Bilkent University
Department of Economics

A Simple Introduction to Vector Autoregressions

Mahmut S. İpek

Bilkent University, Department of Economics

August 27, 2026

Purpose. This note gives an intuitive route from autoregression to structural VARs, impulse responses, and two common identification strategies. It keeps the visual examples and derivations that make the main ideas concrete, while remaining deliberately selective and easy to follow.

1 The question

Monetary economists often want to measure how a policy action affects output, inflation, and interest rates. Three features matter:

- **Magnitude:** How large is the response?
- **Timing:** When does the response appear?
- **Persistence:** How long does it last?

The main difficulty is endogeneity. The policy rate affects the economy, but policymakers also change the policy rate in response to the economy. A simple correlation, therefore, does not isolate a monetary policy shock.

Key idea

A vector autoregression (VAR) summarizes the joint dynamics of several variables. A structural VAR (SVAR) adds assumptions that allow us to interpret some movements as economic shocks.

1.1 What is a shock?

A plastic globe represents the Earth while deliberately leaving out most of its details. Economic models work similarly: they include only the variables needed for their purpose and treat unexpected influences from outside the model as shocks.

A shock is, therefore, an unexpected force that comes from outside the model. In a hypothetical model containing every relevant variable and every causal mechanism, there would be no shock. Constructing such a model, however, is neither possible nor useful.

For example, COVID-19 can be treated as an external shock in a VAR containing only economic variables because it is not explained by the past behavior of those variables. A model containing epidemiological mechanisms might explain part of the same event internally.

In monetary policy research, a monetary policy shock is the part of a policy decision that cannot be predicted from the central bank's usual response to macroeconomic conditions. Such deviations may reflect some unobservables: changing beliefs about the economy, shifting policy preferences, temporary objectives, political considerations, or the persuasiveness of internal rhetoric (Romer and Romer, AER, 2004).

A structural shock is, therefore, not necessarily unexplained in an absolute sense. It is an exogenous and unexpected disturbance relative to the variables and relationships included in the model. Once a shock has been isolated, the next question is how the economy responds to it over time. This is the role of an impulse response function.

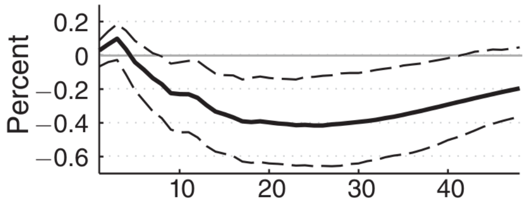
1.2 Our destination: impulse responses

The main object we want to estimate is an **impulse response function (IRF)**. An IRF traces how an economic variable evolves after a shock. In monetary economics, for example, we may ask how output and the price level respond after an unexpected monetary policy tightening.

IRFs help us answer three questions:

- **Magnitude:** How large is the response?
- **Timing:** When does the response appear?
- **Persistence:** How long does the response remain visible?

Figure 1: Ind. prod. response to a MP shock



Source: Figure 1 in Gertler and Karadi (2015). x-axis shows the number of months after the shock. The monetary policy shock is 25 basis points (bps).

An IRF traces the response of a variable over time after a one-time structural shock. The horizontal axis reports the number of periods after the shock, while the vertical axis measures the size of the response relative to the path that would have occurred without the shock.

Reading Figure 1: Following a 25-bps monetary policy shock, industrial production becomes significantly negative after about nine months; before then, the confidence band includes zero. The response reaches its largest decline of ≈ 0.4 percent after two years and then gradually fades; i.e., it returns to zero again around 40-months after the shock.

The following figures preview the type of result that we ultimately want to produce. The arrows connect each variable in the model to its estimated response following a monetary policy shock.

Figure 2: Impulse responses from the three-variable model.

