PROGRAMA DE DOCTORADO EN ESTADÍSTICA, OPTIMIZACIÓN Y MATEMÁTICA APLICADA (EOMA)

TESIS DOCTORAL:

Estudio teórico y práctico de Ecuaciones Simultáneas Multinivel

Autora:

Rocío Hernández Sanjaime

Director:

José Juan López Espín

Universidad Miguel Hernández de Elche

2023

UNIVERSITAS Miguel Hernández
INSTITUTO DE INVESTIGACIÓN

Financiación

Esta investigación ha sido financiada por la Generalitat Valenciana y el Fondo Social Europeo a través de la ayuda predoctoral ACIF/2018/219.

El Dr. D. José Juan López Espín, director de la tesis doctoral titulada **"***Estudio teórico y práctico de Ecuaciones Simultáneas Multinivel***"**

INFORMA:

Que D. Rocío Hernández Sanjaime ha realizado bajo mi supervisión el trabajo titulado **"***Estudio teórico y práctico de Ecuaciones Simultáneas Multinivel***"** conforme a los términos y condiciones definidos en su Plan de Investigación y de acuerdo al Código de Buenas Prácticas de la Universidad Miguel Hernández de Elche, cumpliendo los objetivos previstos de forma satisfactoria para su defensa pública como tesis doctoral.

Lo que firmo para los efectos oportunos, en Elche a 1 de Septiembre de 2023

Director de la tesis

Dr. D José Juan López Espín

El Dr. D. Domingo Morales González, Coordinador del Programa de Doctorado en Estadística, Optimización y Matemática Aplicada (EOMA)

INFORMA:

Que Dña. Rocío Hernández Sanjaime ha realizado bajo la supervisión de nuestro Programa de Doctorado el trabajo titulado *"Estudio teórico y práctico de Ecuaciones Simultáneas Multinivel"* conforme a los términos y condiciones definidos en su Plan de Investigación y de acuerdo al Código de Buenas Prácticas de la Universidad Miguel Hernández de Elche, cumpliendo los objetivos previstos de forma satisfactoria para su defensa pública como tesis doctoral.

Lo que firmo para los efectos oportunos, en Elche a 1 de Septiembre de 2023

Prof. Dr. D. Domingo Morales González

Coordinador del Programa de Doctorado en Estadística, Optimización y Matemática Aplicada

A Jessi y a mamá

Agradecimientos

Esta tesis representa el trabajo de varios años y momentos que no siempre han sido fáciles, llenos de incertidumbre y dudas que hubieran sido una carga sin las personas que me han acompañado a lo largo de toda esta etapa.

En primer lugar, he de expresar mi agradecimiento a las instituciones que han apoyado mi formación como investigadora y que han financiado esta tesis para que fuera posible: la Generalitat Valenciana y el Fondo Social Europeo mediante la ayuda predoctoral ACIF/2018/219, el Ministerio de Economía y Competitividad a través del proyecto TIN2016-8056-R y a la Universidad Miguel Hernández de Elche a través de las ayudas para la asistencia a cursos, congresos o jornadas.

En segundo lugar, esta tesis no hubiera existido sin el empeño de mi director Jose Juan para persuadirme a matricularme en los estudios de doctorado cuando una tesis no entraba en mis planes. Gracias por tu apoyo, dedicarme tu tan valioso y escaso tiempo, horas en llamadas, tu paciencia y comprensión en todo momento incluyendo los más difíciles. Gracias por preocuparte no solo por la tesis, sino también por mí. Pero sobre todo, gracias por tu carácter alegre y por tu extraordinario optimismo que han equilibrado mi visión objetiva (no diré negativa) para obtener una media realista de todo este proceso.

Por otro lado, esta tesis tampoco hubiera sido posible sin la ayuda inestimable de Martín, que de forma colateral ha invertido muchas horas delante de un ordenador por mi culpa. Y por supuesto, el resto de compañeros de servicio y UMH, que me han acompañado en estos años haciendo que trabajar fuera más divertido, en especial, aquellos de la fila random Roberto y Yaro de quienes me llevo una bonita amistad.

Y en la prosa no pueden faltar las personas con las que me adentré en las matemáticas Bea, María Pilar, Pily y Ana que hicieron que la carrera fuera más llevadera. En especial, he de agradecer a Bea, con quien desde la distancia compartía inquietudes y me dió la confianza y el empujón que me faltaba para empezar esta tesis.

Por último, me queda agradecer a toda aquellas personas de mi familia que siempre han estado presentes de una manera u otra todo este tiempo. A mi hermana Jessica, a quien tenía muy claro que dedicaría esta tesis, porque siempre escribía mi nombre junto al suyo y ahora me tocaba a mí. Por su amor incondicional y por transmitirme su afán de superación y sus ganas de aprender sin importar las limitaciones. A mi madre, simplemente por todo, porque es imposible resumir todo lo que le agradezceré siempre. Por toda la huella que deja en mi. Por enseñarme a ver más allá y luchar, por su energía, complicidad y por su apoyo todos estos años de trabajo. A mi hermano Jorge, que siempre se ha preocupado por mis estudios desde pequeña aunque no hiciera

falta y a mi padre, que aunque no entienda muy bien cómo funciona todo esto de la tesis también me apoya. A mis tías Coti y Amparo, por siempre estar ahí, preocuparse y quererme tanto, y por tantos y tantos cafés.

Tesis por compendio de publicaciones

La presente tesis doctoral titulada *"Estudio teórico y práctico de ecuaciones simultáneas multinivel"* se ha elaborado siguiendo la normativa de Estudios de Doctorado de la Universidad Miguel Hernández de Elche para la presentación de tesis doctorales en la modalidad de compendio de publicaciones (Artículo 18 de la Normativa de Estudios de Doctorado publicada el 31 de mayo de 2022). Las referencias completas de los artículos que constituyen el cuerpo de la tesis son:

- Hernández-Sanjaime, R.; González, M.; Peñalver, A.; and López-Espín, J. J. (2021). Estimating Simultaneous Equation Models through an Entropy-Based Incremental Variational Bayes Learning Algorithm. Entropy 2021, 23(4), 384. *DOI: 10.3390/e23040384*

- Hernández-Sanjaime, R.; González, M.; and López-Espín, J. J. (2020). Multilevel simultaneous equation model: A novel specification and estimation approach. Journal of Computational and Applied Mathematics, 366, 112378. *DOI: 10.1016/j.cam.2019.112378*

- Hernández-Sanjaime, R.; González, M.; and López-Espín, J. J. (2020). Estimation of Multilevel Simultaneous Equation Models through Genetic Algorithms. Mathematics, 8(12), 2098. *DOI: 10.3390/math8122098*

Índice general

Summary

Simultaneous Equation Models (SEMs) have been developed to reflect the presence of jointly dependent variables in a system of regression equations. In other words, these models allow to take account of the simultaneity between the set of variables integrating the model. SEMs are multi-equation models widely used in econometrics and social sciences like for example in the supply and demand model or the Keynesian model.

These models assume that intertemporally error terms are uncorrelated; however, if data present multilevel or grouped structure, this assumption does not always hold. Traditionally, the literature addresses the estimation of simultaneous equation models whether not testing this assumption or assuming that a bias may be induced in the estimation. Nevertheless, in practice, this simplification is often unrealistic and likely to produce misleading results. The three studies included in this work tackle this matter.

The objective of the present thesis is to analyse simultaneous equation models when data are grouped. To this end, this thesis is divided in three articles addressing different aspects of the mentioned problem:

- **Article 1.-** In this first study, estimation of simultaneous equation models in the presence of heterogeneity in data is reviewed. Within the framework of sequential approaches, a two-step strategy using a clustering method based on entropy is proposed in order to enhance SEM estimates in this context. The efficiency of the algorithm to identify group membership is examined through an exhaustive computational study. Moreover, the advantages of the proposed methodology are analysed and it is applied to a macroeconomic problem.

- **Article 2.-** In this contribution, a new model referred as to Multilevel Simultaneous Equation Model (MSEM) is introduced, that is, a SEM in which data present grouped structure; its characteristics are studied and an estimator for the parameters of the model is presented. Additionally, the application of an optimisation solver is suggested to obtain an approximation of the proposed theoretical estimator and results are compared with regard to other estimation methods such as Two-Stage Least Squares (2SLS).

- **Article 3.-** In this third work, we delve more deeply into the calculation of the proposed theoretical estimator for multilevel simultaneous equation models. In particular, the case in which covariance matrices are unknown is addressed. For

that purpose, a hybrid metaheuristic incorporating the joint use of a genetic algorithm with the preceding optimisation solver is developed. A simulation study evaluating the parameterized schema of the metaheuristic in order to select the optimal combination of values in the MSEM context is carried out. Lastly, the performance of the calculated estimates and the improvement in goodness of fit offered by the metaheuristic are analysed with regard to other estimation methods.

Resumen

Los Modelos de Ecuaciones Simultáneas (MES) tienen como finalidad reflejar la presencia de variables mutuamente dependientes en un sistema de ecuaciones de regresión, es decir, permiten tener en cuenta la simultaneidad entre el conjunto de variables que intervienen en el modelo. Se trata de modelos multiecuacionales ampliamente utilizados en econometría y ciencias sociales tales como el modelo de la oferta y la demanda o el modelo Keynesiano.

Estos modelos asumen que los términis de error están intertemporalmente incorrelacionados, pero cuando los datos presentan estructura multinivel o agrupada este supuesto no siempre se cumple. En la literatura, tradicionalmente los modelos de ecuaciones simultáneas se estiman bien sin comprobar este supuesto o asumiendo que puede cometerse un sesgo en la estimación. Sin embargo, en la práctica, esta simplificación es frecuentemente poco realista y puede conducir a resultados erróneos. Las tres aportaciones incluidas en esta tesis abordan esta problemática.

El objetivo de la presente tesis es analizar los modelos de ecuaciones simultáneas cuando los datos se encuentran agrupados. Para ello, esta tesis se ha estructurado en tres artículos que tratan distintos aspectos de este problema:

- **Artículo 1.-** En este primer trabajo, se revisa la estimación de los modelos de ecuaciones simultáneas en presencia de heterogeneidad en los datos. Desde el enfoque de los procedimientos secuenciales, se propone una estrategia en dos etapas que utiliza un algoritmo de clustering basado en entropía para mejorar la estimación de los MES en este contexto. Se estudia la eficacia del algoritmo para identificar agrupaciones mediante un completo estudio computacional, se analizan las ventajas de la metodología propuesta y se aplica a un problema macroeconómico.

- **Artículo 2.-** En esta aportación introducimos un nuevo modelo al que hemos denominado Modelo de Ecuaciones Simultáneas Multinivel (MESM), esto es, un MES en el que los datos se encuentran agrupados; se estudian sus características y se propone un estimador de los parámetros del modelo. Adicionalmente, se sugiere el uso de un *solver* de optimización cuya aplicación permite obtener una aproximación del estimador teórico propuesto y se comparan los resultados con otros métodos de estimación como mínimos cuadrados en dos etapas (MC2E).

- **Artículo 3.-** En este tercer trabajo se profundiza en la obtención del estimador teórico propuesto para los modelos de ecuaciones simultáneas multinivel y se

aborda el caso en el que las matrices de covarianzas no son conocidas. Para ello, se plantea una metaheurística híbrida que añade el uso de un algoritmo genético de forma conjunta al *solver* de optimización y se estudia mediante simulación el esquema parametrizado del algoritmo para así seleccionar la combinación óptima de valores en el contexto de los MESM. Finalmente, se analiza la calidad de las estimaciones obtenidas y la mejora en la bondad de ajuste de la metaheurística frente a otros métodos de estimación.

Capítulo 1

Introducción

La presente tesis se enmmarca dentro de la rama de la estadística dedicada a la econometría, disciplina definida en Griliches e Intriligator (1984) como la aplicación de las matemáticas y los métodos estadísticos al análisis de datos económicos [8]. En particular, este trabajo se centra en un tipo de modelo multiecuacional denominado Modelo de Ecuaciones Simultáneas (MES).

Además de definir e introducir la línea de investigación, en este capítulo se especifican los objetivos de esta tesis y se establece la relación existente entre los distintos artículos que la conforman. Comenzaremos esta introducción describiendo las caracterísicas de los modelos de ecuaciones simultáneas. A continuación, justificamos la necesidad de desarrollar un nuevo modelo estadístico que será en el que centraremos la investigación: los modelos de ecuaciones simultáneas multinivel. Finalmente, se describen brevemente los aspectos trabajados en cada una de las distintas aportaciones de la presente tesis así como la relación entre estas contribuciones.

1.1. Modelos de Ecuaciones Simultáneas

Los modelos estadísticos son herramientas que nos permiten adentrarnos en la complejidad de los fenómenos que investigamos. Ahora bien, las técnicas que elijamos deben reproducir de forma realista la complejidad del mundo que intentamos comprender. Tradicionalmente, los modelos estadísticos más extendidos y utilizados con este propósito han sido los modelos de regresión. No obstante, estos modelos en muchas ocasiones resultan ser modelos estadísticos demasiado sencillos o muy restrictivos.

En los modelos de regresión uniecuacionales, una variable (la variable dependiente), Y, se expresa como función lineal de una o más variables (las variables independientes o explicativas), X, dada por la expresión

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$$

La relación causa-efecto contemplada por estos modelos va de las variables explicativas X (causa) a la variable dependiente Y (efecto) de forma unidireccional. Sin

embargo, en determinadas situaciones esta relación se desdibuja porque existe una influencia bidireccional entre ambos tipos de variables. Es decir, la relación causa-efecto que va de las X a Y pierde sentido y se convierte en una relación de interdependencia.

Definición 1.1. Modelos de Ecuaciones Simultáneas
Los Modelos de Ecuaciones Simultáneas (MES) son modelos estadísticos formados por un sistema de ecuaciones de regresión que reflejan la *simultaneidad* entre el conjunto de variables que intervienen en el modelo [7].

En otras palabras, la característica única de los modelos de ecuaciones simultáneas es que en ellos una variable dependiente en una ecuación puede aparecer como variable explicativa en otra ecuación. Por consiguiente, los MES se utilizan cuando existen efectos de retroalimentación entre variables *mutuamente dependientes* y nos encontramos ante la necesidad de utilizar un modelo multiecuacional.

1.1.1. Descripción del modelo

La relación en dos sentidos entre las variables de los modelos de ecuaciones simultáneas hace que la denominación de variables dependientes y explicativas no sea la más precisa. En su lugar, en un modelo de ecuaciones simultáneas distinguimos tres tipos de variables:

- **Variables endógenas:** Aquellas cuyos valores están determinados *dentro* del modelo. Es decir, estas variables influyen en el modelo, pero a su vez se ven influenciadas por él. Se representan por Y_t y se consideran variables aleatorias. El número de variables endógenas determina el número de ecuaciones del modelo.

- **Variables predeterminadas**. Se dividen a su vez en dos categorías:

  - **Exógenas** (tanto presentes como rezagadas). Aquellas determinadas *fuera* del modelo. Son variables que influyen en el modelo, pero no se ven influenciadas por él. Se representan por X_t.
  - **Endógenas rezagadas**. Son variables endógenas rezagadas en el tiempo, es decir, su valor es conocido y no está por tanto, determinado por el modelo en el periodo de tiempo presente. Pueden entenderse como una extensión de las variables exógenas.

  Por simplificar la notación, todas las variables predeterminadas, ya sean exógenas o endógenas rezagadas se denotarán por X_t y se consideran varibales no aleatorias.

- **Variables de error**. Se denotan por u_t y son variables de ruido blanco, es decir, variables idénticamente distribuidas siguiendo una normal de media cero y desviación típica constante.

Formalmente, el modelo general de m ecuaciones con m variables endógenas y k variables predeterminadas es:

$$\begin{aligned}
y_{1t} = \quad & \alpha_{21}y_{2t} + \alpha_{31}y_{3t} + ... + \alpha_{m1}y_{mt} + \beta_{11}x_{1t} + \beta_{21}x_{2t} + ... + \beta_{k1}x_{kt} + u_{1t} \\
y_{2t} = \quad & \alpha_{12}y_{1t} + \quad + \alpha_{32}y_{3t} + ... + \alpha_{m2}y_{mt} + \beta_{12}x_{1t} + \beta_{22}x_{2t} + ... + \beta_{k2}x_{kt} + u_{2t} \\
& \vdots \\
y_{mt} = \quad & \alpha_{1m}y_{1t} + \alpha_{2m}y_{2t} + ... + \alpha_{m-1,m}y_{m-1t} + \beta_{1m}x_{1t} + \beta_{2m}x_{2t} + ... + \beta_{km}x_{km} + u_{mt}
\end{aligned} \tag{1.1}$$

donde $y_1, y_2, ..., y_m$ son m variables endógenas

$x_1, x_2, ..., x_k$ son k variables predeterminadas (una de estas variables puede tomar un valor unitario para incorporar el término independiente de cada ecuación)

$u_1, u_2, ..., u_m$ son m perturbaciones aleatorias

$t = 1, 2, ..., n$ número total de observaciones

α los coeficientes de las variables endógenas

β los coeficientes de las variables predeterminadas

La variable situada a la izquierda de la igualdad se denomina endógena principal. Cuando la endógena principal está expresada en cada ecuación respecto al resto de variables endógenas y exógenas, decimos que el modelo está expresado en su **forma estructural** y los coeficientes α y β se conocen como parámetros estructurales.

Equivalentemente, el sistema (1.1) puede escribirse en forma matricial de la siguiente manera:

$$Y = YA + XB + U \tag{1.2}$$

donde

$$Y_{nxm} = \begin{pmatrix} y_{11} & y_{21} & ... & y_{m1} \\ y_{12} & y_{22} & ... & y_{m2} \\ & & \vdots & \\ y_{1n} & y_{2n} & ... & y_{mn} \end{pmatrix} = \begin{pmatrix} y_1 & y_2 & ... & y_m \end{pmatrix}, A_{mxm} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & ... & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & ... & \alpha_{2m} \\ & & \vdots & \\ \alpha_{m1} & \alpha_{m2} & ... & \alpha_{mm} \end{pmatrix}$$

$$X_{nxk} = \begin{pmatrix} x_{11} & x_{21} & ... & x_{k1} \\ x_{12} & x_{22} & ... & x_{k2} \\ & & \vdots & \\ x_{1n} & x_{2n} & ... & x_{kn} \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & ... & x_k \end{pmatrix}, B_{kxm} = \begin{pmatrix} \beta_{11} & \beta_{12} & ... & \beta_{1m} \\ \beta_{21} & \beta_{22} & ... & \beta_{2m} \\ & & \vdots & \\ \beta_{k1} & \beta_{k2} & ... & \beta_{km} \end{pmatrix}$$

$$U_{nxm} = \begin{pmatrix} u_{11} & u_{21} & ... & u_{m1} \\ u_{12} & u_{22} & ... & u_{m2} \\ & & \vdots & \\ u_{1n} & u_{2n} & ... & u_{mn} \end{pmatrix} = \begin{pmatrix} u_1 & u_2 & ... & u_m \end{pmatrix}$$

Obsérvese que por convenio para esta expresión en la que las endógenas principales están despejadas $\alpha_{ii} = 0, \ \forall i = 1, ..., m$.

A partir de las ecuaciones estructurales (1.1) se puede transformar el modelo de ecuaciones simultáneas para obtener otra expresión del mismo, la **forma reducida**. En la forma reducida, las variables endógenas se expresan únicamente respecto a las variables predeterminadas del sistema, esto es:

$$\begin{aligned} y_{1t} &= \pi_{11}x_{1t} + \cdots + \pi_{k1}x_{kt} + v_{1t} \\ y_{2t} &= \pi_{12}x_{1t} + \cdots + \pi_{k2}x_{kt} + v_{2t} \\ &\vdots \\ y_{mt} &= \pi_{1m}x_{1t} + \cdots + \pi_{km}x_{kt} + v_{tm} \end{aligned} \tag{1.3}$$

donde π_{ij} con $i = 1, ..., k, j = 1, ..., m$ son los parámetros de la forma reducida
$v_1, v_2, ..., v_m$ son m perturbaciones aleatorias de la forma reducida

En forma matricial:

$$Y = X\Pi + V \tag{1.4}$$

donde

$$\Pi_{kxm} = B(I-A)^{-1} = \begin{pmatrix} \pi_{11} & \pi_{12} & \dots & \pi_{1m} \\ \pi_{21} & \pi_{22} & \dots & \pi_{2m} \\ & & \vdots & \\ \pi_{k1} & \pi_{k2} & \dots & \pi_{km} \end{pmatrix}$$

$$V_{nxm} = U(I-A)^{-1} = \begin{pmatrix} v_{11} & v_{21} & \dots & v_{m1} \\ v_{12} & v_{22} & \dots & v_{m2} \\ & & \vdots & \\ v_{1n} & v_{2n} & \dots & v_{mn} \end{pmatrix}$$

EJEMPLO. Modelo Keynesiano

Función de consumo: $C_t = \beta_0 + \beta_1 Y_t + u_t \qquad 0 < \beta_1 < 1$ (1.5)

Identidad de ingreso: $Y_t = C_t + I_t$ (1.6)

En este modelo, tenemos que:

- C_t (consumo) e Y_t (ingreso) son las variables endógenas.
- I_t (gasto de inversión) y 1_t (la constante) son las variables predeterminadas.

Siguiendo la notación anterior,

$$Y = \begin{pmatrix} C_t & Y_t \end{pmatrix} \quad X = \begin{pmatrix} 1_t & I_t \end{pmatrix} \quad U = \begin{pmatrix} u_t & \mathbf{0} \end{pmatrix} \quad A = \begin{pmatrix} 0 & \beta_1 \\ 1 & 0 \end{pmatrix} \quad B = \begin{pmatrix} \beta_0 & 0 \\ 0 & 1 \end{pmatrix}$$

Y la forma reducida del modelo sería:

$$Y_t = \quad \pi_0 + \pi_1 I_t + v_{1t}$$

$$C_t = \quad \pi_2 + \pi_3 I_t + v_{2t}$$

siendo los coeficientes y perturbaciones de la forma reducida:

$$\Pi = B(I-A)^{-1} = \begin{pmatrix} \pi_0 & \pi_2 \\ \pi_1 & \pi_3 \end{pmatrix} = \begin{pmatrix} \dfrac{\beta_0}{1-\beta_1} & \dfrac{\beta_0\beta_1}{1-\beta_1} \\ \dfrac{1}{1-\beta_1} & \dfrac{1}{1-\beta_1} \end{pmatrix}$$

$$V = U(I-A)^{-1} = \begin{pmatrix} v_1 & v_2 \end{pmatrix} = \begin{pmatrix} \dfrac{u_t}{1-\beta_1} & \dfrac{\beta_1 u_t}{1-\beta_1} \end{pmatrix}$$

Obsérvese que estos coeficientes son combinaciones no lineales de los coeficientes estructurales.

1.2. El problema de identificación

Una vez planteado un modelo de ecuaciones simultáneas el objetivo es estimar los parámetros estructurales α_{11}, ..., α_{mm}, β_{11}, ..., β_{km}. Sin embargo, en los modelos de ecuaciones simultáneas hemos de tener en cuenta que puede que no todas las ecuaciones del sistema se puedan resolver. Este problema se conoce como el problema de la identificación y precede al problema de estimación.

El problema de identificación pretende establecer si las estimaciones numéricas de los coeficientes de una ecuación estructural (1.1) pueden ser obtenidas a partir de los coeficientes estimados de la forma reducida (1.3). Si esto puede hacerse para una ecuación, decimos que dicha ecuación está identificada, en caso contrario la ecuación considerada está subidentificada o no identificada.

El problema surge cuando una ecuación dada en forma reducida puede ser compatible con diferentes ecuaciones estructurales, es decir, con diferentes modelos y puede ser difícil saber con cuál se está trabajando. En otras palabras, el problema de la identificación aparece porque un mismo conjunto de información puede ser compatible con diferentes conjuntos de coeficientes estructurales, es decir, diferentes modelos. Por ello, distinguimos tres situaciones y decimos que una ecuación puede estar:

- **Subidentificada**, si no se puede obtener valores numéricos para los parámetros estructurales o equivalentemente, si el número de parámetros de la forma reducida es *menor* que el número de parámetros de la forma estructural.

- **Exactamente identificada**, si pueden obtenerse valores numéricos únicos de los parámetros estructurales, o equivalentemente, si el número de parámetros de la forma reducida es *igual* que el número de parámetros de la forma estructural.

- **Sobreidentificada**, si puede obtenerse más de un valor numérico para algunos de los parámetros estructurales, o equivalentemente, si el número de parámetros de la forma reducida es *mayor* que el número de parámetros de la forma estructural.

Si la ecuación está exactamente identificada o sobreidentificada, diremos que la ecuación está identificada y un modelo está identificado si cada una de sus ecuaciones lo está. Nótese además la diferencia entre subdidentificación y sobreidentificación. En el primer caso, es imposible obtener estimaciones de los parámetros estructurales y la ecuación no se puede resolver, mientras que en el último, puede haber diversas estimaciones de uno o más coeficientes estructurales y la solución no es única.

Resumiendo, para poder calcular la matriz Π en (1.3), en primer lugar hemos de tener en cuenta que el número de ecuaciones planteadas $(n\ x\ m)$ sea igual o mayor que el número de parámetros a estimar, que en este caso serían los elementos de la matriz Π $(k\ x\ m)$, con lo que se deduce que para poder obtener la expresión reducida de un MES necesariamente $k \leq n$.

Para poder calcular las matrices A y B del modelo estructural en (1.1) a partir de las estimaciones de los coeficientes de la forma reducida (1.3), necesitamos que el número de ecuaciones, es decir, ahora el número de coeficientes de la forma reducida $(k\ x\ m)$ sea igual o mayor que el número de incógnitas, que en este caso son los coeficientes estructurales no nulos, como máximo $m\ x\ (m-1)$ para la matriz A y $k\ x\ m$ para la matriz B.

1.2.1. Reglas para la identificación

Para determinar si una ecuación en un sistema de ecuaciones simultáneas está identificada o no, puede pasarse el sistema a forma reducida y tratar de recuperar los parámetros estructurales en función de los parámetros de la forma reducida. Sin embargo, este procedimiento puede llegar a ser muy laborioso. Para estudiar la identificación de una ecuación, en la práctica se utilizan dos condiciones que ofrecen una rutina sistemática: la **condición de orden** y la **condición de rango**.

La condición de rango establece si la ecuación bajo consideración está identificada o no, mientras que la condición de orden determina si dicha ecuación está exactamente identificada o sobreidentificada. La condición de orden es una condición necesaria pero no suficiente para la identificación, mientras que la condición de rango es necesaria y suficiente.

Para introducir las condiciones de orden y rango adoptamos la siguiente notación:

$m =$ número de variables endógenas en el modelo.
$m_i =$ número de variables endógenas en la ecuación i.
$k =$ número de variables predeterminadas en el modelo.
$k_i =$ número de variables predeterminadas en la ecuación i.

Condición de orden

En un modelo de m ecuaciones simultáneas, una ecuación podrá estimarse si el número de igualdades que aportan los coeficientes de la forma reducida (k) es mayor o igual al número de incógnitas en dicha ecuación ($m_i + k_i - 1$). En otras palabras, para que la ecuación i-ésima esté identificada, el número de variables predeterminadas excluidas en esa ecuación no debe ser menor que el número de variables endógenas incluidas en la ecuación menos una, es decir, $k - k_i \geq m_i - 1$.

$k - k_i > m_i - 1 \Rightarrow$ la ecuación está sobreidentificada.
$k - k_i = m_i - 1 \Rightarrow$ la ecuación está exactamente identificada.

Equivalentemente, esta condición establece que para que una ecuación esté identificada debe excluir *al menos* $m-1$ variables (endógenas y predeterminadas) de las que aparecen en el modelo. Si exlcuye exactamente $m-1$ variables, la ecuación está exactamente identificada. Si exlcuye más de $m-1$ variables, la ecuación está sobreidentificada.

Condición de rango

En un modelo que contiene m ecuaciones con m variables endógenas, la ecuación i-ésima está identificada, si y sólo si, puede construirse al menos un determinante de orden $(m-1)\ x\ (m-1)$ distinto de cero, a partir de los coeficientes de las variables endógenas y predeterminadas excluidas de esa ecuación particular, pero incluidas en las otras ecuaciones del modelo.

Resumiendo, el estudio de las condiciones de orden y de rango para la identificación conduce a las siguientes conclusiones:

Principios generales de identificabilidad de una ecuación estructural en un sistema de m ecuaciones simultáneas

1. $k - k_i > m_i - 1$ y $rango(A) = m - 1 \Rightarrow$ ecuación i-ésima está sobreidentificada.
2. $k - k_i = m_i - 1$ y $rango(A) = m - 1 \Rightarrow$ ecuación i-ésima está exactamente identificada.
3. $k - k_i > m_i - 1$ y $rango(A) < m - 1 \Rightarrow$ ecuación i-ésima está subidentificada.
4. $k - k_i < m_i - 1$ la ecuación estructural i-ésima no está identificada y en este caso, $rango(A) < m - 1$.

