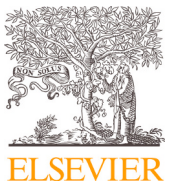
Computational Statistics and Data Analysis 219 (2026) 108348



Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda





Automated specification search for composite-based structural equation modeling: A genetic approach

Laura Trinchera [a], Gloria Pietropolli [b], Mauro Castelli [c], Florian Schuberth [d,*]

[a] NEOMA Business School, 1 Rue du Maréchal Juin, Mont-Saint-Aignan, 76130, France
[b] University of Trieste, Via Alfonso Valerio 12, Trieste, 34127, Italy
[c] NOVA Information Management School (NOVA IMS), Universidade NOVA de Lisboa, Campus de Campolide, Lisboa, 1070-312, Portugal
[d] University of Twente, P.O. Box 217, Enschede, 7500 AE, the Netherlands

ARTICLE INFO

Keywords:
SEM
Genetic algorithms
PLS-SEM
Simulation study
Model search
Exploratory approach

ABSTRACT

Structural Equation Modeling (SEM) is primarily employed as a confirmatory approach for empirically testing theoretical models by assessing how well they fit collected data. In practice, researchers frequently take a more exploratory approach and manually assess alternative models. Although automated search techniques have been developed for factor-based SEM to identify the best-fitting model, automated specification search remains largely unexplored in composite-based SEM. To address this gap, a new method is introduced: Automated Genetic Algorithm Specification Search for Partial Least Squares Path Modeling (AGAS-PLS). The proposed algorithm combines partial least squares path modeling with a genetic algorithm to identify the "best" structural model. A Monte Carlo simulation was conducted to assess the ability of AGAS-PLS to accurately identify the structural model of the data-generating process under various conditions, including different sample sizes and levels of model complexity. The practical applicability of AGAS-PLS was further illustrated using empirical data.

1. Introduction

Partial Least Squares Path Modeling (PLS-PM, Wold, 1982; Tenenhaus et al., 2005), also known as Partial Least Squares Structural Equation Modeling (PLS-SEM, Hair et al., 2011), is an approach to composite-based structural equation modeling (SEM, Henseler, 2021). It is widely used in various fields of social sciences (Cook and Forzani, 2023), including marketing (Sarstedt et al., 2022a), information systems research (Benitez et al., 2020; Evermann and Rönkkö, 2023), human resource management (Ringle et al., 2020), and tourism research (Müller et al., 2018). Over the last decade, PLS-PM has undergone numerous enhancements (see, e.g., Khan et al., 2019). For example, it has been extended to consistently estimate linear and non-linear structural equation models containing composites and common factors (Dijkstra and Schermelleh-Engel, 2014; Dijkstra and Henseler, 2015a,b). Moreover, modifications to PLS-PM have been proposed to deal with ordinal observed variables (Schuberth et al., 2018), outliers (Schamberger et al., 2020), and multicollinearity issues (Jung and Park, 2018). Furthermore, overall model fit criteria have been introduced (Schuberth et al., 2023a) and approaches to evaluate out-of-sample predictive power have been proposed (Shmueli et al., 2016; Schamberger et al., 2023). In addition, tests for composite invariance and multigroup analysis have been developed (Henseler et al., 2016; Klesel et al.,

* Corresponding author.
E-mail addresses: laura.trinchera@neoma-bs.fr (L. Trinchera), gloria.pietropolli@units.it (G. Pietropolli), mcastelli@novaims.unl.pt (M. Castelli), f.schuberth@utwente.nl (F. Schuberth).

https://doi.org/10.1016/j.csda.2026.108348
Received 14 March 2025; Received in revised form 26 January 2026; Accepted 26 January 2026
Available online 29 January 2026
0167-9473/© 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

2022). Despite these numerous enhancements to PLS-PM, *specification search* (MacCallum, 1986; Long, 1983) is a topic that has received limited attention in the PLS-PM literature. Specification search refers to the process of modifying an initially specified model to improve its quality, e.g., its fit to the empirical data or its out-of-sample predictive power. Although some progress has been made in manual specification search (e.g., Sharma et al., 2019; Danks et al., 2020), automated specification search remains a largely unexplored topic in the PLS-PM context.

To address this gap, a Genetic Algorithm (GA, Holland, 1992) specifically designed for PLS-PM, named Automated Genetic Algorithm Specification Search for PLS-PM (AGAS-PLS), is introduced. It leverages the effectiveness of GAs in solving complex, high-dimensional optimization problems to conduct automated model specification searches within the PLS-PM framework. In particular, it uses a GA to automatically evaluate alternative structural model specifications and identify the linear recursive structural model that best fits the data, given a predefined set of exogenous constructs and a fixed outer model.

The remainder of the paper is structured as follows. Section 2 gives a concise overview of factor-based SEM, composite-based SEM and PLS-PM. Section 3 reviews existing approaches to automated specification search that have been proposed in the SEM literature. Section 4 details how AGAS-PLS combines PLS-PM and GA. Section 5 presents the results of a Monte Carlo simulation study conducted to evaluate the ability of AGAS-PLS to identify the structure of the structural model of the data-generating process (DGP). Section 6 presents an empirical illustration of AGAS-PLS using a dataset originally collected by Bergami and Bagozzi (2000), demonstrating the practical applicability of the proposed approach. Finally, Section 7 presents conclusions and directions for future research.

2. Approaches to structural equation modeling and partial least squares path modeling

SEM is a family of multivariate statistical techniques that allows researchers to model relations between constructs. A structural equation model typically comprises two sub-models (Lohmöller, 1989): (i) an outer model, which links constructs to their observed variables; and (ii) a structural model (also called inner model), which defines the relations between constructs.

In SEM, there are different types of outer model (e.g., Yu et al., 2021; Schuberth et al., 2023b). Arguably, the two main types of outer model are the common factor model and the composite model (Cho et al., 2022a). In the common factor model, also known as the reflective measurement model, a construct is modeled as a common factor, which is a latent variable that causes a set of observed variables, also called manifest variables (Jöreskog, 1970). Alternatively, the composite model can be used to model a construct (Grace and Bollen, 2008; Henseler and Schuberth, 2021; Hwang et al., 2021; Schuberth, 2023). In this approach, commonly referred to as a *formative specification* (Diamantopoulos and Winklhofer, 2001), the construct ξ_j is modeled as a composite, i.e., an emergent variable (Henseler and Schuberth, 2020) formed as a weighted linear combination of observed variables:

$$\xi_j = \boldsymbol{w}_j' \boldsymbol{y}_j, \quad j = 1, \ldots, J, \tag{1}$$

where J denotes the total number of constructs and the vector $\boldsymbol{w}_j$ contains the weights assigned to the observed variables $\boldsymbol{y}_j$. Both $\boldsymbol{w}_j$ and $\boldsymbol{y}_j$ are vectors of length K_j, representing the total number of observed variables used to form the composite ξ_j. In line with PLS-PM, each observed variable is connected to only one composite.

The structural model establishes the relations between the constructs:

$$\boldsymbol{\xi} = \boldsymbol{B}\boldsymbol{\xi} + \boldsymbol{\zeta}, \tag{2}$$

where $\boldsymbol{\xi}$ is the vector of composites, i.e., $\xi_j = \boldsymbol{w}_j' \boldsymbol{y}_j$, $\boldsymbol{B}$ is the matrix of the path coefficients, and $\boldsymbol{\zeta}$ contains the structural error terms. In line with PLS-PM conventions, only linear recursive structural models are considered. Additionally, the structural error terms are assumed to be mutually uncorrelated and uncorrelated with the other exogenous variables in the model.

Several approaches have been proposed for estimating the parameters of structural equation models. Among these, the Maximum-Likelihood (ML) estimator (Jöreskog, 1970) is the most widely used method for models in which all constructs are specified as common factors, i.e., in factor-based SEM. On the other hand, alternative approaches have been developed for structural equation models with composites only, i.e., composite-based SEM. Prominent examples of these approaches include Generalized Structured Component Analysis (GSCA, Hwang and Takane, 2004) and Partial Least Squares Path Modeling (PLS-PM, Wold, 1975). Despite recent contributions enabling flexible modeling of composites in factor-based SEM (Schuberth, 2023; Yu et al., 2023), PLS-PM remains the most widely applied approach to composite-based SEM.

PLS-PM proceeds in two stages to obtain the parameter estimates of a structural equation model (Wold, 1982; Lohmöller, 1989). First, the iterative PLS algorithm is applied to estimate the weight vectors $\boldsymbol{w}_j$, which are then used to form the composites. Second, Ordinary Least Squares (OLS) regressions are performed to estimate the structural relations among the resulting composites. As is standard in PLS-PM, all observed variables are standardized prior to estimation, resulting in each variable having a mean of zero and a variance of one.

The PLS algorithm iterates through three steps at each iteration (i) to estimate the weights used to form the composites. In the first step, an outer approximation of each composite ξ_j is computed as follows:

$$\hat{\boldsymbol{\xi}}_j^{(i)} = \boldsymbol{Y}_j \hat{\boldsymbol{w}}_j^{(i)}, \tag{3}$$

where $\boldsymbol{Y}_j$ is the data matrix containing the N observations of the standardized observed variables $\boldsymbol{y}_j$ associated with the composite ξ_j. The weights $\hat{\boldsymbol{w}}_j^{(i)}$ are scaled in each iteration by $\left(\hat{\boldsymbol{w}}_j^{(i)\prime} \boldsymbol{S}_{jj} \hat{\boldsymbol{w}}_j^{(i)}\right)^{0.5}$, where $\boldsymbol{S}_{jj}$ denotes the sample correlation matrix of $\boldsymbol{y}_j$. This ensures that each outer approximation has unit variance and zero mean. Typically, unit weights are used as the initial weights ($\hat{\boldsymbol{w}}_j^{(0)}$).

In the second step, the inner approximation of ξ_j is obtained as a weighted linear combination of the outer approximations:

$$\tilde{\xi}_j^{(i)} = \sum_{l=1}^{J} e_{jl}^{(i)} \hat{\xi}_l^{(i)}. \tag{4}$$

The inner weights $e_{jl}^{(i)}$ are usually calculated using one of the following schemes: centroid, factorial, or path weighting scheme (see, e.g., Tenenhaus et al., 2005). These schemes typically produce very similar results (Noonan and Wold, 1982). The resulting inner approximations of ξ_j are subsequently scaled to have unit variance.

In the third step, the new weights $\boldsymbol{w}_j^{(i+1)}$ are obtained. Two modes are commonly used for this purpose. In Mode A, the weights of the observed variables $\boldsymbol{y}_j$ are obtained as the coefficients of K_j simple OLS regressions, where each observed variable is regressed on the corresponding inner approximation $\tilde{\xi}_j^{(i)}$. In contrast, in Mode B, the weights are obtained as the coefficients of a multiple OLS regression of the inner approximation $\tilde{\xi}_j^{(i)}$ on the corresponding observed variables $\boldsymbol{y}_j$. The weights $\hat{\boldsymbol{w}}_j^{(i+1)}$ are then scaled by $\left(\hat{\boldsymbol{w}}_j^{(i+1)\prime} \boldsymbol{S}_{jj} \hat{\boldsymbol{w}}_j^{(i+1)}\right)^{0.5}$ to ensure that the resulting composite $\hat{\xi}_j^{(i+1)}$ has unit variance.

These three steps are repeated until the difference between the successive sets of weights falls below a predefined tolerance level. The weights at convergence, $\hat{\boldsymbol{w}}_j$, are then used to compute the composite scores according to Eq. (3).

In the second stage of PLS-PM, OLS regressions are performed using the composite scores in accordance with the specified structural model to estimate the path coefficients. Further methodological details about PLS-PM can be found in Wold (1982), Dijkstra (1985), Tenenhaus et al. (2005) and Schuberth et al. (2018).

3. Automated specification search in SEM

Jöreskog and Sörbom (1993) identified three distinct scenarios when applying SEM: (i) the strictly confirmatory scenario, where a single model is specified and either retained or rejected; (ii) the alternative or competing models scenario, where multiple models are specified and one is selected based on a predefined comparison criterion; and (iii) the model-generating scenario, where an initial model is specified and iteratively revised when it fails to adequately fit the data. The strictly confirmatory scenario is rarely encountered in practice (Marcoulides et al., 1998). Instead, researchers often investigate different model specifications (Gefen et al., 2011; Rigdon et al., 2017; Henseler and Schuberth, 2020). These alternative specifications may arise from searching for alternative or more parsimonious explanations of the phenomenon under study (Sharma et al., 2019; MacCallum, 1986). Similarly, alternative models are considered when the quality of the originally specified model is deemed unacceptable (Long, 1983). In the SEM literature, the process of finding alternative model specifications is known as *specification search*.

