

Person-Centered Data Analysis With Covariates and the R-Package `confreq`

Mark Stemmler¹ , Jörg-Henrik Heine², Susanne Wallner¹

[1] Department of Psychology, Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Erlangen, Germany. [2] Centre for International Student Assessment, School of Education, Technical University Munich (TUM), Munich, Germany.

Methodology, 2021, Vol. 17(2), 149–167, <https://doi.org/10.5964/meth.2865>

Received: 2020-02-25 • Accepted: 2021-06-15 • Published (VoR): 2021-06-30

Corresponding Author: Mark Stemmler, Lehrstuhl für Psychologische Diagnostik, Methodenlehre und Rechtspsychologie, Nägelsbachstr. 49c, 6. OG, 91052 Erlangen, Germany. E-mail: mark.stemmler@fau.de

Supplementary Materials: Data [see [Index of Supplementary Materials](#)]



Abstract

Configural Frequency Analysis (CFA) is a useful statistical method for the analysis of multiway contingency tables and an appropriate tool for person-oriented or person-centered methods. In complex contingency tables, patterns or configurations are analyzed by comparing observed cell frequencies with expected frequencies. Significant differences between observed and expected frequencies lead to the emergence of Types and Antitypes. Types are patterns or configurations which are significantly more often observed than the expected frequencies; Antitypes represent configurations which are observed less frequently than expected. The R-package `confreq` is an easy-to-use software for conducting CFAs; another useful shareware to run CFAs was developed by Alexander von Eye. Here, CFA is presented based on the log-linear modeling approach. CFA may be used together with interval level variables which can be added as covariates into the design matrix. In this article, a real data example and the use of `confreq` are presented. In sum, the use of a covariate may bring the estimated cell frequencies closer to the observed cell frequencies. In those cases, the number of Types or Antitypes may decrease. However, in rare cases, the Type-Antitype pattern can change with new emerging Types or Antitypes.

Keywords

configural frequency analysis (CFA), log-linear modeling (LLM), person-oriented research, CFA with covariates, R-package `confreq`

This article describes the use of the package `confreq` (Heine, Alexandrowicz, & Stemmler, 2020) for the R language and environment for statistical computing (R Core Team, 2020). The R-package `confreq` allows for computation of Configural Frequency



Analysis (CFA; von Eye, 2002; von Eye & Wiedermann, in press) together with covariates. The CFA was developed by Gustav A. Lienert (1920–2001) as a multivariate non-parametric method for detecting types and syndromes, and analyzes multidimensional contingency tables (Lienert & Krauth, 1975). CFA compares observed (f_o) with expected (f_e) frequencies and searches for over- or underfrequent cells. Configurations that are observed more frequently than the expected frequencies ($f_o > f_e$) are called *Types* and configurations that are observed less often than expected frequencies ($f_o < f_e$) are called *Antitypes* (Lienert & Krauth, 1975).

In the English speaking literature CFA is closely related to the name Alexander von Eye (von Eye, 1990, 2002). He revealed the close relationship of CFA to log-linear modeling and he is responsible for a number of further statistical improvements of the CFA like, for instance, the parametric CFA (Spiel & von Eye, 1993), the auto-association CFA (von Eye, Mun, & Bogat, 2008) and the CFA based on the Bayes theorem (von Eye & Gutierrez-Peña, 2004).

Theoretical Background

In the field of developmental psychology, person-oriented research is mainly represented by David Magnusson and Lars Bergman from the University of Stockholm in Sweden (see also Stemmler & Heine, 2017). They call this approach *holistic and interactionistic*, where development is investigated in functionally organized systems (Bergmann & Magnusson, 1997). Systems or individuals are embedded and strongly connected with their context. The individual is seen as a self-organizing unit, which functions and develops as an irreducible whole. This wholeness evolves out of the inter-dynamics between the different structures and elements of the system. The respective structures and processes of the individual encompass psychological constructs like behavior, perceptions, goals, plans of action, social norms, motives, values, and biological functioning in the brain and physiological system. They obtain their roles and meanings as part of the interaction between the structures of the systems within the whole individual. “A certain element derives its significance not from its structure but from its functional role in the system of which it forms a part” (Magnusson & Mahoney, 2001, p. 5). Development cannot be understood by studying single factors in isolation from other simultaneously operating factors. The individual and the environment influence each other. The individual is seen as an active agent or producer of his or her own development (Lerner & Busch-Rossnagel, 1981; Silbereisen & Noack, 2006). Therefore, the human being is seen as a complex dynamic system, which could be understood ideally only under the holistic interactionistic approach (Bergman, Magnusson, & El Khouri, 2003; Magnusson & Allen 1983).

