algorithms

MDPI

Article

Implementation Aspects in Regularized Structural Equation Models

Alexander Robitzsch [1,2]

[1] IPN–Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany; robitzsch@leibniz-ipn.de
[2] Centre for International Student Assessment (ZIB), Olshausenstraße 62, 24118 Kiel, Germany

Abstract: This article reviews several implementation aspects in estimating regularized single-group and multiple-group structural equation models (SEM). It is demonstrated that approximate estimation approaches that rely on a differentiable approximation of non-differentiable penalty functions perform similarly to the coordinate descent optimization approach of regularized SEMs. Furthermore, using a fixed regularization parameter can sometimes be superior to an optimal regularization parameter selected by the Bayesian information criterion when it comes to the estimation of structural parameters. Moreover, the widespread penalty functions of regularized SEM implemented in several R packages were compared with the estimation based on a recently proposed penalty function in the Mplus software. Finally, we also investigate the performance of a clever replacement of the optimization function in regularized SEM with a smoothed differentiable approximation of the Bayesian information criterion proposed by O'Neill and Burke in 2023. The findings were derived through two simulation studies and are intended to guide the practical implementation of regularized SEM in future software pieces.

Keywords: structural equation modeling; confirmatory factor analysis; regularized estimation; Bayesian information criterion

Citation: Robitzsch, A. Implementation Aspects in Regularized Structural Equation Models. *Algorithms* **2023**, *16*, 446. https://doi.org/10.3390/a16090446

Academic Editors: Eugene Semenkin, Todor Ganchev and Predrag S. Stanimirovic

Received: 21 August 2023
Revised: 12 September 2023
Accepted: 14 September 2023
Published: 18 September 2023

1. Introduction

Confirmatory factor analysis (CFA) and structural equation models (SEM) are among of the most important statistical approaches for analyzing multivariate data in the social sciences [1–7]. In these models, a multivariate vector $\boldsymbol{X} = (X_1, \ldots, X_I)$ of I observed variables (also referred to as items) is modeled as a function of a vector of latent variables (i.e., factors) $\boldsymbol{\eta}$. SEMs represent the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of the random variable $\boldsymbol{X}$ as a function of an unknown parameter vector θ. In this sense, they apply constrained estimation for the moment structure of the multivariate normal distribution [8].

SEMs impose a measurement model that relates the observed variables $\boldsymbol{X}$ to latent variables $\boldsymbol{\eta}$:

$$\boldsymbol{X} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon} \, . \tag{1}$$

In addition, we denote the covariance matrix $\mathsf{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi}$, and $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ are multivariate normally distributed random vectors. Moreover, $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ are uncorrelated random vectors. The issue of model identification has to be evaluated on a case-by-case basis [9,10]. We now describe two different specifications: the CFA and the more general SEM approach.

In the CFA approach, the multivariate normal (MVN) distribution is represented as $\boldsymbol{\eta} \sim \mathsf{MVN}(\boldsymbol{\alpha}, \boldsymbol{\Phi})$ and $\boldsymbol{\epsilon} \sim \mathsf{MVN}(\boldsymbol{0}, \boldsymbol{\Psi})$. Hence, one can represent the mean vector $\boldsymbol{\mu}(\theta)$ and the covariance matrix $\boldsymbol{\Sigma}(\theta)$ in CFA as a function of an unknown parameter vector θ as

$$\boldsymbol{\mu}(\theta) = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\alpha} \;\text{ and }\; \boldsymbol{\Sigma}(\theta) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi} \, . \tag{2}$$

The parameter vector θ contains freely estimated elements of $\boldsymbol{\nu}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\alpha}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$.

In the general SEM approach, a matrix $\boldsymbol{B}$ of regression coefficients is specified such that

$$\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\eta} + \boldsymbol{\xi}, \text{ with } \mathsf{E}(\boldsymbol{\xi}) = \boldsymbol{\alpha} \text{ and } \mathsf{Var}(\boldsymbol{\xi}) = \boldsymbol{\Phi} . \tag{3}$$

Note that (3) can be rewritten as

$$\boldsymbol{\eta} = (\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\xi}, \text{ with } \mathsf{E}(\boldsymbol{\xi}) = \boldsymbol{\alpha} \text{ and } \mathsf{Var}(\boldsymbol{\xi}) = \boldsymbol{\Phi} , \tag{4}$$

where $\boldsymbol{I}$ denotes the identity matrix. Hence, the mean vector and the covariance matrix are represented in SEM as

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\nu} + \boldsymbol{\Lambda}(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\alpha} \text{ and } \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Phi}[(\boldsymbol{I} - \boldsymbol{B})^{-1}]^{\top}\boldsymbol{\Lambda}^{\top} + \boldsymbol{\Psi} , \tag{5}$$

The estimation of SEM often follows an ideal measurement model. For example, a simple-structure factor loading matrix $\boldsymbol{\Lambda}$ is desired in a multidimensional CFA. In this case, an item loads on one and only factor $\boldsymbol{\eta}$, meaning that the number of non-zero entries in a row of $\boldsymbol{\Lambda}$ is one. However, this assumption on the simple-structure loading matrix could be somewhat violated in practice. For this reason, some cross-loadings could be assumed to be different from zero. Such sparsity assumptions on SEM model parameters can be tackled with regularized SEM [11]. Moreover, deviations in the entries of the observed and modeled mean vector (i.e., $\boldsymbol{\mu} - \boldsymbol{\mu}(\boldsymbol{\theta})$) can be quantified in non-zero entries of the vector of item intercepts $\boldsymbol{\nu}$ (see [12]). Again, model errors could be sparsely distributed, which would allow for the application of regularized SEM. In a similar manner, model deviations $\boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\boldsymbol{\theta})$ can be tackled by assuming sparsely distributed entries in the matrix of residual covariances $\boldsymbol{\Psi}$. Notably, regularized SEM estimation is now becoming more popular in the social sciences and is recognized as an important approach in the machine learning literature [13].

In this article, we review several implementation aspects in estimating regularized SEMs with single and multiple groups. A recent article by Orzek et al. [14] recommended avoiding differentiable approximations for the non-differentiable optimization function in regularized SEM. We critically evaluate the credibility of this statement. Furthermore, we compare the currently used regularization estimation approach in most software, such as the regsem R package [15], with a recently proposed optimization function in the commercial SEM software package Mplus [16]. Finally, we also investigate the performance of a clever replacement of the optimization function in regularized SEM with a smoothed differentiable approximation of the Bayesian information criterion [17]. The findings were derived through two simulation studies. They are intended to provide guidance for the practical implementation of regularized SEM in future software pieces.

The remainder of the article is organized as follows. Different approaches of regularized maximum likelihood estimation methods of SEMs are reviewed in Section 2. In Section 3, research questions are formulated that are addressed in two subsequent simulation studies. In Section 4, results from a simulation study involving a multiple-group CFA model with violations of measurement invariance in item intercepts are presented. Section 5 reports findings from a simulation study of a single-group CFA in the presence of cross-loadings. In Section 6, the findings of the two simulation studies are summarized, and the research questions from Section 3 are answered. Finally, the article closes with a discussion in Section 7.

2. Estimation of Regularized Structural Equation Models

We now describe regularized maximum likelihood (ML) estimation approach for multiple-group SEMs. Note that some identification constraints must be imposed to estimate the covariance structure model (5) (see [2]). For modeling multivariate normally distributed data without missing values, the empirical mean vector $\overline{\boldsymbol{x}}$ and the empirical covariance matrix $\boldsymbol{S}$ are sufficient statistics for estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Hence, they are also sufficient statistics for the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$.

Now, assume that there exist G groups with sample sizes N_g and empirical means $\bar{x}_g$ and covariance matrices S_g ($g = 1, \ldots, G$). The population mean vectors and covariance matrices are denoted by μ_g and Σ_g, respectively. The model-implied mean vectors and covariance matrices are denoted by $\mu_g(\theta)$ and $\Sigma_g(\theta)$, respectively. Note that the parameter vector θ does not have an index g to indicate that there can be common and unique parameters across groups. In a multiple-group CFA, equal factor loadings and item intercepts across groups are frequently imposed (i.e., measurement invariance holds).

Let $\xi_g = (\bar{x}_g, S_g)$ be the sufficient statistics of group g. The combined vector containing all sufficient statistics for the multiple-group SEM is denoted by $\xi = (\xi_1, \ldots, \xi_G)$. The negative log-likelihood function l for the multiple-group SEM (see [2,4]) is given by

$$l(\theta; \xi) = \sum_{g=1}^{G} \frac{N_g}{2} \Big(-I \log(2\pi) + \log|\Sigma_g(\theta)| + \operatorname{tr}(S_g \Sigma_g(\theta)^{-1}) + (\bar{x}_g - \mu_g(\theta))^\top \Sigma_g(\theta)^{-1} (\bar{x}_g - \mu_g(\theta)) \Big) \,. \tag{6}$$

In empirical applications, the model-implied mean vectors covariance matrices will frequently be misspecified [18–20], and θ can be interpreted as a pseudo-true parameter defined as the maximizer of the fitting function l in (6).

In regularized SEM estimation, a penalty function is added to the log-likelihood function that imposes some sparsity assumption on a subset of model parameters [11,12,21]. Frequently, the penalty function $\mathcal{P}$ is non-differentiable in order to impose sparsity. We define a known parameter $\delta_k \in \{0, 1\}$ for all parameters θ_k, where $\delta_k = 1$ indicates that for the kth entry θ_k in θ, a penalty function is applied. The penalized log-likelihood function is given by

$$l_{\text{pen}}(\theta, \lambda; \xi) = l(\theta; \xi) + N^* \sum_{k=1}^{K} \delta_k \mathcal{P}(|\theta_k|^p, \lambda) \,, \tag{7}$$

where λ is a nonnegative regularization parameter, and N^* a scaling factor that frequently equals the total sample size $N = \sum_{g=1}^{G} N_g$. The power p in the penalty function usually takes values in $[0, 2]$. Most of the literature on regularized SEMs employs the power $p = 1$, but $p = 0.5$ has been recently suggested [16] (but see also [22]). The minimizer of $l(\theta)$ is denoted as the regularized (or penalized) ML estimate.

