NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

***National Institute for Applied Statistics Research Australia***

**University of Wollongong, Australia**

**Working Paper**

03-24

# A General Similarity Measure for Simple Correspondence Analysis

Eric J. Beh and Rosaria Lombardo



National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522, Australia T: +61 2 42215076. E: karink@uow.edu.au

# A General Similarity Measure for Simple Correspondence Analysis

Eric J. Beh

*National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Australia*
*&*
*Centre for Multi-Dimensional Data Visualisation (MuViSU), Stellenbosch University, South Africa*

Rosaria Lombardo

*Department of Economics, University of Campania, Italy*

**Abstract**

This paper presents a general similarity measure for comparing different variants of simple correspondence analysis when analysing the association using the Cressie-Read family of divergence statistics (Beh and Lombardo, 2024, *International Statistical Review*). It includes, as special cases, the similarity measures that have been proposed in the correspondence analysis literature for assessing the similarities and differences between the traditional approach to simple correspondence analysis and new approaches like the log-ratio analysis, and the Hellinger distance method. This paper describes six further properties that show how the proposed general similarity measure can be expanded upon.

*Keywords:* Divergence residual, Cressie-Read family of divergence statistics, Log-ratio analysis, Hellinger distance.

*Email address:* ericb@uow.edu.au (Eric J. Beh)

*March 6, 2024*

## 1. Introduction

For about the last 30 years there has been considerable attention given in the (simple) correspondence analysis ((S)CA) literature to the power transformation of the cell frequencies (or some function of them) of a contingency table. Of particular interest is the general framework of Cuadras and Cuadras (2006) who called their method *parametric correspondence analysis*. This involves applying a power transformation, $\delta$, to the elements of the row and column profiles. They found the most suitable transformation was the square-root ($\delta = 1/2$) and called this variant of SCA the "Hellinger decomposition" (HD) method; this choice of $\delta$ was also advocated by Rao (1995) for the purposes of performing SCA. A similar approach to SCA, but one made independently of Cuadras and Cuadras (2006), was described in Beh et al. (2018), who showed that the Freeman-Tukey statistic (Freeman and Tukey, 1950) is the numerical foundations of HD while Beh et al. (2024) showed that these foundations can be obtained using a power transformation in reciprocal averaging. Another special case of *parametric* CA is when $\delta \to 0$ and is referred to as *log-ratio analysis* (LRA); see Cuadras et al. (2006) and Greenacre (2009, 2010) for details.

A broader framework for performing SCA where a power transformation is applied to the elements of the row and column profiles was proposed by Beh and Lombardo (2024) who showed the role played by the Cressie-Read family of divergence statistics (Cressie and Read, 1984; Read and Cressie, 1988). Beh and Lombardo (2024) showed that the traditional approach to SCA (which implies the partition of Pearson's chi-squared statistic) (Greenacre, 1984; Lebart et al., 1984; Beh, 2004; Beh and Lombardo, 2014, 2021) is a special case of their framework. They also showed that HD and LRA are also special cases where the Freeman-Tukey statistic and modified log-likelihood ratio statistic, respectively, serve as the numerical foundations.

Cuadras and Cuadras (2006) also showed that SCA can be compared with

any other variant from their parametric family by quantifying a *similarity measure*. A large similarity measure shows that SCA and a variant from their parametric SCA approach are very different while a zero measure shows that the two SCA variants produce identical results (ignoring the case where complete independence is observed between the variables of the contingency table).

This paper presents a more general similarity measure whose foundations lie with the SCA framework described by Beh and Lombardo (2024). While the SCA approach of Cuadras and Cuadras (2006) is confined to comparing SCA with any special case derived from their parametric approach to SCA, a benefit of the general similarity measure proposed in this paper is that any two variants can be compared. To discuss the proposed measure, this paper is divided into six further sections. Section 2 gives a brief overview of the SCA technique described in Beh and Lombardo (2024) while the two similarity measures described by Cuadras et al. (2006) are presented in Section 3; these measures are $\theta$ (for comparing SCA and HD) and $\phi$ (for comparing SCA and LRA). The measure that serves as the core of this paper is presented in Section 4 and we show that $\theta$ and $\phi$ are special cases. Section 5 gives six properties of the proposed measure and a demonstration of it is made in Section 6 where the smoking data of Greenacre (1984) is analysed. Cuadras et al. (2006) also study this data and so we empirically show the links between the proposed general similarity measure and their measures. Some final remarks will be left for Section 7.

## 2. On the Variants of Correspondence Analysis

### *2.1. Notation*

Consider an $I \times J$ contingency table, $\mathbf{N}$, where the $(i, j)$th (non-negative) cell count is denoted by $n_{ij}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Let the grand total of $\mathbf{N}$ be $n$ and define the $(i, j)$th cell proportion by $p_{ij} = n_{ij}/n$

so that $\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$. Define the $i$th row marginal proportion by $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$. Similarly, define the $j$th column marginal proportion by $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$.

*2.2. The Divergence Residual*

Before we define and describe the general similarity measure that is central to this paper – see Section 4.1 – we first provide a brief account of the SCA method (Beh and Lombardo, 2024) on which its foundations rest. This method considers the case where the elements of the centred row and column profiles are raised to a power $\delta$. Here, the $i$th centred row profile is

$$\left(\left(\frac{p_{i1}}{p_{i\bullet}}\right)^{\delta} - p_{\bullet j}^{\delta}, \left(\frac{p_{i2}}{p_{i\bullet}}\right)^{\delta} - p_{\bullet j}^{\delta}, \ldots, \left(\frac{p_{ij}}{p_{i\bullet}}\right)^{\delta} - p_{\bullet j}^{\delta}, \ldots, \left(\frac{p_{iJ}}{p_{i\bullet}}\right)^{\delta} - p_{\bullet j}^{\delta}\right)$$

and the $j$th centred column profile is

$$\left(\left(\frac{p_{1j}}{p_{\bullet j}}\right)^{\delta} - p_{i\bullet}^{\delta}, \left(\frac{p_{2j}}{p_{\bullet j}}\right)^{\delta} - p_{i\bullet}^{\delta}, \ldots, \left(\frac{p_{ij}}{p_{\bullet j}}\right)^{\delta} - p_{i\bullet}^{\delta}, \ldots, \left(\frac{p_{Ij}}{p_{\bullet j}}\right)^{\delta} - p_{i\bullet}^{\delta}\right).$$

See Beh et al. (2024) for further details. To compare these centred profiles in the context of SCA, the approach of Beh and Lombardo (2024) uses the Cressie-Read family of divergence statistics as the basis for assessing the association structure between the variables of $\mathbf{N}$. It involves defining

$$r_{ij}(\delta) = \frac{\sqrt{p_{i\bullet}p_{\bullet j}}}{\delta}\left(\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right)^{\delta} - 1\right) \tag{1}$$

to be the *divergence* residual of the $(i, j)$th cell of $\mathbf{N}$, for some value of $\delta \in (-\infty, \infty)$. Bishop et al. (1975, p.490) showed that $\sqrt{n}r_{ij}(\delta)$ is an asymptotic standard normally random variable, while its sum-of-square is

$$\mathrm{CR}(\delta) = n\sum_{i=1}^{I}\sum_{j=1}^{J} r_{ij}^{2}(\delta). \tag{2}$$

