Least square estimation of non-linear structural models

Reinhard Oldenburg

Angaben zur Veröffentlichung / Publication details:

Oldenburg, Reinhard. 2024. "Least square estimation of non-linear structural models." *Statistics, Optimization & Information Computing* 12 (2): 281–97. https://doi.org/10.19139/soic-2310-5070-1868.

STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **12**, March 2024, pp 281–297.
Published online in International Academic Press (www.IAPress.org)

Least Square Estimation of Non-linear Structural Models

Reinhard Oldenburg

Department of Mathematics, Augsburg University, Germany

Abstract A new method for estimating a wide class of structural equation models (SEM) is proposed and evaluated. A weighted least squares approach is used that estimates parameters and latent variables. This new approach is flexible enough to handle non-linear and non-smooth models and allows us to model various constraints. The method includes various strategies to deal with the problem of choosing weights. The principle strengths and weaknesses of this approach are discussed, and simulation studies are performed to reveal the problems and potential of this approach.

Keywords Structural equation model, Simulation study, Nonlinear regression model, error estimation

AMS 2010 subject classifications 62P15

DOI: 10.19139/soic-2310-5070-1868

1. Introduction

It is needless to say anything about the importance of structural equational modeling (SEM). A good overview of its development and application is given by [13]. However, despite the many virtues of SEM and the many variations that have been developed over time, there are a number of reasons to look for ways to extend the framework and gain more flexibility in modeling. This paper will introduce and explore a very elementary least squares estimation procedure which is, however, computationally demanding. The traditional approach described, e.g., in [4], is characterized by calculating a parameter-dependent covariance matrix from the model equations and fitting this to the empirical covariance matrix of the data. In this calculation step, the latent variables themselves are eliminated, only their (co)variances remain. This has practical (simplicity and performance of estimation) and theoretical benefits (proofs of properties, reducing the danger of over-fitting, reflecting model assumptions, e.g., on independence of errors). Fitting the model equations (this sloppy phrase is a shortcut for fitting the parameters of the model equations such that residuals of equations become small by some measure) as proposed in this paper, however, is a strategy that applies to a much wider range of models, and hence it is worth investigating the benefits (and the drawbacks) of this approach. First, this introduction shows three model classes that motivate the work, then, it discusses the literature and finally gives an overview of the rest of the paper.

1.1. Examples

The following three examples motivate the development of the new method. They will be taken up later when simulation studies are performed.

1.1.1. Example 1: A competence model. The first example of a class of models that needs more flexible estimation methods than available in established SEM approaches is competency modeling. It has become a widespread

*Correspondence to: Reinhard Oldenburg (Email: reinhard.oldenburg@math.uni-augsburg.de). Department of Mathematics, Augsburg University. Universitätsstraße 14, 86159 Augsburg, Germany.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

assumption in research on the learning of mathematics that one can distinguish knowledge of domain-specific facts and techniques (e.g., in geometry G and statistics S) and general mathematical abilities like argumentation A or modeling M (see, e.g., [23], for an example of a more complicated model with the same structure). These constructs should be modeled by latent variables with some items that try to measure them individually. However, most items will involve at least one domain and at least one general competence. Thus, one is tempted to write an equation for the partial credit on item I with an interaction term like $I = c_0 + c_1 \cdot S + c_2 \cdot M + c_3 \cdot S \cdot M$. To make the interaction term sensible in this case, it must be assumed that the latent variables are positive. The interaction can be interpreted most easily if latent values are from the unit interval $[0, 1]$, such that the product can be interpreted in the lines of fuzzy logic [34] as a conjunction. A practical application of these ideas has been published in [24].

1.1.2. Example 2: Implicative relations. Correlations capture no directional information between variables. To retrieve directional information, other means such as implicative relations are needed. One successful approach is statistical implicative analysis [11], which works, however, only with observed variables and does not allow mixing implicative and correlative relations.

Implications between two centered, numerical variables $x \Rightarrow y$ are not affected by negative values of x, but positive values of y are expected for positive values of x. My suggestion is to incorporate implications using the function $\theta(x) := (x + |x|)/2$ or a step function H defined by $x < 0 \Rightarrow H(x) := 0 \wedge x \geq 0 \Rightarrow H(x) := 1$. One approach is to express the implication $x \Rightarrow y$ by the equation $\theta(x) \cdot (y - a \cdot x) = 0$, where a is a kind of regression coefficient that gives the correlation only in the implicative sense (see estimations of this kind of model for more details). Another possibility is to use θ as a linking function, but there are yet more options. Further research will be needed to evaluate these methods in comparison with established ones. For the present paper, the aim is only to demonstrate that such equations can be used as models.

1.1.3. Example 3: Multiplicative bias in Likert scales. Many psychological studies use Likert scales to measure constructs. Typically, there will be several Likert scales $x_{k,j}, j = 1, \ldots, m$ to measure each construct ξ_k according to a measurement model $x_{k,1} = \xi_k + \epsilon_{k,1}, x_{k,j} = \lambda_{k,j} \cdot \xi_k + \epsilon_{k,j}, j = 2..m$. $x_{k,j}$ may be the individuals' agreement with a given statement. Now, some individuals may have a tendency to express their level of agreement very strongly, while others are reluctant to choose high values on Likert scales. Such an individual bias would increase the correlation of estimated latent variables beyond the true values. Thus, it might be sensible to consider a modified measurement model $x_{k,1} = \chi \cdot \xi_k + \epsilon_{k,1}, x_{k,j} = \lambda_{k,j} \cdot \chi \cdot \xi_k + \epsilon_{k,j}, j = 2..m$, where χ is a latent real variable that describes the expressiveness factor by which an individual over or under expresses agreement on the Likert scale for every individual given their true value (the product $\chi \cdot \xi$ is to be interpreted point-wise over the cases). The values of χ should have a mean of 1 when averaged over the individuals, while values of ξ_k should be in the unit interval.

1.2. Literature

The standard approach to linear SEM is based on the covariance structure [4, pp. 319]. Various methods are available to fit empirical and model-implied covariance matrices and to assess the model fit. The three examples given above show, however, that there are models of interest outside the scope of this method.

The approach presented in this paper estimates latent variable values (factor scores) directly. I see this as an advantage. Surely, in many applications, an individual's factor values of latent variables are not of interest, but having estimates for them allows for further investigations into the model fit. In general, estimates for latent variables can be obtained in various ways: A posteriori factor score methods (e.g., regression scores and Bartlett scores, see [7] and [33]) first estimate an SEM model and then use this information to arrive at estimates for the latent variables, the factor scores. Obviously, this limits the determination of factor scores to cases for which SEM can be estimated. Moreover, this separates the estimation process into two steps, while it might seem better to have one coherent estimation step. Composite methods in a sense take the opposite direction, i.e., they first determine composites as weighted sums of observed variables and substitute them for the latent variables. A popular composite method is that of partial least squares PLS ([8] provides an overview). Related methods such as general structure component analysis (GSCA) have been developed to include measurement error equations [14] and to

allow the modeling of common factors [15]. However, these approaches incorporate two models (measurement and composite), leading to a double determination of the latent variable which somewhat obscures the meaning of the estimates. Factor score regression [6], or more generally the SAM approaches to SEM [28], is somewhat similar in spirit to the current work. However, their methods are not applicable, e.g., to the competence model given above, because they need to separate the measurement and the structural model, while in the competence model these are intrinsically mixed.

Bayesian variants of SEM as realized, e.g., in the blavaan software package [22], are more flexible and can give estimates for latent variables directly. They use Monte Carlo methods (MCMC) to obtain distributions of parameters and latent variables. Moreover, they can be adopted to estimate nonlinear models, see, e.g., [20]. Hence, they can estimate many of the models that motivated the present paper. Therefore, Bayesian estimation is considered as an alternative method in this paper.

There are some further approaches that overcome the restriction to linear models of traditional SEM estimation. Suggestions for nonlinear models are, e.g., given in [29] and [31]. However, many of them lack the ease of use that linear models have. Some approaches are limited to very special generalizations, e.g., quadratic terms. While this has the advantage that some distribution theoretic results can be achieved (e.g., [17]), it could mean that the modeling flexibility is insufficient to match certain situations, as seen in the motivating examples. In fact, it seems that there is no approach that can handle the three examples given above. Non-linearity is often associated with growth curve models (e.g., [13, p. 532]), but in these models, the relations between latent variables remain linear. An approach that may be flexible enough to handle piece-wise and nonlinear models uses splines [12], but it seems that there is no easily accessible implementation.

