Psychological Methods

Regularized Structural Equation Modeling With Stability Selection

Xiaobei Li and Ross Jacobucci

Online First Publication, January 28, 2021. http://dx.doi.org/10.1037/met0000389

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

ISSN: 1082-989X

https://doi.org/10.1037/met0000389

Regularized Structural Equation Modeling With Stability Selection

Xiaobei Li and Ross Jacobucci
Department of Psychology, University of Notre Dame

Abstract

Regularization methods such as the least absolute shrinkage and selection operator (LASSO) are commonly used in high dimensional data to achieve sparser solutions. Recently, methods such as regularized structural equation modeling (SEM) and penalized likelihood SEM have been proposed, trying to transfer the benefits of regularization to models commonly used in social and behavioral research. These methods allow researchers to estimate large models even in the presence of small sample sizes. However, some drawbacks of the LASSO, such as high false positive rates (FPRs) and inconsistency in selection results, persist at the same time. We propose the application of stability selection, a method based on repeated resampling of the data to select stable coefficients, to regularized SEM as a mechanism to overcome these limitations. Across 2 simulation studies, we find that stability selection greatly improves upon the LASSO in selecting the correct paths, specifically through reducing the number of false positives. We close the article by demonstrating the application of stability selection in 2 empirical examples and presenting several future research directions.

Translational Abstract

With the rapid increase in data collection in empirical science, larger number of variables are available for researchers to explore, leading to a pervasive challenge of variable selection. Data mining techniques such as regularization have been receiving increasing attention in psychological studies in recent years, however, the results from blindly applying those data mining techniques could lead to number of problems. For example, a large number of uninformative variables could be selected and treated as informative, and the selection result may be different from the targeted set of variables even with a large enough sample size. We found that using stability selection is a good way to alleviate those problems, and to find more trustworthy selection results. We detail this algorithm under the structural equation modeling framework and assess its use in simulations and 2 empirical examples.

Keywords: structural equation modeling, regularization, LASSO, stability selection, machine learning

Supplemental materials: https://doi.org/10.1037/met0000389.supp

With the rapid increase in data collection in psychological science, larger number of variables are available for researchers to explore, providing a pervasive challenge of variable selection. Although including more variables in the model could typically increase the flexibility of the model and thus improve the fit, this flexibility comes at the cost of interpretability, as a complex relationship is typically harder to interpret. Moreover, overfitting could occur when the model attempts to capture noise in the sample data rather than the true signal, making the model less generalizable. From a bias-variance tradeoff perspective, more flexible methods reduce the bias, but increase the variance (e.g., see Yarkoni & Westfall, 2017). Thus, when inference is the goal, there are clear advantages to using models with simpler structures. In addition to the increased interpretability of simpler models, parameter estimates often have smaller standard errors, allowing for more precise substantive conclusions (Raykov & Marcoulides, 1999).

Stemming from statistical learning, regularized regression methods, such as the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996), have been shown to effectively produce sparse solutions. This is achieved by explicitly penalizing the complexity (parameter coefficients) of the model. To be specific, a penalty term is added to the fit function that sums the freely estimated parameters, so that the new fit function would yield a different set of parameter estimates, which are biased toward zero. When the "summing" technique is carefully chosen (L_1 norm in the case of LASSO), some of the less relevant parameters are forced to zero, thus performing variable selection, or creating a simpler model structure (detailed later). In terms of total generalization error, this bias may not be a bad thing, as the LASSO trades off an increase in bias with a decrease in variance. For an overview, see Yarkoni and Westfall (2017).

Although regularization is a widely adopted method in regression and has been shown to possess many desirable properties, it was not

Xiaobei Li https://orcid.org/0000-0002-8266-9490
Ross Jacobucci https://orcid.org/0000-0001-7818-7424
This has been posted as a preprint: https://psyarxiv.com/t4d5y.
Correspondence concerning this article should be addressed to Xiaobei Li, Department of Psychology, University of Notre Dame, 412 Corbett Family Hall, Notre Dame, IN 46556, United States. Email: xli29@nd.edu

until recently that regularization has been applied to the structural equation modeling (SEM) framework in regularized structural equation modeling (regularized SEM [RegSEM]; Jacobucci et al., 2016) and the penalized likelihood (PL) method for structural equation models (Huang et al., 2017). Through penalizing specific parameters in an SEM, RegSEM attempts to transfer the benefits of regularization to models with latent variables. The goals of this include creating simpler models, although still demonstrating good fit, or overcoming problems with maximum likelihood estimation (MLE), for instance, "Heywood cases" (Bentler & Chou, 1988), nonconvergence, underidentification, and many others. With a penalty term applied to different parameters, the regularization process could take many different forms, for example, selecting among multiple predictors of a latent variable (Jacobucci et al., 2019), simplifying factor structure by removing cross-loadings (Scharf & Nestler, 2019), determining whether additional linear terms are necessary in longitudinal models, and many others.

The idea of applying regularization/penalization[1] to obtain sparser solutions has been applied broadly in psychometrics. For example, LASSO regularization has been applied in item response theory (IRT) to achieve latent variable selection (Sun et al., 2016). More recently, Bauer et al. (2019) proposed a regularization approach to moderated nonlinear factor analysis estimation, penalizing the likelihood for differential item functioning (DIF) parameters, to simplify the assessment of measurement invariance. Liang and Jacobucci (2019) studied the performance of several regularization methods for detecting measurement bias, with focus put on complex SEM with small sample sizes. In confirmatory factor analysis (CFA), Pan et al. (2017) proposed a post hoc model modification method using the Bayesian LASSO. In exploratory factor analysis (EFA), besides the traditional two-step approach of applying factor rotation to the least squares or maximum likelihood estimators to obtain sparser solutions, penalized least squares (Sitek et al., 2002) and penalized maximum likelihood (Hirose & Yamamoto, 2013; Hirose & Yamamoto, 2015; Zou et al., 2010) techniques have been proposed. Furthermore, Scharf and Nestler (2019) showed that though both rotated and regularized EFA tended to underestimate cross-loadings and inflate factor correlations when the factor structures are complex, regularized EFA was able to recover loading patterns as long as some of the items followed a simple structure.

In CFA, the hypothesized model is usually constructed based on theory and/or previous research, whereas in EFA the researcher typically has no a priori hypotheses about latent factor composition or patterns of the measured variables. Regularized SEM, however, considers the case in between, where the researcher has part of the model constructed based on theory, but another part of the model that the researcher wishes to explore. In this case, the researcher could place penalties only on those uncertain parts, thereby specifying a model that exists in between purely confirmatory and exploratory modeling. After the selection procedure, the resulting model should then be verified on a new dataset.

Historically, modification indices (MacCallum et al., 1992) are used when the data does not fit the hypothesized model well. Specifically, the chi-square difference (D) test has been used to compare nested models in CFA, path analysis, and SEM. Easier-to-apply chi-square test procedures such as the Lagrange multiplier (LM) test and Wald (W) test (Bentler, 1986; Chou & Bentler, 1990) were shown to be asymptotically equivalent to the D test (Lee & Bentler, 1980). The LM test is used to evaluate the effect of adding free parameters, whereas the W test is used to evaluate the effect of dropping free parameters, which is more relevant to the circumstances in which regularized SEM typically applies. Although regularized SEM performs variable or structure selection in one step (all parameters entering the model are selected at once), the W test could take several rounds. This practice could be hazardous when a large modification are to be made, and should be implemented with care. For more detailed discussion of LM test and W tests, see chapter 6 of Bentler (2006).

Current applications of regularized SEM, and regularization in psychometrics more broadly, typically use the LASSO penalty because of its ability to perform variable selection and its computational simplicity. However, two prominent problems with the LASSO have been noted in previous studies. The first problem is with the consistency of LASSO. A variable selection procedure is consistent if the probability of selecting the correct model tends to one as sample size goes to infinity. However, it has been shown that nontrivial conditions (irrepresentable condition; Yuan & Lin, 2006; Zhao & Yu, 2006; Zou, 2006; and "neighborhood stability" condition; Meinshausen & Bühlmann, 2006) are needed for the LASSO to provide consistent selection results. As a stricter property, a method is said to possess the oracle property if it is consistent in the variable selection procedure, and it estimates the nonzero parameters as efficiently as could be possible if we knew which variables were uninformative ahead of time. In short, an oracle estimator needs to be consistent in both variable selection and parameter estimation. Moreover, Fan and Li (2001) proved that the oracle properties do not hold for LASSO by arguing that large coefficients are biased downward when estimated with LASSO, which could be suboptimal in terms of estimation rate. Besides, the LASSO depends upon a regularizing parameter λ, which controls the amount of shrinkage applied to the coefficients. This value must be chosen in some fashion, for instance, by cross-validation. However, it has been noted that the result could be sensitive to "fold assignment" (Bovelstad et al., 2007; Datta & Zou, 2019; Roberts & Nowak, 2014). On top of that, Meinshausen and Bühlmann (2006) showed the conflict of optimal prediction and consistent variable selection in the LASSO. When choosing the optimal tuning parameter λ, a criterion like prediction accuracy is needed, which, in general does not give consistent variable selection result (Leng et al., 2006).

An additional problem with the LASSO is the high false positive rates (FPRs), that is, null features or noise variables being selected. Meinshausen (2007) pointed out that many noise variables are selected by LASSO regression if the estimators are chosen by cross-validation. Moreover, false discoveries may occur very early on the LASSO path[2] (Fan & Song, 2010), and true features and null features are always interspersed on the LASSO path, no matter how strong the effect sizes are (Su et al., 2017), leading to a practical tradeoff between FPRs and FNRs. This pattern is also observed in regularized SEM. We demonstrate this through fitting

[1] Throughout the article, we use the term *regularization* and *penalization* interchangeably.

[2] We distinguish the LASSO path from paths in SEM: The LASSO paths are the profiles of the coefficients across the vector of penalties. As the tuning parameter gets larger, the coefficient estimates generally decreases at different rates, with some decreasing to 0 earlier than others. With a particular tuning parameter (penalty), only those who did not hit zero should remain in the model.

a regression model with 9 predictors using regularized SEM. The data are simulated to have 3 different coefficient sizes, 0.8, 0.2 and 0, each with 3 of the 9 predictors. The penalties are applied to the 6 predictors with simulated coefficient 0.2 and 0 when running the regularization model. This process is repeated on 100 bootstrap samples of the simulated data, and the resulting FPRs and FNRs are displayed in Figure 1. It is observed that although regularized SEM generally produces very low averaged false negative rates (FNRs) on small samples, the averaged FPRs is usually relatively high (e.g., Jacobucci et al., 2019). That is, when the LASSO penalty is used, regularized SEM does not tend to overpenalize signals (true features), but it may not give enough penalization, so that a large number of noise variables (null features) are often included. Nevertheless, this high averaged FPRs stays even with large enough sample size. For instance, in Jacobucci et al. (2019), an average of over 35% of noise variables are selected to the final set by regularized SEM with small sample sizes, and the average FPRs stays over 30% even with sample size of 2000. This is in line with the performance of LASSO in regression (Drysdale et al., 2019; Su et al., 2017), tracing back to the problem of inconsistency. In many other applications in psychometrics, similar patterns are observed. For example, when LASSO is used in regularized DIF (Bauer et al., 2019), with the tuning parameter determined by minimizing the Bayesian information criteria (BIC; Schwarz, 1978), the FPRs could be as high as 48% when DIF is pervasive (66%) and of large magnitude, at sample size of 2,000, though this rate is lower than using the item response theory likelihood ratio DIF detection procedure (Thissen et al., 1993). As discussed earlier, regularized SEM has both a confirmatory part and an exploratory part, although its selection procedure is data driven. For such methods, the best we could hope for is it get to the "true model" asymptotically when the model is structured in the correct way. That is, we want the probability of selecting the correct model to increase as sample size increases. However, the LASSO penalty fails this aim (consistency). In addition, with high FPRs, the final model would contain many noise variables/paths, and researchers may get different sets of selection results each time rerunning the model, because different sets of noise variables/paths might have been falsely chosen each time. Moreover, without consistency, this would still happen even with very large sample sizes. Both drawbacks of the LASSO could drastically change researchers' assessment of the trustworthiness of the results.

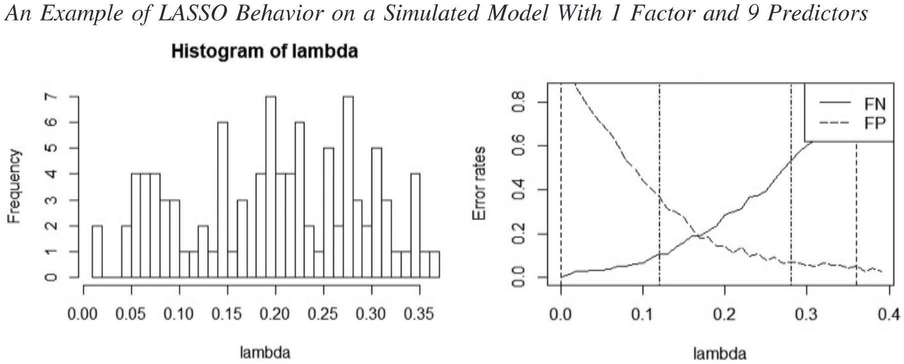
To demonstrate issues with LASSO in SEM, we used a subset of the model detailed in Empirical Example 1. In this, we specified a two-factor CFA model for the Grit Scale (Duckworth et al., 2007), while using 10 Conscientiousness items and a randomly selected sample of size 1,000, to predict both latent factors, resulting in a MIMIC model. Our goal in applying the LASSO is to remove a subset of the 20 regression coefficients, resulting in a simplified understanding of the relationship between constructs. Further, we ran the model with 100 bootstrap samples to assess the stability of the penalties selected, as well as the resulting variability to the regression paths that were selected.

