Psychology Science, Volume 45, 2003 (2), p. 330 - 338

Type structures in CFA

JOACHIM KRAUTH[1]

Summary

Configural Frequency Analysis (CFA) is a procedure in data analysis which was introduced by G.A. Lienert (1969). This procedure is based on a specific statistical definition of types. Due to this definition results can be obtained which may, at first glance, contradict intuition. In particular, for small contingency tables it is observed that only a small percentage of imaginable type structures, i.e. patterns of types, antitypes and others, is actually possible. Here, we try to give bounds for the percentage of possible type structures and discuss consequences for the performance of a CFA.

Key words: Type structure, Configural Frequency Analysis (CFA)

[1] Prof. Dr. J. Krauth, Department of Experimental Psychology, University of Düsseldorf, D-40225 Düsseldorf, Germany; E-mail: J.Krauth@uni-duesseldorf.de

1. Types and Antitypes in CFA

A random sample of subjects is selected from a population and for each of these subjects it is observed which values are obtained by $t \geq 2$ nominal scaled variables $X_1,...,X_t$. The variable X_i can take the $r_i \geq 2$ unordered categories $(C_{i1},...,C_{ir_i})$ as possible values, for i=1,...,t. Thus, each subject produces a t-dimensional vector $(C_{1j_1},...,C_{tj_t})$ of categories with $1 \leq j_i \leq r_i, i = 1, ..., t.$ Instead of this explicit notation we will rather use the short version $(j_1,..., j_t)$. Lienert (1969) denoted the vector $(j_1,..., j_t)$ as a *configuration*. The number of subjects $(f_{j_1 \cdots j_t})$ in the random sample which exhibit the configuration $(j_1,..., j_t)$ is denoted by Lienert (1969) as a *configural frequency*. In case of high configural frequencies which cannot be explained by a random combination of such categories which occur all with high probabilities, Lienert (1969) assumes the possible existence of *types* of subjects. Such *types* are to be identified by means of the so-called *configural frequency analysis* (C*FA*).

Let $\pi_{j_1 \cdots j_t}$ be the probability of the occurrence of the configuration $(j_1,...,j_t)$ if a subject is randomly sampled from the population, for j_i = 1,...,r_i , i = 1,...,t. Altogether

$$(1) \qquad r = \prod_{i=1}^{t} r_i$$

different configurations are possible. Because to each subject in the population corresponds exactly one of the r possible configurations, the sum of the r *configural probabilities* $\pi_{j_1 \cdots j_t}$ must yield 1:

$$(2) \qquad \sum_{j_1=1}^{r_1} \cdots \sum_{j_t=1}^{r_t} \pi_{j_1 \cdots j_t} = 1.$$

If we fix category j_i for variable X_i and sum over all categories of the other variables, we derive the corresponding onedimensional marginal probability

$$(3) \qquad \pi_{\dots j_i \dots} = *_i \sum_{j_1=1}^{r_1} \cdots \sum_{j_t=1}^{r_t} \pi_{j_1 \cdots j_t} \quad \text{for} \quad j_i = 1, ..., r_i, i = 1, ..., t.$$

where $*_i$ means, that the sum $\sum_{j_i=1}^{r_i}$ is not considered in the summation. The marginal probability $\pi_{\dots j_i \dots}$ is the probability that a subject which is randomly selected from the population exhibits the category C_{ij_i} for variable X_i irrespective of the categories which this subject exhibits for the other variables.

The concept of types or antitypes, respectively, which was formulated by Lienert (1969) can be formalized in the following way. We consider the differences (so-called *residuals*)

$$(4) \qquad \delta_{j_1 \cdots j_t} = \pi_{j_1 \cdots j_t} - \prod_{s=1}^{t} \pi_{\dots j_s \dots} \quad \text{for } j_i = 1, ..., r_i, i = 1, ..., t.$$

Here, we subtract from the configural probability $\pi_{j_1 \cdots j_t}$ that probability for the configuration $(j_1,..., j_t)$ which would result if the t variables $X_1,...,X_t$ would be independent, i.e. if the t categories $C_{j_1},...,C_{j_t}$ would be randomly combined to form a configuration.

