

## Shrinking Small Sample Problems in Multilevel Structural Equation Modeling via Regularization of the Sample Covariance Matrix

Julia-Kim Walther, Martin Hecht & Steffen Zitzmann

To cite this article: Julia-Kim Walther, Martin Hecht & Steffen Zitzmann (2025) Shrinking Small Sample Problems in Multilevel Structural Equation Modeling via Regularization of the Sample Covariance Matrix, Structural Equation Modeling: A Multidisciplinary Journal, 32:1, 46-65, DOI: [10.1080/10705511.2024.2380919](https://doi.org/10.1080/10705511.2024.2380919)

To link to this article: <https://doi.org/10.1080/10705511.2024.2380919>



© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC



Published online: 09 Aug 2024.



Submit your article to this journal [↗](#)



Article views: 711



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 2 View citing articles [↗](#)

## Shrinking Small Sample Problems in Multilevel Structural Equation Modeling via Regularization of the Sample Covariance Matrix

Julia-Kim Walther<sup>a</sup> , Martin Hecht<sup>b</sup>  and Steffen Zitzmann<sup>c</sup> 

<sup>a</sup>University of Tübingen; <sup>b</sup>Helmut Schmidt University; <sup>c</sup>Medical School Hamburg

### ABSTRACT

Small sample sizes pose a severe threat to convergence and accuracy of between-group level parameter estimates in multilevel structural equation modeling (SEM). However, in certain situations, such as pilot studies or when populations are inherently small, increasing samples sizes is not feasible. As a remedy, we propose a two-stage regularized estimation approach designed for scenarios with both a small number of groups and small group sizes, and a low ICC. The method employs the wide format approach to multilevel SEM, where, at first, the sample covariance matrix is replaced by a shrinkage estimate, and then, this estimate is used to fit the SEM. By means of a simulation study, we evaluated the effectiveness of our two-stage approach. Our findings demonstrate that this method consistently ensures model convergence, provides more accurate between-level estimates, and even improves accuracy of within-level estimates in cases of very small group sizes.



### KEYWORDS

ICC; multilevel SEM; regularization; small samples

In psychology and the education sciences, observational units are often nested within higher-level units, such as students (level-1 units) within classes (level-2 units). Multilevel structural equation modeling (SEM) is a powerful tool for estimating parameters across these different levels. In the within-between framework used by common statistical software (e.g., Mplus and *lavaan*), parameters are decomposed into within-group level (e.g., student) and between-group level (e.g., class) components. Challenges arise when sample sizes are small at any level, leading traditional maximum likelihood estimation (MLE) methods to either fail to converge or produce highly inaccurate estimates of between-group level parameters (e.g., Hox et al., 2010; Hox & Maas, 2001; Lüdtke et al., 2008, 2011; McNeish & Stapleton, 2016; Meuleman & Billiet, 2009; Shin & Raudenbush, 2010; Stegmueller, 2013; Zitzmann, 2018; Zitzmann et al., 2015). However, collecting larger samples can be costly, time-consuming, or impractical for specific study designs, such as pilot studies with limited classes and students, or for certain populations, such as school boards. Moreover, small variances at the class (between-group) level, often expressed in relation to large variances at the student (within-group) level as low Intra Class Correlation (ICC), further lower convergence rates (Lüdtke et al., 2011; Zitzmann, 2018), and accuracy of class (between-group) level parameter estimates (Hox & Maas, 2001; Lüdtke et al., 2011; Muthen & Satorra,

1995; Zitzmann et al., 2021). Therefore, in scenarios with small samples and low ICCs, there is a need for an alternative approach that is straightforward to implement and can mend both convergence and accuracy issues.<sup>1</sup>

A broad category of methods, known as *regularization* encompasses techniques aimed at enhancing convergence and accuracy in statistical analyses. Originally developed by Tikhonov (1943) to address stability issues in inverse matrix problems, the concept of regularization was swiftly adopted by the statistical community to adapt traditional MLE to produce “reasonable answers in unstable situations” (Bickel et al., 2006, p. 272). “Unstable situations” cover a variety of scenarios, among them small sample sizes, where the goal is typically to minimize the chances of encountering degenerate matrices, inadmissible solutions, and models that either do not converge or yield highly inaccurate outcomes. Techniques commonly used involve refining traditional maximum likelihood estimation (MLE) by incorporating approaches such as shrinkage, constraints, fixed parameters, or penalties. Overall, the goals and techniques of regularization approaches differ considerably fairly. For instance, shrinkage estimation of the covariance matrix (e.g., Touloumis, 2015) aims at obtaining a well-behaved eigenstructure and more accurate estimates, whereas penalizing the objective function of estimators (e.g., P.-H. Huang et al., 2017; Jacobucci et al., 2016) is motivated by the goal of achieving more parsimonious models. Despite

**CONTACT** Julia-Kim Walther  [julia-kim.walther@uni-tuebingen.de](mailto:julia-kim.walther@uni-tuebingen.de)  Hector Research Institute of Education Sciences and Psychology, University of Tübingen, 72072 Tübingen, Germany.

<sup>1</sup>Note that, when research questions concern only within-group variation, fixed effect approaches might be an alternative solution for handling small sample settings. For example, one could dummy-code the groups such that one obtains fixed effects estimates for each group (see e.g., Allison, 2009; Muthen & Satorra, 1995; Snijders & Bosker, 2012). However, if one is interested in the variation of the between-group parameters, or if the number of groups is too large (given the separate modeling of between-group parameters for any group), then mixed-effects modeling, such as multilevel SEM, might be a more sensible choice.

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

their differences, all approaches share a common feature: they introduce a moderate amount of bias into estimation. This is driven by the principle of a “bias-variance tradeoff,” where the strategy is to reduce variance by accepting increased bias, thereby enhancing the overall accuracy of the estimation. We understand regularization as umbrella term for biased estimators that are employed in unstable situations.<sup>2</sup>

Within the general SEM framework, several regularization techniques have been employed to alleviate estimation problems encountered with small sample sizes. For instance, techniques such as constrained MLE and Bayesian estimation introduce bias into MLE by either limiting the range of possible parameter values, such as setting latent variances to one, or using weakly informative priors. These methods, which include contributions from Anderson and Gerbing (1984), Chen et al. (2001), and Zitzmann et al. (2022) for constrained MLE, and Depaoli and Clifton (2015), Zitzmann et al. (2016), and Ulitzsch et al. (2023) for Bayesian estimation, specifically target the calculation of model parameters, essentially the “output” of a structural equation model (SEM). However, in situations where non-convergence and poor estimation accuracy is contingent on a distorted eigenstructure of the sample covariance matrix<sup>3</sup>, essentially the “input” of a SEM, simply adjusting model parameters through regularization will not suffice.

In such cases, regularizing the sample covariance matrix itself may prove to be a more effective solution. The ridge method, widely used for addressing eigenstructure issues in the sample covariance matrix (Kamada et al., 2014; Yuan et al., 2011; Yuan & Chan, 2008), involves a subtle yet impactful adjustment: it adds a small value to its diagonal elements (i.e., variances of the observed variables). This technique has been demonstrated to significantly improve both the rate of convergence and, potentially, also the accuracy of estimations (Kamada et al., 2014; Kamada & Kano, 2012; Yuan & Bentler, 2017). However, employing techniques beyond ridging, which primarily yields a well-behaved eigenstructure in the sample covariance matrix, may enhance the accuracy of estimation a fortiori.

A considerable number of methods has been developed to regularize the sample covariance matrix in statistics, and applied research fields such as portfolio selection in finance and estimation of large covariance matrices in genomics. *Shrinkage estimation*, a key approach among these, has its origin in the work of Stein (1956), who highlighted the bias

in eigenvalues of the sample covariance matrix in small samples. The Steinian (or Stein-type) shrinkage technique creates a weighted average of the sample covariance matrix and a predetermined target matrix, which imposes a specific structure. For instance, using the identity matrix as the target suggests that variances are one, covariances are zero, and eigenvalues are one. The weighting shrinks the sample covariance matrix and their eigenvalues towards those of the target matrix. These approaches vary by the choice of target matrix and how the weighting (or shrinkage) parameter is calculated. In unstable scenarios with small sample size paired with a large number of observed variables (“small  $N$ , large  $p$ ”), shrinkage estimation has been shown to surpass traditional MLE in maintaining eigenstructure and improving accuracy (e.g., Ledoit & Wolf, 2004; Touloumis, 2015). Ledoit and Wolf (2012) concluded that without additional information on the true covariance matrix’s structure, shrinkage estimation has been arguably the most effective method so far (for an overview of shrinkage estimation see, e.g., Ledoit & Wolf, 2020).

Even though particularly promising, shrinkage estimation of the covariance matrix has been barely scrutinized in the context of SEM. Notable exceptions include studies by Arruda and Bentler (2017) and De Jonckere and Rosseel (2023), who explored its application in single-level SEM, and found it to enhance overall model evaluation, convergence and accuracy without significant computational costs. Despite these findings, evidence remains sparse, and in multilevel SEM, it is even more so. Here, Zitzmann et al. (2021) applied shrinkage (Bayesian) estimation to the between-group variance of the predictor in a bivariate two-level model which led to more accurate model parameters at the between-group level in small samples. The present article aims to examine the effectiveness of shrinkage estimation of the covariance matrix for handling small sample sizes and low ICCs in multilevel SEMs in a proof of concept manner. More specifically, it scrutinizes whether integrating shrinkage estimation into a two-stage SEM estimation approach improves convergence rates and the accuracy of between-group level parameter estimates. To explore this, we examine balanced, continuous two-level data using two-level intercept-only models by means of a simulation study. The article is structured as follows. Firstly, as we use the single-level CFA approach to multilevel SEM (Barendse & Rosseel, 2020; Mehta & Neale, 2005; Walther et al., 2024), which utilises the data in wide format (WF), we offer a concise overview of this approach. Secondly, we detail the shrinkage approach proposed by Touloumis (2015) that we have adopted in this study, and we elaborate on how it modifies the (co)variances at both levels when applied in multilevel SEM. Thirdly, we present the outcomes of our simulation study, discuss its implications, and suggest directions for future research.

## 1. Multilevel Structural Equation Modeling

Suppose, we observed four classes (number of groups  $g = 4$ ) with two students within each class (balanced group size  $n = 2$ ). This yields a total sample size of four students ( $N = g \cdot n = 4$ ). We investigate two observed variables

<sup>2</sup>Note that in psychology and the educational sciences, we might be more familiar with terms other than regularization. “Stabilization” is often used in the context of accuracy and model selection (e.g., Breiman, 1996; Ulitzsch et al., 2023; Zitzmann, 2018). “Smoothing” is a prominent term in the context of improving the eigenstructure of covariance matrices (e.g., Lorenzo-Seva & Ferrando, 2021; Wothke, 1993). More recently, “regularization” found its way into the mainstream literature to denote matters related to improper solutions, model sparsity, and overfitting (e.g., Arruda & Bentler, 2017; Jacobucci et al., 2016; Jung & Takane, 2007; Liang & Jacobucci, 2020; Orzek & Voelke, 2023; Williams & Rodriguez, 2022). However, there is no strict usage of the terms, and we are not aware of any consistent taxonomy.

<sup>3</sup>This means that the sample eigenvalues are more spread out compared to their population counterparts which makes non-invertible (i.e., singular, degenerate), non-positive definite matrices with large condition numbers more likely.

( $p = 2$ ), namely, engagement during class ( $x_1$ ), and performance in a test ( $x_2$ ). The whole *data set* is depicted in Panel A in Figure 1. We are interested in whether students within the same class show more similar engagement during class ( $x_1$ ) and performance in the test ( $x_2$ ) than students across different classes. In other words, we scrutinize whether variance at the class (between-group) level is substantially large compared to the student (within-group) level; that is, whether  $ICC > 0$ . The population models are depicted to the right in Panel A. For both variables,  $ICC = 0.05$ .

To analyze the data, we use a two-level intercept-only model. This model can be estimated through two different multilevel SEM approaches that mainly differ by their required *data format*: the long format (LF) approach (Muthén, 1990, 1994), and the wide format (WF) approach (Barendse & Rosseel, 2020; Mehta & Neale, 2005). The analytical and empirical equivalence of both approaches with MLE in terms of estimation accuracy has been demonstrated (Barendse & Rosseel, 2020; Mehta & Neale, 2005; Walther et al., 2024), and both methods can be implemented using the SEM package *lavaan* in R. However, given that the lesser-known WF approach is critical for our two-stage approach, we will focus on this approach in the following, highlighting the differences of the data format in terms of *data matrix*, *sample covariance matrix*, and the *model specification*. Nevertheless, we will consider the unregularized, standard LF approach in the simulation study. Details on model estimation and fitting functions for continuous variables are available in existing literature (e.g., Mehta & Neale, 2005).

### 1.1. The Wide Format (WF) Approach

The wide format (WF) approach essentially uses a single-level restricted CFA which is fitted to the total (two-level) data matrix in WF (WF-T). In WF-T, every observed variable  $p$  is split into every  $n^{\text{th}}$  unit (see Panel B), which we call “specific-units” variables in contrast to the  $p$  “all-units” variables in the LF approach. The rationale underneath is that “people [ $n$ ] are variables too” (Mehta & Neale, 2005, p. 1). For instance,  $x_{1,2}$  is engagement during class ( $x_1$ ) for every 2nd student in class. The sample covariance matrix is estimated by the MLE for single-level data. Thus, we obtain a single-level represented two-level sample covariance matrix  $\mathbf{S}_{WF-T}$  from the  $p \cdot n$  “specific-units” variables in WF-T (see Panel C).

