Local Projections or VARs?
A Primer for Macroeconomists*

José Luis Montiel Olea
Cornell University

Mikkel Plagborg-Møller
Princeton University

Eric Qian
Princeton University

Christian K. Wolf
MIT & NBER

Preliminary and incomplete – comments welcome!

January 23, 2025

Abstract: What should applied macroeconomists know about local projection (LP) and vector autoregression (VAR) impulse response estimators? The two methods share the same estimand, but in finite samples lie on opposite ends of a bias-variance trade-off. While the low bias of LPs comes at a quite steep variance cost, this cost must be paid to achieve robust uncertainty assessments. VARs should thus only be used with long lag lengths, ensuring equivalence with LP. For LP estimation, we provide guidance on selection of lag length and controls, bias correction, and standard error construction.

Keywords: dynamic causal effect, impulse response, local projection, misspecification, vector autoregression. *JEL codes:* C22, C32.

1 Introduction

Applied macroeconomists routinely seek to estimate the dynamic causal effects—or impulse response functions—of aggregate shocks to economic policies or fundamentals. The two

*Email: montiel.olea@gmail.com, mikkelpm@princeton.edu, ericqian@princeton.edu, and ckwolf@mit.edu. Prepared for the NBER Macroeconomics Annual 2025. We are grateful for comments from Isaiah Andrews, Jim Stock, and Iván Werning. Plagborg-Møller acknowledges that this material is based upon work supported by the NSF under Grant #2238049 and by the Alfred P. Sloan Foundation, and Wolf does the same for NSF Grant #2314736.

dominant empirical methods for doing so are structural vector autoregressions (VARs), going back to Sims (1980), and local projections (LPs), as introduced by Jordà (2005). The primary objectives of this paper are, first, to review how to think conceptually about the choice between these two estimation methods, and second, to offer concrete recommendations for applied practice.

A first important observation is that the choice between LPs and VARs has absolutely nothing to do with questions of identification: for any given LP and the economic identifying assumptions that it implements, there exists an equivalent VAR, and *vice versa*. At their core, typical identification schemes in empirical macroeconomics propose to recover the causal effects of aggregate shocks or policies as some (sometimes quite complicated) function of the autocovariances of time series data. Conceptually, LPs and VARs are simply two ways of estimating their common large-sample estimand, though they differ in how they exploit a given finite data set.

In the short samples typical of applied work in macroeconomics, the choice between LPs and VARs is one of navigating a bias-variance trade-off. At one end of this spectrum, lag-augmented LPs (or equivalently, VARs with many lags) robustly have low bias, though at the cost of materially elevated variance. At the other end, (short-lag) VARs tend to deliver very sizable precision gains, but at the cost of often substantial biases. The intuition is that short-lag VARs extrapolate based only on the first few autocovariances of the data, while LPs flexibly estimate autocovariances at all horizons, without extrapolation. In practice, the variance cost of LPs tends to be so substantial that VARs are typically preferable in terms of mean squared error. The VAR's bias, however, very severely threatens the accuracy of its uncertainty assessments, while LP confidence intervals instead accurately reflect statistical uncertainty, by virtue of the LP's robustly small bias.

Our overall recommendation is that, since accurate communication of estimation uncertainty is integral in applied macroeconometric practice, researchers should estimate LPs (or equivalently very long-lag VARs). We discourage the use of the short-lag VAR specifications often encountered in the empirical literature, because of the inevitable fragility of their uncertainty assessments. We conclude with a "how-to" list of practical recommendations for LP estimation, covering the selection of control variables and lag length, bias correction, and standard error construction.

Due to our focus on simple, concrete take-aways for applied practice, our review of the literature leaves out several specialized or new topics. These include panel data (Almuzara and Sancibrian, 2024), nonlinear specifications (Caravello and Martinez-Bruera, 2024; Gonçalves, Herrera, Kilian, and Pesavento, 2024), simultaneous confidence bands (Montiel Olea and Plagborg-Møller, 2019), variance decompositions (Plagborg-Møller and Wolf, 2022), and certain more technical structural shock identification schemes. Other excellent reviews include Kilian and Lütkepohl (2017), Stock and Watson (2018), and Jordà and Taylor (2024).

Outline. The organization of the paper follows the outline above: we first review questions of identification, then discuss and quantify the bias-variance trade-off between LPs and VARs, next ask how to navigate that trade-off in practice, and finally close with a list of concrete practical recommendations. Throughout, our analysis will be deliberately simple, emphasizing clarity over comprehensiveness; for the interested reader we give references to the original literature. To structure our discussion, we summarize our main insights as "lessons" that should be viewed as guidelines rather than as formal propositions. Supplementary details—in particular for our empirically calibrated simulations—are provided in the online appendices.

2 Identification

In this section we review the basic definition of the LP and VAR estimators, and we establish that they share the exact same estimand (i.e., large-sample limit) when the estimation lag length is large. The purpose of economic identifying assumptions is to establish that this common estimand is interesting. It follows that the choice between LPs and VARs is completely orthogonal to questions of identification.

2.1 What do LPs estimate?

We begin with a discussion of the estimand of LPs. LP practitioners seek to estimate the dynamic causal effects of macroeconomic "shocks" on aggregate outcomes. To this end, they run linear regressions of the following form, estimated separately by ordinary least squares (OLS) for each horizon $h = 0, 1, 2, \ldots$:

$$y_{t+h} = \mu_h + \theta_h^{\text{LP}} x_t + \gamma_h' r_t + \sum_{\ell=1}^{p} \delta_{h,\ell}' w_{t-\ell} + \xi_{h,t}. \tag{2.1}$$

They then report as the impulse responses of interest the regression coefficients $\{\theta_h^{\text{LP}}\}_h$. In the above regression, y_t is a (scalar) outcome of interest, x_t is a (scalar) impulse variable, r_t

is a vector of time series that is included as contemporaneous controls, $w_t = (r_t', x_t, y_t, q_t')'$ collects all variables included as lagged controls, with q_t a potential additional control vector, and $\xi_{h,t}$ is the multi-step forecast error in the regression. Here and in the rest of this section, we consider the limit where the empirical sample size is infinitely large, to abstract from finite-sample issues that will be the focus of the following sections.

By the Frisch-Waugh-Lovell theorem, the regression coefficient θ_h^{LP} has a simple interpretation: it equals the coefficient from a projection of the outcome y_{t+h} on a "shock" $\tilde{x}_t$ that is given by the residual from a projection of x_t on the control variables in the regression (2.1):

$$\tilde{x}_t \equiv x_t - \text{proj}(x_t \mid r_t, w_{t-1}, \ldots, w_{t-p}). \tag{2.2}$$

Lesson 1. *In an LP, we are estimating impulse responses with respect to a shock that is defined as the residual from projecting the impulse variable on the control variables.*

This LP estimand is economically interesting because, under some assumptions, it can be given a structural interpretation. The argument is most transparent when a researcher directly observes an unpredictable shock (or a valid proxy/instrument for that shock), e.g., by observing the high-frequency responses of an asset price in narrow time windows around policy announcements (as in Känzig, 2021). In this case, she can straightforwardly use the LP (2.1) to estimate the causal effects of that shock: the impulse variable x_t is the observed shock, the contemporaneous control vector r_t is empty, and finally further controls q_t are not necessary for consistent estimation, but should be included for efficiency and robustness reasons, as discussed in Section 5.[1] Note that in this case the residualized shock $\tilde{x}_t$ simply equals the shock x_t itself, by virtue of being unpredictable.

In another common class of applications, the impulse variable x_t equals a policy instrument (such as the Federal Funds Rate), the contemporaneous controls r_t equal endogenous variables in the policymaker's reaction function (such as output and inflation), and further lagged controls q_t may then be included as well. Then the residualized shock $\tilde{x}_t$ equals the disturbance in the policy rule (such as a monetary policy shock), under the classic timing assumption that the endogenous variables r_t do not respond within the period to this disturbance (Christiano, Eichenbaum, and Evans, 1999).[2] Thus, the LP again measures the causal

[1] If the researcher is willing to assume an underlying *linear* Structural Vector Moving Average (SVMA) model, then the LP regression coefficients equal the shock's true impulse responses (up to scale). If instead she assumes a more general non-linear causal model, then the estimand is a particular weighted average of marginal effects (see Plagborg-Møller and Wolf, 2021; Kolesár and Plagborg-Møller, 2024).

[2] By defining the shock as the residual in a policy rule where all variables are observed by the econome-

effects of this disturbance. We will elaborate on other, more involved ways of identifying macroeconomic shocks in Section 2.3.

2.2 What do VARs estimate?

We now consider the second estimation method: VARs. The first step in VAR analysis is to estimate a reduced-form VAR in the vector of all observed time series w_t (using the same definition as above):

$$w_t = c + \sum_{\ell=1}^{p} A_\ell w_{t-\ell} + u_t. \tag{2.3}$$

VAR practitioners translate this reduced-form model into structural impulse response functions using identification conditions. We begin with one popular way of doing so: orthogonalizing the reduced-form residuals u_t using what is known as a recursive ordering, and then propagating forward the orthogonalized residuals through the VAR (2.3). Let $BB' = \Sigma$ be the Cholesky decomposition of the covariance matrix $\Sigma \equiv \text{Var}(u_t)$ of the forecast errors u_t, where B is a lower-triangular matrix with positive diagonal entries. Given that we are interested in the effect of the orthogonalized shock to the impulse variable x_t on the outcome y_t, let e_x and e_y denote the two unit vectors such that $x_t = e_x' w_t$ and $y_t = e_y' w_t$. Then the structural VAR impulse response estimate equals

$$\theta_h^{\text{VAR}} \equiv e_y' C_h B e_x,$$

where the reduced-form impulse responses satisfy the recursion $C_h = \sum_{\ell=1}^{\min\{h,p\}} A_\ell C_{h-\ell}$, with C_0 equal to the identity matrix. Evidently, the VAR impulse responses are obtained by extrapolation: the parameters of the reduced-form VAR (2.3) are estimated from observed autocovariances out to lag p, and then the parametric structure of the model is exploited to compute responses at all horizons h, including $h > p$.

By standard properties of Cholesky decompositions, this procedure isolates as the "shock" the (rescaled) residual in a projection of $u_{x,t} \equiv e_x' u_t$ on $u_{r,t} \equiv e_r' u_t$, where e_r is the unit vector such that $r_t = e_r' w_t$. By definition of the residuals u_t, this shock is the same as the residual in a projection of x_t on r_t and p lags of the observables w_t—which in turn is precisely the shock $\tilde{x}_t$ in the LP (2.1). Differently from the LP, however, the impulse response coefficients $\{\theta_h^{\text{VAR}}\}_h$ are now not direct projection coefficients of future outcomes $\{y_{t+h}\}_h$ on this common

trician, we are implicitly imposing the assumption of invertibility, i.e., that the policy shock is spanned by current and lagged observed variables. We will comment further on this structural assumption in Section 2.3.

