Quantitative Economics 14 (2023), 1199–1220 1759-7331/20231199

# Local projections, autocorrelation, and efficiency

Amaze Lusompa
Research Department, Federal Reserve Bank of Kansas City

It is well known that Local Projections (LP) residuals are autocorrelated. Conventional wisdom says that LP have to be estimated by OLS and that GLS is not possible because the autocorrelation process is unknown and/or because the GLS estimator would be inconsistent. I show that the autocorrelation process of LP can be written as a Vector Moving Average (VMA) process of the Wold errors and impulse responses and that autocorrelation can be corrected for using a consistent GLS estimator. Monte Carlo simulations show that estimating LP with GLS can lead to more efficient estimates.

Keywords. Impulse responses, local projections, autocorrelation, GLS.
JEL classification. C32, C36.

## 1. Introduction

Vector Autoregressions (VARs) and Local Projections (LP) are major tools used in empirical macroeconomic analysis, primarily being used for causal analysis and forecasting through the estimation of impulse response functions. The two methods often give different results when applied to the same problem (Ramey (2016)), and the choice of whether to use impulse responses from LP or VARs can be thought of as the bias-variance tradeoff problem with VARs and LP lying on a spectrum of small sample bias variance choices.[1]

It is well known that LP residuals are autocorrelated. Practitioners exclusively estimate LP via OLS (Ramey (2016)). Jordà (2005) argues that since the true data-generating process (DGP) is unknown, Generalized Least Squares (GLS) estimation is not possible. Hansen and Hodrick (1980) claim that direct forecast regressions (LP) cannot be esti-

Amaze Lusompa: amaze.lusompa@kc.frb.org
I thank Regis Barnichon, Bill Branch, Todd Clark, Ivan Jeliazkov, Òscar Jordà, Lutz Kilian, Daniel Lewis, Fabio Milani, Eric Swanson, Jonathan Wright, two anonymous referees, and seminar participants at several venues for helpful comments, discussions, and/or suggestions. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant 1106401, the Federal Reserve Bank of San Francisco's Thomas J. Sargent Dissertation Fellowship, and the Federal Reserve Bank of Boston under the American Economic Association Summer Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation, the Federal Reserve Bank of San Francisco, the Federal Reserve Bank of Boston, the Federal Reserve Bank of Kansas City, or the Federal Reserve System.

[1] In the case of stationary time series, Plagborg-Møller and Wolf (2021) show linear time-invariant VAR($\infty$) and LP($\infty$) estimate the same impulse responses.

 https://doi.org/10.3982/QE1988

mated by GLS because estimates would be inconsistent.[2] I show that under standard time series assumptions, the autocorrelation process can be written as a Vector Moving Average (VMA) process of the Wold errors and impulse responses and can be corrected for using GLS. Consistency and asymptotically normality of the LP GLS estimator is proved, though efficiency of LP GLS relative to LP OLS is not proved uniformly in this paper. Efficiency of LP GLS relative to LP OLS is only shown for the homoskedastic AR(1) case. Monte Carlo simulations for a wide range of models highlight the benefits of LP GLS and give us an idea of where it lies on the VAR-LP frontier.

The paper is outlined as follows: Section 2 contains the core result showing that the autocorrelation process of LP can be written as a VMA process of the Wold errors and impulse responses and illustrates why GLS is possible. Section 3 explains how to estimate LP GLS and presents the asymptotic properties of the estimator. Section 4 discusses some relative efficiency results of LP estimated by OLS versus LP GLS. Section 5 contains Monte Carlo evidence of the small sample properties of LP GLS, and Section 6 concludes. The Online Appendix in the Supplementary Material (Lusompa (2023)) contains most proofs, additional Monte Carlo evidence, discussion of bootstrapping theory and inference with LP GLS, an empirical application to Gertler and Karadi (2015), as well as a "how to" section for the code. Replication materials can be downloaded from the Quantitative Economics website.

*Some notation* $N(\cdot,\cdot)$ is the normal distribution. *plim* is the probability limit, $\xrightarrow{p}$ is converges in probability, and $\xrightarrow{d}$ is converges in distribution. $\xrightarrow{p^*}$ is converges in probability, and $\xrightarrow{d^*}$ is converges in distribution with respect to the bootstrap probability measure. vec is the vector operator and $\otimes$ is the Kronecker product.

## 2. The autocorrelation process and OLS

Section 2.1 discusses how LP work, drawbacks of OLS estimation with LP, and how GLS estimation can improve upon them. Section 2.2 presents the core result: the autocorrelation process of LP can be written as a VMA process of the Wold errors and impulse responses and can be corrected for via GLS.

### 2.1 *LP and OLS*

To illustrate how LP work, take the simple VAR(1) model

$$y_{t+1} = A_1 y_t + \varepsilon_{t+1},$$

where $y_t$ is a demeaned $r \times 1$ vector of endogenous variables and $\varepsilon_t$ is an $r \times 1$ vector white noise process with $E(\varepsilon_t) = 0$ and $\text{var}(\varepsilon_t) = \Sigma$. Assume that the eigenvalues of $A_1$ have moduli less than unity and $A_1 \neq 0$. Iterating forward leads to

$$y_{t+h} = A_1^h y_t + A_1^{h-1}\varepsilon_{t+1} + \cdots + A_1\varepsilon_{t+h-1} + \varepsilon_{t+h}.$$

[2] Hansen and Hodrick (1980) assume strict exogeneity (which neither LP or VARs satisfy) is a necessary condition for GLS. See Hayashi (2000) for a counterexample for why strict exogeneity is not a necessary condition.

To estimate the impulse responses of a VAR, one would estimate $A_1$ from equation (1) and then use the delta method, bootstrapping, or Monte Carlo integration to perform inference on the impulse responses: $\{A_1, \ldots, A_1^h\}$. To estimate impulse responses using LP, one would estimate the impulse responses directly at each horizon with separate regressions

$$\begin{aligned} y_{t+1} &= B_1^{(1)} y_t + e_{t+1}^{(1)}, \\ &\vdots \\ y_{t+h} &= B_1^{(h)} y_t + e_{t+h}^{(h)}, \end{aligned}$$

where $h$ is the horizon, and when the true DGP is a VAR(1), $\{B_1^{(1)}, \ldots, B_1^{(h)}\}$ and $\{A_1, \ldots, A_1^h\}$ are equivalent. Even if the true DGP is not a VAR(1), $B_1^{(1)} = A_1$ because the horizon 1 LP is a VAR. In practice, it is common for more than one lag to be used. A VAR($k$) and the horizon $h$ LP($k$) can be expressed as

$$y_{t+1} = A_1 y_t + \cdots + A_k y_{t-k+1} + \varepsilon_{t+1},$$

and

$$y_{t+h} = B_1^{(h)} y_t + \cdots + B_k^{(h)} y_{t-k+1} + e_{t+h}^{(h)},$$

respectively. Bear in mind that any VAR($k$) can be written as a VAR(1) (companion form), so results and examples involving the VAR(1) can generally be extended to higher-order VARs.

LP have at least two drawbacks. One, because the dependent variable is a lead, a total of $h$ observations are lost from the original sample when estimating projections for horizon $h$. Two, the error terms in LP for horizons greater than 1 are inherently autocorrelated. Assuming the true model is a VAR(1), it is obvious that autocorrelation occurs because the LP residuals follow a VMA($h-1$) process of the residuals in equation (1). That is,

$$e_{t+h}^{(h)} = A_1^{h-1} \varepsilon_{t+1} + \cdots + A_1 \varepsilon_{t+h-1} + \varepsilon_{t+h},$$

or written in terms of LP

$$e_{t+h}^{(h)} = B_1^{(h-1)} \varepsilon_{t+1} + \cdots + B_1^{(1)} \varepsilon_{t+h-1} + \varepsilon_{t+h}.$$

Inference generally proceeds using nonparametric Heteroskedasticity and Autocorrelation Consistent (HAC) standard errors, which will yield asymptotically correct standard errors in the presence of autocorrelation and heteroskedasticity of unknown forms. Autocorrelation can be corrected for explicitly by including $\{\varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$ in the conditioning set of the horizon $h$ LP. Obviously, $\{\varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$ are unobserved and would have to be estimated. This will affect the asymptotic distribution of the estimator and will be formally discussed in Section 3, but for now this issue can be ignored.

One major advantage of correcting for autocorrelation explicitly is that it fixes what I dub the "increasing variance problem." To my knowledge, the increasing variance

TABLE 1. Asymptotic variance of residuals for LP horizons.

| Horizons | 5 | 10 | 20 | 40 |
|---|---|---|---|---|
| LP OLS | 5.7093 | 9.9683 | 17.3036 | 28.2102 |
| LP GLS | 1 | 1 | 1 | 1 |

problem has not been noticed in the literature. If the true model is a VAR(1), then $\text{var}(e_{t+h}^{(h)}) = \sum_{i=0}^{h-1} A_1^i \Sigma A_1^{i\prime} < \infty$, which is increasing in $h$.[3] Macro variables tend to be persistent, so $A_1^i$ may decay slowly leading to the increase in the variance to be pretty persistent as $h$ increases. To illustrate, let the true model be an AR(1) with

$$y_{t+1} = 0.99 y_t + \varepsilon_{t+1},$$

where $\text{var}(\varepsilon_t) = 1$. The $\text{var}(e_{t+h}^{(h)}) = \sum_{i=0}^{h-1} A_1^i \Sigma_\varepsilon A_1^{i\prime} = \sum_{i=0}^{h} 0.99^{2i}$. Table 1 presents the asymptotic variance of the residuals for different horizons when estimated by OLS versus LP estimated with GLS.

