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.gonzalez@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

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

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 heterogeneous 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

$$Y = YA + XB + U, \quad (1)$$

where $Y = [y^1, \dots, y^m]$ is a $N \times m$ matrix of N observations of m endogenous variables, $X = [x^1, \dots, x^k]$ is a $N \times k$ matrix of N observations of k predetermined variables and $U = [u^1, \dots, u^m]$ 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, $a_{ii} = 0, i = 1, 2, \dots, m$).

The error terms u_t . ($t = 1, \dots, N$) are assumed to be normally distributed:

$$u_{t'}' \sim N(0, \Sigma), \quad E(u_{t'}', u_{t''}') = \delta_{t't''} \Sigma \quad t, t' = 1, 2, \dots, N \quad (2)$$

where $\delta_{t't''}$ 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

$$Y = X\Pi + V, \quad (3)$$

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

The random vectors v_t . have the following properties:

$$v_{t'}' \sim N(0, \Omega), \quad v_t = u_t(I - A)^{-1}, \quad \Omega = ((I - A)')^{-1} \Sigma (I - A)^{-1}, \quad (4)$$

$$\text{Cov}(v_{t'}', v_{t''}') = \delta_{t't''} \Omega \quad t, t' = 1, 2, \dots, N$$

In SEMs, parameter estimation can only be accomplished when the system is identified. An equation i ($i = 1, \dots, m$) 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.

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

where $|\hat{\Sigma}_e|$ is the determinant of error covariance matrix e_i ($i = 1, \dots, m$) 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 $X = \{x_1, \dots, x_N\}$ 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 $i = 1, \dots, K$, 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

$$P(X, \mu, T, z|\pi) = P(X|\mu, T, z)P(z|\pi)P(\mu)P(T) \quad (6)$$

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 $Q(\Theta)$ in the log marginal likelihood expression:

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

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

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

$$Q(\Theta) = Q_z(z) Q_\mu(\mu) Q_T(T) \quad (8)$$

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 $\mathcal{L}(Q)$ can be evaluated as

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

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.

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

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

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

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

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

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 VBgm 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:

$$YA^* + XB + U = 0, \quad (13)$$

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 $k \geq m$:

$$B = QR, \quad (14)$$

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 $\text{rank}(B) = n$, so that R is nonsingular:

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

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, $X = [X_1 \ X_2]$ 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

$$\begin{aligned} -YA^*R^{-1} &= [X_1 \ X_2] \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} = X_1 Q_1 + X_2 Q_2 \\ X_1 &= -[YA^*R^{-1} + X_2 Q_2] Q_1^{-1} \end{aligned} \quad (16)$$

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]:

$$C_t = \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \alpha_3 (W_t^p + W_t^g) + \varepsilon_{1t}$$

$$I_t = \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 K_{t-1} + \varepsilon_{2t}$$

$$W_t^p = \gamma_0 + \gamma_1 X_t + \gamma_2 X_{t-1} + \gamma_3 A_t + \varepsilon_{3t}$$

$$X_t = C_t + I_t + G_t \quad (\text{equilibrium demand})$$

$$P_t = X_t - T_t - W_t^p \quad (\text{private profits})$$

$$K_t = K_{t-1} + I_t \quad (\text{capital stock})$$

where

C_t = Consumption expenditure

I_t = Investment

P_t = Private profits

K_t = Capital stock

G_t = Government nonwage spending

T_t = Indirect business taxes plus net exports

W_t^g = Government wages

W_t^p = Private-sector wages

X_t = Total demand

A_t = Time trend measured as years from 1931

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. 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

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.

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., Serrián, 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

