Humble VARs*

Régis Barnichon — FRB San Francisco
Geert Mesters — FRB New York
Emi Nakamura — UC Berkeley
Jón Steinsson — UC Berkeley

July 13, 2026

Abstract

Macroeconomic VARs are often interpreted as models of the entire economy, yet in practice they are typically small systems with few variables and lags. As a result, estimated impulse responses are often highly sensitive to lag length and variable selection. We argue that VARs for impulse response analysis should therefore be viewed more modestly: rather than representing the whole economy, they should only be expected to approximately capture the transmission of the shock of interest. We propose Conditional Cross-Validation (CCV), a criterion for choosing VARs optimized for impulse response estimation. Different VARs can be selected based on the outcome variable, horizon (short- vs. long-run), or the researcher's loss function.

*Preliminary and Incomplete. The views expressed in this paper are the sole responsibility of the authors and do not necessarily reflect the views of the Federal Reserve Bank of San Francisco, the Federal Reserve Bank of New York or the Federal Reserve System.

1 Introduction

A researcher is handed a sequence of macroeconomic shocks, say monetary shocks. How can they best estimate the transmission of this shock through the economy? A prominent method in macroeconomics is to use a vector autoregressive (VAR) model.[1] A $\text{VAR}_n(p)$ models the evolution of n variables $\mathbf{y}_t$ in recursive fashion as being a function of p lags of $\mathbf{y}_t$ and n shocks $\boldsymbol{\varepsilon}_t$:

$$\mathbf{y}_t = \mathbf{A}_1\mathbf{y}_{t-1} + \ldots + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{u}_t \ , \qquad \text{with} \quad \mathbf{u}_t = \mathbf{B}\boldsymbol{\varepsilon}_t \ , \tag{1}$$

where $\mathbf{A}_1, \ldots, \mathbf{A}_p$ are the autoregressive matrices that capture the lagged effects and $\mathbf{B}$ is the impact matrix of the standardized shocks $\boldsymbol{\varepsilon}_t$. Using a $\text{VAR}_n(p)$ model to estimate the transmission of a shock involves modeling 1-step ahead changes in $\mathbf{y}_t$ and iterating these forward to predict h-step ahead changes.

Sims (1980) provided a powerful motivation for VARs as a flexible—essentially atheoretical—representation of the economy, in contrast to more tightly parameterized large-scale structural macroeconometric models of the 1970s, such as Fair (1974). But while VARs are often presented as flexible statistical models (Lütkepohl, 2013), the VARs *actually* employed by applied researchers often embed assumptions that are quite strong—much stronger, for example, than the surrogacy assumption recently proposed in the dynamic treatment effects literature (e.g., Athey et al., 2025).

The reason for this is that VARs used in practice are typically small-order systems, i.e., systems with a small number of endogenous variables and a small number of lags. A canonical example is the popular $\text{VAR}_3(4)$ with $n = 3$ endogenous variables (e.g., output, inflation and the interest rate) and $p = 4$ quarterly lags. The short sample sizes typically encountered in macro (50 years of quarterly data, often much less) naturally lead researchers to limit the number of parameters they need to estimate.

For VARs to be considered atheoretical representations of the economy, they must provide a complete statistical model of the economy. This is a remarkably ambitious goal: to model the entire unconditional autocovariance structure of the economy with only p lags of n variables. It means taking account of *all* variables and *all* shocks that affect the economy at a macro level. Is it plausible that a VAR with four quarterly lags of three variables can achieve this?

Consider the myriad shocks that have affected the U.S. economy over the past 75 years. There are the "standard" ones: monetary shocks, fiscal shocks, productivity shocks, oil shocks, etc. But there are also many other shocks: 9/11, Covid, the Nixon price controls, the 1980 credit controls, various strikes, the Korean War, the 1975 tax rebate, the savings

[1] See e.g., Sims (1980); Hamilton (1994); Lütkepohl (2005); Canova (2007); Kilian and Lütkepohl (2017).

and loans crisis, the failure of Lehman Brothers, the rise of female labor supply, and the list could go on. For a VAR to be a complete statistical model of the economy, it must capture the effects of all of these shocks.[2]

When the dynamic evolution of the economy is affected by more variables or more shocks than one is able to include in a VAR, one can "solve out" for the additional variables and shocks.[3] But this typically transforms one's original finite order VAR representation of the economy into a VAR(∞) in the variables that one has data on. Since it is not feasible to estimate a VAR(∞), one must truncate the VAR in practice. This means that the VAR is no longer a complete statistical model of the economy. It becomes misspecified, and more misspecified the smaller is the chosen VAR.

To illustrate the problem, consider the transmission of monetary policy. A researcher is interested in estimating the effects of monetary policy on GDP growth, inflation, and the federal funds rate. They have access to a series of exogenous changes in the policy rate. With the shock series in hand, they can construct the empirical distribution of forecast errors following a unit monetary policy shock.

The top row of Figure 1 does this. The data used are quarterly changes in GDP and the PCE price index, as well as the federal funds rate. The sample period is 1969q1-2007q4. The monetary shock series is from Romer and Romer (2004). These forecast error distributions can be seen as a method to visualize the "raw data" underlying impulse response estimates.[4] By construction, the mean of the forecast error distribution coincides with the local projection point estimate (Jordà, 2005).

The "raw data" already make out a few stylized facts about the effect of monetary policy shocks: a persistent rise in the policy rate, an initial drop in GDP growth followed by a rebound, and a delayed but large and persistent decline in inflation. Note however the large dispersion in the forecast error distributions.

Consider instead using a standard VAR$_3$(4) to estimate the transmission of monetary policy. Since a set of policy shocks has already been identified, a simple way to estimate dynamic causal effects is to add the shock series to the VAR, order it first, and compute impulse responses to innovations to the shock series (e.g., Plagborg-Møller and Wolf, 2021*a*). The bottom row of Figure 1 overlays impulse responses estimated in this manner on the raw estimated forecast error distributions.

Notice that the 95% bootstrap confidence bands from the VAR are substantially narrower than the forecast error distributions. By imposing restrictions on impulse responses—across

[2]Each of these shock types should actually be counted as several different shocks to the extend that different instances have different expected dynamics. For example, a forward guidance shock is a different shock from a target rate monetary shock.

[3]I.e., iteratively substitute for these variables throughout the VAR model ("solve backward").

[4]Intuitively, given a forecast based on data observed before a monetary shock, we can compute the distribution of forecast errors following a unit monetary shock. See the Appendix for details.

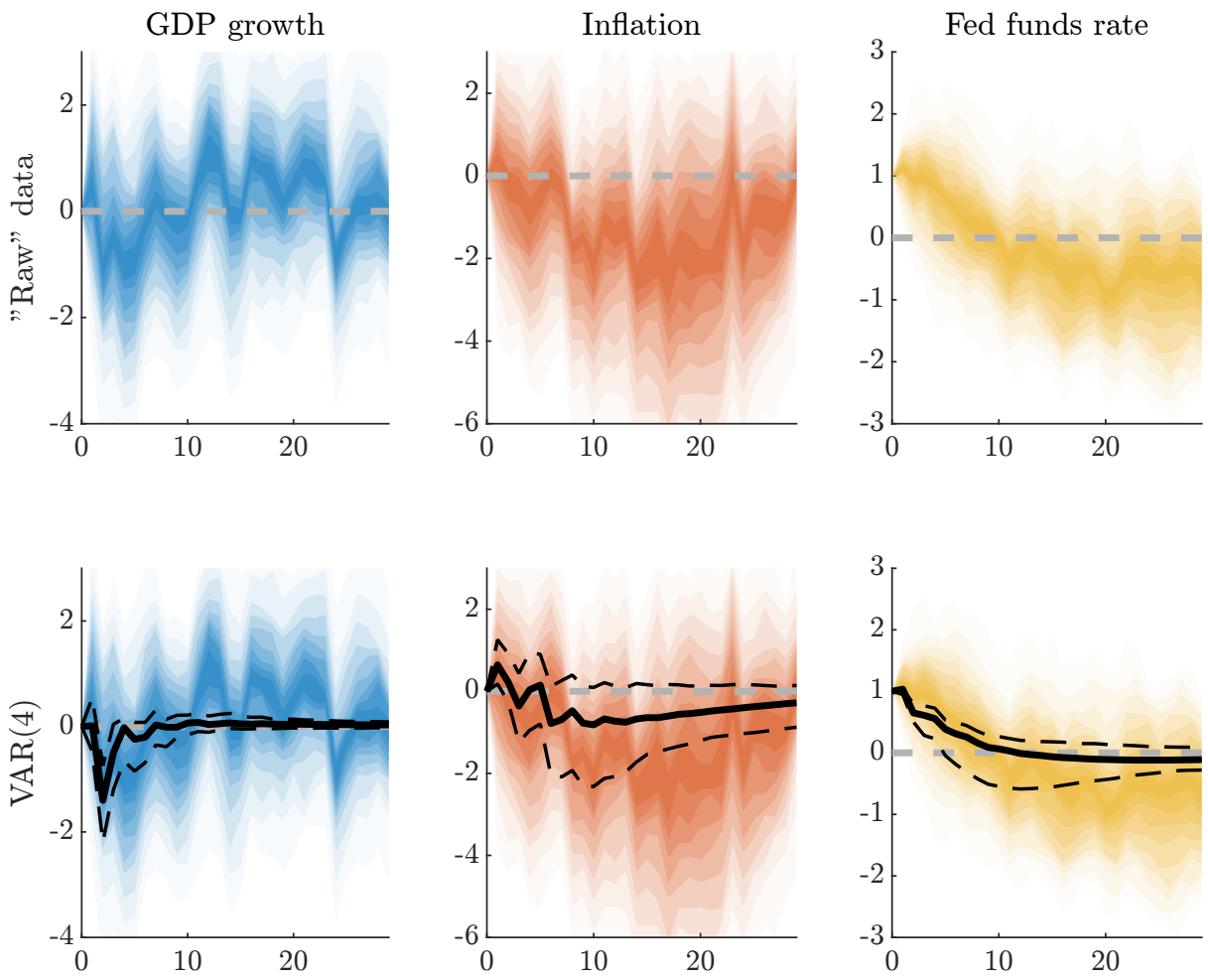


Figure 1: Illustration: Effect of monetary policy

Note: Quantiles (5 percent intervals) of realized outcomes for GDP growth, inflation and the fed funds rate conditional on a unit monetary shock at h=0. Darker colors indicate quantiles closer to median realization. Thick black lines are the VAR(4) points estimates along with the 95 percent confidence bands (dashed lines).

horizons, variables, and shocks—the $\text{VAR}_3(4)$ achieves substantial power gains and enables informative inference. At the same time, however, the VAR impulse responses are at odds with the raw data in several important respects. In other words, the $\text{VAR}_3(4)$ shows signs of substantial bias: the GDP growth response does not display any sign of rebound after two years, and the response of inflation is small and less persistent than the forecast error distribution.[5] In section 3, we show that these VAR impulse responses are highly sensitive to the number of variables and number of lags in the VAR.

Small-order VARs are almost certainly misspecified. They can therefore not be thought

[5]Since an LP is a consistent estimator of the transmission of a shock, the distance between the VAR-based IR point estimate and the mean of the forecast error distribution provides an indication of the size of the bias associated with the VAR-based IR estimates.

of as atheoretical representations of the economy's dynamics and likely yield biased impulse response estimates. A natural response to this issue is to turn to less parametric methods, such as local projections, as recommended by Olea et al. (2025). Local projections estimate impulse responses variable by variable and horizon by horizon. This makes them less vulnerable to dynamic misspecification.

Local projections (LP) are therefore a natural benchmark for dynamic causal inference. However, given the short samples often encountered in macroeconomics, local projections sometimes yield estimates that are sufficiently imprecise as to make the analysis uninformative. This raises the question of whether efficiency gains from making stronger assumptions may in some cases be worth the risk of introducing some bias.

We propose that, for the purpose of impulse response estimation, VARs should not be used to construct complete statistical models of the economy. Rather, their use should have a more humble objective: to improve the power of inference, i.e., as "power tools." Crucially, an impulse response is a conditional forecast. Consequently, to improve the precision of impulse response estimates, a VAR need not capture the full unconditional autocovariance structure of the economy. It needs only to approximately capture the transmission of the shock of interest. Since an impulse response is a forecast conditional on a particular shock, only the VAR's ability to forecast *conditional* on the shock of interest is relevant to improving the impulse response estimator.

This "conditional" insight implies that, for the purpose of impulse response estimation, VARs should be tailored to the specific impulse response they are being used to estimate. We propose a method for explicitly confronting the specification question of (i) which variables to include?, and (ii) how many lags? When used as a power tool (as opposed to as a complete statistical model of the economy), VARs should be specified with a "conditional objective": to select the VAR specification that provides the "best" conditional forecast according to the researcher's objective function. This objective function can be to minimize bias—in which case it will select a high order VAR (or equivalently an LP)—but it could also be to minimize mean-square-error, confidence band coverage, or some other desirable feature.

We propose a Conditional Cross-Validation (CCV) criterion to select the best model based on a researcher's objective. The CCV model selection procedure consists of choosing the VAR specification that minimizes the *out-of-sample* difference between the VAR-based impulse response and the impulse response estimated from a linear projection on the shock (i.e., an LP). Since VARs parameterize impulse responses as a sum of damped cosine basis functions, one can think of the CCV methodology as selecting the set of basis functions that produces the best out-of-sample predictions.

Our CCV procedure bears some resemblance to the Hausman test, but differs in that the two IR estimates (VAR vs. LP) are estimated over different samples. This is crucial since

otherwise the procedure would increase the size of the VAR to fit the local projection impulse response exactly (at the cost of overfitting out of sample). Comparing the two estimators over different samples avoids this pitfall.

We show that CCV can be specified to be consistent for either the population bias or mean squared error of the impulse response, allowing the researcher to choose her prefer target criteria. The result holds when the true data generating process is a stationary process subject to mild regularity conditions. A set of simulations verify the properties of CCV selection for finite samples.

Literature review

Since the seminal contribution of Sims (1980), VARs have become a central tool for empirical macroeconomic analysis. Early influential contributions include the perspective of policy analysis of Sims (1986), the structural applications of VAR of Bernanke (1986) and Blanchard and Quah (1989), as well as the Bayesian literature on VAR of Doan, Litterman and Sims (1984) and Litterman (1986).[6] The classical perspective of VAR views the VAR as a reduced-form approximation to the joint dynamics of the macroeconomy that can be used for forecasting, policy analysis and impulse response estimation.

Since the VAR inception, two views of VARs have co-existed among the macro and econometrics literature. On the hand hand, there is the view that the VAR is a reduced-form model of the economy, that is a model of the underlying Data Generating Process (DGP). Informally, this idea goes back to the Wold decomposition theorem. As any covariance-stationary vector time series process can be written as an invertible vector moving-average (VMA), one can invert the VMA, obtain a VAR(∞) representation of the data, which is then truncated in order to be estimable on finite dataset. In this perspective, the VAR is viewed as a statistical model of the underlying true DGP. This approach has been extensively used by the structural VAR literature, where the VAR is used to identify structural economic shocks using short-run restrictions (e.g., Sims, 1980; Bernanke, 1986; Christiano, Eichenbaum and Evans, 1999), long-run restrictions (e.g., Blanchard and Quah, 1989), or sign restrictions (e.g., Uhlig, 2005).

On the other hand however, the macro-econometric literature has long been aware that VARs are potentially mis-specified models, which can nonetheless be useful for forecasting or even impulse response estimation by trading some bias for lower variance (e.g., Lewis and Reinsel, 1985; Lütkepohl and Poskitt, 1991; Poskitt, 1992; Inoue and Kilian, 2002; Kuersteiner, 2005; Schorfheide, 2005*a*; Plagborg-Møller and Wolf, 2021*b*; Olea et al., 2026; González-Casasús and Schorfheide, 2025*a*; Olea et al., 2025). The Bayesian VAR exemplifies

[6]The statistical theory of VAR models was largely developed at that time and expounded in textbooks like Hannan (1970).

this view, where large VARs are routinely estimated in combination with a prior (e.g., Minnesota or other long-run priors), see Baumeister (2025).

In this work, we argue that this latter view —the VAR is a power tool— should be embraced to its full conclusion: any finite lag VAR is mis-specified and the choice of VAR specification should be guided not by the desire to capture the true DGP but rather by the researcher's objective of best estimating the impulse response of interest.

A large literature studies the conditions under which VARs can recover structural shocks from observed macroeconomic variables. Building on the non-fundamentalness and invertibility literature (e.g. Lippi and Reichlin, 1993; Fernández-Villaverde et al., 2007; Forni and Gambetti, 2014; Canova and Ferroni, 2022), these works emphasize that finite-dimensional VARs may fail to recover structural shocks when the observed variables do not span the relevant information set of economic agents. This concern motivated information-rich specifications such as factor-augmented VARs (Bernanke, Boivin and Eliasz, 2005), large VARs (Banbura, Giannone and Reichlin, 2010), and external instrument approaches (e.g. Stock and Watson, 2008; Mertens and Ravn, 2013; Gertler and Karadi, 2015). Our paper studies a complementary problem: even when the shock is externally identified or directly observed, omitted variables may still distort the propagation dynamics of the economy and thereby bias estimated impulse responses.

Given this concern, it is tempting to move away from recursively propagated VAR impulse responses toward direct estimation methods (Dufour and Renault, 1998; Dufour, Pelletier and Renault, 2006; Jordà, 2005). The most prominent example is the local projection (LP) estimator of Jordà (2005), which estimates impulse responses horizon by horizon and therefore imposes fewer restrictions on the dynamics of the economy. Recent work shows that LPs and VARs with sufficiently many lags share the same population estimand (Plagborg-Møller and Wolf, 2021*a*), while differing in finite samples through the bias–variance trade-off (Kilian and Kim, 2011; Brugnolini, 2018; Choi and Chudik, 2019; Bruns and Lütkepohl, 2022; Li, Plagborg-Møller and Wolf, 2024; Olea et al., 2024, 2025; Ludwig, 2024). Our paper proposes a method to navigate that trade-off with the conditional cross-validation criterion, which relates to the literature on lag-length and model selection in dynamic systems. Classical procedures such as sequential testing and information criteria are primarily designed for forecasting performance or recovery of the correct dynamic representation, see Lütkepohl (2005) and Kilian and Lütkepohl (2017) for textbook treatments. Several papers recognize that these objectives may differ from accurate impulse response estimation. Hsiao (1979) emphasized variable-specific specification choices, while Schorfheide (2005*b*), Lohmeyer et al. (2019), and González-Casasús and Schorfheide (2025*b*) study lag-selection procedures targeting multistep prediction or impulse response risk. Our paper contributes to this literature by proposing a conditional cross-validation (CCV) approach for model specification in im-

pulse response analysis. The key idea is to condition the evaluation criterion on the shock of interest, thereby targeting the ability of a specification to recover the relevant impulse responses rather than its unconditional forecasting performance. The framework is flexible and can be applied to lag-length selection, variable selection, surrogate horizon selection, regularization choices, and combinations of direct and recursive estimators. More broadly, the paper relates to the out-of-sample forecasting and cross-validation literature in econometrics and statistics (e.g. Elliot and Timmermann, 2016, Part III), while adapting these ideas to shock-specific impulse response estimation.

2 VARs and the Omitted Variable Bias: Illustration

Consider a simple macro model with a dynamic Phillips curve, IS curve and policy rule. We have a Phillips curve with adaptive expectations ($\mathbb{E}_{t-1}\pi_t = \pi_{t-1}$)

$$\pi_t = \rho_\pi \pi_{t-1} + \lambda x_t + \nu_t \ , \tag{2}$$

where x_t is the output gap, ν_t is an iid cost-push shock. The dynamic IS equation is

$$x_t = \rho_x x_{t-1} - \sigma^{-1}(i_t - \pi_t) + \zeta_t \ , \tag{3}$$

with ζ_t a natural interest rate shock, and the interest rate rule is

$$i_t = \rho_i i_{t-1} + \phi_\pi \pi_t + \varepsilon_t \ , \tag{4}$$

where ε_t is a monetary shock. All coefficients on lagged terms are strictly less than one (the ρ's).

A researcher is interested in estimating the dynamic causal effects of a change in the central bank policy rate: the impulse responses to a policy shock. A time series of exogenous changes in the policy rate —a sequence of shocks— has already been identified, so that the question is strictly about choosing the best estimation method: How to obtain an accurate and precise estimate of the impulse responses.

One popular approach is to rely on local projections (LP Jordà, 2005), which impose little restrictions on the underlying data generating process except for linearity. However, we can immediately see that LP is not the most efficient approach here. In this example, the data *are* highly structured, as the structural macro model can be captured by a VAR(1) with 3 endogenous variables (π_t, x_t, i_t). Therefore, given a series for monetary shocks ε_t (or a proxy), a researcher interested in learning the effects of monetary policy should estimate a VAR(1) with $(\varepsilon_t, \pi_t, x_t, i_t)$. In short samples, a VAR(1) can deliver large efficiency gains

relative to local projections, because local projections do not exploit the underlying structure of the data.

The VAR must be correctly specified however; this choosing the correct set of endogenous variables and the correct number of lags. Unfortunately, the correct specification of the VAR is very fragile. To illustrate this, we will consider two simple twists to the model: (i) adding an endogenous variable (leverage), which meditates the effect of monetary policy, and (ii) adding an exogenous non-iid process (oil price inflation) to the determinants of inflation.

The set of endogenous variables

Consider allowing for a credit channel of monetary policy. Specifically, changes in monetary policy can affect the quantity of available credit and thus leverage, which affects the output gap (e.g., Bernanke and Gertler, 1995; Gilchrist and Zakrajšek, 2012). The IS equation becomes

$$x_t = \rho_x x_{t-1} - \sigma^{-1}(i_t - \pi_{t-1}) - \delta \ell_{t-1} + \zeta_t \tag{5}$$

where ℓ_t is the leverage–GDP ratio with

$$\ell_t = \rho_\ell \ell_{t-1} + \beta i_t \tag{6}$$

with $\rho_\ell < 1$. Note how the policy rate affects leverage, such that the effect of monetary policy on the economy is *mediated* through the amount of leverage. Collecting all variables in $\tilde{\mathbf{y}}_t = (\pi_t, x_t, \ell_t, i_t)$, it is easy to see that $\tilde{\mathbf{y}}_t$ follows a VAR(1). Thus, estimating a VAR(1) with with $(\varepsilon_t, \tilde{\mathbf{y}}_t)$ will deliver consistent IR estimates.

Imagine however that the researcher omits leverage from the list of explanatory variables and considers describing this economy with a VAR for $\mathbf{y}_t = (\pi_t, x_t, i_t)$. Simple substitution of ℓ_t in the IS curve the immediately delivers that the underlying VAR becomes a VAR(∞). In other words, omitting one variable from the VAR makes the correct lag specification go from $p = 1$ to $p = \infty$.[7] In that case, there simply does not exist a finite order VAR for $\mathbf{y}_t$ that will consistently estimates the dynamic effects of monetary policy, *even if* a series of monetary shocks has already been identified.

In other words, a researcher using a VAR($p < \infty$) with $(\varepsilon_t, \mathbf{y}_t)$ will not get a consistent estimate for the effect of monetary shocks past horizon p, because she did not control for

[7] To see that, just combine (5) and (6) and iterate backwards to get

$$\begin{aligned} x_t &= \rho_x x_{t-1} - \sigma^{-1}(i_t - \pi_{t-1}) - \delta(\rho_\ell \ell_{t-2} + \beta i_{t-1}) + \zeta_t \\ &= \rho_x x_{t-1} - \sigma^{-1}(i_t - \pi_{t-1}) - \delta(\rho_\ell(\rho_\ell \ell_{t-3} + \beta i_{t-2}) + \beta i_{t-1} + \varkappa_{t-1}) + \zeta_t \\ &= \ldots = A(L)(\pi_t, x_t, i_t) + \zeta_t \text{ with } A(L) \text{ of infinite order.} \end{aligned}$$

ℓ_t. Intuitively, this is similar to an omitted variable bias, and this example illustrates how correct VAR specification can be seen as an *identifying* assumption; an exclusion restriction. In our example, a VAR(1) for $(\varepsilon_t, \pi_t, x_t, i_t)$ will be correctly specified if the underlying model satisfies the exclusion restriction $\delta = 0$: the effect of monetary policy on output and inflation is not independently mediated by the reaction of leverage. Under that light, the choices of n and p —the VAR specification question— can be seen as important as the identification scheme, the "traditional" identifying assumption. If the exclusion restriction fails, no finite-p VAR for $(\varepsilon_t, \pi_t, x_t, i_t)$ can consistently estimate the dynamic causal effects of monetary shocks. In other words, CLS can be satisfied for a four-variable VAR but it is impossible to satisfy for the three-variable VAR without leverage.

The number of lags p

As another example, consider the case where oil price inflation enters the Phillips curve with

$$\pi_t = \rho_\pi \pi_{t-1} + \lambda x_t + \gamma \pi_{t-1}^{oil} + \nu_t \tag{7}$$

where π_t^{oil} is energy price inflation, which follows an exogenous MA(1) process

$$\pi_t^{oil} = \xi_t^{oil} + \alpha_1 \xi_{t-1}^{oil} \tag{8}$$

with ξ_t^{oil} an iid shock, such that oil price inflation is an exogenous process that follows an MA(q).[8]

This time, the true underlying model is a VARMA(1,1), and including oil price inflation into the VAR will not fix the problem. A VAR with the vector $\tilde{\mathbf{y}}_t = (\pi_t, x_t, \pi_t^{oil}, i_t)$ will still suffer an omitted variable bias, as the only VAR representation of a VARMA(1,1) is a VAR(∞). The nature of the OVB is more challenging to address, because the omitted variable is now a latent-state; the shock ξ_t^{oil}. In additional, unlike with the endogenous "leverage" variable, the problem is not related to omitting a variable that *mediates* the effect of monetary policy. Instead, the problem comes from the fact that there is an exogenous process —oil price inflation—, which does affect the time series properties of inflation but is not explicitly modeled in the VAR. As a result, a VAR with a finite number of lags cannot capture the unconditional autocovariance property of the data: the exogenous process does affect the time series properties of $\mathbf{y}_t$ in a non-autorecursive way, and this will bias the IR estimates of any finite lag length VAR.

[8] Other examples of processes with an MA-type structure include temporary price controls, one-time events like COVID, or processes unrelated to cyclical factors, for instance health-care price inflation (Mahedy and Shapiro, 2017).

Taking stock

These examples illustrate how VARs are highly exposed to mis-specification risk and omitted variable bias: in order to capture the transmission of shocks, a VAR(p) must include *all* relevant endogenous and exogenous variables that affect the outcome variables $\mathbf{y}_t$. Thus, even though identification per se is not an issue when the shock is entered in the VAR, an omitted variable bias can still affect the IR estimates.

To avoid this issue, one strategy would be to rely on local projections.[9] However, this would mean giving up on any underlying structure in the economy that could be exploited for efficiency gains. In the case of leverage, there does exist a (larger) VAR(1) that can capture the underlying DGP, and in the case of oil price inflation there does exist a VARMA(1,1) that can capture the underlying DGP. And while a VARMA(1,1) has only an equivalent VAR(∞) representation, a VAR(p) with $p > 1$ may provide a reasonable approximation of the transmission of monetary policy and yield more efficient estimates than regular LP.

In fact, that latter idea —truncating the VAR(∞) representation of the DGP— informally underlies most VAR applications. Unfortunately, lag selection and choice of endogenous variables are rarely justified in practice, with users reverting to "standard" specifications and typical lag structures of 4 quarters or 12 months. Lag selection methods such as AIC/BIC are rarely used as they are often not informative, as we will see below. More importantly, traditional lag length specification criteria do not target IR estimation as their objectives. Instead, they are based on unconditional one-step ahead forecasting performances objectives. Because they do not explicitly take into account the bias inherent to VAR-based IR estimates, they do not allow to use VARS while "controlling" for bias, as forcefully emphasized by Olea et al. (2025).

The objective of this paper is to provide a method to exploit the underlying structure in the data through a VAR, but taking IR-estimation bias into account. Indeed, even if the VAR is mis-specified, and the VAR-based IR estimates are biased, it may still be possible to control that bias while obtaining efficiency gains relative to conventional local projections.

3 Understanding the sources of the VAR bias

Although it is well known that finite-order VARs are prone to bias (Olea et al., 2025), it is useful to examine the sources of this bias more closely. We consider two complementary perspectives. The first is a VAR-as-DGP perspective, which generalizes the previous ex-

[9]Relatedly, a Bayesian response to this observation is to simply consider large VAR systems with many lags and regularize the estimation of the VAR parameters using a (e.g., Minnesota prior). This is a popular strategy, see comments of Baumeister (2025), that can work well in practice. We will come back to this approach in Section **??**.

amples by formalizing the assumptions under which a VAR consistently estimates impulse responses. The second is a VAR-as-power-tool perspective, in which the VAR is viewed as a parametrization of impulse responses: each impulse response is written as a sum of damped cosine functions at different frequencies. This parametrization can reduce variance, but it can also generate substantial bias when the number of basis functions is too small.

3.1 The VAR-as-DGP perspective

Consider a general representation of the economy: the Slutzky-Frisch paradigm, which represents the path of observed macro variables as arising from current and past shocks (e.g., Stock and Watson, 2018). Specifically, for the vector $\mathbf{d}_t \in \mathbb{R}^d$ we have[10]

$$\mathbf{d}_t = \sum_{j=0}^{\infty} \boldsymbol{\Theta}_j \boldsymbol{\eta}_{t-j} \ , \qquad \boldsymbol{\eta}_t \sim \text{WN}(\mathbf{0}, \mathbf{I}) \ , \tag{9}$$

where $\boldsymbol{\Theta}_j \in \mathbb{R}^{d \times d_\eta}$ are coefficient matrices and $\boldsymbol{\eta}_t \in \mathbb{R}^{d_\eta}$ are the shocks which form a white noise sequence with mean zero and identify variance. We think of $\mathbf{d}_t$ as the universe of variables required for describing the economy and the shocks can be structural — i.e. have clear economic interpretation — but may also capture uninterpretable measurement error.