Here we give the reader a brief overview of the interdisciplinary use of CFA (see also Stemmler, 2020):

1. Researchers from the *field of hydrobiology* (Melcher, Lautsch, & Schmutz, 2012) were interested in spawning habitats of fish, because a sufficient fish stock is important for the ecological system of a river. They found that many European fish prefer the following significant configuration (Type): a shaded habitat with a fine and coarse substrate, depending on high flow velocity.
2. Ilker and Ercan (2018) studied the causes of the death of cattle calves with the help of CFA. They recorded the characteristics of the barn system (separation of mothers and calves or joint rearing), type of disease (intestinal disease, respiratory disease, trauma), vaccination status (vaccinated versus unvaccinated) and gender. The Turkish researchers from *veterinary medicine* found a significant configuration (Type): Cattle calves died more often than expected from intestinal disease, if they were not vaccinated and if the mothers were kept with the calves together in a barn; the gender of the calf was irrelevant.
3. Lannegrand-Willems et al. (2018) used CFA in their research on *developmental psychology*. Adolescence and emerging adulthood were seen as periods in life when individuals question and define their place in society and form their identity. This French research group studied the importance of different forms of civic engagement among late adolescents and emerging adults and found a significant configuration (Type): Students, who lacked identity formation did not participate in any political or civic engagement, neither did they vote and they had no feeling of sense of belonging to any social group; at the same time, they were untroubled by the absence of personal commitment.
4. Brenner et al. (2020) studied race and ethnicity considerations in traumatic brain injuries based on a database of Pennsylvania trauma injuries (i.e., Pennsylvania Trauma Outcome Study [PTOS]). Depending on the functional status at discharge, they used CFA to investigate the discharge destinations. Among several Types, the researchers from *traumatic brain injury research* found that for brain injured patients with a moderate to severe disfunction at discharge, White individuals were overall more likely to receive extended care than individuals in other racial groups who were likely to be sent home.

CFA belongs to the person-oriented analytic approach for the analysis of frequencies in multi-way contingency tables (cf. Stemmler, 2020). In this manuscript, the use of the R-package `confreq` (Heine et al., 2020) is demonstrated. First, the results of a first order CFA are presented and then covariates are added to the design matrix in two consecutive steps. The effects of covariates on the detection of *Types* or *Antitypes* are described. The data used are from a study of juveniles called Chances and Risks in the Life-Course (CURL; Reinecke et al., 2013).

Methodological Background

On the statistical level and for the use of CFA, individuals, animals or objects are grouped in cross-tabulations into disjunct categories based on their respective patterns or configurations (Stemmler & Heine, 2017). Patterns (configurations) with frequencies (o_{ijk}) that occur significantly more often than their corresponding expected cell frequencies (e_{ijk}) constitute CFA *Types*. Configurations occurring significantly less often than predicted under the null hypothesis constitute CFA *Antitypes*. Log-linear modeling (LLM) and CFA are closely related (von Eye & Mun, 2013). LLM is used to identify the structure among the categorical variables. It parameterizes the distribution of cell frequencies, or, to put it differently, the logarithms of cell frequencies, in terms of main effects and interactions. Each CFA base model can be expressed as a LLM model; however, there are LLM models that cannot be used as a CFA base model.

In log-linear modeling the expected frequencies are estimated by using the Generalized Linear Model (GLM). The General Linear Model is a special case of the Generalized Linear Model. The GLM is:

$$f(y) = X\beta \quad (1)$$

The function $f(y)$ is called link function. The link function describes the transformation of the dependent variable. In matrix algebra, the parameters are calculated according to the GLM as

$$\beta = (X^T X)^{-1} X^T Y, \quad (2)$$

where Y is a column vector including the dependent variable. X is the matrix of independent variables, and β is the vector of parameters. In LLM the predictor model can be written as:

$$\log(e) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (3)$$

with $\log(e)$ as the logarithm of the expected frequencies. In LLM one uses \ln (the natural logarithm) with base e (i.e., Euler's constant = 2.7182...).

If we replace the parameters β by λ we obtain the equation for a log-linear modeling:

$$\ln(e_{ijk}) = \lambda_0 + \lambda_i A_i + \lambda_j B_j + \lambda_k C_k \quad (4)$$

in the case of three variables A , B and C . The relation of the parameters is the same as in the GLM (see Equation 2):

$$\lambda = (X^T X)^{-1} X^T \log(e) \quad (5)$$

where λ is the parameter vector, $\log(e)$ is the vector of expected model frequencies; these are the frequencies that are consistent with the log-based model. X is the design matrix and may contain effect-coded main effects, interaction terms as well as covariates plus the constant (intercept). In addition to effect coding, dummy coding and contrast coding are possible. The design matrix X has as many rows as there are cells or configurations. The first λ weight is always the constant, coded with ones. λ comprises the weights of the independent variables and is a one-column vector with as many entries as X has columns.

The basis of CFA is the analysis of frequencies in multi-way contingency tables. Each individual case is cross-tabulated into disjunct categories based on his or her respective pattern or configuration. The underlying logic is the comparison of observed frequencies $f_{(o)}$ with expected frequencies $f_{(e)}$. Therefore, a global chi-square, a goodness-of-fit statistic, is calculated (this following formula is, for didactic reasons, presented for three variables but can easily be extended to any number of variables):

$$\chi_{ijk}^2 = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \frac{(o_{ijk} - e_{ijk})^2}{e_{ijk}} \quad (6)$$

I = number of categories of the first variable ranging from $i = 1, 2, \dots, I$

J = number of categories of the second variable ranging from $j = 1, 2, \dots, J$

K = number of categories of the third variable, ranging from $k = 1, 2, \dots, K$

o_{ijk} = the observed frequencies of pattern ijk

e_{ijk} = the expected frequencies of pattern ijk

and the general formula for the degrees of freedom for a contingency table with main effects is:

$$df = T - \sum_{d=1}^D (v_d - 1) - 1 \quad (7)$$

with T representing the number of cells or configurations, with $d = 1, \dots, D$ representing the number of variables (dimensions), and v_d the number of categories of a variable.