For

$$(5) \qquad \delta_{j_1 \cdots j_t} > 0$$

a *type* (T) is present according to CFA.

For

(6) $$\delta_{j_1 \cdots j_t} < 0$$

an *antitype* (A) is present according to CFA.
Finally, for

(7) $$\delta_{j_1 \cdots j_t} = 0$$

neither a type nor an antitype is present (0) according to CFA.
By summation we get

(8) $$\sum_{j_1=1}^{r_1} \cdots \sum_{j_t=1}^{r_t} \delta_{j_1 \cdots j_t} = 0$$

and

(9) $$*_i \sum_{j_1=1}^{r_1} \cdots \sum_{j_t=1}^{r_t} \delta_{j_1 \cdots j_t} = 0 \text{ for } j_i = 1, \ldots, r_i, i = 1, \ldots, t.$$

Types and antitypes are identified by means of CFA in the following way: For each of the r configurations $(1, \ldots, 1), \ldots, (r_1, \ldots, r_t)$ two statistical tests are performed for finding out which of the three possible situations (T, A or 0) has the highest evidence. We denote an r-dimensional vector with components T, A or 0, as it is identified by CFA, as a *type structure*. Thus, with r configurations each with three possible outcomes (T, A or 0), the maximum number of different type structures is given by

(10) $$M = 3^r.$$

From condition (8) we can conclude that the two type structures (T, …, T), i.e. r types, or (A, …, A), i.e. r antitypes, respectively, are not possible, because a sum consisting only of positive numbers or only of negative numbers, respectively, is not possible because such a sum cannot yield 0. However, a structure of the form (0, …, 0), where neither types nor antitypes are present, is possible.

If we consider in addition the $r_1 + \ldots + r_t$ conditions (9), it is obvious that the number of possible type structures must be smaller than M. In order to identify a type structure by means of CFA, altogether $(2r)$ dependent statistical tests for types and antitypes have to be performed simultaneously. Since this requires an alpha adjustment, it would be interesting to investigate whether there exist situations where we need fewer tests for identifying a type structure, if we know which type structures are actually possible.

2. Type Structures in a Fourfold Scheme

In the case of a fourfold scheme with $t = 2, r_1 = r_2 = 2$, which was already discussed by Krauth (1993, pp. 25 – 28) we get $r = 4$ and the 3 possible type structures given in Table 1.

Table 1:
Possible type structures for a 2×2 scheme.

$j_1\ j_2$	$\pi_{j_1 j_2}$	$\delta_{j_1 j_2}$		$\varepsilon_1 > 0$	$\varepsilon_1 < 0$	$\varepsilon_1 = 0$
1 1	π_{11}	$\delta_{11} = \pi_{11} - \pi_{1.}\pi_{.1}$	ε_1	T	A	0
1 2	π_{12}	$\delta_{12} = \pi_{12} - \pi_{1.}\pi_{.2}$	$-\varepsilon_1$	A	T	0
2 1	π_{21}	$\delta_{21} = \pi_{21} - \pi_{2.}\pi_{.1}$	$-\varepsilon_1$	A	T	0
2 2	π_{22}	$\delta_{22} = \pi_{22} - \pi_{2.}\pi_{.2}$	ε_1	T	A	0

From (9) we get

(11) $$\delta_{11} + \delta_{12} = 0, \delta_{21} + \delta_{22} = 0, \delta_{11} + \delta_{21} = 0, \delta_{12} + \delta_{22} = 0.$$

If we choose the reparameterization $\varepsilon_1 = \delta_{11}$, we get $\delta_{12} = -\varepsilon_1, \delta_{21} = -\varepsilon_1, \delta_{22} = \varepsilon_1,$, i.e. all possible type structures are determined by only one parameter (ε_1). Depending on whether ε_1 is positive, negative or equal to 0, we get one of the 3 possible type structures (T, A, A, T), (A, T, T, A) or (0, 0, 0, 0). Instead of the maximal number of M = 81 structures only 3 structures are possible. Thus, it is possible to identify the inherent type structure by performing only 2 tests (e.g. tests for type and antitype with respect to the configuration (1, 1)) instead of performing 2r = 8 tests as in the usual procedure.