1.3. Estimación por Mínimos Cuadrados Ordinarios (MCO) de los modelos de ecuaciones simultáneas

1.3.1. Inconsistencia del estimador MCO

Como es bien sabido, en ausencia de ecuaciones simultáneas, o presencia del problema de simultaneidad, los estimadores obtenidos por MCO son consistentes y eficientes. Sin embargo, la característica única de los modelos de ecuaciones simultáneas de que la variable dependiente en una ecuación del sistema pueda aparecer como variable explicativa en otra, hace que tal variable explicativa endógena esté correlacionada con el término de error de la ecuación en la cual aparece como variable explicativa. Como consecuencia, el método de mínimos cuadrados ordinarios es, en general, inaplicable porque los estimadores así obtenidos no solamente son sesgados, sino que también son inconsistentes, es decir, no convergen hacia sus verdaderos valores poblacionales [4].

EJEMPLO. Supongamos que se desa estimar la función de consumo del modelo Keynesiano. Bajo los supuestos clásicos de regresión:

$$E(u_t) = 0 \tag{1.7}$$

$$E(u_t^2) = \sigma^2 > 0 \tag{1.8}$$

$$E(u_t, u_{t+j}) = 0 \quad \forall\, j \neq 0 \tag{1.9}$$

$$cov(I_t, u_t) = 0 \tag{1.10}$$

Demostraremos en dos pasos que los estimadores de los parámetros de la función consumo son inconsistentes.

Paso 1 Y_t y u_t están correlacionadas
Sustituimos (1.5) en (1.6),

$$Y_t = \beta_0 + \beta_1 Y_t + u_t + I_t \quad \Rightarrow \quad Y_t = \frac{\beta_0}{1-\beta_1} + \frac{1}{1-\beta_1} I_t + \frac{1}{1-\beta_1} u_t \tag{1.11}$$

Aplicando las propiedades de la esperanza, el supuesto (1.7) y teniendo en cuenta que I_t es una variable exógena, es decir, fija con anterioridad

$$E(Y_t) = \frac{\beta_0}{1-\beta_1} + \frac{1}{1-\beta_1} I_t \tag{1.12}$$

Restando (1.11) y (1.12) se tiene,

$$Y_t - E(Y_t) = \frac{1}{1-\beta_1} u_t \tag{1.13}$$

Por tanto, de (1.13) y (1.7) y utilizando (1.8), resulta

$$cov(Y_t, u_t) = E[(Y_t - E[Y_t])(u_t - E[u_t])] = \frac{E(u_t^2)}{1-\beta_1} = \frac{\sigma^2}{1-\beta_1} \neq 0 \tag{1.14}$$

$\Rightarrow Y_t$ y u_t están correlacionadas, es decir, las perturbaciones están correlacionas con las variables explicativas y se viola el supuesto clásico de regresión.

Paso 2 $\hat{\beta}_1$ es un estimador inconsistente de β_1
Por definición, el estimador MCO de β_1 es

$$\hat{\beta}_1 = \frac{\sum (C_t - \bar{C})(Y_t - \bar{Y})}{\sum (Y_t - \bar{Y})^2} = \frac{\sum c_t y_t}{\sum y_t^2} = \frac{\sum C_t y_t}{\sum y_t^2} \tag{1.15}$$

donde las letras minísculas denotan las desviaciones de las observaciones a la media. Susituyendo C_t de (1.5), se obtiene

$$\hat{\beta}_1 = \frac{\sum (\beta_0 + \beta_1 Y_t + u_t) y_t}{\sum y_t^2} = \beta_1 + \frac{\sum y_t u_t}{\sum y_t^2} \tag{1.16}$$

Tomando el valor esperado a ambos lados de (1.16) se tiene

$$E(\hat{\beta}_1) = \beta_1 + E\left[\frac{\sum y_t u_t}{\sum y_t^2}\right] \tag{1.17}$$

de donde se deduce que a menos que el término $\sum y_t u_t / \sum y_t^2$ sea cero, $\hat{\beta}_1$ es un estimador sesgado de β_1.

Calculando el límite de $\hat{\beta}_1$ y aplicando las reglas básicas de operaciones con límites, resulta que

$$\begin{aligned} \text{plim}(\hat{\beta}_1) &= \text{plim}(\beta_1) + \text{plim}\left(\frac{\sum y_t u_t}{\sum y_t^2}\right) \\ &= \text{plim}(\beta_1) + \text{plim}\left(\frac{\sum y_t u_t / n}{\sum y_t^2 / n}\right) \\ &= \beta_1 + \text{plim}\left(\frac{\sum y_t u_t / n}{\sum y_t^2 / n}\right) \end{aligned} \tag{1.18}$$

Observando que en el segundo sumando el numerador se corresponde con la covarianza muestral de Y y u y el denominador con la varianza muestral de Y y teniendo en cuenta que al tomar límites, la covarianza muestral entre Y y u tiende a la covarianza poblacional, $cov(Y_t, u_t)$; la varianza muestral de Y se aproxima a su varianza poblacional, σ_Y y la expresión (1.14), entonces

$$\text{plim}(\hat{\beta}_1) = \beta_1 + \frac{\sigma^2/(1-\beta_1)}{\sigma_Y^2} = \beta_1 + \frac{1}{1-\beta_1}\left(\frac{\sigma^2}{\sigma_Y^2}\right) \tag{1.19}$$

Dado que $0 < \beta_1 < 1$ y que σ^2 y σ_Y^2 son positivas, se tiene que $\text{plim}(\hat{\beta}_1)$ será siempre mayor que β_1, es decir, $\hat{\beta}_1$ sobreestima siempre el verdadero valor de β_1 y por tanto, $\hat{\beta}_1$ es un estimador sesgado de β_1.

1.3.2. Pruebas de simultaneidad y de exogeneidad

Como hemos visto, en presencia de simultaneidad, los estimadores MCO no son siquiera consisitentes y hemos de considerar técnicas de estimación alternativas que proporcionen estimadores consistentes y eficientes como por ejemplo, los métodos de mínimos cuadrados en dos etapas (MC2E) y de variables instrumentales. Paradójicamente, si estas técnicas alternativas se aplican cuando no hay simultaneidad, las estimaciones así obtenidas son consistentes pero no eficientes. Por consiguiente, antes de descartar MCO en favor de otros métodos se debe verificar el problema de simultaneidad.

Prueba de simultaneidad
El problema de simultaneidad surge cuando algunas de las variables explicativas en una ecuación son endógenas y es probable que estén correlacionadas con el término de error. En una prueba de simultaneidad, el objetivo consiste precisamente en averiguar si una variable explicativa está correlacionada o no con el término de error. Si lo está, existe el problema de simultaneidad y deben utilizarse alternativas a MCO; si no lo está, puede aplicarse MCO. Para llevar a cabo esta comprobación, comúnmente se utiliza la **prueba de especificación de Hausman** [9].

Prueba de exogeneidad
Adicionalmente, la prueba de especificación de Hausman puede utilizarse como prueba estadística de exogeneidad para determinar si una variable o grupo de variables son endógenas o exógenas. No obstante, en general, esta decisión depende de cada problema y de la información *a priori* de la que se disponga y, en la práctica, habitualmente corresponde al criterio del investigador establecer qué variables del modelo son endógenas y cuáles exógenas.

1.4. Métodos de estimación

En un modelo de ecuaciones simultáneas no podemos en general estimar los parámetros de las ecuaciones del modelo por MCO y debemos considerar otras técnicas alternativas de estimación. En este apartado, se enuncia un compendio de los métodos de estimación más utilizados en los MES.

Suponiendo que una ecuación en un modelo de ecuaciones simultáneas está identificada (en forma exacta o sobreidentificada), se dispone de diversos métodos para estimarla. Estos métodos con propiedades estadísticas diversas tradicionalmente se clasifican en dos grupos:

- **Métodos de información limitada (o métodos uniecuacionales**)
  Estiman cada ecuación del sistema individualmente, considerando las restricciones impuestas sólo sobre esa ecuación y una información global del resto de ecuaciones del sistema para fines de estimación.
  - Mínimos Cuadrados Ordinarios (MCO) (sistemas triangulares)

  - Mínimos Cuadrados Indirectos (MCI)
  - Mínimos Cuadrados en Dos Etapas (MC2E)
  - Máxima Verosimilutd con Información Limitada (MVIL)

- **Métodos de información completa (o métodos de sistemas**)
  Estiman todas las ecuaciones del sistema de forma conjunta, teniendo en cuenta la influencia de unas ecuaciones en otras.

  - Mínimos Cuadrados en Tres Etapas (MC3E)
  - Máxima Verosimilitud con Información Completa (MVIC)

Para preservar el espíritu de los modelos de ecuaciones simultáneas, idealmente debería utilizarse los métodos de sistemas. Sin embargo, por razones de economía computacional, errores de especificación, etc. los métodos uniecuacionales son en la práctica los más utilizados.

1.5. Justificación y objetivos de un nuevo modelo

En los modelos de ecuaciones simultáneas se asume que los términos de error son serialmente independientes. Sin embargo, en determinandas situaciones, los datos se presentan jearquizados o agrupados (multinivel) y este supuesto no se cumple dando lugar a estimaciones sesgadas. El objetivo central de la presente tesis doctoral es el estudio teórico y práctico de los modelos de ecuaciones simultáneas multinivel que amplíen el marco de formalización de los modelos de ecuaciones simultáneas tradicionales.

Aunque el modelo de ecuaciones simultáneas multinivel puede parecer una extensión lógica de los MES, la literatura que trata esta problemática es escasa. Básicamente, las consideraciones que se han hecho hasta el momento se abordan desde los modelos multinivel [6], en los que en presencia del problema de simultaneidad entre variables, se propone construir un modelo recursivo o triangular [3]. Otros ejemplos que también adoptan esta solución para superar las dificultades metodológicas causadas por la endogeniedad de algunas de las variables en un modelo multinivel pueden encontrarse en Ciencias de la Educación cuando se pretende estudiar el efecto del tamaño de las clases y el gasto en educación sobre los resultados académicos obtenidos por el alumnado [20, 21]. De nuevo, en ambos trabajos los modelos empleados, aunque algo más complejos, son recursivos. Por tanto, aunque las aplicaciones de esta extensión parecen evidentes, se encuentran a la espera de que se profundice en el desarrollo teórico.

No obstante, en los modelos de ecuaciones simultáneas, cuando los términos de error no son serialmente independientes, frecuentemente el procedimiento habitual empleado consiste en estimar el modelo ignorando que se incumple este supuesto. La presente tesis se ocupa de analizar y plantear dos enfoques distintos desde los que abordar este problema: i) modificando la estimación de los MES para su mejora (mediante correcciones en los estimadores, utilizando técnicas de análisis de datos como clustering...) o ii) modificando este supuesto en los MES.

Así, por una parte, en esta tesis se ha planteado una estrategia en dos etapas para mejorar las estimaciones en los modelos de ecuaciones simultáneas en presencia de heterogeneidad. Este procedimiento combina un algoritmo variacional con el esquema de modelos de ecuaciones simultáneas multigrupo. La principal ventaja del método propuesto es que no necesita ser reejecutado para determinar el número de grupos en los que se agrupan los datos. Los estudios de simulación subrayan la fiabilidad del algoritmo para la identificación y clasificación de los diferentes clusters. Asimismo, el criterio de información de Akaike refuerza el uso de este método frente a otros modelos estimados. Por tanto, los objetivos para esta primera parte preliminar de la tesis son los siguientes:

- Aplicar en una primera etapa un algoritmo variacional bayesiano para agrupar observaciones en el conjunto de variables endógenas de un MES.
- Estimar en una segunda etapa los parámetros en cada uno de los grupos obtenidos en la etapa 1.
- Estudiar la bondad de ajuste del procedimiento en dos etapas.
- Aplicar el método a un problema macroeconómico con datos reales.

Por otra parte, el segundo enfoque desde el que abordar la problemática constituye el eje central de la tesis y se basa en la introducción de una doble estructura para la matriz de varianzas en los MES en los cuales se viola el supuesto de ausencia de correlación intertemporal de los términos de error. Los objetivos para esta línea de investigación son los siguientes:

- Plantear teóricamente un nuevo modelo estadístico: los Modelos de Ecuaciones Simultáneas Multinivel (MESM).
- Plantear el sistema de ecuaciones de máxima verosimilitud para la estimación de los parámetros del modelo y estimación teóricamente de los mismos.
- Estudiar mediante simulación la estimación de los parámetros de un MESM tanto bajo el supuesto de conocer las matrices de varianza y covarianzas como en el caso de que sean desconocidas.
- Analizar los resultados obtenidos y compararlos con otros métodos de estimación tradicionalmente utilizados en los Modelos de Ecuaciones Simultáneas como MC2E.

Como se ha indicado en los anteriores objetivos, esta propuesta ha dado lugar al desarrollo de un nuevo modelo al que hemos denominado Modelos de Ecuaciones Simultáneas Multinivel (MESM) para el que se ha planteado la estimación teórica mediante el método de máxima verosimilitud. Sin embargo, ante la imposibilidad de resolver de manera analítica el sistema de ecuaciones obtenido mediante una solución cerrada, la estimación del MESM se ha dividido en dos fases de trabajo. La primera se ha llevado a

cabo utilizando un *solver* de optimización incluido en el software estadístico R bajo el supuesto de que las matrices de varianzas-covarianzas sean conocidas. En una primera aproximación, elegir las estimaciones proporcionadas por MC2E para las matrices de coeficientes y usarlas como puntos semilla para inicializar el *solver* ha demostrado empíricamente que las soluciones del modelo propuesto, MESM, están más cerca de los valores reales de los parámetros que aquellas que se calculan tradicionalmente ignorando la dependencia serial.

No obstante, en el caso general (cuando las matrices de varianza-covarianzas no son conocidas), aunque inicialmente la idea era preservar la misma línea de trabajo, usar solo un *solver* de optimización no es suficiente para obtener una solución aproximada de los estimadores de máxima verosimilitud. El espacio de parámetros aumenta su complejidad y es necesario recurrir a un procedimiento heurístico. Existen muchas técnicas diferentes de este tipo para calcular los parámetros (algoritmos genéticos, scatter search, GRASP,....).

En la segunda fase de trabajo se propone una heurística híbrida que combina un algoritmo genético estándar y un *solver* de optimización (el mismo utilizado en el caso de varianzas conocidas). Para completar el proyecto queda por estudiar el comportamiento del algoritmo híbrido en distintos contextos de MESMs, así como analizar y comparar los resultados frente a otras alternativas. Además, es importante garantizar que la solución obtenida no corresponde a un óptimo local y para ello se requiere explorar todo el espacio de parámetros. En nuestro problema, la complejidad del espacio de soluciones hace que sea imposible su completa exploración. Por esta razón, se considera un algoritmo genético que genera aleatoriamente una población inicial de cromosomas a lo largo de todo el espacio de soluciones que representan posibles candidatos al estimador de máxima verosimilitud. Esta idea junto la probabilidad de mutación incluida en el genético permite trabajar con un subconjunto diverso y manejable del espacio de soluciones.

Finalmente, se concluye el capítulo exponiendo la relación existente entre los distintos trabajos que integran la presente tesis.

1.6. Relación entre los distintos artículos constitutivos de la tesis

El primer trabajo [12] propone una estrategia en dos etapas para mejorar la estimación de los modelos de ecuaciones simultáneas teniendo en cuenta la heterogeneidad en las variables endógenas. En la primera etapa, mediante la aplicación de un algoritmo variacional incremental basado en el concepto de entropía se lleva a cabo una partición de la muestra en grupos para, a continuación, en la segunda etapa, estimar un MES en cada uno de los grupos obtenidos. A través de un estudio de simulación, se analiza empíricamente el porcentaje de acierto del algoritmo tanto en la detección del número de agrupaciones presentes en los datos como en el porcentaje de elementos bien clasificados. Asimismo, se compara la bondad de ajuste de la metodología frente a

otras técnicas alternativas y finalmente, se aplica a un problema macroeconómico con datos reales.

Otra opción para superar las limitaciones de los modelos de ecuaciones simultáneas cuando los términos de error no están intertemporalmente no correlacionados consiste en modificar los propios supuestos del modelo. El planteamiento de un nuevo marco teórico denominado Modelo de Ecuaciones Simultáneas Multinivel (MESM) que extienda el uso de los modelos de ecuaciones simultáneas cuando se viola este supuesto, así como la estimación teórica de los parámetros del nuevo modelo mediante el método de máxima verosimilitud es el objetivo de la segunda contribución [11]. A partir del sistema de ecuaciones de verosimilitud se aborda la estimación de los MESM bajo el supuesto de matrices de varianzas-covarianzas conocidas. También, se comparan los resultados obtenidos con los que proporcionan otros métodos de estimación tradicionalmente utilizados en los MES.

Finalmente, la estimación de los parámetros del modelo en el caso general, es decir, cuando se desconocen las matrices de varianzas-covarianzas es el tema que trata la tercera aportación. Para ello, en [10] se desarrolla una metaheurística híbrida, se estudia el esquema parametrizado del algortimo para seleccionar la combinación de valores óptimos en el contexto de los MESM y se comparan los resultados obtenidos en distintos escenarios mediante simulación.

Por consiguiente, los tres trabajos [12], [11] y [10] plantean distintas aproximaciones para resolver el problema de estimación en los modelos de ecuaciones simultáneas cuando el supuesto de errores incorrelados falla.

Capítulo 2

Resumen de los artículos

2.1. Estimación de modelos de ecuaciones simultáneas mediante un algoritmo de aprendizaje variacional bayesiano

La presencia de heterogeneidad en los modelos de ecuaciones simultáneas es frecuentemente problemática en las aplicaciones reales de estos modelos. Tradicionalmente, se asume que las observaciones son homogéneas y se estima un único conjunto de parámetros [17]. Sin embargo, en determinadas situaciones este supuesto no siempre se cumple y como advierten diversos autores, esta simplificación implica potencialmente un sesgo en las estimaciones [13]. Este trabajo se centra en mejorar la estimación de los MES en los que las observaciones endógenas tienden a formar agrupaciones. Para ello, se desarrolla una estrategia en dos etapas que primero identifica y forma grupos entre las observaciones endógenas mediante un algoritmo de aprendizaje variacional bayesiano y a continuación, aplica el esquema de modelos de ecuaciones simultáneas multigrupo estimando un MES para cada uno de los grupos obtenidos. Finalmente, se evalúa y compara la eficacia de la metodología propuesta frente a otras alternativas mediante un estudio computacional. Adicionalmente, el procediento desarrollado se aplica a modo ilustrativo a un problema macroeconómico que puede consultarse en la propia publicación.

2.1.1. Descripción del algoritmo de aprendizaje variacional bayesiano

Para afrontar el problema de heterogeneidad en el contexto de los modelos de ecuaciones simultáneas, esta publicación propone utilizar una estrategia en dos etapas. En la primera de ellas, el procedimiento selecciona el número óptimo de grupos y clasifica las observaciones a partir de un algoritmo variacional bayesiano incremental basado en el concepto de entropía (EBIVB, por sus siglas en inglés Entropy-Based Incremental Variational Bayes algorithm). La principal aportación de esta publicación es que tanto la identificación del número óptimo de grupos como la clasificación de observaciones

se basa en el concepto de la deficiencia Gaussiana (GD) y el algoritmo no necesita ser reejecutado variando el número de grupos para determinar cuál es el óptimo que mejor ajusta los datos.

Comúnmente, los procedimientos secuenciales utilizados en la literatura incluyen *k-means* como algoritmo de clustering en la primera etapa, el cual es un algoritmo no jerárquico basado en la distancia. A diferencia de estas propuestas, en nuestro trabajo usamos un algoritmo de clustering jerárquico basado en entropía. Las ventajas más importantes de esta propuesta son dos. En primer lugar, no es necesario fijar el número de grupos *a priori*. En segundo lugar, el uso de la entropía como medida de similaridad en la fase de clustering evita el cálculo de distancias reduciendo así el efecto de outliers en la formación de los grupos que merman la calidad de los resultados.

En particular, el algoritmo utilizado es un modelo de mixtura gaussiano que extiende el método variacional bayesiano introducido en [18]. El algoritmo EBIVB es un procedimiento iterativo que empieza con un solo núcleo ($K = 1$) incialmente obtenido a partir de la muestra y que en cada iteración añade un nuevo grupo a la mixtura dividiendo el actual. El objetivo de esta estrategia es optimizar los coeficientes de la mixtura π maximizando la verosimilitud marginal de los datos $P(X|\pi)$. Para ello, se introduce una distribución $\mathscr{Q}(\Theta)$ en el logaritmo de la función de verosimilitud marginal que permite encontrar una cota inferior de $P(X,\mu,T,z|\pi)$.

$$\begin{aligned}\ln P(X|\pi) &= \ln \sum_z \int P(X,\Theta|\pi)d\Theta = \ln \sum_z \int \mathscr{Q}(\Theta)\frac{P(X,\Theta|\pi)}{\mathscr{Q}(\Theta)}d\Theta \\ &\geq \sum_z \int \mathscr{Q}(\Theta)\ln\frac{P(X,\Theta|\pi)}{\mathscr{Q}(\Theta)}d\Theta = \mathscr{L}(\mathscr{Q})\end{aligned} \tag{2.1}$$

donde $X = \{x_1,\ldots,x_N\}$es el conjunto de observaciones

z son variables binarias latentes tal que $z_{in} = 1 \Leftrightarrow$ la componente i de la mixtura da lugar a la observación x_n y $\sum_{i=1}^{K} z_{in} = 1$ con $i = 1,\ldots,K$, siendo K el número de núcleos en la mixtura

$X|z$ sigue una distribución normal de media μ_i y matriz de covarianza inversa T_i

$\Theta = \{\mu,T,z\}$ para simplificar la notación

La distribución $\mathscr{Q}(\Theta)$ puede factorizarse como $\mathscr{Q}(\Theta) = \mathscr{Q}_z(z)\mathscr{Q}_\mu(\mu)\mathscr{Q}_T(T)$ y entonces, la cota inferior $\mathscr{L}(\mathscr{Q})$ puede evaluarse como

$$\begin{aligned}\mathscr{L}(Q) &= \langle \ln P(X|\mu,T,z)\rangle + \langle \ln P(z)\rangle + \langle \ln P(\mu)\rangle + \langle \ln P(T)\rangle \\ &\quad -\langle \ln \mathscr{Q}_z(z)\rangle - \langle \ln \mathscr{Q}_\mu(\mu)\rangle - \langle \ln \mathscr{Q}_T(T)\rangle\end{aligned} \tag{2.2}$$

De este modo, la estimación de los coeficientes de la mixtura puede simplificarse maximizando la cota en (2.2) respecto de π. Para llevar a cabo la optimización, en

cada iteración el algoritmo EBIVB selecciona la peor componente de la mixtura en términos de su deficiencia Gaussiana, comparando para cada una de ellas la entropía de su función de probabilidad subyacente repecto a la entropía Gaussiana teórica. Por tanto, para el cálculo de la GD de una componente, necesitamos conocer la entropía de dicha componente y para ello, el algoritmo utiliza el estimador de Leonenko [15] basado en la entropía de Shannon. Una vez que se selecciona la peor componente, esta se sustituye por dos nuevas componentes separadas adecuadamente una de la otra. El procedimiento se repite, ahora con $K+1$ núcleos, hasta que se alcanza la convergencia. Es decir, el proceso de división finaliza cuando añadir una nueva componente no proprociona mejor ajuste de los datos y algunos de los coeficientes de la mixtura convergen a cero.

Algorithm 1: Esquema parametrizado de la estrategia en dos etapas

K=1 (número de grupos)

Inicializar los núcleos a partir de la muestra $\rightarrow X$

Calcular GD de X $\rightarrow GD(X)$

while not EndCondition **do**

Seleccionar el núcleo con la mayor GD $\rightarrow X\_sel$

Dividir X_sel en dos núcleos $\rightarrow K = K+1$

Calcular GD de $X_i,\ i = 1,...,K \rightarrow GD(X_1),...,GD(X_K)$

end while

Estimar un MES para cada núcleo $\rightarrow MES(X_i)\ i = 1,...,K$

2.1.2. Resultados computacionales

Para evaluar el comportamiento de la metodología desarrollada y su rendimiento frente a otras alternativas llevamos a cabo dos experimentos de simulación. Los detalles correspondientes a la generación de datos pueden encontrarse en la propia publicación.

En el primer experimento, comparamos la bondad de ajuste de la estrategia propuesta frente a diferentes modelos estimados. En concreto, comparamos los resultados mediante el criterio de información de Akaike (AIC) del modelo agregado (AGG) que ignora la heterogeneidad de los datos, el modelo de grupos conocidos (GM), un modelo porcentual (PM) que partiendo del modelo de grupos conocidos clasifica deliberadamente un porcentaje $p\%$ de las observaciones en un grupo incorrecto y del modelo secuencial (CA). La Tabla 2.1 muestra el valor AIC en un promedio de 10 simulaciones para cada uno de las metodologías anteriores en modelos de ecuaciones simultáneas con diferente número de variables endógenas m. Asimismo, para analizar el sesgo introducido por los procedimientos secuenciales, se calcula el porcentaje de error cometido por el algoritmo en la clasificación de las observaciones, tomando como referencia el modelo de grupos conocidos.

Tabla 2.1: Valor medio de AIC para diferentes modelos estimados en $s = 10$ simulaciones y tasa de error de clasifiación cometida por el algoritmo de clustering respecto al modelo de grupos conocidos.

tamaño	GM	PM			AGG	CA	CA Error de clustering
m	$p = 0\,\%$	$p = 5\,\%$	$p = 10\,\%$	$p = 15\,\%$	Aggregado		%
2	76003.15	76069.52	76200.78	76601.07	82421.45	75964.97	1.65
4	144038.25	148422.05	150889.29	152474.46	153744.05	144130.05	0.77
6	238407.47	244234.13	247418.12	248074.44	249785.10	238794.03	0.67
8	332163.59	338568.06	343189.90	344504.72	349888.61	333404.96	0.39

Analizando la tabla anterior, se observa que el modelo de grupos conocidos obtiene mejores resultados que el modelo agregado y el modelo porcentual, tal como se preveía. De hecho, el modelo agregado es el que muestra siempre los peores valores de AIC y en el caso del modelo porcentual, se puede ver la evolución y mejora de este criterio desde $p = 15\,\%$ a $p = 0\,\%$ para cada uno de los problemas considerados. Obsérvese que el modelo porcentual con $p = 0\,\%$ se corresponde con el modelo de grupos conocidos. Respecto a la metodología secuencial propuesta, es interesante señalar que la calidad relativa del modelo se encuentra entre la del modelo de grupos conocidos y la del modelo porcentual con $p = 5\,\%$, a excepción del caso $m = 2$. Por tanto, podemos concluir que la precisión en la identificación de grupos y clasificación de observaciones de la estrategia en dos etapas propueta es equivalente a la de un modelo porcentual con $p \in (0, 0.05)$. Asimismo, en la tabla puede comprobarse como el error introducido por el algoritmo EBIVB decrece conforme aumenta el tamaño del problema y en todos los casos, el error en la identificación y clasificación de observaciones cometido por el algoritmo es inferior al $2\,\%$.

Por su parte, el objetivo del segundo experimento es estudiar la eficacia del algoritmo para identificar el número óptimo de clústers en los que agrupar las observaciones y por tanto, para clasificarlas correctamente. La Tabla 2.2 muestra la evolución del criterio de información de Akaike al variar progresivamente el número de grupos en el algoritmo EBIVB, partiendo del modelo agregado ($K = 1$) hasta el modelo estimado cuando el algoritmo alcanza el criterio de parada, es decir, cuando detecta el número óptimo de grupos.

Tabla 2.2: Evolución del valor medio de AIC para diferentes modelos estimados.

tamaño	Número de clústers			
m	1	2	3	4
2	83795.08	78139.16	76307.94	75408.18
4	136575.21	136703.96	135996.17	132493.36
6	255276.91	256088.06	251367.12	249979.87
8	348972.97	351896.28	341170.68	331409.30

Según la Tabla 2.2, en los experimentos simulados el algoritmo siempre alcanza el criterio de parada para $K = 4$. En base a los resultados, en presencia de heterogeneidad la bondad de ajuste proporcionada por la estrategia en dos etapas es mejor que la

del modelo agregado. En la literatura, freceuntemente se utilizan procedimientos que requieren preespecificar el número de grupos al ejecutar el algoritmo. Sin embargo, si esta información se desconoce es conveniente ejecutar el algoritmo variando el número de grupos y posteriormente comparar los resultados utilizando criterios de información u otros estadísticos para la selección del modelo. Este hecho implica a su vez analizar la robustez de los diferentes criterios de información (e.g. AIC, CAIC, BIC) como herramienta de selección del número óptimo de grupos. En nuestro caso, no es necesario el uso de estos criterios para determinar el número óptimo de grupos, pero verificamos que los resultados son consistentes con la evolución del criterio de información de Akaike. En la Tabla 2.2 puede confirmarse como el valor del AIC mejora desde el modelo agregado ($K = 1$) hasta el modelo estimado cuando el número de grupos es el óptimo. En todos las simulaciones, se fijó $K = 4$ como el número de agrupaciones para la generación de los datos y el algoritmo EBIVB eligió el mismo número de grupos en el 100 % de los casos. Además, la tasa de éxito puede comprobarse que es estable cuando el número de variables endógenas aumenta. Para concluir, como observación final, cabe señalar que el algoritmo podría elegir un número óptimo de grupos diferente del correspondiente al del modelo real si la configuración obtenida en el clustering es más precisa o plausible que la del modelo original.

2.2. Modelos de Ecuaciones Simultáneas Multinivel

En los modelos de ecuaciones simultáneas [9] se asume que los términos de error son serialmente independientes

$$u_{t.}^T \sim N(0,\Sigma), \quad E(u_{t.}^T, u_{t'.}) = \delta_{tt'}\Sigma \qquad t,t' = 1,2,...,N \tag{2.3}$$

donde $\delta_{tt'}$ es la delta de Kronecker y Σ una matriz definida positiva.

