Local Projection Inference is Simpler and More Robust Than You Think*

José Luis Montiel Olea
Columbia University

Mikkel Plagborg-Møller
Princeton University

First version: March 17, 2020
This version: December 4, 2020

Abstract: Applied macroeconomists often compute confidence intervals for impulse responses using local projections, i.e., direct linear regressions of future outcomes on current covariates. This paper proves that local projection inference robustly handles two issues that commonly arise in applications: highly persistent data and the estimation of impulse responses at long horizons. We consider local projections that control for lags of the variables in the regression. We show that lag-augmented local projections with normal critical values are asymptotically valid uniformly over (i) both stationary and non-stationary data, and also over (ii) a wide range of response horizons. Moreover, lag augmentation obviates the need to correct standard errors for serial correlation in the regression residuals. Hence, local projection inference is arguably both simpler than previously thought and more robust than standard autoregressive inference, whose validity is known to depend sensitively on the persistence of the data and on the length of the horizon.

Keywords: impulse response, local projection, long horizon, uniform inference.

*Email: jm4474@columbia.edu, mikkelpm@princeton.edu. We are grateful for comments from two anonymous referees, Jushan Bai, Otavio Bartalotti, Guillaume Chevillon, Max Dovi, Marco Del Negro, Domenico Giannone, Nikolay Gospodinov, Michael Jansson, Òscar Jordà, Lutz Kilian, Michal Kolesár, Simon Lee, Sophocles Mavroeidis, Konrad Menzel, Ulrich Müller, Serena Ng, Elena Pesavento, Mark Watson, Christian Wolf, Tao Zha, and numerous seminar participants. We would like to especially thank Atsushi Inoue and Anna Mikusheva for a very helpful discussion of our paper. Montiel Olea would like to thank Qifan Han and Giovanni Topa for excellent research assistance. Plagborg-Møller acknowledges that this material is based upon work supported by the NSF under Grant #1851665. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

1 Introduction

Impulse response functions are key objects of interest in empirical macroeconomic analysis. It is increasingly popular to estimate these parameters using the method of *local projections* (Jordà, 2005): simple linear regressions of a future outcome on current covariates (Ramey, 2016; Angrist et al., 2018; Nakamura and Steinsson, 2018; Stock and Watson, 2018). Since local projection estimators are regression coefficients, they have a simple and intuitive interpretation. Moreover, inference can be carried out using textbook standard error formulae, adjusting for serial correlation in the (multi-step forecast) regression residuals.

Despite its popularity, there exist no theoretical results justifying the use of local projection *inference* over autoregressive procedures. From an identification and estimation standpoint, Kilian and Lütkepohl (2017) and Plagborg-Møller and Wolf (2020) argue that neither local projections nor Vector Autoregressions (VARs) dominate the other in terms of mean squared error in finite samples, and in population the two methods are equivalent. However, from an inference perspective, the only available guidance on the relative performance of local projections comes in the form of a small number of simulation studies, which by necessity cannot cover the entire range of empirically relevant data generating processes.

In this paper we show that—in addition to its intuitive appeal—frequentist local projection inference is robust to two common features of macroeconomic applications: highly persistent data and the estimation of impulse responses at long horizons. Key to our result is that we consider *lag-augmented* local projections, which use lags of the variables in the regression as controls. Formally, we prove that standard confidence intervals based on such lag-augmented local projections have correct asymptotic coverage *uniformly* over the persistence in the data generating process and over a wide range of horizons.[1] This means that confidence intervals remain valid even if the data exhibits unit roots, and even at horizons h that are allowed to grow with the sample size T, e.g., $h = h_T \propto T^\eta$, $\eta \in [0, 1)$. In fact, when persistence is not an issue, and the data is known to be stationary, local projection inference is also valid at *long* horizons; i.e., horizons that are a non-negligible fraction of the sample size ($h_T \propto T$).

Lag-augmenting local projections not only robustifies inference, it also simplifies the computation of standard errors by obviating the adjustment for serial correlation in the residuals. It is common practice in the local projections literature to compute Heteroskedas-

[1] We focus on marginal inference on individual impulse responses, not *simultaneous* inference on a vector of several response horizons (Inoue and Kilian, 2016; Montiel Olea and Plagborg-Møller, 2019).

ticity and Autocorrelation Consistent/Robust (HAC/HAR) standard errors (Jordà, 2005; Ramey, 2016; Kilian and Lütkepohl, 2017; Stock and Watson, 2018). Instead, we prove that the usual Eicker-Huber-White heteroskedasticity-robust standard errors suffice for *lag-augmented* local projections. The reason is that, although the regression residuals are serially correlated, the *regression scores* (the product of the residuals and residualized regressor of interest) are serially uncorrelated under weak assumptions. This finding further simplifies local projection inference, as it side-steps the delicate choice of HAR procedure and associated difficult-to-interpret tuning parameters (e.g., Lazarus et al., 2018).

The robustness properties of lag-augmented local projection inference stand in contrast to the well-known fragility of standard autoregressive procedures. Textbook autoregressive inference methods for impulse responses (such as the delta method) are invalid in some cases with near-unit roots or medium-long to long horizons (e.g., $h_T \propto \sqrt{T}$), as discussed further below. We show that lag-augmented local projection inference is valid when the data has near-unit roots and the horizon sequence satisfies $h_T/T \to 0$. Though the method fails in the case of unit roots and very long horizons $h_T \propto T$, existing VAR-based methods that achieve correct coverage in this case are either highly computationally demanding or result in impractically wide confidence intervals. When the data is stationary and interest centers on short horizons, local projection inference is valid but less efficient than textbook AR inference. Thus, the robustness afforded by our recommended procedure is not a free lunch. We provide a detailed comparison with alternative inference procedures in Section 3 below.

Our results rely on assumptions that are similar to those used in the literature on autoregressive inference. In particular, we assume that the true model is a VAR(p) with possibly conditionally heteroskedastic innovations and known lag length. We discuss the choice of lag length p in Section 6. The key assumption that we require on the innovations is that they are conditionally mean independent of both past and *future* innovations (which is trivially satisfied for i.i.d. innovations). Our strengthening of the usual martingale difference assumption is crucial to avoid HAC inference, but we show that the assumption is satisfied for a large class of conditionally heteroskedastic innovation processes. The robustness property of local projection inference only obtains asymptotically if the researcher controls for all p lags of all of the variables in the VAR system. Thus, our paper highlights the advantages of multivariate modeling even when using single-equation local projections.

To illustrate our theoretical results, we present a small-scale simulation study suggesting that lag-augmented local projection confidence intervals achieve a favorable trade-off between coverage and length. Since local projection estimation is subject to small-sample biases just

like VAR estimation (Herbst and Johannsen, 2020), we consider a simple and computationally convenient bootstrap implementation of local projection. The simulations suggest that non-augmented autoregressive procedures with delta method standard errors have more severe under-coverage problems than local projection inference, especially at moderate and long horizons. Autoregressive confidence intervals can be meaningfully *shorter* than lag-augmented local projection intervals in *relative* terms, but in *absolute* terms the difference in length is surprisingly modest. Our simulations also indicate that lag-augmented local projections with heteroskedasticity-robust standard errors have better coverage/length properties than more standard *non-augmented* local projections with off-the-shelf HAR standard errors. Finally, although the lag-augmented autoregressive bootstrap procedure of Inoue and Kilian (2020) achieves good coverage, it yields prohibitively wide confidence intervals at longer horizons when the data is persistent.

Related Literature. It is well known that standard autoregressive (AR) inference on impulse responses requires an auxiliary rank condition to rule out super-consistent limit distributions, thus yielding a $\sqrt{T}$-normal limit with strictly positive variance, see Assumption B of Inoue and Kilian (2020). When this rank condition holds, the textbook AR impulse response estimator is asymptotically normal even in the presence of (near-)unit roots (Inoue and Kilian, 2002). However, there are two common features of the data that lead to violations of the rank condition. First, the condition can fail when some linear combinations of the variables exhibit no persistence (Benkwitz et al., 2000). Second, in the presence of (near-)unit roots, certain linear combinations of the autoregressive coefficients are necessarily super-consistent (Sims et al., 1990). This compromises textbook AR inference for certain combinations of impulse response horizons and parameter values that typically cannot be ruled out *a priori*, especially in AR(1) or VAR(1) models, but also in higher-order autoregressions (Phillips, 1998; Inoue and Kilian, 2020, Remark 3, p. 455). In an important paper, Inoue and Kilian (2020) show that *lag-augmented* autoregressive inference solves the rank problem caused by (near-)unit roots, but data generating processes that lack persistence still need to be ruled out *a priori*. We build on their ideas, which in turn are based on Toda and Yamamoto (1995) and Dolado and Lütkepohl (1996). As we show, the validity of lag-augmented *local projection* (LP) inference does not hinge on auxiliary rank conditions.

Moreover, the validity of textbook AR inference is also compromised when the length of the impulse response horizon is large (Pesavento and Rossi, 2007; Mikusheva, 2012). Standard bootstrap methods rectify some of these problems, but not all. Several papers have

proposed AR-based methods for impulse response inference at *long* horizons $h = h_T \propto T$ (Wright, 2000; Gospodinov, 2004; Pesavento and Rossi, 2007; Mikusheva, 2012; Inoue and Kilian, 2020). With the exception of Mikusheva (2012), the literature on long-horizon inference has exclusively focused on near-unit root processes as opposed to devising uniformly valid procedures. The Hansen (1999) grid bootstrap analyzed by Mikusheva (2012) is asymptotically valid at short and long horizons. However, it is not valid at intermediate horizons (e.g., $h_T \propto \sqrt{T}$), unlike the LP procedure we analyze. Mikusheva argues, though, that the grid bootstrap is *close* to being valid at intermediate horizons, although it is much more computationally demanding than our recommended procedure, especially in VAR models with several parameters. Inoue and Kilian (2020) show that a version of the Efron bootstrap confidence interval, when applied to lag-augmented AR estimators, is valid at long horizons. We show that this procedure delivers impractically wide confidence intervals at moderately long horizons when the data is persistent, unlike lag-augmented LP.

We appear to be the first to prove the *uniform* validity of lag-augmented LP inference. Mikusheva (2007, 2012) and Inoue and Kilian (2020) derive the uniform coverage properties of various AR inference procedures, but they do not consider LP. The *pointwise* properties of LP procedures have been discussed by Jordà (2005), Kilian and Lütkepohl (2017), and Stock and Watson (2018), among others. Kilian and Kim (2011) and Brugnolini (2018) present simulation studies comparing AR inference and LP inference. Brugnolini (2018) finds that the lag length in the LP matters, which is consistent with our theoretical results.

Though the theoretical results in this paper appear to be novel, Dufour et al. (2006, Section 5) and Breitung and Brüggemann (2019) have discussed some of the main ideas presented herein. First, both these papers state that lag augmentation in LP avoids unit root asymptotics, but neither paper considers inference at long horizons or derives uniform inference properties. Second, Breitung and Brüggemann (2019) further argue that HAC inference in LP can be avoided if the true model is a VAR(p), although it is not clear from their discussion what are the assumptions needed for this to be true. Neither of these papers provide results concerning the efficiency of lag-augmented LP inference relative to other lag-augmented or non-augmented inference procedures, as we do in Section 3.

Local projections are closely related to multi-step forecasts. Richardson and Stock (1989) and Valkanov (2003) develop a non-standard limit distribution theory for long-horizon forecasts. Chevillon (2017) proves a robustness property of direct multi-step inference that involves non-normal asymptotics due to the lack of lag augmentation. Phillips and Lee (2013) test the null hypothesis of no long-horizon predictability using a novel approach that

requires a choice of tuning parameters, but yields uniformly-over-persistence normal asymptotics. This test is based on an estimator with a faster convergence rate than ours in the non-stationary case. However, to the best of our knowledge, their approach does not carry over immediately to impulse response inference, and it is not obvious whether the procedure is uniformly valid over both short and long horizons.

Outline. Section 2 provides a non-technical overview of our results in the context of a simple AR(1) model, including an illustrative simulation study. Section 3 provides an in-depth comparison of lag-augmented LP with other inference procedures. Section 4 presents the formal uniformity result for a general VAR(p) model. Section 5 describes a simple bootstrap implementation of lag-augmented local projection that we recommend for practical use. Section 6 concludes. Proofs are relegated to Appendix A and the Online Supplement. Appendices B and C contain further simulation and theoretical results. The supplement and a full Matlab code repository are available online.[2]

2 Overview of the Results

This section provides an overview of our results in the context of a simple univariate AR(1) model. The discussion here merely intends to illustrate our main points. Section 4 presents general results for VAR(p) models.

2.1 Lag-Augmented Local Projection

Model. Consider the AR(1) model for the data $\{y_t\}$:

$$y_t = \rho y_{t-1} + u_t, \quad t = 1, 2, \ldots, T, \quad y_0 = 0. \tag{1}$$

The parameter of interest is a nonlinear transformation of ρ, namely the impulse response coefficient at horizon $h \in \mathbb{N}$. We denote this parameter by $\beta(\rho, h) \equiv \rho^h$. In Section 4 below we argue that the zero initial condition $y_0 = 0$ is not needed for our results to go through. Our main assumption in the univariate model is:

Assumption 1. *$\{u_t\}$ is strictly stationary and satisfies $E(u_t \mid \{u_s\}_{s \neq t}) = 0$ almost surely.*

[2] https://github.com/jm4474/Lag-augmented_LocalProjections

The assumption requires the innovations to be mean independent relative to past and future innovations. This is a slight strengthening of the usual martingale difference assumption on u_t. Assumption 1 is trivially satisfied if $\{u_t\}$ is i.i.d., but it also allows for stochastic volatility and GARCH-type innovation processes.[3]

Local Projections With and Without Lag Augmentation. We consider the local projection (LP) approach of Jordà (2005) for conducting inference about the impulse response $\beta(\rho, h)$. A common motivation for this approach is that the AR(1) model (1) implies

$$y_{t+h} = \beta(\rho, h) y_t + \xi_t(\rho, h), \tag{2}$$

where the regression residual (or *multi-step forecast error*),

$$\xi_t(\rho, h) \equiv \sum_{\ell=1}^{h} \rho^{h-\ell} u_{t+\ell},$$

is generally serially correlated, even if the innovation u_t is i.i.d.

The most straight-forward LP impulse response estimator simply regresses y_{t+h} on y_t, as suggested by equation (2), but the validity of this approach is sensitive to the persistence of the data. Specifically, this standard approach leads to a non-normal limiting distribution for the impulse response estimator when $\rho \approx 1$, since the regressor y_t exhibits near-unit-root behavior in this case. Hence, inference based on normal critical values will not be valid uniformly over all values of $\rho \in [-1, 1]$ even for fixed forecast horizons h. If ρ is safely within the stationary region, then the LP estimator is asymptotically normal, but inference generally requires the use of Heteroskedasticity and Autocorrelation Robust (HAR) standard errors to account for serial correlation in the residual $\xi_t(\rho, h)$.

