Local Projections vs. VARs: Lessons From Thousands of DGPs*

Dake Li
Princeton University

Mikkel Plagborg-Møller
Princeton University

Christian K. Wolf
University of Chicago

Preliminary and incomplete

February 22, 2021

Abstract: We conduct a simulation study of Local Projection (LP) and Vector Autoregression (VAR) estimators of structural impulse responses across thousands of data generating processes (DGPs), designed to mimic the properties of the universe of U.S. macroeconomic data. Our analysis considers various structural identification schemes and several variants of LP and VAR estimators, and we pay particular attention to the role of the researcher's loss function. A clear bias-variance trade-off emerges: Because our DGPs are not exactly finite-order VAR models, LPs have lower bias than VAR estimators; however, LPs have substantially higher variance than VARs at intermediate or long horizons. Unless researchers are overwhelmingly concerned with bias, additional shrinkage via Bayesian VARs or penalized LPs is attractive.

Keywords: external instrument, impulse response function, local projection, proxy variable, structural vector autoregression. *JEL codes:* C32, C36.

1 Introduction

Since Jordà (2005) introduced the popular local projection (LP) impulse response estimator, debate has raged about its benefits and drawbacks relative to Vector Autoregression (VAR)

*Email: dakel@princeton.edu, mikkelpm@princeton.edu, and ckwolf@uchicago.edu. We received helpful comments from Ulrich Müller, Chris Sims, Lumi Stevens, Andrei Zeleneev, and seminar participants at Princeton and UCL. 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.

estimation. Recently, Plagborg-Møller & Wolf (2020b) proved that these two seemingly different methods in fact estimate precisely the same impulse responses asymptotically, provided that the lag length used for estimation tends to infinity. This result holds regardless of identification scheme and regardless of the underlying data generating process (DGP). Nevertheless, the question of which estimator to choose in finite samples remains open. It is also an urgent question, since researchers have remarked that LPs and VARs can give conflicting results when applied to central economic questions such as the effects of monetary or fiscal stimulus (Ramey, 2016; Nakamura & Steinsson, 2018).

Whereas the LP estimator utilizes the sample autocovariances flexibly by directly projecting an outcome at the future horizon h on current covariates, a VAR(p) estimator extrapolates longer-run impulse responses from the first p sample autocovariances. Hence, though the estimates from the two methods agree approximately at horizons $h \leq p$, they can disagree substantially at intermediate and longer horizons.[1] Intuitively, the extrapolation employed by VARs should yield a lower variance but – potentially – a higher bias than for LPs, perfectly analogous to the trade-off between direct and iterated reduced-form forecasts (Schorfheide, 2005; Kilian & Lütkepohl, 2017).[2] How much more should one care about bias than variance to optimally choose the LP estimator over the VAR estimator in realistic sample sizes, and how does the trade-off depend on the DGP? Unfortunately, these questions are challenging to answer analytically, due to the dynamic and nonlinear nature of the time series estimators, as well as the breadth of DGPs encountered in applied practice.

In this paper we illuminate the bias-variance trade-off in impulse response estimation through a comprehensive simulation study of LP and VAR methods applied to thousands of empirically relevant DGPs. Our paper contributes to the literature in four ways. First, we consider a vast number of DGPs rather than a select few. These DGPs are obtained as randomly drawn subsets of variables from an encompassing model that has been fitted to a large data set of commonly encountered U.S. macroeconomic time series. Second, we refine what is often viewed as a binary distinction between "local projections" and "VARs" by additionally analyzing a whole menu of related estimation approaches that trace out the known bias-variance frontier. Third, we present results for three popular structural identification schemes: directly observed shocks, external instruments (IVs, also known as proxies), and recursive identification. Fourth, we examine the crucial role played by the

[1] See Plagborg-Møller & Wolf (2020b, Proposition 2) for a formal result.

[2] The trade-off is also conceptually similar to the relationship between polynomial series estimators and kernel estimators in cross-sectional nonparametric regression.

researcher's loss function by reporting the relative weights on bias and variance necessary to justify the use of one estimator over another.

We obtain our extensive array of DGPs by drawing specifications at random from a large-scale, empirically calibrated dynamic factor model (DFM).[3] In particular, we use the estimated DFM in Stock & Watson (2016), which has been fitted to 207 U.S. macroeconomic time series that span a wide variety of variable categories. As discussed by Stock & Watson, DFMs are known to accurately capture the joint co-movements of conventional macroeconomic data, ensuring that our simulation results will be informative about the universe of U.S. aggregate time series. From the encompassing 207-variable DFM we draw 6,000 random subsets of five variables, whose model-implied time series processes then constitute the set of DGPs in our simulation study. The DGPs vary in terms of how well they can be approximated by low-order VAR models, thus yielding an interesting bias-variance trade-off. Moreover, the DGPs exhibit substantial heterogeneity in persistence, nature of impulse responses, and invertibility of the structural shocks, consistent with the heterogeneity faced by applied researchers. One important caveat to our analysis, which we discuss further below, is that none of our DGPs exhibit (near-)unit roots, as the DFM is fitted to stationarity-transformed data.

We employ several variants of LP and VAR impulse response estimators under three possible structural identification schemes: observed shocks, IVs/proxies, and recursive identification.[4] In addition to the popular least-squares LP and VAR estimators, we enrich the bias-variance possibility frontier by considering: (i) penalized LP (Barnichon & Brownlees, 2019), which smooths out the estimated impulse response function; (ii) Bayesian VAR estimation, which shrinks VAR coefficients toward zero; (iii) model averaging of univariate and multivariate VAR models of various lag lengths (Hansen, 2016); and (iv) bias correction of the VAR coefficients (Pope, 1990; Kilian, 1998). Furthermore, for IV identification, we distinguish between internal IV methods (Ramey, 2011; Plagborg-Møller & Wolf, 2020b) and external IV methods (Stock, 2008; Stock & Watson, 2012; Mertens & Ravn, 2013).

Applying the estimation methods to simulated data from our multitude of DGPs, a clear and unavoidable bias-variance trade-off emerges:

1. Least-squares LP and VAR estimators lie on opposite ends of the bias-variance spectrum:

[3] Our overall approach is inspired by Lazarus et al. (2018). They, of course, are interested in the very different question of how to select among long-run variance estimators.

[4] To emulate standard monetary and fiscal shock applications, we restrict our DGPs to contain either the federal funds rate or aggregate government spending.

small bias and large variance for LPs, and large bias and small variance for VARs. For any given impulse response horizon, there exists a weight on squared bias relative to variance in the loss function that would make a researcher indifferent between the least-squares LP and VAR estimators. Broadly speaking, such indifference requires a researcher to care around four times more about squared bias than about variance.

2. Shrinkage methods dramatically lower the variance of LP and VAR methods, at a moderate cost in bias. Unless researchers care almost exclusively about bias, penalized LP is preferred to least-squares LP, and Bayesian VAR shrinkage is preferred to the least-squares VAR estimator.

3. For any given loss function, no single method dominates at all horizons. Unless the concern for bias is overwhelming, penalized LP is the most attractive estimation method at short horizons, while Bayesian VAR estimation is the most attractive method at intermediate and long horizons.

4. In the case of IV identification, the SVAR-IV estimator is heavily (median-)biased, but provides substantial reduction in dispersion. Depending on the weight attached to bias, it may therefore be justifiable to use external IV methods despite their lack of robustness to non-invertibility (unlike internal IV methods).

These findings provide a novel perspective on recent work emphasizing the dangers of VAR model mis-specification (Ramey, 2016; Nakamura & Steinsson, 2018). Our DGPs do not admit finite-order VAR representations, so VAR methods indeed invariably suffer from larger bias, as cautioned there. Reducing that bias via direct projection, however, incurs a steep cost in terms of increased sampling variance at intermediate and long horizons. Researchers who employ LP estimators should be prepared to pay that price.

Literature. Our large-scale simulation study is inspired by the seminal work of Marcellino et al. (2006) on direct and iterated multi-step forecasts, but we focus instead on structural impulse responses. While simulation studies in this branch of the forecasting literature often consider low-dimensional specifications, we consider systems with several variables, consistent with standard practice in the applied structural macroeconometrics literature. The structural perspective also requires us to contend with issues such as the variety of different shock identification schemes, normalization of impulse responses, and the special role of external instrumental variables.

Unlike our rich set of empirically calibrated DGPs, previous simulation studies of LP and VAR methods consider at most a handful of DGPs. Important examples of such previous studies include Jordà (2005), Meier (2005), Kilian & Kim (2011), Brugnolini (2018), Choi & Chudik (2019), and Austin (2020). These papers either obtain their DGPs from stylized, low-dimensional VARMA models, calibrated DSGE models, and/or a few empirically calibrated VAR models. Our analysis also differs in the following respects: we consider shrinkage estimation procedures as competitors to the familiar least-squares LP and VAR estimators; we study several popular identification schemes; and we examine how our conclusions vary with the impulse response horizon and the researcher's loss function.

Even though the simulation results are at the heart of our paper, we also illustrate the bias-variance trade-off through an analytical example that builds on Schorfheide (2005). That paper develops a general theory of the asymptotic bias and variance of direct and iterated (reduced-form) forecasts under local mis-specification. While these theoretical results are valuable for analytically distilling the forces at work, they do not by themselves resolve the bias-variance trade-off faced by applied practitioners – a trade-off that invariably depends on many features of the DGP.

We stress that our paper focuses solely on point estimation, as opposed to inference or hypothesis testing; we also restrict attention to stationary DGPs. See Inoue & Kilian (2020) and Montiel Olea & Plagborg-Møller (2020) for theoretical and simulation results on VAR and LP confidence interval procedures with highly persistent data. Moreover, we focus exclusively on impulse response estimands, rather than variance decompositions.

Outline. Section 2 illustrates the bias-variance trade-off for LP and VAR estimators using a simple analytical example. Section 3 describes the empirically calibrated dynamic factor model that we use to generate thousands of different DGPs. Section 4 defines the menu of LP- and VAR-based estimation procedures. Section 5 contains our main simulation results. Section 6 concludes by summarizing the lessons for applied researchers and offering guidance for future research. The appendix contains details on theory and implementation. A supplemental appendix with further simulation results and a Matlab code suite are available in our online GitHub repository. The code suite is designed to facilitate the application of the estimation procedures to other user-specified DGPs.[5]

[5] https://github.com/dake-li/lp_var_simul

2 The bias-variance trade-off

This section motivates our simulation study with an analytical discussion of the bias-variance trade-off between LP and VAR impulse response estimators. Section 2.1 begins with a simple model that cleanly illustrates the trade-off, and Section 2.2 connects the discussion to the rest of the paper.

2.1 Illustrative example

Plagborg-Møller & Wolf (2020b) show that the impulse response estimands of VAR and LP estimators with p lags generally differ at horizons $h > p$: the VAR extrapolates from the first p autocovariances of the data, while LP exploits all autocovariances out to horizon $h + p$. This observation suggests the presence of a bias-variance trade-off whenever the true DGP is not a finite-order VAR, perfectly analogous to the choice between "direct" and "iterated" predictions in multi-step forecasting (Marcellino et al., 2006). We here formalize this basic intuition by extending the arguments of Schorfheide (2005) to structural impulse response estimation in a simple DGP.

Model. Consider a simple sequence of drifting DGPs:

$$y_t = \rho y_{t-1} + \varepsilon_{1,t} + \varepsilon_{2,t} + \frac{\alpha}{\sqrt{T}}\varepsilon_{2,t-1}, \tag{1}$$

where $\varepsilon_t \equiv (\varepsilon_{1,t}, \varepsilon_{2,t})'$ is an i.i.d. white noise process with $\mathrm{Var}(\varepsilon_t) = \mathrm{diag}(1, \sigma_2^2)$. The DGP drifts towards an AR(1) process at rate $T^{-1/2}$, where T is the sample size. We will show that this ensures a non-trivial bias-variance trade-off in the limit $T \to \infty$. The DGP captures the notion that finite-order autoregressive models are often a good – but not exact – approximation to the true underlying DGP. Note that the degree of autoregressive misspecification is governed by α.

We are interested in the impulse responses of y_t with respect to a unit impulse in $\varepsilon_{1,t}$. The true impulse response function is evidently $\theta_h \equiv \rho^h$. We assume that the researcher observes $w_t \equiv (\varepsilon_{1,t}, y_t)'$, i.e., she observes the shock $\varepsilon_{1,t}$ but not $\varepsilon_{2,t}$. To evaluate the performance of a given estimator $\hat{\theta}_h$, we will throughout this paper consider loss functions of the form

$$\mathcal{L}_\omega(\theta_h, \hat{\theta}_h) = \omega \times \Big(\mathbb{E}[\hat{\theta}_h - \theta_h]\Big)^2 + (1 - \omega) \times \mathrm{Var}(\hat{\theta}_h). \tag{2}$$

For $\omega = \frac{1}{2}$, this equals the mean squared error (MSE). For $\omega > \frac{1}{2}$, the researcher is more concerned about (squared) bias than variance, and *vice versa*.

ESTIMATORS. For now, we consider two estimators of θ_h.

1. **LP.** The (least-squares) local projection estimator $\hat{\beta}_h$ is obtained from the OLS regression

$$y_{t+h} = \hat{\beta}_h \varepsilon_{1,t} + \hat{\zeta}_h' w_{t-1} + \text{residual}_h, \tag{3}$$

   at each horizon h. Notice that this LP specification controls for one lag of the data.

2. **VAR.** We consider a recursive structural VAR specification in $w_t = (\varepsilon_{1,t}, y_t)'$, again with one lag. Define the usual OLS coefficient estimator $\hat{A} \equiv (\sum_{t=2}^T w_t w_{t-1}')(\sum_{t=2}^T w_{t-1} w_{t-1})^{-1}$ and residual covariance matrix $\hat{\Sigma} \equiv T^{-1} \sum_{t=2}^T \hat{u}_t \hat{u}_t'$, where $\hat{u}_t \equiv w_t - \hat{A} w_{t-1}$. Define the lower triangular Cholesky factor $\hat{C}$, where $\hat{C}\hat{C}' = \hat{\Sigma}$. The un-normalized VAR impulse responses with respect to the first orthogonalized shock at horizon h are given by $\hat{A}^h \hat{C} e_1$, where e_j is the j-th unit vector of dimension 2, $j = 1, 2$. To facilitate comparison with LP, we normalize the impact response of the first variable in the VAR (i.e., $\varepsilon_{1,t}$) with respect to the first shock to be 1. This yields the estimator $\hat{\delta}_h \equiv e_2' \hat{A}^h \hat{\gamma}$, where $\hat{\gamma} \equiv (1, \hat{\kappa})'$ and $\hat{\kappa} \equiv \hat{\Sigma}_{21}/\hat{\Sigma}_{11}$.[6]

Note that at the impact horizon $h = 0$, the two estimators $\hat{\beta}_0$ and $\hat{\delta}_0$ are numerically equal.

TRADE-OFF. How does the optimal choice of estimation method depend on the properties of the DGP (1) and the bias weight ω in the loss function (2)? Along the stated asymptote, the researcher faces a clear bias-variance trade-off:

Proposition 1. *Consider the model (1), and fix $h \geq 0$, $\rho \in (-1, 1)$, $\sigma_2 > 0$, and $\alpha \in \mathbb{R}$. Assume $E(\varepsilon_{j,t}^4) < \infty$ for $j = 1, 2$. Define $\sigma_{0,y}^2 \equiv \frac{1+\sigma_2^2}{1-\rho^2}$. Then, as $T \to \infty$,*

$$\sqrt{T}(\hat{\beta}_h - \theta_h) \xrightarrow{d} N(\text{aBias}_{LP}, \text{aVar}_{LP}), \quad \sqrt{T}(\hat{\delta}_h - \theta_h) \xrightarrow{d} N(\text{aBias}_{VAR}, \text{aVar}_{VAR}), \tag{4}$$

where for all $h \geq 0$,

$$\text{aBias}_{LP} \equiv 0, \quad \text{aVar}_{LP} \equiv \sigma_{0,y}^2(1 - \rho^{2(h+1)}) - \rho^{2h},$$

[6] We have $\hat{C} = \begin{pmatrix} \sqrt{\hat{\Sigma}_{11}} & 0 \\ \hat{\Sigma}_{21}/\sqrt{\hat{\Sigma}_{11}} & \hat{\Sigma}_{22} - \hat{\Sigma}_{21}^2/\hat{\Sigma}_{11} \end{pmatrix}$. We therefore achieve the desired normalization of the impact effect of the shock by dividing $\hat{C} e_1$ by $\sqrt{\hat{\Sigma}_{11}}$. This gives the normalized impulse responses $\hat{A}^h \hat{\gamma}$.