As mentioned above, a point that distinguishes the proposed method (RLSSEM (residual-based least squares SEM)) to the methods mentioned above is that RLSSEM is completely rooted in the data in the sense that the raw data themselves are used in the model, and are not a derived construct as the covariance matrix or higher moments. Taking the observed data directly and using the proposed model in its original form seems to be the most natural approach, because it suggests that the modeled meaning of a construct fits the intended meaning. Similar data-centered approaches have only recently been investigated for SEM in [5]. This approach combines data-centric estimation of factor scores with an objective function that fits the correlation matrix. Thus, it is close in spirit to the approach of this paper, but it is, at least in its original form, restricted to linear models. Other data-centric approaches have been studied in the context of exploratory factor analysis ([32], [1], [2]). Moreover, full information maximum likelihood (FIML), see [3] in [21], uses the data, not the covariance matrix; however, the aim of this method is just to have a likelihood function that is based on all available data without overcoming the scope of ML estimation of linear SEM.

The next chapter introduces the general framework and discusses several strategies for weight selection. Furthermore, a short section describes measures for model fit. The remainder of the paper presents results from simulation studies.

2. Residual-based least squares SEM (RLSSEM)

The models that will be analyzed have the form of equations that relate data, latent variables, and parameters. Statistical error is the residual (difference of left-hand and right-hand side) to which a model equation fails to hold exactly. This setup includes functional models as special cases. The fitting process will estimate latent variables and parameters in such a way that the equations hold as good as possible. This will now be explained in more technical terms.

The data that are analyzed by using the methods described in this paper consist of a numeric $n \times k$ matrix A. The columns are interpreted as column vectors for n cases of k observed variables. The equational model that is to be fitted to these data consists of a set of m equations numbered by the index $l \in \{1, \ldots, m\}$.

$$g_l\left(\{x_j\}, \{\eta_q\}, \{p_s\}\right) = \epsilon_l \approx 0 \tag{1}$$

These equations relate the measured variables $x_j, j \in \{1, \ldots, k\}$, with latent variables $\eta_q, q \in \{1, .., Q\}$, and with parameters $p_s \in \mathbb{R}, s \in \{1, \ldots, S\}$. In the most general approach, one does not impose any further restrictions on

the real-valued functions g_l, but, of course, their choice is crucial and one should investigate if the models under consideration are identified. To date, this setup specifies a very large model class. Concrete models correspond to concrete choices of these functions.

Latent variables depend on the case just like the observed variables, while the parameters are case-independent properties of the model. The errors ϵ_l should vanish if the model fits the data perfectly, and in any case, they are expected to have expectation 0.

In this setup, the residuals (ϵ_l) are random and the parameters and latent variables are fixed variables.

Estimating the parameters and latent variables will be driven by minimizing error variance while error means vanish. Further model assumptions may be that errors are uncorrelated with each other, that errors are uncorrelated with the latent variables, and that errors are uncorrelated with the observed variables. I call these assumptions secondary constraints. The aim of estimating the model is to determine the $n \cdot Q$ numbers that comprise the latent variables $\eta_q \in \mathbb{R}^n$ and the S parameters p_s to ensure the errors are as small as possible. Hence, following the least squares approximation idea, the goal is to minimize the objective function:

$$F_w\left(\{\eta_q\},\{p_s\}\right) := \sum_{l=1}^{m} w_l \cdot \epsilon_l^2 = \sum_{l=1}^{m} w_l \cdot \sum_{i=1}^{n} \left(g_l\left(\{A_{i,j}\},\{\eta_{q,i}\},\{p_s\}\right)\right)^2 \tag{2}$$

where $w \in \mathbb{R}^m$ is a vector of positive real numbers (weights) of the equations.

Furthermore, in the optimization process, one may wish to put some further constraints (they may be added to describe the model more exactly, but they are not needed to obtain an estimate):

- All (some) latent variables are centered: $\forall q \in \{1,\ldots,Q\} : \mathrm{E}(\eta_q) = 0$.
- All (some) latent variables are normalized: $\forall q \in \{1,\ldots,Q\} : \mathrm{var}(\eta_q) = 1$.
- All (some) error covariances are zero: $\forall l \neq l' \in \{1,\ldots,m\} : \mathrm{cov}(\epsilon_l, \epsilon_{l'}) = 0$.
- All (some) latent error covariances are zero: $\forall q \in \{1,\ldots,Q\} : \forall l \in \{1,\ldots,m\} : \mathrm{cov}(\eta_q, \epsilon_l) = 0$.
- All (some) observed-error-covariances are zero: $\forall j \in \{1,\ldots,k\} : \forall l \in \{1,\ldots,m\} : \mathrm{cov}(x_j, \epsilon_l) = 0$.

These additional constraints may either be set strongly in the sense of constrained optimization or more softly by adding a penalty term to F like $P \cdot c^2$, where $P \in \mathbb{R}^+$ is a penalty constant that adjusts how strict the constraint $c = 0$ is to be realized. In practice, it is often advisable to follow the softer strategy because finite samples usually violate these equations to some extent.

The minimizer of (2) gives values for all parameters and also for all latent variables. Plugging these values back into the expressions for ϵ_l, one obtains estimates of all individual errors, and from this error (co)variances can be calculated.

At this level of generality, it is almost impossible to make substantial claims about the estimates obtained in this way. It is not to be expected that parameters of all types of models that fall into this framework can be estimated without bias. One reason is the incidental parameter problem (see [19] for an overview). The latent variable values η_q are incidental parameters in this view, and their existence can lead to biased estimations in certain models, while this is a theoretical problem, there is a simple pragmatic answer: when estimating a model based on real data, one should first conduct simulation studies to investigate if the model at hand shows strong bias.

2.1. Weight estimation and secondary constraints

The presentation above leaves two important questions open: first, how to choose the weights, and second, what secondary constraints should be taken into account. It turns out that these two issues are related. This will be explained by a calculation and will also be supported by the simulation results presented later on.

2.1.1. Interplay between weights and secondary constraints Numerical simulation experiments indicate that estimates may suffer from substantial bias for some models when F_w above is minimized with the wrong set of weights, but estimates are good when the weights are chosen as the inverses of the true error variances. This is, of course, not a surprise because it is one of the assumptions that shows that least squares is equivalent to maximum likelihood. Moreover, the following calculation will reveal a connection with secondary constraints. To

be more precise, it will be shown that under some assumptions, a sub-optimal choice of weights implies a non-zero correlation of errors and vice versa; bringing this correlation down by a penalty term pushes estimates to the correct values. However, the calculations make some assumptions and approximations and thus are not to be seen as proofs but as hints that support the conjectures.

To keep things simple, we assume to have only $m = 2$ equations (g_1, g_2) such that the objective function is $F := F_w = F_{(w_1,w_2)} = w_1 \cdot g_1^2 + w_2 \cdot g_2^2$. Moreover, assume that the equations are twice differentiable. Combine all parameters and latent variable values into one symbol θ. This θ can be thought of as a $(S + Q) \times n$-matrix, where the first S column vectors contain the same value in all rows. $g_j(\theta) \in \mathbb{R}^n$ is then the vector of residuals of equation j. Assume that there is an isolated true solution θ_0 which is obtained when choosing the correct optimal weights w_1^o, w_2^o. Then, θ_0 is a minimizer for

$$F^o(\theta) = w_1^o \cdot g_1(\theta)^2 + w_2^o \cdot g_2(\theta)^2 \tag{3}$$

and hence the derivative (for short notation $'$ denotes the gradient) should vanish:

$$0 = \frac{1}{2}F^{o\prime}(\theta_0) = w_1^o \cdot g_1(\theta_0) \cdot g_1'(\theta_0) + w_2^o \cdot g_2(\theta_0) \cdot g_2'(\theta_0) \tag{4}$$

