Factor Loading Recovery for Smoothed Tetrachoric Correlation Matrices

Justin D. Kracht

September 2020

Introduction

Overview

Tetrachoric correlation matrices are often used when conducting exploratory factor analysis on data sets with dichotomous items, but these matrices are sometimes indefinite (problematic for reasons I will discuss later)

Matrix smoothing algorithms produce a proper “smoothed” matrix from and indefinite matrix

1. Are smoothed matrices better approximations of their corresponding population correlation matrices than indefinite tetrachoric correlation matrices?
2. When used in factor analysis, do smoothed correlation matrices lead to better factor loading estimates than indefinite tetrachoric correlation matrices?
3. Do three commonly-used smoothing algorithms differ with respect to Questions (1) and (2)?
   - Higham (2002)
   - Bentler-Yuan (2011)
   - Knol-Berger (1991)

Previous Work

- Knol & Berger (1991) found no significant differences between factor solutions from smoothed and unsmoothed (indefinite) tetrachoric correlation matrices
  - Very small study; 10 indefinite correlation matrices with 250 subjects and 15 items
- Debelak & Tran (2013) and Debelak & Tran (2016) investigated whether applying matrix smoothing to indefinite tetrachoric/polychoric correlation matrices improved dimensionality estimation using parallel analysis
  - Smoothing improved dimensionality recovery (best results for Bentler-Yuan)
  - Differences were small

Previous Work

- Kracht and Waller (under review) replicated Debelak & Tran (2013) and extended their design
  - Only analyzed indefinite tetrachoric correlation matrices (focused on relative algorithm performance)
  - 1, 3, 5, or 10 major factors
  - Wider range of model error conditions and item characteristics
  - Bentler-Yuan algorithm led to slightly better results that the other methods, but differences were very small, however...
  - Led to somewhat better population correlation matrix recovery than the other methods

Background

By definition, a proper correlation matrix, $\mathbf{R}_{p \times p} = \{r_{ij}\}$, satisfies:

- $r_{ij} = r_{ji}$ (symmetry)
- $\mathrm{diag}(\mathbf{R}) = \mathbf{I}$ (unit diagonal)
- $r_{ij} \in [-1, 1]$ (elements bounded by -1 and 1)
- $\mathbf{R} \succcurlyeq 0$ (positive semidefinite)

Let the eigendecomposition of $\mathbf{R}$ be denoted as

$$\mathbf{R} = \mathbf{V}\Lambda\mathbf{V}'$$

where Λ denotes the diagonal matrix of ordered eigenvalues such that $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ and $\sum \lambda_i = p$.

- Positive definite ($\mathbf{R} \succ 0$): $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p > 0$
- Positive semidefinite ($\mathbf{R} \succcurlyeq 0$): $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p \geq 0$
- Indefinite: $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p < 0$

Spot the impostor:

$$\mathbf{R}_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad \mathbf{R}_2 = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix} \quad \mathbf{R}_3 = \begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Indefinite Correlation Matrices

$$\mathbf{R}_3 = \begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$\lambda = [2, 2, -1]$$

- Item 1 and Item 2 are correlated -1
- Item 1 and Item 3 are correlated 1
- But... Item 2 and Item 3 are correlated 1?

A Geometric Perspective

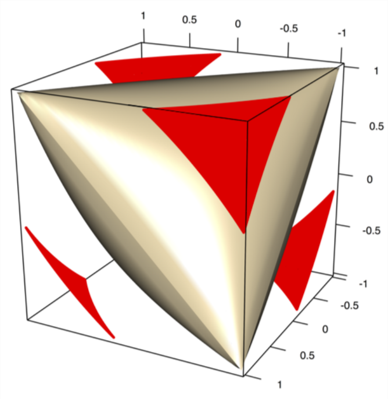


Figure 1: The elliptical tetrahedron representing the space of all PSD 3×3 correlation matrices. The three axes represent the off-diagonal elements r_{12}, r_{13}, and r_{23}. The red patches contain all indefinite 3×3 correlation matrices with a minimum eigenvalue $\lambda_{\text{min}} = -0.5$.