The selection rate of each regression path across the 100 bootstrap samples is summarized in Table 1. Although the bootstrap samples overlap with each other, we see high degrees of instability in a subset of paths. For example, the path "C1→grit1" has a selection probability of 0.94, which means that it is selected in 94 of the 100 bootstrap samples when regularized SEM is applied. This probability is much higher than that of path "C10→grit1" (0.33), which indicates that C1 is more stable than C10 as a predictor to the latent factor grit1. However, a single run of regularized SEM cannot reveal this stability information. Figure 2 demonstrated the cause of this instability in a histogram of the optimal penalty level chosen by LASSO on the 100 bootstrap samples. Similar to what we found in Figure 1, the penalty values varied significantly (from 0 to over 0.3), which leads to different sets of selection results. With this degree of instability, researchers may select a less stable path but leave a more stable path out, and the selection result may vary significantly across slightly different samples, casting serious doubt on the reliability of the selection results.

In statistical learning, a large body of research has shown that instability could be reduced by aggregating over a large set of competing models (e.g., bootstrap aggregating, Breiman, 1996; random

Figure 1

An Example of LASSO Behavior on a Simulated Model With 1 Factor and 9 Predictors



Note. Left panel: Histogram of possible optimal λ values chosen by LASSO regularization on bootstrap samples of the simulated data. Right panel: False positive rates (FPRs) and false negative rates (FNRs) on the full sample based on different λ values. Dashed lines represent the minimum and maximum of the chosen λ values based on bootstrap samples. Dash-dotted lines are the 25th and 75th quantiles of the chosen λ values based on bootstrap samples. Within this interquartile region, the FPRs and FNRs are relatively balanced.

Table 1

Selection Rate of the Regression Path Across the 100 Bootstrap Samples

Path	Selection probability
C1 → grit1	0.94
C2 → grit1	0.74
C3 → grit1	0.47
C4 → grit1	1
C5 → grit1	0.97
C6 → grit1	1
C7 → grit1	0.34
C8 → grit1	1
C9 → grit1	0.98
C10 → grit1	0.33
C1 → grit2	1
C2 → grit2	0.73
C3 → grit2	1
C4 → grit2	0.55
C5 → grit2	1
C6 → grit2	0.7
C7 → grit2	0.61
C8 → grit2	1
C9 → grit2	1
C10 → grit2	1

forest, Breiman, 2001; boosting, Friedman, 2001). Moreover, Barber and Candès (2015) pointed out that effective FPR control helps in improving model consistency and interpretation. Stability selection, as proposed by Meinshausen and Bühlmann (2010), was designed to improve the performance of variable selection and structure estimation algorithms by aggregating the selection results from bootstrap samples. Instead of performing regularization once on the whole sample, stability selection performs it multiple times on multiple bootstrap samples. Then at each penalty value we get a probability of each variable/path being selected by the regularization procedure. The final selection result is then based on those selection probabilities. For instance, suppose we only have two variables, and we only consider one penalty value. Suppose one variable is selected 90% of the time, and the other one is only selected 20% of the time. It is clear that the first variable is more "stable" and should be kept in the final selection result, whereas the other one is more likely a noise variable. The actual algorithm becomes more complicated when multiple penalty values are involved (detailed later), but the idea is the same. By applying stability selection, it provides finite sample control for FPRs and hence a transparent principle to choose a proper amount of regularization. It is a very general technique based on resampling that has a wide range of applicability, for instance, in neuroscience (Ryali et al., 2015; Ye et al., 2012), genetics (Alexander & Lange, 2011), and bioinformatics (Hofner et al., 2015).

In this study, we aim to directly address the inconsistency and high FPRs of LASSO regularization in latent variable models by incorporating stability selection in regularized SEM. We first provide an overview of regularized SEM and stability selection, then present the proposed algorithms to use regularization and stability selection under the SEM framework. We then compare the performances of stability selection as well as a number of extensions based on the LASSO, in terms of FPRs, FNRs, bias, and variance across two simulation studies, focusing on selecting predictors of latent variables, and selecting cross loadings, respectively. Lastly, we perform stability selection on two empirical examples to further discuss the practical advantages and issues of the proposed method. Although we focus on the application of regularized SEM, we believe that our proposed procedure and findings would pertain to other types of psychometric models with regularization applied.

RegSEM

Regularized SEM (Jacobucci et al., 2016) adds a penalty term to the traditional MLE fit function:

$$F_{regsem} = \underbrace{log(|\Sigma|) + tr(C * \Sigma^{-1}) - log(|C|) - p}_{MLE} + \underbrace{\lambda P(\cdot)}_{penalty}, \tag{1}$$

where Σ is the model implied covariance matrix, C is the observed covariance matrix, and p is the number of variables. The regularization parameter λ is a tuning parameter that cannot be determined jointly with the coefficients. It quantifies the influence of the penalty, with larger λ incurring larger penalty, and thus result in greater shrinkage of the coefficient sizes. Its optimal value is often determined through cross-validation or based on information criteria, such as Akaike information criterion (AIC; Akaike, 1974)

$$AIC = 2k - 2\ln(\hat{L}),$$

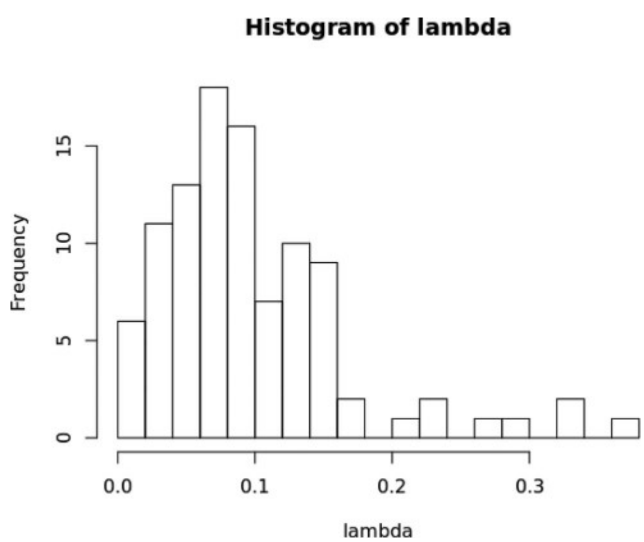
where $\hat{L}$ is the maximum value of the likelihood function for the model, and Bayesian information criteria (BIC; Schwarz, 1978)

$$BIC = \ln(N)k - 2\ln(\hat{L}).$$

As can be seen in the formulas, the AIC does not depend directly on the sample size N, making it prone to overfitting, whereas BIC does depend on sample size, and is more conservative relative to AIC in most cases. In practice, the BIC generally performs well in finding the regularization parameter λ that results in reasonable

Figure 2

Histogram of Possible Optimal λ Values Chosen by Regularized Structural Equation Modeling on Bootstrap Samples of a Subset of the Model in Empirical Example 1



parameter estimates for regularized factor analysis models (Hirose & Yamamoto, 2015; Jacobucci et al., 2016).

$P(\cdot)$ is a general penalty function which reflects the sum of all selected coefficients, and can take many different forms like the LASSO ($\|\cdot\|_1$), ridge ($\|\cdot\|_2$), and elastic net ($(1 - \alpha)\|\cdot\|_1 + \alpha\|\cdot\|_2$). Although the LASSO is widely used in regression, using only one value of λ could lead to predictable bias in the resulting estimates, particularly when the scale of the variables are dramatically[3] different (Fan & Li, 2001). For this reason, a two-stage procedure termed the "relaxed LASSO" (Meinshausen, 2007) is usually performed when the LASSO is chosen as the penalty type in regularized SEM. That is, we first perform variable/path selection using LASSO, and then we fit an SEM model with only the selected variables/paths to get their unbiased estimates. It has been shown that the relaxed LASSO produces unbiased parameter estimates when applied to mediation models estimated under the SEM framework (Serang et al., 2017).

Another method proposed to overcome this bias of the LASSO is through using the adaptive LASSO (ALASSO; Zou, 2006). When the ALASSO is applied in SEM, each parameter to be penalized is first scaled by the MLE estimates before LASSO is performed ($\|\beta_{ML}^{-1} * \beta_{pen}\|_1$). With this scaling, smaller parameters would receive larger penalties, while limiting the bias in the estimation of larger parameters. However, because MLE parameter estimates are needed when applied to SEM, the ALASSO may not be applicable when the model has a large number of variables in relation to sample size, or other estimation difficulties have occurred.

As we have discussed previously in the introduction, besides bias, the high FPRs and inconsistency of the selection results are two major concerns of LASSO. Besides the ALASSO, other penalties have also been proposed to deal with the above two problems, for example, the smoothly clipped absolute deviation (SCAD; Fan & Li, 2001), and minimax concave penalty (MCP; Zhang, 2010) have been shown to possess the oracle properties. However, some penalties may involve minimization of nonconvex function, which could be computationally complex (Fan & Peng, 2004; Meinshausen, 2007). Jacobucci (2017) and Huang et al. (2017) demonstrated the algorithmic details of both the MCP and SCAD, however, in practice they are difficult to estimate, particularly in complex SEMs.

In summary, regularization is performed by biasing the parameter estimates, with the LASSO penalty being one of the most popular choices. To overcome the bias in the final solution, the two-step "relaxed LASSO" procedure is recommended when the LASSO penalty is used (Meinshausen, 2007). That is, first perform variable/path selection using LASSO penalty, then fit model with only the selected variables/paths to get their unbiased estimates. There are also many other types of penalties to overcome the drawbacks of the LASSO, and the ALASSO penalty is a good option when the starting model is identified.

Stability Selection

Stability selection uses resampling for the purpose of selecting variables. The method is very general and can be paired with multiple variable selection or structural estimation techniques. Before we introduce the idea of stability selection, we first formalize the variable selection (or, structure estimation) problem: Suppose we have a p-dimensional vector β, with s ($s < p$) components being nonzero. We denote the set of nonzero values by $S = \{k\colon \beta_k \neq 0\}$, and the set of zero coefficients by $V = \{k\colon \beta_k = 0\}$. The goal of variable selection (or, structure estimation) is then to infer the set S from the p possible components. For every tuning parameter $\lambda \in \Lambda \subseteq \mathbb{R}^+$ that determines a specific amount of regularization, we could get a structure estimate $\hat{S}^\lambda \subseteq \{1, \ldots, p\}$. The goal is then to determine whether there exists a $\lambda \in \Lambda$ such that $\hat{S}^\lambda = S$ with high probability, and to achieve (or get a close approximation) of this λ.

Stability selection does not focus on finding a single λ, rather, it aggregates across a range Λ of regularization values. The procedure is as follows: Let I be a random sample of $\{1, \ldots, n\}$ of size $\lfloor \frac{n}{2} \rfloor$ drawn without replacement ($\lfloor \frac{n}{2} \rfloor$ is chosen as it mostly closely resembles bootstrap, while allowing for a computationally efficient implementation). For every set $K \subseteq \{i, \ldots, p\}$, the probability of being in the selected set is $\hat{\Pi}_K^\lambda = P^*(K \subseteq \hat{S}^\lambda(I))$. Then for each predictor $k = 1, \ldots, p$, the stability path is given by the selection probabilities $\hat{\Pi}_K^\lambda$, $\lambda \in \Lambda$. For each cutoff $\pi_{thr} \in (0, 1)$, and a set of regularization parameters Λ, the set of stable predictors is then defined as $\hat{S}^{stable} = \{k\colon \max_{\lambda \in \Lambda} \hat{\Pi}_k^\lambda \geq \pi_{thr}\}$. Predictors with high selection probabilities are kept (e.g., > 0.8 or 0.9) and those with low selection probabilities are discarded. The steps for stability selection combined with regularized SEM are summarized and discussed in the next section.

To see the advantage of stability selection, let Λ denote the set of regularization parameters. To perform variable selection, traditional techniques pick one model from the set $\{\hat{S}^\lambda\colon \lambda \in \Lambda\}$, which may lead to the two potential problems. First is that the correct model S may not be contained in the above set. Even in the absence of this problem, the second problem is that even if the correct model S is contained in the set, we may not be able to accurately determine a correct degree of regularization λ to select exactly S, or even a close approximation of S. Perturbing the data multiple times and choosing structures or variables that occur with high probability in the selection set allows for more flexible choices which are not limited to one single model from the set (Meinshausen & Bühlmann, 2010), and thus, potentially eliminating the two identified problems. It has also been shown that stability selection can control the familywise Type I error rate (or, FPR) in multiple testing for finite sample size and thus markedly improve the structure estimation or selection methods (Meinshausen & Bühlmann, 2010).