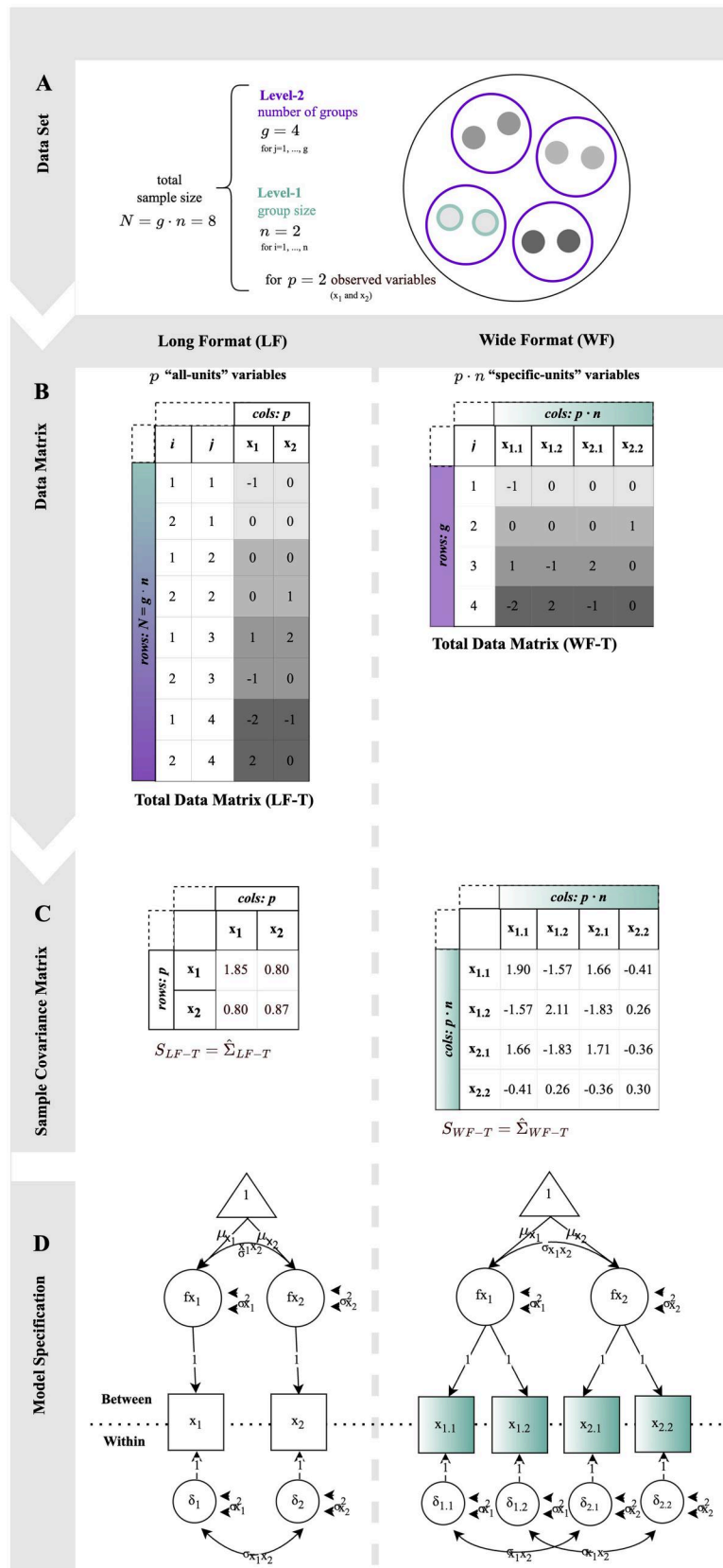
In the model, class (between-group) level parameters are modelled by common factors, and student (within-group) level parameters are modelled by unique factors that are equality constrained. The means and (co)variances of the common factors are estimated freely to obtain the class (between-group) level parameters. Variances of the unique factors of each common factor are equality constrained to estimate student (within-group) level variances. Covariances among unique factor of every  $n$ -th observed variable of each  $p$  are equality constrained to estimate student (within-group) level covariances (see Panel D). The equality constraints represent the homoscedasticity assumption (of the

specific-units variables). For our example, we would have two common factors (because of  $p = 2$  observed variables) with two observed variables for each common factor (because of  $p \cdot n$  specific-units variables in WF-T). Thus, two means, two variances, and one covariance of common factors for the class (between-group) level, and two variances, and one covariance of unique factors for the student (within-group) level are estimated freely.

The implementation of traditional MLE in SEM software such as *lavaan* requires a positive definite sample covariance matrix (Hamaker et al., 2003; Singer, 2010; Van Montfort et al., 2018; Voelkle et al., 2012). Amongst other things, this necessitates a data matrix whose number of columns is less than or equal to the number of rows because otherwise, at least one sample eigenvalue becomes zero and the sample covariance matrix turns non-positive definite (e.g., Duncan et al., 1997; Gorsuch, 1983; Wothke, 1993). In the WF approach,  $cols \leq rows$  translates to  $(p \cdot n) \leq g$ . Alternatively, the raw data formulation of MLE, full information maximum likelihood (FIML), may be used, which circumvents the problem (Hamaker et al., 2003; Trendafilov & Unkel, 2011; Unkel & Trendafilov, 2010; Voelkle et al., 2012). In *lavaan*, FIML estimation could be applied by setting ‘missing = “fiml”’. However, since we aim to replace the sample covariance matrix with a shrinkage estimate that has an improved eigenstructure, we must use traditional MLE instead of FIML. Moreover, we must use single-level SEM (i.e., the WF approach), because in multilevel SEM, such as implemented in *lavaan* version 0.6–15, we cannot provide a covariance matrix instead of a data matrix. We will turn towards shrinkage estimation in the subsequent section.

## 2. Shrinkage Estimation of the Covariance Matrix

In shrinkage estimation, the population covariance matrix  $\Sigma$  is estimated as a weighted average of the sample covariance matrix and a pre-specified target matrix. The amount of weighting is controlled by the shrinkage parameter  $\lambda \in [0, 1]$ . If  $\lambda = 0$ , no shrinkage is applied, and the sample covariance matrix will be kept. If  $\lambda = 1$ , we obtain the target matrix as the estimate of  $\Sigma$ . In linear shrinkage, which we focus on, the same shrinkage intensity is applied to every element of the covariance matrix. To avoid misunderstanding: “shrinkage” does not necessarily mean that the elements get smaller, but they are shrunken towards a certain value (of the target matrix). For example, if  $\sigma_S = 0.1$  and  $\sigma_T = 1$ , then 0.1 is “shrunken” towards 1. The target matrix is chosen for its well-behaved eigenstructure, making shrinkage estimates more likely to be positive definite, non-singular, and well-conditioned, often resulting in greater accuracy compared to the traditional ML sample covariance matrix, as demonstrated in studies such as Ledoit and Wolf (2004, 2020). Additionally, shrinkage estimation can be viewed as a form of Bayesian estimation with weakly informative priors, a perspective supported by Ledoit and Wolf (2004), and others. In the next section, we will briefly review the linear shrinkage estimator proposed by Touloumis (2015), which we term *covshrink* for convenience.



**Figure 1.** Data and model in the long format (LF) and wide format (WF) approach.

*Note. Data set:* the data collected in a given setting. *Data Matrix:* the data set in matrix form, where columns refer to observed variables and rows to observed units. *Data Format:* one of two possible formats of the data matrix, long format (LF) or wide format (WF). In WF, every observed variable  $p$  is split for every unit in the group  $n$ . For instance,  $x_{1,2}$  is  $x_1$  for every 2<sup>nd</sup> unit in the group. *Sample Covariance Matrix:* a symmetric matrix which contains (co)variances of the observed variables. *Model Specification:* representation of the model to be estimated, here, a two-level intercept-only model. Between-group parameters are located above the dashed line; within-group parameters below. At the within-group level, identical parameters indicate equality constraints. Data matrix or sample covariance matrix, and model specification are input to *lavaan*. Example data set with number of groups  $g = 4$ , group size  $n = 2$ , and number of observed variables  $p = 2$ . The R code to generate data and model is available in [Appendix A](#).

## 2.1. Covshrink: A Linear Shrinkage Estimator of the Covariance Matrix

Touloumis (2015) refined the popular linear shrinkage estimator by Ledoit and Wolf (2004) through (1) extending the set of target matrices, and (2) deriving consistent closed form solutions of the shrinkage parameters in “small  $N$ , large  $p$ ” settings. This new family of estimators has demonstrated improved estimation compared to preceding methods, indicated by the simulated percentage relative improvement in average loss (SPRIAL), which compares the MSE of the target estimator to that of a baseline estimator (e.g., the sample covariance matrix), in such settings (Touloumis, 2015). The general equation for the linear shrinkage estimator is expressed as:

$$\hat{\mathbf{S}}^* = (1 - \hat{\lambda})\mathbf{S} + \hat{\lambda}\mathbf{T}, \quad (1)$$

where  $\mathbf{S}$  is the unbiased MLE of the (single-level)  $p \times p$  population covariance matrix,  $\mathbf{T}$  is the target matrix, and  $\hat{\lambda}$  is the shrinkage parameter, which depends on the choice of the target matrix. The target matrix can be one of three diagonal matrices: the equal target matrix  $\hat{\nu}\mathbf{I}_p$  with the mean of the sample variances in the diagonal (originally proposed by Ledoit & Wolf, 2004), the identity matrix  $\mathbf{I}_p$  with ones in the diagonal, or the unequal target matrix  $\mathbf{D}_S$  with the sample variances in the diagonal. Across all types of target matrices, off-diagonal elements (i.e., covariances) of the shrinkage estimate are systematically pulled towards zero. However, the specific non-zero value to which on-diagonal elements (i.e., variances) are pulled varies depending on the target matrix employed. When using the equal target matrix  $\hat{\nu}\mathbf{I}_p$ , variances are pulled towards the mean of the sample variances. Meanwhile, the identity matrix  $\mathbf{I}_p$  pulls variances towards one, while the unequal target matrix  $\mathbf{D}_S$  leaves the variances unchanged. The closed form solution of the shrinkage parameter of the equal matrix  $\hat{\nu}\mathbf{I}_p$ , where  $\hat{\nu} = Y_{1N/p}$ , is:

$$\hat{\lambda}_E = \frac{Y_{2N} + Y_{1N}^2}{NY_{2N} + Y_{1N}^2 + \frac{p-N+1}{p}Y_{1N}^2}, \quad (2)$$

for the shrinkage parameter of the identity matrix  $\mathbf{I}_p$ :

$$\hat{\lambda}_I = \frac{Y_{2N} + Y_{1N}^2}{NY_{2N} + Y_{1N}^2 - (N-1)(2Y_{1N} - p)}, \quad (3)$$

and for the shrinkage parameter of the unequal target matrix  $\mathbf{D}_S$ :

$$\hat{\lambda}_U = \frac{Y_{2N} + Y_{1N}^2 - 2Y_{3N}}{NY_{2N} + Y_{1N}^2 - (N-1)Y_{3N}}, \quad (4)$$

where  $Y_{1N}$ ,  $Y_{2N}$ , and  $Y_{3N}$  are combinations of U-statistics (for their estimation, see Touloumis, 2015, pp. 5, 12). According to Touloumis (2015), the optimal shrinkage intensity, which minimizes the MSE between the population covariance matrix and the respective shrinkage estimator, is approximated by sample-based unbiased and ratio-consistent estimators. The resulting biased shrinkage estimators of  $\mathbf{\Sigma}$  are  $\hat{\mathbf{S}}_E^*$  (equal target matrix),  $\hat{\mathbf{S}}_I^*$  (identity target matrix), and  $\hat{\mathbf{S}}_U^*$  (unequal target matrix). Because the shrinkage parameters have a closed form, the approach is

computationally fast, regardless of the number of observed variables  $p$ . Moreover, the obtained estimates are non-singular and well-conditioned. These are useful properties for convergence (e.g., *lavaan* requires a positive definite sample covariance matrix in single-level SEM), and estimation accuracy (e.g., large condition numbers have been linked to the less stable estimates; Y. Huang & Bentler, 2015; Kelley, 1995; Lange et al., 1999; Yuan & Bentler, 2017).

## 3. Shrinkage Estimation of the Covariance Matrix in Multilevel Structural Equation Modeling

Shrinkage estimation of the covariance matrix is part of our two-stage approach.<sup>4</sup> At the first stage, the sample covariance matrix is replaced by a shrinkage estimate of  $\mathbf{\Sigma}$ . At the second stage, the model is estimated based on this refined estimate. Touloumis (2015) shrinkage estimator was optimized for “small  $N$ , large  $p$ ” scenarios which, can be translated to “small  $g$ , small  $n$ , large  $p$ ” configurations in the context of multilevel analysis. While this two-stage approach appears to be a resource-efficient strategy for addressing issues such as non-convergence and inaccurate between-group level parameter estimates resulting from small samples or low ICCs, only a limited body of research has investigated the performance of such methods within the SEM framework (Arruda & Bentler, 2017; De Jonckere & Rosseel, 2023; Zitzmann et al., 2021). Existing evidence suggests that similar two-stage approaches can indeed enhance convergence and estimation accuracy. However, such an approach has not yet been proposed and investigated in the context of multilevel SEM. In the subsequent section, we will delve deeper into how the (co)variances in the shrinkage estimate differ from those in the sample covariance matrix, and elucidate the implications for model parameters.