To robustify and simplify inference, we will instead consider a *lag-augmented* local projection, which uses y_{t-1} as an additional control variable. In the autoregressive literature, "lag augmentation" refers to the practice of using more lags for estimation than suggested by the true autoregressive model. Define the covariate vector $x_t \equiv (y_t, y_{t-1})'$. Given any horizon $h \in \mathbb{N}$, the lag-augmented LP estimator $\hat{\beta}(h)$ of $\beta(\rho, h)$ is given by the coefficient on

[3]For example, consider processes $u_t = \tau_t \varepsilon_t$, where ε_t is i.i.d. with $E(\varepsilon_t) = 0$, and for which one of the following two sets of conditions hold: (a) $\{\tau_t\}$ and $\{\varepsilon_t\}$ are independent processes; or (b) τ_t is a function of lagged values of ε_t^2, and the distribution of ε_t is symmetric. Assumption 1 is in principle testable, but that is outside the scope of this paper.

y_t in a regression of y_{t+h} on y_t and y_{t-1}:

$$\begin{pmatrix} \hat{\beta}(h) \\ \hat{\gamma}(h) \end{pmatrix} \equiv \left(\sum_{t=1}^{T-h} x_t x_t' \right)^{-1} \sum_{t=1}^{T-h} x_t y_{t+h}. \tag{3}$$

Here $\hat{\beta}(h)$ is the impulse response estimator of interest, while $\hat{\gamma}(h)$ is a nuisance coefficient.

The purpose of the lag augmentation is to make the effective regressor of interest stationary even when the data y_t has a unit root. Note that equations (1)–(2) imply

$$y_{t+h} = \beta(\rho, h) u_t + \beta(\rho, h+1) y_{t-1} + \xi_t(\rho, h). \tag{4}$$

If u_t were observed, the above equation suggests regressing y_{t+h} on u_t, while controlling for y_{t-1}. Intuitively, this will lead to an asymptotically normal estimator of $\beta(\rho, h)$, since the regressor of interest u_t is stationary by Assumption 1, and we control for the term that involves the possibly non-stationary regressor y_{t-1}. Fortunately, due to the linear relationship $y_t = \rho y_{t-1} + u_t$, the coefficient $\hat{\beta}(h)$ on y_t in the feasible lag-augmented regression (3) on (y_t, y_{t-1}) precisely equals the coefficient on u_t in the desired regression on (u_t, y_{t-1}). This argument for why lag-augmented LP can be expected to have a uniformly normal limit distribution even when $\rho \approx 1$ is completely analogous to the reasoning for using lag augmentation in AR inference (Sims et al., 1990; Toda and Yamamoto, 1995; Dolado and Lütkepohl, 1996; Inoue and Kilian, 2002, 2020). In the LP case, lag augmentation has the additional benefit of simplifying the computation of standard errors, as we now discuss.

Standard Errors. We now define the standard errors for the lag-augmented LP estimator. We will show that, contrary to conventional wisdom (e.g., Jordà, 2005, p. 166; Ramey, 2016, p. 84), HAR standard errors are *not* needed to conduct inference on *lag-augmented* LP, despite the fact that the regression residual $\xi_t(\rho, h)$ is serially correlated. Instead, it suffices to use the usual heteroskedasticity-robust Eicker-Huber-White standard error of $\hat{\beta}(h)$:[4]

$$\hat{s}(h) \equiv \frac{(\sum_{t=1}^{T-h} \hat{\xi}_t(h)^2 \hat{u}_t(h)^2)^{1/2}}{\sum_{t=1}^{T-h} \hat{u}_t(h)^2}, \tag{5}$$

[4]This is computed by the `regress, robust` command in Stata, for example. The usual homoskedastic standard error formula suffices if u_t is assumed to be i.i.d.

where we define the lag-augmented LP residuals

$$\hat{\xi}_t(h) \equiv y_{t+h} - \hat{\beta}(h)y_t - \hat{\gamma}(h)y_{t-1}, \quad t = 1, 2, \ldots, T-h, \tag{6}$$

and the residualized regressor of interest

$$\hat{u}_t(h) \equiv y_t - \hat{\rho}(h)y_{t-1}, \quad t = 1, 2, \ldots, T-h,$$

$$\hat{\rho}(h) \equiv \frac{\sum_{t=1}^{T-h} y_t y_{t-1}}{\sum_{t=1}^{T-h} y_{t-1}^2}.$$

As mentioned in the introduction, the fact that we may avoid HAR inference simplifies the implementation of LP inference, as there is no need to choose amongst alternative HAR procedures or specify tuning parameters such as bandwidths (Lazarus et al., 2018).

Why is it not necessary to adjust for serial correlation in the residuals? Since lag-augmented LP controls for y_{t-1}, equation (4) suggests that the estimator $\hat{\beta}(h)$ is asymptotically equivalent with the coefficient in a linear regression of the (population) residualized outcome $y_{t+h} - \beta(\rho, h+1)y_{t-1}$ on the (population) residualized regressor $u_t = y_t - \rho y_{t-1}$:

$$\begin{aligned}\hat{\beta}(h) &\approx \frac{\sum_{t=1}^{T-h}\{y_{t+h} - \beta(\rho, h+1)y_{t-1}\}u_t}{\sum_{t=1}^{T-h} u_t^2} \\ &= \beta(\rho, h) + \frac{\sum_{t=1}^{T-h} \xi_t(\rho, h)u_t}{\sum_{t=1}^{T-h} u_t^2}.\end{aligned}$$

The second term in the decomposition above determines the sampling distribution of the lag-augmented local projection. Although the multi-step regression residual $\xi_t(\rho, h)$ is serially correlated on its own, the *regression score* $\xi_t(\rho, h)u_t$ is serially uncorrelated under Assumption 1.[5] For any $s < t$,

$$\begin{aligned}E[\xi_t(\rho, h)u_t\xi_s(\rho, h)u_s] &= E[E(\xi_t(\rho, h)u_t\xi_s(\rho, h)u_s \mid u_{s+1}, u_{s+2}, \ldots)] \\ &= E[\xi_t(\rho, h)u_t\xi_s(\rho, h)\underbrace{E(u_s \mid u_{s+1}, u_{s+2}, \ldots)}_{=0}].\end{aligned} \tag{7}$$

Thus, the heteroskedasticity-robust (but not autocorrelation-robust) standard error $\hat{s}(h)$ suffices for doing inference on $\hat{\beta}(h)$.[6] Notice that this result crucially relies on (i) lag-

[5] Breitung and Brüggemann (2019) make this same observation, but they appear to claim that it is sufficient to assume that $\{u_t\}$ is white noise, which is incorrect.

[6] Stock and Watson (2018, p. 152) mention a similar conclusion for the distinct case of LP with an

augmenting the local projections and (ii) the strengthening in Assumption 1 of the usual martingale difference assumption on $\{u_t\}$ (as remarked above, the strengthening still allows for conditional heteroskedasticity and other plausible features of economic shocks).[7]

Though lag augmentation robustifies and simplifies local projection inference, it is not necessarily a free lunch. We show in Section 3 that the relative efficiency of non-augmented and lag-augmented local projection estimators depends on ρ and h.

Lag-Augmented Local Projection Inference. Define the nominal $100(1-\alpha)\%$ lag-augmented LP confidence interval for the impulse response at horizon h based on the standard error $\hat{s}(h)$:

$$\hat{C}(h,\alpha) \equiv \left[\hat{\beta}(h) - z_{1-\alpha/2}\,\hat{s}(h)\,,\ \hat{\beta}(h) + z_{1-\alpha/2}\,\hat{s}(h)\right],$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

Our main result shows that the lag-augmented LP confidence interval above is valid regardless of the persistence of the data, i.e., whether or not the data has a unit root. Crucially, the result does not break down at moderately long horizons h. We provide a formal result for VAR(p) models in Section 4 and for now just discuss heuristics. Consider any upper bound $\bar{h}_T$ on the horizon which satisfies $\bar{h}_T/T \to 0$. Then Proposition 1 below implies that

$$\inf_{\rho\in[-1,1]} \inf_{1\le h\le \bar{h}_T} P_\rho\left(\beta(\rho,h) \in \hat{C}(h,\alpha)\right) \to 1-\alpha \quad \text{as } T\to\infty, \tag{8}$$

where P_ρ denotes the distribution of the data $\{y_t\}$ under the AR(1) model (1) with parameter ρ. In words, the result states that, for sufficiently large sample sizes, LP inference is valid even under the *worst-case* choices of parameter $\rho \in [-1,1]$ and horizon $h \in [1,\bar{h}_T]$. As is well known, such *uniform* validity is a much stronger result than *pointwise* validity for fixed ρ and h. In fact, if we restrict attention to only the stationary region $\rho \in [-1+a, 1-a]$, $a \in (0,1)$, then the statement (8) is true with the upper bound $\bar{h}_T = (1-a)T$ on the horizon. That is, if we know the time series is not close to a unit root, then local projection inference is valid even at long horizons h that are non-negligible fractions of the sample size T.

---

instrumental variable, under some conditions on the instrument.

[7]The nuisance coefficient $\hat{\gamma}(h)$ is not interesting *per se*, but note that inference on this coefficient would generally require HAR standard errors, and its limit distribution is in fact non-standard when $\rho \approx 1$.

2.2 Illustrative Simulation Study

We now present a small simulation study to show that lag-augmented LP achieves a favorable trade-off between robustness and efficiency relative to other procedures. For clarity, we continue to assume the simple AR(1) model (1) with known lag length. Our baseline design considers homoskedastic innovations $u_t \overset{i.i.d.}{\sim} N(0,1)$. In Appendix B.1 we present results for ARCH innovations.

We stress that, although we use the AR(1) model for illustration here, the central goal of this paper is to develop a procedure that is feasible even in realistic VAR(p) models. Thus, we avoid computationally demanding procedures, such as the AR grid bootstrap, which are difficult to implement in applied settings. We provide an extensive theoretical comparison of various inference procedures in Section 3.

Table 1 displays the coverage and median length of impulse response confidence intervals at various horizons. We consider several versions of AR inference and LP inference, either implemented using the bootstrap or using delta method standard errors. "LP" denotes local projection and "AR" autoregressive inference. "LA" denotes lag augmentation. The subscript "b" denotes bootstrap confidence intervals constructed from a wild recursive bootstrap design (Gonçalves and Kilian, 2004), as described in Section 5 (for LP we use the percentile-t confidence interval). Columns without the "b" subscript use delta method standard errors. For LA-LP, we always use Eicker-Huber-White standard errors as discussed in Section 2.1, whereas non-augmented LP always uses HAR standard errors.[8] The column "AR-LA" is the Efron bootstrap confidence interval for *lag-augmented* AR estimates developed by Inoue and Kilian (2020) and discussed further in Section 3.[9] All estimation procedures include an intercept. The sample size is $T = 240$. We consider data generating processes (DGPs) $\rho \in \{0, .5, .95, 1\}$ and horizons h up to 60 periods (25% of the sample size, which is not unusual in applied work). The nominal confidence level is 90%. We use 5,000 Monte Carlo repetitions, with 2,000 bootstrap draws per repetition.

Consistent with our theoretical results, the bootstrap version of lag-augmented local projection (column 1) achieves coverage close to the nominal level in almost all cases, whereas the competing procedures either under-cover or return impractically wide confidence inter-

[8] As an off-the-shelf, state-of-the-art HAR procedure, we choose the Equally Weighted Cosine (EWC) estimator with degrees of freedom as recommended by Lazarus et al. (2018, equations 4 and 10). The degrees of freedom depend on the effective sample size $T - h$ and thus differ across horizons h.

[9] We use the Pope (1990) bias-corrected AR estimates to generate the bootstrap samples, as recommended by Inoue and Kilian (2020).

Table 1: Monte Carlo results: homoskedastic innovations

	Coverage						Median length					
h	LP-LA$_b$	LP-LA	LP$_b$	LP	AR-LA$_b$	AR	LP-LA$_b$	LP-LA	LP$_b$	LP	AR-LA$_b$	AR
						$\rho = 0.00$						
1	0.902	0.892	0.912	0.889	0.891	0.894	0.218	0.211	0.233	0.215	0.211	0.210
6	0.908	0.899	0.916	0.898	0.000	1.000	0.219	0.214	0.233	0.220	0.000	0.000
12	0.909	0.900	0.903	0.897	0.000	1.000	0.222	0.217	0.230	0.226	0.000	0.000
36	0.903	0.895	0.903	0.898	0.000	1.000	0.235	0.229	0.244	0.239	0.000	0.000
60	0.898	0.886	0.894	0.889	0.000	0.979	0.252	0.244	0.261	0.255	0.000	0.000
						$\rho = 0.50$						
1	0.906	0.896	0.912	0.885	0.897	0.897	0.219	0.212	0.205	0.187	0.211	0.184
6	0.895	0.886	0.906	0.875	0.897	0.832	0.252	0.245	0.293	0.266	0.046	0.032
12	0.906	0.894	0.903	0.887	0.897	0.766	0.255	0.248	0.293	0.280	0.002	0.001
36	0.900	0.889	0.901	0.884	0.897	0.643	0.271	0.262	0.309	0.296	0.000	0.000
60	0.905	0.891	0.903	0.880	0.897	0.595	0.291	0.279	0.333	0.316	0.000	0.000
						$\rho = 0.95$						
1	0.892	0.878	0.842	0.827	0.882	0.850	0.220	0.212	0.076	0.072	0.212	0.075
6	0.903	0.838	0.851	0.789	0.882	0.810	0.523	0.452	0.395	0.345	1.011	0.318
12	0.889	0.806	0.853	0.752	0.882	0.769	0.678	0.550	0.644	0.518	1.744	0.430
36	0.885	0.814	0.865	0.674	0.882	0.656	0.728	0.625	0.859	0.612	6.567	0.272
60	0.892	0.833	0.892	0.693	0.882	0.595	0.731	0.651	0.942	0.641	23.050	0.095
						$\rho = 1.00$						
1	0.895	0.874	0.820	0.554	0.877	0.532	0.219	0.211	0.040	0.040	0.210	0.039
6	0.875	0.777	0.836	0.503	0.877	0.494	0.564	0.498	0.243	0.222	1.206	0.214
12	0.843	0.676	0.827	0.429	0.877	0.459	0.821	0.671	0.477	0.385	2.553	0.379
36	0.741	0.428	0.755	0.200	0.877	0.348	1.338	0.950	1.200	0.592	21.107	0.670
60	0.642	0.276	0.712	0.156	0.877	0.295	1.434	0.978	1.667	0.637	161.250	0.731

Coverage probability and median length of nominal 90% confidence intervals at different horizons. AR(1) model with $\rho \in \{0, .5, .95, 1\}$, $T = 240$, i.i.d. standard normal innovations. 5,000 Monte Carlo repetitions; 2,000 bootstrap iterations.

vals. In contrast, non-augmented LP (columns 3 and 4) exhibits larger coverage distortions in almost all cases. As is well known, textbook AR delta method confidence intervals (column 6) severely under-cover when $\rho > 0$ and the horizon is even moderately large.

