THE UNIVERSITY OF TEXAS AT SAN ANTONIO, COLLEGE OF BUSINESS

Working Paper SERIES

Date August 15, 2014

WP # 0006MSS-253-2014

# A two-stage principal component analysis of symbolic data using equicorrelated and jointly equicorrelated covariance structures

Anuradha Roy _
Department of Management Science and Statistics
The University of Texas at San Antonio
One UTSA Circle,
San Antonio, TX 78249 USA

# A two-stage principal component analysis of symbolic data using equicorrelated and jointly equicorrelated covariance structures

Anuradha Roy *
Department of Management Science and Statistics
The University of Texas at San Antonio
One UTSA Circle
San Antonio, TX 78249, USA

**Abstract**

A new approach to derive the principal components of symbolic data is proposed in this article. This is done in two stages: first getting eigenblocks and eigenmatrices of the variance-covariance matrix, and then analyzing these eigenblocks and the corresponding principal vectors together in some seemly sense to get the adjusted eigenvalues and the corresponding eigenvectors of the interval data. The proposed method is very efficient in two-level and three-level symbolic data sets. Results illustrating the accuracy and appropriateness of the new method over the existing methods are presented. We have clearly shown with the help of examples that our proposed method for principal component analysis (PCA) of three-level symbolic data generalizes the commonly used PCA for multivariate data.

*Keywords:* Jointly equicorrelated covariance structure; symbolic data; Two-stage principal component analysis

**JEL Classification:** C13, C30

## 1 Introduction

Advances in computing power in the past few decades greatly encouraged the collection of tensor data sets in all fields of science, biomedical, medical, social science, engineering and business. In many of these areas, recent technological advances allow for the collection of massive datasets with interval-valued variables which occur naturally and is very common these days. In these situations right thing to do is to model the symbolic data, especially the interval data which captures the variability of events, rather than classical data. Moreover, in many real world applications the available information is imprecise and ambiguous, and therefore cannot be expressed by a single numerical data. In these cases it is better to summarize the information using interval-valued data.

*Correspondence to: Anuradha Roy, Department of Management Science and Statistics, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249, USA

Lauro and Palumbo (2000), Palumbo and Lauro (2003), Giordan and Kiers (2006) among many others developed methods of conducting principal component analysis (PCA) on symbiloc data for which each symbolic variable $y_j$, $j = 1, \ldots, p$, takes interval values $[y_{i\_j}^-, y_{i\_j}^+]$ for each observation $i = 1, \ldots, n$, and where each observation may even represent the aggregation of $n_i$ individuals.

Thus, an interval-valued symbolic random variable is one that takes values in an interval. Billard and Diday (2006) says, "It is the presence of this internal variation which necessitates the need for new techniques for analysis which in general will differ from those for classical data". But, in this paper we consider the interval valued data as two repeated measurements at the lower and upper bounds of an interval, and develop a new method using some recently developed classical multi-level multivariate techniques (Leiva and Roy, 2011; Roy and Fonseca, 2012) carefully and prudently to analyze the interval data. In this paper we especially develop a new method to derive principal components (PCs) of symbolic data, and consider the Fruit Juice data from Giordani and Kiers (2006, Table 4) which is reproduced here as Table 1 to show the performance of our new method. This interval data set describing 16 fruit juices evaluated by a group of judges on six features, namely, Appearance, Smell, Taste, Naturalness, Sweetness and Density. More specifically, there are eight fruit juices (apple, apricot, banana, pineapple, grapefruit, orange, peach and peer) and two brands for each juice. We guess the same fruit juice of the two different brands should have some common factors like Appearance and Smell. Unfortunately, Giordani and Kiers (2006) did not use this brand information in deriving the principal components in their paper. We use this brand information in this article to derive the principal components and as a consequence there is a substantial improvement in the result (Roy, 2014a,b). Without the brand information first two PCs account for an apparent proportion of 85.54%, whereas the use of brand information improves it to 91.51%.

All the judges evaluated the Appearance and the Smell before tasting and the remaining characteristics later. It reveals the fact that these two features, Appearance and Smell, bring together the first impression of the fruit juices. To evaluate each attribute, a scale, whose values are from 1 to 10, is used. Unfortunately, the inter individual differences in judges were unknown. With respect to the rating pertaining to the $i$th juice and the $j$th variable, only the lower bound ($y_L^-$), the upper bound $y_R^+$, the mean value ($m$) and the standard deviation ($s$) are known. Every interval datum is then constructed as $(y_L^-, y^- = m - s, y^+ = m + s, y_R^+)$. Since each attribute was evaluated on a scale whose values are from 1 to 10, we may assume that the variance-covariance matrix of the six features is same at the two ends of the intervals.

In this article we introduce a new method using a two-stage principal component analysis exploiting multivariate equicorrelated (exchangeable) and jointly equicorrelated (doubly exchangeable) covariance matrix (Roy and Leive, 2011, 2007) as defined in Section 2.2.1 to the Fruit juice interval data considering it as two-level and three-level data respectively. We show in Example 1 in Section 3 that the midpoints and the midranges of the six interval valued variables are the first two principal vectors of the equicorrelated covariance matrix considering the Fruit juice data as two-level. And, grand midpoints and grand midranges of the six interval valued variables are the first and the third principal vectors of the jointly equicorrelated covariance matrix considering the Fruit juice data as three-level; the brand difference of the juices turns out to be the second principal vector (see Example 2 in Section 3). The introduction of our new method needs some preliminaries, which we present in the next section.

# 2 Preliminaries

## 2.1 Matrices of Intervals

Let $I[\boldsymbol{Y}]$ represents the interval valued data matrix having $p$ columns and n rows, where $p$ denotes the number of features/variables and $n$ denotes the number of sampling units. So, we may write $I[\boldsymbol{Y}]$ as

$$I(\boldsymbol{Y}) = \begin{pmatrix} \boldsymbol{y}_1' \\ \vdots \\ \boldsymbol{y}_n' \end{pmatrix} = \begin{pmatrix} [y_{1\_1}^-, y_{1\_1}^+] & \cdots & [y_{1\_p}^-, y_{1\_p}^+] \\ \vdots & \ddots & \vdots \\ [y_{n\_1}^-, y_{n\_1}^+] & \cdots & [y_{n\_p}^-, y_{n\_p}^+] \end{pmatrix}, \tag{1}$$

where each component is an interval. The $i$th row of $I(\boldsymbol{Y})$ pertains to the $i$th observation unit, $i = 1, \ldots, n$. As each observation unit is characterized by $p$ (interval valued) variables, it can be represented as a hyperrectangle in $\in \mathbb{R}^p$ and the number of vertices of each hyperrectangle is $2^p$. If $p = 1$, each hyper rectangle reduced to a segment, a rectangle if $p = 2$, a parallelepiped or paralleletope in case of $p = 3$ and $p > 3$ respectively.

The extension of the PCA to the interval data has been proposed by Cazes, Chouakria, Diday and Schektman (1997) and by Chouakria, Diday and Cazes (1999) as "Vertices Principal Component Analysis" (V-PCA). See Bock and Diday (2000) too. V-PCA does not directly summarize the interval data in (1). Each interval valued row is however transformed into the numerical matrix $\boldsymbol{Y}_i$, such that each row in $\boldsymbol{Y}_i$ refers to the $i$th hyperrectangle. Therefore, $\boldsymbol{Y}_i$'s $i = 1, \ldots, n$, has $2^p$ rows and $p$ columns. By stacking one below the other the matrices $\boldsymbol{Y}_i$'s $i = 1, \ldots, n$, we get the new numerical valued data matrix $\boldsymbol{Y}_{\text{V-PCA}}$ with $(n2^p \times p)-$dimension as

