Università di Pavia

Impulse Response Functions

Eduardo Rossi

VAR(p):

$$\mathbf{y}_t = (y_{1t}, \ldots, y_{Nt})'$$

$$\mathbf{\Phi}(L)\mathbf{y}_t = \boldsymbol{\epsilon}_t$$

$$\mathbf{\Phi}(L) = \mathbf{I}_N - \mathbf{\Phi}_1 L - \ldots - \mathbf{\Phi}_p L^p$$

$$\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{Nt})'$$

$$\boldsymbol{\epsilon}_t \sim \text{independent } VWN(\mathbf{0}, \mathbf{\Omega})$$

The process is stable if

$$det(\mathbf{I}_N - \mathbf{\Phi}_1 z - \ldots - \mathbf{\Phi}_p z^p) \neq 0 \text{ for } |z| \leq 1$$

On the assumption that the process has been initiated in the infinite past ($t = 0, \pm 1, \pm 2, \ldots$) it generates stationary time series that have time-invariant means, variances, and covariances.

Because VAR models represent the correlations among a set of variables, they are often used to analyze certain aspects of the relationships between the variables of interest.

Granger (1969) has defined a concept of *causality* which,under suitable conditions, is fairly easy to deal with in the context of VAR models. Therefore it has become quite popular in recent years. The idea is that *a cause cannot come after the effect.*

If a variable x affects a variable z, the former should help improving the predictions of the latter variable.

Ω_t is the information set containing all the relevant information in the universe available up to and including period t.
$z_t(h|\Omega_t)$ be the optimal (minimum MSE) h-step predictor of the process z_t at origin t, based on the information in Ω_t.

The corresponding forecast MSE: $\Sigma_t(h|\Omega_t)$.

The process x_t is said to cause z_t in Granger's sense if

$$\Sigma_t(h|\Omega_t) < \Sigma_t(h|\Omega_t \, \{x_s : s \leq t\}) \qquad \text{for at least one} \qquad h = 1, 2, \ldots$$

$\Omega_t \, \{x_s : s \leq t\}$ is the set containing all the relevant information in the universe except for the information in the past and present of the x_t process.

If z_t can be predicted more efficiently if the information in the x_t process is taken into account in addition to all other information in the universe, then x_t is *Granger-causal* for z_t.

If $\mathbf{x}_t : (N \times 1)$ causes $\mathbf{z}_t : (M \times 1)$ and $\mathbf{z}_t$ also causes $\mathbf{x}_t$ the process $(\mathbf{z}_t', \mathbf{x}_t')'$ is called a *feedback system*.

We say that there is *instantaneous causality* between z_t and x_t if

$$\Sigma_z(1|\Omega_t \cup \mathbf{x}_{t+1}) \neq \Sigma_z(1|\Omega_t)$$

In other words, in period t, adding $\mathbf{x}_{t+1}$ to the information set helps to improve the forecast of $\mathbf{z}_{t+1}$.

This concept of causality is really symmetric, that is, if there is instantaneous causality between $\mathbf{z}_t$ and $\mathbf{x}_t$, then there is also instantaneous causality between $\mathbf{x}_t$ and $\mathbf{z}_t$.

A possible criticism of the foregoing definitions could relate to the choice of the MSE as a measure of the forecast precision. Of course, the choice of another measure could lead to a different definition of causality.

Equality of the MSEs will imply equality of the corresponding predictors.

In that case a process $\mathbf{z}_t$ is not Granger-caused by $\mathbf{x}_t$ if the optimal predictor of $\mathbf{z}_t$ does not use information from the $\mathbf{x}_t$ process. This result is intuitively appealing.

A more serious practical problem is the choice of the information set Ω_t. Usually all the relevant information in the universe is not available to a forecaster and, thus, the optimal predictor given Ω_t cannot be determined. Therefore a less demanding definition of causality is often used in practice.

Instead of all the information in the universe, only the information in the past and present of the process under study is considered relevant and Ω_t is replaced by $\{\mathbf{z}_s, \mathbf{x}_s | s \leq t\}$. Furthermore, instead of optimal predictors, *optimal linear predictors* are compared.

Granger-causality may not tell us the complete story about the interactions between the variables of a system.
In applied work, it is often of interest to know the response of one variable to an impulse in another variable in a system that involves a number of further variables as well.
One would like to investigate the impulse response relationship between two variables in a higher dimensional system. Of course, if there is a reaction of one variable to an impulse in another variable we may call the latter causal for the former.