RegSEM With Stability Selection

We now provide the details on pairing RegSEM with stability selection:

1. Standardize the variables to place them on the same scale.

2. Select a reasonable range for Λ (discussed below).

3. For each $\lambda \in \Lambda$, obtain M (e.g., $M = 100$) bootstrap samples of $I = \{1, \ldots, n\}$.

[3] It is typically recommended to standardize variables prior to analysis, however, there are some latent variable models where this is inappropriate.

a. For each bootstrap sample T, perform regularized SEM with this specific amount of regularization λ, and obtain a set of selected predictors $\hat{S}^{\lambda}(T) \subseteq \{1, \ldots, p\}$.

b. Consider the M sets of selected predictors altogether and obtain the selection probabilities $\hat{\Pi}_k^{\lambda}$ for each k in $1, \ldots, p$, at this particular λ.

4. Select the final set $\hat{S}^{stable} = \{k: \max_{\lambda \in \Lambda} \hat{\Pi}_k^{\lambda} \geq \pi_{thr}\}$ based on the choice of probability cutoff π_{thr}.

a. Here, we select all predictors that have selection probability greater than or equal to π_{thr} at some $\lambda \in \Lambda$.

5. Refit the SEM model without any regularization, using only the subset of selected predictors.

One issue with the traditional form of stability selection is that although it is said that the empirical results vary little for threshold values in the range (0.6, 0.9) and are not sensitive to choices of Λ for high dimension cases (Meinshausen & Bühlmann, 2010), it is not the case in low dimension settings (Huang & Marchetti-Bowick, 2014). Thus, when applied in psychological research where the dimension is usually not high, selecting a proper range Λ and a proper threshold value π_{thr} becomes crucial.

Choosing the probability cutoff π_{thr} is an extremely important component of this procedure as it determines the final selection result. This value could be predetermined as a fixed value by researchers (e.g., 0.8), or it could be tuned based on model fit. To be specific, we could refit the SEM model based on a different cutoff value π_{thr}, then compare the corresponding model fit, choosing the threshold that corresponds to the best fitted model.

Because we are aggregating the maximum selection probabilities $\max_{\lambda \in \Lambda} \hat{\Pi}_k^{\lambda}$ of each variables across a range of regularization Λ, this range Λ needs to be carefully chosen. It could either be determined by theory or by prerun results. For example, we could prerun RegSEM on bootstrap samples of the data to get the empirical distribution (left panel of Figure 1) of the final λ values chosen by regularized SEM. Then we could use the interquartile range of the prerun λ values as the range for stability selection. As demonstrated in the right panel of Figure 1, the region between the two dash-dotted lines corresponds to the interquartile range of the λ values for the example data. In this region, the FPR and FNR are relatively balanced, and the cross point of the two error rate curves falls in this region most of the time. This range is used throughout this study, while we leave it open for better choices.

Simulation Studies

Through the use of two simulation studies, our goal is to examine whether stability selection can improve the selection results (in terms of selection consistency and controlling FPRs) in regularized SEM. In addition, we aim to assess whether the improvements are worth the extra computational costs. We examine how severe the two previously stated problems with the LASSO (i.e., inconsistency and high FPRs) are when the LASSO penalty is used in regularized SEM. Further, we assess the bias and variance of the relaxed LASSO to judge whether the drawbacks with the LASSO would pose problems in real applications. Moreover, as stated in the introduction, some extensions of LASSO are developed to overcome drawbacks of LASSO, and are available in the R package `regsem` (v1.5.0; Jacobucci et al., 2020). We also include the ALASSO as a representative of those LASSO extensions, and additional computational methods such as cross-validation and bootstrapping to see whether there are computationally cheaper solutions. The performance of the different methods are compared in terms of FPRs, FNRs, root mean square error (RMSE) and relative bias (RB). We hypothesize that stability selection will have lower FPR than other methods, without increasing FNRs by too large of an extent, and could provide consistent selection results. With its improved ability to select the correct set of variables/paths, we also expect the RMSE and RB of the parameter estimates from the model constructed by stability selection to be lower.

Method

In each study, nine methods are compared: Besides (a) the traditional p value based selection with MLE (using 0.05 as a cutoff) and (b) RegSEM-LASSO procedure, we also tested the performance of (c) RegSEM-ALASSO. For computational methods, we tested the performance of (d) fivefold cross validated RegSEM-LASSO using the one-standard-error (1se) rule[4] and (e) RegSEM-LASSO based on one hundred bootstrap samples.[5] Both resampling methods are paired with the BIC as a fit index. For RegSEM with stability selection, we tried to vary the penalty types and the fit indices used for tuning the probability threshold π_{thr}. Four combinations are considered: (f) stability selection-LASSO-BIC, (g) stability selection-LASSO-AIC, (h) stability selection-ALASSO-BIC, and (i) stability selection-ALASSO-AIC. For stability selection, 50 RegSEM preruns are performed on bootstrap samples of the data to determine the range Λ for the regularization values in each case. The range Λ is then divided to 10 equal distant values of λ, and the selection probability at each of those λ values are obtained from 100 replications of RegSEM with degree of penalty fixed at this λ on bootstrap samples. For all the methods stated above, whenever RegSEM is performed, its optimal λ value is chosen based on the BIC (Jacobucci et al., 2016). The two studies consider different application scenarios. The first study considers selecting among multiple predictors of a latent variable, following the general form of Jacobucci et al.

[4] To perform five-fold cross validation, we randomly split the sample into five nonoverlapping groups. For each of the five iterations, we use four groups as training set to perform regularized SEM using the LASSO penalty, calculating the BIC on the remaining group with the model achieved at each λ value. We repeat this procedure for all five combinations, average the BIC values at each λ and obtain the standard error (standard deviation across folds) of the lowest averaged BIC values. As in regression, regularized SEM with LASSO penalty demonstrated high FPRs, thus we follow the convention of applying the 1se rule to select a larger λ value to lower the FPRs. This is achieved by not using the λ value corresponding to the lowest averaged BIC but using the largest λ value that has averaged BIC value not greater than the lowest averaged BIC plus one standard error.

[5] In line with the selection procedure of stability selection, the bootstrap method we are using here involves first computing the selection probability of each path/variable, then selecting a probability threshold to determine the final selection result. The probability threshold is tuned using the AIC. With this, we want to see whether the sophisticated stability selection procedure have an advantage over simply repeating and aggregating the regularization procedures on bootstrap samples.

(2019) to determine whether our proposed methods can improve upon the results of their simulation study. The data generating model is a multiple indicators multiple causes (MIMIC; Joreskog & Goldberger, 1975) model with one latent factor F_1. There are in total six indicators ($Y_1, \ldots, Y_6$), and 30 predictors of F_1. The penalty is applied to all 30 regression paths. For identification purposes, the variance of the latent factor is fixed at 1, whereas all other factor loadings and regression paths are freely estimated. The mean structure is not estimated in this model. We used such a model to represent the common cases where a large number of potential covariates are available but we lack strong a priori expectations. Bollen and Pearl (2013) stated that strong causal relationships are made by imposing zero coefficients or imposing zero covariances in the model. Such assumptions without proper justification or theoretical foundations undermine the causal relationship in the hypotheses (Shipley, 2016). Thus, exploratory approaches may be more appropriate than confirmatory approaches in such scenarios.

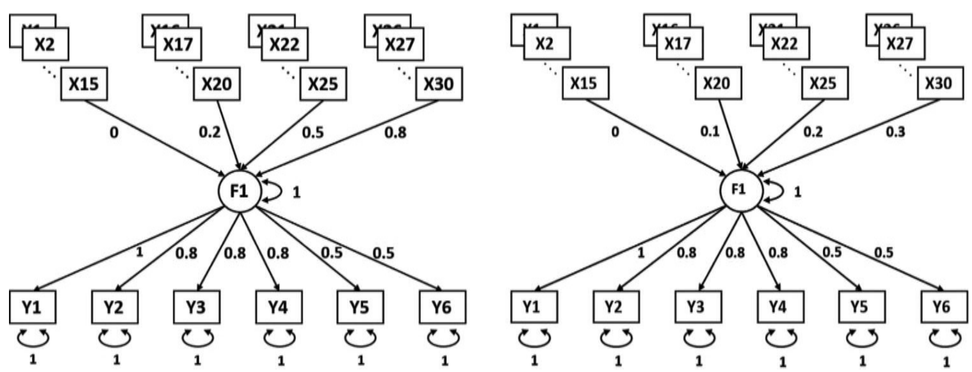
We tested two models with different effect sizes: the first model tested a wider range (0.2 to 0.8) of regression coefficients (Jacobucci et al., 2019), whereas the second model considers more realistic situations where all regression coefficients are relatively small in magnitude (0.1 to 0.3). The corresponding path diagrams are depicted in Figure 3. In both conditions, $X_1, \ldots, X_{15}$ are simulated to be uninformative "noise" variables, with coefficients simulated as 0; $X_{16}, \ldots, X_{20}$ are simulated to be predictors with relatively small effect sizes, having coefficients 0.2 in the large effect size condition and 0.1 in the small effect size condition; $X_{21}, \ldots, X_{25}$ are relatively medium effect size predictors, having coefficients 0.5 in the large effect size condition and 0.2 in the small effect size condition; $X_{26}, \ldots, X_{30}$ are relatively large effect size predictors, with coefficient 0.8 and 0.3 in the large effect size condition and small effect size condition, respectively. We tested across sample sizes of 75, 150, 400, 800, and 2,000. Our aim was to examine performance in small sample sizes (N = 75, 150) where variable selection may be a necessity given the high number of parameters, more typical sample sizes in applied research (N = 400, 800), as well as to examine expected asymptotic behavior (N = 2,000).

The second study considers selecting cross loadings. Asparouhov & Muthén (2009) stated the disadvantages of using CFA measurement modeling in SEM: the CFA approach of fixing many or all cross-loadings at zero may force a researcher to specify a more parsimonious model than is suitable for the data, leading to poor model fit, and the use of extensive model modification. Moreover, misspecification of zero loadings in CFA tends to give distorted factors. Those practices would weaken the believability and replicability of the final model. In such scenarios, exploratory approaches appear preferable to confirmatory approaches (Browne, 2001), and we see regularized SEM as a good tool given this. In this study, a two-factor model is generated with simulated parameters as depicted in Figure 4. Each factor has five large loadings of 0.8, three loadings of 0.6, and two loadings of 0.4. Although the loadings in solid lines are considered certain and are always included in the model, the paths in dashed lines are the cross loadings that we are uncertain about and are considered for removal to make the final model more parsimonious. The dash-dotted paths have a simulated loading of 0.3, although the dashed paths are the true "noise" which are simulated to have loadings of zero. Penalties are only put on those 14 paths. As we mentioned earlier, RegSEM when combined with LASSO penalty tends to give not enough penalization, so that the FPR are very high. Our goal here is to try to remove those true "noise" cross loadings, while keeping those nonzero loadings. In the data generating model, each factor is simulated to have a variance of 1 and the correlation of the two factors is simulated to be 0.3. For this study, sample sizes of 40, 50, 75, 150, 400, 800, and 2,000 are considered. Across each of the simulation condition, each cell is replicated 100 times.

Data for the two simulation studies were generated using R package `lavaan` (multivariate normal by default), all regularization related procedures were implemented through R package `regsem`. The function `cv_regsem` was used to perform regularized SEM across multiple penalty values. Stability selection was

Figure 3

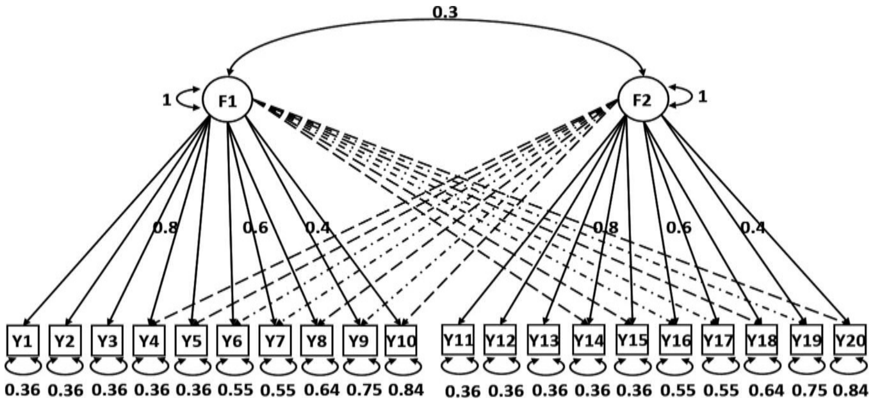
Path Diagram of the Data Generating Model for Simulation Study 1



Note. A multiple indicators multiple causes (MIMIC) model with 30 causes of one latent variable is simulated. Among those 30 causes, the first 15 are simulated noise variables, the other 15 variables are simulated to have relatively small, medium, and large (5 each) effect on the latent variable. Two settings are considered: the left panel follows from the design used in Jacobucci et al. (2019), and has larger effect sizes than the right panel; the right panel considers more realistic conditions, where all the regression coefficients are smaller than 0.3.

Figure 4

Path Diagram of the Data Generating Model for Simulation Study 2



Note. A two-factor model is simulated, with factor correlation of 0.3. Each factor has five large loadings of 0.8, three loadings of 0.6, and two loadings of 0.4, marked by solid line. Paths in dashed lines are the cross loadings that are considered for removal. The dashed paths are simulated to have zero loadings, and the dash-dotted paths have simulated loading of 0.3.

performed by function `stabsel`. A new function `stabsel_par` is also available in the package (v1.5.2), which performs the procedure in parallel to speed up the procedure when multiple cores are available. The simulation code can be found at https://osf.io/xdg3p/.