Table 1: Fruit juices interval data

| Fruit juices | Appearance | Smell | Taste | Naturalness | Sweetness | Density |
|---|---|---|---|---|---|---|
| Apple1 | (6.78,6.78,7.50,7.52) | (5.47,5.59,6.49,6.59) | (7.40,7.40,8.17,8.40) | (5.66,5.77,6.86,7.20) | (7.27,7.27,7.99,8.29) | (5.81,5.81,6.7,6.74) |
| Apple2 | (6.60,6.79,7.64,7.72) | (6.28,6.34,7.23,7.40) | (6.31,6.32,7.33,7.43) | (5.72,5.87,6.91,7.12) | (6.67,6.67,7.57,7.65) | (5.47,5.55,6.53,6.59) |
| Apricot1 | (6.82,6.82,7.50,7.68) | (7.87,7.87,8.45,8.68) | (7.60,7.60,8.36,8.54) | (7.35,7.51,8.25,8.47) | (7.42,7.46,8.11,8.40) | (7.03,7.04,7.82,8.15) |
| Apricot2 | (7.32,7.53,8.15,8.16) | (7.09,7.09,7.89,8.19) | (5.17,5.42,6.42,6.71) | (4.66,4.81,5.82,6.06) | (4.90,5.15,6.15,6.31) | (5.79,5.87,6.72,6.77) |
| Banana1 | (4.96,5.24,6.21,6.37) | (3.92,4.14,5.20,5.60) | (3.64,4.13,5.20,5.32) | (4.27,4.63,5.68,5.95) | (4.76,4.98,5.92,6.16) | (3.62,3.78,4.73,4.74) |
| Banana2 | (5.27,5.46,6.46,6.67) | (3.68,3.98,5.08,5.36) | (3.26,3.58,4.69,4.94) | (3.92,4.15,5.18,5.46) | (4.23,4.57,5.63,5.91) | (3.65,3.83,4.77,4.77) |
| Grapefruit1 | (6.28,6.30,7.26,7.40) | (6.52,6.65,7.59,7.65) | (5.17,5.46,6.58,6.85) | (6.00,6.16,7.20,7.33) | (2.45,2.65,3.39,3.39) | (3.64,3.84,4.72,4.76) |
| Grapefruit2 | (6.31,6.42,7.21,7.43) | (5.63,5.83,6.70,6.75) | (6.35,6.46,7.30,7.47) | (6.11,6.12,6.96,7.23) | (4.14,4.14,5.02,5.19) | (3.06,3.38,4.34,4.46) |
| Orange1 | (6.64,6.64,7.44,7.59) | (7.12,7.15,7.97,8.24) | (6.39,6.39,7.29,7.44) | (5.67,5.74,6.70,6.72) | (5.75,5.75,6.57,6.67) | (3.64,3.80,4.76,4.97) |
| Orange2 | (6.89,6.93,7.55,7.55) | (6.06,6.09,6.87,6.90) | (6.82,6.82,7.66,7.94) | (5.60,5.75,6.69,6.72) | (5.93,5.93,6.89,7.13) | (3.88,4.06,4.98,4.98) |
| Peach1 | (7.09,7.21,7.81,7.93) | (6.94,6.94,7.69,7.78) | (6.42,6.52,7.44,7.54) | (5.70,5.89,6.86,7.10) | (6.69,6.75,7.56,7.68) | (5.03,5.03,5.92,5.92) |
| Peach2 | (6.98,7.01,7.74,7.82) | (6.22,6.29,7.11,7.11) | (7.38,7.38,8.15,8.38) | (6.83,6.83,7.60,7.72) | (6.83,6.96,7.74,7.81) | (4.99,4.99,5.83,5.85) |
| Peer1 | (6.89,6.89,7.67,7.76) | (7.19,7.28,8.04,8.24) | (7.14,7.17,7.99,8.19) | (6.44,6.47,7.33,7.49) | (7.59,7.59,8.37,8.54) | (7.22,7.34,8.06,8.27) |
| Peer2 | (7.52,7.52,8.20,8.20) | (6.32,6.40,7.28,7.44) | (7.69,7.69,8.33,8.57) | (6.72,6.72,7.48,7.63) | (7.71,7.71,8.45,8.62) | (6.72,6.72,7.60,7.67) |
| Pineapple1 | (6.61,6.77,7.51,7.66) | (5.74,5.74,6.54,6.66) | (6.18,6.21,7.10,7.31) | (5.45,5.52,6.52,6.85) | (5.63,5.82,6.71,6.75) | (3.92,4.16,5.00,5.00) |
| Pineapple2 | (6.66,6.66,7.42,7.59) | (5.90,6.19,7.09,7.30) | (5.65,5.84,6.76,6.98) | (5.23,5.52,6.48,6.56) | (5.52,5.62,6.62,6.92) | (3.28,3.69,4.67,4.69) |

follows:

$$\boldsymbol{Y}_{\text{V-PCA}} = \begin{pmatrix} \boldsymbol{Y}_1' \\ \vdots \\ \boldsymbol{Y}_n' \end{pmatrix}. \tag{2}$$

V-PCA perform PCA on (2). The matrix $\boldsymbol{Y}_{\text{V-PCA}}$ is now treated as though it represents classical $p-$variate data for $n2^p$ individual units. Chouakria (1998) has shown that the basic theory for a classical analysis carries through; hence a traditional PCA can be applied. Regrettably, V-PCA requires the analysis of the data matrix $\boldsymbol{Y}_{\text{V-PCA}}$, the dimension of which increases exponentially with the number of variables.

An exploratory tool in order to summarize interval valued data sets is centers principal component analysis (C-PCA), as proposed by Cazes et al. (1997). Like V-PCA, C-PCA also transforms the interval valued data matrix into a new single valued matrix – midpoint or center of the interval at hand. Thus, C-PCA method basically consists of a PCA on the centers of the intervals, whereas the V-PCA method uses all vertices of the hyperrectangle defined by the intervals for all variables for each observation. A generic interval $I[y]_{ij} \equiv [y_{i\_j}^-, y_{i\_j}^+] \, \forall \, i = 1, \ldots, n, \; j = 1, \ldots, p$ and $y_{i\_j}^- \leq y_{i\_j}^+$. The interval $I[y]_{ij}$ can also be expressed by the couple $\{y_{i\_j}^c, y_{i\_j}^r\}$ where $y_{i\_j}^c = \frac{1}{2}(y_{i\_j}^- + y_{i\_j}^+)$ and $y_{i\_j}^r = \frac{1}{2}(y_{i\_j}^+ - y_{i\_j}^-)$. As mentioned in the Introduction, these are actually two principal vectors of the equicorrelated covariance structure (see Example 1 in Section 3), but not normalized for two-level interval data. These two principal vectors together account for the total variance of the data. So, just performing the principal component analysis on the new midpoint variables $y_{i\_j}^c$, $j = 1, 2, \ldots, p$, which are actually the components of first principal vector, separately and calculating the percent eigenvalues separately is not correct. The eigenvalues calculated on the new midpoint variables $y_{i\_j}^c$, $j = 1, 2, \ldots, p$ are the second stage eigenvalues. Palumbo and Lauro (2003) developed a PCA method for interval-valued data based on midpoints $(y_{i\_j}^c)$ and midranges $(y_{i\_j}^r)$ for $j = 1, 2, \ldots, p$ separately, and calculated the percent eigenvalues and percent cumulative eigenvalues separately or partially. Giordani and Kiers (2006) also calculated PCs separately on midpoint variables, which account for 99.70% of the total variation of the Fruit juice data. Consequently, 0.30% of the total variation of the data is accounted for by midrange variables, which are the components of the second principal vector. The eigenvalues calculated by these authors are the second stage eigenvalues on the first stage principal vectors. Regrettably, they did not adjust the percent eigenvalues and percent cumulative eigenvalues to the total variance, but worked on a partial basis. We show in this article that one should not perform PCA on midpoint (C-PCA) by itself, as it only accounts for the partial variance and

does not account for the total variance of the data.

We assume that the interval variables in $I[\boldsymbol{Y}]$ have been centered with respect to their mean interval $I[\bar{y}]_j$ for $j = 1, \ldots, p$. Thus, the interval valued data matrix can be written as $I[\boldsymbol{Y}] \equiv \{\boldsymbol{Y}^c, \boldsymbol{Y}^r\}$, where

$$\boldsymbol{Y}_{\text{C-PCA}} = \boldsymbol{Y}^c = \begin{pmatrix} y^c_{1\_1} & \cdots & y^c_{1\_p} \\ \vdots & \ddots & \vdots \\ y^c_{n\_1} & \cdots & y^c_{n\_p} \end{pmatrix}, \tag{3}$$

$$\text{and} \quad \boldsymbol{Y}^r = \begin{pmatrix} y^r_{1\_1} & \cdots & y^r_{1\_p} \\ \vdots & \ddots & \vdots \\ y^r_{n\_1} & \cdots & y^r_{n\_p} \end{pmatrix}.$$

C-PCA perform PCA on (3). Palumbo and Lauro (2003) mentioned that PCA on interval-valued data can be then resolved in terms of midpoints, midranges. The terms $(\boldsymbol{Y}'^c\boldsymbol{Y}^c)$ and $(\boldsymbol{Y}'^r\boldsymbol{Y}^r)$ are two standard variance-covariance matrices computed on single-valued data. Palumbo and Lauro (2003) also mentioned that two independent PCA's could be singly exploited on these two matrices that do not cover the whole variance. Then they introduced residual variance-covariance matrix to solve the issue and defined the global variance-covariance matrix. We show in this article that these two independent PCs really cover the whole variance. However, Palumbo and Lauro (2003) have correspondingly introduced standardized interval matrix $I(\boldsymbol{Z}) \equiv \{\boldsymbol{Y}^c\boldsymbol{\Sigma}^{-1}, \boldsymbol{Y}^r\boldsymbol{\Sigma}^{-1}\}$ to get both midpoint and midranges PCA, which admit an independent representation, where the square diagonal $(p \times p)$ matrix $\boldsymbol{\Sigma}$ has the generic term $\sigma_j$, with $\sigma_j^2$ represents the generic diagonal term of the global variance-covariance matrix. It is to be noticed that both the Equations (6) and (7) on Page 6 in their paper are not properly represented, as $(\boldsymbol{Y}^c\boldsymbol{\Sigma}^{-1})$ is not a $(p \times p)$ matrix, it should be rather $(\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}'^c\boldsymbol{Y}^c\boldsymbol{\Sigma}^{-1})$. The same correction also holds for the Equation (7).

In the literature we see that C-PCA is widely used by many authors (Cazes et al., 1997; Palumbo and Lauro, 2003; Giordani and Kiers; 2006). Among them Palumbo and Lauro (2003) used midpoints (C-PCA) and midranges separately, and Cazes et al. (1997), and Giordani and Kiers (2006) used C-PCA by itself. This is however not appropriate and thus not accurate. One has to accomplish PCA for interval data together with midpoint (C-PCA) and midrange variables to calculate percent and percent cumulative eigenvalues. Otherwise, some part of the total variance would be unaccountable. Thus, in this article we propose adjusted percent eigenvalues and adjusted percent cumulative eigenvalues. We show in Section 5.1 that to get adjusted percent eigenvalues we must divide each eigenvalue corresponding to midpoint variables and midrange variables by a sum of total midpoint eigenvalues and total midrange eigenvalues, because the total

variance in the data is the sum of total midpoint eigenvalues and total midrange eigenvalues for two-level interval data. We then extend this concept to adjusted percent eigenvalues and adjusted percent cumulative eigenvalues for three-level interval data in Section 5.2.

