Humble VARs*

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

March 26, 2026

Abstract

Macroeconomic VARs aim to capture the economy's full autocovariance structure. Yet VARs used for impulse response analysis are typically small systems with few variables and lags, and the estimated impulse responses are highly sensitive to lag length and variable choice, indicating misspecification. We argue that VARs used for impulse response estimation should be framed around a more humble objective: the VAR only needs to capture the conditional autocovariance structure of the data. Building on this insight, we propose Conditional Out-of-Sample selection (COOS), 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. Do not circulate. 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 we best estimate the dynamic causal effects of that shock? A prominent method in macroeconomics is to use a vector autoregressive (VAR) model.[1] A VAR with p lags and n endogenous variables consists in modeling the dynamic behavior of the n variables in a recursive fashion; modeling 1-step ahead changes as a function of p lagged observations and iterating 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 structural models.

While VARs are often presented as flexible statistical models (Lütkepohl, 2013), the VARs *actually* employed by applied researchers often embed assumptions that are actually quite strong—much stronger, for example, than the surrogacy assumption sometimes made in the treatment effects literature (e.g., Athey et al., 2019). The reason is that these VARs are typically small-order systems, i.e., systems with a small number of endogenous variables and a small number of lags, a canonical example being the popular monetary VAR with $n = 3$ endogenous variables (e.g., output, inflation and the interest rate) and $p = 4$ quarterly lags. Indeed, with the short sample size typically encountered in macro (50 years of quarterly data, often much less), practitioners naturally aim at limiting the numbers of parameters to estimate and use a reduced number of lags and endogenous variables.

However, it is easy to forget that these canonical small-order VARs aim to achieve a remarkably ambitious goal: modeling the entire unconditional autocovariance structure of the macroeconomy with only p autocovariance matrices. To do this, the VAR must provide a *complete* statistical model of the economy, including *all* variables that can affect the macroeconomy or sufficient lags to encode this information. In other words, in a canonical VAR, four quarterly lags of three variables must encode the full state and dynamics of the macroeconomy, including the lingering effects of financial booms and busts, 9-11 and Covid. Is this plausible? We provide evidence it is not. Using a canonical VAR application —the estimation of impulse responses to monetary shocks (Romer and Romer, 2004)— we find that the estimated impulse responses are highly sensitive to the inclusion of additional variables or lags.[2]

Given this risk of mis-specification, it is tempting to turn to less parametric methods, most prominently Local Projections (Jordà, 2005; Olea et al., 2025). In short macro samples macro, however, efficiency concerns are paramount, and the dimensionality reductions offered

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

[2] We are by no means the first researchers to point out the sensitivity of VAR predictions to the lag length or the set of endogenous variables. However, there has been surprisingly little work documenting the high sensitivity of impulse response estimates to the VAR specification. See Bernanke, Gertler and Watson (1997) for an important exception in the context of oil shocks.

by VARs could be invaluable.

In this work, we argue that VARs should be framed around a more humble objective. Rather than attempting to model the entire economy, VARs should be viewed as power tools for specific tasks. To estimate impulse responses, a VAR does not need to capture the full unconditional autocovariance structure of the macroeconomy; it only needs to capture the *conditional* autocovariance structure. The intuition is simple: since an impulse response is a forecast conditional on a particular shock, accurately estimating an impulse response only requires that the VAR correctly captures the conditional autocovariance structure of the macroeconomy.

Guided by this "conditional" insight, we propose two innovations to the VAR apparatus. First, the VAR should be specified with a "conditional objective" in mind — to "best" estimate the conditional autocovariance of the data—, and we propose a Conditional Out-of-Sample selection (COOS) criterion to select the model with the best impulse responses (IRs) estimates.[3] Second, we argue that a "data conditioning" step prior to the VAR estimation could remove irrelevant but hard-to-model features of the data and yield efficiency improvements by reducing the dimension of the "best" VAR.

The first innovation is to explicitly confront the VAR specification question in the context of IR estimation: (i) which variables to include?, and (ii) how many lags? The COOS model selection procedure consists in choosing the VAR specification that minimizes the *out-of-sample* difference between the VAR-based IR and the IR estimated from a linear projection on the shock (a Local Projection or LP). Since VARs parameterize IRs as a sum of damped cosine basis functions, one can think of the COOS methodology as selecting the set of basis functions that produces the best out-of-sample predictions. The 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.

The second innovation is to propose a "data conditioning" step prior to the VAR estimation. The idea is simple: since the objective is to recover the conditional autocovariance of the data, the VAR does not need to be fitted on the unconditional data. In fact, by projecting the macro data on the space spanned by the shock of interest one can eliminate exogenous nuisance factors —"exogenous confounders"— that are difficult to capture with a small-order VAR: exogenous processes, noise, one-off events like COVID (Lenza and Primiceri, 2022), hard-to-detect MA coefficients (Olea et al., 2025) or even low-frequency

[3] "Best" can be defined in terms of mean-square-error, bias or confidence bands coverage, among other possibilities. The researcher can choose its preferred objective.

movements orthogonal to the shock of interest. By projecting out these dynamic processes, the resulting "cleansed" DGP may be dramatically simpler, presenting the small-order VAR with a more achievable task.

In practice, this conditional subspace may be too "small" to conduct reliable inference,[4] but the idea can be applied more generally: a data conditioning step can involve (a) projecting *out* certain factors difficult to model but independent of the shock of interest,[5] and/or (b) projecting *on* a subspace spanned by well known shocks; not only the shock of interest but possibly also other canonical macro shocks.[6] Ultimately, the idea is to find a subspace that is both small enough to remove hard-to-model but irrelevant features of the data and yet large enough to leave enough variation to estimate the VAR with sufficient precision. Interesting, the same COOS model specification criterion can be used to find the best data-conditioning step.

We evaluate the success of our proposed innovations (i) COOS and (ii) data-conditioning in a large Monte Carlo study. Across large set of DGPs, we obtain large gains in the precision of estimated impulse responses, both over canonical VARs as well as over Local Projections: COOS-specified VARS are successful as "power tools" in the estimation of impulse responses.

Literature Review

In a macro context, LPs and VARs have been proposed as the main estimators of dynamic causal effects (see Sims, 1980; Jordà, 2005; Olea et al., 2025). While LPs and VARs with infinite number of lags share the same estimand —the same population estimate (Plagborg-Møller and Wolf, 2021)—, they lie on opposite sides of the bias–variance trade-off. While VARs effectively model h-step ahead effects by recursively iterating on one-step ahead forecasts, LP imposes no dynamic structural on the impulse responses. As such, LP estimates have no asymptotic biases and is more robust to mis-specification (Olea et al., 2024).[7]

The recent recommendation from Olea et al. (2025) is to not rely on (short-lags) VARs because the risk of bias is too high for regular short lag length VARs. While the VAR error bands are small, they may not reflect the true uncertainty in the estimates, as the VAR

[4]For instance, the subspace spanned by (past and present) monetary shocks alone is likely "small" as monetary shocks account for a small share of the total variance of macro variables.

[5]This step includes popular (though rarely explicitly justified) procedures: band-pass filtering the data, projecting out polynomial trends, dummying out certain episodes.

[6]A related popular procedure is to project the macro variables on common factors as in the Factor Augmented-VAR literature (e.g., Stock and Watson, 2016).

[7]A growing literature has recently explored (both theoretically and by simulation) the relative properties of LPs and VARs in the specific context of dynamic causal inference. For theoretical explorations, see Jordà (2005); Kilian and Kim (2011); Ramey (2016); Brugnolini (2018); Choi and Chudik (2019); Bruns and Lütkepohl (2022); Li, Plagborg-Møller and Wolf (2024*a*); Olea et al. (2024, 2025); Ludwig (2024), and for performance evaluation in a controlled simulation setting, see Kilian and Kim (2011); Brugnolini (2018); Choi and Chudik (2019); Bruns and Lütkepohl (2022); Li, Plagborg-Møller and Wolf (2024*a*); Olea et al. (2024, 2025).

bias can lead to severe under-coverage of error bands. Since "accurate communication of estimation uncertainty in integral in applied macro", Olea et al. (2025) recommend relying on LP estimates or VARs with very long-lags ($p = H$, the IRF horizon), in which case the VAR is simply LP.

In the short samples typically encountered in macro however, efficiency is paramount, and a small amount of bias may be worth paying in these instances. Traditionally however, VARs have been used without much consideration for mis-specification risk, and practitioners have paid little attention to the identification assumption underlying the appropriate use of VAR(p)s. And while a large VAR literature has studied the risk of VAR-misspecification, it is from a different perspective —the VAR invertibility problem—: under which conditions can a VAR be used to identify (i.e., recover) structural shocks from a set of lagged endogenous variables?[8] We focus on a separate problem: given some exogenous treatment/shock series, how can we use VARs to more efficiently estimate the causal effects of interest? How important is the lag order choice, and how does this choice affects the IR estimates? Surprisingly, there is little work on this topic.

As this paper highlights, lag length selection is of crucial importance for successful causal analysis using VARs. Popular lag length selection procedures —sequential testing (i.e. VAR(p) vs VAR($p+1$)), tests for serial correlation in the residuals, information criteria and out-of-sample predictions (e.g. Lütkepohl, 2005)— are all designed for forecasting or specifying the entire VAR correctly, and the literature on this topic has remained relatively quiet over the past decades. An important exception however is the work of Schorfheide (2005); González-Casasús and Schorfheide (2025) who suggest using a multistep loss function for selecting the lags of the VAR based on estimates of prediction or IRF estimation risks. Last, our insight that the VAR specification procedure can be optimized depending on the shock, outcome variable and horizon of interest, relates to the work of Hsiao (1979) who suggested using the AIC criteria separately for the different variables in the VAR.

While the empirical macro literature paid little attention to the sensitivity of the IR estimates to lag length, there is an important exception in the oil shocks literature, where Hamilton and Herrera (2004) showed how Bernanke, Gertler and Watson (1997)'s IR estimates could change substantially depending on the number of lags included in the VAR.

[8] See for instance Lippi and Reichlin (1993); Forni and Gambetti (2014); Fernández-Villaverde et al. (2007); Canova and Ferroni (2022). Solutions focused on expanding the information set (and/or the dimension) of the VAR, e.g., expectations-augmented (e.g., Ramey, 2016) or factor-augmented VARs (Bernanke, Boivin and Eliasz, 2005).

2 Illustration: a VAR but which VAR?

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.

A direct approach would consist in directly regressing the variable of interest at time $t+h$ on the shock at time t, a so-called local projection Jordà (2005). To estimate the dynamic responses of n variables over H horizons, there would then be nH coefficient to estimates. An alternative is to use a VAR, a dynamic model for some vector of variables $\mathbf{y}_t$ satisfying

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

where $\mathbf{y}_t$ is a $(n \times 1)$ vector of macro variables, $\mathbf{A}_j \in \mathbb{R}^{n\times n}$ are the autoregressive coefficient matrices and $\mathbf{u}_t$ is a vector of (white noise) error terms.

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.[9]

However an important, and to date largely unresolved, question is "Which VAR to use?"; how should the researcher specify her VAR to obtain "best" impulse response estimates: (i) Which variables to include in the VAR and how many? (ii) How many lags (p)? Using two simple examples —an empirical setting and a simulation—, we argue that the VAR specification choice is central for VAR-based impulse response estimation in terms of (i) accuracy and (ii) precision, *even* when the shock of interest has already been identified.

We first illustrate how a researcher could draw different conclusions on the effects of monetary policy —obtain different point estimates— depending on the VAR specification, even when a monetary shock series has already been identified. Second, we use a simple simulation to show how the VAR specification decision can have substantial implications for inference, and how an "well-specified" VAR can deliver substantial efficiency gains at

[9] 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.

relatively little bias cost, making an "optimally designed" VAR an attractive power tool — substantial efficiency gains at relatively little bias cost— for the short sample environments encountered in macro.

2.1 VAR specification and point estimates

A researcher is interested in estimating the effects of monetary policy on GDP growth, inflation and the fed funds rate,[10] and she has access to a series of exogenous policy changes in the policy rate. Here, we will use the Romer and Romer (2004) monetary shock series.

With the shock series in hand, it is possible to construct the empirical distribution of forecast errors following a unit monetary policy shock. Intuitively, given a forecast based on data observed before a monetary shock, we can compute the distribution of forecast errors following a monetary shock. The top row of Figure 1 plots the empirical distribution of such forecast errors over multiple horizons. This chart can be seen as a method to visualize the "raw data" underlying impulse response estimates: the relation between a unit monetary shock and the realized forecast errors over H horizons.[11]

The raw data already make out a few stylized facts about the effect of monetary 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.

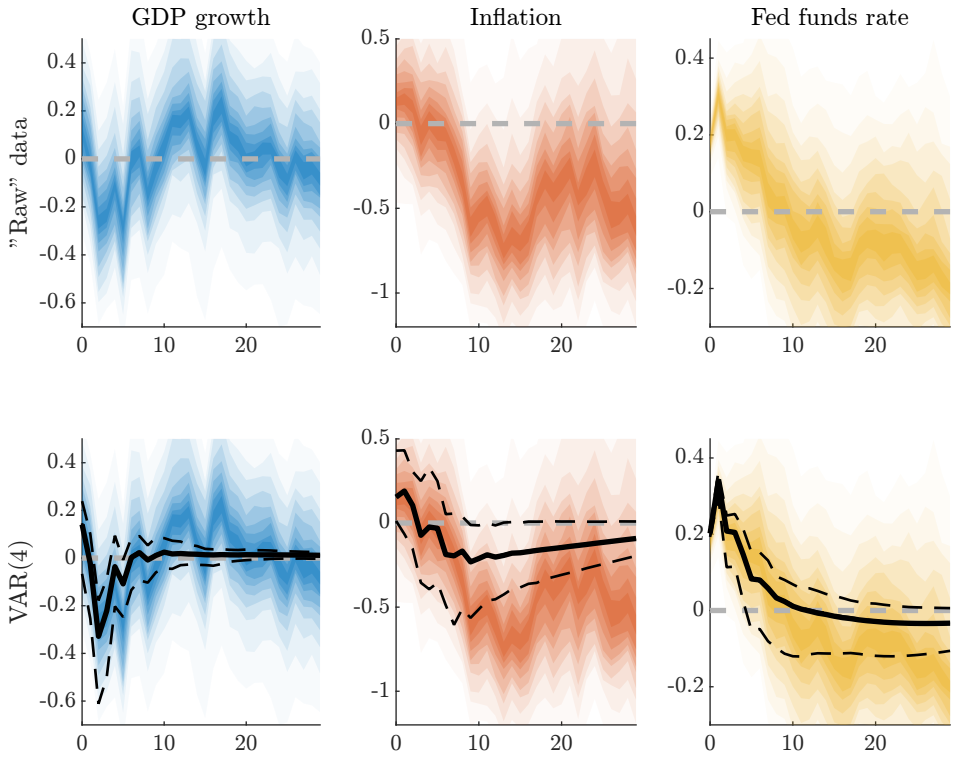
Consider instead using a VAR to estimate the transmission of the policy shock. A popular choice for monetary VARs going back to Sims (1980) is a system with $n = 3$ endogenous variables (output growth, PCE inflation and the fed funds rate) and $p = 4$ lags for quarterly data. Since a shock proxy has already been identified, a simple way (e.g., Plagborg-Møller and Wolf, 2021) to obtain the impulse responses to a monetary shock is then to add the Romer and Romer (2004) shock to the VAR, order it first, and compute the impulse responses to innovations to the Romer and Romer (2004) shock.

The bottom row shows the estimated impulse responses to a monetary shock. Notice how the VAR confidence bands are substantially narrower than the forecast error dispersion. VARs are power tools: by imposing restrictions on impulse responses —across horizons, variables, and shocks— the VAR(4) achieves substantial power gains —the confidence bands are "narrow"— and enables informative inference. At the same time however, we can immediately notice that the VAR impulse responses seem at odds with the raw data; for instance,

[10] GDP growth is quarter-to-quarter change in GDP, and inflation is quarter-to-quarter change in the PCE price index. The data cover 1969q1-2007q4.

[11] Except for an "average shock effect normalization", we did not impose any structure on the data. By construction, the mean of that forecast error distribution coincides with the local projection point estimate. See the Appendix for details.