We now discuss typical choices of the penalty function $\mathcal{P}$. For a scalar parameter x, the least absolute shrinkage and selection operator (LASSO) penalty is a popular penalty function used in regularization [23], and it is defined as

$$\mathcal{P}_{\text{LASSO}}(x, \lambda) = \lambda |x| \,, \tag{8}$$

where λ is a nonnegative regularization parameter that controls the extent of sparsity in the obtained parameter estimate. Note that the LASSO penalty function combined with $p = 0.5$ is equivalent to the alignment loss function (ALF [16]):

$$\mathcal{P}_{\text{ALF}}(x, \lambda) = \lambda \sqrt{|x|} \,. \tag{9}$$

It is known that the LASSO penalty introduces bias in estimated parameters. To circumvent this issue, the smoothly clipped absolute deviation (SCAD [24]) penalty has been proposed.

$$\mathcal{P}_{\text{SCAD}}(x, \lambda) = \begin{cases} \lambda |x| & \text{if } |x| < \lambda \\ -(x^2 - 2a\lambda|x|^2 + \lambda^2)(2(a-1))^{-1} & \text{if } \lambda \le |x| \le a\lambda \\ (a+1)\lambda^2 & \text{if } |x| > a\lambda \end{cases} \tag{10}$$

In many studies, the recommended value of $a = 3.7$ (see [24]) has been adopted (e.g., [25,26]). The SCAD penalty retains the penalization rate and the induced bias of the lasso for model parameters close to zero, but continuously relaxes the rate of penalization as the absolute value of the model parameters increases. Note that $\mathcal{P}_{\text{SCAD}}$ has the property of the lasso penalty around zero, but has zero derivatives for x values strongly differing from zero.

Note that the minimizer of l_{pen} is a function of the fixed regularization parameter λ; that is,

$$\widetilde{\theta}(\lambda) = \arg\min_{\theta} l_{\text{pen}}(\theta, \lambda; \xi) \, . \tag{11}$$

Hence, the parameter estimate $\widetilde{\theta}(\lambda)$ of θ depends on a parameter that must be known. To circumvent this issue, the regularized SEM can be repeatedly estimated on a finite grid of regularization parameters λ (e.g., on an equidistant grid between 0.01 and 1.00 with increments of 0.01). The Bayesian information criterion (BIC), defined by $\text{BIC} = 2l(\theta, \xi) + \log(N)H$, where H denotes the number of parameters, can be used to select an optimal regularization parameter. Because the minimization of BIC is equivalent to the minimization of BIC/2, the final parameter estimate $\widehat{\theta}$ is obtained as

$$\widehat{\theta} = \widetilde{\theta}(\widehat{\lambda}) \text{ with } \widehat{\lambda} = \arg\min_{\lambda} \left\{ l(\widetilde{\theta}(\lambda), \xi) + \frac{\log(N)}{2} \left(\sum_{k=1}^{K} \delta_k \chi_0(\widetilde{\theta}_k(\lambda)) \right) \right\} , \tag{12}$$

where the function χ_z as an indicator whether $|x|$ is larger than z for any $z \geq 0$:

$$\chi_z(x) = \begin{cases} 1 & \text{if } |x| > z \\ 0 & \text{if } |x| \leq z \end{cases} \, . \tag{13}$$

In particular, the quantity $\sum_{k=1}^{K} \delta_k \chi_0(\widetilde{\theta}_k(\lambda))$ in (12) counts the number of parameter estimates $\widetilde{\theta}_k(\lambda)$ for $k = 1, \ldots, K$ for which the penalty function is applied (i.e., $\delta_k = 1$) and which differ from 0.

Note that the minimization of the BIC depends on two components. First, the model fit can be improved by minimizing the negative log-likelihood function while freely estimating more parameters. Second, sparse models are preferred in BIC minimization because the second term in (12) minimizes the number of estimated model parameters that are different from zero. Hence, there is always a trade-off between model fit improvement and parsimonious model estimation.

It should be emphasized that BIC is frequently preferred over the Akaike information criterion (AIC) in regularized estimation [11,27]. In typical sample sizes, BIC imposes stronger penalization of the number of estimated parameters than AIC. In fact, alternative information criteria with even stronger penalization are discussed in regularization [25,28,29].

Regularized estimation of single-group and multiple-group SEMs are widespread in the methodological literature [11,21,30–34]. In these applications, cross-loadings, entries in the covariance matrix of residuals, or the vector of item intercepts are regularized. Applying regularized estimation in SEMs allows for flexible yet parsimonious model specifications.

2.1. Regularized SEM Estimation Approaches

Regularized estimation of (11) typically involves a non-differentiable optimization function because the penalty function is non-differentiable. In [14], exact and approximate solutions are distinguished for minimizing the penalized log-likelihood function l_{pen} in (11).

Exact estimation operates on the non-differentiable penalized log-likelihood function. In coordinate descent (CD), the penalized log-likelihood function is cyclically minimized across all entries of the parameter vector θ (see [23]). If the function l_{pen} is minimized in the kth coordinate θ_k of θ, the remaining entries in θ are fixed to the estimate from the previous iteration. This coordinate-wise estimation can be repeated for all parameters and iterated until convergence is reached. The advantage of CD when using the LASSO or the SCAD penalty is that regularized parameters are exactly zero, while nonregularized parameters differ from zero. Hence, a sparse estimate θ is obtained. However, CD can be computationally demanding [14]. In addition, it can also not be generally ensured that a global minimum (instead of a local minimum) is found with CD estimation.

Alternatively, the non-differentiable optimization function can be replaced by a differentiable one [12,35–38]. The penalty function involves the non-differentiable absolute value function that can be replaced by

$$|x| \approx (x^2 + \varepsilon)^{1/2} \text{ or more generally } |x|^p \approx (x^2 + \varepsilon)^{p/2} \tag{14}$$

for a sufficiently small $\varepsilon > 0$, such as $\varepsilon = 10^{-3}$ or $\varepsilon = 10^{-4}$. Fortunately, general-purpose optimizers that rely on derivatives can be relied on when using differentiable approximations based on the penalized log-likelihood function. These optimizers are widely available in software and are reliable if good starting values are available. The disadvantage of the differentiable approximation (DA) approach is that there are no estimated parameters that are exactly zero. To determine a parameter estimate $\widehat{\theta}$, a threshold τ [14] must be specified that defines which small parameter entries should be set to zero. Hence, the final parameter estimate in DA is given by

$$\widehat{\theta} = \widetilde{\theta}(\widehat{\lambda}) \text{ with } \widehat{\lambda} = \arg\min_{\lambda} \left\{ l(\widetilde{\theta}(\lambda), \xi) + \frac{\log(N)}{2} \left(\sum_{k=1}^{K} \delta_k \chi_\tau(\widetilde{\theta}_k(\lambda)) \right) \right\}. \tag{15}$$

Note that the threshold τ is typically a function of ε [14], and τ should be (much) larger than ε. In general, the penalized ML estimate based on DA defined in (15) relies on two tuning parameters, ε and τ, that must be properly chosen. Orzek et al. [14] argue that there is typically not enough knowledge on how to choose these tuning parameters in practical applications. Therefore, they generally prefer CD over DA.

2.2. Direct BIC Minimization Approach of O'Neill and Burke

The estimation approaches described in Section 2.1 require repeatedly fitting a SEM on a grid of regularization parameters λ. Such an approach is computationally demanding, in particular for SEMs with a large number of parameters. The final parameter estimate is obtained by minimizing the BIC across all estimated regularized SEMs. A naïve idea might be directly minimizing the BIC to avoid introducing the penalty function and the unknown regularization parameter λ in the optimization. Only a subset of parameters for which sparsity should be imposed is relevant in the BIC computation. Hence, a parameter estimate by minimizing the BIC is given by

$$\widehat{\theta} = \arg\min_{\theta} \left\{ l(\theta, \xi) + \frac{\log(N)}{2} \left(\sum_{k=1}^{K} \delta_k \chi_0(\theta_k) \right) \right\}. \tag{16}$$

The optimization function in (16) employs a L_0 penalty function [39–41] with a fixed regularization parameter $\log(N)/2$. This optimization function contains the non-differentiable indicator function χ_0. However, like in the DA of the non-differentiable penalty function, the function χ_0 could also be replaced by a differentiable approximation. O'Neill and Burke [17] had the brilliant idea of approximating the indicator function χ_0 by

$$\mathcal{N}(x) = \frac{x^2}{x^2 + \varepsilon} \tag{17}$$

for a sufficiently small $\varepsilon > 0$. Hence, the minimization problem (16) can be replaced by

$$\widehat{\theta} = \arg\min_{\theta} \left\{ l(\theta, \xi) + \frac{\log(N)}{2} \left(\sum_{k=1}^{K} \delta_k \mathcal{N}(\theta_k) \right) \right\}. \tag{18}$$

The estimation approach from (18) is referred to as the smoothed direct BIC minimization (DIR) approach. This estimation approach has been applied to distributional regression models [17].

2.3. Standard Error Estimation

We now describe the computation of the variance matrix of parameter estimates $\widehat{\theta}$ from penalized ML estimation for a fixed regularized parameter λ or the direct BIC minimization approach. Both estimation approaches minimize a differentiable (or differentiable approximation) function $F(\theta, \xi)$ with respect to θ as a function of sufficient statistics ξ (see also [6]). The vector of sufficient statistics $\widehat{\xi}$ is approximately normally distributed (see [3]); that is,

$$\widehat{\xi} - \xi_0 \sim \mathsf{MVN}(\mathbf{0}, \boldsymbol{V}_{\xi}) \tag{19}$$

for a true population parameter ξ_0 of sufficient statistics. We denote by $F_\theta = (\partial F)/(\partial \theta)$ the vector of partial derivatives with respect to θ. The parameter estimate $\widehat{\theta}$ is given as the root of the non-linear equation

$$F_\theta(\theta, \widehat{\xi}) = \mathbf{0} \, . \tag{20}$$

General M-estimation theory (i.e., the delta method [18]) can be applied to derive the variance matrix of $\widehat{\theta}$. Assume that there exists a (pseudo-)true parameter θ_0 such that

$$F_\theta(\theta_0, \xi_0) = \mathbf{0} \, . \tag{21}$$

We now derive the covariance matrix of $\widehat{\theta}$ by utilizing a Taylor expansion of F_θ. We denote by $F_{\theta\theta}$ and $F_{\theta\xi}$ the matrices of second-order partial derivatives of F_θ with respect to θ and ξ, respectively. We obtain

$$F_\theta(\widehat{\theta}, \widehat{\xi}) = F_\theta(\theta_0, \xi_0) + F_{\theta\theta}(\theta_0, \xi_0)(\widehat{\theta} - \theta_0) + F_{\theta\xi}(\theta_0, \xi_0)(\widehat{\xi} - \xi_0) = \mathbf{0} \, . \tag{22}$$

As the parameter estimate $\widehat{\theta}$ is a non-linear function of $\widehat{\xi}$, the Taylor expansion (22) provides the approximation

$$\widehat{\theta} - \theta_0 = -F_{\theta\theta}(\theta_0, \xi_0)^{-1} F_{\theta\xi}(\theta_0, \xi_0)(\widehat{\xi} - \xi_0) \, . \tag{23}$$

By defining $\boldsymbol{A} = -F_{\theta\theta}(\theta_0, \xi_0)^{-1} F_{\theta\xi}(\theta_0, \xi_0)$, we get by using the multivariate delta formula [18]:

$$\mathsf{Var}(\widehat{\theta}) = \boldsymbol{A}\boldsymbol{V}_{\xi}\boldsymbol{A}^{\top} . \tag{24}$$

An estimate of $\boldsymbol{A}$ is obtained as $\widehat{\boldsymbol{A}} = -F_{\theta\theta}(\widehat{\theta}, \widehat{\xi})^{-1} F_{\theta\xi}(\widehat{\theta}, \widehat{\xi})$. This approach is ordinarily used for differentiable discrepancy functions in the SEM literature [3,7,42]. Standard errors for entries in $\widehat{\theta}$ can be obtained by taking the square root of diagonal elements of $\mathsf{Var}(\widehat{\theta})$ computed from (24).

