genpathmox: An R Package to Tackle Numerous Categorical Variables and Heterogeneity in Partial Least Squares Structural Equation Modeling

by Giuseppe Lamberti

Abstract Partial least squares structural equation modeling (PLS-SEM), combined with an analysis of the effects of categorical variables after estimating the model, is a well-established statistical approach to the analysis of complex relationships between variables. However, the statistical methods and software packages available are limited when it comes to assessing the effects of several categorical variables and trying to shape different groups following different models. Following the framework established by Lamberti et al. (2016), we have developed the **genpathmox** *R* package for handling a large number of categorical variables when faced with heterogeneity in PLS-SEM. The package has various functions that allow a wide range of assessments, including estimation, visualization, and hypothesis testing. In this paper, we describe the implementation of **genpathmox** in detail and demonstrate its usefulness by analyzing employee satisfaction data.

Introduction

Partial least squares structural equation modeling (PLS-SEM) (Wold, 1985) is a method for estimating causal relationships between observed and latent variables (Evermann and Rönkkö, 2021). This method was developed initially as an alternative to the classical covariance based-structural equation modeling (CB-SEM) that estimates latent variables (LVs) as common factors that explain co-variation between the associated indicators (Hair Jr et al., 2017a). However, in the last fifteen years has become a reference for estimating causal model in particular in marketing and management methodology (Becker et al., 2022; Sarstedt et al., 2022a,b), supported by several publications (Hair Jr et al., 2021; Henseler, 2020; Evermann and Rönkkö, 2021).

Simultaneous with the affirmation of the PLS-SEM approach to estimating causal models is the development of a parallel line of research into the problem of heterogeneity in parameter estimation, when the existence of different models is assumed for data characterized by differences in the coefficients that explain causal relationships between LVs. In this scenario, a single model could provide a biased view of those causal relationships. The literature describes several approaches to tackling heterogeneity that can be classified in terms of observed heterogeneity and non-observed heterogeneity (see Klesel et al. (2022) for a detailed review of the analysis of the heterogeneity in PLS-SEM).

Observed heterogeneity is based on the hypothesis that differences in models depend on certain categorical variables (CVs), e.g., sociodemographics, such that the data can be separated into groups and different models can be estimated according to a specific CV. The goal here is to verify, using statistical testing, whether differences in coefficients estimated for groups are significant or non-significant (Hair Jr et al., 2017b). Belonging to this category are the classical PLS multigroup tests, including the parametric (Keil et al., 2000), permutation (Chin and Dibbern, 2010), and Henseler (Henseler et al., 2009), as well as the more recent approach proposed by Klesel et al. (2019). As for non-observed heterogeneity, this is present when differences are inherent to the data. In this scenario, a different approach is required, and different models are typically identified following a latent class (Sarstedt et al., 2022c) or a clustering (Esposito Vinzi et al., 2008) approach.

Methodological advances and the increase in applications have led to the development of numerous *R* packages, starting with **plspm** (Sanchez et al., 2015), subsequently followed by **cSEM** (Rademaker and Schuberth, 2020) and **SEMinR** (Ray et al., 2020), both of which incorporate recent developments in the PLS approach, including improved estimation (consistent PLS (PLSc) (Dijkstra and Henseler, 2015)), improved validation criteria (the PLSpredict approach to prediction (Shmueli et al., 2016, 2019)), new reliability measures (Dijkstra and Henseler, 2015; Hair Jr et al., 2019), and also **matrixpls** (Rönkkö, 2017), particularly used for simulation studies. Concerning heterogeneity, the **cSEM** package allows multigroup analysis with several embedded tests. Pathmox analysis (Sanchez and Aluja, 2006; Lamberti et al., 2016, 2017) was proposed as a method useful when observed heterogeneity is assumed, but several potential CVs exist that could define different groups and models. This method explores and identifies, using an iterative algorithm, the most significantly different groups associated with significant differences in models. The algorithm follows a segmentation tree approach,

where each node is a PLS-SEM model. Differences are compared and partitions are chosen that define the most significant divergences between coefficients (Lamberti et al., 2016). The algorithm has been further improved, first by including a statistical test capable of identifying, for each split, the model coefficient responsible for the partitions (Lamberti et al., 2017), and then by combining the pathmox algorithm with classical multigroup analysis in a new approach called hybrid multigroup analysis (Lamberti, 2021).

In this paper, we describe the **genpathmox** package (Lamberti, 2022) to implement classical pathmox analysis and the more recently developed hybrid multigroup analysis (Lamberti, 2021). The package, initially developed in 2014, has been updated to include recent methodological advances (among them, PLSc (Dijkstra and Henseler, 2015) for estimating models) and has also been modified to work jointly with the **cSEM** package to increase analytical flexibility. Below we first review the framework formalized by Lamberti et al. (2016), then provide an overview of the **genpathmox** package, following which we demonstrate an application to a real-world analysis of employee satisfaction data and work climate drivers.

Overview of the pathmox methodology

Pathmox analysis (Sanchez and Aluja, 2006; Lamberti et al., 2016, 2017) was introduced to handle observed heterogeneity in PLS-SEM when several CVs exist. Unlike classical methods for tackling observed heterogeneity in PLS-SEM, instead of testing whether a CV produces a significant difference in model coefficients (i.e., a confirmatory approach), an exploratory approach is adopted. That is, pathmox explores and aims to identify, if they exist, the more significantly different groups associated with the more different PLS models. The algorithm applies a binary partition tree approach. It first estimates a single global model for the entire dataset to define the root note of the tree, and then explores all possible binary partitions for each CV. Differences in coefficients are statistically evaluated by the *F*-global test (Lamberti et al., 2016). This test provides a global measure of the degree of difference between partitions (i.e., a p-value). Comparisons are then sorted in descending order of the p-values, and the partition chosen as optimal is that corresponding to the smallest p-value (i.e., reflecting the greatest difference between models). This procedure is iterated and results in a tree structure of PLS-SEM models.

The degree of difference between models (the split criterion) is determined by applying a test, inspired by Chow (1960) and Lebart et al. (1979), that compares differences between the coefficients of two linear regression models. In pathmox, the difference between two PLS-SEM models is determined by comparing structural model coefficients, i.e., by comparing, as in the case of Chow (1960) and Lebart et al. (1979), restricted deviance and unrestricted deviance, defined, respectively, as the deviance calculated for the whole sample considering a single model valid for all the observations, and the sum of the model deviances estimated for each group of observations defined by the CVs.