### 3.1. WFcovshrink: Shrinkage Estimation of the Covariance Matrix in the WF Approach

Recall that the WF approach is a single-level SEM approach that utilises the single-level represented two-level sample covariance matrix  $\mathbf{S}_{WF-T}$  where the (co)variances of  $p \cdot n$  “specific-units” variables are contained (revisit Figure 1 for more details). The normal theory derived, biased MLE reads:

$$\mathbf{S}_{WF-T} = \frac{1}{g} \sum_{j=1}^g (\mathbf{X}_{\cdot j} - \overline{\mathbf{X}}_{\cdot}) (\mathbf{X}_{\cdot j} - \overline{\mathbf{X}}_{\cdot})^T, \quad (5)$$

where  $\mathbf{X}_{\cdot j}$  denotes the data matrix in WF (WF-T) and  $\overline{\mathbf{X}}_{\cdot}$  denotes a row vector with grand mean estimates. The sample covariance matrix is the estimate of the population covariance matrix,  $\mathbf{S}_{WF-T} = \hat{\mathbf{\Sigma}}_{WF-T}$ . When shrinkage estimation of the covariance matrix is applied, the (single-level)  $p \times p$  dimensioned  $\mathbf{S}$  is replaced by the (single-level

<sup>4</sup>Note that in fact, every SEM is a two-stage approach as the sample covariance matrix has to be estimated in order to estimate the model parameters. Nonetheless, usually users supply the data matrix and the software estimates the sample covariance matrix automatically. Thus, from a user perspective, standard SEM can be considered a one-stage approach.

represented two-level)  $(p \cdot n) \times (p \cdot n)$  dimensional  $\mathbf{S}_{WF-T}$ <sup>5</sup> in Equation (1), and  $N$  by  $g$ , and  $p$  by  $p \cdot n$  in Equations (2), (3), and (4). Within the present study, we scrutinize all three target matrices, resulting in the shrinkage estimators with the equal target matrix,  $\hat{\mathbf{S}}_E^*$ , the identity target matrix,  $\hat{\mathbf{S}}_I^*$ , and the unequal target matrix,  $\hat{\mathbf{S}}_U^*$ .

For an illustration of the effect of the shrinkage estimation by Touloumis (2015) in the WF approach (WFcovshrink) in the following, we focus on  $\hat{\mathbf{S}}_E^*$ , see Figure 2. In Panel A, it is highlighted how  $\hat{\Sigma}_{WF-T}$  is used to model  $\hat{\theta}$ . In Panel B, the principle of how the shrinkage estimate  $\hat{\mathbf{S}}_E^*$  alters  $\hat{\theta}$  is explained. In Panel C, a concrete example is presented.

In Panel A, leftmost, we see the (earlier introduced) model specification of the two-level intercept-only model in the WF approach. A restricted CFA is fitted to the  $p \cdot n$  “specific-units” variables in the data matrix in WF. In the middle, the (co)variances of these  $p \cdot n$  “specific-units” variables in  $\mathbf{S}_{WF-T}$  are shown. To the right, these  $p \cdot n$  “specific-units” (co)variances are reformulated as the  $p$  “all-units” (co)variances that are modelled thereof. (Co)variances of  $x_{1,1}$  and  $x_{1,2}$  (see upper left, green block) are used to model the variances of one common and two unique factors which correspond to the between-group and within-group level variances of  $x_1$ . Their variances contribute to the between-group and within-group level variances, whereas their covariance contributes only to the between-group level variance via the common factor. Similarly, (co)variances of  $x_{2,1}$  and  $x_{2,2}$  (see lower right, green block) are used to model between-group and within-group level variances of  $x_2$ . The covariances of  $x_{1,1}$  and  $x_{1,2}$  with  $x_{2,1}$  and  $x_{2,2}$ , respectively (see lower left or upper right, orange block), are used to model the covariances of the two common factors, and every  $n$ -th unique factor of each common factor which correspond to between-group and within-group level covariances of  $x_1$  and  $x_2$ .

This reformulation helps to understand the principle of how shrinkage estimation with the equal target matrix alters the estimates of the two-level intercept-only model ( $\hat{\theta}$ ), which is illustrated in Panel B. To the left, the reformulated  $\mathbf{S}_{WF-T}$  is shown again. The on-diagonal elements of  $\mathbf{S}_{WF-T}$  (grey bar) are averaged ( $\hat{\nu}$ ) and used as the on-diagonal elements (“equal variances”) in the equal target matrix  $\hat{\nu}\mathbf{I}_{p \cdot n}$ . In reformulated terms,  $\hat{\nu}$  is the grand mean of the total variances of both variables  $x_1$  and  $x_2$  ( $\hat{\sigma}_B^2 + \hat{\sigma}_W^2$ ). The off-diagonal elements of  $\hat{\nu}\mathbf{I}_{p \cdot n}$  are zero. To the right, we see an overview of the directions in which the sample (co)variances in  $\mathbf{S}_{WF-T}$  are pulled by shrinkage estimation. Generally, on-diagonal elements are pulled towards the mean of the diagonal elements,

and off-diagonal elements are pulled towards zero. Using the reformulation, this means that total variances ( $\hat{\sigma}_B^2 + \hat{\sigma}_W^2$ ) are pulled towards the grand mean of the total variances ( $\overline{\hat{\sigma}_B^2 + \hat{\sigma}_W^2}$ ), and between-group variances ( $\hat{\sigma}_B^2$ ) are pulled towards zero. Consequently, within-group variances ( $\hat{\sigma}_W^2$ ) are pulled towards the grand mean of the total variances ( $\overline{\hat{\sigma}_B^2 + \hat{\sigma}_W^2}$ ), too. Between-group covariances ( $\hat{\sigma}_B$ ) and within-group covariances ( $\hat{\sigma}_W$ ) are pulled towards zero. The expected biases in  $\hat{\theta}$  are depicted in the rightmost table. It is expected that between-group level variances, and between-group and within-group covariances, have downward biases, whereas within-group variances have an upward biases. Therefore, estimates of ICC will be more conservative than those derived by the unregularized WF approach.

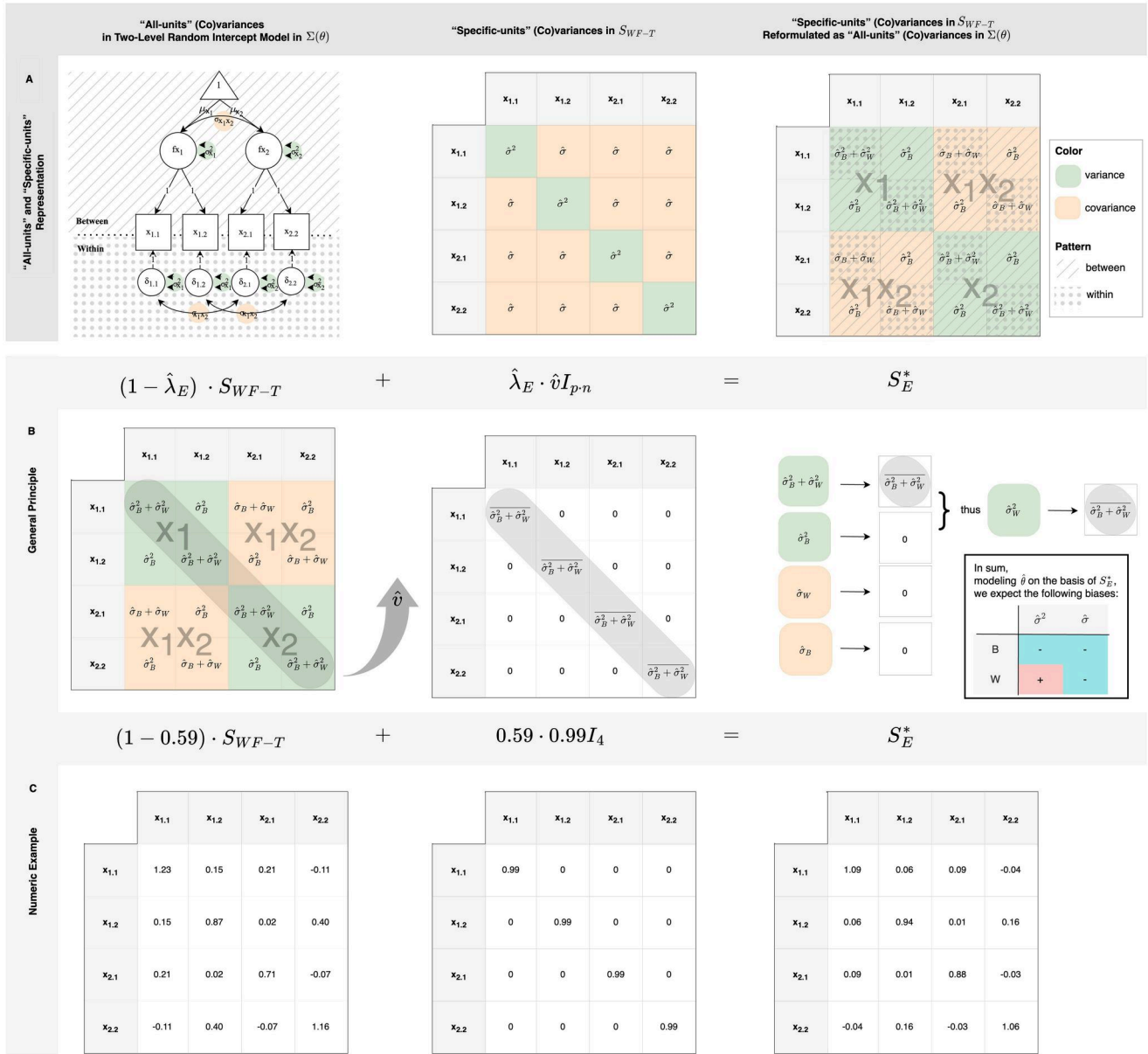
Let us consider this more concretely. In Panel C, WFcovshrink is illustrated by means of an example data set (in which  $g = 50$  in contrast to the earlier example data set). Leftmost,  $\mathbf{S}_{WF-T}$  (estimated by the unbiased MLE) is depicted. The middle of the panel shows the equal target matrix  $\hat{\nu}\mathbf{I}_{p \cdot n}$  where  $\hat{\nu} = 0.99$  (mean of the sample variances in  $\mathbf{S}_{WF-T}$ , or reformulated, grand mean of the total variances  $\overline{\hat{\sigma}_B^2 + \hat{\sigma}_W^2}$  of  $x_1$  and  $x_2$ ). In the present medium  $g$ , small  $n$ , large  $p$  setting, the shrinkage parameter is  $\hat{\lambda} = 0.59$ , and thus, the shrinkage estimate is to a large extent influenced by the target matrix. To the right, the resulting shrinkage estimate  $\mathbf{S}_E^*$  is presented. To comprehend how  $\mathbf{S}_E^*$  alters  $\hat{\theta}$ , we compare the model parameter estimates retrieved from the WF approach and the WFcovshrink approach (the R code is available in the Appendix A). In this example, population parameters are  $\sigma_B^2 = 0.05$ ,  $\sigma_W^2 = 0.95$ , thus  $ICC = 0.05$ , and  $\sigma_B = 0.015$ , and  $\sigma_W = 0.285$  for both variables  $x_1$  and  $x_2$ .

It can be seen in Table 1 that for the between-group level, over- and underestimation was decreased (by pulling the estimates closer to zero). For the within-group level, underestimation of the variance of  $x_1$  was decreased but overestimation slightly increased for  $x_1$  (by pulling the estimates of the variances closer to their grand mean), and overestimation of their covariance was decreased (by pulling the estimate closer to zero).

In both approaches, one estimate of variances at the between-group level was negative.<sup>6</sup> Concerning the resulting ICCs, the estimates of the WFcovshrink approach ( $0.06/(0.06 + 0.95) = 0.06$  and  $-0.04/(-0.04 + 1.02) = -0.04$ ) were more accurate than those of the unregularized

<sup>5</sup>Note that in the implementation of the shrinkage estimation in R (*ShrinkCovMat* package), only the unbiased MLE, which has  $g-1$  in the denominator, can be used. In contrast, the default of single-level SEM in *lavaan* (i.e., WF approach) is the normal theory derived, biased MLE in Equation 5 (Rosseel et al., 2023, reference manual p.81 accessed on 16 September 2023, `lav_matrix_cov` function). We run the unregularized WF approach with both the unbiased and the biased MLE to check whether they differ substantially. In Figure B1 in the Appendix we see that for convergence there were no differences, and for estimation accuracy, there were negligible differences in using the unbiased or biased estimator of the sample covariance matrix.