Sin embargo, este supuesto no siempre es válido cuando los datos presentan estructura jerárquica o agrupada y puede introducir un sesgo en las estimaciones, tal y como hemos podido comprobar en el ejemplo del Modelo de Klein estudiado en el primero de los trabajos que constituyen esta tesis [12].

La principal contribución de esta segunda publicación consiste en considerar una ditrisbución matricial para los términos de error que permita tener en cuenta la correlación temporal en los modelos de ecuaciones simultáneas en este tipo de situaciones. Partiendo de un MES en el que las observaciones están agrupadas en l grupos independientes, introducir la distribución normal matricial [1] da lugar al desarrollo de un nuevo modelo, al que hemos denominado Modelos de Ecuaciones Simultáneas Multinivel (MESM).

$$Y_j = Y_j A + X_j B + E_j \qquad E_j \sim N_{n,m}(0,U,\Sigma) \qquad j = 1,...,l \tag{2.4}$$

con 0, U y Σ matrices de tamaño $n \times m$, $n \times n, m \times m$, respectivamente y U y Σ matrices definidas positivas.

A continuación, se plantea el estimador de máxima verosimilitud (MLE) para los MESM de modo que el logaritmo de la función de verosimilitud viene dado por

$$\begin{aligned} L = &-\frac{nml}{2}ln(2\pi) - \frac{ml}{2}ln|U| - \frac{nl}{2}ln|((I-A)^{-1})^T\Sigma(I-A)^{-1}| \\ &-\frac{1}{2}\sum_{j=1}^{l} tr\left(U^{-1}(Y_j - X_jB(I-A)^{-1})(I-A)\Sigma^{-1}(I-A)^T(Y_j - X_jB(I-A)^{-1})^T\right) \end{aligned} \tag{2.5}$$

Finalmente, aplicando derivadas matriciales se obtiene el sistema de ecuaciones de máxima verosimilitud

$$\begin{aligned} \frac{\partial L}{\partial U} = &-mlU^{-1} + \frac{ml}{2}diag(U^{-1}) + \sum_{j=1}^{l}\left(U^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1}(Y_j(I-A) - X_jB)^TU^{-1}\right) \\ &-\frac{1}{2}\sum_{j=1}^{l} diag\left(U^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1}(Y_j(I-A) - X_jB)^TU^{-1}\right) = 0 \end{aligned} \tag{2.6}$$

$$\begin{aligned} \frac{\partial L}{\partial \Sigma} = &-nl\Sigma^{-1} + \frac{nl}{2}diag(\Sigma^{-1}) + \sum_{j=1}^{l}\left(\Sigma^{-1}(Y_j(I-A) - X_jB)^TU^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1}\right) \\ &-\frac{1}{2}\sum_{j=1}^{l} diag\left(\Sigma^{-1}(Y_j(I-A) - X_jB)^TU^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1}\right) = 0 \end{aligned} \tag{2.7}$$

$$\frac{\partial L}{\partial B} = \sum_{j=1}^{l}(X_j^{-T}U^{-1}Y_j)(I-A)\Sigma^{-1} - \sum_{j=1}^{l}(X_j^TU^{-1}X_j)B\Sigma^{-1} = 0 \tag{2.8}$$

$$\frac{\partial L}{\partial (I-A)} = nl((I-A)^{-1})^T - \sum_{j=1}^{l}(Y_j^TU^{-1}Y_j(I-A)\Sigma^{-1} - Y_j^TU^{-1}X_jB\Sigma^{-1}) = 0 \tag{2.9}$$

Por definición, el estimador de máxima verosimilitud es el máximo global de (2.5). El procedimiento habitual para obtener este estimador consiste en resolver el sistema de ecuaciones de máxima verosimilitud anterior igualando cada derivada a cero. Sin embargo, maximizar esta función es un problema no lineal y en este caso, no puede resolverse analíticamente a través de una fórmula cerrada.

Para obtener el estimador de máxima verosimilitud, en esta publicación se plantea utilizar un *solver* de optimización genérico. Básicamente, el objetivo es obtener una primera aproximación de este estimador. Por ello, en este punto hemos distinguido dos situaciones:

1. Estimación de las matrices de coeficientes A y B cuando las matrices de covarianzas U y Σ son conocidas.

2. Estimación de las matrices de coeficientes A y B cuando las matrices de covarianzas U y Σ son desconocidas.

En esta publicación, se aborda el primer caso y se procede a la búsqueda del estimador de máxima verosimilitud bajo el supuesto de matrices de varianza conocidas. No obstante, puesto que los métodos númericos tienden a ser sensibles a los valores iniciales, se establece $\hat{A}_0 = A_{2SLS}$ y $\hat{B}_0 = B_{2SLS}$, pues aunque MC2E asume observaciones no correlacionadas intertemporalmente, estos valores constituyen en general un punto de partida adecuado. Es decir, cualquier solución obtenida por el *solver* nunca empeorará el valor de la función de verosimilitud que proporcionan las estimaciones dadas por MC2E, método que se usa generalmente asumiendo que en estos casos se comete un sesgo.

Para demostrar la mejora que supone incorporar esta distribución con una doble estructura de varianza en los errores, se diseñan diversos experimentos de simulación. En este resumen comentaremos brevemente la Tabla 2.3. La descripción de los distintos parámetros utilizados, así como el resto de pruebas computacionales pueden consultarse en la propia publicación.

Tabla 2.3: Distancias euclídeas medias $||\hat{A}-A||_{2,s}$ y $||\hat{B}-B||_{2,s}$ entre la estimación $\hat{A}$ y el parámetro A y entre la estimación $\hat{B}$ y el parámetro B, en $s = 10$ simulaciones. Valor medio de la función de fitness y porcentaje de ejecuciones en las que MLE mejora el valor de fitness de MC2E. $U = (u_{ij}) \in [-5, 5]$

tamaño			MC2E		MLE_{nlm}		Fitness		%
m	k	l	$\|\|\hat{A}-A\|\|$	$\|\|\hat{B}-B\|\|$	$\|\|\hat{A}-A\|\|$	$\|\|\hat{B}-B\|\|$	2SLS	MLE_{nlm}	Mejora
2	3	5	$1{,}55_{1,73}$	$1{,}80_{1,92}$	$1{,}50_{1,62}$	$1{,}67_{1,74}$	−737,53	−267,43	100 %
2	3	10	$2{,}22_{2,49}$	$3{,}50_{4,75}$	$1{,}42_{0,94}$	$1{,}91_{2,06}$	−5019,17	−760,49	80 %
2	3	25	$0{,}98_{1,46}$	$1{,}82_{2,84}$	$0{,}94_{1,39}$	$1{,}67_{2,37}$	−1207,75	−296,99	100 %
2	3	50	$1{,}02_{1,42}$	$1{,}29_{2,09}$	$1{,}04_{1,42}$	$1{,}39_{2,09}$	−50338,95	−672,09	90 %
8	12	5	$6{,}41_{1,50}$	$10{,}89_{3,19}$	$6{,}40_{1,46}$	$10{,}82_{3,15}$	−474794	−31080,5	90 %
8	12	10	$8{,}62_{8,29}$	$13{,}03_{12,04}$	$8{,}63_{8,03}$	$12{,}99_{12,02}$	−932483	−435080	100 %
8	12	25	$7{,}68_{3,90}$	$11{,}07_{4,67}$	$7{,}56_{3,79}$	$11{,}08_{4,69}$	−4814615,7	−603524,25	100 %
8	12	50	$4{,}31_{3,24}$	$5{,}70_{3,00}$	$4{,}32_{3,23}$	$5{,}69_{2,99}$	−1182940	−1028962,6	100 %
10	15	5	$9{,}09_{3,95}$	$16{,}60_{8,00}$	$9{,}12_{4,03}$	$16{,}54_{7,96}$	−950988,07	−80876,12	100 %
10	15	10	$6{,}71_{2,36}$	$9{,}20_{2,75}$	$6{,}68_{2,39}$	$9{,}18_{2,75}$	−3281584,5	−755413,76	100 %
10	15	25	$5{,}16_{1,96}$	$7{,}21_{2,73}$	$5{,}17_{1,95}$	$7{,}19_{2,73}$	−684998313	−424518142	100 %
10	15	50	$5{,}03_{2,68}$	$6{,}32_{2,56}$	$5{,}03_{2,69}$	$6{,}31_{2,57}$	−55048498	−19943667	100 %
15	20	5	$15{,}21_{2,35}$	$26{,}65_{7,54}$	$15{,}20_{2,35}$	$26{,}64_{7,54}$	−32944034	−13236945	100 %
15	20	10	$13{,}13_{3,23}$	$20{,}00_{7,75}$	$13{,}13_{3,23}$	$20{,}00_{7,75}$	−3953024	−1599824,4	100 %
15	20	25	$11{,}60_{3,17}$	$15{,}45_{4,77}$	$11{,}60_{3,17}$	$15{,}43_{4,77}$	−12677072	−1599824,4	100 %
15	20	50	$9{,}91_{1,98}$	$12{,}27_{3,44}$	$9{,}91_{1,98}$	$12{,}27_{3,43}$	−1394508,6	−707533,98	100 %

La Tabla 2.3 muestra la dispersión de las estimaciones obtenidas por MC2E y mediante el *solver* de optimización *nlm* respecto a los coeficientes reales de las matrices de parámetros A y B para 10 simulaciones cuando la matriz de covarianzas U (correlación intertemporal) toma sus valores en el intervalo [-5, 5]. Además, se incluye en la tabla el valor medio de la función de verosimilitud obtenido por cada uno de estos métodos y el porcentaje de simulaciones en las que las estimaciones de máxima

verosimilitud obtenidos por el *solver* mejoran en términos de fitness a MC2E.

Basándonos en los resultados computacionales, el porcentaje de simulaciones en las que el estimador de máxima verosimilitud hallado por el *solver* supera estrictamente en fitness a MC2E es prácticamente del 100 % para los distintos tamaños de problemas considerados, a excepción de problemas pequeños para los que este porcentaje se reduce ligeramente. Se observa que la distancia euclídea media de las matrices de coeficientes proporcionada por ambos métodos de estimación es muy similiar aunque en general, el estimador de máxima verosimilitud muestra valores de dispersión más pequeños. En ambos casos, los coeficientes asociados a las variables exógenas son los que alcanzan mayor dispersión. A partir de problemas con m=10 endógenas, cuando el número de grupos l aumenta, esta dispersión se reduce progresivamente. Por tanto, $\hat{A}$ y $\hat{B}$ son estimadores consistentes de A y B.

Todo lo anterior sugiere que introducir una doble estructura en la varianza de los MES cuando se viola el supuesto de términos de error intertemporalmente no correlacionados mejora la precisión de las estimaciones. Además, obsérvese que cuando $U = I$, es decir, no hay correlación serial, el modelo propuesto se reduce a un modelo de ecuaciones simultáneas y seguiría siendo válido. En esta publicación, se demuestra empíricamente que en presencia de datos agrupados, los resultados obtenidos al maximizar la función de verosimilitud del nuevo modelo desarrollado están más cerca de los parámetros reales del modelo que las estimaciones dadas por MC2E que ignoran la dependencia serial de los errores.

2.3. Estimación de modelos de ecuaciones simultáneas multinivel mediante algoritmos genéticos

En esta publicación se aborda la estimación de los modelos de ecuaciones simultáneas multinivel definidos en [11] en el caso general, es decir, cuando las matrices de varianzas-covarianzas asociadas a la distribución de los errores son desconocidas. Para ello, se propone una metaheurística híbrida que consiste en un algoritmo genético optimizado. A partir de un estudio de simulación se determinan los parámetros de ajuste del algoritmo teniendo en cuenta diversos aspectos relativos tanto a las soluciones como al tiempo de ejecución del algoritmo. Finalmente, se aplica la metaheurística a problemas de MESM de diferentes características, se evalúan las estimaciones obtenidas y se comparan los resultados experimentales frente a otros métodos de estimación.

2.3.1. Descripción del algortimo genético para la estimación de MESM

La maximización de la función de verosimilitud de los MESM implica la resolución de un sistema de ecuaciones no lineal. La complejidad del sistema de derivadas y la falta de una solución analítica cerrada conduce al desarrollo de una metaheurística basada en un algoritmo genético optimizado. Los cromosomas del algoritmo, compuestos por las matrices A, B, U y Σ, representan posibles soluciones al problema de

optimización y por tanto, cada cromosoma consituye un candidato para el estimador de máxima verosimilitud. Las distintas funciones y parámetros del algoritmo genético propuesto se describen a continuación:

Inicialización y Condición de término

La población inicial se genera aleatoriamente y su tamaño (*PopSize*) se define al principio. El proceso generacional se repite hasta que se alcanza un número fijado de iteraciones ($Maxiter$).

Evaluación y Selección

La aptitud de cada cromosoma se calcula utilizando la función de verosimilitud (2.5). Una vez evaluados todos los cromosomas se crea un conjunto de referencia (*BenchSet*) compuesto por los cromosomas con mejor aptitud que se actualiza en cada generación.

Cruzamiento

En cada genereación, se selecciona una parte de la población existente de entre los individuos con mejor aptitud para crear una nueva generación (*RepSize*). Las soluciones individuales se emparejan aleatoriamente de forma que un mismo cromosoma puede cruzarse con mútliples cromosomas dando lugar a un conjunto de nuevos individuos (*CrossSize*).

La recombinación opera sobre cada par de cromosomas padres y genera un descendiente que hereda cada una de las matrices A, B, U y Σ íntegramente de uno de los progenitores de forma aleatoria.

Mutación

En cada generación, se considera una probabilidad de mutación (P_{mut}) de los cromosomas que afecta únicamente a los elementos de la diagonal de las matrices de covarianza U y Σ. Si un cromosoma de la nueva generación muta, todos los elementos en la diagonal de una de estas matrices modifiarán su valor numérico.

Mejora de los cromosomas

Después de la mutación, un porcentaje de los cromosomas descendientes, mutados o no, son seleccionados para entrar en un proceso de mejora con probabilidad P_{imp}. Un solver de optimización utiliza los cromosomas como solución inicial para maximizar la función de verosimilitud (2.5). Si el solver encuentra una solución mejor, el cromosoma se actualiza. En caso contrario, el cromosma no se modifica.

Finalmente, una vez concluidas las fases de mutación y mejora, los descendientes pueden formar parte del conjunto de referencia si su aptitud es mejor que cualquiera de la de los individuos en este conjunto, produciéndose un reemplazo generacional.

Optimización

Una vez que el proceso generacional ha concluido, se selecciona un conjunto de entre los mejores cromosomas (*OptSize*) para entrar de nuevo en un último proceso de optimización. Finalmente, la metaheurística devuelve las soluciones correspondientes al mejor individuo.

Algorithm 2: Esquema parametrizado para la metaheurística híbrida

Initialize(ParamIni) $\rightarrow$ *S_ini*

Computefitness(S_ini, ParamIni)

Select(S_ini, ParamSelIni) $\rightarrow$ *S_ref*

while not EndCondition(S_ref, ParamEndCon) **do**

  Select(S_ref, ParamSel) $\rightarrow$ *S_sel*

  Combine(S_sel, ParamCom) $\rightarrow$ *S_com*

  Mutate(S_com, ParamMut) $\rightarrow$ *S_mut*

  Improve(S_com, S_mut, ParamImp) $\rightarrow$ *S_imp_com*, *S_imp_mut*

  Computefitness(S_com,S_mut,S_imp_com, S_imp_mut, ParamCom)

  Include(S_com,S_mut,S_imp_com, S_imp_mut, S_ref, ParamInc) $\rightarrow$ *S_ref*

end while

Improve(S_ref, ParamImpPost) $\rightarrow$ *S_opt*

Select(S_opt, ParamSelBest)

2.3.2. Configuración de los parámetros de ajuste de la metaheurística

En la configuración de los parámetros de la metaheurística distinguimos dos tipos de parámetros. Por una parte, los asociados a un algoritmo genético estándar y por otra, aquellos relacionados con la hibridación introducida en este trabajo. Los parámetros del primer grupo se fijan a valores predefinidos como se indica a continuación. La población inicial, *PopSize*, está integrada por 300 cromosomas y en cada generación, se determina un conjunto de referencia, *BenchSet*, integrado por 100 cromosomas. El número de indiviudos seleccionados para ser cruzados es *RepSize*= 20 y en cada generación, se crean *CrossSize*= 25 nuevos cromosomas con una probabilidad de mutación, P_{mut} del 25 %. El resto de parámetros de la metaheurística se determinan a partir de un estudio experimental cuyos resultados se presentan en la Tabla 2.4.

La Tabla 2.4 muestra para distintos modos de optimización programados, es decir, para distintas configuraciones de parámetros, el valor de fitness que devuelve la metaheurística y el tiempo de ejecución en segundos antes de entrar en la fase de opti-

mización (*Alg.Prev*, *Prev*) y una vez concluida esta etapa (*Alg.End*, *Post*), si se lleva a cabo.

Tabla 2.4: Valor medio de fitness y tiempo medio de ejecución (en 10 simulaciones) para diferentes parámetros de configuración de la metaheurística híbrida.

Mode			**Fitness**		**Time**	
MaxIter	*OptSize*	P_{imp}	**Alg.Prev**	**Alg.End**	**Prev**	**Post**
10	0	0	−16,726.56	−16,726.56	0.32	0.32
50	0	0	−412,005.68	−12,005.68	0.90	0.90
100	0	0	−12,028.44	−12,028.44	1.60	1.60
10	10	0	−17,927.11	−10,568.38	0.32	1751.52
50	10	0	−12,098.04	−10,346.21	0.90	1732.36
100	10	0	−11,753.68	−10,319.02	1.60	1739.39
10	0	10	−9944.31	−9944.31	3663.19	3663.19
50	0	10	−8949.24	−8949.24	17,441.91	17,441.91
100	0	10	−8488.46	−8488.46	33,893.78	33,893.78
10	10	5	−10,364.71	−9992.68	2024.01	3776.39
50	10	5	−8814.78	−8756.06	8983.55	10,519.69
100	10	5	−8953.06	−8838.36	17,051.52	18,454.82

El análisis de la Tabla 2.4 concluye que incluir un solver de optimización tanto después de la mutación ($P_{imp} \neq 0$) como al final del algoritmo genético ($OptSize \neq 0$) mejora en cualquier caso el valor de fitness proporcionado por el algoritmo genético estándar como puede comprobarse en la columna $Alg.End$. Los resultados señalan que insertar la optimización después de la fase de mutación (P_{imp}=10 y $OptSize$=0) tiene mayor impacto que optimizar únicamente los mejores individuos al final del algoritmo (P_{imp}=0 y $OptSize$=10). Además, el estudio de simulación ilustra que no hay una diferencia significativa entre aplicar el solver únicamente después de la mutación y utilizar ambas opciones de optimización (P_{imp}=5 y $OptSize$=10). Sin embargo, esta última configuración permite un margen de mejora de las soluciones asumiendo un coste de tiempo despreciable. Por otro lado, para cualquiera de los modos del algoritmo genético optimizado, los diferentes valores de condición de fin proporcionan una fitness similar pero los tiempos de ejecución se disparan conforme aumenta el número de iteraciones.

A partir de los resultados experimentales se fija la probabilidad de mejora, P_{imp}= 5 % en cada iteración y el número de cromosomas a optimizar de entre los más aptos una vez concluido el algoritmo genético es $OptSize$= 10. Obsérvese que la combinación P_{imp}= 0 y $OptSize$= 0, se corresponde con un algoritmo genético estándar. El proceso generacional se repite hasta que se alcanza un número máximo de iteraciones, $MaxIter$=10. La Figura 2.1 resume el proceso iterativo y los valores seleccionados para los parámetros.

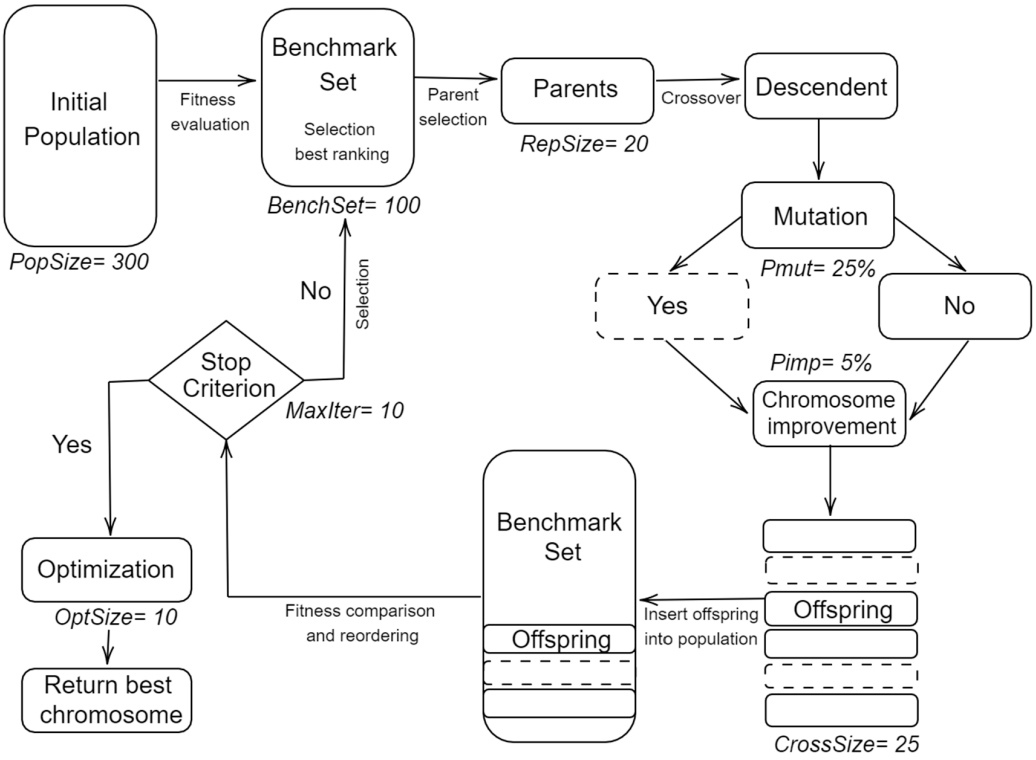

Figura 2.1: Esquema del Algoritmo Genético Híbrido.

2.3.3. Resultados computacionales

Una vez establecidos los parámetros de la metaheurística a partir de los experimentos de simulación anteriores, se consideran MESM de diferentes características y se les aplica el algoritmo. Adicionalmente, se ajusta la matriz de covarianza U introduciendo un parámetro adicional, λ, que divide el intervalo en el que se generan los valores de entrada de esta matriz entre 100, 10, 1, 0.1 o 0.01. Finalmente, las soluciones obtenidas por la metaheurística híbrida se comparan en términos de fitness, tiempo y precisión con otros métodos de estimación como MC2E y un algortimo genético estándar con un número máximo de iteraciones de 10000. La Tabla 2.5 resume todos estos resultados en función de los distintos parámetros para un promedio de 5 pruebas. Los detalles correspondientes a la simulación de los datos de cada problema MESM y la generación de cromosomas en el espacio de soluciones pueden consultarse en la propia publicación.

A partir de los resultados, se concluye que la fitness proporcionada tanto por la metaheurística híbrida (F_{HM}) como por el algoritmo genético estándar (F_{GA}) siempre mejoran la puntuación de fitness dada por MC2E (F_{2SLS}). Además, el valor de fitness de la función de verosimilitud obtenido por la metaheurística híbrida mejora el valor obtenido utilizando únicamente un algoritmo genético. Por tanto, el método propuesto en esta hibridación mejora la fitness sobre cualquiera de los procedimientos alternativos considerados.

Tabla 2.5: Promedio de resultados para fitness, tiempo de ejecución y precisión (en 5 simulaciones) de diferentes métodos de estimación.

Tamaño	λ	F_{HM}	F_{GA}	F_{2SLS}	**S.prev**	**S.post**	S_{GA}	$\|\|Y-\hat{Y}_{HM}\|\|$	$\|\|Y-\hat{Y}_{GA}\|\|$	$\|\|Y-\hat{Y}_{2SLS}\|\|$
	100	−3163.36	−3635.03	−411,983.91	434.99	922.77	57.25	537.39	116.48	89.19
	10	−3811.79	−4440.59	−57,824.25	584.70	1087.31	59.51	559.08	302.24	280.04
8 12 5 30	1	−4094.60	−5748.64	−34,293.66	504.84	976.59	59.57	1123.97	948.76	899.87
	0.1	−7248.25	−10,943.99	−48,767,714.85	522.52	1030.72	59.67	2264.33	2658.97	2596.77
	0.01	−10,069.02	−61,501.99	−881,112.47	507.95	1008.34	59.51	7696.10	8519.65	8368.07
	100	−7341.87	−8668.81	−141,301.20	1894.89	3680.11	142.04	1270.48	408.67	186.20
	10	−8411.78	−10,104.22	−7,104,505.59	1781.75	3547.47	141.76	1464.06	841.10	608.62
15 20 5 30	1	−11,014.00	−12,694.56	−1,283,916.42	2310.64	4140.55	136.27	2563.91	2433.26	2076.20
	0.1	−12,097.40	−22,999.82	−327,573.88	2188.60	3900.75	127.63	5406.90	6179.77	6306.63
	0.01	−47455.24	−170,551.28	−21,252,787.68	1346.84	3189.38	130.23	14,815.85	20,587.63	18,881.66
	100	−14,003.91	−16,217.86	−3,125,727.39	3851.21	8308.45	270.85	3922.10	2720.13	344.21
	10	−14,556.42	−17,178.54	−2,558,133.93	5222.54	9589.12	270.27	2274.98	2417.81	917.13
22 28 5 30	1	−20,391.21	−28,325.12	−6,585,294.85	5454.31	9287.51	271.69	4343.76	5204.45	3823.36
	0.1	−22,702.70	−33,346.31	−18,947,400.33	3153.00	7526.67	270.77	8860.07	10,557.65	9396.87
	0.01	−48,689.37	−327,171.61	−15,177,293.12	5968.90	10,372.43	271.31	31,943.36	42,114.08	36,997.67

Para medir la precisión de las estimaciones se considera la norma de Frobenius. Es interesante señalar que MC2E muestra mejores resultados para valores de U pequeños, pero a medida que el rango de los valores de entrada de la matriz U aumenta, la metaheurística híbrida supera a MC2E para cualquier tamaño del problema MSEM. No obstante, este resultado no se generaliza a un algoritmo genético estándar. De hecho, en este caso MC2E siempre produce mejores estimaciones que el algoritmo genético (salvo la excepción 15 20 5 30 λ=0.1). Si comparamos la norma de ambos algoritmos, puede observarse que para problemas pequeños y medianos (i.e tamaños 8 12 5 30 y 15 20 5 30), la metaheurística híbrida produce mejores resultados que el algoritmo genético estándar para valores grandes de U, es decir, para λ = 0.1 y 0.01. Sin embargo, para problemas de gran tamaño (i.e. 22 28 5 30), nuestro algoritmo se convierte en un método de estimación mejor respecto al algoritmo genético estándar incluso para valores pequeños de U. Por consiguiente, siempre que la metaheurística híbrida mejora las estimaciones dadas por MC2E, también mejora las del algoritmo genético estándar.

En conclusión, MC2E es el método de estimación más apropiado cuando lo valores de U son pequeños, es decir, cuando los errores son serialmente incorrelados (o al menos la dependencia serial no es muy fuerte). No obstante, a medida que los valores de U aumentan y no se verifica este supuesto, la metaheurística híbrida plantea una opción preferible a MC2E. Además, es importante señalar que cuando la técnica propuesta proporciona las mejores soluciones, una condición de fin $MaxIter$ =10 combinado con un porcentaje de optimización pequeño es suficiente para superar los resultados devueltos por un algoritmo genético estándar con una número de iteraciones considerablemente grande. Obviamente, la carga computacional y tiempos de ejecución de la metaheurística es mayor, pero también proporciona las mejores soluciones.

Capítulo 3

Conclusiones y trabajos futuros

En este capítulo se recopilan las principales conclusiones de la investigación realizada durante el periodo doctoral y se presentan los problemas a abordar en futuros trabajos. Las líneas de investigación que aquí se plantean deben entenderse como una consecución natural del trabajo contemplado en esta memoria.

3.1. Conclusiones

En los trabajos que componen esta tesis se analizan dos líneas de actuación para abordar la estimación de los modelos de ecuaciones simultáneas multinivel.

La primera de ellas es un estudio preliminar que busca tener en cuenta la heterogeneidad que puediera existir en el conjunto de variables endógenas de los modelos tradicionales de ecuaciones simultáneas a fin de obtener mejores estimaciones. Este análisis se centra en las variables endógenas ya que son las que básicamente se utilizan *a posteriori* para realizar predicciones, extraer conclusiones o tomar decisiones. Para ello, en una primera aportación se estima un modelo de ecuaciones simultáneas multigrupo aplicando un algoritmo variacional basado en entropía. Este estudio resalta la mejora en las estimaciones en modelos de ecuaciones simultáneas de diferentes tamaños en presencia de heterogeneidad. Por el contrario, se evidencia también el problema de la dimensionalidad, i.e. el número de variables que intervienen en el modelo, como la principal limitación computacional del procedimiento propuesto.

