General Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Advances in Methods and Practices in Psychological Science
2019, Vol. 2(1) 55–76


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245919826527
www.psychologicalscience.org/AMPPS

A Practical Guide to Variable Selection in Structural Equation Modeling by Using Regularized Multiple-Indicators, Multiple-Causes Models

Ross Jacobucci[1], Andreas M. Brandmaier[2,3], and Rogier A. Kievit[3,4]

[1]Department of Psychology, University of Notre Dame; [2]Max Planck Institute for Human Development, Berlin, Germany; [3]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, Germany, and London, United Kingdom; and [4]Medical Research Council Cognition and Brain Sciences Unit, University of Cambridge

Abstract

Methodological innovations have allowed researchers to consider increasingly sophisticated statistical models that are better in line with the complexities of real-world behavioral data. However, despite these powerful new analytic approaches, sample sizes may not always be sufficiently large to deal with the increase in model complexity. This difficult modeling scenario entails large models with a limited number of observations given the number of parameters. Here, we describe a particular strategy to overcome this challenge: *regularization*, a method of penalizing model complexity during estimation. Regularization has proven to be a viable option for estimating parameters in this small-sample, many-predictors setting, but so far it has been used mostly in linear regression models. We show how to integrate regularization within structural equation models, a popular analytic approach in psychology. We first describe the rationale behind regularization in regression contexts and how it can be extended to regularized structural equation modeling. We then evaluate our approach using a simulation study, showing that regularized structural equation modeling outperforms traditional structural equation modeling in situations with a large number of predictors and a small sample size. Next, we illustrate the power of this approach in two empirical examples: modeling the neural determinants of visual short-term memory and identifying demographic correlates of stress, anxiety, and depression.

Keywords

regularization, structural equation models, MIMIC, lasso, variable selection, open materials

Received 2/16/18; Revision accepted 12/14/18

The empirical sciences have seen a rapid increase in data collection, both in the number of studies conducted and in the richness of data within each study. With large numbers of variables available, researchers often want to go beyond what their hypothesis-driven models tested and explore which variables are most informative in explaining observed variability in the outcome of interest. Typical questions asked are "What is the importance of my variables for predicting the outcome of interest?" and, ultimately, "What subset of variables is most predictive of (or most relevant for) that outcome?"

How to select a subset of variables for modeling purposes (*variable selection*) is a pervasive challenge in applied statistics. In the field of statistical learning (also known as machine learning or data mining), a large amount of attention has been dedicated to the topic of how predictors can be optimally selected when there is little or no prior knowledge. Statistical approaches to variable selection range from the notorious stepwise variable-selection procedures (cf. Thompson, 1995) to more complex and comprehensive approaches, such as

Corresponding Author:
Ross Jacobucci, Department of Psychology, University of Notre Dame, 390 Corbett Hall, Notre Dame, IN 46556
E-mail: rjacobuc@nd.edu

support vector machines and random forests. One particularly fruitful approach is *regularized regression*, a method that solves the variable-selection problem by adding a penalty term that penalizes solutions, effectively producing sparse solutions in which only few predictors are allowed to be "active." Regularization approaches vary in their precise specifications and include methods such as ridge (Hoerl & Kennard, 1970), lasso (least-absolute-shrinkage-and-selection operator; Tibshirani, 1996), and elastic-net (Zou & Hastie, 2005) regression.

Despite their strengths, these regularization approaches are generally developed in a context of models that include only observed indicators and consequently do not allow for modeling measurement error. However, incorporation of measurement error is central to many approaches in psychology. The most dominant approach to incorporating measurement error in psychology and related fields is the use of structural equation modeling (SEM). SEM offers a general framework in which hypotheses can be formulated at the construct (latent) level and explicit measurement models link the observed variables to the latent constructs. Latent-variable models account for measurement error, assess reliability and validity, and often have greater generalizability and statistical power than methods based on observed variables (e.g., Brandmaier, Wenger, Raz, & Lindenberger, 2018; Little, Lindenberger, & Nesselroade, 1999). Here we describe a novel approach called *regularized* SEM, which incorporates the strengths of regularization into the SEM framework, allowing researchers to estimate sparse model solutions and implicitly solve large-scale variable selection in SEM by introducing a penalized likelihood function. We use simulations and two empirical data sets to illustrate the performance of regularized SEM and discuss practical aspects of using the method for modeling empirical data. First, though, we outline the general principles of regularization and discuss how to extend these principles to SEM.

Regularization Overview

Regularization in the context of regression

To set the stage for discussing the use of regularization (e.g., shrinkage or penalized estimation) in structural equation models, we give a brief overview in the context of regression. (For more detail, interested readers may consult McNeish, 2015, or Helwig, 2017.) We use ordinary least squares (OLS) estimation as a basis for our discussion. Given N continuous observations of P predictors in matrix X and associated continuous outcome Y, one can estimate the regression coefficients by minimizing the residual sum of squares (RSS) as follows:

$$\text{RSS} = \Sigma_{i=1}^{N} \left(Y_i - \beta_0 - \Sigma_{j=1}^{P} \beta_j X_{ij} \right)^2. \quad (1)$$

For coefficients, one estimates an intercept β_0 along with β_j coefficients (one for each of the P predictors). However, there may be instances when a simpler model—that is, one that includes fewer predictors—is preferred. To select the variables for this simpler model, one can use the lasso (Tibshirani, 1996). Lasso regularization builds upon Equation 1, incorporating a penalty for each parameter (larger parameter values incur a larger penalty):

$$\text{lasso} = \underbrace{\text{RSS}}_{\text{OLS}} + \lambda \Sigma_{j=1}^{P} |\beta_j|. \quad (2)$$

The lasso penalty includes the traditional RSS as in Equation 1, but introduces two new components. First and foremost, it introduces a new penalty term that reflects the sum of all beta coefficients (the right-hand term in Equation 2). In this manner, much as how a traditional regression attempts to minimize the squared residuals, the lasso penalty tries to drive parameters to zero, thus implicitly performing variable selection. Second, as can be seen in Equation 2, the sum of the absolute values of the β_j coefficients is multiplied by a hyperparameter, λ. This term quantifies the influence of the lasso penalty on the overall model fit and thus weights the importance of the least squares fit versus the importance of the lasso penalty: As λ increases, a stronger penalty is incurred for each parameter, which results in greater *shrinkage* of the coefficient sizes. The λ term is called a hyperparameter because it cannot be estimated jointly with the β_j coefficients (this is not the case in Bayesian regularization, which we return to shortly). As there is no generally optimal value for λ, it is common to test a range of λ values, combined with cross-validation, to examine what the most appropriate degree of regularization is for a given data set.

Another type of regularization is ridge regularization (Hoerl & Kennard, 1970). In contrast to the lasso, the ridge sums the *squared* coefficients. Whereas the lasso penalty will push the betas all the way to zero (as any nonzero beta will contribute to the penalty term), the ridge penalty will instead shrink the betas, but not necessarily all the way to 0 (as the squaring operation means that small betas incur negligible penalties). One benefit of ridge regularization is that it better handles multicollinearity among predictors.

In an effort to combine the variable-selection aspects of the lasso with the ridge regularization's ability to handle collinearity, Zou and Hastie (2005) proposed the elastic net. Through the use of a *mixing* parameter, α, the elastic net combines ridge and lasso regularization:

$$\text{elastic net} = \text{RSS} + (1-\alpha)\underbrace{\lambda\Sigma_{j=1}^{P}\beta_j^2}_{\text{ridge}} + \alpha\underbrace{\lambda\Sigma_{j=1}^{P}|\beta_j|}_{\text{lasso}}. \quad (3)$$

Much as different values of λ, combined with cross-validation, can be tested to choose a final model, different values of α can be tested. Generally, the values tested range from zero (equivalent to the ridge penalty) to 1 (equivalent to the lasso penalty).

Extensions

Originating from the application of ridge regression as a way to improve the results of OLS when predictors are correlated (Hoerl & Kennard, 1970), a large number of alternative forms of regularization have been proposed. In the case of high-dimensional research scenarios, sparser versions of the lasso have been proposed. These include the adaptive lasso (Zou, 2006), the smoothly clipped absolute deviation penalty (Fan & Li, 2001), and the minimax concave penalty (Zhang, 2010), to name a few. Methods such as these have been shown to produce more optimal results than the lasso when only a small number of predictors among thousands of candidates or more are desired to have nonzero coefficients. In general, there is no optimal type of regularization, as each type is optimal under different assumptions.

An additional way that regularization methods have been extended is with Bayesian estimation. In Bayesian regression, prior distributions are placed on all the coefficients in the model. When these priors are diffuse (large variances), the observed data have a large influence on the posterior distribution of each parameter. Regularization as applied to Bayesian estimation entails placing different types of prior distributions on those parameters of interest and constraining the width of these priors to shrink the coefficients toward zero. Thus, prior knowledge, as applied through strong priors, carries greater weight in determining the posterior distribution for each parameter when regularization is applied. Placing normal-distribution priors has been shown to be equivalent to ridge regression (Kyung, Gill, Ghosh, & Casella, 2010; T. Park & Casella, 2008; Tibshirani, 1996), whereas the lasso corresponds to Laplace distribution priors (T. Park & Casella, 2008; Tibshirani, 1996). Particularly when variable selection is desired, a number of more advanced forms of Bayesian regularization have been found to perform better than the Bayesian version of the lasso (see van Erp, Oberski, & Mulder, 2018, for an overview).

The Rationale for Regularization

Traditionally, a test statistic (and associated *p* value) is used to determine the significance of a parameter. In regression with regularization, one instead tests a sequence of penalties, compares models to choose a best-fitting model, and examines whether the parameter estimates in this best model are nonzero. Nonzero coefficients can be thought of as *important* (e.g., see Laurin, Boomsma, & Lubke, 2016). This approach stands in stark contrast to the use of *p* values, as using regularization to label parameters as important does not rely on any asymptotic foundations (there are no statements with regard to a population). In particular, because the regularized estimates move away from the point of maximum likelihood, asymptotic distributions of parameter estimates do not hold any more. Commonly paired with cross-validation, regularization attempts to identify which parameters are likely to be nonzero not only in the current sample, but also in a hold-out sample.

Sparsity is an implicit conceptual assumption of regularization methods, such as the lasso, that set parameters to zero (e.g., Hastie, Tibshirani, & Wainwright, 2015). In other words, this approach reflects the hypothesis that the true underlying model has few nonzero parameters. However, in psychological research, this is unlikely to be true. Instead, most variables in a data set likely have small correlations among themselves (e.g., the "crud" factor—Meehl, 1990). As a result, the use of regularization in psychological research will impart some degree of bias into the results; as with all procedures, there is no such thing as a "free lunch" (Wolpert & Macready, 1997). Although this may at first seem to be an undesirable side effect, we argue that there are common situations in which the benefits of reduced variance outweigh the drawbacks of nonzero degrees of bias. First, we provide a brief overview of the bias-variance trade-off.

