Correspondence Analysis of Longitudinal Data

Mark de Rooij*

LEIDEN UNIVERSITY, LEIDEN, NETHERLANDS

Peter van der G. M. Heijden

UTRECHT UNIVERSITY, UTRECHT, NETHERLANDS

*Corresponding author (rooijm@fsw.leidenuniv.nl)

Keywords: *categorical data, contingency table, latent class analysis, superindicator matrix, Burt matrix, event history data*

Abstract: Correspondence analysis is an exploratory tool for the analysis of associations between categorical variables, the results of which may be displayed graphically. For longitudinal data two types of analysis can be distinguished: the first focusses on transitions, whereas the second investigates trends. For transitional analysis with two time points, an analysis of the transition matrix (showing the relative frequencies for pairs of categories) provides insight into the structure of departures from independence in the transitions. Transitions between more than two time points can also be studied simultaneously. In trend analyses often the trajectories of different groups are compared. Examples for all these analyses are provided.

Correspondence analysis is an exploratory tool for the analysis of association(s) between **categorical** variables. Usually, the results are displayed in a **graphical** way.

There are many interpretations of correspondence analysis. Here we make use of two of them. A first interpretation is that the observed categorical data are collected in a **matrix**, and correspondence analysis approximates this matrix by a matrix of lower rank [19]. This lower rank approximation of, say, rank $M + 1$ is then displayed graphically in a M-dimensional representation in which each row and each column of the matrix is displayed as a point. The difference in rank between the rank $M +$

1 matrix and the rank *M* representation is matrix of rank 1, and this matrix is the product of the marginal counts of the matrix, that is most often considered uninteresting. This brings us to the second interpretation, that is, that when the two-way matrix is a **contingency table**, correspondence analysis decomposes the departure from a matrix where the row and column variables are independent [15, 16]. Thus, correspondence analysis is a tool for **residual** analysis. This interpretation holds because for a contingency table estimates under the independence model are obtained from a product of the margins of the table (divided by the total sample size).

Longitudinal data are data where observations (e.g. individuals) are measured at least twice using the same variables. We consider here only categorical (i.e. nominal or ordinal) variables, as only this kind of variables is analyzed in standard applications of correspondence analysis [14]. We first discuss correspondence analysis for the analysis of transitions. Thereafter, we consider analysis of trends with canonical correspondence analysis.

1 Transitional Analysis

1.1 Two Time Points

When there is one categorical variable measured at two time points, a so-called transition matrix can be constructed [1]. In this transition matrix, the row variable is the categorical variable measured at time 1, and the column variable is the categorical variable at time 2. The aim of a correspondence analysis of a transition matrix is to get an insight into the transitions from time 1 to time 2. Different questions about these transitions exist, and these lead to different form of correspondence analysis.

We index the levels of the row variable (time 1) with i, ($i = 1, \ldots, I$) and the levels of the column variable (time 2) with j, ($j = 1, \ldots, J$). We denote relative frequencies by p_{ij}, probabilities by π_{ij}, and estimates of probabilities by $\hat{\pi}_{ij}$. Marginal elements are found by replacing the index by "+", for example, row marginal elements of the matrix with relative frequencies are p_{i+} and column marginal elements are p_{+j}.

A first analysis would be a standard correspondence analysis of the contingency table with elements p_{ij}. The interpretation discussed above shows that the resulting graphic display can be interpreted as showing a

decomposition of the residuals from the independence model, that is, $\hat{\pi}_{ij} = p_{i+}\ p_{+j}$ [14–16].

A problem with this standard analysis is that often interest goes out to the off-diagonal elements (i.e. the cells for which $i \neq j$) in the contingency table, as these represent the individuals that change. In a standard correspondence analysis, the view on these cells might be blurred by the diagonal cells, especially, when $p_{ij} \gg p_{i+}p_{+j}$ (which is the case when many individuals remain in the same level of the categorical variable from time point 1 to 2). A solution to this problem is not to study the residuals from the independence model, but from the so-called quasi-independence model, defined here as $\pi_{ii} = p_{ii}$ for $i = j$ and $\pi_{ii} = \alpha_i\beta_j$ for $i \neq j$ [1]. It is possible to adjust correspondence analysis so that residuals from quasi-independence are decomposed. This can be done in two ways: by adjusting the computer program or by changing the input data. The last option seems most simple, and the way to do it is as follows: the diagonal elements p_{ii} have to be replaced by elements for which independence holds. This can be accomplished by filling in elements $p_{i+}p_{+i}$ for the diagonal. By doing this, the margins of the new table have changed so that the elements on the diagonal are not independent, and therefore, using the new margins, again elements $p_{i+}p_{+i}$ have to be filled in. After a few iterations, these elements have stabilized, and a correspondence analysis of the resulting table can be interpreted as a decomposition of quasi-independence [14, 15, 18].