The number of parameters to be considered (in this case: 1, i.e. ε_1) corresponds to the number (F) of degrees of freedom as given in Lienert (1969). We have

(12) $$F = \prod_{i=1}^{t} r_i - \sum_{i=1}^{t} r_i + t - 1,$$

in the general case and

(13) $$F = 2^t - t - 1$$

in the special case with r_1 =...= r_t= 2. Thus, we get for the fourfold scheme with $t = 2, r_1 = r_2 = 2$ the value F = 1.

Since ε_1 is the difference of 2 probabilities, this parameter cannot be chosen arbitrarily large or small. In view of

(14) $$\max\ \{0, \pi_{.1} - \pi_{2.}\} \le \pi_{11} \le \min\ \{\pi_{1.}, \pi_{.1}\}$$

following from the definition of onedimensional marginal probabilities we get for

(15) $$\varepsilon_1 = \pi_{11} - \pi_{1.}\pi_{.1}$$

the inequalities

(16) $$\max\ \{0, \pi_{.1} - \pi_{2.}\} - \pi_{1.}\pi_{.1} \le \varepsilon_1 \le \min\ \{\pi_{1.}, \pi_{.1}\} - \pi_{1.}\pi_{.1}.$$

The value of ε_1 may be chosen arbitrarily within the interval defined by (16). However, if at least one of the marginal probabilities is equal to 0, the interval degenerates to a single point (ε_1 = 0) In this case we must choose ε_1 = 0 and only the structure (0, 0, 0, 0) is possible.

At first glance it may astonish that only the 3 type structures given in Table 1 should be possible for a fourfold scheme. Obviously, it is possible, that, e.g., the configurations (1, 1) and (1, 2) are exhibited by many subjects while the configurations (2, 1) and (2, 2) are exhibited only by few subjects. Assume for example that the configural probabilities are given by $\pi_{11} = .40, \pi_{12} = .50, \pi_{21} = .03$ and $\pi_{22} = .07$. In this case we might be inclined to identify (1, 1) and (1, 2) as obvious types and (2, 1) and (2, 2) as obvious antitypes. However, this interpretation is not in accordance with the definition of a type by Lienert (1969) because we have the marginal probabilities

$$\pi_{1.} = .90, \pi_{2.} = .10, \pi_{.1} = .43, \pi_{.2} = .57$$

and as a consequence

$$\delta_{11} = .40 - .90 \times .43 = .013, \delta_{12} = .50 - .90 \times .57 = -.013,$$
$$\delta_{21} = .03 - .10 \times .43 = -.013, \delta_{22} = .07 - .10 \times .57 = .013.$$

Thus, we find the antitypes (1, 2) and (2, 1) and the types (1, 1) and (2, 2). The configuration (1, 2) would constitute an antitype in spite of the high configural probability $\pi_{12} = .50$ because π_{12} is smaller than the probability $.90 \times .57 = .513$ in case of a random combination of the categories. By a similar argument we find that (2, 2) is a type in spite of the small configural probability ($\pi_{22} = .07$) because π_{22} is larger than the probability $.10 \times .57 = .057$ in case of a random combination of the categories. In this example, we get the type structure (T, A, A, T) which was shown to be one of the three possible structures.

3. Type Structures in Other Simple Situations

For $t = 2, r_1 = 2, r_2 = 3$ we get $r = 6, F = 2$ and the possible type structures depicted in Table 2. Here, the reparameterization $\varepsilon_1 = \delta_{11}, \varepsilon_2 = \delta_{23}$ was chosen, and altogether 13 possible type structures result compared with a maximal number of $M = 729$ structures.

Table 2:
Possible type structures for a 2×3 scheme.