As θ_0 is a minimizer of a differentiable function, one expects the second derivative to be positive definite:

$$w_1^o \cdot g_1'(\theta_0)^2 + w_2^o \cdot g_2'(\theta_0)^2 + w_1^o \cdot g_1(\theta_0) \cdot g_1''(\theta_0) + w_2^o \cdot g_2(\theta_0) \cdot g_2''(\theta_0) > 0 \tag{5}$$

Now, assume that we shift the weights away from the optimal value by taking (while retaining their sum): $w_1 = w_1^o + v, w_2 = w_2^o - v$. Then, the minimizer with these weights may shift away from θ_0 to $\theta = \theta_0 + \Delta$. The new critical point equation is:

$$0 = \frac{1}{2}F'(\theta) = (w_1^o + v) \cdot g_1(\theta_0 + \Delta) \cdot g_1'(\theta_0 + \Delta) + (w_2^o - v) \cdot g_2(\theta_0 + \Delta) \cdot g_2'(\theta_0 + \Delta) \tag{6}$$

Now, expand to first order, i.e., set $g_j(\theta_0 + \Delta) = g_j(\theta_0) + g_j'(\theta_0) \cdot \Delta$ and $g_j'(\theta_0 + \Delta) = g_j'(\theta_0) + g_j''(\theta_0) \cdot \Delta$. Then, by subtracting (4) from (6), one arrives, after a tedious computation (which has been checked by using computer algebra), at:

$$\begin{aligned} vg_1(\theta_0)g_1'(\theta_0) + vg_1'(\theta_0)^2 \cdot \Delta + w_1^o \cdot g_1'(\theta_0)^2 \cdot \Delta - vg_2(\theta_0)g_2'(\theta_0) - vg_2'(\theta_0)^2 \cdot \Delta + \\ w_2^o g_2'(\theta_0)^2 \cdot \Delta + vg_1(\theta_0)g_1''(\theta_0) \cdot \Delta + w_1^o g_1(\theta_0)g_1''(\theta_0) \cdot \Delta - vg_2(\theta_0)g_2''(\theta_0) \cdot \Delta + \\ w_2^o g_2(\theta_0)g_2''(\theta_0) \cdot \Delta + vg_1'(\theta_0)g_1''(\theta_0) \cdot \Delta^2 + w_1^o g_1'(\theta_0)g_1''(\theta_0) \cdot \Delta^2 \\ -vg_2'(\theta_0)g_2''(\theta_0) \cdot \Delta^2 + w_2^o g_2'(\theta_0)g_2''(\theta_0) \cdot \Delta^2 = 0 \end{aligned} \tag{7}$$

This approximate equation for the critical point of the optimization with shifted weights allows us to draw several conclusions. First, we ask under what conditions the new minimizer will be the optimal one, that is, $\Delta = 0$. Putting this into the above equation, one obtains $v \cdot g_1(\theta_0)g_1'(\theta_0) - v \cdot g_2(\theta_0)g_2'(\theta_0) = 0$. If the weights were not the optimal ones, i.e., $v \neq 0$, then this would mean that $g_1(\theta_0)g_1'(\theta_0) = g_2(\theta_0)g_2'(\theta_0)$ and that would mean that the gradient is parallel to the level curve and hence θ_0 is not isolated—in contradiction with the assumption above.

What is shown by the above calculation for the low rank expansion above is conjectured to hold in general at least approximately: the correct θ_0 is conjectured to be the minimizer only if the correct weights are used ($v = 0$). Can correct weights ($v = 0$) lead to the wrong estimate, i.e., $\Delta \neq 0$? Setting $v = 0$ in (7), ignoring higher order terms, and factoring out Δ, we find that the same expression that was required to be positive definite in (5) should be zero. Thus, under the assumptions made here, correct weights can be expected to lead to correct estimates.

Now I investigate the secondary constraint that errors are independent, so that the scalar product should vanish, i.e., $g_1(\theta_0) \cdot g_2(\theta_0) = 0$. When the weights deviate from w^o, the scalar product will deviate from 0, and hence the function $G(\theta) := (g_1(\theta) \cdot g_2(\theta))^2$ will have a local minimum in θ_0, i.e.,

$$0 = \frac{1}{2}G'(\theta_0) = g_1(\theta_0) \cdot g_2(\theta_0) \cdot (g_1'(\theta_0) \cdot g_2(\theta_0) + g_1(\theta_0) \cdot g_2'(\theta_0)) \tag{8}$$

Therefore, the idea is to move from the objective function F to an augmented version

$$F^a_{(w_1,w_2)} := F_{(w_1,w_2)} + P \cdot (g_1(\theta_0) \cdot g_2(\theta_0))^2 \tag{9}$$

where P is a large number as a penalty weight.

Now, using $\frac{1}{2}{F^o_{(w_1^o,w_2^o)}}'(\theta_0) = 0$ from (4), one calculates

$$\frac{1}{2}{F^a}_{(w_1^o+v,w_2^o-v)}'(\theta_0+\Delta) = \frac{1}{2}{F^a}_{(w_1^o+v,w_2^o-v)}'(\theta_0+\Delta) - \frac{1}{2}{F^o_{(w_1^o,w_2^o)}}'(\theta_0) \tag{10}$$

Ignoring quadratic terms in Δ or v yields an equation of the following form:

$$A_1 + P \cdot A_2 + \Delta \cdot (B_1 + P \cdot B_2) = 0 \tag{11}$$

Here, A_1, A_2, B_1, B_2 are free of Δ and free of P. Of interest is A_2, which factors into a product with one factor: $g_1(\theta_0) \cdot g_2'(\theta_0) + g_2(\theta_0) \cdot g_1'(\theta_0)$. This factor is zero according to (8). Thus, (11) collapses to $A_1 + \Delta \cdot (B_1 + P \cdot B_2) = 0$ with $B_2 \neq 0$ and, by increasing the penalty P, one can make the deviation Δ small.

Summarizing this subsection, we have two conjectures: First, the local minimizer of F_w should be expected to give unbiased estimates only when the weights are chosen correctly. Second, by applying additional penalties to realize secondary constraints, one may approximate the true solution even if the weights deviate from the optimal values. Note that all calculations above are executed using linear Taylor expansions. Thus, the support for the conjectures is only strong when deviations from the correct solution are small. Experiments showed that choosing initial values can be crucial indeed.

2.1.2. Strategies to choose weights One pathway to make the above framework applicable is to postulate weight-choosing strategies. If the true error variances $\sigma(\epsilon_l)^2$ are known, it would be sensible to set $w_l := 1/\sigma(\epsilon_l)^2$ because this would turn multiplication of an equation with an arbitrary factor into an invariance operation. Moreover, it is easy to see that for any $w_l := c/\sigma(\epsilon_l)^2, c \in \mathbb{R}^+$ and under the additional assumption that the errors are independent and normally distributed, the minimizer of F is the maximum-likelihood estimate, because then for the normalized equations $g_l' := g_l/\sigma(\epsilon_l)$, all errors $\epsilon_l' := \epsilon_l/\sigma(\epsilon_l)$ are distributed in a standard normal distribution. Hence, they have the same variance, and the probability density f of a particular observation $A_{i,\cdot} \in \mathbb{R}^k$ is given by

$$f(A_{i,\cdot}) \sim \prod_{l=1}^{m} exp\left(\frac{-g_l\left(A_{i,\cdot}, \{\eta_{q,i}\}, \{p_s\}\right)^2}{2}\right) \tag{12}$$

and by standard arguments, the least squares minimizer maximizes likelihood (the argument is literally the same calculation as in [30, p. 32]).

In general, knowledge of the true $\sigma(\epsilon_l)$ will not be available and thus the following strategies can be used:

- Strategy W_1: For unweighted least squares, one simply chooses $w_l = 1$. The objective function (just like F_{ULS} in SEM) requires equation residuals to be measured on the same scale.
- Strategy W_n: Assuming that the measured data are of good quality, one may assume that such equations that relate just one latent variable to the observed data are of high importance, while equations that relate many latent variables are of a more hypothetical nature and thus might be less important. This motivates the choice of $w_l = \frac{1}{n^{L_l}}$, where L_l is the number of latent variables in equation g_l.
- Strategy $W_w(W_1)$: This is a two-step-strategy. First, use strategy W_1 to obtain estimates of all variables (and thus of the error variances) and then use the reciprocal error variances as weights in the second step.
- Strategy $W_w(W_n)$: just as before, but now the first round is conducted using W_n.

