NOVA

University of Newcastle Research Online

nova.newcastle.edu.au

Beh, Eric J.; D'Ambra, Luigi; Simonetti, Biagio "Correspondence analysis of cumulative frequencies using a decomposition of Taguchi's statistic" Communications in Statistics: Theory and Methods Vol. 40, Issue 9, p. 1620-1632 (2011)

Available from: <http://dx.doi.org/10.1080/03610921003615880>

This is an Accepted Manuscript of an article published in Communications in Statistics: Theory and Methods on 16/03/2011, available online:

<http://www.tandfonline.com/10.1080/03610921003615880>

Accessed from: <http://hdl.handle.net/1959.13/1053783>

Correspondence Analysis of Cumulative Frequencies using a Decomposition of Taguchi's Statistic

Eric J. Beh

School of Mathematical and Physical Sciences,
University of Newcastle, Callaghan, Australia

Luigi D'Ambra

Dipartimento di Matematica e Statistica,
University "Federico II" of Naples, Napoli, Italy

Biagio Simonetti

Dipartimento di Analisi dei Sistemi Economici e Sociali,
University of Sannio, Benevento, Italy

Abstract

Taguchi's statistic has long been known to be a more appropriate measure of association for ordinal variables than the Pearson chi-squared statistic. Therefore, there is some advantage in using Taguchi's statistic for performing correspondence analysis when a two-way contingency table consists of one ordinal categorical variable. This paper will explore the development of correspondence analysis using a decomposition of Taguchi's statistic.

Keywords: Cumulative Chi-Squared Statistic; Correspondence Analysis; Non-Symmetrical Correspondence Analysis; Taguchi's Statistic.

1 Introduction

Correspondence analysis (CA) of a two-way contingency table is a popular statistical tool for graphically identifying the nature of the association between the categorical variables. It has been used by data analysts from a variety of disciplines over the past 50 years, however little attention has been paid to the

case where the variables are ordinal. Although, there have been some contributions that deal with ordinal categorical variables, including those of Parsa and Smith (1993), Ritov and Gilula (1993), Schriever (1983) and Beh (1997). Generally these procedures involve constraining the output obtained from applying singular value decomposition (SVD) so that the coordinates in the first dimension have an ordered structure. An alternative approach involves using moment decomposition (Beh, 1997) or hybrid decomposition (Beh, 2004).

When the Pearson chi-squared statistic is used to study the association between categorical variables, it can perform poorly when dealing with ordinal structures - see Agresti (2007, Section 2.5) and Barlow, Bartholomew, Bremner and Brunk, (1972) for a discussion of this issue and some possible remedies. To address this problem, Taguchi's statistic (Taguchi, 1966, 1974) may be considered. Taguchi's statistic takes into account the presence of an ordinal categorical variable by considering the cumulative frequency of cells in the contingency table. There are two advantages that make Taguchi's statistic a more appropriate measure of association than the more commonly used Pearson chi-squared statistic. Firstly, the Pearson statistic does not adequately take into account the ordinal structure of a variable (Nair, 1987) while Taguchi's statistic does by considering cumulative cell frequencies. Secondly, Taguchi's statistic is more suitable in studies (such as clinical trials) where the number of categories within a variable is equal to, or larger than, 5 (Hirotsu, 1990). Therefore one may consider in these cases the CA of cumulative frequencies using Taguchi's statistic. One may also consider Cuadras (2002) who proposes a way of performing CA using cumulative cell frequencies.

This paper explores the application of Taguchi's statistic for performing CA. Section 2 defines Taguchi's statistic for a two-way contingency table consisting of nominal row categories and ordered column categories. A link between it and the Goodman-Kruskal tau index (Goodman and Kruskal, 1954) is also established. An example of the relationship between Taguchi's statistic and the Goodman-Kruskal tau statistic is also provided in this section. Section 3 considers the CA of a two-way contingency table using Taguchi's statistic. In particular we will consider the partition of the statistic using the SVD, and further establish the relationship between Taguchi's statistic and the Goodman-Kruskal tau index. A discussion of the properties of CA using Taguchi's statistic is given in Section 4 where attention is given to the coordinates for obtaining a graphical summary of the association, the reconstitution formula and distance measures. Section 5 demonstrates the application of this approach and a comparison is made with results obtained using non-symmetrical CA (NSCA). Some final remarks are left for Section 6.

