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

Dake Li
Princeton University

Mikkel Plagborg-Møller
Princeton University

Christian K. Wolf
MIT & NBER

June 24, 2022

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, designed to mimic the properties of the universe of U.S. macroeconomic data. Our analysis considers various identification schemes and several variants of LP and VAR estimators. A clear bias-variance trade-off emerges: LP estimators have lower bias than VAR estimators but substantially higher variance at intermediate and long horizons. Consequently, unless researchers are overwhelmingly concerned with bias, 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.

*Email: dakel@princeton.edu, mikkelpm@princeton.edu, and ckwolf@mit.edu. We received helpful comments from Isaiah Andrews, Régis Barnichon, Gabe Chodorow-Reich, Viet Hoang Dinh, Òscar Jordà, Helmut Lütkepohl, Massimiliano Marcellino, Pepe Montiel Olea, Ulrich Müller, Emi Nakamura, Chris Sims, Lumi Stevens, Jim Stock, Mark Watson, Andrei Zeleneev, and numerous seminar and conference participants. Plagborg-Møller acknowledges that this material is based upon work supported by the NSF under Grant #1851665. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

1 Introduction

Since Jordà (2005) introduced the popular local projection (LP) impulse response estimator, there has been a debate about its benefits and drawbacks relative to Vector Autoregression (VAR) estimation (Sims, 1980). Recently, Plagborg-Møller & Wolf (2021) proved that these two 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 (e.g., 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 instead 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 long 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, applying LP and VAR methods to thousands of empirically relevant DGPs. Our goal is to identify which estimators perform well *on average* across many DGPs and thus may serve as practical default procedures. Rather than insisting on the usual binary distinction between "local projections" and "VARs", we furthermore consider an entire menu of related estimation approaches. We find that the usual least-squares LP estimator tends to have lower bias than the least-squares VAR estimator, as

[1] See Plagborg-Møller & Wolf (2021, 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.

expected, but this bias reduction incurs a substantial cost of higher variance. Consequently, unless researchers are almost exclusively concerned with bias, least-squares LP is not optimal, and shrinkage estimation via Bayesian VARs or penalized LPs is usually attractive.

Our simulation study considers an extensive array of DGPs, obtained by drawing specifications at random from a large-scale, empirically calibrated dynamic factor model (DFM).[3] 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 well-known to accurately capture the joint co-movements of conventional macroeconomic data, thus ensuring that our simulation results will be informative about the universe of standard U.S. time series. From the encompassing 207-variable DFM we draw 6,000 random subsets of five variables (subject to certain constraints that emulate applied practice), whose model-implied time series processes constitute the set of DGPs we consider for our simulation study. Importantly, these DGPs vary in terms of how well they can be approximated by low-order VAR models, thus yielding a non-trivial bias-variance trade-off between LPs and VARs. Moreover, the DGPs exhibit substantial heterogeneity in persistence, shape of impulse response functions, and invertibility of the structural shocks, consistent with the heterogeneity faced by applied researchers. An estimation method that works well across this multitude of DGPs therefore carries substantial promise as an attractive default procedure.

We study the ability of several variants of LP and VAR methods to accurately estimate impulse response functions. 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 white noise; (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). For each estimation method, we consider three classical structural identification schemes: observed shocks, instrumental variables (IVs)/proxies, and recursive identification. For IV identification, we distinguish between internal IV methods (Ramey, 2011; Plagborg-Møller & Wolf, 2021) and external IV methods (Stock, 2008; Stock & Watson, 2012; Mertens & Ravn, 2013).

Applying the various estimation methods to simulated data from our multitude of DGPs, a clear and unavoidable bias-variance trade-off emerges. We highlight four main lessons:

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

1. Least-squares LP and VAR estimators lie on opposite ends of the bias-variance spectrum: small bias and large variance for LPs, and large bias and small variance for VARs. Thus, 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. Yet, the *slope* of this trade-off is stark: indifference between the two methods requires a researcher to care on average around four times more about squared bias than about variance.

2. Shrinkage methods dramatically lower the variance of LP and VAR methods, at a usually moderate cost in terms of bias. In particular, unless researchers care almost exclusively about bias, penalized LP is preferred to least-squares LP, and Bayesian VAR shrinkage is preferred to or at least competitive with the least-squares VAR estimator. However, when the data is very persistent, the performance of the Bayesian VAR estimator is sensitive to the choice of prior at long horizons.

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 least-squares or Bayesian VAR estimation are the most attractive methods 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).

Our findings provide a novel perspective on recent work emphasizing the potential dangers of VAR model mis-specification (Ramey, 2016; Nakamura & Steinsson, 2018). We consider DGPs that do not admit finite-order VAR representations, so VAR methods indeed 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 conventional 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, though we focus instead on structural impulse responses. While simulation studies in 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 popular shock identification schemes, normalization of impulse responses, and the special role of external instrumental variables.

Our large-scale set-up differs from prior simulation studies of LP and VAR methods, which have considered at most a handful of DGPs. Examples include Jordà (2005), Meier (2005), Kilian & Kim (2011), Brugnolini (2018), Choi & Chudik (2019), Austin (2020), and Bruns & Lütkepohl (2021). 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 least-squares estimators; we study several popular structural identification schemes; and we examine how our conclusions vary with the impulse response horizon and the researcher's loss function. All these features are essential to the above-mentioned main lessons that we draw from our results.

Even though the simulation results are at the heart of our analysis, we start off by illustrating 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 practitioners, as this trade-off invariably depends in a complicated fashion on many features of the DGP.

Finally, we stress that our paper focuses solely on point estimation, as opposed to inference or hypothesis testing. See Inoue & Kilian (2020) and Montiel Olea & Plagborg-Møller (2021) for theoretical and simulation results on VAR and LP confidence interval procedures. Moreover, we focus exclusively on impulse response estimands, rather than variance decompositions or historical 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 our many DGPs. Section 4 defines the menu of LP- and VAR-based estimation procedures. Section 5 contains our main simulation results and robustness checks. Section 6 summarizes the lessons for applied researchers and offers guidance for future research. The appendix contains implementation details. A supplemental appendix with proofs and further simulation results and a Matlab code suite are available online (`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 (2021) 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 is proportional to the mean squared error (MSE).[4] 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

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

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

2. **VAR.** We consider a recursive VAR specification in $w_t = (\varepsilon_{1,t}, y_t)'$, again with one lag. Define the usual least-squares 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 $\hat{\kappa} \equiv \hat{\Sigma}_{21}/\hat{\Sigma}_{11}$.[5]

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 $T \to \infty$,*

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

[4] The objective function (2) is not a loss function in the usual decision theoretic sense (which would call it a risk function when $\omega = \frac{1}{2}$). We proceed with the non-standard terminology for ease of exposition.

[5] 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 $\hat{A}^h \hat{\gamma}$.

where for all $h \geq 0$,

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

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 Supplemental Appendix F. □

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 $\rho \neq 0$.[6]

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 persistence and severe mis-specification ($\rho = 0.9$ and $\alpha = 5$).[7] Both set $\sigma_2 = 1$. Figure 2 then 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 asymp-

[6] 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 $w_t = Aw_{t-1} + u_t$, $u_t \overset{i.i.d.}{\sim} N(0, \Sigma)$, 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.

[7] To interpret the magnitude of α, let $\hat{\tau}$ denote the t-statistic for testing the significance of the second lag in a univariate AR(2) regression for $\{y_t\}$. Supplemental Appendix F.3 shows that $\hat{\tau} \xrightarrow{d} N\left(-\rho\frac{\sigma_2^2}{1+\sigma_2^2}\alpha, 1\right)$. We remind the reader, however, that the purpose of this section is illustration, not quantitative realism.

