This postprint was originally published by Taylor & Francis as:
Friemelt, B., Bloszies, C., Ernst, M. S., Peikert, A., Brandmaier, A. M., & Koch, T. (2023). On the performance of different regularization methods in bifactor-(S-1) models with explanatory variables: Caveats, recommendations, and future directions. *Structural Equation Modeling, 30*(4), 560–573. https://doi.org/10.1080/10705511.2022.2140664

On the Performance of Different Regularization Methods in Bifactor-(S-1) Models with Explanatory Variables—Caveats, Recommendations, and Future Directions

Benedikt Friemelt[a], Christian Bloszies[a], Maximilian S. Ernst[b,c], Aaron Peikert[b,c], Andreas M. Brandmaier[b,d,e] & Tobias Koch[a]

[a]Friedrich-Schiller Universität Jena; [b]Max-Planck-Institute for Human Development; [c]Humboldt-Universität zu Berlin; [d]Max-Planck UCL Centre for Computational Psychiatry and Ageing Research; [e]MSB Medical School Berlin Germany.

CONTACT Mr. Benedikt Friemelt; benedict.friemelt@uni-jena.de; Friedrich-Schiller Universität Jena, Psychology, Jena, 07743 Germany

ABSTRACT

Regularization methods in linear regression models with manifest variables have been shown to be effective in selecting key predictors from a set of many variables, while improving predictions for novel observations. Regularization methods are particularly attractive for the analysis of complex multidimensional data when theory development is the primary goal; for example when researchers attempt to predict general or specific factors in bifactor models using many potentially relevant predictors. However, applications of regularization methods in such models are still scarce. In a simulation study, we examined the performance of different regularization methods in bifactor-(S-1) models, varying the number of predictors, the correlations with the outcome (effect size), the underlying structure of multicollinearity as well as the sample size, the type of penalty, and a single-step versus a two-step approach. We explore potential caveats in the use of regularization methods in bifactor-(S-1) models, provide practical recommendations, and discuss future directions.

KEYWORDS

Bifactor-(S-1) model; CTC(M-1) model; regularization methods; residual approach

Regularization is a well-known technique in predictive modeling (see Hastie et al., 2009). It is characterized by adding some sort of penalty term to the loss functions of typical estimation methods (e.g., ordinary least square or maximum likelihood estimation). In linear regression analysis, for example, this leads to a bias in an otherwise asymptotically unbiased estimator. However, through a substantial reduction in variance of the estimated parameters, a lower meansquared-error (MSE) is achieved (Hastie et al., 2009).

$$MSE\left(\hat{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = Var(\hat{\theta}) + Bias^2\left(\hat{\theta}\right) \qquad (1)$$

In Equation 1, $\hat{\theta}$ denotes the vector of all estimated model parameters, Y_i the values of the dependent variable, and $\hat{Y}_i$ the respective predicted values. Note that it is generally impossible to reduce one part of the MSE without increasing the other, a concept known as the bias-variance-tradeoff (Hastie et al., 2009). In practice, researchers must decide how much bias is acceptable to ensure a more stable estimation process, more robust predictions, and increased efficiency.

This question leads us to the distinction between explanatory and predictive modeling. As both Shmueli (2010) and Sainani (2014) note, the goal of explanatory modeling is mainly to test causal relationships within the data, based on an already developed theory. Estimating parameters with as little bias as possible is thus crucial to ensure the validity of findings and interpretations. These results are then often used to design interventions or treatments for modifiable risk factors, which rely on the assumed causal relationship between the model variables. Bullain and Doody (2020) give an overview of potential risk factors for the development of Alzheimer's disease and possible interventions derived by explanatory modeling approaches. A notable caveat in the use of theoretically unbiased estimators, however, lies in the conditions needed to fulfill this property. This means that there must not be any omitted variables, all relationships between variables must be linear (in the context of linear regression or other linear modeling approaches), and all distributional assumptions need to hold. Arguably, these assumptions are hardly ever met in practice, which limits the promises of this kind of estimators.

In contrast, predictive modeling is not concerned about explaining relationships in the observed data, but rather tries to make the best possible predictions for new observations without conforming to a specified underlying theory. Biased estimates of parameters are therefore much less of a problem since the resulting function is first and foremost a tool and not the entity of interest (Shmueli, 2010). Instead, prediction errors, like the MSE, for new observations are minimized by choosing a favorable balance between bias and variance in estimations. While psychological research is often interested in substantive interpretations, certain fields can still profit from a predictive approach. Diagnosing mental illnesses serves as a good example for when the accuracy of a prediction should be as high as possible to avoid potential harm for the individual case. But apart from situations

in which prediction is the primary goal, predictive modeling can complement explanatory modeling in various ways, for example, by providing a baseline model with which an explanatory model must compete with in terms of predictive accuracy, or by supporting researchers in refining theoretical models based on empirical data (Brandmaier et al., 2016). In the following, we will shortly describe the mechanism and properties of two well-known regularization approaches in their original context of manifest linear regression. The outlined concepts are, however, equally applicable in latent regression analysis.

For ridge regression, the penalized loss function for univariate regression models takes the following form:

$$L_{ridge} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p}\beta_j^2 \qquad (2)$$

where λ denotes a tuning parameter that determines the influence of the penalty term (the sum of the squared regression weights) relative to the ordinary least squares (OLS) estimator.

A second approach is the so-called least absolute shrinkage and selection operator (lasso, Tibishirani, 1996).

$$L_{lasso} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \qquad (3)$$

Note that, if $\lambda > 0$, both functions only reduce to the original least square estimator when all β_j are 0. Therefore, the sum of all either squared regression weights (ridge) or their absolute values (lasso) need to be lower compared with OLS in order to minimize the regularized loss function.

Both of these regularization methods (see Equations 2 and 3) have been applied to a wide variety of regression problems (Hastie et al., 2009), providing distinct strengths and weaknesses (Fu, 1998; Tibishirani, 1996; Zou & Hastie, 2005). Ridge regression is typically better at handling high multicollinearity between predictors than the lasso. The latter often fails to exhibit a grouping effect in these circumstances, meaning a selection of only one of multiple highly correlated predictors while the others are shrunken to 0. This ability of variable selection can in return be seen as an advantage over ridge regression, because it can actually produce parsimonious and therefore much more interpretable models.

In the social and behavioral sciences, structural equation modeling (SEM) is widely used and allows researchers to model and test complex relationships based on latent variables (i.e., free of measurement error). Hence, SEM can be conceived as a combination of linear regression analysis (or path analysis) and confirmatory factor analysis:

$$\mathbf{y} = \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon} \qquad (4)$$

$$\mathbf{x} = \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta} \qquad (5)$$

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \qquad (6)$$

where Equations 4 and 5 refer to two measurement model relating the observed variables (denoted by the vectors $\mathbf{y}$ and $\mathbf{x}$) to the latent variables (denoted by the vectors $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$). Equation 6 represents the structural model relating the latent variables to each other, for example, in terms of a latent linear regression equation. So far, only few studies (e.g. Brandmaier & Jacobucci, in press) examined the performance of regularization methods in complex structural equation models for multidimensional data. Jacobucci et al. (2019) originally applied regularization methods in a multiple indicators multiple causes (MIMIC) model by shrinking the parameter estimates of regressions from latent variables onto manifest covariates not involved in the measurement of the construct. In principle, all paths between two latent or manifest variables in an SEM can be subject to regularization. Scharf and Nestler (2019) compared regularization models with conventional rotation techniques for explanatory factor analyses. In their simulation design, both factor loadings as well as factor covariances were regularized. More recently, Scharf et al. (2021) complemented this line of research by investigating the performance of regularization methods for regressing a latent variable onto multiple other measured factors.

Regularizing certain paths in an SEM can be understood as a compromise of the earlier mentioned explanatory and predictive modeling approach, making it particularly useful as a structured approach to theory development (Brandmaier & Jacobucci, in press). Regularized SEM promises to find stable, generalizable but not necessarily “true” models that expand or refine a baseline model by incorporating additional variables and their multivariate relationships.