2 Association Measures and Cumulative Frequencies

2.1 Taguchi's Statistic

To determine the nature of the association between the variables of a contingency table, tests involving the Pearson chi-squared statistic are generally considered. However, the statistic does not take into account the structure of ordered categorical variables (Agresti, 2007, page 41). To overcome this problem, Taguchi (1966, 1974) developed a simple statistic that does take into consideration the structure of an ordered categorical variable. It does so by considering the cumulative frequency of the cells of the contingency table across the ordered variable. Thus, the focus of this paper is to consider the decomposition of Taguchi's statistic for the purposes of performing correspondence analysis. We will focus our attention on a two-way contingency table where one of the variables consists of ordered responses.

Consider a two-way contingency table, $\mathbf{N}$, that cross-classifies n individuals/units according to I row categories and K ordered (ascending) column categories. Denote n_{ik} as the (i, k) th joint cell frequency with a relative frequency of $p_{ik} = n_{ik}/n$ which is the (i, k) th element of the matrix $\mathbf{P}$. Let $n_{i\bullet}$ and $n_{\bullet k}$ be the i th row and k th column marginal frequencies, respectively. Also, let $p_{i\bullet} = n_{i\bullet}/n$ be the (i, i) th element of the diagonal matrix $\mathbf{D}_I$ and $p_{\bullet k} = n_{\bullet k}/n$ be the k th element of the vector $\mathbf{c} = (p_{\bullet 1}, p_{\bullet 2}, \dots, p_{\bullet K})^T$. Denote $Z_{ik} = \sum_{j=1}^k n_{ij}$ so that $n_{i1} = Z_{i1} \leq Z_{i2} \leq \dots \leq Z_{iK} = n_{i\bullet}$. Here Z_{ik} is the cumulative frequency of the i th row category up to the k th column category and their consideration provides a way of ensuring that the ordinal structure of the column categories is preserved. Similarly, denote $d_k = \sum_{j=1}^k n_{\bullet j}/n = \sum_{j=1}^k p_{\bullet j}$ as the cumulative relative frequency up to the k th column category, respectively so that $p_{\bullet 1} = d_1 \leq d_2 \leq \dots \leq d_K = 1$. Taguchi (1966) proposed the statistic

$$T = \sum_{k=1}^{K-1} w_k \left[\sum_{i=1}^I n_{i\bullet} \left(\frac{Z_{ik}}{n_{i\bullet}} - d_k \right)^2 \right] \quad (1)$$

as a measure of the association between the row and ordered column variables. Nair (1987) considers the case where $w_k = [d_k(1 - d_k)]^{-1}$ and showed that the distribution of T can be approximated using Satterthwaite's (1946) method. Unlike the Pearson chi-squared statistic Taguchi's statistic, T , is finitely bounded by the interval $[0, n(I - 1)]$.

Suppose we consider $Z_{ik}/n_{i\bullet} - d_k$ in (1). It can be alternatively expressed as $\sum_{j=1}^k (p_{ij}/p_{i\bullet} - p_{\bullet j})$. Ignoring the summation term for the moment, this is the difference between the unconditional marginal prediction of a column category, $p_{\bullet j}$, and the conditional prediction of that category, $p_{ij}/p_{i\bullet}$. Note that Taguchi's statistic is a cumulative version of the numerator of the Goodman-Kruskal tau

index (Goodman and Kruskal, 1954)

$$\tau_N = \sum_{i=1}^I \sum_{k=1}^K p_{i\bullet} \left(\frac{p_{ik}}{p_{i\bullet}} - p_{\bullet k} \right)^2 .$$

which is identical to Light and Margolin’s (1971) R^2 and has a proportion of explained variation interpretation.

Note that τ_N does not involve the cumulative frequencies of the contingency table and therefore treats the variables as having a nominal structure. Thus, Taguchi’s statistic considers a uni-directional association between the variables instead of a two-way association structure that the Pearson chi-squared statistic considers. For the contingency table, $\mathbf{N}$, considered in this paper, the statistic (1) implies that the row variable is treated as a predictor variable while the (ordered) column variable is treated as a response variable.