We will study this type of causality by tracing out the effect of an exogenous shock or innovation in one of the variables on some or all of the other variables.

This kind of impulse response analysis is called *multiplier analysis*.

For instance, in a system consisting of an inflation rate and an interest rate, the effect of an increase in the inflation rate may be of interest.

In the real world, such an increase may be induced exogenously from outside the system by events like the increase of the oil price in 1973/74 when the OPEC agreed on a joint action to raise prices.

Alternatively, an increase or reduction in the interest rate may be administered by the central bank for reasons outside the simple two variable system under study.

Three-variables system:

$$\mathbf{y}_t = \begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{bmatrix} \qquad \begin{array}{l} \text{investment} \\ \text{income} \\ \text{consumption} \end{array}$$

the effect of an innovation in investment. To isolate such an effect, suppose that all three variables assume their mean value prior to time $t = 0$, $\mathbf{y}_t = \boldsymbol{\mu}, t < 0$, and investment increases by one unit in period $t = 0$, i.e. $\epsilon_{10} = 1$.

Now we can trace out what happens to the system during periods $t = 1, 2, \ldots$ if no further shocks occur, that is,

$$\epsilon_{20} = \epsilon_{30} = 0$$

and

$$\boldsymbol{\epsilon}_2 = \mathbf{0}, \boldsymbol{\epsilon}_3 = \mathbf{0}, \ldots$$

We assume that all three variables have mean zero and set $\mathbf{c} = \mathbf{0}$, then

$$\mathbf{y}_t = \boldsymbol{\Phi}_1 \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t$$

$$\mathbf{y}_t = \begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{bmatrix} = \begin{bmatrix} .5 & 0 & 0 \\ .1 & .1 & .3 \\ 0 & .2 & .3 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \epsilon_{3,t} \end{bmatrix}$$

Tracing a unit shock in the first variable in period $t = 0$ in this system we get

$$\mathbf{y}_0 = \begin{bmatrix} y_{1,0} \\ y_{2,0} \\ y_{3,0} \end{bmatrix} = \begin{bmatrix} \epsilon_{1,0} \\ \epsilon_{2,0} \\ \epsilon_{3,0} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbf{y}_1 = \begin{bmatrix} y_{1,1} \\ y_{2,1} \\ y_{3,1} \end{bmatrix} = \mathbf{\Phi}_1 \mathbf{y}_0 = \begin{bmatrix} 0.5 \\ 0.1 \\ 0 \end{bmatrix}$$

$$\mathbf{y}_2 = \begin{bmatrix} y_{1,2} \\ y_{2,2} \\ y_{3,2} \end{bmatrix} = \mathbf{\Phi}_1 \mathbf{y}_1 = \mathbf{\Phi}_1^2 \mathbf{y}_0 = \begin{bmatrix} 0.5 \\ 0.06 \\ 0.02 \end{bmatrix}$$

it turns out that $\mathbf{y}_i = (y_{1,i}, y_{2,i}, y_{3,i})'$ is just the first column of $\mathbf{\Phi}_1^i$

$$\mathbf{y}_i = \mathbf{\Phi}_1^i \mathbf{u}_i$$

where

$$\mathbf{u}_i' = (0, 0, \ldots, 1, \ldots, 0)$$

the elements of $\mathbf{\Phi}_1^i$ represent the effects of unit shocks in the variables of the system after i periods. They are called *impulse responses* or *dynamic multipliers*.

Recall that $\mathbf{\Phi}_1^i = \mathbf{\Psi}_i$ just the i-th coefficient matrix of the MA representation of a VAR(1) process.
The MA coefficient matrices contain the impulse responses of the system. This result holds more generally for higher order VAR(p) processes as well.
$VMA(\infty)$ representation:

$$\mathbf{y}_t = \sum_{i=0}^{\infty} \mathbf{\Psi}_i \boldsymbol{\epsilon}_{t-i} \quad \mathbf{\Psi}_0 = \mathbf{I}_n$$

Impulse-response function

$$\mathbf{y}_{t+n} = \sum_{i=0}^{\infty} \mathbf{\Psi}_i \boldsymbol{\epsilon}_{t+n-i}$$

$$\{\mathbf{\Psi}_n\}_{i,j} = \frac{\partial y_{it+n}}{\partial \epsilon_{jt}}$$

the response of $y_{i,t+n}$ to a one-time impulse in $y_{j,t}$ with all other variables dated t or earlier held constant.