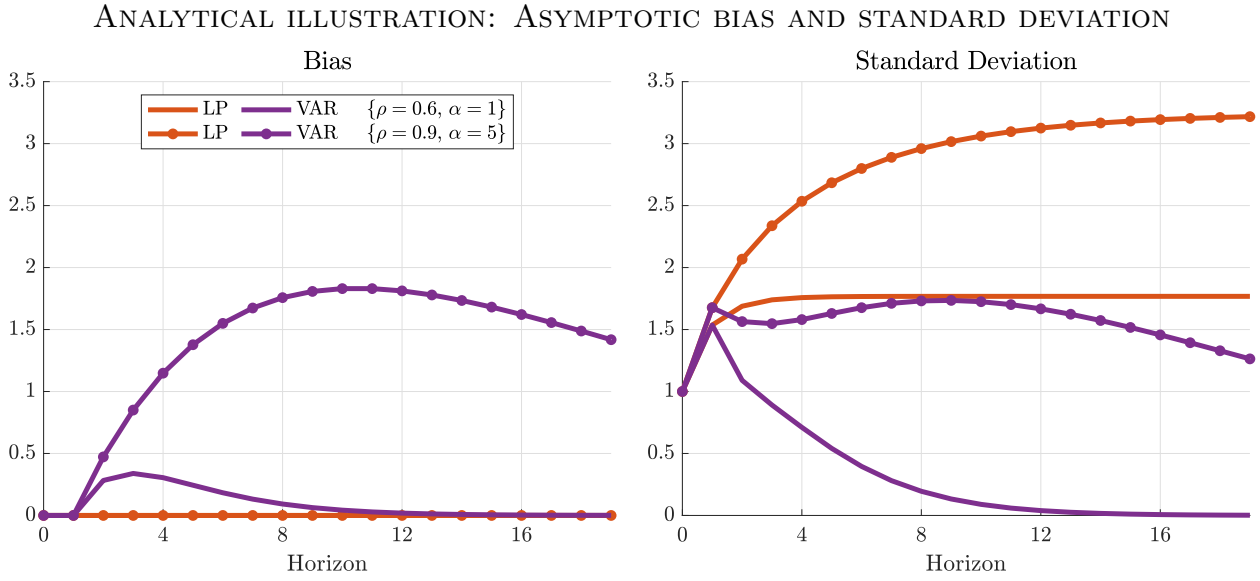

Figure 1: Asymptotic bias and standard deviation for LP (red) and VAR (purple) in the DGP (1) with $\sigma_1 = 1$ and $\{\rho = 0.6, \alpha = 1\}$ (no markers) or $\{\rho = 0.9, \alpha = 5\}$ (markers).

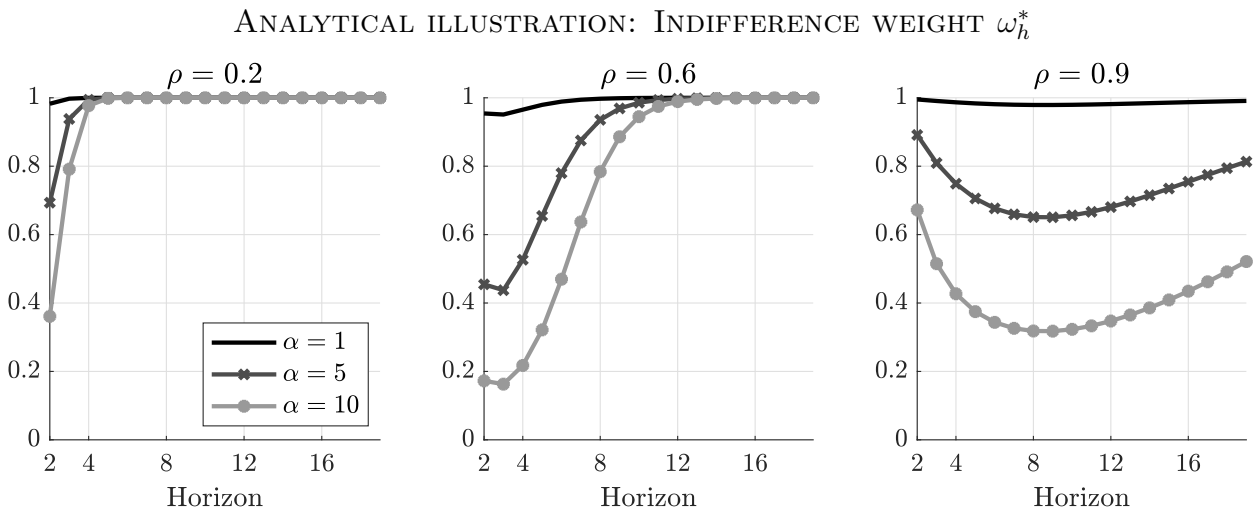

Figure 2: Weight $\omega = \omega_h^*$ in the asymptotic loss function $\omega \times \text{aBias}^2_{\hat{\theta}_h} + (1-\omega) \times \text{aVar}_{\hat{\theta}_h}$ that yields indifference between the LP and VAR estimator. LP is preferred whenever $\omega \geq \omega_h^*$. The three panels correspond 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.

totic 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, 2021).

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 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.[8]

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.

To summarize, even in this simple model environment, the researcher's evaluation of VAR and LP estimators is shaped by several parameters: the DGP's persistence ρ, the degree of mis-specification α, the horizon h of interest, and the weight ω attached to squared bias rather than variance in the loss function.

2.2 Outlook

Because analytical bias-variance calculations invariably end up depending on a multitude of parameters, the rest of this paper will use simulations to explore the trade-off across a rich

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

and empirically relevant set of 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 practically relevant bias weight ω necessary to justify the use of one linear projection technique over another. Moreover, we will also consider several variants of the least-squares LP and VAR estimators that use shrinkage approaches to further trace out the bias-variance possibility frontier.

3 Data generating processes

This section presents our DGPs. We define the empirically calibrated encompassing model in Section 3.1, from which we then 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. Various modifications to this baseline set of DGPs are considered later in Section 5.5.

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).

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 components. 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 component $v_{i,t}$ 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 $\xi_{i,t}$ 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 components 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, chosen to mimic as closely as possible popular structural 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, $i_t \in \bar{w}_t$ 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 impulse responses of an outcome variable y_t with respect to a one standard deviation (i.e., one unit) innovation to $\varepsilon_{1,t}$:

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

   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 index ι_y corresponds to the location of y_t in the vector $\bar{w}_t$.

   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 $w_t = (z_t, \bar{w}_t)'$. As is standard in IV applications, we here adopt the "unit effect" normalization of Stock & Watson (2016), so the object of interest becomes

$$\theta_h \equiv \frac{\bar{\Lambda}_{\iota_y,\bullet}\Theta^f_{\bullet,1,h}}{\bar{\Lambda}_{\iota_i,\bullet}\Theta^f_{\bullet,1,0}}, \quad h = 0, 1, 2, \ldots, \tag{10}$$

where the index ι_i corresponds to the location of some normalization variable i_t in the vector $\bar{w}_t$. The above unit effect normalization defines the magnitude of the shock $\varepsilon_{1,t}$ such that it raises the normalization variable i_t by one unit on impact.

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 $w_t = \bar{w}_t$. 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 in the VAR(∞) process for $\bar{w}_t$ implied by the DFM. The ordering of the variables for the recursive identification is described below in Section 3.3. Note that the recursively orthogonalized innovation differs across DGPs, and it generally does not equal any of the structural shocks $\varepsilon_{j,t}$ in the DFM. We nevertheless consider this impulse response estimand due to its popularity in applied work. We relegate the mathematical definition of the recursive impulse response estimand to Supplemental Appendix C.

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 and refer to Stock & Watson for further details. The full vector of observables X_t contains quarterly observations on 207 time series for 1959Q1–2014Q4, mostly consisting of real activity variables, price measures, interest rates, asset and wealth variables, and productivity series.[9] 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 component equation (7). The reduced-form parameters are estimated by principal components and least-squares procedures. This pins down all parameters of the DFM except for the structural impact response matrix H, which we discuss below.