2.2 Example

To demonstrate the application of Taguchi’s statistic and the Goodman-Kruskal tau numerator, consider Table 1 which summarises the taste-testing data based on example 3 of Bradley, Katti and Coons (1962). This table represents a sample of 210 individuals that were asked their impression of five different foods and to judge it on a five point scale - 1 (terrible) to 5 (excellent). Therefore, given a food that someone is asked to test, there are two questions that interest us - firstly “is it possible to predict how that food will be rated, when compared against other foods?” If such a question can be adequately addressed the second question is “is there a perceived difference between the ratings (column categories) when considering the ordered structure of the variables using cumulative frequencies?” Therefore, the foods will be considered as the predictor variable in the study and the ratings will be treated as the response variable.

————— Table 1 about here —————

The Goodman-Kruskal tau numerator is 0.081. Therefore, using the approach proposed by Light and Margolin (1971) to formally test the significance of this measure, the foods studied do (statistically) significantly influence the ratings they receive. Therefore NSCA can be used to graphically depict how the five different foods influence the rating it receives (see Section 5). However, since the rating (column) variable is ordered, one can also determine how similar or different cumulate categories are perceived by those individuals in the study. One can also identify those food groups which dominate these perceived rating differences. Therefore, rather than using the Goodman-Kruskal statistic, we instead use Taguchi’s statistic. Doing so yields $T = 110.78$, where, for $w_k = [d_k(1 - d_k)]^{-1}$, $d_1 = 0.20$, $d_2 = 0.38$, $d_3 = 0.56$ and $d_4 = 0.92$. Using

the procedure outlined in Nair (1987, section 3.2), the critical value of Taguchi's statistic at the 0.05 level of significance is 69.90 thereby confirming the conclusion obtained using the Goodman-Kruskal tau index, but taking into account the ordinal structure of the response variable. Table 2 illustrates the mathematical property that the Pearson chi-squared statistic of each subtable formed by aggregating column categories sums to T .

————— Table 2 about here —————

Now that it has been established that by comparing the different foods, a food type can help predict the rating its receives, we can now investigate the nature of this one-way association. This can be done by not only predicting the rating a food will receive (using the Goodman-Kruskal tau index) but also identify any perceived difference between cumulate ratings and, if so, what food dominates such differences (Taguchi's statistic). To delve more deeply into this issue, the next section will explore the role of Taguchi's statistic with CA.

3 Decomposing Taguchi's Statistic

3.1 Taguchi's Statistic Revisited

Consider Taguchi's statistic (1). If we denote

$$y_{ik} = \sqrt{w_k n_{i\bullet}} \left(\frac{Z_{ik}}{n_{i\bullet}} - d_k \right) \quad (2)$$

then this is the (i, k) th element of the matrix

$$\mathbf{Y} = \sqrt{n} \times \mathbf{D}_I^{1/2} \left(\mathbf{D}_I^{-1} \mathbf{P} - \mathbf{1}_I \mathbf{c}^T \right) \mathbf{U} \mathbf{W}^{1/2} \quad (3)$$

Here, $\mathbf{U}$ is an upper triangular matrix of 1's with the K th column removed so that it is of dimension $K \times (K - 1)$. The presence of $\mathbf{U}$ ensures that the cumulative sum of the cells of the contingency table are formed. Also, $\mathbf{W}$ is the diagonal matrix of w_k values and $\mathbf{1}_I$ is a $I \times 1$ vector with elements 1. The (i, k) th element of the matrix $\mathbf{Y}$ reflects the deviation

$$H_0 : \pi_{ik} = \frac{Z_{ik}}{n_{i\bullet}} - d_{\bullet k} = \sum_{j=1}^k \left(\frac{p_{ij}}{p_{i\bullet}} - p_{\bullet j} \right) = 0$$

which, taking into consideration the cumulative nature of the ordered column categories, is the hypothesis of zero predictability of the column variable given the presence of the row variable. Such a hypothesis is consistent with performing NSCA (D'Ambra and Lauro, 1989). By considering (3), Taguchi's statistic can be expressed in matrix form by