Figure 1: Illustration: Effect of monetary policy



Notes: 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).

the GDP growth response does not display any sign of rebound after two years, and the inflation response displays a seemingly too small and too early decline.

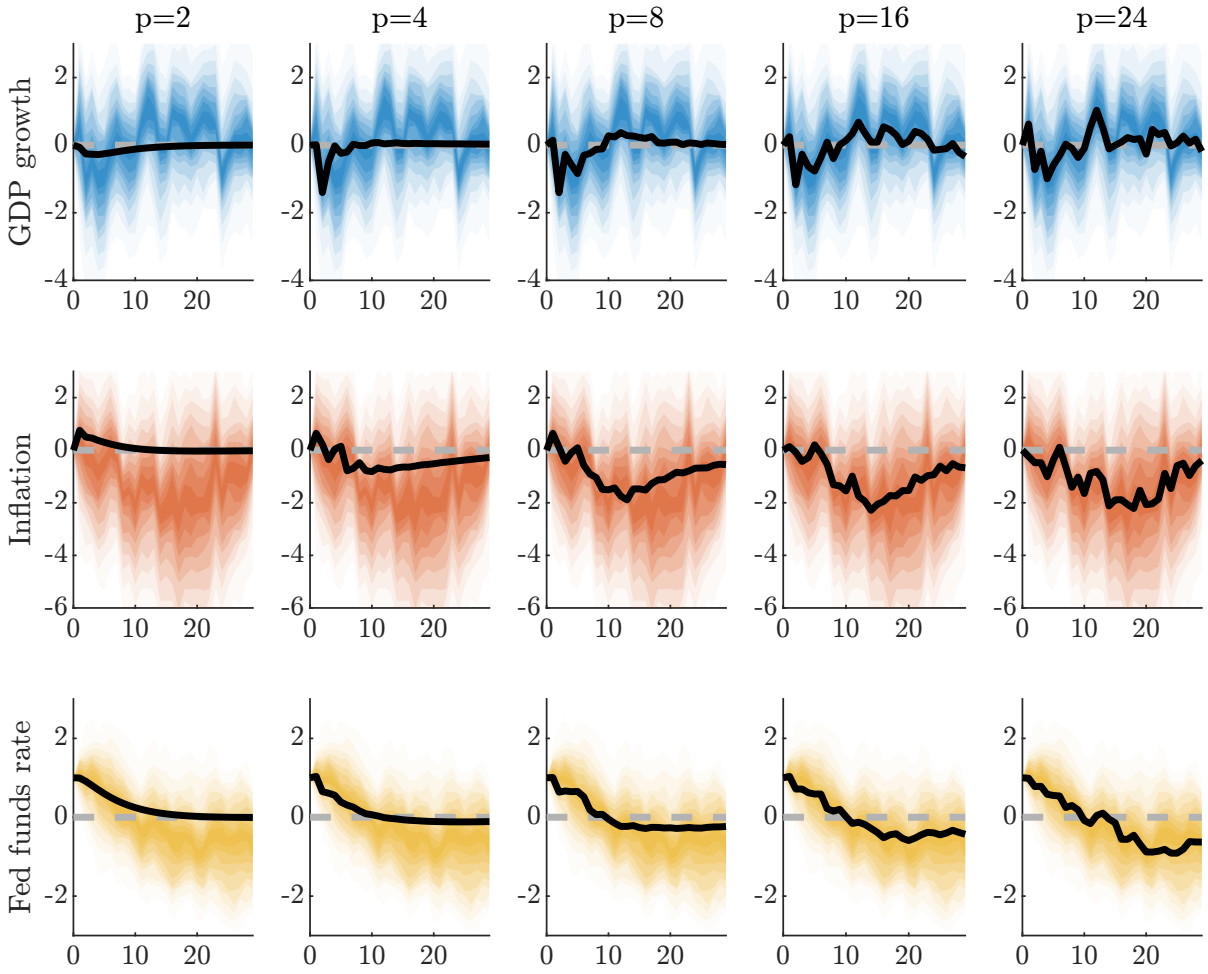
Could it be that this popular VAR specification is *too* restrictive? That using $p = 4$ lags is too short a lag length? or that using only $n = 3$ endogenous variables is too small a VAR size? But if so, which variables (and how many) should be included? And what should be the lag length? Unfortunately, the impulse responses are non-linear transformations of the VAR coefficients, and it is not always clear how restrictions on the VAR size (the number of lags or the set of endogenous variables) translate into restrictions on the underlying impulse responses.

To illustrate the influence of the VAR specification, we will assess the sensitivity of IR estimates to the VAR specification, first to the number of lags p (holding n fixed), and then to the number of endogenous variables n (holding p fixed).

The lag length choice

Figure 2 plots the estimated IRs as we increase the VAR lag length from $p = 2$ to 4, 8, 16 and 24.

Figure 2: Lag length sensitivity



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

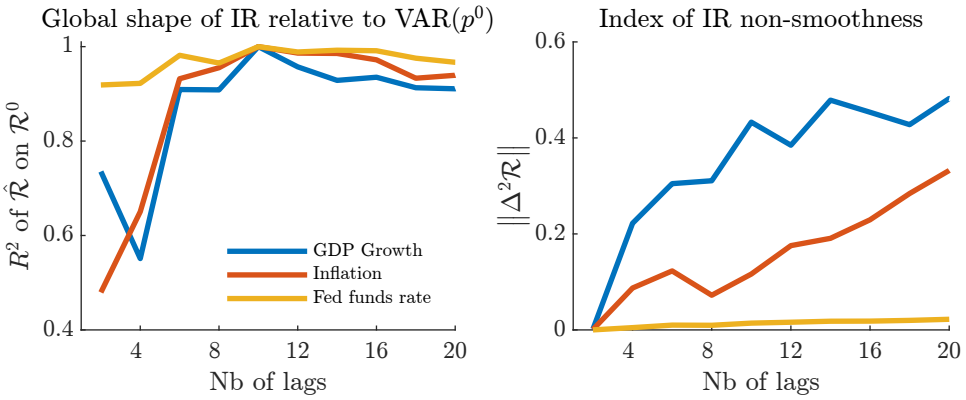
Notice how much the estimated IRs change as we increase the lag length of the VAR. Across p two types of changes occur in the IRs: (i) the global shape changes and (ii) the smoothness of the irfs changes. Figure 10 shows two metrics describing (i) and (ii). The first metric is the R^2 of a regression of the VAR(p)-estimated impulse response on the impulse response estimated with a VAR(p^0) and smoothed in order to capture the "global shape" by removing high-frequency movements (using an HP filter with $\lambda = 100$). This R^2 metric serves to measure how the "global shape" of the estimated IR changes as we increase p above $p^0 = 10$.[12] A high R^2 indicates that the global shape of the IR remains stable as we increase

[12]The level of p^0 is arbitrary, but it is meant to capture a long lag length for macro applications.

the VAR lag length above p^0. In this example, we took $p^0 = 10$. The second metric is a measure of IR non-smoothness and is defined as the sum-of-squares of the second difference of $\mathbf{\Psi}$: $\|\Delta^2\mathbf{\Psi}\|$. Intuitively, this index captures the extent of choppiness in the impulse response, as jagged impulse responses will feature high average values of the squared second derivative.

In Figure 2, notice how the estimated IR change dramatically as we increase the VAR lag order. 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, possibly a sign of overfitting as we will see in the next section. A similar pattern can be seen in the response of inflation. As we initially increase the lag length, the decline in inflation 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. Figure 3 confirms this visual inspection: the R^2 metrics are close to one and stable after about $p = 8$. However, as we increase p further, the estimated impulse response becomes more and more jagged: the degree of non-smoothness keeps increasing (Figure 3, right panel). 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.

Figure 3: IR lag length sensitivity: Summary Measures



Notes: Left panel: R^2 of a regression of the VAR(p)-estimated IR on the smoothed IR estimated with a VAR(p^0), $p^0 = 10$. The R^2 metric measures how the VAR(p)-estimated IR differs from the (smoothed) IR estimated with a VAR(p^0). A high R^2 indicates stability in the IR global shape. Right panel: measure of IR non-smoothness; defined as the sum-of-squares of the second derivative of $\mathbf{\Psi}$: $\|\Delta^2\mathbf{\Psi}\|$.

The set of endogenous variables

Although rarely discussed in applied work, the set of endogenous variables can also dramatically affect the estimated IRs.

We repeat our p-sensitivity experiment, but where we vary to set number of endogenous variables, progressively increasing n, from the baseline minimalist form with $n = 3$ to a large VAR with $n = 13$ endogenous variables. Throughout, we keep the lag length fixed at $p = 4$.

The variables we successively entered in addition to the baseline 3 endogenous variables (output growth, inflation and the fed funds rate) are (i) the unemployment rate, (ii) consumption growth and investment growth, (iii) commodity price inflation, (iv) consumption growth split into durables and non-durables growth, and inflation split into its main components service inflation, durable goods inflation and non-durables goods inflation, and also core inflation, (v) the BAA-AAA spread, changes in total reserves or the exchange rate.

Figure 4 shows the results with a similar pattern to the p-sensitivity exercise: as we increase the number of endogenous variables, the rebound in GDP growth starts appearing, and the decline in inflation becomes more marked. And as with the p-sensitivity exercise, the larger the VAR size (pn^2) the less smooth the estimated impulse responses (though the effect is stronger with p).

In fact, while the VAR impulse response for GDP growth approaches that of VAR with large p as the researcher raises n (suggesting that a user might be able to compensate a too short lag length with a larger system —"substituting n for p"—. However, this is not the case for the inflation response: the impulse response remains substantially below the LP estimate even with $n = 13$, and increasing p has a stronger effect on the inflation IR than increasing n does. Of course, this could be due to the set of endogenous variables missing some central mediator/surrogate for the inflation response.

As it turns out, the traditional "extra" variables entered in monetary VARs (commodity price inflation, BAA-AAA spread or EBP, changes in total reserves or the exchange rate) do little to the estimated IRs. These variables may be necessary for identification but *not* for propagation estimation. Instead, the components of GDP and the components of inflation seem most promising to explain the transmission of monetary policy, with durable consumption and investment being responsible for the rebound in GDP growth after 2-3 years and service inflation displaying a protracted and persistent decline. In other words, these persistent variables help the VAR generate richer impulse response dynamics at longer horizons.

The reason for obtaining different conclusions to the question of "which variables to include" goes back to the a simple point: shock identification and dynamic effect estimation are two different tasks. The sensitivity to the set of endogenous variables —"n-sensitivity" for short— for propagation estimation is *different* from the role of controls in identification

strategies.

The literature has so far mostly focused on the influence of the n-choice on the credibility of the identification scheme. Traditionally, the n-choice is discussed in the context of shock identification: the number of endogenous variables must be large enough to span the agents' information set and be able to recover the true shocks from the Wold innovations —the invertibility problem—. It is in that context that variables such as commodity prices or forward-looking variables like stock prices were included in the VAR: to better capture the agents' information set and thus to not mistakenly consider an interest rate change as exogenous when it is in fact in reaction to developments occurring in the oil market or in financial markets.

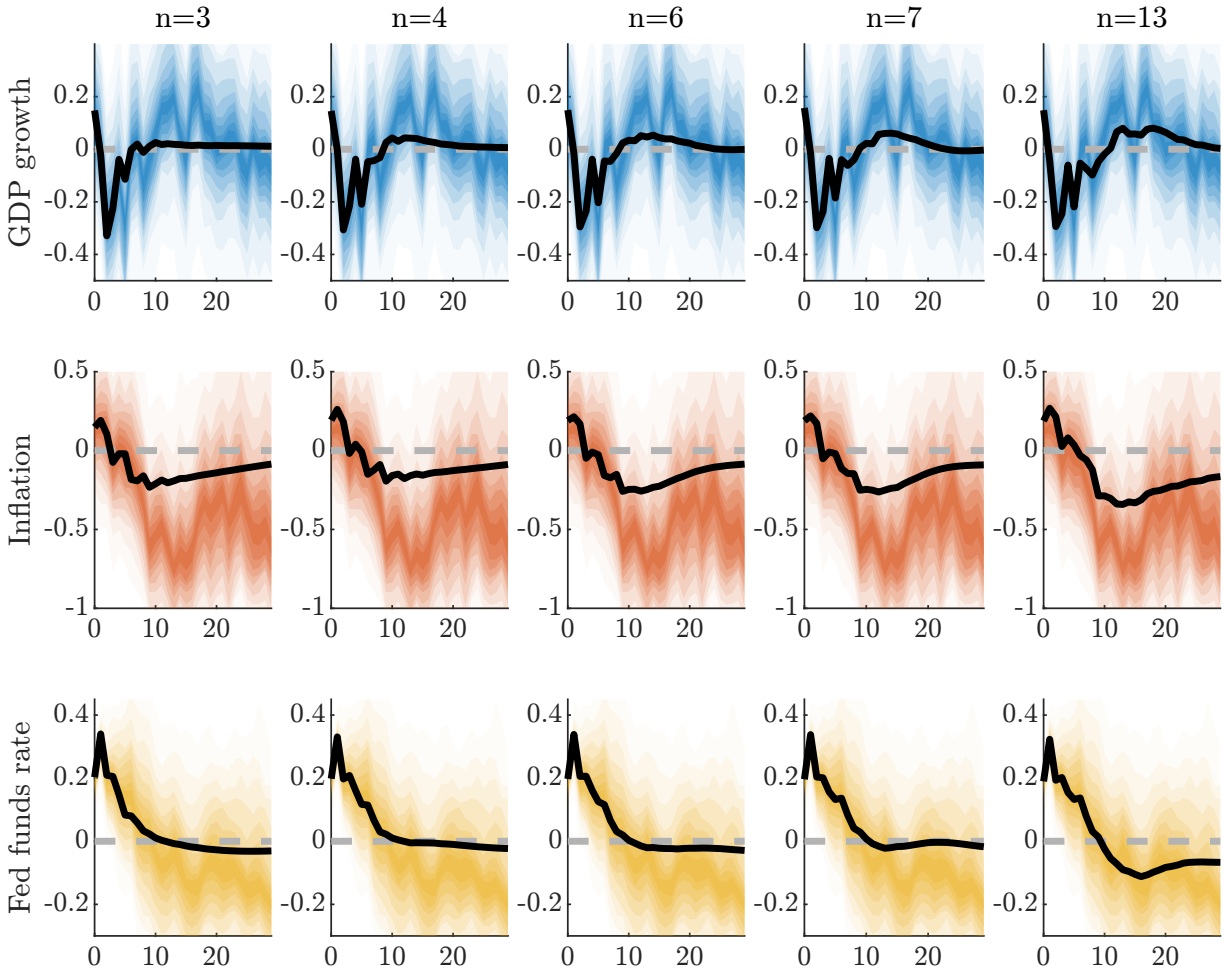
Freed from the issue of identification (here, thanks to the Romer and Romer (2004) narrative series), we can focus solely on the propagation question, and highlight the sensitivity of the IR point estimates to the set of controls in VAR.

More generally, this n-sensitivity exercise illustrates how the most appropriate set of endogenous variables can be very different depending on the task at hand: shock identification vs propagation estimation. For propagation estimation a researcher must include the variables that are either mediators or surrogates for the long-run effects of the shock on the economy. For instance, if a monetary shock affects the unemployment rate, which then affects households' cash-on-hand and then consumption decisions, it is important to include the unemployment rate into the VAR. Similarly, if the rebound in economic activity is driven by an investment rebound, it may be important to include investment in the VAR, as the I response mediates the effect of the monetary shock on output. This can be very different from the need to include variables like commodity price inflation (Sims, 1992) or non-borrowed reserves (Bernanke and Mihov, 1998) for the purpose of shock identification.

Possible candidates for endogenous variables thus include the components of GDP or the components of inflation. To the extent that the dynamic response of disaggregated data is different from the aggregate, using disaggregated data that have different time series properties (here different degrees of persistence) can help the VAR better predict (i.e., estimate) the dynamic effect of the monetary shock. This point relates more generally to the debate of whether one should forecast with aggregate variables or dis-aggregated variables (Marcellino, Stock and Watson, 2003). In fact, we will see that the solution to the n and p choice is precisely related to a forecasting performance criterion.

An interesting metric in that context is which variable leads to largest changes in the shape of the impulse response, in particular studying which variables are most influential when estimating the impulse responses of output growth and inflation.