"shock;" instead, the dynamic effects of the shock are iteratively propagated forward through the estimated VAR model (2.3).

The simple linear regression logic that we just used to interpret the LP and VAR estimands reveals that the methods are intimately connected: they deliver the dynamic causal effects of the exact same implicit "shock" $\tilde{x}_t$, just in one case through direct projection (LP), in the other through iterative propagation (VAR). On impact ($h = 0$), they are thus necessarily the same.[3] Furthermore, as the lag length p becomes arbitrarily large, the extrapolative reduced-form VAR model (2.3) is sufficiently flexible to perfectly capture all autocovariance properties of the data, making iterative forecasts equivalent to direct projections; thus, the LP and VAR estimands are identical at *all* horizons h (Plagborg-Møller and Wolf, 2021; Xu, 2023).[4] Finally, for general but finite estimation lag length p, the extrapolative VAR model still captures the relevant autocovariance properties well out to lag p, so that the LP and VAR estimands are typically very close at horizons $h \leq p$, but not necessarily for $h > p$.[5]

Lesson 2. *An impulse response from an LP can equivalently be viewed as having been obtained from a VAR that controls for a large number p of lags. In particular, impulse responses from LPs and VARs tend to be very similar at horizons $h \leq p$.*

We provide a visual illustration of the second lesson in Figure 2.1, which plots LP and VAR impulse response estimands for a recursively identified monetary policy shock. The underlying data generating process (DGP) is the dynamic factor model that we study later in this paper (see Supplemental Appendix E.3 for implementation details); importantly, that GDP does *not* satisfy a finite-order VAR model. The three panels then plot the estimands for estimation lag lengths $p = \{2, 6, 12\}$ (in red dashed and blue dashed-dotted, respectively), as well as the (common) population estimand ($p = \infty$, in grey). Consistent with our discussion above, LP and VAR estimands are close to each other at horizons $h \leq p$, before then deviating; in particular, for p sufficiently large, they are identical throughout.

[3] Here we abstract from an inessential issue: the implied LP and VAR shocks may have different variances. But once rescaled to have the same units (e.g., standard deviation equal to 1), equivalence follows.

[4] Though this result pertains to linear estimation methods, the equivalence is *nonparametric* in the sense that it does not impose any particular parametric model on the underlying data generating process.

[5] See Plagborg-Møller and Wolf (2021) for a general discussion of the finite-p case. As discussed there, the LP and VAR estimands at $h \leq p$ for finite p in general differ slightly because the computed impulse responses depend on autocovariances up to lag $p + h$, not just up to p.

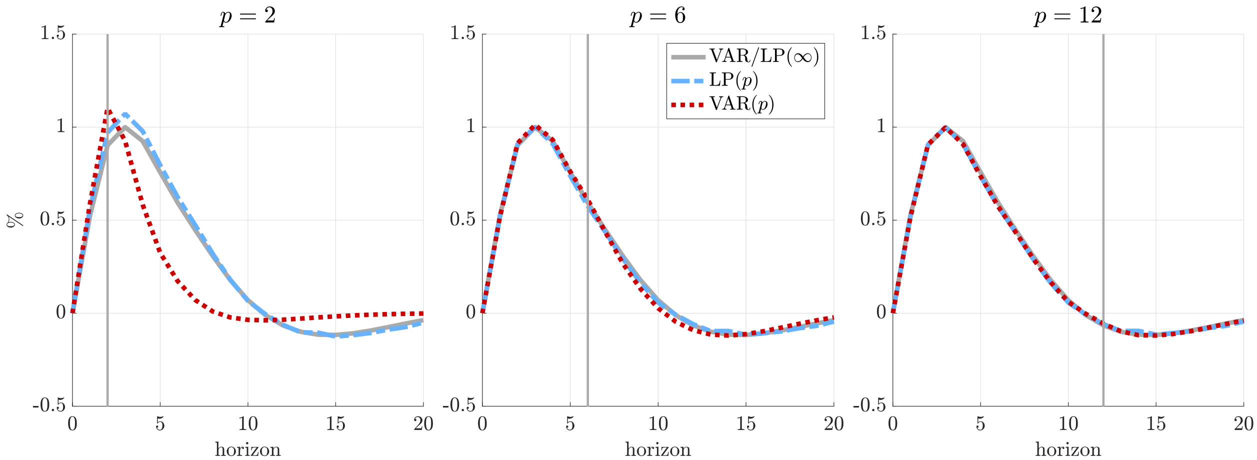


Figure 2.1: LP (blue, dashed-dotted) and VAR (red, dashed) impulse response estimands in a dynamic factor model. The panels show the response of output to a recursively identified monetary policy shock, with lag lengths $p = \{2, 6, 12\}$. The grey line is the $p = \infty$ population estimand (for both LP and VAR), and the horizontal line marks the lag length p.

2.3 More general identification approaches

The large-sample equivalence between LPs and VARs extends beyond the observed shock and recursive identification schemes discussed above. A common approach to identifying macroeconomic shocks begins with the assumption of "invertibility", i.e., that the true shocks are spanned by current and lagged time series observables w_t (see Fernández-Villaverde, Rubio-Ramírez, Sargent, and Watson, 2007, for a standard treatment). Identification schemes like the recursive ordering of Christiano, Eichenbaum, and Evans (1999)—which, as we showed above, can equivalently be implemented using either LPs or VARs—combine the assumption of invertibility with short-run timing restrictions on the shocks. Invertibility may however also be combined with other restrictions to identify shocks, including for example long-run restrictions (Blanchard and Quah, 1989) or sign restrictions (Uhlig, 2005). In all these identification approaches, the structural shock of interest ultimately equals $\beta' u_t$, where u_t is the reduced-form forecast error from (2.3), and the weight vector β is a (potentially complicated) function of the researcher's identifying assumptions. A structural VAR practitioner would then report the impulse response estimate $\theta_h^{\text{VAR}} = e_y' C_h \Sigma \beta$. But by our earlier Frisch-Waugh-Lovell logic, we can alternatively compute impulse responses with respect to the same shock $\beta' u_t$ from the more general LP

$$y_{t+h} = \mu_h + \theta_h^{\text{LP}} (\beta' w_t) + \sum_{\ell=1}^{p} \delta_{h,\ell}' w_{t-\ell} + \xi_{h,t}. \tag{2.4}$$

Hence, either VARs or LPs can be used to compute this shared estimand of interest, and the two methods will agree in large samples when the estimation lag length p is large.

Lesson 3. *When it comes to identification, anything you can do with VARs, you can do with LPs, and* vice versa.[6]

In conclusion, the choice between LPs and VARs has nothing to do with the economic identifying assumptions necessary to isolate a macroeconomic shock of interest. Instead, the question is simply which estimation method is better at recovering their common estimand in finite samples. This question of *estimation*—and not *identification*—is the focus of the remainder of the paper.

3 The bias-variance trade-off

We now dig deeper into the key econometric properties of LPs and VARs. First, we motivate our analysis by demonstrating that, despite the equivalence between LPs and VARs when the lag length is very large, the choice between the two estimators does matter in practice when using a small or a moderate lag length, as typically seen in applied work. Second, we present simple illustrative simulations of the properties of LP and VAR estimators. Finally, we review the available econometric theory. The overarching takeaway of this section will be that, in finite data sets, there is a clear bias-variance trade-off between LPs and (short-lag) VARs: small bias and large variance for LPs, and *vice versa* for VARs.

3.1 VARs vs. LPs in empirical work

We first establish that, in finite samples and with a moderate estimation lag length, LPs and VARs can provide meaningfully different estimates of their common estimand (that would obtain in large samples if we were able to control for very many lags). Our analysis is based on the literature synthesis of Ramey (2016). We replicate four of the headline applications of that paper, for shocks to monetary policy, taxes, government purchases, and technology, respectively. We then use LPs and VARs to estimate impulse responses for several variables and at several horizons, throughout staying as closely as possible to the exact specifications considered in Ramey (2016). Across all response variables and horizons, we then compare

[6] See Plagborg-Møller and Wolf (2021) for a detailed discussion of how exactly identification schemes like long-run and sign restrictions map into β.

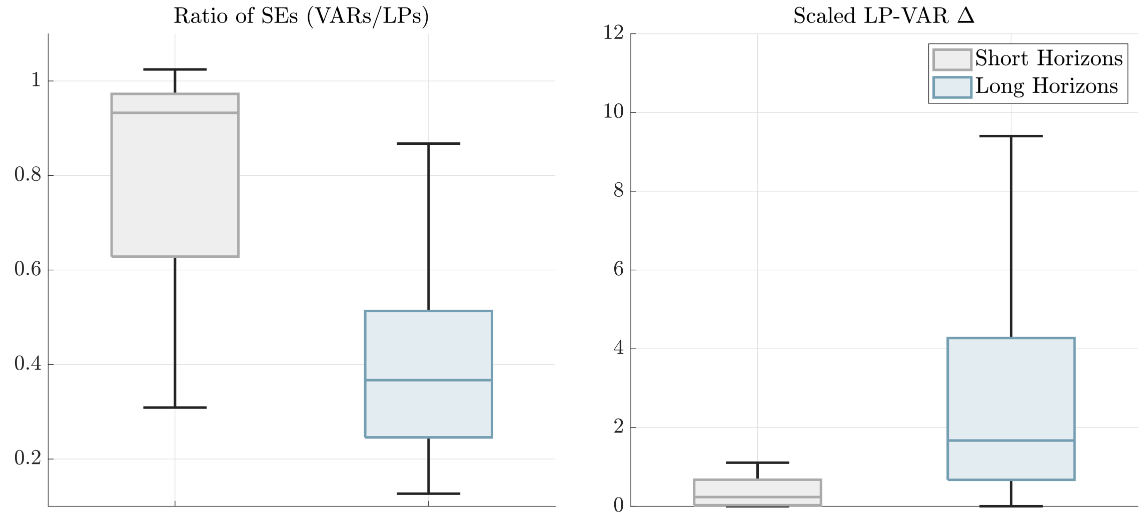


Figure 3.1: Box plots of VAR-to-LP standard error ratios (left panel) and normalized point estimate differences (right panel) for short horizons (blue) and for long horizons (grey). The central mark indicates the median, the bottom and top edges of the box indicate the 25th and 75th percentiles, and the whiskers extend to the most extreme data points that are not more than 1.5 times the interquartile range away from the bottom or top of the box (i.e., ignoring outliers). Based on applications in Ramey (2016). See Supplemental Appendix C for details.

standard errors and compute the differences in point estimates. Implementation details are provided in Supplemental Appendix C.

