AUSTRIAN JOURNAL OF STATISTICS
Volume 37 (2008), Number 2, 161–173

A Functional Approach to Configural Frequency Analysis

Alexander von Eye[1], Patrick Mair[2]

[1]Michigan State University
[2]Wirtschaftsuniversität Wien

Abstract: Standard Configural Frequency Analysis (CFA) is a one-step procedure that determines which cells of a cross-classification contradict a base model. The results are possible types/antitypes depending on whether the observed cell frequencies are significantly lower/higher with respect to the base model. Selecting these cells out does not guarantee that the base model fits. Therefore, the role played by these cells for the base model is unclear, and interpretation of types and antitypes can be problematic. In this paper, *functional CFA* is proposed. This model of CFA pursues two goals simultaneously. First, cells are selected out that constitute types and antitypes. Second, the base model is fit to the data. This is done using an iterative procedure that blanks out individual cells one at a time, until the base model fits or until there are no more cells that can be blanked out. In comparison to standard CFA, functional CFA is shown to be more parsimonious, that is, fewer types and antitypes need to be selected out. The methods are illustrated and compared using data examples from the literature.

Zusammenfassung: Die Lienert'sche Konfigurationsfrequenzanalyse (KFA) ist ein 1-Schritt Verfahren, mit Hilfe dessen man bestimmen kann, welche Zellen einer Kreuztabellierung einem Basismodell widersprechen. Dabei resultieren mögliche Typen/Antitypen, je nachdem ob die Häufigkeiten in Bezug auf das Basismodell signifikant über-/unterbesetzt sind. Nimmt man diese Zellen aus dem Modell, ist nicht garantiert, dass das Modell die Daten auch gut beschreibt. Deshalb ist die Rolle, die diese Zellen für das Basismodell spielen, unklar, und die Interpretation von Typen und Antitypen kann problematisch werden. In dieser Arbeit wird der Ansatz einer funktionalen KFA vorgestellt. Dabei werden zwei Ziele verfolgt: Erstens werden die Zellen identifiziert, die Typen und Antitypen konstituieren. Zweitens wird das Basismodell an die Daten angepasst. Diese Ziele werden mit einer iterativen Prozedur verfolgt, die eine Zelle nach der anderen aus der Kreuztabellierung entfernt. Die Iteration endet, wenn das Basismodell die Daten gut beschreibt, oder wenn keine Zellen mehr entfernt werden können. Im Vergleich zur Lienert'schen KFA ist die funktionale sparsamer, d.h., es werden weniger Typen und Antitypen identifiziert. Die Methoden werden mit Hilfe von Datenbeispielen aus der Literatur illustriert.

Keywords: Configural Frequency Analysis, Functional CFA, Kieser-Victor CFA.

1 Introduction

Configural Frequency Analysis (CFA; Lienert, 1968; von Eye and Gutiérrez-Peña, 2004) allows the researcher to identify those cells in a cross-tabulation that contradict a particular base model. Existing approaches to CFA have approached the identification process from three directions. The first and classical approach specifies a base model and, then, examines either all cells or a selection of cells with the goal of finding those that contradict the base model. This approach assumes under the null hypothesis that each case in the table was drawn from the same population. The typical base model is a simple, hierarchical log-linear model the expected frequencies of which can be estimated using closed forms. More complex base models have also been discussed (von Eye, 2002). The second approach (Kieser and Victor, 1991, 1999, 2000) proceeds under the assumption that the cases in those cells that belong to a CFA type (a review of the concepts of CFA types and antitypes follows in the next section of this article) were drawn from different populations. Therefore, estimation of expected cell frequencies must exclude these cases. The typical base model is a quasi-independence log-linear model for which, in most cases, closed forms do not exist. The third approach is Bayesian (Gutiérrez-Peña and von Eye, 2000).

In this article, a fourth approach to CFA is proposed. This approach will also be frequentist, and will be compared to the first two approaches. It is functional in the sense that types and antitypes are defined by the role they play for the base model. Iteratively, cells will be blanked out that contradict the base model. The iteration concludes as soon as the base model can be retained. A corresponding implementation in R (R Development Core Team, 2007) is provided.