### 2.2 Formulation of the problem

As mentioned in the introduction we use the recently developed multi-level covariance structure (Roy and Leive, 2011, 2007) to derive the PCs of the Fruit juice data. We consider the six variables as the first level, the lower and upper bounds of an interval as the second level and the two brands as the third level of the Fruit juice data. Thus each of the eight observation units in this data has information on three multivariate level. Let a typical sample in the Fruit juice data looks like $\big((y_{11\_1}, y_{12\_1}), \ldots, (y_{11\_6}, y_{12\_6}), (y_{21\_1}, y_{22\_1}), \ldots, (y_{21\_6}, y_{22\_6})\big)$. The first subscript from the right represents the variable. The second subscript: if it is 1, then it is the lower bound of an interval, and if it is 2, it is the upper bound of an interval. The third subscript represents the brand of the data set: for example, if it is 1 then it represents Brand 1, and if 2 it represents Brand 2. We can write the interval samples as a $(12 \times 1)$ dimensional vector form as $(y_{11\_1}, y_{12\_1}, \ldots, y_{11\_6}, y_{12\_6}, y_{21\_1}, y_{22\_1}, \ldots, y_{21\_6}, y_{22\_6})'$. We then rearrange this vector by grouping together first the six lower bounds of the intervals and then the six upper bounds of the intervals for each of the brands as follows

$$\boldsymbol{y} = (y_{11\_1}, \ldots, y_{11\_6}, y_{12\_1}, \ldots, y_{12\_6}, y_{21\_1}, \ldots, y_{21\_6}, y_{22\_1}, \ldots, y_{22\_6})'.$$

In this article we consider the two bounds of an interval valued variable as two repeated measurements of that variable, and since the values from 1 to 10 are used to evaluate each variable we may assume that the variance covariance matrix of the six variables are the same at the lower bound as well as at the upper bound. If we do not consider brand as separate level, we may assume the Fruit juices data with 16 observations as two-level data set with multivariate equicorrelated covariance structure in deriving its PCs. As mentioned before we may consider two brands as the third level of the data set and use jointly equicorrelated covariance structure in deriving the PCs of the Fruit juice data. In the next two sections we talk about equicorrelated covariance structure and the jointly equicorrelated covariance structure that are suitable for the Fruit Juice data in deriving its principal components.

### 2.2.1 Equicorrelated and jointly equicorrelated covariance structures

We will first consider the Fruit juice data as two-level data set, i.e., we do not consider brand as a separate level like Giordani and Kiers (2006). The $(12 \times 12)-$dimensional equicorrelated covariance structure suitable for the Fruit juice data deeming it as two-level is described as:

$$\begin{aligned} \boldsymbol{\Gamma}_{\boldsymbol{y}}^{(2)} &= \begin{bmatrix} \boldsymbol{U}_0 & \boldsymbol{U}_1 \\ \boldsymbol{U}_1 & \boldsymbol{U}_0 \end{bmatrix} \\ &= \mathbf{I}_2 \otimes (\boldsymbol{U}_0 - \boldsymbol{U}_1) + \mathbf{J}_2 \otimes \boldsymbol{U}_1, \end{aligned} \tag{4}$$

where $\boldsymbol{I}_u$ is the $u \times u$ identity matrix, and $\boldsymbol{J}_u = \mathbf{1}_u \mathbf{1}_u'$ with $\mathbf{1}_u$ is a $u \times 1$ vector of ones. Clearly this data set has six variables and two repeated measurements, and number of samples is 16. Thus, the $6 \times 6-$dimensional blocks $\boldsymbol{U}_0$ in (4) represent the variance-covariance matrix of the six feature variables at the lower as well as at the upper bounds of the intervals, whereas the $6 \times 6-$dimensional off-diagonal blocks $\boldsymbol{U}_1$ in (4) represent the covariance matrix of the six features between the lower and the upper bounds of the intervals. The matrix $\boldsymbol{U}_0$ is positive definite and the matrix $\boldsymbol{U}_1$ is just symmetric. The variance covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(2)}$ is then said to have an equicorrelated covariance structure with equicorrelation parameters $\boldsymbol{U}_0$ and $\boldsymbol{U}_1$. The matrices $\boldsymbol{U}_0$ and $\boldsymbol{U}_1$ are unstructured.

Nevertheless, there are two brands for each juice, and most likely these two brands must have some common factors like the Appearance and smell, and we must take this information too in the modeling of the Fruit juice data. In others words, we may say that these two brands are correlated and we may consider brand as the third level of the data. So, the same Fruit juice data set can be considered as three-level data set with six variables, two bounds and the two brands as the three levels, and consequently, the number of samples reduces to 8. Therefore, the $(24 \times 24)-$dimensional jointly equicorrelated covariance structure suitable for Fruit juice data is described as follows

$$\begin{aligned} \boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)} &= \left[\begin{array}{cc|cc} \boldsymbol{U}_0 & \boldsymbol{U}_1 & \boldsymbol{W} & \boldsymbol{W} \\ \boldsymbol{U}_1 & \boldsymbol{U}_0 & \boldsymbol{W} & \boldsymbol{W} \\ \hline \boldsymbol{W} & \boldsymbol{W} & \boldsymbol{U}_0 & \boldsymbol{U}_1 \\ \boldsymbol{W} & \boldsymbol{W} & \boldsymbol{U}_1 & \boldsymbol{U}_0 \end{array}\right], \\ &= \boldsymbol{I}_4 \otimes (\boldsymbol{U}_0 - \boldsymbol{U}_1) + \boldsymbol{I}_2 \otimes \boldsymbol{J}_2 \otimes (\boldsymbol{U}_1 - \boldsymbol{W}) + \boldsymbol{J}_4 \otimes \boldsymbol{W}, \end{aligned} \tag{5}$$

where $\boldsymbol{U}_0$ is a positive definite symmetric $6 \times 6$ matrix, and $\boldsymbol{U}_1$ and $\boldsymbol{W}$ are symmetric $6 \times 6$ matrices. The variance covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)}$ is then said to have a jointly equicorrelated covariance structure with equicorrelation parameters $\boldsymbol{U}_0, \boldsymbol{U}_1$ and $\boldsymbol{W}$. The matrices $\boldsymbol{U}_0, \boldsymbol{U}_1$ and $\boldsymbol{W}$ are all unstructured.

Thus, the $6 \times 6-$dimensional blocks $\boldsymbol{U}_0$ in (5) represent the variance-covariance matrix of the six features at the lower as well as at the upper bounds of the intervals, whereas the $6 \times 6-$dimensional off-diagonal blocks $\boldsymbol{U}_1$ in (5) represent the covariance matrix of the six features between the lower and the upper bounds of the intervals, and it is same for the two brands. $\boldsymbol{U}_0$ is same for the two bounds and as well for the two brands. The $6 \times 6$ block off diagonals $\boldsymbol{W}$ represent the covariance matrix of the six response variables between any two brands and it is assumed to be the same for any bound (lower or upper) or between the two bounds. If we do not consider brand as a separate level, i.e., there is only one brand, the above jointly equicorrelated matrix (5) reduces to equicorrelated matrix (4).

In this article we derive the principal components in two stages using our new method considering the data as two-level and show that the results are same as that of Giordani and Kiers (2006). We will then derive principal components using our new method considering the data as three-level and show that there is a huge improvement in the result. Not only this, considering the data as three-level gives the information about the difference between the two brands

At the first stage we derive 'eigenblocks' and 'eigenmatrices' of equicorrelated and jointly equicorrelated covariance structures as described in (4) and (5), and the corresponding 'principal vectors'. At the second stage we derive the eigenvalues and eigenvectors of the 'eigenblock' with the components of the principal vectors derived at the first stage as the variables, and then derive the adjusted eigenvalues and the principal components as the linear combination of the components of the first stage principal vectors.

## 3 Eigenblocks and eigenmatrices of equicorrelated and jointly equicorrelated covariance structures

Eigenblocks and eigenmatrices were first introduced by Roy and Fonseca (2012) without properly naming it as eigenblocks and eigenmatrices. They have worked with eigenblocks and eigenmatrices, and therefore worked with a set of independent principal vectors with eigenblocks as their variance-covariance matrices while working on linear models with jointly equicorrelated (doubly exchangeable) distributed errors. In this paper we are interested in seeing the usefulness and interpretations of the independent principal vectors to derive PCs of interval data.

### 3.1 Eigenblocks and eigenmatrices of equicorrelated covariance structure

When we consider the Fruit juice data as two-level we use equicorrelated covariance structure as defined in (4). Let us consider the orthogonal matrix $\boldsymbol{\Gamma}_0 = (\boldsymbol{P}_2' \otimes \boldsymbol{I}_6)$ where

$$\boldsymbol{P}_2' = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}. \tag{6}$$

Clearly $\boldsymbol{P}_2'$ is an orthogonal matrix, so is $\boldsymbol{\Gamma}_0$, and $\boldsymbol{\Gamma}_0$ is not function of either $\boldsymbol{U}_0$ or $\boldsymbol{U}_1$. This $(12\times12)-$dimensional orthogonal matrix $\boldsymbol{\Gamma}_0$ diagonalizes the $(12\times12)-$dimensional equicorrelated matrix $\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(2)}$ such that

$$\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(2)}\boldsymbol{\Gamma}_0' = \text{Diag}\left[\boldsymbol{\Delta}_{2,2}; \boldsymbol{\Delta}_{2,1}\right], \tag{7}$$

where the $6 \times 6$ block diagonals $\boldsymbol{\Delta}_{2,2}$ and $\boldsymbol{\Delta}_{2,1}$ are given by

$$\begin{aligned} \boldsymbol{\Delta}_{2,2} &= \boldsymbol{U}_0 + \boldsymbol{U}_1, \\ \text{and} \quad \boldsymbol{\Delta}_{2,1} &= \boldsymbol{U}_0 - \boldsymbol{U}_1. \end{aligned}$$

See Lemma 3.1 in Roy and Fonseca (2012) for detail. Therefore,

$$\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(2)} = \boldsymbol{\Gamma}_0'\text{Diag}\left[\boldsymbol{\Delta}_{2,2}; \boldsymbol{\Delta}_{2,1}\right]\boldsymbol{\Gamma}_0.$$

Since $\boldsymbol{\Gamma}_0$ is an orthogonal matrix

$$\begin{aligned} \text{tr}\big(\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(2)}\big) &= \text{tr}\big(\boldsymbol{\Gamma}_0'\text{Diag}\left[\boldsymbol{\Delta}_{2,2}; \boldsymbol{\Delta}_{2,1}\right]\boldsymbol{\Gamma}_0\big) \\ &= \text{tr}\big(\text{Diag}\left[\boldsymbol{\Delta}_{2,2}; \boldsymbol{\Delta}_{2,1}\right]\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0'\big) \\ &= \text{tr}\big(\text{Diag}\left[\boldsymbol{\Delta}_{2,2}; \boldsymbol{\Delta}_{2,1}\right]\big) \\ &= \text{tr}\big(\boldsymbol{\Delta}_{2,2}\big) + \text{tr}\big(\boldsymbol{\Delta}_{2,1}\big). \end{aligned}$$

Thus, the total population variance $\text{tr}\big(\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(2)}\big) = \text{tr}\big(\boldsymbol{\Delta}_{2,2}\big) + \text{tr}\big(\boldsymbol{\Delta}_{2,1}\big)$. Therefore, the trace of the variance-covariance matrix of the data is the sum of the traces of the two eigenblocks.

