Annals of Operations Research
https://doi.org/10.1007/s10479-022-04718-z

ORIGINAL RESEARCH

Weighted cumulative correspondence analysis based on a particular cumulative power divergence family

Antonello D'Ambra[1] · **Giovanni Meccariello**[2] · **Livia Della Ragione**[2]

Accepted: 4 April 2022
© The Author(s) 2022

Abstract
The Pearson's X^2 statistic and the likelihood ratio statistic G^2 are most frequently used for testing independence or homogeneity, in two-way contingency table. These indexes are members of a continuous family of Power Divergence (PD) statistics, but they perform badly in studying the association between ordinal categorical variables. Taguchi's and Nair's statistics have been introduced in the literature as simple alternatives to Pearson's index for contingency tables with ordered categorical variables. It's possible to show, using a parameter, how to link Taguchi's and Nair's statistics obtaining a new class called Weighted Cumulative Chi-Squared (WCCS-type tests). Therefore, the main aim of this paper is to introduce a new divergence family based on cumulative frequencies called Weighted Cumulative Power Divergence. Moreover, an extension of Cumulative Correspondence Analysis based on WCCS and further properties are shown.

Keywords Power divergence family · Cumulative index · Contingency table · Generalized singular value decomposition · Correspondence analysis

1 Introduction

The Pearson's X^2 statistic and the likelihood ratio statistic G^2 are most frequently used for testing independence or homogeneity, in two-way contingency table, in many research areas (Cressie and Read 1989). These indexes are members of a continuous family of Power Divergence (PD) statistics defined by Cressie and Read (1984). It's important to note that the PD family does not perform well when the rows/columns of the table are ordered, as reported in

✉ Giovanni Meccariello
giovanni.meccariello@stems.cnr.it

Antonello D'Ambra
antonello.dambra@unicampania.it

Livia Della Ragione
livia.dellaragione@stems.cnr.it

[1] Department of Economics, University of Campania "L.Vanvitelli", Via Gran Priorato di Malta, Capua, CE, Italy

[2] Institute of Sciences and Technologies for Sustainable Energy and Mobility, Via Guglielmo Marconi, 4, Naples, NA, Italy

Agresti (2007) and Barlow et al. (1972). This is partially due to the low power for ordered alternatives to the null hypothesis. In Barlow et al. (1972) authors discuss several exact and approximate likelihood ratio procedures for testing in these situations. Unfortunately, the distribution theory underlying these procedures can be complex. To address this problem, the class of tests called Cumulative Chi-Squared-type (CCS-type; Taguchi 1966, 1974; Nair 1987) may be considered. These CCS-type tests take into account the presence of an ordinal categorical variable considering the cumulative frequency of the cells in the contingency table. Moreover, CCS-type statistic is more suitable for studying situations where the number of categories within a variable is greater than or equal to 5 (Hirotsu 1990) (such as clinical trials). Further properties of this index are investigated in detail in Takeuchi and Hirotsu (1982), Nair (1986) and Hirotsu (1986); these CCS-type test can be obtained, in particular case, as sum (or weighted sum) of PD. Concerning to the Correspondence Analysis (CA) (Horst 1935; Fisher 1940), many authors (Beh 2001; D'Ambra et al. 2005; Sarnacchiaro and D'Ambra 2007; Lombardo et al. 2011) have proposed methods for analyzing a two-way contingency table preserving the information present in an ordinal variable using the orthogonal polynomials by means of the recurrence formula of Emerson (Emerson 1968) obtaining a Hybrid Decomposition (HD) or Bivariate Moment Decomposition (BMD), or by means of cumulative analysis: Cumulative Correspondence Analysis (CCS) (Beh et al. 2007, 2011; Sarnacchiaro 2011; D'Ambra et al. 2011; D'Ambra and Amenta 2011) and Doubly Cumulative Correspondence Analysis (DCCA)(D'Ambra et al. 2014; Camminatiello et al. 2021). The purpose of this work is to relate CCS type tests using a β parameter in order to obtain a new class called Weighted Cumulative Chi-Squared (WCCS) and to further extend this class to the PD family by obtaining a new index called Weighted Cumulative Power Divergence (WCPD). Moreover, for one particular family and after choosing particular β values, a variant of weighted CA based on cumulative frequencies called Weighted Cumulative Correspondence Analysis (hereafter WCCA) and its properties, by means of Generalized Singular Value Decomposition (GSVD), are shown. The subsequent contents of this article are organized as shown in the following. The used notation is defined in Sect. 2, whilst the WCPD-type index is introduced in Sect. 3. In Sect. 4 the extension of CA using the WCCS-type statistic is shown. Moreover, an additional property of this extension is introduced in Sect. 5. In Sect. 6 the confidence circle of the CA extension described in Sect. 4, is shown. In Sect. 7 our approach is illustrated by means of two empirical studies about the satisfaction with university curriculum counselor service and the CO_2 light-duty vehicle evaluation relating to different type approval. Some final remarks on this approach are highlights in the final section.

2 Notation

Assume that X and Y be categorical variables with $i = 1, \ldots, I$ and $j = 1, \ldots, J$ categories, respectively, and denote $(X_1, Y_1), \ldots, (X_n, Y_n)$ a random sample of the random vector (X, Y) where n is the fixed and known total number of observations. Also pose that N_{ij} be the random variable which counts the number of observations that fall into the cross-category $i \times j$, while $N_{i\bullet}$ and $N_{\bullet j}$ represent the counts for the categories i and j, respectively. Under a multinomial model with an ordinal variable, let $\mathbf{N} = (n_{ij})$ of size $I \times J$. The column variable could be assumed to be ordinal with increasing scores, without loosing generality. Moreover, we can indicate p_{ij} the proportion of observations that fall in the i-th row and j-th column ($j = 1, \ldots, J$) of the table and the generic element of matrix $\mathbf{P}$. Consequently,

we denote $\mathbf{D}_I$ and $\mathbf{D}_J$ as the diagonal matrices of the row and column marginal proportions $p_{i\bullet}$ and $p_{\bullet j}$, respectively, where $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$ and $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$. F_s represents the cumulative total of Y evaluated in s, that is $F_Y(s) = Pr(Y \leq s) = \sum_{j=1}^{s} p_{\bullet j} = p_s$. Finally, let $C_{is} = \sum_{j=1}^{s} N_{ij}$ and $C_s = \sum_{j=1}^{s} N_{\bullet j}$ be the cumulative count and the cumulative column total up to the s-th column category, respectively, where $s = 1, \ldots, J-1$.

3 Weighted cumulative power divergence family

Taguchi's statistic (Taguchi 1966, 1974) is a simple alternative to Pearson's test to measure the association between categorical variables in the case of one of them possesses ordered categories. However, the Pearson's statistic does not take into account the structure of ordered categorical variables (Agresti 2007). To deal with this issue, Taguchi (1966, 1974) developed a statistic that takes into consideration the structure of an ordered categorical variable. To assess the association between the nominal and ordered column variables, Taguchi (1966, 1974) proposed the following statistic

$$T_E = \sum_{s=1}^{J-1} \frac{1}{d_s(1-d_s)} \sum_{i=1}^{I} N_{i\bullet} \left(\frac{C_{is}}{N_{i\bullet}} - d_s \right)^2$$

with $0 \leq T_E \leq n(J-1)$ and $d_s = C_s/n$ is the cumulative column proportion up to s-th column. Moreover, in Takeuchi and Hirotsu (1982) and Nair (1986) the authors explain that the T_E statistic is linked to the Pearson chi-squared statistic $T_E = \sum_{s=1}^{J-1} X_s^2$ where X_s^2 is Pearson's chi-squared for the $I \times 2$ contingency tables obtained by aggregating the first s column categories and the remaining categories $(s+1)$ to J, respectively. For this reason, the Taguchi's statistic T_E is called the Cumulative Chi-Squared (CCS) statistic. It's possible to define the class of WCCS-type tests (Weighted Cumulative Chi-Squared) considering a given set of weights $w_s > 0$

$$T_{WCCS} = \sum_{s=1}^{J-1} w_s \sum_{i=1}^{I} N_{i\bullet} \left(\frac{C_{is}}{N_{i\bullet}} - d_s \right)^2$$

In Nair (1986) this class is called CCS-type tests. The choice of different weighting schemes defines the members of this class. Possible choices for w_s could be obtained assigning constant weights to each term (i.e. $w_s = 1/J$) or assume it proportional to the inverse of the conditional expectation of the s-th term under the null hypothesis of independence (i.e. $w_s = [d_s(1-d_s)]^{-1}$). It is evident that T_{CCS} subsumes T_E as special case. Moreover, author in Nair (1986) illustrates that T_{CCS} with $w_s = 1/J$ defined as

$$T_N = \sum_{s=1}^{J-1} \frac{1}{J} \sum_{i=1}^{I} N_{i\bullet} \left(\frac{C_{is}}{N_{i\bullet}} - d_s \right)^2$$

has good power against ordered alternatives with $0 \leq T_N \leq \frac{n}{J} \sum_{s=1}^{J-1} d_s(1-d_s)$. Moreover, it's possibile to generalize the T_{WCCS} using the parameter $\beta \in \Re$, in the following manner

$$T_{WCCS}^{(\beta)} = \sum_{s=1}^{J-1} w_s^{\beta} \sum_{i=1}^{I} N_{i\bullet} \left(\frac{C_{is}}{N_{i\bullet}} - d_s \right)^2$$