1 1	ε_1		T	T	A	T	A	A	0	0	0	A	A	T	T
1 2	$-\varepsilon_1$	$+\varepsilon_2$	T	A	T	A	T	A	0	A	T	0	T	0	A
1 3		$-\varepsilon_2$	A	T	T	A	A	T	0	T	A	T	0	A	0
2 1	$-\varepsilon_1$		A	A	T	A	T	T	0	0	0	T	T	A	A
2 2	ε_1	$-\varepsilon_2$	A	T	A	T	A	T	0	T	A	0	A	0	T
2 3		ε_2	T	A	A	T	T	A	0	A	T	A	0	T	0

With the usual procedure 12 simultaneous tests for types and antitypes had to be performed in order to identify the structure. However, if one has identified the type structure of the first 3 configurations by means of 6 tests, the total structure is known because the possible structures of the second 3 configurations follow from those of the first 3 by means of the transformation T→ A, A →T, 0 → 0. However, in this case we have only a saving of 50% in comparison to the fourfold table where we had a saving of 75%.

For t = 2, r_1 = 2, r_2 = 4, we get r = 8, F = 3. One possible reparameterization is given in Table 3. Here, 51 possible type structures result in comparison with the maximal number of M = 6561 structures. With the usual procedure 16 simultaneous tests are to be performed to identify the type structure. However if the identification has been performed for the first 4 configurations by means of 8 tests, the identification for the remaining configurations is possible via the transformation T→ A, A →T, 0 → 0.

Table 3:
Reparameterization for a 2×4 scheme.

1 1	ε_1		
1 2	$-\varepsilon_1$	$+\varepsilon_2$	
1 3		$-\varepsilon_2$	$+\varepsilon_3$
1 4			$-\varepsilon_3$
2 1	$-\varepsilon_1$		
2 2	ε_1	$-\varepsilon_2$	
2 3		ε_2	$-\varepsilon_3$
2 4			ε_3

For t = 2, r_1 = r_2 = 3 we get r = 9, F = 4. A possible reparameterization is given in Table 4. Here we have 493 possible type structures in comparison with a maximal number of M = 19683. In this case, no tests are saved in comparison with the usual procedure.

Table 4:
Reparameterization for a 3×3 scheme.

1 1	ε_1			
1 2	$-\varepsilon_1$	$+\varepsilon_2$		
1 3		$-\varepsilon_2$		
2 1	$-\varepsilon_1$		$+\varepsilon_3$	
2 2	ε_1	$-\varepsilon_2$	$-\varepsilon_3$	$+\varepsilon_4$
2 3		ε_2		$-\varepsilon_4$
3 1			$-\varepsilon_3$	
3 2			ε_3	$-\varepsilon_4$
3 3				ε_4

For t = 3, r_1 = r_2 = r_3 = 2 we get r = 8, F = 4. A possible reparameterization is given in Table 5. Here, we have 985 possible type structures in comparison with a maximal number of M = 6561. In this case, again no tests are saved in comparison with the usual procedure.

Table 5:
Reparameterization for a $2\times2\times2$ scheme.

1 1 1	ε_1			
1 1 2	$-\varepsilon_1$			$+\varepsilon_4$
1 2 1	$-\varepsilon_1$	$+\varepsilon_2$		$-\varepsilon_4$
1 2 2	ε_1	$-\varepsilon_2$		
2 1 1		$-\varepsilon_2$	$+\varepsilon_3$	
2 1 2		ε_2	$-\varepsilon_3$	$-\varepsilon_4$
2 2 1			$-\varepsilon_3$	$+\varepsilon_4$
2 2 2			ε_3	

4. Bounds for the Number of Possible Type Structures

Even if the knowledge of the possible type structures in most situations will not have any effect on the number of statistical tests which are necessary for identifying a type structure, it might be of interest to know how many possible type structures exist in a given situation. For a 2×2 scheme these were 3 out of 81, i.e. 3.7%, for a 2×3 scheme 13 out of 729, i.e. 1.8%, and for a 2×4 scheme 51 out of 6561, i.e. 0.8%.