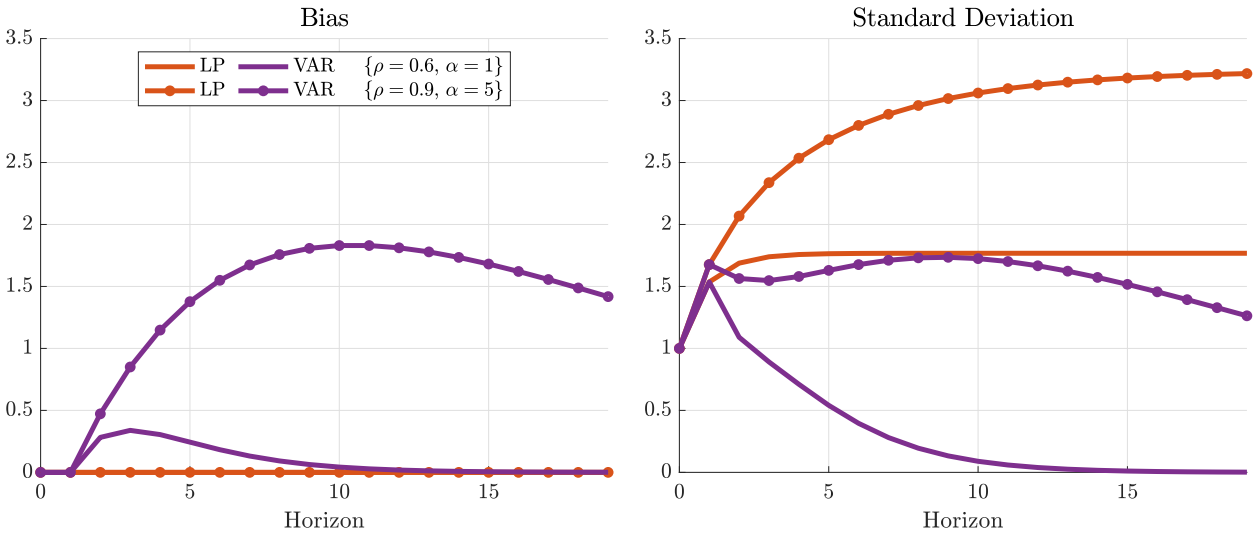

Figure 1: Asymptotic bias and standard deviation for LP (red) and VAR (purple) in the DGP (1) with $\sigma_1 = 1$ and $\{\rho = 0.6, \alpha = 1\}$ (no markers) or $\{\rho = 0.9, \alpha = 5\}$ (markers).

and for $h \geq 1$,

$$\text{aBias}_{VAR} \equiv \rho^{h-1}(h-1)\frac{\alpha\sigma_2^2}{\sigma_{0,y}^2 - 1}, \quad \text{aVar}_{VAR} \equiv \rho^{2(h-1)}(1-\rho^2)\sigma_{0,y}^2\left(1 + \frac{(h-1)^2}{\sigma_{0,y}^2 - 1}\right) + \rho^{2h}\sigma_2^2.$$

Proof. Please see Appendix C. □

Proposition 1 implies a ranking of the asymptotic biases and variances. First, we clearly have $|\,\text{aBias}_{\text{VAR}}\,| > |\,\text{aBias}_{\text{LP}}\,| = 0$ whenever $h \geq 2$, $\rho \neq 0$, and $\alpha \neq 0$. Second, as in Schorfheide (2005), the asymptotic variances in Proposition 1 are the same as in the well-specified model with $\alpha = 0$. It then follows from the Cramér-Rao bound that $\text{aVar}_{\text{VAR}} \leq \text{aVar}_{\text{LP}}$ for all h, and the inequality is strict when $h \geq 2$ and $\rho \neq 0$.[7]

Figure 1 plots the asymptotic bias and standard deviation for two parametrizations: one with low persistence and modest mis-specification ($\rho = 0.6$ and $\alpha = 1$), and one with high

[7]The argument goes as follows. Both asymptotic variances in Proposition 1 are the same as they would be when estimating the parameter $\tilde{\theta}_h \equiv e_2' A^h \text{chol}(\Sigma) e_1 / \Sigma_{11}^{1/2}$ in the well-specified bivariate VAR(1) model $w_t = A w_{t-1} + u_t$, $u_t \overset{i.i.d.}{\sim} N(0, \Sigma)$, when $A = \left(\begin{smallmatrix} 0 & 0 \\ 0 & \rho \end{smallmatrix}\right)$ and $\Sigma = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1+\sigma_2^2 \end{smallmatrix}\right)$ ("chol" denotes the lower triangular Cholesky factor). The VAR estimator is the MLE in this model. It can be verified that the LP estimator (which regresses $w_{2,t+h}$ on $w_{1,t}$, controlling for w_{t-1}) is a consistent and asymptotically regular estimator of $\tilde{\theta}_h$ in this model (regardless of A and Σ). Hence, the Cramér-Rao bound implies $\text{aVar}_{\text{VAR}} \leq \text{aVar}_{\text{LP}}$. Moreover, when $h \geq 2$ and $\rho \neq 0$, LP is strictly inefficient since its asymptotic correlation with the VAR estimator is not 1, according to the proof of Proposition 1.

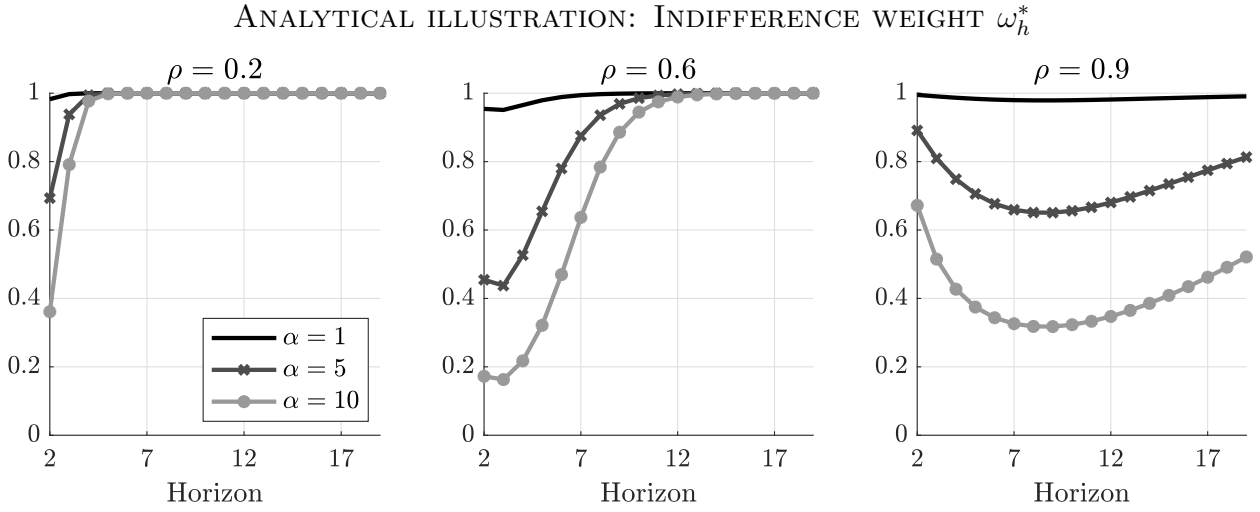

Figure 2: Weight $\omega = \omega_h^*$ in the asymptotic loss function $\omega \times \text{aBias}^2_{\hat{\theta}_h} + (1-\omega) \times \text{aVar}_{\hat{\theta}_h}$ that yields indifference between the LP and VAR estimator. LP is preferred whenever $\omega \geq \omega_h^*$. The three panels correspond to different values of $\rho \in \{0.2, 0.6, 0.9\}$; the three curves in each panel correspond to different values of $\alpha \in \{1, 5, 10\}$. All results are computed with $\sigma_2 = 1$. The figure omits the horizons $h \in \{0, 1\}$, at which the two estimation methods are (asymptotically) equivalent.

persistence and severe mis-specification ($\rho = 0.9$ and $\alpha = 5$).[8] Figure 2 plots the weight $\omega = \omega_h^*$ on bias in the asymptotic analogue of the loss function (2) that yields indifference between the LP and VAR estimator. That is, LP is preferred over VAR if and only if $\omega \geq \omega_h^*$. The three panels correspond to different degrees of persistence $\rho \in \{0.2, 0.6, 0.9\}$, while the three curves in each panel correspond to different degrees of mis-specification $\alpha \in \{1, 5, 10\}$.

We draw the following four conclusions from Proposition 1 and Figures 1 and 2:

1. For $h \in \{0, 1\}$, there actually is no bias-variance trade-off: on impact, the two estimators are numerically equivalent; at $h = 1$, both are asymptotically unbiased, and the asymptotic variance coincides. Thus, in Figure 1, the red and purple lines coincide initially, for both parametrizations. Intuitively, the equivalence at $h = 1$ reflects the fact that the VAR(1) estimator does not extrapolate, instead reporting the direct projection of y_{t+1} on w_t, exactly as LP does (Plagborg-Møller & Wolf, 2020b).

2. For $h \geq 2$, the bias-variance trade-off is non-trivial. LP directly projects y_{t+h} on the shock $\varepsilon_{1,t}$, which is uncorrelated with any lagged controls, so the asymptotic bias is always zero. In contrast, the VAR(1) estimator extrapolates the response at horizon h from the sample

[8] Both set $\sigma_2 = 1$. Assuming sample size $T = 200$, the mis-specification MA(1) term $\frac{\alpha}{\sqrt{T}}\varepsilon_{2,t-1}$ accounts for around 4% of the total variance of y_t in the first parametrization, and 28% in the second.

autocovariances at horizon 1. Though this tight parametric extrapolation yields a low variance relative to LP, it incurs a bias due to dynamic mis-specification when $\alpha \neq 0$. It follows that the indifference weight ω_h^* in Figure 2 is always in $(0, 1)$: LP can be justified if the concern for bias is sufficiently high, and *vice versa* for VAR.

3. The indifference weight ω_h^* is decreasing in the degree of mis-specification α. That is, the greater the mis-specification, the less the researcher needs to care about bias to prefer LPs over VARs. This result follows immediately from the observation that the asymptotic variances are independent of α, while the asymptotic bias of the VAR is increasing in α.[9]

4. For $h \to \infty$, the bias-variance trade-off is invariably resolved in favor of the VAR estimator: both asymptotic bias and variance for the VAR tend to zero, while the asymptotic variance of LP tends to $\sigma_{0,y}^2 > 0$. The practical relevance of this result hinges crucially on the persistence of the DGP, since Figure 2 shows that the indifference weight ω_h^* converges more slowly to 1 as $h \to \infty$ when the DGP is more persistent. The figure also shows that ω_h^* need not be a monotone function of the horizon h. It is therefore impossible to say at which horizons the VAR estimator is preferred over LP without carefully taking into account the horizon h, the persistence ρ, and the degree of mis-specification α.

2.2 Outlook

The simple analytical example shows that, at intermediate horizons, a researcher estimating the DGP (1) inevitably faces a non-trivial bias-variance trade-off between least-squares LP and VAR estimators. Resolving that trade-off requires taking a stand on the loss function and knowing something about the persistence of the DGP and degree of mis-specification of the finite-order VAR specification.

Motivated by these observations, we will in the rest of the paper explore the bias-variance trade-off across a much richer set of empirically relevant DGPs. In the language of Section 2.1, these DGPs will tell us about empirically plausible degrees of mis-specification α and persistence ρ, and so about the bias weight ω necessary to justify the use of one projection technique over another. Moreover, rather than insisting on the binary comparison of least-squares LP and VAR estimators, we will expand the bias-variance possibility frontier by considering several variants of these estimation procedures.

[9] Schorfheide (2005), shows that this result goes through in much more general DGPs, though he focuses on reduced-form estimators.

3 Data generating processes

This section presents our DGPs. We define the empirically calibrated encompassing model in Section 3.1, from which we draw thousands of DGPs with corresponding structural impulse response estimands, as described in Section 3.2. We discuss implementation details in Section 3.3, and provide summary statistics for the DGPs in Section 3.4.

3.1 Encompassing model

We construct our simulation DGPs from an encompassing model that is known to accurately describe the population of U.S. macroeconomic time series: the large-scale dynamic factor model (DFM) of Stock & Watson (2016). As we illustrated in Section 2, the nature of the bias-variance trade-off for structural impulse response estimation is necessarily sensitive to properties of the underlying DGP. This is why it is crucial to employ an encompassing model that is known to accurately capture the co-movements among a large number of commonly used U.S. macroeconomic time series.

The DFM postulates that a large-dimensional $n_X \times 1$ vector X_t of observed macroeconomic time series is driven by a low-dimensional $n_f \times 1$ vector f_t of latent factors, as well as an $n_X \times 1$ vector v_t of idiosyncratic shocks. The latent factors are assumed to follow a stationary VAR(p) process

$$f_t = \Phi(L) f_{t-1} + H\varepsilon_t, \tag{5}$$

where $\varepsilon_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{n_f,t})'$ is an $n_f \times 1$ vector of aggregate shocks, which are i.i.d. and mutually uncorrelated, $\text{Var}(\varepsilon_t) = I_{n_f}$. The $n_f \times n_f$ matrix H determines the impact impulse responses of the factors with respect to the aggregate shocks. The observed macroeconomic aggregates X_t are given by

$$X_t = \Lambda f_t + v_t, \tag{6}$$

where the idiosyncratic innovation $v_{i,t}$ for macro observable $X_{i,t}$ follows the AR(q) process

$$v_{i,t} = \Delta_i(L) v_{i,t-1} + \Xi_i \xi_{i,t}, \tag{7}$$

with ξ_{it} being i.i.d. across t and i. We assume all shocks and innovations are jointly normal and homoskedastic. In Section 3.3 we describe how the parameters of the DFM are calibrated to the Stock & Watson (2016) data set, but first we describe how we construct our lower-dimensional DGPs from the encompassing large-scale DFM.

3.2 DGPs and impulse response estimands

We use the encompassing model (5)–(7) to build thousands of lower-dimensional DGPs for our simulation study. Specifically, for each DGP, we draw a random subset of $n_{\bar{w}}$ variables $\bar{w}_t$ from the large vector X_t, i.e., $\bar{w}_t \subset X_t$. The variables $\bar{w}_t$ follow the time series process implied by the encompassing model (5)–(7). In particular, $\bar{w}_t$ is driven by some combination of aggregate structural shocks ε_t and idiosyncratic innovations v_t. We draw thousands of such random combinations of variables, thus yielding thousands of lower-dimensional DGPs. The details of how we select the variable combinations are postponed until Section 3.3.

For each DGP drawn in this way, we consider three types of structural impulse response estimands, mimicing popular identification schemes in applied macroeconometrics (Ramey, 2016; Stock & Watson, 2016). In the following, $y_t \in \bar{w}_t$ denotes a response variable of interest in the DGP, x_t is an impulse variable (which might equal an aggregate structural shock $\varepsilon_{1,t}$, or might be one variable in $\bar{w}_t$), $i_t \in \bar{w}_t$ is a variable used to normalize the scale of the shock (to be defined below), z_t is an external instrument (if applicable), and w_t denotes the vector of all observed time series in the DGP.

1. **Observed shock identification.** In this identification scheme we assume that the econometrician observes both the endogenous variables $\bar{w}_t$ and the first structural shock $\varepsilon_{1,t}$, so the full vector of observables is $w_t = (\varepsilon_{1,t}, \bar{w}_t)'$. The objects of interest are the scaled impulse responses of an outcome variable y_t with respect to $\varepsilon_{1,t}$:

$$\theta_h \equiv \frac{\bar{\Lambda}_{\iota_y,\bullet}\Theta^f_{\bullet,1,h}}{\bar{\Lambda}_{\iota_i,\bullet}\Theta^f_{\bullet,1,0}}, \quad h = 0, 1, 2, \ldots, \tag{8}$$

   where $\Theta^f(L) = (I - \Phi(L))^{-1}H$ are the impulse responses of the factors f_t to the structural shocks ε_t implied by (5), while $\bar{\Lambda}$ are those rows of Λ that correspond to the observables $\bar{w}_t$. The indices ι_y and ι_i correspond to the locations of y_t and i_t in the vector $\bar{w}_t$, respectively. Notice that we use a "unit effect" normalization (Stock & Watson, 2016): We consider a shock $\varepsilon_{1,t}$ of a magnitude that raises the variable i_t by one unit on impact.