This is the second order Taylor series approximation of the family of Cressie-Read divergence statistics about $\left(p_{ij}/\left(p_{i\bullet}p_{\bullet j}\right)\right)^{\delta}=1$; see Cressie and Read (1984, Section 4.3) and Read and Cressie (1988, pp. 94–95). $\mathrm{CR}\left(\delta\right)$ can be used to assess the statistical significance of the association between variables of $\mathbf{N}$ by comparing it with the $100\left(1-\alpha\right)\%$ quantile of the chi-squared distribution with $\left(I-1\right)\left(J-1\right)$ degrees of freedom where $\alpha$ is the level of significance. When $\delta=0$, $1/2$ and $1$, (2) gives *exactly* the modified log-likelihood ratio statistic ($M^2$), the Freeman-Tukey statistic ($T^2$) and Pearson's chi-squared statistic ($X^2$), respectively, where

$$\begin{aligned} M^2 &= \mathrm{CR}\left(0\right)=2n\sum_{i=1}^{I}\sum_{j=1}^{J}p_{i\bullet}p_{\bullet j}\ln\left(\frac{p_{i\bullet}p_{\bullet j}}{p_{ij}}\right), \\ T^2 &= \mathrm{CR}\left(\frac{1}{2}\right)=4n\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\sqrt{p_{ij}}-\sqrt{p_{i\bullet}p_{\bullet j}}\right)^2, \\ X^2 &= \mathrm{CR}\left(1\right)=n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij}-p_{i\bullet}p_{\bullet j}\right)^2}{p_{i\bullet}p_{\bullet j}}. \end{aligned}$$

Beh and Lombardo (2024) show that these three measures serve as the numerical foundations of:

- LRA developed by Greenacre (2009, 2010) when $\delta=0$.
- HD developed by Cuadras and Cuadras (2006) when $\delta=1/2$. An equivalent SCA approach was proposed by Beh et al. (2018) who showed that $T^2$ is the measure of association used to define the total inertia.
- SCA when $\delta=1$.

Another special case of (2) is the second order approximation of the Cressie-Read statistic when $\delta=2/3$ (Cressie and Read, 1984, p. 463):

$$\mathrm{CR}^{*}=\mathrm{CR}\left(\frac{2}{3}\right)=\frac{3n}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}p_{i\bullet}p_{\bullet j}\left(\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right)^{2/3}-1\right)^2. \qquad (3)$$

Since the SCA technique of Beh and Lombardo (2024) is dependent on a given value of $\delta$, we refer to the analysis that is performed as SCA $(\delta)$. For example, SCA $(1)$ refers to the traditional approach to SCA while SCA $(0)$ and SCA $(1/2)$ refer to LRA and HD, respectively. However, we will use SCA, LRA and HD when referring to these variants.

## 3. A Review of Two Similarity Measures

### *3.1. Similarity Measure for SCA and HD*

Cuadras and Cuadras (2006, p.68) defined

$$\theta(\delta) = 1 - \sum_{i=1}^{I}\sum_{j=1}^{J} p_{i\bullet}^{1-\delta} p_{ij}^{\delta} p_{\bullet j}^{1-\delta}. \tag{4}$$

for some value of the "parameter" $\delta$. They defined this as their *measure of agreement* between SCA and any member of their parametric approach to CA. This measure can be alternatively expressed as

$$\begin{aligned} \theta(\delta) &= -\sum_{i=1}^{I}\sum_{j=1}^{J} p_{i\bullet}p_{\bullet j}\left(\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right)^{\delta} - 1\right) &(5)\\ &= -\delta\sum_{i=1}^{I}\sum_{j=1}^{J}\sqrt{p_{i\bullet}p_{\bullet j}}r_{ij}(\delta) &(6)\end{aligned}$$

with bounds (Cuadras and Cuadras, 2006, p. 68)

$$0 \leq \theta(\delta) \leq 1 - \frac{1}{\min(I,\, J)^{1-\delta}}. \tag{7}$$

Equations (5) and (6) show that $\theta(\delta)$ is the weighted sum of the divergence residual, (1) so that, under complete independence, $\theta(\delta) = 0$ for all values of $\delta$. Note that substituting $\delta = 1$ into (5) and (6) gives $\theta(1) = 0$. This make sense since this result means that there is no difference in the output from SCA when compared with itself. However, substituting $\delta = 0$ into (5), and

(6), means that $\theta(0) = 0$. This suggests there is no difference between SCA and LRA. However, this is not the case unless there is complete independence between the variables of $\mathbf{N}$. Section 4.3 provides an alternative expression for $\theta(0)$.

The goal of Cuadras and Cuadras (2006) for defining $\theta(\delta)$ was to compare SCA with HD, which can be made by substituting $\delta = 1/2$ into (5):

$$\theta \equiv \theta\left(\frac{1}{2}\right) = -\sum_{i=1}^{I}\sum_{j=1}^{J} p_{i\bullet}p_{\bullet j}\left(\sqrt{\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}} - 1\right).$$

Thus, ignoring the case of complete independence, $\theta = 0$ means that SCA and HD will produce equivalent numerical and visual summaries of the association. Although, it is important to note that this will only occur when the two configuration of points are located at the origin of the optimal correspondence plot which is where all points lie when there is complete independence in $\mathbf{N}$.

*3.2. Similarity Measure for SCA and LRA*

Cuadras et al. (2006) derived their similarity measure between SCA and LRA using two approximations. The first involves the first-order Taylor series approximation of the square root of $p_{ij}/(p_{i\bullet}p_{\bullet j})$:

$$\sqrt{\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}} - 1 \approx \frac{1}{2}\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}} - 1\right).$$

The second is the first order Taylor series approximation of the (natural) logarithm of this ratio so that, for $0 < p_{ij}/(p_{i\bullet}p_{\bullet j}) < 2$,

$$\ln\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right) \approx \frac{p_{ij}}{p_{i\bullet}p_{\bullet j}} - 1\,.$$

Therefore, Cuadras et al. (2006, p. 455) gave the measure

$$\phi = -\sum_{i=1}^{I}\sum_{j=1}^{J} p_{i\bullet}p_{\bullet j}\ln\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right) \tag{8}$$

as a first order approximation of $2\theta$. Cuadras et al. (2006, p. 454) discussed that this approximation works well when there is *almost* independence between the row and column variables of $\mathbf{N}$; in this case LRA, HD and SCA "may provided quite similar graphic displays". It should be noted that, for these "graphic displays", $\phi$ and $2\theta$ will be *almost* zero and the configuration of points in a LRA, HD and SCA of $\mathbf{N}$ will be "quite close" since they will be located very close to the origin. In general, the approximation $\phi \approx 2\theta$ is not guaranteed and worsens as the strength of the association in $\mathbf{N}$ increases.

Greenacre (2009) then shows that, when using the Box-Cox transformation

$$\lim_{\delta \to 0} \frac{1}{\delta}\left(\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right)^{\delta} - 1\right) = \ln\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right), \tag{9}$$

then $\phi$ serves as a similarity measure between SCA and LRA. Equation (9) plays a key part in the development of LRA, as it does for showing that the framework described by Beh and Lombardo (2024) leads to LRA when $\delta = 0$.

## 4. The General Similarity Measure

### *4.1. The Measure*

Suppose we wish to compare the results obtained from performing SCA $(\delta)$ and SCA $(\delta')$ on $\mathbf{N}$, for $\delta \geq \delta'$. Then we define our *general similarity measure* between these two SCA variants as

$$\varphi(\delta,\, \delta') = \sum_{i=1}^{I}\sum_{j=1}^{J} \sqrt{p_{i\bullet}p_{\bullet j}}\left(r_{ij}(\delta) - r_{ij}(\delta')\right) \tag{10}$$

so that $\varphi(\delta,\, \delta') \geq 0$ where $r_{ij}(\delta)$ is defined by (1). This measure is interpreted as follows: except in the case of complete independence, $\varphi(\delta,\, \delta') = 0$ implies that SCA $(\delta)$ and SCA $(\delta')$ yield equivalent numerical and visual summaries, otherwise $\varphi(\delta,\, \delta') > 0$. As SCA $(\delta)$ and SCA $(\delta')$ become more different, (10) increases.