The response of variable i to a unit shock (forecast error) in variable j is sometimes depicted graphically to get a visual impression of the dynamic interrelationships within the system.

If the variables have different scales, it is sometimes useful to consider innovations of one standard deviation rather than unit shocks.

For instance, instead of tracing an unexpected unit increase in investment in the investment/income/consumption system with $Var(\epsilon_{1,t}) = 2.25$, one may follow up on a shock of $\sqrt{2.25} = 1.5$ units because the standard deviation of $\epsilon_{1,t}$ is 1.5.
Of course, this is just a matter of rescaling the impulse responses.

The impulse responses are zero if one of the variables does not Granger-cause the other variables taken as a group.

An innovation in variable k has no effect on the other variables if the former variable does not Granger-cause the set of the remaining variables.

A problematic assumption in this type of impulse response analysis is that a shock occurs only in one variable at a time. Such an assumption may be reasonable if the shocks in different variables are independent.

If they are not independent one may argue that the error terms consist of all the influences and variables that are not directly included in the set of $\mathbf{y}$ variables.

Thus, in addition to forces that affect all the variables, there may be forces that affect variable 1, say, only. If a shock in the first variable is due to such forces it may again be reasonable to interpret the $\mathbf{\Psi}_i$ coefficients as dynamic responses.

On the other hand, correlation of the error terms may indicate that a shock in one variable is likely to be accompanied by a shock in another variable.

In that case, setting all other errors to zero may provide a misleading picture of the actual dynamic relationships between the variables.

This is the reason why impulse response analysis is often performed in terms of the MA representation:

$$\mathbf{\Omega} = \mathbf{P}\mathbf{P}' \qquad \text{Cholesky decomposition}$$

$\mathbf{P}$ is a lower triangular matrix. From

$$\mathbf{y}_t = \boldsymbol{\epsilon}_t + \mathbf{\Psi}_1 \boldsymbol{\epsilon}_{t-1} + \mathbf{\Psi}_2 \boldsymbol{\epsilon}_{t-2} + \ldots \qquad \mathbf{\Psi}_0 = \mathbf{I}_n$$

to

$$\mathbf{y}_t = \mathbf{\Theta}_0 \mathbf{w}_t + \mathbf{\Theta}_1 \mathbf{w}_{t-1} + \mathbf{\Theta}_2 \mathbf{w}_{t-2} + \ldots$$

with

$$\mathbf{\Theta}_i = \mathbf{\Psi}_i \mathbf{P}$$

$$\mathbf{w}_t = \mathbf{P}^{-1} \boldsymbol{\epsilon}_t$$

$$E[\mathbf{w}_t \mathbf{w}_t'] = \mathbf{I}_N$$

It is reasonable to assume that a change in one component of $\mathbf{w}_t$ has no effect on the other components because the components are orthogonal (uncorrelated).

Moreover, the variances of the components are one. Thus, a unit innovation is just an innovation of size one standard deviation.

The elements of the $\boldsymbol{\Theta}_i$ are interpreted as responses of the system to such innovations.

$$\{\boldsymbol{\Theta}_i\}_{jk}$$

is assumed to represent the effect on variable j of a unit innovation in the k-th variable that has occurred i periods ago.

To relate these impulse responses to a VAR model, we consider the zero mean VAR(p)process:

$$\mathbf{\Phi}(L)\mathbf{y}_t = \boldsymbol{\epsilon}_t$$

This process can be rewritten in such a way that the disturbances of different equations are uncorrelated. For this purpose, we choose a decomposition of the white noise covariance matrix

$$\mathbf{\Omega} = \mathbf{W}\mathbf{\Sigma}\mathbf{W}'$$

$\mathbf{\Sigma}$ is a diagonal matrix with positive diagonal elements and $\mathbf{W}$ is a lower triangular matrix with unit diagonal.

