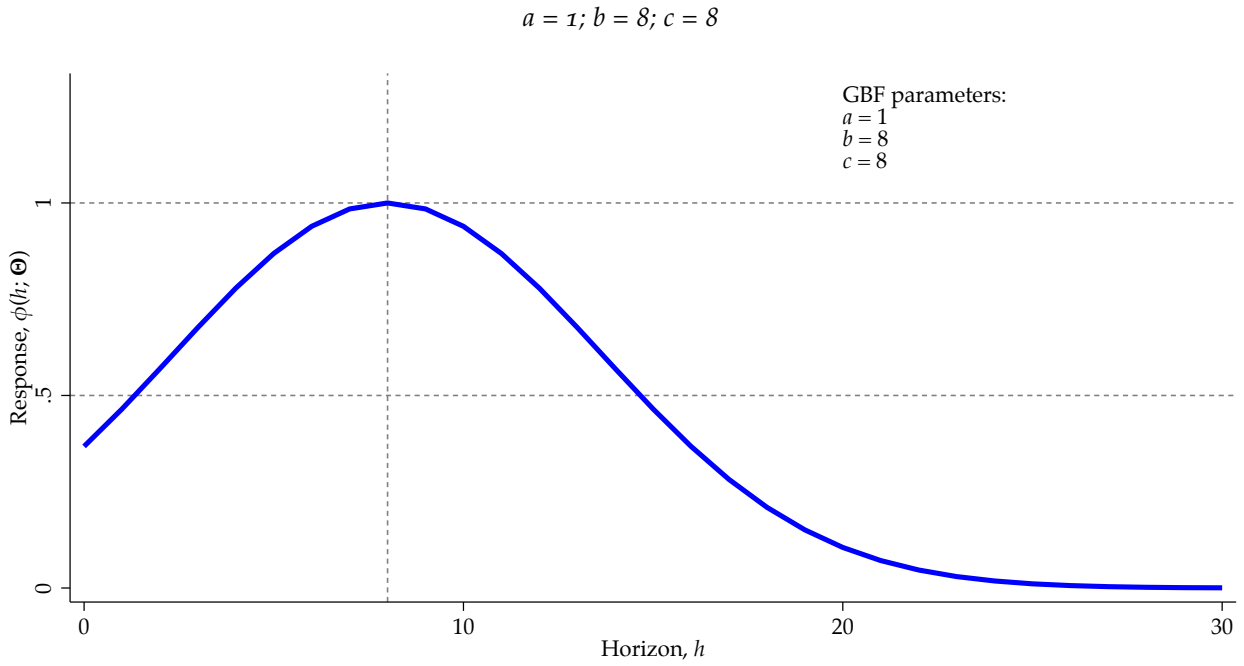
$$Q(\hat{\boldsymbol{\Theta}}) = (\hat{\boldsymbol{\beta}} - \phi(\hat{\boldsymbol{\Theta}}))' \hat{\Omega}_\beta^{-1} (\hat{\boldsymbol{\beta}} - \phi(\hat{\boldsymbol{\Theta}})) \xrightarrow{d} \chi^2_k, \tag{22}$$

which provides an overidentifying restrictions test with which to evaluate the quality of the approximation provided by $\phi(\boldsymbol{\Theta})$. Alternatively, one can write down $\phi(h; \boldsymbol{\Theta})$ as the coefficient of the local projection and estimate $\phi(h; \boldsymbol{\Theta})$ directly using, for example, GMM. This is how Figure 6 is constructed, for example, later in the paper.

Two examples from the literature have received the most attention. In Barnichon and Brownlees

[10] For example, in STATA one could use the `tssmooth` command.

Figure 3: *A typical Gaussian basis function*



(2019), the authors use the B-spline method of Eilers and Marx (1996). B-splines take the form

$$\phi_j(h;\boldsymbol{\theta}_j) = c_{1j}\sin\left(\tfrac{2\pi}{H}j\,h\right) + c_{2j}\cos\left(\tfrac{2\pi}{H}j\,h\right); \quad j = 1,\ldots,J\,.$$

Hence, local projections can be estimated by penalized least-squares as

$$y_{t+h} = \phi_1(h;\boldsymbol{\theta}_1)s_t + \ldots + \phi_J(h;\boldsymbol{\theta}_J)s_t + v_{t+h}\,, \tag{23}$$

where for simplicity, we omit the constant term and $\boldsymbol{x}_t$. The B-spline method can smooth the response over a wide variety of shapes using convenient least-squares methods. However, because the analyst is required to choose a tuning parameter for regularization, the theoretical justification for how to construct standard errors formally has not yet been developed.

In Barnichon and Matthes (2018), the authors use Gaussian basis functions, specifically $\phi(h;\boldsymbol{\Theta}) = a\exp[-((h-b)^2/c^2)]$ for $\boldsymbol{\Theta} = (a,b,c)'$. This single basis function approximates single-humped responses very well. Figure 3 shows an example of what this function looks like. The parameters a, b, and c have an interesting and useful interpretation as is shown in Figure 3. The parameter a measures the height of the "hump"; b measures how many periods from the initial shock until the response reach its peak; and $c\sqrt{\ln 2}$ measures the half-life of the peak response.

As an example below, we will apply this approach later to the response of the unemployment rate to a shock in the policy interest rate in Figure 6, to be discussed in detail later. Importantly, when $H+1 \gg 3$, we will see that there is substantial reduction in the dimension of the parameter vector, resulting in considerably more efficient responses, as Figure 6 shows.

Still, although many typical macroeconomic responses are single-humped in shape, and also take all positive (or all negative) values, i.e., are single-signed, many other responses exhibit more than one hump, or have shapes that shift from positive to negative values and vice versa. Such shapes require expanding the basis function approximation by at least one term, which reduces the benefits of this approach considerably (since the parameters of each basis function are harder to identify separately). In such cases, approximation with B-splines becomes more attractive.

6. Pointwise inference

Unlike more traditional settings in econometrics, we shall see that conducting inference on responses calculated with LPs has a few interesting wrinkles. In this section, after introducing basic ideas, we discuss robust pointwise inference based on adding extra lags in the regression (lag-augmentation) that is uniformly valid over both stationary and non-stationary data and over a wide range of response horizons (Montiel Olea and Plagborg-Møller, 2021). Moreover, when the true lag is unknown and possibly infinite, LPs are semi-parametrically efficient if the controlled lag diverges with the sample, which means that the efficiency loss of local projections relative to other methods vanishes asympotically (Xu, 2023). Plagborg-Møller, Montiel-Olea, Qian, and Wolf (2024) further show that LP confidence intervals are surprisingly robust to misspecification, offering the correct probability coverage relative to VARs that are only mildly misspecified.

As always, and even as with VARs, we must be alert to small-sample problems. Since impulse responses are functions of VAR parameters, they will inherit small-sample biases (see, e.g., Kilian, 1998, 1999, for a VAR bootstrap procedure to correct small-sample inference). We wait until the next section to consider issues of simultaneous inference, when we are interested in characterizing the uncertainty of the impulse response path rather than individual elements of the impulse response.

6.1. The moving-average residual structure of local projections

The basic features of LP inference are easily shown with a simple AR(1) example,[11] such as $w_t = \phi w_{t-1} + u_t$. By repeated substitution, e.g. as in Equation 5 earlier for a VAR(1), the LP is

$$w_{t+h} = \phi^{h+1} w_{t-1} + v_{t+h}\,; \quad v_{t+h} = u_{t+h} + \phi\, u_{t+h-1} + \ldots + \phi^h u_t\,; \quad h = 1, \ldots, H\,. \tag{24}$$

Thus, the regression of w_{t+h} on w_{t-1} will have serially dependent residuals (though dated $t+h$ thru t rather than depending on past values), in this case, a moving-average of order h or MA(h). A simple solution proposed by Jordà (2005) is to use a heteroskedasticity and autocorrelation consistent (HAC) covariance estimator, such as Newey-West (Newey and West, 1987). This semi-parametric correction obviates the need to assume that the particular dependence of the residuals is known.

[11] We omit the constant term, deterministic trends and other features to keep the exposition simple.

Much of the literature appears to follow a similar strategy, even when it comes to panel data, where the Driscoll-Kraay (Driscoll and Kraay, 1998) covariance estimator is used instead.

6.2. The basic issues

In this section and the next we take a very stylized model to explain the main issues. In particular, suppose the DGP is a simple AR(1) model given by

$$w_t = \phi w_{t-1} + u_t\,; \quad t = 1, \dots, T\,; \quad w_0 = 0\,, \tag{25}$$

where u_t is strictly stationary and we further assume $E(u_t|\{u_s\}_{s\neq t}) = 0$ *almost surely*. We make this assumption to follow the setup in Montiel Olea and Plagborg-Møller (2021), later used to present estimation of LPs with lag-augmentation. Using also the notation in that paper as well, let $\beta(\phi, h)$ denote the LP parameter used to estimate the impulse response ϕ^h, that is

$$w_{t+h} = \beta(\phi, h)\, w_t + \xi_t(\phi, h)\,; \qquad \xi_t(\phi, h) \equiv \sum_{l=1}^{h} \phi^{h-l}\, u_{t+l}\,. \tag{26}$$

As we remarked earlier, the moving average form of $\xi_t(\phi, h)$ led Jordà (2005) to recommend HAC-robust standard errors. In addition, note that when $\phi \to 1$, $\hat{\beta}(\phi, h)$ will have a near-unit root distribution. The resulting downward bias in the estimator of the impulse response is well-known (see, e.g., Pesavento and Rossi, 2006, 2007, in the context of impulse responses estimated with VARs with roots near to unity).

Near-to-unity asymptotic results indicate that inference based on critical normal values will not be valid uniformly over all values of $\phi \in [-1, 1]$ even for fixed h. However, when ϕ is inside the stationary region, the LP estimator is asymptotically normal.[12]

6.3. Lag-augmented local projections

A simple extension to the traditional local projection estimator turns out to simplify inference considerably. In particular, Montiel Olea and Plagborg-Møller (2021) suggest adding w_{t-1} as an additional regressor to Equation 26. The purpose of this *lag augmentation* is to make the effective regressor of interest stationary even if the data w_t has a unit root. Montiel Olea and Plagborg-Møller (2021) show that, with rearranging, the lag-augmented local projection can be written as

$$w_{t+h} = \beta(\phi, h)\, u_t + \beta(\phi, h+1) w_{t-1} + \xi_t(\phi, h)\,. \tag{27}$$

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

[12] In order to stay in the strictly stationary region, we may assume $\phi = 1 - c_T/T$ such that $2 > c_T/T > 0$ as $T \to \infty$, for example.

between w_t and u_t, the feasible local projection onto (w_t, w_{t-1}) provides an estimate of $\beta(\phi, h)$ precisely equal to the one that would be obtained from the projection onto (u_t, w_{t-1}). Thus, the actual regression to be estimated is

$$w_{t+h} = \beta(h)w_t + \beta(h+1)w_{t-1} + \xi_t(h) \quad \rightarrow \quad \hat{\beta}(h), \hat{\xi}_t(h)\,. \tag{28}$$

Lag-augmentation has two benefits. As Montiel Olea and Plagborg-Møller (2021) show, the distribution of $\hat{\beta}(h)$ of this feasible lag-augmented local projection is uniformly normal in $\phi \in [-1, 1]$ using similar arguments as lag-augmentation in AR inference (see, e.g., Sims, Stock, and Watson, 1990; Toda and Yamamoto, 1995; Dolado and Lütkepohl, 1996; Inoue and Kilian, 2002, 2020). The second benefit is that it simplifies the computation of standard errors.

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 26? From Equation 27, note that u_t was assumed to be uncorrelated with past and future values of itself, and therefore the regression *score* $\xi_t(\phi, h)u_t$ is serially uncorrelated. To see this, note that the standard error formula in the ideal regression of Equation 27 would be

$$\hat{\mathbf{s}}_h = \frac{(\sum_{t=1}^{T-h} \hat{\xi}_t(\phi, h)^2 \hat{u}_t^2)^{1/2}}{\sum_{t=1}^{T-h} \hat{u}_t^2}\,. \tag{29}$$

But by similar linearity arguments used to justify the feasible augmented local projection, it can be calculated directly from Equation 28 using White corrected standard errors as indicated. In addition, Montiel Olea and 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 $h_T/T \to 0$).

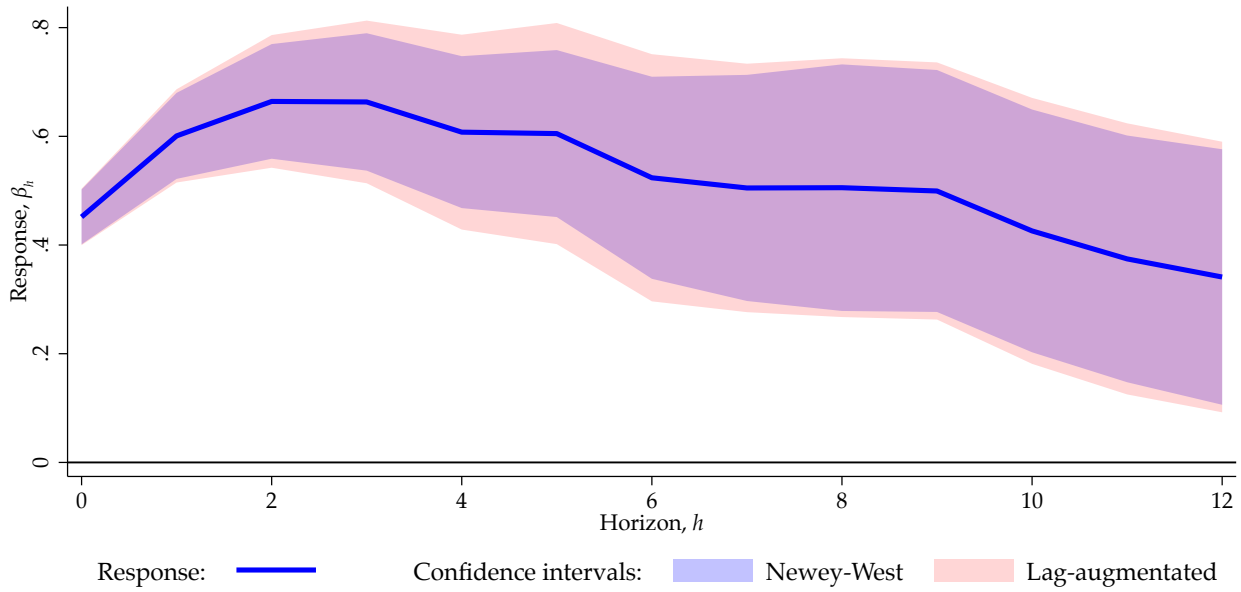
Moreover, lag-augmentation is shown to work more generally when the DGP is assumed to be a VAR(p) or a vector error correction model (VECM), though we are not aware that similar results have been derived for panel data in settings where the time dimension is larger than the cross section dimension, i.e., $T \gg N$. 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. Naturally, lag-augmentation can also be applied to identified LPs (Plagborg-Møller and Wolf, 2021; Montiel Olea and Plagborg-Møller, 2021).

As an illustration of how Newey-West and lag-augmented confidence intervals compare, Figure 4 shows the results from a simple simulation based on the bivariate model

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} 0.7 & 0.2 \\ 0.2 & 0.7 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} u_t^y \\ u_t^x \end{pmatrix}; \qquad u_t^y = e_t^y + e_t^x\,, \quad u_t^x = e_t^x\,; \quad e_t^y, e_t^x \sim N(0,1)\,,$$

with a sample of 300 observations (after disregarding 500 initial observations). The figure shows that

Figure 4: *Comparing Newey-West versus lag-augmented confidence bands*



Notes: Data generated from a bivariate VAR(1). The simulated sample size is 300 observations after disregarding 500 initialization observations. Response shown with Newey-West (blue shaded region) versus lag-augmented (orange shaded region) 95% confidence bands. See text.

both methods generate similar confidence intervals. In fact, several experiments (not 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 and Plagborg-Møller (2021) provide bootstrap procedures that we briefly sketch here though the reader should go to the original source for details. Suppose that you want to provide inference for an impulse response estimated with lag-augmented LPs for which you also obtain the standard error as described earlier (i.e., using White corrected standard errors). Montiel Olea and Plagborg-Møller (2021) then suggest estimating the corresponding VAR(p).[13] 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 and 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.