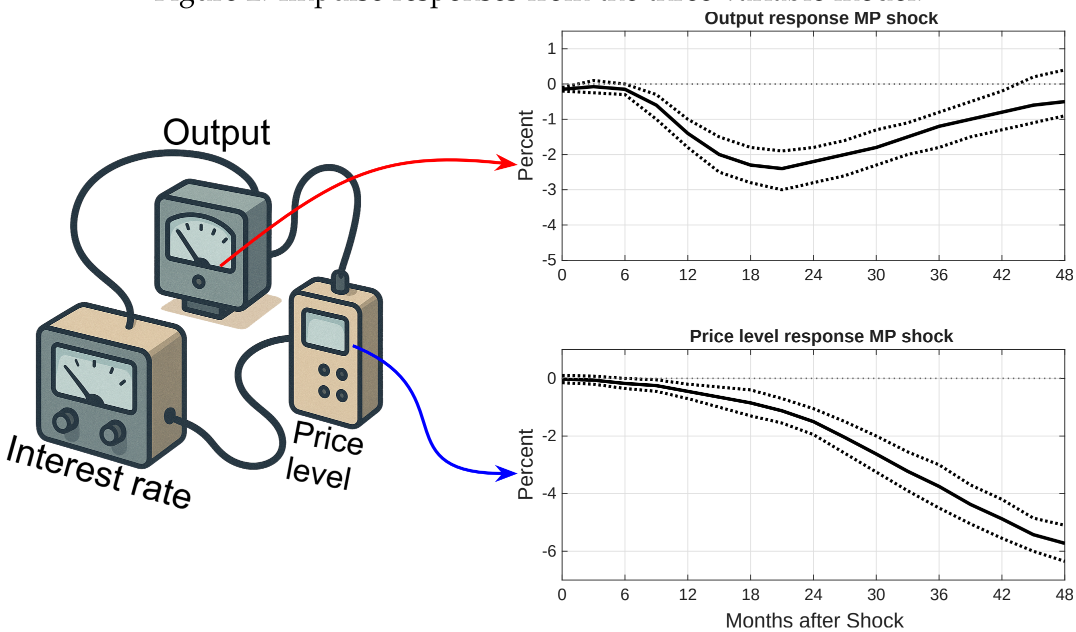
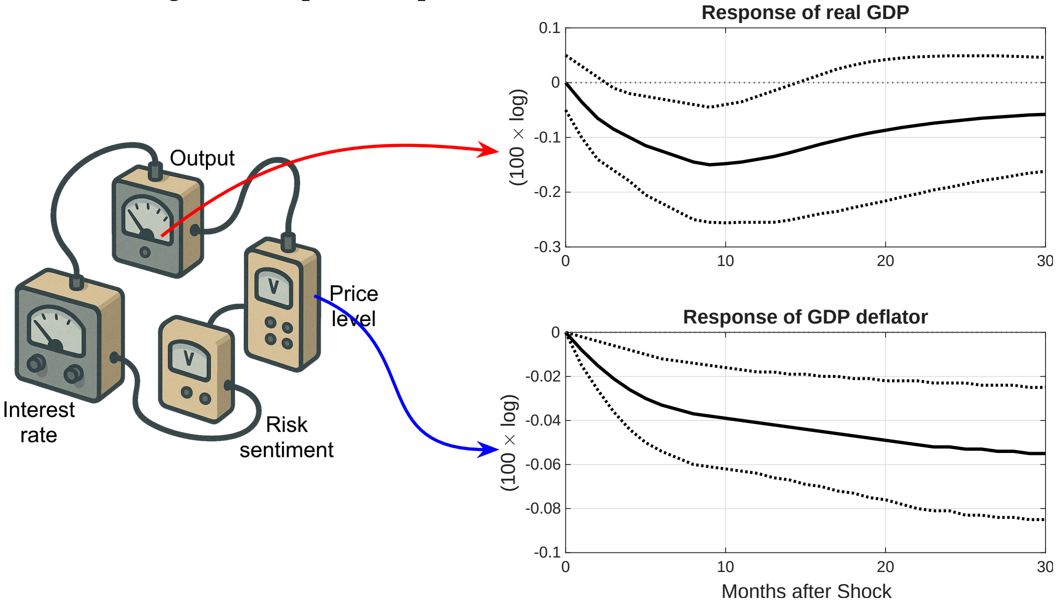


Figure 3: Impulse responses from the four-variable model.



The two models above investigate the same monetary policy question, but they produce different impulse responses. An IRF is not estimated independently of the model: the variables included in the system, their transformations, the lag length, and the identifying assumptions can all affect the result. Once we have developed the methodology, we will be able to understand why different models can produce different IRFs.

2 Start with an autoregression

The name becomes intuitive if we begin with the ordinary meaning of *regress*: to go back.

regress (n.)
late 14c., regresse, "a return, passage back, act of going back," from Latin *regressus* "a return, retreat, a going back," noun use of past participle of *regredi* "to go back," from re- "back" (see re-) + *gradi* "to step, walk" (from PIE root *ghredh- "to walk, go"). More common in legal language. Mental sense of "act of working back from an effect to a cause" is from 1610s.
also from late 14c. *Source:* *etymonline.com*

An early form of linear regression appeared in Newtons work around 1700, the modern least-squares method was later published by Legendre (1805) and Gauss (1809).

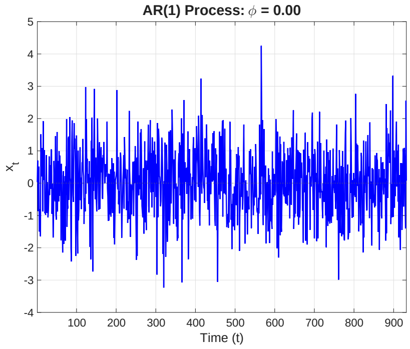
A usual regression and an autoregression differ in what appears on the right-hand side:

$$\begin{aligned}\text{Usual regression:}\quad & y_t = \beta x_t + e_t \\ \text{Autoregression:}\quad & y_t = \phi y_{t-1} + \tilde{e}_t\end{aligned}$$

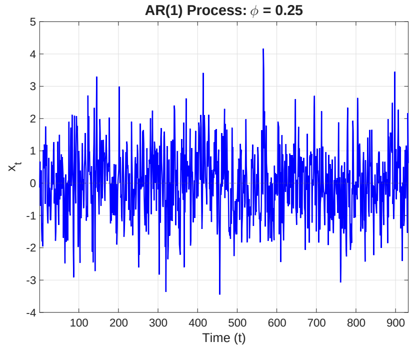
In a usual regression, y_t is related to other explanatory variables collected in x_t. In an autoregression, y_t is related to its own past. The term $\tilde{e}_t$ is the new, unpredictable part, and ϕ determines how persistent the series is.