An important alternative goodness-of-fit statistic to the Pearson's chi-square is the Likelihood Ratio chi-square (LR):

$$LR = 2 \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K o_{ijk} \ln \frac{o_{ijk}}{e_{ijk}} \quad (8)$$

The global chi-square tests the following statistical hypotheses (H_0 and H_1). Again, the following formulas are, for didactic reasons, presented to three variables but may be extended to any number of variables easily:

$$H_0: \pi_{ijk} = \pi_{i..} \pi_{.j.} \pi_{..k} \quad (9)$$

$$H_1: \pi_{ijk} \neq \pi_{i..} \pi_{.j.} \pi_{..k} \quad (10)$$

π_{ijk} = defines the cell probabilities at the population level, $\pi_{i..} \pi_{.j.} \pi_{..k}$ = define the marginal probabilities at the population level.

In semantic terms, the null (H_0) and alternative hypothesis (H_1) are expressed as follows:

H_0 : There are no significant (local) associations between the variables involved or the variables are independent of each other.

H_1 : There are significant (local) associations between the variables involved or the variables are not independent of each other.

The alternative hypothesis includes also higher-order associations. In non-hierarchical log-linear models, lower-order associations are omitted (cf. Rindskopf, 1990). From the perspective of log-linear modeling, leaving out the lower-order association effect parameters can be problematic, because the effects coded in the design matrix may no longer be independent of each other. Subsequently, this makes the interpretation of the effect parameters more complex (cf. Mair & von Eye, 2007; von Eye & Mun, 2013).

The expected frequencies were calculated according to the assumption of independence:

$$e_{ijk} = \frac{o_{i..} \times o_{.j.} \times o_{..k}}{n^2} \quad (11)$$

A CFA that is based on the assumption of independence is called first order CFA. In addition, we differentiate between a *local* and a *global* chi-square value. A significant global chi-square, which is a goodness-of-fit statistic, is a necessary but not a sufficient condition for a significant local chi-square. It can be that the null hypothesis is rejected, but cell-wise model-data discrepancies may not be extreme enough to result in Types or Antitypes (von Eye & Wiedermann, in press). A significant local chi-square indicates a local association between variables; it is calculated by

$$\chi_{ijk}^2 = \frac{(o_{ijk} - e_{ijk})^2}{e_{ijk}} \quad (12)$$

with 1 degree of freedom.

Significant local chi-square values represent *Types* or *Antitypes*. Another valuable statistic in the search of Types or Antitypes is the chi-square approximation to the z-test:

$$z_{ijk(\frac{\alpha}{2})}^2 = \chi_{ijk\alpha}^2 \quad (13)$$

$$z_{ijk(\frac{\alpha}{2})}^2 = \frac{(o_{ijk} - e_{ijk})^2}{e_{ijk}} = \frac{(o_{ijk} - np_{ijk})^2}{np_{ijk}} =$$

$$z_{ijk} = \frac{(o_{ijk} - e_{ijk})}{\sqrt{np_{ijk}}} \quad (14)$$

CFA also allows the use of continuous variables as covariates. “The use of covariates typically carries the estimated cell frequencies closer to the observed cell frequencies, because more information is used in the estimation procedure (von Eye & Niedermeier, 1999)” (von Eye, 2002, p. 309). Note, CFA tests are never fully independent (von Eye, Mair, & Mun, 2010) and an alpha protection is required (e.g., Bonferroni’s adjustment or Holm’s procedure).

Method

Study Subjects

The data for the present paper relate to the project “CURL” (see Reinecke et al., 2013 for an overview). They include 1248 students from 5th grade; 189 (15.1%) of whom had reported at least one crime in the last year. The longitudinal data for t_1 to t_2 (time gap: two years) included 775 juveniles with complete data with regard to delinquency. Of the 189 offenders at t_1 , 114 (ca. 60%) remained in the longitudinal data file, and about one half (48.2%) reported of having committed another crime at t_2 .

Study Variables

The selection of variables, here possible risk factors, for the following analyses was based on a publication with the title “Risk factors for the development of antisocial behavior in childhood and youth” (German translation: Risikofaktoren für die Entwicklung dissozialen Verhaltens in der Kindheit und Jugend; Stemmler et al., 2018). In this chapter, which included an introduction to the concept of risk factors and their characteristics, also data from the project “CURL” were analyzed. For the following analyses, all bivariate associations were included that showed any significant correlation between the risk factors and delinquent behavior. Delinquent behavior encompassed behavior that is forbidden under the penal law; this includes property crime, vandalism and violence. An “offender” was defined as a person that reported having committed at least one crime in the past year. A “non-offender” was a study person who had not committed any crime in the past year.

The design of the study and a detailed description of the measures can be taken from Weiss and Wallner (2019).

Results

Results of the First Order CFA

In 5th grade *Antisocial Attitudes* together with *Delinquent Peers* were significantly associated with *Offender Status* two years later (cf. Stemmler & Wallner, 2019). The results of the first order CFA can be found in Table 1.

Table 1

Results of a First Order Configurational Frequency Analysis (CFA) With the Risk Factors Antisocial Attitude and Delinquent Peers in Combination With Offender Status (2 Years Later)

Patterns							
Att	Peer	Offend	$f_{(o)}$	$f_{(e)}$	z-statistic	p	Type/Antitype
-	-	-	486	449.67	1.71	.043	
-	-	+	130	149.34	-1.58	.057	
-	+	-	8	24.01	-3.27	.001	Antitype
-	+	+	7	7.97	-0.35	.365	
+	-	-	46	70.55	-2.92	.002	Antitype
+	-	+	31	23.43	1.56	.059	
+	+	-	8	3.77	2.18	.015	
+	+	+	14	1.25	11.40	.000	Type

Note. Att = Antisocial Attitudes; Peer = Delinquent Peers; Offend = Offender Status. “-” = not present/no; “+” = present/yes. Type (overfrequent cell): $f_{(o)} > f_{(e)}$; Antitype (underfrequent cell): $f_{(o)} < f_{(e)}$.