From a graphical standpoint, pathmox is little different from a classical segmentation tree – as the algorithm produces a tree with a root, intermediate nodes, and terminal nodes – other than that each node is associated with a PLS model.

The split criterion

The split criterion used to define the tree partitions is a critical aspect of the pathmox algorithm. Below we describe the *F*-global test, the formulation of the null and alternative hypotheses, and the statistic used to test the null hypothesis (further details are available in Lamberti et al. (2016)).

Consider a simple structural model with one dependent LV denoted by the Greek letter η, and explained by a generic set of independent LVs denoted by the matrix $\mathbf{X} = \{\xi_{ip}\}$, where $i = 1, \ldots, n$ refers to the observations (n denotes their number), and where $p = 1, \ldots, P$ refers to the independent LVs. Its generalization into a more complex model is straightforward.

Using the matrix form, the model can be expressed as:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of coefficients for the regression of $\boldsymbol{\eta}$, and where $\boldsymbol{\varepsilon}$ is the disturbance term. Let the data be partitioned by rows, where the partition is determined by a CV with m categories (i.e., segments or groups); the number of units in group g ($g = 1, \ldots, m$) is denoted by n_g, and the total sample size can be expressed as $n = \sum_{g=1}^{m} n_g$. The *F*-global test compares model coefficients only by considering binary partitions. This means that the number of comparisons depends on the nature of the CVs (dummy = 1 comparison; nominal = $2^{m-1} - 1$ comparisons; ordinal = $m - 1$ comparisons).

The logic of the test is to compare the coefficients of two models considering two different

scenarios. Under the null hypothesis we assume one model is valid for all observations. This implies that one coefficient for each indipendet LV is enough to explain the dependent LV. Under the null hypothesis we assume that one model is valid for all observations. This implies that one coefficient for each independent LV is enough to explain the dependent LV. If we consider the simplest case of a dummy CV (m=2), denoting the two groups as A and B, the null and alternative hypotheses can be formulated as:

$$H_0 : \boldsymbol{\beta}_A = \boldsymbol{\beta}_B \tag{2}$$
$$H_1 : \boldsymbol{\beta}_A \neq \boldsymbol{\beta}_B \tag{3}$$

According to the null and alternative hypotheses, and following Lebart et al. (1979), we can rearrange equation 1 as follows:

$$\begin{bmatrix} \boldsymbol{\eta}_A \\ \boldsymbol{\eta}_B \end{bmatrix} \quad \begin{bmatrix} \mathbf{X}_A \\ \mathbf{X}_B \end{bmatrix} \quad [\ \boldsymbol{\beta}\ ] \quad + \begin{bmatrix} \boldsymbol{\varepsilon}_A \\ \boldsymbol{\varepsilon}_B \end{bmatrix} \tag{4}$$

$$\begin{bmatrix} \boldsymbol{\eta}_A \\ \boldsymbol{\eta}_B \end{bmatrix} \quad \begin{bmatrix} \mathbf{X}_A & 0 \\ 0 & \mathbf{X}_B \end{bmatrix} \quad \begin{bmatrix} \boldsymbol{\beta}_A \\ \boldsymbol{\beta}_B \end{bmatrix} \quad + \begin{bmatrix} \boldsymbol{\varepsilon}_A \\ \boldsymbol{\varepsilon}_B \end{bmatrix} \tag{5}$$

We calculate the deviance (the sum of squared residuals, SSR) for both models (under the null hypothesis (SSR_{H0}) and the alternative hypothesis (SSR_{H1})), and test the null hypothesis using the following statistic:

$$F = \frac{(SSR_{H_0} - SSR_{H_1}) \Big/ p}{SSR_{H_1} \Big/ n - 2p} \tag{6}$$

which follows an F distribution with p and $(n - 2p)$ degrees of freedom, where p is the number of explanatory LVs for the dependent LV, and where $n = n_A + n_B$ is the total number of observations.

Improving tree partition interpretation: the F-coefficient test

The F-coefficient test was an important improvement in the algorithm introduced in Lamberti et al. (2017). The F-global test used in pathmox as a split criterion is a global criterion that establishes whether or not the CV reflects a significant difference. However, it does not provide information as to which coefficients are responsible for that difference. The F-coefficient test complements the split criterion in pathmox by providing information on the coefficients responsible for the partition.

Rearranging the model formulated by Eq.1, we consider the particular case of one dependent LV denoted η, and two predictor LVs denoted ξ_1 and ξ_2:

$$\boldsymbol{\eta} = \boldsymbol{\xi}_1\beta_1 + \boldsymbol{\xi}_2\beta_2 + \varepsilon \tag{7}$$

Let us assume that a significant difference exists between the models estimated for the two groups, A and B, as defined by a generic dummy variable. Applying the F-global test (Lamberti et al., 2016) we cannot determine whether the difference between the two models depends on ξ_1 or ξ_2, or depends on both. However the null hypotheses for β_1 and β_2, and the corresponding alternative hypotheses, can be reformulated to determine whether the coefficients estimated for the predictors are significantly different, as follows:[1]

$$H_0 : \beta_{iA} = \beta_{iB} \quad \text{with i } = 1, 2 \tag{8}$$
$$H_1 : \boldsymbol{\beta}_A \neq \boldsymbol{\beta}_B \tag{9}$$

According to the null and alternative hypotheses, and following Lebart et al. (1979), we can rearrange equations 4 and 5 as:

$$\begin{bmatrix} \boldsymbol{\eta}_A \\ \boldsymbol{\eta}_B \end{bmatrix} \quad \begin{bmatrix} \boldsymbol{\xi}_{1A} & 0 & 0 \\ 0 & \boldsymbol{\xi}_{2A} & 0 \\ \boldsymbol{\xi}_{1B} & 0 & 0 \\ 0 & 0 & \boldsymbol{\xi}_{2B} \end{bmatrix} \quad \begin{bmatrix} \beta_1 \\ \beta_{2A} \\ \beta_{2A} \end{bmatrix} \quad + \begin{bmatrix} \boldsymbol{\varepsilon}_A \\ \boldsymbol{\varepsilon}_B \end{bmatrix} \tag{10}$$

[1] Note that the alternative hypothesis is the same for both ξ_1 and ξ_2.