<sup>6</sup>The unregularized WF approach that uses the ML sample covariance matrix has high variability and low bias in small samples, whereas the shrinkage estimate in the WFcovshrink approach reduces variability by means of bias. Thus, in other data sets, the unregularized WF approach may yield non-negative, overestimated between-group level variances. There are different procedures to deal with inadmissible, negative between-group level variances (so called “Heywood cases”), that we would expect (and empirically found in the present study) more often in the downwardly biased WFcovshrink approach, for instance, setting them to zero (see e.g., Zitzmann et al., 2022, who justify the procedure by the very definition of MLE). However, non-negative, overestimated between-group level variance, which we might expect more often in the unregularized WF approach, are taken at face value. Thus, the downward bias in the WFcovshrink approach might be dealt with better (in addition to its estimates being more accurate).



**Figure 2.** Shrinkage estimation of the covariance matrix in the WF approach.

*Note.*  $S_{WF-T}$  contains (co)variances of  $p \cdot n$  "specific-units" variables. In Panel A, these are reformulated as (co)variances of  $p$  "all-units" variables modelled in the two-level intercept-only model. Panel B introduces the principle of how the shrinkage estimate with the equal target matrix alters estimates of the two-level intercept-only model. In Panel C, a numeric example with the earlier data set (number of groups  $g = 50$ , group size  $n = 2$ , and number of observed variables  $p = 2$ ) is given. The R code to generate the (unbiased) sample covariance matrix and apply shrinkage estimation is available in [Appendix A](#).

**Table 1.** Model parameter estimates of two-level intercept-only model for example data set.

Approach	Between			Within		
	$\sigma_{x_1}^2 = 0.05$	$\sigma_{x_2}^2 = 0.05$	$\sigma_{x_1x_2} = 0.015$	$\sigma_{x_1}^2 = 0.95$	$\sigma_{x_2}^2 = 0.95$	$\sigma_{x_1x_2} = 0.285$
$\hat{\theta}_{WF}$	0.15	-0.08	-0.05	0.88	1.01	0.35
$\hat{\theta}_{WFcovshrink\epsilon}$	0.06	-0.04	-0.02	0.95	1.02	0.15

*Note.* Example data set with number of groups  $g = 50$ , group size  $n = 2$ , and number of observed variables  $p = 2$ . The R code to generate data and estimate the models is available in [Appendix A](#).

WF approach  $(0.15)/(0.15 + 0.88) = 0.14$  and  $-0.08/(-0.08 + 1.01) = -0.08$ . In sum, all but one estimate are closer to their population counterparts in WFcovshrink compared to the unregularized WF approach. Nevertheless, this was just one example data set. Whether WFcovshrink yields empirically reliable similar gains in performance in other settings, and by means of other target matrices than the equal target matrix

$\hat{v}I_{p \cdot n}$ , remains to be put to test. We addressed these questions with a simulation study, which we will present next.

#### 4. Simulation Study

With this simulation study, we aimed to investigate whether applying shrinkage estimation, as part of the two-stage SEM



estimation approach, would increase convergence and estimation accuracy in multilevel SEM when small samples at any level or small ICCs are present. The idea is to obtain a biased but more precise estimate of the covariance matrix  $\Sigma$  that yields more accurate model parameters  $\hat{\theta}$  in turn. Specifically, we applied the shrinkage estimator by Touloumis (2015) to the WF approach in multilevel SEM. In the following, we outline the method of our study before presenting and discussing the main findings.

#### 4.1. Method

The computations were conducted on an AMD Ryzen Threadripper PRO 3975WX 32-cores (3.50 GHz) CPU on a Windows 10 (Version 20H2) platform utilising R version 4.3.1 (R Core Team, 2023), along with several R packages: *cowplot* version 1.1.1 (Wilke, 2020), *DescTools* version 0.99.50 (Signorell et al., 2024), *dplyr* version 1.1.2 (Wickham et al., 2023), *ggplot2* version 3.4.2 (Wickham et al., 2023), *huxtable* version 5.5.6 (Hugh-Jones, 2022), *lavaan* version 0.6-15 (Rosseel et al., 2023), *patchwork* version 1.1.2 (Pedersen, 2022), *ShrinkCovMat* version 1.4.0 (Touloumis, 2019), *tidyr* version 1.3.0 (Wickham et al., 2022), and *xlsx* version 0.6.5 (Dragulescu & Arendt, 2020). The R code for data generation, analysis, table, and figures is available at <https://github.com/demianJK/WFcovshrink>.

##### 4.1.1. Data Generation

We varied different factors that we allocate to either *sample characteristics* or *population characteristics* to facilitate interpretation. We make this distinction to emphasize what we can modify (by our study design) and what not. Sample characteristics comprise the number of groups  $g$ , the group size  $n$ , and the number of observed variables  $p$ . We included the following numbers of groups: 4, 10, 30, 50, and 100. The smallest number of groups is given by the minimum sample size that the R function for shrinkage estimation can deal with (which relates to  $g$  in the WF approach). The maximum number of groups was chosen to see how the WFcovshrink approaches perform in samples large enough to achieve good performance by the unregularized LF and WF approaches to multilevel SEM. The group size was varied between 2, 5, and 10. We restrained the upper group size to 10, because the WF approach is rather advised for smaller  $n$  scenarios (Barendse & Rosseel, 2020; Walther et al., 2024), because of larger computational costs, and preliminary simulations supported that this holds true for WFcovshrink as well. As numbers of observed variables  $p$ , we selected 2, 5, and 10. The population characteristics encompass the variances and covariances of the population covariance matrix at both the between- and within-group level. The variance at both levels was determined by the ICC, which is defined as the ratio of between-group variance to the total variance (Hox et al., 2017),  $\sigma_B^2/(\sigma_B^2 + \sigma_W^2)$ . Two levels of the ICC were included, 0.05 and 0.25, which represent the lower and upper levels of realistic ICCs in the social sciences (Adams et al., 2004; Gulliford et al., 1999).

The total variances were set to 1, and thus,  $ICC = \sigma_B^2$ . The covariances were determined by the correlation at each level. Correlations of .10 and .30 were chosen, inspired by meta-analytically derived small and large correlations in the social sciences (Gignac & Szodorai, 2016). Covariances were calculated through the variance and the correlation. The combination of all factor levels in our simulation study resulted in a fully-crossed design with 360 conditions. For each condition, 1000 data sets were simulated.

##### 4.1.2. Data Analysis

*Two-Level Intercept-Only Model:* As pointed out earlier, we considered only the two-level intercept-only model or, put differently, a model that estimates the (co)variances of the  $p$  all-units variables at the between-group and within-group levels, and the means of the between-group level, freely. We did so because various structured models (e.g.,  $x_1$  as predictor of  $x_2$  or the other way around) have the same underlying covariance matrix, and we were primarily interested in examining the effects of shrinkage estimation of the covariance matrix on model performance.

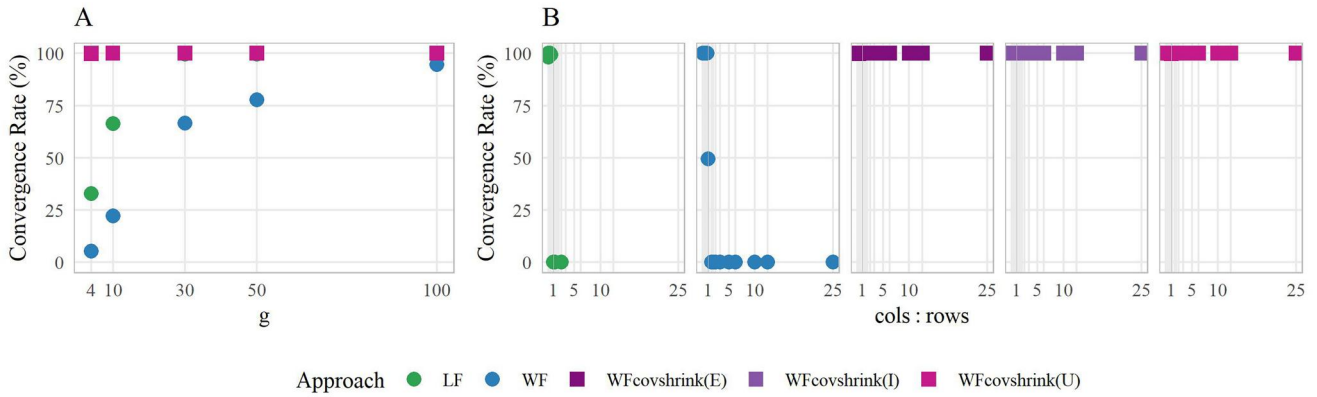
*Approaches:* We compared the performance of the proposed two-stage estimation WFcovshrink to the unregularized WF approach, and the unregularized, standard LF approach. For WFcovshrink, we scrutinized a consistent usage of all target matrices: the equal target matrix in WFcovshrink(E), the identity matrix in WFcovshrink(I), and the unequal target matrix in WFcovshrink(U).

##### 4.1.3. Evaluation Criteria

We conducted comparisons of model performance based on convergence and estimation accuracy. A model was deemed converged if the optimizer indicated that it had found a solution. The convergence rate represents the percentage of converged models out of the total number of estimated models per condition. Estimation accuracy was evaluated in terms of bias and overall accuracy (which incorporates both bias and variance of an estimator). We considered the relative bias,  $\sum(\hat{\theta} - \theta)/\theta \cdot 100\%$ , and the relative root mean squared error (RMSE),  $\sqrt{\sum(\hat{\theta} - \theta)^2/\theta} \cdot 100\%$ .

#### 4.2. Results

Hereinafter, we will delve into the key findings of the simulation study. We will commence by examining convergence, followed by a discussion on estimation accuracy (bias and overall accuracy). Readers interested in further results are referred to the supplementary materials provided in the Appendix. To summarize, we found evidence that the input type (data or sample covariance matrix) and the type of MLE of the sample covariance matrix (biased or unbiased) did not influence the performance of the WF approach substantially (Figure B1), that the WFcovshrink approaches had no severely increased computation times in contrast to the WF approach (Figure B2), and that the WFcovshrink approaches yielded higher percentages of negatively estimated between-group level



**Figure 3.** Convergence rates by sample characteristics.

Note.  $g$  = number of groups; WFcovshrink(E) = equal target matrix; WFcovshrink(I) = identity target matrix; WFcovshrink(U) = unequal target matrix. The *cols:rows* refer to those of long format between-group data matrix LF-B ( $p : g$ ) and the wide format total data matrix WF-T ( $(p \cdot n) : g$ ) for the LF and WF approaches, respectively.

variances and ICC (which we will link later to an increased downward bias).

#### 4.2.1. Convergence

In terms of sample characteristics, the sample size at level-2,  $g$ , proved to be the most influential factor affecting convergence. As illustrated in Panel A of Figure 3, convergence rates aggregated by  $g$  revealed a typical observation: for the LF and WF approaches, convergence rates increased with increasing sample size. In contrast, the WFcovshrink approach consistently converged across all sample sizes. Previous research (Walther et al., 2024) has highlighted the significance of the relationship between the columns and rows of the data matrices in understanding convergence rates, as depicted in Panel B. Replicating earlier findings, we observed that  $cols < rows$  and  $cols \leq rows$  are required for converging models in the LF and WF approaches, respectively. Additionally, the LF approach tended to converge in more diverse settings because satisfying  $p < g$  (in the long format between-group data matrix LF-B) is easier than satisfying  $(p \cdot n) \leq g$  (in the wide format total data matrix WF-T) (Walther et al., 2024). Notably, this restriction did not apply to the WFcovshrink approaches, regardless of *cols:rows* of WF-T. It is interesting to note that convergence rates did neither significantly differ by the number of observed variables ( $p$ ) nor the population characteristics.