We were not able to derive an explicit formula for the number of possible type structures in a given situation. However, rather crude upper and lower bounds for this number are derived in the following.

We consider the variable X_i with r_i categories. For each category j_i with $1 \le j_i \le r_i$ there exist

(17) $s_i = r / r_i$

configurations which contain this category. With respect to these s_i configurations we can consider 3^{s_i} structures (cf. formula (10)). Among these structures is the structure which contains only zeros (i.e. neither types nor antitypes) and which is always possible. Of the remaining (3^{si} -1) structures all those structures with only T or 0 (altogether (2^{si} -1) structures) or those structures with only A or 0 (altogether (2^{si} -1) structures) are no possible type structures. The remaining structures with at least one T and one A are possible.

Thus, we get

(18) $3^{s_i} - 2(2^{s_i} - 1) = 3^{s_i} - 2^{s_i+1} + 2$

possible type structures.

If we select from the r_i categories (r_i - 1) categories with altogether (r_i - 1)s_i configurations, we can assign to each of these categories an arbitrary possible structure of those which are enumerated in formula (18). This yields

(19) $(3^{s_i} - 2^{s_i+1} + 2)^{r_i - 1}$

possible type structures for (r_i -1)s_i selected configurations. For the remaining category with s_i configurations it is not possible to assign an arbitrary possible structure of those enumerated in (18) because the restrictions (9) must be obeyed. It is even possible that the possible structure for the last category is completely fixed by the preceding assignments. Thus we find that by (19) a lower bound for the number (n_{TS}) of possible type structures is given. If not all numbers (r_i) of categories are identical, we can improve the lower bound (19) by selecting the variable X_i in such a way that the bound becomes maximum, i.e. we use the lower bound

(20) $$L_{TS} = \max_{1 \le i \le t}(3^{s_i} - 2^{s_i+1} + 2)^{r_i - 1}.$$

For the situations with $t = 2$ and min $\{r_1, r_2\} = 2$ the possible structure for the last (here: the second) category is completely fixed by the preceding assignment, i.e. we have $n_{TS} = L_{TS}$.

In all other cases we have neither a completely fixed structure for the last category nor is it possible to select a structure for the last category without any restriction. A selection without any restriction yields the number

(21) $$(3^{s_i} - 2^{s_i+1} + 2)^{r_i}$$

of possible structures. Thus we get by (21) an upper bound for n_{TS}. An improved upper bound is given by

(22) $$U_{TS} = \min_{1 \le i \le t}(3^{s_i} - 2^{s_i+1} + 2)^{r_i}.$$

In the special case $r_1 = ... = r_t =: r_0$ we get

(23) $$L_{TS}^{(r_0)} = (3^{r_0^{t-1}} - 2^{r_0^{t-1}+1} + 2)^{r_0 - 1}, U_{TS}^{(r_0)} = (3^{r_0^{t-1}} - 2^{r_0^{t-1}+1} + 2)^{r_0}.$$

For $r_0 = 2$ this yields

(24) $$L_{TS}^{(2)} = (3^{2^{t-1}} - 2^{2^{t-1}+1} + 2), U_{TS}^{(2)} = (3^{2^{t-1}} - 2^{2^{t-1}+1} + 2)^2.$$

By means of the bounds which were derived for n_{TS} we can derive bounds for the percentage n_{TS} % of possible structures by dividing the bounds for n_{TS} by the maximal number $M = 3^r$ (cf. (10)) and multiplying by 100:

(25) $$\frac{L_{TS}}{M}100 \le n_{TS}\% \le \frac{U_{TS}}{M}100.$$

For the case with $r_1 = ... = r_t = 2$, i.e. for 2^t schemes, we can give a slightly improved lower bound. To that end we consider only the two first and the two last configurations

$$K_1 \;\; = (j_1 = 1, ..., j_{t-1} = 1, j_t = 1),\ K_2 = (j_1 = 1, ..., j_{t-1} = 1, j_t = 2),$$
$$K_{r-1} = (j_1 = 2, ..., j_{t-1} = 2, j_t = 1), K_r = (j_1 = 2, ..., j_{t-1} = 2, j_t = 2).$$