In the present study, we focus on the first example by Jacobucci et al. (2019) but extend the simulation design to a more complex measurement model, consisting of two structurally different measurements of the same trait. In particular, we evaluate the performance of regularization methods for the explanation of general and specific factors in bifactor-(S-1) models using a large set of predictor variables, building on the method presented by Koch et al. (2018). Bifactor-(S-1) models have been proposed for measurement designs with fixed (or structurally different) methods as opposed to random (or interchangeable) methods (see Eid et al., 2017 for details). The key idea of the bifactor-(S-1) model is to specify a reduced number of specific factors and contrast structurally different methods (e.g., parent reports, implicit tests, or physiological data) against a reference method (e.g., self-reports). Today, the bifactor-(S-1) model is frequently applied in many psychological research fields (Burns et al., 2020; Chanal & Guay, 2015; Danay & Ziegler, 2011; Eid, 2020; Gäde et al., 2017; Heinrich et al., 2020).

Formally, the bifactor-(S-1) model is based on a latent regression approach in which one method (or facet) is chosen to represent the trait of interest and is then used as a latent predictor for the true-score variables of the remaining methods in a latent regression analysis (Eid et al., 2017). The resulting residuals of this latent regression are defined as latent method variables and share the same properties of residual variables. By definition, the latent method variables have an expectation of zero and are necessarily uncorrelated with the trait factor of the same trait-method unit. However, method variables can be correlated with each

Originally published in: Structural Equation Modeling, 30(4), 2023, p. 561

other within and across different trait-method units. The bifactor-(S-1) model allows for a separation of trait- and method-specific influences for all indicators, except for the reference method indicators (see Figure 1). Hence, the bifactor-(S-1) model enables researchers to decompose the total variance of each nonreference indicator (e.g., parent report, $k \neq 1$) into three parts (see Eid et al., 2017): a part that is shared with the reference trait (quantified by the consistency coefficient), a part that is *not* shared with the reference trait but that is attributable to method- or facet-specific influences (quantified by specificity coefficients), and a part that is due to measurement error (quantified by 1 minus the reliability coefficient). The basic idea of the bifactor-(S-1) model is to contrast structurally different methods (or facets) against a reference method (or facet). A reference method is preferably chosen based on theory, for example, to ease the interpretation of the results, to answer concrete research questions, to replicate previous results, or if there is a gold standard that is best suited to represent the trait of interest. Indicators for which no method factors are assumed are considered reference indicators and determine the meaning of the trait factor as the trait measured by the reference method. Through its directed path of TR_1 onto all nonreference indicators, it follows ($\forall\ k \neq 1$):

$$Y_{ik} = \alpha_{ik} + \lambda_{ik}\overbrace{(\beta_{0k} + \beta_{1k}TR_1 + M_k)}^{TR_k} + \varepsilon_{ik} \qquad (7)$$

$$Y_{ik} = \underbrace{a_{ik} + \lambda_{ik}\beta_{0k}}_{a_{Tik}} + \underbrace{\lambda_{ik}\beta_{1k}}_{\lambda_{Tik}} TR_1 + \lambda_{ik}M_k + \varepsilon_{ik} \qquad (8)$$

$$Y_{ik} = \alpha_{Tik} + \lambda_{Tik}T_1 + \lambda_{Mik}M_k + \varepsilon_{ik} \qquad (9)$$

where Y_{ik} is the i-th indicator measured by the k-th method. M_k represents a method factor and the trait T_1 is defined as trait measured by the reference method $T_1 \coloneqq TR_1$. In Equation 7. TR_k represents the true-score variable of a nonreference method (e.g., parents reported loneliness of a child) that can be linearly decomposed as follows: $\beta_{0k} + \beta_{1k}TR_1 + M_k$. Solving Equation 7 yields Equation 8, which shows how α_{Tik} and λ_{Tik} must be constrained so that the model in Equation 9 (see Figure 2) matches the original latent regression model in Equation 7.

In the above parametrization, the intercepts and loadings of the trait factor on the nonreference indicators are constrained ($\alpha_{Tik} \coloneqq \alpha_{ik} + \lambda_{ik}\beta_{0k}$ and $\lambda_{Tik} \coloneqq \lambda_{ik}\beta_{1k}$) to mimic

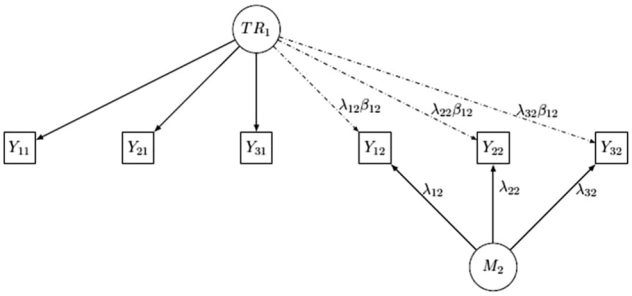

Figure 1. Restricted Bi-factor-(S-1) model. Trait influences on (nonreference) method indicators (dash-dotted) are restricted to be the product of their loading onto a specific factor (λ_{ik}) and one common parameter (β_{1k}). Error terms and mean structures are omitted for clarity.

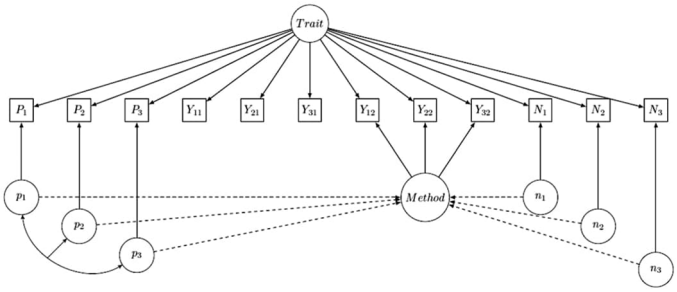

Figure 2. Simplified simulation model with only 3 informative (p_i) and 3 noise (n_i) predictors. Dotted paths are subject to regularization. The reference method is referred to as the trait factor, while all nonreference methods (i.e., only one) are referred to as method factors. Error terms and mean structures are omitted for clarity.

the indirect effects in Equations 7 and 8, and thus the model can be termed *restricted* bifactor-(S-1) model (cf. Geiser et al., 2008). In practice, the intercepts and loadings are typically not constrained in any way, in which case we obtain an unrestricted (or conventional) bifactor-(S-1) model (see Eid et al., 2017):

$$Y_{ik} = \mu_{Tik} + \delta_{Tik} T_1 + \lambda_{Mik} M_k + \varepsilon_{ik} \qquad (10)$$

where μ_{Tik} is the unrestricted intercept, δ_{Tik} is the unrestricted trait loading for the reference trait T_1, λ_{Mik} is the method loading pertaining to the method factor M_k, and ε_{ik} denotes the residual (measurement error) variable. In the above model, the intercept and loadings are freely estimated, except for the one loading per factor, which is fixed to 1 to set the scale ($\delta_{T1k} = \lambda_{M1k} = 1$). The unrestricted bifactor-(S-1) model often shows a superior fit as compared with the restricted version. Nevertheless, the restricted bifactor-(S-1) model is statistically equivalent to a correlated first-order factor model or a latent regression model. Moreover, the fit of the restricted bifactor-(S-1) model is invariant against the choice of the reference method, which may be attractive to applied researchers (see Geiser et al., 2008). In the following, we focus on the conventional (unrestricted) bifactor-(S-1) model, as this is more commonly used in practice.

Consider a multi-rater study in which each target person is rated on several personality traits using self-reports and parent reports. For simplicity, we select the self-report as a reference method and define it as the reference trait. Consequently, the parent report serves as the nonreference method and is contrasted against the self-reports. As shown in Figure 1, the trait factor reflects the true self-report whereas the method factor reflects the parent specific view on the child that is not shared with the child's self-report.

Similar to the MIMIC model used by Jacobucci et al. (2019), multiple covariates could be entered into the model to predict the method factor. For example, researchers may be interested in identifying key predictors of the parent-specific view. However, in measurement models where general and specific factors are necessarily uncorrelated (i.e., g-factor models), covariates should not be linked simultaneously to the general and specific factors as this would directly lead to a suppression structure and potential parameter bias (see Koch et al., 2018, for details). To avoid model misspecification and parameter bias, it is recommended to correct all covariates from confounding trait influences before predicting the specific (method) factors. This strategy has been termed residual approach (Koch et al., 2018), as confounding trait variance is partialed out from the covariates and solely the residualized (trait-free) covariates are used as predictors for the specific (method) factors. In our example, the specific view of parents (i.e., method factor) is explained by multiple external variables (e.g., attachment, emotional stability, age) while correcting for the child's self-report (see Figure 2). Conversely, one could also correct the covariates from confounding method variance to explain the trait factor (see Koch et al., 2018).