It is only when both $\rho = 1$ and $h \geq 36$ that lag-augmented local projection exhibits serious coverage distortions, again consistent with our theory. However, even in these cases, the coverage distortions are similar to or less pronounced than those for non-augmented LP and for delta method AR inference.

Although the Inoue and Kilian (2020) lag-augmented AR bootstrap confidence interval (column 5) achieves correct coverage for $\rho > 0$ at all horizons, this interval is extremely wide in the problematic cases where ρ is close to 1 and the horizon h is intermediate or long. We explain this fact theoretically in Section 3. Confidence intervals with median width greater than 1 would appear to be of little practical use, since the true impulse response parameter is bounded above by 1 in the AR(1) model.[10] Note also that the Inoue and Kilian (2020) interval severely under-covers when $\rho = 0$ at all even (but not odd) horizons h, as explained theoretically in Section 3.[11]

Although outperformed by bootstrap procedures, the lag-augmented local projection delta method interval (column 2) performs well among the group of delta method procedures. Its coverage distortions are much less severe than textbook AR delta method inference (column 4) and non-augmented LP inference with HAR standard errors (column 6). Recall that the lag-augmented LP confidence interval is at least as easy to compute as these other delta method confidence intervals. The reason why the bootstrap improves on the coverage properties of the delta method procedures is related to the well-known finite-sample bias of AR and LP estimators (Kilian, 1998; Herbst and Johannsen, 2020).[12]

Table 1 illustrates the fact that the robustness of lag-augmented local projection inference entails an efficiency loss relative to AR inference when ρ is well below 1, although this loss is not large in absolute terms. In *percentage* terms, local projection confidence intervals are much wider than AR-based confidence intervals when $\rho \ll 1$ and the horizon h is intermediate or long, since AR procedures mechanically impose that the impulse response function tends to 0 geometrically fast with the horizon. Yet, in *absolute* terms, the median length of the LP confidence intervals is not so large as to be a major impediment to applied research.

[10] In the AR(1) model, we could intersect all confidence intervals with the interval $[-1, 1]$. In this case, the median length of the Inoue and Kilian (2020) confidence interval is close to 1, cf. Appendix B.2.2.

[11] Inoue and Kilian (2020) assume $\rho \neq 0$ and discuss why this restriction is necessary in their case.

[12] Our bootstrap implementation of *non-augmented* LP also appears to be quite effective at correcting the most severe coverage distortions of the delta method procedure.

The relative efficiency of lag-augmented LP vs. non-augmented LP cannot be ranked and depends on the DGP and on the horizon. When ρ is close to 1, lag-augmented LP intervals are sometimes (much) narrower than lag-augmented AR intervals. We analytically characterize the various efficiency trade-offs in Appendix B.2.1.

Supplemental Appendix D shows that the preceding qualitative conclusions extend to richer models. There we consider a bivariate VAR(4) model with varying degrees of persistence, as well as two empirically calibrated VAR(12) models with four or five observables.

3 Comparison With Other Inference Procedures

The simulations and theoretical results in this paper suggest that lag-augmented local projection is the only known confidence interval procedure that achieves uniformly valid coverage over the DGP and over a wide range of horizons, while preserving reasonable average length and remaining computationally feasible in realistic settings. However, the simulations also suggest that lag-augmented local projection inference is less efficient than standard AR inference when the data is stationary. In this section we discuss in more detail the coverage and length properties of alternative confidence interval procedures for impulse responses. We review the well-known drawbacks of textbook AR inference, provide new results on the relative length of lag-augmented LP vs. non-augmented LP and lag-augmented AR, and discuss the computational challenges of the AR grid bootstrap. We refer the reader back to the small-scale simulation study in Section 2.2 for illustrations of the following arguments.

Textbook Autoregressive Inference. The uniformity result (8) for lag-augmented LP stands in stark contrast to textbook AR inference on impulse responses, which suffers from several well-known issues. First, for the standard OLS AR estimator, the usual asymptotic normal limiting theory is invalid when the derivative of the impulse response parameter with respect to the AR coefficients has a singular Jacobian matrix. In the AR(1) model, this occurs at all horizons $h \geq 2$ in the white noise case $\rho = 0$ (Benkwitz et al., 2000). Second, as with non-augmented LP, textbook AR inference is not uniformly valid when the data is nearly non-stationary, unless one further restricts the parameter space (Phillips, 1998, Remark 2.5; Inoue and Kilian, 2002; Inoue and Kilian, 2020, Remark 3, p. 455).[13] Third, pre-testing for the presence of a unit root does not yield uniformly valid inference and can

[13]This is well known in the AR(1) model. In the AR(2) model, a non-normal limit arises at $h = 2$ when there is a unit root and the autoregressive coefficients are equal (Inoue and Kilian, 2020, Remark 3, p. 455).

lead to poor finite sample performance (e.g., Mikusheva, 2007, p. 1412). Fourth, plug-in AR inference with normal critical values must necessarily break down at medium-long horizons $h = h_T \propto T^{1/2}$ and at long horizons $h_T \propto T$, due to the severe nonlinearity of the impulse response transformation at such horizons (Mikusheva, 2012, Section 4.3). Wright (2000) and Pesavento and Rossi (2006, 2007) construct confidence intervals for persistent processes at long horizons $h = h_T \propto T$ by inverting the non-standard AR limit distribution, but these tailored procedures do not work uniformly over the parameter space or over the horizon.

The severe under-coverage of delta method AR inference is starkly illustrated in Section 2.2 (see column 6 of Table 1). As discussed in detail by Inoue and Kilian (2020), standard bootstrap approaches to AR inference do not solve all the uniformity issues.

We must emphasize, however, that if we restrict attention to stationary processes and short-horizon impulse responses, the standard OLS AR impulse response estimator is more efficient than lag-augmented LP. Hence, there is a trade-off between efficiency in benign settings and robustness to persistence and longer horizons, as is also clear in the simulation results in Section 2.2. We expand upon the efficiency properties of the standard AR estimator in Appendix B.2.1.

Lag-Augmented AR Inference. The above-mentioned non-uniformity of the textbook AR inference method in the case of near-non-stationary data can be remedied by lag augmentation (Inoue and Kilian, 2020). In the case of an AR(1) model, the lag-augmented AR estimator $\hat{\beta}_{\text{ARLA}}(h)$ is given by $\hat{\rho}_1^h$, where $(\hat{\rho}_1, \hat{\rho}_2)$ are the OLS coefficients from a regression of y_t on (y_{t-1}, y_{t-2}) (i.e., we estimate an AR(2) model). The intuition why this guarantees a normal limiting distribution even in the unit root case is the same as in Section 2.1. Lag-augmented AR and lag-augmented LP coincide at horizon $h = 1$, but not at longer horizons. Lag augmentation involves a loss of efficiency: The lag-augmented AR estimator is strictly less efficient than the non-augmented AR estimator except when the true process is white noise (see Appendix B.2.1). Note that lag augmentation by itself does not solve the above-mentioned issues that occur when the Jacobian of the impulse response transformation is singular, or when doing inference at medium-long or long horizons.[14]

The bootstrap confidence interval for lag-augmented AR proposed by Inoue and Kilian (2020) has valid coverage even at long horizons. Specifically, Inoue and Kilian (2020) show

[14]The AR(1) simulations in Section 2.2 show that the coverage of the Inoue and Kilian (2020) confidence interval is 0 at all *even* horizons when $\rho = 0$. This is because the true impulse response is 0, but the bootstrap samples of $\hat{\rho}_1^h$ are all strictly positive. Their procedure achieves uniformly correct coverage at *odd* horizons.

that the *Efron* bootstrap confidence interval—applied to recursive AR bootstrap samples of $\hat{\beta}_{\text{ARLA}}(h)$—has valid coverage even at long horizons $h = h_T \propto T$, as long as the largest autoregressive root is bounded away from 0.[15]

Unfortunately, we show in Appendix B.2.2 that the expected length of the lag-augmented AR interval is prohibitively large when the data is persistent and the horizon is long. Precisely, in the case of an AR(1) model, $\hat{\beta}_{\text{ARLA}}(h) = \hat{\rho}_1^h$ is *inconsistent* for sequences of DGPs $\rho = \rho_T$ and horizons $h = h_T$ such that $h_T \propto T^\eta, \eta \in [1/2, 1]$, and $h_T(1 - \rho_T) \to a \in [0, \infty)$. The reason is that the lag-augmented coefficient estimator $\hat{\rho}_1$ converges at rate $T^{-1/2}$ even in the unit root case, implying that the estimation error in $\hat{\rho}_1$ is not negligible when raising the estimator to a power of $h = h_T$. This implies that the Efron bootstrap confidence interval is inconsistent (i.e., its length does not shrink to 0 in probability) for such sequences ρ_T and h_T. In fact, when $\eta > 1/2$, the width of the confidence interval for the h_T impulse response is almost equal to the entire positive part of the parameter space $[0, 1]$ with probability equal to the nominal confidence level. This contrasts with the lag-augmented LP confidence interval, which is consistent for any sequence $\rho_T \in [-1, 1]$ and any sequence h_T such that $h_T/T \to 0$. The large width of the Inoue and Kilian (2020) interval is illustrated in the simulations in Section 2.2 (see the second-to-last column in Table 1).

Interestingly, *if we restrict attention to stationary processes and short horizons, the relative efficiency of lag-augmented AR and lag-augmented LP inference is ambiguous.* In the context of a stationary, homoskedastic AR(1) model with a fixed horizon h of interest, Figure 1 shows that lag-augmented AR is more efficient than lag-augmented LP when ρ is small or when the horizon h is large, and vice versa. For any horizon h, there exists some cut-off value for $\rho \in (0, 1)$, above which lag-augmented LP is more efficient. Intuitively, the nonlinear impulse response transformation $\rho \mapsto \rho^h$ is highly sensitive to values of ρ near 1 whenever h is large, which compounds the effects of estimation error in $\hat{\rho}$, whereas LP is a purely linear procedure.

AR Grid Bootstrap and Projection. The grid bootstrap of Hansen (1999) represents a computationally intensive approach to doing valid inference at fixed and long horizons, regardless of persistence, but it is invalid at intermediate horizons, as shown by Mikusheva

[15]For intuition, consider the AR(1) case. The Efron bootstrap preserves monotonic transformations, and the bootstrap transformation $\beta(\rho, h) = \rho^h$ is monotonic (if we restrict attention to $\rho \in (0, 1]$ or $\rho \in [-1, 0)$). Hence, the Efron confidence interval is valid for ρ^h if it is valid for ρ itself. In more general VAR(p) models, the same argument can be applied at long horizons, since here only the largest autoregressive root matters for impulse responses (if the roots are well-separated).

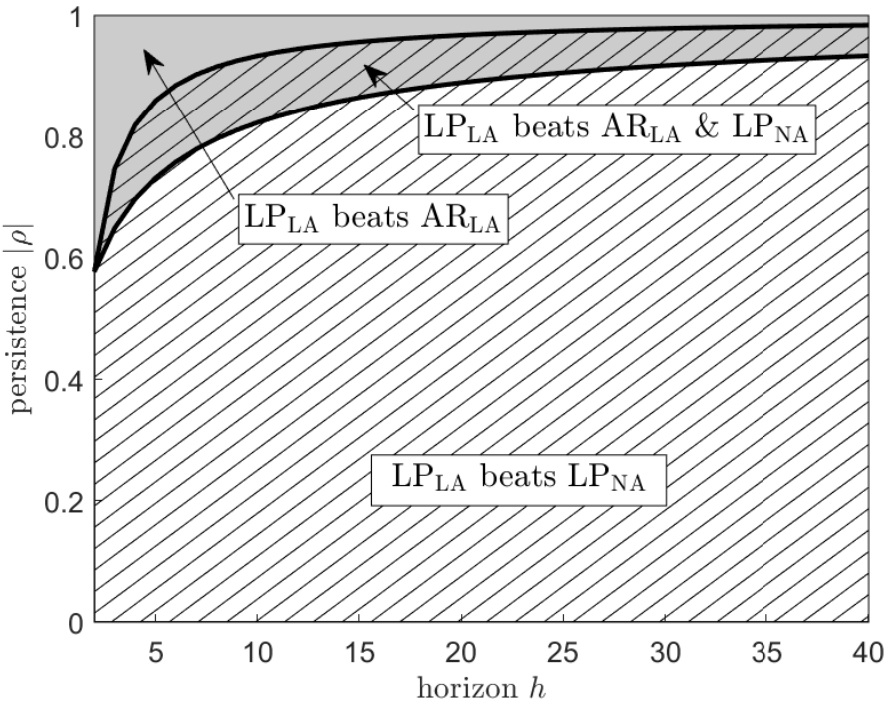

Figure 1: Efficiency ranking of three different estimators of the fixed impulse response $\beta(\rho, h) = \rho^h$ in the homoskedastic AR(1) model: lag-augmented LP ($\mathrm{LP_{LA}}$), non-augmented LP ($\mathrm{LP_{NA}}$), and lag-augmented AR ($\mathrm{AR_{LA}}$). Gray area: combinations of $(|\rho|, h)$ for which $\mathrm{LP_{LA}}$ is more efficient than $\mathrm{AR_{LA}}$. Thatched area: $\mathrm{LP_{LA}}$ is more efficient than $\mathrm{LP_{NA}}$. See Appendix B.2.1 for analytical derivations of the indifference curves (thick lines).

(2012). The grid bootstrap is based on test inversion, so it requires running an autoregressive bootstrap on each point in a fine grid of potential values for the impulse response parameter of interest. It also requires estimating a constrained OLS estimator that imposes the hypothesized null on the impulse response at each point in the grid. Recall that lag-augmented LP inference is computationally simple and valid at any horizon $h = h_T$ satisfying $h_T/T \to 0$. However, in the case of unit roots and long horizons $h_T \propto T$, lag-augmented LP inference with normal critical values is not valid, while the grid bootstrap is valid (Mikusheva, 2012).

Another computationally intensive approach is to form a uniformly valid confidence set for the AR parameters and then map it into a confidence interval for impulse responses by projection. Although doable in the AR(1) model, this approach would appear to be computationally infeasible and possibly highly conservative in realistic VAR(p) settings, unlike lag-augmented LP (see Section 4).

Other Local Projection Approaches. Non-augmented LP is not robust to non-stationarity, as already discussed in Section 2.1. *If the data is stationary and the horizon h is fixed, the relative efficiency of non-augmented LP and lag-augmented LP is generally ambiguous*, as shown in Figure 1 in the case of a homoskedastic AR(1) model. There are two competing forces. On the one hand, as shown in Section 2.1, non-augmented LP uses the regressor y_t, which has higher variance than the effective regressor u_t in the lag-augmented case. By itself, this suggests that non-augmented LP should be more efficient. On the other hand, absent lag augmentation, the LP regression scores are serially correlated and thus have a larger long-run variance. On balance, Appendix B.2.1 shows that lag-augmented LP is relatively more efficient the smaller is ρ and the larger is h.

