Automated model specification search in CFA using metaheuristics

Master Thesis
Applied Data Science

Author:
Oscar Kromhof
8170045

Supervisor:
David Goretzko,
Methodology and Statistics

July 7, 2023

Utrecht University

Contents

Abstract

Confirmatory Factor Analysis is an essential tool in psychometrics to indirectly measure abstract psychological constructs. It is therefore important to have a well fitting CFA model on empirical data-sets. It turns out that in practice many theoretical models don't fit the empirical data well which could potentially be resolved by model re-specification. Research on the topic has led to the use of meta-heuristics for this task. Numerous meta-heuristics have been proposed for model specification. From previous research the most promising meta-heuristics seem to be the Particle Swarm Optimisation and the Genetic Algorithm. The final goal is to be able to automate the process of model specification.
In this thesis we investigate which set of parameters in the objective functions (hyperparameters) for each of the named metaheuristics yield the best result for model specification according to certain fit-measures *(CFI, RMSEA, SRMR)* that are specific to Structural Equation Models. From the analysis of the selected data-sets from the OSF-website we conclude that having the model-complexity penalty term, λ, in the objective-function in the range of $[0, 0.2]$ will lead to descent results. For the slope-parameter we find that values beyond a threshold $-b < -100$ give the better values for the fit-measures. The threshold values/centring values c_1, c_2, c_3 seem to have a less significant impact on the resulting fit-measures. None of these conclusions should be interpreted as hard cut-off values but rather suggestions for ***avoiding*** bad performance of the algorithms in the context of CFA. We also gathered evidence that the Genetic Algorithm is consistently performing better than the Particle Swarm Optimisation on the selected data-sets while taking up less computing time, suggesting that the Genetic Algorithm might be preferred over the PSO for model specification in Structural Equation Modelling.

KEYWORDS

metaheuristics, particle swarm optimization, genetic algorithm, structural equation modeling, model specification search, hyperparameter tuning

Preface

This paper was written for the completion of my Master's Degree in Applied Data Science at Utrecht University. I want to thank my supervisor David Goretzko for his excellent guidance and support during this project.

1 Introduction

The main goal of this thesis is to assess different metaheuristic objective-function parameter settings, to improve the *out of sample model fit* of certain CFA models. In this process we hope to obtain consistent results for different empirical data-sets so we can suggest a set of reasonable parameter settings for the Genetic Algorithm and the PSO-algorithm on empirical data-sets. Since the aforementioned text contains quite some technical terminology, we will brake it down in the following sub-paragraphs with a brief explanation of each component.

Confirmatory Factor Analysis (CFA)

Confirmatory Factor Analysis is a statistical analysis method where one tries to measure an unobserved hypothesised latent variable by measuring associated indicator variables. The word "confirmatory" as an adjective is there because the researcher is interested in testing a particular factor model based on a certain theory. It was first introduced by Charles Spearman [1]. He came up with the method in the context of cognitive research on school children. Spearman noticed that the children's test scores on different school-subjects were correlated with each other, so he hypothesised that this was due to a general intelligence factor which influenced all the performances on the different tests. Later on it was discovered that this is just a special case with one factor (the so called g-factor) and that CFA is actually part of a bigger body of statistical models called Structural Equation Models (SEM) which is one of the main psychometric tools in modern psychology. In this thesis we will constrain our scope of analysis only to the more basic CFA cases, although it could be extended to more complex SEM-models.
The CFA models that will be analysed in this thesis can be represented mathematically in matrix-vector form as follows,

$$y = \mathbf{\Lambda}\xi + \epsilon \tag{1}$$

Where y represents the measured indicators and $\mathbf{\Lambda}$ represents an $n \times m$-matrix with n the number of indicators and m the number of latent factors, the matrix elements λ_{ij} represent then the cross-loadings, ξ represents the vector which contains the factor/latent variables and ϵ represents the residual noise/-variance.

It can mathematically be shown, for example as done in [2], that the variance-covariance matrix of the model can be decomposed as,

$$\mathbf{\Sigma} = \mathbf{\Lambda\Phi\Lambda'} + \mathbf{\Theta} \tag{2}$$

Here $\mathbf{\Lambda}$ is defined as in Equation 1 (and $\mathbf{\Lambda'}$ is its transpose), $\mathbf{\Phi}$ represents the $m \times m$ variance-covariance-matrix for the latent factors and finally $\mathbf{\Theta}$ is the variance-covariance matrix of the residuals. In our case, this will just be a diagonal matrix since we assume there is no correlation between the residuals ϵ. So the parameters to be estimated are the factor loading's λ_{ij}, the factor

variance/co-variances ϕ_{ij} and the residual variances θ_i. In this thesis we will use maximum-likelihood as the estimation procedure.

Model Specification and Metaheuristics

Let's first consider the difference between model specification and parameter estimation. Parameter estimation is part of classical statistics where one has a given statistical model and a given data-set and where one tries to optimise the parameters of that statistical model for the given data-set, for example with ML-methods, Bayesian estimation, Least Squares, Method of Moments, etc.
Model specification operates one level higher where one tries to find the optimal *model structure* within a particular class of possible models. In the CFA case described above this means which parameters of the statistical model should be freely estimated (1) and which parameters should be fixed to zero (0). In this way the problem can be seen as a combinatorial optimisation problem where one tries to find the "best" binary vector according to some metric. Of course the naive way of solving this is trying out all the possible combinations of binary vectors and evaluate these solutions according to the specified *performance metric*. These type of procedures will quickly become computationally infeasible, for example if we only consider the factor loadings, λ_{ij}, and define $N = n \times m$ then the possible amount of combinations of parameters that can be turned "off/on" will be given by 2^N.
To tackle this problem local optimisation methods have been used but they run the risk of converging to a local optimum which is discussed by Muroshashi and Toyoda in [3]. This is why they resorted to so called metaheuristics to attempt to find better (but not perfect) solutions to the optimisation problem. In next paragraph meta-heuristics will be explained.
Heuristics, or "rules of thumb", are problem-specific strategies that focus on finding a good solution quickly to that specific problem, while metaheuristics are general-purpose strategies (often based on natural phenomena) that can be used to find good solutions across a broad range of problems. Numerous meta-heuristics have been developed and studied in the setting of model-spefication in CFA for example the Genetic Algorithm (GA) in [3], Hybrid Ant Colony Optimization Algorithm (hACO) [4] or Bee Swarm Optimisation (BSO) [5]. In this thesis the focus will be on the Genetic Algorithm (GA) [6] and the Particle Swarm Optimisation (PSO). Some desirable properties/characteristics of meta-heuristics listed in [7] are:

- They can handle problems with non-linear and non-differentiable objective functions, discrete variables and constraints. In our case we deal with discrete variables, binary vectors, $\mathbf{v} \in \{0,1\}^d$. Where d is the maximal amount of parameters that could be present in the choosen model structure.

- Meta-heuristics aim to explore the search space comprehensively and avoid getting stuck in local optima.

- Minimal Problem-Specific Knowledge: Unlike problem-specific algorithms that require detailed domain knowledge, meta-heuristics operate based on general principles and heuristics. They do not rely heavily on problem-

specific information, making them more widely applicable and easier to implement.

Some more undesirable properties are:

- Lack of Guarantee for Global Optimality: While meta-heuristics aim for global optimisation, they do not provide any guarantee of finding the global optimum. There is always a possibility of settling for sub-optimal solutions, depending on the problem complexity and search space characteristics.
- Parameter Tuning: Most meta-heuristics involve multiple parameters that need to be fine-tuned for optimal performance. Finding the right parameter values can be challenging and time-consuming. Inappropriate parameter settings may lead to poor results or slow convergence. *This will be the main objective of this project.*
- The minimal Problem-Specific knowledge comes at a cost of, Lack of Problem-Specific Exploitation, metaheuristics are general-purpose algorithms that focus on exploration rather than exploiting problemspecific knowledge. In domains where detailed domain-specific knowledge is available, problem-specific algorithms may outperform meta-heuristics.

In the context of Confirmatory Factor Analysis it is important to note that there exists a 1-to-1 mapping from the set of model-structures to a set of binary vectors say $\zeta : M \to \{0,1\}^d$, this will be explained further in Section 2. M represents the set of possible model structures.

Goodness of Fit-Indices and optimisation criteria

We will optimise the model selection procedure by combining the following three fit indices for factor analysis models: SRMR, CFI and RMSEA into one optimisation function and use meta-heuristic algorithms to approximate the optimal solution according to the objective function that which be discussed in the next section.

The objective function: A three-parameter logistic regression model

The fit indices are specifically combined in a three parameter logistic regression function which is explained in more detail in [8]. The three-parameter logistic regression (3PL) model is commonly used in the field of psychometric, specifically in item response theory. The three-parameter logistic regression in our specific case is given by the expression,

$$\frac{(1-\lambda)\left(\left(1-\frac{1}{1+\exp(-b(c_1-\text{SRMR}))}\right)+\frac{1}{1+\exp(-b(c_2-\text{CFI}))}\right)}{3} + \frac{(1-\lambda)\left(1-\frac{1}{1+\exp(-b(c_3-\text{RMSEA}))}\right)}{3}+\lambda\left(\frac{n_{m.params}}{n_{t.params}}\right)$$

where $n_{m.params}$ and $n_{t.params}$ are the models estimated parameters within the search space and the search spaces' total potentially estimated parameters respectively. With the parameter b the slope is defined and thus the range of sensitivity and c_i is used for centring around a specific value as explained in [9]. $\lambda \in [0, 1]$ represents the penalty for model complexity i.e. higher values for λ gives a higher penalty for selecting higher model complexity. The meaning of the terms SRMR, CFI, RMSEA will be explained in the next sub-section. The default set of parameter values for the function during this project are listed in Table 1. If it is not mentioned otherwise these will be the parameter values.

λ	b	c_1	c_2	c_3
0.5	-55	0.08	0.95	0.06

Table 1: *Default parameter values for the objective function.*

This function will be used in the meta-heuristics to determine the fitness of a certain solution, which will be further discussed in the Section 2.

Descriptive Measures of Overall Model Fit in Structual Equation Modelling

These criteria are based on the difference between the empirical sample covariance matrix $\mathbf{S}$ (which represent the correlations that are observed among the different indices) and the model implied covariance matrix $\mathbf{\Sigma}(\hat{\theta})$ where $\hat{\theta}$ represents the vector of parameter estimates as discussed in [10]. In this thesis the two metrics that are considered are the RMSEA and the SRMR. Rules of thumb to assess the model fit with these metrics are given by:

$$\text{Good fit: } 0 \leq SRMR \leq 0.05 \text{ — Reasonable fit: } 0.05 \leq SRMR \leq 0.1$$

$$\text{Good fit: } 0 \leq RMSEA \leq 0.05 \text{ — Reasonable fit: } 0.05 \leq RMSEA \leq 0.08$$

Descriptive Measures Based on Model Comparisons

The basic idea of comparison indices is that the fit of a model of interest is compared to the fit of some baseline model according to [10]. Even though any model nested hierarchically under the target model (the model of interest) may serve as a comparison model, the independence model is used most often. The independence model assumes that the observed variables are measured without error, i.e., all error variances are fixed to zero and all factor loadings are fixed to one, and that all variables are uncorrelated. This baseline model is a very restrictive model in which only the variances, (θ_i), of the variables, have to be estimated. For the range of the CFI we have $0 \leq CFI \leq 1$.

$$\text{Good fit: } 0.97 \leq CFI \leq 1.00 \text{ — Reasonable fit: } 0.95 \leq CFI \leq 0.97$$

and hence any values below 0.95 are considered poor model fit.