With a small ϕ, the effect of a new disturbance disappears quickly. As ϕ moves closer to one, disturbances fade more slowly, and the series tends to remain above or below its usual level for longer stretches. If $|\phi| > 1$, disturbances grow rather than fade, so the process is explosive.

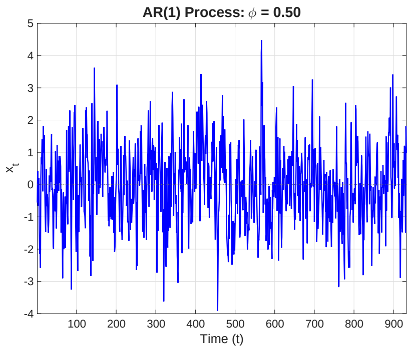
Figure 4: As ϕ approaches one, disturbances become increasingly persistent.



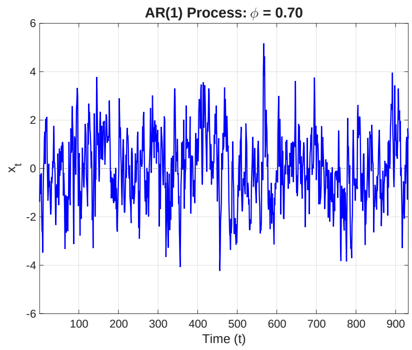
$\phi = 0$: no persistence.



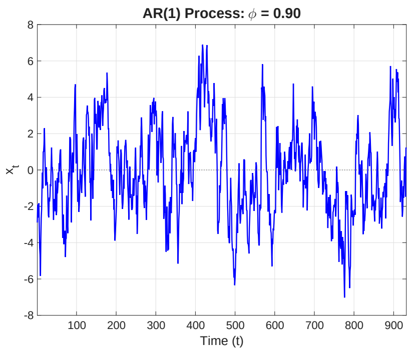
$\phi = 0.25$: short-lived movements.



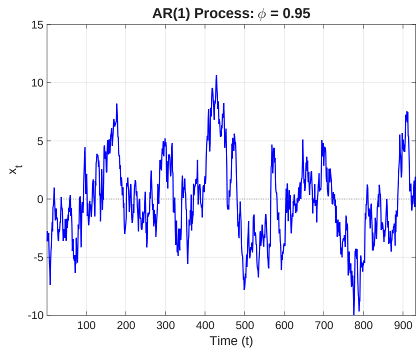
$\phi = 0.50$: moderate persistence.



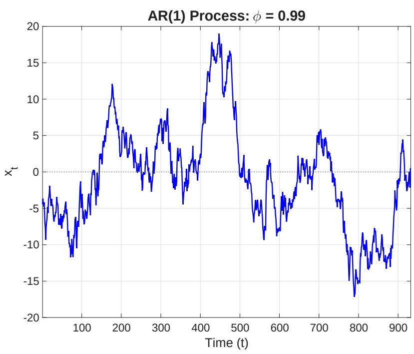
$\phi = 0.70$: longer-lasting movements.



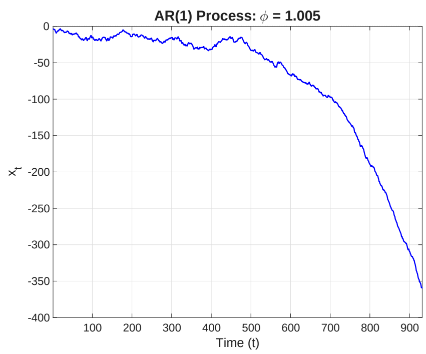
$\phi = 0.90$: disturbances fade slowly.



$\phi = 0.95$: persistent movements.



$\phi = 0.99$: disturbances barely fade.

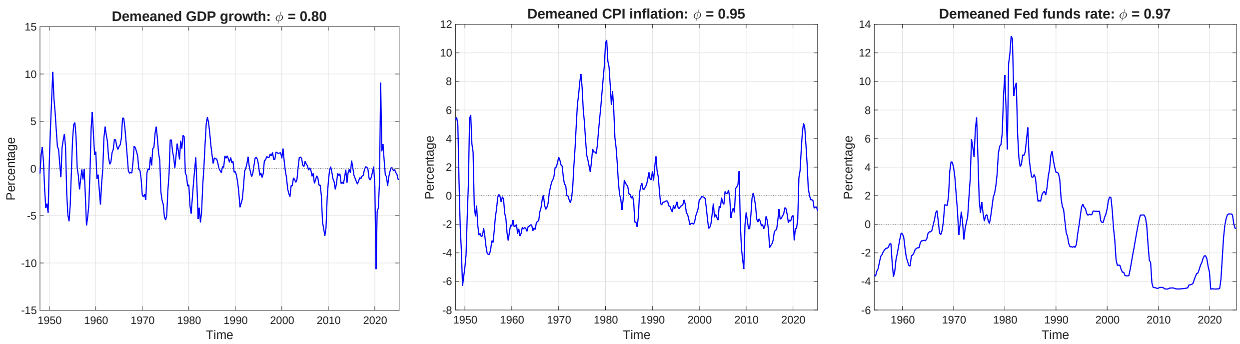


$\phi = 1.005$: explosive movements.