The aim of this study is to compare the performance of ridge and lasso regularization when regularizing the latent regression parameters from the latent (residualized) covariates to the method factor (see dotted lines in Figure 2). The performance of the two regularization methods was evaluated by means of the overall prediction accuracy as well as the estimation bias in comparison with ordinary Maximum-Likelihood (ML) estimation (i.e., without regularization methods). We varied the number of the predictors as well as the amount of multicollinearity between the predictor variables. Additionally, we varied the informativeness of the predictors in the model as it is often the case in exploratory studies to select a small subset of key predictors out of large group of variables (Jacobucci et al., 2019).

As the literature on the performance of regularization in complex SEM (such as bifactor models) is still sparse, our study aims to contribute to the question of how well regularization methods can be applied to SEM. Many articles use the RegSem-package (Jacobucci et al., 2016), the lsl- or its successor the lslx-package (Huang, 2020; Huang et al., 2017). Both Huang and Jacobucci detail examples of using the implemented algorithms for a mixture of exploratory and confirmatory factor analysis designs. The strength of this approach lies within not only exploring or confirming a certain model structure, but also giving more stable estimates (even though biased) for the sign and magnitude of single path coefficients.

Referring to our example above (see Figure 2), researchers are typically not only interested in the overall model fit but want to identify the subset of relevant (residualized) covariates that have an effect on the outcome as well as their effect sizes. Especially in designs with many predictors and a small sample size, ML point estimates will become exceedingly unreliable. Regularization methods provide a way to reduce the variance of these estimates with the caveat of introducing some bias. Therefore, the regularized estimates deviate on average from the actual population parameters, but carry much less risk of misrepresenting them completely due to sample inflation in magnitude or possibly sign reversal, which is common in linear regression with small sample sizes (Gelman & Carlin, 2014).

Scharf et al. (2021) examined regularization within an SEM consisting of four latent variables that also function as predictors for a fifth latent factor. Like in the current article, the degree of multicollinearity between the latent predictors was varied to uncover possible problems in the estimation with standard ML compared with regularization via ridge or the elastic net, a hybrid method combining ridge and lasso penalization (Zou & Hastie, 2005). In the current study, we used a total of nine predictors in clusters of three predictor sizes and varied the amount of multicollinearity within the clusters of predictors. Additionally, we included several noise parameters, that is, predictors with a true regression coefficient of 0. In the study by Scharf et al. (2021), only one predictor served as a noise parameter. To measure the influence of noise parameters, up to 25 noise parameters were included into the model. These parameters did not contribute to multicollinearity in the population level because they were all independent of each other and the informative predictors. Furthermore, our method for

introducing multicollinearity between latent predictors differed from the method by Scharf et al. (2021) as they adhere to Grewal et al. (2004) who adopted a method proposed by Mason and Perreault (1991) for simulating multicollinearity in manifest regression models based on R^2 (multiple R-squared). In this paradigm, the regression weights for the predictors are chosen freely, while the correlation between the predictors can be set for each individual pair. The regression weights will however be unstandardized, which makes it difficult to compare the results regarding different effect sizes. In the current simulation, we chose possible population parameters for predictor size and multicollinearity, which in turn facilitate comparisons between population and estimated parameters on a standardized scale. Further details on this method are described in the section below.

1. Simulation Design

We chose the simulation conditions based on the study by Jacobucci et al. (2019), but additionally examined the influence of different predictor effect sizes and noise parameters. To keep the simulation design, and thus the computational cost, within manageable limits, we slightly reduced the number of different values for sample size and multicollinearity among predictors, but still ensured that we covered a range that is common for applied research in psychology.

For sample size, we chose values of 100, 250, 500, and 1,000. For correlations among predictors, we chose values of 0, 0.2, 0.5, and 0.9, spanning a range from no to severe multicollinearity (although thresholds signifying low or high multicollinearity are disputed (see Grewal et al., 2004)). Because the underlying multicollinearity structure limits the range of possible effect sizes, we varied the effect sizes depending on the respective collinearity condition. Specifically, we first fixed the overall proportion of variance explained R^2, an overall effect size of the model) following Equation 11. This bears the advantage that the classification by Cohen (1988) can be used to compare the effect sizes to ordinary linear regression models with manifest variables. Second, we computed the effect sizes for the informative predictors assuming equal correlations among the informative predictors (see Equation 12):

$$R^2 = \boldsymbol{c}^T \boldsymbol{A}^{-1} \boldsymbol{c} \qquad (11)$$

$$r_{yx} = \sqrt{\frac{R^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}}} \qquad (12)$$

where $\mathbf{c}$ represents the vector of first order correlations between predictors and outcome and $\boldsymbol{c}^{\boldsymbol{T}}$ is its transpose. $\mathbf{A}^{-1}$ is the inverted correlation matrix of all predictors, with a_{ij} as its respective entries (elements).

We varied the amount of explained variance R^2 with values of .1, .2, .4 and .6 instead of the individual effect sizes of the predictors across different conditions. This corresponds to the following effect size by Cohen (1988): $\phi^2 = 0.11$, 0.25, 0.67, and 1.5. As the individual predictor effect sizes are directly related, they will become larger with an increasing R^2. To further distinguish between effect sizes, we defined three clusters of uncorrelated predictors each being responsible for a certain percentage of the respective explained variance (small = 11%, medium = 33%, large = 55%). Finally, we varied the number of (noninformative) noise predictors. To this end, we used values of 0, 10, and 25 resulting in 192 simulation conditions. For each condition, we simulated 200 data sets resulting in 192 x 200 = 38,400 data sets.

2. Procedure

All analyses were conducted in R (R Core Team, 2021) using lavaan (Rosseel, 2012) for ML and lslx (Huang, 2020)[1] for the regularization methods. The code is available on the OSF website: https://osf.io/4nz6j/. We specified the loading matrix according to the number of total predictors in the model. Each standardized trait loading for each covariate was set to a value of 0.2 across all conditions while the rest of the variance was explained by the respective latent residual variables and therefore had a loading of $\sqrt{1 - 0.2^2} \approx 0.98$. Because the predictors were used as independent variables in a regression of the method factor, all loadings from the manifest predictor variables onto the nonreference method factor were set to 0.

Following the residual approach by Koch et al. (2018), we fixed the correlation between the latent predictors and the trait to 0. The values for the collinearity between the predictors were chosen based on the respective condition. In order to differentiate different predictor sizes, we subdivided the nine informative predictors into groups of three with varying amounts of total predictive power (see Table 1). Only predictors within the same subgroup had a collinearity deviating from 0.

2.1. Performance Criteria

The statistical performance of the regularization method was evaluated with regard to the following criteria. We calculated the parameter estimation bias (peb) for each predictor size cluster (Koch et al., 2014):

$$peb(c) = \frac{1}{n_c}\sum_{p=1}^{n_c}(\left|M_{pc} - e_{pc}\right|) \qquad (13)$$

where n_c denotes the number of parameters in cluster c (i.e., 3), M_{pc} denotes the estimated parameter p in cluster c averaged over all 200 replications, and e_{pc} represents the corresponding true population parameter. For brevity, only the peb for the large effect size cluster is reported. Plots for the

[1]Preliminary analyses have been conducted with the RegSem-package (Jacobucci et al., 2016), however multiple problems occurred. First, the computational effort exceeded our expectations especially in conditions with a high number of noise variables rendering the simulation infeasible. Second, many converged models produced theoretically impossible values indicating improper solutions (Rindskopf, 1984). Third, after employing a rather simple and somewhat liberal indicator to avoid these theoretically impossible values, close to one third of all models indicated improper solutions. This indicator marked all models in which the sum of the squared regression coefficients (ridge) or the sum of their absolute values (lasso) increased beyond the respective value obtained with ML estimation.

Table 1. Standardized regression weights dependent on R^2, level of multicollinearity and effect size cluster.

		Standardized regression weights		
R^2	Multicollinearity	Small	Medium	Large
0.1	0	0.061	0.105	0.136
	0.2	0.051	0.089	0.115
	0.5	0.043	0.075	0.096
	0.9	0.036	0.063	0.081
0.2	0	0.086	0.149	0.192
	0.2	0.072	0.126	0.163
	0.5	0.060	0.105	0.136
	0.9	0.051	0.089	0.115
0.4	0	0.122	0.211	0.272
	0.2	0.103	0.178	0.230
	0.5	0.086	0.149	0.192
	0.9	0.073	0.126	0.162
0.6	0	0.149	0.258	0.333
	0.2	0.126	0.218	0.282
	0.5	0.105	0.183	0.236
	0.9	0.089	0.154	0.199

small and medium cluster can be attained with the help of the supplementary scripts at the OSF repository.

