Model Specification Search with Metahuristics Focusing on Practical Use

Hiroto Murohashi
(School of Letters, Arts and Sciences, Waseda University)

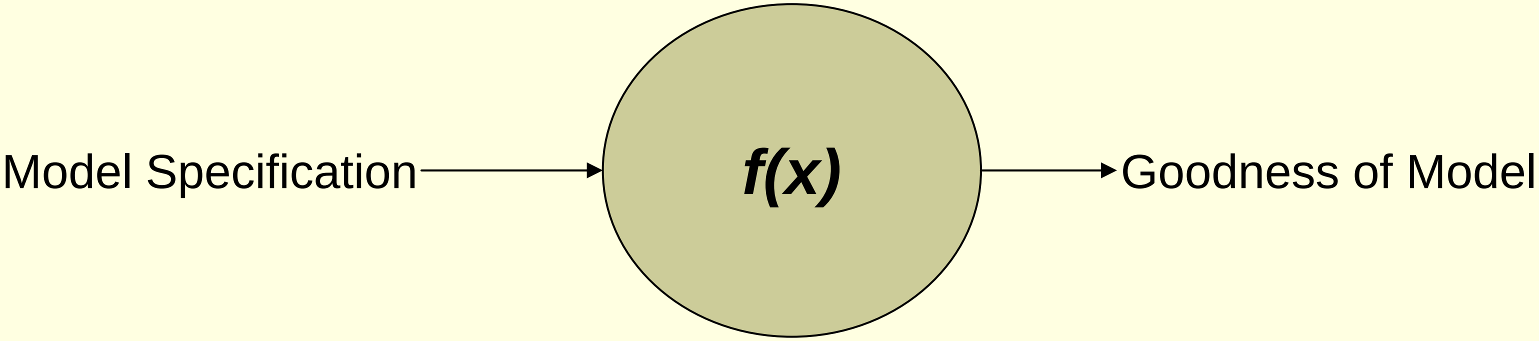
Hideki Toyoda
(School of Letters, Arts and Sciences, Waseda University)

Importance of Model Specification Search in SEM

- Model specification search is needed especially in application study
  - To adjust model to collected data
  - To explore with their data to find appropriate model

Problem Establishment

- Model specification search is a combinational optimization problem to optimize function
  - which search space is very vast



Performance limitations of existing methods (1)

- Greedy method
  - All possible subset selection approach (Joreskog & Sorborn, 1990)

- It works well only for the situations where the search space is small
  - In most cases, it is impractical because they need too much calculating time

Performance limitations of existing methods (2)

- Local search method
  - LM test
  - Wald test

- It only works well when the proposed model is located near the true model
  - The result depends on the initial value
  - There's no guarantee that it can find globally optimal solution

One Solution to Problem

- To use Metaheuristics
  - It can search more large space than local search method in relatively simple algorithm
- Metahuristics showed good result in model specification study
  - Marcoulides, Drezner & Schumacker, 1998
  - Marcoulides & Drezner, 2001
  - Harwood & Scheines, 2002
  - Marcoulides & Drezner, 2003
  - Murohashi & Toyoda, 2004

Purpose of Study

- To propose algorithm for model specification search of factor analysis model using metahuristics
    - Focus on the specific situation
    - Contrive ways to make searching to be faster and to be more accurate

Framework of the method

- 1st step: search for simple structure factor analysis model with genetic algorithm
- 2nd step: search for usual factor analysis model (release the restriction)

- In 1st step, we limit search space to simple structure model because:
  - In practical situation, simpler models are usually recommended
  - To make the search more efficient (fast and accurate)

1^{st} Step: Overview

- Use Genetic Algorithm (Holland, 1975)
  - Algorithm based on the mechanism of natural selection
  - Directly operate “chromosomes”, not the solution itself

1. Generate individuals at random to form the initial population
2. Calculate the goodness of fit of each individual in the population
3. Create new population from existing population reflecting the goodness of each individual
4. Back to step 2 until the criterion is satisfied

1st step: Encoding(1)

- Our target is simple structure factor analysis model
- Assume as follows:
  - Always estimate all covariance among factors
  - Fix each factor's variance at 1.0 for model identification
  - Don't estimate error covariance among observed variables