Both goodness-of-fit statistics suggested a poor fit: LR = 77.72, $df = 4$, $p < .001$; $\chi^2 = 161.90$, $df = 4$, $p < .001$; AIC = 127.570; BIC = 127.888. Therefore, with respect to the calculated expected frequencies which were determined under the assumption of the null hypothesis that interaction effects do not exist, one *Type* and two *Antitypes* emerged. The *Antitypes* suggested that there were fewer observed frequencies than expected under the null hypothesis of independence. Based on the expectancy of the null hypothesis the pattern “- + -” was an Antitype, meaning that there were fewer juveniles than expected to have no *Antisocial Attitudes* but being associated with *Delinquent Peers* and not being an *Offender*. In addition, another Antitype emerged for configuration “+ - -”, indicating that there were fewer juveniles than expected but who were not an *Offender* and not being associated with *Delinquent Peers* but having *Antisocial Attitudes*. The Type “+ + +” was more interesting in terms of criminological research: More juveniles than

expected under the null hypothesis committed an *Offense* who also showed *Antisocial Attitudes* and who spent their leisure time with *Delinquent Peers*. At the same time the configuration “+ + -” was almost a Type with $p = .015$ (it missed the Bonferroni adjusted alpha level), showing that there were juveniles with *Antisocial Attitudes* who socialized with *Delinquent Peers* but reported not having committed an *Offense*, maybe those juveniles underreported their committed delinquent acts.

The R-Package `confreq` and Other CFA Software

Alexander von Eye (Michigan State University) has written a CFA program (von Eye, 1998) which is available as a shareware. This program was written in FORTRAN 90 and runs on the DOS level and is therefore suitable only for Windows PCs. The program starts by double-clicking on the file `cfa.exe`. The “von Eye program” is controlled by typing numbers into the program. After it starts, the user needs to proclaim whether the data will be entered via a file `<= 1 >` or interactively `<= 2 >`. The “von Eye program” can display a design matrix, if requested (without the constant); it is easy to use and allows to run two sample CFAs in addition to zero order and first order CFAs.

Funke, Mair, and von Eye (2007) wrote the first R package called CFA. However, this R package has not been updated for use in newer major R base versions. We, therefore, recommend that the new R package `confreq` should be used. `confreq` is the abbreviation for **conf**igural **freq**uencies. The package was written by Jörg-Henrik Heine (Heine et al., 2020); it is constantly updated and maintained. The name `confreq` avoids a mix-up with Confirmatory Factor Analysis which is also often abbreviated as CFA. The package `confreq` is now available (Version 1.5.5-2) from the repositories on CRAN¹ and therefore suitable for the latest R version 4.0 (R Core Team, 2020).

Within R, one can read in a frequency table by typing in the pattern and their frequencies into a spreadsheet file. Such form of data are typically named as *tabulated data*, where the rows represent all possible combinations of the variables and the rightmost column holds the respective frequencies. To prepare the data to be imported into R, save the spreadsheet as an `csv`-file into your current R workspace directory by naming it for example as “5thgrade.csv” (additional materials, including the R syntax and the Excel files, are provided in [Supplementary Materials](#)). For correct processing the tabulated data with `confreq` the header of the rightmost column holding the pattern frequencies must be named “Freq”.

The following R syntax will lead to the results of [Table 1](#).

1) See <https://cran.r-project.org/web/packages/confreq/citation.html>

```
# reading in an EXCEL file in csv-format
#
order1 <- read.table("5thgrade.csv", sep=";", header=TRUE, quote="\")
order1
# you need to load the R-package confreq
# do not use zeros as configural patterns!
library("confreq")
# convert the data to patterned frequencies
order1pat<-dat2fre(fre2dat(order1))
order1pat
# first order CFA
resd1 <- CFA(order1pat,alpha=0.05, form="~ Offender + Delinqpeer + Attitude")
summary(resd1)
# inspect the design matrix of the first order CFA
resd1$designmatrix
```

The resulting design matrix for the base model (see last syntax line in the box above) looks like the following:

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{pmatrix}$$

In the first column one can see the constant, followed by the effect coded main effects for *Antisocial Attitude*, *Delinquent Peers* and *Offender Status*.

Results of the First Order CFA With One Covariate

The underlying idea is that covariates are employed in the loglinear base model to compute the expected frequencies. As the first covariate we added *Parental Engagement*, a scale from the Alabama Parenting Questionnaire (Frick, 1991). While expressing CFA in terms of LLM, the covariate is added by simply extending the GLM Equation 1 (see Glück & von Eye, 2000):

$$\log(e) = X\beta + c\beta_c \quad (15)$$

with c = covariate vector, β_c = parameter for the covariate.

The resulting model belongs to the family of nonstandard log-linear models. In the literature there exist caution with the ambiguous interpretation of parameters from such nonstandard models. Mair (2007) offers a solution by looking at the effects coded in the design matrix and determining the numerical contribution of single effects.