$$\begin{aligned}
T &= \text{trace}(\mathbf{Y}\mathbf{Y}^\top) \\
&= n \times \text{trace} \left[\mathbf{D}_I^{1/2} (\mathbf{D}_I^{-1} \mathbf{P} - \mathbf{1}_I \mathbf{c}^\top) \mathbf{U} \mathbf{W} \mathbf{U}^\top (\mathbf{D}_I^{-1} \mathbf{P} - \mathbf{1}_I \mathbf{c}^\top)^\top \mathbf{D}_I^{1/2} \right]
\end{aligned}$$

If we replace $\mathbf{U} \mathbf{W} \mathbf{U}^\top$ by $\mathbf{I}$ then Taguchi's statistic, T , is the numerator of the Goodman-Kruskal tau index (Goodman and Kruskal, 1954).

3.2 CA using the T Statistic

Suppose we let $\tilde{\mathbf{Y}} = \mathbf{Y}/\sqrt{n}$. Correspondence analysis using Taguchi's statistic may be performed by applying the SVD of (3) such that

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{A}} \mathbf{D}_\lambda \tilde{\mathbf{B}}^\top \quad (4)$$

where $\tilde{\mathbf{A}}$ is a $I \times M$ matrix of singular vectors associated with the row categories and $\tilde{\mathbf{B}}$ is a $(K - 1) \times M$ matrix of singular vectors associated with the difference between cumulate column categories. For both matrices $M = \min(I, K - 1)$ and they are subject to the constraint $\tilde{\mathbf{A}}^\top \tilde{\mathbf{A}} = \mathbf{I}$ and $\tilde{\mathbf{B}}^\top \tilde{\mathbf{B}} = \mathbf{I}$. The diagonal elements of $\mathbf{D}_\lambda$ are real and positive and are the first $M = \min(I, K - 1)$ singular values of $\mathbf{Y}$. They are arranged in descending order so that

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_M.$$

Taguchi's inertia, defined as T/n , may be expressed as the sum of squares of these singular values since

$$\begin{aligned}
\frac{T}{n} &= \text{trace} [\tilde{\mathbf{Y}}^\top \tilde{\mathbf{Y}}] \\
&= \text{trace} [\tilde{\mathbf{B}} \mathbf{D}_\lambda \tilde{\mathbf{A}}^\top \tilde{\mathbf{A}} \mathbf{D}_\lambda \tilde{\mathbf{B}}^\top] \\
&= \text{trace} [\tilde{\mathbf{B}} \mathbf{D}_\lambda^2 \tilde{\mathbf{B}}^\top] \\
&= \text{trace} [\mathbf{D}_\lambda^2].
\end{aligned}$$

It can be similarly shown that $T = \text{trace} [\tilde{\mathbf{Y}} \tilde{\mathbf{Y}}^\top] = \text{trace} [\mathbf{D}_\lambda^2]$. Therefore, Taguchi's inertia can be partitioned into the sum of squares of the singular values in much the same way as NSCA involves such a partition of the numerator of the Goodman-Kruskal tau index and classical CA involves partitioning the Pearson chi-squared statistic.

4 Properties

By considering Taguchi's statistic, this section describes some of the features that make CA an important tool in categorical data analysis. In particular, we present coordinates to visualise the association between the two variables of $\mathbf{N}$, a reconstitution formula and distance measures.

4.1 Graphical Representation

To graphically represent the association between the two variables using the SVD of Taguchi's statistic, the row and column coordinates may be defined as

$$\mathbf{F} = \mathbf{D}_I^{-1/2} \tilde{\mathbf{A}} \mathbf{D}_\lambda = \mathbf{D}_I^{-1/2} \tilde{\mathbf{Y}} \tilde{\mathbf{B}} \quad (5)$$

$$\mathbf{G} = \mathbf{W}^{-1/2} \tilde{\mathbf{B}} \mathbf{D}_\lambda = \mathbf{W}^{-1/2} \tilde{\mathbf{Y}}^\top \tilde{\mathbf{A}} \quad (6)$$