$$\begin{bmatrix} \eta_A \\ \eta_B \end{bmatrix} \quad \begin{bmatrix} \xi_{1A} & 0 & 0 \\ 0 & \xi_{2A} & 0 \\ 0 & 0 & \xi_{1B} \\ 0 & \xi_{2B} & 0 \end{bmatrix} \quad \begin{bmatrix} \beta_{1A} \\ \beta_{1B} \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_A \\ \varepsilon_B \end{bmatrix} \tag{11}$$

$$\begin{bmatrix} \eta_A \\ \eta_B \end{bmatrix} \quad \begin{bmatrix} \xi_{1A} & 0 & 0 & 0 \\ 0 & \xi_{2A} & 0 & 0 \\ 0 & 0 & \xi_1 1B & 0 \\ 0 & 0 & 0 & \xi_{2B} \end{bmatrix} \quad \begin{bmatrix} \beta_{1A} \\ \beta_{1B} \\ \beta_{2A} \\ \beta_{2B} \end{bmatrix} + \begin{bmatrix} \varepsilon_A \\ \varepsilon_B \end{bmatrix} \tag{12}$$

We calculate again the deviance for both models (under the null hypothesis (SSR_{H0}) and the alternative hypothesis (SSR_{H1})) and test the null hypothesis using the following statistic:

$$F_i = \frac{\left(SSR_{H_{0_{\beta_i}}} - SSR_{H_1}\right) \Big/ 1}{SSR_{H_1} \Big/ 2\left(n - \sum p\right)} \text{ with i=1,2} \tag{13}$$

which follows an F distribution with 1 and $2\left(n - \sum p\right)$ degrees of freedom, and where p is the number of explanatory LVs, and $n = n_A + n_B$ is the total number of observations.

Note that both the F-global and the F-coefficient are implemented in the **genpathmox** package.

Stop criteria

Since pathmox is an iterative algorithm, its convergence depends on the specific stop criteria adopted by the user. Three criteria (all implemented in the **genpathmox** package) are proposed in Lamberti et al. (2016):

1. A more significant partition is not found. This means that the null hypothesis is not rejected in any of the candidate partitions, and as the obtained models are similar to each other, it make no sense to continue splitting the data. This condition is also strictly related to the significance threshold of the p-value chosen by the user, usually set to 0.05 (a typical p-value threshold in PLS-SEM applications).
2. Maximum tree depth is achieved. This is related to the number of terminal nodes required by the user, a choice based on the complexity of the model and the number of CVs used. Generally speaking, trees of 2-3 levels (with a maximum of 4-8 associated terminal nodes) are preferred.[2]
3. A node has too few observations to be partitioned. PLS-SEM works well with a relatively small number of observations, but it is recommended to fix a threshold of a relatively large number of observations to ensure nodes that are as representative as possible. For exploratory purposes, the recommended number of observations is between 50 and 100.

Pathmox to reduce the number of comparisons before running multigroup analysis: hybrid multigroup analysis

A criticism of the multigroup approach is that differences between coefficients could be difficult to interpret when the number of comparisons is high. This could happen when we have to simultaneously analyze more than one CV, or when the CV has more than 3 or 4 levels.

The pathmox algorithm does not realize an a posteriori statistical comparison of the coefficients of the models associated with the terminal nodes, nor does it establish the invariance between groups that is an important aspect of comparing PLS-SEM models (Henseler et al., 2016).[3] However, pathmox

[2] A greater tree depth results in a higher number of terminal nodes, with the direct consequence of the need to make a higher number of comparisons between model coefficients, and with results that may not always be easy to interpret.

[3] Invariance ensures that a dissimilar group-specific model estimate does not depend on diverse LV content and meanings across groups. A specific procedure to verify measurement invariance in the PLS-PM framework – proposed by Henseler et al. (2016) – is measurement invariance of composite models (MICOM), consisting of three hierarchical steps: (1) configural invariance, which ensures the same LV specifications when LVs are equally parameterized and estimated across groups; (2) compositional invariance, which ensures that LV scores reflect the same construct across groups; and (3) equality of latent variables, which means that values and variances ensure that data can be pooled across groups. If all three steps are confirmed, full measurement invariance is established, while if only the first two steps are confirmed partial measurement invariance is established, which is a necessary condition for performing multigroup analysis. A practical guideline on applying MICOM is provided by (Hair Jr

can be used just to reduce the number of groups to compare before running a classical multigroup analysis. Instead of using the original CV, the multigroup comparison uses a new CV defined by the CV groups resulting from the tree partitions. This is called the hybrid segmentation variable (Lamberti, 2021), which is used for the hybrid multigroup analysis.

The hybrid multigroup analysis consists of sequential steps as follows:

1. Use pathmox to identify the most significantly different groups
2. Use multigroup analysis to compare the groups:
   (a) Test the invariance of the constructs among groups using the MICOM procedure (Henseler et al., 2016)
   (b) Test the statistical differences between models using a criterion proposed by the literature (Klesel et al., 2022).

Note that the **genpathmox** package does not includes any function to automatically run the hybrid multigroup analysis. Rather, this analysis is done, as will be shown below, by combining the **genpathmox** and **cSEM** packages (Rademaker and Schuberth, 2020) (using the functions `testMICOM()` to test invariance, and `testMGD()` to compare coefficients).

Pathmox advantages and limitations

A first advantage of pathmox is that, given a set of CVs, it yields the most significantly different groups associated with the most significantly different models. The algorithm reduces the number of groups to be compared and analyzed, with the direct consequence of the user merely having to interpret the differences. A second advantage is that it ranks CVs by their importance in the split process (as in other classical tree partitioning procedures). This is important because an analysis of differences in PLS-SEM with more CVs involves not only comparing groups, but also establishing the most significant sources of heterogeneity in defining differences.

The main limitation of pathmox is related to the split criteria. The fact that the algorithm realizes an exhaustive search over unadjusted p-values to determine the best partition could potentially produce biased results (Loh and Shih, 1997). A possible solution would be to apply a Bonferroni correction for multiple comparisons, as yet this is not available in the current version of the package. As with the F-global and the F-coefficients use a parametric test based on a classical parametric statistic: the F statistic. This supposes the normality assumption of the perturbation terms with equal variance in all dependent constructs, even though the assumptions are rarely met in practice. Nevertheless, the sensitivity of the F statistic is guaranteed by larger sample size, lower levels of random perturbations, and clearer differences in the segments as shown by the simulations performed by Lamberti et al. (2016, 2017) . Another important limitation is that pathmox focuses only on the problem of detecting the path coefficients that are responsible for differences between PLS-PM models by adapting the measurement model to each segment. This leads to the problem of invariance that greatly increases in importance when we analyze data considering potential sources of heterogeneity and fitting one model to each segment. In this situation, it could become difficult to guarantee that each construct in each segment is measuring the same latent concept.

