

Title: Signal cancellation factor analysis

Author: André Achim

**Abstract:** Signal cancellation provides a radically new and efficient approach to exploratory factor analysis, without matrix decomposition nor presetting the required number of factors. Its current implementation requires that each factor has at least two unique indicators. Its principle is that it is always possible to combine two indicator variables exclusive to the same factor with weights that cancel their common factor information. Successful combinations, consisting of noise only, are recognized by their null correlations with all remaining variables. The optimal combinations of two multifactorial indicators, though, typically retain correlations with some other variables. Their signal, however, can be cancelled through combinations with unifactorial indicators of their contributing factors. The loadings are estimated from the relative signal cancellation weights of the variables involved along with their observed correlations. The factor correlations are obtained from those of their unifactorial indicators, corrected by their factor loadings. The method is illustrated with synthetic data from a complex six-factor structure that even includes two doublet factors. Another example using actual data documents that signal cancellation can rival confirmatory factor analysis.

**Keywords:** Factor analysis; Rotation-free; no Heywood cases; unambiguous doublet factors.

*Preprint DOI : [10.31234/osf.io/h7qwg](https://doi.org/10.31234/osf.io/h7qwg)*

Exploratory factor analysis (EFA) applies a common factor model to account for observed correlations among variables. It does so by decomposing the data correlation or covariance matrix into a smaller number of dimensions and applying a rotation scheme to provide a sound hypothesis of the relationship between factors and variables. The EFA process requires decisions for setting the number of common factors, for estimating the unique variance to be subtracted from the data matrix, and for rotating the principal components of the reduced matrix to meaningful latent variables that hopefully correspond to the underlying factors. A new approach to the common factor model is here proposed to recover the common factors without predetermining the number of required dimensions, decomposing the correlation matrix, nor requiring heuristic rotation.

Alternate graph-network methods (Cox & Wermuth, 1993, Golino & Epskamp, 2017) were proposed that do not depend on factoring a reduced correlation matrix. These methods conceptualize the correlations as links between pairs of nodes (i.e., variables). Their algorithms form communities of variables sharing high within-community partial correlations while having as few as possible strong between-community partial correlations. The partial correlation of two variables,  $A$  and  $B$ , with respect to a third variable  $C$  is meant to exclude from both  $A$  and  $B$  their common variance also shared with  $C$ , leaving mostly preexisting common variance unique to the pair. There are two caveats to this. Unless  $C$  perfectly reflects the underlying common factor, subtracting from  $A$  and  $B$  least-squares determined multiples of  $C$  only incompletely cancels their information shared with  $C$ . Furthermore, “decorrelating” both  $A$  and  $B$  from  $C$  injects into each some portions of the noise part (unique variance) of  $C$ , inducing some correlation due to such shared noise. Decorrelating  $A$  from  $C$  amounts to projecting  $A$  on  $C$ , keeping only the residual (orthogonal to  $C$ ), optionally raising it to unit length. When both variables are normalized to unit length, the projection of  $A$  on  $C$  amounts to  $r_{AC}C$ , where  $r_{AC}$ , the correlation of  $A$  and  $C$ , is the product of their loadings on the common factor. What is subtracted from  $A$  in “decorrelating” it from  $C$  does not remove all of  $A$ ’s loading on the common factor, but a fraction of this that corresponds to  $C$ ’s loading on the common factor. The factor dependent part of  $A$  (its signal) is thus incompletely cancelled. This observation was the seed to the new approach introduced here.

Signal cancellation.

The purpose of Signal Cancellation Factor Analysis (SCFA) is to decipher the factor structure behind the data. The only restriction for its success is that each factor is reflected in at least two exclusive variables; further variables may reflect multiple factors. The basic principle of signal cancellation is that it is always possible to combine two variables exclusive to the same factor such that their respective information from the factor (i.e. their signal part) cancels. For instance, if  $A$  loads 0.6 on the common factor and  $C$  loads 0.8, subtracting  $(6/8)*C$  from  $A$  cancels the factor contribution to both  $A$  and  $C$ . Such cancellation leaves only a combination of the unique variance (noise) of the two combined variables. Indeed, looking at  $A$  and  $C$  as having respective coordinates  $(.6, .8, 0)$  and  $(.8, 0, .6)$  on three orthogonal axes, one signal and two noise, the weighted sum  $A-(6/8)C$  has coordinates  $(0, .8, -.45)$ . This loads 0 on the common factor and consists only in a combination of two orthogonal noise dimensions.

