ANÁLISIS DE OBSERVABILIDAD E IDENTIFICABILIDAD ESTRUCTURAL DE MODELOS NO LINEALES: APLICACIÓN A LA VÍA DE SEÑALIZACIÓN JAK/STAT

Alejandro Fernández Villaverde, Julio Rodríguez Banga
Grupo de Ingeniería de Procesos, IIM-CSIC. Eduardo Cabello 6, Vigo 36208, Galicia, España,
{afvillaverde, julio}@iim.csic.es

Resumen

La observabilidad e identificabilidad estructural son propiedades que se corresponden con la posibilidad teórica de determinar los vectores de estado y de parámetros de un modelo a partir de su salida. Recientemente hemos desarrollado una herramienta MATLAB, STRIKE-GOLDD, que analiza la observabilidad e identificabilidad estructural local de modelos no lineales de ecuaciones diferenciales ordinarias empleando una metodología simbólica. Si bien STRIKE-GOLDD es una metodología de propósito general, su desarrollo está motivado por el estudio de modelos biológicos, para los que estos análisis son de especial relevancia. En este trabajo se analiza la observabilidad e identificabilidad estructural de un modelo dinámico de la vía de señalización celular JAK/STAT. Este sistema de transducción de señales juega un papel importante en procesos de inmunidad, proliferación y división celular, apoptosis y oncogénesis. Usando este modelo de relevancia biomédica como caso de estudio, el presente trabajo proporciona un tutorial sobre el uso de STRIKE-GOLDD, ilustrando sus capacidades y limitaciones.

Palabras clave: Modelado dinámico, Sistemas no lineales, Observabilidad, Identificabilidad, Biosistemas, Métodos computacionales.

1. INTRODUCCIÓN

El concepto de observabilidad describe la posibilidad teórica de inferir el vector de estado de un sistema a partir de observaciones del vector de salida. De forma similar, el concepto de identificabilidad estructural se refiere a la posibilidad teórica de determinar los valores de los parámetros de un modelo observando la salida.

La palabra "teórica" se usa aquí para especificar que tanto la observabilidad como identificabilidad estructural tienen en cuenta sólo las ecuaciones dinámicas del modelo, incluyendo la definición de entradas y salidas, pero no las características de las medidas experimentales tales como el número de muestras o el nivel de ruido. Así pues, si un modelo tiene parámetros estructuralmente no identificables o estados no observables, la causa de estas fallas se encuentra en la estructura de las ecuaciones (por ejemplo, simetrías que impiden distinguir entre variables), y no en limitaciones experimentales. Para tener en consideración estas últimas limitaciones existen los conceptos de identificabilidad práctica y observabilidad práctica, si bien este último término es poco usado.

Tanto la observabilidad como la identificabilidad estructural han sido estudiadas para modelos no lineales durante el último medio siglo, y los fundamentos teóricos de su análisis están bien establecidos. El concepto de observabilidad fue originalmente introducido por R.E. Kalman para sistemas lineales en 1960 [13], y posteriormente extendido a modelos no lineales [15, 12]. Asimismo, el concepto de identificabilidad estructural fue inicialmente formulado para sistemas lineales por Bellman y Åström en 1970 [5] y desarrollado después para el caso no lineal por otros autores [17, 7].

Los parámetros de un modelo – es decir, sus constantes desconocidas – pueden ser considerados como variables de estado cuya derivada es cero. De esa forma el estudio de la identificabilidad estructural se convierte en un caso particular del análisis de observabilidad [21]. A pesar de esta equivalencia, es interesante notar que, mientras que el concepto de observabilidad fue concebido y desarrollado por la comunidad de control automático, el de identificabilidad estructural fue estudiado principalmente en un contexto biomédico, inicialmente en modelado fisiológico y en las últimas décadas también en farmacocinética y de biología de sistemas. (Bien es cierto que en ambos casos la mayor parte de los investigadores eran ingenieros de sistemas y control.) Esta diferencia en la atención prestada a uno y otro concepto se explica por la diferente importancia de su análisis en ambas comunidades: típicamente los modelos biológicos tienen mayor incertidumbre paramétrica y más limitaciones experimentales que los sistemas mecánicos, eléctricos o químicos tradicionalmente estudiados en ingeniería de control. En consecuencia, su identificación presenta más dificultades; un es-

https://doi.org/10.17979/spudc.9788497497169.631

tudio publicado en Automatica en 1978 [4] ya alertaba de que en sistemas biológicos "the low signal to noise ratios, high variability, and measurement difficulties also make estimation procedures unreliable", y a pesar de los avances alcanzados en las últimas décadas, muchas dificultades permanecen [24]. Ello hace que la cuestión de la identificabilidad estructural sea especialmente relevante en modelos biológicos. Una panorámica reciente sobre las diferentes técnicas relacionadas con la identificabilidad en este área se puede encontrar en [26]. Un libro de referencia sobre identificación paramétrica es [30]. La relación entre identificabilidad estructural, observabilidad y otras propiedades como la controlabilidad ha sido comentada para modelos lineales en [8, 7] y más recientemente para no lineales en [25]. Para una visión más general sobre modelado dinámico en biología de sistemas se puede consultar por ejemplo [9, 16].