Without loss of generality we define $\mathbf{y}_t \equiv \mathbf{d}_{1:d_y,t} \in \mathbb{R}^{d_y}$ and η_{1t} as the outcome variables and structural shock of interest, respectively. We aim to recover the causal effect of η_{1t} on the outcome variables $\mathbf{y}_{t+h}$ at horizon h. That is

$$\boldsymbol{\theta}_h \equiv \boldsymbol{\Theta}_{1:d_y,1,h} = \mathbb{E}(\mathbf{y}_{t+h}|\eta_{1t} = 1) - \mathbb{E}(\mathbf{y}_{t+h}|\eta_{1t} = 0) \ , \tag{10}$$

To estimate the dynamic causal effects θ_h, we consider a $\text{VAR}_n(p)$ —equation (1)— with n endogenous variables and p lags for the vector $\mathbf{y}_t \in \mathbb{R}^n$.[11] Whether or not this VAR is able to recover the impulse responses of interest $\boldsymbol{\theta}_h$ depends on two conditions: (i) whether the shock of interest is included in the VAR, i.e. whether $\varepsilon_{1t} = \eta_{1t}$, noting that the ordering of the shocks is arbitrary, and (ii) whether the lag length of the VAR is correct.

Condition (i) combines invertibility — i.e. the shock of interest is included in the linear span of the VAR variables — and identification — i.e. the first column of $\mathbf{B}$ is correctly identified. There exists a large VAR literature that discusses the various approaches by which conditions (i) can be ensured to hold (Ramey, 2016). A simple way to ensure that

[10] We implicitly assume that suitable transformations have been made that ensure that $\{\mathbf{d}_t\}$ is a covariance stationary and purely non-deterministic process.

[11] Note that we assume that the vector of outcome variables is the same as the vector of endogenous variables in the VAR. This could be easily relaxed by separating the vector $\mathbf{w}_t \in \mathbb{R}^n$ to be included in the VAR, and the vector $\mathbf{y}_t$ of outcome variables that are included in $\mathbf{w}_t$. We abstract from this distinction to ease on notations.

this condition holds is by including a measure of the shock in the VAR as we did in the previous section. We will not discuss Invertibility in any more detail, as this is not the focus of this paper.

Condition (ii), however, has received much less attention in the literature and takes center stage in this paper. The condition requires that the lag length of the VAR is correctly selected, i.e., the errors are not serially correlated:

$$\mathbb{E}(\mathbf{u}_t|\mathbf{u}_{t-1},\ldots,\mathbf{u}_{t-p}) = 0\ . \tag{CLS}$$

Although this correct lag specification (CLS) requirement is clearly stated in more methodological treatments of VARs (e.g., Lütkepohl, 2005), its importance is often neglected in macroeconomic applications, where lag lengths are frequently chosen in an ad hoc manner—for example, using four lags with quarterly data or twelve lags with monthly data. Leaving serial correlation in the VAR innovations biases the resulting impulse responses, because the VAR fails to account for the additional dynamics contained in $\mathbf{u}_t$.

In other words, even if the identification problem has been solved by e.g. including an observed shock, a VAR remains prone to omitted variable bias through the interacted choices for n and p. Getting these choices wrong leads to biases and thus sensitive impulse response estimates as documented in the previous section. And importantly, as the previous macro example showed, there may simply not even exist a finite-order VAR that can capture the underlying DGP.

In that context, specifying the correct VAR is a difficult problem. Most traditional VAR specification methods (i.e., lag selection methods, see the appendix for a brief review) start from the premise that there exists a finite-p VAR that captures the underlying true model, i.e. a VAR of the form (1) is the true data generation process, and the CLS assumption holds. For instance, under regularity conditions the BIC then asymptotically selects the true lag order (e.g., Kilian and Lütkepohl, 2017). Unfortunately, the attractive properties of traditional lag length selection methods disappear when the true DGP is not a VAR with finite lag length. More generally, just as in the applied and treatment effect literature where unobserved characteristics are a common threat to identification (Angrist and Pischke, 2015), it seems hard to justify a priori that no omitted variables exist for a given VAR representation, and existing methods for detecting such biases are sensitive to the choice for the true data generating process (e.g., Olea et al., 2025).

3.2 The IR parametrization perspective

Instead of viewing a VAR as a model of the economy, we can equivalently see a VAR as a parametrization of the impulse responses to shocks, where each impulse response is written

as a sum of damped cosines of different frequencies. This perspective makes clear that (i) a VAR can be seen as an IR estimator, which lies on the bias-variance spectrum depending on the n and p choices —a VAR is just a power tool rather than a complete statistical model of the economy—, and (ii) a parsimonious VAR may or may not be an efficient IR estimator depending on the transmission of the shock. VARs are great at parsimoniously capturing geometrically decaying (possibly oscillating) transmission channels, but they are less good at capturing protracted responses, overshooting patterns with increasing magnitude or delayed reversals.

While a VAR model (1) is a system of linear difference equations —a parametrization of the laws of motion of the endogenous macro variables—, a $\text{VAR}_n(p)$ can also be seen as imposing restrictions on the shapes of the impulse responses to shocks. To see that, multiply (1) by the shock $\varepsilon_{s,t-h}$ and take expectations to get[12]

$$\mathbf{IR}_{s,h} = \begin{cases} \mathbb{E}(\mathbf{y}_t \varepsilon_{s,t-h}) & \text{if} \quad h \in \{0, ..., p\} \\ \mathbf{A}_1 \mathbf{IR}_{s,h-1} + \ldots + \mathbf{A}_p \mathbf{IR}_{s,h-p} & \text{if} \quad h \in \{p+1, \ldots, H\} \end{cases}, \tag{12}$$

which states that the IR of $\mathbf{y}_t$ to the shock ε_{st} —$\mathbf{IR}_{s,h}$— satisfies the system of difference equation (12) with initial conditions given by $\mathbb{E}(\mathbf{y}_t \varepsilon_{s,t-h})$ where the expectation is computed under (1) assuming that the errors are serially uncorrelated.

Expression (12) highlight two important points: First, the initial conditions imply that the VAR-based impulse responses do not impose any restrictions on the autocovariances until horizon p, which explains how the VAR estimates the IR consistently up until horizon p (see e.g., Plagborg-Møller and Wolf, 2021*a*). However, beyond horizon p the VAR propagates the effect of the shock through the difference equation, and restrictions are imposed. Second, the same difference equations (11) must hold for *any* shock. This means that a $\text{VAR}_n(p)$ imposes that the impulse responses of $\mathbf{y}_t$ to *any* shock follow the *same* difference equation. In other words, the VAR imposes restrictions on the impulse responses across variables, across horizons and across shocks, resulting in a substantial dimensionality reduction. In particular, by imposing that the same difference equation governs the IRs to all shocks, it is possible to estimate the VAR on unconditional data —this is the benefit of modeling the entire system governing $\mathbf{y}_t$, in contrast to a local projection which only estimates the IR of interest (and thus faces a larger unexplained residual).[13] Clearly, if the true impulse

[12] Specifically, we have

$$\mathbb{E}\mathbf{y}_t \varepsilon_{1,t-h} = \mathbf{A}_1 \mathbb{E}\mathbf{y}_{t-1} \varepsilon_{1,t-h} + ... + \mathbf{A}_p \mathbb{E}\mathbf{y}_{t-p} \varepsilon_{1,t-h} + \mathbf{B} \underbrace{\mathbb{E}\varepsilon_t \varepsilon_{1,t-h}}_{=0}. \tag{11}$$

[13] Specifically, the reader may wonder how a VAR performs better than a "parametrized-LP"-type estimator where the LP-based IR estimate is constrained as as a sum of damped cosines. The reason lies in the joint modeling aspect of VARs: a VAR simultaneously models the joint dynamics of multiple variables.

responses satisfy these restrictions the VAR is attractive over LP.

To better understand how the VAR restrictions affect the shape of VAR-based impulse responses, we can explicitly solve the difference equations (11).

Lemma 1. The solution to the p-th difference equation (11) is given by

$$\boldsymbol{\varphi}_h = \sum_{b=1}^{B} \mathbf{c}_b \rho_b^h \cos(2\pi h/\tau_b + \varphi_b) \tag{13}$$

where $\rho_b < 1$ is the geometric decay term, τ_b the period of the cosine, φ_b the phase, and $\mathbf{c}_b$ the $n \times 1$ vector of loadings on the basis function of frequency $2\pi/\tau_b$. The number of basis functions satisfies $B \leq np/2$. □

A proof is provided in the appendix.

In words, the lemma states that positing a VAR consists in parameterizing the impulse responses with specific basis functions —geometrically decaying cosines—, and a VAR$_n(p)$ will model the impulse responses with at most $np/2$ cosines of different frequencies. The larger p, the larger the number of cosines, and the larger the set of impulse response functions that the VAR can capture.

The functional form of the basis functions is particularly instructive. Because of the geometric decay term, a low-dimensional, short-lag VAR provides a parsimonious basis for short-lived responses that oscillate but decay geometrically (at least approximately). This makes small VARs well suited for capturing impulse responses that peak in the first horizonss after a shock and then mean-reverts toward zero. The same parametrization, however, can be restrictive when the true dynamics are more persistent or more irregular. Protracted responses require roots close to the unit circle, while multiple rebound effects of increasing magnitude or delayed reversals require combining multiple damped cosines. Since a VAR$_n(p)$ has at most $np/2$ cosines of different frequencies, a VAR with small number of variables and lags only has limited set of basis functions to parametrize the impulse responses. As a result, it may approximate long-lived or non-monotonic impulse responses only poorly, either by forcing the response to decay too rapidly or by generating artificial oscillations. In this sense, the parsimony of a small VAR comes at the cost of a strong restriction on the admissible shapes of the impulse responses.

To illustrate this point with a concrete example, we can go back to our monetary example from the introduction.[14] Figure 2 decomposes the IRs estimated from our monetary VAR$_3(4)$.

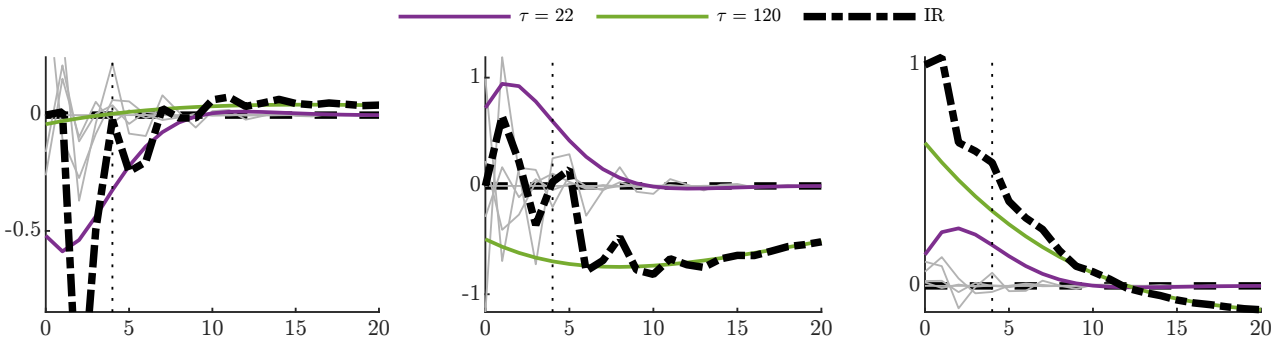
This introduces cross-equation restrictions, which (i) reduce the dimensionality of the estimation problem, but (ii) most importantly allow to exploit the full unconditional variance of the data to estimate the IR parameters. See the Appendix for more details.

[14]The appendix provides another example (from a simulated DGP) to further illustrate this point.

It shows all the basis functions in gray, and highlights the most important ones in color. Note that the basis functions are weighted by their appropriate loading (from the vector $\mathbf{c}_b$).

Two points are worth noting.[15] First, we can see that only two basis functions (with periods $\tau = 22$ and $\tau = 120$ quarters in purplus and green respectively) do most of the work at parameterizing the impulse responses. This is the reasons for the efficiency gains from the VAR: the same two damped cosines are used to propagate the effects of the monetary shocks to GDP growth, inflation and the fed funds rate. In fact, at longer horizons the impulse responses are entirely driven by only one damped cosine.[16] In this application, the *same* cosine with period $\tau = 120$ quarters (green line) drives the rebound of GDP growth, the mean reversion of inflation, *and* the mean reversion of the fed funds rate.

Figure 2: VAR IR estimates ($p = 4$) and Basis Function Decompositions



Notes: Estimated IRs along with their basis functions decompositions (grey lines), with the two largest basis functions in thick color (τ denoting their respective periodicity). In each panel, the grey and colored lines sum to the estimated IR. The dotted vertical line denotes the lag order p of the VAR. Past this horizon p, the VAR is not guaranteed to consistently estimate the IR.

Second, the VAR uses many of its basis functions to fit the short-run dynamics of the impulse responses. In the VAR$_3$(4), four of the six damped cosines are devoted to these short-

[15] Another lesson is a small-order VAR *can* capture persistent impulse responses, *even* slowly building cycles. Here for instance the estimated IR of inflation peaks much later than $h = p = 4$: the peak response of inflation occurs at $h = 12$ (a full two years after the VAR lag length). However, that persistent effect can only be of a very specific type: as a combination of damped cosines. This point can already be seen from a simple AR(2) $y_t = a_1 y_{t-1} + a_2 y_{t-2}$. That process can capture an arbitrarily slow build up of a treatment effect, because it can capture any impulse response of the form $rcos(\omega h + \phi)$ with $r \in [0, 1[$ and $\omega \in [0, 2\pi[$. For that, just use $\phi_2 = -r^2$ and $\phi_1 = 2rcos(\omega)$ (and φ given by the initial conditions). Thus, an AR(2) can capture an IR with arbitrarily slow build up, i.e., an arbitrarily low frequency ω. However, the shape of that slow building effect is severely restricted: it must follow a damped cosine.

[16] Because of the exponential decay, the number of basis functions determining the IR shrinks as only the slowest decaying basis functions —the largest ρ_b's—, survive as we increases the IR horizon. As $h \to \infty$, the IR will be driven by one basis function alone —the only basis function that has not yet decayed to zero—. And since all variables in the VAR share the same set of basis functions, this means that for h large enough the VAR(p) imposes that all IRs are given by the *same* damped cosine.

run dynamics—the gray lines in Figure 2. This is consistent with the fact that the VAR impulse-response estimator is consistent for $h \leq 4$: many basis functions are used to capture these first four horizons accurately. Thus, although a $\text{VAR}_n(p)$ may appear to provide a rich set of basis functions, many of them may be "spent" on fitting short-run dynamics. In this sense, a low-order VAR can be viewed as sacrificing long-run impulse-response dynamics in order to achieve greater accuracy at short horizons.

Increasing the number of lags

As we increase the number of VAR parameters (n or p), the number of basis functions increases, and the VAR is able to capture more complex IR structures. To see that, Figure 3 plots the estimated IRs as we increase the VAR lag length from $p = 2$ to 4, 8, 16 and 24.

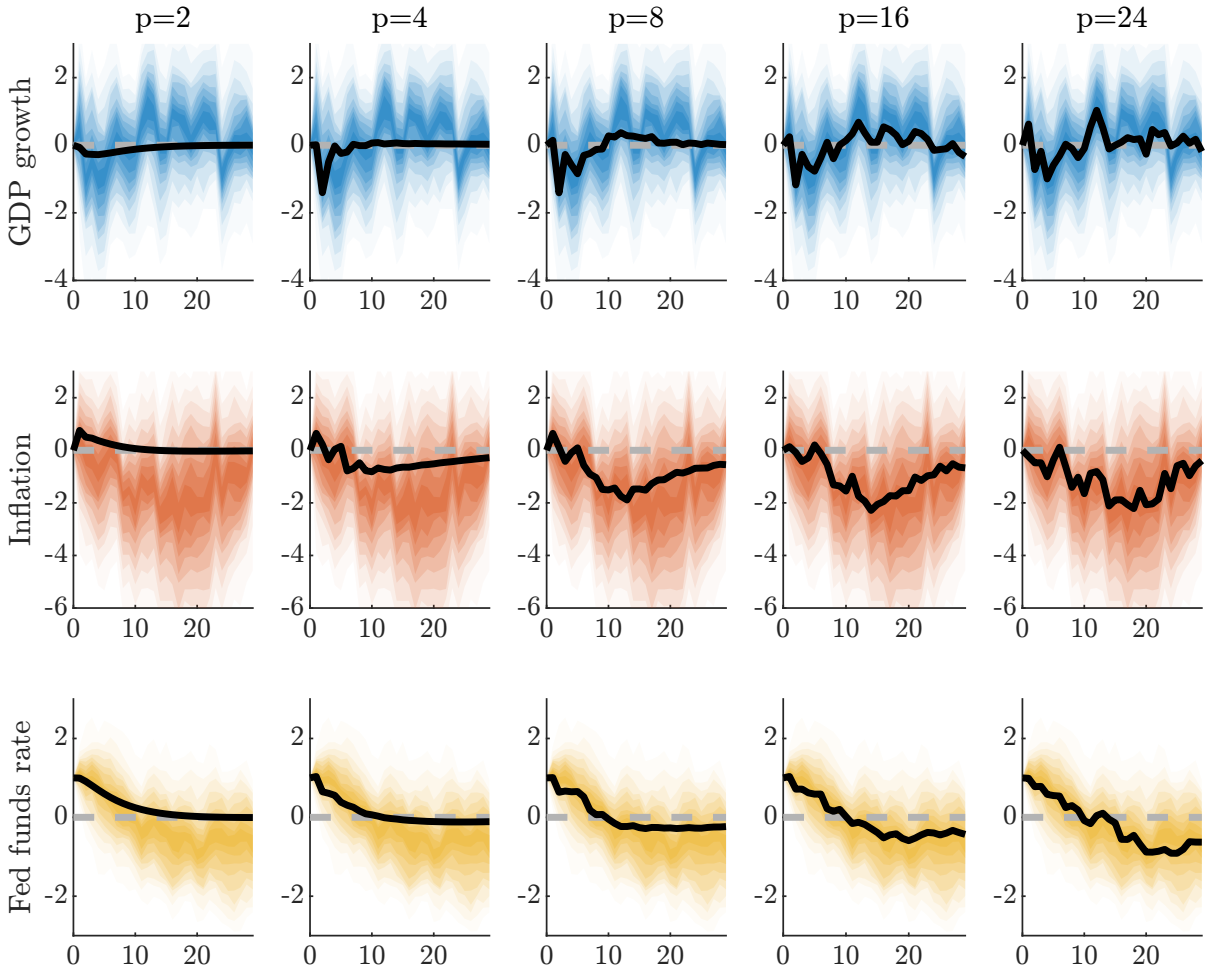
Importantly, recall that the shaded areas depict the deciles of the forecast error distribution with the mean corresponding the LP point estimate. Since LP is a consistent IR estimataor, we can informally use the distance between the VAR point estimate and the center of the forecast error distribution as a measure of the bias of the $\text{VAR}_n(p)$.

Notice how the estimated IR changes dramatically as we increase the VAR lag order with the bias allegedly becoming smaller. For inflation, the VAR initially features a strong price puzzle with no inflation decline. As the lag length increases, the price puzzles disappears, and an inflation decline starts appearing, becoming progressively more pronounced but also more delayed. Then as we increase the lag length further, the *overall* shape of the IR appears to stabilize, though the response starts becoming more jagged and choppy, typically a sign of overfitting. A similar pattern can be seen in the response of GDP growth. As we initially increase the lag length, the decline in GDP growth becomes larger and more persistent, and as we increase p further, a rebound in GDP growth starts appearing, becoming progressively larger and more persistent. After $p \approx 10$, the overall shape of the IR also appears to stabilize, but the IR starts becoming choppy. Turning to the IR of the fed funds rate, the response is more stable over the lag length choice, but it also becomes jagged for higher p values.

This example illustrates how for too small p, the VAR does not have the flexibility to capture the rich dynamics of the IRs to monetary shocks, here the rebound in GDP growth and the very delayed response of inflation. Then, as we increase p as the number of basis functions increase, the VAR can capture these IR features. As we increase p further however, the IRs become jagged, and we obtain IR estimates close to those of LP —with little bias, but also low efficiency—.[17]

[17]For illustration, Figure 7 in the appendix plots the basis functions from the VAR(16) underlying the IRs estimated in Figure 2. Note how many basis functions are "spent" on high-frequency features of the impulse responses.

Figure 3: Lag length sensitivity



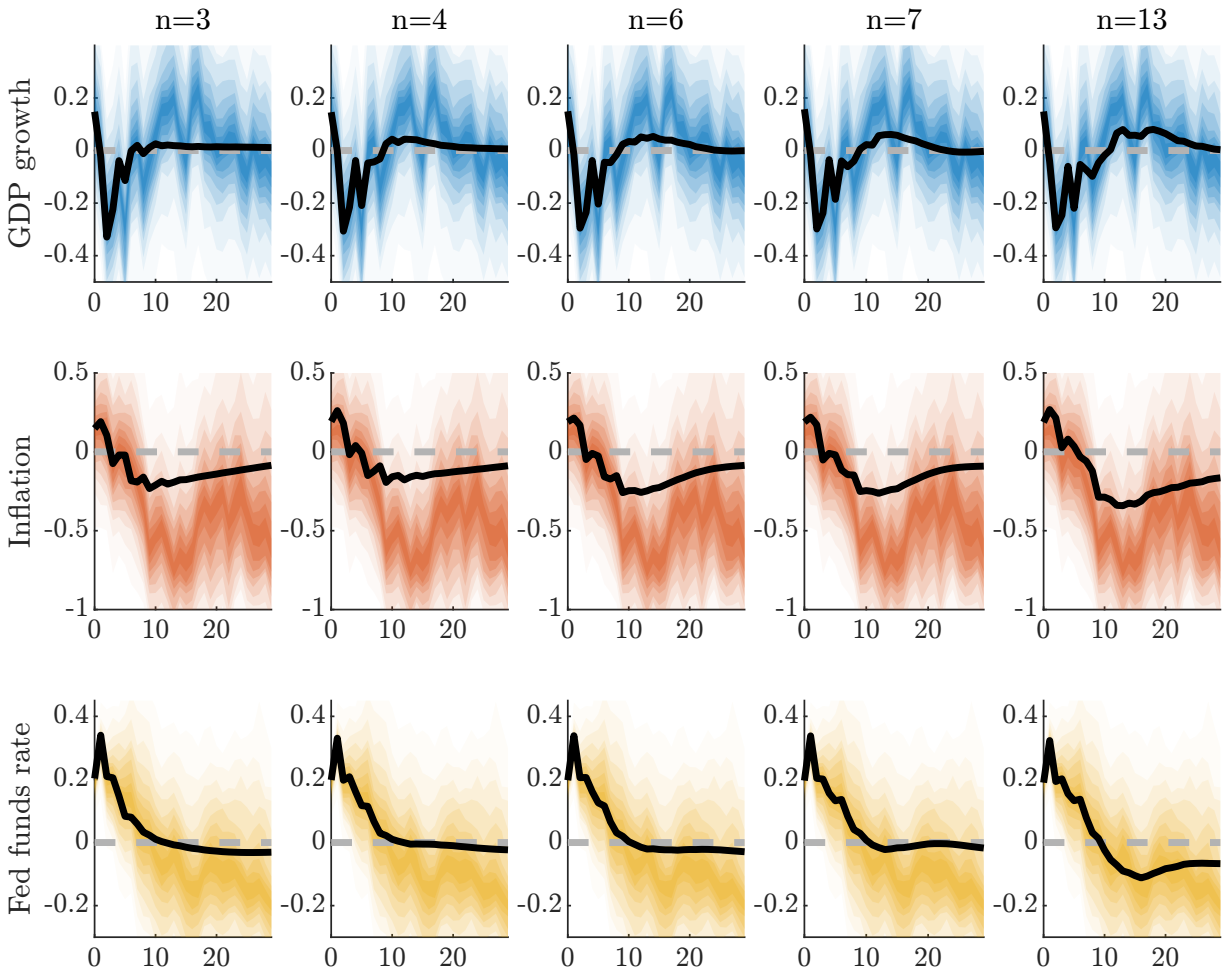
Notes: Each IR represents the IR estimated from a VAR(p) of increasing order $p = 2, 4, 8, 16, 24$.

Increasing the set of endogenous variables

The set of endogenous variables can also dramatically affect estimated impulse responses. We repeat the p-sensitivity experiment, but now vary the number of endogenous variables, increasing n from a baseline VAR with output growth, inflation, and the federal funds rate ($n = 3$) to a large VAR with $n = 13$, while keeping $p = 4$. The additional variables include the unemployment rate; consumption and investment growth; commodity price inflation; durables and non-durables consumption growth; services, durable goods, non-durable goods, and core inflation; as well as the BAA–AAA spread, total reserves, and the exchange rate. Figure 4 shows a pattern similar to the p-sensitivity exercise: as n increases, the rebound in GDP growth becomes more pronounced and the decline in inflation stronger. At the same time, larger VARs become less smooth, although this effect is more pronounced when increasing p. Interestingly, increasing n can partly compensate for a short lag length in the

GDP response, but not for inflation, where increasing p remains more important.[18]

Figure 4: VAR SIZE SENSITIVITY



Notes: Each IR represents the IR estimated from a VAR(4) with an increasing number of endogenous variables $n = 3, 4, 6, 7, 13$.

[18]Interestingly, the traditional "extra" variables in monetary VARs — such as commodity prices, spreads, reserves, or exchange rates — have little effect on the impulse responses. By contrast, disaggregated components of GDP and inflation appear much more important. Durable consumption and investment help generate the medium-run rebound in GDP growth, while service inflation produces a persistent decline in inflation. This leads to a broader point: shock identification and propagation estimation are fundamentally different tasks. The literature has mostly discussed the choice of endogenous variables in the context of identification, where the goal is to span the agents' information set and recover structural shocks from Wold innovations. Here, however, identification is handled externally using the Romer and Romer (2004) narrative shocks, allowing us to focus solely on propagation.

4 VAR selection by Conditional Cross Validation

While in general VARs are highly exposed to mis-specification risk, the previous p- and n-sensitivity exercises suggest that for some impulse responses it may be possible to select a smaller n, p pair to reduce the variance without increasing bias, or to reduce the mean squared error. To select n, p to satisfy a specific objective that the researcher has in mind for the VAR we introduce Conditional Cross-Validation (CCV) criterion as general methodology to select the "best" VAR model given a researcher's objective. The method is inspired by the literature on out-of-sample model selection, but tailored to our impulse selection objective.

4.1 Recap: pseudo out-of-sample forecasting to select the VAR

To ease into our CCV methodology, it is useful to first recap how one would select n, p using out-of-sample forecasting which is a popular approach when aiming to select the best forecasting model (e.g. Elliot and Timmermann, 2016).

Suppose that we have a sample consisting of T observations and a set of candidate VAR models $\mathcal{M}$ that is defined as

$$m \in \mathcal{M} \qquad \text{with} \qquad m = (n_m, p_m) \ ,$$

where n_m indexes the set of included variables in $\mathbf{y}_t$ and p_m defines the lag length of model m. In total the researcher considers all combinations collected in the set M. To select the preferred VAR model for forecasting we initially estimate the models under consideration on the first S observations of the sample, $S < T$, and compare the VAR prediction for the variable of interest y in period $S + h$ with its realized value. We then expand the initial sample by estimating the models under consideration on the first $S + 1$ observations and construct the prediction error for $S + 1 + h$. We continue this procedure until the estimation window extends to $T - h$. This procedure results in a sequence of $T - S - h$ prediction errors. We denote this sample of prediction error as the validation sample $\mathcal{V}$, and we denote the samples used for estimation as the training sample $\mathcal{T}$.

With the pseudo out-of-sample forecasts in hand the researcher can choose any preferred metric for selecting the model of interest based on the validation sample $\mathbf{V}$. For example, suppose a researcher is interested in forecasting the variable w_{t+h} and wants to minimize the mean-square forecast error (MFSE). She can choose the lag order corresponding to the model with the lowest recursive MSFE for the variable w, i.e.

$$(n^*, p^*) = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \frac{1}{|\mathcal{V}|} \sum_{j \in \mathcal{V}} \left(w_{j+h|j}(p) - w_{j+h} \right)^2 \ ,$$

where $w_{j+h|j}(m)$ is the h-step ahead VAR forecast for variable w.[19]

In general, out-of-sample forecasting is attractive as it allows the researcher to choose the variables, horizon and metric of interest. For instance, mean squared error can be replaced by bias or a desired quantile of the forecast error distribution. However, if impulse response estimation is the object of interest, it is clear that targeting the forecast error is inappropriate.

4.2 Conditional cross-validation

To adapt the pseudo out-of-sample forecasting approach for impulse response estimation we recall that impulse responses are conditional forecasts, where the conditioning is on the shock that defines the impulse response. We exploit this representation to construct an out-of-sample criteria that directly targets the impulse response of interest.

Set-up

To set this up, suppose that the true data generating process is given by the vector moving average model (9). For a variable w_t included in $\mathbf{d}_t$ we define the impulse response of interest as

$$\theta_h \equiv \mathbf{\Theta}_{h,w1} = \mathbb{E}(w_{t+h} \mid \eta_{1t} = 1) - \mathbb{E}(w_{t+h} \mid \eta_{1t} = 0) \ .$$

The method can be easily extended to allow for multiple variables, horizons and even shocks.

We approximate θ_h by a VAR impulse response. We assume that the shock of interest is observed and discuss the case where only a noisy measure is available below. An arbitrary subset of VAR variables is is denoted by $\mathbf{y}_t = (\eta_{1t}, w_t, \mathbf{y}_{3:n,t})$, where we always include the variable w_t and the shock η_{1t} in each subset. We consider the approximating VAR

$$\mathbf{y}_t = \mathbf{A}_1\mathbf{y}_{t-1} + \ldots + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{\Sigma}^{1/2}\mathbf{u}_t \ , \qquad \mathbb{E}\mathbf{u}_t = \mathbf{0} \ , \qquad \mathrm{Var}(\mathbf{u}_t) = \mathbf{I}_n \ ,$$

where p is the lag length. The matrices $\mathbf{A}_1, \ldots, \mathbf{A}_p$ are defined as the least squares projection coefficients. As the iid shock series η_{1t} is included in the VAR, we have that $\mathbf{\Sigma}_1 = \mathrm{Cov}(\mathbf{y}_t, \eta_{1t})$ where $\mathbf{\Sigma}_1$ is the first column of $\mathbf{\Sigma} = \mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2'}$. The approximating VAR model can be defined for different choices n, p that we collect $\mathcal{M}$.

