Notas Económicas Regionales

Secretaría Ejecutiva del Consejo Monetario Centroamericano

No. 130, Marzo 2022

Una razón para evitar la descomposición de Cholesky en la identificación de un VAR estructural

Randall Romero Aguilar* †
rromero@secmca.org

Para muchos economistas que no están suficientemente familiarizados con los vectores autorregresivos (VAR), la identificación de shocks estructurales por medio de una descomposición de Cholesky resulta demasiado abstracto, por lo que en discusiones de trabajos empíricos que presentan funciones de impulso-respuesta les resulta un tanto misterioso que los resultados sean sensibles al ordenamiento de las variables del VAR.

En esta nota explico por qué el uso de la descomposición de Cholesky es innecesario en el cálculo de las funciones de impulso-respuesta, mostrando que puede sustituirse por un procedimiento conceptualmente equivalente pero mucho más explícito y sencillo de entender. Este procedimiento alternativo permite además comprender por qué debemos tener mucha cautela a la hora de interpretar las funciones de impulso-respuesta.

Un poco de teoría econométrica

Para entender mejor el problema que tratamos en esta nota, vale la pena repasar algunos conceptos de la teoría econométrica. Para facilitar la discusión, ilustraremos esos conceptos con ayuda de un sencillo VAR estructural de un rezago y dos variables endógenas p_t y q_t, que representan el precio y la cantidad de un bien:

$$\alpha_{11}p_t + \alpha_{12}q_t = \beta_{11}p_{t-1} + \beta_{12}q_{t-1} + \varepsilon_{1t}$$
$$\alpha_{21}p_t + \alpha_{22}q_t = \beta_{21}p_{t-1} + \beta_{22}q_{t-1} + \varepsilon_{2t}$$

o bien, en notación matricial[1],

$$\mathbf{A}\mathbf{y}_t = \mathbf{B}\mathbf{y}_{t-1} + \varepsilon_t$$

*Economista de la Secretaría Ejecutiva del Consejo Monetario Centroamericano (SECMCA). Doctor en Economía por la Ohio State University.

†Las opiniones expresadas son las del autor y no necesariamente representan la posición de la SECMCA, ni de los miembros del CMCA.

[1]En esta nota omitiremos el intercepto de cada ecuación, para facilitar la exposición. Los resultados acá presentados no cambian si se incluyen los interceptos.

donde

$$\mathbf{y}_t = \begin{bmatrix} p_t \\ q_t \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} \qquad \varepsilon_t = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$$

En esencia, este VAR estructural lo que indica es que el precio y la cantidad se determinan de manera simultánea, y que dependen de sus valores pasados.

Para poder trabajar con este modelo, es necesario conocer los valores de sus coeficientes α_{ij}, β_{ij} para $i, j \in \{1, 2\}$, los cuales en la práctica estimamos a partir de datos históricos de p_t y de q_t. Un primer paso, indispensable para estimar este modelo, es *normalizar* cada una de las ecuaciones, es decir, asignar al coeficiente de una variable endógena el valor de uno. En la literatura de los VAR generalmente se normaliza una variable distinta en cada ecuación, por lo que en el ejemplo anterior tendríamos $\alpha_{11} = \alpha_{22} = 1$:

$$p_t + \alpha_{12} q_t = \beta_{11} p_{t-1} + \beta_{12} q_{t-1} + \varepsilon_{1t} \tag{1a}$$
$$\alpha_{21} p_t + q_t = \beta_{21} p_{t-1} + \beta_{22} q_{t-1} + \varepsilon_{2t} \tag{1b}$$

En este caso particular, en la literatura de los VAR se interpretaría a ε_{1t} como un *shock de precio* y a ε_{2t} como un *shock de cantidad*.