We now partition (horizontal, side by side) the orthogonal matrix $\boldsymbol{\Gamma}_0'$ as two $12 \times 6$ blocks as $\boldsymbol{\Gamma}_0' = [\boldsymbol{E}_{2,1} : \boldsymbol{E}_{2,2}]$. So,

$$\boldsymbol{\Gamma}_0 = \begin{bmatrix} \boldsymbol{E}_{2,1}' \\ \boldsymbol{E}_{2,2}' \end{bmatrix}.$$

Therefore, the eigenblock-eigenmatrix pairs of $\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(2)}$ are $(\boldsymbol{\Delta}_{2,2}, \boldsymbol{E}_{2,1})$ and $(\boldsymbol{\Delta}_{2,1}, \boldsymbol{E}_{2,2})$, where $\text{tr}\boldsymbol{\Delta}_{2,2} \geq \text{tr}(\boldsymbol{\Delta}_{2,1})$. Therefore, the spectral decomposition of the equicorrelated matrix $\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(2)}$ is

$$\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(2)} = \boldsymbol{E}_{2,1}\boldsymbol{\Delta}_{2,2}\boldsymbol{E}_{2,1}' + \boldsymbol{E}_{2,2}\boldsymbol{\Delta}_{2,1}\boldsymbol{E}_{2,2}',$$

where $\boldsymbol{E}_{2,1}$ is the eigenmatrix corresponding to the eigenblock $\boldsymbol{\Delta}_{2,2}$ and $\boldsymbol{E}_{2,2}$ is the eigenmatrix corresponding to the eigenblock $\boldsymbol{\Delta}_{2,1}$. Let

$$\boldsymbol{z} = \boldsymbol{\Gamma}_0 \boldsymbol{y} = \begin{bmatrix} \boldsymbol{E}'_{2,1} \\ \boldsymbol{E}'_{2,2} \end{bmatrix} \boldsymbol{y} = \begin{bmatrix} \boldsymbol{E}'_{2,1}\boldsymbol{y} \\ \boldsymbol{E}'_{2,2}\boldsymbol{y} \end{bmatrix}.$$

Therefore, $\text{Var}(\boldsymbol{z}) = \boldsymbol{\Gamma}_0 \text{Var}(\boldsymbol{y}) \boldsymbol{\Gamma}'_0 = \boldsymbol{\Gamma}_0 \boldsymbol{\Gamma}_{\boldsymbol{y}}^{(2)} \boldsymbol{\Gamma}'_0 = \text{Diag}\left[\boldsymbol{\Delta}_{2,2}; \boldsymbol{\Delta}_{2,1}\right]$ from (7). Thus, $\boldsymbol{E}'_{2,1}\boldsymbol{y}$ and $\boldsymbol{E}'_{2,2}\boldsymbol{y}$ are independent and $\text{Var}(\boldsymbol{E}'_{2,1}\boldsymbol{y}) = \boldsymbol{\Delta}_{2,2}$ and $\text{Var}(\boldsymbol{E}'_{2,2}\boldsymbol{y}) = \boldsymbol{\Delta}_{2,1}$.

Thus, when we consider the Fruit juice data as two-level, the $(12 \times 12)-$dimensional equicorrelated matrix has a total of two $(6 \times 6)-$dimensional eigenblocks: both the eigenblocks $\boldsymbol{\Delta}_{2,2}$ and $\boldsymbol{\Delta}_{2,1}$ are with multiplicity one. Therefore, the two $(6 \times 1)-$dimensional principal vectors for the equicorrelated matrix $\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(2)}$ are $\boldsymbol{E}'_{2,1}\boldsymbol{y}$ and $\boldsymbol{E}'_{2,2}\boldsymbol{y}$, and these two principal vectors are independent. The first principal vector has the variance-covariance matrix $\boldsymbol{\Delta}_{2,2}$ and the second one has the variance-covariance matrix $\boldsymbol{\Delta}_{2,1}$. The interpretation of these two $(6 \times 1)-$dimensional principal vectors is given in the following example.

**Example 1**: For the sake of simplicity, we consider each observation in the data has information on three variables in a two-level interval data set. Thus, a typical sample looks like $\big((y_{11\_1}, y_{12\_1}), (y_{11\_2}, y_{12\_2}), (y_{11\_3}, y_{12\_3})\big)$. We write the interval samples as a vector form $(y_{11\_1}, y_{12\_1}, y_{11\_2}, y_{12\_2}, y_{11\_3}, y_{12\_3})'$. As mentioned before we then rearrange the vector by grouping together first the three lower bounds of the intervals and then the three upper bounds of the intervals as $\boldsymbol{y} = (y_{11\_1}, y_{11\_2}, y_{11\_3}, y_{12\_2}, y_{12\_1}, y_{12\_3})'$. Now, premultiplying $\boldsymbol{y}$ by the orthogonal

matrix $\boldsymbol{\Gamma}_0$ we get

$$\boldsymbol{\Gamma}_0\boldsymbol{y} = \left( \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \otimes \boldsymbol{I}_3 \right) \begin{bmatrix} y_{11\_1} \\ y_{11\_2} \\ y_{11\_3} \\ y_{12\_1} \\ y_{12\_2} \\ y_{12\_3} \end{bmatrix} = \left[\begin{array}{ccc:ccc} \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} \\ \hdashline \frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{\sqrt{2}} \end{array}\right] \begin{bmatrix} y_{11\_1} \\ y_{11\_2} \\ y_{11\_3} \\ y_{12\_1} \\ y_{12\_2} \\ y_{12\_3} \end{bmatrix}$$

$$= \left[\begin{array}{c} (y_{11\_1} + y_{12\_1})/\sqrt{2} \\ (y_{11\_2} + y_{12\_2})/\sqrt{2} \\ (y_{11\_3} + y_{12\_3})/\sqrt{2} \\ \hdashline (y_{11\_1} - y_{12\_1})/\sqrt{2} \\ (y_{11\_2} - y_{12\_2})/\sqrt{2} \\ (y_{11\_3} - y_{12\_3})/\sqrt{2} \end{array}\right].$$

Now, the six rows in the above matrix are the two $(3 \times 1)-$dimensional independent principal vectors of the variance-covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(2)}$. Therefore, the first principal vector is

$$\boldsymbol{y}_{2,1} = \begin{bmatrix} (y_{11\_1} + y_{12\_1})/\sqrt{2} \\ (y_{11\_2} + y_{12\_2})/\sqrt{2} \\ (y_{11\_3} + y_{12\_3})/\sqrt{2} \end{bmatrix},$$

and the second principal vector is

$$\boldsymbol{y}_{2,2} = \begin{bmatrix} (y_{11\_1} - y_{12\_1})/\sqrt{2} \\ (y_{11\_2} - y_{12\_2})/\sqrt{2} \\ (y_{11\_3} - y_{12\_3})/\sqrt{2} \end{bmatrix}.$$

The $(3 \times 1)-$dimensional first principal vector $\boldsymbol{y}_{2,1}$ corresponding to eigenblock $\boldsymbol{\Delta}_{2,2}$ represents the midpoints between the lower bounds and the corresponding upper bounds of the intervals. Similarly, the $(3 \times 1)-$dimensional second principal vector $\boldsymbol{y}_{2,2}$ corresponding to eigenblock $\boldsymbol{\Delta}_{2,1}$ represents the midranges between the lower bounds and the corresponding upper bounds of the intervals.

Since orthonormal eigenvectors and the corresponding percent eigenvalues do not change if we multiply the above principal vectors by some constant, we prefer to work with $\boldsymbol{y}_{2,i}$ instead of $\frac{1}{\sqrt{2}}\boldsymbol{y}_{2,i}$ for $i = 1, 2$, even though these represent the true midpoints and true midranges (Palumbo and Lauro, 2003).

Now, we work independently with these principal vectors and their corresponding variance-covariance matrices, i.e., the corresponding eigenblocks at the second stage to get the eigenvalues

and eigenvectors of $\mathbf{\Gamma}_{\boldsymbol{y}}^{(2)}$. We use *Factor Procedure* of *SAS* (*SAS* Institute Inc., 2012) with Method= Prin Priors=One with Cov and Rotate=Varimax option to get varimax rotated PCs of the components of the first principal vector with variance-covariance matrix $\mathbf{\Delta}_{2,2}$. Similarly, PCs of the components of the second principal vector with variance-covariance matrix $\mathbf{\Delta}_{2,1}$. Even though we get the eigenvalues independently from the two eigenblocks using *Proc Factor*, all the percent eigenvalues and percent cumulative eigenvalues are recalculated as if their total is $\big(\text{tr}(\mathbf{\Delta}_{2,2})+\text{tr}(\mathbf{\Delta}_{2,1})\big)$, which indeed is the total variance of the equicorrelated covariance structure. As mentioned in the Introduction these recalculated percent eigenvalues and percent cumulative eigenvalues are adjusted percent eigenvalues and adjusted percent cumulative eigenvalues.