To assess overall prediction accuracy, we also computed the mean squared error (MSE) as a compound measure for estimation bias and variance (see Equation 1). Note that, to enable more visual differentiation between conditions, we chose the root mean squared error (RMSE) to display in our figures. For ridge as well as lasso regularization, the values for lambda were chosen via the lowest bayesian information criterion (BIC) (Jacobucci et al., 2019; Schwarz, 1978). Since ridge regression cannot set coefficients to 0 and therefore change (i.e., increase) degrees of freedom, robust degrees of freedom are used in this case (see Huang, 2020, Appendix A). Lambda values started at 0, meaning no penalization, and went up to 0.5 in steps of 0.01 akin to Jacobucci et al. (2019), who used a maximal value of 0.3. In order to keep the computational effort to a reasonable level, we did not extend the value range because preliminary analyses showed no best fitting results beyond this threshold.

2.2. Results

Table 2 shows the nonconvergence rates across all multicollinearity conditions for estimation via ridge, lasso and ML. While there were no convergence issues encountered in conditions with ordinary maximum likelihood estimation, both ridge and lasso show severe convergence issues, especially in conditions with high multicollinearity close to 0.9 and in conditions with many (25) noise parameters.

Nearly all nonconverged models were in the highest condition of multicollinearity, as indicated by nonconvergence warnings from the extraction functions of the lslx-package. For 25 noise parameters, no multicollinearity of 0.9 has been analyzed because estimated computation times would exceed a reasonable timeframe. Our results suggest that, in the 25-noise parameter condition, an onset of convergence problems is to be expected at already lower levels of multi- collinearity (see Table 2).

Table 2. Nonconvergence rates for different values of multicollinearity and estimation methods.

		Estimation method		
Multicollinearity	No. of noise parameters	Ridge	Lasso	ML
0	0	0.000%	0.000%	0.000%
	10	0.000%	0.000%	0.000%
	25	0.094%	0.094%	0.000%
0.2	0	0.000%	0.000%	0.000%
	10	0.000%	0.000%	0.000%
	25	0.031%	0.063%	0.000%
0.5	0	0.000%	0.000%	0.000%
	10	0.000%	0.000%	0.000%
	25	0.531%	0.594%	0.000%
0.9	0	68.594%	69.594%	0.000%
	10	67.750%	68.656%	0.000%
	25	NA*	NA*	0.000%

*NA = Not Available. For 25 noise parameters, no models with multicollinearity condition of 0.9 have been fitted because of exceedingly high computation times.

To avoid distorting the parameter estimates from the converged models, all nonconverged models were excluded from the subsequent data analysis for corresponding conditions. This reduced the available data significantly in conditions of highest multicollinearity between predictors.

To study the overall efficiency of estimating the regression parameters, we calculated the RMSE (Figure 3). To focus on the most illustrative results, we report only the results for the large effect sizes as well as for the model with 10 noise parameters because herein we obtained estimates for all levels of multicollinearity and included noise parameters, which benefit from the shrinkage effect of regularization in any simulated sample in respect to their estimation bias. Nonetheless, we provide a comprehensive collection of R-scripts for producing plots for all conditions in the online supplemental material (https://osf.io/4nz6j/).

The estimation methods only differed to a relatively small amount when multicollinearity was low. However, for the condition of high multicollinearity, discrepancies in estimation efficiency emerge between the three methods. While all methods become more efficient with increasing sample size, ML estimation consistently produces higher RMSEs than the regularization methods, as expected. Overall, ridge regression performs almost uniformly throughout all conditions of sample size, R^2, and multicollinearity, and produces the lowest RMSE throughout the simulation. In comparison, the lasso is a little less efficient, especially when the sample size is 500 or below. For a fair comparison with ML estimation, we only included datasets, which converged for both ridge and lasso. Still, ML is clearly outperformed in terms of RMSE, particularly when multicollinearity is at 0.9 and sample size is at or below 250.

Next, we evaluated the absolute parameter estimation bias (peb, Figure 4). In line with prior expectation, both lasso and ridge introduce some bias while ML estimates are relatively unbiased. Again, ridge regression performs rather uniformly whereas lasso shows significantly lower biases with increasing sample size. For differences in R^2 and

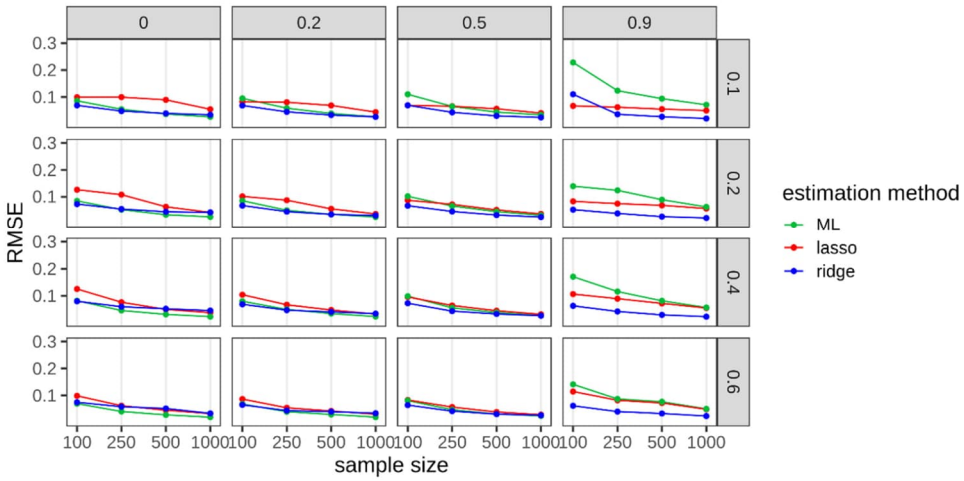

Figure 3. Comparison of RMSE for ML, lasso, and ridge estimation for large predictors and 10 noise parameters. Columns indicate multicollineaity, rows indicate R^2, and color indicates the estimation method.

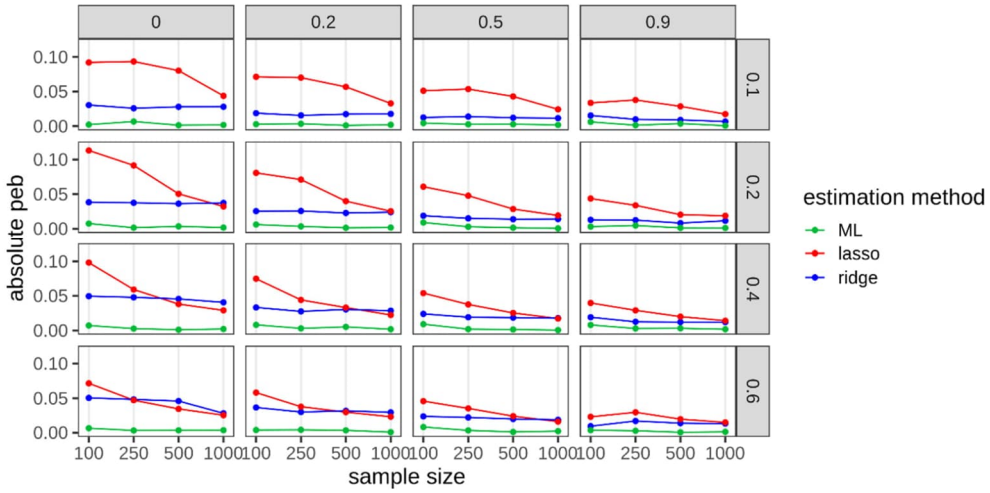

Figure 4. Absolute parameter estimation bias for the large effect size cluster with 10 noise parameters. Columns indicate multicollinearity, rows indicate R^2, and color indicates the estimation method.

therefore effect sizes, a trend in opposite directions occurred. While the lasso clearly produces larger biases for smaller R^2, biases in the ridge tend to increase with increasing R^2, although less pronounced. Additionally, a pattern of decreasing bias for increasing multicollinearity can be observed for both regularization methods. As shown in Table 1, higher multicollinearity also corresponds to smaller effect sizes; however, as already mentioned, ridge and lasso do not necessarily react the same way to variation in true regression weights.