When do Indefinite Correlation Matrices Occur?

Indefinite correlation matrices will never occur when calculating Pearson correlation matrices from complete data.

They can occur when forming correlation matrices:

- Using pairwise deletion with missing data
- From correlations calculated using different data sets (i.e., composite correlation matrices)
- From tetrachoric (polychoric) correlations

The Problem with Indefinite Correlation Matrices

$\mathbf{R}_{\text{tet}}$: The tetrachoric correlation matrix

$\mathbf{R}_{\text{Pop}}$: The population correlation matrix estimated by $\mathbf{R}_{\text{tet}}$

Problems:

- An indefinite $\mathbf{R}_{\text{tet}}$ is not in the set of possible $\mathbf{R}_{\text{Pop}}$ matrices
- Some multivariate analysis procedures require PSD correlation matrices (i.e., maximum likelihood factor analysis)
- Can lead to nonsensical interpretations (e.g., negative component variance in PCA)

A matrix smoothing algorithm is a procedure that modifies an indefinite correlation matrix to produce a correlation matrix that is at least PSD.

- The Higham Alternating Projections algorithm (APA; Higham, 2002)
- The Bentler-Yuan algorithm (BY; Bentler & Yuan, 2011)
- The Knol-Berger algorithm (KB; Knol & Berger, 1991)

Intuition: Find the closest PSD correlation matrix ($\mathbf{R}_{\mathrm{APA}}$) to a given indefinite correlation matrix ($\mathbf{R}_{\_}$) by iteratively projecting between two sets:

- $\mathcal{S}$: The set containing all possible $p \times p$ symmetric matrices that are PSD
- $\mathcal{U}$: The set containing all possible $p \times p$ symmetric matrices that have a unit diagonal

The Higham Alternating Projections Algorithm (2002)

For symmetric matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$, define two projection functions:

- $P_S(\mathbf{A}) = \mathbf{V}\mathrm{diag}(\max(\lambda_i, 0))\mathbf{V}'$: Project $\mathbf{A}$ onto $\mathcal{S}$ by replacing all negative eigenvalues with zero in the eigendecomposition.
- $P_U(\mathbf{A})$: Project $\mathbf{A}$ onto $\mathcal{U}$ by replacing the diagonal elements of $\mathbf{A}$ with ones.

The Higham Alternating Projections Algorithm (2002)

Initialize $\mathbf{A}_0$ as the indefinite correlation matrix $\mathbf{R}_{\_}$. Repeat the operation

$$\mathbf{A}_{k+1} = P_U(P_S(\mathbf{A}_k))$$

until convergence occurs or the maximum number of iterations is exceeded.

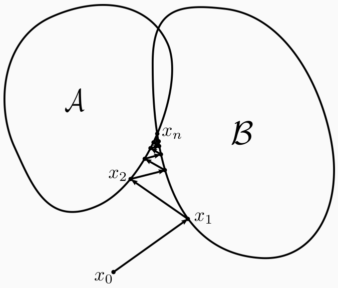


Figure 2: Simplified illustration of the method of alternating projections.

Intuition: Shrink the correlations involving variables with minimum trace factor analysis (MTFA; Jamshidian & Bentler, 1998) estimated communalities ≥ 1.

The Bentler-Yuan Algorithm (2011)

$$\mathbf{R}_{\text{BY}} = \Delta \mathbf{R}_0 \Delta + \mathbf{I}$$

$\mathbf{R}_0 = \mathbf{R}\_ - \mathbf{I}$
Δ^2 is a diagonal matrix with elements δ_i^2,

$$\delta_i^2 = \begin{cases} 1 & \text{if } h_i < 1 \\ k/h_i & \text{if } h_i \geq 1. \end{cases}$$

$k \in (0, 1)$ is a constant chosen by the user
h_i is the MTFA communality estimate for the ith item

The Knol-Berger Algorithm (1991)

Intuition: Replace all negative eigenvalues with a small positive constant in the eigenvalue decomposition and then scale the result to a correlation matrix.