Correcting for autocorrelation explicitly is asymptotically more efficient because $\text{var}(\varepsilon_{t+h}) \leq \text{var}(e_{t+h}^{(h)})$, where the equality only binds when $A_1 = 0$. The increasing variance problem cannot only cause standard errors to be larger than they have to be, but the larger variance is one of the reasons why LP impulse responses are sometimes erratic.[4]

### 2.2 *The autocorrelation process of LP*

First, I will show that even when the true DGP is not a VAR, including the horizon 1 LP residuals (or equivalently, VAR residuals), $\{\varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$, in the horizon $h$ conditioning set will eliminate autocorrelation as long as the data follow standard regularity conditions and the horizon 1 LP residuals are uncorrelated. Second, I will show that the autocorrelation process of $e_{t+h}^{(h)}$ is a moving average process with a known structure.

ASSUMPTION 1. *The data $\{y_t\}$ are covariance stationary and purely nondeterministic, with an everywhere nonsingular spectral density matrix, and absolutely summable Wold representation coefficients. So, there exists an invertible Wold representation*

$$y_t = \varepsilon_t + \sum_{i=1}^{\infty} \Theta_i \varepsilon_{t-i}.$$

Assumption 1 implies that by the Wold representation theorem, there exists a linear and time-invariant VMA representation of the uncorrelated one-step ahead fore-

[3]This is a major reason why Kilian and Kim (2011) found that LP had excessive average length relative to the bias-adjusted bootstrap VAR interval in their Monte Carlo simulations. I provide Monte Carlo evidence of this in the Online Appendix.

[4]Obviously, eliminating the increasing variance problem would not prevent erratic behavior of LP impulse responses since they are not restricted.

cast errors $\{\varepsilon_t\}$. Assumption 1 guarantees that the Wold representation can be inverted into a VAR or LP process. It follows that $\varepsilon_t = y_t - \text{Proj}(y_t|y_{t-1}, y_{t-2}, \ldots)$ where $\text{Proj}(y_t|y_{t-1}, y_{t-2}, \ldots)$ is the (population) orthogonal projection of $y_t$ onto $\{y_{t-1}, y_{t-2}, \ldots\}$.

Consider for each horizon $h = 1, 2, \ldots$ the infinite lag linear LP,

$$y_{t+h} = B_1^{(h)} y_t + B_2^{(h)} y_{t-1} + \cdots + e_{t+h}^{(h)}.$$

PROPOSITION 1. *Under Assumption 1, including $\{\varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$ in the conditioning set of the horizon h LP will eliminate autocorrelation in the horizon h LP residuals.*

PROOF. I first show that

$$\begin{aligned}&\text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}, y_t, y_{t-1}, \ldots)\\ &\quad = \text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}, y_{t+h-1}, y_{t+h-2}, \ldots).\end{aligned}$$

From the Wold representation, we know that $\varepsilon_{t+h-1} = y_{t+h-1} - \text{Proj}(y_{t+h-1}|y_{t+h-2}, y_{t+h-3}, \ldots)$, which implies that $\{\varepsilon_{t+h-1}, y_{t+h-1}, y_{t+h-2}, \ldots\}$ are linearly dependent. This implies that $y_{t+h-1}$ can be dropped from $\text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}, y_{t+h-1}, y_{t+h-2}, \ldots)$ since it contains redundant information. Therefore,

$$\begin{aligned}&\text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}, y_{t+h-1}, y_{t+h-2}, \ldots)\\ &\quad = \text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}, y_{t+h-2}, y_{t+h-3}, \ldots).\end{aligned}$$

Similarly, $\varepsilon_{t+h-2} = y_{t+h-2} - \text{Proj}(y_{t+h-2}|y_{t+h-3}, y_{t+h-4}, \ldots)$, which implies that $\{\varepsilon_{t+h-2}, y_{t+h-2}, y_{t+h-3}, \ldots\}$ are linearly dependent. This implies that $y_{t+h-2}$ can be dropped from $\text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}, y_{t+h-2}, y_{t+h-3}, \ldots)$ since it contains redundant information. Therefore,

$$\begin{aligned}&\text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}, y_{t+h-2}, y_{t+h-3}, \ldots)\\ &\quad = \text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}, y_{t+h-3}, y_{t+h-4}, \ldots).\end{aligned}$$

This process is repeated until $y_{t+1}$ is being dropped due to linear dependence yielding

$$\text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}, y_{t+1}, y_t, \ldots) = \text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}, y_t, y_{t-1}, \ldots).$$

Therefore, if Assumption 1 is satisfied and the horizon 1 LP residuals are uncorrelated,

$$\begin{aligned}&\text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}, y_t, y_{t-1}, \ldots)\\ &\quad = \text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}, y_{t+h-1}, y_{t+h-2}, \ldots).\end{aligned}$$

It follows that

$$\begin{aligned}&\big[y_{t+h} - \text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}, y_t, y_{t-1}, \ldots)\big]\\ &\quad \perp \big[y_{t+h-i} - \text{Proj}(y_{t+h-i}|\varepsilon_{t+h-i-1}, \ldots, \varepsilon_{t-i+1}, y_{t-i}, y_{t-i-1}, \ldots)\big] \quad \forall i \geq 1,\end{aligned}$$

where $\perp$ is the orthogonal symbol. □

Therefore, if the data satisfy Assumption 1, autocorrelation can be eliminated in the horizon $h$ LP by including $\{\varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$ in the conditioning set. Of course, if the true model requires only finitely many lags in the LP specification, then the proof above applies to that case as well, since the extraneous lags will all have coefficients of zero in population.

THEOREM 1. *Under Assumption 1, the autocorrelation process of the horizon h LP residuals* $(e_{t+h}^{(h)})$ *is a* VMA$(h-1)$ *process of the Wold errors and impulse responses. That is,* $e_{t+h}^{(h)} = \Theta_{h-1}\varepsilon_{t+1} + \cdots + \Theta_1\varepsilon_{t+h-1} + \varepsilon_{t+h}$.

PROOF. We know from the Wold representation that $\varepsilon_t \perp y_{t-1}, y_{t-2}, \ldots$, hence $\varepsilon_t \perp \varepsilon_s$ for $t \neq s$. Recall that the infinite lag horizon $h$ LP is

$$y_{t+h} = B_1^{(h)} y_t + B_2^{(h)} y_{t-1} + \cdots + e_{t+h}^{(h)} = \text{Proj}(y_{t+h}|y_t, y_{t-1}, \ldots) + e_{t+h}^{(h)}.$$

By Proposition 1, including $\{\varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$ in the conditioning set eliminates autocorrelation, so the horizon $h$ LP can be rewritten as

$$y_{t+h} = \text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}, y_t, y_{t-1}, \ldots) + u_{t+h}^{(h)},$$

where $u_{t+h}^{(h)} = e_{t+h}^{(h)} - \text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}) = e_{t+h}^{(h)} - \text{Proj}(y_{t+h}|\varepsilon_{t+h-1}) - \cdots - \text{Proj}(y_{t+h}|\varepsilon_{t+1})$. The Proj can be broken up additively because $\{\varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$ are orthogonal to each other and to $\{y_t, y_{t-1}, \ldots\}$. By Proposition 1, $u_{t+h}^{(h)}$ is not autocorrelated. By the Wold representation, we know that $\text{Proj}(y_{t+h}|\varepsilon_t) = \Theta_h\varepsilon_t$. This implies that the horizon $h$ LP can be written as

$$y_{t+h} = B_1^{(h+1)} y_t + B_2^{(h+1)} y_{t-1} + \cdots + \Theta_{h-1}\varepsilon_{t+1} + \cdots + \Theta_1\varepsilon_{t+h-1} + u_{t+h}^{(h)},$$

which implies

$$e_{t+h}^{(h)} = \Theta_{h-1}\varepsilon_{t+1} + \cdots + \Theta_1\varepsilon_{t+h-1} + u_{t+h}^{(h)}.$$

Using the same linear dependence arguments as in Proposition 1, it can be shown that

$$\text{Proj}(y_{t+h}|\varepsilon_{t+h-1}, \ldots, \varepsilon_{t+1}, y_t, y_{t-1}, \ldots) = \text{Proj}(y_{t+h}|y_{t+h-1}, y_{t+h-2}, \ldots),$$

which implies that $u_{t+h}^{(h)} = \varepsilon_{t+h}$, in population. As a result, the autocorrelation process of $e_{t+h}^{(h)}$ is a VMA$(h-1)$ process of Wold errors and coefficients. □

Thus, in population, the error process for the horizon $h$ LP can be written as a VMA$(h-1)$ process even if the true model is not a VAR. In population $B_1^{(h)} = \Theta_h$, which implies

$$e_{t+h}^{(h)} = B_1^{(h-1)}\varepsilon_{t+1} + \cdots + B_1^{(1)}\varepsilon_{t+h-1} + \varepsilon_{t+h}.$$