   This set-up captures those empirical studies in which the researcher has constructed a plausible direct measure of the shock of interest; examples include the monetary shock series of Romer & Romer (2004) or the fiscal shock series of Ramey (2011).

2. **IV/proxy identification.** In this scheme, rather than directly observing the structural

shock $\varepsilon_{1,t}$, the econometrician observes the noisy proxy

$$z_t = \rho_z z_{t-1} + \varepsilon_{1,t} + \nu_t, \tag{9}$$

where ν_t is an i.i.d. process (independent of all shocks and innovations in the DFM) with $\text{Var}(\nu_t) = \sigma_\nu^2$. The full vector of observables is thus $w_t = (z_t, \bar{w}_t)'$. The structural object of interest is the same scaled impulse response function as in (8). One example of an IV z_t is the high-frequency change in futures prices around monetary policy announcements employed by Gertler & Karadi (2015) to identify the effects of monetary policy shocks.

3. **Recursive identification.** For our third and final structural estimand, the researcher observes only the endogenous variables $\bar{w}_t \subset X_t$, with no further direct or noisy shock measures. Thus, the total vector of observables is $w_t = \bar{w}_t$. It is straightforward to show from (5)–(7) that w_t follows an *infinite-order* VAR process,

$$w_t = \sum_{\ell=1}^{\infty} A_\ell^w w_{t-\ell} + u_t^w, \tag{10}$$

where u_t^w is a white noise process. We provide formulas for $\{A_\ell^w, \Sigma_u^w\}$ as a function of DFM primitives in Appendix A.4. Consistent with a large literature on recursive shock identification in VARs (e.g., Christiano et al., 1999; Blanchard & Perotti, 2002), we take as the estimand the impulse responses with respect to a recursive orthogonalization of the reduced-form (Wold) forecast errors u_t^w. Define the Cholesky decomposition $\text{Var}(u_t^w) = \Sigma_u^w = B^w(B^w)'$, with B^w lower triangular, and define $C^w(L) \equiv A^w(L)^{-1}$, where $A^w(L) = I_{n_{\bar{w}}} - \sum_{\ell=1}^{\infty} A_\ell^w L^\ell$. Then we define the recursive impulse response estimands as

$$\theta_h \equiv \frac{C^w_{\iota_y,\bullet,h} B^w_{\bullet,\iota_x}}{C^w_{\iota_i,\bullet,0} B^w_{\bullet,\iota_x}}, \quad h = 0, 1, 2, \ldots, \tag{11}$$

where ι_y, ι_x, and ι_i are the indices corresponding to y_t, x_t, and i_t in the vector $\bar{w}_t$, respectively. Unlike the observed shock and IV estimands considered above, the estimand (11) might not equal the model-implied *structural* impulse response of the variable y_t with respect to any aggregate shock $\varepsilon_{j,t}$ in the DFM.[10] In other words, the expression (11) is the impulse response with respect to a potentially non-structural innovation. We

[10] A necessary condition for the impulse responses (11) to equal structural impulse responses from (5)–(7) is that $\varepsilon_{j,t} \in \text{span}(\{\bar{w}_{t-\ell}\}_{\ell=0}^{\infty})$ for at least one shock j. A sufficient condition for $\varepsilon_t \in \text{span}(\{\bar{w}_{t-\ell}\}_{\ell=0}^{\infty})$ is that $n_{\bar{w}} = n_f$, $\bar{\Lambda}$ is non-singular, and $\Xi_i = 0$ for all i in $\bar{w}_t$.

nevertheless consider the estimand (11) due to its popularity in applied work.

To interpret the estimand (11), consider two popular applied identification schemes. First, in the monetary policy shock identification scheme of Christiano et al. (1999), y_t may be aggregate output, $x_t = i_t$ is the nominal interest rate, and the nominal rate is typically ordered *after* all other observables. The recursive estimand is then the impulse response of output to a residualized interest rate innovation, normalized by the impact response of interest rates. Second, for the fiscal policy shock identification procedure in Blanchard & Perotti (2002), we may again take y_t to be aggregate output and let $x_t = i_t$ be aggregate government spending. In this case, reduced-form innovations in the government spending equation are treated as structural shocks, and so (11) gives impulse responses to those innovations, normalized by the impact response of government spending.

3.3 Implementation

This section first discusses the empirical calibration of the DFM and then specifies the particular DGPs and structural impulse responses that we consider in the simulation study.

DFM PARAMETERS. We parametrize the DFM (5)–(7) based on the empirical reduced-form parameter estimates from Stock & Watson (2016), using the same specification as in Lazarus et al. (2018). We provide a brief summary here, with details in Appendix A.1 and in the original work of Stock & Watson. The full vector of observables X_t contains quarterly observations on 207 time series, mostly consisting of real activity variables, price measures, interest rates, asset and wealth variables, and productivity series. Each series is seasonally adjusted and – importantly – transformed to approximate stationarity. Following the formal dimensionality tests of Stock & Watson, we allow for $n_f = 6$ factors and two lags in the factor equation (5). Finally, we allow for two lags in the idiosyncratic innovation equation (7). The reduced-form parameters are estimated by principal components and least-squares procedures. This pins down all parameters of the DFM, except the structural impact response matrix H, which we discuss below.

Given the large number and wide coverage of the macro series included in X_t, we believe that our simulation study will be informative about properties of the *universe* of U.S. time series that are frequently encountered in applied work.

DGP AND ESTIMAND SELECTION. To provide a comprehensive picture of the bias-variance trade-off, we select thousands of different possible sets of observables $\bar{w}_t \subset X_t$. We consider

two protocols for selecting these observables: one aimed at mimicking monetary policy shock applications, and one aimed at fiscal policy shock applications. Specifically, for each estimand and each type of policy shock, we randomly draw $n_{\text{spec}} = 3{,}000$ configurations of $n_{\bar{w}} = 5$ macro observables $\bar{w}_t$. Thus, we end up with a total of 6,000 DGPs. For the monetary policy DGPs we restrict $\bar{w}_t$ to always contain the federal funds rate, while for the fiscal policy DGPs we restrict $\bar{w}_t$ to contain federal government spending. These two series are chosen as the normalization variables i_t. The remaining four variables in $\bar{w}_t$ are selected uniformly at random from X_t, except we impose that at least one variable should be a measure of output, and at least one other variable a measure of prices. The impulse response variable y_t is selected uniformly at random from the four series (other than i_t).

For each of the 6,000 DGPs, we define the three structural impulse response estimands in Section 3.2 as follows:

1. **Observed shock.** We select the structural impact response matrix H in the factor equation (5) so as to maximize the impact effect of the shock $\varepsilon_{1,t}$ on the federal funds rate (for monetary shocks) and government spending (for fiscal shocks), subject to the constraint that H is consistent with our estimate of the reduced-form innovation variance-covariance matrix for the factors.[11] This ensures that monetary and fiscal shocks account for substantial short-run variation in nominal interest rates and government spending, respectively. Additionally, this construction avoids issues related to division by near-zeros when normalizing the impulse responses.

2. **IV.** The matrix H is defined as in the "observed shock" case. As for the IV parameters in equation (9), we set $\rho_z = 0.1$ (i.e., the IV is close to white noise). To ensure an empirically plausible signal-to-noise ratio, we calibrate σ_ν^2 using real-world IV data series. Specifically, we base the calibration on the Romer & Romer (2004) series for monetary policy DGPs and the Ben Zeev & Pappa (2017) series for fiscal policy DGPs. See Appendix A.3 for details. We show below in Section 3.4 that the strength of these IVs ranges from somewhat weak to moderately strong, as measured by the first-stage F-statistic.[12]

3. **Recursive.** For monetary policy DGPs, we order the federal funds rate last, as in Christiano et al. (1999); this restricts the other included variables to not respond contempo-

[11] This criterion only pins down the first column of H. The remaining columns are selected arbitrarily subject to the constraints imposed by the estimated reduced-form parameters, since only their reduced-form implications matter for our simulations.

[12] In future work we hope to add additional heterogeneity to the IV parameters.

DGP SUMMARY STATISTICS

Percentile	min	10	25	50	75	90	max
Data and shocks							
trace(long-run var)/trace(var)	0.42	0.93	0.98	1.14	2.29	4.78	18.09
Largest VAR eigenvalue	0.82	0.84	0.84	0.84	0.84	0.86	0.91
Fraction of VAR coef's $\ell \geq 5$	0.02	0.10	0.15	0.23	0.34	0.44	0.84
Degree of shock invertibility	0.14	0.16	0.19	0.28	0.41	0.47	0.65
IV first stage F-statistic	9.49	9.61	9.70	17.29	27.58	28.20	29.10
Impulse responses up to $h = 20$							
No. of interior local extrema	1	2	2	2	3	4	6
Horizon of max abs. value	0	0	0	0	1	2	8
Average/(max abs. value)	-0.44	-0.16	-0.09	-0.02	0.06	0.11	0.46
R^2 in regression on quadratic	0.01	0.11	0.23	0.49	0.71	0.84	0.98

Table 1: Quantiles of various population parameters across the 6,000 DGPs for observed shock and IV identification. "long-run var": long-run variance. "Fraction of VAR coef's $\ell \geq 5$": $\sum_{\ell=5}^{50} \|A_\ell^w\| / \sum_{\ell=1}^{50} \|A_\ell^w\|$, where $\|\cdot\|$ is the Frobenius norm. IV first stage F-statistic: $T \times R^2/(1-R^2)$, where $T = 200$ and R^2 is the population R-squared in a projection of i_t on $(z_t - E(z_t \mid z_{t-1}))$. "Average/(max abs. value)": $(\frac{1}{20}\sum_{h=0}^{19} \theta_h)/\max_h\{|\theta_h|\}$. "$R^2$ in regression on quadratic": R-squared from a regression of the impulse response function $\{\theta_h\}_{h=0}^{19}$ on a quadratic polynomial in h.

raneously to the monetary innovation. For fiscal policy DGPs, we order the government expenditure series first, as in Blanchard & Perotti (2002); this restricts the fiscal authority to respond to other innovations in the recursive VAR with a lag.

3.4 Summary statistics

Consistent with the experience of applied researchers, our DGPs exhibit substantial heterogeneity along several dimensions. Table 1 displays the distribution of various population parameters across our 6,000 DGPs. The table focuses on impulse responses with respect to directly observed monetary policy and government spending shocks, though results for recursively defined shocks are similar, as shown in Supplemental Appendix E.

First of all, the DGPs feature varying degrees of persistence. Our primary measure of the persistence of the DGP is given by $\text{trace}(LRV(\bar{w}_t))/\text{trace}(\text{Var}(\bar{w}_t))$, where "LRV" denotes

the long-run variance matrix.[13] This measure varies widely across the DGPs, with more than 25% of the DGPs being anti-persistent on average (i.e., the ratio is less than 1), and the 90th percentile nearly equal to 5. The largest eigenvalue of the VAR companion matrix, another measure of persistence, has a median of 0.84 and does not vary as much across DGPs.[14]

Second, the DGPs are heterogeneous in terms of how well they can be approximated by a low-order VAR. Table 1 reports the ratio $\sum_{\ell=5}^{50} \|A_\ell^w\| / \sum_{\ell=1}^{50} \|A_\ell^w\|$, which measures the relative magnitude of the VAR coefficients at or after lag 5 (we use the Frobenius matrix norm). The 10th and 90th percentiles equal 0.10 and 0.44, respectively. Hence, the analysis in Section 2 suggests that the bias of low-order VAR procedures will vary substantially across our various DGPs.

Third, for the IV specifications, the DGPs differ in terms of the degree of invertibility of the shock and the strength of the IV. The degree of invertibility is defined as the R-squared value in a population projection of the shock of interest on current and lagged macro observables $\{\bar{w}_{t-\ell}\}_{\ell=0}^{\infty}$. The bias of some SVAR-based external instrument procedures depends on how far below 1 this measure is, as discussed further in Section 4. The table shows that 90% of the DGPs have degrees of invertibility below 47%, i.e., substantial non-invertibility. This is not surprising: The DFM (5)–(7) features a realistic amount of idiosyncratic noise v_t, making it challenging to accurately back out the aggregate shock of interest $\varepsilon_{1,t}$ from a small number of observed time series $\bar{w}_t$. The strength of the IV is somewhat weak to moderate, as the population first stage F-statistic (from a regression of the normalization variable i_t on the IV z_t) varies between approximately 10 and 30, given sample size $T = 200$.

Finally, the true values of our impulse response estimands exhibit a wide variety of shapes. Table 1 shows that, though most impulse response functions peak at horizons $h = 0$ or $h = 1$, they are typically not simple monotonically decaying or even hump-shaped functions: the median number of interior local extrema of the impulse response functions is 2 (a monotonic function would have 0; a hump-shaped function would have 1). Many impulse response functions change sign at some horizon, as evidenced by the average response (across horizons) typically being much smaller than the the maximal response. Finally, the smoothness of the impulse response functions varies substantially: The R-squared value in a regression of the impulse responses $\{\theta_h\}_{h=0}^{19}$ on a quadratic polynomial $b_0 + b_1 \times h + b_2 \times h^2$ has 10th and 90th percentiles given by 0.11 and 0.84, respectively. Figure 3 displays the true values of

[13] This measure equals $(1 + \rho)/(1 - \rho)$ for an AR(1) process with coefficient ρ.

[14] This is because the interest rate and government spending series are quite persistent, and all DGPs include one of these two series.

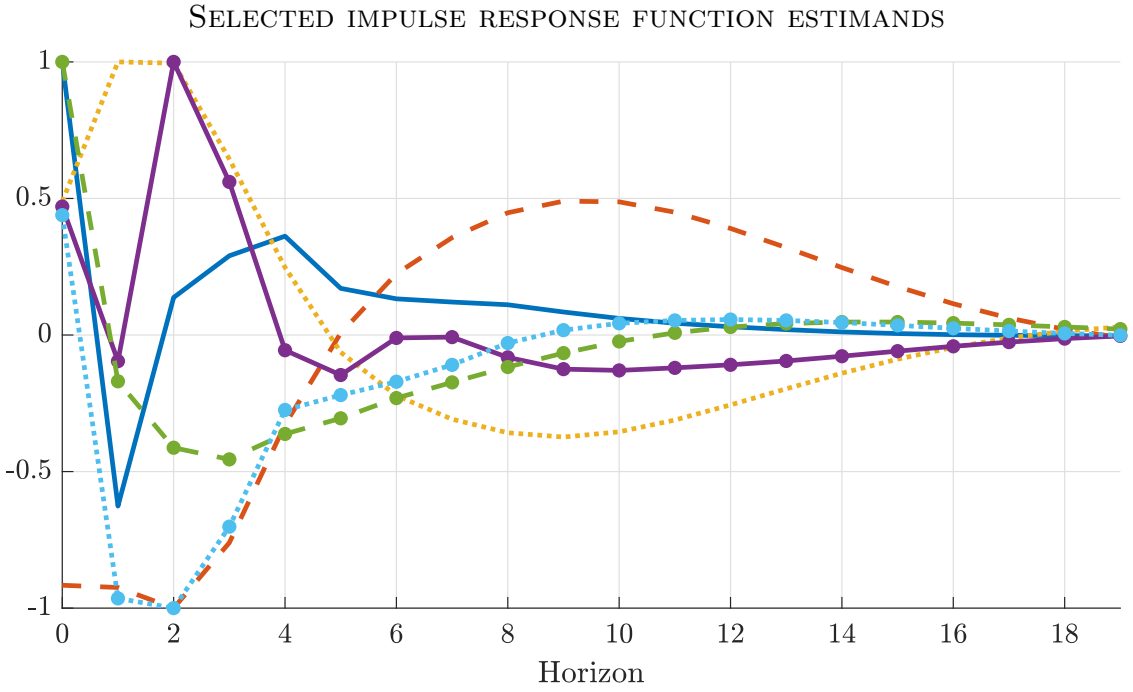

Figure 3: Selected impulse responses of macro observables to monetary and fiscal policy shocks. Here the impulse response functions are normalized to have a maximum value 1 or -1.

six impulse response functions that provide a representative picture of the heterogeneity. Note, however, that all of them tend to zero when the horizon gets large, as required by the stationarity of the DGPs.

4 Estimation methods

We now give a brief overview of the different VAR- or LP-based estimation methods that we consider in the simulation study. Though all these methods aim at estimating the same population impulse responses (8) and (11), the methods differ in terms of their bias/variance properties, and in terms of their robustness to non-invertibility. Mathematical definitions and implementation details are relegated to Appendix B. All estimators include an intercept.