Bias-variance trade-off

Although regularization is often used when variable selection is desired to achieve a parsimonious level of description, or when the number of predictors is larger than the sample size, one of the fundamental motivations behind regularization concerns the bias-variance trade-off. Bias refers to whether estimates or predictions are, on average (across many random draws from the population), equal to the true values in the population. Variance, on the other hand, refers to the variability, or precision, of these estimates (see Yarkoni & Westfall, 2017, for further discussion). Practically speaking, researchers want bias to be absent and variance to be low (e.g., the Gauss-Markov theorem guarantees that least squares estimation yields unbiased estimates with the lowest variance among all unbiased linear estimators); however, it can be difficult to achieve both goals in practice. Regularization plays a role when one wishes to allow for some bias in order to achieve a larger decrease in variance. When the sample size may be insufficient to adequately test the number of predictors

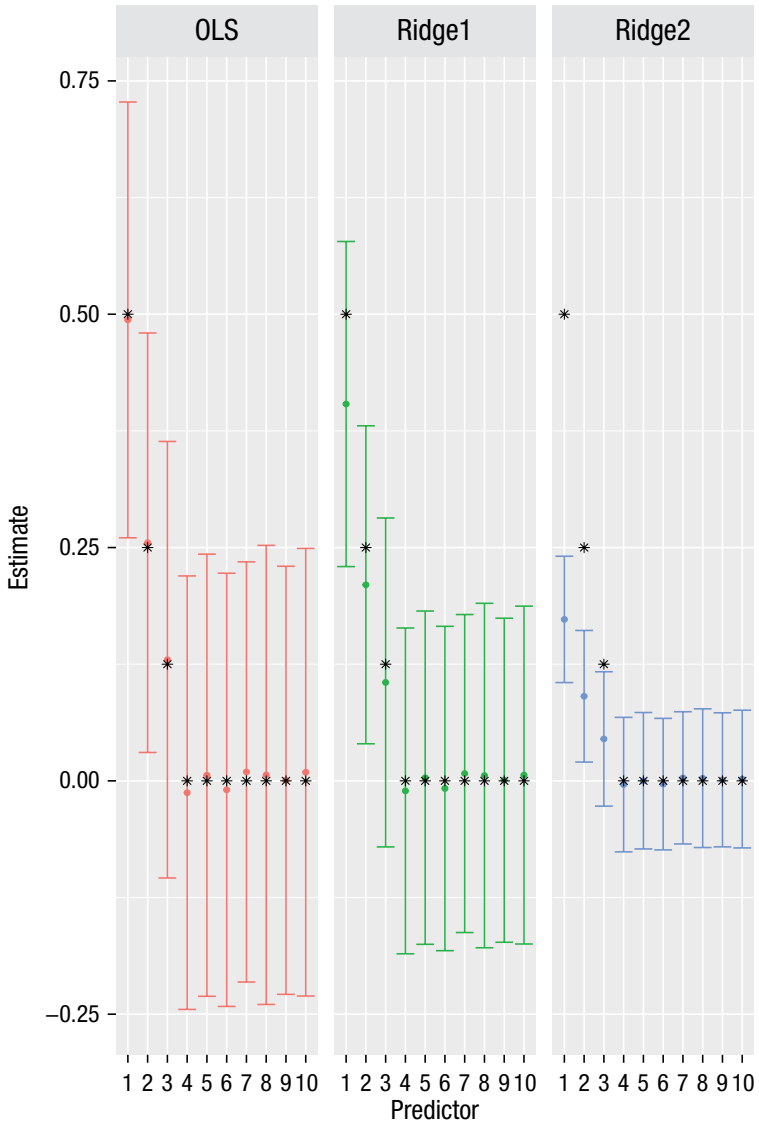

Fig. 1. Illustration of the bias-variance trade-off: parameter estimates from models fit to simulated data. The simulation involved 1,000 repetitions of 30 observations with 10 predictors of a normally distributed outcome variable. Results from an ordinary least squares (OLS) model and two ridge regressions (a weaker penalty of 5 for Ridge1 and a stronger penalty of 50 for Ridge2) are shown. The error bars show ±1 *SD* of the estimate for each parameter. Asterisks denote the simulated parameter estimates.

the researcher desires to include in the model, regularization will systematically bias the regression coefficients toward zero, as the variance of the estimator will be high because of the low sample size. Such an approach will prove particularly beneficial when the true model is sparse (i.e., only few predictors are important).

To provide a simple example, we simulated 30 observations with 10 predictors of a normally distributed outcome variable. This 3-to-1 (30-to-10) ratio is far below recommended guidelines for regression models. Across 1,000 repetitions, the first predictor was simulated to have the strongest regression coefficient (.5), the second predictor was simulated to be half as strong (.25), and the third predictor was simulated to be half as strong as the second (.125). The other 7 predictors had simulated coefficients of zero. The resultant coefficients from both OLS and ridge regression models are displayed in Figure 1.

The figure shows that the estimates from the OLS model are unbiased (i.e., the mean parameter estimates correspond with the simulated parameter estimates). However, the absence of bias comes at the expense of variance, as the OLS coefficients have a large degree of variability. This is to be expected given the results

of methodological work on sample sizes in linear regression (Green, 1991). However, instead of restricting the number of predictors entered into the model in order to address a small fixed sample size (e.g., testing only 2 predictors when testing all 10 is desired), researchers can use regularization to impart bias as a mechanism to decrease the variance of the estimates. In contrast to the OLS results, the ridge parameter estimates are biased toward zero (i.e., they are lower than the means in the data-generating mechanism), and this bias is greater when the penalty is larger. Higher regularization imparts more bias toward zero, while also reducing the variance of the parameter estimates. Particularly when the sample size is small or the number of variables is large (compared with the sample size), this is a desirable property of regularization.

Rationale for accepting bias to reduce variance

Even though there may be a confluence of small effects in a data set, researchers may not value including every nonzero parameter into the model, as it complicates estimation and renders interpretation difficult. In such cases, researchers care more about what could be termed *functional* sparsity. They specifically want to develop a parsimonious model that facilitates interpretation and generalization of the most important parameters.

One of the main motivations for developing regularization methods is for use with data sets that have more variables than observations. In such cases, OLS regression cannot be used. Although settings in which the number of parameters exceeds N may still be uncommon, the benefits generalize to settings in which the ratio of observations to predictors is small, which can be said to pose a sample-size challenge (e.g., Bakker, Van Dijk, & Wicherts, 2012). Adequate power to detect a given parameter requires a suitably large sample size (depending on the magnitude of the effect), and when multiple effects are considered, either separately or in the context of a multivariate model, the required sample size can increase rapidly. If a sample size is small for practical or principled reasons, one strategy for testing a complex model is to reduce the dimensionality of the model. Most commonly this means using some method, such as stepwise regression, to reduce the number of coefficients in a regression model, which can be highly problematic (e.g., Harrell, 2015).

Regularization in Structural Equation Modeling

In psychological research, it is common to have more than one outcome of interest, each specified as a latent variable. Usually, researchers want to model not only latent variables, but also predictors of these factors. One strategy is to estimate factor scores in a confirmatory factor analysis (CFA), extract the factor estimates, and treat those as outcomes in a traditional OLS regression. However, this can be problematic (e.g., Devlieger & Rosseel, 2017; Grice, 2001), inducing issues such as biased estimates of the regression parameters and factor-score indeterminacy. In contrast, one can stay within the latent-variable framework and include predictors of all outcomes of interest in a single analysis. This allows for richer analysis; for example, one can test the equality of relationships across time, assess fit (through various fit indices), and test for directed relationships between latent variables. Pairing regularization with a multivariate model of this type requires a generalization of the types of univariate regularization methods we discussed earlier.

Regularization has been extended in a number of directions beyond linear regression. These extensions have been applied to, for example, generalized linear models (e.g., M. Y. Park & Hastie, 2007), network-based models (e.g., Epskamp, Rhemtulla, & Borsboom, 2017), item-response-theory models (Chen, Li, Liu, & Ying, 2018; Sun, Chen, Liu, Ying, & Xin, 2016), differential item functioning (Magis, Tuerlinckx, & De Boeck, 2015; Tutz & Schauberger, 2015), educational assessment (Culpepper & Park, 2017), and factor analysis (e.g., Hirose & Yamamoto, 2015), to name just a few. Specific to our purposes is what we refer to as regularized SEM, or RegSEM (Jacobucci, Grimm, & McArdle, 2016; see also Huang, Chen, & Weng, 2017).

RegSEM directly builds different types of regularization into the estimation of structural equation models, by expanding the traditional maximum likelihood estimation (MLE) to include a penalty term, as follows:

$$F_{\text{RegSEM}} = \underbrace{\log(|\Sigma|) + \text{tr}(\mathbf{C} * \Sigma^{-1}) - \log(|\mathbf{C}|) - P}_{\text{MLE}} + \underbrace{\lambda P(\cdot)}_{\text{penalty}}, \quad (4)$$

where $\mathbf{C}$ is the sample covariance matrix, Σ is the model implied covariance, matrix, and P is the number of variables. This adds a penalty term, $\lambda P(\cdot)$, to the traditional MLE fit function. Just as in regularized regression, λ is the penalty; $P(\cdot)$ is a general function for summing parameters. In the case of the lasso, $P(\cdot)$ sums the absolute values of the specific parameter estimates. The same goal is accomplished for ridge penalties, the elastic net, and other extensions (see Jacobucci, 2017). An important step is selecting which parameter estimates should be included (i.e., which parameters are penalized) in $P(\cdot)$. Because this form of regularization takes place in the estimation of structural equation models, regularization can be selectively applied to one or more subsets of parameters, including factor loadings (e.g.,

to select items from a questionnaire to create a short form), variances or covariances (e.g., to test whether the addition of residual covariances is necessary), or—of specific interest to us here—regression paths.[1] For each of these penalized parameters in the model, it is important to standardize the corresponding variables prior to the analysis. By standardizing the variables, one ensures that all penalized parameters contribute equally to model fit.

When the type of regularization is either the lasso or the elastic net (or another sparse penalty), the number of effective degrees of freedom can change as the penalty increases. Most notably, as the penalty increases, each parameter that is set to zero increases the degrees of freedom (see Jacobucci et al., 2016, for additional information). Thus, increasing the penalty often results in an improvement in fit as assessed by those fit indices that include the number of parameters in the equation (e.g., root mean square error of approximation, or RMSEA; comparative-fit index, or CFI; and information criteria). Note, however, that some fit indices are derived under the assumption that the point estimate is maximum likelihood; thus, it may be preferable to evaluate prediction error in a test set rather than to use classic in-sample test statistics (see Yarkoni & Westfall, 2017).

RegSEM combines confirmatory aspects of SEM with an exploratory search for important predictors. The confirmatory and exploratory aspects can take place in either the measurement or the structural parts of a structural equation model. In many situations, researchers may have some a priori idea of how some variables relate to each other. For instance, imagine that a group of researchers have constructed an initial CFA model with five indicators of a single latent variable, such as fluid intelligence. This confirmatory formulation may be based on previous research support for a single latent dimension underlying the covariance among the five indicators. Figure 2 displays the addition of three predictors (say, volumetric measures of different brain regions; cf. Kievit et al., 2014) to the initial CFA model. The resulting model is called a multiple-indicators (factor loadings), multiple-causes (regression parameters directed to the latent variable) model (MIMIC model; Jöreskog & Goldberger, 1975). Once the model is run, traditional techniques, such as the Wald test (and associated test statistics), can be used to determine which predictors have nonzero population values.

This kind of model is commonly used to simultaneously estimate the joint influence of a set of presumed causal influences on one or more latent variables. However, given the constraints of traditional SEM approaches, the predictors are usually selected a priori on the basis of theoretical or empirical considerations (cf. Kievit et al., 2014). Now imagine an alternative scenario in

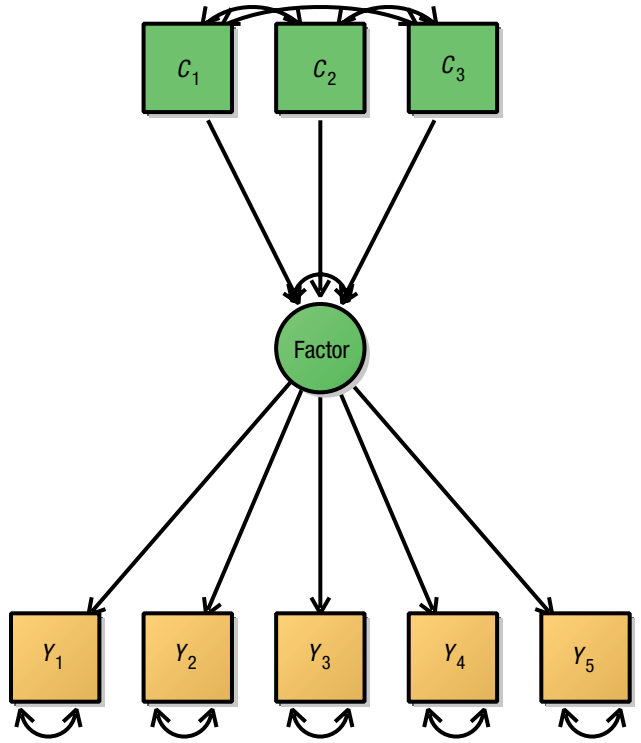

Fig. 2. A simple multiple-indicators, multiple causes (MIMIC) model with five indicators (yellow boxes) and three predictors (green boxes) of a latent factor. Note that the variances of C_1, C_2, and C_3 have been omitted.

which the researchers have a much larger number of predictors they wish to test (e.g., gray-matter volume in all regions identified in an atlas). None of these additional relationships may be based on previous hypotheses. The researchers may be relatively uncertain about which covariates in their data set are important predictors of the fluid-intelligence latent factor, either because they do not have strong a priori expectations or because there are a large number of candidates (e.g., genetic markers, brain variables). In this case, an *exploratory search* would be conducted. Traditional tools are no longer as suitable in such a scenario, as the model may not converge, or estimates may be imprecise, because of problems using MLE with large numbers of variables when the sample size is limited (e.g., see Hastie et al., 2015). Although previous research has examined the influence of large models on test statistics (Yuan, Yang, & Jiang, 2017), less attention has been paid to strategies that produce more accurate parameter estimates in large models. Here, in an effort to reduce this gap in the literature, we propose and evaluate the use of regularization.