La segunda línea de investigación constituye el cuerpo central de esta tesis y aborda el problema desde otro punto de vista diferente. Esta vía propone modificar los supuestos de los modelos de ecuaciones simultáneas de manera que puedan contemplarse situaciones donde los términos de error estén intertemporalmente correlacionados. Esta estrategia supone introducir un nuevo modelo denominado modelo de ecuaciones simultáneas multinivel (MESM) y se ha dividido en dos fases de trabajo: cuando se conocen las matrices de covarianzas y en el caso general, cuando son desconocidas.

La segunda de las aportaciones recoge la definición y estimación teórica del nuevo modelo desarrollado. La introducción de una doble estructura en la matriz de covarianzas mejora considerablemente el ajuste de las estimaciones de los parámetros cuando

los errores no son serialmente independientes. La mayor contribución de este trabajo es, por tanto, el tratamiento de la dependencia serial de los errores en la propia formulación del modelo que evita cometer errores de estimación en los MES. De igual modo, se muestra el impacto que originan diferentes valores de dependecia serial en las estimaciones y los primeros resultados de simulación dejan entrever futuras limitaciones computacionales importantes.

Por último, la tercera aportación aborda la parte más compleja de la investigación. En esta fase se propone una metaheurística híbrida para la estimación de los MESM en el caso general. Los parámetros de ajuste del algoritmo se seleccionan tras un estudio experimental considerando diferentes opciones, todas pensadas para superar las dificultades computacionales emergentes en la fase anterior. Finalmente, el algoritmo propuesto se compara frente a otras alternativas y se reafirma la idoneidad de los parámetros de ajuste de la metaheurística seleccionados a pesar de ser valores modestos, lo que contribuye a aliviar considerablemente el coste computacional.

En síntesis, el objetivo de esta investigación es la mejora de las estimaciones en los modelos de ecuaciones simultáneas cuando se viola el supuesto de ausencia de correlación intertemporal en los términos de error. La técnica propuesta se presenta como mejor alternativa frente a otros métodos de estimación de los MES tradicionales bajo estas condiciones, principalmente MC2E, no diseñados para tal fin. De hecho, los experimentos muestran cómo se impone el modelo propuesto a medida que la correlación serial de los errores en los modelos de ecuaciones simultáneas es más fuerte.

3.2. Trabajos futuros

- **Utilización de derivadas en la estimación de los MESM.** La estimación teórica por máxima verosimilitud de los modelos de ecuaciones simultáneas multinivel está resuelta en el segundo de los trabajos que consitituyen esta tesis (véase [11]). Sin embargo, tanto la imposibilidad de aislar las matrices de parámetros A, B, U, Σ para su estimación sustituyendo en una fórmula la información conocida, como de calcular de manera secuencial o triangular las estimaciones de estas matrices condujo a la búsqueda de métodos alternativos para su obtención. En los trabajos presentados en esta tesis se optó por la utilización *solvers* de optimización y metaheurísticas, pero en ningún caso se hizo uso de la información teórica que podrían aportar las derivadas (3.2.3), (3.2.4), (3.2.5) y (3.2.6) obtenidas en [11].

  El objetivo en esta dirección de investigación es el desarrollo de un procedimiento que incorpore la información del sistema de ecuaciones del método de máxima verosimilitud a fin de reducir el número de parámetros a estimar y en consecuencia, la complejidad del problema. Este objetivo pensamos que puede abordarse desde distintas aproximaciones. Una estrategía sería centrarnos únicamente en el sistema de ecuaciones de verosimilitud y crear un método numérico iterativo para obtener las matrices de parámetros. Procedimientos de este tipo han sido anteriormente sugeridos para obtener el estimador de máxima verosimilitud en el campo de los modelos de ecuaciones simultáneas [19, 2].

Por otro lado, la información de las derivadas podría integrarse también en la heurística del algoritmo híbrido ya diseñado. Evidentemente, en cualquiera de los dos casos, el impacto de la utilización de esta información se traduciría directamente en una reducción de la carga computacional del problema. Por útlimo, cabe señalar que no está claro que todas las ecuaciones de verosimilitud puedan ser fácilmente implementadas, lo cual también conduciría a plantear distintas fases de trabajo dependiendo de la ecuación o ecuaciones que consideremos incorporar al procedimiento.

- **Inspección del espacio de parámetros en los MESM**. En el tercer trabajo, se desarrolló una metaheurística híbrida para la estimación de los MESM. Recordemos que con este método se perseguía reducir la complejidad del espacio de búsqueda de soluciones preservando su diversidad. Sin embargo, esta vía para afrontar tal fin no es única y por ello, sería necesario examinar el espacio de parámetros mediante otras ideas o algoritmos, así como analizar y comparar los resultados obtenidos con cada uno de ellos estudiando distintos aspectos importantes como:

  – El punto semilla utilizado para inicializar el algoritmo.

  – La convergencia y velocidad de convergencia del algoritmo según los distintos puntos semillas.

  – Estabilidad del algoritmo.

  Además, la elevada carga computacional y tiempos de ejecución de la metaheurística híbrida en [10] limita el uso de esta propuesta. Es interesante el desarrollo de una versión de memoria compartida y el estudio de otras técnicas heurísticas (Scatter Search, GRASP...). También, es conveniente profundizar en la investigación de diferentes métodos y funciones de optimización valorando las ventajas e inconvenientes de cada uno de ellos, para finalmente elegir el que mejor se adapte a nuestro problema.

- **Otros paradigmas de programación.** En la segunda y tercera aportación de esta tesis nos hemos centrado en comparar métodos de estimación. No obstante, de manera paralela a las técnicas matemáticas utilizadas para resolver el problema, es interesante también programar el algoritmo en paralelo utilizando dos paradigmas diferentes como memoria compartida (OpenMP) o memoria distribuida (MPI) con el objetivo de comparar eficiencia y rendimiento y poder paralelizar el código, otras de las cuestiones fundamentales a estudiar para reducir el coste computacional.

Hasta aquí los problemas propuestos son una continuación directa de los trabajos llevados a cabo en cada una de las publicaciones que constituyen esta tesis. Sin embargo, es importante considerar otras nuevas líneas de investigación también relacionadas.

- **Aplicación del modelo de ecuaciones simultáneas multinivel en ciencias biomédicas.** Debido a las características recogidas en los supuestos de los modelos de ecuaciones simultáneas multinivel, pensamos que una de las aplicaciones más interesantes sería su uso en ciencias biomédicas. Existen precedentes de modelos multinivel en los que se ajusta la endogenidad de alguna de las variables incluyendo ecuaciones adicionales que dan lugar a un modelo de ecuaciones simultáneas recursivo y que han sido aplicados en la asignación de recursos en los sistemas de salud y de educación (véase, por ejemplo [3, 21]). Los MESM abarcarían este tipo problemáticas y además, el caso general de simultaneidad de variables no recursivas.

- **Búsqueda del mejor MESM.** La investigación de esta tesis doctoral se centra en el desarrollo de un nuevo modelo estadístico, los modelos de ecuaciones simultáneas multinivel, y su estimación. Sin embargo, una pregunta que abre otra línea de trabajo y que es inevitable plantearse es si el MESM estimado es el que mejor se adapta y modeliza nuestro conjunto de datos. Una cuestión importante es garantizar que el MESM obtenido es el mejor de entre aquellos MESM que son isomorfos y compatibles para un mismo conjunto de datos, o al menos que se encuentra entre los mejores más allá de la información que puedan proporcionar criterios de comparación como AIC, BIC, criterios de entropía, etc.

  El objetivo de esta línea de investigación sería encontrar un modelo a partir de un conjunto de datos mediante algoritmos de búsqueda, es decir, sin desarrollar teóricamente ningún modelo de ecuaciones simultáneas previo. Esta idea ha sido trabajada anteriormente por López- Espín en [16] y aparece aplicada para el modelo Keynesiano Simple y un modelo para la preeclampsia (véase [5]), pero no han sido desarrolladas para los MESM. En este punto, proponemos extender a los MESM los algoritmos de búsqueda obtenidos para MES.

- **Redes neuronales en el contexto de los modelos de ecuaciones simultáneas.** Las redes neuronales son, por definición, un grupo interconectado de nodos (neuronas) entre los que se produce un intercambio de información mediante señales imitando la sinapsis en un cerebro biológico. En cada conexión, la información de salida de una neurona es la información de entrada de otra hasta que se obtienen unos valores finales. En otras palabras, las redes neuronales pueden interpretarse como un mapeo entre un conjunto de *inputs* y *outpus*, en nuestro caso serían, las variables exógenas y endógenas, respectivamente. Desde este punto de vista, la interrelación entre las variables endógenas característica de los MES es captada de forma natural por la interacción entre nodos de la red neuronal. Existen ejemplos de modelos de ecuaciones simultáneas planteados mediante redes neurales [14]. Siguiendo a L.R. Kumar, las principales razones para considerar esta aproximación son las siguientes:

  a) La mayoría de los métodos de sistemas son sensibles a la especificación del modelo y se restringen a sistemas lineales. Las redes neuronales multicapa

únicamente requieren la especificación de los *inputs* y de los *outpus*, pero no la relación de dependencia entre ellos.

b) La estimación de modelos con gran número de variables y ecuaciones puede requerir diferentes pasos cuando se utilizan métodos de estimación uniecuacionales mientras que solo se necesitaría una única red neuronal.

c) La arquitectura de las redes neuronales es paralela, lo cual conduce a una implementación más eficiente en problemas de grandes dimensiones.

Por ello, partiendo de esta idea pensamos que los modelos de ecuaciones simultáneas multinivel podrían plantearse como un problema de mapeo de redes neuronales. El objetivo a largo plazo en esta línea de investigación sería en primer lugar, profundizar en el uso de redes neuronales en MES y en segundo lugar, el planteamiento de los MESM mediante redes neuronales. Además, este nuevo enfoque permitiría por ejemplo incorporar información procedente de la investigación médica en los modelos mediante identidades e incluso fusionar información procedente de *deep learning*.

Conclusions and future work

This chapter summarises the main conclusions of the research made in the doctoral period and it lists the problems that will be addressed as future work. The research lines here suggested should be understood as a natural continuation of the work presented in this document.

Conclusions

The studies integrating this thesis analyse two lines of approach intended for the estimation of multilevel simultaneous equation models.

The former is a preliminary study aimed at taking into account the possible existence of heterogeneity in the set of endogenous variables of traditional simultaneous equation models in order to improve estimates. This analysis is focused on endogenous variables since these variables are basically the ones ultimately used for making predictions, drawing conclusions or making decisions. For that purpose, in a first approach, a variational algorithm based on entropy is applied to estimate multigroup simultaneous equation model. This work highlights the improvement in simultaneous equation models estimates in the presence of heterogeneity for different model sizes. However, it also evidences the curse of dimensionality problem, i.e. the number of variables involved in the model, as the main computational shortcoming of the suggested procedure.

The second line of research constitutes the main body of the thesis and it tackles the problem described above from a different point of view. This approach proposes modifying simultaneous equation model assumptions so as to intertemporally correlated error terms can be considered. This strategy entails the introduction of a new model denominated Multilevel Simultaneous Equation Model (MSEM) and it has been divided in two work phases: when covariance matrices are known and the general case, when these matrices are unknown.

The second paper presents the definition and theoretical estimation of the new developed model. The introduction of a double covariance matrix structure substantially improves parameter estimates when error terms are not serially independent. The greatest contribution of this work is, therefore, the treatment of the serial dependence of the error terms in the formulation of the model itself preventing misestimation in the SEM. Besides, the paper shows the impact in parameter estimates originated by different values of serial dependence and preliminary simulation results betray potential

important computational limitations.

Finally, the last study addresses the most complex part of the research. In this phase, a hybrid metaheuristic is proposed for MSEM estimation in the general case. The tuning parameters of the algorithm are selected after running an experimental study examining different options, all thought to overcome the computational difficulties emerged in the previous phase. To conclude, the proposed algorithm is compared with regard to other alternatives and it reinforces the suitability of the tuning parameters despite being modest values, which contributes to alleviate computational burden significantly.

In summary, the goal of this research is the improvement of simultaneous equation models estimates when the assumption of intertemporally uncorrelated error terms is violated. The technique proposed here offers a better alternative than other estimation methods used in traditional SEM under these conditions, mainly 2SLS, which are not designed for this purpose. Actually, the experiments carried out evidence that the developed model outperforms traditional procedures as error serial correlation in simultaneous equation models is stronger.

Future work

- **Use of derivatives in MSEM estimation.** Theoretical estimation of multilevel simultaneous equation model via maximum likelihood method is solved in the second paper integrating this thesis (see [11]). Nevertheless, both the impossibility of isolating matrices of parameters A, B, U, Σ for their estimation by substituting a priori information in a formula and the impossibility of obtaining these matrices estimates either sequentially or triangularly lead to search alternative methods for their calculation. In the studies gathered during this thesis, we opted for optimisation solvers and metaheuristics, but additional theoretical information that might provide derivatives (3.2.3), (3.2.4), (3.2.5) and (3.2.6) obtained in [11] was used in neither case.

  The aim of this line of research is to develop a procedure incorporating the information in the system of maximum likelihood equations so as to reduce the number of parameters to estimate and then, the complexity of the problem. This objective can be tackled from different approaches. One strategy could be to focus on the system of likelihood equations and to create an iterative numerical method to obtain the matrices of parameters estimates. Similar procedures have been previously suggested to obtain maximum likelihood estimator in the field of simultaneous equation models [19, 2].

  On the other hand, the information in the derivatives could also be integrated into the heuristic of the hybrid algorithm already designed. Evidently, in any case, the impact of using this information would directly result in a reduction of the computational burden of the problem. Lastly, it is worth noting that it is unclear whether all likelihood equations could be easily implemented. This fact

would also lead to set out different work scenarios depending on the equation or equations we consider to incorporate into the procedure.

- **Exploration of the space of parameters in MSEM**. In the third study, a hybrid metaheuristic was developed to estimate MSEM. It recalls that this method sought to reduce the complexity of the space of solutions preserving its diversity. However, the proposed approach for meeting this challenge is not unique and therefore, it would be necessary to examine the space of parameters by using other ideas or algorithms and to analise and compare the obtained results studying several important aspects such as:

  - Initial seed selection for the algorithm.
  - Convergence and convergence speed of the algorithm depending on different seed points.
  - Stability of the algorithm.

  Moreover, the high computational burden and execution times of the hybrid metaheuristic in [10] restrict the use of this proposal. It would be highly interesting to develop a shared memory version and to examine other heuristic techniques (Scatter Search, GRASP, ...). In addition, it would be convenient to deepen in alternative optimisation methods and functions assessing the advantages and disadvantages of each of them to eventually, choose the one that best suits our problem.

- **Other programming paradigms.** In the second and third contributions of this thesis, we have focused on the comparison of estimation methods. Nonetheless, in parallel to the mathematical techniques used to solve the problem it would be interesting to programme the algorithm in parallel using two different paradigms such as share memory (OpenMP) or MPI (distributed memory) and to compare the performance and efficiency of the algorithm in each case. Additionally, to reduce execution times, a parallel version of the code is one of the fundamental aspects to further study.

So far, the proposed problems are a straight continuation of the studies carried out throughout this predoctoral period. However, related new lines of research also need to be considered.

- **Application of multilevel simultaneous equation model in the biomedical sciences**. Because of the specific features of multilevel simultaneous equation model assumptions, we think that one of the most interesting applications of these models would be in biomedical sciences. There are precedents of multilevel models in which the endogeneity of some of the variables is adjusted including additional equations that lead to a recursive simultaneous equation model. These previous works can be mainly found in studies analysing resources allocation in the health and educational systems (see for example [3, 21]). MSEM would

embrace this type of situations and, on top of that, the general case when simultaneous relationship between variables is not recursive.

- **Search for the best MSEM**. The research carried out in this thesis is focused on the development of a new statistical model, the multilevel simultaneous equation model, and its estimation. However, an inevitable question that opens another line of work is that if the estimated MSEM is the one that best fits our data set. A fundamental point is to guarantee that the obtained MSEM is the best among those isomorphic MSEMs that are compatible for a same data set. In any case, the obtained MSEM has to be, at least, among the best models beyond the information that might be provided by comparison criteria such as AIC, BIC, entropy criteria, etc.

  The objective of this line of research would be finding a model from a data set by using searching algorithms, that is to say, without developing theoretically any preceding simultaneous equation model. This idea has been previously studied by López- Espín in [16]. In fact, it has been applied to the Simple Keynesian model and to a model for the preeclampsia (see [5]), but the idea has not been developed to MSEM. At this point, we propose to extend the obtained searching algorithms in SEMs to MSEMs.

- **Neural networks in simultaneous equation models framework.** Neural networks are, by definition, an interconnected group of nodes (neurons) exchanging information through signals that mimic the synapses in a biological brain. In each connection, the output information in one neuron is the input information of another one until certain final values are obtained. In other words, neural networks can be interpreted as a mapping between a set of inputs and a set of outputs, in our case, these will be the exogenous and endogenous variables, respectively. From this point of view, the interrelationship between the endogenous variables that defines the SEMs is captured in a natural way by the interaction between nodes of the neural network. There are examples of simultaneous equation models developed by using neural networks [14]. Citing L.R. Kumar, the main reasons for adopting this approach are the following:

  a) Most systems methods are sensitive to model specification and are restricted to linear systems. Multilayer neural networks only require the specification of the inputs and outputs, but not the depedent relationship between them.

  b) The estimation of models involving a large number of variables and equations might require different steps when uniequational methods are used whereas it would need a single neural network.

  c) Neural network architecture is parallel, which would lead to a more efficient implementation in high dimensional problems.

For all these reasons, we think that multilevel simultaneous equation model could be considered as a neural network mapping problem. The long-term objective in this line of research would be firstly, study the use of neural network in SEMs and secondly, the development of MSEMs by using this technique. Moreover, this new approach would allow to incorporate in the models information coming from medical research such as identities or even merging information from deep learning.

Anexos

Anexo I. Estimating Simultaneous Equation Models through an Entropy-Based Incremental Variational Bayes Learning Algorithm

- Hernández-Sanjaime, R., González, M., Peñalver, A., and López-Espín, J. J. (2021). Estimating Simultaneous Equation Models through an Entropy-Based Incremental Variational Bayes Learning Algorithm. Entropy 2021, 23(4), 384.
- https://doi.org/10.3390/e23040384

entropy

MDPI

Article

Estimating Simultaneous Equation Models through an Entropy-Based Incremental Variational Bayes Learning Algorithm

Rocío Hernández-Sanjaime *, Martín González, Antonio Peñalver and Jose J. López-Espín

Center of Operations Research, Miguel Hernández University, 03202 Elche, Spain; martin.gonzaleze@umh.es (M.G.); a.penalver@umh.es (A.P.); jlopez@umh.es (J.J.L.-E.)
* Correspondence: rocio.hernandezs@umh.es

Abstract: The presence of unaccounted heterogeneity in simultaneous equation models (SEMs) is frequently problematic in many real-life applications. Under the usual assumption of homogeneity, the model can be seriously misspecified, and it can potentially induce an important bias in the parameter estimates. This paper focuses on SEMs in which data are heterogeneous and tend to form clustering structures in the endogenous-variable dataset. Because the identification of different clusters is not straightforward, a two-step strategy that first forms groups among the endogenous observations and then uses the standard simultaneous equation scheme is provided. Methodologically, the proposed approach is based on a variational Bayes learning algorithm and does not need to be executed for varying numbers of groups in order to identify the one that adequately fits the data. We describe the statistical theory, evaluate the performance of the suggested algorithm by using simulated data, and apply the two-step method to a macroeconomic problem.

Keywords: computational econometrics; simultaneous equation model; clustering; variational algorithms; Shannon entropy; Leonenko estimator

Citation: Hernández-Sanjaime, R.; González, M.; Peñalver, A.; López-Espín, J.J. Estimating Simultaneous Equation Models through an Entropy-Based Incremental Variational Bayes Learning Algorithm. *Entropy* **2021**, *23*, 384. https://doi.org/10.3390/e23040384

Academic Editor: Stelios Bekiros

Received: 14 February 2021
Accepted: 18 March 2021
Published: 24 March 2021

1. Introduction

Simultaneous equation models (SEMs) constitute the reference statistical methodology in the analysis of jointly dependent variables [1]. Most applications are primarily found in econometrics [2,3], but also in medicine [4] and even in the study of divorce rates [5]. These models can be seen as multivariate regression models that reflect the simultaneity in structural relations in a system of multiple endogenous variables. Traditionally, it is common to assume homogeneity in the structural relations across observations and to estimate a single set of structural parameters. However, many practical situations in a wide range of disciplines (e.g., economics, finance, marketing, or sociology) involve structural changes in the studied variables or unobserved heterogeneity in data. Consequently, this simplification is often unrealistic and likely to produce misleading results.

Over the past few years, the problem of heterogeneity in data was studied for regression models. Many papers in the social sciences modelled anomalies by segmented regression. In this context, some aspects of change were investigated through formal statistical testing procedures [6,7]. Alternatively, structural-break inference was discussed using information criterion methods based on modifications of the Schwarz criterion (1978; BIC) [8,9], the Akaike criterion (1974; AIC) [10], and, more recently, the penalty term of several information criteria [11]. In particular, these models were widely applied in macroeconomics, where government interventions and policy changes at specific time points can affect both economic and market structures; examples include the analysis of regional stock exchanges in the United States [12] or the Euro area monetary policy [11]. Nevertheless, segmented regression also arises in a natural way in the context of industrial chemistry [13], agricultural and biological sciences [14], or climatology [15].

In the structural modelling framework, which embeds simultaneous equation models, researchers are also prone to estimating the model as if data belong to a single population [16]. However, this assumption does not always hold, and several authors cautioned against pooling data that may come from different segments [17,18]. Furthermore, traditional fit statistics do not alert about the presence of unaccounted heterogeneity in the model. Problems stemmed from failure to handle heterogeneity in structural equation models are illustrated in Jedidi et al. [19].

Typically, the procedure used to overcome these statistical difficulties is denominated multigroup structural equation modelling [20,21]. This supposes that the sample can be partitioned into G groups (also referred to as kernels from now on), assumed to be known a priori, and it estimates separate structural equation models for each of the G groups. In principle, models of different forms can be specified for each of the G groups. Nevertheless, the most relevant drawback of this a priori segmentation approach is that, in many situations, researchers do not know the number of groups that account for heterogeneity and do not have enough information to form segments a priori.

Thus, groups have to be determined from the data post hoc, and different strategies are possible. One option is to implement the finite-mixture structural equation model (STEMM) developed in Jedidi et al. [22] for simultaneously detecting and treating unobserved heterogeneity via the expectation-maximisation (EM) algorithm. This model generalises the multigroup structural equation model to the case in which group membership cannot be established a priori. In particular, it encompasses several specialised models, including finite mixture simultaneous equation models [19,23]. Alternatively, a sequential two-step process that first forms groups using some clustering algorithm applied to all variables, and then implements the multigroup structural equation modelling methodology to estimate each of the resulting groups can be considered. Note that for either approach, the numbers of groups must be prespecified when running the procedure, and this information is usually unknown. Methodologically, the finite-mixture model is robust and superior in goodness of fit to sequential data analysis strategies. However, one must consider that the clustering algorithm commonly used in the two-step scheme is K-means, which is not satisfactory, especially when the groups significantly overlap. Therefore, a two-step approach may be problematic [24,25], but its performance heavily depends on the type of clustering algorithm used in the first step.

Clustering procedures have two significant downsides. First, most clustering algorithms struggle with high-dimensional data. This weakness is recognised in the literature as the curse of dimensionality [26]. Data-reduction methods (e.g., principal component analysis) were discussed in the context of clustering and structural equation modelling [27,28]. Nevertheless, in this work, we restricted dimensions to a manageable number. Second, a caveat is that very large samples are required to perform cluster analysis. In the structural equation model framework, if the sample size is modest, the researcher may have no choice but to use other approaches such as the MIMIC method [16] or the STEMM approach. However, in many studies, we expect the sample size to be reasonably large.

This paper proposes a two-step procedure that in the first step picks the appropriate number of groups and classifies observations using an entropy-based incremental variational Bayes (EBIVB) algorithm. Traditionally, the two-step strategy used in the literature when struggling with heterogenous observations includes the standard K-means algorithm in the first step, which is a nonhierarchical distance-based algorithm. Unlike this approach, we used an entropy-based hierarchical clustering. There are two main advantages in our proposal. First, the number of clusters is not fixed and does not need to be specified by the researcher. Second, the use of entropy as the similarity measurement in the clustering step avoids distance calculation, reducing the outlier effect on cluster quality. Moreover, previous studies in the structural equation modelling context analysed the robustness of different model selection criteria (e.g., CAIC, BIC) in choosing the correct number of groups. To our knowledge, this study is the first using a clustering algorithm

that identifies the unknown groups based on the Gaussian deficiency (GD). On the whole, these factors are expected to improve model estimation. The rest of the paper is divided as follows: Section 2 provides a brief overview of the used statistical model and clustering algorithm. In Section 3, the two-step method is tested in a simulation study, and the obtained results are compared with other sequential approaches. Additionally, performance comparisons on real data are illustrated in Section 4. Lastly, main conclusions and future work are listed in Section 5.

2. Methodology

2.1. Simultaneous Equation Models (SEMs)

Simultaneous equation models consist of a system of linear regression equations with jointly dependent variables. In SEMs, variables can be classified as (i) endogenous if they are explained through a set of variables, i.e., dependent variables; or (ii) predetermined if they are independent nonrandom variables, i.e., exogenous and lagged endogenous variables. The main property of a simultaneous equation model is that the model allows for endogenous variables to be incorporated as explanatory variables in other equations. This way, SEMs contemplate the interdependent relations across variables. Formally, the structural form of the model for a system with m equations, m endogenous variables, and k predetermined variables is

$$Y = YA + XB + U, \tag{1}$$

where $Y = [y^1, \ldots, y^m]$ is a $N \times m$ matrix of N observations of m endogenous variables, $X = [x^1, \ldots, x^k]$ is a $N \times k$ matrix of N observations of k predetermined variables and $U = [u^1, \ldots, u^m]$ is a $N \times m$ matrix of the structural disturbances of the system. Matrices A $(m \times m)$ and B $(k \times m)$ are the endogenous and exogenous unknown coefficient matrices, respectively (by convention, $a_{ii} = 0, i = 1, 2, \ldots, m$).

The error terms $u_{t\cdot}$ $(t = 1, \ldots, N)$ are assumed to be normally distributed:

$$u'_{t\cdot} \sim N(0, \Sigma), \quad E(u'_{t\cdot}, u_{t'\cdot}) = \delta_{tt'}\Sigma \qquad t, t' = 1, 2, \ldots, N \tag{2}$$

where $\delta_{tt'}$ is the Kronecker delta and Σ a positive definite matrix.

Moreover, we assume that error terms are uncorrelated with the predetermined variables of the system, and that there is no linear dependence among the predetermined variables. Lastly, assuming that $(I - A)$ is nonsingular, the reduced form of the system is given by

$$Y = X\Pi + V, \tag{3}$$

where $\Pi = B(I - A)^{-1}$, $V = U(I - A)^{-1}$.

The random vectors $v_{t\cdot}$ have the following properties:

$$\begin{gathered} v'_{t\cdot} \sim N(0, \Omega), \quad v_{t\cdot} = u_{t\cdot}(I - A)^{-1}, \quad \Omega = ((I - A)')^{-1}\Sigma(I - A)^{-1}, \\ Cov(v'_{t\cdot}, v_{t'\cdot}) = \delta_{tt'}\Omega \qquad t, t' = 1, 2, \ldots, N \end{gathered} \tag{4}$$

In SEMs, parameter estimation can only be accomplished when the system is identified. An equation i $(i = 1, \ldots, m)$ is identified if the order condition is fulfilled, i.e., $m_i - 1 \leq k - k_i$ where m_i and k_i are the number of endogenous and exogenous variables in equation i, respectively [29]. Then, if the model can be estimated, indirect least-squares (ILS), two-stage least-squares (2SLS), three-stage least-squares (3SLS), or maximum-likelihood (ML) methods are the customary calculation approaches [30].

The goodness of the model can be assessed by evaluating diverse global measures of fit, such as information criteria. The Akaike information criterion (AIC); its corrected version AICc [31]; or the Bayesian information criterion (BIC) were adapted to SEMs [32]. Here, AIC was chosen for a SEM, expressed as follows.

$$AIC = N\ln|\hat{\Sigma}_e| + 2\sum_{i=1}^{m}(m_i + k_i - 1) + m(m+1), \tag{5}$$

where $|\hat{\Sigma}_e|$ is the determinant of error covariance matrix e_i ($i = 1, \ldots, m$) and e_i the difference between y_i and its estimation given equation i.

If the information criterion of one model is lower than the information criterion of another, the former model is considered better than the second. Because AIC is based on the maximum-likelihood method, the covariance matrix of the prediction errors has to be minimised.

2.2. *Entropy-Based Incremental Variational Bayes Learning Algorithm (EBIVB)*

Gaussian mixture models are frequently used as a formal approach to clustering [33]. These statistical pattern-recognition techniques were designed to deal with complex probability density functions, and those based on Gaussian kernels are especially useful for modelling data when forming clustering structures as being generated by a set of different kernels. In this latter approach, the estimation of the parameters of each kernel can be carried out by using different methods: maximum likelihood (ML), maximum a posteriori (MAP), or Bayesian inference.

The third estimation option is based on a fully Bayesian inference model [34]. This method can be computationally demanding and may involve intractable integrals as the complexity of the model increases. Thus, some alternatives emerged to overcome these downsides: the Laplacian method [35], Markov chain Monte Carlo (MCMC) [36], and variational methods [37].