La geometría diferencial proporciona herramientas para el análisis de la observabilidad e identificabilidad estructural de sistemas no lineales [6, 23]. Si bien sus fundamentos matemáticos fueron establecidos hace tiempo [22, 20], su aplicación a modelos de tamaño medio o grande no resulta trivial, y hasta tiempos recientes no había implementaciones disponibles de algoritmos adecuados para este tipo de problemas. La publicación de STRIKE-GOLDD [27] supuso un avance en este sentido. STRIKE-GOLDD es una toolbox de MATLAB, disponible en acceso abierto, que implementa un conjunto de algoritmos basados en geometría diferencial aplicables a modelos de ecuaciones diferenciales ordinarias no lineales.

En este artículo aplicamos STRIKE-GOLDD a un modelo de relevancia biomédica, la vía de señalización celular JAK/STAT. Las vías o cascadas de señalización celular son redes de reacciones bioquímicas que transmiten información desde el exterior de la célula a su interior y crean la respuesta adecuada [1]. La vía JAK/STAT está involucrada en procesos de inmunidad, proliferación y división celular, apoptosis y oncogénesis. Un modelo clásico de esta vía fue publicado por Bachmann et al. en 2011 [3]. En este artículo analizamos la observabilidad e identificabilidad estructural de este modelo. Si bien su identificabilidad paramétrica había sido analizada numéricamente en la publicación original, hasta donde tenemos conocimiento éste es el primer estudio que realiza este análisis desde un punto de vista estructural y de forma simbólica. Asimismo estudiamos la observabilidad de este modelo, reportando una falla de observabilidad teórica e indicando formas de corregirla. Como contribución adicional, la descripción detallada del análisis de este caso de estudio proporciona un tutorial sobre el uso de STRIKE-GOLDD, al mismo tiempo que ilustra sus capacidades y limitaciones.

2. MÉTODOS Y MODELO

2.1. ANÁLISIS DE OBSERVABILIDAD

2.1.1. Observabilidad lineal

A modo ilustrativo se presenta en esta sección la formulación del problema de observabilidad para el caso de sistemas lineales invariantes en el tiempo, que es una simplificación del caso más general de sistemas no lineales. En este caso los modelos a analizar se pueden representar como:

$$M_L : \begin{cases} \dot{x}(t) &= A(\theta)\cdot x(t) + B(\theta)\cdot u(t), \\ y(t) &= C(\theta)\cdot x(t), \\ x_0 &= x(t_0,\theta) \end{cases} \tag{1}$$

donde $\theta \in \mathbb{R}^q$ es el vector de parámetros, $u(t) \in \mathbb{R}^r$ el vector de entradas, $x(t) \in \mathbb{R}^n$ el vector de estados, y $y(t) \in \mathbb{R}^m$ el de salidas. $A(\theta)$, $B(\theta)$ y $C(\theta)$ son matrices constantes de dimensiones $n\times n$, $n\times r$ y $m \times n$, respectivamente. Omitiremos la dependencia de θ por simplicidad. La observabilidad de M_L se determina mediante la siguiente condición:

Teorema 1. *Un modelo lineal invariante en el tiempo definido por (1) es observable si y sólo si el rango de su matriz de observabilidad lineal, $\mathcal{O}^L$, es igual al número de estados n, siendo $\mathcal{O}^L = \left(C|C\cdot A|C\cdot A^2|\ldots|C\cdot A^{n-1}\right)^T$ [13].*

2.1.2. Observabilidad no lineal

Consideremos ahora el caso más general de modelos no lineales definidos por un sistema de ecuaciones diferenciales ordinarias:

$$M_{NL} : \begin{cases} \dot{x}(t) &= f(x(t),u(t),\theta), \\ y(t) &= g(x(t),\theta), \\ x_0 &= x(t_0,\theta) \end{cases} \tag{2}$$

donde f y g son vectores de funciones analíticas. La correspondiente matriz de observabilidad no lineal, $\mathcal{O}^{NL}$, se calcula mediante derivadas de Lie. Para el caso de entradas dependientes del tiempo, $u(t)$, las derivadas de Lie se definen como:

Definición 1. *La derivada de Lie de $g(x)$ respecto de $f(x)$ es:*

$$L_f g(x) = \frac{\partial g(x)}{\partial x} f(x,u) \tag{3}$$

donde $u^{(j)}$ es la derivada de orden j de la entrada u. Las derivadas de orden superior a uno se

calculan recursivamente:

$$L_f^i g(\tilde{x}) = \frac{\partial L_f^{i-1} g(\tilde{x})}{\partial \tilde{x}} f(\tilde{x}, u) + \sum_{j=0}^{j=i-1} \frac{\partial L_f^{i-1} g(\tilde{x})}{\partial u^{(j)}} u^{(j+1)} \quad (4)$$

Las derivadas obtenidas mediante las fórmulas (3,4) se denominan a veces derivadas *extendidas* de Lie [14], para remarcar que se consideran entradas posiblemente variantes en el tiempo, $u(t)$.