As mentioned above, variable selection is a prominent advantage of the lasso in comparison to ridge regression. Of course, the selection process is not perfect as sometimes noise predictors will still be included in the model (false-positive) as well as informative predictors will be falsely set to 0 (false-negative). Figure 5 depicts the average false-positive rate while Figure 6 shows the average false-positive rate over all conditions.

Because, in the condition of 0 noise parameters, a false-positive result is impossible, there are only results for false-positives for 10 and 25 noise predictors. Figure 5 shows a substantial increase in false-positive rates when R^2 and therefore also effect sizes of the informative predictors become larger. This might seem surprising at first because the signal to noise ratio increases with higher R^2 which in turn should facilitate the identification of truly informative and noninformative predictors. However, findings by Su et al. (2017) show that the lasso often selects some noise predictors after selecting a few informative predictors with comparatively high effect size. High multicollinearity (0.9) on the other hand seems to increase false inclusions of predictors mostly for a small R^2 and small sample sizes. For a sample size of 250 and above, however, higher multicollinearity is associated with a lower false-positive rate. Finally, an increase in noise parameters appears to be associated with a slight reduction in false-positive rates in nearly all conditions.

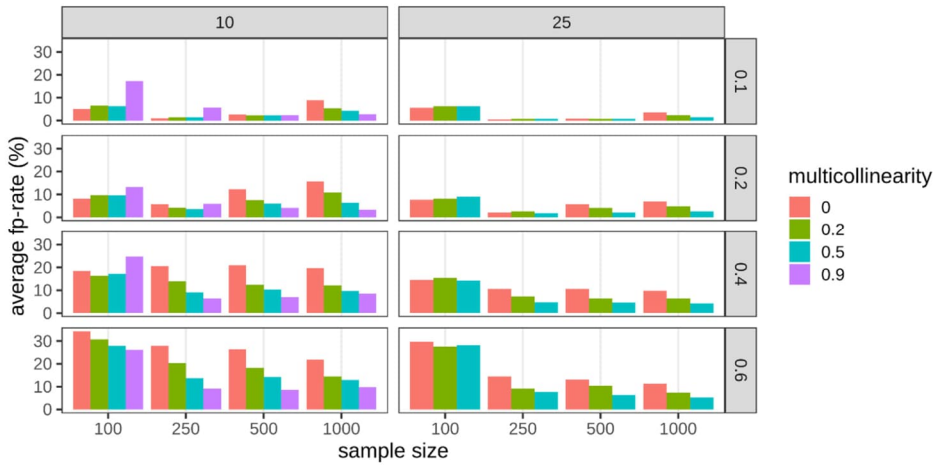

Figure 5. Average false-positive rate for the lasso. Columns indicate amount of noise predictors, rows indicate R^2, and color indicates the amount of multicollinearity. For 25 noise parameters, no data is available for a multicollinearity of 0.9.

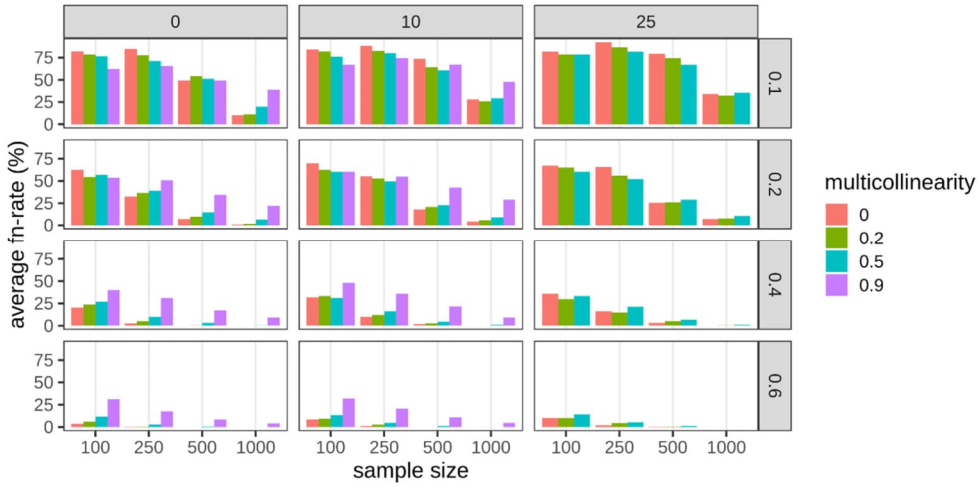

Figure 6. Average false-negative rate for the lasso. Columns indicate amount of noise predictors, rows indicate R^2, and color indicates the amount of multicollinearity. For 25 noise parameters, no data is available for a multicollinearity of 0.9.

Figure 6 shows a complementary picture to Figure 5 in the sense that whenever false-positive rates are low, false-negative rates are high and vice versa. This constitutes a trade-off described in Su et al. (2017), which makes it impossible for the lasso to produce both parsimonious models and recover all potentially relevant predictors. Nonetheless, it can offer this trade-off without further inference while ML or ridge regression either retain all variables in the model or would have to rely on significance testing to classify predictor estimates as (ir)relevant. In addition, the diagram shows the correspondence between high multicollinearity and a higher rate of false-negative results. This can be interpreted as a result of the absent grouping effect in the lasso (Zou & Hastie, 2005), which means that, for groups of highly correlated predictors, the lasso tends to only retain one of them in the model and shrinks the remaining ones to 0.

The high false-negative rates in some conditions imply that coefficient sizes are subject to high amounts of shrinkage, which can be undesirable when a model is not characterized by a few highly influential covariates, but rather many covariates with a small to moderate association. In these situations, it may be helpful to use a refitting strategy[2]. For this purpose, the lasso would only serve as a model or variable selection procedure while, in a second step, a different estimator such as conventional OLS is applied. Chzhen et al. (2019) discuss the associated issue of a possible sign reversal of coefficients and how to combat this problem.

[2]We thank an anonymous reviewer for their remark regarding this topic.

Originally published in: Structural Equation Modeling, 30(4), 2023, p. 567

However, the interpretability of the resulting models can still be questionable as, for example, p-values will be biased as the selection of variables and their parameter estimates are based on the same data. Moreover, as shown here, the selection process via the lasso can be imperfect, which raises the question in which way bias and variance of the parameters in a refitted model are influenced. Therefore, we cannot make a clear recommendation of whether a refitting strategy is an effective solution.

3. Discussion of Study 1

For overall prediction accuracy, ridge and lasso differ the most from regular ML estimation in conditions of high multicollinearity as well as low to moderate sample sizes. The improvement in estimation efficiency measured by the (R)MSE shows potential when stability of parameter estimates is more important than overall unbiasedness. In cases when researchers and practitioners still value the latter, ridge regression outperformed the lasso by producing considerably less bias, particularly in smaller sample sizes. When parsimony is preferred on the other hand, the lasso is to be favored over ridge regression because of its inherent ability of variable selection.

It is worth noting that the observed convergence rates in conditions of high multicollinearity was far from satisfactory and severely attenuates the possible benefits in particularly these conditions. In addition, estimating the most complex models proved to be comparatively time demanding (150–160 seconds per model on average in the condition of 10 noise parameters) considering the rather quickly computed ML solutions (about a fifth of a second per model on average)[3].

Because regularization, especially ridge regression, is theoretically well equipped to handle high multicollinearity, these results of a problematic estimation process led us to speculate whether they may originate from implementation issues rather than from the regularization methods themselves. Therefore, we conducted a second simulation in order to investigate if using a different software package could solve the convergence issues of the regularization methods in case of high collinearity. Because lslx and RegSem were to our knowledge the only available R-packages for regularization within structural equation models in a single step, we settled on a two-step process by first obtaining factor scores and then using them as predictors in a regularized manifest regression. For the second step, we chose the glmnet-package (Friedman et al., 2010), as it provides regularized solutions for manifest regressions almost at the same speed as OLS or ML estimation and is also generally known to not be susceptible to convergence problems.

3.1. Study 2

For the second simulation, the same population parameters and datasets as described earlier were used. However, the estimating process was split into two steps. First, the factor scores for covariates as well as the method-factor were estimated based on the results of the ML solution via the regression method as it is implemented in the lavaan-package (Rosseel, 2012). In a second step, these values then served as parts of regularized linear regression using the glmnet-package (Friedman et al., 2010). Instead of the BIC, 10-fold cross-validation was used to determine the best-fit lambda. Again, lasso and ridge regression were compared using the same metrics as in study 1.

