Title: Signal cancellation recovery of factors

Author: André Achim

Abstract:

Signal cancellation recovery of factors (SCRoF) produces sparse exploratory factor analysis solutions without matrix decompositionnor requiring the number of factors. The only assumption of SCRoF is that each factor has at least two unique indicators since the signal vectors of unique indicators are colinear with the origin, they can always be combined with suitable weights to cancel their common factor, leaving only a their unique variances, uncorrelated with all remaining variables. A sparse factor structure results from clustering variables that pairwise cancel their common signal. Since two variables with proportional loadings on two factors would also cluster together; such clusters must be detected and excluded as bona fide factor dimensions. Cancellation of the signal of multifactorial indicators is achieved by variables respectively exclusive to the factors involved. Factor loadings are estimated from the relationship of the signal cancellation weights with the observed correlations. Factor correlations are obtained from the correlations of all pairs of their unifactorial indicators along with their respective factor loadings. Finally, the individually estimated parameters of these sparse factor solutions are globally optimized for maximum likelihood, yielding χ2 assessment of the model success. SCRoF is illustrated with synthetic data from a complex six-factor structure that even includes two doublet factors. Another example uses actual data to document that SCRoF can benefit confirmatory factor analysis when the initial model does not fit the data well. Stand alone, Matlab and R versions of SCRof are available.

Keywords: Signal cancellation; Exploratory factor analysis; Sparse solutions; Rotation-free

#### Preprint DOI : 10.31234/osf.io/h7qwg

Abstract:

Geometrically, the common factor model views factors as directions in the multivariate data space. The signal of any variable is a vector within the factor subspace produced by adding its contributing factors with lengths given by factor loadings. Since the signal parts of variables all exclusively informed by the same factor are colinear with the origin along the factor, such variables can be combined to cancel their common factor direction, leaving only a compound of their unique variances uncorrelated with any remaining variable. After testing all pairs of variables for possible signal cancellation, clustering those that mutually cancel their signal identifies all potential factors having at least two unique indicator variables. Although signal cancellation would also succeed for two variables with proportional loadings on the same pair of factors, this can be recognized by cancellation of these variable signals by pairs of variables representing the two factors. The signal cancellation recovery of factors (SCRoF) algorithm implements these principles, only requiring that each factor has at least two unique indicators, not even requiring having estimated the number of factors. A two significance-threshold strategy leads to exploring alternate sparse matrix solutions. The individually estimated factor loadings and factor correlations of each potential solution are globally optimized for maximum likelihood, yielding a χ2 indication of compatibility with observed data. SCRoF is first illustrated with synthetic data from a complex six-factor structure. Actual data then document that SCRoF can benefit confirmatory factor analysis when the initial model appears inadequate.

Exploratory factor analysis (EFA, e.g., Thurstone, 1947; Tucker & MacCallum ,1997; Mulaik, 2009; Achim, 2020) applies a common factor model to account for observed correlations among variables in terms of a few common factors that inform these variables. Its main interest is that it provides suggestions about the relationship of the variables to their common factors without requiring a previous hypothesis of such relationship, except for the number of factors. This is especially useful in poorly developed knowledge domains. Yet, simple solutions are often hard to achieve as the EFA mathematical model allows all factors to inform all variables, leaving it to the user to apply a subsequent rotation to derive a subjectively acceptable and hopefully parsimonious solution (e.g., Howard, 2016).

EFA typically starts with estimating each variable’s unique variance, which is then excluded from the correlation or covariance matrix. The resulting reduced matrix is then decomposed into the preset number of orthogonal dimensions that hopefully delimit the complete and sufficient common factor space. The user must then apply a suitable rotation scheme to produce a sound hypothesis of the relationship between factors and variables, sometimes modifying the number of factors to achieve this. Different users may reach different solutions for the same data due to different decisions on (a) the algorithm to estimate the unique variances subtracted from the data matrix, (b) the number of common factors, (c) rotating the principal components of the reduced matrix to meaningful latent variables, or (d) nullifying low factor loadings or low factor correlations.

In principle, sparse matrix solutions (i.e. pattern matrices with many null loadings) should be easier to interpret. A few penalization-based procedures to obtain sparse matrix EFA solutions were suggested (e.g., Trendafilov, Fotanella & Adachi, 2017). They however all include at least one parameter to be tuned to the data at hand, which carries an extra level of subjectivity.

Confirmatory factor analysis (CFA; Jöreskog, 1967, 1978) also implements a common factor model, but the structure of this model must be pre-specified. Its purpose is to validate whether the data matrix could emanate from the specified structure with its parameters optimized. The optimized model becomes the null hypothesis for the observed data, which is assessed by a χ2 fit value that should not be significant for the model to be declared consistent with the data. For complex domains, where not all relevant sources of information are already known, an approximate model may be accepted despite the χ2 fit suggests it’s rejection. CFA wisdom even includes not having too large sample sizes, because these promote the detection of small inconsistencies between the data and the specified model. When the initial model is rejected, exploratory investigations, guided by data try to bring the χ2 fit index to non-significance. Other fit indices were developed to help decide on the acceptability of models that, although rejected by the χ2 test, may nevertheless be useful approximations.

Contrary to EFA solutions, most models subjected to CFA already specify that each factor directly informs only a subset of variables, thus typically corresponding to a sparse matrix solution to be tested against the data. Correlations among the factors themselves then account for the correlations between variables exclusively informed by different factors.

In brief, sparse factor structures are desirable when possible. CFA often test sparse models but requires pre-specifying the factor structure. It also benefits from a χ2 fit index useful to judge of the consistency of the optimized model with the observed data. EFA only requires specifying the number of factors but typically returns ambiguous dense solution matrices.

Signal Cancellation Recovery of Factors (SCRoF) is proposed herein as a new approach to the common factor model, hybrid between EFA and CFA to recover the common factors. SCRoF aims at deciphering the actual factor structure behind the observed correlated variables. Like EFA, it does not require previous knowledge of the solution, but unlike EFA, it does not require prespecifying the number of factors. Its signal cancellation approach does not proceed by decomposing the correlation matrix and does not need any final heuristic rotation. Its only specific requirement to recover the underlying factors and their correlations is that each factor is expressed in at least two exclusive variables. The approach consists in explaining all variables by cancelling their signal, i.e., of the information they received from their common factors.

SCRoF applies the following steps further detailed below: Orphan variables are first excluded since variables with no solid correlation would incorrectly mimic true pairwise signal cancellation. All possible pairs of retained variables are then tested for mutual signal cancellation and clusters are formed of variables able to mutually cancel their respective signal. These subsets of variables are temporarily considered unique indicators of a corresponding factor; a good approximation of their loadings is obtained. All triplets of clusters are then tested for coplanarity, meaning that any pair of variables from two clusters can cancel the signal of those from the remaining cluster. When three clusters are thus found to occupy only two dimensions, one of these clusters is excluded as factor; its variables are deemed multifactorial, just as any other variable that entered no cluster. The loadings of all multifactorial variables are then obtained through cancellation of their signal by a suitable subset of variables representing different factors. The factor correlation matrix is estimated from all correlations between the exclusive variables of the respective factors, along with their factor loadings; non-significant correlations are nullified. The resulting sparse pattern factor and correlation matrices are finally globally optimized for maximum likelihood yielding a χ2 fit index to assess compatibility of the observed data with the model.

SCRoF was inspired by graph-network approaches to EFA (Cox & Wermuth, 1993, Golino & Epskamp, 2017) that do not proceed by factorization of a reduced correlation matrix. These approaches conceptualize the correlations as links between pairs of nodes (i.e., variables). Their various algorithms form communities of variables sharing high within-community partial correlations while having as few as possible strong between-community partial correlations. Partial correlation is therefore at the heart of these approaches.

In general, 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, along with residual noise. There are two caveats to this. Unless *C* perfectly reflects the underlying common factor, subtracting from both *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 of the noise part (unique variance) of *C*, inducing some correlation between *A* and *B* due to such shared noise. While the residuals of both *A* and *B* become orthogonal to variable *C* (its signal plus its noise), simulations with 50% signal variance in all three variables easily show that these residuals remain correlated despite a single common factor being involved (i.e., without an extra source of variance exclusive to *A* and *B*).

In general, decorrelating *A* from *C* amounts to projecting *A* on *C*, keeping only the residual part, which is orthogonal to *C* (as a whole) but not to its signal part. When both variables are normalized to unit length, the projection of *A* on *C* amounts to *rACC*, where *rAC*, the correlation[[1]](#footnote-1) of *A* and *C*, is the product of their respective loadings, namely *a* and *c*, on their common dimension. Properly removing this shared signal from *A* would rather require subtracting *aC*/*c* from *A*, where *C/c* brings the shared signal in *C* to unit length and multiplication by *a* brings it back to the corresponding signal length in *A*. In brief, partial correlation weight *C* by *rAC*=*ac* (both less than 1.0) while the required weight is larger, at *a/c*. Thus, what is subtracted from *A* in “decorrelating” it from *C* does not remove all of *A*’s loading on the common dimension, but only a fraction that is *c*2 too small (although the fraction *c* is unknown) thus incompletely cancelling the part of *A* on the shared dimension (its signal). This observation, along with the need not to inject some of the predictor variance into both to-be-decorrelated variables, was the seed to the new approach introduced here.