Local projection approaches. The basic idea behind local projections, as proposed by Jordà (2005), is to estimate the impulse responses separately at each horizon by a direct regression of the future outcome on current covariates. We consider two such approaches:

1. **Least-squares LP.** OLS regression of the response variable y_{t+h} on the innovation variable x_t, controlling for p lags of all data series w_t. The innovation variable equals

$x_t = \varepsilon_{1,t}$ for "observed shock" identification and $x_t = z_t$ for IV identification. In the case of recursive identification, we additionally control for the contemporaneous values of the variables that are ordered before x_t in the system (Plagborg-Møller & Wolf, 2020b). Since least-squares LP does not mechanically impose any functional form on the relationship between impulse responses at different horizons h, the bias tends to be small. However, the estimated impulse response functions tend to look jagged in finite samples and tend to be estimated with high variance at longer horizons.

2. **Penalized LP** (abbreviated "Pen LP"). To lower the variance of least-squares LP at the expense of potentially increasing the bias, Barnichon & Brownlees (2019) propose a penalized regression modification of LP. The estimator minimizes the sum of squared forecast residuals (across both horizons and time) and a penalty term that encourages the estimation of smooth impulse responses. This is a type of shrinkage estimation: The unrestricted least-squares estimate is pushed in the direction of a smooth quadratic function of the horizon. The degree of shrinkage is chosen by cross-validation.

In the case of IV identification, we apply the LP-IV estimation approach of Stock & Watson (2018), which is robust to non-invertibility. This approach is easily adapted to penalized LP, as discussed in Appendix B.1.

VAR APPROACHES. Like local projections, a VAR with lag length p flexibly estimates the impulse responses out to horizon p; however, the VAR extrapolates the responses at longer horizons $h > p$ using only the sample autocovariances out to lag p. As suggested by the analysis in Section 2, this tends to generate impulse response estimates with lower variance but higher bias than LP estimates at intermediate horizons. In our stationary class of DGPs, VAR impulse response estimates eventually tend to zero when the horizon gets large (with high probability), unlike LP estimates; this feature will be important for the bias/variance trade-off at long horizons. We consider four VAR-based approaches:

1. **Least-squares VAR.** Standard VAR impulse response estimates based on equation-by-equation OLS estimates of the reduced-form coefficients.

2. **Bias-corrected VAR** (abbreviated "BC VAR"). As above, but using the formula in Pope (1990) to analytically correct the first-order bias of the reduced-form coefficients caused by persistent data.

3. **Bayesian VAR** (abbreviated "BVAR"). As above, but where the reduced-form coefficients are posterior mean estimates under a Minnesota-style prior.[15] Due to the stationarity of the DGP, we shrink towards independent white noise processes (rather than independent unit root processes, which is conventional for highly persistent data). The prior variance hyper-parameters follow the recommendations in Canova (2007).

4. **VAR model averaging** (abbreviated "VAR Avg"). Hansen (2016) develops a data-driven method for averaging across the impulse response estimates produced by several different VAR specifications. We construct a weighted average of 40 different specifications, each of which is estimated by OLS: univariate AR(1) to AR(20) models, and multivariate VAR(1) to VAR(20) models. The weights are chosen to minimize an empirical estimate of the final impulse response estimator's MSE.

Observed shock identification is carried out by simply ordering the shock first in the recursive VAR. We consider two different approaches to IV estimation:

i) **Internal instruments.** Proceed as if the IV were equal to the true shock of interest, i.e., order the IV first in the VAR and compute responses to the first orthogonalized innovation (Ramey, 2011). Plagborg-Møller & Wolf (2020b) prove that this approach consistently estimates the *normalized* structural impulse responses (8) even if the IV is contaminated with measurement error as in (9), and even if the shock is non-invertible.

ii) **SVAR-IV** (also known as proxy-SVAR). Exclude the IV from the reduced-form VAR, and estimate the structural shock by projecting the IV on the reduced-form VAR innovations (Stock, 2008; Stock & Watson, 2012; Mertens & Ravn, 2013; Gertler & Karadi, 2015). This estimator is consistent if the shock of interest is invertible, but not otherwise (Forni et al., 2019; Miranda-Agrippino & Ricco, 2019; Plagborg-Møller & Wolf, 2020a). We shall see that the SVAR-IV estimator tends to have lower variance than the "internal instruments" estimator due to the smaller dimension of the VAR system, potentially justifying the use of the former despite its lack of robustness to non-invertibility.

We implement the "internal instruments" approach using all four types of VAR estimation techniques described earlier. For brevity, we only consider the least-squares version of the SVAR-IV estimator.

[15] Note that, due to computational constraints, we do not report the posterior means of the impulse responses, but instead iterate on the posterior means of the VAR parameters.

To visualize the various estimation methods, Supplemental Appendix D plots the estimated impulse response functions in a few data sets simulated from a single DGP.

5 Results

We now present our main simulation results. We summarize the results through four takeaways, one in each subsection. The first three takeaways focus on observed shock identification, though results for recursive identification are very similar, as shown in Supplemental Appendix E. The fourth takeaway is concerned with IV identification.

Lag length and simulation settings. The evidence presented in this section focuses on estimators that use $p = 4$ lags for estimation (except VAR model averaging, which uses many different lag lengths, as described in Section 4). In our DGPs the Akaike Information Criterion almost always selects very short lag lengths $\hat{p}_{AIC}$. Thus, for all intents and purposes our results may be interpreted as having been generated by the lag length selection rule $p = \max\{\hat{p}_{AIC}, 4\}$. Our reading of applied practice is that researchers typically include at least 4 lags in quarterly data. Results for 8 lags are presented in Supplemental Appendix F and discussed below when relevant.

Throughout this section we consider the 6,000 monetary and fiscal policy shock DGPs *jointly* rather than separately. Separate results for the two kinds of DGPs are relegated to Supplemental Appendix G, as they are qualitatively similar. For each DGP, we simulate time series of length $T = 200$ quarters and approximate the population bias and variance of the estimators by averaging across 5,000 Monte Carlo simulations. The results presented in this section take about two weeks to produce in Matlab on a server with 25 parallel cores.

5.1 There is a clear bias-variance trade-off between LP and VAR

Our first takeaway is that researchers invariably face a bias-variance trade-off: Because most of our DGPs are not well approximated by finite-order VAR models, least-squares LPs tend to have lower bias, while least-squares VAR estimators tend to have lower variance. This agrees with the simple analytical example in Section 2. We focus on the baseline least-squares LP and VAR estimators in this subsection, leaving other procedures for later.

Figures 4 and 5 depict the bias-variance trade-off at various horizons. These figures show the median (across our 6,000 DGPs) of the absolute bias or standard deviation, respectively, as a function of the horizon. The different lines correspond to different estimators, with

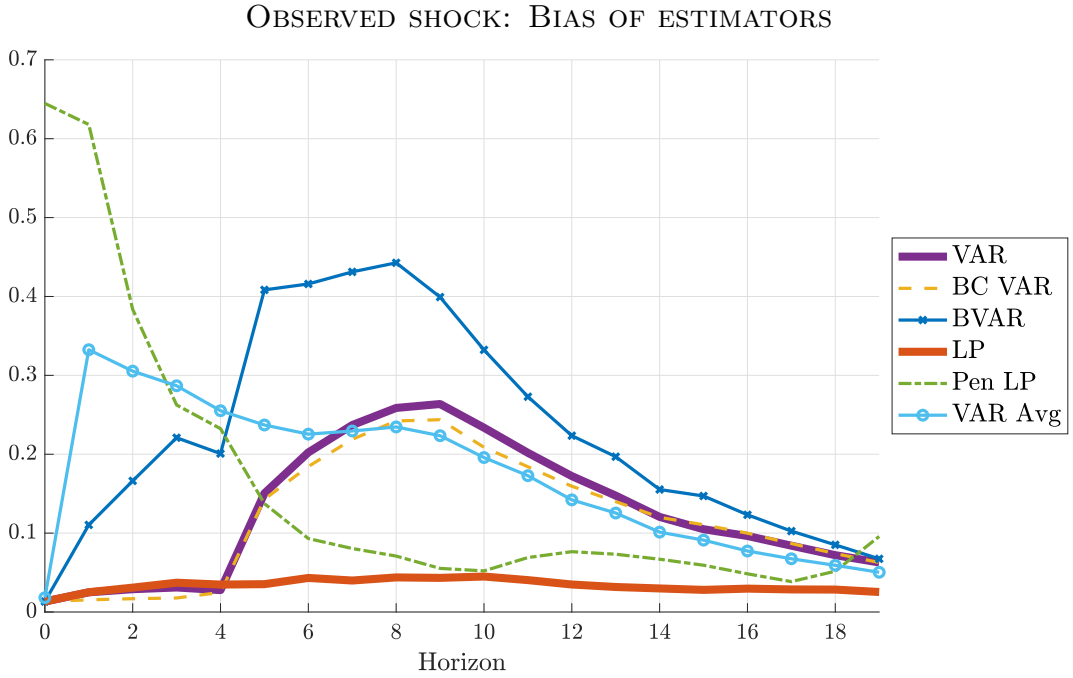

Figure 4: Median (across DGPs) of absolute bias of the different estimation procedures, relative to $\sqrt{\frac{1}{20}\sum_{h=0}^{19}\theta_h^2}$.

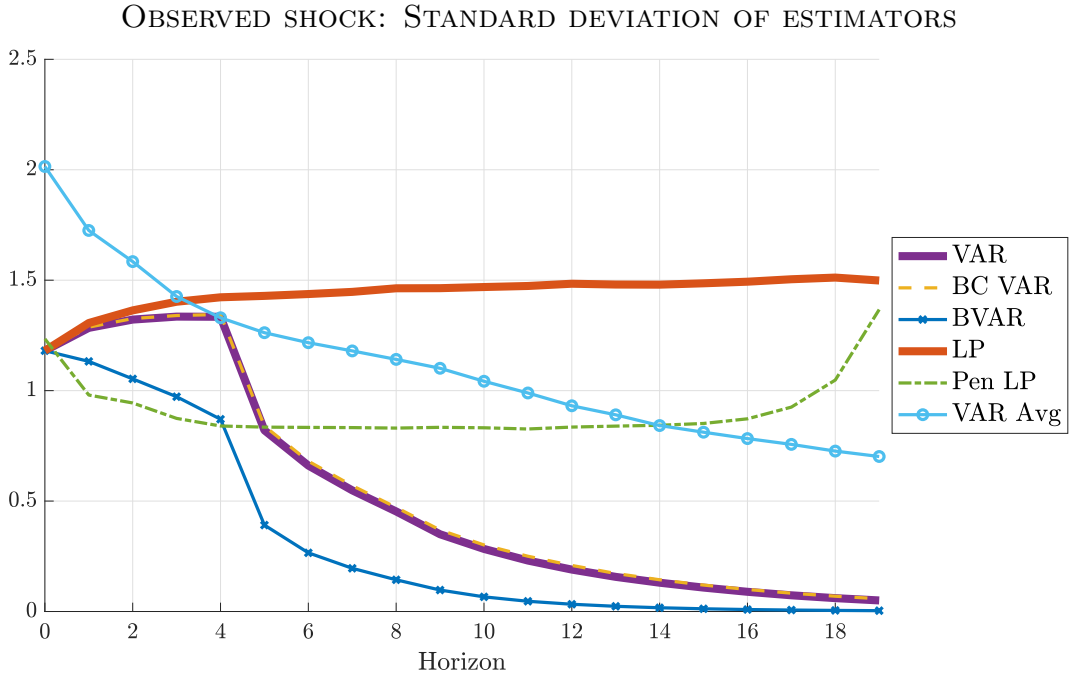

Figure 5: Median (across DGPs) of standard deviation of the different estimation procedures, relative to $\sqrt{\frac{1}{20}\sum_{h=0}^{19}\theta_h^2}$.

least-squares LP and VAR being thick lines. Before taking the median, we cancel out the units of the response variables by dividing the bias and standard deviation by $\sqrt{\frac{1}{20}\sum_{h=0}^{19}\theta_h^2}$, i.e., the root mean squared value of the *true* impulse response function out to horizon 19. Note that the scale of the vertical axis differs between the bias and standard deviation plots.

The figures show that least-squares LP and VAR estimators have similar bias and variance at horizons $h \leq p = 4$, but not at longer horizons $h > p$. The median LP bias is close to zero at all horizons, yet the variance is high and does not decrease with the horizon.[16] In contrast, the median VAR variance decays quickly toward zero as a function of the horizon, at the cost of elevated bias at intermediate horizons (i.e., horizons h that are moderately larger than the estimation lag length p). We show in Supplemental Appendix F that the near-equivalent performance of least-squares LP and VAR estimators at horizons $h \leq p$ continues to hold when $p = 8$. All these observations are consistent with the theoretical results in Section 2, Schorfheide (2005), and Plagborg-Møller & Wolf (2020b).

Figure 6 shows that least-squares VAR is preferred over least-squares LP for almost all horizons and almost all researcher loss functions; the only exceptions are when the weight ω on squared bias in the loss function (2) is near 1 and the horizon h is intermediate. The figure uses grey shadings to indicate the *fraction* of DGPs for which, at a given h and ω, the loss for least-squares LP is smaller than that for least-squares VAR. For $h \leq p = 4$, the VAR is preferred almost uniformly because its variance is slightly smaller; however, the difference is negligible. In the region with a non-trivial bias-variance trade-off – i.e., $h > 4$ – a stark picture emerges: For the researcher to prefer the LP method, her relative weight on squared bias will usually have to be at least four times larger than that on variance. This is because, as already revealed by Figures 4 and 5, the relatively low bias of the LP method comes at a very steep cost in terms of increased sampling variance. Viewed through the lens of the analytical illustration in Figure 2, the findings coming out of our large number of empirically disciplined DGPs are consistent with a variant of our simple model (1) with (i) moderate persistence (ρ) and (ii) non-zero but limited VAR mis-specification (α). It is important to note that, if we had instead considered DGPs with higher average persistence, the bias of the VAR procedure at longer horizons would have likely been worse.

[16] Kilian & Kim (2011) find in simulations that LP does not have lower bias than VAR estimators, but they consider a different variant of LP that uses an auxiliary VAR to identify the structural shocks.

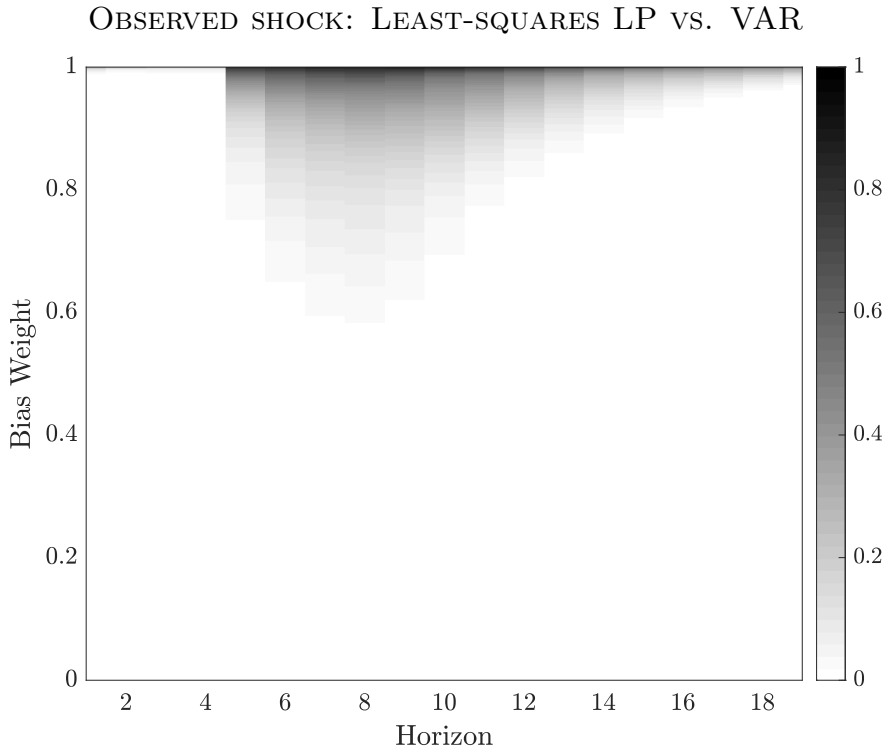

Figure 6: Fraction of DGPs for which the least-squares LP estimator has a lower loss than the VAR estimator. Darker areas correspond to regions where LP is preferred more often. Horizontal axis: impulse response horizon h. Vertical axis: weight ω on squared bias in the loss function (2). The loss function is normalized by the scale of the impulse response function, as in Figures 4 and 5. The impact horizon $h = 0$ is omitted due to numerical equivalence between the estimators.