Over the past few decades, various algorithms have been combined with factor-based SEM approaches to automatically search for the "best" model specification. Early work proposed combining approaches to factor-based SEM with simulated annealing (Kirkpatrick et al., 1983; Marcoulides and Drezner, 1999) and Tabu search (Marcoulides et al., 1998; Glover and Laguna, 1998). Additionally, Ant Colony Optimization (ACO) algorithms and GAs have been combined with the ML estimator to perform automated specification searches in the context of factor-based SEM (Marcoulides and Drezner, 2001, 2003; Leite et al., 2008). Furthermore, a penalized ML estimator has been proposed for factor-based SEM, enabling the identification of relevant latent variables (Zhou et al., 2014). More recently, a hybrid approach combining ACO with the Tabu search algorithm has been proposed to improve accuracy and efficiency (Jing et al., 2022). Finally, Bee Swarm Optimization (Akbari et al., 2010) has been proposed for automated specification search (Schroeders et al., 2024).

In contrast, the PLS-PM literature has only proposed rudimentary forms of specification search, such as manually fitting alternative models and comparing their quality (Callaghan et al., 2007; Sharma et al., 2019; Danks et al., 2020; Sharma et al., 2021). So far, no fully automated approach to specification search currently exists in the PLS-PM context.

Depending on the type of research (Henseler, 2018), researchers can choose from a variety of criteria to evaluate and compare the quality of a model estimated by PLS-PM (Sharma et al., 2019; Chin et al., 2020). Common measures include the Stone-Geisser Q^2 (Geisser, 1974; Stone, 1974; Wold, 1985) and the coefficient of determination (R^2, Chin, 1998), which assess the in-sample predictive power of a model estimated by PLS-PM. In addition, the Goodness-of-Fit (GoF) and relative GoF indices have been proposed as additional evaluation tools (Tenenhaus et al., 2004; Henseler and Sarstedt, 2013). The Standardized Root Mean Square Residual (SRMR, Bentler, 1995) is also widely used in the PLS-PM literature to judge the overall fit of an estimated model (Henseler et al., 2014; Dijkstra and Henseler, 2015a; Schuberth et al., 2023a), i.e., the discrepancy between the sample correlation matrix and the model-implied counterpart. The quality of a model can also be compared using information-theoretic model selection criteria, such as the Akaike Information Criterion (AIC, Akaike, 1998), the Bayesian Information Criterion (BIC, Schwarz, 1978), and the Hannan-Quinn criterion (Hannan and Quinn, 1979). Moreover, additional criteria have been proposed to evaluate the out-of-sample predictive power of a model estimated by PLS-PM (Shmueli et al., 2016; Schamberger et al., 2023). A Cross-Validated Predictive Ability Test (CVPAT) has also been suggested to enable comparisons of the out-of-sample predictive performance of alternative models (Liengaard et al., 2021).

Despite the number of criteria that can be used for manual model comparison, PLS-PM is underdeveloped when it comes to automated specification search, i.e., letting an algorithm determine the model specification that enhances model quality.

4. Automated genetic algorithm specification search for partial least squares path modeling

This section introduces AGAS-PLS, a new method that enables researchers using PLS-PM to perform a fully automated structural model specification search. This search is restricted to recursive structural models and the outer model is treated as fixed. Consequently, the observed variables must be assigned to their corresponding constructs before starting the specification search. Similarly, AGAS-PLS limits its search to linear recursive structural models. These search constraints are consistent with practical applications, as researchers usually have prior knowledge of the number of constructs, how constructs relate to their observed variables, and which constructs serve as foundational drivers or antecedents within a theoretical model. Moreover, in recursive structural models such as those considered in AGAS-PLS, at least one exogenous construct must be identified to ensure an acyclic dependency structure.

Among the various approaches to automated specification search presented in Section 3, a GA is selected due to its effectiveness in solving optimization problems characterized by complex, noisy, and high-dimensional search spaces (Almeida and Awruch, 2009; Ruehle, 2017; Hamdia et al., 2021; Castelli and Vanneschi, 2014; Gondro and Kinghorn, 2007; Rashid et al., 2015; Rubio-Largo et al., 2018; Ferreira et al., 2023; Leporati et al., 2023; Nadizar and Pietropolli, 2023).

A typical GA consists of four key components: (i) a *population of candidate solutions*, called individuals, which are commonly represented by chromosomes. Each chromosome is a string of genes, where a gene is often binary (0 or 1), and forms a unique point in the search space; (ii) a *selection operator* based on a specified fitness function that selects a certain number of candidate solutions; (iii) a *crossover operator* that combines the selected candidate solutions, called parents, to generate new candidate solutions, called offspring; and (iv) a *mutation operator* that introduces random variations in the candidate solutions to maintain chromosome diversity and explore new regions of the solution space. In general, GAs evolve populations of candidate solutions iteratively by selecting candidate solutions based on their fitness values, recombining their chromosomes, and applying mutations to maintain diversity. This process balances exploitation (i.e., refining promising solutions) and exploration (i.e., searching for new potential solutions), enabling GAs to efficiently navigate complex, high-dimensional search spaces. Over multiple generations, GAs converge towards optimal or near-optimal solutions, making them particularly effective for optimization problems where traditional analytical methods struggle (Gen and Lin, 2023; Pietropolli et al., 2022; Tonelli et al., 2024; Jorgensen et al., 2024).

In AGAS-PLS, each candidate solution represents the structure of a structural model and constitutes a potential solution to the specification search. The structure of the structural model can be represented by a binary indicator matrix $\mathbf{\Pi}$, which has the same dimensions as the path coefficient matrix $\boldsymbol{B}$. The element π_{jl} of $\mathbf{\Pi}$ is set to 1 if construct l influences construct j, and 0 otherwise. Consequently, rows associated with exogenous constructs contain only zeros. In addition, the diagonal of $\mathbf{\Pi}$ consists of zeros because a construct is not allowed to affect itself. To ensure that AGAS-PLS only searches for recursive structural models based on a given set of exogenous constructs, the algorithm initializes the candidate solutions accordingly. Moreover, AGAS-PLS applies crossover and mutation operators that preserve the predefined set of exogenous constructs during the entire specification search. Furthermore, candidate solutions that fall outside the valid search space, i.e., those that do not satisfy the predefined constraints of the structural model, are penalized with the low fitness value $\bar{f}$ during the evaluation of the model fitness.

Algorithm 1 Automated genetic algorithm specification search for PLS-PM (AGAS-PLS).

Input: Data set $\mathcal{D}$; Binary indicator matrix $\boldsymbol{L}$; Set of constructs that are exogenous $\boldsymbol{\xi}_{exo}$; Population size $\mathcal{P}$; Number of generations G; Mutation probability p_m; Crossover probability p_c; Low fitness value $\bar{f}$
Output: Structure of the structural model $\mathbf{\Pi}^*$
Initialize population $\mathcal{P} \leftarrow \{\mathbf{\Pi}_{10}, \dots, \mathbf{\Pi}_{P0}\}$
for each $\mathbf{\Pi}_{i0} \in \mathcal{P}$ **do**
    Set rows in $\mathbf{\Pi}_{i0}$ of constructs $\in \boldsymbol{\xi}_{exo}$ to zero
    Set diagonal of $\mathbf{\Pi}_{i0}$ to zero
    Compute fitness $f_{i0} \leftarrow$ **Fitness Evaluation**($\mathbf{\Pi}_{i0}$, $\boldsymbol{L}$, $\mathcal{D}$, $\bar{f}$)
end for
for generation $g = 1$ to G **do**
    Selection: Select $\mathcal{P}$ parents using linear rank selection
    Crossover: Apply crossover with probability p_c
    Mutation: Apply mutation with probability p_m, preserving constraints
    for each $\mathbf{\Pi}_{ig} \in \mathcal{P}$ **do**
        Compute fitness $f_{ig} \leftarrow$ **Fitness Evaluation**($\mathbf{\Pi}_{ig}$, $\boldsymbol{L}$, $\mathcal{D}$, $\bar{f}$)
    end for
end for
Return: $\mathbf{\Pi}^*$ corresponding to the highest fitness value

As shown in Algorithm 1, AGAS-PLS begins with the initialization of a population of $\mathcal{P}$ candidate solutions. For this purpose, a set of $\mathcal{P}$ binary indicator matrices $\mathbf{\Pi}_{i0}$, $i = 1, \dots, \mathcal{P}$, are generated randomly. Then, the diagonal elements of each candidate solution are set to zero to ensure that the constructs do not influence themselves. Similarly, the elements in the rows corresponding to the constructs in the set of exogenous constructs $\boldsymbol{\xi}_{exo}$ are set to zero. This ensures that the set of exogenous constructs remains unchanged. The $\mathcal{P}$ candidate solutions ($\mathbf{\Pi}_{i0}$) together with the predefined binary indicator matrix of the outer model $\boldsymbol{L}$, define a set of $\mathcal{P}$ candidate models ($\mathcal{M}_{i0}$), whose fitness is evaluated as described in Algorithm 2. The population of $\mathcal{P}$ initial candidate solutions evolves over G

generations by iteratively applying selection, crossover, and mutation operators. Note that the binary indicator matrix of the outer model $\boldsymbol{L}$ does not change during the specification search and remains the same for all candidate models.

Algorithm 2 Fitness evaluation.

Input: Candidate solution $\boldsymbol{\Pi}_{ig}$; Binary indicator matrix $\boldsymbol{L}$; Dataset $\mathcal{D}$; Low fitness value $\bar{f}$
Output: Fitness value f_{ig}
if any construct is isolated, i.e., entire row and column are zeros **then**
    Repair $\boldsymbol{\Pi}_{ig}$ by setting a random entry to one respecting the specification constraints
end if
Convert $\boldsymbol{\Pi}_{ig}$ to graph and check for cycles using depth-first search
if cycles detected **then**
    $\boldsymbol{\Pi}_{ig}$ is assigned the low fitness value $f_{ig} \leftarrow \bar{f}$
else
    Use $\boldsymbol{\Pi}_{ig}$, $\boldsymbol{L}$ and $\mathcal{D}$ to estimate a model $\mathcal{M}_{ig}$ with PLS-PM
    Check whether PLS-PM algorithm converged
    if no convergence of the PLS-PM algorithm **then**
        $\boldsymbol{\Pi}_{ig}$ is assigned the low fitness value $f_{ig} \leftarrow \bar{f}$
    else
        $\boldsymbol{\Pi}_{ig}$ is assigned the fitness value $f_{ig} \leftarrow -\text{BIC}(\mathcal{M}_{ig})$
    end if
end if
Return f_{ig}

At each generation g, a selection operator forms $\mathcal{P}/2$ pairs of parents candidate solutions. The selection operator follows a linear rank selection: candidate solutions are first ranked according to their fitness values, and higher-ranked solutions are assigned higher selection probabilities to increase their chances of passing on their structure to the next generation. Parent pairs are then drawn with replacement according to these probabilities from the population of $\mathcal{P}$ candidate solutions of generation $g-1$.

Linear rank selection provides a moderate and consistent selection pressure, defined as the degree to which the algorithm favors higher-quality individuals for reproduction, compared to other selection operators such as fitness-proportional methods (Back, 1994). This characteristic helps preserve population diversity while still guiding the search toward better solutions (Shukla et al., 2015). It is important to note that a candidate solution may be selected multiple times by the selection operator and can therefore act as a parent in multiple pairs.

Subsequently, a crossover operator is used to combine the parents of each pair. Specifically, the two parents of a pair exchange portions of their structures with a probability of p_c to generate two new candidate solutions, called offspring. Pairs that do not undergo the crossover operation simply transfer their structure to their offspring. In particular, a single-point crossover is used to combine the parents, i.e., a random point is selected along the parents' structures, and the segments are swapped beyond that point (Umbarkar and Sheth, 2015). Fig. 1 illustrates how the crossover operator works. Since crossover only involves combining existing entries from the parents, each offspring is guaranteed to have zeros on its diagonal. Similarly, the constructs initially defined as exogenous remain exogenous throughout the specification search. At the end of the crossover process, $\mathcal{P}$ offspring are generated, each representing a new candidate solution $\boldsymbol{\Pi}_{ig}$.

Once the $\mathcal{P}$ offspring are obtained, they are mutated with probability p_m. If an offspring $\boldsymbol{\Pi}_{ig}$ mutates, one of its entries is randomly changed from 1 to 0 or vice versa. This process is also known as bit-flip mutation. Mutation introduces variability into the set of candidate solutions, and helps the population of candidate solutions escape local minima that might otherwise lead to suboptimal solutions.

