Research Master's programme:
Methodology & Statistics for the Behavioural, Biomedical & Social Sciences
Utrecht University, the Netherlands

MSc Thesis Pascal van Luit (5594103)
Improving Generalizability of Structural Equation Models through
Cross-Validated Model Modifications
May 2020

Supervisors:
Dr. Daniel Oberski
Erik-Jan van Kesteren MSc

Second Grader:
Dr. Remco Feskens

Preferred journal of publication: Structural Equation Modelling: A Multidisciplinary Journal
Word count: 6241

Improving Generalizability of Structural Equation Models through Cross-Validated Model Modifications

Research Master Thesis by Pascal van Luit

Supervised by Erik-Jan van Kesteren & Daniel Oberski

Abstract

Structural equation modelling (SEM) is a popular modelling tool in the behavioral and social sciences. SEM models are often modified with the aid of modification indices; which are computed based on a sample dataset. This practice comes with a risk for overfitting to the sample dataset. To prevent overfitting, this paper proposes SEM model modification methods in combination with cross-validation. A simulation study is run to assess the performance of the standard modification method with the proposed cross-validation methods. Results indicate that cross-validating model modifications in SEM can be effective in obtaining a lower Mean Squared Error of estimating a parameter of interest.

Keywords: cross-validation, structural equation modelling, generalizability, model modification, model specification search

Introduction

Structural equation modeling (SEM) is a popular method for creating a model and investigating hypotheses about parameters of interest. More specifically, SEM is often used to model latent variables; variables that are not directly observable but are inferred from variables that are directly observable. An example of a parameter of interest could be whether there is a covariance between social media usage and social media benefit (Abrahim, Mir, Suhara, Mohamed, & Sato, 2019). Both variables are not easily directly observable but they can be inferred from observable variables such as perceived enjoyment, perceived usefulness, educational usage of social media and more (Abrahim et al., 2019).

Ideally, the model matches the data generating process (DGP) that is present in the population. Achieving that, however, is a difficult task because SEM models are built using only a sample from a population. The execution of this task comes with the risk of model misspecification: specifying a model which does not match the DGP. Models remain an approximation to the truth and therefore never perfectly resemble the truth (Burnham & Anderson, 2004). Misspecifying models may also lead to parameter estimates which are biased (Yuan, Marshall, Bentler, Yuan, & Bentler, 2003). Having models that contain bias is undesirable as researchers want parameter estimates which are as close as possible to true population parameter values.

Many solutions for the problem of model misspecification exist – this is broadly referred to as model specification search. The solutions aim to adjust a baseline model to make it fit well to a dataset (Marcoulides & Falk, 2018). A popular method is to decide which model parameters should be freely estimated based on modification indices (MIs). MIs are an estimate of how much a model fit would improve if a certain parameter is freely estimated instead of restricted to zero. MI values are data-driven which makes the procedure of model adjustment according to MIs susceptible to capitalizing on chance characteristics of the data (MacCallum, Roznowski, & Necowitz, 1992). Consequently, models are prone to *overfitting*; the model fits exceptionally well to the data at hand, but fails to generalize well to new data from the same DGP. Models in which many parameters are added according to MI's also run the risk of becoming overparameterized and having increased variance - parameter estimates vary a lot depending on the sample on which the model is fit (Cudeck & Browne, 1983).

When specifying a model, the trade-off between the bias and the variance of parameter estimates needs to be considered. One the one hand, it is desirable for the model to have minimized bias so that parameter estimates are as close as possible to the theoretical true parameter value. The dataset is a sample from the DGP, which means that parameters values of the dataset resemble the parameter values of the DGP to a certain extent. When building a model one would therefore try to decrease bias to have model estimates match parameter values in the sample. On the other hand, there comes a point at which bias reduction leads to an increase in variance of parameter estimates (James, Witten, Hastie, & Tibshirani, 2013). This is the point at which a model starts overfitting to a dataset; sample characteristics start to become overly present in the model parameter estimates. An overparameterized model does not generalize well to new samples from the same DGP, because samples all differ from each other and do not perfectly match the DGP. Ultimately, it is most desirable for a model to have minimized bias and minimized variance.

The field of statistical learning presents several techniques to optimise the bias-variance trade-off. Most prominently, cross-validation can be applied to estimate the generalizability of a model (Browne & Cudeck, 1989; Cudeck & Browne, 1983). Cross-validation allows a model to be both built and tested for generalizability using only one dataset. The technique splits the dataset into separate sets; sets which are used for specifying a model and sets which are used for testing the model. The performance of a model on the sets which were not used to specify the model, can then be indicative of how well a model performs on data which was not used to specify the model.

In this paper we present a method to combine SEM model specification with cross-validation to obtain parameters of interest with low generalization error. In this method, the estimated generalization error found during cross-validation acts as a stopping mechanism on the model specification search. The search is stopped when the optimal model is obtained. This method is shown to be more conservative in adding modifications and is thus less prone to specifying an overfitted SEM model.

This paper is organised as follows. First, the necessary background information on SEM, model specification search and cross-validation is provided. Second, the method for obtaining the optimal model using cross-validation is introduced. Third, the performance of this method is shown relative to existing methods of using MI's. Lastly, a conclusion is provided with suggestions for further research.

Background

Model Specification Search in SEM

SEM presents the possibility to investigate several parameters concerning relationships between observed variables in a dataset and underlying latent constructs. Such investigation is performed by searching for a model that has a covariance structure which matches the covariance structure present between the observed variables in a dataset (Hox, 1999). There are several types of SEM, such as path analysis, confirmatory factor analysis and latent growth modeling.

In this paper, a two-factor confirmatory factor analysis (CFA) model is taken under the scope. CFA models are very common in SEM (Khine, 2013; Preacher, 2010; Cudeck & Browne, 1983), and are therefore chosen as a running and motivating example. In research practice, CFA's often have several latent factors. The simulation study in this paper is performed with two latent factors for simplicity of analysis and interpretation. This method can easily be extended to more complex models at a later stage.