The *genpathmox* package

Overview

The **genpathmox** package is based on one main function called `pls.pathmox()`, which implements the pathmox algorithm and provides results for analysis. Four additional functions are a `summary()` function, and three plot functions (`plot()`, `bar_impvar()`, `bar_terminal()`) that help the user interpret results. In practice, users should first apply the main function `pls.pathmox()` to generate a "`plstree`" object. The components of this object include tree partition results, fitted coefficients of the PLS models for each terminal node, and relevant components to be used for relevant analysis. The "`plstree`" object plays an instrumental role, as it is a necessary input for the other functions in the package. This design is convenient, as details of data, PLS model, and tree split rules need only be specified once in `pls.pathmox()`, and are passed to other functions once the "`plstree`" object is called up.

The `summary()` function provides a complete output of all results, `plot()` provides the segmentation tree plot, `bar_impvar()` provides a bar plot of the ranking by importance of the CVs that participate in the split process, and `bar_terminal()` produces a bar plot of the coefficients of the PLS terminal nodes of the tree, enabling intuitive analysis of the differences between them.

et al., 2017b), while Henseler et al. (2016) provide more details on methodological aspects.

The **genpathmox** package has been designed to interact with the **cSEM** package (Rademaker and Schuberth, 2020), one of the latest and most complete packages for PLS-SEM analysis. This package can be used to analyze each model associated with the terminal nodes identified by pathmox and to run the hybrid multigroup analysis. To that end, the "`plstree`" object also contains a list of datasets called `.hybrid`, corresponding to lists of datasets of observations belonging in the tree terminal nodes.

Using the `hybrid` list combined with the **cSEM** package function `csem()`, each terminal node can be easily and completely analysed in terms of model validation, coefficient estimation, and inference. The resulting object generated by `csem()` can then be passed to `testMICOM()` to verify the invariance of the model constructs for the terminal nodes, and to `testMGD()` to compare the coefficients. The hybrid multigroup approach (Lamberti, 2021) can then be implemented. Further details on how to use the **cSEM** package are available in Rademaker and Schuberth (2020).

Figure 1 illustrates how to use the **genpathmox** package. On the left, the grey block contains the input elements, i.e., the data, the model, and the tree rules. Calling up `pls.pathmox()` generates the "`plstree`" object, as shown in the central orange block, which yields estimation and visualization results, as shown in the two orange blocks on the right. Finally, `plstree$hybrid` is used as the input parameter of the **cSEM** package that yields full results for the terminal nodes and the multigroup analysis, as shown in the blue blocks.

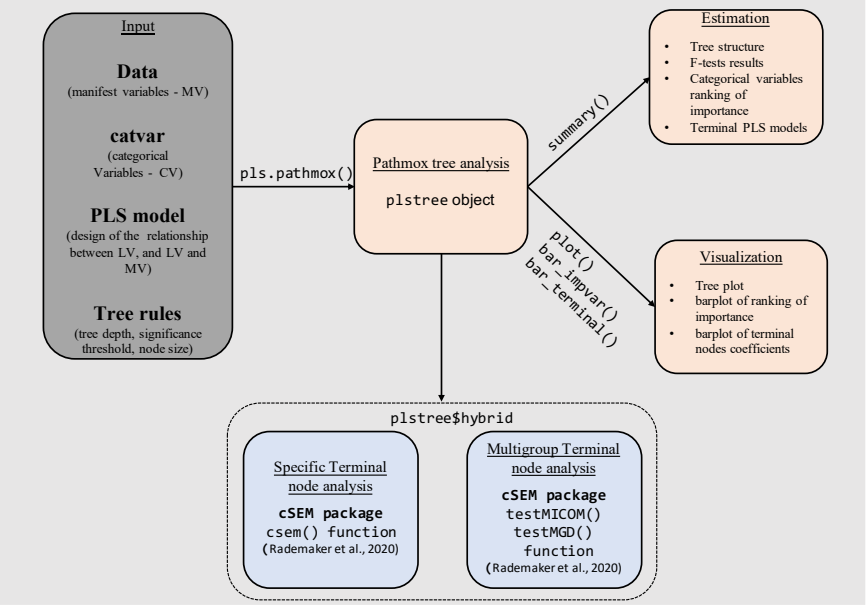

Figure 1: Illustration of **genpathmox** package functions

Implementation of main functions

Estimation function: `pls.pathmox`

To apply the `pls.pathmox` function, users need to specify at least three arguments:

1. `.model`. A description of the user-specified model using the lavaan syntax (Rosseel, 2012). Structural and measurement models are defined by enclosure between double quotes. The directional link between constructs is defined using the ("∼") operator. The dependent LV is on the left-hand side of the operator, and the explanatory LVs, separated by the ("+") operator, are on the right-hand side. As for the outer model, LVs are defined by listing the corresponding indicators after the operator ("=∼") if latent variable is modelled as a common factor, or the operator("<∼") if latent variable is modelled as a composite. On the left-hand side of the operator is the LV, and on the right-hand side are the indicators separated by the ("+") operator. Variable labels cannot contain ("."), following the lavaan syntax (for details of the meaning of modes A and B, see Hair Jr et al. (2016)).
2. `.data`. A matrix or data frame containing the indicators.
3. `.catvar`. A single factor or a set of factors organized as a data frame containing the CVs used as sources of heterogeneity.

Other input parameters have default values. Table 1 reports the meaning of each parameter and the admissible and default values.