In many applied fields such as genetics, cognitive neuroscience, and epidemiology, the ratio of predictors to the available sample size may be large. Indeed, one could argue that the absence of regularization methods

may help explain why fields such as cognitive neuroscience rely on mass univariate approaches (i.e., a relationship between an outcome and neural data is tested thousands of times, separately for each brain region or voxel). However, multivariate approaches generally paint a richer, more realistic picture of the true data structure, and also allow the researcher to investigate which effects are redundant, and which may be partially independent and complementary. To examine the possible benefits of regularization in the SEM context, we conducted three studies. In Study 1, we examined the effectiveness of both MLE and regularization in the context of complex structural equation models. In Studies 2 and 3, we applied regularized SEM to large existing data sets.

Disclosures

The scripts for the simulation and applied analyses in this article are available online at https://osf.io/z2dtq/.

Study 1: Simulation

Method

To evaluate the effectiveness of the RegSEM lasso, we designed simulation conditions that researchers may commonly face when evaluating a large number of predictors (e.g., a property such as cortical thickness measured across many brain regions). We varied our simulations across two dimensions: sample size and predictor collinearity. The template model for our simulation is depicted in Figure 3. The model included six indicators (Y_1–Y_6) of the latent variable, *f*. The factor loadings of these indicators differed in their simulated population values (see Fig. 3). As predictors of *f*, there were 70 uninformative (noise) variables (C_{n1}–C_{n70}), with simulated population coefficients of zero. Additionally, there were three sets of 10 predictors each; a set of predictors with small effect sizes (0.20, C_{s1}–C_{s10}), a set of predictors with medium effect sizes (0.50; C_{m1}–C_{m10}), and a set of predictors with large effect sizes (0.80; C_{l1}–C_{l10}). Taken together, there were 100 potential predictors of *f*, each treated as a fixed effect. The variance of the latent variable was fixed to 1 for identification purposes, so that each factor loading could be freely estimated (we did not estimate a mean structure).

After creating simulated data according to the model in Figure 3, we tested a model that included 112 free parameters: 100 regression coefficients, 6 factor loadings, and 6 residual variances. Although rules of thumb are inherently limited, common guidelines suggest a ratio of 10:1 between sample size and the number of free parameters (e.g., Kline, 2015) to obtain stable estimates. In this case, that would mean a minimum *N* of

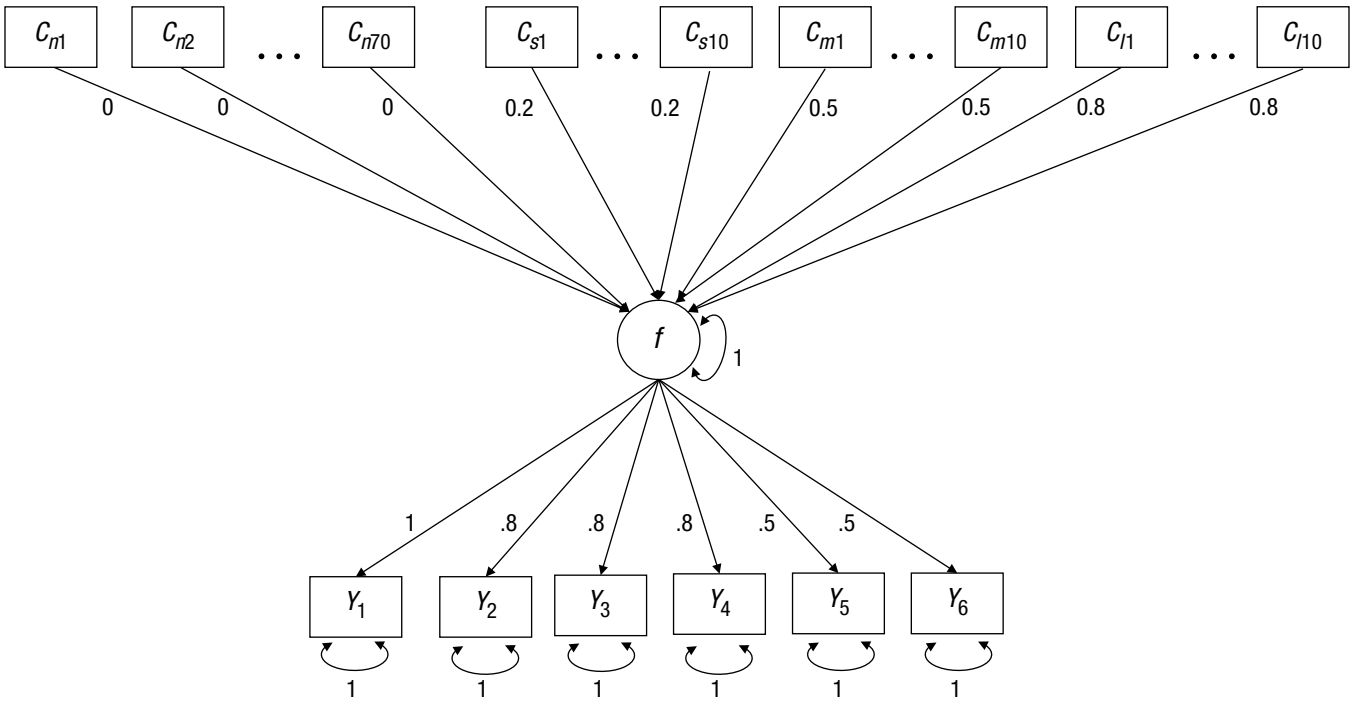

Fig. 3. Template model for the simulation in Study 1. This multiple-indicators, multiple-causes (MIMIC) model included a single latent factor *f*, six indicators (Y_1–Y_6) with factor loadings from .5 to 1 and unique error variances, and 100 potential predictors. Some predictors were uninformative (C_{n1}–C_{n70}), and others had a small effect (C_{s1}–C_{s10}), a moderate effect (C_{m1}–C_{m10}), or a strong effect (C_{l1}–C_{l10}).

1,200. Given that many researchers may wish to test models of this size, but may not have the requisite sample size, we wanted to test a variety of sample sizes to examine when the performance of MLE degrades and when the use of regularization is beneficial. Therefore, we tested sample sizes[2] of 150, 250, 350, 500, 800, and 2,000.

In most psychological studies that examine the influence of a variety of predictors, these predictors have correlations among themselves. This complicates the interpretation of the results. For instance, it becomes challenging to determine the relative contribution of individual predictors (Grömping, 2009). Moreover, high degrees of collinearity can result in problematic estimation. Therefore, we also investigated the effect of predictor collinearity in our simulation by simulating data with correlations of 0, .20, .50, .80, and .95 among all predictors. We expected that bias in both MLE and regularized estimation would increase as the correlation among predictors increased. Because lasso regularization is problematic with high degrees of collinearity, we also tested the elastic-net estimator. Finally, we examined the prevalence of Type I errors (wrongly including a noise predictor in the final model) and Type II errors (wrongly excluding a true predictor) across the sample sizes and effect sizes.

To test each form of estimation, we used a different package in the R statistical environment (R Core Team, 2018). For MLE, we used the lavaan package (Version 0.5-23.1097; Rosseel, 2012). For RegSEM, we used the regsem package (Version 1.0.6; Jacobucci, Grimm, Brandmaier, & Serang, 2017). Both lasso and elastic-net regularization are implemented in regsem, along with a host of additional penalties (Jacobucci, 2017). We varied λ (see Equation 4) across 30 values, ranging from 0 to 0.29 in equal increments. In initial preruns, higher penalty values were included, but they always resulted in worse fit. To choose a final model among the 30 models run, we used the Bayesian information criterion (BIC; Schwarz, 1978). Each cell in the simulation's design was replicated 200 times.

Results

Instead of giving a detailed analysis of the simulation's results, we provide a high-level overview. We compare the performance of the RegSEM lasso with the performance of MLE using three metrics: root mean square error (RMSE; averaged across each set of parameters), relative bias (averaged across each set after taking the absolute value of each parameter), and error rate (for Type I and Type II errors, respectively). For each performance metric, we discuss how results varied across sample sizes and predictor collinearities. We do not present the results for RegSEM elastic-net estimation, as those results were almost identical to the results for the RegSEM lasso.

Parameter estimates. First we discuss the precision of parameter recovery, quantified as RMSE (Fig. 4) and relative bias (Fig. 5). At high sample sizes, the results for RMSE showed that MLE performed similarly to the lasso, and MLE's performance was better than the lasso's when performance was measured by relative bias. This difference in results for the two metrics was expected, because, as we discussed earlier, the lasso imparts bias to reduce variance. RMSE measures both bias and variance, whereas relative bias measures only bias; thus, with the lasso, the increase in bias is somewhat offset by a decrease in variance. At small sample sizes, and particularly a sample size of 150, the lasso performed better than MLE with regard to both RMSE and relative bias. In conditions with only 150 observations, MLE was highly unstable in its estimation of parameters; that is, parameter estimates were drastically larger than their simulated values.

The lasso produced better RMSE results than MLE in most conditions, particularly with sample sizes of 150 and 250. When the correlation among all predictors was extremely high (.95), the lasso produced a large amount of RMSE in the factor loadings. This condition is probably the reason why the RMSE values for the estimates of the factor loadings were higher for the lasso than for MLE across sample sizes. This poor performance can most likely be explained by covariance expectations and by the fact that correlations among predictors create a complicated web of relationships (see the appendix for further details). Fortunately, collinearity of predictors in the range of .95 is unlikely to be observed in real data sets.

The results for relative bias were more mixed. The lasso produced less relative bias than MLE for the sample size of 150, but MLE produced less relative bias with larger samples. An extreme degree of collinearity resulted in a large increase in relative bias for the lasso, much as extreme collinearity increased RMSE for the lasso. This secondary effect of collinearity was much less evident with MLE. Collinearity had a U-shaped effect on relative bias of the regression coefficients when the lasso was used: Both small (.00) and extreme (.95) correlations among predictors resulted in the highest relative bias. In contrast, this relationship did not hold for MLE. Together, our simulations show that regularized SEM outperforms traditional MLE in accuracy of parameter estimation when sample sizes are small and the number of predictors is large.