2 A Review of Lienert's Classical CFA

When applying CFA, researchers, in a first step, specify a base model, also called chance model. In the present context we focus on log-linear base models. The standard base model thus has the form $\log \mathbf{m} = \mathbf{X}\boldsymbol{\lambda}$, where $\mathbf{m}$ is the vector of model frequencies, $\mathbf{X}$, is the design matrix, and $\boldsymbol{\lambda}$ is the parameter vector. CFA examines individual cells. Let the observed frequency of Cell c be n_c, and the corresponding expected frequency, m_c, be estimated under some chance model, where c goes over all cells in the table. CFA tests, typically for each cell, the null hypothesis under which $\mathrm{E}(n_c - m_c) = 0$. If $\mathrm{E}(n_c - m_c) > 0$, cell c is said to constitute a CFA *type*. If $\mathrm{E}(n_c - m_c) < 0$, cell c is said to constitute a CFA *antitype*. If $\mathrm{E}(n_c - m_c) = 0$, cell c is said to constitute neither a type nor an antitype. In brief, *types* occur more frequently than one would expect by chance, and *antitypes* occur less frequently than one would expect by chance.

For the decision as to whether a cell constitutes a CFA type or antitype, a number of tests has been proposed (for an overview, see von Eye, 2002). Each of these tests can be used to examine individual cells of a cross-classification. Tests for the examination of groups of cells have also been proposed. CFA tests are either exact or asymptotic, and they either can be used under any sampling scheme or require product-multinomial sampling. The binomial test is exact and can be used under any sampling scheme. The z-test and the X^2-test are asymptotic and can also be used under any sampling scheme. The ex-

Table 1: Standard CFA for Weather (W) and Persons Waiting (P) cross-classification.

Configuration c	n_c	m_c	r_c	p-value	Type/Antitype?
11	173	142.10	5.2366	.0000	Type
12	76	106.89	-5.2366	.0000	Antitype
21	50	54.21	-0.9555	.3393	–
22	45	40.78	0.9555	.3393	–
31	15	15.40	-0.1627	.8707	–
32	12	11.59	0.1627	.8707	–
41	51	61.06	-2.1765	.0295	–
42	56	45.93	2.1765	.0295	–
51	27	39.94	-3.3343	.0009	Antitype
52	43	30.05	3.3343	.0009	Type
61	15	18.26	-1.1985	.2307	–
62	17	13.73	1.1985	.2307	–

act and asymptotic hypergeometric tests (Lehmacher, 1981) require product-multinomial sampling. These tests are the most powerful of all current CFA tests, by far.

Base models of CFA contain all effects that are *not of interest* to the researcher (von Eye, 2004). Thus, if a base model is rejected, (1) the data are bound to reflect types or antitypes, and (2) these types and antitypes reflect the effects that are of interest to the researcher. In the present article, we focus on log-linear base models. These models have, in standard frequentist CFA, been mostly simple models, that is, models for which closed forms exist for the estimation of the expected frequencies.

In the present article, the group of log-linear base models will be extended to enable the *functional approach to CFA*. The new log-linear base models will not be in the class of simple hierarchical models any more. Instead, they will be non-standard (Mair, 2007; Mair and von Eye, 2007). That is, these models will contain terms that identify cells as *structural* in a sense comparable to structural zeros. Adding these terms changes standard hierarchical CFA base models into nonstandard models.

Data example: The following data example is presented to illustrate the application of Lienert's classical approach to CFA. It uses data from Wurzer (2005, p. 98). In a 6×2 table the variable weather (W) is cross-classified with persons waiting at a public Internet terminal (P). W was scored as 1 = dry and warm, 2 = dry and cold, 3 = raining and warm, 4 = raining and cold, 5 = snowing and warm, 6 = snowing and cold; P was scored as 1 = yes and 2 = no. The results of ordinary CFA are given in Table 1. Note that the standardized Pearson residuals r_c and the corresponding $N(0, 1)$-approximation were used (see Section 3.1) with a Bonferroni-protected $\alpha^* = 0.00417$. Obviously, one type is constituted by persons that are waiting when the weather is dry and warm; one antitype is constituted by persons that are not waiting when the weather is dry and warm. The second type is constituted by individuals who do not wait when snow falls and it is warm. The second antitype is constituted by individuals who do wait under these weather conditions.