To prevent changes to the set of exogenous constructs and ensure that the constructs do not affect themselves, certain elements of the binary indicator matrix $\boldsymbol{\Pi}_{ig}$ that are constrained to zero are restricted from modification, i.e., the elements on the diagonal and the elements in the rows corresponding to the constructs in the initially specified set of exogenous constructs $\boldsymbol{\xi}_{exo}$.

After applying the crossover and mutation operators, the fitness of the newly generated candidate solutions is evaluated as detailed in Algorithm 2. First, it is verified that each candidate solution, $\boldsymbol{\Pi}_{ig}$, does not correspond to a structural model with isolated constructs, i.e., constructs whose corresponding rows and columns in the matrix $\boldsymbol{\Pi}_{ig}$ contain only zeros. This check is essential because PLS-PM parameter estimates are not identified for models with isolated constructs (Dijkstra, 1985). If isolated constructs are detected, one entry from the corresponding row or column is randomly selected and set to one while adhering to the specification constraints. Subsequently, each candidate solution $\boldsymbol{\Pi}_{ig}$ is assessed for feedback loops, which indicate a non-recursive structural model. For this purpose, the candidate solution $\boldsymbol{\Pi}_{ig}$ is converted to a graph, and the depth-first search algorithm (DFS, Tarjan, 1972) is applied to detect cycles indicating feedback loops. If a cycle is detected, the candidate solution is assigned the low fitness value $\bar{f}$, reducing its likelihood of being selected as a parent in the next generation. Conversely, if no cycles are found, a candidate model, defined by the candidate solution $\boldsymbol{\Pi}_{ig}$ and the predefined binary indicator matrix $\boldsymbol{L}$, is estimated using PLS-PM and the dataset $\mathcal{D}$. If the PLS-PM algorithm failed to converge, the corresponding candidate solution is penalized with the low fitness value, $\bar{f}$; otherwise, the fitness value, f_{ig}, is computed. As a result, each generation ends with a population of $\mathcal{P}$ candidate solutions, ensuring that only those satisfying the following criteria persist: a linear, recursive structural model; the absence of isolated constructs; and adherence to the constraint

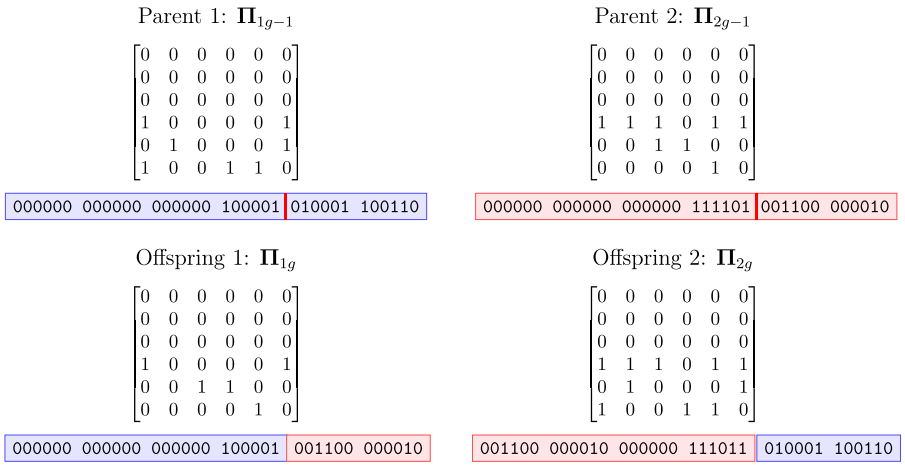


Fig. 1. Example of single-point crossover. Note: The top panel illustrates two parents $\mathbf{\Pi}_{1g-1}$ and $\mathbf{\Pi}_{2g-1}$, and the bottom panel shows their corresponding offspring $\mathbf{\Pi}_{1g}$ and $\mathbf{\Pi}_{2g}$. Below each binary indicator matrix, its vectorized version is shown. The red vertical line marks the swapping point, i.e., the crossover point between the parents' structures. The offspring inherit the same exogenous constructs and preserve zeros along the diagonal, just like their parents. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

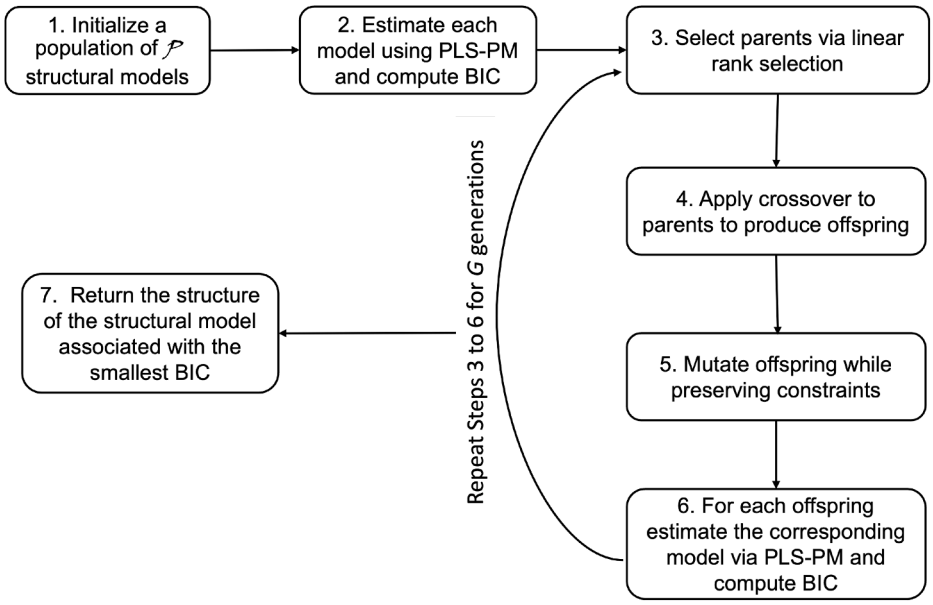


Fig. 2. Simplified illustration of AGAS-PLS detailed in Algorithm 1.

on the set of exogenous constructs. This iterative process fosters progressive improvement in model quality over generations. After G generations, AGAS-PLS returns the candidate solution $\mathbf{\Pi}^*$ associated with the highest fitness value, which determines the final structural model. Fig. 2 provides a simplified illustration of the iterative procedure described in Algorithm 1. It shows how PLS-PM and evolutionary optimization work together throughout the automated search process. The GA generates candidate solutions, and the PLS-PM produces parameter estimates that are used as input for the fitness function, which directs the evolutionary refinement.

As discussed in Section 2, various criteria can be used to assess and compare models estimated by PLS-PM. In principle, all of these criteria qualify as a fitness function in AGAS-PLS. In line with Sharma et al. (2019), a criterion derived from information theory (Akaike, 1981) is used in AGAS-PLS to support the selection of a model that balances parsimony and goodness of fit, as recommended by Myung (2000). Specifically, AGAS-PLS employs the BIC (Schwarz, 1978) and the fitness value for a valid candidate solution $\mathbf{\Pi}_{ig}$ is computed such that models with lower BIC values are favored:

$$f_{ig} = -\text{BIC}(\mathcal{M}_{ig}, D), \tag{5}$$

where $\mathcal{M}_{ig}$ represents a model estimated by PLS-PM. Its structural model is specified by $\mathbf{\Pi}_{ig}$ and its outer model by $\boldsymbol{L}$. To calculate the BIC for each $\mathcal{M}_{ig}$, the formula known from the ML estimator in factor-based SEM is used (e.g., Kline, 2023, Chapter 11): BIC = $q\log(N) - 2\log(L)$, where q is the number of estimated model parameters, N denotes the sample size, and $\log(L)$ denotes the logarithm of the likelihood function.

In the present setting, the logarithm of the likelihood function is given by the following:

$$\log(L) = -N/2(q\log(2\pi) + \mathrm{tr}(\boldsymbol{S}\hat{\boldsymbol{\Sigma}}^{-1}) + \log(|\hat{\boldsymbol{\Sigma}}|)), \tag{6}$$

where tr() denotes the trace operator, $|\boldsymbol{A}|$ denotes the determinant of matrix $\boldsymbol{A}$, and matrix $\hat{\boldsymbol{\Sigma}}$ is the variance-covariance matrix implied by the estimated composite model (Schuberth et al., 2023a).

The use of the BIC as a fitness function is supported by its inherent properties. The BIC is consistent (Wasserman, 2000; Huang, 2017), meaning that as the sample size increases the probability of selecting a model with the same structure as the model of the DGP, if it is among the considered models and ignoring statistical equivalent models, approaches unity. Additionally, the BIC penalizes model complexity and thus effectively limits the selection of saturated models (i.e., models with zero degrees of freedom and thus a perfect overall model fit), favoring more parsimonious models, even if this results in a slightly lower overall model fit. Given that the search space explored by AGAS-PLS always includes a model with the same structure as the model of the DGP, and that the objective is to identify simple, interpretable models with strong generalizability, the use of BIC as the model selection criterion is well-justified.

5. Monte Carlo simulation

5.1. Simulation design

A Monte Carlo simulation was conducted to evaluate the performance of AGAS-PLS. The assessment focuses on AGAS-PLS's ability to recover the structure of the structural model of the DGP, i.e., the population model from which a specific dataset was drawn. For this purpose, several structural models were considered that closely mirror the structure and complexity of those employed in previous PLS-PM simulation studies (e.g., Dijkstra and Henseler, 2015a; Reinartz et al., 2009).

All structural population models comprised six constructs: three exogenous constructs and three endogenous constructs. Each construct was modeled as a composite with three observed variables. The weights to form the six composites were specified as follows: $\boldsymbol{w}_1' = (0.6 \quad 0.4 \quad 0.2)$, $\boldsymbol{w}_2' = (0.3 \quad 0.5 \quad 0.6)$, $\boldsymbol{w}_3' = (0.4 \quad 0.5 \quad 0.5)$, $\boldsymbol{w}_4' = (0.4 \quad 0.2 \quad 0.6)$, $\boldsymbol{w}_5' = (0.4 \quad 0.4 \quad 0.5)$, and $\boldsymbol{w}_6' = (0.7 \quad 0.6 \quad 0.5)$. The correlations between the observed variables forming a composite ranged from 0.15 to 0.80.

To account for varying levels of complexity, a full factorial design was employed comprising three factors: (i) structural model structure, (ii) effect size, and (iii) sample size. This resulted in 36 simulation conditions: four structural model structures × three effect sizes × three sample sizes. Each simulation factor is described in detail below.

Structural model structure

Four structures of the structural model were considered, representing varying levels of complexity, as illustrated in Fig. 3. As AGAS-PLS only searches for linear recursive structural models, the structures considered in the simulation were limited to this type. Importantly, these structures are representative of real empirical applications, as evidenced by a comprehensive review of PLS-PM usage in marketing research (Sarstedt et al., 2022b). Similarly, the structures align with those found in influential empirical studies, such as the Unified Theory of Acceptance and Use of Technology (UTAUT, Venkatesh et al., 2003; Venkatesh and Bala, 2008) in the IS literature and the American Customer Satisfaction Index (ACSI) model, which is widely used in marketing research (Fornell et al., 1996).

Structure 1 represents a full mediation model (Baron and Kenny, 1986), where the effects of ξ_1, ξ_2, ξ_3 on ξ_5 and ξ_6 are fully mediated by ξ_4 and ξ_5, respectively, see Fig. 3(a). This structure is the sparsest among those considered; accordingly, AGAS-PLS is expected to perform best under this condition. Structures 2 and 3 represent partial mediation models that extend Structure 1 by incorporating additional direct effects. Specifically, in Structure 2, both ξ_1 and ξ_3 have direct effects on ξ_4 and ξ_6, respectively, see Fig. 3(b). Similarly, in Structure 3, ξ_1 and ξ_3 show direct effects on ξ_6, see Fig. 3(c). Structure 4 represents a saturated structural model, i.e., no constraints are imposed on the correlation matrix of the composites, see Fig. 3(d). Structure 4 is expected to represent the most challenging scenario for AGAS-PLS, as a saturated model admits a large number of equivalent models - each with a different structure but an identical fit, thereby making correct identification inherently difficult.

Effect size

Three levels of effect size were considered to account for the varying strengths of relations between composites. In particular, for each structural model structure, the population path coefficients were varied. The path coefficients were chosen to correspond to small ($0.02 \leq f^2 < 0.15$), medium ($0.15 \leq f^2 < 0.35$) and large ($f^2 \geq 0.35$) effect sizes, as defined by Cohen (1992). Cohen's f^2 was chosen as the effect size measure because it is commonly used in the PLS-PM context (e.g., Benitez et al., 2020). Table A.6 in the Appendix shows the population path coefficients, including the effect sizes.

Sample size

Finally, to assess the performance of AGAS-PLS under different sample sizes, the sample size N was varied across 100, 250, to 500 observations. These sample sizes reflect those commonly encountered in empirical studies using PLS-PM (e.g., Ringle et al., 2012).