In this paper, we use an incremental extension of the variational Bayes (VB) method first introduced in Peñalver and Escolano [38]. This is an iterative procedure that starts with only one kernel ($K = 1$), which is initially provided by the sample, and at each iteration incorporates a new component into the mixture by splitting one of the current kernels.

Formally, given a dataset $X = \{x_1, \ldots, x_N\}$ of N observations, mixture distribution can be interpreted as a latent variable model that for each observation x_n introduces a set of binary latent variables describing which kernel from the mixture generated the observation, $z_{in} \in \{0, 1\}$ where $i = 1, \ldots, K$, and K is the number of kernels. In this way, $z_{in} = 1 \Leftrightarrow$ component i gave rise to observation x_n and $\sum_{i=1}^{K} z_{in} = 1$.

The conditional probability density function of observations X given $z = \{z_{in}\}$ is normal with mean μ_i and inverse covariance matrix T_i, stated as $P(X|\mu, T, z) = \prod_{n=1}^{N}\prod_{i=1}^{K}\mathcal{N}(x_n|\mu_i, T_i)^{z_{in}}$, where the prior distribution of z is the product of multinomials $P(z|\pi) = \prod_{i=1}^{K}\prod_{n=1}^{N}\pi_i^{z_{in}}$. In addition, the observations are assumed to be independently generated, and the conjugate priors over the means and inverse covariances are, respectively, $P(\mu) = \prod_{i=1}^{K}\mathcal{N}(\mu_i|0, \beta I)$ and $P(T) = \prod_{i=1}^{K}\mathcal{W}(\nu, V)$, where β is a small fixed parameter corresponding to a wide prior over μ, I is the identity matrix, $\mathcal{W}$ represents Wishart distribution, and V and ν stand for the scale matrix and the degrees of freedom for a wide prior over T. Lastly, the joint distribution is specified as

$$P(X, \mu, T, z|\pi) = P(X|\mu, T, z)P(z|\pi)P(\mu)P(T) \tag{6}$$

The objective is to optimise mixing coefficients π by maximising data marginal likelihood. This is analytically unfeasible, but variational methods [39] can provide a lower bound of $P(X, \mu, T, z|\pi)$ by introducing a distribution $\mathcal{Q}(\Theta)$ in the log marginal likelihood expression:

$$\begin{aligned}\log P(X|\pi) &= \log\sum_z\int P(X, \Theta|\pi)d\Theta = \log\sum_z\int \mathcal{Q}(\Theta)\frac{P(X, \Theta|\pi)}{\mathcal{Q}(\Theta)}d\Theta \\ &\geq \sum_z\int \mathcal{Q}(\Theta)\log\frac{P(X, \Theta|\pi)}{\mathcal{Q}(\Theta)}d\Theta = \mathcal{L}(\mathcal{Q})\end{aligned} \tag{7}$$

where $\Theta = \{\mu, T, z\}$ to simplify notation.

Because the true log likelihood is independent of $\mathcal{Q}$, we simply need to maximise the lower bound of the true marginal likelihood. If a mean-field approximation [39,40] is adopted, then $\mathcal{Q}$ can be factorised over the variables in Θ; so,

$$\mathcal{Q}(\Theta) = \mathcal{Q}_z(z)\mathcal{Q}_\mu(\mu)\mathcal{Q}_T(T) \tag{8}$$

Starting from the expressions for the joint distribution in (6), the lower bound in (7), the factorisation in (8) and the equations for the variational factors as described in Peñalver and Escolano [38], the lower bound $\mathcal{L}(\mathcal{Q})$ can be evaluated as

$$\begin{aligned}\mathcal{L}(Q) &= \langle \log P(X|\mu, T, z)\rangle + \langle \log P(z)\rangle + \langle \log P(\mu)\rangle + \langle \log P(T)\rangle \\ &\quad -\langle \log \mathcal{Q}_z(z)\rangle - \langle \log \mathcal{Q}_\mu(\mu)\rangle - \langle \log \mathcal{Q}_T(T)\rangle\end{aligned} \tag{9}$$

The estimation of the mixing coefficients in the mixture is obtained by the maximisation of the bound in (9) with respect to π and an expectation-maximisation procedure is needed. After each EM iteration, bound $\mathcal{L}(Q)$ should not decrease. This fact and the possible convergence of some coefficients to zero are used both as model order selection and stopping criteria.

At each iteration, this EBIVB approach compares the entropy of the underlying probability density function of each kernel in the mixture regarding the theoretical entropy of a Gaussian [41], and the worst component in terms of its Gaussian deficiency (GD) is selected. Therefore, this component is supposed to be the one with greater entropy difference with a true Gaussian. Then, the worst-adjusted kernel is removed and replaced by two new kernels adequately separated from each other.

For the calculation of the GD of a component, it is necessary to estimate the entropy of that component. The Leonenko estimator [42] is a k neural-network (NN) entropy estimator based on Shannon entropy formula $H(X) = -\int f(x)\log f(x)dx$, which may be interpreted as an average of the $\log f(x)$, being $f(x)$ an existing probability density function.

In order to estimate $H(X)$, probability distribution $P_k(\epsilon)$ of the distance between a sample x_i and its k-NN is examined. This way, if we consider a ball of diameter ϵ centred at x_i, and there is a point within distance $\epsilon/2$, then there are $k-1$ other points closer to x_i, and $N-k-1$ points farther from it.

$$P_k(\epsilon)d\epsilon = k\binom{N-1}{k}\frac{dp_i(\epsilon)}{d\epsilon}p_i^{k-1}(1-p_i)^{N-k-1}, \tag{10}$$

where p_i is the mass of the ϵ-ball and $p_i(\epsilon) = \int_{||\xi - x_i|| < \epsilon/2} f(\xi)d\xi$.

The expectation of $\log p_i(\epsilon)$ is expressed as

$$E(\log p_i) = \int_0^\infty P_k(\epsilon)\log p_i(\epsilon)d\epsilon = \psi(k) - \psi(N), \tag{11}$$

where $\psi(\cdot)$ is the digamma function.

If $f(x)$ is assumed to be constant all over the ϵ-ball, we can consider approximation $p_i(\epsilon) \approx \frac{V_d}{2^d}\epsilon^d\mu(x_i)$, where d is the dimension, and V_d is the volume of the unit ball $\mathcal{B}(0,1)$, defined as $V_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$. Then, approximation $\log f(\epsilon) \approx \psi(k) - \psi(N) - dE(\log\epsilon) - \log\frac{V_d}{2^d}$ can be formulated, and lastly, the estimation of $H(X)$ is given by

$$\hat{H}(X) = \psi(N) - \psi(k) + \log\frac{V_d}{2^d} + \frac{d}{N}\sum_{i=1}^{N}\log\epsilon_i \tag{12}$$

where $\epsilon_i = 2||x_i - x_j||$ is twice the distance between sample x_i and its k-NN x_j.

Once the worst component is selected and superseded by two new ones, a new standard step of the VBgmm method [43] with $K+1$ components is conducted to maximise the marginal likelihood for the updated number of components. This process is repeated until convergence is reached. The split fails when the new component does not provide a better fit to the data, and some of the mixing coefficients tend to zero. In this case, it is eliminated, and the algorithm ends with a mixture of K kernels instead of $K+1$.

The split proceeding is an ill-posed problem since it may have more than one solution that discontinuously depends upon the initial data. Furthermore, a new EM-like method is needed every time a split test is completed. Therefore, we ensure that the number of splits are controlled, and a particular case of the procedure described in Dellaportas and Papageorgiou [44] is used with that purpose. The overall process is linear with the number of kernels (one split per iteration) and not prone to initialisation. This fact accelerates convergence and prevents the algorithm from falling into a local maximum of the marginal likelihood function, improving the performance of other current variational methods.

3. Experiment Results

We conducted two simulation experiments to test the performance of the proposed two-step estimation strategy in handling heterogeneity for correctly specified simultaneous equation models. The first experiment compares the Akaike information criterion of different estimated models: aggregate analysis (AGG), which ignores heterogeneity; known group membership (GM); and percentage model (PM), which classifies a proportion p% of the observations in an incorrect group deliberately. The purpose of the second experiment is to study the behaviour of the solution estimates of the suggested sequential method when varying the number of groups in exactly identified and overidentified SEM, assuming that the distributional form is properly specified.

Four different values for endogenous variables were examined, m = 2, 4, 6, and 8, and the number of exogenous variables was fixed to $k = 10$. To approximate real-life applications, we used a sufficiently large random sample of 1000 observations per group in all simulations. This fixed sample size was generated assuming that data in each group have a multivariate normal distribution and the clusters were constructed to be either non overlapping or slightly overlapping. The next section describes in detail the data-generation process in the simulation studies. For each problem size, we performed 10 replications to reduce the chance of outliers.

Experiments were run in a parallel NUMA node with 4 Intel hexacore Nehalem-EX EC E7530 with 24 cores at 1.87 GHz and 32 GB of RAM. All tests were implemented in C code with the exception of the data-generation process and the clustering algorithm. R statistical package GNU R version 3.5.2. was used in the data simulation, and MATLAB library R2016b in the clustering procedure.

3.1. Data-Generation Process

Consider the simultaneous equation model (1) rewritten as follows:

$$YA^* + XB + U = 0, \tag{13}$$

where Y, X, B, and U correspond to the matrices described in (1), and $A^* = (A - I)$ has each main diagonal entry set to -1, so that each equation contains its main endogenous variable. Consider the observations of the endogenous (Y) and predetermined (X) variables clustered into G groups, that is,

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_G \end{bmatrix} \in \mathbb{R}^{N \times m} \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_G \end{bmatrix} \in \mathbb{R}^{N \times k}$$

In all experiments, the simultaneous equation model was constructed to satisfy the order condition, rejecting any codification that led to an underidentified model. This procedure ensured the identification of the system. For the structural model, each element in matrices A^* and B was randomly generated by following a uniform distribution over the interval from -10 to 10. The same coefficient matrices, A^* and B, were considered for all groups in the data-generation process, thereby assuming the model was initially the same across groups. However, once established, the model may be disrupted by unpredictable external events such as government intervention in the economy or changes in variables resulting from accidents or disasters.

In order to reflect changes due to unexpected exogenous factors that affect the endogenous variables, we directly induced shocks to the endogenous variables in the generation process. To this end, the sample data of the endogenous variables were created for G groups by varying the mean and dispersion parameters for each group. Hence, we considered the data in each of the clusters as the response of the set of endogenous variables at the time of a specific shock and at subsequent times. To avoid undersized samples, we set a fixed sample size of 1000 observations per group in all simulations. Thus, we also provided a sufficiently large sample size for the clustering algorithm. The endogenous variables were assumed to have multivariate normal distribution. For each group, $g \in G$, the mean of each endogenous variable was randomly chosen from a uniform distribution from -8 to 8 for $m = 2$, and -10 to 10 in the rest of the cases. The elements of the variance–covariance matrix of the endogenous variables were also randomly selected from uniform distribution from -8 to 8, -10 to 10, -15 to 15, and -20 to 20 for $m = 2, 4, 6, 8$, respectively. These ranges were chosen to allow for some degree of separation among groups. The objective was to create nonoverlapping or slightly overlapping clusters. We computed rmvnorm in R to generate random values from multivariate normal distribution, which constituted the observations of each of the groups.

In all experiments, once the coefficient matrices and endogenous variables had been generated, the exogenous variables were calculated. To this effect, we considered the QR decomposition of exogenous coefficient matrix $B \in \mathbb{R}^{k \times m}$, with $k \geq m$:

$$B = QR, \tag{14}$$

where $Q \in \mathbb{R}^{k \times m}$ with $Q^T Q = I_m$ and $R \in \mathbb{R}^{m \times m}$ is an upper triangular matrix.

Using the QR decomposition of B, and assuming $rank(B) = n$, so that R is nonsingular:

$$-YA^*R^{-1} = XQ = X\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}, \tag{15}$$

where $Q_1 \in \mathbb{R}^{m \times m}$ and $Q_2 \in \mathbb{R}^{(k-m) \times m}$.

Similarly, X could be partitioned into two submatrices, $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$ where $X_1 \in \mathbb{R}^{N \times m}$ and $X_2 \in \mathbb{R}^{N \times (k-m)}$, and structural form (13) can be expressed as

$$-YA^*R^{-1} = \begin{bmatrix} X_1 & X_2 \end{bmatrix}\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} = X_1Q_1 + X_2Q_2$$

$$X_1 = -[YA^*R^{-1} + X_2Q_2]\ Q_1^{-1} \tag{16}$$

To obtain the sample data of the set of exogenous variables, matrix X_2 was randomly generated following a uniform distribution over the interval from 0.1 to 1.1, and X_1 was calculated by substituting in (16), assuming Q_1 was nonsingular.

3.2. Simulation Experiments

3.2.1. First Experiment—Model Selection

This experiment compares the goodness of fit of different estimated models. Previous studies addressed the problem of unobserved heterogeneity in simultaneous equation

models from two different approaches. One option is to estimate a single-group simultaneous equation model by using aggregate data. However, by doing so, one implicitly assumes data homogeneity. The second option is to theorise that there are different groups that could follow different models allowing for parameter values to differ across groups.

Under this second strategy, if group membership is known a priori, the researcher can use standard multigroup methods. The main limitation of this approach is that, in most situations, groups are unknown and have to be determined. To illustrate the bias introduced by two-step procedures that apply a clustering algorithm followed by multigroup simultaneous equation analysis, a percentage model was included in the experiment. This model considers the known group membership model as starting point and selects a percentage of total observations $p\%$ to be assigned to an incorrect group. These observations are chosen from among those in each group falling out of the ellipsoid of equal concentration associated with a probability of 0.75. For each group, the ellipsoid is determined by the corresponding group multivariate normal distribution [45]. Lastly, these observations are reallocated into the group that minimises the distance from the observation coordinates to the cluster centroids, without considering their group of origin.

In order to test the robustness of the proposed clustering algorithm (CA), in this first experiment we examined the performance of the aggregate model (AGG), the known group membership model (GM), and the percentage model (PM) where $p = 5\%, 10\%,$ and 15%. Note that the known group membership model corresponds to the percentage model with $p = 0\%$. For ease of comparison, the known group membership (GM) model was considered to be the benchmark. Table 1 shows the experiment outcomes for the true number of groups $G = 4$.

Table 1. Akaike information criterion (AIC) mean value for different estimated models over $s = 10$ simulation runs and clustering classification error rate committed by the algorithm regarding known group membership model. Note: GM, group membership; PM, percentage model; AGG, aggregate model; CA, proposed clustering algorithm.

Size	**GM**	**PM**			**AGG**	**CA**	**CA Clustering Error**
m	$p = 0\%$	$p = 5\%$	$p = 10\%$	$p = 15\%$	Aggregate		%
2	76,003.15	76,069.52	76,200.78	76,601.07	82,421.45	75,964.97	1.65
4	144,038.25	148,422.05	150,889.29	152,474.46	153,744.05	144,130.05	0.77
6	238,407.47	244,234.13	247,418.12	248,074.44	249,785.10	238,794.03	0.67
8	332,163.59	338,568.06	343,189.90	344,504.72	349,888.61	333,404.96	0.39

Expectedly, the known group membership model outperformed aggregate and percentage analysis. In all cases, the aggregate model recorded the worst result. The AIC value decreased from the the percentage model with $p = 15\%$ to the known group membership model, i.e., $p = 0\%$. Interestingly, the goodness of fit provided by the sequential analysis was between the known group membership model (GM) and the percentage model (PM) with $p = 5\%$ for all cases, except for $m = 2$ when the clustering configuration found by the algorithm enhanced the known group membership results. Therefore, the clustering classification accuracy was equivalent to the performance of the percentage model with $p \in (0, 0.05)$. Furthermore, the percentage of error in clustering classification decreased when the dimension of the problem determined by the number of endogenous variables increased.

3.2.2. Second Experiment—Solution-Estimate Analysis

This experiment examines the performance of the clustering technique to select the optimal number of groups and to classify observations into the most accurate group. The principal goal is to determine the ability of the clustering algorithm to recover group membership. The proposed sequential analysis first applies the clustering algorithm to all

endogenous variables, and a simultaneous equation model is then estimated within each of the resulting groups. Alternatively, we could have estimated the simultaneous equation model for different numbers of clusters and selected the model returning the best fit with regard to a given information criteria.

To assess the strength of the clustering algorithm in detecting the correct number of groups, we calculated the AIC value of some randomly selected problems. Table 2 shows the evolution of the Akaike information criterion from the aggregate model ($G = 1$) up to the model estimated when the clustering algorithm stopped; in other words, up to the model with the optimal number of groups detected by the algorithm.

Table 2. Evolution of AIC mean value for different estimated models.

Size	**Number of Clusters**			
m	1	2	3	4
2	83,795.08	78,139.16	76,307.94	75,408.18
4	136,575.21	136,703.96	135,996.17	132,493.36
6	255,276.91	256,088.06	251,367.12	249,979.87
8	348,972.97	351,896.28	341,170.68	331,409.30

According to Table 2, for the selected problems, the clustering algorithm always stopped at $G = 4$. On the basis of the computational results, the goodness of fit value provided by the two-step strategy enhanced the score of the one-population model ($G = 1$). The Akaike information criteria improved from the aggregate model to the model estimated with the optimal number of clusters. Thus, the algorithm results were consistent with the evolution of the AIC. Overall, for $G = 4$ being the true number of groups, the algorithm picked the same number of groups as the known group membership model (GM) in 100% of the runs. The success rate of the algorithm was stable when the number of endogenous variables increased. As a final remark, note that the algorithm could pick an optimal number of groups different from the number of groups corresponding to the true model if the resulting clustering configuration was more accurate or plausible than the one in the original model.

4. Empirical Application

This section presents an illustrative application of our method to Klein's Model I. Klein [2] developed three Keynesian macroeconomic models to study the United States economy for the 1921–1941 period, and examine the consequences of different political measures. The smallest of these three models, known as Model I, describes the workings of the U.S. economy in terms of a simultaneous six-equation system: three behavioural equations, an equilibrium condition, and two accounting identities. The model may be written following the notation set out in Greene [46]:

$$\begin{aligned} C_t &= \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \alpha_3 (W_t^p + W_t^g) + \varepsilon_{1t} \\ I_t &= \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 K_{t-1} + \varepsilon_{2t} \\ W_t^p &= \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1} + \gamma_3 A_t + \varepsilon_{3t} \\ X_t &= C_t + I_t + G_t \quad \text{(equilibrium demand)} \\ P_t &= X_t - T_t - W_t^p \quad \text{(private profits)} \\ K_t &= K_{t-1} + I_t \quad \text{(capital stock)} \end{aligned}$$

where

$$
\begin{aligned}
C_t &= \text{Consumption expenditure} \\
I_t &= \text{Investment} \\
P_t &= \text{Private profits} \\
K_t &= \text{Capital stock} \\
G_t &= \text{Government nonwage spending} \\
T_t &= \text{Indirect business taxes plus net exports} \\
W_t^g &= \text{Government wages} \\
W_t^p &= \text{Private-sector wages} \\
X_t &= \text{Total demand} \\
A_t &= \text{Time trend measured as years from 1931}
\end{aligned}
$$

In the above model, we have 6 endogenous variables (C_t, I_t, P_t, W_t^p, X_t, and K_t), 3 lagged endogenous variables (P_{t-1}, X_{t-1} and K_{t-1}) and 4 exogenous variables, including the time trend (G_t, T_t, W_t^g and A_t). The three former equations linearly describe the consumption, investment, and private-sector wage bill, respectively. Additionally, we have three identities that express the total demand according to all production undertaken in the economy, the profits net of taxes, and the capital stock in any period, respectively.

Whether the data are homogeneous or not, a widespread practice is to estimate a single set of aggregate-level structural parameters. However, such aggregate estimates are not fully acceptable if there is significant heterogeneity in the data because no variation in the structural relations is permitted.

To test for heterogeneity, the proposed two-step procedure was used to obtain estimates of the structural parameters for different numbers of groups beyond the aggregate case of $G = 1$. For purposes of illustrating our algorithm, the application was conducted using time-series data from 1921 to 2000 in order to provide a reasonably sample size. All variables were measured in USD billions of 1996 [47] except for time A_t, measured in calendar years from 1961.

Moreover, to demonstrate the benefits of the proposed methodology, it was compared with other state-of-the-art algorithms, namely, it was contrasted with the two-step scheme that uses K-means as clustering algorithm. The resulting AIC values of the different approaches are shown in Table 3, departing from $G = 1$ until the stop criterion explained in the preceding sections was reached.

Table 3 offers two main conclusions. First, the proposed algorithm stopped when the sample is divided into three groups, suggesting that $G = 3$ is the optimal solution. Furthermore, the AIC criterion for model selection (AIC_{CA}) reinforced the choice of the three-group model provided by the stop criterion implemented in our methodology. The AIC value was the minimum for $G = 3$ groups, also pointing out that the three-group model was the best option to describe the data. The evolution and improvement of the AIC values for the different number of groups from $G = 1$ to $G = 3$ are shown in Table 3. Second, if other two-step techniques are used, and in particular a nonhierarchical clustering such as K-means, the number of clusters need to be specified at the beginning. After inspection of the AIC values ($AIC_{k-means}$) when varying the number of groups from 1 through 3, the best solution is also $G = 3$. However, as one may expect, the group membership configuration provided by the EBIVB method outperforms clustering obtained by K-means; thus, it offers a better model for any value of G.

In the EBIVB method, the three groups comprise 48.10%, 11.39%, and 40.51% of the sample, respectively. In contrast to these percentages, K-means proportions are approximately 53.16%, 27.85%, and 18.99%, respectively. The results indicate that the proposed algorithm obtained two groups for observations registered prior to 1969 (without following definite temporal classification), whereas the last group comprised observations

from 1969 to 2000. Instead, the K-means approach suggested two structural breaks in 1963 and 1985, which split the sample into three groups. Nonetheless, the economic implications of these findings go beyond the scope of this paper.

Table 3. Statistical Criteria for Model Selection

Number of Clusters	$AIC_{k-means}$	AIC_{CA}
1	2004.415	2004.415
2	1856.893	1855.913
3	1811.696	1638.756

5. Conclusions

Aggregate analysis yields to incorrect results when estimating simultaneous equation models in the presence of considerable heterogeneity in the data. Traditional sequential approaches that estimate separate models for distinct clusters obtained either by a priori assignment or via a clustering method such as K-means may also lead to unsatisfactory outcomes. An alternative two-step strategy for handling heterogeneity in the SEM context was introduced. An incremental variational Bayes clustering algorithm and multigroup simultaneous equation model methodology were combined to study the structural variations of the model. The main advantage of the proposed procedure is that the estimation algorithm does not need any reruns for determining the optimal number of groups and obtaining the clustering of the dataset. The number of groups does not need to be specified a priori, and as a novelty, the groups are formed on the basis of the Gaussian deficiency.

To assess the goodness of fit of this approach, a percentage model was included in the simulation experiments. The study highlighted the good reliability of the EBIVB algorithm in the identification and classification of the different clusters, with performance that was equivalent to a percentage model with p lower than 5%. Additionally, the Akaike information criteria of the two-step method reinforced the use of this option over other estimated models. Because of the variety of statistical criteria available for model selection, alternative information heuristics (e.g., CAIC and BIC) need to be examined, and other appealing global measures of fit such as entropy must be explored.

Although the EBIVB algorithm showed good behaviour in the experimental study, its properties could underperform in other situations with ill-conditioned data. For example, estimation problems may arise in the presence of substantially overlapping clusters or large numbers of endogenous variables. The main shortcomings may appear in pattern recognition when finding the prior distribution and mixing coefficients of the Bayesian method. Despite such potential limitations, the suggested procedure is preferable to other traditional approaches for estimating simultaneous equation models when struggling with heterogeneous observations.

Author Contributions: conceptualisation, R.H.-S. and J.J.L.-E.; methodology, J.J.L.-E. and A.P.; investigation, R.H.-S.;software, M.G.; writing—original-draft preparation, R.H.-S. and A.P.; writing—review and supervision, J.J.L.-E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministerio de Economía y Competitividad of Spain under Grant TIN2016-8056-R and a predoctoral contract from the Generalitat Valenciana and the European Social Fund to R.H.-S. under Grant ACIF/2018/219.

Data Availability Statement: The data presented in this study are available in Carnero, B. S., Seriñán, P. R., & García, M. M. (2002). El Modelo Klein I y los ciclos económicos. [Klein's Model I and economic cycles]. Review on Economic Cycles, 4(1).

Acknowledgments: The authors gratefully acknowledge the computer resources provided by the Scientific Computing and Parallel Programming Group of the University of Murcia for the

simulation study. The authors really appreciate positive comments and suggestions from the anonymous reviewers.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Hausman, J.A. Specification and estimation of simultaneous equation models. *Handb. Econom.* **1983**, *1*, 391–448.
2. Klein, L.R. *Economic Fluctuations in the United States, 1921–1941*; John Wiley & Sons, Inc.: New York, NY, USA, 1950.
3. Dornbusch, R.; Fischer, S. *Macroeconomics*, 3rd ed.; Little, Brown: New York, NY, USA 1984.
4. King, T.M. Using simultaneous equation modeling for defining complex phenotypes. *BMC Genet. BioMed Cent.* **2003**, *4*, S10.
5. Ressler, R.W.; Waters, M.S. Female earnings and the divorce rate: a simultaneous equations model. *Appl. Econ.* **2000**, *32*, 1889–1898.
6. Andrews, D.W. Tests for parameter instability and structural change with unknown change point. *Econom. J. Econom. Soc.* **1993**, *61*, 821–856.
7. Bai, J.; Perron, P. Estimating and testing linear models with multiple structural changes. *Econometrica* **1998**, *66*, 47–78.
8. Yao, Y.C. Estimating the number of change-points via Schwarz'criterion. *Stat. Probab. Lett.* **1988**, *6*, 181–189.
9. Liu, J.; Wu, S.; Zidek, J.V. On segmented multivariate regression. *Stat. Sin.* **1997**, *7*, 497–525.
10. Ninomiya, Y. Information criterion for Gaussian change-point model. *Stat. Probab. Lett.* **2005**, *72*, 237–247.
11. Hall, A.R.; Osborn, D.R.; Sakkas, N. Inference on structural breaks using information criteria. *Manch. Sch.* **2013**, *81*, 54–81.
12. McZgee, V.E.; Carleton, W.T. Piecewise regression. *J. Am. Stat. Assoc.* **1970**, *65*, 1109–1124.
13. Dunicz, B. Discontinuities in the surface structure of alcohol-water mixtures. *Kolloid-Zeitschrift und Zeitschrift für Polymere* **1969**, *230*, 346–357.
14. Sprent, P. Some hypotheses concerning two phase regression lines. *Biometrics* **1961**, *17*, 634–645.
15. Werner, R.; Valev, D.; Danov, D.; Guineva, V. Study of structural break points in global and hemispheric temperature series by piecewise regression. *Adv. Space Res.* **2015**, *56*, 2323–2334.
16. Muthén, B.O. Latent variable modeling in heterogeneous populations. *Psychometrika* **1989**, *54*, 557–585.
17. Kohli, A.K. Effects of supervisory behavior: The role of individual differences among salespeople. *J. Mark.* **1989**, *53*, 40–50.
18. Day, R.L. Extending the concept of consumer satisfaction. *ACR N. Am. Adv.* **1977**, *4*, 149–154.
19. Jedidi, K.; Jagpal, H.S.; DeSarbo, W.S. Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Mark. Sci.* **1997**, *16*, 39–59.
20. Jöreskog, K.G. Simultaneous factor analysis in several populations. *Psychometrika* **1971**, *36*, 409–426.
21. Sörbom, D. A general method for studying differences in factor means and factor structure between groups. *Br. J. Math. Stat. Psychol.* **1974**, *27*, 229–239.
22. Jedidi, K.; Jagpal, H.S.; DeSarbo, W.S. STEMM: A general finite mixture structural equation model. *J. Classif.* **1997**, *14*, 23–50.
23. Jedidi, K.; Ramaswamy, V.; DeSarbo, W.S.; Wedel, M. On estimating finite mixtures of multivariate regression and simultaneous equation models. *Struct. Equ. Model. A Multidiscip. J.* **1996**, *3*, 266–289.
24. Aitkin, M.; Anderson, D.; Hinde, J. Statistical modelling of data on teaching styles. *J. R. Stat. Soc. Ser. A Gen.* **1981**, *144*, 419–448.
25. McLachlan, G.J.; Basford, K.E. Mixture models. Inference and applications to clustering. In *Statistics: Textbooks and Monographs*; Dekker: New York, NY, USA, 1988.
26. Bernal-Rusiel, J.L.; Greve, D.N.; Reuter, M.; Fischl, B.; Sabuncu, M.R.; Initiative, A.D.N. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. *NeuroImage* **2013**, *66*, 249–260.
27. McLachlan, G. *Discriminant Analysis and Statistical Pattern Recognition*; John Wiley and Sons: Hoboken, NJ, USA, 2004; Volume 544.
28. Chang, W.C. On using principal components before separating a mixture of two multivariate normal distributions. *Appl. Stat.* **1983**, *32*, 267–275.
29. Damodar, N. *Basic Econometrics*; The Mc-Graw Hill: New York, NY, USA, 2004.
30. Dhrymes, P.J. *Econometrics: Statistical Foundations and Applications*; Springer-Verlag: New York Inc., 1974.
31. Hurvich, C.M.; Tsai, C.L. Regression and time series model selection in small samples. *Biometrika* **1989**, *76*, 297–307.
32. Gorobets, A.A. *The Optimal Prediction Simultaneous Equations Selection*; Technical Report; 2004.
33. Jain, A.; Dubes, R.; Mao, J. Statistical Pattern Recognition: A Review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 4–38.
34. Box, G.E.; Tiao, G.C. *Bayesian Inference in Statistical Analysis*; John Wiley and Sons: Hoboken, NJ, USA, 2011; Volume 40.
35. Husmeier, D. The Bayesian Evidence Scheme for Regularizing Probability-Density Estimating Neural Networks. *Neural Comput.* **2000**, *12*, 2685–2717.
36. MacKay, D. *Introduction to Monte Carlo Methods*; Learning in Graphical Models; Jordan, M.I., Ed.; MIT Press: Cambridge, MA, USA, 1999; pp. 175–204.
37. Nasios, N.; Bors, A. Variational Learning for Gaussian Mixtures. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2006**, *36*, 849–862.
38. Peñalver, A.; Escolano, F. Entropy-Based Incremental Variational Bayes Learning of Gaussian Mixtures. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *23*, 534–540.