3. Research Questions

In the following two simulation studies, several implementation and algorithmic aspects of regularized SEM estimation are investigated. Five research questions (RQ) are imposed in this section that will be answered by means of the simulation studies.

The research questions are tackled through two simulation studies. The first, Simulation Study 1, considers the case of regularized multiple-group SEM estimation with noninvariant item intercept. In the second, Simulation Study 2, regularized SEM estimation is applied for data simulated from a two-factor model in the presence of cross-loadings.

3.1. RQ1: Fixed or Estimated Regularization Parameter λ?

In the first research question, RQ1, we consider the choice of the regularization parameter λ regarding statistically efficient parameter estimation if structural parameters, such as factor means or factor correlations, are the primary analytical focus. We study whether an optimal regularization parameter is obtained by information criteria or a pre-chosen regularization parameter. Using only a fixed value of the regularization parameter instead of estimating the regularized SEM on a sequence of regularization parameters would decrease the computational burden of the estimation.

3.2. RQ2: Exact Optimization or Differentiable Approximation?

In the second research question, RQ2, we compare exact optimization and approximate optimization approaches based on differentiable approximations for regularized SEMs. Previous work argued that the exact approach should be generally preferred. We thoroughly investigate whether this preference is justified. Notably, approximate optimization with differentiable optimization functions is easier to implement because general-purpose optimizers are widely available and provide reliable convergence guarantees if adequate starting values are used in the estimation.

3.3. RQ3: Direct BIC Minimization or Minimizing BIC Using a Grid of λ Values?

The third research question, RQ3, investigates whether the direct one-step BIC minimization approach provides comparable results to the indirect estimation approach that requires the estimation of the regularized SEM on a grid of regularization parameters. If the one-step BIC minimization approach provides similar findings to the indirect approach, substantial computational gains would be achieved, which eases the application of regularized SEM.

3.4. RQ4: Always Choosing the Power $p = 1$ in the Penalty Function?

In the fourth research question, RQ4, we investigate whether there are considerable differences in the choice of the power p in the penalty. While the majority of regularization approaches employ the absolute value function $p = 1$, a recent implementation in the popular Mplus software utilizes $p = 0.5$. The outcome of this comparison gives hints on how future regularized SEM software should be implemented.

3.5. RQ5: Does the Delta Method Work for Standard Error Estimation?

Finally, in the fifth research question, RQ5, the quality of standard error estimation in terms of coverage rates (see Section 2.3) is studied. It is interesting whether the standard errors based on the delta method are reliable if they are applied for differentiable approximations of the optimization function in regularized SEM.

4. Simulation Study 1: Noninvariant Item Intercepts (DIF)

In Simulation Study 1, we investigated the impact of group-specific item intercepts in a multiple-group one-dimensional factor model. In the data-generating model (DGM), measurement invariance was violated.

4.1. Method

The setup of the simulation study mimics the one presented in [43]. Datasets were simulated from a one-dimensional factor model involving five items and three groups. The factor variable η_1 was normally distributed with group means $\alpha_{1,1} = 0$, $\alpha_{2,1} = 0.3$, and $\alpha_{3,1} = 0.8$. The group variances were set to $\phi_{1,11} = 1$, $\phi_{2,11} = 1.5$, and $\phi_{3,11} = 1.2$, respectively. All factor loadings were set to 1, and all measurement error variances were set to 1 in all groups and uncorrelated with each other. The factor variable, as well as the residual variables, were normally distributed.

Some non-zero group-specific item intercepts were simulated that indicate measurement noninvariance. These differential item functioning (DIF [44]) effects in item intercepts were simulated in one and only one of the five items in each group. In the first group, the fourth item intercepts had a DIF effect δ. In the second group, the first item had a DIF effect $-\delta$, while the second item had a DIF effect $-\delta$ in the third group. The DIF effect δ was chosen as either 0.3 or 0.6. The sample size per group was chosen as $N = 500$ or $N = 1000$.

A regularized multiple-group one-dimensional SEM was specified as the analysis model. In this model, invariant factor loadings were assumed. For identification reasons, the mean of the factor variable in the first group was fixed at 0, and the standard deviation in the first group was fixed at 1. The SCAD penalty function was imposed on group-

specific item intercepts. In the penalty function, the powers $p = 1$ and $p = 0.5$ were investigated. The SEM was estimated on a grid of regularization parameters between 0.025 and 0.40, with increments of 0.025. The exact estimation approach was implemented by the coordinate descent (CD). In the differentiable approximation (DA) of the non-differentiable SCAD penalty function, we chose $\varepsilon = 10^{-4}$. The optimal regularization parameter λ was obtained by minimizing the BIC. Because no estimated item intercepts are exactly set to 0 in the estimation, the thresholds $\tau = 0.01$, 0.02, and 0.05 were chosen as the cutoff values for treating model parameters as a value of 0 in the BIC computation. Furthermore, the smoothed direct BIC minimization (DIR) approach of O'Neill and Burke was carried out using $\varepsilon = 0.01$. This relatively large value was found to be optimal in preliminary simulation studies in which the tuning parameter ε was varied as 0.1, 0.01, 0.001, and 0.0001. The lowest RMSE and a small bias for parameter estimates were obtained for $\varepsilon = 0.01$.

For the direct BIC minimization method DIR and regularized ML estimation for a set of fixed regularization parameters λ, standard errors were computed by means of the delta method described in Section 2.3. Confidence intervals were calculated based on the normal distribution assumption (i.e., the confidence interval of an estimate $\hat{\theta}$ was computed as $[\hat{\theta} - 1.96 \cdot \mathsf{SE}(\hat{\theta}), \hat{\theta} + 1.96 \cdot \mathsf{SE}(\hat{\theta})]$, where $\mathsf{SE}(\hat{\theta})$ is the estimated standard error).

In total, 1,000 replications were conducted for all 2 (DIF effect size δ) $\times$ 2 (sample size N) = 4 conditions of the simulation study. We investigated the estimation quality of factor means and factor variances. Bias and root mean square error (RMSE) were utilized to assess the performance of different estimators. Let $\hat{\theta}_r$ be a model parameter estimate in replication $r = 1, \ldots, R$. The bias was estimated by

$$\text{Bias}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}_r - \theta) \,, \tag{25}$$

where θ denotes the true parameter value. The RMSE was estimated by

$$\text{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}_r - \theta)^2} \,. \tag{26}$$

Coverage rates at the confidence level of 95% were computed as the percentage of the events that a computed confidence interval covers the true parameter value. The models were estimated using the `sirt::mgsem()` function in the R [45] package sirt [46]. Replication material can be found in the directory "*Simulation Study 1*" located at https://osf.io/7kzgb (accessed on 21 August 2023) .

4.2. Results

Figures 1 and 2 display the absolute bias and the RMSE of the factor mean $\alpha_{2,1}$ in the second group as a function of the regularization parameter λ for the two sample sizes, $N = 500$ and $N = 1000$, and the two powers of the penalty function $p = 1$ and $p = 0.5$, respectively. It can be seen in the two figures that there is a range of values of the regularization parameter λ, which results in unbiased and least variable (i.e., in terms of RMSE) estimates. The optimal fixed regularized parameter was larger for $p = 0.5$ than for $p = 1$. However, the minimal RMSE was similar for $p = 1$ and $p = 0.5$ in Simulation Study 1. Furthermore, the RMSE of the factor mean estimate based on the optimal regularization parameter selected by the minimal BIC did not generally outperform a well-chosen fixed regularization parameter λ.

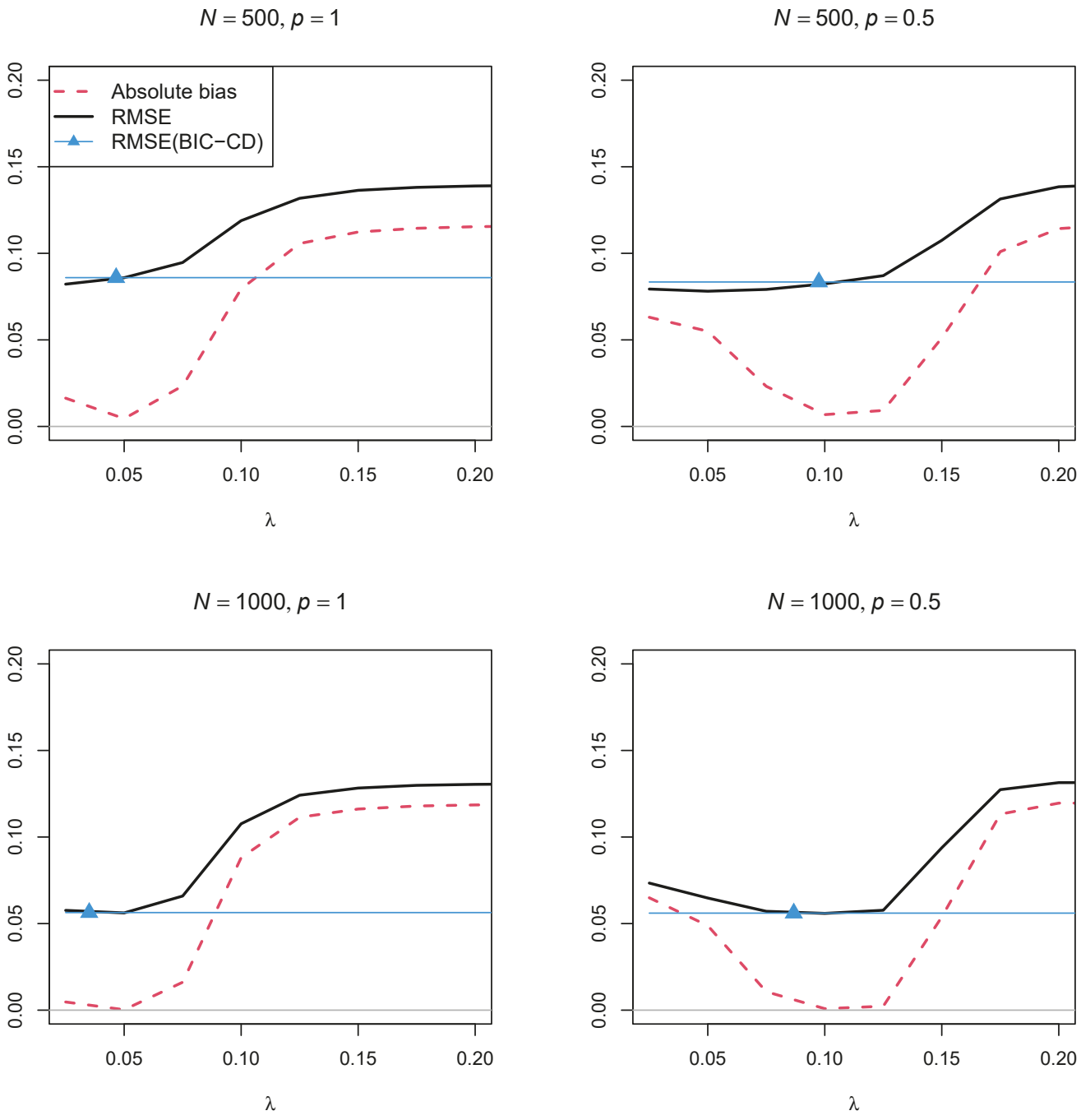

Figure 1. Simulation Study 1: Absolute bias and root mean square error (RMSE) of the factor mean $\alpha_{2,1}$ in the second group as a function of the regularization parameter λ for a DIF effect of the item intercept of $\delta = 0.3$ for sample sizes $N = 500$ and $N = 1000$ and powers $p = 1$ and $p = 0.5$ of the penalty function. The RMSE of the estimate obtained by the optimal BIC and coordinate descent (BIC-CD) is displayed by the blue line. The location of the average optimal regularization parameter obtained by BIC-CD is displayed by the blue triangle.