This decomposition is obtained from the Choleski decomposition $\mathbf{\Omega} = \mathbf{P}\mathbf{P}'$ by defining a diagonal matrix $\mathbf{D}$ which has the same main diagonal as $\mathbf{P}$ and by specifying

$$\begin{aligned} \mathbf{W} &= \mathbf{P}\mathbf{D}^{-1} \\ \mathbf{\Sigma} &= \mathbf{D}\mathbf{D}' \end{aligned}$$

then from

$$\mathbf{\Omega} = \mathbf{P}\mathbf{P}'$$

and

$$\mathbf{P} = \mathbf{W}\mathbf{D}$$

$$\mathbf{\Omega} = \mathbf{P}\mathbf{P}' = \mathbf{W}\mathbf{D}\mathbf{D}'\mathbf{W}' = \mathbf{W}\mathbf{\Sigma}\mathbf{W}'$$

Premultiplying the VAR(p) by $\mathbf{A} \equiv \mathbf{W}^{-1}$

$$\mathbf{A}\mathbf{y}_t = \mathbf{A}_1^* \mathbf{y}_{t-1} + \ldots + \mathbf{A}_p^* \mathbf{y}_{t-p} + \mathbf{e}_t$$

where

$$\mathbf{A}_i^* = \mathbf{A}\boldsymbol{\Phi}_i$$

$$\mathbf{e}_t = \mathbf{A}\boldsymbol{\epsilon}_t$$

$\mathbf{e}_t$ has a diagonal var-cov matrix:

$$E[\mathbf{e}_t\mathbf{e}_t'] = \mathbf{A}E[\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t']\mathbf{A}' = \mathbf{A}\boldsymbol{\Omega}\mathbf{A}'$$

Adding $(\mathbf{I}_N - \mathbf{A})\mathbf{y}_t$ to both sides gives

$$(\mathbf{I}_N - \mathbf{A})\mathbf{y}_t + \mathbf{A}\mathbf{y}_t = (\mathbf{I}_N - \mathbf{A})\mathbf{y}_t + \mathbf{A}_1^*\mathbf{y}_{t-1} + \ldots + \mathbf{A}_p^*\mathbf{y}_{t-p} + \mathbf{e}_t$$

$$\mathbf{y}_t = \mathbf{A}_0^*\mathbf{y}_t + \mathbf{A}_1^*\mathbf{y}_{t-1} + \ldots + \mathbf{A}_p^*\mathbf{y}_{t-p} + \mathbf{e}_t$$

Because $\mathbf{W}$ is lower triangular with unit diagonal, the same is true for $\mathbf{A}$.

Hence,

$$
\begin{aligned}
\mathbf{A}_0^* &= (\mathbf{I}_N - \mathbf{A}) \\
&= \begin{bmatrix}
0 & 0 & \dots & 0 & 0 \\
\beta_{12} & 0 & \dots & 0 & 0 \\
\vdots & \ddots & & \ddots & \vdots \\
\vdots & & & & \vdots \\
\beta_{N,1} & \beta_{N,2} & \dots & \beta_{N,N-1} & 0
\end{bmatrix}
\end{aligned}
$$

is a lower triangular matrix with zero diagonal and, thus, in the representation of the VAR(p) process, the first equation contains no instantaneous y's on the right-hand side. The second equation may contain $y_{1,t}$ and otherwise lagged y's on the right-hand side. More generally,the k-th equation may contain $y_{1,t}, \dots, y_{k-1,t}$ and not $y_{k,t}, \dots, y_{N,t}$ on the right-hand side.

Thus, if the VAR(p) with instantaneous effects reflects the actual ongoings in the system, y_{st} cannot have an instantaneous impact on y_{kt} for $k < s$.

In the econometrics literature such a system is called a *recursive model* (Theil (1971)). Wold has advocated these models where the researcher has to specify the instantaneous ”causal” ordering of the variables. This type of causality is therefore sometimes referred to as *Wold-causality*.

If we trace e_{it} innovations of size one standard error through the system, we just get the $\mathbf{\Theta}$ impulse responses. This can be seen by solving the system for $\mathbf{y}_t$:

$$(\mathbf{I}_N - \mathbf{A}_0^*)\mathbf{y}_t = \mathbf{A}_1^*\mathbf{y}_{t-1} + \ldots + \mathbf{A}_p^*\mathbf{y}_{t-p} + \mathbf{e}_t$$

$$\mathbf{y}_t = (\mathbf{I}_N - \mathbf{A}_0^*)^{-1}\mathbf{A}_1^*\mathbf{y}_{t-1} + \ldots + (\mathbf{I}_N - \mathbf{A}_0^*)^{-1}\mathbf{A}_p^*\mathbf{y}_{t-p} + (\mathbf{I}_N - \mathbf{A}_0^*)^{-1}\mathbf{e}_t$$

Noting that $(\mathbf{I}_N - \mathbf{A}_0^*)^{-1} = \mathbf{W} = \mathbf{P}\mathbf{D}^{-1}$ shows that the instantaneous effects of one-standard deviation shocks (e_{it}'s of size one standard deviation) to the system are represented by the elements of

$$\mathbf{W}\mathbf{D} = \mathbf{P} = \mathbf{\Theta}_0$$

because the diagonal elements of $\mathbf{D}$ are just standard deviations of the components of $\mathbf{e}_t$.