As the Equation 15 shows, the covariate is simply added as a column to the design matrix of the log-linear model; there is one score per covariate for each cell. Usually, the cell means of the continuous covariate are used; however, any other statistics may also be applied, for example, medians, percentages, probabilities or even categorical covariates and interactions of covariates with other variables are possible. If a cell has t cells and design matrix X contains k vectors (including the constant), the maximum number of covariates is $t - k - 1$.

Let's have a look at the R-syntax with one covariate:

```
##### the covariates from CURL 5th Grade -----
co <- read.csv2(file = "covariate.csv", header = TRUE)
co
# to run a CFA with one covariate, here Apq_pe Parental Engagement
erg4_PE <- CFA(order1pat, cova = co$Apq_pe)
# 'cova = co$Apq_pe' adds the covariate to the design matrix
summary(erg4_PE, showall = T, type = "pChi")
# have a closer look at the design matrix
erg4_PE$designmatrix
```

The resulting design matrix with the means of *Parental Engagement*, in the far right column, looks like the following:

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 3.54 \\ 1 & -1 & 1 & 1 & 3.48 \\ 1 & 1 & -1 & 1 & 3.25 \\ 1 & -1 & -1 & 1 & 3.64 \\ 1 & 1 & 1 & -1 & 3.20 \\ 1 & -1 & 1 & -1 & 3.00 \\ 1 & 1 & -1 & -1 & 3.33 \\ 1 & -1 & -1 & -1 & 3.48 \end{pmatrix}$$

The results of the first order CFA with one covariate can be taken from Table 2. It is also possible to add the covariate to a base model which also includes interactions.

The fit is still not perfect; there are significant differences between the observed and expected frequencies; however, the AIC and BIC were reduced and we lose one degree of freedom: LR = 48.32, $df = 3$, $p < .001$; $\chi^2 = 56.71$, $df = 3$, $p < .001$; AIC = 100.173; BIC = 100.570. The Antitype “- + -” vanished because the expected frequencies got closer to the observed one. A new Type evolved: “+ - +” because the expected and observed frequencies are deviated further apart; the difference between the two changed from 7.57 to 16.86. The interpretation of the new Type needs to involve the configuration's covariate, meaning that all juveniles in this cell are adjusted to the covariate *Parental Engagement*. It maybe that controlling for parents' engagement leads to *Offenders with Antisocial Attitudes* who do not associate as much with *Delinquent Peers* as expected under the null hypothesis. The remaining Antitype and Type stayed the same. It is

Table 2

Results of a First Order CFA Plus the Covariate Parental Engagement With the Risk Factors Antisocial Attitude and Delinquent Peers in Combination With Offender Status (2 Years Later)

Patterns							
Att	Peer	Offend	$f_{(o)}$	$f_{(e)}$	z-statistic	p	Type/Antitype
-	-	-	486	458.35	1.29	.098	
-	-	+	130	147.39	-1.43	.076	
-	+	-	8	9.61	-0.52	.302	
-	+	+	7	15.65	-2.19	.014	
+	-	-	46	73.08	-3.17	.001	Antitype
+	-	+	31	14.14	4.46	.000	Type
+	+	-	8	6.96	0.39	.347	
+	+	+	14	4.78	4.22	.000	Type

Note. Att = Antisocial Attitudes; Peer = Delinquent Peers; Offend = Offender Status. “-” = not present/no; “+” = present/yes. Type (overfrequentated cell): $f_{(o)} > f_{(e)}$; Antitype (underfrequentated cell): $f_{(o)} < f_{(e)}$.

necessary to correct for multiple testing. In `confreq` either the Bonferroni adjustment or the Holm’s alpha protection can be applied (cf. [Stemmler, 2020](#)).

Results of the First Order CFA With Two Covariates

Next to *Parental Engagement*, another covariate, the use of *Corporal Punishment*, was added to the design matrix (see far right column). The resulting design matrix with the means of *Parental Engagement* and *Corporal Punishment* looks like the following:

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 3.54 & 1.27 \\ 1 & -1 & 1 & 1 & 3.48 & 1.44 \\ 1 & 1 & -1 & 1 & 3.25 & 1.64 \\ 1 & -1 & -1 & 1 & 3.64 & 1.90 \\ 1 & 1 & 1 & -1 & 3.20 & 1.70 \\ 1 & -1 & 1 & -1 & 3.00 & 1.48 \\ 1 & 1 & -1 & -1 & 3.33 & 2.00 \\ 1 & -1 & -1 & -1 & 3.48 & 1.79 \end{pmatrix}$$

Let’s have a look at the R-syntax with two covariates:

```
##### the covariates from CURL 5th Grade -----
co <- read.csv2(file = "covariate.csv", header = TRUE)
co
# to run a CFA with two covariates, here Parental Engagement and Corporal Punishment
erg5_CP <- CFA(order1pat, cova = cbind(co$Apq_pe, co$Apq_cp))
summary(erg5_CP, showall = T, type = "pChi")
# have a closer look at the design matrix
erg5_CP$designmatrix
```

The results of the first order CFA with two covariates can be found in [Table 3](#).