Bayesian Information Criteria (BIC)

In order to compare performance between different metaheuristics we're going to make use of the Bayesian Information Criteria given by,

$$\text{BIC} = k\ln(n) - 2\ln(\hat{L}), \tag{3}$$

where $\hat{L} = p(x|\hat{\theta}, M)$ is the maximised value of the likelihood function of the particular CFA model. $\hat{\theta}$ is the estimated vector of parameters which maximises the likelihood function given the data x, n represents the sample size and k represents the number of parameters in the model.
The BIC introduces a penalty term for over-fitting the model to the data, models that generate a lower BIC are preferred over models that result in a higher BIC as stated in [11]. The BIC will be used for the comparison of performance between the Genetic Algorithm and the Particle Swarm Optimisation in Section 5.3.

Overview and Research question

As mentioned earlier the main goal of this thesis is to develop an idea of the functional hyperparameters that are best suited for the respective meta-heuristics/optimisation algorithms. During the project we will only consider the meta-heuristics Particle Swarm Optimisation (PSO) and the Genetic Algrotihm (GA).

2 The Genetic Algorithm and the Particle Swarm Optimisation

In this section a brief explanation of the two meta-heuristics, Genetic Algorithm and the PSO, will be given. The focus will be more on the general principles of the meta-heuristics rather than a detailed explanation of all the technicalities on how these algorithms should be implemented in the context of model specification in SEM.

Genetic-Algorithm (GA)

In this paragraph the optimisation procedure of the Genetic Algorithm is explained mainly based on [6]. The Genetic Algorithm is a special case of a broader class of meta-heuristics named Evolutionary Algorithms. It is inspired by natural selection in populations. The algorithm is initialised with a starting population, which is generated by randomly mutating the initial model structure, in this project this will be a set of binary vectors which represent the different model structures. This starting population will be evaluated on it's fitness and then by some selection mechanism, where the individuals with higher fitness will have a higher probability of survival. These selected individuals will generate the next generation of individuals by cross-over and mutation. This process

then continuous till how many generations in the algorithm are pre-specified, after which the best fitted solution in the final population is returned as a result of the process. The mutation rate and the cross-over rate can all be pre-specified. The non-objective function hyperparameters for this algorithm are: n_{pop}, $n_{parents}$, n_{mut}, n_{gen}. They represent respectively: the amount of individuals in the starting population, the amount of individuals or solutions selected from the population to be used for reproduction in the creation of the next generation, the mutation rate for the newly created population and the total amount of generations before the algorithm terminates.

Particle Swarm Optimisation (PSO)

The Particle Swarm Optimisation is also a population-based algorithm only now the population should not be imagined as a static set of individuals but a dynamic population that respond to their own (local) situation as well as to each others' behaviour (global). The original inventors of the algorithm, Kennedy & Eberhart [12], were also inspired by bird flocking and fish schooling.
Based on an initial given model-structure the initial population is generated. Here each individual particle in the population represents a potential solution from which the following metrics are computed:

- The particle with the overall best position, best_{ind}. This is the position that corresponds to the position where the particle assumed its best fitting value in its trajectory during the optimisation procedure.
- The overall best position of the swarm, best_{pop}, i.e. of the population of assumed solutions.
- The particles best current position current_{ind}. The current fitness refers to the fitness value of a particle at the current iteration or time step of the algorithm. It represents the quality of the particle's current position in the search space.

These calculations are done after every genertion/iteration of the PSO at initiation we have The final ingredient is introducing some random variation in the velocity update such that the search space is sufficiently explored. If we combine all this information we get the following expression for the velocity update for an individual particle i,

$$V_{i+1} = V_i + \Phi_1(\text{best}_{ind} - \text{current}_{ind}) + \Phi_2(\text{best}_{pop} - \text{current}_{ind})$$

where Φ_i is uniformly distributed random variables on the interval $[0, 1]$, i.e. $\Phi_1, \Phi_2 \sim U([0, 1])$. In this manner there is a general tendency for the particle population to move towards an optimal solution in the search space. There is also an extra parameter added to the algorithm which puts a maximum value on the speed the particles can attain.
The non-objective function hyperparameters for this algorithm are: n_{pop}, v_{max},

n_{gen} the size of the population, the maximum velocity of the particles and the number of generations respectively, n_{gen} puts a maximum on the amount of iteration the PSO performs.

3 Data description

In this section we will discuss the empirical data-sets that will be utilised to perform the analysis of the hyperparameter settings and test the Genetic Algorithm and the Particle Swarm Optimisation.

3.1 Short Dark Triad Scale

This data-set concerns the dark triad model which is a psychological construct of personality types [13]. The data consists of questionnaires where participants had to respond on a scale from 1-6. There were $n = 1100$ row/participants in the data-set. After the deletion of individuals containing NA values $n = 802$ participants remained. This operation could lead to bias in the data if values were not missing completely at random this issue will be later addressed in the discussion section. There were also some measurement that were inversely scaled which had to be reversed again for the data analysis. For further information and access to the data one could visit the OSF-website by following the link `https://osf.io/g2m4z`. For our analysis we partitioned the data in an 80% train set and a 20% validation set by selecting rows from the data-set at random.

3.2 International Personality Item Pool (IPIP300-data): Five-factor model

The International Personality Item Pool (IPIP300) data-set is a collection of items for use in personality tests, originally created by Lewis Goldberg [14]. One could access the data via the OSF-website by the following link `https://osf.io/wxvth/?view_only=`. In this project the file IPIP300.por was used, which consists of test data of 3071313 individuals witch each 310 recorded variables. The first 10 variables denote demographic information and the other variables denote the item scores on the scale 1-5. Some data entries have the value 0 which indicates a missing values, individuals with missing values were deleted from the data-set. After this operation only 125102 individuals remain in the data-set, so only 40% is left from the original size. This could potentially introduce bias in the remaining data due to the fact that the data might not be missing at random. We also used other personality traits that are present in the IPIP data-set, these will have exactly the same structure as the extra-version one.

3.3 Ethical considerations

Since the used data-sets are from large scientific studies that are available in the public domain (via the OSF-website) where no sensitive variables were recorded and participants are anonymized there won't be any ethical issues, i.e. privacy considerations, bias towards certain ethnic groups, etc. during this project.

4 Methods

In this project we're going to make use of the empirical data-sets discussed in the previous section. Also the necessary transformations are executed on the data sets before they are ready for use. The initial model structure based on theory are given by Figure 1 and 2: the Dark Triad and the extra-version personality construct respectively, where we assume there are no correlations among the residuals. Other personality constructs from the IPIP data-set will also be used for analysis although not all results will be included in the thesis due to an overflow of figures that will be generated from that.

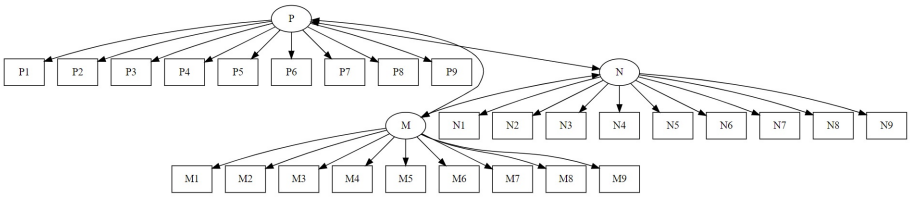


Figure 1: *Visualisation of the hypothesised CFA-model for the dark triad personality construct. This model will be used as the initialisation for the PSO and the GA algorithm on the Dark Triad data. Mechiavellianism ="M", Narcissism ="N", Psychopathy ="P".*

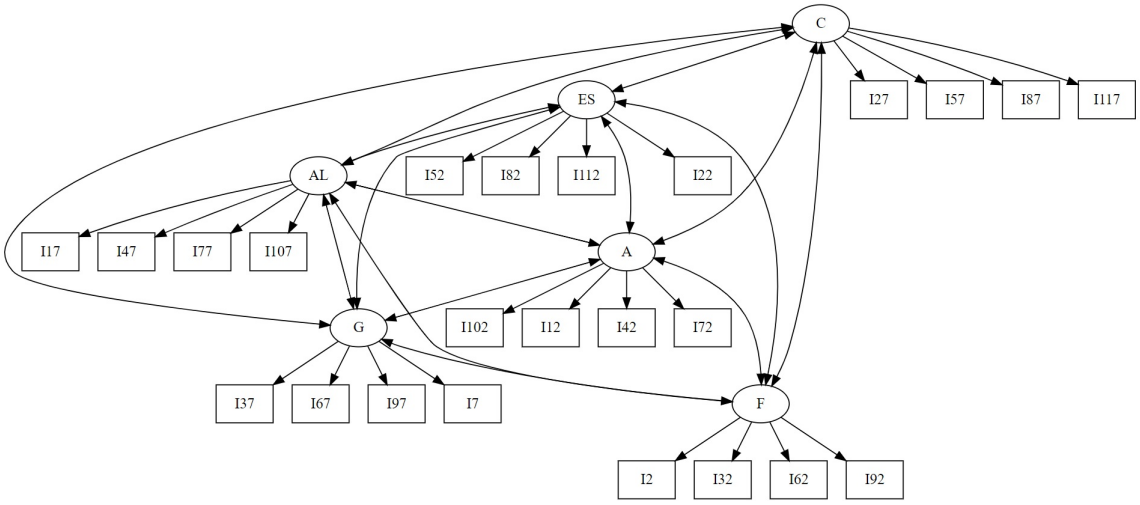


Figure 2: *Visualisation of the initial CFA model for the Extraversion personality scale. This model will be used as the initialisation for the PSO and the GA algorithm on the IPIP300 data. The latent variables are coded as: Friendliness ="F", Gregariousness = "G", Assertiveness = "A", Activity Level = "AL", Excitement Seeking = "ES", Cheerfulness ="C".*

For the data-analysis the statistical software R is used in concert with the following packages: lavaan [15], lavaanPlot [16], dplyr [17], tidyr [18], ggplot2 [19]. We also make use of the R-functions from the Gitlab repository `https://gitlab.com/KarikSiemund/specification-search-via-combinatorial-optimzation` to run the Particle Swarm Optimisation and Genetic Algorithm in the Structural Equation Modelling context. To gain access to the repository one should contact one of the moderators. The scripts that were created for the data analysis and the creation of the figures can be found on `https://gitlab.com/oscarkromhof/ads-master-thesis-script`. To request access to the scripts one could send an email to `oscarkromhof@hotmail.com`. The seeds for the pseudo-random-number generator that are going to be used in the scripts to generate the plots in the results sections are either 123, 1223443 or 38201.

To obtain results on how changing the objective function parameters influences the performance of GA and PSO we're going to marginally vary the parameters of the objective function which should be sufficient to draw conclusions about which parameters have significant impact on the *out of sample fit-measures* and how the variation influences the results generated by the meta-heuristics. We will also re-sample our training and validation set to see if any patterns that emerge when the parameters are marginally varied are caused by

training set specific features or show a genuine indication for a good parameter value.
Finally we will investigate the bi-variate variation of the λ parameter and the slope $-b$ to see if the results are consistent with the marginal variation. We might be able to identify a region of reasonable values for the objective function hyperparameters.