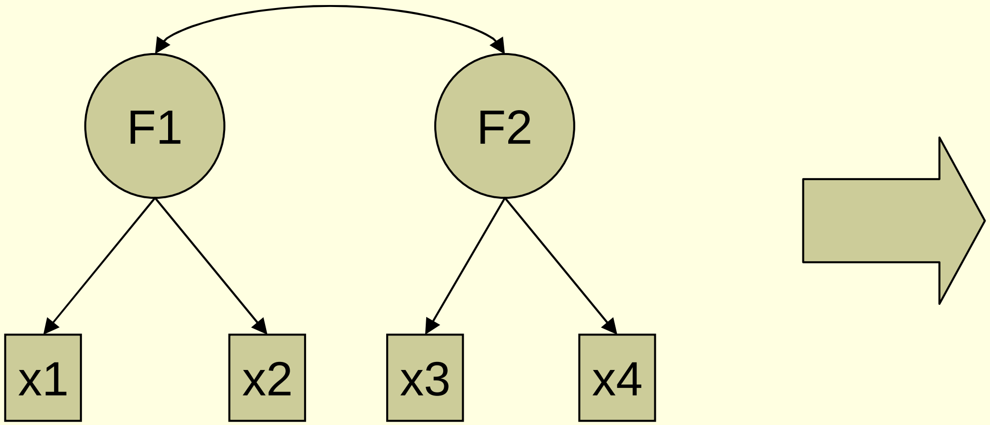
1st step: Encoding(2)

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{f}}\mathbf{A}' + \boldsymbol{\Sigma}_{\mathbf{e}}$$

- Σ f
  - Diagonal elements are to be fixed at one
  - Non-diagonal elements are to be free
- Σ e
  - Diagonal elements are to be free
  - Non-diagonal elements are to be fixed at zero
- A
  - Every observed variable measures only one factor

1st step: Encoding(3)

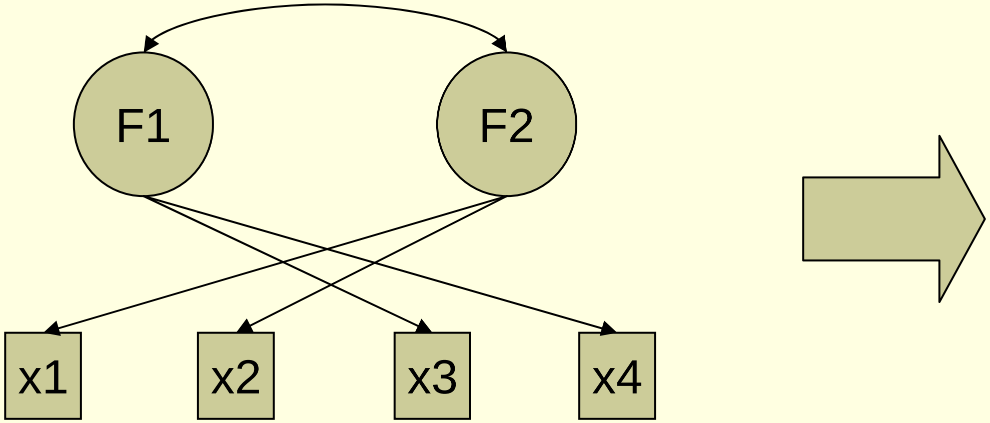
- The information that which factor the observed variables measures can be treated as chromosome



[1 1 2 2]

1st step: Factor reordering(1)

- It is unidentifiable which factor should be the “first” one and so on
  - Same model can have different chromosome



[2 2 1 1]

1st step: Factor reordering(2)

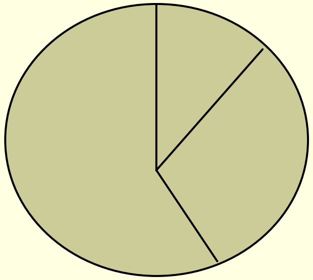
- Reorder factors in the ascending order corresponding to the observed variables which assigned the youngest number among the observed variables measuring same factor

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

[3 3 2 1 2] [1 1 2 3 2]

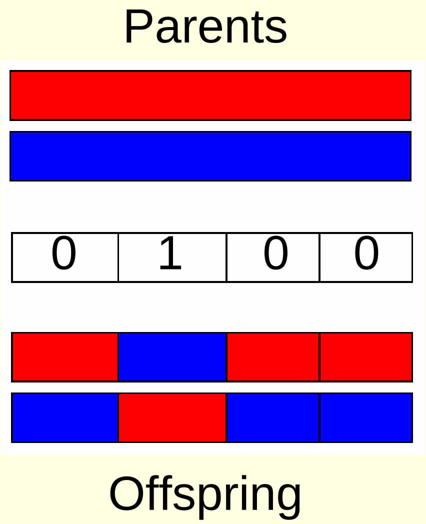
1st step: Selection

- To evaluate each chromosomes goodness
  - (1-RMSEA)^2
- To select the parents to make children
  - Roulette Wheel Selection Algorithm
    - Select populations from virtual roulette wheel that represents the proportion of fitness function's value of all the population members



1st step: Reproduction

- Crossover
  - Uniform crossover algorithm
    - Generate offspring from the parents' chromosome, based on a randomly generated 0/1 mask
- Mutation
  - each gene of chromosome is checked individually and if mutation occurs, the value of that gene will be changed to another value.