5.2 Shrinkage dramatically lowers variance, at some cost of bias

Our second main takeaway is that shrinkage methods often allow a substantial reduction in variance, at a moderate cost in terms of bias. In this subsection we focus on two shrinkage methods in particular: penalized LP, which shrinks the conventional least-squares LP estimates towards smooth impulse response functions, and the Bayesian VAR (BVAR) estimator, which shrinks the least-squares VAR coefficients towards independent white noise.

The dash-dotted and x-marked lines in Figure 4 show that penalized LPs and BVARs have uniformly higher bias than least-squares LPs and VARs, respectively. This is an inevitable cost of shrinkage, since the stylized prior information used for the shrinkage is never completely accurate. For penalized LP, the increase in bias is particularly stark at short horizons, since this estimator uses the longer-horizon impulse response estimates to inform the estimation of the short-run impulse responses (through the prior belief in smoothness).

Figure 5, however, shows that shrinkage can lead to large reductions in variance: The

median standard deviation of penalized LP lies everywhere below that of standard LP (and is in fact the lowest of all methods for $h \leq p$), and the standard deviation of the BVAR estimator is always strictly below that of least-squares VAR.

Figures 7 and 8 depict the head-to-head loss-based comparison of penalized LP vs. least-squares LP, and of BVAR vs. least-squares VAR, respectively. Given the relatively high bias of penalized LP at short horizons, the trade-off for $h \leq p$ is non-trivial: the researcher's optimal choice of least-squares LP vs. penalized LP is sensitive to the bias weight ω. At intermediate horizons, however, the case for penalized LP becomes overwhelming: substantial decreases in variance are achieved at the cost of moderate increases in bias, so a preference for conventional LP methods requires an overwhelming concern for bias.[17] In the comparison of BVARs and least-squares VARs, the bias-variance trade-off is similarly non-trivial, as shrinkage offers a very competitive reduction in variance in return for its increased bias, so justification of least-squares VAR requires relatively high bias weight ω. Note, however, that if we focus on the longest horizons in Figures 4 and 5, the difference between least-squares VAR and BVAR is minimal, unlike when comparing the two LP procedures. This is because any kind of VAR-estimated impulse response function tends to zero as $h \to \infty$.

5.3 No method dominates, but shrinkage is generally welcome

Our third takeaway is that, for a fixed loss function, no single method is best at *all* horizons. Nevertheless, regardless of the horizon, one of the LP or VAR shrinkage procedures is preferred by most loss functions, unless bias receives a very high weight.

Figure 9 shows the optimal estimation method as a function of the horizon h and the bias weight ω. We report the estimation method that minimizes the *average* loss (2) across DGPs, after normalizing the loss to cancel out units as in Figures 4 and 5.

The attractive performance of the shrinkage estimators is evident visually, as most of the figure is either cross-hatched green (penalized LP) or solid-dotted blue (BVAR). If interest centers solely on very short impulse response horizons, penalized LP is almost always the best choice, except when ω is very close to 1 (i.e., an almost exclusive concern for bias). At horizons longer than the estimation lag length $p = 4$, some kind of VAR method is generally preferable. In particular, BVAR shrinkage is optimal if the weight on bias is not too large, while least-squares VAR or bias-corrected VAR are preferred if the weight on bias is high.

[17] Penalized LP performs similarly to least-squares LP at horizon $h = 19$. The reason is mechanical: The Barnichon & Brownlees (2019) smoothness penalty described in Appendix B.1 affects estimates near the "boundary" (i.e., the largest penalized horizon) less than at the other horizons.

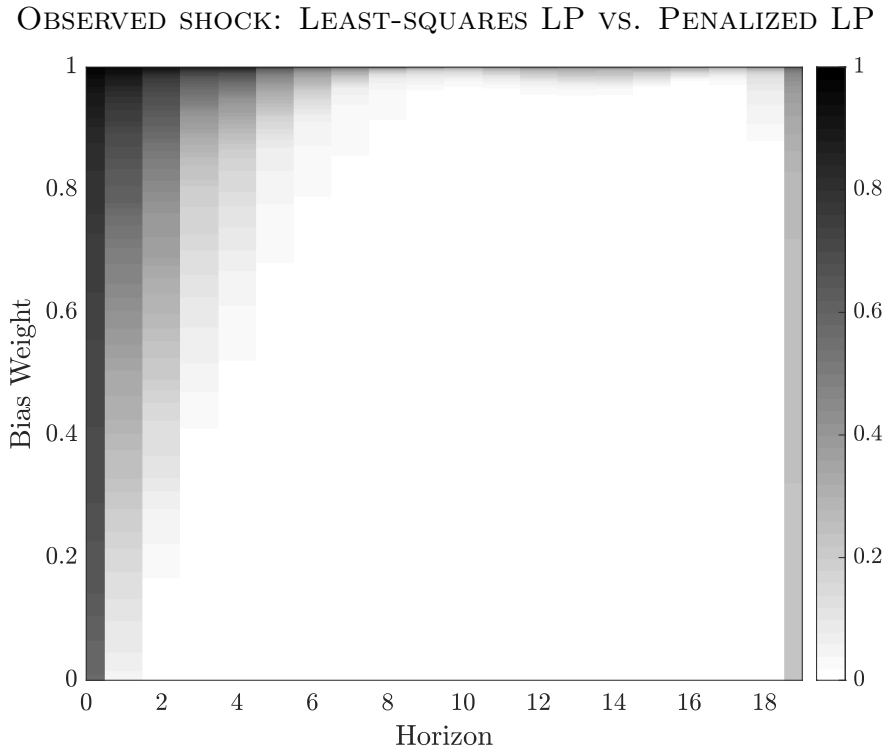

Figure 7: Fraction of DGPs for which the least-squares LP estimator has a lower loss than penalized LP. Darker areas correspond to regions where least-squares LP is preferred more often. See caption for Figure 6.

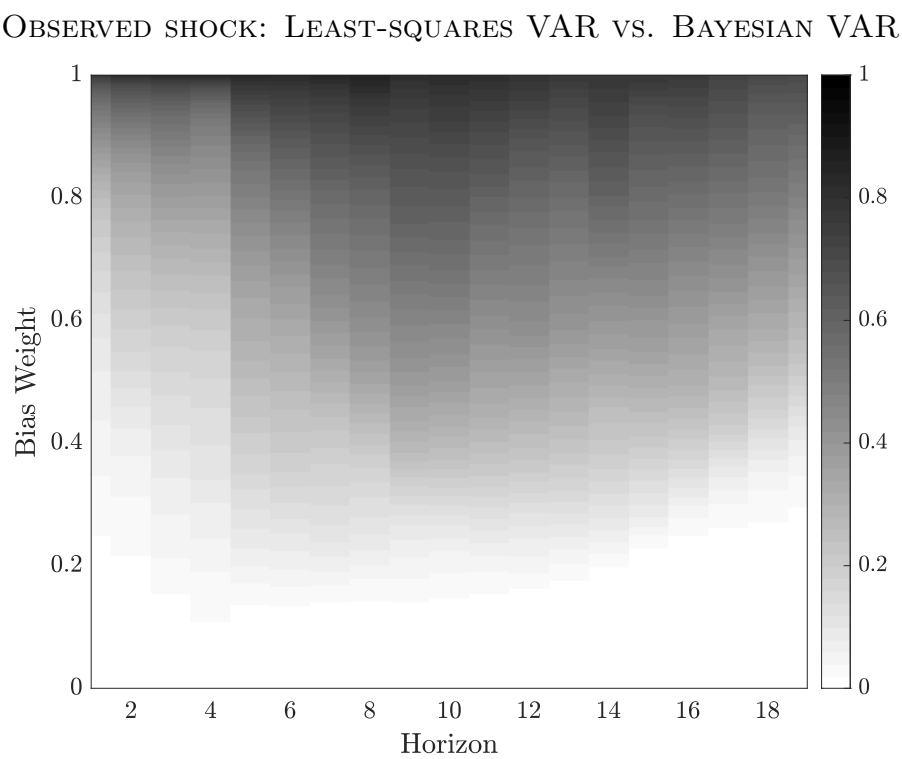

Figure 8: Fraction of DGPs for which the least-squares VAR estimator has a lower loss than the BVAR estimator. Darker areas correspond to regions where least-squares VAR is preferred more often. See caption for Figure 6. The impact horizon $h = 0$ is omitted due to numerical equivalence between the estimators.

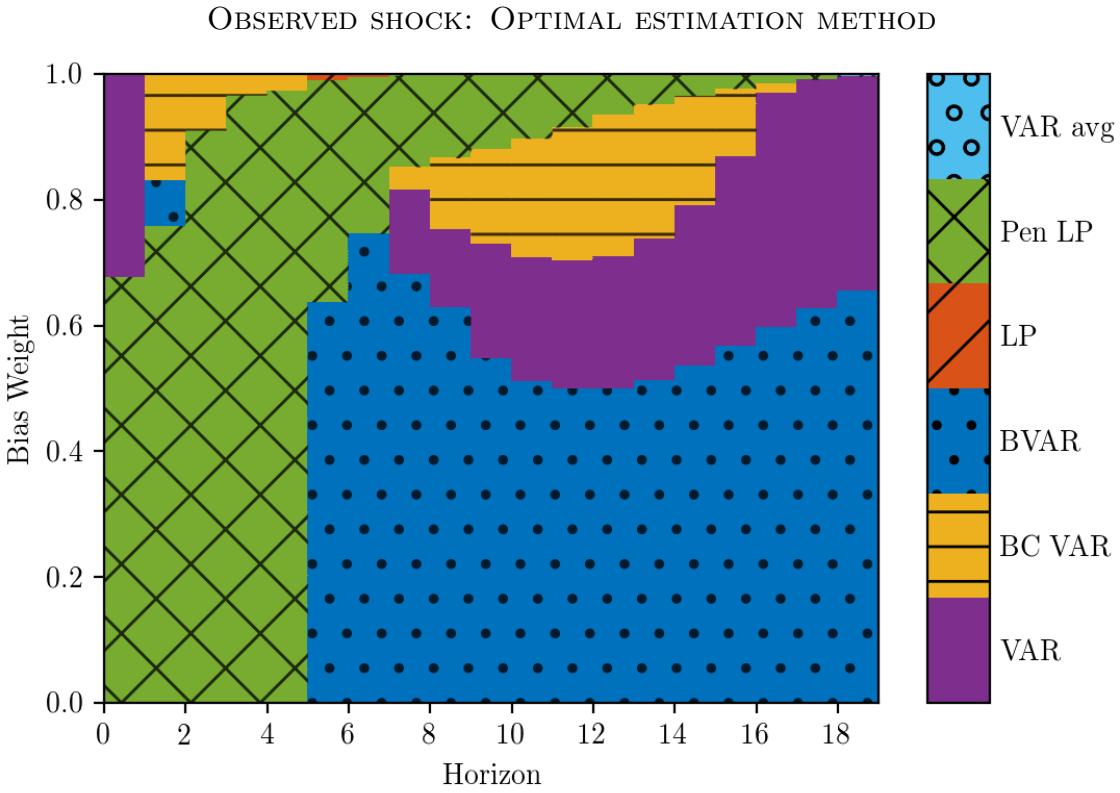

Figure 9: Method that minimizes the average (across DGPs) loss function (2). Horizontal axis: impulse response horizon. Vertical axis: weight on squared bias in loss function. The loss function is normalized by the scale of the impulse response function, as in Figures 4 and 5. At $h = 0$, VAR and LP are numerically identical; we break the tie in favor of VAR.

In Supplemental Appendix F we show that the preceding conclusions remain valid when the estimation lag length is increased to 8, except in this case penalized LP is optimal for most ω at horizons $h \leq p = 8$.

Though bias-corrected VAR (yellow with horizontal lines) looks attractive in Figure 9 when $\omega \geq 0.8$, Figure 4 shows that in practice the reduction in bias relative to the conventional least-squares VAR estimator is small. This finding is due to the modest persistence of our DGPs, cf. Section 3.4.[18]

Figure 9 also shows that least-squares LP (orange with diagonal lines) and the VAR model averaging estimator (light blue with hollow circles) are essentially never optimal.[19] Least-squares LP is essentially dominated because its low bias is not much lower than that of penalized LP at intermediate horizons, so the high variance of least-squares LP is difficult

[18]The Pope (1990) bias correction corrects for bias due to persistence, not due to mis-specification.

[19]Least-squares LP is in fact strictly preferred in a very thin region with intermediate p and $\omega \approx 1$.

to justify. The reason why VAR model averaging does poorly is due to its high median standard deviation relative to other VAR-based estimators, cf. Figure 5. Closer inspection reveals that the high standard deviation is a consequence of the estimator having a very fat-tailed sampling distribution, with a non-negligible probability of erratic estimates.[20]

5.4 SVAR-IV is heavily biased, but has relatively low dispersion

Our last takeaway is concerned with IV/proxy estimation procedures. Among the robust "internal instruments" procedures, the bias-variance trade-off is very similar to that discussed above for the case of an observed shock. The "external instruments" SVAR-IV procedure, however, contributes very starkly to the trade-off: It can be severely biased due to its lack of robustness to non-invertibility, but at the same time it has substantially lower dispersion than the "internal instruments" procedures.

Figures 10 and 11 show the median bias and interquartile range of the various IV estimators. We here report median bias instead of (mean) bias and the interquartile range instead of standard deviation, because the sampling distributions of the IV impulse response estimates is fat-tailed, as is often the case with moderately strong IVs.[21]

If we ignore the dotted burgundy line for SVAR-IV, these figures are qualitatively similar to those presented in Section 5.1. However, SVAR-IV stands out by exhibiting especially high median bias at short horizons, but low interquartile range at short and intermediate horizons. This is consistent with the existing theoretical work referenced in Section 4: Unlike the "internal instruments" procedures, SVAR-IV is asymptotically biased when the shock is not invertible, and we saw in Section 3.4 that the degree of invertibility is generally low in our DGPs. On the other hand, the SVAR-IV procedure has fewer parameters to estimate, as it excludes the IV z_t from the reduced-form VAR regression, causing a reduction in dispersion relative to the other procedures. Though we view the high median bias of SVAR-IV across our DGPs as a worrying finding, its low dispersion is intriguing and may in some cases actually trump the bias concerns.[22]

[20] We use Hansen's (2016) code off the shelf. It would be interesting to investigate whether the procedure could be modified to avoid erratic estimates.

[21] For completeness, (mean) bias and standard deviation figures are reported in Supplemental Appendix H.

[22] Since our estimand of interest is the normalized impulse response function (8), the SVAR-IV bias in Figure 10 does *not* reflect the scaling bias discussed by Plagborg-Møller & Wolf (2020a) for the estimation of forecast variance decompositions. For the purposes of forecast variance estimation, the performance of SVAR-IV is likely to be substantially worse than the situation considered in the present paper.

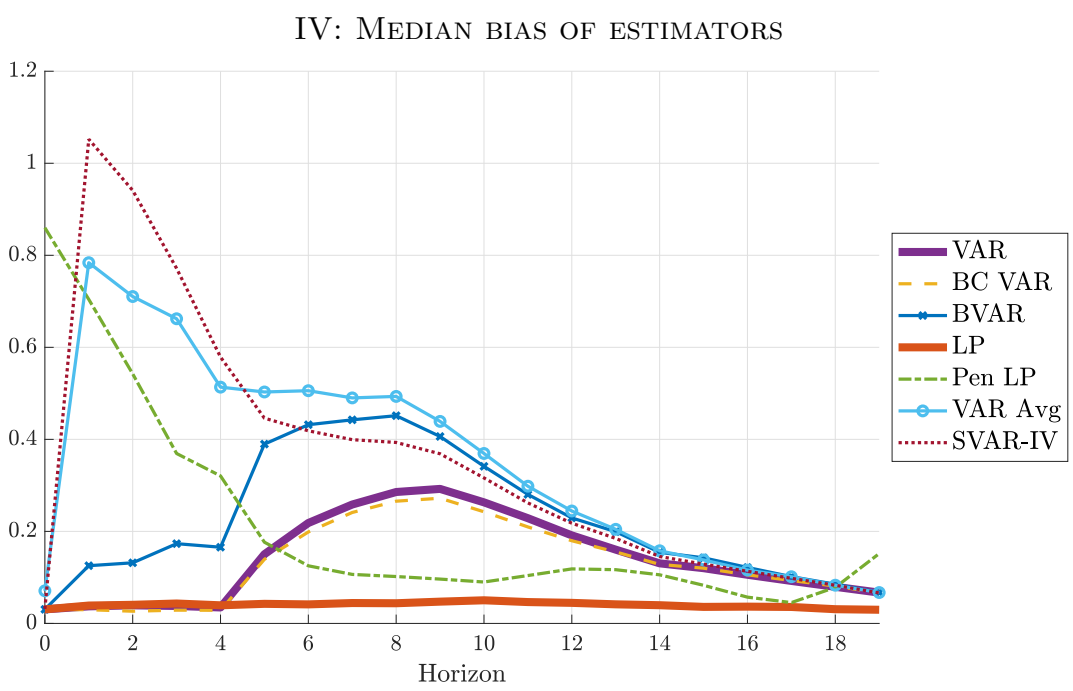

Figure 10: Median (across DGPs) of absolute median bias of the different estimation procedures, relative to $\sqrt{\frac{1}{20}\sum_{h=0}^{19}\theta_h^2}$.