5 Results and Analysis

5.1 Marginal variation of the λ parameter

Results Genetic-Algorithm (GA)

In Figure 3 the development of the fit-indices, CFI, RMSEA and SRMR, as the λ parameter varies in the objective function, is visualised. The model specified by the Genetic-Algorithm is evaluated on the 20% hold-out data. Other parameter values of the Algorithm were set to $n_{pop} = 30$, $n_{parents} = 8$, $n_{mut} = 6$, $n_{gen} = 50$, and for the objective function they were set as in Table 1. λ is varied from 0.1 to 0.9 with increments of 0.2. Also the case $\lambda = 0$ was evaluated separately which gave a CFI of 0.84 (rounded to two decimals). The black dotted line in Figure 3 represents the CFI for the initial model, as displayed in Figure 1 on the out of sample data.

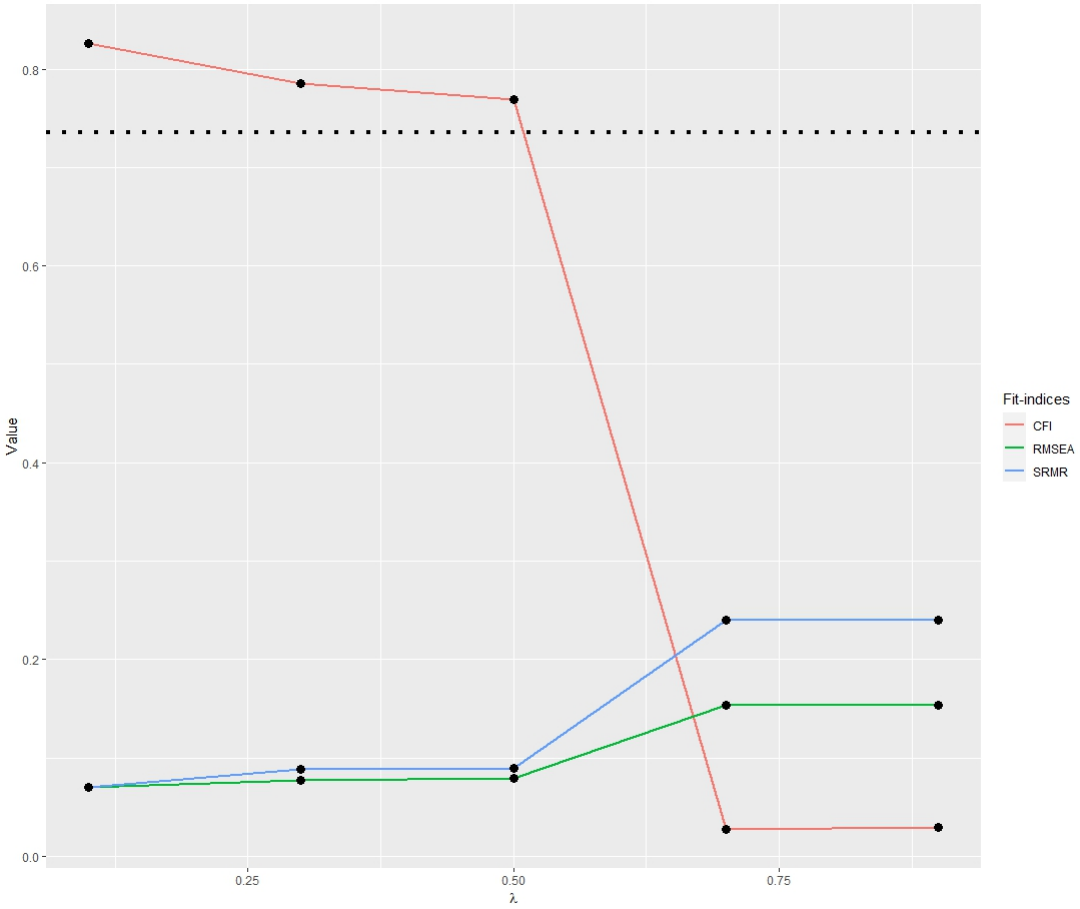


Figure 3: *Fit-indices for the Genetic Algorithm on the* ***Dark Triad validation data*** *for different values of* λ *in the objective-function. Other parameter settings were set to* $n_{pop} = 30$*,* $n_{parents} = 8$*,* $n_{mut} = 6$*,* $n_{gen} = 50$*. The black dotted line represents the* ***CFI*** *of the initial model on the out of sample data.*

The GA algorithm was also performed multiple times on different re-samples of test and validation data (k-fold cross-validation), and for each iteration we find the same general structure that around $\lambda = 0.5$ there is a steep decline in all the fit measures. The specific fit-measure values vary a bit from sample to sample but the general structure is the same. For the rest of this section the procedure is the same as described above only with different changes in the data-set and meta-heuristic used.

In Figure 4 we apply the same idea to the IPIP-extraversion data, again λ is varied and the black dotted line represents the CFI of the initial model on the out of sample data. The parameters that are not part of the objective function are given in the caption of Figure 4. Note that the amount of generations is changed from $n_{gen} = 50$ to $n_{gen} = 5$, this was done to reduce the computation time where we checked that the additional amount of generations didn't add much value to the improvement of the fit-measures.

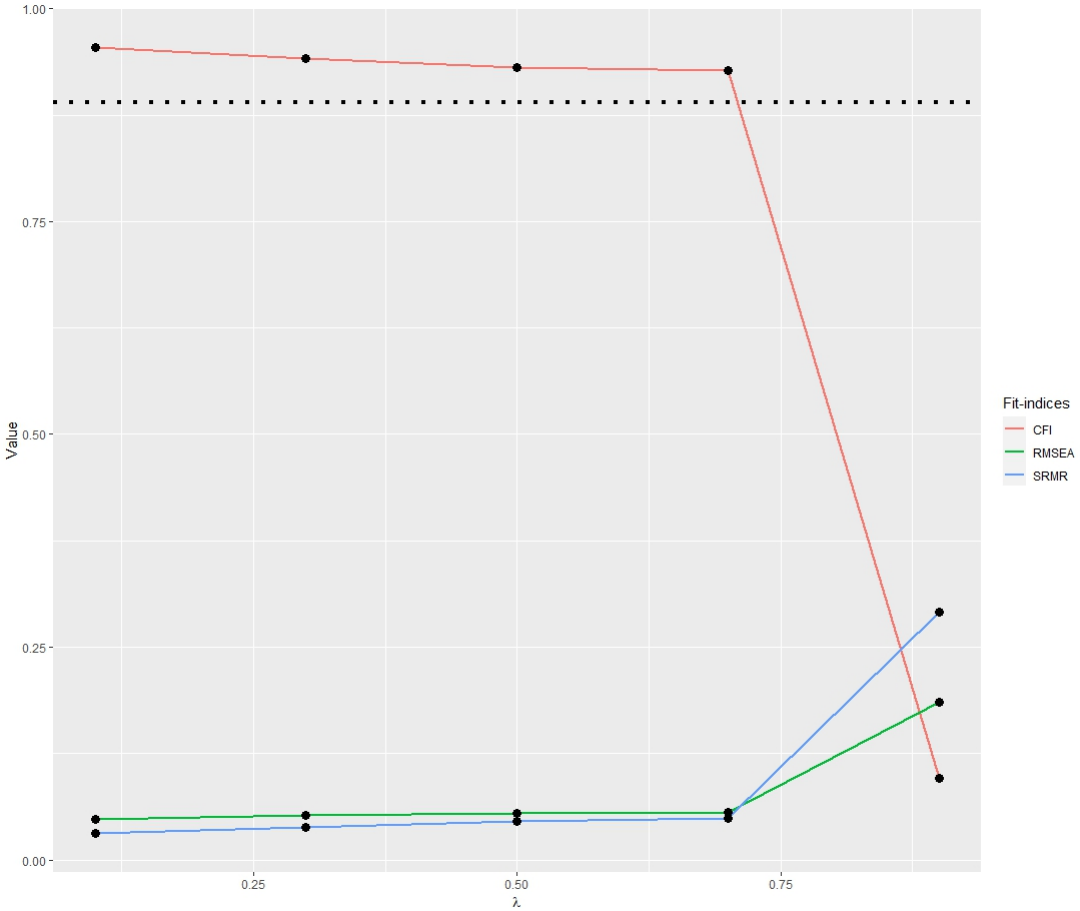


Figure 4: *Fit-indices for the Genetic Algorithm on the* ***IPIP extra-version personality*** *validation data for different values of λ in the objective-function. Other parameter settings were set to $n_{pop} = 30$, $n_{parents} = 8$, $n_{mut} = 6$, $n_{gen} = 5$. The black dotted line represents the* ***CFI*** *of the initial model on the out of sample data.*

The results are analogously constructed as for the Dark Triad data. The main difference is the point λ where the quality of the fit-measures changes, and the fact that better fit measures are reached.

Results Particle Swarm Optimisation (PSO).

The development of the fit-measures by changing λ generated with the Particle Swarm Optimisation is displayed in Figure 5.

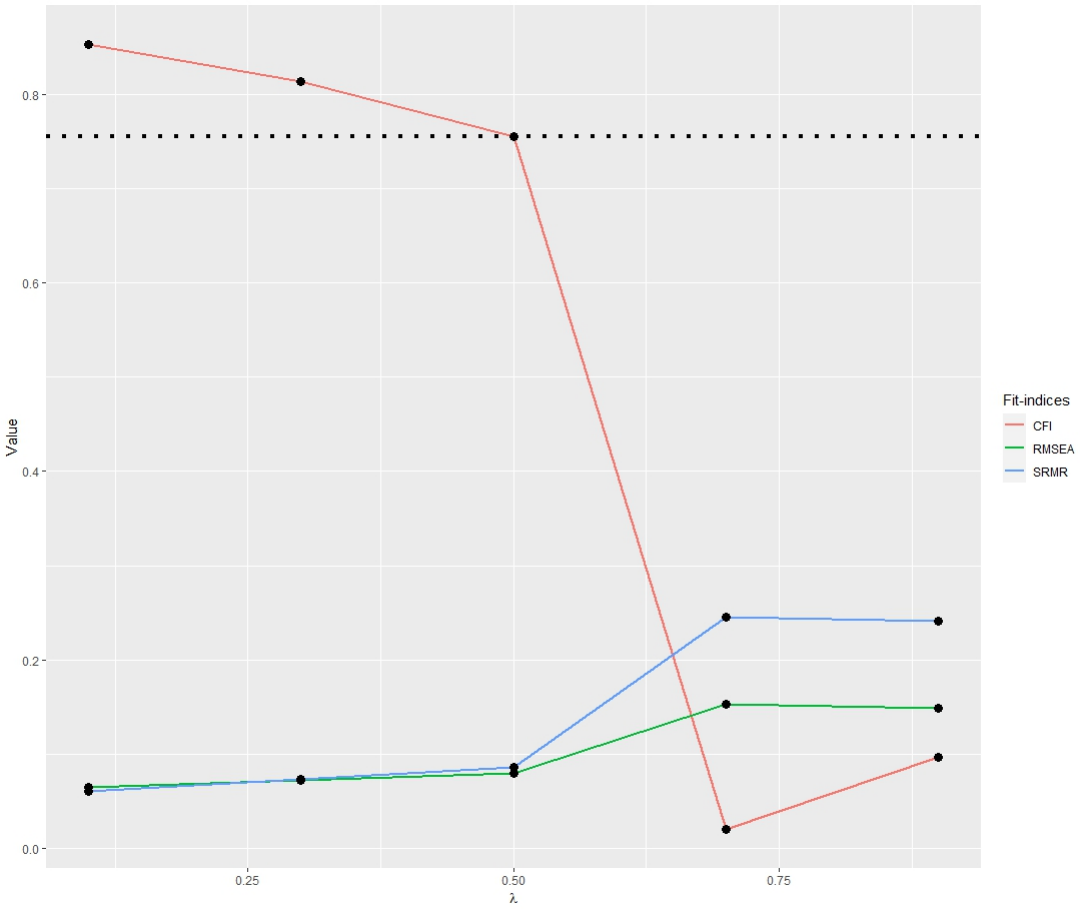


Figure 5: *Fit-indices generated by the Particle Swarm Optimisation (PSO) on the* ***Dark Triad*** *validation data for different values of* λ *in the objective-function. Other parameter settings were set to* $n_{pop} = 30$, $v_{max} = 6$, $n_{gen} = 50$. *The black dotted line represents the* ***CFI*** *of the initial model on the out of sample data.*

In Figure 6 the development of the fit-indices for the Extraversion IPIP-data on the validation data-set as λ varies is displayed.