Table 3

Results of a First Order CFA Plus the Two Covariates Parental Engagement and Corporal Punishment With the Risk Factors Antisocial Attitude and Delinquent Peers in Combination With Offender Status (2 Years Later)

Patterns							
Att	Peer	Offend	$f_{(o)}$	$f_{(e)}$	z-statistic	p	Type/Antitype
-	-	-	486	486.759	-0.034	.486	
-	-	+	130	127.086	0.258	.398	
-	+	-	8	8.110	-0.038	.485	
-	+	+	7	9.047	-0.680	.248	
+	-	-	46	47.592	-0.231	.409	
+	-	+	31	31.564	-0.100	.460	
+	+	-	8	5.541	1.045	.148	
+	+	+	14	14.303	-0.008	.468	

Note. Att = Antisocial Attitudes; Peer = Delinquent Peers; Offend = Offender Status. “-” = not present/no; “+” = present/yes. Type (overfrequent cell): $f_{(o)} > f_{(e)}$; Antitype (underfrequent cell): $f_{(o)} < f_{(e)}$.

With two covariates, the significant differences between the observed and expected frequencies vanished. We invested another degree of freedom but we have a reasonable fit: LR = 1.60, $df = 2$, $p = .449$; $\chi^2 = 1.69$, $df = 2$, $p = .428$; AIC = 55.45; BIC = 55.92. No Types or Antitypes emerged. Obviously, high or low covariate values corresponded to high or low observed cell frequencies pulling the observed and expected frequencies together. Although it is not a perfect association, high values of *Parental Engagement* were mainly present for juveniles with no *Antisocial Attitudes* and high *Corporal Punishment* was associated with *Delinquent Peers*.

In CFA, covariates which correlate with the residuals decrease the differences between the observed and expected cell frequencies. However, covariates which do not correlate can lead to the emergence of new Types and Antitypes (Glück & von Eye, 2000; von Eye, Mair, & Mun, 2010).

Conclusions

We demonstrated the use of CFA with covariates. CFA is a very useful tool in the realm of person-oriented research which is related to other statistical methods which analyze patterns or configuration of information, like latent-class analysis (LCA), latent profile analysis and general growth mixture models (GGMM). GGMM are basically growth curve models performed for different latent classes (Stemmler & Lösel, 2015). All methods have in common, that they try to explain unobserved heterogeneity in groups. The appropriateness of such models is usually tested using goodness-of-fit measures such as information indices, for example, the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and their derivatives (von Eye & Wiedermann, *in press*).

Compared to other person-oriented data analysis approaches, CFA can be distinguished according to different aspects. First, in comparison to LCA and the mixture models associated with it (Yamamoto & Everson, 1995), a central difference is that these models are probabilistic models in which probabilities are modeled or estimated, while in CFA observed pattern frequencies are compared with expected ones. The goal of modeling in such probabilistic models is to assign each person to a predefined number of latent classes on the basis of his or her response pattern with a certain probability, whereby a maximum assignment probability can usually be determined for one of the latent classes. The (ideal) model conception consists here in the assumption of disjunctive and exhaustive person classes of initially unknown size. In this sense, LCA (and other mixture models) can be viewed as a procedure for model-based data clustering, through the application of which individual units of study (persons or objects) within the total sample can thus be grouped into subgroups (Fraley & Raftery, 2002). Thus, second, with LCA and mixture models, the primary goal is optimal model fit, whereas with CFA the focus is primarily on non-fitting models and the interest is primarily in residuals analysis. With LCA and mixture models, the patterns of association or structure of dependence between variables are supposed to disappear by assuming a fitting number of latent classes, thus explaining the associations between variables. CFA, on the contrary, focuses on over- or underfrequent configurations (patterns) and, to that extent, requires a non-fitting model to identify types and/or antitypes and thus engages in residual analysis.

The use of additional covariates makes CFA even more flexible. In particular, if one investigates variables of different scale levels (e.g., categorical and interval level variables). In the person-oriented research, a covariate which is significantly related to the variables under investigation brings the observed frequencies closer to the expected frequencies; this results in a diminishing number of Types and Antitypes. Moreover, this disappearance is probably causally related. von Eye and Wiedermann (2016) wrote “specifically, whenever Types or Antitypes disappear after the design matrix was extended, the hypothesis can be entertained that the add-on effects are explanatory for the disappeared Types or Antitypes” (p. 168). In some cases, a new pattern of emerging Types and Antitypes appears, depending on the correlation of the covariates with the

residuals of the model without continuous variables. Although usually the mean or median of a single continuous variable is added to represent the cases in a cell. Notwithstanding, CFA with covariates still belong to the person-oriented approach, because the person or objects in a cell are still considered to be indivisible, only more information is added; moreover, it is also still possible to add a covariate as another categorical variable functioning, for instance, as a stratification variable (cf. von Eye, 2002, Chapter 10).

Using the log-linear modeling (LLM) approach to CFA, covariates are simply added to the design matrix of a first order CFA by adding columns of means, medians or even percentages.

In addition, the use of the R-package `confreq` was demonstrated. When reading in the patterned frequencies in an Excel sheet, the use of covariates is straight forward. One can use as many covariates as one wishes, depending solely on the spare degree of freedoms left. Together with `confreq` CFA is a very powerful statistical tool in person-oriented research, however, it should be mentioned that `confreq` does not allow one yet to perform a Bayesian CFA. Next, to the first order CFA, other versions of the CFA are available for example, the two-sample CFA, prediction CFA (P-CFA) in longitudinal data, Configural Mediator Model (Stemmler, 2020) or functional CFA (fCFA; von Eye & Mair, 2008), which enable to blank out extreme outlier cells (cf. Stemmler & Heine, 2017) and CFA is even a complimentary tool for analyzing tree structures based on CHAID (Stemmler, Heine, & Wallner, 2019).