Figure 3.1 shows that there are meaningful differences in both precision and location of the two estimators: VAR impulse responses often have substantially lower standard errors than LPs (left panel), and the two sets of point estimates can be quite far apart (right panel). In the left panel we display a box plot of the ratio of VAR and LP standard errors across all different shocks and outcome variables, separately for short horizons (blue, $\leq$ one year) as well as long horizons (grey, $>$ one year). At short horizons, VARs are only somewhat more precise than LPs, with the median standard error ratio only slightly below one, consistent with our theoretical discussion in Section 2.[7] At long horizons, VARs are instead materially more precise, yet again consistent with the preceding discussion. The right panel then shows the difference in point estimates between the two methods, normalized by the VAR standard error. We see that differences can be quite material, often being of much greater magnitude than the VAR standard error. Furthermore, also echoing our preceding discussion, the gaps are smaller at short impulse response horizons.

[7] Asymptotically, VAR standard errors are weakly smaller than LP standard errors, so the standard error ratio is bounded above by 1, as we will explain later. This need not be the case in small samples, however.

Lesson 4. *At intermediate and long horizons, LP and VAR impulse response estimates in empirical applications are often materially different, and VAR estimates typically have much lower standard errors than LP.*

These empirical results can of course not directly tell us whether the LP or the VAR estimates tend to be closer to the truth. For this, we will next turn to a simulation exercise, before presenting the econometric theory.

3.2 A numerical illustration

As a first step toward gaining intuition for the econometric properties of LPs and VARs, we present some illustrative simulations. For pedagogical reasons, we here use a simple univariate ARMA$(1, 1)$ data generating process, while leaving empirically realistic simulations to later sections. We consider the DGP

$$y_t = \rho y_{t-1} + \varepsilon_t + \alpha \varepsilon_{t-1}, \quad \varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma^2), \tag{3.1}$$

where y_t is an observed scalar outcome variable, ε_t is an unobserved scalar shock, and ρ and α are parameters. We are interested in the impulse response of y_t with respect to ε_t at horizon h, which has the formula $\theta_h \equiv \rho^h + \alpha\rho^{h-1}$ for $h \geq 1$ (and $\theta_0 = 1$).

Figure 3.2 shows that, in this DGP, there is a stark bias-variance trade-off between LPs and (short-lag) VARs. The figure shows density plots of the LP (blue) as well as VAR (red) estimators of the impulse response at horizon $h = 2$ for $\rho = 0.85$, $\alpha = 0.1$, and with a sample size of $T = 240$. Both estimators control for one lag of the data, with larger lag lengths to be considered below. We can see that the LP estimator is centered close to the true value of the impulse response. Intuitively, because LP just directly projects the future outcome y_{t+h} of interest on y_t (controlling for y_{t-1}), it is able to pick up on the full ARMA dynamics of the DGP (3.1). The VAR estimator, in contrast, suffers from extrapolation bias: its impulse response estimate $\hat{\rho}^h$ is obtained by first estimating the autoregressive parameter $\hat{\rho}$ through a *one*-period-ahead forecast regression of y_t on y_{t-1}, and then iterating forward h steps using the parametric AR(1) model. The VAR estimator therefore is not directly informed by the sample autocovariances at horizons 2 and longer, causing it to miss out on the more intricate moving average dynamics in the DGP (3.1). Yet, the figure also shows that the parametric extrapolation has a clear benefit: because the first-order autoregressive coefficient $\hat{\rho}$ is more precisely estimated than longer-horizon autocovariances, the sampling distribution of the VAR impulse response estimator is less dispersed than the LP estimator. In summary, the

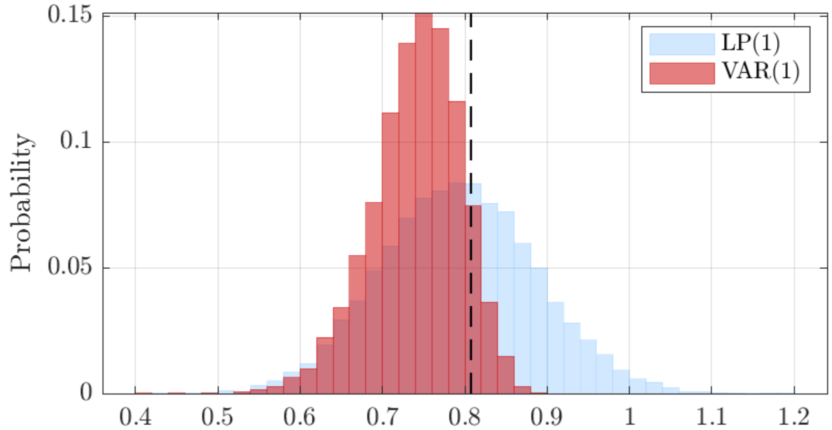
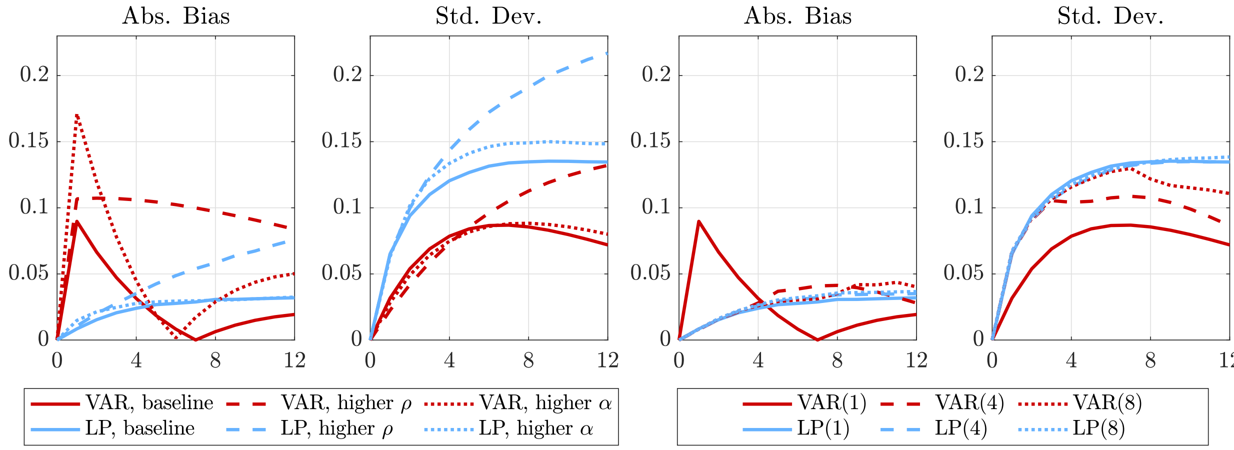


Figure 3.2: Histogram of VAR(1) and LP(1) impulse response estimates for $h = 2$, $\rho = 0.85$, and $\alpha = 0.1$ relative to the true impulse response θ_h (dashed line).

sampling distribution of the LP estimator is well-centered but dispersed, while that of the VAR estimator is more tightly concentrated but centered incorrectly.

Panel (a) of Figure 3.3 exhibits how the nature of the bias-variance trade-off differs across DGPs and impulse response horizons. On impact, $h = 0$, the two estimators are numerically identical (they both equal 1). However, at intermediate horizons the bias-variance trade-off is stark, for the reasons discussed earlier. The larger the moving average coefficient α, the more misspecified is the one-lag VAR specification, and consequently the larger is the VAR bias. The VAR bias then decreases at long horizons, since in stationary DGPs the impulse responses must converge to zero as $h \to \infty$, a feature that is mechanically enforced by the VAR estimator $\hat{\rho}^h$ (but not by the LP estimator). However, the figure also reveals that the speed of convergence of the impulse responses to zero depends on the persistence parameter ρ, so it is not clear *a priori* at what horizons we should expect the impulse responses to be small—and thus when we can stop worrying about the VAR bias.

Finally, panel (b) of Figure 3.3 shows how changes in the estimation lag length p shape the bias-variance trade-off, demonstrating that the bias of the VAR estimator at shorter horizons is reduced when the lag length is increased. In this panel, the LP and VAR estimators now control for p lags of the data $y_{t-1}, \dots, y_{t-p}$ rather than just one. When we control for $p = 4$ lags, the VAR impulse response estimator is now nearly unbiased out to horizon $h = 4$. This comes at a cost, however: the VAR variance increases substantially, to essentially equal that of LP at those horizons. When we control for even more lags, the VAR bias is reduced at even longer horizons, but again at the cost of higher variance. These results again illustrate our



(a) Alternative DGPs (b) Alternative lag lengths

Figure 3.3: Absolute bias and standard deviation of LP and VAR estimators, as a function of horizon h. In the left panel, VAR and LP are estimated with 1 lag. Compared to the baseline DGP ($\rho = 0.85$ and $\alpha = 0.1$), "higher ρ" takes $\rho = 0.95$ and "higher α" takes $\alpha = 0.2$. The right panel varies the estimation lag length p for the baseline DGP.

second lesson: an LP is essentially a VAR that controls for a lot of lags. Consistent with this interpretation, the bias and standard deviation of the VAR estimator depend dramatically on the choice of lag length, while this choice matters much less for the LP estimator.

In the next section we will establish that these simulation-based conclusions in fact hold up theoretically in a much wider class of multivariate models. Subsequent sections will then use large-scale, empirically calibrated simulation studies to demonstrate that our takeaways are practically relevant, in addition to being theoretically well-founded.

3.3 Theoretical insights

To gain a deeper understanding of the bias-variance trade-off, we review some of the main theoretical results from Montiel Olea, Plagborg-Møller, Qian, and Wolf (2024).

Univariate model. We begin our analytical investigations with the univariate ARMA$(1, 1)$ model (3.1). Montiel Olea, Plagborg-Møller, Qian, and Wolf show that the VAR and LP estimators of the true impulse response θ_h are both approximately normally distributed in large samples, though with differing bias and variance:

$$\hat{\theta}_{h,\text{VAR}} \overset{\cdot}{\sim} N\left(\theta_h + b_h(p), \tau^2_{h,\text{VAR}}(p)\right), \quad \hat{\theta}_{h,\text{LP}} \overset{\cdot}{\sim} N\left(\theta_h, \tau^2_{h,\text{LP}}\right), \tag{3.2}$$

where $\dot\sim$ denotes "approximate distribution in large samples" (in the usual formal sense of asymptotic normality), $b_h(p)$ is a VAR bias term that depends on the estimation lag length p, and the large-sample variances satisfy $\tau^2_{h,\text{LP}} \geq \tau^2_{h,\text{VAR}}(p)$; we will return later to the observation that the LP large-sample variance does not depend on the estimation lag length. We note that the formal result requires the moving average coefficient α in the DGP (3.1) to be relatively small.[8] Thus, our analysis gives the VAR estimator the benefit of the doubt by viewing the true DGP as close to, but not exactly equal to, an AR(1), thereby yielding a non-trivial trade-off between model misspecification and statistical sampling error. The popularity and the documented empirical success of autoregressive estimators in the existing forecasting literature suggest that carefully implemented autoregressive specifications are often not grossly misspecified, though few economists would likely believe that such simple parametric models can in fact successfully capture every aspect of the real world.

We now study in greater detail the bias and variance expressions in (3.2).