2.1 Persistence in macroeconomic data

Separate AR(1) estimates offer a first descriptive look at the data. In the examples below, output growth is persistent, while inflation and the policy rate are even more persistent.

Figure 5: Estimated persistence differs across macroeconomic variables.



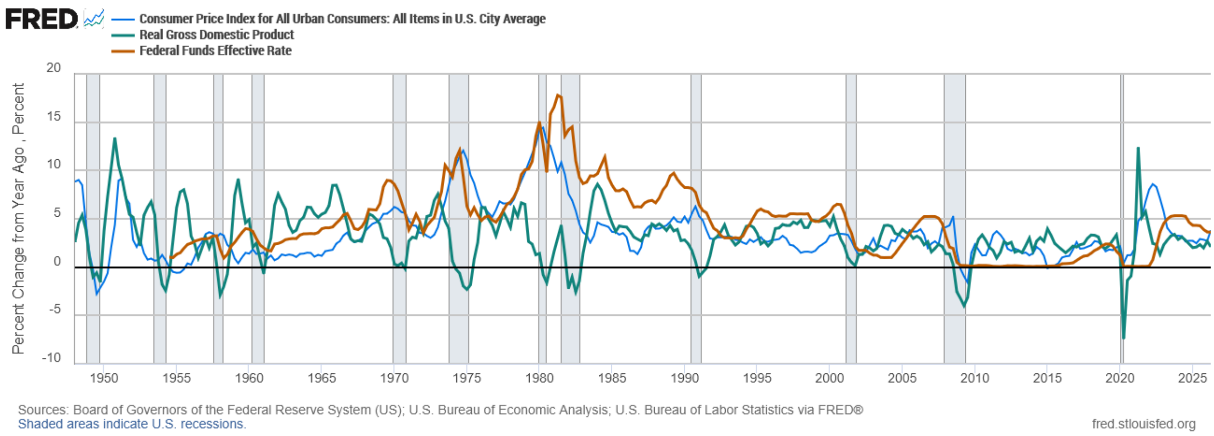
GDP growth: $\widehat{\phi} = 0.80$. Inflation: $\widehat{\phi} = 0.95$. Policy rate: $\widehat{\phi} = 0.97$.

Note: Persistence describes predictability, not causality.

3 From separate autoregressions to a VAR

Y_t is output[1], π_t is inflation, and i_t is the policy rate. The data suggest two features: each series depends on its own history, and the series also move together. This joint dependence motivates a vector autoregression.

Figure 6: A joint view of the macroeconomic series



[1] In the empirical application, Y_t would typically denote $\Delta \log Y_t$. We write Y_t for brevity.

3.1 Guided exercise: from three ARs to a VAR

Start with only one lag. Indeed, all VARs can be written in one-lag[2].

Start with separate autoregressions.

$$Y_t = \phi_Y Y_{t-1} + u_{1,t} \tag{1}$$
$$\pi_t = \phi_\pi \pi_{t-1} + u_{2,t} \tag{2}$$
$$i_t = \phi_i \, i_{t-1} + u_{3,t} \tag{3}$$

These equations describe persistence separately but rule out interdependency.

Add one lag of every variable to each equation.

$$Y_t = \phi_{11} Y_{t-1} + \phi_{12} \pi_{t-1} + \phi_{13} i_{t-1} + u_{1,t} \tag{4}$$
$$\pi_t = \phi_{21} Y_{t-1} + \phi_{22} \pi_{t-1} + \phi_{23} i_{t-1} + u_{2,t} \tag{5}$$
$$i_t = \phi_{31} Y_{t-1} + \phi_{32} \pi_{t-1} + \phi_{33} i_{t-1} + u_{3,t} \tag{6}$$

Each equation now uses the same information set. The terms $u_{1,t}, u_{2,t}, u_{3,t}$ are one-step-ahead reduced-form innovations.

Stack the scalar equations.

$$\underbrace{\begin{bmatrix} Y_t \\ \pi_t \\ i_t \end{bmatrix}}_{x_t} = \underbrace{\begin{bmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{bmatrix}}_{A_1} \underbrace{\begin{bmatrix} Y_{t-1} \\ \pi_{t-1} \\ i_{t-1} \end{bmatrix}}_{x_{t-1}} + \underbrace{\begin{bmatrix} u_{1,t} \\ u_{2,t} \\ u_{3,t} \end{bmatrix}}_{u_t} \tag{7}$$

Write the compact reduced form.

$$x_t = A_1 x_{t-1} + u_t \tag{8}$$

A_1 collects the lag coefficients, and u_t collects the reduced-form innovations. It is important to note that a reduced-form innovation is not an economic shock. It is what the lagged variables failed to predict; it does not label the economic cause. The next section elaborates on that.