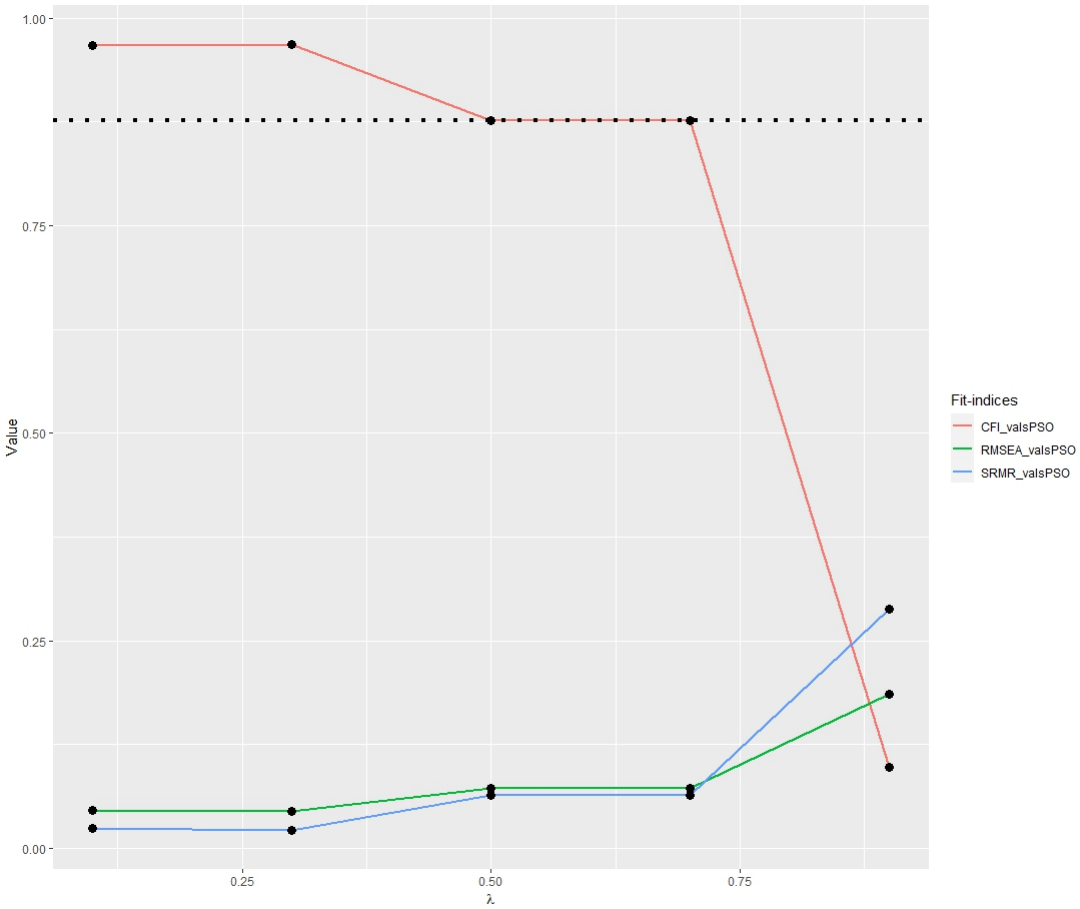


Figure 6: *Fit-indices for the Particle Swarm Optimisation on the* ***IPIP Extraversion data validation*** *for different values of λ in the objective-function. Other parameter settings were set to $n_{pop} = 30$, $v_{max} = 6$, $n_{parents} = 8$, $n_{gen} = 5$. The black dotted line represents the* ***CFI*** *of the initial model on the out of sample data.*

Other personality constructs from the IPIP personality data: Neuroticism, Openness To Experience, Conscientiousness, Agreeableness were also evaluated and yielded similar results.

5.2 Marginal variation of the slope $-b$

In Figure 7 the development of the fit indices is plotted against the slope parameter $-b$. The procedure for generating the plots equivalent to that of the generation of the plots where we vary λ.

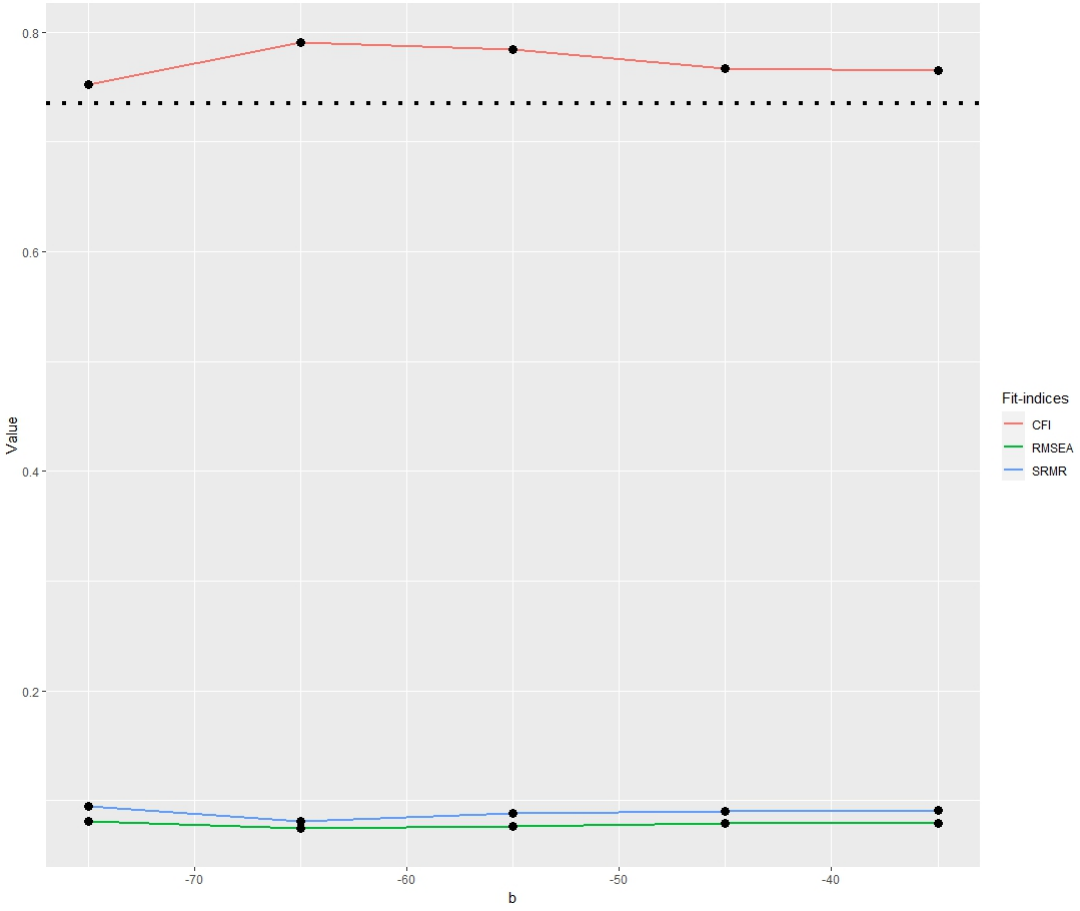


Figure 7: *Fit-indices for the Genetic Algorithm (GA) on the* ***Dark Triad*** *validation data for different values of b in the objective-function. Other parameter settings were set to* $n_{pop} = 30$, $n_{mut} = 6$, $n_{gen} = 5$. *The black dotted line represents the* ***CFI*** *of the initial model on the out of sample data.*

In Figure 8 the development of the fit indices is plotted against the slope parameter of the objective function $-b$. The construction of Figure 8 is analogous to the previous subsection on λ.

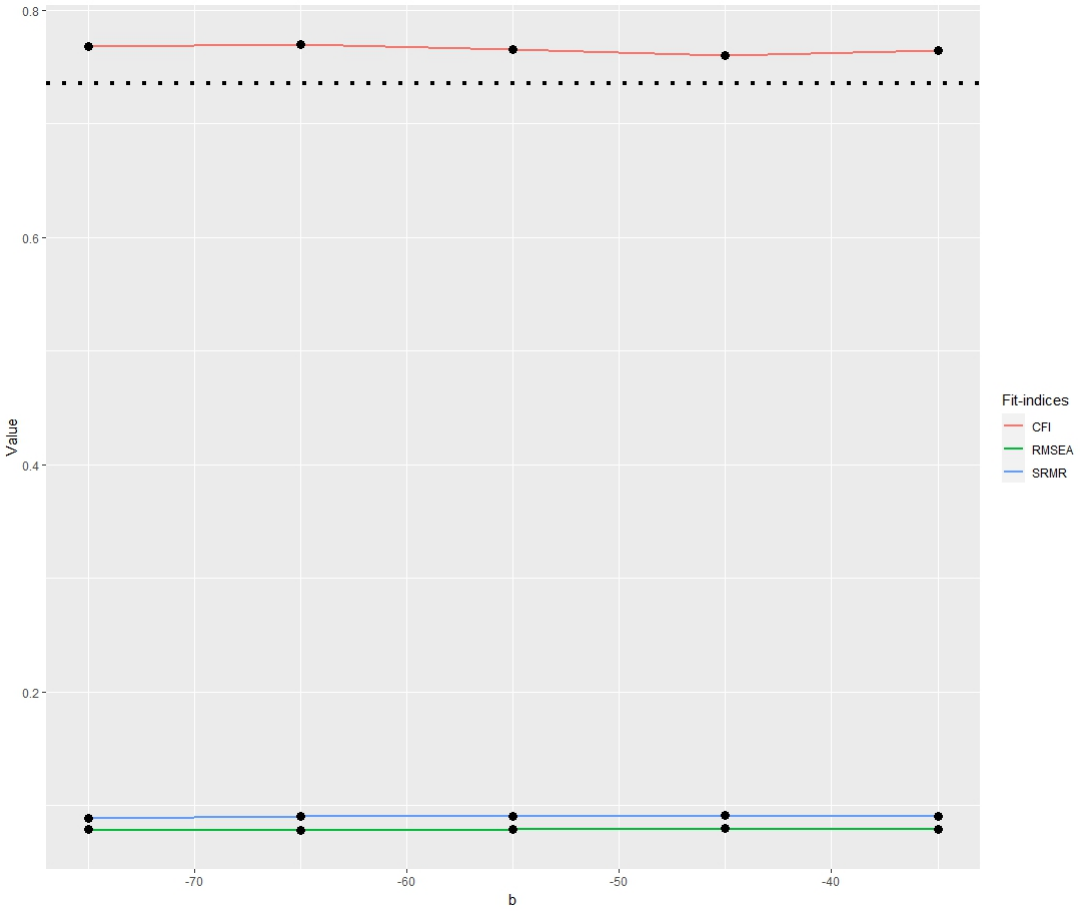


Figure 8: *Fit-indices for the Particle Swarm Optimisation (PSO) on the* ***Dark Triad*** *validation data for different values of λ in the objective-function. Other parameter settings were set to $n_{pop} = 30$, $v_{max} = 6$, $n_{gen} = 5$, $n_{parents} = 8$.*

In Figure 9 the scale on the x-axis is extended from -75 to -35 to -400 to 0 to investigate if more negative values would have an impact on the fit-measures.