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

6.4. Robustness

In a recent paper, Plagborg-Møller, Montiel-Olea, Qian, and Wolf (2024) provide analytical results showing that conventional local projection inference is surprisingly robust to large amounts of misspecification when the data are generated by a local-to-VAR process[14] In contrast, VAR confidence intervals vastly undercover even with small misspecifications that are hard to detect in practice. VARs are generally specified with too few lags.

Intuitively, a VAR parsimoniously approximates the DGP from which an impulse response is then derived. This results in better mean-squared error (MSE) properties and smoother looking impulse responses. In contrast, a local projection approximates the impulse response itself. This results in lower bias, though possibly higher MSE (Li, Plagborg-Møller, and Wolf, 2024). However, LPs will result in valid confidence intervals and are, in that sense, superior to VARs from a robustness standpoint. Moreover, LPs can also be smoothed, if desired, as we previously discussed in Section 5.

7. Joint inference

Impulse responses describe the trajectories of outcome variables following an intervention. Getting a sense of the uncertainty about the shape of the estimated impulse response is akin to a multiple hypothesis test. Because estimates of the response coefficients are correlated (except under the previous null), it is not enough to rely on individual hypothesis tests. In fact, this correlation can generate wide point-wise bands even when the joint null of significance is soundly rejected, much like classic regression with collinearity.

Figure 5 illustrates these issues. The experiment in the figure consists of an intervention measured by a Romer and Romer (2004) monetary shock (extended to 2007Q4)[15] where the response of interest is the cumulative change in the log level of the Consumer Price Index (CPI) in Figure 5a, and the rate of inflation (i.e., the first difference) in Figure 5b, using as controls two lags of CPI inflation, two lags of real GDP growth, and two lags of the federal funds rate.

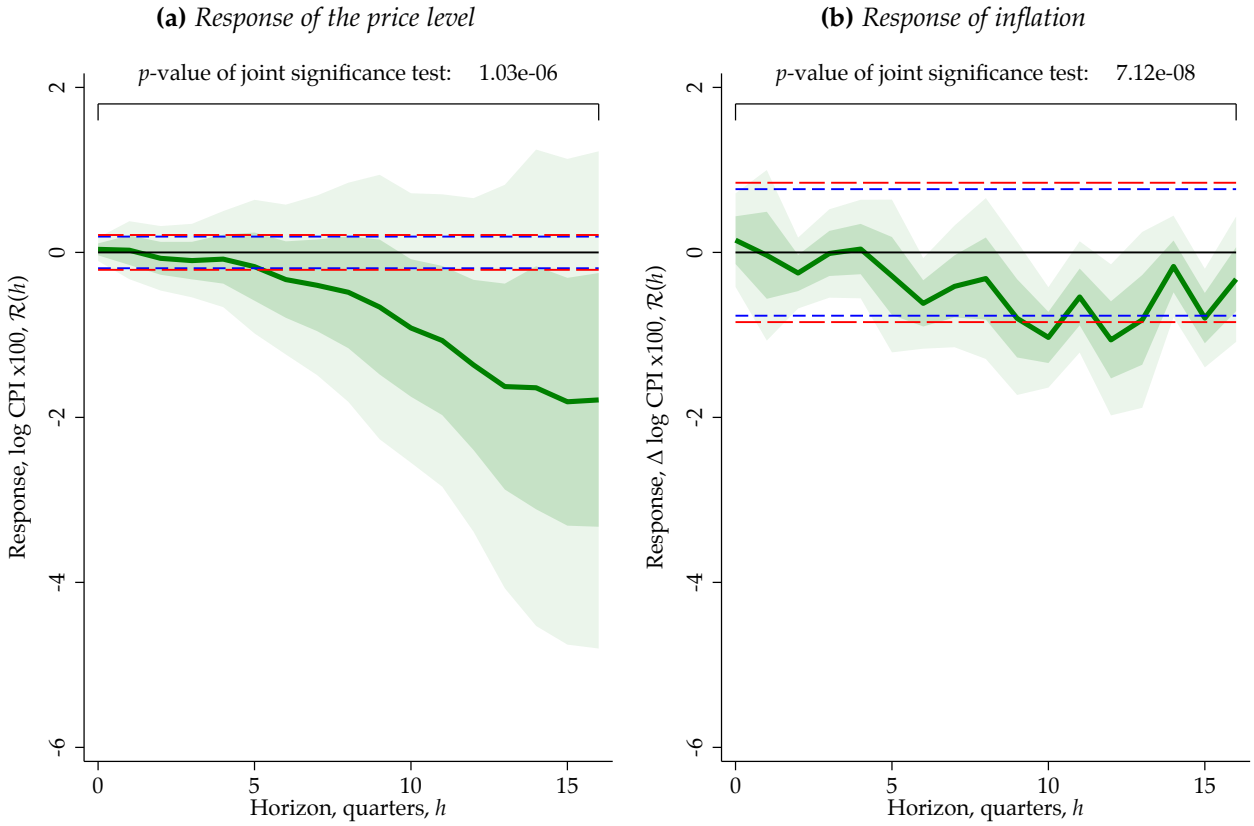
Figure 5a shows that in response to a 1 percentage point Romer and Romer (2004) monetary shock, the price level does not respond for about two years. Thereafter, the price level begins to decline. At the three-year mark, prices are 1.13% lower than at the start (or a rate of deflation of about 0.44% per year). However, point-wise error bands suggest that these dynamics are not statistically significant for any of the three years (12 quarters) displayed. Figure 5b is noisier but suggests that *changes* in the rate of inflation were negative and different from zero in several of the responses after the 2-year mark approximately.

What is the correct interpretation of the evidence? If we recognized that the price level remained unchanged for about 2 years, would the right conclusion be that thereafter inflation remained

[14]Meaning, a DGP that is approximately a VAR with moving-average terms that are "small" in the asymptotic sense.

[15]Data extended by Wieland and Yang (2020)

Figure 5: *Response of the price level and inflation to a Romer-Romer monetary shock*



Notes: The outcome is the cumulative change of 100 times the log of the consumer price index (CPI) in the left panel and the first difference of the same variable in the right panel. We used two lags of CPI inflation, six lags of real GDP growth, and two lags of the federal funds rate as additional controls. The intervention is a Romer and Romer (2004) shock (extended to 2007). Shaded areas are one and two standard deviation pointwise confidence bands using heteroscedasticity robust standard errors. Dashed lines are point-wise 95% *significance bands* (see later). The blue dashed lines are obtained analytically whereas the red long-dashed lines are obtained using the wild block bootstrap. Sample: 1969Q1–2007Q4. See text.

unchanged as well? One would expect that if monetary policy had no effect on prices, we would see, with roughly equal probability, positive and negative values of the price response. Cleary, this is not the case. Let's find out how best to proceed.

7.1. Local projections as a GMM problem

When hypothesis tests involve LP parameters over several horizons (as the next section does), it is necessary to estimate LPs as a system to obtain the appropriate covariance matrix. We illustrate how to do this using the Generalized Method of Moments or GMM. We present the main results using the outcomes and the intervention only, which you can think of as having been previously projected onto the controls (and relying on the Frisch-Waugh-Lovell theorem and a linear model). This allows us to focus on the important results with minimal extra notation.

Let $\boldsymbol{y}_t(H) = (y_t, \ldots, y_{t+h})'$ be an $(H+1) \times 1$ vector collecting all the left-hand side outcome variables. Construct $S_t = I_{H+1} \otimes s_t$ where I_{H+1} is the identity matrix of order $H+1$, and s_t is the intervention.[16] Collect all the error terms in $\boldsymbol{v}_t(H) = (v_t, \ldots, v_{t+H})'$. Let $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_H)'$ collect all the impulse response coefficients. Finally, we entertain the possibility that we have $l \geq 1$ external instruments in the $1 \times l$ vector $\boldsymbol{z}_t$ and thus we construct $Z_t = I_{H+1} \otimes \boldsymbol{z}_t$. In the absence of instruments, if one can appeal to identification based on selection-on-observables arguments (or if the interventions are exogenous), one can simply set $Z_t = S_t$.

Using these definitions, the population moment condition for the system of $H+1$ local projections is

$$E\left[Z_t'(\boldsymbol{y}_t(H) - S_t\boldsymbol{\beta})\right] = 0\,. \tag{30}$$

Thus, the corresponding sample GMM problem can be specified as

$$\min_{\boldsymbol{\beta}} \left[\frac{1}{N}\sum_{t=p+1}^{T-H} Z_t'(\boldsymbol{y}_t(H) - S_t\boldsymbol{\beta})\right]' \hat{\Lambda}^{-1} \left[\frac{1}{N}\sum_{t=p+1}^{T-H} Z_t'(\boldsymbol{y}_t(H) - S_t\boldsymbol{\beta})\right], \tag{31}$$

with $N = (T-H) - (p+1)$. When choosing $\hat{\Lambda} = I_{H+1}$, the estimator is referred to as the equally-weighted estimator and yields consistent estimates of $\boldsymbol{\beta}$. However, the optimal weighting matrix is

$$\hat{\Lambda} = \frac{1}{N}\sum_{t=p+1}^{T-H} Z_t'\tilde{\boldsymbol{v}}_t(H)\tilde{\boldsymbol{v}}_t(H)'Z_t\,, \tag{32}$$

where $\tilde{\boldsymbol{v}}_t(H)$ refers to the residuals based on the equally weighted estimator. We do not enter into a discussion of two-step versus optimally iterated estimators of the weighted matrix (and hence the estimates of $\boldsymbol{\beta}$), for which a discussion can be found in traditional textbooks such as Cameron and Trivedi (2005) and Wooldridge (2010).

Before proceeding further and as a preview of our discussion on identification with external instruments, we discuss the assumptions needed. These differ from the usual relevance and exogeneity conditions, as Stock and Watson (2018) and Plagborg-Møller and Wolf (2022) show. As a reminder, we have omitted control variables for simplicity so the assumptions we are about to state should be understood to be for instruments, interventions, outcomes, and residuals projected onto explanatory variables other than those in S_t. Hence, we assume:

Assumption 1

- ***Relevance:*** $E(Z_t' S_t) \neq \mathbf{0}$;
- ***Lead-lag exogeneity:*** $E(Z_{t+l}'\,\boldsymbol{v}_t(H)) = 0 \;\; \forall\, l$.

[16] The notation $\otimes$ refers to the Kronecker product.

It is important to note that these conditions are derived from rather standard assumptions about the DGP. In recent work, Rambachan and Shephard (2019b) provide conditions based on a flexible, fully nonparametric foundation using a potential outcome time series framework. These conditions are too technical for this review and we refer the reader to their paper for more details.

Based on Assumption 1 and relatively general conditions, the estimate of the impulse response $\mathcal{R}_{sy} = \boldsymbol{\beta}$ can be obtained from

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{N} \sum_{p+1}^{T-H} S_t' Z_t \hat{\Lambda}^{-1} Z_t' S_t \right)^{-1} \left(\frac{1}{N} \sum_{p+1}^{T-H} S_t' Z_t \hat{\Lambda}^{-1} Z_t' \boldsymbol{y}_t(H) \right), \tag{33}$$

which will be consistent and asymptotically normal with approximate covariance matrix given by

$$\hat{\Omega}_\beta = \left[\left(\frac{1}{N} \sum_{p+1}^{T-H} S_t' Z_t \right) \hat{\Lambda}^{-1} \left(\frac{1}{N} \sum_{p+1}^{T-H} Z_t' S_t \right) \right]^{-1}. \tag{34}$$

Usually, a correction for heteroskedasticity and autocorrelation would be indicated and this can accomplished as is common with, for example, a Bartlett correction such that

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

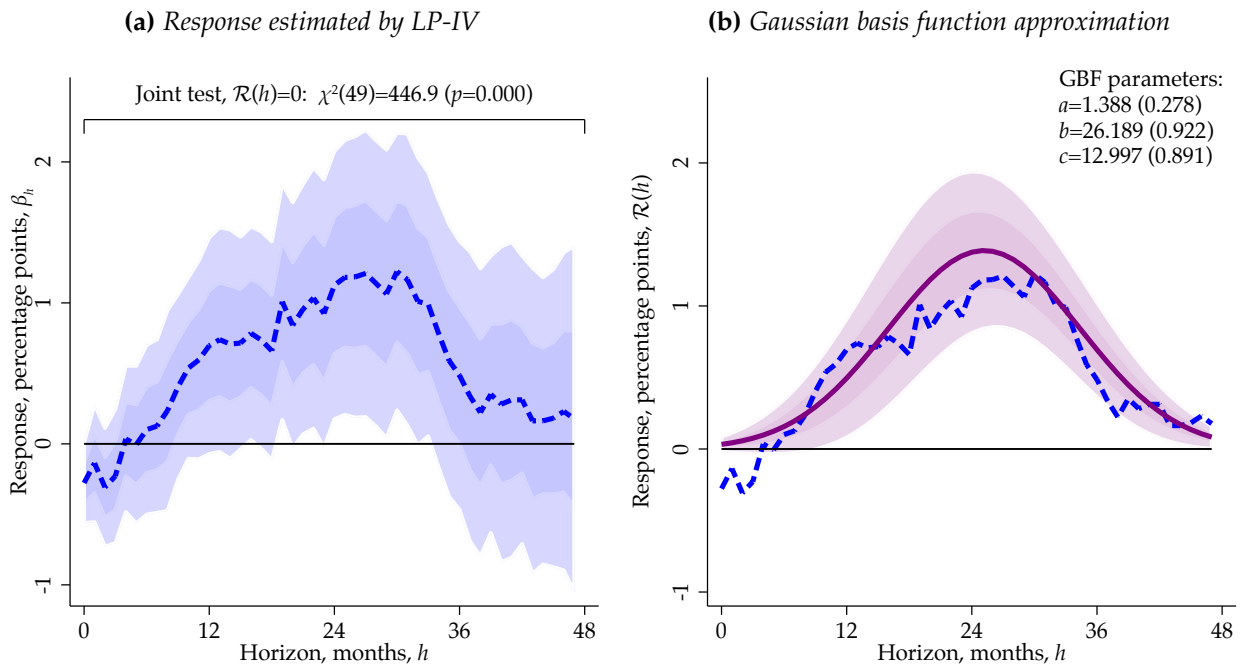
One may conjecture that this traditional form of the covariance matrix would be correct with lag-augmentation, though this particular result is not specifically proven in Montiel Olea and Plagborg-Møller (2021). Thus, standard system estimators using instrumental variables can be used to obtain these results (including corrections for heteroskedasticity and autocorrelation). An estimate of Ω_β plays an important role in the next two sections and for simultaneous inference of the impulse response in general.

The left panel of Figure 6 provides an example of system LPs estimated by GMM. The estimates shown are the response of the U.S. unemployment rate when using a Romer-Romer shock as an instrument for the federal funds rate.

7.2. Simultaneous inference

There will be times when practitioners are interested in assessing the impulse response over a subset range of periods. As a case in point, Figure 6 shows the response of the unemployment rate to a monetary shock using monthly data. The response of the unemployment rate to a monetary shock has an almost perfect bell shape. The unemployment rate gradually increases to about 1¼%, approximately two years after impact, and returns back to zero about four years after impact. Hence one may be interested in assessing the significance of the change in the unemployment rate between years 1 and 3 (or between 12 to 36 months), say.

Figure 6: *Response of the unemployment rate to a Romer-Romer monetary shock*



Notes: Local projection estimated using the Romer-Romer shock as an instrument for the federal funds rate. The local projection contains 6 lags of the funds rate, the unemployment rate, and PCE inflation. The intervention is the Romer and Romer (2004) shock. Left panel shows 95% pointwise confidence bands using heteroscedasticity robust standard errors based on the raw LP estimates based on GMM. The right panel shows the same response next to the fitted Gaussian basis function using GMM. The standard errors shown are directly calculated using GMM. Sample: 1985M1–1999M12. See text.