To define the VAR impulse responses, it is useful to use the companion form representa-

[19] This procedure may be adapted to allow for forecasts of more than one VAR variable by evaluating e.g. the trace of the MSPE matrix. Inoue and Kilian (2006) establish that under the assumption that S is a fixed fraction of T, the predictive least squares method is asymptotically equivalent to the AIC.

tion. For model $m \in \mathcal{M}$ we write

$$Y_t(m) = \mathbf{A}_m Y_{t-1}(m) + \mathbf{S}_m \mathbf{u}_t \ , \qquad Y_t(m) = \begin{bmatrix} \mathbf{y}_t \\ \vdots \\ \mathbf{y}_{t-p+1} \end{bmatrix}, \qquad \mathbf{S}_m = \begin{bmatrix} \mathbf{\Sigma}^{1/2} \\ 0 \\ \vdots \end{bmatrix}.$$

The $\text{VAR}_n(p)$ forecast for w_{t+h} and impulse response of w_{t+h} with respect to η_{1t}, given known approximating coefficients, are given by

$$w_{t+h|t}(m) = \mathbf{e}_2' \mathbf{A}_m^h Y_t(m) \qquad \text{and} \qquad \varphi_h(m) = \mathbf{e}_2' \mathbf{A}_m^h \mathbf{S}_m \mathbf{e}_1 \ ,$$

where $\mathbf{e}_i$ is a vector with zeros everywhere except for a vector in position i. It is useful to note that post-multiplying the forecast by the shock and taking expectations given exactly the impulse response of interest: $\mathbb{E}(w_{t+h|t}(m)\eta_{1t}) = \varphi_h(m)$. That is, the impulse response is equal to the forecast after projecting it on the shock.

When the coefficients are estimated using data up to time T the forecast and impulse response are denoted by

$$\widehat{w}_{t+h|t}(m) = \mathbf{e}_2' \widehat{\mathbf{A}}_{m,T}^h Y_t(m) \qquad \text{and} \qquad \widehat{\varphi}_{h,T}(m) = \mathbf{e}_2' \widehat{\mathbf{A}}_{m,T}^h \widehat{\mathbf{S}}_{m,T} \mathbf{e}_1 \ .$$

We do not take a stand on how the coefficients should be estimated. In fact, we only impose high level convergence requirements below, thus allowing for popular Bayesian estimators as well as conventional least squares estimates.

Targets

We want to select the specification $m \in \mathcal{M}$ that performs best for estimating the impulse response θ_h. The notion of "best" depends on the researcher's objective. Here we consider two standard criteria: bias and mean squared error.

The population bias criterion is given by

$$m_b^* = \underset{m \in \mathcal{M}}{\text{argmin}} \, |C_{b,h}(m)|, \qquad \text{with} \qquad C_{b,h}(m) = \mathbb{E}(\widehat{\varphi}_{h,T}(m)) - \theta_h \ . \tag{14}$$

The preferred model m_b^* minimizes the bias of the impulse response estimator at horizon h. More elaborate objectives can be constructed by averaging over horizons or variables. In our simulations and empirical work, we often target $H^{-1} \sum_{h=0}^{H} |C_{b,h}(m)|$ which corresponds to the average bias across horizons.

Similarly, we can target the mean squared error:

$$m_e^* = \underset{m \in \mathcal{M}}{\operatorname{argmin}}\, C_{e,h}(m), \qquad \text{with} \qquad C_{e,h}(m) = \mathbb{E}\left[(\widehat{\varphi}_{h,T}(m) - \theta_h)^2\right] . \tag{15}$$

These objectives differ fundamentally from traditional lag-length selection criteria such as AIC or BIC, which are typically designed to optimize one-step-ahead predictive accuracy. Schorfheide (2005*b*) recognized that this objective may be inappropriate when the researcher is interested in dynamic responses and proposed criteria targeting h-step-ahead prediction errors. More recently, González-Casasús and Schorfheide (2025*b*) extend this perspective to impulse responses by studying criteria related to $C_{e,h}(m)$.

Conditional cross-validation criteria

To approximate the population criteria (14) and (15), two challenges must be addressed. First, the expectations $\mathbb{E}(\cdot)$ must be approximated. Second, the unknown target parameter θ_h must be replaced by an observable proxy.

We address the first challenge through sample splitting, similar as for out-of-sample forecasting this generates multiple out-of-sample impulse response estimates which can be combined to compute an estimate for the expected value. To address the second challenge, we exploit the fact that the structural shock (or a valid proxy for it) is observed. Under model (9) we have that

$$\mathbb{E}(w_{t+h}\eta_{1t}) = \theta_h,$$

so that $w_{t+h}\varepsilon_{1t}$ provides a noisy but unbiased proxy for the impulse response at horizon h.

Let $\pi \in (0,1)$ denote the sample splitting fraction, and define $S_\pi = \pi T$ and $S_{1-\pi} = (1-\pi)T$. The CCV estimator for the bias criterion is

$$\text{CCV}_{b,h}(m) = \left| \frac{1}{F_\pi} \sum_{s=S_\pi}^{T-h} \hat{\varphi}_{h,1:s}(m) - w_{s+h}\eta_{1s} \right| + \left| \frac{1}{F_{1-\pi}} \sum_{s=h+1}^{S_{1-\pi}} \hat{\varphi}_{h,s+1:T}(m) - w_s \eta_{1,s-h} \right|, \tag{16}$$

where F_π and $F_{1-\pi}$ denote the number of evaluation observations in each subsample.

The first term evaluates impulse response estimates using a forward expanding window, while the second uses a backward expanding window. The motivation is practical: identified macroeconomic shocks are often unevenly distributed over time. For example, monetary policy shocks may be concentrated in specific historical episodes. Using both forward and backward evaluation windows reduces the risk that the criterion ignores important shock realizations. In practice, we generally set $\pi = 0.5$.

Similarly, for the mean squared error criterion we define

$$\mathrm{CCV}_{e,h}(m) = \frac{1}{F_\pi} \sum_{s=S_\pi}^{T-h} (\hat{\varphi}_{h,1:s}(m) - w_{s+h}\eta_{1s})^2 + \frac{1}{F_{1-\pi}} \sum_{s=h+1}^{S_{1-\pi}} (\hat{\varphi}_{h,s+1:T}(m) - w_s \eta_{1,s-h})^2. \quad (17)$$

In our simulations, these cross-validation schemes perform well in balancing accurate recovery of the optimal specification with computational feasibility. More elaborate splitting procedures are also possible; see Racine (2000) for a broader discussion of cross-validation methods.

4.3 Some theoretical results

We briefly discuss the theoretical properties of the CCV criteria. Broadly speaking, the analysis requires specifying the underlying data generating process. Throughout, we assume that the economy is generated by the vector moving average representation in (9).

Our main consistency result is based on the following conditions.

Assumption 1. For the data generating process (9) we assume that

1. $\eta_{1t} = \varepsilon_{1t}$
2. $\sum_{j=0}^{\infty} \|\boldsymbol{\Theta}_j\| < \infty$
3. $\{\boldsymbol{\eta}_t\}$ is an iid sequence with $\mathbb{E}\boldsymbol{\eta}_t = 0$, $\mathrm{Var}(\boldsymbol{\eta}_t) = \mathbf{I}_{d_\eta}$ and $\mathbb{E}\|\boldsymbol{\eta}_t\|^4 < \infty$.
4. For each (m),
$$\hat{\theta}_{h,1:t}(m) \xrightarrow{a.s.} c_h(m) \qquad \text{as } t \to \infty,$$
and
$$\mathbb{E}\left[\left(\hat{\theta}_{h,1:T}(m) - \theta_h\right)^2\right] \to (c_h(m) - \theta_h)^2 \qquad \text{as } T \to \infty.$$

□

Part 1 imposes the identification condition: the observed shock ε_{1t} correctly identifies the structural shock associated with the impulse response of interest. In the appendix we extend the analysis to the case where only a proxy for the shock is available. Parts 2 and 3 impose standard regularity conditions ensuring that laws of large numbers and central limit arguments apply to $\{w_{s+h}\varepsilon_{1s}\}$. Part 4 requires that the impulse response estimator converges to some pseudo-true limit $c_h(m)$. Importantly, this limit need not coincide with the true impulse response θ_h, thereby allowing for misspecification.

Under these conditions we obtain the following consistency result.

Theorem 1. Under Assumption 1, for each $m \in \mathcal{M}$, as $T \to \infty$ we have

$$\text{CCV}_{b,h}(m) \xrightarrow{a.s.} C_{b,h}(m)$$

and

$$\text{CCV}_{e,h}(m) \xrightarrow{a.s.} C_{e,h}(m) \ ,$$

where δ does not depend on m. □

The theorem shows that, as the sample size grows and the out-of-sample evaluation windows become large, the CCV criteria converge to their population counterparts. Hence, asymptotically, the cross-validation procedure selects the same specifications that would be preferred if the true population criteria were known.

In the appendix we also study the behavior of the CCV criteria under local misspecification, where some coefficients of the VMA representation are small and cannot be statistically distinguished from zero. This framework is similar to those studied in (e.g. Schorfheide, 2005*b*; Müller and Stock, 2011; Lohmeyer et al., 2019; Olea et al., 2024; González-Casasús and Schorfheide, 2025*b*). In this case, the limits of the CCV criteria become random variables centered around the population targets, paralleling related results for information criteria in misspecified environments.

5 Empirical results: revisiting the monetary VAR

We can now revisit our monetary VAR application and use CCV to select the optimal lag length and set of endogenous variables. To illustrate how a different VAR is optimal depending on the horizon and variable of interest, we will cater our selection procedure to different variables and horizons of interest. Specifically, we optimize the VAR to best estimate (in an MSE sense) the IR of inflation, the IR of gdp growth, the IR of the policy rate, and then all three impulse responses jointly (with equal weights). In addition, we consider p selection based on the long run and based on the sum across all horizons, again imposing equal weighting. For all cases we split the sample halfway between 1960 and 2007 and we do the exercise using quarterly and monthly data setting $p^{\max} = 20$ or 60.

Table 5 reports the selected lag lengths. The top panel reports the results for the quarterly frequency. We find that if the objective is to fit the inflation impulse response a large lag length is preferred, for fitting the average across horizons we select $p^* = 7$ whereas for the long run $p^* = 15$ is selected. If the interest is in the impulse response of gdp growth we find $p^* = 1$ and $p^* = 7$, implying that for the average impulse response only one lag is sufficient, only if we care about the long run we select a larger lag length that ensures that the impulse response oscillates.

Table 1: CCV LAG LENGTH SELECTION — MONETARY

Quarterly	Average	Long run ($H = 20$)	AIC	BIC
π	7	15	7	1
y	1	7	7	1
i	1	11	7	1
$\pi + y + i$	7	7	7	1
			7	1
Monthly	Average	Long run ($H = 60$)		
π	37	31	60	2
y	1	33	60	2
i	25	38	60	2
$\pi + y + i$	12	28	60	2

Next, we use CCV to determine the set of endogenous variables to add to the baseline VAR. We have 10 candidate variables: the components of GDP (Durable consumption, Non-durable consumption, Investment), Components on inflation (Durables PCE Inflation, Non-durables PCE inflation), as well as popular monetary VAR variables: commodity price inflation, the BAA-AAA spread, changes in total reserves and the exchange rate. We iterate CCV over all possible combinations of max five variables. Similar as for lag length selection we report the model selections for different variables of interest ($\pi, y, i, \pi+y+i$) and different target horizons (average, long run). Table 5 displays the findings. First, the baseline VAR is never selected, and additional variables are always added for reducing the conditional out-of-sample forecast errors. Second, overall the number of lags that is selected with the additional variables is more stable around 7-9 lags across criteria. The only exception is the IR of GDP growth on average over horizons which continues to select $p = 1$. Third, in terms of the variables selected we note that durable and non-durable consumption are found important variables.

Table 2: CCV variable and lag length selection — Monetary

Quarterly, Average

	p	vars
π	8	Δnon-dur con, log S&P, ΔPCE-non-dur
y	1	Δnon-dur con, ΔI, π^{comm}, log S&P, ΔPCE-non-dur
i	8	ΔI, ΔPCE-dur, ΔPCE-non-dur
$\pi + y + i$	8	UR, Δdur con, Δnon-dur con, ΔPCE-dur

Quarterly, Long run ($H = 20$)

	p	vars
π	9	Δdur con, ΔI
y	7	Δdur con, Δnon-dur con, π^{comm}, Δ log S&P
i	8	Δnon-dur con, ΔI, ΔPCE-dur
$\pi + y + i$	7	UR, Δdur con, ΔI

References

Angrist, Joshua D., and Jörn-Steffen Pischke. 2015. *Mastering 'Metrics: The Path from Cause to Effect.* Princeton, NJ:Princeton University Press.

Athey, Susan, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 2025. "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely." *The Review of Economic Studies*, rdaf087.

Banbura, Marta, Domenico Giannone, and Lucrezia Reichlin. 2010. "Large Bayesian Vector Auto Regressions." *Journal of Applied Econometrics*, 25(1): 71–92.

Baumeister, Christiane. 2025. "Discussion of "Local Projections or VARs? A Primer for Macroeconomists"." *NBER Macroeconomics Annual.*

Bernanke, Ben S. 1986. "Alternative Explanations of the Money-Income Correlation." *Carnegie-Rochester Conference Series on Public Policy*, 25: 49–99.

Bernanke, Ben S, and Mark Gertler. 1995. "Inside the black box: the credit channel of monetary policy transmission." *Journal of Economic perspectives*, 9(4): 27–48.

Bernanke, Ben S, Jean Boivin, and Piotr Eliasz. 2005. "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach." *The Quarterly journal of economics*, 120(1): 387–422.

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

Brugnolini, Luca. 2018. "About local projection impulse response function reliability."

Bruns, Martin, and Helmut Lütkepohl. 2022. "Comparison of local projection estimators for proxy vector autoregressions." *Journal of Economic Dynamics and Control*, 134: 104277.

Canova, Fabio. 2007. *Methods for Applied Macroeconomic Research.* Princeton University Press.

Canova, Fabio, and Filippo Ferroni. 2022. "Mind the gap! Stylized dynamic facts and structural models." *American Economic Journal: Macroeconomics*, 14(4): 104–135.

Choi, Chi-Young, and Alexander Chudik. 2019. "Estimating impulse response functions when the shock series is observed." *Economics Letters*, 180: 71–75.

Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans. 1999. "Monetary Policy Shocks: What Have We Learned and to What End?" In *Handbook of Macroeconomics.* Vol. 1A, , ed. John B. Taylor and Michael Woodford, 65–148. Elsevier.

Doan, Thomas, Robert Litterman, and Christopher Sims. 1984. "Forecasting and conditional projection using realistic prior distributions." *Econometric Reviews*, 3(1): 1–100.

Dufour, Jean-Marie, and Eric Renault. 1998. "Short Run and Long Run Causality in Time Series: Theory." *Econometrica*, 66(5): 1099–1125.

Dufour, Jean-Marie, Denis Pelletier, and Eric Renault. 2006. "Short run and long run causality in time series: inference." *Journal of Econometrics*, 132(2): 337–362.

Elliot, Graham, and Allan Timmermann. 2016. *Economic Forecasting.* Princeton University Press.

Fair, Ray C. 1974. *A Model of Macroeconomic Activity.* Cambridge, MA:Ballinger.

Fernández-Villaverde, Jesús, Juan F Rubio-Ramírez, Thomas J Sargent, and Mark W Watson. 2007. "ABCs (and Ds) of understanding VARs." *American economic review*, 97(3): 1021–1026.

Forni, Mario, and Luca Gambetti. 2014. "Sufficient information in structural VARs." *Journal of Monetary Economics*, 66: 124–136.

Gertler, Mark, and Peter Karadi. 2015. "Monetary Policy Surprises, Credit Costs, and Economic Activity." *AEJ: Macroeconomics*, 7: 44–76.

Gilchrist, Simon, and Egon Zakrajšek. 2012. "Credit Spreads and Business Cycle Fluctuations." *American Economic Review*, 102(4): 1692–1720.

González-Casasús, Oriol, and Frank Schorfheide. 2025*a*. "Misspecification-Robust Shrinkage and Selection for VAR Forecasts and IRFs." National Bureau of Economic Research Working Paper 33474.

González-Casasús, Oriol, and Frank Schorfheide. 2025*b*. "Misspecification-robust shrinkage and selection for var forecasts and irfs." National Bureau of Economic Research.

Hamilton, James Douglas. 1994. *Time series analysis.* Princeton Univ. Press.

Hannan, E. J. 1970. *Multiple Time Series.* John Wiley & Sons.

Hsiao, Cheng. 1979. "Autoregressive Modeling of Canadian Money and Income Data." *Journal of the American Statistical Association*, 74(367): 553–560.

Inoue, Atsushi, and Lutz Kilian. 2002. "Bootstrapping Autoregressive Processes with Possible Unit Roots." *Econometrica*, 70(1): 377–391.

Inoue, Atsushi, and Lutz Kilian. 2006. "On the selection of forecasting models." *Journal of Econometrics*, 130(2): 273–306.

Ivanov, Ventzislav, and Lutz Kilian. 2005. "A Practitioner's Guide to Lag Order Selection For VAR Impulse Response Analysis." *Studies in Nonlinear Dynamics & Econometrics*, 9(1): 1–36.

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

Kilian, Lutz, and Helmut Lütkepohl. 2017. *Structural Vector Autoregressive Analysis.* Cambridge University Press.

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

Kuersteiner, Guido M. 2005. "Automatic Inference for Infinite Order Vector Autoregressions." *Econometric Theory*, 21(1): 85–115.

Lewis, Richard A., and Gregory C. Reinsel. 1985. "Prediction of Multivariate Time Series by Autoregressive Model Fitting." *Journal of Multivariate Analysis*, 16(3): 393–411.

Li, Dake, Mikkel Plagborg-Møller, and Christian K Wolf. 2024. "Local projections vs. vars: Lessons from thousands of dgps." *Journal of Econometrics*, 244(2): 105722.

Lippi, Marco, and Lucrezia Reichlin. 1993. "The dynamic effects of aggregate demand and supply disturbances: Comment." *The American Economic Review*, 83(3): 644–652.

Litterman, Robert B. 1986. "Forecasting with Bayesian Vector Autoregressions: Five Years of Experience." *Journal of Business Economic Statistics*, 4(1): 25–38.

Lohmeyer, Jan, Franz Palm, Hanno Reuvers, and Jean-Pierre Urbain. 2019. "Focused information criterion for locally misspecified vector autoregressive models." *Econometric Reviews*, 38(7): 763–792.

Ludwig, Julian F. 2024. "Local Projections Are VAR Predictions of Increasing Order." *Available at SSRN*.

Lütkepohl, Helmut. 2005. *New introduction to multiple time series analysis.* Springer.

Lütkepohl, Helmut. 2013. "Vector autoregressive models." In *Handbook of research methods and applications in empirical macroeconomics*. 139–164. Edward Elgar Publishing.

Lütkepohl, Helmut, and D. S. Poskitt. 1991. "Estimating Orthogonal Impulse Responses via Vector Autoregressive Models." *Econometric Theory*, 7(4): 487–496.

Mahedy, Tim, and Adam Hale Shapiro. 2017. "What's Down with Inflation?" *FRBSF Economic Letter*, , (2017-35).

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

Müller, Ulrich K., and James H. Stock. 2011. "Forecasts in a Slightly Misspecified Finite Order VAR." working paper.

Olea, José Luis Montiel, Mikkel Plagborg-Møller, Eric Qian, and Christian K Wolf. 2024. "Double Robustness of Local Projections and Some Unpleasant VARithmetic." National Bureau of Economic Research.

Olea, José Luis Montiel, Mikkel Plagborg-Møller, Eric Qian, and Christian K. Wolf. 2026. "Double Robustness of Local Projections and Some Unpleasant VARithmetic." *Econometrica*. forthcoming.

Olea, José Luis Montiel, Mikkel Plagborg-Moller, Eric Qian, and Christian) Wolf. 2025. "Local Projections or VARs? A Primer for Macroeconomists." Working Paper.

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

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

Poskitt, D. S. 1992. "Identification of Echelon Canonical Forms for Vector Linear Processes Using Least Squares." *The Annals of Statistics*, 20(1): 195–215.

Racine, Jeff. 2000. "Consistent cross-validatory model-selection for dependent data: hv-block cross-validation." *Journal of Econometrics*, 99(1): 39–61.

Ramey, Valerie. 2016. "Macroeconomic Shocks and Their Propagation." In *Handbook of Macroeconomics*. , ed. J. B. Taylor and H. Uhlig. Amsterdam, North Holland:Elsevier.

Romer, Christina D, and David H Romer. 1989. "Does monetary policy matter? A new test in the spirit of Friedman and Schwartz." *NBER macroeconomics annual*, 4: 121–170.

Romer, Christina D., and David H. Romer. 2004. "A New Measure of Monetary Shocks: Derivation and Implications." *American Economic Review*, 94: 1055–1084.

Schorfheide, Frank. 2005*a*. "VAR Forecasting under Misspecification." *Journal of Econometrics*, 128(1): 99–136.

Schorfheide, Frank. 2005*b*. "VAR forecasting under misspecification." *Journal of Econometrics*, 128(1): 99–136.

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

Sims, Christopher A. 1986. "Are forecasting models usable for policy analysis?" *Quarterly Review, Federal Reserve Bank of Minneapolis*, 10: 2–16.

Stock, James H., and Mark W. Watson. 2018. "Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments." *The Economic Journal*, 128(610): 917–948.

Stock, James H, and M Watson. 2008. "What is new in econometrics: Time series, lecture 7." *Short Course Lectures: NBER Summer Institute. NBER. https://www. nber. org/minicourse_2008. html*.

Uhlig, Harald. 2005. "What Are the Effects of Monetary Policy on Output? Results from an Agnostic Identification Procedure." *Journal of Monetary Economics*, 52(2): 381–419.

Appendix A: proofs

Conditional distribution of forecast errors

To get a glimpse at the "raw data" underlying an IR estimation, it is helpful to report the distribution of realized forecast errors following a unit monetary shock.

To construct the forecasts, we first project $Y_t = (y_t, \pi_t, i_t)$ on the time t information set captured by a set of controls W_t (for instance lags of Y_t) to obtain

$$Y_{t+h}^{\perp} = Y_{t+h} - \text{Proj}(Y_{t+h}|W_t).$$

$Y_t^{\perp}$ can be seen as the forecast error for Y_t given the time t information set captured by W_t. The forecast is then the predicted value $\hat{Y}_{t+h}^{\perp}$.

Next, we can project $Y_{t+h}^{\perp}$ on ε_t

$$Y_{t+h}^{\perp} = \beta_h \varepsilon_t + \eta_{t+h}$$

to obtain the effect of monetary shocks on forecasts errors at horizons h. This regression (effectively a local projection) allows to re-normalize the forecast errors by the size of the monetary shock. The "raw IR data" are then the forecast errors at horizon h conditional on experiencing a unit monetary shock:

$$e_{t+h} = \hat{\beta}_h + \eta_{t+h}.$$

The series e_{t+h} allows to compute the empirical distribution of forecast errors at horizon h following a unit monetary shock. By construction, the mean of that distribution is the LP point estimate.

Intuitively, this exercise is similar in spirit to Romer and Romer (1989) but where we re-scaled the forecast revision by the size of the shock at t. In Romer and Romer (1989), all the shocks had unit size, so no normalization was necessary.

Frequency representation of VAR-based IRs

The VAR(p) can be re-written in companion form as a VAR(1)

$$\mathbf{Y}_t = \mathbf{F}\mathbf{Y}_{t-1} + \mathbf{v}_t$$

where

$$\mathbf{Y}_t = \begin{bmatrix} Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-p+1} \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} \phi_1 & \cdots & & \cdots & \phi_{p-1} & \phi_p \\ \mathbf{I}_m & \mathbf{0} & & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \mathbf{I}_m & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_m & \end{bmatrix}, \quad \mathbf{v}_t = \begin{bmatrix} \epsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Note that the matrix $\mathbf{F}$ is of dimension $mp \times mp$.

The impulse responses of this system are given by the response of Y_t to an impulse ϵ_t.

To obtain the impulse response (i.e, solve the homogeneous difference equation), we diagonalize $\mathbf{F}$ (for exposition purposes, we posit distinct eigenvalues inside the unit circle,[20] so that there are mp distinct eigenvalues) to write $\mathbf{F} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$ with $\mathbf{P}$ a diagonal matrix of eigenvalues and $\mathbf{P}$ the matrix of corresponding eigenvectors so we can write

$$\mathbf{Y}_h = \mathbf{P}\mathbf{D}^h\mathbf{P}^{-1}\mathbf{Y}_0 \tag{18}$$

where we switched to horizon time subscript h to highlight the focus on response functions to a unit impulse at date $t = 0$ with $\mathbf{Y}_0 = B(\epsilon_0, 0, ..., 0)'$.

From (18), we can see that the impulse response function $\mathcal{R}_{is}(\cdot)$ —the impulse response of y_{it} to a unit shock to the sth element of ϵ_0 : $\mathcal{R}_{is}(h) = \frac{\partial y_{i,t+h}}{\partial \epsilon_{s,t}}$— is given by

$$\mathcal{R}_{is}(h) = \sum_{b=1}^{mp} \lambda_b^h c_b$$

with λ_b the eigenvalues of $\mathbf{F}$ and c_b are constant terms that depend on the initial conditions, i.e., $\mathbf{Y}_0$. In fact, we have

$$c_b = v_{ib} v_{bs}^{-1}$$

where v_{ab} are the entries of $\mathbf{P}$ and v_{ab}^{-1} are the entries of $\mathbf{P}^{-1}$ and s is the index of the shock of interest, so that $s \in \{1, .., m\}$.

Writing the eigenvalues in polar coordinates $\lambda_j = R_b e^{i\mathcal{R}_b}$ where $R_b = |\lambda_b|$ and $\mathcal{R}_b = \arccos\left(Re(\lambda_b)/R_b\right)$ with the convention that $\mathcal{R}_j = 0$ when λ_j is a real variable, we get

$$\mathcal{R}_{is}(h) = \sum_{b=1}^{mp} v_{ib} v_{bs}^{-1} R_b^h e^{i\mathcal{R}_b h}.$$

Further since complex eigenvalues always come as pairs of complex conjugates, we can combine them to write

$$\mathcal{R}_{is}(h) = \sum_{b=1}^{mp} v_{ib} v_{bs}^{-1} R_b^h 2\left(\cos(\mathcal{R}_b h) - \sin(\mathcal{R}_b h)\right).$$

or in more compact form

$$\mathcal{R}_{is}(h) = \sum_{b=1}^{B} 2R_b^h \left(\cos(\frac{2\pi}{T_b} h + \varphi_b)\right) \tag{19}$$

with T_b the period of the cosine and $\varphi_b = \arctan(\frac{\text{Im}(v_{ib} v_{bs}^{-1})}{\text{Re}(v_{ib} v_{bs}^{-1})})$ its phase.[21]

Expression (19) makes clear that the impulse responses of a VAR(p) are given by a sum of B basis functions, which are either decaying exponentials or decaying cosines. A VAR(p) with m variables has at most $B \leq pm/2$ cosine functions. Or stated differently, damped

[20] It's easy to generalize to the case with non-distinct roots.

[21] To obtain (19), use the trigonometric identity $\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)$ with $\alpha = \mathcal{R}_b h$ and $\beta = \arccos(\text{Re}(v_{ib} v_{bs}^{-1}))$ (or equivalently $\beta = \arcsin(\text{Im}(v_{ib} v_{bs}^{-1}))$. For real eigenvalues, we use the convention $\mathcal{R}_b = 0$ and $\varphi_b = 0$.

cosines are the *basis functions* used by VARs to capture IRFs.

We can thus see a VAR(p) variables as a dimension-reduction technique whereby we impose that the IRFs (the coefficient of the structural VMA representation) can be written as a sum of damped cosines or exponentials.

Importantly, a VAR(p) with m variables capture all the impulse responses with $\frac{pm}{2}$ basis functions given by damped cosines of different frequencies. The parameters p and m are thus key inputs that control the flexibility of the VAR approximation. As $p \to \infty$, the number of basis functions is infinite and the VAR(∞) can capture any impulse response function.

Appendix B: traditional lag length selection methods

Sequential testing

To find out whether including less (or more) lags is needed we may conduct hypothesis tests. Indeed comparing the VAR(p) model against the VAR($p+1$) amounts to testing whether $H_0 : \mathbf{A}_{p+1} = \mathbf{0}$, if this is the case the VAR($p+1$) becomes equivalent to the VAR(p) and we may be satisfied with using the VAR(p) model. A commonly used test statistic is the likelihood ratio test

$$\text{LR}(p) = n[\log(\det(\widehat{\mathbf{\Sigma}}_u(p))) - \log(\det(\widehat{\mathbf{\Sigma}}_u(p+1)))] \ , \tag{20}$$

where $\widehat{\mathbf{\Sigma}}_u(p) = \frac{1}{n}\sum_{t=1}^{n} \hat{\mathbf{u}}_t \hat{\mathbf{u}}_t'$ with $\hat{\mathbf{u}}_t$ the OLS residual from fitting the VAR(p). When the sample size tends to infinite under H_0 this test statistic converges to a random variable that follows a $\chi^2(n^2)$ distribution. We may compare LR(p) to the critical values of this distribution to determine whether the null should be rejected.

Different sequential procedures can be used to determine the lag length using the LR test. For instance, top-down sequential testing starts from some large $p^{\max}$ and sequentially computes LR($p^{\max}-1$), LR($p^{\max}-2$), and so on. When the test statistic starts exceeding the critical value at p, the null is rejected and the lag length cannot be further reduced making $p+1$ the chosen lag length.