Here $w_s^\beta = [\frac{1}{d_s(1-d_s)}]^\beta$ and $-\infty < \beta < +\infty$, with reasonable values of $\beta \in [-3; 3]$. In particular we have $T_{WCCS}^{(-3)} \leq T_{WCCS}^{(0)} = JT_N \leq T_{WCCS}^{(1)} \equiv T_{CCS} = T_E \leq T_{WCCS}^{(3)}$. It's possible to explain the link between $T_{WCCS}^{(\beta)}$ and Pearson's chi-squared for the $I \times 2$ contingency tables obtained by aggregating the first s column categories and the remaining categories $(s+1)$ to J as in the following formula

$$T_{WCCS}^{(\beta)} = \sum_{s=1}^{J-1} w_s^\beta d_s(1-d_s)X_s^2 = \sum_{s=1}^{J-1} h_s^{(\beta)} X_s^2 \tag{1}$$

where $h_s^{(\beta)} = w_s^\beta d_s(1-d_s) = [d_s(1-d_s)]^{(1-\beta)}$. The authors have been shown in D'Ambra et al. (2018) that T_E is like Leti's unlikability coefficient (Leti 1983) $\tilde{D} = 2n\sum_{s=1}^{J-1} w_s F_s(1-F_s)$. In the previous equation, the cumulative total of Y (F_s) is written as a weighted sum of the cumulative distributions of the conditional variable $(Y|X=i)$ evaluated in s

$$F_{s|i} = Pr(Y \leq s|X=i) = \sum_{j=1}^{s} \frac{p_{ij}}{p_{i\bullet}} = \frac{p_{is}}{p_{i\bullet}}$$

where

$$F_s = p_s = \sum_{i=1}^{I} p_{is} = \sum_{i=1}^{I} p_{i\bullet} F_{s|i}$$

is the weighted mean of the $F_{s|i}$. Indeed, $\tilde{D}/2$ can be partitioned in a sum of two orthogonal components according to the well-known principle of between and within group variance (D'Ambra et al. 2018). This result can be extended also to our case, in particular:

$$\begin{aligned}
\frac{\tilde{D}^{(\beta)}}{2} &= n\sum_{s=1}^{J-1} w_s^\beta F_s(1-F_s) \\
&= \underbrace{n\sum_{i=1}^{I} p_{i\bullet}\sum_{s=1}^{J-1} w_s^\beta F_{s|i}(1-F_{s|i})}_{D_W^{(\beta)}} + \underbrace{n\sum_{i=1}^{I} p_{i\bullet}\sum_{s=1}^{J-1} w_s^\beta (F_{s|i}-F_s)^2}_{D_B^{(\beta)}} \\
&= D_W^{(\beta)} + n\sum_{s=1}^{J-1} w_s^\beta \sum_{i=1}^{I} p_{i\bullet}\left(\frac{p_{is}}{p_{i\bullet}} - F_s\right)^2 \\
&= D_W^{(\beta)} + T_{WCCS}^{(\beta)}
\end{aligned}$$

Thus, we can defined the following relative index

$$R_{(\beta)}^2 = \frac{T_{WCCS}^{(\beta)}}{\tilde{D}^{(\beta)}/2}$$

It's easy to show that when $\beta = 0$ we have

$$R_{(0)}^2 = \frac{JT_N}{n\sum_{s=1}^{J-1} d_s(1-d_s)} = \frac{T_N}{\frac{n}{J}\sum_{s=1}^{J-1} d_s(1-d_s)}$$

while if $\beta = 1$ we have

$$R^2_{(1)} = \frac{T_E}{n(J-1)}$$

It's possible to note that the $R^2_{(\beta)}$ can be seen as a weighted mean of $\phi^2_s = X^2_s/n$. Indeed

$$R^2_{(\beta)} = \frac{T^{(\beta)}_{WCCS}}{\tilde{D}^{(\beta)}/2} = \frac{\sum_{s=1}^{J-1} h^{(\beta)}_s \phi^2_s}{\sum_{s=1}^{J-1} h^{(\beta)}_s} = \sum_{s=1}^{J-1} \omega^{(\beta)}_s \phi^2_s$$

with $\omega^{(\beta)}_s = \frac{h^{(\beta)}_s}{\sum_{s=1}^{J-1} h^{(\beta)}_s}$ and $\sum_{s=1}^{J-1} \omega^{(\beta)}_s = 1$. For this reason $\min[\phi^2_s] \leq R^2_{(\beta)} \leq \max[\phi^2_s]$ where equalities hold in case of non-association or perfect dependence. Following the Power Divergence Family (PD) approach proposed by Cressie and Read (1989), the formula (1) could be extended by introducing the parameter λ, with $\lambda \in \Re$. In this way we could define a new family of cumulative indices called Weighted Cumulative Divergence of Power (WCPD):

$$\begin{aligned} T^{(\lambda;\beta)}_{WCPD} &= \frac{2n}{\lambda(\lambda+1)} \left\{ \left[\sum_{s=1}^{J-1} h^{(\beta)}_s \sum_{i=1}^{I} \frac{p^{(1+\lambda)}_{is}(1-d_s)^{\lambda} + (p_{i\bullet} - p_{is})^{(1+\lambda)} d^{\lambda}_s}{p^{\lambda}_{i\bullet} d^{\lambda}_s (1-d_s)^{\lambda}} \right] - 1 \right\} \\ &= \sum_{s=1}^{J-1} h^{(\beta)}_s \Im^{(\lambda)}_s \end{aligned}$$

where $\Im^{(\lambda)}_s$ is the PD family for s sub-table with $T^{(1;\beta)}_{WCPD}$ and $T^{(0;\beta)}_{WCPD}$ defined by continuity. It's easy to demonstrate that when $\lambda = 1$

$$T^{(1;\beta)}_{WCPD} \equiv T^{(\beta)}_{WCCS}$$

This new family includes for $\lambda = 1$ the weighted sum of Pearson's X^2_s ($T^{(\beta)}_{WCCS}$), for $\lambda = -2$ the weighted sum of Neyman's called Weighted Cumulative Neyman's statistic ($T^{(\beta)}_{WCN} = \sum_{s=1}^{J-1} w^{(\beta)}_s \tilde{X}^2_s$), for $\lambda = -1$ the weighted sum of Likelihood Ratio Modified called Weighted Cumulative Likelihood Ratio Modified statistic ($T^{(\beta)}_{WCLRm} = \sum_{s=1}^{J-1} w^{(\beta)}_s \tilde{G}^2_s$), for $\lambda = -1/2$ the weighted sum of Freeman-Tukey's statistic called Weighted Cumulative Freeman-Tukey's statistic ($T^{(\beta)}_{WCFT} = \sum_{s=1}^{J-1} w^{(\beta)}_s T^2_s$), for $\lambda = 0$ the weighted sum of Likelihood Ratio called Weighted Cumulative Likelihood Ratio statistic ($T^{(\beta)}_{WCLR} = \sum_{s=1}^{J-1} w^{(\beta)}_s G^2_s$). Finally for $\lambda = 2/3$ the weighted sum of Cressie and Read solution called Weighted Cumulative Cressie and Read statistic ($T^{(\beta)}_{WCCR} = \sum_{s=1}^{J-1} w^{(\beta)}_s CR^2_s$) is defined. We note that for $\beta = 1$ we have obtained, for different λ values, the unweighted indices called Cumulative Power Divergence (CPD)

$$T^{(\lambda)}_{CPD} \equiv T^{(\lambda;1)}_{WCPD} = \sum_{s=1}^{J-1} w_s d_s (1-d_s) \Im^{(\lambda)}_s = \sum_{s=1}^{J-1} \Im^{(\lambda)}_s$$

4 Correspondence analysis based on $T^{(\beta)}_{WCCS}$

In this section the variant of weighted CA based on cumulative frequencies C_{is} called Weighted Cumulative Correspondence Analysis (hereafter WCCA) is shown. In this con-

text we consider only value $\lambda = 1$, since in this case we have a spectral decomposition of the $T^{(\beta)}_{WCCS}$. In other case, when the $\lambda \neq 1$ the Generalize Singular Value Decomposition (GSVD) cannot be used. The aim is to obtain a graphical representation in a reduced space of the row and column categories. In fact, this new approach graphically determines the similar cumulative categories C_{is} with respect to nominal ones, according to the choice of w^{β}_{s}. We denote it by the following formula

$$\mathbf{B}^{(\beta)} = \mathbf{D}_I^{-1}\mathbf{P}\mathbf{X}^T(\mathbf{W}^{\beta})^{\frac{1}{2}} \tag{2}$$

where $\mathbf{W}^{\beta}$ is diagonal square matrix of dimension $[(J-1)\times(J-1)]$ of general term w^{β}_{s} and $\mathbf{X}$ is a matrix of dimension $[(J-1)\times J)]$ formed in this way:

$$\mathbf{X} = \mathbf{L} - [\mathbf{D}(\mathbf{1}_{J-1}\mathbf{1}_J^T)]$$

where $\mathbf{L}$ is a lower triangular matrix of dimension $[(J-1)\times J)]$, and $\mathbf{1}$ is a vector of ones of appropriate dimension and $\mathbf{D} = diag(d_s)$. The Correspondence Analysis based on cumulative frequencies provides the following:

$$GSVD[\mathbf{B}^{(\beta)}]_{\mathbf{D}_I;\mathbf{I}} \Rightarrow \mathbf{B}^{(\beta)} = \mathbf{U}^{(\beta)}\mathbf{\Lambda}^{(\beta)}[\mathbf{V}^{(\beta)}]^T$$

where $\mathbf{U}^{(\beta)}$ is the $I \times M$ matrix of left singular vectors (with M=rank$[\mathbf{B}^{(\beta)}]$) such that $[\mathbf{U}^{(\beta)}]^T\mathbf{D}_I\mathbf{U}^{(\beta)} = \mathbf{I}$ and $\mathbf{V}^{(\beta)}$ is the $(J-1)\times M$ matrix of right singular vectors such that $[\mathbf{V}^{(\beta)}]^T\mathbf{V}^{(\beta)} = \mathbf{I}$, and $\mathbf{\Lambda}^{(\beta)}$ is a positive definite diagonal matrix of order M where the (m,m)-th element $\lambda^{(\beta)}_m$ is the m-th singular vector of $\mathbf{B}$. In particular, the total inertia can be expressed in terms of $\mathbf{B}$ so that

$$\frac{T^{(\beta)}_{WCCS}}{n} = \sum_{m=1}^{M}[\lambda^{(\beta)}_m]^2 = trace\{[\mathbf{B}^{(\beta)}]^T\mathbf{D}_I\mathbf{B}^{(\beta)}\} = \sum_{i=1}^{I}\sum_{s=1}^{J-1} p_{i\bullet}[b^{(\beta)}_{is}]^2 \tag{3}$$