- *Bias.* The first important implication of (3.2) is that, while the LP estimator has approximately zero bias in large samples, the VAR estimator is generally biased. When a single lag $p = 1$ is used for estimation, the VAR bias has the simple formula

$$b_h(1) = \underbrace{h\rho^{h-1}}_{\frac{\partial(\rho^h)}{\partial\rho}} \underbrace{(1-\rho^2)\alpha}_{\approx\frac{\text{Cov}(y_{t-1},\alpha\varepsilon_{t-1})}{\text{Var}(y_{t-1})}} - \underbrace{\rho^{h-1}\alpha}_{\theta_h-\rho^h}.$$

  The product of the first two factors on the right-hand side is due to the endogeneity bias in the estimated autoregressive coefficient $\hat\rho$, here caused by the lagged error term $\alpha\varepsilon_{t-1}$ in the regression of y_t on y_{t-1}. The second factor is the bias of $\hat\rho$, while the first factor maps ρ into the impulse response of interest. The last term in the bias expression is due to the VAR estimator then relying on a misspecified functional form to transform $\hat\rho$ into impulse responses $\hat\rho^h$. It can be shown that both these sources of

[8] Formally, α needs to be proportional to the magnitude of the standard deviation of the estimators (i.e., $\alpha = \alpha_T \propto T^{-1/2}$). If we instead analyzed the asymptotic properties of fixed-lag VAR estimators under the DGP with fixed parameter α (as in Braun and Mittnik, 1993), the conclusions would be stark but empirically uninteresting: for any moving average coefficient $\alpha \neq 0$, the VAR estimator would be inconsistent for the true impulse response due to misspecification, so for large sample sizes the ratio of the bias to the standard deviation of the estimator diverges to infinity, yielding a trivial bias/variance trade-off in the limit. The "local-to-zero" modeling device of setting $\alpha \propto T^{-1/2}$ should not be viewed as a literal description of reality, but rather as a technical device intended to tractably and accurately capture key finite-sample phenomena, similar to the econometric literatures on weak instruments or near-unit roots. Our later simulations will show that the lessons learned from these local-to-zero asymptotics are borne out in realistic DGPs.

bias can be reduced by increasing the estimation lag length p; however, as we will see, this necessarily comes at the cost of higher variance.

The key to the favorable bias property of the LP is that we control for a lag y_{t-1} of the data. Intuitively, by controlling for a lag of the data, we are effectively regressing y_{t+h} on the residualized "shock" $\tilde{x}_t = y_t - \text{proj}(y_t \mid y_{t-1})$. Since this residualized shock correlates only weakly with further lags of the data under the ARMA model (3.1) and with small α, the textbook omitted variable bias formula for regression implies that the LP bias will be negligible. Increasing the lag length further has no effect on the asymptotic bias, which remains zero; in contrast, an LP estimator that regresses y_{t+h} on y_t (i.e., without a lagged control) would be subject to bias of the same magnitude as its standard deviation, and so should be avoided.

- *Variance.* While LP dominates VAR in terms of bias, the large-sample variance $\tau^2_{h,\text{LP}}$ of LP necessarily exceeds the variance $\tau^2_{h,\text{VAR}}(p)$ of the VAR, and typically strictly so. Intuitively, since we assume the model to be rather "close" to a finite-order VAR (i.e., only mild misspecification), the large-sample variance of LPs and VARs can be shown to be the same as in a correctly specified VAR model. In that correctly specified model, however, the VAR estimator is efficient by virtue of being the quasi-maximum-likelihood estimator. The high variance of LP is easily spotted in practice: LP-estimated impulse response functions tend to look much more jagged or erratic (as a function of the horizon) than VAR-estimated impulse response functions. This jaggedness is the price to pay for low bias. Increasing the LP lag length has no further effect on the asymptotic variance, while increasing the VAR lag length increases variance, delivering small bias at the cost of impulse responses ultimately as jagged and volatile as those of LP.

It turns out that the greater sensitivity of the VAR estimator to dynamic specification is a general feature of the estimation approach, as we now discuss.

Multivariate generalization. Montiel Olea, Plagborg-Møller, Qian, and Wolf show that all the above-mentioned qualitative lessons from the ARMA$(1,1)$ model (3.1) go through in a much wider class of multivariate VARMA(p_0, q_0) models, where p_0 is some finite true autoregressive order, while the moving average lag length q_0 could be infinite. As before, this set-up gives the benefit of the doubt to VAR practitioners by modeling the moving average coefficients as "small." This is an empirically relevant class of models, for two reasons. First, the class is consistent with typical linearized structural macroeconomic models.[9] Second, it captures several kinds of empirically relevant types of dynamic misspecification: even if the true DGP were a finite-order VAR, if the econometrician accidentally omits some lags or relevant control variables from the empirical specification, this is tantamount to introducing moving average terms (Granger and Morris, 1976).

It turns out that, even in this much larger class of models, the large-sample distributions (3.2) continue to apply, and thus so does the bias-variance trade-off, as long as we consider LP and VAR estimators that control for at least $p \geq p_0$ lags of the data. If we fail to control for at least p_0 lags, either estimator could be badly biased in large samples. The way to interpret this result in practice is that it is important to control for those variables and lags that are *strongly* predictive of *either* the outcome variable of interest *or* the impulse variable (in the present theoretical setting, this amounts to controlling for p_0 lags of the data vector, since the moving average coefficients are small relative to the autoregressive coefficients). But once we control for these strong predictors, the LP estimator has zero asymptotic bias and constant variance as a function of p, even though further lags of the data do have some modest remaining predictive power. In other words, LP is relatively insensitive to the omission of moderately important control variables and lags (modeled through the moving average terms). By contrast, the VAR estimator generally has nonzero bias even once we control for the most important predictors, though it has lower variance than LP.

Lesson 5. *In finite samples, there is a stark bias-variance trade-off between LPs and VARs with small-to-moderate lag lengths. VARs extrapolate, and so have low variance but potentially high bias. LPs instead do not extrapolate and so have low bias and high variance.*

There is a simple way to "bias-correct" the VAR estimator: use a very large estimation lag length. In particular, in the present model set-up, it can be shown that the large-sample bias $b_h(p)$ of the VAR estimator at horizon h is zero when the estimation lag length p exceeds $p_0 + h$. However, when the lag length is chosen this large, then the variance is necessarily inflated to that of LP—in other words, the VAR bias is zero *only because* long-lag VAR and LP estimators are equivalent in large samples. In practice, we of course do not know exactly how large the true autoregressive lag length p_0 is, so to be certain that we included enough lags to eliminate meaningful biases, we would need to compare the VAR results with comparable LP results and verify that they are close to each other.

[9] Linearized DSGE models in macroeconomics almost always have a VARMA representation, though they typically do not have an exact finite-order VAR representation.

Is the VAR bias concerning? As the final step in our theoretical analysis we show that even relatively minor amounts of dynamic misspecification can yield economically large VAR biases. This discussion will rationalize much of what we find in our later simulations.

Our approach here is to ask how large the VAR bias $b_h(p)$ can possibly be, as a function of the magnitude of the dynamic misspecification. Montiel Olea, Plagborg-Møller, Qian, and Wolf (2024) show that the ratio of the absolute bias to the standard deviation of the VAR estimator is (asymptotically) bounded by

$$\frac{|b_h(p)|}{\tau_{h,\mathrm{VAR}}(p)} \leq \sqrt{T \times \mathcal{M}} \times \sqrt{\frac{\tau_{h,\mathrm{LP}}^2}{\tau_{h,\mathrm{VAR}}^2(p)} - 1}, \tag{3.3}$$

where T is the sample size and $\mathcal{M}$ is a measure of the magnitude of the misspecification of the VAR model, namely the fraction of the variance of the moving average residual that is explained by the lagged shocks in the VARMA model. In a well-specified finite-order VAR model there are no lagged shocks that enter into the residual, and so in this case we have $\mathcal{M} = 0$. In the simple ARMA(1,1) model (3.1), this residual variance fraction is approximately equal to α^2. For a concrete example, suppose that $\mathcal{M} = 0.01$, so that lagged shocks account for a mere 1% of the variance of the residual, and furthermore assume that $T = 100$ and the VAR standard error is half as large as the LP standard error, $\tau_{h,\mathrm{VAR}}(p)/\tau_{h,\mathrm{LP}} = 0.5$, as frequently encountered in applications (recall Figure 3.1). Then the upper bound (3.3) equals $\sqrt{3} \approx 1.73$, so that the bias of the VAR estimator can be nearly twice as large as its standard error, despite the minor degree of misspecification in this example. Biases of this magnitude are obviously worrisome when drawing statistical inferences from conventional VAR regression output. Though the formula (3.3) represents an upper bound on the extent of bias, Montiel Olea, Plagborg-Møller, Qian, and Wolf show that there always exists a residual moving average process (with small coefficients) that achieves this bound. Moreover, the particular form of this "least favorable" residual process does not appear to be unreasonable *ex ante* based on economic theory, and conventional statistical tests have low power to detect it *ex post* in the data.

To summarize, the worst-case bias (3.3) implies that the *only* way for VAR practitioners to *guarantee* that the bias is negligible is to include so many lags p in the specification that the standard error becomes equal to that of LP. Conversely, for any VAR specification that delivers material efficiency gains relative to LP, there is reason to worry that the VAR bias could be large relative to the standard error.

Lesson 6. *There is* no free lunch *for VAR practitioners: whenever VARs have small standard* *errors relative to LP, there is good reason to worry about large biases. Conventional model selection procedures or specification tests do not adequately guard against this risk. The only way to ensure that VAR impulse responses have low bias is to control for so many lags that they become equivalent with LP.*

The absence of a free lunch has particular bite at long horizons. If the DGP is stationary, impulse responses must be close to zero at very long horizons, but it is rare to possess precise prior knowledge about exactly at what rate the impulse responses decay. Unless the lag length is very large, the VAR estimator simply extrapolates the long-run responses based on the short-run empirical autocovariances, eventually enforcing an exponential rate of decay of the impulse responses in stationary environments. The VAR-estimated impulse responses at long horizons will therefore tend to have very small standard errors relative to LP (which, as we have discussed, does not enforce exponential decay). However, the potential for VAR bias is then high: if the estimated rate of decay is inaccurate, this can dramatically affect the magnitude of estimated long-horizon responses, and their bias can be several times larger than their standard errors, as indicated by the bias bound (3.3). In turn, this can lead to large errors in estimating such key features as the half-life or quarter-life of the impulse response function, or its cumulative value. LPs remain approximately unbiased at long horizons, and their large standard errors accurately reflect the fundamental issue that any finite data set only has limited information about what happens in the long run, in the absence of prior information. In particular, LPs simply do not allow the estimation of impulse responses at ultra-long horizons h beyond the observed sampled size T (or rather, the effective sample size $T - p$); VARs instead produce such estimates by pure extrapolation, with potential for severe bias. The fact that an LP at horizon h uses only $T - p - h$ data points (as opposed to $T - p$ for a VAR) is not a drawback of the procedure; it is a necessary consequence of the desire to avoid extrapolation.