Or we can return to the example shown in Figure 5, where we may be interested in assessing the significance of the inflation response after 3 years (12 quarters). Properly speaking, point-wise error bands will not provide the correct coverage to make these assessments. The correct approach is therefore to construct error bands that account for the simultaneous nature of the implied hypotheses. This problem was pointed out in Jordà (2009), who provided a solution based on Scheffé's multiple comparison approximation (see also Wolf and Wunderli, 2015, for an application of the same idea to direct forecasts.).

More recently, Montiel-Olea and Plagborg-Møller (2019) propose a different approximation based on the sup-t procedure that can be easily implemented by simulation methods, the bootstrap, or using a Bayesian approach. Asymptotically, the sup-t procedure is shown to produce the narrowest bands of other commonly used methods of simultaneous inference, including Scheffé's. That said, these bands will tend to be relatively conservative since they accommodate unspecified nulls. Naturally, when the investigator proposes a specific null, an appropriate test can be constructed and inverted to generate narrower bands.

For now, we introduce the basics of the sup-t bands. Let $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_H)'$. Under relatively

general conditions, $\hat{\boldsymbol{\beta}} \to N(0, \Omega_\beta)$. Hence, we then define the one-parameter confidence band,

$$\hat{\mathcal{B}}(c) = [\hat{\beta}_0 - \hat{\sigma}_0 c, \hat{\beta}_0 + \hat{\sigma}_0 c] \times \ldots \times [\hat{\beta}_H - \hat{\sigma}_H c, \hat{\beta}_H + \hat{\sigma}_H c] = \bigcap_{h=0}^{H} [\hat{\beta}_h - \hat{\sigma}_h \hat{q}_{1-\alpha}, \hat{\beta}_h + \hat{\sigma}_h \hat{q}_{1-\alpha}], \quad (35)$$

where it has been traditional to choose $c = 1.96$ for point-wise 95% confidence intervals. However, in order to account for all possible joint tests that a practitioner may want to implement (as in the examples described earlier based on Figure 5 and Figure 6), notice that

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

The distribution of $\max_{h=0,\ldots,H} \left|\sigma_h^{-1} V_h\right|$ in unknown but it is easy enough to simulate any desired quantile of this distribution.

Accordingly, we can adapt the algorithm proposed in Montiel-Olea and Plagborg-Møller (2019) to our LP problem as follows:

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 $\hat{q}_{1-\alpha}$ as the empirical $1-\alpha$ quantile of $\max_h \left|\sigma_h^{-1} \hat{V}_h^{(m)}\right|$ across $m = 1, \ldots, M$.
- **Step 3:** Then $\hat{\mathcal{B}}(\hat{q}_{1-\alpha}) = \bigcap_{h=0}^{H} [\hat{\beta}_h - \hat{\sigma}_h \hat{q}_{1-\alpha}, \hat{\beta}_h + \hat{\sigma}_h \hat{q}_{1-\alpha}]$.

Note that the second step in this algorithm can be substituted easily when bootstrap/Bayesian methods are used. These extensions are discussed in Montiel-Olea and Plagborg-Møller (2019).

7.3. Significance bands

In a traditional randomized controlled trial the key hypothesis of interest is whether the treatment is effective. In other words, whether there is sufficient statistical evidence against the null hypothesis that the treatment has zero effect on the outcome. Similarly, in plotting an impulse response, the researcher often wants to assess whether the response is statistically different from zero. In the previous section we argued that, because response coefficients are correlated and this is a joint hypothesis test, relying on point-wise confidence intervals would provide incorrect probability coverage for our hypothesis. The previous sections show how to provide bounds to address this problem.

However, is there a better approach? The problem of assessing the significance of an impulse response is a familiar one. Think about the way we assess the autocorrelation function in a time series. A plot of the autocorrelations at different horizons is accompanied by two confidence interval, or error bands, on either side of the zero line. These *significance bands* straddle the null hypothesis and are, themselves, constructed under the null. If the autocorrelation at a given horizon strays

outside the significance bands, we conclude that such a coefficient must be different from zero, in a statistical sense.

The key insight is that one can use the *Lagrange multiplier* (LM) principle to simplify the construction of the error bands. In fact, the error bands for the autocorrelation function do not depend on the variance of the data, only on the sample size. For a 95% confidence, the bands are $\pm 1.96/\sqrt{T}$. We can apply the same logic to think about the significance of the impulse response and hence construct *significance* bands. The approach described next relies on Inoue, Jordà, and Kuersteiner (2024). A simple example conveys the intuition and is easy to generalize.

Consider the local projection $y_{t+h} = \beta_h s_t + v_{t+h}$ for $h = 0, \ldots, H$. The controls are omitted for simplicity though we can again exploit the Frisch-Waugh-Lovell theorem and then we can think of y_{t+h} and s_t as the result of having projected out the controls. Further suppose that we have an instrument, z_t (we could have more than one, but the main results are more easily grasped with a single instrument), such that it is relevant and meets the lead-lag exogeneity condition in, e.g., Stock and Watson (2018).

Let $n = T - h$, then the asymptotic distribution of the instrumental variable estimator for the LP can be derived as usual, in particular,

$$\sqrt{n}(\hat{\beta}_h - \beta_h) = \frac{n^{-1/2}\sum_{t=1}^{n} z_t y_{t+h}}{n^{-1}\sum_{t=1}^{n} z_t s_t}\,. \tag{36}$$

Under relatively general conditions,

$$\frac{1}{n}\sum_{t=1}^{n} z_t s_t \xrightarrow{p} E(z_t s_t) = \gamma_{zs}\,; \qquad \frac{1}{n^{1/2}}\sum_{t=1}^{n} z_t y_{t+h} \xrightarrow{d} N(0, g)\,. \tag{37}$$

That is, the denominator will converge in probability to its population moment and the numerator will be driving the asymptotic distribution. Next, using the LM principle, we will exploit the null hypothesis to simplify how g is calculated. Specifically, we can note that

$$\begin{aligned} g = \text{Var}\left(\frac{1}{n^{1/2}}\sum_{t=1}^{n} z_t y_{t+h}\right) &\approx \sum_{j=-\infty}^{\infty} E(z_t y_{t+h} z_{t-j} y_{t+h-j}) \\ &= \sum_{j=-\infty}^{\infty} E(z_t z_{t-j}) E(y_{t+h} y_{t+h-j}) \\ &= \sum_{j=-\infty}^{\infty} \gamma_z(j)\gamma_y(j)\,, \end{aligned} \tag{38}$$

where we exploit the assumption that under the null, s_t and y_{t+h} are unrelated and hence so are z_t and y_{t+h}. Thus, based on Equation 37 and Equation 38 we have that

$$\sqrt{n}(\hat{\beta}_h - 0) \xrightarrow{d} N(0, \sigma_h^2)\,; \quad \sigma_h^2 = \frac{\sum_{j=-\infty}^{\infty} \gamma_z(j)\gamma_y(j)}{\gamma_{zs}^2} = \frac{g}{\gamma_{zs}^2}\,. \tag{39}$$

In practice, g cannot be directly estimated as it involves infinite terms but it can be approximated with a Bartlett correction, such as with the Newey-West estimator.

The intuition for this result can be best provided with a simple example. Consider the univariate AR(1) case, where $z = s = y$, then $g = \gamma_y^2$, the variance of y, which is the same as the denominator since $\gamma_{zs}^2 = \gamma_y^2$ and therefore $\sigma_h^2 = 1$. In that case, the LP is simply the estimate of the autocorrelation function. Hence, Equation 39 recovers the well-known bands for the autocorrelogram of y. In the special case where y is a white noise then $\sqrt{n}(\hat{\beta}_1 - 0) \to N(0,1)$, which leads to the well-known case in which the asymptotic 95% confidence bands are $\pm 1.96/\sqrt{n}$. Figure 5 provides an illustration by showing the usual confidence bands alongside the significance bands just described in the general setting.

The construction of significance bands can be summarized as follows:

Practical construction of significance bands

- **Step 1:** Calculate the sample average of the product $s_t z_t$. Call this $\hat{\gamma}_{sz}$.
- **Step 2:** Construct the auxiliary variable $\eta_t = y_t z_t$. Then regress η_t on a constant. The Newey-West estimate of the standard error of the intercept of this auxiliary regression is then an estimate of $g^{1/2}$. Call this $\hat{g}^{1/2}$.
- **Step 3:** Hence, an estimate of $\sigma_h/\sqrt{n}$, is then $\hat{s}_\beta = \hat{g}^{1/2}/\hat{\gamma}_{sz}$.
- **Step 4:** Construct the significance bands as: $[\xi_{\alpha/2H}\hat{s}_\beta, \xi_{1-\alpha/2H}\hat{s}_\beta]$ where $\xi_{\alpha/2H}$ is the critical value of a standard normal random variable at significance level $\alpha/2H$. We use the Bonferroni value $2H$ rather than 2 since we account for the significance of all the coefficients in the response simultaneously.

Inoue, Jordà, and Kuersteiner (2024) also provide a complementary wild block bootstrap procedure that is easy to implement in practice. We refer readers to that paper for more details.

7.4. Summary of best practices: Inference

As discussed earlier, inference in dynamic settings can be tricky. In small samples, serial correlation can generate estimation biases. This is true whether one estimates impulse responses with LPs or with VARs, as the literature has showed (see, e.g., Pope, 1990; Kilian and Lütkepohl, 2017; Piger and Stockwell, 2023; Herbst and Johannsen, 2024). The presence of unit roots or near unit roots can also make inference complicated (see, e.g., Pesavento and Rossi, 2006, 2007). However, as Piger and Stockwell (2023) show, small sample biases appear to be considerably reduced when using long-differencing (as we saw in Figure 1).

In this section we brought several new points to the fore. First, error bands constructed by inverting point-wise t-ratios (as the literature currently does) should be understood as providing a sense of the precision with which each coefficient is estimated. Like a typical regression with nearly collinear regressors, standard errors for individual coefficients can be quite large, even when an

F-test would overwhelmingly reject the null that they are jointly zero. Since a common hypothesis in any impulse response analysis is to assess whether the response is statistically different from zero, we think current practice could be extended to display significance bands alongside error bands.

Second, when estimating LPs using individual regressions, as is often done in empirical practice, estimation of standard errors with lag-augmented specifications and White corrected standard errors provide a simple solution with correct uniform probability coverage under a wide range of scenarios (stationarity, near unit roots, non-stationarity) and even for long distance horizons (as long as the sample size is large enough relative to the horizon). Moreover, these standard errors compare well with those based on VAR impulse responses.

Third, more conservative bounds based on the sup-t method can be reported instead of pointwise error bands when one is interested in providing a summary graphical representation that the reader can use to assess different multiple hypotheses of interest (usually relating to the significance nulls over subsets of horizons).

Fourth, of course, any formal multiple hypothesis test can be constructed using an estimate of the covariance matrix of the response coefficients. This can be done by setting up the simultaneous GMM problem as we showed earlier, which can be based on multiple instrumental variables (as we will discuss below in Section 8). Finally, there are other alternatives currently being developed. Lusompa (2018) proposes a feasible GLS procedure where the idea is to parametrically adjust for the moving-average structure of the residuals using the residuals from the first local projection and estimates of subsequent impulse response coefficient estimates. Lusompa (2018) also provides results based on a time-varying parameter Bayesian approach. Following on this last line of research, Tanaka (2020); Ferreira, Miranda-Agrippino, and Ricco (2023) provide Bayesian estimation routines for LPs and hence inferential procedures based on the posterior distribution of these estimators.

8. Causality

Local projections, by themselves, do not solve the problem of identification or rather, the ability to uncover causal relations. In this section we visit available methods to move the analysis from correlation to causation. The definition of an impulse response in Equation 1 (repeated here for convenience) consists of a counterfactual difference in mean outcomes

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

where the key to identification is to establish how interventions in s_t are determined. In practice, s_t may not be randomly assigned (to use the potential outcomes language), or it may not be *exogenously* determined outside the model. Up to this point, most of the presentation has set aside this issue, which we now tackle head on.

8.1. Selection-on-observables: LP-OLS

A simple and common approach to identification is *selection-on-observables*: that is, to assume that conditional on $\boldsymbol{x}_t$, variation in s_t is as good as random. Suppose for a moment that $s_t \in \{0,1\}$ and that s_t is randomly assigned as it would be in a randomized control trial. In that case, the $\boldsymbol{x}_t$ play no role in achieving identification (though they improve efficiency) and the the impulse response $\mathcal{R}_{sy}(h) = E[y_{t+h}|s_t = 1] - E[y_{t+h}|s_t = 0]$, could be directly estimated as:

$$\hat{\mathcal{R}}_{sy}(h) \;=\; \frac{\sum_{t=h}^{T} y_{t+h} s_t}{\sum_{t=h}^{T} s_t} \;-\; \frac{\sum_{t=h}^{T} y_{t+h}(1-s_t)}{\sum_{t=h}^{T}(1-s_t)}; \quad h = 0, 1, \ldots, H\,. \tag{40}$$

In fact, the previous expression can be estimated as a simple local projection: $y_{t+h} = \mu_h + \beta_h s_t + u_{t+h}$ where $\hat{\mathcal{R}}_{sy}(h) = \hat{\beta}_h$.

In practice, s_t is usually not randomly assigned, but rather determined endogenously. Naturally, the most direct approach is to include observable information, $\boldsymbol{x}_t$ as right-hand side variables in a typical LP, specifically,

$$y_{t+h} = \alpha_h + \beta_h\, s_t + \boldsymbol{\gamma}_h\, \boldsymbol{x}_t + v_{t+h}\,; \quad h = 0, 1, \ldots, H\,.$$

As an example, note that in the context of a VAR DGP, the traditional Cholesky decomposition of the reduced-form error covariance based on a Wold causal ordering of the variables in the VAR has a direct correspondence to how such an assumption is implemented in an LP: one simply has to add as additional controls the appropriate contemporaneous values of system variables into $\boldsymbol{x}_t$. Specifically, in addition to lagged values of all the variables in the system, one should include the contemporaneous values of the variables ordered first in the Cholesky causal chain. Asymptotically these are equivalent: in large samples, both approaches (the Cholesky VAR and the analogous LP) will recover the same impulse responses (Plagborg-Møller and Wolf, 2021).

8.2. Inverse propensity scores: LP-IPW and LP-IPWRA

However, the covariates $\boldsymbol{x}_t$ may affect s_t non-linearly and this would in principle complicate matters considerably—the specific type of nonlinearity is usually unknown. The applied micro literature has solved this issue by reweighting the sample averages in Equation 40 using *inverse propensity scores*. The use of the propensity score goes back to Horvitz and Thompson (1952) and Rosenbaum and Rubin (1983). In economics, early references include Hirano, Imbens, and Ridder (2003) with the first applications to local projections by Angrist, Jordà, and Kuersteiner (2016) and Jordà and Taylor (2016), denoted LP-IPW.

So what is a propensity score? In the setting where $s_t \in \{0,1\}$, it refers to $p_t = P(s_t = 1|\boldsymbol{x}_t)$, which in practice can be obtained from a logit or probit first stage estimation.[17] Reweighting

[17] In turn, such a two-step estimator will require adjusting the calculation of the standard errors in the

Equation 40 with the inverse of the propensity score leads to the following expression,

$$\hat{\mathcal{R}}_{sy}(h) = \frac{\sum_{t=h}^{T} y_{t+h} s_t}{\sum_{t=h}^{T} p_t} - \frac{\sum_{t=h}^{T} y_{t+h}(1-s_t)}{\sum_{t=h}^{T}(1-p_t)}; \quad h = 0, 1, \ldots, H. \tag{41}$$

When interventions are binary, as in our example, inverse propensity score weighting offers a flexible alternative to achieving identification based on conditioning on observable covariates. Moreover, one can build on this estimator by also including controls $\boldsymbol{x}_t$ linearly on the right-hand side of the LP (i.e., regression adjustment) and using weighted least squares based on the propensity score. We can call this LP-IPWRA and it is a *doubly-robust* estimator. The literature on doubly-robust estimators is quite extensive and we refer the reader to Jordà and Taylor (2016) for the appropriate references to get started and for an example of application.