Ahora bien, si resolvemos el sistema de ecuaciones simultáneas, es decir, si despejamos las variables p_t y q_t, obtenemos un VAR en forma reducida:

$$p_t = \pi_{11} p_{t-1} + \pi_{12} q_{t-1} + \upsilon_{1t} \tag{2a}$$
$$q_t = \pi_{21} p_{t-1} + \pi_{22} q_{t-1} + \upsilon_{2t} \tag{2b}$$

o bien, en notación matricial,

$$\mathbf{y}_t = \Pi \mathbf{y}_{t-1} + \upsilon_t \tag{2c}$$

donde la matriz de coeficientes reducidos es

$$\Pi = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} = \mathbf{A}^{-1}\mathbf{B} \tag{2d}$$

y los shocks reducidos son

$$\upsilon_t = \begin{bmatrix} \upsilon_{1t} \\ \upsilon_{2t} \end{bmatrix} = \mathbf{A}^{-1}\varepsilon_t = \underbrace{\frac{1}{1 - \alpha_{12}\alpha_{21}} \begin{bmatrix} 1 \\ -\alpha_{21} \end{bmatrix}}_{\mathbf{A}_1^{-1}} \varepsilon_{1t} + \underbrace{\frac{1}{1 - \alpha_{12}\alpha_{21}} \begin{bmatrix} -\alpha_{12} \\ 1 \end{bmatrix}}_{\mathbf{A}_2^{-1}} \varepsilon_{2t} \tag{2e}$$

donde $\mathbf{A}_k^{-1}$ corresponde a la columna k-ésima de la inversa de $\mathbf{A}$. Como puede apreciarse en (2e), los shocks reducidos en un VAR son combinaciones lineales de los shocks estructurales: esto implica que si, por ejemplo, duplicamos el tamaño de los shocks estructurales, entonces los shocks reducidos también se duplican.

La función de impulso-respuesta

Uno de los atractivos principales de los modelos VAR es que permiten, de una forma muy sencilla, cuantificar los efectos dinámicos de un shock al sistema. Para ello, sustituimos recursivamente la ecuación (2c) del VAR reducido para mostrar que

$$\mathbf{y}_{t+s} = \Pi^{s+1}\mathbf{y}_{t-1} + \Pi^{s}\upsilon_t + \Pi^{s-1}\upsilon_{t+1} + \cdots + \Pi\upsilon_{t+s-1} + \upsilon_{t+s}$$

por lo que, si deseamos ver el efecto de una perturbación *transitoria* reducida υ_t en t conforme avanza el tiempo, basta con fijar $\mathbf{y}_{t-1} = \upsilon_{t+1} = \cdots = \upsilon_{t+s-1} = \upsilon_{t+s} = 0$ en la expresión anterior y con ello calcular $\Pi^s\upsilon_t$ como una función del tiempo transcurrido s, a lo que conocemos como la función de impulso-respuesta.

Vemos entonces, que para calcular la función de impulso-respuesta es necesario contar con los valores de los parámetros del VAR en forma reducida, Π, así como la magnitud de los impulsos que se desean simular, υ_t. En la práctica, cuando se trabaja con los modelos VAR, lo que se desea simular es el efecto de un shock a una de las ecuaciones estructurales (1), lo cual en principio es muy sencillo si partimos de la ecuación (2e): por ejemplo, si deseamos ver el efecto de un shock de precio de una unidad, fijamos $\varepsilon_{1t} = 1, \varepsilon_{2t} = 0$ y por ello calculamos la función de impuso respuesta con un shock reducido de $\upsilon_t = \mathbf{A}_1^{-1}$ y la respuesta de las variables sería

$$\Pi^s\mathbf{A}_1^{-1}$$

Por lo tanto, para calcular el efecto dinámico de un shock estructural, es necesario conocer o *identificar* los valores de los coeficientes de la matriz $\mathbf{A}$. Así, en resumen, debemos estimar tanto la forma reducida del VAR como la matriz de efectos contemporáneos de su forma estructural.