In some empirical settings, the researcher may directly observe the autoregressive innovation, or some component of the innovation, for example by constructing narrative measures of economic shocks (Ramey, 2016). For concreteness, consider the AR(1) model (1) and assume we observe the innovation u_t. In this case, it is common in empirical practice to simply regress y_{t+h} on u_t, without controls. Although this strategy provides consistent impulse response estimates when the data is stationary, it is inefficient relative to lag-augmented LP, since the latter approach additionally controls for the variable y_{t-1}, which would otherwise show up in the error term in the representation (4). Thus, lag augmentation is desirable on robustness and efficiency grounds even if some shocks are directly observed.

Summary. Existing and new theoretical results confirm the main message of our simulations in Section 2.2: Lag-augmented LP is the only known procedure that is computationally feasible in realistic problems and can be shown to have valid coverage under a wide range of DGPs and horizon lengths, without achieving such valid coverage by returning a confidence interval that is impractically wide. This robustness does come at the cost of a loss of efficiency relative to non-robust AR methods. However, the efficiency loss is large in *relative* terms only in stationary, short-horizon cases, where lag-augmented LP confidence intervals do well in *absolute* terms, as illustrated in Section 2.2. Based on these results, we believe that it is only in the case of highly persistent data and very long horizons $h = h_T \propto T$ that the use of alternative robust procedures should be considered, such as the computationally demanding AR grid bootstrap.

4 General Theory for the VAR(p) Model

This section presents the inference procedure and theoretical uniformity result for a general VAR(p) model. In this case, the lag-augmented LP procedure controls for p lags of all the time series that enter into the VAR model. We follow Mikusheva (2012) and Inoue and Kilian (2020) in assuming that the lag length p is finite and known. We also assume that the VAR process has no deterministic dynamics for simplicity. See Section 6 for further discussion of these assumptions.

4.1 Model and Inference Procedure

Consider an n-dimensional VAR(p) model for the data $y_t = (y_{1,t}, \ldots, y_{n,t})'$:

$$y_t = \sum_{\ell=1}^{p} A_\ell y_{t-\ell} + u_t, \quad t = 1, 2, \ldots, T, \quad y_0 = \cdots = y_{1-p} = 0, \tag{9}$$

Let $A \equiv (A_1, \ldots, A_p)$ denote the $n \times np$ matrix collecting all the autoregressive coefficients. The assumption of zero pre-sample initial conditions $y_0 = \cdots = y_{1-p} = 0$ is made for notational simplicity and can be relaxed, as discussed below in the remarks after Proposition 1. As in the AR(1) case, we assume that the n-dimensional innovation process $\{u_t\}$ satisfies the strengthening of the martingale difference condition in Assumption 1 (which from now on will refer to the vector process $\{u_t\}$).

We seek to do inference on a scalar function of the reduced-form impulse responses of the VAR model. Generalizations to *structural* impulse responses and *joint* inference require more notation but are otherwise straight-forward, see Section 6. Let $\beta_i(A, h)$ denote the $n \times 1$ vector containing each of variable i's reduced-form impulse responses at horizon $h \geq 0$. Without loss of generality, we focus on the impulse responses of the first variable $y_{1,t}$. Thus, we seek a confidence interval for the scalar parameter $\nu'\beta_1(A, h)$, where $\nu \in \mathbb{R}^n \backslash \{0\}$ is a user-specified vector. For example, the choice $\nu = e_j$ (the j-th unit vector) selects the horizon-h response of $y_{1,t}$ with respect to the j-th reduced-form innovation $u_{j,t}$.

Local projection estimators of impulse responses are motivated by the representation

$$y_{1,t+h} = \beta_1(A, h)' y_t + \sum_{\ell=1}^{p-1} \delta_{1,\ell}(A, h)' y_{t-\ell} + \xi_{1,t}(A, h), \tag{10}$$

see Jordà (2005) and Kilian and Lütkepohl (2017, Chapter 12.8). Here $\delta_{1,\ell}(A, h)$ is an $n \times 1$

vector of regression coefficients that can be obtained by iterating on the VAR model (9). The model-implied multi-step forecast error in this regression is

$$\xi_{1,t}(A,h) \equiv \sum_{\ell=1}^{h} \beta_1(A,h-\ell)' u_{t+\ell}. \tag{11}$$

Multivariate Lag-Augmented Local Projection. The lag-augmented LP estimator corresponding to the VAR model (9) is motivated by (10). We regress $y_{1,t+h}$ on the n variables y_t, using the np variables $(y'_{t-1},\ldots,y'_{t-p})$ as additional controls. According to equation (10), the population regression coefficients on the last n control variables y_{t-p} equal zero. Thus, we are including one additional lag in the estimation of the impulse response coefficients. Given any horizon $h \in \mathbb{N}$, the lag-augmented LP estimator $\hat{\beta}_1(h)$ of $\beta_1(A,h)$ is given by the vector of coefficients on y_t in the regression of $y_{1,t+h}$ on $x_t \equiv (y'_t, y'_{t-1},\ldots,y'_{t-p})'$:

$$\begin{pmatrix} \hat{\beta}_1(h) \\ \hat{\gamma}_1(h) \end{pmatrix} \equiv \left(\sum_{t=1}^{T-h} x_t x_t' \right)^{-1} \sum_{t=1}^{T-h} x_t y_{1,t+h}, \tag{12}$$

where $\hat{\beta}_1(h)$ is a vector of dimension $n \times 1$.

The usual (Eicker-Huber-White) heteroskedasticity-robust standard error for $\nu'\hat{\beta}_1(h)$ is defined as

$$\hat{s}_1(h,\nu) \equiv \frac{1}{T-h} \left\{ \nu' \hat{\Sigma}(h)^{-1} \left(\sum_{t=1}^{T-h} \hat{\xi}_{1,t}(h)^2 \hat{u}_t(h) \hat{u}_t(h)' \right) \hat{\Sigma}(h)^{-1} \nu \right\}^{1/2},$$

where

$$\hat{\xi}_{1,t}(h) \equiv y_{1,t+h} - \hat{\beta}_1(h)' y_t - \hat{\gamma}_1(h)' X_t, \quad X_t \equiv (y'_{t-1},\ldots,y'_{t-p})',$$

$$\hat{u}_t(h) \equiv y_t - \hat{A}(h) X_t, \quad \hat{A}(h) \equiv \left(\sum_{t=1}^{T-h} y_t X_t' \right) \left(\sum_{t=1}^{T-h} X_t X_t' \right)^{-1},$$

and

$$\hat{\Sigma}(h) \equiv \frac{1}{T-h} \sum_{t=1}^{T-h} \hat{u}_t(h) \hat{u}_t(h)'.$$

The $1-\alpha$ confidence interval for $\nu'\beta_1(A,h)$ is defined as

$$\hat{C}_1(h,\nu,\alpha) \equiv \left[\nu'\hat{\beta}_1(h) - z_{1-\alpha/2}\, \hat{s}_1(h,\nu)\ ,\ \nu'\hat{\beta}_1(h) + z_{1-\alpha/2}\, \hat{s}_1(h,\nu) \right].$$

Parameter Space. We consider a class of VAR processes with possibly multiple unit roots combined with arbitrary stationary dynamics. Specifically, we will prove that the confidence interval $\hat{C}_1(h,\nu,\alpha)$ has uniformly valid coverage over the following parameter space. Let $\|M\| \equiv \sqrt{\text{trace}(M'M)}$ denote the Frobenius matrix norm, and let I_n denote the $n \times n$ identity matrix.

Definition 1 (VAR parameter space)**.** *Given constants $a \in [0,1)$, $C > 0$, and $\epsilon \in (0,1)$, let $\mathcal{A}(a,C,\epsilon)$ denote the space of autoregressive coefficients $A = (A_1,\ldots,A_p)$ such that the associated p-dimensional lag-polynomial $A(L) = I_n - \sum_{\ell=1}^{p} A_\ell L^\ell$ admits the factorization*

$$A(L) = B(L)(I_n - \text{diag}(\rho_1,\ldots,\rho_n)L), \tag{13}$$

where $\rho_i \in [a-1, 1-a]$ for all $i = 1,\ldots,n$, and $B(L)$ is a lag polynomial of order $p-1$ with companion matrix $\mathbf{B}$ satisfying $\|\mathbf{B}^\ell\| \leq C(1-\epsilon)^\ell$ for all $\ell = 1,2,\ldots$.[16]

This parameter space contains any stationary VAR process (for sufficiently small a,ϵ and sufficiently large C) as well as many—but not all—non-stationary processes. Lag polynomials $A(L)$ in this parameter space imply that the process $\{y_t\}$ can be written in the form $y_t = \text{diag}(\rho_1,\ldots,\rho_n)y_{t-1} + \tilde{y}_t$, where $\tilde{y}_t \equiv B(L)^{-1}u_t$ is a stationary process whose impulse responses at horizon ℓ decay at the geometric rate $(1-\epsilon)^\ell$. We allow all the roots $\rho_1,\ldots,\rho_n$ to be potentially close to or equal to 1. Mikusheva (2012, Section 4.2) considers the same class of processes but with $\rho_2 = \cdots = \rho_n = 0$. We are not aware of other uniform inference results that allow multiple near-unit roots. Although the parameter space in Definition 1 appears more restrictive than the local-to-unity framework of Phillips (1988, Eqn. 2), we argue below that our uniform coverage result applied to the parameter space $\mathcal{A}(a,C,\epsilon)$ immediately implies an extended result that also covers processes with cointegration among the control variables $y_{2,t},\ldots,y_{n,t}$. However, we do impose the restriction that the response variable of interest $y_{1,t}$ has at most one root near unity, as in Wright (2000), Pesavento and Rossi (2006), Mikusheva (2012), and Inoue and Kilian (2020).

4.2 Additional Assumptions

Our main result requires two further technical assumptions in addition to Assumption 1. Let $\lambda_{\min}(M)$ denote the smallest eigenvalue of a symmetric positive semidefinite matrix M.

[16] See Appendix A for the standard definition of a companion matrix.

Assumption 2.

i) $E(\|u_t\|^8) < \infty$, and there exists $\delta > 0$ such that $\lambda_{\min}(E[u_t u_t' \mid \{u_s\}_{s<t}]) \geq \delta$ almost surely.

ii) The process $\{u_t \otimes u_t\}$ has absolutely summable cumulants up to order 4.

Part (i) of Assumption 2 is a common requirement for consistent estimation of regression standard errors with possibly heteroskedastic residuals. Part (ii) is a standard weak dependence restriction on the second moments of u_t (Brillinger, 2001, Chapter 2.6).

We will write $\rho(A) = (\rho_1(A), \ldots, \rho_n(A))'$ to represent any of the possible vectors of roots $\rho_1, \ldots, \rho_n$ corresponding to a collection of autoregressive coefficients $A = (A_1, \ldots, A_p) \in \mathcal{A}(0, C, \epsilon)$. This is a slight abuse of notation, since the mapping from $A(L)$ to ρ_i's is one-to-many. Define $g(\rho, h)^2 \equiv \min\{\frac{1}{1-|\rho|}, h\}$ and $\rho_i^*(A, \epsilon) \equiv \max\{|\rho_i(A)|, 1 - \epsilon/2\}$. Define also the $np \times np$ diagonal matrix $G(A, h, \epsilon) \equiv I_p \otimes \text{diag}(g(\rho_1^*(A, \epsilon), h), \ldots, g(\rho_n^*(A, \epsilon), h))$.

Assumption 3. *For any $C > 0$ and $\epsilon \in (0, 1)$,*

$$\lim_{K\to\infty} \lim_{T\to\infty} \inf_{A \in \mathcal{A}(0,C,\epsilon)} P_A \left(\lambda_{\min} \left(G(A, T, \epsilon)^{-1} \left[\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right] G(A, T, \epsilon)^{-1} \right) \geq 1/K \right) = 1.$$

This high-level assumption ensures that the properly scaled (matrix) "denominator" in the VAR OLS estimator $\hat{A}(h)$ is uniformly non-singular asymptotically, so the estimator is uniformly well-defined with high probability in the limit. Hence, the assumption is essentially necessary for our result.

How can Assumption 3 be verified? $G(A_T, T, \epsilon)^{-1} \left[\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right] G(A_T, T, \epsilon)^{-1}$ is known to converge in distribution in a *pointwise* sense to an almost surely positive definite (perhaps stochastically degenerate) random matrix under stationary, local-to-unity, or unit root sequences $\{A_T\}$ (e.g., Phillips, 1988; Hamilton, 1994).[17] Assumption 3 requires that such convergence obtains for *all* possible sequences $\{A_T\}$. In Appendix C we illustrate how the assumption can be verified in the AR(1) model under an additional weak condition on the innovation process.

[17] Note that the diagonal entries of $G(A, T, \epsilon)^{-1}$ are constants for stationary VAR coefficient matrices A, whereas these diagonal entries are proportional to $T^{-1/2}$ under local-to-unity or unit root sequences.

4.3 Main Result

We now state the result that the LP estimator $\nu'\hat{\beta}_1(h)$ is asymptotically normally distributed uniformly over the parameter space in Definition 1, even at long horizons h. Let P_A denote the probability measure of the data $\{y_t\}$ when it is generated by the VAR(p) model (9) with coefficients $A \in \mathcal{A}(a, C, \epsilon)$. The distribution of the innovations $\{u_t\}$ is fixed.

Proposition 1. *Let Assumptions 1 to 3 hold. Let $C > 0$ and $\epsilon \in (0, 1)$.*

i) Let $a \in (0, 1)$. For all $x \in \mathbb{R}$,

$$\sup_{A \in \mathcal{A}(a,C,\epsilon)} \sup_{1 \le h \le (1-a)T} \left| P_A \left(\frac{\nu'[\hat{\beta}_1(h) - \beta_1(A, h)]}{\hat{s}_1(h, \nu)} \le x \right) - \Phi(x) \right| \to 0.$$

ii) Consider any sequence $\{\bar{h}_T\}$ of nonnegative integers such that $\bar{h}_T < T$ for all T and $\bar{h}_T/T \to 0$. Then for all $x \in \mathbb{R}$,

$$\sup_{A \in \mathcal{A}(0,C,\epsilon)} \sup_{1 \le h \le \bar{h}_T} \left| P_A \left(\frac{\nu'[\hat{\beta}_1(h) - \beta_1(A, h)]}{\hat{s}_1(h, \nu)} \le x \right) - \Phi(x) \right| \to 0.$$

Proof. See Appendix A. □

The uniform asymptotic normality established above immediately implies that the confidence interval $\hat{C}_1(h, \nu, \alpha)$ has uniformly valid coverage asymptotically. Part (i) considers stationary VAR processes whose largest roots are bounded away from 1; then inference is valid even at long horizons $h = h_T \propto T$. Part (ii) allows all or some of the n roots $\rho_1, \ldots, \rho_n$ to be near or equal to 1, but then we require $h_T/T \to 0$.