Testing serial correlation

The most direct way to ensure that the CLS assumption holds is to test whether the reduced form errors $\mathbf{u}_t$ have serial correlation. A popular test is the Portmanteau test for $H_0 : \mathbb{E}(\mathbf{u}_t \mathbf{u}_{t-j}) = \mathbf{0}$ for $j = 1, \ldots, c$, where c is chosen by the researcher. The Portmanteau test statistic is given by

$$Q_c = n \sum_{j=1}^{c} \text{Tr}(\widehat{\mathbf{\Gamma}}_j' \widehat{\mathbf{\Gamma}}_0^{-1} \widehat{\mathbf{\Gamma}}_j \widehat{\mathbf{\Gamma}}_0^{-1}) \ , \tag{21}$$

where $\widehat{\mathbf{\Gamma}}_j' = \frac{1}{n}\sum_{t=j+1}^{n} \hat{\mathbf{u}}_t \hat{\mathbf{u}}_{t-j}'$. Under regularity conditions when n is large the distribution of Q_c under H_0 can be approximated by a $\chi^2(n^2(c-p))$ random variable.

To find the lag length of the VAR based on Q_c we start with a VAR with a small number of lags, say $p = 0$, test for serial correlation, and if the test is rejected (there is evidence for serial correlation), we increase the lag length and continue.

Information criteria

Arguably the most popular lag length selection procedure is to use an information criteria like AIC or BIC/SIC. In general, an information criteria takes the following form

$$\mathrm{C}(p) = \log(\det(\widehat{\mathbf{\Sigma}}_u(p))) + c_n f(p) \ , \tag{22}$$

where $\widehat{\mathbf{\Sigma}}_u(p) = \frac{1}{n}\sum_{t=1}^{n} \hat{\mathbf{u}}_t \hat{\mathbf{u}}_t'$ captures the outer sum of squared residuals, c_n is a rate parameter that depends on the sample size n and $f(p)$ captures the number of parameters in the VAR(p) model. In our set-up $f(p) = n^2 p$, but when a constant is included we have $n + n^2 p$.

As the lag order p increases, the VAR in-sample fit improves, and the first term —the one-step ahead forecast errors— gets smaller, but the second term —a function of the number of VAR parameters ($f(p)$)— increases. This creates a trade-off between model fit (one-step ahead forecasting performances) and model complexity. The idea is to select the lag length that minimizes $\mathrm{C}(p)$ for $p = 0, \dots p^{\max}$. Different choices for the rate parameter are consider:

AIC	$c_n = 2/n$
SIC/BIC	$c_n = \log n/n$
HQC	$c_n = 2\log\log n/n$

Under regularity conditions the SIC/BIC selects the true lag order when the sample increases if the true data generating process admits a VAR representation with finite lag length. (should the data be described by a VAR(p^0), $p^0 < \infty$), and the AIC chooses the lag order that minimizes the estimated one-step ahead forecasting error (though not necessarily the correct lag-order, and $\hat{p}^{aic} > \hat{p}^{bic}$). A large scale simulation study by Ivanov and Kilian (2005) find that in finite sample the AIC criteria is often preferable.

Lag length selection and mis-specification

More generally, most available lag selection methods have one common feature: they are designed for recovering the correct VAR model as opposed to recovering the IR of interest. Clearly, having the correct VAR implies getting the correct IR, but the reverse may not be true. As a result, when the VAR(p) is mis-specified, such methods may not select the "best" lag order to estimate an impulse response of interest; "best" in the sense of generating the most accurate IR estimate (e.g., lowest mean-squared error) among competing VAR lag orders.

In fact, even modest deviations from a VAR(p), i.e. deviations that are hard to detect using a statistical test, can fool traditional lag length selection methods that aim to find the true VAR. We illustrate this point with a simple example:

$$\begin{cases} x_t = \rho x_{t-1} + \varepsilon_t^x \\ y_t = \beta x_{t-8} + \varepsilon_t^y \end{cases}$$

It is clear that the true lag length $p^* = 8$ correctly captures the underlying DGP and the IR of interest. But what lag length will BIC select? For T large, we have

$$\mathrm{BIC}(1) \simeq \sigma_x^2\sigma_x^2\left(1 + \frac{\beta^2}{1-\rho^2}\frac{\sigma_x^2}{\sigma_y^2}\right) + 2^2\frac{\ln T}{T} \qquad \text{and} \qquad \mathrm{BIC}(8) \simeq \sigma_x^2\sigma_x^2 + 2^2 p^0 \frac{\ln T}{T}$$

thus

$$\text{BIC}(1) - \text{BIC}(8) \simeq \frac{\beta^2}{1-\rho^2}\frac{\sigma_x^2}{\sigma_y^2} - 2^2(p^0-1)\frac{\ln T}{T} \ .$$

Because the second term is negative, BIC may select a model with only one lag if the first term is small enough. In particular, if the shock of interest ε_t^x accounts for a relatively small share of the variance —$\frac{\sigma_x}{\sigma_y}$ is of order $O(1/\sqrt{T})$ —, we can have $\text{BIC}(1) - \text{BIC}(8) < 0$ and BIC will select only one lag. This form of local mis-specification is problematic when realizing that the macro shocks of interest (like monetary and fiscal shocks) often explain a small share of the variance in macro variables. Alternatively, if the VAR lag coefficient β is small (so that $\beta = O(1/\sqrt{T})$), we can also have that $\text{BIC}(1) - \text{BIC}(8) < 0$ and BIC will again conclude that a model with one lag is optimal.

The key feature of this simple example is that whenever the variance of the shock of interest is small, whenever important lags are small, conventional lag length selection methods may select the wrong lag length. In the econometrics literature models with these types of small coefficients are typically referred to as locally mis-specified and there exists a modern literature that aims to derive alternative information criteria that directly target the impulse response of interest to select the VAR (Lohmeyer et al., 2019; González-Casasús and Schorfheide, 2025*b*).

Appendix C: VAR vs parametrized LP

A VAR reduce the dimension of the estimation problem by imposing dynamic restrictions across variables, a VAR(p) with n variables can parametrize the Hn impulse response coefficients with pn^2 parameters (pn parameters per equation and there are n equations). Provided that $np << H$, the VAR can deliver large efficiency gains. With quarterly data, a 3-variable VAR(4) only estimates $pn^2 = 36$ parameters instead of $Hn = 90$ parameters to compute the impulse responses up until horizon $H = 30$ quarters.[22]

That said, the best way to understand the VAR-as-PT idea is through the IR parametrization idea of section **??**. Specifically, recall that a VAR imposes that the IRs to *all* shocks follow the *same* difference equation.

A related useful perspective on the VAR-as-PT vs LP, is to ask the question as of why a VAR is a more efficiency power tool than a parametrized LP, where each IR is written as a sum of damped cosines, exactly like in a VAR.

A useful way to understand the dimensionality reduction delivered by a VAR is to compare it with horizon-by-horizon local projections, including smooth or parametrized LPs. Consider the VAR

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t.$$

Once the matrices $A_1, \ldots, A_p$ are estimated, they completely determine the propagation of

[22] A VAR can provide even larger efficiency gains for shocks that account for a small share of the total variance, as is often the case in macro —monetary or fiscal shocks typically explain a small share of the total variance in macro variables. Unlike LP a VAR will estimate the parameters using the full auto-covariance structure of the data, while LP will only exploit the variation coming from the shock of interest. So if the shock of interest explains α percent of the variance in the data, the *effective* parameter/sample-size ratio is only $\frac{\alpha Hn}{T}$ for local projection (while that ratio for the VAR remains $\frac{pn^2}{T}$). Thus, provided that $\frac{pn}{H\alpha} << 1$, the VAR can provide even larger power gains over LP.

shocks through the system. In particular, if $IR_{s,h}$ denotes the impulse response to shock s at horizon h, then the impulse responses satisfy the same difference equation as the data:

$$IR_{s,h} = A_1 IR_{s,h-1} + \cdots + A_p IR_{s,h-p}.$$

This is a strong restriction. The responses to any shock must be generated by the same autoregressive dynamics. Hence the VAR restricts impulse responses jointly across horizons, variables, and shocks.

By contrast, a parametrized LP imposes restrictions more directly on the shape of each impulse response. For example, one may approximate the response as a sum of damped cosine terms,

$$IR_{s,h} = \sum_b c_b \rho_b^h \cos\left(2\pi h/\tau_b + \varphi_b\right).$$

This parametrization imposes smoothness and persistence across horizons, and thereby reduces the number of horizon-specific coefficients relative to an unrestricted LP. However, the restriction is primarily a restriction across horizons for a given response. Unless further structure is imposed, the coefficients governing the shape of the response may vary freely across variables and shocks.

The VAR imposes a broader set of restrictions. Since all responses are generated by the same matrices $A_1, \ldots, A_p$, the dynamic behavior of one response is informative about the dynamic behavior of other responses. In the damped-cosine representation, this means that the frequencies, damping rates, and phases are not chosen independently response by response. They are tied together by the common roots of the VAR. Thus the VAR disciplines the full impulse-response system, rather than only smoothing individual responses across horizons.

The resulting efficiency gains come from two sources. First, restrictions across variables and horizons reduce the number of free parameters needed to describe the impulse-response system. In raw parameter-count terms, however, this reduction need not be very large. A VAR with n variables and p lags contains n^2p autoregressive parameters, while a parametrized LP designed to recover a comparable set of responses may involve on the order of $2n^2p$ parameters. Thus the pure parameter-count advantage is only moderate.

The more important gain is that the VAR's restrictions operate across shocks. Because the same dynamic system governs the responses to all shocks, the VAR can be estimated from the full unconditional second-moment structure of the data. The researcher does not need to isolate each shock-specific response using only moments conditional on that shock or instrument. Instead, the autoregressive coefficients are estimated from the unconditional dynamics of y_t, exploiting all comovements among the variables. Once these coefficients are known, the impulse responses to all shocks are recovered by iterating the estimated system.

This is the central sense in which the VAR achieves dimensionality reduction. It is not merely that the VAR has fewer parameters than a smooth or parametrized LP. Rather, the VAR imposes a common law of motion on the entire impulse-response system. Restrictions across variables and horizons reduce the number of coefficients, while restrictions across shocks allow the model to use the full unconditional variation in the data. This second channel is especially important: it is what gives the VAR an efficiency advantage relative to smooth LPs or parametrized LPs, which typically estimate shock-specific responses from a narrower set of conditional moments.

Equivalently, a parametrized LP regularizes the impulse responses by choosing a low-dimensional representation of each response curve. A VAR regularizes the whole dynamic system. The same transition matrices determine every horizon, every variable, and every shock. The effective dimensionality reduction therefore comes less from a dramatic fall in the number of estimated coefficients, and more from the fact that those coefficients are estimated using the full covariance structure of the multivariate process.

Appendix F: VAR mis-specification and IR bias, Example #2

Here, we will use a simple simulation to illustrate how VAR under-parametrization —having too few damped cosine basis functions— can bias the IR estimates, with the VAR "trying" to fit the IR of interest, but unable to with a single damped cosine —the dynamic effect being too protracted and building up late.

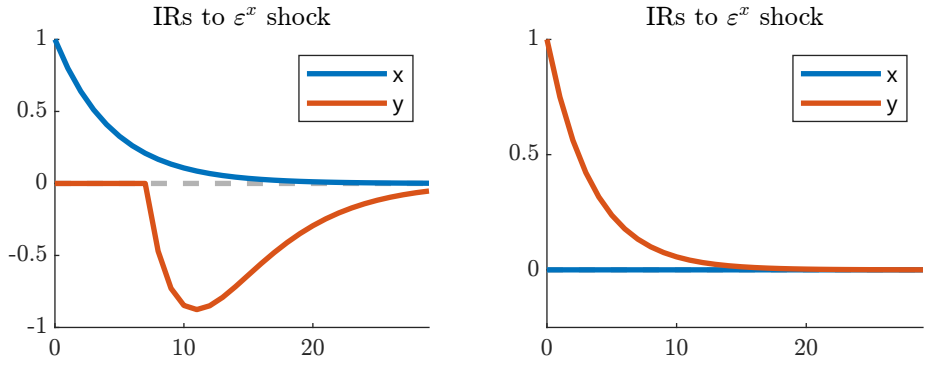
Specifically, We simulate a simple bivariate DGP for $\mathbf{y}_t = (y_t, x_t)'$ that captures some stylized features of the monetary impulse responses seen previously:

$$\begin{cases} x_t = \rho_y x_{t-1} + \varepsilon_t^x \\ y_t = \rho_y y_{t-1} + \beta x_{t-8} + \varepsilon_t^y \end{cases} \tag{23}$$

with $\rho_x < 1$, $\rho_y < 1$ and $-1 < \beta < 0$. We have $\sigma_{\varepsilon^x} = 1$ and $\sigma_{\varepsilon^y} = 1$.

The IR of interest is the impulse response of y_t to a shock to x_t. Figure 5 plots the corresponding IRs. This DGP is meant to mimic (in a loose way) the delayed response of inflation to a monetary shock. Importantly, the DGP can be represented as a VAR(p), so a VAR is well-specified, but it needs to have enough lags p. Here we must have $p \geq p^0 = 8$.

Figure 5: Simulated impulse responses

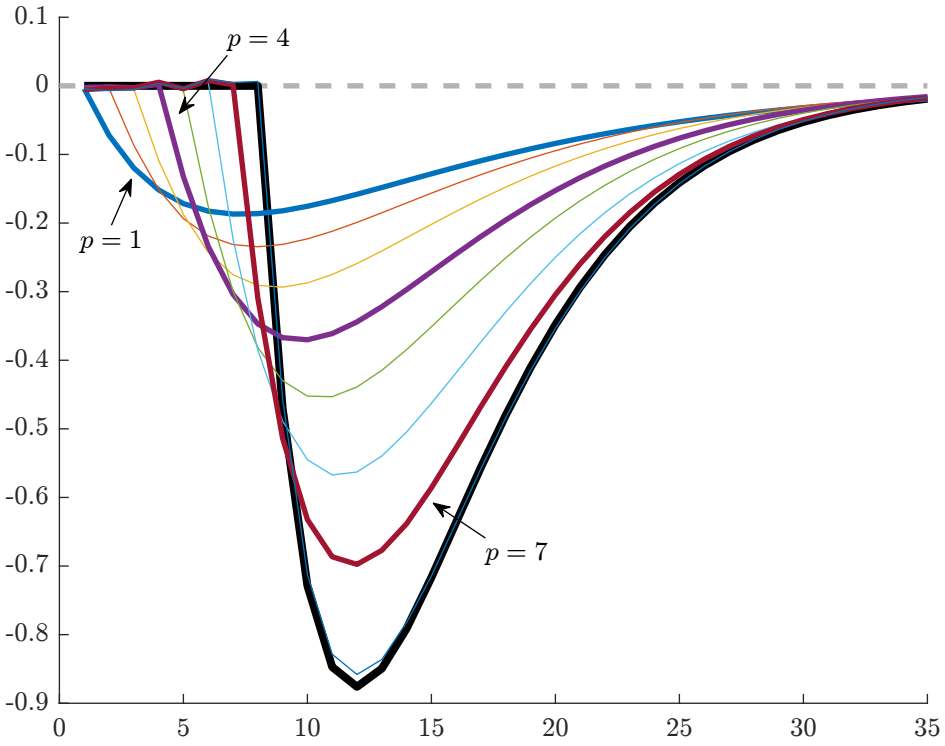


We will illustrate the effects of VAR *under*-parametrization in population. Similar to Figure 3, Figure 6 plots the VAR(p)-based IRs as we increase the lag order p.[23] We can see that a mis-specified VAR with $p < 8$ faces a tension: it is trying to capture the hump-shaped response of w_t but its dynamics are not rich enough to capture the protracted nature of the IR (no response for 7 periods). Thus for $p < 8$, the VAR systematically feature a too small

[23] We use a very large sample size to focus on population estimates.

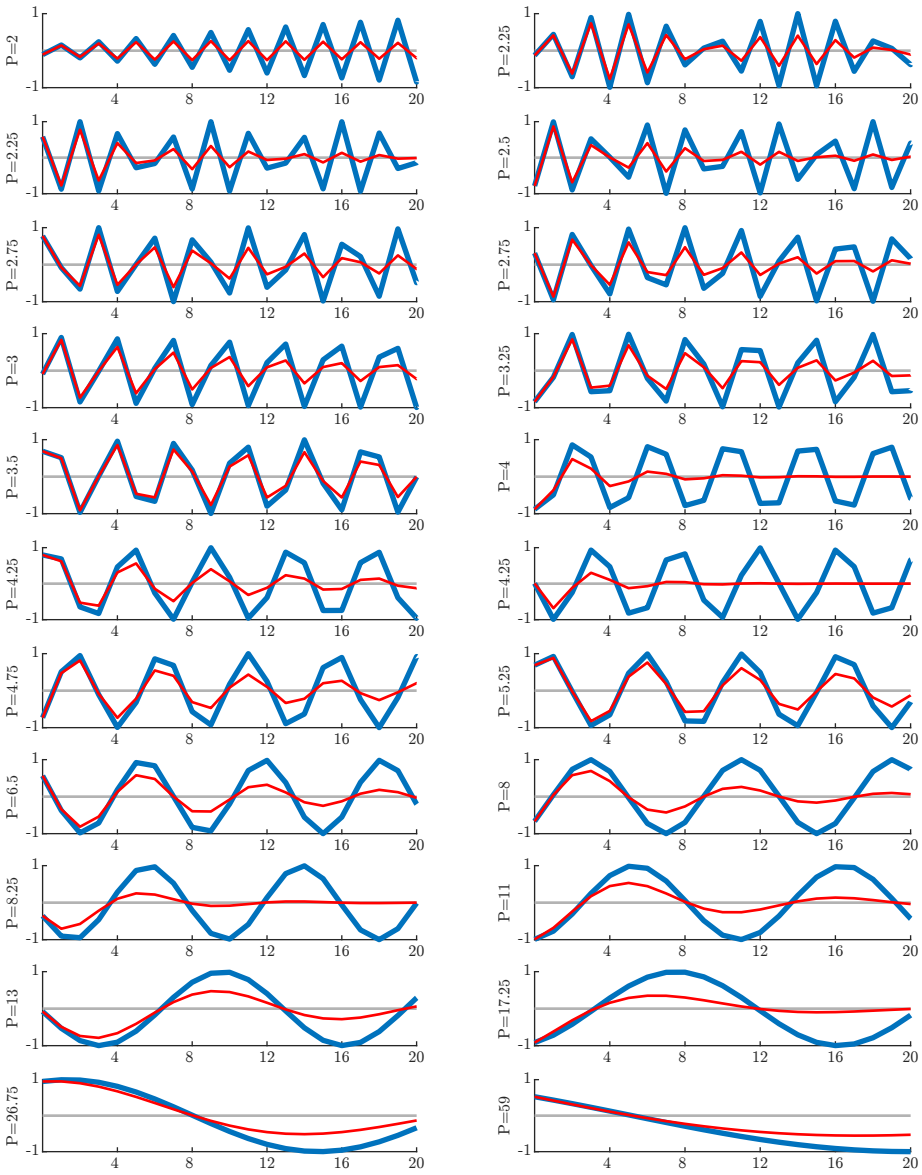
and too early decline in w. This is the same pattern that we observed for the IR of inflation in the monetary VAR (Figure 3, middle row)

Figure 6: VAR-BASED IRS



Notes: Top panel: Each IR represents the IR estimated from a VAR(p) of increasing order $p = 1, 2, ..., 8$.

Figure 7: VAR-BASED BASIS FUNCTIONS FOR $p = 16$



Notes: The red lines are the basis functions underlying the VAR representation of the impulse responses, organized from high frequency (low periodicity P) in top rows to low frequencies (high P). The blue lines report the cosines of different frequencies (without damping), and the red lines report the same cosines with the damping factor. The "jagged" aspect of the high frequency cosines comes from the discrete nature of the VAR, and the too low sampling rate.

Surrogacy in Time Series*

Régis Barnichon — FRB San Francisco
Geert Mesters — FRB New York
Emi Nakamura — UC Berkeley
Jón Steinsson — UC Berkeley

July 13, 2026

Abstract

This paper proposes a new approach to impulse response estimation based on the idea of *surrogacy*. We introduce an approximate surrogacy assumption under which, conditional on intermediate outcomes observed shortly after a shock, the original shock no longer contains important additional predictive information for future outcomes. We show that this condition naturally arises in important classes of time series models. Building on this insight, we develop two estimators — Surrogate LP and Surrogate VAR — that combine short-run local projections with long-run predictive propagation. The proposed estimators preserve the robustness advantages of local projections while substantially improving long-horizon efficiency. Simulation evidence shows large variance reductions relative to conventional local projections without inheriting the misspecification biases of VARs for realistic VARMA-like data generating processes.

*Preliminary and Incomplete. The views expressed in this paper are the sole responsibility of the authors and do not necessarily reflect the views of the Federal Reserve Bank of San Francisco, the Federal Reserve Bank of New York or the Federal Reserve System.

1 Introduction

In time series analysis, two broad approaches are available for estimating the dynamic transmission —or impulse response functions— of forecast innovations or structural shocks: (i) vector autoregressions (VARs), going back to Sims (1980), and (ii) local projections (LPs), as introduced by Jordà (2005).

Vector autoregressions (VARs) impose a fully specified recursive structure and thus exploit cross-equation restrictions to improve statistical efficiency. The downside, however, is that misspecification of the short-run dynamics can propagate throughout the entire impulse response function, leading to substantial biases. In contrast, LPs estimate impulse responses directly, horizon by horizon, and therefore impose no dynamic restrictions. This makes impulse response estimates more robust to dynamic misspecification, but also more prone to large variance (Jordà, 2005; Olea et al., 2026).

With the two methods lying on opposite sides of the bias–variance trade-off, researchers often face a difficult choice: recursive models are statistically efficient but fragile to misspecification, whereas direct estimators are robust but often imprecise; see Olea et al. (2025) for a detailed review of the literature.

In this paper, we propose a new class of impulse response estimators that offers a middle ground between these two extremes. The class exploits the idea of *surrogacy*, a concept originating in the medical statistics literature (Prentice, 1989) and also used in the dynamic treatment effects literature (e.g., Hernan and Robins, 2025, Part III).

The central idea underlying a surrogacy assumption is that there exist some short-run outcomes —the surrogates— that summarize *all* information relevant for the long-run effect of a treatment. In other words, once the response of the surrogates is observed, the original treatment contains no additional predictive information about the long-run outcome.

We argue that a closely related idea applies naturally to impulse response analysis in time series settings. While the short-run transmission of a shock may depend on complicated latent states, frictions, or omitted propagation channels, these mechanisms gradually become reflected in observable outcomes over time. Once a sufficiently informative block of intermediate outcomes has been observed, the original structural shock may no longer contain important additional predictive information for future outcomes. We refer to these intermediate outcomes as *surrogates*, and to the horizon after which this approximate conditional independence emerges as the *surrogacy horizon*. Formally, we introduce an approximate surrogacy condition for time series that can be interpreted as a form of approximate Granger non-causality: conditional on the surrogate variables, the original structural shock no longer contains substantial predictive information for future outcomes. We show that this condition naturally arises in many state space models.

The surrogacy condition allows long-horizon impulse responses to be decomposed into two components: (i) the causal effects of the shock on the surrogates and (ii) the predictive effects of the surrogates on the long-run outcomes. Thus, rather than estimating each long-horizon response directly from the shock —which is often only weakly related to distant outcomes— surrogacy allows the researcher to use short-run outcomes as a predictive bridge to longer horizons. Since that predictive bridge is estimated from the unconditional variation in the time series, rather than only from the variation directly induced by the shock (as in standard LPs), a surrogacy-based estimator can deliver more precise long-run estimates than LPs while avoiding the stronger recursive restrictions of a VAR and the associated biases. We show that the surrogacy assumption is weaker than the assumption of a correct VAR specification.

Concretely, we develop two new estimators for impulse responses. The first, *Surrogate LP*, combines the conventional local projection moment condition with a new surrogacy-based moment condition that combines short-run local projections with long-run direct forecasting equations estimated conditional on the surrogate variables. The resulting GMM estimator relies only on the surrogacy condition itself and does not impose additional recursive restrictions on the post-surrogacy dynamics. The second, *Surrogate VAR*, combines the conventional local projection moment condition with an additional VAR based surrogacy moment conditions. The surrogacy VAR conditions combine short-run local projections with autoregressive propagation imposed only after the surrogacy horizon. Relative to Surrogate LP, this estimator requires an additional assumption: after the surrogacy horizon, the relevant propagation dynamics must admit a VAR representation. Importantly, this assumption is imposed only on the post-surrogacy dynamics rather than globally from impact onward, and is therefore distinct from the surrogacy condition itself. Both estimators preserve the robustness advantages of local projections at short horizons where misspecification is most severe, while still exploiting additional dynamic structure to improve efficiency at longer horizons.

For the Surrogate LP estimator, we derive the asymptotic distribution and prove that under the approximate surrogacy condition the estimator asymptotically improves upon conventional local projections in efficiency. For the Surrogate VAR estimator, we derive moment conditions that identify the recursive propagation parameters using lagged instruments dated sufficiently far before the surrogacy horizon, allowing the post-surrogacy propagation law to be estimated by IV or GMM methods even when the short-run dynamics are misspecified.

To select the surrogate variables and the surrogacy horizon, we propose a data-driven procedure similar in spirit to an information criterion. We stress that a conservative selection rule is important in this setting because under-selecting the surrogate horizon is often substantially more costly than over-selecting it. If recursive propagation is imposed too

early, any remaining dynamic misspecification may subsequently propagate throughout the entire impulse response path. In contrast, selecting a surrogate horizon that is somewhat too large primarily reduces efficiency because more horizons are estimated directly. Our procedure is based directly on the defining implication of surrogacy: after conditioning on a valid surrogate state, the original shock should no longer improve forecasts of future outcomes. We therefore construct a horizon-wise prediction-ratio criterion that measures the remaining predictive content of the shock after conditioning on candidate surrogate variables, and combine it with an information-type penalty that deliberately favors larger surrogate states whenever the evidence against surrogacy remains non-negligible.

We study the finite-sample behavior of the proposed methods in simulation environments designed to capture the types of misspecification encountered in macroeconomic applications. We consider both stylized VARMA designs and empirically realistic dynamic factor model environments taken from Li, Plagborg-Møller and Wolf (2024). We find that the Surrogate LP and Surrogate VAR estimators achieve bias comparable to that of conventional LP estimators, while delivering substantially lower variance at longer horizons. These findings hold across all data-generating processes considered. Typically, the variance gains of the Surrogate VAR estimator are somewhat larger relative to those of Surrogate LP.

Our empirical application combines a quarterly dataset of macroeconomic and financial variables with externally identified structural shocks. We consider four outcome variables — real output growth, inflation, unemployment, and the short-term interest rate — and study their responses to monetary policy shocks, government spending shocks, oil-price shocks, and technology shocks. The candidate surrogate variables include the outcome variables themselves together with a broader set of indicators capturing financial conditions, inflation expectations, expenditure components, asset prices, and disaggregated inflation measures. For each shock–outcome pair, the empirical analysis identifies the surrogate state that summarizes the relevant transmission mechanism of the shock. We then compare impulse response estimates across the different empirical specifications and study how the selected surrogate variables vary across shocks and outcomes.

The remainder of this paper is organized as follows. We finish the introduction with a brief overview of some related literature. Section 2 provides an illustrative example that highlights how surrogates naturally arise in conventional macro models and how we may exploit them to obtain reliable and efficient impulse response estimates. Section 3 formalizes our assumption of approximate surrogacy for time series models and Section 4 discusses surrogate LP and VAR estimators that can be formulates based on this assumption. Section 5 discusses the selection of the surrogate variables. Section (6) evaluate the performance of the surrogate LP and VAR estimators. Finally, Section **??** documents surrogate variables for different macro shocks and outcomes.

Related literature

Our paper contributes to the large literature comparing VAR and local projection estimators for impulse response analysis. Since the seminal contribution of Sims (1980), VARs have become a central tool for empirical macroeconomics and structural shock identification. The classical perspective views the VAR as a reduced-form approximation to the data generating process obtained by truncating an underlying VAR(∞) representation implied by the Wold decomposition. At the same time, the literature has long recognized that finite-order VARs are generally misspecified approximations that may nevertheless improve estimation accuracy by trading bias for variance (e.g., Lewis and Reinsel, 1985; Lütkepohl and Poskitt, 1991; Schorfheide, 2005; Plagborg-Møller and Wolf, 2021; Olea et al., 2024; González-Casasús and Schorfheide, 2025*a*; Olea et al., 2025). These concerns have motivated direct estimators such as local projections (Jordà, 2005), which avoid recursively propagating misspecified dynamics. Recent work studies the resulting robustness–efficiency trade-off between LPs and VARs (e.g., Kilian and Kim, 2011; Brugnolini, 2018; Li, Plagborg-Møller and Wolf, 2024; Olea et al., 2025, 2024). Our paper contributes to this literature by introducing surrogate LP and surrogate VAR estimators that impose dynamic structure only after the surrogacy horizon, thereby combining the robustness advantages of LPs with the efficiency gains of recursive propagation.

Our framework also relates to the surrogacy and dynamic treatment effect literature. Early work by Prentice (1989) introduced the idea that treatment effects on long-run outcomes may be summarized through intermediate surrogate variables. Related ideas appear in the mediation literature (e.g., Pearl, 2001; Imai, Keele and Tingley, 2010); see Hernan and Robins (2025, Part III) for a textbook treatment. Recent work in econometrics and statistics has formalized surrogacy using conditional independence restrictions in the potential outcome framework Athey et al. (2025), studied efficiency gains from surrogate information (Chen and Ritzwoller, 2023; Kallus and Mao, 2025) and conducted sensitivity analysis under imperfect surrogacy (Fan et al., 2026). Our paper adapts these ideas to macroeconomic impulse response analysis. In our setting, the intermediate outcomes correspond to observable post-shock states that summarize the relevant propagation channels of the economy after the initial transmission phase. The surrogate LP and surrogate VAR therefore combine short-run causal estimation with long-run predictive propagation, linking the treatment-effect literature on surrogates to the macroeconometric literature on impulse response estimation.

