Make Some Noise: Generating Data From Imperfect Models

Justin Kracht

September 7, 2022

University of Minnesota

Presentation Overview

1. Background: What is model error, and why does it matter?
2. Simulating Model Error: What methods are available to simulate model error?
3. Quantifying Model Error: What model fit indices are used to quantify model error, and how do they differ?
4. Study Aims
5. Methods
6. Results

Background

Model Error and Covariance Structure Models

Covariance structure models allow a structured covariance matrix to be represented as a function of a vector of parameters,

$$\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\gamma}), \tag{1}$$

where $\boldsymbol{\Omega}$ is a $p \times p$ model-implied covariance matrix and γ is a vector of free parameters.

> *"All models are wrong, but some are useful"*
> *— Box (1987, p. 424)*

What is Model Error?

Psychological phenomena are complex; no population covariance matrix will be perfectly represented by $\boldsymbol{\Omega}(\boldsymbol{\gamma})$ in practice.

Model error can be grouped into two basic categories (Meehl, 1990).

- Incompleteness: The model is too simple to adequately reflect reality (e.g., there are more common factors than are modeled).
- Falsity: There are contradictions between the model and the world (e.g., non-linear relationships are modeled as linear).

Representing Model Error

An population covariance matrix with model error can be represented by

$$\boldsymbol{\Sigma} = \boldsymbol{\Omega} + \mathbf{E}, \tag{2}$$

where $\mathbf{E}$ is a symmetric error matrix representing the effects of model error.

Why Should We Care About Model Error?

Monte Carlo simulation studies that sample from $\boldsymbol{\Omega}$ (rather than $\boldsymbol{\Sigma}$) are likely to produce overly-optimistic results.

Simulation work has shown that model error can affect:

- Parameter estimation for exploratory factor analysis (Briggs, 2003).
- Dimensionality identification with parallel analysis (Kracht, 2020).
- Behavior of confidence regions and fungible parameter estimates for structural equation models (Pek, 2012).
- ...And other statistical procedures (Beauducel, 2016; de Winter, 2016, Gnambs, 2016; Hsu, 2015; Trichtinger, 2020).

Problem: How can we simulate $\boldsymbol{\Sigma}$ with "realistic" model error?

Simulating Model Error

Existing Model Error Methods

Three of the most prominent model error methods are:

1. Tucker, Koopman, and Linn (TKL; 1969)
2. Cudeck and Browne (CB; 1992)
3. Wu and Browne (WB; 2015)

The Tucker, Koopman, and Linn Method

The TKL method is based on the common factor analysis model for k common factors:

$$\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}, \tag{3}$$

where

- $\boldsymbol{\Omega}_{p\times p}$: model-implied covariance matrix.
- $\boldsymbol{\Lambda}_{p\times k}$: factor-pattern matrix.
- $\boldsymbol{\Phi}_{k\times k}$: common factor covariance matrix.
- $\boldsymbol{\Psi}_{p\times p}$: diagonal matrix containing the unique variances.

Let all common factors be standardized in the population so that $\boldsymbol{\Omega}$ has a unit diagonal (i.e., is a correlation matrix).

The Tucker, Koopman, and Linn Method

Model error is represented as the effect of numerous minor common factors such that $\mathbf{W}_{p \times q}$ is the matrix of minor factor loadings for the $q \succeq k$ minor common factors,

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}' \tag{4}$$

Minor common factors are "...far too many and far too minor to be retained in a factor analysis of empirical data" (MacCallum, 2003, p. 135).

The Tucker, Koopman, and Linn Method: User Parameters

Two user-specified parameters affect the characteristics of $\mathbf{W}$:

- $\nu_E \in [0, 1]$: Proportion of unique variance allocated to the minor common factors.
- $\epsilon \in [0, 1]$: Controls how variance is distributed among the minor common factors. Values close to zero result in relatively equipotent minor factors whereas values of close to one result in error variance primarily being distributed to the first minor factor.

The Tucker, Koopman, and Linn Method

Advantages

- Straightforward interpretation of model error arising from un-modeled minor common factors.
- Quite flexible; two parameters (ν_E and ϵ) to manipulate.
- Relatively easy to implement.

Disadvantages

- Cannot be used for all types of covariance structure models (factor analysis models only).
- No clear guidelines for reasonable/realistic values of ν_E and ϵ.
- No direct control of degree of model fit (in terms of e.g., RMSEA) as traditionally implemented.

The Cudeck and Browne Method

Cudeck and Browne developed a model error method for any covariance structure model, $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\gamma})$.

For a particular vector of model parameters $\boldsymbol{\gamma}_0$, let $\boldsymbol{\Sigma}_0 = \boldsymbol{\Omega}(\boldsymbol{\gamma}_0) + \mathbf{E}$

Their method seeks to find an $\mathbf{E}$ matrix such that

1. $F(\boldsymbol{\Sigma}_0, \boldsymbol{\Omega}(\boldsymbol{\gamma}))$ is minimized when $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$.
2. $F(\boldsymbol{\Sigma}_0, \boldsymbol{\Omega}(\boldsymbol{\gamma})) = \delta$ when $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ for some user-specified δ.