## 3. LP GLS and its properties

Since $e_{t+h}^{(h)}$ can be written as

$$e_{t+h}^{(h)} = B_1^{(h-1)}\varepsilon_{t+1} + \cdots + B_1^{(1)}\varepsilon_{t+h-1} + u_{t+h}^{(h)},$$

GLS can be used to eliminate autocorrelation in LP while avoiding increasing the number of parameters by including $\{\varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$ in the horizon $h$ conditioning set. To understand how, I will first explain what happens when $\{\varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$ is included in the conditioning set. Just like it is impossible to estimate a VAR($\infty$) in practice, one cannot estimate LP with infinite lags since there is insufficient data. In practice, truncated LP are used where the lags are truncated at $k$. The proofs of consistency and asymptotic normality discuss the rate at which $k$ needs to grow with the sample size to ensure consistent estimation of the impulse responses. In practice, $k$, needs to be large enough that the estimated residuals from the horizon 1 LP are uncorrelated, which is what will be assumed for now. From Theorem 1, we know the horizon $h$ LP is

$$y_{t+h} = B_1^{(h)}y_t + \cdots + B_k^{(h)}y_{t-k+1} + B_1^{(h-1)}\varepsilon_{t+1} + \cdots + B_1^{(1)}\varepsilon_{t+h-1} + u_{t+h,k}^{(h)},$$

where $u_{t+h,k}^{(h)}$ is the lag $k$ analogue of $u_{t+h}^{(h)}$. Due to $\{\varepsilon_{t+1}, \ldots, \varepsilon_{t+h-1}\}$ being unobserved, the estimates $\{\hat{\varepsilon}_{t+1,k}, \ldots, \hat{\varepsilon}_{t+h-1,k}\}$ from the horizon 1 LP/VAR with $k$ lags must be used instead. Estimates of the impulse responses are still consistent (will be shown in Theorem 2), however, even if the sample size is large, textbook formulas for GLS standard errors underrepresent uncertainty because $\{\hat{\varepsilon}_{t+1,k}, \ldots, \hat{\varepsilon}_{t+h-1,k}\}$ are generated regressors and because the textbook formulas for GLS assume strict exogeneity is satisfied. In order to do valid inference, one must use formulas that take into account that the generated regressors were estimated, which the textbook GLS estimator does not.[5]

Including $\{\hat{\varepsilon}_{t+1,k}, \ldots, \hat{\varepsilon}_{t+h-1,k}\}$ in the conditioning set increases the number of parameters in each equation in the system by $(h-1)\times r$. If consistent estimates of $\{B_1^{(h-1)}, \ldots, B_1^{(1)}\}$ are obtained in previous horizons, one can do a Feasible GLS (FGLS) transformation. Let $\widetilde{y}_{t+h}^{(h)} = y_{t+h} - \hat{B}_1^{(h-1),GLS}\hat{\varepsilon}_{t+1,k} - \cdots - \hat{B}_1^{(1),OLS}\hat{\varepsilon}_{t+h-1,k}$. Then one can estimate horizon $h$ via the following equation:

$$\widetilde{y}_{t+h}^{(h)} = B_1^{(h)}y_t + \cdots + B_k^{(h)}y_{t-k+1} + \tilde{u}_{t+h,k}^{(h)}.$$

$\widetilde{y}_{t+h}^{(h)}$ is just a FGLS transformation that eliminates the autocorrelation in the LP residuals without having to sacrifice degrees of freedom and $\tilde{u}_{t+h,k}^{(h)}$ is the error term corresponding to this FGLS transformation. If the impulse responses are estimated consistently, then by the continuous mapping theorem, $\widetilde{y}_{t+h}^{(h)}$ converges in probability to the true GLS transformation $y_{t+h}^{(h)} = y_{t+h} - B_1^{(h-1)}\varepsilon_{t+1} - \cdots - B_1^{(1)}\varepsilon_{t+h-1}$ asymptotically. For clarification, LP can be estimated sequentially, horizon by horizon, as follows. First, estimate the horizon 1 LP/VAR

$$y_{t+1} = B_1^{(1)}y_t + \cdots + B_k^{(1)}y_{t-k+1} + \varepsilon_{t+1,k}.$$

[5]In the proof of asymptotic normality of the limiting distribution, it can be seen that the impact of the generated regressors does not disappear asymptotically.

$\hat{B}_1^{(1),OLS}$ and $\hat{\varepsilon}_{t,k}$ are estimates of $B_1^{(1)}$ and $\varepsilon_{t,k}$, respectively. Horizon 2 can be estimated as

$$\tilde{y}_{t+2}^{(2)} = B_1^{(2)} y_t + \cdots + B_k^{(2)} y_{t-k+1} + \tilde{u}_{t+2,k}^{(2)},$$

where $\tilde{y}_{t+2}^{(2)} = y_{t+2} - \hat{B}_1^{(1),OLS}\hat{\varepsilon}_{t+1,k}$, and $\hat{B}_1^{(2),GLS}$ is the FGLS estimate of $B_1^{(2)}$. Horizon 3 can be estimated as

$$\tilde{y}_{t+3}^{(3)} = B_1^{(3)} y_t + \cdots + B_k^{(3)} y_{t-k+1} + \tilde{u}_{t+3,k}^{(3)},$$

where $\tilde{y}_{t+3}^{(3)} = y_{t+3} - \hat{B}_1^{(2),GLS}\hat{\varepsilon}_{t+1,k} - \hat{B}_1^{(1),OLS}\hat{\varepsilon}_{t+2,k}$, and $\hat{B}_1^{(3),GLS}$ is the FGLS estimate of $B_1^{(3)}$, and so on.

A closely related working paper by Breitung and Brüggemann (2023) was developed independently of the present paper and released after earlier drafts of the present paper instead use the transformation $\tilde{y}_{t+h}^{(h)} = y_{t+h} - \hat{\varepsilon}_{t+h,k}$ and include $\{\hat{\varepsilon}_{t+2,k}, \ldots, \hat{\varepsilon}_{t+h-1,k}\}$ as regressors. This correction would also leave the LP residual uncorrelated. They show that in the AR(1) case this correction is asymptotically as efficient as the VAR. However, since the estimator is approximately equivalent to the VAR in finite samples, it is probably the case their estimator exhibits the same issues with truncation bias as the VAR.[6]

The LP GLS estimator has desirable properties. But first, some assumptions need to be introduced.

ASSUMPTION 2. *Let $y_t$ satisfy the Wold representation as presented in Assumption 1. Assume that in addition (i) $\varepsilon_t$ is strictly stationary and ergodic such that $E(\varepsilon_t|\mathcal{F}_{t-1}) = 0$ a.s., where $\mathcal{F}_{t-1} = \sigma(\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots)$ is the sigma field generated by $\{\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$ and (ii) $E(\varepsilon_t\varepsilon_t') = \Sigma$ is positive definite.*

Note that for any Wold representation $\det\{\Theta(z)\} \neq 0$ for $|z| \leq 1$ where $\Theta(z) = \sum_{h=0}^{\infty}\Theta_h z^h$. It follows from Assumption 2 that the Wold representation can be written as an infinite-order VAR representation $y_t = \sum_{j=1}^{\infty} A_j y_{t-j} + \varepsilon_t$, with $\sum_{j=1}^{\infty}\|A_j\| < \infty$ where $\|A_j\|^2 = \text{tr}(A_j'A_j)$ and $A(z) = I_r - \sum_{h=1}^{\infty} A_h z^h = \Theta(z)^{-1}$. By recursive substitution,

$$y_{t+h} = B_1^{(h)} y_t + B_2^{(h)} y_{t-1} + \cdots + \varepsilon_{t+h} + \Theta_1\varepsilon_{t+h-1} + \cdots + \Theta_{h-1}\varepsilon_{t+1},$$

where $B_1^{(h)} = \Theta_h$, $B_j^{(h)} = \Theta_{h-1}A_j + B_{j+1}^{(h-1)}$ for $h \geq 1$ and with $B_{j+1}^{(0)} = 0$; $\Theta_0 = I_r$ for $j \geq 1$. The standard horizon $h$ LP consists of estimating $\Theta_h$ from a least squares estimate of $A_1^h$

[6]The approximate equivalence is to the order $O_p(T^{-1})$. They do not write their proofs for more general stationary processes (i.e., VAR($\infty$)), so it is not clear if their higher-order equivalence argument holds more generally. Though, consistency and asymptotic normality of their estimator should hold in the general case based off of proofs in this paper. Their estimator does not avoid the degrees of freedom issue discussed above since they include the residuals in the conditioning set, but this problem can be fixed as discussed above. Bruns and Lütkepohl (2022) show that with a moving block bootstrap the Breitung and Brüggemann (2023) estimator can have nontrivial coverage distortions relative to the estimator proposed in this paper (they use VAR(1) data generating processes). It may be worth exploring in future research if there is a way to construct a new estimator, which is a combination of the two estimators and is at least weakly better than both.

with truncated regression

$$y_{t+h} = B_1^{(h)} y_t + \cdots + B_k^{(h)} y_{t-k+1} + e_{t+h,k}^{(h)},$$
$$\text{where } e_{t+h,k}^{(h)} = \sum_{j=k+1}^{\infty} B_j^{(h)} y_{t-j+1} + \varepsilon_{t+h} + \sum_{l=1}^{h-1} \Theta_l \varepsilon_{t+h-l}.$$