Estimando un modelo VAR

La versión reducida del VAR, es decir las ecuaciones (2a) y (2b), como es bien conocido, puede estimarse de manera insesgada con mínimos cuadrados ordinarios (MCO), ecuación por ecuación, por tratarse de un sistema de ecuaciones aparentemente no relacionadas[2] (modelo SUR, por sus siglas en inglés).

En contraste, por tratarse de un sistema de ecuaciones simultáneas, en general no es posible estimar el VAR estructural (1a) - (1b) de manera insesgada por el método de mínimos cuadrados ordinarios, debido a que el término de error estará correlacionado con los regresores: por ejemplo, si consideramos a p_t como la variable endógena de la ecuación

[2]En el sitio https://randall-romero.github.io/econometria/07-sur/02-sur.html muestro por qué se cumple este resultado. Una exposición completa de los modelos SUR aparece en Greene (2012)

(1a), de manera que q_t sea una de sus variables explicativas, tenemos que

$$\begin{aligned}
\text{Cov}(\epsilon_{1t}, q_t) &= \text{Cov}(\epsilon_{1t}, \pi_{21}p_{t-1} + \pi_{22}q_{t-1} + \upsilon_{2t}) && \text{(sustituyendo } q_t \text{ con (2b))} \\
&= \text{Cov}\left(\epsilon_{1t}, \pi_{21}p_{t-1} + \pi_{22}q_{t-1} + \frac{\varepsilon_{2t} - \alpha_{21}\varepsilon_{1t}}{1 - \alpha_{12}\alpha_{21}}\right) && \text{(sustituyendo } \varepsilon_{2t} \text{ con (2e))} \\
&= \text{Cov}\left(\epsilon_{1t}, \frac{\varepsilon_{2t} - \alpha_{21}\varepsilon_{1t}}{1 - \alpha_{12}\alpha_{21}}\right) && \text{(asume } \epsilon_{1t} \text{ independiente de } p_{t-1}, q_{t-1}) \\
&\neq 0 && \text{(porque ambos términos dependen de } \varepsilon_{1t})
\end{aligned}$$

por lo que un cambio en ϵ_{1t} afectará a q_t, porque en la segunda ecuación q_t depende de p_t.

Un problema aún mayor es que los parámetros del VAR estructural (1) no pueden estimarse, con ningún método, porque el modelo no está *identificado*. Lo que esto significa es que existen numerosos valores de los parámetros estructurales que son compatibles con los mismos parámetros reducidos, por lo que no tendríamos forma, utilizando únicamente los datos, de discernir cuales son los valores "verdaderos" de los parámetros. Por ello, para poder cuantificar los valores de los parámetros a_{12}, a_{21} es necesario imponer supuestos adicionales respecto al VAR estructural. Consideraremos acá dos enfoques alternativos.

Identificación con exclusión de variables exógenas

Una primera opción, compatible con la metodología desarrollada por la Comisión Cowles para la estimación de modelos de ecuaciones simultáneas, es contar con variables exógenas *distintas* en cada ecuación, de suerte que movimientos de una variable en una ecuación permita identificar los parámetros de la *otra* ecuación. Por ejemplo, si pensamos, arbitrariamente, en la ecuación (1a) como una curva de oferta y en la (1b) como una curva de demanda, y definimos m_t y k_t como el ingreso de los consumidores y el stock de capital de las empresas, podemos reescribir el modelo (1) como

$$p_t + \alpha_{12}q_t = \beta_{11}p_{t-1} + \beta_{12}q_{t-1} + \gamma_{11}k_t + \cancelto{=0}{\gamma_{12}} m_t + \varepsilon_{1t} \tag{3a}$$

$$\alpha_{21}p_t + \quad q_t = \beta_{21}p_{t-1} + \beta_{22}q_{t-1} + \cancelto{=0}{\gamma_{21}} k_t + \gamma_{22}m_t + \varepsilon_{2t} \tag{3b}$$

Otra manera de interpretar lo anterior es que hemos planteado un modelo con más variables explicativos, y al *excluir* al ingreso m_t de (3a) y al capital k_t de (3b) efectivamente hemos impuesto dos restricciones, a saber $\gamma_{12} = 0$ y $\gamma_{21} = 0$.