4 Why we need a structural VAR

A reduced-form VAR (Equation (8)) is estimable directly from data and is useful for describing how variables forecast one another over time, but its innovations do not generally have a direct economic interpretation. The reduced-form forecast error (u_t) tells us that something unexpected happened at time t, but it does not specify what economic shock caused it. For example, an unexpected increase in the policy rate need not be a monetary policy shock. The policy rate may rise unexpectedly because the central bank responds within the period to an unexpected increase in inflation or output.

[2]see Section 9: Companion form.

The purpose of the structural VAR representation is to make the problem of causal identification explicit. To give the innovations a causal interpretation, we therefore distinguish between the estimable reduced-form forecast errors (u_t) and a vector of underlying structural shocks (ε_t).

The distinction arises because economic variables can interact contemporaneously. Consider Equation (8). Lag coefficients (A_1) cannot capture the contemporaneous relationships. A structural representation, therefore, includes contemporaneous variables directly: for example, b_{12} captures the contemporaneous effect of inflation on output, and b_{13} captures the contemporaneous effect of the policy rate on output.[3] We change the coefficients from ϕ to φ as we separate contemporaneous terms.

$$Y_t = \varphi_{11}Y_{t-1} + \varphi_{12}\pi_{t-1} + \varphi_{13}i_{t-1} + \underbrace{(-b_{12}\pi_t - b_{13}i_t + \varepsilon_{Y,t})}_{u_{1,t}+\text{predictable parts}}$$
$$\pi_t = \varphi_{21}Y_{t-1} + \varphi_{22}\pi_{t-1} + \varphi_{23}i_{t-1} + \underbrace{(-b_{21}Y_t - b_{23}i_t + \varepsilon_{\pi,t})}_{u_{2,t}+\text{predictable parts}}$$
$$i_t = \varphi_{31}Y_{t-1} + \varphi_{32}\pi_{t-1} + \varphi_{33}i_{t-1} + \underbrace{(-b_{31}Y_t - b_{32}\pi_t + \varepsilon_{i,t})}_{u_{3,t}+\text{predictable parts}}$$

Notice that the parts predictable by lagged variables in parentheses will be absorbed by lagged terms with φs[4]. The remaining unpredictable parts will be a mix of structural shocks comprising $u_{k,t}$, which will be explicitly shown on the next page. Moving the contemporaneous variables back to the left-hand side gives the structural equations.

$$\begin{aligned}
1\,Y_t + b_{12}\pi_t + b_{13}i_t &= \varphi_{11}Y_{t-1} + \varphi_{12}\pi_{t-1} + \varphi_{13}i_{t-1} + \varepsilon_{Y,t} \\
b_{21}Y_t + 1\,\pi_t + b_{23}i_t &= \varphi_{21}Y_{t-1} + \varphi_{22}\pi_{t-1} + \varphi_{23}i_{t-1} + \varepsilon_{\pi,t} \\
b_{31}Y_t + b_{32}\pi_t + 1\,i_t &= \varphi_{31}Y_{t-1} + \varphi_{32}\pi_{t-1} + \varphi_{33}i_{t-1} + \varepsilon_{i,t}
\end{aligned}$$

The blue coefficients on the contemporaneous variables form the matrix B_0. Thus, B_0 is simply a compact representation of the within-period relationships:

$$B_0x_t = B_1x_{t-1} + \varepsilon_t, \qquad \text{where} \qquad B_0 = \begin{bmatrix} 1 & b_{12} & b_{13} \\ b_{21} & 1 & b_{23} \\ b_{31} & b_{32} & 1 \end{bmatrix} \qquad \varepsilon_t = \begin{bmatrix} \varepsilon_{Y,t} \\ \varepsilon_{\pi,t} \\ \varepsilon_{i,t} \end{bmatrix} \tag{9}$$

The vector ε_t contains the structural shocks. We commonly normalize these shocks so that they have zero means, unit variances, and, by definition, they do not correlate with one another:

$$\mathbb{E}[\varepsilon_t] = 0 \qquad \mathbb{E}[\varepsilon_t\varepsilon_t'] = I \tag{10}$$

The normalization fixes the scale of the structural shocks. It does not, by itself, identify the contemporaneous relationships in B_0.

[3] They are not recoverable via OLS due to endogeneity.

[4] Hence the difference between ϕs and φs: namely, reduced-form and structural-form coefficients, respectively.

Equation (9) cannot generally be estimated equation by equation with OLS because the contemporaneous variables are endogenous. Multiplying the structural system by B_0^{-1} gives

$$x_t = \underbrace{B_0^{-1}B_1}_{A_1} x_{t-1} + \underbrace{B_0^{-1}\varepsilon_t}_{u_t} \tag{11}$$

Therefore,

$$u_t = B_0^{-1}\varepsilon_t \tag{12}$$

Multiplying by B_0^{-1} does not eliminate the contemporaneous relationships. Instead, those relationships become embedded in the mapping from the structural shocks ε_t to the reduced-form forecast errors u_t.