3.4 Outlook

In Sections 2 and 3 we have made broad conceptual points on identification as well as on the basic bias-variance trade-off that is at the heart of the choice between LPs and VARs. In the remainder of the paper we give recommendations on how to exactly navigate the bias-variance trade-off in empirical practice. Guided by our earlier discussion of the bias-variance trade-off, we first discuss specification choices and implementation details for LP and VAR estimators in Section 4. Section 5 then provides a quantitative, empirically calibrated assessment of the bias-variance properties of these estimators, while Section 6 discusses implications for inference. Section 7 summarizes our recommendations for applied practice.

4 LP and VAR specification choices

We discuss two of the main practical considerations involved in estimating impulse responses through either LPs or VARs: selection of control variables and lags, and bias correction.

4.1 Control variables and lag length

In applications of LP or VAR estimators, a ubiquitous question is which variables and how many lags we should control for. There are four main factors to consider.

1. The controls should ensure valid identification of an interpretable shock. As discussed in Section 2, both LPs and VARs ultimately project the outcomes of interest on a measure of a shock, which by the Frisch-Waugh-Lowell theorem is simply given by the residual after regressing the impulse variable on the full list of controls. That list of controls, and the number of lags of these, should be sufficiently rich such that the residualized shock is unpredictable from other external variables or further lags of the data. In applications of recursive/Cholesky identification, the residualized shock often can be directly interpreted as the residual in a policy rule (e.g., a Taylor rule residual), such that economic theory can be brought to bear on what to control for (e.g., variables in the central bank's reaction function). If the impulse variable is a credibly unpredictable "shock", for example because it is obtained from high-frequency asset price movements over short time windows around policy announcements, then control variables are not needed to ensure correct identification. However, even in this case it is important to include controls for the reasons outlined next.

2. The second purpose of control variables is to increase efficiency. As in a randomized control trial, even when control variables are not correlated with the treatment (here the shock), they still soak up residual variation in the outcome, which typically lowers standard errors. While this argument seems to suggest including a very large number of controls, to preserve degrees of freedom, it is advisable to use economic theory to select a more limited potential list of controls that are likely strong predictors of the outcome. The final set of controls and lag length can be selected based on conventional

information criteria or model specification tests. In the simulations below we use the Akaike Information Criterion (AIC) for this purpose.

The remaining two factors are specific to LP estimation.

3. The third reason for including controls is that it helps robustify LP estimation, as we discussed in Section 3. If we have access to a credible shock measure, it is tempting to simply project the outcome on this shock without controls, but that is a mistake. The theory reviewed in Section 3.3 shows that once we control for variables that are strong predictors of *either* the outcome *or* the impulse variable (as we should be doing anyway for the two reasons outlined above), then LP estimation is relatively robust to dynamic misspecification, i.e., the omission of controls or lags with small-to-moderate predictive power. Hence, in addition to increasing efficiency, controls help inoculate against minor imperfections of the shock measure.

4. The fourth reason for using control variables is that it simplifies and robustifies the construction of LP confidence intervals. We will review the reasons for this in Section 6.

Practical takeaways. VAR practitioners tend to select control variables and lags with an eye towards their identifying assumptions as well as predictive power, consistent with the first two factors reviewed above. The precise lag length is then often selected using the AIC or by using conventional fixed lengths such as 4 for quarterly data and 12 for monthly data, consistent with the recommendations of Kilian and Lütkepohl (2017). We will review the performance of these VAR specification choices in our later simulations.

For LPs, there is relatively little consensus on how to select controls and lags, and so we suggest the following procedure, guided by the above considerations. In general, researchers should always control for (i) variables that are central to their identification scheme, (ii) the outcome and impulse variables, and (iii) any additional variables that strongly predict *either* the outcome *or* impulse variables (or both), as suggested by past experience or economic theory. Among this set, the precise choice of lags and controls can be guided by the AIC or other selection criteria as follows: first, estimate an auxiliary VAR that includes the outcome and impulse variables and other potential controls; second, use the AIC to select the set of controls and lag length as in conventional VAR practice (e.g., as discussed in Kilian and Lütkepohl, 2017, Section 2.6); third, use the selected variables and lags as controls in subsequent LPs.[10] Note that the role of the auxiliary VAR is merely to identify key predictors and lags; we do not use the auxiliary VAR to compute impulse responses because it lacks the robustness properties of LP (unless we control for many more lags than standard model selection criteria would select).

While some LP practitioners run regressions with long differences $y_{t+h} - y_{t-1}$ of the outcome variable on the left-hand side, this is actually redundant when using lagged controls. By standard OLS algebra, this outcome transformation has no impact whatsoever on the estimated LP coefficient once we control for at least one lag of the outcome, which we have argued is advisable.

4.2 Bias correction

The theoretical analysis reviewed in Section 3.3 focused on one key source of bias: dynamic misspecification. In the small data sets typical in applied macroeconomics, however, LPs and VARs are also subject to another source of bias: persistence of the data. In particular, impulse response estimates are typically biased towards displaying less persistent effects than the actual truth. For VARs, Kilian and Lütkepohl (2017) recommend applying the Pope (1990) bias correction to the estimated VAR coefficients to partially remove this source of bias. For LPs, Herbst and Johannsen (2024) find that a qualitatively similar bias is present and propose a simple bias correction. Though Li, Plagborg-Møller, and Wolf (2024) find that the LP bias correction increases the variance of the estimator without entirely removing bias, it is advisable to apply the correction since the justification for using LP over VAR prioritizes bias over variance anyway. We demonstrate the advantages of the bias correction in practice in simulations below.

It is important to remember that neither the Pope (1990) correction for VARs nor the Herbst and Johannsen (2024) correction for LPs deal with the bias that is due to dynamic misspecification. Hence, all our earlier points about the trade-off between LPs and VARs continue to apply to the bias-corrected estimators.

[10] Alternatively, we could do two separate model selection exercises for a single-equation one-step-ahead forecast of the outcome and a single-equation one-step-ahead forecast of the impulse variable. We should then use the *union* of the two sets of selected control variables and lags in subsequent LPs.

5 The bias-variance trade-off in practice

This section presents empirically calibrated simulation evidence on the quantitative nature of the bias-variance trade-off between various implementations of LPs and VARs. The simulation evidence here complements our earlier theoretical treatment, revealing what shape the bias-variance trade-off is likely to take in practice, and so laying the groundwork for our later practical recommendations.

5.1 The simulation experiment

Following Li, Plagborg-Møller, and Wolf (2024), we consider a large menu of possible data-generating processes (DGPs) estimated to mimic as closely as possible the properties of the universe of U.S. macroeconomic data. We will only provide a high-level overview of the DGPs here, with implementation details relegated to the original work of Li, Plagborg-Møller, and Wolf and to Supplemental Appendix E.

Data generating processes. Our simulations are based on a dynamic factor model (DFM) fitted to a large number of U.S. macroeconomic time series. We use the data set of Stock and Watson (2016), which consists of 207 quarterly U.S. time series spanning a range of variable categories, from quantities to prices. We then fit two separate DFMs to this data set: a non-stationary variant that allows for cointegrating relationships among the latent factors, and a stationary one, with all data series pre-transformed to ensure stationarity. The resulting DFMs imply complex and varied dynamics for macroeconomic time series of the sort encountered in applied practice.

Given these DFMs, we proceed to construct a varied array of lower-dimensional DGPs by considering hundreds of different subsets of time series, with variable selection closely emulating applied practice. Specifically, each DGP consists of a set of five observable time series, selected at random from the DFM's series, and restricting attention to those variables that are most commonly used in applied practice. We then consider a researcher who observes data of those selected time series, and identifies monetary and fiscal policy shocks through either a recursive ordering of these observable time series, or by also additionally measuring the policy shock, as in our discussion of identification in Section 2. For the monetary policy DGPs we restrict the vector of observables to always contain the federal funds rate, while for fiscal policy DGPs we always include government spending; for recursive identification these two policy variables are ordered last and first, respectively. The researcher is then interested in the response of one of the other four variables to the identified shock. She estimates this response using several different estimation strategies, to be discussed in detail below. Given that the true DFM is known to us, we can compute estimator biases, variances, and mean squared errors (for this section) as well as confidence interval properties (for the uncertainty assessments in the next section).

We stress that the DFMs and the resulting DGPs that we construct do not admit finite-order VAR representations, yielding a non-trivial bias-variance trade-off between LPs and VARs. The latent nature of the macroeconomic factors and the idiosyncratic measurement errors induce moving average dynamics in the processes for the observed time series, consistent with our discussion of dynamic misspecification in Section 3.3.

Our overall objective is to provide an empirically grounded assessment of which estimators are likely to perform well on average in settings typically encountered in applied work, and thus can serve as attractive default procedures. While we cannot claim that our empirically calibrated DFMs capture all aspects of the real world, we contend that a minimum condition for an econometric procedure to be useful for applied work is that it should at least perform well on the kinds of DGPs that we consider. The empirically grounded DGPs will furthermore showcase the quantitative bite of the general theoretical considerations reviewed in Section 3.

Estimators. We consider several variants of the LP and VAR estimators introduced in Section 2—variants that differ in the selection of the lag length p, the choice of control variables w_t, and whether we apply the small-sample bias corrections discussed in Section 4. All estimators include an intercept. We only provide a relatively brief overview here, with further technical details in Supplemental Appendix D.

1. **LP.** The researcher estimates the LP (2.1) to implement her observed shock or recursive identifying assumptions. We consider several different LP variants.

    - *Lag length.* We either fix the lag length at conventional values, or estimate it using the Akaike Information Criterion (AIC) following the procedure that was outlined in Section 4.1.
    - *Control variables.* The researcher observes a list of macroeconomic aggregates. Under recursive shock identification, all of those observables need to be included as controls. Under observed shock identification, we will consider two specifications: one that controls for lags of all of the observed variables, and one that controls for only lags of the measured shock and the outcome of interest.

- *Bias correction.* We implement the bias correction of Herbst and Johannsen (2024).

2. **VAR.** The researcher estimates the VAR (2.3) to implement her identifying assumptions. We also consider several different VAR variants.

   - *Lag length.* The lag length is either fixed at conventional values or estimated using the AIC for the reduced-form VAR, consistent with our discussion in Section 4.1.
   - *Control variables.* Exactly as for LP, we include the full set of observables as controls in the case of recursive identification, while for observed shock identification we include either the full set or just estimate a bivariate system in shock and outcome.
   - *Bias correction.* We estimate VARs with the Pope (1990) analytical correction for biases caused by persistent data, consistent with the recommendation of Kilian and Lütkepohl (2017), and as discussed in Section 4.2.

5.2 Results

In this section we present simulation results for 200 stationary and 200 non-stationary DGPs, for observed shock identification and averaging across monetary and fiscal shocks, with 100 DGPs for each. Results separately by the type of shock and for recursive identification are broadly similar, and relegated to Supplemental Appendices F.2 and F.3. We approximate population biases and variances by averaging across 1,000 Monte Carlo simulations per DGP.

