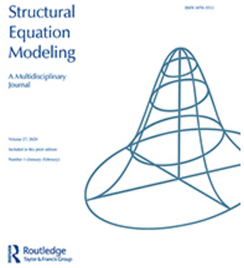



Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

# Model Selection with Lasso in Multi-group Structural Equation Models

Jonas Christoffer Lindstrøm & Fredrik A. Dahl

**To cite this article:** Jonas Christoffer Lindstrøm & Fredrik A. Dahl (2020) Model Selection with Lasso in Multi-group Structural Equation Models, Structural Equation Modeling: A Multidisciplinary Journal, 27:1, 33-42, DOI: 10.1080/10705511.2019.1638262

**To link to this article:** https://doi.org/10.1080/10705511.2019.1638262



Published online: 14 Aug 2019.

*Structural Equation Modeling: A Multidisciplinary Journal*, 27: 33–42, 2020

ISSN: 1070-5511 print / 1532-8007 online
DOI: https://doi.org/10.1080/10705511.2019.1638262

OPEN ACCESS

# Model Selection with Lasso in Multi-group Structural Equation Models

Jonas Christoffer Lindstrøm[1,2] and Fredrik A. Dahl[1,2]
[1]*Health Services Research Unit, Akershus University Hospital, Norway*
[2]*Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Norway*

A Structural Equations Modeling analysis of multiple groups often involves specification of cross-group parameter equality constraints. In this paper, we present a technique for estimating the differences and equalities in parameters between groups using L1-penalized estimation (also known as the Lasso). We present the general model formulation and provide an algorithm for estimating the parameters across a range of penalization levels and a procedure for determining the amount of penalization. We also provide two case studies, one with a model including only observed variables, and one with a model with latent variables. Further, we conduct a simulation study to investigate some properties of the method.

**Keywords**: Lasso, structural equations modelling, variable selection, regularization, multiple group analysis

## INTRODUCTION

Structural Equation Models (SEM) are a family of models used to analyze complex relationships between variables (Mulaik, 2009). In one way, the SEM framework can be seen as an extension of the familiar regression model, where we are not limited to having a single response variable and response variables can themselves be predictors for other variables. In this way, both direct and indirect associations can be modeled. Another important part of the SEM methodology is the possibility to incorporate latent variables. Latent variable models include both regression-type models where some variables can be latent and factor models where the focus is the relationship between observed variables and unobserved constructs.

A SEM analysis is usually confirmatory in nature, where the two-fold goal is to estimate the parameters of a prespecified model and to investigate how well the model fits the data. As such it can be of interest to do subgroup analyses to see if the model specification holds across groups and parameter estimates are similar (Jöreskog, 1971). A multigroup analysis can involve adding constraints to some of the parameters, so that they are estimated to be equal across groups, while others are left to be estimated freely. This must be done in a way that considers the balance between model fit and model complexity. The freeing and constraining of the parameters across groups is usually determined by the researcher, guided by a combination of theoretical considerations, goodness-of-fit tests, and various model fit indexes (Cheung & Rensvold, 2002; Mulaik, 2009). A multi-group SEM analysis will therefore often be more exploratory in nature than a single group analysis.

Another strategy for balancing model complexity and model fit is to use penalized estimation procedures. Penalized estimation techniques extend the estimation criterion by an additional term that is a function of the magnitudes of the parameters, so that the parameter space becomes constrained. This type of constraint differs from the hard equality constraints that are usually employed, whether it be equality across groups or equality to 0, since the constraint implied by the penalty applies to the

Correspondence should be addressed to Jonas Christoffer Lindstrøm, Health Services Research Unit, Akershus University Hospital, Lørenskog 1478, Norway. E-mail: jol7@ahus.no

overall structure of the parameters and not to specific parameters. The amount of penalty, or the amount of constraint, can be adjusted via a few *tuning parameters*. These tuning parameters then smoothly shift the model estimates from good fit in complex models (possibly overfitted) to poorer fit in less complex models that hopefully generalizes better on new data. The tuning can be done using a data-driven procedure such as cross-validation.

Different types of functions, with different properties, can be used as penalties in the parameter estimation. The two most popular penalty functions are the ridge and the lasso. The ridge penalty is simply the sum of the squared parameters, and the lasso penalty is the sum of the absolute values of the parameters. Although they are superficially similar, their properties differ in important ways. While both tend to decrease the overall magnitude of the parameters, the lasso tend to estimate some parameters to be exactly 0, while leaving others non-zero. This property is sometimes desirable as this can be thought of as a form of model selection. This is usually not the case when the ridge penalty is used.

Penalized estimation has been used extensively in big data applications such as machine learning and high-throughput genomics, but has not been used much together with the SEM methodology. A penalized estimate was recently been developed for two-stage least squares estimators (Jung, 2013) and for covariance estimators (Jacobucci, Grimm, & McArdle, 2016) (Huang, Chen, & Weng, 2017), and there have been applications in mediation analysis (Serang, Jacobucci, Brimhall, & Grimm, 2017).

In this paper, we present a new data-driven methodology for identifying group differences in structural equation models using penalized estimation procedures.

## THE STRUCTURAL EQUATIONS MODEL

In the rest of this paper we will use the RAM notation to describe and specify structural equations models (McArdle & McDonald, 1984). With the RAM notation, the model is specified using two matrices, A and S, and a vector M, which contains all parameters. If the model contains $p$ variables, including both observed and latent variables, then the A matrix is a $p \times p$ matrix encoding of the asymmetric relations in the model graph. The asymmetric relations are in this context regression coefficients and factor loadings. In FIGURE 1 this is illustrated for a simple example model. Similarly, the $p \times p$ matrix S is the symmetric adjacency matrix of the undirected part of the model graph, which specifies the variances (on the diagonal) and the residual covariances (the off-diagonal elements).

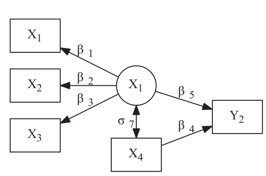
The $p$ vector M contains the means (for the exogenous variables) and intercepts (for the endogenous variables). An additional $k \times p$ filter matrix F helps extract the $k$ observed variables of the model implied covariance matrix, so that it can be compared to the observed $k \times k$ data covariance matrix $C$.