This approach can be extended further by adjusting correspondence analysis so that it can decompose residuals from the symmetry model or from the quasi-symmetry model [14, 15]. Another development is to use statistical models instead of the exploratory approach described here. There are also close connections between correspondence analysis and latent class analysis [15].

We give a small example to illustrate an analysis of the departure from independence. Space limitations withhold us from a detailed interpretation, and for interpretation principles, we refer to [15]. The data are 5 import car types out of 16 car types published in [15]: subcompacts (subi), small specialties (smai), compacts (comi), midsize (midi), and luxury (luxi). In the rows of Table 1, we find the cars disposed of, and in the columns the new cars. Notice the dominant observed frequencies on the diagonal. These values dominate the first dimensions of a correspondence analysis (see Figure 1), especially, the diagonal luxi-cell compared with the rest. In a second analysis, we decompose the residuals from quasi-independence. Such an analysis can be accomplished by filling in "independent" values for the diagonal. These values are 12 790, 1381, 1033, 503 and 71. The interpretation of this correspondence analysis uses

the same principles as for standard correspondence analysis of the table with the adjusted margins. For the margins, the residuals are zero, and therefore, the graph only shows car type changes. The car order for cars disposed off is luxi, midi, comi, subi, and smai, but for new cars it is luxi, midi, smai, comi, and subi (see Figure 2). Notice, for example, the different position of smai. It is due to asymmetries in the data that become visible now that the dominance of the diagonal elements has been suppressed. For example, when people dispose of a smai, they buy a luxi very often (relative to the margins of the adjusted table, i.e. observed 459 but predicted by margins 239) but the reverse does not hold (observed 341 but predicted by margins 413).

Figure 1 Ordinary CA of car changing data [fg001.eps]

Figure 2 Generalized CA decomposing residuals from quasi-symmetry [fg002.eps]

Table 1 1979 Car Changing Data

	subi	**smai**	**comi**	**midi**	**luxi**	**Total**
subi	25 986	5400	2257	1307	288	35 238
smai	3622	5249	738	1070	459	11 138
comi	6981	1023	1536	1005	127	10 672
midi	2844	772	565	3059	595	7835
luxi	997	341	176	589	3124	5227
Total	40 430	12 785	5272	7030	4593	70 110

Rows denote cars disposed, columns denote new cars. Abbreviations are in the text.

1.2 More than Two Time Points

When there is one categorical variable measured at more than two time points, it is usual to code the response profiles into a so-called superindicator matrix. Correspondence analysis of a superindicator matrix is also known as multiple correspondence analysis. A superindicator matrix has *N* individuals in the rows and the categories for each of the time points in the columns. This correspondence analysis has the aim to get insight into the transitions between all time points simultaneously. The analysis also yields quantifications for the individuals, and the quantifications for an individual can be considered as summaries of the response profile of this

individual that can be used, but it can also be used to obtain a classification of the response profiles of the individuals [2, 5, 8, 10, 13, 14, 17].

As an example, we give a superindicator matrix of one dichotomous variable measured at three time points for $N = 101$ individuals (see Table 2). (In many computer programmes, the column vector with frequencies cannot be specified, but instead a matrix with 101 rows will serve as the data input file.) The matrix can be made larger in a straightforward way when the number of categories is larger than two, when there are more time points, or when there are more individuals. A correspondence analysis of this matrix will yield a three-dimensional display with 101 points, one for each individual, and a graphical display with 8 points, one for each category at each time point. Without going into technical details (see [10, 17]), individuals with similar profiles will be close together, categories that are often used by the same individuals will be close together, and, when we overlay the two graphs, individuals will be close to the categories that they use. It is also important to notice that, since correspondence analysis displays the departure from the row and from the column margin of a table, it follows that correspondence analysis will *not* show the trend in "a" and "b" over the three time points. This trend can be studied from the counts in the 3 × 2 table of time points by categories [14, 17]. See next Section.