The effect coded design matrix $\mathbf{X}$ for the data example in Table 1 is of the following

form:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & -1 \\ 1 & -1 & -1 & -1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix}$$

From left to right, this design matrix contains in the first column the constant vector, the vectors for the main effects of W, and the vector for the main effect of P. There are no vectors for interactions. Therefore, the CFA types and antitypes suggest that interactions exist (LR-$X^2 = 31.82$, $df = 5$, $p < 0.0001$). But, as was indicated above, CFA is not interested in identifying these interactions. Instead, CFA focuses on the interpretation of those cells (configurations) that stand out as types and antitypes.

Note that the model used for the present data example is equivalent to the model $\log \mathbf{m} = \lambda + \lambda^H + \lambda^T + \lambda^{HT} + \lambda^P$ where H indicates humidity and T temperature. The type and antitype thus can be interpreted as "weather conditions predict waiting behavior" (for prediction CFA, see von Eye, Mair, and Bogat, 2005).

3 Functional CFA

3.1 Basic principles of functional CFA

One characteristic that, with the exception of Kieser and Victor CFA (KV-CFA; more detail follows below), all CFA approaches share is that they are one-step methods. One base model is specified, and the analysis is performed in one run. The result is expressed in terms of local deviations from the base model. However, as Victor (1989) notes, due to dependencies between the cells in a table, so called *phantom types* can occur: If a certain cell constitutes a type/antitype it can result that a neighbor cell becomes a type/antitype as well; without being actually a type. Thus, stepwise approaches can be advantageous since type/antitype cells are excluded one-by-one and the model is re-fitted after each step.

Functional CFA (fCFA) asks questions concerning the deviations from a base model. However, it combines the goals of modeling with the goals of CFA. fCFA asks what role particular configurations play for a base model. If a configuration contradicts a base model, it is removed from the table and the base model is fitted again. This process is repeated until either no cells can be removed any more or the base model fits. Thus, the researcher can extract and interpret types/antitypes successively and fit the model after each step without biasing the results due to phantom types. At the end, when no types can be detected anymore, the base model fits.

fCFA thus uses base models that differ from standard CFA. The base models of fCFA consist of two parts. The first is identical to the base model of standard CFA, that is, $\log \mathbf{m} = \mathbf{X}_s \boldsymbol{\lambda}_s$. This part is *structural* in the sense that it specifies the variable relationships considered in the base model. The second is the part used to blank out type and antitype cells. This part is termed *functional* as it serves to mark those cells that contradict the base model and, thus, constitute types and antitypes. The base model thus changes to $\log \mathbf{m} = \mathbf{X}_s \boldsymbol{\lambda}_s + \mathbf{X}_f \boldsymbol{\lambda}_f$. The functional part of the model is created in an iterative process (see below).

If the iteration comes to an end before the pool of cells that can be removed is depleted (i.e, $df = 0$), the results are the following:

1. A selection of cells that constitute CFA types and antitypes. The interpretation of these cells proceeds as in standard CFA. However, the base model needs to be kept in mind.
2. A fitting final model. This model describes the variable relationships within an incomplete table, that is, a table without the type and antitype cells. These cells have been removed by way of declaring them *structural zeros*.

In contrast to standard CFA which practically always yields types or antitypes, fCFA can yield the statement that a base model cannot be fitted to a table. In this case, the types and antitypes that were constituted by the cells removed during the iteration cannot be interpreted, because the goal of fitting the base model was not reached.

To describe the procedure of fCFA, consider a CFA base model that is specified as $\log \mathbf{m} = \mathbf{X}_s \boldsymbol{\lambda}_s$. Let this base model meet the criteria set up by von Eye and Schuster (2000). Then, the iteration that is performed in fCFA involves the following steps:

1. Inspect the cell-wise discrepancies from this base model and identify the largest.
2. Blank out the cell with the largest discrepancy and re-fit the base model.
3. Repeat steps 1 and 2 until either the base model fits or the table becomes impossible to re-analyze because too many cells have been blanked out and the base model still does not fit.