Given a set of RAM matrices, the model implied covariance matrix $\Sigma(\theta)$ is given as (Jacobucci et al., 2016)

$$\Sigma(\theta) = F(I - A)^{-1} S (F(I - A)^{-1})^{\mathrm{T}} \tag{1}$$

and the model implied mean vector is given as

$$\mu(\theta) = F(I - A)^{-1} M \tag{2}$$




(a)

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \beta_1 \\ 0 & 0 & 0 & 0 & 0 & \beta_2 \\ 0 & 0 & 0 & 0 & 0 & \beta_3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \beta_4 & 0 & \beta_5 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(b)

$$S = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4 & 0 & \sigma_7 \\ 0 & 0 & 0 & 0 & \sigma_5 & 0 \\ 0 & 0 & 0 & \sigma_7 & 0 & \sigma_6 \end{bmatrix}$$

(c)

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \beta_1 + \delta_1 \\ 0 & 0 & 0 & 0 & 0 & \beta_2 + \delta_2 \\ 0 & 0 & 0 & 0 & 0 & \beta_3 + \delta_3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \beta_4 + \delta_4 & 0 & \beta_5 + \delta_5 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(d)

FIGURE 1 The A matrix encoding the directed relationships in an example model. (a) A graph displaying a toy model, excluding the self-arrows that signify variances. (b) The A matrix encoding the directed relationships in the example model. The layout of the non-zero elements of A correspond to the adjacency matrix of the model graph. (c) The S matrix encoding the undirected relationships in the example model, including the variances along the diagonal. (d) Illustration of the A matrix in a group with the difference parameterization.

The parameters are estimated by minimizing the negative log-likelihood criterion for a multivariate normal model

$$f(\Sigma,\mu,\theta) = \log(|C|) + Tr(C * \Sigma(\theta)^{-1}) - \log(C) - r + (\mu - \mu(\theta))^T \Sigma(\theta)^{-1} (\mu - \mu(\theta)) \quad (3)$$

where $r$ is the number of parameters.

## Group SEM

A common situation is that the data can be divided into $G$ subgroups and it is of interest to see if the different groups yield similar parameter estimates and model fit. The estimation criterion can be extended as follows to account for the group structure:

$$f(C,\mu,\theta) = f(C_1,\mu_1,\theta_1) + f(C_2,\mu_2,\theta_2) + ... + f(C_G,\mu_G,\theta_G) \quad (4)$$

where $C_g$ is the observed covariance matrix, $\mu_g$ is the vector of observed means and $\theta_g$ is the parameter vector, for group $g$.

A typical group SEM analysis will proceed in an iterative fashion, where several models are fitted, with some of the parameters constrained to be equal across groups, while others are left to be estimated independently in each group. Determining which parameters should be constrained and free are decided by the researcher, and is typically judged using a combination of goodness of fit indices and the researcher's knowledge of the subject that is studied.

# THE LASSO GROUP SEM

Instead of relying on model fit indices and manual specification of the constraints in the group SEM analysis, we will in this section develop a penalized estimation procedure, using the Lasso penalty (Hastie, Tibshirani, & Wainwright, 2015), to identify the parameters that differ between subgroups.

Let, arbitrarily, the first group be the reference group. We reparameterize the part of the model belonging to the other groups to reflect how much the parameters depart from what is the case in the reference group. Specifically, let

$$\theta_g = \theta_1 + \delta_g \quad (5)$$

for $g \geq 2$. The $\delta_g$ parameters represent how much the parameters in a group deviates from the corresponding parameters in the reference group. When $\delta_g = 0$, then $\theta_g = \theta_1$ will be the same in the two groups. See FIGURE 1(d) for an example of how the A matrix would look like with this setup. The $\delta_g$ and the parameter parameters will be the parameters of interest in the rest of the analysis.

With this parameterization, we can rewrite the criterion in Equation 4 and extend it with the Lasso penalty

$$f(\Sigma,\mu,\theta) = f(C_1,\mu_1,\theta_1) + \sum_{g=2}^{G} f(C_g,\mu_g,\theta_g) + \lambda_g \sum_{j}^{J} |\delta_{gj}| \quad (6)$$

where $\lambda_g \geq 0$ is a constant and a *tuning parameter* that is not directly estimated, and J is the number of $\delta_g$ parameters. The consequence of the additional penalization term is that the absolute magnitude of the estimates of $\delta_{gi}$ will be smaller. For some values of $\lambda_g$ some of the parameters will be estimated to be exactly 0, thus estimating those effects to be the same as in the reference group. $\lambda_g$ acts as a tuning parameter that balances between model fit and model complexity. When $\lambda_g = 0$ for $g = 1, ..., G$ the criterion become equivalent (apart from the parameterization) with the unpenalized criterion in Equation 4, and the estimates will be ordinary ML estimates. The amount of penalty is the same for each $\delta$ parameter within a group, which require the parameters to be on the same scale. To ensure this the variables must be scaled before the parameters are estimated.

As pointed out in (Jacobucci et al., 2016), there is no point in penalizing the parameters in both the A and S matrices. Furthermore, we can consider the means and intercepts as nuisance parameters, and remove them from our analysis by centering the variables in each group prior to the analysis. In the following, we will therefore only penalize the parameters in the A matrix, which we denote $\delta^\alpha$, and leave the parameters in the S matrix, $\delta^\sigma$, unpenalized. [It can sometimes be worthwhile to fix the parameters in S so that these do not vary when the penalty changes, as this can compensate for lack of fit caused by penalizing the parameters in A.]

In the rest of the paper, we discuss only the two-group case, unless otherwise stated, and drop the group-indicating subscripts. Extension to several groups is straightforward, but with a caution in regard to the choice of reference group. In the two-group case, the choice of reference group does not matter, as should be evident from equation 5. Unfortunately, this is not the case when there are more than two groups.

## Parameter estimation

Here we describe an algorithm to minimize the penalized log-likelihood criterion in equation 6 for a range of $\lambda$ values, and then discuss how to find a good value for $\lambda$. The algorithm builds on the cyclic coordinate descent method where each parameter estimate is updated in turn, until convergence. This method is known to be efficient for computing the parameters for a range of $\lambda$'s when the estimates from a given value of $\lambda$ are used as the initial estimates for the next value of $\lambda$ (Hastie et al., 2015).

