UNIVERSITAT ROVIRA I VIRGILI

Testing Sweet Smoothing algorithm via a simulation study

Urbano Lorenzo-Seva
Pere J. Ferrando

Tarragona 2020

Please reference this document as:

Lorenzo-Seva, U., & Ferrando, P.J. (2020). *Testing Sweet Smoothing algorithm via a simulation study. Technical Report*. Department of Psychology, Universitat Rovira i Virgili, Tarragona.

Document available at: http://psico.fcep.urv.cat/utilitats/factor/

Testing Sweet Smoothing algorithm via a simulation study

Least-squares exploratory factor analysis based on tetrachoric/polychoric correlations is a robust, defensible and widely used approach for performing item analysis, especially in the first stages of scale development. A relatively common problem in this scenario, however, is that the inter-item correlation matrix fails to be positive definite.

In the following simulation study, we compare four alternatives in order to convert a not positive definite correlation matrix to a positive definite correlation matrix. In the simulation we include Sweet Smoothing, that is a refinement of a previous methodology developed by Bentler & Yuan (2011) that we label Straight Smoothing.

Straight smoothing proposed by Bentler and Yuan

Bentler and Yuan (2011) observed that the strategies to solve that are not positive definite (NPD) correlation matrices typically impacted all the variables in the matrix. To prevent this from happening, they proposed focusing the smoothing procedure only on the problematic variables (i.e., the ones that would potentially produce a Heywood case in the factor solution). Their proposal is to extract all the possible factors in the common factor space and to check which variables have communalities larger than 1 in the factor solution. Once these variables have been detected, the correlation estimates to which they are related are decreased by a low value k, so the smoothed $\tilde{\mathbf{R}}$ matrix is positive definite (PD). The decreasing factor is arbitrary and depends on the correlation matrix at hand. In their numerical example, they used a value of k=.96 but any other value can be used as long as the information lost in the smoothed $\tilde{\mathbf{R}}$ matrix is minimum, and the matrix $\tilde{\mathbf{R}}$ is PD. Whenever we computed this approach, we implemented their method iteratively in order to find an optimal constant value k. In the first iteration, we multiplied the corresponding

values of **R** by a value of k=1-0.0010/√(N), and decreased k progressively with the value 0.0010/√(N), until the smoothed correlation matrix $\tilde{\mathbf{R}}$ was PD.

It must be noted that the information lost during the smoothing procedure only has an effect on a (possibly small) number of variables. In the most extreme situations, all the information in the smoothed variables could be lost, which would be equivalent to removing these variables from the analysis. Again, the amount of information lost in $\tilde{\mathbf{R}}$ can be quantified using expression

$$v = \frac{\sum_{i \neq j} r_{ij} - \sum_{i \neq j} \tilde{r}_{ij}}{\sum_{i \neq j} r_{ij}} \quad (1)$$

where r_{ij} are the off-diagonal elements of the correlation matrix that is NPD, and $\tilde{r}_{ij}$ are the off-diagonal elements of the smoothed correlation matrix that is PD. A value of v of one would mean that no information has survived the smoothing procedure, while a value close to zero would mean that the amount of information destroyed is minimum. In addition to v, the same information could be computed for each variable in the correlation matrix. This index v_j would be reporting the amount of information destroyed in each variable.

As Bentler and Yuan did not explicitly name their approach, here we refer to it as *straight smoothing*. This label refers to the fact that the method attacks all the possible annoying variables at once.

A new proposal: Sweet Smoothing

We propose a new approach that is essentially equivalent to that of Bentler and Yuan, but applied very carefully, so that the amount of lost information in $\tilde{\mathbf{R}}$ is minimal. A detailed description can be found in Lorenzo-Seva and Ferrando (2020). Our approach can be summarized with this iterative algorithm:

Step 1. Set the number of factors to be extracted $r = 1$.

Step 2. Extract r factors from **R**, and check for Heywood cases.

Step 3. If no Heywood cases are observed, increase r in 1 and go to step 2. Otherwise, go to next step.