After obtaining the selection results, we focus our comparison on FPRs and FNRs across different sample sizes. We want to examine how high the FPRs are with RegSEM-LASSO and if the other methods could lower the FPRs, without raising up the FNRs by too much. Moreover, by considering selection results at larger sample sizes, we could see whether both FPRs and FNRs decrease to 0 as sample size gets larger, thereby getting a sense of the consistency of each methods. FPRs and FNRs are computed as follows:

$$FPR_i = \frac{Number\ of\ selected\ noise\ variables\ in\ group\ i}{Total\ number\ of\ noise\ variables\ in\ group\ i} \quad (2)$$

$$FNR_j = \frac{Number\ of\ omitted\ variables\ with\ simulated\ effect\ size\ j}{Total\ number\ of\ variables\ with\ simulated\ effect\ size\ j} \quad (3)$$

We summarize the FPRs and FNRs by group and/or effect size. For Simulation 1, there is only one group of noise variables, so no separation is needed for FPRs, but we simulated three different effect sizes, so there are three FNRs calculated. For Simulation 2, there are three groups of noise loadings: $Y4 \rightarrow F2$, $Y5 \rightarrow F2$, $Y14 \rightarrow F1$ and $Y15 \rightarrow F1$, each having relatively large simulated effect sizes (with simulated loading of 0.8) on the other factor; $Y8 \rightarrow F2$ and $Y18 \rightarrow F1$ have medium simulated effect sizes (with simulated loading of 0.6) on the other factor; $Y10 \rightarrow F2$ and $Y20 \rightarrow F1$ have relatively small simulated effect sizes (with simulated loading of 0.4) on the other factor. Similarly, for signal loadings, there are two groups: $Y6 \rightarrow F2$, $Y7 \rightarrow$ F2, $Y16 \rightarrow F1$ and $Y17 \rightarrow F1$ each having medium simulated effect size on the other factor; $Y9 \rightarrow F2$ and $Y19 \rightarrow F1$ have relatively small simulated effect size on the other factor. Those rates are then averaged within each group and across the 100 replications per simulation condition (i.e., the paths with same simulated loadings are treated as the same and averaged).

Besides TPRs and FPRs, we also analyze the bias and variance components of each method. We compute the average RMSE and average RB for each parameter estimates across the 100 replications for each simulation condition by:

$$RMSE_i = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}_{ir} - \beta_i\right)^2} \quad (4)$$

and

$$RB_i = \frac{1}{R}\sum_{r=1}^{R}\frac{\hat{\beta}_{ir} - \beta_i}{\beta_i}. \quad (5)$$

where R here is the number of replications (100), β is the true population parameter, and $\hat{\beta}_r$ is the sample estimate for β in the rth dataset. Those estimates are then averaged within each group of parameter with the same simulated condition and value in the data generating model. RMSE incorporates both bias and efficiency, making it a rough proxy for the overall accuracy of a given method. Comparing those RMSE values, we would be able to tell each method's ability of recovering the true population parameters. RB assesses the bias relative to the value of the population parameter. According to Muthén and colleagues (Muthén et al., 1987), values greater than 15% are considered problematic. For the noise variables which have simulated coefficient of 0, the standard bias $\left(\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}_{ir} - \beta_i\right)\right)$ is computed and averaged instead.

Results

The results of TPRs and FPRs of the two simulation studies are summarized in Tables 2 and 3 and Table 4, respectively.[6] Because the overall patterns of the results are very similar across the two studies, we give an overview of the results together for each of the nine methods considered.

As expected, the FPRs of MLE models falls around 0.05 in most cases. However, the high FNRs (for small effect size predictors

[6] Corresponding plots are available in online supplementary materials.

Table 2

Average False Positive Rates (FPR) and False Negative Rates (FNR) of Simulation Study 1, When the Effect Sizes of Causes in the MIMIC Model is Large

			FNR		
Method	Sample size	FPR	Small	Moderate	Large
MLE	75	0.127	0.634	0.076	0
	150	0.075	0.426	0.002	0
	250	0.068	0.204	0	0
	400	0.063	0.072	0	0
	800	0.051	0.002	0	0
	2000	0.045	0	0	0
RegSEM (LASSO)	75	0.459	0.306	0.008	0
	150	0.391	0.194	0	0
	250	0.432	0.050	0	0
	400	0.396	0.012	0	0
	800	0.405	0	0	0
	2000	0.393	0	0	0
RegSEM (adaptive LASSO)	75	0.174	0.552	0.056	0
	150	0.117	0.380	0.004	0
	250	0.110	0.204	0	0
	400	0.099	0.056	0	0
	800	0.047	0.004	0	0
	2000	0.001	0	0	0
Cross validated RegSEM (LASSO)	75	0.525	0.276	0.007	0
	150	0.455	0.166	0.002	0
	250	0.133	0.35	0	0
	400	0.041	0.196	0	0
	800	0.079	0.008	0	0
	2000	0.165	0	0	0
Bootstrap RegSEM (LASSO)	75	0.410	0.396	0.037	0
	150	0.193	0.274	0	0
	250	0.190	0.098	0	0
	400	0.164	0.026	0	0
	800	0.161	0	0	0
	2000	0.151	0	0	0
Stability selection (LASSO)—BIC	75	0.107	0.676	0.092	0
	150	0.056	0.532	0.002	0
	250	0.039	0.328	0	0
	400	0.027	0.124	0	0
	800	0.017	0.006	0	0
	2000	0.023	0	0	0
Stability selection (adaptive LASSO)—BIC	75	0.106	0.667	0.079	0
	150	0.037	0.544	0.004	0
	250	0.030	0.372	0	0
	400	0.018	0.142	0	0
	800	0.009	0.008	0	0
	2000	0.003	0	0	0
Stability selection (LASSO)—AIC	75	0.277	0.508	0.036	0
	150	0.203	0.262	0.002	0
	250	0.191	0.092	0	0
	400	0.169	0.020	0	0
	800	0.160	0	0	0
	2000	0.144	0	0	0
Stability selection (adaptive LASSO)—AIC	75	0.267	0.464	0.028	0
	150	0.197	0.268	0.002	0
	250	0.189	0.104	0	0
	400	0.162	0.024	0	0
	800	0.078	0.002	0	0
	2000	0.003	0	0	0

Note. MIMIC = multiple indicators multiple causes; MLE = maximum likelihood estimation; RegSEM = regularized structural equation modeling; LASSO = least absolute shrinkage and selection operator; BIC = Bayesian information criterion; AIC = Akaike information criterion.

in simulation 1, and for cross loading detections in Simulation 2) shows that the power could be an issue. That is, paths with true signals are often disregarded, especially for cases with small sample sizes.

Similar to what was found in previous studies (Jacobucci et al., 2016), RegSEM-LASSO displays high FPRs, while having relatively low FNRs. This means that RegSEM-LASSO is able to include most signal variables/paths, however a large number of noise variables/paths are included at the same time. With respect to consistency, although the FNRs could be reduced to nearly 0 when sample size gets large enough, the FPRs stay high, demonstrating the two problems of the LASSO, namely, high FPRs and inconsistency, as previous studies have already noted.

Table 3
Average False Positive Rates (FPR) and False Negative Rates (FNR) of Simulation Study 1, When the Effect Sizes of Causes in the MIMIC Model is Small

Method	Sample size	FPR	FNR Small	FNR Moderate	FNR Large
MLE	75	0.133	0.780	0.684	0.390
	150	0.095	0.798	0.456	0.148
	250	0.057	0.704	0.222	0.024
	400	0.053	0.562	0.074	0.002
	800	0.050	0.338	0.004	0
	2000	0.050	0.030	0	0
RegSEM (LASSO)	75	0.347	0.568	0.404	0.146
	150	0.207	0.620	0.268	0.080
	250	0.162	0.558	0.152	0.008
	400	0.172	0.404	0.044	0.004
	800	0.216	0.174	0.002	0
	2000	0.266	0.012	0	0
RegSEM (adaptive LASSO)	75	0.094	0.840	0.730	0.490
	150	0.064	0.838	0.586	0.256
	250	0.073	0.690	0.226	0.020
	400	0.075	0.548	0.092	0
	800	0.041	0.368	0.010	0
	2000	0.002	0.260	0	0
Cross validated RegSEM (LASSO)	75	0.438	0.476	0.31	0.129
	150	0.322	0.545	0.232	0.071
	250	0.202	0.594	0.188	0.012
	400	0.024	0.73	0.166	0.004
	800	0.018	0.672	0.046	0
	2000	0.049	0.074	0	0
Bootstrap RegSEM (LASSO)	75	0.262	0.634	0.515	0.264
	150	0.177	0.670	0.306	0.096
	250	0.151	0.554	0.136	0.012
	400	0.146	0.404	0.024	0.002
	800	0.158	0.180	0	0
	2000	0.156	0.010	0	0
Stability selection (LASSO)—BIC	150	0.070	0.872	0.762	0.506
	250	0.043	0.884	0.608	0.236
	400	0.024	0.826	0.358	0.050
	800	0.013	0.764	0.168	0.002
	2000	0.007	0.532	0.012	0
	2000	0.007	0.122	0	0
Stability selection (adaptive LASSO)—BIC	75	0.073	0.870	0.770	0.500
	150	0.044	0.878	0.602	0.226
	250	0.018	0.820	0.358	0.042
	400	0.014	0.770	0.160	0.004
	800	0.008	0.560	0.016	0
	2000	0.003	0.184	0	0
Stability selection (LASSO)—AIC	75	0.259	0.672	0.514	0.242
	150	0.197	0.642	0.278	0.084
	250	0.178	0.506	0.118	0.010
	400	0.172	0.368	0.016	0
	800	0.157	0.178	0	0
	2000	0.153	0.01	0	0
Stability selection (adaptive LASSO)—AIC	75	0.294	0.630	0.488	0.246
	150	0.174	0.648	0.318	0.100
	250	0.157	0.528	0.116	0.008
	400	0.150	0.390	0.020	0.002
	800	0.081	0.282	0.002	0
	2000	0.003	0.184	0	0

Note. MIMIC = multiple indicators multiple causes; MLE = maximum likelihood estimation; RegSEM = regularized structural equation modeling; LASSO = least absolute shrinkage and selection operator; BIC = Bayesian information criterion; AIC = Akaike information criterion.

Table 4
Average False Positive Rates (FPR) and False Negative Rates (FNR) of Simulation Study 2

Method	Sample size	FPR Y_4, Y_5, Y_{14}, Y_{15}	FPR Y_8, Y_{18}	FPR Y_{10}, Y_{20}	FNR Y_6, Y_7, Y_{16}, Y_{17}	FNR Y_9, Y_{19}
MLE	40	0.068	0.075	0.045	0.438	0.500
	50	0.075	0.090	0.035	0.358	0.440
	75	0.048	0.075	0.055	0.220	0.300
	150	0.048	0.035	0.055	0.020	0.035
	250	0.040	0.060	0.055	0	0
	400	0.055	0.050	0.055	0	0
	800	0.040	0.030	0.060	0	0
	2,000	0.045	0.055	0.065	0	0
RegSEM (LASSO)	40	0.462	0.365	0.295	0.188	0.265
	50	0.482	0.295	0.315	0.123	0.180
	75	0.348	0.285	0.205	0.073	0.175
	150	0.392	0.220	0.210	0.005	0.010
	250	0.342	0.285	0.230	0	0
	400	0.365	0.215	0.210	0	0
	800	0.285	0.165	0.120	0	0
	2,000	0.298	0.145	0.160	0	0
RegSEM (adaptive LASSO)	40	0.110	0.125	0.085	0.422	0.495
	50	0.110	0.120	0.070	0.360	0.455
	75	0.100	0.105	0.090	0.202	0.320
	150	0.100	0.090	0.120	0.020	0.030
	250	0.068	0.080	0.060	0	0
	400	0.052	0.050	0.075	0	0
	800	0.002	0.010	0.010	0	0
	2,000	0.008	0.025	0.025	0	0
Cross validated RegSEM (LASSO)	40	0.613	0.581	0.5	0.11	0.14
	50	0.655	0.543	0.495	0.06	0.109
	75	0.524	0.462	0.398	0.038	0.097
	150	0.185	0.055	0.04	0.025	0.12
	250	0.162	0.11	0.03	0.007	0.03
	400	0.182	0.09	0.075	0	0
	800	0.195	0.1	0.06	0	0
	2,000	0.208	0.085	0.09	0	0
Bootstrap RegSEM (LASSO)	40	0.222	0.185	0.160	0.295	0.370
	50	0.215	0.172	0.136	0.194	0.288
	75	0.180	0.170	0.140	0.100	0.185
	150	0.205	0.130	0.145	0.007	0.010
	250	0.140	0.170	0.110	0	0
	400	0.195	0.125	0.160	0	0
	800	0.158	0.150	0.095	0	0
	2,000	0.158	0.105	0.130	0	0
Stability selection (LASSO)—BIC	40	0.118	0.080	0.075	0.425	0.510
	50	0.088	0.070	0.050	0.368	0.505
	75	0.058	0.065	0.040	0.235	0.360
	150	0.035	0.015	0.035	0.027	0.100
	250	0.025	0.040	0.020	0	0.015
	400	0.020	0.005	0.015	0	0
	800	0.008	0.010	0.015	0	0
	2,000	0.010	0.005	0.005	0	0
Stability selection (adaptive LASSO)—BIC	40	0.080	0.085	0.070	0.448	0.525
	50	0.080	0.080	0.040	0.397	0.495
	75	0.043	0.060	0.040	0.235	0.345
	150	0.023	0.015	0.030	0.027	0.095
	250	0.018	0.040	0.020	0	0.015
	400	0.018	0.005	0.015	0	0
	800	0.000	0.010	0.015	0	0
	2000	0.003	0	0	0	0