Before the model estimation, the data should be standardized. Within each group, subtract the mean and divide by the standard deviation of all the variables. This removes the mean and intercept parameters from further considerations and bring all variables to a common scale. This ensures that the coefficients are directly comparable. Otherwise, we would have to use a separate λ for each coefficient.

1. Fit the model with the difference parameterization, but without any penalization ($\lambda = 0$). This gives us the estimates $\hat{\theta}_1$ and $\hat{\delta}_{\lambda_0}$.
   a. Use this to find $\lambda^{max}$ (see below). For a desired number of distinct values of λ, select evenly spaced values between $\lambda^{max}$ and 0.
2. Set all $\delta_j^{\alpha} = 0$
3. Starting with $\lambda^{max}$ and continue with successively lower values of λ, do the following steps:
   a. With all $\delta_j^{\alpha}$ parameters fixed to their current estimates, estimate the parameters in S, $\delta^{\sigma}$.
   b. For each parameter $\delta_j^{\alpha}, j = 1, ..., J$, do the following steps until convergence:
      i. Find a provisional estimate $\tilde{\delta}_j^{\alpha}$ of $\delta_j^{\alpha}$ by unpenalized ML, with all other $\delta^{\alpha}$ parameters fixed to their current estimates.
      ii. (Soft Tresholding) Apply the soft thresholding function before updating the current estimates:

$$\hat{\delta}_j = S_{\lambda}(\tilde{\delta}_j) \tag{7}$$

where

$$S_{\lambda}(x) = \text{sign}(x) \times \max(0, |x| - \lambda) \tag{8}$$

The soft thresholding step of the algorithm (Equation 8) is what causes the shrinking of the parameters. The soft thresholding shrinks the magnitude of the provisional estimate $\tilde{\delta}_j^{\alpha}$ by λ. If the intermediate estimate is smaller in magnitude than λ the new estimate becomes 0. The soft thresholding function is shown in FIGURE 2.

### *Determining* $\lambda^{max}$

It is useful to determine an upper bound of the value of λ such that all $\delta^{\alpha}$ are estimated to be 0. Any value of λ greater than the upper bound, $\lambda^{max}$, will just give the same result as $\lambda^{max}$. To find $\lambda^{max}$ we start by fitting the group model with the difference parameterization using maximum likelihood. This step should be done anyway since it corresponds to fitting the model with $\lambda = 0$. Since the soft thresholding function in equation 8 shrinks all parameters of smaller magnitude than λ to 0, then it is clear that values of λ greater than $\max(|\delta^{\alpha}|)$ will yield parameter estimates that are all 0. Therefore, the upper bound $\lambda^{max} = \max(|\delta^{\alpha}|)$. In our experience, however, the greatest value of λ that yields at least one non-zero estimate is often lower than $\lambda^{max}$.

## Model tuning

While it is useful to estimate the model for a range of λ values and see how the parameters change with it, this does not give us any indication of which values of λ give a reasonable balance between model fit and complexity. One way to determine the quality of the different

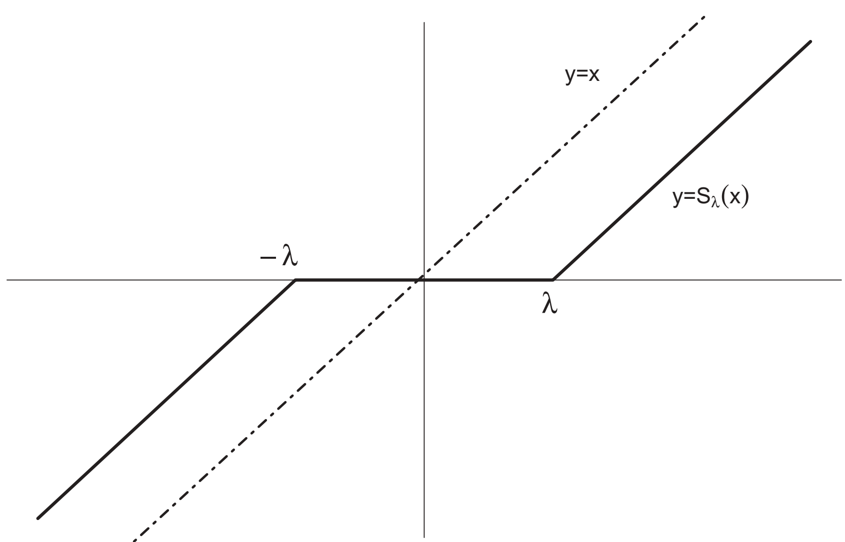



FIGURE 2 The soft threshold function. The soft threshold function shrinks the magnitude of the parameter estimates towards 0 by λ. If the parameter is smaller than λ then the parameter is estimated to be exactly 0.

estimates is to estimate how well the model will fit a new data set that was not used for estimation. To do this we use a cross-validation strategy where the data are split randomly in k pieces and the model is estimated, for a range of λs, on k-1 of the pieces. The last piece is withheld from model estimation and is used to gauge how well the model generalizes outside the data used in the estimation. To evaluate the out-of-sample model fit we use the unpenalized likelihood function in equation 4.

Since the model fitting relies on the observed covariance matrix, we want to avoid calculating this for small subsets of the data since it is known to be unstable when the sample size is small. We therefore split the data set in two (k = 2), and repeat the procedure on different random two-splits of the data (say, 5 times).

The data set is split within groups, so when k = 2, about 50% of the observations in each group are selected. This maintains the relative sample sizes of each group and ensures that one of the groups does not, by chance, become smaller than necessary.

## EXAMPLE ANALYSIS

### Example from affect valuation theory

In this example analysis, we fit a model from Affect Valuation Theory, where the influences and consequences about how people want to feel and how they actually feel varies between cultures (Tsai, Knutson, & Fung, 2006). The data set consists of responses to questionnaires gathered from a sample of European Americans (n = 65), Asian American (n = 67) and Hong Kong Chinese (n = 83). The model is shown in FIGURE 3(a).

In this example we selected the European American group as the reference group; thus, the $\delta$ parameters are the differences in the effects between the Asian American and Hong Kong Chinese samples and the European American sample.