Therefore, we see that by premultiplying the observation vector by orthogonal matrix we can transform the data to a set of independent principal vectors that represent midpoints and midranges of the interval data. Note that if the lower bounds and the corresponding upper bounds become equal, i.e., if $y_{11\_1} = y_{12\_1}, y_{11\_2} = y_{12\_2}$ and $y_{11\_3} = y_{12\_3}$, the interval data just becomes a $(3 \times 1)-$dimensional multivariate observation and hence we may just consider a $(3 \times 1)-$dimensional multivariate observation $\boldsymbol{y}_{\text{Tra}} = (y_{11\_1}, y_{11\_2}, y_{11\_3})'$ with a covariance matrix $\boldsymbol{U}_0$ and perform the traditional PCA.

On the other hand with $y_{11\_1} = y_{12\_1}, y_{11\_2} = y_{12\_2}$ and $y_{11\_3} = y_{12\_3}$, the first principal vector becomes

$$\boldsymbol{y}_{2,1} = \begin{bmatrix} \sqrt{2}\, y_{11\_1} \\ \sqrt{2}\, y_{11\_2} \\ \sqrt{2}\, y_{11\_3} \end{bmatrix} = \sqrt{2}\, \boldsymbol{y}_{\text{Tra}},$$

with eigenblock $\mathbf{\Delta}_{2,2} = \boldsymbol{U}_0 + \boldsymbol{U}_0 = 2\boldsymbol{U}_0$, as $\boldsymbol{U}_1 = \boldsymbol{U}_0$ in this case. The second principal vector becomes a null vector with null eigenblock. Therefore, the total variance $= \mathbf{\Delta}_{2,2} + \mathbf{\Delta}_{2,1} = 2\boldsymbol{U}_0$. Also, $\text{Var}(\boldsymbol{y}_{2,1}) = \text{Var}(\sqrt{2}\, \boldsymbol{y}_{\text{Tra}}) = 2\text{Var}(\boldsymbol{y}_{\text{Tra}}) = 2\boldsymbol{U}_0 = \mathbf{\Delta}_{2,2}$. Hence, if we perform PCA on $\boldsymbol{y}_{\text{Tra}}$ or on the first principal vector $\boldsymbol{y}_{2,1}$, the percent eigenvalues and the percent adjusted eigenvalues would be the same. Furthermore, the normalized eigenvectors would also be the same in both the cases. Thus, we see that our proposed method of PCA for two-level interval data generalizes the commonly used PCA for multivariate data.

### 3.2 Eigenblocks and eigenmatrices of jointly equicorrelated covariance structure

When we consider the Fruit juice data as three-level we use equicorrelated covariance structure as defined in (5). Let us consider the following two orthogonal matrices

$$\begin{aligned} \boldsymbol{\Gamma}_1 &= \boldsymbol{P}_2' \otimes \boldsymbol{I}_{12} \\ \text{and} \quad \boldsymbol{\Gamma}_2 &= \boldsymbol{I}_2 \otimes \boldsymbol{\Gamma}_0, \end{aligned}$$

where $\boldsymbol{\Gamma}_0$ and $\boldsymbol{P}_2'$ are orthogonal matrices as defined in Section 3.1 with $\boldsymbol{P}_2'$ defined in (6) ($\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ are not function of either $\boldsymbol{U}_0, \boldsymbol{U}_1$ or $\boldsymbol{W}$). Since the product of orthogonal matrices is an orthogonal matrix,

$$\begin{aligned} \boldsymbol{\Gamma}_{21} &= \boldsymbol{\Gamma}_2 \boldsymbol{\Gamma}_1 \\ &= \big(\boldsymbol{I}_2 \otimes (\boldsymbol{P}_2' \otimes \boldsymbol{I}_6)\big)\big(\boldsymbol{P}_2' \otimes \boldsymbol{I}_{12}\big) \\ &= \boldsymbol{P}_2' \otimes \boldsymbol{P}_2' \otimes \boldsymbol{I}_6, \end{aligned}$$

is an orthogonal matrix. In Lemma 3.1 in Roy and Fonseca (2012) it is shown that

$$\boldsymbol{\Gamma}_{21} \boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)} \boldsymbol{\Gamma}_{21}' = \begin{bmatrix} \boldsymbol{\Delta}_{3,3} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Delta}_{3,1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Delta}_{3,2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Delta}_{3,1} \end{bmatrix},$$

where the three $6 \times 6$ eigenblocks $\boldsymbol{\Delta}_{3,3}, \boldsymbol{\Delta}_{3,2}$ and $\boldsymbol{\Delta}_{3,1}$ are given by

$$\begin{aligned} \boldsymbol{\Delta}_{3,3} &= \boldsymbol{U}_0 + \boldsymbol{U}_1 + 2\boldsymbol{W} = (\boldsymbol{U}_0 + \boldsymbol{W}) + (\boldsymbol{U}_1 + \boldsymbol{W}), \\ \boldsymbol{\Delta}_{3,2} &= \boldsymbol{U}_0 + \boldsymbol{U}_1 - 2\boldsymbol{W} = (\boldsymbol{U}_0 - \boldsymbol{W}) + (\boldsymbol{U}_1 - \boldsymbol{W}), \\ \text{and} \quad \boldsymbol{\Delta}_{3,1} &= \boldsymbol{U}_0 - \boldsymbol{U}_1. \end{aligned}$$

Thus, the $(24 \times 24)-$dimensional orthogonal matrix $\boldsymbol{\Gamma}_{21}$ diagonalizes the $(24 \times 24)-$dimensional jointly equicorrelated matrix $\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)}$ such that

$$\boldsymbol{\Gamma}_{21} \boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)} \boldsymbol{\Gamma}_{21}' = \text{Diag}\,[\boldsymbol{\Delta}_{3,3}; \boldsymbol{\Delta}_{3,1}; \boldsymbol{\Delta}_{3,2}; \boldsymbol{\Delta}_{3,1}]. \tag{8}$$

Finally, note that premultiplying and postmultiplying the above equation (8) by the following permutation matrix (Harville, 1997) $\boldsymbol{K}_6$ and its transpose, where

$$\boldsymbol{K}_6 = \begin{bmatrix} \boldsymbol{I}_6 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_6 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_6 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_6 \end{bmatrix},$$

we get

$$\boldsymbol{K}_6\boldsymbol{\Gamma}_{21}\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)}\boldsymbol{\Gamma}_{21}'\boldsymbol{K}_6' = \text{Diag}\left[\boldsymbol{\Delta}_{3,3}; \boldsymbol{\Delta}_{3,2}; \boldsymbol{\Delta}_{3,1}; \boldsymbol{\Delta}_{3,1}\right]. \tag{9}$$

Therefore,

$$\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)} = \boldsymbol{\Gamma}_{21}'\boldsymbol{K}_6'\text{Diag}\big[\boldsymbol{\Delta}_{3,3}; \boldsymbol{\Delta}_{3,2}; \boldsymbol{\Delta}_{3,1}; \boldsymbol{\Delta}_{3,1}\big]\boldsymbol{K}_6\boldsymbol{\Gamma}_{21}.$$

Since the permutation matrix is an orthogonal matrix, $\boldsymbol{K}_6$ and $\boldsymbol{\Gamma}_{21}$ are both orthogonal matrices. Therefore,

$$\begin{aligned}
\text{tr}\big(\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)}\big) &= \text{tr}\big(\boldsymbol{\Gamma}_{21}'\boldsymbol{K}_6'\text{Diag}\big[\boldsymbol{\Delta}_{3,3}; \boldsymbol{\Delta}_{3,2}; \boldsymbol{\Delta}_{3,1}; \boldsymbol{\Delta}_{3,1}\big]\boldsymbol{K}_6\boldsymbol{\Gamma}_{21}\big) \\
&= \text{tr}\big(\text{Diag}\left[\boldsymbol{\Delta}_{3,3}; \boldsymbol{\Delta}_{3,1}; \boldsymbol{\Delta}_{3,2}; \boldsymbol{\Delta}_{3,1}\right]\boldsymbol{K}_6\boldsymbol{\Gamma}_{21}\boldsymbol{\Gamma}_{21}'\boldsymbol{K}_6'\big) \\
&= \text{tr}\big(\text{Diag}\left[\boldsymbol{\Delta}_{3,3}; \boldsymbol{\Delta}_{3,2}; \boldsymbol{\Delta}_{3,1}; \boldsymbol{\Delta}_{3,1}\right]\big) \\
&= \text{tr}\big(\boldsymbol{\Delta}_{3,3}\big) + \text{tr}\big(\boldsymbol{\Delta}_{3,2}\big) + 2\text{tr}\big(\boldsymbol{\Delta}_{3,1}\big).
\end{aligned}$$

Thus, the total population variance $\text{tr}\big(\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)}\big) = \text{tr}\big(\boldsymbol{\Delta}_{3,3}\big) + \text{tr}\big(\boldsymbol{\Delta}_{3,2}\big) + 2\text{tr}\big(\boldsymbol{\Delta}_{3,1}\big)$. Therefore, the trace of the variance-covariance matrix of the data is the sum of traces of its eigenblocks.