Type I and Type II errors. An alpha criterion of .05 was used to determine parameter significance in the MLE

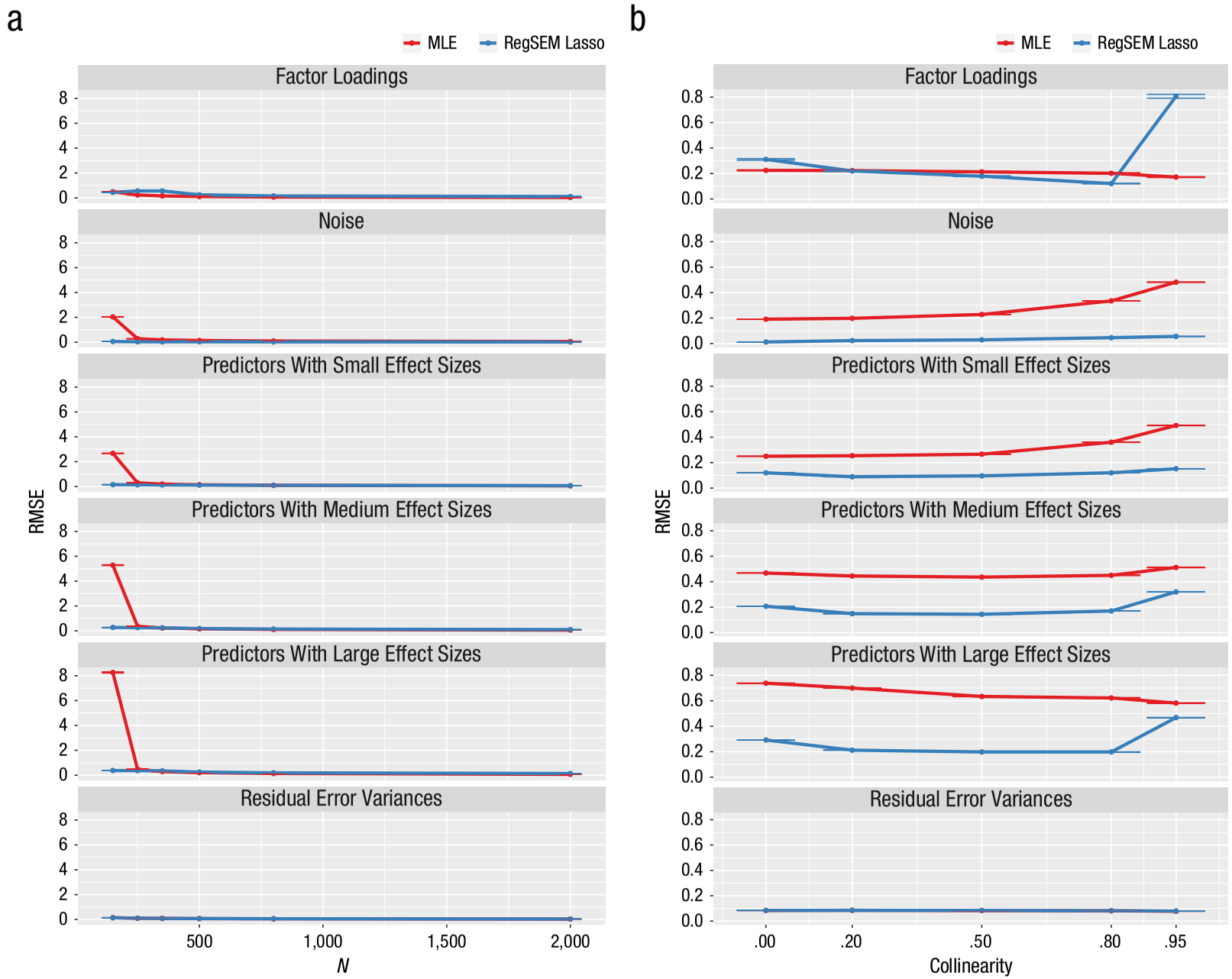

Fig. 4. Results from Study 1: root mean square error (RMSE) as a function of (a) sample size and (b) predictor collinearity. Results are shown separately for maximum likelihood estimation (MLE) and the RegSEM lasso. From top to bottom, the graphs show results for the factor loadings, the uninformative predictors (noise), the informative predictors of different effect sizes (small, medium, and large), and the residual error variances. Error bars represent ±1 Monte Carlo standard error. Note that some of the error bars are barely visible because of very small values.

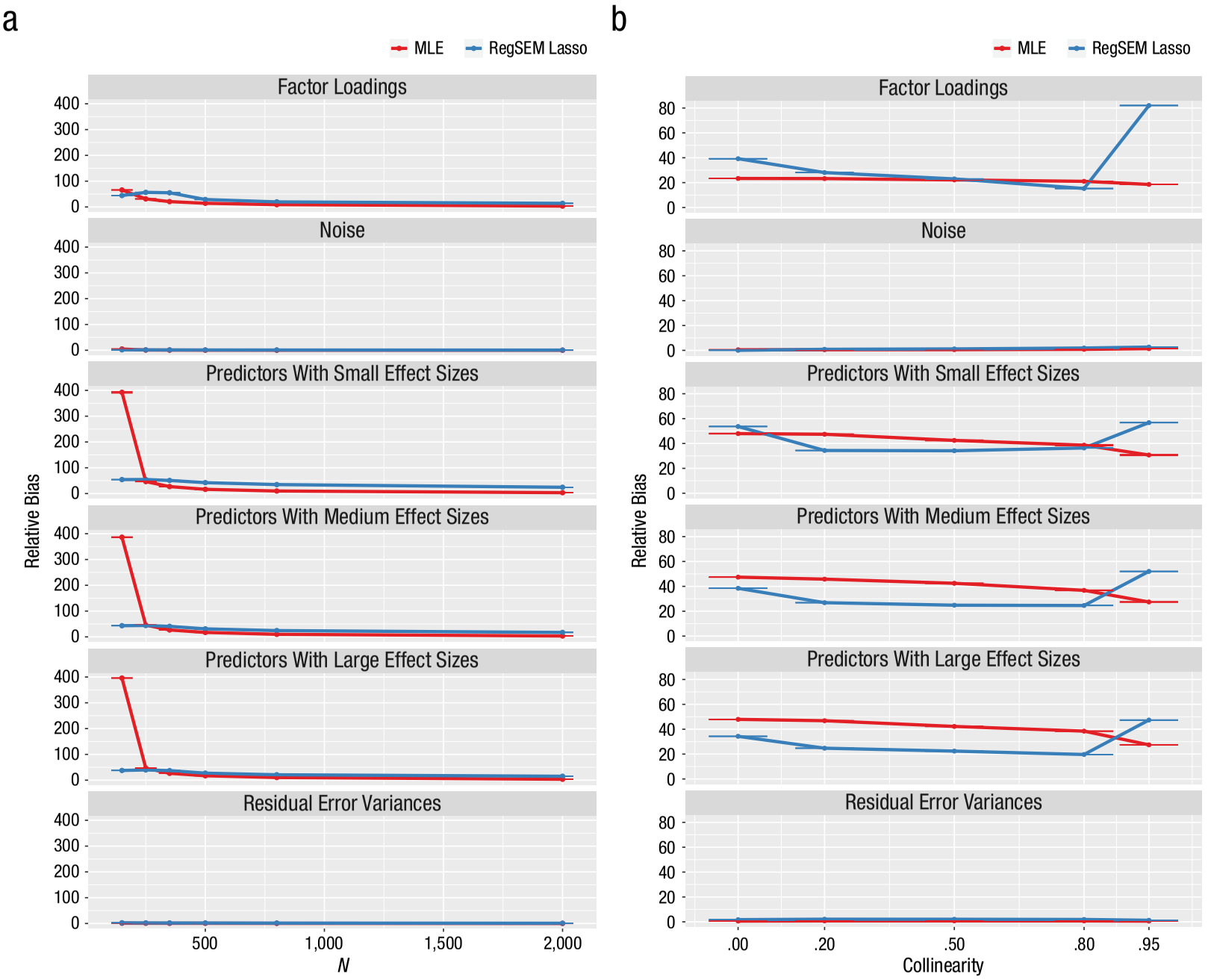

Fig. 5. Results from Study 1: relative bias as a function of (a) sample size and (b) predictor collinearity. Results are shown separately for maximum likelihood estimation (MLE) and the RegSEM lasso. From top to bottom, the graphs show results for the factor loadings, the uninformative predictors (noise), the informative predictors of different effect sizes (small, medium, and large), and the residual error variances. Error bars represent ±1 Monte Carlo standard error. Note that some of the error bars are barely visible because of very small values.

models. Thus, a Type I error occurred when the coefficient for a noise variable had a p value less than .05. Results for the frequency of Type I errors (see Fig. 6) showed that sample size had a larger effect than did collinearity in the MLE models. For a sample size of 150, there was a 17% chance for a noise variable to be incorrectly identified as a significant parameter. Although the Type I error rates were higher than .05 across the range of collinearities tested, this can mostly be attributed to the influence of sample size. A Type II error occurred when the coefficient for a predictor with a small, medium, or large simulated effect size had a p value greater than .05. Results for Type II errors (see Fig. 6) were alarming in that they indicated there was low power to detect the effects of the parameters with small and medium simulated effect sizes. As collinearity increased, so did the Type II error rates for these parameters (i.e., power decreased); the inverse relationship held for sample size. Even for the parameters with a simulated effect size of 0.8, larger-than-expected numbers of Type II errors were committed when sample sizes were small and when collinearity was high.

Overall, the lasso models committed far more Type I errors than the MLE models, but also had much lower Type II error rates (i.e., they rarely omitted a truly predictive variable).

Summary. Across our simulations, the MLE models estimated parameters more accurately than the lasso models when sample sizes were large, and the lasso models were more accurate than the MLE models when sample sizes were small. Overall, the MLE models had less relative bias (as expected), but the lasso models improved upon the MLE models with respect to the RMSE. The RegSEM lasso committed more Type I errors then MLE did, but achieved much higher rates of power (lower Type II error rates) across conditions. These results are in line with previous findings, such as those of Serang, Jacobucci, Brimhall, and Grimm (2017), who found a similar trade-off between regularization and other forms of estimation in the context of mediation models. The optimal method for a given research context depends on the relative importance of decreasing parameter bias and decreasing parameter variance. Within a given sample, an MLE model may produce more accurate results than a model using regularization, but the MLE model may not generalize as well. We note that the contrast between these forms of estimation may not apply beyond the small selection of models we discuss here.

Study 2: White-Matter Determinants of Visual Short-Term Memory

Many features of brain structure and function may have complementary cognitive effects. Thus, a challenge in cognitive neuroscience is how to reconcile the dimensionality constraints of covariance-based methodologies such as SEM with the richness of the imaging metrics (which may include hundreds of measures per individual). Here, we describe an illustrative example in which we used regularized SEM to model visual short-term memory (VSTM) as a function of white-matter microstructure. The data came from the Cambridge Study of Cognition, Aging and Neuroscience (Cam-CAN, www.cam-can.org; Shafto et al., 2014), a study of a large, population-derived cohort of healthy individuals

The Sample

The data for this empirical illustration are from 627 adults (320 female; age range from 18 to 88, $M = 54.18$, $SD = 18.42$) who participated in a large battery of cognitive tests, demographic and life-style measurements, and MRI scans (for more details on the cohort and sampling methodology, see Taylor et al., 2017). We focus on participants who had complete data for a specific cognitive task (the VSTM task) and a common index of white-matter microstructure (fractional anisotropy). Analyses of subsets of these MRI data (but not the data for this cognitive task) have previously been reported (e.g., Henson et al., 2016; Kievit et al., 2016; Kievit et al., 2014).

The VSTM task

The VSTM task in the Cam-CAN battery was developed to quantify the capacity and precision of VSTM. The task consists of three phases: an encoding phase, during which participants view one to four colored circles (targets); a brief blank screen (900 ms); and a cue in the same spatial location as one of the target circles (see Fig. 7). Participants are asked to use a color wheel to pick the color of the circle that previously appeared at the cued location, as well as to rate their confidence in their judgment. They performed a total of 224 trials across two blocks; position of the targets, set size, and cue (i.e., which target was cued) were counterbalanced across blocks. We focus here on the effects of set size for set sizes 2 through 4 (to avoid the ceiling effects associated with the simplest version of the task). Each participant's mean performance for each of these three set sizes was scored; the scores for each set size ranged between 0 and the maximum number of circles for that set size.