A two-factor CFA model is a model consisting of two latent factors which account for the variance found in the observed variables. When specifying such a model in SEM, the assumption is made that the DGP is also a two-factor CFA. The exact relationships between the latent factors and the observed variables is what one then tries to identify. The objective is to specify a two-factor CFA model which best resembles the DGP. The model modification procedure starts with a baseline model that specifies independence among the observed variables and that each observed variable has a loading from only one latent factor. Next, the idea is to find a better fitting model by freeing selected restricted parameters (Marcoulides & Falk, 2018). By doing so, the model tries to free the parameter(s) which are restricted in the baseline model. An example of this is a cross-loading which is present in the DGP but not in the baseline model.

Model Modification and Modification Indices

Model adjustment in SEM is often performed according to MI's and theory about the variables in the dataset (Marcoulides & Falk, 2018). An MI is an estimate of the amount by which the chi-square goodness of fit statistic, a popular measure of model fit, is expected to decrease if a parameter is freed in the model. MIs can thus be used to identify parameters which have been restricted in the model and which, if freely estimated, would improve the fit of the model to the data. Employing MI's to make decisions about which parameters to freely estimate in the model aids in the search for the correct model specification as is present in the DGP. However, this method is accompanied with a risk for overfitting. MI values are data-driven and aid in specifying a model which fits well to the data at hand (Browne & Cudeck, 1992). This method puts models at risk of capitalizing on chance characteristics of the sample (MacCallum et al., 1992).

Overfitting

Methods to reduce the risk of overfitting when using MIs do exist. For example, it is common practice to examine the MI according to a theoretical framework to determine the plausibility of freeing a certain parameter (Whittaker, 2012). An MI can recommend a researcher to add a free parameter, and the researcher must decide whether freeing the parameter is in accordance with theory. This avoids including freely estimated parameters which go against established theory. In some cases however, there is no well established theory to go by and an exploratory factor analysis may be performed (Morin et al., 2013). In such a case, there is no literature theory to help in preventing overfitting of the model.

Furthermore, the MI method is generally implemented with a minimum MI requirement for adding a free parameter. A relevant MI cutoff is 3.84 as this corresponds to the critical value of a chi-square distribution with 1 degree of freedom at $\alpha = .05$. This means that freeing a parameter with an MI of greater than 3.84 would significantly improve the fit of the model to the data. SEM manuals often advise 10 as a minimum value for adding a free parameter (Muthén & Muthén, 2010). Using different minimum values can lead to models with different parameters and can therefore influence the fit of the model which is finally specified.

Adding multiple free parameters in a model heightens the complexity of the model and can lead to models which are overparameterized (Cudeck & Browne, 1983). In SEM, the risk for overfitting is

suppressed by having preference for parsimonious models (Preacher, 2010). This means that between models which have similar fit, the model with fewer free parameters is preferred. The strive for parsimony is also reflected in other SEM model fit measures such as the Tucker Lewis Index (TLI) and Root Mean Square Error of Approximation (RMSEA) (Hooper, Coughlan, & Mullen, 2008), which simultaneously lower model complexity and reduce the susceptibility to overfitting to a small extent.

The Bias-Variance Trade-Off

The trade-off between bias and variance is an indicator of the underfitting or overfitting of a model. Adding freely estimated parameters to a model brings parameter estimates closer to the true parameter values of a sample (Kolenikov, 2011). This procedure increases the complexity of a model, and reduces bias of the estimates.

However, continuing to add free parameters (based on MI values of a sample) to a model also leads to increased variance of the estimate of the parameter of interest. This is because the characteristics of different samples determine the MI values and thus heavily influence the specification of the model and its freely estimated parameters. Continuing a procedure that adds free parameters without any stopping mechanism can therefore be detrimental to the generalizability of a model to new datasets.

Cross-Validation

Cross-validation presents the option to assess model generalizability within a single sample taken from a population (Barrett, 2007; Schreiber, 2006). The performance of a model on new unseen data can be estimated by testing model performance within the dataset on which the model is built.

The specific class of cross-validation used in this paper is called *k-fold* cross-validation. *k-fold* cross-validation enables the assessment of model generalizability by splitting a dataset up into k groups. *k - 1* groups are used as a training set to build a model and the model is then assessed by its performance on the k^{th} group; the k^{th} group is referred to as the validation set. Model performance on the validation set gives an idea for how well a model generalizes to data which was not used to build it. This is repeated k times so that each set is used once as the validation set. The average performance can then be used as an indicator for the generalizability of a model. This technique does however come with some computational and time costs because model fitting needs to be performed k times instead of just once.

Recommendations to apply the technique of cross-validation in SEM have been made before so that the process of model selection is prudent (Yuan, Marshall, & Weston, 2002). Such model selection is prudent because a stopping mechanism can be enforced on the procedure of adding free parameters. This practice can be combined with the practice of selecting free model parameters based on theory and reason (Whittaker, 2012). The next section proposes a method to perform SEM utilizing the tools of cross-validation and MI model adjustment.

Methods[1]

Standard Procedure of Model Modification

To assess the proposed methods of model modifications, their performance is compared to the performance of using only MIs. This method can also be referred to as a "greedy search" as it frees any parameters which the dataset at hand suggests would improve the model fit. In practice, MIs are often used in combination with existing theory from literature. In this simulation study, theory is omitted and modifications are made solely based on MIs. This is done for reasons of analysis simplicity, and also because in practice, datasets can suggest parameters with large MIs but without support from theory in literature. When this happens, a researcher is faced with making a difficult decision about whether or not to free a parameter. Given that model fit would improve by freeing the parameter, it is a realistic scenario that the parameter is freed. To keep this decision process consistent, the factor of literature theory is taken out of the decision process in this study. Possible implications of this are also highlighted in the discussion section.