We also added elastic net regression to the comparison (Zou & Hastie, 2005) as it combines the strengths of ridge and lasso regularization. The loss function for the elastic net can be expressed as

$$L_{enet} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + (1-\alpha)\lambda\sum_{j=1}^{p}\beta_j^2 + \alpha\lambda\sum_{j=1}^{p}|\beta_j| \qquad (14)$$

where the additional hyperparameter α determines the split between ridge and lasso type penalties. Its values range from 0 (equivalent to pure ridge), to 1 (equivalent to pure lasso). The R-package glmnetUtils (Ooi, 2021) offers an easy way to cross-validate the α and λ parameter at the same time.

3.2. Results

In contrast to study 1, all models in study 2 converged without problems. Additionally, the computation time for the regularization was vastly reduced as the two-step regularized regression found their solutions on average in about a fifth of a second per model in the 10 noise predictor condition. In the single-step method conducted via lslx, the average run time per model was about 150–160 seconds, as mentioned above.

The RMSE shows a similar pattern as already observed in the first study with lasso and ridge regression both exhibiting lower levels than ML in conditions of high collinearity and low to moderate sample sizes, although the differences are less pronounced (Figure 7). High multicollinearity now affects the regularization methods more than in the single-step approach, which could be attributed to the now fully available datasets for a collinearity of 0.9. In addition, ridge and the lasso now perform more similarly than in study 1. The elastic net performs somewhere between ridge and lasso, which is unsurprising given its definition as a flexible mixture of both these types of penalty terms.

Concerning the parameter estimation bias, both the lasso and ridge regression showed reduced levels in comparison with the single-step approach. The lasso now profits from an increase in sample size much earlier than before and assumes a similar trajectory to ridge regression, which before was rather unaffected by changes in the sample size. Higher R^2 now appears to generally correspond to lower biases for both regularization methods in contrast to the results from study 1. Higher collinearity on the other hand is still associated with a lower bias, even though effect sizes become smaller. The elastic net introduces about as much bias as ridge regularization over all conditions (Figure 8).

[3]Calculations were performed under SMP Debian 5.10.113-1 (2022-04-29) x86_64 GNU/Linux running on a machine with an AMD Ryzen Threadripper 3970X 32-Core Processor with 2195 Mhz and 1/16/128 MB L1/L2/L3-Cache respectively as well as a maximum of 126 GiB of RAM.

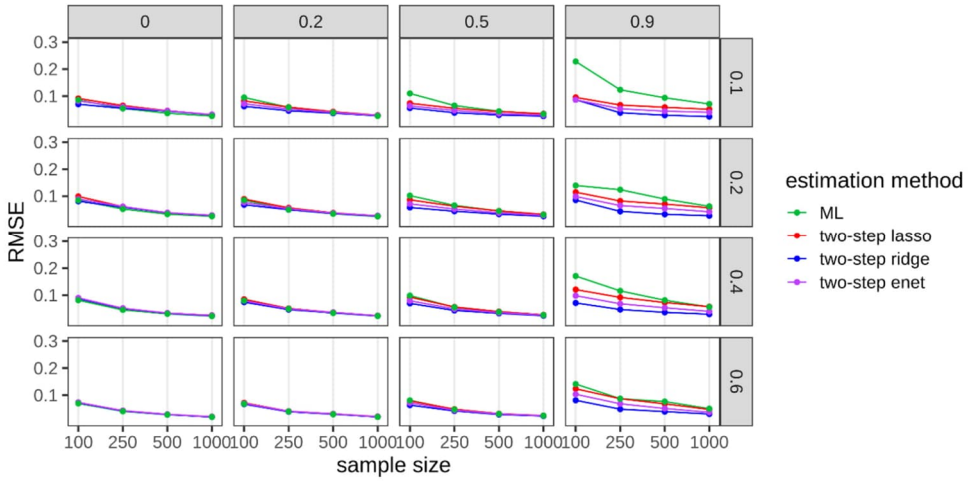

Figure 7. Comparison of RMSE for ridge, lasso, and elastic net (two-step method) and ML estimation for large predictors and 10 noise parameters. Columns indicate multicollinearity, rows indicate R^2, and color indicates the estimation method.

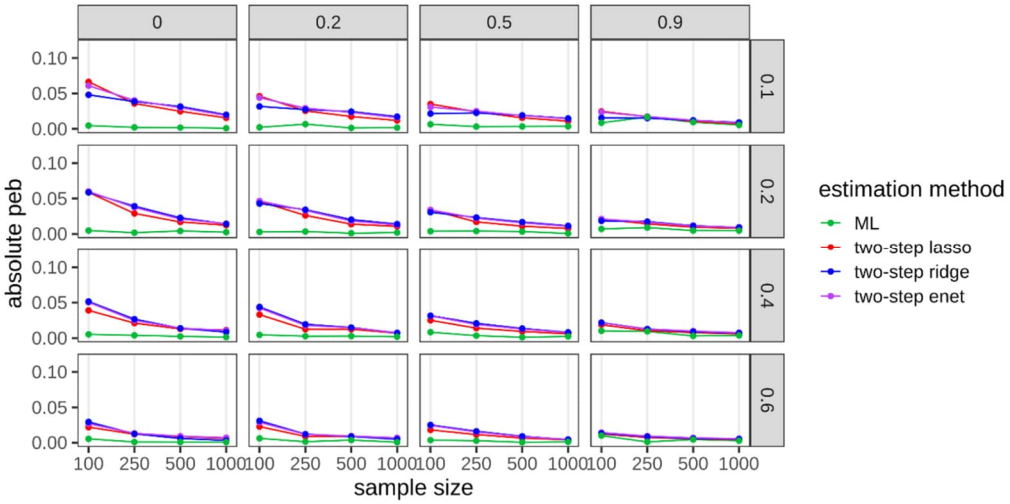

Figure 8. Absolute parameter estimation bias for the large effect size cluster with 10 noise parameters in the two-step approach. Columns indicate multicollinearity, rows indicate R^2, and color indicates the estimation method.

Finally, the false-positive as well as the false-negative rates for the lasso in the two-step approach are presented in Figures 9 and 10, respectively. Overall, the false positive rate is considerably higher than compared with the single-step approach with no condition under 20%, which implies the choice of a lower penalty parameter.

This discrepancy to the single-step approach is likely due to the different selection criteria (BIC vs. *k*-fold cross-validation) available with the lslx- and glmnet-package. The influence of R^2 and multicollinearity on the other hand mimics the results from the single-step as false positives tend to increase with higher R^2 and decrease with higher multicollinearity. The influence of the increase in noise parameters is still noticeable, but less pronounced as an overall smaller false-positive rate can be observed with 25 noise parameters compared with 10.

Similar to study 1, the false-negative rates complement the false-positive rates. In accordance, false-negatives occur less frequently in the two-step compared with the single-step approach. This again is likely to be a result of a smaller penalty parameter, which leads to more predictors retained in the model. As already depicted in Figure 6, high multicollinearity coincides with a higher false-negative rate, which can be attributed to the absent grouping effect in the lasso.

4. Discussion of Study 2

By using factor score regression as well as the well-established glmnet package, the convergence issues found in study 1 could be successfully circumvented. Parameter estimates were now available for all levels of multicollinearity. They allowed for a more robust interpretation of the overall

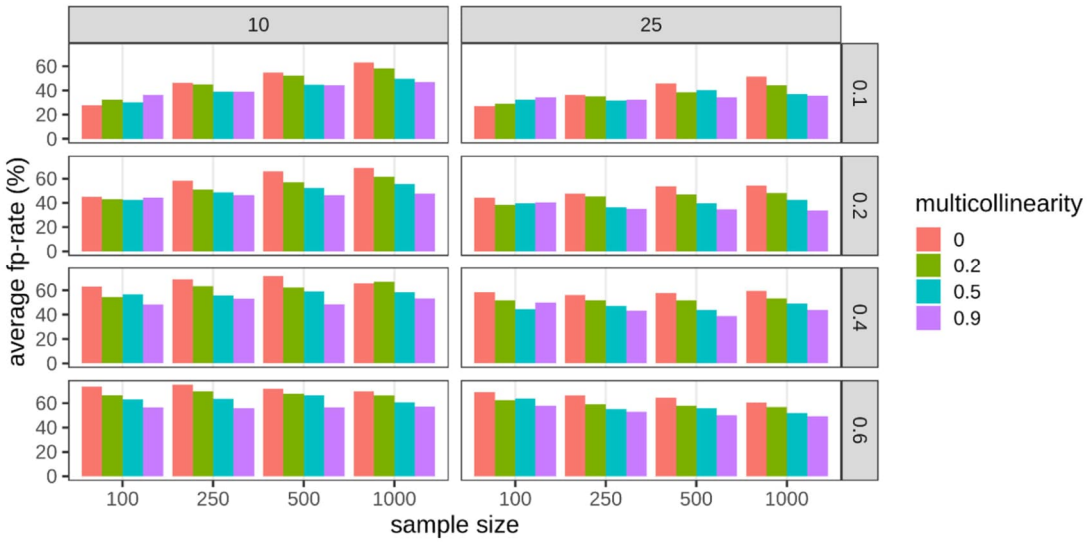

Figure 9. Average false-positive rate for the lasso (two-step). Columns indicate amount of noise predictors, rows indicate R^2, and color indicates the amount of multicollinearity. For 25 noise parameters, no data is available for a multicollinearity of 0.9.