Based on the initial analysis we found that $\lambda^{max}$ in this sample was approximately 0.58, but the smallest value of λ that estimated all $\delta$ parameters to zero was about 0.45 for the Hong Kong Chinese group and 0.27 for the Asian American group. The evolution of the estimates for the $\delta$ parameters for a range of 20 values of λ is shown in FIGURE 3(b). We used the 50% data splitting repeated 5 times to find the optimal λ, which was found to be $\lambda = 0.17$ for both groups, see FIGURE 3(c). The ML estimates and the penalized estimates with the optimal λ on the full data set are shown in TABLE 1.

The procedure identified three parameters that differed between the European Americans and Hong Kong Chinese and four that differed between the European Americans and Asian Americans. In all cases, including the non-zero $\delta$'s, the estimates are shrunk towards zero, as expected.

The ML estimates of the $\delta$ parameters for temperament on rigorous activities (−0.4) in the Asian American group are the parameter of second greatest magnitude in that group, but was shrunk to 0 at the optimal λ, while several of the other parameters were shrunk, but not to 0. In other words, the penalized $\delta$ parameters do not seem to be a simple function of the magnitude of the ML estimates.

As a sensitivity analysis of the method on these data set we sampled the data with replacement and repeated the estimation and cross-validation 100 times. The median λ that gave the least cross-validation error was 0.27 (IQR:0.23–0.33) for the Asian American group and 0.22 (IQR: 0.15–0.27) for the Hong Kong Chinese group. The proportion of times each of the $\delta$ parameters was estimated to be non-zero are listed in TABLE 1. It shows that the parameters with non-zero estimates in the original data set were the most commonly selected coefficients, and that this applies to both groups. It is also interesting to note that there is not a complete correspondence between the magnitude of the ML estimates and the probability of having penalized estimates shrunk to zero. This is especially seen in the coefficients for the Hong Kong Chinese group, where for instance the $\delta$-coefficient from Ideal Affect to Rigorous Activities has an ML estimate of −0.20 and has about half the probability of being selected as the coefficient for Temperament to Depression, which has a ML estimate of −0.16, which is closer to zero.

### A model with latent variables

In this example, we analyze data from the Norwegian election survey that was conducted after the 2013 general election (Kleven, Aardal, Berghm, Hesstvedt, & Hindenes, 2015). We study how different levels of political interest and participation influence how easy the respondents (n = 1522) felt it was to decide who to vote on. The respondents were asked how often they discuss politics with friends and family, participated in discussions online, how interested they are in politics, whether they participated in a demonstration, sign petitions and how much they cared about which party won. The model has two latent variables, political interest and political participation, and one observed outcome variable. The model graph is shown in FIGURE 4(a).

We fitted the model stratified on gender (756 women, 766 men), with men as the reference group. The evolution of the $\delta$ parameters is shown in FIGURE 4(b). From the cross-validation we found that the best out of sample fit was achieved with $\lambda = 0.28$ (FIGURE 4(c)).

The estimated parameters are shown in TABLE 2. Almost all $\delta$ parameters were estimated to be 0, including the parameters relating the latent variables to their observed indicators. The only nonzero $\delta$ parameter is the one for the effect of participation on how easy it was to decide. This parameter even changed sign between the ML estimate and the penalized estimate, going from −0.07 to 0.03. This is