Bias-variance trade-off. Figure 5.1 shows that, in practice, there indeed is a stark bias-variance trade-off between LPs and VARs, consistent with our earlier theoretical discussion. The left panels of the figure show the median absolute bias (across monetary and fiscal shock DGPs) for the stationary (top) and non-stationary (bottom) encompassing DFM, while the right panels show the standard deviation, in both cases normalized by the overall scale of the true impulse response. Red lines correspond to VAR estimators, and blue lines indicate LP estimators; as outlined earlier, the various line styles correspond to several different options for selection of lag length and controls. While all LP estimators have relatively low bias across horizons, all VAR estimators suffer from substantial bias at intermediate horizons, as well as at long horizons in persistent DGPs. Quantitatively, we see that the bias reduction afforded by LP estimators comes at relatively steep variance cost; we will

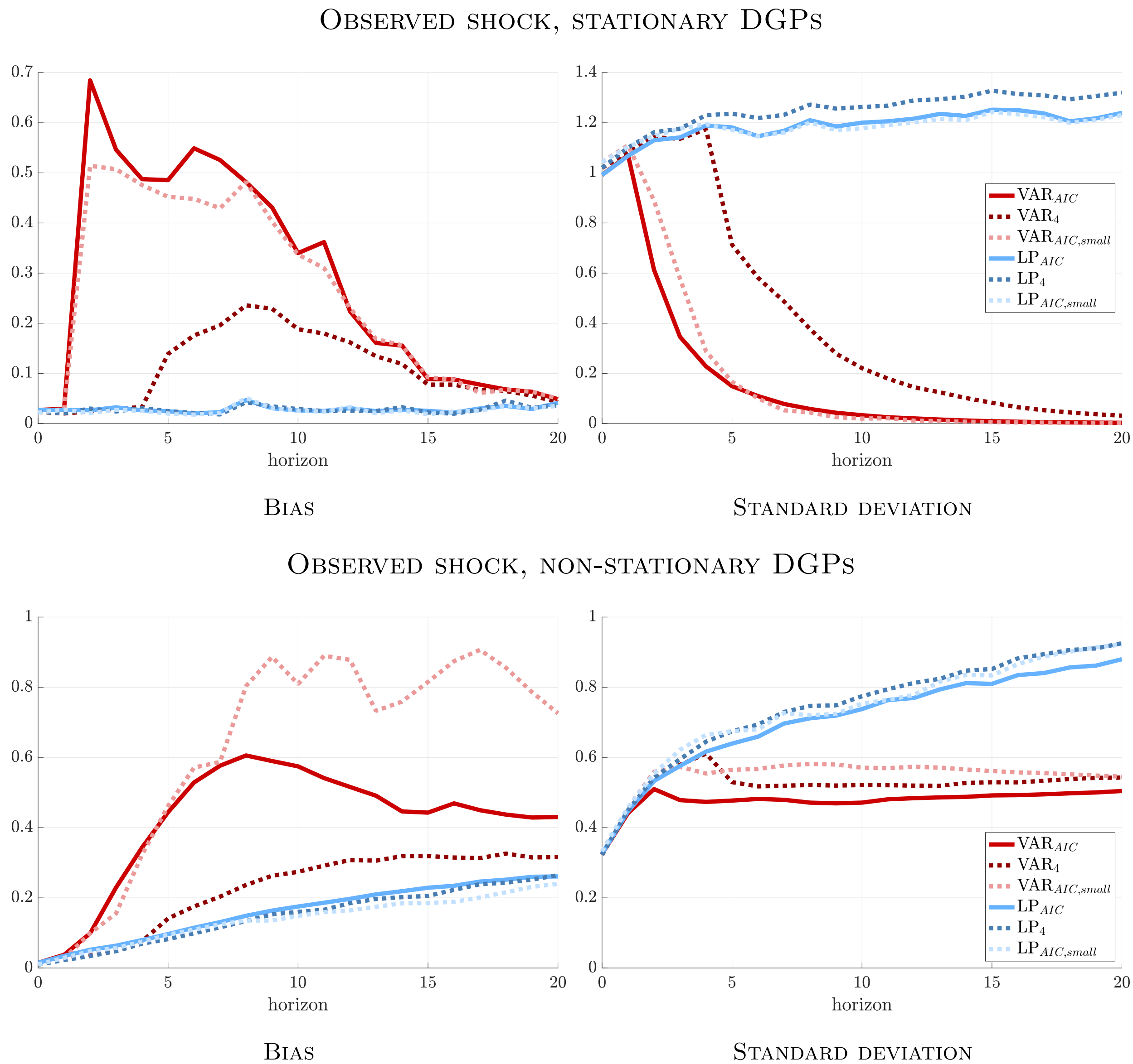


Figure 5.1: Median (across DGPs) of absolute bias $|\mathbb{E}[\hat{\theta}_h - \theta_h]|$ (left panels) and standard deviation $\sqrt{\mathrm{Var}(\hat{\theta}_h)}$ (right panels) of the different estimation procedures, relative to $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$. The subscript "AIC" indicates lag length selection via the AIC, "4" indicates four lags, and "small" indicates a small system, containing only shock and outcome of interest.

elaborate on this observation later, when discussing implications for mean squared error.[11]

[11] We only report results for bias-corrected LPs and VARs, consistent with our discussion in Section 4.2. Bias correction barely matters in the stationary DGPs, but LP biases would be materially larger in non-stationary DGPs in the absence of bias correction, particularly at medium and long horizons.

While LPs are relatively insensitive to dynamic specification, the bias-variance properties of VAR estimators are shaped decisively by lag length selection and the horizon of interest, and somewhat less so by the choice of controls. The AIC typically selects a very short lag length in our DGPs (the mean lag length selected equals 1.88), and so we see a sharp bias-variance trade-off already at short impulse horizons. Manually increasing the VAR lag length to 4 (dark red, dashed) aligns LPs and VARs up to horizon $h = 4$, again consistent with the theory reviewed earlier. The performance of LPs is, on the other hand, virtually unaffected by the precise number of lags included (solid blue vs. dashed dark blue). Furthermore, since in the stationary DGPs all impulse responses converge to zero relatively fast, both VAR bias and standard deviation are small at long horizons, dominating LP. In practice, however, the speed of this convergence is uncertain; in particular, in the non-stationary DGPs (in the bottom panel), the bias-variance trade-off remains stark even at long horizons. Finally, comparing the solid and light-dashed lines, we see that the effect of including either few or many controls is negligible for LP but moderately important for VAR. This last conclusion depends on the identification scheme, however: controls play a key role for recursive identification using either estimation approach, as shown in Supplemental Appendix F.2.

Mean squared error. We now further assess the quantitative nature of the bias-variance trade-off using the familiar mean squared error (MSE) criterion. For a given estimator $\hat{\theta}_h$ of the true impulse response θ_h, the MSE is defined as

$$\text{MSE}(\hat{\theta}_h) = \underbrace{\mathbb{E}\left[\hat{\theta}_h - \theta_h\right]^2}_{\text{Bias}(\hat{\theta})^2} + \text{Var}(\hat{\theta}_h).$$

In words, MSE weights equally an estimator's (squared) bias and its variance. Comparing the magnitudes on the y-axes of the top and bottom panels of Figure 5.1, we see that the bias reduction of LP comes at more than a one-to-one cost in terms of the standard deviation; for completeness, we report the implied MSE plots in Supplemental Appendix F.1. We can conclude that, at least when evaluated through the lens of the familiar MSE criterion for point estimation, the bias reduction of LP appears exceedingly costly.[12]

[12] Leveraging the worst-case bias result (3.3) in Section 3.3, it is straightforward to establish that even a minor amount of misspecification $\mathcal{M} \geq 1/T$ could in principle suffice for VAR MSE to exceed LP MSE (see Montiel Olea, Plagborg-Møller, Qian, and Wolf, 2024, Section 4.1). In the DGPs that we consider, the magnitude of misspecification is indeed quite material (e.g., for the stationary DFM, the average $\sqrt{T \times \mathcal{M}}$ across DGPs is 2.64 for $p = 4$ VAR lags), yet the VAR MSE is typically smaller, suggesting that the *type* of misspecification is not close to the worst case on average. We thank Isaiah Andrews for raising this point.

Lesson 7. *Empirically calibrated simulations reveal that, in practice, there is a sharp bias-variance trade-off between LPs and VARs at intermediate and long horizons. LPs attain their low bias at significant variance cost. If the researcher objective is to minimize mean squared error, then she will typically be unwilling to incur that variance cost.*

Of course, some researchers might be much more concerned with minimizing bias over minimizing variance, in which case LPs are preferrable over VARs. In the next section we will argue that the construction of valid confidence intervals *endogenously* forces researchers to heavily prioritize bias in this way.

6 Uncertainty assessments

Applied macroeconomists not only report point estimates of dynamic causal effects, they also want to quantify the statistical uncertainty of those estimates. This is an important task, since the standard errors of impulse responses are often of roughly the same magnitude as the estimates themselves. The accuracy of uncertainty assessments is conventionally evaluated by the *coverage* of the implied confidence interval: the probability that the reported interval covers the true impulse response should be at least 90% (say), at every horizon and regardless of the shape of the true impulse response function.

This section argues that LPs—or VARs with very long lag lengths—are the only known procedures that can robustly achieve satisfactory coverage of confidence intervals in practice. The intuition is that even quite small amounts of bias (relative to the standard error) can cause severe coverage distortions, and as we have seen, only LPs (or equivalently VARs with very long lags) robustly achieve low bias across a range of empirically relevant DGPs. Hence, though the previous section showed that VAR estimators can be attractive for estimation purposes to researchers who weight bias and variance approximately equally (such as under the MSE criterion), we are forced to heavily prioritize bias if we want accurate uncertainty assessments in the conventional sense of controlling the confidence interval coverage.

6.1 Confidence intervals for impulse responses

We begin by reviewing methods for computing standard errors for LP estimation of impulse responses. We will not review inference methods for VAR impulse responses, since these are already covered in textbooks such as Kilian and Lütkepohl (2017).