The Knol-Berger Algorithm (1991)

$$\mathbf{R}_{\_} = \mathbf{V}\Lambda\mathbf{V}'$$

$$\Lambda_{+} = \mathsf{diag}[\mathsf{max}(\lambda_i, 0)],\ i \in \{1, \ldots, p\}$$

$$\mathbf{R}_{\mathrm{KB}} = [\mathrm{dg}(\mathbf{V}\Lambda_{+}\mathbf{V}')]^{-1/2}\mathbf{V}\Lambda_{+}\mathbf{V}'[\mathrm{dg}(\mathbf{V}\Lambda_{+}\mathbf{V}')]^{-1/2}$$

Example: Matrix Smoothing Algorithms

$$\mathbf{R}_{\_} = \begin{bmatrix} 1 & 0.48 & 0.64 & 0.48 & 0.65 & 0.83 \\ 0.48 & 1 & 0.52 & 0.23 & 0.68 & 0.75 \\ 0.64 & 0.52 & 1 & 0.60 & 0.58 & 0.74 \\ 0.48 & 0.23 & 0.60 & 1 & 0.74 & 0.80 \\ 0.65 & 0.68 & 0.58 & 0.74 & 1 & 0.80 \\ 0.83 & 0.75 & 0.74 & 0.80 & 0.80 & 1 \end{bmatrix}$$

Eigenvalues: (4.21, 0.77, 0.52, 0.38, 0.18, -0.06)

Communalities: (1.029, 1.122, 0.557, 1.299, 0.823, 0.997)

Variables 1, 2, and 4 have estimated communalities > 1.

Example: Matrix Smoothing Algorithms

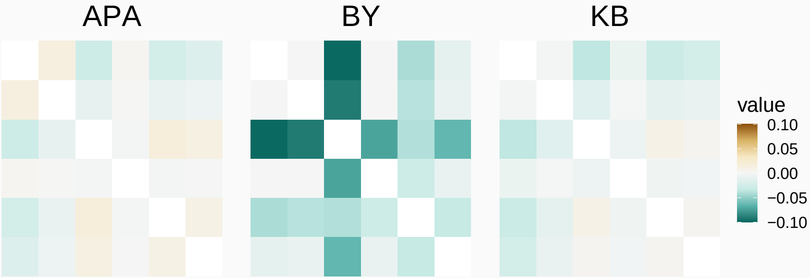


Figure 3: Differences between the elements of the $\mathbf{R}_{\mathrm{Sm}}$ and $\mathbf{R}_{-}$ matrices for the Higham, Bentler-Yuan, and Knol-Berger algorithms.

$$\mathbf{P} = \mathbf{F}\Phi\mathbf{F}' + \Theta^2 \tag{1}$$

- $\mathbf{P}$: $p \times p$ population correlation matrix
- $\mathbf{F}$: $p \times m$ factor loading matrix
- Φ: $m \times m$ factor correlation matrix
- Θ^2: $p \times p$ matrix of unique item variances

Common Factor Model with Model Approximation Error

Tucker et al. (1969)

$$\mathbf{P} = \mathbf{F}\Phi\mathbf{F}' + \Theta^2 + \mathbf{W}\mathbf{W}' \tag{2}$$

- $\mathbf{P}$: $p \times p$ population correlation matrix
- $\mathbf{F}$: $p \times m$ factor loading matrix
- Φ: $m \times m$ factor correlation matrix
- Θ^2: $p \times p$ matrix of unique item variances
- $\mathbf{W}$: $p \times q$ minor factor loading matrix for the $q \gg m$ minor common factors