Finally, our paper relates to the literature on Granger causality, invertibility, and state-space methods. The surrogacy condition can be interpreted as a form of approximate Granger non-causality (Granger, 1969; Sims, 1972), or even closer subspace granger non-causality (Al-Sadoon, 2014): conditional on the surrogate variables, the structural shock no longer contains substantial additional predictive information for future outcomes. Our framework is also

closely connected to state-space smoothing ideas (e.g. Durbin and Koopman, 2012). In many macroeconomic environments, latent propagation states are only imperfectly observed contemporaneously but become increasingly revealed through future observables. The surrogate variables therefore act as a finite-dimensional representation of the latent state relevant for long-run propagation. In this sense, our approach is also related to the non-fundamentalness and invertibility literature (e.g., Lippi and Reichlin, 1993; Fernández-Villaverde et al., 2007; Forni and Gambetti, 2014; Canova and Ferroni, 2022).

2 Surrogacy for time series: motivation

Before introducing surrogacy in a time-series setting, it is useful to begin with the cross-sectional version of the idea, as it appears in the medical literature (Prentice, 1989) and, more recently, in the treatment-effects literature (e.g. Athey et al., 2025). The basic problem is the following. A researcher wants to estimate the long-run effect of a treatment (ε) on an outcome (w). In an experimental dataset, treatment is randomly assigned, so the causal effect of (ε) is identified. However, the long-run outcome (w) may not be observed for the treatment sample, either because individuals are followed only for a limited time or because the relevant outcome takes many years to materialize. In those cases, to recover the long-run effect, one option is to use some observed short-run outcomes as bridge from the treatment to the long-run outcomes, exploiting the treatment dataset for the effect of the treatment on the short run outcomes and some other dataset, e.g. an administrative dataset, for the effect of the short run outcomes on the long run outcomes. The surrogacy assumption ensures the validity of such bridge construction.

The surrogacy assumption

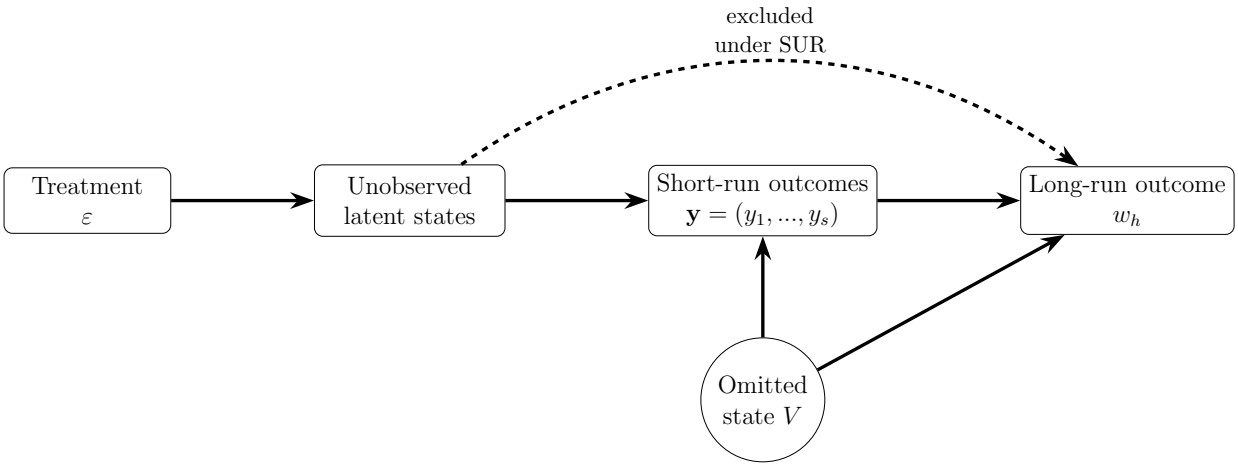
Surrogacy states that there exists a set of observed variables $\mathbf{y}$, which contain all the information that is relevant for predicting the long-run effect of treatment. Once the surrogates have been observed, treatment ε itself should contain no additional information about the outcome w. Formally, the surrogacy condition can be stated as

$$\text{Proj}(w \mid \mathbf{y}, \varepsilon) = \text{Proj}(w \mid \mathbf{y}) \ , \tag{1}$$

where $\text{Proj}(a \mid b)$ denotes the linear projection of a onto b. We note that typically the surrogacy condition is stated in terms of conditional expectations instead of linear projections, but since we work in a linear time series environment below we keep the projection notation throughout.

Although any variable could in principle act as a surrogate for an outcome, in the case

Figure 1: Surrogacy as post-treatment sufficiency



of estimating *long-run* effects, attractive surrogate candidates are *short-term* responses that are in the treatment dataset. To illustrate the idea, denote by w_h the outcome variable h periods after treatment, and assume that $\mathbf{y} = (y_1, ..., y_s) = \mathbf{y}_{1:s}$ can act as surrogates for w_h for $h > s$ where s is the surrogacy horizon. Then the surrogacy condition can be restated as

$$\text{Proj}(w_h \mid \mathbf{y}_{1:s}, \varepsilon) = \text{Proj}(w_h \mid \mathbf{y}_{1:s}) \ . \tag{2}$$

Expression (2) states that after one conditions on the observed short-run response of the outcome, treatment assignment no longer helps predict the outcome at longer horizons. To help develop some intuition, Figure 1 summarizes the 1 assumption with a DAG. Under surrogacy, treatment can only affect the long-run outcome only through the short-run surrogate vector $\mathbf{y}_{1:s}$. Thus, once $\mathbf{y}_{1:s}$ is observed, treatment assignment contains no additional information about the long-run outcome w_h.

Intuitively, surrogacy can be viewed as a weaker "post-treatment" version of identification through controls for long horizons. Rather than requiring the researcher to control for *all* factors that jointly affect the outcome variables and the treatment variable —an often elusive objective in the presence of unobserved heterogeneity— surrogacy requires controlling for a set of short-run post-treatment outcomes that are sufficient for the long-run response. The idea is that short-run responses to treatment may contain enough information to absorb the effects of omitted variables —the omitted states— on the final outcome. In other words, the hope underlying the surrogacy assumption is that the long-run consequences of treatment are mediated by a relatively low-dimensional vector of state variables, and that the observed short-run responses are rich enough to reveal the outcome-relevant component of that state.

In that sense, surrogacy shifts the burden from modeling all causal pathways to validating an informational condition: once the surrogates are conditioned on, treatment should have no remaining predictive content for the final outcome.

From causal to predictive for long horizon causal effects

The key implication of the surrogacy condition (2) is that it provides a predictive bridge from short-run responses to long-run outcomes. Specifically, denoting by θ_h a causal effect of ε on w_h, 2 allows to decompose the long-run effect of treatment into a product of two terms: (i) the short run causal effects of treatment on the surrogates —the effect of ε on $\mathbf{y}_{t:t+s}$—, and (ii) a prediction of the surrogates on the long-run outcome —the effect of $\mathbf{y}_{t:t+s}$ on w_h—, which can hopefully be estimated from a different dataset. Specifically,

$$\theta_h = \boldsymbol{\beta}_{hs}\boldsymbol{\tau}_{0:s} \ ,$$

where $\boldsymbol{\beta}_{hs}$ is the coefficient vector from the projection $\text{Proj}(w_h \mid \mathbf{y}_{0:s})$ and $\boldsymbol{\tau}_{0:s} = \text{Proj}(\mathbf{y}_{0:s} \mid \varepsilon = 1) - \text{Proj}(\mathbf{y}_{0:s} \mid \varepsilon = 0)$.[1]

In other words, under SUR, a predictive model is sufficient to estimate long-run causal effects based on propagating short-run causal effects. Importantly, the researcher is not interpreting the regression of w_h on $\mathbf{y}_{0:s}$ as a structural causal relationship. Instead, the regression is used as a predictive bridge from the observed short-run response to the long-run outcome. Surrogacy is precisely the assumption that this bridge is valid for the causal effect of treatment: treatment affects the long-run outcome only through information already summarized by the surrogate vector $\mathbf{y}_{0:s}$. In practice, the treatment dataset can be used to estimate the short run causal effects whereas a second, e.g. administrative, dataset can be used to estimate the predictive long run coefficients.

Example: Blood markers and cancer survival

As illustration,[2] consider a clinical trial evaluating the effect of a cancer treatment on five-year survival. Since survival takes years to observe, researchers often consider earlier clinical variables, such as tumor response, progression-free survival, or a blood marker, as surrogate endpoints.

Surrogacy says that two patients with the same value of the surrogates should have the same expected five-year survival, regardless of whether one received the treatment and the

[1] Under (2), $\text{Proj}(w_h \mid \mathbf{y}_{0:s}, \varepsilon) = \text{Proj}(w_h \mid \mathbf{y}_{0:s}) = \boldsymbol{\beta}_{hs}\mathbf{y}_{0:s}$. Using $\theta_h = \text{Proj}(w_h \mid \varepsilon = 1) - \text{Proj}(w_h \mid \varepsilon = 0)$ and $\text{Proj}(w_h \mid \varepsilon) = \text{Proj}(\text{Proj}(w_h \mid \mathbf{y}_{0:s}, \varepsilon) \mid \varepsilon)$, we obtain $\theta_h = \boldsymbol{\beta}_{hs}\left(\text{Proj}(\mathbf{y}_{0:s} \mid \varepsilon = 1) - \text{Proj}(\mathbf{y}_{0:s} \mid \varepsilon = 0)\right) = \boldsymbol{\beta}_{hs}\boldsymbol{\theta}_{0:s}$.

[2] See the Appendix for additional examples.

other did not. If this condition holds, then treatment affects five-year survival *only* by changing the short-run distribution of the marker. In that case, it is possible to estimate long-run survival from the short-run responses of the marker.

Importantly, the marker-survival relationship need not itself be causal. The marker may be predictive because it reflects tumor burden, immune response, frailty, genetics, or other latent biological states. What matters is that, conditional on the marker, treatment carries no additional information about survival. The prediction model is valid because the marker *summarizes* the treatment-relevant information about the long-run outcome.[3]

Surrogacy in time series: an ARMA(1,1) example

We can now turn to the main topic of our paper: surrogacy in a time series context. Unlike in experimental settings, there is typically only one dataset and long-run outcomes are observed directly. However, statistical inference becomes increasingly difficult at longer horizons because the effective sample size shrinks with the horizon and the explanatory power of shocks often becomes weak. In the limit, when the horizon approaches the sample size, essentially no long-run variation remains for inference. A conventional approach to improving efficiency is to impose parametric restrictions on the shape of the impulse response function, for instance through a VAR model. Such recursive approaches can substantially reduce variance, but they also risk introducing misspecification bias and coverage distortions when the underlying dynamics are not well approximated by a finite-order VAR (Olea et al., 2024).

To illustrate how surrogacy can address this trade-off, consider an econometrician interested in estimating the impulse response of a variable w_{t+h} to an exogenous iid shock ε_t, namely

$$\theta_h = \text{Proj}(w_{t+h} \mid \varepsilon_t = 1) - \text{Proj}(w_{t+h} \mid \varepsilon_t = 0).$$

We use a simple ARMA(1,1) model to motivate the main ideas:

$$w_t = \rho w_{t-1} + \varepsilon_t + b\varepsilon_{t-1}, \qquad |\rho| < 1, \qquad \varepsilon_t \overset{iid}{\sim} (0, \sigma^2). \tag{3}$$

The impulse response function is

$$\theta_0 = 1, \qquad \theta_h = (\rho + b)\rho^{h-1}, \qquad h > 0.$$

[3]Conversely, if the set of surrogates does not contain *all* short-run responses relevant to predict long-run effects of treatments, then SUR will fail. For instance, this would be the case if the cancer treatment also increases toxicity or resistance, which independently affect long-run outcomes.

Exact surrogacy as a benchmark. Before discussing the general ARMA(1,1) case, it is useful to consider the special case $b = 0$, which reduces (3) to an AR(1). In that benchmark, any outcome w_{t+s} for $s = 0, \ldots, h$ is itself a sufficient state variable for the future evolution of the process. As a result, surrogacy holds exactly for any surrogate horizon $s \geq 0$. For instance, for $s = 0$ we have

$$\text{Proj}(w_{t+h} \mid w_t, \varepsilon_t) = \text{Proj}(w_{t+h} \mid w_t).$$

The impulse response for $h > s = 0$ can therefore be decomposed exactly as

$$\theta_h = \beta_{h0}\theta_0, \qquad \beta_{h0} = \text{Proj}(w_{t+h} \mid w_t) = \rho^h.$$

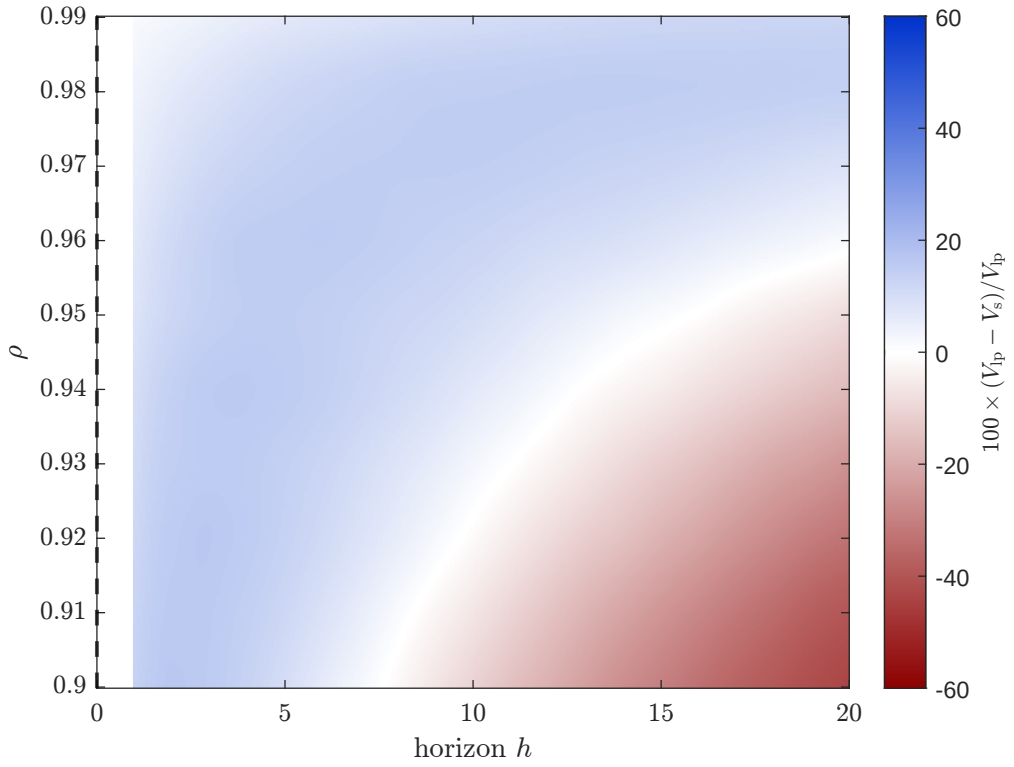
Thus, once the contemporaneous response w_t is observed, the shock itself contains no additional predictive information for future outcomes. Long-horizon impulse responses can therefore be recovered by combining the short-run causal effect θ_0 with a predictive propagation coefficient estimated from the unconditional dynamics of the time series. This simple benchmark illustrates the central idea underlying surrogacy: long-run causal effects can sometimes be recovered from predictive propagation of short-run causal responses.

Interestingly, even in this simple case we can build a surrogacy-based estimator that is guaranteed to have a lower asymptotic variance than the conventional local projection estimator. The central idea is to optimally combine the surrogacy-based estimator with the conventional LP estimator as these estimators are efficient in different parts of the parameter space and combining them leads leads to a uniform reduction in variance. To show this, suppose that the researcher considers $\hat{\theta}^{\text{s}}_h = \hat{\beta}_{h0}\hat{\theta}_0$, where $\hat{\beta}_{h0}$ is obtained by regressing w_{t+h} on w_t. This surrogate estimator can outperform the conventional LP estimator $\hat{\theta}^{\text{lp}}$, but only in certain parts of the parameter space where the direct regression of w_{t+h} on w_t is informative.

To illustrate the variance trade-off, Figure 2 shows the percent asymptotic variance reduction from $\hat{\theta}^{\text{s}}_h$ relative to conventional LP across AR(1) persistence values and horizons. Blue regions indicate that the surrogacy estimator is more precise; red regions indicate that LP is more precise. Surrogacy can lower variance, but only when persistence is high and the short-run outcomes remain informative about future outcomes. However, as persistence falls or the horizon grows, the response dies out more quickly, the surrogate w_t contains less information about the long-horizon response, and the SUR variance advantage fades and can ultimately reverse.[4]

[4]The variance reduction from SUR can reverse because the SUR-based estimator is a product estimator, in that it must estimate both the projection coefficient β_{h0} *and* the short-run response θ_0. When the impulse response coefficient $\theta_h = \rho^h$ is small, either because persistence is low or because the horizon h

Figure 2: Variance gains from surrogacy in an AR(1)



Notes: The figure compares the variance of regular LP estimates with the direct surrogacy estimator in the AR(1) model $w_t = \rho w_{t-1} + \varepsilon_t$. We report the asymptotic variance reduction from $\hat{\theta}_h^{\text{s}}$ relative to $\hat{\theta}_h^{\text{lp}}$. Positive values indicate lower variance than LP; negative values indicate that LP has lower variance.

To ensure asymptotic variance reduction uniformly over the parameter space we combine the two estimators; the surrogate LP estimator is defined as

$$\hat{\theta}_h^{\text{sur}} = \hat{\omega}\hat{\theta}_h^{s} + (1 - \hat{\omega})\hat{\theta}_h^{\text{lp}} \ ,$$

where $\hat{\omega}$ is an estimate for the weight that is chosen to minimize the asymptotic variance of $\hat{\theta}_h^{\text{sur}}$.

There exist many alternative surrogate estimators. For instance, if the researcher considers $\hat{\theta}_h^{\text{s}} = \hat{\rho}^h \hat{\theta}_0$, where $\hat{\rho}$ is obtained by regressing w_t on w_{t-1}, the asymptotic variance of this surrogacy estimator is even lower when compared to the direct regression based estimator, but an explicit stance on the recursive propagation model is required. Below we refer to this type of estimator as a surrogate VAR estimator.

is large, the surrogate projection contains little useful information about the long-horizon response. The potential variance gain from using the surrogate is then small, and the additional sampling uncertainty from estimating θ_0 remains. In this region, the extra uncertainty introduced by the surrogate step can dominate the gain from avoiding a direct long-horizon LP estimate, causing $\hat{\theta}_h^{\text{sur}}$ to have higher variance than LP.

Approximate surrogacy. The more interesting case arises when $b \neq 0$. In this case, w_t is no longer a sufficient state variable because the latent innovation ε_t also affects future outcomes through the moving-average component. Exact surrogacy therefore fails for any finite surrogate horizon s. Intuitively, in an invertible ARMA(1,1), the innovation ε_t can only be recovered from the *infinite* history of observables, implying that no finite vector $(w_t, \ldots, w_{t+s})$ perfectly spans the latent state.

However, exact recovery of the latent state is not necessary for accurate long-horizon propagation. What matters instead is whether the economically relevant component of the latent state becomes increasingly spanned by observable short-run outcomes. In the ARMA(1,1) model, the omitted moving-average component affects future outcomes only transiently. As the surrogate horizon s increases, the predictive content of the original shock conditional on $(w_t, \ldots, w_{t+s})$ becomes increasingly small.

Formally, define the surrogacy deviation term

$$\xi_t^{hs} = \mathrm{Proj}(w_{t+h} \mid w_{t:t+s}, \varepsilon_t) - \mathrm{Proj}(w_{t+h} \mid w_{t:t+s}).$$

For the ARMA(1,1) model, this deviation is governed by the part of ε_t that remains unrecovered after observing the surrogate vector. In particular, we have [5]

$$\|\xi_t^{hs}\|_2 \lesssim |b|^{s+1}|\rho|^{h-s-1}\sigma^2 \ . \tag{4}$$

Thus, under invertibility, $|b| < 1$, the surrogacy deviation shrinks geometrically with both the surrogate horizon s and the forecasting horizon h. Increasing the surrogate horizon reveals progressively more information about the latent innovation ε_t, reducing the component of the original shock that remains unspanned by the observed short-run outcomes. At the same time, for fixed s, the residual effect of that unrecovered component on future outcomes decays with the forecasting horizon because the autoregressive propagation term $|\rho|^{h-s-1}$ becomes smaller as h increases.

Approximate surrogacy can therefore be interpreted as a form of short-lived omitted-

[5] To see this, set the shock date to zero and write

$$w_t = \rho w_{t-1} + \varepsilon_t + b\varepsilon_{t-1}.$$

Then

$$u_j \equiv w_j - \rho w_{j-1} = \varepsilon_j + b\varepsilon_{j-1}.$$

Iterating backward gives

$$\varepsilon_s = u_s - bu_{s-1} + \cdots + (-b)^{s-1}u_1 + (-b)^s\varepsilon_0.$$

Thus, conditional on $w_{0:s}$, the remaining dependence of the future state ε_s on ε_0 is proportional to $(-b)^s$. Since, for $h > s$, the effect of ε_s on w_h is $b\rho^{h-s-1}$, the surrogacy deviation is proportional to $b\rho^{h-s-1}(-b)^s$, and hence has magnitude $|b|^{s+1}|\rho|^{h-s-1}$.

state dependence. The latent moving-average component is not directly observed, but its influence gradually becomes encoded in the observed short-run dynamics and eventually dies out. Consequently, once a sufficiently rich block of intermediate outcomes has been observed, or once sufficiently long horizons are considered, the original shock contains little additional predictive information for future outcomes beyond what is already summarized by the surrogate vector.

More generally, approximate surrogacy holds whenever the components of the latent state not fully captured by the observed short-run outcomes have only transitory effects on future outcomes. In the ARMA(1,1) example, the omitted state is the innovation ε_t. Even though a finite vector of short-run outcomes does not perfectly reveal that innovation, the remaining approximation error decays geometrically because the omitted moving-average component becomes progressively less relevant for predicting distant outcomes. Thus, a modest number of short-run outcomes can be sufficient for accurate long-run propagation when the effect of the latent state decays sufficiently quickly over time. In our general treatment below we formalize what is sufficiently quickly decaying by making the approximation depend on the sample size, allowing for local-to-zero parametrization of the model parameters.

Approximately unbiased propagation. Under approximate surrogacy, the long-horizon impulse response admits the decomposition

$$\theta_h = \boldsymbol{\beta}_{hs}\boldsymbol{\theta}_{0:s} + \Delta_{s,h},$$

where $\boldsymbol{\theta}_{0:s} = (\theta_0, \ldots, \theta_s)'$ collects the short-run impulse responses, $\boldsymbol{\beta}_{hs}$ denotes the predictive projection coefficient from regressing w_{t+h} onto the surrogates $(w_t, \ldots, w_{t+s})$ and $\Delta_{s,h}$ is the bias term. Hence,

$$\theta_h \simeq \boldsymbol{\beta}_{hs}\boldsymbol{\theta}_{0:s}$$

whenever the approximation error $\Delta_{s,h}$ is sufficiently small.

This decomposition implies that long-horizon impulse responses can be estimated by combining short-run causal effects with predictive propagation estimated from the unconditional variation in the data. Importantly, the predictive regression is not interpreted causally. Rather, it acts as a forecasting bridge linking the observed short-run responses to the long-run outcome.

A key insight is that, even when the unconditional dynamics of the observables do not admit a simple finite-order recursive representation, the *predictive propagation* after the surrogacy horizon may still have a much simpler structure. In the ARMA(1,1) example, the unconditional dynamics are not governed by a finite-order autoregression because the latent moving-average component remains part of the state. Nevertheless, once the surrogate vector

$(w_t, \ldots, w_{t+s})$ has absorbed most of the economically relevant information about the latent innovation, the remaining predictive propagation coefficients inherit a simple autoregressive structure. In particular, for $h > s$,

$$\boldsymbol{\beta}_{hs} = \rho^{h-s-1}\boldsymbol{\beta}_{s+1,s}.$$

Thus, although the unconditional dynamics are ARMA, the post-surrogacy predictive propagation evolves recursively according to the stable autoregressive component. Intuitively, once the short-run dynamics have revealed most of the latent state, the remaining long-run propagation becomes substantially simpler than the original data-generating process.

The resulting estimator need not be exactly unbiased because surrogacy only holds approximately. However, if the surrogate horizon is chosen sufficiently conservatively, the approximation error becomes negligible and the resulting estimator remains highly accurate. In our formal analysis below, we make this idea precise by allowing the approximation error to vanish asymptotically with the sample size.

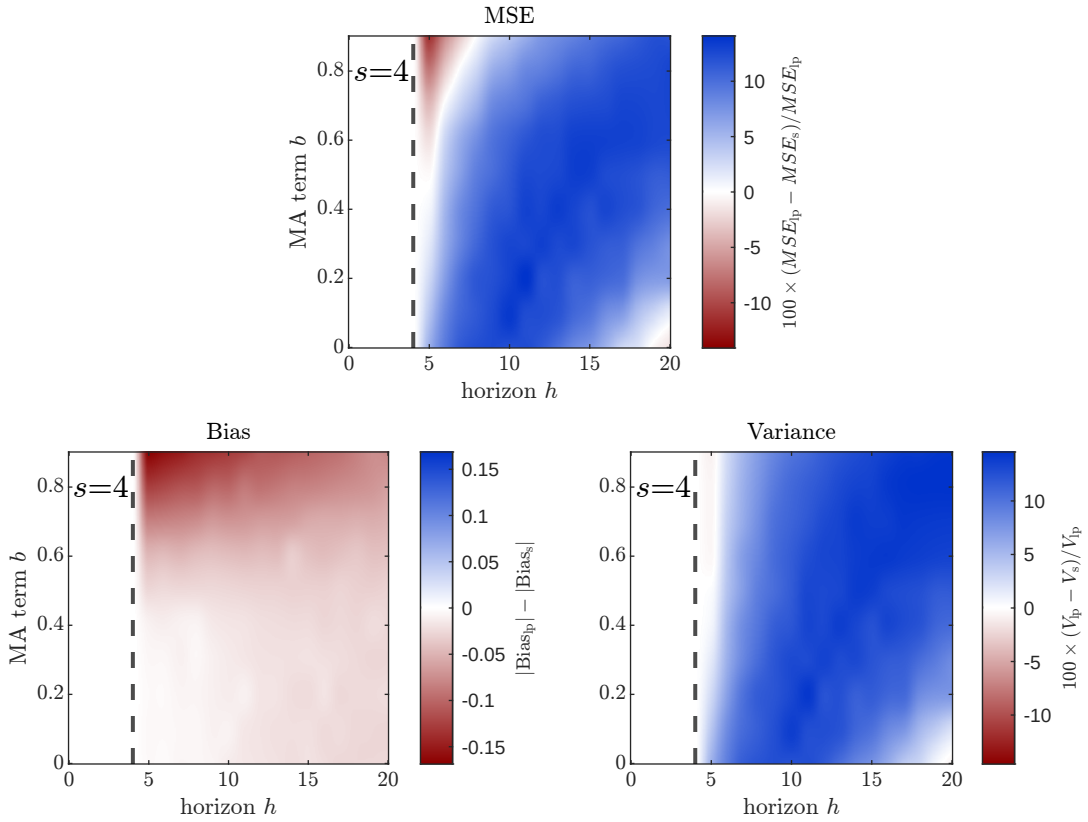
Variance reduction and approximation error. Figure 3 illustrates the bias and variance implied by approximate surrogacy in the ARMA(1,1) environment. The figure shows the relative asymptotic bias and variance of the conventional LP estimator to that of the SUR estimator based on $\hat{\theta}_h^{\text{sur}} = \hat{\omega}\hat{\theta}_h^s + (1-\hat{\omega})\hat{\theta}_h^{\text{lp}}$, with $\hat{\theta}_h^s = \hat{\boldsymbol{\theta}}_{0:s}\hat{\boldsymbol{\beta}}_{hs}$ and $\hat{\boldsymbol{\beta}}_{hs}$ the regression coefficients from regressing w_{t+h} on $w_{t:t+s}$. We fix the surrogate horizon at $s = 4$.

For moderate values of the moving-average coefficient b, the figure for bias is predominantly white, indicating similar bias across the two methods. At the same time the variance figure is largely blue, showing substantial gains in variance from surrogacy. In this region, the short-run outcomes capture most of the economically relevant component of the latent state, so the approximation error remains small while the predictive propagation substantially improves precision relative to conventional LPs.

As the moving-average component becomes larger, the approximation error from finite-dimensional surrogates becomes more important at intermediate horizons because the surrogate vector no longer spans enough of the latent innovation. This generates the red region in the upper part of the bias figure. Nevertheless, the deterioration is limited and tends to disappear again at longer horizons because the omitted moving-average component is transitory and its effect gradually dies out through the autoregressive propagation term $|\rho|^{h-s-1}$.

The figure therefore illustrates the central motivation for surrogacy-based impulse response estimation. Even when the underlying dynamics are not governed by a finite-order autoregression and exact surrogacy fails, a relatively small number of short-run outcomes can still capture enough of the latent state to substantially reduce long-horizon variance

Figure 3: Gains from surrogacy in an ARMA(1,1)



Notes: The figure compares LP and SUR estimates in the ARMA(1,1) model $w_t = \rho w_{t-1} + \varepsilon_t + b\varepsilon_{t-1}$, with $\rho = 0.95$ and $s = 4$. The upper panel reports $100(MSE_h^{LP} - MSE_h^{SUR})/MSE_h^{LP}$; the lower-left panel reports $|\text{Bias}_h^{LP}| - |\text{Bias}_h^{SUR}|$; and the lower-right panel reports $100(V_h^{LP} - V_h^{SUR})/V_h^{LP}$. Positive values indicate gains from SUR; red regions indicate cases where LP performs better. Differences are set to zero for $h \leq s$ by construction.

while introducing only limited approximation error.

An alternative surrogacy estimator can be based on exploiting the autoregressive structure in the predictive $\boldsymbol{\beta}_{hs}$ coefficients. Such estimators can further reduce the variance relative to the LP estimate, but they run more risk of introducing additional bias due to the additional structure assumed.

