Structural equation modeling with time dependence: an application comparing Brazilian energy distributors

Vinícius Diniz Mayrink[1], **Renato Valladares Panaro**[1] **and Marcelo Azevedo Costa**[2]

[1] Departamento de Estatística, ICEx, Universidade Federal de Minas Gerais.

[2] Departmento de Engenharia de Produção, Escola de Engenharia, Universidade Federal de Minas Gerais.

Abstract

This study proposes a Bayesian structural equation model (SEM) to explore financial and economic sustainability indicators, considered by the Brazilian energy regulator (ANEEL), to evaluate the performance of energy distribution companies. The methodology applies confirmatory factor analysis for dimension reduction of the original multivariate data set into few representative latent variables (factors). In addition, a regression structure is defined to establish the impact of the factors over the response "indebtedness" of the companies; this is a central aspect regularly discussed within ANEEL to identify whether a distributor may have difficulty to manage the concession. Most of the variables in this study are collected for 8 different years (2011-2018), therefore, a time dependence is inserted in the analysis to correlate observations. The SEM approach has several advantages in this context: it avoids using criticizable deterministic formulations to measure non-observable aspects of the distributors, it allows a broad statistical analysis exploring elements that cannot be investigated through the simple descriptive studies currently developed by the regulator and, finally, it provides tools to properly rank and compare distances between companies. The Bayesian view is a powerfull option to handle the SEM fit here, since convergence issues, due to sample size and high dimensionality, may be experienced via classical alternatives based on maximization.

keywords: Confirmatory factor analysis; Regression; Multivariate analysis; NUTS.

1 Introduction

Methodologies for multivariate statistical studies are, in general, designed to handle situations where the analyst has a data set containing multiple variables with different structures of associations among them. Dimension reduction is usually an interesting strategy for the analysis, in this case, fewer non-observable (latent) variables are created to summarize the main information in the whole data set. Two important methods to reach this goal are: the principal component analysis (PCA) and the factor analysis (FA); see Johnson and Wichern (2007) for details. In particular, the FA can be used in two versions: confirmatory (CFA) and exploratory (EFA). The first case is usually employed when the researcher imposes a relationship between a group of variables and a target latent factor to be constructed; see Brown (2015). The notion of which variables should be connected to each factor is a pre-requisite for the method. In constrast, the EFA is considered when the user does not know how to segregate the variables, indicating which ones are important to explain each unobserved factor.

The structural equation model (Skrondal and Rabe-Hesketh, 2004; Hoyle, 2014; Keith, 2019), hereafter denoted by SEM, is a technique combining CFA and multiple regression to explore the existing interrelated dependence between measured variables and latent constructs. In other words, a SEM can be fully specified in two parts: (i) the measurement model which associates the measured variables to latent factors and (ii) the structural model which defines associations of the type "factors explaining factors" and "factors explaining observed responses". For an overview of SEMs, the reader should consider Sanchez et al. (2005); this reference is focused on applications in environmental epidemiology. The methodology is also widely used in Psychology to deal with features that cannot be directly quantified through instruments, such as: intelligence or happiness. The references Anderson and Gerbing (1988) and MacCallum and Austin (2000) provide a review about the use of SEM in psychology researches.

Address for correspondence: Vinícius D. Mayrink, Departamento de Estatística, ICEx, Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Belo Horizonte, MG, Brazil, 31270-901. E-mail: vdm@est.ufmg.br

Computers advances in the past decades have determined an increase in the number of applications using SEMs to deal with large datasets in complex and unstructured problems. In order to allow inferences here, the Bayesian framework has some advantages over the classical maximum likelihood (ML) estimation; see Palomo et al. (2007) for an overview of Bayesian SEMs. The first advantage is the exact estimation of posterior distributions, for the many SEM parameters, via Markov Chain Monte Carlo (MCMC) algorithms (Gamerman and Lopes, 2006). This contrasts with the ML case, where the outputs are point estimates with asymptotic standard errors. The second advantage is the fact that asymptotic assumptions in ML may not be valid depending on the sample size; Lee and Song (2003) is an interesting reference showing the better behavior of the Bayesian SEM compared to the ML case in small samples. One last advantage is the possibility to incorporate prior knowledge, which brings flexibility and configures an element to handle the necessary constraints to ensure model identifiability (this is a key issue in SEM).

When working with multidimensional longitudinal data, it is obviously important that one should account for the within-subject dependence due to repeated measurements. In this case, joint models accounting for dependency within and across time points are appealing. A variety of frameworks for this purpose can be found in Diggle et al. (2002). A commom solution is to consider a random effects model to establish a connection between the observations from the same subject. Random effects models can be seen as part of a broader class of latent variable models that may be considered to treat longitudinal data; Dunson (2007) provides an overview. As an alternative to the random effects option, the within-subject dependence may also be imposed through a temporal structure that correlates the nearest time points. This is a natural strategy in dynamic linear modeling for time series data (West and Harrison, 1997), which can also be seen as a spatial model with regions regularly disposed in a line (Mayrink and Gamerman, 2009). The spatio-temporal FA approach is dicussed in Gamerman et al. (2008) and Lopes et al. (2011), where temporal and spatial associations are defined for the scores and loadings matrices, respectively. A SEM framework assuming this type of dependence structure, either in the score or loadings matrix, is not a topic fully explored in the literature.

The main goal of the present paper is to develop a SEM capable of investigating hidden relationships in a data set of financial and economic indicators (F-E indicators) reported from companies of energy distribution regulated by the *Agência Nacional de Energia Elétrica* (ANEEL) in Brazil. According to ANEEL, the F-E indicators have a fundamental importance in the task of supervising the energy companies. In brief, they provide key information to evaluate the performance of the distributors. This supervision is intended to prevent the degradation of the service and to identify possible issues in the national energy distribution administration. The F-E indicators are monitored every 3 months and data bases from different years are freely available in the regulator homepage; see `www.aneel.gov.br/informacoes-tecnicas`. In the present paper, we consider observations in a 8-year period (2011–2018). Currently, the ANEEL quartely evaluations are guided by the so called "Technical Note Number 111/2016" approved in 2016. This document defines 20 variables (indicators) that can be used to explain 6 latent dimensions for each company: indebtedness, efficiency, investments, profitability, shareholder return and operational features. Some attention is devoted to the dimension indebtedness, which is closely related to the sustainability of the distributor in terms of services and activities. The regulator is particularly concerned with the scenario where a distributor tends to experience, in the medium term, difficulties to manage the concession.

In line with the previous discussions, the main contributions of the present work are listed as follows:

- Propose a SEM to build the mentioned 6 latent dimensions (factors) and then explain, in a second level of the hierarchical model, the indebtedness of the energy companies. Due to the multiple year data collection, time dependence is inserted in the model to handle the similarities between near time points. Our Bayesian SEM, with time dependence imposed for groups of loadings, can be seen as a new approach in the related literature. Some references addressing longitudinal data with other approaches involving the SEM context are: Frese et al. (2007), Song et al. (2008), Barbieri et al. (2018), Seddig and Leitgob (2018) and Usami et al. (2019).
- Develop a comprehensive simulation study, based on artificial data, to show that the proposed framework behaves well.
- Discuss and compare the results from the real data application involving the ANEEL F-E indicators. We highlight the fact that this data set is novel and has never been investigated elsewhere in the context of SEM.

This work is organized as follows. Section 2 provides further details about the real data set to be used in our illustration. Section 3 presents the SEM model with time dependence. Section 4 shows a simulation study with artificial data to evaluate the performance of the proposed model. Section 5 develops the main application based on the ANEEL data set. Finally, Section 6 indicates the main conclusions and final remarks.

2 The data

The ANEEL data set to be used in our real analysis contains information of 20 different F-E indicators; 17 of them observed for 8 different years and 3 of them having a single observation for the whole period. Most of these indicators are measurements based on monetary amounts in the currency Brazilian Real; exceptions: sectors in constitution (Setoriais-C), shareholder flow (Fluxo-A) and number of consumers (N-Consum.). Table 1 shows details regarding each indicator; the symbol $*$ identifies the cases without replications across the years. These indicators are collected for 53 Brazilian energy distribution companies with concession covering different parts of the country's territory; overlaps do not occur between areas.

Table 1: Identification of the 20 F-E indicators collected by the Brazilian regulator (ANEEL). Acronyms are exhibited in the second column as a reference to the notation reported in the Technical Note 111/2016 (`https://www.aneel.gov.br/informacoes-tecnicas`). All indicators are observed for 8 different years (2011–2018), except the ones with acronyms marked with the symbol * ; these cases have a single observation for each company in the whole period.

Indicator	Acronym	Meaning
1	DLR	Net debt with financial assets. and liabilities.
2	EBITDA-A	Adjusted earns before interest, taxes, depreciation and amortization.
3	QRR	Regulatory reintegration quota.
4	VPB	Value of the tariff (portion B).
5	PMSO-A	Adjusted expenses with personnel, materials, third party services and others.
6	PMSO-R	Regulatory expenses with personnel, materials, third party services and others.
7	CAPEX-U4/5A*	Capital expenditures, sum of the last 4 or 5 years (tariff cicle size).
8	QRR-U4/5A*	Regulatory reintegration quota, sum of the last 4 or 5 years (tariff cicle size).
9	EBIT-A	Adjusted earns before interest and taxes.
10	EBIT-R	Regulatory earns before interest and taxes.
11	BRL	Net remuneration basis.
12	Setoriais-C*	Sectors in constitution.
13	Fluxo-A	Shareholder flow.
14	BRLK-P	Net remuneration basis with owned capital.
15	Resultado-Liq.	Net profit.
16	DGC	Global continuity performance.
17	Perdas-Rea.	Real losses.
18	Perdas-Reg.	Regulatory losses.
19	Mercado-Gwh	Gwh market.
20	N-Consum.	Number of consumers.

The current guidelines defined by the Brazilian regulator, in their Technical Note 11/2016, establish some formulations intended to measure 6 different dimensions of interest. The expressions are presented in Table 2. Note that some of them are ratios, other cases are differences or the value from a single F-E indicator. An important point to be highlighted is the definition of 2 or even 4 formulations to represent some of the dimensions. This multiplicity of elements complicates the analysis, forcing the analyst to choose or weight the different expressions to summarize the dimension. This aspect motivates the application of a CFA approach that uses groups of F-E indicators to build the target dimensions; these dimensions are to be treated here as

latent factors. Using the connections represented in Table 2, one can easily appoint the groups of observable F-E indicators that might be relevant to explain each factor. In line with this idea, consider the associations in Table 3. A superscript sign mark is inserted in the acronym identification to emphasize which indicators are expected to have a positive or negative correlation with the latent factor under construction; indicators in numerators or in the left hand side of subtractions are the ones with expected positive association. Table 3 also shows a simpler notation, $X_1, \cdots, X_{19}$, representing each F-E indicator.