To visually summarise the association between the row and the column categories, we define the row and column principal coordinates by

$$\mathbf{F}^{(\beta)} = \mathbf{U}^{(\beta)}\mathbf{\Lambda}^{(\beta)} \tag{4}$$

and

$$\mathbf{G}^{(\beta)} = \mathbf{V}^{(\beta)}\mathbf{\Lambda}^{(\beta)} \tag{5}$$

Here, $\mathbf{F}^{(\beta)}$ and $\mathbf{G}^{(\beta)}$ are of size $I \times M$ and $(J-1)\times M$. The s-th row of matrix $\mathbf{G}^{(\beta)}$ contains the coordinates of category $y(1:s)$ (see Table 1) in the M-dimensional space. However, it is well known that the proximity evaluation between a row principal coordinate and a column principal coordinate has been long questioned, because $\mathbf{F}^{(\beta)}[\mathbf{G}^{(\beta)}]^T \neq \mathbf{B}^{(\beta)}$ (see (2)). Authors in Gabriel (1971) proposed a rescaling of (4) and (5) yielding a biplot display. They obtained

$$\mathbf{F}^{(\beta)} = \mathbf{U}^{(\beta)}[\mathbf{\Lambda}^{(\beta)}]^{\delta}$$

and

$$\mathbf{G}^{(\beta)} = \mathbf{V}^{(\beta)}[\mathbf{\Lambda}^{(\beta)}]^{1-\delta}$$

for $0 \leq \delta \leq 1$. In Lombardo et al. (1996) authors referred to such plots as "column isometric", "symmetric" and "row isometric" factorisations, respectively. The "row isometric" plot ($\delta = 1$) is then achieved by jointly plotting the row principal coordinates (4) and the column

Table 1 Cross-classification of 493 students satisfaction with university curriculum counselor service and its use

	Satisfaction level				
	(I)	(II)	(III)	(IV)	Total
Use 1	6	24	31	5	66
Use 2–4	5	42	49	5	101
Use 5–6	6	87	57	5	155
Use 7–8	11	73	44	7	135
Use 8+	6	14	10	6	36
	33	240	191	28	493

standard coordinates $\mathbf{G}^{(\beta)} = \mathbf{V}^{(\beta)}$. Similarly, the "column isometric" biplot is obtained by using the row standard $\mathbf{F}^{(\beta)} = \mathbf{U}^{(\beta)}$ and the column principal coordinates (5). In Nair (1986, 1987) it is pointed out that the T_{WCCS} statistic can be approximated using Satterthwaite's method (Satterthwaite 1946). Indeed, letting $\mathbf{\Gamma}^{(\beta)}$ be the $(J-1)\times(J-1)$ diagonal matrix of the nonzero eigen values $\gamma_s^{(\beta)}$ of matrix $\mathbf{X}^T\mathbf{W}^\beta\mathbf{X}\mathbf{D}_J$, then

$$GSVD(\mathbf{X})_{\mathbf{W}^\beta;\mathbf{D}_J} \Rightarrow \mathbf{X} = \mathbf{Z}^{(\beta)}[\mathbf{\Gamma}^{(\beta)}]^{\frac{1}{2}}[\mathbf{Q}^{(\beta)}]^T$$

where $[\mathbf{Z}^{(\beta)}]^T\mathbf{W}^\beta\mathbf{Z}^{(\beta)} = [\mathbf{Q}^{(\beta)}]^T\mathbf{D}_J\mathbf{Q}^{(\beta)} = \mathbf{I}$ and $\mathbf{X}^T\mathbf{W}^\beta\mathbf{X} = [\mathbf{Q}^{(\beta)}]^T\mathbf{\Gamma}^{(\beta)}\mathbf{Q}^{(\beta)}$. Thus, for $\mathbf{K}^{(\beta)} = \mathbf{D}_I^{\frac{1}{2}}\mathbf{P}\mathbf{Q}^{(\beta)}$

$$\begin{aligned}
T_{WCCS}^{(\beta)} &= n \times trace\left\{\mathbf{D}_I^{\frac{1}{2}}\mathbf{P}\mathbf{X}^T\mathbf{W}^\beta\mathbf{X}\mathbf{P}^T\mathbf{D}_I^{\frac{1}{2}}\right\} \\
&= n \times trace\left\{\mathbf{D}_I^{\frac{1}{2}}\mathbf{P}[\mathbf{Q}^{(\beta)}]^T\mathbf{\Gamma}^{(\beta)}\mathbf{Q}^{(\beta)}\mathbf{P}^T\mathbf{D}_I^{\frac{1}{2}}\right\} \\
&= n \times trace\left\{\mathbf{K}^{(\beta)}\mathbf{\Gamma}^{(\beta)}[\mathbf{K}^{(\beta)}]^T\right\} \\
&= n\sum_{s=1}^{J-1}\gamma_s^{(\beta)}\sum_{i=1}^{I}[k_{is}^{(\beta)}]^2
\end{aligned}$$

In Nair (1986, 1987) authors evidence that, as $n \to \infty$ the quantity $n \times \sum_{i=1}^{I}[k_{is}^{(\beta)}]^2$ is an asymptotically (central) chi-squared distribution for s-th component with $(I-1)$ d.f. under the null hypothesis: $n \times \sum_{i=1}^{I}[k_{is}^{(\beta)}]^2 \sim \chi^2_{(I-1)}(s)$. Consequently, as $n \to \infty$, the limiting distribution of $T_{WCCS}^{(\beta)}$ is then a linear combination of iid chi-squared random variables

$$T_{WCCS}^{(\beta)} \to_{H_0}^{D} \sum_{s=1}^{J-1}\gamma_s^{(\beta)}\chi^2_{(I-1)}(s)$$

In Nair (1987) the author indicates that, by using Satterthwaite's two-moment approximation (Satterthwaite 1946), this distribution can be approximated as

$$T_{WCCS}^{(\beta)} \sim r^{(\beta)}(I-1)\chi^2_{v^{(\beta)}}$$

with

$$r^{(\beta)} = \frac{1}{(I-1)}\frac{\sum_{s=1}^{J-1}(\gamma_s^{(\beta)})^2}{\sum_{s=1}^{J-1}\gamma_s^{(\beta)}} \qquad v^{(\beta)} = \frac{1}{r^{(\beta)}}\sum_{s=1}^{J-1}\gamma_s^{(\beta)}$$

For this reason, we can be defined the following statistical test

$$\tilde{T}^{(\beta)}_{WCCS} \sim \chi^2_{v^{(\beta)}}$$

with $\tilde{T}^{(\beta)}_{WCCS} = \frac{T^{(\beta)}_{WCCS}}{r^{(\beta)}(I-1)}$. In particular the degree of freedom can be written in this way:

$$v^{(\beta)} = \frac{(I-1)(J-1)}{\eta^{(\beta)}}$$

with $\eta^{(\beta)} = \frac{(J-1)\sum_{s=1}^{J-1}(\gamma_s^2)^{(\beta)}}{(\sum_{s=1}^{J-1}\gamma_s^{(\beta)})^2} > 1$. Moreover, given $\tilde{T}^{(\beta)}_{WCCS}$ and the contingency table composed by I independent rows, it's possible to measure which of them is statistically significant. In particular:

$$\tilde{T}^{(\beta)}_{WCCS}(i) = n\frac{1}{r^{(\beta)}(I-1)}\sum_{s=1}^{J-1} w_s^{\beta} N_{i\bullet}\left(\frac{C_{is}}{N_{i\bullet}} - d_s\right)^2 \sim \chi^2_{\frac{(J-1)}{\eta^{(\beta)}}} \tag{6}$$

Regards to WCPD, the degrees of freedom are invariant for different values of λ (Cressie and Read 1984, 1989). In this context the s-th components are a solution of the following expression

$$T^{(\lambda;\beta)}_{WCPD} - \sum_{s=1}^{J-1}\gamma_s^{(\beta)}\Im^{(\lambda)}(s) = 0$$

with constraint $\sum_{s=1}^{J-1}\Im^{(\lambda)}(s) = \Im^{(\lambda)}$, where $\Im^{(\lambda)}$ is the PD family calculated on the original contingency table $\mathbf{N}$. Thus, we can define the following statistic:

$$\tilde{T}^{(\lambda;\beta)}_{WCPD} = \frac{T^{(\lambda;\beta)}_{WCPD}}{r^{(\beta)}(I-1)} \sim \chi^2_{\frac{(I-1)(J-1)}{\eta^{(\beta)}}}$$

In this context, one of the fundamental problem to solve is choosing an appropriate value of β. A possible approach for defining the β choice is to draw a plot that shows the trend of the statistic $\tilde{T}^{(\lambda;\beta)}_{WCPD}$ and the critical value $\chi^2_{\frac{(I-1)(J-1)}{\eta^{(\beta)}};1-\alpha}$ as a function of the β parameter. The values of the β parameter will be the one that maximizes the difference between $\tilde{T}^{(\lambda;\beta)}_{WCPD}$ and $\chi^2_{\frac{(I-1)(J-1)}{\eta^{(\beta)}};1-\alpha}$. In this way, the β value will be defined to obtain the statistic $\tilde{T}^{(\lambda;\beta)}_{WCPD}$ furthest from the hypothesis of independence.