The $\mathbf{\Theta}_i$ may provide response functions that are quite different from the $\mathbf{\Psi}_i$ responses.

Note that $\mathbf{\Theta}_0 = \mathbf{P}$ is lower triangular and some elements below the diagonal will be nonzero if $\mathbf{\Omega}$ has nonzero off-diagonal elements.

For the investment/income/consumption example

$$\mathbf{\Theta}_0 = \mathbf{P} = \begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0.5 & 0.7 \end{bmatrix}$$

indicates that an income (y_2) innovation has an immediate impact on consumption (y_3).

If the white noise covariance matrix $\mathbf{\Omega}$ contains zeros, some components of $\boldsymbol{\epsilon}_t$ are contemporaneously uncorrelated. Suppose,for instance, that ϵ_{1t} is uncorrelated with $\epsilon_{it}, i = 2, \ldots, N$.
In this case, $\mathbf{A} = \mathbf{W}^{-1}$ and thus $\mathbf{A}_0^*$ has a block of zeros so that y_1 has no instantaneous effect on $y_i, i = 2, \ldots, N$. In the example,investment has no instantaneous impact on income and consumption because

$$\mathbf{\Omega} = \begin{bmatrix} 2.25 & 0 & 0 \\ 0 & 1 & 0.5 \\ 0 & 0.5 & 0.74 \end{bmatrix}$$

ϵ_{1t} is uncorrelated with ϵ_{2t} and ϵ_{3t}. This, of course, is reflected in the matrix of instantaneous effects $\mathbf{\Theta}_0$ (*impact multipliers*).

The fact that $\mathbf{\Theta}_0$ is lower triangular shows that the ordering of the variables is of importance, that is, it is important which of the variables is called y_1 and which one is called y_2 and so on.

One problem with this type of impulse response analysis is that the ordering of the variables cannot be determined with statistical methods but has to be specified by the analyst. The ordering has to be such that the first variable is the only one with a potential immediate impact on all other variables.

The second variable may have an immediate impact on the last $N-2$ components of $\mathbf{y}_t$ but not on y_{1t} and so on. To establish such an ordering may be a quite difficult exercise in practice.

The choice of the ordering, the Wold causal ordering, may, to a large extent, determine the impulse responses and is therefore critical for the interpretation of the system.

For the investment/income/consumption example it may be reasonable to assume that an increase in income has an immediate effect on consumption while increased consumption stimulates the economy and, hence, income with some time lag.

Besides specifying the relevant impulses to a system, there are a number of further problems that render the interpretation of impulse responses difficult.
A major limitation of our systems is their potential incompleteness.Although in real economic systems almost everything depends on everything else, we will usually work with low-dimensional VAR systems.
All effects of omitted variables are assumed to be in the innovations. If important variables are omitted from the system,this may lead to major distortions in the impulse responses and makes them worthless for structural interpretations.

Consider a system $\mathbf{y}_t$ which is partitioned in vectors $\mathbf{z}_t$ and $\mathbf{x}_t$, with VMA representation

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{z}_t \\ \mathbf{x}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Psi}_{11}(L) & \boldsymbol{\Psi}_{12}(L) \\ \boldsymbol{\Psi}_{21}(L) & \boldsymbol{\Psi}_{22}(L) \end{bmatrix} \begin{bmatrix} \boldsymbol{\epsilon}_{1t} \\ \boldsymbol{\epsilon}_{2t} \end{bmatrix}$$

If the $\mathbf{z}_t$ variables are considered only and the $\mathbf{x}_t$ variables are omitted from the analysis, we get a prediction error MA representation:

$$\mathbf{z}_t = \boldsymbol{\mu}_1 + \sum_{i=0}^{\infty} \boldsymbol{\Psi}_{11,i} \boldsymbol{\epsilon}_{1t-i} + \sum_{i=1}^{\infty} \boldsymbol{\Psi}_{12,i} \boldsymbol{\epsilon}_{2t-i}$$

$$\mathbf{z}_t = \boldsymbol{\mu}_1 + \sum_{i=0}^{\infty} \mathbf{F}_i \mathbf{v}_{t-i}$$

The actual reactions of the $\mathbf{z}_t$ components to innovations $\boldsymbol{\epsilon}_{1t}$ may be given by the $\boldsymbol{\Psi}_{11,i}$ matrices.