Here, $F(\boldsymbol{\Sigma}_0, \boldsymbol{\Omega}(\boldsymbol{\gamma}))$ is a discrepancy function (ML or OLS).

The Cudeck and Browne Method

Advantages

- Adaptable to any covariance structure model.
- Allows a target RMSEA value to be directly specified (because δ is related to RMSEA; more on that later).
- No structural assumptions about $\mathbf{E}$; any suitable $\mathbf{E}$ matrix will do.

Disadvantages

- Can result in indefinite $\boldsymbol{\Sigma}$ matrices if δ is too large.
- Relatively difficult to implement.
- Inflexible; only one parameter to control.
- Implies that $\boldsymbol{\gamma}_0$ will be perfectly recovered from a sample covariance matrix as sample size goes to ∞ (is this reasonable?).

The Wu and Browne Method

Model error is considered to be a random effect due to differences between the operational population and some ideal population for which the theory is hypothesized.

Specifically, $\boldsymbol{\Sigma}$ is considered to be a random sample from an inverse-Wishart distribution such that

$$(\boldsymbol{\Sigma}|\boldsymbol{\Omega}) \sim W_p^{-1}(m\boldsymbol{\Omega}, m), \tag{5}$$

where $m = 1/v$ is a continuous precision parameter such that $m > p - 1$.

> *"[The ideal] population need not have an explicit empirical description. In this sense, the general population is defined by the model rather than by its empirical nature."*
> *— MacCallum and O'Hagan (2015, p. 605)*

The Wu and Browne Method

Advantages

- Adaptable to any covariance structure model.
- Fast and relatively easy to implement.
- Allows a target RMSEA value to be specified.

Disadvantages

- Inflexible; only one parameter to control.
- Resulting RMSEA values are often not close to target values when target values are (relatively) large.
- Possible RMSEA values limited by $m > p - 1$.
- Based on a specific theory of model error; assumes that error-perturbed covariance matrices are distributed according to an inverse Wishart distribution.

Evaluating Model Error: Fit Indices

Many population model fit indices have been used to quantify the discrepancy between:

- $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$
- $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Omega}}$ (the implied covariance matrix from an analysis of $\boldsymbol{\Sigma}$)

Absolute Fit Indices

- Root Mean Square Error of Approximation (RMSEA; Steiger, 1990)
- Standardized Root Mean Square Residual (SRMR; Hu & Bentler, 1999)
- Correlation Root Mean Square Residual (CRMR; Bollen, 1989; Ogasawara, 2001)

Incremental Fit Indices

- Comparative Fit Index (CFI; Bentler, 1990)
- Tucker-Lewis Index (TLI; Tucker & Lewis, 1973)

Evaluating Model Error: Disagreement Among Indices

Different model fit indices can lead to different qualitative interpretations of model fit when cut-off values are used to categorize model fit.

Example: RMSEA($\boldsymbol{\Sigma}$, $\boldsymbol{\Omega}$) $=$ 0.04; CFI($\boldsymbol{\Sigma}$, $\boldsymbol{\Omega}$) = 0.78.

- RMSEA values less than 0.05 generally represent good model fit.
- CFI values less than 0.90 generally represent unacceptably poor model fit.

> *"[T]here is no such thing as a magical, single-number summary that says everything worth knowing about model fit."*
> —*Kline (2011, p. 193)*

Controlling the Amount of Model Error

Given that fit indices can lead to different qualitative interpretations of model fit, researchers should use and report multiple model fit indices in Monte Carlo simulation studies.

- Can model error methods be used as-is or modified to produce $\boldsymbol{\Sigma}$ matrices with model-fit statistic values close to the target values? How effective are they in this task?
- How do the model fit methods differ in terms of CFI, TLI, and SRMR/CRMR for error-perturbed covariance matrices, when holding RMSEA values approximately equal?

TKL: Specifying the Amount of Model Error

Problems

- There are no empirically-supported guidelines for appropriate values of ν_E and ϵ.
- Choosing values of ν_E and ϵ that result in a specific model fit index value is difficult.
- Choosing values of ν_E and ϵ that result in multiple specific model fit index values (e.g., RMSEA and CFI) can be *very* difficult (sometimes impossible).

Solution

- Create an optimization procedure to find values of ν_E and ϵ that give RMSEA and/or CFI values that are as close as possible to target values.

TKL: Optimization Procedure (RMSEA & CFI)

Use the L-BFGS-B (Zhu et al., 1997) algorithm to minimize the function:

$$G(\nu_E, \epsilon) = b_1 \left[\varepsilon - \varepsilon_T\right]^2 + b_2 \left[CFI - CFI_T\right]^2 + \mathbf{1}_W \lambda \tag{6}$$

- b_1 and b_2: User-specified weights that sum to one.
- ε_T: User-specified target RMSEA value.
- CFI_T: User-specified target CFI value.
- $\mathbf{1}_W$: Indicator function that equals one if any minor factor has more than two factor loadings $\geq .3$ in absolute value.
- λ: User-specified penalty.