respectively. Therefore, if there is approximately zero predicability of the column categories given the row categories then $\mathbf{F} \approx 0$ and $\mathbf{G} \approx 0$ since $\mathbf{Y} \approx 0$. If one considers the expression of $\mathbf{Y}$ given by (3) then these coordinates may be alternatively expressed as

$$\mathbf{F} = \mathbf{D}_I^{-1} \mathbf{P} \mathbf{U} \mathbf{W}^{-1/2} \tilde{\mathbf{B}} \quad \mathbf{G} = \mathbf{U}^\top \mathbf{P}^\top \mathbf{D}_I^{-1/2} \tilde{\mathbf{A}}.$$

By considering such scaling, Taguchi's inertia can be expressed as

$$\frac{T}{n} = \text{trace}(\mathbf{F}^\top \mathbf{D}_I \mathbf{F}) = \text{trace}(\mathbf{G}^\top \mathbf{W} \mathbf{G}) = \text{trace}(\mathbf{D}_\lambda^2)$$

so that the metric of the nominal row coordinates is $\mathbf{D}_I$ and the metric of the column categories is $\mathbf{W}$.

Considering these coordinates provides a unique interpretation of the plot obtained by performing CA. One must note that for the column coordinates $\mathbf{G}$ a coordinate does not reflect the position of each ordered column category. Instead it provides an insight into establishing how different cumulate (ordered) column categories are given that they are subject to the presence of the row categories.

To provide a more discriminating view of the difference between each cumulate rating category, one may consider rescaling the row and column profile coordinates to obtain biplot-type coordinates (Goodman, 1986):

$$\begin{aligned} \mathbf{F}_\alpha &= \mathbf{D}_I^{-1/2} \tilde{\mathbf{B}} \mathbf{D}_\lambda^\beta \\ \mathbf{G}_\alpha &= \mathbf{W}^{-1/2} \tilde{\mathbf{A}} \mathbf{D}_\lambda^{1-\beta} \end{aligned}$$

for some $0 \leq \beta \leq 1$. Note that these coordinates are related to the factorisation (for categorical data) proposed by Gabriel (1971) for the construction of the biplot.

4.2 Modelling Taguchi's Statistic

If we consider (3) and (4) we get

$$\mathbf{P} \mathbf{U} = \mathbf{D}_I (\mathbf{1}_I \mathbf{c}^\top \mathbf{U} + \mathbf{D}_I^{-1/2} \tilde{\mathbf{A}} \mathbf{D}_\lambda \tilde{\mathbf{B}}^\top \mathbf{W}^{-1/2}) \quad (7)$$

and is the reconstitution formula for CA using Taguchi's statistic. Note that, one if disregards $\mathbf{U}$ and $\mathbf{W}$ from (7) then we end up with the reconstitution formula for NSCA.

Alternatively one may reconstitute the cumulative sum of the cells by considering the row and profile coordinates defined by (5) and (6). Doing so yields the correspondence model

$$\mathbf{PU} = \mathbf{D}_I \left(\mathbf{1}_I \mathbf{c}^\top \mathbf{U} + \mathbf{F} \mathbf{D}_\lambda^{-1} \mathbf{G}^\top \right). \quad (8)$$

If $\mathbf{F} \approx 0$ and/or $\mathbf{G} \approx 0$, then $\mathbf{PU} = \mathbf{D}_I \mathbf{1}_I \mathbf{c}^\top \mathbf{U}$ which can be alternatively written as $\left(\mathbf{D}_I^{-1} \mathbf{P} - \mathbf{1}_I \mathbf{c}^\top \right) \mathbf{U}$ which is the matrix form of H_0 in section 3.1. Therefore row coordinates near the origin are not useful predictors of the column response categories. They are also not useful for identifying whether there is perceived difference between cumulate column response categories. Similarly, those row coordinates far from the origin are helpful in discriminating between those cumulate column categories whose column coordinate is in close proximity. Also column coordinates near the origin suggest that there is very little perceived difference between those cumulate categories. The next section describes more on how to interpret the distance of a row/column coordinate from the origin.