Table 2: Formulations, defined by ANEEL in the Technical Note 111/2016, to measure 6 dimensions of interest based on the F-E indicators. The first dimension "indebtedness" is critical for the analysis, since the regulator uses this information to evaluate the companies in terms of sustainability. Table 1 shows a glossary of the indicators. Consider for simplicity: DIF1 = EBITDA-A − QRR, DIF2 = EBIT-A − EBIT-R and DIF3 = Perdas-Rea. − Perdas-Reg.

	Dimension	Formulations			
1	Indebtedness	$\frac{\text{DLR}}{\text{DIF1}}$			
2	Efficiency	$\frac{\text{EBITDA-A}}{\text{VPB}}$	$\frac{\text{PMSO-A}}{\text{PMSO-R}} - 1$		
3	Investments	$\frac{\text{CAPEX-U4/5A}}{\text{QRR-U4/5A}} - 1$			
4	Profitability	$\frac{\text{DIF2}}{\text{BRL}} - 1$	$\frac{\text{Setoriais-C}}{\text{EBIT-R}} - 1$		
5	Shareholder return	$\frac{\text{Fluxo-A}}{\text{BRLK-P}}$	Resultado-Liq.		
6	Operational features	DGC	DIF3	Mercado-Gwh	N-Consum.

Table 3: Summary of F-E indicators connected to each dimension. Consider for simplicity: DIF1 = EBITDA-A − QRR, DIF2 = EBIT-A − EBIT-R and DIF3 = Perdas-Rea. − Perdas-Reg. Indicators marked with $(-)$ are found in denominators or in the right hand side of subtractions (they are expected to have a negative correlation with the corresponding latent dimension). Indicators marked with $(+)$ are found in numerators or in the left hand side of subtractions (they are expected to have a positive correlation with the target dimension).

Dimension	Associated F-E indicators			
1 Indebtedness	$X_1 = \text{DLR}^{(+)}$	$X_2 = \text{DIF1}^{(-)}$		
2 Efficiency	$X_3 = \text{EBITDA-A}^{(+)}$	$X_4 = \text{VPB}^{(-)}$	$X_5 = \text{PMSO-A}^{(+)}$	$X_6 = \text{PMSO-R}^{(-)}$
3 Investments	$X_7 = \text{CAPEX-U4/5A}^{(+)}$	$X_8 = \text{QRR-U4/5A}^{(-)}$		
4 Profitability	$X_9 = \text{DIF2}^{(+)}$	$X_{10} = \text{BRL}^{(-)}$	$X_{11} = \text{Setoriais-C}^{(+)}$	$X_{12} = \text{EBIT-R}^{(-)}$
5 Shareholder return	$X_{13} = \text{Fluxo-A}^{(+)}$	$X_{14} = \text{BRLK-P}^{(-)}$	$X_{15} = \text{Resultado-Liq.}^{(+)}$	
6 Operational features	$X_{16} = \text{DGC}^{(+)}$	$X_{17} = \text{DIF3}^{(+)}$	$X_{18} = \text{Mercado-Gwh}^{(+)}$	$X_{19} = \text{N-Consum.}^{(+)}$

The summary provided in Table 3 contains the technical connections to be used in the CFA analysis within the SEM model being proposed in this paper. We again reinforce to the reader the fact that all indicators, except X_7, X_8 and X_{11}, are observed with replications for 8 different years. The Bayesian SEM detailed in the next section has a temporal structure to associate values from near time points. Given that ANEEL is particularly interested in supervising the sustainability of the energy companies, with focus on the dimension indebtedness calculated via DLR/DIF (see Table 2), we choose to include this quantity in our analysis as a response to be explained by the latent factors. This will be discussed ahead in the model description.

Missing values (NAs) can be found among the F-E indicators. The percentage of missing values in the whole data set is 33.09%. The variable with the largest number of NAs is the number of consumers (indicator X_{19}, year 2011), having 7 cases out of 53. Dealing with missing data under the Bayesian inference is relatively simple. These are unknown quantities for which we can set a prior to be updated into a posterior distribution through the Bayes rule. In the present work, we choose to follow this path of NA estimation, and thus avoid the much criticized strategies of removing the whole variable (wasting information) or imputing a fixed number (e.g. mean) to maintain the observed part of the variable.

Pre-processing the data before fitting a factor model is a stantard step in many applications. This is particularly important in studies related to gene expression data involving measurements of how active are the genes in different samples for the same type of tumor in cancer research. The equipment designed to measure the gene activity may not be well adjusted when scanning some of the sample units. In this case, the final data set is composed by unbalanced samples, i.e., some tumors were not properly measured. See Mayrink and Lucas (2015) and references therein for further details related to pre-processing gene expression data in FA. The mentioned distortion among the sample units has a strong impact on the FA results. The problematic samples may create a pattern in the data that will mislead the final conclusions. This concern is clearly related to the ANEEL data set due to the presence of energy companies with distinct sizes and financial strengths. Most of the indicators in Table 3 are highly correlated with the size of the company. In order to illustrate this "company size" effect, we compute the scores of the first principal component (pc_1) in a PCA analysis (Johnson and Wichern, 2007) involving all 131 variables (16 indicators with 8 replications = years and 3 indicators being unique for the whole period 2011-2018). In brief, the pc_1 is a new variable (size 53) that best summarizes the existing global pattern shared by all 131 variables (combinations "F-E indicator and time" across 53 companies). Figure 1 shows the magnitude of the pc_1 scores for each energy company. As it can be seen, some high scores (say $> 50{,}000$) are obtained and a data inspection shows that these high scores are related to large companies with great financial strengh. The cases ENEL-SP, CEMIG and Light are recognized as large energy companies in Brazil and their scores are amog the highest in the graph. On the other hand, the AmE (with concession in the Amazon region) is considered a small-sized company and its pc_1 score is located at the bottom of the graph.

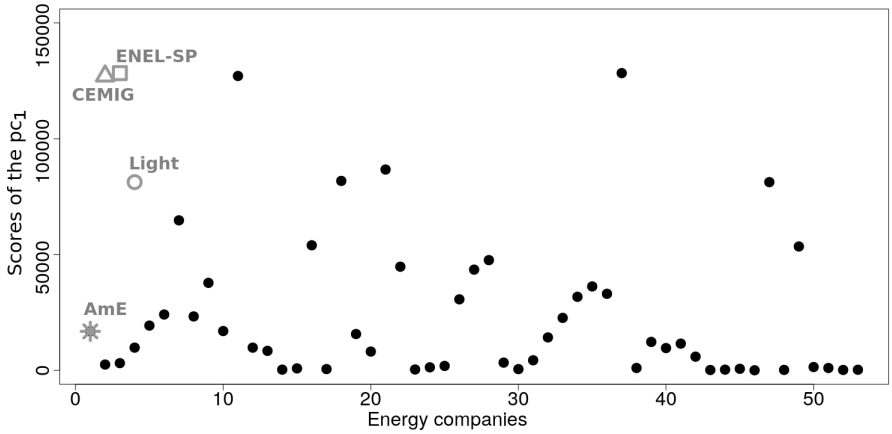

Figure 1: Magnitude of the pc_1 scores summarizing the existing global pattern related to the 131 variables in the real data set. Each point in the graph corresponds to an energy company. The cases AmE (Amazon), CEMIG (Minas Gerais), ENEL-SP (São Paulo) and Light (Rio de Janeiro) are highlighted in the graph. PCA is applied to the raw measurements without transformations.

The effect "company size" must be removed from the data before fitting the CFA model. In order to reach this goal, first denote $X^{s_1}_{[it]j}$ as the observed value for the i-th F-E indicator ($i = 1, \cdots, 19$), the t-th year ($t = 1, \cdots, 8$) and the j-th energy company ($j = 1, \cdots, 53$). Remark: if $i = 7, 8$ or 11, set $t = 1$ due to the non-longitudinal configuration of these indicators. The superscript s_1 denotes that the data set is in a "stage 1", i.e., the original version without modifications. Further, assume that $X^{s_1}_{\bullet j}$ is a column vector containing all 131 values related to the j-th company. We choose to apply a simple first pre-processing procedure where the elements in $X^{s_1}_{\bullet j}$ are standardized by subtracting the sample mean and dividing by the sample standard deviation of the vector. This transformation is intended to balance the distinctions "between companies".

Let $X^{s_2}_{[it]\bullet} = (X^{s_2}_{[it]1}, X^{s_2}_{[it]2}, \cdots, X^{s_2}_{[it]53})$ be the row vector containing values of the variable $[it]$ for all j. The superscript s_2 identifies the data in a "stage 2", where the observations were standardized as previously described. The second pre-processing step is the standardization of $X^{s_2}_{[it]\bullet}$. This is critical in FA, otherwise the variables $[it]$ with large variances will dominate the results and hide important conclusions. This transformation in the second step is supposed to treat the differences "between variables".

We emphasize that if the data set contains missing values, the adopted strategy is to replace those elements by the mean of the corresponding variable $X^{s_1}_{[it]\bullet}$. This is done with the data in "stage 1" and

before applying the first standardization step previously mentioned.

3 SEM with time dependence

Let X be a $N_v \times N_e$ matrix of pre-processed data incorporating observed variables in the rows and sample elements in the columns. In the context of the ANEEL data set, $N_v = 131$ is the number of variables (combination "F-E indicator and time") and $N_e = 53$ is the number of energy companies. In addition, suppose that K is the total number of latent factors in the model; this quantity is usually fixed by the analyst when fitting a CFA.

Now assume that the rows of X can be partitioned in N_b blocks such that each block contains the replications of one variable for different years. In this sense, consider $X_1, \cdots, X_{N_b}$ to represent the blocks, with X_i being a $T_i \times N_e$ matrix, where T_i is the number of years collected for the i-th indicator. Note that, the model under construction admits that the blocks in X may not contain the same number of rows (time points). This means that the longitudinal study can have an unbalanced configuration for the number of years. According to the described structure, we also denote $X_{[it]\bullet} = (X_{[it]1}, X_{[it]2}, \cdots, X_{[it]N_e})$ as the row vector with observations of the variable $[it]$ (F-E indicator i and time t), which is one row of matrix X. The CFA model, seen as the first part of the SEM, has the following representation:

$$X \;=\; \alpha\,\lambda \;+\; \epsilon, \tag{1}$$

where α is the factor loadings matrix with dimension $N_v \times K$, λ is the factor scores matrix of size $K \times N_e$, each row of λ is a latent factor, and ϵ is a $N_v \times N_e$ matrix of independent errors. A common assumption for this type of model is $\epsilon_{[it]j} \sim N(0, \sigma^2_{[it]})$. In other words, the variability may differ from row to row in X.