Blanking out cells uses the same methods as declaring cells structural zeros. In each case, no model-specific probability density mass is placed into these cells, and these cells are excluded from the estimation of both overall fit and cell-specific residuals. Several types of residuals within the GLM framework can be taken into account. The classical definition are the *Pearson residuals* e_i which for a Poisson GLM are defined by

$$e_i = \frac{n_i - m_i}{\sqrt{m_i}} \,.$$

Asymptotic theory states that *standardized Pearson residuals* given by

$$r_i = \frac{n_i - m_i}{\sqrt{m_i(1 - h_{ii})}}$$

follow more closely the standard normal distribution. The elements h_{ii} ($0 \leq h_{ii} \leq 1$) are the main diagonal elements of the *hat matrix*

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X} \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{W}^{1/2} \,.$$

For Poisson GLM the elements w_{ii} of the diagonal matrix $\mathbf{W}$ are the model frequencies m_i (see e.g. Agresti, 2002, p. 139). If $n_i = m_i$, no standard error can be computed. In fCFA this occurs for cells already blanked out. However, this issue does not affect the appropriateness of the solution since the residual values are used in a descriptive manner in terms of blanking out $\max |r_i|$ within each iteration l.

Another technical matter concerns the correction of the α-level. It is an important issue in ordinary CFA since we have a *simultaneous testing* situation: Each cell is tested on the base of the residuals whether it constitutes a type or an antitype (see von Eye, 2002, Section 3.10). What is basically done in fCFA is that at each step the LR-test is carried out and if the model does not fit, the maximal residual value $e_i^{\max}$ is said to constitute a type if $\operatorname{sgn} e_i^{\max} = 1$ and an antitype if $\operatorname{sgn} e_i^{\max} = -1$. Thus, no test is carried out at an individual residual level. Please note that the individual tests and the protected alpha mentioned in the analysis of the data in Table 1 are needed for classical CFA only, but not for fCFA. For fCFA, the largest residual is used, and the significance test is not performed.

However, since fCFA is a stepwise approach where, after each step, the model fit is tested, we have the situation of *sequential testing*. A corresponding α-correction becomes relevant in the case of (large) tables where many iterations have to be performed in order to achieve a model that does not contradict with the log-linear base model. As in Kieser and Victor (1999), a corresponding procedure for multiple testing of nested hypothesis proposed by Bauer and Hackl (1987) can be applied.

3.2 A comparison between fCFA and KV-CFA

As was mentioned in the last section, the version of CFA proposed by Kieser and Victor (1999, 2000) is the only approach other than fCFA, that involves a stepwise selection procedure. Kieser and Victor (1999) propose the following steps for their exploratory forward inclusion routine.

1. Starting from a log-linear base model, contrasts for structural zeros are sequentially included.
2. Select the parameter for which the corresponding LR-value is minimal. (Note that this step involves removing cells from the table.)
3. Repeat steps 1 and 2 until the goodness-of-fit test is non-significant.

KV-CFA differs in two central points from fCFA. First, the authors aim at minimizing the overall LR statistic. The blanking out of cells is a means toward this goal. In contrast, in functional CFA the identification of "outlandish cells" is the goal. The fact that fCFA typically yields a model that fits, is a byproduct. However, this byproduct is a condition for an admissible solution. Second, to find an optimal solution, KV-CFA uses the overall goodness-of-fit LR-criterion. In contrast, fCFA blanks those cells out that are extreme based on the magnitude of residual scores. Solutions based on different statistics will differ depending on the discrepant characteristics of these statistics (see, e.g., von Eye and Mun, 2003; von Weber, von Eye, and Lautsch, 2004). In the following applications, it becomes clear that the different criteria for imposing structural zero contrasts and, thus, blanking out types/anti-type cells will typically result in different solutions.

For problems with large and high-dimensional tables involved (as, e.g., in *data mining*) it is straightforward to show that fCFA performs much faster than KV-CFA: Let us denote the iteration steps by $l = 0, \ldots, L$ and the total number of cells by C. Within each iteration step, KV-CFA computes $C - l$ models. For a simple example of a cross-classification of 6 variables each of them having 4 categories the total number of cells is $C = 4096$. Thus, in step 0, KV-CFA computes 4096 models, in the second step 4095, etc. fCFA is far more efficient since within each step l, only one model is fitted, i.e., the model which blanks out the cell with the largest residual. Thus, in total only L models have to be computed.

