Factor Loading Recovery for Smoothed Non-positive Definite Correlation Matrices

Justin D. Kracht[1]

[1] University of Minnesota

Author Note

Enter author note here.

Correspondence concerning this article should be addressed to Justin D. Kracht, Department of Psychology, University of Minnesota, N218 Elliott Hall 75 East River Road, Minneapolis, MN 55455. E-mail: krach018@umn.edu

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words “**here we show**” or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: keywords

Word count: X

Factor Loading Recovery for Smoothed Non-positive Definite Correlation Matrices

Tetrachoric correlation matrices (Olsson, 1979) are often recommended for use in item factor analysis (Wirth & Edwards, 2007). However, tetrachoric correlation matrices are frequently *non-positive definite* (NPD), having one or more negative eigenvalues.

Matrix Smoothing Algorithms

Higham Alternating Projections Algorithm (APA; 2002).

Bentler-Yuan Algorithm (BY; 2011).

Knol-Berger Algorithm (KB; 1991).

Factor Estimation Methods

Principal Axes Factor Analysis.

Least-Squares Factor Analysis.

Maximum-Likelihood Factor Analysis.

Methods

I designed and ran a simulation study to evaluate four approaches to dealing with NPD tetrachoric correlation matrices in the context of exploratory factor analysis. Namely, I conducted exploratory factor analyses on tetrachoric correlation matrices smoothed using the Higham (2002), the Bentler-Yuan (2011), Knol-Berger (1991) algorithms and on unsmoothed (NPD and PSD) tetrachoric correlation matrices. I designed the simulation study with two primary purposes in mind. First, I wanted to know which smoothing method (Higham, Bentler-Yuan, Knol-Berger, or None) produced (possibly) smoothed correlation matrices ($\mathbf{R}_{\text{Sm}}$) that most closely approximated the corresponding population correlation matrices. Second, I wanted to know which smoothing method produced correlation matrices that led to the best estimates of the population factor loading matrix when used in exploratory factor analyses. With these two purposes in mind, I conducted our simulation study as follows.

In the first step of the simulation, I generated random sets of binary data randomly generated from a variety of orthogonal factor models. The factor models had varying numbers of major common factors, Factors $\in \{1, 3, 5, 10\}$. Following the procedure of (Tucker, Koopman, & Linn, 1969), I also incorporated the effects of model approximation error into the data by including 150 minor common factors in each population model. In total, these 150 minor common factors accounted for 0%, 10%, or 30% (Error $\in \{0, .1, .3\}$) of the uniqueness variance of the error-free model (i.e., the model with only the major common factors). According to Briggs and MacCallum (2003), these conditions represent models with perfect, good, or moderate model fit. Including these three levels of model approximation error in the simulation ensured that both ideal (Error $= 0$) and the more empirically-plausible levels of model approximation error (Error $\in \{.1, .3\}$) were considered in this study.

In addition to systematically varying the number of major factors and the proportion of variance accounted for by model approximation error, I also systematically varied the number of factor indicators (i.e., items loading on each factor) , Items/Factor $\in \{5, 10\}$, and the number of subjects per item, Subjects/Item $\in \{5, 10, 15\}$. The total numbers of items and sample sizes for each factor number condition can be found in Table 1. Each item loaded on only one factor and item factor loadings were uniformly fixed at one of three levels, Loading $\in \{.3, .5, .8\}$. Though “rules-of-thumb” for factor loadings vary, Hair, Andersen, Tatham, and Black (1998, p. 111) note that “factor loadings greater than ±0.3 are considered to meet the minimal level … if the loadings are ±0.5 or greater, they are considered practically significant." Factor loadings of ±0.8 are considered to be high (e.g., MacCallum, Widaman, Preacher, & Hong, 2001). Thus, the three factor loadings investigated in this study were chosen to represent low, moderate, and high levels of factor saliency.

The combinations of the independent variables specified above resulted in a fully-crossed design with
4 (Factors) × 3 (Error) × 2 (Items/Factor) × 3 (Subjects/Item) × 3 (Loading) = 216 unique conditions. For each of these conditions, I used the `simFA` function in the R `fungible` library (R Core Team, 2019, @waller2019a) to generate 1,000 random sets of data in accordance with the factor model corresponding to that condition. To obtain binary responses from continuous observed scores, items were assigned classical item difficulties (p; i.e., the expected proportion of correct responses, Crocker & Algina, 1986) at equal intervals between 0.15 and 0.85. For example, items in a five-item data set were assigned classical item difficulties of .150, .325, .500, .675, and .850. The classical item difficulties were used to obtain threshold values, t, such that $P(X > t) = p$ where $X \sim N(0, 1)$. I then used the thresholds to dichotomize the continuous observed scores and obtain simulated binary response data. If a data set had any homogeneous item response vectors (i.e., had one or more items with zero variance), the data set was discarded and a new sample of data was generated until all items had non-homogeneous response vectors. This procedure was necessary to calculate tetrachoric correlation matrices in the next step of the simulation.

In the second step of the simulation procedure, I calculated a tetrachoric correlation matrix for each simulated binary data set. Tetrachoric correlation matrices were calculated using the `tetcor` function in the R `fungible` library (Waller, 2019), which computes maximum likelihood tetrachoric correlation coefficients corrected for bias using the method of Brown and Benedetti (1977). If a tetrachoric correlation matrix was NPD, the Higham (2002), Benler-Yuan (2011), and Knol-Berger (1991) matrix smoothing algorithms were applied to the NPD tetrachoric correlation matrix to produce three smoothed, PSD correlation matrices.

In the third and final step of the simulation procedure, I applied three exploratory factor extraction algorithms (principal axes [PA], ordinary least squares [OLS], and maximum likelihood [ML] factor analysis) to each of the PSD and NPD tetrachoric correlation matrices and the smoothed correlation matrices. Each of the factor solutions were then rotated using a quartimin rotation and aligned to match the corresponding population factor loading matrix such that the least squares discrepency between the matrices was minimized. The alignment step ensured that the elements of each estimated factor loading matrix were matched (in order and sign) to the elements of the corresponding population factor loading matrix.

Results

Discussion

References

Briggs, N. E., & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research*, *38*(1), 25–56.

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* ERIC.

Hair Jr., J. F., Anderson, R. E., Tatham, R. L., & William, C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behav. Res.*, *36*(4), 611–637.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460.

R Core Team. (2019). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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.

Waller, N. G. (2019). *Fungible: Psychometric functions from the Waller lab.*

Wirth, R., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58.

Table 1

Number of items and subjects resulting from each combination of number of factors (Factors), number of items per factor (Items/Factor), and subjects per item (Subjects/Item).

Factors	Items/Factor	Subjects/Item	Items	Sample Size
1	5	5	5	25
3	5	5	15	75
5	5	5	25	125
10	5	5	50	250
15	5	5	75	375
1	10	5	10	50
3	10	5	30	150
5	10	5	50	250
10	10	5	100	500
15	10	5	150	750
1	5	10	5	50
3	5	10	15	150
5	5	10	25	250
10	5	10	50	500
15	5	10	75	750
1	10	10	10	100
3	10	10	30	300
5	10	10	50	500
10	10	10	100	1000
15	10	10	150	1500

1	5	15	5	75
3	5	15	15	225
5	5	15	25	375
10	5	15	50	750
15	5	15	75	1125
1	10	15	10	150
3	10	15	30	450
5	10	15	50	750
10	10	15	100	1500
15	10	15	150	2250