Procedurally, non-linear optimization attempts signal cancellation on each pair of variables by minimizing the correlations of the combined variables with all remaining variables. These correlations are transformed into  $\chi^2$  values that provide both the between variable distances for clustering the variables into corresponding factors and a threshold where to stop the hierarchical clustering. Given that all pairs of variables within a cluster (factor) should be able to mutually cancel the signal from their common factor, any significant  $\chi^2$ , adjusting for their number tested, indicates failure of signal cancellation, which implies signal from distinct factors forbidding cluster fusion.

Individual variables informed by more than one factor generally do not enter any such cluster, by virtue of a significant  $\chi^2$ . An exception would be for two variables having proportional loadings on a pair of factors, creating a two-variable cluster coplanar with two other clusters. When this is detected, the most likely pair of variables is rejected as a factor, its variables becoming listed among the remaining multi-factor dependent variables.

Each variable not already associated with a factor is then explained through cancellation of its signal by variables representative of distinct factors, again asserting signal cancellation success by lack of correlation of the combination with any other variable. Whether the signal cancelling variables (or their respective factors) are correlated or not has no effect on this procedure.

The factor loadings are derived from the observed correlations along with the relative weights that achieved signal cancellation<sup>1</sup>. The factor correlations are estimated from the correlations of factor-exclusive variables, corrected for the loading on their respective factors<sup>2</sup>.

Signal cancellation implementation.

The present implementation of SCFA uses MATLAB R2023a, although an R version should be available very shortly. All its steps proceed without requiring any informed decision from the user. The application returns a message if some variables cannot be explained through signal cancellation. This would likely be due to some factors not being represented by at least two exclusive indicator variables. After normal termination, differences between solution implied and observed correlations are reported along with corresponding difference z scores to help appreciate the solution.

Although SCFA could be applied to the complete but normalized data, the operations for forming the weighted combinations of variables and for calculating their correlations with the remaining variables are more efficient when applied to the Cholesky transform of the correlation matrix. This yields an upper triangular matrix with variables as columns and rows containing the minimal

---

<sup>1</sup> For signal cancellation within a pair of normalized variables  $A$  and  $B$ , let their respective loadings be  $a$  and  $b$ . The expected  $r_{AB}$  correlation is then the product of  $a$  and  $b$ . That weight  $w$  cancels the signal in the combination  $wA-B$  implies that  $wa=b$ . Substituting  $wa$  for  $b$  in  $r_{AB}=ab$ , one gets  $r_{AB}=wa^2$ ,  $a=\sqrt{r_{AB}/w}$ , and  $b=r_{AB}/a$ . For a variable  $V$  whose signal is cancelled by variables  $B$  and  $G$  that respectively load  $b$  and  $g$  on their respective factors, the successful sum is  $w_bB+w_gG-V$  such that the loadings are obtained by  $w_bb$  and  $w_gg$ . This is because the optimal weights act by scaling the signal parts of the two cancelling variables.

<sup>2</sup> The correlation between two unifactorial variables is the product of their respective loadings. To assess the correlation of their respective factors, the variable correlation is corrected by dividing it by the product of the variable loadings on their respective factor. Obviously, for instance, this yields a correlation of 1.0 for the factor with itself when the two variables are indicators of the same factor.

information to reproduce all correlations with previous variables and to bring to 1.0 the sum of squares of each variable. Weighted sums of these columns have the same projections (correlations) on the matrix columns as if applied to the complete normalized dataset.

Signal cancellation involves finding weights for combining variables to minimize a set of correlations with the weighted sum. It also requires a statistical test for judging of signal cancellation success. The current implementation choice for SCFA is to set the weights by minimizing the largest absolute correlation, but to combine all correlations for statistical detection of imperfect signal cancellation. As a rule, expected null values divided by their standard error become  $z$  scores. For expected null correlations, the standard error is  $1/\sqrt{N-1}$ , yielding  $z=r\sqrt{N-1}$ . Also, a squared  $z$  score, notably here  $z^2=(N-1)r^2$ , is a  $\chi^2(1)$  (i.e. chi-square with one degree of freedom). The sum of  $k$  independent  $\chi^2(1)$  is distributed as  $\chi^2(k)$  under the null hypothesis. Here  $k$  would be the number of correlations of the optimized signal cancellation combination with the remaining variables.