The standard procedure runs an algorithm which takes as input: a baseline model, a dataset, and a minimum MI requirement. The algorithm fits the baseline model to the dataset and obtains MIs of the restricted parameters. These MIs serve as the indicator for whether or not an additional parameter is freed. The algorithm checks if the largest MI value is higher than the minimum MI requirement; if

[1] All methods utilize functionality made possible by the *lavaan* R package maintained by Yves Rosseel (Rosseel, 2012).

the maximum MI is above this threshold, the restricted parameter with the largest MI value is freed, and the new model is fit to the dataset. Subsequently new MI values are obtained for the remaining restricted parameters, and the process is repeated. This algorithm continues until the largest MI value is no longer greater than the minimum MI value specified at the beginning. Finally, the final model, which is the baseline model plus any additionally freed parameters, is provided in the output. The procedure is summarized in **Algorithm 1**.

Algorithm 1: Model Modification with Modification Indices

```
fit model
obtain MIs
while max(MI) > min.mi do
    free parameter with largest MI
    refit new model and recompute MIs
end
print final model
```

Cross-Validated Model Modifications

Two novel methods for modifying structural equation models are presented in this paper. The methods are developed using a combination of cross-validation, modification indices and chi-square model fit statistics. Recall that MIs are an estimate of the improvement in model fit based on a single dataset, and are not necessarily an estimate of the improvement of a model to the true DGP. Thus, to a certain extent MIs are useful as a basis for making modifications to a model. However, a braking mechanism is also necessary to prevent a modification procedure from getting carried away and overfitting to the data at hand. The next two sections each explain a newly proposed method for SEM model specification. The methods each have their own braking mechanism which aims to halt and prevent the modification procedure from overfitting, and thus specify models which generalize better to new unseen data.

Cross-Validation using Modification Indices

The first method tries to correct for the aspect that MIs are based on a single dataset. The method tries to do so by using out-of-sample (OOS) MIs to determine which additional parameter should be freed in a model. More specifically, the method computes k OOS MIs. To use the method the following is specified: a baseline model, a dataset, a minimum MI requirement and k. A dataset is split into k groups and the baseline model is fit k times, each time on *k-1* groups. The group which is left out is then used to compute OOS MI values by fitting the model which was found based on the *k-1* groups and preserving the model parameters found. Subsequently, MIs are provided based on the OOS group. This is an applied form of *k-fold cross-validation*.

This procedure is performed k times as each group is left out once. All OOS MI values are combined and their mean is computed. The mean OOS MI value serves as the value which is compared to the minimum MI requirement. Similar to the standard procedure of model modification, the restricted parameter with largest OOS MI value is freed in the model if its value is greater than the minimum MI requirement. This procedure is summarized in **Algorithm 2**.

Algorithm 2: Model Modification based on OOS MIs

```
split data into k groups
fit model k times on k - 1 groups
obtain MI values k times from group which is left out
take mean of OOS MI values
while max (mean OOS MI) > min.mi do
    free parameter with largest mean OOS MI
    repeat steps 1 to 4
end
print final model
```

Cross-Validation using Chi-Square Model Fit Statistic

The second proposed method also utilizes *k-fold cross-validation*, by combining MIs and the chi-square model fit statistic (χ^2). This method tries to implement a braking mechanism on the model modification procedure by continuously checking model fit on OOS data. Similarly, a baseline model, a dataset, a minimum MI requirement, k (number of groups), and a significance level α are specified.

A dataset is split into k groups and the baseline model is fit k times, each time on k - *1* groups. MI values are computed based on those k - *1* groups and the restricted parameter with the largest MI, given that it is greater than the minimum MI requirement, is freed in the model. Following this, the new model is fit to the group which was left out. From this group, a chi-square fit is obtained to indicate how well the new model fits on data which was not used to compute MI values. Once again, the procedure is performed k times as each group is left out once. All OOS fits are combined and their mean is computed. If the mean OOS fit is a significant χ^2 fit, then the additionally freed parameter is accepted. If the mean OOS fit is not a significant improvement, the new free parameter is rejected. This procedure is summarized in **Algorithm 3**.

Algorithm 3: Model Modification based on OOS χ^2 Fits

```
split data into k groups
fit model k times on k - 1 groups
obtain MI values from those k - 1 groups
obtain k χ² fits from groups which are left out
take mean of OOS χ²
while mean OOS χ² is significant do
    free parameter with largest MI
    repeat steps 1 to 5
end
print final model
```

Simulation Study[2]

To compare the performance of cross-validation assisted model modifications to existing model modification procedures, a simulation study is performed.

Data in the simulation study is generated based on a confirmatory factor analysis (CFA) model with two latent factors and three observed variables per factor, plus a cross-loading. This is visualized in Figure 1.

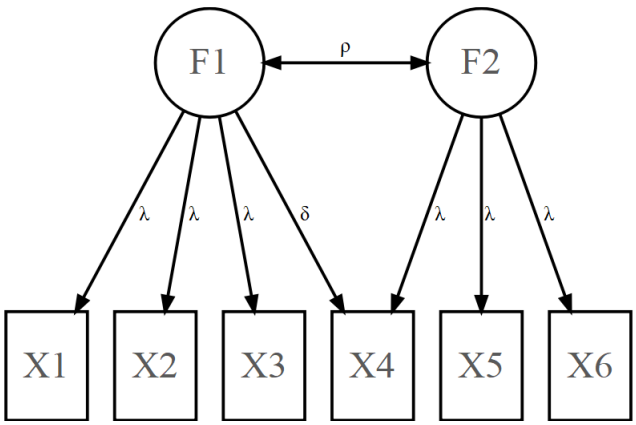


Figure 1: DGP: Two-factor CFA model with a cross-loading

[2]This study is approved by the Ethics Committee of the Faculty of Social and Behavioural Sciences of Utrecht University (file number: 19-222)

Figure 1 includes three types of parameters:

- ρ represents the covariance strength between the two latent factors. This is the **parameter of interest**, for example representing a covariance between social media usage and social media benefit (Abrahim et al., 2019).
- λ represents the factor loading strength of the observed variable on the respective latent factors. The factor loading is the correlation between the two variables. This is kept equal for all observed variables for simplicity of analysis.
- δ represents the cross-loading strength which is present in the DGP but is omitted in the baseline SEM model which is specified in the model specification search algorithms.