White-matter predictors

For the neural indicators, we used a common metric of white-matter organization called fractional anisotropy. This metric quantifies the dispersion of water molecules

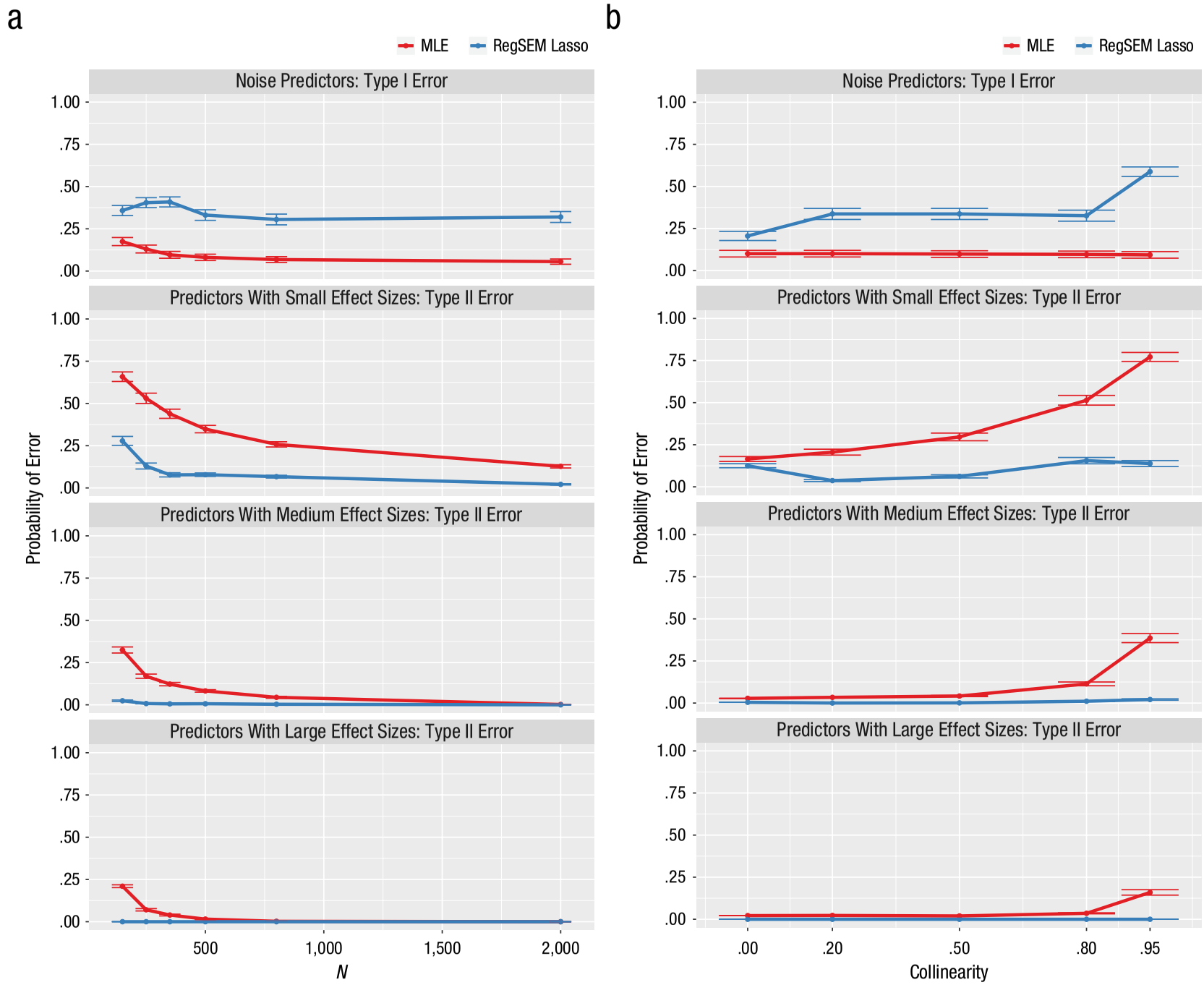

Fig. 6. Results from Study 1: probability of committing Type I and Type II errors as a function of (a) sample size and (b) predictor collinearity. Results are shown separately for maximum likelihood estimation (MLE) and the RegSEM lasso. From top to bottom, the graphs show results for noise predictors and predictors with small, medium, and large effect sizes. Error bars represent ±1 Monte Carlo standard error. Note that some of the error bars are barely visible because of very small values.

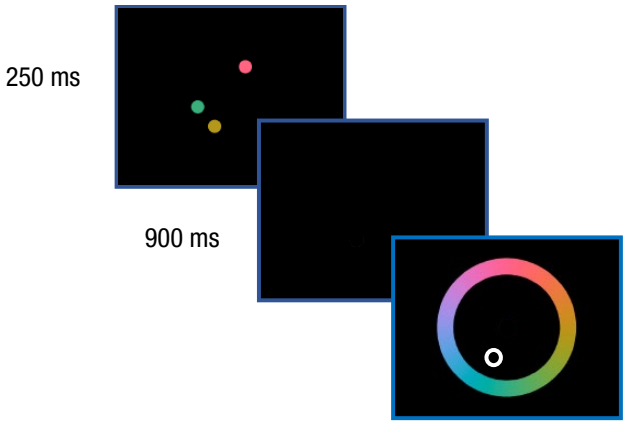

Fig. 7. The visual short-term memory task in the Cambridge Study of Cognition, Aging and Neuroscience (Shafto et al., 2014). On each trial, participants viewed from one to four target circles for 250 ms and then a 900-ms blank screen. Finally, a cue for one of the previous targets was presented, and participants were asked to use a color wheel to indicate which hue most closely matched that of the cued target.

and the extent to which this dispersion is constrained by the organization of white-matter structures. Fractional anisotropy is a complex and indirect measure with various limitations, and its relationship to white-matter health is not yet fully understood (Bender, Prindle, Brandmaier, & Raz, 2016; Jones, Knösche, & Turner, 2013). Nonetheless, fractional anisotropy is widely used, as it has been shown to be associated with individual differences in a range of cognitive domains, especially in old age (Madden et al., 2009). We focused on mean fractional anisotropy for each tract in the ICBM-DTI-81 atlas (Mori et al., 2008), which parcellates the human white-matter skeleton into 48 tracts. Although our previous work used white-matter atlases of lower dimensionality (e.g., Kievit et al., 2016, and de Mooij, Henson, Waldorp, & Kievit, 2018), we intentionally used a more high-dimensional white-matter atlas for this example to illustrate the benefit of regularization. (For more details regarding the analysis pipeline, see Kievit et al., 2016.)

The MIMIC model

To examine the neural determinants of VSTM, we fit a MIMIC model (Jöreskog & Goldberger, 1975). Our model captured the hypothesis that VSTM (the latent variable), which was measured by scores on the VSTM task (the multiple indicators), was in turn affected by the fractional anisotropy of various white-matter tracts (the multiple causes; see Kievit et al., 2012, for a comparison of the MIMIC model with competing representations). First, we specified a measurement model in which the latent VSTM variable was measured by memory scores at each of the three set sizes (2, 3 and 4). Next, we simultaneously regressed this latent variable on the fractional anisotropy for all 48 white-matter tracts. This model tested the joint prediction of the latent variable by all 48 white-matter tracts, which allowed us to determine whether one or more tracts helped predict individual differences in VSTM.

Model estimation and results

We estimated the regularized model across a range of λ values and used the BIC (also known as the Schwarz criterion) to compare the fit of these models (see Jacobucci, Grimm, & McArdle, 2016, for details on alternative strategies for selecting a final model). The BIC balances the increased parsimony achieved by regularizing parameters to zero with the concurrent decrease in explanatory power. As we had a strong a priori hypothesis about the measurement model, we regularized the structural parameters (i.e., the joint prediction of the latent variable by the 48 tracts), but not the factor loadings or residual variances. As Figure 8 shows, the best solution was obtained with a λ value of 0.18. This model yielded an acceptable RMSEA of 0.0321. Figure 9 shows the beta estimates across a range of λ values, as well as the location of the 6 tracts with nonzero beta estimates in the final model.

As Figure 9 shows, the beta estimates for six tracts remained nonzero in the regularized MIMIC model. Strikingly, three of these tracts are subdivisions of the fornix (the column body, as well as the cres), and all

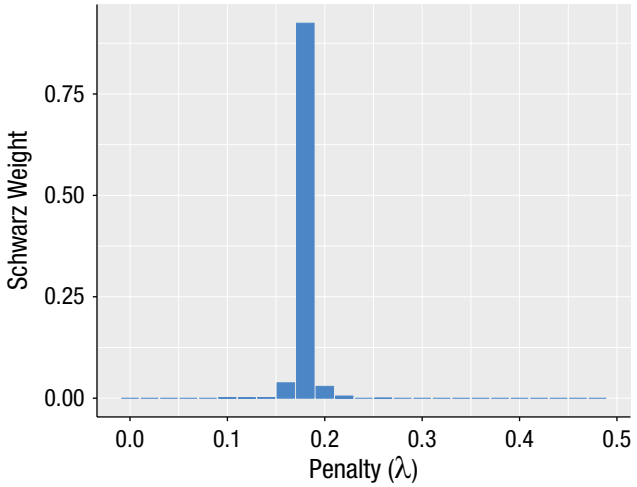

Fig. 8. Results from Study 2: Schwarz weights (cf. Wagenmakers & Farrell, 2004) of the models as a function of the penalty (λ) value. Higher weights correspond to lower values of the Bayesian information criterion and thus indicate a better-fitting model.

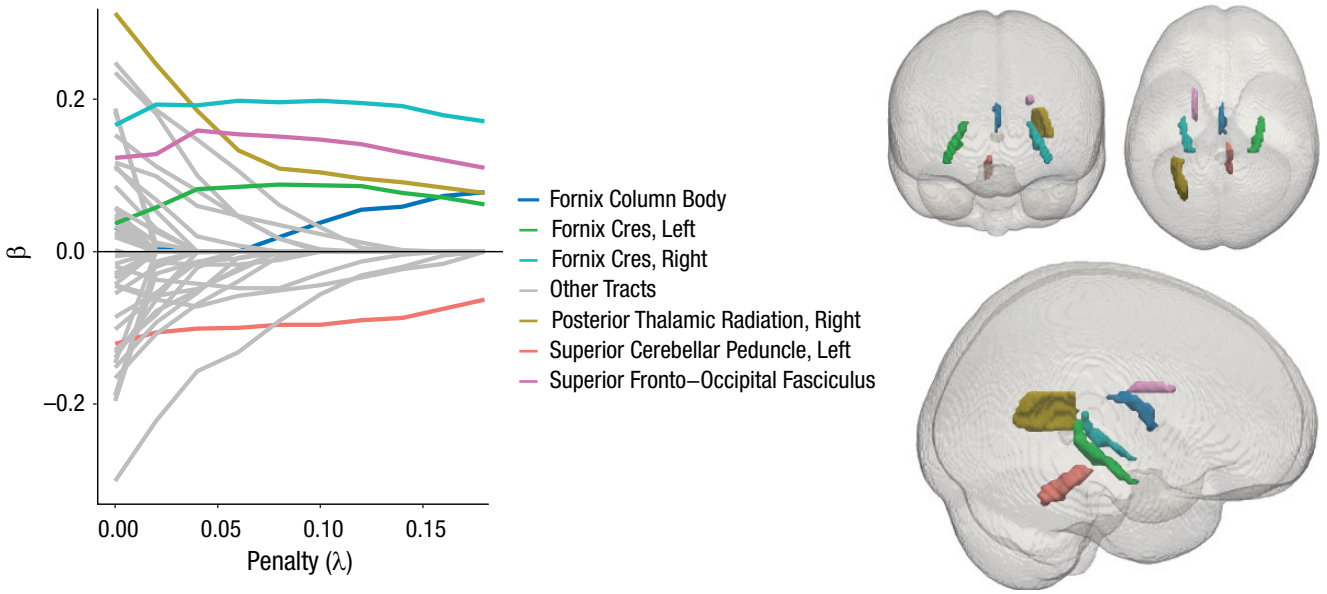

Fig. 9. Results from Study 2: beta estimates for the white-matter tracts as a function of the penalty (λ) value and brain maps showing the location of the six tracts with nonzero estimates in the final model. Note that the tracts with effects regularized to zero are shown in gray.

showed positive effects (i.e., greater white-matter microstructure was associated with better VSTM performance). The fornix, which connects the hippocampus to other brain regions, has long been associated with various aspects of memory, usually autobiographic memory (e.g., Hodgetts et al., 2017) but also, in the Cam-CAN cohort, subdomains such as recollection, familiarity, and priming (Henson et al., 2016). Notably, there have been some Phase I trials suggesting that deep-brain stimulation to the fornix may alleviate memory complaints in patients with early Alzheimer's disease (Laxton et al., 2010). The posterior thalamic radiations (see Fig. 9) have been posited as crucial for focusing and allocating attention in demanding tasks (Menegaux et al., 2017). Evidence from infants suggests an association between greater white-matter organization in the posterior thalamic radiations and better performance on the VSTM task (Menegaux et al., 2017). Finally, we observed a positive association between VSTM performance and white-matter microstructure of the superior fronto-occipital fasciculus, which was positively associated with children's spatial working memory in previous research (Vestergaard et al., 2011).