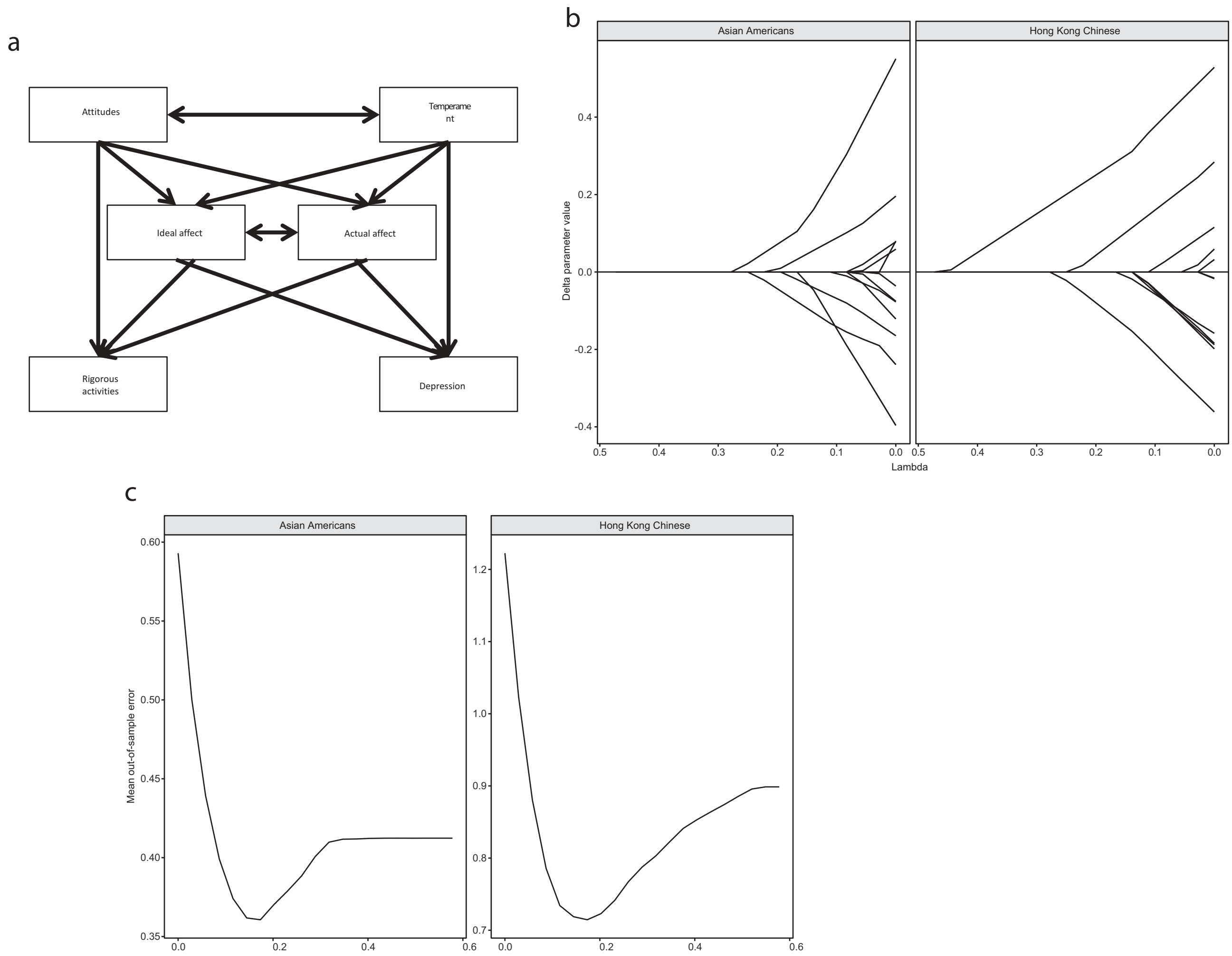



FIGURE 3 The affect valuation model. (a) Path diagram of the Affect Valuation Theory model. (b) The are represented with ellipses, while the $\delta$ parameters for a range of λ. The rightmost part of the plot, where $\lambda = 0$ are the ordinary ML estimates. The vertical gray lines indicate the values of λ that were evaluated. (c) Cross-validation results. The line shows the average out of sample error after five rounds of cross-validation for a range of λ values. The optimal λ is 0.17.

probably related to the effect of the other predictor of the response, "Interest", which was drastically shrunk from 0.36 to zero.

## A SIMULATION STUDY

To investigate some properties of the estimation and cross-validation procedure under different settings and sample sizes we did a simulation study. In order to have a realistic model, we used the Affect Valuation Theory model and data as the basis to simulate new data. We used the same analysis setup as before, by using two-fold cross-validation 5 times repeated with 20 values of λ.

In the first simulation study, we investigated the procedure under a typical null hypothesis setting where there are no differences between groups. To simulate data for this study we first fitted the Affect Valuation Theory model to the complete scaled data set ignoring the groups. Then, 100 data sets were simulated for two groups with the same parameters in each, with balanced sample sizes of 71, 150, 250 and 500 in each group. We chose to use a sample size of 71 since this is the average group size in the original AVT data set. The true parameter values used in this simulation are shown in the top part of TABLE 3.

the second simulation study, we investigated the procedure in the situation when most, but not all parameters, are the same across groups. To generate the true model we fitted the AVT model to the scaled data from the European American and Hong Kong Chinese groups, with the parameters estimated freely across the two groups. The parameter estimates from

TABLE 1
Parameter Estimates for the Affect Valuation Theory Model on Scaled and Centered Data. The Column β Shows the Regression Coefficients for the European American Group. The Column $\delta_{\lambda=0}$ Shows the Addition to the β Parameters that Yield the ML Estimates for the Hong Kong Chinese Group. The Estimates Using the Optimal λ Is Given in the $\delta_{\lambda=0.15}$ Column. Four Parameters Were Estimated to Be Non-zero. The Rightmost Column Show the Proportion of Times the Parameter Were Estimated to Be Non-zero across 50 Bootstrap Samples

| | | | *Asian American* | | | *HK Chinese* | | |
|---|---|---|---|---|---|---|---|---|
| *Response* | *Predictor* | $\beta_{EA}$ | $\delta_{\lambda=0}$ | $\delta_{\lambda=0.17}$ | $P(\delta\neq0)$ | $\delta_{\lambda=0}$ | $\delta_{\lambda=0.17}$ | $P(\delta\neq0)$ |
| Ideal affect | Attitudes | 0.19 | −0.16 | −0.01 | 0.35 | 0.03 | – | 0.21 |
| Actual affect | Attitudes | 0.01 | 0.20 | 0.03 | 0.27 | 0.28 | 0.07 | 0.53 |
| Rigorous activities | Attitudes | −0.18 | 0.06 | – | 0.10 | 0.12 | – | 0.25 |
| Depression | Attitudes | 0.03 | −0.04 | – | 0.08 | 0.06 | – | 0.14 |
| Ideal affect | Temperament | 0.31 | 0.08 | – | 0.13 | −0.18 | – | 0.31 |
| Actual affect | Temperament | 0.61 | −0.08 | – | 0.03 | −0.36 | −0.11 | 0.68 |
| Rigorous activities | Temperament | 0.37 | −0.40 | – | 0.21 | −0.19 | – | 0.44 |
| Depression | Temperament | −0.40 | 0.08 | – | 0.20 | −0.16 | – | 0.46 |
| Rigorous activities | Ideal affect | 0.21 | −0.12 | – | 0.08 | −0.20 | – | 0.22 |
| Depression | Ideal affect | 0.05 | −0.08 | – | 0.17 | −0.02 | – | 0.22 |
| Rigorous activities | Actual affect | −0.38 | 0.55 | 0.10 | 0.53 | 0.53 | 0.28 | 0.94 |
| Depression | Actual affect | −0.03 | −0.24 | −0.07 | 0.34 | −0.02 | – | 0.21 |

TABLE 2
Parameter Estimates for the Election Survey Model. The Column β Shows the Regression Coefficients for the Model Fitted to the Responses from Men (The Reference Group). The Column $\delta_{\lambda=0}$ Shows the Addition to the β Parameters that Yield the ML Estimates for Women. The Estimates Using the Optimal λ Is Given in the $\delta_{\lambda=0.28}$ Column. Four Parameters Were Estimated to Be Non-zero. *: Parameters Fixed to 1 to Make the Latent Variables Identifiable

| *Response* | *Predictor* | β | $\delta_{\lambda=0}$ | $\delta_{\lambda=0.28}$ |
|---|---|---|---|---|
| Interest | Interested in politics | 1* | – | – |
| Interest | Discuss politics with friends and family | 0.79 | − 0.02 | – |
| Interest | Care about who wins | 0.76 | 0.06 | – |
| Participation | Online discussion | 1* | – | – |
| Participation | Signed petition | 1.69 | 0.01 | – |
| Participation | Demonstration | 1.41 | −0.18 | – |
| Participation | Interest | 0.20 | 0.09 | – |
| Easy to decide | Participation | −0.29 | −0.07 | 0.03 |
| Easy to decide | Interest | 0.09 | 0.36 | – |

the European American group were then used as the true parameters in both our simulated groups, except for three parameters, in which the estimates from the Hong Kong Chinese group were used: The coefficients from Attitudes and Temperament to Actual affect, and the coefficient from Actual affect to Rigorous activities. We then simulated 100 data sets using the same sample sizes as in the first simulation study. The true parameter values used in this simulation are shown in the middle part of TABLE 3.