1st step: Local search(1)

- Genetic algorithm is basically for global search
  - Can search large area efficiently and fast
  - Risk of overlooking good model near existing individual
- To use Genetic Local Search Algorithm is recommended
  - Combination of genetic algorithm & local search algorithm

1st step: Local search(2)

- Wald & LM Test: for all individual
  1. Find the observed variable which has the lowest value of Wald statistic
  2. Change the factor which that observed variable measures to the one which has the highest value of LM statistic
  3. If the new model has better fit than old one, replace individual
- 1-flip neighborhood search: only for the best individual
  1. Changed the value of gene one at a time from the head of chromosome
  2. If the new model has better fit than old one, replace individual

1st step: Replacement

- Eliminate individual which represents model specification that causes problem
  - Improper solution
- Keep each pattern of chromosome one
  - To keep the variety of information
- Employed Elitism
  - To keep some individuals which have good chromosomes preferentially

2nd step: Overview

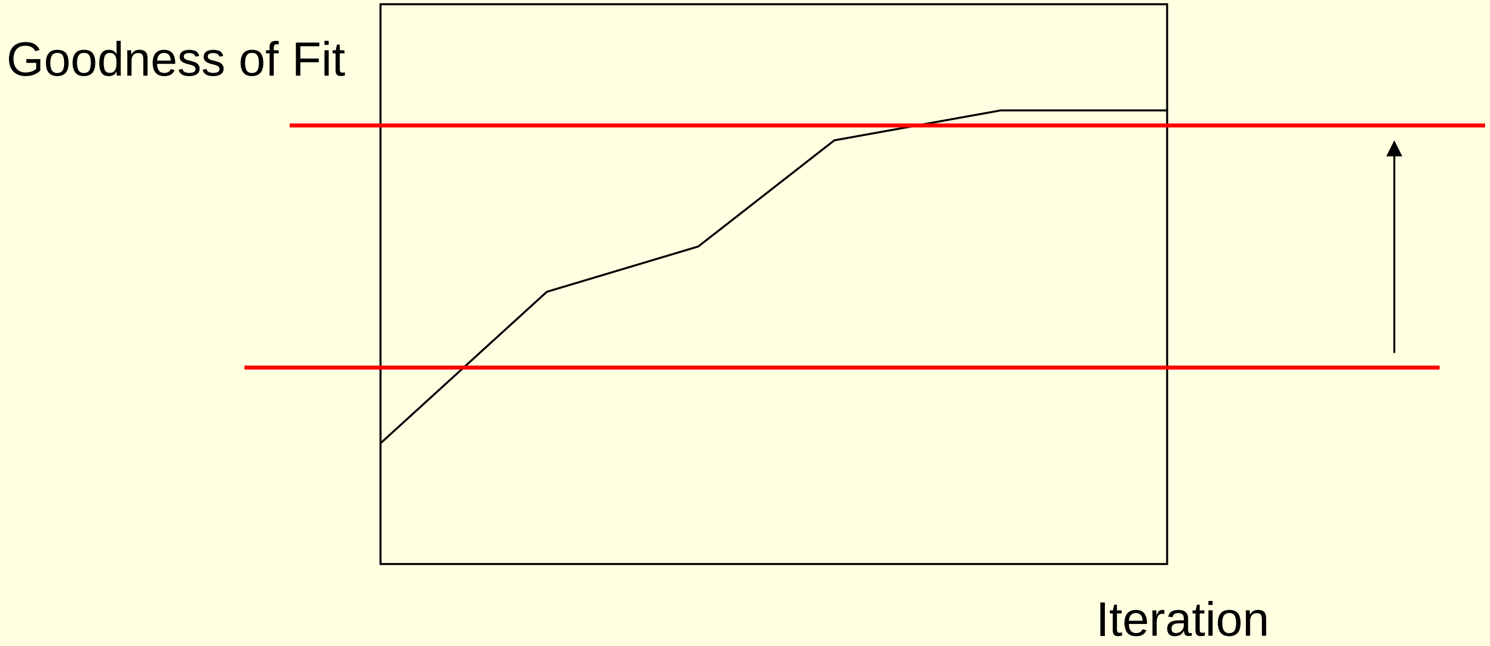
- Search the neighborhood of the model which is found in 1st step
  - Don’t limit to simple structure
  - But don’t go too far from the starting point
- We can use existing local search method
  - Modification Indices
  - TETRAD
- Here, we propose to use great deluge algorithm