Table 2 A Small Example of a Categorical Data Matrix (Panel A) and its Superindicator Matrix (Panel B)

Panel A		**Panel B**		
t		t_1	t_2	t_3
1 2 3	**Freq**	**a b**	**a b**	**a b**
a a a	40	1 0	1 0	1 0
a a b	16	1 0	1 0	0 1
a b a	4	1 0	0 1	1 0
a b b	12	1 0	0 1	0 1
b a a	8	0 1	1 0	1 0
b a b	3	0 1	1 0	0 1
b b a	6	0 1	0 1	1 0
b b b	12	0 1	0 1	0 1

Another way to interpret this analysis is when we realize that a correspondence analysis of the superindicator matrix, say $\boldsymbol{G}$, is mathematically related to a correspondence analysis of the so-called Burt matrix $\boldsymbol{G'G}$. The Burt matrix for this example is shown in Table 3. This matrix is a concatenation of a two-way contingency table for each pair of

time points, and diagonal matrices with marginal frequencies. This shows that the solution of correspondence analysis only uses two-way **interactions**, and ignores higher-way interactions. Thus, a Burt matrix contains sufficient information for a nonstationary **Markov chain** (the table of time points 1 and 3 is the matrix product of the tables of time points 1 and 2, and 2 and 3) [14, 17].

Table 3 The Burt Matrix for the Example in Table 2

		t_1		t_2		t_3	
		a	**b**	**a**	**b**	**a**	**b**
t_1	a	72	0	56	16	44	28
	b	0	29	11	18	14	15
t_2	a	56	11	67	0	48	19
	b	16	18	0	34	10	24
t_3	a	44	14	48	10	58	0
	b	28	15	19	24	0	43

Examples of such analyses can be found in [2, 5, 8, 14, 17]. If the number of individuals is not very large, the estimates for the category points will be unstable. More stability is obtained by constraining category points of adjacent time points to be the same. Such a solution can be obtained by adding up the indicator matrices of the adjacent time points [13]. This is also the way to go when the data to be analyzed are **event history** data, where the observations are in continuous time, or career data. Examples of unconstrained and constrained analyses are in [5, 8, 10, 13, 14, 17].

2 Trend Analysis

In the previous section the interest was in how individuals change from one category to another. Another type of analysis answers the question which categories become more popular over time and which less popular. In such an analysis the focus is often on the comparison how one group (treatment) develops compared to another (control). Take as an example, data [7] from a randomized controlled trial where approximately half of the subject receive a treatment and the other half no treatment. The subjects are mentally ill homeless participants living in San Diego, US. The treatment is a certificate designed to make it easier for those subjects to start living independently. Over the period of two years four repeated

measurements are taken at baseline, after 6 months, 12 months, and 24 months, where the housing condition of the subjects was assessed using three categories: living on the Street, Living in a Community House, and Living Independently. The data are represented in Table 4.

Table 4 The housing data

	S	C	I
C0	100	61	19
C6	30	93	38
C12	13	85	48
C24	18	66	61
T0	80	75	26
T6	15	45	101
T12	19	23	115
T24	19	36	103

Canonical correspondence analysis [11] is equal to standard correspondence analysis but with linear constraints on the row and/or column points. In the application on longitudinal data, constraints are placed on the coordinates of the rows (of Table 4), representing the two treatment groups at the four time points. Canonical correspondence analysis then decomposes the residuals from the hierarchical loglinear model with an association between time and treatment and a main effect of housing status. The types of constraints that are used are equal to those in growth curve models. The coordinates may, for example, be constrained to be linear functions of the time variable (with scores 0, 6, 12, and 24). The trajectories for both groups might be parallel (only a main effect of group) or non parallel (an interaction between time and group). Higher order functions of time can also be used, for example quadratic functions.

For our application introduced above separate quadratic functions for both groups fit the data well, explained inertia equals 96.3%. The graphical representation is shown in Figure 3 where it can be seen that the

trajectories for both groups already start at different points (the randomization did not work as was hoped for). The treatment group has a somewhat favorable position, that is, further away from the Living on the Street condition. Furthermore, the shape of the trajectories is rather different. The treatment group rapidly moves into the direction of Independent Housing, indicating that this category is becoming more popular over time for this group, and only in the end of the study there is a trend backwards. The trajectory of the control group first moves into the direction of Community Housing and later on in the study into the direction of Independent Housing.

Figure 3 Canonical CA for housing data [fg003.eps]

The analysis shown can be very useful in the interpretation of a marginal (GEE) or subject specific model (GLMM) for multinomial outcome variables. As was discussed in [6], the interpretation of regression coefficients in multinomial models is not simple, especially in cases with interactions and/or higher order treatment of variables. These cases are often encountered in longitudinal studies, because in these studies we are interested in the differential development of different groups over time. The interpretation of regression coefficients in multinomial models is further complicated because the coefficients refer to contrasts of categories of the response variable with a baseline category [6]. A graphical representation of the data as shown in Figure 3 provides a graphical representation of data under similar restrictions as employed in such multinomial models.