Figure 4: VAR SIZE SENSITIVITY



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

2.2 VAR specification and inference

We will use a simple simulation to illustrate how the VAR specification can have substantial implications for inference, and how a "well-specified" VAR can deliver substantial efficiency gains at relatively little bias cost.

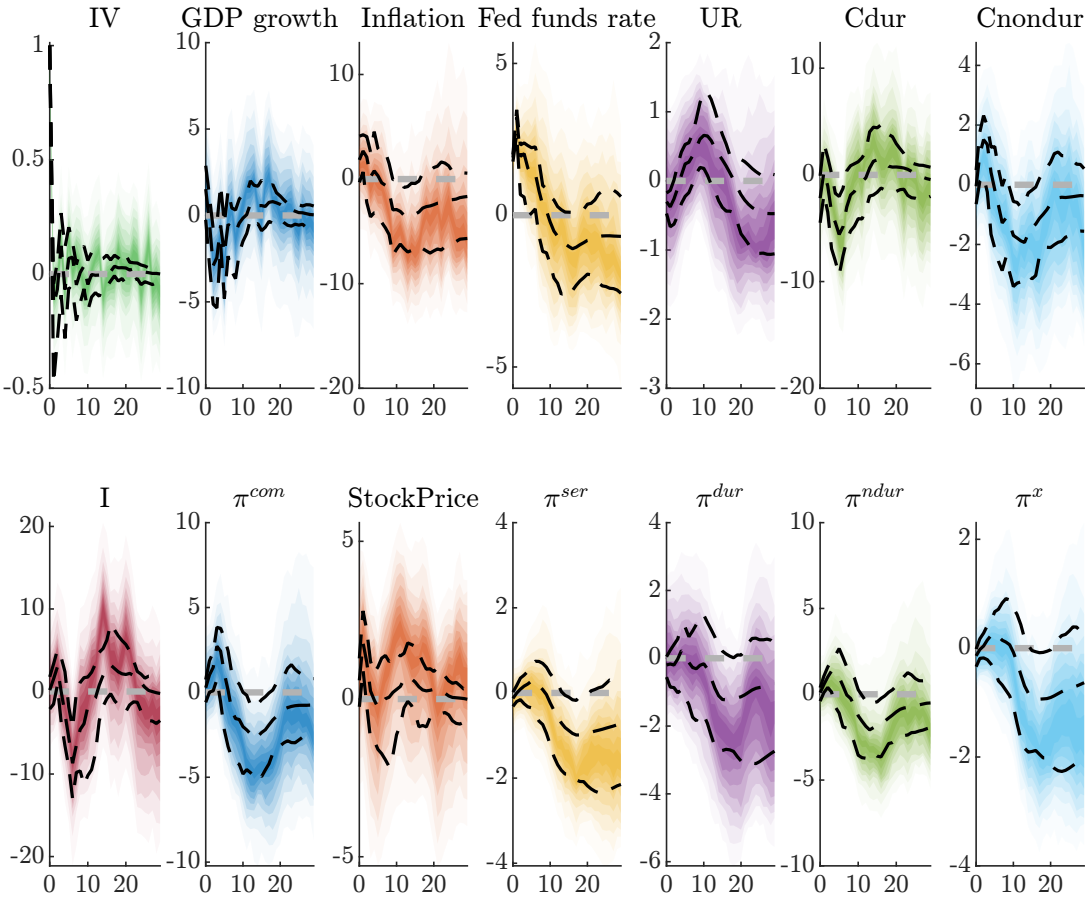
We simulate a simple bivariate DGP for $\mathbf{y}_t = (y_t, x_t)'$ that captures some stylized features of the monetary impulse responses seen previously:

$$\begin{cases} x_t = \rho_y x_{t-1} + \varepsilon_t^x \\ y_t = \rho_y y_{t-1} + \beta x_{t-8} + \varepsilon_t^y \end{cases} \tag{2}$$

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 6 plots the

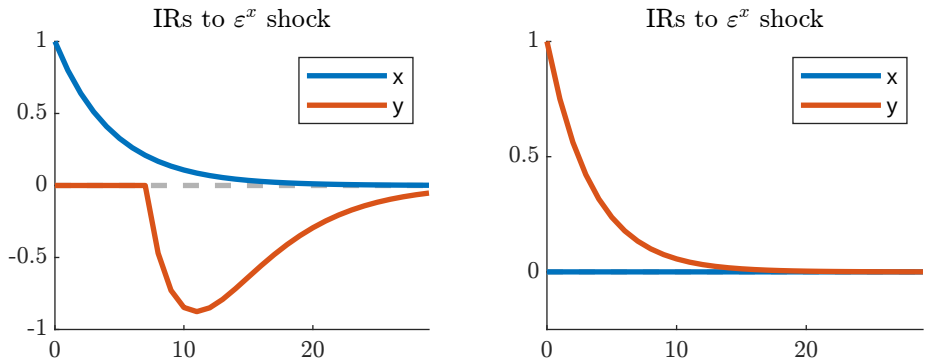
Figure 5: IR estimates from large VAR(4) with $n = 13$



Notes: IR estimated from a VAR(4) with $n = 13$ endogenous variables.

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 6: Simulated impulse responses



We first illustrate the effects of VAR *under*-parametrization and then study the effects of VAR *over*-parametrization.

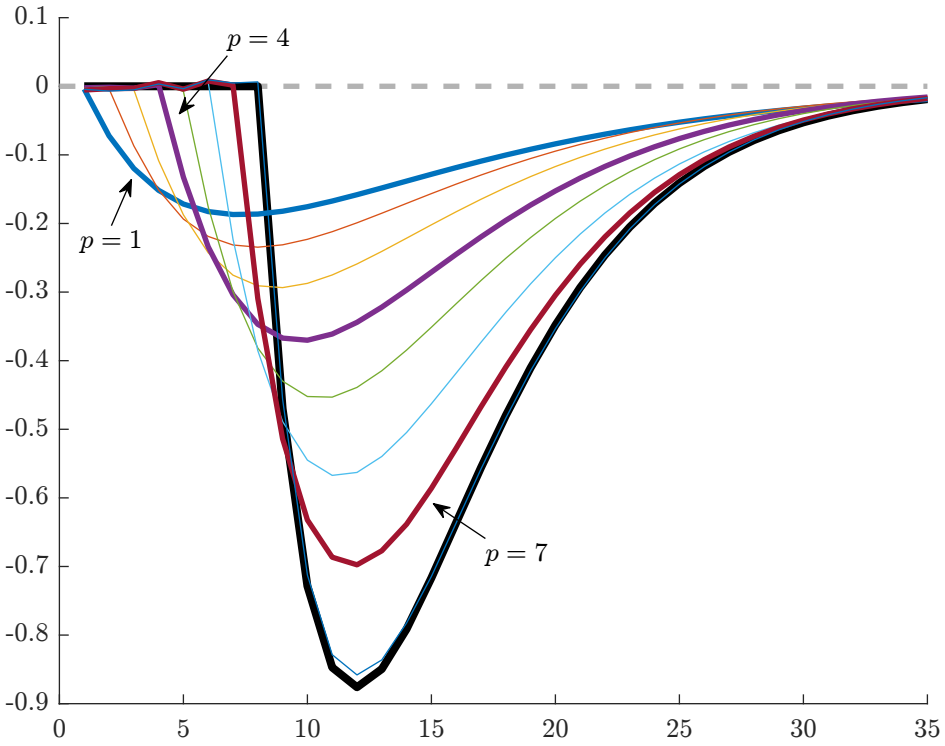
On the effects of VAR under-parametrization We first focus on the effects of under-parametrizing —a too short lag length p— in population. Similar to Figure 2, Figure 7 plots the VAR(p)-based IRs as we increase the lag order p.[13] We can see that a mis-specified VAR with $p < 8$ faces a tension: it 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 and too early decline in w. This is the same pattern that we observed for the IR of inflation in the monetary VAR (Figure 2, middle row)

On the effects of VAR over-parametrization To illustrate how overfitting can affect estimation precision and inference, we now consider a finite sample setting with $T = 200$, and we study how performances are affected when we vary p, notably when $p > p^0$ where $p^0 = 8$ is the correct lag length. Figure 8 compares four estimators: a VAR(4) —a short-VAR—, a VAR(8) —an intermediate VAR—, a a VAR(20) —a long VAR— and a Local Projection with 20 lags used as controls, it shows the MSE of IR estimates at different horizons, along with IR confidence band length and confidence interval coverage.

Notice the large deterioration in performance when going from a VAR(8) —an "efficient estimator" since the DGP *is* a VAR(8)— to a VAR(20) —an over-parametrized estimator. The MSE is 3 times larges at horizon 15 (see orange line vs yellow line, left panel, Figure 8).

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

Figure 7: VAR-BASED IRS



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

More generally, Figure 9 shows how MSE performances deteriorates rapidly when the VAR becomes over-parametrized with $p > 8$: estimation variance (red line) increases monotonically with the number of VAR parameters. This is the pitfall of over-parametrization.[14],[15]

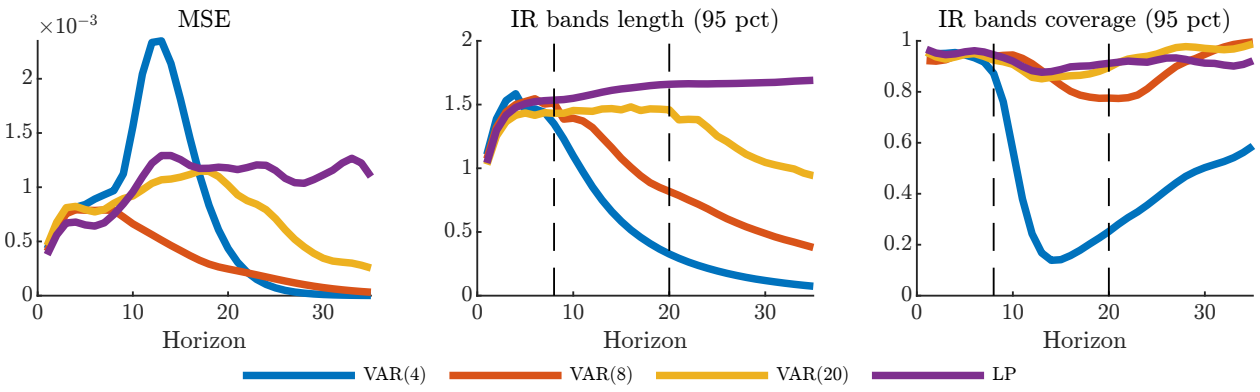
To further explore the consequences of over-parametrization (and help understand our monetary VAR results), Figure 10 shows the same two metrics that we used in the monetary VAR, see Figure 3. The global shape of the estimated IR is stable past horizon $p \approx 8$ —once the VAR becomes sufficiently parametrized—, but the degree of IR non-smoothness rises with $p > p^0$ —as the VAR becomes over-parametrized. This can explain the deterioration in MSE performances.

Interestingly, note how similar are Figures 10 to the picture we obtained for the monetary

[14]That said, notice also that the VARs deliver substantial power improvements over LP. At horizon $h = 10$ for instance, the VAR(8)'s MSE is 40 percent smaller than LP's, and it is as much as 60 percent smaller at horizon $h = 20$, growing ever larger as we increase the horizon. Even a VAR(20) delivers large gains over LP for $h > 20$. This is because the VAR imposes mean reversion, whereas LP does not.

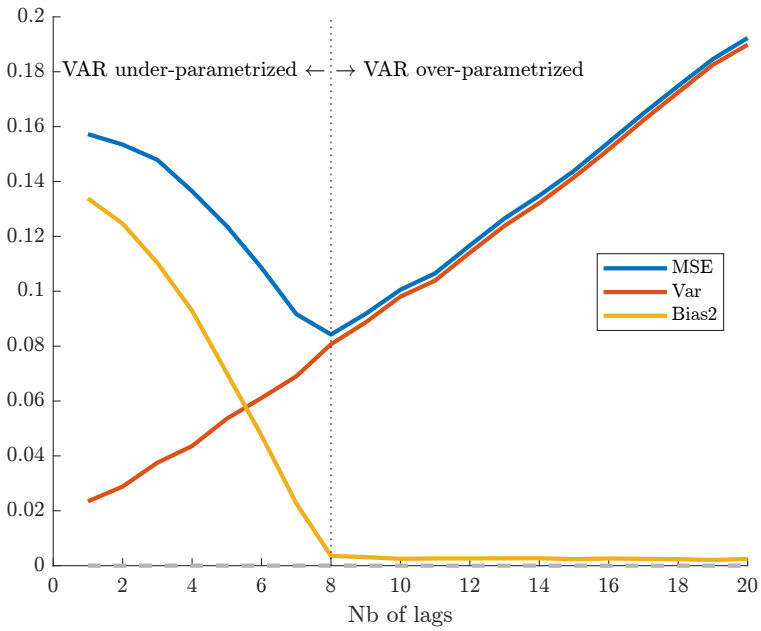
[15]The finite sample simulation also illustrates the pitfalls of under-parametrization. Setting p too small as with the VAR(4), and the VAR becomes highly biased. The MSEs over horizons 8 to 16 shoot up compared to the other estimators. Worse, while the IR bands of the VAR(4) are narrow, they have very poor coverage —counter-factually rejecting the null of no effect 70 percent of the time instead of the nominal 5 percent—. A mis-specified VAR can thus give a false sense of confidence in the point estimates, a point forcefully made by Olea et al. (2025).

Figure 8: VAR(p) and IR estimation performance: Simulation



Notes: Confidence bands were computed by non-parametric bootstrap of a long VAR with $p = 20$.

Figure 9: IR lag length sensitivity in Simulation: MSE decomposition

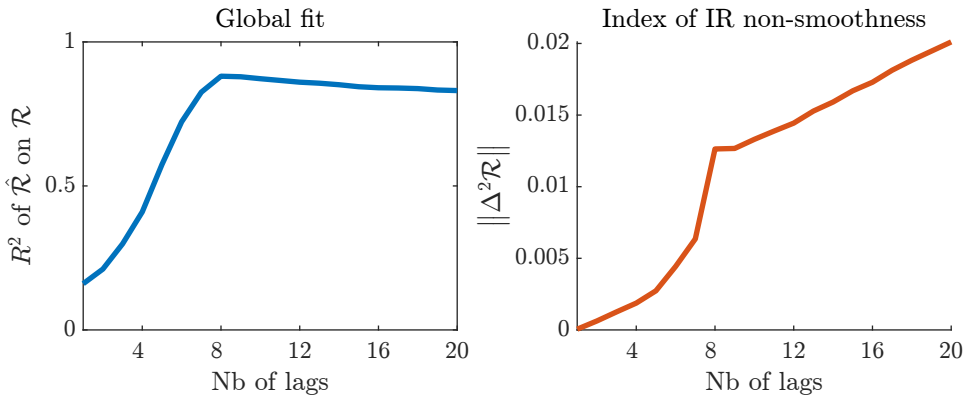


Notes: XXX.

VAR. In our simulation, the jagged nature of the IR is indeed a sign of overfitting. Thus, overfitting can explain why the IR estimates of our empirical monetary VAR becomes jagged for p above about 10.

To sum up, by imposing dynamic restrictions across horizons and across variables, a well-designed VAR can achieve large efficiency gains over semi non-parametric methods like LP. However, navigating the bias/variance trade-off of VARs is difficult: too small a VAR

Figure 10: IR lag length sensitivity in Simulation: Summary Measures



Notes: XXX.

and the dimensionality reduction will have a large bias cost, and too large a VAR and the efficiency gains will deteriorate rapidly.

2.3 Take-away

The main take-away of these illustrative examples is that an important, but often overlooked, question for VAR users is that of the VAR specification; how to choose the set of endogenous variables (and the set's size n) and how to choose the lag length (p)? We thus turn to these issues in the next two sections, first from the traditional VAR-as-DGP perspective —under which assumptions does an n-VAR(p) consistently estimate an IR of interest?—, and second from a less common, but we think more fruitful perspective: the bias/variance trade-off —how can we use a VAR to estimate an IR of interest *even if* the VAR provides a poor approximation of the underlying DGP.[16] In other words, how should a researcher choose p and n to optimize the VAR power tool for the task at hand: the estimation of dynamic causal effects.