3.3 Application examples for fCFA and KV-CFA

In this section we compute various examples and compare the results from fCFA and KV-CFA. All computations are performed using the R package `cfa` (Funke, Mair, and von Eye, 2007). The corresponding functions are `fCFA()` and `kvCFA()`. The R call is of the following structure: `n.i` is the observed frequency vector, `X` the design matrix, and `tabdim` a vector denoting the dimensions of the table. These three arguments are required to specify. In addition, the user can select the residual type by means of `restype` and the α-level using `alpha`.

Example 1: We start with the application of fCFA and KV-CFA on the dataset in Table 1. The R code chunk is

```
> #----------- define design matrix, data, dimensions ---------
> dd <- data.frame(a1=gl(6,2), b1=gl(2,1,12))
> X <- model.matrix(~a1+b1, dd, contrasts = list(a1 = "contr.sum",
+                   b1 = "contr.sum"))
> n.i <- c(173,76,50,45,15,12,51,56,27,43,15,17)
> tabdim <- c(6,2)
> #------------------ Kieser-Victor CFA -----------------------
> (res1 <- kvCFA(n.i, X, tabdim = tabdim))
> summary(res1)
> #-------------------- functional CFA ------------------------
> (res2 <- fCFA(n.i, X, tabdim = tabdim))
> summary(res2)
```

The results for fCFA are given in Table 2 and those for KV-CFA in Table 3.

Functional CFA and KV-CFA blank out cell 12 (antitype) and cell 11 (type), respectively. Ordinary CFA (see Table 1) has identified four types/antitypes. In tables that are spanned by dichotomous variables, using standardized residuals has the effect that the sum of the residuals in cells with complementary indices is zero. Therefore, standardized residuals will not allow one to select types or antitypes solely based on the magnitude of the residual. We recommend considering one of the following strategies. First, if researchers are mainly interested in types, blank out type-constituting cells only.

Second, if types and antitypes are equally interesting, use information that is provided by other measures of discrepancy. We can use the hypergeometric test proposed by Lehmacher (1981) which uses the exact variance instead of an asymptotic variance for the denominator of the formula for the standard normal z. This test is known as *Lehmacher's*

Table 2: fCFA for Internet Terminal Data.

Step	Cell blanked out	LR-X^2	χ^2	*df*	p-value
0	none	31.82	31.43	5	< 0.0001
1	11	3.97	3.94	4	0.4107

Table 3: KV-CFA for Internet Terminal Data.

Step	Cell blanked out	LR-X^2	χ^2	*df*	p-value
0	none	31.82	31.43	5	< 0.0001
1	11	3.97	3.94	4	0.4107

z. Alternatively, the squared Pearson residuals X^2 as well as log-linear interaction terms (unweighted and weighted) as defined in Goodman (1991) can be taken into account.

A third approach is to exclude both cells at the same time. The effect of blanking out a particular cell c (of a dichotomous variable) can be that in the next fCFA step $n_c = m_c$. Let c' denote the corresponding complementary cell. Since the margins are fixed it follows necessarily that $n_{c'} = m_{c'}$. Concerning the example above, if cell 11 is considered as a type, cell 12 should be declared simultaneously as an antitype. Otherwise cell 12 fits perfectly due to the fact that cell 11 was excluded. Subsequent versions of the `cfa` package will include corresponding strategies and options for the treatment of dichotomous variables.

Example 2: A stepwise CFA (i.e., fCFA and KV-CFA) is performed on a dataset from Aksan et al. (1999) (see also von Eye, 2002). Their children's temperament data describe Control (C), Negative Affect (A), and Approach (H). Each of the variables was classified into 3 levels: C = 1 indicating low control, C = 2 average control, C = 3 high control; A = 1 low score negative effect, A = 2 average, A = 3 high; and H = 1 high score in approach, H = 2 average, and H = 3 low. The base model is again a log-linear main effects model. The fCFA results are given in Table 4. The corresponding results for the KV-CFA can be found in Table 5.