Conditions for the Simulation Study

The conditions for the simulation differ in four aspects: sample size (100, 200, 500), strength of δ (0.1, 0.3, 0.5), strength of λ (0.1, 0.3, 0.5) and strength of ρ (0.1, 0.3, 0.5). The (in)ability of a model specification search to identify δ is indicative of how well the modelling procedure is able to identify a restricted parameter that should actually be freely estimated. The factor loading strength (λ) is referred to as the reliability because it indicates how strong the correlation between the factors and variables is. A higher λ corresponds to more variance in an observed variable being explained by the factor. δ represents the loading F1 has on X4 - a loading which is not specified in the baseline SEM model.

To summarize the conditions there are three sample sizes (100, 200, 500), three cross-loading strengths, three reliability values and three strengths of the parameter of interest. This makes a total of 81 (3 sample sizes $\times$ 3 δ's $\times$ 3 λ's $\times$ 3 ρ's) conditions. For each condition, 50 simulated datasets are computed. This makes for a total of 4050 simulated datasets.

Comparison of Specification Search Methods

On each dataset, four approaches of model modification are performed: no modification, standard procedure of model modification, model modification with cross-validated MIs, and model modification with cross-validated OOS χ^2 fits. For each method of model modification, two minimum MI criteria are used: MI4 and MI10. MI4 is derived from rounding 3.84 (the cutoff point which corresponds with 1 degree of freedom at $\alpha = .05$) for simplicity of analysis, and the modelling procedures can still be compared fairly and effectively. This makes for a total of seven different methods. The methods are summarized in Table 1.

Table 1: Summary of Methods

Method	**Description**
No Mod	The baseline model is fit to the data.
MI4	Algorithm 1 is used with a minimum MI of 4.
MI10	Algorithm 1 is used with a minimum MI of 10.
MI-CV4	Algorithm 2 is used with a minimum MI of 4.
MI-CV10	Algorithm 2 is used with a minimum MI of 10.
CHI-CV4	Algorithm 3 is used with a minimum MI of 4.
CHI-CV10	Algorithm 3 is used with a minimum MI of 10.

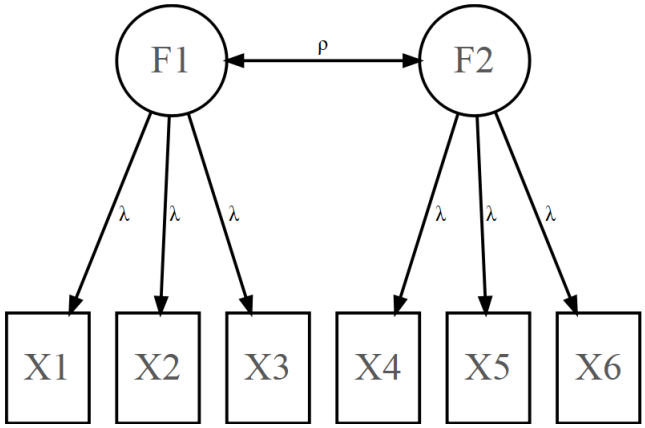
The performances of these methods are compared in their ability to:

1. reproduce the DGP covariance matrix
2. correctly estimate the parameter of interest (ρ).

All methods begin with the same baseline model which differs from the DGP model in one way: there is no cross-loading included. This is visualized in Figure 2. Modelling procedures are therefore challenged in their ability to recover the true DGP. The ability to recover the true DGP is measured by comparing the modelled covariance matrix with the true covariance matrix present in the DGP. The distance between the modelled covariance matrix and the true covariance matrix provides a measure for how well a procedure performs in finding a model that fits well to the data. Throughout this paper,

this measure is denoted as Σ. Furthermore, a good model specification procedure will free only the cross-loading parameter (present in the DGP, and not in the baseline model) and no other parameter.

Figure 2: Baseline Model



ρ is the defined as the parameter of interest - meaning that in this case we have the greatest interest in estimating this parameter as accurately as possible. We want the estimated ρ to be as close as possible to the ρ present in the DGP. To measure a modelling procedure's ability in correctly estimating ρ the mean squared error (MSE) of predictions is computed. This is computed according to the following formula:

$$1/n \cdot \sum(\rho - \hat{\rho})^2 \tag{1}$$

where ρ is the latent factor covariance in the population and $\hat{\rho}$ is the estimated latent factor covariance in a model. n is the number of samples on which a SEM method is performed - in this study, n is always 50 because 50 datasets are simulated for each condition and each SEM method is performed on each dataset, giving a total of 50 estimates of ρ under each condition.

The formula for the Mean Square Error (MSE) is used because this metric penalizes large errors more severely than small errors. Each condition will have seven MSE values, one for each method of model specification. A lower MSE value indicates better generalizability as the average estimated parameter of interest is closer to the true parameter value present in the population.

Results

This section is organized as follows. First the results are reported on a general level; meaning the results are interpreted in a compiled manner and conditions are not separately looked into. Second, the results are categorized and reported under certain conditions.

For the MSE of the parameter of interest (ρ), the median is used in reporting. This is because there were a number of highly influential outliers which skewed the mean. For the covariance matrix distance measures (Σ), the mean is reported because results for this performance measure were normally distributed, making the mean a representative statistic. Furthermore, tables include results for all 7 methods. Figures only include results for methods with a minimum MI of 10 (MI10, MI-CV10 and CHI-CV10) and the results of the No Mod method. Only these four methods are included in the figures because results with a minimum MI of 4 show very similar trends as their minimum MI 10 counterparts. With four methods, figures are clearer for interpretation.