3 The VAR-as-DGP perspective

A popular starting point when writing down a VAR is that the VAR is meant to provide a reasonable description of the macroeconomy (ADD example CITE), that is that the VAR can provide a good description of the underlying DGP. From that perspective, specifying the VAR then amounts to finding the set of variables that can capture *all* the important

[16] As we will see, the key is that the VAR does provide a good approximation of the *conditional* DGP, that is a good approximation of the economy's behavior *conditional on* the macro shock of interest.

interactions taking place at the macroeconomic level, and then recovering the true lag length through an information criterion or alternative methods (e.g. Kilian and Lütkepohl, 2017).

In the context of monetary policy for instance, practitioners often posit a VAR with three variables —output growth, inflation and the policy rate—, consistent with the canonical (though stylized) IS/MP–AD/AS framework, in which these three variables are sufficient to characterize the macroeconomic equilibrium and its response to macro shocks.

The key (often implicit) assumption underlying this approach is that there exists a finite size VAR that can fully characterize *all* the dynamic relationships between macro variables, or loosely speaking that there exists a finite size VAR, which can provide a "complete" statistical model of the economy. In this section, we will argue that this is a very strong assumption, much stronger than the corresponding assumption made in other fields, in particular the treatment effect literature.

3.1 VAR assumptions and IR estimation consistency

Environment

To formally lay down the assumptions underlying the VAR, we first briefly formalize the setting in which our study takes place, specifying the underlying data generating process, the dynamic causal effect of interest and the VAR estimand.

We 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[17]

$$\mathbf{d}_t = \sum_{j=0}^{\infty} \boldsymbol{\Theta}_j \boldsymbol{\eta}_{t-j} \ , \qquad \boldsymbol{\eta}_t \sim \mathrm{WN}(\mathbf{0}, \mathbf{D}) \ , \tag{3}$$

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 diagonal variance $\mathbf{D}$. 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 $y_t \equiv \mathbf{d}_{1t}$ and η_{1t} as the outcome variable and structural shock of interest, respectively. We aim to recover the causal effect of η_{1t} on the outcome variable y_{t+h} at horizon h, for $h = 0, \ldots, H$. That is

$$\theta_h \equiv \boldsymbol{\Theta}_{11,h} = \mathbb{E}(y_{t+h} | \eta_{1t} = 1) - \mathbb{E}(y_{t+h} | \eta_{1t} = 0) \ , \qquad h = 0, \ldots, H \ . \tag{4}$$

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

Since the scale of the structural shock is not pinned down by the model (3), we adopt the unit effect normalization (e.g. Stock and Watson, 2018). For this let $x_t \equiv \mathbf{d}_{2t}$ be the normalizing variable on which η_{1t} has a unit effect on impact, i.e. we normalize $\mathbf{\Theta}_{21,0} = 1$. As an example, suppose that η_{1t} is the monetary policy shock, we can take x_t as the policy rate which implies that a one unit increase in the shock is equal to a one percentage point increase in the policy rate.

To estimate the dynamic causal effect θ_h, we consider a VAR for the vector $\mathbf{y}_t \in \mathbb{R}^n$ with $\mathbf{y}_t = (y_t, x_t, \mathbf{w}_t)$ where $\mathbf{w}_t$ captures variables in the VAR beyond y_t and x_t.from now on as the vector $\mathbf{w}_t$.

The generic VAR model is

$$\begin{aligned} \mathbf{y}_t &= \mathbf{A}_1\mathbf{y}_{t-1} + \ldots + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{u}_t \qquad \text{where} \qquad \mathbf{u}_t = \mathbf{B}\boldsymbol{\varepsilon}_t & (5) \\ &= \mathbf{A}(L)\mathbf{y}_{t-1} + \mathbf{B}\boldsymbol{\varepsilon}_t & (6) \end{aligned}$$

with $\mathbf{A}_j \in \mathbb{R}^{n\times n}$ the autoregressive coefficient matrices and $\mathbf{B} \in \mathbb{R}^{n\times n}$ relates the reduced form population residuals $\mathbf{u}_t$ to the, possibly structural, shocks $\boldsymbol{\varepsilon}_t$, which are mean zero and mutually uncorrelated such that $\text{var}(\boldsymbol{\varepsilon}_t) = \mathbf{D}_\varepsilon$ is diagonal. Formally, the coefficient matrices $\mathbf{A}_1, \ldots, \mathbf{A}_p$ are defined as the best linear prediction coefficients such that, by construction $\mathbb{E}(\mathbf{u}_t\mathbf{y}'_{t-j}) = 0$ for $j = 1, \ldots, p$.

The impulses responses of the VAR are defined in the usual way: the response at horizon h of variable i to shock j is

$$\mathbf{\Psi}_{ij,h} = (\mathbf{C}_h\mathbf{B})_{i,j} \ , \qquad h = 0, \ldots, H \ , \tag{7}$$

where the i and j subscripts respectively denote the i-th row and j-th column of a matrix and the matrix $\mathbf{C}_h$ is defined recursively for $h = 0, ..., H$ by $\mathbf{C}_h = \sum_{m=1}^{h} \mathbf{C}_{h-m}\mathbf{A}_m$, with $\mathbf{C}_0 = \mathbf{I}_n$ (e.g. Kilian and Lütkepohl, 2017, Chapter 4).

VAR assumptions: Beyond Invertibility

Although rarely discussed, it is helpful to explicitly lay down the assumptions under which an n-variable VAR(p) can recover, i.e., consistently estimate, the impulse response of interest θ_h.

For pedagogical purposes, we start from an intuitive justification for using a VAR: as the "inversion" of the economy's SVMA representation (3), where the VAR-IR consistency rests on two assumptions:

Assumption 1. *Given the economy* (3) *the VAR model for* $\mathbf{y}_t = (y_t, x_t, \mathbf{w}_t)$*, as defined in* (5)*, satisfies the following*

INV $\boldsymbol{\varepsilon}_t = \boldsymbol{\eta}_t$ *and* $\boldsymbol{\eta}_t \in \text{span}\{\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots\}$

CLS for a known or estimable finite p *we have* $\mathbb{E}(\mathbf{u}_t|\mathbf{u}_{t-1}, \mathbf{u}_{t-2}, \ldots) = 0$.

The first assumption —Invertibility (INV)— states that there is a one-to-one mapping between VAR innovations and structural shocks affecting the economy; meaning that it is possible to recover all structural shocks from past observations of $\mathbf{y}_t$. The term "Invertibility" comes from the possibility of inverting the economy's SVMA representation (3) and turning it into a VAR, i.e., that there exists a matrix lag polynomial $\mathbf{A}(L)$ such that $\boldsymbol{\Theta}(L)^{-1}(\mathbf{I} - \mathbf{A}(L)) = \mathbf{I}$. In other words, the autocovariance properties of of $\mathbf{y}_t$ can be fully captured by a VAR (of possibly infinite lag length) $\mathbf{y}_t = \mathbf{A}(L)\mathbf{y}_{t-1} + \mathbf{B}\boldsymbol{\varepsilon}_t$. The second assumption —Correct Lag Specification (CLS)— states that that VAR has finite lag length: a VAR with a finite number of lags is sufficient to capture the entire autocovariance structure of $\mathbf{y}_t$.

Taken together, these two assumptions imply there exists a VAR(p) that can provide a complete structural model of the economy. The CLS assumption states that there exists a finite lag length VAR that provides a complete *statistical* model of the economy (a complete characterization of the autocovariance structure of $\mathbf{y}_t$), and the Invertibility assumption allows to convert that statistical model into a structural model. In other words, under Assumption 1, there exists a VAR(p) that can serve as model of the underlying DGP; a "VAR-as-DGP" perspective. With the Invertibility and CLS conditions, it is then easy to show that the VAR impulse response correspond to the impulse response of interest, see for example Stock and Watson (2016) for a proof.

The VAR-as-DGP perspective embodied by Assumption 1 has guided most model selection methods (more on this below) as well as the first generation of structural VARs and the aim to identify all the underlying structural shocks affecting a system (e.g., Gali, 1992; Blanchard and Perotti, 2002; Kilian, 2009).[18] This perspective is best illustrated with the seminal Blanchard and Quah (1989) bivariate VAR designed to identify both aggregate demand and aggregate supply shocks. Intuitively, if the macroeconomy is hit by only two types of aggregate shocks (aggregate demand and aggregate supply), which can be recovered from past values of two macro variables (unemployment and real output) —Invertibility in Assumption 1—, and if the autocovariance properties of these variables can be fully described with a finite lag length ($p = 8$ with monthly variables) —CLS in Assumption 1—, then it is possible to use a VAR to consistently estimate the dynamic effects of aggregate demand and supply shocks.

Since the Invertibility imposes the strong assumption that *all* shock must be recoverable from lags of macro variables,[19] the more recent VAR literature moved to a more modest goal

[18] See also sign restrictions to identify all underlying structural shocks (e.g., Furlanetto et al., 2025).

[19] A large literature has explored the plausibility of the Invertibility assumption as well as the consequences of its failure (e.g., Beaudry and Portier, 2014).

of identifying the dynamic effects only one (or a subset) of all possible structural shocks, replacing the Invertibility assumption with the weaker Partial Invertibility assumption (Stock and Watson, 2018): $\boldsymbol{\eta}_{1t} \in \text{span}\{\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots\}$, i.e., only the shock of interest needs to be spanned by variables in the VAR. The focus on "Partial Invertibility" was spurred by independent work on the identification of structural shocks *outside* the VAR (so called, "external instruments" (Stock and Watson, 2008*b*)) for instance from a narrative approach (Romer and Romer, 1989) or from high-frequency windows (Kuttner, 2001). Indeed, if a researcher has available a proxy z_t for the shock of interest, then we can state the (weaker) assumptions underlying a VAR with IVs as

Assumption 2. *Given the economy* (3) *the VAR(p) model for* $\mathbf{y}_t = (y_t, x_t, \mathbf{w}_t)$*, as defined in* (5)*, satisfies the following*

(IV) the instrument $z_t \in \{\mathbf{d}_t\}$ *is given by* $z_t = \gamma\eta_{1t} + \sum_{j=1}^{\infty} \boldsymbol{\Theta}_{z,j}\boldsymbol{\eta}_{t-j}$ *with* $\gamma \neq 0$

(CLS) for a known or estimable finite p we have $\mathbb{E}(\mathbf{u}_t|\mathbf{u}_{t-1}, \mathbf{u}_{t-2}, \ldots) = 0$.

The IV assumption replaces the previous Invertibility assumption. The key feature of the proxy series z_t, or instrumental variable, is that it contemporaneously only depends on the shock of interest.[20] The correct lag length assumption remains the same. Indeed, a mis-specified VAR(p) with a proxy for η_{1t} in $\mathbf{y}_t$ can only consistently estimate dynamic treatment effects up until horizon p (Plagborg-Møller and Wolf, 2021). Consistency of VAR-based impulse response estimates beyond horizon p require correct lag length specification.

With this assumption in hand, the VAR will consistently estimate the IR of interest θ_h. Formally, we have

Proposition 1. *Given the economy* (3)*, the VAR model for* $\mathbf{y}_t = (z_t, y_t, x_t, \mathbf{w}_t)$ *as defined in* (5) *satisfying assumption 2, the VAR impulse response coincides with the true impulse response* θ_h*:*

$$\frac{\boldsymbol{\Psi}_{21,h}}{\boldsymbol{\Psi}_{31,0}} = \theta_h \qquad \text{for all } h = 0, \ldots, p, \ldots, H \ .$$

where $\boldsymbol{\Psi}_{ij,h}$ *is defined in* (7).

Proposition 1 underlies most recent approaches to estimate structural impulse responses with VARs.

While the independent identification of shocks and the use of Partial Invertibility have substantially relaxed the full-invertibility assumption underlying the first generation of structural VARs, the second assumption —Correct Lag Specification— has received comparatively little attention.

[20]Note that we do allow the series to depend on lagged structural shocks as most instruments in the literature are not strictly exogenous (e.g. Fieldhouse and Mertens, 2023).

And yet, the Correct Lag Specification assumption is crucial. While (partial) invertibility ensures that there exists a VAR that can capture the underlying DGP, there is no guarantee that that VAR has finite lag length, and therefore that it is estimable in finite sample. CLS ensures that such a finite lag length VAR exists, i.e., that there exists a finite lag length VAR that can provide a "complete" *statistical* model of all variables in the VAR (in the sense of captuing the entire autocovariance property of $\mathbf{y}_t$). Loosely speaking, CLS requires that the n-VAR(p) postulated by a researcher is a "complete" statistical model of the economy, i.e., that it is rich enough to captures all predictable movements in the variables in $\mathbf{y}_t$, and all remaining unexplained variation is unpredictable noise.

3.2 How strong is the CLS assumption?

CLS is a very strong assumption. To see that, we will take two perspectives: (i) by illustrating the fragility of CLS in a simple New-Keynesian model, and (ii) by contrasting CLS with comparable assumptions in the treatment-effect literature.

CLS in a stylized New-Keynesian model

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

$$\pi_t = \rho_\pi \pi_{t-1} + \lambda x_t + \nu_t \ , \tag{8}$$

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

$$x_t = \rho_x x_{t-1} - \sigma^{-1}(i_t - \pi_t) + \zeta_t \ , \tag{9}$$

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

$$i_t = \rho_i i_{t-1} + \phi_\pi \pi_t + \varepsilon_t \ , \tag{10}$$

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

Note that this 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 can simply estimate a VAR(1) with $(\varepsilon_t, \pi_t, x_t, i_t)$: CLS is satisfied.

The CLS assumption —the VAR(1) is well specified— is fragile however. To see that, consider adding a simple twist to the model by 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

$$x_t = \rho_x x_{t-1} - \sigma^{-1}(i_t - \pi_{t-1}) - \delta \ell_{t-1} + \zeta_t \tag{11}$$

where ℓ_t is the leverage–GDP ratio with

$$\ell_t = \rho_\ell \ell_{t-1} + \beta i_t \tag{12}$$

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 $(z_t, \tilde{\mathbf{y}}_t)$ will deliver consistent IR estimates (where z_t is a proxy for the monetary shock ε_t).

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 $p = \infty$.[21]

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 ℓ_t. Loosely speaking, this is similar to an Omitted Variable Bias, and this example illustrates how positing CLS 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 CLS assumption— can be seen as important as the identification scheme, the "traditional" identifying assumption.[22] 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

[21] To see that, just combine (11) and (12) and iterate backwards to get

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

[22] This example also illustrates how a (well-chosen) VAR can yield substantial efficiency gains, which can be crucial given the short samples often encountered in macro. In this example, the data *are* highly structured —the true DGP is a VAR(1) for $\mathbf{y}_t = (\pi_t, x_t, \ell_t, i_t)$—, and one can do a lot better than semi non-parametric methods (like local projections) in terms of efficiency.

other words, CLS can be satisfied for a four-variable VAR but it is impossible to satisfy for the three-variable VAR without leverage.[23]

CLS vs. Surrogacy

It is also instructive to contrast the CLS assumption with comparable identifying assumptions made in the dynamic treatment effect literature to infer the causal effects of treatment on long-term outcomes. When an experiment or identification strategy only allows to estimate the short-run effects of treatments, it is possible to predict longer-term effects provided that the short-run outcome variables can act as *surrogates* for longer-term outcomes (e.g. Prentice, 1989).[24] Surrogacy was first made explicit in the medical literature by Prentice (1989), with the aim of estimating the impact of interventions on long-term outcomes when only short-term outcomes are available. The surrogacy assumption states that the effect of a specific intervention on the long-run outcome variable can be entirely predicted from a subset of these short-run outcomes: the "surrogates". Restated in our context the surrogacy assumption becomes

Assumption 3.

$$\text{SUR(p)} \qquad \mathbb{E}(\mathbf{y}_{t+h}|\mathbf{y}_{t+1:t+p}, \eta_{1t}) = \mathbb{E}(\mathbf{y}_{t+h}|\mathbf{y}_{t+1:t+p}) \ , \qquad \textit{for} \ \ h = p+1, \ldots, H \ ,$$

where $\mathbf{y}_{t+1:t+p} = (\mathbf{y}_{t+1}, \mathbf{y}_{t+1}, \ldots, \mathbf{y}_{t+p})$.