For LP impulse responses, heteroskedasticity-robust standard errors suffice for accurate uncertainty assessments. In the LP regression (2.1), the residual $\xi_{h,t}$ is typically serially correlated because it is a multi-step forecast error, which would suggest the use of Heteroskedasticity and Autocorrelation Consistent (HAC) standard errors, such as Newey-West. Fortunately, Montiel Olea and Plagborg-Møller (2021) show that HAC corrections are unnecessary under a weak assumption on the shocks, and conventional heteroskedasticity-robust standard errors for OLS suffice.[13] The reason is that, while the LP regression residual $\xi_{h,t}$ is indeed serially correlated, what matters for the distribution of the LP estimator is the *product* of the residual and the residualized shock $\tilde{x}_t$ defined in (2.2). This product is serially uncorrelated (though typically heteroskedastic) under the natural assumption that the shocks are unpredictable (i.e., conditionally mean independent) from their past and future values. And even if this assumption is slightly violated, it is still likely that the well-documented practical challenges of HAC estimation will make conventional heteroskedasticity-robust standard errors a better choice for applied work.[14]

Given a standard error $\hat{\tau}_h^{\text{LP}}$ for the LP impulse response estimate $\hat{\theta}_h^{\text{LP}}$, a level-$(1-a)$ confidence interval can be obtained with the usual formula:

$$\hat{\theta}_h^{\text{LP}} \pm \hat{\tau}_h^{\text{LP}} z_{1-a/2},$$

where $z_{1-a/2}$ is the $(1-a/2)$ quantile of the standard normal distribution (e.g., $z_{1-a/2} \approx 1.64$ for a $1-a = 90\%$ confidence interval). Montiel Olea and Plagborg-Møller (2021) propose an alternative bootstrap algorithm for computing an LP confidence interval that is asymptotically equivalent to the heteroskedasticity-robust confidence interval for OLS, but which slightly improves coverage in finite samples.[15]

Lesson 8. *When computing standard errors and associated confidence intervals for LP impulse responses, heteroskedasticity-robust standard errors typically perform as well or better than more complicated HAC corrections. Simulation evidence suggests that further improvements in the finite-sample performance can be obtained by employing the bootstrap algorithm proposed by Montiel Olea and Plagborg-Møller (2021).*

The next sections compare—first theoretically, then through simulations—the quality of

[13] This is particularly attractive because HAC standard errors are known to often understate the true uncertainty in sample sizes typical in applied macroeconomics, see for example Lazarus, Lewis, Stock, and Watson (2018) and Herbst and Johannsen (2024).

[14] Xu (2023) proposes a standard error formula that is valid more generally and can be used for robustness.

[15] See Velez (2024) for results on the asymptotic validity of related bootstrap procedures.

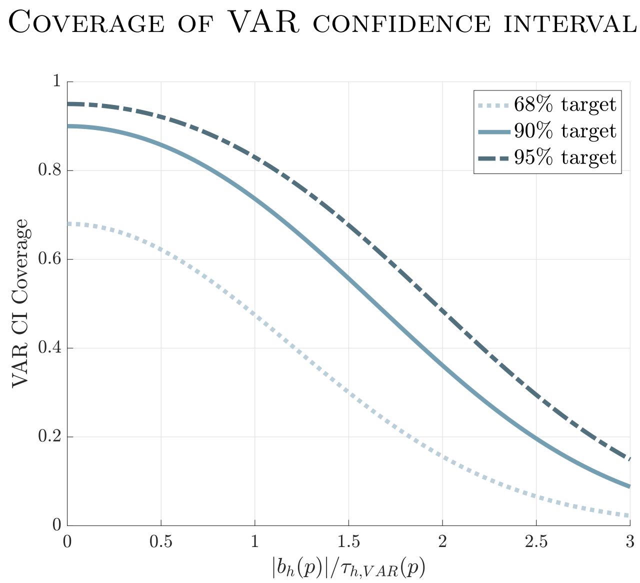


Figure 6.1: VAR coverage as a function of relative bias $|b_h(p)|/\tau_{h,\text{VAR}}(p)$, for target coverage levels of 68 per cent (dotted), 90 per cent (solid), and 95 per cent (dashed-dotted).

uncertainty assessments based on either LP or VAR inference.

6.2 Bias is very costly for coverage

In this section we establish theoretically that conventional VAR confidence intervals can exhibit severe coverage distortions even under small amounts of dynamic misspecification, while LP intervals are instead much more robust. Our analysis here builds on and extends our review of the bias-variance trade-off in Section 3.3, in particular the approximate distribution of the estimators stated in equation (3.2).

How bias affects coverage. The approximate VAR sampling distribution in (3.2) together with a straightforward calculation reveals that the coverage probability of the conventional VAR confidence interval is a decreasing function of the bias/standard-error ratio $|b_h(p)|/\tau_{h,\text{VAR}}(p)$; specifically, it equals

$$P(|Z| \leq z_{1-a/2}), \quad \text{where } Z \sim N(|b_h(p)|/\tau_{h,\text{VAR}}(p), 1). \tag{6.1}$$

Figure 6.1 plots the coverage probability (6.1), showing that even a quite moderate ratio of bias to standard error yields large coverage distortions for the VAR confidence interval. In Section 3.3 we gave a numerical example in which a small amount of VAR misspecification caused the bias to be around 1.73 times the standard error; this ratio would cause a putative 90% confidence interval to cover the true impulse responses with probability less than 50%! Panel (b) of Figure 3.1 indeed suggests that the bias-to-standard-error ratio of VAR exceeds 2 at long horizons in many applications.[16] Figure 6.1 thus implies that a researcher who is interested in guaranteeing that the coverage of a reported confidence interval is not too low relative to the target coverage must necessarily prioritize bias over variance.

MSE VS. COVERAGE. The implicit bias-variance preferences that emerge from coverage considerations are not only qualitatively different from the explicit bias-variance preferences based on MSE—they also differ very meaningfully *quantitatively*. Based on the approximate distributions (3.2) of LP and VAR, a simple calculation shows that VAR is preferred to LP on MSE grounds if and only if

$$|b_h(p)|/\tau_{h,\mathrm{VAR}}(p) \le \sqrt{\tau^2_{h,\mathrm{LP}}/\tau^2_{h,\mathrm{VAR}}(p) - 1}.$$

Once again making reference to panel b) of Figure 3.1, and focusing on long horizons, we may expect the left-hand side of the above inequality to be close to 2 in many applications. Panel a) of the same figure shows that the median value of $\tau_{h,\mathrm{LP}}/\tau_{h,\mathrm{VAR}}(p)$ can be close to 0.4 in practice, which corresponds to a value of $\sqrt{5.25} \approx 2.29$ for the right-hand side of the above inequality. Thus, consistent with the simulations in Section 5, VAR estimators may be preferable from the MSE perspective in "typical" applications, even though their bias could cause severe coverage distortions in these same applications, as argued above.

But because bias is so costly for coverage, the results of Sections 3 and 5 imply that LP confidence intervals—or equivalently VAR confidence intervals with a very large estimation lag length—are preferred over short-lag VARs when it comes to uncertainty assessments. Recall that, in the face of minor amounts of misspecification, we can only guarantee a low bias/standard-error ratio for VARs when we use very many lags (usually many more lags than indicated by conventional model selection or evaluation procedures). In practice, it thus only appears to be reasonable to trust VAR confidence intervals if they are approximately as wide as corresponding LP intervals.

Lesson 9. *The coverage of conventional VAR confidence intervals is highly sensitive to*

[16]The scaled difference $(\hat{\delta}_h^{\mathrm{LP}} - \hat{\delta}_h^{\mathrm{VAR}})/\tau_{h,\mathrm{VAR}}(p)$ of LP and VAR estimates is an unbiased estimate of $b_h(p)/\tau_{h,\mathrm{VAR}}(p)$. An important caveat is that Figure 3.1 reports the absolute value of the scaled difference, which is an overestimate of the absolute scaled bias.

dynamic misspecification, unlike LP intervals. To robustify VAR intervals, one must control for so many lags that the intervals become approximately the same as the LP intervals.

We furthermore stress that one does not need to believe that the (short-lag) VAR bias is large in *most* applications in order to prefer LP over VAR confidence intervals. In conventional statistical and econometric practice, we seek to control coverage *uniformly* across a wide range of empirically relevant DGPs, and not just for the "typical" DGP. In other words, our uncertainty assessment should reliably indicate that uncertainty is high *whenever* that is in fact the case. Hence, the mere fact that VARs can *sometimes* be badly biased, as shown theoretically in Section 3.3 and then practically in Section 5, would militate against short-lag VAR confidence intervals.

6.3 Challenges at long horizons and with persistent data

While the previous section discussed the fragility of VAR inference under mild misspecification, LP confidence intervals can also have more accurate coverage than conventional VAR confidence intervals *even if the VAR model is correctly specified.* We review the intuition here and refer to Montiel Olea and Plagborg-Møller (2021) for technical details.

Montiel Olea and Plagborg-Møller show that, assuming a correctly specified VAR model, LP confidence intervals control coverage more robustly at long horizons and across a wider range of persistence of the underlying DGP than VAR intervals. Intuitively, at long horizons h, the VAR impulse response transformation is a highly nonlinear transformation of the estimated parameters (e.g., the exponential formula $\hat{\rho}^h$ in the AR(1) case). This spells trouble for traditional VAR inference procedures because they ultimately rely on a linearization argument for asymptotic validity. By contrast, LP relies on linear regressions. Moreover, in DGPs with (near-)unit roots, it is well known that autoregressive coefficient estimates can have non-normal distributions, which dramatically complicates the calculation of appropriate critical values for impulse responses. LP, however, is equivalent to a projection of the outcome on the residualized shock $\tilde{x}_t$ in (2.2), and the latter is robustly stationary, so that the LP coefficient will typically still have a normal distribution even if the data itself has stochastic trends. In particular, due to their robustness to high persistence, it is advisable to estimate LPs in levels rather than transforming the data to first differences.

Lesson 10. *LP confidence intervals are more robust than conventional VAR confidence intervals to the length of the impulse response horizon and the persistence of the data, even if the VAR model is correctly specified.*

6.4 Simulation evidence

We now complement the theoretical insights of the previous two sections with simulation results. Our simulations are based on the large menu of DGPs described in Section 5.

Confidence intervals. We report the coverage of confidence intervals constructed from several variants of LP and VAR estimators, as in Section 5. We set the target confidence level to be 90%.

1. **LP.** We consider two approaches to construct LP confidence intervals.
   - *Analytical.* This analytical confidence interval is constructed using conventional heteroskedasticity-robust standard errors for OLS, deliberately ignoring HAC corrections, as discussed in Section 6.1 and Montiel Olea and Plagborg-Møller (2021).[17]
   - *Bootstrap.* We implement the percentile-t bootstrap confidence interval recommended by Montiel Olea and Plagborg-Møller (2021). The bootstrap samples are generated from an auxiliary VAR model.[18]
2. **VAR.** We also consider two approaches to construct confidence intervals based on VAR estimators.
   - *Analytical.* We use textbook formulae to compute delta method standard errors and associated confidence intervals for the VAR estimators.[19]
   - *Bootstrap.* Following the recommendation of Inoue and Kilian (2020), we report the Efron bootstrap confidence interval, based on the same bootstrap procedure described above for LP.

Results. As in Section 5 we present simulation results for 200 stationary and 200 non-stationary DGPs, for observed shock identification and averaging across monetary and fiscal policy shocks, with 100 DGPs for each. Coverage results separately by type of shock and for recursive identification are similar, and relegated to Supplemental Appendices F.2 and F.3. Coverage is approximated by averaging over 1,000 Monte Carlo simulations per DGP.