Any missing observation in X is treated as an unknown quantity to be estimated in the Bayesian analysis. Let $\alpha_{[it]\bullet}$ be the row of α corresponding to the variable $[it]$ and $\lambda_{\bullet j}$ is the j-th column of λ. Suppose that $X_{[it]j}$ is an NA case related to the variable indexed by $[it]$ and the sample unit j. We use the structure in (1) to write the prior for the missing value as follows: $X_{[it]j} \sim N(\alpha_{[it]\bullet}\lambda_{\bullet j}, \sigma^2_{[it]})$.

The confirmatory aspect of this FA is imposed to the model through prior specifications related to the elements of α. The practitioner is supposed to use technical information about the application to define K disjoint groups of variables that better explain each latent factor to be estimated. The loadings in α measure the association between variables in the rows of X and factors in the rows of λ. A loading $\alpha_{[it]k} \approx 0$ suggests that the variable $X_{[it]\bullet}$ does not explain the k-th factor (notation: $\lambda_{k\bullet}$ is the k-th row of λ). On the other hand, a non-null $\alpha_{[it]k}$ indicates a significant relationship between variable $X_{[it]\bullet}$ and factor $\lambda_{k\bullet}$. The sign of $\alpha_{[it]k}$ is important to interpret this relationship, i.e., a positive loading means positive correlation between $X_{[it]\bullet}$ and $\lambda_{k\bullet}$ and a negative loading means negative correlation.

The mentioned disjoint groups of variables are represented by $G_1, G_2, \cdots, G_K$. Define $G_k = \{\text{indices } [it] : X_{[it]\bullet} \text{ is related to } \lambda_{k\bullet}\}$. We suppose $\alpha_{[it]k} \neq 0$ for all $[it] \in G_k$ and $\alpha_{[it]k} \approx 0$ for all $[it] \notin G_k$, with $k \in \{1, \cdots, K\}$. The reader must note that it is reasonable to include the whole block X_i (all years), for $i = 1, \cdots, N_b$, within the same group G_k. In other words, it does not make sense to assume that $X_{[i1]\bullet}$ (year 1) is connected to the k-th factor and $X_{[i2]\bullet}$ (year 2) is connected to another factor. Given this restriction, one can see that the groups $G_1, \cdots, G_K$ are in fact a partition of the list of blocks $X_1, \cdots, X_{N_b}$. This partition is supposed to be chosen by the researcher.

In line with the previous discussion, denote $\alpha_{[i\bullet]k} = (\alpha_{[i1]k}, \alpha_{[i2]k}, \cdots, \alpha_{[iT_i]k})^\top$ as the vector of loadings associated to the block X_i such that $[it] \in G_k$. We set:

$$\alpha_{[i\bullet]k} \sim N_{T_i}(\mathbf{0}, \nu_{1i}\Sigma_i), \tag{2}$$

where $\mathbf{0}$ is a vector $(T_i \times 1)$ of zeros, $\nu_{1i} > 0$ is a scalar fixed by the analyst, $\Sigma_i = (D_i - \rho W_i)^{-1}$ is a matrix of temporal dependence (size $T_i \times T_i$) with $D_i = \text{diag}\{1, 2, \cdots, 2, 1\}$ (number of neighbors in the diagonal), ρ is a scalar chosen so that $D_i - \rho W_i$ can be inverted, W_i is a $T_i \times T_i$ binary band diagonal matrix such that: $W_{ir,c} = 1$ (row r, column c) when $|r - c| = 1$ (0 otherwise). This structure is basically a CAR model (Besag, 1974; Besag et al., 1991; Banerjee et al., 2014), well known in spatial statistics, inserted here to treat the temporal structure. The strategy of fixing ν_{1i} is important to ensure that this parameter is not too close to zero, which in turn would lead to an unwanted weak relationship between the variables and the target factor. Note that, if $T_i = 1$, Σ_i becomes the scalar 1 in (2).

The prior specification for the remaining loadings in α is set to ensure a null connection between the corresponding variables and factors. In other words, we assume independence *a priori* and admit $\alpha_{[it]k} \sim N(0, \nu_2)$, with $\nu_2 > 0$ fixed and small (say 0.0001), for all indices $[it] \notin G_k$. This completes the full specification of priors for α in (1).

The next stage of the model description is related to the choice of priors for the factor scores in λ. This is a key aspect for the SEM model, since one or more regression equations are built to relate "factors with other factors" or "factors with an observed response"; denote the observed response by $Y = (Y_1, Y_2, \cdots, Y_{N_e})^\top$. The number of regression equations and the choice of explanatory/response variables vary depending on the target application. For the sake of simplicity, we discuss here two types of regression settings that may appear in a SEM model. These are particular cases, but they encompass the most important features to be evaluated in the methodology proposed in this paper.

First regression setting: It is related to the ANEEL illustration motivating this paper. In this context, a single regression equation is defined using the K factors to explain Y. Under this version, we are free to assume *a priori* that $\lambda_{kj} \sim N(0,1)$, independently for all $k \in \{1, 2, \cdots, K\}$ and $j \in \{1, 2, \cdots, N_e\}$. This choice for the factor scores is standard in FA and it can be seen as a strategy to avoid identifiability issues due to the product $\alpha\lambda$ in (1); note that one can have the same likelihood by increasing α and descreasing λ accordingly (or vice versa). The regression equation in the SEM model is given by:

$$Y_j \;=\; \beta_0 + \beta^\top \lambda_{\bullet j} + \varepsilon_j \quad \text{with} \quad \varepsilon_j \sim N(0, \tau). \tag{3}$$

Consider $\beta = (\beta_1, \cdots, \beta_K)^\top$. The prior uncertainty about the regression coefficients is specified through the distributions: $\beta_0 \sim N(m_0, v_0)$ and $\beta \sim N_K(m_\beta, S_\beta)$. The hyperparameters $m_0 \in \mathbb{R}$ (mean) and $S_0 \in \mathbb{R}^+$ (variance) are scalars. In addition, m_β is a $K \times 1$ mean vector and S_β is a $K \times K$ covariance matrix. The hyperparameters $\{m_0, v_0, m_\beta, S_\beta\}$ are supposed to be chosen by the analyst.

Second regression setting: Without loss of generality, let $\Omega_1 = \{k_{11}, k_{12}, \cdots, k_{1q_1}\}$ and $\Omega_2 = \{k_{21}, k_{22}, \cdots, k_{2q_2}\}$ represent two subsets of indices contained by the larger set $\Omega_0 = \{1, 2, \cdots, K\}$. These subsets may have an intersection, but they are not equal. In addition, assume that $\Omega_0 \setminus (\Omega_1 \cup \Omega_2) = \{1, 2\}$. Three regression equations are defined:

$$\lambda_{1j} \;=\; \beta_1^{r_1}\lambda_{k_{11}j} + \beta_2^{r_1}\lambda_{k_{12}j} + \cdots + \beta_{q_1}^{r_1}\lambda_{k_{1q_1}j} + \varepsilon_j^{r_1} \quad \text{with} \quad \varepsilon_j^{r_1} \sim N(0,1), \tag{4}$$

$$\lambda_{2j} \;=\; \beta_1^{r_2}\lambda_{k_{21}j} + \beta_2^{r_2}\lambda_{k_{22}j} + \cdots + \beta_{q_2}^{r_2}\lambda_{k_{2q_2}j} + \varepsilon_j^{r_2} \quad \text{with} \quad \varepsilon_j^{r_2} \sim N(0,1), \tag{5}$$

$$Y_j \;=\; \beta_0^{r_3} + \beta_1^{r_3}\lambda_{1j} + \beta_2^{r_3}\lambda_{2j} + \varepsilon_j^{r_3} \quad \text{with} \quad \varepsilon_j^{r_3} \sim N(0, \tau). \tag{6}$$

The superscripts (r_1, r_2 or r_3) in the coefficients and error terms indicate their corresponding regression equation. Denote: $\beta^{r_1} = (\beta_1^{r_1}, \beta_2^{r_1}, \cdots, \beta_{q_1}^{r_1})^\top$, $\beta^{r_2} = (\beta_1^{r_2}, \beta_2^{r_2}, \cdots, \beta_{q_2}^{r_2})^\top$ and $\beta^{r_3} = (\beta_0^{r_3}, \beta_1^{r_3}, \beta_2^{r_3})^\top$. The prior distributions adopted for these vectors are: $\beta^{r_1} \sim N_{q_1}(m_1, S_1)$, $\beta^{r_2} \sim N_{q_2}(m_2, S_2)$ and $\beta^{r_3} \sim N_3(m_3, S_3)$, where m_1, m_2 and m_3 are mean vectors and S_1, S_2 and S_3 are covariance matrices to be chosen by the analyst. Using (4) and (5), one can easily identify normal distributions as prior specifications for λ_{1j} and λ_{2j}; the regression linear predictor is the mean and the variance is 1. We further assume that $\lambda_{kj} \sim N(0,1)$, independently for all $k \in (\Omega_1 \cup \Omega_2)$. Note that two restrictions are imposed in (4) and (5) to comply with assumptions to avoid identifiability issues in the FA. The first one is the absence of an intercept, which is required to prevent λ_{1j} and λ_{2j} to have great deviations from the center zero. The second restriction is the fixed variance = 1 for the error terms $\varepsilon_j^{r_1}$ and $\varepsilon_j^{r_2}$. Assuming this variance as unknown will weaken the magnitude constraint for λ_1 and λ_2, which in turn will cause estimation problems due to the product $\alpha\lambda$ in (1). In contrast with the situation in which a latent factor is treated as the response variable, the regression (6) has an observed response, therefore, the presence of an interpect and the estimation of the variance τ do not imply in identifiability issues. We emphasize again that the regression setting dicussed here is a particular case where the factors 1 and 2 are explained by other factors and Y_j is explained by the factors 1 and 2; however, this is not mandatory and other configurations can be investigated.

The final prior specifications, completing the model description, are based on Inverse Gamma (IG) distributions to represent the initial information about variance parameters. Consider:

$$\sigma^2_{[it]} \sim IG(a_{\sigma^2}, b_{\sigma^2}) \quad \text{and} \quad \tau \sim IG(a_\tau, b_\tau), \tag{7}$$

where $a_{\sigma^2} > 0$, $b_{\sigma^2} > 0$, $a_\tau > 0$ and $b_\tau > 0$ are fixed shape and scale hyperparameters to be chosen by the researcher. Although $\sigma^2_{[it]}$ may differ for each variable indexed by $[it]$, the same prior is set for all cases as an additional simplification step. In terms of notation, assume that σ^2 is a column vector containing all $\sigma^2_{[it]}$, for $i = 1, \cdots, N_b$ and $t = 1, \cdots, T_i$.