Interestingly, Figure 2 illustrates in the condition of a larger DIF effect $\delta = 0.6$ that too small regularization parameters λ resulted in biased parameter estimates. The issue occurred both for $p = 1$ and $p = 0.5$. Moreover, by comparing Figures 1 and 2, it is evident that the optimal regularization parameter is a function of the size of DIF effects δ. That is, larger DIF effects δ resulted in larger regularization parameters λ.

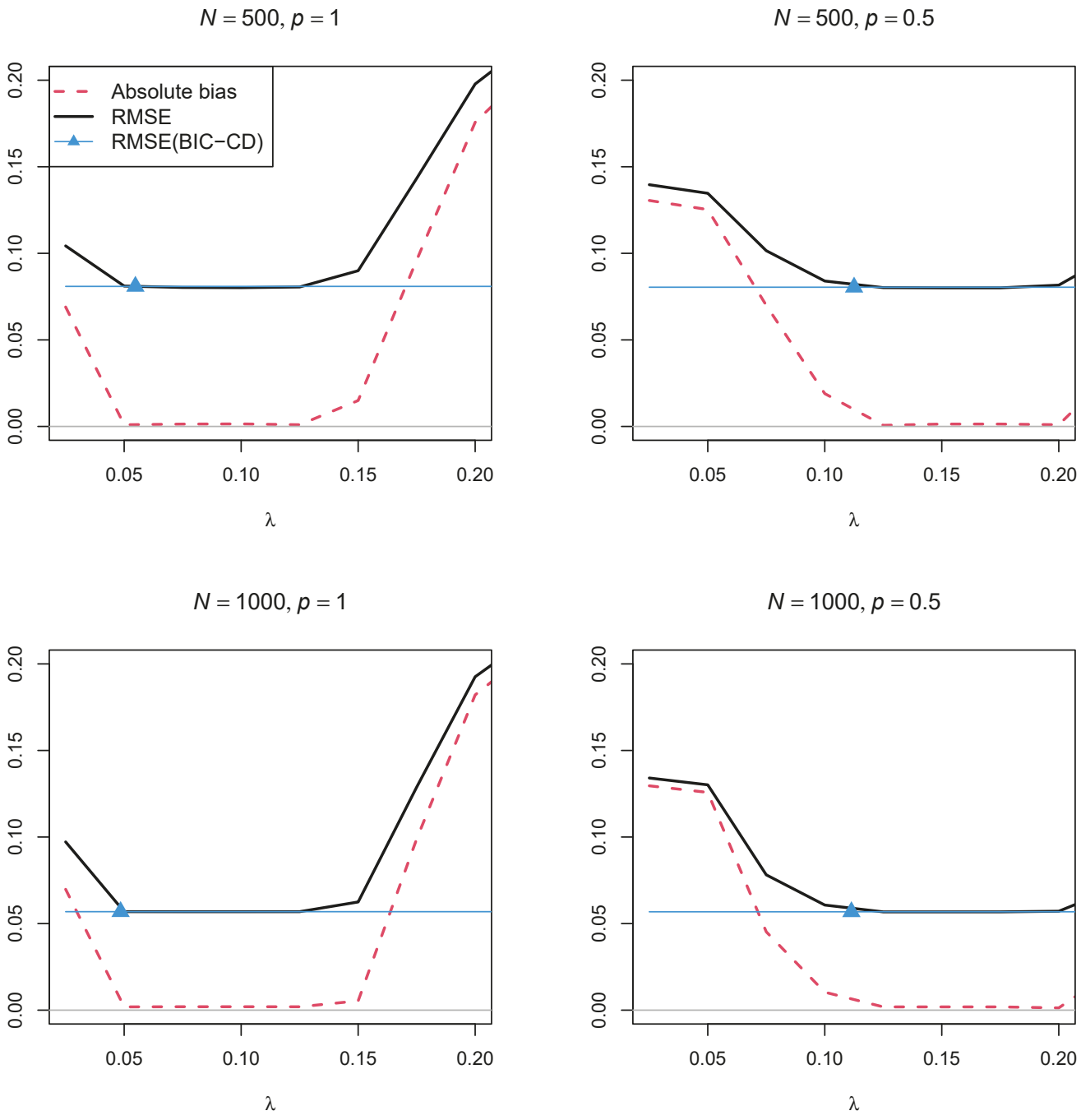

Figure 2. Simulation Study 1: Absolute bias and root mean square error (RMSE) of the factor mean $\alpha_{2,1}$ in the second group as a function of the regularization parameter λ for a DIF effect of the item intercept of $\delta = 0.6$ for sample sizes $N = 500$ and $N = 1000$ and powers $p = 1$ and $p = 0.5$ of the penalty function. The RMSE of the estimate obtained by the optimal BIC and coordinate descent (BIC-CD) is displayed by the blue line. The location of the average optimal regularization parameter obtained by BIC-CD is displayed by the blue triangle.

Table 1 presents the bias and the RMSE of the estimated group means of the second and the third group. In this table, the direct BIC minimization approach DIR is compared with the exact estimation approach (CD) and the differentiable approximation (DA) using the optimal regularization parameter λ based on the minimal BIC, as well as for fixed regularization parameters $\lambda = 0.05$ for the power $p = 1$ and $\lambda = 0.10$ for $p = 0.5$. The DA estimation approach is shown using the threshold $\tau = 0.02$. Values of the fixed

regularization parameters were chosen based on the findings in Figures 1 and 2. Overall, there was only a negligible bias in factor mean estimates. Regarding RMSE, all different estimation methods resulted in relatively similar estimates. Furthermore, there were essentially no differences between the exact solution (CD) and the approximate solution (DA). If researchers seek toknow a close-to-optimal regularization parameter λ, regularized ML estimation must not involve the choice of an optimal λ based on a minimal BIC. Finally, the computationally cheap direct BIC minimization approach (DIR) performed similarly to BIC estimation that requires fitting a regularized SEM at a sequence of regularization parameters λ. A slight increase in RMSE of the DIR method was only observed for $N = 500$ and $\delta = 0.3$, which was the consequence of slightly biased parameter estimates.

Table 1. Simulation Study 1: Bias and root mean square error (RMSE) of factor means as a function of sample size N and the size of the DIF effect of item intercepts δ.

				p = 1				p = 0.5			
			BIC	BIC		λ = 0.05		BIC		λ = 0.10	
Par	N	δ	**DIR**	**CD**	**DA**	**CD**	**DA**	**CD**	**DA**	**CD**	**DA**
			Bias								
$\alpha_{2,1}$	500	0.3	−0.012	−0.005	−0.005	−0.004	−0.005	−0.005	−0.006	−0.007	−0.007
		0.6	0.000	0.000	0.001	−0.001	−0.001	0.000	0.000	−0.019	−0.019
	1000	0.3	−0.005	−0.001	0.000	0.000	0.000	−0.001	−0.001	−0.001	−0.001
		0.6	0.001	0.002	0.002	0.002	0.002	0.002	0.002	−0.010	−0.010
$\alpha_{3,1}$	500	0.3	−0.013	−0.007	−0.006	−0.006	−0.006	−0.007	−0.007	−0.008	−0.008
		0.6	0.001	0.001	0.002	0.000	0.000	0.001	0.001	−0.017	−0.017
	1000	0.3	−0.004	0.000	0.001	0.001	0.001	0.000	0.000	0.000	0.000
		0.6	0.004	0.004	0.004	0.004	0.004	0.004	0.004	−0.008	−0.008
			RMSE								
$\alpha_{2,1}$	500	0.3	0.090	0.086	0.086	0.086	0.086	0.083	0.083	0.082	0.082
		0.6	0.081	0.081	0.081	0.081	0.081	0.080	0.080	0.084	0.084
	1000	0.3	0.057	0.056	0.056	0.056	0.056	0.056	0.056	0.056	0.056
		0.6	0.057	0.057	0.057	0.057	0.057	0.057	0.057	0.060	0.061
$\alpha_{3,1}$	500	0.3	0.096	0.090	0.090	0.090	0.090	0.087	0.087	0.086	0.086
		0.6	0.083	0.082	0.082	0.082	0.082	0.082	0.082	0.083	0.084
	1000	0.3	0.058	0.058	0.058	0.058	0.058	0.057	0.057	0.057	0.057
		0.6	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.061	0.061

Note . Par = parameter; $\alpha_{g,1}$ = factor mean in group $g = 2, 3$; estimation using the optimal regularization parameter based on the BIC; p = power used in the penalty function; λ = fixed regularized parameter; DIR = direct BIC minimization using the differentiable approximation of O'Neill and Burke (2023) with $\varepsilon = 0.01$; CD = coordinate descent; DA = differentiable approximation using the threshold parameter $\tau = 0.02$.

Table 2 compares the average number of regularized parameters of the exact approach (CD) and the differentiable estimation approach (DA) using the thresholds τ of 0.01, 0.02, and 0.04. It turned out that the number of regularized item intercepts in the selected models was very similar. Only for $N = 500$ and $\delta = 0.3$ was the number of regularized parameters slightly too low. Notably, lower values of thresholds such as $\tau = 0.005$ or $\tau = 0.001$ would result in a substantially lower average number of regularized parameters.

Table 2. Simulation Study 1: Average number of regularized item intercepts using the optimal regularized parameter λ based on the BIC as a function of sample size N and the size of the DIF effect of item intercepts δ.

		p = 1				p = 0.5			
			DA with τ =				DA with τ =		
N	δ	CD	0.01	0.02	0.04	CD	0.01	0.02	0.04
500	0.3	11.86	11.45	11.93	12.08	11.95	11.96	11.96	11.96
	0.6	11.94	11.96	11.97	11.99	11.97	11.97	11.97	11.97
1000	0.3	11.95	11.87	11.98	12.00	11.95	11.96	11.96	11.96
	0.6	11.99	12.00	12.00	12.00	11.98	11.99	11.99	11.99

Note. p = power used in the penalty function; CD = coordinate descent; DA = differentiable approximation using a threshold parameter τ.