*4.2. Link with $\theta(\delta)$, and between SCA and HD*

Equation (10) can be used to assess how SCA compares with any other SCA variant from the framework of Beh and Lombardo (2024) or the parametric CA approach of Cuadras et al. (2006). Therefore, $\theta(\delta)$ is linked to (10) as follows:

$$\begin{aligned}
\varphi(1,\,\delta) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sqrt{p_{i\bullet}p_{\bullet j}}\left(r_{ij}(1)-r_{ij}(\delta)\right) \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sqrt{p_{i\bullet}p_{\bullet j}}\left(\frac{p_{ij}-p_{i\bullet}p_{\bullet j}}{\sqrt{p_{i\bullet}p_{\bullet j}}}-\frac{\sqrt{p_{i\bullet}p_{\bullet j}}}{\delta}\left(\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right)^{\delta}-1\right)\right) \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\left((p_{ij}-p_{i\bullet}p_{\bullet j})-\frac{p_{i\bullet}p_{\bullet j}}{\delta}\left(\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right)^{\delta}-1\right)\right) \\
&= -\frac{1}{\delta}\sum_{i=1}^{I}\sum_{j=1}^{J}p_{i\bullet}p_{\bullet j}\left(\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right)^{\delta}-1\right) \\
&= \frac{\theta(\delta)}{\delta} \qquad (11)
\end{aligned}$$

where $\delta \neq 0$; we discuss what happens when $\delta = 0$ in Section 4.3. Thus, the similarity measure $\theta(\delta)$ is related to (10) by

$$\theta(\delta) = \delta\,\varphi(1,\,\delta)\ .$$

For example, the similarity measure between SCA and HD that was described by Cuadras et al. (2006) is a special case of (10) since

$$\theta \equiv \theta\left(\frac{1}{2}\right) = \frac{1}{2}\varphi\left(1,\,\frac{1}{2}\right)\ . \qquad (12)$$

*4.3. Link between SCA and LRA*

The similarity measure between SCA and LRA, $\phi$, defined by (8), is a special case of (10) when $\delta = 1$ and $\delta' = 0$ so that

$$\phi = \varphi(1,\,0)\ . \qquad (13)$$

This equivalency can be derived using (5), (11) and the Box-Cox transformation, (9), as follows

$$\begin{aligned}
\varphi(1,\,0) &= \lim_{\delta' \to 0} \varphi(1,\,\delta') \\
&= \lim_{\delta' \to 0} \frac{\theta(\delta')}{\delta'} \\
&= -\lim_{\delta' \to 0} \frac{1}{\delta'} \sum_{i=1}^{I} \sum_{j=1}^{J} p_{i\bullet} p_{\bullet j} \left( \left( \frac{p_{ij}}{p_{i\bullet} p_{\bullet j}} \right)^{\delta'} - 1 \right) \\
&= -\sum_{i=1}^{I} \sum_{j=1}^{J} p_{i\bullet} p_{\bullet j} \left\{ \lim_{\delta' \to 0} \frac{1}{\delta'} \left( \left( \frac{p_{ij}}{p_{i\bullet} p_{\bullet j}} \right)^{\delta'} - 1 \right) \right\} \\
&= -\sum_{i=1}^{I} \sum_{j=1}^{J} p_{i\bullet} p_{\bullet j} \ln \left( \frac{p_{ij}}{p_{i\bullet} p_{\bullet j}} \right) \\
&= \phi
\end{aligned}$$

as required.

Recall that in Section 3.1 we said that substituting $\delta = 0$ into (4) and (5) yields $\theta(0) = 0$ which implies that there is no difference between SCA and LRA, even when there is an association between the variables of $\mathbf{N}$. It may appear that this problem can be overcome by substituting $\delta = 0$ into (6). However, doing so leaves us with a $\theta(0)$ that does not exist since $r_{ij}(0)$ does not exist. Since $\phi$ reflects any difference between LRA and SCA, while $\varphi(1,\,0) = \phi$, we could have simply defined

$$\theta(0) = \phi\,. \tag{14}$$

Although, a stronger case for this equality comes from the second line of the derivation of (13):

$$\theta(0) = \lim_{\delta' \to 0} \frac{\theta'(\delta)}{\delta'} = \phi\,.$$

*4.4. Link between $\phi$ and $\theta(\delta)$*

Suppose we wish to compare LRA with SCA $(\delta)$ where $\delta > 0$. Then

$$\varphi(\delta, 0) = \phi - \frac{\theta(\delta)}{\delta} \tag{15}$$

so that $\phi > \varphi(\delta, 0)$ when $\delta > 0$. Equation (15) can be derived by starting with

$$\begin{aligned} \varphi(\delta, 0) &= \lim_{\delta' \to 0} \varphi(\delta, \delta') \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} \sqrt{p_{i\bullet}p_{\bullet j}} r_{ij}(\delta) - \sum_{i=1}^{I}\sum_{j=1}^{J} \sqrt{p_{i\bullet}p_{\bullet j}} \lim_{\delta' \to 0} r_{ij}(\delta') \end{aligned}$$

Using the Box-Cox transformation, (9), for the second term on the right-hand sides gives

$$\begin{aligned} \varphi(\delta, 0) &= \sum_{i=1}^{I}\sum_{j=1}^{J} \sqrt{p_{i\bullet}p_{\bullet j}} r_{ij}(\delta) - \sum_{i=1}^{I}\sum_{j=1}^{J} p_{i\bullet}p_{\bullet j} \ln\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right) \\ &= -\frac{1}{\delta}\left(-\sum_{i=1}^{I}\sum_{j=1}^{J} \sqrt{p_{i\bullet}p_{\bullet j}} r_{ij}(\delta)\right) + \left(-\sum_{i=1}^{I}\sum_{j=1}^{J} p_{i\bullet}p_{\bullet j} \ln\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right)\right) \\ &= -\frac{\theta(\delta)}{\delta} + \phi \end{aligned}$$

which completes the proof of (15). Using (14), (15) can be alternatively written as $\varphi(\delta, 0) = \theta(0) - \theta(\delta)/\delta$. So $\varphi(0, 0) = 0$ as expected when comparing LRA with itself.

*4.5. Link between $\phi$ and $2\theta$*

Consider again (15) which we now write as

$$\phi = \varphi(\delta, 0) + \frac{\theta(\delta)}{\delta}.$$

For example, when $\delta = 1$, $\varphi(1, 0) = \phi - \theta(1)/1 = \phi$ since $\theta(1) = 0$ and confirms (13). When $\delta = 1/2$ then

$$\begin{aligned} \phi &= \varphi\left(\frac{1}{2}, 0\right) + 2\theta\left(\frac{1}{2}\right) \\ &= 2\theta + \varphi\left(\frac{1}{2}, 0\right). \end{aligned} \tag{16}$$

Therefore, while Cuadras et al. (2006) show that $\phi \approx 2\theta$ when there is *almost* independence between the variables of $\mathbf{N}$, more generally $\phi > 2\theta$. The difference between the two measures is due to $\varphi(1/2, 0)$, the similarity measure between HD and LRA. Thus, any difference between LRA and SCA will always be bigger than any difference between LRA and HD. This also makes intuitive sense since a comparison of LRA with SCA requires using two values of $\delta$ (0 and 1, respectively) that are more different than when comparing LRA and HD (0 and 1/2, respectively).