5 Further properties

In this section, the proprieties of WCCA are showed. Let's consider now the i-th row of (4). Its generic value $f^{(\beta)}_{im}$ is the principal coordinate of the i-th row along the m-th dimension of the correspondence plot

$$f^{(\beta)}_{im} = u^{(\beta)}_{im}\lambda^{(\beta)}_m$$

and it is centred at the origin of the space $\Re^M$. Therefore, the squared Euclidean distance of the i-th row coordinate from the origin of the plot is

$$d^2(i,0) = \sum_{m=1}^{M} [f_{im}^{(\beta)}]^2 = \sum_{s=1}^{J-1} w_s^{\beta} \left(\frac{C_{is}}{N_{i\bullet}} - d_s \right)^2$$

Note that Taguchi's inertia can be expressed in terms of this distance so that

$$\frac{T_{WCCS}^{(\beta)}}{n} = \sum_{i=1}^{I} p_{i\bullet} d^2(i,0)$$

We can also obtain, a similar expression for the ordered column categories. To prove this result, we examine the s-th row of (5). The coordinate of the s-th pair of cumulative categories on the m-th principal dimension of the correspondence plot is given by

$$g_{sm}^{(\beta)} = v_{sm}^{(\beta)} \lambda_m^{(\beta)}$$

Therefore, the squared Euclidean distance of the s-th column coordinate from the origin of the plot is:

$$d^2(s,0) = \sum_{m=1}^{M} [g_{sm}^{(\beta)}]^2 = h_s^{(\beta)} \phi_s^2 = h_s^{(\beta)} \frac{X_s^2}{n} \tag{7}$$

so that $T_{WCCS}^{(\beta)}$ inertia is

$$\sum_{s=1}^{J-1} d^2(s,0) = \frac{T_{WCCS}^{(\beta)}}{n}$$

Equation (7) points out that the squared Euclidean distance of the s-th cumulated category from the origin amounts to the weighted $\phi_s^2 = \frac{X_s^2}{n}$ computed on the s-th contingency table of size $I \times 2$. This implies that

$$T_{WCCS}^{(\beta)} = n \sum_{s=1}^{J-1} \sum_{m=1}^{M} [g_{sm}^{(\beta)}]^2 \tag{8}$$

Each X_s^2 is asymptotically a chi-squared random variable with $(I-1)$ df, under the null hypothesis of independence. It can be proofed that it is possible to partition this $T_{WCCS}^{(\beta)}$ statistic (7) into orthogonal components, each of which is a chi-squared random variable with degree of freedom equal to 1

$$\frac{n[g_{sm}^{(\beta)}]^2}{h_s^{(\beta)}} \rightarrow_{H_0}^{d} \chi_1^2$$

This result allows us to identify which are the significant cumulative categories by constructing confidence circles, for each cumulative category. If it includes the origin of the axes, then it is not significant. In this way we will refer to the independence regions inside a circle in a two-dimensional plot. It highlights the cumulative categories that haven't a statistically significant contribution to the association between the row and column variables at a predefined α (confidence circle). The centre of this independence region is g_{sm} while the radii

length of the $100(1-\alpha)\%$ confidence circle in a two-dimensional correspondence plot is given by

$$r_s^{(\beta)} = \sqrt{\frac{h_s^{(\beta)} \chi_\alpha^2}{n}} \tag{9}$$

Here χ_α^2 is the $1-\alpha$ percentile of a chi-squared distribution with two degrees of freedom (Lebart et al. (1984); Beh and D'Ambra (2010)). The radii length r_s of the s-th cumulated categories confidence circle is defined by (9). These circles provide a way to identify which cumulative categories are consistent with the independence hypothesis and which are not. Indeed, if the confidence circle $100(1-\alpha)$ includes the origin of the axes, the cumulative s-th category is consistent with the independence hypothesis. Moreover, let $\mathbf{g}_s^{(\beta)} = (g_{s1}^{(\beta)}, \ldots, g_{sM}^{(\beta)})$ and $\mathbf{g}_{s'}^{(\beta)} = (g_{s'1}^{(\beta)}, \ldots, g_{s'M}^{(\beta)})$ be the vector of coordinates of the cumulated categories s (first column of sub-table $\mathbf{N}_s$), and s' (first column of sub-table $\mathbf{N}_{s'}$) respectively, in the M-dimensional space. In this way, it is possible to calculate the similarity between the cumulated categories s and s' by computing Tucker's congruence coefficient $\psi_{ss'}$ referred to as the cosine similarity (Tucker (1951); Abdi (2007))

$$\psi_{ss'}^{(\beta)} = \frac{(\mathbf{g}_s^{(\beta)})^T \mathbf{g}_{s'}^{(\beta)}}{||\mathbf{g}_s^{(\beta)}|| \times ||\mathbf{g}_{s'}^{(\beta)}||} = \cos(\theta) \tag{10}$$

where θ is the angle between $\mathbf{g}_s$ and $\mathbf{g}_{s'}$. Here, ψ is the correlation of vectors $\mathbf{g}_s$ and $\mathbf{g}_{s'}$ about their origin (or "zero"), whereas Pearson's correlation coefficient is based on the deviations from their respective means. If $\psi_{ss'}$ is near to 1 (almost collinear), then the row profiles of sub-tables $\mathbf{N}_s$ and $\mathbf{N}_{s'}$ appear to be proportional whilst this characteristic is lost when $\psi_{ss'}$ is near to 0 (almost orthogonal vectors). When $\psi_{ss'}$ is near to 1 or -1, then all of the row profiles play the same (or inverse) role in explaining the cumulated categories s and s'. Different roles are played when $\psi_{ss'}$ is near to 0. Instead, the row profiles appear inversely proportional when $\psi_{ss'}$ is near to -1. Moreover, it is important to point out that the Tucker's congruence coefficient $\psi_{ss'}^{(\beta)}$ is invariant for several values of β.

6 Confidence circles

In the typical CA, the confidence circles proposed by Lebart et al. (1984) are a useful tool to check if a particular row category is significant. Generally, if the origin lies outside the confidence circle for a particular category, then that category contributes to the dependency between the row and column categories of the contingency table. If the origin lies within the circle for a particular category, then that category does not contribute to the dependence between the variables. These circles are similar to the regions that in Mardia et al. (1982) authors derived for canonical variate analysis, while in Ringrose (1992, 1996) they also explored their use for CA by means of bootstrap procedure. In classical Cumulative Correspondence Analysis (CCA), based on decomposition of Taguchi's index (Taguchi 1966, 1974), the confidence circles have been shown by D'Ambra et al. (2021). This concept can also be extended in WCCA. Suppose that a two-way contingency table consists of a row (predictor) and column (response) variable that is asymmetrically structured. The $\tilde{T}_{WCCS}^{(\beta)}$

can be expressed in terms of the predictor coordinates such that

$$\tilde{T}^{(\beta)}_{WCCS} = \frac{n}{r^{(\beta)}(I-1)} \sum_{i=1}^{I} \sum_{m=1}^{M} p_{i\bullet}[f^{(\beta)}_{im}]^2 \sim \chi^2_{\frac{(I-1)(J-1)}{\eta^{(\beta)}}}$$

For the i-th row coordinate

$$\frac{n}{r^{(\beta)}(I-1)} \sum_{m=1}^{M} p_{i\bullet}[f^{(\beta)}_{im}]^2 \sim \chi^2_{\frac{(J-1)}{\eta^{(\beta)}}} \tag{11}$$

Equation (11) is equivalent to (6) defined in Sect. 4. Therefore, it results that

$$\sum_{m=1}^{M} p_{i\bullet}[f^{(\beta)}_{im}]^2 \sim \frac{r^{(\beta)}(I-1)}{n} \chi^2_{\frac{(J-1)}{\eta^{(\beta)}}}$$

When the variables of a bivariate table are considered symmetrically related, as in the case of CA, the confidence circles approach is used to identify those categories that contribute most to the hypothesis independence test (Lebart et al. 1984). These circles are similar to those used in canonical analysis (Mardia et al. 1982). Another approach to calculate them is based on a bootstrap procedure. In the case that the bivariate table concerns to ordinal variables, it has been described how the radii of these circles are identical to those in Lebart et al. (1984). The relationship between the i-th column coordinate for the two more important components of a two-dimensional plot is

$$p_{i\bullet}[f^{(\beta)}_{i1}]^2 + p_{i\bullet}[f^{(\beta)}_{i2}]^2 \sim \frac{r^{(\beta)}(I-1)}{n} \chi^2_{\frac{2}{\eta^{(\beta)}}}$$

At the α level of significance, this can be expressed as

$$[f^{(\beta)}_{i1}]^2 + [f^{(\beta)}_{i2}]^2 \sim \frac{r^{(\beta)}(I-1)}{N_{i\bullet}} \chi^2_{\frac{2}{\eta^{(\beta)}},\alpha}$$

Therefore, the $1-\alpha$ confidence circle for the i-th column coordinate in the two-dimensional plot has a radii length

$$r^{(\beta)}_i = \sqrt{\frac{r^{(\beta)}(J-1) \times \chi^2_{\frac{2}{\eta^{(\beta)}}\alpha}}{N_{i\bullet}}} \tag{12}$$

Note that it depends on the i-th marginal proportion classified into that category. Thus, for a very small ranking in the i-th predictor category, the radii length will be relatively large. Likewise, for a relatively broad ranking, the radii length will be relatively small.