8.3. Traditional VAR identification schemes for local projections

We should note that VARs have had a 25-year running start over local projections when it comes to the issue of identification. However, recent work by Plagborg-Møller and Wolf (2021) formally derives the equivalence of VARs and local projections under some of the more traditional identification approaches.

We have already commented on the implementation of Cholesky-type identification in LPs. The right method here is simply to include the contemporaneous variables in the Wold causal order as right-hand side variables in the LP. In this subsection we touch on two other popular approaches, starting with identification through long-run restrictions, introduced by Blanchard and Quah (1989).

Using the same set-up as Blanchard and Quah (1989), Plagborg-Møller and Wolf (2021) show that to implement the long-run identification in that paper one can follow a two step procedure. Consider a bivariate set-up with GDP and the unemployment rate. First, based on the assumption that movements of output in the long-run are only explained by supply shocks, one can estimate a local projection of the long difference (for a *large* value of H chosen by the experimenter) of the log of real gdp, say $y_{t+H} - y_{t-1}$ on Δy_t, and the unemployment rate, say U_t (and the lags of both as additional covariates). Call β_H the response coefficients associated with Δy_t and U_t in this LP. Then the supply shock is essentially the linear combination $s_t = \beta_H^y \Delta y_t + \beta_H^U U_t$. In the second step one simply estimates the local projection using s_t as the impulse.

Plagborg-Møller and Wolf (2021) discuss other identification alternatives, such as identification with sign restrictions. However, in general traditional methods suffer from the inability to test the validity of the identification assumptions, and in the case of sign restrictions, inference can be quite complicated as one usually only achieves set identification, not point identification.

second stage.

8.4. Instrumental variables: LP-IV

Last but not least, identification of LPs via the use of instrumental variables is perhaps the most intuitive approach. This is now an established method, referred to as LP-IV (since its first appearance in Jordà, Schularick, and Taylor, 2015) and has found many uses in applied macroeconomics that are too numerous to mention.

As is usually the case with instrumental variables, conditions must be satisfied. One will need a *relevance* assumption (that is, the instrument is correlated with the endogenous variable, in this case the intervention) and an *exogeneity* assumption (the instrument is uncorrelated with the residuals). However, the latter is slightly different than in static settings, as Stock and Watson (2018) and Plagborg-Møller and Wolf (2021) show and as we have already discussed in Section 7.1. Formally, the assumptions one needs, as stated in Stock and Watson (2018), are those already stated in Assumption 1 above, or for more general nonparametric settings, those as discussed in Rambachan and Shephard (2019b).

Instrumental variable estimation of LPs offers another important advantage over VARs as highlighted by Plagborg-Møller and Wolf (2022), which is that identification is achieved even in *non-invertible* settings. Non-invertibility, loosely speaking, is a situation where the variables in a system are determined by an even larger number of shocks. An example would be news shocks about future technology, but there are many others in macroeconomics. In such a setting, structural shocks cannot be recovered solely as a function of current and lagged values of the variables in the system. As a result, VAR identification methods based on the covariance matrix of reduced-form residuals will not work directly, which poses a challenge. These issues have been highlighted in, for example Fernández-Villaverde, Rubio-Ramírez, Sargent, and Watson (2007), and more specifically by Stock and Watson (2018) and Plagborg-Møller and Wolf (2022) when discussing invertibility in the context of VARs and LPs. Recent developments to achieve identification in non-invertible VARs have been proposed in, e.g., Chahrour and Jurado (2022).

9. Indirect inference: Impulse response matching estimators

In many settings, the structural parameters $\boldsymbol{\theta}$ of an economic model can be expressed as functions of some auxiliary parameters $\boldsymbol{\pi}$ that can be estimated more easily with an auxiliary model. An example of such an approach are the impulse response matching estimators used in Rotemberg and Woodford (1997); Christiano, Eichenbaum, and Evans (2005) and Iacoviello (2005).

Specifically, suppose that we can express $\boldsymbol{\theta} = g(\boldsymbol{\pi})$ where, in particular, at the true value $\boldsymbol{\theta}_0 = g(\boldsymbol{\pi}_0)$. If in addition, $g(\boldsymbol{\pi})$ is locally identified and differentiable, and $\sqrt{T}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \to N(0, \Omega_\pi)$, to state the basic assumptions, then, the classical minimum distance problem

$$\min_{\theta} (\hat{\boldsymbol{\theta}} - g(\boldsymbol{\theta}))' W_T (\hat{\boldsymbol{\theta}} - g(\boldsymbol{\theta})) \; ; \quad \dim(\boldsymbol{\theta}) = q < r = \dim(\boldsymbol{\theta}) \, ,$$

with W_T a weighting function, can be shown to deliver

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N(0, \Omega_\theta) \; ; \quad \Omega_\theta = (G' \Omega_\pi^{-1} G)^{-1} \, ,$$

when setting $W_T = \Omega_\pi^{-1}$, the optimal weighting matrix, and where G refers to the Jacobian of g with respect to $\boldsymbol{\pi}$ (see, e.g., Newey and McFadden, 1986, for a careful statement of the assumptions). This is a well known result with a long history in statistics and with many generalizations, including empirical likelihood estimation, for example (see, e.g., Owen, 1988).

In this section we review several settings in which this principle can be put to work to estimate traditional time series models and rational expectations or DSGE models models more generally; and its relation to system projection IV methods (Lewis and Mertens, 2022) and to evaluate deviations from optimal policy paths (Barnichon and Mesters, 2023).

9.1. Projection Minimum Distance

Jordà and Kozicki (2011) propose a simple approach via the method of Projection Minimum Distance (PMD) to estimate time series models whose likelihoods would usually require maximization with numerical optimization routines. The same principles can also be applied to estimate a wide range of rational expectations models or even DSGE models (Castellanos and Cooper, 2023).

The following simple example illustrates the point. Suppose that our interest is in estimating the reduced-form ARMA(1,1) model

$$y_t = \rho y_{t-1} + \epsilon_t + \theta \epsilon_{t-1} \, , \tag{42}$$

where the constant is omitted for simplicity. Further, suppose that $\mathcal{R}_h$ is the impulse response coefficient for $h = 0$ to H from a local projection of y onto itself (hence for clarity we omit the subscripts in $\mathcal{R}_h$). Let $\hat{\mathcal{R}} = (\hat{\mathcal{R}}_0, \ldots, \hat{\mathcal{R}}_h, \ldots, \hat{\mathcal{R}}_H)'$ and let the corresponding estimate of the covariance matrix be $\hat{\Omega}_{\mathcal{R}}$. Then ρ and θ can be estimated using the following two-step process:

Projection Minimum Distance method

- Step 1: Obtain $\hat{\mathcal{R}}$ and $\hat{\Omega}_{\mathcal{R}}$ as usual using, for example, GMM as shown previously.
- Step 2: Estimate the OLS pseudo-regression (to implement the minimum distance step):

$$\underbrace{\begin{pmatrix} \hat{\mathcal{R}}_1 \\ \hat{\mathcal{R}}_2 \\ \vdots \\ \hat{\mathcal{R}}_H \end{pmatrix}}_{\hat{\mathcal{R}}_y} = \underbrace{\begin{pmatrix} 1 & \hat{\mathcal{R}}_0 \\ 0 & \hat{\mathcal{R}}_1 \\ \vdots & \\ 0 & \hat{\mathcal{R}}_{H-1} \end{pmatrix}}_{\hat{\mathcal{R}}_x} \underbrace{\begin{pmatrix} \rho \\ \theta \end{pmatrix}}_{\delta} \quad \to \quad \hat{\delta} = (\hat{\mathcal{R}}_x' \hat{\mathcal{R}}_x)^{-1} (\hat{\mathcal{R}}_x' \hat{\mathcal{R}}_y) \, .$$

  where the variance-covariance matrix of the parameter estimates $V(\hat{\delta})$ can be obtained using classical minimum distance results based on $\hat{\mathcal{R}}_y$, $\hat{\mathcal{R}}_x$ and $\hat{\Omega}_{\mathcal{R}}$.

To see this method in practice, consider estimation of the following standard, generic rational expectations expression (a good example of such an expression is the Phillips curve),

$$y_t = E_t \boldsymbol{w}_{t+1} \boldsymbol{\theta}_e + \boldsymbol{w}_t \boldsymbol{\theta}_c + u_t \,, \tag{43}$$

where $\boldsymbol{w}_t$ is a vector of forcing variables. Note that extending this specification with more lags or when a vector of left-hand side variables is considered, would be reasonably straightforward.

Shifting time and taking expectations on both sides of Equation 43, it is easy to see that

$$E[y_{t+h}|s_t; \boldsymbol{x}_t] = E[\boldsymbol{w}_{t+h+1}|s_t; \boldsymbol{x}_t]\boldsymbol{\theta}_e + E[\boldsymbol{w}_{t+h}|s_t; \boldsymbol{x}_t]\boldsymbol{\theta}_c \,; \quad h = 0, 1, \ldots, H \,. \tag{44}$$

Taking the difference in expectations when $s_t = s_0 + 1$ versus when $s_t = s_0$, we have

$$\mathcal{R}_{sy}(h) = \mathcal{R}_{\boldsymbol{s}w}(h+1)\boldsymbol{\theta}_e + \mathcal{R}_{\boldsymbol{s}w}(h)\boldsymbol{\theta}_c \,; \quad h = 0, 1, \ldots, H \,. \tag{45}$$

If s_t is not identified, then as Barnichon and Mesters (2020) and more recently Lewis and Mertens (2022) propose, one can use instrumental variables.[18]

More generally, for settings linear in the parameters such as Equation 43, and without detailing all the usual assumptions for brevity, we can state the problem as follows,

$$\underset{H\times 1}{\mathcal{R}} = \underset{H\times k}{\mathcal{G}} \; \underset{k\times 1}{\boldsymbol{\theta}} \,. \tag{46}$$

The corresponding minimum distance problem is therefore

$$\min_{\theta} (\hat{\mathcal{R}} - \hat{\mathcal{G}}\boldsymbol{\theta})' W (\hat{\mathcal{R}} - \hat{\mathcal{G}}\boldsymbol{\theta}) \,, \tag{47}$$

where we may assume that $\sqrt{T}(\hat{\mathcal{R}} - \mathcal{R}) \to N(0, \Omega_{\mathcal{R}})$, which will be the case in most standard applications. The first order conditions are:

$$-\hat{\mathcal{G}}' W (\hat{\mathcal{R}} - \hat{\mathcal{G}}\boldsymbol{\theta}) = 0 \,. \tag{48}$$

Using a mean value expansion around the first order conditions, we have that

$$\sqrt{T}(\hat{\mathcal{R}} - \hat{\mathcal{G}}\boldsymbol{\theta}) = \sqrt{T}(\hat{\mathcal{R}} - \mathcal{R}_0) - \overline{\mathcal{G}}' W \sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \,; \qquad \overline{\mathcal{G}}, \hat{\mathcal{G}} \to \mathcal{G}_0 \,. \tag{49}$$

Thus, plugging this mean value expansion back into the first order conditions in Equation 48, one can easily show that

$$\Omega_\theta = (\mathcal{G}' W \mathcal{G})^{-1} (\mathcal{G}' W \Omega_{\mathcal{R}} W \mathcal{G}) (\mathcal{G}' W \mathcal{G})^{-1} \,, \tag{50}$$

[18] Lewis and Mertens (2022) also provide methods for inference with weak instruments.

which simplifies to $\Omega_\theta = (\mathcal{G}'\Omega_{\mathcal{R}}^{-1}\mathcal{G})^{-1}$ when choosing the optimal weighting matrix $W = \Omega_{\mathcal{R}}^{-1}$. In finite samples one would approximate $\mathcal{G}$ with $\hat{\mathcal{G}}$ and $\Omega_{\mathcal{R}}$ with $\hat{\Omega}_{\mathcal{R}}$.

This approach, of course, does not require the user to use LPs to estimate the impulse responses and their covariance matrix. Guerrón-Quintana, Inoue, and Kilian (2017) formally derive the asymptotic properties of matching estimators based on VARs. Relatedly, Hall, Inoue, Nason, and Rossi (2012) propose an information criterion to determine the optimal number of horizons of the impulse response that balances fit with the increased uncertainty of responses estimated at far horizons. The formula for their criterion is rather simple, given by

$$\hat{H} \;=\; \text{argmin}_{h\in\{h_{min},\ldots,h_{max}\}} \;\ln(|\hat{\Omega}_\theta|) + h\frac{\ln(\sqrt{T/k})}{(\sqrt{T/k})}\,, \tag{51}$$

where k is the truncation lag in the LP specification and Ω_θ refers to the covariance matrix of the structural parameters.

As an illustration of the practical application of the PMD method, we present an example based on the estimation of the Phillips curve for the U.K. Here, Equation 45 takes the form

$$\mathcal{R}_{s\pi}(h) = \mathcal{R}_{s\pi}(h+1)\theta_\pi + \mathcal{R}_{sx}(h)\theta_u\,,$$

where π_t is 12-month CPI inflation in log form and $x_t = u_t - u_t^*$ is the unemployment gap relative to the NAIRU, where the latter is extracted from a very low frequency bandpass filter. The identified monetary policy innovation s_t is from Cloyne and Hürtgen (2016). Using PMD we find $\hat{\theta}_\pi = 0.838(0.593)$ and $\hat{\theta}_x = -1.990(0.180)$. Figure 7 shows the two impulse responses used to estimate these parameters, with panel (a) displaying the response of inflation to a monetary shock, $\mathcal{R}_{s\pi}(h)$, and panel (b) showing the response of the unemployment gap instead, $\mathcal{R}_{sx}(h)$. Panels (c) and (d) show the partial scatters which correspond to how the parameters $\hat{\theta}_\pi$ and $\hat{\theta}_x$ are calculated.