The derivation of (16) can be verified using (10) as follows:

$$\begin{aligned} \phi &= \varphi(1, 0) \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} \sqrt{p_{i\bullet}p_{\bullet j}}\left(r_{ij}(1) - r_{ij}(0)\right) \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} \sqrt{p_{i\bullet}p_{\bullet j}}\left(r_{ij}(1) - r_{ij}\left(\frac{1}{2}\right) + r_{ij}\left(\frac{1}{2}\right) - r_{ij}(0)\right) \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} \sqrt{p_{i\bullet}p_{\bullet j}}\left(r_{ij}(1) - r_{ij}\left(\frac{1}{2}\right)\right) \\ &\qquad + \sum_{i=1}^{I}\sum_{j=1}^{J} \sqrt{p_{i\bullet}p_{\bullet j}}\left(r_{ij}\left(\frac{1}{2}\right) - r_{ij}(0)\right) \\ &= \varphi\left(1, \frac{1}{2}\right) + \varphi\left(\frac{1}{2}, 0\right) \\ &= 2\theta\left(\frac{1}{2}\right) + \varphi\left(\frac{1}{2}, 0\right) \end{aligned}$$

as required.

## 5. Six Further Properties

Now that we have described how the general similarity measure, (10), is linked to the similarity measures discussed by Cuadras and Cuadras (2006) and Cuadras et al. (2006), we turn our attention to describing six more properties of (10). Some of these properties may seem obvious but they expand upon the those given in Section 4. While $\delta \in (-\infty, \infty)$, when a comparison is made of SCA $(\delta)$ and SCA $(\delta')$, we will assume that $\delta > \delta' \geq 0$. Some of our discussion will consider the special case where $1 \geq \delta > \delta' \geq 0$ so that a demonstration of these properties can be made in terms of SCA, HD and LRA.

### *5.1. Property 1*

The first property we show is that the general similarity measure of SCA $(\delta)$ with itself is zero so that

$$\varphi(\delta, \delta) = 0\,.$$

The proof of this result is trivial but, using (10), is

$$\varphi(\delta, \delta) = \sum_{i=1}^{I}\sum_{j=1}^{J} \sqrt{p_{i\bullet}p_{\bullet j}}\,(r_{ij}(\delta) - r_{ij}(\delta)) = 0\,.$$

For example $\varphi(0, 0) = 0$ since there is no difference in comparing the results of LRA with itself. Similarly, $\varphi(1/2, 1/2) = 0$ and $\varphi(1, 1) = 0$.

### *5.2. Property 2*

Interchanging the order of $\delta$ and $\delta'$ leads to

$$\varphi(\delta, \delta') = -\varphi(\delta', \delta)\ .$$

This result can also be easily shown since

$$\begin{aligned}\varphi(\delta,\,\delta') &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sqrt{p_{i\bullet}p_{\bullet j}}\left(r_{ij}(\delta)-r_{ij}(\delta')\right)\\ &= -\sum_{i=1}^{I}\sum_{j=1}^{J}\sqrt{p_{i\bullet}p_{\bullet j}}\left(r_{ij}(\delta')-r_{ij}(\delta)\right)\\ &= -\varphi(\delta',\,\delta)\ .\end{aligned}$$

Therefore, since $\varphi(\delta,\,\delta') > 0$ when $\delta > \delta'$ then $\varphi(\delta',\,\delta)$ is negative. The sign does not change the interpretation of (10) but shows that the magnitude remains unchanged; a change in sign reflects a change in order of $\delta$ and $\delta'$.

*5.3. Property 3*

Equation (10) can be expressed in terms of the measure of agreement proposed by Cuadras and Cuadras (2006) so that

$$\varphi(\delta,\,\delta') = \frac{\theta(\delta')}{\delta'} - \frac{\theta(\delta)}{\delta}\,, \tag{17}$$

where $\delta > \delta'$. This can be proven as follows:

$$\begin{aligned}\varphi(\delta,\,\delta') &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sqrt{p_{i\bullet}p_{\bullet j}}r_{ij}(\delta) - \sum_{i=1}^{I}\sum_{j=1}^{J}\sqrt{p_{i\bullet}p_{\bullet j}}r_{ij}(\delta')\\ &= \left(-\sum_{i=1}^{I}\sum_{j=1}^{J}\sqrt{p_{i\bullet}p_{\bullet j}}r_{ij}(\delta')\right) - \left(-\sum_{i=1}^{I}\sum_{j=1}^{J}\sqrt{p_{i\bullet}p_{\bullet j}}r_{ij}(\delta)\right)\ .\end{aligned}$$

Substituting (6) in for the terms in parentheses on the right-hand side completes the proof. For example

$$\varphi(1,\,\delta') = \frac{\theta(\delta')}{\delta'} - \theta(1) = \frac{\theta(\delta')}{\delta'}$$

since $\theta(1) = 0$ and confirms (11). Also

$$\varphi(\delta,\,0) = \lim_{\delta' \to 0}\ \frac{\theta(\delta')}{\delta'} - \frac{\theta(\delta)}{\delta} = \phi - \frac{\theta(\delta)}{\delta}$$

thereby confirming (15). Thus, when $\delta = 1$, or when $\delta' \to 0$, then $\varphi(1,\, 0) = \phi$ confirming the derivation of (13).

Using (11) then (17) can alternatively, but equivalently, be expressed as

$$\varphi(\delta,\, \delta') = \varphi(1,\, \delta') - \varphi(1,\, \delta)$$

so that

$$\varphi(1,\, \delta') = \varphi(1,\, \delta) + \varphi(\delta,\, \delta') \tag{18}$$

for $1 > \delta > \delta'$. This partition shows that the general similarity measure between SCA and any other variant where $\delta' < 1$ can be partitioned into two terms, where each term is also a similarity measure between two different SCA variants derived using the Cressie-Read family of divergence statistics (Beh and Lombardo, 2024). Using (17) shows that the same partition can be achieved using the similarity measure discussed by Cuadras and Cuadras (2006). A more general partition of (10) is now discussed (in *Property 4*).

*5.4. Property 4*

Suppose that $\delta'' \geq \delta \geq \delta'$. Then (10) can be partitioned by

$$\varphi(\delta'',\, \delta') = \varphi(\delta'',\, \delta) + \varphi(\delta,\, \delta') \tag{19}$$

so that all three terms are non-negative. To derive this property note that (10) can be expressed by

$$\begin{aligned}
\varphi(\delta'',\, \delta') &= \sum_{i=1}^{I}\sum_{j=1}^{J} \sqrt{p_{i\bullet}p_{\bullet j}}\,(r_{ij}(\delta'') - r_{ij}(\delta')) \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J} \sqrt{p_{i\bullet}p_{\bullet j}}\,[(r_{ij}(\delta'') - r_{ij}(\delta)) + (r_{ij}(\delta) - r_{ij}(\delta'))] \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J} \sqrt{p_{i\bullet}p_{\bullet j}}\,(r_{ij}(\delta'') - r_{ij}(\delta))
\end{aligned}$$

$$
\begin{aligned}
&\quad + \sum_{i=1}^{I}\sum_{j=1}^{J} \sqrt{p_{i\bullet}p_{\bullet j}}\left(r_{ij}\left(\delta\right) - r_{ij}\left(\delta'\right)\right) \\
&= \quad \varphi\left(\delta'',\, \delta\right) + \varphi\left(\delta,\, \delta'\right)
\end{aligned}
$$

as required. While much of our discussion has focused $\delta = 0,\, 1/2$ and/or 1, (19) shows that (10) can be partitioned for any three values of $\delta \in (-\infty,\, \infty)$. A special case of this partition is (18) where $1 > \delta > \delta'$. By letting $\delta'' = 1$, $\delta = 1/2$ and $\delta' = 0$ then (19) simplies to

$$
\varphi\left(1,\, 0\right) = \varphi\left(1,\, \frac{1}{2}\right) + \varphi\left(\frac{1}{2},\, 0\right)
$$

which is equivalent to (16).