When several tests are applied that could reject the same null hypothesis, it is appropriate to adjust the per-test criterion for the overall risk of a type I error to remain at the selected nominal .05 alpha level. For all of  $k$  tests, equivalently for the largest of  $k$  values, not to be significant when  $H_0$  prevails, the critical value may be set at the  $1-.95^{1/k}$ . This principle is used at several places in SCFA. This approach could not be applied, though, for directly assessing the significance of the minimized largest absolute correlation, because the minimization typically brings the largest correlation equal with the second largest.

Having identified clusters of unifactorial indicators of each factor, the correlation between pairs of factors may be estimated by averaging the correlations between all pairs of unifactorial variables, one from each factor, corrected by the inverse product for the factor loadings of the contributing variables. A factor correlation could alternately be estimated from that of their merged indicators<sup>3</sup>, also corrected by the inverse product of their factor loadings. The current implementation of SCFA uses the former approach. Although the relative merit of the two approaches remains to be investigated, the merged indicator approach would not work for explaining a single multifactorial indicator in the presence of only two factors, as there would be no remaining variable that should have expected null correlation of the last to-be-explained variable with the combinations of the two merged indicators.

The voluntarily complex 6-factor example used below to illustrate SCFA was made to include two doublet factors, where only one of them correlates with other factors. In conventional EFA, doublet factors require estimating two loadings constrained by a single correlation, such that, say, a correlation of 0.49 may be reproduced by any pair of loadings. Acceptable values range from (0.49, 1.0) to (1.0, 0.49), with no preference for (0.7, 0.7) and much less for that actual pair of loadings. The parameter search may even reach solutions where one normalized variable is assigned a loading larger than 1.0, as so-called Heywood case. The purpose of having two

---

<sup>3</sup> In merging  $k$  unifactorial indicators of the same factor, each is first scaled for unit noise (unique) variance. The expected noise variance of their sum is therefore  $k$ . Subtracting  $k$  from the sum of squares of the resulting merged indicator, before it is rescaled to unit variance, and taking the square root gives its unscaled *signal* amplitude, which becomes the merged indicator factor loading when the indicator is rescaled to unit variance.

doublet factors, is to illustrate that signal cancellation should not retain this indeterminacy for a doublet factor correlated with remaining factors, as yet incomplete cancellation between its two variables would imply correlations with variables from these other factors. For an independent doublet factor, fortuitous sample correlations would likely settle the loadings within reasonable limits. Presence or lifting of the indeterminacy for factors that load on only two variables may thus be judged from the factor correlations.

Illustrative Example: a difficult factor structure.

SCFA will first be illustrated with an arbitrarily defined factor structure defined to illustrate features of SCFA and to be deliberately challenging for existing factoring techniques, including graph-network methods. In specifying this factor structure, two concessions were made, namely that each factor should have at least two indicators, which is a current prerequisite for SCFA, and that each variable should have a community of at least 0.25. Besides the two doublet factors included in the test factor structure, one variable (at rank 12) cross-loads on three of the six common factors, and those at rank 6 and 7 have proportional loadings on two common factors. Furthermore, two orphan variables, i.e., that share no common information with any remaining variable, are included. Table 1 lists the factor loadings, and Table 2 the associated more or less arbitrary factor correlation matrix.

Table 1. A challenging structure of six factors, including two doublet factors and two orphan variables (with null loadings on all common factors).

Variable rank	Common factor					
	1	2	3	4	5	6
1	<b>.5</b>	0	0	0	0	0
2	<b>.6</b>	0	0	0	0	0
3	<b>.55</b>	0	0	0	0	0
4	0	<b>.5</b>	0	0	0	0
5	0	<b>.6</b>	0	0	0	0
6	0	<b>.5</b>	<b>.75</b>	0	0	0
7	0	<b>-.4</b>	<b>-.6</b>	0	0	0
8	0	0	<b>.5</b>	0	0	0
9	0	0	<b>.6</b>	0	0	0
10	0	0	0	<b>.5</b>	0	0
11	0	0	0	<b>.6</b>	0	0
12	0	<b>.4</b>	<b>.6</b>	<b>.4</b>	0	0
13	0	0	0	0	<b>.5</b>	0
14	0	0	0	0	<b>.8</b>	0
15	0	0	0	0	0	<b>.5</b>
16	0	0	0	0	0	<b>.8</b>
17	0	0	0	0	0	0
18	0	0	0	0	0	0

Table 2. Arbitrary correlations applied to the factors.