Remarks.

1. The proof of Proposition 1 shows that the uniform convergence rate of $\hat{\beta}_1(h_T)$ is $O_p((h_T/T)^{1/2})$ if $h_T/T \to 0$. This rate may be slower than that of the possibly superconsistent non-augmented LP estimator, which is the price to pay for uniformity. If we restrict attention to the stationary parameter space $\mathcal{A}(a, C, \epsilon)$, $a > 0$, the convergence rate of $\hat{\beta}_1(h_T)$ is $O_p(T^{-1/2})$ provided that $h_T \le (1 - a)T$.

2. There are three main challenges in establishing the uniform validity of local projection inference.

a) The variance of the regression residual $\xi_{1,t}(A,h)$ is increasing in the horizon h and also depends on A. Thus, the simplest laws of large numbers and central limit theorems for stationary processes do not apply. We instead apply a central limit theorem for martingale difference sequences and derive uniform bounds on moments of relevant variables. The central limit theorem is delicate, since the regression scores $\xi_{1,t}(A,h)u_t$ are not a martingale difference sequence with respect to the natural filtration generated by past u_t's. However, it is possible to "reverse time" in a way that makes the scores a martingale difference sequence with respect to an alternative filtration, see the proof of the auxiliary Lemma A.1.

b) To handle both unit roots, stationary processes, and everything in between, we must consider various kinds of sequences of drifting parameters $A = A_T$, following the general logic of Andrews et al. (2019). This is primarily an issue when showing consistency of the standard error $\hat{s}_1(h,\nu)$, which requires deriving the convergence rates of the various estimators along drifting parameter sequences. We do this by explicit calculation of moment bounds that are uniform in the both the DGP and the horizon.

c) Our proof requires bounds on the rate of decay of impulse response functions that are uniform in both the DGP and the horizon. Though the AR(1) case is trivial due to the monotonically decreasing exponential functional form $\beta(\rho,h) = \rho^h$, the bounds for the general VAR(p) case require more work, see especially Lemma E.4 in Supplemental Appendix E.2. These results may be of independent interest.

3. Proposition 1 does not cover the case where $h \propto T$ and some of the roots ρ_i are local-to-unity or equal to unity. Simulation evidence and analytical calculations along the lines of Hjalmarsson and Kiss (2020) strongly suggest that even in the AR(1) model the asymptotic normality of lag-augmented local projections does *not* go through when $\rho = 1$ and $h = \kappa T$ for $\kappa \in (0,1)$. Indeed, in this case the sample variance of the regression scores $\xi_t(\rho,h)u_t$ appears not to converge in probability to a constant, thus violating the conclusion of the key auxiliary Lemma A.6 below. As discussed in Section 3, the behavior of plug-in autoregressive impulse response estimators is also non-standard when $\rho \approx 1$ and $h \propto T$.

4. A corollary of our main result is that we can allow for cointegrating relationships to exist among the control variables $y_{2,t},\ldots,y_{n,t}$. This is because both the LP estimator and the reduced-form impulse responses are equivariant with respect to non-singular linear

transformations of these $n-1$ variables. For example, consider a 3-dimensional process $(y_{1,t}, y_{2,t}, y_{3,t})$ that follows a VAR model in the parameter space in Definition 1 with $\rho_2 = 1, \rho_3 = 0$. Now consider the transformed process $(y_{1,t}, \tilde{y}_{2,t}, \tilde{y}_{3,t}) = (y_{1,t}, y_{2,t} + y_{3,t}, -y_{2,t} + y_{3,t})$. The variables $\tilde{y}_{2,t}$ and $\tilde{y}_{3,t}$ are cointegrated with cointegrating vector $(1,1)'$. Since $(\tilde{y}_{2,t}, \tilde{y}_{3,t})$ is a non-singular linear transformation of $(y_{2,t}, y_{3,t})$, the conclusions of Proposition 1 apply also to the transformed data vector.

5. If the vector of innovations u_t were observed, an alternative estimator would regress $y_{1,t+h}$ onto u_t and $y_{t-1}, \ldots, y_{t-p}$. As discussed in Section 2.1, this estimator is numerically equivalent with $\hat{\beta}_1(h)$, so the uniformity result carries over.

6. It is easily verified in our proofs that, rather than initializing the process at zero, we can allow the initial conditions $y_0, \ldots, y_{1-p}$ to be random variables that are independent of the innovations $\{u_t\}_{t\geq 1}$, as long as $E[\|y_\ell\|^4] < \infty$ for $\ell \leq 0$.

5 Bootstrap Implementation

In this section we describe the bootstrap implementation of lag-augmented local projection that we recommend for practical use. We find in simulations that the bootstrap procedure is effective at correcting small-sample coverage distortions. These distortions arise primarily due to the small-sample bias of local projection, which Herbst and Johannsen (2020) show is analogous to the well-known bias of the VAR OLS estimator (Kilian, 1998).

Our baseline algorithm is based on a wild autoregressive bootstrap design, which allows for heteroskedastic VAR innovations (Gonçalves and Kilian, 2004) as in our theoretical results. Guided by simulation evidence, we construct the bootstrap confidence interval using the equal-tailed percentile-t method, which has a built-in bias correction (Kilian, 1998; Kilian and Lütkepohl, 2017, Chapter 12.2.6).

The bootstrap procedure for computing a $1-\alpha$ confidence interval proceeds as follows, assuming a VAR(p) model:

1. Compute the impulse response estimate of interest $\nu'\hat{\beta}_1(h)$ and its standard error $\hat{s}_1(h,\nu)$ by lag-augmented local projection as in Section 4.1.

2. Estimate the VAR(p) model by OLS without lag augmentation. Compute the corresponding VAR residuals $\hat{u}_t$. Bias-adjust the VAR coefficients using the formula in Pope (1990) (this adjustment is optional, but improves finite-sample performance).

3. Compute the impulse response of interest implied by the VAR model estimated in step 2. Denote this impulse response by $\nu'\hat{\beta}_{1,\text{VAR}}(h)$.

4. For each bootstrap iteration $b = 1, \ldots, B$:

   i) Generate bootstrap residuals $\hat{u}_t^* \equiv U_t \hat{u}_t$, $t = 1, \ldots, T$, where $U_t \overset{i.i.d.}{\sim} N(0,1)$ are computer-generated random variables that are independent of the data.

   ii) Draw a block of p initial observations $(y_1^*, \ldots, y_p^*)$ uniformly at random from the $T - p + 1$ blocks of p observations in the original data.

   iii) Generate bootstrap data y_t^*, $t = p+1, \ldots, T$, by iterating on the bias-corrected VAR(p) model estimated in step 2, using the innovations $\hat{u}_t^*$.

   iv) Apply the lag-augmented LP estimator to the bootstrap data $\{y_t^*\}$. Denote the impulse response estimate and its standard error by $\nu'\hat{\beta}(h)^*$ and $\hat{s}_1(h,\nu)^*$, respectively.

   v) Store $\hat{T}_b^* \equiv (\nu'\hat{\beta}_1(h)^* - \nu'\hat{\beta}_{1,\text{VAR}}(h))/\hat{s}_1(h,\nu)^*$.[18]

5. Compute the $\alpha/2$ and $1-\alpha/2$ quantiles of the B draws of $\hat{T}_b^*$, $b = 1, \ldots, B$. Denote these by $\hat{Q}_{\alpha/2}$ and $\hat{Q}_{1-\alpha/2}$, respectively.

6. Return the percentile-t confidence interval[19]

$$[\nu'\hat{\beta}_1(h) - \hat{s}_1(h,\nu)\hat{Q}_{1-\alpha/2}, \nu'\hat{\beta}_1(h) - \hat{s}_1(h,\nu)\hat{Q}_{\alpha/2}].$$

Instead of the above recursive VAR design, it is also possible to use the standard fixed-design pairs bootstrap, as in any linear regression with serially uncorrelated scores.[20] In this case, the usual Efron bootstrap confidence interval is valid, like the percentile-t interval. However, simulations suggest that the pairs bootstrap procedure is less accurate in small samples than the above recursive bootstrap design, mirroring the results in Gonçalves and Kilian (2004) for autoregressive inference.

[18] It is critical that the bootstrap t-statistic $\hat{T}_b^*$ is centered at the VAR-implied impulse response $\nu'\hat{\beta}_{1,\text{VAR}}(h)$ rather than the LP-estimated impulse response $\nu'\hat{\beta}_1(h)$. This is because the former estimate is the pseudo-true parameter in the recursive bootstrap DGP, and the latter estimate differs from the former by an amount that is not asymptotically negligible.

[19] It is not valid to use the Efron bootstrap confidence interval based on the bootstrap quantiles of $\hat{\beta}(h)^*$. This is because the bootstrap samples are asymptotically centered around $\hat{\beta}_{\text{VAR}}(h)$, not $\hat{\beta}(h)$.

[20] This is the bootstrap carried out by Stata's `bootstrap` command with standard settings.

Our online code repository implements the above recommended bootstrap procedure, as well as several alternative LP- and VAR-based procedures, see Footnote 2.

6 Conclusion and Directions for Future Research

Local projection inference is already popular in the applied macroeconomics literature. The simple nature of local projections has allowed the methods of causal analysis in macroeconomics to connect with the rich toolkit for program evaluation in applied microeconomics; see for example Angrist et al. (2018), Nakamura and Steinsson (2018), Stock and Watson (2018), and Rambachan and Shephard (2019). We hope the novel results in this paper on the statistical properties of local projections may further this convergence.

Recommendations for Applied Practice. The simplicity and statistical robustness of *lag-augmented* local projection inference makes it an attractive option relative to existing inference procedures. We recommend that applied researchers conduct inference based on lag-augmented local projections with heteroskedasticity-robust (Eicker-Huber-White) standard errors. This procedure can be implemented using any regression software and has desirable theoretical properties relative to textbook delta method autoregressive inference and to non-augmented local projection methods. In particular, we showed that confidence intervals based on lag-augmented local projections that use robust standard errors with standard normal critical values are uniformly valid over the persistence in the data and for a wide range of horizons. We also suggested a simple bootstrap implementation in Section 5, which seems to achieve even better finite-sample performance.

Conventional VAR-based procedures deliver smaller standard errors than local projections in many cases, but this comes at the cost of fragile coverage, especially at longer horizons. In our opinion, there are only two cases in which the lag-augmented local projection inference method is inferior to competitors: (i) If the data is known to be at most moderately persistent and interest centers on very short impulse response horizons, in which case textbook VAR inference is valid and efficient. (ii) When the data has (near-)unit roots and interest centers on horizons that are a substantial fraction of the sample size, in which case the computationally demanding AR grid bootstrap may be deployed if feasible (Hansen, 1999; Mikusheva, 2012). In all other cases, lag-augmented local projection inference appears to achieve a competitive trade-off between robustness and efficiency.

How should the VAR lag length p be chosen in practice? Naive pre-testing for p causes

uniformity issues for subsequent inference (Leeb and Pötscher, 2005). Though we leave the development of a formal procedure for future research (see below), our theoretical analysis yields three insights. First, users of local projection should worry about the choice of p in order to obtain robust inference, just as users of VAR methods do. Second, p should be chosen conservatively, as is conventional in VAR analysis (Kilian and Lütkepohl, 2017, Chapter 2.6.5). In our framework there is no asymptotic efficiency cost of controlling for more than p_0 lags if the true model is a VAR(p_0), and the simulation results in Supplemental Appendix D confirm that the cost is also small in finite samples. Third, the logic of Section 2.1 suggests that in realistic models where the higher-lag VAR coefficients are relatively small, it is not crucial to get p exactly right: What matters is that we include enough control variables so that the effective regressor of interest approximately satisfies the conditional mean independence condition (Assumption 1).

Directions for Future Research. It would be interesting to relax the assumption of a finite lag length p by adopting a VAR(∞) framework. We are not aware of existing work on uniform inference in such settings. One possibility would be to base inference on a sieve VAR framework that lets the lag length used for estimation tend to infinity at an appropriate rate as in Gonçalves and Kilian (2007). A second possibility is to impose *a priori* bounds on the rate of decay of the VAR coefficients, and then take the resulting worst-case bias of finite-p local projection estimators into account when constructing confidence intervals (as in the "honest inference" approach of Armstrong and Kolesar, 2018).

Due to space constraints, we leave a proof of the validity of the suggested bootstrap strategy to future work. It appears straight-forward, albeit tedious, to prove its pointwise validity. Proving uniform validity requires extending the already lengthy proof of Proposition 1.

Several extensions of the results in this paper could be pursued by adopting techniques from the VAR literature. First, the results of Plagborg-Møller and Wolf (2020) suggest straight-forward ways to generalize our results on reduced-form impulse response inference to *structural* inference. Second, our assumption of no deterministic dynamics in the VAR model could presumably be relaxed using standard arguments. Third, by considering linear system estimators rather than single-equation OLS, our results on scalar inference could be extended to simultaneous inference on several impulses (Inoue and Kilian, 2016; Montiel Olea and Plagborg-Møller, 2019). Finally, whereas we adopt a frequentist perspective in this paper, it remains an open question whether local projection inference is relevant from a Bayesian perspective.

A Proof of Proposition 1

NOTATION. We first introduce some additional notation. For $p \geq 1$, the *companion matrix* of the VAR(p) model (9) is the $np \times np$ matrix given by

$$\mathbf{A} = \begin{bmatrix} A_1 & A_2 & \dots & A_{p-1} & A_p \\ I_n & 0 & \dots & 0 & 0 \\ 0 & I_n & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & I_n & 0 \end{bmatrix}, \tag{14}$$

where $A_1, \dots, A_p$ are the slope coefficients of the autoregressive model (Kilian and Lütkepohl, 2017, p. 25). The companion matrix of a VAR with no lags is defined as a the $n \times n$ matrix of zeros.

Recall that $\|M\| \equiv \sqrt{\text{trace}(M'M)}$ denotes the Frobenius norm of the matrix M. This norm is sub-multiplicative: $\|M_1 M_2\| \leq \|M_1\| \times \|M_2\|$. We use $\lambda_{\min}(M)$ to denote the smallest eigenvalue of the symmetric positive semidefinite matrix M.

Denote $\Sigma \equiv E(u_t u_t')$, and note that this matrix is positive definite by Assumption 2(i). Define, for any collection of autoregressive coefficients A, for any $h \in \mathbb{N}$, and for an arbitrary vector $w \in \mathbb{R}^n$:

$$v(A, h, w) \equiv \{E[\xi_{1,t}(A, h)^2 (w' u_t)^2]\}^{1/2}, \tag{15}$$

where

$$\xi_{i,t}(A, h) \equiv \sum_{\ell=1}^{h} \beta_i(A, h-\ell)' u_{t+\ell}, \quad i = 1, \dots, n. \tag{16}$$

The $n \times 1$ vector $\beta_i(A, h)$ contains each of variable i's impulse response coefficients at horizon $h \geq 1$:

$$\beta_i(A, h)' \equiv e_i(n)' J \mathbf{A}^h J', \tag{17}$$

where $J \equiv [I_n, 0_{n \times n(p-1)}]$ and $e_i(n)$ is the i-th column of the identity matrix of dimension n.