ASSUMPTION 3. *Let Assumption 2 hold. Assume that, in addition,* (*i*) *the r-dimensional* $\varepsilon_t$ *has absolutely summable cumulants up to order* 8. *That is,*

$$\sum_{i_2=-\infty}^{\infty} \cdots \sum_{i_h=-\infty}^{\infty} \left|\kappa_{\boldsymbol{a}}(0, i_2, \ldots, i_j)\right| < \infty \quad \text{for } j = 2, \ldots, 8,$$

$a_1, \ldots, a_j \in \{1, \ldots, r\}$, $\boldsymbol{a} = \{a_1, \ldots, a_j\}$, *and* $\kappa_{\boldsymbol{a}}(0, i_2, \ldots, i_j)$ *denotes the jth joint cumulant of* $\varepsilon_{0,a_1}, \varepsilon_{i_2,a_2}, \ldots, \varepsilon_{i_j,a_j}$. *In particular, this condition includes the existence of the eight moments of* $\varepsilon$. (*ii*) $L_r E(\operatorname{vec}(\varepsilon_t \varepsilon_{t-j}')\operatorname{vec}(\varepsilon_t \varepsilon_{t-j}')')L_r'$ *is positive definite for all j, and* $L_r$ is a finite $r(r+1)/2 \times r^2$ elimination matrix, which is defined as $L_r \operatorname{vec}(A) = \operatorname{vech}(A)$. (*iii*) $k$ *satisfies* $\frac{k^4}{T} \to 0$; $T, k \to \infty$. (*iv*) $k$ *satisfies* $(T-k-H)^{1/2} \sum_{j=k+1}^{\infty} \|A_j\| \to 0$; $T, k \to \infty$.

THEOREM 2. *Under Assumption 3, the LP GLS estimator is consistent. In particular,*

$$\hat{B}_1^{(h),GLS} \xrightarrow{p} \Theta_h, \quad \text{and more generally} \quad \left\|\hat{B}(k,h,GLS) - B(k,h)\right\| \xrightarrow{p} 0,$$

*where*

$$\underbrace{\hat{B}(k,h,GLS)}_{r \times kr} = \left(\hat{B}_1^{(h),GLS}, \ldots, \hat{B}_k^{(h),GLS}\right) = (T-k-H)^{-1} \sum_{t=k}^{T-h} \tilde{y}_{t+h}^{(h)} X_{t,k}' \hat{\Gamma}_k^{-1},$$
$$\underbrace{B(k,h)}_{r \times kr} = \left(B_1^{(h)}, \ldots, B_k^{(h)}\right), \qquad \underbrace{X_{t,k}}_{kr \times 1} = \left(y_t', \ldots, y_{t-k+1}'\right)',$$
$$\underbrace{\hat{\Gamma}_k}_{kr \times kr} = (T-k)^{-1} \sum_{t=k}^{T} X_{t,k} X_{t,k}', \qquad \underbrace{\Gamma_k}_{kr \times kr} = E\left(X_{t,k} X_{t,k}'\right).$$

REMARK 1. Goncalves and Kilian (2007) use these assumptions to show consistency and asymptotic normality of the VAR($\infty$) when there is conditional heteroskedasticity. These assumptions are more general versions of the ones used by Lewis and Reinsel (1985) and Jordà and Kozicki (2011) who show consistency and asymptotic normality of the VAR($\infty$) and the LP($\infty$), respectively, in the i.i.d. case.[7]

[7] These are sufficient conditions. Some of the proofs can be written under weaker conditions (e.g., Theorem 2 with $\frac{k^2}{T} \to 0$), but for sake of brevity these will suffice.

PROOF. See the Online Appendix. □

As noted earlier, the parameters used in the GLS correction are not known, and their uncertainty must be taken into account in order to do valid inference. To take into account the uncertainty in the generated regressors, one can use bootstrapping, multistep estimation, or joint estimation. The bootstrap estimator is fleshed out in detail in the Online Appendix. Standard errors for the multistep estimator will be discussed next.

Following Brüggemann, Jentsch, and Trenkler (2016), a mixing condition is imposed for structural inference.

ASSUMPTION 4. *Let Assumption 3 hold. Assume that, in addition, $\varepsilon_t$ is strong $(\alpha)$ mixing, with $\alpha(m)$ of size $-4(\nu+1)/\nu$ for some $\nu > 0$, where $\alpha(m) = \sup_{A\in\mathcal{F}_{-\infty}^{0}, B\in\mathcal{F}_{m}^{\infty}} |P(A\cap B) - P(A)P(B)|$ for $m = 1, 2, \ldots$ denote the $\alpha$ mixing process of $\varepsilon_t$ where $\mathcal{F}_{-\infty}^{0} = \sigma(\ldots, \varepsilon_{-2}, \varepsilon_{-1}, \varepsilon_0)$ and $\mathcal{F}_{m}^{\infty} = \sigma(\varepsilon_m, \varepsilon_{m+1}, \ldots)$. Lastly, $V(k,H)$ is positive definite where $H$ is the max horizon and*

$$V(k,H) = \begin{bmatrix} V_{11}(k,H) & V_{12}(k,H) \\ V_{21}(k,H) & V_{22} \end{bmatrix}, \qquad V_{11}(k,H) = \sum_{p=-\infty}^{\infty} \operatorname{cov}\big(Score_{t+H}^{(H)}, Score_{t+H-p}^{(H)}\big),$$

$$V_{22} = L_r' \left\{ \sum_{p=-\infty}^{\infty} E\big(\operatorname{vec}\big(\varepsilon_{t+1}\varepsilon_{t+1}'\big), \operatorname{vec}\big(\varepsilon_{t+1-p}\varepsilon_{t+1-p}'\big)'\big) - \operatorname{vec}(\Sigma)\operatorname{vec}(\Sigma)' \right\} L_r,$$

$$V_{12}(k,H) = V_{21}(k,H)' = \sum_{p=-\infty}^{\infty} \operatorname{cov}\big(Score_{t+H}^{(H)}, \operatorname{vec}\big(\varepsilon_{t+1-p}\varepsilon_{t+1-p}' - \Sigma\big)' L_r\big),$$

$$Score_{t+H}^{(H)} = l(k,H)' \begin{bmatrix} \big(\Gamma_k^{-1} X_{t,k} \otimes I_r\big)\varepsilon_{t+H} + s_{k,H}\big(\Gamma_k^{-1} X_{t,k} \otimes I_r\big)\varepsilon_{t+1} \\ \vdots \\ \big(\Gamma_k^{-1} X_{t,k} \otimes I_r\big)\varepsilon_{t+2} + s_{k,2}\big(\Gamma_k^{-1} X_{t,k} \otimes I_r\big)\varepsilon_{t+1} \\ \big(\Gamma_k^{-1} X_{t,k} \otimes I_r\big)\varepsilon_{t+1} \end{bmatrix},$$

$$s_{k,h} = \left( \sum_{l=1}^{h-1} \big\{ \Gamma_k^{-1} \Gamma_{(h-l-1),k}' \otimes \Theta_l \big\} \right), \quad \Gamma_{(m-n),k} = E\big(X_{m,k} X_{n,k}'\big),$$

and $l(k,H)$ *is a sequence of* $kr^2H \times 1$ *vectors such that* $0 < M_1 \le \|l(k,H)\|^2 \le M_2 < \infty$.

REMARK 2. $l(k,H)$ is simply a Cramer–Wold device, which is used to show that any linear combinations of the parameters that satisfy the condition have asymptotically normal limiting distributions. Assumption 4 is slightly to somewhat stronger than the condition used in Brüggemann, Jentsch, and Trenkler (2016), and is probably a stronger condition than necessary. However, it allows for the use of mixingale inequalities, which in turn allows for a straightforward proof using the mixingale Central Limit Theorem

(CLT). The proof can probably be written using weaker conditions and following a proof similar to Theorem 3.1 in Brüggemann, Jentsch, and Trenkler (2016).