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 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 for the IV and recursive estimands. 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 real activity, and at least one other variable a measure of prices.[10] 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 implement the three structural impulse response estimands in Section 3.2 as follows:

[9] Table 1 and the Data Appendix of Stock & Watson (2016) list all variables and their categories.

[10] Real activity series correspond to categories 1–3 in the classification in Table 1 of Stock & Watson (2016), while price series correspond to category 6.

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. This ensures that monetary and fiscal shocks account for substantial short-run variation in nominal interest rates and government spending, respectively. Additionally, we avoid issues related to division by near-zeros when normalizing the impulse responses for the IV estimand. See Appendix A.1 for further details.

2. **IV.** The matrix H is defined as in the "observed shock" case. As for the IV parameters in equation (9), we draw ρ_z uniformly at random from the set $\{0, 0.25, 0.5\}$.[11] To ensure an empirically plausible signal-to-noise ratio, we calibrate σ_ν^2 using real-world IV 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.2 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.

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 contemporaneously 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.4.

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.[12] This measure varies widely across the DGPs, with more than

[11] The external IVs used in empirical practice tend to have low to moderate autocorrelation (Ramey, 2016), consistent with our assumptions on ρ_z.

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

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	9.28	9.40	9.48	15.31	23.38	23.91	24.66
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)	-0.42	-0.16	-0.08	-0.02	0.06	0.11	0.43
R^2 in regression on quadratic	0.01	0.09	0.20	0.46	0.69	0.83	0.97

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)": $(\frac{1}{21}\sum_{h=0}^{20} \theta_h)/\max_h\{|\theta_h|\}$. "$R^2$ in regression on quadratic": R-squared from a regression of the impulse response function $\{\theta_h\}_{h=0}^{20}$ on a quadratic polynomial in h.

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.[13] We consider an alternative set of more persistent DGPs in a robustness check in Section 5.5.

Second, the DGPs are heterogeneous in terms of how well they can be approximated by a relatively 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 coefficient matrices $\{A_\ell^w\}_\ell$ in the VAR(∞) representation for $\{\bar{w}_t\}$ at or after lag 5 (with $\|\cdot\|$ denoting 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 the various DGPs that we consider in our simulations.

Third, for the IV specifications, the DGPs differ in terms of the degree of invertibility of

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

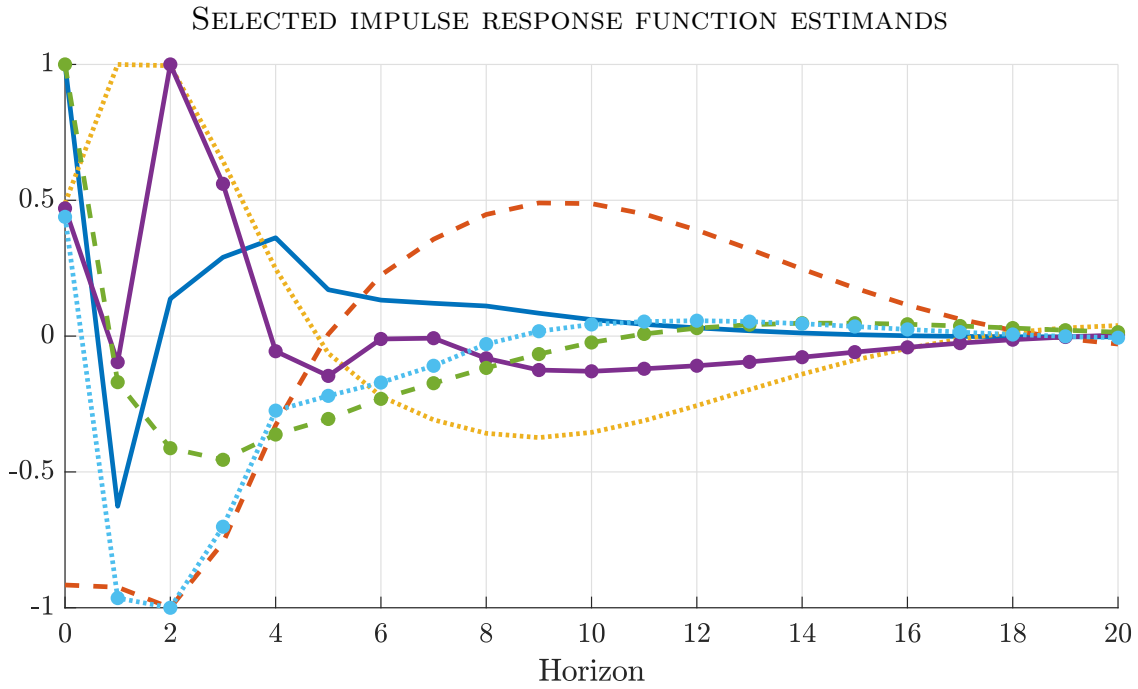

Figure 3: Selected impulse responses of macro observables to monetary and fiscal policy shocks. Here the impulse response functions are normalized to have a maximum value 1 or -1.

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 25, 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 maximal response. Finally, the smoothness of the

impulse response functions varies substantially: the R-squared value in a regression of the impulse responses $\{\theta_h\}_{h=0}^{20}$ on a quadratic polynomial $b_0 + b_1 \times h + b_2 \times h^2$ has 10th and 90th percentiles given by 0.09 and 0.83, respectively. Figure 3 displays the true values of 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- and LP-based estimation methods that we consider in the simulation study. Though all these methods aim at estimating the same population impulse responses defined in Section 3.2, the methods differ in terms of their bias/variance properties, and in terms of their robustness to non-invertibility. Implementation details are relegated to Appendix B.[14] 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, 2021). 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) plus a penalty term that encourages

[14] 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.

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.

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.

[15] We report posterior means of the impulse responses based on 100 draws.

The VAR model averaging estimator effectively includes LP among the list of candidate estimators (as in the related approach of Miranda-Agrippino & Ricco, 2021a). This is because the candidate VAR(20) model gives results similar to LP with several lagged controls, at all horizons considered in our study (Plagborg-Møller & Wolf, 2021).

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 (2021) prove that this approach consistently estimates the *normalized* structural impulse responses (10) 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, 2021b; Plagborg-Møller & Wolf, 2022). We shall see that the SVAR-IV estimator tends to exhibit lower dispersion 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.

Lag length selection. As a baseline, the LP and VAR estimators use $p = 4$ lags for estimation (except VAR model averaging, which uses many different lag lengths). In our DGPs, the Akaike Information Criterion almost always selects very short lag lengths $\hat{p}_{AIC}$, as discussed further in Section 5.6 below. 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 $p = 8$ are discussed in Section 5.5.

5 Results

This section presents our simulation results. We summarize the results through four takeaways, one in each subsection. The first three takeaways focus on observed shock identification. The fourth takeaway is concerned with IV identification. We show in Section 5.5 that these conclusions are qualitatively robust to several specification choices (including recursive identification). Finally, in Section 5.6, we justify our focus on the *average* performance of estimators across DGPs, by arguing that there is limited scope for selecting among estimators in a data-dependent way.

Throughout this section we present results for our 6,000 monetary and fiscal policy shock DGPs considered jointly rather than separately. 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. Our baseline results take about one week to produce in Matlab on 16 servers with 25 parallel cores each.

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 least-squares LP and VAR being the thick lines. Before taking the median, we cancel out the units of the response variables by dividing the bias and standard deviation by $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$, i.e., the root mean squared value of the *true* impulse response function out to horizon 20. 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,

[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.