Finally, recall the notation $\rho_i(A)$, $g(\rho, h)$, $\rho_i^*(A, \epsilon)$, and $G(A, h, \epsilon)$ introduced in Section 4.2.

In the proofs below we simplify notation by omitting the subscript A (which indexes the data generating process) from expectations, variances, covariances, and so on.

PROOF. We have defined the lag-augmented local projection estimator of $\beta_1(A,h)$ as the vector of coefficients on y_t in the regression of $y_{1,t+h}$ on y_t with controls $X_t \equiv (y'_{t-1},\ldots,y'_{t-p})$. By the Frisch-Waugh theorem, we can also obtain the coefficient of interest by regressing $y_{1,t+h}$ on the VAR residuals:

$$\hat{\beta}_1(h) \equiv \left(\sum_{t=1}^{T-h} \hat{u}_t(h)\hat{u}_t(h)'\right)^{-1} \sum_{t=1}^{T-h} \hat{u}_t(h) y_{1,t+h}, \tag{18}$$

where we recall the definitions

$$\hat{u}_t(h) \equiv y_t - \hat{A}(h)X_t, \quad \hat{A}(h) \equiv \left(\sum_{t=1}^{T-h} y_t X_t'\right)\left(\sum_{t=1}^{T-h} X_t X_t'\right)^{-1}.$$

Recall also from (10) that

$$\begin{aligned} y_{1,t+h} &= \beta_1(h,A)'y_t + \sum_{\ell=1}^{p-1} \delta_{1,\ell}(A,h)'y_{t-\ell} + \xi_{1,t}(A,h) \\ &= \beta_1(h,A)'y_t + \gamma_1(A,h)'X_t + \xi_{1,t}(A,h) \\ &\quad \text{(where the last } n \text{ entries of } \gamma_1(A,h) \text{ are zero)} \\ &= \beta_1(h,A)'(y_t - AX_t) + \underbrace{(\beta_1(h,A)'A + \gamma_1(A,h)')}_{\equiv \eta_1(A,h)'} X_t + \xi_{1,t}(A,h). \end{aligned} \tag{19}$$

Using the definition (18) of the lag-augmented local projection estimator, we have

$$\begin{aligned} \hat{\beta}_1(h) &= \left(\sum_{t=1}^{T-h} \hat{u}_t(h)\hat{u}_t(h)'\right)^{-1} \sum_{t=1}^{T-h} \hat{u}_t(h) y_{1,t+h} \\ &= \left(\sum_{t=1}^{T-h} \hat{u}_t(h)\hat{u}_t(h)'\right)^{-1} \sum_{t=1}^{T-h} \hat{u}_t(h)[u_t'\beta_1(A,h) + X_t'\eta_1(A,h) + \xi_{1,t}(A,h)] \\ &\quad \text{(by equation (19))} \\ &= \left(\sum_{t=1}^{T-h} \hat{u}_t(h)\hat{u}_t(h)'\right)^{-1} \sum_{t=1}^{T-h} \hat{u}_t(h)[u_t'\beta_1(\rho,h) + \xi_{1,t}(A,h)] \\ &\quad \text{(because } \textstyle\sum_{t=1}^{T-h} \hat{u}_t(h)X_t' = 0 \text{ by definition of } \hat{u}_t(h)) \\ &= \beta_1(A,h) + \left(\sum_{t=1}^{T-h} \hat{u}_t(h)\hat{u}_t(h)'\right)^{-1} \sum_{t=1}^{T-h} \hat{u}_t(h)[(u_t - \hat{u}_t(h))'\beta_1(A,h) + \xi_{1,t}(A,h)] \\ &= \beta_1(A,h) + \left(\sum_{t=1}^{T-h} \hat{u}_t(h)\hat{u}_t(h)'\right)^{-1} \sum_{t=1}^{T-h} \hat{u}_t(h)\xi_{1,t}(A,h), \end{aligned}$$

where the last equality uses $u_t - \hat{u}_t(h) = (\hat{A}(h) - A)X_t$ and again $\sum_{t=1}^{T-h} \hat{u}_t(h)X_t' = 0$ by definition of $\hat{u}_t(h)$. Define $\hat{\nu}(h) \equiv \hat{\Sigma}(h)^{-1}\nu$ and $\tilde{\nu} \equiv \Sigma^{-1}\nu$. Then

$$\begin{aligned}\frac{\nu'[\hat{\beta}_1(h) - \beta_1(A,h)]}{\hat{s}_1(h,\nu)} &= \frac{\hat{\nu}(h)' \sum_{t=1}^{T-h} \hat{u}_t(h)\xi_{1,t}(A,h)}{(T-h)\hat{s}_1(h,\nu)} \\ &= \left(\frac{\hat{\nu}(h)' \sum_{t=1}^{T-h} \xi_{1,t}(A,h)u_t}{(T-h)^{1/2}v(A,h,\tilde{\nu})} + \frac{\hat{\nu}(h)' \sum_{t=1}^{T-h} [\hat{u}_t(h) - u_t]\xi_{1,t}(A,h)}{(T-h)^{1/2}v(A,h,\tilde{\nu})} \right) \\ &\quad \times \frac{v(A,h,\tilde{\nu})}{(T-h)^{1/2}\hat{s}_1(h,\nu)}.\end{aligned}$$

Using the drifting parameter sequence approach of Andrews et al. (2019), both statements (i) and (ii) of the proposition follow if we can show the following: For any sequence $\{A_T\}$ of autoregressive coefficients in $\mathcal{A}(0, C, \epsilon)$, and for any sequence $\{h_T\}$ of nonnegative integers satisfying $h_T \leq (1-a)T$ for all T and $g(\max_i\{|\rho_i(A)|\}, h_T)^2/(T - h_T) \to 0$, we have:

i) $\frac{\sum_{t=1}^{T-h_T} \xi_{1,t}(A_T,h_T)(w'u_t)}{(T-h_T)^{1/2}v(A_T,h_T,w)} \xrightarrow[P_{A_T}]{d} N(0,1)$, for any $w \in \mathbb{R}^n \backslash \{0\}$.

ii) $\frac{(T-h_T)^{1/2}\hat{s}_1(h_T,\nu)}{v(A_T,h_T,\tilde{\nu})} \xrightarrow[P_{A_T}]{p} 1$.

iii) $\frac{\sum_{t=1}^{T-h} [\hat{u}_t(h) - u_t]\xi_{1,t}(A,h)}{(T-h_T)^{1/2}v(A_T,h_T,w)} \xrightarrow[P_{A_T}]{p} 0$, for any $w \in \mathbb{R}^n \backslash \{0\}$.

iv) $\hat{\nu}(h_T) \xrightarrow[P_{A_T}]{p} \nu$.

Result (i) follows from Lemma A.1 below. Result (ii) follows from Lemma A.2 below. Result (iii) follows by bounding

$$\begin{aligned}&\frac{\left\| \sum_{t=1}^{T-h_T} \xi_{1,t}(A_T,h_T)[\hat{u}_t(h_T) - u_t] \right\|}{(T-h_T)^{1/2}v(A_T,h_T,w)} \\ &\leq (T-h_T)^{1/2} \left\| [\hat{A}(h_T) - A_T]G(A_T, T-h_T, \epsilon) \right\| \\ &\quad \times \left\| \frac{\sum_{t=1}^{T-h_T} G(A_T, T-h_T, \epsilon)^{-1} X_t \xi_{1,t}(A_T,h_T)}{(T-h_T)v(A_T,h_T,w)} \right\|.\end{aligned}$$

The first factor on the right-hand side above is $O_{P_{A_T}}(1)$ by Lemma A.3(iii) below. The second factor on the right-hand side above tends to zero in probability by Lemma A.4 below. Thus, result (iii) follows.

Finally, result (iv) follows immediately from Lemma A.5 below and the fact that Σ is positive definite by Assumption 2(i). □

Lemma A.1 (Central limit theorem for $\xi_{i,t}(A,h)(w'u_t)$)**.** *Let Assumptions 1 and 2 hold. Let $i = 1, \ldots, n$. Let $\{A_T\}$ be a sequence of autoregressive coefficients in the parameter space $\mathcal{A}(0, \epsilon, C)$, and let $\{h_T\}$ be a sequence of nonnegative integers satisfying $T - h_T \to \infty$ and $g(\rho_i(A), h_T)^2/(T - h_T) \to 0$. Then*

$$\frac{\sum_{t=1}^{T-h_T} \xi_{i,t}(A_T, h_T)(w'u_t)}{(T-h_T)^{1/2} v(A_T, h_T, w)} \xrightarrow[P_{A_T}]{d} N(0,1),$$

for any $w \in \mathbb{R}^n \backslash \{0\}$.

Proof. The definition of the multi-step forecast error implies

$$\sum_{t=1}^{T-h_T} \xi_{i,t}(A_T, h_T)(w'u_t) = \sum_{t=1}^{T-h_T} \left(\beta_i(A_T, h_T - 1)'u_{t+1} + \ldots + \beta_i(A_T, 0)'u_{t+h_T}\right)(w'u_t). \quad (20)$$

The summands above do not form a martingale difference sequence with respect to a conventionally defined filtration of the form $\sigma(u_{t+h_T}, u_{t+h_T-1}, u_{t+h_T-2}, \ldots)$, even if $\{u_t\}$ is i.i.d. Instead, we will define a process that "reverses time". For any T and any time period $1 \le t \le T - h_T$, define the triangular array and filtration

$$\chi_{T,t} = \frac{\xi_{i,T-h_T+1-t}(A_T, h_T)(w'u_{T-h_T+1-t})}{(T-h)^{1/2} v(A_T, h_T, w),}$$
$$\mathcal{F}_{T,t} = \sigma(u_{T-h_T+1-t}, u_{T-h_T+2-t}, \ldots).$$

We say that we have reversed time because $\chi_{T,1}$ corresponds to the (scaled) last term that appears in the summation (20); the term $\chi_{T,2}$ to the second-to-last term, and so on. By reversing time we have achieved three things. First, the sequence of σ-algebras is a *filtration*:

$$\mathcal{F}_{T,1} \subseteq \mathcal{F}_{T,2} \subseteq \ldots \subseteq \mathcal{F}_{T,T-h_T}.$$

Second, the process $\{\chi_{T,t}\}$ is adapted to the filtration $\{\mathcal{F}_{T,t}\}$, as $\chi_{T,t}$ is measurable with respect to $\mathcal{F}_{T,t}$ for all t. Third, the pair $\{\chi_{T,t}, \mathcal{F}_{T,t}\}$ form a martingale difference array:

$$\begin{aligned} E[\chi_{T,t} \mid \mathcal{F}_{T,t-1}] &\propto E[(\beta_i(A_T, h_T - 1)'u_{T-h_T+2-t} \ldots + \beta_i(A_T, 0)'u_{T+1-t})(w'u_{T-h_T+1-t}) \\ &\qquad \mid u_{T-h_T+2-t}, u_{T-h_T+3-t}, \ldots] \\ &= (\beta_i(A_T, h_T - 1)'u_{T-h_T+2-t} \ldots + \beta_i(A_T, 0)'u_{T+1-t}) \\ &\quad \times E[(w'u_{T-h_T+1-t}) \mid u_{T-h_T+2-t}, u_{T-h_T+3-t}, \ldots] \end{aligned}$$

$$= 0,$$

where the last equality follows from Assumption 1.

Thus, we can apply the martingale central limit theorem in Davidson (1994, Thm. 24.3) to show that

$$\sum_{t=1}^{T-h_T} \chi_{T,t} \xrightarrow{d} N(0,1),$$

which is the statement of the lemma. We now verify the conditions of this theorem. First, by definition of $v(A,h,w)$,

$$\sum_{t=1}^{T-h_T} E[\chi_{T,t}^2] = 1.$$

Second, in Lemma A.6 below we show (by means of Chebyshev's inequality)

$$\sum_{t=1}^{T-h_T} \chi_{T,t}^2 = \frac{\sum_{t=1}^{T-h_T} \xi_{i,t}(A_T,h_T)^2 (w'u_t)^2}{(T-h_T)v(A_T,h_T,w)^2} \xrightarrow{p} 1.$$

Finally, we argue that $\max_{1\leq t\leq T-h_T} |\chi_{T,t}(A_T,h_T)| \xrightarrow{p} 0$. By Davidson (1994, Thm. 23.16), it is sufficient to prove that, for arbitrary $c>0$, we have

$$(T-h_T)E\left[\chi_{T,t}^2 \mathbb{1}(|\chi_{T,t}|>c)\right] \to 0.$$

Indeed,

$$\begin{aligned}
&(T-h_T)E\left[\chi_{T,t}^2 \mathbb{1}(|\chi_{T,t}|>c)\right] \\
&\leq (T-h_T)E\left[\chi_{T,t}^2 \mathbb{1}(|\chi_{T,t}|>c) \times \frac{\chi_{T,t}^2}{c^2}\right] \\
&\leq (T-h_T)\frac{E[\chi_{T,t}^4]}{c^2} \\
&= \frac{1}{(T-h_T)c^2} E\left[\left|v(A_T,h_T,w)^{-1}\xi_{i,T-h_T+1-t}(A_T,h_T)(w'u_{T-h_T+1-t})\right|^4\right] \\
&\leq \frac{6E(\|u_t\|^8)}{(T-h_T)\times\delta^2\times\lambda_{\min}(\Sigma)^2\times c^2},
\end{aligned}$$

where the last inequality uses Lemma A.7 below (recall that δ is the constant in Assumption 2(i)). The right-hand side tends to zero as $T-h_T\to\infty$, as required. □

Lemma A.2 (Consistency of standard errors.)**.** *Let Assumptions 1 to 3 hold. Let the sequence $\{A_T\}$ of elements in $\mathcal{A}(0,C,\epsilon)$ and the sequence $\{h_T\}$ of non-negative integers satisfy*

$T - h_T \to \infty$ *and* $g(\max_i\{|\rho_i(A_T)|\}, h_T)^2/(T - h_T) \to \infty$. *Define* $\tilde{\nu} \equiv \Sigma^{-1}\nu$. *Then*

$$\frac{(T - h_T)^{1/2}\hat{s}(h_T, \nu)}{v(A_T, h_T, \tilde{\nu})} \xrightarrow[P_{A_T}]{p} 1.$$