7 Empirical studies

7.1 Empirical study 1—satisfaction with university curriculum counselor service

The Table 1 shows a contingency table between the satisfaction of university curriculum counselor service and its use. The row variable, named Use, is the X categorical variables (with $I = 5$) and it represents the use of university curriculum counselor service. It is on ordinal scale with five categories: one time (Use 1), from 2 to 4 times (Use 2–4), 5 or 6 times

Table 2 $T^{(\lambda)}_{CPD}$ as sum of Power Divergence index of the sub-tables

s	$\tilde{X}^2_s$ $\lambda=-2$	$\tilde{G}^2_s$ $\lambda=-1$	T^2_s $\lambda=-0.5$	G^2_s $\lambda=0$	CR_s $\lambda=2/3$	X^2_s $\lambda=1$
1	7.767	7.608	7.713	7.959	8.549	8.982
2	9.846	9.756	9.733	9.719	9.719	9.727
3	6.638	7.002	7.439	8.106	9.485	10.459
$T^{(\lambda)}_{CPD}$	24.252	24.369	24.886	25.784	27.753	29.168

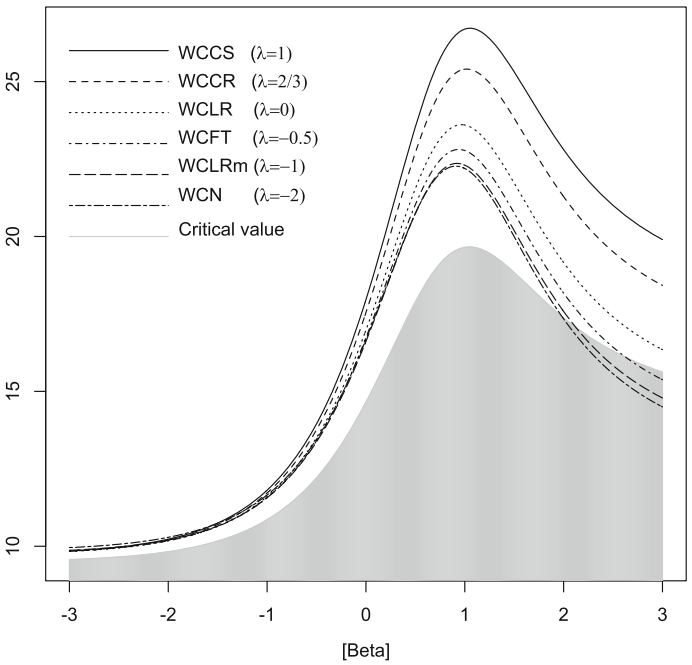

Fig. 1 $\tilde{T}^{(\lambda;\beta)}_{WCPD}$ statistic and critical value for $1-\alpha=0.95$ in function of β

(Use 5–6), 7 or 8 times (Use 7–8), more than 8 times (Use 8+). The column variable is the Y categorical variables (with $J=4$) an it defines the satisfaction of university curriculum counselor service. The variable is on ordinal scale from 1 (dissatisfied) to 4 (extremely satisfied). This dataset was examined also by Camminatiello et al. (2021).

Table 2 shows the $T^{(\lambda)}_{CPD}$ indices for different λ values. It is clear that the optimal subtable that has the greatest contribution is (I: II) vs (III: IV) ($s=2$) for each λ, except for the last index where the optimal subtable is (I: III) versus (IV) ($s=3$). Moreover, the critical value with $1-\alpha=0.95$ and 4 degree of freedom is 9.488, therefore only statistical index of the optimal sub-tables is significant. The last index value produces an exception because in this case the sub-tables (I:II) vs (III: IV) (s=2) and (I: III) vs (IV) (s=3) are significant.

In Fig. 1 the value for statistic $\tilde{T}^{(\lambda;\beta)}_{WCPD}$ with $\lambda\in(-2,-1,-0.5,0,2/3,1)$ and critical value, with $1-\alpha=0.95$ as function of β, is shown. In particular, the statistic $\tilde{T}^{(\lambda;\beta)}_{WCPD}$ results significative for each value of β with $\lambda\in(0,2/3,1)$ underlining a strong statisti-

Table 3 $T_{WCPD}^{(\lambda;\beta)}$ as Weighted sum of Power Divergence index of the sub-tables

s	$h_s^{(0.66)}\tilde{X}_s^2$ $\lambda=-2$	$h_s^{(0.69)}\tilde{G}_s^2$ $\lambda=-1$	$h_s^{(0.75)}T_s^2$ $\lambda=-0.5$	$h_s^{(0.83)}G_s^2$ $\lambda=0$	$h_s^{(0.98)}CR_s$ $\lambda=2/3$	$h_s^{(1.06)}X_s^2$ $\lambda=1$
1	3.054	3.248	3.883	4.990	8.092	10.590
2	6.120	6.325	6.861	7.662	9.451	10.578
3	2.454	2.826	3.579	4.929	8.946	12.467
$T_{WCPD}^{(\lambda;\beta)}$	11.628	12.400	14.323	17.582	26.489	33.636

Table 4 $T_{WCPD}^{(\lambda;\beta)}$ statistic significance

Index	λ	β	$T_{WCPD}^{(\lambda;\beta)}$	$r^{(\beta)}(I-1)$	$\tilde{T}_{WCPD}^{(\lambda;\beta)}$	$\eta^{(\beta)}$	d.f.	p-value
WCN	-2	0.66	11.628	0.536	21.683	1.162	10.326	0.0197
WCLRm	-1	0.69	12.400	0.567	21.854	1.151	10.425	0.0195
WCTF	-0.5	0.75	14.323	0.638	22.466	1.132	10.600	0.0174
WCLR	0	0.83	17.582	0.751	23.427	1.113	10.785	0.0139
WCCR	2/3	0.98	26.489	1.043	25.388	1.094	10.973	0.0079
WCCS	1	1.06	33.636	1.259	26.722	1.092	10.991	0.0050

Table 5 $T_{WCCS}^{(1.06)}$ inertia decomposition

Axis	Inertia	%	Cumulative %
1	0.04635	67.9377	67.9377
2	0.02091	30.6495	98.5871
3	0.00097	1.4129	100.0000
$\frac{T_{WCCS}^{(1.06)}}{n}$	0.06823	100	

cally significative association between the rows and column-aggregated sub-tables. On the contrary for $\lambda \in (-2,-1,-0.5)$ the statistic $\tilde{T}_{WCPD}^{(\lambda;\beta)}$ doesn't result significative for high β value. The β values that maximize the distance between $T_{WCPD}^{(\lambda;\beta)}$ and critical value for $\lambda \in [-2;-1;-1/2;0;2/3;1]$ are $\beta=0.66$, $\beta=0.69$, $\beta=0.75$, $\beta=0.83$, $\beta=0.98$ and $\beta=1.06$, respectively. The Table 3 explains the weighted statistics for each λ value related to all the subtables, the optimal sub-table doesn't change for each value of λ, but the importance of the sub-table is stronger in forming $T_{WCPD}^{(\lambda;\beta)}$. Table 4 shows $T_{WCPD}^{(\lambda;\beta)}$ index and $\tilde{T}_{WCPD}^{(\lambda;\beta)}$ statistic based on β values that maximize the distance between $\tilde{T}_{WCPD}^{(\lambda;\beta)}$ statistic and p value with $1-\alpha=0.95$, for $\lambda \in \{-2;-1;-0.5;0;2/3;1\}$. For $\lambda=1$ the Fig. 2 shows the row statistic calculated by (6) and critical value calculated with $1-\alpha=0.95$ in function of β. All the categories are not significant for $\beta<0.34$, while for $\beta \geq 0.34$ only the category "Use 8+" results significant.

From the point of view of CA based on spectral decomposition of $T_{WCCS}^{(\beta)} \equiv T_{WCPD}^{(1;\beta)}$, the choice of β falls in the value of 1.06, a value that maximizes the difference between $\tilde{T}_{WCCS}^{(\beta)}$ and critical values as shown in Fig. 1 and Table 4. By partitioning the inertia using $T_{WCCS}^{(1.06)}$, we obtain the principal inertia values which are summarised in Table 5.