1	.4	-.3	0	.4	0
.4	1	0	-.3	.3	0
-.3	0	1	0	-.5	0
0	-.3	0	1	0	0
.4	.3	-.5	0	1	0
0	0	0	0	0	1

The resulting signal eigenvalues (excluding the unique variances) are 2.51, 1.50, 0.89, 0.77, 0.51, and 0.26. The expected eigenvalues of the full correlation matrix are 2.96, 2.13, 1.42, 1.40, 1.13, 1.00, 1.00, 0.93, 0.73, 0.72, 0.70, 0.70, 0.67, 0.60, 0.59, 0.58, 0.47, and 0.29. The two eigenvalues of 1.00 are due to the orphan variables.

Data were initially generated using this factor structure with  $N = 1000$ . Following indications that the solution was not reliable, a new sample of size  $N=2000$  was generated. The required larger sample size reflects the complexity of this illustrative example, not a general requisite of SCFA. To illustrate that signal cancellation does not require larger samples than alternate methods, SCFA will also be applied to data used by Tabachnick & Fidell (5<sup>th</sup> edition, 2007, p. 735) to illustrate confirmatory factor analysis, where sample size was  $N=175$ .

Challenging simulation,  $N=1000$ .

The current implementation of SCFA takes data (or correlation matrix and sample size) as input and returns a structure with many fields, some of which containing partial or temporary results of little relevance to users. Another command, SCFAreport, takes this structure as input and prints the (pattern) factor matrix and the factor correlation matrix. It also reports absolute value z-score of the differences between the Fisher (atanh) transformation of observed and solution-implied correlations in decreasing order of significance (all  $|z| > 2$ , with a minimum of five values are listed). The header indicates the critical value adjusted for number-of-test both per variable and globally; variables declared orphan are not counted among number of tests. The first 10 data lines of this part of the output are listed in Table 3. This leaves no doubt that the data would be incompatible with the obtained solution if it provided a close approximation of the population correlation matrix. The largest difference is for the observed correlation of 0.346 between variables 5 and 6, which the solution expects null.

Table 3. Statistical comparison of observed to solution-implied correlations. The critical  $|z\_diff|$  values at the .05 level, corrected for number of tests, are given in the header line.

r	c	Robs	Rsol	z_diff	z_crit(/var,gobal) = (2.9478, 3.5226)
5	6	+0.346	+0.000	11.3790	
5	7	-0.300	+0.000	9.7581	
8	9	+0.270	+0.520	9.4659	
5	12	+0.179	-0.106	9.0689	
5	9	+0.023	-0.255	8.9633	
6	9	+0.479	+0.665	8.8382	
9	12	+0.376	+0.575	8.1828	
4	6	+0.245	+0.000	7.9001	
4	9	+0.037	-0.203	7.6822	
4	12	+0.128	-0.084	6.7332	

Challenging simulation, N=2000.

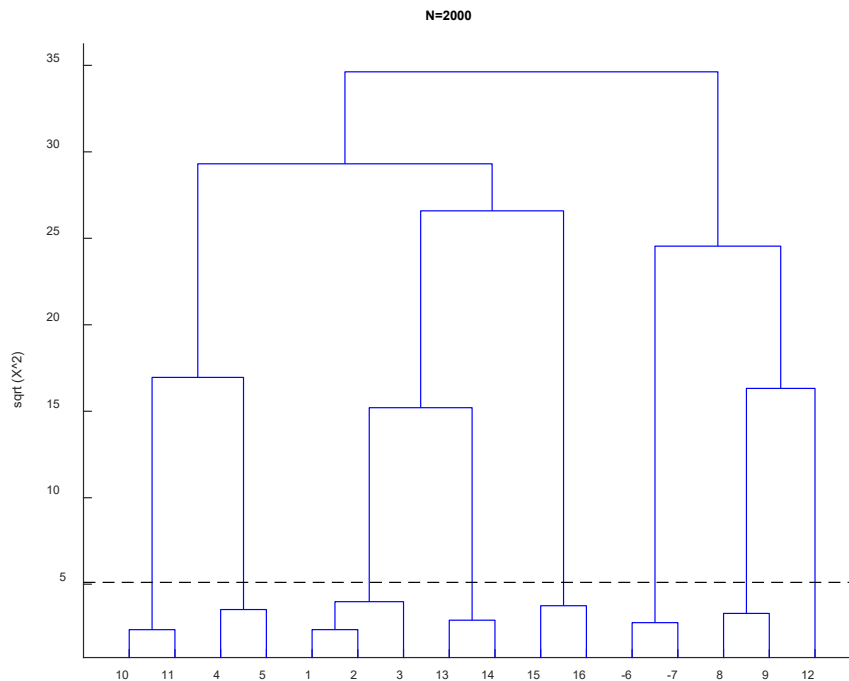
The SCFA solution for the same factor structure but with N=2000 appears reliable. The observed correlations closely correspond to the solution-implied correlations, with no significant residuals, even at the individual variable level threshold of 2.95. The largest  $|z|$  value was 2.86, for variables 7 and 12, for which the observed correlation is -0.468 and the solution estimate is -0.516. Although the  $|z|$  value calculations incorporate sample size, their critical values only depend on the number of (non-orphan) variables.