Conditioning on the first p leads of the outcome variable $\mathbf{y}_t$, the shock has no further effect on the outcomes. In words, knowing the short-term response of $\mathbf{y}_t$ to the shock $\eta_{1,t}$ over horizon 1 to p is sufficient to trace out the *entire* dynamic effect of the shock: the expected long-term effect of the shock can be entirely characterized by its short-term (up until horizon p) effects.

To help understand the Surrogacy assumption, it is useful to restate it as a decomposition of the long-term causal effect. Specifically, the response at horizon h (θ_h) can be decomposed as the product of a short-term causal effect and a "short-term to long-term" prediction term. Specifically, we have the Corollary

[23]This is *not* to say that a finite-p VAR for $(\varepsilon_t, \pi_t, x_t, i_t)$ cannot provide valuable estimates of the effects of monetary shocks. Even if the VAR is mis-specified, a well-chosen finite-p VAR can help strike an attractive bias-variance trade-off for IR estimates. In fact, this is precisely the message of the paper —a VAR should be seen a power tool— and the subject of the next section.

[24]In the micro literature the identification of dynamic treatment effects when long-run outcomes are not directly measured rests on three assumptions: (i) Unconfoundedness, (ii) Surrogacy and (iii) Comparability, see Athey et al. (2019); Chen and Ritzwoller (2023) The "shock identification" assumption in macro is the "unconfoundedness" assumption found in the treatment effect literature —treatment is independent of potential outcome variables (possibly conditional on covariates)—, (ii) Correct Lag Specification maps with the Surrogacy assumption, and (iii) Comparability is automatically satisfied in a stationary time series context.

Corollary 1 (Surrogacy and IR decomposition)**.** *The Surrogacy assumption* SUR(p) *implies*

$$\theta_h = \boldsymbol{\beta}_h' \boldsymbol{\tau}_{t+1:t+p} \tag{13}$$

where $\boldsymbol{\tau}_{t+1:t+p} = \mathbb{E}(\mathbf{y}_{t+1:t+p}|\eta_{1t} = 1) - \mathbb{E}(\mathbf{y}_{t+1:t+p}|\eta_{1t} = 0)$ *is the short term causal effect of the shock on* $\mathbf{y}_t$ *and* $\boldsymbol{\beta}_h$ *is the projection coefficient vector in* $\mathbb{E}(y_{t+h}|\mathbf{y}_{t+1:t+p}) = \boldsymbol{\beta}_h' \mathbf{y}_{t+1:t+p}$.

In words, the causal effect at horizon h is the product of (i) $\boldsymbol{\tau}_{t+1:t+p}$, the causal effects of the shock on all variables in the VAR from horizons 1 to p times (ii) $\boldsymbol{\beta}_h$, the coefficients of a predictive regression of long-term outcomes on short-term outcomes.

The VAR CLS assumption is stronger than the surrogacy condition: the CLS Assumption implies the Surrogacy assumption, but the converse is false.

Proposition 2.

$$\text{CLS} \;\Rightarrow\; \text{SUR(p)} \qquad \textit{but} \qquad \text{CLS} \;\not\Leftarrow\; \text{SUR(p)}$$

Compared to Surrogacy, the VAR CLS assumption 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.

Restated in terms of the long-term effect decomposition (28), CLS also imposes that (i) the long-term effect decomposition holds for all variables in the VAR in response to any shock, and (ii) the prediction coefficient vector satisfies the VAR-recursion $\boldsymbol{\beta}_h = \mathbf{A}(L)\boldsymbol{\beta}_{h-1}$.

In sum, the strong CLS assumption implies that the VAR(p) must be a complete *statistical* model of the macroeconomy, in that it must includes all endogenous and exogenous variables that help predict the macro variables of interest, i.e., the VAR(p) must fully characterize the autocovariance structure of $\mathbf{y}_t$. 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.[25]

3.3 The pitfalls of conventional lag length selection methods

In line with the "VAR-as-DGP" perspective, traditional lag length selection methods (see the appendix for a brief review) start from the premise that there exists a finite-p VAR that does capture the underlying DGP. For instance, under regularity conditions the BIC asymptotically selects the true lag order *if* the true data generating process admits a VAR

[25] 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 6.

representation with finite lag length (e.g., Kilian and Lütkepohl, 2017). More generally, available lag selection methods have one common feature: they are all 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 CLS assumption fails —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.

We can illustrate this point with our simple introductory example of Section 2.1 and restated for convenience:

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

It is clear that the true lag length $p_0 = 8$ correctly captures the underlying DGP and the IR of interest. But what lag length will BIC select?

For T large, we have

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

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.

4 Take-away

The VAR-as-DGP perspective needs very strong assumptions to hold, making standard VAR specification methods fragile to mis-specification and not optimized to "best" estimate impulse responses of interest. In the next section, we thus propose a different perspective —the VAR-as-Power-Tool—, where the objective of the VAR is *not* to provide a complete statistical model of the macro economy. Rather, the purpose of the VAR is to provide a

dimension reduction tool to best estimate the dynamic causal effects of macro shocks.

5 The VAR-as-Power-Tool perspective

Why do VARs require such stronger assumptions than found in the treatment effect literature? Intuitively, the main reason is that a VAR(p) has a very ambitious aim: to model the *entire* unconditional autocovariance structure of $\mathbf{y}_t$ with only p autocovariance matrices. This requires to include in the VAR all possible factors that affect the macro-economy, a likely elusive objective that makes VAR users exposed to mis-specification risk.

In that context, we propose to consider the VAR from a different perspective: not as a model of the true underlying DGP, but instead as a power tool; a tool that can be tailored to the research task at hand, here the estimation of dynamic causal effects. In line with that perspective, model selection should target the objective of the researcher, and we will propose a VAR specification selection procedure that can be used to select both the number of lags p as well as the set of endogenous variables to include in the VAR.

The simple insight underlying our approach is that, instead of seeing a VAR as a model of the economy, we can equivalently see a VAR as a parametrization of the impulse responses to shocks. From that perspective, we will see that a mis-specified VAR *can* be a powerful tool to estimate an IR of interest, *even if* the VAR provides a poor description of the economy as a whole. Instead, the key is that the VAR restrictions are can parsimoniously capture *the* IRs of interest.

5.1 VAR and IR parametrization

Instead of viewing a VAR as a representation of the economy, we can equivalently see a VAR as a representation of the IRs to shocks, that is as a parametrization of the coefficients of the underlying SVMA (3). Specifically, we will now see how a VAR model restricts the shape of the IRs —across variables, across horizons and across shocks—, and how increasing p and n relaxes these restrictions.

IR restrictions across variables, horizons and shocks

VARs are systems of linear difference equations —parametrizations of the law of motion of the endogenous macro variables—. As such a VAR(p) imposes strong restrictions on the shapes of the impulse responses to shocks. To see that, multiply (5) by a shock $\varepsilon_{s,t-h}$ (the

sth element of the h-period lagged shock ε_t) and take expectations to get[26]

$$\mathbf{\Psi}_{s,h} = \begin{cases} \mathbb{E}(\mathbf{y}_t \varepsilon_{s,t-h}) & \text{if} \quad h \in \{0, ..., p\} \\ \mathbf{A}_1 \mathbf{\Psi}_{s,h-1} + \ldots + \mathbf{A}_p \mathbf{\Psi}_{s,h-p} & \text{if} \quad h \in \{p+1, \ldots, H\} \end{cases} \tag{15}$$

which states that the estimated IR of $\mathbf{y}_t$ to shock s —$\mathbf{\Psi}_{s,h}$— satisfies the system of difference equation (15) with initial conditions given by $\mathbb{E}(\mathbf{y}_t \varepsilon_{s,t-h})$.

Expression (15) highlight two important points: First, the initial conditions imply that the VAR-based IRs do not impose any restrictions on the autocovariances until horizon p. Beyond horizon p, the VAR propagates the effect of the shock through the difference equation, and restrictions are imposed. Second, the same difference equations (14) must hold for *any* shock s. This means that a VAR(p) imposes that the IRs of $\mathbf{y}_t$ to *any* shocks follow the *same* difference equation.[27] In other words, the VAR imposes restrictions on the IRs across variables, across horizons and across shocks.

We can unpack this last observation, as the VAR efficiency gains come from two separate angles: the restrictions across horizons and variables allow to reduce IR dimension problem from nH to pn^2, while restrictions across shocks means that it is possible to use the unconditional series to estimate the IRs.[28]

IRs as combinations of damped cosines

To better understand how the VAR restrictions affect the shape of VAR-based IRs, we can explicitly solve the difference equations (14). Formally, the solution to the p-th difference equation (14) is given by

$$\mathbf{\Psi}_{s,h} = \sum_b \mathbf{c}_b \rho_b^h \cos(2\pi h/\tau_b + \varphi_b) \tag{16}$$

with $\mathbf{\Psi}_{s,h}$ the IR of $\mathbf{y}_t$ to shock s at horizon h, $\rho_b < 1$ the geometric decay term, τ_b the period of the cosine, φ_b its phase, and $\mathbf{c}_b$ the $(n \times 1)$ vector of loadings on the basis function of frequency $2\pi/\tau_b$. See the appendix for details of the derivation.

In words, positing a VAR consists in parametrizing the impulse responses with specific basis functions —decaying cosines—, and a VAR(p) will model the IRs with at most $np/2$

[26]Specifically, we have

$$\mathbb{E}\mathbf{y}_t \varepsilon_{s,t-h} = \mathbf{A}_1 \mathbb{E}\mathbf{y}_{t-1} \varepsilon_{s,t-h} + ... + \mathbf{A}_p \mathbb{E}\mathbf{y}_{t-p} \varepsilon_{s,t-h} + \mathbf{B} \underbrace{\mathbb{E}\varepsilon_t \varepsilon_{s,t-h}}_{=0} \tag{14}$$

[27]Representation (15) can be seen as a *conditional* Yule-Walker equation, where the autocovariances of the originalYule Walker equation are conditional on a particular structural shock.

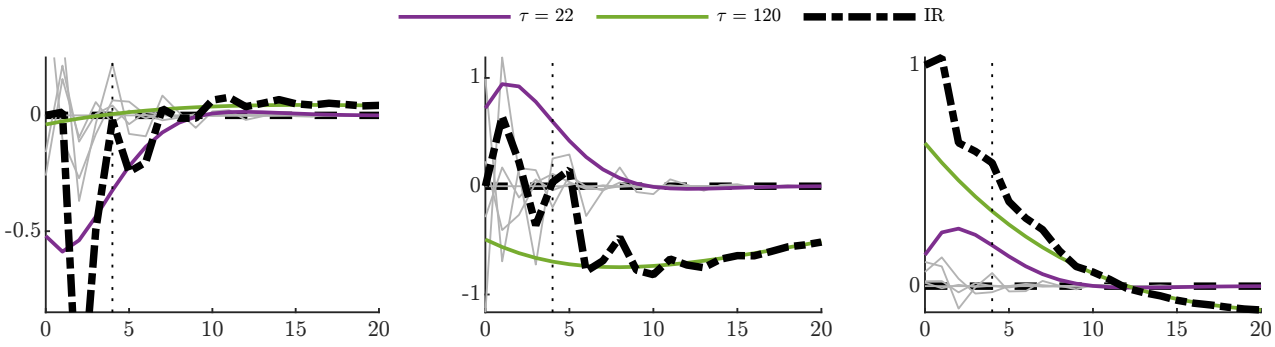
[28]Without restrictions across shocks, the variance available would be restricted to the variance explained by the shock of interest alone. So if shocks account for 20 percent of the variance of output say, this means the VAR is able to exploit 5 times more variation than a LP-IV estimator.

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.

Basis function interpretation of the monetary VAR To illustrate this basis function interpretation, we first go back to our empirical exercise with a monetary VAR. Figure 11 decomposes the IRs estimated from our monetary VAR(4). It shows all basis functions in grey, 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. First, a short lag length VAR *can* capture persistent impulse responses, *even* slowly building cycles (contrary to earlier statements (Kilian and Lütkepohl, 2017, , p64)).[29] Here, 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.

Figure 11: 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.

In fact, at longer horizons the impulse responses are entirely driven by only one damped cosine: 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

[29]This 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.

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. In this application, the *same* cosine with period $P = 120$ quarters (green line) drives the rebound of GDP growth, the mean reversion of inflation, *and* the mean reversion of the fed funds rate.

Of course, as we increase the number of VAR parameters (n or p), the number of basis functions increase and the VAR is able to capture more complex IR structures. For illustration, Figure 14 plots the basis functions from the VAR(16) underlying the IRs estimated in Figure 2.

5.2 Conditional Out-Of-Sample (COOS) model selection

Based on the discussion above; if we consider the VAR as an approximation to some underlying, more general and complicated, data generating process, selecting the lag length should be approached differently. Since recovering the true model is infeasible, the lag length or set of endogenous variables is better selected based on the objective of the researcher.[30] In addition to targeting a specific object, different researchers may care about different performance metrics, e.g. mean squared error, or bias or even coverage. Ideally, the model selection procedure would select the best model for a given objective and performance criteria.

To achieve this, we build on the out-of-sample forecast evaluation literature. To set the stage, we first briefly describe the standard lag selection approach based on recursive mean-squared predication error.

Suppose we have a sample consisting of T observations and candidate VAR models of order $p \in 0, ..., p^{\max}$. 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 $\mathcal{T}$.

The recursive mean-square prediction error (MPSE) of the VAR over the validation sample $\mathcal{V}$ is then the average of its squared recursive prediction errors.[31] The MPSE selection procedure then consists in choosing the lag order corresponding to the model with the lowest

[30] In this paper, we focus on impulse response estimation, but the principle applies to other goals as well such as constructing a variance decomposition or a constructing forecast.

[31] While we focus on MPSE, other metrics are possible, for instance mean absolute error. Similarly, while we describe a recursive estimation procedure whereby we progressively expand the estimation sample, a rolling estimation sample could also be used.

recursive MSPE or

$$p^*_{\text{mpse}} = \underset{p\in\{0,\dots,p^{\max}\}}{\operatorname{argmin}} \frac{1}{V}\sum_{j\in\mathcal{V}} \left(y_{j+h|j}(p) - y_{j+h}\right)^2 \ ,$$

where V is the validation sample and $y_{j+h|j}(p)$ is the h-step ahead VAR forecast for variable y.[32]

With this as background, recall that we are interested in impulse response estimation, but the out-of-sample selection procedure sketched above targets the wrong objective (the unconditional forecast error). To fix this this, the key insight is to think of an impulse response as a prediction *conditional* on one shock, while an unconditional forecast is a prediction based on all shocks affecting the system. To isolate the prediction based on the shock of interest we can then project the VAR forecast error on the identified shock, i.e., compute a *conditional* forecast error: the difference between the VAR-predicted effect of a shock —the effect predicted by the VAR estimated over the training sample— and the average realized effect of the shock over the validating sample. We then select the model which produces the smallest out-of-sample conditional forecast errors. [33]

With z_t a proxy for the shock of interest (see Assumption 2 for formal definition), and with $y_{j+h|t}$ and y_{j+h} a sequence of forecasts and realized outcomes over the validation sample $\mathcal{V}$, we can choose the lag length from

$$p^*_{\text{coos}} = \underset{p\in\{0,\dots,p^{\max}\}}{\operatorname{argmin}} \frac{1}{V}\sum_{j\in\mathcal{V}} \left(z_j\left(y_{j+h|j}(p) - y_{j+h}\right)\right)^2 \ .$$

We will refer to this model selection procedure as the Conditional Out-Of-Sample criterion or COOS.[34]

To get some intuition on this selection method, note that the projection term $\sum_{j\in\mathcal{V}} z_j y_{j+h}$ is identical (modulo a scaling factor) to the local projection coefficient estimate $\sum_{j\in\mathcal{V}} z_j y_{j+h} / \sum_{j\in\mathcal{V}} z_j^2$. As a result, the COOS procedure can be seen as comparing the VAR-based IR estimated

[32]This procedure may be adapted to allow for forecasts of more than one VAR variable by evaluating 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.

[33]The criteria for selecting p is related to the notion of conditional predictive ability proposed in Giacomini and White (2006). Adopting their approach in our setting amounts to projecting the squared forecast error on z_t, i.e. they would select p^* by minimizing $\sum_{s=S}^{n-H} z_s(y_{s+h} - y_{s+h|s}(p))^2$. This does not directly target the mean squared error of the impulse response however.