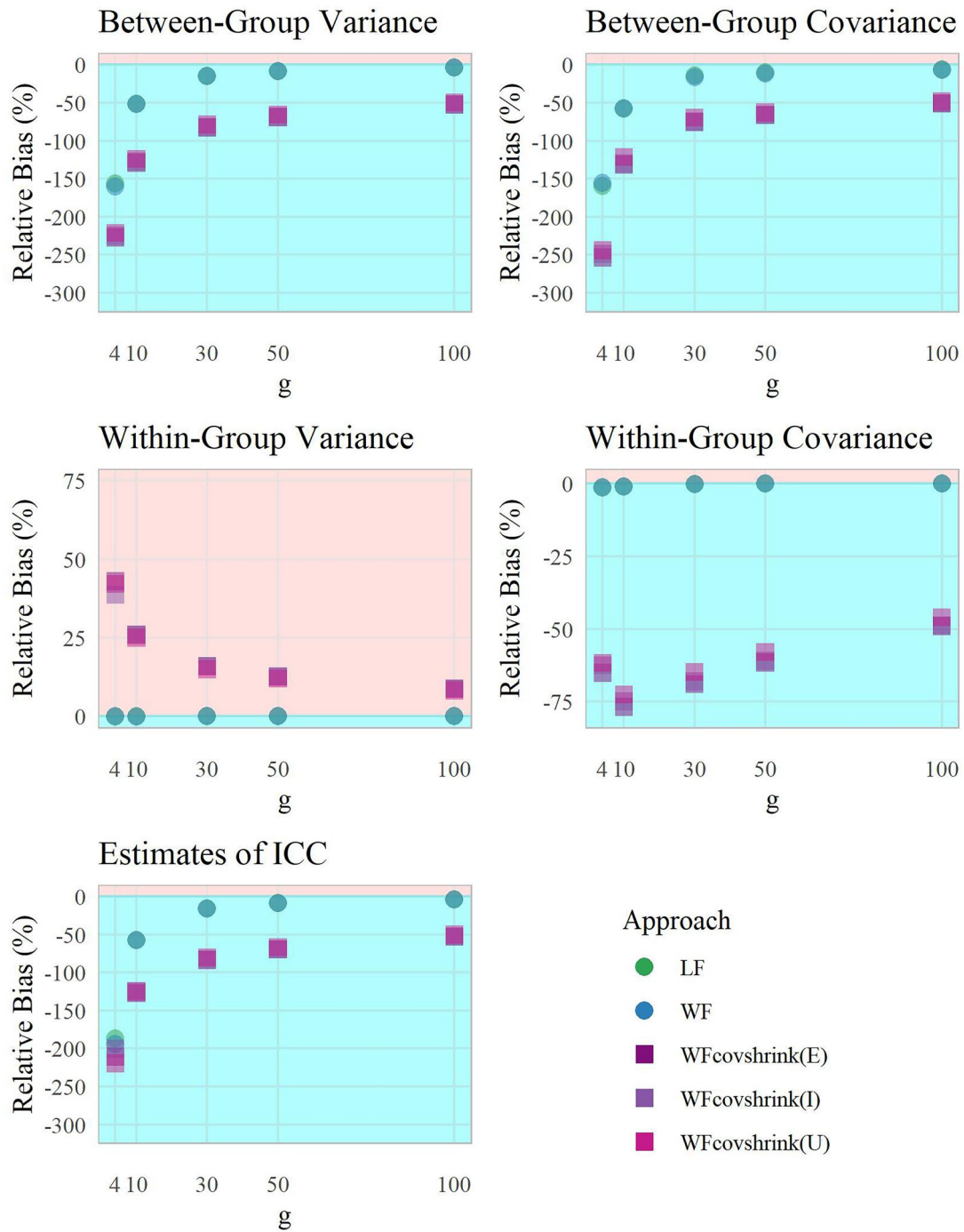
#### 4.2.2. Estimation Accuracy

In the following, we review the estimation accuracy of parameters derived from the LF, WF, and WFcovshrink approaches. Firstly, we examine the relative bias for the model parameters, and secondly, the overall estimation accuracy (relative RMSE) at the between- and within-group levels and the thereof estimated ICCs. Note that for all estimation accuracy parameters, we only considered settings resulting in convergence rates greater than zero across all approaches. Given the WF approach's tendency to exhibit the lowest convergence rates, this implies that we exclusively considered settings where  $p \cdot n \leq g$ . Otherwise, comparing estimation accuracy measures aggregated by different

settings would lead to unfair comparisons, as the WFcovshrink approaches consistently converged, even in practically unrealistic, highly inaccurate settings (e.g.,  $g = 4$ ,  $n = 2$ , and  $p = 10$ ).

*Bias:* We focused on four types of parameters of the random-intercept models, variances and covariances at the between- and within-group level, respectively, and the thereof estimated ICCs. These are depicted in Figure 4. Overall we see, as expected from the known bias-variance tradeoff, that the WFcovshrink approaches had increased biases in contrast to the unregularized approaches. Moreover, when comparing the hypothesized direction of biases in Panel B of Figure 2 with the actual empirical observations, we found a match between our hypotheses and the observed evidence. More specifically, at the between-group level, both variances and covariances exhibited a tendency towards underestimation. Conversely, at the within-group level, the unregularized approaches were unbiased regardless of the sample size (number of groups  $g$ ), while the WFcovshrink approaches introduced an upward bias in variances, and a downward bias in covariances. Following from this, all approaches exhibited a downward bias in the estimates of the ICC, and the WFcovshrink approaches consistently yielded a more conservative underestimation. This downward bias trend in the between-group level variances and the estimates of the ICC was further evidenced by the significant proportion of negatively estimated variances and ICCs in the WFcovshrink approach, as illustrated in Figure B3 in the Appendix.

*Overall Estimation Accuracy:* In the upper panel of Figure 5, the relative RMSE of the between-group parameters aggregated by number of groups  $g$  and group size  $n$  is shown. Overall, smaller numbers of groups  $g$  and group sizes  $n$  resulted in less accurate estimates. However, the WFcovshrink approaches consistently yielded more accurate estimates, with the most significant improvements observed in settings with small  $g$  and especially small  $n$  samples. For example, in a more realistic scenario with a group size of 5 and 50 groups, the unregularized approaches produced relative RMSEs of approximately 200%, while the WFcovshrink approaches reduced it by half. In the middle panel, which

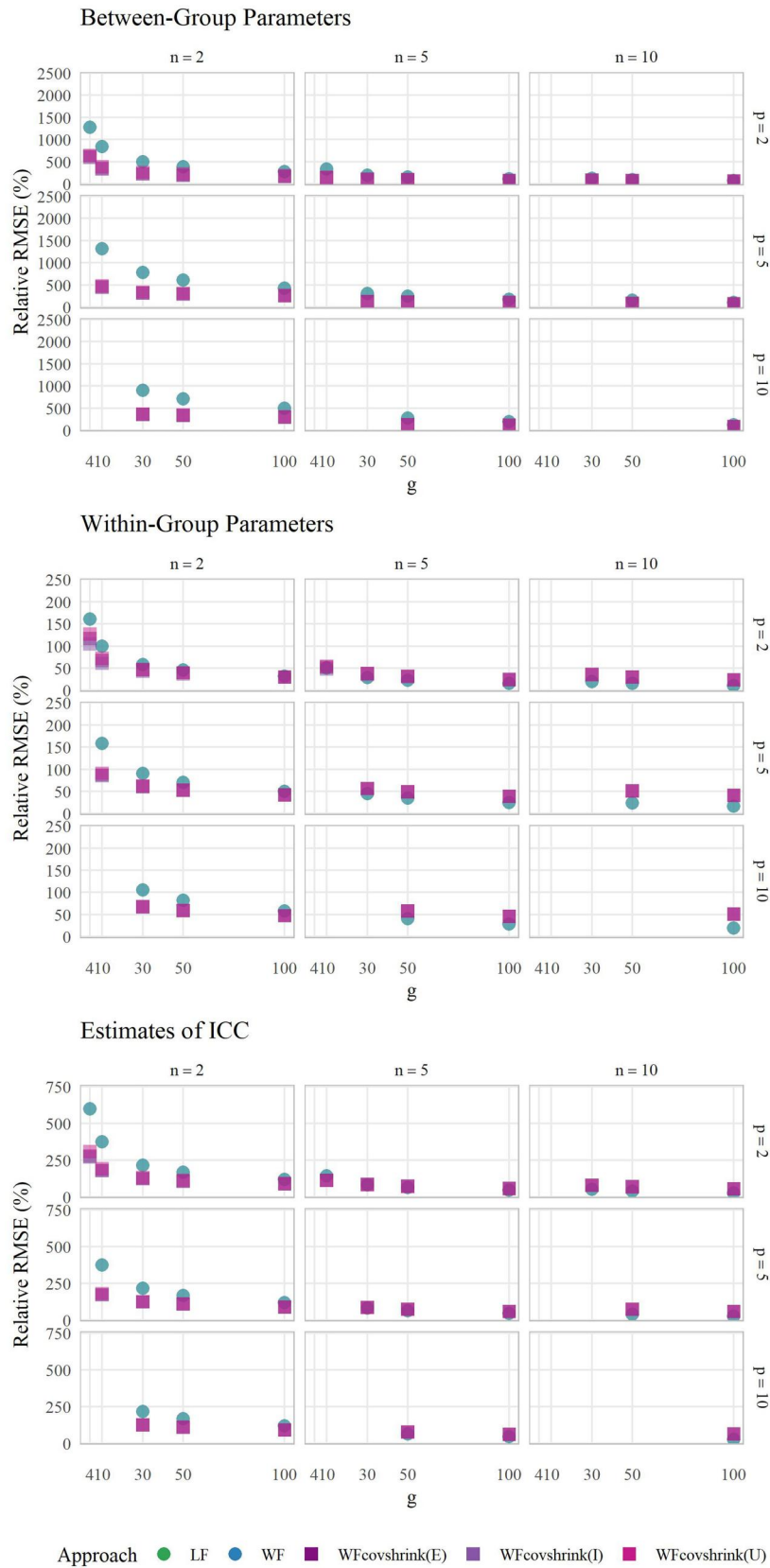


**Figure 4.** Relative bias of parameter estimates.

Note.  $g$  = number of groups; WFcovshrink(E) = equal target matrix; WFcovshrink(I) = identity target matrix; WFcovshrink(U) = unequal target matrix.

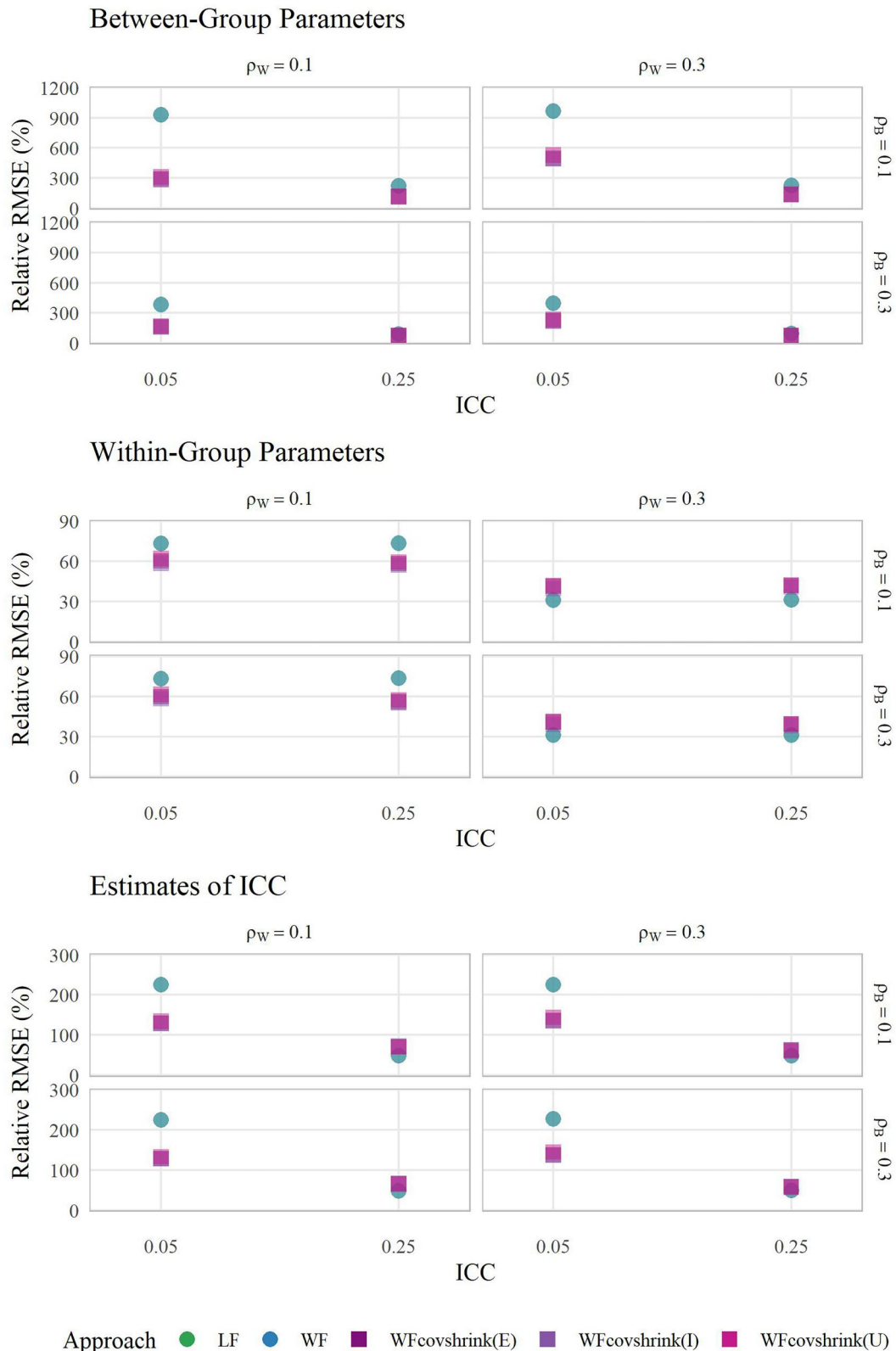
focuses on the within-group level, it can be seen that the WFcovshrink approaches were generally more accurate than the unregularized approaches only when the group size was very small ( $n = 2$ ). However, as the group sizes increased, the estimates from the WFcovshrink approaches tended to be somewhat less accurate. Returning to the setting with a group size of 5 and 50 groups, the unregularized approaches exhibited an average relative RMSE of approximately 30%, whereas the WFcovshrink approaches showed an average relative RMSE of around 45%. The relative RMSE of the ICC estimates are shown in the lower panel. They combine

the results of the between- and within-group level: when the group size was very small ( $n = 2$ ), the WFcovshrink approaches were more accurate, particularly, the smaller the number of groups  $g$  were, but when the group size  $n$  became larger, the unregularized approaches became more accurate. In the setting with a group size of 5 and 50 groups, the estimated ICCs showed an average relative RMSE of approximately 75% in the WFcovshrink approaches, and approximately 65% in the unregularized approaches. In sum, the benefit of the WFcovshrink approaches appealed to the between-group parameters in all



**Figure 5.** Overall estimation accuracy by sample characteristics.

*Note.*  $g$  = number of groups;  $n$  = group size;  $p$  = number of observed variables; WFcovshrink(E) = equal target matrix; WFcovshrink(I) = identity target matrix; WFcovshrink(U) = unequal target matrix.