Methods

Simulation Conditions

- Major common factors: $m \in \{1, 3, 5, 10\}$
- Items per factor: $p/m \in \{5, 10\}$
- Subjects per item: $N/p \in \{5, 10, 15\}$
- Factor Loading: Loading $\in \{0.4, 0.6, 0.8\}$
    - Orthogonal models with simple structure
- Model Error: $\upsilon_E \in \{0.0, 0.1, 0.3\}$
    - Proportion of uniqueness variance apportioned to minor common factors
- Classical item difficulties ranged from 0.15 to 0.85 at equal intervals

Fully-crossed design with 216 unique conditions

Simulation Procedure

For each of the 216 unique conditions, conduct 1,000 replications of the following steps:

1. Generate binary response data using Equation (1)
2. Compute the tetrachoric correlation matrix
3. If the matrix is PSD, next; Else, smooth using:
    - Higham (2002)
    - Bentler-Yuan (2011)
    - Knol-Berger (1991)
4. For each of the three smoothed correlation matrices and the unsmoothed matrix, estimate factor loadings using:
    - Principal Axes factor extraction (PA)
    - Ordinary Least Squares (OLS)
    - Maximum Likelihood (ML)

$\mathrm{R}_{\mathbf{Pop}}$ Recovery Criterion

Given two $p \times p$ symmetric matrices, $\mathbf{A} = \{a_{ij}\}$ and $\mathbf{B} = \{b_{ij}\}$,

$$\mathrm{D_s}(\mathbf{A}, \mathbf{B}) = \sqrt{\sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \frac{\left(a_{ij} - b_{ij}\right)^2}{p(p-1)/2}}.$$

- $\mathbf{R}_{\mathrm{Sm}} \in \{\mathbf{R}_{-}, \mathbf{R}_{\mathrm{APA}}, \mathbf{R}_{\mathrm{BY}}, \mathbf{R}_{\mathrm{KB}}\}$
- $\mathbf{R}_{\mathrm{Pop}} = \mathbf{F}\Phi\mathbf{F}' + \Theta^2 + \mathbf{W}\mathbf{W}'$

Evaluate recovery of $\mathbf{R}_{\mathrm{Pop}}$ using $\mathrm{D_s}(\mathbf{R}_{\mathrm{Sm}}, \mathbf{R}_{\mathrm{Pop}})$

Lower is better

$\mathbf{F}$ Recovery Criterion

Evaluate how well the factor loading matrix, $\mathbf{F}$, was recovered using:

$$\mathrm{RMSE}(\mathbf{F}, \hat{\mathbf{F}}) = \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{m} \frac{\left(f_{ij} - \hat{f}_{ij}\right)^2}{pm}}$$

Lower is better

Results

Indefinite Matrix Frequency

124,346 (57.6%) of 216,000 tetrachoric correlation matrices were indefinite

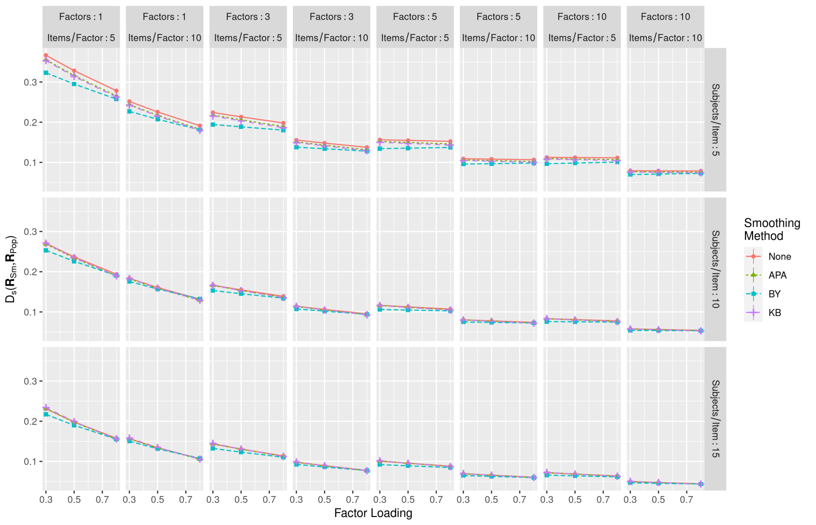
Indefinite matrices were most common in conditions with:

- Many factors/items per factor (i.e., total number of items)
- Few subjects per items
- Large factor loadings

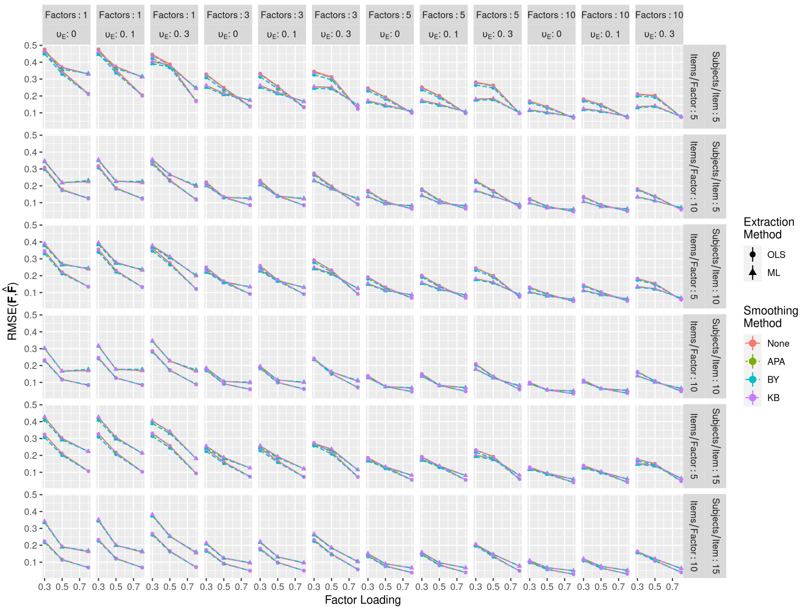
		Factors			
N/p	Loading	1	3	5	10
5	0.3	46.2	98.9	100.0	100.0
5	0.5	52.8	99.7	100.0	100.0
5	0.8	56.4	100.0	100.0	100.0
10	0.3	8.1	22.9	33.0	43.4
10	0.5	16.5	47.7	66.1	85.7
10	0.8	49.1	99.3	100.0	100.0
15	0.3	1.0	0.6	0.4	0.5
15	0.5	2.6	3.7	6.4	16.2
15	0.8	32.8	86.0	96.4	100.0

Note: Percent of indefinite matrices conditioned on number of subjects per item (N/p), factor loading, and number of factors.

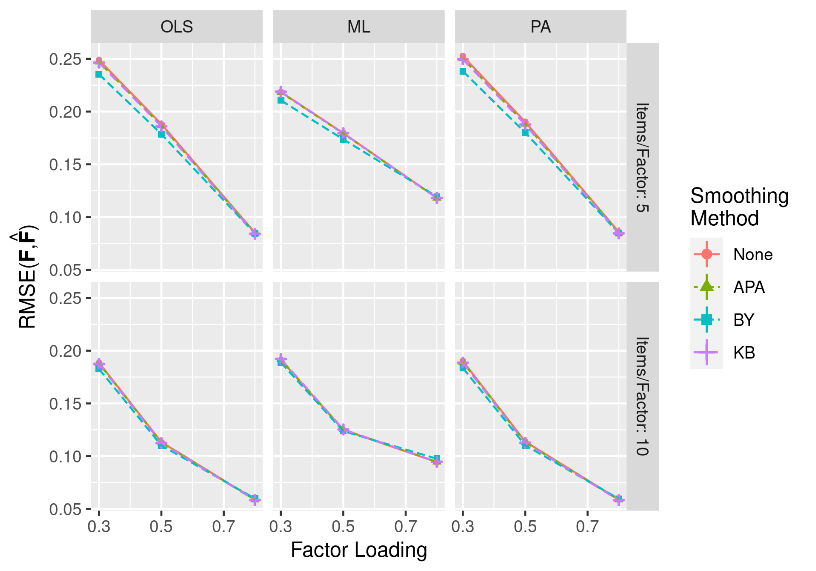
Population Correlation Matrix ($\mathbf{R}_{\mathbf{Pop}}$) Recovery