Suppose now we are interested in a comparison of LRA, SCA and SCA $(2/3)$; for the latter of these three SCA variants the total inertia is in terms $\mathrm{CR}^{*}$ defined by (3). Then $\phi$ can be partitioned so that

$$
\phi = \varphi\left(1,\, 0\right) = \varphi\left(1,\, \frac{2}{3}\right) + \varphi\left(\frac{2}{3},\, 0\right).
$$

The two terms on the right-hand side is the general similarity measure between SCA and SCA $(2/3)$, and the measure between SCA $(2/3)$ and LRA. One may also partition $2\theta$ so that

$$
2\theta \equiv \varphi\left(1,\, \frac{1}{2}\right) = \varphi\left(1,\, \frac{2}{3}\right) + \varphi\left(\frac{2}{3},\, \frac{1}{2}\right)
$$

where the first term on the right-hand side is the general similarity measure between SCA and SCA $(2/3)$ while the second term is just (10) for SCA $(2/3)$ and HD. Beh and Lombardo (2024) showed that, generally, there will be very little difference between the SCA variants when $\delta = 1/2$ and $\delta = 2/3$. Therefore it is to be expected that $\varphi\left(2/3,\, 1/2\right)$ will be much smaller than $\varphi\left(1,\, 2/3\right)$ highlighting that any differences that exist between HD and SCA $(2/3)$ will be much smaller than any difference that exists between SCA and SCA $(2/3)$.

*5.5. Property 5*

*Property 4* can be generalised so that $\varphi(\delta,\, \delta')$ is partitioned for $m$ values of $\delta$; $m$ could be three (like in *Property 4*) but it could also be that $m > 3$. This partition is now described.

Suppose that the largest (and first) value of $\delta$ is denoted by $\delta = \delta_{(1)}$ while the smallest (and $m$th) value is denoted by $\delta' = \delta_{(m)}$, respectively. Further, suppose that $\delta_{(k)} > \delta_{(k+1)}$ for $k = 1, \ldots, m-1$. Then, using *Property 4*,

$$\varphi\left(\delta_{(1)},\, \delta_{(m)}\right) \;=\; \varphi\left(\delta_{(1)},\, \delta_{(2)}\right) + \varphi\left(\delta_{(2)},\, \delta_{(m)}\right) .$$

We can also apply *Property 4* to $\varphi\left(\delta_{(2)},\, \delta_{(m)}\right)$ and all subsequent last terms of the partition so that

$$\begin{aligned}
\varphi\left(\delta_{(1)},\, \delta_{(m)}\right) &= \varphi\left(\delta_{(1)},\, \delta_{(2)}\right) + \varphi\left(\delta_{(2)},\, \delta_{(3)}\right) + \varphi\left(\delta_{(3)},\, \delta_{(m)}\right) \\
&= \ldots \\
&= \varphi\left(\delta_{(1)},\, \delta_{(2)}\right) + \varphi\left(\delta_{(2)},\, \delta_{(3)}\right) + \ldots \\
&\qquad\qquad + \varphi\left(\delta_{(m-1)},\, \delta_{(m)}\right)
\end{aligned}$$

Therefore,

$$\varphi\left(\delta_{(1)},\, \delta_{(m)}\right) = \sum_{k=1}^{m-1} \varphi\left(\delta_{(k)},\, \delta_{(k+1)}\right)$$

so that all terms on the right-hand side of this equation are non-negative. Such a partition allows for a comparison of multiple SCA $(\delta)$ variants to be simultaneously made.

*5.6. Property 6*

Recall that $\min\left[\varphi(\delta,\, \delta')\right] = 0$. To determine the maximum value that (10) can take, from (17) we have

$$\max\left[\varphi(\delta,\, \delta')\right] \;=\; \max\left(\frac{\theta(\delta')}{\delta'} - \frac{\theta(\delta)}{\delta}\right)$$

$$
\begin{aligned}
&= \max\left(\frac{\theta\left(\delta'\right)}{\delta'}\right) - \min\left(\frac{\theta\left(\delta\right)}{\delta}\right) \\
&= \frac{1}{\delta'}\max\left(\theta\left(\delta'\right)\right) - \frac{1}{\delta}\min\left(\theta\left(\delta\right)\right)
\end{aligned}
$$

Using (7) this simplifies to

$$
\max\left[\varphi\left(\delta,\, \delta'\right)\right] = \frac{1}{\delta'}\left(1 - \frac{1}{\min\left(I,\, J\right)^{1-\delta'}}\right).
$$

Note that when $\delta' = 0$,

$$
\max\left[\varphi\left(\delta,\, 0\right)\right] = \lim_{\delta' \to 0} \frac{1}{\delta'}\left(1 - \frac{1}{\min\left(I,\, J\right)^{1-\delta'}}\right) \to \infty\,.
$$

For example, $\max\left(\phi\right) = \max\left[\varphi\left(1,\, 0\right)\right]$ has no upper bound, just as Cuadras et al. (2006, p. 455) showed.

## 6. Example: Greenacre's Artificial Smoking Data

### *6.1. The Data*

Suppose we consider the same contingency table that was examined by Cuadras et al. (2006, Table 2). The data they examined is summarised in Table 1 and is artificial in nature; it was first introduced by Greenacre (1984, Table 3.1) to describe the key algebraic and visual features of SCA. Table 1 classifies 193 fictitious staff according to how often they smoke cigarettes (*Smoking*) and their position (*Position*) within a fictitious company.

### *6.2. A Preliminary Analysis*

A chi-squared test of independence shows that Table 1 has a Pearson chi-squared statistic of $X^2 = 16.442$ which has a p-value of 0.172. Therefore, there is no evidence of a statistically significant association existing between (*Smoking*) and (*Position*). The Freeman-Tukey statistic of Table 1 is $T^2 = 17.549$ (p-value = 0.130) while the modified chi-squared statistic is $M^2 =$

Table 1: Artificial contingency table of 193 workers classified according to smoking status and staff group

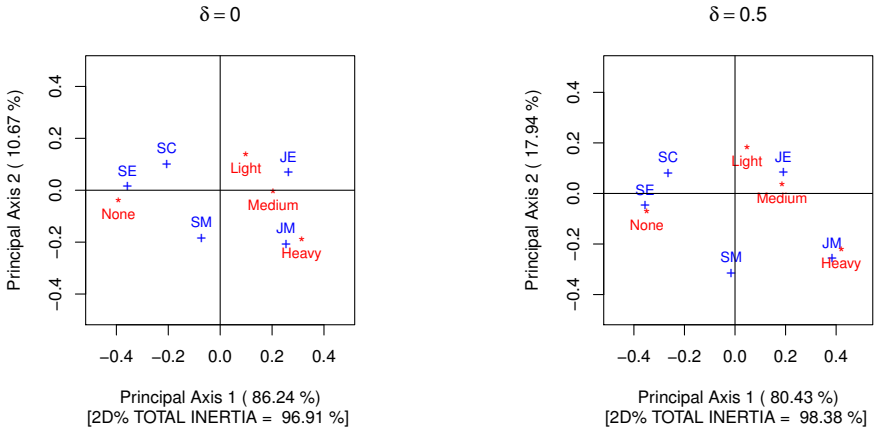
| | *Smoking* status | | | | |
|---|---|---|---|---|---|
| *Position* | None | Light | Medium | Heavy | Total |
| Senior Manager | 4 | 2 | 3 | 2 | 11 |
| Junior Manager | 4 | 3 | 7 | 4 | 18 |
| Senior Employer | 25 | 10 | 12 | 4 | 51 |
| Junior Employer | 18 | 24 | 33 | 13 | 88 |
| Secretary | 10 | 6 | 7 | 2 | 25 |
| Total | 61 | 45 | 62 | 25 | 193 |

17.259 (p-value = 0.140); both statistics help to confirm the conclusions reached from $X^2$. Cramér's V of 0.169, which is bounded by [0, 1], shows that there is indeed a very weak association between *Smoking* and *Position*. Despite the lack of a statistically significant association, we continue with our analysis of Table 1 to highlight the properties of the general similarity measure, (10), and show how they relate to the observations made by Cuadras et al. (2006).