**Figure 6.** Overall estimation accuracy by population characteristics.

*Note.* ICC = Intraclass Correlation;  $\rho_B$  = correlation at between-group level;  $\rho_W$  = correlation at within-group level; WFcovshrink(E) = equal target matrix; WFcovshrink(I) = identity target matrix; WFcovshrink(U) = unequal target matrix.

settings, including small group sizes ( $n \leq 10$ ) for small to moderate numbers of groups ( $g \leq 100$ ), but to within-group and ICC parameters only in very small group sizes settings ( $n = 2$ ) for small to moderate number of groups ( $g \leq 100$ ).

Figure 6 illustrates the estimation accuracy influenced by the ICC and the correlations of variables in the population. The upper panel, which focuses on the between-group parameters, shows that smaller ICC values (indicating

smaller variances at the between-group level) led to less accurate estimates across all approaches. Notably, the WFcovshrink approaches consistently yielded more accurate estimates compared to the unregularized approaches here. Furthermore, an interaction effect was evident between ICC and correlation at both levels, impacting the overall estimation accuracy of the between-group level parameters. Settings with a low ICC and small correlations at the between-group level resulted in the least accurate estimates across all approaches (approximately 900% in the unregularized approaches and approximately 400% in the WFcovshrink approaches). Notably, in these settings, larger correlations at the within-group level additionally decreased the accuracy of the between-group level parameters. Conversely, settings with a high ICC and large correlations at the between-group level yielded the most accurate estimates across all approaches (approximately 6% in the unregularized approaches, slightly less in the WFcovshrink approaches). Here, the correlations at the within-group level had no substantial influence. Similar to scenarios with small sample sizes, the WFcovshrink approaches proved most effective in addressing the more challenging settings. As indicated in the middle panel, which shows the accuracy of the within-group level parameters, we observed that smaller correlations at the within-group level led to less accurate estimates across all approaches. It appeared that the WFcovshrink approaches resulted in more accurate estimates at the within-group level when these correlations at the within-group level were small but not when they were large. Thus, once again, we observed that the WFcovshrink approach was most effective in handling the more problematic settings. In the lower panel, showing the estimates of the ICCs, we found that the ICC in population had the strongest influence. Small ICCs resulted in the least accurate estimates throughout all approaches, but the WFcovshrink approaches were the most effective here again.

## 5. Discussion

Small sample sizes, such as small group sizes (level-1 units) and small numbers of groups (level-2 units), often pose challenges to multilevel SEM, including difficulties in achieving convergence and inaccuracies in estimating between-group level parameters. To tackle these issues, our research investigated the effectiveness of a two-stage estimation approach, WFcovshrink, which replaces the sample covariance matrix by an estimate of the linear shrinkage estimator introduced by Touloumis (2015). Unlike the traditional unregularized long format (LF) and wide format (WF) approaches, the WFcovshrink methods consistently achieved convergence, regardless of the sample size or the ratio of columns to rows in the data matrix. In terms of accuracy, WFcovshrink outperformed the other approaches in estimating between-group level parameters across all sample sizes tested. Regarding within-group level accuracy, WFcovshrink proved superior only in scenarios with extremely small group sizes ( $n = 2$ ), but even when the number of groups reached up to 100. Our approach also

delivered more accurate ICC estimates by exhibiting a conservative downward bias compared to the typically overestimated ICCs found in unregularized methods in cases with small ICCs (0.05) and very small group sizes ( $n = 2$ ). Given that in psychology and the education sciences, the ICCs commonly encountered are usually at the lower end (Adams et al., 2004; Gulliford et al., 1999), the conservative nature of WFcovshrink's estimates might be preferred. WFcovshrink showed its greatest efficacy in the most challenging conditions: small samples at any level, low ICCs, and minor correlations at the between- or within-group level. The performance of the three target matrices within WFcovshrink was largely similar. In sum, incorporating shrinkage estimation of the sample covariance matrix into a two-stage approach for multilevel SEM significantly mitigated the issues of non-convergence and inaccurate parameter estimates at the between-group level, and for very small group sizes, it effectively shrunk the issue of imprecise within-group level parameter estimates.

However, we must acknowledge that the proposed two-stage approach is only a partial success at this time. While it proves the concept, it remains limited in practical application for two main reasons. Firstly, settings with very small group sizes ( $n = 2$ ) combined with small ICCs are relatively rare. These may appear in pilot studies, but few other research areas consider such settings. Secondly, and closely related to the first point, more customized target matrices need to be considered. The employed target matrices were designed for single-level data, not for single-level representations of multilevel data. Thus, the multilevel nature of the data was not adequately accounted for. Future research calls for more customized target matrices, as without these, the approach is rarely applicable in any realistic setting.

Further points that limit the generalizability of our findings need to be addressed. Firstly, in each simulation scenario, the variances of all observed variables at each level, and consequently the ICCs, were identical. This uniformity might have led to overly optimistic results when using the equal target matrix for shrinkage estimation. Closely related, the total variance of each observed variable was set to 1 in the population. Thus, the results may be limited to situations with variables having unit variances and future research is needed to investigate observed variables with other metrics. However, when practitioners have data with other than unit variance, two pragmatic ways to use our approach may be (1) to first standardize the variances and then use any target matrix or (2) to use the equal or unequal target (but not the identity) target matrices. Secondly, we limited the simulation study to balanced data (i.e., equal group size), but unbalanced data is often the case in practice. How our approach might be used with missing data deserves further attention. Missing values can be imputed ad hoc, for example, with multiple imputation techniques such as MICE (Buuren & Groothuis-Oudshoorn, 2011), and the resulting complete data matrix can then be used for regularization of the covariance matrix. Though, in small sample scenarios, imputation ought to be carefully considered as it might introduce bias (Grund et al., 2018). Another idea

would be to use the pairwise complete data to estimate the regularized covariance matrix. In any case, standard errors are likely to be incorrect because of the varying sample sizes for each (co)variance and we would have to account for this fact. Moreover, these are just ideas that need to be empirically studied. Till then, the application of our approach is limited to balanced data (e.g., experimental data). Thirdly, we only investigated scenarios with small group sizes  $n$  because of larger computational costs of the  $p \cdot n$  “specific-units” variables in the WF approach. The model size and syntax grows with both  $p$  and  $n$  as well. To formulate the constraints in the WF approach more efficiently one could use Kronecker product constraints as suggested by Oort (2001, 2009). In this regard, however, using (proprietary) software with more advanced support for matrix algebra in SEM, such as openMX, is suggested. Fourthly, our investigation focused exclusively on two-level intercept-only models. It remains to be tested how regularized estimators of unstructured covariance matrices perform with more structured models. Arruda and Bentler (2017) also used a regularized estimator of unstructured covariance matrices, but, unlike our approach, applied it to the weight matrix in generalized least squares (GLS) estimation – which is commonly the sample covariance matrix. They found that their approach improved overall model evaluation (test statistics, rejection rates) in small samples compared to standard GLS and MLE for a common CFA in simulation studies, which includes three latent factors, each with five manifest variable indicators and unique error variances. Although these results suggest that the benefits of regularized estimators of unstructured covariance matrices, such as the one we employed by Touloumis (2015), could extend to more structured models, a thorough exploration of this possibility would be a valuable avenue for future research. Lastly, the accuracy of standard errors produced by WFcovshrink and the development of potential corrections warrant further investigation in order to ensure the approach’s broader reliability and applicability.

In conclusion, the application of shrinkage estimation to the covariance matrix within multilevel structural equation modeling (SEM) is a relatively new and evolving field. Our study stands out as one of the pioneering efforts to integrate this shrinkage estimation of covariance matrices in the SEM framework, and, to the best of our knowledge, it is the first to examine this method specifically in the multilevel modeling context. However, before the approach can be applied broadly in practice, more research needs to be done. Still, we believe this method merits consideration by the research community, offering a valuable tool for enhancing the accuracy and convergence of multilevel SEM analyses in small sample size scenarios.

## Disclosure statement

The authors report there are no competing interests to declare.

## ORCID

Julia-Kim Walther  <http://orcid.org/0000-0001-5758-1211>  
 Martin Hecht  <http://orcid.org/0000-0002-5168-4911>  
 Steffen Zitzmann  <http://orcid.org/0000-0002-7595-4736>

## References

- Adams, G., Gulliford, M. C., Ukoumunne, O. C., Eldridge, S., Chinn, S., & Campbell, M. J. (2004). Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology*, *57*, 785–794. <https://doi.org/10.1016/j.jclinepi.2003.12.013>
- Allison, P. D. (2009). *Fixed Effects Regression Models*. SAGE publications.
- Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, *49*, 155–173. <https://doi.org/10/cwnzr3>
- Arruda, E. H., & Bentler, P. M. (2017). A regularized GLS for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 657–665. <https://doi.org/10/gcmfhs>
- Barendse, M., & Rosseel, Y. (2020). Multilevel modeling in the ‘wide format’ approach with discrete data: A solution for small cluster sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 696–721. <https://doi.org/10.1080/10705511.2019.1689366>
- Bickel, P. J., Li, B., Tsybakov, A. B., Van De Geer, S. A., Yu, B., Valdés, T., Rivero, C., Fan, J., & Van Der Vaart, A. (2006). Regularization in statistics. *Test*, *15*, 271–344. <https://doi.org/10.1007/BF02607055>
- Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, *24*, 2350–2383. <https://doi.org/10.1214/aos/1032181158>
- Buuren, S. v., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67. <https://doi.org/10.18637/jss.v045.i03>
- Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research*, *29*, 468–508. <https://doi.org/10/cs43xwZSCC:0000626>
- De Jonckere, J., & Rosseel, Y. (2023). A model-based shrinkage target to avoid non-convergence in small sample SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*, 941–955. <https://doi.org/10.1080/10705511.2023.2171420>
- Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 327–351. <https://doi.org/10.1080/10705511.2014.937849>
- Dragulescu, A., & Arendt, C. (2020). November 10). *Xlsx: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files* (Version 0.6.5). Retrieved March 3, 2023, from <https://CRAN.R-project.org/package=xlsx>
- Duncan, T. E., Duncan, S. C., Alpert, A., Hops, H., Stoolmiller, M., & Muthen, B. (1997). Latent variable modeling of longitudinal and multilevel substance use data. *Multivariate Behavioral Research*, *32*, 275–318. [https://doi.org/10.1207/s15327906mbr3203\\_3](https://doi.org/10.1207/s15327906mbr3203_3)
- Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. <https://doi.org/10.1016/j.paid.2016.06.069>
- Gorsuch, R. L. (1983). *Factor analysis*. Lawrence Earlbaum Associates.
- Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: simulations and recommendations. *Organizational Research Methods*, *21*, 111–149. <https://doi.org/10.1177/1094428117703686>
- Gulliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: Data from the health survey for England 1994. *American Journal of Epidemiology*, *149*, 876–883. <https://doi.org/10/gn2gxn>
- Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. M. (2003). ARMA-based SEM when the number of time points  $T$  exceeds the number of cases  $N$ : Raw data maximum likelihood. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 352–379. ZSCC: 0000049. <https://doi.org/10/dkmbnp>
- Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small