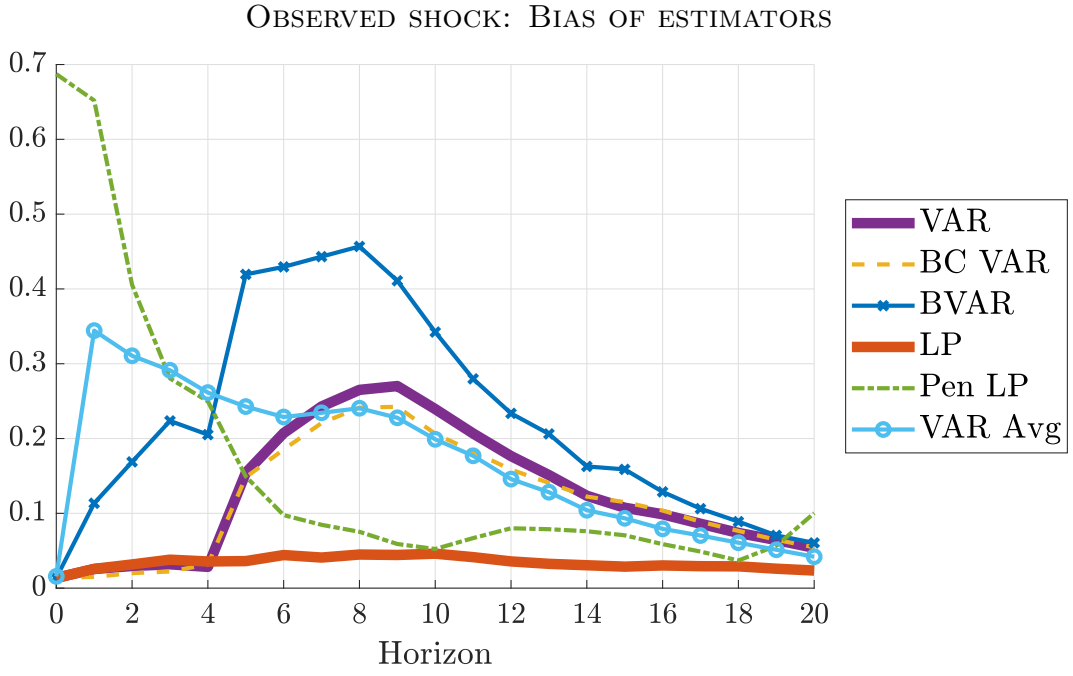

Figure 4: Median (across DGPs) of absolute bias of the different estimation procedures, relative to $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$.

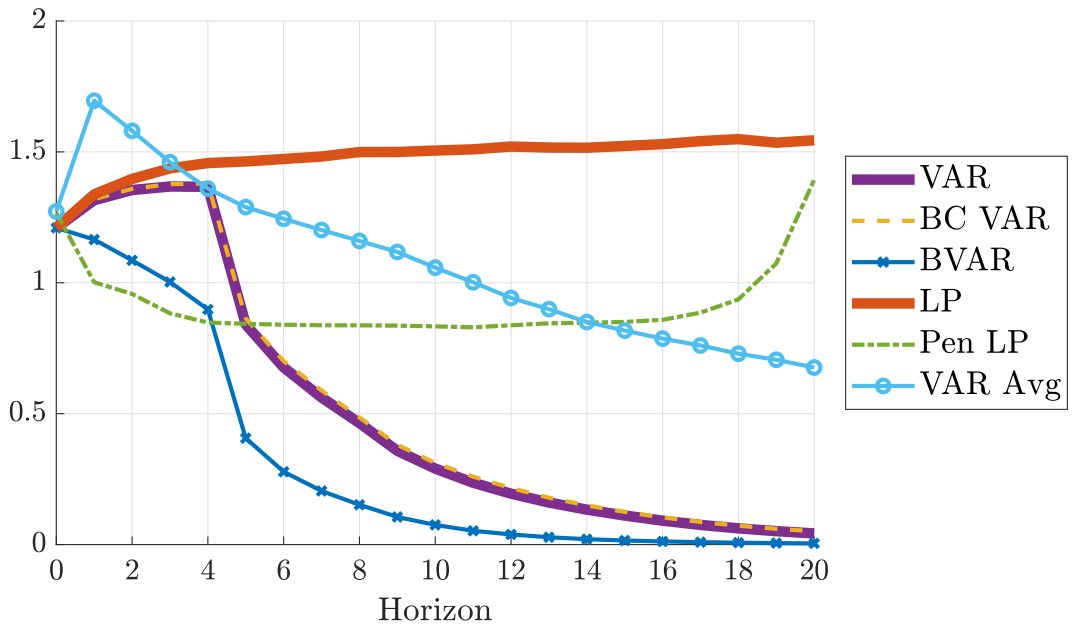

Figure 5: Median (across DGPs) of standard deviation of the different estimation procedures, relative to $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$.

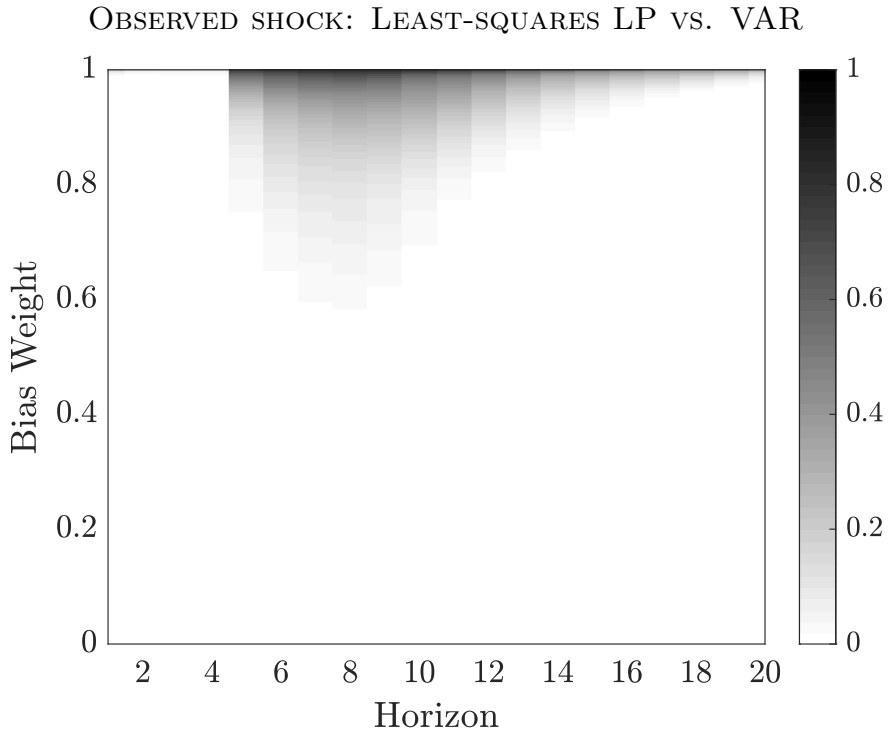

Figure 6: Fraction of DGPs for which the least-squares LP estimator has a lower loss than the VAR estimator. Darker areas correspond to regions where LP is preferred more often. Horizontal axis: impulse response horizon h. Vertical axis: weight ω on squared bias in the loss function (2). The loss function 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.

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). These observations are consistent with the theoretical results in Section 2, Schorfheide (2005), and Plagborg-Møller & Wolf (2021).

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 calibration of our simple model in Section 2 with (i) moderate persistence (ρ) and (ii) non-zero but limited VAR mis-specification (α).

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: shrinkage offers a very competitive reduction in variance in return for its increased bias, and so justification of least-squares VAR requires a relatively high bias weight ω. Note that if we

[17]Penalized LP performs similarly to least-squares LP at horizon $h = 20$. The reason is mechanical: the Barnichon & Brownlees (2019) smoothness penalty affects estimates near the "boundary" (i.e., the largest penalized horizon) less than at the other horizons.