To see this explicitly, define

$$C = B_0^{-1} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \tag{13}$$

Then

$$\begin{aligned} u_{Y,t} &= c_{11}\varepsilon_{Y,t} + c_{12}\varepsilon_{\pi,t} + c_{13}\varepsilon_{i,t} \\ u_{\pi,t} &= c_{21}\varepsilon_{Y,t} + c_{22}\varepsilon_{\pi,t} + c_{23}\varepsilon_{i,t} \\ u_{i,t} &= c_{31}\varepsilon_{Y,t} + c_{32}\varepsilon_{\pi,t} + c_{33}\varepsilon_{i,t} \end{aligned} \tag{14}$$

Each reduced-form forecast error is therefore generally a weighted average of several structural shocks. For example, an unexpected movement in the output equation need not represent a pure output shock. It may also contain the contemporaneous effects of inflation and monetary policy shocks.

This mixing also explains why reduced-form forecast errors can be correlated, even when the structural shocks are mutually uncorrelated:

$$\begin{aligned} \Sigma_u &\equiv \mathbb{E}[u_t u_t'] \\ &= B_0^{-1}\mathbb{E}[\varepsilon_t \varepsilon_t'](B_0^{-1})' \\ &= B_0^{-1}(B_0^{-1})'. \end{aligned} \tag{15}$$

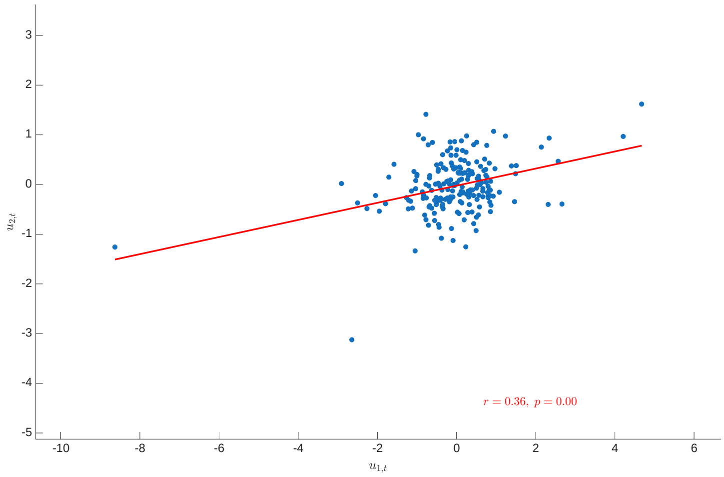


Figure 7: Reduced-form GDP and CPI forecast errors can be contemporaneously correlated because they are mixtures of the same underlying structural shocks.

The identification problem

The reduced-form coefficients A_1 and forecast errors u_t can be estimated equation by equation. The difficult step is recovering the economically meaningful structural shocks ε_t from those forecast errors. This requires identifying the contemporaneous impact matrix B_0^{-1}.

5 Impulse response functions

An impulse response function traces the effect of one structural shock while holding other structural shocks at zero. For now, assume that we recovered B_0^{-1}; the next sections will elaborate on ways to do so. Suppose the economy operates at its normalized steady state, $x_{-1} = 0$, and a unit monetary policy shock occurs in period 0:

$$\varepsilon_{MP,0} = e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{16}$$

The impact response is

$$x_0 = B_0^{-1} e_3 \tag{17}$$

One period later,

$$x_1 = A_1 B_0^{-1} e_3 \tag{18}$$

and after another period,

$$x_2 = A_1^2 B_0^{-1} e_3 \tag{19}$$

Therefore, for the timing convention used here,

$$\mathrm{IRF}_t^{(MP)} = A_1^t B_0^{-1} \varepsilon_{MP,0} \tag{20}$$

The impact matrix B_0^{-1} determines what happens immediately. The autoregressive matrix A_1 determines how the initial effect propagates.

Why two models can produce different IRFs

Changing the variables or restrictions can change the estimated impact vector $B_0^{-1}e_3$. Changing the variables, transformations, or lag length can also change the propagation matrices $A_1, \ldots, A_p$. Therefore two models may disagree both about the initial effect and about its persistence.

For a three-variable model, the impact response to a monetary policy shock is simply the third column of B_0^{-1}:

$$x_0 = B_0^{-1}e_3 = \begin{bmatrix} c_{13} \\ c_{23} \\ c_{33} \end{bmatrix} \tag{21}$$

The next-period response combines all three elements through A_1:

$$x_h = A_1^h x_1 = \left(\begin{bmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{bmatrix} \right)^h \begin{bmatrix} c_{13} \\ c_{23} \\ c_{33} \end{bmatrix} \tag{22}$$

This multiplication is the essence of propagation: the impact on each variable becomes an input into every equation one period later. So, if A_1 changes, the whole propagation does too.

If we have more than one lag, a reduced-form VAR(p) is

$$x_t = A_1 x_{t-1} + A_2 x_{t-2} + \cdots + A_p x_{t-p} + u_t \qquad u_t = B_0^{-1}\varepsilon_t \tag{23}$$

Let $R_0 = B_0^{-1}e_3$ be the impact response. Later responses satisfy

$$R_1 = A_1 R_0 \tag{24}$$

$$R_2 = A_1 R_1 + A_2 R_0 \tag{25}$$

$$R_3 = A_1 R_2 + A_2 R_1 + A_3 R_0 \tag{26}$$

and, in general,

$$R_h = \sum_{j=1}^{\min\{h,p\}} A_j R_{h-j} \tag{27}$$

Each lag matrix creates another path through which the shock can affect future values.