Parameter	Purpose	Possible values	Default values
`.scheme`	inner weighting scheme	`"centroid"`, `"factorial"`, or `" path"`	`"path"`
`.consistent`	consistent PLS estimation (Dijkstra and Henseler, 2015) is used instead of classical approach (Wold, 1985)	`TRUE` or `FALSE`	`TRUE`
`.alpha`	minimum threshold significance	values belonging the interval $[0, 1]$	0.05
`.deep`	maximum tree depth	an integer ≥ 1	2
`.size`	minimum proportion of total sample admissible for a node size	value belonging the interval $[0, 1]$	0.10
`.candidate size`	minimum proportion admissible size for a candidate node	an integer ≥ 0	50
`.tree`	logic parameter to show the tree plot	`TRUE` or `FALSE`	`TRUE`

Table 1: Input parameters with default values

Once the split process is complete, results are saved in the object `"plstree"`, which now contains all the results necessary to interpret the pathmox analysis (see Table 2)

Results	Use
`MOX`	provides information on the tree structure: node type (intermediate or terminal), node size, binary split
`terminal_paths`	allows visualization of path coefficients and R^2 for each terminal node
`var_imp`	provides a ranking of the CVs used in the split process
`Fg.r`	identifies which CV is responsible for the partition
`Fc.r`	identifies the path coefficient responsible for the partition
`hybrid`	datasets associated with each terminal node

Table 2: `"plstree"` results

Visualization functions: `plot`, `bar_impvar` and `bar_terminal`

Three types of plots are possible in the **genpathmox** package: a pathmox tree plot, a bar plot reflecting the ranking of the CVs, and a bar plot of the PLS-SEM coefficients of the terminal nodes. The tree plot is obtained by applying the `plot()` function , which returns a tree structure with root, intermediate, and terminal nodes. For each partition, the F-global test p-value is reported with the associated CV, and each node reports the number of observations. The plot is implemented using functions of the **diagram** package (Soetaert, 2020). The plot of the CV ranking is obtained using the `bar_impvar()` function. This function uses the `barplot()` function to visualize the importance of the CVs. The importance of each CV is based on the F-statistic of the F-global test calculated for each CV in each tree node. Finally, the plot of the coefficients of the PLS-SEM model for each terminal node is obtained using the `bar_terminal()` function, also based on the `barplot()` function and allowing a more intuitive comparison of the coefficients of the terminal nodes.

The user can choose between two bar plot visualizations: (1) a plot of all the coefficients of the same model in the same plot, which is useful for comparing the terminal nodes models; and (2) a plot of the same coefficients for all terminal nodes in the same plot (lines correspond to the coefficients and bars report the coefficient effects), useful for a more direct comparison of a specific coefficient between models. In the former, the bar plot depicted for each model also plots the R^2. Visualization

options are selected by modifying the `.bycoef` parameter. By default, this is set to `FALSE`, meaning that the function implements the first option. We also need to specify for which dependent LV we want to visualize the predictor effect by fixing the parameter `.LV = ""`, which we do by indicating the dependent LV between the quotation marks.

Application: analysis of employee satisfaction in terms of work climate drivers

The **genpathmox** package is illustrated using real data on employee satisfaction in an international Spanish bank. In the financial sector, the impact of work climate on the relationship between strategic human resource management and organizational performance is crucial, in particular among younger employees (Kollmann et al., 2020). Another issue of relevance is that different groups of employees may respond in different ways to specific human resource management practices (Lamberti et al., 2020). The data referring to the younger employees ($\leq$30 years) of a Spanish bank contain measures regarding satisfaction (SAT), loyalty (LOY), and five work climate constructs: empowerment (EMP), company reputation (REP), leadership (LEAD), pay (PAY), and work conditions (WC). Our model relates the five work climate constructs with SAT, and SAT with LOY. Each construct is represented by a specific set of indicators. Information is also available on gender (female 53.36% vs male 46.64%), job level (intermediate in the company hierarchy, 52.01%), and seniority (length of service less than 5 years, 66.81%).

Full details of indicators and LVs are available in the **genpathmox** manual, and details of the theoretical framework are provided in Lamberti et al. (2020).

Our objectives were (1) to identify defining characteristics of different groups of employees, and (2) to analyze differences in the models for those groups.

Estimation

We use the `pls.pathmx()` function to partition the tree according to the CVs. The specified order was (1) the model (`.model`), (2) the data (`.data`), and (3) the CVs (`.catvar`). The other parameters were left at the default values. Using the lavaan syntax, we defined the structural model relating the five work climate constructs (EMP, REP, LEAD, PAY, WC) with SAT, and SAT with LOY, and we then related each construct to its own set of indicators (measurement model).

```
# load genpathmox package
library(genpathmox)

# load data
data(climate)

# define del model
climate_model  =  "
        # structural model
        SAT ~ EMP + REP + PAY + WC + LEAD
        LOY ~   SAT
        # measurement model
        EMP =~ Empo1 + Empo2 + Empo3 + Empo4 + Empo5
        REP  =~ Imag1 + Imag2 + Imag3
        PAY  =~ Pay1+ Pay2 + Pay3 + Pay4
        WC =~ Work1 + Work2 + Work3
        LEAD =~ Lead1 + Lead2 + Lead3 + Lead4 + Lead5
        SAT =~ Sat1 + Sat2 + Sat3 + Sat4 + Sat5 + Sat6
        LOY =~ Loy1 + Loy2 + Loy3
        "
# define the set of categorical variables
climate_catvar = climate[,1:3]

# run the pls.pathmox() function
climate.pathmox = pls.pathmox(
        .model = climate_model,
        .data = climate,
        .catvar = climate_catvar)
```

```
PLS-SEM PATHMOX ANALYSIS

-------------------------------------------
Info parameters algorithm
  parameters algorithm value
1    threshold signif.  0.05
2   node size limit(\%) 0.10
3     tree depth level  2.00

-------------------------------------------
Info segmentation variables
          nlevels ordered treatment
Level           3    TRUE   ordinal
Seniority       2    TRUE    binary
Gender          2   FALSE    binary
```