The analysis shown here is closely related to analyses using an ideal point model in a (quasi) maximum likelihood framework [3, 4]. There is a different formalization of the two types of analysis, but results are often very similar [20]. The maximum likelihood makes it possible to test different effects, which can be helpful in confirmatory research settings. However, such test requires distributional assumptions.

We believe that both applications of correspondence analysis, i.e. transitional and trend analysis provide great insight into the data and into the patterns of change. It should be noted that we did not deal with drop out in any way. The analysis assumes missing values are of the Missing Completely at Random type [9]. If this is not the case, analysis should be preceded by multiple imputations [12].

References

[1] Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975) *Discrete multivariate analysis*. M.I.T.Press, Cambridge.

[2] De Leeuw, J., van der Heijden, P. G. M. & Kreft, I. (1985) Homogeneity analysis of event history data, *Methods of Operations Research* **50**: 299–316.

[3] De Rooij, M. (2009). Trend vector models for the analysis of change in continuous time for multiple groups. *Computational Statistics and Data Analysis* **53**: 3209 – 3216.

[4] De Rooij, M. & Schouteden, M. (2012). The mixed effect trend vector model. *Multivariate Behavioral Research* **47**: 635-664.

[5] Deville, J. -C. & Saporta, G. (1983) Correspondence analysis, with an extension towards nominal time series, *Journal of econometrics* **22**: 169–189.

[6] Fox, J. and Anderson, R. (2006). Effect displays for multinomial and proportional-odds logit models. *Sociological Methodology* **36**: 225 - 255.

[7] Hedeker, D & Gibbons, R.D. (2006). *Longitudinal data analysis.* John Wiley and Sons, Inc., Hoboken, New Jersey.

[8] Martens, B. (1994) Analyzing event history data by cluster analysis and multiple correspondence analysis: an example using data about work and occupations of scientists and engineers, in *Correspondence analysis in the social sciences*, M. Greenacre & J. Blasius, eds. Academic Press, London, 233–251.

[9] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

[10] Saporta, G. (1985) Data analysis for numerical and categorical individual time-series, *Applied stochastic models and data analysis* **1**: 109–119.

[11] Ter Braak, C.J.F. (1986) Canonical Correspondence Analysis: A new eigenvector technique for multivariate direct gradients. *Ecology*, **67**: 1167-1179.

[12] Van Buuren, S. (2012), *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton, FL.

[13] Van Buuren, S. & de Leeuw, J. (1992) Equality constraints in multiple correspondence analysis, *Multivariate behavioral research* **27**: 567–583.

[14] van der Heijden, P. G. M. (1987) *Correspondence analysis of longitudinal categorical data*. D.S.W.O.-Press, Leiden.

[15] van der Heijden, P. G. M., de Falguerolles, A. & de Leeuw, J. (1989) A combined approach to contingency table analysis using correspondence analysis and loglinear analysis, *Applied Statistics* **38**: 249–292.
[16] van der Heijden, P. G. M. & de Leeuw, J. (1985) Correspondence analysis used complementary to loglinear analysis, *Psychometrika* **50**: 429–447.
[17] van der Heijden, P. G. M. & de Leeuw, J. (1989) Correspondence analysis, with special attention to the analysis of panel data and event history data, in *Sociological Methodology 1989*, C. C. Clogg, ed. Basil Blackwell, Oxford, 43–87.
[18] van der Heijden, P. G. M., de Vries, H. & van Hooff, J. A. R. A. M. (1990) Correspondence analysis of transition matrices, with special attention to missing entries and asymmetry, *Animal Behaviour* **40**: 49–64.
[19] van der Heijden, P. G. M., Gilula, Z. & van der Ark, L. A. (1999) An extended study into the relationships between correspondence analysis and latent class analysis, in *Sociological Methodology 1999*, M. Sobel & M. Becker eds. Blackwell, Cambridge, pp 147–186.
[20] van der Heijden, P. G. M., Mooijaart, A., & Takane, Y. (1994). Correspondence analysis and contingency table models, in *Correspondence analysis in the social sciences: Recent developments and applications,* Greenacre, M. J., & Blasius, J. Eds. New York: Academic Press, pp. 79-111.