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Supplementary Materials

For this article the following Supplementary Materials are available via the PsychArchives repository (for access see [Index of Supplementary Materials](#) below):

- R scripts and Excel data files of the presented data examples.

Index of Supplementary Materials

Stemmler, M., Heine, J.-H., & Wallner, S. (2021). *Supplementary materials to: Person-centered data analysis with covariates and the R-package confreq* [Code]. PsychOpen GOLD.

<https://doi.org/10.23668/psycharchives.4946>

References

- Bergman, L. R., & Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology*, *9*(2), 291-319.
<https://doi.org/10.1017/s095457949700206x>
- Bergman, L. R., Magnusson, D., & El Khouri, B. M. (2003). *Studying individual development in an interindividual context: A person-oriented approach* (Vol. 4). Hove, United Kingdom: Psychology Press.
- Brenner, E. K., Grossner, E. C., Johnson, B. N., Bernier, R. A., Soto, J., & Hillary, F. G. (2020). Race and ethnicity considerations in traumatic brain injury research: Incidence, reporting, and outcome. *Brain Injury*. Advance online publication.
<https://doi.org/10.1080/02699052.2020.1741033>.
- Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*(458), 611-631.
<https://doi.org/10.1198/016214502760047131>
- Frick, P. J. (1991). *The Alabama Parenting Questionnaire* [Unpublished Manuscript]. Department of Psychology, University of Alabama.
- Funke, S., Mair, P., & von Eye, A. (2007). *The CFA package: Program module in R*. Retrieved from <http://cran.r-project.org/>
- Glück, J., & von Eye, A. (2000). Including covariates in Configural Frequency Analysis. *Psychologische Beiträge*, *42*(3), 405-417.
- Heine, J. H., Alexandrowicz, R., & Stemmler, M. (2020). *confreq: Configural Frequencies Analysis using Log-linear Modeling* (R package version 1.5.5-2). Retrieved from <https://cran.r-project.org/web/packages/confreq>
- Ilker, M., & Ercan, E. (2018). Investigation of local association in animal research for multiway cross tabulated count data. *Scholars Journal of Agriculture and Veterinary Sciences (SJAVS)*, *5*(6), 343-351. Retrieved from <http://saspjournals.com/wp-content/uploads/2018/07/SJAVS-56-343-351-c.pdf>
- Lannegrand-Willems, L., Chevrier, B., Perchec, C., & Carrizales, A. (2018). How is civic engagement related to personal identity and social identity in late adolescents and emerging adults? A person-oriented approach. *Journal of Youth and Adolescence*, *47*(4), 731-784.
<https://doi.org/10.1007/s10964-018-0821-x>
- Lerner, R. M., & Busch-Rossnagel, N. A. (Eds.). (1981). *Individuals as producers of their development: A life-span perspective*. New York, NY, USA: Academic Press.
- Lienert, G. A., & Krauth, J. (1975). Configural Frequency Analysis as a statistical tool for defining types. *Educational and Psychological Measurement*, *35*(2), 231-238.
<https://doi.org/10.1177/001316447503500201>
- Magnusson, D., & Allen, V. L. (Eds.). (1983). *Human development, an interactional perspective*. New York, NY, USA: Academic Press.

- Magnusson, D., & Mahoney, J. L. (2001). *Reports from the project Individual Development and Adaptation. A holistic person approach for research on positive development* (Technical report No. 76). IDA I Department of Psychology, Stockholm university.
- Mair, P. (2007). A Framework to Interpret Nonstandard Log-Linear Models. *Austrian Journal of Statistics*, 36(2), 89-103. <https://doi.org/10.17713/ajs.v36i2.323>
- Mair, P., & von Eye, A. (2007). Application scenarios for nonstandard log-linear models. *Psychological Methods*, 12(2), 139-156. <https://doi.org/10.1037/1082-989X.12.2.139>
- Melcher, A., Lautsch, E., & Schmutz, S. (2012). Non-parametric methods – Tree and P-CFA – for the ecological evaluation and assessment of suitable aquatic habitats: A contribution to fish psychology. *Psychological Test and Assessment Modeling*, 54(3), 293-306.
- R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R foundation for statistical computing. Retrieved from <https://www.R-project.org/>.
- Reinecke, J., Stemmler, M., Arnis, M., El-Kayed, N., Meinert, J., Pöge, A., Schepers, D., Stünkel, Z., Uysal, B., Wallner, S., Weiss, M., & Wittenberg, J. (2013). Entstehung und Entwicklung von Kinder- und Jugenddelinquenz: erste Ergebnisse einer Längsschnittstudie [Origin and development of child and adolescent delinquency: first results of a longitudinal study]. *Neue Kriminalpolitik* [New criminal policy], 25(3).
- Rindskopf, D. (1990). Nonstandard log-linear models. *Psychological Bulletin*, 108(1), 150-162. <https://doi.org/10.1037/0033-2909.108.1.150>
- Silbereisen, R. K., & Noack, P. (2006). Kontexte und Entwicklung. In W. Schneider & F. Wilkening (Eds.), *Zyklus der Psychologie, Serie V, Bd. 1 Theorien, Modelle und Methoden der Entwicklungspsychologie* [Encyclopedia of Psychology, Series V, Vol. 1 Theories, models and methods of developmental psychology] (pp. 311–368). Göttingen, Germany: Hogrefe.
- Spiel, C., & von Eye, A. (1993). Configural frequency analysis as a parametric method for the search of types and antitypes. *Biometrical Journal*, 35(2), 151-164. <https://doi.org/10.1002/bimj.4710350206>
- Stemmler, M. (2020). *Person-centered methods: Configural Frequency Analysis (CFA) and other methods for the analysis of contingency tables – Second Edition*. New York, NY, USA: Springer.
- Stemmler, M., & Heine, J.-H. (2017). Using configural frequency analysis as a person-centered analytic approach with categorical data. *International Journal of Behavioral Development*, 41(5), 632-646. <https://doi.org/10.1177/0165025416647524>
- Stemmler, M., Heine, J.-H., & Wallner, S. (2019). Analyzing tree structures with configural frequency Analysis and the R-package confreq. *Psychological Test and Assessment Modeling*, 61(4), 419-433. Retrieved from https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2019-4/03_Stemmler_Heine.pdf
- Stemmler, M., & Lösel, F. (2015). Developmental pathways of externalizing behavior from preschool age to adolescence: An application of general growth mixture modeling. In M. Stemmler, A. von Eye, & W. Wiedermann (Eds.), *Dependent data in social sciences research - forms, issues, and methods of analysis* (pp. 91–106). New York, NY, USA: Springer.