Table 3 focuses on the coverage rates of selected model parameters. It can be seen that coverage rates for a model parameter were acceptable for both powers $p = 1$ and $p = 0.5$ if the respective parameter estimate was approximately unbiased. Interestingly, the coverage rates were also satisfactory for the direct BIC minimization approach (DIR).

Table 3. Simulation Study 1: Bias and coverage rates as a function of sample size N and the size of the DIF effect of item intercepts δ.

			Bias					Coverage				
				$p = 1$		$p = 0.5$			$p = 1$		$p = 0.5$	
			BIC	DA with λ =		DA with λ =		BIC	DA with λ =		DA with λ =	
Par	N	δ	DIR	0.05	0.10	0.10	0.15	DIR	0.05	0.10	0.10	0.15
$\alpha_{2,1}$	500	0.3	−0.02	−0.01	**−0.08**	−0.01	**−0.06**	94.1	93.7	**81.9**	92.4	**82.8**
		0.6	0.00	0.00	0.00	−0.01	0.00	95.4	95.0	95.1	95.2	95.3
	1000	0.3	0.00	0.00	**−0.09**	0.00	**−0.06**	95.3	95.5	**67.5**	95.3	**74.2**
		0.6	0.00	0.00	0.00	0.00	0.00	94.5	94.5	94.5	94.5	94.5
$\alpha_{3,1}$	500	0.3	−0.01	0.00	**−0.08**	−0.01	**−0.05**	94.1	94.7	**81.8**	92.9	**85.1**
		0.6	0.00	0.00	0.00	0.00	0.00	95.5	94.9	95.0	94.5	94.9
	1000	0.3	0.00	0.00	**−0.09**	0.00	**−0.05**	95.8	95.6	**66.0**	95.2	**75.7**
		0.6	0.00	0.00	0.00	0.00	0.00	95.3	95.1	95.2	95.1	95.2
$\phi_{2,11}$	500	0.3	0.01	0.01	0.01	0.01	0.01	95.4	95.3	95.3	95.2	95.3
		0.6	0.01	0.01	0.01	0.00	0.00	95.6	95.4	95.5	95.2	95.3
	1000	0.3	0.01	0.01	0.01	0.00	0.01	95.2	95.0	95.0	94.9	95.0
		0.6	0.00	0.00	0.00	0.00	0.00	95.5	95.2	95.2	95.2	95.2
$\phi_{3,11}$	500	0.3	0.01	0.01	0.01	0.01	0.01	94.8	94.6	94.7	94.4	94.6
		0.6	0.01	0.01	0.01	0.01	0.01	95.2	95.2	95.2	95.1	95.1
	1000	0.3	0.00	0.00	0.00	0.00	0.01	95.2	95.2	95.4	95.4	95.3
		0.6	0.00	0.00	0.00	0.00	0.00	95.8	95.3	95.3	95.2	95.2
$\nu_{2,1}$	500	0.3	0.02	0.01	**0.16**	0.02	**0.12**	94.5	93.9	**35.1**	91.6	**48.3**
		0.6	0.00	0.00	0.00	0.03	0.00	95.3	94.3	95.2	**85.7**	95.2
	1000	0.3	0.01	0.00	**0.18**	0.00	**0.12**	96.0	95.7	**32.1**	94.8	**52.0**
		0.6	0.00	0.00	0.00	0.02	0.00	95.1	94.7	94.5	**85.3**	94.5

Table 3. *Cont.*

			Bias					Coverage				
				$p = 1$		$p = 0.5$			$p = 1$		$p = 0.5$	
			BIC	DA with λ =		DA with λ =		BIC	DA with λ =		DA with λ =	
Par	N	δ	**DIR**	**0.05**	**0.10**	**0.10**	**0.15**	**DIR**	**0.05**	**0.10**	**0.10**	**0.15**
$\nu_{3,2}$	500	0.3	0.02	0.01	**0.17**	0.02	**0.12**	94.2	93.5	**29.6**	92.0	**48.4**
		0.6	0.00	0.00	0.00	0.03	0.00	95.8	95.1	95.4	**86.1**	95.4
	1000	0.3	0.00	0.00	**0.19**	0.00	**0.11**	95.6	95.2	**23.7**	94.5	**52.1**
		0.6	0.00	0.00	0.00	0.02	0.00	95.3	95.0	94.9	**85.9**	95.0

Note. Par = parameter; $\alpha_{g,1}$ = factor mean in group $g = 2, 3$; $\phi_{g,11}$ = factor variance in group $g = 2, 3$; $\nu_{g,i}$ = item intercept of item i in group $g = 2, 3$; estimation using the optimal regularization parameter based on the BIC; p = power used in the penalty function; λ = fixed regularized parameter; DIR = direct BIC minimization using the differentiable approximation of O'Neill and Burke (2023) with $\varepsilon = 0.01$; DA = differentiable approximation using the threshold parameter $\tau = 0.02$; Absolute biases larger than 0.04 are printed in bold. Coverage rates smaller than 91 or larger than 98 are printed in bold.

5. Simulation Study 2: Two-Dimensional Factor Model with Cross-Loadings

In Simulation Study 1, regularized ML estimation of a two-dimensional factor model with cross-loadings was investigated.

5.1. Method

The data-generating method involves a two-dimensional factor model involving ten manifest variables $X_1, \ldots, X_{10}$ (i.e., items), and two latent (factor) variables η_1 and η_2. The data-generating model is graphically presented in Figure 3. The first five items load on the first factor, while the last five items load on the second factor. Three cross-loadings for items X_1, X_9, and X_{10} were introduced.

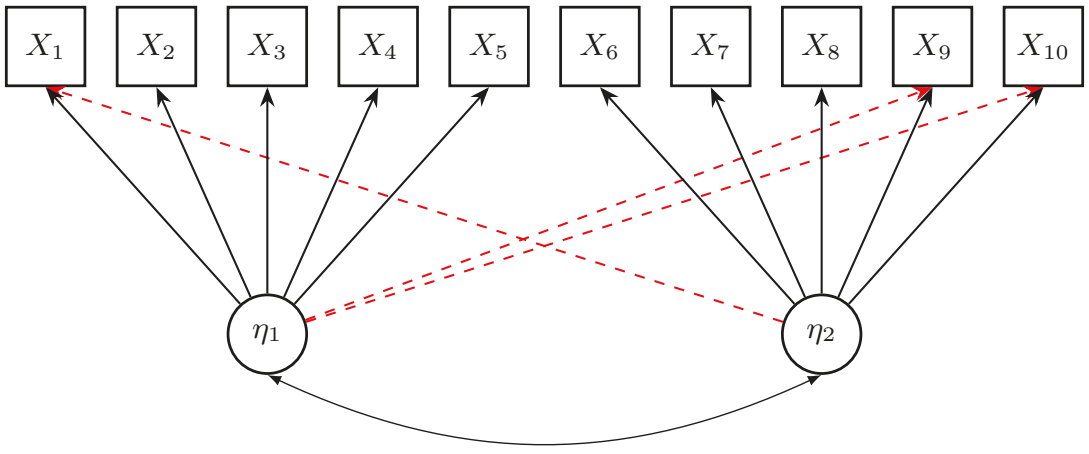

Figure 3. Simulation Study 2: Data-generating model.

All variables had zero means and were normally distributed. Furthermore, the latent variables η_1 and η_2 were standardized (i.e., they had a true variance of 1). The true factor correlation ϕ_{12} of the two factor variables was set to 0.5. The primary factor loadings of the ten items were 1.000, 0.858, 0.782, 0.877, 0.888, 1.000, 0.815, 0.721, 0.880, and 0.749. The variances of the normally distributed residual error variables were chosen as 0.115, 0.464, 0.572, 0.345, 0.411, 0.122, 0.536, 0.680, 0.383, and 0.627.

All cross-loadings were simulated with the size δ. In the simulation, δ was chosen as 0.2 or 0.4. Furthermore, the sample size N was chosen to be either 500 or 1000.

The two-dimensional factor model with the SCAD penalty function on the cross-loadings was specified as the analysis model. For identification reasons, the variances of the

factor variables were fixed to 1. The estimation method followed those used in Simulation Study 1. Again, $\varepsilon = 10^{-4}$ was employed in the differentiable approximation method (DA) utilizing thresholds $\tau = 0.01$, 0.02, and 0.04 in the BIC minimization. The smoothed direct BIC minimization approach (DIR) was again conducted with $\varepsilon = 0.01$.

In total, 1000 replications were conducted for all 2 (size of cross-loadings δ) $\times$ 2 (sample size N) = 4 conditions of the simulation study. We analyzed the estimation quality of model parameter estimates through bias, RMSE, and coverage rates.

The models were again estimated using the `sirt::mgsem()` function in the R [45] package sirt [46]. Replication material and the data-generating parameters can be found in the directory "*Simulation Study 2*" located at https://osf.io/7kzgb (accessed on 21 August 2023) .

5.2. Results

Figures 4 and 5 display the absolute bias and the RMSE of the factor correlation $\phi_{2,11}$ of the two factors as a function of the regularization parameter λ for the two sample sizes, $N = 500$ and $N = 1000$, and the two powers of the penalty function $p = 1$ and $p = 0.5$, respectively. In contrast to Simulation Study 1, the two figures showed that using a regularization parameter λ that is smaller than the optimal λ selected by BIC resulted in parameter estimates with a smaller RMSE. This property somehow questions the standard procedure of searching for a parsimonious model according to the BIC if a structural parameter should be estimated with low variance. Furthermore, parameter estimates with the power $p = 0.5$ in the penalty function resulted in a lower RMSE than using the power $p = 1$.

Table 4 presents the bias and the RMSE of the estimated factor correlation and the four factor loadings for the first two items. According to the findings in Figures 4 and 5, we also display the parameter estimates for the fixed regularization parameter $\lambda = 0.025$. The estimated parameters were unbiased. Hence, we focus on differences between the estimation approaches regarding the RMSE. Like in Simulation Study 1, the exact estimation approach (CD) performed similarly to the differentiable estimation approach (DA). The only exception was the case of $p = 0.5$ and the fixed regularization parameter $\lambda = 0.025$. In this case, the DA approach resulted in fewer variables estimated than the CD approach for the factor correlation $\phi_{2,11}$. Moreover, the direct BIC minimization approach (DIR) was similar or superior to the CD and DA estimation approaches based on BIC. This is an interesting finding because repeatedly fitting the regularized SEM on a grid of regularization parameters is not required in DIR.

Table 4. Simulation Study 2: Bias and root mean square error (RMSE) of factor means as a function of sample size N and the size of cross-loadings δ.