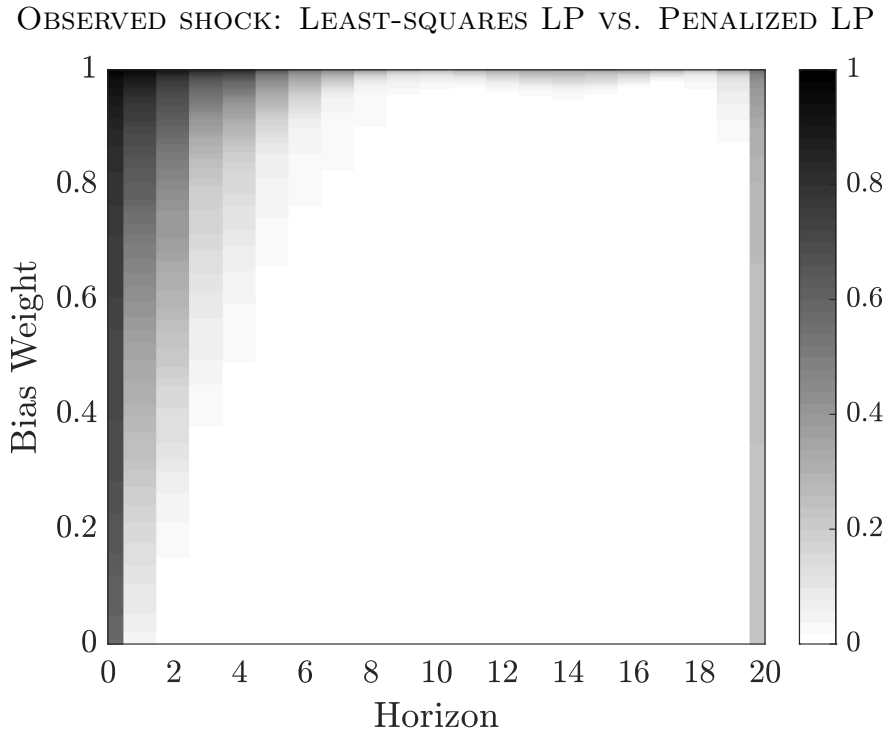

Figure 7: Fraction of DGPs for which the least-squares LP estimator has a lower loss than penalized LP. Darker areas correspond to regions where least-squares LP is preferred more often. See caption for Figure 6.

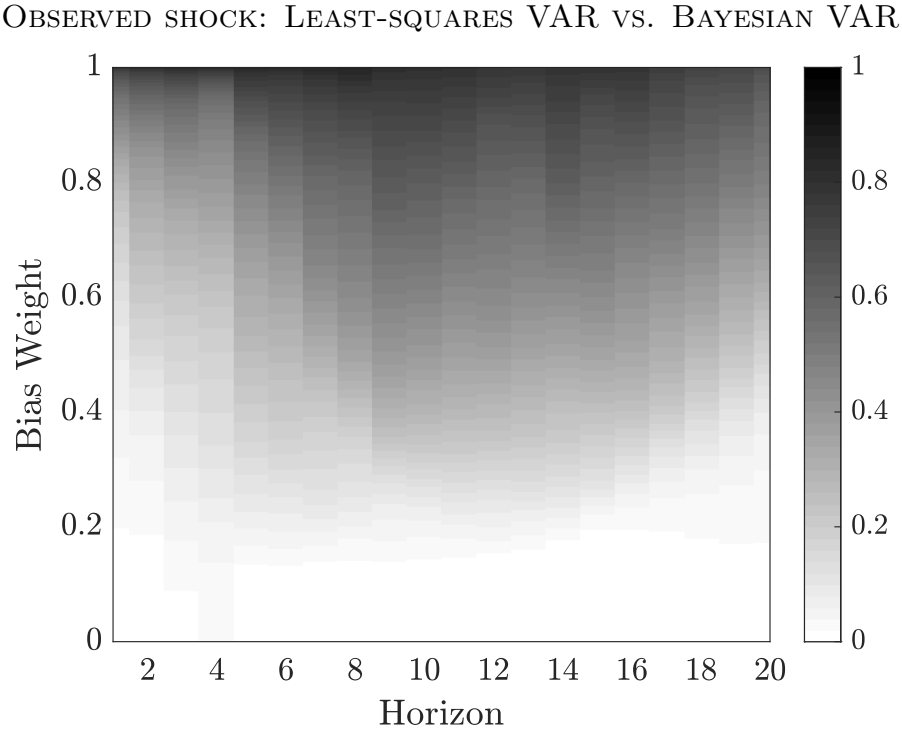

Figure 8: Fraction of DGPs for which the least-squares VAR estimator has a lower loss than the BVAR estimator. Darker areas correspond to regions where least-squares VAR is preferred more often. See caption for Figure 6. The impact horizon $h = 0$ is omitted due to numerical equivalence between the estimators.

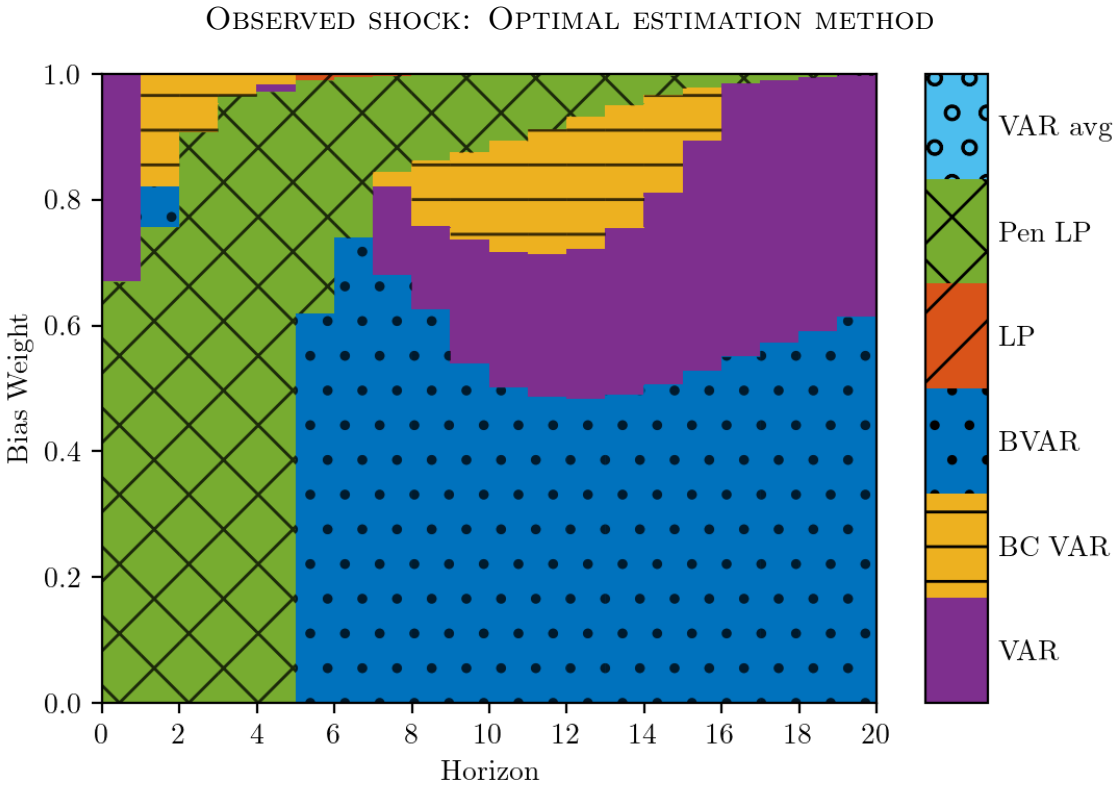

Figure 9: Method that minimizes the average (across DGPs) loss function (2). Horizontal axis: impulse response horizon. Vertical axis: weight on squared bias in loss function. The loss function is normalized by the scale of the impulse response function, as in Figures 4 and 5. At $h = 0$, VAR and LP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

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 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 short impulse response horizons, penalized LP is almost always the best choice, except when ω is close to 1 (i.e., an almost exclusive concern for bias). At horizons longer than the 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.