CB: Specifying the Amount of Model Error (RMSEA)

The CB method allows a user to generate an error-perturbed covariance matrix with a specified RMSEA value.

$$\delta = \varepsilon_T df, \tag{7}$$

where *df* denotes the model degrees of freedom.

- $\boldsymbol{\Sigma}$ can be indefinite if δ is too large.
- If δ is large, θ_0 might correspond to a saddle point.

WB: Specifying the Amount of Model Error (RMSEA)

The error-perturbed covariance matrix $\boldsymbol{\Sigma}$ is sampled from:

$$(\boldsymbol{\Sigma}|\boldsymbol{\Omega}) \sim W_p^{-1}(m\boldsymbol{\Omega}, m), \tag{8}$$

The precision parameter, $m = 1/v$, is related to RMSEA such that $v \approx \varepsilon^2$.

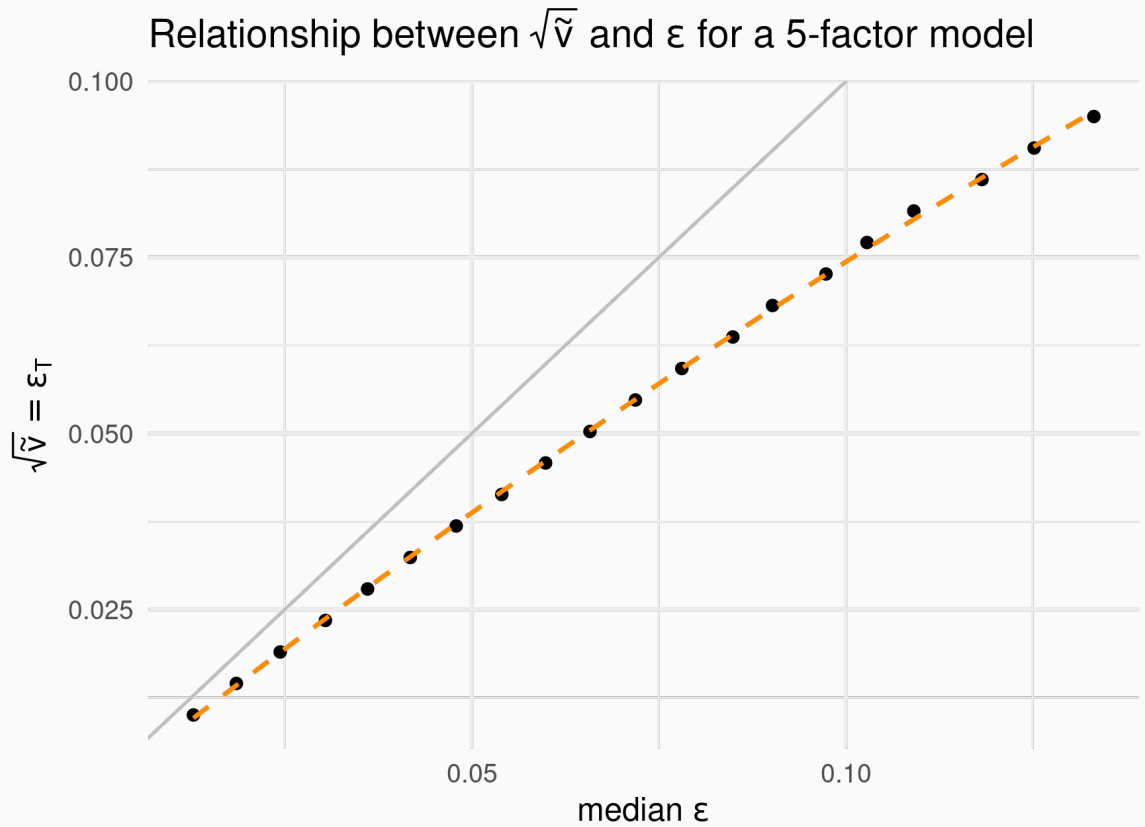
The approximation gets worse as ε^2 increases; simply using $m = 1/\tilde{v}$ is unlikely to lead to RMSEA values close to ε_T when ε_T is not very small.

WB: Fit a Regression Model for v

Find an appropriate value of v such that RMSEA$(\boldsymbol{\Sigma}, \boldsymbol{\Omega}) \approx \varepsilon_T$:

1. Create a vector of $\tilde{v}$ values for ε_T values in a reasonable range (e.g., 20 values between 0.01 and 0.095).
2. For each value of $\tilde{v}$:
   - Sample a number of covariance matrices (e.g., 50) from the corresponding inverse-Wishart distribution.
   - Calculate the median RMSEA value for the sampled covariance matrices.
3. Regress $\tilde{v}$ on the median RMSEA and squared median RMSE values.
4. Use the fitted model from Step 3 to find a value of v that is likely to lead to error-perturbed covariance matrices with RMSEA values close to ε_T.

WB: Example



The solid gray line indicates where the target RMSEA and median RMSEA values would be equal.
The dashed orange line indicates the predicted value of v given an RMSEA value.