[34]As with any out-of-sample forecast evaluation procedure, the COOS procedure requires choosing a sample split point and whether to use a rolling or increasing estimation window for parameter estimation. To reduce the influence of this choice we may iterate the procedure over many possible sample split points and take the average impulse response, see also e.g. Rossi and Inoue (2012) for similar ideas in the forecasting literature.

over the training sample $\mathcal{T}$ with the LP-based IR estimated over the validation sample $\mathcal{V}$:[35]

$$p^*_{\text{coos}} = \underset{p\in\{0,\dots,p^{\max}\}}{\text{argmin}} \left(\hat{\theta}^{\text{var}}_{\mathcal{T}} - \hat{\theta}^{\text{lp}}_{\mathcal{V}}\right)^2 .$$

The first term is the average effect of shock as *predicted* by a VAR estimated over the training sample, while the second term is the *realized* average effect of the shock over the validation sample.

Alternatively, the COOS procedure can also be seen as an out-of-sample Hausman-type specification test; comparing the efficient but potentially inconsistent VAR-based IR estimate with the consistent but potentially less efficient LP-based IR estimate. However and this is key, the comparison uses the VAR and LP estimates from separate samples. Otherwise, the approach would lead to exact fitting (asymptotically); the procedure wanting to increase the lag length in order to fit the LP exactly, leading to over-fitting and canceling any VAR efficiency gains. Comparing estimates from two separate samples avoids that pitfall.

Importantly, while we presented the COOS method as a procedure to select the VAR lag length, we can apply the exact same approach to select which variables to include in the VAR. This is particularly useful since, unlike with lag selection methods, there are few formal procedure to select the set of variables to include in the VAR, and practitioners have little guidance.[36]

Specifically, we seek to determine whether a particular vector of variables $\mathbf{y}_t$ (from a set of vector candidates $\mathcal{Y}$) should be included in the VAR. We index each model with a subscript i , and let $y_{j+h|j}(p,i)$ be the h-period ahead forecast for a VAR(p) model with variables according to label i. A COOS algorithm can then be run for each lag length p *and* each vector candidate i. The algorithm below formally describe the COOS model selection procedure.

Algorithm COOS: As preliminary inputs choose the max lag length to be considered $p^{\max}$, the set $\mathcal{Y}$ of endogenous variables to consider for $\mathbf{y}$, the maximum horizon of the IR of interest H and a splint point S to define the training sample $\mathcal{T}$ and the validation sample $\mathcal{V}$ for the out-of-sample forecasts.

□ for each lag length $p = 0, \dots, p^{\max}$ and each candidate vector $\mathbf{y}_i \in \mathcal{Y}$

[35]Equivalently, the COOS procedure can be seen as comparing the local projection IR coefficients estimated on the realized data in sample $\mathcal{V}$ with a local projection IR coefficients estimated on *forecasted data* generated by a VAR(p) estimated on the sample $\mathcal{T}$.

[36]One approach relies on the concept of Granger causality, which can be seen as a test of the CLS assumption underlying the VAR: if the VAR is correctly specified and CLS holds, Granger causality tests should reject any residual predictability (e.g. Jarociński and Maćkowiak, 2017). This approach is rarely used in practice, with practitioners rarely discussing their choice of endogenous variable.

- for each $j \in \mathcal{V} = S+1, ..., T-H$
  - Estimate the VAR(p) using observations $\mathbf{y}_1, \ldots, \mathbf{y}_j$.
  - Compute the forecasts $y_{j+1|j}(p), \ldots, y_{j+H|j}(p)$.
- For each h project the forecasts $\{y_{j+h|j}(p)\}_{j\in\mathcal{V}}$ and observations $\{y_{j+h}\}_{j\in\mathcal{V}}$ on $\{z_j\}_{j\in\mathcal{V}}$ and denote the projected series by $\hat{y}_{j+h|j}(p)$ and $\hat{y}_{j+h}$.
- Compute the evaluation metric $d(\hat{y}_{j+h|j}(p,i), \hat{y}_{j+h})$ for each j, h pair (for instance the squared distance), and aggregate across j, h as desired to get $\hat{\delta}(p,i)$, e.g. $\hat{\delta}(p,i) = \sum_{j\in\mathcal{V}} \sum_{h=1}^{H} d(\hat{y}_{j+h|j}(p,i), \hat{y}_{j+h})$.

☐ Select $(p^*_{\text{coos}}, \mathbf{w}^*_{\text{coos}}) = \text{argmin}_{\mathbf{w}_i} \hat{\delta}(p)$.

Two-step COOS

In practice, we do not always identify shocks using instruments z_t. A popular alternative is to impose parametric restrictions, e.g. short run restrictions, on the coefficients of the structural VAR. When such identification scheme is used, we cannot directly apply COOS. Instead we first need to construct z_t. The important insight is that z_t can be noisy and does not need to be a perfect measure for the structural shock of interest. In light of this we follow Chang and Sakata (2007) and use a long-VAR with $p^{\max}$ lags to extract a measure for the shock of interest, i.e. η_{1t}, using the preferred parametric identification scheme. With this as step one we can then proceed to execute the COOS algorithm in a second step.

5.3 On the optimal VAR size: a basis functions perspective

In Section 5.1, we saw that the VAR size can be restated as the number of basis functions used to parametrize impulse responses. Effectively, COOS will select the VAR size —the set of basis functions— that best predicts a causal effect of interest, and a different VAR may be "optimal" depending on the variable of interest and depending on the horizon of interest.

This can be seen from the IR basis function decomposition perspective. For certain variables, the impulse response may already be very well approximated by a few damped cosines, in which case a small-p VAR can already provide a very good approximation of the impulse response. Estimating the dynamic effects at longer horizons however may require a larger VAR if the long-run dynamics are richer than a simple dampened oscillation. For instance, larger VARs will be needed to capture strong but delayed overshooting effect. Similarly, while the dynamic response of some variables may be well approximated with only a few damped cosines, some variables with complicated impulse response features may only be described with a large number of damped cosines and as such would require a large number of lags.

The key point is that an optimal use of the VAR power tool may require the estimation of multiple VARs, each optimized for a specific variable and horizon of interest. Although quite different from the current practice, this is a natural implication of the VAR-as-power-tool perspective. In fact, we will shortly see that the same conclusion holds for the set of endogenous variables in the VAR; each impulse response of interest may have a different "optimal" VAR specification with p lags and n endogenous variables.

The same intuition holds for using COOS to select the set of endogenous variables, though in that case the difference equation is more intuitive.

5.4 Monte Carlo simulations

In the appendix, we describe a set of Monte Carlo simulations to assess and compare the performances of COOS against popular model selection procedures.

We find that lag length selection using COOS generally gives the smallest errors. Exceptions are the settings where (i) the true β and σ_x are so close to zero that their influence can be ignored, and (ii) the true β and σ_x are large such that AIC can detect the 8th lag. For the mis-specified VAR of Carriero, Clark and Marcellino (2015) the results are shown in Table 8. Here we find that in all settings where the VAR is mis-specified, i.e. $\alpha \neq 0$, the COOS criteria leads to the most accurate impulse response estimates.

5.5 Revisiting the monetary VAR

We can now revisit our monetary VAR application and use COOS 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 5.5 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: COOS lag length selection — Monetary

Quarterly	Average	Long run ($H = 20$)	AIC	BIC
π	7	15	20	1
y	1	7	20	1
i	1	11	20	1
$\pi + y + i$	7	7	20	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 COOS 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 COOS 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 5.5 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.

6 Shrinkage for parameter estimation

Having outlined lag length and variable selection we turn to parameter estimation. So far our simulation and empirical results have used conventional least squares methods to estimate the VAR parameters. This was done to abstract away from other choices, such as priors, penalties and tuning parameters. In general, introducing some form of regularization can be useful for estimating impulse responses and the choices for regularization will interact with the other choices for lag length and variables.

A large part of the macro VAR literature has adopted Bayesian methods for inference

Table 2: COOS 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

in VAR models, see for instance Kilian and Lütkepohl (2017, Chapter 5). Indeed, given that stable macro samples are relatively short when compared to the complexity of the VAR model, it can be attractive to impose additional structure when estimating the VAR parameters. A Bayesian approach introduces such structure by means of specifying prior densities for the parameters of the VAR. For example, priors that shrink least squares estimates toward specific values may help reduce the variance of the conventional OLS estimators. Bayesian methods can be a useful tool for formally incorporating such additional information.

Bayesian VAR with large p

In practice, researchers often select a large lag length — to avoid violating the CLS assumption — and impose a prior on the VAR coefficients to reduce parameter estimation uncertainty. The attractiveness of this approach is clearly demonstrated in the works of De Graeve and Westermark (2025) and Baumeister (2025) who show that for various simulation designs imposing a large lag length out-performs using a default short lag length and also more non-parametric local projections. Improvements are documented for both point estimation accuracy and coverage. One concern with this approach is that the typical priors are not optimized for impulse response estimation. Instead they are designed for obtaining the best fit or forecast under a particular scenario, e.g. near unit root processes. This is just to say that we might be able to improve upon the "fix a large lag length" approach if the true data generating process is at odds with the chosen prior.

Here we compare a conventional Bayesian approach, which fixes the lag length at $p^{\max} = 20$ and imposes a Minnesota prior, with the same Bayesian estimator where the lag length

is selected by COOS. For comparison purposes we also report the least squares results for COOS that were obtained above. Table 6 shows the results for the bivariate simulation design based on equation (2). Two findings stand out. First, using a shrinkage estimator by means of the Minnesota prior uniformly improves the estimation results. The gains are often large: in the order of 30-50% MSE reductions. Second, fixing the lag length at $p^{\max}$ does not deteriorate the estimates: COOS does not lead to improvements over fixing the lag length.

To investigate the robustness of this finding we consider different parameter values, changing the shape of the impulse response in different ways. Figure 12 shows the shapes considered and the accompanying estimation results are plotted in Table 4 6. We find that the good performance of fixing the lag length depends on the underlying DGP: specifically when the true VAR order is small simply fixing the lag length at $p^{\max}$ results in sizable losses of up to 25% when compared to using COOS.

Bayesian VAR with large n

Pioneering work by Banbura, Giannone and Reichlin (2010) showed that it was possible to use Bayesian methods to estimate large VAR models where the number of variables is in the order of hundreds. Carefully designed priors building on Litterman (1986) were developed that enable reliable estimation. In this section we compare using such "one-for-all" model against using COOS for selecting a possibly much smaller VAR when the interest is in a specific impulse response.

7 Conditional VARs

So far we have approached the problem of specifying the VAR by minimizing the conditional the out-of-sample forecast errors for the variable of interest. The conditioning was on the shock of interest and the VAR forecasts were constructed using least squares or Bayesian methods. The main idea was to obtain a sequence of estimate for error of the impulse response of interest.

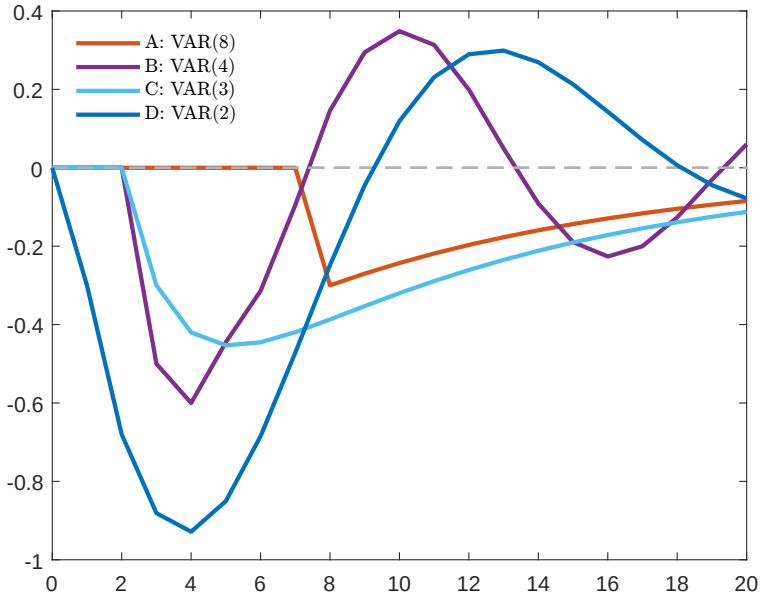
A closely related, yet distinct approach, is based on projecting the observed variables directly on the identified shock and its lags. Given the projected series we can then conduct model selection, i.e. selecting the variables and lag length, using the projected variables that only depend on the shocks of interest. Intuitively, a conventional unconditional approach for VAR inference requires modeling the *entire* unconditional autocovariance structure of $\mathbf{y}_t$ with only p autocovariance matrices. This requires to include in the VAR all possible factors that affect the economy. In contrast, when we are interested in the response of a

Table 3: Simulation results Bayesian VARs

MSE	n	β	σ_x	COOS-OLS	COOS-BAY	LARGE-BAY
	200	-0.3	0.1	0.037	**0.023**	0.025
	200	-0.3	0.2	0.021	**0.013**	**0.013**
	200	-0.5	0.1	0.045	0.029	**0.028**
	200	-0.5	0.2	0.026	0.017	**0.016**
	500	-0.3	0.1	0.017	**0.011**	0.012
	500	-0.3	0.2	0.009	**0.006**	**0.006**
	500	-0.5	0.1	0.018	**0.013**	**0.013**
	500	-0.5	0.2	0.009	**0.007**	**0.007**

MAE	n	β	σ_x	COOS-OLS	COOS-BAY	LARGE-BAY
	200	-0.3	0.1	0.141	**0.112**	0.117
	200	-0.3	0.2	0.106	**0.081**	0.083
	200	-0.5	0.1	0.176	**0.130**	**0.130**
	200	-0.5	0.2	0.124	**0.101**	**0.101**
	500	-0.3	0.1	0.087	**0.073**	0.074
	500	-0.3	0.2	0.061	**0.054**	0.055
	500	-0.5	0.1	0.116	**0.102**	**0.102**
	500	-0.5	0.2	0.079	**0.073**	0.073

Figure 12: Simulation Design — Impulse response sensitivity



Notes: XXX.

Table 4: Simulation results Bayesian VARs - sensitivity

MSE	n	case	COOS-OLS	COOS-BAY	LARGE-BAY
	200	A	0.021	**0.012**	0.013
	200	B	0.026	**0.017**	**0.017**
	200	C	0.032	**0.026**	0.031
	200	D	0.051	**0.037**	0.049
	500	A	0.009	**0.006**	**0.006**
	500	B	0.010	**0.007**	**0.007**
	500	C	0.010	**0.009**	0.011
	500	D	0.020	**0.016**	0.020

specific shock it makes sense to condition on the shock and model directly the *conditional* autocovariances of the data, that is to only fit the impulse response of interest.

We first motivate the conditional VAR with three examples, illustrating how a "conditioning" step help a smaller VAR fit the data and capture the IRs of interest. The following section formally establishes the assumptions underlying our approach.

7.1 Examples

A New-Keynesian model with exogenous oil inflation

Consider again our same baseline NK model, but modified to include energy price inflation in the Phillips curve with

$$\pi_t = \rho_\pi \pi_{t-1} + \lambda x_t + \gamma \pi_{t-1}^{oil} + \nu_t \tag{17}$$

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

$$\pi_t^{oil} = \rho_{oil} \pi_{t-1}^{oil} + \xi_t^{oil} \tag{18}$$

with ξ_t^{oil} an iid shock.

As with leverage, the vector $\tilde{\mathbf{y}}_t = (\pi_t, x_t, \pi_t^{oil}, i_t)$ follows a VAR(1), but the vector $Y_t = (\pi_t, x_t, i_t)$ follows a VAR(∞), and a VAR($p < \infty$) will suffer from omitted variable biases. The problem is that a VAR(1) for Y_t will propagate the monetary shock based on the unconditional autocovariances of π_t, x_t and i_t, but that unconditional covariance is affected by the oil inflation process. To correctly estimate the transmission of monetary shocks, the VAR(1) should hold oil price inflation fixed, i.e., control for it, or else suffer from an omitted variable bias.

This problem is again an omitted variable bias, but of a different kind than with endogenous leverage: it is an omitted *exogenous variable* bias. If the VAR does not capture *all* the

exogenous factors (including shocks) that affect Y_t, a VAR(p) will not consistently estimate the IR to a known shock ε_t.