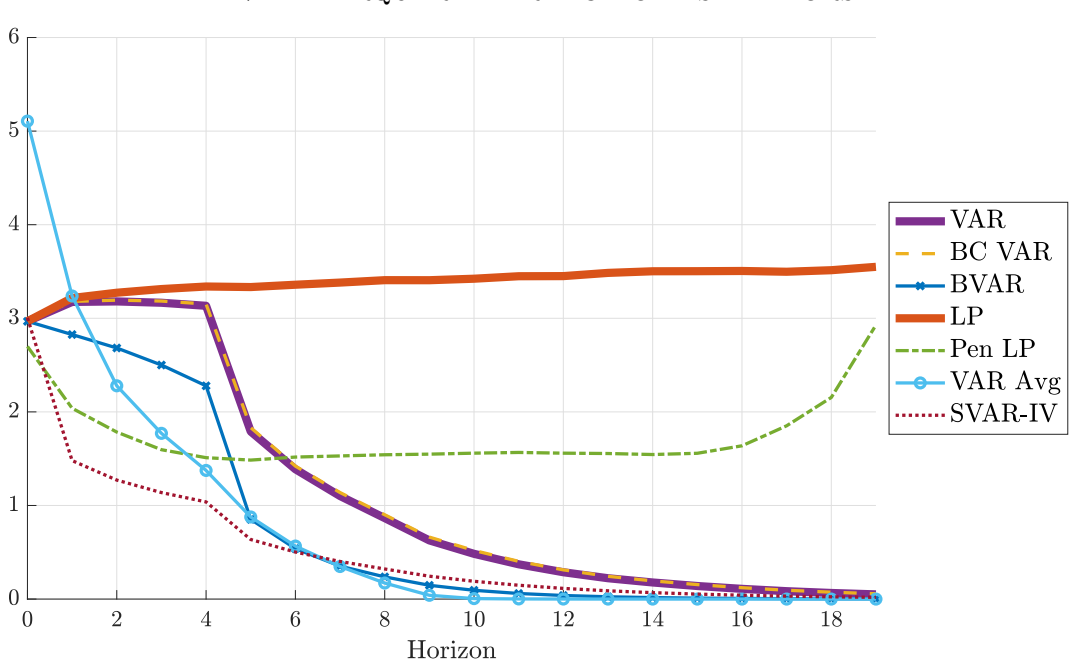

Figure 11: Median (across DGPs) of interquartile range of the different estimation procedures, relative to $\sqrt{\frac{1}{20}\sum_{h=0}^{19}\theta_h^2}$.

6 Conclusion and directions for future research

We conducted a large-scale simulation study of the performance of LP and VAR structural impulse response estimators, as well as several variants of these methods. By drawing thousands of DGPs from an empirically estimated encompassing dynamic factor model, we ensured that our results apply to a wide range of settings faced by applied researchers. We drew the following four conclusions.

1. As predicted by theory, there is a non-trivial bias-variance trade-off between least-squares LP and VAR estimators. Empirically relevant DGPs are unlikely to admit finite-order VAR representations, so mis-specification of VAR estimators is a valid concern, as discussed by Ramey (2016) and Nakamura & Steinsson (2018). Reducing this bias by going all the way to LP estimation, however, tends to come at a surprisingly steep cost in terms of increased sampling variance. An important caveat is that we focus on DGPs with stationarity-transformed variables, so the persistence of the data is moderate. Future research could explore whether DGPs with near-unit root persistence impart much higher bias in VAR estimators.[23]

2. Shrinkage procedures – such as the penalized LP procedure of Barnichon & Brownlees (2019) and Bayesian VAR estimation – can dramatically lower the variance at the cost of moderately larger bias. The researcher must have an overwhelming concern about bias to justify the use of conventional least-squares estimators over shrinkage procedures. The literature on shrinkage estimators in macroeconometrics is not yet saturated, and we believe our findings are encouraging for further work in this area.

3. No method dominates at all response horizons, though one of the shrinkage procedures is almost always optimal or near-optimal. Penalized LP is especially attractive at short horizons, while BVAR estimation is attractive at intermediate and long horizons. In future work, we hope to investigate how the optimal choice of estimation method can be guided by diagnostics computed in the data set at hand. It is not obvious that data-dependent model selection criteria will perform well; for example, we found the VAR model averaging procedure of Hansen (2016) to have some trouble accurately estimating the model weights.

[23] Montiel Olea & Plagborg-Møller (2020) recommend the use of lag-augmented LP for *confidence interval* construction when the data is persistent.

4. In the case of IV identification, the popular SVAR-IV (or proxy-SVAR) procedure can be badly biased at short horizons, but it has substantially lower dispersion at all horizons than "internal instruments" procedures such as LP-IV. The high (median) bias of SVAR-IV is due to its lack of robustness to non-invertibility, which is a pervasive and realistic feature of our DGPs. If possible, it would be useful to develop alternative non-invertibility-robust estimation procedures that capture some of the variance improvement enjoyed by SVAR-IV, perhaps by exploiting known features of the time series process for the IV.

Appendix A DGPs

This section provides further details on our various DGPs and structural impulse response function estimands. Appendix A.1 discusses how we estimate the encompassing factor model (5)–(7), and Appendix A.2 then gives details on how we select individual DGPs from this encompassing model. In Appendices A.3 and A.4, we provide details on our IV and recursive shock estimands.

A.1 DFM estimation

Our calibration of the parameters of the DFM uses the empirical estimates of Stock & Watson (2016) and Lazarus et al. (2018). We use the same data as those papers. The DFM is estimated on quarterly data of 207 different time series, spanning a wide variety of real activity, price, as well as financial variables. Series are temporally averaged (when necessary), seasonally adjusted, and transformed to ensure integration of order zero; outliers and low-frequency trends are removed; and each series is standardized to have zero mean and unit variance. The resulting data set spans 1959Q1–2014Q4. A detailed list of all variables and their categories is provided in Table 1 and the Data Appendix of Stock & Watson (2016).

As in Lazarus et al. (2018), we choose six static factors as well as two lags each in the factor equation (5) and the measurement error equation (7). Given this model specification, the factors f_t and loading matrix Λ are estimated using principal components, and the parameters of the measurement error equations are estimated using least-squares. Detailed replication codes are provided in our GitHub repository.

A.2 DGP selection

For each structural identification approach (observed shock, IV, and recursive) and each policy experiment (fiscal or monetary), we randomly draw 3,000 sets of macro observables $\bar{w}_t$ from X_t, where $n_{\bar{w}} = 5$, in line with standard practice in the applied macroeconometrics literature. For fiscal experiments, one of the five observables is always real government expenditure, while for monetary experiments we always include the federal funds rate. For the other four observables, we always draw one output series (i.e., categories 1-3 in the classification in Table 1 of Stock & Watson (2016)) and one price series (category 6); the remaining two variables are drawn randomly from all categories. The response variable y_t is then drawn randomly from those four series.

Our definition of the structural shock of interest, ε_{1t}, ensures that it has the largest possible contemporaneous effect on the normalization variable i_t. Letting $\eta_t = H\varepsilon_t$, the shock is thus defined through the solution of the following constrained optimization problem:

$$\max_{H} \quad \bar{\Lambda}_{\iota_i,\bullet} H e_1 \quad \text{s.t.} \quad HH' = \Sigma_\eta,$$

where $\Sigma_\eta \equiv \text{Var}(\eta_t)$, and e_1 selects the first column of H.[24] The problem above can be simplified if we decompose H as $H = BR$, where $\Sigma_\eta = BB'$ with B lower-triangular. We can then write the problem as

$$\max_{R_{\bullet,1}} \quad \bar{\Lambda}_{\iota_i,\bullet} B R_{\bullet,1} \quad \text{s.t.} \quad R'_{\bullet,1} R_{\bullet,1} = 1.$$

This yields $R_{\bullet,1} = B'\bar{\Lambda}'_{\iota_i,\bullet}(\bar{\Lambda}_{\iota_i,\bullet}\Sigma_\eta\bar{\Lambda}'_{\iota_i,\bullet})^{-1/2}$, and hence $H_{\bullet,1} = \Sigma_\eta\bar{\Lambda}'_{\iota_i,\bullet}(\bar{\Lambda}_{\iota_i,\bullet}\Sigma_\eta\bar{\Lambda}'_{\iota_i,\bullet})^{-1/2}$.

Finally note that, since the same normalization variables i_t are included for each fiscal and monetary experiment, the resulting structural shocks $\varepsilon_{1,t}$ are the same for each of the randomly drawn DGPs.

A.3 IV process calibration

To calibrate the IV equation (9), we use actual, empirically measured shock series. Specifically, we use the series of Romer & Romer (2004) and Ben Zeev & Pappa (2017) for monetary and fiscal simulation DGPs, respectively. We choose the noise level σ_ν^2 for the simulated IV z_t to match the explanatory power of the DFM's factor innovations for these two real-world shock series. Though this could be done in various ways, we proceed as follows. First, we regress each shock series $\hat{z}_t$ on leads and lags of the estimated factor innovations $\hat{\eta}_t$:

$$\hat{z}_t = \hat{c} + \sum_{l=-p}^{p} \hat{a}'_l \hat{\eta}_{t-l} + \text{residual}_t.$$

We choose the number p of leads and lags based on the BIC. If the underlying structural shock measured by $\hat{z}_t$ is recoverable with respect to the factors in the DFM (Plagborg-Møller & Wolf, 2020a), then the R^2 in the above two-sided regression will pin down the signal-to-noise ratio of $\hat{z}_t$. We calibrate the signal-to-noise ratio of the simulation IV (9) to be consistent

[24] Note that, by construction, this optimization problem only pins down the first column of H, since that column is the only relevant one for the structural impulse responses to $\varepsilon_{1,t}$. The remaining columns then just need to satisfy the variance-covariance constraint.

with this estimated signal-to-noise ratio. Specifically, $\sigma_\nu^2 = \frac{1}{R^2} - 1$. Encouragingly, this approach yields IVs with empirically realistic F-statistics (see Table 1).

A.4 Recursive shock estimand

To characterize the recursive shock estimand, we first express the reduced-form VAR representation (10) in terms of factor model primitives. We can reduce the encompassing factor model to only include the selected observables $w_t = \bar{w}_t$ as

$$\begin{aligned} f_t &= \Phi(L) f_{t-1} + H\varepsilon_t, && \text{(A.1)} \\ w_t &= \Lambda^w f_t + v_t^w, && \text{(A.2)} \\ v_t^w &= \Delta^w(L) v_{t-1} + \Xi^w \xi_t^w. && \text{(A.3)} \end{aligned}$$

Combining (A.2) and (A.3), we get

$$[I - \Delta^w(L)L] w_t = [I - \Delta^w(L)L]\Lambda^w f_t + \Xi^w \xi_t^w. \tag{A.4}$$

We can then write the system in state-space form (Fernández-Villaverde et al., 2007),

$$\begin{aligned} s_{t+1} &= A s_t + B\zeta_t, && \text{(A.5)} \\ w_t^* &= C s_t + D\zeta_t, && \text{(A.6)} \end{aligned}$$

where $s_t \equiv [f_t', f_{t-1}', \ldots, f_{t-q}']'$, $w_t^* = [I - \Delta^w(L)L] w_t$, $\zeta_t \equiv (\varepsilon_{t+1}', \xi_t')'$, A is the companion matrix for the lag polynomial $\Phi(L)$, $B \equiv \left(\begin{smallmatrix} H' & 0 & \cdots \\ 0 & 0 & \cdots \end{smallmatrix}\right)'$, $C \equiv [I, -\Delta_1^w, \ldots, -\Delta_q^w] \left(\begin{smallmatrix} \Lambda^w & 0 & \cdots \\ 0 & \Lambda^w & \cdots \\ \vdots & \vdots & \ddots \end{smallmatrix}\right)$, and $D \equiv (0, \Xi^w)$.[25]

Given the state-space system (A.5)–(A.6), straightforward manipulations give the standard reduced-form VAR representation for w_t^*:

$$w_t^* = C[I - (A - KC)L]^{-1} K w_{t-1}^* + u_t^{w^*}, \tag{A.7}$$

[25] This state-space form is valid if $p \leq q+1$, which is true in our encompassing factor model. When $p > q+1$, we need to accordingly change $s_t = [f_t', f_{t-1}', \ldots, f_{t-p+1}']'$ and $C = [I, -\Delta_1^w, \ldots, -\Delta_q^w] \left(\begin{smallmatrix} \Lambda^w & & & \\ & \Lambda^w & & \\ & & \ddots & \\ & & & 0 \end{smallmatrix}\right)$.

where $u_t^{w^*}$ is the Wold innovation for w_t^*, $\text{Var}(u_t^{w^*}) = C\Sigma C' + DD'$, and (K, Σ) satisfy

$$\begin{aligned} \Sigma &= A\Sigma A' + BB' - (A\Sigma C' + BD')(C\Sigma C' + DD')^{-1}(A\Sigma C' + BD')', \\ K &= (A\Sigma C' + BD')(C\Sigma C' + DD')^{-1}. \end{aligned}$$

Given (A.7) as the VAR(∞) representation for $w_t^* = [I - \Delta^w(L)L]w_t$, we obtain the VAR($\infty$) representation for w_t itself as[26]

$$\{I - C[I - (A - KC)L]^{-1}KL\}[I - \Delta^w(L)L]w_t = u_t^w.$$

The representation (10) then gives the recursive shock estimand (11), as already discussed in Section 3.2.

[26]Note that these two Wold innovations are equivalent, i.e., $u_t^{w^*} \equiv u_t^w$.

Appendix B Estimation procedures

This section provides further details on the various estimation procedures sketched in Section 4 and used for the simulation study in Section 5. Appendix B.1 begins with LP-based approaches, while Appendix B.2 considers VAR-based approaches.

B.1 LP-based approaches

Least-squares LP. The least-squares LP estimator of the impulse response at horizon h is based on the coefficient $\hat{\beta}_h$ in the h-step-ahead OLS regression

$$y_{t+h} = \hat{\mu}_h + \hat{\beta}_h x_t + \hat{\zeta}_h q_t + \sum_{\ell=1}^{p} \hat{\varphi}_{h,\ell} w_{t-\ell} + \text{residual}_h, \tag{B.1}$$

that is, we regress on the variable x_t, with controls given by the vector q_t as well as p lags of all of the data w_t. The estimands of Section 3.2 can now be recovered as follows:

1. **Observed shock.** We set x_t equal to the observed shock ε_t and omit the contemporaneous controls q_t (we still control for lagged data).[27] To estimate the relative impulse response defined in Section 3.2, we divide $\hat{\beta}_h$ by the corresponding coefficient in a regression as in (B.1) that has i_t on the left-hand side instead of y_{t+h}.

2. **IV.** We estimate a Two-Stage Least Squares (2SLS) version of (B.1), setting x_t equal to the normalization variable i_t, and instrumenting for this variable with the IV z_t. We omit q_t in this specification (but still include lagged controls). This is numerically the same as doing a LP of y_{t+h} on z_t (with lagged controls), and dividing this coefficient by the LP coefficient in a regression of i_t on z_t (with lagged controls), see Stock & Watson (2018) and Plagborg-Møller & Wolf (2020b).

3. **Recursive identification.** x_t is the innovation variable, while q_t are the variables ordered before x_t in the identification scheme (Plagborg-Møller & Wolf, 2020b).

Penalized LP. Barnichon & Brownlees (2019), taking a cue from Shiller (1973), propose to lower the variance of LP by exploiting a prior belief in smoothness of the impulse response function across horizons. This comes at the cost of increasing the bias, unless the true impulse response function is perfectly consistent with the imposed type of smoothness.

[27] The lags are not needed for consistency in this case, but they often improve efficiency.