Given the high dimensionality and complexity of the joint posterior distribution related to the SEM model proposed in this section, we take advantage of an MCMC algorithm to allow indirect sampling. The chosen algorithm, known as No-U-Turn Sampling or NUTS (Hoffman and Gelman, 2014), is based on the Hamiltonian dynamics and it can be seen as an extension of the so called Hamiltonian Monte Carlo method. The NUTS can be easily implemented through the platform `Stan` (`mc-stan.org`), which has an interface with different programming languages, including the software `R` (R Core Team, 2019) with the package `rstan` (Stan Development Team, 2019).

4 Simulation study with artificial data

In this section, we develop a simulation study to evaluate the performance of the proposed SEM assuming the regression setting defined in (4), (5) and (6). First, the analysis is focused on a single artificial data set. Later, a Monte Carlo (MC) structure with 100 replications (data sets generated under the same conditions) is applied for a broader investigation of the proposed SEM.

Consider $K = 5$ factors, $\Omega_1 = \{3, 4\}$, $\Omega_2 = \{4, 5\}$; naturally, we have $\Omega_0 = \{1, 2, 3, 4, 5\}$. In addition, admit that the data matrix X is 100×50, meaning that $N_v = 100$ variables and $N_e = 50$ sample units. We also assume the existence of 5 missing values randomly positioned in X. The procedure to generate the data is as follows:

1. The $N_v = 100$ rows of X are partitioned in 5 groups such that, rows 1–20 are related to G_1, 21–40 are with G_2, 41–60 are with G_3, 61–80 are with G_4 and 81–100 are with G_5. The group G_k is then split in two subgroups with 10 elements each. The first 10 elements form one subgroup and the remaining 10 are related to the other subgroup. These subgroups are the blocks ($N_b = 10$) explained in Section 3. Given the longitudinal feature of the model, the rows incorporated in each block will be seen as an F-E indicator with 10 replications observed for different time points ($T_i = 10$ for $i = 1, \cdots, 10$).
2. Set true values in the regression part:
   $\beta^{r1} = (1, -0.5)^\top$, $\beta^{r2} = (-1, 0.5)^\top$, $\beta^{r3} = (1.5, 0.5, -1)^\top$ and $\tau = 0.36$.
3. Generate the true scores of the factors 3, 4 and 5:
   $\lambda_{3j} \sim N(0, 1)$, $\lambda_{4j} \sim N(0, 1)$ and $\lambda_{5j} \sim N(0, 1)$, for $j = 1, \cdots, N_e$.
4. Generate the true scores of the factors 1 and 2:
   $\lambda_{1j} \sim N(\beta_1^{r_1}\lambda_{3j} + \beta_2^{r_1}\lambda_{4j}, 1)$ and $\lambda_{2j} \sim N(\beta_1^{r_2}\lambda_{4j} + \beta_2^{r_2}\lambda_{5j}, 1)$, for $j = 1, \cdots, N_e$.
5. Generate the observed response variable Y_j using the equation (6), for $j = 1, \cdots, N_e$.
6. Generate $\sigma^2_{[it]} \sim U(0.1, 0.9)$, for $i = 1, \cdots, 10$ and $t = 1, \cdots, 10$.
7. Generate the errors $\epsilon_{[it]j} \sim N(0, \sigma^2_{[it]})$, for $i = 1, \cdots, 10$, $t = 1, \cdots, 10$ and $j = 1, \cdots, N_e$.
8. Obtain the non-null loadings in α as described near equation (2). Consider $\rho = 0.9$ and $\nu_{1i} = 2/\max\{\text{diag}(\Sigma_i)\}$, which implies that the largest variance in $\nu_{1i}\Sigma_i$ is 2 ($i = 1, \cdots, 10$). For all $[it] \notin G_k$, set $\alpha_{[it]k} = 0$.
9. Calculate X using the expression (1).
10. Randomly select 5 values in X to be treated as missing values.

The next step of the analysis is to define the prior specifications representing our initial uncertainty about the unknown parameters in the SEM model. Equation (2) indicates the prior for the non-null loadings in α. We set $\rho = 0.9$ and $\nu_{1i} = 9/\max\{\text{diag}(\Sigma_i)\}$ so that the largest variance *a priori* in $\nu_{1i}\Sigma_i$ is 9. In addition, consider $\nu_2 = 0.0001$ to induce a null loading as explained in Section 3. The structure of priors related to λ can be found in (4), (5) and in the discussion below these equations. Assume $a_{\sigma^2} = a_\tau = 2.1$ and $b_{\sigma^2} = b_\tau = 1.1$ in (7), that is, the mean is 1 and the variance is 10 in the IG prior. Finally, m_1, m_2 and m_3 are chosen to be vectors of zeros and S_1, S_2 and S_3 are diagonal covariance matrices with variance 9 for all regression coefficients.

The MCMC algorithm in `Stan` is configured with 7,000 iterations and a burn-in period involving the first 4,000 iterations. Only 1 chain is obtained for each parameter, therefore, the final posterior sample is composed by 3,000 observations. In terms of initial values of the chains we choose: zero for all regression coefficients, 0.5 for τ and all $\sigma^2_{[it]}$, 0.1 for all factor scores in λ and, finally, we set 1, 0 or -1 for $\alpha_{[it]k}$ by looking at the sign of the true value. Using the true sign to start the chains of the loadings is an strategy to induce the model to converge to the sign configuration of the true α, thus avoiding the possible sign exchange in the product $\alpha\lambda$. In practice, these starting values for α can be chosen based on the expected relationship between the variables $X_{[it]\bullet}$'s and each factor $\lambda_{k\bullet}$. The option $+1$ means positive correlation, 0 represents no association and -1 indicates negative correlation. The analyst might consider the prior knowledge to define these quantities in real applications such as the one presented ahead in Section 5.

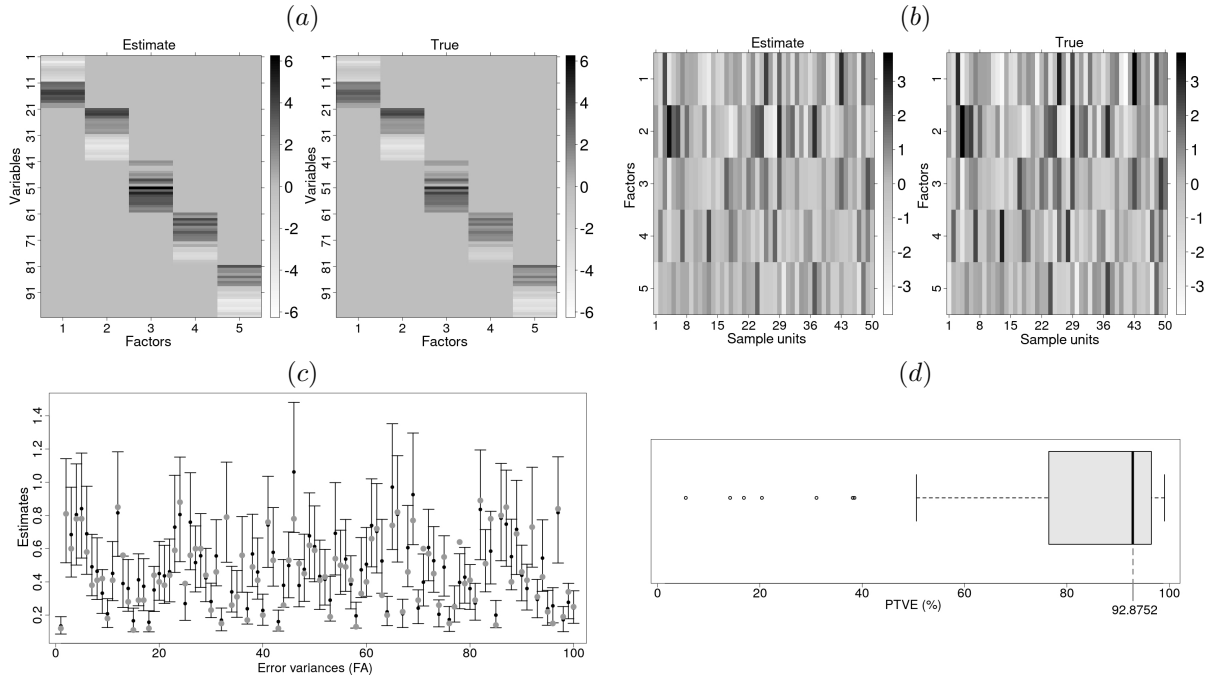

Figure 2: Exploring the SEM with artificial data. Panel (a) compares the posterior means and the true loadings in α. Panel (b) compares the posterior means and the true scores in λ. The heat maps in (a) and (b) show negative, zero and positive values in white, grey and black, respectively. Panel (c) presents the 95% Highest Probability Density (HPD) intervals for all error variances in σ^2 related to the CFA part; small black points are the posterior means and the greys points are the true values. Finally, Panel (d) displays a boxplot for the percentage of the total variability explained (PTVE) by the CFA, which is computed for each variable $[it]$.

The Panels (a) and (b) in Figure 2 compare the posterior means and the true values of the matrices α and λ, respectively. Note that the pattern of values in the estimated matrix is quite close to the one exhibited in the true matrix. This result clearly indicates that the SEM behaves well to estimate the loadings and the scores. Panel (c) shows the 95% Highest Probability Density (HPD) intervals for the error variances in σ^2. In this case, the true values are represented by grey points and the posterior mean are the black points. As it can be seen, almost all true values are captured by the credibility interval, which is another indication of good performance of the CFA part within the SEM.

Figure 2 (d) presents a boxplot summarizing the distribution of a diagnostic statistic called PTVE (percentage of the total variability explained by the CFA). This statistic is a vector $N_v \times 1$ with elements related to the variables in the rows of X. The calculation is based on the formulation $100[h/(h+\sigma^2)]$, which is applied using the current values of the parameters at each iteration of the MCMC. The means of the chains created in this process provide the data represented in the boxplot. The term h is a vector $N_v \times 1$ of comunalities obtained from the diagonal of the matrix $\alpha\alpha^\top$. The comunalities can be interpreted as the part of the variabilities, in the rows of X, that are explained by the CFA. The non-explained portions are accomodated by the variances forming the vector σ^2. Note that the median of the boxplot is located above

the 80% level, implying that most variables are well explained by the CFA.