39. Corduneau, A.; Bishop, C. *Variational Bayesian Model Selection for Mixture Distributions*; Morgan Kaufmann: Burlington, MA, USA, 2001; pp. 27–34.
40. Ghahramani, Z.; Beal, M.J. Variational Inference for Bayesian Mixtures of Factor Analysers. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2000; pp. 449–455.
41. Peñalver, A.; Escolano, F.; Sáez, J.M. Learning gaussian mixture models with entropy-based criteria. *IEEE Trans. Neural Netw.* **2009**, *20*, 1756–1772.
42. Leonenko, N.; Pronzato, L. A class of rényi information estimators for multi-dimensional densities. *Ann. Stat.* **2008**, *36*, 2153–2182.
43. Constantinopoulos, C.; Likas, A. Unsupervised Learning of Gaussian Mixtures Based on Variational Component Splitting. *IEEE Trans. Neural Netw.* **2007**, *18*, 745–755.
44. Dellaportas, P.; Papageorgiou, I. Multivariate mixtures of normals with unknown number of components. *Stat. Comput.* **2006**, *16*, 57–68.
45. Mardia, K.; Kent, J.; Bibby, J. Multivariate analysis. In *Probability and Mathematical Statistics*; Academic Press Inc.: Cambridge, MA, USA, 1979.
46. Greene, W.H. *Econometric Analysis*; Pearson Education India: Delhi, India, 2003.
47. Carnero, B.S.; Seriñán, P.R.; García, M.M. El Modelo Klein I y los ciclos económicos [Klein's Model I and economic cycles]. *Rev. Econ. Cycles* **2002**, *4(1)*.

Anexo II. Multilevel Simultaneous Equation Model: A novel specification and estimation approach

- Hernández-Sanjaime, R., González, M., and López-Espín, J. J. (2020). Multilevel simultaneous equation model: A novel specification and estimation approach. Journal of Computational and Applied Mathematics, 366, 112378.
- https://doi.org/10.1016/j.cam.2019.112378

Multilevel simultaneous equation model: A novel specification and estimation approach

Rocío Hernández-Sanjaime *, Martín González, Jose J. López-Espín

Center of Operations Research, Miguel Hernández University, Elche, Alicante, Spain

ARTICLE INFO

Article history:
Received 25 February 2019
Received in revised form 31 July 2019

Keywords:
Multilevel simultaneous equation model
Maximum likelihood estimation
Matrix normal distribution
Simultaneous equation model
Multilevel model

ABSTRACT

Conventional simultaneous equation models assume that the error terms are serially independent. In some situations, data may present hierarchical or grouped structure and this assumption may be invalid. A new multivariate model referred as to Multilevel Simultaneous Equation Model (MSEM) is developed under this motivation. The maximum likelihood estimation of the parameters of an MSEM is considered. A matrix-valued distribution, namely, the matrix normal distribution, is introduced to incorporate an among-row and an among-column covariance matrix structure in the specification of the model. In the absence of an analytical solution of the system of likelihood equations, a general-purpose optimization solver is employed to obtain the maximum likelihood estimators. In a first approach to the solution of the problem, the adequacy of the matrix normal distribution is evaluated empirically in the case in which the double covariance structure is known. Using simulated data under the model assumptions, the performance of the maximum likelihood estimator (MLE) is assessed with regard to other conventional alternatives such as two-stage least squares estimator (2SLS).

1. Introduction

The limitations of classic statistical models to accurately reproduce the complexity of problems in which data are hierarchically structured or there is endogeneity between variables make the use of new methodological techniques necessary. Multilevel models and simultaneous equation models (SEM) have been developed for the statistical analysis of hierarchy and simultaneity, respectively. Nevertheless, models combining endogeneity and hierarchically structured data open a new line of research. The literature handling this mixture of factors is scarce and limited to recursive models.

The present paper addresses the estimation of the parameters of a Multilevel Simultaneous Equation Model (MSEM), that is, a SEM in which observed data are clustered into independent groups. An among-row and among-column covariance matrix structure is considered in order to take into account data correlation within groups. The matrix normal distribution allows to incorporate this specific patterned matrix in the estimation process and seems to be appropriate for this purpose. Further details of this distribution will be described in Section 3. Alternative matrix non-normal distributions, for example the matrix Student-t distribution, have also been assumed in recent studies to estimate parameters incorporating variability among individuals [1].

Previously, matrix normal distribution has been applied to the analysis of multivariate repeated measurements [2]. In this context, one can encompass an m-variate response observed on n occasions, either m variables measured at n time points for one subject or m variables measured for n subjects that belong to the same group, yielding in both situations

* Corresponding author.
E-mail addresses: rocio.hernandezs@umh.es (R. Hernández-Sanjaime), martin.gonzaleze@umh.es (M. González), jlopez@umh.es (J.J. López-Espín).

https://doi.org/10.1016/j.cam.2019.112378

an $n \times m$ observation matrix X. It should be noted that univariate repeated measurements and growth curves (also known as latent trajectory models where the repeated measurements are viewed as outcomes that depend on some metric of time (e.g. age, day or wave of measurement)) correspond to the $m = 1$ case. In these types of model analysing change, one variable is observed on n occasions and the degenerate matrix normal distribution is applied [3].

In the SEM estimation framework, it is usual to assume that errors are generated by a multivariate procedure with uncorrelated observations. In general, the errors have been supposed to follow a multivariate normal distribution [4]. Multilevel models allow dealing with grouped data but have been scarcely developed for multivariate response [5]. The aim of this paper is to merge multivariate response models with simultaneity and clustered data. Given the double covariance matrix structure, it is possible to bring these two relevant situations together: multivariate response as in a SEM and grouped correlated observations as in multilevel models.

Simultaneous equation models have been traditionally used in Econometrics, the best-known examples are the Klein's Model [6] or the macroeconomic IS-LM models [7]. However, their use has also been recently extended to other fields such as health sciences for modelling complex phenotypes [8] or even transport research for modelling the air traffic in the New York area [9]. Multilevel models have been widely implemented in cross-sectional studies from social and biomedical sciences in which units are naturally grouped at different levels (e.g. students in schools, voters in districts, etc.) or in longitudinal data such as clinical trials, when the same individual or unit response is repeatedly measured at several time points [10–12].

Applications of multilevel simultaneous equation modelling can be mainly found in studies analysing resources allocation in health or educational systems [13,14]. Nonetheless, the approach adopted basically consists of a multilevel model in which the endogeneity of some of the variables is adjusted including a second equation that creates a recursive simultaneous equation model. The extension proposed in this paper would not be confined to recursive models and aims to expand such practical situations.

Under the matrix normal distribution assumption, a random sample of independent and identically distributed (i.i.d.) groups provides the basis to derive the joint density of the new model. The parameters estimation is carried out via the maximum likelihood method. In the absence of a closed solution, a data sample is simulated and the maximum likelihood estimator is examined calling the R optimization function *nlm* from the stats package [15].

The rest of the article is organized as follows: Section 2 includes a brief overview of the statistical models employed and their most relevant characteristics. In Section 3, the MSEM is defined and its estimation via the maximum likelihood method is introduced. The simulation experiment proposed for solving the model in absence of analytic solutions is described in Section 4. This section also summarizes the numerical results obtained by using simulated data. Finally, main conclusions are listed in Section 5.

2. Statistical models

2.1. Simultaneous equation models (SEM)

Simultaneous equation model [16] consists of a system of linear regression equations that reflects the presence of jointly endogenous variables, i.e. the simultaneity between the set of variables of the model. Unlike single-equation models in which a dependent variable is a function of a set of independent variables, a SEM is a multi-equation model in which the dependent variable can appear as an explanatory variable in other equations. Formally, the structural form of the model

$$Y = YA + XB + U \tag{2.1.1}$$

where $Y = [y^1, \ldots, y^m]$ is a $N \times m$ matrix of N observations of m endogenous variables, $X = [x^1, \ldots, x^k]$ is a $N \times k$ matrix of N observations of k non-random predetermined variables which contains both exogenous and lagged endogenous variables, and $U = [u^1, \ldots, u^m]$ is a $N \times m$ matrix of the structural disturbances of the system. The matrices A $(m \times m)$ and B $(k \times m)$ are the endogenous and exogenous unknown coefficient matrices, respectively.[1]

The error terms $u_{t\cdot}$ $(t = 1, \ldots, N)$ are assumed to be serially independent random vectors normally distributed with 0 mean vector and covariance matrix Σ. Thus, the errors may be *contemporaneously* correlated but are *intertemporally* uncorrelated. That is, the rows of U, denoted $u_{t\cdot}$, have the properties:

$$u'_{t\cdot} \sim N(0, \Sigma), \quad E(u'_{t\cdot}, u_{t'\cdot}) = \delta_{tt'}\Sigma \qquad t, t' = 1, 2, \ldots, N \tag{2.1.2}$$

$\delta_{tt'}$ being the Kronecker delta and Σ a positive definite matrix.

Extensions to non-normal errors are possible [see [17] and [18]] but not considered in this work.

In addition, it is assumed that error terms are uncorrelated with the predetermined variables of the system, and there is no linear dependence among the predetermined variables so that the model has a unique interpretation in terms of its unknown parameters:

$$E(X'U) = 0 \quad \text{and} \quad rank(X) = k \tag{2.1.3}$$

[1] By convention, $a_{ii} = 0$, $i = 1, 2, \ldots, m$.

Finally, the coefficient matrix $(I - A)$ is assumed to be non-singular and the reduced form of the system becomes

$$Y = X\Pi + V \quad \text{where} \quad \Pi = B(I - A)^{-1} \quad \text{and} \quad V = U(I - A)^{-1} \tag{2.1.4}$$

The rows of V, $v_{t\cdot}$, are independent identically distributed (i.i.d.) random vectors with 0 mean vector and covariance matrix Ω :

$$\begin{aligned} &v'_{t\cdot} \sim N(0, \Omega), \quad v_{t\cdot} = u_{t\cdot}(I - A)^{-1} \quad \text{and} \quad \Omega = ((I - A)')^{-1}\Sigma(I - A)^{-1} \\ &Cov(v_{t\cdot}{}', v_{t'\cdot}) = \delta_{tt'}\Omega \qquad t, t' = 1, 2, \ldots, N \end{aligned} \tag{2.1.5}$$

The estimation of the parameters of the model can be tackled in two different ways: using limited information methods (single-equation methods) or full information techniques (system methods). The first approach that includes estimators such as indirect least squares (ILS), two-stage least squares (2SLS) or limited information maximum likelihood (LIML) treats each equation of the system in isolation. System methods such us three-stage least squares (3SLS) or full information maximum likelihood (FIML) estimate all the unknown parameters of the system simultaneously [19].

2.2. Multilevel models

Multilevel models, also called hierarchical linear models or linear mixed models among other denominations, are statistical techniques suitable for handling data that have a hierarchical, nested or clustered structure [5]. The existence of such dependent data structures implies that members of the same group share a set of features that derives in an intraclass correlation. Group effects describe how strongly units in the same group tend to resemble and influence each other. The statistical problems of ignoring these relationships may render invalid statistical conclusions [20].

For simplicity's sake, we will consider the 2-level model hereafter. More levels in the model and complex hierarchical structures shall be consulted in [5]. Model specification for multilevel models can be formulated in two different but equivalent approaches. One is based on a single equation that involves both fixed and random effects [21] while the other approach explicitly specifies the model in two levels with two different equations [22]. In this paper, we adopt the former representation expressed as follows:

$$y_i = X_i\beta + Z_i\mu_i + \varepsilon_i \qquad i = 1, \ldots, l \tag{2.2.1}$$

where y_i represents the n_i-response vector for the ith group of n_i individuals in cross-sectional data whereas it represents the n_i repeated measurements of the ith subject in longitudinal data, X_i is the $n_i \times p$ design matrix of the fixed effects, β is the p-vector of the fixed effects coefficients to be estimated, Z_i is the $n_i \times q$ design matrix of the random effects, μ_i is the q-vector of random effects for the ith group and ε_i is the n_i-vector of residuals [23]. It should be noted that X_i combines both level-1 and level-2 explanatory variables and $Z_i's$ columns are a subset of $X_i's$ $(q \leq p)$ incorporating random effects μ_i to y_i. That is, any component of β can be allowed to vary randomly by simply including the corresponding columns of X_i in Z_i [24].

The following distributional assumptions are made:

$$\begin{aligned} &\mu_i \sim N(0, D) \\ &\varepsilon_i \sim N(0, R_i) \\ &\mu_1, \ldots, \mu_l, \varepsilon_1, \ldots, \varepsilon_l \quad \text{independent} \end{aligned} \tag{2.2.2}$$

where μ_i reflects how the subset of regression coefficients for group i deviates from those of the population and ε_i comprises the residuals not explained by fixed or random effects. D and R_i are the covariance matrices of the multivariate normal distributions and l is the total number of groups [25]. The between variance component D is the same for all groups while R_i may vary across units.

Formally, the introduction of random effects helps to distinguish the conditional (group-specific) mean $E(y_i|\mu_i)$ and marginal (population-average) mean $E(y_i)$ as well as group-specific covariance $Cov(y_i|\mu_i)$ and population-average covariance $Cov(y_i)$:

$$\begin{aligned} E(y_i|\mu_i) &= X_i\beta + Z_i\mu_i \\ E(y_i) &= X_i\beta \\ Cov(y_i|\mu_i) &= R_i \\ Cov(y_i) &= Z_iDZ_i{}' + R_i \end{aligned} \tag{2.2.3}$$

Therefore, for each group

$$y_i \sim N(X_i\beta, Z_iDZ_i{}' + R_i) \tag{2.2.4}$$

This model structure allows units of the same group to be positively correlated, i.e. to account for intra-subject variability, and each group to diverge from the population allowing for inter-subject variability [25].

Finally, let consider the general model by stacking up all the groups, y_i, into a single column vector

$$y = X\beta + Z\mu + \varepsilon \tag{2.2.5}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_l \end{bmatrix} \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_l \end{bmatrix} \quad Z = \begin{bmatrix} Z_1 & 0 & \dots & 0 \\ 0 & Z_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Z_l \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_l \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_l \end{bmatrix}$$

Hence,

$$\begin{aligned} &y \sim N(X\beta, V) \quad \text{with} \quad V = ZGZ' + R \\ &G = diag(D, D, \dots, D) \quad R = diag(R_1, R_2, \dots, R_l) \end{aligned} \tag{2.2.6}$$

Standard estimation methods in multilevel models are maximum likelihood (ML) [26] and restricted maximum likelihood (REML) [27]. The aim is to calculate the fixed effects coefficients β as well as the variance components involved in V. The estimation of the fixed effects given the variance components is straightforward. Unfortunately, solutions to the variance components are not easy to handle computationally.

Maximum likelihood estimation methods require the maximization of the likelihood function which involves solving nonlinear equations. Historically, obtaining the estimators was a challenging computationally task. Nowadays, most statistical software have integrated routines for linear mixed models estimation in their packages: HLM [22], MLwiN [5] or nlme [23,28].

3. Multilevel simultaneous equation model

3.1. Definition of the Multilevel Simultaneous Equation Model (MESM)

Consider again a simultaneous equation model specified as in (2.1.1), but with observed data clustered into l independent groups

$$Y_j = Y_j A + X_j B + E_j \qquad j = 1, \dots, l \quad \text{independent groups} \tag{3.1.1}$$

Bearing in mind that ignoring groupings may invalidate many of the traditional statistical techniques, model assumptions of a SEM shall be modified. The error terms are no longer generated by a multivariate procedure with *intertemporally* uncorrelated observations. Therefore, distributional assumptions need to be reformulated.

A first approach to deal with this problem is to consider a double covariance matrix structure. The incorporation of the among-row and among-column covariance matrices allows specifying a covariance matrix for the variables and a covariance matrix for the group autocorrelation. This separable variance–covariance structure will provide the error distribution and will lead to more efficient inference.

Prior to introducing the MLE for the model proposed, the matrix normal distribution must be presented. Let X be an $n \times m$ random matrix and M, U, Σ $n \times m$, $n \times n$, $m \times m$ matrices, respectively, with U and Σ non-negative definite. Matrix M will represent the mean of the distribution whereas U and Σ the temporal autocorrelation and contemporaneous covariance matrices, respectively. By definition [29], X follows a matrix normal distribution with parameters M, U and Σ, denoted by $X \sim N_{n,m}(M, U, \Sigma)$, if X has the moment-generating function:

$$M_X(T) = exp\left\{ tr(M'T) + \frac{1}{2} tr(T'UTV) \right\} \quad \text{with } T \text{ an } n \times m \text{ matrix} \tag{3.1.2}$$

An equivalent definition involving the Kronecker product $\otimes$ and the vec operator is specified as:

$$X \sim N_{n,m}(M, U, \Sigma) \quad \text{if} \quad \text{vec}(X) \sim N_{np}(\text{vec}(M), \Sigma \otimes U) \tag{3.1.3}$$

Being U and Σ positive definite matrices, the distribution of X is said to be regular if X has the probability density function

$$f_X(X) = c^{-1} exp\left[-\frac{1}{2} tr\{U^{-1}(X - M)\Sigma^{-1}(X - M)^T\} \right] \tag{3.1.4}$$

with $c = (2\pi)^{nm/2} |U|^{m/2} |\Sigma|^{n/2}$

For model (3.1.1), the condition that each group has the same number of units will be imposed, so that the matrix U is common to all groups. The notation $n_1 = n_2 = \dots = n_l = n$ will be used hereafter.

Consider again a SEM with clustered data and applying the normal matrix distribution exposed above, for each group it results

$$Y_j = Y_j A + X_j B + E_j \qquad E_j \sim N_{n,m}(0, U, \Sigma) \tag{3.1.5}$$

with 0, U and Σ an $n \times m$, $n \times n$, $m \times m$ matrix, respectively.

And applying some basic properties:

$$Y_j = X_jB(I-A)^{-1} + E_j(I-A)^{-1} \quad \text{and} \quad W_j = Y_j - X_jB(I-A)^{-1} \tag{3.1.6}$$

we have that,

$$W_j \sim N(0, U, ((I-A)^{-1})^T \Sigma (I-A)^{-1}) \tag{3.1.7}$$

3.2. The MLE for a MESM

Under the normality assumption, a random sampling of l groups provides $n \times m$ i.i.d. matrices $E_1, \ldots, E_l$, from which the parameters estimators can be derived using maximum likelihood methods by formulating the appropriate function.

In view of the error distribution (3.1.6) and replacing W_j by the observable quantities $Y_j - X_jB(I-A)^{-1}$, the form of the joint likelihood function is stated by

$$\begin{aligned} f(W_1, \ldots, W_l) &= \prod_{j=1}^{l} f_j(W_j) = (2\pi)^{-\frac{nml}{2}} |U|^{-\frac{ml}{2}} |((I-A)^{-1})^T \Sigma (I-A)^{-1}|^{-\frac{nl}{2}} \\ exp &\left\{ -\frac{1}{2} \sum_{j=1}^{l} tr\left(U^{-1}(Y_j - X_jB(I-A)^{-1})(I-A)\Sigma^{-1}(I-A)^T(Y_j - X_jB(I-A)^{-1})^T \right) \right\} \end{aligned} \tag{3.2.1}$$

The logarithm of the likelihood function, $L = log\, f(W_1, \ldots, W_l)$, is given by

$$\begin{aligned} L = &-\frac{nml}{2} ln(2\pi) - \frac{ml}{2} ln|U| - \frac{nl}{2} ln|((I-A)^{-1})^T \Sigma (I-A)^{-1}| \\ &- \frac{1}{2} \sum_{j=1}^{l} tr\left(U^{-1}(Y_j - X_jB(I-A)^{-1})(I-A)\Sigma^{-1}(I-A)^T(Y_j - X_jB(I-A)^{-1})^T \right) \end{aligned} \tag{3.2.2}$$

The application of matrix derivatives [30–32] provides the system of likelihood equations:

$$\begin{aligned} \frac{\partial L}{\partial U} = &-mlU^{-1} + \frac{ml}{2} diag(U^{-1}) + \sum_{j=1}^{l} \left(U^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1}(Y_j(I-A) - X_jB)^T U^{-1} \right) \\ &- \frac{1}{2} \sum_{j=1}^{l} diag\left(U^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1}(Y_j(I-A) - X_jB)^T U^{-1} \right) = 0 \end{aligned} \tag{3.2.3}$$

$$\begin{aligned} \frac{\partial L}{\partial \Sigma} = &-nl\Sigma^{-1} + \frac{nl}{2} diag(\Sigma^{-1}) + \sum_{j=1}^{l} \left(\Sigma^{-1}(Y_j(I-A) - X_jB)^T U^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1} \right) \\ &- \frac{1}{2} \sum_{j=1}^{l} diag\left(\Sigma^{-1}(Y_j(I-A) - X_jB)^T U^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1} \right) = 0 \end{aligned} \tag{3.2.4}$$

$$\frac{\partial L}{\partial B} = \sum_{j=1}^{l} (X_j^T U^{-1} Y_j)(I-A)\Sigma^{-1} - \sum_{j=1}^{l} (X_j^T U^{-1} X_j) B \Sigma^{-1} = 0 \tag{3.2.5}$$

$$\frac{\partial L}{\partial (I-A)} = nl((I-A)^{-1})^T - \sum_{j=1}^{l} (Y_j^T U^{-1} Y_j (I-A)\Sigma^{-1} - Y_j^T U^{-1} X_j B \Sigma^{-1}) = 0 \tag{3.2.6}$$

Let $\hat{U}$, $\hat{\Sigma}$, $\hat{A}$ and $\hat{B}$ denote the maximum likelihood estimators of U, Σ, A, B, respectively. If we isolate some of the parameters above, it results from (3.2.5) and (3.2.6) that

$$\hat{B} = \left[\sum_{j=1}^{l} X_j^T \hat{U}^{-1} X_j \right]^{-1} \left[\sum_{j=1}^{l} X_j^T \hat{U}^{-1} Y_j \right] (I - \hat{A}) \tag{3.2.7}$$

$$\hat{\Sigma} = \frac{1}{nl} (I-\hat{A})^T \left\{ -\sum_{j=1}^{l} Y_j^T \hat{U}^{-1} X_j \hat{B} (I-\hat{A})^{-1} + \sum_{j=1}^{l} Y_j^T \hat{U}^{-1} Y_j \right\} (I - \hat{A}) \tag{3.2.8}$$

Replacing (3.2.7) and (3.2.8) in (3.2.3):

$$
\begin{aligned}
&-ml\hat{U}^{-1}+\frac{ml}{2}diag(\hat{U}^{-1})\\
&+\sum_{j=1}^{l}\left(\hat{U}^{-1}\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)\right.\\
&\left.\times\ V^{-1}\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)^T\hat{U}^{-1}\right)\\
&-\frac{1}{2}\sum_{j=1}^{l}diag\left(\hat{U}^{-1}\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)\right.\\
&\left.\times\ V^{-1}\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)^T\hat{U}^{-1}\right)=0
\end{aligned}
\tag{3.2.9}
$$

where

$$
V=\frac{1}{nl}\left\{-\sum_{j=1}^{l}Y_j^T\hat{U}^{-1}X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]+\sum_{j=1}^{l}Y_j^T\hat{U}^{-1}Y_j\right\}
$$

Replacing (3.2.7) and (3.2.8) in (3.2.4):

$$
\begin{aligned}
&-nl\hat{\Sigma}^{-1}+\frac{nl}{2}diag(\hat{\Sigma}^{-1})\\
&+\sum_{j=1}^{l}\left(\hat{\Sigma}^{-1}(I-\hat{A})^T\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)^T\right.\\
&\left.\times\ \hat{U}^{-1}\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)(I-\hat{A})\hat{\Sigma}^{-1}\right)\\
&-\frac{1}{2}\sum_{j=1}^{l}diag\left(\hat{\Sigma}^{-1}(I-\hat{A})^T\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)^T\right.\\
&\left.\times\ \hat{U}^{-1}\left(Y_j-X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]\right)(I-\hat{A})\hat{\Sigma}^{-1}\right)=0
\end{aligned}
\tag{3.2.10}
$$

where

$$
\hat{\Sigma}=\frac{1}{nl}(I-\hat{A})^T\left\{-\sum_{j=1}^{l}Y_j^T\hat{U}^{-1}X_j\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}X_j\right]^{-1}\left[\sum_{j=1}^{l}X_j^T\hat{U}^{-1}Y_j\right]+\sum_{j=1}^{l}Y_j^T\hat{U}^{-1}Y_j\right\}(I-\hat{A})
$$

By replacing (3.2.7) and (3.2.8) in (3.2.3) and also in (3.2.4) the four equation system is reduced to a two equation system that depends on U and $(I-A)$. System (3.2.9)–(3.2.10) has not a closed analytic solution and the estimation of the two matrices of parameters U and $(I-A)$ needs to be solved iteratively by designing a two-stage algorithm. Once these two matrices have been estimated, the pair $\hat{B}$ and $\hat{\Sigma}$ can be obtained by substitution in (3.2.7) and (3.2.8).

4. Numerical results

By definition, the MLE is the global maximum of the (log)-likelihood function. The standard way to proceed to obtain this estimator implies solving the system of likelihood equations described in Section 3 by setting each derivative equal

Table 1
Mean Euclidean distances $||\hat{A}-A||_{2,s}$ and $||\hat{B}-B||_{2,s}$ between estimate $\hat{A}$ and parameter A and between estimate $\hat{B}$ and parameter B, over $s = 10$ simulation runs. Mean fitness value and percentage of runs MLE improves 2SLS fitness score. $U = (u_{ij}) \in [-5, 5]$.

Size			2SLS		MLE_{nlm}		Fitness		%
m	k	l	$\|\|\hat{A}-A\|\|$	$\|\|\hat{B}-B\|\|$	$\|\|\hat{A}-A\|\|$	$\|\|\hat{B}-B\|\|$	2SLS	MLE_{nlm}	Improvement
2	3	5	$1.55_{1.73}$	$1.80_{1.92}$	$1.50_{1.62}$	$1.67_{1.74}$	−737.53	−267.43	100%
2	3	10	$2.22_{2.49}$	$3.50_{4.75}$	$1.42_{0.94}$	$1.91_{2.06}$	−5019.17	−760.49	80%
2	3	25	$0.98_{1.46}$	$1.82_{2.84}$	$0.94_{1.39}$	$1.67_{2.37}$	−1207.75	−296.99	100%
2	3	50	$1.02_{1.42}$	$1.29_{2.09}$	$1.04_{1.42}$	$1.39_{2.09}$	−50 338.95	−672.09	90%
8	12	5	$6.41_{1.50}$	$10.89_{3.19}$	$6.40_{1.46}$	$10.82_{3.15}$	−474 794	−31 080.5	90%
8	12	10	$8.62_{8.29}$	$13.03_{12.04}$	$8.63_{8.03}$	$12.99_{12.02}$	−932 483	−435 080	100%
8	12	25	$7.68_{3.90}$	$11.07_{4.67}$	$7.56_{3.79}$	$11.08_{4.69}$	−4 814 615.7	−603 524.25	100%
8	12	50	$4.31_{3.24}$	$5.70_{3.00}$	$4.32_{3.23}$	$5.69_{2.99}$	−1 182 940	−1 028 962.6	100%
10	15	5	$9.09_{3.95}$	$16.60_{8.00}$	$9.12_{4.03}$	$16.54_{7.96}$	−950 988.07	−80 876.12	100%
10	15	10	$6.71_{2.36}$	$9.20_{2.75}$	$6.68_{2.39}$	$9.18_{2.75}$	−3 281 584.5	−755 413.76	100%
10	15	25	$5.16_{1.96}$	$7.21_{2.73}$	$5.17_{1.95}$	$7.19_{2.73}$	−684 998 313	−424 518 142	100%
10	15	50	$5.03_{2.68}$	$6.32_{2.56}$	$5.03_{2.69}$	$6.31_{2.57}$	−55 048 498	−19 943 667	100%
15	20	5	$15.21_{2.35}$	$26.65_{7.54}$	$15.20_{2.35}$	$26.64_{7.54}$	−32 944 034	−13 236 945	100%
15	20	10	$13.13_{3.23}$	$20.00_{7.75}$	$13.13_{3.23}$	$20.00_{7.75}$	−3 953 024	−1 599 824.4	100%
15	20	25	$11.60_{3.17}$	$15.45_{4.77}$	$11.60_{3.17}$	$15.43_{4.77}$	−12 677 072	−1 599 824.4	100%
15	20	50	$9.91_{1.98}$	$12.27_{3.44}$	$9.91_{1.98}$	$12.27_{3.43}$	−1 394 508.6	−707 533.98	100%

to zero. Instead, the scheme here suggested on finding the MLE is to use a generic optimization solver based on numerical methods. The idea, in this paper, is simply to obtain a first approach to the MLE by setting up starting parameter values for the log-likelihood function and computing the *nlm* optimization solver included in the statistical software R.