Step 4. Set the correction value k=1 - 0.0001.

Step 5. Decrease the correlation values of the variables that showed communalities larger than 1 using the value k, in order to obtain the smoothed correlation matrix $\tilde{\mathbf{R}}$.

Step 6. Check if the matrix $\tilde{\mathbf{R}}$ is PD: in this case, the algorithm ends and matrix $\tilde{\mathbf{R}}$ is the smoothed correlation matrix that removes the minimum information. Otherwise, go to next step.

Step 7. Decrease k with the value .0001.

Step 8. If the value of k is lower than .5 and r is lower than the maximum possible number of factors in the common factor space, then it is considered that too much information is to be removed from the variables at hand: in this case, increase r in 1 and go to step 2. Otherwise, go to step 5.

While all the information in some of the smoothed variables could be lost, the algorithm is expected to find the minimum number of variables that need to be smoothed, and aims to produce the minimum loss of information. In order to achieve it, the smoothing of the NPD matrix is done progressively and very carefully, removing a very low amount of information in each iteration. It could be said that we manipulated the NPD matrix in a loving way with the aim of damaging it as few as possible. This is why we label it *sweet smoothing*.

Simulation study

The simulation study had two main goals. The first was to assess the amount of information that is lost when the five approaches discussed so far are used. The second was to assess the extent to which the smoothing procedures recover the model at the population level. The study was based on binary variables and used a full factorial design with 3×2×5=30 conditions with 500 replicas per condition. The independent variables were:

1. Number of factors (1, 2 and 3). In the population, the size of factor salient loadings ranged between .42 to .77; and the size of non-salient loading ranged between -.15 to .15.
2. Number of items (18 and 24).
3. Size of negative eigenvalues in the NPD correlation matrices. The sizes were *very low* (in the range -0.001 and -0.05); *low* (in the range -0.06 and -0.10); *medium* (in the range -0.11 and -0.15); *large* (in the range -0.16 and -0.20); and *very large* (in the range -0.21 and -0.25).

The main idea in the simulation study was to produce the responses to sets of binary items that would show the appropriate factor solution for the population. The total number of responses was 100,000, and this was the population from which random small samples were drawn. The size of the samples was N=200 when the number of items was 18; and N=300 when the number of items was 24. In order to manipulate the size of the negative eigenvalues, samples were drawn at random until a tetrachoric correlation matrix was found that was NPD, and the negative eigenvalue of which was in the appropriate range.

We simulated data using the underlying variable approach. We used two thresholds: one for non-biased items (a threshold value between -.5 and .5), and another for biased items (a value threshold between 0.5 and 1.5). In each data generated, each item was at

random selected as symmetric item or as biased item. We must point out that our simulated dataset (N=100,000) must produce a PD polychoric matrix: it was when drawing samples of reduce size that the NPD polychoric matrices were expected.

For each condition, the following analyses were computed:

1. The population (tetrachoric) correlation matrix (that was PD) was computed and analyzed using Robust Unweighted Least Squares (RULS).
2. The sample (tetrachoric) correlation matrix (that was NPD) was computed and factor analyzed using RULS. In the factor analysis, items were allowed to show communalities larger than one.
3. The sample correlation matrix that was NPD was corrected using four approaches: Knol and ten Berge (1989), non-linear smoothing (Devlin et al., 1975); linear smoothing (Jöreskog & Sörbom, 1981); straight smoothing (Bentler & Yuan, 2011), and sweet smoothing. Each corrected correlation matrix was factor analyzed using RULS.
4. For each factor model, communality was computed, and the index *v* (i.e., the proportion of destroyed information in the smoothed correlation matrix) was computed for the smoothed correlation matrices. The goodness-of-fit of the factor model was assessed using CFI, GFI and RMSR.

Results

The procedure proposed by Knol and ten Berge (1989) failed to converge in 70% of the matrices. The convergence was largest when there were 24 items and a single factor was expected (91.2% of convergence). However, it never converged when there were 18 items and the number of expected factors was 3. As the ratio of convergence for this method was very low in our simulation study, we do not report the outcomes of this method.