Es casualmente este tipo restricciones las que Sims (1980) critica en su artículo seminal, al considerarlas "increíbles". Además, en su motivación para proponer el VAR como herramienta de análisis, Sims muestra su insatisfacción con la distinción a veces arbitraria entre

variables endógenas y exógenas, al señalar que "debe ser posible estimar modelos macro de gran escala como formas reducidas sin restricciones, tratando todas las variables como endógenas" (Sims 1980, p.15).

Identificación con exclusión de variables endógenas

La segunda opción que discutiremos para resolver el problema de identificación fue planteada precisamente por Sims (ibíd.). Para empezar, denotemos por $\Sigma = \mathrm{E}(\varepsilon_t \varepsilon_t')$ a la matriz de covarianza de ε_t y por $\Omega = \mathrm{E}(\upsilon_t \upsilon_t')$ a la de υ_t. En (2e) encontramos que $\upsilon_t = \mathbf{A}^{-1}\varepsilon_t$, de manera que las dos matrices de covarianza están relacionadas por

$$\Omega = A^{-1}\Sigma A'^{-1} \tag{4}$$

donde A' denota la transpuesta de A. Lo que propone Sims es asumir que (i) el modelo VAR estructural es recursivo, es decir, que la primera variable no depende contemporáneamente de la segunda ($\alpha_{12} = 0$), lo que da por resultado una matriz A triangular inferior, y (ii) que no hay correlación entre los distintos shocks estructurales, con lo que la matriz Σ es diagonal:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ \alpha_{21} & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \tag{5}$$

En este caso, el VAR estructural toma la forma:

$$p_t = \beta_{11} p_{t-1} + \beta_{12} q_{t-1} + \varepsilon_{1t} \tag{6a}$$

$$\alpha_{21} p_t + q_t = \beta_{21} p_{t-1} + \beta_{22} q_{t-1} + \varepsilon_{2t} \tag{6b}$$

Al sustituir (5) en (4) encontramos que

$$\begin{aligned} \Omega &= \begin{bmatrix} 1 & 0 \\ \alpha_{21} & 1 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} 1 & \alpha_{21} \\ 0 & 1 \end{bmatrix}^{-1} \\ &= \begin{bmatrix} 1 & 0 \\ -\alpha_{21} & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} 1 & -\alpha_{21} \\ 0 & 1 \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} \sigma_1 & 0 \\ -\alpha_{21}\sigma_1 & \sigma_2 \end{bmatrix}}_{\mathbf{P}} \underbrace{\begin{bmatrix} \sigma_1 & -\alpha_{21}\sigma_1 \\ 0 & \sigma_2 \end{bmatrix}}_{\mathbf{P}'} \end{aligned}$$

Es decir, hemos factorizado a Ω como el producto de una matriz triangular inferior **P** y su transpuesta **P'**, lo que se conoce como factorización de Cholesky. En la práctica, esto significa que para calcular la respuesta del VAR a una perturbación de una desviación estándar en la j-ésima variable del VAR, basta con tomar la j-ésima columna de la matriz

de Cholesky **P** como el impulso en forma reducida. Ahora bien, para obtener los valores de **P** es necesario estimar la matriz de covarianza Ω, lo cual podemos hacer de manera sencilla a partir de los residuos de las ecuaciones del VAR reducido (2a)-(2b). Así, la respuesta de la variable i transcurridos s períodos desde una perturbación de una desviación estándar en la variable j se calcula simplemente como el elemento ij de la matriz $\Pi^s\mathbf{P}$