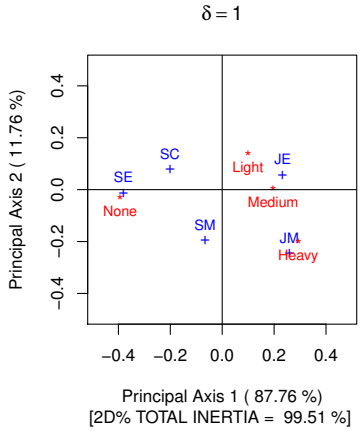
### *6.3. Comparison of LRA, HD and SCA*

Suppose that a LRA, HD and a SCA are performed on Table 1. The resulting two-dimensional correspondence plot from the LRA is given by Figure 1(a). Similarly, Figure 1(b) and Figure 1(c), respectively, is the two-dimensional correspondence plot from applying a HD and SCA to Table 1. All three plots have the same horizontal and vertical scales - ranging between -0.4 and 0.4 - and show a very similar configuration of points. All visualise at least 95% of the (lack of) association between the variables and are therefore excellent visual depictions of how the row and column categories of Table 1 are linked. The very similar configurations in all three plots suggests that the

general similarity measure, (10), for each pair of these three SCA variants should be very small. We now turn our attention to comparing the measures from these analyses.




(a) (b)




(c)

Figure 1: Correspondence plot of Table 1 using CR ($\delta$) (a) LRA ($\delta = 0$), (b) HD ($\delta = 1/2$), and (c) SCA ($\delta = 1$)

*6.4. Comparison of the Similarity Measures*

While (10) can be used to quantify how similar or different two SCA variants are in an optimal correspondence plot, the third and higher dimensions of such a space provide very little extra information (no more than 5%) on how the rows and columns of Table 1 are linked. Therefore, Figure 1 provides an excellent proxy for anything not visually depicted. Thus, (10) will reflect extremely well any similarities (or differences) that exist when comparing the three analyses.

The general similarity measure for the LRA-HD, HD-SCA, and LRA-SCA pairs are summarised in Table 2 to 8 decimal places (dp). Such a level of precision in these values is required since there exists very little (or, at the very least, no statistically significant) association in Table 1 and so the measures are all quite small.

Table 2: General similarity measures, (10), when comparing LRA, HD and SCA for Table 1

| Paired | General Similarity Measure | | | | |
|---|---|---|---|---|---|
| Comparison | $\delta$ | $\delta'$ | Measure | Value | MC P-value |
| HD & LRA | 1/2 | 0 | $\varphi\,(1/2, 0)$ | 0.02191635 | 0.343 |
| SCA & HD | 1 | 1/2 | $2\theta$ | 0.02132735 | 0.462 |
| SCA & LRA | 1 | 0 | $\phi$ | 0.04324370 | 0.378 |

Table 2 shows that the general similarity measure between SCA and LRA is

$$\phi = \varphi\,(1,\, 0) = 0.04324370$$

which is what Cuadras et al. (2006, p. 456) obtained, although they summarised their measure to 4dp. Cuadras et al. (2006, p.455) also showed that, for HD and SCA, $\theta \equiv \theta\,(1/2) = 0.0107$; this is indeed a very small value

since $\theta \in [0,\, 0.5]$ using (7). Using (10) we find that

$$2\theta = \varphi\left(1,\, \frac{1}{2}\right) = 0.02132735\,.$$

so that, using (12),

$$\theta = \frac{1}{2}\varphi\left(1,\, \frac{1}{2}\right) = \frac{1}{2} \times 0.02132735 = 0.01066367$$

as expected. Also, notice that $\phi > 2\theta$ as we showed would be the case in general in Section 4.5. These results show that the differences that exist between the SCA and LRA of Table 1 is twice any difference that exists between performing a SCA and HD. Thus, the approximation $\phi = 2\theta$ holds only if one considers that these measures are all very small; a feature that arises since the association between *Smoking* and *Position* is very weak and not statistically significant. However, a better approximation of $\phi$ for Table 1 is $\phi = 4\theta$; we see that $4\theta = 2 \times 0.02132735 = 0.0426547$ which differs from $\phi$ by only 0.000589, or 1.38% of $\phi$. Although we must keep in mind that $\phi \approx 4\theta$ is only for the data in Table 1 and is not guaranteed for all two-way contingency tables.

*6.5. Further Visual Depictions*

While our discussion of the choice of $\delta$ has focused on 0, 1/2 and 1 so that comparisons can be directly made of LRA, HD and SCA, respectively, recall that $\delta$ can take on any value in the interval $(-\infty,\, \infty)$. This is because the Cressie-Read family of divergence statistics permits such values of $\delta$ to be selected. Beh and Lombardo (2024, Section 8) discussed a range of criteria on which $\delta$ can be chosen. Options they discussed include selecting $\delta$ so that one avoids the possibility of obtaining negative expected cell frequencies when reconstituting them based on the information contained in a one or two-dimensional correspondence plot. One may also choose the value of $\delta$ that avoids the presence of overdispersion, a feature of contingency tables

discussed by Haberman (1973) and Agresti (2013, p. 80). While there may be many criteria on which to select $\delta$, Beh and Lombardo (2024) recommended that $\delta \in [0, 3/2]$ or, if one is interested in identifying potential outlying categories then a large value of $\delta$, such as $\delta = 3$, might be chosen. In general though, they showed that $\delta = 1/2$ is an excellent choice since it satisfies the key criteria they described, with $\delta = 2/3$ being an another appropriate choice. Another criteria on which Beh and Lombardo (2024) recommended that $\delta$ be chosen is the value that maximises the amount of association that can be depicted using only the first two dimensions. This is very much in line with the recommendation made by Cuadras and Cuadras (2006, p. 72) who said

> . . . we propose the choice of $[\delta]$ such that it accounts for the maximum percentage of inertia in a [SCA $(\delta)$] plot in low dimension, generally two dimensions. The best value of $[\delta]$ may provide a relatively high percentage.

Therefore, we shall continue our comparison of the various SCA $(\delta)$ variants and the general similarity measures that can be obtained by expanding our selection of $\delta$. So, in addition to $\delta = 0$, $1/2$ and 1 we shall also consider the $\delta$ values of $2/3$ and 3. We also consider the value of $\delta$ where the percentage of the total inertia captured in the first two dimensions is maximised; in this case $\delta = 1.15$ so that 99.576% of the total inertia is depicted using the first two dimensions. Since selecting $\delta$ based on the amount of association that can be depicted using the first two dimensions is just one possible criterion for choosing $\delta$, further investigation into the impact this choice has on the distribution of the divergence residuals will be the subject of new research to be undertaken.

Thus, to gain an understanding of the impact that changing $\delta \in [0, 3]$ has on the two-dimensional correspondence plot of Table 1 see Figure 2. It

visualises the percentage contribution of the first dimension (the dashed line), the second dimension (the dotted line) and the combined inertia of these two dimensions (the solid line) to the total inertia. The vertical grey line in Figure 2 corresponds to $\delta = 1.15$. While it is not shown, the percentage of the total inertia depicted using the first two dimensions approaches 100% as $\delta \to \infty$.