Orphan exclusion. The first SCFA step after the correlation matrix is produced, is inspection for likely orphan variables. For this, the maximum absolute correlation of each variable with all others is squared, multiplied by N-1, and compared to  $\chi^2_{\text{inv}}(.95^{1/17}, 1)$ , the constant 17 being the number of variables minus 1. This is equivalent to declaring significantly different from 0 a correlation of absolute value above .0663, correcting for 17 correlations that could appear significant for a variable that would be independent of all others. A variable is thus declared orphan if none of its observed correlations differs from 0 at the .05 level after adjusting for the number of tests. The smallest two  $|z|$  maxima are observed at 0.0356 and 0.0429 for variables 17 and 18 respectively, followed by a clearly significant correlation of 0.3086 between variables 10 and 11. The former two were declared orphan and removed from the following operations, although they will appear in the factor matrix solution with null loadings.

Paired signal cancellation. When only two variables are involved in signal cancellation, it does not matter which one is cancelling and which is cancelled. Each of the 16 remaining variables is paired with each 15 others, for a total of 120 pairs. For each pair, the weight of the first variable is optimized, the other having a fixed weight of -1. The two thus weighted variables are additively combined, the sum scaled for unit sum of squares and all (here 14) remaining normalized variables are projected on it to obtain their correlations with the weighted sum. The optimization criterion to minimize is the maximum (of the 14) absolute correlations. The sum of all 14 squared correlations (multiplied by N-1 to further allow statistical testing) becomes a figure of merit quantifying the distance between the two variables for purpose of clustering. Each pair thus yields an optimized weight and an associated  $\chi^2$  value.

Variable clustering into factors. The  $\chi^2$  values are used as mutual distances between all variables for the application of complete clustering. Complete clustering means that the distance between two partial clusters is the maximal distances between their respective elements. When this maximal distance is a non-significant  $\chi^2$ , corrected for the number of between sub-cluster pairs, we accept that their grouping only includes variables belonging to the same factor. Although the hierarchical clustering continues until all variables form a single cluster, the decision threshold is the last critical  $\chi^2$  value that allowed clustering of sub-clusters assessed as reflecting the same factor. Figure 1 provides the dendrogram of the hierarchical clustering achieved, with a dashed horizontal line representing the last (square root) critical  $\chi^2$  allowing homogeneous variable grouping. Note that variable 12, that loads on three factors, is not clustered below the dashed threshold, but that variables 6 and 7, with proportional loadings on two factors are clustered below the threshold. At this point, only variable 12 is listed as to-be-explained through cancellation of its signal by unifactorial variables from distinct factors.

Figure 1. Clustering dendrogram with square root  $\chi^2$  distance as vertical scale.



Coplanar cluster identification. The dendrogram indicates 7 clusters, all but one consisting of only two variables. All 35 ( $=7*6*5/(3*2)$ ) triplets of clusters were then tested for coplanarity. For a given triplet, each variable in turn represents its cluster and the representative variables of the first two clusters are set to cancel the signal of a variable from the third cluster. Here, this third variable receives a fixed weight of -1 and weights are optimized for the other two variables, minimizing the largest absolute correlation of the weighted sum with the remaining 13 variables. The sum of all squared correlations is multiplied by N-1 to transform it into  $\chi^2$ , here with 13 degrees of freedom. When all clusters involved in coplanar assessment consist in pairs of variables, this yields eight signal cancelling attempts, and eight corresponding  $\chi^2$ . Coplanarity is declared when none of these is significant at the .05 level, corrected for their number (here 8). When so, the mean correlation of variables from each pair of these clusters is obtained and the least correlated pair of clusters is kept as distinct factors while the variables of the remaining cluster are presumed multifactorial and become listed as to-be-explained (their signal to be cancelled) by unifactorial variables from distinct factors. Variable 12, that clustered way above the dashed horizontal line, was already thus listed. The coplanar cluster is also removed from the groupings that correspond to distinct factors. To identify variables from declared coplanar clusters, their ranks are given in negative in the dendrogram.