The naming convention is to use W for compound symbols that denote a strategy for weight choice. The subscript 1 means that all weights are constant while n means that the weight depends on the number of cases and the number of latent equations in a relation. Two-step strategies are denoted by a repeated $W_w()$, with the first step inside the parentheses.

In the simulation studies reported below, the methods W_1, W_n, and $W_w(W_n)$ are included, i.e., $W_w(W_1)$ is omitted because it was never better than $W_w(W_n)$.

2.1.3. Incorporating secondary constraints and combination of methods Here, we will use the second conjecture from Section 2.1.1. If the model contains assumptions that account for secondary assumptions, they may be incorporated into the optimization process. For example, if the model assumptions include $\text{cor}(\epsilon_l, x_j) = 0$, $\text{cor}(\epsilon_l, \eta_q) = 0, \text{cor}(\epsilon_l, \epsilon_{l'}) = 0, l \neq l' \Rightarrow \text{cor}(\epsilon_l, \epsilon_{l'}) = 0$. One can either incorporate them as strict constraints or by penalties, but the former choice is usually not adequate because the random sample may violate this. According to the classical work in [9], if a zero correlation in the population is estimated by a sample correlation of size n, the variance of the estimate is given by $\frac{1}{n-3}$. Hence, it is sensible to form the augmented objective function:

$$\begin{aligned} F_w^a\left(\{\eta_q\},\{p_s\}\right) := & \sum_{l=1}^{m} w_l \cdot \epsilon_l^2 + \\ & (n-3)\cdot\left(\sum_{l=1}^{m-1}\sum_{l'=l+1}^{m}\text{cor}(\epsilon_l,\epsilon_{l'}) + \sum_{l=1}^{m}\sum_{q=1}^{Q}\text{cor}(\epsilon_l,\eta_q) + \sum_{l=1}^{m}\sum_{j=1}^{k}\text{cor}(\epsilon_l,x_j)\right) \end{aligned} \tag{13}$$

Now, the last correlation summands all have the same error variance, namely 1. What remain to be determined are the weights. Combining with the results above, we obtain the following strategies, that all have a superscript a to indicate that they use F_w^a:

- Strategy $W^a(1)$: one simply chooses $w_l = 1$ in minimizing F_w^a.
- Strategy $W^a(W_1)$: choose w_l according to a first round with strategy W_1.
- Strategy $W^{2a} = W^a(W^a(W_1))$: iterate the weight estimation process twice.
- Strategy W^{2a0}: iterate the weight estimation process twice, but in the last round, omit the correlation terms, i.e., use F_w instead of F_w^a.

Out of these methods, only W^{2a}, W^{2a0} will be reported, as they proved to work especially well.

2.1.4. Further comments on estimation If the parameter space is compact and all g_l are continuous, then general principles guarantee the existence of a minimizer $\{\widehat{\eta_q}\},\{\widehat{p_s}\}$, i.e., values such that the function value of the objective function equals the infimum over the total parameter space, $F_1\left(\{\widehat{\eta_q}\},\{\widehat{p_s}\}\right) = \inf F_1\left(\{\eta_q\},\{p_s\}\right)$. Compactness can be deduced if there is an argument that the absolute values of latent variables and parameters are bounded by some (maybe very large) constant. Hence, existence of a minimizer is guaranteed in most situations. Uniqueness, however, is not guaranteed. If the functions g_l are twice differentiable, it can be checked (Hesse matrix) if the minimizer is locally unique. However, it is a question for the optimization method which minimizer may be found.

Regarding maximum likelihood estimation of nonlinear models, [26] has shown that the widespread belief that non-normality of errors always leads to non-consistency is false. However, in general, consistency cannot be assumed. [16, ch. 1] gives a number of examples where nonlinear least square estimators are not consistent. Hence, theoretical investigations of consistency can only be expected for special models. In this paper, only empirical evidence from simulation studies will be given.

2.2. Fitting measures for RLSSEM

The fitting process determines estimates for all parameters and latent variables. This allows calculation of the errors (residuals) ϵ_l of all equations for each case, and hence, a lot of checks to assess model fit can be performed. For example, one can check if errors are approximately zero and if their variances are small. A simple one number fit measure is the mean of residuals $F_{min} := \frac{F_w^{min}}{n}$, or the square root of its per equation average, i.e., $R := \sqrt{\frac{F_w^{min}}{n\cdot m}}$. This latter form is especially useful when comparing different models.

Another and very informative method to judge model fit is to inspect the a posteriori correlation matrices of errors with errors, errors with latents, and errors with observed variables. If some of those are unexpectedly large, one may consider changing the model or adding additional penalties to bring them down. Again, of course, what is substantially large depends on distributional assumptions.

Experience showed that an informative one-number measure can be calculated as follows. Take the average of the largest third of all absolute values of all error–error correlations. Moreover, calculate the same average for error–latent and for error–data correlations. Finally, a goodness-of-fit measure, GOF, is defined as the average of these three numbers.

2.3. Notes on implementation

The approach described here has been in implemented in Mathematica (Wolfram Research). Full source code is available from https://myweb.rz.uni-augsburg.de/~oldenbre/sem/cbsem.zip. This implementation offers more strategies and options than used in this paper. It contains two implementations, one which is tailored for fast estimation but does not support all strategies described above, and one that is slower but more flexible. The numerical optimization step benefits from a good initial guess. If there is a measurement equation $x_j = \eta_q + \epsilon$, it is sensible to use the values of the observed variable x_j as an initial guess for η_q. Interestingly, the impact on the final estimation quality was small, but the run-time was reduced. Therefore, this strategy for selecting initial values was applied whenever possible. Issues of the practical implementation of an older version are given in [25].

3. Case studies

Here, we present studies of the performance of RLSSEM on various problems with simulated data. All simulation results are presented in tables with an identical structure. Unless noted otherwise, all studies were conducted from $N = 100$ simulations. The sample size n was 100 or 300 and is given in the first column. For some models, high values of n are also reported to investigate bias. However, the run-time is approximately quadratic in n so this was not performed for all models.

The parameters and their true values, $p = t$, used in the simulation are given as an equation in the second column. The remaining columns contain the average error (i.e., bias) of the estimates of various methods over N simulation rounds. Below, for each parameter, there is a row (sd) that contains the standard deviations of the estimation errors in parentheses. The last row presents the root mean square error (RMSE) $\frac{1}{N}\sqrt{\sum_{k=1}^{N}(\hat{p}_k - t)^2}$. The row at the bottom reports the average running time in seconds.

3.1. Bollen's democracy model and a variant

A classic example of a non-trivial SEM model is Bollen's model of democracy and industrialization [4, p. 332]. This model can be estimated perfectly with the standard SEM approach, where ML estimation gives consistent unbiased estimates. Surprisingly, this model is the one that shows the strongest sensitivity for the choice of weights. While for many other models, the uniform weight strategy W_1 gives good results, here, this simple strategy will be disappointing, as we shall see, but more advanced strategies can estimate it well.

For the following study, the original real-world data provided by Bollen are not used, instead data are generated from a simulation to be able to compare to the true values.