THEOREM 3. *Define* $\hat{\Sigma} = (T-k)^{-1}\sum_{t=k}^{T-H}\hat{\varepsilon}_{t,k}\hat{\varepsilon}_{t,k}'$. *Under Assumption* 4,

$$\left(l(k,H)'\begin{bmatrix}\sqrt{T-k-H}\,\mathrm{vec}\big[\hat{B}(k,H,GLS)-B(k,H)\big]\\ \vdots\\ \sqrt{T-k-H}\,\mathrm{vec}\big[\hat{B}(k,2,GLS)-B(k,2)\big]\\ \sqrt{T-k-H}\,\mathrm{vec}\big[\hat{B}(k,1,OLS)-B(k,1)\big]\end{bmatrix}\right) \xrightarrow{d} N\big(0,V(k,H)\big).$$
$$\sqrt{T-k-H}\,\mathrm{vech}[\hat{\Sigma}-\Sigma]$$

PROOF. See the Online Appendix. □

Note that even though the FGLS procedure makes the residuals asymptotically uncorrelated, the regression score for reduced form impulse responses, $Score_{t+H}^{(H)}$, is still serially correlated. This is because the impact of the estimated residuals used in the FGLS correction does not disappear in the limiting distribution.[8] Fortunately, since the autocorrelation process was explicitly derived, one does not need to choose a HAC procedure and associated difficult-to-interpret tuning parameters to calculate $V_{11}(k,H)$.[9] To see why, note that the regression score can be "rearranged." Define

$$Rscore_{t+1}^{(H)} = l(k,H)'\begin{bmatrix}\big(\Gamma_k^{-1}X_{t-H+1,k}\otimes I_r\big)\varepsilon_{t+1}+s_{k,H}\big(\Gamma_k^{-1}X_{t,k}\otimes I_r\big)\varepsilon_{t+1}\\ \vdots\\ \big(\Gamma_k^{-1}X_{t-1,k}\otimes I_r\big)\varepsilon_{t+1}+s_{k,2}\big(\Gamma_k^{-1}X_{t,k}\otimes I_r\big)\varepsilon_{t+1}\\ \big(\Gamma_k^{-1}X_{t,k}\otimes I_r\big)\varepsilon_{t+1}\end{bmatrix},$$

where $Rscore_{t+1}^{(H)}$ is the rearranged score where the time subscripts of $\varepsilon$ line up. Note that for finite $H$,

$$\sup_{s\in\mathbb{R}}\left|P\left[(T-k-H)^{-1/2}\sum_{t=k}^{T-H}Score_{t+H}^{(H)}\le s\right]-P\left[(T-k-H)^{-1/2}\sum_{t=k}^{T-H}Rscore_{t+1}^{(H)}\le s\right]\right|$$

converges in probability to zero. Due to the martingale difference assumption and the fact that the horizon $h$ LP residuals are only correlated up to $h-1$ horizons, $V_{11}(k,H) = \sum_{p=-\infty}^{\infty}\mathrm{cov}(Score_{t+H}^{(H)}, Score_{t+H-p}^{(H)}) = \sum_{p=-H+1}^{H-1}\mathrm{cov}(Score_{t+H}^{(H)}, Score_{t+H-p}^{(H)}) = \mathrm{var}(Rscore_{t+1}^{(H)})$. Therefore, the limiting distribution in Theorem 3 is not affected by substituting $Rscore_{t+1}^{(H)}$ for $Score_{t+H}^{(H)}$.

If one is only interested in marginal distribution for the impulse responses, note that

$$\sqrt{T-k-H}\,l(k)'\mathrm{vec}\big[\hat{B}(k,h,GLS)-B(k,h)\big]\xrightarrow{d}N\big(0,\Omega(k,h,GLS)\big),$$

[8] If the true errors were known, the score would be uncorrelated, but they obviously have to be estimated.

[9] See Lazarus, Lewis, Stock, and Watson (2018), footnote 1, for a summary of the different types of HAC estimators.

where $l(k)$ is a $kr^2 \times 1$ Cramer–Wold device satisfying $0 < M_1 \leq \|l(k)\|^2 \leq M_2 < \infty$, and

$$\begin{aligned}\Omega(k,h,GLS) = l(k)'\big\{&E\big[\big(\Gamma_k^{-1}X_{t-h+1,k}\otimes I_r\big)\varepsilon_{t+1}\varepsilon_{t+1}'\big(\Gamma_k^{-1}X_{t-h+1,k}\otimes I_r\big)'\big]\\&+E\big[s_{k,h}\big(\Gamma_k^{-1}X_{t,k}\otimes I_r\big)\varepsilon_{t+1}\varepsilon_{t+1}'\big(\Gamma_k^{-1}X_{t,k}\otimes I_r\big)'s_{k,h}'\big]\\&+E\big[\big(\Gamma_k^{-1}X_{t-h+1,k}\otimes I_r\big)\varepsilon_{t+1}\varepsilon_{t+1}'\big(\Gamma_k^{-1}X_{t,k}\otimes I_r\big)'s_{k,h}'\big]\\&+E\big[s_{k,h}\big(\Gamma_k^{-1}X_{t,k}\otimes I_r\big)\varepsilon_{t+1}\varepsilon_{t+1}'\big(\Gamma_k^{-1}X_{t-h+1,k}\otimes I_r\big)'\big]\big\}l(k).\end{aligned}$$

Even though deriving the autocorrelation process makes calculation of $V_{11}(k,H)$ (and hence $\Omega(k,h,GLS)$) simple and straightforward, structural inference is more complicated since a nonparametric or sieve parametric HAC estimators would in general be needed to calculate $V_{12}(k,H) = V_{21}(k,H)'$ and $V_{22}$. This is also true in the VAR case as noted in Brüggemann, Jentsch, and Trenkler (2016).[10] Consistent estimation of $V(k,H)$ combined with the delta method would lead to asymptotically valid joint inference. I do not explore which type of HAC estimators perform best for $V_{12}(k,H) = V_{21}(k,H)'$ and $V_{22}$. I instead propose a block wild bootstrap estimator due to its simplicity (details can be found in the Online Appendix).

## 4. LP GLS and relative efficiency

It is not proved in this paper that LP GLS is uniformly at least as efficient as LP OLS, but to give a sense of potential efficiency gains of estimating LP via GLS, I will compare the asymptotic relative efficiency of the LP GLS estimator and the LP OLS estimator when the true model is a homoskedastic AR(1). Take the simple AR(1) model

$$y_{t+1} = ay_t + \varepsilon_{t+1},$$

where $|a| < 1$, $a \neq 0$, and $\varepsilon_t$ is an i.i.d. error process with $E(\varepsilon_t) = 0$ and $\operatorname{var}(\varepsilon_t) = \sigma^2$. Define $\{b^{(1)},\ldots,b^{(h)}\}$ as the LP impulse responses for the AR(1) model. For simplicity, assume the lag length is known. By Proposition 6 in the Online Appendix, the limiting distribution of the LP GLS impulse response at horizon $h$ is

$$\sqrt{T}\big(\hat{b}^{(h),GLS} - a^h\big) \xrightarrow{d} N\big(0,\big[1+\big(h^2-1\big)a^{2h-2}\big]\big(1-a^2\big)\big).$$

The limiting distribution of the LP OLS impulse response at horizon $h$ is

$$\sqrt{T}\big(\hat{b}^{(h),OLS} - a^h\big) \xrightarrow{d} N\big(0,\big(1-a^2\big)^{-1}\big[1+a^2-\{2h+1\}a^{2h}+\{2h-1\}a^{2h+2}\big]\big)$$

(Bhansali (1997)).

Theorem 4 (FGLS Efficiency). *Assume the true model is an* AR(1) *as specified above. Then*

$$\lim\big(\operatorname{var}\big(\sqrt{T}\big(\hat{b}^{(h),GLS} - a^h\big)\big)\big) \leq \lim\big(\operatorname{var}\big(\sqrt{T}\big(\hat{b}^{(h),OLS} - a^h\big)\big)\big).$$

[10] This is due to the more general assumption of conditional heteroskedasticity for the errors. In the i.i.d. case, a HAC estimator would not be needed for $V_{12}(k,H)$.

TABLE 2. Relative efficiency of LP (GLS) to LP (OLS).

| Autocorrelation Coefficient | Horizons | | | | | |
|---|---|---|---|---|---|---|
| | 3 | 5 | 10 | 20 | 30 | 40 |
| $a = 0.99$ | 0.993 | 0.979 | 0.945 | 0.88 | 0.818 | 0.759 |
| $a = 0.975$ | 0.983 | 0.948 | 0.864 | 0.713 | 0.580 | 0.464 |
| $a = 0.95$ | 0.966 | 0.896 | 0.735 | 0.475 | 0.288 | 0.165 |
| $a = 0.9$ | 0.931 | 0.792 | 0.508 | 0.179 | 0.061 | 0.029 |
| $a = 0.75$ | 0.827 | 0.53 | 0.195 | 0.123 | 0.123 | 0.123 |
| $a = 0.5$ | 0.727 | 0.496 | 0.45 | 0.45 | 0.45 | 0.45 |
| $a = 0.25$ | 0.854 | 0.828 | 0.827 | 0.827 | 0.827 | 0.827 |
| $a = 0.01$ | 1 | 1 | 1 | 1 | 1 | 1 |

PROOF. See the Online Appendix. □

The relative efficiency between the LP GLS and LP impulse responses (given by the ratio of the variances) determines how much more efficient one specification is relative to another. Note that the relative efficiency not only depends on the persistence, $a$, but on the horizon as well. Table 2 presents the relative efficiency between the LP GLS and LP OLS impulse responses for different values of $a$. The gains from LP GLS can be large but they are not necessarily monotonic. This is because if the persistence is not that high, the impulse responses decay to zero quickly making the variance of the impulse responses small, and the gains from correcting for autocorrelation are not as large.

## 5. MONTE CARLO EVIDENCE

In this section, I present Monte Carlo evidence of the finite sample properties of the LP GLS bootstrap and the multistep (analytical) LP GLS estimator. I compare the following six estimators: LP GLS bootstrap (LP GLS Boot), Bias-adjusted LP GLS bootstrap (LP GLS Boot BA), Analytical LP GLS estimator (LP GLS), Analytical VAR estimator (VAR), Bias-adjusted VAR bootstrap (VAR Boot BA), LP OLS with equal-weighted cosine HAC standard errors (Lazarus et al. (2018)) (LP OLS). The abbreviations in the parentheses are what the estimators are referred to in the figures.