4.3 Distances

4.3.1 Distance from Origin

Consider the (i, m) th element of (5). The i th row coordinate on the m th dimension of the correspondence plot is

$$f_{im} = \frac{\tilde{a}_{im}}{\sqrt{p_{i\bullet}}} \lambda_m$$

and is centred around the origin of the M -dimensional correspondence plot. Therefore the squared Euclidean distance of the i th row coordinate from the origin of the plot is

$$\begin{aligned} d_I^2(i, 0) &= \sum_{m=1}^M f_{im}^2 \\ &= \sum_{k=1}^{K-1} \frac{1}{p_{i\bullet}} \left(\sum_{m=1}^M \tilde{a}_{im} \lambda_m \tilde{b}_{jm} \right)^2 \\ &= \sum_{k=1}^{K-1} \frac{1}{p_{i\bullet}} \left(\sqrt{w_k p_{i\bullet}} \left(\frac{Z_{ik}}{n_{i\bullet}} - d_k \right) \right)^2 \\ &= \sum_{k=1}^{K-1} w_k \left(\frac{Z_{ik}}{n_{i\bullet}} - d_k \right)^2 \end{aligned}$$

so that $\{f_{im} : m = 1, \dots, M\}$ is the coordinate of the i th cumulative row profile in an M -dimensional subspace. Note that Taguchi's inertia can be expressed in terms of this distance by

$$\frac{T}{n} = \sum_{i=1}^I p_{i\bullet} d_I^2(i, 0) .$$

Therefore, if the i th row coordinate is far from the origin then it is an important category for predicting the column responses. Similarly, a row coordinate situated close to the origin indicates that that particular category does not contribute to the predictability of the column responses.

We can also obtain a similar expression for the ordered column categories. Consider the (k, m) th element of (6), for $k = 1, 2, \dots, K-1$. The coordinate of the k th pair of cumulate categories on the m th dimension of the correspondence plot,

$$g_{km} = \frac{\tilde{b}_{km}}{\sqrt{w_k}} \lambda_m .$$

Then the squared Euclidean distance of this coordinate in an M -dimensional plot is

$$d_K^2(k, 0) = \sum_{m=1}^M g_{km}^2 = \sum_{i=1}^I \frac{1}{w_k} \left(\sum_{m=1}^M \tilde{a}_{im} \lambda_m \tilde{b}_{km} \right)^2 = \sum_{i=1}^I p_{i\bullet} \left(\frac{Z_{ik}}{n_{i\bullet}} - d_k \right)^2$$

so that Taguchi's inertia is

$$\frac{T}{n} = \sum_{k=1}^{K-1} w_k d_K^2(k, 0)$$

This distance measure can be used to identify those cumulate column categories that have a similar or different profile, while preserving the ordinal structure of the column categories and given the uni-directional association nature of the two variables. For example, suppose that a position of an ordered pair of cumulate column categories lies some distance from the origin. Then there is a difference between the profiles of the two ordered categories. Its proximity to a row coordinate will reflect how well that row category influences the outcome of these ordered column responses.

4.3.2 Intra-Variable Distance

To measure the intra-variable distance consider the squared Euclidean distance between the i th row and i' th row categories in the full M -dimensional correspondence plot

$$d_I^2(i, i') = \sum_{m=1}^M (f_{im} - f_{i'm})^2$$

It can be shown that, by considering the property of $\tilde{a}_{im}$ and Z_{ik} ,

$$d_I^2(i, i') = \sum_{k=1}^{K-1} w_k \left(\frac{Z_{ik}}{n_{i\bullet}} - \frac{Z_{i'k}}{n_{i'\bullet}} \right)^2.$$

Therefore two cumulative row profiles, $Z_{ik}/n_{i\bullet}$ and $Z_{i'k}/n_{i'\bullet}$ that are approximately the same will share a very similar position in the correspondence plot. Similarly, two cumulative row profiles that are different will be at a distance from one another in the plot. It can be shown that the squared Euclidean distance between the k th and k' th column coordinates is

$$d_K^2(k, k') = \sum_{i=1}^I \frac{1}{n^2 p_{i\bullet}} (Z_{ik} - Z_{ik'})^2$$

5 Taste-Testing Data Revisited

Suppose we reconsider Table 1. A NSCA of the table produces the correspondence plot given by Figure 1.