In other words, in a VAR(1) that omits oil price inflation, the propagation of the monetary shock is *confounded* by the effects of oil price inflation on the economy. Unlike with endogenous variables like leverage however, exogenous variables or shocks do not *mediate* the effect of monetary policy. In that case, we will now show that it is possible to *project out* this confounding.

The key is to realize that the IRs for Y_t *are* in fact described by a simple VAR(1), just not one that can be estimated by OLS. To see that, take derivatives of the model's equations with respect to the monetary shock $\varepsilon_{1,t-h}$, to get:[37]

$$\frac{\partial \pi_t}{\partial \varepsilon_{1,t-h}} = \rho_\pi \frac{\partial \pi_{t-1}}{\partial \varepsilon_{1,t-h}} + \lambda \frac{\partial x_t}{\partial \varepsilon_{1,t-h}} + \gamma \underbrace{\frac{\partial \pi_t^{oil}}{\partial \varepsilon_{1,t-h}}}_{=0} + \underbrace{\frac{\partial \nu_t}{\partial \varepsilon_{1,t-h}}}_{=0} \tag{19}$$

or

$$\mathcal{R}_h^\pi = \rho_\pi \mathcal{R}_{h-1}^\pi + \lambda \mathcal{R}_h^x$$

where $\mathcal{R}_j$ denotes the impulse response of variable j to a monetary shock. Notice how oil price inflation drops out of the equation: any factor that does not mediate the effect of monetary policy drops out. Proceeding similarly for the IS and policy equations, we get

$$\begin{aligned} \mathcal{R}_h^x &= \rho_x \mathcal{R}_{h-1}^x - \sigma^{-1}(\mathcal{R}_h^i - \mathcal{R}_h^\pi) \\ \mathcal{R}_h^i &= \rho_i \mathcal{R}_{h-1}^i + \phi_\pi \mathcal{R}_h^\pi. \end{aligned}$$

Collecting equations, we have a system of the form

$$\mathcal{R}_h^Y = \mathrm{A} \mathcal{R}_{h-1}^Y \tag{20}$$

with $\mathcal{R}_h^Y = (\mathcal{R}_h^x, \mathcal{R}_h^\pi, \mathcal{R}_h^i)'$ and A the matrix of coefficients on the lagged impulse response.[38]

Expression (20) shows that the impulse responses $\mathcal{R}_h^Y$ to a monetary shock do follow a VAR(1): conditional on monetary shocks, the economy follows a VAR(1). In other words, the effect of monetary policy *is* described by a VAR(1) on Y_t, but that VAR(1) cannot be estimated consistently by OLS. We will refer to the VAR(1) describing $\mathcal{R}_h^Y$ as a *conditional* VAR.

While OLS cannot consistently estimate this conditional VAR(1), it is possible to estimate

[37] or take conditional expectations $E(.|\varepsilon_{1,t})$ or $Proj(.|\varepsilon_{1,t})$.

[38] Specifically, we have $\mathrm{A} = \begin{bmatrix} 1 & -\sigma^{-1} & \sigma^{-1} \\ -\lambda & 1 & 0 \\ 0 & -\phi_\pi & 1 \end{bmatrix}^{-1} \begin{bmatrix} \rho_x & 0 & 0 \\ 0 & \rho_\pi & 0 \\ 0 & 0 & \rho_i \end{bmatrix}$.

this VAR by IVs, effectively using lags of monetary shocks as instrumental variables; where the first stage is a regression of Y_t on $\{\varepsilon_{t-h}\}_{h=0}^{\infty}$. Intuitively, projecting on the space spanned by past monetary shocks projects out the exogenous confounding factors that biased the VAR(p) OLS estimates. This approach can be seen as a VAR regression in *impulse response space*, i.e., by regressing the impulse responses on their lags.[39]

In our example, projecting Y_t on past monetary shocks can address confounding issues by projecting out the influence of oil inflation, allowing the VAR to capture the *conditional* auto-covariance matrices of Y_t, which are the covariances of interest. In other words, a first-stage projection allows to shrink the dimension of the VAR needed to capture the IRs of interest. In this case, from a VAR(∞) to a VAR(1).

A mis-specified VAR with local-to-zero coefficients

In practice, it is often hard to determine whether coefficients are actually zero or merely close to zero, or whether there is no serial correlation or whether it is merely small. At first side, one may be inclined to simply treat the small stuff as zero and choose a smaller lag length, but this may be sub-optimal impulse response estimation, or even forecasting, as the total contribution of the small coefficients to the impulse response of interest can be meaningful.

To model "small coefficients" a recent literature has developed methods for mis-specified VAR models where the CLS assumption fails, allowing for serial correlation which is small relative to the sample size (e.g. Schorfheide, 2005; Müller and Stock, 2011; Lohmeyer et al., 2019; Olea et al., 2024; González-Casasús and Schorfheide, 2025). For instance, consider the DGP

$$\mathbf{y}_t = \mathbf{A}_1\mathbf{y}_{t-1} + \ldots + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{u}_t \ , \qquad \mathbf{u}_t = \frac{c}{\sqrt{n}}\sum_{j=0}^{\infty}\mathbf{B}_j\mathbf{e}_{t-j} \ , \tag{21}$$

where $\mathbf{e}_t \sim \text{WN}(0, \mathbf{\Sigma}_e)$. In this VAR($p$) the error term is modeled as a moving average process scaled by $1/\sqrt{n}$.[40] The scaling implies that traditional estimation and lag length selection methods cannot detect the serial correlation in the errors.

By projecting on the space spanned by a shock of interest, we automatically get rid of (project out) these small (hard-to-detect) coefficients and eliminate this source of mis-specification.

[39] Although never used for VARs (i.e., for vector models), this procedure has been implemented to consistently estimate single equations, notably structural macro equations like the Phillips curve, see Barnichon and Mesters (2020); Lewis and Mertens (2022).

[40] An alternative way to set this up is by modeling the larger lags of the VAR(p) as small (e.g. Lohmeyer et al., 2019).

Dealing with one-off events

Often times, the economy is affected by one-off events that can severely distort the time series properties (in finite sample) and push the required VAR specification to longer lags (or call for a VARMA specification as the previous example). Think for instance of the COVIV pandemic, the idiosyncrati price controls of the 70s or banking troubles in the mid-70s, early 80s (Latin American crisis), and again in late 80s (S&L crisis).

Avoiding these dates in the VAR conditioning step automatically addresses these issues.

7.2 Inference in Conditional VARs

Having motivated the intuition underlying a conditional VAR we outline the general approach for conducting inference in a conditional VAR. We are given the selected VAR variables $\mathbf{y}_t$ and separately the instrument series z_t which correlates with the shock of interest. In addition, we postulate that we may have an additional set of what we call *clean up* instruments, denoted by the vector $\mathbf{v}_t$. These instruments are not necessarily structural in the sense that they only correlate with one specific type of structural shock, but they are required to only correlate with $\mathbf{u}_t$ and not with the lags $\mathbf{u}_{t-1}, \mathbf{u}_{t-2}, \ldots$. Intuitively, we will use these instruments to "clean up" the VAR observations by removing shocks that are outside of the economic mechanisms that we want to capture with the VAR.

For clarify it is useful to write the VAR as a linear regression model for the individual VAR variables:

$$y_{it} = \boldsymbol{\gamma}_i' \mathbf{m}_t + u_{it} \ , \tag{22}$$

where $\mathbf{m}_t = (\mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-p}')'$ and $\boldsymbol{\gamma}_i'$ is the ith row of $\boldsymbol{\Gamma} = (\mathbf{A}_1, \ldots, \mathbf{A}_p)$. Recall that throughout we have imposed that all variables are demeaned and note that $\mathbf{m}_t$ is of dimension $pn \times 1$. The general idea of the conditional VAR is to estimate $\boldsymbol{\gamma}_i$ using $z_{t-1}, \mathbf{v}_{t-1}, \ldots, z_{t-h}, \mathbf{v}_{t-h}$ as instruments. The lagged shock proxies should be correlated with the lagged endogenous variables and under the assumption that they have been correctly identified they should be uncorrelated with the shocks u_{it} in setting where (i) u_{it} is serially uncorrelated — i.e. the CLS assumption holds —, but (ii) also when u_{it} is serially correlated but due to exogenous processes. The latter allows to relax the CLS assumption by allowing for violations coming from factors that are orthogonal to $z_t, \mathbf{v}_t$. To satisfy the rank condition we require that $kh \geq pn$ where k is the number of variables in $z_t, \mathbf{v}_t$.

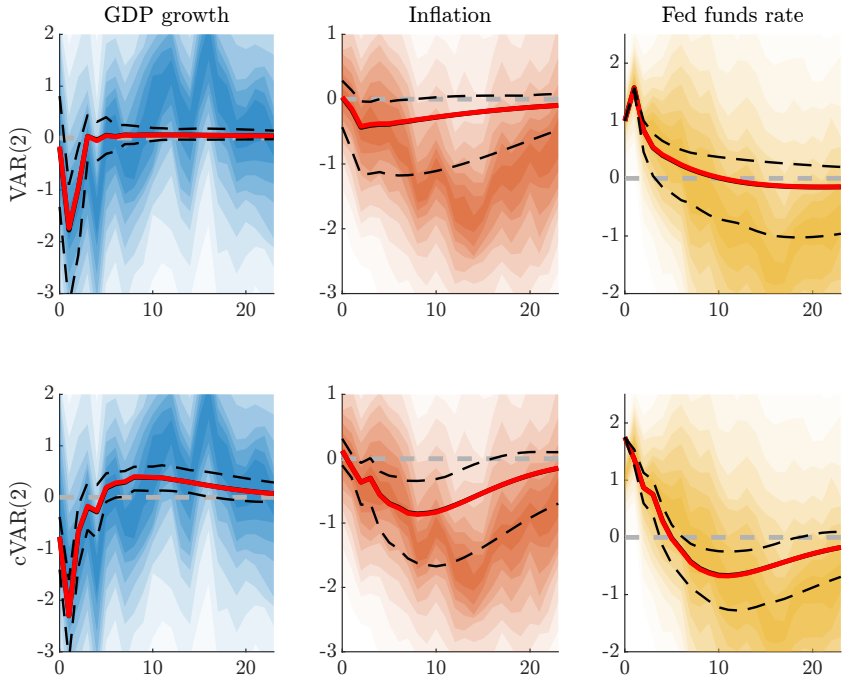
Given the IV estimates $\boldsymbol{\gamma}_i$ for $i = 1, \ldots, n$, we can use the contemporaneous instrument z_t to identify $\mathbf{B}_1$ which maps $\mathbf{u}_t = \mathbf{B}\boldsymbol{\varepsilon}_t$. This second step is identical to SVAR-IV as originally proposed by Stock and Watson (2008*a*) and Mertens and Ravn (2013). The difference is that conditional VAR would estimate the lagged autoregressive matrices using IV.

Trade-off

The conditional VAR can reduce model complexity by projecting out dynamic processes that are orthogonal to the shock of interest. This would result in a lower p and n. The price to pay for this is that less variation is available for parameter estimation: (i) a hard restriction $kh \geq pn$ and (ii) a softer restriction that requires the shock sequences to explain sufficient variation in $\mathbf{m}_t$. For many application it will be hard to balance the trade-off using only z_t as this severely restricts the number of variables and lags that can be considered in the VAR. Therefore introducing smart clean up instruments can be helpful.

The analogy to a structural model is useful. In a structural model only a few sources of exogenous variation appear. This is on purpose to keep the model interpretable and able to provide insights. A conventional VAR instead includes all shocks that affect $\mathbf{y}_t$, many of these shocks are typically not of interest, e.g. variation due to covid, the weather, etc, and it can be helpful not to consider them in order to relax the CLS assumption.

Figure 13: Illustration: Effect of monetary policy, VAR vs cVAR



Notes: 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. Upper row: thick black lines are the point-estimates from a VAR(2) with (dy, π, i). Bottom row: thick black lines are the point-estimates from a VAR(2) with (dy, π, i). Dashed lines mark the 95 percent confidence bands.

8 Conclusion

Macroeconomic VARs aim to capture the economy's full autocovariance structure. Yet VARs used for impulse response analysis are typically small systems with few variables and lags, and the estimated impulse responses are highly sensitive to lag length and variable choice, indicating misspecification. We argue that VARs used for impulse response estimation should be framed around a more humble objective: the VAR only needs to capture the conditional autocovariance structure of the data. Building on this insight, we propose Conditional Out-of-Sample selection (COOS), 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.

Appendix

Looking at the raw data: the conditional distribution of forecast errors

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

First, to construct forecasts, we 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 regress $Y_t^{\perp}$ on ε_t or

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

or

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

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, that is

$$\frac{Y_{t+h}^{\perp} - Proj(Y_{t+h}^{\perp}|\varepsilon_t)}{\varepsilon_t}.$$

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,[41] 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

$$\mathbf{Y}_h = \mathbf{P}\mathbf{D}^h\mathbf{P}^{-1}\mathbf{Y}_0 \tag{23}$$

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 (23), 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 $\mathcal{R}_b =$

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

$\arccos\left(Re(\lambda_b)/R_b\right)$ 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

$$\mathcal{R}_{is}(h) = \sum_{b=1}^{B} 2R_b^h \left(\cos(\frac{2\pi}{T_b} h + \varphi_b)\right) \tag{24}$$

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.[42]

Expression (24) 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 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.

Conventional 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

[42] To obtain (24), 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 $\varphi_b = 0$.

likelihood ratio test

$$\mathrm{LR}(p) = n[\log(\det(\widehat{\mathbf{\Sigma}}_u(p))) - \log(\det(\widehat{\mathbf{\Sigma}}_u(p+1)))] \ , \tag{25}$$

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

$$Q_c = n \sum_{j=1}^{c} \mathrm{Tr}(\widehat{\mathbf{\Gamma}}_j' \widehat{\mathbf{\Gamma}}_0^{-1} \widehat{\mathbf{\Gamma}}_j \widehat{\mathbf{\Gamma}}_0^{-1}) \ , \tag{26}$$

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

$$\mathrm{C}(p) = \log(\det(\widehat{\mathbf{\Sigma}}_u(p))) + c_n f(p) \ , \tag{27}$$

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 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.

Evaluating model selection procedures: some Monte Carlo simulations

In this section, we discuss the results from Monte Carlo studies that compare the performance of different lag length selection procedures.

Simulation designs

We consider three simulation designs: (i) the bivariate VAR from (2), (ii) the 6 variable mis-specified VAR from Carriero, Clark and Marcellino (2015), which was also adopted in González-Casasús and Schorfheide (2025), and (iii) the dynamic factor model design of Li, Plagborg-Møller and Wolf (2024*b*).

Bivariate VAR First we consider the baseline VAR given by (2). We fix $\rho = 0.9$ and vary the coefficient $\beta = -0.3, -0.5$. Further we simulate the errors from a normal distribution with mean zero and variances set to $\sigma_w = 1$ and $\sigma_x = 0.1, 0.2$. This specifications, with small coefficients, capture the idea that the monetary policy shock is hard to identify. We consider sample sizes $n = 200, 500$ and or each specification we simulate 1000 datasets.

Mis-specified VAR from Carriero, Clark and Marcellino (2015) The second DGP is taken from Carriero, Clark and Marcellino (2015), which was also adopted in González-Casasús and Schorfheide (2025) and is based on the mis-specified six variable VAR given in

equation (21). We take $p = 1$ and use the same coefficient values as in González-Casasús and Schorfheide (2025). The degree of mis-specification is given by $\alpha = 0, 1, 2$, such that for $\alpha = 0$ we obtain a correctly specified VAR(1). The shock of interest is identified via a choleski decomposition.

DFM from Li, Plagborg-Møller and Wolf (2024*b*) The last DGP is taken from Li, Plagborg-Møller and Wolf (2024*b*) and simulates 6000 different VARs from a dynamic factor model.

Methods and evaluation For each simulated dataset we compute the different lag length estimators discussed above. Given a p estimate we estimate the VAR parameters and compute the impulse response of ε_t^x on w_{t+h} for $h = 0, \dots, H = 20$. We always use conventional least squares estimates to compute the structural impulse response. We evaluate the estimated impulse responses based on average mean squared and absolute error over impulse horizons and simulations. Importantly, we are not interested recovering the true lag length; we are interested in finding the best impulse response estimate.