On the other hand,the $\mathbf{F}_i$ or corresponding orthogonalized impulse response are likely to be interpreted as impulse responses if the analyst does not realize that important variables have been omitted. The $\mathbf{F}_i$ will be equal to the $\mathbf{\Psi}_{11,i}$, if and only if $\mathbf{x}_t$ does not Granger-cause $\mathbf{z}_t$.

The stable VAR(p) has a Wold MA representation:

$$\mathbf{y}_t = \boldsymbol{\epsilon}_t + \boldsymbol{\Psi}_1 \boldsymbol{\epsilon}_{t-1} + \boldsymbol{\Psi}_2 \boldsymbol{\epsilon}_{t-2} + \dots$$

where

$$\boldsymbol{\Psi}_s = \Sigma_{j=1}^{s} \boldsymbol{\Psi}_{s-j} \boldsymbol{\Phi}_j \qquad s = 1, 2, \dots$$

with $\boldsymbol{\Psi}_0 = \mathbf{I}_K$. The elements of the $\boldsymbol{\Theta}_j$ are *the forecast error impulse responses*.

Different ways to orthogonalize the impulses. Cholesky decomposition of $\boldsymbol{\Omega}$. Such an approach is arbitrary, unless there are special reasons for a recursive structure.

Different ways to use nonsample information in specifying unique innovations and hence unique i-r.

Structural form model

$$\mathbf{A}\mathbf{y}_t = \mathbf{\Phi}_1\mathbf{y}_{t-1} + \mathbf{\Phi}_2\mathbf{y}_{t-2} + \ldots + \mathbf{\Phi}_p\mathbf{y}_{t-p} + \mathbf{B}\mathbf{e}_t$$

The matrix $\mathbf{A}$ contains the instantaneous relations between the left-hand-side variables. It has to be invertible.

The reduced forms are obtained by premultiplying with $\mathbf{A}^{-1}$

$$\mathbf{y}_t = \mathbf{A}^{-1}\mathbf{\Phi}_1\mathbf{y}_{t-1} + \mathbf{A}^{-1}\mathbf{\Phi}_2\mathbf{y}_{t-2} + \ldots + \mathbf{A}^{-1}\mathbf{\Phi}_p\mathbf{y}_{t-p} + \mathbf{A}^{-1}\mathbf{B}\mathbf{e}_t$$

$$\boldsymbol{\epsilon_t} = \mathbf{A}^{-1}\mathbf{B}\mathbf{e}_t$$

which relates the reduced-form disturbances $\boldsymbol{\epsilon_t}$ to the underlying structural shocks $\mathbf{e}_t$.

To identify the structural form parameters, we must place restrictions on the parameter matrices.

Even if $\mathbf{A} = \mathbf{I}_N$ the assumption of orthogonal shocks

$$E\left[\mathbf{e}_t \mathbf{e}_t'\right] = \mathbf{I}_N$$

is not sufficient to achieve identification.

For a N-dimensional system, $\frac{N(N-1)}{2}$ restrictions are necessary for orthogonalizing the shocks because there are $\frac{N(N-1)}{2}$ potentially different instantaneous covariances.

An example of such an identification scheme is the triangular (or recursive) identification suggested by Sims (1980). Such a scheme is also called *Wold causal chain system* and is often associated with a causal chain from the first to the last variable in the system.

Because the i-r functions computed from these models depend on the ordering of the variables, nonrecursive identification schemes that also allow for *instantaneous effects* of the variables

$$\mathbf{A} \neq \mathbf{I}_N$$

have been suggested in the literature (Sims (1986), Bernanke (1986)). Restrictions on the long-run effects of some shocks are also used to identify SVAR models (Blanchard and Quah (1989), Galì (1999), King, Plosser, Stock, and Watson (1991)).

$$\mathbf{A}\boldsymbol{\epsilon_t} = \mathbf{B}\mathbf{e}_t$$

The most popular kind of restrictions:

- $\mathbf{B} = \mathbf{I}_N$, the A-model.
- $\mathbf{A} = \mathbf{I}_N$, the innovation is $\boldsymbol{\epsilon_t} = \mathbf{B}\mathbf{e}_t$, the B-model.
- the AB-model
- Prior information.