(*table continues*)

Table 4 (*continued*)

Method	Sample size	FPR			FNR	
		Y_4, Y_5, Y_{14}, Y_{15}	Y_8, Y_{18}	Y_{10}, Y_{20}	Y_6, Y_7, Y_{16}, Y_{17}	Y_9, Y_{19}
Stability selection (LASSO)—AIC	40	0.192	0.192	0.154	0.359	0.410
	50	0.237	0.132	0.158	0.178	0.329
	75	0.184	0.105	0.132	0.099	0.158
	150	0.211	0.145	0.184	0.013	0
	250	0.158	0.171	0.158	0	0
	400	0.184	0.132	0.158	0	0
	800	0.151	0.158	0.105	0	0
	2000	0.171	0.118	0.237	0	0
Stability selection (adaptive LASSO)—AIC	40	0.203	0.175	0.185	0.273	0.320
	50	0.193	0.170	0.170	0.195	0.250
	75	0.158	0.170	0.185	0.090	0.140
	150	0.158	0.150	0.175	0.008	0.005
	250	0.123	0.170	0.105	0	0
	400	0.160	0.140	0.170	0	0
	800	0.055	0.035	0.060	0	0
	2000	0.003	0	0	0	0

Note. MLE = maximum likelihood estimation; RegSEM = regularized structural equation modeling; LASSO = least absolute shrinkage and selection operator; BIC = Bayesian information criterion; AIC = Akaike information criterion.

When the ALASSO is used instead, parameters with smaller MLE estimates are given larger penalties, thus the FPRs becomes lower than RegSEM-LASSO. However, similar to using MLE, the FNRs becomes high for small effect paths when sample size is small. One nice feature of RegSEM-ALASSO is that the FPRs and FNRs continues to drop as sample size gets larger. When sample size is large, the FPRs could get to lower than 0.05 (i.e., better than MLE), with FNR close to 0. This shows that, ALASSO gives consistent selection results under the two simulation conditions.

As for fivefold cross-validation using the LASSO and the 1se rule, the FPRs are actually higher than RegSEM-LASSO when sample size is small. This may be because cross-validation used the averaged out-of-sample BICs and could select a lower λ value than just applying LASSO, even if 1se rule has been applied. Besides, there is no clear pattern of decrease in both FPRs and FNRs as sample size gets larger. In Table 2, we see a large decrease (0.455 to 0.133) in averaged FPRs from sample size 150 to 250. However, the FNRs of the relative small effect predictors grows from 0.166 to 0.350. Similar patterns are also observed in Table 3. Though cross-validation with 1se rule appears to perform better than RegSEM-LASSO at larger sample sizes, the resulting pattern of this method is unclear. Moreover, because sample splitting is required in the cross-validation process, it may not be applicable when sample size is small, hence high rates of nonconvergence may occur in small sample sizes.

Bootstrapping did help decrease the FPRs, slightly sacrificing the FNRs, but again, the selection results are inconsistent with the nondecreasing FPRs as sample size gets larger. Given the larger computational resources needed, the results do not seem promising.

Stability selection-LASSO-BIC and stability selection-ALASSO-BIC both give similar results as MLE when sample size is small, and they perform better than MLE in terms of FPRs when sample size is large. But again, like MLE, the FNRs is very high for small samples, indicating that the selection is too conservative. When AIC is used for determining π_{thr} instead, a smaller penalty is incurred for model complexity, and thus gives more balanced FPRs and FNRs, as what we are targeting at (e.g., FPRs and FNRs as within the dash-dotted region in right panel of Figure 1). The use of stability selection-ALASSO-AIC affords the nice property of decreasing both FPRs and FNRs as sample size gets larger, showing consistency in selection results.

Besides analyzing the simulation results descriptively, we also performed regression analyses using the simulation outcomes as dependent variables, and simulation condition as independent variables. The results are summarized in Tables 5–7. Compared to RegSEM-LASSO, most of the methods considered in the simulation studies were able to significantly decrease FPRs, while not increasing FNRs significantly in the larger effect size case of simulation Study 1 and simulation Study 2. However, for the more realistic case of the small effect size MIMIC model in simulation Study 1, stability selection-ALASSO-AIC is the only one that keeps this pattern, which further confirms the advantage of our proposed method of combining stability selection with regularized SEM.

Besides looking at the selection results of each method, we also studied the parameter estimates in terms of bias and variance. The results of RMSE and RB across the two simulations are summarized in Supplemental Tables S1 and S2 of the online supplementary materials. Again, the patterns of each methods' performance turn out similar across the two simulation studies, so we give a brief overview of the results together here. First, regularization methods such as the LASSO achieve variable selection by biasing some of the parameter estimates to 0. Given this, all the methods considered in the simulations except MLE used the two-step "relaxed LASSO" procedure of first performing variable selection than refitting the model with only the selected variables/paths to reduce the bias. Thus, to make the comparison fair, we calculated the RMSE and RB for the MLE method in three different ways: (a) MLE1 used the raw SEM output using MLE; (b) MLE2 treated those nonsignificant coefficients as 0; (c) MLE3 used the same two-step procedure, which removes those paths with nonsignificant coefficient estimates, and refitted the model with only those significant ones. It turns out that this extra step in MLE3 helps lower the RMSE and RB in both simulation studies. This is not the

Table 5
Regression Results of Simulation Study 1, When the Effect Sizes of Causes in the MIMIC Model is Large

Predictor	b	b 95% CI [*LL*, *UL*]	sr^2	sr^2 95% CI [*LL*, *UL*]	Fit
Regression results using FPR as the criterion					
(Intercept)	0.45**	[0.39, 0.51]			
Sample size	−0.00**	[−0.00, −0.00]	.10	[.01, .18]	
MLE	−0.34**	[−0.42, −0.26]	.37	[.19, .54]	
RegSEM (ALASSO)	−0.32**	[−0.40, −0.24]	.33	[.16, .49]	
CV RegSEM (LASSO)	−0.24**	[−0.32, −0.16]	.18	[.06, .30]	
Bootstrap RegSEM (LASSO)	−0.20**	[−0.28, −0.12]	.13	[.03, .23]	
Stability selection (LASSO)—BIC	−0.37**	[−0.45, −0.29]	.43	[.24, .61]	
Stability selection (ALASSO)—BIC	−0.38**	[−0.46, −0.30]	.45	[.27, .64]	
Stability selection (LASSO)—AIC	−0.22**	[−0.30, −0.14]	.16	[.04, .27]	
Stability selection (ALASSO)—AIC	−0.26**	[−0.34, −0.18]	.22	[.08, .35]	
					R^2 = .783** 95% CI [.59, .82]
Regression results using FNR (small) as the criterion					
(Intercept)	0.22**	[0.07, 0.36]			
Sample size	−0.00**	[−0.00, −0.00]	.33	[.14, .51]	
MLE	0.13	[−0.07, 0.33]	.02	[−.03, .06]	
RegSEM (ALASSO)	0.11	[−0.09, 0.30]	.01	[−.03, .05]	
CV RegSEM (LASSO)	0.42**	[0.22, 0.62]	.18	[.03, .32]	
Bootstrap RegSEM (LASSO)	0.04	[−0.16, 0.24]	.00	[−.01, .01]	
Stability selection (LASSO)—BIC	0.18	[−0.01, 0.38]	.03	[−.03, .10]	
Stability selection (ALASSO)—BIC	0.20	[−0.00, 0.39]	.04	[−.03, .10]	
Stability selection (LASSO)—AIC	0.05	[−0.14, 0.25]	.00	[−.02, .02]	
Stability selection (ALASSO)—AIC	0.05	[−0.15, 0.25]	.00	[−.01, .02]	
					R^2 = .584** 95% CI [.28, .65]
Regression results using FNR (medium) as the criterion					
(Intercept)	0.02	[−0.02, 0.05]			
Sample size	−0.00**	[−0.00, −0.00]	.07	[−.02, .16]	
MLE	0.01	[−0.04, 0.06]	.00	[−.01, .02]	
RegSEM (ALASSO)	0.01	[−0.04, 0.06]	.00	[−.01, .01]	
CV RegSEM (LASSO)	0.16**	[0.11, 0.20]	.35	[.17, .53]	
Bootstrap RegSEM (LASSO)	0.00	[−0.04, 0.05]	.00	[−.01, .01]	
Stability selection (LASSO)—BIC	0.01	[−0.03, 0.06]	.00	[−.01, .02]	
Stability selection (ALASSO)—BIC	0.01	[−0.03, 0.06]	.00	[−.01, .02]	
Stability selection (LASSO)—AIC	0.00	[−0.04, 0.05]	.00	[−.01, .01]	
Stability selection (ALASSO)—AIC	0.00	[−0.04, 0.05]	.00	[−.00, .00]	
					R^2 = .641** 95% CI [.36, .70]
Regression results using FNR (large) as the criterion					
(Intercept)	0.00	[−0.01, 0.02]			
Sample size	−0.00	[−0.00, 0.00]	.02	[−.04, .09]	
MLE	0.00	[−0.02, 0.02]	.00	[.00, .00]	
RegSEM (ALASSO)	0.00	[−0.02, 0.02]	.00	[.00, .00]	
CV RegSEM (LASSO)	0.04**	[0.02, 0.06]	.19	[.01, .36]	
Bootstrap RegSEM (LASSO)	−0.00	[−0.02, 0.02]	.00	[.00, .00]	
Stability selection (LASSO)—BIC	0.00	[−0.02, 0.02]	.00	[.00, .00]	
Stability selection (ALASSO)—BIC	0.00	[−0.02, 0.02]	.00	[.00, .00]	
Stability selection (LASSO)—AIC	0.00	[−0.02, 0.02]	.00	[.00, .00]	
Stability selection (ALASSO)—AIC	0.00	[−0.02, 0.02]	.00	[.00, .00]	
					R^2 = .353* 95% CI [.02, .43]

Note. MIMIC = multiple indicators multiple causes; FPR = false positive rate; FNR = false negative rate; MLE = maximum likelihood estimation; RegSEM = regularized structural equation modeling; LASSO = least absolute shrinkage and selection operator; BIC = Bayesian information criterion; AIC = Akaike information criterion; CI = confidence interval. The dependent variables are FPRs and FNRs (corresponding to small [0.2], medium [0.5], and large [0.8] simulated parameters), independent variables are sample size and methods. Reference level of methods is RegSEM LASSO.
* $p < .05$. ** $p < .01$.

Table 6

Regression Results of Simulation Study 1, When the Effect Sizes of Causes in the MIMIC Model is Small