The data were simulated using the simulateData function in the lavaan package.

A summary of the results from the two simulation studies is shown in TABLE 3. In the first simulation study, the proportion of times the procedure estimated the δ-coefficients to something other than exactly 0 varied between around 55% and 70%, and the proportion does not seem to be related to the sample size. In other words, there seem to be at least a 30% chance that a true 0 δ parameter is estimated to something other than 0. The Mean Squared Error (MSE) and bias observed in the simulations are small for all sample sizes, indicating that the magnitude of the non-zero δ's are usually small.

In the second simulation study, the proportion of times the procedure estimated the δ-coefficients to something other than 0 depended strongly on whether the δ parameters were truly zero or not. The truly non-zero δ's were usually estimated to be non-zero at least 97% of the times. The truly zero δ's were however not estimated to exactly zero as often as in the first simulation study, usually in the range of 10 to 30 percent.

In both simulation studies, the mean squared errors of the parameter estimates were small, but since the main reason for using the Lasso technique is to estimate some parameters *exactly* to 0, the results of the second simulation study show that the optimal λ based on cross-validation might not penalize the estimates enough. This is a known issue in the lasso literature, where a "one-standard-error" rule might be used instead to set the λ (Hastie et al., 2015). The rule is to set a stricter λ that is within one standard error of the cross-validation error. We therefore did a variant of the second simulation study using this rule of thumb. The results are shown in the bottom part of TABLE 3. The proportion of times the true δ's was estimated to exactly 0 were greater than in the second study, but still lower than in the first, typically in the 30 to 55 percent range.

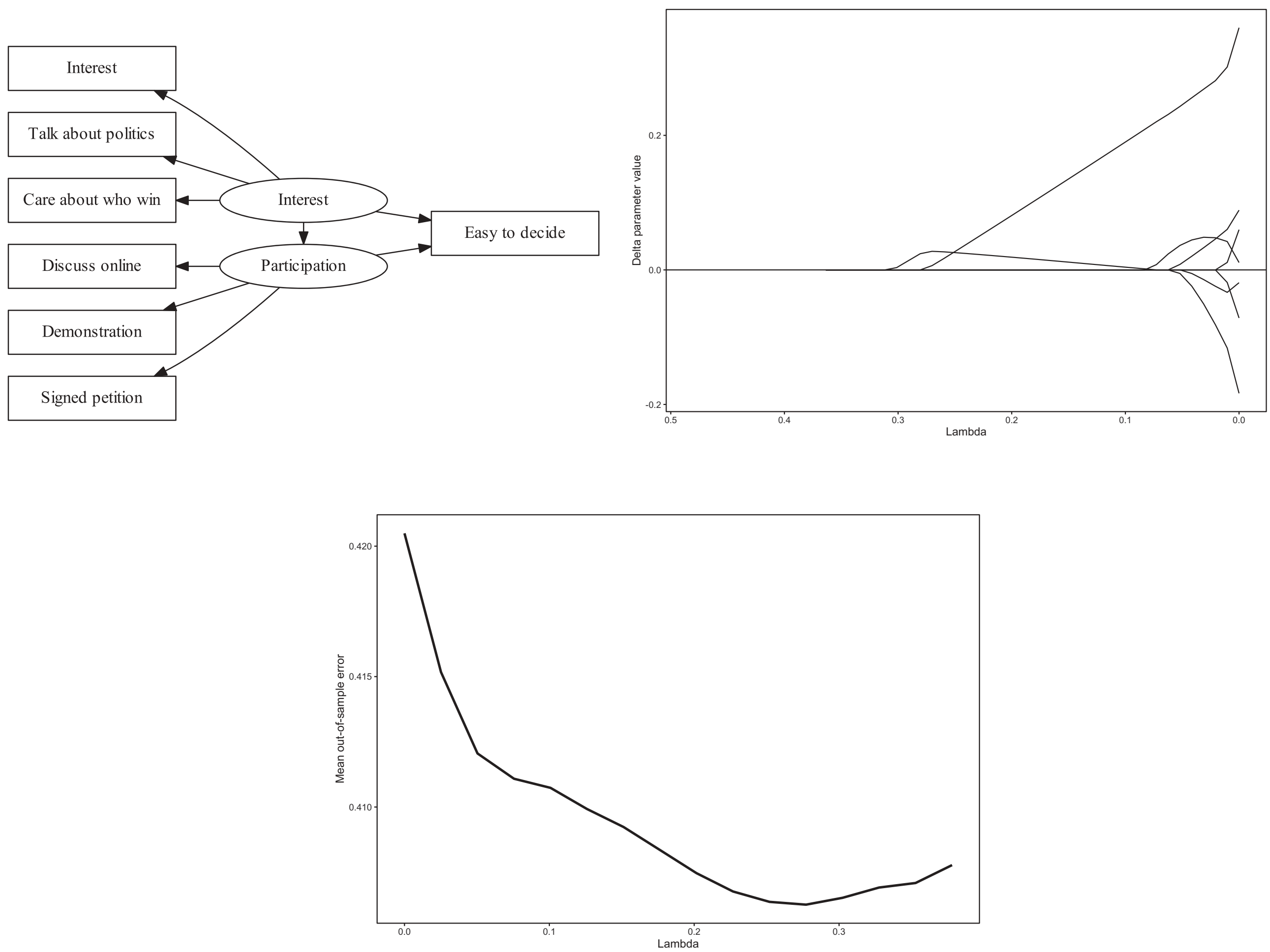



FIGURE 4 The Norwegian election survey model. (a) The path diagram for the model. The latent variables are represented with ellipses, while the observed are rectangular. (b) The evolution of the estimates for the $\delta$ parameters for a range of $\lambda$. (c) Average out of sample error after five rounds of cross-validation for a range of $\lambda$ values. The minimum out of sample error was achieved when $\lambda = 0.28$.

## DISCUSSION

In this paper, we have presented a method for doing group analyses of structural equation models, with the goal of identifying regression coefficients that differ between groups. The method is based on the L1-penalization technique which we employ by parameterizing the coefficients in a group by how much they differ from the coefficients in an arbitrarily chosen reference group. The consequence of the L1-penalization is that some of these differences are estimated to be 0, indicating that the coefficients are equal in the two groups.