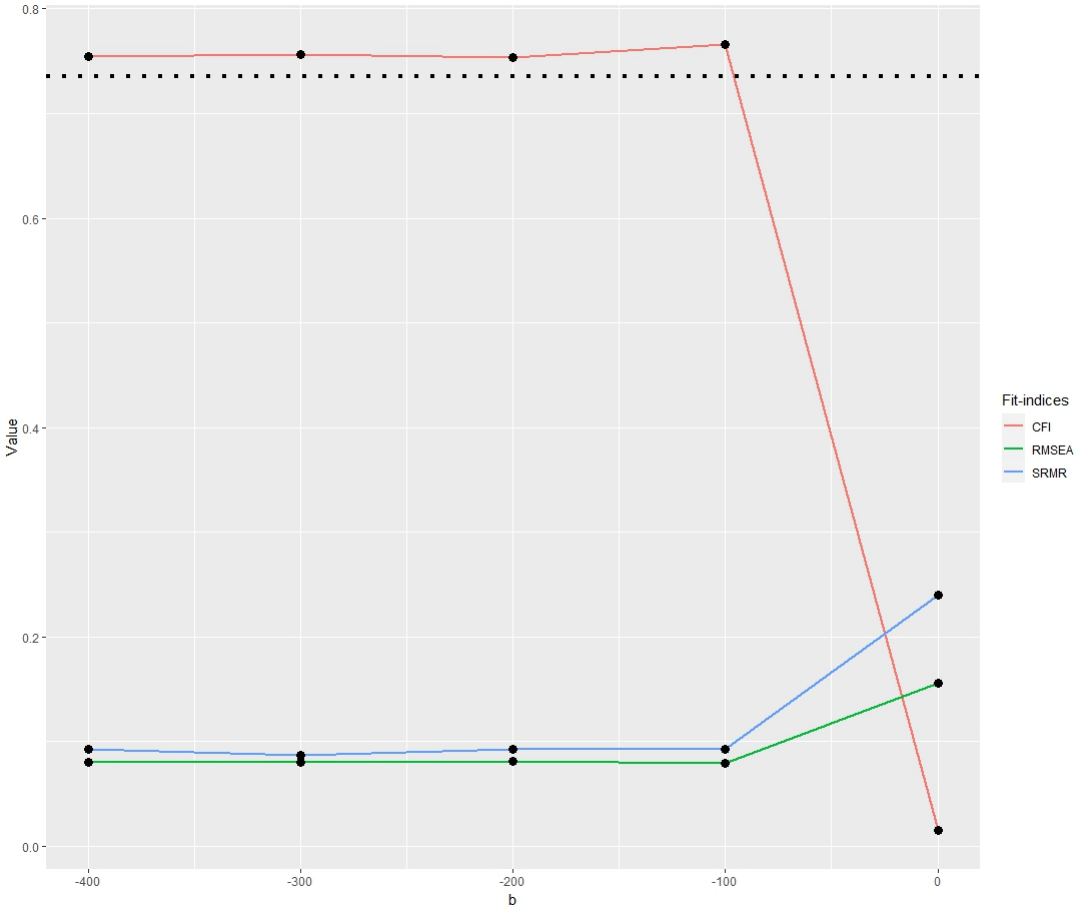


Figure 9: *Fit-indices for the Particle Swarm Algorithm (PSO) on the* ***Dark Triad*** *validation data for different values of* λ *in the objective-function. Other parameter settings were set to* $n_{pop} = 30$, $v_{max} = 6$, $n_{mut} = 6$, $n_{gen} = 5$, $n_{parents} = 8$.

The IPIP data was also tested for which similar constant plots for the fit-measures as a function $-b$ were observed.

5.3 Marginal variation of the threshold/centring parameters c_1, c_2, c_3.

In this section the results for the marginal variation of c_1, c_2, c_3 are displayed. In Figure 10 the CFI of the marginal variation of c_1 is displayed, where the black dotted line represents the CFI value on the out of sample data of the initial theoretical model.

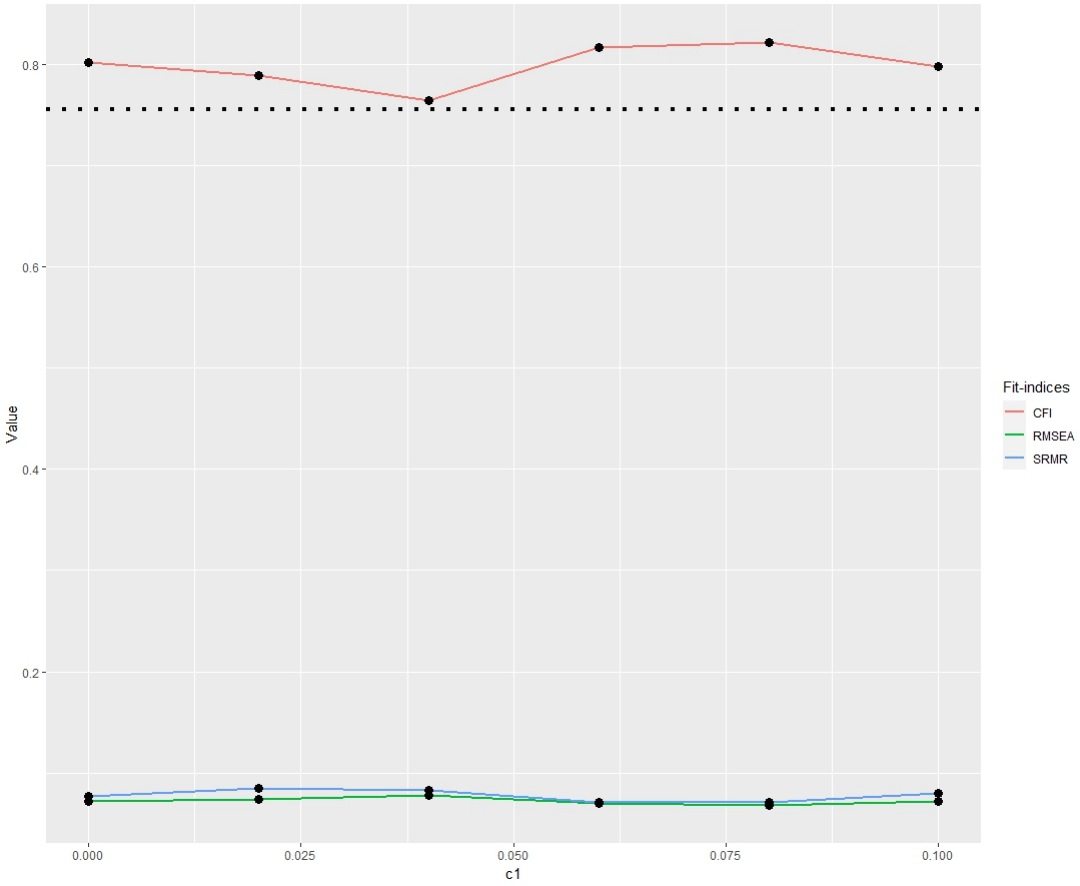


Figure 10: *Evolution of the fit indices for the Genetic Algorithm as the value of* c_1 *varies on the out of sample* ***Dark Triad*** *validation data-set.* $n_{pop} = 30$, $n_{mut} = 6$, $n_{gen} = 5$. *The black dotted line represents the CFI value on the out of sample data of the initial theoretical model*

In Figure 11 the fit-measures of the marginal variation of c_2 are displayed where the procedure is exactly analogous to the previous sub-sections.

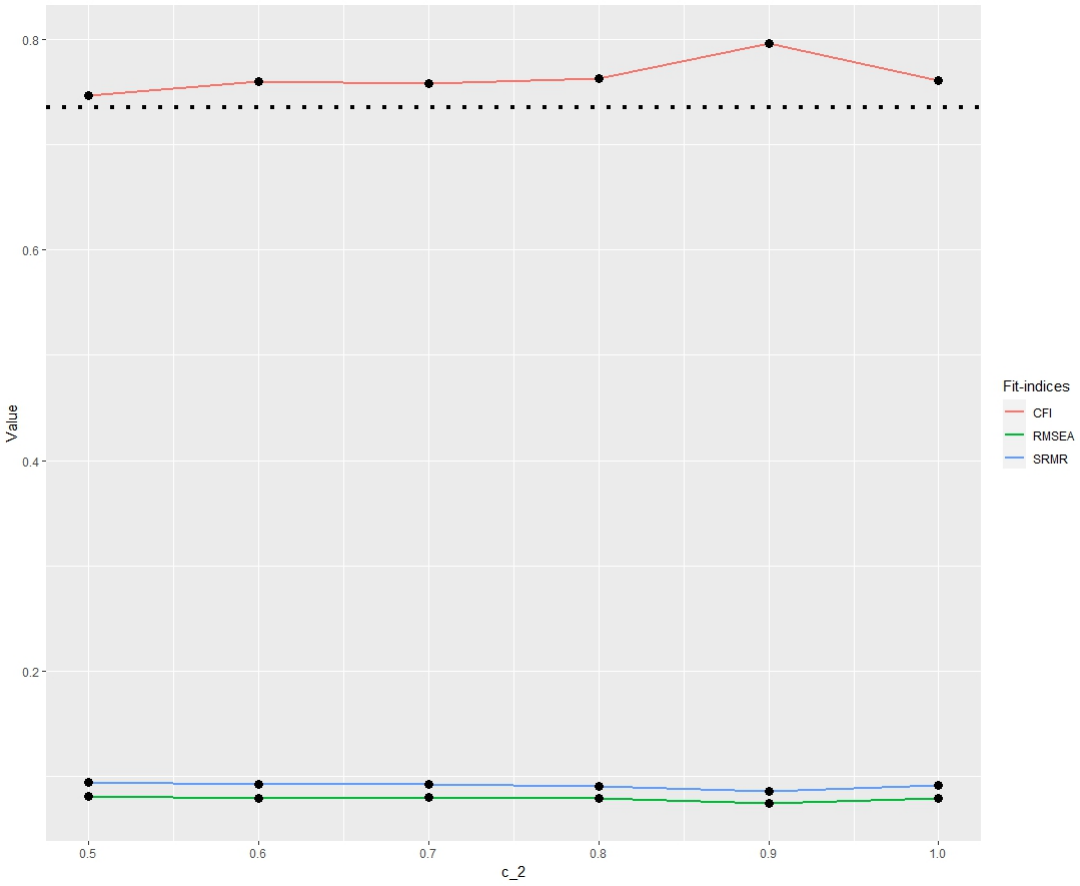


Figure 11: *Evolution of the fit indices generated by the* ***Genetic Algorithm*** *as the value of* c_2 *varies on the out of sample* ***Dark Triad*** *validation data-set.*$n_{pop} = 30$, $n_{mut} = 6$, $n_{gen} = 50$. *The black dotted line represents the CFI value on the out of sample data of the initial theoretical model*

Figure 12 displays the CFI of the marginal variation of c_3 on the dark triad data. The procedure is exactly analogous as the previous sub-sections.