				$p = 1$				$p = 0.5$			
			BIC	**BIC**		**$\lambda = 0.025$**		BIC		**$\lambda = 0.025$**	
Par	***N***	***δ***	**DIR**	**CD**	**DA**	**CD**	**DA**	**CD**	**DA**	**CD**	**DA**
			Bias								
ϕ_{12}	500	0.2	0.000	0.011	0.009	−0.001	−0.001	0.001	0.001	0.000	0.000
		0.4	0.000	0.001	0.000	0.000	−0.001	0.001	0.001	0.000	0.000
	1000	0.2	0.000	0.002	0.001	−0.001	−0.001	0.000	0.000	0.000	0.000
		0.4	−0.001	0.000	−0.001	−0.001	−0.001	0.000	0.000	0.000	0.000
λ_{11}	500	0.2	−0.001	−0.002	−0.002	−0.001	−0.001	0.000	−0.001	−0.001	−0.001
		0.4	0.000	−0.001	−0.001	0.000	0.000	0.000	0.000	−0.001	−0.001
	1000	0.2	0.000	0.000	−0.001	0.000	0.000	0.000	0.000	0.000	0.000
		0.4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	−0.001

Table 4. *Cont.*

				$p = 1$				$p = 0.5$			
			BIC	**BIC**		**λ = 0.025**		BIC		**λ = 0.025**	
Par	***N***	***δ***	**DIR**	**CD**	**DA**	**CD**	**DA**	**CD**	**DA**	**CD**	**DA**
λ_{12}	500	0.2	−0.001	−0.006	−0.005	0.000	0.000	−0.001	−0.001	0.000	0.000
		0.4	−0.001	−0.002	−0.001	−0.001	−0.001	−0.002	−0.001	0.000	0.000
	1000	0.2	−0.001	−0.002	−0.001	0.000	0.000	−0.001	0.000	0.000	0.000
		0.4	−0.002	−0.002	−0.002	−0.002	−0.001	−0.002	−0.002	−0.001	−0.001
λ_{21}	500	0.2	−0.001	−0.003	−0.002	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001
		0.4	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.002
	1000	0.2	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001
		0.4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
λ_{22}	500	0.2	0.000	0.000	0.000	0.001	0.001	0.000	0.000	0.001	0.001
		0.4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002
	1000	0.2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001
		0.4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			RMSE								
ϕ_{12}	500	0.2	0.040	0.049	0.048	0.044	0.043	0.041	0.041	0.028	0.024
		0.4	0.039	0.040	0.040	0.043	0.043	0.040	0.040	0.029	0.024
	1000	0.2	0.028	0.030	0.030	0.030	0.029	0.028	0.028	0.022	0.017
		0.4	0.028	0.028	0.028	0.030	0.029	0.028	0.028	0.023	0.017
λ_{11}	500	0.2	0.040	0.042	0.042	0.041	0.041	0.040	0.040	0.034	0.034
		0.4	0.040	0.041	0.041	0.041	0.041	0.040	0.040	0.034	0.034
	1000	0.2	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.024	0.023
		0.4	0.029	0.029	0.029	0.030	0.029	0.029	0.029	0.025	0.024
λ_{12}	500	0.2	0.031	0.038	0.037	0.035	0.035	0.032	0.032	0.029	0.027
		0.4	0.032	0.033	0.033	0.036	0.035	0.033	0.033	0.029	0.027
	1000	0.2	0.022	0.024	0.023	0.023	0.023	0.022	0.022	0.022	0.019
		0.4	0.023	0.023	0.023	0.025	0.024	0.024	0.024	0.022	0.020
λ_{21}	500	0.2	0.042	0.043	0.043	0.045	0.045	0.042	0.042	0.043	0.042
		0.4	0.042	0.042	0.042	0.045	0.045	0.042	0.042	0.043	0.043
	1000	0.2	0.030	0.030	0.030	0.031	0.031	0.030	0.030	0.030	0.030
		0.4	0.030	0.029	0.030	0.031	0.031	0.030	0.030	0.030	0.030
λ_{22}	500	0.2	0.010	0.018	0.018	0.029	0.029	0.013	0.012	0.036	0.035
		0.4	0.011	0.009	0.010	0.029	0.029	0.011	0.010	0.036	0.035
	1000	0.2	0.006	0.009	0.009	0.015	0.014	0.007	0.007	0.025	0.024
		0.4	0.007	0.005	0.004	0.016	0.015	0.008	0.007	0.026	0.024

Note. Par = parameter; ϕ_{12} = factor correlation; λ_{id} = factor loading of ith item on the dth factor; estimation using the optimal regularization parameter based on the BIC; p = power used in the penalty function; λ = fixed regularized parameter; DIR = direct BIC minimization using the differentiable approximation of O'Neill and Burke (2023) with $\varepsilon = 0.01$; CD = coordinate descent; DA = differentiable approximation using the threshold parameter $\tau = 0.02$.

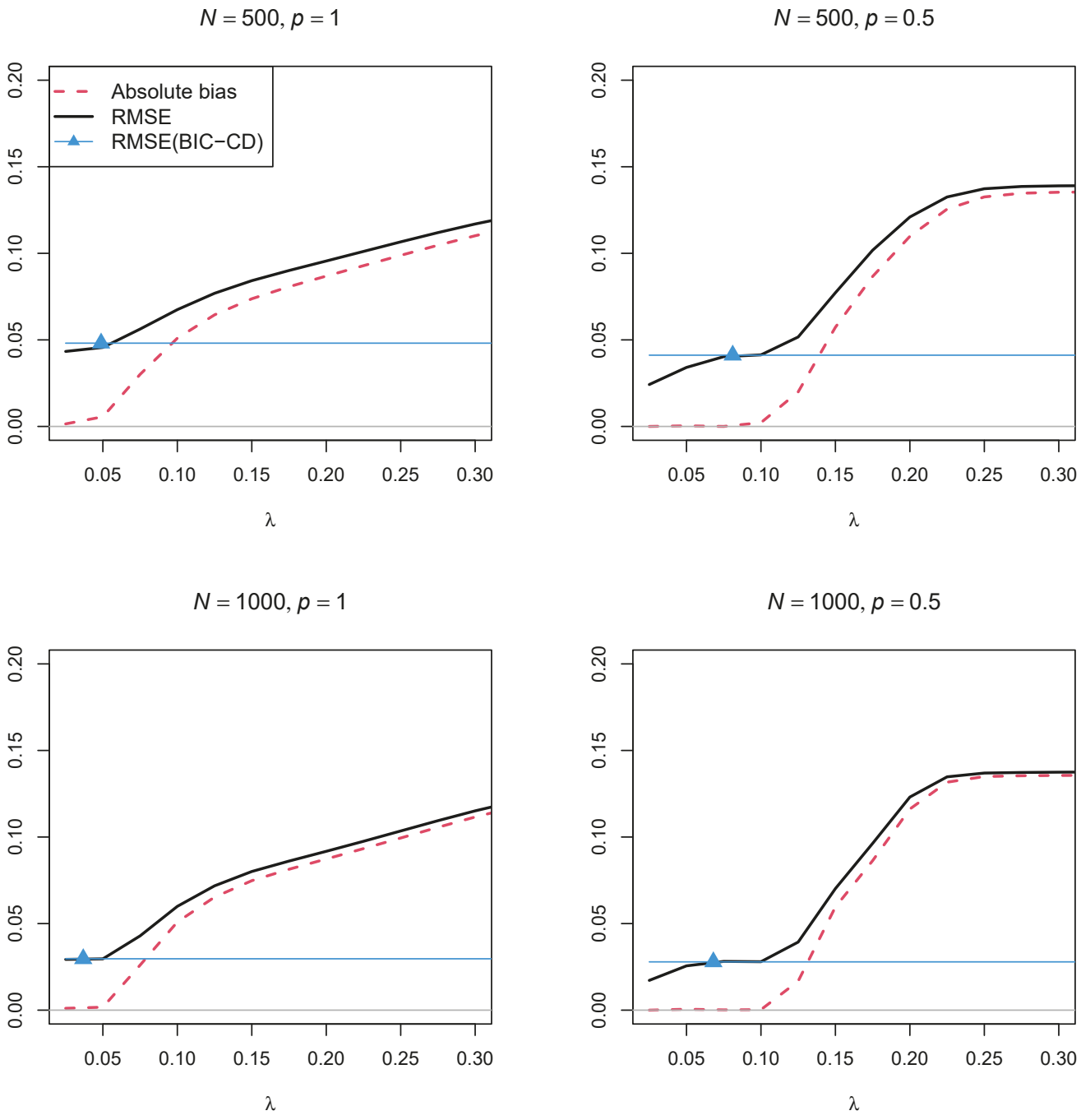

Figure 4. Simulation Study 2: Absolute bias and root mean square error (RMSE) of the factor correlation ϕ_{12} as a function of the regularization parameter λ for a cross-loading of $\delta = 0.2$ for sample sizes $N = 500$ and $N = 1000$ and powers $p = 1$ and $p = 0.5$ of the penalty function. The RMSE of the estimate obtained by the optimal BIC and coordinate descent (BIC-CD) is displayed by the blue line. The location of the average optimal regularization parameter obtained by BIC-CD is displayed by the blue triangle.