A compilation of the results and the code for reproducing the results is available on my GitHub profile[3].

General

In the case of a two-factor Confirmatory Factor Analysis (CFA) model with one cross-loading being present in the DGP, fitting the baseline model (a two-factor CFA model without the cross-loading) and applying Algorithm 3 leads to the best results. This means that fitting the baseline model (No Mod) or utilizing Algorithm 3 (CHI-CV4 & CHI-CV10) obtains the most favourable MSE of the parameter of interest and the modelled covariance matrices which are closest to the true covariance matrices (Median MSE = 0.0403, Mean Σ = 0.6361). Algorithm 3 performs almost identically with minimum MI 4 and minimum MI 10. When looking past the fifth decimal place, there are very minor negligible differences between these three methods. These differences do provide evidence that CHI-CV4 and CHI-CV10 do sometimes add a modification and do not always just do the same as No Mod.

On a general level, when using standard model modification (Algorithm 1) a minimum MI of 10 leads to a better MSE value than when a minimum MI of 4 is set. This is indicative that with a minimum MI of 4, overfitting takes place. The median MSE found when using a minimum MI of 10 is 0.043 and with a minimum MI of 4 is 0.049. This finding is also evident in the Σ measure where a minimum MI of 10 leads to a smaller mean distance from the true covariance matrix than a minimum MI of 4. See Table 2 for the exact results.

Algorithm 2 leads to worse performance measures that Algorithm 1. The median MSE of Algorithm 2 is 0.066 with a minimum MI of 4 and 0.065 with a minimum MI of 10. The mean Σ is 0.642 with a minimum MI of 4 and 0.641 with a minimum MI of 10.

To conclude the general results, using cross-validation of OOS χ^2 model fits (Algorithm 3 (CHI-CV4 & CHI-CV10)), yields the best performance measures of the three modification algorithms. This was achieved at a significance level of $\alpha = .05$. Given that Algorithm 3 yielded almost identical results as when no modifications were made, there is a suggestion that the conservative nature of Algorithm 3 helped it to perform as well as it did in this case - the case where the baseline model is only a small mismatch from the DGP in terms of free parameters.

Table 2: General Results

Method	MSE (median)	Σ (mean)
No Mod	0.0403	0.6361
MI4	0.0486	0.6453
MI10	0.0429	0.6366
MI-CV4	0.0663	0.6418
MI-CV10	0.0647	0.6411
CHI-CV4	0.0403	0.6361
CHI-CV10	0.0403	0.6361

[3]https://github.com/pascalvanluit/Master-Thesis

Results According to Different Sample Sizes

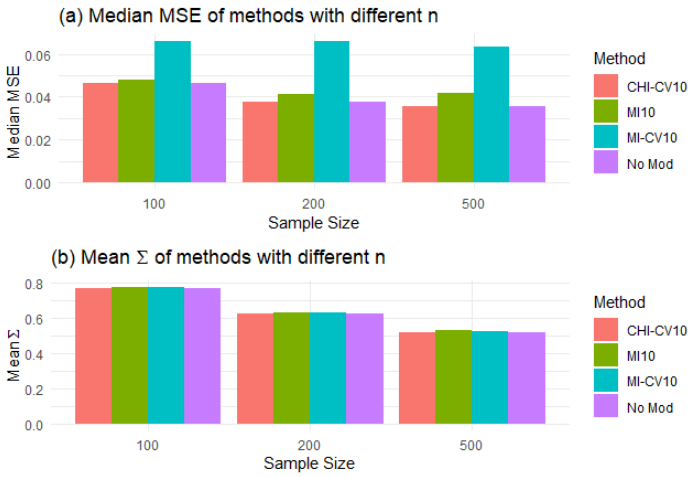
When looking at more specific circumstances it is evident that the performance of methods is related to sample size. Table 3 presents the results of the performance measures according to the different sample sizes. As is the case on the general level, No Mod, CHI-CV4 and CHI-CV10 perform the best within each sample size. Results can be found in Table 3 and are visualized in Figure 3.

Larger sample sizes lead to a smaller median MSE in almost all cases. An exception being in the case of Algorithm 1 (the standard model modification procedure). For this method, the performance improves when n increases from 100 to 200. However, for Algorithm 1 the median MSE measure worsens when n is further increased to 500. This goes against the expectation that a larger sample size improves the MSE of the estimate of the parameter of interest. When Algorithm 2 has a minimum MI of 4, the median MSE also slightly worsens when n increases from 200 to 500. Algorithm 3 continually improves as n increases, for both a minimum MI of 4 and 10. In all cases, larger sample size leads to a smaller Mean Σ. These findings are visualized in Figure 3(b). The improvement of the Mean Σ measure is greater between n = 100 to n = 200 than it when n is increased from 200 to 500.