If we set $K_1 = T$, $K_2 = A$, $K_{r-1} = A$, $K_r = T$, we get a possible structure irrespective of the status of the remaining r - 4 = 2^t - 4 configurations. The same is true if we set $K_1 = A$, $K_2 = T$, $K_{r-1} = T$, $K_r = A$. From this follows that we have for 2^t schemes at least

(26) $$L_{TS}^{*} = 2 \times 3^{2^t - 4} + 1$$

possible structures where the additional possible structure (corresponding to the last term in (26)) is that structure where the status 0 is assigned to all configurations.

For the 2×2 scheme we get $L^*_{TS} = 3$, i.e. the true value. For the $2\times2\times2$ scheme we get $L^*_{TS} = 163$ and for the $2\times2\times2\times2$ scheme $L^*_{TS} = 1062883$.

Extending this approach we can successively improve the lower bounds for 2^t schemes if t is sufficiently large. Here, we demonstrate only the next stage of the approach: If we have $t \geq 3$, we get additional possible structures if we consider the 3 first and the 3 last configurations. Then, we set either $K_1 = T$, $K_2 = T$, $K_3 = A$, $K_{r-2} = A$, $K_{r-1} = T$, $K_r = T$ or $K_1 = A$, $K_2 = A$, $K_3 = T$, $K_{r-2} = A$, $K_{r-1} = A$, $K_r = A$. Here, we can choose the status of the remaining $r - 6 = 2^t - 6$ configurations in an arbitrary way. From this we get the improved lower bound

$$(27) \qquad L^{**}_{TS} = 2\times3^{2^t-4} + 2\times3^{2^t-6} + 1.$$

For the $2\times2\times2$ scheme we get $L^{**}_{TS} = 181$ and for the $2\times2\times2\times2$ scheme $L^{**}_{TS} = 1180981$.

5. Examples

For the examples from sections 2 and 3 we obtain the following bounds for n_{TS}:

1. $t = 2$, $r_1 = r_2 = 2$: $L_{TS} = L^*_{TS} = 3$, $U_{TS} = 9$, $n_{TS} = 3$
2. $t = 2$, $r_1 = 2$, $r_2 = 3$: $L_{TS} = 13$, $U_{TS} = 27$, $n_{TS} = 13$
3. $t = 2$, $r_1 = 2$, $r_2 = 4$: $L_{TS} = 51$, $U_{TS} = 81$, $n_{TS} = 51$
4. $t = 2$, $r_1 = r_2 = 3$: $L_{TS} = 169$, $U_{TS} = 2197$, $n_{TS} = 493$
5. $t = 3$, $r_1 = r_2 = r_3 = 2$: $L_{TS} = 51$, $L^*_{TS} = 163$, $L^{**}_{TS} = 181$, $U_{TS} = 2601$, $n_{TS} = 985$

6. Discussion

We have seen that it may be problematical to interpret types and antitypes which were identified by means of CFA independently of each other. This is particularly true for contingency tables with few variables and few categories. The specific definition of a type which is inherent to CFA may have the consequence that types or antitypes for certain configurations entail corresponding types and antitypes for other configurations. These latter types or antitypes might be considered as artifacts which should not be given an empirical interpretation.

References

1. Krauth, J.: Einführung in die Konfigurationsfrequenzanalyse (KFA). Ein multivariates nichtparametrisches Verfahren zum Nachweis und zur Interpretation von Typen und Syndromen. Weinheim/Basel: Beltz, Psychologie-Verlags-Union, 1993.
2. Lienert, G. A.: Die ”Konfigurationsfrequenzanalyse” als Klassifikationsmethode in der klinischen Psychologie. In: Irle, M. (Ed.) Bericht über den 26. Kongress der Deutschen Gesellschaft für Psychologie, Tübingen 16.9 – 19.9.1968, 244 – 253. Göttingen: Hogrefe, 1969.