Predictor	b	b 95% CI [*LL*, *UL*]	sr^2	sr^2 95% CI [*LL*, *UL*]	Fit
Regression results using FPR as the criterion					
(Intercept)	0.26^{**}	[0.20, 0.32]			
Sample size	-0.00^{**}	[−0.00, −0.00]	.10	[−.01, .20]	
MLE	-0.16^{**}	[−0.24, −0.08]	.13	[.01, .26]	
RegSEM (ALASSO)	-0.17^{**}	[−0.25, −0.09]	.16	[.03, .30]	
CV RegSEM (LASSO)	−0.05	[−0.13, 0.03]	.02	[−.03, .06]	
Bootstrap RegSEM (LASSO)	−0.05	[−0.13, 0.03]	.02	[−.03, .06]	
Stability selection (LASSO)—BIC	-0.20^{**}	[−0.28, −0.12]	.23	[.07, .38]	
Stability selection (ALASSO)—BIC	-0.20^{**}	[−0.28, −0.12]	.23	[.07, .38]	
Stability selection (LASSO)—AIC	−0.04	[−0.12, 0.04]	.01	[−.02, .04]	
Stability selection (ALASSO)—AIC	-0.09^{*}	[−0.17, −0.01]	.04	[−.03, .11]	
					$R^2 = .613^{**}$ 95% CI [.32, .67]
Regression results using FNR (small) as the criterion					
(Intercept)	0.59^{**}	[0.50, 0.68]			
Sample size	-0.00^{**}	[−0.00, −0.00]	.69	[.52, .87]	
MLE	0.15^{*}	[0.02, 0.27]	.02	[−.01, .04]	
RegSEM (ALASSO)	0.20^{**}	[0.08, 0.32]	.03	[−.01, .07]	
CV RegSEM (LASSO)	0.13^{*}	[0.00, 0.25]	.01	[−.01, .04]	
Bootstrap RegSEM (LASSO)	0.02	[−0.10, 0.14]	.00	[−.00, .00]	
Stability selection (LASSO)—BIC	0.28^{**}	[0.15, 0.40]	.06	[.01, .12]	
Stability selection (ALASSO)—BIC	0.29^{**}	[0.17, 0.41]	.07	[.01, .13]	
Stability selection (LASSO)—AIC	0.01	[−0.12, 0.13]	.00	[−.00, .00]	
Stability selection (ALASSO)—AIC	0.05	[−0.07, 0.18]	.00	[−.01, .01]	
					$R^2 = .864^{**}$ 95% CI [.74, .89]
Regression results using FNR (medium) as the criterion					
(Intercept)	0.28^{**}	[0.12, 0.44]			
Sample size	-0.00^{**}	[−0.00, −0.00]	.40	[.20, .60]	
MLE	0.10	[−0.12, 0.31]	.01	[−.03, .05]	
RegSEM (ALASSO)	0.13	[−0.09, 0.35]	.02	[−.03, .07]	
CV RegSEM (LASSO)	0.01	[−0.20, 0.23]	.00	[−.00, .00]	
Bootstrap RegSEM (LASSO)	0.02	[−0.20, 0.24]	.00	[−.01, .01]	
Stability selection (LASSO)—BIC	0.17	[−0.04, 0.39]	.03	[−.04, .10]	
Stability selection (ALASSO)—BIC	0.17	[−0.04, 0.39]	.03	[−.04, .10]	
Stability selection (LASSO)—AIC	0.01	[−0.21, 0.23]	.00	[−.00, .00]	
Stability selection (ALASSO)—AIC	0.01	[−0.20, 0.23]	.00	[−.00, .00]	
					$R^2 = .485^{**}$ 95% CI [.15, .56]
Regression results using FNR (large) as the criterion					
(Intercept)	0.10	[−0.01, 0.21]			
Sample size	-0.00^{**}	[−0.00, −0.00]	.21	[.02, .40]	
MLE	0.05	[−0.09, 0.20]	.01	[−.03, .05]	
RegSEM (ALASSO)	0.09	[−0.06, 0.24]	.02	[−.05, .09]	
CV RegSEM (LASSO)	−0.00	[−0.15, 0.14]	.00	[−.00, .00]	
Bootstrap RegSEM (LASSO)	0.02	[−0.13, 0.17]	.00	[−.02, .02]	
Stability selection (LASSO)—BIC	0.09	[−0.06, 0.24]	.03	[−.05, .10]	
Stability selection (ALASSO)—BIC	0.09	[−0.06, 0.24]	.02	[−.05, .09]	
Stability selection (LASSO)—AIC	0.02	[−0.13, 0.16]	.00	[−.01, .01]	
Stability selection (ALASSO)—AIC	0.02	[−0.13, 0.17]	.00	[−.01, .02]	
					$R^2 = .285$ 95% CI [.00, .36]

Note. MIMIC = multiple indicators multiple causes; FPR = false positive rate; FNR = false negative rate; MLE = maximum likelihood estimation; RegSEM = regularized structural equation modeling; LASSO = least absolute shrinkage and selection operator; BIC = Bayesian information criterion; AIC = Akaike information criterion; CI = confidence interval. The dependent variables are FPRs and FNRs (corresponding to small [0.1], medium [0.2], and large [0.3] simulated parameters), independent variables are sample size and methods. Reference level of methods is RegSEM LASSO. A significant *b*-weight indicates the semi-partial correlation is also significant. *b* represents unstandardized regression weights. sr^2 represents the semi-partial correlation squared. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.
* $p <. 05$. ** $p <. 01$.

Table 7
Regression Results of Simulation Study 2

Predictor	*b*	*b* 95% CI [*LL*, *UL*]	sr^2	sr^2 95% CI [*LL*, *UL*]	Fit
Regression results using FPR [Avg.(Y4, Y5, Y14, Y15)] as the criterion					
(Intercept)	0.40**	[0.34, 0.45]			
Sample size	−0.00**	[−0.00, −0.00]	.06	[−.00, .11]	
MLE	−0.32**	[−0.40, −0.24]	.28	[.14, .41]	
RegSEM (ALASSO)	−0.30**	[−0.38, −0.23]	.25	[.12, .38]	
CV RegSEM (LASSO)	−0.03	[−0.11, 0.04]	.00	[−.01, .01]	
Bootstrap RegSEM (LASSO)	−0.19**	[−0.26, −0.11]	.10	[.02, .17]	
Stability selection (LASSO)—BIC	−0.33**	[−0.40, −0.25]	.29	[.15, .43]	
Stability selection (ALASSO)—BIC	−0.34**	[−0.41, −0.26]	.31	[.17, .46]	
Stability selection (LASSO)—AIC	−0.19**	[−0.26, −0.11]	.09	[.02, .17]	
Stability selection (ALASSO)—AIC	−0.24**	[−0.32, −0.16]	.16	[.06, .26]	
					R^2 = .756** 95% CI [.60, .80]
Regression results using FPR [Avg.(Y8, Y18)] as the criterion					
(Intercept)	0.27**	[0.22, 0.33]			
Sample size	−0.00**	[−0.00, −0.00]	.10	[.01, .20]	
MLE	−0.19**	[−0.27, −0.11]	.15	[.04, .27]	
RegSEM (ALASSO)	−0.17**	[−0.25, −0.09]	.13	[.02, .23]	
CV RegSEM (LASSO)	0.01	[−0.07, 0.09]	.00	[−.00, .00]	
Bootstrap RegSEM (LASSO)	−0.10*	[−0.18, −0.02]	.04	[−.02, .10]	
Stability selection (LASSO)—BIC	−0.21**	[−0.29, −0.13]	.19	[.06, .32]	
Stability selection (ALASSO)—BIC	−0.21**	[−0.29, −0.13]	.19	[.06, .32]	
Stability selection (LASSO)—AIC	−0.10*	[−0.18, −0.02]	.05	[−.02, .11]	
Stability selection (ALASSO)—AIC	−0.12**	[−0.20, −0.04]	.06	[−.01, .14]	
					R^2 = .577** 95% CI [.34, .64]
Regression results using FPR [Avg.(Y10, Y20)] as the criterion					
(Intercept)	0.24**	[0.18, 0.29]			
Sample size	−0.00**	[−0.00, −0.00]	.06	[−.02, .13]	
MLE	−0.16**	[−0.24, −0.09]	.14	[.02, .26]	
RegSEM (ALASSO)	−0.15**	[−0.23, −0.07]	.12	[.01, .23]	
CV RegSEM (LASSO)	−0.01	[−0.08, 0.07]	.00	[−.01, .01]	
Bootstrap RegSEM (LASSO)	−0.08*	[−0.16, −0.01]	.04	[−.02, .10]	
Stability selection (LASSO)—BIC	−0.19**	[−0.26, −0.11]	.18	[.05, .32]	
Stability selection (ALASSO)—BIC	−0.19**	[−0.27, −0.11]	.19	[.05, .33]	
Stability selection (LASSO)—AIC	−0.06	[−0.13, 0.02]	.02	[−.03, .06]	
Stability selection (ALASSO)—AIC	−0.09*	[−0.16, −0.01]	.04	[−.02, .10]	
					R^2 = .513** 95% CI [.26, .58]
Regression results using FNR [Avg.(Y6, Y7, Y16, Y17)] as the criterion					
(Intercept)	0.10*	[0.00, 0.19]			
Sample size	−0.00**	[−0.00, −0.00]	.20	[.04, .36]	
MLE	0.08	[−0.05, 0.21]	.02	[−.03, .07]	
RegSEM (ALASSO)	0.08	[−0.05, 0.20]	.02	[−.03, .07]	
CV RegSEM (LASSO)	−0.02	[−0.15, 0.11]	.00	[−.01, .01]	
Bootstrap RegSEM (LASSO)	0.03	[−0.10, 0.15]	.00	[−.02, .02]	
Stability selection (LASSO)—BIC	0.08	[−0.04, 0.21]	.02	[−.04, .08]	
Stability selection (ALASSO)—BIC	0.09	[−0.04, 0.22]	.02	[−.04, .08]	
Stability selection (LASSO)—AIC	0.03	[−0.10, 0.16]	.00	[−.02, .02]	
Stability selection (ALASSO)—AIC	0.02	[−0.11, 0.15]	.00	[−.01, .02]	
					R^2 = .271* 95% CI [.01, .35]
Regression results using FNR [Avg.(Y9, Y19)] as the criterion					
(Intercept)	0.14*	[0.03, 0.26]			
Sample size	−0.00**	[−0.00, −0.00]	.23	[.07, .40]	
MLE	0.08	[−0.08, 0.24]	.01	[−.03, .05]	
RegSEM (ALASSO)	0.08	[−0.07, 0.24]	.01	[−.03, .06]	
CV RegSEM (LASSO)	−0.02	[−0.17, 0.14]	.00	[−.01, .01]	

(*table continues*)

Table 7 (*continued*)

Predictor	b	b 95% CI [*LL*, *UL*]	sr^2	sr^2 95% CI [*LL*, *UL*]	Fit
Bootstrap RegSEM (LASSO)	0.03	[−0.13, 0.18]	.00	[−.01, .02]	
Stability selection (LASSO)—BIC	0.11	[−0.05, 0.26]	.02	[−.03, .08]	
Stability selection (ALASSO)—BIC	0.11	[−0.05, 0.26]	.02	[−.03, .08]	
Stability selection (LASSO)—AIC	0.03	[−0.12, 0.19]	.00	[−.02, .02]	
Stability selection (ALASSO)—AIC	0.01	[−0.15, 0.17]	.00	[−.01, .01]	
					$R^2 = .299^{**}$ 95% CI [.03, .38]

Note. The dependent variables are false positive rates (FPRs) and false negative rates (FNRs), independent variables are sample size and methods. MLE = maximum likelihood estimation; RegSEM = regularized structural equation modeling; LASSO = least absolute shrinkage and selection operator; BIC = Bayesian information criterion; AIC = Akaike information criterion. A significant b-weight indicates the semi-partial correlation is also significant. b represents unstandardized regression weights. sr^2 represents the semi-partial correlation squared. *LL* and *UL* indicate the lower and upper limits of a confidence interval (CI), respectively. Reference level of methods is RegSEM LASSO.
$^{*}p < .05$. $^{**}p < .01$.

usual practice for CFA, but when the modeling procedure tends to be more exploratory, this two-step procedure may be preferable even for MLE. Second, regularized SEM either with LASSO or ALASSO, and cross-validated regularized SEM appears to have larger RMSE and RB values than the other computationally intensives methods. Further, the large RMSE and RB values do not decrease as sample size grows. This is in line with the inconsistent behavior of the LASSO, that is, the selection results do not improve as sample size grows, so the bias and variance both remain large. Lastly, the stability selection methods achieve levels of RMSE and RB value that are close to that from MLE, and those values decrease as sample size increases. This result was in line with our expectations that improved selection of variables/paths would result in lower RMSE and RB.

To summarize, our simulation results evidence that the high FPR and inconsistency problems of the LASSO preserve in regularized SEM, and that the use of stability selection could significantly lower the FPRs without raising the FNRs too much. And the resulting models have similar levels of RMSE and RB as using MLE on the parameter estimates. Based on the simulation results of the current two studies, stability selection-ALASSO-AIC shows benefit over both MLE and regularized SEM and provides consistent selection results. It is thus preferred when computational resources permit. For models with similar size to our simulated models or Empirical Example 2 (detailed later), the procedure of stability selection would take hours. But for larger models, such as in our Empirical Example 1 (detailed later), a single run of regularized SEM itself could take an hour, making the stability selection procedure several days. Depending on the number of cores used, the process could be sped up using function `stabsel_par`, which performs the stability selection procedure on different bootstrap samples in parallel. With limited computational resources, it may be more appropriate to just use RegSEM-ALASSO instead of RegSEM-LASSO. However, it is possible in real applications that the full model is unidentified, and thus MLE estimates are not available. In such scenarios, ALASSO is not applicable, and researchers could consider using stability selection-LASSO-BIC.

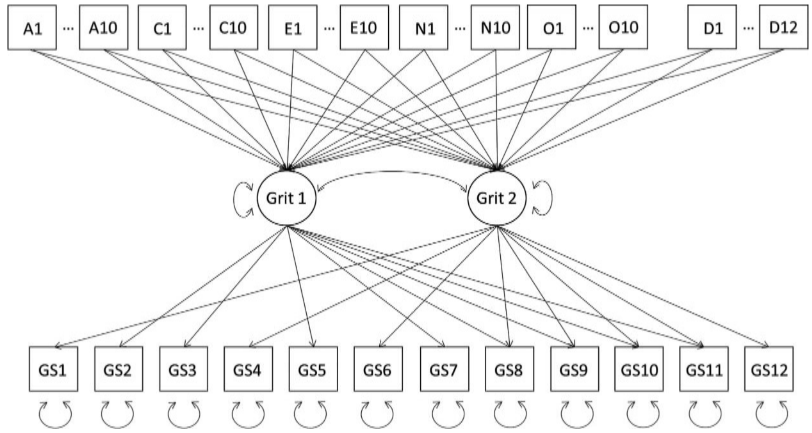
Empirical Example 1

To demonstrate the use of stability selection with regularized SEM, our first empirical example makes use of openly available data from Open Psychometrics (https://openpsychometrics.org/_rawdata/), which included responses from 4,270 participants who completed online surveys. The analyzed data were restricted to a random sample of 1,000 participants who had complete data on the Grit Scale (Duckworth et al., 2007) and the International Personality Item Pool Big-5 Personality Inventory (Goldberg, 1999). In this model, we have 10 predictors each from the Extraversion, Conscientiousness, Openness, Neuroticism, and Agreeableness facets of the Big Five theory of personality. In addition, we have measures of education, religion, age, race/ethnicity, and family (12 total). We specified a two-factor CFA model for the grit indicators, while including all 62 predictors of both latent factors, resulting in a MIMIC model (displayed in Figure 5). Starting with the goal to use this large set of predictors to predict two facets of grit, we see two possible motivations for the application of regularization, and specifically a form of regularization that performs variable (parameter) selection. The first is that there are 124 regression coefficients, making a sample size of 1,000 a bare minimum for accurate parameter estimation. Second is that we could have a large enough sample size but want to explicitly perform variable selection to simplify our inferences, or use the results of the variable selection to inform the specification of a simpler model, or administration of a reduced set of items in a follow up study. With this as our motivation, we applied stability selection-ALASSO with the AIC to select among the regression coefficients.