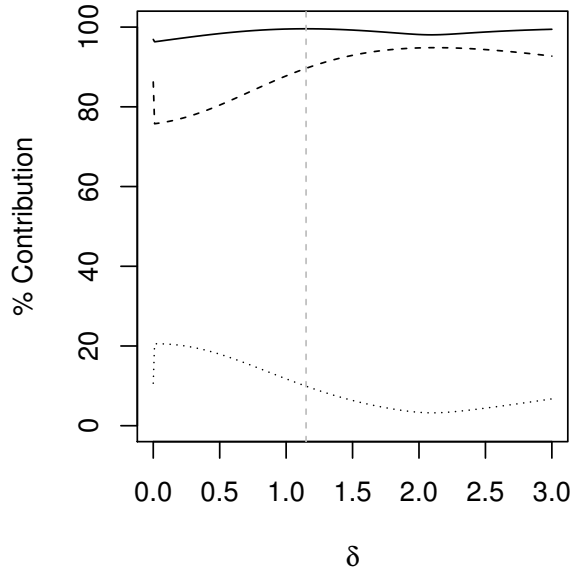
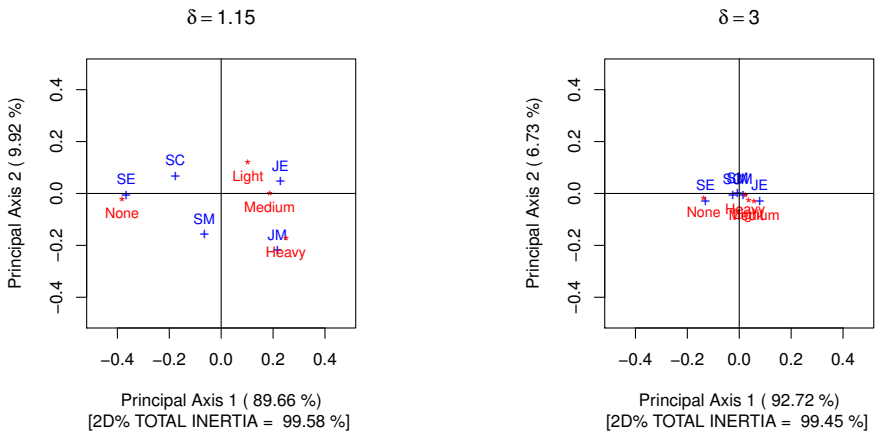



Figure 2: Solid line is the percentage of the contribution to the total inertia of a two-dimensional correspondence plot for $\delta \in [0, 3]$. The dashed line refers to the first dimension and the dotted line refers to the second dimension. The vertical dashed line is at $\delta = 1.15$ which gives the optimal two-dimensional correspondence plot.

The correspondence plots obtained by performing SCA (1.15) and SCA (3) are given by Figure 3(a) and Figure 3(b) respectively. We have ensured that the scale of the vertical and horizontal axes are identical to those in Figure 1 so that a direct comparison can be made of the configurations for different values of $\delta$. Note that Figure 3(b) is designed to identify any potential outlying categories and shows that there are none. Comparing the correspondence

plot of Figure 3(a) with those of Figure 1 shows a fairly similar configuration and so we would expect that calculating (10) with $\delta = 1.15$ and each of $\delta = 0$, $1/2$ and 1 to remain quite small. However, Figure 3(b) shows a correspondence plot that is very different when compared with those of Figure 1 and so we would expect $\varphi(3, \delta)$ for $0 \leq \delta < 1$ to be relatively large. We now turn our attention to calculating these measures.




(a) (b)

Figure 3: Correspondence plot of Table 1 using CR $(\delta)$ where (a) $\delta = 1.15$, and (b) $\delta = 3$

### *6.6. Revisiting the Comparison of the Similarity Measures*

Table 3 summarises the general similarity measures when $\delta = 0$, $1/2$, $2/3$, 1, 1.15 and 3. Comparing each of the general similarity measures shows that the largest difference in SCA variants is between LRA and SCA $(3)$; here $\varphi(0, 3) = 0.131$ to 3dp. This result should not be surprising since we are comparing analyses generated using two very different values of $\delta$. The least different two SCA variants is when a comparison is made of SCA and SCA $(1.15)$ so that $\varphi(1, 1.15) = 0.006$ to 3dp. Therefore, of the values of $\delta$ considered here, the traditional approach to SCA gives almost the opti-

Table 3: General similarity measures when comparing SCA ($\delta$) variants for Table 1; $\delta = 0$, 1/2, 2/3, 1, 1.15 and 3

| Paired | Summaries | | | |
|---|---|---|---|---|
| Comparison | $\delta$ | $\delta'$ | $\varphi(\delta, \delta')$ | MC P-value |
| HD & LRA | 1/2 | 0 | 0.02191635 | 0.343 |
| SCA (2/3) & HD | 2/3 | 1/2 | 0.00715332 | 0.418 |
| SCA & SCA (2/3) | 1 | 2/3 | 0.01417403 | 0.277 |
| SCA (1.15) & SCA | 1.15 | 1 | 0.00634730 | 0.212 |
| SCA (3) & SCA (1.15) | 3 | 1.15 | 0.08169089 | 0.177 |
| SCA (3) & LRA | 3 | 0 | 0.13128189 | 0.268 |

mal two-dimensional correspondence plot. Figure 4 provides a simple visual summary of the $\varphi(\delta, \delta')$ values that are summarised in Table 3.

While the Table 3 does not summarise the general similarity measure between LRA and SCA it can be easily determined using *Property 4* since

$$\begin{aligned}
\phi &= \varphi(1, 0) \\
&= \varphi\left(1, \frac{2}{3}\right) + \varphi\left(\frac{2}{3}, \frac{1}{2}\right) + \varphi\left(\frac{1}{2}, 0\right) \\
&= 0.014174028 + 0.007153319 + 0.021916352 \\
&= 0.04324370
\end{aligned}$$

which is exactly the measure given in Table 2. Of these three measures, there is little difference between HD and SCA (2/3) since $\varphi(2/3, 1/2) = 0.007$ (to 3dp). This result is also not surprising since the two $\delta$ values are relatively close to each other; see also Beh and Lombardo (2024). In fact, there is about twice as much of a difference between SCA and SCA (2/3) than there is between HD and SCA (2/3) since $\varphi(1, 2/3) = 0.014$. Similarly, there is about three times more of a difference between HD and LRA than there is between HD and SCA (2/3); note that for such a comparison $\varphi(1/2, 0) = 0.022$.

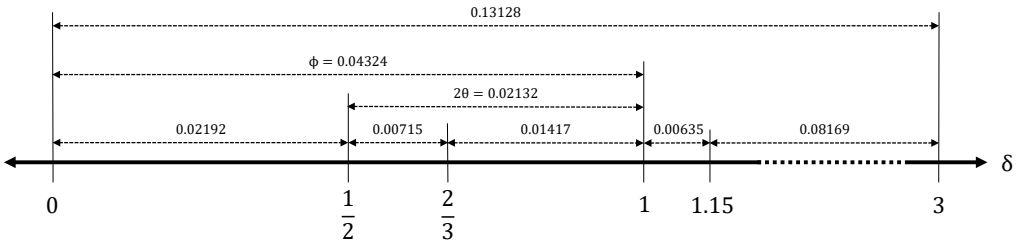



Figure 4: A visual depiction of the links between the $\varphi(\delta,\, \delta')$ values summarised in Table 3 for $\delta = 0,\, 1/2,\, 2/3,\, 1,\, 1.15$ and 3

One can also determine $2\theta$ using the summaries in Table 3 since

$$\begin{aligned} 2\theta &= \varphi\left(1,\, \frac{1}{2}\right) \\ &= \varphi\left(1,\, \frac{2}{3}\right) + \varphi\left(\frac{2}{3},\, \frac{1}{2}\right) \\ &= 0.014174028 + 0.007153319 \\ &= 0.02132735 \end{aligned}$$

which is identical to the value of $2\theta$ given in Table 2.