9.2. Optimal Policy Evaluation

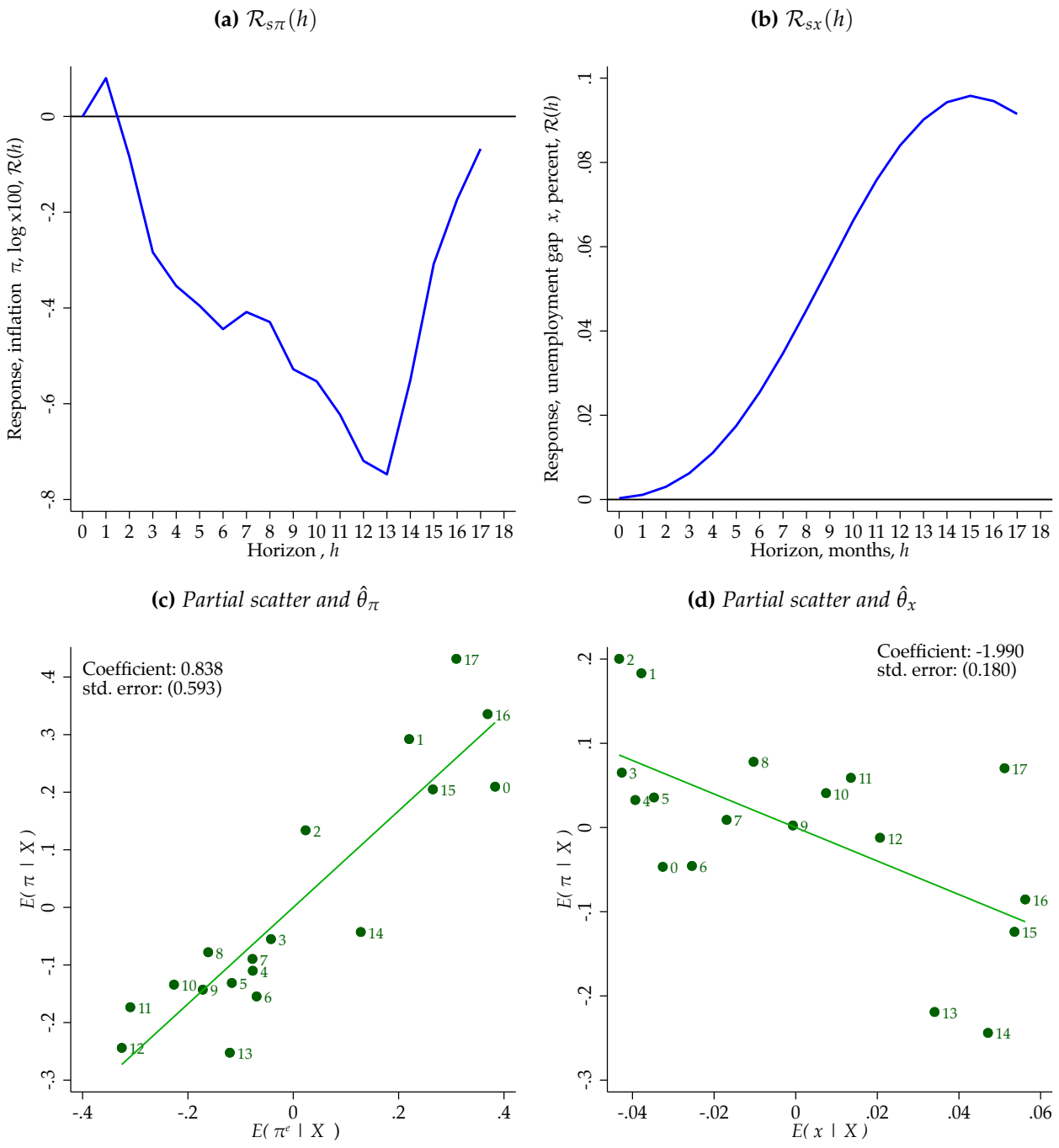
Barnichon and Mesters (2023) offer a clever approach to evaluating policy around the optimal path when policy is obtained as a linear rule from minimizing quadratic loss. Specifically, suppose the policymaker is interested in minimizing deviations of inflation from target as well as deviations of the unemployment rate from the natural rate, over a given horizon. Hence we can define

$$\underset{1\times H}{\eta_{\pi,t}^{H\,\prime}} = [(E_t\pi_{t+1} - \pi^*)\ldots(E_t\pi_{t+H} - \pi^*)]'\,, \tag{52}$$

$$\underset{1\times H}{\eta_{u,t}^{H\,\prime}} = [(E_t u_{t+1} - u^*)\ldots(E_t u_{t+H} - u^*)]'\,, \tag{53}$$

where π_t and π^* refer to inflation and its target; and u_t and u^* refer to the unemployment rate and its natural rate, for example. Assume the policymaker's goal is to minimize the quadratic loss given

Figure 7: *Using Projection Minimum Distance to estimate the Phillips curve for the U.K.*



Notes: Sample: 1975m1–2007m12. X denotes controls. See text.

by

$$\min_{s} \mathcal{L} = \frac{1}{2} \left(\eta_{\pi,t}^{H\,\prime} \;\; \eta_{u,t}^{H\,\prime} \right) W \begin{pmatrix} \eta_{\pi,t}^{H} \\ \eta_{u,t}^{H} \end{pmatrix} , \tag{54}$$

where W is a weighting matrix that may reflect how the policymaker weighs deviations in one period relative to others, and minimization is based on choosing the policy variable s.

Notice that linearity allows us to write

$$\frac{\partial \eta_{\pi,t}^{H}}{\partial s} = \mathcal{R}_{s\pi}^{H}; \quad \frac{\partial \eta_{u,t}^{H}}{\partial s} = \mathcal{R}_{su}^{H} , \tag{55}$$

where $\mathcal{R}_{s\pi}^{H}$ and $\mathcal{R}_{su}^{H}$ are the responses of inflation and the unemployment rate to a policy shock. Hence, the first order conditions, $\nabla \hat{\mathcal{L}}(\hat{s}) = 0$ are

$$\nabla \hat{\mathcal{L}}(\hat{s}) \;=\; \underbrace{(\mathcal{R}_{s\pi}^{H\,\prime} \;\; \mathcal{R}_{su}^{H\,\prime})}_{\mathcal{R}'} W \underbrace{\begin{pmatrix} \eta_{\pi,t}^{H} \\ \eta_{u,t}^{H} \end{pmatrix}}_{\boldsymbol{\eta}} \;=\; \mathcal{R}' W \boldsymbol{\eta} = 0 . \tag{56}$$

Now consider a mean value expansion around the optimal $\hat{s}$ calculated in a finite sample given by

$$\underbrace{\nabla \hat{\mathcal{L}}(\hat{s})}_{=0 \text{ by F.O.C.}} \;=\; \underbrace{\nabla \hat{\mathcal{L}}(s_0)}_{=\mathcal{R}_0 W \boldsymbol{\eta}_0} + \underbrace{\nabla^2 \hat{\mathcal{L}}(\bar{s})}_{=\overline{\mathcal{R}}' W \overline{\mathcal{R}}} (\hat{s} - s_0) \; ; \quad \bar{s} \in [\hat{s}, s_0] , \quad \hat{s} \to s_0 , \tag{57}$$

where s_0 is the population optimal value of policy and hence we may write $\hat{\delta} = \hat{s} - s_0$ as the change in current policy that would get us closer to the true optimal policy. In the linear-quadratic setting that we have entertained so far, note that $\overline{\mathcal{R}}' W \overline{\mathcal{R}} \to \mathcal{R}_0' W \mathcal{R}_0$ and, hence,

$$\hat{\delta} = -(\mathcal{R}_0' W \mathcal{R}_0)^{-1} (\mathcal{R}_0' W \boldsymbol{\eta}_0) . \tag{58}$$

Using typical minimum distance results, and denoting the covariance matrix of $\hat{R}$ as $\Sigma_{\mathcal{R}}$, the variance of $\hat{\delta}$ is

$$V(\hat{\delta}) = (\mathcal{R}_0' W \mathcal{R}_0)^{-1} (\mathcal{R}_0' W \Sigma_{\mathcal{R}} W \mathcal{R}_0) (\mathcal{R}_0' W \mathcal{R}_0)^{-1} , \tag{59}$$

which simplifies to $V(\hat{\delta}) = (\mathcal{R}_0' \Sigma_{\mathcal{R}}^{-1} \mathcal{R}_0)^{-1}$ in the special case where the policymaker chooses $W = \Sigma_{\mathcal{R}}^{-1}$. In practice, of course, all the population items can be substituted with their finite sample estimates. Thus, $\hat{\mathcal{R}}$ and $\hat{\boldsymbol{\eta}}$ can be obtained by local projections, for example, or from a VAR.

Equation 58 and Equation 59 thus allow one to test, for example, the null hypothesis that policy is approximately at its optimal level, i.e., $H_0 : \delta = 0$. When this hypothesis is rejected, $\hat{\delta}$ provides the policymaker the direction in which to modify policy toward the optimal value. Naturally, the responses embedded in $\hat{\mathcal{R}}$ need to be estimated causally, say, using an LP-IV or other identification

approach, and hence a natural estimate of $\Sigma_{\mathcal{R}}$ can be easily obtained with GMM as shown in Section 7.1. Importantly, note that nowhere in the discussion did we have to explicitly write down the policy rule.

10. Counterfactual paths

As we discussed in the introduction, an impulse response is a counterfactual comparison of means based on switching on and off a policy variable in the initial period. For example, the response of the unemployment rate presented in panel (b) of Figure 6 shows what such a response looks like based on a Romer and Romer (2004) shock. The response reflects the effect of the initial shock on the unemployment rate, as well as the effect of the shock on the monetary policy variable (say the federal funds rate) over time, and how it feeds back into the unemployment response. Thus we may ask, what would happen to the unemployment rate response if the policy path itself were to deviate from its usual pattern? This is the question that we try to answer in this section.

The Lucas critique (Lucas, 1976) suggests that such an experiment would be fraught. A deviation from the usual policy path would result in agents modifying their behavior accordingly, which would shift the latent economic environment and hence invalidate the analysis. In order to try to avoid this well-known problem, the approach that we follow in this section is similar to the *modest policy interventions* studied by Leeper and Zha (2003). That paper entertains modifications of the policy path that are sufficiently *modest* that they are unlikely to trigger a substantial revision in the agents' expectations and thus a transformation of the economic environment. We will follow the same intuition and therefore, the response modified by these modest interventions is probably best interpreted as a *derivative*.

Under relatively general conditions, Section 7.1 above showed that impulse responses estimated by local projections are asymptotically normal with a given variance-covariance matrix. Denote with $\boldsymbol{\beta}_u$ and $\boldsymbol{\beta}_r$ the $H \times 1$ response of the unemployment rate and the funds rate respectively to a monetary shock.

Hence, we may write

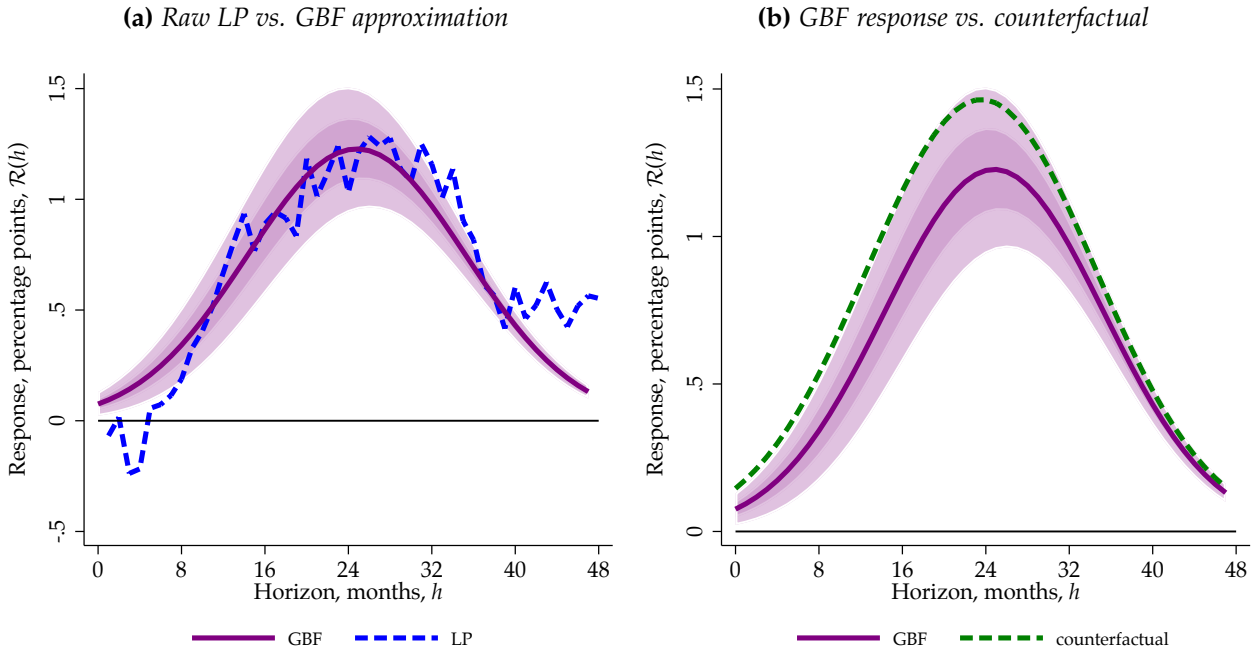
$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_u \\ \hat{\boldsymbol{\beta}}_r \end{pmatrix} \rightarrow \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\beta}_u \\ \boldsymbol{\beta}_r \end{pmatrix}; \begin{pmatrix} \Omega_{uu} & \Omega_{ur} \\ \Omega_{ru} & \Omega_{rr} \end{pmatrix} \right). \tag{60}$$

Denote by $\boldsymbol{\beta}_r^c$ a counterfactual response of the funds rate. Based on the rules of the multivariate normal distribution, we can then calculate the unemployment rate response conditional on $\boldsymbol{\beta}_r^c$ as follows,

$$\boldsymbol{\beta}_u^c = \hat{\boldsymbol{\beta}}_u + \Omega_{ur}\Omega_{rr}^{-1}(\boldsymbol{\beta}_r^c - \hat{\boldsymbol{\beta}}_r), \tag{61}$$

$$\Omega_{uu}^c = \Omega_{uu} - \Omega_{ur}\Omega_{rr}^{-1}\Omega_{ru}. \tag{62}$$

Figure 8: *The response of the unemployment rate to a counterfactual funds rate response*



Notes: Sample: 1985m1–2000m12. Response of the unemployment rate and the federal funds rate to a shock in the latter, instrumented with a Romer and Romer (2004) monetary shock. Both responses estimated using a Gaussian basis function using GMM as shown in Section 5. See text.

Assessing whether β_r^c represents a *modest* enough deviation from $\hat{\beta}_r$ can be accomplished using the Mahalanobis distance, which given the assumptions maintained will have an approximate χ^2 distribution. That is,

$$\mathcal{M} = (\beta_r^c - \hat{\beta}_r)' \Omega_{rr}^{-1} (\beta_r^c - \hat{\beta}_r) \to \chi_H^2 \,. \tag{63}$$

Thus, when β_r^c is relatively close to $\hat{\beta}_r$, the statistic $\mathcal{M}$ will be small and the *null* that the proposed counterfactual path β_r^c is indistinguishable from β_r, will not be rejected based on a χ_H^2 metric. We will interpret failure to reject the null as evidence in favor of a modest couterfactual.

As an example, Figure 8 shows how this approach can be used in practice. Panel (a) of the figure replicates panel (b) of Figure 6. It shows the response of the unemployment rate estimated using LP-IV and using a Gaussian basis function (GBF) approximation.

We may not have strong views on the value of the individual coefficients of an impulse response. However, the coefficients of the GBF have a nice interpretation that we can exploit for our purposes. These coefficients are also normally distributed so that we can apply the same calculations as in Equation 61. Specifically, let $\beta_u = \phi(a_u, b_u, c_u)$ where ϕ denotes the GBF and a_u, b_u, c_u are its corresponding coefficients for the smoothed unemployment rate response. Similarly we write $\beta_r = \phi(a_r, b_r, c_r)$. Now we can use Equation 61 to determine $\beta_u^c = \phi(a_u^c, b_u^c, c_u^c)$ based on some counterfactual assumption on the path for β_r.

As an example, in panel (b) of Figure 8 we consider a counterfactual in which the response of the federal funds rate is approximated with a Gaussian basis function with parameters $a_r = 2.17\,(0.04); b_r = 4.66\,(0.12)$; and $c_r = 6.10\,(0.14)$, where the numbers in parenthesis are standard errors. To keep things simple, we then experiment with a counterfactual path for the funds rate where we reduce the parameter b_r by one standard deviation. Recall that this is the parameter associated with the timing of the peak response. Thus, by reducing b_r, we move the peak funds response 1 standard deviation earlier. This policy experiment results in the counterfactual path for the unemployment rate response displayed as a dashed green line in panel (b) of Figure 8.

The figure repeats the original GBF response estimate of the unemployment rate (as a solid line) along with its counterfactual response (as a dashed line). As expected, the counterfactual experiment results in the unemployment rate being higher earlier on, and peaking slightly sooner, before returning back to 0. Here, the Mahalanobis distance statistic $\mathcal{M}$ has a p-value of 0.08, indicating that this is a borderline *modest* intervention and thus the numerical results should be interpreted with some caution.

11. State-dependent responses: A decomposition

LPs are well suited to the analysis of state-dependent impulse responses, that is where the impulse response may be allowed to vary across regimes determined by one or more state variables.

Stratification For example, many studies have examined whether the impact of a monetary policy shock depends on the boom-bust or recession-expansion state of the economy (e.g., Tenreyro and Thwaites, 2016; Angrist, Jordà, and Kuersteiner, 2016; Jordà, Singh, and Taylor, 2024). Likewise, another literature focuses on whether the impact of a fiscal policy shock is also dependent on the state of the cycle (e.g., Auerbach and Gorodnichenko, 2012a; Jordà and Taylor, 2016; Ramey and Zubairy, 2018).