The model has three latent variables, $ind60$, $dem60$, $dem65$, and eleven observed variables, $x_1, \ldots, x_3, y_1, \ldots, y_8$. The model equations are (with error variables δ, ϵ, γ and intercept variables s_i, t_i):

$$\begin{aligned}
x_1 &= 1{\cdot}ind60 + t_1 + \delta_1 \\
x_2 &= c_2{\cdot}ind60 + t_2 + \delta_2 \\
x_3 &= c_3{\cdot}ind60 + t_3 + \delta_3 \\
y_1 &= 1{\cdot}dem60 + s_1 + \epsilon_1 \\
y_2 &= d_2{\cdot}dem60 + s_2 + \epsilon_2 \\
y_3 &= d_3{\cdot}dem60 + s_3 + \epsilon_4 \\
y_4 &= d_4{\cdot}dem60 + s_4 + \epsilon_4 \\
y_5 &= 1{\cdot}dem65 + s_5 + \epsilon_5 \\
y_6 &= d_6{\cdot}dem65 + s_6 + \epsilon_6 \\
y_7 &= d_7{\cdot}dem65 + s_7 + \epsilon_7 \\
y_8 &= d_8{\cdot}dem65 + s_8 + \epsilon_8 \\
dem60 &= b_1{\cdot}ind60 + \gamma_1 \\
dem65 &= b_2{\cdot}ind60 + b_3{\cdot}dem60 + \gamma_2
\end{aligned}$$

Furthermore, all errors are simulated to be independent. The algorithm produces

$$X_1, X_2, X_3, Y_1, Y_2, \ldots, Y_8, ind60, dem60, dem65 \in \mathbb{R}^n$$

from the following input: sample size $n \in \mathbb{N}$, parameters $b_1, b_2, b_3, c_2, c_3,\ d_2, d_3, d_4,\ d_6, d_7, d_8 \in \mathbb{R}$, and variance parameters $\sigma_{X1}, \sigma_{X2}, \ldots, \sigma_1, \sigma_2 \in \mathbb{R}$. Starting from normally distributed $ind60 \sim N(0,1)$, all values are determined according to the model. The parameters were chosen to be

$$(b_1, b_2, b_3, c_2, c_3, d_2, \ldots, d_4, d_6, \ldots, d_8) = (1.2, 0.5, 0.8, 0.7, 0.9, 0.3, 0.9, 1.7, 0.6, 0.4, 1.3)$$

and for the error standard deviations

$$\begin{aligned}
(\sigma_{X1}, \sigma_{X2}, \sigma_{X3}, \sigma_{Y1}, \sigma_{Y2}, \sigma_{Y3}, \sigma_{Y4}, \sigma_{Y5}, \sigma_{Y6}, \sigma_{Y7}, \sigma_{Y8}, \sigma_1, \sigma_2) = \\
(0.1, 0.2, 0.3, 0.2, 0.1, 0.2, 0.3, 0.2, 0.1, 0.2, 0.3, 0.3, 0.2)
\end{aligned}$$

This choice reflects the idea that the error variances should differ and be of moderate size.

The first insight gained from the simulations was that all methods give good estimates for the loadings in the measurement model (i.e., $c_2, c_3, d_2, \ldots, d_4, d_6, \ldots, d_8$ are estimated without noticeable error), but parameters in the inner structural model are subject to larger deviations. Thus, in the following, we will focus on the estimation of the three parameters b_1, b_2, b_3 in the structural model.

Table 1 gives results for this model where all model assumptions (independent errors, normality of latent variables and errors) are fulfilled. The results show that for this model, W_1 has problems in estimating b_2, b_3. A closer analysis shows that this is due to the fact that ind60 and dem60 correlate strongly and thus in estimating dem65 from them, it is difficult to analyze the individual contribution to dem65. Strategy W_1 overestimates b_2 and underestimates b_3. The other strategies perform much better, although the run-times of the best strategies are about 300 times higher than estimation with lavaan [27], which is given in the table under the label ML.

In further studies, it was investigated whether replacing the normal distribution of errors and of latent variables by uniform distributions and simulating some correlations between errors have any influence, but nothing surprising was found; all methods perform only slightly worse than in the normally distributed case. As is well known, this model can also be estimated well with Bayesian methods, and this has been verified with stan (accessed through the rstan interface in R) and according to the description in [20, ch. 3]. The results were very close to ML estimates, but the run-time was, of course, much longer.

Table 1. Bollen's democracy model

n	**parameter**	**ML**	W_1	W_n	$Ww(W_n)$	W^{2a}	W^{2a0}
300	$b_1 = 1.2$	0.	0.	-0.21	-0.01	0.04	-0.03
	sd	(0.02)	(0.02)	(0.42)	(0.04)	(0.02)	(0.05)
	RMSE	0.02	0.02	0.47	0.04	0.05	0.06
	$b_2 = 0.5$	0.	0.4	-0.07	-0.26	0.	-0.05
	sd	(0.07)	(0.23)	(0.16)	(5.58)	(0.07)	(0.21)
	RMSE	0.07	0.46	0.18	5.55	0.07	0.21
	$b_3 = 0.8$	0.	-0.33	-0.13	0.2	0.01	0.05
	sd	(0.05)	(0.18)	(0.27)	(4.43)	(0.06)	(0.19)
	RMSE	0.05	0.37	0.3	4.41	0.06	0.19
	GOF	nc	0.243	0.374	0.247	0.137	0.239
	time	0.6	61.	9.4	19.2	167.6	226.6
1000	$b_1 = 1.2$	0.	0.	-0.01	-0.02	0.05	-0.02
	sd	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	RMSE	0.01	0.01	0.02	0.02	0.05	0.02
	$b_2 = 0.5$	0.01	0.39	0.	0.14	0.	-0.07
	sd	(0.05)	(0.12)	(0.03)	(0.05)	(0.05)	(0.07)
	RMSE	0.05	0.41	0.03	0.15	0.05	0.1
	$b_3 = 0.8$	0.	-0.32	0.	-0.11	0.01	0.06
	sd	(0.03)	(0.1)	(0.02)	(0.04)	(0.04)	(0.06)
	RMSE	0.03	0.33	0.02	0.12	0.04	0.08
	GOF	nc	0.241	0.241	0.243	0.129	0.233
	time	1.	55.2	54.1	113.8	338.5	338.5

3.1.1. A nonlinear variant The next test was a non-linear version of Bollen's democracy model. The last equation of the model was changed to $dem65 = b_2 \cdot ind60 + b_3 \cdot dem60^2 + \gamma_2$. Moreover, all distributions (latent and errors) were changed to uniform distributions of the same variance. The results are given in Table 2. Again, W_1 has problems with this model, but the other strategies perform quite well.

This model allows for a comparison with the Bayesian approach (MCMC) and the results are given under the heading Bayes in the results table. The realization was similar to the linear version, i.e., all parameters and latent variables were estimated. The initial values of all offset variables were 0 and for all slopes they were 1, and initial values for the three latent variables were their indicators x_1, y_1, y_5. Variance parameters' prior distributions were chosen to be an inverse gamma distribution [**?**, see]ch. 3]LeeSong. Iteration and further parameters were increased until in most runs no warnings were produced by stan.

With this setup, the Bayesian estimation is very good, but avoiding divergent transitions required a large number of iterations and this resulted in high run-times, e.g., 367 seconds on average for a sample size of $n = 100$. Therefore, the results in the Bayes column of Table 2 were calculated from only 20 runs of the simulation, and for $n = 1000$ the Bayesian model was not applied at all (and, for the same reason, for the other methods only 10 simulations were run, in this case).

3.2. A real-world example

This section gives a small example of the kinds of problems this method was developed for, namely, the problem described in Section 1.1.1. The simulation used the following true parameters: $c_{11} = 0.4, c_{13} = 0.6, c_{22} = 0.3, c_{23} = 0.7, c_3 = 0.9, c_4 = 0.8, c_5 = 0.7, c_6 = 0.7$. The simulation algorithm is given below. Here, $\mathcal{U}(a, b)^n$ stands for a vector of n uniformly distributed real numbers from the interval $[a, b]$ and similarly $\mathcal{N}(\mu, \sigma)^n$ is a normally distributed random vector.