La matriz de observabilidad no lineal, $\mathcal{O}^{NL}$, se calcula a partir de las derivadas de Lie como:

$$\mathcal{O}^{NL}(x) = \begin{pmatrix} \frac{\partial}{\partial x} g(x) \\ \frac{\partial}{\partial x}(L_f g(x)) \\ \frac{\partial}{\partial x}(L_f^2 g(x)) \\ \vdots \\ \frac{\partial}{\partial x}(L_f^{n-1} g(x)) \end{pmatrix} \quad (5)$$

Es posible entonces formular la siguiente condición de observabilidad no lineal:

Teorema 2. *Sea un modelo M_{NL} dado por (2). Si el rango de su matriz de observabilidad $\mathcal{O}^{NL}$ es igual al número de estados (n) en un entorno de x_0, con $\mathcal{O}^{NL}(x_0)$ definida por (5), el modelo es localmente observable alrededor de x_0 [12, 22].*

2.1.3. Identificabilidad estructural

Los teoremas de observabilidad presentados en los apartados precedentes no consideran modelos con parámetros desconocidos. Si existe un vector de parámetros desconocidos, θ, debe ser tenido en cuenta a la hora de analizar la observabilidad. Asimismo, es de interés en este caso analizar si dichos parámetros son identificables desde un punto de vista estructural. Las siguientes definiciones formalizan este concepto.

Definición 2. *Un parámetro θ_i de un modelo M_{NL} dado por (2) es estructuralmente identificable localmente (s.l.i. en sus siglas en inglés, "structurally locally identifiable") si para casi cualquier vector $\theta^* \in \mathbb{R}^q$ existe un entorno $\mathcal{N}(\theta^*)$ tal que:*

$$\hat{\theta} \in \mathcal{N}(\theta^*),\ g(x, \hat{\theta}) = g(x, \theta^*) \Rightarrow \hat{\theta}_i = \theta_i^* \quad (6)$$

Definición 3. *Un parámetro θ_i es estructuralmente no identificable (s.u. en sus siglas en inglés, "structurally unidentifiable") si la condición (6) no se cumple en ningún entorno de θ^*.*

Definición 4. *Un modelo M_{NL} es s.l.i. si todos sus parámetros son s.l.i..*

Definición 5. *Un modelo M_{NL} is s.u. si al menos uno de sus parámetros es s.u..*

Si los parámetros son considerados como variables de estado de dinámica cero, la identificabilidad estructural se puede estudiar de la misma forma que la observabilidad, como un caso particular [21, 19, 2]. Para ello se aumenta el vector de estados con los parámetros desconocidos:

$$\tilde{x} = \begin{bmatrix} x \\ \theta \end{bmatrix} \quad (7)$$

De esta forma, al igual que la matriz de observabilidad no lineal (5), se puede definir una matriz de observabilidad-identificabilidad $\mathcal{O}_I^{NL}(\tilde{x})$:

$$\mathcal{O}_I^{NL}(\tilde{x}) = \begin{pmatrix} \frac{\partial}{\partial \tilde{x}} g(\tilde{x}) \\ \frac{\partial}{\partial \tilde{x}}(L_f g(\tilde{x})) \\ \frac{\partial}{\partial \tilde{x}}(L_f^2 g(\tilde{x})) \\ \vdots \\ \frac{\partial}{\partial \tilde{x}}(L_f^{n+q-1} g(\tilde{x})) \end{pmatrix} \quad (8)$$

Y asimismo se formula la condición de observabilidad-identificabilidad correspondiente:

Teorema 3. *Si un modelo M_{NL} dado por (2) es tal que rank$(\mathcal{O}_I^{NL}(\tilde{x}_0)) = n + q$, con $\mathcal{O}_I^{NL}(\tilde{x}_0)$ dado por (8), entonces M_{NL} es observable y estructuralmente identificable localmente en un entorno $\mathcal{N}(\tilde{x}_0)$ de $\tilde{x}_0$.*

2.2. IMPLEMENTACIÓN: LA TOOLBOX STRIKE-GOLDD

La metodología descrita en la sección 2.1 permite en principio el análisis de la observabilidad e identificabilidad estructural de cualquier modelo en ecuaciones diferenciales ordinarias, siempre y cuando sus entradas sean funciones conocidas e infinitamente diferenciables. En la práctica, sin embargo, tanto la construcción de la matriz de observabilidad-identificabilidad como el cálculo de su rango son operaciones simbólicas cuya complejidad puede resultar muy elevada para modelos de gran tamaño o para los que sea necesario calcular un gran número de derivadas de Lie, especialmente para ciertos tipos de no-linealidades.

La herramienta software STRIKE-GOLDD [27, 28, 29] implementa una metodología que integra los cálculos descritos en la sección 2.1 con una serie de mejoras algorítmicas que facilitan su aplicación a modelos de complejidad no trivial. Entre otras características, implementa: un flujo de trabajo que calcula el mínimo número de derivadas de Lie necesarias, incrementándolo iterativamente; un procedimiento para calcular la observabilidad o identificabilidad de estados y parámetros individuales; la posibilidad de descomponer el modelo en submodelos más fácilmente tratables, descomposición que puede ser definida por el usuario o