In summary, I find that the LP GLS bootstrap estimators minimizes downside risk. The bootstrap VAR had the shortest confidence interval on average, but coverage can vary widely depending on the DGP. LP OLS typically had at least decent coverage, but coverage typically did not exceed that of its GLS counterparts, and it was relatively inefficient compared to the GLS estimators. The analytical LP GLS estimator does not perform as well as the LP GLS bootstraps. The performance of the analytical VAR could vary greatly from one DGP to the next.

Unless stated otherwise, all simulations use a sample size of 250, which is representative of a quarterly data set dating back to 1960.[11] The comprehensive McCracken and

[11]Even though the most prominent macro variables such as GDP, inflation, and unemployment date back to at least 1948, many do not date back that far.

Ng (2016) data set goes back to 1959 for quarterly and monthly data; so a sample size of 250 would be a representative lower bound for practically most quarterly macroeconomic variables of interest.[12] All of the Monte Carlos are generated using normally distributed errors. All of the methods use the same lag length for each simulation. The LP GLS methods requires that the VAR residuals are white noise. Unless stated otherwise, the simulations lag lengths are chosen using a lag length criteria (e.g., AIC, BIC, HQIC) and then the VAR residuals are tested for autocorrelation using the Ljung–Box Q-test. The baseline lag length used is AIC, but results are not sensitive to other choices. If the null of white noise is rejected, a lag is added, the model is reestimated, and the new residuals are tested for autocorrelation. This process is repeated until the null of white noise is not rejected for the VAR residuals. This lag length is then used for all of the estimators.

Simulations were conducted 1000 times, and bootstraps have 5000 replications each. Unless stated otherwise, only the Wold impulse responses are estimated. Coverage and average length for 95% confidence intervals are calculated. That is, for each simulation, I estimated the model for each desired horizon using all of the estimation methods and then check if the 95% confidence intervals contain the true impulse response. I then calculate the probability that the 95% confidence interval contains the true impulse response over the Monte Carlo simulations, which gives me the coverage for each method and horizon. I use Efron percentile confidence intervals for all of the bootstrap estimators. For each simulation draw, I also save the length of the 95% interval for the the different methods for each horizon. The lengths are then averaged over each Monte Carlo simulation for each method and horizon to get the respective average length of the 95% confidence intervals for each method and horizon. Fifteen horizons are analyzed, which would be representative of analyzing 4 years of impulse responses for quarterly data.

Even though LP GLS can improve efficiency relative to LP OLS over a wide range of DGP, this section will focus on situations where truncation bias could be an issue. As highlighted in Plagborg-Møller and Wolf (2021), truncation bias can hamper the performance of a VAR. The data generating processes (DGP) used in the Monte Carlos were chosen to highlight that in situations where the true DGP may not be well approximated by a VAR (e.g., situations where partial autocorrelations of the true DGP decay sufficiently slowly), LP GLS is a viable alternative. The first DGP is the following ARMA(1, 1) from Kilian and Kim (2011),

$$y_{t+1} = 0.9y_t + \varepsilon_{t+1} + m\varepsilon_t,$$

where $m \in \{0, 0.25, 0.5, 0.75\}$, and $\varepsilon_t \sim N(0, 1)$. Though simple, it is easy to control the magnitude of the AR and MA component making it instructive. Select results can be found in Figures 1 and 2. The bias-adjusted VAR bootstrap and analytical VAR perform the best in terms of coverage, but LP GLS bootstraps perform well, with coverage of at least approximately 90% at all horizons. LP OLS has slightly better coverage than the

[12]Even if one is doing IV, the major instruments used in macroeconomics are available for at least 200 observations (Ramey (2016)).

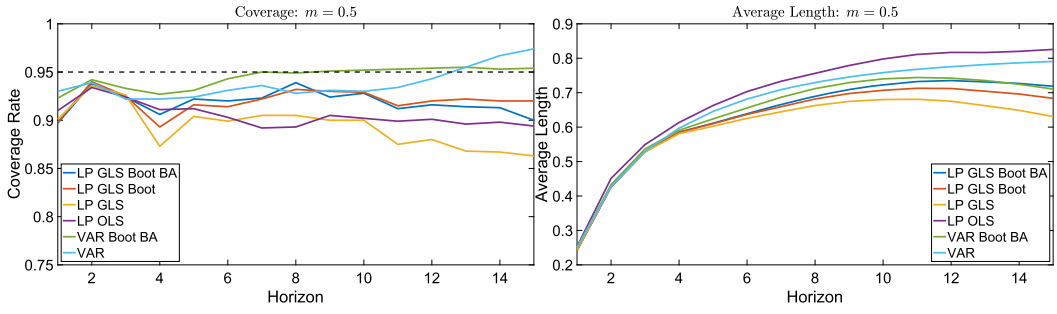


FIGURE 1. Coverage rates for 95% confidence intervals and average length for ARMA(1, 1) models.

analytical LP GLS estimator, but not the bootstraps.[13] The analytical LP GLS estimator is the most efficient, but this is at least partly due to it having the worst coverage. The LP GLS bootstrap estimators are more efficient than the LP OLS estimator with most of the efficiency gains coming at the longer horizons. Moreover, the LP GLS estimators are competitive with the VAR bootstrap in terms of efficiency and are more efficient than the analytical VAR. The analytical VAR has the second largest average length (second to LP OLS).[14]

Though arma models can be pedagogical since they are simple and it is easy to control the magnitude of the AR and MA component, they may not be able to replicate what one may encounter in practice. Therefore, I include some empirically calibrated models. Next, I will present evidence for two empirically calibrated Monte Carlos for a fiscal VAR and technology VAR. But before I get into the details, it is important that I first discuss a potential shortcoming in the way we calibrate empirical models in the impulse response literature. Jordà (2005) and Kilian and Kim (2011) each empirically calibrated a VAR(12) to give an empirically relevant Monte Carlo that would give a gauge of what is happen-

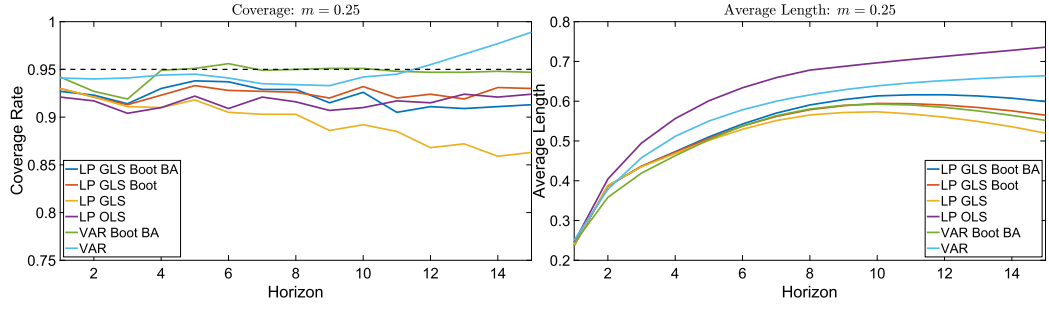


FIGURE 2. Coverage rates for 95% confidence intervals and average length for ARMA(1, 1) models (continued).

[13]As noted by a referee, LP OLS coverage could improve by being bootstrapped as well.

[14]Note that for DGP that do not admit finite lag VAR representations, I use the infinite-order analytic VAR confidence intervals (see Lütkepohl (1990) for details). This choice does not need to be made with the VAR bootstrap (Goncalves and Kilian (2007)).

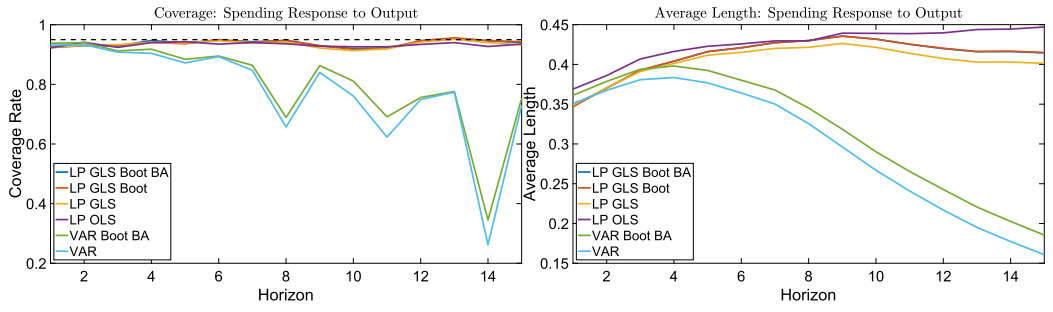


FIGURE 3. Coverage rates for 95% confidence intervals and average length for fiscal VAR.

ing in practice. The problem with empirically calibrated VARs is that if the model used to generate the data suffers from truncation bias, the Monte Carlo simulation may mask the impact truncation bias could have in practice, say, a VAR with truncation bias is used to generate data in a Monte Carlo. If the coverage in that Monte Carlo is great, one might conclude that the truncation bias is not an issue. Obviously, we know truncation bias can be an issue because we know what the true impulse responses are. In practice, when empirically calibrating a Monte Carlo, the results of the Monte Carlo are only as good as the calibration, and in practice we do not know how good the calibration is. To protect against this problem, I estimate empirically calibrated VARs with lag lengths longer than what is typically used in practice in order to protect against truncation bias.[15]

For the quarterly data sets, researchers will typically not include more than 1 or 2 years worth of lags, so I estimate a VAR(16). The first empirically calibrated VAR is a fiscal VAR that includes growth rates of real GDP per capita and real spending per capita, which are the baseline variables used in fiscal multiplier analysis (Ramey and Zubairy (2018)). The data runs from 1947Q2–2019Q4. Select results are presented in Figure 3. There is little to no efficiency gain, if not a slight efficiency loss, from using the LP GLS estimators relative to LP OLS for this DGP. The VAR estimators can be much more efficient than the LP estimators, particularly at longer horizons. However, this is partly due to the VAR estimators having coverage distortions while the LP GLS estimators as well as the LP OLS estimator have approximately nominal coverage throughout. The VAR estimators had coverage drop below 60% for most of impulse responses by horizon 15.