Although the results for these five tracts align well with previous literature, we also observed a single (surprising) negative effect: Greater white-matter integrity of the superior cerebellar peduncle was associated with poorer VSTM performance. However, closer inspection suggested that this pattern was likely an artifact of image registration, as the integrity of this tract, unlike others, *increases* (bilaterally) with age. A likely explanation is that the relatively deep location of this tract within the brain makes it vulnerable to registration challenges, such as partial volume effects (Alexander, Hasan, Lazar, Tsuruda, & Parker, 2001). We suggest that this negative pattern is more likely due to an imaging artifact than to a true association.

It should be noted that the regularized model solution does not imply that the white-matter microstructure of all the other tracts is uncorrelated to VSTM. When predictors are collinear, the predictors that emerge in a regularized solution are likely to be the most representative of broader sets of correlated predictors (in the present case, a single tract may capture most or all of the predictive power across a network of tracts). In the case of collinear predictors, regularizing groups of predictors with the *group lasso* (Friedman, Hastie, & Tibshirani, 2010) may be more appropriate than regularizing individual predictors; however, this approach has not yet been generalized to SEM.

To summarize, a regularized structural equation/MIMIC model was able to model the relation between cognitive performance and imaging metrics, taking a high-dimensional set of predictors and reducing this set to create a relatively parsimonious representation of key tracts previously implicated in VSTM performance. These results demonstrate the viability of this methodology in cognitive neuroscience in general and in research with aging and developmental cohorts in particular.

Study 3: Modeling the Determinants of Depression, Anxiety, and Stress

The Sample

Previous work has suggested that there are many distinct predictors of individual differences in depression, anxiety, and stress (e.g., Sümer, Poyrazli, & Grahame, 2008), but the extent to which these determinants of mental health are separable or collinear (nonunique) is unclear. For our second empirical example, we attempted to answer this question using a large (N = 27,835) publicly available data set containing answers to the Depression Anxiety Stress Scales (DASS; Lovibond & Lovibond, 1995). This data set was collected from an online sample and is freely available at https://openpsychometrics.org/_rawdata/. The 42-item DASS captures latent variables of depression, anxiety, and stress (each with 14 indicators), and the data set also includes a set of personality and demographic covariates, which we subjected to regularization. These covariates include responses on the Ten-Item Personality inventory (TIPI; Gosling, Rentfrow, & Swann, 2003), which uses 7-point Likert scales (from *disagree strongly* to *agree strongly*). Other covariates are highest level of education (ranging from 1, *less than high school*, to 4, *graduate degree*), gender (1 = male, 2 = female), age (in years), handedness (1 = right, 2 = left), voter record (1 = *I have voted in the last year*, 2 = *I have not voted in the last year*), and family size ("Including you, how many children did your mother have"; participants filled in the correct number). These covariates are included for illustrative, rather than conceptual, reasons.

Results

First, we fit a three-factor measurement model to all the DASS data. This model fit the data well, $\chi^2(816) = 64{,}490.79$, $p < .001$; RMSEA = 0.053, 90% confidence interval = [0.053, 0.053]; CFI = 0.897; standardized root mean square residual = .040, and all factor loadings were moderate to strong (range = .50–.84, M = .70). Despite considerable covariance among the latent variables (all correlations > .7), a three-factor model fit considerably better than a competing unidimensional account (in which all items were taken to measure a single latent variable), $\Delta\chi^2 = 29{,}414$, $\Delta df = 3$, $p < .001$. Next, we fit a MIMIC model in which the three latent factors were simultaneously regressed on the 16 predictors. Results reported here are based on a random subsample (N = 1,000) of the full cohort. Model fit was good, $\chi^2(1440) = 4{,}348.61$, $p < .001$; RMSEA = 0.045, 90% confidence interval = [0.044, 0.046]; CFI = 0.884; standardized root mean square residual = .038, and the joint covariates predicted a large amount of variance (depression: 40.7%; anxiety: 41.9%; stress: 52.3%). In this same subsample, MLE yielded 4 predictors that were nominally significant for stress, 7 that were significant for anxiety, and 6 that were significant for depression (see Fig. 10).

Next, we refit the model using the RegSEM lasso. Forty penalty values were tested, and Figure 11 summarizes the results. The optimal BIC solution was observed with a λ value of 0.15. This penalty regularized the paths of 27 of the 48 structural parameters to zero, yielding the most parsimonious model representation. Table 1 shows the fully standardized parameter estimates for the MLE model as well as the regularized model. Across all three factors, two TIPI items—4 ("easily upset, anxious") and 9 ("calm, emotionally stable")—had strong associations (r = ~.3) with the three mental-health outcomes. Note that the strong associations between the latent factor anxiety and TIPI Item 4 is unsurprising given their partial content overlap. However, both the MLE and the lasso estimates demonstrate that a considerable number of additional predictors, including education and "reserved, quiet" (TIPI Item 6) personality, explained unique variance in individual differences in mental health. Moreover, the DASS and TIPI purportedly capture distinct domains (mental health and personality), and this model is meant for illustrative purposes only. We present these parameters for readers to interpret accordingly.

Note that the predictors with the largest Wald-test z values in Table 1 do not consistently correspond to the predictors selected as having nonzero coefficients in the lasso model. One thing to keep in mind when interpreting lasso parameter estimates is that they are biased toward zero because of the shrinkage (Tibshirani, 1996). To address this bias, one can refit the model without any penalty in a second stage that includes only the chosen subset of predictors. This procedure, which is referred to as the relaxed lasso (Meinshausen, 2007), has been shown to perform favorably compared with best-subset selection, forward stepwise selection, and the lasso without the second stage (Hastie, Tibshirani, & Tibshirani, 2017; Serang et al., 2017). Because we did not follow this two-stage approach, the regularized coefficients are best interpreted only as zero or nonzero.

Discussion

We have argued that regularized SEM is a powerful and underutilized method for researchers who want to examine a (relatively) large number of predictors, or who have a relatively modest sample size combined

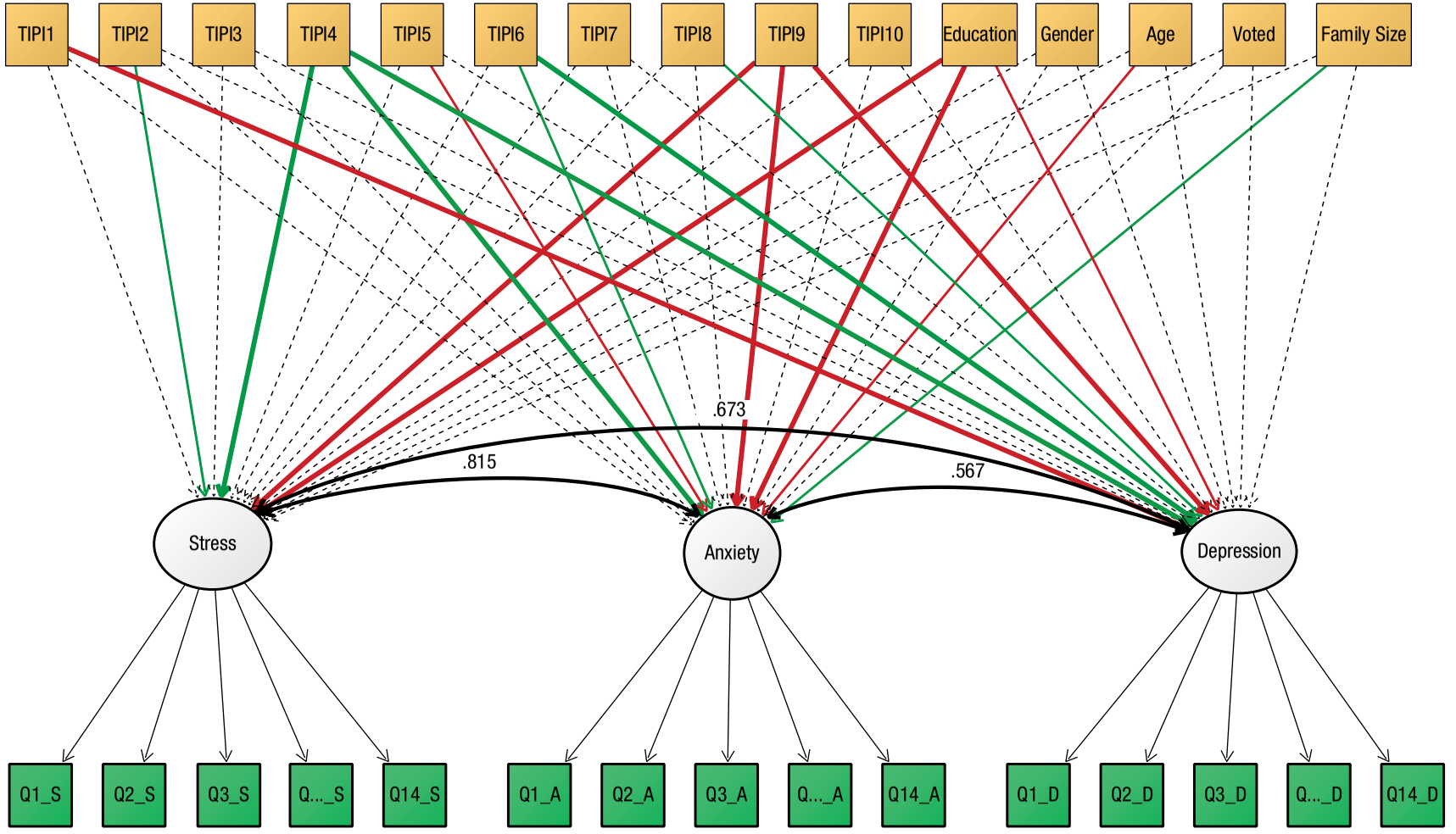

Fig. 10. Results from Study 3: multiple-indicators, multiple-causes (MIMIC) model of stress, anxiety, and depression. Dashed lines indicate nonsignificant paths, and solid lines indicate significant paths ($\alpha < .05$) in the maximum likelihood model. The color of the solid lines indicates whether effects are positive (green) or negative (red), and the lines' thickness indicates the strength of the effects (thick = $z > 3$, thin = $z < 3$). TIPI1 through TIPI10 refer to the items on the Ten-Item Personality inventory (Gosling, Rentfrow, & Swann, 2003). Q1_S through Q14_D (in the green boxes) refer to the items on the Depression Anxiety Stress Scales (Lovibond & Lovibond, 1995). For purposes of clarity, results for the handedness predictor are not shown. Variances are omitted from this figure.