As an example of how one can implement stratification, let D_{t-r} be a binary indicator variable for some measure of the state of the economy at time $t-r$ for $r > 0$ prior to intervention. In principle, if the state is determined prior to intervention and the intervention itself is not influenced by the state or other factors (i.e., is as good as if randomly determined), then one can, for example, estimate two long-difference LPs,

$$y_{t+h} - y_{t-1} = \alpha_h^j + \beta_h^j \Delta s_t + \boldsymbol{\gamma}_h^j \Delta \boldsymbol{x}_t + v_{t+h}\,; \quad D_{t-r} = j \in \{0,1\}\,, \quad r > 0\,, \quad h = 0,1,\ldots,H\,, \tag{64}$$

where the controls $\Delta \boldsymbol{x}_t$ might include lag differences of the outcome and lags of the intervention. Here, β_h^j would capture the coefficients of the response in regime $j = 0, 1$.

Why is this approach needed? When impulse responses are state-dependent, estimating a traditional local projection by conditioning on past information without also conditioning on the state, will mix up the state-specific responses to yield only an overall average response. The correct

approach is to condition on the state as well, and to estimate a state-dependent LP. In general, this will require the full set of interactions of the state variable with all controls for past information.

How should one interpret a state-dependent impulse response? The answer depends on the method used. In a state-dependent VAR, if one derives the response as usual by using state-specific VAR parameters and deriving the response as usual, the implicit assumption is that the economy will remain in that particular state forever into the future. This is usually unrealistic. In practice, the economy may, and likely will, switch states in the future and may do so more than once. Thus, the correct impulse response given the state will usually require simulation methods to then average across all possible future trajectories that allow the state to shift as time goes on.

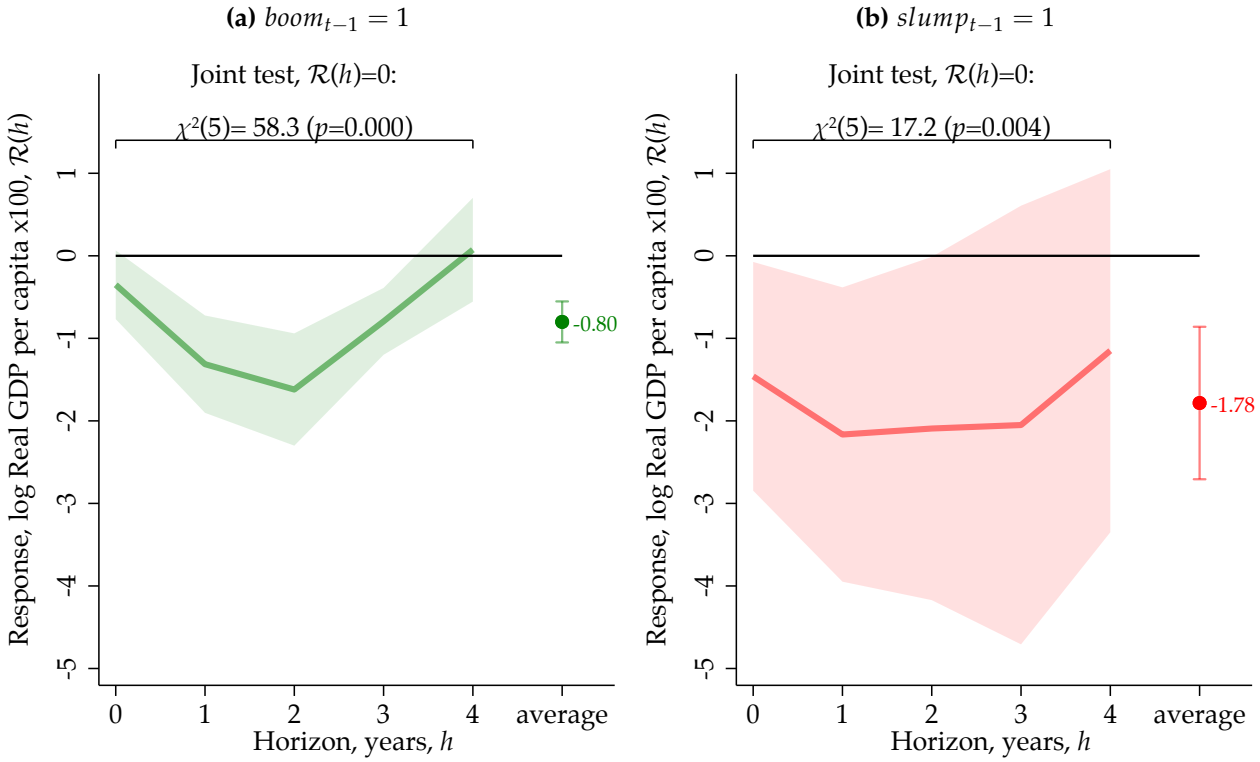
LPs do not require such simulations. By construction, conditional on today's state, LPs directly estimate the average response across all possible trajectories that the economy may follow in the future, including possible future shifts in the state, given today's state and conditional on controls.

As an illustration, following the earlier Figure 2 based on Jordà and Taylor (2016), in Figure 9 we present an example. Recall that the outcome is real GDP and the policy shock is a fiscal consolidation, for an OECD annual panel from 1978 to 2019 sample based on the data constructed by Guajardo, Leigh, and Pescatori (2014) and updated to 2019. The stratification variable D_t takes the value 1 in a boom (or 0 in a slump), defined, respectively, as periods when the HP-detrended cyclical component of output is positive (or negative). Critically, a key assumption is that consolidations are not influenced by whether the economy is in a boom or a slump.

The figure shows updated results comparable to the main findings in Jordà and Taylor (2016). Fiscal consolidations are contractionary over horizon years 0 to 4, in both split samples. Tests of both the average response and the joint test of non-zero response indicate that the differences are statistically significant. However, the output response is much larger when fiscal consolidations are implemented during slumps, as compared to booms. The estimated slump response is imprecisely estimated, but the result becomes clearer when the average response $\frac{1}{H}\sum_0^4 \beta_h$ is calculated, which amounts to -1.78 in slumps and -0.80 in booms. When the estimated multipliers are similarly calculated with stratification (not shown) they are also negative, and twice as large in slumps as compared to booms.

Lastly, note that difficulties arise if the state includes current information (unlike Equation 64). Then, policy interventions are influenced by the current state and vice versa. For example, interest rates are set low in slumps and high in booms so a naïve stratification could result in confounding, e.g., a finding of weak responses to monetary shocks. (This problem might be less severe for fiscal policy, which may react with a lag). Ideally, both the policy intervention and state would be determined exogenously in a quasi-experiment, where one does not influence the other. Thus we may require instruments for both the intervention and the state. A similar point has been raised by Gonçalves, Herrera, Kilian, and Pesavento (2024); they show that large interventions are unlikely to represent the true population response and that a conservative interpretation is to view the estimated responses as *derivatives*, i.e., what would happen with an infinitesimally small intervention.

Figure 9: *State-dependent cumulative fiscal impulse response* $\mathcal{R}_{yf}(h)$



Notes: Outcome y_{it} is log real GDP per capita from Jordà and Taylor (2016), and f denotes a fiscal shock, a treatment Δs_{it} is dCAPB from Guajardo, Leigh, and Pescatori (2014), updated to 2019, instrument z_{it} is GLP2 size of fiscal consolidation Guajardo, Leigh, and Pescatori (2014), updated to 2019. OECD sample, 1978–2019. Control variables are two lags of treatment, two lags of outcome, lag change in the public debt to GDP ratio, and lag of HP-filtered cyclical component of log real GDP per capita. 95% confidence bands are shown and the joint test.

Heterogeneity Linearity, whether in the context of VARs or LPs, makes complex models tractable, but the restrictions that it imposes are sometimes forbidding. We saw earlier that linearity means that the sign of the intervention is irrelevant as it simply flips the sign of the impulse response. It also means that the size (or dose) of the intervention simply scales the response proportionately, but does not change its shape—a 50 basis points (bps) change in interest rates would be expected to have twice the impact of a 25 bps change. And the state of the world when the intervention takes place has no effect on the response—an interest rate hike during a boom would be expected to cool the economy as during a bust. These and other features of linearly estimated responses seem too restrictive and our fiscal policy example seems to bear this out.

Abandoning linearity usually comes at a steep cost in complexity, at least when working with VARs. However, since LPs are a single equation method, these costs tend to be lower. In fact, a great degree of heterogeneity can be achieved with specifications that remain linear in parameters and hence easy to estimate with standard methods. In this section we rely on recent work by Cloyne, Jordà, and Taylor (2023) to explain some of these extensions and their interpretation. We

refer the reader to that paper for the in-depth exposition of what follows. More complex forms of nonlinearities, of course, will require nonlinear estimation methods and appropriate care in computing the impulse response as discussed later in Section 12.

Consider a departure from the binary example discussed earlier, where the economy can be either in a boom or a slump. Instead, think of the economy as being in a continuum of states. There are a number of ways of approaching this problem, perhaps the simplest one is where the state of the economy is determined by the vector $\boldsymbol{x}_t$ of controls. The previous boom/slump example would be a special case where, say, an indicator variable $x_t \in \{0, 1\}$ determines the state. Yet another approach would be to use a factor variable that summarizes the state of the economy as a function of a vector of variables.

We may therefore be interested in comparing the responses resulting from moving from, say $\boldsymbol{x}_t = \boldsymbol{x}_0$ (such as, for example, $\boldsymbol{x}_0 = \overline{\boldsymbol{x}}$), where as before, $\boldsymbol{x}_t$ denotes lags of the outcome, the intervention, and other exogenous and pre-determined variables. The state of interest is some deviation $\boldsymbol{\delta}_x$ from this equilibrium state. It may be easier to think of a setting where all entries in $\boldsymbol{\delta}_x$ are zero, except for one variable of interest characterizing the state though this is, of course, not necessary. As before, let s_t denote the policy variable that will be shifted from s_0 to $s_0 + \delta_s$ (in non-linear models, the effect depends on where it is evaluated). This is a scenario similar to that in Auerbach and Gorodnichenko (2012b,a) and Tenreyro and Thwaites (2016), for example.

The researcher is thus usually interested in evaluating the effectiveness of an intervention in a given state, via

$$\mathcal{R}_{sy|x}(h) = E[y_{t+h}|s_t = s_0 + \delta_s; \boldsymbol{x}_t = \boldsymbol{x}_0 + \boldsymbol{\delta}_x] - E[y_{t+h}|s_t = s_0; \boldsymbol{x}_t = \boldsymbol{x}_0 + \boldsymbol{\delta}_x], \tag{65}$$

where δ_s is the only difference between these two expectations. This response can be further decomposed by adding and subtracting $E[y_{t+h}|s_t = s_0 + \delta_s; \boldsymbol{x}_t = \boldsymbol{x}_0]$ and $E[y_{t+h}|s_t = s_0; \boldsymbol{x}_t = \boldsymbol{x}_0]$. Simple manipulations allow us to decompose Equation 65 into

$$\begin{aligned}
\mathcal{R}_{sy|x}(h) = & \underbrace{E[y_{t+h}|s_t = s_0 + \delta_s; \boldsymbol{x}_t = \boldsymbol{x}_0 + \boldsymbol{\delta}_x] - E[y_{t+h}|s_t = s_0 + \delta_s; \boldsymbol{x}_t = \boldsymbol{x}_0]}_{\mathcal{R}_{xy|s=s_0+\delta_s}(h)} \\
& - \underbrace{E[y_{t+h}|s_t = s_0; \boldsymbol{x}_t = \boldsymbol{x}_0 + \boldsymbol{\delta}_x] - E[y_{t+h}|s_t = s_0; \boldsymbol{x}_t = \boldsymbol{x}_0]}_{\mathcal{R}_{xy|s=s_0}(h)} \\
& + \underbrace{E[y_{t+h}|s_t = s_0 + \delta_s; \boldsymbol{x}_t = \boldsymbol{x}_0] - E[y_{t+h}|s_t = s_0; \boldsymbol{x}_t = \boldsymbol{x}_0]}_{\mathcal{R}_{sy|x=x_0}(h)}.
\end{aligned} \tag{66}$$

What do we learn from this decomposition? First, the effect of a policy intervention that happens when the state is at $\boldsymbol{x}_0 + \boldsymbol{\delta}_x$ reflects components that seemingly have nothing to do with the intervention, as is captured by $\mathcal{R}_{xy|s=s_0+\delta_s}(h)$ and $\mathcal{R}_{xy|s=s_0}(h)$. We say seemingly because, although the only element in the conditioning information set that is shifting is $\boldsymbol{x}_t$, the state is related to the policy variable s_t in general. For example, lower interest rates are generally an endogenous

response to a weak economy. Thus, the decomposition highlights that identification requires not only exogenous variation in s_t but also in $\boldsymbol{x}_t$ (or at least, in the subset of variables in $\boldsymbol{x}_t$ implicated in determining the state).

Based on these simple derivations Cloyne, Jordà, and Taylor (2023) propose the following extension to the usual LP linear specification,

$$y_{t+h} = \underbrace{\mu_{0h} + \beta_h(s_t - s_0) + \boldsymbol{\gamma}_h(\boldsymbol{x}_t - \boldsymbol{x}_0)}_{\text{usual local projection}} + \underbrace{\boldsymbol{\theta}_h(s_t - s_0)(\boldsymbol{x}_t - \boldsymbol{x}_0)}_{\text{extension}} + v_{t+h}\,; \quad h = 0, 1, \ldots, H; \quad t = h, \ldots, T\,, \tag{67}$$

where note that a common choice would be to set $s_0 = \bar{s}$ and $\boldsymbol{x}_0 = \bar{\boldsymbol{x}}$ though this is done for convenience and clearly is not the only normalization one could arrange. Note that Equation 67 is still linear in parameters and therefore easy to estimate.

Going back to the decomposition of Equation 66, note the terms involving a shift in the state, $\mathcal{R}_{xy|s=s_0+\delta_s}(h) - \mathcal{R}_{xy|s=s_0}(h) = \boldsymbol{\theta}_h\delta_s\boldsymbol{\delta}_x$, whereas the term directly related to the policy intervention, $\mathcal{R}_{sy|x=x_0}(h) = \beta_h\delta_s$, which is the usual impulse response coefficient. The sum of the two is the state-dependent response where now clearly the term $\boldsymbol{\theta}_h\delta_s\boldsymbol{\delta}_x$ will attenuate/amplify the original response $\beta_h\delta_s$ depending on the sign of $\boldsymbol{\theta}_h$. Cloyne, Jordà, and Taylor (2023) call $\mathcal{R}_{sy|x=x_0}(h) = \beta_h\delta_s$ the *direct effect* of the intervention on the outcome and the term $\mathcal{R}_{xy|s=s_0+\delta_s}(h) - \mathcal{R}_{xy|s=s_0}(h) = \boldsymbol{\theta}_h\delta_s\boldsymbol{\delta}_x$, the *indirect effect*. This is because the latter captures how intervention shifts the way covariates affect the outcome.

This last term plays an important role. First, compared to the usual stratification of impulse responses based on a given state variable, Equation 67 suggests that such specifications may incur an omitted variable bias—stratification could also be required of other elements in $\boldsymbol{x}_t$. Second, as Fortin, Lemieux, and Firpo (2011) explain, the decomposition in Equation 67 for static regressions (also know as the Kitagawa-Oaxaca-Blinder[19] decomposition) is a partial equilibrium decomposition. In other words, the covariates themselves are correlated and hence not usually identified, an observation also made in Cloyne, Jordà, and Taylor (2023) and later by Gonçalves, Herrera, Kilian, and Pesavento (2024), as we previously discussed. Thus, the second lesson is that one requires identification not only for s_t, but also for the elements of $\boldsymbol{x}_t$ whose stratification one is interested in characterizing.