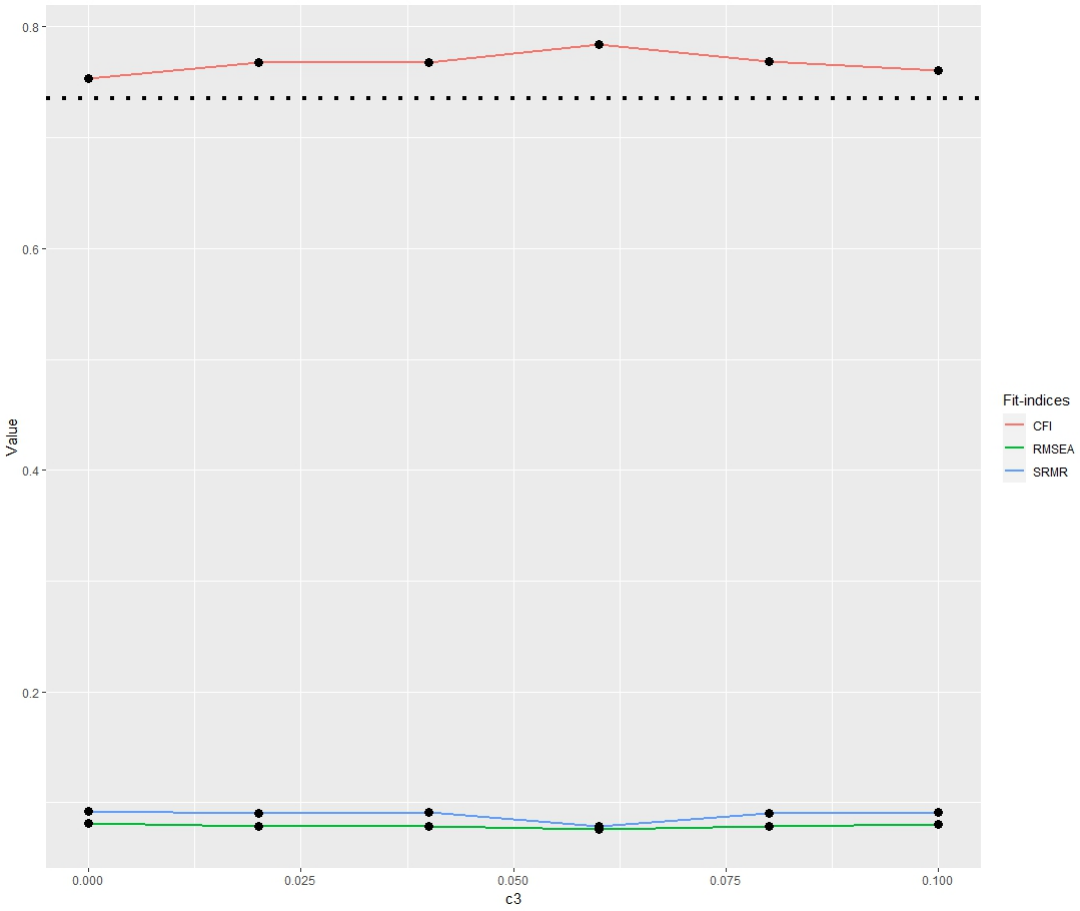
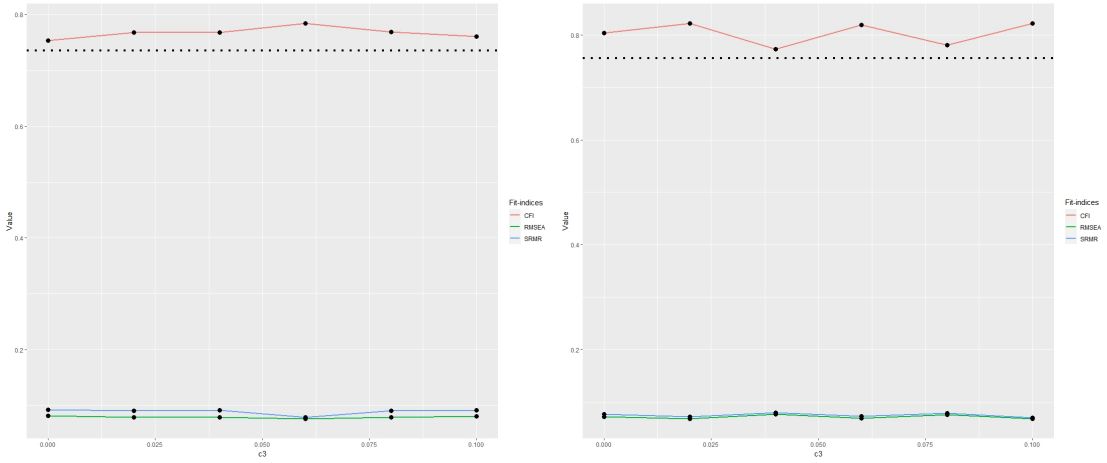


Figure 12: *Evolution of the fit indices for the Genetic Algorithm as the value of c_3 varies on the out of sample* ***Dark Triad*** *validation data-set $n_{pop} = 30$, $n_{mut} = 6$, $n_{gen} = 50$. The black dotted line represents the CFI value on the out of sample data of the initial theoretical model.*

Figure 13 compares the development of different validation and training sets.



(a) (b)

Figure 13: *Comparison of re-sampling the test and validation set by changing the seed. The development of the fit-measures is depicted as the c_3 parameter in the objective function varies for different test and validation sets. The black dotted line represents the CFI of the out of sample fit on the* ***Dark Triad*** *data-set. The Genetic Algorithm was used for the generation of this plot.*

The variation of c_1, c_2 and c_3 was also performed on the IPIP-data. This yields similar results in the sense that variation of these parameters did not result in significant changes in the values of the fit-indices. The only difference being that n_{gen} was changed from 50 to 5 due to computational constraints.

The simultaneous variation of λ and $-b$.

In Figure 14 the filled contour plot of the CFI is depicted generated by the GA where on the $x-$axis the slope, $-b$, is varied and on the y-axis the λ value is varied.

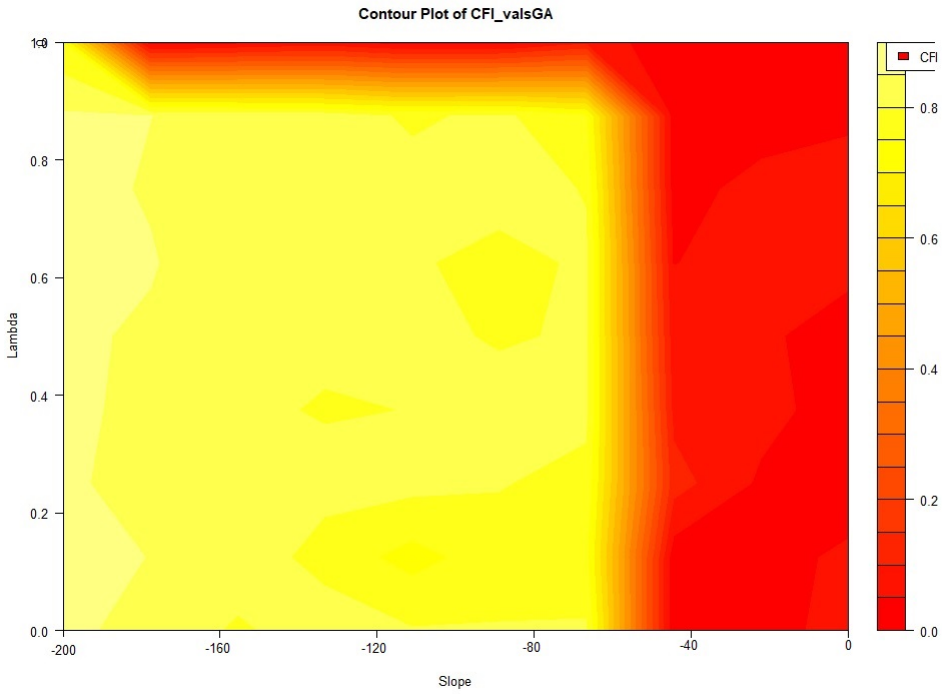


Figure 14: *Filled contour plot of the CFI, generated by the GA. On the x-axis the slope, $-b$ varies between -200 and 0 and on the y-axis λ varies between 0 and 1. Lighter colours yellow colours correspond to better CFI values where darker red colours correspond to worse CFI values. The CFI was calculated on the* out of sample data *of the* ***Dark Triad*** *model.*

In Figure 15 the filled contour plot of the CFI generated by the **PSO** is depicted where on the $x-$axis the slope, $-b$, is varied and on the y-axis the λ value is varied.

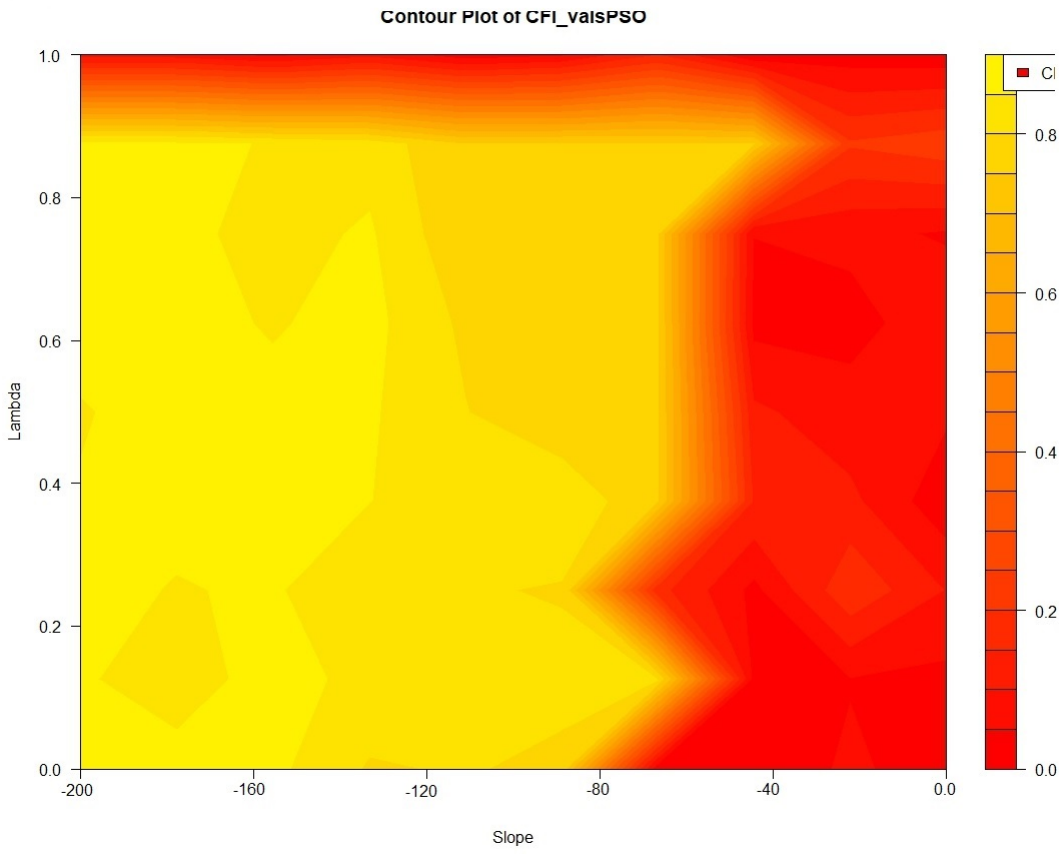


Figure 15: *Filled contour plot of the CFI, generated by the GA. On the x-axis the slope , $-b$ varies between -200 and 0 and on the y-axis λ varies between 0 and 1. Lighter colours yellow colours correspond to better CFI values where darker red colours correspond to worse CFI values. The CFI was calculated on the out of sample data of the* ***Dark Triad*** *data.*

The generation of the plots for the IPIP-data took too much computation time to complete, so they were aborted and are not included in the results of this thesis.