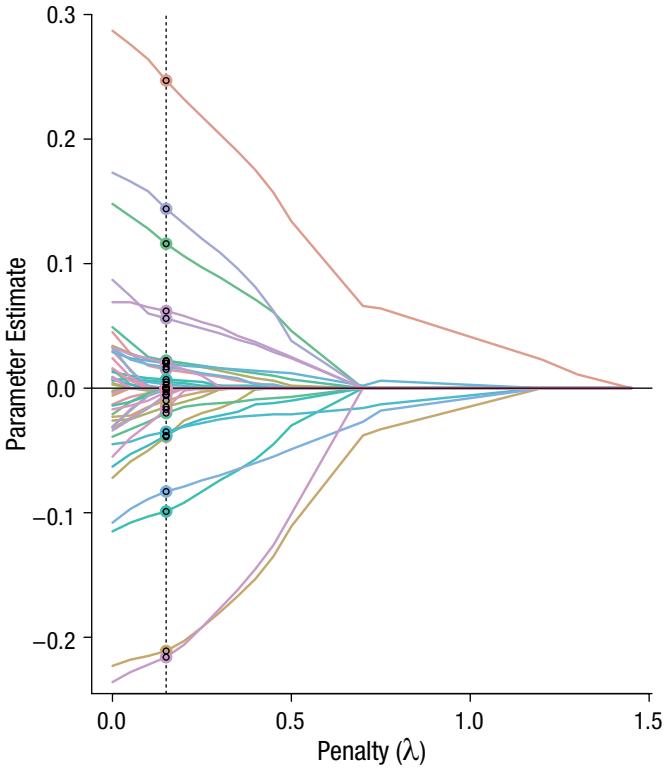

Fig. 11. Parameter trajectory plot from Study 3. The graph shows the values of the 48 regression coefficients as a function of the penalty value. The dashed vertical line highlights the penalty value yielding the model with the best fit (i.e., the lowest Bayesian information criterion).

with a model of moderate complexity. We have described regularization as applied to both regression and SEM, and have evaluated its use in high-dimensional MIMIC models. In our simulation study, models with lasso penalties incurred less error than MLE models when sample sizes were small and demonstrated higher power to detect effects of small and medium magnitude. Our results illustrate how sample size and the correlation among regressors influence the accuracy of parameter estimates and how variable selection is performed in an extremely complex model. Regularized SEM was applied to modeling VSTM as a function of white-matter microstructure in a large existing data set. Starting with a complex model of 48 distinct white-matter tracts, the regularized model identified 6 tracts as determinants of VSTM. Finally, in our last example, we used a regularized model to identify a broad set of variables that explain individual differences in stress, anxiety, and depression.

Our simulation study showed that regularized SEM may be a viable option for researchers looking to identify relatively low-dimensional sets of predictors in fields with broad sets of candidate variables, such as cognitive neuroscience and behavior genetics. Notably, this technique goes beyond traditional methods used

Table 1. Results From Study 3: Fully Standardized Regression Parameters From the Maximum Likelihood and Lasso Models

	Maximum likelihood model		
Predictor	Standardized estimate	Wald-test z	Lasso model's estimate
Predictors of the stress latent variable			
TIPI1	−0.018	−0.55	−0.003
TIPI2	0.062	2.14	0.029
TIPI3	−0.043	−1.54	−0.007
TIPI4	0.394	10.65	0.134
TIPI5	−0.044	−1.42	−0.002
TIPI6	0.041	1.37	0
TIPI7	−0.032	−1.28	0
TIPI8	−0.009	−0.31	0
TIPI9	−0.307	−8.3	−0.103
TIPI10	0.004	0.14	0
Education	−0.099	−3.3	0
Gender	0.047	1.88	0
Age	−0.035	−1.52	−0.012
Handedness	−0.002	−0.08	0
Voter record	0.005	0.19	0
Family size	0.012	0.46	0
Predictors of the anxiety latent variable			
TIPI1	−0.007	−0.2	0
TIPI2	0.029	0.97	0
TIPI3	−0.041	−1.24	0
TIPI4	0.293	7.51	0.069
TIPI5	−0.078	−2.29	−0.016
TIPI6	0.098	3.06	0.003
TIPI7	−0.027	−0.87	0
TIPI8	0.013	0.42	0.014
TIPI9	−0.228	−5.7	−0.05
TIPI10	0.026	0.84	0
Education	−0.125	−3.68	0
Gender	0.014	0.5	0.017
Age	−0.089	−2.02	−0.013
Handedness	0.015	0.54	0
Voter record	0.057	1.97	0
Family size	0.064	2.21	−0.006
Predictors of the depression latent variable			
TIPI1	−0.14	−4.12	−0.04
TIPI2	0.042	1.5	0
TIPI3	−0.041	− 1.32	0
TIPI4	0.224	6.22	0.046
TIPI5	−0.044	−1.33	0
TIPI6	0.113	3.53	0.027
TIPI7	0.012	0.43	0
TIPI8	0.089	2.87	0.043
TIPI9	−0.306	−8.5	−0.118
TIPI10	0.042	1.45	0
Education	−0.071	−2.37	0
Gender	−0.022	−0.85	0
Age	0.01	0.43	−0.009
Handedness	−0.005	−0.17	0
Voter record	0.02	0.74	0
Family size	0.031	1	0

Note: TIPI1 through TIPI10 refer to the items on the Ten-Item Personality inventory (Gosling, Rentfrow, & Swann, 2003).

to correct for multiple comparisons in neuroimaging studies. Methods that correct for multiple comparisons, such as those based on the false discovery rate and Gaussian random-field theory (for an accessible introduction, see Brett, Penny, & Kiebel, 2003), are generally still implemented to correct (mass) univariate tests, rather than joint simultaneous prediction across voxels (regions of interest). It may be possible to combine regularized SEM with methods of joint comparisons, such as principal component regression, to estimate the joint predictive value of multiple components across many voxels even in cases with modest sample sizes (e.g., Wager, Atlas, Leotti, & Rilling, 2011).

Limitations and challenges

Although we have illustrated several benefits of regularization in regression and SEM when sample sizes are small, we did not include any conditions with sample sizes below 100 in our simulation study. This was mostly due to the complexity of our model, as we were unable to achieve stable estimates at a sample size of 120 or below. In regularized regression, it is possible to test models in which there are more predictors than observations; however, to our knowledge, methods of testing models with more predictors than observations have not been extended to SEM, and we were unsuccessful in our attempt to apply regularized SEM to such cases in our simulation study. A possible solution is to use Bayesian SEM, in which strongly informative priors or hierarchal models with sparsity-inducing priors can achieve stable estimation even in such extreme cases (see Jacobucci & Grimm, 2018). Given that Bayesian estimation is increasingly being used in cases of small numbers of observations (McNeish, 2016), and that pairing of Bayesian SEM and regularization has been more widely applied than pairing of frequentist SEM and regularization (see Brandt, Cambria, & Kelava, 2018; Feng, Wu, & Song, 2017; Lu, Chow, & Loken, 2016), we expect to see more research in this area in the future. Other avenues for future work include investigating bias when regularization is used in factor-score regression (Devlieger & Rosseel, 2017), a method that may help overcome the current limitation of needing more observations than predictors. Additionally, bias induced by high degrees of collinearity may be reduced by first creating factor scores, and thus fixing the factor loadings.

Frequentist software for regularized SEM currently requires complete cases. As it is rare for psychological data to have no missing values, this requirement is currently a considerable weakness of regularized SEM. One strategy for modeling data with missing values is multiple imputation. The main issue with using this strategy with regularized SEM concerns how to combine the results. In traditional multiple imputation for SEM, parameter estimates can be aggregated across 10 to 20 data sets (or more) by averaging the parameter estimates and correcting the standard errors for the lack of randomness in the process. However, regularization is most often used to perform variable selection, and this necessitates a way to aggregate a set of 0 to 1 decisions across imputed data sets. Although some research has addressed how to aggregate results in regression (Liu, Wang, Feng, & Wall, 2016), this work has not been generalized to SEM.

Lockhart, Taylor, Tibshirani, and Tibshirani (2014) have derived sampling distributions to calculate p values that take into account the adaptive nature of the lasso regression model, but this work has not been extended to SEM with the lasso. When parameter estimates are not accompanied by p values or confidence intervals, researchers may feel uncertain in making inferences. Consequently, inference can be more challenging with regularized structural equation models than with regularized regression models, particularly given the inherent bias in estimation. One proposed method for overcoming this challenge is the relaxed lasso (Meinshausen, 2007), which has been shown to produce unbiased parameter estimates when applied to mediation models (Serang et al., 2017).

It may be difficult to change the mind-set of relying on p values and instead to characterize nonzero paths as important. To overcome this difficulty, we recommend thinking in terms of generalizing to an alternative sample. Although researchers may incur bias when using regularization, the more important aim is generalization, which is achieved by reducing variance and preferring models of a complexity that is afforded by the observed data. This holds true particularly for exploratory studies, which are less concerned with within-sample inference and more concerned with informing future research.

In our simulation, we found a trade-off between MLE and the RegSEM lasso with respect to Type I and Type II errors: The RegSEM lasso kept more variables in the model (more Type I and fewer Type II errors), whereas MLE was more restrictive with respect to which variables were deemed significant (fewer Type I and more Type II errors). In exploratory studies, we generally recommend a liberal stance; that is, more emphasis should be given to the inclusion of potentially important variables, and the possibility of including variables that do not have either predictive or inferential value should be of less concern. In an ideal setting, researchers would apply regularized SEM to data from a pilot or initial study in the hopes of being maximally efficient in identifying what variables should be included in a future, possibly larger study. Our simulation study

supports the idea that applying MLE when the sample is small and the number of variables is large will result in the exclusion of potentially relevant variables. Note, however, that our conclusions depended not only on the method of regularization applied but also on the specific heuristic for choosing the penalty (i.e., relying on fit indices rather than domain expertise). The penalty values that align with the goals of researchers who want to be relatively inclusive in variable selection (i.e., who can tolerate more Type I errors) may be different from the penalty values that align with the goals of researchers who want to be more exclusive (i.e., who can tolerate more Type II errors; (see also Lakens et al., 2018).

Related approaches

Regularized SEM is only one of the new methods developed for SEM with large data sets. Particularly in the area of variable selection, SEM trees (Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013) and forests (Brandmaier, Prindle, McArdle, & Lindenberger, 2016) are alternative methods. SEM trees directly use the observed covariates to partition observations, and in the process, only a subset of covariates are used to create the model. This allows researchers to uncover nonlinearities and interactions. Additional methods include the use of heuristic search algorithms (e.g., Marcoulides & Ing, 2012), various methods for identifying group differences (Frick, Strobl, & Zeileis, 2015; Kim & von Oertzen, 2018; Tutz & Schauberger, 2015), and the use of graphical models for identifying latent variables (e.g., Epskamp et al., 2017). Given the increasing amounts of data sharing, facilitated by various new tools for data storage and sharing, such as the Open Science Framework (https://osf.io/) and OpenfMRI (https://openfmri.org/), we can envision the utility of testing models much larger than our template simulation model. One of the biggest challenges to such work is software implementation. In this regard, we expect Bayesian estimation (see Jacobucci & Grimm, 2018), discussed earlier, to be a particularly fruitful alternative to our frequentist approach, especially in the creation of new sampling methods such as those in the Stan software package (Carpenter et al., 2016). Interfaces for specifying models (see Merkle & Rosseel, 2015) are sure to be more widely used among psychological researchers as they become easier to use.

Concluding thoughts

We encourage researchers to think of regularization as an approach that can combine confirmatory and exploratory modeling. Regularization gives researchers more flexibility to make both their uncertainty and their knowledge concrete. It is particularly suitable when researchers hope to use a principled approach to go beyond the limitations of their theory to identify potentially fruitful avenues for future study. In both our simulation and our empirical examples, we conducted exploratory searches for important predictors in relation to a confirmatory latent-variable model. This is only one example of how these types of modeling can be fused, and we look forward to seeing new areas of application. We hope that this article sheds light on a new family of statistical methods that have much utility for psychological research.

Appendix: Dependency Among Parameters in MIMIC Models

With two predictors (y_1 and y_2), and just one indicator (x_1) of a single latent variable, the covariance between x_1 and y_1 ($cov_{x1,y1}$) is

$$cov_{x1,y1} = \lambda_{x1}\beta_{y1}\sigma_{y1} + \lambda_{x1}\beta_{y2}\sigma_{y1,y2},$$

where λ_{x1} is the factor loading for x_1, β_{y1} is the regression coefficient for y_1, σ_{y1} is the variance of y_1, and $\sigma_{y1,y2}$ is the covariance between y_1 and y_2.