Proof. See Supplemental Appendix E.2. □

Lemma A.3 (Convergence rates of estimators)**.** *Let the conditions of Lemma A.2 hold. Let* $w \in \mathbb{R}^n\backslash\{0\}$. *Then the following statements all hold:*

i) $\frac{\|\hat{\beta}_1(h_T) - \beta_1(A_T, h_T)\|}{v(A_T, h_T, w)} \xrightarrow[P_{A_T}]{p} 0.$

ii) $\frac{\|G(A_T, T - h_T, \epsilon)[\hat{\eta}_1(A_T, h_T) - \eta_1(A_T, h_T)]\|}{v(A_T, h_T, w)} \xrightarrow[P_{A_T}]{p} 0.$

iii) $(T - h_T)^{1/2}\|(\hat{A}(h_T) - A_T)G(A_T, T - h_T, \epsilon)\| = O_{P_{A_T}}(1).$

Proof. See Supplemental Appendix E.3. □

Lemma A.4 (OLS numerator)**.** *Let Assumptions 1 and 2 hold. Let* $\{A_T\}$ *be a sequence of autoregressive coefficients in* $A_T \in \mathcal{A}(0, \epsilon, C)$, *and let* $\{h_T\}$ *be a sequence of nonnegative integers satisfying* $T - h_T \to \infty$ *and* $g(\max_i\{|\rho_i(A)|\}, h_T)^2/T \to 0$. *Then, for any* $w \in \mathbb{R}^n\backslash\{0\}$, $i, j \in \{1, \ldots, n\}$, *and* $r \in \{1, \ldots, p\}$,

$$\frac{\sum_{t=1}^{T-h_T} \xi_{i,t}(A_T, h_T) y_{j,t-r}}{(T - h_T)v(A_T, h_T, w)g(\rho_j^*(A_T, \epsilon), T - h_T)} \xrightarrow[P_{A_T}]{p} 0.$$

Proof. See Supplemental Appendix E.4. □

Lemma A.5 (Consistency of $\hat{\Sigma}(h)$.)**.** *Let Assumptions 1 to 3 hold. Let the sequence* $\{h_T\}$ *of non-negative integers satisfy* $T - h_T \to \infty$. *Then both the following statements hold:*

i) $\frac{1}{T - h_T}\sum_{t=1}^{T-h_T} u_t u_t' \xrightarrow{p} \Sigma.$

ii) Assume the sequence $\{A_T\}$ *in* $\mathcal{A}(0, C, \epsilon)$ *and* $\{h_T\}$ *satisfy* $g(\max_i\{|\rho_i(A_T)|\}, h_T)^2/(T - h_T) \to \infty$. *Then* $\hat{\Sigma}(h_T) - \frac{1}{T - h_T}\sum_{t=1}^{T-h_T} u_t u_t' \xrightarrow[P_{A_T}]{p} 0.$

Proof. See Supplemental Appendix E.5. □

Lemma A.6 (Consistency of the sample variance of $\xi_{i,t}(A_T, h)(w'u_t)$)**.** *Let the conditions of Lemma A.1 hold. Then*

$$\frac{\sum_{t=1}^{T-h_T} \xi_{i,t}(A_T, h_T)^2 (w'u_t)^2}{(T - h_T)v(A_T, h_T, w)^2} \xrightarrow[P_{A_T}]{p} 1.$$

Proof. See Supplemental Appendix E.6. □

Lemma A.7 (Bounds on the fourth moments of $\xi_{i,t}(A,h)(w'u_t)$ and $\xi_{i,t}(A,h)$)**.** *Let Assumption 1 and Assumption 2(i) hold. Then*

$$E\left[\left(v(A,h,a)^{-1}\xi_{i,t}(A,h)(w'u_t)\right)^4\right] \leq \frac{6E(\|u_t\|^8)}{\delta^2\lambda_{\min}(\Sigma)^2}$$

and

$$E\left[\left(v(A,h,w)^{-1}\xi_{i,t}(A,h)\right)^4\right] \leq \frac{6E(\|u_t\|^4)}{\delta^2\lambda_{\min}(\Sigma)^2\|w\|^4}$$

for all $h \in \mathbb{N}$, $A \in \mathcal{A}(0,\epsilon,C)$, *and* $w \in \mathbb{R}^n\backslash\{0\}$.

Proof. See Supplemental Appendix E.7. □

B Comparison of Inference Procedures

B.1 AR(1) Simulation Study: ARCH Innovations

Consider the AR(1) model (1) with innovations u_t that follow an ARCH(1) process

$$u_t = \tau_t\varepsilon_t, \quad \tau_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2, \quad \varepsilon_t \overset{i.i.d.}{\sim} N(0,1). \tag{21}$$

These innovations satisfy Assumption 1. In our simulations, we set $\alpha_1 = .7$ and $\alpha_0 = (1-\alpha_1)$.[21] Table 2 presents the results, which are qualitatively similar to the i.i.d. case discussed in Section 2.2.

B.2 Analytical Results

In this subsection we provide details on the relative efficiency of lag-augmented LP versus other procedures. Throughout, we focus on the tractable AR(1) model in Section 2.

B.2.1 Relative Efficiency of Lag-Augmented LP

Here we compare the efficiency of lag-augmented LP relative to (i) non-augmented AR, (ii) lag-augmented AR, and (iii) non-augmented LP. We restrict attention to a stationary, homoskedastic AR(1) model and to a fixed impulse response horizon h.

[21] This value of α_0 ensures $\mathbb{E}[\tau_t^2] = 1$.

Table 2: Monte Carlo results: ARCH innovations

	Coverage						Median length					
h	LP-LA$_b$	LP-LA	LP$_b$	LP	AR-LA$_b$	AR	LP-LA$_b$	LP-LA	LP$_b$	LP	AR-LA$_b$	AR
						$\rho = 0.00$						
1	0.892	0.861	0.916	0.804	0.831	0.868	0.386	0.356	0.406	0.316	0.337	0.360
6	0.913	0.903	0.910	0.865	0.000	1.000	0.211	0.207	0.218	0.195	0.000	0.000
12	0.901	0.895	0.896	0.874	0.000	1.000	0.209	0.205	0.211	0.204	0.000	0.000
36	0.903	0.894	0.899	0.890	0.000	1.000	0.222	0.217	0.221	0.215	0.000	0.000
60	0.899	0.889	0.904	0.887	0.000	0.991	0.236	0.229	0.237	0.231	0.000	0.000
						$\rho = 0.50$						
1	0.891	0.865	0.908	0.806	0.836	0.874	0.387	0.357	0.330	0.257	0.336	0.294
6	0.900	0.892	0.908	0.843	0.837	0.776	0.246	0.238	0.272	0.232	0.090	0.048
12	0.904	0.895	0.896	0.879	0.837	0.689	0.240	0.233	0.265	0.250	0.008	0.001
36	0.897	0.887	0.894	0.869	0.837	0.579	0.254	0.246	0.277	0.265	0.000	0.000
60	0.901	0.885	0.902	0.879	0.837	0.540	0.273	0.262	0.300	0.283	0.000	0.000
						$\rho = 0.95$						
1	0.897	0.859	0.823	0.824	0.838	0.856	0.392	0.359	0.084	0.079	0.335	0.086
6	0.896	0.819	0.854	0.788	0.838	0.806	0.621	0.519	0.381	0.327	1.746	0.355
12	0.880	0.785	0.850	0.747	0.838	0.758	0.724	0.560	0.604	0.489	3.942	0.467
36	0.869	0.788	0.859	0.667	0.838	0.643	0.717	0.596	0.816	0.596	64.319	0.291
60	0.881	0.825	0.885	0.692	0.838	0.579	0.711	0.615	0.900	0.625	1032.604	0.095
						$\rho = 1.00$						
1	0.896	0.860	0.841	0.579	0.839	0.560	0.386	0.356	0.040	0.041	0.330	0.041
6	0.879	0.759	0.859	0.543	0.839	0.513	0.686	0.585	0.240	0.228	2.035	0.223
12	0.854	0.662	0.845	0.454	0.839	0.468	0.902	0.715	0.473	0.396	5.510	0.391
36	0.731	0.424	0.752	0.213	0.839	0.352	1.384	0.935	1.170	0.609	177.260	0.669
60	0.640	0.279	0.697	0.164	0.839	0.294	1.475	0.964	1.647	0.642	5593.663	0.729

Coverage probability and median length of nominal 90% confidence intervals at different horizons. AR(1) model with $\rho \in \{0, .5, .95, 1\}$, $T = 240$, innovations as in equation (21). 5,000 Monte Carlo repetitions; 2,000 bootstrap iterations.

Specifically, we here assume the AR(1) model (1) with $\rho \in (-1, 1)$ and where the innovations u_t are assumed to be i.i.d. with variance σ^2. This provides useful intuition, even though the main purpose of this paper is to develop methods that work in empirically realistic settings with several variables/lags, high persistence, and longer horizons.

Comparison with non-augmented AR. In a stationary and homoskedastic AR(1) model, the non-augmented AR estimator is the asymptotically efficient estimator among all regular estimators that are consistent also under heteroskedasticity. This follows from standard semiparametric efficiency arguments, since the non-augmented AR estimator simply plugs the semiparametrically efficient OLS estimator of ρ into the smooth impulse response transformation ρ^h. In particular, non-augmented AR is weakly more efficient than (i) lag-augmented AR, (ii) non-augmented LP, and (iii) lag-augmented LP. As we have discussed in Section 3, however, standard non-augmented AR inference methods perform poorly in situations outside of the benign stationary, short-horizon case.

To gain intuition about the efficiency loss associated with lag augmentation, consider the first horizon $h = 1$. At this horizon, the lag-augmented LP and lag-augmented AR estimators coincide. These estimators regress y_{t+1} on y_t, while controlling for y_{t-1}. As discussed in Section 2.1, this is the same as regressing y_{t+1} directly on the innovation u_t, while controlling for y_{t-1} (which is uncorrelated with u_t). In contrast, the non-augmented AR estimator just regresses y_{t+1} on y_t without controls. Note that (i) the regressor y_t has a higher variance than the regressor u_t, and (ii) the residual in both the augmented and non-augmented regressions equals u_{t+1}. Thus, the usual homoskedastic OLS asymptotic variance formula implies that the non-augmented AR estimator is more efficient than the lag-augmented AR/LP estimator.

Comparison with lag-augmented AR. The relative efficiency of the lag-augmented AR and lag-augmented LP impulse response estimators is ambiguous. In the homoskedastic AR(1) model, the proof of Proposition 1 implies that the asymptotic variance of the lag-augmented LP estimator $\hat{\beta}(h)$ is

$$\text{AsyVar}_\rho(\hat{\beta}(h)) = \frac{E[u_t^2 \xi_t(\rho, h)^2]}{[E(u_t^2)]^2} = \frac{\sigma^2 E[\xi_t(\rho, h)^2]}{\sigma^4} = \frac{\sigma^2 \sum_{\ell=0}^{h-1} \rho^{2\ell} \sigma^2}{\sigma^4} = \sum_{\ell=0}^{h-1} \rho^{2\ell}. \tag{22}$$

We want to compare this to the asymptotic variance of the plug-in AR estimator $\hat{\beta}_{\text{ARLA}}(h) \equiv \hat{\rho}_{\text{LA}}^h$, where $\hat{\rho}_{\text{LA}}$ is the coefficient estimate on the first lag in a regression with *two* lags (Inoue

and Kilian, 2020). Note that $\hat{\rho}_{\text{LA}} = \hat{\beta}(1)$ by definition. By the delta method, the asymptotic variance of $\hat{\beta}_{\text{ARLA}}(h)$ is given by

$$\text{AsyVar}_\rho(\hat{\beta}_{\text{ARLA}}(h)) = (h\rho^{h-1})^2 \times \text{AsyVar}_\rho(\hat{\rho}_{\text{LA}}) = (h\rho^{h-1})^2 \times \text{AsyVar}_\rho(\hat{\beta}(1)) = (h\rho^{h-1})^2.$$

To rank the LP and ARLA estimators in terms of asymptotic variance, note that

$$\text{AsyVar}_\rho(\hat{\beta}(h)) \leq \text{AsyVar}_\rho(\hat{\beta}_{\text{ARLA}}(h)) \iff \sum_{\ell=0}^{h-1} \rho^{2(\ell-h+1)} \leq h^2 \iff \sum_{m=0}^{h-1} \rho^{-2m} \leq h^2.$$

Consider the inequality on the far right of the above display. For $h \geq 2$, the left-hand side is monotonically decreasing from ∞ to h as $|\rho|$ goes from 0 to 1. Hence, there exists an indifference function $\underline{\rho}\colon \mathbb{N} \to (0,1)$ such that

$$\text{AsyVar}_\rho(\hat{\beta}(h)) \leq \text{AsyVar}_\rho(\hat{\beta}_{\text{ARLA}}(h)) \iff |\rho| \geq \underline{\rho}(h).$$

Figure 1 in Section 3 plots the indifference curve between lag-augmented LP standard errors and lag-augmented AR standard errors (lower thick line).

Comparison with non-augmented LP. The non-augmented LP estimator $\hat{\beta}_{\text{LPNA}}(h)$ is obtained from a regression of y_{t+h} on y_t without controls. As is clear from the representation (2), the asymptotic variance of this estimator is given by

$$\begin{aligned}
\text{AsyVar}_\rho(\hat{\beta}_{\text{LPNA}}(h)) &= \frac{\sum_{\ell=-\infty}^{\infty} E[y_t \xi_t(\rho,h) y_{t-\ell} \xi_{t-\ell}(\rho,h)]}{[E(y_t^2)]^2} \\
&= \frac{\sum_{\ell=-h+1}^{h-1} \rho^{|\ell|} E[y_{t-|\ell|}^2] E[\xi_t(\rho,h)\xi_{t-|\ell|}(\rho,h)]}{[E(y_t^2)]^2} \\
&= \frac{\sum_{\ell=-h+1}^{h-1} \sum_{m=|\ell|}^{h-1} \rho^{2m}}{E(y_t^2)/\sigma^2} = (1-\rho^2) \sum_{\ell=-h+1}^{h-1} \sum_{m=|\ell|}^{h-1} \rho^{2m} \\
&= \sum_{\ell=-h+1}^{h-1} (\rho^{2|\ell|} - \rho^{2h}) = \sum_{\ell=0}^{h-1} \rho^{2\ell} + \sum_{\ell=1}^{h-1} \rho^{2\ell} - (2h-1)\rho^{2h}.
\end{aligned}$$

Thus, using (22), we find that

$$\text{AsyVar}_\rho(\hat{\beta}(h)) \leq \text{AsyVar}_\rho(\hat{\beta}_{\text{LPNA}}(h)) \iff \sum_{\ell=1}^{h-1} \rho^{2\ell} \geq (2h-1)\rho^{2h} \iff \sum_{\ell=1}^{h-1} \rho^{-2\ell} \geq (2h-1).$$

The last equivalence assumes $\rho \neq 0$, since lag-augmented and non-augmented LP are clearly equally efficient when $\rho = 0$. For $h = 1$, the last inequality above is never satisfied. This is because at this horizon lag-augmented and non-augmented LP reduce to lag-augmented and non-augmented AR, respectively, and the latter is more efficient, as discussed previously. For $h \geq 2$, the left-hand side of the last inequality above decreases monotonically from ∞ to $h - 1$ as $|\rho|$ goes from 0 to 1. Thus, there exists an indifference function $\overline{\rho}\colon \mathbb{N} \to (0, 1)$ such that

$$\text{AsyVar}_\rho(\hat{\beta}(h)) \leq \text{AsyVar}_\rho(\hat{\beta}_{\text{LPNA}}(h)) \iff |\rho| \leq \overline{\rho}(h).$$

Figure 1 in Section 3 plots the indifference curve between lag-augmented LP and non-augmented LP (upper thick line).