Table 3: Results According to Different Sample Sizes

	n = 100		n = 200		n = 500	
	MSE (median)	Σ (Mean)	MSE (median)	Σ (Mean)	MSE (median)	Σ (Mean)
No Mod	0.0463	0.7664	0.0377	0.6230	0.0358	0.5191
MI4	0.0535	0.7803	0.0443	0.6315	0.0468	0.5242
MI10	0.0478	0.7667	0.0410	0.6236	0.0417	0.5195
MI-CV4	0.0675	0.7726	0.0645	0.6299	0.0670	0.5229
MI-CV10	0.0662	0.7735	0.0659	0.6285	0.0633	0.5213
CHI-CV4	0.0463	0.7664	0.0377	0.6230	0.0358	0.5191
CHI-CV10	0.0463	0.7664	0.0377	0.6230	0.0358	0.5191

Figure 3: Results According to Different Sample Sizes



Results According to Different Factor Covariance Strengths

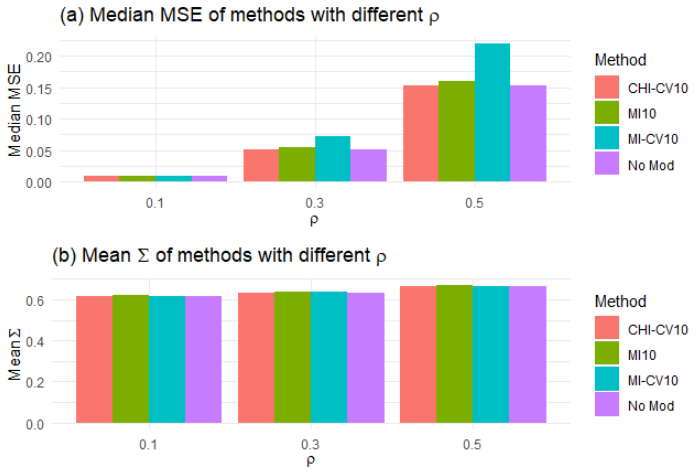
The performance of methods in estimating the parameter of interest also appears to be related to the strength of the latent factor covariance in the population (ρ). Once again, No Mod, CHI-CV4 and CHI-CV10 lead to the best results on the selected performance measures within each strength of ρ. The results are presented in Table 4 and visualized in Figure 4.

Results show that when ρ is larger in the DGP the median MSE is greater for all methods, indicating that methods are better at estimating ρ when it is smaller. The performance of methods also slightly deteriorates in terms of the mean Σ measure. As ρ gets larger, the Mean Σ also increases, indicating that all methods become less good in reproducing the covariance matrix present in the DGP. All methods perform very similarly in their ability to reproduce the covariance matrix present in the DGP.

Table 4: Results with different ρ values in the DGP

	$\rho = 0.1$		$\rho = 0.3$		$\rho = 0.5$	
	MSE (median)	Σ (Mean)	MSE (median)	Σ (Mean)	MSE (median)	Σ (Mean)
No Mod	0.00884	0.6154	0.0509	0.6300	0.1533	0.6630
MI4	0.00931	0.6239	0.0599	0.6403	0.1726	0.6717
MI10	0.00889	0.6163	0.0552	0.6305	0.1600	0.6630
MI-CV4	0.00879	0.6191	0.0731	0.6372	0.2241	0.6692
MI-CV10	0.00851	0.6191	0.0728	0.6358	0.2190	0.6685
CHI-CV4	0.00884	0.6154	0.0509	0.6300	0.1533	0.6630
CHI-CV10	0.00884	0.6154	0.0509	0.6300	0.1533	0.6630

Figure 4: Results with different ρ values in the DGP



Results According to Different Cross-loading Strengths

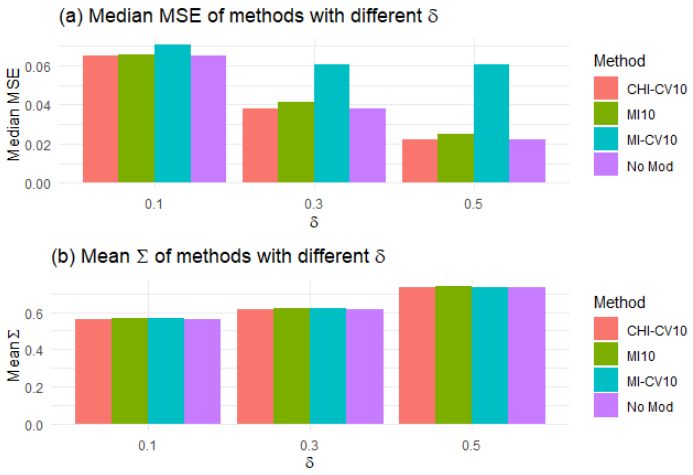
With larger δ values in the DGP all methods had lower Median MSE results, indicating that all methods perform better in estimating ρ when δ is larger in the DGP. This result is suggestive that the degree of mismatch between the baseline model and the DGP influences the ability of a modification procedure to correctly estimate the parameter of interest. The larger the mismatch, the better the estimation of the parameter of interest. This may be because a greater mismatch makes it easier for a modelling procedure to identify correct modifications to add to a model. No Mod, CHI-CV4 and CHI-CV10 lead to the best results in terms of median MSE within each strength of δ. The improvement in median MSE is least pronounced for MI-CV10, between $\delta = 0.3$ and $\delta = 0.5$ the median MSE only improves by 0.0001.

Concerning the measure of mean Σ, all methods perform very similarly under all values of δ. The methods No Mod, CHI-CV4 and CHI-CV10 all perform equally well according to the measures of median MSE (0.0648 under $\delta = 0.1$, 0.0377 under $\delta = 0.3$, and 0.0223 under $\delta = 0.5$) and mean Σ (0.5609 under $\delta = 0.1$, 0.6159 $\delta = 0.3$, and 0.7316 under $\delta = 0.5$). These results are summarized in Table 5 and visualized in Figure 5.