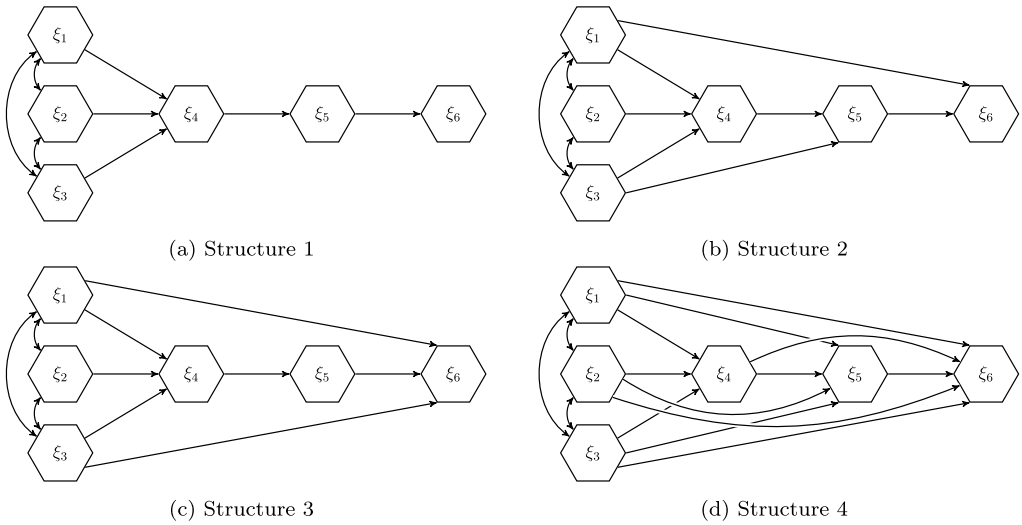


Fig. 3. Different structural model structures. Note: Single-headed arrows represent regression coefficients and double-headed arrows represent correlations.

5.2. Data generation and software used

For each of the 36 simulation conditions, 200 simulation runs were performed, resulting in a total of 7200 datasets. In each run, a sample was drawn from a multivariate normal distribution with a mean of zero and a variance-covariance matrix implied by the population model. The population variance-covariance matrices of the observed variables were generated following the procedure proposed by Dijkstra (2017).

For the GA, a population size of $\mathcal{P} = 100$ was employed, and the population was evolved over $G = 200$ generations. Moreover, consistent with conventional settings in binary genetic algorithms (Back, 1996; Holland, 1992), the crossover and mutation probabilities were set to $p_c = 0.8$ and $p_m = 0.2$, respectively. Preliminary experiments confirmed that these values yield stable performance under the considered problem setting. Moreover, the lowest fitness value $\bar{f}$ was set to -10^5. Finally, the composites ξ_1, ξ_2, and ξ_3 were assigned to the set of exogenous composites, $\boldsymbol{\xi}_{exo}$.

The complete simulation study was carried out in the statistical programming environment R (Version 4.3.0 Team, 2023). The R package cSEM.DGP (Schamberger, 2023) was used to generate the samples. AGAS-PLS was implemented using user-written functions based on the cSEM package (Rademaker and Schuberth, 2020) for estimating models via PLS-PM and the GA package (Scrucca, 2013) for designing the genetic algorithm. The default cSEM settings for PLS-PM were adopted: Mode B was used to estimate the weights, and the path weighting scheme was used to compute the inner weights. The complete R code is available at: https://osf.io/2nqze/.

5.3. Results

The performance of AGAS-PLS was evaluated using the following five key metrics: (i) *Accuracy Rate* (Acc. R.); (ii) *Parsimony Rate* (Pars. R.); (iii) *Success Rate* (Succ. R.); (iv) *Incorrect Rate* (Inc. R.); and (v) *Path Selection Frequency*. Additionally, Table 1 reports the median computation time (in seconds) for each simulation condition. All experiments were run sequentially on a single CPU core of a MacBook Air equipped with an Apple M2 processor and 8 GB of RAM. No GPU acceleration was used. Computational times ranged from 305 seconds to 507 seconds, depending on sample size, effect size, and model complexity.

Accuracy rate

The accuracy rate is defined as the proportion of cases in which AGAS-PLS correctly identifies the structure of the structural model of the DGP. According to Table 1, the performance of AGAS-PLS with respect to the accuracy rate clearly depended on the effect size and complexity of the structural model, whereas the sample size seemed to play only a minor role. For the simplest structural model structure, i.e., Structure 1, the accuracy rate was consistently high across all conditions, reaching up to 0.87 for large effect sizes. For more complex models, i.e., Structures 2 and 3, the accuracy rate depended heavily on the effect size. In these cases, AGAS-PLS accurately identified the structure only for medium and large effect sizes, achieving accuracy rates between 0.57 and 0.77 for models with a medium effect size and between 0.77 and 0.88 for those with a large effect size. In contrast, AGAS-PLS struggled to correctly identify the structures with a small effect size, with accuracy rates below 0.35. As expected, AGAS-PLS was unable to identify the structure of a saturated structural model, i.e., Structure 4.

Table 1
Performance metrics across sample sizes, structural models, and effect sizes. The final column reports the AGAS-PLS computation time (in seconds - median values on 200 runs).

Sample	Structure	Effect Size	Acc. R.	Pars. R.	Succ. R.	Inc. R.	Time (s)
100	Structure 1	Low	0.63	0.00	0.63	0.37	486.25
		Medium	0.76	0.00	0.76	0.24	370.78
		High	0.86	0.00	0.90	0.10	334.64
	Structure 2	Low	0.20	0.08	0.28	0.72	378.74
		Medium	0.77	0.01	0.78	0.22	350.75
		High	0.81	0.00	0.81	0.19	304.49
	Structure 3	Low	0.12	0.43	0.55	0.45	401.87
		Medium	0.56	0.00	0.56	0.44	333.27
		High	0.88	0.00	0.88	0.11	352.53
	Structure 4	Low	0.00	0.01	0.01	0.99	378.86
		Medium	0.00	0.00	0.00	1.00	341.73
		High	0.00	0.44	0.44	0.56	325.59
250	Structure 1	Low	0.63	0.00	0.63	0.37	506.00
		Medium	0.75	0.00	0.75	0.25	368.64
		High	0.87	0.00	0.87	0.13	335.37
	Structure 2	Low	0.27	0.09	0.36	0.64	381.67
		Medium	0.75	0.01	0.76	0.24	391.25
		High	0.86	0.00	0.86	0.14	346.29
	Structure 3	Low	0.09	0.36	0.45	0.55	427.90
		Medium	0.56	0.00	0.56	0.44	398.47
		High	0.83	0.00	0.83	0.17	353.36
	Structure 4	Low	0.00	0.03	0.03	0.97	341.91
		Medium	0.00	0.00	0.00	1.00	347.52
		High	0.02	0.45	0.47	0.53	316.89
500	Structure 1	Low	0.65	0.00	0.65	0.35	507.35
		Medium	0.80	0.00	0.80	0.20	361.86
		High	0.86	0.00	0.86	0.14	348.23
	Structure 2	Low	0.35	0.07	0.42	0.58	402.69
		Medium	0.70	0.00	0.70	0.30	356.36
		High	0.82	0.00	0.82	0.18	336.44
	Structure 3	Low	0.08	0.40	0.48	0.52	431.81
		Medium	0.60	0.0	0.60	0.40	393.50
		High	0.77	0.00	0.77	0.23	352.31
	Structure 4	Low	0.00	0.03	0.03	0.97	370.58
		Medium	0.00	0.00	0.00	1.00	345.24
		High	0.02	0.43	0.43	0.57	321.21

Parsimony rate

The parsimony rate is the percentage of cases in which AGAS-PLS identified a structural model more parsimonious than that of the DGP, i.e., a model retaining only a subset of the paths present in the model of the DGP. A parsimonious model can be considered a simplified version of the model of the DGP. Although these solutions do not fully represent the true model, they can still offer a meaningful approximation of the underlying structure. Moreover, parsimonious models that fit the data well tend to be more scientifically replicable, easier to interpret, and often exhibit higher predictive accuracy (Bentler and Mooijaart, 1989; Shmueli and Koppius, 2011). Similar to the accuracy rate, the parsimony rate was largely invariant to the sample size. Parsimonious solutions were rarely identified for Structures 1 and 2, where the structural model of the DGP is already quite parsimonious. For Structure 3 and a low effect size, the parsimony rate ranged from 0.36 to 0.43, suggesting that AGAS-PLS frequently opted for more parsimonious models when effect sizes were small. Finally, for Structure 4 (the saturated model) with large effect sizes, AGAS-PLS achieved a parsimony rate of 0.45, indicating that in nearly half of the runs the algorithm selected a structural model that was simpler than the fully saturated structure. This behavior was partly driven by the specification of the exogenous constructs, which restricted the search space and thereby encouraged the algorithm to favor more parsimonious structural models over the fully saturated one.

Success rate

Following Sharma et al. (2019), the success rate of AGAS-PLS was computed for each simulation condition as the combined percentage of correctly identified models (Acc. R.) and parsimonious models (Pars. R.). These rates reflect the ability of AGAS-PLS to

Table 2
Path frequency in selected models over 200 runs for Structure 1.

Path	True	n=100			n=250			n=500		
		Small	Med	High	Small	Med	High	Small	Med	High
$\xi_1 \rightarrow \xi_4$	1.00	0.99	1.00	1.00	0.99	1.00	1.00	0.98	1.00	1.00
$\xi_2 \rightarrow \xi_4$	1.00	0.96	1.00	1.00	0.96	1.00	1.00	0.98	1.00	1.00
$\xi_3 \rightarrow \xi_4$	1.00	1.00	1.00	1.00	0.98	1.00	1.00	0.99	1.00	1.00
$\xi_5 \rightarrow \xi_4$	0.00	0.07	0.04	0.00	0.09	0.02	0.00	0.07	0.01	0.00
$\xi_6 \rightarrow \xi_4$	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00
$\xi_1 \rightarrow \xi_5$	0.00	0.04	0.06	0.00	0.04	0.03	0.01	0.04	0.02	0.03
$\xi_2 \rightarrow \xi_5$	0.00	0.06	0.03	0.03	0.07	0.03	0.03	0.04	0.04	0.01
$\xi_3 \rightarrow \xi_5$	0.00	0.04	0.01	0.04	0.04	0.01	0.01	0.04	0.04	0.03
$\xi_4 \rightarrow \xi_5$	1.00	0.90	0.96	1.00	0.88	0.96	1.00	0.91	0.98	1.00
$\xi_6 \rightarrow \xi_5$	0.00	0.11	0.07	0.02	0.12	0.08	0.01	0.11	0.05	0.03
$\xi_1 \rightarrow \xi_6$	0.00	0.07	0.06	0.01	0.07	0.06	0.03	0.06	0.05	0.02
$\xi_2 \rightarrow \xi_6$	0.00	0.08	0.04	0.04	0.05	0.06	0.03	0.10	0.03	0.02
$\xi_3 \rightarrow \xi_6$	0.00	0.09	0.06	0.04	0.09	0.07	0.02	0.06	0.03	0.01
$\xi_4 \rightarrow \xi_6$	0.00	0.04	0.03	0.01	0.07	0.04	0.01	0.04	0.02	0.04
$\xi_5 \rightarrow \xi_6$	1.00	0.82	0.92	0.98	0.80	0.91	0.99	0.82	0.95	0.96

identify the structural paths present in the model of the DGP. Success rates were generally high, particularly for simpler models and conditions with large and medium effect sizes. As with the accuracy and parsimony rates, the success rate was hardly affected by the sample size. Moreover, the success rate was low for Structure 4 (the saturated model), except in the case of a large effect size.

Incorrect rate

The incorrect rate is the percentage of cases in which AGAS-PLS identifies a model that includes at least one path not present in the model of the DGP. Consequently, the incorrect rate is equal to one minus the success rate. As shown in Table 1, the incorrect rate decreased as the effect size increased and reached its highest values when attempting to identify the structure of a saturated structural model, i.e., Structure 4.

Path selection frequency

Finally, the frequency with which AGAS-PLS identified paths between composites, including those not present in the model of the DGP, was examined. Tables 2–5 show the detailed results. A value of 1 in the "True" column indicates that the corresponding path was present in the structural model of the DGP. Similarly, a value of 1 in the other columns indicates that a specific path appeared in all the structural models identified by AGAS-PLS. The values reported in these tables are interpreted as the probability that a given path was included in the model selected by AGAS-PLS. Note that models containing paths pointing to ξ_1, ξ_2, and ξ_3 were excluded from the search space, as these constructs were defined as exogenous and therefore could not have incoming paths.