- samples. *Structural Equation Modeling: A Multidisciplinary Journal*, 8, 157–174. [https://doi.org/10.1207/S15328007SEM0802\\_1](https://doi.org/10.1207/S15328007SEM0802_1)
- Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, 64, 157–170. <https://doi.org/10.1111/j.1467-9574.2009.00445.x>
- Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.
- Huang, P.-H., Chen, H., & Weng, L.-J. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika*, 82, 329–354. <https://doi.org/10/gbm65j>
- Huang, Y., & Bentler, P. M. (2015). Behavior of asymptotically distribution free test statistics in covariance versus correlation structure analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 22, 489–503. <https://doi.org/10/gcz6zpj>
- Hugh-Jones, D. (2022). *Huxtable: Easily Create and Style Tables for LaTeX, HTML and Other Formats* (Version 5.5.6). Retrieved March 3, 2023, from <https://CRAN.R-project.org/package=huxtable>
- Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 23, 555–566. <https://doi.org/10/gcmfh8>
- Jung, S., & Takane, Y. (2007). Regularized common factor analysis. *New Trends in Psychometrics*, 141–149. Retrieved February 14, 2022, from <https://www.semanticscholar.org/paper/Regularized-Common-Factor-Analysis-Jung-Takane/2df36e88cacc510b1f900960b6e784df886f0fb>
- Kamada, A., & Kano, Y. (2012, July 1–4). *Statistical inference in structural equation modeling with a near singular covariance matrix* [Paper presentation]. 2nd Institute of Mathematical Statistics Asia Pacific Rim Meeting, Tsukuba, Japan.
- Kamada, A., Yanagihara, H., Wakaki, H., & Fukui, K. (2014). Selecting a shrinkage parameter in structural equation modeling with a near singular covariance matrix by the GIC minimization method. *Hiroshima Mathematical Journal*, 44, 315–326. <https://doi.org/10/gpgn8b>
- Kelley, C. T. (1995). *Iterative methods for linear and nonlinear equations*. Society for Industrial and Applied Mathematics. Retrieved June 9, 2022, from <https://epubs.siam.org/doi/book/10.1137/1.9781611970944>
- Lange, K., Chambers, J., & Eddy, W. (1999). *Numerical analysis for statisticians* (Vol. 2). Springer.
- Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88, 365–411. [https://doi.org/10.1016/S0047-259X\(03\)00096-4](https://doi.org/10.1016/S0047-259X(03)00096-4)
- Ledoit, O., & Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40. <https://doi.org/10.1214/12-AOS989>
- Ledoit, O., & Wolf, M. (2020). The power of (non-)linear shrinking: A review and guide to covariance matrix estimation. <https://doi.org/10.5167/UZH-170642>
- Liang, X., & Jacobucci, R. (2020). Regularized structural equation modeling to detect measurement bias: Evaluation of lasso, adaptive lasso, and elastic net. *Structural Equation Modeling: A Multidisciplinary Journal*, 27, 722–734. ZSCC: 0000023. <https://doi.org/10.1080/10705511.2019.1693273>
- Lorenzo-Seva, U., & Ferrando, P. J. (2021). Not positive definite correlation matrices in exploratory item factor analysis: Causes, consequences and a proposed solution. *Structural Equation Modeling: A Multidisciplinary Journal*, 28, 138–147. <https://doi.org/10.1080/10705511.2020.1735393>
- Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods*, 16, 444–467. <https://doi.org/10.1037/a0024376>
- Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, 13, 203–229. <https://doi.org/10/c8qhd2>
- McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, 28, 295–314. <https://doi.org/10.1007/s10648-014-9287-x>
- Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, 10, 259–284. <https://doi.org/10.1037/1082-989X.10.3.259>
- Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods*, 3, 45–58.
- Muthén, B. O. (1990). *Mean and covariance structure analysis of hierarchical data*. Department of Statistics, UCLA.
- Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, 22, 376–398. <https://doi.org/10.1177/0049124194022003006>
- Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, 25, 267. <https://doi.org/10.2307/271070>
- Oort, F. J. (2001). Three-mode models for multivariate longitudinal data. *The British Journal of Mathematical and Statistical Psychology*, 54, 49–78. <https://doi.org/10.1348/000711001159429>
- Oort, F. J. (2009). Three-mode models for multitrait-multimethod data. *Methodology*, 5, 78–87. <https://doi.org/10.1027/1614-2241.5.3.78>
- Orzek, J. H., & Voelkle, M. C. (2023). Regularized continuous time structural equation models: A network perspective. *Psychological Methods*, 28, 1286–1320. <https://doi.org/10.1037/met0000550>
- Pedersen, T. L. (2022). *Patchwork: The composer of plots* (Version 1.1.2). Retrieved March 3, 2023, from <https://CRAN.R-project.org/package=patchwork>
- R Core Team. (2023). *R: A language and environment for statistical computing*. <https://www.R-project.org/>
- Rosseel, Y., Jorgensen, T. D., Rockwood, N., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Scharf, F., & Du, H. (2023). *Lavaan: Latent Variable Analysis* (Version 0.6-15). Retrieved March 3, 2023, from <https://CRAN.R-project.org/package=lavaan>
- Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics*, 35, 26–53. <https://doi.org/10/c63vdm>
- Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., Arachchige, C., Arppe, A., Baddeley, A., Barton, K., Bolker, B., Borchers, H. W., Caiiro, F., Champely, S., Chessel, D., Chhay, L., Cooper, N., Cummins, C., Dewey, M., Doran, H. C., & Zeileis, A. (2024). *DescTools: Tools for Descriptive Statistics* (Version 0.99.50). Retrieved March 21, 2024, from <https://cran.r-project.org/web/packages/DescTools/index.html>
- Singer, H. (2010). SEM modeling with singular moment matrices Part I: ML-estimation of time series. *The Journal of Mathematical Sociology*, 34, 301–320. <https://doi.org/10.1080/0022250X.2010.509524>
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2. ed). SAGE.
- Stegmüller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and bayesian approaches. *American Journal of Political Science*, 57, 748–761. <https://doi.org/10.1111/ajps.12001>
- Stein, C. (1956). *Some problems in multivariate analysis, Part I*. Stanford Univ CA.
- Tikhonov, A. N. (1943). On the stability of inverse problems. *Doklady Akademii Nauk SSSR*, 39, 195–198.
- Touloumis, A. (2015). Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings. *Computational Statistics & Data Analysis*, 83, 251–261. <https://doi.org/10.1016/j.csda.2014.10.018>
- Touloumis, A. (2019). *ShrinkCovMat: Shrinkage Covariance Matrix Estimators* (Version 1.4.0). Retrieved March 21, 2024, from <https://cran.r-project.org/web/packages/ShrinkCovMat/index.html>
- Trendafilov, N. T., & Unkel, S. (2011). Exploratory factor analysis of data matrices with more variables than observations. *Journal of Computational and Graphical Statistics*, 20, 874–891. <https://doi.org/10.1198/jcgs.2011.09211>
- Ullrich, E., Lüdtke, O., & Robitzsch, A. (2023). Alleviating estimation problems in small sample structural equation modeling—A comparison of constrained maximum likelihood, Bayesian estimation, and fixed reliability approaches. *Psychological Methods*, 28, 527–557. <https://doi.org/10.1037/met0000435.supp>



- Unkel, S., & Trendafilov, N. T. (2010). Simultaneous parameter estimation in exploratory factor analysis: An expository review. *International Statistical Review*, 78, 363–382. <https://doi.org/10.1111/j.1751-5823.2010.00120.x>
- Van Montfort, K., Oud, J. H., & Voelkle, M. C. (2018). *Continuous time modeling in the behavioral and related sciences*. Springer.
- Voelkle, M. C., Oud, J. H. L., von Oertzen, T., & Lindenberger, U. (2012). Maximum likelihood dynamic factor modeling for arbitrary  $N$  and  $T$  using SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, 19, 329–350. <https://doi.org/10.1080/10705511.2012.687656>
- Walther, J.-K., Hecht, M., Nagengast, B., & Zitzmann, S. (2024). To be long or to be wide: How data format influences convergence and estimation accuracy in multilevel structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 0, 1–16. <https://doi.org/10.1080/10705511.2024.2320050>
- Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D, RStudio. (2023). *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* (Version 3.4.2). Retrieved March 3, 2023, from <https://CRAN.R-project.org/package=ggplot2>
- Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D., & Posit, P. B. C. (2023). *Dplyr: A Grammar of Data Manipulation* (Version 1.1.2). Retrieved March 3, 2023, from <https://CRAN.R-project.org/package=dplyr>
- Wickham, H., Girlich, M., & RStudio. (2022). *Tidyr: Tidy Messy Data* (Version 1.3.0). Retrieved June 14, 2022, from <https://CRAN.R-project.org/package=tidyr>
- Wilke, C. O. (2020). *Cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'* (Version 1.1.1). Retrieved May 19, 2023, from <https://cran.r-project.org/web/packages/cowplot/index.html>
- Williams, D. R., & Rodriguez, J. E. (2022). Why overfitting is not (usually) a problem in partial correlation networks. *Psychological Methods*, 27, 822–840. <https://doi.org/10/gqb2r6>
- Wothke, W. (1993). Nonpositive Definite Matrices in Structural Modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). Sage Publications.
- Yuan, K.-H., & Bentler, P. M. (2017). Improving the convergence rate and speed of Fisher-scoring algorithm: Ridge and anti-ridge methods in structural equation modeling. *Annals of the Institute of Statistical Mathematics*, 69, 571–597. <https://doi.org/10/gpgn83>
- Yuan, K.-H., & Chan, W. (2008). Structural equation modeling with near singular covariance matrices. *Computational Statistics & Data Analysis*, 52, 4842–4858. <https://doi.org/10.1016/j.csda.2008.03.030>
- Yuan, K.-H., Wu, R., & Bentler, P. M. (2011). Ridge structural equation modelling with correlation matrices for ordinal and continuous data: Ridge SEM with correlation matrices. *The British Journal of Mathematical and Statistical Psychology*, 64, 107–133. <https://doi.org/10/cwd74t>
- Zitzmann, S. (2018). A computationally more efficient and more accurate stepwise approach for correcting for sampling error and measurement error. *Multivariate Behavioral Research*, 53, 612–632. <https://doi.org/10/gpgn86>
- Zitzmann, S., Lüdtke, O., & Robitzsch, A. (2015). A Bayesian approach to more stable estimates of group-level effects in contextual studies. *Multivariate Behavioral Research*, 50, 688–705. <https://doi.org/10/gg5fg2>
- Zitzmann, S., Lüdtke, O., Robitzsch, A., & Hecht, M. (2021). On the performance of bayesian approaches in small samples: A comment on Smid, McNeish, Miocevic, and van de Schoot (2020). *Structural Equation Modeling: A Multidisciplinary Journal*, 28, 40–50. ZSCC: 0000027. <https://doi.org/10.1080/10705511.2020.1752216>
- Zitzmann, S., Lüdtke, O., Robitzsch, A., & Marsh, H. W. (2016). A Bayesian approach for estimating multilevel latent contextual models. *Structural Equation Modeling: A Multidisciplinary Journal*, 23, 661–679. <https://doi.org/10.1080/10705511.2016.1207179>
- Zitzmann, S., Wagner, W., Hecht, M., Helm, C., Fischer, C., Bardach, L., & Göllner, R. (2021). How many classes and students should ideally be sampled when assessing the role of classroom climate via student ratings on a limited budget? An optimal design perspective. *Educational Psychology Review*, 34, 511–536. ZSCC: NoCitationData[s0]. <https://doi.org/10.1007/s10648-021-09635-4>
- Zitzmann, S., Walther, J.-K., Hecht, M., & Nagengast, B. (2022). What is the maximum likelihood estimate when the initial solution to the optimization problem is inadmissible? The case of negatively estimated variances. *Psych*, 4, 343–356. <https://doi.org/10.3390/psych4030029>

## Appendices

### Appendix A

#### R Code

```
### Example Code
# - for data, sample covariance matrix, and two-level random-intercept model in Figure 1 (with g=4)
# - for shrinkage estimate of the sample covariance matrix, and two-level random-intercept model in Figure 2,
  and estimates in Table 1 (with g=50)

# Note that the code only works for p=2 and n=2.
# If you want to examine other settings, check out the
# the code for the simulation study on Github:
# https://github.com/demianJK/WFcovshrink

## (0) preparation #####

# load required packages
library(lavaan) # for model estimation
library(tidyr) # for reformatting LF to WF with pivot_wider()
library(ShrinkCovMat) # for shrinkage estimation
# (code runs in ShrinkCovMat 1.4.0 which is the latest on CRAN, Feb 21st 2024)

# set random number seed to obtain example data
set.seed(4395)
```

```

## (1) Population Characteristics #####

# We use the lavaan syntax to set the population models.
popModel_B <- "x1~~0.05*x1; x2~~0.05*x2; x1~~0.015*x2" # between level
popModel_W <- "x1~~0.95*x1; x2~~0.95*x2; x1~~0.285*x2" # within level
# means are zero by default

# We have two variables x1 and x2.
p <- 2
# The variances for both variables are the same at each level.
# The variance at the between level is 0.05.
# The variance at the within level is 0.95.
# Thus, the ICC=0.25.
# The correlation of the two variables is the same at both levels (.3).
# The covariances differ.
# Transform the correlation formula to get the covariances.
# corr_x1x2=cov_x1x2 / (sd_x1 * var_x2)
# |* (sd_x1 * sd_x2) and sd_x1=sd_x2 thus |* var_x1
# corr_x1x2 * var_x1=cov_x1x2

## (2) Sample Characteristics #####

g <- 50 # number of groups (you may change this)
n <- 2 # group size (balanced data)
N <- g * n # total sample size

# the data sampling is done in long format (LF)
sample_B <- simulateData(popModel_B, sample.nobs=g,
  model.type = "lavaan") # between level
sample_W <- simulateData(popModel_W, sample.nobs=N, # within level
  model.type = "lavaan")
groups <- rep(1:g, each=n) # group numbers ("j" in Figure 1)
LF_T <- sample_W # create data frame with the same dimensions
LF_T[,] <- 0 #. and clear all entries
  for (j in unique(groups)) { # merge the sampled data from both levels
    for (i in min(which(groups == j)):max(which(groups == j)))
      LF_T[i,] <- sample_W[i,] + sample_B[j,]
  }
LF_T$persons <- rep(1:n, g) # unit numbers ("i" in Figure 1)
LF_T$groups <- as.factor(groups)
LF_T <- cbind(LF_T[, (p+1):(p+2)], LF_T[, 1:p]) # rearrange columns
# LF-T is the total data matrix in long format (LF)..
round(LF_T[, 3:(3+p-1)], 0) # note that we round for Figure 1 and 2
#.. and the total covariance matrix is estimated by the unbiased estimator (see Muthén, 1994)
Sigma_LF_T <- cov(LF_T[, 3:4])
round(Sigma_LF_T, 2)

# Now we reformat to wide format (WF).
WF_T <- pivot_wider(LF_T, names_from = "persons", values_from=3:4, names_sep = ".")
round(WF_T[, 2:(2+(p*n)-1)], 0)
varnames <- colnames(WF_T[, 2:(2+(p*n)-1)])

# shrinkage estimate S*_E with equal target Matrix vI_p
# note that unbiased S is employed

WF_T_trans <- t(WF_T[, -1]) # transpose because ShrinkCovMat(data, .) expects
# that rows correspond to variables and columns to observations

# estimate S*_E (note that the approach uses unbiased S_WF-T)
Wfcovshrink_E <- ShrinkCovMat::shrinkcovmat.equal(data=WF_T_trans,
  centered=FALSE)

round(Wfcovshrink_E$Sigmasample, 2) # unbiased S, round for Figures 1 and 2
round(Wfcovshrink_E$STarget, 2) # vI_p, round for Figure 2
round(Wfcovshrink_E$Sigmahat, 2) # S*_E, round for Figure 2
round(Wfcovshrink_E$lambdahat, 2) # lambda_E, round for Figure 2
# names of covariance matrix required for lavaan
colnames(Wfcovshrink_E$Sigmahat) <- varnames
rownames(Wfcovshrink_E$Sigmahat) <- varnames