Obviously, fCFA blanks out 7 cells. The set of types/antitypes is

$$\Theta^{(f)} = \{313, 131, 132, 312, 323, 121, 122\}\,.$$

For the KV-CFA the set of types/antitypes is

$$\Theta^{(KV)} = \{313, 131, 112, 332, 322, 121, 132\}\,,$$

where again 7 cells are blanked out. The common types/antitypes for both methods are

$$\Theta^{(f)} \cap \Theta^{(KV)} = \{313, 131, 132, 121\}\,.$$

The first 2 iteration steps are basically the same for both methods, at $l = 3$ fCFA identifies cell 312 as type whereas KV-CFA cell 112.

Example 3: For a further analysis of the behavior of fCFA against KV-CFA data from Netter (1983) are used (see also von Eye, 2002). In an experiment on stress responses, a sample of 162 subjects worked under two stress conditions. The first condition was a

Table 4: fCFA for the Children's Temperament Data.

Step	Cell blanked out	LR-X^2	χ^2	*df*	p-value
0	none	171.34	185.16	20	< 0.0001
1	313	116.70	119.10	19	< 0.0001
2	131	85.68	83.77	18	< 0.0001
3	132	67.18	68.58	17	< 0.0001
4	312	49.82	49.86	16	< 0.0001
5	323	39.39	38.15	15	0.0005
6	121	29.93	29.48	14	0.0078
7	122	20.31	20.56	13	0.0878

Table 5: KV-CFA for the Children's Temperament Data.

Step	Cell blanked out	LR-X^2	χ^2	*df*	p-value
0	none	171.34	185.16	20	< 0.0001
1	313	116.70	119.10	19	< 0.0001
2	131	85.68	83.77	18	< 0.0001
3	112	65.97	65.21	17	< 0.0001
4	332	51.16	50.97	16	< 0.0001
5	322	39.96	39.61	15	0.0005
6	121	30.52	30.08	14	0.0065
7	132	19.77	19.80	13	0.1011

response time task, and the second condition a verbal fluency task. Under each condition, plasma samples were taken to measure two levels of adrenaline ($A_1 \in \{1, 2\}$, i.e., for each condition; $A_1 \in \{1, 2\}$) and two levels of noradrenaline ($N_1 \in \{1, 2\}$; $N_1 \in \{1, 2\}$). An additional variable pertains to the participant classification (P) into hypertonic ($P = 1$) and normal ($P = 2$). It results a 2^5 cross-classification of $A_1 \times A_2 \times N_1 \times N_2 \times P$. This time, we do not have a main-effects log-linear base model but rather a two-way interaction model with respect to adrenalin/noradrenalin for both levels, i.e., $[A_1N_1][A_2N_2][P]$ in bracket notation.

The results for fCFA are given in Table 6 and the results for KV-CFA in Table 7. The set of types/antitypes found by fCFA is

$$\Theta^{(f)} = \{22221, 11112, 12122, 21212, 21211, 11212, 21112, 11111, 21111\},$$

whereas for KV-CFA

$$\Theta^{(KV)} = \{22221, 11112, 21122, 11111, 12212, 21121, 22222, 22121\}.$$

Correspondingly,

$$\Theta^{(f)} \cap \Theta^{(KV)} = \{22221, 11112, 11111\}.$$

Again, after step 2 the procedures diverge. However, both methods need $L = 9$ iterations.

Table 6: fCFA for Netter's adrenaline data.

Step	Cell blanked out	LR-X^2	χ^2	df	p-value
0	none	112.87	113.71	24	< 0.0001
1	22221	93.26	87.57	23	< 0.0001
2	11112	81.85	74.39	22	< 0.0001
3	12122	70.34	66.27	21	< 0.0001
4	21212	60.82	56.86	20	< 0.0001
5	21211	49.69	45.41	19	0.0002
6	11212	42.88	39.93	18	0.0008
7	21112	37.50	34.79	17	0.0029
8	11111	31.29	29.50	16	0.0124
9	21111	24.74	22.35	15	0.0535

Table 7: KV-CFA for Netter's adrenaline data.