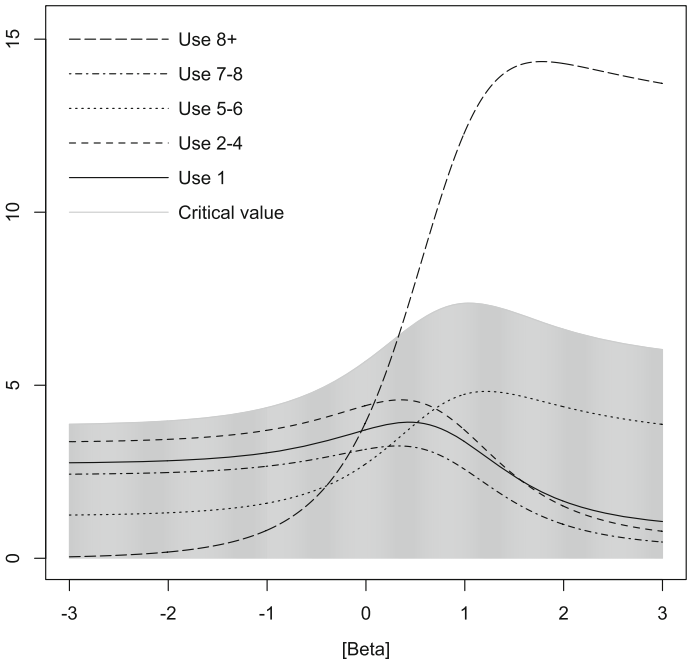

Fig. 2 $\tilde{T}^{(\beta)}_{WCCS}(i) \equiv \tilde{T}^{(1;\beta)}_{WCPD}(i)$ statistic and critical value for $1 - \alpha = 0.95$ in function of β

Table 6 Radii length of indipendence regions

	(I)	(I:II)	(I:III)
Ray	0.120	0.115	0.120

Figure 2 illustrates the WCCA plot (dimensions 1–2) of the cumulated column categories (principal coordinates) and the overlapped independence circle. This plot depicts 98.59% of the association that exists between the two variables in WCCA. For the column categories, the label "(I)" reflects the cumulative total of rating "(I)" and "(II:IV)" with those of ratings (II), (III) and (IV). Labels "(I:II)" and "(III:IV)" reflect instead the comparison related to the cumulative total of ordered rating from "(I)" to "(II)" with "(III)" and "(IV)". The remaining labels can be interpreted in a similar manner.

Figure 3 indicates that all contrasts based on the cumulative categories illustrate a valuable weight for the analysis because the independence region, with the radii length defined by (9) and summarized in Table 6, doesn't include the origin. Figure 3 highlights a strong opposite congruence between the first and last aggregated sub-tables. This configuration points out that the row categories, for the first and last sub-tables, show an opposite behaviour concerning the cumulated categories. This is not true for the second aggregated sub-table. The meaning of Fig. 3 is confirmed by reading Table 7 which indicates the similarity matrix amongst the s-th contingency tables of order $I \times 2$ according to (10). This matrix reveals that the values of the Tucker congruence coefficient for the first and second sub-tables are close to -1, with the exception of the second sub-table. This implies that all the row profiles of the first and second sub-tables play an inverse role in explaining the cumulated categories s and s'. The second sub-table plays a not clear role (values near to 0).

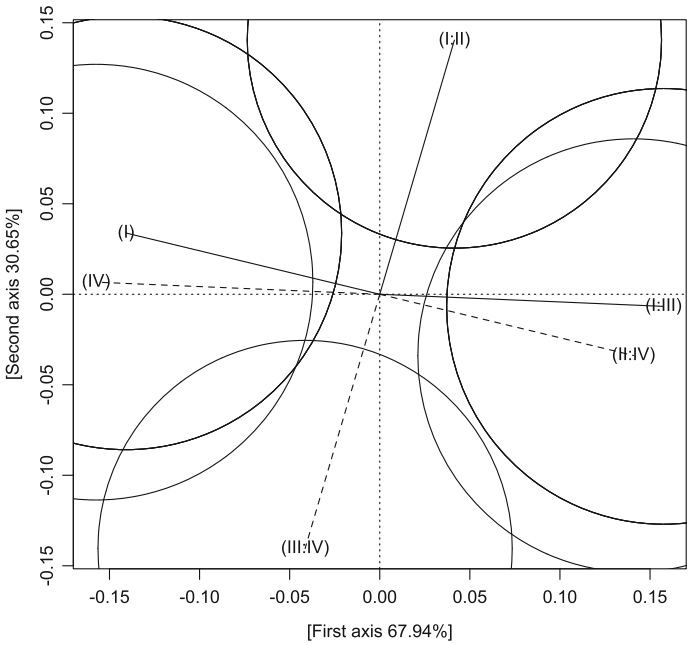

Fig. 3 WCCA plot: cumulated column categories and independence regions

Table 7 Similarity matrix

	(I)	(I:II)	(I:III)
(I)	1	–	–
(I:II)	− 0.055	1	–
(I:III)	− 0.941	0.236	1

In Fig. 4 we represent a two-dimensional biplot of the asymmetric association between the row and the column categories of Table 1. This plot is a row isometric biplot where the origin represents the column marginal distribution (that is the independence of the ordered columns from the rows). The confidence circles of the row categories, with radii length defined by (12) and summarised in Table 8, are then superimposed on the predictor categories. The Fig. 4 communicates its significance in explaining the relationship because its circle doesn't include the origin. Considering the rules of the row isometric biplot (Lombardo et al. 1996), Fig. 4 highlights the role played by "Use 8+" category, that is "use of university curriculum counselor service more than 8 times", in predicting both the low category (I) than the high category to (IV). Moreover, it endorses that the optimal subtable (I:II) vs (III: IV) (s = 2) is not characterized by category "Use 8+".

7.2 Empirical study 2—CO_2 light-duty vehicle evaluation relating to different type approval

The dataset comes from an experimental campaign carried out on urban roads of the Naples city areas (Meccariello and Della Ragione 2017) with a vehicle of the same segment for

Table 8 Rows significance and radii length of confidence circles

	Ray	$T^{(1.06)}_{WCCS}(i)$	d.f.	p value
Use 1	0.329	3.257	2.747	0.312
Use 2–4	0.266	3.551	2.747	0.276
Use 5–6	0.214	4.791	2.747	0.161
Use 7–8	0.230	2.458	2.747	0.463
Use 8+	0.445	12.723	2.747	0.004

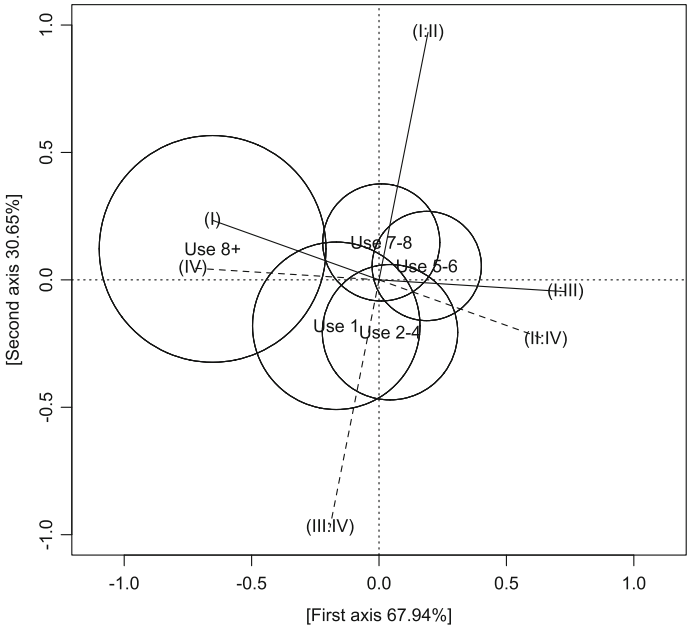

Fig. 4 Row isometric WCCA biplot ($\delta = 1$) with superimposed confidence circles

each Euro. The identified path crosses the Naples city center, greatly influenced by road traffic, whose characteristics are 22 km long and mainly flat. Vehicles have been equipped for on-road tests basically by an On-Board Diagnostics (OBD) tool to obtain engine operating parameters (velocity, revolutions per minute, temperature, car gear), a GPS tool to collect the geographic point and a PEMS Semtech-DS gas analyzer to acquire (CO), nitrogen oxides (NOx), carbon dioxide (CO_2) emissions at 1Hz. The Table 9 shows a contingency table between the CO_2 emission values and type approval directive.

The row variable, named Euro, is the X categorical variable (with $I = 3$) and it represents the evolution of the European Union emission regulations for the light duty vehicles. The type approval technology changed from 2004 to now, as technology increased, the CO_2 emission limits became lower. The variable Euro is on an ordinal scale with three categories: Euro 4 standards (2000/2005) with Directive 98/69/EC[1], Euro 5/6 standards (2009/2014) with regulation 715/2007[2]. The column variable, named CO_2 is the Y categorical variables (with $J = 7$) and it is split up in seven classes. The variable is on ordinal scale, with a range from

[1] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A31998L0069

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32007R0715

Table 9 Contingency table between the CO_2 emission values and type approval directive

	CO_2 level							
	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	Total
Euro 4	182	137	442	535	469	578	1077	3420
Euro 5	148	198	828	653	466	537	1223	4053
Euro 6	16	1247	655	305	195	278	630	3326
	346	1582	1925	1493	1130	1393	2930	10,799

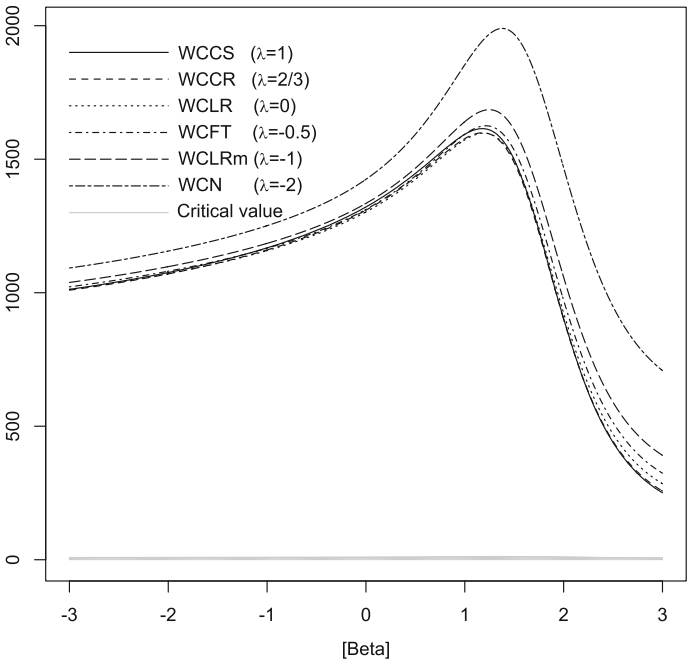

Fig. 5 $\tilde{T}^{(\lambda;\beta)}_{WCPD}$ statistic and critical value for $1-\alpha = 0.95$ in function of β

0 g/s (low emission) to 15g/s (high emission). The classes are divided as follows: $CO_2(I)$ values less than 0.4 g/s, $CO_2(II)$ values greater than 0.4 g/s to 0.6 g/s; $CO_2(III)$ values greater than 0.6 g/s to 0.8 g/s, $CO_2(IV)$ values greater than 0.8 g/s to 1 g/s, $CO_2(V)$ values greater than 1 g/s to 1.2 g/s, $CO_2(VI)$ values greater than 1.2 g/s to 1.5 g/s, $CO_2(VII)$ values greater than 1.5 g/s.