Table 4 presents the true values and estimates related to the posterior distributions of the missing values, regression coefficients and the variance τ defined in the SEM fitted to the artificial data set. Note that all true values were captured by the corresponding HPD interval and they are close to the posterior means and medians. Larger posterior variances are detected for the missing values when compared to the regression parameters. The posterior uncertainty (see again the variance) for the coefficients in the regression r_3 (response is Y) is lower than those in r_1 and r_2 (response is a latent factor). In summary, the results in Table 4 provide evidence of a good performance of the SEM to estimate missing observations and to handle the regression part. A usual diagnostic for the regression (6) within the SEM structure is the adjusted coefficient of determination (commonly denoted by $R^2_{Adj.}$ in the literature). This statistic is calculated in each iteration of the MCMC and the result is reported as the posterior mean or median of the chain. We obtain $R^2_{Adj.} = 90.3155$ (mean) and 90.4206 (median), indicating that the regression is able to explain a high percentage of the variability in Y.

Table 4: Analysis of a simulated data set. True values and estimates from the posterior distribution of missing observations and parameters in the regression part of the SEM. The last two columns indicate the 95% Highest Posterior Density (HPD) intervals.

	true	mean	median	variance	HPD inf.	HPD sup.
$X_{[2,9]6}$	1.2875	0.2518	0.2533	0.3352	-0.9140	1.3392
$X_{[9,10]10}$	-0.0509	-0.4337	-0.4398	0.4542	-1.7698	0.8161
$X_{[4,2]26}$	4.0511	4.3378	4.3407	0.1874	3.4806	5.1478
$X_{[7,8]45}$	0.0190	-0.3664	-0.3656	0.6735	-1.9367	1.2768
$X_{[1,10]49}$	-2.2889	-2.1956	-2.2024	0.2321	-3.0765	-1.2087
$\beta_1^{r_1}$	1.0000	1.1058	1.0949	0.0445	0.6873	1.5033
$\beta_2^{r_1}$	-0.5000	-0.2959	-0.2937	0.0267	-0.5956	0.0435
$\beta_1^{r_2}$	-1.0000	-1.3232	-1.3066	0.0594	-1.8011	-0.8583
$\beta_2^{r_2}$	0.5000	0.4862	0.4844	0.0275	0.1929	0.8338
$\beta_0^{r_3}$	1.5000	1.4858	1.4846	0.0088	1.2980	1.6680
$\beta_1^{r_3}$	0.5000	0.6739	0.6665	0.0086	0.5036	0.8646
$\beta_2^{r_3}$	-1.0000	-1.1681	-1.1662	0.0161	-1.4196	-0.9318
τ	0.3600	0.4243	0.4127	0.0082	0.2642	0.5971

A final remark to be emphasized about the simulated data analysis related to Figure 2 and Table 4 is the fact that the visual inspection of some chains indicates the convergence to the target distribution. As a consequence, there is no need to increase the number of iterations and the burn-in period in the MCMC algorithm. The next step of the study is to consider a MC structure in which 100 artificial data sets, generated under the same conditions, are explored for a deeper evaluation of the proposed SEM. Consider again the procedure to generate data previously described in this section. Two scenarios are investigated here: $N_e = 50$ and $N_e = 100$. The idea behind this choice is to evaluate the gain in terms of inference when the sample size for each variable $[it]$ is doubled. Apart from the modification of N_e in the second scenario, all remaining settings defined for the previous simulated data analysis are kept the same in this MC study; this includes: size of the model components, true values, prior specifications, initial values and MCMC setup. As a computational strategy for faster results through parallel computing, we use the `R` package `snowfall` (Knaus, 2015) to manage the workflow in the MC scheme.

In order to evaluate the quality of the estimates, we choose to report the relative bias (RB) for each parameter of the SEM. This quantity is obtained through the expression $\text{RB}(\theta) = 100\ (\hat{\theta} - \theta_{\text{true}})\ /\ |\theta_{\text{true}}|$, where θ is a generic parameter, $\hat{\theta}$ is the posterior mean and θ_{true} is the true value. The RB is basically the ratio between the estimation error and the magnitude of the true value. Negative and positive results indicate underestimation and overestimation, respectively. The fraction is multiplied by 100 to adjust scale leading to a quantity indicating a percentage representing how big is the error with respect to the magnitude of the true value.

Figure 3 compares results from the scenarios $N_e = 50$ and $N_e = 100$. The first five panels show the median and interquartile range of the 100 replications of the RB for each parameter in the MC scheme. As it can be seen from Panels (c) and (e), the model tends to overestimate the variances σ^2 and τ; however,

the interquartile ranges capture the level zero. This overestimation might be explained by the asymmetric shape of the posterior distributions, that is, the heavy right tail affects the mean leading to a positive RB. In Panels (a), (b) and (d), one can note that the median RB's are located around the level zero providing a more balanced amount of estimates above and below the true value. In all panels displaying the RB, note that the distances between zero and the black points (median) seem smaller when $N_e = 100$. In addition, all interquartile ranges are clearly shorter in the scenario $N_e = 100$. Panel (f) presents the boxplots related to the 100 measures of PTVE and $R^2_{Adj.}$. The four graphs are entirely located above 60%, suggesting that the proposed SEM fits well all artificial data sets. The boxplot of PTVE's does not show great alteration between the scenarios $N_e = 50$ and $N_e = 100$. On the other hand, the graph related to the $R^2_{Adj.}$ is slightly higher and has lower variability when $N_e = 100$.

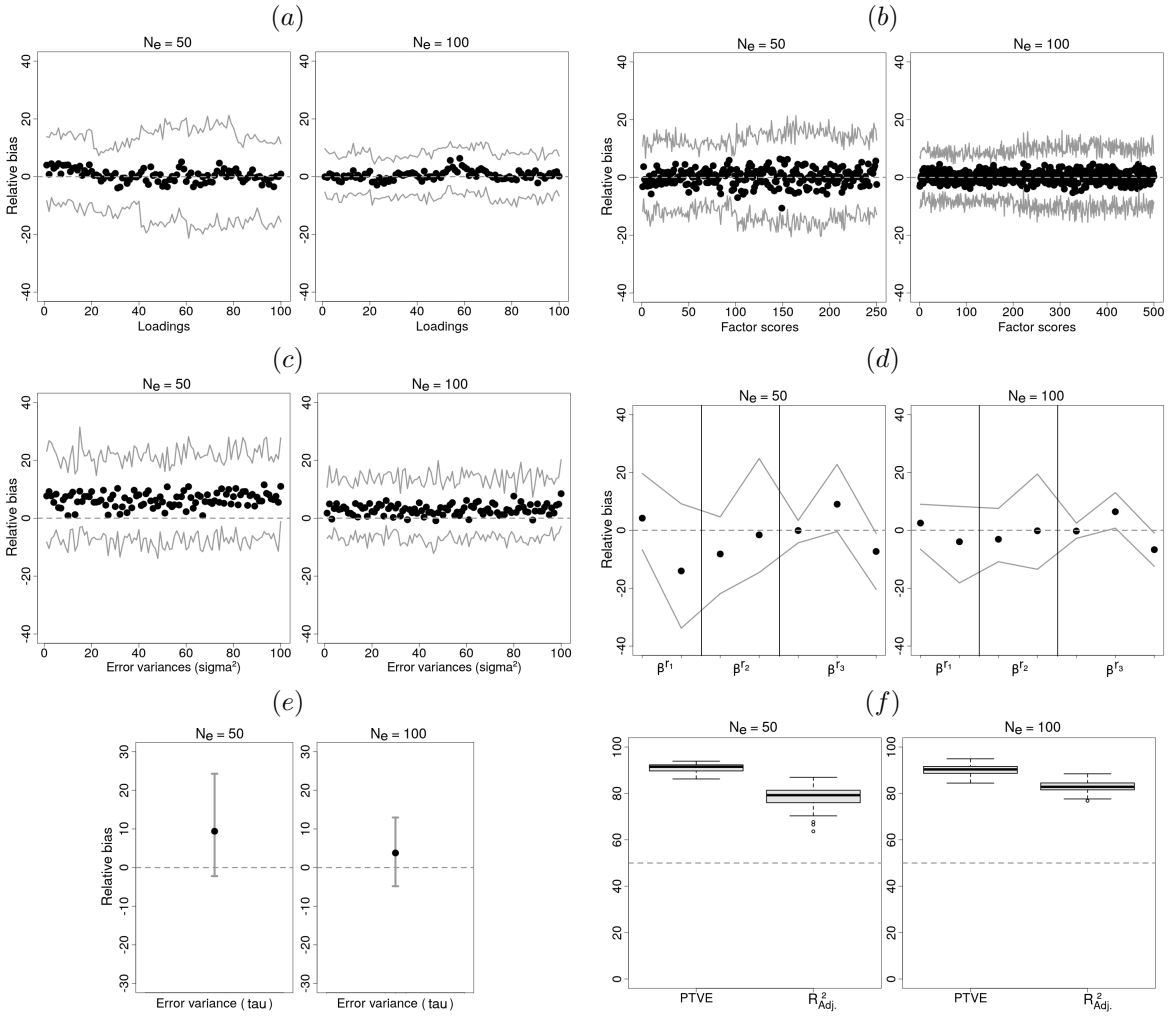

Figure 3: Results confronting $N_e = 50$ against $N_e = 100$ in the MC simulation study with 100 replications. Panels (a), (b), (c), (d) and (e) show the median relative bias (black point) and the interquartile range (in grey); the dashed horizontal grey line indicates the reference level zero. Panel (f) displays boxplots representing the distribution of the statistics PTVE and $R^2_{Adj.}$ for quality assessment; the dashed horizontal grey line identifies the 50% level. In each MCMC iteration, we calculate the median of the PTVE's (median of N_v values) and the $R^2_{Adj.}$ (for the 3^{rd} regression equation). The chains of the median PTVE and $R^2_{Adj.}$ are then summarized through the mean for each MC replication. The boxplots in Panel (f) are based on the corresponding 100 averages from the MC scheme.

Figure 4 exhibits another interesting result that confirms the good performance of the proposed SEM. Here, we explore for each parameter the percentages of MC replications in which the fitted model provides a 95% HPD interval that incorporates the corresponding true value. All panels clearly indicate coverage

percentages near the nominal level of 95%. Panels (a) and (b) suggest that the result seems to improve (i.e. points concentrating near the 95% level) in the scenario with $N_e = 100$.

The simulation study developed in this section is now completed with important conclusions obtained about the behavior of the SEM proposed in this work. In the next section, the analyses are focused on the real data set previously discussed in Section 2.