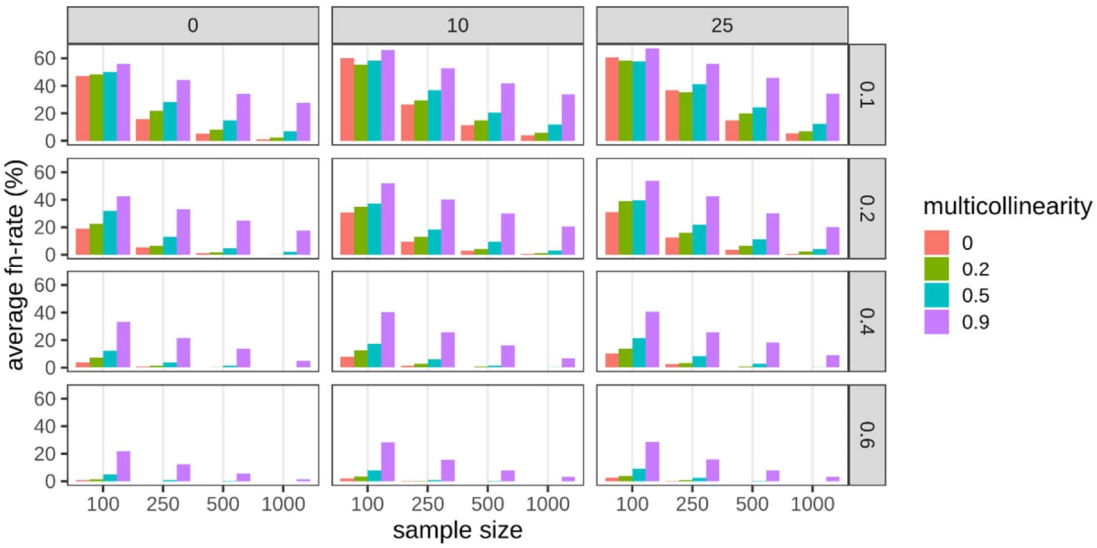

Figure 10. Average false-negative rate for the lasso (two-step). Columns indicate amount of noise predictors, rows indicate R^2, and color indicates the amount of multicollinearity. For 25 noise parameters, no data is available for a multicollinearity of 0.9.

results and performance of ridge and lasso regression. While the two-step method necessarily introduces some uncertainty because the relevant factor scores are themselves only estimates, there appears to be little concern about a drop in overall estimation accuracy for the regression weights of interest.

An important limitation to the two-step method comes in its comparatively low versatility. While packages like lslx offer to regularize almost all directed and undirected paths in a structural equation model, the workaround of first estimating factor scores and then using them in a regularized regression is limited to latent regressions and cannot be performed for factor loadings or covariances. Additionally, the two-step regularized regressions will be optimized solely in respect to the model consisting of the covariates and the latent variable of interest and therefore not take into account the overall fit of the specified SEM. As mentioned earlier, for the single-step method, the overall model fit represented by the BIC was used to select an optimal solution. Depending on the context, both these methods could be justified, but it ultimately remains for the applying researcher to decide.

5. General Discussion

In this article, we discussed the application of regularization methods for predicting latent specific (method) factors in bifactor S-1 models. To assess the overall performance of two popular regularization methods, the ridge and the lasso,

we varied the collinearity between the informative predictors as well as their overall effect size, the amount of noninformative covariates in the model, and the sample size. We compared the results obtained by performing regular ML estimation.

We expected the regularization methods and in particular the ridge regression to handle multicollinearity among predictors rather well because it has been developed to address this problem (Hoerl & Kennard, 1970). However, the results showed high rates (in some conditions over 80%) of non-convergence for models with a high degree of multicollinearity, while ML estimation always converged. For the computation of the (R)MSE as a measure for overall estimation accuracy, only convergent models were considered. In line with other findings in the literature, regularization is particularly promising when sample size is low and multicollinearity is strong.

To further investigate the excess rate of nonconverging models, we tried a different approach next. In a two-step procedure, we first estimated the factor scores from the ML results for the latent covariates and the specific factor and then used them for a regularized regression with the glmnet-package (Friedman et al., 2010). Then, all estimated models converged even in the condition of highest multicollinearity while mostly maintaining their improvements (in terms of lowered (R)MSE) in comparison with the ML procedure. Note that this kind of manifest analysis is also possible to run with the lslx-package with the same result of complete absence of convergence problems. However, in contrast with glmnet, lslx is not primarily designed for manifest analyses, which resulted in higher computation times.

The difference in run time between the full model estimation via lslx and the reduced version via glmnet is quite substantial. While lslx in the example of 10 noise parameters averaged a run time of 164 seconds for a single model with lasso regularization, glmnet computed one such model within about .2 seconds. When using the two-step approach via lslx, the computation time for a single model was still at about 2 seconds and therefore still considerably slower. The lavaan solution needed for the first step averaged a run time of again approximately .2 seconds. For estimating a single model, as it would often times be the case in applied research, these differences in computational effort do not pose a problem. However, when there is interest in computing a series of models with highly correlated variables (high collinearity), the run times might exceed a reasonable threshold because an increase in the number of estimated variables also appears to drastically extend the durations of computation. The two-step method in contrast provided very quick solutions with no particular sensitivity to model size or amount of multicollinearity. Furthermore, computational efficiency is not only an issue in simulation studies but also in studies that use hyperparameter tuning to select an optimal penalty by computational approaches such as cross-validation.

Concluding from these observations, the two-step method in comparison with the single-step method delivers much more stable estimates for the kind of model specified in this article in conditions of very high multicollinearity. While it cannot optimize the fit of the overall model, the manifest regression between estimated factor scores can function as a reasonable substitute when models via lslx fail to converge for any of the pre-specified penalty parameter values.

Generally, the packages used to perform regularization within SEMs focus on selecting the penalty parameters by optimizing the overall model fit via information criterions like the AIC or BIC. However, for manifest regressions, Hastie et al. (2009) recommend cross-validation, which seeks to find a penalty parameter that minimizes the out of sample MSE or produces the most sparse model within one standard error of this best-fit parameter. It is still unclear which path or possibly combination of paths is most suitable for penalty parameter selection in structural equation models. Information criteria offer a global model optimization, but it is yet unclear how regularization reduces degrees-of-freedoms, specifically in ridge regression. The MSE for a specific regression on the other hand is unable to assure a globally best fitting model, but has no trouble accounting for the improved stability of parameter estimates.

Currently, we recommend researchers avoid the single-step approach with RegSem and lslx for bifactor-(S-1)-type models when a sizable amount of multicollinearity (of greater than 0.9) is to be expected within their model. Our simulation has shown that even in small clusters consisting of three predictors, high multicollinearity can lead to serious convergence issues. However, it needs to be noted that the two-step method is only feasible when a regression between latent variables shall be regularized and, thus, considerably limits the applicability of regularization in SEM. Some researchers might be more interested in regularizing factor loadings or factor covariances in order to perform what Huang (2020) calls “semi-confirmatory structural equation modeling”. Scharf and Nestler (2019) discuss this method in more detail.

Splitting the SEM estimation and regularizing the prediction of one factor through covariates into two parts outperformed the single-step method in our simulation and can function as a reliable substitute until the optimizers have been refined to deal with problematic conditions such as high multicollinearity and low to moderate sample size.