Step	Cell blanked out	LR-X^2	χ^2	df	p-value
0	none	112.87	113.71	24	< 0.0001
1	22221	93.26	87.57	23	< 0.0001
2	11112	81.85	74.39	22	< 0.0001
3	21122	66.34	62.44	21	< 0.0001
4	11111	58.51	51.13	20	< 0.0001
5	12212	41.30	39.39	19	0.0022
6	21121	35.01	33.24	18	0.0094
7	22222	28.21	26.37	17	0.0425
8	22121	21.58	18.66	16	0.1573

4 Discussion

The new version of CFA proposed in this article, functional CFA, selects types and antitypes iteratively, based on the contribution to the base model that is made by the cells that constitute the types and antitypes. Over the course of the iteration, the role played by individual cells changes. Therefore, the results of functional CFA can be expected to differ from the results from standard CFA in three important respects.

First, the number of types and antitypes is typically smaller in functional CFA. With each iteration step, the discrepancies from the base model can be expected to become smaller, and not all cells that constitute types and antitypes in the first step of the iteration – this step is identical to standard frequentist CFA – need to be declared structural cells. Therefore, the number of types and antitypes from functional CFA can be smaller.

Second, the pattern of types and antitypes identified by functional CFA can differ from the pattern from standard CFA. The reason is that model-data discrepancies are model-specific. Although the structural part of the base model does not change over the course of the iterative search for types and antitypes, the functional part will change because, with each iteration step, the design matrix will include additional vectors. These vectors are needed to specify which cells are blanked out. Because of these additional vectors, the standardized residuals for the non-structural cells can change, and, thus, their role as

type- or antitype constituting.

Third, functional CFA can fail. In contrast to standard CFA which always yields results functional CFA can fail when the number of cells that need to be declared structural is so large that the base model cannot be fit again. In this case, no cell can be said to constitute a type or antitype, and researchers may consider a different variant of CFA.

The question arises when to select functional CFA over standard CFA. From our perspective, functional CFA does not replace standard CFA. The relationship of functional to standard CFA is analogous to that of stepwise regression to standard regression. Functional CFA is a stepwise, exploratory procedure for the search for types and antitypes. The model is re-fit at each step of the iteration. Functional CFA is the method of choice in exploratory research. In confirmatory research, standard CFA (or confirmatory CFA by Kieser and Victor, 1999) can be used.

Functional CFA improves on standard CFA in three elementary ways. First, in standard CFA, the situation can occur that types and antitypes contradict a model that does not even fit. In these cases, the status of types and antitypes as contradicting a base model is doubtful. Second, in most cases, it can be expected that functional CFA is more parsimonious in that fewer cells need to be marked as constituting types and antitypes. Third, functional CFA can fail in the sense that the base model cannot be improved to the extent that it fits the cells that are not marked as structural. The main reason for this is that the number of structural cells has becomes too large.

It should be noted that, for the current analyses and comparisons, the exploratory version of Kieser and Victor's CFA was used. The authors have also proposed a confirmatory version that begins with blanking out an a priori determined cell. It is obvious that this version can lead to dramatically different appraisals of the type/antitype structure in a table because this cell is not necessarily the one with the largest discrepancy or the one that leads, when blanked out, to the greatest reduction in the overall goodness-of-fit score.

This article presents the first step in the development of functional CFA. There are many areas that need to be developed further. The following areas seem to be most important, at this point. First, optimal selection procedures for types and antitypes need to be developed. From the application of stepwise procedures for the development of regression models (see Neter, Kutner, Nachtsheim, and Li, 2004; von Eye and Schuster, 1998), we know that many methods are not guaranteed to provide optimal solutions. Specifically, the fact that regression parameter estimates can change in the presence/absence of certain variables poses problems for the final selection of a parsimonious solution. In an analogous fashion, the selection of cells to be blanked out can have an effect on the final solution. For the present article, the largest residual was used as the sole criterion. Alternative criteria are conceivable, for example the criterion that the number of eventually retained types and antitypes be smallest, or the criterion that the largest residual for step $i + 1$ be maximized/minimized at step i. Kieser and Victor use a strategy that focuses on the overall goodness-of-fit. Hybrid criterion sets are conceivable.

Furthermore, non-log-linear base models can be considered. As was demonstrated by von Eye (2002, 2004), classes of base models exist that are not log-linear. Examples of such models include models that use a priori probabilities. Future research will have to determine the usability of the functional CFA approach under these classes of base models as well as in a Bayesian context.

Acknowledgements

The authors are indebted to Eun Young Mun for helpful discussions of earlier versions of this article.

References

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). Hoboken, NJ: Wiley.