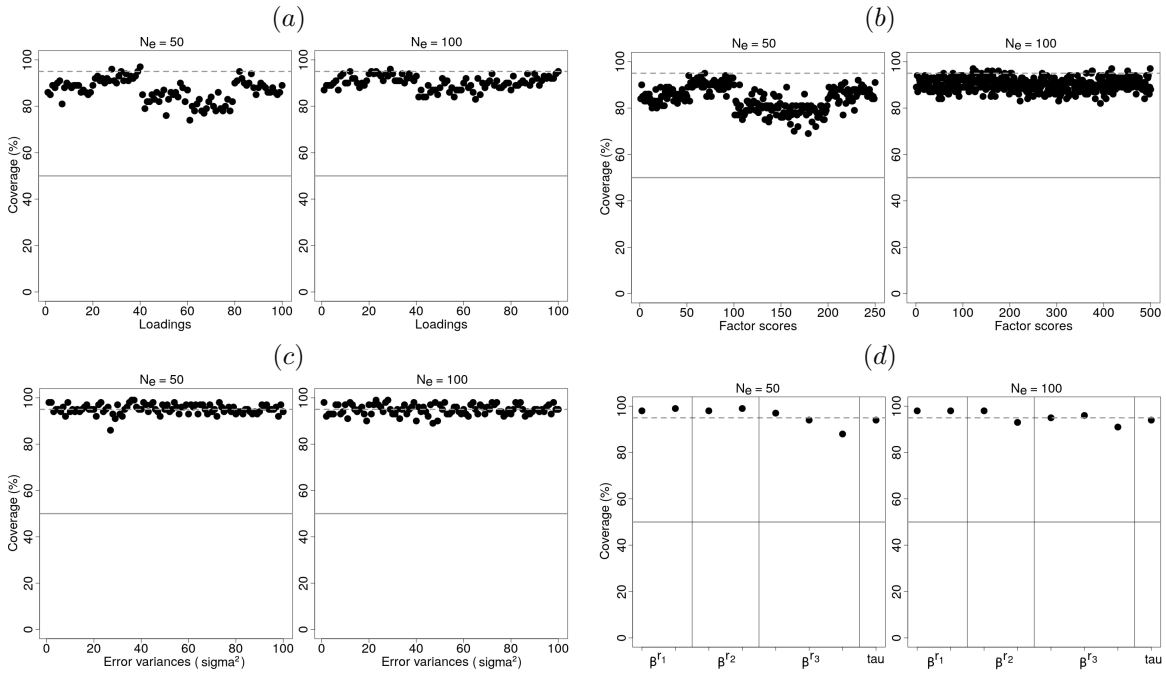

Figure 4: Comparison $N_e = 50$ versus $N_e = 100$ in terms of coverage percentages for 95% HPD intevals. The black points in the graphs represent the proportion of MC replications providing an HPD interval that includes the true value of the corresponding parameter. The nominal level (95%) is identified by the horizontal grey dashed line. The 50% level is indicaded by the horizontal grey continuous line.

5 Results for the ANEEL data

This section shows the results of the proposed SEM, with time dependence, fitted to the real data set of F-E indicators described in Section 2. This illustration has the following configurations: $N_v = 131$ variables (combining indicators and years), $N_e = 53$ energy companies, $K = 6$ factors to be estimated. The rows of Table 3 identify how the blocks $X_1, X_2, \cdots, X_{19}$ are partitioned into 6 disjoint groups $G_1, \cdots, G_6$ to be considered in the CFA. Recall that this separation is established based on the ANEEL technical note that provides the formulations in Table 2.

Consider here the first regression setting given by equation (3) in Section 3. We admit the next prior specifications:

- $\rho = 0.9$ and $\nu_{1i} = 9/\max\{\text{diag}(\Sigma_i)\}$ in (2). These choices provide a covariance matrix $\nu_{1i}\Sigma_i$ with diagonal variances ranging from 4.879 to 9.
- $\nu_2 = 0.0001$ to indicate a distribution centered at zero and having small variance for any loading connecting a factor k with a variable having index $[it] \notin G_k$.
- $m_0 = 0$ and $m_\beta = \mathbf{0}$ to avoid choosing a sign in our initial information about the regression coefficients. Further, let $v_0 = 9$ and $S_\beta = 9\ I_K$ to express high initial uncertainty; I_K is a $K \times K$ identity matrix.
- $a_{\sigma^2} = a_\tau = 2.1$ and $b_{\sigma^2} = b_\tau = 1.1$. This configuration of shape and scale parameters (IG distribution) provides expected value = 1 and variance = 10, which can be considered weakly informative in the present illustration.

The 6 latent factors, created in the CFA within the SEM model, will be used in the regression part to explain a response variable Y of interest for the application. This response is the indebtedness of the energy companies obtained via the formulation (DLR/DIF1) written in Table 2. The F-E indicators DLR and DIF1 have replications for the 8-year period. We choose to work with a single response variable, therefore, the median of the 8 replications are computed for each company. The median is chosen for this summarization as an strategy to avoid atypical values causing deviations from the normality assumption. Again, the ANEEL is particularly concerned with the sustainability of the energy companies and it tends to monitor, with great attention, this specific quantity to evaluate their performances.

In terms of MCMC settings, the algorithm is once again configured to run 7,000 iterations with a burn-in period involving the first 4,000 observations. Only 1 chain is built for each parameter. The starting value for the chain of $\alpha_{[it]k}$ is -1, 0 or 1, i.e., we consider the sign mark, indicated near the acronyms in Table 3 to choose one of these options. Recall that these signs represent the expected direction of the correlation between the corresponding variable and factor. The other parameters are initialized using the same options detailed in Section (4). Convergence issues are not detected in this application of the NUTS; visual inspection of traceplots indicates that the burn-in period is appropriate and the chains behave well with good mixing.

Figure 5 (a) presents a heat map image of the data matrix X without the first pre-processing step (standardization of columns in X) justified at the end of Section (2). As it can be seen, several rows (e.g. 1–50) exhibit a similar pattern of values, which suggests a positive correlation between variables. In addition, some columns (companies) have a strong dark color indicating large positive observations for most variables. Panel (b) shows the heat map for the fully pre-processed data matrix X. The correlations detected between rows in Panel (a) are less apparent in Panel (b). The image in Panel (b) displays the data that will be fitted through the SEM proposed in this paper. In the regression part, the response Y is supposed to follow the normal distribution as indicates the equation (3) in Section 3. The histogram in Panel (c) provides the distribution of the response "median indebtedness" obtained in accordance to the ANEEL formulation. The visual inspection of the histogram shape does not suggest that the normality is an unreasonable assumption.

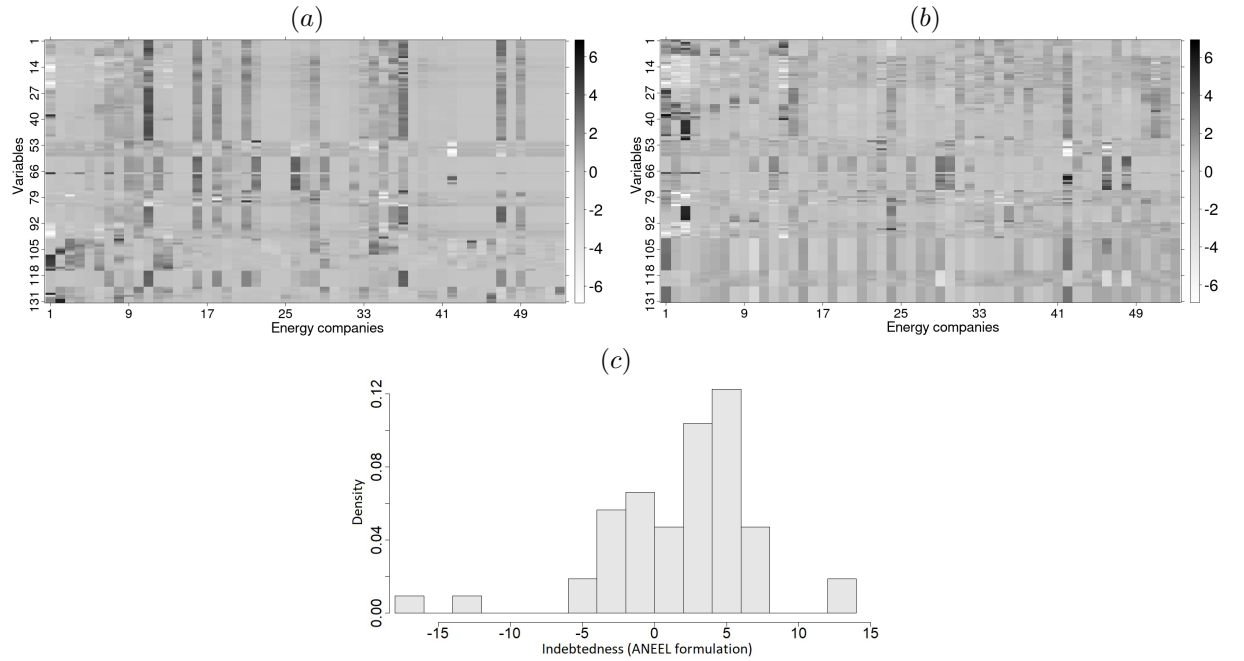

Figure 5: Descriptive graphs summarizing somes aspects about the ANEEL data. Panel (a) shows the matrix X without the first pre-processing correction to balance the observations across the energy companies (rows were standardized). Panel (b) presents the fully pre-processed matrix X (columns and rows were standardized). Remark: Panels (a) and (b) are built with missing observations replaced by the sample mean of the corresponding variable. The histogram in Panel (c) exhibits the distribution of the median values of indebtedness according to the ANEEL formulation.

Figure 6 (a) shows heat maps related to the analysis of the loadings. The left image on Panel (a)

exhibits the matrix of signs used as starting values for the chains of α. Recall that these signs were chosen as a strategy to provide a direction for the model in terms of the expected relationship between variables and latent factors. The heat map on the right of Panel (a) represents the posterior mean of the loadings matrix. As it can be seen, the estimated pattern of signs has some similarities with the image on the left, meaning that several variable-factor correlations are in accordance with the initial expectation. Due to the confirmatory setting imposed to the model, many loadings are estimated near zero (in grey), implying that the group of variables connected with one factor cannot be related to any other factor. Non-null estimates are observed as blocks positioned in the matrix diagonal.