The penalized method we have presented provides a data-driven alternative to the commonly used manual specification of the cross-group constraints. This represents one possible application of the framework introduced in (Jacobucci et al., 2016).

In our presentation of our two case studies, we have emphasized the identification of the non-zero differences between two groups. We also considered the resulting point estimates and how they differed from the non-penalized ML estimates. While these are interesting, they should be interpreted with care. It is for instance difficult to obtain standard errors from penalized estimates, which means that standard methods for inference do not apply. Our method is therefore best suited for exploratory analysis, and should not be considered a substitute for testing specific hypotheses.

On both case studies, the non-zero coefficients were not simply the coefficients with the greatest magnitudes of the ML estimates. This is because the penalized estimation also takes into account the co-linearity among the variables. This phenomenon can be witnessed in the plots showing the evolution of the parameters when the penalization is changed. This is most evident in the analysis of the election survey data, where the trajectories change as new non-zero parameters enter the model. Accounting for co-linearity when the model is manually re-specified with new

TABLE 3
Summary of the Simulation Studies. The True Coefficients from Group 1 are Listed in the Columns $\theta_1$ and the $\delta$ Parameters are Shown in the $\delta$ Column. The Top Part Show the Results of Study 1, Where All $\delta$ Parameters are Exactly 0. The Middle Part Show the Result of a Study Where Some of the $\delta$'s are Not 0. The Bottom Part Shows the Results of a Variant of the Second Study, but the Where the Final $\lambda$ Is Set Using the Stricter "one-standard-error" Rule

| | | | | *N = 71* | | | *N = 150* | | | *N = 250* | | | *N = 500* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Response* | *Predictor* | $\theta_1$ | $\delta$ | $P(\delta \neq 0)$ | *MSE* | *Mean bias* | $P(\delta \neq 0)$ | *MSE* | *Mean bias* | $P(\delta \neq 0)$ | *MSE* | *Mean bias* | $P(\delta \neq 0)$ | *MSE* | *Mean bias* |
| Depression | Actual Affect | −0.08 | – | 0.64 | 0.01 | −0.00 | 0.69 | 0.00 | −0.00 | 0.71 | 0.00 | −0.00 | 0.63 | 0.00 | 0.00 |
| Rigorous activties | Actual Affect | −0.00 | – | 0.59 | 0.01 | 0.00 | 0.67 | 0.00 | −0.00 | 0.60 | 0.00 | 0.00 | 0.52 | 0.00 | −0.00 |
| Actual affect | Attitudes | 0.19 | – | 0.69 | 0.00 | −0.00 | 0.70 | 0.00 | −0.01 | 0.68 | 0.00 | −0.00 | 0.61 | 0.00 | −0.00 |
| Depression | Attitudes | 0.04 | – | 0.63 | 0.01 | 0.01 | 0.66 | 0.00 | −0.00 | 0.58 | 0.00 | 0.00 | 0.52 | 0.00 | −0.00 |
| Ideal affect | Attitudes | 0.17 | – | 0.57 | 0.01 | 0.01 | 0.66 | 0.00 | 0.00 | 0.55 | 0.00 | −0.01 | 0.59 | 0.00 | 0.00 |
| Rigorous activties | Attitudes | −0.10 | – | 0.63 | 0.00 | −0.00 | 0.63 | 0.00 | 0.00 | 0.64 | 0.00 | 0.00 | 0.58 | 0.00 | −0.00 |
| Depression | Ideal Affect | 0.02 | – | 0.57 | 0.01 | −0.01 | 0.69 | 0.00 | 0.00 | 0.65 | 0.00 | −0.00 | 0.57 | 0.00 | −0.01 |
| Rigorous activties | Ideal Affect | 0.11 | – | 0.60 | 0.01 | −0.00 | 0.60 | 0.00 | −0.01 | 0.59 | 0.00 | 0.00 | 0.56 | 0.00 | 0.00 |
| Actual affect | Temperament | 0.44 | – | 0.64 | 0.00 | 0.01 | 0.61 | 0.00 | 0.01 | 0.61 | 0.00 | −0.00 | 0.71 | 0.00 | 0.00 |
| Depression | Temperament | −0.47 | – | 0.69 | 0.01 | 0.00 | 0.73 | 0.00 | 0.00 | 0.70 | 0.00 | 0.00 | 0.62 | 0.00 | 0.00 |
| Ideal affect | Temperament | 0.25 | – | 0.58 | 0.01 | −0.01 | 0.67 | 0.00 | 0.01 | 0.63 | 0.00 | −0.00 | 0.60 | 0.00 | −0.00 |
| Rigorous activties | Temperament | 0.15 | – | 0.61 | 0.01 | 0.01 | 0.62 | 0.00 | 0.01 | 0.57 | 0.00 | −0.00 | 0.59 | 0.00 | 0.00 |
| Depression | Actual Affect | −0.03 | – | 0.26 | 0.02 | −0.02 | 0.20 | 0.01 | 0.01 | 0.17 | 0.01 | −0.00 | 0.09 | 0.01 | 0.00 |
| Rigorous activties | Actual Affect | −0.38 | 0.53 | 0.01 | 0.04 | 0.08 | 0.00 | 0.02 | 0.05 | 0.00 | 0.01 | 0.04 | 0.00 | 0.01 | 0.03 |
| Actual affect | Attitudes | 0.01 | 0.28 | 0.01 | 0.02 | 0.03 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | −0.01 | 0.00 | 0.00 | −0.02 |
| Depression | Attitudes | 0.03 | – | 0.26 | 0.02 | 0.01 | 0.27 | 0.01 | −0.01 | 0.24 | 0.01 | −0.00 | 0.18 | 0.00 | 0.00 |
| Ideal affect | Attitudes | 0.19 | – | 0.27 | 0.01 | −0.02 | 0.14 | 0.01 | 0.00 | 0.20 | 0.00 | 0.00 | 0.17 | 0.00 | 0.01 |
| Rigorous activties | Attitudes | −0.18 | – | 0.21 | 0.02 | −0.04 | 0.21 | 0.01 | −0.01 | 0.13 | 0.00 | −0.02 | 0.16 | 0.00 | −0.01 |
| Depression | Ideal Affect | 0.05 | – | 0.17 | 0.03 | −0.02 | 0.22 | 0.01 | −0.00 | 0.23 | 0.01 | −0.00 | 0.09 | 0.00 | −0.01 |
| Rigorous activties | Ideal Affect | 0.21 | – | 0.30 | 0.02 | 0.02 | 0.27 | 0.01 | 0.01 | 0.18 | 0.01 | 0.01 | 0.17 | 0.00 | −0.01 |
| Actual affect | Temperament | 0.61 | −0.36 | 0.03 | 0.03 | −0.09 | 0.00 | 0.01 | −0.06 | 0.00 | 0.01 | −0.04 | 0.00 | 0.00 | −0.04 |
| Depression | Temperament | −0.40 | – | 0.22 | 0.02 | −0.00 | 0.18 | 0.01 | −0.00 | 0.25 | 0.01 | 0.00 | 0.12 | 0.00 | −0.00 |
| Ideal affect | Temperament | 0.31 | – | 0.24 | 0.01 | −0.01 | 0.28 | 0.01 | −0.01 | 0.24 | 0.01 | 0.01 | 0.12 | 0.00 | 0.00 |
| Rigorous activties | Temperament | 0.37 | – | 0.23 | 0.03 | −0.01 | 0.22 | 0.01 | 0.01 | 0.27 | 0.01 | 0.02 | 0.14 | 0.00 | 0.02 |
| Depression | Actual Affect | −0.03 | – | 0.62 | 0.01 | −0.01 | 0.46 | 0.00 | 0.01 | 0.34 | 0.01 | −0.01 | 0.30 | 0.00 | 0.00 |
| Rigorous activties | Actual Affect | −0.38 | 0.53 | 0.05 | 0.06 | 0.16 | 0.00 | 0.03 | 0.10 | 0.00 | 0.01 | 0.07 | 0.00 | 0.01 | 0.05 |
| Actual affect | Attitudes | 0.01 | 0.28 | 0.15 | 0.03 | 0.13 | 0.02 | 0.01 | 0.07 | 0.00 | 0.01 | 0.03 | 0.00 | 0.00 | 0.00 |
| Depression | Attitudes | 0.03 | – | 0.64 | 0.01 | −0.00 | 0.56 | 0.00 | −0.01 | 0.42 | 0.00 | −0.00 | 0.44 | 0.00 | 0.00 |
| Ideal affect | Attitudes | 0.19 | – | 0.56 | 0.01 | −0.01 | 0.47 | 0.00 | 0.00 | 0.41 | 0.00 | 0.01 | 0.43 | 0.00 | 0.01 |
| Rigorous activties | Attitudes | −0.18 | – | 0.45 | 0.01 | −0.03 | 0.44 | 0.00 | −0.01 | 0.38 | 0.00 | −0.02 | 0.44 | 0.00 | −0.01 |
| Depression | Ideal Affect | 0.05 | – | 0.52 | 0.01 | −0.01 | 0.49 | 0.00 | 0.00 | 0.42 | 0.00 | −0.00 | 0.34 | 0.00 | −0.00 |
| Rigorous activties | Ideal Affect | 0.21 | – | 0.55 | 0.01 | 0.00 | 0.55 | 0.00 | −0.00 | 0.39 | 0.00 | 0.00 | 0.40 | 0.00 | −0.01 |
| Actual affect | Temperament | 0.61 | −0.36 | 0.15 | 0.05 | −0.18 | 0.02 | 0.02 | −0.12 | 0.00 | 0.01 | −0.07 | 0.00 | 0.01 | −0.06 |
| Depression | Temperament | −0.40 | – | 0.49 | 0.01 | −0.00 | 0.46 | 0.01 | −0.00 | 0.45 | 0.00 | 0.00 | 0.39 | 0.00 | −0.00 |
| Ideal affect | Temperament | 0.31 | – | 0.54 | 0.00 | −0.01 | 0.59 | 0.00 | −0.00 | 0.45 | 0.00 | 0.00 | 0.32 | 0.00 | 0.00 |
| Rigorous activties | Temperament | 0.37 | – | 0.47 | 0.01 | −0.01 | 0.51 | 0.01 | 0.00 | 0.45 | 0.00 | 0.01 | 0.40 | 0.00 | 0.01 |