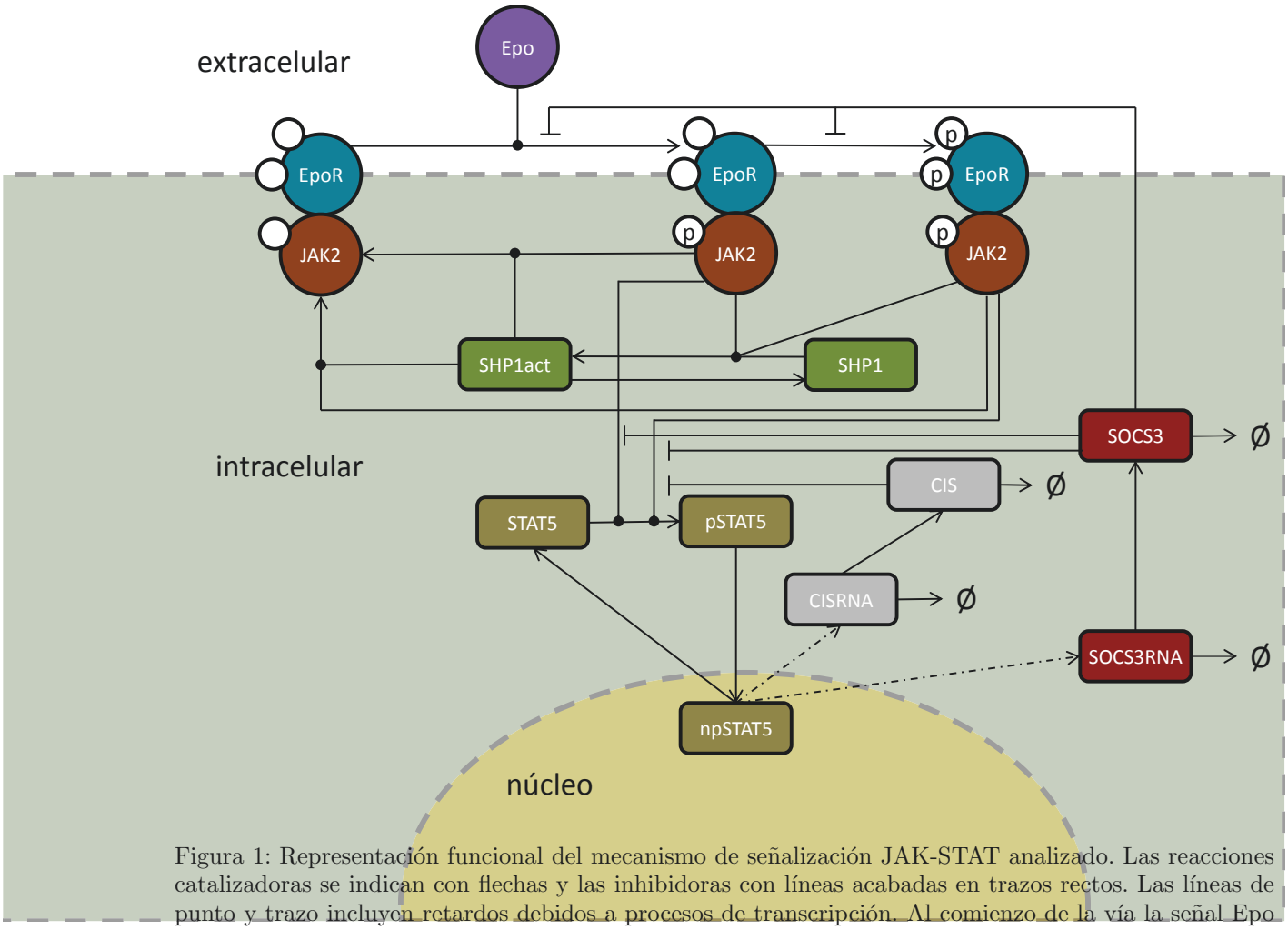

Figura 1: Representación funcional del mecanismo de señalización JAK-STAT analizado. Las reacciones catalizadoras se indican con flechas y las inhibidoras con líneas acabadas en trazos rectos. Las líneas de punto y trazo incluyen retardos debidos a procesos de transcripción. Al comienzo de la vía la señal Epo activa JAK2; al final del proceso, las proteínas SOCS3 y CIS participan en una realimentación negativa que inhibe el mecanismo.

mediante un algoritmo de optimización combinatoria [10]; y la consideración de entradas cuyas derivadas de un determinado orden se anulan, lo que permite analizar la observabilidad para entradas de tipos genéricos (constantes, rampas, etc). Más detalles de la metodología se proporcionan en [27, 28, 29].

STRIKE-GOLDD es una toolbox de MATLAB desarrollada en acceso abierto, cuyo histórico de versiones es accesible en `https://sites.google.com/site/strikegolddtoolbox/`. La última versión se encuentra disponible también en GitHub: `https://github.com/afvillaverde/strike-goldd_2.1`.

Para facilitar la reproducción de los cálculos realizados en este trabajo, hemos añadido la implementación del caso de estudio descrito en la Subsección 2.3 y analizado en la Sección 3 a la última versión de STRIKE-GOLDD, dentro de la carpeta 'models'.

2.3. CASO DE ESTUDIO: LA VÍA DE SEÑALIZACIÓN CELULAR JAK-STAT

Una forma de comunicación entre células es la liberación de moléculas que actúan como señales bioquímicas. Una célula puede detectar una determinada señal si tiene el receptor adecuado. La activación de dicho receptor se denomina transducción de señal. A continuación la señal es transmitida al interior de la célula y propagada por la vía de señalización correspondiente, hasta que eventualmente se desencadena una respuesta.

JAK-STAT son las siglas de "JAnus Kinase-Signal Transducer and Activator of Transcription", es decir, una vía de transducción de señales a través de Janus quinasas transductoras de señal y activadoras de la transcripción. La vía JAK-STAT participa en procesos de inmunidad, proliferación y división celular, apoptosis y oncogénesis. Su desregulación está asociada a enfermedades tales como síndromes de deficiencia inmunitaria y cánceres.