Aims of the Proposed Study

Two main aims:

1. Determine how error-perturbed covariance matrices produced by the TKL, CB, and WB model error methods differ in terms of CFI, TLI, and SRMR when matched (approximately) on RMSEA.
2. Determine how well (a) the proposed optimization procedure for controlling RMSEA and CFI with the TKL method and (b) the regression procedure for controlling RMSEA with the WB method work.

Method

Design Variables: Population Models

- Population models with 1, 3, 5 or 10 major common factors.
- Data sets with either 5 or 15 items per factor ($p \in [5, 15, 25, 45, 50, 75, 150]$).
- Common factor correlations either 0.0 (orthogonal), 0.3, or 0.6.
- Three factor loading structure types (simple structure): "Strong", "Moderate", and "Weak", corresponding to factor loadings of 0.8, 0.6, and 0.4 (Hair et al., 2018).

Design Variables: Model Error

- Model error methods
  - TKL (using target RMSEA, target CFI, and target RMSEA/CFI values; with and without constraints)
  - CB
  - WB
- Model fit targets
  - $\varepsilon_T \in [0.025, 0.065, 0.090]$.
  - $CFI_T \in [0.99, 0.95, 0.90]$.
  - ε_T and CFI_T pairs correspond to very good, fair, and poor model fit (MacCallum et al., 2001; Myers et al., 2015).

Note that target CFI only affects the TKL (RMSEA/CFI) procedure.

Design Variables: Summary

Table 1: Design Variables and Levels.

Variable	Levels
Factors	1, 3, 5, 10
Items/Factor	5, 15
Factor Correlation (ϕ)	0.0, 0.3, 0.6
Loadings	0.4, 0.6, 0.8
Target Model Fit	Very Good ($\varepsilon_T = 0.025$, $CFI_T = 0.99$), Fair ($\varepsilon_T = 0.065$, $CFI_T = 0.95$), Poor ($\varepsilon_T = 0.090$, $CFI_T = 0.90$)
Error Method	TKL (six variants), CB, WB

Study Design Variables: Summary

- 42 (error-free) population models.
- 1,440 crossed conditions.
- 500 reps for all model-error methods.
- $500 \times 1,440 = 720,000$ simulated $\boldsymbol{\Sigma}$ matrices.

Study Design: Simulation Procedure

Data generation

- For each (error-free) population model, generate the model-implied correlation matrix ($\boldsymbol{\Omega}$).
- For each model-implied correlation matrix:
  - Generate 500 error-perturbed covariance matrices ($\boldsymbol{\Sigma}$) for each of the TKL, CB, and WB model-error methods at each ε_T and CFI_T value pair.

Evaluation

For each error-perturbed correlation ($\boldsymbol{\Sigma}$) matrix:

- Compute RMSEA, CFI, TLI, and CRMR for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ (and for $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Omega}}$).
- Compute:

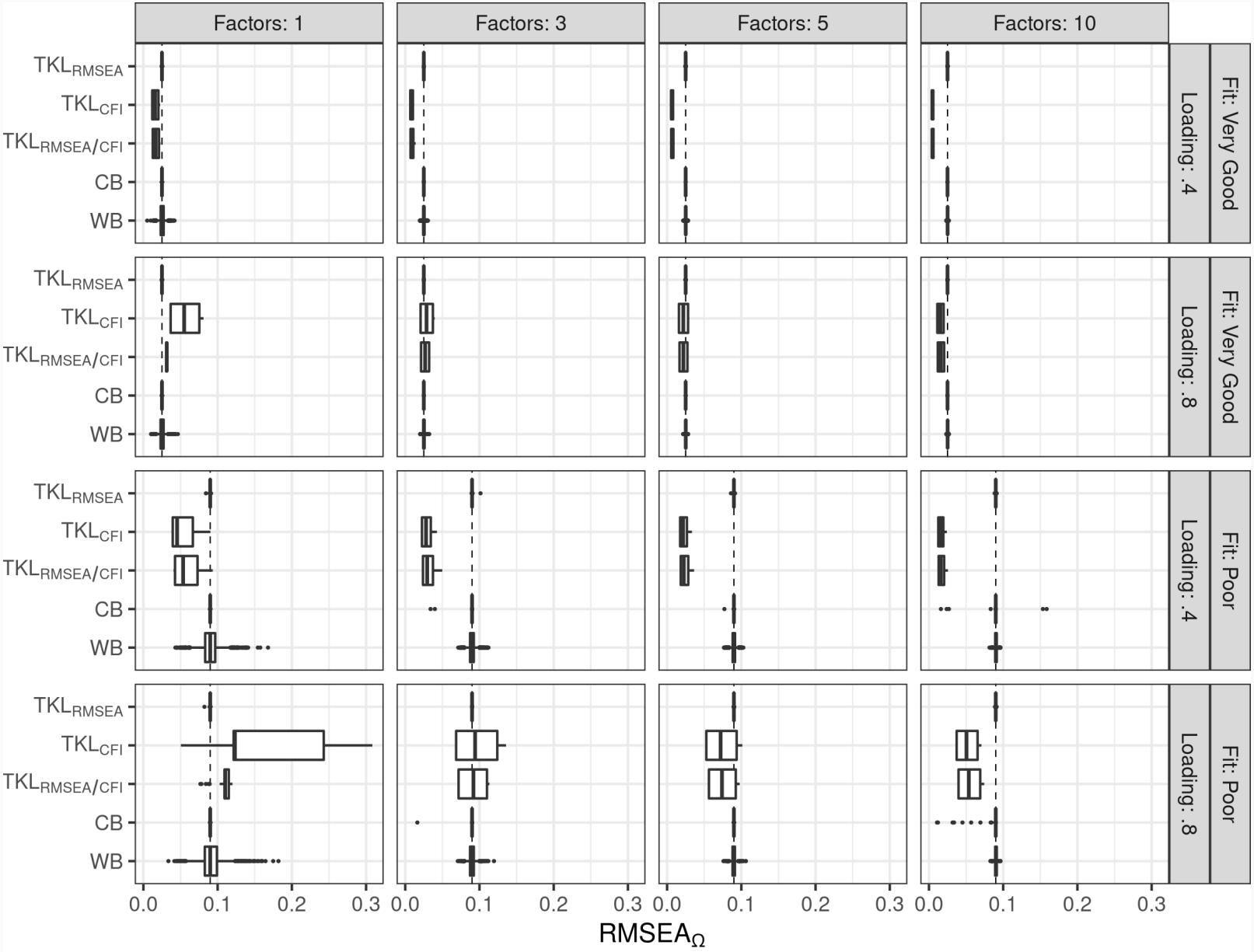
$$D = |\mathrm{RMSEA}_{\mathrm{obs}} - \mathrm{RMSEA}_{\mathrm{T}}| + |\mathrm{CFI}_{\mathrm{obs}} - \mathrm{CFI}_{\mathrm{T}}| \tag{9}$$