2nd step: Procedure(1)

1. Calculate Wald and LM statistic for factor loading matrix
2. Compare the result by p-value and choose three positions which has relatively bad fit
3. At each position, consider modification plan as follows:
    1. If the position is not freely estimated, set the parameter free
    2. If the position is already freely estimated, constrain the parameter to be zero
    3. If and only if the position is already estimated and is the only parameter which is freely estimated in that row, constrain that parameter to be zero and set the parameter which has the lowest LM statistic free
4. Combine three plans and finally consider 2^3-1=7 pattern alternatives to improve the goodness of fit
5. Choose the alternative which improves the goodness of fit the most

2nd step: Procedure(2)

- Additional condition to avoid the model being too complicated
  - Make the criteria strict along with the increase of iteration



Example Analysis: Data

- Test data from Holzinger & Swineford(1939)
  - Examinees were Grant-White Elementary School’s 7th and 8th grade students
  - N=145
- Test consisted of 26 items from 5 different areas, and we used 24 variables (without item 3, 4)
  - Spatial
  - Verbal
  - Speed
  - Memory
  - Mathematical ability

Example Analysis: 1st step(1)

- The set up was as follows:
  - Number of factor = 4
  - Size of population = 30
  - Mutation rate = 0.05
  - Number of elite individual = 3
- The whole algorithm was implemented using *proc iml* of SAS v8.02

Example Analysis: 1st step(2)

- The result was

[1 2 3 3 3 3 3 4 4 4 4 2 2 1 2 4 2 2 1 2 2 2 1 1]

  - Fit indices was as follows:
    - RMSEA = 0.0724
    - AGFI = 0.7530
    - RMR = 0.0767
- The number of genetic cycle elapsed was 12

Example Analysis: 2nd step(1)

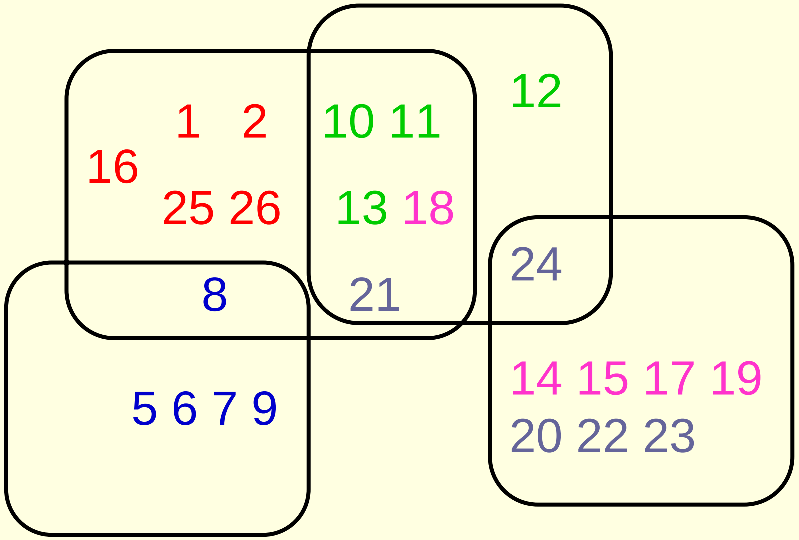
- Finishing criteria for *i*-th iteration was as follows:

$$R_{i-1} - R_{i-1} \times \left(\sqrt{Log[i]} - 1\right)$$

  - i: Number of iteration
  - R_{i-1} : RMSEA of the last iteration

Iteration	RMSEA	Criteria
0	0.0724	–
1	0.0647	0.1448
2	0.0586	0.075534
3	0.0552	0.055779
4	0.0518	0.045407

Example Analysis: 2nd step(2)



- Goodness of fit is as follows:
  - RMSEA = 0.0552
  - AGFI = 0.7967
  - RMR = 0.0610

Conclusion

- The proposed 2-step algorithm showed result in model specification search of factor analysis model
- In 1st step, by restricting search space to simple structure:
  - The search space itself was reduced
  - It become easy to deal with
    - Equivalent model
    - Model parsimony
- In 2nd step, by using ??? Method:
  - We could find reasonably complicated model which balance goodness of fit and interpretability

Further Topics

- Extend 2nd step to the search for error covariance
- After this proposed procedure, execute another searching procedure which do path analysis among factors