Factor Loading Recovery for Smoothed Tetrachoric Correlation Matrices

Justin D. Kracht

Fall 2020

Introduction

Item Factor Analysis

We often want to conduct exploratory factor analysis on binary response data

- The assumption of continuous outcomes required by the common linear factor model is violated when data are binary
- Tetrachoric correlation matrices (Brown & Benedetti, 1977; Divgi, 1979) are often used to estimate the correlations between the normally-distributed, continuous latent variables often assumed to underlie observed binary data

Indefinite Tetrachoric Correlation Matrices

- Tetrachoric correlation matrices can be indefinite, particularly when computed from data sets with:
  - Few subjects
  - Many items
  - Extreme items (high factor loadings, extreme item difficulties)

A correlation matrix, $\mathbf{R}_{p \times p}$ with elements $r_{ij} = r_{ji}$, $i, j \in \{1, \dots, p\}$, is a symmetric matrix that:

1. Has unit diagonal
2. Has all $|r_{ij}| \leq 1$
3. Is positive semidefinite (PSD), having eigenvalues $\lambda_i \geq 0$ $\forall\ i \in \{1, \dots, p\}$

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 do not work with indefinite correlation matrices (i.e., maximum likelihood factor analysis)
- Can lead to nonsensical interpretations (e.g., negative component variance in PCA)

Matrix Smoothing Algorithms

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)

The Higham Alternating Projections Algorithm (2002)

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

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.

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.

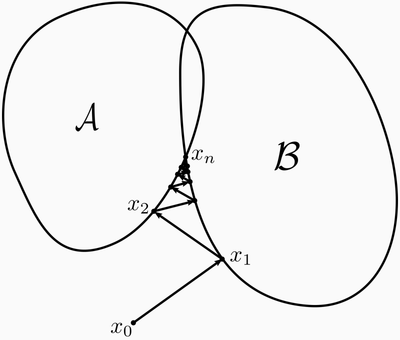


Figure 1: Simplified illustration of the method of alternating projections.

The Bentler-Yuan Algorithm (2011)

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

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

$\mathbf{R}_0 = \mathbf{R}_{\_} - \mathbf{I}$
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.

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

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

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

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

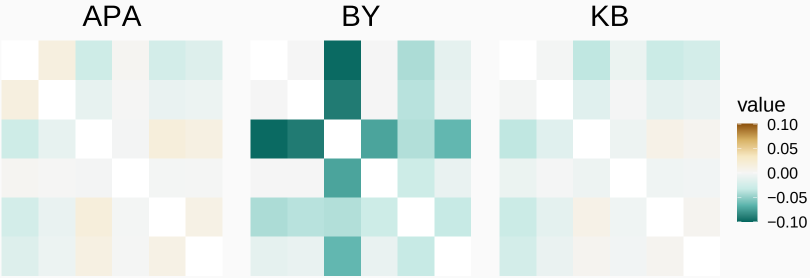


Figure 2: Differences between the elements of the $\mathbf{R}_{\mathrm{Sm}}$ and $\mathbf{R}_{\_}$ matrices for the Higham, Bentler-Yuan, and Knol-Berger algorithms.

Common Factor Model

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

- $\mathbf{P}$: $p \times p$ 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 $q \gg p$ 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 $\in \{0.4, 0.6, 0.8\}$
- Model Error: $\upsilon_E \in \{0.0, 0.1, 0.3\}$
    - Proportion of uniqueness variance apportioned to minor common factors

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 $\mathbf{B} = \{b_{ij}\}$,

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

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

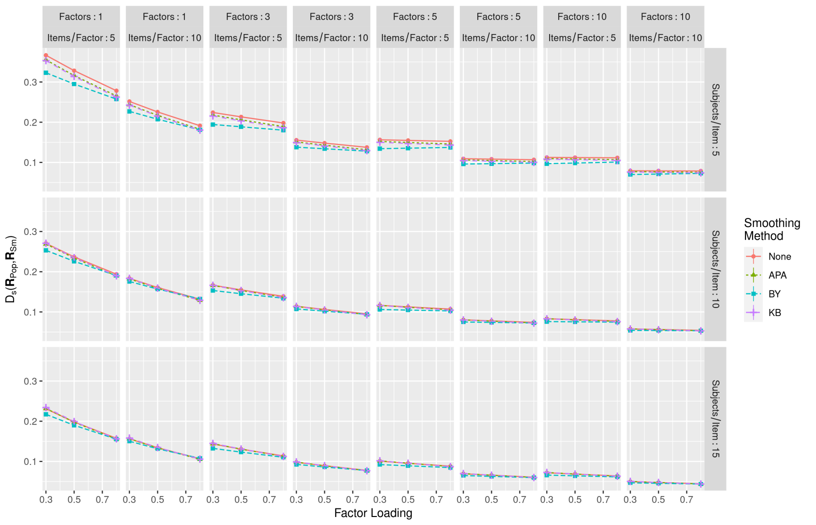
Evaluate recovery of $\mathbf{R}_{\mathrm{Pop}}$ using $\mathrm{D_s}(\mathbf{R}_{\mathrm{Pop}}, \mathbf{R}_{\mathrm{Sm}})$

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

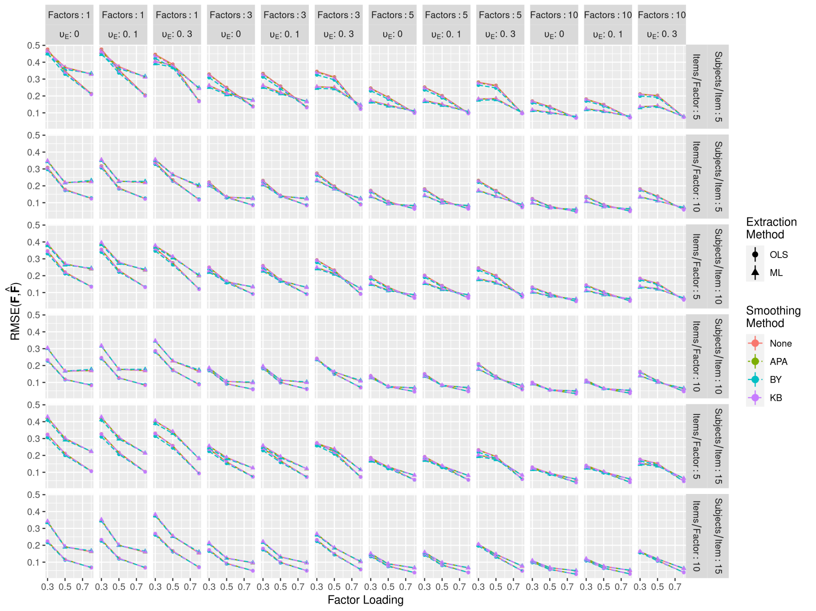
$$\mathrm{RMSE}(\mathbf{F}, \hat{\mathbf{F}}) = \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{m} \frac{\left(f_{ij} - \hat{f}_{ij}\right)^2}{pm}}$$

Results

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

Limitations

Future Directions

Appendix

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, $\mathbf{R}_{\text{APA}} = \mathbf{Y}_k$.

Bentler, P., & 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

Brown, M. B., & Benedetti, J. K. (1977). On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika*, *42*(3), 347–355. https://doi.org/10.1007/BF02293655

Divgi, D. R. (1979). Calculation of the tetrachoric correlation coefficient. *Psychometrika*, *44*(2), 169–172. https://doi.org/10.1007/BF02293968

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. (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

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