In Fig. 5 the value for statistic $T^{(\lambda;\beta)}_{WCPD}$ with $\lambda \in (-2; -1; -0.5; 0; 2/3; 1)$ and critical value with $1-\alpha = 0.95$ in function of β is shown. In particular, the statistic $T^{(\lambda;\beta)}_{WCPD}$ results significative for each value of β underlining a strong statistically significant association between the rows and column aggregated sub-tables. Table 10 explains the construction of the indexes $T^{(\lambda;\beta)}_{WCPD}$. In the construction of the index, for $\lambda = -2$, the first sub-table is the most important, with a weighting factor of about 23%. On the other hand, for values other than -2, the sub-table does not have a relevant weight in the construction of the indices.

Table 10 $T_{WCPD}^{(\lambda;\beta)}$ as Weighted sum of Power Divergence index of the sub-tables

s	$h_s^{(1.38)}\tilde{X}_s^2$ $\lambda=-2$	$h_s^{(1.25)}\tilde{G}_s^2$ $\lambda=-1$	$h_s^{(1.21)}T_s^2$ $\lambda=-0.5$	$h_s^{(1.19)}G_s^2$ $\lambda=0$	$h_s^{(1.18)}CR_s$ $\lambda=2/3$	$h_s^{(1.18)}X_s^2$ $\lambda=1$
1	2050.569	627.104	421.079	322.937	260.995	245.882
2	2729.642	1986.397	1824.688	1773.299	1813.753	1875.332
3	1931.219	1522.385	1417.687	1369.764	1356.711	1367.160
4	1149.485	913.368	848.471	813.841	793.042	790.467
5	664.425	524.239	484.892	462.898	447.638	444.194
6	357.801	272.002	247.943	234.047	223.492	220.511
$T_{WCPD}^{(\lambda;\beta)}$	8883.141	5845.495	5244.761	4976.785	4895.631	4943.546

Table 11 $T_{WCPD}^{(\lambda;\beta)}$ statistic significance

Index	λ	β	$T_{WCPD}^{(\lambda;\beta)}$	$r^{(\beta)}(I-1)$	$\tilde{T}_{WCPD}^{(\lambda;\beta)}$	$\eta^{(\beta)}$	d.f.	p-value
WCN	-2.00	1.38	8883.140	4.464	1989.753	2.088	5.748	< 0.001
WCLRm	-1.00	1.25	5845.495	3.468	1685.441	2.126	5.643	< 0.001
WCTF	-0.50	1.21	5244.761	3.227	1625.240	2.145	5.594	< 0.001
WCLR	0.00	1.19	4976.785	3.115	1597.438	2.156	5.566	< 0.001
WCCR	2/3	1.18	4895.631	3.062	1598.968	2.161	5.552	< 0.001
WCCS	1.00	1.18	4943.546	3.062	1614.617	2.161	5.552	< 0.001

Overall, in the construction of the index, the greatest weight is given by sub-table 2 as it assumes a high weight for each λ value.

Table 11 shows $T_{WCPD}^{(\lambda;\beta)}$ index and $\tilde{T}_{WCPD}^{(\lambda;\beta)}$ statistic based on β values that maximize the distance between $\tilde{T}_{WCPD}^{(\lambda;\beta)}$ statistic and $p$value with $1-\alpha = 0.95$, for $\lambda \in (-2; -1; 0.5; 0; 2/3; 1)$.

For $\lambda = 1$ the Fig. 6 presents the row statistic calculated by (6) and critical value calculated with $1-\alpha = 0.95$ in function of β. All categories are significant for the entire set of β values because they are above the critical region.

Respect to the correspondence analysis based on spectral decomposition of $T_{WCCS}^{(\beta)} = T_{WCPD}^{(1;\beta)}$, the choice of β falls in the value of 1.18. By partitioning the inertia using $T_{WCCS}^{(1.18)}$, we obtain the principal inertia values which are summarized in Table 12. From Table 12 it can be seen that the first axis explains 98% of the total variability.

Figure 7 illustrates the WCCA plot (dimensions 1–2) of the cumulated column categories (principal coordinates) and the overlapped independence circle. In Fig. 7 the total inertia is 100% given the number of categories of the Euro variable.

Figure 7 indicates that all contrasts based on the cumulative categories play an important role in interpreting the analysis because the independence region, with the radii length defined by (9), does not include the origin. Also it highlights a strong opposite congruence between the first and last aggregated sub-tables. In Table 13 we can observe that the first sub-table is opposed to all the others. In fact, the values of the latter are all negative. Instead, the other sub-tables are all strongly correlated with each other. In Fig. 8 a two-dimensional biplot of the asymmetric association between the row and the column categories of Table 9 is shown. This

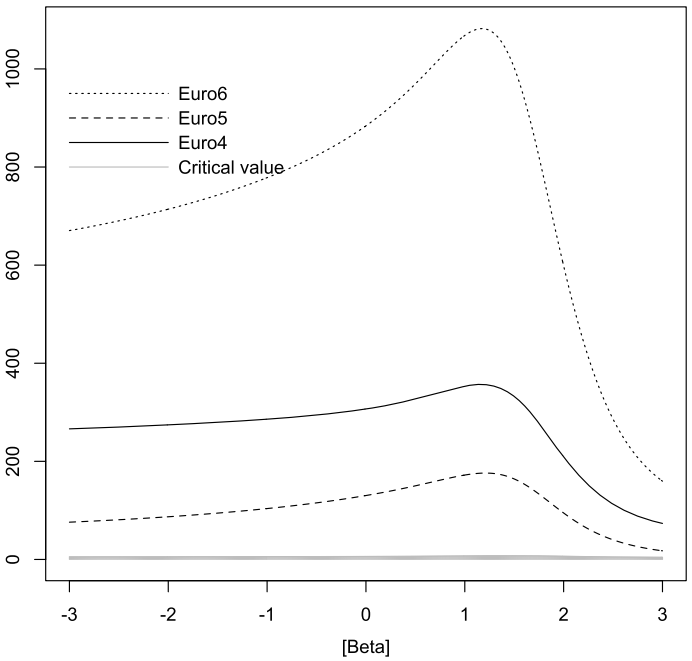

Fig. 6 $\tilde{T}^{(\beta)}_{WCCS}(i) \equiv \tilde{T}^{(1;\beta)}_{WCPD}(i)$ statistic and critical value for $1-\alpha = 0.95$ in function of β

Table 12 $T^{(1.18)}_{WCCS}$ inertia decomposition

Axis	Inertia	%	Cumulative %
1	0.45045	98.40006	98.40006
2	0.00732	1.59994	100.00000
$\frac{T^{(1.18)}_{WCCS}}{n}$	0.45778	100	

plot is a row isometric biplot in which the origin represents the column marginal distribution. Figure 8 evidences that the all categories Euro are significant in explaining the relationship because their circles don't include the origin.

Moreover, the Fig. 8 highlights the role played by the "Euro 6" respect to Euro 4 and Euro 5. In fact, they are on the opposite side. It has to point out that the technological improvement of the Euro 6 is much higher than that which took place between the improvement from Euro 4 to Euro 5. Compared to CO_2 emissions, the transition from Euro 4 to 5 was much lower than the transition to Euro 6. In fact, from Euro 4 to Euro 5, the greatest variations occurred in NOx and CO emissions. CO_2 for the Euro 4 and 5 technologies is about $202 g/km$ compared to the first version of the Euro 6 which is about $120 g/km$ (DIRECTIVE 98/69/EC of 1998, for Euro 4, and REGULATION (EC) No. 715/2007 of 2007, for Euro 5 and 6, of the European Parliament and of the Council). Furthermore, it must be taken into consideration that the emissions values used are derived from vehicle in real use, which normally produce different emission values with respect to the regulatory limits. The latter are very close to each other, and so the effect that we can see on pollutant emissions is less noticeable and weaker.