Table 2. Non-linear variant of Bollen's model

n	**parameter**	**Bayes**	W_1	W_n	$Ww(W_n)$	W^{2a}	W^{2a0}
100	$b_1 = 1.2$	-0.01	0.	0.	-0.01	0.05	-0.01
	sd	(0.01)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	RMSE	0.02	0.04	0.04	0.05	0.06	0.05
	$b_2 = 0.5$	0.10	-0.06	-0.06	-0.06	-0.06	-0.05
	sd	(0.11)	(0.25)	(0.24)	(0.24)	(0.26)	(0.24)
	RMSE	0.15	0.25	0.24	0.24	0.27	0.25
	$b_3 = 0.8$	0.00	0.04	-0.02	0.	0.02	0.04
	sd	(0.02)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	RMSE	0.02	0.05	0.04	0.04	0.04	0.06
	GOF	na	0.255	0.266	0.262	0.171	0.243
	time	367	1.6	1.6	2.7	43.8	34.8
300	$b_1 = 1.2$	-0.01	0.	-0.01	-0.02	0.05	-0.02
	sd	(0.01)	(0.04)	(0.03)	(0.04)	(0.03)	(0.04)
	RMSE	0.02	0.04	0.03	0.04	0.06	0.04
	$b_2 = 0.5$	0.00	-0.03	-0.03	-0.03	-0.04	-0.03
	sd	(0.06)	(0.2)	(0.19)	(0.19)	(0.21)	(0.19)
	RMSE	0.06	0.2	0.19	0.19	0.21	0.2
	$b_3 = 0.8$	0.00	0.04	-0.02	0.01	0.02	0.04
	sd	(0.01)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
	RMSE	0.01	0.05	0.04	0.03	0.04	0.05
	GOF	na	0.243	0.252	0.249	0.158	0.231
	time	1096	4.6	4.7	8.5	123.9	90.5
2000	$b_1 = 1.2$		0.0	-0.01	-0.02	0.04	-0.02
	sd		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	RMSE		0.01	0.01	0.02	0.05	0.02
	$b_2 = 0.5$		0.01	0.01	0.01	0.01	0.01
	sd		(0.05)	(0.04)	(0.04)	(0.05)	(0.05)
	RMSE		0.05	0.04	0.04	0.05	0.05
	$b_3 = 0.8$		0.04	-0.03	0.00	0.02	0.05
	sd		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	RMSE		0.04	0.03	0.01	0.02	0.05
	GOF		0.22	0.225	0.222	0.125	0.198
	time		193.2	191.9	404.1	1328.1	1611.2

1. $C := \mathcal{U}(0, 0.1)^n, M := C + \mathcal{U}(0, 0.9)^n, S := C + \mathcal{U}(0, 0.8)^n$.
2. $x_1 := c_{11} \cdot S + c_{13} \cdot S \cdot M + \mathcal{N}(0, 0.2)^n$.
3. $x_2 = c_{22} \cdot M + c_{23} \cdot S \cdot M + \mathcal{N}(0, 0.1)^n$.
4. $x_3 := c_3 \cdot S + \mathcal{N}(0, 0.2)^n, \quad x_5 := c_5 \cdot S + \mathcal{N}(0, 0.2)^n$.
5. $x_4 := c_4 \cdot M + \mathcal{N}(0, 0.1)^n, \quad x_6 := c_6 \cdot M + \mathcal{N}(0, 0.1)^n$.

The data set generated by this algorithm was fit to the following model:

$$\begin{aligned} x_1 &= u_1 + c_{11} \cdot S + c_{13} \cdot S \cdot M \\ x_2 &= u_2 + c_{22} \cdot M + c_{23} \cdot S \cdot M \\ x_3 &= u_3 + c_3 \cdot S, \quad x_5 = u_5 + c_5 \cdot S \\ x_4 &= u_4 + c_4 \cdot M, \quad x_6 = u_6 + c_6 \cdot M \\ 0 &\leq S_i \leq 1, \quad 0 \leq M_i \leq 1, \quad \forall i = 1, \ldots, n \\ 0 &\leq u_i \leq 1, \quad \forall i = 1, \ldots, n \end{aligned}$$

Due to the particularities of this model, there is no natural way to fix the scale of the latent variables; only the range is restricted. To extract sensible information from the estimate, it is thus useful to normalize parameters via multiplication resp. division by using the standard deviations of the quantities they link. To compare to the right value, the effective true normalized parameters from the simulation have to be standardized as well. The results of $c_{11}, c_{13}, c_{22}, c_{23}$ are reported in Table 3 (the other coefficients are less critical and usually estimated well by all methods) .

The restriction of all variables to the unit interval effects the estimation methods. As the variables are bounded, so are their variances and hence the error variances. This makes more sophisticated methods for weight determination unnecessary. A second effect is that an equation like $x = c \cdot \xi + \epsilon$, with $x, \xi, \epsilon \in [0, 1]$, implies that for cases $i \in \{1, \ldots, n\}$ with large ξ_i, the signed individual error ϵ_i tends to be low; hence, a small negative correlation between ξ and ϵ is to be expected. As a result, methods that aim to bring correlations between errors and other variables down may be at a disadvantage.

The results in Table 3 confirm that for such models the simple methods $W_1, W_n, W_w(W_n)$ perform reasonably well. From the more advanced methods, W^{2a0} performed well.

For this model, a Bayesian approach has been evaluated as well and is reported in the "Bayes" column in the result table. Initial values for all offset variables were 0, for all slopes they were 0.5, and the initial values for the two latent variables were their indicators x_3, x_4. Variance parameters' prior distributions were chosen as an inverse gamma distribution [**?**, see]ch. 3]LeeSong. A total of 20000 iterations were necessary in order to ensure mixing of chains and other convergence warnings from stan. The results show very good estimates, comparable to W^{2a0}, which was the best method here. However, the run-time for the MCMC approach was by far the highest in this comparison.

3.3. A model with implications

This example takes up the ideas from Section 1.1.3 of the introduction.

The data were generated by the following algorithm:

1. $X \sim \mathcal{U}(-1, 1)^n$.
2. if $X_i < 0$ then $Y_i \sim \mathcal{U}(-1, 1)$ else $Y_i = 0.9 \cdot X + \mathcal{U}(-0.1, 0.1)^n$.
3. $x_1 := X + \mathcal{N}(0, 0.3)^n, x_2 := 0.7 \cdot X + \mathcal{N}(0, 0.2)^n$.
4. $y_1 := Y + \mathcal{N}(0, 0.1)^n, y_2 := 0.4 \cdot Y + \mathcal{N}(0, 0.3)^n$.

The equations of the measurement model are, accordingly, $x_1 = X, x_2 = c_2 \cdot X + u_2, y_1 = Y, y_2 = c_4 \cdot Y + u_4$. The implication between the latent variable was modeled with the following equation: $t(X) \cdot (Y - a \cdot X) = 0$, with a smoothed step function $t(x) = \frac{1}{2} + \frac{1}{\pi} \cdot \arctan(500x)$. The results are summarized in Table 4. All strategies show some difficulties, especially for smaller samples. However, even for $n = 300$, method $W^a(W_1)$ showed a bias of 0.09 for c_4 from the true value 0.4, which amounts to 22%.

In implicative analysis, besides the estimation of parameters, an important question is if the direction of implication is determined correctly. To assess this issue, two further models were used with 25 data sets generated from the same algorithm. The first modified model is a purely correlative model, where the last equation of the model above is replaced by a linear regression $Y = a \cdot X + b$. The other alternative used was $t(Y) \cdot (X - a \cdot Y) = 0$, i.e., it reversed the direction of the implication. Model comparison is shown in Table 5 for a sample size of 300

Table 3. Simulated competence model

n	**parameter**	**Bayes**	W_1	W_n	$Ww(W_n)$	W^{2a}	W^{2a0}
100	$c_{11} = 0.4$	-0.02	0.00	0.09	0.08	0.05	0.07
	sd	(0.08)	(0.18)	(0.12)	(0.13)	(0.21)	(0.15)
	RMSE	0.08	0.18	0.12	0.12	0.2	0.14
	$c_{13} = 0.6$	0.08	0.12	-0.03	0.01	0.17	0.02
	sd	(0.12)	(0.17)	(0.1)	(0.11)	(0.21)	(0.13)
	RMSE	0.14	0.18	0.26	0.22	0.2	0.22
	$c_{22} = 0.3$	-0.01	0.11	0.15	0.1	0.06	0.04
	sd	(0.06)	(0.06)	(0.05)	(0.05)	(0.09)	(0.07)
	RMSE	0.06	0.19	0.23	0.18	0.16	0.13
	$c_{23} = 0.7$	0.07	-0.08	-0.14	-0.06	0.09	0.03
	sd	(0.11)	(0.07)	(0.06)	(0.06)	(0.08)	(0.08)
	RMSE	0.13	0.2	0.26	0.18	0.08	0.11
	GOF	na	0.422	0.475	0.466	0.341	0.419
	time	68	0.7	0.7	1.5	7.4	7.3
300	$c_{11} = 0.4$	-0.02	-0.05	0.05	0.04	0.01	0.04
	sd	(0.04)	(0.08)	(0.06)	(0.06)	(0.1)	(0.06)
	RMSE	0.05	0.13	0.06	0.06	0.11	0.06
	$c_{13} = 0.6$	0.07	0.16	-0.01	0.02	0.21	0.03
	sd	(0.07)	(0.08)	(0.06)	(0.06)	(0.12)	(0.07)
	RMSE	0.10	0.09	0.23	0.19	0.12	0.19
	$c_{22} = 0.3$	-0.01	0.08	0.12	0.07	0.02	0.01
	sd	(0.04)	(0.04)	(0.04)	(0.04)	(0.07)	(0.05)
	RMSE	0.04	0.16	0.2	0.15	0.12	0.1
	$c_{23} = 0.7$	0.08	-0.06	-0.13	-0.04	0.13	0.05
	sd	(0.05)	(0.04)	(0.04)	(0.05)	(0.07)	(0.05)
	RMSE	0.09	0.18	0.24	0.16	0.07	0.08
	GOF	na	0.403	0.46	0.451	0.339	0.408
	time	229	3.1	3.2	6.4	27.2	26.7