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, as discussed in Section 3.4. Figure 9 also shows that least-squares LP (orange with diagonal lines) as well as the VAR model averaging estimator (light blue with hollow circles) are almost never optimal.[18] Least-squares LP is essentially dominated because its low bias is not much lower than that of penalized LP at intermediate horizons, and so the high variance of least-squares LP is difficult to justify. The reason why VAR model averaging does poorly is due to its high median standard deviation relative to other VAR-based estimators (see 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.[19]

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.[20] If we ignore the dotted burgundy line for SVAR-IV, these figures are qualitatively similar to those presented

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

[19] 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.

[20] For completeness, (mean) bias and standard deviation are reported in Supplemental Appendix E.1.

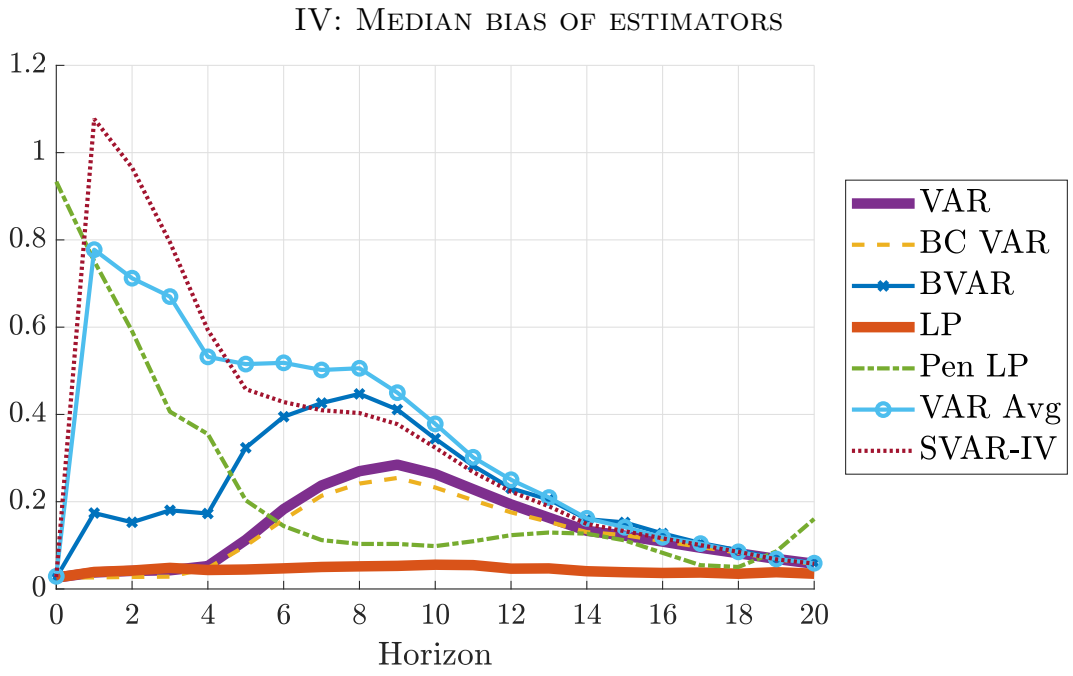

Figure 10: Median (across DGPs) of absolute median bias of the different estimation procedures, relative to $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$.

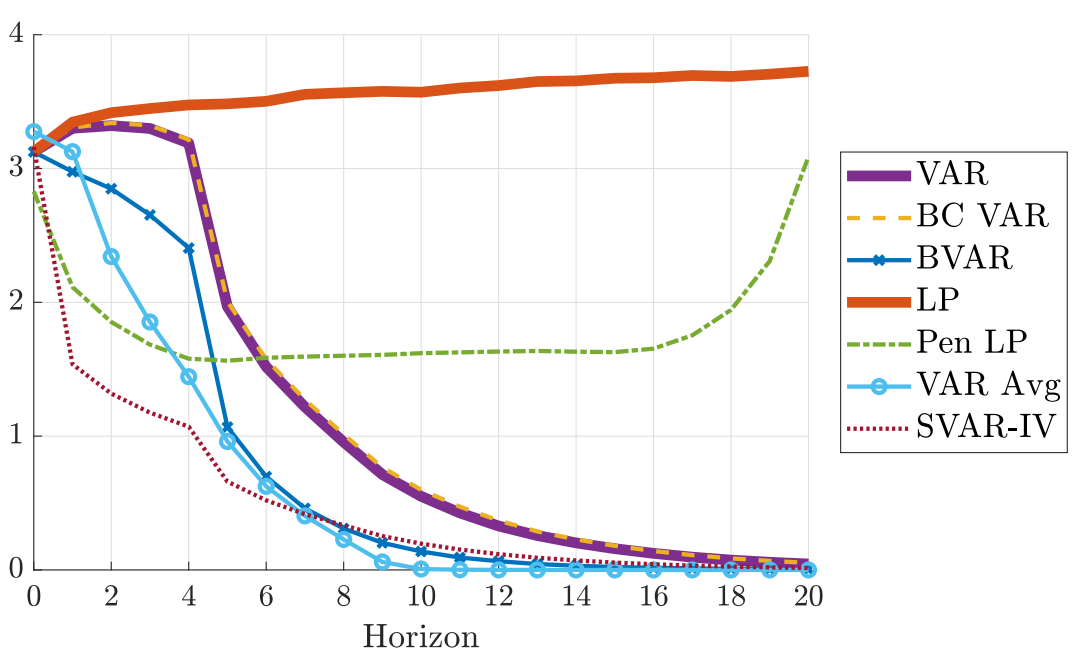

Figure 11: Median (across DGPs) of interquartile range of the different estimation procedures, relative to $\sqrt{\frac{1}{21}\sum_{h=0}^{20}\theta_h^2}$.

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 in fact trump the bias concerns.

5.5 Robustness

This section argues that our main conclusions in Sections 5.1 to 5.4 are robust to several alterations of our baseline simulation specification. Since our baseline DGPs are not highly persistent (see Section 3.4), we here pay particular attention to robustness exercises that consider DGPs with greater persistence. Various other robustness checks are listed subsequently, with details relegated to Supplemental Appendix E.

Persistence. We consider an alternative set of DGPs with higher persistence than our baseline specification. This is achieved by adjusting the VAR(2) process for the factors f_t so that the largest absolute eigenvalue of the VAR polynomial equals 0.95, implying that some impulse responses in each DGP decay with a half-life of approximately 3.4 years. We correspondingly scale the factor innovations to keep the trace of the factor variance-covariance matrix constant. See Appendix A.3 for details on the adjustment.

Higher persistence does not change our qualitative conclusions, except that the performance of the Bayesian VAR procedure becomes more sensitive to the prior at long horizons. Figure 12 shows the optimal method choice for this more persistent set of DGPs; other results are presented in Supplemental Appendix E.2. At short horizons, the results are similar to the baseline results in Figure 9, with penalized LP an attractive option unless the concern for bias is overwhelming. At long horizons, the least-squares or bias-corrected VAR procedures do relatively better than the Bayesian VAR procedure compared to our baseline. This is because the BVAR procedure we consider shrinks towards white noise, which incurs a substantial bias for some persistent DGPs. The natural alternative of centering the prior at independent random walks produces only slightly lower bias. We thus conclude that,