Table 1 shows the mean of index *v* for each smoothing method. As expected, sweet smoothing destroyed least information in the smoothed correlation matrix. On the other hand, non-linear smoothing destroyed a large amount of variance. It must be said that the amount of information destroyed was minimum for all the methods when the negative eigenvalue was low (i.e., eigenvalues between -0.001 and -0.05): for example, in this situation index *v* showed values of .04 and .03 for straight and sweet smoothing, respectively. Conversely, the amount of variance destroyed was large when the size of the negative eigenvalue was large (i.e., eigenvalues between -0.21 and -0.25): for example, in this situation index *v* showed values of .23 and .14 for straight and sweet smoothing, respectively.

In addition, the table shows the mean bias of the estimated communality (i.e., the difference of the observed communality minus the population communality). As can be observed in the table, when no smoothing was applied, the communality was overestimated, while the smoothing procedures systematically underestimated the communality in the population. However, sweet smoothing was the method that most accurately reproduced the communality in the population.

Table 1. Indices of the decrease in variance produced by the smoothing algorithms

Index	No smoothing	Non-linear	Linear	Straight	Sweet
V	--	.321	.192	.140	.089
Bias in h	0.151	-1.393	-0.668	-0.421	-0.187

Table 2 shows the mean of the goodness-of-fit indices of the exploratory factor analysis for the population correlation matrix (that was PD), and the sample correlation matrix (that was NPD) with no smoothing at all. It can be observed that the analysis of this sample correlation matrix has the effect of underestimating the true value of the goodness-of-fit indices. The table also shows the mean of the goodness-of-fit indices of the

exploratory factor analysis after the various smoothing procedures have been applied. The worst fits were obtaining using the non-linear smoothing, while sweet smoothing recovered the true values in the population quite well.

Table 2. Goodness-of-fit indices related to the simulation study

	CFI	GFI	RMSR
Population	.942	.902	.075
No smoothing	.931	.894	.080
Smoothing techniques			
Non-linear	.923	.863	.069
Linear	.941	.894	.072
Straight	.940	.896	.073
Sweet	.943	.901	.074

The conclusion of the simulation study is that not to correct the sample correlation matrix when it happens to be NPD is a bad research option. In addition, sweet smoothing is a suitable strategy for correcting sample correlation matrices that fail to be PD.

As the short simulation below suggests, our proposal seems to improve the behavior not only of the original proposal on which is based (Straight Smoothing) but also of the existing alternatives as well.

For the benefit of the interested reader, we shall briefly describe the basis of the RULS approach used in the simulation. The parameter estimation is the same as in the standard ULS case. However, the standard errors and the goodness-of-fit indices derived from the chi-square statistic are corrected (i.e. robust corrections), using the information contained in the asymptotic covariance matrix. Useful further readings on the procedure are Li (2016) and Muthén (1993).

References

Bentler, P. M., & Yuan, K. H. (2011). Positive definiteness via off-diagonal scaling of a symmetric indefinite matrix. *Psychometrika*, *76*(1), 119-123. doi: 10.1007/s11336-010-9191-3

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, *62*, 531-545. doi:10.2307/2335508

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL 5: analysis of linear structural relationships by maximum likelihood and least squares methods*; [user's guide]. University of Uppsala.

Knol, D. L., & ten Berge, J. M. (1989). Least-squares approximation of an improper correlation matrix by a proper one. *Psychometrika*, *54*(1), 53-61.

Li, C. H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods, 21*(3), 369–387. doi:10.1037/met0000093

Lorenzo-Seva, U. & Ferrando, P. J. (2020). Not Positive Definite Correlation Matrices in Exploratory Item Factor Analysis: Causes, Consequences and a Proposed Solution. *Structural Equation Modeling: A Multidisciplinary Journal*, DOI: 10.1080/10705511.2020.1735393

Muthén, B. O. (1993). Goodness of fit with categorical and nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205-234). Newbury Park, CA: Sage.