As shown above, the default output of the `pls.pathmx()` function is a small table containing the stop criteria and list of CVs used in the split partitions. Below we use the `summary()` function to interpret the results.

```
summary(climate.pathmox)

PLS-SEM PATHMOX ANALYSIS

-------------------------------------------
Info parameters algorithm:
  parameters algorithm value
1     threshold signif  0.05
2   node size limit(%)  0.10
3     tree depth level  2.00
-------------------------------------------
Info tree:
        parameters tree value
1             deep tree     2
2 number terminal nodes     3
-------------------------------------------
Info nodes:
  node parent depth  type terminal size      % variable   category
1    1      0     0  root       no  669 100.00     <NA>       <NA>
2    2      1     1  node       no  476  71.15    Level low/medium
3    3      1     1 least      yes  193  28.85    Level       high
4    4      2     2 least      yes  258  38.57   Gender     Female
5    5      2     2 least      yes  218  32.59   Gender       Male
-------------------------------------------
Info splits:

Variable:
  node variable     g1.mod g2.mod
1    1    Level low/medium   high
2    2   Gender     Female   Male

Info F-global test results (global differences):
     node F value Pr(>F)
[1,]    1  6.9711 <2e-16 ***
[2,]    2  3.0647 0.0021 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Info F-coefficient test results (coefficient  differences) :

Node 1 :
            F value Pr(>F)
EMP -> SAT   2.5902 0.1078
REP -> SAT   0.4056 0.5243
PAY -> SAT   3.6390 0.0567 .
WC -> SAT    0.7342 0.3917
LEAD -> SAT  4.1333 0.0422 *
SAT -> LOY   0.1044 0.7467
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Node 2 :
            F value Pr(>F)
EMP -> SAT   0.0229 0.8797
REP -> SAT   0.6333 0.4263
PAY -> SAT   0.1874 0.6652
WC -> SAT    0.9907 0.3198
LEAD -> SAT  2.5447 0.1110
SAT -> LOY  17.9754 <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
-------------------------------------------
Info variable importance ranking:
   variable   ranking
2     Level 0.3949974
3 Seniority 0.3101112
1    Gender 0.2948914
-------------------------------------------
Info terminal nodes PLS-SEM models (path coeff. & R^2):
          node 3 node 4 node 5
EMP->SAT  0.1233 0.2051 0.2725
REP->SAT  0.3037 0.1548 0.1238
PAY->SAT  0.0798 0.2863 0.1290
WC->SAT   0.4222 0.1333 0.3768
LEAD->SAT 0.2283 0.3283 0.1545
SAT->LOY  0.6934 0.7582 0.8806
R^2 SAT   0.6831 0.6863 0.6597
R^2 LOY   0.4808 0.5749 0.7754
```

According to `summary()` output, we can interpret the results as follows:

1. Pathmox indicates that different groups of employees exist that define SAT and LOY differently.
2. The variables that intervene in the split process are, first, job level (F-statistic = 6.971, p-value <0.001), and second, gender (F-statistic = 3.065, p-value = 0.002), i.e., employees are first partitioned into low/intermediate level employees versus high level employees, and low/intermediate level employees are then partitioned according to gender.
3. The coefficients responsible of the first split are LEAD–>SAT (F-statistic = 4.133, p-value = 0.042), and for the second split, SAT–>LOY (F-statistic = 17.975, p-value < 0.001).
4. Pathmox ultimately identifies three groups: high level employees (node 3), female low level employees (node 4), and male low level employees (node 5).
5. The CV ranking reveals that the most important differentiating characteristic for SAT and LOY is job level, followed by seniority, and finally gender.
6. In terms of work climate drivers defining SAT, the model comparisons indicate that:
   (a) High level employees (node 3) are least motivated by EMP ($\beta = 0.123$) and most motivated by WC $\beta = 0.422$) and REP ($\beta = 0.303$).
   (b) Female low level employees (node 4) are most motivated by LEAD ($\beta = 0.328$), PAY ($\beta = 0.286$), and EMP ($\beta = 0.205$).
   (c) Male low level employees (node 5) are most motivated by WC ($\beta = 0.377$) and least motivated by REP ($\beta = 0.124$), and also are the employees with the highest R^2 for LOY (0.775).

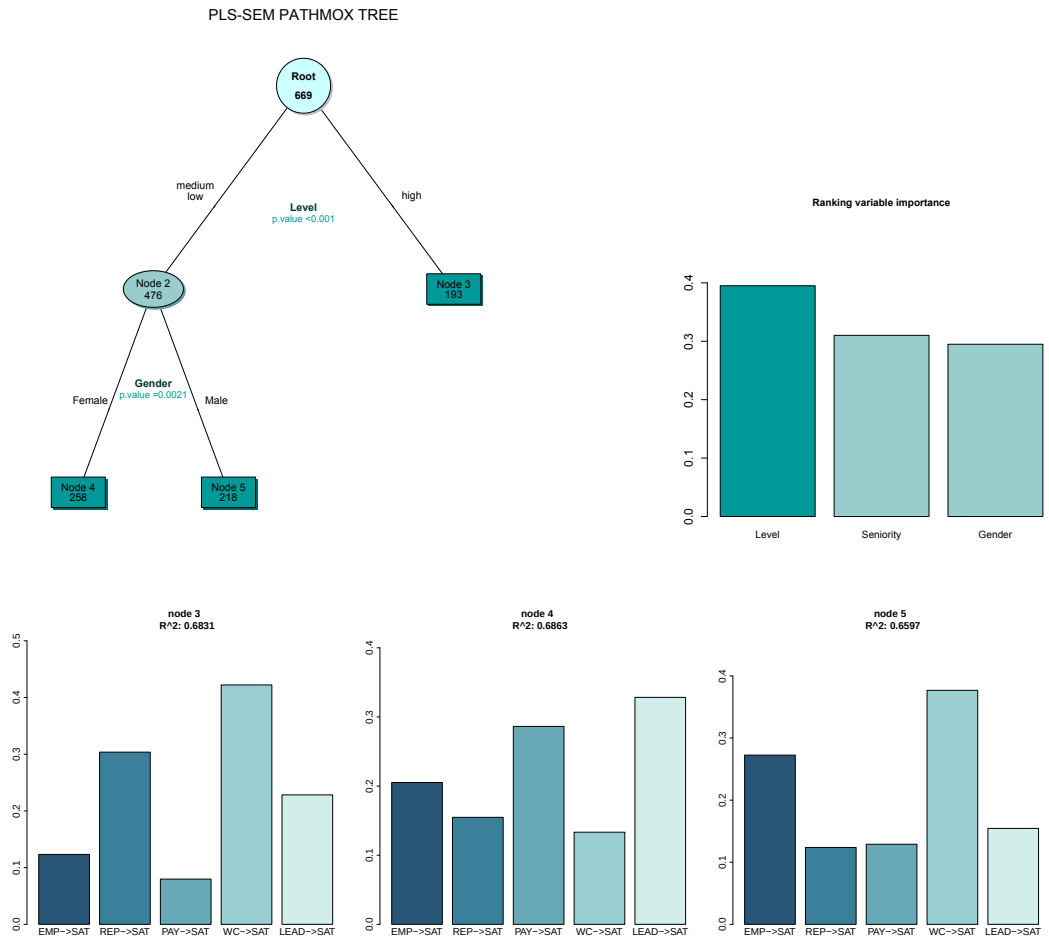

Figure 2: Visualization tools in the genpathmox package

Visualization

The `summary()` output can be complemented by plots. First, the `plot()` function, as applied to the object "`plstree`" generated from the `pls.pathmox` function, produces the tree plot. The `bar_impvar()` and `bar_terminal()` functions allow graphical visualization of the ranking of CVs and a comparison of the coefficients (default value `bycoef = FALSE`, and `LV = "SAT"` to show the predictors of SAT most relevant for the analysis of work climate drivers). The three plots are shown in Figure 2.

```
# treeplot
plot(climate.pathmox)
# ranking of CVs
bar_impvar(climate.pathmox)
# coefficients comparison
bar_terminal(climate.pathmox, .LV = "SAT")
```