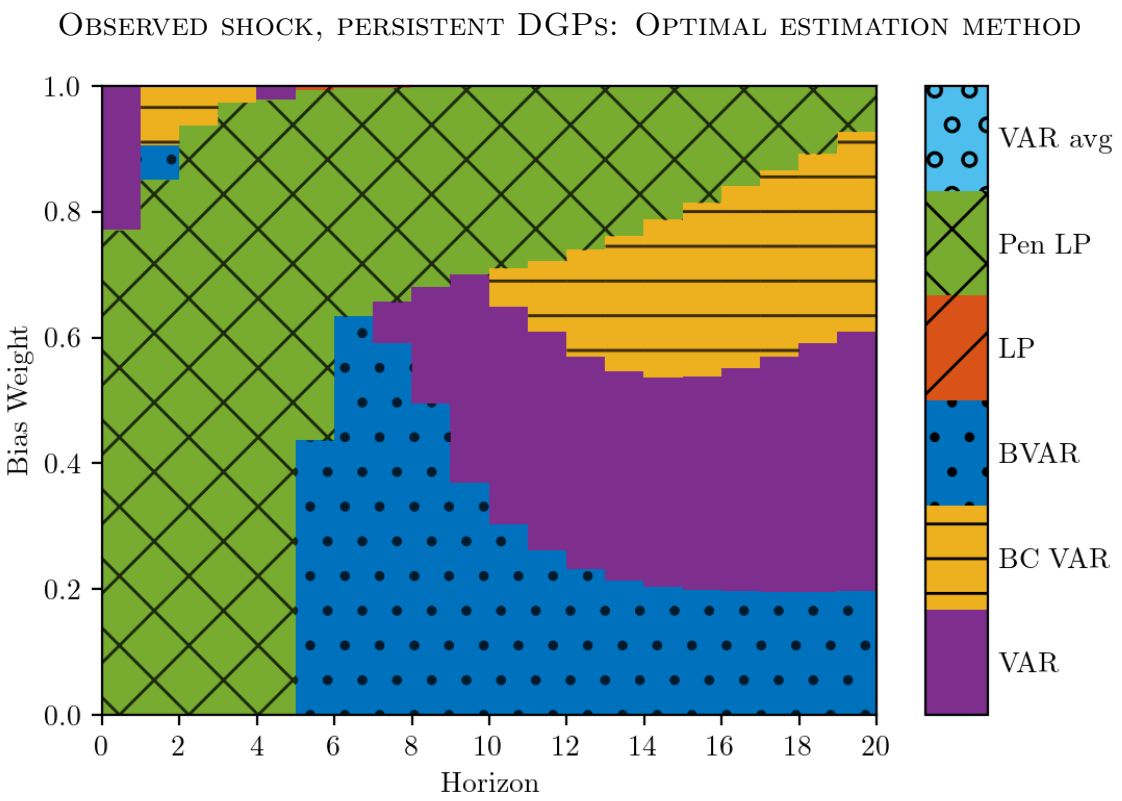

Figure 12: Method that minimizes the average (across DGPs) loss function (2) for the alternative persistent DFM. Horizontal axis: impulse response horizon. Vertical axis: weight on squared bias in loss function. The loss function is normalized by the scale of the impulse response function, as in Figures 4 and 5. At $h = 0$, VAR and LP are numerically identical; we break the tie in favor of VAR.

when the data is very persistent and interest centers on long horizons, prior elicitation for Bayesian VARs should receive careful attention and should not be left to convention. Another finding is that the bias correction of the VAR estimator has more bite in the persistent set of DGPs, though the performance of this estimator remains qualitatively comparable to the un-corrected least-squares VAR estimator.

Rather than directly changing the persistence of the DGPs, we also consider an alternative exercise that *cumulates* impulse responses across horizons. Starting with our baseline results, we compute cumulated impulse responses for those time series in the Stock & Watson (2016) data set that have been differenced prior to estimation. Thus, the resulting impulse responses can be interpreted as the responses of the levels of these variables. With this change in estimand, both bias and variance of the estimates now tend to increase with the horizon, as shown in Supplemental Appendix E.3. Yet our takeaways regarding the *relative* magnitudes

of bias and variance are not materially affected. In particular, unless the concern for bias is substantial, the bias-variance trade-off continues to be best addressed by employing shrinkage techniques such as penalized LPs or BVARs.

Other robustness checks. The following modifications to our baseline simulation specification all leave our main conclusions qualitatively unchanged.

- **Recursive identification:** Results for recursively identified shocks are strikingly similar to those for observed shock identification. See Supplemental Appendix E.4.
- **Monetary vs. fiscal shocks:** If we consider the monetary shock DGPs separately from the fiscal shock DGPs, the bias-variance trade-off is qualitatively similar to when we consider the DGPs jointly. See Supplemental Appendix E.5.
- **Lag length:** If the estimation lag length p is set to 8 instead of 4, our conclusions go through as long as we reinterpret "short horizons" to mean $h \leq 8$ instead of $h \leq 4$. See Supplemental Appendix E.6.
- **Sample size:** Halving the sample size to $T = 100$ quarters tends to increase the estimator standard deviations more than the biases, so shrinkage techniques look even more desirable than in our baseline. See Supplemental Appendix E.7.
- **Variable categories:** We find little evidence that the biases or standard deviations of individual impulse response estimators depend systematically on which categories of time series are included in the DGP (e.g., how many financial variables are used). See Supplemental Appendix E.8.
- **Salient observables:** Our results remain essentially unchanged if we restrict attention to a subset of 18 oft-used, salient macroeconomic time series out of the 207 ones in the Stock & Watson (2016) data set. We consider the exhaustive list of *all* five-variable DGPs that can be formed from these 18 series, subject to the selection rules in Section 3.3. See Supplemental Appendix E.9.
- **Near-worst-case performance:** Whereas our baseline results pertain to the median performance of estimators across DGPs, some researchers may instead prefer to focus on ensuring acceptable performance for particularly challenging DGPs. To this end, Supplemental Appendix E.10 reports the 90th percentiles of the bias and standard deviation across DGPs. Interestingly, adopting this "near-worst-case" perspective does

not alter much the *relative* magnitudes of bias and standard deviation across estimation procedures. Hence, none of the estimation procedures seem to have a particular advantage in ensuring robustness to challenging environments, over and above their performance on typical DGPs.

5.6 Discussion: can we select the estimator based on the data?

Whereas the preceding results have focused on the performance of estimators *on average* across our many DGPs, it is natural to ask whether the choice of estimator can be guided by the data at hand. We now show that this appears to be difficult, as conventional model selection or evaluation criteria are unable to detect even substantial mis-specification of the VAR(4) model in the vast majority of our DGPs. These findings are consistent with the previously documented poor performance of the VAR model averaging estimator. For simplicity, we focus again on the "observed shock" DGPs.

First, the Akaike Information Criterion (AIC) tends to select very short lag lengths $\hat{p}_{AIC}$ in our DGPs, as mentioned earlier. The 90th percentile of $\hat{p}_{AIC}$ (across simulations) does not exceed 2 in any of our 6,000 DGPs, and it equals 2 in only 11.0% of the DGPs. This frequently used model selection tool therefore essentially never indicates that the VAR(4) specification is mis-specified.

Second, the Lagrange Multiplier test of residual serial correlation has low power in most of our DGPs. We carry out this test by regressing the VAR sample residuals on their first lags, controlling for four lags of the observed variables, and employing the specific likelihood ratio test defined in Johansen (1995). Using a 10% significance level for the test, less than 0.1% of the DGPs exhibit a rejection probability above 25%. Hence, this conventional specification test of the VAR(4) model is severely under-powered, despite the fact that many of our DGPs are in fact not well approximated by a VAR(4) model in population (see Section 3.4).

It is of course possible that other model selection criteria or specification tests will work better. However, at a minimum, the performance of the VAR model averaging estimator discussed in Section 5.3 and the evidence presented in this subsection together suggest that it is not straightforward to develop effective data-dependent estimator selection rules for use on conventional macroeconomic time series data. We therefore leave the issue of data-driven estimator selection to future research.

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 generating 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 at intermediate and long horizons.

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 often only moderately larger bias. Typically the researcher must care overwhelmingly about bias to justify the use of conventional least-squares estimators over shrinkage procedures. An important caveat is that, if the data is highly persistent and interest centers on long horizons, shrinkage via Bayesian VARs requires careful prior elicitation. The literature on shrinkage estimators of impulse responses 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 usually optimal or near-optimal. Penalized LP is especially attractive at short horizons, while BVAR estimation is generally attractive—or at least competitive—at intermediate and long horizons. One might hope to achieve the best of both worlds by selecting among estimators in a data-dependent way, but preliminary evidence suggests that the scope for this is limited. In particular, we find that conventional model selection or evaluation tools are unable to detect substantial VAR mis-specification in realistic sample sizes. Despite our negative results in this area, we view data-dependent estimator selection as an important topic for future research.