and two selected methods. The correct implicative model shows the best fit and thus one may conclude that the method in fact can detect the correct direction of an implication.

An attempt was undertaken to estimate the implicative model using Bayesian estimation based on MCMC. The same approach was taken as in the previous example, but in this case, no satisfactory result could be achieved. Divergent transitions after startup and the problem of non-mixing chains remained even after tweaking different parameters and the number of iterations, until a single run for a sample size of $n = 100$ exceeded one day. However, estimates were still so far away from the true values that they were useless. The failure of the MCMC approach for the implicative model results from the difficulty in modeling $\epsilon := t(X) \cdot (Y - a \cdot X) = 0$. This equation cannot easily be used to predict a variable. Thus, in stan, it was modeled as target$+ =$ normal_lpdf$(\epsilon|0, sd_N)$ but, as mentioned, the results were poor.

Table 4. Implication model

n	**parameter**	W_1	W_n	$Ww(W_n)$	W^{2a}	W^{2a0}
100	$c_2 = 0.7$	-0.06	-0.06	-0.1	0.	0.
	sd	(0.05)	(0.05)	(0.06)	(0.06)	(0.08)
	RMSE	0.08	0.08	0.11	0.06	0.08
	$c_4 = 0.4$	0.14	0.15	0.04	0.12	0.01
	sd	(0.08)	(0.08)	(0.07)	(0.07)	(0.06)
	RMSE	0.16	0.17	0.09	0.14	0.06
	$a = 0.9$	-0.12	-0.15	-0.12	-0.01	0.05
	sd	(0.06)	(0.06)	(0.06)	(0.08)	(0.1)
	RMSE	0.13	0.16	0.14	0.08	0.11
	GOF	0.496	0.537	0.49	0.334	0.463
	time	0.5	0.6	1.1	7.2	7.4
300	$c_2 = 0.7$	-0.06	-0.06	-0.1	0.	-0.01
	sd	(0.04)	(0.04)	(0.04)	(0.05)	(0.07)
	RMSE	0.07	0.08	0.11	0.05	0.07
	$c_4 = 0.4$	0.13	0.15	0.04	0.11	0..00
	sd	(0.07)	(0.07)	(0.06)	(0.06)	(0.06)
	RMSE	0.15	0.16	0.07	0.13	0.06
	$a = 0.9$	-0.12	-0.16	-0.12	-0.02	0.04
	sd	(0.05)	(0.05)	(0.05)	(0.06)	(0.09)
	RMSE	0.13	0.16	0.14	0.07	0.10
	GOF	0.494	0.537	0.489	0.332	0.46
	time	1.6	1.6	3.2	21.5	21.9
1000	$c_2 = 0.7$	-0.05	-0.06	-0.09	0.	0.
	sd	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
	RMSE	0.05	0.06	0.09	0.02	0.03
	$c_4 = 0.4$	0.13	0.14	0.03	0.1	-0.01
	sd	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	RMSE	0.13	0.15	0.04	0.1	0.02
	$a = 0.9$	-0.12	-0.16	-0.12	-0.02	0.03
	sd	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)
	RMSE	0.12	0.16	0.12	0.04	0.04
	GOF	0.488	0.537	0.485	0.328	0.45
	time	14.4	14.1	29.6	83.1	92.4
5000	$c_2 = 0.7$	-0.05	-0.06	-0.1	0.	0.
	sd	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
	RMSE	0.05	0.06	0.1	0.01	0.02
	$c_4 = 0.4$	0.13	0.15	0.03	0.11	-0.01
	sd	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)
	RMSE	0.13	0.15	0.04	0.11	0.01
	$a = 0.9$	-0.12	-0.16	-0.13	-0.02	0.03
	sd	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)
	RMSE	0.12	0.16	0.13	0.04	0.04
	GOF	0.487	0.537	0.486	0.327	0.449
	time	182.7	200.2	420.3	479.6	510.5

Table 5. Model comparsion for implicative data

Strategy $W_w(W_n)$	**Correlation model**	**Correct model**	**Inverse direction**
F_{min}	0.30	0.16	0.28
GOF	0.67	0.49	0.65
a (true 0.9)	0.38	0.80	0.32
Strategy W^{2a}	**Correlation model**	**Correct model**	**Inverse direction**
F_{min}	0.48	0.22	0.53
GOF	0.45	0.33	0.41
a (true 0.9)	0.79	0.88	1.41

3.4. A model for multiplicative Likert scale bias correction

This section takes up the third example from the introduction. The simulated model consists of five latent variables $\xi_1, \ldots, \xi_5$ that were simulated to take values distributed uniformly in $[0, 0.8]$ with a mutual correlation of $\text{cor}(\xi_j, \xi_k) = 0.2$. For each of these latent variables, six indicator variables were simulated according to the equation $x_{k,1} = \xi_k + \epsilon_{k,1}, x_{k,j} = \lambda_{k,j} \cdot \xi_k + \epsilon_{k,j}, j = 2..m$. The weights $\lambda_{k,j}$ were chosen randomly from the set $\{0.5, 0, 6 \ldots, 1.2\}$ and the error standard deviations were in the range 0.1 to 0.3. The latent variable χ that gives the multiplicative bias was simulated as uniformly distributed around the interval $[0.6, 1.4]$ (and uncorrelated with the rest).

From these simulated data, sum scores for the scales (averages over j of the $x_{k,j}$) were calculated. Due to the multiplicative bias, they overestimate the correlation of latent variables. With the parameters chosen in the study, the sum score estimates are 0.36, thus overestimating the true correlation by 0.16 (i.e., 80%.).

The model that was fitted to the data specified further that the latent variables ξ_k are restricted to the unit interval. Furthermore, to avoid numerical issues, χ was restricted to $[0.5, 1.5]$ with a mean of 1. It turned out that all methods other than W^{2a} gave results that showed unexpectedly high correlations between χ and the other latent variables. Thus, the additional constraint that these correlations should vanish was added to the model, and this is reflected in their names by adding a 0 in the exponent in Table 6. This results table shows differences between the mean correlations calculated from the estimates of the latent variables and the true value of the correlations (0.2). Due to long run-times, only 20 simulations (each of sample size 300) were performed.

The insights gathered here are that multiplicative bias can have a substantial influence on estimates and that a nonlinear model can be used to correct for most of this bias. All methods performed similarly well, with the Bayes approach being slightly worse than the RLSSEM methods. Moreover, the Bayesian approach was most time-intensive.

Table 6. Latent correlations estimated from simulated Likert scales

n	**parameter**	**Sum scores**	**Bayes**	W_1^*	$Ww(W_n)^*$	W^{2a}	W^{2a0*}
300	cor	0.16	0.06	0.03	0.05	0.03	0.04
	0.2	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
	RMSE	0.16	0.07	0.04	0.05	0.04	0.05
	time	0	3909	33	72	1686	2610

3.5. Further models

Of course, the methods were checked with more models than can be reported here in detail. For example, as a first almost trivial example, a linear regression model with errors both in the independent and dependent variable

was considered. Thus, one has two observed variables x, y and one latent variable X with model equations $x = X + u_1 + \epsilon_1$ and $y = a \cdot X + u_2 + \epsilon_2$. Of course, this model is not identified as long as no further information about error variances is available. For this model, strategy W_1 reproduces orthogonal regression [10]. Adding a second observed variable makes the model identified and all methods estimate it well.