Signal cancellation principle and procedure outline.

The signal of a variable that reflects a single factor can be represented as a vector in the direction of the factor, with its length from the origin expressing the factor loading. Two variables exclusive to the same factor thus have their respective signal colinear with the origin. The basic principle of signal cancellation is that, because of this feature, it is always possible to combine two such variables with suitable weights to cancel their signal, leaving nothing in the direction of the factor. The successful combination then lies in the plane of the two unique variances which is expected orthogonal to all other sources of variances.

For instance, if *A* loads 0.6 on the common factor and *B* loads 0.8, the difference between *A* and (6/8)\**B* (in either direction) cancels the factor contribution to both *A* and *B*. Indeed, looking at *A* and *B* as having respective coordinates (.6, .8, 0) and (.8, 0, .6) on three orthogonal axes, one signal followed by two noise dimensions, the weighted sum *A*-(6/8)*B*, called a *contrast*, has coordinates (0, .8, ‑.45). The successful contrast of two variables reflecting the same factor thus loads 0 on the common factor and consists only in a combination of their two orthogonal noise components.

Since the signature of successful signal cancellation is that the contrast consists of noise only, this contrast should not correlate with any remaining variable. In general, the signal of a variable informed by more than one factor cannot be cancelled by any other single variable. Two variables found to cancel their mutual signal are thus likely exclusively informed by a common factor, but this is not necessary. Two variables that would load proportionally on the same two factors are also colinear with the origin and can thus mutually cancel their respective composite signals.

Procedurally, SCRoF first identifies and sets aside orphan variables, i.e., those that show no significant correlation. Such variables are excluded from further processing and given null loadings on all factors. This exclusion is important since a pairwise contrast involving an orphan variable would inherit all the non-significant correlations of the orphan variable simply by giving it a huge weight relative to the other variable. Such apparent signal cancellation involving an orphan variable paired with any other variable would mislead the grouping of mutually cancellable pairs of variables as exclusive indicators of a common factor. SCRoF currently excludes the variables having no correlation significant at *p* ≤ .001, after correcting for the number of correlations of the variable.

Non-linear optimization then attempts signal cancellation on all pairs of retained variables through contrasts of the form *wA-B*, finding the weight *w* that minimizes the correlations of the contrast with all remaining variables. The actual minimization bears on the largest absolute value of these correlations, following which all correlations are combined into a χ2 value[[2]](#footnote-2) that provides both the between variable distances for subsequent clustering of the variables into common factors, and a probability threshold where to stop their hierarchical clustering.

The clustering requirement is that all pairs of variables within a cluster (a putative factor) should be able to mutually cancel their respective signal from a common factor. This calls for complete clustering for which the distance between two sub-clusters is the maximal distance between pairs of variables from each cluster. The fusion of two sub-clusters is therefore forbidden by any significant χ2 between variables from different sub-clusters after adjusting for the number of tests. Indeed, a failure of pairwise signal cancellation, documented by a significant χ2, implies that at least two common factors are involved in that pair, which forbids the fusion of their respective clusters as a set of variables all exclusively influenced by the same factor.

As already mentioned, successful variable clustering does not guarantee that they share signal from the same common factor. A cluster of two variables having proportional loadings on two factors must however be coplanar with the two clusters of variables exclusive to one or the other of these two bona fide factors. This implies that the number of factors might be less than the total number of clusters emanating from the hierarchical clustering based on pairwise signal cancellation.

To identify the presence of coplanar clusters, all triplets of clusters, at least one of which has only two variables, are tested for possible coplanarity, which involves that any pair of variables from two of these clusters can cancel the signal of any variable from the remaining cluster. Failure of cancellation implies that the three clusters tested occupy three dimensions and thus form three distinct factors. When collinearity is detected between three clusters, the variables of a two-variable clusters are declared multifactorial dependent. If two or all three coplanar clusters have only two variables, parallel scenarios are developed in which a different two-variable coplanar cluster loses its putative factor status.

The variables not already associated with a factor are then individually explained through cancellation of their signal by two or more variables that represent distinct factors, again asserting signal cancellation success by lack of correlation of the optimized contrast with the remaining variables. For instance, for a variable *V* whose signal is a sum of two factors respectively represented by variables *A* and *B* and by variables *E*, *F* and *G*, cancellation of the signal of *V* would be achieved by a contrast opposing *V* to a suitably weighted sum of *A* and *E* or of any other pair representing each factor. Whether the signal cancelling variables (or their respective factors) are correlated or not has no effect on this procedure[[3]](#footnote-3).

All factor loadings are derived from the optimal signal cancellation weights along with the observed correlations[[4]](#footnote-4). The average is used when the same variable loading is multiply estimated from its signal cancellation with all possible alternate variables. At this point, the sparse pattern matrix is completed with good estimates of the respective loadings.

Elaboration of the factor correlation matrix relies on the SCRoF requirement that each factor is represented by at least two exclusive variables. Although correlation significance is based on the first significance test of the canonical correlation applied to the exclusive indicators of a pair of factors, the corresponding first canonical correlation underestimates the correlation between the two underlying factors due to the presence of residual noise. Rather, following the principle that the correlation between two variables is the product of their respective loading further multiplied by the correlation between their factors, a good correlation estimate is obtained from the weight that brings the vector of products of between-factor loadings closest to the vector of between-factor variable correlations[[5]](#footnote-5).

Non-significant correlations are then nullified using a false discovery rate strategy (FDR; Benjamini & Hochberg, 1995) along with two statistical thresholds. FDR implies ordering the *p* values for all pairs of factors from the most to the least significant and correcting each *p* value for assessing the maximum correlation among all not yet assessed correlations[[6]](#footnote-6). As discussed below, two significance thresholds are used, bracketing a zone of higher risk of incorrect decision. All clearly non-significant correlations are first nullified while all clearly significant correlations are maintained. All subsets of correlations that have an adjusted *p* between the two limits, if any, are nullified in alternate factor correlation matrices.

Global parameter optimization is finally applied to all non-zero entries of a factor correlation matrix and of the associated factor pattern matrix (exclusive of orphan variables) using their current values as initial parameter estimates. This maximum likelihood optimization mimics CFA in refining the parameter estimates and providing a χ2 fit index to assess compatibility with the data. To prevent Heywood cases, the criterion evaluation function returns a huge value for parameter combinations that either brings a variable community above .98 or results in a population correlation matrix with a negative determinant. An explored solutions (pattern and associated correlation matrix) that results in a community thus topped at or just below .98 are likely incorrect. This would be acknowledged by the corresponding χ2 fit indicating incompatibility between that solution and the data.

Additional considerations in SCRoF implementation.

SCRoF operates on the data correlation matrix along with sample size, but the input may also consist of the raw data or of the covariance matrix along with sample size. All its steps proceed without requiring any informed decision from the user. Since more than one explored scenario might appear consistent with the data, the user might have, however, to discuss the relative merits of different statistically acceptable models, and openly opt, or not, for one of them.

Although SCRoF’s signal cancellation computations 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 with rows containing the minimal information required to reproduce all correlations with previous variables and to bring to 1.0 the sum of squares of each variable (i.e., column). Weighted sums of these columns have the same projections on the remaining matrix columns as if applied to the complete normalized dataset. Thus, when a contrast is normalized to unit sum of squares, its sums of cross products with the remaining variables directly assess the corresponding correlations.

Alternate factor structure scenarios have already been mentioned about which coplanar clusters to disqualify as a bona fide factor. Alternate scenarios were also implied in nullifying subsets of factor correlations with uncertain significance. The two-threshold strategy also creates alternative factor structure solutions at the variable grouping stage that specifies the factor pattern matrix, i.e. for deciding on variable clustering and on coplanarity detection.

To resist both type I (false positive) and type II (false negative) errors, SCRoF uses two statistical thresholds, namely .001 and .25. Obtaining p ≤ .001 for a relevant statistical test causes rejection of the current null hypothesis. Similarly, *p* ≥ .25 makes the scenario consider that the null hypothesis holds. When .001 < *p* < .25, both decisions are considered in alternate processing scenarios. For instance, two sub-clusters with their worst signal cancellation pair yielding p = .09 will be grouped together in one scenario and kept as distinct clusters in another.

When several cluster unifications are thus statistically undecided, all subsets of statistical decisions are considered in separate scenarios. This could however substantially reduce the number of factors when several undecided groupings are simultaneously excluded. To prevent exploring solutions with clearly too few factors, scenarios are rejected if involving fewer than the minimum number of required dimensions as assessed with the Next Eigenvalue Sufficiency Test (NEST; Achim, 2017). This statistically based method was documented not to exceed its nominal type I error rate. The latter is here set at .002, meaning that a k-factor model is rejected as insufficient only if the empirical probability of the next eigenvalue (i.e., at rank k+1) is less than this low threshold, which strongly limits the risk of overestimating the number of required dimensions.