We now partition (horizontal, side by side) the orthogonal matrix $\boldsymbol{\Gamma}_{21}'\boldsymbol{K}_6'$ as four $24 \times 6$ blocks $\boldsymbol{\Gamma}_{21}'\boldsymbol{K}_6' = [\boldsymbol{E}_{3,1} : \boldsymbol{E}_{3,2} : \boldsymbol{E}_{3,3} : \boldsymbol{E}_{3,4}]$. So,

$$\boldsymbol{K}_6\boldsymbol{\Gamma}_{21} = \begin{bmatrix} \boldsymbol{E}_{3,1}' \\ \boldsymbol{E}_{3,2}' \\ \boldsymbol{E}_{3,3}' \\ \boldsymbol{E}_{3,4}' \end{bmatrix}.$$

Therefore, the eigenblock-eigenmatrix pairs of $\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)}$ are $(\boldsymbol{\Delta}_{3,3}, \boldsymbol{E}_{3,1})$, $(\boldsymbol{\Delta}_{3,2}, \boldsymbol{E}_{3,2})$, $(\boldsymbol{\Delta}_{3,1}, \boldsymbol{E}_{3,3})$ and $(\boldsymbol{\Delta}_{3,1}, \boldsymbol{E}_{3,4})$, where $\text{tr}(\boldsymbol{\Delta}_{3,3}) \geq \text{tr}\boldsymbol{\Delta}_{3,2} \geq \text{tr}(\boldsymbol{\Delta}_{3,1})$. Therefore, the spectral decomposition of equicorrelated matrix $\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)}$ is

$$\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)} = \boldsymbol{E}_{3,1}\boldsymbol{\Delta}_{3,3}\boldsymbol{E}_{3,1}' + \boldsymbol{E}_{3,2}\boldsymbol{\Delta}_{3,2}\boldsymbol{E}_{3,2}' + \boldsymbol{E}_{3,3}\boldsymbol{\Delta}_{3,1}\boldsymbol{E}_{3,3}' + \boldsymbol{E}_{3,4}\boldsymbol{\Delta}_{3,1}\boldsymbol{E}_{3,4}',$$

where $\boldsymbol{E}_{3,1}$ is the eigenmatrix corresponding to eigenblock $\boldsymbol{\Delta}_{3,3}$, $\boldsymbol{E}_{3,2}$ is the eigenmatrix corresponding to eigenblock $\boldsymbol{\Delta}_{3,2}$ and $\boldsymbol{E}_{3,3}$ and $\boldsymbol{E}_{3,4}$ are the two eigenmatrices corresponding to the eigenblock $\boldsymbol{\Delta}_{3,1}$. Let

$$\boldsymbol{z} = \boldsymbol{K}_6\boldsymbol{\Gamma}_{21}\boldsymbol{y} = \begin{bmatrix} \boldsymbol{E}_{3,1}' \\ \boldsymbol{E}_{3,2}' \\ \boldsymbol{E}_{3,3}' \\ \boldsymbol{E}_{3,4}' \end{bmatrix}\boldsymbol{y} = \begin{bmatrix} \boldsymbol{E}_{3,1}'\boldsymbol{y} \\ \boldsymbol{E}_{3,2}'\boldsymbol{y} \\ \boldsymbol{E}_{3,3}'\boldsymbol{y} \\ \boldsymbol{E}_{3,4}'\boldsymbol{y} \end{bmatrix}.$$

Therefore, $\text{Var}(\boldsymbol{z}) = \boldsymbol{K}_6\boldsymbol{\Gamma}_{21}\text{Var}(\boldsymbol{y})\boldsymbol{\Gamma}_{21}'\boldsymbol{K}_6' = \boldsymbol{K}_6\boldsymbol{\Gamma}_{21}\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)}\boldsymbol{\Gamma}_{21}'\boldsymbol{K}_6' = \text{Diag}\Big[\boldsymbol{\Delta}_{3,3}; \boldsymbol{\Delta}_{3,2}; \boldsymbol{\Delta}_{3,1}; \boldsymbol{\Delta}_{3,1}\Big]$ from (9). Thus, $\boldsymbol{E}_{3,1}'\boldsymbol{y}$, $\boldsymbol{E}_{3,2}'\boldsymbol{y}$, $\boldsymbol{E}_{3,3}'\boldsymbol{y}$ and $\boldsymbol{E}_{3,4}'\boldsymbol{y}$ are independent and $\text{Var}(\boldsymbol{E}_{3,1}'\boldsymbol{y}) = \boldsymbol{\Delta}_{3,3}$, $\text{Var}(\boldsymbol{E}_{3,2}'\boldsymbol{y}) = \boldsymbol{\Delta}_{3,2}$, and $\text{Var}(\boldsymbol{E}_{3,3}'\boldsymbol{y}) = \text{Var}(\boldsymbol{E}_{3,4}'\boldsymbol{y}) = \boldsymbol{\Delta}_{3,1}$.

When we consider the Fruit juice data as three-level, the $(24 \times 24)-$ dimensional jointly equicorrelated matrix has a total of four eigenblocks: the eigenblocks $\boldsymbol{\Delta}_{3,3}$ and $\boldsymbol{\Delta}_{3,2}$ are with multiplicity one, and the eigenblock $\boldsymbol{\Delta}_{3,1}$ is with multiplicity two. Therefore, the four $(6 \times 1)-$dimensional independent principal vectors for the jointly equicorrelated covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)}$ are $\boldsymbol{E}_{3,1}'\boldsymbol{y}$, $\boldsymbol{E}_{3,2}'\boldsymbol{y}$, $\boldsymbol{E}_{3,3}'\boldsymbol{y}$ and $\boldsymbol{E}_{3,4}'\boldsymbol{y}$. The interpretation of these four $(6 \times 1)-$dimensional principal vectors is given in the following example.

**Example 2**: As before for the sake of simplicity, let us again consider that each observation unit has information on three variables in a three-level interval data set. Thus, a typical sample looks like $\big((y_{11\_1}, y_{12\_1}), (y_{11\_2}, y_{12\_2}), (y_{11\_3}, y_{12\_3}), (y_{21\_1}, y_{22\_1}), (y_{21\_2}, y_{22\_2}), (y_{21\_3}, y_{22\_3})\big)$. As before, we rearrange the vector by grouping together first the three lower bounds of the intervals and then the three upper bounds of the intervals at the first brand and then the same at the second brand as $\boldsymbol{y} = (y_{11\_1}, y_{11\_2}, y_{11\_3}, y_{12\_2}, y_{12\_1}, y_{12\_3}, y_{21\_1}, y_{21\_2}, y_{21\_3}, y_{22\_2}, y_{22\_1}, y_{22\_3})'$. Thus, premultiplying $\boldsymbol{y}$ by $\boldsymbol{\Gamma}_{21} = \boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_1$ we get all four $(3 \times 1)-$dimensional principal vectors in one vector. We first premultiply $\boldsymbol{y}$ by $\boldsymbol{\Gamma}_1$, and then premultiply the result by $\boldsymbol{\Gamma}_2$. Now,

$$\boldsymbol{\Gamma}_1\boldsymbol{y} = \left(\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \otimes \boldsymbol{I}_6\right) \begin{bmatrix} y_{11\_1} \\ y_{11\_2} \\ y_{11\_3} \\ y_{12\_1} \\ y_{12\_2} \\ y_{12\_3} \\ y_{21\_1} \\ y_{21\_2} \\ y_{21\_3} \\ y_{22\_1} \\ y_{22\_2} \\ y_{22\_3} \end{bmatrix} = \begin{bmatrix} (y_{11\_1} + y_{21\_1})/\sqrt{2} \\ (y_{11\_2} + y_{21\_2})/\sqrt{2} \\ (y_{11\_3} + y_{21\_3})/\sqrt{2} \\ (y_{12\_1} + y_{22\_1})/\sqrt{2} \\ (y_{12\_2} + y_{22\_2})/\sqrt{2} \\ (y_{12\_3} + y_{22\_3})/\sqrt{2} \\ (y_{11\_1} - y_{21\_1})/\sqrt{2} \\ (y_{11\_2} - y_{21\_2})/\sqrt{2} \\ (y_{11\_3} - y_{21\_3})/\sqrt{2} \\ (y_{12\_1} - y_{22\_1})/\sqrt{2} \\ (y_{12\_2} - y_{22\_2})/\sqrt{2} \\ (y_{12\_3} - y_{22\_3})/\sqrt{2} \end{bmatrix}.$$

Therefore,

$$\boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_1\boldsymbol{y} = \left(\boldsymbol{I}_2 \otimes \left(\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \otimes \boldsymbol{I}_3\right)\right) \begin{bmatrix} (y_{11\_1}+y_{21\_1})/\sqrt{2} \\ (y_{11\_2}+y_{21\_2})/\sqrt{2} \\ (y_{11\_3}+y_{21\_3})/\sqrt{2} \\ (y_{12\_1}+y_{22\_1})/\sqrt{2} \\ (y_{12\_2}+y_{22\_2})/\sqrt{2} \\ (y_{12\_3}+y_{22\_3})/\sqrt{2} \\ (y_{11\_1}-y_{21\_1})/\sqrt{2} \\ (y_{11\_2}-y_{21\_2})/\sqrt{2} \\ (y_{11\_3}-y_{21\_3})/\sqrt{2} \\ (y_{12\_1}-y_{22\_1})/\sqrt{2} \\ (y_{12\_2}-y_{22\_2})/\sqrt{2} \\ (y_{12\_3}-y_{22\_3})/\sqrt{2} \end{bmatrix}$$

$$= \left[\begin{array}{cccccc|cccccc} \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{\sqrt{2}} \end{array}\right] \begin{bmatrix} (y_{11\_1}+y_{21\_1})/\sqrt{2} \\ (y_{11\_2}+y_{21\_2})/\sqrt{2} \\ (y_{11\_3}+y_{21\_3})/\sqrt{2} \\ (y_{12\_1}+y_{22\_1})/\sqrt{2} \\ (y_{12\_2}+y_{22\_2})/\sqrt{2} \\ (y_{12\_3}+y_{22\_3})/\sqrt{2} \\ (y_{11\_1}-y_{21\_1})/\sqrt{2} \\ (y_{11\_2}-y_{21\_2})/\sqrt{2} \\ (y_{11\_3}-y_{21\_3})/\sqrt{2} \\ (y_{12\_1}-y_{22\_1})/\sqrt{2} \\ (y_{12\_2}-y_{22\_2})/\sqrt{2} \\ (y_{12\_3}-y_{22\_3})/\sqrt{2} \end{bmatrix}$$