constraints is a difficult task, and the penalized procedure we have presented here seems to handle it.

Our simulation studies based on the Affect Valuation Theory case study explore some properties of the procedure. They show that non-zero $\delta$'s will often be estimated to be non-zero, but whether exactly zero $\delta$'s will be estimated as such, is not necessarily easy to answer and depend on both the data and model, as well as on the amount of penalization. In light of the discussion of the case studies, it is worth repeating that the procedure presented here is not intended as a replacement for classical hypothesis testing, but could provide some interesting exploratory results.

The R scripts and data sets used to compute the results in this paper are available from www.github.com/jcl7/group_sem_lasso.

## ACKNOWLEDGMENTS

We would like to thank Professor Jeanne L. Tsai for allowing us to use her data set for the Affect Valuation Theory model.

## REFERENCES

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233–255. doi:10.1207/S15328007SEM0902_5

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The Lasso and generalizations*. (Number 143 in Monographs on statistics and applied probability). Boca Raton, FL: CRC Press. OCLC: 921942690.

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

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: a Multidisciplinary Journal*, *23*, 555–566. doi:10.1080/10705511.2016.1154793

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409–426. doi:10.1007/BF02291366

Jung, S. (2013). Structural equation modeling with small sample sizes using two-stage ridge least-squares estimation. *Behavior Research Methods*, *45*, 75–81. doi:10.3758/s13428-012-0206-0

Kleven, Y., Aardal, B., Berghm, J., Hesstvedt, S., & Hindenes, D. (2015). *Valgundersøkelsen 2013. dokumentasjons- og tabellrapport* (Technical report). Norway: Statistics Norway.

McArdle, J. J., & McDonald, R. P. (1984). Some algebraic properties of the reticular action model for moment structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 234–251. doi:10.1111/bmsp.1984.37.issue-2

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Chapman & Hall/CRC statistics in the social and behavioral sciences series. Boca Raton, FL: CRC Press/Taylor & Francis. OCLC: 836828151.

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

Tsai, J. L., Knutson, B., & Fung, H. H. (2006). Cultural variation in affect valuation. *Journal of Personality and Social Psychology*, *90*, 288–307. doi:10.1037/0022-3514.90.2.288