SCRoF tests for the maximum of some statistics. For instance, in deciding if a sub-cluster of two variables may be aggregated with another of three variables, it will require that all six signal cancellation χ2 involving one variable from each sub-cluster be not significant. The test is thus applied on the most significant of these six χ2. The associated probability *p* is then converted to a net probability as 1-(1-*p*)*s*, where *s*, here six, is the number of statistics over which the maximum was retained. For instance, with ten variables, the contrast of each pairwise signal cancelling attempt is correlated with the eight remaining variables; these correlations are transformed into z2 which are summed to form a χ2 with eight degrees of freedom. If the maximum of the six χ2(8) was, say, 27.0 which has an associated *p* of .0007, this raw probability would become a net probability of 1-.99936 = .0042 that all six χ2 are at or below the observed maximal value. In such a case, this net probability would cause the aggregation of the two clusters to be both allowed and prevented in parallel scenarios.

The two-threshold approach also applies to coplanarity testing, where an in-between probability results in separate scenarios, one in which all three clusters involved will be treated as three distinct factors and other scenarios with each two-variable cluster being excluded, one at the time, as a bona fide factor. This could yield four parallel scenarios if all three possibly coplanar clusters consist of only two variables.

The list of relevant explored scenarios is printed at SCRoF termination along with their χ2 statistics but excluding those with *p* < .0001[[7]](#footnote-7). Scenarios with the same pattern matrix but alternate correlation matrices are listed consecutively, with extra information on the line that introduces the factor pattern matrix. For instance  
 4: p=0.13522 X2(94)=109.202 VG6 FC1 6f mf:6,7 Coplan VG1  
designates scenario #4 with its probability and χ2 values; this scenario is internally designated as variable grouping 6 with factor correlation variant 1. All lines have this information, the rest of the above line is unique to a new variable grouping. Here, ‘6f mf:6,7’ indicates a 6-factor solution with variables 6 and 7 made multifactorial; the end of the line indicates that the current VG6 group emanated from the coplanar detection function when processing variable grouping 1. These lines would terminate with the ranks of variables that could not be explained through signal cancellation, if any. These incomplete solutions are listed last. As their fit statistics ignore the unexplained variables, such incomplete solutions would only be considered when no complete solution is statistically acceptable.

The main SCRoF output however consists in a data structure that contains the details of all explored scenarios along with several intermediate results. An associated command, SCRoFreport, receives the data structure name along with a line number from the SCRoF printout. It prints the factor pattern matrix and factor correlation matrix of the corresponding solution; it also paints the dendrogram depicting the pairwise clustering of variables in which the variables are identified by their ranks; the variables clustered together in the scenario are linked together by thick black lines. The tick lines of a coplanar cluster excluded as a bona fide factor are rather painted in grey.

Illustrative Example: a difficult factor structure.

Condition. SCFA will first be illustrated with an arbitrarily defined 6-factor structure deliberately challenging for standard EFA techniques. It contains two orphan variables, two doublet factors (one correlated and the other orthogonal to remaining factors), a pair of variables with proportional loadings on two factors and a variable loading on three factors. In specifying this factor structure, two concessions were made, namely that each factor should have at least two exclusive indicators, as required for SCRoF, and that each variable should have a community of at least 0.25. The arbitrary ‘population’ factor loadings are given in the left part of Table 1 and their arbitrary correlations in the left part of Table 2.

The population signal eigenvalues (exclusive of unique variances) are 2.51, 1.50, 0.89, 0.77, 0.52, and 0.26. The expected eigenvalues of the population 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 1.00 for the sixth and seventh population eigenvalues are due to the orphan variables.

Given the complexity of this factor structure, sample size was set at N=2000 to clearly illustrate SCRoF. Details of the SCRoF solution with this sample are provided in Appendix A while Appendix B further discusses SCRoF output for five independent samples with N=1000.

Results. With N=2000, the initial clustering of pairwise cancellation gave seven clusters, but no seven-factor solution made it with *p* > .05. One six-factor solution occurred identically twice, once where the clustering of v6 with v7 was rejected and once when their clustering was accepted but its variables made multifactorial following coplanarity detection. Out of thirteen different solutions reported, nine had a χ2 fit with *p* > .05. Six of these implied declaring variables v8 and v9 as multivariate, the remaining three declared v6 and v7 as multivariate. None of six scenarios considering v4 and v5 as multivariate made it to the SCRoF output as each was rejected by the χ2 fit statistics with p < .0001. Inspection of the pattern of correlations between the initial seven clusters indicated that the clusters of v4, v5 and that of v8, v9 were orthogonal while the cluster of v6 and v7 correlated respectively .55 and .90 with the former two clusters. A correlation of .90 between two factors would be quite suspect. This is likely the reason why declaring v4 and v5 as multifactorial was systematically rejected. If these were real data, it would remain possible that the clusters of v4, v5 and that of v6, v7 are the two factors (correlating .55), but it remains more likely that the clusters of v4, v5 and of v8, v9 are the two factors, given their orthogonality. Although the contents of v6 to v9, forming the clusters among which one is not a factor, could argue otherwise, preference should go to the latter factor structure. We will therefore consider the cluster of v4, v5 as factor F2 and that of v8, v9 as factor F3, labeling the factors not by their extraction order but by the corresponding population factors.

Aside from the solution obtained twice, there are three solutions with v6 and v7 as multifactorial. They all differ by the faith of v1. In one scenario, v1 is clustered with F1 (made of v2 and v3) and loads .53 on it. This scenario has χ2(94) = 109.20, *p* = .14. It happens to be the correct solution although we would not know it. The other two solutions rejected the grouping of v1 with F1, thus treating it as multifactorial. Their common dendrogram of pairwise clustering is presented in Figure 1. In these two scenarios, v1 loaded respectively .54 and .48 on F1 and loaded .05 on factor F6 represented by v15 and v16. Since its signal cancellation by only two factors had a net probability of .11, this was both accepted and rejected. The model with v1 depending on two factors had χ2(93) = 103.30, *p* = .22; that with v1 further informed by factor F5 of v13, v14 (with loading .10) had χ2(92) = 92.50, *p* = .47. Although suspicious for v16 harboring a loading of .99, this solution would likely be blindly preferred and is the one presented in the right parts of Tables 1 and 2. The side-by-side presentation of population and solution matrices was meant to facilitate result appreciation.

Table 1. Population (left) and SCRoF ‘preferred’ solution (right) for synthetic data with N=2000.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable rank | Population common factors | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | |
| 1 | **.5** | 0 | 0 | 0 | 0 | 0 | |
| 2 | **.6** | 0 | 0 | 0 | 0 | 0 | |
| 3 | **.55** | 0 | 0 | 0 | 0 | 0 | |
| 4 | 0 | **.5** | 0 | 0 | 0 | 0 | |
| 5 | 0 | **.6** | 0 | 0 | 0 | 0 | |
| 6 | 0 | **.5** | **.75** | 0 | 0 | 0 | |
| 7 | 0 | **-.4** | **-.6** | 0 | 0 | 0 | |
| 8 | 0 | 0 | **.5** | 0 | 0 | 0 | |
| 9 | 0 | 0 | **.6** | 0 | 0 | 0 | |
| 10 | 0 | 0 | 0 | **.5** | 0 | 0 | |
| 11 | 0 | 0 | 0 | **.6** | 0 | 0 | |
| 12 | 0 | **.4** | **.6** | **.4** | 0 | 0 | |
| 13 | 0 | 0 | 0 | 0 | **.5** | 0 | |
| 14 | 0 | 0 | 0 | 0 | **.8** | 0 | |
| 15 | 0 | 0 | 0 | 0 | 0 | **.5** | |
| 16 | 0 | 0 | 0 | 0 | 0 | **.8** | |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 18 | 0 | 0 | 0 | 0 | 0 | 0 | |
| Variable rank | SCRoF solution common factors | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | **.48** | 0 | 0 | 0 | **.10** | **.05** |
| 2 | **.60** | 0 | 0 | 0 | 0 | 0 |
| 3 | **.55** | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | **.47** | 0 | 0 | 0 | 0 |
| 5 | 0 | **.58** | 0 | 0 | 0 | 0 |
| 6 | 0 | **.47** | **.76** | 0 | 0 | 0 |
| 7 | 0 | **-.35** | **-.63** | 0 | 0 | 0 |
| 8 | 0 | 0 | **.53** | 0 | 0 | 0 |
| 9 | 0 | 0 | **.61** | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | **.45** | 0 | 0 |
| 11 | 0 | 0 | 0 | **.59** | 0 | 0 |
| 12 | 0 | **.39** | **.60** | **.42** | 0 | 0 |
| 13 | 0 | 0 | 0 | 0 | **.45** | 0 |
| 14 | 0 | 0 | 0 | 0 | **.81** | 0 |
| 15 | 0 | 0 | 0 | 0 | 0 | **.38** |
| 16 | 0 | 0 | 0 | 0 | 0 | **.99** |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 2. Arbitrary correlations applied to the factors in the population (left) and in the SCRoF ‘preferred’ solution (right).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 |
| 1 | **.47** | **-.23** | 0 | **.29** | 0 |
| .47 | 1 | 0 | **-.29** | **.29** | 0 |
| -.23 | 0 | 1 | 0 | **-.49** | 0 |
| 0 | -.29 | 0 | 1 | 0 | 0 |
| .29 | .29 | -.49 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 |

Figure 1. Variable clustering dendrogram based on pairwise signal cancellation, for the two solutions in which v1, v6, v7 and v12 are multifactorial. The initial cluster of v6 and v7 is marked by grey lines.