For Structure 1, the probability that AGAS-PLS selected a path that is present in the model of the DGP remained stable across all sample sizes and effect sizes. The paths present in the model of the DGP were almost always identified by AGAS-PLS. The lowest selection rate for these paths was 0.80, which was observed for the path from ξ_5 to ξ_6 in the case of a small sample size and a small effect size. As with the other performance metrics, the sample size did not affect the ability of AGAS-PLS to correctly identify a path present in the model of the DGP. In the case of Structure 2, i.e., the model that includes partial mediation effects, AGAS-PLS correctly identified most paths in more than 82% of the simulation runs. However, in the case of a small effect size, AGAS-PLS incorrectly identified a few paths (i.e., $\xi_5 \rightarrow \xi_4$ and $\xi_6 \rightarrow \xi_5$). This shows that AGAS-PLS is not always capable of identifying the correct direction of relations between two composites. However, when comparing the percentage of reversed relations to the percentage of correctly identified relations for medium to large sample sizes, it becomes evident that AGAS-PLS tended to favor the directions specified in the model of the DGP. This behavior was less pronounced in small sample sizes. A similar pattern was observed for Structure 3. Finally, as expected, when considering Structure 4, AGAS-PLS was unable to correctly identify the paths present in the model of the DGP, except for those originating from the set of exogenous components to ξ_6. These results indicate that, as expected, when the structural model was saturated, AGAS-PLS was unable to discriminate among equivalent models and tended to select incorrect paths more frequently.

5.4. Sensitivity analysis

To gain further insight into the computational behavior of AGAS-PLS, a sensitivity analysis was conducted, focusing on two key GA tuning parameters: the population size and the number of generations. Specifically, two population sizes, $\mathcal{P} \in 50, 100$, were considered and the number of generations was varied as $G \in 50, 100, 200, 300$. All other GA tuning parameters were kept identical to those used in the Monte Carlo simulation study.

Table 3
Path frequency in selected models over 200 runs for Structure 2.

Path	True	n=100			n=250			n=500		
		Small	Med	High	Small	Med	High	Small	Med	High
$\xi_1 \to \xi_4$	1.00	0.98	0.98	0.98	1.00	1.00	1.00	1.00	1.00	0.99
$\xi_2 \to \xi_4$	1.00	0.95	0.98	0.98	0.92	1.00	1.00	0.95	1.00	0.99
$\xi_3 \to \xi_4$	1.00	0.65	0.91	0.93	0.64	0.94	0.96	0.73	0.90	0.94
$\xi_5 \to \xi_4$	0.00	0.52	0.16	0.05	0.43	0.12	0.05	0.40	0.18	0.07
$\xi_6 \to \xi_4$	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00
$\xi_1 \to \xi_5$	0.00	0.04	0.04	0.09	0.06	0.03	0.01	0.03	0.04	0.06
$\xi_2 \to \xi_5$	0.00	0.09	0.06	0.06	0.09	0.03	0.05	0.06	0.07	0.07
$\xi_3 \to \xi_5$	1.00	0.92	1.00	1.00	0.92	0.99	1.00	0.96	1.00	1.00
$\xi_4 \to \xi_5$	1.00	0.48	0.84	0.95	0.56	0.88	0.95	0.60	0.82	0.92
$\xi_6 \to \xi_5$	0.00	0.38	0.09	0.04	0.34	0.14	0.04	0.32	0.13	0.03
$\xi_1 \to \xi_6$	1.00	0.97	1.00	1.00	0.96	1.00	1.00	0.96	1.00	1.00
$\xi_2 \to \xi_6$	0.00	0.04	0.02	0.00	0.10	0.04	0.02	0.04	0.06	0.03
$\xi_3 \to \xi_6$	0.00	0.09	0.06	0.06	0.07	0.06	0.06	0.06	0.07	0.05
$\xi_4 \to \xi_6$	0.00	0.03	0.02	0.03	0.03	0.00	0.04	0.04	0.01	0.02
$\xi_5 \to \xi_6$	1.00	0.50	0.91	0.96	0.58	0.86	0.96	0.61	0.87	0.97

Table 4
Path frequency in selected models over 200 runs for Structure 3.

Path	True	n=100			n=250			n=500		
		Small	Med	High	Small	Med	High	Small	Med	High
$\xi_1 \to \xi_4$	1.00	0.98	1.00	1.00	0.97	1.00	0.99	1.00	1.00	1.00
$\xi_2 \to \xi_4$	1.00	0.94	1.00	1.00	0.98	1.00	0.99	0.95	1.00	1.00
$\xi_3 \to \xi_4$	1.00	0.98	1.00	1.00	0.94	0.99	0.99	0.95	1.00	1.00
$\xi_5 \to \xi_4$	0.00	0.17	0.28	0.06	0.30	0.24	0.03	0.22	0.28	0.08
$\xi_6 \to \xi_4$	0.00	0.04	0.02	0.00	0.03	0.03	0.00	0.01	0.01	0.00
$\xi_1 \to \xi_5$	0.00	0.05	0.05	0.03	0.00	0.01	0.05	0.00	0.04	0.01
$\xi_2 \to \xi_5$	0.00	0.04	0.05	0.01	0.03	0.05	0.06	0.08	0.03	0.06
$\xi_3 \to \xi_5$	0.00	0.03	0.08	0.03	0.04	0.06	0.03	0.03	0.03	0.08
$\xi_4 \to \xi_5$	1.0	0.80	0.68	0.94	0.69	0.70	0.95	0.76	0.67	0.92
$\xi_6 \to \xi_5$	0.00	0.36	0.32	0.00	0.45	0.33	0.05	0.40	0.33	0.07
$\xi_1 \to \xi_6$	1.00	0.74	0.99	1.00	0.71	1.00	1.00	0.72	1.00	1.00
$\xi_2 \to \xi_6$	0.00	0.07	0.03	0.04	0.03	0.04	0.01	0.08	0.01	0.04
$\xi_3 \to \xi_6$	1.00	0.54	0.99	1.00	0.57	1.00	1.00	0.53	1.00	1.00
$\xi_4 \to \xi_6$	0.00	0.01	0.05	0.01	0.03	0.03	0.03	0.03	0.01	0.01
$\xi_5 \to \xi_6$	1.00	0.57	0.68	0.96	0.46	0.67	0.95	0.50	0.67	0.93

Fig. 4 illustrates the effect of $\mathcal{P}$ and G on the success rate (top row) and computation time (bottom row) across three effect size conditions (small, medium, and large). Each condition is based on 30 independent runs and was evaluated for Structure 1 with a sample size of 100, i.e., under the most challenging simulation scenario. The reported computation times correspond to the median wall-clock time (in seconds) required for a full sequential execution of AGAS-PLS on a single CPU core.

As expected, increasing the number of generations G led to higher success rates, as the algorithm was allowed to explore the search space more thoroughly. However, this improvement came at the cost of substantially increased computation time, since more evolutionary iterations had to be performed. Similarly, increasing the population size ($\mathcal{P}$) from 50 to 100 improved performance by allowing each generation to explore a broader range of candidate solutions. This also approximately doubled the computation time because twice as many individuals had to be evaluated.

Interestingly, the performance gains from increasing the number of generations (G) beyond 100 were relatively modest. In particular, increasing G from 100 to 300 only marginally improved the success rate, while nearly tripling the computation time. This suggests diminishing returns when extending the number of generations beyond moderate values. Similarly, although $\mathcal{P} = 100$ consistently outperformed $\mathcal{P} = 50$, this advantage became less significant when considering the associated computational cost for large numbers of generations.

Overall, these results suggest that choosing $\mathcal{P} = 100$ and $G = 100$ provides a balanced and computationally efficient configuration, offering a favorable trade-off between runtime and success rate when computational resources or time are limited. However, when a

Table 5
Path frequency in selected models over 200 runs for Structure 4.

Path	True	n=100			n=250			n=500		
		Small	Med	High	Small	Med	High	Small	Med	High
$\xi_1 \to \xi_4$	1.00	0.59	0.06	0.34	0.56	0.03	0.30	0.56	0.05	0.26
$\xi_2 \to \xi_4$	1.00	0.22	0.03	0.28	0.14	0.00	0.28	0.13	0.01	0.22
$\xi_3 \to \xi_4$	1.00	0.40	0.06	0.27	0.42	0.03	0.26	0.42	0.04	0.23
$\xi_5 \to \xi_4$	0.00	0.55	0.27	0.38	0.56	0.24	0.34	0.54	0.28	0.36
$\xi_6 \to \xi_4$	0.00	0.89	0.96	0.20	0.88	0.98	0.20	0.87	0.94	0.16
$\xi_1 \to \xi_5$	1.00	0.33	0.27	0.44	0.28	0.26	0.42	0.28	0.28	0.48
$\xi_2 \to \xi_5$	1.00	0.36	0.56	0.40	0.44	0.49	0.40	0.44	0.50	0.42
$\xi_3 \to \xi_5$	1.00	0.62	0.43	0.42	0.74	0.34	0.41	0.74	0.38	0.46
$\xi_4 \to \xi_5$	1.00	0.43	0.46	0.22	0.40	0.50	0.22	0.41	0.46	0.24
$\xi_6 \to \xi_5$	0.00	0.91	0.90	0.32	0.85	0.90	0.29	0.84	0.87	0.32
$\xi_1 \to \xi_6$	1.00	0.98	0.98	0.95	0.96	0.99	0.92	0.96	0.98	0.92
$\xi_2 \to \xi_6$	1.00	0.98	0.98	0.95	0.98	0.99	0.92	0.98	0.98	0.92
$\xi_3 \to \xi_6$	1.00	0.94	0.98	0.95	0.89	0.99	0.92	0.88	0.98	0.92
$\xi_4 \to \xi_6$	1.00	0.06	0.04	0.80	0.07	0.03	0.74	0.08	0.06	0.77
$\xi_5 \to \xi_6$	1.00	0.07	0.10	0.72	0.12	0.10	0.71	0.12	0.12	0.69

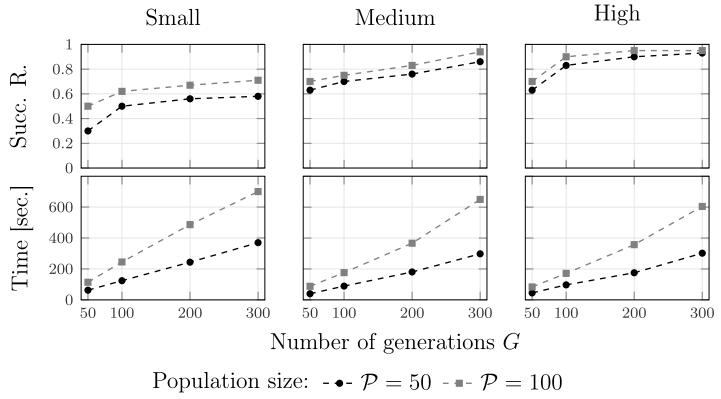


Fig. 4. Results of the sensitivity analysis. Note: The top row shows the success rate and the bottom row the median computation time (in seconds) for Structure 1, evaluated with a sample size of 100 observations across the three effect sizes (small, medium, and high). Each point corresponds to the success rate and computation time, respectively, obtained for a specific number of generations, G. Results are based on 30 runs.

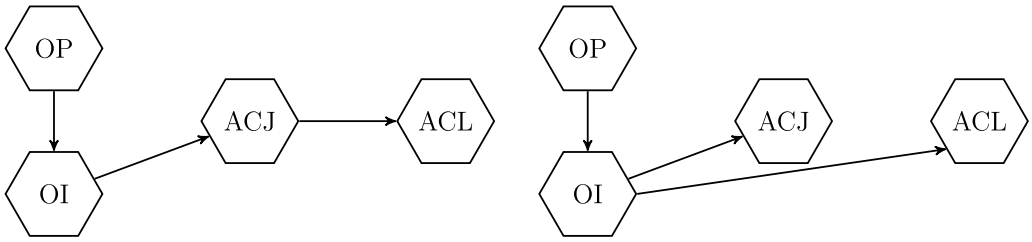
slight gain in performance justifies additional computational effort, increasing the number of generations to $G = 200$ may be preferable. Beyond $G = 200$, increases in the number of generations did not lead to substantial improvements in AGAS-PLS's performance, despite substantially increasing computation time; therefore, this should be avoided in practice.

Finally, consistent with the findings of the Monte Carlo study, success rates were higher for models with larger effect sizes. Interestingly, larger effect sizes were also associated with slightly lower computation times. One possible explanation is that stronger structural relations enabled AGAS-PLS to identify the optimal model structure (in terms of a small BIC) earlier in the evolutionary process, thereby reducing the time spent on repairing invalid individuals (see Line 4 of Algorithm 2).