Surrogacy in time series: some special cases

The surrogacy-based impulse response decomposition, i.e. $\theta_h \simeq \boldsymbol{\beta}_{hs}\boldsymbol{\theta}_{0:s}$ for $h > s$, includes a few special cases that are well known in economics. First, and somewhat obvious, when s is larger than the largest impulse horizon of interest the impulse response function becomes equal to that of the conventional local projection for all horizons. This is important, as for some data generating processes the local projection may simply be the best possible

estimator and our data driven selection rule rule for s, discussed in Section 5, allows the surrogate estimator to coincide with the local projection.

Second, any structural VAR impulse response estimate is a special case of a surrogate estimator with $s = 0$. Specifically, recall that the impulse response of variable i to shock j in a structural VAR can be written as

$$\hat{\theta}_h^{\text{svar}} = \mathbf{e}_i' \mathbf{A}^h \mathbf{B} \mathbf{e}_j \ ,$$

where $\mathbf{e}_i$ has zeros everywhere except in position i where there is a one, $\mathbf{A}$ is the autoregressive matrix of the VAR in companion form and $\mathbf{B}$ is the impact matrix. The decomposition shows that the contemporaneous impact of the jth shock on the variables of the SVAR — i.e. $\mathbf{B}\mathbf{e}_j$ — is propagated to the ith variable h periods ahead using the predictive coefficients $\mathbf{e}_i'\mathbf{A}^h$. Our surrogate VAR generalizes this approach by allowing for direct effects of the shock on multiple horizons s and not just the contemporaneous horizon. This flexibility allows the surrogate VAR to avoid dynamic mis-specification for the first s periods.

3 Surrogacy in time series

In this section, we formally define our approximate surrogacy assumption and derive the implications of approximate surrogacy for impulse response estimation. Let ε_t denote a structural shock, $\mathbf{w}_t$ a vector of outcome variables and $\mathbf{x}_{t-1}$ a vector of pre-determined control variables. We are interested in the effect of ε_t on $\mathbf{w}_{t+h}$ for $h = 0, \ldots, H$ after controlling for $\mathbf{x}_{t-1}$. We assume that all variables are observed for time periods $t = 1, \ldots, T$. The case in which the shock is not observed directly, but an informative proxy is available, is discussed in the appendix.

Assumption 1. We assume that there exists a vector $\mathbf{y}_{t:t+s} = (\mathbf{y}_t', \ldots, \mathbf{y}_{t+s}')'$ such that, for all $h = s+1, \ldots, H$,

$$\text{Proj}(\mathbf{w}_{t+h} \mid \mathbf{y}_{t:t+s}, \mathbf{x}_{t-1}, \varepsilon_t) = \text{Proj}(\mathbf{w}_{t+h} \mid \mathbf{y}_{t:t+s}, \mathbf{x}_{t-1}) + \boldsymbol{\xi}_t^{hs},$$

where

$$\|\boldsymbol{\xi}_t^{hs}\|_2 = O(T^{-a}), \qquad \text{for some } a > \frac{1}{2} \ .$$

In words, after the *surrogacy horizon* s, and conditional on the *surrogates* $\mathbf{y}_{t:t+s}$, the shock ε_t has only a very small, asymptotically negligible, effect on the outcomes. The surrogacy condition is closely related to the notion of Granger non-causality (Granger, 1969; Sims, 1972): conditional on the surrogate variables, the shock no longer contains substantial additional predictive information for future outcomes. Allowing for mild deviations from

exact surrogacy, i.e. the $O(T^{-a})$ term, is important in time series settings because exact finite-dimensional surrogacy rarely holds, as illustrated by the ARMA(1,1) example in the previous section. Modeling the $\boldsymbol{\xi}_t^{hs}$ term as a vanishing function of the sample size naturally aligns with the local-to-zero framework commonly used in econometrics to model small departures from idealized benchmark models (e.g. Newey, 1985; Staiger and Stock, 1997). The following example illustrates the link.

Example 1. Recall the ARMA$(1,1)$ process discussed in the previous section,

$$w_t = \rho w_{t-1} + \varepsilon_t + \frac{b}{\sqrt{T}}\varepsilon_{t-1} \ ,$$

where the moving-average coefficient is now modeled local-to-zero, reflecting that moving-average components are often small and difficult to detect in finite samples in macro time series (e.g. Schorfheide, 2005; Müller and Stock, 2011; Olea et al., 2024; González-Casasús and Schorfheide, 2025*b*).

As shown in the previous section, exact surrogacy fails because the innovation ε_t cannot be perfectly recovered from a finite set of observables. However, under the local-to-zero parameterization this failure becomes asymptotically negligible. Define $c_T = b/\sqrt{T}$. Then

$$(1-\rho L)w_t = \varepsilon_t + c_T\varepsilon_{t-1}.$$

Since the process is invertible,

$$\varepsilon_t = (1+c_T L)^{-1}(1-\rho L)w_t = \left(1 - c_T L + c_T^2 L^2 - \cdots\right)(1-\rho L)w_t.$$

Equivalently,

$$\varepsilon_{t+s} = r_{t+s} - c_T r_{t+s-1} + \cdots + (-c_T)^{s-1} r_{t+1} + (-c_T)^s \varepsilon_t,$$

where $r_t = (1-\rho L)w_t$. Hence, conditional on the surrogate block $w_{t:t+s} = (w_t, \ldots, w_{t+s})$, the only remaining dependence on ε_t enters through the final term $(-c_T)^s\varepsilon_t$.

As shown in the illustrative discussion above, future outcomes depend on the omitted innovation only through the moving-average propagation term. Therefore, for every $h > s$,

$$\text{Proj}(w_{t+h} \mid w_{t:t+s}, \varepsilon_t) = \text{Proj}(w_{t+h} \mid w_{t:t+s}) + \xi_t^{hs},$$

with $\|\xi_t^{hs}\|_{L^2} = O(T^{-(s+1)/2})$. Therefore Assumption 1 holds with $a = (s+1)/2$. In particular, any $s \geq 1$ implies $a > 1/2$. □

This example extends naturally to VARMA(p,q) models, an important class of time series models arising from dynamic stochastic general equilibrium models. As in the illustrative

discussion above, the surrogacy horizon is determined by how quickly the omitted moving-average component becomes negligible. When the moving-average coefficients drift to zero at rate $1/\sqrt{T}$, the approximation error vanishes sufficiently quickly once the surrogate horizon is chosen large enough relative to the moving-average order. As we will see below, the benefits of surrogacy are typically greatest when the surrogate horizon s remains small relative to the maximum horizon H of interest.

The surrogacy condition implies that the dynamic causal effect of ε_t on $\mathbf{w}_{t+h}$ for horizons $h > s$ can be approximately decomposed into a short-run causal effect and a long-run prediction. The following lemma formalizes this.

Lemma 1. Define the dynamic causal effect

$$\boldsymbol{\theta}_h = \text{Proj}(\mathbf{w}_{t+h} \mid \mathbf{x}_{t-1}, \varepsilon_t = 1) - \text{Proj}(\mathbf{w}_{t+h} \mid \mathbf{x}_{t-1}, \varepsilon_t = 0) \ .$$

Suppose Assumption 1 holds. Then, for $h = s+1, \ldots, H$,

$$\boldsymbol{\theta}_h = \boldsymbol{\beta}_{hs}\boldsymbol{\tau}_{0:s} + O(T^{-a}) \ ,$$

where $\boldsymbol{\beta}_{hs}$ are coefficients from the projection $\text{Proj}(\mathbf{w}_{t+h} \mid \mathbf{y}_{t:t+s}, \mathbf{x}_{t-1}) = \boldsymbol{\beta}_{hs}\mathbf{y}_{t:t+s} + \boldsymbol{\gamma}_{hs}\mathbf{x}_{t-1}$, and $\boldsymbol{\tau}_{0:s} = (\boldsymbol{\tau}_0', \ldots, \boldsymbol{\tau}_s')'$, with $\boldsymbol{\tau}_j = \text{Proj}(\mathbf{y}_{t+j} \mid \mathbf{x}_{t-1}, \varepsilon_t = 1) - \text{Proj}(\mathbf{y}_{t+j} \mid \mathbf{x}_{t-1}, \varepsilon_t = 0)$.. □

The lemma decomposes the impulse response of interest into the short-run causal effects $\boldsymbol{\tau}_{0:s}$, which capture the effect of the shock on the surrogates, and the long-run predictive coefficients $\boldsymbol{\beta}_{hs}$, which capture the predictive effect of the surrogates on the outcome variables. The decomposition is not exact as a result of the approximation in the surrogacy condition, but the distortion vanishes with the sample size.

Surrogacy vs VAR assumptions

It is instructive to compare the surrogacy assumption to the assumptions underlying impulse response analysis in a VAR model. Consider

$$\mathbf{y}_t = \mathbf{A}_1\mathbf{y}_{t-1} + \ldots + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{u}_t \ , \tag{5}$$

where the VAR assumes that the errors are serially uncorrelated, i.e. $\text{Cov}(\mathbf{u}_t, \mathbf{u}_{t-j}) = 0$. To compare the serially uncorrelated error assumption, we also assume that the errors are mean zero, which is without loss of generality, and that ε_t is equal to $\mathbf{u}_{1t}$, i.e. partial invertibility holds in the VAR.

Proposition 1. For the VAR model (5) we have that

$$\text{Cov}(\mathbf{u}_t, \mathbf{u}_{t-j}) = \mathbf{0} \quad \Rightarrow \quad \text{Proj}(\mathbf{y}_{t+h} \mid \mathbf{y}_{t:t+s}, \varepsilon_t) = \text{Proj}(\mathbf{y}_{t+h} \mid \mathbf{y}_{t:t+s}) \ ,$$

but the converse is not true. □

Compared to surrogacy, the no serial correlation assumption is substantially stronger, as it adds three requirements: (i) Surrogacy holds for all types of treatment/shock that can affect the variables in the VAR, (ii) all variables in the VAR are surrogates for the other variables in the VAR, a form of "circular surrogacy", and (iii) the "effects" of the surrogates on long-term outcomes can be computed recursively.

To sum up, as discussed in the introduction local projections impose no dynamic restrictions on the impulse responses and thus requires no additional assumption, while the VAR relies on the strong no serial correlation assumption to bring additional efficiency gains. Surrogacy opens up a middle road between these two extremes: under the weaker surrogacy assumption we can exploit Lemma 1 and model impulse responses by combining short run causal effects with long run prediction to improve the efficiency of LP estimates without paying a large bias cost.

4 Surrogate LPs and VARs

We impose the surrogacy assumption 1 and discuss different estimation methods that can be adopted for constructing surrogacy based estimators. Specifically, we first discuss the estimation of the short-run causal effects $\boldsymbol{\tau}_{0:s}$ and the long-run predictive coefficients $\boldsymbol{\beta}_{hs}$. For horizons $h > s$, we then combine these with the conventional local projection moment conditions for the impulse responses $\boldsymbol{\theta}_h$ in a joint GMM estimator that imposes the surrogacy restriction $\boldsymbol{\theta}_h = \boldsymbol{\beta}_{hs}\boldsymbol{\tau}_{0:s}$. The surrogacy horizon s and the surrogate variables are treated as given throughout this section. Their selection is discussed in Section 5.

4.1 Short-run causal effects

For simplicity, assume that $\mathbf{w}_t$ is a subset of $\mathbf{y}_t$, such that for horizons $h \leq s$, the impulse responses $\boldsymbol{\theta}_h$ are a subset of the short-run causal effects $\boldsymbol{\tau}_h$ defined in Lemma 1. These effects can be estimated using standard local projections. Specifically, when the shock ε_t is observed, we consider

$$\mathbf{y}_{t+h} = \boldsymbol{\tau}_h \varepsilon_t + \mathbf{B}_h \mathbf{x}_{t-1} + \boldsymbol{\zeta}_{t+h}, \qquad h = 0, \ldots, s. \tag{6}$$

When only a proxy for ε_t is observed, the same regression can be estimated using IV methods after selecting a suitable normalizing variable for the instrument. We provide the details for the IV estimator in Appendix B.

Going back to the intuition, after a shock occurs many different variables respond dynamically. Many of these responses cannot be measured by the econometrician and are therefore omitted from the analysis. As shown by Olea et al. (2024), local projections possess attractive robustness properties in such dynamically misspecified environments. Alternatively, one could estimate a VAR for $(\mathbf{y}_t, \varepsilon_t)$ with lag length at least equal to horizon s, which yields asymptotically equivalent impulse response estimates over these horizons.

Under standard regularity conditions, the least-squares estimator $\hat{\boldsymbol{\tau}}_h$ satisfies

$$\sqrt{T}(\hat{\boldsymbol{\tau}}_h - \boldsymbol{\tau}_h) \xrightarrow{d} N(0, \Omega_h),$$

as $T \to \infty$, where Ω_h denotes the asymptotic covariance matrix. Inference therefore proceeds exactly as in conventional local projections using heteroskedasticity and serial-correlation robust variance estimators. Alternatively, lag-augmentation methods as in Montiel Olea and Plagborg-Møller (2021) may be employed, in which case heteroskedasticity-robust covariance estimators remain valid. Confidence intervals can then be constructed in the standard way.

The regression (6) also defines the moment conditions for the short-run causal effects $\boldsymbol{\tau}_{0:s}$. These moment conditions will later be combined with the long-run predictive and local projection moment conditions in a joint GMM estimator for the long-run impulse responses. Nevertheless, for horizons $h \leq s$, inference remains identical to conventional local projection inference and does not rely on the surrogacy assumption.

4.2 Long-run predictive and causal effects

For horizons $h > s$, the impulse response $\boldsymbol{\theta}_h$ can be estimated using three complementary sources of information. First, the short-run local projections (6) identify the causal effects $\boldsymbol{\tau}_{0:s}$. Second, predictive models identify the long-run predictive coefficients $\boldsymbol{\beta}_{hs}$. Third, the conventional local projection directly identifies the impulse response $\boldsymbol{\theta}_h$. We consider two different predictive models: direct forecasts, leading to Surrogate LPs, and recursive VAR forecasts, leading to Surrogate VARs. In both cases, these moment conditions are combined in a joint GMM estimator that imposes the surrogacy restriction $\boldsymbol{\theta}_h = \boldsymbol{\beta}_{hs}\boldsymbol{\tau}_{0:s}$.

Surrogate LP

The most straightforward way to implement the surrogacy predictions is by direct forecasting. That is, we consider the predictive regression

$$\mathbf{w}_{t+h} = \boldsymbol{\beta}_{hs}\mathbf{y}_{t:t+s} + \boldsymbol{\gamma}_{hs}\mathbf{x}_{t-1} + \boldsymbol{\eta}_{t+h}, \qquad h = s+1, \ldots, H, \tag{7}$$

based on which $\boldsymbol{\beta}_{hs}$ can be estimated using least squares. Alternatively, one may consider a penalized estimator such as Lasso or Ridge.

For each horizon $h > s$, we also consider the conventional local projection

$$\mathbf{w}_{t+h} = \boldsymbol{\theta}_h\varepsilon_t + \boldsymbol{\delta}_h\mathbf{x}_{t-1} + \mathbf{e}_{t+h}, \tag{8}$$

which directly estimates the impulse response.

The regressions (6), (7), and (8) define three sets of moment conditions. Let

$$\mathbf{z}_t^{\tau} = (\varepsilon_t, \mathbf{x}_{t-1}')', \qquad \mathbf{z}_t^{\beta} = (\mathbf{y}_{t:t+s}', \mathbf{x}_{t-1}')', \qquad \mathbf{z}_t^{\theta} = (\varepsilon_t, \mathbf{x}_{t-1}')'$$

denote the corresponding vectors of regressors. The associated moment conditions are, for $j = 0, \ldots, s$,

$$\mathbb{E}[\mathbf{z}_t^{\tau}\boldsymbol{\zeta}_{t+j}'] = \mathbf{0}, \qquad \mathbb{E}[\mathbf{z}_t^{\beta}\boldsymbol{\eta}_{t+h}'] = \mathbf{0}, \qquad \mathbb{E}[\mathbf{z}_t^{\theta}\mathbf{e}_{t+h}'] = \mathbf{0}.$$

Together these moments identify the unknown parameters $\boldsymbol{\tau}_{0:s}$, $\boldsymbol{\beta}_{hs}$, and $\boldsymbol{\theta}_h$. Rather than estimating these parameters separately, we estimate them jointly by GMM while imposing the surrogacy restriction $\boldsymbol{\theta}_h = \boldsymbol{\beta}_{hs}\boldsymbol{\tau}_{0:s}$. Specifically, let $\hat{\mathbf{g}}_T(\boldsymbol{\psi}_{hs})$ denote the stacked sample analogue of the moment conditions above, with $\boldsymbol{\psi}_{hs} = (\boldsymbol{\theta}_h, \boldsymbol{\beta}_{hs}, \boldsymbol{\tau}_{0:s})$. The *Surrogate LP* estimator is defined as

$$(\hat{\boldsymbol{\theta}}_h^{\text{sur}}, \hat{\boldsymbol{\beta}}_{hs}, \hat{\boldsymbol{\tau}}_{0:s}) = \arg\min_{\boldsymbol{\psi}_{hs}} \hat{\mathbf{g}}_T(\boldsymbol{\psi}_{hs})'\mathbf{W}_T\hat{\mathbf{g}}_T(\boldsymbol{\psi}_{hs}), \qquad \text{subject to} \qquad \boldsymbol{\theta}_h = \boldsymbol{\beta}_{hs}\boldsymbol{\tau}_{0:s}, \tag{9}$$

where $\mathbf{W}_T$ is a positive semidefinite weighting matrix.

Compared to conventional LP, which estimates $\boldsymbol{\theta}_h$ solely from the local projection moment conditions, the surrogate estimator combines two complementary sources of information about the long-run impulse response. The conventional LP moments directly identify $\boldsymbol{\theta}_h$, while the predictive moments exploit the surrogacy restriction to link $\boldsymbol{\theta}_h$ to the short-run causal effects and the long-run predictive coefficients. The resulting constrained GMM estimator therefore combines the information contained in both approaches.

Proposition 2. Fix $h > s$ and suppose Assumptions 1 and 3 hold. Let $\hat{\boldsymbol{\theta}}_h^{\text{sur}}$ denote the

constrained GMM estimator defined in (9). Then

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_h^{\text{sur}} - \boldsymbol{\theta}_h\right) \xrightarrow{d} N(0, \Omega_h^{\text{sur}}),$$

where Ω_h^{sur} is the asymptotic covariance matrix of the constrained GMM estimator. Moreover, if $\mathbf{W}_T$ converges to the efficient GMM weighting matrix,

$$\Omega_h^{\text{sur}} \leq \text{Avar}\left(\hat{\boldsymbol{\theta}}_h^{LP}\right),$$

□

The predictive moments can be substantially more informative than the conventional LP moments. A key reason is that the unconditional direct projection in (7) can have much smaller residuals than the local projection, since the right-hand side variables are not just ε_t, but the entire surrogate state $\mathbf{y}_{t:t+s}$, which includes (i) the contribution of all shocks, and (ii) future outcomes in addition to the contemporaneous observation. Intuitively, conventional LP only exploits the variation explained by the shock ε_t to estimate the impulse response. If the shock explains little of the variation in the outcome variable of interest, as is often the case for monetary policy shocks, the LP estimator can be imprecise. Surrogacy alleviates this limitation because, once one conditions on the outcomes over the first s horizons after treatment, the shock no longer has predictive content for subsequent outcomes. This makes it possible to estimate forecasting models using the unconditional data.

Surrogate VAR

While combining short-run LPs with long-run direct forecasts can improve efficiency relative to conventional LP, it still leaves the impulse response unrestricted across horizons. Additional efficiency gains may be obtained by imposing parametric restrictions on the post-surrogacy dynamics. We therefore consider a Surrogate VAR estimator that replaces the unrestricted post-surrogacy dynamics by a finite-order VAR while combining the resulting moment conditions with those introduced for the Surrogate LP estimator.

Conditional on the surrogacy horizon s defined by Assumption 1, we impose additional structure on the post-surrogacy dynamics. Specifically, suppose that, for $h > s$ and some $p \geq 1$, the impulse responses satisfy

$$\boldsymbol{\theta}_h = \mathbf{A}_1\boldsymbol{\theta}_{h-1} + \cdots + \mathbf{A}_p\boldsymbol{\theta}_{h-p}. \tag{10}$$

Equivalently, the post-surrogacy response is generated by the companion dynamics associated with the coefficient matrices $\mathbf{A}_1, \ldots, \mathbf{A}_p$.

The motivation for imposing a VAR representation only after the surrogacy horizon is that, while the short-run dynamics of a treatment effect may be complex, the propagation of the shock may simplify beyond horizon s, allowing a parsimonious VAR to deliver additional efficiency gains. Returning to the ARMA$(1,1)$ example from Section 2, the impulse responses decay geometrically after the surrogacy horizon. In that case, a simple VAR(1) fitted only to capture the autocovariance structure beyond horizon s provides a parsimonious approximation to the long-run dynamics that can be estimated more efficiently than unrestricted local projections.

To characterize the additional moment conditions implied by (10), define the VAR innovations

$$\mathbf{u}_t^A = \mathbf{y}_t - \mathbf{A}_1\mathbf{y}_{t-1} - \cdots - \mathbf{A}_p\mathbf{y}_{t-p}.$$

We impose the following additional surrogacy condition.

Assumption 2. For all $h > s$, suppose (10) holds and

$$\mathrm{Proj}\big(\mathbf{y}_{t+h} \mid \mathbf{y}_{t+h-p:t+h-1}, \mathbf{x}_{t-1}, \mathbf{u}_t^A\big) = \mathrm{Proj}(\mathbf{y}_{t+h} \mid \mathbf{y}_{t+h-p:t+h-1}, \mathbf{x}_{t-1}) + \boldsymbol{\xi}_{t,h}^A,$$

where $\|\boldsymbol{\xi}_{t,h}^A\|_2 = O(T^{-a})$ for some $a > 1/2$.

This assumption strengthens the baseline surrogacy condition in two respects. First, it is imposed for the full vector $\mathbf{y}_{t+h}$ rather than only the outcome variables of interest. Second, it rules out additional linear predictive content in the VAR innovation $\mathbf{u}_t^A$ after conditioning on the post-surrogacy state and controls. Nevertheless, it remains substantially weaker than the conventional VAR assumption, since it only restricts the dynamics after the surrogacy horizon and therefore permits arbitrary short-run misspecification.

Assumption 2 implies a set of orthogonality conditions for the propagation coefficients.

Lemma 2. Suppose Assumption 2 holds. Then

$$\mathbb{E}\left[\mathbf{y}_{t-h-p:t-h}\mathbf{u}_t^{A\prime}\right] = O(T^{-a}), \qquad h > s.$$

□

The lemma provides a set of IV moment conditions for the VAR propagation coefficients. Specifically, the lagged vectors $\mathbf{y}_{t-h-p:t-h}$ may be used as instruments for estimating $\mathbf{A}_1, \ldots, \mathbf{A}_p$. Equivalently, the same coefficients satisfy the higher-order Yule–Walker equations

$$\Gamma_{h+1} = \mathbf{A}_1\Gamma_h + \cdots + \mathbf{A}_p\Gamma_{h-p+1}, \qquad h > s,$$

where $\Gamma_h = \mathbb{E}[\mathbf{y}_t\mathbf{y}_{t-h}']$.

To formulate the additional moment conditions, collect the VAR coefficients in

$$\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_p),$$

define

$$\mathbf{X}_{t-1}^p = (\mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-p}')',$$

and let $\mathbf{Z}_t = (\mathbf{y}_{t-\ell_1}', \ldots, \mathbf{y}_{t-\ell_L}')'$ denote a vector of lagged instruments satisfying $\ell_j > s + p$. Then Lemma 2 implies the orthogonality conditions

$$\mathbb{E}[\mathbf{Z}_t \mathbf{u}_t^{A\prime}] = O(T^{-a}),$$

where

$$\mathbf{u}_t^A = \mathbf{y}_t - \mathbf{A}\mathbf{X}_{t-1}^p.$$

The Surrogate VAR estimator combines these IV moment conditions with the short-run local projection moments (6) and the long-run local projection moments (8). To describe the nonlinear restriction, define the recursive mapping $g_h^V(\boldsymbol{\psi}^V)$, where

$$\boldsymbol{\psi}^V = \begin{pmatrix} \mathrm{vec}(\mathbf{A}_1)' & \cdots & \mathrm{vec}(\mathbf{A}_p)' & \boldsymbol{\tau}_0' & \cdots & \boldsymbol{\tau}_s' \end{pmatrix}',$$

by

$$g_j^V(\boldsymbol{\psi}^V) = \mathbf{S}_w \boldsymbol{\tau}_j, \qquad j = 0, \ldots, s,$$

and recursively, for $j > s$,

$$g_j^V(\boldsymbol{\psi}^V) = \mathbf{A}_1 g_{j-1}^V(\boldsymbol{\psi}^V) + \cdots + \mathbf{A}_p g_{j-p}^V(\boldsymbol{\psi}^V).$$

Thus, under (10),

$$\boldsymbol{\theta}_h = g_h^V(\boldsymbol{\psi}^V) + O(T^{-a}).$$

Let $\hat{\mathbf{g}}_T^V$ denote the stacked sample moment vector obtained by augmenting the moment conditions of the Surrogate LP estimator with the IV moment conditions above. We define the *Surrogate VAR* estimator as the constrained GMM estimator

$$\hat{\boldsymbol{\theta}}_h^{\mathrm{surV}} = \arg\min \hat{\mathbf{g}}_T^{V\prime} \mathbf{W}_T \hat{\mathbf{g}}_T^V, \qquad \text{subject to} \qquad \boldsymbol{\theta}_h = g_h^V(\boldsymbol{\psi}^V), \tag{11}$$

where $\mathbf{W}_T$ is a positive semidefinite weighting matrix.

Compared to the Surrogate LP estimator, the Surrogate VAR estimator imposes a finite-dimensional VAR representation on the post-surrogacy dynamics. This additional structure can deliver further efficiency gains whenever the propagation of the impulse response is well

approximated by a low-order VAR. The estimator nests the Surrogate LP estimator when no post-surrogacy restrictions are imposed and nests the conventional VAR estimator when the VAR representation is correctly specified from the impact horizon onward.

The following proposition characterizes the limiting distribution of the Surrogate VAR estimator.

Proposition 3. Fix $h > s$ and suppose Assumptions 1–2 and 4 hold. Let $\hat{\boldsymbol{\theta}}_h^{\text{surV}}$ denote the constrained GMM estimator defined in (11). Then

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_h^{\text{surV}} - \boldsymbol{\theta}_h\right) \xrightarrow{d} N(0, \Omega_h^{\text{surV}}),$$

where Ω_h^{surV} denotes the asymptotic covariance matrix of the constrained GMM estimator. Moreover, if $\mathbf{W}_T$ converges to the efficient GMM weighting matrix,

$$\Omega_h^{\text{surV}} \leq \text{Avar}\left(\hat{\boldsymbol{\theta}}_h^{LP}\right).$$

□

As in the Surrogate LP case, inference may alternatively be conducted using a block wild bootstrap that re-estimates all stages of the Surrogate VAR procedure within each bootstrap replication.

5 Selecting the surrogates

We discuss the selection of the surrogate variables $\mathbf{y}_{t:t+s} = (\mathbf{y}_t', \mathbf{y}_{t+1}', \ldots, \mathbf{y}_{t+s}')'$, which can be alternatively restated as selecting (s, n), where selecting n is understood as selecting the variables included in $\mathbf{y}$. Let $\mathcal{S} = \{(s, n)\}$ denote the set of all candidate surrogate variables considered by the researcher.

Our objective is to select the smallest set of surrogate variables that removes the predictive content of the shock ε_t at post-surrogacy horizons. A natural approach would be to formulate this as a sequence of hypothesis tests and select the first candidate surrogate for which the null hypothesis of no remaining predictive content is not rejected. However, such a procedure is conservative in the wrong direction for our application. Since failure to reject may simply reflect low power, hypothesis-testing approaches tend to select surrogate variables that are too small, especially in finite samples. From the perspective of impulse response estimation, under-selecting s is particularly costly because any remaining misspecification is subsequently propagated recursively through the surrogate forecasting model; see Section 6.1 for an illustration of this point.

For this reason we instead adopt an information-criterion-type approach that explicitly penalizes remaining predictive content of the shock. The resulting procedure is conservative in the opposite direction: it favors larger surrogate variable sets whenever the evidence against surrogacy remains non-negligible. This asymmetry is desirable in our setting because over-selecting s can only increase variance, whereas under-selecting s can propagate dynamic misspecification throughout the impulse response path.

Our selection approach is based directly on the surrogacy definition. Assumption 1 implies that, once we condition on a valid surrogate state, the shock ε_t should no longer have additional predictive content for future outcomes beyond horizon s. To formalize this idea, consider the forecasting regressions

$$\mathbf{w}_{t+h} = \boldsymbol{\gamma}_h \varepsilon_t + \boldsymbol{\delta}_h \mathbf{y}_{t:t+s} + \boldsymbol{\kappa}_h \mathbf{x}_{t-1} + \eta_{i,t+h}, \qquad h = s+1, \ldots, H,$$

where $\mathbf{w}_{t+h}$ denotes the response variables of interest. Under the surrogacy condition we should therefore have $\boldsymbol{\gamma}_h = 0$ for $h = s+1, \ldots, S$.

For each candidate surrogate $(s, n) \in \mathcal{S}$, we compare a restricted forecasting model that excludes ε_t to an unrestricted model that includes it. Let $RSS_R(s, n, h)$ and $RSS_U(s, n, h)$ denote the residual sum of squares from the restricted and unrestricted forecasting regressions at horizon h, respectively. We define the corresponding likelihood-ratio statistic as $LR_h(s, n) = T_h \log(RSS_R(s, n, h)/RSS_U(s, n, h))$, where T_h denotes the effective sample size at horizon h. Intuitively, $LR_h(s, n)$ measures the additional predictive content of the shock ε_t after conditioning on the candidate surrogate variables. If the candidate surrogate is too small, then including ε_t substantially improves forecast accuracy and the likelihood-ratio statistic will be large. Conversely, if the surrogacy condition approximately holds, then the restricted and unrestricted models should have similar fit and the statistic will be small.