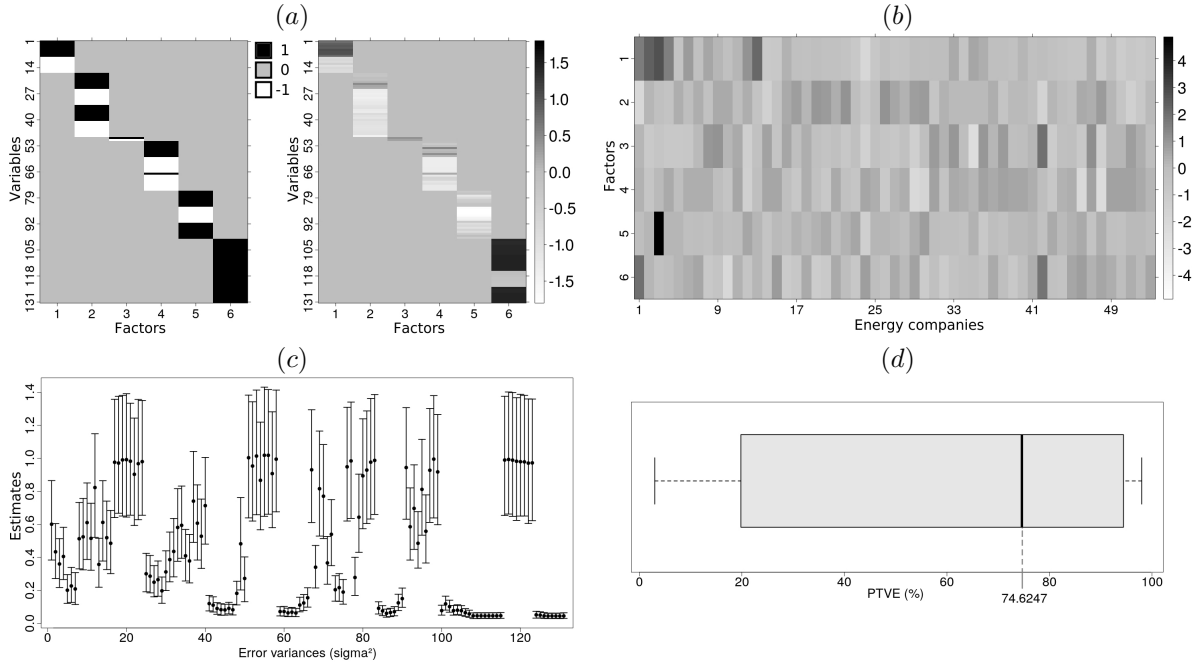

Figure 6: Results of the SEM fitted to the ANEEL data set with $K = 6$ factors. Panel (a) shows the initial values (left) and the posterior mean (right) of the loadings matrix α. Panel (b) presents the heat map for the posterior mean of the factor scores matrix λ. The factors in each row of λ are: 1 = indebtedness, 2 = efficiency, 3 = investiments, 4 = profitability, 5 = shareholder return and 6 = operational features. Panel (c) exhibits the 95% HPD intervals and the posterior mean (black point) for the error variances $\sigma^2_{[it]}$. Panel (d) shows a boxplot of the PTVE for each variable in X.

The posterior mean of the factor scores in λ are presented in the heat map given in Figure 6 (b). The estimates are ranging between -2.4173 and 4.8870, with the extreme value $\hat{\lambda}_{53} = 4.8870$ detected for the small company CEA (third column) from Amapá (north of Brazil). The first factor (first row of λ) is a new variable created based on a close relationship with the F-E indicators considered by ANEEL in the formulation to access the dimension indebtedness. The reader must see that there are two indebtedness variables being explored in this analysis, one of them is the estimated $\lambda_{1\bullet}$ (heat map in Figure 6 (b)) and the other is the median of the results obtained per year via the expression in Table 2.

Figure 6 (c) shows the posterior mean (black points) and the 95% HPD intervals for the error variances in σ^2 related to the CFA. The estimates are ranging between 0.0451 and 1.0187. There is a clear distinction in terms of amplitude when comparing the HPD intervals. The posterior uncertainty seems higher for the cases where $\sigma^2_{[it]}$ is estimated near 1. On the other hand, short intervals are found when $\sigma^2_{[it]}$ is estimated near 0. The reader is reminded that, in this real application, some variables have 8 replications (years) and the presence of these blocks explains the visual aspect of similar intervals near each other in the graph.

In the analysis of the statistics PTVE and $R^2_{Adj.}$, we follow the same principles described in Section 4 to calculate these quantities, i.e., the values are obtained at each iteration of the MCMC and the resulting chains are summarized through point estimates. Figure 6 (d) presents the distribution of the PTVE obtained for the 131 variables. In this case, the boxplot is wide with an interquartile range between 19.8044 and 94.4355. The median ($74.6247 > 50\%$) indicates that most variables are well explained by the CFA. As a diagnostic

for the regression part of the SEM, the model fit with $K = 6$ factors provides the posterior mean 62.3358 and the median 90.2369 for the chain of $R^2_{Adj.}$. These two estimates above 50% suggest that the regression structure is representing well the variability of the response in Figure 5 (c).

Table 5 presents some results corresponding to the regression part of the SEM. The only coefficient having a negative estimate is β_1 (indebtedness), however, this parameter is detected as non-significant (0 is inside the HPD interval). Three coefficients were estimated as positive and detected as significant according to the HPD criterion, they are β_0 (intercept), β_2 (efficiency) and β_3 (investments). In summary, the model fit indicates that increasing the scores of the factors efficiency and investments has an impact of increasing the median indebtedness calculated through the ANEEL formulation. Finally, the last parameter reported in Table 5 is the regression error variance τ. Note that the posterior distribution has a large variability in this case, which gives the idea of high posterior uncertainty about this parameter. As a consequence of the asymmetric shape of the posterior distribution, the mean and median are quite different.

Table 5: Estimates based on the posterior distribution of the parameters in the regression part of the SEM with $K = 6$ factors. The last two columns indicate the 95% Highest Posterior Density (HPD) intervals.

Parameter	Factor	Mean	Median	Variance	HPD inf.	HPD sup.
β_0		1.9658	1.9786	0.4301	0.6700	3.2367
β_1	Indebtedness	-0.3355	-0.3647	0.6919	-1.9178	1.3613
β_2	Efficiency	4.1310	4.1182	1.9629	1.3504	6.8425
β_3	Investments	3.9198	4.5246	2.9990	0.3657	6.3953
β_4	Profitability	1.1470	1.1373	1.2724	-1.1795	3.2523
β_5	S. return	0.1005	0.0487	1.2440	-2.0521	2.3493
β_6	Operational	0.9887	0.9968	1.3277	-1.3842	3.2234
τ		8.2194	2.2036	82.3089	0.1342	24.6697

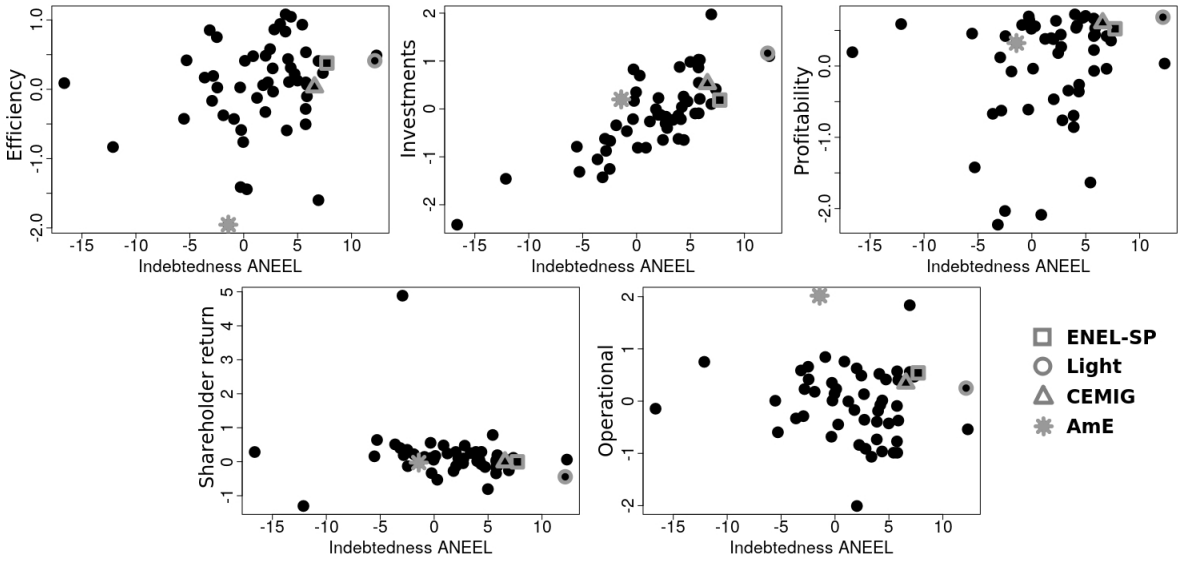

Figure 7: Dispersion diagrams confronting the posterior mean of the factor scores in λ against the response Y = indebtedness calculated through the ANEEL formulation in Technical Note 111/2016. The points representing the energy companies: AmE (Amazon), CEMIG (Minas Gerais), ENEL-SP (São Paulo) and Light (Rio de Janeiro) are highlighted with different symbols in grey.

Figure 7 shows dispersion graphs comparing the factors 2, 3, 4, 5 and 6 against the median indebtedness computed via the ANEEL formulation in Table 2. Four companies are identified with symbols in grey: three of them have concession in the southeast of Brazil, being classified as itermediate to large sized energy distributors. The company AmE has activities in the Amazon region and it is considered to be small sized. Note that the closer a point is to the bottom right corner of the graphs, the higher is the ANEEL indebtedness and lower is the factor score. Among the four companies under inspection, we have that Light tends to have

higher indebtedness, since the grey circle is always in the right hand side. In addition, AmE (grey asterisk) tends to have lower indebtedness due to the position on the left side of the panels. In general, CEMIG (grey triangle) and ENEL-SP (grey square) are near each other in all cases.

Note that, in the previous model fit, the latent factor indebtedness was treated as a covariate to explain the response in the regression part of the SEM structure. Table 5 shows that β_1 is not significant, therefore, the latent indebtedness does not have an impact on the formal indebtedness analyzed by ANEEL. As a simplification to the previous SEM, we investigate the model fit with $K = 5$ factors, that is, without the latent indebtedness. As it can be seen from Figure 8 and Table 6, few changes are observed when comparing these results with the corresponding ones for the case $K = 6$. The dispersion diagrams related to $K = 5$ are also quite similar to those in Figure 7; they are not shown here due to this great similarity. In terms of quality of the model fit, a small improvement can be detected when removing the latent indebtedness. The median PTVE is slightly larger, having the value 80.0346 for $K = 5$ (the result was 74.6247 for $K = 6$). In addition, the $R^2_{Adj.}$ also indicates an increase, with mean = 65.0194 and median = 91.3854 for $K = 5$ (the result was mean = 62.3358 and median = 90.2369 for $K = 6$).