Results

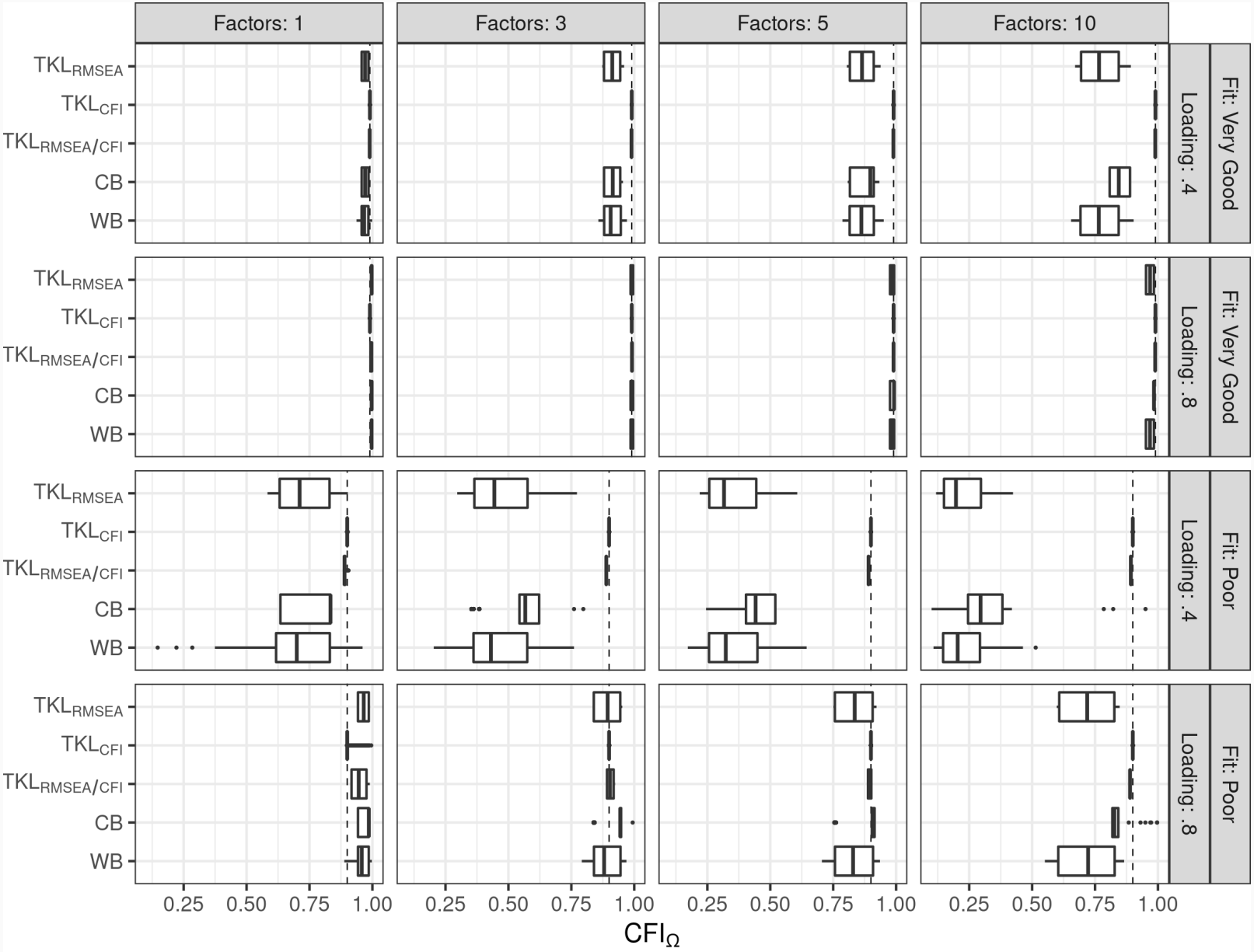
Section Overview

- First topic
- Second topic
- Third topic

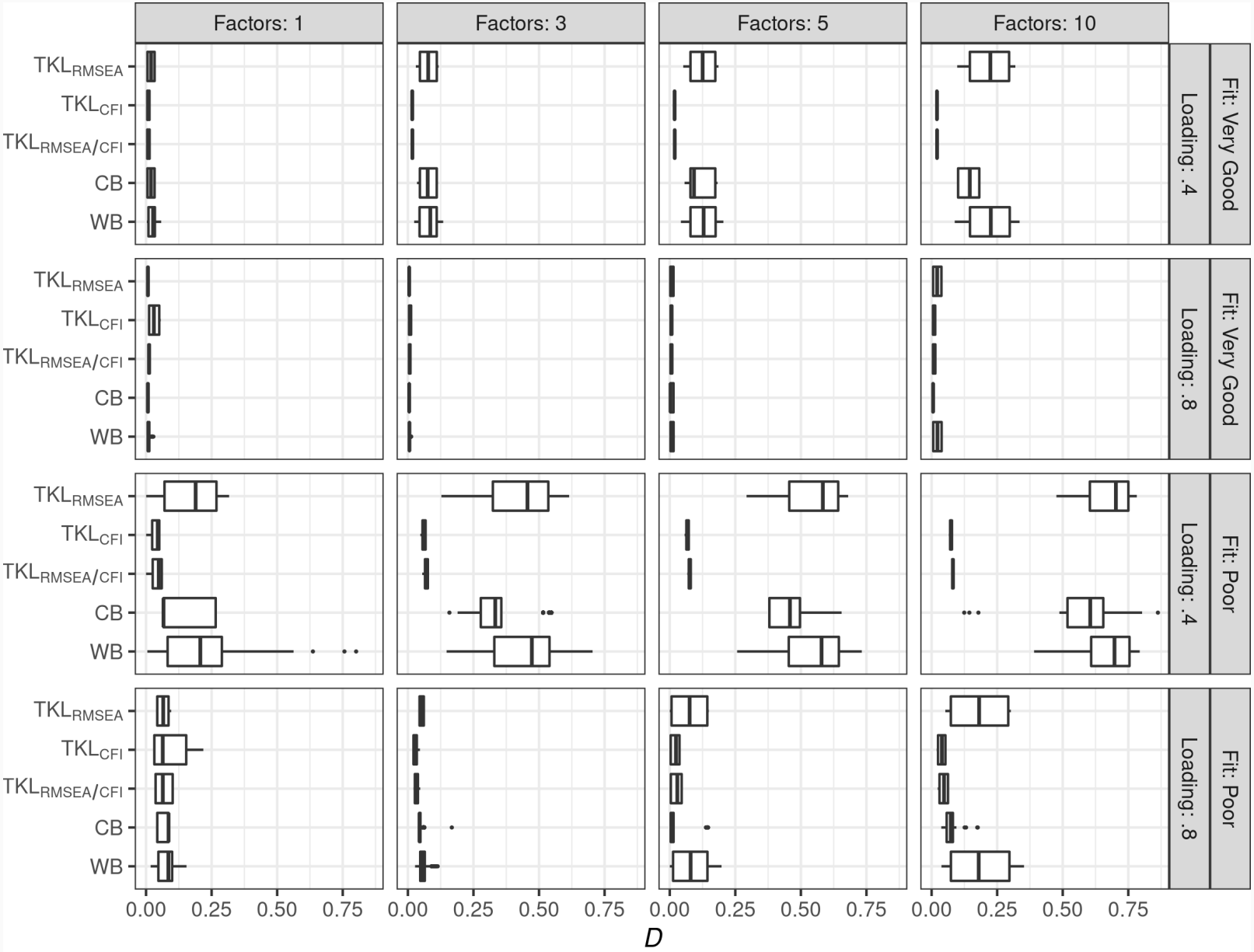
Distribution of RMSEA_{Ω}



Distribution of $\mathbf{CFI}_{\Omega}$



Fit Index Agreement: *D*



Backup Slides

Population Model Fit Indices

$$\text{RMSEA} = \varepsilon = \sqrt{\frac{F_h}{df_h}}, \tag{10}$$

$$\text{CFI} = 1 - \frac{F_h}{F_b}, \tag{11}$$

$$\text{TLI} = 1 - \frac{F_h/df_h}{F_b/df_b} \tag{12}$$

- F_h and df_h denote the discrepancy function value and degrees of freedom for the full model.
- F_b and df_b denote the discrepancy function value and degrees of freedom for the baseline (independence) model.