Table 5: Results with different δ values in the DGP

	$\delta = 0.1$		$\delta = 0.3$		$\delta = 0.5$	
	MSE (median)	Σ (Mean)	MSE (median)	Σ (Mean)	MSE (median)	Σ (Mean)
No Mod	0.0648	0.5609	0.0377	0.6159	0.0223	0.7316
MI4	0.0671	0.5712	0.0452	0.6262	0.0302	0.7386
MI10	0.0654	0.5613	0.0412	0.6167	0.0252	0.7318
MI-CV4	0.0718	0.5682	0.0591	0.6209	0.0647	0.7363
MI-CV10	0.0705	0.5677	0.0607	0.6207	0.0606	0.7350
CHI-CV4	0.0648	0.5609	0.0377	0.6159	0.0223	0.7316
CHI-CV10	0.0648	0.5609	0.0377	0.6159	0.0223	0.7316

Figure 5: Results with different δ values in the DGP



Results According to Different Factor Loading Strengths

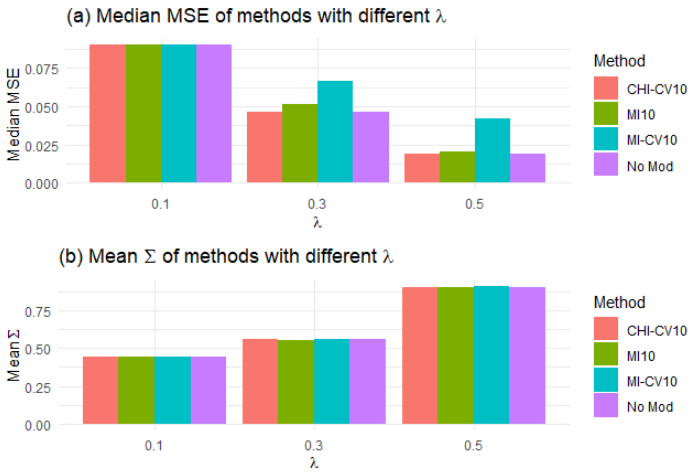
Results indicate that larger values of λ are associated with smaller median MSE's and with larger mean Σ's. The results can be found in Table 6 and they are visualized in Figure 6. A notable result is that the median MSE is the same for each method under the condition that $\lambda = 0.1$ - when the relationship between the observed variables and latent factors is weak in the DGP, all methods perform very similarly on the selected measures.

As λ increases, the median MSE decreases for all methods. Most of all for No Mod, CHI-CV4 and CHI-CV10 - these 3 methods have the most favourable results once again. On the contrary, as λ increases, the mean Σ worsens for all methods. All methods perform very similarly in terms of this measure under all conditions of λ. The results are compiled in Table 6 and visualized in Figure 6.

Table 6: Results with different λ values in the DGP

	$\lambda = 0.1$		$\lambda = 0.3$		$\lambda = 0.5$	
	MSE (median)	Σ (Mean)	MSE (median)	Σ (Mean)	MSE (median)	Σ (Mean)
No Mod	0.09	0.4448	0.0464	0.5612	0.0187	0.9025
MI4	0.09	0.4466	0.0565	0.5683	0.0229	0.9211
MI10	0.09	0.4445	0.0516	0.5607	0.0201	0.9046
MI-CV4	0.09	0.4452	0.0653	0.5634	0.0437	0.9169
MI-CV10	0.09	0.4453	0.0662	0.5633	0.0421	0.9148
CHI-CV4	0.09	0.4448	0.0464	0.5612	0.0187	0.9025
CHI-CV10	0.09	0.4448	0.0464	0.5612	0.0187	0.9025

Figure 6: Results with different λ values in the DGP



Discussion

In this thesis, and overview of current model modification methods in SEM is provided. Following this, an alternative approach for modifying models in SEM is suggested. Namely, applying cross-validation whilst making model modifications in the SEM model specification search. The current method and suggested methods were then assessed in their ability to estimate a parameter of interest: the covariance between two latent factors in a two-factor confirmatory factor analysis. They were also assessed in their ability to reproduce the covariance matrix present in the DGP.

Results from the simulation study indicate that model modifications utilizing k-fold cross-validation in the form of out-of-sample χ^2 model fits leads to the most favourable results amongst the modification procedures. Of the modification algorithms, Algorithm 3 obtains the greatest stability and smallest median MSE in the estimates of the parameter of interest (POI). On the contrary, model modifications utilizing cross-validation by obtaining out-of-sample MIs led to the worst median MSE in the estimates of the POI. The standard model modification procedure obtained performance measures which lay in between those of the two cross-validation procedures. MI-CV and the standard procedure both led to unstable estimates; this is evident in the number of outliers in MSE estimates for these two algorithms. Obtaining influential outliers is highly undesirable and is part of the reason why it remains important to check whether adding a modification to a model is supported by theory in the relevant literature. In practice, only one model is built, and it is thus very important to be cautious about adding modifications to a baseline model. One wants to minimize the risk of specifying a model which overfits to the sample data. Especially in cases where literature theory is not able to provide sufficient clarification for the decision of whether or not to free a particular additional parameter.

The median MSE results indicate that making no modification (No Mod) and applying Algorithm 3 (CHI-CV10: model modification based on out-of-sample χ^2 fits) leads to best and almost equal estimates of the POI (ρ). This is likely due to the fact that the cross-validation in Algorithm 3 provides a suitable braking mechanism during the model modification procedure. Using the mean of k OOS fits appears to be effective in preventing sample characteristics from imposing too much influence on modifications added to the baseline model. Influential outliers can be a consequence of sample characteristics and it is beneficial for a researcher to prevent such outliers from influencing the accuracy of their estimates for their parameter of interest. For example, this can aid a researcher in estimating the covariance between social media usage and social media benefit (Abrahim et al., 2019).

There are some limitations to the conclusions which can be drawn from this study. In the simulation study, the baseline model was not very different from the model present in the DGP - the DGP only had one additional free parameter compared to the baseline model. This led to very good performance measures of the non-modifying procedure of fitting the baseline model (No Mod). This was expected due to the great similarity. However, this finding does not help for supporting to make modifications at all in these particular circumstances. Despite that, in reality we do not know what is present in the DGP, and it is likely that some sort of model modification procedure will be adopted by a researcher in practice.