Since the surrogacy condition must hold for all horizons beyond s, we aggregate the horizon-specific likelihood-ratio statistics using the maximum statistic

$$LR_{\max}(s, n) = \max_{h=s+1,\ldots,H} LR_h(s, n).$$

The use of the maximum statistic ensures that the candidate surrogate is rejected whenever the shock retains predictive content at any post-surrogacy horizon.

We then select the smallest surrogate variable for which the remaining predictive content of ε_t is sufficiently small relative to a penalty sequence $c_T(s, n)$. Specifically, we define

$$(\hat{s}, \hat{n}) = \min_{(s,n)\in\mathcal{S}} \left\{ LR_{\max}(s, n) \leq c_T(s, n) \right\},$$

where the minimum is taken with respect to model dimension. Our baseline specification

uses the BIC penalty $c_T(s,n) = \log T$. We also report results for the corresponding AIC-type rule, $c_T(s,n) = 2$.

The consistency argument follows from a simple separation property. If a candidate surrogate is invalid, then the shock ε_t retains predictive content at some horizon beyond s. In that case, the unrestricted regression improves the population forecast relative to the restricted regression, and the likelihood-ratio statistic diverges at rate T. If a candidate surrogate is valid, then Assumption 1 implies that the restricted and unrestricted regressions are asymptotically equivalent, because the remaining contribution of ε_t is of order $O(T^{-a})$ with $a > 1/2$. Hence the likelihood-ratio statistic remains bounded in probability. A BIC-type penalty, which diverges but does so more slowly than T, therefore rejects invalid candidates while retaining valid candidates asymptotically.

Proposition 4 (Consistency of surrogate selection)**.** Suppose that Assumption 1 and the regularity conditions in Assumption 5 (Appendix C) hold. If the penalty sequence satisfies $c_T(s,n) \to \infty$ and $c_T(s,n) = o(T)$ then

$$P\left((\hat{s},\hat{n}) = (s_0,n_0)\right) \to 1.$$

□

The proposition applies to any penalty sequence satisfying $c_T(s,n) \to \infty$ and $c_T(s,n) = o(T)$. In particular, the BIC penalty $c_T(s,n) = \log T$ satisfies these conditions and therefore consistently selects the smallest valid surrogate. In contrast, the AIC-type penalty $c_T(s,n) = 2$ remains bounded and therefore does not generally deliver consistent selection. Nevertheless, in finite samples the AIC-type rule may be attractive when the researcher prefers a more conservative choice of s that further guards against dynamic misspecification.

Once the surrogate horizon s has been selected, we separately select the forecasting model used for post-surrogacy propagation. In our baseline examples this amounts to selecting either the forecasting variables for the direct forecast or the lag length of the conditional VAR. For this second step we use conventional information criteria such as AIC. For the surrogate VAR it is important to keep in mind that the recursive parameters are estimated using IV.

6 Simulation evidence

This section evaluates the surrogate LP and surrogate VAR estimators in environments where the true data generating process follows a VARMA process. Our objective is to assess whether surrogate methods can improve the variance of local projection estimators while not paying the cost in terms of additional bias that conventional VARs typically incur.

We consider two complementary classes of simulation designs. The first consists of stylized low-dimensional VARMA models that transparently illustrate the mechanics of surrogate estimation and the role of the surrogate horizon s. In these designs the econometrician observes the full vector of variables, but the dynamics nevertheless fail to admit an exact finite-order VAR representation because of the moving-average component. The second class consists of empirically realistic dynamic factor model (DFM) environments similar to those studied by Li, Plagborg-Møller and Wolf (2024). In these designs the econometrician observes only a small subset of the variables governing the true propagation mechanism, so the observed subsystem behaves approximately like a VAR(∞) due to omitted latent state variables.

In both environments, conventional recursive VAR estimators attempt to summarize the entire propagation mechanism using a low-dimensional autoregressive representation. Short recursive specifications may therefore propagate misspecification bias throughout the impulse response path, while richer recursive specifications can become highly variable in finite samples. Direct estimators such as local projections avoid imposing recursive restrictions on the propagation mechanism, but may suffer from substantial variance at longer horizons.

6.1 Illustrative VARMAs

We first consider stylized simulation designs based on a five-variable VARMA$(3, 2)$ model for $\mathbf{y}_t$. These designs are intended to transparently illustrate the mechanics of surrogate estimation and the role of the surrogate horizon s. Although the econometrician observes the full vector of variables, the moving-average component implies that the observed process does not admit an exact finite-order VAR representation. Conventional recursive VAR estimators are therefore misspecified even though all relevant variables are observed.

We consider four different parametrizations chosen to generate impulse responses with distinct shapes and persistence patterns for the effect of the first shock ε_t on the second variable $w_t = \mathbf{y}_{2t}$. For each parametrization we simulate 5000 datasets of length $T = 200$. To illustrate the workings of the surrogate estimators we report results for several choices of the surrogate horizon s each selecting $\mathbf{y}_{t:t+s}$ as the surrogates. For the surrogate VAR we use $p = 3$ lags and six lags of instruments for estimation. In the larger scale study DFM study below we investigate data-driven selection of s and p, as well as the sensitivity of the results to the choice of instruments.

Figure 4 summarizes the main findings. We find that the conventional VAR behaves most erratically across designs. In some cases the AIC criterion selects a lag length capable of approximating the true impulse response well (e.g. DGP 2), while in others the estimated VAR responses are substantially distorted. In contrast, the LP and surrogate estimators

(when s is large enough) behave more systematically. For the first s horizons all surrogate estimators coincide with the LP because they rely on direct estimation. Beyond horizon s, the performance depends critically on the choice of the surrogate horizon.

When $s = 3$ or 6 for some designs, the surrogate LP estimator has average bias equal to the LP estimator, but its variance is generally lower after horizon s. The gains in variance are between 20-30% after horizon 10 and across DGPs. We omit the surrogate VAR for clarity of the figure. The main message is that choosing s correctly is essential for the surrogate estimators; when we select $s = 1$ the estimator starts showing substantial bias comparable to the conventional VAR.

6.2 Dynamic factor model

Next, we consider empirically realistic simulation environments based on the dynamic factor model framework of Li, Plagborg-Møller and Wolf (2024). The data generating process is a large-scale dynamic factor model estimated on U.S. macroeconomic data, generating persistent common dynamics, strong cross-sectional dependence, and long-run propagation patterns that are difficult to summarize using small recursive systems.

Following the setup of Li, Plagborg-Møller and Wolf (2024), the econometrician observes only a small (randomly selected) subset of the variables of the dynamic factor model which behave approximately like a VAR(∞). This setting provides a natural laboratory for evaluating the surrogate estimators. Conventional recursive VARs attempt to summarize the entire latent propagation mechanism using a small autoregressive representation, which may induce substantial bias at medium and long horizons. In contrast, local projections remain robust to these recursive misspecifications but can become highly variable at long horizons. The surrogate estimators are designed to combine the strengths of both approaches. Short-run responses are estimated directly, allowing the estimator to absorb short-run misspecification arising from omitted latent factors. Beyond the surrogate horizon, the responses are propagated recursively using forecasting relationships estimated on the selected surrogate variables. The recursive component therefore exploits the approximate low-dimensional structure that emerges once the observed variables contain sufficient information about the relevant latent state dynamics.

We follow the implementation of Li, Plagborg-Møller and Wolf (2024) and simulate $S = 5000$ datasets of sample size $T = 200$ for 100 different model specifications for each shock type, which can be a monetary or fiscal policy shock, using 30 different random seed. Each data generating process, 6000 ($100 \times 2 \times 30$) has five variables that we denote by $\mathbf{y}_t$. We consider both the case where the shock is observed and the case where an instrument for the shock is available.

To implement the surrogate methods we use the likelihood-ratio-based scoring rule with the AIC penalty as developed in Section 5 to select the surrogacy horizon s based on $\mathbf{y}_{t:t+s}$. Alternatively, we could also select $\mathbf{y}_t$ from the variables of the dynamic factor model, but the computational costs are prohibitive to do this in such large simulation study. Therefore we keep the same variables $\mathbf{y}_t$ and only select s, which also facilitates a more transparent comparison with the conventional VAR.

For each design we compare the conventional LP estimator and a conventional VAR with lag length selected by AIC to their surrogate counterparts. For each horizon we compute the average bias and variance (across the 5000 draws). We report the median of the average bias and variance across the 6000 different data generating processes after normalizing by the L_2 norm of the true impulse response.

The results are shown in Figure 5. The top panel shows the results for bias and the bottom the variance results. In terms of bias we find that the surrogate LP is generally very close to the bias of the conventional LP. Around horizons 4-10 the bias is slightly larger whereas around horizons 12-17 the bias is somewhat lower. The surrogate VAR shows a bit more bias relative to the conventional LP, but remains comfortably away from the bias of the VAR. In terms of variance the both the surrogate LP and VAR dominate the LP from horizon 8 onward. The persistence in this DGP is one of the main reasons that the variance gains start later when compared to the VARMA simulation presented above.

7 Conclusion

In this paper we proposed a new approach to impulse response estimation based on the idea of approximate surrogacy. The key insight is that, although the short-run transmission of structural shocks may be dynamically complex and difficult to model recursively, future outcomes may become predictable from intermediate outcomes observed shortly after the shock. This perspective allows us to decompose impulse response estimation into a short-run causal problem and a long-run prediction problem. Building on this decomposition, we introduced Surrogate LP and Surrogate VAR estimators that preserve the robustness advantages of local projections during the initial transmission phase while exploiting additional predictive structure to improve long-horizon efficiency. We showed that the approximate surrogacy condition naturally arises in state-space and VARMA environments, developed data-driven procedures for selecting surrogate variables and horizons, and demonstrated in simulations and empirical applications that the proposed methods can substantially reduce long-horizon variance without inheriting the large misspecification biases often associated with conventional VARs.

Several directions for future research appear promising. First, the present paper focuses

on two specific surrogate-based estimators built around direct forecasting and autoregressive propagation. More generally, the surrogacy framework opens the door to a much broader class of propagation methods. For example, factor models, sparse high-dimensional forecasting methods, machine learning approaches, or nonlinear state-space representations may provide even more effective ways to summarize the post-surrogacy dynamics of the economy. Second, while our empirical applications focused on macroeconomic impulse responses, the underlying idea is much more general. Many dynamic systems outside macroeconomics feature complex short-run transmission mechanisms together with lower-dimensional long-run propagation. Potential applications therefore include finance, labor economics, epidemiology, industrial organization, climate economics, and other settings in which intermediate outcomes may summarize the relevant state of the system after an initial shock or intervention.

References

Al-Sadoon, Majid M. 2014. "Geometric and long run aspects of Granger causality." *Journal of Econometrics*, 178: 558–568.

Athey, Susan, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 2025. "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely." *The Review of Economic Studies*, rdaf087.

Brugnolini, Luca. 2018. "About local projection impulse response function reliability."

Canova, Fabio, and Filippo Ferroni. 2022. "Mind the gap! Stylized dynamic facts and structural models." *American Economic Journal: Macroeconomics*, 14(4): 104–135.

Chen, Jiafeng, and David M. Ritzwoller. 2023. "Semiparametric estimation of long-term treatment effects." *Journal of Econometrics*, 237(2, Part A): 105545.

Durbin, J., and S. J. Koopman. 2012. *Time Series Analysis by State Space Methods; 2nd Edition.* Oxford:Oxford University Press.

Fan, Yanqin, Carlos A. Manzanares, Hyeonseok Park, and Yuan Qi. 2026. "A Sensitivity Analysis of the Surrogate Index Approach for Estimating Long-Term Treatment Effects."

Fernández-Villaverde, Jesús, Juan F Rubio-Ramírez, Thomas J Sargent, and Mark W Watson. 2007. "ABCs (and Ds) of understanding VARs." *American economic review*, 97(3): 1021–1026.

Forni, Mario, and Luca Gambetti. 2014. "Sufficient information in structural VARs." *Journal of Monetary Economics*, 66: 124–136.

González-Casasús, Oriol, and Frank Schorfheide. 2025*a*. "Misspecification-Robust Shrinkage and Selection for VAR Forecasts and IRFs." National Bureau of Economic Research Working Paper 33474.

González-Casasús, Oriol, and Frank Schorfheide. 2025*b*. "Misspecification-robust shrinkage and selection for var forecasts and irfs." National Bureau of Economic Research.

Granger, C. W. J. 1969. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods." *Econometrica*, 37(3): 424–438.

Hernan, Miguel A., and James M. Robins. 2025. *Causal Inference: What If.* Boca Raton:CRC Press.

Imai, K., L. Keele, and D. Tingley. 2010. "A general approach to causal mediation analysis." *Psychological methods*, 15(4): 309–334.

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

Kallus, Nathan, and Xiaojie Mao. 2025. "On the role of surrogates in the efficient estimation of treatment effects with limited outcome data." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2): 480–509.

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

Lewis, Richard A., and Gregory C. Reinsel. 1985. "Prediction of Multivariate Time Series by Autoregressive Model Fitting." *Journal of Multivariate Analysis*, 16(3): 393–411.

Li, Dake, Mikkel Plagborg-Møller, and Christian K Wolf. 2024. "Local projections vs. vars: Lessons from thousands of dgps." *Journal of Econometrics*, 244(2): 105722.

Lippi, Marco, and Lucrezia Reichlin. 1993. "The dynamic effects of aggregate demand and supply disturbances: Comment." *The American Economic Review*, 83(3): 644–652.

Lütkepohl, Helmut, and D. S. Poskitt. 1991. "Estimating Orthogonal Impulse Responses via Vector Autoregressive Models." *Econometric Theory*, 7(4): 487–496.

Montiel Olea, José Luis, and Mikkel Plagborg-Møller. 2021. "Local Projection Inference is Simpler and More Robust Than You Think." *Econometrica*, 89(4): 1789–1823.

Müller, Ulrich K., and James H. Stock. 2011. "Forecasts in a Slightly Misspecified Finite Order VAR." working paper.

Newey, Whitney K. 1985. "Maximum Likelihood Specification Testing and Conditional Moment Tests." *Econometrica*, 53(5): 1047–1070.

Olea, José Luis Montiel, Mikkel Plagborg-Møller, Eric Qian, and Christian K Wolf. 2024. "Double Robustness of Local Projections and Some Unpleasant VARithmetic." National Bureau of Economic Research.

Olea, José Luis Montiel, Mikkel Plagborg-Møller, Eric Qian, and Christian K. Wolf. 2026. "Double Robustness of Local Projections and Some Unpleasant VARithmetic." *Econometrica*. forthcoming.

Olea, José Luis Montiel, Mikkel Plagborg-Moller, Eric Qian, and Christian) Wolf. 2025. "Local Projections or VARs? A Primer for Macroeconomists." Working Paper.

Pearl, Judea. 2001. "Direct and indirect effects." *UAI'01*, 411–420. San Francisco, CA, USA:Morgan Kaufmann Publishers Inc.

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

Prentice, Ross L. 1989. "Surrogate endpoints in clinical trials: definition and operational criteria." *Statistics in medicine*, 8(4): 431–440.

Schorfheide, Frank. 2005. "VAR Forecasting under Misspecification." *Journal of Econometrics*, 128(1): 99–136.

Sims, Christopher A. 1972. "Money, Income, and Causality." *The American Economic Review*, 62(4): 540–552.

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

Staiger, Douglas, and James H. Stock. 1997. "Instrumental variables regression with weak instruments." *Econometrica*, 65: 557–586.

White, Halbert L. Jr. 2000. *Asymptotic Theory for Econometrians - 2nd edition.* San Diego, California:Academic Press.

Appendix A: Other examples of Surrogacy

We briefly discuss a few other example for the surrogacy condition in cross-sectional designs.

Example 2: Education and long-run wages

Consider the effect of education on wages. A central omitted variable in regressions of wages on education is ability. Ability affects educational attainment and also affects future earnings, so a regression of wages on education suffers from omitted-variable bias. If years of education are randomized, then causal effect of education can be identified in an experimental sample. However, long-run wages, say ten years after schooling, may be unobserved. In that case, one can use surrogacy to infer the long-run effect from short-run wage outcomes.

For instance, the early income path could act as surrogates for the long-run income path: if the early wage path is sufficient for predicting long-run income, it means that the early wage path reveals ability or other persistent determinants of later earnings, and a predictive regression can be used to estimate the long-run effect of education on wages from short-run effects.[6]

Conversely, the assumption would fail if education had a delayed effect on wages that was not visible in the early wage path. For example, education might generate credentials, networks, or promotion opportunities that only pay off after several years. If those channels are not captured by the short-run wage response, then treatment status would still help predict long-run income after conditioning on early wages, and surrogacy would be invalid.

Example 3: Long-run effects of vaccination

As last example, consider the long-run effect offered by vaccination. Let (T) denote vaccination, let (S) denote short-run antibody levels, and let (Y) denote later protection against infection. The short-run antibody response may be generated by many biological mechanisms that are difficult to model. However, suppose that once the short-run antibody level is observed, the future path of protection is governed the level of antibodies, which itself follows a simple geometric decay process. Then, to estimate the long-run effect of vaccination, it is enough to (i) observe the short-run antibody response, and (ii)predict how antibody levels evolve over time, for instance with an AR structure.

Appendix B: proofs

Proof of Lemma 1. Fix a horizon $h > s$. By Assumption 1,

$$\text{Proj}(\mathbf{w}_{t+h} \mid \mathbf{y}_{t:t+s}, \mathbf{x}_{t-1}, \varepsilon_t) = \boldsymbol{\beta}_{hs}\mathbf{y}_{t:t+s} + \boldsymbol{\gamma}_{hs}\mathbf{x}_{t-1} + \boldsymbol{\xi}_t^{hs},$$

where $\|\boldsymbol{\xi}_t^{hs}\|_2 = O(T^{-a})$. Now project both sides onto $(\mathbf{x}_{t-1}, \varepsilon_t)$. By the tower property of linear projections,

$$\text{Proj}(\mathbf{w}_{t+h} \mid \mathbf{x}_{t-1}, \varepsilon_t) = \boldsymbol{\beta}_{hs}\text{Proj}(\mathbf{y}_{t:t+s} \mid \mathbf{x}_{t-1}, \varepsilon_t) + \boldsymbol{\gamma}_{hs}\mathbf{x}_{t-1} + \text{Proj}(\boldsymbol{\xi}_t^{hs} \mid \mathbf{x}_{t-1}, \varepsilon_t).$$

[6] Again, the prediction coefficient need not be causal. A regression of long-run wages on early wages may reflect ability, family background, occupation, geography, and other omitted variables. That is acceptable as the regression is used only as a predictive bridge.

Evaluating at $\varepsilon_t = 1$ and $\varepsilon_t = 0$, and subtracting, gives

$$\boldsymbol{\theta}_h = \boldsymbol{\beta}_{hs} \left[\text{Proj}(\mathbf{y}_{t:t+s} \mid \mathbf{x}_{t-1}, \varepsilon_t = 1) - \text{Proj}(\mathbf{y}_{t:t+s} \mid \mathbf{x}_{t-1}, \varepsilon_t = 0) \right] + \Delta_{\xi,t},$$

where

$$\Delta_{\xi,t} = \text{Proj}(\boldsymbol{\xi}_t^{hs} \mid \mathbf{x}_{t-1}, \varepsilon_t = 1) - \text{Proj}(\boldsymbol{\xi}_t^{hs} \mid \mathbf{x}_{t-1}, \varepsilon_t = 0).$$

The term involving $\mathbf{x}_{t-1}$ cancels because $\mathbf{x}_{t-1}$ is fixed in both counterfactual comparisons. By definition,

$$\text{Proj}(\mathbf{y}_{t:t+s} \mid \mathbf{x}_{t-1}, \varepsilon_t = 1) - \text{Proj}(\mathbf{y}_{t:t+s} \mid \mathbf{x}_{t-1}, \varepsilon_t = 0) = \boldsymbol{\tau}_{0:s}.$$

Finally, projections are contractions in L^2, so

$$\|\text{Proj}(\boldsymbol{\xi}_t^{hs} \mid \mathbf{x}_{t-1}, \varepsilon_t)\|_2 \leq \|\boldsymbol{\xi}_t^{hs}\|_2 = O(T^{-a}),$$

which implies $\Delta_{\xi,t} = O(T^{-a})$. Therefore, $\boldsymbol{\theta}_h = \boldsymbol{\beta}_{hs}\boldsymbol{\tau}_{0:s} + O(T^{-a})$ as claimed. □

Proof of Proposition 1. For $h > s$, repeated substitution of the VAR recursion gives

$$\mathbf{y}_{t+h} = \mathbf{B}_{hs}\mathbf{y}_{t:t+s} + \mathbf{r}_{t+h} \ , \qquad \mathbf{r}_{t+h} = \sum_{\ell=s+1}^{h} \mathbf{D}_{\ell h}\mathbf{u}_{t+\ell} \ ,$$

for suitable coefficient matrices $\mathbf{B}_{hs}$ and $\mathbf{D}_{\ell h}$. If the VAR innovations are serially uncorrelated, then

$$\text{Cov}(\mathbf{u}_{t+\ell}, \mathbf{u}_t) = \mathbf{0}, \qquad \ell = s+1, \ldots, h.$$

Since $\varepsilon_t = u_{1t}$ is one component of $\mathbf{u}_t$, it follows that

$$\text{Cov}(\mathbf{r}_{t+h}, \varepsilon_t) = \mathbf{0}.$$

Therefore ε_t is orthogonal to the residual from projecting $\mathbf{y}_{t+h}$ on $\mathbf{y}_{t:t+s}$, which implies

$$\text{Proj}(\mathbf{y}_{t+h} \mid \mathbf{y}_{t:t+s}, \varepsilon_t) = \text{Proj}(\mathbf{y}_{t+h} \mid \mathbf{y}_{t:t+s}).$$

The converse is not true. Consider the scalar process

$$y_t = u_t, \qquad u_t = \eta_t + c\eta_{t-1},$$

where η_t is serially uncorrelated with variance σ_η^2 and $c \neq 0$. Let $\varepsilon_t = u_t$ and set $s = 0$. Since $y_t = u_t = \varepsilon_t$, the shock is already contained in the surrogate variable, so that

$$\text{Proj}(y_{t+h} \mid y_t, \varepsilon_t) = \text{Proj}(y_{t+h} \mid y_t), \qquad h > 0.$$

Thus the surrogacy projection restriction holds exactly. However, $\text{Cov}(u_t, u_{t-1}) = c\sigma_\eta^2 \neq 0$, so the VAR innovation is serially correlated. Hence the converse implication fails. □

Proof of Proposition 2. Let

$$\boldsymbol{\psi}_h = \begin{pmatrix} \text{vec}(\boldsymbol{\beta}_{hs}) \\ \boldsymbol{\tau}_{0:s} \end{pmatrix}, \qquad \hat{\boldsymbol{\psi}}_h = \begin{pmatrix} \text{vec}\,\hat{\boldsymbol{\beta}}_{hs}) \\ \hat{\boldsymbol{\tau}}_{0:s} \end{pmatrix}.$$

Under Assumption 3, the least-squares estimators in (7) and (6) satisfy the regularity conditions for the least-squares central limit theorem for mixing observations; see e.g. Exercise 5.21 in White (2000). Applied to the stacked system, this implies

$$\sqrt{T}\left(\hat{\boldsymbol{\psi}}_h - \boldsymbol{\psi}_h\right) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_h)\ ,$$

where $\boldsymbol{\Sigma}_h = \mathbf{A}_h^{-1}\mathbf{V}_h\mathbf{A}_h^{-1}$, $\mathbf{V}_h$ is the long-run covariance matrix of the stacked score $\mathbf{q}_{ht}$ defined in Assumption 3 part 3, and

$$\mathbf{A}_h = \begin{pmatrix} \mathbf{M}_1 \otimes I_{d_w} & 0 \\ 0 & I_{s+1} \otimes \mathbf{M}_2 \otimes I_{d_y} \end{pmatrix}\ . \tag{12}$$

Define $g(\boldsymbol{\psi}_h) = \boldsymbol{\beta}_{hs}\boldsymbol{\tau}_{0:s}$. By Lemma 1, $\boldsymbol{\theta}_h = g(\boldsymbol{\psi}_h) + O(T^{-a})$ for some $a > 1/2$, while $\hat{\boldsymbol{\theta}}_h^{\text{sur}} = g(\hat{\boldsymbol{\psi}}_h)$. Hence,

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_h^{\text{sur}} - \boldsymbol{\theta}_h\right) = \sqrt{T}\left(g(\hat{\boldsymbol{\psi}}_h) - g(\boldsymbol{\psi}_h)\right) + o(1),$$

because $\sqrt{T}O(T^{-a}) = o(1)$. Since g is continuously differentiable, a first-order Taylor expansion gives

$$\sqrt{T}\left(g(\hat{\boldsymbol{\psi}}_h) - g(\boldsymbol{\psi}_h)\right) = \mathbf{D}_h\sqrt{T}\left(\hat{\boldsymbol{\psi}}_h - \boldsymbol{\psi}_h\right) + o_p(1),$$

where $\mathbf{D}_h = (\boldsymbol{\tau}_{0:s}' \otimes I_{d_w}, \boldsymbol{\beta}_{hs})$. Therefore,

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_h^{\text{sur}} - \boldsymbol{\theta}_h\right) \xrightarrow{d} N(0, \Omega_h^{\text{sur}}), \qquad \Omega_h^{\text{sur}} = \mathbf{D}_h\boldsymbol{\Sigma}_h\mathbf{D}_h',$$

which proves the result. □

Proof of Proposition ??. By the Frisch–Waugh–Lovell theorem, we may work with variables residualized with respect to $\mathbf{x}_{t-1}$. To simplify notation, write these residuals as $\mathbf{w}_{t+h}$, $\mathbf{y}_{t:t+s}$, and ε_t throughout the proof. Then the relevant population projections are

$$\mathbf{w}_{t+h} = \boldsymbol{\beta}_{hs}\mathbf{y}_{t:t+s} + \boldsymbol{\eta}_{t+h}, \qquad \text{and} \qquad \mathbf{y}_{t:t+s} = \boldsymbol{\tau}_{0:s}\varepsilon_t + \boldsymbol{\zeta}_{t:t+s}\ .$$

By Assumption 3 part 2, $\mathbb{E}[\mathbf{y}_{t:t+s}\boldsymbol{\eta}_{t+h}'] = 0$ and $\mathbb{E}[\varepsilon_t\boldsymbol{\zeta}_{t:t+s}'] = 0$. Assumption 1 further implies $\mathbb{E}[\varepsilon_t\boldsymbol{\eta}_{t+h}'] = O(T^{-a})$, and therefore $\mathbb{E}[\boldsymbol{\zeta}_{t:t+s}\boldsymbol{\eta}_{t+h}'] = O(T^{-a})$. Since $a > 1/2$, these terms are negligible at the $\sqrt{T}$ rate. Substituting the second projection into the first gives

$$\mathbf{w}_{t+h} = \boldsymbol{\beta}_{hs}\boldsymbol{\tau}_{0:s}\varepsilon_t + \boldsymbol{\beta}_{hs}\boldsymbol{\zeta}_{t:t+s} + \boldsymbol{\eta}_{t+h}\ .$$

Since Assumption 1 implies $\boldsymbol{\theta}_h = \boldsymbol{\beta}_{hs}\boldsymbol{\tau}_{0:s} + O(T^{-a})$, the conventional LP estimator has influence function

$$Q_\varepsilon^{-1}\varepsilon_t\left(\boldsymbol{\beta}_{hs}\boldsymbol{\zeta}_{t:t+s} + \boldsymbol{\eta}_{t+h}\right)$$

up to terms that are $o_p(1)$ after summation and multiplication by $T^{-1/2}$. For the surrogate estimator, the first-stage contribution from estimating $\boldsymbol{\tau}_{0:s}$ has influence function $Q_\varepsilon^{-1}\varepsilon_t\boldsymbol{\beta}_{hs}\boldsymbol{\zeta}_{t:t+s}$, while the second-stage contribution from estimating $\boldsymbol{\beta}_{hs}$ has influence function $\boldsymbol{\eta}_{t+h}\mathbf{y}'_{t:t+s}\mathbf{M}_S^{-1}\boldsymbol{\tau}_{0:s}$. Hence the surrogate influence function is

$$Q_\varepsilon^{-1}\varepsilon_t\boldsymbol{\beta}_{hs}\boldsymbol{\zeta}_{t:t+s} + \boldsymbol{\eta}_{t+h}\mathbf{y}'_{t:t+s}\mathbf{M}_S^{-1}\boldsymbol{\tau}_{0:s}.$$

The $\boldsymbol{\zeta}_{t:t+s}$ contribution is therefore the same in the conventional LP and surrogate influence functions. The only difference is the contribution of $\boldsymbol{\eta}_{t+h}$: the conventional LP weights it by $Q_\varepsilon^{-1}\varepsilon_t$, whereas the surrogate estimator weights it by $\mathbf{y}'_{t:t+s}\mathbf{M}_S^{-1}\boldsymbol{\tau}_{0:s}$. Therefore the difference between the asymptotic variance of the conventional LP estimator and the asymptotic variance of the surrogate estimator is exactly

$$\sum_{\ell=-\infty}^{\infty}\mathbb{E}\left[\boldsymbol{\eta}_{t+h}\boldsymbol{\eta}'_{t-\ell+h}\left\{Q_\varepsilon^{-1}\varepsilon_t\varepsilon_{t-\ell}Q_\varepsilon^{-1} - \mathbf{y}'_{t:t+s}\mathbf{M}_S^{-1}\boldsymbol{\tau}_{0:s}\boldsymbol{\tau}'_{0:s}\mathbf{M}_S^{-1}\mathbf{y}_{t-\ell:t-\ell+s}\right\}\right],$$

again up to terms that are asymptotically negligible because $a > 1/2$. By condition (**??**), this matrix is positive semidefinite. Hence $\mathrm{Avar}(\hat{\boldsymbol{\theta}}_h^{LP}) - \mathrm{Avar}(\hat{\boldsymbol{\theta}}_h^{\mathrm{sur}}) \succeq 0$. □

Proof of Lemma 2. Let

$$\mathbf{u}_t^A = \mathbf{y}_t - \mathbf{A}_1\mathbf{y}_{t-1} - \cdots - \mathbf{A}_p\mathbf{y}_{t-p}.$$

Assumption 2 implies that, for all horizons $h > s$, the VAR innovation $\mathbf{u}_t^A$ has no additional linear predictive content for $\mathbf{y}_{t+h}$ once we condition on the surrogate VAR state. Equivalently, the linear projection coefficient on $\mathbf{u}_t^A$ in the projection of $\mathbf{y}_{t+h}$ on $(\mathbf{y}_{t+h-p:t+h-1}, \mathbf{u}_t^A)$ is $O(T^{-a})$.