————— Figure 1 about here —————

Figure 1 indicates that foods C and D receive a poorer rating than the remaining three foods studied. Food D also receives a relatively large proportion of excellent ratings, food E is perceived as relatively good (associated with rating = 4) while food B is judged as average (associated with rating = 3). The figure also suggests that there is very little difference between rankings 1 and 2, but a large difference between rankings 3 and 4 and rankings 4 and 5. The two-dimensional display of Figure 1 explains most of the association that exists between the two variables (95.9%). Table 3 summarises the inertias for each of the $M = \min(5, 5) - 1 = 4$ dimensions needed to represent all of the association.

————— Table 3 about here —————

Since the ratings are ordered, we partition Taguchi's inertia (of 0.5275) and obtain the row and column coordinates defined by (5) and (6), respectively. These coordinates are jointly displayed in Figure 2. For the column categories, the label "1/2" reflects the comparison made of the cumulative total of ordered rating 1 with those of ratings 2, 3, 4 and 5 given the five foods. Similarly, the labels "2/3" reflects the comparison made of the cumulative total of the ordered ratings 1 and 2 with those of the remaining predictor categories. The remaining labels, "3/4" and "4/5" can be interpreted in a similar manner. To calculate the inertia and the coordinates we again consider $w_k = [d_k(1 - d_k)]^{-1}$ where $d_1 = 0.20$, $d_2 = 0.38$, $d_3 = 0.56$ and $d_4 = 0.92$.

Figure 2 about here

Figure 2 shows that the relative position of food categories is the same as depicted in the NSCA plot of Figure 1. This is because, for both analyses the row variable is treated as the predictor variable, helping to predict how well a food is judged. However, while NSCA allows for one to identify how each row category helps predict a column response, Taguchi's analysis allows for one to identify how similar (or different) cumulate ordered column response categories are. Consider Figure 2 which graphically depicts 98.9% of the association that exists between the two variables (note that by taking into account the ordered structure of the response variable, this figure accounts for more of the association than the plot obtain using NSCA). Figure 2 indicates that the difference between ratings 1 and 2 are negligible since the position of this point is closest to the origin. This is not surprising since these two column categories in Figure 1 are close together. Therefore, for both analyses, the categories of these two rating levels could be merged to form a single rating level. However, of the four pairs of cumulate ratings, the difference between rating 2 and rating 3, and the difference between rating 4 and rating 5, are quite distinct, indicating that there is a perceived difference between these cumulate categories. Therefore the merging of either of these pairs to form a single response rating should be avoided. The source of the variation between these ratings is dominated by food *A*, food *C* and food *D*. The apparent difference between ratings 4 and 5 (the two most positive food ratings) can be attributed to foods *A* and food *B*.

By partitioning Taguchi's inertia we obtain the inertia values summarised in Table 4.

Table 4 about here

It is therefore apparent that the rating categories are influenced by the five different foods considered in the study. Although such an interpretation is not immediately clear when viewing Figure 2 since the column points appear clustered close to the origin. To discriminate more clearly the proximity of these points one may view the row isometric plot of Figure 3.

Figure 3 about here

6 Discussion

This paper has discussed the use of Taguchi's statistic for performing CA. Making use of the statistic has the advantage of taking into consideration the structure of ordinal categorical variables by determining the cumulative sum of the cell frequencies across the variable. The correspondence plots derived from the analysis of Table 1 graphically show how the row categories are asymmetrically associated

with the ordinal column variable. Such plots also reveal how CA using Taguchi's statistic can determine how similar, or dissimilar, cumulate response categories are. Since Taguchi's statistic (1) implies an asymmetric association structure between a nominal and ordinal variable of a two-way contingency table, this paper has also shown the link between CA using Taguchi's statistic and NSCA which considers the Goodman-Kruskal tau index. Such a link exists by the presence, or absence of the matrices $\mathbf{U}$ and $\mathbf{W}$ since if they are not incorporated into Taguchi's analysis then NSCA is performed.

Future work may reveal generalisations of Taguchi's statistic for the CA of a doubly ordered two-way contingency table as well as the development of multivariate versions of the statistic for multiple CA.