Since the maximization of the log-likelihood function is a nonlinear problem, calculations for obtaining the MLE of the model proposed are cumbersome and numerical procedures are often sensitive to initial values. At this point, two situations will be distinguished: (1) estimation of coefficient matrices A and B for known covariance matrices U and Σ and (2) estimation of A and B with an unknown covariance structure.

In this paper, we focus on the estimation in MSEMs with known covariance matrices U and Σ. In the absence of *a priori* information, the choice of $\hat{A}_0 = A_{2SLS}$ and $\hat{B}_0 = B_{2SLS}$ will generally constitute a suitable initial solution for $\hat{A}$ and $\hat{B}$ although it postulates intertemporally uncorrelated observations.

The experiment aims to compare 2SLS algorithm and the optimization function *nlm* for different sizes of an MSEM in order to determine the most efficient method in each case. These two techniques differ in nature, 2SLS is based on least squares and thus minimizes the sum of squared residuals while the *nlm* function is applied to obtain the maximum of the likelihood function. In a SEM, 2SLS and limited information maximum likelihood estimators are asymptotically equivalent [19]. We seek to analyse whether the MLE for the new model proposed obtains better estimates than the 2SLS estimator in presence of serial dependence.

Experiments have been executed in a parallel NUMA node with 4 Intel hexa-core Nehalem-EX EC E7530, with 24 cores, at 1.87 GHz and 32 GB of RAM. All tests were carried out with a C code, including the call of the optimization function nlm of R. Namely, the R statistical package used is GNU R version 3.5.2.

Four different values for endogenous and exogenous variables were considered and $l = 5$, 10, 25, 50. Whatever the problem size, the number of observations in each group is $n = 5$. Tables 1 and 2 show the experiment outcomes for the same among-column covariance matrix Σ, but two different among-row covariance matrices U of the error terms distribution. In both cases, error disturbances E_j in Eq. (3.1.5) were generated by using the property stated in [29]: If $Z_j = (z_{st})$ denotes an $n \times m$ random matrix with z_{st} $(s = 1, \ldots, n; t = 1, \ldots, m)$ i.i.d. $N(0, 1)$, then

$$E_j = U^{1/2} Z_j \Sigma^{1/2} \sim N_{n,m}(0, U, \Sigma) \qquad j = 1, \ldots, l$$

On the basis of the simulation results, the fitness value of the likelihood function provided by the optimization solver is the same as the likelihood value given by the 2SLS estimator or enhances it in all cases. Tables 1 and 2 show the percentage of simulation runs in which the fitness value of likelihood function calculated by the optimization solver purely outperforms the 2SLS fitness likelihood value and also the mean fitness value in each case. As a measure of the dispersion of the estimator around the parameter, the mean Euclidean distance between estimate and parameter is evaluated.

For example in Table 1, for a problem size $m = 8$, $k = 12$, $l = 5$ and $n = 5$ the optimization solver score outperforms fitness 2SLS value 90% of the simulation runs and 10% of the times both techniques obtain the same likelihood value. The mean fitness value illustrates this improvement being −474 794 for 2SLS and −31 080.5 for the maximum likelihood method. Moreover, the mean Euclidean norm of the coefficient matrices is closer to the parameters using the maximum likelihood estimator. For the endogenous variables, the distance between estimate and parameter over $s = 10$ simulation runs is 6.41 with a standard deviation of 1.50 with 2SLS and 6.40 with a standard deviation of 1.46 with the MLE. The same situation is repeated for the exogenous variables.

From the results, one can gather that the MLE tends to outperform the 2SLS fitness score for small values of U, that is when the serial dependence is not very strong, as shown in Table 1. Nevertheless, the greater the U values are, the more

Table 2
Mean Euclidean distances $||\hat{A} - A||_{2,s}$ and $||\hat{B} - B||_{2,s}$ between estimate $\hat{A}$ and parameter A and between estimate $\hat{B}$ and parameter B, over $s = 10$ simulation runs. Mean fitness value and percentage of runs MLE improves 2SLS fitness score. $U = (u_{ij}) \in [-500, 500]$.

Size			2SLS		MLE_{nlm}		Fitness		%
m	k	l	$\|\|\hat{A} - A_0\|\|$	$\|\|\hat{B} - B_0\|\|$	$\|\|\hat{A} - A_0\|\|$	$\|\|\hat{B} - B_0\|\|$	2SLS	MLE_{nlm}	Improvement
2	3	5	$1.60_{1.46}$	$4.10_{4.72}$	$1.63_{1.48}$	$4.20_{4.64}$	−4355.14	−446.46	60%
2	3	10	$1.40_{1.61}$	$2.06_{1.85}$	$1.31_{1.51}$	$1.90_{1.79}$	−19 732.70	−786.22	70%
2	3	25	$0.70_{0.94}$	$1.69_{2.66}$	$0.69_{0.94}$	$1.67_{2.68}$	−2659.69	−1661.99	40%
2	3	50	$0.90_{0.95}$	$2.56_{5.38}$	$0.98_{1.01}$	$2.63_{5.37}$	−6870.14	−5600.67	20%
8	12	5	$8.27_{3.09}$	$16.86_{8.73}$	$8.16_{3.07}$	$16.77_{8.60}$	−1.18E+10	−90 849 708	100%
8	12	10	$7.02_{4.05}$	$10.78_{5.81}$	$6.95_{4.03}$	$10.75_{5.75}$	−2 409 591 512	−12 709 197	80%
8	12	25	$4.89_{2.01}$	$6.70_{2.42}$	$4.96_{2.08}$	$6.72_{2.45}$	−4 461 250.85	−761 458.57	100%
8	12	50	$3.17_{1.25}$	$4.27_{1.75}$	$3.20_{1.25}$	$4.28_{1.75}$	−6 406 963.37	−501 370.69	80%
10	15	5	$9.32_{1.90}$	$16.05_{3.04}$	$9.31_{1.90}$	$16.03_{3.05}$	−12 121 082	−146 919.62	90%
10	15	10	$11.13_{4.35}$	$16.89_{8.39}$	$11.09_{4.31}$	$16.89_{8.38}$	−180 931 364	−18 311 430	90%
10	15	25	$8.78_{4.85}$	$13.80_{7.67}$	$8.77_{4.85}$	$13.77_{7.69}$	−6 033 087.5	−10 361 773.7	100%
10	15	50	$5.46_{2.22}$	$7.42_{1.89}$	$5.45_{2.22}$	$7.40_{1.88}$	−20 594 215	−10 283 613	100%
15	20	5	$15.26_{2.57}$	$33.78_{17.26}$	$15.24_{2.60}$	$33.75_{17.23}$	−825 513.65	−236 981.94	100%
15	20	10	$14.65_{3.27}$	$24.08_{8.00}$	$14.65_{3.28}$	$24.07_{7.99}$	−793 610.27	−455 702.25	100%
15	20	25	$11.38_{3.01}$	$16.63_{5.64}$	$11.37_{3.02}$	$16.60_{5.63}$	$-1.17E+11$	−994 995 734	80%
15	20	50	$9.33_{2.18}$	$11.57_{2.42}$	$9.33_{2.19}$	$11.56_{2.42}$	−3 518 530.8	−2 077 883.1	60%

cases the MLE obtains the same 2SLS fitness score, as shown in Table 2. One of the reasons is attributed to increasing difficulties in the calculations needed for the implementation of the optimization solver.

Expectedly, according to the tables above, dispersion decreases when the sample size l increases, so $\hat{A}$ and $\hat{B}$ are consistent estimators of A and B, respectively. In each problem, there is little difference in the mean Euclidean distance between estimate and parameter for the 2SLS algorithm and the ML method. However, in general, the MLE shows lower values of dispersion. In both tables, the exogenous coefficient matrices show the largest values in the mean Euclidean norm.

5. Conclusions

The introduction of a double variance structure in a SEM in which the assumption of intertemporally uncorrelated error terms is violated lays the basis for the development of a modified model that we referred to as MSEM. The maximum likelihood (ML) estimation of an MSEM has been set out. In the absence of an analytical solution of the system of likelihood equations, the estimation of an MSEM has been carried out using a general-purpose optimization solver with simulated data under the assumption of known variance–covariance matrices.

In a first approach, selecting the 2SLS estimates of the coefficient matrices as starting values for the optimization solver has empirically proved that the obtained estimates are closer to the parameter than those calculated when the serial dependence of the errors is ignored. However, limitations in the optimization method integrated in the solver currently used to find the maximum of the likelihood function might underperform MLE. For this reason, other alternatives need to be explored and other general-purpose optimization solvers not based on gradient methods should be examined.

The estimation of the variance matrices is not straightforward and in our humble opinion, we consider that it requires a deep and separate study. Thus, the development of a complete methodology for estimating the parameters of a MSEM, including the variance components of the model in the case in which these parameters are unknown must be incorporated as future work. Moreover, it is interesting to include the development of restricted maximum likelihood method (REML) for MSEM and to compare estimates of variance components with maximum likelihood results. Finally, extensions of MSEM to a matrix non-Gaussian distribution of errors must be considered as further work.

Acknowledgements

This research was partially supported by a grant from the Ministerio de Economía y Competitividad of Spain (TIN2016-8056-R) and a predoctoral contract from the Generalitat Valenciana and the European Social Fund (ACIF/2018/219) to R.H. The authors gratefully acknowledge the computer resources and assistance provided by the Scientific Computing and Parallel Programming Group of the University of Murcia for the simulation study.

References

[1] J.E. Contreras-Reyes, F.O.L. Quintero, R. Wiff, Bayesian modeling of individual growth variability using back-calculation: Application to pink cusk-eel (genypterus blacodes) off Chile, Ecol. Model. 385 (2018) 145–153.
[2] P. Dutilleul, The mle algorithm for the matrix normal distribution, J. Stat. Comput. Simul. 64 (2) (1999) 105–123.

[3] M. Byakagaba, Apport de la matrice normale aux modèles d'analyse de la variance et des mesures rèpètèes (Unpublished Doctoral Dissertation), Université catholique de Louvain, Louvain-la-Neuve, Belgium, 1987, Faculty of Science.
[4] P.C. Phillips, Exact small sample theory in the simultaneous equations model, Handb. Econom. 1 (1983) 449–516.
[5] H. Goldstein, Multilevel Statistical Models, Vol. 922, John Wiley & Sons, 2011.
[6] L.R. Klein, Economic Fluctuations in the United States, Wiley, 1950, pp. 1921–1941.
[7] R. Dornbusch, S. Fischer, Macroeconomics, third ed., McGraw-Hill, New York, 1984.
[8] T.M. King, Using simultaneous equation modeling for defining complex phenotypes, in: BMC Genetics, Vol. 4, BioMed Central, 2003, p. S10.
[9] I. Lu, J. Peixoto, W. Taam, A simultaneous equation model for air traffic in the new york area, in: Air Transport Research Society World ConferenceAir Transportation Research SocietyAir Transport Research Society, 2003.
[10] S.A. Graham-Bermann, L. Miller-Graff, Community-based intervention for women exposed to intimate partner violence: A randomized control trial., J. Family Psychol. 29 (4) (2015) 537.
[11] J.H. Shin, Application of repeated-measures analysis of variance and hierarchical linear model in nursing research, Nurs. Res. 58 (3) (2009) 211–217.
[12] I.L. Simone, A. Ceccarelli, C. Tortorella, A. Bellacosa, F. Pellegrini, I. Plasmati, M.F. De Caro, M. Lopez, F. Girolamo, P. Livrea, Influence of interferon beta treatment on quality of life in multiple sclerosis patients, Health Qual. Life Outcomes 4 (1) (2006) 96.
[13] F. Steele, A. Vignoles, A. Jenkins, The effect of school resources on pupil attainment: a multilevel simultaneous equation modelling approach, J. R. Stat. Soc. A 170 (3) (2007) 801–824.
[14] R. Blundell, F. Windmeijer, Cluster effects and simultaneity in multilevel models, Health Econ. 6 (4) (1997) 439–443.
[15] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 2012. URL: http://www.R-project.org/.
[16] J.A. Hausman, Specification and estimation of simultaneous equation models, Handb. Econom. 1 (1983) 391–448.
[17] P. Phillips, Finite sample theory and the distributions of alternative estimators of the marginal propensity to consume, Rev. Econom. Stud. 47 (1) (1980) 183–224.
[18] J. Knight, Non-Normality of disturbances and the k-class structural estimator (Unpublished manuscript), Univ. of New South Wales, 1981.
[19] P.J. Dhrymes, Econometrics: Statistical Foundations and Applications, Springer Science & Business Media, 2012.
[20] M. Aitkin, D. Anderson, J. Hinde, Statistical modelling of data on teaching styles (with discussion), J. R. Stat. Soc. A (1981) 419–461.
[21] N.M. Laird, J.H. Ware, Random-effects models for longitudinal data, Biometrics (1982) 963–974.
[22] S.W. Raudenbush, A.S. Bryk, Hierarchical Linear Models: Applications and Data Analysis Methods, Vol. 1, Sage, 2002.
[23] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, et al., Linear and nonlinear mixed effects models, R package version 3 (2014).
[24] J.L. Bernal-Rusiel, D.N. Greve, M. Reuter, B. Fischl, M.R. Sabuncu, A.D.N. Initiative, et al., Statistical analysis of longitudinal neuroimage data with linear mixed effects models, Neuroimage 66 (2013) 249–260.
[25] X. Zhang, A Tutorial on Restricted Maximum Likelihood Estimation in Linear Regression and Linear Mixed-Effects Model, 2015.
[26] H.O. Hartley, J.N. Rao, Maximum-likelihood estimation for the mixed analysis of variance model, Biometrika 54 (1–2) (1967) 93–108.
[27] D.A. Harville, Maximum likelihood approaches to variance component estimation and to related problems, J. Amer. Statist. Assoc. 72 (358) (1977) 320–338.
[28] J.C. Pinheiro, D.M. Bates, Mixed-effects models in s and s-plus, 2011.
[29] S.F. Arnold, The Theory of Linear Models and Multivariate Analysis, Wiley, New York, 1981.
[30] D.A. Harville, Matrix Algebra from a Statistician's Perspective, Vol. 1, Springer, 1997.
[31] P.S. Dwyer, Some applications of matrix derivatives in multivariate analysis, J. Amer. Statist. Assoc. 62 (318) (1967) 607–625.
[32] K.B. Petersen, M.S. Pedersen, The matrix cookbook, Tech. Univ. Denmark 7 (15) (2008) 510.

Anexo III. Estimation of Multilevel Equation Models through Genetic Algorithms

- Hernández-Sanjaime, R., González, M., and López-Espín, J. J. (2020). Estimation of Multilevel Simultaneous Equation Models through Genetic Algorithms. Mathematics, 8(12), 2098.
- https://doi.org/10.3390/math8122098

mathematics

MDPI

Article

Estimation of Multilevel Simultaneous Equation Models through Genetic Algorithms

Rocío Hernández-Sanjaime *, Martín González and Jose J. López-Espín

Center of Operations Research, Miguel Hernández University, 03202 Elche, Spain; martin.gonzaleze@umh.es (M.G.); jlopez@umh.es (J.J.L.-E.)

* Correspondence: rocio.hernandezs@umh.es

Received: 21 October 2020; Accepted: 18 November 2020; Published: 24 November 2020

Abstract: Problems in estimating simultaneous equation models when error terms are not intertemporally uncorrelated has motivated the introduction of a new multivariate model referred to as Multilevel Simultaneous Equation Model (MSEM). The maximum likelihood estimation of the parameters of an MSEM has been set forth. Because of the difficulties associated with the solution of the system of likelihood equations, the maximum likelihood estimator cannot be obtained through exhaustive search procedures. A hybrid metaheuristic that combines a genetic algorithm and an optimization method has been developed to overcome both technical and analytical limitations in the general case when the covariance structure is unknown. The behaviour of the hybrid metaheuristic has been discussed by varying different tuning parameters. A simulation study has been included to evaluate the adequacy of this estimator when error terms are not serially independent. Finally, the performance of this estimation approach has been compared with regard to other alternatives.

Keywords: multilevel simultaneous equation model; maximum likelihood estimation; genetic algorithms; optimization

1. Introduction

Simultaneous equation models (SEM) [1] and multilevel models [2] have been widely discussed in the statistical literature and many applications arise in econometrics [3,4], medicine [5,6] or social sciences [7,8]. Simultaneous equation models contemplate jointly dependent or endogenous variables in a system of regression equations whereas multilevel models allow dealing with hierarchical or clustered data. However, clustered data may be problematic in the simultaneous equation models framework yielding to misleading results.

Indeed, grouped correlated observations may be precisely one of the causes that accounts for unobserved heterogeneity in simultaneous equation models. Typically, these models assume serially independent data, but it is well-known that ignoring the intraclass correlation across observations induced by groupings can lead to significant bias in the parameter estimates [9]. Under the motivation of overcoming limitations of traditional statistical models, it is necessary to think of new methodologies.

Multilevel simultaneous equation models combining both simultaneity and hierarchically structured data emerge as a valuable solution in this sort of situation. However, the literature exploring the confluence of these two factors in a single statistical models is scarce and it is worth investigating. To the best of our knowledge, theory and practice of multilevel simultaneous equation modelling are basically confined to recursive systems. So far, the existing methodology considers multilevel models in which the endogeneity of some of the variables has been adjusted including additional equations that create a recursive simultaneous equation model. Applications of these types of systems are essentially found in studies analysing resources allocation in the educational system or factors influencing women fertility [10,11].

Alternatively, starting out from a different point of view, a novel modelling framework denominated Multilevel Simultaneous Equation Model (MSEM) has been first introduced in Hernández-Sanjaime et al. [12]. This proposed approach is not restricted to recursive systems and could be gainfully used for estimating statistical models when faced with endogenous variables and temporal correlated observations. Potential applications would comprehend extensions of simultaneous equations models to problems in which error terms are temporally correlated. For instance, economic problems with temporal data, social science studies with hierarchical data (e.g., students in schools, people in districts...) or longitudinal research such as clinical trials, as long as these applications also entail jointly dependent variables.

This latter approach introduces a matrix distribution with a double variance structure in a SEM in which the assumption of intertemporally uncorrelated error terms is violated. The incorporation of an among-row and an among-column covariance matrix pattern aims to solve the misspecification of the model when this circumstance occurs. The maximum likelihood estimation of the parameters of an MSEM has been already established theoretically and the adequacy of the model has been assessed empirically when the double covariance strucuture is known [12].

Previously, under the assumption of known variance-covariance matrices, solutions to parameter estimates have been explored using a general-purpose optimization solver that selects 2SLS estimates of the coefficient matrices as starting values. The present paper addresses the estimation of the parameters of a multilevel simultaneous equation model in the general case, that is, when the double covariance structure is unknown. Although wanting to keep the same line of work in the general case, we realised that using only an optimization procedure would not be sufficient. Because of the absence of an analytical solution of the system of likelihood equations and the increasing complexity of the search space, a heuristic is required. Namely, a genetic algorithm that creates multiple sets of random parameters is considered.

Genetic algorithms are metaheuristics commonly used to generate high-quality solutions to optimization problems as an alternative to exhaustive search procedures [13,14]. In particular, genetic algorithms have been applied for estimating simultaneous equation models [15] and in the problem of finding the best SEM from a set of variables [16]. Likewise, the use of other metaheuristics or hybridations of several methods has been examined in the SEM context. A unified shared-memory metaheuristic scheme and parallel versions of different metaheuristics (GRASP, genetic algorithms, scatter search and their combinations) for obtaining a SEM from a dataset of variables have been presented and compared in Almeida et al. [17].

In this paper, a hybrid metaheuristic is developed to solve the problem of estimating an MSEM. The metaheuristic consists of a standard genetic algorithm that includes a call to the *nlm* general-purpose optimization function comprised in the statistical software R [18]. The obtained solution is efficient and its computational cost is considerably lower than finding a solution through exhaustive search procedures.

The rest of the paper is organised as follows. Section 2 reviews the main characteristics of simultaneous equation models. In Section 3, the multilevel simultaneous equation models is defined and its estimation via the maximum likelihood method is established. Section 4 describes the basis of the hybrid metaheuristic proposed for obtaining the parameters of an MSEM. Section 5 discusses different options for model estimation and reports the simulation results. Finally, the main conclusions, methodology limitations and future research are outlined in Section 6.

2. Simultaneous Equation Models

In the general linear regression model, one deals with a situation in which a dependent variable is expressed as a function of a set of explanatory variables that are either non-stochastic or at least independent of the error terms. Frequently, in some problems this assumption cannot hold. Many times we are interested in a model in which the dependent variable in one equation can also determine some of the explanatory variables by entering another equation (or equations), that is, a model in which a

number of variables are mutually and simultaneously determined. It is these types of situations that are considered in the theory of simultaneous equation models.

The structural form of a simultaneous equation model for a system with m equations, m endogenous variables (which influence and are influenced by other variables) and k predetermined variables (which influence but are not influenced by the system) is

$$Y = YA + XB + E \tag{1}$$

where $Y = [y^1, ..., y^m]$ is a $N \times m$ matrix of N observations of m endogenous variables, $X = [x^1, ..., x^k]$ is a $N \times k$ matrix of N observations of k non-random predetermined variables which contains both exogenous and lagged endogenous variables, and $E = [e^1, ..., e^m]$ is a $N \times m$ matrix of the structural disturbances of the system. The matrices A ($m \times m$) and B ($k \times m$) are the endogenous and exogenous unknown coefficient matrices, respectively (by convention, $a_{ii} = 0$, $i = 1, 2, ..., m$). The rows of E, denoted $e_{t.}$, have the properties

$$e_{t.}^T \sim N(0, \Sigma) \quad E(e_{t.}^T, e_{t'.}) = \delta_{tt'}\Sigma \qquad t, t^{'} = 1, 2, ..., N \tag{2}$$

$\delta_{tt'}$ being the Kronecker delta and Σ a positive definite matrix.

Thus, the error terms $e_{t.}$ $(t = 1, ..., N)$ are assumed to be serially independent random vectors normally distributed with 0 mean vector and covariance matrix Σ. Additionally, we make the customary assumptions that error terms are uncorrelated with the predetermined variables of the system and there is no linear dependence among the predetermined variables

$$E(X^T E) = 0 \quad \text{and} \quad rank(X) = k \tag{3}$$

Finally, assuming that $(I - A)$ is non-singular, the endogenous variables can be uniquely determined in terms of the predetermined and random variables of the system and the reduced form of the model becomes

$$Y = X\Pi + V \quad \text{where} \quad \Pi = B(I - A)^{-1} \quad \text{and} \quad V = E(I - A)^{-1} \tag{4}$$

Before considering the estimation of the model, we shall examine the identification problem. If the order condition $m_i - 1 \leq k - k_i$ holds, where m_i and k_i are the number or endogenous and exogenous variables in the i-th equation ($i = 1, ..., m$), an equation is identified. It is only in such case that parameters can be calculated and two-stage least squares (2SLS), indirect least squares (ILS), three-stage least squares (3SLS) or maximum likelihood (ML) are the most commonly estimation methods [19].

The preceding distributional assumptions about the structural disturbances involve that error terms may be contemporaneously correlated but are intertemporally uncorrelated. Nevertheless, the property that characterizes clustered data is the intraclass correlation and in the presence of that type of data structure, we cannot assert that the random terms of the system are intertemporally uncorrelated. Thus, we can hardly expect to obtain parameter estimates having desirable properties. In such case, the estimation techniques fail and it is necessary to establish alternative methods that provide at least consistent, unbiased and possibly efficient estimators.

3. Multilevel Simultaneous Equation Model

3.1. Definition of the Model

Consider a general simultaneous equation model specified as in (1) with observed data clustered into l independent groups

$$Y_j = Y_j A + X_j B + E_j \qquad j = 1, ..., l \quad \text{independent groups} \tag{5}$$

where Y_j, X_j denote the observations of endogenous and predetermined variables in group j and E_j the matrix of structural disturbances of the system in group j. Note that A, B are the same for all groups assuming that the observed data are generated by a single set of parameter values representing a unique structural model.

As it is well known, if groupings are ignored during the estimation process, one runs the risk of drawing erroneous statistical conclusions. Typically, multilevel models would be used to allow for data organized into groups, but these models are not designed to deal with endogenous variables. One option is to modify simultaneous equation model distributional assumptions so that error terms can be serially dependent, thereby taking into account data correlation within groups. This point and the fact that A and B are the same across groups will differentiate this proposal from previous research approaches when handling heterogeneity in simultaneous equation models framework.

The matrix normal distribution [20] incorporates an among-row and among-column covariance matrix structure that has motivated the development of a novel multivariate approach referred as to Multilevel Simultaneous Equation Model (MSEM) [12]. This separable variance-covariance patterned matrix appears to be adequate for the double objective targeted in this work and will be considered in the error terms distribution.

For model (5), imposing the condition that each group has the same number of units, n, we assume that error terms follow a matrix normal distribution

$$E_j \sim N_{n,m}(0, U, \Sigma) \qquad j = 1, ..., l \tag{6}$$

where 0 ($n \times m$) is the mean of the distribution and U ($n \times n$) and Σ ($m \times m$) are symmetric positive definite matrices specifying the covariance between units of the same group and the covariance between variables, i.e., the temporal and contemporaneous correlation, respectively (note that the condition of equal number of units in all groups let us suppose that the U covariance matrix is common across groups).

Applying some basic properties to express the values of the endogenous variables in the reduced form

$$Y_j = X_j B(I - A)^{-1} + E_j(I - A)^{-1} \quad \text{and} \quad W_j = E_j(I - A)^{-1} \tag{7}$$

we have that,

$$W_j \sim N(0, U, ((I - A)^{-1})^T \Sigma (I - A)^{-1}) \qquad j = 1, ..., l \tag{8}$$

Therefore, replacing W_j by the observable quantities $Y_j - X_j B(I - A)^{-1}$, the matrix normal density function is

$$f_j(W_j) = c^{-1} exp\left[-\frac{1}{2} tr\left(U^{-1}(Y_j(I - A) - X_j B)\Sigma^{-1}(Y_j(I - A) - X_j B)^T\right)\right] \tag{9}$$

with $c = (2\pi)^{nm/2}|U|^{m/2}|((I - A)^{-1})^T \Sigma (I - A)^{-1}|^{n/2}$.

3.2. Maximum Likelihood Estimation

Given the above distributional assumptions, the log-likelihood function for a random sample of l independent and identically distributed (i.i.d.) groups is

$$\begin{aligned} L(A, B, U, \Sigma) = &- \frac{nml}{2} ln(2\pi) - \frac{ml}{2} ln|U| - \frac{nl}{2} ln|((I - A)^{-1})^T \Sigma (I - A)^{-1}| \\ &- \frac{1}{2} \sum_{j=1}^{l} tr\left(U^{-1}(Y_j(I - A) - X_j B)\Sigma^{-1}(Y_j(I - A) - X_j B)^T\right) \end{aligned} \tag{10}$$

The problem is to maximize L with respect to the parameters A, B, U and Σ given the sample data (X_j, Y_j) $(j = 1, ..., l)$ and a fixed number of groups l.

Applying matrix derivatives [21–23], we obtain the system of likelihood equations:

$$\frac{\partial L()}{\partial U} = -mlU^{-1} + \frac{ml}{2} diag(U^{-1}) + \sum_{j=1}^{l} \left(U^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1}(Y_j(I-A) - X_jB)^T U^{-1} \right) - \frac{1}{2} \sum_{j=1}^{l} diag \left(U^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1}(Y_j(I-A) - X_jB)^T U^{-1} \right) = 0 \quad (11)$$

$$\frac{\partial L()}{\partial \Sigma} = -nl\Sigma^{-1} + \frac{nl}{2} diag(\Sigma^{-1}) + \sum_{j=1}^{l} \left(\Sigma^{-1}(Y_j(I-A) - X_jB)^T U^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1} \right) - \frac{1}{2} \sum_{j=1}^{l} diag \left(\Sigma^{-1}(Y_j(I-A) - X_jB)^T U^{-1}(Y_j(I-A) - X_jB)\Sigma^{-1} \right) = 0 \quad (12)$$

$$\frac{\partial L()}{\partial B} = \sum_{j=1}^{l} (X_j^T U^{-1} Y_j)(I-A)\Sigma^{-1} - \sum_{j=1}^{l} (X_j^T U^{-1} X_j) B\Sigma^{-1} = 0 \quad (13)$$

$$\frac{\partial L()}{\partial (I-A)} = nl((I-A)^{-1})^T - \sum_{j=1}^{l} (Y_j^T U^{-1} Y_j (I-A)\Sigma^{-1} - Y_j^T U^{-1} X_j B\Sigma^{-1}) = 0 \quad (14)$$

The maximum likelihood estimator (MLE) is defined as the global maximum of the (log)-likelihood function. The usual method to calculate this estimator involves solving the system of likelihood equations described above by setting each derivative equal to zero. Unfortunately, in this case the system has not a closed solution and it is not possible to isolate all the unknown parameters implicated.