Population Model Fit Indices

$$\text{SRMR} = \sqrt{\frac{1}{p(p+1)/2} \sum_{i \leq j} \left(\frac{\sigma_{ij} - \omega_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \right)^2} \tag{13}$$

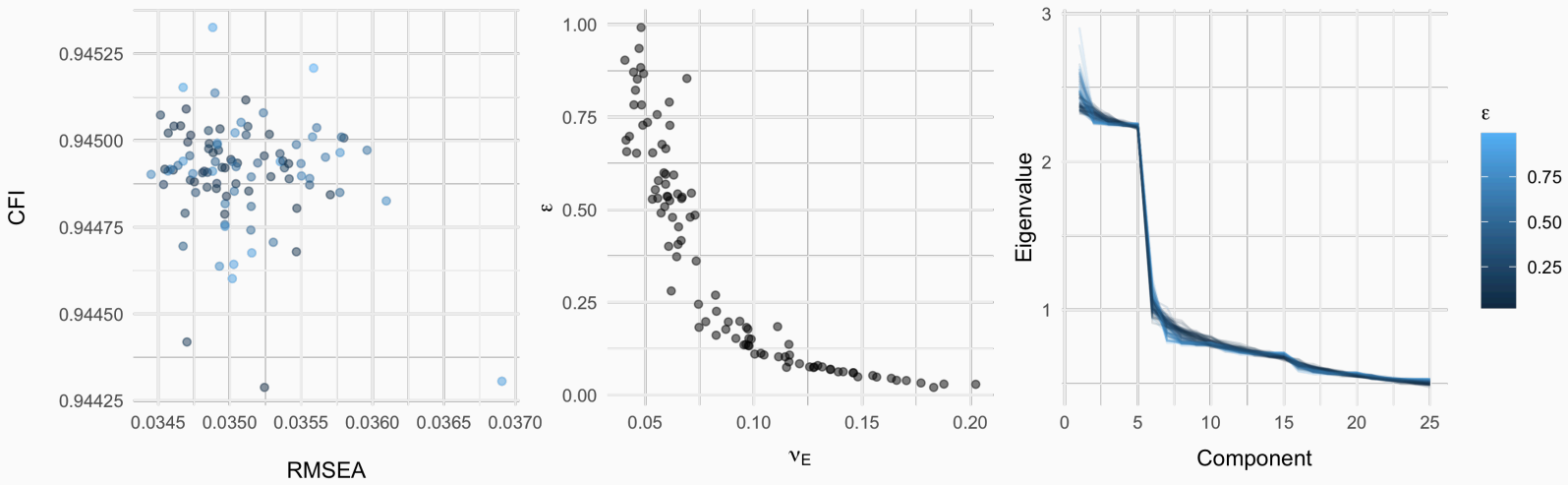
$$\text{CRMR} = \sqrt{\frac{1}{p(p-1)/2} \sum_{i < j} \left(\sigma_{ij} - \omega_{ij} \right)^2} \tag{14}$$

- p is the number of observed variables.
- σ_{ij} and ω_{ij} are the i, jth elements of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$, respectively.