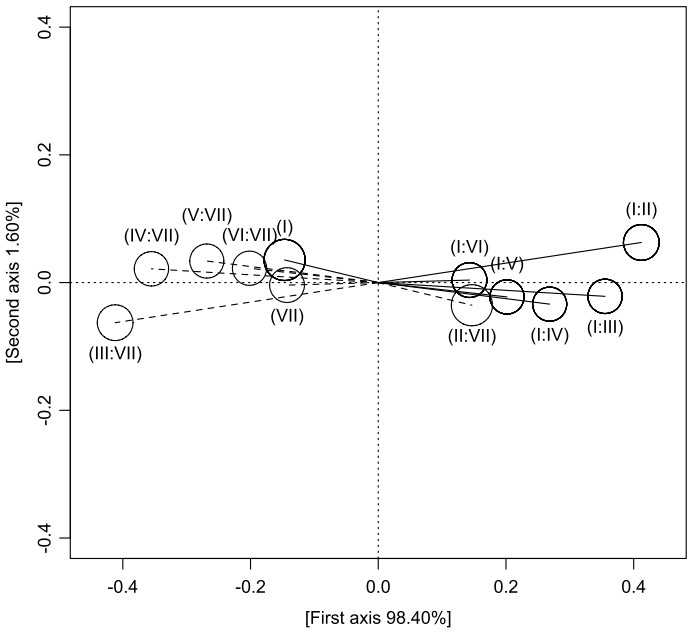

Fig. 7 WCCA plot: cumulated column categories and independence regions

Table 13 Similarity matrix

	(I)	(I:II)	(I:III)	(I:IV)	(I:V)	(I:VI)
(I)	1.000	− 0.926	− 0.984	− 0.994	− 0.992	− 0.965
(I:II)	− 0.926	1.000	0.978	0.962	0.966	0.992
(I:III)	− 0.984	0.978	1.000	0.998	0.999	0.996
(I:IV)	− 0.994	0.962	0.998	1.000	0.999	0.988
(I:V)	− 0.992	0.966	0.999	0.999	1.000	0.990
(I:VI)	− 0.965	0.992	0.996	0.988	0.990	1.000

8 Conclusion

The PD family proposed by Cressie and Read (1984, 1989) allows to link different indexes known in literature. However, it can be observed how these indexes behave badly for studying of association between categorical ordinal variables. To improve their behaviours, a new class of CCS-type based on Weighted Cumulate Chi-Squared (WCCS-type test) was defined. It was obtained with the introduction of a β parameter. Furthermore, for a particular value of the parameter, the Correspondence Analysis based on cumulative frequencies was performed and its properties were shown. Finally, the proposed methodology was used in two particular case studies concerning satisfaction with the university curricular consultancy service and

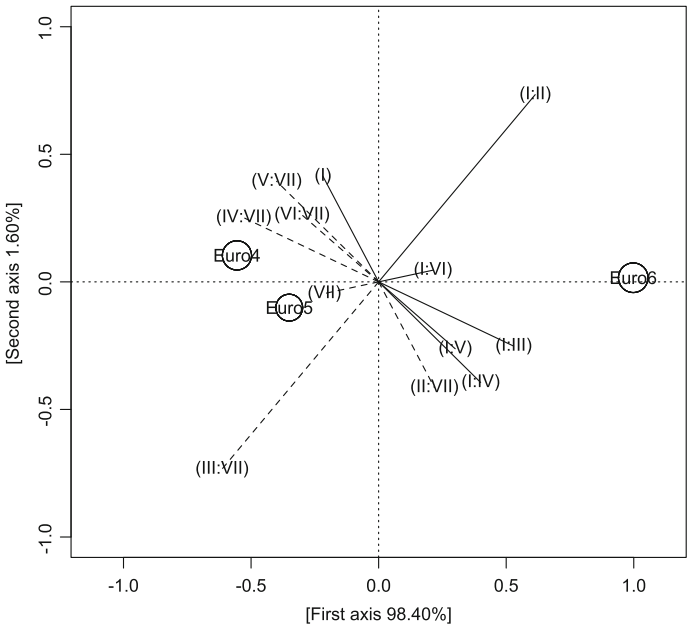

Fig. 8 Row isometric WCCA biplot ($\delta = 1$) with superimposed confidence circles

the assessment of CO_2 emissions produced by light-duty vehicles in relation to different type approvals.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Abdi, H. (2007). Rv coefficient and congruence coefficient. In *Encyclopedia of measurement and statistics* (pp. 849–853). Sage.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Wiley.

Barlow, R., Bartholomew, D., Bremner, J., & Brunk, H. (1972). *Statistical inference under order restrictions*. Wiley.

Beh, E. J. (2001). Confidence circles for correspondence analysis using orthogonal polynomials. *Journal of Applied Mathematics and Decision Sciences, 5,* 35–45.

Beh, E. J., & D'Ambra, L. (2010). Non-symmetrical correspondence analysis with concatenation and linear constraints. *Australian and New Zealand Journal of Statistic,52,* 27–44.

Beh, E. J., D'Ambra, L., & Simonetti, B. (2007). *Ordinal correspondence analysis based on cumulative Chisquared test, correspondence analysis and related methods*. Rotterdam: CARME.

Beh, E. J., D'Ambra, L., & Simonetti, B. (2011). Correspondence analysis of cumulative frequencies using a decomposition of Taguchi's Statistic. *Communication in Statistics - Theory Methods,40,* 1620–1632.

Camminatiello, I., D'Ambra, A., & D'Ambra, L. (2021). The association in two-way ordinal contingency tables through global odds ratios. *Metron*. https://doi.org/10.1007/s40300-021-00224-7.
Cressie, N., & Read, R. C. T. (1989). Pearson's X^2 and the loglikelihood ratio statistic G^2: A comparative review. *International Statistical Review,57,* 19–43.
Cressie, N., & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society - Series B, 46,* 440–464.
D'Ambra, A., & Amenta, P. (2011). Correspondence analysis with linear constraints of ordinal cross classifications. *Journal of Classification,28,* 70–92.
D'Ambra, A., Amenta, P., & Beh, E. J. (2021). Confidence regions and other tools for an extension of correspondence analysis based on cumulative frequencies. *AStA Advances in Statistical Analysis,105,* 405–429.
D'Ambra, L., Amenta, P., & D'Ambra, A. (2018). Decomposition of cumulative chi-squared statistics, with some new tools for their interpretation. *Statistical Methods & Applications,27,* 297–318.
D'Ambra, L., Beh, E. J., & Amenta, P. (2005). Catanova for two-way contingency tables with ordinal variables using orthogonal polynomials. *Communication in Statistics - Theory Methods,34,* 1755–1969.
D'Ambra, L., Beh, E. J., & Camminatiello, I. (2014). Cumulative correspondence analysis of two-way ordinal contingency tables. *Communication in Statistics - Theory Methods,43,* 1099–1113.
D'Ambra, L., Koskoy, O., & Simonetti, B. (2011). Cumulative correspondence analysis of ordered categorical data from industrial experiments. *Journal of Applied Statistics,36,* 1315–1328.
Emerson, P. L. (1968). Numerical construction of orthogonal polynomials from a general recurrence formula. *Biometrics, 24,* 696–701.
Fisher, R. A. (1940). The precision of discriminant functions. *Annals of Eugenics, 10,* 422–429.
Gabriel, K. R. (1971). The biplot graphic display with application to principal component analysis. *Biometrika, 58,* 453–467.
Hirotsu, C. (1986). Cumulative chi-squared statistic as a tool for testing goodness of fit. *Biometrika, 73,* 165–173.
Hirotsu, C. (1990). A critical look at accumulation analysis and related methods: Discussion. *Technometrics, 32,* 133–136.
Horst, P. (1935). Measuring complex attitudes. *The Journal of Social Psychology, 6,* 369–374.
Lebart, L., Warwick, K. M., & Morineau, A. (1984). *Multivariate descriptive statistical analysis*. Wiley.
Leti, G. (1983). *Statistica descrittiva*. Il Mulino
Lombardo, R., Beh, E. J., & D'Ambra, A. (2011). Studying the dependence between ordinal-nominal categorical variables via orthogonal polynomials. *Journal of Applied Statistics,38,* 2119–2132.
Lombardo, R., Carlier, A., & D'Ambra, L. (1996). Nonsymmetric correspondence analysis for three-way contingency tables. *Methodologica,4,* 59–80.
Mardia, K. V., Bibby, J. T., & Kent, J. M. (1982). *Multivariate analysis*. Academic Press.
Meccariello, G., & Della Ragione, L. (2017). *Statistical determination of local driving cycles based on experimental campaign as WLTC real approach*. SAE Technical Paper. https://doi.org/10.4271/2017-24-0138
Nair, V. N. (1986). Testing in industrial experiments with ordered categorical data. *Technometrics, 28,* 283–291.
Nair, V. N. (1987). Chi-squared type tests for ordered alternatives in contingency tables. *Journal of the American Statistical Association, 82,* 283–291.
Ringrose, T. J. (1992). Bootstrapping and correspondence analysis in archaeology. *Journal of Archaeological Science, 19,* 615–629.
Ringrose, T. J. (1996). Alternative confidence regions for canonical variate analysis. *Biometrika, 83,* 575–587.
Sarnacchiaro, P., & D'Ambra, A. (2007). Explorative data analysis and Catanova for ordinal variables: An integrated approach. *Journal of Applied Statistics, 34,* 1035–1050.
Sarnacchiaro, P., & D'Ambra, A. (2011). Cumulative correspondence analysis to improve the public train transport. *Electronic Journal of Applied Statistical Analysis, 2,* 15–24.
Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrika Bulletin, 2,* 110–114.
Taguchi, G. (1966). *Statistical analysis*. Maruzen.
Taguchi, G. (1974). A new statistical analysis for clinical data, the accumulating analysis, in contrast with the chi-square test. *Saishin Igaku, 29,* 806–813.
Takeuchi, K., & Hirotsu, C. (1982). The cumulative chi square method against ordered alternative in two-way contingency tables. *Report of Statistical Research, Japanese Union of Scientist and Engineers, 29,* 1–13.
Tucker, L. R. (1951). *A method for the synthesis of factor analysis studies*. (Personnel Research Section Report No. 984). Department of the Army, Washington, DC.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.