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. An interesting question is whether it is possible to develop alternative non-invertibility-robust estimation procedures that capture some of the variance improvement enjoyed by SVAR-IV.

Whereas our simulation DGPs are calibrated to stationarity-transformed macroeconomic time series, future research could fruitfully analyze the bias-variance trade-off in other data environments. First, due to space constraints, we have not considered applying the estimation procedures directly to time series data with unit roots. Second, we conjecture that the nature of the bias-variance trade-off may differ somewhat in panel data settings, to the extent that the availability of a large cross section reduces the sampling variance of the estimators for a given time dimension.

Appendix A Details on DGP definitions

A.1 Shock definition

Our definition of the structural shock of interest, $\varepsilon_{1,t}$, ensures that it has the largest possible contemporaneous effect on nominal interest rates (for monetary shocks) and government spending (for fiscal shocks). Letting $\eta_t = H\varepsilon_t$ and ι^* denote the index of the policy instrument, the shock is thus defined through the solution of the following problem:

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

where $\Sigma_\eta \equiv \text{Var}(\eta_t)$, and e_1 selects the first column of H. The solution is given by $H_{\bullet,1} = \Sigma_\eta \bar{\Lambda}'_{\iota^*,\bullet}(\bar{\Lambda}_{\iota^*,\bullet}\Sigma_\eta \bar{\Lambda}'_{\iota^*,\bullet})^{-1/2}$.[21]

A.2 IV process calibration

We calibrate the parameters of the IV equation (9) using the shock series of Romer & Romer (2004) and Ben Zeev & Pappa (2017) for monetary and fiscal simulation DGPs, respectively.

[21] The remaining columns in H are chosen arbitrarily to satisfy the variance-covariance constraint; these columns only matter for the simulation results through the implications for reduced-form dynamics.

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, 2022), 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 with this estimated signal-to-noise ratio. Specifically, $\sigma_\nu^2 = \frac{1}{R^2} - 1$.

A.3 More persistent factor model

The more persistent encompassing factor model studied in Section 5.5 adjusts the baseline factor model as follows. We replace the polynomial $\Phi(L) = \Phi_1 + \Phi_2 L$ in the VAR(2) process (5) for the factors by the polynomial $\tilde{\Phi}(L) = \tilde{\Phi}_1 + \tilde{\Phi}_2 L$, where $\tilde{\Phi}_1 \equiv b \times \Phi_1$, $\tilde{\Phi}_2 \equiv b^2 \times \Phi_2$, $b \equiv 0.95/\lambda_{\max}$, and $\lambda_{\max}$ is the absolute value of the largest eigenvalue of the original VAR polynomial $I - L\Phi(L)$. It is straightforward to verify that this adjustment scales all eigenvalues up by a factor of b. In particular, the largest absolute eigenvalue of $I - L\tilde{\Phi}(L)$ equals 0.95. To hold constant the explanatory power of the factors in the DFM, we multiply the matrix H by a scalar factor such that the trace of $\text{Var}(f_t)$ is left unchanged.

Appendix B Details on estimation procedures

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 estimated as follows:

1. **Observed shock.** We set x_t equal to the observed shock $\varepsilon_{1,t}$ and omit the contemporaneous controls q_t (we still control for lagged data).[22]

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

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 (2021).

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, 2021).

Penalized LP. The Barnichon & Brownlees (2019) estimator lowers the variance of LP by exploiting a prior belief in smoothness of the impulse response function across horizons. Following their preferred implementation, we model the impulse response function using B-spline basis functions. The jaggedness penalty function penalizes deviations from a quadratic function of the horizon h. We penalize impulse responses up to horizon $\bar{h} = 20$. The penalty parameter is selected in a data-dependent way using 5-fold cross-validation. We do not penalize the coefficients on the control variables in the LP. When reporting relative impulse responses (10), we divide by the *least-squares* LP estimate of the impact response of the normalization variable i_t to the structural shock.

Least-squares VAR. The least-squares VAR coefficient estimates are obtained through equation-by-equation OLS regressions. We perform a conventional Cholesky decomposition of the forecast error variance-covariance matrix and compute impulse response functions with respect to the orthogonalized shocks. 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.

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 (Plagborg-Møller & Wolf, 2021). The relative impulse response (10) is obtained by dividing by the impact response of the normalization variable i_t.

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 the policy instrument i_t.

In contrast to the above internal instruments approach, the SVAR-IV (or "proxy-SVAR") estimator of Stock (2008) is obtained by computing the reduced-form impulse responses $\hat{\Psi}_h$ ($h = 0, 1, \dots$) corresponding to a VAR in $\bar{w}_t$ (i.e., excluding z_t), and then reporting relative impulse responses (10) corresponding to the absolute structural impulse responses $\hat{\Psi}_h \hat{\gamma}$, where $\hat{\gamma}$ is the sample covariance vector of the reduced-form VAR residuals $\hat{u}_t$ and the IV z_t.

Bayesian VAR. Our BVAR implementation assumes 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). The prior is a version of the popular "Minnesota prior". Following Kilian & Lütkepohl (2017, chapter 5.2.3), we adapt this prior to a stationary setting by centering the prior of the VAR reduced-form coefficients at zero, which would imply that the time series are independent white noise processes. The prior variance hyperparameters are given by the default choices in Canova (2007, chapter 10.2.2).[23]

VAR model averaging. 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. Let $\hat{\delta}_h(r)$ denote the un-normalized, least-squares recursive impulse response estimate at some horizon h for model $r = 1, \dots, R$. We estimate $\hat{\delta}_h(r)$ from $R = 40$ 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 $p = 20$. As in Hansen (2016), the variance-covariance matrix of innovations Σ and thus the impact effect δ_0 are fixed across candidate models and treated as known without error.[24] 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 $\hat{M}(\omega_1, \dots, \omega_R) \approx E[T(\sum_{r=1}^{R} \omega_r \hat{\delta}_h(r) - \delta_h)^2]$, 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).[25] We run this optimization for the weights separately at each impulse response horizon. Relative impulse responses (10) are

[23] 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.

[24] To match the impact effect estimate in our benchmark method of least-squares VAR, we use $\hat{\Sigma}$ from the $p = 4$ VAR estimate as the true value across all the candidate models.

[25] 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).

computed by dividing the absolute impulse response by the least-squares VAR(4) impact impulse response estimate of i_t with respect to the identified shock.

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. *Quarterly Journal of Economics*, *117*(4), 1329–1368.

Brugnolini, L. (2018). About Local Projection Impulse Response Function Reliability. CEIS Research Paper, Vol. 16, Issue 6, No. 440.

Bruns, M. & Lütkepohl, H. (2021). Comparison of Local Projection Estimators for Proxy Vector Autoregressions. DIW Berlin Discussion Paper 1949.

Canova, F. (2007). *Methods for Applied Macroeconomic Research*. 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 1A chapter 2, (pp. 65–148). Elsevier.

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.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.

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. (2021a). Bayesian Local Projections. Warwick Economics Research Papers 1348.

Miranda-Agrippino, S. & Ricco, G. (2021b). Identification with External Instruments in Structural VARs. Manuscript, University of Warwick.

Montiel Olea, J. L. & Plagborg-Møller, M. (2021). Local Projection Inference Is Simpler and More Robust Than You Think. *Econometrica*, *89*(4), 1789–1823.

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

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

Plagborg-Møller, M. & Wolf, C. K. (2022). Instrumental Variable Identification of Dynamic Variance Decompositions. *Journal of Political Economy*. 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.

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, *48*(1), 1–48.

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 chapter 8, (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.