$$= \left[\begin{array}{c} ((y_{11\_1}+y_{21\_1})+(y_{12\_1}+y_{22\_1}))/2 \\ ((y_{11\_2}+y_{21\_2})+(y_{12\_2}+y_{22\_2}))/2 \\ ((y_{11\_3}+y_{21\_3})+(y_{12\_3}+y_{22\_3}))/2 \\ \hline ((y_{11\_1}+y_{21\_1})-(y_{12\_1}+y_{22\_1}))/2 \\ ((y_{11\_2}+y_{21\_2})-(y_{12\_2}+y_{22\_2}))/2 \\ ((y_{11\_3}+y_{21\_3})-(y_{12\_3}+y_{22\_3}))/2 \\ \hline ((y_{11\_1}-y_{21\_1})+(y_{12\_1}-y_{22\_1}))/2 \\ ((y_{11\_2}-y_{21\_2})+(y_{12\_2}-y_{22\_2}))/2 \\ ((y_{11\_3}-y_{21\_3})+(y_{12\_3}-y_{22\_3}))/2 \\ \hline ((y_{11\_1}-y_{21\_1})-(y_{12\_1}-y_{22\_1}))/2 \\ ((y_{11\_2}-y_{21\_2})-(y_{12\_2}-y_{22\_2}))/2 \\ ((y_{11\_3}-y_{21\_3})-(y_{12\_3}-y_{22\_3}))/2 \end{array}\right].$$

As a result,

$$\boldsymbol{K}_3\boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_1\boldsymbol{y} = \begin{bmatrix} \boldsymbol{I}_3 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_3 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_3 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_3 \end{bmatrix} \begin{bmatrix} ((y_{11\_1}+y_{21\_1})+(y_{12\_1}+y_{22\_1}))/2 \\ ((y_{11\_2}+y_{21\_2})+(y_{12\_2}+y_{22\_2}))/2 \\ ((y_{11\_3}+y_{21\_3})+(y_{12\_3}+y_{22\_3}))/2 \\ \hline ((y_{11\_1}+y_{21\_1})-(y_{12\_1}+y_{22\_1}))/2 \\ ((y_{11\_2}+y_{21\_2})-(y_{12\_2}+y_{22\_2}))/2 \\ ((y_{11\_3}+y_{21\_3})-(y_{12\_3}+y_{22\_3}))/2 \\ \hline ((y_{11\_1}-y_{21\_1})+(y_{12\_1}-y_{22\_1}))/2 \\ ((y_{11\_2}-y_{21\_2})+(y_{12\_2}-y_{22\_2}))/2 \\ ((y_{11\_3}-y_{21\_3})+(y_{12\_3}-y_{22\_3}))/2 \\ \hline ((y_{11\_1}-y_{21\_1})-(y_{12\_1}-y_{22\_1}))/2 \\ ((y_{11\_2}-y_{21\_2})-(y_{12\_2}-y_{22\_2}))/2 \\ ((y_{11\_3}-y_{21\_3})-(y_{12\_3}-y_{22\_3}))/2 \end{bmatrix}$$

$$= \begin{bmatrix} ((y_{11\_1}+y_{21\_1})+(y_{12\_1}+y_{22\_1}))/2 \\ ((y_{11\_2}+y_{21\_2})+(y_{12\_2}+y_{22\_2}))/2 \\ ((y_{11\_3}+y_{21\_3})+(y_{12\_3}+y_{22\_3}))/2 \\ \hline ((y_{11\_1}-y_{21\_1})+(y_{12\_1}-y_{22\_1}))/2 \\ ((y_{11\_2}-y_{21\_2})+(y_{12\_2}-y_{22\_2}))/2 \\ ((y_{11\_3}-y_{21\_3})+(y_{12\_3}-y_{22\_3}))/2 \\ \hline ((y_{11\_1}+y_{21\_1})-(y_{12\_1}+y_{22\_1}))/2 \\ ((y_{11\_2}+y_{21\_2})-(y_{12\_2}+y_{22\_2}))/2 \\ ((y_{11\_3}+y_{21\_3})-(y_{12\_3}+y_{22\_3}))/2 \\ \hline ((y_{11\_1}-y_{21\_1})-(y_{12\_1}-y_{22\_1}))/2 \\ ((y_{11\_2}-y_{21\_2})-(y_{12\_2}-y_{22\_2}))/2 \\ ((y_{11\_3}-y_{21\_3})-(y_{12\_3}-y_{22\_3}))/2 \end{bmatrix}.$$

Now, the 12 rows in the above matrix are the four $(3 \times 1)-$dimensional principal vectors of the variance-covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)}$, and they are as follows:

$$\boldsymbol{y}_{3,1} = \begin{bmatrix} ((y_{11\_1}+y_{21\_1})+(y_{12\_1}+y_{22\_1}))/2 \\ ((y_{11\_2}+y_{21\_2})+(y_{12\_2}+y_{22\_2}))/2 \\ ((y_{11\_3}+y_{21\_3})+(y_{12\_3}+y_{22\_3}))/2 \end{bmatrix},$$

$$\boldsymbol{y}_{3,2} = \begin{bmatrix} ((y_{11\_1}-y_{21\_1})+(y_{12\_1}-y_{22\_1}))/2 \\ ((y_{11\_2}-y_{21\_2})+(y_{12\_2}-y_{22\_2}))/2 \\ ((y_{11\_3}-y_{21\_3})+(y_{12\_3}-y_{22\_3}))/2 \end{bmatrix},$$

$$\boldsymbol{y}_{3,3} = \begin{bmatrix} ((y_{11\_1}+y_{21\_1}-(y_{12\_1}+y_{22\_1}))/2 \\ ((y_{11\_2}+y_{21\_2}-(y_{12\_2}+y_{22\_2}))/2 \\ ((y_{11\_3}+y_{21\_3}-(y_{12\_3}+y_{22\_3}))/2 \end{bmatrix},$$

and

$$\boldsymbol{y}_{3,4} = \begin{bmatrix} ((y_{11\_1}-y_{21\_1})-(y_{12\_1}-y_{22\_1}))/2 \\ ((y_{11\_2}-y_{21\_2})-(y_{12\_2}-y_{22\_2}))/2 \\ ((y_{11\_3}-y_{21\_3})-(y_{12\_3}-y_{22\_3}))/2 \end{bmatrix}.$$

The first principal vector $\boldsymbol{y}_{3,1}$ corresponding to eigenblock $\boldsymbol{\Delta}_{3,3}$ represents the total grand midpoints of the feature variables. The second principal vector $\boldsymbol{y}_{3,2}$ corresponding to eigenblock $\boldsymbol{\Delta}_{3,2}$ represents the difference between the two brands of fruit juices. For example $(y_{11\_1} - y_{21\_1})$ provides the difference between the first brand and the second brand of the first variable at the lower bound of an interval. And, $(y_{12\_1} - y_{22\_1})$ provides the difference between the first brand and the second brand of the first variable at the upper bound of the same interval. So, $\big((y_{11\_1} - y_{21\_1}) + (y_{12\_1} - y_{22\_1})\big)/2$ represents the average difference between the first brand and the second brand of the first feature variable. Similarly, for the other two components of the second principal vector. The third and the fourth principal vectors correspond to the same eigenblocks $\boldsymbol{\Delta}_{3,1}$. And, these two principal vectors are independent. The average of these two blocks, which represents the grand midranges has the variance-covariance matrix $\boldsymbol{\Delta}_{3,1}$.

Now, we work independently with these principal vectors and their corresponding variance-covariance matrices, i.e., the corresponding eigenblocks at the second stage to get the eigenvalues and eigenvectors of $\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)}$. Even though we get the eigenvalues independently from the eigenblocks, all the percent eigenvalues and percent cumulative eigenvalues are recalculated as if their total is $\mathrm{tr}\big((\boldsymbol{\Delta}_{3,3}) + \mathrm{tr}(\boldsymbol{\Delta}_{3,2}) + 2\mathrm{tr}(\boldsymbol{\Delta}_{3,1})\big)$, which indeed is the total variance of the jointly equicorrelated covariance structure. As before, we call these recalculated percent eigenvalues and percent cumulative eigenvalues as adjusted percent eigenvalues and adjusted percent cumulative eigenvalues.

Now, we have seen that if the data have equicorrelated covariance structure or jointly equicorrelated covariance structure, premultiplying and postmultiplying it by some orthogonal matries yield blockdiagonal matrix with eigenblocks of size $(6 \times 6)$ and the first principal vector corresponding to the first eigenblock represents to the grand midpoints of the interval data.

Therefore, we see that by premultiplying three-level observation vector by an orthogonal matrix we can transform the data to a set of independent principal vectors that represent grand midpoints, brand difference and grand midranges. In general for multi-level interval data we can premultiply the data vector by some suitable multiple orthogonal matrices and get the representation of grand midpoints and midranges.

Note that if there is only one brand, following the same way as in Example 1, one can show that the four principal vectors of three-level interval data reduces to two principal vectors of two-level interval data as in Example 1. Furthermore, in Example 1 we have shown that our proposed method for PCA for two-level interval data generalizes the commonly used PCA for multivariate data. Thus, we can say that our suggested method extends the traditional PCA not only to the two-level interval data, but also to three-level interval data.

Now, for the Fruit juice data the population variance-covariance matrix is unknown, thus in the next section we derive an unbiased estimate of jointly equicorrelated covariance matrix considering the data are three-level.