The 50 regularized SEM preruns determined a regularization range from 0.11 to 0.215. Based on AIC of the post selection models, the final threshold is chosen to be 0.85. That is, those paths with a maximum selection probability greater than 0.85 within the range [0.11, 0.215] are selected for the final model. With this, out of the 124 penalized paths, 72 were removed. The selection results are displayed in Table 8. Of the 10 predictors from each domain of the Big Five, only six Agreeableness, nine Openness, six Extraversion, and nine Neuroticism regression paths remained in the model. In contrast, a majority of Conscientiousness paths (14) remained in the model, which is in line with prior research on the similarity between Conscientiousness and Grit (Duckworth et al., 2007).

In addition, the selection procedure resulted in a sparser model with similar fit. Before selection, the model with all predictors included had a CFI of 0.832, TLI of 0.796, and

Figure 5
Path Diagram of Empirical Example 2



Note. A two-factor multiple indicators multiple causes (MIMIC) model is specified. Penalty is put on all paths from the indicators to the two latent factors. In this model, we have 10 predictors each from the Extraversion (E1 to E10), Conscientiousness (C1 to C10), Openness (O1 to O10), Neuroticism (N1 to N10), and Agreeableness (A1 to A10) facets of the Big Five theory of personality. In addition, we have 12 demographic variables (D1 to D12).

RMSEA of 0.037 (90% CI 0.035, 0.039). After removing the 72 predictors, the new model had a CFI of 0.852, TLI of 0.831, and RMSEA of 0.040 (90% CI 0.038, 0.043). Given that we had additional data at our disposal, we tested the new model on a second sample of 1,000 observations to assess the replication of this new factor structure. The resulting model fit equally well, with a CFI of 0.856, TLI of 0.835, and RMSEA of 0.040 (90% CI 0.037, 0.043). Although additional model modification would be necessary to achieve good fit with the CFI and TLI, we were able to significantly reduce the number of predictors included.

Empirical Example 2

Given that the application of regularized SEM has predominantly been applied to regression parameters in a MIMIC model or to factor loadings in a CFA model, our goal in the empirical example was to apply penalties in a unique way. For this, we applied penalties to residual covariances as a mechanism to overcome poor fit and identify possible latent variables that explain a facet of manifest variables not captured by a general factor. We tested this using personality data collected online and freely available at https://openpsychometrics.org/_rawdata/ . Specifically, we analyzed responses to the 50-item International Personality Item Pool Five-Factor Model measure (Goldberg, 1999). The dataset consists of 4,270 observations, but for the purposes of this analysis we randomly sampled 1,000 cases. Of the five scales, we focused on Conscientiousness, fitting a one factor model to the 10 observed variables. The item content is detailed in the Appendix. The one-factor model did not fit well, with a CFI of 0.907, TLI of 0.880, and RMSEA of 0.080 (90% CI [0.071, 0.089]). The unstandardized and standardized factor loadings are displayed in Table 9.

We view the approach of applying regularization to the residual covariances as falling in between the use of modification indices and the application of exploratory factor analysis. Even in the presence of well-defined theory, CFA models may not fit well because of the presence of correlated residuals (Muthén & Asparouhov, 2012). Specifically, our aim could be to extract a single factor model, while also learning or exploring the presence of method effects that result in nontrivial residual covariances (e.g., Brown, 2013; Marsh & Yeung, 1996). As a first step, we fit a model with all possible residual covariances. Given that this model is not identified (−1 degrees of freedom), we do not report parameter estimates. As a second step, we added LASSO penalties to each of the residual covariances as a mechanism to identify which may be nonzero. This procedure is similar to the Bayesian LASSO approach of Pan et al. (2017) and the residual network modeling approach of Epskamp et al. (2017).

Note that we do not use the ALASSO, even though it had superior performance in the simulation study, as the MLE parameter estimates are unstable given the nonidentification. Therefore, we paired the stability selection-LASSO with BIC to select residual covariances, as this resulted in the best performance for LASSO penalties.

Table 9 shows the selected residual covariances after running the stability procedure. After modifying the initial model to include those residual covariances, the new model evidenced far superior fit, with a CFI of 0.976, TLI of 0.963, and RMSEA of 0.044 (90% CI [0.034, 0.055]). Given that we had additional data at our disposal, we tested the new model on a second sample of 1,000 observations to assess the replication of this new factor structure. Although we could just estimate the model in the holdout data, as is a common form of replication in factor analysis research, a stricter test is to use the expected covariance matrix from the first sample, then treat this as fixed for the assessment of fit in the holdout. Reestimating the model in the holdout sample resulted in an almost identical fit (unsurpris-

Table 8
Selection Results of Empirical Example 1 From Stability Selection

Indicators		Factors	Select	Indicators		Factors	Select
education	→	grit1	FALSE	education	→	grit2	TRUE
urban	→	grit1	FALSE	urban	→	grit2	TRUE
gender	→	grit1	FALSE	gender	→	grit2	TRUE
engnat	→	grit1	FALSE	engnat	→	grit2	FALSE
age	→	grit1	TRUE	age	→	grit2	TRUE
hand	→	grit1	FALSE	hand	→	grit2	FALSE
religion	→	grit1	FALSE	religion	→	grit2	TRUE
orientation	→	grit1	FALSE	orientation	→	grit2	FALSE
race	→	grit1	FALSE	race	→	grit2	TRUE
voted	→	grit1	FALSE	voted	→	grit2	FALSE
married	→	grit1	FALSE	married	→	grit2	FALSE
familysize	→	grit1	FALSE	familysize	→	grit2	TRUE
A1	→	grit1	FALSE	A1	→	grit2	FALSE
A2	→	grit1	FALSE	A2	→	grit2	TRUE
A3	→	grit1	TRUE	A3	→	grit2	FALSE
A4	→	grit1	FALSE	A4	→	grit2	TRUE
A5	→	grit1	FALSE	A5	→	grit2	FALSE
A6	→	grit1	TRUE	A6	→	grit2	FALSE
A7	→	grit1	FALSE	A7	→	grit2	FALSE
A8	→	grit1	FALSE	A8	→	grit2	FALSE
A9	→	grit1	FALSE	A9	→	grit2	FALSE
A10	→	grit1	TRUE	A10	→	grit2	TRUE
C1	→	grit1	TRUE	C1	→	grit2	TRUE
C2	→	grit1	FALSE	C2	→	grit2	FALSE
C3	→	grit1	TRUE	C3	→	grit2	TRUE
C4	→	grit1	TRUE	C4	→	grit2	FALSE
C5	→	grit1	TRUE	C5	→	grit2	TRUE
C6	→	grit1	TRUE	C6	→	grit2	FALSE
C7	→	grit1	FALSE	C7	→	grit2	TRUE
C8	→	grit1	TRUE	C8	→	grit2	TRUE
C9	→	grit1	TRUE	C9	→	grit2	TRUE
C10	→	grit1	FALSE	C10	→	grit2	TRUE
O1	→	grit1	FALSE	O1	→	grit2	TRUE
O2	→	grit1	TRUE	O2	→	grit2	FALSE
O3	→	grit1	TRUE	O3	→	grit2	FALSE
O4	→	grit1	FALSE	O4	→	grit2	TRUE
O5	→	grit1	FALSE	O5	→	grit2	TRUE
O6	→	grit1	TRUE	O6	→	grit2	FALSE
O7	→	grit1	FALSE	O7	→	grit2	FALSE
O8	→	grit1	FALSE	O8	→	grit2	TRUE
O9	→	grit1	FALSE	O9	→	grit2	FALSE
O10	→	grit1	TRUE	O10	→	grit2	TRUE
E1	→	grit1	FALSE	E1	→	grit2	FALSE
E2	→	grit1	FALSE	E2	→	grit2	FALSE
E3	→	grit1	FALSE	E3	→	grit2	FALSE
E4	→	grit1	FALSE	E4	→	grit2	FALSE
E5	→	grit1	FALSE	E5	→	grit2	TRUE
E6	→	grit1	TRUE	E6	→	grit2	TRUE
E7	→	grit1	TRUE	E7	→	grit2	TRUE
E8	→	grit1	FALSE	E8	→	grit2	FALSE
E9	→	grit1	FALSE	E9	→	grit2	FALSE
E10	→	grit1	TRUE	E10	→	grit2	FALSE
N1	→	grit1	TRUE	N1	→	grit2	TRUE
N2	→	grit1	FALSE	N2	→	grit2	FALSE
N3	→	grit1	FALSE	N3	→	grit2	TRUE
N4	→	grit1	FALSE	N4	→	grit2	TRUE
N5	→	grit1	FALSE	N5	→	grit2	FALSE
N6	→	grit1	FALSE	N6	→	grit2	FALSE
N7	→	grit1	TRUE	N7	→	grit2	FALSE
N8	→	grit1	FALSE	N8	→	grit2	TRUE
N9	→	grit1	TRUE	N9	→	grit2	TRUE
N10	→	grit1	FALSE	N10	→	grit2	TRUE

Note. In the select column, true means the path is selected to the final model, whereas false means not.

Table 9
Results of Empirical Example 2

Model parameter	Original model	New model
Factor loadings		
C1	0.767	0.762
C2	0.749	0.627
C3	0.395	0.375
C4	0.781	0.672
C5	0.780	0.785
C6	0.890	0.791
C7	0.605	0.568
C8	0.598	0.533
C9	0.747	0.797
C10	0.481	0.537
Residual variances		
C1	0.724	0.667
C2	1.282	1.432
C3	0.797	0.812
C4	1.014	1.108
C5	0.928	0.923
C6	1.194	1.325
C7	0.905	0.920
C8	0.894	0.922
C9	0.989	0.874
C10	0.749	0.688
Residual covariances		
C2 WITH C4		0.257
C2 WITH C5		0.132
C2 WITH C6		0.429
C4 WITH C6		0.231
C4 WITH C8		0.176
C7 WITH C9		0.151

Note. All parameters were significant at $p < .05$.

ingly given that the sample characteristics are identical). This stricter form of replication resulted in only a slight decrement in fit, with a CFI of 0.966, TLI of 0.948, and RMSEA of 0.053.[7]

In examining the residual covariances in Table 9, we can see that C2, "I leave my belongings around," accounts for the majority of relationships. In examining the other three items, we can see that they all generally assess an additional factor which could be described as messiness, a subfacet of Conscientiousnous. These results are further confirmed in the use of exploratory factor analysis, where a two-factor solution (using oblimin rotation) has one factor that is mainly indicated by those same four variables.

General Discussion

The purpose of this article was to highlight two of the main drawbacks of using LASSO regularization in psychometric models, namely inconsistency and high FPRs, while detailing and arguing for the application of stability selection and other LASSO extensions. We demonstrated this in two simulation studies, which considered selecting predictors of latent variables and cross loadings in SEM. The FPRs using the LASSO penalty are very high and remain high even as sample size increases. There is no clear sign that using cross-validation with the 1se rule or bootstrapping would help with the two problems of the LASSO in regularized SEM. The use of stability selection indeed could help lower the FPRs, but would not help with the inconsistency, that is, though the FPRs of stability selection with the LASSO is lower than just using the LASSO, the rates do not decrease as sample size increases. Conversely, the ALASSO provides consistent selection results, but would provide FPRs as high as LASSO when sample size is relatively small. Stability selection with the ALASSO penalty and using the AIC to choose the probability threshold gives superior performance to other regularization methods, both in lowering the FPRs and providing consistent results. We demonstrated the application of stability selection across two empirical examples, one selecting among a large set of personality predictors in a MIMIC model, the other selecting among residual covariances in a CFA model that had poor initial fit.

As with most things in statistics (and machine learning, e.g., the no free lunch theorem; Wolpert & Macready, 1997), no one method performed optimally across the simulation conditions. Given this, we put the onus on the researcher for justifying their choice of methodology. In many situations, the use of MLE works extremely well, particularly if the FPR is of utmost concern. If parameters need to be added based on poor fit, the use of modification indices can work well as long as the number added is small. However, when the sample size is small relative to the number of parameters, estimation problems occur, or the number of parameters to be removed (or added and tested to see if they make a meaningful contribution) is large, a number of the proposed methods are recommended. When MLE parameter estimates can be reliably estimated, the stability selection-ALASSO-AIC demonstrated the best performance, whereas the ALASSO can also be used when computational resources are limited. When MLE exhibits problems, or the parameter estimates may be inflated because of other data characteristics, we recommend the stability selection-LASSO-BIC.