[17]In our simulations, *a priori* knowledge of the true DGP justifies the use of *homoskedastic* standard errors, though this will be adjusted in future versions of this draft.

[18]The current version of the paper draws the VAR residuals i.i.d., thus imposing homoskedasticity, but we will change this to a wild bootstrap in future versions of this draft.

[19]As discussed above, given our simulation designs, we currently impose homoskedasticity.

Observed shock: Confidence interval coverage

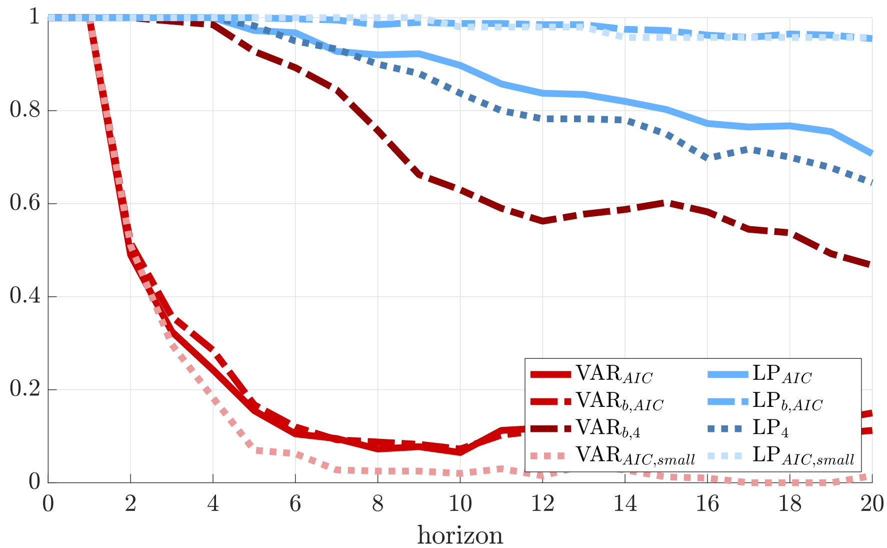


Figure 6.2: Fraction of DGPs (both stationary and non-stationary) for which the confidence interval coverage probability is above 80 per cent, by estimation procedure; the b subscript in the figure legend indicates bootstrap confidence intervals. The subscript "AIC" indicates lag length selection via the AIC, "4" indicates four lags, and "small" indicates a small system, containing only shock and outcome of interest.

As expected from the theoretical discussions in Sections 6.2 and 6.3, LPs tend to provide robust uncertainty assessments, while VARs do not. To establish this, Figure 6.2 shows the fraction of DGPs at which a given method delivers a coverage probability above 80% (and thus close to the target level of 90%), with VARs in red and LPs in blue. LP confidence intervals attain accurate coverage levels for a large fraction of the DGPs, especially the bootstrap version of the confidence interval. By contrast, for the majority of DGPs that we consider here, VARs with AIC-selected lag length have a coverage probability below 80% at all horizons $h > 2$. The coverage of longer-lag, bootstrapped VARs is somewhat better (yet again as expected theoretically), but we see that the coverage of even the best-performing VAR confidence interval is substantially worse than the LP bootstrap interval with lag length selected by AIC.

Lesson 11. *In empirically calibrated simulations, only conventional LPs or (equivalently) VARs with very large lag length deliver robust uncertainty assessments. The best performance is attained with bootstrapped standard errors.*

7 Summary of recommendations for applied practice

Based on the 11 lessons we have drawn from the available theory and simulation evidence on the econometric properties of LPs and VARs, we make the following summary recommendations for applied researchers who seek to perform inference on the dynamic causal effects of macroeconomic shocks to policies or fundamentals.

a) Neither LPs nor VARs solve any identification problems in and of themselves. Regardless of which method you use, you are projecting on something that is supposedly an economic shock. The first step to any empirical analysis is to be transparent about what this shock represents and to convince the reader that this interpretation is sensible.

b) The choice between LPs and VARs amounts to choosing between two different finite-sample estimators of the exact same large-sample estimand. LPs have low bias at the expense of high variance, while VARs (with small to moderate lag length) rely on extrapolation to produce low-variance impulse response estimates at the expense of potentially large bias. The choice of estimation method is likely to matter more the longer is the impulse response horizon of interest.

c) Uncertainty assessment is a central part of empirical analyses in macroeconomics, and there are only two known procedures that can provide confidence intervals with accurate coverage across a wide range of empirically relevant DGPs: conventional LPs, and VARs with very large lag length.

d) If using VARs, the lag length should be chosen so large that the confidence intervals approximately coincide with LP intervals. That is, an LP robustness check is an indispensable part of any VAR analysis. Do not rely on AIC or other conventional selection criteria to select the VAR lag length, and do not expect conventional model specification tests to guard sufficiently against the deleterious effects of dynamic misspecification.

e) We discourage the use of VAR specifications with the relatively short lag lengths (or strong shrinkage through a Minnesota prior) that have hitherto been common in applied work. There is nothing wrong with reporting the *point estimates* from such procedures, and indeed these estimates often have low MSE relative to LP. But since the associated confidence intervals have very fragile coverage, it is necessary to complement the results with either an LP analysis or a VAR analysis that uses a much longer lag length.

f) If using LPs, the choice of controls is as important as in VAR analysis. One should at least control for (i) variables that are central to the identification argument, (ii) several lags of the outcome and impulse variables, and (iii) other strong predictors of either of the latter two variables. See Section 4 for the specific procedure that we recommend for selecting the set of controls and the number of lags for LP.

g) The LP estimator should be bias-corrected using the procedure of Herbst and Johannsen (2024). Instead of reporting HAC standard errors for LP, simply report the conventional heteroskedasticity-robust standard errors—or even better, the bootstrap confidence interval proposed by Montiel Olea and Plagborg-Møller (2021).

8 Next steps

In future versions of this paper we will furthermore add the following points.

1. **Heteroskedasticity.** We currently report simulation results for inference methods that impose homoskedasticity, consistent with our underlying DGPs. To more closely emulate applied practice—and consistent with our general recommendations in Section 6—we will switch to heteroskedasticity-robust inference.

2. **Alternative estimation methods.** So far we only consider standard LP and VAR estimators (and some minor variants thereof). We will further enrich the bias-variance frontier by also considering penalized LPs and Bayesian VARs.

3. **Alternative notions of coverage.** In judging the quality of uncertainty assessments, we so far use the criterion of correct coverage horizon by horizon. We plan to study how the different estimation methods fare with alternative, weaker notions of coverage, e.g., requiring correct probability *on average* across horizons, or only for a smooth projection of the true impulse response function.

References

Almuzara, M., and V. Sancibrian (2024): "Micro responses to macro shocks," Staff Report 1090, Federal Reserve Bank of New York.

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

Braun, P. A., and S. Mittnik (1993): "Misspecifications in vector autoregressions and their effects on impulse responses and variance decompositions," *Journal of Econometrics*, 59(3), 319–341.

Caravello, T., and P. Martinez-Bruera (2024): "Disentangling Sign and Size Nonlinearities," Discussion Paper 4704050, SSRN.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (1999): "Monetary policy shocks: What have we learned and to what end?," in *Handbook of Macroeconomics*, ed. by J. Taylor, and M. Woodford, vol. 1, chap. 2, pp. 65–148. Elsevier.

Fernández-Villaverde, J., J. F. Rubio-Ramírez, T. J. Sargent, and M. W. Watson (2007): "ABCs (and Ds) of Understanding VARs," *American Economic Review*, 97(3), 1021–1026.

Gonçalves, S., A. M. Herrera, L. Kilian, and E. Pesavento (2024): "State-dependent local projections," *Journal of Econometrics*, 244(2), 105702.

Granger, C. W. J., and M. J. Morris (1976): "Time Series Modelling and Interpretation," *Journal of the Royal Statistical Society. Series A (General)*, 139(2), 246–257.

Herbst, E. P., and B. K. Johannsen (2024): "Bias in local projections," *Journal of Econometrics*, 240(1), 105655.

Inoue, A., and L. Kilian (2020): "The uniform validity of impulse response inference in autoregressions," *Journal of Econometrics*, 215(2), 450–472.

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

Jordà, O., and A. M. Taylor (2024): "Local Projections," Working Paper 32822, National Bureau of Economic Research.

Känzig, D. R. (2021): "The Macroeconomic Effects of Oil Supply News: Evidence from OPEC Announcements," *American Economic Review*, 111(4), 1092–1125.

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

Kolesár, M., and M. Plagborg-Møller (2024): "Dynamic Causal Effects in a Nonlinear World: the Good, the Bad, and the Ugly," *arXiv preprint arXiv:2411.10415*.

Lazarus, E., D. J. Lewis, J. H. Stock, and M. W. Watson (2018): "HAR Inference: Recommendations for Practice," *Journal of Business & Economic Statistics*, 36(4), 541–559.

Li, D., M. Plagborg-Møller, and C. K. Wolf (2024): "Local projections vs. VARs: Lessons from thousands of DGPs," *Journal of Econometrics*, 244(2), 105722, Themed Issue: Macroeconometrics.

Montiel Olea, J. L., and M. Plagborg-Møller (2019): "Simultaneous confidence bands: Theory, implementation, and an application to SVARs," *Journal of Applied Econometrics*, 34(1), 1–17.

——— (2021): "Local Projection Inference Is Simpler and More Robust Than You Think," *Econometrica*, 89(4), 1789–1823.

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

Plagborg-Møller, M., and C. K. Wolf (2021): "Local Projections and VARs Estimate the Same Impulse Responses," *Econometrica*, 89(2), 955–980.

——— (2022): "Instrumental Variable Identification of Dynamic Variance Decompositions," *Journal of Political Economy*, 130(8), 2164–2202.

Pope, A. L. (1990): "Biases of Estimators in Multivariate Non-Gaussian Autoregressions," *Journal of Time Series Analysis*, 11(3), 249–258.

Ramey, V. A. (2016): "Macroeconomic Shocks and Their Propagation," in *Handbook of Macroeconomics*, ed. by J. B. Taylor, and H. Uhlig, vol. 2, chap. 2, pp. 71–162. Elsevier.

SIMS, C. A. (1980): "Macroeconomics and Reality," *Econometrica*, 48(1), 1–48.

STOCK, J. H., AND M. W. WATSON (2016): "Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics," in *Handbook of Macroeconomics*, vol. 2, chap. 8, pp. 415–525. Elsevier.

STOCK, J. H., AND M. W. WATSON (2018): "Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments," *The Economic Journal*, 128(610), 917–948.

UHLIG, H. (2005): "What are the effects of monetary policy on output? Results from an agnostic identification procedure," *Journal of Monetary Economics*, 52(2), 381–419.

VELEZ, A. (2024): "The Local Projection Residual Bootstrap for AR(1) Models," arXiv:2309.01889.

XU, K.-L. (2023): "Local Projection Based Inference under General Conditions," Manuscript, Indiana University Bloomington.