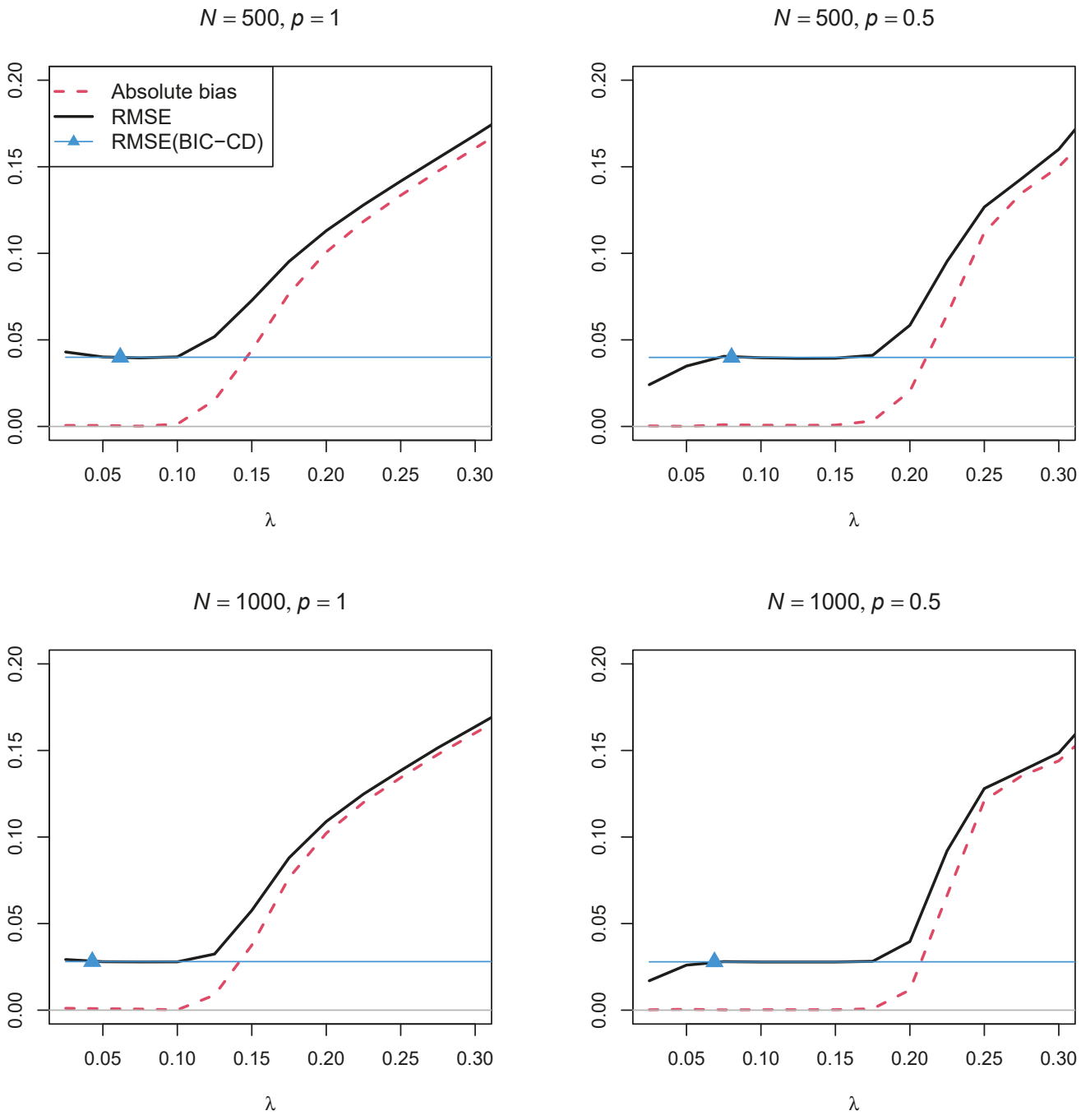

Figure 5. Simulation Study 2: Absolute bias and root mean square error (RMSE) of the factor correlation ϕ_{12} as a function of the regularization parameter λ for a cross-loading of $\delta = 0.4$ for sample sizes $N = 500$ and $N = 1000$ and powers $p = 1$ and $p = 0.5$ of the penalty function. The RMSE of the estimate obtained by the optimal BIC and coordinate descent (BIC-CD) is displayed by the blue line. The location of the average optimal regularization parameter obtained by BIC-CD is displayed by the blue triangle.

Table 5 displays the average number of regularized cross-loadings. It turned out that the choice of the threshold parameter τ was less critical for $p = 0.5$ than for $p = 1$. Furthermore, using $\tau = 0.02$ in the DA approach resulted in a similar average number of regularized parameters to the exact approach (CD).

Table 5. Simulation Study 2: Average number of regularized cross-loadings using the optimal regularized parameter λ based on the BIC as a function of sample size N and the size of cross-loadings δ.

		$p = 1$				$p = 0.5$			
			DA with $\tau =$				DA with $\tau =$		
N	δ	CD	0.01	0.02	0.04	CD	0.01	0.02	0.04
500	0.3	6.50	5.85	6.64	7.03	6.98	7.01	7.01	6.99
	0.6	6.92	6.72	6.94	6.98	6.94	6.95	6.95	6.94
1000	0.3	6.72	6.35	6.86	6.99	6.95	6.96	6.96	6.91
	0.6	6.97	6.95	6.99	6.99	6.95	6.95	6.95	6.90

Note. p = power used in the penalty function; CD = coordinate descent; DA = differentiable approximation using a threshold parameter τ.

Finally, Table 6 displays the coverage rates and bias of the estimated factor correlation and the factor loading of the first item. Coverage rates were satisfactory for the DIR approach as well as for the DA approach with fixed regularization parameters. There was a tendency of overcoverage for the power $p = 0.5$ for a small regularization parameter $\lambda = 0.025$.

Table 6. Simulation Study 2: Bias and coverage rates as a function of sample size N and the size of cross-loadings δ.

			Bias					Coverage				
				$p = 1$		$p = 0.5$			$p = 1$		$p = 0.5$	
			BIC	DA with λ =		DA with λ =		BIC	DA with λ =		DA with λ =	
Par	N	δ	DIR	0.025	0.05	0.025	0.05	DIR	0.025	0.05	0.025	0.05
ϕ_{12}	500	0.2	0.00	0.00	0.00	0.00	0.00	94.2	96.2	93.8	97.3	97.9
		0.4	0.00	0.00	0.00	0.00	0.00	94.3	96.0	94.7	98.0	98.3
	1000	0.2	0.00	0.00	0.00	0.00	0.00	95.1	95.9	94.7	97.0	97.0
		0.4	0.00	0.00	0.00	0.00	0.00	95.1	95.8	95.1	97.7	96.9
λ_{11}	500	0.2	0.00	0.00	0.00	0.00	0.00	94.4	95.4	94.6	**99.6**	96.7
		0.4	0.00	0.00	0.00	0.00	0.00	94.9	95.8	95.0	**99.7**	96.9
	1000	0.2	0.00	0.00	0.00	0.00	0.00	94.7	94.9	94.5	**99.8**	95.9
		0.4	0.00	0.00	0.00	0.00	0.00	95.1	95.4	95.1	**99.9**	96.4

Note. Par = parameter; ϕ_{12} = factor correlation; λ_{id} = factor loading of ith item on the dth factor; estimation using the optimal regularization parameter based on the BIC; p = power used in the penalty function; λ = fixed regularized parameter; DIR = direct BIC minimization using the differentiable approximation of O'Neill and Burke (2023) with $\varepsilon = 0.01$; DA = differentiable approximation using the threshold parameter $\tau = 0.02$; Absolute biases larger than 0.04 are printed in bold. Coverage rates smaller than 91 or larger than 98 are printed in bold.

6. Summary of Simulation Findings

In this section, the main findings of the two simulation studies are discussed regarding the research questions posed in Section 3.

6.1. RQ1: Using a Fixed Regularization Parameter λ Can Be Advantageous Regarding Bias and RMSE

First, the findings of Simulation Study 2 demonstrated that using a fixed regularized parameter λ instead of an optimally chosen λ by means of minimizing BIC can result in more efficient estimates of structural parameters (research question RQ1). This finding undermines the fact that obtaining efficient parameter estimates is not necessarily related to the search for a parsimonious model in terms of minimal information criteria.

6.2. RQ2: Differentiable Approximations of Penalty Functions Generally Work

Second, differentiable approximation approaches of the non-differentiable penalty function in regularized estimation using appropriate tuning parameters performed similarly to exact estimation approaches that employed coordinate descent (research question RQ2). This result contradicts recommendations in the recent literature that differentiable approximation approaches should generally be avoided. Note that the latter approaches can utilize widely availably general-purpose optimizers. Moreover, regularized estimation with the differentiable approximation of the penalty function is frequently faster than specialized optimizers.

6.3. RQ3: Direct BIC Minimization Is a Competitive Estimation Method

Third, even if differentiable approximation were used in regularized estimation, selecting the optimal regularization parameter λ based on the minimal BIC requires repeated estimation of the SEM which can be computationally demanding. A recently proposed smooth direct BIC minimization approach by O'Neill and Burke [17] avoids the specification of a regularization parameter and directly minimizes a smoothed version of the BIC. In our simulation studies that involved SEMs, the direct BIC minimization approach performed surprisingly well and had similar performance to the ordinarily employed indirect BIC minimization approach that requires repeated estimations (research question RQ3). This finding is remarkable because it could change the practice of the implementation of regularized estimation.

6.4. RQ4: The Power $p = 0.5$ in the Penalty Function Can Be Sometimes Beneficial

Fourth, the choice of the power p of the penalty function is ordinarily $p = 1$, but a recent implementation used $p = 0.5$ in SEM. Our simulations demonstrated similar performance of both power values (research question RQ4). However, in Simulation Study 2, $p = 0.5$ with a particular fixed regularization parameter λ outperformed the estimation based on $p = 1$. Moreover, the determination of the number of estimated parameters depended less on a chosen threshold for $p = 0.5$ than $p = 1$ if a differentiable approximation of regularized estimation was utilized.

6.5. RQ5: Reliable Standard Error Estimation Using the Delta Method

Fifth, our simulation studies demonstrated that standard error computation based on the delta method was satisfactory for the direct BIC minimization approach as well as for regularized estimation with a fixed regularization parameter λ (research question RQ5).

7. Discussion and Conclusions

In this article, implementation aspects of regularized maximum likelihood estimation of SEMs were investigated. We obtained some insights into how regularized SEMs could be efficiently implemented in practice. In contrast to statements in the literature, differentiable approximations of the non-differentiable penalty functions in regularized SEM perform comparably well to specialized estimation methods if tuning parameters in these approximations are thoughtfully chosen.

Our preliminary conclusion for regularized SEM estimation from our simulation studies is that the direct BIC minimization approach or the fixed regularization parameter approach should deserve more attention in future research. By focusing on these approaches, the computational burden of regularized SEM is noticeably reduced. Future research might investigate whether the findings obtained for SEMs transfer to other models involving latent variables such as item response models [47–52], latent class models [28,53–55], or mixture models [56,57].

In this article, we focused on a differentiable approximation of the BIC. However, the same approximation technique could be applied to estimating regularized SEMs that minimize the AIC. However, we have preliminary simulation evidence that convergence

issues appeared more frequently when minimizing the differentiable approximation of the AIC compared to BIC.

Hopefully, the availability of the direct BIC minimization approach could lead to more widespread use of regularized estimation. Nevertheless, the regularization approaches discussed in this paper still hinge on the assumption that there is sparsity with respect to regularized model parameters. Such sparse models or parameter deviations might not always be appropriate for modeling real-world datasets.

The simulation studies showed that the optimal regularization parameter λ regarding the bias and RMSE of the model parameters of interest does not necessarily coincide with the optimal λ obtained by minimizing the BIC. Determining the optimal regularization parameter λ for particular regularized SEMs is, therefore, difficult for researchers. Maybe only simulation studies that involve a similarly complex model and a similar sample size could help to determine an appropriate λ. If the researcher's interest lies in the interpretation of model parameters in a regularized SEM, it is uncertain as to why model fitting is aimed at minimizing a prediction error, as in BIC, because such a criterion can only be weakly related to estimating optimal model parameters (see [58]).