Aksan, N., Goldsmith, H. H., Smider, N. A., Essex, M. J., Clark, R., Hyde, J. S., et al. (1999). Derivation and prediction of temperamental types among preschoolers. *Developmental Psychology*, *35*, 958-971.

Bauer, P., and Hackl, P. (1987). Multiple testing in a set of nested hypotheses. *Statistics*, *18*, 345-349.

Funke, S., Mair, P., and von Eye, A. (2007). `cfa`: R package for the analysis of configuration frequencies [Computer software manual]. URL: `http://cran.R-project.org`.

Goodman, L. A. (1991). Measures, models, and graphical displays in the analysis of cross-classified data. *Journal of the American Statistical Association*, *86*, 1085-1111.

Gutiérrez-Peña, E., and von Eye, A. (2000). A Bayesian approach to configural frequency analysis. *Journal of Mathematical Sociology*, *24*, 151-174.

Kieser, M., and Victor, N. (1991). A test procedure for an alternative approach to configural frequency analysis. *Methodika*, *5*, 87-97.

Kieser, M., and Victor, N. (1999). Configural frequency analysis (CFA) revisited - a new look at an old approach. *Biometrical Journal*, *41*, 967-983.

Kieser, M., and Victor, N. (2000). An alternative approach for the identification of types in contingency tables. *Psychologische Beiträge*, *42*, 402-404.

Lehmacher, W. (1981). A more powerful simultaneous test procedure in configural frequency analysis. *Biometrical Journal*, *23*, 429-436.

Lienert, G. A. (1968). Die Konfigurationsfrequenzanalyse als Klassifikationsmethode in der klinischen Psychologie. [Configural frequency analysis as classification method in clinical psychology.]. *Paper presented at the 26. Kongress der Deutschen Gesellschaft für Psychologie in Tübingen*.

Mair, P. (2007). A framework to interpret nonstandard log-linear models. *Austrian Journal of Statistics*, *36*, 1-15.

Mair, P., and von Eye, A. (2007). Application scenarios for nonstandard log-linear models. *Psychological Methods*, *12*, 139-156.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Li, W. (2004). *Applied Linear Statistical Models*. Chicago: Irwin Press.

Netter, P. (1983). Typen sympathomedullärer Aktivität und ihrer psychischen Korrelate. In H. Studt (Ed.), *Psychosomatik in Forschung und Praxis* (p. 216-233). München: Urban & Schwarzenberg.

R Development Core Team. (2007). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org` (ISBN 3-900051-07-0)

Victor, N. (1989). An alternative approach to configural frequency analysis. *Methodika*, *3*, 61-73.

von Eye, A. (2002). *Configural Frequency Analysis: Methods, Models, and Applications*. Mahwah, NJ: Lawrence Erlbaum.

von Eye, A. (2004). Base models for configural frequency analysis. *Psychology Science*, *46*, 150-170.

von Eye, A., and Gutiérrez-Peña, E. (2004). Configural frequency analysis - the search for extreme cells. *Journal of Applied Statistics*, *31*, 981-997.

von Eye, A., Mair, P., and Bogat, G. A. (2005). Prediction models for configural frequency analysis. *Psychology Science*, *47*, 342-355.

von Eye, A., and Mun, E. Y. (2003). Characteristics of measures for 2×2 tables. *Understanding Statistics*, *2*, 243-266.

von Eye, A., and Schuster, C. (1998). *Regression Analysis for Social Sciences: Models and Applications*. San Diego, CA: Academic Press.

von Eye, A., and Schuster, C. (2000). Configural frequency analysis under retrospective and prospective sampling schemes: Frequentist and Bayesian approaches. *Psychologische Beiträge*, *42*, 428-447.

von Weber, S., von Eye, A., and Lautsch, E. (2004). The Type II error of measures for the analysis of 2×2 tables. *Understanding Statistics*, *3*, 259-282.

Wurzer, M. (2005). *An Application of Configural Frequency Analysis: Evaluation of the Uage of Internet Terminals*. Unpublished master's thesis, University of Vienna.

Corresponding Author's Address:

Alexander von Eye
Department of Psychology
Michigan State University
316 Psychology Building
East Lansing, MI 48826-1116
USA

E-mail: voneye@msu.edu