Time-varying responses However, there is another interesting feature of Equation 67. As long as $\boldsymbol{\theta}_h \neq \mathbf{0}$, then the impulse response will vary depending on the value that $\boldsymbol{x}_t$ takes in relation to $\boldsymbol{x}_0$. That is, the impulse response is time-varying. Previous papers have reported time-varying responses (e.g., Cogley and Sargent, 2005; Primiceri, 2005), however, these are usually based on low-dimensional time-varying VARs where the parameters of the model are allowed to follow a latent unit root process. Estimation is done using Bayesian methods. Importantly, time variation in the response is linked to the latent drift in the parameters though direct economic interpretation

[19] See Kitagawa (1955); Blinder (1973); Oaxaca (1973).

of what caused the drift is indirect, by looking at how the drift correlates with other economic aggregates. In contrast, Equation 67 ties the time-variation of the responses directly to the state of the economy characterized by the value of $\boldsymbol{x}_t$ at each point in the sample, which may be very useful.

In practice what this means is that one can answer the question: How effective is a policy intervention likely to be given the current state of the economy characterized by observable information? Moreover, this question can be answered without specifically giving a label to what that state is. This seems to be a question of first order importance for policymakers. We postpone discussion of how instruments can be used to achieve identification to Section 12. In that section, we discuss nonlinearities more broadly and that seems a better place for such a discussion.

For now, we provide a simple simulation exercise to illustrate the main features of the Cloyne, Jordà, and Taylor (2023) approach. Assume that there are two exogenous variables of interest, s_t will be the primary intervention of interest whereas x_t will be a secondary exogenous variable. You can think of it as a secondary intervention, such as when one examines fiscal policy given monetary policy. The DGP is as follows,

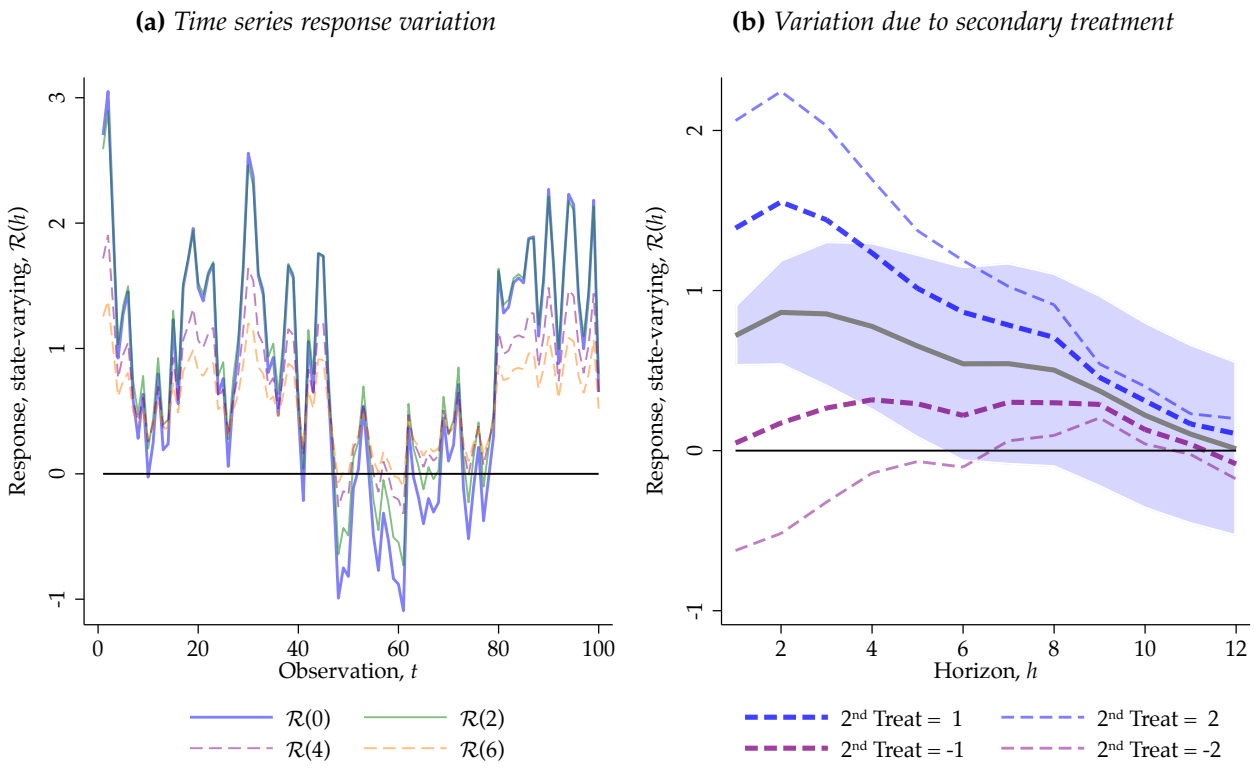
$$\begin{cases} s_t &= 0.75\, s_{t-1} + v_{s,t}\,, \\ x_t &= 0.75\, x_{t-1} + v_{x,t}\,, \\ y_t &= 0.75\, y_{t-1} + \gamma x_{t-1} + I(|s_t| > 1)\,(\beta s_t + \theta x_t s_t) + v_{y,t}\,; \end{cases} \qquad v_{i,t} \sim N(0,1) \text{ for } i = y, s, x\,. \quad (68)$$

Hence, s_t and x_t are exogenous by construction. We activate the primary treatment s_t only when $|s_t| > 1$ as indicated by the notation $I(|s_t| > 1)$. We assume that the internal propagation dynamics captured by y_{t-1} remain the same whether or not $|s_t| > 1$.

For the simulation, we set $\gamma = 0.75$ and $\beta = \theta = 0.5$ to keep the simulation simple. We initialize the data with 500 burn-in replications that we disregard and study instead the subsequent 500 observations. Then we estimate LPs as described in Equation 67. The results are displayed in Figure 10.

The figure is arranged in two panels. In panel (a) we show the response coefficients at horizons 0, 2, 4, 6 for the first 100 observations in the sample to highlight the time-variation generated by the interaction of the primary and secondary treatment variables. Even with this simple set-up, the effect on impact can fluctuate considerably: it is mostly positive for the first 50 observations, mostly negative for the next 25, before turning positive again. In panel (b) we show the average impulse response (which is the figure shown in most analyses) along with the attenuation (in purple)/amplification (in blue) generated by x_t for $x_t = -2, -1, 1, 2$. The response on average begins around 0.75 on impact and by period 12 it has died off to zero. When $x_t = 2$ the response on impact can be as large as 2 whereas when $x_t = -2$ the response on impact can be as low as about -0.75.

Figure 10: *Variation in the impulse response due to secondary treatment*



Notes: Data simulated from the model in Equation 68. Panel (a) shows the coefficients for the impulse response $\mathcal{R}(h)$ at horizons $h = 0, 2, 4, 6$ for the first 50 observations in the sample. Panel (b) shows the average response $\mathcal{R}(h)$ over the sample (along with two standard error bands) as well as the attenuation/amplification of the response when the secondary treatment takes on values -2, -1, 0, 1, 2. See text.

12. Nonlinearities

Nonlinearities are inherently difficult to handle as the range of possible specifications is vast. In practice, nonlinear specifications are usually motivated by specific objectives. Generally speaking, nonlinearities are difficult to implement in a VAR. The parametric load increases very rapidly, and nonlinear estimation methods quickly become cumbersome. LPs help alleviate this problem by virtue of being single equation methods. That said, nonlinearities also require the practitioner to form assumptions about the DGP to interpret the coefficients. The reader is directed to the work of Rambachan and Shephard (2019a,b) for more details.

In this section we review a general observation about non-linear LPs and highlight a few of the studies from the literature. Illustrating the main issues can be done with a simple motivating example. Hence consider the following nonlinear (separable) local projection

$$y_{t+h} = \mu_h(s_t; \boldsymbol{x}_t; \boldsymbol{\theta}) + v_{t+h}\,; \quad h = 0, 1, \ldots, H\,. \tag{69}$$

The corresponding impulse response will be

$$\mathcal{R}_{sy}(h, s_0, \delta;\, \boldsymbol{x}_t) = \mu_h(s_t = s_0 + \delta;\, \boldsymbol{x}_t;\, \boldsymbol{\theta}) - \mu_h(s_t = s_0;\, \boldsymbol{x}_t;\, \boldsymbol{\theta})\,. \tag{70}$$

Several features are worth remarking. First, note that the functional form $\mu_h(\,.\,)$ is allowed to vary for each horizon h. Second, note that the response function $\mathcal{R}$ not only depends on h. It now depends on the benchmark counterfactual, $s_t = s_0$, on the size and sign of the intervention, δ, and the value of the conditioning set $\boldsymbol{x}_t$. These are features we observed earlier when discussing stratification.

Moreover, care must be used when using instrumental variables to achieve identification, as we foreshadowed in the previous section. As is well known (see, e.g., Newey, 1990), even if we have an instrument for s_t, it is desirable to instrument any nonlinear transformation of s_t instead of using a first stage regression of s_t on z_t. Intuitively, the moment conditions that we want to exploit are

$$E\left[(y_{t+h} - \mu_h(s_t;\, \boldsymbol{x}_t;\, \theta))\boldsymbol{z}_t\right] = 0\,; \quad h = 0, 1, \ldots, H\,. \tag{71}$$

Simply put, Jensen's inequality would advise against running a first stage regression of s_t on $\boldsymbol{z}_t$ and $\boldsymbol{x}_t$ and then estimating

$$y_{t+h} = \mu_h(\hat{s}_t;\, \boldsymbol{x}_t;\, \boldsymbol{\theta}) + v_{t+h}\,; \quad h = 0, 1, , \ldots, H\,.$$

Now, for example, consider the following local projection

$$y_{t+h} = \beta_{1h} s_t + \beta_{2h} s_t^2 + \beta_{3h} s_t\, x_t + \gamma_h x_t + v_{t+h}\,; \quad h = 0, 1, \ldots, H\,. \tag{72}$$

The corresponding response is $\mathcal{R}_{sy}(h, s_0, \delta;\, x_t) = \beta_1 + \beta_2(\delta^2 + 2s_0\delta) + \beta_3\delta x_t$. This response is no longer symmetric (since δ^2 is always positive); it also varies with the size of the intervention, δ; it further depends on where the response is benchmarked, $s_t = s_0$; and lastly it will vary depending on the value of the control, x_t. However, it is worth noting that this particular specification is still linear in the parameters, which means that it can be estimated by simple least-squares methods. Recent applications of nonlinear local projections are numerous, indicating the usefulness of this technique: for example, estimation of quantile local projections (Linnemann and Winkler, 2016; Adrian, Grinberg, Liang, Malik, and Yu, 2022; Jordà, Kornejew, Schularick, and Taylor, 2022); and local projections when the outcome variable is binary (Drehmann, Patton, and Sorensen, 2007; Barattieri and Cacciatore, 2023), to cite a few.

Finally, going back to our discussion of IV estimation, suppose that an instrument z_t for s_t is available. Instead of using z_t in a first stage regression for s_t, the correct approach is to estimate Equation 72 by using as instruments z_t, z_t^2 and $z_t\, x_t$, perhaps in addition to other nonlinear transformations (see, e.g., Newey, 1990). These could then be use to construct the moment conditions in Equation 71.

13. Panel data

Increasingly, empirical economic analysis relies on longitudinal or panel data. Local projections are well-suited to handle this type of data. Estimating a single panel regression is far more convenient than estimating a system of panel regressions, as would be necessary with a vector autoregression. In addition to having potentially more observations with which to increase the precision of the response estimates, panel data will have implications for inference and open up additional methods of identification. Thus, a typical panel data local projection could be specified as

$$y_{it+h} = \mu_{ih} + \delta_{th} + \beta_h s_{it} + \boldsymbol{\gamma}_h \boldsymbol{x}_{it} + v_{it+h}\,, \tag{73}$$

where the main differences versus earlier specifications are the presence of individual and time-fixed effects, and a sample of $i = 1, \ldots, N$ individual units observed over $t = 1, \ldots, T$ time periods.

Specification, identification and analysis using local projections along the lines discussed in previous sections remain largely the same and many of the same methods are directly applicable to panel data. There is, however, two areas worth discussing in more detail: inference and difference-in-differences identification.

Inference In typical panel applications, inference depends in a fundamental manner on the dimensions N and T of the panel. In addition to the moving-average structure of the residuals in a local projection, it is natural to consider individual-level correlation across individual units. Recent developments in this area, specially regarding clustered standard errors are worth discussing.

At a basic level and taking a similar approach to that originally proposed in Jordà (2005), one could adjust for heteroscedasticity and autocorrelation using Driscoll-Kraay robust standard errors (Driscoll and Kraay, 1998), i.e., the direct analog of Newey-West standard errors for panels.[20] The asymptotic justification for this method relies on $T \to \infty$ with N fixed, or N growing at a slower rate than T.

A cluster-robust approach could be used in situations where $N \to \infty$ with T fixed to correct for autocorrelation. However, if T is relatively small, a recommended correction for heteroscedasticity is to use the wild cluster bootstrap (see Cameron, Gelbach, and Miller, 2008; Canay, Santos, and Shaikh, 2021; Roodman, Nielsen, MacKinnon, and Webb, 2019).[21] Importantly, note that the asymptotic distribution of the response coefficient in panels with large N relative to T will be dominated by the cross-sectional dimension, which will greatly remove concerns about distortions generated when there are roots near unity.

[20] In STATA this can be implemented with the command `xtscc`. The command allows one to select the maximum lag considered, the default option being set at $m(T) = \lfloor 4(T \div 100)^{2/9} \rfloor$.

[21] This type of bootstrap can be implemented in STATA with the user supplied command `boottest`. We thank Colin Cameron for useful comments on this section.

Difference-in-differences estimation A popular identification approach when selection into treatment is endogenously determined—but based on characteristics that are time-invariant—is difference-in-difference (DiD) estimation. In the simplest setting, with two periods and two groups (one of which is treated and the other one which is not) and under suitable conditions (e.g. *no anticipation* of treatment selection and *parallel trends*, i.e., treated and control units would have evolved along their pre-treatment trends absent treatment), the DiD estimator identifies the *average treatment effect on the treated*.

The literature has, however, evolved from this simple setting—usually based on the well-known two-way fixed effects (TWFE) estimator—to include more complex situations. This includes cases where more than one group of individuals receives treatment and this treatment is perhaps not administered at the same time (i.e., it is *staggered*). Moreover, treatment effects may vary across groups depending on when treatment is received (i.e., they are *heterogeneous*) and the effects may also change over time after treatment (i.e., they are *dynamic*). These extensions have generated an extensive new literature, well summarized in the surveys by Roth, Sant'Anna, Bilinski, and Poe (2023) and De Chaisemartin and d'Haultfoeuille (2023), for example.

What might be the connection between LPs and DiD? It is very deep. Dube, Girardi, Jordà, and Taylor (2023) show that most of the extensions to the basic two-period, two groups setting can be accommodated with a simple modification of an LP estimator under standard assumptions. In particular, using similar notation as before, let y_{it} denote the outcome variable, let the treatment indicator s_{it} equal 0 before treatment, 1 when treatment is administered and thereafter (i.e., treatment is an absorbing state), and let $\boldsymbol{x}_{it}$ denote a vector of covariates. That is, here, entering treatment is measured by Δs_{it}.

Hence the LP-DiD estimator of Dube, Girardi, Jordà, and Taylor (2023) can be expressed as

$$y_{i,t+h} - y_{i,t-1} = \delta_{th} + \beta_h \, \Delta s_{it} + \sum_{j=1}^{p} \rho_{jh} \, \Delta y_{i,t-j} + \gamma_h \, \boldsymbol{x}_{it} + v_{i,t+h} \,, \tag{74}$$

where δ_{th} are time fixed-effects (individual fixed effects are absorbed through the long differencing) and where—crucially—the estimation sample is restricted to observations that correspond to either $\Delta s_{it} = 1$ (newly treated units), or $s_{i,t+h} = 0$ (not yet treated units) to ensure *clean controls*, that is, to avoid comparisons between newly treated units and previously treated units. Dube, Girardi, Jordà, and Taylor (2023) show that many of the estimators proposed in the DiD literature and reviewed in the surveys by Roth, Sant'Anna, Bilinski, and Poe (2023) and De Chaisemartin and d'Haultfoeuille (2023) will fit into this convenient regression framework.