The second empirical Monte Carlo is a technology VAR that includes growth rates of labor productivity, real GDP per capita, real stock prices per capita, and total factor productivity. These are the baseline variables used in Ramey (2016). The data runs from 1947Q3–2015Q2. Again, I estimate a VAR(16) and use that to generate the data. Select results are presented in Figure 4. Similar to the fiscal VAR, there is little to no efficiency gain, if not a slight efficiency loss, from using the LP GLS estimators relative to LP OLS for this DGP. The VAR estimators are more efficient, particularly at longer horizons, but again the VARs have severe coverage distortions, where by horizon 15, most impulse responses had coverage rates drop below 50%. The LP estimators have coverage of at least

[15]Alternatively, I could also generate the model using LP to estimate the Wold coefficients up to $q$ horizons out and generate data based off of that MA($q$). Unfortunately, it would be more cumbersome since it would require more choices in the setup, and hence more robustness checks.

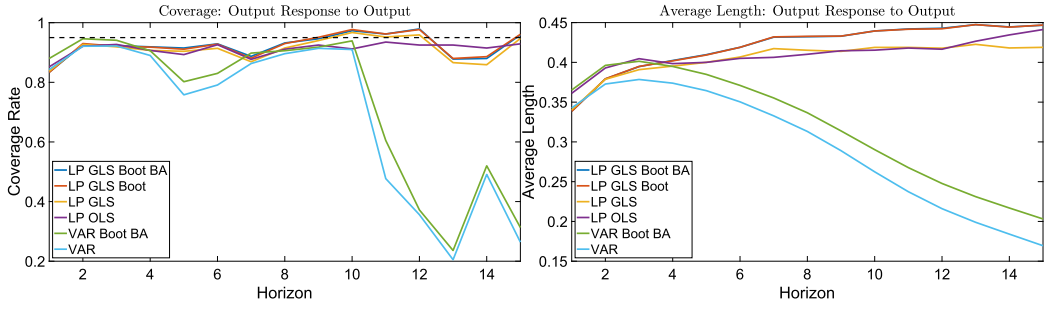


FIGURE 4. Coverage rates for 95% confidence intervals and average length for technology VAR.

90% for essentially all impulse responses at all horizons, with the occasional response dipping into the mid to high 80s.

Though including more lags may help protect against truncation bias when calibrating the VAR, it could be the case that it still does not accurately approximate the true DGP, making empirically calibrated VAR Monte Carlos not that informative about what may occur in practice. As an alternative, I include an estimated structural model. The following is the standard 3 equation New Keynesian model of a short term interest rate, output, and inflation from An and Schorfheide (2007). Select results are presented in Figure 5. All of the estimators generally had at least 90% coverage for all parameters. The average length for the LP GLS estimators are quite a bit shorter than the LP OLS estimator, and they are competitive with the VAR bootstrap. The average length of the analytic VAR confidence intervals are much wider than the bootstrap VAR confidence intervals and even the LP GLS intervals.

As discussed earlier, Monte Carlos are only as informative about what may occur in practice as they are good approximations of actual DGP. Another way to analyze the impact truncation bias may have on inference would be to generate data from theoretical models that have notoriously given VARs trouble. One theoretical model shown to have potential truncation bias issues is the standard medium scale DSGE model augmented with news shocks about future productivity from Sims (2012). The model is too

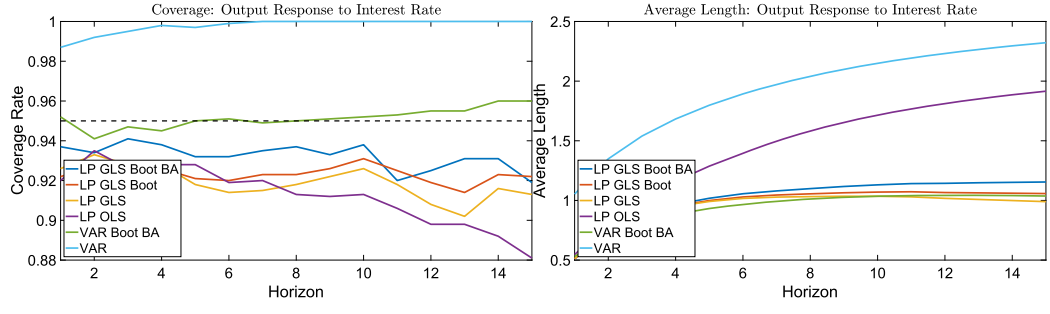


FIGURE 5. Coverage rates for 95% confidence intervals and average length for the 3-equation New Keynesian model.

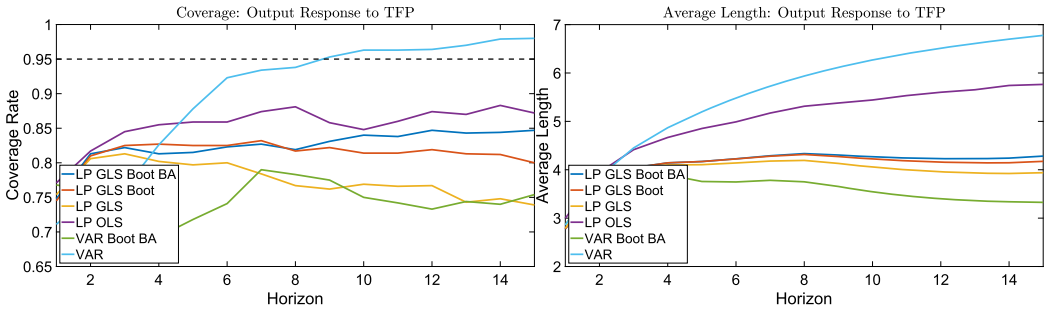


FIGURE 6. Coverage rates for 95% confidence intervals and average length for medium scale DSGE model augmented with news.

detailed to be expanded on here, but details can be found in Sims (2012).[16] Select results are presented in Figure 6. With the exception of the analytical VAR estimator, all of the estimators have below nominal coverage throughout, though coverage rates are comparable for most of those estimators. In terms of efficiency, the analytical VAR is the least efficient, but this is probably due in part to the analytical VAR sometimes having proper coverage while the other estimators do not. The LP GLS bootstrap estimators are more efficient than LP OLS, with large efficiency gains coming at longer horizons. The analytical LP GLS estimator is the most efficient of the LP estimators, but it also has the worst coverage. The VAR bootstrap estimator is the most efficient of them all, but arguably has the worst coverage.[17]

It is well known that VARs have trouble approximating real business cycle models (RBC) with technology shocks and long-run restrictions (see Chari, Kehoe, and McGrattan (2008), Poskitt and Yao (2017), and references therein). For the final DGP, I estimate the VARMA(1, 1) from the RBC model used in Poskitt and Yao (2017). Specific details of the RBC model can be found in Poskitt and Yao (2017), but the VARMA(1, 1) is

$$y_{t+1} = A_1 y_t + \varepsilon_{t+1} + M_1 \varepsilon_t, \quad \varepsilon_t \sim N(0, \Sigma),$$

where

$$A_1 = \begin{bmatrix} 0.9413 & 1.0446 \\ 0.0006 & 0.8045 \end{bmatrix}, \qquad M_1 = \begin{bmatrix} -0.2498 & -0.9173 \\ -0.1924 & 0.7065 \end{bmatrix},$$

$$\Sigma = \begin{bmatrix} 0.5186 & 0.4058 \\ 0.4058 & 0.4009 \end{bmatrix} \times 10^{-3}.$$

The literature in general only calculates the bias in the impulse responses for these models and does not analyze coverage or average length of different estimation methods, so this analysis will be informative about the properties of different estimators for these models. In order to impose long-run restrictions for LP, one must decide on the max

[16] Specifically, "The Full Model with One Period Anticipation" is used.

[17] Though not shown here, simply including 2 years worth of lags improves coverage for all estimators.