Conclusion. For this complex synthetic example, we retain that the correct solution was among the scenarios qualified at *p* > .05. Nearby solutions might appear preferable by their χ2 fit, which reminds us that better ML fit does not guarantee closer to truth. Replication would be urged if the differences between the three solutions implied non-trivial theorical consequences.

Real data example

We next illustrate (a) that large sample sizes are not a particular requisite of SCRoF compared to alternate EFA methods and (b) that signal cancellation can bring extra constraints on CFA solutions when the original model needs modification.

In Chapter 14 of Tabachnik & Fidell (2019), author J.B. Ullman used 11 WISC subscales data from 177 learning disabled children to illustrate structural equation modelling. Two outlier cases, one univariate and one multivariate, were removed. The data were initially modelled as reflecting two correlated factors, where the first six subscales are taken as pure indicators of Verbal intelligence and the last five as pure indicators of Performance intelligence. This model did not fit the data, with χ2(43) = 70.2 (*p* =.0054). LISREL suggested adding a link from Performance to Comprehension, which brought the fit to χ2(42) = 60.3 (*p* =.033). As the Coding variable did not load significantly on its Performance factor, its further removal yielded the final model with χ2(33) = 45.0 (*p* = .08).

Adopting the signal cancellation point of view, this solution implies that no single other variable could cancel the bifactorial signal of Comprehension, it being the only multifactorial variable in the model. SCRoF should confirm this prediction if the solution is correct.

Results. SCRoF first eliminated Coding as an orphan variable and reported three scenarios. One corresponded closely to the original two-factor model, but without the Coding variable, in which all indicators are unique to their factor. As SCRoF ignores the orphan variables in assessing model fitness, it reported χ2(34) = 55.0, *p* = .013 for this model. Two other solutions acknowledged that Comprehension and Similarities could mutually cancel their respective signal. Their common dendrogram is presented in Figure 2. In one of these scenarios (χ2(32) = 37.7, *p* = .23), the cluster of Comprehension and Similarities deemed these variables as combinations of the Verbal and Performance factors. In the third scenario (χ2(32) = 38.282, *p* = .21), coplanarity was rejected and this two-variable cluster was maintained as a third factor. Acknowledging three factors is not only less parsimonious than a two-factor solution, but the extra factor would correlate .91 with Verbal and .75 with Performance; these two would still correlate .48. The preferred two-factor solution is presented in Table 3, along with the final textbook solution.

Figure 2. Variable clustering dendrogram expressing pairwise signal cancellation among the 10 retained WISC subscales, which clearly shows the mutual signal cancellation of Comprehension (v2) and Similarities (v4).



Table 3. Final textbook CFA solution to the 10 WISC subscales (after deleting Coding) and best SCRoF solution. The factor correlations are respectively .59 and .48.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | | CFA | | SCRoF | |
| Rank | Name | Verbal | Performance | Verbal | Performance |
| 1 | Information | **.78** | 0 | **.79** | 0 |
| 2 | Comprehension | **.50** | **.30** | **.47** | **.35** |
| 3 | Arithmetic | **.56** | 0 | **.57** | 0 |
| 4 | Similarities | **.70** | 0 | **.55** | **.25** |
| 5 | Vocabulary | **.78** | 0 | **.78** | 0 |
| 6 | Digit Span | **.40** | 0 | **.41** | 0 |
| 7 | Picture Completion | 0 | **.62** | 0 | **.63** |
| 8 | Picture Arrangement | 0 | **.45** | 0 | **.45** |
| 9 | Block Design | 0 | **.67** | 0 | **.64** |
| 10 | Object Assembly | 0 | **.58** | 0 | **.58** |

Conclusion. By finding signal cancellation of Comprehension by another variable, namely Similarities, SCRoF contradicts the earlier solution and indicates that Comprehension and Similarities have proportional enough loadings on both the Verbal and Performance factors for these to mutually cancel their signal.

**Discussion**

The principle of signal cancellation was introduced as a radically new approach to EFA in which factors are individually identified from pairs of their unique indicator variables. Once the unifactorial variables are associated with their common factors through pairwise signal cancellation, the remaining multivariate variables are explained by cancelling their composite signals by subsets of unifactorial variables. Contrary to the standard EFA model that allows all factors to affect all variables, this approach naturally produces sparse solutions. SCRoF rotation-free solutions are produced without user intervention, not event to specify the number of factors. The two-threshold approach, however, may yield a few solutions consistent with the data.

Along with the χ2 fit probability of the data given the model, assessing the merits of alternate solutions should consider model parsimony and the between-factors correlations. The 6-factor synthetic example illustrated that the correct solution was among those not rejected by the χ2 test, although other solutions presented better fits. The solution reported in Table 1 as the one that “would likely be blindly preferred” improved the χ2 fit, over the solution that we know to be the correct one, by 16.7 for 2 degrees of freedom, which counterbalanced having to accept its factor loading of .99 on variable 16.

More experience and debates among experts will be required on the ensuing question of how large a fit improvement must be, between two otherwise statistically acceptable models, to carry its preference over the more parsimonious alternative. We here have an example for which knowing the correct solution indicates that parsimony should have prevailed. With real data, the credibility of differences between statistically acceptable solutions could be decided by new data, especially if the extra parameters have important theoretical consequences. Such reports would usefully feed the reflection on this issue.

SCRoF solutions are expected to closely reproduce the data 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, by SCRoF or by CFA. Although similar skewness should not prevent pairwise signal cancellation of same factor indicators in SCRoF, opposite surface skewness for positively correlated indicators, already known to require an extra factor in CFA (e.g. Brandenburg, 2024), could be equally misleading for SCRoF. Preliminary inspection of the data and eventually variable transformations remain strongly recommended.

SCRoF is less indeterminate than other EFA 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 sampling error. As the solution presented in Table 1 did not recognize F6 as a doublet factor, the doublet factor point is best illustrated by the completely correct solution (that with *p* = .14). In testing for significant correlations

When v1 was clustered within F1, SCRoF recognized both . 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)) is available on github (<https://github.com/quantmeth/SCFA>) and eventually on CRAN (the SCFA package).

The present implementation of SCRoF was developed in MATLAB R2023a. A stand-alone version of SCRoF is also available. An R version should be available by the time of this publication.

Future versions of SCRoF may come to deal with such situations.

Comparison of SCRoF with other methods to assess the required number of factors using a difficult condition from Haslbeck, & van Bork, Psychological Methods, 2024, 29(1): 48-64 could also be added.)

Thus, while SCRoF operates without any decision from its user, it may present alternate solutions consistent with the data. When more than one solution is reported, the recommended practice is to report all those compatible with the data, and to discuss the reason to prefer one over the others, like parsimony on the number of factors involved or very large factor correlations.

When several scenarios were investigated, one of these solutions often stands out either for having the only non-significant χ2 or for being more parsimonious than an alternate non-significant solution. The most parsimonious non-rejected scenarios are flagged as the suggested variant of a given factor structure.

This 6-factor example includes 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, resulting in a so-called Heywood case. The purpose of having two doublet factors, is to illustrate that signal cancellation should not retain this indeterminacy for a doublet factor correlated with remaining factors; indeed, incomplete cancellation between its two variables would imply correlations of the contrast with the variables from these other factors. SCRoF detects independent doublet factors (their two indicators have no other significant correlation) and assigns them equal loadings, as Onyx does,

Since it is possible for a scenario to correspond to an impossible population correlation matrix (i.e. with a negative determinant) or to include a variable with a communality above unity, such scenarios would be listed with *p*=0 and χ2(0) = 0.

This would likely be due to some factors not being represented by at least two exclusive indicator variables. If no other solution is consistent with the data, removing those variables from the analysis could be considered.

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

### Achim A. (2017). Testing the number of required dimensions in exploratory factor analysis. *The Quantitative Methods for Psychology*, 13(1), 64–74. doi:10.20982/tqmp.13.1.p064.