Factor Loading Recovery



Factor Loading Recovery



Discussion

Summary: Population Correlation Matrix ($\mathbf{R}_{\mathbf{Pop}}$) Recovery

- $\mathbf{R}_{\text{Pop}}$ recovery was better in conditions with:
  - High factor loadings
  - Many major factors
  - Many items per factor
  - Many subjects per item
- The Bentler-Yuan (2011) algorithm led to slightly better recovery in conditions with:
  - Low factor loadings
  - Few major factors
  - Few items per factor
  - Few subjects per item

Summary: Factor Loading Recovery

- Factor loading recovery was better in conditions with:
  - High factor loadings
  - Many items per factor
  - Small amounts of model approximation error
  - Under these conditions, OLS and PA led to better results than ML
- Bentler-Yuan (2011) led to slightly better results in conditions with:
  - Low factor loadings
  - Few items per factor
  - ML factor extraction

Limitations & Future Directions

- Only orthogonal models with fixed factor loadings
- Investigated only indefinite tetrachoric correlation matrices
    - Polychoric correlation matrices
    - Composite correlation matrices
    - Correlation matrices calculated from missing data
- Investigate methods that avoid the problem
    - Remove problematic items
    - Full-information factor analysis
    - Bayesian/penalized tetrachoric estimation

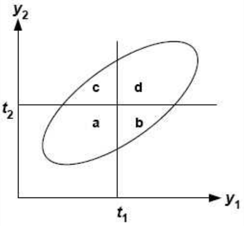
Simulation Code

https://z.umn.edu/matrix_smoothing

https://github.umn.edu/krach018/masters_thesis

Backup Slides

Let y_1^* and y_2^* denote binary variables obtained by dichotomizing continuous, normally-distributed variables y_1 and y_2 (with correlation r) using thresholds t_1 and t_2, respectively.



Objective: Estimate r

Tetrachoric Correlation

1. $\hat{t_i} = \Phi^{-1}(p_i - 1), i \in \{1, 2\}$
   - p_i: Proportion of correct responses (i.e., $y_i^* = 1$) for y_i^*
   - $\Phi^{-1}(*)$: Inverse standard normal cumulative distribution function
2. Solve for r
   - p_{11}: proportion of correct responses for both y_1^* and y_2^*

$$
\begin{aligned}
L\left(\hat{t}_1, \hat{t}_2, r\right) &= \frac{1}{2\pi\sqrt{1-r^2}} \int_{\hat{t}_2}^{\infty} \int_{\hat{t}_1}^{\infty} e^{\left[-\frac{y_1^{*2} + y_2^{*2} - 2ry_1^* y_2^*}{2\left(1-r^2\right)}\right]} dy_1^* dy_2^* \\
&= p_{11}
\end{aligned}
$$

Algorithm 1: For an indefinite correlation matrix $\mathbf{R}\_$, find the nearest PSD correlation matrix

Initialize $\mathbf{S}_0 = 0$; $\mathbf{Y}_0 = \mathbf{R}\_$

for $k = 1, 2, \ldots$ **do**

$$\mathbf{Z}_k = \mathbf{Y}_{k-1} - \mathbf{S}_{k-1}$$
$$\mathbf{X}_k = P_S(\mathbf{Z}_k)$$
$$\mathbf{S}_k = \mathbf{X}_k - \mathbf{Z}_k$$
$$\mathbf{Y}_k = P_U(\mathbf{X}_k)$$

end

The algorithm continues until convergence occurs or the maximum number of iterations is exceeded. If the algorithm converges, $\mathbf{R}_{\text{APA}} = \mathbf{Y}_k$.