The method works by flexibly parametrizing the impulse response function and then penalizing its jaggedness. First, we model the impulse response function as a flexible parametric function of the horizon h: $\beta_h = \sum_{k=1}^{K} b_k B_k(h)$, where $B_1(\cdot), \ldots, B_K(\cdot)$ are K basis functions. These could for example be chosen to be simple polynomials, but we follow Barnichon & Brownlees (2019) and use the computationally attractive cubic B-spline basis functions.[28] Second, we choose the coefficients $b_1, \ldots, b_K$ to minimize the sum of squared residuals, subject to a penalty function $P(\cdot)$ that discourages jagged-looking impulse responses:

$$\min_{b,\mu,\zeta,\varphi} \sum_{h=0}^{\bar{h}} \sum_{t=p+1}^{T-h} \left(y_{t+h} - \mu_h - \sum_{k=1}^{K} b_k B_k(h) x_t - \zeta_h q_t - \sum_{\ell=1}^{p} \varphi_{h,\ell} w_{t-\ell} \right)^2 + \lambda P(b_1, \ldots, b_k)$$

The final un-normalized impulse response estimate is $\hat{\beta}_h = \sum_{k=1}^{K} \hat{b}_k B_k(h)$, where $\hat{b}_1, \ldots, \hat{b}_K$ solve the above minimization. We penalize impulse responses up to horizon $\bar{h} = 19$.

We choose the penalty function as $P(b_1, \ldots, b_k) = \sum_{k=4}^{K} \Delta^3 b_k$, where Δ denotes the finite difference operator. Due to the properties of B-splines, this penalty function penalizes the third-order "derivative" of β_h with respect to h, and thus shrinks the impulse response function β_h towards a quadratic function of h. The scalar $\lambda \geq 0$ governs the amount of shrinkage. If $\lambda = 0$, then the above minimization reduces to unrestricted, horizon-by-horizon least-squares LP, defined above. As $\lambda \to \infty$, the estimated impulse response function converges to a quadratic function of the horizon h. For $\lambda \in (0, \infty)$, the estimated impulse response function is intermediate between the jagged unconstrained LP estimator and a perfect quadratic function of the horizon. To choose λ in a data-dependent way, we use 5-fold cross-validation as in Barnichon & Brownlees (2019). Notice that we do not penalize the coefficients on the control variables in the LP.

Given the un-normalized shrinkage LP impulse response estimate $\hat{\beta}_h$, we obtain the relative impulse response by dividing with the *least-squares* LP estimate of the impact impulse response of i_t. The three identification schemes are implemented in the same way as for least-squares LP, described above. In particular, the IV identification case is handled by setting $w_t = (z_t, \bar{w}_t')'$ and $x_t = z_t$ with no contemporaneous controls q_t in the penalized LP, and then normalizing with the least-squares LP estimate of the impact impulse response of i_t to the structural shock.

[28] We also set K equal to $\bar{h} + 2$, where the left-most inner knot of each $B_k(h)$ goes from -2 to $\bar{h} - 1$ to cover all the horizons of interest.

B.2 VAR-based approaches

LEAST-SQUARES VAR. The least-squares VAR estimator (henceforth just "VAR") flexibly captures the first p sample autocovariances through the equation-by-equation OLS regression

$$w_t = \hat{\nu} + \sum_{\ell=1}^{p} \hat{A}_\ell w_{t-\ell} + \hat{u}_t. \tag{B.2}$$

The *reduced-form* impulse responses $\hat{\Psi}_0, \hat{\Psi}_1, \ldots$ are obtained through the familiar recursive formula

$$\hat{\Psi}_0 = I, \quad \hat{\Psi}_h = \sum_{\ell=1}^{\min\{p,h\}} \hat{A}_\ell \hat{\Psi}_{h-\ell}, h = 1, 2, \ldots \tag{B.3}$$

Let $\hat{\Sigma}$ be the sample variance-covariance matrix of the least-squares residuals $\hat{u}_t$, and let $\hat{\Sigma} = \hat{C}\hat{C}'$ be the Cholesky decomposition, where $\hat{C}$ is lower triangular with positive diagonal elements. The un-normalized "structural" impulse response estimate $\hat{\delta}_h$ at horizon h is given by the (j, k) element of the matrix $\hat{\Psi}_h \hat{C}$. Here the index j equals the index of the response variable y_t in the data vector w_t. The choice of "innovation" index k and the ordering of variables in w_t both depend on the choice of structural identification scheme, as described below. Finally, the desired relative impulse response estimate is obtained by dividing $\hat{\delta}_h$ by the impact impulse response of i_t with respect to the k-th innovation, i.e., the (ℓ, k) element of $\hat{\Psi}_0 \hat{C}$, where ℓ is the index of i_t in the vector w_t.

The estimands of Section 3.2 can now be estimated as follows:

1. **Observed shock.** The shock $\varepsilon_{1,t}$ is ordered first in w_t, and we compute responses to the first innovation ($k = 1$).

2. **IV.** We initially consider an "internal instruments" approach as in Ramey (2011). That is, we include the IV z_t in the data vector w_t, order the IV first, and compute responses with respect to the first innovation ($k = 1$). In essence, this approach treats the IV z_t *as if* it were equal to the shock (plus potential lagged variables). Plagborg-Møller & Wolf (2020b) prove that this approach leads to consistent estimates of *relative* impulse responses, even if the IV is merely a *noisy* measure of the actual shock, as in our simulation DGP.

3. **Recursive identification.** The ordering of variables in w_t equals the ordering of the desired population impulse response estimand (cf. Section 3.2), and we compute responses to the innovation of variable i_t (i.e., k is the index of i_t in vector w_t).

In the case of IV identification, there exists a popular alternative to the above-mentioned "internal instruments" estimator, namely the SVAR-IV (or "proxy-SVAR") estimator of Stock (2008), which was further developed by Stock & Watson (2012), Mertens & Ravn (2013), and Gertler & Karadi (2015). The SVAR-IV estimator requires an extra assumption to be consistent: the identified shock needs to be *invertible* with respect to the variables w_t included in the reduced-form VAR model, where – importantly – the data vector w_t now *excludes* the IV z_t.

Estimating the VAR in w_t and computing reduced-form impulse responses as in (B.2)–(B.3), the un-normalized SVAR-IV impulse response estimator $\hat{\delta}_h$ is given by the j-th element of $\hat{\Psi}_h\hat{\gamma}$, where j is the index of y_t in w_t, and $\hat{\gamma}$ is the sample covariance vector of the reduced-form VAR residuals $\hat{u}_t$ and the IV z_t. The relative impulse response estimate of interest is obtained by dividing $\hat{\delta}_h$ by the ℓ-th element of $\hat{\Psi}_h\hat{\gamma}$, where ℓ is the index of i_t in w_t. The SVAR-IV estimator tends to have lower variance than the "internal instruments" VAR estimator, but higher bias if the shock of interest is non-invertible. By excluding the IV z_t from the data vector w_t, the SVAR-IV approach has to estimate fewer reduced-form VAR coefficients, which reduces the variance of the impulse response estimator. However, the SVAR-IV estimator is generally inconsistent unless the shock is invertible (Forni et al., 2019; Miranda-Agrippino & Ricco, 2019; Plagborg-Møller & Wolf, 2020a,b).

The various alternative VAR-based estimation procedures below will only consider the more robust "internal instruments" IV approach. That is, the only SVAR-IV estimation procedure that we consider uses the least-squares VAR coefficient estimates described above.

Bias-corrected VAR. It is well known that VAR coefficient estimates are biased in small samples, usually in the direction of spuriously indicating less persistence than in the true DGP. The bias tends to be larger for more persistent true DGPs.

To remove part of this bias, we follow Kilian (1998) and consider a modification of the least-squares VAR estimator that applies the Pope (1990) asymptotic bias correction to the reduced-form VAR coefficient matrices. Once the bias-corrected reduced-form coefficients $\hat{A}_1, \ldots, \hat{A}_p$ are obtained, the remaining impulse response formulas are the same as in the least-squares case discussed above.

Bayesian VAR. Though VARs tightly constrain long-run impulse responses, they model short-run impulse responses very flexibly. Bayesian VARs seek to reduce the variance of the VAR estimates by exploiting prior information about the time series dynamics. We use the posterior mean estimator implied by a Gaussian VAR(p) model with a Gaussian

prior on the VAR coefficients (the residual variance-covariance matrix $\hat{\Sigma}$ is fixed at the unconstrained least-squares estimate). This effectively shrinks the unrestricted least-squares estimates toward certain baseline coefficient values, described below. Once the reduced-form VAR coefficients are obtained, the remaining formulas are the same as earlier.

The prior is a version of the popular "Minnesota prior", adapted to a stationary setting. As recommended by Kilian & Lütkepohl (2017, chapter 5.2.3), we center the prior of the VAR reduced-form coefficients at zero, which implies the time series are independent white noise processes.[29] We set the prior variance hyper-parameters, which determine how severely we shrink, in accordance with the default choices listed by Canova (2007, chapter 10.2.2).[30]

VAR MODEL AVERAGING. In practice, the appropriate specification of the lag length and variables in a VAR is usually ambiguous. Rather than reporting results for a single VAR specification, an alternative approach is to average across results for multiple specifications.

Hansen (2016) proposes a data-dependent procedure for averaging across impulse responses estimates produced by a collection of different AR and VAR models with different lag lengths. The averaging weights are then chosen to minimize an estimate of the MSE of the resulting averaged estimator. Formally, let $\hat{\delta}_h(r)$ denote the un-normalized, least-squares recursive impulse response estimate at some horizon h for model $r = 1, \ldots, R$. We then consider $R = 40$ candidate models: First, univariate AR models for y_t with lag lengths from $p = 1$ up to $p = 20$; and second, VAR models in w_t with lag lengths from $p = 1$ up to $p = 20$.[31] The VAR model averaging estimator is given by $\sum_{r=1}^{R} \hat{\omega}_r \hat{\delta}_h(r)$, where the weights $\{\hat{\omega}_r\}_{r=1}^{R}$ are chosen to minimize the data-dependent approximated MSE estimate $\hat{M}(\omega_1, \ldots, \omega_R) \approx E[T(\sum_{r=1}^{R} \omega_r \hat{\delta}_h(r) - \delta_h)^2]$, subject to the constraints that all weights are nonnegative and $\sum_{r=1}^{R} \omega_r = 1$. Details of the MSE estimate are given in Hansen (2016, Section 6).[32] To go from the un-normalized weighted impulse response estimate $\sum_{r=1}^{R} \hat{\omega}_r \hat{\delta}_h(r)$ to the desired relative impulse response, we divide by the least-squares impact impulse response estimate of i_t with respect to the identified shock, computed using the Cholesky decomposition of $\hat{\Sigma}$ in the VAR(20) model.

[29] The conventional Minnesota prior for non-stationary data instead shrinks towards independent unit root processes.

[30] Specifically, for each lag ℓ, the prior variance is $0.04/\ell^2$ for lags of the same variable, $0.01/\ell^2$ for lags of other variables, and 4000 for the intercept.

[31] For simplicity, $\hat{\Sigma}$ is fixed at the $p = 20$ VAR estimate, and it is treated as known without error, as in Hansen (2016).

[32] The object of interest, $\hat{\delta}_h(r)$, is a scalar, which allows us to omit the weighting matrix required in the MSE estimate in Hansen (2016).

The VAR model averaging estimator effectively includes LPs among the list of candidate estimators. This is because, as previously mentioned, a VAR(p) model gives similar impulse response estimates as a LP with p lags. In Section 5 we only consider impulse responses up to horizon $h = 20$, so the candidate VAR(20) model will give similar results to a LP with many lagged controls, at all horizons considered (Plagborg-Møller & Wolf, 2020b).

Appendix C Proofs

C.1 Auxiliary lemmas

Before proving Proposition 1, we state and prove some auxiliary lemmas. All lemmas below impose the assumptions of Proposition 1.

Lemma C.1. *Define the process* $\tilde{y}_t \equiv (1-\rho L)^{-1}(\varepsilon_{1,t}+\varepsilon_{2,t})$ *for all* t. *Then for all* $j=1,2$ *and* $\ell \geq 0$,

$$\frac{1}{T}\sum_{t=1}^{T}(y_t-\tilde{y}_t)^2 = O_p(T^{-1}), \quad \frac{1}{T}\sum_{t=1}^{T}(y_t-\tilde{y}_t)\varepsilon_{j,t+\ell} = O_p(T^{-1}).$$

Proof. From the DGP (1) we have

$$y_t-\tilde{y}_t = \frac{\alpha}{\sqrt{T}}B(L)\varepsilon_{2,t}, \quad B(L)\equiv(1-\rho L)^{-1}L.$$

Since the moving average coefficients of $B(L)$ are geometrically decaying, the first statement of the lemma follows from Phillips & Solo (1992, Theorem 3.7) and the assumption of finite variances of the shocks. The second statement of the lemma follows from Chebyshev's inequality and the fact that the process $\varepsilon_{j,t+\ell} \times B(L)\varepsilon_{2,t}$ is serially uncorrelated under our assumptions on the shocks. □

In the following, we define $\tilde{w}_t \equiv (\varepsilon_{1,t}, \tilde{y}_t)'$, where $\tilde{y}_t$ was defined in Lemma C.1. Recall also the definitions of θ_h, $\hat{\beta}_h$, $\hat{\delta}_h$, $\hat{A}$, $\hat{\kappa}$, and the unit vector e_j from Section 2.

Lemma C.2. *We have*

$$\hat{\beta}_h-\theta_h = \frac{1}{T}\sum_{t=2}^{T-h}\left\{\sum_{\ell=1}^{h}\rho^{h-\ell}\varepsilon_{1,t+\ell}+\sum_{\ell=0}^{h}\rho^{h-\ell}\varepsilon_{2,t+\ell}\right\}\varepsilon_{1,t}+o_p(T^{-1/2}). \tag{C.1}$$

Proof. Let $\hat{\varepsilon}_{1,t} \equiv \varepsilon_{1,t}-\hat{b}'w_{t-1}$ be the residual from an auxiliary regression of $\varepsilon_{1,t}$ on w_{t-1}. Using Lemma C.1, it is straightforward to show that

$$\hat{b} = \left\{E(\tilde{w}_t\tilde{w}_t')^{-1}+o_p(1)\right\}\left\{\frac{1}{T}\sum_{t=2}^{T-h}\tilde{w}_{t-1}\varepsilon_{1,t}+O_p(T^{-1})\right\} = O_p(T^{-1/2}),$$

where the last step applies Chebyshev's inequality to the sample average of the serially uncorrelated process $\tilde{w}_{t-1}\varepsilon_{1,t}$, using the assumption $E(\varepsilon_{j,t}^4)<\infty$.

By the Frisch-Waugh theorem and sample orthogonality of $\hat{\varepsilon}_{1,t}$ and w_{t-1}, we may write

$$\hat{\beta}_h = \theta_h + \frac{\frac{1}{T}\sum_{t=2}^{T-h}(y_{t+h} - \theta_h\hat{\varepsilon}_{1,t})\hat{\varepsilon}_{1,t}}{\frac{1}{T}\sum_{t=2}^{T-h}\hat{\varepsilon}_{1,t}^2} = \theta_h + \frac{\frac{1}{T}\sum_{t=2}^{T-h}(y_{t+h} - \theta_h\varepsilon_{1,t} - \theta_h y_{t-1})\hat{\varepsilon}_{1,t}}{\frac{1}{T}\sum_{t=2}^{T-h}\hat{\varepsilon}_{1,t}^2}. \tag{C.2}$$

Lemma C.1 and $\hat{b} = o_p(1)$ yield $\frac{1}{T}\sum_{t=2}^{T-h}\hat{\varepsilon}_{1,t}^2 \xrightarrow{p} E(\varepsilon_{1,t}^2) = 1$. We can therefore focus on the numerator in the fraction in (C.2), which we decompose as

$$\frac{1}{T}\sum_{t=2}^{T-h}(y_{t+h} - \theta_h\varepsilon_{1,t} - \theta_h y_{t-1})\varepsilon_{1,t} + \frac{1}{T}\sum_{t=2}^{T-h}(y_{t+h} - \theta_h\varepsilon_{1,t} - \theta_h y_{t-1})(\hat{\varepsilon}_{1,t} - \varepsilon_{1,t}). \tag{C.3}$$

We first show that the first term above equals the sum on the right-hand side of (C.1). Iteration on the DGP (1) implies

$$y_{t+h} - \theta_h\varepsilon_{1,t} - \theta_h y_{t-1} = \sum_{\ell=1}^{h}\rho^{h-\ell}\varepsilon_{1,t+\ell} + \sum_{\ell=0}^{h}\rho^{h-\ell}\left(\varepsilon_{2,t+\ell} + \frac{\alpha}{\sqrt{T}}\varepsilon_{2,t+\ell-1}\right).$$

The desired conclusion then follows from

$$\frac{1}{T}\sum_{t=2}^{T-h}\sum_{\ell=0}^{h}\rho^{h-\ell}\frac{\alpha}{\sqrt{T}}\varepsilon_{2,t+\ell-1}\varepsilon_{1,t} = \frac{\alpha}{T^{3/2}}\sum_{t=2}^{T-h}\sum_{\ell=0}^{h}\rho^{h-\ell}\varepsilon_{2,t+\ell-1}\varepsilon_{1,t} = O_p(T^{-1}),$$

which can be verified using Chebyshev's inequality and the fact that the summand in the sum over t is a serially uncorrelated process.