- Hausman, J.A. Specification and estimation of simultaneous equation models. *Handb. Econom.* **1983**, *1*, 391–448.
- Klein, L.R. *Economic Fluctuations in the United States, 1921–1941*; John Wiley & Sons, Inc.: New York, NY, USA, 1950.
- Dornbusch, R.; Fischer, S. *Macroeconomics*, 3rd ed.; Little, Brown: New York, NY, USA 1984.
- King, T.M. Using simultaneous equation modeling for defining complex phenotypes. *BMC Genet. BioMed Cent.* **2003**, *4*, S10. [[CrossRef](#)] [[PubMed](#)]
- Ressler, R.W.; Waters, M.S. Female earnings and the divorce rate: A simultaneous equations model. *Appl. Econ.* **2000**, *32*, 1889–1898. [[CrossRef](#)]
- Andrews, D.W. Tests for parameter instability and structural change with unknown change point. *Econom. J. Econom. Soc.* **1993**, *61*, 821–856. [[CrossRef](#)]
- Bai, J.; Perron, P. Estimating and testing linear models with multiple structural changes. *Econometrica* **1998**, *66*, 47–78. [[CrossRef](#)]
- Yao, Y.C. Estimating the number of change-points via Schwarz’ criterion. *Stat. Probab. Lett.* **1988**, *6*, 181–189. [[CrossRef](#)]
- Liu, J.; Wu, S.; Zidek, J.V. On segmented multivariate regression. *Stat. Sin.* **1997**, *7*, 497–525.
- Ninomiya, Y. Information criterion for Gaussian change-point model. *Stat. Probab. Lett.* **2005**, *72*, 237–247. [[CrossRef](#)]
- Hall, A.R.; Osborn, D.R.; Sakkas, N. Inference on structural breaks using information criteria. *Manch. Sch.* **2013**, *81*, 54–81. [[CrossRef](#)]
- McZgee, V.E.; Carleton, W.T. Piecewise regression. *J. Am. Stat. Assoc.* **1970**, *65*, 1109–1124. [[CrossRef](#)]
- Dunicz, B. Discontinuities in the surface structure of alcohol-water mixtures. *Kolloid-Z. Z. Polymere* **1969**, *230*, 346–357. [[CrossRef](#)]
- Sprent, P. Some hypotheses concerning two phase regression lines. *Biometrics* **1961**, *17*, 634–645. [[CrossRef](#)]
- 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. [[CrossRef](#)]
- Muthén, B.O. Latent variable modeling in heterogeneous populations. *Psychometrika* **1989**, *54*, 557–585. [[CrossRef](#)]
- Kohli, A.K. Effects of supervisory behavior: The role of individual differences among salespeople. *J. Mark.* **1989**, *53*, 40–50. [[CrossRef](#)]
- Day, R.L. Extending the concept of consumer satisfaction. *ACR N. Am. Adv.* **1977**, *4*, 149–154.
- 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](#)]
- Jöreskog, K.G. Simultaneous factor analysis in several populations. *Psychometrika* **1971**, *36*, 409–426. [[CrossRef](#)]
- 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. [[CrossRef](#)]
- Jedidi, K.; Jagpal, H.S.; DeSarbo, W.S. STEMM: A general finite mixture structural equation model. *J. Classif.* **1997**, *14*, 23–50. [[CrossRef](#)]
- 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. [[CrossRef](#)]
- Aitkin, M.; Anderson, D.; Hinde, J. Statistical modelling of data on teaching styles. *J. R. Stat. Soc. Ser. A Gen.* **1981**, *144*, 419–448. [[CrossRef](#)]
- McLachlan, G.J.; Basford, K.E. Mixture models. Inference and applications to clustering. In *Statistics: Textbooks and Monographs*; Dekker: New York, NY, USA, 1988.
- 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. [[CrossRef](#)] [[PubMed](#)]
- McLachlan, G. *Discriminant Analysis and Statistical Pattern Recognition*; John Wiley and Sons: Hoboken, NJ, USA, 2004; Volume 544.
- Chang, W.C. On using principal components before separating a mixture of two multivariate normal distributions. *Appl. Stat.* **1983**, *32*, 267–275. [[CrossRef](#)]
- Damodar, N. *Basic Econometrics*; The Mc-Graw Hill: New York, NY, USA, 2004.
- Dhrymes, P.J. *Econometrics: Statistical Foundations and Applications*; Springer: New York, NY, USA, 1974.
- Hurvich, C.M.; Tsai, C.L. Regression and time series model selection in small samples. *Biometrika* **1989**, *76*, 297–307. [[CrossRef](#)]
- Gorobets, A.A. *The Optimal Prediction Simultaneous Equations Selection*; Technical Report; 2004. Available online: <https://repub.eur.nl/pub/1839> (accessed on 24 March 2021).
- Jain, A.; Dubes, R.; Mao, J. Statistical Pattern Recognition: A Review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 4–38. [[CrossRef](#)]

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. [[CrossRef](#)] [[PubMed](#)]
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. [[CrossRef](#)] [[PubMed](#)]
38. Peñalver, A.; Escolano, F. Entropy-Based Incremental Variational Bayes Learning of Gaussian Mixtures. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *23*, 534–540. [[CrossRef](#)]
39. Corduneanu, 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. [[CrossRef](#)] [[PubMed](#)]
42. Leonenko, N.; Pronzato, L. A class of rényi information estimators for multi-dimensional densities. *Ann. Stat.* **2008**, *36*, 2153–2182. [[CrossRef](#)]
43. Constantinopoulos, C.; Likas, A. Unsupervised Learning of Gaussian Mixtures Based on Variational Component Splitting. *IEEE Trans. Neural Netw.* **2007**, *18*, 745–755. [[CrossRef](#)] [[PubMed](#)]
44. Dellaportas, P.; Papageorgiou, I. Multivariate mixtures of normals with unknown number of components. *Stat. Comput.* **2006**, *16*, 57–68. [[CrossRef](#)]
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.; Serinán, 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*. Available online: https://www.researchgate.net/profile/Basilio-Sanz/publication/24121320_El_Modelo_Klein_I_y_los_Ciclos_Economicos/links/02e7e52bc9ddd79aa2000000/El-Modelo-Klein-I-y-los-Ciclos-Economicos.pdf (accessed on 24 March 2021).