Minimum Trace Factor Analysis

Given a population covariance (correlation) matrix, Σ, minimum trace factor analysis seeks to find the diagonal matrix of unique variances, $\Psi = \mathrm{diag}(\Psi_{11}, \dots, \Psi_{pp})$ to solve the optimization problem:

$$\operatorname*{Min}_{\Psi} \mathrm{tr}(\Sigma - \Psi) \text{ subject to } \Sigma - \Psi \succeq 0 \tag{3}$$

The greatest lower bound of reliability is then defined as:

$$\rho := 1 - \frac{\mathrm{tr}\,\bar{\Psi}}{1_p' \Sigma 1_p}$$

where $\bar{\Psi} = \bar{\Psi}(\Sigma)$ is the optimal solution of Equation (3) (Shapiro & ten Berge, 2002).

$\mathbf{H}_0 = \mathrm{diag}(h_1, \dots, h_p)$
- h_i is the estimated communality for Item i

Algorithm 2: Extract principal axes factor solution

Initialize $\mathbf{R}_0^* = \mathbf{R} - \mathbf{I} + \mathbf{H}_0$

for $k = 1, 2, \dots$ **do**

$\quad \mathbf{R}_{k-1}^* = \mathbf{V}_{k-1}\Lambda_{k-1}\mathbf{V}_{k-1}'$

$\quad \mathbf{R}_k^* = \mathbf{R}_{k-1}^* - \mathbf{I} + \Lambda_{k-1}$

$\quad \epsilon = |\,\mathrm{diag}\,\Lambda_k - \mathrm{diag}\,\Lambda_{k-1}|$

end

Stop when $\epsilon \leq \delta$.

$\hat{\mathbf{P}}$: Implied correlation matrix from the estimated factor model

$\mathbf{R}$: Observed correlation matrix

Minimize the discrepancy function:

$$F_{OLS}(\mathbf{R}, \hat{\mathbf{P}}) = \frac{1}{2} \operatorname{tr} \left[(\mathbf{R} - \hat{\mathbf{P}})^2 \right]$$

Maximum Likelihood Factor Extraction

Minimize the discrepancy function:

$$F_{ML}(\mathbf{R}, \hat{\mathbf{P}}) = \log|\hat{\mathbf{P}}| - \log|\mathbf{R}| + \operatorname{tr}\left(\mathbf{S}\hat{\mathbf{P}}^{-1}\right) - p$$

Bentler, P. M., & Yuan, K.-H. (2011). Positive Definiteness via Off-diagonal Scaling of a Symmetric Indefinite Matrix. *Psychometrika*, *76*(1), 119–123. https://doi.org/10.1007/s11336-010-9191-3

Debelak, R., & Tran, U. S. (2016). Comparing the effects of different smoothing algorithms on the assessment of dimensionality of ordered categorical items with parallel analysis. *PLOS ONE*, *11*(2), 1–18. https://doi.org/10.1371/journal.pone.0148143

Debelak, R., & Tran, U. S. (2013). Principal Component Analysis of Smoothed Tetrachoric Correlation Matrices as a Measure of Dimensionality. *Educational and Psychological Measurement*, *73*(1), 63–77. https://doi.org/10.1177/0013164412457366

Higham, N. J. (2002). Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis*, *22*(3), 329–343. https://doi.org/10.1093/imanum/22.3.329

Jamshidian, M., & Bentler, P. M. (1998). A quasi-newton method for minimum trace factor analysis. *Journal of Statistical Computation and Simulation*, *62*(1-2), 73–89. https://doi.org/10.1080/00949659808811925

Knol, D. L., & Berger, M. P. F. (1991). Empirical Comparison Between Factor Analysis and Multidimensional Item Response Models. *Multivariate Behavioral Research*, *26*(3), 457–477. https://doi.org/10.1207/s15327906mbr2603_5

Shapiro, A., & ten Berge, J. M. F. (2002). Statistical inference of minimum rank factor analysis. *Psychometrika*, *67*(1), 79–94. https://doi.org/10.1007/BF02294710

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, *34*(4), 421–459. https://doi.org/10.1007/BF02290601