### Achim A. (2020). Esprit et enjeux de l’analyse factorielle exploratoire [Spirit and issues in exploratory factor analysis]. *The Quantitative Methods for Psychology*, 16(4), 213-247. doi :10.20982/tqmp.16.4.p213

### Benjamini Y. & Hochberg Y, (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1) 289-300.

### Brandenburg N. (2024) Factor retention in ordered categorical variables: Benefits and costs of polychoric correlations in eigenvalue‑based testing. *Behavior Research Methods*, https://doi.org/10.3758/s13428-024-02417-0

### Cox D.R. & Wermuth N. (1993) [Linear dependencies represented by chain graphs](https://www.jstor.org/stable/2245958?casa_token=e0Pua6i52tYAAAAA:wLi7Ty-peUsjZ49pWRga2BPJ4Qcp4H_aIv3Nq0n0xmvarAzo9O1dCxCTuw8GHp2w8FPs0NQA3EpTEKf4KbFnhwkXbsE_6tXYKsknfJmmTstY4EuE-Deo). *Statistical Science*, 8: 204-218.

Golino H.F. & 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 R.A. & Lundy M.E. (1994) PARAFAC: Parallel factor analysis. *Computational Statistics & Data Analysis*, *18*(1), 39-72.

Horn J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185. <https://doi.org/10.1007/BF02289447>

Howard MC (2016) A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *Intl. Journal of Human–Computer Interaction*, 32: 51–62. DOI: 10.1080/10447318.2015.1087664

Jöreskog K.G. (1967) [Some contributions to maximum likelihood factor analysis](https://www.diva-portal.org/smash/record.jsf?pid=diva2:49861). *Psychometrika*, 32, 443-482. https://doi.org/10.1007/BF02289658.

Jöreskog K.G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, *43*, 443-477.

Kiers H.A.L. & Giordani P. (2020) Candecomp/Parafac with zero constraints at arbitrary positions in a loading matrix. *Chemometrics and Intelligent Laboratory Systems* 207, 104145

Mulaik S.L. (2009) Foundations of Factor Analysis (2nd ed.). Chapman & Hall/CRC, Boca Raton.

Spearman C. (1904) The proof and measurement of association between two things. *The American Journal of Psychology*, 15, 72–101. Reproduced in: *International Journal of Epidemiology*, 2010, 39:1137–1150. doi:10.1093/ije/dyq191

Tabachnick B.G. & Fidell L.S. (2019) Using Multivariate Statistics, 7th 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>

Tucker L.R. & MacCallum R.C. (1997) [Exploratory factor analysis](http://www.ffzg.unizg.hr/psihologija/phm/nastava/Book_Exploratory%20Factor%20Analysis.PDF). Unpublished manuscript, Ohio State University, Columbus (available from Google Scholar)

Thurstone L.L. (1947) Multiple-factor Analysis: A Development and Expansion of The Vectors of the Mind. University of Chicago Press.

Trendafilov N.T., Fontanella, S. & Adachi, K. (2017). Sparse Exploratory Factor Analysis. *Psychometrika*, 82(3) pp. 778–794.

Challenging simulation, N=2000.

Figure 1 reproduces the call to SCRoF with its text output. Statistically acceptable but incomplete solutions would be relevant only if no complete solution was available. Although descriptive text is in French, the equivalent English meaning should be obvious. The cryptic description part of acceptable complete solutions (no variable excluded) was underlined to ease the present discussion. This is seen for lines 4, 7, 9 and 10. It is also seen in line 14, because the solution with 7 factors (marked by ‘7f’) will help decide between scenarios that differ by which variables of a pairwise grouping were considered multifactorial (e.g., ‘mf:6,7’). Preference is a priori directed toward fewer factors satisfactorily explaining all variables.

Figure 1. Matlab use of SCRoF with the synthetic data for N=2000.

>> AA=SCRoF(dat);

NEST indique au moins 5 facteurs

Scénarios explorés:

1: p=0.00623 X2(94)=131.739 VG1 FC1 7f initial VG0

2: p=0.13522 X2(94)=109.202 VG2 FC1 6f Grappes VG1

3: p=0.01321 X2(93)=125.851 VG3 FC1 7f Grappes VG1

4: p=0.13522 X2(94)=109.202 VG6 FC1 6f mf:6,7 Coplan VG1

5: p=0.07517 X2(95)=115.481 VG7 FC1 6f mf:8,9 Coplan VG1

6: p=0.21849 X2(93)=103.297 VG9 FC1 6f mf:6,7 Coplan VG3

7: p=0.13007 X2(94)=109.566 VG10 FC1 6f mf:8,9 Coplan VG3

8: p=0.00665 X2(93)=130.159 VG11 FC1 7f MultiSatur VG1

9: p=0.01250 X2(92)=125.044 VG13 FC1 7f MultiSatur VG3

10: p=0.08368 X2(94)=113.478 VG15 FC1 6f mf:8,9 MultiSatur VG7

11: p=0.46578 X2(92)=92.499 VG18 FC1 6f mf:6,7 MultiSatur VG9

12: p=0.29095 X2(93)=100.016 VG19 FC1 6f mf:8,9 MultiSatur VG10

13: p=0.14324 X2(93)=107.577 VG20 FC1 6f mf:8,9 MultiSatur VG10

14: p=0.32924 X2(92)=97.435 VG23 FC1 6f mf:8,9 MultiSatur VG19

The first three scenarios were produced before coplanar testing. Scenario #1 uses all seven initial clusters, including that of v6 with v7. This scenario is useful whenever coplanarity is detected and at least two of the clusters consist of a pair of variables, as it allows to see the correlations among all clusters considered as factors. Scenario #2 differs from #1 by considering the clustering of v6 with v7 as not significant; these two variables therefore are considered multifactorial. This scenario turns out identical to #4, in which the clustering of v6 with v7 was considered significant but their cluster is excluded as part of a coplanar trio. Scenario #3 implies one extra parameter (has one less degree of freedom) by virtue of variable v1 not clustering with variables v2 and v3. Its solution has an extra loading of .0503 on the factor of v15 and v16. Although the difference in χ2 with #1 of 5.89 for the loss of one degree of freedom indicates significant fit improvement, the equivalent scenarios #2 and #4 have a much better χ2 fit.

It may be observed in Figure 1 that the six factor variants involved only v6, v7 or v8, v9 even though v4, v5 also constitute a two-variable grouping with which the former two groups were found colinear. A scenario with the latter as multifactorial variables was also followed, but the loadings of v5 were limited to their sum of squares (communality) of .98. The resulting model χ2 fit can be seen in the output structure, here named AA, as χ2(95)=264, with *p* calculated as essentially null.

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 chi2inv(.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 χ2 value.

Variable clustering into factors. The χ2 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 χ2, 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 χ2 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 χ2 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 χ2 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 χ2, 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 χ2. 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 χ2 criterion, itself squared for more effective weighting, thus presuming that better signal cancellation provides better loading estimates.

Table 1.   
Left: The challenging structure of six factors, including two doublet factors and two orphan variables (with null loadings on all common factors).   
Right: The sparse SCRoF solution for a sample of N=2000

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable rank | Common factor | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | |
| 1 | **.5** | 0 | 0 | 0 | 0 | 0 | |
| 2 | **.6** | 0 | 0 | 0 | 0 | 0 | |
| 3 | **.55** | 0 | 0 | 0 | 0 | 0 | |
| 4 | 0 | **.5** | 0 | 0 | 0 | 0 | |
| 5 | 0 | **.6** | 0 | 0 | 0 | 0 | |
| 6 | 0 | **.5** | **.75** | 0 | 0 | 0 | |
| 7 | 0 | **-.4** | **-.6** | 0 | 0 | 0 | |
| 8 | 0 | 0 | **.5** | 0 | 0 | 0 | |
| 9 | 0 | 0 | **.6** | 0 | 0 | 0 | |
| 10 | 0 | 0 | 0 | **.5** | 0 | 0 | |
| 11 | 0 | 0 | 0 | **.6** | 0 | 0 | |
| 12 | 0 | **.4** | **.6** | **.4** | 0 | 0 | |
| 13 | 0 | 0 | 0 | 0 | **.5** | 0 | |
| 14 | 0 | 0 | 0 | 0 | **.8** | 0 | |
| 15 | 0 | 0 | 0 | 0 | 0 | **.5** | |
| 16 | 0 | 0 | 0 | 0 | 0 | **.8** | |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 18 | 0 | 0 | 0 | 0 | 0 | 0 | |
| Variable rank | Common factor | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | **.54** | 0 | 0 | 0 | 0 | 0 |
| 2 | **.58** | 0 | 0 | 0 | 0 | 0 |
| 3 | **.54** | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | **.47** | 0 | 0 | 0 | 0 |
| 5 | 0 | **.58** | 0 | 0 | 0 | 0 |
| 6 | 0 | **.47** | **.76** | 0 | 0 | 0 |
| 7 | 0 | **-.36** | **-.62** | 0 | 0 | 0 |
| 8 | 0 | 0 | **.53** | 0 | 0 | 0 |
| 9 | 0 | 0 | **.61** | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | **.45** | 0 | 0 |
| 11 | 0 | 0 | 0 | **.59** | 0 | 0 |
| 12 | 0 | **.39** | **.59** | **.41** | 0 | 0 |
| 13 | 0 | 0 | 0 | 0 | **.45** | 0 |
| 14 | 0 | 0 | 0 | 0 | **.81** | 0 |
| 15 | 0 | 0 | 0 | 0 | 0 | **.61** |
| 16 | 0 | 0 | 0 | 0 | 0 | **.61** |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 2.  
Left: Arbitrary correlations applied to the factors. Right: Factor correlations recovered by SCRoF.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | F1 | F2 | F3 | F4 | F5 | F6 |
| F1 | 1 | .4 | -.3 | 0 | .4 | 0 |
| F2 | .4 | 1 | 0 | -.3 | .3 | 0 |
| F3 | -.3 | 0 | 1 | 0 | -.5 | 0 |
| F4 | 0 | -.3 | 0 | 1 | 0 | 0 |
| F5 | .4 | .3 | -.5 | 0 | 1 | 0 |
| F6 | 0 | 0 | 0 | 0 | 0 | 1 |
|  | F1 | F2 | F3 | F4 | F5 | F6 |
| F1 | 1 | .48 | -.27 | 0 | .35 | 0 |
| F2 | .48 | 1 | 0 | -.29 | .29 | 0 |
| F3 | -.27 | 0 | 1 | 0 | -.50 | 0 |
| F4 | 0 | -.29 | 0 | 1 | 0 | 0 |
| F5 | .35 | .29 | -.50 | 0 | 1 | 0 |
| F6 | 0 | 0 | 0 | 0 | 0 | 1 |