Su importancia ha ocasionado un gran interés en el modelado de la vía JAK-STAT por parte de las comunidades biomédica y de biología de sistemas. Un modelo ampliamente usado es el presentado en [3], una variante del cual ha sido incluida recientemente en una colección de benchmarks de estimación de parámetros [11]. Un esquema gráfico de este modelo se muestra en la Fig. 1. Se trata de un modelo con 25 estados dinámicos, 14 salidas, 5 entradas constantes, y 27 parámetros desconocidos. Las 14 salidas son funciones no lineales de los estados y parámetros. Sus ecuaciones son:

$$
\begin{aligned}
\dot{x}_1 &= x_{2345}x_8\theta_{11}/\theta_{26} - k_5x_1\theta_{10}/M_1,\\
\dot{x}_2 &= k_5x_1\theta_{10}/M_1 - x_2\theta_7/M_1 - x_2x_8\theta_{11}/\theta_{26}\ldots\\
&\ldots\ -3x_2\theta_7/((\theta_8x_6+1)M_1),\\
\dot{x}_3 &= \theta_7x_2/M_1 - \theta_{11}x_8x_3/\theta_{26}\ldots\\
&\ldots\ -3\theta_7x_3/((\theta_8x_6+1)M_1),\\
\dot{x}_4 &= 3x_2\theta_7/((\theta_8x_6+1)M_1) - \theta_7x_4/M_1\ldots\\
&\ldots\ -\theta_{11}x_8x_4/\theta_{26},\\
\dot{x}_5 &= \theta_7x_4/M_1 - \theta_{11}x_8x_5/\theta_{26}\ldots\\
&\ldots\ +3\theta_7x_3/((\theta_8x_6+1)M_1),\\
\dot{x}_6 &= -x_6(\theta_9/\theta_{25})(x_5+x_3),\\
\dot{x}_7 &= \theta_{13}x_8 - x_7(\theta_{12}/\theta_{25})x_{2345},\\
\dot{x}_8 &= x_7(\theta_{12}/\theta_{25})x_{2345} - \theta_{13}x_8,\\
\dot{x}_9 &= k_6\theta_{23}x_{11}/k_7 - x_9(\theta_{22}/\theta_{25})x_{2345}/M_1 - \ldots\\
&\ldots\ x_9\theta_{21}(x_5+x_3)^2/((x_{18}\theta_3/\theta_1+1)M_1\theta_{25}^2),\\
\dot{x}_{10} &= x_9\theta_{22}x_{2345}M_1/\theta_{25} - \theta_{24}x_{10}\ldots\\
&\ldots\ +x_9\theta_{21}(x_5+x_3)^2/(\theta_{25}^2(x_{18}\theta_3/\theta_1+1)M_1),\\
\dot{x}_{11} &= k_7\theta_{24}x_{10}/k_6 - \theta_{23}x_{11},\\
\dot{x}_{12} &= -x_{12}\theta_4 - \theta_5x_{11}(k_1-1)/\theta_{27},\\
\dot{x}_{13} &= x_{12}\theta_4 - x_{13}\theta_4,\\
\dot{x}_{14} &= x_{13}\theta_4 - x_{14}\theta_4,\\
\dot{x}_{15} &= x_{14}\theta_4 - x_{15}\theta_4,\\
\dot{x}_{16} &= x_{15}\theta_4 - x_{16}\theta_4,\\
\dot{x}_{17} &= x_{16}\theta_4k_6/k_7 - x_{17}\theta_5,\\
\dot{x}_{18} &= x_{17}\theta_1\theta_6 - x_{18}\theta_6 + k_2\theta_6\theta_2\theta_1,\\
\dot{x}_{19} &= -x_{19}\theta_{18} - \theta_{19}x_{11}(k_1-1)/\theta_{27},\\
\dot{x}_{20} &= x_{19}\theta_{18} - x_{20}\theta_{18},\\
\dot{x}_{21} &= x_{20}\theta_{18} - x_{21}\theta_{18},\\
\dot{x}_{22} &= x_{21}\theta_{18} - x_{22}\theta_{18},\\
\dot{x}_{23} &= x_{22}\theta_{18} - x_{23}\theta_{18},\\
\dot{x}_{24} &= k_6x_{23}\theta_{18}/k_7 - x_{24}\theta_{19},\\
\dot{x}_{25} &= x_{24}\theta_{15}\theta_{20} - x_{25}\theta_{20} + k_3\theta_{20}\theta_{16}\theta_{15}\\
y_1 &= 2(x_2+x_3+x_4+x_5)\theta_{25},\\
y_2 &= 16(x_3+x_4+x_5)\theta_{25},\\
y_3 &= x_{18}\theta_1,\\
y_4 &= x_{25}/\theta_{14},\\
y_5 &= (x_9+x_{10})/\theta_{27},\\
y_6 &= x_{10}\theta_{27},\\
y_7 &= x_9,\\
y_8 &= x_7+x_8,\\
y_9 &= x_{18},\\
y_{10} &= x_{25},\\
y_{11} &= 100x_{10}/(x_{10}+x_9),\\
y_{12} &= x_{24},\\
y_{13} &= x_{17},\\
y_{14} &= (x_7+x_8)(1+(k_4\theta_{27}))/\theta_{26},
\end{aligned}
\tag{9}
$$