Comparison PSO and GA for different λ values

In Figure 16 and 17 the comparison between the PSO and the GA are displayed in terms of the Bayesian Information Criteria as computed by Equation 3.

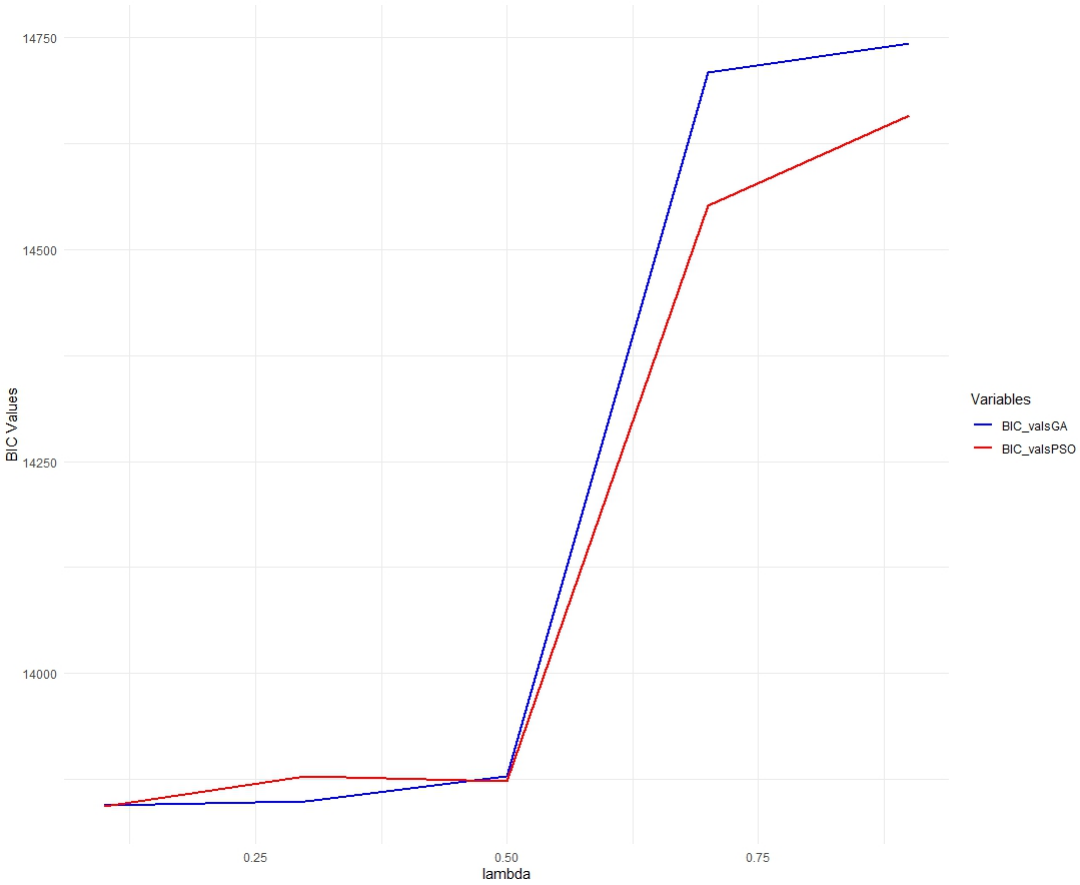


Figure 16: *Bayesian Information Criteria (BIC) values as a function of* λ *for the Particle Swarm Optimisation (PSO) and the Genetic Algorithm (GA) on the* ***Dark Triad*** *validation data. Other parameter settings were set to* $n_{pop} = 30$, $n_{parents} = 8$, $n_{mut} = 6$, $n_{gen} = 50$ *for GA and to* $n_{pop} = 30$, $v_{max} = 6$, $n_{mut} = 6$, $n_{gen} = 50$ *for the PSO.*

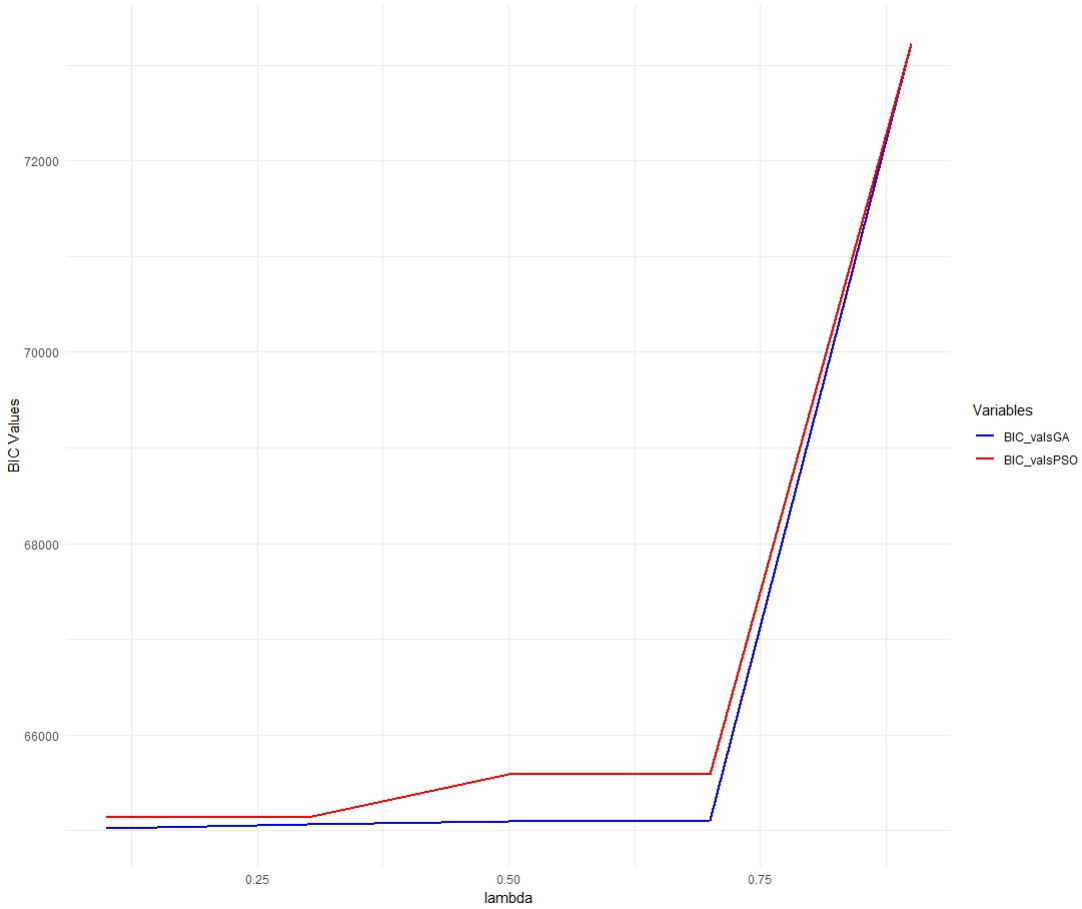


Figure 17: *Fit-indices for the Particle Swarm Optimisation (PSO) on the* ***Extraversion*** *validation data for different values of* λ *in the objective-function. Other parameter settings were set to* $n_{pop} = 30$, $v_{max} = 6$, $n_{mut} = 6$, $n_{gen} = 5$ *and* $n_{pop} = 30$, $n_{mut} = 6$, $n_{gen} = 5$.

Variation of $-b$, c_1, c_2, c_3 to compute the BIC yielded similar results in the sense that their behaviour is similar as for the marginal variation that was conducted in this section. The Genetic Algorithm yielding small but consistently better results than for the Particle Swarm Optimisation. Note that the non-objective function hyperparameters are the same i.e. $n_{pop} = 30$ and $n_{gen} = 50$. The comparison is also made on the other personality constructs present in the personality data, Neuroticism, Openness To Experience and Conscientiousness. This gave similar results.

Analysis

Analysis of marginally varying λ

When we analyse Figure 3, 4, 5 and 6, we first note that in every case the CFI value is monotonically decreasing as a function of λ. There seems to be a data-set dependent value λ_i where we go from a domain of reasonable values for the fit-measures followed by a steep drop to bad values for the fit-measures.

Analysis of marginally varying the slope $-b$

In Figure 7 and Figure 8 we see that varying the value of the slope has very little effect on the fitness of the model on the validation data. Furthermore the little variation that is visible for different values of b is due to the fact that different test and train samples were used to generate the different plots hence this difference is only due to specific patterns in the used training-data and gives no indication that one value is preferred over another to enhance out of sample fit performance.

Analysis of marginally varying the parameters c_1, c_2, c_3

In Figure 10, 11 and 12 we only notice small differences in the fit-measures as the c_i values vary. In Figure 13 the comparison between different training and validation sets is shown. We see some small changes in the optimal fit value which are likely due to training set specific noise. From the Figures it is not clear that there is an optimal value, or a range of values, for the hyper-parameters c_1, c_2 and c_3.

Analysis of the simultaneous variation of λ and $-b$

In Figure 14 we see the same pattern as for the individual variation of λ and $-b$. We see that there is a point when the CFI sharply drops around $b = 55$ till $b = 60$. We see the same pattern in Figure 15 (PSO), only now the boundary is not a straight line anymore but rather depends on the value of λ. We also observe that there is a steep decline in the CFI-values around $\lambda = 0.9$ for all values of b. In the simultaneous variation of λ and $-b$ we did observe the same patterns as in the marginal variation of λ and $-b$ in the sense that there are regions where there is a steep decline in the CFI value, the main difference being that these areas of steep decline are both shifted in Figure 14 and 15 compared to Figure 7 and 8. There is another inconsistent observation between the marginal variation and the simultaneous variation of the parameters λ and $-b$, for the marginal variation we choose $-b = -50$ which gave reasonable CFI values. But if we inspect Figure 14, 15 we find that the CFI values around $-b = -50$ only yield poor results. From this observation it seems wise to pick rather larger values for b, in the spirit of better save than sorry, since Figure 7 and 8 indicate that the specific slope value doesn't impact the performance.

Comparison of PSO and GA performance

By inspecting Figure 16 and Figure 17 we can see that in both cases that in the range of lower BIC-values the GA performance is better than the PSO algorithm measured by the BIC-criteria. For the worse BIC-values this relationship is reversed. Suggesting that the Genetic Algorithm is preferred over the PSO in the context of Confirmatory Factor Analysis.

6 Conclusion and Discussion

6.1 Conclusions

Main conclusions on hyperparameter selection

We found that in the uni-variate variation without exception that $\lambda = 0$ values had the best performing fit-measures. There is also a data-set dependent turning point say λ_i where we go from an interval of reasonable fit-measures values to an interval with bad performing fit-measures. For the PSO and the GA we get the following recommendations for the hyper-paramters-objective-functions values. In order of importance:

- λ values in $[0, 0.2]$ lead to reasonable results, avoid values in the neighbourhood $\lambda = 1$. A corollary to this is that the empirical data-sets apparently demand these higher degrees of complexity (smaller values for λ i.e. smaller punishment for model complexity) in order to able to provide a good model fit on the *out of sample data*.

- For the slope $-b$, one should at least choose, $b > 100$, so $-b < -100$. Choosing larger values for b leads the objective function to be more sensitive to differences of models whose fit is close to the threshold values used for centring, since larger values for the slope tend to perform better these higher sensitivity levels appear to be necessary for the tested data-sets.

- c_1, c_2, c_3 don't seem to have a significant effect on the fit-measures within a reasonable domain of the threshold values.