6. Empirical illustration

To illustrate the practical application of AGAS-PLS, a dataset originally collected by Bergami and Bagozzi (2000) was analyzed. This dataset comprises responses from 157 male and 148 female employees working in the electronics division of a large South Korean conglomerate and is publicly available at https://www.smartpls.com/documentation/sample-projects/organizational-identification-variations-pls. Following Hwang and Takane (2004), the analysis focused on the male employees and the 21 items associated with four constructs: organizational prestige (OP) formed by eight items, organizational identification (OI) formed by six items, affective commitment-joy (ACJ) formed by four items, and affective commitment-love (ACL) formed by three items.

For AGAS-PLS, the same settings as in the Monte Carlo simulation were adopted. Specifically, $G = 200$ generations and a population size of $\mathcal{P} = 100$ candidate solutions were used, with crossover and mutation probabilities set to $p_c = 0.8$ and $p_m = 0.2$, respectively.



(a) Structure proposed by AGAS-PLS (b) Hwang and Takane (2004) structure

Fig. 5. AGAS-PLS applied to the Organizational Identification dataset collected by Bergami and Bagozzi (2000).

Mode B was employed for weight estimation, and the path weighting scheme was used to calculate the inner weights. Furthermore, consistent with previous studies (Bergami and Bagozzi, 2000; Hwang and Takane, 2004), OP was specified as an exogenous construct. To perform the AGAS-PLS, we used the `doModelSearch()` function of the cSEM R package (Rademaker and Schuberth, 2020). The R code is available at: https://osf.io/2nqze/. The AGAS-PLS specification search required 167.53 seconds of computation on a single CPU core of a MacBook Air equipped with an Apple M2 processor and 8 GB of RAM. No GPU acceleration was used.

Fig. 5 compares the original structural model proposed by Hwang and Takane (2004) with the structural model recovered by AGAS-PLS. The original model (right) yields a BIC of 7836.36, while AGAS-PLS selected an alternative specification (left) with a lower BIC of 7806.50. In the original model, OP is the only exogenous construct, and its effect on ACJ is fully mediated by OI. Similarly, in the structural model identified by AGAS-PLS, OP is the only exogenous construct and reveals a direct effect of OI on ACJ. Moreover, the original model shows a direct effect of OI on ACL. In contrast, in the model identified by AGAS-PLS, the direct effect of OI on ACL is absent. Instead, the effect of OI on ACL is fully mediated by ACJ.

7. Conclusion and future research

AGAS-PLS was introduced as a new tool for automated specification search in composite-based SEM, and an implementation is provided in the R package cSEM (Rademaker and Schuberth, 2020). The Monte Carlo simulation results demonstrated the effectiveness of AGAS-PLS in identifying the structure of the structural model of the DGP under different conditions, including varying sample size and levels of model complexity. In particular, AGAS-PLS demonstrated high accuracy in detecting the true structure of the structural model. Larger effect sizes were found to enhance accuracy, whereas increased model complexity reduced it. These findings were independent of sample size: AGAS-PLS successfully identified the structure of the structural model of the DGP for models including full and partial mediation with medium to large effect sizes, even with a small sample size of 100 observations. The observed effect of sample size is consistent with the results reported by Sharma et al. (2019).

As model complexity increased, AGAS-PLS tended to identify more parsimonious models, particularly when effect sizes were small. This behavior compensated for its limited ability to accurately recover highly complex structural model structures. The tendency of AGAS-PLS to favor more parsimonious solutions in complex models when effect sizes were small, together with its ability to correctly identify most structural paths in simpler scenarios, further highlighted its potential to recover the true structural model (i.e., the structural model of the DGP).

The presence of equivalent models is a well-documented issue in the SEM literature (see Raykov and Penev, 1999, for a discussion on this topic). Equivalent models are alternative models with the same number of constructs, degrees of freedom, and model fit, but with a different structure. Since AGAS-PLS, like other automated specification search procedures, is driven by model fit criteria, it cannot differentiate between the structural model of the DGP and misspecified models that fit the data equivalently well. However, providing a set of exogenous constructs arguably helped AGAS-PLS to identify more parsimonious models, i.e., models that are nested within the model of the DGP.

Although tools like AGAS-PLS can greatly facilitate and streamline the model selection process, researchers should view them as aids for narrowing down potential models to a “shortlist” of candidates (Marcoulides and Drezner, 2003). Any structural model identified through automated specification searches must undergo replication or cross-validation before drawing final conclusions about its validity. This ensures the robustness and generalizability of the findings, maintaining scientific rigor in empirical research.

Several limitations of AGAS-PLS offer promising avenues for future research. First, AGAS-PLS requires defining at least one exogenous construct. Although this effectively helps identify recursive structural models and also aligns with many real-world applications, it may limit the applicability of AGAS-PLS in scenarios where it is unclear which constructs are exogenous. Moreover, AGAS-PLS is currently limited to linear, recursive structural models. Expanding the specification search to non-recursive models, e.g., models with reciprocal effects, would make AGAS-PLS applicable to more complex model structures. Furthermore, AGAS-PLS can currently only be used to recover the structure of the structural model. Therefore, the proposed approach could be extended to enable automated specification searches at the level of the outer model, i.e., selecting the relevant observed variables that make up the composites.

Moreover, AGAS-PLS relies on PLS-PM, which produces inconsistent parameter estimates for models involving common factors (Hui and Wold, 1982; Schuberth et al., 2024). Therefore, AGAS-PLS could be extended by using Consistent Partial Least Squares (PLSc), which enables the estimation of models incorporating both composites and common factors (see, e.g., Dijkstra and Henseler, 2015a,b;

Rademaker et al., 2019; Schuberth et al., 2023b). Furthermore, AGAS-PLS currently explores all admissible linear, recursive structural models. However, in many applications, certain structural relations are either theoretically justified or theoretically implausible. Therefore, another promising direction for future research is to incorporate prior substantive knowledge directly into the specification search. This information could be used to constrain selected elements of the structural matrix (i.e., set to zero or restricted to be nonzero) before the search begins. This would reduce the search space, decrease computational time, and ensure that the resulting model remains consistent with established theory or domain expertise.

Finally, future research could explore the use of alternative fitness functions, such as those focused on out-of-sample predictive performance, as proposed in the context of GSCA (Cho et al., 2022b). Moreover, the performance of AGAS-PLS when the set of exogenous constructs and/or the outer model structure are incorrectly specified should be studied. Addressing these limitations and pursuing these research avenues will strengthen the position of AGAS-PLS as a flexible and robust tool for specification searches in composite-based SEM, broadening its applicability in empirical research.

CRediT authorship contribution statement

Laura Trinchera: Conceptualization, Data curation, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing; **Gloria Pietropolli:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Mauro Castelli:** Methodology, Resources, Software, Supervision, Validation, Writing – original draft, Writing – review & editing; **Florian Schuberth:** Conceptualization, Investigation, Methodology, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

Portions of this manuscript were prepared with the assistance of OpenAI's GPT-5 language model (ChatGPT, version 2025), Google's Gemini, and DeepL Write. The authors used these AI tools to improve the clarity, structure, and style of the text. The authors carefully reviewed, edited, and verified all AI-generated content to ensure its accuracy, originality, and compliance with scholarly and ethical standards. No generative AI tools were used for conceptualization, literature review, data analysis, statistical modeling, or the generation or interpretation of research results. The authors take full responsibility for the content of the manuscript and the views, opinions, and conclusions presented in this paper are solely those of the authors.

Acknowledgment

L. Trinchera gratefully acknowledges the support of the Area of Excellence in AI, Data Science & Business at NEOMA Business School. This article is based upon work from COST Action Randomised Optimisation Algorithms Research Network (ROAR-NET), CA22137, supported by COST (European Cooperation in Science and Technology). This work was partially supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project - UID/04152/2025 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS - https://doi.org/10.54499/UID/04152/2025 (2025-01-01/2028-12-31) and UID/PRR/04152/2025. https://doi.org/10.54499/UID/PRR/04152/2025 (2025-01-01/ 2026-06-30)

Appendix A.

Table A.6
Path coefficients, including effect sizes, used in the Monte Carlo simulation.

	Structure 1						Structure 2					
	Small		Medium		Large		Small		Medium		Large	
Path	β	f^2	β	f^2	β	f^2	β	f^2	β	f^2	β	f^2
$\xi_1 \rightarrow \xi_4$	0.30	0.11	0.38	0.29	0.30	0.51	0.30	0.10	0.35	0.21	0.36	0.37
$\xi_2 \rightarrow \xi_4$	0.20	0.05	0.35	0.28	0.50	1.59	0.20	0.05	0.30	0.17	0.38	0.46
$\xi_3 \rightarrow \xi_4$	0.25	0.06	0.30	0.17	0.40	0.84	0.25	0.06	0.32	0.16	0.37	0.36
$\xi_1 \rightarrow \xi_5$	-	-	-	-	-	-	-	-	-	-	-	-
$\xi_2 \rightarrow \xi_5$	-	-	-	-	-	-	-	-	-	-	-	-
$\xi_3 \rightarrow \xi_5$	-	-	-	-	-	-	0.28	0.09	0.36	0.15	0.48	0.42
$\xi_1 \rightarrow \xi_6$	-	-	-	-	-	-	0.30	0.10	0.38	0.19	0.50	0.58
$\xi_2 \rightarrow \xi_6$	-	-	-	-	-	-	-	-	-	-	-	-
$\xi_3 \rightarrow \xi_6$	-	-	-	-	-	-	-	-	-	-	-	-
$\xi_4 \rightarrow \xi_5$	0.35	0.14	0.45	0.25	0.80	1.78	0.35	0.13	0.40	0.19	0.44	0.35
$\xi_4 \rightarrow \xi_6$	-	-	-	-	-	-	-	-	-	-	-	-
$\xi_5 \rightarrow \xi_6$	0.25	0.07	0.40	0.20	0.50	0.33	0.25	0.07	0.34	0.15	0.45	0.47
	Structure 3						Structure 4					
	Small		Medium		Large		Small		Medium		Large	
Path	β	f^2	β	f^2	β	f^2	β	f^2	β	f^2	β	f^2
$\xi_1 \rightarrow \xi_4$	0.30	0.10	0.35	0.21	0.36	0.36	0.30	0.10	0.38	0.29	0.30	0.24
$\xi_2 \rightarrow \xi_4$	0.20	0.05	0.30	0.17	0.38	0.46	0.20	0.05	0.35	0.28	0.40	0.49
$\xi_3 \rightarrow \xi_4$	0.25	0.06	0.32	0.16	0.37	0.36	0.25	0.06	0.30	0.17	0.40	0.40
$\xi_1 \rightarrow \xi_5$	-	-	-	-	-	-	0.15	0.03	0.25	0.20	0.28	0.84
$\xi_2 \rightarrow \xi_5$	-	-	-	-	-	-	0.18	0.05	0.28	0.29	0.35	1.23
$\xi_3 \rightarrow \xi_5$	-	-	-	-	-	-	0.20	0.05	0.25	0.20	0.28	0.69
$\xi_1 \rightarrow \xi_6$	0.20	0.04	0.32	0.17	0.37	0.41	0.20	0.06	0.22	0.27	0.22	7.85
$\xi_2 \rightarrow \xi_6$	-	-	-	-	-	-	0.20	0.07	0.20	0.24	0.20	5.01
$\xi_3 \rightarrow \xi_6$	0.15	0.02	0.33	0.18	0.36	0.39	0.15	0.03	0.20	0.23	0.20	5.80
$\xi_4 \rightarrow \xi_5$	0.35	0.14	0.40	0.19	0.55	0.43	0.35	0.15	0.35	0.25	0.30	0.44
$\xi_4 \rightarrow \xi_6$	-	-	-	-	-	-	0.20	0.05	0.25	0.22	0.25	5.91
$\xi_5 \rightarrow \xi_6$	0.25	0.07	0.35	0.25	0.39	0.53	0.25	0.08	0.28	0.17	0.30	2.50

References

Akaike, H., 1981. Likelihood of a model and information criteria. J. Econom. 16 (1), 3–14.

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle. In: Parzen, E., Tanabe, K., Kitagawa, G. (Eds.), Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, pp. 199–213. https://doi.org/10.1007/978-1-4612-1694-0_15

Akbari, R., Mohammadi, A., Ziarati, K., 2010. A novel bee swarm optimization algorithm for numerical function optimization. Commun. Nonlinear Sci. Numer. Simul. 15 (10), 3142–3155. https://doi.org/10.1016/j.cnsns.2009.11.003

Almeida, F.S., Awruch, A.M., 2009. Design optimization of composite laminated structures using genetic algorithms and finite element analysis. Compos. Struct. 88 (3), 443–454. https://doi.org/10.1016/j.compstruct.2008.05.004

Back, T., 1994. Selective pressure in evolutionary algorithms: a characterization of selection mechanisms. In: Proceedings of the First IEEE Conference on Evolutionary Computation. IEEE World Congress on Computational Intelligence. IEEE, pp. 57–62. https://doi.org/10.1109/icec.1994.350042

Back, T., 1996. Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms. Oxford University Press, Oxford.