donde se han usado las variables auxiliares $x_{2345} = x_2+x_3+x_4+x_5$ y $M_1 = x_{25}\theta_{17}/\theta_{15}+1$. Los estados $x_1, x_2, \ldots, x_{25}$ son, respectivamente, las especies EpoRJAK2, EpoRpJAK2, p1EpoRpJAK2, p2EpoRpJAK2, p12EpoRpJAK2, EpoRJAK2_CIS, SHP1, SHP1Act, STAT5, pSTAT5, npSTAT5, CISnRNA1, CISnRNA2, CISnRNA3, CISnRNA4, CISnRNA5, CISRNA, CIS, SOCS3nRNA1, SOCS3nRNA2, SOCS3nRNA3, SOCS3nRNA4, SOCS3nRNA5, SOCS3RNA y SOCS3. Para los 27 parámetros desconocidos, escritos como θ_i en (9), se usaron las siguientes abreviaturas en la publicación original [3]: CISEqc, CISEqcOE, CISInh, CISRNADelay, CISRNATurn, CISTurn, EpoRActJAK2, EpoRCISInh, EpoRCISRemove, JAK2ActEpo, JAK2EpoRDeaSHP1, SHP1ActEpoR, SHP1Dea, SHP1ProOE, SOCS3Eqc, SOCS3EqcOE, SOCS3Inh, SOCS3RNADelay, SOCS3RNATurn, SOCS3Turn, STAT5ActEpoR, STAT5ActJAK2, STAT5Exp, STAT5Imp, init_EpoRJAK2, init_SHP1, y init_STAT5. El modelo tiene 7 constantes conocidas (k_1–k_7), 5 de las cuales se corresponden con condiciones experimentales que se pueden considerar como entradas constantes (k_1–k_5), entre las que se incluye la señal de entrada externa ($k_5 \equiv$ Epo).

En [3] se analizó numéricamente la identificabilidad de los parámetros del modelo usando el método del "profile likelihood" [18], concluyendo que seis de ellos ($\theta_5 \equiv$ CISRNATurn, $\theta_8 \equiv$ EpoRCISInh, $\theta_{11} \equiv$ JAK2EpoRDeaSHP1, $\theta_{12} \equiv$ SHP1ActEpoR, $\theta_{18} \equiv$ SOCS3RNADelay y $\theta_{20} \equiv$ SOCS3Turn) son prácticamente no identificables. Hasta la fecha no conocemos ningún estudio que haya realizado un análisis simbólico de la identificabilidad estructural de los parámetros de este modelo, ni de la observabilidad de sus estados.

3. RESULTADOS

Como primer análisis con STRIKE-GOLDD consideramos el problema sin descomponer el modelo, y especificamos un tiempo máximo corto (100 segundos) para el cálculo de cada derivada de Lie. Asimismo, y dado que las entradas son típicamente constantes para cada condición experimental, ponemos a cero el número de derivadas de las entradas distintas de cero (ésta es una elección conservadora, ya que la observabilidad e identificabilidad podrían mejorar considerando entradas variantes en el tiempo). Para ello modificamos los siguientes campos en el fichero `options.m`:

```
modelname = 'BachmannJAKSTAT';
opts.maxLietime = 100;
opts.nnzDerU = [0 0 0 0 0];
```

Al inspeccionar el modelo encontramos que uno de los parámetros a estimar, $\theta_{27} \equiv$ init_STAT5, es la condición inicial de un estado observado, $y_7 = x_9 \equiv$ STAT5. Dado que conceptualmente la identificabilidad estructural no contempla ningún tipo de restricción en cuanto a la precisión de las medidas disponibles, se puede asumir que este parámetro es directamente medible y por tanto conocido. En lugar de modificar el modelo para reflejar este conocimiento, lo especificamos en `options.m` como:

```
syms init_STAT5
prev_ident_pars = init_STAT5;
```

Una vez modificado el fichero de opciones, corremos el algoritmo ejecutando `STRIKE_GOLDD.m`. Esta primera ejecución determina que al menos 17 de los 26 parámetros desconocidos son estructuralmente identificables (todos menos $\theta_4 \equiv$ CISRNADelay, $\theta_5 \equiv$ CISRNATurn, $\theta_8 \equiv$ EpoRCISInh, $\theta_{11} \equiv$ JAK2EpoRDeaSHP1, $\theta_{14} \equiv$ SHP1ProOE, $\theta_{18} \equiv$ SOCS3RNADelay, $\theta_{19} \equiv$ SOCS3RNATurn, $\theta_{25} \equiv$ init_EpoRJAK2 y $\theta_{26} \equiv$ init_SHP1), y 4 de los estados son observables ($x_7 \equiv$ SHP1, $x_8 \equiv$ SHP1Act, $x_{10} \equiv$ pSTAT5 y $x_{11} \equiv$ npSTAT5), además de los 5 estados directamente medidos ($x_9 \equiv$ STAT5, $x_{17} \equiv$ CISRNA, $x_{18} \equiv$ CIS, $x_{24} \equiv$ SOCS3RNA y x_{25} SOCS3). El algoritmo no se pronuncia sobre el resto de variables, ya que necesitaría calcular la matriz de observabilidad-identificabilidad con más derivadas de Lie, pero el límite de tiempo impuesto para el cálculo de cada derivada (100 segundos) interrumpe su ejecución después de 5 derivadas. Para intentar obtener resultados sobre los parámetros restantes, repetimos el análisis con dos cambios:

1. Especificamos en el fichero `options.m` los 17 parámetros identificables dentro de la variable `prev_ident_pars`, tal y como hicimos anteriormente con `init_STAT5`.
2. Forzamos la descomposición del modelo (`opts.forcedecomp= 1`) e indicamos que los submodelos sean determinados por el algoritmo de optimización, no por el usuario (`opts.decomp_user= 0`). Para ello es necesaria la toolbox MEIGO [10].