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIC	Akaike information criterion
BIC	Bayesian information criterion
CD	coordinate descent
CFA	confirmatory factor analysis
DA	differentiable approximation
DGM	data-generating model
DIF	differential item functioning
LASSO	least absolute shrinkage and selection operator
ML	maximum likelihood
RMSE	root mean square error
SCAD	smoothly clipped absolute deviation
SEM	structural equation model

References

1. Bartholomew, D.J.; Knott, M.; Moustaki, I. *Latent Variable Models and Factor Analysis: A Unified Approach*; Wiley: New York, NY, USA, 2011. [CrossRef]
2. Bollen, K.A. *Structural Equations with Latent Variables*; Wiley: New York, NY, USA, 1989. [CrossRef]
3. Browne, M.W.; Arminger, G. Specification and estimation of mean-and covariance-structure models. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences*; Arminger, G., Clogg, C.C., Sobel, M.E., Eds.; Springer: Boston, MA, USA, 1995; pp. 185–249. [CrossRef]
4. Jöreskog, K.G.; Olsson, U.H.; Wallentin, F.Y. *Multivariate Analysis with LISREL*; Springer: Basel, Switzerland, 2016. [CrossRef]
5. Kaplan, D. *Structural Equation Modeling: Foundations and Extensions*; Sage: Thousand Oaks, CA, USA, 2009. [CrossRef]
6. Shapiro, A. Statistical inference of covariance structures. In *Current Topics in the Theory and Application of Latent Variable Models*; Edwards, M.C., MacCallum, R.C., Eds.; Routledge: London, UK, 2012; pp. 222–240. [CrossRef]
7. Yuan, K.H.; Bentler, P.M. Structural equation modeling. In *Handbook of Statistics*; Rao, C.R., Sinharay, S., Eds.; Elsevier: Amsterdam, The Netherlands, 2007; Volume 26, pp. 297–358. [CrossRef]
8. Magnus, J.R.; Neudecker, H. *Matrix Differential Calculus With Applications in Statistics and Econometrics*; Wiley: New York, NY, USA, 2019. [CrossRef]
9. Bollen, K.A.; Davis, W.R. Two rules of identification for structural equation models. *Struct. Equ. Model.* **2009**, *16*, 523–536. [CrossRef]

10. Drton, M.; Foygel, R.; Sullivant, S. Global identifiability of linear structural equation models. *Ann. Stat.* **2011**, *39*, 865–886. [CrossRef]
11. Jacobucci, R.; Grimm, K.J.; McArdle, J.J. Regularized structural equation modeling. *Struct. Equ. Model.* **2016**, *23*, 555–566. [CrossRef] [PubMed]
12. Robitzsch, A. Model-robust estimation of multiple-group structural equation models. *Algorithms* **2023**, *16*, 210. [CrossRef]
13. Brandmaier, A.M.; Jacobucci, R.C. Machine learning approaches to structural equation modeling. In *Handbook of Structural Equation Modeling*; Hoyle, R.H., Ed.; Guilford Press: New York, NY, USA, 2023; pp. 722–739.
14. Orzek, J.H.; Arnold, M.; Voelkle, M.C. Striving for sparsity: On exact and approximate solutions in regularized structural equation models. *Struct. Equ. Model.* **2023**, *Epub ahead of print*. [CrossRef]
15. Li, X.; Jacobucci, R.; Ammerman, B.A. Tutorial on the use of the regsem package in R. *Psych* **2021**, *3*, 579–592. [CrossRef]
16. Asparouhov, T.; Muthén, B. Penalized Structural Equation Models. Technical Report. 2023. Available online: https://rb.gy/tbaj7 (accessed on 28 March 2023).
17. O'Neill, M.; Burke, K. Variable selection using a smooth information criterion for distributional regression models. *Stat. Comput.* **2023**, *33*, 71. [CrossRef]
18. Boos, D.D.; Stefanski, L.A. *Essential Statistical Inference*; Springer: New York, NY, USA, 2013. [CrossRef]
19. Kolenikov, S. Biases of parameter estimates in misspecified structural equation models. *Sociol. Methodol.* **2011**, *41*, 119–157. [CrossRef]
20. White, H. Maximum likelihood estimation of misspecified models. *Econometrica* **1982**, *50*, 1–25. [CrossRef]
21. Huang, P.H.; Chen, H.; Weng, L.J. A penalized likelihood method for structural equation modeling. *Psychometrika* **2017**, *82*, 329–354. [CrossRef]
22. Xu, Z.; Chang, X.; Xu, F.; Zhang, H. $L_{1/2}$ regularization: A thresholding representation theory and a fast solver. *IEEE Trans. Neur. Net. Lear.* **2012**, *23*, 1013–1027. [CrossRef]
23. Hastie, T.; Tibshirani, R.; Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*; CRC Press: Boca Raton, FL, USA, 2015. [CrossRef]
24. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
25. Fan, J.; Li, R.; Zhang, C.H.; Zou, H. *Statistical Foundations of Data Science*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2020. [CrossRef]
26. Zhang, H.; Li, S.J.; Zhang, H.; Yang, Z.Y.; Ren, Y.Q.; Xia, L.Y.; Liang, Y. Meta-analysis based on nonconvex regularization. *Sci. Rep.* **2020**, *10*, 5755. [CrossRef] [PubMed]
27. Huang, P.H. A penalized likelihood method for multi-group structural equation modelling. *Br. Math. Stat. Psychol.* **2018**, *71*, 499–522. [CrossRef] [PubMed]
28. Chen, Y.; Li, X.; Liu, J.; Ying, Z. Regularized latent class analysis with application in cognitive diagnosis. *Psychometrika* **2017**, *82*, 660–692. [CrossRef] [PubMed]
29. Zhang, Y.; Li, R.; Tsai, C.L. Regularization parameter selections via generalized information criterion. *J. Am. Stat. Assoc.* **2010**, *105*, 312–323. [CrossRef] [PubMed]
30. Chen, J. Partially confirmatory approach to factor analysis with Bayesian learning: A LAWBL tutorial. *Struct. Equ. Model.* **2022**, *22*, 800–816. [CrossRef]
31. Geminiani, E.; Marra, G.; Moustaki, I. Single- and multiple-group penalized factor analysis: A trust-region algorithm approach with integrated automatic multiple tuning parameter selection. *Psychometrika* **2021**, *86*, 65–95. [CrossRef]
32. Hirose, K.; Terada, Y. Sparse and simple structure estimation via prenet penalization. *Psychometrika* **2022**, *Epub ahead of print*. [CrossRef]
33. Huang, P.H. lslx: Semi-confirmatory structural equation modeling via penalized likelihood. *J. Stat. Softw.* **2020**, *93*, 1–37. [CrossRef]
34. Scharf, F.; Nestler, S. Should regularization replace simple structure rotation in exploratory factor analysis? *Struct. Equ. Model.* **2019**, *26*, 576–590. [CrossRef]
35. Battauz, M. Regularized estimation of the nominal response model. *Multivar. Behav. Res.* **2020**, *55*, 811–824. [CrossRef] [PubMed]
36. Oelker, M.R.; Tutz, G. A uniform framework for the combination of penalties in generalized structured models. *Adv. Data Anal. Classif.* **2017**, *11*, 97–120. [CrossRef]
37. Robitzsch, A. Comparing the robustness of the structural after measurement (SAM) approach to structural equation modeling (SEM) against local model misspecifications with alternative estimation approaches. *Stats* **2022**, *5*, 631–672. [CrossRef]
38. Tutz, G.; Gertheiss, J. Regularized regression for categorical data. *Stat. Model.* **2016**, *16*, 161–200. [CrossRef]
39. Oelker, M.R.; Pößnecker, W.; Tutz, G. Selection and fusion of categorical predictors with L_0-type penalties. *Stat. Model.* **2015**, *15*, 389–410. [CrossRef]
40. Phan, D.T.; Idé, T. l_0-regularized sparsity for probabilistic mixture models. In Proceedings of the 2019 SIAM International Conference on Data Mining, Calgary, AB, Canada, 2–4 May 2019; pp. 172–180. [CrossRef]
41. Shen, X.; Pan, W.; Zhu, Y. Likelihood-based selection and sharp parameter estimation. *J. Am. Stat. Assoc.* **2012**, *107*, 223–232. [CrossRef]

42. Shapiro, A. Statistical inference of moment structures. In *Handbook of Latent Variable and Related Models*; Lee, S.Y., Ed.; Elsevier: Amsterdam, The Netherlands, 2007; pp. 229–260. [CrossRef]
43. Asparouhov, T.; Muthén, B. Multiple-group factor analysis alignment. *Struct. Equ. Model.* **2014**, *21*, 495–508. [CrossRef]
44. Millsap, R.E. *Statistical Approaches to Measurement Invariance*; Routledge: New York, NY, USA, 2011. [CrossRef]
45. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2023. Available online: https://www.R-project.org/ (accessed on 15 March 2023).
46. Robitzsch, A. *sirt: Supplementary Item Response Theory Models*; R package version 3.13-228. 2023. Available online: https://CRAN.R-project.org/package=sirt (accessed on 11 August 2023).
47. Belzak, W.C. The multidimensionality of measurement bias in high-stakes testing: Using machine learning to evaluate complex sources of differential item functioning. *Educ. Meas.* **2023**, *42*, 24–33. [CrossRef]
48. Chen, Y.; Li, C.; Ouyang, J.; Xu, G. DIF statistical inference without knowing anchoring items. *Psychometrika* **2023**, *Epub ahead of print*. [CrossRef]
49. Robitzsch, A. Comparing robust linking and regularized estimation for linking two groups in the 1PL and 2PL models in the presence of sparse uniform differential item functioning. *Stats* **2023**, *6*, 192–208. [CrossRef]
50. Sun, J.; Chen, Y.; Liu, J.; Ying, Z.; Xin, T. Latent variable selection for multidimensional item response theory models via L_1 regularization. *Psychometrika* **2016**, *81*, 921–939. [CrossRef] [PubMed]
51. Tutz, G.; Schauberger, G. A penalty approach to differential item functioning in Rasch models. *Psychometrika* **2015**, *80*, 21–43. [CrossRef] [PubMed]
52. Zhang, S.; Chen, Y. Computation for latent variable model estimation: A unified stochastic proximal framework. *Psychometrika* **2022**, *87*, 1473–1502. [CrossRef] [PubMed]
53. Chen, Y.; Liu, J.; Xu, G.; Ying, Z. Statistical analysis of Q-matrix based diagnostic classification models. *J. Am. Stat. Assoc.* **2015**, *110*, 850–866. [CrossRef] [PubMed]
54. Robitzsch, A. Regularized latent class analysis for polytomous item responses: An application to SPM-LS data. *J. Intell.* **2020**, *8*, 30. [CrossRef] [PubMed]
55. Xu, G.; Shang, Z. Identifying latent structures in restricted latent class models. *J. Am. Stat. Assoc.* **2018**, *113*, 1284–1295. [CrossRef]
56. Robitzsch, A. Regularized mixture Rasch model. *Information* **2022**, *13*, 534. [CrossRef]
57. Wallin, G.; Chen, Y.; Moustaki, I. DIF analysis with unknown groups and anchor items. *arXiv* **2023**, arXiv:2305.00961. [CrossRef]
58. Browne, M.W. Cross-validation methods. *J. Math. Psychol.* **2000**, *44*, 108–132. [CrossRef]