Promising results in this regard were achieved by a reanalysis of some of the problematic conditions with the recent Julia (Bezanson et al., 2017) packages StructuralEquationModels.jl (Ernst & Peikert, 2022b) and ProximalSEM.jl (Ernst & Peikert, 2022a). For the condition of 10 noise parameters as well as the highest multicollinearity level of 0.9, all models converged for ridge with the RMSE showing a very similar trajectory as discussed in both studies of this article. Furthermore, computation time was on average even lower than for the ML solution. The respective plots can be found in the appendix.

In conclusion, we suggest a cautious approach when facing high multicollinearity within regularized multidimensional structural equation models. While packages like lslx and RegSem are promising with regard to improving

Originally published in: Structural Equation Modeling, 30(4), 2023, p. 571

generalizability of models while also offering formal venues for theory development, there still seem to be fundamental issues with the current implementations and off-shelf optimizers used that need to be addressed before the single-step approach is recommendable.

ORCID

Maximilian S. Ernst http://orcid.org/0000-0003-2237-6255
Aaron Peikert http://orcid.org/0000-0001-7813-818X
Andreas M. Brandmaier http://orcid.org/0000-0001-8765-6982
Tobias Koch http://orcid.org/0000-0002-8143-3566

References

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*, 65–98. https://doi.org/10.1137/141000671

Brandmaier, A. M., & Jacobucci, R. C. (in press). Machine-learning approaches to structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd rev. ed.). New York: Guilford Press.

Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods*, *21*, 566–582. https://doi.org/10.1037/met0000090

Bullain, S., & Doody, R. (2020). What works and what does not work in Alzheimer's disease? From interventions on risk factors to anti-amyloid trials. *Journal of Neurochemistry, 155*, 120–136. https://doi.org/10.1111/jnc.15023

Burns, G. L., Geiser, C., Servera, M., Becker, S. P., & Beauchaine, T. P. (2020). Application of the bifactor S–1 model to multisource ratings of ADHD/ODD symptoms: An appropriate bifactor model for symptom ratings. *Journal of Abnormal Child Psychology, 48*, 881–894. https://doi.org/10.1007/s10802-019-00608-4

Chanal, J., & Guay, F. (2015). Are autonomous and controlled motivations school-subjects-specific? *PloS One, 10,* e0134660. https://doi.org/10.1371/journal.pone.0134660

Chzhen, E., Hebiri, M., & Salmon, J. (2019). On lasso refitting strategies. Bernoulli, 25, 3175–3200. https://doi.org/10.3150/18-BEJ1085

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Routledge. https://doi.org/10.4324/9780203771587

Danay, E., & Ziegler, M. (2011). Is there really a single factor of personality? A multirater approach to the apex of personality. *Journal of Research in Personality, 45*, 560–567. https://doi.org/10.1016/j.jrp.2011.07.003

Eid, M. (2020). Multi-faceted constructs in abnormal psychology: Implications of the bifactor S-1 model for individual clinical assessment. *Journal of Abnormal Child Psychology, 48*, 895–900. https://doi.org/10.1007/s10802-020-00624-9

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods, 22*, 541–562. https://doi.org/10.1037/met0000083

Ernst, M. S., & Peikert, A. (2022a). ProximalSEM.jl [Computer Software.] https://doi.org/10.5281/zenodo.7223919 (Original work published 2021.)

Ernst, M. S., & Peikert, A. (2022b). StructuralEquationModels.jl [Computer software.] https://doi.org/10.5281/zenodo.6719627

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*, 1–22. https://doi.org/10.18637/jss.v033.i01

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics, 7*, 397–416. https://doi.org/10.2307/1390712

Gäde, J. C., Schermelleh-Engel, K., & Klein, A. G. (2017). Disentangling the common variance of perfectionistic strivings and perfectionistic concerns: A bifactor model of perfectionism. *Frontiers in* Psychology, 8, 160. https://doi.org/10.3389/fpsyg.2017.00160

Geiser, C., Eid, M., & Nussbeck, F. W. (2008). On the meaning of the latent variables in the CT-C (M-1) model: A comment on Maydeu-Olivares and Coffman. *Psychological Methods, 13*, 49–57. https://doi.org/10.1037/1082-989X.13.1.49

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 9*, 641–651. https://doi.org/10.1177/1745691614551642

Grewal, R., Cote, J. A., & Baumgartner, H. (2004). *Multicollinearity and measurement error in structural equation models: Implications for theory testing. Marketing Science*, *23*, 519–529. https://doi.org/10.1287/mksc.1040.0070

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer Science & Business Media. https://doi.org/10.1007/978-0-387-84858-7

Heinrich, M., Zagorscak, P., Eid, M., & Knaevelsrud, C. (2020). Giving G a meaning: An application of the bifactor-(S-1) approach to realize a more symptom-oriented modeling of the Beck depression inventory–II. *Assessment*, 27, 1429–1447. https://doi.org/10.1177/1073191118803738

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55–67. https://doi.org/10.2307/1267351

Huang, P.-H. (2020). lslx: Semi-confirmatory structural equation modeling via penalized likelihood. *Journal of Statistical Software, 93*, 1–37. https://doi.org/10.18637/jss.v093.i07

Huang, P.-H., Chen, H., & Weng, L.-J. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika, 82,* 329–354. https://doi.org/10.1007/s11336-017-9566-9

Jacobucci, R., Brandmaier, A. M., & Kievit, R. A. (2019). A practical guide to variable selection in structural equation modeling by using regularized multiple-indicators, multiple-causes models. *Advances in Methods and Practices in Psychological Science, 2*, 55–76. https://doi.org/10.1177/2515245919826527

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/10.1080/10705511.2016.1154793

Koch, T., Holtmann, J., Bohn, J., & Eid, M. (2018). Explaining general and specific factors in longitudinal, multimethod, and bifactor models: Some caveats and recommendations. *Psychological Methods, 23*, 505–523. https://doi.org/10.1037/met0000146

Koch, T., Schultze, M., Eid, M., & Geiser, C. (2014). A longitudinal multilevel CFA-MTMM model for interchangeable and structurally different methods. *Frontiers in Psychology, 5,* 311. https://doi.org/10.3389/fpsyg.2014.00311

Mason, C. H., & Perreault, W. D., Jr. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research, 28,* 268–280. https://doi.org/10.2307/3172863

Ooi, H. (2021). *glmnetUtils: Utilities for 'Glmnet'.* In https://CRAN.R-project.org/package=glmnetUtils

R Core Team. (2021). *R: A Language and Environment for Statistical Computing.* In R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rindskopf, D. (1984). Structural equation models: Empirical identification, Heywood cases, and related problems. *Sociological Methods & Research, 13,* 109–119. https://doi.org/10.1177/0049124184013001004

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software, 48,* 1–36. https://doi.org/10.18637/jss.v048.i02

Sainani, K. L. (2014). Explanatory versus predictive modeling. *PM & R: The Journal of Injury, Function, and Rehabilitation, 6*, 841–844. https://doi.org/10.1016/j.pmrj.2014.08.941

Scharf, F., & Nestler, S. (2019). Should regularization replace simple structure rotation in exploratory factor analysis? *Structural Equation Modeling: A Multidisciplinary Journal, 26,* 576–590. https://doi.org/10.1080/10705511.2018.1558060

Scharf, F., Pförtner, J., & Nestler, S. (2021). Can ridge and elastic net structural equation modeling be used to stabilize parameter estimates when latent factors are correlated? *Structural Equation*

Modeling: A Multidisciplinary Journal, 28, 928–940. https://doi.org/10.1080/10705511.2021.1927736
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464. https://doi.org/10.1214/aos/1176344136
Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25,* 289–310. https://doi.org/10.1214/10-STS330
Su, W., Bogdan, M., & Candes, E. (2017). False discoveries occur early on the lasso path. *The Annals of Statistics, 45,* 2133–2150. https://doi.org/10.1214/16-AOS1521
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Appendix

The above plot compares the results from Study 1 with the results obtained by ridge regularization with the StructuralEquationModels.jl package by Ernst and Peikert (2022b). Again, columns indicate multicollinearity and rows indicate R^2. This time, the RMSE is averaged over all predictor size clusters. We can clearly see that the trend of improving the RMSE through a single-step approach when multicollinearity is high remains unchanged, but it is now much more reliably interpretable because the Julia-package has no trouble computing models in the highest multicollinearity condition without encountering any convergence issues. For more information on recreating this simulation within Julia, see the OSF-Repository: https://osf.io/4nz6j/