De esta forma determinamos la identificabilidad de 5 parámetros adicionales ($\theta_8 \equiv$ EpoRCISInh, $\theta_{11} \equiv$ JAK2EpoRDeaSHP1, $\theta_{14} \equiv$ SHP1ProOE, $\theta_{25} \equiv$ init_EpoRJAK2 y $\theta_{26} \equiv$ init_SHP1).

Al quedar sólo 4 parámetros cuya identificabilidad no ha sido determinada, la dimensión del problema se ha reducido considerablemente. Probamos pues a realizar de nuevo el análisis completo, incluyendo los 22 parámetros ya determinados dentro del vector `prev_ident_pars`, y aumentamos el tiempo de cálculo a `opts.maxLietime = 5000`. De esta forma, el algoritmo es capaz de calcular 6 derivadas de Lie. Tanto con la derivada número 5 como con la 6 el rango de la matriz es el mismo (28), lo que indica que se puede realizar un análisis concluyente. Se determina así que todos los parámetros son estructuralmente identificables y que 2 estados, p1EpoRpJAK2 (x_3) y p12EpoRpJAK2 (x_5) son no observables. Estos estados corresponden a diferentes estados de fosforilación del receptor Epo por parte de las proteínas JAK. La falta de observabilidad es causada por la estructura del modelo y la función de salida, ya que se mide el conjunto p1EpoRpJAK2 + p2EpoRpJAK2 + p12EpoRpJAK2, pero no cada uno de los estados fosforilados por separado. Al no ser observables, no es posible usar el modelo para monitorizar o predecir en detalle la fosforilación del receptor.

En ocasiones la falta de observabilidad o identificabilidad puede ser remediada con entradas de dinámica más rica. STRIKE-GOLDD permite explorar esta posibilidad [28], ya que es posible especificar el número de derivadas distintas de cero de las entradas mediante la opción `opts.nnzDerU`. Sin embargo, en sistemas de señalización como el descrito las limitaciones experimentales a menudo no permiten aplicar entradas variantes en el tiempo. Otra posibilidad, también mencionada en [28], es utilizar varios experimentos con entradas constantes de distintos valores, en lugar de un único experimento. Si bien esta posibilidad no está implementada automáticamente en STRIKE-GOLDD, sí que es posible realizar este análisis. Para ello, se redefine el modelo creando tantas copias de los estados y entradas como experimentos se quiera considerar. Para el modelo JAK-STAT estudiado, la repetición del análisis con entradas variables en el tiempo o con múltiples experimentos de entradas constantes no altera el resultado: ambos estados siguen siendo no observables.

Añadiendo alguno de los dos estados, p2EpoRpJAK2 o p12EpoRpJAK2, como salida adicional al modelo, el análisis confirma que en caso de ser posible medir directamente alguno de los dos estados directamente el otro pasaría a ser observable.

Alternativamente, una inspección atenta de las ecuaciones dinámicas del modelo revela que es posible agrupar los dos estados en uno sólo. Si reformulamos el modelo de esta forma,

$$x_{\text{nuevo}} = \text{p1EpoRpJAK2+p12EpoRpJAK2}, \quad (10)$$

se obtiene un modelo de idéntica dinámica que el original, con una dimensión menos, y totalmente observable y estructuralmente identificable.

4. CONCLUSIONES

En este trabajo se ha analizado la observabilidad e identificabilidad de un modelo no lineal de una vía de señalización celular de relevancia biomédica, la JAK-STAT. El análisis, realizado de forma simbólica con la toolbox de MATLAB STRIKE-GOLDD, ha revelado que el modelo es estructuralmente identificable localmente pero posee dos estados no observables. Dicha falta de observabilidad es fácilmente remediable mediante una reformulación del modelo.

Es conveniente analizar la observabilidad e identificabilidad estructural antes de usar un modelo con fines predictivos, para descartar conclusiones potencialmente erróneas. El presente trabajo ha ilustrado la aplicación de una metodología general para este propósito, mostrando cómo interpretar los resultados y cómo tomar las acciones necesarias para solucionar faltas de observabilidad. En el caso estudiado dichas acciones pueden ser una reformulación de las ecuaciones dinámicas del modelo, reduciendo el número de estados, o bien un incremento del número de estados medidos.

Agradecimientos

Este trabajo ha recibido financiación del programa Horizonte 2020 de la Unión Europea, a través del proyecto 686282 (CanPathPro), y del proyecto del Ministerio de Ciencia, Innovación y Universidades SYNBIOCONTROL (DPI2017-82896-C2-2-R).

English summary

STRUCTURAL IDENTIFIABILITY AND OBSERVABILITY ANALYSIS OF A JAK/STAT SIGNALLING PATHWAY MODEL

Abstract

Observability and structural identifiability are classic systemic properties that describe whether it is theoretically possible to determine the state and parameter vectors of a model from its output. Recently we developed a MATLAB tool, STRIKE-GOLDD, which adopts a symbolic approach for analysing observability and structural identifiability of nonlinear ODE models. While being a general purpose tool, the motivation behind the development of STRIKE-GOLDD was the study of biological models, for which these analyses can provide very valuable insight. This paper demonstrates the application of STRIKE-GOLDD to a dynamic model of the JAK/STAT signaling pathway, which plays an important role in processes such as immunity, cell division, cell death and tumour formation. Using this model of biomedical relevance as a case study, the present paper provides a tutorial on the use of the STRIKE-GOLDD methodology, highlighting its capabilities and limitations.