Terminal node outputs

Specific terminal nodes can be analyzed using the **cSEM** package function `csem()`. By default the `csem()` function needs two parameter datasets that include all indicators (`.data`) and PLS-SEM model relationships (`.model`). As we are interested in the results of the terminal nodes, we pass the `hybrid` list in the "`plstree`" object to the `.data` parameter, and use the same model object defined for the `pls.pathmox()` function. Below we reproduce the code, but not the output, as not directly related with the **genpathmox** package.

```
# load cSEM package
library(cSEM)

# identify terminal nodes
terminal_nodes_data = climate.pathmox$hybrid

# terminal nodes results
terminal_nodes_results = csem(.data = terminal_nodes_data,
                                                    .model = climate_model)
```

Hybrid multigroup approach (Lamberti, 2021): invariance and multigroup analysis

For the invariance and the multigroup comparison of the terminal nodes identified by pathmox, we pass the object generated by the `csem()` function to the `testMICOM()` and `testMGD()` functions. Note that, for the multigroup comparison, we need to indicate which coefficients to compare by fixing the parameter `.parameters_to_compare`. We generate the work climate model, again using the lavaan syntax, but in this case only indicating the causal relationship between LVs. Finally, we indicate which statistical test to use for the comparison using the `.approach_mgd` function (for our example, the permutation test, `.approach_mgd = "Chin"`). Below we reproduce the code to show how **genpathmox** interfaces with **cSEM**, but not the results, as not generated by the **genpathmox** package.

```
# MICOM procedure
climateMICOM = testMICOM(terminal_nodes_results)



# define the relationship between LVs
climateMICOM = testMICOM(terminal_nodes_results)

climate_innermodel = "
        # Structural model
        SAT ~ EMP + REP + PAY + WC + LEAD
        LOY ~   SAT
        "
# multigroup analysis
climateMGA = testMGD(terminal_nodes_results,
                                        .parameters_to_compare = climate_innermodel,
                                        .approach_mgd = "Chin")
```

Summary

The **genpathmox** *R* package (Lamberti, 2022) handles observed heterogeneity in PLS-SEM when the number of CVs is high and we do not know what should be the most significant groupings are. Development of **genpathmox** reflects the statistical framework described in Lamberti (2021), and Lamberti et al. (2017, 2016), and the package has several functions that enable estimation and visualization of tree partitions. The **genpathmox** package quickly obtains a full picture of the effects of heterogeneity in PLS-SEM and identifies the groups that are most significantly different.

Bibliography

J. M. Becker, J. H. Cheah, R. Gholamzade, C. M. Ringle, and M. Sarstedt. Pls-sem's most wanted guidance. *International Journal of Contemporary Hospitality Management*, 2022. URL https://doi.org/10.1108/IJCHM-04-2022-0474. [p1]

W. W. Chin and J. Dibbern. An introduction to a permutation based procedure for multi-group pls analysis: Results of tests of differences on simulated data and a cross cultural analysis of the sourcing of information system services between germany and the usa. In *Handbook of partial least squares*, pages 171–193. Springer, Berlin Heidelberg, 2010. URL https://doi.org/10.1007/978-3-540-32827-8_8. ISBN 978-3-540-32825-4. [p1]

G. C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society*, pages 591–605, 1960. URL https://doi.org/10.2307/1910133. [p2]

T. K. Dijkstra and J. Henseler. Consistent partial least squares path modeling. *MIS quarterly*, 39(2): 297–316, 2015. URL https://www.jstor.org/stable/26628355. [p1, 2, 7]

V. Esposito Vinzi, L. Trinchera, S. Squillacciotti, and M. Tenenhaus. Rebus-pls: A response-based procedure for detecting unit segments in pls path modelling. *Applied Stochastic Models in Business and Industry*, 24(5):439–458, 2008. URL https://doi.org/10.1002/asmb.728. [p1]

J. Evermann and M. Rönkkö. Recent developments in pls. *Communications of the Association for Information Systems*, 44:123–132, 2021. URL https://doi.org/10.17705/1CAIS.044XX. [p1]

J. F. Hair Jr, G. T. M. Hult, C. Ringle, and M. Sarstedt. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. saGe publications, Los Angeles, 2016. URL https://doi.org/10.1007/978-3-030-80519-7. ISBN 978-1-5443-9640-8. [p6]

J. F. Hair Jr, L. M. Matthews, R. L. Matthews, and M. Sarstedt. Pls-sem or cb-sem: updated guidelines on which method to use. *International Journal of Multivariate Data Analysis*, 1(2):107–123, 2017a. URL https://doi.org/10.1504/IJMDA.2017.087624. [p1]

J. F. Hair Jr, M. Sarstedt, C. M. Ringle, and S. P. Gudergan. *Advanced issues in partial least squares structural equation modeling*. saGe publications, Los Angeles, 2017b. ISBN 9781483377391. [p1, 4]

J. F. Hair Jr, J. J. Risher, M. Sarstedt, , and C. M. Ringle. When to use and how to report the results of pls-sem. *European business review*, 31(1):2–24, 2019. URL https://doi.org/10.1108/EBR-11-2018-0203. [p1]

J. F. Hair Jr, G. T. M. Hult, C. M. Ringle, M. Sarstedt, N. P. Danks, and S. Ray. *Partial least squares structural equation modeling (PLS-SEM) using R: A workbook*. Springer Nature, Los Angeles, 2021. URL https://doi.org/10.1007/978-3-030-80519-7. ISBN 9783030805197. [p1]

J. Henseler. *Composite-based structural equation modeling: Analyzing latent and emergent variables*. Guilford Publications, New York, 2020. ISBN 9781462545605. [p1]

J. Henseler, C. M. Ringle, and R. R. Sinkovics. The use of partial least squares path modeling in international marketing. In *New challenges to international marketing*, pages 277–319. Emerald Group Publishing Limited, Bingley, 2009. URL https://doi.org/10.1108/S1474-7979(2009)0000020014. ISBN 978-1-84855-468-9. [p1]