*6.7. A Simulation Study*

To assess whether there is a statistically significant difference between the various pairs of SCA $(\delta)$ variants we assume that the marginal totals of Table 1 are fixed and known. We can then determine the Monte Carlo p-value for each pair by simulating, using the `r2dtable()` function in `R`, 1000 contingency tables with the same marginal totals as Table 1. For $\delta = 0,\, 1/2$ and 1 these Monte Carlo p-values are summarised in the fifth column of Table 2. It shows that the p-value between LRA and HD is 0.343 while the p-value between SCA and HD is 0.462. Therefore, there is no evidence that either pair of SCA variants are statistically significantly different from each

other. The same conclusion can also be made when comparing SCA and LRA, which produces a Monte Carlo p-value of 0.378. These p-values have also been calculated for the pairs of SCA variants summarised in Table 3 and appear in its fifth column; note that there is no evidence that any of six pairs of SCA ($\delta$) we studied are (statistically) significantly different from each other. This is not surprising since there does not exist a statistically significant association between the variables of Table 1.

We could have provided a more detailed examination of the simulation results by comparing and contrasting the empirical distribution and numerical summaries for each $\varphi(\delta, \delta')$. However, for simplicity we have not done so but this will be the subject for further investigation at a later date.

## 7. Discussion

This paper has presented a general similarity measure, (10), that allows for a comparison to be made of any two SCA variants that are special cases of the technique outlined in Beh and Lombardo (2024). Equivalently, using (17), this similarity measure can also be used to compare these two methods using the parametric CA approach described by Cuadras and Cuadras (2006), despite the parametric approach being designed to compare SCA with one other variant of SCA. We have also shown that (10) can be partitioned into any number of general similarity measures that provide a simultaneous comparison of not just two or three SCA variants, but any number of variants.

There are other areas of further research that can be undertaken to show how (10) can be expanded. Cuadras and Cuadras (2006) briefly considered an adaptation of their $\theta(\delta)$ that allowed for a similarity measure to be quantified for each dimension of the correspondence plot. This is certainly an area that investigated for (10). This suggests that a similarity measure for each row and column can also be quantified. This is advantageous since understanding

the shifts in position of each point in the correspondence plot is important for understanding where there are similarities and where there are differences in the correspondence plots.

While we have confined ourselves to the analysis of a two-way contingency table, just as Cuadras and Cuadras (2006) and Cuadras et al. (2006) did, there is room to expand (10) for the CA of a multi-way contingency table. To show how this could be achieved suppose we have, for the sake of simplicity, a three-way contingency table, $\underline{\mathbf{N}}$ consisting of $I$ rows, $J$ columns and $K$ tubes. If the $i$th row, $j$th column and $k$th tube marginal proportion is defined as $p_{i\bullet\bullet}$, $p_{\bullet j\bullet}$ and $p_{\bullet\bullet k}$, respectively, then (10) can be extended for $\underline{\mathbf{N}}$ such that

$$\varphi(\delta,\, \delta') = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sqrt{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}}\left(r_{ijk}(\delta) - r_{ijk}(\delta')\right) \tag{20}$$

where $p_{ijk}$ is the $(i,\, j,\, k)$th proportion and, for a given $\delta \in (-\infty,\, \infty)$,

$$r_{ijk}(\delta) = \frac{\sqrt{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}}}{\delta}\left(\left(\frac{p_{ijk}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^{\delta} - 1\right).$$

The similarity measure (20) is applicable when comparing the configuration of points from a three-way correspondence analysis (Carlier and Kroonenberg, 1996; Lombardo et al., 2021, 2023) of the three-way contingency table and can be expanded for any multi-way contingency table.

Another adaptation of (10) that can be made is when studying stratified data or time-dependent data where the same two (or more) categorical variables are studied at each strata or time period. Suppose we define $r_{ij(s)}(\delta)$ to be the divergence residual for the $s$th strata/time period, for $s = 1,\, 2$. Then, for a given value of $\delta$ we can define the similarity measure

$$\varphi(\delta) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sqrt{p_{i\bullet}p_{\bullet j}}\left(r_{ij(1)}(\delta) - r_{ij(2)}(\delta)\right).$$

The difference between this measure and (10) is that rather than considering two different SCA variants, here we are comparing the same variant across the two strata/time periods.

These and other areas of study that expand our understanding of (10) will be left for future consideration.

## References


Agresti, A., 2013. Categorical Data Analysis (3rd ed). Wiley, New York.

Beh, E., Lombardo, R., Wang, T.W., 2024. Power transformations and reciprocal averaging, in: Beh, E.J., Lombardo, R., Clavel, J. (Eds.), Analysis of Categorical Data from Historical Perspectives: Essays in Honour of Shizuhiko Nishisato. Springer, Singapore, pp. 173 – 199.

Beh, E.J., 2004. Simple correspondence analysis: A bibliographic review. International Statistical Review 72, 257 – 284.

Beh, E.J., Lombardo, R., 2014. Correspondence Analysis: Theory, Practice and New Strategies. Wiley, Chichester.

Beh, E.J., Lombardo, R., 2021. An Introduction to Correspondence Analysis. Wiley, Chichester.

Beh, E.J., Lombardo, R., 2024. Correspondence analysis and the Cressie-Read family of divergence statistics. International Statistical Review , (in press).

Beh, E.J., Lombardo, R., Alberti, G., 2018. Correspondence analysis and the freeman-tukey statistic: A study of archaeological data. Computational Statistics and Data Analysis 128, 73 – 86.

Bishop, Y.M., Fienberg, S.E., Holland, P.W., 1975. Discrete Multivariate Analysis: Theory and Practice. MIT Press.

Carlier, A., Kroonenberg, P.M., 1996. Decompositions and biplots in three-way correspondence analysis. Psychometrika 61, 355 – 373.

Cressie, N.A.C., Read, T.R.C., 1984. Multinomial goodness-of-fit tests. Journal of the Royal Statistical Society, Series B 46, 440 – 464.

Cuadras, C.M., Cuadras, D., 2006. A parametric approach to correspondence analysis. Linear Algebra and its Applications 417, 64 – 74.

Cuadras, C.M., Cuadras, D., Greenacre, M., 2006. A comparison of different methods for representing categorical data. Communications in Statistics - Simulation and Computation 35, 447 – 459.

Freeman, M.F., Tukey, J.W., 1950. Transformations related to the angular and square root. The Annals of Mathematical Statistics 21, 607 – 611.

Greenacre, M., 2009. Power transformations in correspondence analysis. Computational Statistics and Data Analysis 53, 3107 – 3116.

Greenacre, M., 2010. Log-ratio analysis is a limiting case of correspondence analysis. Mathematical Geosciences 42, 129 – 134.

Greenacre, M.J., 1984. Theory and Application of Correspondence Analysis. Academic Press, London.

Haberman, S., 1973. The analysis of residuals in cross-classified tables. Biometrics 75, 457 – 467.

Lebart, L., Morineau, A., Warwick, K.W., 1984. Multivariate Descriptive Statistical Analysis. Wiley, New York, USA.

Lombardo, R., Beh, E.J., Kroonenberg, P.M., 2021. Symmetrical and non-symmetrical variants of three-way correspondence analysis for ordered variables. Statistical Science 36, 542 – 561.

Lombardo, R., van de Velden, M., Beh, E.J., 2023. Three-way correspondence analysis in R. The R Journal 15/2, 237 – 262.

Rao, C.R., 1995. A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. Qüestiió 19, 23 – 63.

Read, T.R.C., Cressie, N.A.C., 1988. Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer-Verlag.