- Stemmler, M., & Wallner, S. (2019). Die Vorhersage von Jugenddelinquenz im Rahmen des personenorientierten Ansatzes – Analysen mit der Konfigurationsfrequenzanalyse (KFA) [The prediction of juvenile delinquency in the context of the person-oriented approach - analyses with the configuration frequency analysis (CFA)]. In S. Wallner, M. Weiss, J. Reinecke, & M. Stemmler (Eds.), *Devianz und Delinquenz in Kindheit und Jugend - Neue Ansätze der kriminologischen Forschung* [Deviance and Delinquency in Childhood and Adolescence - New Approaches in Criminological Research] (pp. 141-156). Wiesbaden, Germany: Springer VS.
- Stemmler, M., Wallner, S., & Link, E. (2018). Risikofaktoren für die Entwicklung dissozialen Verhaltens in der Kindheit und Jugend [Risk factors for the development of dissocial behaviour in childhood and adolescence]. In D. Hermann, & A. Pöge (Eds.), *Kriminal-soziologie: Handbuch für Wissenschaft und Praxis* [Criminal Sociology: Handbook for Science and Practice] (pp. 247–262). Baden-Baden, Germany: Nomos Verlagsgesellschaft.
- von Eye, A., (1990). *Introduction to Configural Frequency Analysis: The search for types and antitypes in cross-classifications*. Cambridge, United Kingdom: Cambridge University Press.
- von Eye, A. (1998). *Configural Frequency Analysis - A program for 32 Bit Windows operating systems* [Program manual]. East Lansing, MI, USA: Michigan State University, Department of Psychology.
- von Eye, A., (2002). *Configural Frequency Analysis: Methods, models, and applications*. Mahwah, NJ, USA: Lawrence Erlbaum.
- von Eye, A., & Gutiérrez-Penã, E. (2004). Configural frequency analysis: The search for extreme cells. *Journal of Applied Statistics*, 31(8), 981-997. <https://doi.org/10.1080/0266476042000270545>
- von Eye, A., & Mair, P. (2008). A functional approach to configural frequency analysis. *Austrian Journal of Statistics*, 37(2), 161-173. <https://doi.org/10.17713/ajs.v37i2.297>
- von Eye, A., Mair, P., & Mun, E.-Y. (2010). *Advances in Configural Frequency Analysis*. New York, NY, USA: Guilford Press.
- von Eye, A., & Mun, E.-Y. (2013). *Log-Linear Modeling: Concepts, interpretation, and application*. New York, NY, USA: John Wiley & Sons.
- von Eye, A., Mun, E.-Y., & Bogat, G. A. (2008). Temporal patterns of variable relationships in person-oriented research – Longitudinal models of Configural Frequency Analysis. *Developmental Psychology*, 44(2), 437-445. <https://doi.org/10.1037/0012-1649.44.2.437>
- von Eye, A., & Niedermeier, K. E. (1999). *Statistical analysis of longitudinal categorical data – An introduction with computer illustrations*. Mahwah, NJ, USA: Lawrence Erlbaum.
- von Eye, A., & Wiedermann, W. (in press). *Configural Frequency Analysis*. New York, NY, USA: Springer Nature.
- von Eye, A., & Wiedermann, W. (2016). Local associations in latent class analysis: Using configural frequency analysis for model evaluation. *Journal for Person-Oriented Research*, 2(3), 155-170.
- Weiss, M., & Wallner, S. (2019). Methodik der Studie [Methodology of the study]. In S. Wallner, M. Weiss, J. Reinecke, & M. Stemmler (Eds.), *Devianz und Delinquenz in Kindheit und Jugend: Neue Ansätze der kriminologischen Forschung* [Deviance and delinquency in childhood and youth: New approaches in criminological research] (pp. 17–38). Wiesbaden, Germany: Springer VS.

Yamamoto, K., & Everson, H. T. (1995). Modeling the mixture of IRT and pattern responses by a modified hybrid model. *ETS Research Report Series*, 1995(1), i-26.

<https://doi.org/10.1002/j.2333-8504.1995.tb01651.x>



Methodology is the official journal of the European Association of Methodology (EAM).



leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.