For the present data, the maximal absolute correlation was above .25 for the first 16 variables (all net *p* < 1e-15) and respectively 0.064 and 0.033 (net *p* = 0.0669 and 0.9292), so that these two variables were declared orphan by both criteria and given null loadings on all factors.

Following pairwise signal cancellation attempts, hierarchical clustering of the complete type is applied to form groupings of variables each capable of cancelling the signal of the others in the same cluster. Although this is based on the χ2 values acting as distances between variables, the limit of allowed clusters is set by transforming the maximum across-clusters χ2 into a net probability, to account for the number of pairs of between-cluster variables involved.

Having identified clusters of unifactorial indicators of each factor, the correlation between pairs of clusters (potential factors) is 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. Statistical testing and nullifying of non-significant correlations, correcting for the number of tests, are postponed to a later step, as the number of factors is not definitively set.

Indeed, when at least three clusters were retained, SCRoF tests all trios of clusters for possible coplanarity. This involves taking each combination of individual variables from each cluster, weighting the first two to cancel the signal of the third one. The worse of all the associated cancellation χ2 is transformed into a net probability that is compared with the two statistical thresholds. The result, for any trio of clusters, imply either that the three clusters share one plane, that they constitute a three-dimensional space, or both in parallel scenarios. For each scenario admitting coplanarity of a trio of clusters, the most orthogonal pair is retained as two bona fide factors. As already mentioned, the variables of the remaining cluster become muti-factorial indicators informed by these two factors.

If variables remain that did not cluster for lack of pairwise cancellation of their signal, SCRoF proceeds with attempts at cancelling their signal by pairs, then triplets, etc., of variables exclusive to their respective factor. Here also, all combinations of variables representing their respective factors are used for signal cancellation, although this is primarily to obtain all possible loadings estimates of the target multi-factor variable on its informing factors. Each loading is obtained from the average of its various estimates, weighted by the inverse of the associated χ2 to give more weight to estimates associated with better signal cancellation.

Deciding on significant factor correlations is based on averaging all *k* pairwise correlations between variables exclusive to each involved factor. The variance of this average is thus 1/k the variance of individual correlations, which is 1/(N-1). A Bonferroni correlation for the total number of correlations tested is also applied and the resulting probabilities under the null hypothesis are compared to the two statistical thresholds. For each clearly significant mean correlation of variables, the corresponding factor correlation (i.e., adjusted for the variable loadings on their respective factors) is retained. For each clearly not significant mean correlation of variables, the factor correlation is set to 0. All combinations of probability in-between the two thresholds give rise to parallel scenario branches.

Each scenario is the combination of three branching levels, respectively relative to kept variables (KV), i.e. after exclusion, or not, of orphan variables, to variable grouping (VG), from parallel clustering decisions or exclusion of coplanar clusters, and to significant factor correlations (FC). Each scenario is expressed in a sparse pattern factor loading matrix and in the factor correlation matrix possibly with some null values. Even though the parameters of each scenario were obtained through separate non-linear optimizations, the credibility of each scenario is assessed by the same χ2 fit index as in CFA (Jöreskog, 1967), where non-zero scenario parameters determine the associated degrees of freedom. This constitutes an initial assessment of the scenarios. The parameters of any scenario may then be refined by transmitting its pattern and correlation matrices to a SEM application. For the upcoming illustrations using Matlab code, this was done manually using Onyx (Ref). The upcoming R version will automatically call Lavaan (Ref).

Illustrations

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, a so-called Heywood case. The purpose of having two 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, although SEM refinement of the solution parameters may make these unconstrained loadings equal. The presence of doublet factor indeterminacy may thus be judged from the factor correlations.

>> for k=1:5,k,SCRoFreport(aa(k));end

k =

1

Scénarios explorés:

1: p=0.93717 X2(47)=33.129 VG:22 FC:1 4f mf:8,9 excluant v14 v-15 v-16 v-13

2: p=0.40404 X2(48)=49.735 VG:22 FC:2

3: p=0.92293 X2(47)=33.951 VG:21 FC:1 4f mf:6,7 excluant v14 v-15 v-16 v-13

4: p=0.00840 X2(48)=74.526 VG:21 FC:2

5: p=0.89335 X2(47)=35.377 VG:14 FC:1 5f excluant v14 v15 v16 v-13

6: p=0.34117 X2(48)=51.425 VG:14 FC:2

7: p=0.00224 X2(49)=81.873 VG:14 FC:3

8: p=0.00016 X2(50)=94.081 VG:14 FC:4

9: p=0.89246 X2(70)=55.753 VG:18 FC:1 5f mf:8,9 excluant v14 v-13

10: p=0.43280 X2(71)=72.359 VG:18 FC:2

11: p=0.14042 X2(72)=84.992 VG:18 FC:3

12: p=0.87682 X2(70)=56.575 VG:17 FC:1 5f mf:6,7 excluant v14 v-13

13: p=0.02139 X2(71)=97.150 VG:17 FC:2

14: p=0.00024 X2(72)=121.563 VG:17 FC:3

15: p=0.84646 X2(70)=58.001 VG:13 FC:1 6f excluant v14 v-13

16: p=0.37900 X2(71)=74.049 VG:13 FC:2

17: p=0.00742 X2(72)=104.496 VG:13 FC:3

18: p=0.00088 X2(73)=116.705 VG:13 FC:4

19: p=0.80068 X2(55)=46.013 VG:11 FC:1 4f mf:6,7 excluant v15 v16 v13

20: p=0.00494 X2(56)=87.049 VG:11 FC:2

21: p=0.59687 X2(94)=90.025 VG:6 FC:1 6f mf:8,9

22: p=0.21632 X2(95)=105.520 VG:6 FC:2

23: p=0.05467 X2(96)=119.171 VG:6 FC:3

24: p=0.56640 X2(94)=91.069 VG:5 FC:1 6f mf:6,7

25: p=0.00735 X2(95)=131.900 VG:5 FC:2

26: p=0.55766 X2(82)=79.500 VG:2 FC:1 6f excluant v13

27: p=0.22938 X2(83)=92.204 VG:2 FC:2

28: p=0.50204 X2(93)=92.265 VG:1 FC:1 7f

29: p=0.15578 X2(94)=107.843 VG:1 FC:2

30: p=0.00248 X2(95)=138.344 VG:1 FC:3

31: p=0.00017 X2(96)=153.645 VG:1 FC:4

32: p=0.39123 X2(80)=82.861 VG:8 FC:1 5f mf:6,7 excluant v13

33: p=0.01498 X2(81)=111.055 VG:8 FC:2

k =

2

Scénarios explorés:

1: p=0.47867 X2(57)=56.904 VG:3 FC:1 5f excluant v12 v10 v11

2: p=0.46083 X2(58)=58.393 VG:3 FC:2

3: p=0.47770 X2(24)=23.720 VG:4 FC:1 4f excluant v10 v11 v12 v6 v7 v8 v9

4: p=0.22907 X2(81)=90.102 VG:11 FC:2

5: p=0.22215 X2(80)=89.357 VG:11 FC:1 6f mf:8,9 excluant v12

6: p=0.09055 X2(82)=99.592 VG:11 FC:3

7: p=0.01366 X2(83)=113.968 VG:11 FC:4

8: p=0.14865 X2(80)=93.187 VG:9 FC:2

9: p=0.13328 X2(79)=93.065 VG:9 FC:1 7f excluant v12

10: p=0.01322 X2(81)=111.825 VG:9 FC:3

11: p=0.00603 X2(82)=117.664 VG:9 FC:4

12: p=0.04060 X2(94)=119.212 VG:7 FC:2

13: p=0.03489 X2(93)=119.200 VG:7 FC:1 6f mf:6,7

14: p=0.00083 X2(95)=144.353 VG:7 FC:3

15: p=0.03459 X2(94)=120.394 VG:2 FC:2

16: p=0.03342 X2(93)=119.512 VG:2 FC:1 6f

17: p=0.00083 X2(95)=144.319 VG:2 FC:3

18: p=0.03601 X2(94)=120.100 VG:8 FC:2

19: p=0.03059 X2(93)=120.148 VG:8 FC:1 6f mf:8,9

20: p=0.01174 X2(95)=128.944 VG:8 FC:3

21: p=0.00045 X2(96)=148.790 VG:8 FC:4

22: p=0.01630 X2(93)=124.476 VG:1 FC:2

23: p=0.01435 X2(92)=124.154 VG:1 FC:1 7f

24: p=0.00135 X2(94)=140.492 VG:1 FC:3

25: p=0.00016 X2(95)=152.650 VG:1 FC:4

k =

3

Scénarios explorés:

1: p=0.39656 X2(94)=96.961 VG:2 FC:1 6f

2: p=0.08732 X2(95)=114.216 VG:2 FC:2

3: p=0.00747 X2(96)=132.980 VG:2 FC:3

4: p=0.37261 X2(93)=96.818 VG:3 FC:1 6f

5: p=0.07684 X2(94)=114.193 VG:3 FC:2

6: p=0.00624 X2(95)=132.909 VG:3 FC:3

7: p=0.17261 X2(94)=106.823 VG:1 FC:1 7f

8: p=0.05626 X2(95)=117.829 VG:1 FC:2

9: p=0.00570 X2(96)=134.639 VG:1 FC:3

k =

4

Scénarios explorés:

1: p=0.92602 X2(94)=74.948 VG:3 FC:1 6f mf:6,7

2: p=0.34941 X2(95)=99.746 VG:3 FC:2

3: p=0.92565 X2(94)=74.979 VG:4 FC:1 6f mf:8,9

4: p=0.41632 X2(95)=97.265 VG:4 FC:2

5: p=0.04653 X2(96)=120.428 VG:4 FC:3

6: p=0.91470 X2(93)=74.954 VG:1 FC:1 7f

7: p=0.23485 X2(94)=103.557 VG:1 FC:2

8: p=0.01566 X2(95)=127.056 VG:1 FC:3

9: p=0.77477 X2(95)=84.351 VG:2 FC:1 6f mf:4,5

10: p=0.02997 X2(96)=123.706 VG:2 FC:2

k =

5

Scénarios explorés:

1: p=0.88068 X2(58)=45.636 VG:7 FC:1 5f mf:6,7 excluant v12 v10 v11

2: p=0.17698 X2(59)=68.918 VG:7 FC:2

3: p=0.85746 X2(58)=46.647 VG:8 FC:1 5f mf:8,9 excluant v12 v10 v11

4: p=0.85094 X2(94)=79.839 VG:4 FC:1 6f mf:6,7

5: p=0.26814 X2(95)=103.080 VG:4 FC:2

6: p=0.02604 X2(96)=124.710 VG:4 FC:3

7: p=0.83189 X2(95)=81.742 VG:5 FC:1 6f mf:8,9

8: p=0.26822 X2(96)=104.122 VG:5 FC:2

9: p=0.71271 X2(94)=85.878 VG:1 FC:1 7f

10: p=0.05757 X2(95)=117.647 VG:1 FC:2

11: p=0.00917 X2(96)=131.695 VG:1 FC:3

>> k=1;aa(k)=SCRoF(dat(1000\*k-999:1000\*k,:));

% following warning issued 21 times

Exiting: Maximum number of function evaluations has been exceeded

- increase MaxFunEvals option.

Current function value: 11.003045

Scénarios explorés:

1: p=0.93717 X2(47)=33.129 VG:22 FC:1 4f mf:8,9 excluant v14 v-15 v-16 v-13

2: p=0.40404 X2(48)=49.735 VG:22 FC:2 4f

3: p=0.92293 X2(47)=33.951 VG:21 FC:1 4f mf:6,7 excluant v14 v-15 v-16 v-13

4: p=0.00840 X2(48)=74.526 VG:21 FC:2 4f

5: p=0.89335 X2(47)=35.377 VG:14 FC:1 5f excluant v14 v15 v16 v-13

6: p=0.34117 X2(48)=51.425 VG:14 FC:2 5f

7: p=0.00224 X2(49)=81.873 VG:14 FC:3 5f

8: p=0.00016 X2(50)=94.081 VG:14 FC:4 5f

9: p=0.89246 X2(70)=55.753 VG:18 FC:1 5f mf:8,9 excluant v14 v-13

10: p=0.43280 X2(71)=72.359 VG:18 FC:2 5f

11: p=0.14042 X2(72)=84.992 VG:18 FC:3 5f

12: p=0.87682 X2(70)=56.575 VG:17 FC:1 5f mf:6,7 excluant v14 v-13

13: p=0.02139 X2(71)=97.150 VG:17 FC:2 5f

14: p=0.00024 X2(72)=121.563 VG:17 FC:3 5f

15: p=0.84646 X2(70)=58.001 VG:13 FC:1 6f excluant v14 v-13

16: p=0.37900 X2(71)=74.049 VG:13 FC:2 6f

17: p=0.00742 X2(72)=104.496 VG:13 FC:3 6f

18: p=0.00088 X2(73)=116.705 VG:13 FC:4 6f

19: p=0.80068 X2(55)=46.013 VG:11 FC:1 4f mf:6,7 excluant v15 v16 v13

20: p=0.00494 X2(56)=87.049 VG:11 FC:2 4f

21: p=0.59687 X2(94)=90.025 VG:6 FC:1 6f mf:8,9

22: p=0.21632 X2(95)=105.520 VG:6 FC:2 6f

23: p=0.05467 X2(96)=119.171 VG:6 FC:3 6f

24: p=0.56640 X2(94)=91.069 VG:5 FC:1 6f mf:6,7

25: p=0.00735 X2(95)=131.900 VG:5 FC:2 6f

26: p=0.55766 X2(82)=79.500 VG:2 FC:1 6f excluant v13

27: p=0.22938 X2(83)=92.204 VG:2 FC:2 6f

28: p=0.50204 X2(93)=92.265 VG:1 FC:1 7f

29: p=0.15578 X2(94)=107.843 VG:1 FC:2 7f

30: p=0.00248 X2(95)=138.344 VG:1 FC:3 7f

31: p=0.00017 X2(96)=153.645 VG:1 FC:4 7f

32: p=0.39123 X2(80)=82.861 VG:8 FC:1 5f mf:6,7 excluant v13

33: p=0.01498 X2(81)=111.055 VG:8 FC:2 5f

>> SCRoFreport(aa(1),24);

Fct:

1 0.586 0.000 0.000 0.000 0.000 0.000

2 0.620 0.000 0.000 0.000 0.000 0.000

3 0.581 0.000 0.000 0.000 0.000 0.000

4 0.000 0.508 0.000 0.000 0.000 0.000

5 0.000 0.640 0.000 0.000 0.000 0.000

6 0.000 0.477 0.739 0.000 0.000 0.000

7 0.000 -0.400 -0.582 0.000 0.000 0.000

8 0.000 0.000 0.497 0.000 0.000 0.000

9 0.000 0.000 0.562 0.000 0.000 0.000

10 0.000 0.000 0.000 0.526 0.000 0.000

11 0.000 0.000 0.000 0.581 0.000 0.000

12 0.000 0.382 0.616 0.390 0.000 0.000

13 0.000 0.000 0.000 0.000 0.487 0.000

14 0.000 0.000 0.000 0.000 0.723 0.000

15 0.000 0.000 0.000 0.000 0.000 0.638

16 0.000 0.000 0.000 0.000 0.000 0.615

17 0.000 0.000 0.000 0.000 0.000 0.000

18 0.000 0.000 0.000 0.000 0.000 0.000

fCorr:

1 2 3 4 5 6

1 1.000 0.401 -0.305 0.000 0.441 0.000

2 0.401 1.000 0.000 -0.287 0.328 0.000

3 -0.305 0.000 1.000 0.000 -0.512 0.000

4 0.000 -0.287 0.000 1.000 0.000 0.000

5 0.441 0.328 -0.512 0.000 1.000 0.000

6 0.000 0.000 0.000 0.000 0.000 1.000

>> SCRoFreport(aa(1),28);

Fct:

1 0.586 0.000 0.000 0.000 0.000 0.000 0.000

2 0.619 0.000 0.000 0.000 0.000 0.000 0.000

3 0.581 0.000 0.000 0.000 0.000 0.000 0.000

4 0.000 0.510 0.000 0.000 0.000 0.000 0.000

5 0.000 0.641 0.000 0.000 0.000 0.000 0.000

6 0.000 0.000 0.880 0.000 0.000 0.000 0.000

7 0.000 0.000 -0.706 0.000 0.000 0.000 0.000

8 0.000 0.000 0.000 0.501 0.000 0.000 0.000

9 0.000 0.000 0.000 0.568 0.000 0.000 0.000

10 0.000 0.000 0.000 0.000 0.524 0.000 0.000

11 0.000 0.000 0.000 0.000 0.580 0.000 0.000

12 0.000 0.000 0.728 0.000 0.399 0.000 0.000

13 0.000 0.000 0.000 0.000 0.000 0.487 0.000

14 0.000 0.000 0.000 0.000 0.000 0.723 0.000

15 0.000 0.000 0.000 0.000 0.000 0.000 0.648

16 0.000 0.000 0.000 0.000 0.000 0.000 0.604

17 0.000 0.000 0.000 0.000 0.000 0.000 0.000

18 0.000 0.000 0.000 0.000 0.000 0.000 0.000

fCorr:

1 2 3 4 5 6 7

1 1.000 0.417 0.000 -0.275 0.000 0.435 0.000

2 0.417 1.000 0.548 0.000 -0.287 0.336 0.000

3 0.000 0.548 1.000 0.824 -0.167 -0.244 0.000

4 -0.275 0.000 0.824 1.000 0.000 -0.495 0.000

5 0.000 -0.287 -0.167 0.000 1.000 0.000 0.000

6 0.435 0.336 -0.244 -0.495 0.000 1.000 0.000

7 0.000 0.000 0.000 0.000 0.000 0.000 1.000

>> for k=2:5,k,aa(k)=SCRoF(dat(1000\*k-999:1000\*k,:)); end

k = 2

% warning issued 4 times

Exiting: Maximum number of function evaluations has been exceeded

- increase MaxFunEvals option.

Current function value: 12.932554

Scénarios explorés:

1: p=0.47867 X2(57)=56.904 VG:3 FC:1 5f excluant v12 v10 v11

2: p=0.46083 X2(58)=58.393 VG:3 FC:2 5f

3: p=0.47770 X2(24)=23.720 VG:4 FC:1 4f excluant v10 v11 v12 v6 v7 v8 v9

4: p=0.22907 X2(81)=90.102 VG:11 FC:2 6f

5: p=0.22215 X2(80)=89.357 VG:11 FC:1 6f mf:8,9 excluant v12

6: p=0.09055 X2(82)=99.592 VG:11 FC:3 6f

7: p=0.01366 X2(83)=113.968 VG:11 FC:4 6f

8: p=0.14865 X2(80)=93.187 VG:9 FC:2 7f

9: p=0.13328 X2(79)=93.065 VG:9 FC:1 7f excluant v12

10: p=0.01322 X2(81)=111.825 VG:9 FC:3 7f

11: p=0.00603 X2(82)=117.664 VG:9 FC:4 7f

12: p=0.04060 X2(94)=119.212 VG:7 FC:2 6f

13: p=0.03489 X2(93)=119.200 VG:7 FC:1 6f mf:6,7

14: p=0.00083 X2(95)=144.353 VG:7 FC:3 6f

15: p=0.03459 X2(94)=120.394 VG:2 FC:2 6f

16: p=0.03342 X2(93)=119.512 VG:2 FC:1 6f

17: p=0.00083 X2(95)=144.319 VG:2 FC:3 6f

18: p=0.03601 X2(94)=120.100 VG:8 FC:2 6f

19: p=0.03059 X2(93)=120.148 VG:8 FC:1 6f mf:8,9

20: p=0.01174 X2(95)=128.944 VG:8 FC:3 6f

21: p=0.00045 X2(96)=148.790 VG:8 FC:4 6f

22: p=0.01630 X2(93)=124.476 VG:1 FC:2 7f

23: p=0.01435 X2(92)=124.154 VG:1 FC:1 7f

24: p=0.00135 X2(94)=140.492 VG:1 FC:3 7f

25: p=0.00016 X2(95)=152.650 VG:1 FC:4 7f

k =

3

Scénarios explorés:

1: p=0.39656 X2(94)=96.961 VG:2 FC:1 6f

2: p=0.08732 X2(95)=114.216 VG:2 FC:2 6f

3: p=0.00747 X2(96)=132.980 VG:2 FC:3 6f

4: p=0.37261 X2(93)=96.818 VG:3 FC:1 6f

5: p=0.07684 X2(94)=114.193 VG:3 FC:2 6f

6: p=0.00624 X2(95)=132.909 VG:3 FC:3 6f

7: p=0.17261 X2(94)=106.823 VG:1 FC:1 7f

8: p=0.05626 X2(95)=117.829 VG:1 FC:2 7f

9: p=0.00570 X2(96)=134.639 VG:1 FC:3 7f

k =

4

Scénarios explorés:

1: p=0.92602 X2(94)=74.948 VG:3 FC:1 6f mf:6,7

2: p=0.34941 X2(95)=99.746 VG:3 FC:2 6f

3: p=0.92565 X2(94)=74.979 VG:4 FC:1 6f mf:8,9

4: p=0.41632 X2(95)=97.265 VG:4 FC:2 6f

5: p=0.04653 X2(96)=120.428 VG:4 FC:3 6f

6: p=0.91470 X2(93)=74.954 VG:1 FC:1 7f

7: p=0.23485 X2(94)=103.557 VG:1 FC:2 7f

8: p=0.01566 X2(95)=127.056 VG:1 FC:3 7f

9: p=0.77477 X2(95)=84.351 VG:2 FC:1 6f mf:4,5

10: p=0.02997 X2(96)=123.706 VG:2 FC:2 6f

k =

5

Scénarios explorés:

1: p=0.88068 X2(58)=45.636 VG:7 FC:1 5f mf:6,7 excluant v12 v10 v11

2: p=0.17698 X2(59)=68.918 VG:7 FC:2 5f

3: p=0.85746 X2(58)=46.647 VG:8 FC:1 5f mf:8,9 excluant v12 v10 v11

4: p=0.85094 X2(94)=79.839 VG:4 FC:1 6f mf:6,7

5: p=0.26814 X2(95)=103.080 VG:4 FC:2 6f

6: p=0.02604 X2(96)=124.710 VG:4 FC:3 6f

7: p=0.83189 X2(95)=81.742 VG:5 FC:1 6f mf:8,9

8: p=0.26822 X2(96)=104.122 VG:5 FC:2 6f

9: p=0.71271 X2(94)=85.878 VG:1 FC:1 7f

10: p=0.05757 X2(95)=117.647 VG:1 FC:2 7f

11: p=0.00917 X2(96)=131.695 VG:1 FC:3 7f

1. Spearman (1904) discussed that the correlation of two observed variables systematically underestimates the correlation between the “true objective value” (‘signal’ in this manuscript) of the variables each measured with error. He proposed two ways to correct the attenuation, both based on the correlations of independent replications of each variable. He further illustrated that each approach brings the corrected correlation of the true objective value of the two variables to 1.0, which is to be expected in absence of noise (although not when one variable reflects a unique source of variance besides the common one). [↑](#footnote-ref-1)
2. As a rule, expected null values divided by their standard error become *z* scores. For expected null correlations, the standard error is 1/√(*N*-1), yielding *z* = *r* √(*N*‑1). Also, a squared *z* score, notably here *z*2 = (*N*-1) *r*2, is a χ2(1) (i.e., chi-square with one degree of freedom). The sum of *k* independent χ2(1) is distributed as χ2(*k*) under the null hypothesis. Here *k* would be the number of correlations of the optimized signal cancellation contrast with the remaining variables, i.e., *k* = 8 for pairwise signal cancellation attempts within a dataset of 10 variables. [↑](#footnote-ref-2)
3. Geometrically, signal cancellation amounts to reaching the signal vector of *V* by combining suitable lengths along the two factors that inform *V*, and then subtracting *V* from the weighted sum of the variables whose respective signal has the direction of their own factor. [↑](#footnote-ref-3)
4. For signal cancellation within a pair of normalized variables *A* and *B*, let their respective loadings be *a* and *b*. The expected *rAB* 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 *rAB*=*ab*, one gets *rAB*=w*a*2, *a*=√(*rAB* /*w*), and *b*= *rAB /a*. For a variable *V* whose signal is cancelled by variables *B* and *G*, already established to load *b* and *g* on their respective factors, the successful contrast is *wbB*+*wgG*-*V*, such that the loadings are obtained by *wbb* and *wgg*. This is because the optimal weights act by scaling the signal parts of the two cancelling variables. [↑](#footnote-ref-4)
5. Placing the *observed* cross correlations of the variables representing the two factors in a vector *O* and the corresponding *products* of loadings in a vector *P*, the factor correlation *r* minimizes the sum of squares of O-*rP*. [↑](#footnote-ref-5)
6. For say five factors, there would be 10 correlations. The most significant observed probability, *p*10, is transformed into 1-(1-*p*10)10. The next most significant probability becomes 1-(1-*p*9)9, and so on down to a not significant *p* value. All following tests are automatically also declared not significant. [↑](#footnote-ref-6)
7. As an exception, the initial solution with all acceptable clusters is always presented first, irrespective of its fit value to allow users to examine the correlations among all possible clusters. [↑](#footnote-ref-7)