$$\frac{\partial y_{i,t+s}}{\partial y_{j,t}} = (\Pi^s\mathbf{P})_{ij}$$

Es importante observar que este procedimiento tiene implícito un *ordenamiento* de las variables del VAR. En nuestro ejemplo, hemos asumido que un shock de precio ε_{1t} provoca un cambio inmediato tanto en el precio p_t (ecuación (6a)) como en la cantidad q_t (ecuación (6b), porque q_t depende contemporáneamente de p_t), pero, un shock de cantidad ε_{2t} provoca un cambio contemporáneo en la cantidad q_t pero no en el precio p_t (porque q_t no aparece en la ecuación (6a), ya que habíamos impuesto la restricción $\alpha_{12} = 0$).

Antes de acabar esta sección, es importante mencionar que existen enfoques alternativos para estimar un VAR estructural más allá de los dos enfoques descritos acá. Por ejemplo, (i) métodos que imponen otras restricciones contemporáneas, (ii) métodos narrativos, (iii) identificación con datos de alta frecuencia, (iv) uso de variables instrumentales externas, (v) métodos que imponen restricciones de largo plazo, (vi) restricciones de signos, (vii) VARs aumentados con factores, y (viii) métodos basados en la estimación de modelos estocásticos de equilibrio general (DSGE). Ramey (2016) presenta una breve discusión de estos métodos alternativos.

¿Por qué evitar la descomposición de Cholesky?

Lo irónico de todo esto es que el nivel de abstracción del procedimiento de Sims es completamente innecesario. Recordemos que en esencia el problema principal de estimar directamente un modelo VAR estructural como el (1) es, como lo mencionamos anteriormente, que la estimación de los parámetros estaría sesgada por la simultaneidad de las variables. No obstante, ¡este problema desaparece precisamente si estamos dispuestos a asumir los mismos supuestos que impusimos para implementar la identificación a través de la descomposición de Cholesky! Veamos el por qué: en la ecuación (6a) solo hay como regresores variables rezagadas (p_{t-1} y q_{t-1}), que no tienen correlación con el shock contemporáneo ε_{1t}, por lo que esta ecuación puede estimarse apropiadamente con MCO. Pasamos ahora a la ecuación (6b), cuya variable dependiente es q_t; en ella sí aparece un regresor contemporáneo, a saber p_t. No obstante, en este caso no habría sesgo de simultaneidad si estimamos la ecuación con MCO, porque el shock ε_{2t} no estaría correlacionado con este regresor, precisamente porque hemos asumido que q_t no aparece en la ecuación (6a).

Así, en resumen, vemos que al implementar la identificación por medio de la descomposición de Cholesky, asumiendo que A es triangular inferior, básicamente estamos asumiendo que el VAR estructural es recursivo, por lo que no sufre del problema de sesgo de simultaneidad que motivó precisamente la implementación del método de Sims. Además, al haber asumido en el método de Sims que los errores de las distintas ecuaciones estructurales no están correlacionados, entonces no habría ninguna ganancia de eficiencia que justifique estimar el sistema (6) de manera conjunta[3].

Aparte de ser innecesario, el uso de la matriz de Cholesky para fijar el tamaño de los impulsos implica definir, como lo indica Sims (1980, p.21), el impulso como una desviación estándar en cada una de las ecuaciones del sistema, definición que en la práctica implementan muchos paquetes econométricos. El problema práctico de esta definición del tamaño del impulso es que complica considerablemente la interpretación cuantitativa de las respuestas de las variables, porque obliga a comparar su magnitud con la del impulso. Afortunadamente, esto resulta innecesario también: dado que las funciones de impulso-respuesta dependen linealmente de los impulsos, se pueden simular impulsos unitarios[4] dividiendo cada columna de $\mathbf{A}^{-1}$ entre su respectivo elemento diagonal.