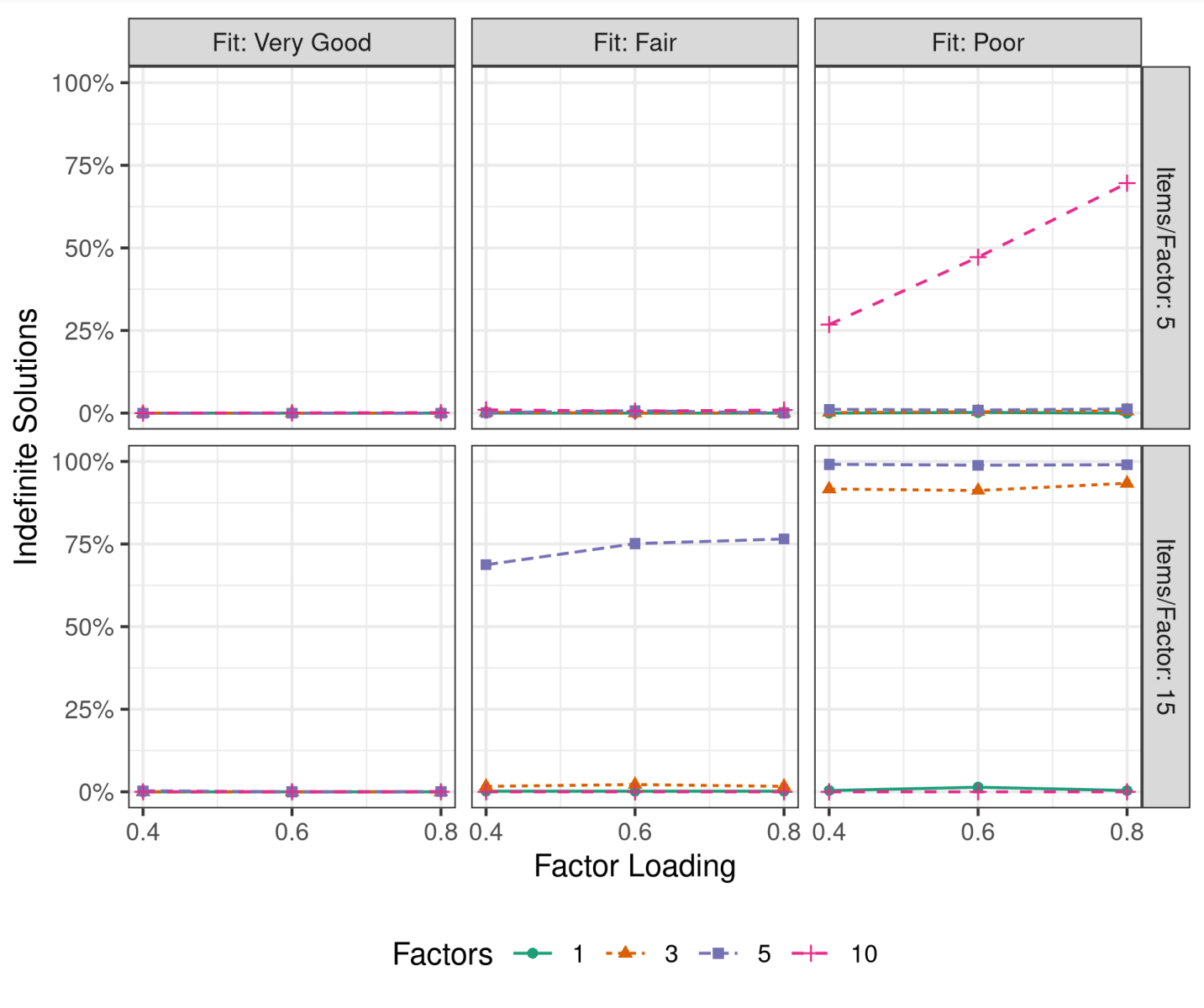
Local Optima

Results for 100 random starts for a five-factor, 25 item model

Target RMSEA of 0.05; Target CFI of 0.95; No constraints on minor factor loadings

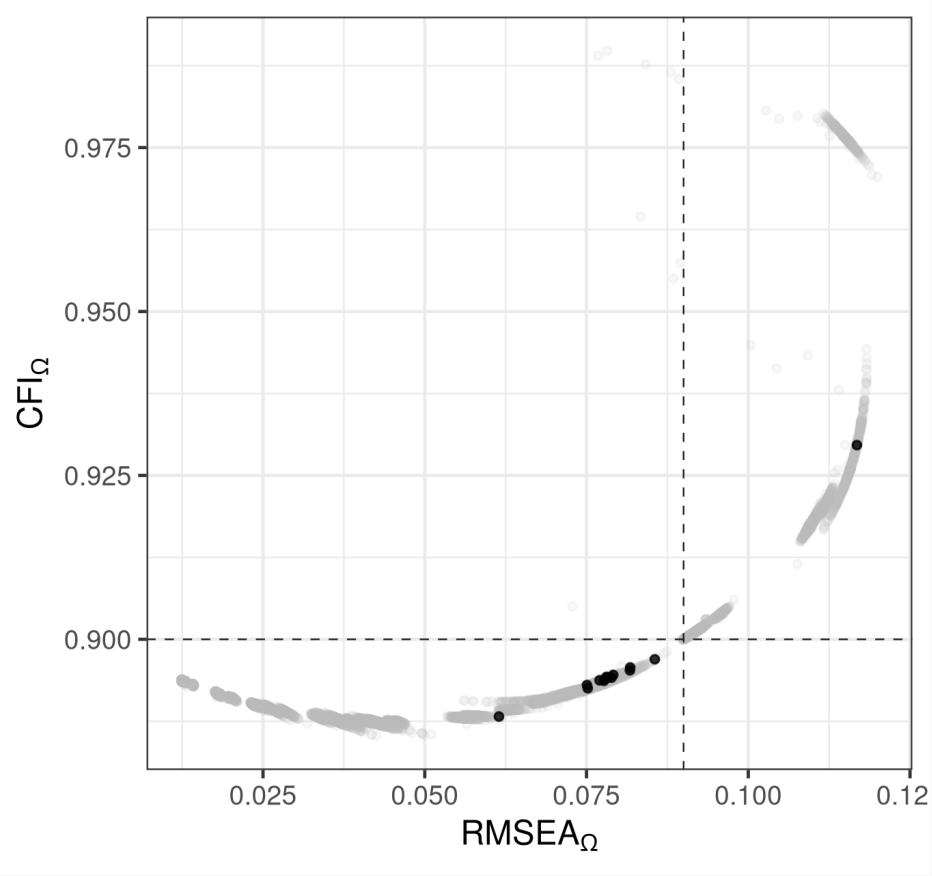


Indefinite Matrices



L-BFGS-B Non-convergence

- Non-convergence only occurred when the $\mathrm{TKL}_{\mathrm{RMSEA/CFI}}$ method was used.
- Only 14 of 90,000 (<1%) of $\mathrm{TKL}_{\mathrm{RMSEA/CFI}}$ cases failed to converge using L-BFGS-B.



Major Minor Factors