6 The identification problem

The covariance matrix of reduced-form innovations is

$$\Sigma_u = \mathrm{Cov}(u_t) = B_0^{-1}(B_0^{-1})' \tag{28}$$

For three variables, B_0^{-1} has nine elements. The symmetric matrix Σ_u has only six unique elements. We therefore need three additional restrictions. In general, if there are k variables, we need

$$k^2 - \frac{k(k+1)}{2} = \frac{k(k-1)}{2} \tag{29}$$

additional restrictions to recover the impact matrix.

To make the counting argument explicit, write $c = B_0^{-1}$ and normalize the structural shocks to have identity covariance. Then

$$\Sigma_u = cc' = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix} \tag{30}$$

The diagonal moments are

$$\sigma_{11} = c_{11}^2 + c_{12}^2 + c_{13}^2 \tag{31}$$
$$\sigma_{22} = c_{21}^2 + c_{22}^2 + c_{23}^2 \tag{32}$$
$$\sigma_{33} = c_{31}^2 + c_{32}^2 + c_{33}^2 \tag{33}$$

and the three covariance moments are

$$\sigma_{12} = c_{11}c_{21} + c_{12}c_{22} + c_{13}c_{23} \tag{34}$$
$$\sigma_{13} = c_{11}c_{31} + c_{12}c_{32} + c_{13}c_{33} \tag{35}$$
$$\sigma_{23} = c_{21}c_{31} + c_{22}c_{32} + c_{23}c_{33} \tag{36}$$

These are six equations for nine unknown impact coefficients. The missing three restrictions cannot be learned from Σ_u alone.

Identification in one sentence

The data estimate the reduced-form covariance matrix. Economic assumptions select one structural decomposition of that covariance matrix.

Common choices include short-run zero restrictions, long-run restrictions, sign restrictions, and external instruments. This note focuses on the first and last.

7 Recursive identification: Cholesky

The Cholesky approach orders variables by how quickly they can respond within the period. With the ordering

$$Y_t \to \pi_t \to i_t \tag{37}$$

the interpretation is:

- output does not respond within the period to inflation or interest-rate shocks;
- inflation may respond within the period to output shocks;
- the interest rate may respond within the period to both output and inflation shocks.

The recursive restriction makes B_0 lower triangular:

$$B_0 = \begin{bmatrix} 1 & 0 & 0 \\ b_{21} & 1 & 0 \\ b_{31} & b_{32} & 1 \end{bmatrix} \tag{38}$$

In practice, estimate $\widehat{\Sigma}_u$ and compute its Cholesky factor P:

$$\widehat{\Sigma}_u = PP', \qquad P \text{ lower triangular.} \tag{39}$$

Under the recursive identification, P is the estimated impact matrix. Structural shocks and impulse responses are then

$$u_t = P\varepsilon_t, \qquad \mathrm{IRF}_t^{(k)} = A_1^{t-1} P e_k. \tag{40}$$

7.1 A numerical illustration

Suppose the estimated reduced-form covariance matrix is

$$\widehat{\Sigma}_u = \begin{bmatrix} 1.0 & 0.3 & 0.2 \\ 0.3 & 0.8 & 0.4 \\ 0.2 & 0.4 & 1.2 \end{bmatrix} \tag{41}$$

Its lower-triangular Cholesky factor is approximately

$$P = \begin{bmatrix} 1.000 & 0 & 0 \\ 0.300 & 0.843 & 0 \\ 0.200 & 0.404 & 0.999 \end{bmatrix} \qquad \widehat{\Sigma}_u = PP' \tag{42}$$

The first column says that a one-unit first structural shock moves the three variables on impact by 1.0, 0.3, and 0.2. The second shock cannot move the first variable on impact; the third shock can move only the third variable on impact. Those zeros do not come from the data alone. They encode the chosen recursive ordering.

7.2 Ordering is a sensitivity check

Different orderings imply different contemporaneous assumptions. For example,

$Y_t \to \pi_t \to i_t$: output is the slow variable and policy is the fast variable,

$i_t \to Y_t \to \pi_t$: policy does not react within the period to output or inflation innovations,

$\pi_t \to Y_t \to i_t$: inflation is contemporaneously predetermined.

Each ordering selects a different triangular factor and can therefore generate different impulse responses, even when the reduced-form VAR is unchanged.

Keep in mind

Variable ordering is an economic assumption, not a harmless computational choice. Different orderings can produce different impulse responses.

8 External instruments: Proxy VAR

Suppose we care only about a monetary policy shock. A valid external instrument z_t can identify the relevant column of the impact matrix without identifying every structural shock.