J. Henseler, C. M. Ringle, and M. Sarstedt. Testing measurement invariance of composites using partial least squares. *International marketing review*, 3(3):405–431, 2016. URL https://doi.org/10.1108/IMR-09-2014-0304. [p4, 5]

M. Keil, B. C. Tan, K. K. Wei, T. Saarinen, V. Tuunainen, and A. Wassenaar. A cross-cultural study on escalation of commitment behavior in software projects. *MIS quarterly*, pages 299–325, 2000. URL https://doi.org/10.2307/3250940. [p1]

M. Klesel, F. Schuberth, J. Henseler, and B. Niehaves. A test for multigroup comparison using partial least squares path modeling. *Internet research*, 29(3):464–477, 2019. URL https://doi.org/10.1108/IntR-11-2017-0418. [p1]

M. Klesel, F. Schuberth, B. Niehaves, and J. Henseler. Multigroup analysis in information systems research using pls-pm: A systematic investigation of approaches. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 53(3):26–48, 2022. URL https://doi.org/10.1145/3551783.3551787. [p1, 5]

T. Kollmann, C. Stöckmann, J. M. Kensbock, and A. Peschl. What satisfies younger versus older employees, and why? an aging perspective on equity theory to explain interactive effects of employee age, monetary rewards, and task contributions on job satisfaction. *Human Resource Management*, 59(1):101–115, 2020. URL https://doi.org/10.1002/hrm.21981. [p8]

G. Lamberti. Hybrid multigroup partial least squares structural equation modelling: an application to bank employee satisfaction and loyalty. *Quality & Quantity*, pages 1–23, 2021. URL https://doi.org/10.1007/s11135-021-01096-9. [p2, 5, 6, 12]

G. Lamberti. *genpathmox: Pathmox Approach Segmentation Tree Analysis*, 2022. URL https://CRAN.R-project.org/package=genpathmox. R package version 0.9. [p2, 12]

G. Lamberti, T. B. Aluja, and G. Sanchez. The pathmox approach for pls path modeling segmentation. *Applied Stochastic Models in Business and Industry*, 32(4):453–468, 2016. URL https://doi.org/10.1002/asmb.2168. [p1, 2, 3, 4, 5, 12]

G. Lamberti, T. Banet Aluja, and G. Sanchez. The pathmox approach for pls path modeling: Discovering which constructs differentiate segments. *Applied Stochastic Models in Business and Industry*, 33(6): 674–689, 2017. URL https://doi.org/10.1002/asmb.2270. [p1, 2, 3, 5, 12]

G. Lamberti, T. Aluja Banet, and J. Rialp Criado. Work climate drivers and employee heterogeneity. *The International Journal of Human Resource Management*, 33(3):472–504, 2020. URL https://doi.org/10.1080/09585192.2020.1711798. [p8]

L. Lebart, A. Morineau, and J. P. Fenelon. *Traitement des donnees statistiques*. Dunod, Paris, 1979. ISBN 10. 2040107878. [p2, 3]

W. Y. Loh and Y. S. Shih. Split selection methods for classification trees. *Statistica sinica*, pages 815–840, 1997. URL https://www.jstor.org/stable/24306157. [p5]

M. E. Rademaker and F. Schuberth. *cSEM: Composite-Based Structural Equation Modeling*, 2020. URL https://CRAN.R-project.org/package=cSEM. Package version: 0.4.0. [p1, 5, 6]

S. Ray, N. P. Danks, and A. C. Valdez. *seminr: Building and Estimating Structural Equation Models*, 2020. URL https://CRAN.R-project.org/package=seminr. Package version: 0.4.0. [p1]

M. Rönkkö. *matrixpls: Matrix-based Partial Least Squares Estimation*, 2017. URL https://github.com/mronkko/matrixpls. R package version 1.0.5. [p1]

Y. Rosseel. lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, 48(2): 1–36, 2012. URL https://doi.org/10.18637/jss.v048.i02. [p6]

G. Sanchez and T. Aluja. Pathmox: a pls-pm segmentation algorithm. *Proceedings of KNEMO*, 69, 2006. [p1, 2]

G. Sanchez, L. Trinchera, and G. Russolillo. *plspm: Tools for Partial Least Squares Path Modeling (PLS-PM)*, 2015. URL https://github.com/gastonstat/plspm. R package version 0.4.9. [p1]

M. Sarstedt, J. F. Hair, M. Pick, B. D. Liengaard, L. Radomir, and C. M. Ringle. Progress in partial least squares structural equation modeling use in marketing research in the last decade. *Psychology & Marketing*, 39(5):1035–1064, 2022a. URL https://doi.org/10.1002/mar.21640. [p1]

M. Sarstedt, J. F. Hair Jr, and C. M. Ringle. Pls-sem: indeed a silver bullet–retrospective observations and recent advances. *Journal of Marketing Theory and Practice*, pages 1–15, 2022b. URL https://doi.org/10.1080/10696679.2022.2056488. [p1]

M. Sarstedt, L. Radomir, O. I. Moisescu, and C. M. Ringle. Latent class analysis in pls-sem: A review and recommendations for future applications. *Journal of Business Research*, 138:398–407, 2022c. URL https://doi.org/10.1016/j.jbusres.2021.08.051. [p1]

G. Shmueli, S. Ray, J. M. V. Estrada, and S. B. Chatla. The elephant in the room: Predictive performance of pls models. *Journal of Business Research*, 69(10):4552–4564, 2016. URL https://doi.org/10.1016/j.jbusres.2016.03.049. [p1]

G. Shmueli, M. Sarstedt, J. F. Hair, J. H. Cheah, H. Ting, S. Vaithilingam, and C. M. Ringle. Predictive model assessment in pls-sem: guidelines for using plspredict. *European journal of marketing*, 53(11): 2322–2347, 2019. URL https://doi.org/10.1108/EJM-02-2019-0189. [p1]

K. Soetaert. *diagram: Functions for Visualising Simple Graphs (Networks), Plotting Flow Diagrams*, 2020. URL https://CRAN.R-project.org/package=diagram. R package version 1.6.5. [p7]

H. Wold. *Partial Least Squares*, volume 6, pages 581–591. John Wiley and Sons, Ltd, 1985. URL https://doi.org/10.1002/0471667196.ess1914.pub2. [p1, 7]

Giuseppe Lamberti
Department of Economics and Statistics
University of Naples, Federico II
Italy
http://orcid.org/0000-0002-8666-796X
giuseppe.lamberti@unina.it