This equation means that when predictor covariance is high, the estimation of the second regression coefficient plays a large role. Thus, whenever parameters are overpenalized (either because the sample is not large enough to estimate them or because sparsity is desired), this bias not only is incurred in the regression, but also trickles down to the factor loadings. This is always a problem, as the model will try to "make up for" the downward bias of β_{y1} in the value assigned to λ_{x1}, but it can be exacerbated to a large extent when there is covariance among predictors. Adding in a large numbers of predictors makes the problem much worse.

Action Editor

Jennifer L. Tackett served as action editor for this article.

Author Contributions

R. Jacobucci, A. M. Brandmaier, and R. A. Kievit generated the idea for the studies and developed the simulation specification. R. Jacobucci ran the analyses. All three authors analyzed the results, generated the figures, and wrote the manuscript. All the authors approved the final submitted version of the manuscript.

ORCID iDs

Andreas M. Brandmaier https://orcid.org/0000-0001-8765-6982
Rogier A. Kievit https://orcid.org/0000-0003-0700-4568

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

R. A. Kievit is supported by the Sir Henry Wellcome Trust (Grant 107392/Z/15/Z) and by an MRC Programme Grant (SUAG/014/RG91365). This project has also received funding from the European Union's Horizon 2020 Research and Innovation program (Grant 732592).

Open Practices

All materials have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/z2dtq/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245919826527. This article has received the badge for Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Prior Versions

A prior version of this article was posted as a preprint at https://psyarxiv.com/bxzjf/.

Notes

1. Note that regression paths can be penalized regardless of which variables they connect. For example, paths from manifest variables to latent variables can be penalized, as can the reversed paths, paths between latent variables, and paths between manifest variables. In fact, lasso regression can be seen as a subset of the RegSEM lasso method.
2. We tested a sample size of 120 as well, but with this sample size the models generated by the regsem package failed to converge at a high rate. Therefore, we did not include these results.

References

Alexander, A. L., Hasan, K. M., Lazar, M., Tsuruda, J. S., & Parker, D. L. (2001). Analysis of partial volume effects in diffusion-tensor MRI. *Magnetic Resonance in Medicine*, *45*, 770–780. doi:10.1002/mrm.1105

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554.

Bender, A. R., Prindle, J. J., Brandmaier, A. M., & Raz, N. (2016). White matter and memory in healthy adults: Coupled changes over two years. *NeuroImage*, *131*, 193–204. doi:10.1016/j.neuroimage.2015.10.085

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.

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*, 71–86.

Brandmaier, A, M., Wenger, E., Raz, N., & Lindenberger, U. (2018). *Assessing reliability in neuroimaging research through intra-class effect decomposition (ICED)*. Manuscript submitted for publication.

Brandt, H., Cambria, J., & Kelava, A. (2018). An adaptive Bayesian lasso approach with spike-and-slab priors to identify multiple linear and nonlinear effects in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 946–960.

Brett, M., Penny, W., & Kiebel, S. (2003). An introduction to random field theory. In R. S. J. Frackowiak, K. J. Friston, C. D. Frith, R. J. Dolan, C. J. Price, S. Zeki, J. T. Ashburner, & W. D. Penny (Eds.), *Human brain function* (2nd ed., pp. 867–879). London, England: Academic Press.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, *20*, 1–37.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2018). Robust measurement via a fused latent and graphical item response theory model. *Psychometrika*, *83*, 538–562.

Culpepper, S. A., & Park, T. (2017). Bayesian estimation of multivariate latent regression models in large-scale educational assessments: Gauss versus Laplace. *Journal of Educational and Behavioral Statistics*, *42*, 591–616.

de Mooij, S. M. M., Henson, R. N. A., Waldorp, L. J., & Kievit, R. A. (2017). Age differentiation within gray matter, white matter, and between memory and white matter in an adult life span cohort. *Journal of Neuroscience*, *38*, 5826–5836.

Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis. *Methodology*, *13*, 31–38.

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*, *82*, 904–927.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Feng, X. N., Wu, H. T., & Song, X. Y. (2017). Bayesian regularized multivariate generalized latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 341–358.

Frick, H., Strobl, C., & Zeileis, A. (2015). Rasch mixture models for DIF detection: A comparison of old and new score specifications. *Educational and Psychological Measurement*, *75*, 208–234.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *arXiv*. Retrieved from https://arxiv.org/abs/1001.0736

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big Five personality domains. *Journal of Research in Personality*, *37*, 504–528.

Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research*, *26*, 499–510.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*, 430–450.

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician*, *63*, 308–319.

Harrell, F. E., Jr. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). New York, NY: Springer.

Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv*. Retrieved from https://arxiv.org/abs/1707.08692

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton, FL: CRC Press.

Helwig, N. E. (2017). Adding bias to reduce variance in psychological results: A tutorial on penalized regression. *Tutorials in Quantitative Methods for Psychology*, *13*, 1–19.

Henson, R. N., Campbell, K. L., Davis, S. W., Taylor, J. R., Emery, T., Erzinclioglu, S., . . . Kievit, R. A. (2016). Multiple determinants of lifespan memory differences. *Scientific Reports*, *6*, Article 32527. doi:10.1038/srep32527

Hirose, K., & Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing*, *25*, 863–875.

Hodgetts, C. J., Postans, M., Warne, N., Varnava, A., Lawrence, A. D., & Graham, K. S. (2017). Distinct contributions of the fornix and inferior longitudinal fasciculus to episodic and semantic autobiographical memory. *Cortex*, *94*, 1–14.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, *12*, 69–82.

Huang, P.-H., Chen, H., & Weng, L.-J. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika*, *82*, 329–354.

Jacobucci, R. (2017). regsem: Regularized structural equation modeling. *arXiv*. Retrieved from https://arxiv.org/abs/1703.08489

Jacobucci, R., & Grimm, K. J. (2018). Comparison of frequentist and Bayesian regularization in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 639–649.

Jacobucci, R., Grimm, K. J., Brandmaier, A. M., & Serang, S. (2017). regsem: Regularized structural equation modeling (R package Version 1.0.6) [Computer software]. Retrieved from https://cran.r-project.org/package=regsem

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 555–566.

Jones, D. K., Knösche, T. R., & Turner, R. (2013). White matter integrity, fiber count, and other fallacies: The do's and don'ts of diffusion MRI. *NeuroImage*, *73*, 239–254.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*, 631–639.

Kievit, R. A., Davis, S. W., Griffiths, J., Correia, M. M., Cam-CAN, & Henson, R. N. (2016). A watershed model of individual differences in fluid intelligence. *Neuropsychologia*, *91*, 186–198. doi:10.1016/j.neuropsychologia.2016.08.008

Kievit, R. A., Davis, S. W., Mitchell, D., Taylor, J. R., Duncan, J., Cam-CAN, & Henson, R. N. (2014). Distinct aspects of frontal lobe structure mediate age-related differences in fluid intelligence and multitasking. *Nature Communications*, *5*, 1–10.

Kievit, R. A., van Rooijen, H., Wicherts, J. M., Waldorp, L. J., Kan, K.-J., Scholte, H. S., & Borsboom, D. (2012). Intelligence and the brain: A model-based approach. *Cognitive Neuroscience*, *3*, 89–97. doi:10.1080/17588928.2011.628383

Kim, B., & von Oertzen, T. (2018). Classifiers as a model-free group comparison test. *Behavior Research Methods*, *50*, 416–426.

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, *5*, 369–411.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., . . . Buchanan, E. M. (2018). Justify your alpha. *Nature Human Behaviour*, *2*, 168–171.

Laurin, C., Boomsma, D., & Lubke, G. (2016). The use of vector bootstrapping to improve variable selection precision in Lasso models. *Statistical Applications in Genetics and Molecular Biology*, *15*, 305–320.

Laxton, A. W., Tang-Wai, D. F., McAndrews, M. P., Zumsteg, D., Wennberg, R., Keren, R., . . . Lozano, A. M. (2010). A phase I trial of deep brain stimulation of memory circuits in Alzheimer's disease. *Annals of Neurology*, *68*, 521–534.

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods*, *4*, 192–211.

Liu, Y., Wang, Y., Feng, Y., & Wall, M. M. (2016). Variable selection and prediction with incomplete high-dimensional data. *Annals of Applied Statistics*, *10*, 418–450.

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. *Annals of Statistics*, *42*, 413–468.

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*, 335–343.

Lu, Z. H., Chow, S. M., & Loken, E. (2016). Bayesian factor analysis as a variable-selection problem: Alternative priors and consequences. *Multivariate Behavioral Research*, *51*, 519–539.

Madden, D. J., Spaniol, J., Costello, M. C., Bucur, B., White, L. E., Cabeza, R., . . . Huettel, S. A. (2009). Cerebral white matter integrity mediates adult age differences in cognitive performance. *Journal of Cognitive Neuroscience*, *21*, 289–302. doi:10.1162/jocn.2009.21047

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*, 111–135.

Marcoulides, G. A., & Ing, M. (2012). Automated structural equation modeling strategies. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 690–704). New York, NY: Guilford Press.

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, *50*, 471–484.

McNeish, D. M. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 750–773. doi:10.1080/10705511.2016.1186549

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195–244.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, *52*, 374–393. doi:10.1016/j.csda.2006.12.019

Menegaux, A., Meng, C., Neitzel, J., Bäuml, J. G., Müller, H. J., Bartmann, P., . . . Sorg, C. (2017). Impaired visual short-term memory capacity is distinctively associated with structural connectivity of the posterior thalamic radiation and the splenium of the corpus callosum in preterm-born adults. *NeuroImage*, *150*, 68–76. doi:10.1016/j.neuroimage.2017.02.017

Merkle, E. C., & Rosseel, Y. (2015). blavaan: Bayesian structural equation models via parameter expansion. *arXiv*. Retrieved from https://arxiv.org/abs/1511.05604

Mori, S., Oishi, K., Jiang, H., Jiang, L., Li, X., Akhter, K., . . . Mazziotta, J. (2008). Stereotaxic white matter atlas based on diffusion tensor imaging in an ICBM template. *NeuroImage*, *40*, 570–582. doi:10.1016/j.neuroimage.2007.12.035

Park, M. Y., & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*, 659–677.

Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, *103*, 681–686.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2).

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Serang, S., Jacobucci, R., Brimhall, K., & Grimm, K. J. (2017). Exploratory mediation analysis via regularization. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 733–744.

Shafto, M. A., Tyler, L. K., Dixon, M., Taylor, J. R., Rowe, J. B., Cusack, R., . . . Henson, R. N. (2014). The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: A cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurology*, *14*, Article 204. doi:10.1186/s12883-014-0204-1

Sümer, S., Poyrazli, S., & Grahame, K. (2008). Predictors of depression and anxiety among international students. *Journal of Counseling & Development*, *86*, 429–437.

Sun, J., Chen, Y., Liu, J., Ying, Z., & Xin, T. (2016). Latent variable selection for multidimensional item response theory models via L1 regularization. *Psychometrika*, *81*, 921–939.

Taylor, J. R., Williams, N., Cusack, R., Auer, T., Shafto, M. A., Dixon, M., . . . Henson, R. N. (2017). The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *NeuroImage*, *144*, 262–269. doi:10.1016/j.neuroimage.2015.09.018

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, *55*, 525–534.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *58*, 267–288.

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*, 21–43.

van Erp, S., Oberski, D. L., & Mulder, J. (2018). *Shrinkage priors for Bayesian penalized regression*. Retrieved from https://osf.io/cg8fq

Vestergaard, M., Madsen, K. S., Baaré, W. F., Skimminge, A., Ejersbo, L. R., Ramsøy, T. Z., . . . Jernigan, T. L. (2011). White matter microstructure in superior longitudinal fasciculus associated with spatial working memory performance in children. *Journal of Cognitive Neuroscience*, *23*, 2135–2146.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192–196.

Wager, T. D., Atlas, L. Y., Leotti, L. A., & Rilling, J. K. (2011). Predicting individual differences in placebo analgesia: Contributions of brain activity during anticipation and pain experience. *Journal of Neuroscience*, *31*, 439–452.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*, 67–82.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122.

Yuan, K.-H., Yang, M., & Jiang, G. (2017). Empirically corrected rescaled statistics for SEM with small *N* and large *p*. *Multivariate Behavioral Research*, *52*, 673–698.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, *38*, 894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.

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.