Baron, R.M., Kenny, D.A., 1986. The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J. Pers. Soc. Psychol. 51 (6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Benitez, J., Henseler, J., Castillo, A., Schuberth, F., 2020. How to perform and report an impactful analysis using partial least squares: guidelines for confirmatory and explanatory IS research. Inf. Manag. 2 (57), 103168. https://doi.org/10.1016/j.im.2019.05.003

Bentler, P.M., 1995. EQS Structural Equations Program Manual. Vol. 6. Multivariate Software, Inc., Encino, CA.

Bentler, P.M., Mooijaart, A.B., 1989. Choice of structural model via parsimony: a rationale based on precision. Psychol. Bull. 106 (2), 315–317. https://doi.org/10.1037/0033-2909.106.2.315

Bergami, M., Bagozzi, R.P., 2000. Self-categorization, affective commitment and group self-esteem as distinct aspects of social identity in the organization. Br. J. Social Psychol. 39 (4), 555–577. https://doi.org/10.1348/014466600164633

Callaghan, W., Wilson, B., Ringle, C.M., Henseler, J., 2007. Exploring causal path directionality for a marketing model: using Cohen's path method. In: Martens, H., Næs, T. (Eds.), Proceedings of PLS'07 - The 5th International Symposium on PLS and Related Methods. PLS, Norway: Matforsk, As, pp. 57–61.

Castelli, M., Vanneschi, L., 2014. Genetic algorithm with variable neighborhood search for the optimal allocation of goods in shop shelves. Oper. Res. Lett. 42 (5), 355–360. https://doi.org/10.1016/j.orl.2014.06.002

Chin, W., Cheah, J.-H., Liu, Y., Ting, H., Lim, X.-J., Cham, T.H., 2020. Demystifying the role of causal-predictive modeling using partial least squares structural equation modeling in information systems research. Indus. Manag. Data Syst. 120 (12), 2161–2209. https://doi.org/10.1108/imds-10-2019-0529

Chin, W.W., 1998. The partial least squares approach for structural equation modeling. In: Marcoulides, G.A. (Ed.), Modern Methods for Business Research. Psychology Press, Hove, pp. 295–336.

Cho, G., Hwang, H., Sarstedt, M., 2022a. A comparative evaluation of factor- and component-based structural equation modeling approaches under (in)correct construct representations. Br. J. Math. Stat. Psychol. 75 (2), 220–251. https://doi.org/10.1111/bmsp.12255

Cho, G., Hwang, H., Sarstedt, M., Ringle, C.M., 2022b. A prediction-oriented specification search algorithm for generalized structured component analysis. Struc. Eq. Model. Multidiscip. J. 29 (4), 611–619. https://doi.org/10.1080/10705511.2022.2057315

Cohen, J., 1992. A power primer. Psychol. Bull. 112 (1), 155–159. https://doi.org/10.1037//0033-2909.112.1.155
Cook, R.D., Forzani, L., 2023. On the role of partial least squares in path analysis for the social sciences. J. Bus. Res. 167, 114132. https://doi.org/10.1016/j.jbusres.2023.114132
Danks, N.P., Sharma, P.N., Sarstedt, M., 2020. Model selection uncertainty and multimodel inference in partial least squares structural equation modeling (PLS-SEM). J. Bus. Res. 113, 13–24. https://doi.org/10.1016/j.jbusres.2020.03.019
Diamantopoulos, A., Winklhofer, H.M., 2001. Index construction with formative indicators: an alternative to scale development. J. Market. Res. 38 (2), 269–277. https://doi.org/10.1509/jmkr.38.2.269.18845
Dijkstra, T.K., 1985. Latent Variables in Linear Stochastic Models: Reflections on "Maximum Likelihood" and "Partial Least Squares" Methods. Vol. 1. Sociometric Research Foundation, Amsterdam.
Dijkstra, T.K., 2017. A perfect match between a model and a mode. In: Latan, H., Noonan, R. (Eds.), Partial Least Squares Path Modeling: Basic Concepts, Methodological Issues and Applications. Springer, Cham, pp. 55–80.
Dijkstra, T.K., Henseler, J., 2015a. Consistent and asymptotically normal PLS estimators for linear structural equations. Comput. Stat. Data Anal. 81, 10–23. https://doi.org/10.1016/j.csda.2014.07.008
Dijkstra, T.K., Henseler, J., 2015b. Consistent partial least squares path modeling. MIS Quart. 39 (2), 297–316. https://doi.org/10.25300/MISQ/2015/39.2.02
Dijkstra, T.K., Schermelleh-Engel, K., 2014. Consistent partial least squares for nonlinear structural equation models. Psychometrika 79 (4), 585–604. https://doi.org/10.1007/s11336-013-9370-0
Evermann, J., Rönkkö, M., 2023. Recent developments in PLS. Commun. Assoc. Inf. Syst. 52, 663–667. https://doi.org/10.17705/1CAIS.05229
Ferreira, J., Castelli, M., Manzoni, L., Pietropolli, G., 2023. A self-adaptive approach to exploit topological properties of different GAs' crossover operators. In: Pappa, G., Giacobini, M., Vasicek, Z. (Eds.), Genetic Programming. Springer Nature Switzerland, Cham, pp. 3–18. https://doi.org/10.1007/978-3-031-29573-7_1
Fornell, C., Johnson, M.D., Anderson, E.W., Cha, J., Bryant, B.E., 1996. The american customer satisfaction index: nature, purpose, and findings. J. Mark. 60 (4), 7–18. https://doi.org/10.2307/1251898
Gefen, D., Straub, D.W., Rigdon, E.E., 2011. An update and extension to SEM guidelines for administrative and social science research. MIS Quart. 35 (2), iii–xiv. https://doi.org/10.2307/23044042
Geisser, S., 1974. A predictive approach to the random effect model. Biometrika 61 (1), 101–107. https://doi.org/10.1093/biomet/61.1.101
Gen, M., Lin, L., 2023. Genetic algorithms and their applications. In: Pham, H. (Ed.), Springer Handbook of Engineering Statistics. Springer, London, pp. 635–674. https://doi.org/10.1007/978-1-4471-7503-2_33
Glover, F., Laguna, M., 1998. Tabu Search. Springer, New York.
Gondro, C., Kinghorn, B.P., 2007. A simple genetic algorithm for multiple sequence alignment. Gen. Mol. Res. 6 (4), 964–982.
Grace, J.B., Bollen, K.A., 2008. Representing general theoretical concepts in structural equation models: the role of composite variables. Environ. Ecol. Stat. 15 (2), 191–213. https://doi.org/10.1007/s10651-007-0047-7
Hair, J.F., Ringle, C.M., Sarstedt, M., 2011. PLS-SEM: indeed a silver bullet. J. Market. Theory Prac. 19 (2), 139–152. https://doi.org/10.2753/MTP1069-6679190202
Hamdia, K.M., Zhuang, X., Rabczuk, T., 2021. An efficient optimization approach for designing machine learning models based on genetic algorithm. Neural Comput. Appl. 33 (6), 1923–1933. https://doi.org/10.1007/s00521-020-05035-x
Hannan, E.J., Quinn, B.G., 1979. The determination of the order of an autoregression. J. R. Stat. Soc. Ser. B (Methodol.) 41 (2), 190–195. https://doi.org/10.1111/j.2517-6161.1979.tb01072.x
Henseler, J., 2018. Partial least squares path modeling: Quo vadis? Qual. Quant. 52 (1), 1–8. https://doi.org/10.1007/s11135-018-0689-6
Henseler, J., 2021. Composite-Based Structural Equation Modeling: Analyzing Latent and Emergent Variables. Guilford Press, New York, NY.
Henseler, J., Dijkstra, T.K., Sarstedt, M., Ringle, C.M., Diamantopoulos, A., Straub, D.W., Ketchen, D.J., Hair, J.F., Hult, G. T.M., Calantone, R.J., 2014. Common beliefs and reality about PLS: comments on Rönkkö and Evermann (2013). Organ. Res. Methods 17 (2), 182–209. https://doi.org/10.1177/1094428114526928
Henseler, J., Ringle, C.M., Sarstedt, M., 2016. Testing measurement invariance of composites using partial least squares. Int. Market. Rev. 33 (3), 405–431. https://doi.org/10.1108/IMR-09-2014-0304
Henseler, J., Sarstedt, M., 2013. Goodness-of-fit indices for partial least squares path modeling. Comput. Stat. 28 (2), 565–580. https://doi.org/10.1007/s00180-012-0317-1
Henseler, J., Schuberth, F., 2020. Using confirmatory composite analysis to assess emergent variables in business research. J. Bus. Res. 120, 147–156. https://doi.org/10.1016/j.jbusres.2020.07.026
Henseler, J., Schuberth, F., 2021. Auxiliary theories. In: Henseler, J. (Ed.), Composite-based Structural Equation Modeling: Analyzing Latent and Emergent Variables. The Guilford Press, London, New York. chapter 2, pp. 25–37.
Holland, J.H., 1992. Genetic algorithms. Sci. Am. 267 (1), 66–73.
Huang, P.-H., 2017. Asymptotics of AIC, BIC, and RMSEA for model selection in structural equation modeling. Psychometrika 82 (2), 407–426. https://doi.org/10.1007/s11336-017-9572-y
Hui, B.S., Wold, H., 1982. Consistency and consistency at large of partial least squares estimates. In: Jöreskog, K.G., Wold, H. (Eds.), Systems under Indirect Observation: Causality, Structure, Prediction Part II. North-Holland, Amsterdam, pp. 119–130.
Hwang, H., Cho, G., Jung, K., Falk, C., Flake, J., Jin, M.J., Lee, S.H., 2021. An approach to structural equation modeling with both factors and components: integrated generalized structured component analysis. Psychol. Methods 26 (3), 273–294. https://doi.org/10.1037/met0000336
Hwang, H., Takane, Y., 2004. Generalized structured component analysis. Psychometrika 69 (1), 81–99. https://doi.org/10.1007/BF02295841
Jing, Z., Kuang, H., Leite, W.L., Marcoulides, K.M., Fisk, C.L., 2022. Model specification searches in structural equation modeling with a hybrid ant colony optimization algorithm. Struc. Eq. Model. Multidiscip. J. 29 (5), 655–666. https://doi.org/10.1080/10705511.2021.2020119
Jöreskog, K.G., 1970. A general method for estimating a linear structural equation system. ETS Res. Bull. Ser. 1970 (2), i–41. https://doi.org/10.1002/j.2333-8504.1970.tb00783.x
Jöreskog, K.G., Sörbom, D., 1993. Lisrel 8: structural equation modeling with the SIMPLIS command language. Scient. Softw. Int. Lawrence Erlbaum Associates, Inc.
Jorgensen, S., Nadizar, G., Pietropolli, G., Manzoni, L., Medvet, E., O'Reilly, U.-M., Hemberg, E., 2024. Large language model-based test case generation for GP agents. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 914–923. https://doi.org/10.1145/3638529.3654056
Jung, S., Park, J., 2018. Consistent partial least squares path modeling via regularization. Front. Psychol. 9. https://doi.org/10.3389/fpsyg.2018.00174
Khan, G.F., Sarstedt, M., Shiau, W.-L., Hair, J.F., Ringle, C.M., Fritze, M.P., 2019. Methodological research on partial least squares structural equation modeling (PLS-SEM): an analysis based on social network approaches. Internet Res. 29 (3), 407–429. https://doi.org/10.1108/IntR-12-2017-0509
Kirkpatrick, S., Gelatt Jr, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220 (4598), 671–680. https://doi.org/10.1126/science.220.4598.671
Klesel, M., Schuberth, F., Niehaves, B., Henseler, J., 2022. Multigroup analysis in information systems research using PLS-PM: a systematic investigation of approaches. Data Base Adv. Inf. Syst. 53 (3), 26–48. https://doi.org/10.1145/3551783.3551787
Kline, R.B., 2023. Principles and Practice of Structural Equation Modeling. The Guilford Press, New York. 5th ed.
Leite, W.L., Huang, I.-C., Marcoulides, G.A., 2008. Item selection for the development of short forms of scales using an ant colony optimization algorithm. Multivariate Behav. Res. 43 (3), 411–431. https://doi.org/10.1080/00273170802285743
Leporati, A., Manzoni, L., Mauri, G., Pietropolli, G., Zandron, C., 2023. Inferring p systems from their computing steps: an evolutionary approach. Swarm Evol. Comput. 76, 101223. https://doi.org/10.1016/j.swevo.2022.101223
Liengaard, B.D., Sharma, P.N., Hult, G. T.M., Jensen, M.B., Sarstedt, M., Hair, J.F., Ringle, C.M., 2021. Prediction: coveted, yet forsaken? introducing a cross-validated predictive ability test in partial least squares path modeling. Decis. Sci. 52 (2), 362–392. https://doi.org/10.1111/deci.12445
Lohmöller, J.-B., 1989. Latent Variable Path Modeling with Partial Least Squares. Physica, Heidelberg.
Long, J.S., 1983. Covariance Structure Models: An Introduction to LISREL. Sage.
MacCallum, R., 1986. Specification searches in covariance structure modeling. Psychol. Bull. 100 (1), 107–120. https://doi.org/10.1037/0033-2909.100.1.107
Marcoulides, G.A., Drezner, Z., 1999. Using simulated annealing for model selection in multiple regression analysis. Multiple Linear Regres. Viewpoints 25 (2), 1–4.