```

```
## (3) estimate models #####
model_WF <- paste0(# Level: 1 (unique factors)
  "x1.1~~Vx1_w*x1.1; x1.2~~Vx1_w*x1.2; x2.1~~Vx2_w*x2.1;
  x2.2~~Vx2_w*x2.2; x1.1~~Cx12_w*x2.1; x1.2~~Cx12_w*x2.2;",
  # these are the desired within variances and covariances
  # Vx1_w, Vx2_w, and Vx12_w are equality constraints
  # Level: 2 (common factors)
  "x1.1~0*1; x1.2~0*1; ; x2.1~0*1; x2.2~0*1;",
  # if level-2 variables are aggregates of level-1 variables,
  # intercepts at level-1 have to be fixed to 0
  "fx1=~1*x1.1+1*x1.2; fx2=~1*x2.1+1*x2.2;",
  # measurement model with factor loadings set to 1
  "fx1~~fx1; fx2~~fx2; fx1~~fx2;",
  # these are the desired between variances and covariances
  "fx1~1; fx2~1") # between means
fit_WF <- sem(model=model_WF,
  data=WF_T)
summary(fit_WF)

fit_WFcovshrink_E <- sem(model_WF,
  sample.cov=WFcovshrink_E$SigmaHat,
  sample.cov.rescale=FALSE,
  # rescale sample.cov with (g-1/g)?
  sample.nobs=g,
  sample.mean=colMeans(WF_T[, -1]))
summary(fit_WFcovshrink_E)

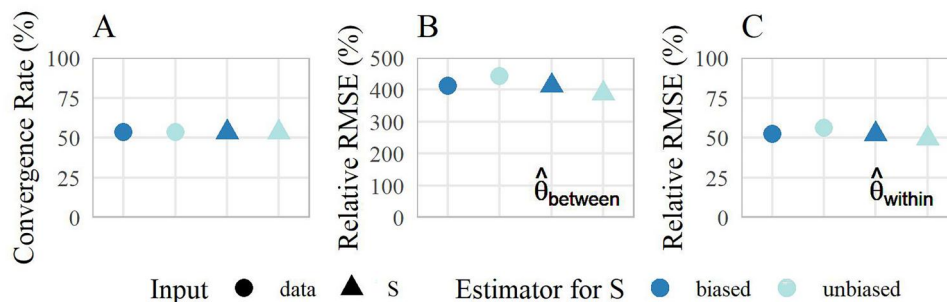
## (4) estimate ICCs #####
# ICC in population: 0.05 (see "Population Characteristics")
# the ICCs are estimated by the parameters of the model-implied matrices

## WF
fit_WF@Fit@x[7]/(fit_WF@Fit@x[7]+fit_WF@Fit@x[1]) # x1
fit_WF@Fit@x[8]/(fit_WF@Fit@x[8]+fit_WF@Fit@x[3]) # x2

## WFcovshrink(E)
fit_WFcovshrink_E@Fit@x[7]/(fit_WFcovshrink_E@Fit@x[7]+fit_WFcovshrink_E@Fit@x[1]) # x1
fit_WFcovshrink_E@Fit@x[8]/(fit_WFcovshrink_E@Fit@x[8]+fit_WFcovshrink_E@Fit@x[3]) # x2
```

## Appendix B

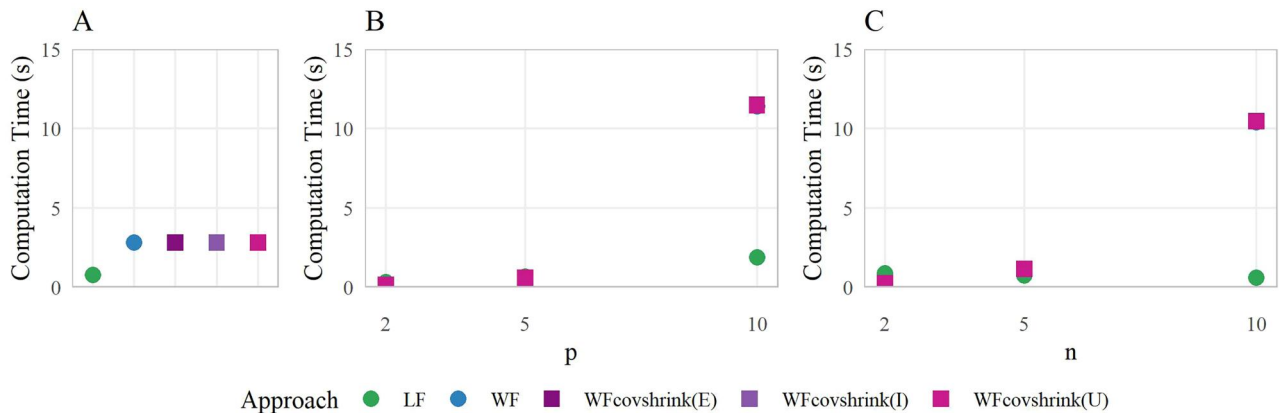
### Supplemental Analysis



**Figure B1.** The WF approach and its different input and sample covariance matrix estimator possibilities.

Note. Data = input was data matrix; S = input was sample covariance matrix; biased = normal theory derived ML of sample covariance matrix used (default); unbiased sample covariance matrix was used ("sample.cov.rescale = TRUE").

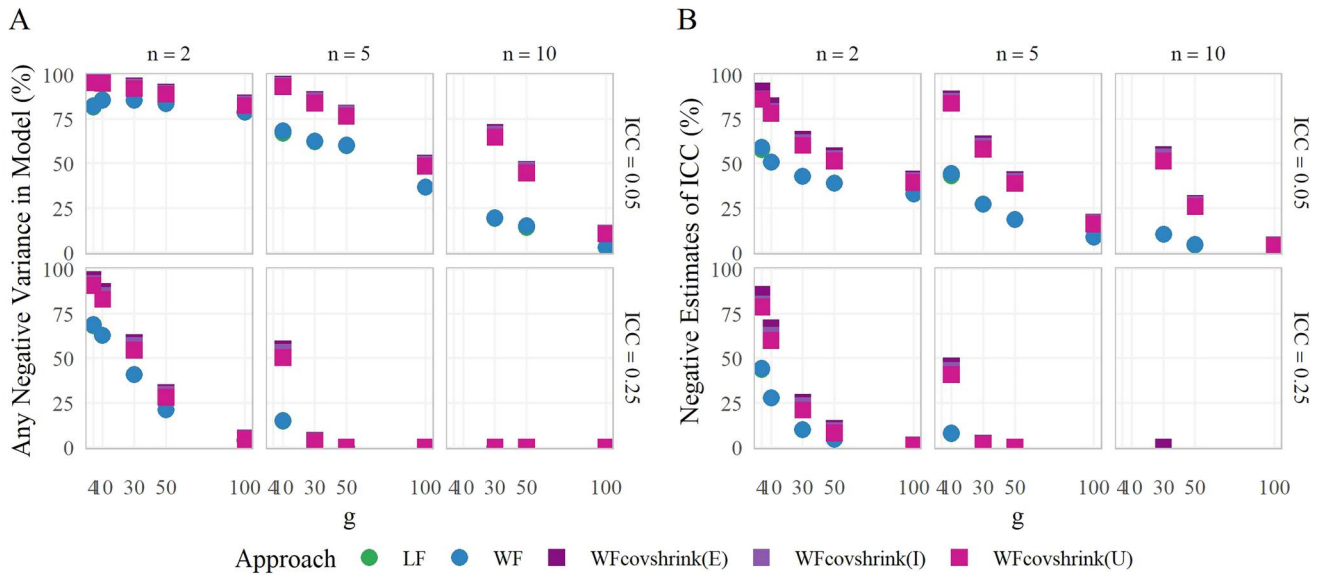
We wanted to control for two differences in the WF and WFcovshrink approaches to check whether performance gains are solely attributable to the proposed two-stage approach. Firstly, we wanted to make sure that convergence gains in the WFcovshrink approach are not due to the input type, or more specifically, supplying a sample covariance matrix instead of a data matrix. The data matrix has to satisfy that the number of columns (number of observed variables) is smaller than the number of rows (number of observations), which is rooted in the implementation of traditional MLE in *lavaan* (see 1.1. The Wide Format (WF) Approach), and we wanted to check whether *lavaan* has similar constraints when supplying a sample covariance matrix (without sweeping the whole source code). Secondly, we wanted to ensure that accuracy gains in the WFcovshrink approach are not due to using different MLE of the sample covariance matrix; in particular, the unbiased one in WFcovshrink in contrast to the biased one that is the default in the WF approach. Figure B1 depicts the results of the conjugate analysis. With regard to convergence rate (Panel A), there were no differences. For overall estimation accuracy of between-group level (Panel B) and within-group level parameter estimates (Panel C), we found marginal differences. There seemed to have been some kind of interaction between the input and estimator. More specifically, the most accurate estimations were derived by supplying the sample covariance estimated by unbiased MLE. In contrast, supplying data and using the unbiased MLE as well yielded the least accurate estimations. However, overall these differences might be negligible. Thus, convergence and estimation accuracy gains can be mostly attributed to replacing the sample covariance matrix by a shrinkage estimate in the WFcovshrink approaches.



**Figure B2.** Computation time by sample characteristics.

*Note.*  $n$  = group size;  $p$  = number of observed variables; WFcovshrink(E) = equal target matrix; WFcovshrink(I) = identity target matrix; WFcovshrink(U) = unequal target matrix.

The computation time of a model is the time the optimizer needed to find a solution. In the WFcovshrink approaches, the time for the shrinkage estimation, which was marginally small for all three target matrices, was added. Figure B2 shows computation times for different aggregation levels. Panel A depicts the overall average computation times, which were smallest in the LF approach ( $\approx 1$ s), but not substantially different in the WF and WFcovshrink approaches ( $\approx 3$ s). In greater detail, Panel B and C depict that the computation time in all WF approaches magnified fairly by the number of observed variables  $p$  and the group size  $n$ . Recall that both quantities determined the number of model parameters that are freely estimated ( $p$ ) and equality constrained ( $n$ ). Again, there was no substantial difference in the WFcovshrink approaches compared to the unregularized WF approach. In the LF approach, the number of freely estimated in the LF approach was only determined by the number of observed variables  $p$ . Thus, the computation time of the LF approach was not influenced by the group size  $n$ , and its computation times were on average smaller. This could be explained by smaller dimensions of the covariance matrix in LF ( $p \cdot p$ ) compared to the WF approaches ( $(p \cdot n) \cdot (p \cdot n)$ ). Note that the population characteristics did not result in substantial differences in computation times. To put this finding into practical context: the larger computation times of any WF approach might be of little consequence if we only estimate a small number of models.



**Figure B3.** Negatively estimated variances at the between-group level and ICC.

Note.  $g$  = number of groups;  $n$  = group size; ICC = Intraclass Correlation; WFcovshrink(E) = equal target matrix; WFcovshrink(I) = identity target matrix; WFcovshrink(U) = unequal target matrix. In Panel A, percentages of negative variance in *any model* are shown. Note that negative variances were only present at the between-group level. In Panel B, percentages of negative estimates of the ICC of *any observed variable* are depicted. Thus, percentages depicted in Panel A are larger than those in Panel B.

Two types of inadmissibly negative estimates are depicted in Figure B3: between-group level variances and ICCs (i.e., quotient of between-group and total variances). Note that at the within-group level, no negative variances were encountered. The percentages of *at least one* negative variances at the between-group level in a model are shown in Panel A. Across all approaches, the percentage was larger when the number of groups  $g$ , the group size  $n$ , or the ICC was smaller. Overall, the WFcovshrink approaches yielded higher percentages of models with negative variances at the between-group level than the unregularized approaches. In Panel B, percentages of negatively estimated ICC for *every* observed variable in a model are shown. A similar picture emerged: percentages soared when the number of groups  $g$ , the group size  $n$ , or the ICC was smaller, and the percentages of the WFcovshrink approaches were higher. The increase in these negative estimates in the WFcovshrink approaches is probably related to the amplification of downward bias of between-group level estimates (see Figure 4).