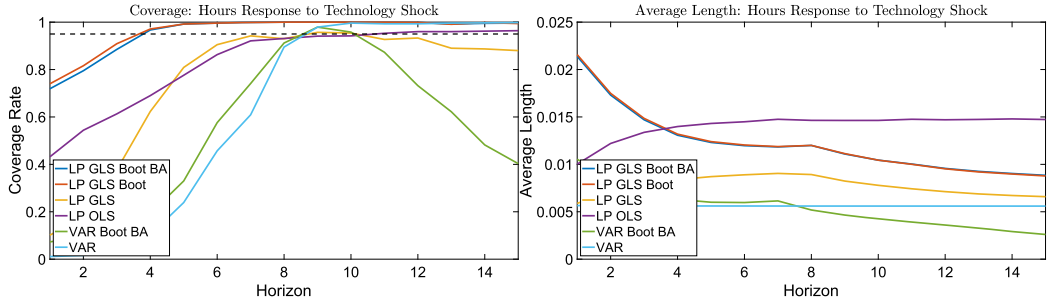


FIGURE 7. Coverage rates for 95% confidence intervals and average length for RBC VARMA.

horizon to be used when calculating the long-run restrictions.[18] There is a dearth of research on the topic, and the problem will not be solved in this paper. To choose the max horizon to be used when calculating the long-run restrictions, I used the ad hoc method of calculating the cumulative impulse responses (of the true Wold coefficients) at different horizons, and gauging how many horizons it takes for the cumulative impulse responses to essentially converge to the true long-run cumulative impulse responses. Taking into account that the sample size for the DGP is 250, I decided on using a max horizon of 150.[19] Due to it already being well established, this DGP can cause truncation bias issues. I do not do lag length selection and just include 2 years worth of lags.[20] Hours response to a technology shock is presented in Figure 7.

All of the estimators had below nominal coverage for at least the first couple of horizons. The LP GLS bootstrap estimators have close to above nominal coverage the rest of the way while the VAR bootstrap and analytical VAR estimator essentially have coverage distortions throughout. The analytical LP GLS and LP OLS estimators perform worse than the LP GLS bootstrap estimators in terms of coverage. The VAR estimators have the shortest average length for all of the estimators, followed by the analytical LP GLS estimator. The LP GLS bootstrap estimators are more efficient than the LP OLS estimator for all but the first 3 horizons.

In summary, I find that the LP GLS bootstrap estimators are the best at minimizing downside risks. They were generally more efficient than the LP OLS estimator (unless the LP OLS estimator was underestimating uncertainty).[21] The analytical LP GLS estimator did not perform as well as its bootstrap counterparts, but it was generally more efficient than LP OLS. The VAR bootstrap is the most efficient out of the estimators, but coverage can vary widely depending on the DGP. As highlighted in the empirically calibrated

[18]This is not an issue with the VAR since the long-run cumulative impulse response can be calculated with just the estimated VAR coefficients.

[19]Decreasing the max horizon made coverage worse for a couple of the impulse responses at some of the earlier horizons (e.g., 1–4), but the qualitative results were not sensitive to using 60, 70, or 100 horizons for example.

[20]Poskitt and Yao (2017) show that popular lag length criteria are essentially worthless when it comes to this DGP. Monte Carlo evidence indicates that there are sever coverage distortions for all estimators when using AIC, BIC, and HQIC.

[21]There were cases where the OLS estimator had shorter confidence intervals, but the coverage was below the nominal level.

Monte Carlos, it can easily be the case that the VAR has truncation bias issues that leads to poor coverage rates. Performance of the analytical VAR varied widely, sometimes having excessive average confidence interval length, comparable or greater than LP OLS.

## 6. Concluding remarks

I show that the autocorrelation process of LP residuals can be written as a VMA process of the Wold errors and impulse responses, and I derive a consistent GLS estimator for LP. In Monte Carlo simulations, I show that estimating LP with GLS can lead to more efficient estimates than LP OLS. The efficiency of LP GLS relative to LP OLS is not proved uniformly in this paper, though it is shown for the homoskedastic AR(1) case.

The results in this paper have many potential extensions. Since the autocorrelation process for the horizon $h$ LP is shown to be a VMA$(h-1)$ process of the Wold errors and impulse responses, this could be used to improve LP OLS bootstrapping. This knowledge could also be used to avoid choosing a HAC procedure and associated difficult-to-interpret tuning parameters for LP OLS by rearranging the score as was done in the GLS case. It may also be useful to extend LP GLS to a nonlinear (in the variables) or nonparametric setting. One potential solution would be to extend polynomial LP, which are motivated by a nonlinear version of the Wold representation (see Jordà (2005, Section 3) for more details). If one does not want to make assumptions about the functional form or the model, the second potential solution would be to extend nonparametric LP. Lastly, since LP are direct multistep forecasts, the results in this paper have the potential to improve the forecast accuracy of direct multistep forecasts.

## References

An, Sungbae and Frank Schorfheide (2007), "Bayesian analysis of DSGE models." *Econometric Reviews*, 26 (2–4), 113–172. [1215]

Bhansali, Raj J. (1997), "Direct autoregressive predictors for multistep prediction: Order selection and performance relative to the plug in predictors." *Statistica Sinica*, 7 (2), 425–449. [1210]

Breitung, Jorg and Ralf Brüggemann (2023), "Projection estimators for structural impulse responses." *Oxford Bulletin of Economics and Statistics* (forthcoming). https://doi.org/10.1111/obes.12562. [1206]

Brüggemann, Ralf, Carsten Jentsch, and Carsten Trenkler (2016), "Inference in vars with conditional heteroskedasticity of unknown form." *Journal of Econometrics*, 191 (1), 69–85. [1208, 1209, 1210]

Bruns, Martin and Helmut Lütkepohl (2022), "Comparison of local projection estimators for proxy vector autoregressions." *Journal of Economic Dynamics and Control*, 134. [1206]

Chari, Varadarajan, Patrick J. Kehoe, and Ellen R. McGrattan (2008), "Are structural vars with long-run restrictions useful in developing business cycle theory?" *Journal of Monetary Economics*, 55 (8), 1337–1352. [1216]

Gertler, Mark and Peter Karadi (2015), "Monetary policy surprises, credit costs, and economic activity." *American Economic Journal: Macro*, 7 (1), 44–76. [1200]

Goncalves, Sílvia and Lutz Kilian (2007), "Asymptotic and bootstrap inference for $\mathrm{ar}(\infty)$ processes with conditional heteroskedasticity." *Econometric Reviews*, 26 (6), 609–641. [1207, 1213]

Hansen, Lars Peter and Robert J. Hodrick (1980), "Forward exchange rates as optimal predictors of future spot rates: An econometric analysis." *Journal of Political Economy*, 88 (5), 829–853. [1199, 1200]

Hayashi, Fumio, ed. (2000), *Econometrics*, Princeton University Press, Princeton, NJ. [1200]

Jordà, Òscar (2005), "Estimation and inference of impulse responses by local projections." *American Economic Review*, 95 (1), 161–182. [1199, 1213, 1218]

Jordà, Òscar and Sharon Kozicki (2011), "Estimation and inference by the method of projection minimum distance: An application to the new Keynesian hybrid Phillips curve." *International Economic Review*, 52 (2), 461–487. [1207]

Kilian, Lutz and Yun Jung Kim (2011), "How reliable are local projection estimators of impulse responses?" *Review of Economics and Statistics*, 93 (4), 1460–1466. [1202, 1212, 1213]

Lazarus, Eben, Daniel J. Lewis, James H. Stock, and Mark W. Watson (2018), "Har inference: Recommendations for practice." *Journal of Business and Economic Statistics*, 36 (4), 541–559. [1209, 1211]

Lewis, Richard and Gregory C. Reinsel (1985), "Prediction of multivariate time series by autoregressive model fitting." *Journal of Multivariate Analysis*, 16 (3), 393–411. [1207]

Lusompa, Amaze (2023), "Supplement to 'Local projections, autocorrelation, and efficiency'." *Quantitative Economics Supplemental Material*, 14, https://doi.org/10.3982/QE1988. [1200]

Lütkepohl, Helmut (1990), "Asymptotic distributions of impulse response functions and forecast error variance decompositions of vector autoregressive models." *Review of Economics and Statistics*, 72 (1), 116–125. [1213]

McCracken, Michael and Serena Ng (2016), "Fred-md: A monthly database for macroeconomic research." *Journal of Business and Economic Statistics*, 34 (4), 574–589. [1211, 1212]

Plagborg-Møller, Mikkel and Christian K. Wolf (2021), "Local projections and vars estimate the same impulse responses." *Econometrica*, 89 (2), 955–980. [1199, 1212]

Poskitt, Donald S. and Wenying Yao (2017), "Vector autoregressions and macroeconomic modeling: An error taxonomy." *Journal of Business and Economic Statistics*, 35 (3), 407–419. [1216, 1217]

Ramey, Valerie A. (2016), "Macroeconomic shocks and their propagation." In *Handbook of Macroeconomics*, Vol. 2 (John B. Taylor and Harald Uhlig, eds.), 71–162. [1199, 1212, 1214]

Ramey, Valerie A. and Sarah Zubairy (2018), "Government spending multipliers in good times and in bad: Evidence from U.S. historical data." *Journal of Political Economy*, 126 (2), 850–901. [1214]

Sims, Eric R. (2012), "News, non-invertibility, and structural VARs." In *Advances in Econometrics*, Vol. 28 (Nathan Balke, Fabio Canova, Fabio Milani, and Mark A. Wynne, eds.), 81–135, Emerald Group Publishing Limited, Bingley, U.K. [1215, 1216]

---

Co-editor Tao Zha handled this manuscript.

Manuscript received 22 September, 2021; final version accepted 6 July, 2023; available online 25 July, 2023.