Références

- Agresti, A. (2007). *An Introduction to Categorical Data Analysis (Second Edition)*, N.J: Wiley.
- Barlow, R. E., Bartholomew, D. J., Bremner, J. M., Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions*. New York: Wiley.
- Beh, E. J. (1997). Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials. *Biometrical Journal* 39: 589-613.
- Beh, E. J. (2004). Simple correspondence analysis: A bibliographic review. *International Statistical Review* 72: 257-284.
- Bradley, R. A., Katti, S. K., Coons, I. J. (1962). Optimal scaling for ordered categories. *Psychometrika* 27: 355-374.
- Cuadras, C. M. (2002). Correspondence analysis and diagonal expansions in terms of distribution functions. *Journal of Statistical Planning and Inference* 103: 137-150.
- D'Ambra, L., Lauro, N. (1989). Non symmetrical analysis of three-way contingency tables. In: Coppi, R., Bolasco, S., eds. *Multiway Data Analysis*, Elsevier Science Publishers B. V., North-Holland, 301-315.
- Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58: 453-467.
- Goodman, L. A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review* 54: 243-309.
- Goodman, L. A., Kruskal, W. H. (1954). Measures of association for cross-classifications. *Journal of the American Statistical Association* 49: 732-764.
- Hirotsu, C. (1990). A critical look at accumulation analysis and related methods: Discussion. *Technometrics* 32: 133-136.
- Light, R., Margolin, B. (1971). An analysis of variance for categorical data. *Journal of the American Statistical Association* 66: 534-544.

- Nair, V. N. (1987). Chi-squared type tests for ordered alternatives in contingency tables. *Journal of the American Statistical Association* 82: 293-291.
- Parsa, A. R., Smith, B. S. (1993). Scoring under ordered constraints in contingency tables. *Communications in Statistics - Theory and Methods* 22: 3537-3551.
- Ritov, Y., Gilula, Z. (1993). Analysis of contingency tables by correspondence models subject to ordered constraints. *Journal of the American Statistical Association* 88: 1380-1387.
- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* 2: 110-114.
- Schriever, B. F. (1983). Scaling of order dependent categorical variables with correspondence analysis. *International Statistical Review* 51: 225-238.
- Taguchi, G. (1966). *Statistical Analysis (In Japanese)*. Tokyo: 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.

Foods	Rating					Total
	1 (terrible)	2	3	4	5 (excellent)	
A	9	5	9	13	4	40
B	7	3	10	20	4	44
C	14	13	6	7	0	40
D	11	15	3	5	8	42
E	0	2	10	30	2	44
Total	41	38	38	75	18	210

Table 1: Taste-testing study of five different foods rated on a five point scale

Aggregated Column Categories									
Foods	1	2+3+4+5	1+2	3+4+5	1+2+3	4+5	1+2+3+4	5	Total
A	9	31	14	26	23	17	36	4	
B	7	37	10	34	20	24	40	4	
C	11	26	27	13	33	7	40	0	
D	11	31	26	16	29	13	34	8	
E	0	44	2	42	12	32	42	2	
		$X_1^2=18.55$	$X_2^2=50.56$		$X_3^2=31.01$		$X_4^2=10.66$		110.78

Table 2: The partition of Taguchi's statistic from Table 1 into Pearson chi-squared statistics

Dimension	Principal Inertia	% of T/n	Cumulative %
1	0.07228	89.34	89.34
2	0.00532	6.57	95.91
3	0.00327	4.04	99.95
4	0.00327	0.05	100.00
Total	0.08091	100.00	

Table 3: Principal inertia values for the NSCA of Table 1

Dimension	Principal Inertia	% of T/n	Cumulative %
1	0.46965	89.03	89.03
2	0.05166	9.79	98.82
3	0.00615	1.16	99.98
4	0.00009	0.02	100.00
Total	0.52755	100.00	

Table 4: Principal inertia values using Taguchi's statistic for Table 1

Figure 1: Non-symmetrical correspondence plot of Table 1

Figure 2: Correspondence plot by partitioning Taguchi's inertia

Figure 3: Row isometric plot by partitioning Taguchi's inertia