Order the variables as $x_t = [i_t \;\; Y_t \;\; \pi_t]'$, so the monetary policy shock is first. The instrument must satisfy:

$$\text{Relevance:} \quad \mathrm{Cov}(z_t, \varepsilon_{i,t}) \neq 0, \tag{43}$$

$$\text{Exogeneity:} \quad \mathrm{Cov}(z_t, \varepsilon_{Y,t}) = 0, \qquad \mathrm{Cov}(z_t, \varepsilon_{\pi,t}) = 0. \tag{44}$$

Because $u_t = c\varepsilon_t$,

$$\mathbb{E}[u_t z_t] = \mathrm{Cov}(\varepsilon_{i,t}, z_t) \begin{bmatrix} c_{i1} \\ c_{Y1} \\ c_{\pi 1} \end{bmatrix} \tag{45}$$

Writing the components gives

$$\mathrm{Cov}(u_{i,t}, z_t) = \mathrm{Cov}(\varepsilon_{i,t}, z_t) c_{i1} \tag{46}$$

$$\mathrm{Cov}(u_{Y,t}, z_t) = \mathrm{Cov}(\varepsilon_{i,t}, z_t) c_{Y1} \tag{47}$$

$$\mathrm{Cov}(u_{\pi,t}, z_t) = \mathrm{Cov}(\varepsilon_{i,t}, z_t) c_{\pi 1} \tag{48}$$

If we normalize $c_{i1} = 1$, the identified impact vector is

$$c_1 = \frac{\mathrm{Cov}(u_t, z_t)}{\mathrm{Cov}(u_{i,t}, z_t)} \tag{49}$$

This normalization scales responses to a one-unit innovation in the policy residual. Scaling the shock to one standard deviation requires an additional rescaling.

Applications and related uses of external instruments include Stock and Watson (2012), Mertens and Ravn (2013), and Gertler and Karadi (2015).

9 Companion form

A VAR(p),

$$x_t = A_1 x_{t-1} + A_2 x_{t-2} + \cdots + A_p x_{t-p} + u_t \tag{50}$$

can be rewritten as a VAR(1) by stacking current and lagged values:

$$X_t = \begin{bmatrix} x_t & x_{t-1} & \cdots & x_{t-p+1} \end{bmatrix}' \tag{51}$$

Then

$$X_t = FX_{t-1} + U_t \tag{52}$$

where

$$F = \begin{bmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{bmatrix}, \qquad U_t = \begin{bmatrix} u_t \; 0 \; \cdots \; 0 \end{bmatrix}' \tag{53}$$

The companion form is useful because any VAR(p) can now be analyzed as a VAR(1). It is simply a first-order state-space representation of the VAR(p).

10 Uncertainty: a simple wild bootstrap

An estimated impulse response is only one sample estimate. A bootstrap asks how much it would change across plausible alternative samples.

For a fitted VAR(p), retain the estimated coefficients and residuals $\widehat{u}_t$. For each bootstrap replication:

$$\widehat{u}_t = x_t - \widehat{A}_1 x_{t-1} - \widehat{A}_2 x_{t-2} - \cdots - \widehat{A}_p x_{t-p}. \tag{54}$$

1. Draw independent signs

$$v_t = \begin{cases} +1 & \text{with probability } 0.5, \\ -1 & \text{with probability } 0.5. \end{cases} \tag{55}$$

2. Form bootstrap innovations $u_t^* = v_t \widehat{u}_t$.
3. Generate a bootstrap sample recursively:

$$x_t^* = \widehat{A}_1 x_{t-1}^* + \cdots + \widehat{A}_p x_{t-p}^* + u_t^* \tag{56}$$

4. Re-estimate the VAR and recompute the impulse responses.
5. Repeat many times and use the empirical quantiles as confidence bands.

The sign flip preserves the magnitude and timing of each residual vector while changing its direction. Narrow bands indicate a precisely estimated response; wide bands indicate substantial sampling uncertainty.

Keep in mind

Confidence bands measure statistical uncertainty under the bootstrap design. They do not validate the economic identification assumptions.

What to remember

1. A VAR model represents the joint dynamics of several variables using their own lags and the lags of the other variables.
2. The reduced-form VAR can be estimated directly, but its innovations generally mix several underlying structural shocks.
3. Structural identification is needed to recover economically interpretable shocks from the reduced-form innovations.
4. Identification requires additional restrictions that cannot be obtained from the reduced-form VAR alone.
5. Impulse responses trace the dynamic effects of an identified structural shock; the impact matrix determines the contemporaneous response, while the VAR coefficients determine its propagation over time.
6. Cholesky identification imposes a recursive contemporaneous ordering, whereas a proxy VAR uses an external instrument to identify a structural shock.
7. Bootstrap confidence bands quantify sampling uncertainty in the estimated impulse responses; they do not establish whether the identifying assumptions are credible.

Further reading

- Kilian, L. and H. Luetkepohl. *Structural Vector Autoregressive Analysis*. Cambridge University Press.
- Stock, J. H. and M. W. Watson (2012). "Disentangling the Channels of the 2007-2009 Recession." *Brookings Papers on Economic Activity*, Spring 2012: 81–135. Link.
- Mertens, K. and M. O. Ravn (2013). "The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States." *American Economic Review*, 103(4): 1212–1247. Link.
- Gertler, M. and P. Karadi (2015). "Monetary Policy Surprises, Credit Costs, and Economic Activity." *American Economic Journal: Macroeconomics*, 7(1): 44–76. Link.
- Lunsford, K. G. (2015) "Identifying Structural VARs with a Proxy Variable and a Test for a Weak Proxy." Link.