Factor loadings. At this step, the number of factors is known. The factor pattern matrix is created with 18 rows and 6 columns and populated with the estimated loadings of the different factors on their identified unifactorial variables, as per footnote 1. These will be required for estimating the loadings of the remaining multifactorial variables. Table 4 provides the estimated factor loadings of all variables, although those for the multifactorial variables will be obtained at a later step.



Table 4. SCFA solution pattern matrix for complex example with N=2000.

0.575	0	0	0	0	0
0.619	0	0	0	0	0
0.539	0	0	0	0	0
0	0.513	0	0	0	0
0	0.647	0	0	0	0
0	0.496	0.760	0	0	0
0	-0.410	-0.633	0	0	0
0	0	0.525	0	0	0
0	0	0.600	0	0	0
0	0	0	0.464	0	0
0	0	0	0.666	0	0
0	0.417	0.616	0.356	0	0
0	0	0	0	0.431	0
0	0	0	0	0.914	0
0	0	0	0	0	0.733
0	0	0	0	0	0.577
0	0	0	0	0	0
0	0	0	0	0	0

Factor correlations. The factor correlations, presented in Table 5, are then assessed, although this could be done later. For a given pair of factors, all (unifactorial) variables of one factor are in turn correlated with all variables of the other factor. When their average does not correspond to a significant correlation given sample size, the factor correlation is entered as null. Otherwise, each variable correlation is corrected by the inverse product of their loadings on their respective factor, as per footnote 2, and their average becomes the estimated factor correlation.

Table 5, Factor correlation matrix for complex example with N=2000.

1	0.4363	-0.3659	0	0.4229	0
0.4363	1	0	-0.3059	0.3226	0
-0.3659	0	1	0	-0.4210	0
0	-0.3059	0	1	0	0
0.4229	0.3226	-0.4210	0	1	0
0	0	0	0	0	1

Multifactorial loadings. The last step tries to cancel the signal of yet unaccounted for variables by pairs, trios, etc. (up to the total number of factors) of variables representing distinct factors. A message would follow if this left any variable unexplained, which would be most likely due to the lack of at least two exclusive indicators for each factor involved but might also follow from an insufficient sample size given the complexity of the population factor structure.

This step was initially implemented by merging the unifactorial variables of each factor into compound indicators, as per footnote 3. Although these provide the best signal/noise ratio to represent a factor, the approach severely limits, as previously mentioned, the number of remaining variables that should express no correlation with the weighted sum when cancellation is achieved. Indeed, any variable contributing some noise to a merged indicator is expected to correlate with the weighted sum by virtue of shared noise.

The current approach assesses the loadings of multifactorial indicators on its signal cancelling predictors as a weighted mean of the respective factor loadings estimated from all combinations of individual variable representing their factor. For instance, cancelling the signal of variable 6 should succeed only when pairing it with one variable from each of the 4-5 and of the 8-9 groupings, providing four estimates for the variable 6 loadings on the factors behind variables 4-5 and 8-9 respectively. For each of these four successful signal cancellation combinations, factor loadings are obtained from the optimized signal cancellation weights along with the observed correlations, as per footnote 2. The various estimates of the same factor loadings are weighted by the inverse of their signal cancellation  $\chi^2$  criterion, itself squared for more effective weighting, thus presuming that better signal cancellation provides better loading estimates.

SCFA of 11 WISC subtests, N=175

SCAF may also take a covariance or correlation matrix, along with sample size, as input. To illustrate CFA, Tabachnik & Fidell (2007) used a covariance matrix of 11 WISC subscales from 177 learning disabled children. Two outlier cases, one univariate and one multivariate, were removed. The data are modelled as reflecting two correlated factors, where the first six subscales are taken as pure indicators of verbal IQ and the last five as pure indicators of performance IQ. Their factor loading solution is reported in Table 6, along with that of SCFA, where variable 11 (coding) was deemed orphan. The 10-variable SCFA dendrogram is reported as Figure 2. The correlation between the two factors was reported as 0.589 for CFA and is found at 0.564 by SCFA. The five least-well fit observed correlations are reported in Table 7. Only two of 45 differences between model and observed correlations have  $|z|$  scores above 1.96. These do not even involve a common variable to suggest a possible lack of fit for a specific variable.