Contrary to the great difference in stability of the parameter of interest estimates, all modelled co-

variance matrices had similar mean distances to the covariance matrices present in the population DGP. All model modification procedures achieved covariance matrices of a very similar distance to the true covariance matrix. This provides evidence that all modelling procedures are similar in their ability to produce a model that fits to the data. There is no strong indication that any method is better than other methods in reproducing the variance-covariance matrix present in the DGP. Cross-validation does therefore not appear to be a promising avenue to pursue if one wants to decrease the distance between the modelled covariance matrix and DGP covariance matrix.

Suggestions for Future Research

To gain an even better insight of the applicability of cross-validation in the SEM model modification procedure, it would be highly worthwhile to conduct a similar simulation study as is presented in this paper. The recommended difference in the new simulation study is that the baseline model should differ more greatly from the true DGP. In addition to the cross-loading, the simulation could be set up such that there are also a number of covariances between observed variables or more cross-loadings present in the DGP. The approach of making no modifications should then lead to poorer performance measures as the baseline model will differ more greatly from the DGP.

A setup where there is a greater mismatch of baseline model and true DGP model would allow for further exploration into whether it is advisable to use cross-validated model modifications instead of the standard model modification procedure. This would create the possibility to investigate which modification procedure performs the best - if there are many modifications which need to be identified, a reliable modification procedure is crucial. Given the results in this simulation study, Algorithm 3 appears to be the most promising method for finding POI estimates which are closest to the true POI value. On the other hand, it could also be the case that algorithm 3 has an approach which is too conservative in adding modifications to a model. This could be a clarification for why it performed so well in the case of this simulation study. To further explore the performance of the CHI-CV methods, testing its performance under more complex circumstances would be a good avenue to pursue.

Furthermore, in addition to the conditions that were varied in this simulation study (α, ρ, δ and n), an investigation for different k's, α levels and minimum MIs in Algorithm 3 is suggested as well. Differing k's, α levels and minimum MIs will likely impact how conservatively the algorithm behaves when freeing additional parameters in a model. By testing different levels, an optimal combination of α and minimum MI parameter values may be found. Hopefully bringing research another step closer to specifying SEM models which are as close as possible to the DGP.

Acknowledgements

This Research Master thesis was made possible by Erik-Jan van Kesteren MSc and Dr. Daniel Oberski of the Methodology & Statistics department at Utrecht University. I would like to sincerely extend my gratitude to both of these people for their supervision, guidance and feedback throughout the process of writing this thesis. Furthermore I would like to thank my family, MSBBSS coursemates and close friends for their encouragement and support throughout the duration of this project.

Bibliography

Abrahim, S., Mir, B. A., Suhara, H., Mohamed, F. A., & Sato, M. (2019). Structural equation modeling and confirmatory factor analysis of social media use and education. *International Journal of Educational Technology in Higher Education*, *16*.

Barrett, P. (2007). Structural equation modelling : Adjudging model fit. *Personality and Individual Differences*, *42*, 815–824. doi: 10.1016/j.paid.2006.09.018

Browne, M., & Cudeck, R. (1989). Single Sample Cross-Validation Indices for Covariance Structures. *Multivariate Behavioral Research*, *24*(4), 37–41. doi: 10.1207/s15327906mbr2404

Browne, M., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, *21*(2), 230–258. doi: 10.1177/0049124192021002005

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304. doi: 10.1177/0049124104268644

Cudeck, R., & Browne, M. W. (1983). Cross-Validation Of Covariance Structures. *Multivariate Behavioral Research*, *18*(2), 147–167. doi: 10.1207/s15327906mbr1802

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural Equation Modelling : Guidelines for Determining Model Fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60.

Hox, J. J. (1999). An Introduction to Structural Equation Modeling. *Family Science Review*, *11*, 354–373.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. New York: Springer.

Khine, M. S. (Ed.). (2013). *Applying SEM in Educational Research*. Rotterdam: Sense Publishers.

Kolenikov, S. (2011). Biases of Parameter Estimates in Misspecified Structural Equation Models. *Sociological Methodology*, *41*, 119–157.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model Modifications in Covariance Structure Analysis; The Problem of Capitalization on Chance. *Psychological bulletin*.

Marcoulides, K. M., & Falk, C. F. (2018). Model Specification Searches in Structural Equation Modeling with R. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 484–491. Retrieved from `https://doi.org/10.1080/10705511.2017.1409074` doi: 10.1080/10705511.2017.1409074

Morin, A. J. S., Marsh, H. W., Craven, R., Hamilton, L., Liem, G. A., Lüdtke, O., ... Parada, R. (2013). Exploratory Structural Equation Modeling. In *Quantitative methods in education and the behavioral sciences: Issues, research, and teaching. structural equation modeling: A second course* (2nd ed., pp. 395–436). Charlotte, NC: Information Age Publishing, Inc.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus User's Guide* (Sixth ed.). Los Angeles, CA: Muthén & Muthén.

Preacher, K. J. (2010). Quantifying Parsimony in Structural Equation Modeling Quantifying Parsimony in Structural Equation Modeling. *Multivariate Behavioral Research*, *41*(3), 227–259. doi: 10.1207/s15327906mbr4103

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from `http://www.jstatsoft.org/v48/i02/`

Schreiber, J. B. (2006). Modeling and Confirmatory Factor Analysis Results : A Review. *The Journal of Educational Research*, *99*(6).

Whittaker, T. A. (2012). Using the Modification Index and Standardized Expected Parameter Change for Model Modification Using the Modification Index and Standardized Expected Parameter Change for Model Modification. *The Journal of Experimental Education*, *80*(1), 26–44. doi: 10.1080/00220973.2010.531299

Yuan, K.-h., Marshall, L. L., Bentler, P. M., Yuan, K.-h., & Bentler, P. M. (2003). Assessing the Effect of Model Misspecifications on Parameter Estimates in Structural Equation Models. *Sociological Methodology*, *33*, 241–265.

Yuan, K.-h., Marshall, L. L., & Weston, R. (2002). Cross-validation by downweighting in influential cases in structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *55*, 125–143.