Keywords: Computational methods, Dynamic modelling, Nonlinear systems, Observability, Identifiability.

Referencias

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell*. Garland Science, 2002.

[2] E. August and A. Papachristodoulou. A new computational tool for establishing model parameter identifiability. *Journal of Computational Biology*, 16(6):875–885, 2009.

[3] J. Bachmann, A. Raue, M. Schilling, M. E. Böhm, C. Kreutz, D. Kaschek, H. Busch, N. Gretz, W. D. Lehmann, J. Timmer, et al. Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Molecular Systems Biology*, 7(1):516, 2011.

[4] G. Bekey and J. Beneken. Identification of biological systems: a survey. *Automatica*, 14(1):41–47, 1978.

[5] R. Bellman and K. J. Åström. On structural identifiability. *Mathematical Biosciences*, 7(3):329–339, 1970.

[6] M. N. Chatzis, E. N. Chatzi, and A. W. Smyth. On the observability and identifiability of nonlinear structural and mechanical systems. *Structural Control and Health Monitoring*, 22(3):574–593, 2015.

[7] C. Cobelli and J. DiStefano. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. *American Journal of Physiology*, 239(1):R7–R24, 1980.

[8] J. DiStefano III. On the relationships between structural identifiability and the controllability, observability properties. *IEEE Transactions on Automatic Control*, 22(4):652–652, 1977.

[9] J. DiStefano III. *Dynamic systems biology modeling and simulation*. Academic Press, 2015.

[10] J. A. Egea, D. Henriques, T. Cokelaer, A. F. Villaverde, A. MacNamara, D.-P. Danciu, J. R. Banga, and J. Saez-Rodriguez. MEIGO: an open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics. *BMC Bioinf.*, 15:136, 2014.

[11] H. Hass, C. Loos, E. R. Alvarez, J. Timmer, J. Hasenauer, and C. Kreutz. Benchmark problems for dynamic modeling of intracellular processes. *Bioinformatics*, page btz020, 2019.

[12] R. Hermann and A. J. Krener. Nonlinear controllability and observability. *IEEE Transactions on Automatic Control*, 22(5):728–740, 1977.

[13] R. E. Kalman. Contributions to the theory of optimal control. *Boletín de la Sociedad Matemática Mexicana*, 5(2):102–119, 1960.

[14] J. Karlsson, M. Anguelova, and M. Jirstrand. An efficient method for structural identiability analysis of large dynamic systems. In *16th IFAC Symposium on System Identification*, volume 16, pages 941–946, 2012.

[15] Y. M. Kostyukovskii. Simple conditions of observability of nonlinear controlled systems. *Avtomatika i Telemekhanika*, (10):32–41, 1968.

[16] J. Picó, A. Vignoni, E. Picó-Marco, and Y. Boada. Modelado de sistemas bioquímicos: De la ley de acción de masas a la aproximación lineal del ruido. *Revista Iberoamericana de Automática e Informática Industrial RIAI*, 12(3):241–252, 2015.

[17] H. Pohjanpalo. System identifiability based on the power series expansion of the solution. *Mathematical Biosciences*, 41(1):21–33, 1978.

[18] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

[19] A. Sedoglavic. A probabilistic algorithm to test local algebraic observability in polynomial time. *Journal of Symbolic Computation*, 33:735–755, 2002.

[20] E. D. Sontag. *Mathematical control theory: deterministic finite dimensional systems*. Springer, New York, USA, 1998.

[21] E. T. Tunali and T.-J. Tarn. New results for identifiability of nonlinear systems. *IEEE Transactions on Automatic Control*, 32(2):146–154, 1987.

[22] M. Vidyasagar. *Nonlinear systems analysis*. Prentice Hall, Englewood Cliffs, NJ, 1993.

[23] A. F. Villaverde. Observability and structural identifiability of nonlinear biological systems. *Complexity*, 2019:8497093, 2019.

[24] A. F. Villaverde and J. R. Banga. Reverse engineering and identification in systems biology: strategies, perspectives and challenges. *Journal of the Royal Society Interface*, 11(91):20130505, 2014.

[25] A. F. Villaverde and J. R. Banga. Structural properties of dynamic systems biology models: Identifiability, reachability, and initial conditions. *Processes*, 5(2):29, 2017.

[26] A. F. Villaverde and A. Barreiro. Identifiability of large nonlinear biochemical networks. *MATCH Communications in Mathematical and in Computer Chemistry*, 76(2):259–276, 2016.

[27] A. F. Villaverde, A. Barreiro, and A. Papachristodoulou. Structural identifiability of dynamic systems biology models. *PLoS Computational Biology*, 12(10):e1005153, 2016.

[28] A. F. Villaverde, N. D. Evans, M. J. Chappell, and J. R. Banga. Input-dependent structural identifiability of nonlinear systems. *IEEE Control Systems Letters*, 3(2):272–277, 2019.

[29] A. F. Villaverde, N. Tsiantis, and J. R. Banga. Full observability and estimation of unknown inputs, states, and parameters of nonlinear biological models. *Journal of the Royal Society Interface*, 16(156), 2019.

[30] E. Walter and L. Pronzato. *Identification of parametric models from experimental data*. Communications and Control Engineering Series. Springer, London, UK, 1997.