Figure 2. Dendrogram of WISC data.

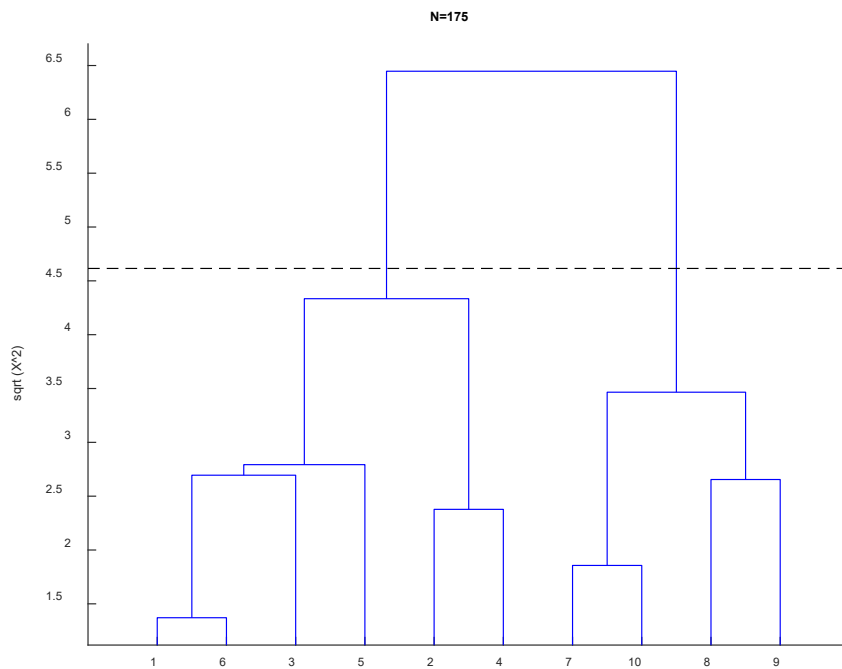


Table 6. CFA and SCFA solutions to the 11 WISC items correlation matrix

variable		CFA		SCFA	
rank	name	Verbal IQ	Performance IQ	Verbal IQ	Performance IQ
1	Information	.76	0	.77	0
2	Comprehension	.69	0	.72	0
3	Arithmetic	.56	0	.58	0
4	Similarities	.70	0	.73	0
5	Vocabulary	.77	0	.73	0
6	Digit Span	.39	0	.37	0
7	Picture Completion	0	.60	0	.62
8	Picture Arrangement	0	.47	0	.49
9	Block Design	0	.68	0	.70
10	Object Assembly	0	.57	0	.51
11	Coding	0	.07 (ns)	0	0

Table 7. Largest five  $|z|$  values of the difference between observed and solution-implied correlations of the WISC data, exclusive of orphan variable 11.

r	c	Robs	Rmodl	$ z\_diff $	$z\_crit(/var,gobal) = (2.7996, 3.2537)$
6	11	+0.173	+0.000	2.2925	
2	7	+0.407	+0.252	2.2832	
2	10	+0.322	+0.206	1.6438	
4	7	+0.369	+0.257	1.6409	
3	10	+0.043	+0.166	1.6297	

### Discussion

Signal cancellation is a simple yet radically new approach to the common factor model. It does not directly address approximating the correlation with fewer dimensions and rather concentrates on explaining each variable by its loading on factors that are identified through variable-pair signal cancellation.

Although SCFA does not aim at reproducing the observed correlations through least squares or maximum likelihood procedures, its solutions closely reproduce the observed correlation matrices under a few requirements, only one being specific to SCFA, namely that each factor should be represented by a minimum of two exclusive indicator variables. Other requirements, common to all factoring approaches, are the additive effect of factors on multivariate indicators and a sufficiently large sample size considering the factor structure. Since skewed variable distributions when the underlying factors are normally distributed are attributable to the measurement instrument (surface skewness), it is good practice to apply symmetrizing transformations before factoring, be it by EFA, SCFA or CFA. SCFA robustness to such surface skewness remains to be explored, but symmetrizing transformations remain recommended.