Marcoulides, G.A., Drezner, Z., 2001. Specification searches in structural equation modeling with a genetic algorithm. In: New Developments and Techniques in Structural Equation Modeling. Psychology Press, pp. 267–288.
Marcoulides, G.A., Drezner, Z., 2003. Model specification searches using ant colony optimization algorithms. Struc. Eq. Model. Multidiscip. J. 10 (1), 154–164. https://doi.org/10.1207/S15328007SEM1001_8
Marcoulides, G.A., Drezner, Z., Schumacker, R.E., 1998. Model specification searches in structural equation modeling using tabu search. Struc. Eq. Model. Multidiscip. J. 5 (4), 365–376. https://doi.org/10.1080/10705519809540112
Müller, T., Schuberth, F., Henseler, J., 2018. PLS Path modeling - a confirmatory approach to study tourism technology and tourist behavior. J. Hospit. Tourism Technol. 9 (3), 249–266. https://doi.org/10.1108/JHTT-09-2017-0106
Myung, I.J., 2000. The importance of complexity in model selection. J. Math. Psychol. 44 (1), 190–204. https://doi.org/10.1006/jmps.1999.1283
Nadizar, G., Pietropolli, G., 2023. A grammatical evolution approach to the automatic inference of p systems. J. Membrane Comput. 5 (3), 129–143.
Noonan, R., Wold, H., 1982. PLS Path modeling with indirectly observed variables: a comparison of alternative estimates for the latent variable. In: Jöreskog, K.G., Wold, H. (Eds.), Systems under Indirect Observation: Causality, Structure, Prediction Part II. North-Holland, Amsterdam, pp. 75–94.
Pietropolli, G., Manzoni, L., Paoletti, A., Castelli, M., 2022. Combining geometric semantic gp with gradient-descent optimization. In: European Conference on Genetic Programming (Part of EvoStar). Springer, pp. 19–33.
Rademaker, M.E., Schuberth, F., 2020. cSEM: Composite-Based Structural Equation Modeling. Package version: 0.6.1.9000. https://floschuberth.github.io/cSEM/.
Rademaker, M.E., Schuberth, F., Dijkstra, T.K., 2019. Measurement error correlation within blocks of indicators in consistent partial least squares. Internet Res. 29 (3), 448–463. https://doi.org/10.1108/intr-12-2017-0525
Rashid, M.A., Khatib, F., Hoque, M.T., Sattar, A., 2015. An enhanced genetic algorithm for Ab initio protein structure prediction. IEEE Trans. Evol. Comput. 20 (4), 627–644. https://doi.org/10.1109/TEVC.2015.2505317
Raykov, T., Penev, S., 1999. On structural equation model equivalence. Multivariate Behav. Res. 34 (2), 199–244. https://doi.org/10.1207/s15327906mb340204
Reinartz, W., Haenlein, M., Henseler, J., 2009. An empirical comparison of the efficacy of covariance-based and variance-based SEM. Int. J. Res. Market. 26 (4), 332–344. https://doi.org/10.1016/j.ijresmar.2009.08.001
Rigdon, E.E., Sarstedt, M., Ringle, C.M., 2017. On comparing results from CB-SEM and PLS-SEM: five perspectives and five recommendations. Marketing ZFP 39 (3), 4–16. https://doi.org/10.15358/0344-1369-2017-3-4
Ringle, C.M., Sarstedt, M., Mitchell, R., Gudergan, S.P., 2020. Partial least squares structural equation modeling in HRM research. Int. J. Human Resour. Manag. 31 (12), 1617–1643. https://doi.org/10.1080/09585192.2017.1416655
Ringle, C.M., Sarstedt, M., Straub, D., 2012. A critical look at the use of PLS-SEM in MIS quarterly. MIS Quart. 36 (1), iii–xiv. https://doi.org/10.2307/41410402
Rubio-Largo, Á., Vanneschi, L., Castelli, M., Vega-Rodríguez, M.A., 2018. Multiobjective metaheuristic to design RNA sequences. IEEE Trans. Evol. Comput. 23 (1), 156–169. https://doi.org/10.1109/TEVC.2018.2844116
Ruehle, F., 2017. Evolving neural networks with genetic algorithms to study the string landscape. J. High Energy Phys. 2017 (38), 1–20. https://doi.org/10.1007/JHEP08(2017)038
Sarstedt, M., Hair, J.F., Pick, M., Liengaard, B.D., Radomir, L., Ringle, C.M., 2022a. Progress in partial least squares structural equation modeling use in marketing research in the last decade. Psychol. Market. 39, 1035–1064. https://doi.org/10.1002/mar.21640
Sarstedt, M., Hair, J.F., Pick, M., Liengaard, B.D., Radomir, L., Ringle, C.M., 2022b. Progress in partial least squares structural equation modeling use in marketing research in the last decade. Psychol. Market. 39 (5), 1035–1064. https://doi.org/10.1002/mar.21640
Schamberger, T., 2023. Conducting monte carlo simulations with PLS-PM and other variance-based estimators for structural equation models: a tutorial using the R package cSEM. Indus. Manag. Data Syst. 123 (6), 1789–1813. https://doi.org/10.1108/imds-07-2022-0418
Schamberger, T., Cantaluppi, G., Schuberth, F., 2023. Revisiting and extending PLS for ordinal measurement and prediction. In: Latan, H., Hair Jr, J.F., Noonan, R. (Eds.), Partial Least Squares Path Modeling -. Springer International Publishing, Cham, 2nd ed. pp. 155–182. https://doi.org/10.1007/978-3-031-37772-3_6
Schamberger, T., Schuberth, F., Henseler, J., Dijkstra, T.K., 2020. Robust partial least squares path modeling. Behaviormetrika 47, 307–334. https://doi.org/10.1007/s41237-019-00088-2
Schroeders, U., Scharf, F., Olaru, G., 2024. Model specification searches in structural equation modeling using bee swarm optimization. Educ. Psychol. Meas. 84 (1), 40–61. https://doi.org/10.1177/00131644231160552
Schuberth, F., 2023. The Henseler-Ogasawara specification of composites in structural equation modeling: a tutorial. Psychol. Methods 28 (4), 843–859. https://doi.org/10.1037/met0000432
Schuberth, F., Henseler, J., Dijkstra, T.K., 2018. Partial least squares path modeling using ordinal categorical indicators. Qual. Quant. 52 (1), 9–35. https://doi.org/10.1007/s11135-016-0401-7
Schuberth, F., Rademaker, M.E., Henseler, J., 2023a. Assessing the overall fit of composite models estimated by partial least squares path modeling. Eur. J. Mark. 57 (6), 1678–1702. https://doi.org/10.1108/EJM-08-2020-0586
Schuberth, F., Schamberger, T., Henseler, J., 2024. More powerful parameter tests? no, rather biased parameter estimates. some reflections on path analysis with weighted composites. Behav. Res. Methods 56, 4205–4215. https://doi.org/10.3758/s13428-023-02256-5
Schuberth, F., Zaza, S., Henseler, J., 2023b. Partial least squares is an estimator for structural equation models: a comment on Evermann and Rönkkö (2021). Commun. Assoc. Inf. Syst. 52, 711–729. https://doi.org/10.17705/1CAIS.05232
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464. https://doi.org/10.1214/aos/1176344136
Scrucca, L., 2013. Ga: a package for genetic algorithms in r. J. Stat. Softw. 53, 1–37. https://doi.org/10.18637/jss.v053.i04
Sharma, P., Sarstedt, M., Shmueli, G., Kim, K.H., Thiele, K.O., 2019. PLS-based model selection: the role of alternative explanations in information systems research. J. Assoc. Inf. Syst. 20 (4), 346–397. https://doi.org/10.17005/1jais.00538
Sharma, P.N., Shmueli, G., Sarstedt, M., Danks, N., Ray, S., 2021. Prediction-oriented model selection in partial least squares path modeling. Decis. Sci. 52 (3), 567–607. https://doi.org/10.1111/deci.12329
Shmueli, G., Koppius, O.R., 2011. Predictive analytics in information systems research. MIS Quart. 35 (3), 553–572.
Shmueli, G., Ray, S., Estrada, J. M.V., Chatla, S.B., 2016. The elephant in the room: predictive performance of PLS models. J. Bus. Res. 69 (10), 4552–4564. https://doi.org/10.1016/j.jbusres.2016.03.049
Shukla, A., Pandey, H.M., Mehrotra, D., 2015. Comparative review of selection techniques in genetic algorithm. In: 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), pp. 515–519. https://doi.org/10.1109/ABLAZE.2015.7154916
Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. Ser. B Stat. Methodol. 36 (2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x
Tarjan, R., 1972. Depth-first search and linear graph algorithms. SIAM J. Comput. 1 (2), 146–160. https://doi.org/10.1137/0201010
Team, R.C., 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.
Tenenhaus, M., Amato, S., Esposito Vinzi, V., 2004. A global goodness-of-fit index for PLS structural equation modelling. In: Proceedings of the XLII SIS Scientific Meeting. Vol. 1. CLEUP Padova, pp. 739–742.
Tenenhaus, M., Vinzi, V.E., Chatelin, Y.-M., Lauro, C., 2005. PLS Path modeling. Comput. Stat. Data Anal. 48 (1), 159–205. https://doi.org/10.1016/j.csda.2004.03.005
Tonelli, T., Pietropolli, G., Sbaiz, G., Manzoni, L., 2024. Genetic programming for the reconstruction of delay differential equations in economics. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion, pp. 2119–2122.
Umbarkar, A.J., Sheth, P.D., 2015. Crossover operators in genetic algorithms: a review. ICTACT J. Soft Comput. 6 (1). https://doi.org/10.21917/ijsc.2015.0150
Venkatesh, V., Bala, H., 2008. Technology acceptance model 3 and a research agenda on interventions. Decis. Sci. 39 (2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x
Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D., 2003. User acceptance of information technology: toward a unified view. MIS Quart. 27 (3), 425–478. https://doi.org/10.2307/30036540
Wasserman, L., 2000. Bayesian model selection and model averaging. J. Math. Psychol. 44 (1), 92–107. https://doi.org/10.1006/jmps.1999.1278

Wold, H., 1975. Path models with latent variables: the NIPALS approach. In: Blalock, H.M., Aganbegian, A., Borodkin, F.M., Boudon, R., Capecchi, V. (Eds.), Quantitative Sociology. Academic Press, New York. International Perspectives on Mathematical and Statistical Modeling. chapter 11, pp. 307–357.
Wold, H., 1982. Soft modeling: the basic design and some extensions. In: Jöreskog, K.G., Wold, H. (Eds.), Systems under Indirect Observation: Causality, Structure, Prediction Part II. North-Holland, Amsterdam, pp. 1–54.
Wold, H., 1985. Partial least squares. In: Kotz, S., Johnson, N.L. (Eds.), Encyclopaedia of Statistical Sciences. Wiley. Vol. 6, pp. 581–591.
Yu, X., Schuberth, F., Henseler, J., 2023. Specifying composites in structural equation modeling: a refinement of the henseler–ogasawara specification. Stat. Anal. Data Min. ASA Data Sci. J. 16 (4), 348–357.
Yu, X., Zaza, S., Schuberth, F., Henseler, J., 2021. Counterpoint: representing forged concepts as emergent variables using composite-based structural equation modeling. ACM SIGMIS Database DATABASE Adv. Inf. Syst. 52 (SI), 114–130. https://doi.org/10.1145/3505639.3505647
Zhou, L., Lin, H., Song, X., Li, Y., 2014. Selection of latent variables for multiple mixed-outcome models. Scand. J. Stat. 41 (4), 1064–1082. https://doi.org/10.1111/sjos.12084