None of these conclusions should be interpreted as hard cut-off values but rather suggestions for ***avoiding*** bad performance of the algorithms in the context of CFA.

Conclusion meta-heuristic comparison of PSO and GA

On all the tested data-sets the GA resulted in better BIC-values than the PSO. We also found that the computation time for the Particle Swarm Optimisation is about a factor 1.5 higher than the Genetic Algorithm. This provides some evidence that the Genetic Algorithm is a better candidate for model specification in SEM than the Particle Swarm Optimisation.

Psychometric conclusions

While psychometric research was not the main research objective of this thesis and the data-sets used were more or less chosen for convenience and accessibility, an interesting observation from the results section is that the big-5-personality-traits in the IPIP data set seem to fit significantly better to before and after model specification than the Dark Triad personality trait. The best obtained CFI-value for the dark triad is 0.88 and the best obtained CFI-value for the IPIP-dataset is 0.96 (for multiple personality traits i.e. extraversion, consciousness, etc.). If we follow this line of reasoning we might conclude that there is evidence that the big-5 model for personality is a "good" personality construct, since we have a CFI > 0.95, after model specification, but the dark triad is not since this threshold wasn't reached after model specification for multiple possible hyperparameter settings.

6.2 Discussion

In this section the conclusions of the previous section are reviewed and recommendation for further research are given.
First of all the non-objective function parameters were not varied and usually chosen based on time and computational constraints. For example, the n_{gen} was reduced from its standard value of 50, since we noted that after 6 iterations the fit-measures didn't change much anymore and the computation time did greatly decrease if we choose a lower number of generations. These changes were mostly necessary for the IPIP data-set. It could always be that for some combinations of hyperparameters we missed an interaction with the non-objective function parameters which would change some of our conclusions. In the section on the comparison between the PSO and the GA, only one set of the non-objective function parameter is used. A subject for further investigation could be to vary the non-objective function parameters to see if they conclusions of the comparison still hold up, although some parameters like the number of generations and population size will always lead to better or equal results so one should probably look at the diminishing returns of an increase in these parameters to find a good stopping criterion, because they will be in direct trade-off with computation time.

This leads us to a second point that if the time and computational resources are available one can always try all the possible combinations of hyperparameters (generated by a fine grid) on the given data-set to reach a more conclusive result. With more computational resources one could evaluate finer and finer grids and one might use this study as a starting point for their further research.

In the simultaneous variation of λ and $-b$ we did observe the same patterns as in the marginal variation of λ and $-b$ but they are both shifted. We also only tested the dark triad data due to time and computational constraints.

Another point is the deletion of the missing values in the data-set which might introduce some form of bias in the data, which could be improved with knowledge on the data-gathering mechanism in combination with data-imputation techniques.

For the psychometric conclusion we should be carefull because it is a cross-sample/construct comparison. It could also be having something to do with the data-quality being worse for the dark triad compared to the big-5-personality construct. By setting up a new study one could gather data on both constructs and see if the difference in fit-measures still significantly persists. This might be interesting since both constructs try to model personality characteristics, so although they might be different constructs in this sense they might be comparable. Another comment we could make on the psychometric part is that we didn't use the full big-5-personality model but only individual sub-branches this was also due to the computational time.

As a final note, it is important to state that this thesis had a rather exploratory nature, which might serve as a building block towards future research in the application of metaheuristics for model specification in SEM.

References

[1] D. J. Bartholomew, "Spearman and the origin and development of factor analysis," *British Journal of Mathematical and Statistical Psychology*, vol. 48, Issue 2, pp. 211–220, 1995.

[2] J. Li, "Confirmatory factor analysis (cfa) in r with lavaan," https://stats.oarc.ucla.edu/r/seminars/rcfa/#s1c, [Online; accessed 16-May-2023].

[3] H. Muroshashi and H. Toyoda, "Model specification search using a genetic algorithm with factor reordering for a simple structure factor analysis model," *British Journal of Mathematical and Statistical Psychology*, vol. 49, No. 2, pp. 179–191, 2007.

[4] Z. Jing, H. Kuang, L. Walter, K. Marcoulides, and C. Fisk, "Model specification searches in structural equation modeling with a hybrid ant colony optimization algorithm," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 29 No. 5, pp. 655–666, 2022.

[5] U. Schroeders, S. Florian, and O. Gabriel, "Model specification searches in structural equation modeling using bee swarm optimization," *Educational and Psychological Measurement*, vol. 49, No. 2, pp. 179–191, 2023.

[6] K. Sourabh, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimedia Tools and Applications*, vol. 80, p. 8091–8126, 2021.

[7] M. Glover and A. Gary, *Handbook of Metaheuristics*. New York: Springer, 2009, doi: https://doi.org/10.1007/978-1-4419-1665-5.

[8] Y. Xiaoli, L. Shaoting, and C. Jiahua, "A three-parameter logistic regression model," *Statistical Theory and Related Fields*, vol. 5 Issue 3, pp. 265–274, 2021.

[9] D. Goretzko, K. Siemund, and C. Heumann, "Specification search in structural equation modeling as a combinatorial optimization problem: Using meta-heuristics to re-specify measurement models."

[10] K. Schermelleh-Engel and H. Moosbrugger, "Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures," *Methods of Psychological Research Online*, vol. 8 No.2, pp. 23–74, 2003.

[11] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21 Issue 4, pp. 36 – 47, 2004.

[12] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, 1995, pp. 39–43.

[13] D. L. Paulhus and M. W. Kevin, "The dark triad of personality: Narcissism, machiavellianism, and psychopathy," *Journal of Research in Personality*, vol. 8 Issue 6, pp. 556–563, 1990.

[14] A. Johnson, "Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120," *Journal of Research in Personality*, vol. 5 Issue 3, pp. 78–89, 2014.

[15] Y. Rosseel, T. D. Jorgensen, and N. Rockwood, *lavaan: Latent Variable Analysis*, 2023, r package version 0.6.12. [Online]. Available: https://CRAN.R-project.org/package=lavaan

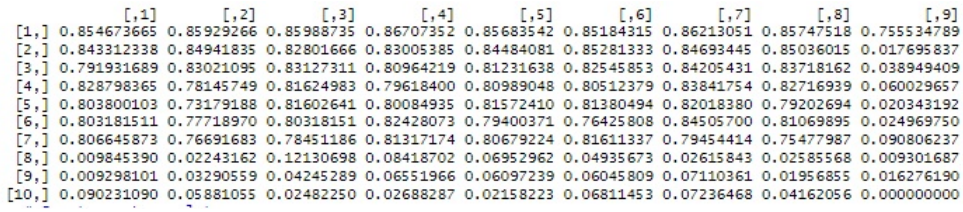
[16] A. Lishinski, *lavaanPlot: Path Diagrams for 'Lavaan' Models via 'DiagrammeR'*, 2023, r package version 0.6.2. [Online]. Available: https://CRAN.R-project.org/package=lavaanPlot

[17] H. Wickham, D. Vaughan, M. Girlich, and K. Ushey, *dplyr: A Grammar of Data Manipulation*, 2023, r package version 1.0.10. [Online]. Available: https://CRAN.R-project.org/package=dplyr

[18] ——, *tidyr: Tidy Messy Data*, 2023, r package version 1.2.1. [Online]. Available: https://CRAN.R-project.org/package=tidyr

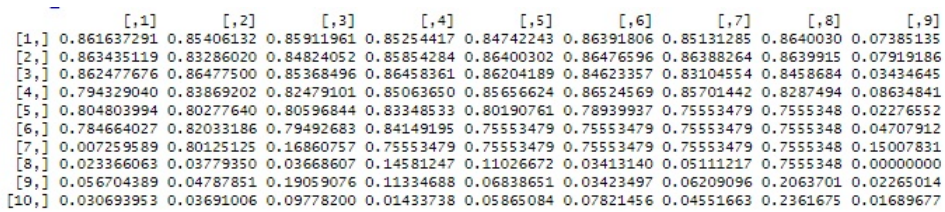
[19] H. Wickham, W. Chang, L. Henry, T. L. Pedersen, K. Takahashi, C. Wilke, K. Woo, H. Yutani, and D. Dunnington, *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, 2023, r package version 3.4.0. [Online]. Available: https://CRAN.R-project.org/package=ggplot2

7 Appendix



	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	0.854673665	0.85929266	0.85988735	0.86707352	0.85683542	0.85184315	0.86213051	0.85747518	0.755534789
[2,]	0.843312338	0.84941835	0.82801666	0.83005385	0.84484081	0.85281333	0.84693445	0.85036015	0.017695837
[3,]	0.791931689	0.83021095	0.83127311	0.80964219	0.81231638	0.82545853	0.84205431	0.83718162	0.038949409
[4,]	0.828798365	0.78145749	0.81624983	0.79618400	0.80989048	0.80512379	0.83841754	0.82716939	0.060029657
[5,]	0.803800103	0.73179188	0.81602641	0.80084935	0.81572410	0.81380494	0.82018380	0.79202694	0.020343192
[6,]	0.803181511	0.77718970	0.80318151	0.82428073	0.79400371	0.76425808	0.84505700	0.81069895	0.024969750
[7,]	0.806645873	0.76691683	0.78451186	0.81317174	0.80679224	0.81611337	0.79454414	0.75477987	0.090806237
[8,]	0.009845390	0.02243162	0.12130698	0.08418702	0.06952962	0.04935673	0.02615843	0.02585568	0.009301687
[9,]	0.009298101	0.03290559	0.04245289	0.06551966	0.06097239	0.06045809	0.07110361	0.01956855	0.016276190
[10,]	0.090231090	0.05881055	0.02482250	0.02688287	0.02158223	0.06811453	0.07236468	0.04162056	0.000000000

Figure 18: *Raw grid values of the CFI that were used to create Figure 14 with the GA. The rows represent lambda and the columns represent $-b$. The total computation-time was $T = 17$ hrs. Element 0,0 represent the grid values (0,-200).*



	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	0.861637291	0.85406132	0.85911961	0.85254417	0.84742243	0.86391806	0.85131285	0.8640030	0.07385135
[2,]	0.863435119	0.83286020	0.84824052	0.85854284	0.86400302	0.86476596	0.86388264	0.8639915	0.07919186
[3,]	0.862477676	0.86477500	0.85368496	0.86458361	0.86204189	0.84623357	0.83104554	0.8458684	0.03434645
[4,]	0.794329040	0.83869202	0.82479101	0.85063650	0.85656624	0.86524569	0.85701442	0.8287494	0.08634841
[5,]	0.804803994	0.80277640	0.80596844	0.83348533	0.80190761	0.78939937	0.75553479	0.7555348	0.02276552
[6,]	0.784664027	0.82033186	0.79492683	0.84149195	0.75553479	0.75553479	0.75553479	0.7555348	0.04707912
[7,]	0.007259589	0.80125125	0.16860757	0.75553479	0.75553479	0.75553479	0.75553479	0.7555348	0.15007831
[8,]	0.023366063	0.03779350	0.03668607	0.14581247	0.11026672	0.03413140	0.05111217	0.7555348	0.00000000
[9,]	0.056704389	0.04787851	0.19059076	0.11334688	0.06838651	0.03423497	0.06209096	0.2063701	0.02265014
[10,]	0.030693953	0.03691006	0.09778200	0.01433738	0.05865084	0.07821456	0.04551663	0.2361675	0.01689677

Figure 19: *Raw grid values of the CFI that were used to create Figure 15 with the PSO. The rows represent lambda and the columns represent $-b$. The total computation-time was $T = 25.5$ hrs. Element (0,0) represent the grid-values (0,-200).*