4. A Genetic Algorithm for MSEM Estimation

The maximization of the log-likelihood function requires the resolution of a nonlinear system of equations that implies cumbersome calculations. In the absence of a closed analytical solution, the computational cost of completely exploring the space of solutions using exhaustive search procedures makes this option operationally unfeasible. As an alternative, the method here suggested for finding the MLE is to use a hybrid metaheuristic technique, which consists of an optimized genetic algorithm. This implementation is able to improve the solution provided by the genetic algorithm by computing the R general-purpose optimization function *nlm* from the stats package.

This section outlines the genetic algorithm programmed for estimating the multilevel simultaneous equation model (MSEM), which is described as follows. A population of chromosomes composed of multiple sets of parameters A, B, U and Σ is explored. Each chromosome determines an initial solution for the MSEM and represents a candidate for the MLE. The population is updated in the crossover and it can be enhanced by using an optimization solver in two different points of the algorithm. First, it can be randomly improved after the mutation function and secondly, once the genetic algorithm has concluded, a predefined set of chromosomes from among the fittest ones will be also optimized. The fitness of a chromosome is calculated evaluating the maximum likelihood function. The general scheme of a genetic algorithm is used (Algorithm 1) and its functions and parameters are described in detail hereunder.

4.1. Initialization and End Conditions

The initial population is randomly created and the population size (*PopSize*) is stated at the beginning. The generational process in the metaheuristic is repeated until it accomplishes the end condition. Namely, the algorithm stops when it reaches the maximum number of iterations (*MaxIter*), which will be studied for different predefined values.

4.2. Evaluating a Chromosome and Selecting the Best Ranking

The fitness value of each chromosome is calculated using Equation (10). For the selection of the fittest individuals, a benchmark set of chromosomes (*BenchSet*) composed of the individuals with higher fitness score is included. In each generation, a comparison of the evaluations of all chromosomes in the population is made and the benchmark set is updated removing those individuals with lower fitness value.

4.3. Crossover

In each iteration a proportion of the existing population is selected to breed a new generation and only a preset number of the individuals in the best ranking of the benchmark set will be chosen for reproduction (*RepSize*).

The individuals selected from the set of chromosomes to produce offspring are randomly paired in *CrossSize* couples, so that the same chromosome can be picked more than once and can be matched with multiple couples. Each pair of ascendent chromosomes (*ascendent*1 and *ascendent*2) will produce a new solution or *descendent*. The number of chromosomes that are created in each generation (*CrossSize*) is also a prespecified parameter.

Different methods can be used to combine the ascendents. In this work, the new individual inherits each of the matrices of parameters A, B, U or Σ completely from one of the ascendents. Therefore, for each new chromosome, a binary code randomly decides if a matrix is inherited from *ascendent*1 encoded with 1 or the *ascendent*2 encoded with 0.

4.4. Mutation

In each iteration, a small probability of mutation denoted by P_{mut} is considered. The mutation will only affect the elements of the diagonal of the covariance matrices U or Σ. If a chromosome from the new offspring is randomly chosen for mutation, all elements in the diagonal of any of these matrices will be subjected to mutation jointly by changing their numerical values.

4.5. Chromosomes Improvement

After mutation, a percentage of the descendent chromosomes is randomly chosen to be improved with a probability called P_{imp}. An optimization solver is included in the algorithm so that the new chromosomes can increment their quality to survive. The solver uses the selected chromosome as initial solution for maximizing the likelihood function (10). If it finds a better set of parameters for the model being estimated, the descendent chromosome is updated. If not, the chromosome remains equal.

Finally, once the descendents have been subjected to mutation and improvement, they can become part of the benchmark set. Each descendent is evaluated and the benchmark set is reordered considering these newly chromosomes. If the fitness score of a descendent is higher than any of the individuals in the benchmark set, the descendent will be included in the benchmark set in the corresponding position determined by its fitness value and the chromosome on the bottom of the ranking will be eliminated.

4.6. Optimization

Once the genetic algorithm has finished, the programme includes again a call to the optimization solver. In this step, a fixed number of the best ranked chromosomes (*OptSize*) is selected to be optimized. As described before, the chosen chromosomes are used as initial solutions by the solver and the fitness values of the resulting optimized chromosomes are calculated. Lastly, these scores are compared and the final solution shown by the metaheuristic corresponds to the parameters stored in the fittest individual. This step guarantees the search of a better final solution starting from a valid good one.

Algorithm 1: Parameterized schema for the hybrid metaheuristic

Initialize(ParamIni) $\rightarrow S\_ini$
Computefitness(S_ini, ParamIni)
Select(S_ini, ParamSelIni) $\rightarrow S\_ref$
while not EndCondition(S_ref, ParamEndCon) **do**
  Select(S_ref, ParamSel) $\rightarrow S\_sel$
  Combine(S_sel, ParamCom) $\rightarrow S\_com$
  Mutate(S_com, ParamMut) $\rightarrow S\_mut$
  Improve(S_com, S_mut, ParamImp) $\rightarrow S\_imp\_com, S\_imp\_mut$
  Computefitness(S_com, S_mut, S_imp_com, S_imp_mut, ParamCom)
  Include(S_com, S_mut, S_imp_com, S_imp_mut, S_ref, ParamInc) $\rightarrow S\_ref$
end while
Improve(S_ref, ParamImpPost) $\rightarrow S\_opt$
Select(S_opt, ParamSelBest)

5. Experiment Results

Experiments have been executed in a DELL PowerEdge R730 node with 2 Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30 GHz, with 20 cores (40 threads) at 2.4 GHz and 25 MB SmartCache memory. Tests were carried out in C code importing the *nlm* optimization function from statistical software R (GNU R version 3.5.2).

In the following simulations, some parameters have been fixed to predefined values. The initial population, *PopSize*, is integrated by 300 chromosomes and in each generation, a benchmark set, *BenchSet*, of 100 chromosomes is considered. The number of individuals selected from the benchmark set to be used in the crossover is *RepSize* = 20 and in each generation, *CrossSize* = 25 new chromosomes are created with a probability of mutation of 25 percent.

The rest of the parameters of the metaheuristic are studied experimentally. Table 1 reports the fitness provided by the algorithm for different values of the tuning parameters when estimating a medium size MSEM problem. Specifically, *Alg.Prev* denotes the fitness returned by the metaheuristic before entering in the optimization step 4.6, while *Alg.End* designates the fitness of a solution after the optimization is accomplished, if applicable. The cost of the algorithm measured in seconds is also presented. Likewise, *Prev* express the time spent by the procedure before entering in the optimization step and *Post* the time after this step is fulfiled.

Table 1. Average fitness and run time (over 10 simulation runs) for different parameters configuration of the hybrid metaheuristic.

Mode			Fitness		Time	
MaxIter	***OptSize***	P_{imp}	***Alg.Prev***	***Alg.End***	***Prev***	***Post***
10	0	0	−16,726.56	−16,726.56	0.32	0.32
50	0	0	−12,005.68	−12,005.68	0.90	0.90
100	0	0	−12,028.44	−12,028.44	1.60	1.60
10	10	0	−17,927.11	−10,568.38	0.32	1751.52
50	10	0	−12,098.04	−10,346.21	0.90	1732.36
100	10	0	−11,753.68	−10,319.02	1.60	1739.39
10	0	10	−9944.31	−9944.31	3663.19	3663.19
50	0	10	−8949.24	−8949.24	17,441.91	17,441.91
100	0	10	−8488.46	−8488.46	33,893.78	33,893.78
10	10	5	−10,364.71	−9992.68	2024.01	3776.39
50	10	5	−8814.78	−8756.06	8983.55	10,519.69
100	10	5	−8953.06	−8838.36	17,051.52	18,454.82

From the results, the rest of the parameters of the hybrid metaheuristic are henceforth established as follows. The probability of improvement, P_{imp}, is fixed to 5% in each iteration. The number of chromosomes to be optimized at the end of the genetic algorithm from among those ranked in the top of the benchmark set, $OptSize$, is fixed to 10. Note that the combination $P_{imp} = 0$ and $OptSize = 0$ corresponds to a standard genetic algorithm. The generational process is repeated until the end condition is reached, $MaxIter = 10$. Figure 1 summarises the iterative process and the parameters values eventually selected in each step of the hybrid metaheuristic which will be used to obtain the MLE.

The analysis in Table 1 indicates that including the call to the *nlm* function after mutation ($P_{imp} \neq 0$) or at the end of the process ($OptSize \neq 0$) outperforms in either case the standard genetic algorithm fitness value, as can be seen in column $Alg.End$. The results also suggest that inserting the optimization after the mutation function ($P_{imp} = 10$ and $OptSize = 0$) has a higher positive impact than just optimizing the fittest individuals at the end of the metaheuristic ($P_{imp} = 0$ and $OptSize = 10$). In fact, this last option does not provide a substantial improvement with regard to a basic genetic algorithm. Moreover, the simulation study illustrates that there is no significant difference between implementing the solver only after mutation and applying both optimizations ($P_{imp} = 5$ and $OptSize = 10$). However, we consider this last configuration a more complete choice since it would allow for at least a minimal enhancement involving a negligible cost in time. On the other hand, for any of the versions of the optimized genetic algorithm, it appears that the different values for the maximum number of iterations provide similar fitness scores. In contrast, the computer time speeds up as the number of iterations increases and leads us to choose $MaxIter = 10$ as end condition.

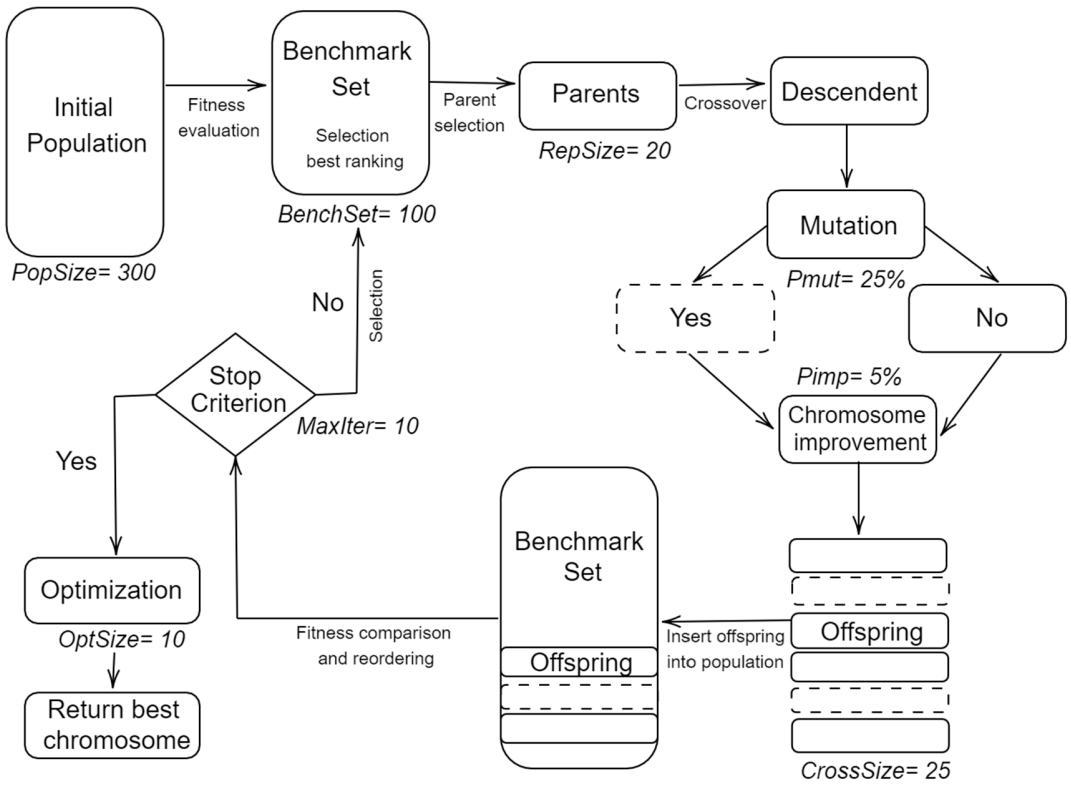

Figure 1. Hybrid genetic algorithm scheme.

Once parameters have been established according to evidences in Table 1, Table 2 shows the experiment outcomes for the hybrid metaheuristic developed in this work. In all experiments, the MSEM was codified so as to satisfy the order condition, thereby ensuring the estimation of the model. Henceforth, we shall assume that the MSEM is identified. The values of the chromosomes

which constitute the entry of the algorithm for the coefficient matrices A and B are, in half of the cases, 2SLS estimates of A and B with a variability from 0% to 0.75%. In the other half, A and B initial entries are selected randomly from the solution space. Although 2SLS method is not designed to deal with temporally correlated observations, 2SLS estimates offer a convenient starting solution for the search algorithm while the variability allows to expand the exploring in the solution space. In this way, we seek to reduce the algorithm dependence on good initial values and speed up the convergence. The starting values for U and Σ are generated randomly in such a way that symmetric positive definite matrices are guaranteed.

Finally, the performance of the hybrid metaheuristic has been examined regarding other estimation methods. The results are compared in terms of fitness, time and accuracy with two-stage least squares (2SLS) and a standard genetic algorithm (GA) with a maximum number of iterations of 10,000 which does not include a call to the *nlm* optimization function. The former method is the one commonly used in simultaneous equation models, which do not allow for intertemporally correlated error terms. Thus, using 2SLS in the MSEM case would involve an estimation bias. It is worth investigating the evolution of 2SLS estimates as the temporal correlation becomes significant. On the other hand, genetic algorithms are customary heuristics widely used in optimization and search problems when exact solutions are necessarily computationally expensive. Heuristics can provide an approximate solution individually or be used as a good baseline to be supplemented with an optimization procedure. This last case is the one considered in the hybrid metaheuristic. Hence, it is interesting to study whether or not the proposed optimized genetic algorithm produces better estimates than a standard genetic algorithm individually (even if a large number of iterations are permitted in the GA when used individually).

Three different values for endogenous and exogenous variables have been considered. Whatever the MSEM size, the number of groups is $l = 5$ and the number of observations in each group is $n = 30$. In addition, the covariance matrix U has been adjusted for five different ranges of values by dividing the generation interval by $\lambda = 100, 10, 1, 0.1$ and 0.01. Note that dividing the generation interval by $\lambda = 1$ returns the original range for the entries of U. For each model configuration of parameters *size* (m k l n) and λ, we have simulated five different models and each of them has been estimated in two different ways using whether the hybrid version or the simple genetic algorithm mentioned above. The exogenous variables have been simulated following a matrix normal distribution. The endogenous variables have been calculated using Equation (7), with the error disturbances E_j generated by using the property stated in Arnold [20]:

If $Z_j = (z_{st})$ denotes an $n \times m$ random matrix with $z_{st} \quad (s = 1, ..., n; t = 1, ..., m)$ i.i.d. $N(0, 1)$, then

$$E_j = U^{1/2} Z_j \Sigma^{1/2} \sim N_{n,m}(0, U, \Sigma) \quad j = 1, ..., l$$

On the basis of the experimental results, the fitness value of the likelihood function obtained either by the hybrid metaheuristic (F_{HM}) or by the standard genetic algorithm (F_{GA}) always outperforms 2SLS fitness score (F_{2SLS}). In addition, the fitness value of the likelihood function obtained by combining the metaheuristic with an optimization procedure clearly enhances the likelihood value achieved by using only the genetic algorithm. Therefore, the solving method proposed in the hybrid version improves the fitness over any of the other alternative procedures.

To assess the accuracy of the different estimation methods, the Frobenius norm between observed data Y and fitted values $\hat{Y}$ is calculated. Interestingly, 2SLS shows better results for small values of U, but as the range of values of U augments, the hybrid metaheuristic improves 2SLS whatever the MSEM size is. Nevertheless, this finding does not occur when using a standard genetic algorithm. In this case, 2SLS always produces better estimates than the genetic algorithm, with just one exception (size 15 20 5 30 and $\lambda = 0.1$).

Table 2. Average results for fitness, run time and accuracy (over 5 simulation runs) of different estimation methods.

Size	λ	F_{HM}	F_{GA}	F_{2SLS}	S.prev	S.post	S_{GA}	$\|\|Y-\hat{Y}_{HM}\|\|$	$\|\|Y-\hat{Y}_{GA}\|\|$	$\|\|Y-\hat{Y}_{2SLS}\|\|$
8 12 5 30	100	−3163.36	−3635.03	−411,983.91	434.99	922.77	57.25	537.39	116.48	89.19
	10	−3811.79	−4440.59	−57,824.25	584.70	1087.31	59.51	559.08	302.24	280.04
	1	−4094.60	−5748.64	−34,293.66	504.84	976.59	59.57	1123.97	948.76	899.87
	0.1	−7248.25	−10,943.99	−48,767,714.85	522.52	1030.72	59.67	2264.33	2658.97	2596.77
	0.01	−10,069.02	−61,501.99	−881,112.47	507.95	1008.34	59.51	7696.10	8519.65	8368.07
15 20 5 30	100	−7341.87	−8668.81	−141,301.20	1894.89	3680.11	142.04	1270.48	408.67	186.20
	10	−8411.78	−10,104.22	−7,104,505.59	1781.75	3547.47	141.76	1464.06	841.10	608.62
	1	−11,014.00	−12,694.56	−1,283,916.42	2310.64	4140.55	136.27	2563.91	2433.26	2076.20
	0.1	−12,097.40	−22,999.82	−327,573.88	2188.60	3900.75	127.63	5406.90	6179.77	6306.63
	0.01	−47455.24	−170,551.28	−21,252,787.68	1346.84	3189.38	130.23	14,815.85	20,587.63	18,881.66
22 28 5 30	100	−14,003.91	−16,217.86	−3,125,727.39	3851.21	8308.45	270.85	3922.10	2720.13	344.21
	10	−14,556.42	−17,178.54	−2,558,133.93	5222.54	9589.12	270.27	2274.98	2417.81	917.13
	1	−20,391.21	−28,325.12	−6,585,294.85	5454.31	9287.51	271.69	4343.76	5204.45	3823.36
	0.1	−22,702.70	−33,346.31	−18,947,400.33	3153.00	7526.67	270.77	8860.07	10,557.65	9396.87
	0.01	−48,689.37	−327,171.61	−15,177,293.12	5968.90	10,372.43	271.31	31,943.36	42,114.08	36,997.67

If one compares the norm of both algorithms, for small and medium MSEM problems (i.e., sizes 8 12 5 30 and 15 20 5 30), the hybrid metaheuristic performs better than the standard genetic algorithm for large values of U, in particular, $\lambda = 0.1$ and 0.01. However, for a big MSEM size i.e., size 22 28 5 30), our algorithm becomes a better estimation method than a standard genetic algorithm even for small values of U. This comparison also illustrates that whenever the hybrid technique improves 2SLS estimates, it also enhances the standard genetic algorithm. Furthermore, it is worth highlighting that when the proposed approach is presented as the best estimation option, an end condition of $MaxIter = 10$ combined with a relatively small percentage of optimization is enough to outperform a standard genetic algorithm with reasonably large number of iterations. This is remarkable since it reinforces the choice of $MaxIter$ tuning parameter from Table 1 and will contribute to alleviate the computational cost.

In summary, these experiments lead to an important conclusion. Overall, 2SLS is a more appropriate estimation method when the U values are small, that is to say, when the error terms are serially uncorrelated (or at least the serial dependence is not very strong). However, as we had postulated, our simulation experiments point out that this technique does not provide satisfactory results when this assumption is violated. As the values of U become greater, the hybrid metaheuristic arises as a preferable option to 2SLS. In that case, our algorithm works best even for a modest sample size and reasonably large models.

Expectedly, as the size of the problem increases the execution time rises and the algorithm can be computationally demanding. The cost of estimating a model is the sum of applying a genetic algorithm and an optimization method. In particular, the last optimization step described in Section 4 originates an evident time impact, as one can compare in the tables above. The time the algorithm needs before entering in the final optimization (S_{prev}) and the total running time once the whole procedure has finished (S_{post}) measured in seconds are displayed. In contrast, the time spent by the standard genetic algorithm (S_{GA}) is also indicated. The hybrid metaheuristic shows the highest execution times but it provides the best outcomes. In addition, the time that the algorithm needs depends heavily on the numerical procedure implemented in *nlm* for optimizing the likelihood of the system and on the complexity of the fitness function.

6. Conclusions and Future Work

This paper proposes a metaheuristic strategy to estimate a multilevel simultaneous equation model (MSEM) when the double covariance structure is unknown. A genetic algorithm is combined with an optimization procedure in order to obtain the maximum likelihood estimator. Different parameters have been experimentally tuned to reduce the complexity of the search space to a more manageable one. However, further simulation studies are necessary to better inspect the space of parameters and avoid falling in local optima.

The simulation experiments suggest that this hybrid metaheuristic produces better outcomes than other estimation methods such as 2SLS when error terms are not intertemporally uncorrelated. According to the results, the algorithm provides satisfactory solutions when the system of likelihood equations cannot be analytically solved and standard optimal procedures are not feasible to implement. Although the hybrid metaheuristic leads to a more efficient inference, some aspects require further analysis. Moreover, as future work extensions to non-normal error distributional forms (e.g., matrix Student's T distribution) need to be investigated.

Because of the computational burden, the simulation results reported above are limited. The execution time grows as the size of the model increases and a parallel version of the algorithm may be a preferable option. The development of a shared memory version may boost the efficiency of processor usage reducing response times in the solution of the problem. Finally, other appealing metaheuristic techniques (Scatter Search, GRASP...) need to be examined and alternative optimization solvers must be explored.

Author Contributions: Conceptualization and Methodology, R.H.-S. and J.J.L.-E.; Software, M.G.; Writing—original draft preparation and Editing, R.H.-S.; Writing—review and Supervision, J.J.L.-E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministerio de Economía y Competitividad of Spain under Grant TIN2016-8056-R and a predoctoral contract from the Generalitat Valenciana and the European Social Fund to R.H.-S. under Grant ACIF/2018/219.

Acknowledgments: The authors gratefully acknowledge the computer resources and assistance provided by the Scientific Computing and Parallel Programming Group of the University of Murcia for the simulation study. The authors would like to thank the anonymous reviewers for their constructive comments and useful suggestions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Hausman, J.A. Specification and estimation of simultaneous equation models. *Handb. Econom.* **1983**, *1*, 391–448.
2. Goldstein, H. *Multilevel Statistical Models*; John Wiley & Sons: Chichester, UK, 2011; Volume 922.
3. Klein, L.R. Economic fluctuations in the United States, 1921–1941. *Econ. J.* **1950**, *61*, 387–389.
4. Dornbusch, R.; Fischer, S. *Macroeconomics*, 3rd ed.; McGraw-Hill: New York, NY, USA, 1984.
5. King, T.M. Using simultaneous equation modeling for defining complex phenotypes. *BMC Genet. Biomed. Cent.* **2003**, *4*, S10. [CrossRef]
6. Simone, I.L.; Ceccarelli, A.; Tortorella, C.; Bellacosa, A.; Pellegrini, F.; Plasmati, I.; De Caro, M.F.; Lopez, M.; Girolamo, F.; Livrea, P. Influence of Interferon beta treatment on quality of life in multiple sclerosis patients. *Health Qual. Life Outcomes* **2006**, *4*, 96. [CrossRef] [PubMed]
7. Ressler, R.W.; Waters, M.S. Female earnings and the divorce rate: A simultaneous equations model. *Appl. Econ.* **2000**, *32*, 1889–1898. [CrossRef]
8. Graham-Bermann, S.A.; Miller-Graff, L. Community-based intervention for women exposed to intimate partner violence: A randomized control trial. *J. Fam. Psychol.* **2015**, *29*, 537. [CrossRef] [PubMed]
9. Jedidi, K.; Jagpal, H.S.; DeSarbo, W.S. Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Mark. Sci.* **1997**, *16*, 39–59. [CrossRef]
10. Steele, F.; Vignoles, A.; Jenkins, A. The effect of school resources on pupil attainment: A multilevel simultaneous equation modelling approach. *J. R. Stat. Soc. Ser. A* **2007**, *170*, 801–824. [CrossRef]
11. Steele, F. Selection effects of source of contraceptive supply in an analysis of discontinuation of contraception: Multilevel modelling when random effects are correlated with an explanatory variable. *J. R. Stat. Soc. Ser. A* **2003**, *166*, 407–423. [CrossRef]
12. Hernández-Sanjaime, R.; González, M.; López-Espín, J.J. Multilevel simultaneous equation model: A novel specification and estimation approach. *J. Comput. Appl. Math.* **2020**, *366*, 112378. [CrossRef]
13. Mitchell, M. *An Introduction to Genetic Algorithms*; MIT Press: Cambridge, MA, USA, 1998.

14. Goldberg, D.E.; Holland, J.H. Genetic algorithms and machine learning. *Mach. Learn.* **1988**, *3*, 95–99. [CrossRef]
15. López, J.J.; Giménez, D. Genetic algorithms for simultaneous equation models. *International Symposium on Distributed Computing and Artificial Intelligence 2008 (DCAI 2008)*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 215–224.
16. López-Espín, J.J.; Giménez, D. Obtaining simultaneous equation models from a set of variables through genetic algorithms. *Procedia Comput. Sci.* **2010**, *1*, 427–435. [CrossRef]
17. Almeida, F.; Giménez, D.; Lopez-Espin, J.J. Obtaining Simultaneous Equation Models through a unified shared-memory scheme of metaheuristics. In Procedings of the 2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum, Anchorage, AK, USA, 16–20 May 2011; pp. 1981–1988.
18. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2012; ISBN 3-900051-07-0.
19. Dhrymes, P.J. *Econometrics: Statistical Foundations and Applications*; Springer: New York, NY, USA, 1974.
20. Arnold, S.F. *The Theory of Linear Models and Multivariate Analysis*; Wiley: New York, NY, USA, 1981.
21. Harville, D.A. *Matrix Algebra from a Statistician's Perspective*; Springer: New York, NY, USA, 1997; Volume 1.
22. Dwyer, P.S. Some applications of matrix derivatives in multivariate analysis. *J. Am. Stat. Assoc.* **1967**, *62*, 607–625. [CrossRef]
23. Petersen, K.B.; Pedersen, M.S. The matrix cookbook. *Tech. Univ. Den.* **2008**, *7*, 510.

Bibliografía

[1] S. F. Arnold. *The theory of linear models and multivariate analysis*. Wiley, New York, NY, 1981.

[2] P. Balestra and J. Varadharajan-Krishnakumar. Full information estimations of a system of simultaneous equations with error component structure. *Econometric Theory*, 3(2):223–246, 1987.

[3] R. Blundell and F. Windmeijer. Cluster effects and simultaneity in multilevel models. *Health Economics*, 6(4):439–443, 1997.

[4] N. Damodar et al. *Basic econometrics*. The Mc-Graw Hill, 2004.

[5] J. J. L. Espın and D. Giménez. *Aspectos computacionales de la resolución y obtención de Modelos de Ecuaciones Simultáneas*. PhD thesis, PhD thesis, Universidad de Murcia, 2009.

[6] H. Goldstein. *Multilevel statistical models*, volume 922. John Wiley & Sons, 2011.

[7] W. H. Greene. *Econometric analysis*. Pearson Education India, 2003.

[8] Z. Griliches and M. D. Intriligatorrs. *Handbook of econometrics*. North Holland, 1984.

[9] J. A. Hausman. Specification and estimation of simultaneous equation models. *Handbook of econometrics*, 1:391–448, 1983.

[10] R. Hernández-Sanjaime, M. González, and J. J. López-Espín. Estimation of multilevel simultaneous equation models through genetic algorithms. *Mathematics*, 8(12):2098, 2020.

[11] R. Hernández-Sanjaime, M. González, and J. J. López-Espín. Multilevel simultaneous equation model: A novel specification and estimation approach. *Journal of Computational and Applied Mathematics*, 366:112378, 2020.

[12] R. Hernández-Sanjaime, M. González, A. Peñalver, and J. J. López-Espín. Estimating simultaneous equation models through an entropy-based incremental variational bayes learning algorithm. *Entropy*, 23(4), 2021.

[13] K. Jedidi, H. S. Jagpal, and W. S. DeSarbo. Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, 16(1):39–59, 1997.

[14] L. R. Kumar. Estimation of simultaneous econometric equations using neural networks. In *Proceedings of the Twenty-Fourth Annual Hawaii International Conference on System Sciences*, volume 4, pages 124–128. IEEE, 1991.

[15] N. Leonenko and L. Pronzato. A class of rényi information estimators for multidimensional densities. *The Annals of Statistics*, 36(5):2153–2182, 2008.

[16] J. J. López-Espín and D. Giménez. Obtaining simultaneous equation models from a set of variables through genetic algorithms. *Procedia Computer Science*, 1(1):427–435, 2010.

[17] B. O. Muthén. Latent variable modeling in heterogeneous populations. *Psychometrika*, 54(4):557–585, 1989.

[18] A. Peñalver and F. Escolano. Entropy-based incremental variational bayes learning of gaussian mixtures. *IEEE Transactions on Neural Networks and Learning Systems*, 23(3):534–540, 2012.

[19] D. S. G. Pollock and D. Pollock. The algebra of econometrics. Technical report, Wiley New York, 1979.

[20] Y. Shin and S. W. Raudenbush. The causal effect of class size on academic achievement: Multivariate instrumental variable estimators with data missing at random. *Journal of Educational and Behavioral Statistics*, 36(2):154–185, 2011.

[21] F. Steele, A. Vignoles, and A. Jenkins. The effect of school resources on pupil attainment: a multilevel simultaneous equation modelling approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(3):801–824, 2007.