We finish the proof by showing that the second term in (C.3) is $O_p(T^{-1})$. This term equals

$$-\frac{1}{T}\sum_{t=2}^{T-h}\left\{\sum_{\ell=1}^{h}\rho^{h-\ell}\varepsilon_{1,t+\ell} + \sum_{\ell=0}^{h}\rho^{h-\ell}\left(\varepsilon_{2,t+\ell} + \frac{\alpha}{\sqrt{T}}\varepsilon_{2,t+\ell-1}\right)\right\}w_{t-1}'\hat{b}.$$

Using $\hat{b} = O_p(T^{-1/2})$ and Lemma C.1, it suffices to show that

$$\frac{1}{T}\sum_{t=2}^{T-h}\varepsilon_{j,t+\ell}\tilde{w}_{t-1} = O_p(T^{-1/2})$$

for all $j = 1, 2$ and $\ell \geq 0$, and

$$\frac{1}{T^{3/2}}\sum_{t=2}^{T-h}\varepsilon_{2,t-1}\tilde{w}_{t-1} = O_p(T^{-1/2}).$$

Both of these statements follow easily from Chebyshev's inequality. □

Lemma C.3. *Define* $A_0 \equiv \left(\begin{smallmatrix} 0 & 0 \\ 0 & \rho \end{smallmatrix}\right)$. *We have*

$$\hat{A} - A_0 = \left(\frac{1}{T} \sum_{t=2}^{T} \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{1,t} + \varepsilon_{2,t} \end{pmatrix} \tilde{w}_{t-1}' + \frac{\alpha \sigma_2^2}{\sqrt{T}} e_2 e_2' \right) E(\tilde{w}_t \tilde{w}_t')^{-1} + o_p(T^{-1/2}),$$
$$\hat{\kappa} - 1 = \frac{1}{T} \sum_{t=2}^{T} \varepsilon_{1,t} \varepsilon_{2,t} + o_p(T^{-1/2}).$$

Proof. By appealing repeatedly to Lemma C.1, and applying standard laws of large numbers and mean-square moment bounds, we get

$$\begin{aligned}
\hat{A} - A_0 &= \left(\frac{1}{T} \sum_{t=2}^{T} \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{1,t} + \varepsilon_{2,t} + \frac{\alpha}{\sqrt{T}} \varepsilon_{2,t-1} \end{pmatrix} w_{t-1}' \right) \left(\frac{1}{T} \sum_{t=2}^{T} w_{t-1} w_{t-1}' \right)^{-1} \\
&= \left(\frac{1}{T} \sum_{t=2}^{T} \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{1,t} + \varepsilon_{2,t} \end{pmatrix} \tilde{w}_{t-1}' + \frac{\alpha}{T^{3/2}} e_2 \sum_{t=2}^{T} \varepsilon_{2,t-1} \tilde{w}_{t-1}' + O_p(T^{-1}) \right) \\
&\quad \times \left(\frac{1}{T} \sum_{t=2}^{T} \tilde{w}_{t-1} \tilde{w}_{t-1}' + o_p(1) \right)^{-1} \\
&= \left(\frac{1}{T} \sum_{t=2}^{T} \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{1,t} + \varepsilon_{2,t} \end{pmatrix} \tilde{w}_{t-1}' + \frac{\alpha}{\sqrt{T}} e_2 E(\varepsilon_{2,t-1} \tilde{w}_{t-1}') + O_p(T^{-1}) \right) \\
&\quad \times \left(E(\tilde{w}_t \tilde{w}_t') + o_p(1) \right)^{-1}.
\end{aligned}$$

Note that $E(\tilde{w}_{t-1} \varepsilon_{2,t-1}) = \sigma_2^2 e_2$. This proves the first statement of the lemma.

Next, by the Frisch-Waugh Theorem, $\hat{\kappa} \equiv \hat{\Sigma}_{21} / \hat{\Sigma}_{11}$ equals the coefficient on $\varepsilon_{1,t}$ in an OLS regression of y_t on $\varepsilon_{1,t}$ and w_{t-1}. In other words, $\hat{\kappa}$ equals the impact LP estimate $\hat{\beta}_0$. The second statement of the lemma then follows from Lemma C.2 applied to $h = 0$. □

C.2 Proof of Proposition 1

We derive the asymptotic distributions of the LP and VAR estimators in that order.

LP. It follows from Lemma C.2 and a standard martingale central limit theorem that

$$\sqrt{T}(\hat{\beta}_h - \theta_h) \xrightarrow{d} N(0, \text{aVar}_{\text{LP}}),$$

where

$$\begin{aligned}
\text{aVar}_{\text{LP}} &= E(\varepsilon_{1,t}^2) E\left(\left\{\sum_{\ell=1}^{h} \rho^{h-\ell}\varepsilon_{1,t+\ell} + \sum_{\ell=0}^{h} \rho^{h-\ell}\varepsilon_{2,t+\ell}\right\}^2\right) \\
&= E\left(\left\{\sum_{\ell=1}^{h} \rho^{h-\ell}\varepsilon_{1,t+\ell} + \sum_{\ell=0}^{h} \rho^{h-\ell}\varepsilon_{2,t+\ell}\right\}^2\right) \\
&= \sum_{\ell=1}^{h} \rho^{2(h-\ell)} E(\varepsilon_{1,t}^2) + \sum_{\ell=0}^{h} \rho^{2(h-\ell)} E(\varepsilon_{2,t}^2) \\
&= \sum_{\ell=0}^{h} \rho^{2(h-\ell)}(1+\sigma_2^2) - \rho^{2h} \\
&= (1+\sigma_2^2)\frac{1-\rho^{2(h+1)}}{1-\rho^2} - \rho^{2h}.
\end{aligned}$$

VAR. We derive the asymptotic distribution of $\hat{\delta}_h$ by appealing to the delta method. Let $f_h(A,\kappa) \equiv e_2' A^h \gamma$, where $\gamma = (1,\kappa)'$, so that $\hat{\delta}_h = f_h(\hat{A},\hat{\kappa})$. We need the Jacobians of this transformation with respect to $\text{vec}(A)$ and κ. According to Lemma C.3, the Jacobians should be evaluated at $\text{plim}\,\hat{A} = A_0 \equiv \left(\begin{smallmatrix} 0 & 0 \\ 0 & \rho \end{smallmatrix}\right)$ and $\text{plim}\,\hat{\kappa} = 1$. Thus, γ should be evaluated at $\gamma_0 \equiv (1,1)'$.

First, for $h \geq 2$,

$$\begin{aligned}
\frac{\partial e_2' A^h \gamma}{\partial\, \text{vec}(A)}\Big|_{A=A_0,\gamma=\gamma_0} &= (\gamma' \otimes e_2') \sum_{j=1}^{h} (A_0')^{h-j} \otimes A_0^{j-1} \\
&= (\gamma' \otimes e_2')\left((A_0')^{h-1} \otimes I + I \otimes A_0^{h-1} + \sum_{j=2}^{h-1} (A_0')^{h-j} \otimes A_0^{j-1}\right) \\
&= (\gamma_0' \otimes e_2')\rho^{h-1}\left(\left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right) \otimes I + I \otimes \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right) + \sum_{j=2}^{h-1} \left(\left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right) \otimes \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)\right)\right) \\
&= (\gamma_0' \otimes e_2')\rho^{h-1}\left(\left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right) \otimes I + I \otimes \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right) + (h-2)\left(\left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right) \otimes \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)\right)\right) \\
&= \rho^{h-1}\Big(e_2' \otimes e_2' + \gamma_0' \otimes e_2' + (h-2)(e_2' \otimes e_2')\Big) \\
&= \rho^{h-1}\Big((\gamma_0 + (h-1)e_2)' \otimes e_2'\Big) \\
&= \rho^{h-1}\Big((1,h) \otimes e_2'\Big),
\end{aligned}$$

where the third-last equality uses $\gamma_0'\left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right) = e_2'\left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right) = e_2'$, and the last equality uses $\gamma_0 = (1,1)'$. It's clear that the final formula is also the correct Jacobian for $h = 1$. Note that

the form of the above Jacobian implies that the only part of $\hat{A}$ that will contribute to the asymptotic distribution of $\hat{\delta}_h$ is the second row $e_2'\hat{A}$.

Second, for any $h \geq 1$,

$$\frac{\partial e_2' A^h \gamma}{\partial \kappa}\Big|_{A=A_0,\gamma=\gamma_0} = e_2' A_0^h e_2 = \rho^h.$$

Next, Lemma C.3 and a standard martingale central limit theorem imply

$$\sqrt{T}(\hat{A} - A_0)' e_2 \xrightarrow{d} N(\text{aBias}(\hat{A}' e_2), \text{aVar}(\hat{A}' e_2)), \quad \sqrt{T}(\hat{\kappa} - 1) \xrightarrow{d} N(0, \text{aVar}(\hat{\kappa})),$$

where

$$\begin{aligned}
\text{aBias}(\hat{A}' e_2) &= \alpha \sigma_2^2 E(\tilde{w}_t \tilde{w}_t')^{-1} e_2, \\
\text{aVar}(\hat{A}' e_2) &= E(\tilde{w}_t \tilde{w}_t')^{-1} (1 + \sigma_2^2), \\
\text{aVar}(\hat{\kappa}) &= \text{Var}(\varepsilon_{1,t} \varepsilon_{2,t}) = \sigma_2^2.
\end{aligned}$$

Moreover, $\hat{A}' e_2$ and $\hat{\kappa}$ are asymptotically independent by Lemma C.3, since

$$\text{Cov}(\tilde{w}_{t-1}(\varepsilon_{1,t} + \varepsilon_{2,t}), \varepsilon_{1,t}\varepsilon_{2,t}) = E(\tilde{w}_{t-1}) E[(\varepsilon_{1,t} + \varepsilon_{2,t})\varepsilon_{1,t}\varepsilon_{2,t}] = 0.$$

Finally, note that

$$E(\tilde{w}_t \tilde{w}_t')^{-1} = \frac{1}{\sigma_{0,y}^2 - 1} \begin{pmatrix} \sigma_{0,y}^2 & -1 \\ -1 & 1 \end{pmatrix}.$$

Given all the preceding ingredients, we can apply the delta method to conclude that

$$\sqrt{T}(\hat{\delta}_h - \theta_h) \xrightarrow{d} N(\text{aBias}_{\text{VAR}}, \text{aVar}_{\text{VAR}}),$$

where

$$\begin{aligned}
\text{aBias}_{\text{VAR}} &= \rho^{h-1}(1, h)\, \text{aBias}(\hat{A}' e_2) + \rho^h \times 0 \\
&= \rho^{h-1}(h-1) \frac{\alpha \sigma_2^2}{\sigma_{0,y}^2 - 1},
\end{aligned}$$

$$\begin{aligned}
\text{aVar}_{\text{VAR}} &= \rho^{2(h-1)}(1, h)\, \text{aVar}(\hat{A}' e_2)(1, h)' + \rho^{2h}\, \text{aVar}(\hat{\kappa}) \\
&= \rho^{2(h-1)} \frac{1 + \sigma_2^2}{\sigma_{0,y}^2 - 1} (\sigma_{0,y}^2 - 2h + h^2) + \rho^{2h} \sigma_2^2
\end{aligned}$$

$$= \rho^{2(h-1)}(1+\sigma_2^2)\left(1+\frac{(h-1)^2}{\sigma_{0,y}^2-1}\right)+\rho^{2h}\sigma_2^2. \qquad \square$$

References

Austin, B. A. (2020). *Essays on Labor Economics and Econometrics*. PhD thesis, Harvard University. Chapter 2: "The trade-off between LP-IV and SVAR-IV estimation".

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

Ben Zeev, N. & Pappa, E. (2017). Chronicle of a war foretold: The macroeconomic effects of anticipated defence spending shocks. *The Economic Journal*, *127*(603), 1568–1597.

Blanchard, O. & Perotti, R. (2002). An empirical characterization of the dynamic effects of changes in government spending and taxes on output. *the Quarterly Journal of economics*, *117*(4), 1329–1368.

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

Canova, F. (2007). *Methods for Applied Macroeconomic Research*. Princeton: Princeton University Press.

Choi, C.-Y. & Chudik, A. (2019). Estimating impulse response functions when the shock series is observed. *Economics Letters*, *180*, 71–75.

Christiano, L., Eichenbaum, M., & Evans, C. (1999). Monetary Policy Shocks: What Have We Learned and to What End? In J. B. Taylor & M. Woodford (Eds.), *Handbook of Macroeconomics, Volume 1, Part A* chapter 2, (pp. 65–148). Elsevier.

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

Forni, M., Gambetti, L., & Sala, L. (2019). Structural VARs and noninvertible macroeconomic models. *Journal of Applied Econometrics*, *34*(2), 221–246.

Gertler, M. & Karadi, P. (2015). Monetary Policy Surprises, Credit Costs, and Economic Activity. *American Economic Journal: Macroeconomics*, *7*(1), 44–76.

Hansen, B. E. (2016). Stein Combination Shrinkage for Vector Autoregressions. Manuscript, University of Wisconsin-Madison.

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

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

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

Kilian, L. & Kim, Y. J. (2011). How Reliable Are Local Projection Estimators of Impulse Responses? *Review of Economics and Statistics*, *93*(4), 1460–1466.

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

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

Marcellino, M., Stock, J. H., & Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, *135*(1–2), 499–526.

Meier, A. (2005). How Big is the Bias in Estimated Impulse Responses? A Horse Race between VAR and Local Projection Methods. Manuscript, European University Institute.

Mertens, K. & Ravn, M. O. (2013). The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States. *American Economic Review*, *103*(4), 1212–1247.

Miranda-Agrippino, S. & Ricco, G. (2019). Identification with External Instruments in Structural VARs under Partial Invertibility. Manuscript, Bank of England.

Montiel Olea, J. L. & Plagborg-Møller, M. (2020). Local Projection Inference is Simpler and More Robust Than You Think. *Econometrica*. Forthcoming.

Nakamura, E. & Steinsson, J. (2018). Identification in Macroeconomics. *Journal of Economic Perspectives*, *32*(3), 59–86.

Phillips, P. C. B. & Solo, V. (1992). Asymptotics for Linear Processes. *Annals of Statistics*, *20*(2), 971–1001.

Plagborg-Møller, M. & Wolf, C. K. (2020a). Instrumental Variable Identification of Dynamic Variance Decompositions. Manuscript, Princeton University.

Plagborg-Møller, M. & Wolf, C. K. (2020b). 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*(3), 249–258.

Ramey, V. A. (2011). Identifying Government Spending Shocks: It's All in the Timing. *Quarterly Journal of Economics*, *126*(1), 1–50.

Ramey, V. A. (2016). Macroeconomic Shocks and Their Propagation. In J. B. Taylor & H. Uhlig (Eds.), *Handbook of Macroeconomics*, volume 2 chapter 2, (pp. 71–162). Elsevier.

Romer, C. D. & Romer, D. H. (2004). A New Measure of Monetary Shocks: Derivation and Implications. *American Economic Review*, *94*(4), 1055–1084.

Schorfheide, F. (2005). VAR forecasting under misspecification. *Journal of Econometrics*, *128*(1), 99–136.

Shiller, R. J. (1973). A Distributed Lag Estimator Derived from Smoothness Priors. *Econometrica*, *41*(4), 775–788.

Stock, J. H. (2008). What's New in Econometrics: Time Series, Lecture 7. Lecture slides, NBER Summer Institute.

Stock, J. H. & Watson, M. W. (2012). Disentangling the Channels of the 2007–09 Recession. *Brookings Papers on Economic Activity*, *2012*(1), 81–135.

Stock, J. H. & Watson, M. W. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of macroeconomics*, volume 2 (pp. 415–525). Elsevier.

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