Finalmente, como se mencionó anteriormente, el procedimiento impone un ordenamiento de las variable. Aunque esto puede parecer trivial en nuestro pequeño ejemplo de solo dos variable, que solo pueden ordenarse de dos maneras distintas, en un VAR de n variables habrán $n!$ ordenamientos distintos. Así, por ejemplo un VAR de tan solo 6 variables, como el presentado por Sims (ibíd.), tendría que considerarse $6! = 720$ posibles ordenamientos de las variables. Esto evidentemente no es práctico.

¿Qué hacer entonces?

Si se desea estimar las respuestas del sistema a shocks estructurales, y se está dispuesto a asumir que el modelo es recursivo, entonces plantear y estimar el modelo VAR en la forma (6) tiene a mi juicio una ventaja evidente: que es completamente transparente cuáles son los supuestos que el autor está imponiendo sobre las relaciones contemporáneas de las variables del VAR.

Una vez estimados los coeficientes del VAR estructural, la matriz de coeficientes contemporáneos $\mathbf{A}$ se invierte fácilmente (por ser triangular), y se calcula las respuesta de las variables a un shock estructural de magnitud λ[5] en la variable k del sistema como $\lambda \Pi^s \mathbf{A}_k^{-1}$

[3] Con mínimos cuadrados generalizados, por ejemplo.

[4] O de cualquier magnitud apropiada, reescalando el impulso unitario con esa magnitud.

[5] En la práctica, típicamente $\lambda = 1$, aunque como se mencionó anteriormente puede fijarse en una magnitud que tenga sentido en las unidades de medición de la variable respectiva.

Un comentario final

Más allá del método que se utilice para calcular las funciones de impulso-respuesta, es importante tener cautela a la hora de interpretar los resultados. Si bien es cierto la intención original de Sims al plantear los VAR como herramienta analítica fue estimar "un sistema dinámico de seis variables sin usar perspectivas teóricas" (Sims 1980, pp 1-2), cuando se interpreta si tienen sentido los resultados de las funciones de impulso-respuesta se hace desde un punto de vista teórico.

Ahora bien, en economía usualmente planteamos teorías como ecuaciones de comportamiento o identidades contables, pero no como la "ecuación de una variable". Por ejemplo, en el ejemplo de las ecuaciones (1a)-(1b), la única razón por la que una ecuación es de precio y la otra de cantidad es porque decidimos normalizar las ecuaciones de esa manera. Pero recordemos que la normalización de las ecuaciones es arbitraria, por lo que no tiene ningún sentido pensar en términos de *shocks* a una variable. En particular, el modelo pudimos normalizarlo con $a_{12} = a_{22} = 1$ y escribir

$$q_t = -\alpha_{11} p_t + \beta_{11} p_{t-1} + \beta_{12} q_{t-1} + \epsilon_{1t}$$
$$q_t = -\alpha_{21} p_t + \beta_{21} p_{t-1} + \beta_{22} q_{t-1} + \epsilon_{2t}$$

en cuyo caso bien podríamos interpretar ϵ_{1t} como un *shock de oferta* y a ϵ_{2t} como un *shock de demanda*, lo cual se acerca más a la manera teórica de interpretar un modelo estructural. En este caso concreto, interpretamos al precio y a la cantidad como la *solución* del sistema, y la única razón por la que "aceptaríamos" un cambio súbito de estas variables es porque se presente un shock de oferta (mal clima que destruye una cosecha, por ejemplo) o de demanda (un cambio en los gustos del consumidor, por ejemplo).

Referencias

Greene, William H. (2012). *Econometric Analysis*. 7ª ed. Prentice Hall. ISBN: 978-0-13-139538-1.

Qin, Duo (2013). *A History of Econometrics. The Reformation from the 1970s*. Oxford University Press. ISBN: 978–0–19–967934–8.

Ramey, Valerie A. (2016). "Macroeconomic Shocks and Their Propagation". En: *Handbook of Macroeconomics*. Vol. 2A, págs. 71-162.

Sims, Christopher A. (1980). "Macroeconomics and Reality". En: *Econometrica* 48.1.