SCFA does not require preliminary estimation of the number of dimensions. Since typical simulation studies comparing different methods to estimate the number of factors use

simple factor structures with all or most variables exclusive to one factor, SCFA is expected to compete very well with such factor structures. One recent such study (Haslbeck & van Bork, 2024) compared the author's new method to several alternatives, including Horn's (1961) parallel analysis, tallying success over 200 datasets per condition. One of their most challenging test conditions was for six factors, all pairs correlated 0.8, with 3 variables per factor and random loadings from 0.28 to 0.71 after normalization. Sample size increased logarithmically from 100 to 5000. For this challenging factor structure, all methods achieved less than 10% success at  $N=847$  while SCFA correctly suggested 6 dimensions in 37 of the 200 trials. At  $N=1208$ , all methods were below 20% while SCFA was at 33.5%. At  $N=1722$ , SCFA exhibited a dip (36%), possibly attributable to sampling error, while the method based on Akaike Information Criterion exhibited a surge (44%) respective to neighboring sample sizes. This is the only sample size condition where SCFA did not surpass all five methods illustrated in the authors' Figure 3. This exploratory comparison Haslbeck and van Bork study was run before statistical comparison of observed and solution-implied correlations was implemented in SCFA. It is likely that most underestimates would have been flagged by large differences on many observed versus solution-implied correlations.

SCFA is less indeterminate than other factoring methods facing doublet factors, as signal cancellation will constrain the two factor loadings, provided that the doublet factor correlates with other factors. This complex 6-factor illustrative example used loadings of 0.5 and .08 for each of two doublet factors, to help appreciate loading recovery despite random error. As expected, the two loadings on the doublet factor correlated with other factors turned in the expected direction, at 0.43 and 0.91, while those for the doublet factor orthogonal to all other factors were 0.73 and 0.58, not reflecting the respective population loadings, where the second loading is larger than the first.

Further work should help selecting possibly better strategies, notably as to the benefit or not of weighted averaging as currently used for loadings of multifactorial variables but not for those of unifactorial loadings. Also factor score estimation is not yet implemented; alternate methods are foreseen, either based on averaging from individual factor-exclusive variables or from merged factor indicators.

Exploration of signal cancellation when some or all factors do not have two exclusive indicator variables is just starting. It seems that the factor space can be reliably delimited through signal cancellation, but that the factors themselves might remain unconstrained within that space. In lack of two unique indicators per factor, signal cancellation might also rely on the PARAFAC model (e.g., Harshman & Lundy, 1994; Kiers & Giordani, 2020) that currently requires unvarying factor correlations across datasets for the solution to be uniquely determined, as signal cancellation is not influenced by factor correlations. Important developments thus appear at the horizon for signal cancellation.

MATLAB code for SCFA is available at [AndreAchim/SCFA: Signal Cancellation Factor Analysis: MATLAB code \(github.com\)](https://github.com/AndreAchim/SCFA). An R version by P.-O. Caron, Université TÉLUQ, ([Pier-Olivier.Caron@Teluq.ca](mailto:Pier-Olivier.Caron@Teluq.ca)) will shortly be available on github (<https://github.com/quantmeth/SCFA>) and eventually on CRAN (the SCFA package).

## Acknowledgements

The author wishes to express gratitude for a scholarship from Epilepsy-Canada/Parke-Davis, back in 1988, that allowed pursuing his career in the development of innovative quantitative methods.

## References

- Cox DR & Wermuth N (1993) Linear dependencies represented by chain graphs. *Statistical Science* , 8: 204-218
- Golino HF & Epskamp S (2017) Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLoS ONE* 12(6): e0174035. <https://doi.org/10.1371/journal.pone.0174035>
- Harshman RA & Lundy ME (1994) PARAFAC: Parallel factor analysis. *Computational Statistics & Data Analysis*, 18(1), 39-72.
- Haslbeck JMB & van Bork R (2024) Estimating the number of factors in exploratory factor analysis via out-of-sample prediction errors. *Psychological Methods*, 29(1), 48–64. <https://doi.org/10.1037/met0000528>
- Horn JL (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185. <https://doi.org/10.1007/BF02289447>
- Kiers HAL & Giordani P (2020) Candecomp/Parafac with zero constraints at arbitrary positions in a loading matrix. *Chemometrics and Intelligent Laboratory Systems* 207, 104145
- Tabachnick BG & Fidrell LS (2007) Using Multivariate Statistics, 5th edition. Pearson Education Inc., Boston, USA.
- The MathWorks Inc. (2023). MATLAB version: 9.14.0 (R2023a), Natick, Massachusetts: The MathWorks Inc. <https://www.mathworks.com>