Now replace t by $t-h$. Then the same restriction implies that the linear projection coefficient on $\mathbf{u}_t^A$ in the projection of $\mathbf{y}_t$ on $(\mathbf{y}_{t-p:t-1}, \mathbf{u}_{t-h}^A)$ is $O(T^{-a})$. Equivalently, after partialling out $\mathbf{y}_{t-p:t-1}$, the covariance between $\mathbf{u}_t^A$ and the lagged variables $\mathbf{y}_{t-h-p:t-h}$ is of order $O(T^{-a})$. Hence

$$\mathbb{E}\left[\mathbf{y}_{t-h-p:t-h}\mathbf{u}_t^{A\prime}\right] = O(T^{-a}).$$

Substituting the definition of $\mathbf{u}_t^A$ gives the desired moment condition. □

Proof of Proposition 3. Let $\hat{\boldsymbol{\psi}}^{\mathrm{V}}$ stack the GMM estimator of $(\mathbf{A}_1, \ldots, \mathbf{A}_p)$ and the least-squares estimators of $(\boldsymbol{\tau}_0, \ldots, \boldsymbol{\tau}_s)$. We first establish the joint limiting distribution of $\hat{\boldsymbol{\psi}}^{\mathrm{V}}$.

By Lemma 2, the propagation coefficients satisfy $\mathbb{E}[\mathbf{Z}_t\mathbf{u}_t^{A\prime}] = O(T^{-a})$ with $a > 1/2$. Therefore the local violation of the IV moment condition is negligible at the $\sqrt{T}$ rate. Together with the mixing, moment, rank, and weighting conditions in Assumption 4, Theorem 5.23 in White implies the asymptotic normality of the IV/GMM estimator of $(\mathbf{A}_1, \ldots, \mathbf{A}_p)$. The same assumptions also imply the joint asymptotic normality of the least-squares estimators $(\hat{\boldsymbol{\tau}}_0, \ldots, \hat{\boldsymbol{\tau}}_s)$. Since the corresponding IV and least-squares scores are stacked in $\mathbf{q}_t^{\mathrm{V}}$, the joint CLT gives

$$\sqrt{T}\left(\hat{\boldsymbol{\psi}}^{\mathrm{V}} - \boldsymbol{\psi}^{\mathrm{V}}\right) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_h^{\mathrm{V}}).$$

By construction of the surrogate VAR recursion, $\hat{\boldsymbol{\theta}}_h^{\mathrm{surV}} = g_h^{\mathrm{V}}(\hat{\boldsymbol{\psi}}^{\mathrm{V}})$. Moreover, Assumption 2 implies that the population impulse response satisfies $\boldsymbol{\theta}_h = g_h^{\mathrm{V}}(\boldsymbol{\psi}^{\mathrm{V}}) + O(T^{-a})$. Since

$a > 1/2$, $\sqrt{T}O(T^{-a}) = o(1)$, and therefore

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_h^{\mathrm{surV}} - \boldsymbol{\theta}_h\right) = \sqrt{T}\left(g_h^{\mathrm{V}}(\hat{\boldsymbol{\psi}}^{\mathrm{V}}) - g_h^{\mathrm{V}}(\boldsymbol{\psi}^{\mathrm{V}})\right) + o_p(1).$$

The map g_h^{V} is continuously differentiable in a neighborhood of the population parameters because, for fixed h, it is generated by finitely many matrix additions and multiplications. Hence a first-order Taylor expansion around $\boldsymbol{\psi}^{\mathrm{V}}$ gives

$$\sqrt{T}\left(g_h^{\mathrm{V}}(\hat{\boldsymbol{\psi}}^{\mathrm{V}}) - g_h^{\mathrm{V}}(\boldsymbol{\psi}^{\mathrm{V}})\right) = \mathbf{D}_h^{\mathrm{V}}\sqrt{T}\left(\hat{\boldsymbol{\psi}}^{\mathrm{V}} - \boldsymbol{\psi}^{\mathrm{V}}\right) + o_p(1),$$

where $\mathbf{D}_h^{\mathrm{V}} = \partial g_h^{\mathrm{V}}(\boldsymbol{\psi}^{\mathrm{V}})/\partial \boldsymbol{\psi}^{\mathrm{V}\prime}$ is evaluated at the population parameters. Combining this expansion with the joint CLT and applying the continuous mapping theorem yields

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_h^{\mathrm{surV}} - \boldsymbol{\theta}_h\right) \xrightarrow{d} N(0, \Omega_h^{\mathrm{surV}}),$$

with $\Omega_h^{\mathrm{surV}} = \mathbf{D}_h^{\mathrm{V}}\boldsymbol{\Sigma}_h^{\mathrm{V}}\mathbf{D}_h^{\mathrm{V}\prime}$. □

Proof of Proposition 4. Fix a candidate surrogate $(s, n) \in \mathcal{S}$ and a horizon $h = s+1, \ldots, H$. By Assumption 5, the sample restricted and unrestricted least-squares problems converge uniformly to their population counterparts. Therefore,

$$T_h^{-1}RSS_R(s, n, h) = \sigma_R^2(s, n, h) + o_p(1), \qquad T_h^{-1}RSS_U(s, n, h) = \sigma_U^2(s, n, h) + o_p(1),$$

uniformly over the finite set of candidates and horizons.

First consider an invalid candidate (s, n). By Assumption 5, there exists at least one horizon $h > s$ such that $\sigma_R^2(s, n, h) > \sigma_U^2(s, n, h)$. Hence

$$\log\left(\frac{RSS_R(s, n, h)}{RSS_U(s, n, h)}\right) = \log\left(\frac{\sigma_R^2(s, n, h)}{\sigma_U^2(s, n, h)}\right) + o_p(1),$$

where the population log-ratio is strictly positive. It follows that

$$LR_h(s, n) = T_h \log\left(\frac{RSS_R(s, n, h)}{RSS_U(s, n, h)}\right) = T_h\Delta_h(s, n) + o_p(T),$$

for some $\Delta_h(s, n) > 0$. Therefore

$$LR_{\max}(s, n) \geq LR_h(s, n) \to_p \infty$$

at rate T. Since $c_T(s, n) = o(T)$ uniformly over $\mathcal{S}$,

$$P\left(LR_{\max}(s, n) \leq c_T(s, n)\right) \to 0.$$

Thus every invalid candidate is excluded with probability approaching one.

Now consider a valid candidate (s, n). By Assumption 1, after conditioning on $\mathbf{y}_{t:t+s}$ and $\mathbf{x}_{t-1}$, the remaining linear predictive content of ε_t for $w_{i,t+h}$ is of order $O(T^{-a})$ for each $h = s+1, \ldots, H$, with $a > 1/2$. Equivalently, the population reduction in mean squared

forecast error from adding ε_t is of order $O(T^{-2a})$. Hence

$$\sigma_R^2(s,n,h) - \sigma_U^2(s,n,h) = O(T^{-2a})$$

uniformly over post-surrogacy horizons. Since $a > 1/2$,

$$T_h \left\{ \sigma_R^2(s,n,h) - \sigma_U^2(s,n,h) \right\} = o(1).$$

Together with the uniform convergence of the sample least-squares criterion, this implies

$$LR_h(s,n) = O_p(1)$$

for each $h = s+1, \ldots, H$. Since the number of horizons is fixed,

$$LR_{\max}(s,n) = \max_{h=s+1,\ldots,H} LR_h(s,n) = O_p(1).$$

Because $c_T(s,n) \to \infty$ uniformly over $\mathcal{S}$,

$$P\left(LR_{\max}(s,n) \leq c_T(s,n)\right) \to 1.$$

Thus every valid candidate is retained with probability approaching one.

Since $\mathcal{S}$ is finite, the preceding statements hold jointly over all candidate surrogates. With probability approaching one, all invalid candidates fail the selection inequality and all valid candidates satisfy it. Therefore the minimization over model dimension selects the smallest valid candidate. By assumption this candidate is unique and equal to (s_0, n_0). Hence

$$P\left((\hat{s}, \hat{n}) = (s_0, n_0)\right) \to 1,$$

as required. □

Appendix C: details surrogate estimators and inference

In this appendix we discuss the regularity conditions for the surrogate estimators, provide implementation details and present the versions for the case where the shock is not observed but a proxy is available.

Details surrogate estimation with observed shocks

Surrogate LP Recall that the regression equations for the short-run causal effect (6) and (7) and the long-run predictive coefficients are given by

$$\begin{aligned} \mathbf{w}_{t+h} &= \boldsymbol{\beta}_{hs}\mathbf{y}_{t:t+s} + \boldsymbol{\gamma}_{hs}\mathbf{x}_{t-1} + \boldsymbol{\eta}_{t+h} \\ \mathbf{y}_{t+j} &= \boldsymbol{\tau}_j\varepsilon_t + \mathbf{B}_j\mathbf{x}_{t-1} + \boldsymbol{\zeta}_{t+j}, \qquad j = 0, \ldots, s. \end{aligned}$$

The coefficients from these regressions are compared to give the surrogate LP estimator as defined in (9). To describe the regularity conditions needed for Proposition 2 we define the

regression vectors

$$\mathbf{r}_{1t} = \begin{pmatrix} \mathbf{y}_{t:t+s} \\ \mathbf{x}_{t-1} \end{pmatrix}, \qquad \mathbf{r}_{2t} = \begin{pmatrix} \varepsilon_t \\ \mathbf{x}_{t-1} \end{pmatrix}.$$

The following regularity conditions are analogous to White (2000, Section 5.4).

Assumption 3. Fix $h > s$. We assume that:

1. The process $\{\mathbf{r}_{1t}, \mathbf{r}_{2t}, \boldsymbol{\eta}_{t+h}, \boldsymbol{\zeta}_t, \ldots, \boldsymbol{\zeta}_{t+s}\}_{t\in\mathbb{Z}}$ is strictly stationary and mixing, with either ϕ-mixing coefficients of size $-r/2(r-1)$ for some $r > 2$, or α-mixing coefficients of size $-r/(r-2)$ for some $r > 2$.
2. $\mathbb{E}[\mathbf{r}_{1t}\boldsymbol{\eta}'_{t+h}] = 0$ and $\mathbb{E}[\mathbf{r}_{2t}\boldsymbol{\zeta}'_{t+j}] = 0$ for $j = 0, \ldots, s$.
3. Defining $\mathbf{q}_{ht} = (\mathrm{vec}(\mathbf{r}_{1t}\boldsymbol{\eta}'_{t+h})', \mathrm{vec}(\mathbf{r}_{2t}\boldsymbol{\zeta}'_t)', \ldots, \mathrm{vec}(\mathbf{r}_{2t}\boldsymbol{\zeta}'_{t+s})')'$, we have $\mathbb{E}|q_{ht,\ell}|^r < \Delta < \infty$ for every element $q_{ht,\ell}$ of $\mathbf{q}_{ht}$.
4. $\mathbf{V}_h = \sum_{\ell=-\infty}^{\infty} \mathbb{E}[\mathbf{q}_{ht}\mathbf{q}'_{h,t-\ell}]$ exists and is positive definite.
5. $\mathbf{M}_1 = \mathbb{E}[\mathbf{r}_{1t}\mathbf{r}'_{1t}]$ and $\mathbf{M}_2 = \mathbb{E}[\mathbf{r}_{2t}\mathbf{r}'_{2t}]$ exist and are positive definite.
6. For some $\delta > 0$, $\mathbb{E}|r_{1t,a}r_{1t,b}|^{r/2+\delta} < \Delta < \infty$ and $\mathbb{E}|r_{2t,a}r_{2t,b}|^{r/2+\delta} < \Delta < \infty$ for all elements $r_{1t,a}, r_{1t,b}$ of $\mathbf{r}_{1t}$ and $r_{2t,a}, r_{2t,b}$ of $\mathbf{r}_{2t}$.
7. There exists $\widehat{\mathbf{V}}_h$ such that $\widehat{\mathbf{V}}_h - \mathbf{V}_h = o_p(1)$.

Assumption 3 collects standard regularity conditions for establishing asymptotic normality of least squares estimators in dependent time series regressions; see White (2000, Section 5.4). The assumptions require that the regression vectors and disturbances are stationary and weakly dependent, satisfy the standard orthogonality conditions implied by linear projections, and possess sufficiently many finite moments. In addition, the long-run covariance matrix and regressor second-moment matrices are assumed to exist and be positive definite to ensure identification and nonsingular asymptotic variance matrices. Finally, we assume the availability of a consistent estimator for the long-run covariance matrix, allowing for feasible asymptotic inference using standard HAC methods.

Given these conditions and the main surrogacy assumption 1, Proposition 2 presents the limiting distribution for the surrogate estimates: $\sqrt{T}(\hat{\boldsymbol{\theta}}_h^{\mathrm{sur}} - \boldsymbol{\theta}_h) \xrightarrow{d} N(0, \Omega_h^{\mathrm{sur}})$ with $\Omega_h^{\mathrm{sur}} = \mathbf{D}_h\boldsymbol{\Sigma}_h\mathbf{D}'_h$, $\boldsymbol{\Sigma}_h = \mathbf{A}_h^{-1}\mathbf{V}_h\mathbf{A}_h^{-1}$ and $\mathbf{A}_h$ is defined (12) and $\mathbf{V}_h$ in Assumption 3 part 4. This asymptotic variance can be estimated using

$$\widehat{\Omega}_h^{\mathrm{sur}} = \widehat{\mathbf{D}}_h\widehat{\mathbf{A}}_h^{-1}\widehat{\mathbf{V}}_h\widehat{\mathbf{A}}_h^{-1}\widehat{\mathbf{D}}'_h\,, \tag{13}$$

where $\widehat{\mathbf{V}}_h$ can be any HAC estimator that satisfies Assumption 3 part 7 and

$$\widehat{\mathbf{D}}_h = (\widehat{\boldsymbol{\tau}}'_{0:s} \otimes I_{d_w} \ \hat{\boldsymbol{\beta}}_{hs})$$
$$\widehat{\mathbf{A}}_h = \begin{pmatrix} \widehat{\mathbf{M}}_1 \otimes I_{d_w} & 0 \\ 0 & I_{s+1} \otimes \widehat{\mathbf{M}}_2 \otimes I_{d_y} \end{pmatrix}$$
$$\widehat{\mathbf{M}}_1 = \frac{1}{T}\sum_{t=1}^{T} \mathbf{r}_{1t}\mathbf{r}'_{1t}$$
$$\widehat{\mathbf{M}}_2 = \frac{1}{T}\sum_{t=1}^{T} \mathbf{r}_{2t}\mathbf{r}'_{2t}$$

In practice, we recommend for $h > s$ reporting $\hat{\boldsymbol{\theta}}_h^{\text{sur}}$ and confidence bands based on $\widehat{\Omega}_h^{\text{sur}}$. For the initial horizons $h = 1, \ldots, s$ conventional local projection inference can be used, see the recommendations in Olea et al. (2025).

Surrogate VAR - details Next we provide the details for the surrogate VAR estimator that is defined in equation (11). Since the propagation coefficients are estimated using instrumental variables we require a different set of regularity conditions.

Assumption 4 (Regularity conditions for the surrogate VAR CLT)**.** Let $\mathbf{X}_{t-1}^p = (\mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p})'$ and let $\mathbf{Z}_t = (\mathbf{y}'_{t-\ell_1}, \ldots, \mathbf{y}'_{t-\ell_L})'$ collect valid instruments with $\ell_j > s + p$. We assume that:

1. The process $\{\mathbf{Z}_t, \mathbf{X}_{t-1}^p, \mathbf{u}_t^A, \mathbf{r}_{2t}, \boldsymbol{\zeta}_t, \ldots, \boldsymbol{\zeta}_{t+s}\}_{t\in\mathbb{Z}}$ is strictly stationary and mixing, with either ϕ-mixing coefficients of size $-r/2(r-1)$ for some $r > 2$, or α-mixing coefficients of size $-r/(r-2)$ for some $r > 2$.

2. $\mathbb{E}[\mathbf{r}_{2t}\boldsymbol{\zeta}'_{t+j}] = 0$ for $j = 0, \ldots, s$.

3. Defining $\mathbf{q}_t^{\text{V}} = (\text{vec}(\mathbf{Z}_t\mathbf{u}_t^{A\prime})', \text{vec}(\mathbf{r}_{2t}\boldsymbol{\zeta}'_t)', \ldots, \text{vec}(\mathbf{R}_{2t}\boldsymbol{\zeta}'_{t+s})')'$, we have $\mathbb{E}|q_{t,\ell}^{\text{V}}|^r < \Delta < \infty$ for every element $q_{t,\ell}^{\text{V}}$ of $\mathbf{q}_t^{\text{V}}$.

4. $\mathbf{V}^{\text{V}} = \sum_{\ell=-\infty}^{\infty} \mathbb{E}[\mathbf{q}_t^{\text{V}}\mathbf{q}_{t-\ell}^{\text{V}\prime}]$ exists and is positive definite.

5. $\mathbf{Q}_{ZX} = \mathbb{E}[\mathbf{Z}_t\mathbf{X}_{t-1}^{p\prime}]$ has full column rank and $\mathbf{M}_2 = \mathbb{E}[\mathbf{r}_{2t}\mathbf{r}'_{2t}]$ exists and is positive definite.

6. For some $\delta > 0$, $\mathbb{E}|Z_{t,a}X_{t-1,b}^p|^{r/2+\delta} < \Delta < \infty$ and $\mathbb{E}|r_{2t,a}r_{2t,b}|^{r/2+\delta} < \Delta < \infty$ for all relevant elements.

7. There exists $\widehat{\mathbf{V}}^V$ such that $\widehat{\mathbf{V}}^V - \mathbf{V}^V = o_p(1)$.

8. $\mathbf{W}_T - \mathbf{W} = o_p(1)$, where $\mathbf{W}$ is symmetric positive definite.

The assumptions are similar as for the surrogate LP, with the exception that we know require the lagged observables to be valid instruments. That is the exogeneity condition already imposed in Assumption 2 together with the relevance conditions impose above in part 5.

Given these conditions, Proposition 3 presents the limiting distribution of the surrogate VAR estimator $\sqrt{T}\left(\hat{\boldsymbol{\theta}}_h^{\text{surV}} - \boldsymbol{\theta}_h\right) \xrightarrow{d} N(0, \Omega_h^{\text{surV}})$ with $\Omega_h^{\text{surV}} = \mathbf{D}_h^V \boldsymbol{\Sigma}_h^V \mathbf{D}_h^{V\prime}$. The asymptotic variance can be estimated by replacing the population moments by their sample equivalents.

Details surrogate estimation with instruments

When the shock ε_t is not observed and we only have an instrument available, say $\hat{\varepsilon}_t$, the short run causal effects can be estimated by LP-IV instead of LP. For this we need to define a normalizing variable on which the structural shock has a unit effect. For example, when the shock is a monetary policy shock the normalizing variable is often the interest rate. In general, let this variable be denoted by p_t and consider the regression

$$\mathbf{y}_{t+h} = \boldsymbol{\tau}_h p_t + \mathbf{B}_h \mathbf{x}_{t-1} + \boldsymbol{\zeta}_{t+h} \qquad h = 0, \ldots, s \ , \tag{14}$$

where we slightly abuse notation by using the same as in equation (6). The coefficients $\boldsymbol{\tau}_h$ can be estimated using LP-IV, say $\hat{\boldsymbol{\tau}}_h^{iv}$, and conventional inference methods apply. The computation of the long run surrogacy estimates then only requires changing $\hat{\boldsymbol{\tau}}_{0:s}$ to $\hat{\boldsymbol{\tau}}_{0:s}^{iv}$ in (9) and (11).

Details surrogate selection

Assumption 5 (Regularity conditions for surrogate selection)**.** Let $\mathcal{S}$ be finite and fix a candidate $(s, n) \in \mathcal{S}$ and a horizon $h = s+1, \ldots, H$. Let

$$\mathbf{Z}_t(s,n) = \left(\mathbf{y}'_{t:t+s}, \mathbf{x}'_{t-1}\right)'$$

denote the restricted forecasting variables and let

$$\mathbf{Z}_t^U(s,n) = \left(\varepsilon_t, \mathbf{Z}_t(s,n)'\right)'$$

denote the unrestricted forecasting variables. We assume that:

1. The process $\{w_{i,t+h}, \varepsilon_t, \mathbf{y}_{t:t+s}, \mathbf{x}_{t-1}\}_{t\in\mathbb{Z}}$ is strictly stationary and mixing, with either ϕ-mixing coefficients of size $-r/2(r-1)$ for some $r > 2$, or α-mixing coefficients of size $-r/(r-2)$ for some $r > 2$.

2. For some $\delta > 0$,

$$\mathbb{E}|w_{i,t+h}|^{2+\delta} < \Delta < \infty, \qquad \mathbb{E}\|\mathbf{Z}_t^U(s,n)\|^{2+\delta} < \Delta < \infty,$$

   uniformly over $(s, n) \in \mathcal{S}$ and $h = s+1, \ldots, H$.

3. The population second-moment matrices

$$\mathbf{M}_R(s,n) = \mathbb{E}[\mathbf{Z}_t(s,n)\mathbf{Z}_t(s,n)'] \quad \text{and} \quad \mathbf{M}_U(s,n) = \mathbb{E}[\mathbf{Z}_t^U(s,n)\mathbf{Z}_t^U(s,n)']$$

   exist and are positive definite, uniformly over $(s, n) \in \mathcal{S}$.

4. The sample second moments converge uniformly to their population counterparts. In particular,

$$T_h^{-1} \sum_t \mathbf{Z}_t(s,n)\mathbf{Z}_t(s,n)' = \mathbf{M}_R(s,n) + o_p(1),$$

$$T_h^{-1} \sum_t \mathbf{Z}_t^U(s,n)\mathbf{Z}_t^U(s,n)' = \mathbf{M}_U(s,n) + o_p(1),$$

and the corresponding sample cross moments with $w_{i,t+h}$ converge uniformly over $(s,n) \in \mathcal{S}$ and $h = s+1, \ldots, H$.

5. For each (s,n) and h, the restricted and unrestricted population linear projections are well defined. Let

$$\sigma_R^2(s,n,h) = \min_b \mathbb{E}\left[\left(w_{i,t+h} - b'\mathbf{Z}_t(s,n)\right)^2\right],$$

and

$$\sigma_U^2(s,n,h) = \min_b \mathbb{E}\left[\left(w_{i,t+h} - b'\mathbf{Z}_t^U(s,n)\right)^2\right].$$

Define a candidate surrogate (s,n) to be valid if Assumption 1 holds for all horizons $h = s+1, \ldots, H$. Otherwise the candidate surrogate is invalid.

For every invalid candidate (s,n), there exists at least one horizon $h > s$ such that

$$\sigma_R^2(s,n,h) > \sigma_U^2(s,n,h).$$

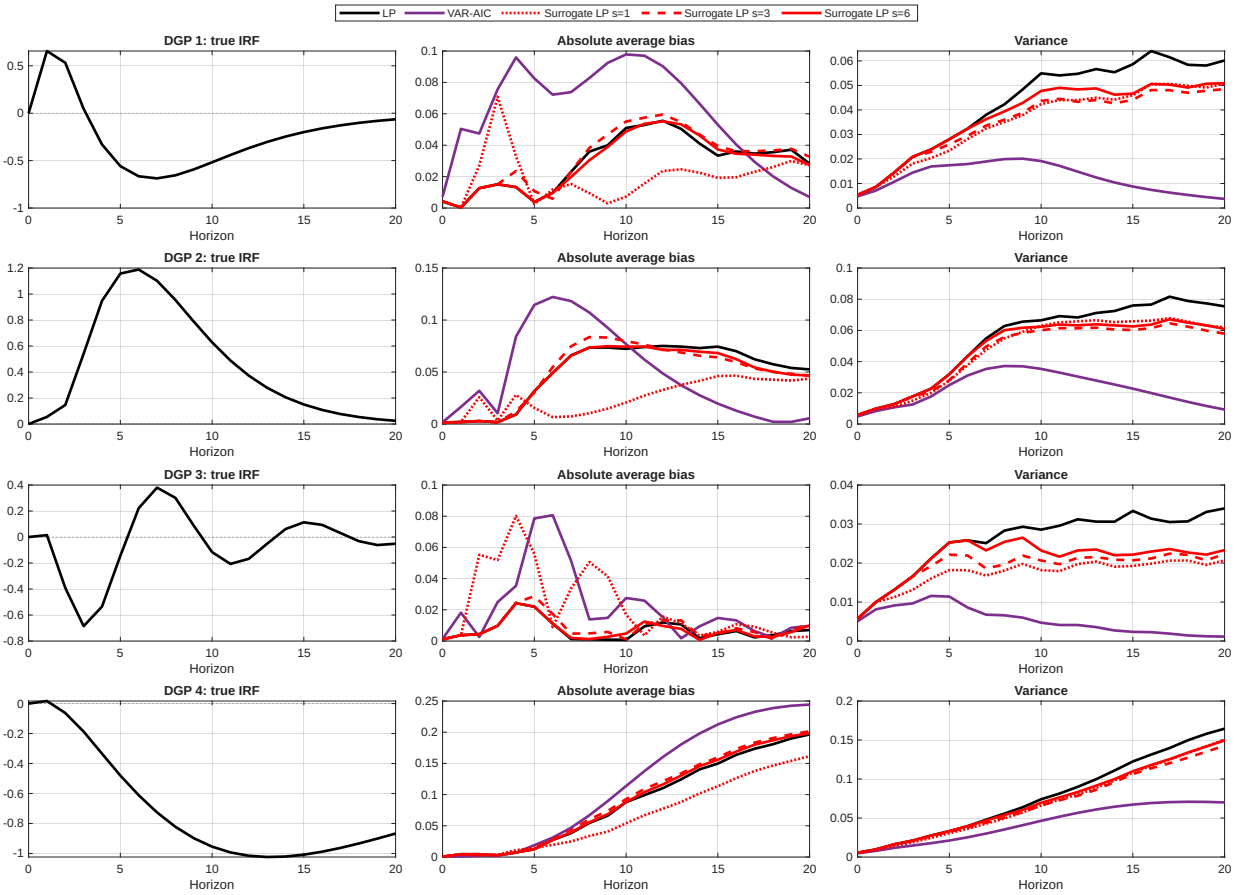
For every valid candidate (s,n),

$$\sigma_R^2(s,n,h) - \sigma_U^2(s,n,h) = O(T^{-2a}), \qquad h = s+1, \ldots, H,$$

with $a > 1/2$.

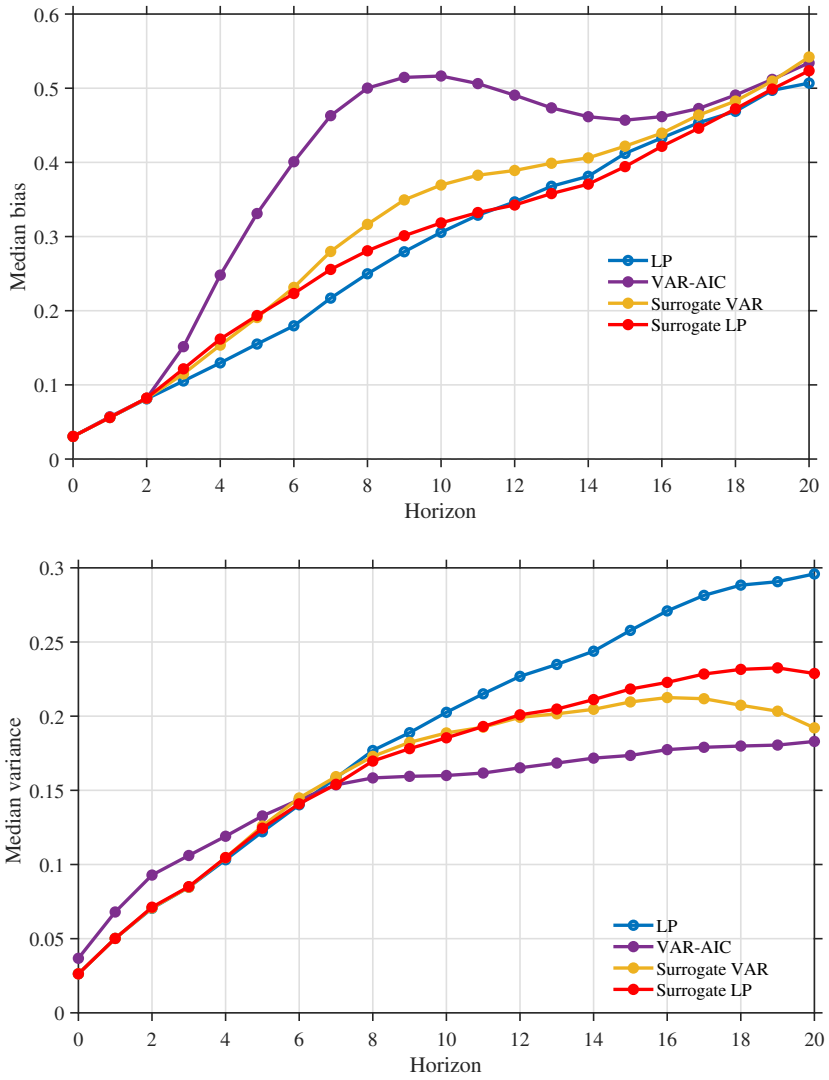
Appendix C: tables and figures

Figure 4: Surrogate LPs and VARs – VARMA simulations



Notes: The figure reports the true impulse responses together with the estimated responses, bias, variance, and mean squared error across the different simulation designs. Results are based on 5000 Monte Carlo replications with sample size $T = 200$. The conventional VAR uses lag length selection by AIC.

Figure 5: Surrogate LPs and VARs – DFM simulations



Notes: The figure reports the median of the average bias and variance across data generating processes derived from the dynamic factor model.