Ganzach's model as given in [18] was investigated and again, the results were similar to those of the non-linear democracy model.

As a further example, a nonlinear mediator model with three latent variables $X \to Y \to Z$ with relations $Y = a_1 \cdot X + u_1, Z = a_2 \cdot Y + a_3 \cdot X^3 + u_2$ and two indicator variables for each latent variable can be estimated best with W^{2a0} (least GOF) with an RMSE between 0.03 and 0.09.

4. Discussion

This paper introduced the idea of estimating non-linear SEM via least squares with corrections for unknown error variances (RLSSEM). Some theoretical arguments were given as to why this works, and the method was evaluated in simulation studies on various models. The proposed methods give considerably better estimates than un-corrected least squares. There are some non-linear models where, apart from RLSSEM, only Bayesian estimation seems applicable, and for these models, RLSSEM was much faster. For some models, the Bayesian approach worked well, but not for all.

This paper gives empirical evidence that RLSSEM works well for moderate sample sizes and allows modeling more flexibly than existing approaches. However, a lot of research questions are still open. On the technical side, it is desirable handle constraints more sophisticatedly to cut down run-times. For practical applications, one needs to develop reliable fit indices and techniques to allow hypothesis testing (extending the ideas sketched above).

This paper has suggested several variants of RLSSEM and simulation studies have shown that there is no single best method. Therefore, the following research attitude is suggested: specifying a model also amounts to specifying a hypothesis for how the data are generated. Thus, when estimating real-world data, one should first specify the data generation algorithm and use it to simulate data sets. Then, one can evaluate which of the methods works best, and moreover, one can obtain an idea about the possible bias of estimates.

Moreover, the possibilities arising from using non-linear fit functions have to be evaluated; piece-wise linear functions and implicative linkings in particular will be analyzed in follow-up research. Theoretical work should further elaborate questions of consistency for some special models. Summing up, this paper provides some new insights and opens up new areas for research.

REFERENCES

1. Adachi, K. Some contributions to data-fitting factor analysis. *J. Jpn. Soc. Comp. Statist* **2012**, *25*, 197.
2. Adachi, K. *Matrix-based Introduction to Multivariate Data Analysis*; Springer: New York, 2020.
3. Arbuckle, J. Full information estimation in the presence of incomplete data. In *Advanced Structural Equation Modeling: Issues and Techniques*; Marcoulides, G.; Schumacker, R., Eds.; Lawrence Erlbaum Associates: Mahwah, 1996; pp. 243–277.
4. Bollen, K.A. *Structural Equations with Latent Variables*; John Wiley: Hoboken, 1989.
5. Cho, G.; Hwang, H. Structured Factor Analysis: A Data Matrix-Based Alternative Approach to Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal* **2022**, *0*, 1–14.
6. Devlieger, I.; Mayer, A.; Rosseel, Y. Hypothesis Testing Using Factor Score Regression: A Comparison of Four Methods. *Educational and Psychological Measurement* **2016**, *76*, 741–770.
7. DiStefano, V.; Zhu, M.; Mindril, D. Understanding and Using Factor Scores: Considerations for the Applied Researcher. *Practical Assessment, Research & Evaluation* **2009**, *14*, 20.
8. Esposito Vinzi, V.; Chin, W.W.; Henseler, J.; Wang, H.E. *Handbook of Partial Least Squares*; Springer: New York, 2010.
9. Fisher, R.A. On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron* **1921**, *1*.
10. Glaister, P. Least Squares Revisited. *The Mathematical Gazette* **2001**, *85*, 104.
11. Gras, R. *Statistical implicative anaylsis*; Springer: New York, 2008.
12. Guo, R.; Zhu, H.; Chow, S.M.; Ibrahim, J.G. Bayesian Lasso for Semiparametric Structural Equation Models. *Biometrics* **2012**, *68*, 567–577.
13. Hoyle, R.H.E. *Handbook of Structural Equation Modeling*; The Guilford Press: New York, 2012.
14. Hwang, H.; Takane, Y.; Jung, K. Generalized Structured Component Analysis with Uniqueness Terms for Accommodating Measurement Error. *Frontiers in Psychology* **2017**, *8*, 2137. https://doi.org/10.3389/fpsyg.2017.02137.

15. Hwang, H.; Cho, G.; Jung, K.; Lee, S.; et al. An approach to structural equation modeling with both factors and components: Integrated generalized structured component analysis. *Psychological Methods* **2020**. https://doi.org/10.1037/met0000336.
16. Ivanov, A.V. *Asymptotic Theory of Nonlinear Regression*; Kluwer: Dordrecht, 1997.
17. Kelava, A.; Werner, C.S.; Schermelleh-Engel, K.; Moosbrugger, H.; Zapf, D.; Ma, Y.; Cham, H.; Aiken, L.S.; West, S.G. Advanced Nonlinear Latent Variable Modeling: Distribution Analytic LMS and QML Estimators of Interaction and Quadratic Effects. *Structural Equation Modeling* **2011**, *18*, 465–491.
18. Kelava, A.; Brandt, H. Estimation of nonlinear latent structural equation models using the extended unconstrained approach. *Review of Psychology* **2009**, *16*, 123–131.
19. Lancaster, T. The incidental parameter problem since 1948. *Journal of Econometrics* **2000**, *95*, 391–413.
20. Lee, S.Y.; Song, X.Y. *Basic and Advanced Bayesian Structural Equation Modeling*; Wiley, 2012.
21. Marcoulides, G.; Schumacker, R., Eds. *Advanced Structural Equation Modeling: Issues and Techniques*; Psychology Press, 1996. https://doi.org/https://doi.org/10.4324/9781315827414.
22. Merkle, E.C.; Rosseel, Y. blavaan: Bayesian Structural Equation Models via Parameter Expansion. *Journal of statistical software* **2015**, *85*, 4.
23. Neumann, I.e.a. Modeling and assessing mathematical competence over the lifespan. *Journal for educational research online* **2013**, *2*, 80–109. https://doi.org/10.25656/01:8426.
24. Oldenburg, R. Do fuzzy-logic non-linear models provide a benefit for the modelling of algebraic competency? *International Journal of Research in Education Methodology* **2022**, *13*, 1 – 10. https://doi.org/10.24297/ijrem.v13i.9198.
25. Oldenburg, R. Structural Equation Modeling – Comparing Two Approaches. *The Mathematica Journal* **2020**.
26. Philips, P.C.B. On the Consistency of Nonlinear FIML. *Econometrica* **1982**, *50*, 5.
27. Rosseel, Y. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* **2012**, *48*, 1–36.
28. Rosseel, Y.; Loh, W.W. A structural after measurement approach to structural equation modeling. *Psychological Methods* **2022**. https://doi.org/10.1037/met0000503.
29. Schumacker, R.E.; Marcoulides, G.A. *Interaction and nonlinear effects in structural equation modeling*; Lawrence Erlbaum Associates: Mahwah, NJ, 1998.
30. Seber, G.A.F.; Wild, C.J. *Nonlinear Regression*; John Wiley: New York, 1988.
31. Umbach, N.; Naumann, K.; Brandt, H.; Kelava, A. Fitting Nonlinear Structural Equation Models in R with Package nlsem. *Journal of Statistical Software* **2017**, *77*, 7.
32. Unkel, S.; Trendafilov, N.T. Simultaneous Parameter Estimation in Exploratory Factor Analysis: An Expository Review. *International Statistical Review* **2010**, *78*, 363–382.
33. Yung, Y.F.; Yuan, K.H. Bartlett Factor Scores: General Formulas and Applications to Structural Equation Models. In *New Developments in Quantitative Psychology*; E., M.R.; van der Ark, L.A.; Bolt, D.M.; Woods, C.M., Eds.; Springer: New York, 2013.
34. Zadeh, L.A. Fuzzy Sets. *Information and Control* **1965**, *8*, 338–353.