B.2.2 Length of Lag-Augmented AR Bootstrap Confidence Interval

Here we prove that the lag-augmented AR bootstrap confidence interval of Inoue and Kilian (2020) is very wide asymptotically when the data is persistent and the horizon is moderately long.

Let $Y^T \equiv (y_1, \ldots, y_T)$ denote a sample of size T generated by the AR(1) model (1). Let P_ρ denote the distribution of the data when the autoregressive parameter equals ρ. Let $\hat{\rho}$ denote the lag-augmented autoregressive estimator of the parameter ρ based on the data Y^T (i.e., the first coefficient in an AR(2) regression). Let $\hat{\rho}^*$ be the corresponding lag-augmented autoregressive estimator based on a bootstrap sample. We use $\mathbb{P}^*(\cdot \mid Y^T)$ to denote the distribution of the bootstrap samples conditional on the data.

By the results in Inoue and Kilian (2020) we will assume that (i) $\sqrt{T}(\hat{\rho} - \rho)$ converges uniformly to $\mathcal{N}(0, \omega^2)$ for some $\omega > 0$, and (ii) the law of $\sqrt{T}(\hat{\rho}^* - \hat{\rho}) \mid Y^T$ also converges to $\mathcal{N}(0, \omega^2)$ (in probability).

We consider a sequence of autoregressive parameters $\{\rho_T\}$ approaching unity as $T \to \infty$, and a sequence of horizons $\{h_T\}$ that increases with the sample size. The restrictions on these sequences are as follows:

$$h_T(1 - \rho_T) \to a \in [0, \infty), \tag{23}$$

$$h_T \propto T^\eta, \quad \eta \in [1/2, 1]. \tag{24}$$

For example, these assumptions cover the cases of (i) local-to-unity DGPs $\rho_T = 1 - a/T$, $a \geq 0$, at long horizons $h_T \propto T$, and (ii) not-particularly-local-to-unity DGPs $\rho_T = 1 - a/\sqrt{T}$,

$a > 0$, at medium-long horizons $h_T \propto \sqrt{T}$.

We now derive an expression for the quantiles of the bootstrap distribution of the impulse response estimates. For any $c \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}^*((\hat{\rho}^*)^{h_T} \leq c \mid Y^T) &= \mathbb{P}^*((\hat{\rho}^*)^{h_T} \leq c \text{ and } \hat{\rho}^* \geq 0 \mid Y_T) + o_{P_{\rho_T}}(1), \\ &= \mathbb{P}^*(\sqrt{T}(\hat{\rho}^* - \hat{\rho}) \leq \sqrt{T}(c^{1/h_T} - \hat{\rho}) \mid Y^T) + o_{P_{\rho_T}}(1).\end{aligned}$$

The equation above implies that the α bootstrap quantile of $(\hat{\rho}^*)^{h_T}$ is given by $c^*_\alpha = \hat{c}_\alpha + o_{P_{\rho_T}}$, where

$$\hat{c}_\alpha = \left(\hat{\rho} + \omega z_\alpha / \sqrt{T}\right)^{h_T}, \tag{25}$$

and z_α is the α quantile of the standard normal distribution. Note that

$$\begin{aligned}\log \hat{c}_\alpha &= h_T \left[\log \hat{\rho} + \frac{\omega z_\alpha}{\sqrt{T}\hat{\rho}} + o_{P_{\rho_T}}(T^{-1/2})\right] \\ &\quad (\text{since } \log(x+y) = \log(x) + y/x + o(y/x)) \\ &= h_T \log \rho_T + \frac{\omega}{\rho_T}\frac{h_T}{\sqrt{T}}\left[\frac{\rho_T}{\omega}\sqrt{T}\log\frac{\hat{\rho}}{\rho_T} + z_\alpha + o_{P_{\rho_T}}(1)\right].\end{aligned}$$

By (23), we have $\rho_T \to 1$ and $h_T \log \rho_T \to -a$. Also, the delta method implies

$$\frac{\rho_T}{\omega}\sqrt{T}\log\frac{\hat{\rho}}{\rho_T} \xrightarrow{d} Z \equiv N(0,1).$$

Since $\sqrt{T}/h_T = O(1)$ by (24), we then conclude that

$$\frac{\sqrt{T}}{h_T}(\log \hat{c}_\alpha + a) \xrightarrow{d} \omega(Z + z_\alpha). \tag{26}$$

This convergence in distribution is joint if we consider several quantiles α simultaneously.

The Inoue and Kilian (2020) lag-augmented AR Efron bootstrap confidence interval is given by $[c^*_{\alpha/2}, c^*_{1-\alpha/2}]$, so its length equals $\hat{c}_{1-\alpha/2} - \hat{c}_{\alpha/2} + o_{P_{\rho_T}}(1)$. We now argue that this length does not shrink to zero asymptotically in two separate cases.

Case 1: $h_T = \kappa\sqrt{T}, \kappa \in (0,1]$. In this case the result (26) immediately implies that the length of the Inoue and Kilian (2020) bootstrap confidence interval converges to a non-degenerate random variable asymptotically (though the confidence interval has correct asymptotic coverage). This contrasts with the lag-augmented LP confidence interval, whose

length shrinks to zero in probability asymptotically.

CASE 2: $h_T \propto T^\eta, \eta \in (1/2, 1]$. In this case $h_T/\sqrt{T} \to \infty$. The result (26) then implies that, for any $\zeta > 0$,

$$\begin{aligned} P_{\rho_T}\Big([\zeta, 1/\zeta] \subset [c^*_{\alpha/2}, c^*_{1-\alpha/2}]\Big) &= P_{\rho_T}\Big(\log \hat{c}_{\alpha/2} \leq \log \zeta \text{ and } \log \hat{c}_{1-\alpha/2} \geq \log(1/\zeta)\Big) + o(1) \\ &= P\Big(Z + z_{\alpha/2} < 0 \text{ and } Z + z_{1-\alpha/2} > 0\Big) + o(1) \\ &= 1 - \alpha + o(1). \end{aligned}$$

This means that, though the Efron bootstrap confidence interval of Inoue and Kilian (2020) has correct coverage, it achieves this at the expense of reporting—with probability $(1-\alpha)$—intervals that asymptotically contain any compact subset of the positive real line $(0, \infty)$. A similar argument shows that if we intersect the Inoue and Kilian (2020) confidence interval with the parameter space $[-1, 1]$ for the impulse response, the confidence interval almost equals $[0, 1]$ with probability $1 - \alpha$. In contrast, as long as $\eta < 1$, the lag-augmented LP confidence interval has valid coverage and length that tends to zero in probability.

C Verification of Assumption 3: AR(1) Case

In the notation of Section 2, and setting $\epsilon = 0$ without loss of generality, it suffices to show: Any sequence $\{\rho_T\} \in [-1, 1]$ has a subsequence (which we will also denote by $\{\rho_T\}$ for simplicity) such that the random variable $\max\{T(1 - |\rho_T|), 1\}\frac{1}{T^2}\sum_{t=1}^T y_{t-1}^2$ converges in distribution along $\{P_{\rho_T}\}$ to a random variable that is strictly positive almost surely. By passing to a further subsequence if necessary, we may assume that $\lim_{T\to\infty} T(1-|\rho_T|)$ exists.

CASE 1: $T(1 - |\rho_T|) \to \infty$. We will argue that $\frac{1-|\rho_T|}{T}\sum_{t=1}^T y_{t-1}^2$ converges in probability to a nonzero constant along some subsequence. This follows from three facts. First, $\rho_T \to \tilde{c} \in [-1, 1]$, at least along some subsequence. Second, direct calculation using Assumption 1 shows that $E[\frac{1-\rho_T^2}{T}\sum_{t=1}^T y_{t-1}^2] \to \sigma^2 = E(u_t^2) > 0$. Third, tedious calculations similar to the proof of Lemma A.4 show that $\text{Var}[\frac{1-\rho_T^2}{T}\sum_{t=1}^T y_{t-1}^2] \to 0$.

CASE 2: $T(1 - |\rho_T|) \to c \in [0, \infty)$. By passing to a further subsequence, we may assume $\rho_T \to 1$ (the case $\rho_T \to -1$ can be handled similarly). We impose the additional assumption that, for "local-to-unity" sequences $\{\rho_T\}$ satisfying $T(1 - \rho_T) \to c \in [0, \infty)$, the sequence of

probability measures $\{P_{\rho_T}\}$ is contiguous to the measure P_1 (i.e., with $\rho = 1$). This is known to hold for i.i.d. innovations $\{u_t\}$ whose density satisfy a smoothness condition (Jansson, 2008), and it also allows for certain types of conditional heteroskedasticity (Jeganathan, 1995, Section 4). Under this extra assumption, we now just need to argue that, when $\rho = 1$ is *fixed*, $\frac{1}{T^2}\sum_{t=1}^{T} y_{t-1}^2$ converges in distribution to a continuously distributed random variable concentrated on $(0, \infty)$. But this is a well-known result from the unit root literature (e.g., Hamilton, 1994, Chapter 17.4), since $\{u_t\}$ satisfies a Functional Central Limit Theorem under Assumptions 1 and 2 (Davidson, 1994, Theorem 27.14).

References

Andrews, D. W. K., X. Cheng, and P. Guggenberger (2019): "Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests," *Journal of Econometrics*, forthcoming.

Angrist, J. D., Ò. Jordà, and G. M. Kuersteiner (2018): "Semiparametric Estimates of Monetary Policy Effects: String Theory Revisited," *Journal of Business & Economic Statistics*, 36, 371–387.

Armstrong, T. B. and M. Kolesar (2018): "Optimal Inference in a Class of Regression Models," *Econometrica*, 86, 655–683.

Benkwitz, A., M. H. Neumann, and H. Lütkepohl (2000): "Problems related to confidence intervals for impulse responses of autoregressive processes," *Econometric Reviews*, 19, 69–103.

Breitung, J. and R. Brüggemann (2019): "Projection estimators for structural impulse responses," University of Konstanz Department of Economics Working Paper Series 2019-05.

Brillinger, D. R. (2001): *Time Series: Data Analysis and Theory*, Classics in Applied Mathematics, SIAM.

Brugnolini, L. (2018): "About Local Projection Impulse Response Function Reliability," Manuscript, University of Rome "Tor Vergata".

Chevillon, G. (2017): "Robustness of Multistep Forecasts and Predictive Regressions at Intermediate and Long Horizons," ESSEC Working Paper 1710.

Davidson, J. (1994): *Stochastic Limit Theory: An Introduction for Econometricians*, Advanced Texts in Econometrics, Oxford University Press.

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

Dufour, J.-M., D. Pelletier, and Éric Renault (2006): "Short run and long run causality in time series: inference," *Journal of Econometrics*, 132, 337–362.

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

——— (2007): "Asymptotic and Bootstrap Inference for AR(∞) Processes with Conditional Heteroskedasticity," *Econometric Reviews*, 26, 609–641.

Gospodinov, N. (2004): "Asymptotic confidence intervals for impulse responses of near-integrated processes," *Econometrics Journal*, 7, 505–527.

Hamilton, J. D. (1994): *Time Series Analysis*, Princeton University Press.

Hansen, B. E. (1999): "The Grid Bootstrap and the Autoregressive Model," *Review of Economics and Statistics*, 81, 594–607.

Herbst, E. P. and B. K. Johannsen (2020): "Bias in Local Projections," Board of Governors of the Federal Reserve System Finance and Economics Discussion Series 2020-010.

Hjalmarsson, E. and T. Kiss (2020): "Long-Run Predictability Tests Are Even Worse Than You Thought," Manuscript.

Inoue, A. and L. Kilian (2002): "Bootstrapping Autoregressive Processes with Possible Unit Roots," *Econometrica*, 70, 377–391.

——— (2016): "Joint confidence sets for structural impulse responses," *Journal of Econometrics*, 192, 421–432.

——— (2020): "The uniform validity of impulse response inference in autoregressions," *Journal of Econometrics*, 215, 450–472.

Jansson, M. (2008): "Semiparametric Power Envelopes for Tests of the Unit Root Hypothesis," *Econometrica*, 76, 1103–1142.

Jeganathan, P. (1995): "Some Aspects of Asymptotic Theory with Applications to Time Series Models," *Econometric Theory*, 11, 818–887.

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

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

Kilian, L. and Y. J. Kim (2011): "How Reliable Are Local Projection Estimators of Impulse Responses?" *Review of Economics and Statistics*, 93, 1460–1466.

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

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

Leeb, H. and B. M. Pötscher (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

Mikusheva, A. (2007): "Uniform Inference in Autoregressive Models," *Econometrica*, 75, 1411–1452.

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

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

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

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

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

Phillips, P. C. B. (1988): "Regression Theory for Near-Integrated Time Series," *Econometrica*, 56, 1021–1043.

——— (1998): "Impulse response and forecast error variance asymptotics in nonstationary VARs," *Journal of Econometrics*, 83, 21–56.

Phillips, P. C. B. and J. H. Lee (2013): "Predictive regression under various degrees of persistence and robust long-horizon regression," *Journal of Econometrics*, 177, 250–264, special issue on "Dynamic Econometric Modeling and Forecasting".

Plagborg-Møller, M. and C. K. Wolf (2020): "Local Projections and VARs Estimate the Same Impulse Responses," *Econometrica*, forthcoming.

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

Rambachan, A. and N. Shephard (2019): "Econometric analysis of potential outcomes time series: instruments, shocks, linearity and the causal response function," ArXiv: 1903.01637.

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

Richardson, M. and J. H. Stock (1989): "Drawing inferences from statistics based on multiyear asset returns," *Journal of Financial Economics*, 25, 323–348.

Sims, C. A., J. H. Stock, and M. W. Watson (1990): "Inference in Linear Time Series Models with Some Unit Roots," *Econometrica*, 58, 113–144.

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

Toda, H. Y. and T. Yamamoto (1995): "Statistical inference in vector autoregressions with possibly integrated processes," *Journal of Econometrics*, 66, 225–250.

Valkanov, R. (2003): "Long-horizon regressions: theoretical results and applications," *Journal of Financial Economics*, 68, 201–232.

Wright, J. H. (2000): "Confidence Intervals for Univariate Impulse Responses With a Near Unit Root," *Journal of Business & Economic Statistics*, 18, 368–373.