The key advantage of LP-DiD relative to panel distributed lag specifications rests on the clean control condition. When a unit enters treatment, it is no longer a valid control for subsequently treated units. In distributed lag specifications, one has to explore algorithmically all valid pairwise comparisons of treated and control units and recent papers in the DiD literature do just that. However, since LPs use forward looking variables, imposing the clean control condition is trivial, as

we have seen. Relative to the rest of the literature, one can then rely directly on regression methods to calculate the response coefficients of interest, which will be a variance-weighted average of the treatment effects for each group (though other user-chosen weights are trivial to implement).[22] In contrast, algorithmic methods based on distributed lag regression generally compute equally-weighted averages. Once again we are confronted with a bias-variance trade-off though in this case, the shoe is on the other foot.

14. Concluding remarks

In this review we tried to cover the most important topics in the rapidly evolving field of local projections. Inevitably, given space constraints, we have had to omit or skim over many new and ongoing areas of research likely to come to fruition in coming years. Our emphasis has been on the main ideas so that researchers can follow best practices. Just as important, we hope to have helped researchers understand how best to adapt the local projections method to their research needs.

An important takeaway from our review should be that local projections help bridge the divide between current best practices in applied microeconomics, and standard time series methods in macroeconomics. We hope to have highlighted the many points of commonality between the two traditions—in both univariate and panel data settings—and how each can benefit the other.

Counterfactual statements about the consequences of interventions are central to applied research and policymaking. Statistics related to such counterfactuals (such as differences in means, differences in quantiles, multipliers, and the like), can be constructed easily using local projections. Central to computing such statistics is identifying the causal channels at work. Though local projections per se do not solve the identification riddle, they incorporate instrumental variable estimation methods naturally, with extensions to nonlinear models, and they provide a decomposition of the indirect channels by which interventions affect outcomes.

Inference occupies a central role in any statistical analysis. Local projections require some degree of care when constructing inference, but once the main issues are understood, designing appropriate inferential procedures is straightforward. Our goal has been to show that local projections are a very flexible yet simple method to investigate dynamic causal properties of the data that have bearing on the problems economists want to investigate.

Local projections offer advantages and simplicity in many respects. But, as highlighted in the introduction and throughout the text, we hope to have provided guidance on how best to implement local projections, leaving the researcher to decide how best to approach individual scenarios given the context and the merits of the method.

[22] Code to implement LP-DiD in STATA can be found here: https://github.com/danielegirardi/lpdid.

References

Adrian, Tobias, Federico Grinberg, Nellie Liang, Sheheryar Malik, and Jie Yu. 2022. The Term Structure of Growth-at-Risk. *American Economic Journal: Macroeconomics* 14(3): 283–323.

Alessandri, Piergiorgio, Òscar Jordà, and Frabrizio Venditti. 2023. Decomposing the Monetary Policy Multiplier. Technical Report 2023-14, Federal Reserve Bank of San Francisco.

Alloza, Mario, Jesús Gonzalo, and Carlos Sanz. 2019. Dynamic effects of persistent shocks. Banco de España Working Paper 1944.

Angrist, Joshua D., Òscar Jordà, and Guido M. Kuersteiner. 2016. Semiparametric Estimates of Monetary Policy Effects: String Theory Revisited. *Journal of Business and Economic Statistics* 36(3): 371–387.

Angrist, Joshua D., and Guido M Kuersteiner. 2011. Causal effects of monetary shocks: Semiparametric conditional independence tests with a multinomial propensity score. *Review of Economics and Statistics* 93(3): 725–747.

Auerbach, Alan J., and Yuriy Gorodnichenko. 2012a. Fiscal multipliers in recession and expansion. In *Fiscal Policy after the Financial Crisis*, edited by Alberto Alesina and Francesco Giavazzi, 63–98. Chicago: University of Chicago Press.

Auerbach, Alan J., and Yuriy Gorodnichenko. 2012b. Measuring the output responses to fiscal policy. *American Economic Journal: Economic Policy* 4(2): 1–27.

Barattieri, Alessandro, and Matteo Cacciatore. 2023. Self-Harming Trade Policy? Protectionism and Production Networks. *American Economic Journal: Macroeconomics* 15(2): 97–128.

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

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

Barnichon, Regis, and Geert Mesters. 2020. Identifying modern macro equations with old shocks. *Quarterly Journal of Economics* 135(4): 2255–2298.

Barnichon, Régis, and Geert Mesters. 2023. A Sufficient Statistics Approach for Macro Policy. *American Economic Review* 113(11): 2809–45.

Blanchard, Olivier J., and Danny Quah. 1989. The Dynamic Effects of Aggregate Demand and Supply Disturbances. *American Economic Review* 79(4): 655–673.

Blinder, Alan S. 1973. Wage Discrimination: Reduced Form and Structural Estimates. *Journal of Human Resources* 8(4): 436–455.

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

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

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

Castellanos, Juan, and Russell Cooper. 2023. Indirect Inference: A Local Projection Approach. Unpublished. https://ssrn.com/abstract=4458439.

Chahrour, Ryan, and Kyle Jurado. 2022. Recoverability and expectations-driven fluctuations. *Review of Economic Studies* 89(1): 214–239.

Chang, Pao-Li, and Shinichi Sakata. 2007. Estimation of impulse response functions using long autoregression. *Econometrics Journal* 10(2): 453–469.

Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans. 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* 113(1): 1–45.

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

Cloyne, James, and Patrick Hürtgen. 2016. The Macroeconomic Effects of Monetary Policy: A New Measure for the United Kingdom. *American Economic Journal: Macroeconomics* 8(4): 75–102.

Cloyne, James, Òscar Jordà, and Alan M. Taylor. 2023. State-Dependent Local Projections: Understanding Impulse Response Heterogeneity. NBER Working Paper 30971.

Cogley, Timothy, and Thomas J. Sargent. 2005. Drifts and volatilities: monetary policies and outcomes in the post WWII US. *Review of Economic Dynamics* 8(2): 262–302.

De Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2023. Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *Econometrics Journal* 26(3): C1–C30.

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

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

Drehmann, Mathias, Andrew J. Patton, and Steffen Sorensen. 2007. Non-linearities and stress testing. In *Risk Measurement and Systemic Risk*, 281–308. Frankfurt: European Central Bank.

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

Dube, Arindrajit, Daniele Girardi, Òscar Jordà, and Alan M Taylor. 2023. A Local Projections Approach to Difference-in-Differences Event Studies. NBER Working Paper 31184.

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

Evans, G. B. A., and N. E. Savin. 1981. Testing For Unit Roots: 1. *Econometrica* 49(3): 753–779.

Fernández-Villaverde, Jesús, Juan F. Rubio-Ramírez, Thomas J. Sargent, and Mark W. Watson. 2007. ABCs (and Ds) of understanding VARs. *American Economic Review* 97(3): 1021–1026.

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

Fortin, Nicole, Thomas Lemieux, and Sergio Firpo. 2011. Decomposition Methods in Economics. In *Handbook of Labor Economics*, edited by Orley Ashenfelter, and David Card, volume 4, 1–102. Amsterdam: Elsevier.

Gonçalves, Sílvia, Ana María Herrera, Lutz Kilian, and Elena Pesavento. 2024. State-Dependent Local Projections. *Journal of Econometrics,* forthcoming.

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

Guajardo, Jaime, Daniel Leigh, and Andrea Pescatori. 2014. Expansionary Austerity? International Evidence. *Journal of the European Economic Association* 12(4): 949–968.

Guerrón-Quintana, Pablo, Atsushi Inoue, and Lutz Kilian. 2017. Impulse response matching estimators for DSGE models. *Journal of Econometrics* 196(1): 144–155.

Hall, Alastair R., Atsushi Inoue, James M. Nason, and Barbara Rossi. 2012. Information criteria for impulse response function matching estimation of DSGE models. *Journal of Econometrics* 170(2): 499–518.

Hamilton, James D. 1994a. State-space models. In *Handbook of Econometrics*, edited by Robert F. Engle and Daniel L. McFadden, volume 4, chapter 50, 3039–3080. Amsterdam: Elsevier.

Hamilton, James D. 1994b. *Time Series Analysis*. Princeton, N.J.: Princeton University Press.

Harvey, Andrew. 1991. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Herbst, Edward, and Benjamin K. Johannsen. 2024. Bias in Local Projections. *Journal of Econometrics* 240(105655).

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4): 1161–1189.

Horvitz, Daniel G., and Donovan J. Thompson. 1952. A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association* 47(260): 663–685.

Iacoviello, Matteo. 2005. House prices, borrowing constraints, and monetary policy in the business cycle. *American Economic Review* 95(3): 739–764.

Inoue, Atsushi, Òscar Jordà, and Guido M. Kuersteiner. 2024. Inference for Local Projections. *The Econometrics Journal,* forthcoming.

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

Inoue, Atsushi, and Lutz Kilian. 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, Martin Kornejew, Moritz Schularick, and Alan M. Taylor. 2022. Zombies at large? Corporate debt overhang and the macroeconomy. *Review of Financial Studies* 35(10): 4561–86.

Jordà, Òscar, and Sharon Kozicki. 2011. Estimation and inference by the method of projection minimum distance: An application to the new Keynesian hybrid Phillips curve. *International Economic Review* 52(2): 461–487.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor. 2015. Betting the house. *Journal of International Economics* 96(S1): 2–18.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor. 2017. Macrofinancial History and the New Business Cycle Facts. *NBER Macroeconomics Annual* 31: 213–263.

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

Jordà, Òscar, and Alan M. Taylor. 2016. The Time for Austerity: Estimating the Average Treatment Effect of Fiscal Policy. *Economic Journal* 126(590): 219–255.

Kendall, M. G. 1954. Note on Bias in the Estimation of Autocorrelation. *Biometrika* 41(3/4): 403–404.

Kilian, Lutz. 1998. Small-sample Confidence Intervals for Impulse Response Functions. *Review of Economics and Statistics* 80(2): 218–230.

Kilian, Lutz. 1999. Finite-sample properties of percentile and percentile-t bootstrap confidence intervals for impulse responses. *Review of Economics and Statistics* 81(4): 652–660.

Kilian, Lutz, and Helmut Lütkepohl. 2017. *Structural Vector Autoregressive Analysis*. Themes in Modern Econometrics. Cambridge: Cambridge University Press.

Kitagawa, Evelyn M. 1955. Components of a difference between two rates. *Journal of the American Statistical Association* 50(272): 1168–94.

Leeper, Eric M., and Tao Zha. 2003. Modest policy interventions. *Journal of Monetary Economics* 50(8): 1673–1700.

Lewis, Daniel J., and Karel Mertens. 2022. Dynamic Identification Using System Projections and Instrumental Variables. CEPR Discussion Paper 17153.

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

Linnemann, Ludger, and Roland Winkler. 2016. Estimating nonlinear effects of fiscal policy using quantile regression methods. *Oxford Economic Papers* 68(4): 1120–45.

Lucas, Robert E. 1976. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy* 1: 19–46.

Lusompa, Amaze B. 2018. U.S. Fiscal Mulitpliers: Time-Varying, Asymmetric, or Both? Unpublished.

Marriott, F. H. C., and J. A. Pope. 1954. Bias in the Estimation of Autocorrelations. *Biometrika* 41(3/4): 390–402.

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

Mikusheva, Anna. 2012. One-Dimensional Inference in Autoregressive Models With the Potential Presence of a Unit Root. *Econometrica* 80(1): 173–212.

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

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

Mountford, Andrew, and Harald Uhlig. 2009. What are the effects of fiscal policy shocks? *Journal of Applied Econometrics* 24(6): 960–992.

Newey, Whitney K. 1990. Efficient instrumental variables estimation of nonlinear models. *Econometrica* 809–837.

Newey, Whitney K., and Daniel McFadden. 1986. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, edited by Engle, R. F., and D. McFadden, volume 4 of *Handbook of Econometrics*, chapter 36, 2111–2245. Amsterdam: Elsevier.

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

Nicholls, Desmond F., and Alun L. Pope. 1988. Bias in the estimation of multivariate autoregressions. *Australian Journal of Statistics* 30(1): 296–309.

Nickell, Stephen. 1981. Biases in Dynamic Models with Fixed Effects. *Econometrica* 49(6): 1417–1426.

Oaxaca, Ronald. 1973. Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review* 14(3): 693–709.

Orcutt, G. H. 1948. A Study of the Autoregressive Nature of the Time Series Used for Tinbergen's Model of the Economic System of the United States, 1919–1932. *Journal of the Royal Statistical Society: Series B (Methodological)* 10(1): 1–45.

Owen, Art B. 1988. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75(2): 237–249.

Pesavento, Elena, and Barbara Rossi. 2006. Small-sample confidence intervals for multivariate impulse response functions at long horizons. *Journal of Applied Econometrics* 21(8): 1135–1155.

Pesavento, Elena, and Barbara Rossi. 2007. Impulse response confidence intervals for persistent data: What have we learned? *Journal of Economic Dynamics and Control* 31(7): 2398–2412.

Phillips, Peter C. B. 1998. Impulse response and forecast error variance asymptotics in nonstationary VARs. *Journal of Econometrics* 83(1-2): 21–56.

Piger, Jeremy, and Thomas Stockwell. 2023. Differences from Differencing: Should Local Projections with Observed Shocks be Estimated in Levels or Differences? Unpublished. https://ssrn.com/abstract=4530799.

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

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

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

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

Primiceri, Giorgio E. 2005. Time varying structural vector autoregressions and monetary policy. *Review of Economic Studies* 72(3): 821–852.

Rambachan, Ashesh, and Neil Shephard. 2019a. Econometric analysis of potential outcomes time series: instruments, shocks, linearity and the causal response function. Unpublished. https://arxiv.org/abs/1903.01637.

Rambachan, Ashesh, and Neil Shephard. 2019b. A nonparametric dynamic causal model for macroeconometrics. Unpublished. https://ssrn.com/abstract=3345325.

Ramey, Valerie A. 2016. Macroeconomic Shocks and Their Propagation. In *Handbook of Macroeconomics*, edited by Taylor, John B., and Harald Uhlig, volume 2, chapter 2, 71–162. Amsterdam: Elsevier.

Ramey, Valerie A., and Sarah Zubairy. 2018. Government Spending Multipliers in Good Times and in Bad: Evidence from US Historical Data. *Journal of Political Economy* 126(2): 850–901.

Romer, Christina D., and David H. Romer. 2004. A new measure of monetary shocks: Derivation and implications. *American Economic Review* 94(4): 1055–1084.

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

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1): 41–55.

Rotemberg, Julio J., and Michael Woodford. 1997. An Optimization-Based Econometric Framework for the Evaluation of Monetary Policy. *NBER Macroeconomics Annual* 12: 297–346.

Roth, Jonathan, Pedro H.C. Sant'Anna, Alyssa Bilinski, and John Poe. 2023. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics* 235(2): 2218–2244.

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

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

Stuart, Alan, and Keith Ord. 2010. *Kendall's Advanced Theory of Statistics, Distribution Theory*, volume 1. New York: Wiley.

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

Tenreyro, Silvana, and Gregory Thwaites. 2016. Pushing on a String: US Monetary Policy Is Less Powerful in Recessions. *American Economic Journal: Macroeconomics* 8(4): 43–74.

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

Uhlig, Harald. 2010. Some fiscal calculus. *American Economic Review* 100(2): 30–34.

White, John S. 1957. Approximate Moments for the Serial Correlation Coefficient. *Annals of Mathematical Statistics* 28(3): 798–802.

Wieland, Johannes F., and Mu-Jeung Yang. 2020. Financial Dampening. *Journal of Money, Credit and Banking* 52(1): 79–113.

Wolf, Michael, and Dan Wunderli. 2015. Bootstrap joint prediction regions. *Journal of Time Series Analysis* 36(3): 352–376.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Mass.: MIT Press, 2nd edition.

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