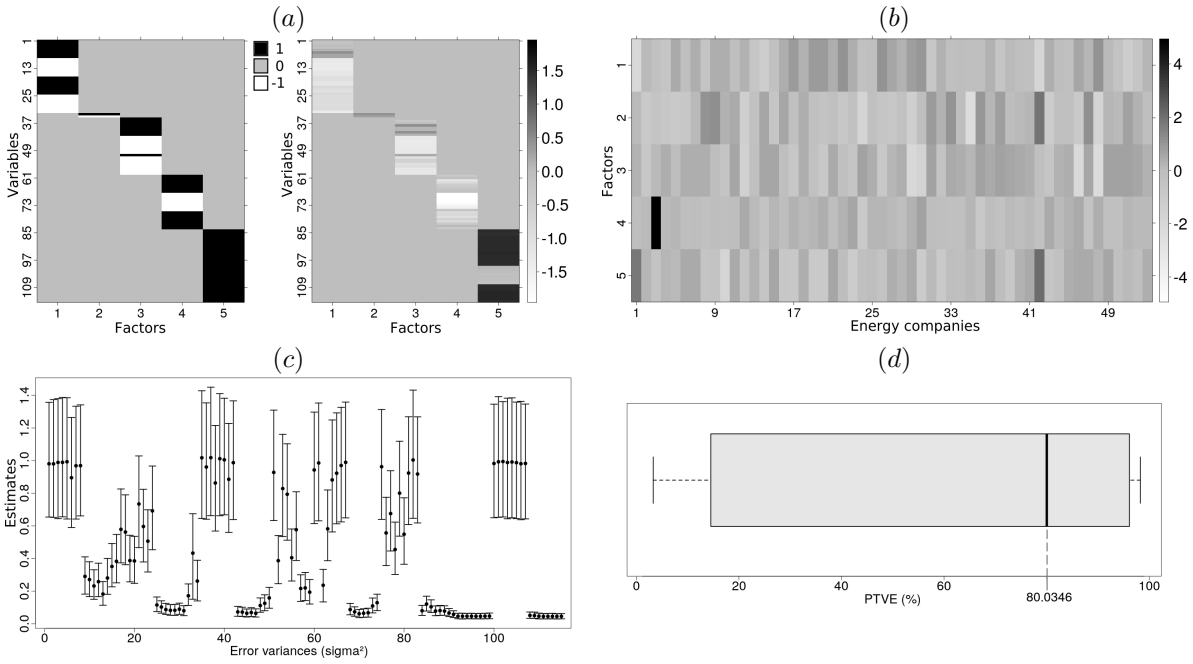

Figure 8: Results of the SEM fitted to the ANEEL data set with $K = 5$ factors. Panel (a) shows the initial values (left) and the posterior mean (right) of the loadings matrix α. Panel (b) presents the heat map for the posterior mean of the factor scores matrix λ. The factors in each row of λ are: 1 = efficiency, 2 = investiments, 3 = profitability, 4 = shareholder return and 5 = operational features. Panel (c) exhibits the 95% HPD intervals and the posterior mean (black point) for the error variances $\sigma^2_{[it]}$. Panel (d) shows a boxplot of the PTVE for each variable in X.

Table 6: Estimates based on the posterior distribution of the parameters in the regression part of the SEM with $K = 5$ factors. The last two columns indicate the 95% Highest Posterior Density (HPD) intervals.

Parameter	Factor	Mean	Median	Variance	HPD inf.	HPD sup.
β_0		2.0690	2.0826	0.3965	0.8999	3.3236
β_1	Efficiency	4.1043	4.0779	1.8212	1.3900	6.6576
β_2	Investments	4.1839	4.6904	2.5290	0.8771	6.5022
β_3	Profitability	0.8782	0.8967	1.2004	-1.3645	2.9022
β_4	S. return	-0.0757	-0.1146	1.0125	-2.0211	1.9651
β_5	Operational	1.5231	1.4966	1.1860	-0.6472	3.5779
τ		7.8101	2.0073	76.1951	0.1302	23.7560

6 Conclusions

The main purpose of the present paper is to develop a SEM with temporal dependence within groups of observed variables considered in the CFA part of the framework. The time association is established via a multivariate Gaussian distribution, commonly used in spatial analyses of areal data. The methodology here was motivated by a real application related to a data set of financial and economic indicators collected every 3 months by the Brazilian energy regulator ANEEL. These indicators are used to calculate six dimensions through formulations (published in the ANEEL Technical Note 111/2016) defined from discussions coordinated by ANEEL in the field of electricity distribution. These dimensions summarize critical information to evaluate the performance of the energy companies under concession for distribution in Brazil. ANEEL pays a special attention to the dimension indebtedness, which is a key element to evaluate the sustainability of the companies in the medium term.

The Bayesian SEM proposed here can be seen as an interesting tool to build latent factors representing the six dimensions documented in the ANEEL technical note. The summarization of the F-E indicators, under our SEM approach, is done by capturing the patterns and contributions of each indicator, as opposed to using one or more determinitic formulations that might be criticized by different specialists. In the regression part of the SEM structure, we assume the factor scores as covariates to explain the median indebtedness computed via the current ANEEL guidelines. The SEM approach provides a structure of relationships that allows us to measure the impact of each F-E indicator over any latent factor and over the target response.

A simulation study was conducted to evaluate the performance of the proposed SEM. The analysis was initially focused on exploring inferences for a single artificial data set and, later, a Monte Carlo setup was considered as a strategy to reach broader conclusions. The results here clearly confirm the good performance of the SEM and they point to the fact that better estimates could be obtained if the number of energy companies were larger.

In the real illustration, our findigs indicate that the dimensions "efficiency" and "investments" of the distributor have significant positive association with the median indebtedness used as response. It seems that increasing the score of these factors leads to an increased indebtedness, which in turn compromises the sustainability of the concession in the medium term. The impacts of the remaining latent factors were detected as non-significant, suggesting that the regulator may require the companies to improve them without undermining the sustainability.

Acknowledgements: The authors would like to thank the support from CEMIG (Companhia Energética de Minas Gerais) through the grant P&D 0636.

References

Anderson, J. C. and D. W. Gerbing (1988). Structural equation modeling in practice: a review and recommended two-step approach. *Psychological Bulletin 103*(3), 411–423.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014). *Hierarchical modeling and analysis for spatial data* (2 ed.). Boca Raton: Chapman and Hall/CRC.

Barbieri, A., M. Tami, X. Bry, D. Azria, S. Gourgou, C. Bascoul-Mollevi, and C. Lavergne (2018). Em algorithm estimation of a structural equation model for the longitudinal study of the quality of life. *Statistics in Medicine 37*(6), 1031–1046.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B 36*(2), 192–236.

Besag, J., J. York, and A. Mollie (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annal of the Institute of Statistical Mathematics 43*, 1–59.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2 ed.). New York: The Guilford Press.

Diggle, P. J., P. Heagerty, K. Y. Liang, and S. L. Zeger (2002). *The analysis of longitudinal data* (2 ed.). Oxford: Oxford University Press.

Dunson, D. B. (2007). Bayesian methods for latent trait modelling of longitudinal data. *Journal of the American Statistical Association 16*(5), 399–415.

Frese, M., H. Garst, and D. Fay (2007). Making things happen: reciprocal relationships between work characteristics and personal initiative in a four-wave longitudinal structural equation model. *Journal of Applied Psychology 92*(4), 1084–1102.

Gamerman, D. and H. F. Lopes (2006). *Markov Chain Monte Carlo: stochastic simulation for Bayesian inference* (2 ed.), Volume 68. London: Chapman and Hall/CRC.

Gamerman, D., H. F. Lopes, and E. Salazar (2008). Spatial dynamic factor analysis. *Bayesian Analysis 3*(4), 759–792.

Hoffman, M. D. and A. Gelman (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research 15*, 1351–1381.

Hoyle, R. H. (2014). *Handbook of structural equation modeling* (1 ed.). New York: The Guilford Press.

Johnson, R. A. and D. W. Wichern (2007). *Applied multivariate statistical analysis* (6 ed.). Upper Saddle River: Pearson/Prenticel Hall.

Keith, T. Z. (2019). *Multiple regression and beyond* (3 ed.). New York: Routledge.

Knaus, J. (2015). *Snowfall: easier cluster computing (based on snow).* R package version 1.84-6.1, `https://CRAN.R-project.org/package=snowfall`.

Lee, S. Y. and X. Y. Song (2003). Model comparison of nonlinear structural equation models with fixed covariates. *Psychometrika 68*, 27–47.

Lopes, H. F., D. Gamerman, and E. Salazar (2011). Generalized spatial dynamic factor models. *Computational Statistics and Data Analysis 55*, 1319–1330.

MacCallum, R. C. and J. T. Austin (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology 51*(1), 201–226.

Mayrink, V. D. and D. Gamerman (2009). On computational aspects of Bayesian spatial models: influence of the neighboring structure in the efficiency of MCMC algorithms. *Computational Statistics 24*, 641–669.

Mayrink, V. D. and J. E. Lucas (2015). Bayesian factor models for the detection of coherent patterns in gene expression data. *Brazilian Journal of Probability and Statistics 29*(1), 1–33.

Palomo, J., D. B. Dunson, and K. A. Bollen (2007). Bayesian structural equation modeling. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models*, Handbook of computing and statistics with applications, pp. 163–188. Amsterdam: North-Holland.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Sanchez, B. N., E. Budtz-Jorgensen, L. M. Ryan, and H. Hu (2005). Structural equation models: a review with applications to environmental epidemiology. *Journal of the American Statistical Association 100*, 1443–1455.

Seddig, D. and H. Leitgob (2018). Approximate measurement invariance and longitudinal confirmatory factor analysis: concept and application with panel data. *Survey Research Methods 12*(1), 29–41.

Skrondal, A. and S. Rabe-Hesketh (2004). *Generalized latent variable modeling: multilevel, longitudinal and structural equation models* (1 ed.). Interdisciplinary statistics. Boca Raton: Chapman and Hall/CRC.

Song, X. Y., S. Y. Lee, and Y. I. Hser (2008). A two-level structural equation model approach for analyzing multivariate longitudinal responses. *Statistics in Medicine 27*(16), 3017–3041.

Stan Development Team (2019). RStan: the R interface to Stan. R package version 2.19.2.

Usami, S., R. Jacobucci, and T. Hayes (2019). The performance of latent growth curve model-based structural equation model trees to uncover population heterogeneity in growth trajectories. *Computational Statistics 34*(1), 1–22.

West, M. and P. J. Harrison (1997). *Bayesian forecasting and dynamic models* (2 ed.). New York: Springer.