We compare our conditional out-of-sample (COOS) lag length selection criteria with a conventional local projection, AIC and BIC criteria. To implement COOS we always split the sample at $n/2$ and use the mean squared or mean absolute loss function for selecting p^*.

Results The simulations results for the bi-variate design are summarized in Table 8. We find that lag length selection using COOS generally gives the smallest errors. Exceptions are the settings where (i) the true β and σ_x are so close to zero that their influence can be ignored, and (ii) the true β and σ_x are large such that AIC can detect the 8th lag. For the mis-specified VAR of Carriero, Clark and Marcellino (2015) the results are shown in Table 8. Here we find that in all settings where the VAR is mis-specified, i.e. $\alpha \neq 0$, the COOS criteria leads to the most accurate impulse response estimates.

Proofs

Proof of Proposition 1. The CLS assumption states that $\mathbf{u}_t$ in (5) is serially uncorrelated. The IV assumption implies that u_{1t} only depends on η_{1t} which is respected when considering $\mathbf{u}_t = \mathbf{B}\boldsymbol{\varepsilon}_t$ when we set $\varepsilon_{1t} = \eta_{1t}$ and take $\mathbf{B}$ lower triangular. The impulse responses of $\mathbf{y}_t$ to η_{1t} are given by $\boldsymbol{\Psi}_{1,h} = \mathbf{C}_h\mathbf{B}_1$ where we note that the first column of $\mathbf{B}$ is identified from the variance of $\mathbf{u}_t$. To ensure that η_{1t} has a unit effect on x_t we divide $\boldsymbol{\Psi}_{31,h} = e_3'\mathbf{C}_h\mathbf{B}_1$ by the impact response on x_t: $\boldsymbol{\Psi}_{21,0} = e_2'B_1$. □

Table 5: Simulation results Bivariate VAR

MSE	n	β	σ_x	LP	AIC	BIC	COOS
	200	-0.3	0.1	0.058	0.024	**0.023**	0.037
	200	-0.3	0.2	0.029	**0.020**	**0.020**	0.021
	200	-0.5	0.1	0.060	0.052	0.054	**0.045**
	200	-0.5	0.2	0.032	0.034	0.051	**0.026**
	500	-0.3	0.1	0.022	0.018	0.019	**0.017**
	500	-0.3	0.2	0.011	0.011	0.018	**0.009**
	500	-0.5	0.1	0.022	0.019	0.049	**0.018**
	500	-0.5	0.2	0.012	**0.006**	0.047	0.009

MAE	n	β	σ_x	LP	AIC	BIC	COOS
	200	-0.3	0.1	0.193	0.132	**0.129**	0.147
	200	-0.3	0.2	0.137	0.122	0.124	**0.112**
	200	-0.5	0.1	0.195	0.194	0.204	**0.164**
	200	-0.5	0.2	0.141	0.150	0.199	**0.126**
	500	-0.3	0.1	0.117	0.116	0.123	**0.098**
	500	-0.3	0.2	0.084	0.083	0.120	**0.071**
	500	-0.5	0.1	0.118	0.102	0.200	**0.101**
	500	-0.5	0.2	0.086	**0.060**	0.197	0.075

Table 6: Simulation results Carriero, Clark and Marcellino (2015) VAR

MSE	n	α	LP	AIC	BIC	COOS
	200	0	0.035	**0.008**	**0.008**	0.015
	200	2	0.085	0.057	0.051	**0.050**
	200	4	0.268	0.182	0.188	**0.181**
	500	0	0.013	**0.003**	**0.003**	0.008
	500	2	0.023	0.020	0.019	**0.018**
	500	4	0.054	0.051	0.057	**0.048**

MAE	n	α	LP	AIC	BIC	COOS
	200	0	0.146	**0.068**	**0.068**	0.091
	200	2	0.226	0.183	0.175	**0.172**
	200	4	0.397	0.328	0.332	**0.325**
	500	0	0.089	**0.044**	**0.044**	0.067
	500	2	0.118	0.109	0.103	**0.102**
	500	4	0.182	0.172	0.180	**0.168**

Proof of Proposition 2. Without loss of generality we set $p = 1$ and note that for any $h = p+1, \ldots, H$ we can write

$$\mathbf{y}_{t+h} = \mathbf{A}_1^{h-1}\mathbf{y}_{t+1} + \mathbf{A}_1^{h-2}\mathbf{u}_{t+2} + \ldots + \mathbf{A}_1\mathbf{u}_{t+h-1} + \mathbf{u}_{t+h} \ ,$$

where, due to the CLS assumption, $\mathbf{u}_{t+2}, \ldots, \mathbf{u}_{t+h}$ are all uncorrelated. With this we consider

$$\begin{aligned}
\mathbb{E}(\mathbf{y}_{t+h}|\mathbf{y}_{t+1}, \eta_{1t}) &= \mathbf{A}_1^{h-1}\mathbf{y}_{t+1} + \mathbf{A}_1^{h-2}\mathbb{E}(\mathbf{u}_{t+2}|\mathbf{y}_{t+1}, \eta_{1t}) + \ldots + \mathbb{E}(\mathbf{u}_{t+h}|\mathbf{y}_{t+1}, \eta_{1t}) \\
&= \mathbf{A}_1^{h-1}\mathbf{y}_{t+1} + \mathbf{A}_1^{h-2}\mathbb{E}(\mathbf{u}_{t+2}|\mathbf{y}_{t+1}) + \ldots + \mathbb{E}(\mathbf{u}_{t+h}|\mathbf{y}_{t+1}) \\
&= \mathbb{E}(\mathbf{y}_{t+h}|\mathbf{y}_{t+1})
\end{aligned}$$

That the converse statement does not hold follows immediately as $\mathbb{E}(\mathbf{u}_{t+j}|\mathbf{y}_{t+1}, \eta_{1t}) \neq \mathbb{E}(\mathbf{u}_{t+j}|\mathbf{y}_{t+1})$ when the prediction errors $\mathbf{u}_t$ are serially correlated. □

Proof of Corollary 1. From the surrogacy condition, we can re-write the dynamic causal effect θ_h as

$$\begin{aligned}
\theta_h &= \mathbb{E}(w_{t+h}|\eta_{1t} = 1) - \mathbb{E}(w_{t+h}|\eta_{1t} = 0) \\
&= \mathbb{E}(\mathbb{E}(w_{t+h}|\mathbf{y}_{t+1:t+p}, \eta_{1t} = 1)|\eta_{1t} = 1) - \mathbb{E}(\mathbb{E}(w_{t+h}|\mathbf{y}_{t+1:t+p}, \eta_{1t} = 0)|\eta_{1t} = 0) \\
&= \mathbb{E}(\mathbb{E}(w_{t+h}|\mathbf{y}_{t+1:t+p})|\eta_{1t} = 1) - \mathbb{E}(\mathbb{E}(w_{t+h}|\mathbf{y}_{t+1:t+p})|\eta_{1t} = 0) \ .
\end{aligned}$$

This shows that surrogacy allows to identify the dynamic causal effect of interest in two steps: (i) compute the predictions for w_{t+h} given $\mathbf{y}_{t+1:t+p}$ and (ii) condition this prediction on the shock of interest. Given the linear environment we have $\mathbb{E}(w_{t+h}|\mathbf{y}_{t+1:t+p}) = \boldsymbol{\beta}_h'\mathbf{y}_{t+1:t+p}$, with prediction coefficients $\boldsymbol{\beta}_h$, and the irf can be written as

$$\theta_h = \boldsymbol{\beta}_h'[\mathbb{E}(\mathbf{y}_{t+1:t+p}|\eta_{1t} = 1) - \mathbb{E}(\mathbf{y}_{t+1:t+p}|\eta_{1t} = 0)] = \boldsymbol{\beta}_h'\boldsymbol{\tau}_{t+1:t+p} \ , \tag{28}$$

where $\boldsymbol{\tau}_{t+1:t+p} = \mathbb{E}(\mathbf{y}_{t+1:t+p}|\eta_{1t} = 1) - \mathbb{E}(\mathbf{y}_{t+1:t+p}|\eta_{1t} = 0)$, i.e. the short term causal effect. □

References

Athey, Susan, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 2019. "The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely." National Bureau of Economic Research.

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

Barnichon, Regis, and Geert Mesters. 2020. "Identifying Modern Macro Equations using Old Shocks." *Quarterly Journal of Economics*, 135: 2255–2298.

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

Beaudry, Paul, and Franck Portier. 2014. "News-driven business cycles: Insights and challenges." *Journal of Economic Literature*, 52(4): 993–1074.

Bernanke, Ben S, and Ilian Mihov. 1998. "Measuring monetary policy." *The quarterly journal of economics*, 113(3): 869–902.

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.

Bernanke, Ben S, Mark Gertler, and Mark Watson. 1997. "Systematic monetary policy and the effects of oil price shocks." *Brookings papers on economic activity*, 1997(1): 91–157.

Blanchard, Olivier, and Roberto Perotti. 2002. "An Empirical Characterization of the Dynamic Effects of Changes in Government Spending and Taxes on Output." *The Quarterly Journal of Economics*, 117: 1329–1368.

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.

Carriero, Andrea, Todd E. Clark, and Massimiliano Marcellino. 2015. "Bayesian VARs: Specification Choices and Forecast Accuracy." *Journal of Applied Econometrics*, 30(1): 46–73.

Chang, Pao-Li, and Shinichi Sakata. 2007. "Estimation of impulse response functions using long autoregression." *The Econometrics Journal*, 10(2): 453–469.

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

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

De Graeve, Ferre, and Andreas Westermark. 2025. "Long-Lag VARs." Working paper.

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.

Fieldhouse, Andrew J., and Karel Mertens. 2023. "The Returns to Government R&D: Evidence from U.S. Appropriations Shocks." working paper.

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

Furlanetto, Francesco, Antoine Lepetit, Ørjan Robstad, Juan Rubio-Ramírez, and Pål Ulvedal. 2025. "Estimating hysteresis effects." *American Economic Journal: Macroeconomics*, 17(1): 35–70.

Gali, Jordi. 1992. "How well does the IS-LM model fit postwar US data?" *The Quarterly Journal of Economics*, 107(2): 709–738.

Giacomini, Raffaella, and Halbert White. 2006. "Tests of Conditional Predictive Ability." *Econometrica*, 74(6): 1545–1578.

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. "Misspecification-robust shrinkage and selection for var forecasts and irfs." National Bureau of Economic Research.

Hamilton, James D. (James Douglas), and Ana Maria Herrera. 2004. "Oil Shocks and Aggregate Macroeconomic Behavior: The Role of Monetary Policy: A Comment." *Journal of Money, Credit, and Banking*, 36(2): 265–286.

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

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. 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.

Jarociński, Marek, and Bartosz Maćkowiak. 2017. "Granger Causal Priority and Choice of Variables in Vector Autoregressions." *The Review of Economics and Statistics*, 99(2): 319–329.

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

Kilian, Lutz. 2009. "Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market." *American economic review*, 99(3): 1053–1069.

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.

Kuttner, Kenneth N. 2001. "Monetary Policy Surprises and Interest Rates: Evidence from the Fed Funds Futures Market." *Journal of Monetary Economics*, 47(3): 523–544.

Lenza, Michele, and Giorgio E Primiceri. 2022. "How to estimate a vector autoregression after March 2020." *Journal of Applied Econometrics*, 37(4): 688–699.

Lewis, Daniel J, and Karel Mertens. 2022. "Dynamic Identification Using System Projections and Instrumental Variables."

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

Li, Dake, Mikkel Plagborg-Møller, and Christian K. Wolf. 2024*b*. "Local Projections vs. VARs: Lessons From Thousands of DGPs." forthcoming.

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 and 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.

Marcellino, Massimiliano, James H. Stock, and Mark W. Watson. 2003. "Macroeconomic forecasting in the euro area: country specific versus area-wide information." *European Economic Review*, 47: 1–18.

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-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. "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.

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.

Rossi, Barbara, and Atsushi Inoue. 2012. "Out-of-Sample Forecast Tests Robust to the Choice of Window Size." *Journal of Business & Economic Statistics*, 30(3): 432–453.

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

Sims, Christopher A. 1980. "Macroeconomics and reality." *Econometrica*, 1–48.

Sims, Christopher A. 1992. "Interpreting the macroeconomic time series facts: The effects of monetary policy." *European economic review*, 36(5): 975–1000.

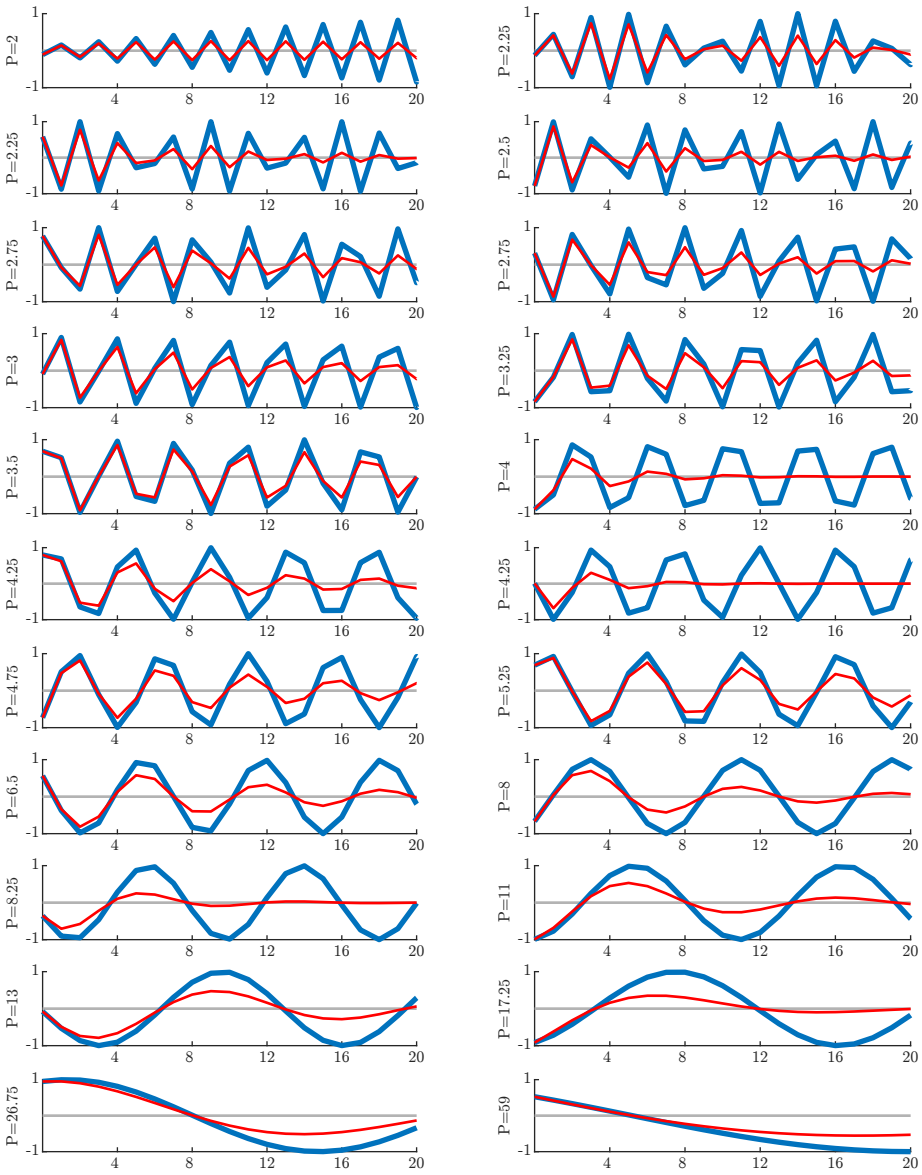
Stock, James H., and Mark W. Watson. 2008*a*. "Lecture 7: Recent Developments in Structural VAR Modeling." Presented at the National Bureau of Economic Research Summer Institute Minicourse: What's New in Econometrics: Time Series, Cambridge, MA, July 15.

Stock, James H., and Mark W. Watson. 2016. "Chapter 8 - Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics." In . Vol. 2 of *Handbook of Macroeconomics*, , ed. John B. Taylor and Harald Uhlig, 415 – 525. Elsevier.

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*b*. "What is new in econometrics: Time series, lecture 7." *Short Course Lectures: NBER Summer Institute. NBER. https://www. nber. org/minicourse_2008. html*.

Figure 14: 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.