Again, we want to reinforce that regularization is a data-driven selection process, and its applicability lies between exploratory and confirmatory research. The high FPRs and inconsistency of the LASSO could cause the selection result to vary on slightly different samples (e.g., bootstrap samples). Stability selection is fundamentally an atheoretical approach to performing variable selection, trying to aggregate across bootstrap samples, thereby stabilizing the selection results. Although new regularization approaches are being developed that incorporate researcher assigned weights (Tay et al., 2020), these have not yet been extended to regularization in SEM.

With stability selection being a two-step procedure, we want to further warn researchers that the inference of the estimated model in the second step (after the set of variables/paths are chosen) should not be done naïvely (e.g., using standard errors), because both the randomness introduced by the selection process and the sample space restriction implied by the chosen model are ignored (Huang, 2020). Results from the process should be applied in new data, particularly if inference into parameter significance is desired. For further discussion on post selection inference, see Huang (2020).

The use of these procedures does not negate the use or influence of theory into the modeling setting. This aspect of using regularized SEM is amplified by the characterization of regularization as falling under the umbrella of machine learning, typically conceptualized as a set of atheoretical algorithms. Instead, we see the utility of applying regularized SEM being amplified when theory

[7] A problem occurred in estimating the 90% CI for the RMSEA.

fails as a mechanism to identify what specific aspects (of a potentially complex theoretical formulation) were missed or erroneously included. In addition, in the nascent stages of theory generation, data may be available for the purposes of hypothesis formulation, to then be followed by a more rigorous testing in a follow up study.

An additional consideration when choosing among various regularization procedures is the computational complexity of each. Stability selection methods require a large number of bootstrap samples, greatly increasing the amount of time it takes to produce results. Particularly for large models, the computational complexity may be prohibitive, thus necessitating the use of MLE or RegSEM-ALASSO depending on the research goals. As one mechanism for decreasing the computation time, parallel processing has been built into the R function that performs stability selection.

Limitations and Future Directions

This study has a number of limitations which can be considered as future research directions. The first one is with respect to the choice of the initial range Λ and the probability threshold π_{thr} of stability selection. Although Meinshausen and Bühlmann (2010) stated that the final results are insensitive to the choice of Λ and vary little for π_{thr} values in (0.6, 0.9), it is not always the case in real applications. Throughout the paper, we determined the range of Λ using the interquartile range of λ values chosen by regularized SEM on 50 bootstrap samples. However, there is no theoretical justification that this application is the best, and we believe there could be more sophisticated choices. As for π_{thr}, we tried treating it as a tuning parameter and selected it based on fit indices. The results, indeed, vary across different choice of fit indices. Although it sounds as if this method yields uncertain results, stability selection does provide a fixed maximal selection probability on Λ for each parameter being penalized, thus the sequence of variable selection is fixed, leading to another possible way of choosing π_{thr} based on theoretical justification. Again, we leave it open for better ways of determining this probability threshold π_{thr}.

Another limitation is that we are only focusing on FPRs and consistency in this study, ignoring other possible problems of LASSO, for example, correlation among variables, and parameter estimation bias. Other penalties such as elastic net have been demonstrated to work well when there is correlation among variables (Zou & Hastie, 2005), however, testing its performance would add another dimension to our simulation study and is out of the scope of the problems concerned in this paper. Investigating these further analysis conditions warrants future work however.

Lastly, in this study, we tried to manipulate several conditions, however, some simulation conditions were ignored. For example, in simulation one, we did not simulate any correlation among the predictors, which is unlikely to happen in most psychological research. In Jacobucci et al. (2019), collinearity was manipulated (0, 0.2, 0.5, 0.8, 0.95), finding that FPRs are indeed higher when collinearity was present. However, the influence was small unless the collinearity is very high (0.95), which is unrealistic in practice. The influence of collinearity on FNRs was negligible. Given these prior results, we didn't add that dimension to our simulation study as we would expect the behavior to be similar when collinearity exists.

Conclusion

Regularization is an approach that allows the incorporation of confirmatory and exploratory modeling. Though it has gained popularity in psychological research in recent years, some of the drawbacks of this method are ignored. In this paper, we focused on two of the most pertinent drawbacks for psychological research: high FPRs and inconsistency. To address these problems with the LASSO, we compared several computational methods in two simulation studies. We found that the use of stability selection combined with the ALASSO as superior performance in terms of both lowering FPRs and providing consistent selection results. It is thus recommended when computational resources permit.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alexander, D. H., & Lange, K. (2011). Stability selection for genome-wide association. *Genetic Epidemiology*, *35*(7), 722–728. https://doi.org/10.1002/gepi.20623

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*(3), 397–438. https://doi.org/10.1080/10705510903008204

Barber, R., & Candès, E. (2015). Controlling the false discovery rate via knockoffs. *Annals of Statistics*, *43*(5), 2055–2085. https://doi.org/10.1214/15-AOS1337

Bauer, D. J., Belzak, W. C., & Cole, V. T. (2019). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling*, *27*(1), 43–55.

Bentler, P. M. (1986). *Lagrange multiplier and Wald tests for EQS and EQSPC*. BMDP Statistical Software.

Bentler, P. M. (2006). *EQS 6, Structural Equations Program Manual*. Multivariate Software Inc.

Bentler, P. M., & Chou, C. P. (1988). Practical issues in structural modeling. In J. S. Long (Ed.), *Common problems/proper solutions* (pp. 161–192). Sage.

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 301–328). Springer. https://doi.org/10.1007/978-94-007-6094-3_15

Bovelstad, H., Nygard, S., Storvold, H., Aldrin, M., Borgan, O., Frigessi, A., & Lingjaerde, O. (2007). Predicting survival from microarray data comparative study. *Bioinformatics 23*(16), 2080–2087. https://doi.org/10.1093/bioinformatics/btm305

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001, October). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, T. A. (2013). *The Oxford handbook of quantitative methods: Vol. 2. Latent variable measurement models*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199934898.013.0013

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150. https://doi.org/10.1207/S15327906MBR3601_05

Chou, C.-P., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier and Wald tests. *Multivariate Behavioral Research*, *25*, 115–136. https://doi.org/10.1207/s15327906mbr2501_13

Datta, A., & Zou, H. (2019). A note on cross-validation for LASSO under measurement errors. *Technometrics*, *62*(4), 549–556. https://doi.org/10.1080/00401706.2019.1668856

Drysdale, E., Peng, Y., Hanna, T. P., Nguyen, P., & Goldenberg, A. (2019). The false positive control LASSO. *ArXiv*. https://arxiv.org/abs/1903.12584

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*, *82*(4), 904–927. https://doi.org/10.1007/s11336-017-9557-x

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *JASA*, *96*, 1348–1360. https://doi.org/10.1198/016214501753382273

Fan, J., & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, *32*(3), 928–961. https://doi.org/10.1214/009053604000000256

Fan, J., & Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *Annals of Statistics*, *38*(6), 3567–3604. https://doi.org/10.1214/10-AOS798

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, *7*, 7–28.

Hirose, K., & Yamamoto, M. (2013). Estimation of oblique structure via penalized likelihood factor analysis. *Computational Statistics & Data Analysis*, *79*, 120–132. https://doi.org/10.1016/j.csda.2014.05.011

Hirose, K., & Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing*, *25*, 863–875. https://doi.org/10.1007/s11222-014-9458-0

Hofner, B., Boccuto, L., & Göker, M. (2015). Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinformatics*, *16*, 144. https://doi.org/10.1186/s12859-015-0575-3

Huang, P. (2020). Postselection inference in structural equation modeling. *Multivariate Behavioral Research*, *55*(3), 344–360. https://doi.org/10.1080/00273171.2019.1634996

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

Huang, S., & Marchetti-Bowick, M. (2014). Summary and discussion of: "Stability Selection." *Statistics Journal Club*, 36–825.

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

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*(1), 55–76. https://doi.org/10.1177/2515245919826527

Jacobucci, R., Grimm, K. J., & Mcardle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling*, *23*(4), 555–566. https://doi.org/10.1080/10705511.2016.1154793

Jacobucci, R., Grimm, K. J., Brandmaier, A. M., Serang, S., Kievit, R. A., Scharf, F., & Li, X. (2020). RegSEM: Regularized structure equation modeling (R package version 1.5.0). https://CRAN.R-project.org/package=regsem

Jöreskog, K., & Goldberger, A. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351), 631–639. https://doi.org/10.2307/2285946

Lee, S.-Y., & Bentler, P. M. (1980). Some asymptotic properties of constrained generalized least squares estimation in covariance structure models. *South African Statistical Journal*, *14*, 121–136.

Leng, C., Lin, Y., & Wahba, G. (2006). A note on the LASSO and related procedures in model selection. *Statistica Sinica*, *16*(4), 1273–1284.

Liang, X., & Jacobucci, R. (2019). Regularized structural equation modeling to detect measurement bias: Evaluation of LASSO, adaptive LASSO, and elastic net. *Structural Equation Modeling*. Advance online publication. https://doi.org/10.1080/10705511.2019.1693273

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Marsh, H. W., & Yeung, A. S. (1996). The distinctiveness of affects in specific school subjects: An application of confirmatory factor analysis with the National Educational Longitudinal Study of 1988. *American Educational Research Journal*, *33*(3), 665–689. https://doi.org/10.3102/00028312033003665

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

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the LASSO. *Annals of Statistics*, *34*(3), 1436–1462. https://doi.org/10.1214/009053606000000281

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B, Statistical Methodology*, *72*(4), 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. https://doi.org/10.1037/a0026802

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*, 431–462. https://doi.org/10.1007/BF02294365

Pan, J., Ip, E. H., & Dubé, L. (2017). An alternative to post hoc model modification in confirmatory factor analysis: The Bayesian LASSO. *Psychological Methods*, *22*(4), 687–704. https://doi.org/10.1037/met0000112

Raykov, T., & Marcoulides, G. A. (1999). On desirability of parsimony in structural equation model selection. *Structural Equation Modeling*, *6*(3), 292–300. https://doi.org/10.1080/10705519909540135

Roberts, S., & Nowak, G. (2014). Stabilizing the LASSO against cross-validation variability. *Computational Statistics & Data Analysis*, *70*, 198–211. https://doi.org/10.1016/j.csda.2013.09.008

Ryali, S., Chen, T., Padmanabhan, A., Cai, W., & Menon, V. (2015). Development and validation of consensus clustering-based framework for brain segmentation using resting fMRI. *Journal of Neuroscience Methods*, *240*, 128–140. https://doi.org/10.1016/j.jneumeth.2014.11.014

Scharf, F., & Nestler, S. (2019). Should regularization replace simple structure rotation in exploratory factor analysis? *Structural Equation Modeling*, *26*(4), 576–590. https://doi.org/10.1080/10705511.2018.1558060

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464. https://doi.org/10.1214/aos/1176344136

Serang, S., Jacobucci, R., Brimhall, K. C., & Grimm, K. J. (2017). Exploratory mediation analysis via regularization. *Structural Equation Modeling*, *24*, 733–744. https://doi.org/10.1080/10705511.2017.1311775

Shipley, B. (2016). *Cause and correlation in biology: A user's guide to path analysis, structural equations and causal inference with R*. Cambridge University Press. https://doi.org/10.1017/CBO9781139979573

Sitek, A., Gullberg, G. T., & Huesman, R. H. (2002). Correction for ambiguous solutions in factor analysis using a penalized least squares objective. *IEEE Transactions on Medical Imaging*, *21*, 216–225. https://doi.org/10.1109/42.996340

Su, W., Bogdan, M., & Candès, E. (2017). False discoveries occur early on the Lasso path. *Annals of Statistics*, *45*, 2133–2150. https://doi.org/10.1214/16-AOS1521

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. https://doi.org/10.1007/s11336-016-9529-6

Tay, J. K., Aghaeepour, N., Hastie, T., & Tibshirani, R. (2020). Feature-weighted elastic net: Using "features of features" for better prediction. *arXiv*: 2006.01395.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Lawrence Erlbaum Associates, Inc.

Tibshirani, R. (1996). *Regression shrinkage and selection via the LASSO*. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on psychological science: A journal of the Association for Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., DiBernardo, A., Narayan, V. A., & Alzheimer's Disease Neuroimaging Initiative. (2012). Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. *BMC Neurology*, *12*, 46. https://doi.org/10.1186/1471-2377-12-46

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B. Methodological*, *68*, 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, *38*(2), 894–942. https://doi.org/10.1214/09-AOS729

Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

Zou, H. (2006). *The adaptive LASSO and its oracle properties*. https://doi.org/10.1198/016214506000000735

Zou, H., Choi, J., & Oehlert, G. (2010). A penalized maximum likelihood approach to sparse factor analysis. *Statistics and Its Interface*, *3*(4), 429–436.

Zou, H., & Hastie, T. J. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

Appendix

Item Content of Empirical Example 2

C1	I am always prepared.
C2	I leave my belongings around.
C3	I pay attention to details.
C4	I make a mess of things.
C5	I get chores done right away.
C6	I often forget to put things back in their proper place.
C7	I like order.
C8	I shirk my duties.
C9	I follow a schedule.
C10	I am exacting in my work.

Received February 10, 2020
Revision received September 23, 2020
Accepted December 4, 2020 ■