## 4 Unbiased estimate of jointly equicorrelated covariance matrix

Let $\boldsymbol{Q}_p = \boldsymbol{I}_p - \boldsymbol{J}_p$ be an orthogonal projector matrix of (any) order $p$. Let $\boldsymbol{Y} = (\boldsymbol{Y}_{1,1}, \boldsymbol{Y}_{1,2}, \boldsymbol{Y}_{2,1}, \boldsymbol{Y}_{2,2})$ be the data matrix from $N_{4p}(\boldsymbol{\mu}, \boldsymbol{\Gamma}_{\boldsymbol{y}}^{(3)})$. In the component matrices in $\boldsymbol{Y}$, the first subscript from the right represents the two bounds of an interval, the second subscript from the right represents the two brands. Therefore, the unbiased estimate of the unstructured variance-covariance matrix $\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{Y}\boldsymbol{Q}_n\boldsymbol{Y}'$ (see Mardia et al., 1979). To obtain unbiased estimates of $\boldsymbol{U}_0, \boldsymbol{U}_1$ and $\boldsymbol{W}$, we expand $\boldsymbol{S}$ blockwise as

$$\begin{aligned}
\boldsymbol{S} &= \frac{1}{n-1}\boldsymbol{Y}'\boldsymbol{Q}_n\boldsymbol{Y} = \frac{1}{n-1}\left(\boldsymbol{Y}_{1,1}, \boldsymbol{Y}_{1,2}, \boldsymbol{Y}_{2,1}, \boldsymbol{Y}_{2,2}\right)'\boldsymbol{Q}_n\left(\boldsymbol{Y}_{1,1}, \boldsymbol{Y}_{1,2}, \boldsymbol{Y}_{2,1}, \boldsymbol{Y}_{2,2}\right) \\
&= \frac{1}{n-1}\begin{pmatrix}
\boldsymbol{Y}'_{1,1}\boldsymbol{Q}_n\boldsymbol{Y}_{1,1} & \boldsymbol{Y}'_{1,1}\boldsymbol{Q}_n\boldsymbol{Y}_{1,2} & \boldsymbol{Y}'_{1,1}\boldsymbol{Q}_n\boldsymbol{Y}_{2,1} & \boldsymbol{Y}'_{1,1}\boldsymbol{Q}_n\boldsymbol{Y}_{2,2} \\
\boldsymbol{Y}'_{1,2}\boldsymbol{Q}_n\boldsymbol{Y}_{1,1} & \boldsymbol{Y}'_{1,2}\boldsymbol{Q}_n\boldsymbol{Y}_{1,2} & \boldsymbol{Y}'_{1,2}\boldsymbol{Q}_n\boldsymbol{Y}_{2,1} & \boldsymbol{Y}'_{1,2}\boldsymbol{Q}_n\boldsymbol{Y}_{2,2} \\
\boldsymbol{Y}'_{2,1}\boldsymbol{Q}_n\boldsymbol{Y}_{1,1} & \boldsymbol{Y}'_{2,1}\boldsymbol{Q}_n\boldsymbol{Y}_{1,2} & \boldsymbol{Y}'_{2,1}\boldsymbol{Q}_n\boldsymbol{Y}_{2,1} & \boldsymbol{Y}'_{2,1}\boldsymbol{Q}_n\boldsymbol{Y}_{2,2} \\
\boldsymbol{Y}'_{2,2}\boldsymbol{Q}_n\boldsymbol{Y}_{1,1} & \boldsymbol{Y}'_{2,2}\boldsymbol{Q}_n\boldsymbol{Y}_{1,2} & \boldsymbol{Y}'_{2,2}\boldsymbol{Q}_n\boldsymbol{Y}_{2,1} & \boldsymbol{Y}'_{2,2}\boldsymbol{Q}_n\boldsymbol{Y}_{2,2}
\end{pmatrix}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\widehat{\boldsymbol{U}}_0 &= \frac{1}{4(n-1)}\left(\sum_{i=1}^{2}\sum_{j=1}^{2}\boldsymbol{Y}'_{i,j}\boldsymbol{Q}_n\boldsymbol{Y}_{i,j}\right), \\
\widehat{\boldsymbol{U}}_1 &= \frac{1}{4(n-1)}\left(\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{\substack{j'=1 \\ j\neq j'}}^{2}\boldsymbol{Y}'_{i,j}\boldsymbol{Q}_n\boldsymbol{Y}_{i,j'}\right),
\end{aligned}$$

$$\text{and} \quad \widehat{\boldsymbol{W}} = \frac{1}{8(n-1)}\left(\sum_{i=1}^{2}\sum_{\substack{i'=1 \\ i\neq i'}}^{2}\sum_{j=1}^{2}\boldsymbol{Y}'_{i,j}\boldsymbol{Q}_n\boldsymbol{Y}_{i',j} + \sum_{i=1}^{2}\sum_{\substack{j=1 \\ i\neq j}}^{2}\boldsymbol{Y}'_{i,j}\boldsymbol{Q}_n\boldsymbol{Y}_{j,i} + \sum_{i=1}^{2}\sum_{\substack{j=1 \\ i\neq j}}^{2}\boldsymbol{Y}'_{i,i}\boldsymbol{Q}_n\boldsymbol{Y}_{j,j}\right).$$

Therefore, an unbiased estimator of $\boldsymbol{\Gamma}$ is

$$\widehat{\boldsymbol{\Gamma}} = \boldsymbol{I}_4 \otimes \left(\widehat{\boldsymbol{U}}_0 - \widehat{\boldsymbol{U}}_1\right) + \boldsymbol{I}_2 \otimes \boldsymbol{J}_2 \otimes \left(\widehat{\boldsymbol{U}}_1 - \widehat{\boldsymbol{W}}\right) + \boldsymbol{J}_4 \otimes \widehat{\boldsymbol{W}}.$$

Also,

$$\begin{aligned}
\widehat{\boldsymbol{\Delta}}_{3,3} &= \widehat{\boldsymbol{U}}_0 + \widehat{\boldsymbol{U}}_1 + 2\widehat{\boldsymbol{W}} = (\widehat{\boldsymbol{U}}_0 + \widehat{\boldsymbol{W}}) + (\widehat{\boldsymbol{U}}_1 + \widehat{\boldsymbol{W}}), \\
\widehat{\boldsymbol{\Delta}}_{3,2} &= \widehat{\boldsymbol{U}}_0 + \widehat{\boldsymbol{U}}_1 - 2\widehat{\boldsymbol{W}} = (\widehat{\boldsymbol{U}}_0 - \widehat{\boldsymbol{W}}) + (\widehat{\boldsymbol{U}}_1 - \widehat{\boldsymbol{W}}), \\
\text{and} \quad \widehat{\boldsymbol{\Delta}}_{3,1} &= \widehat{\boldsymbol{U}}_0 - \widehat{\boldsymbol{U}}_1,
\end{aligned}$$

are unbiased estimators of $\boldsymbol{\Delta}_{3,3}$, $\boldsymbol{\Delta}_{3,2}$ and $\boldsymbol{\Delta}_{3,1}$, respectively.

### 4.1 Unbiased estimate of equicorrelated covariance matrix

Unbiased estimators of equicorrelated covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{y}}^{(2)}$ can be deduced from $\boldsymbol{S}$ by considering $\boldsymbol{Y}$ for only one brand, i.e., $\boldsymbol{Y} = (\boldsymbol{Y}_{1,1}, \boldsymbol{Y}_{1,2})$. Therefore

$$\begin{aligned} \widehat{\boldsymbol{U}}_0 &= \frac{1}{2(n-1)} \left( \sum_{j=1}^{2} \boldsymbol{Y}'_{1,j} \boldsymbol{Q}_n \boldsymbol{Y}_{1,j} \right), \\ \text{and} \quad \widehat{\boldsymbol{U}}_1 &= \frac{1}{2(n-1)} \left( \sum_{j=1}^{2} \sum_{\substack{j'=1 \\ j \neq j'}}^{2} \boldsymbol{Y}'_{1,j} \boldsymbol{Q}_n \boldsymbol{Y}_{1,j'} \right). \end{aligned}$$

Thus,

$$\begin{aligned} \widehat{\boldsymbol{\Delta}}_{2,2} &= \widehat{\boldsymbol{U}}_0 + \widehat{\boldsymbol{U}}_1, \\ \text{and} \quad \widehat{\boldsymbol{\Delta}}_{2,1} &= \widehat{\boldsymbol{U}}_0 - \widehat{\boldsymbol{U}}_1, \end{aligned}$$

are unbiased estimators of $\boldsymbol{\Delta}_{2,2}$ and $\boldsymbol{\Delta}_{2,1}$, respectively.

## 5 Results

In this section we illustrate our proposed method to the Fruit juices data as described in the Introduction. We work on this data (Support data) in their original interval structure $[x_L^-, x_R^+]$ as shown in Table 2. We consider these two measurements as two repeated measures for each of the six variables. We as well work on the interval structure $[y^-, y^+]$, where $y^- = m - s, y^+ = m + s$ and call this data as Core data (Giordan and Kiers, 2006), and consider these two measurements as the two repeated measures for each of the six variables. We will analyze the data set first considering it as two-level and then considering it as three-level, and the results are presented in the following sections.

### 5.1 Results considering the Fruit juice data as two-level

The PCA of Fruit juice data considering it as two-level is performed in this section. The trace(eigenblocks), the percent(%) and the percent(%) cumulative trace(eigenblocks) at the first stage PCA are presented in Table 3. We see the first eigenblock $\boldsymbol{\Delta}_{2,2}$ accounts for 99.7021% of the total variation of the Core data and 98.8744% for the Support data. So, for the next stage PCA we will only consider the components of first principal vector as the variables corresponding

to the first eigenblock $\mathbf{\Delta}_{2,2}$, which is the variance-covariance matrix of the first principal vector for both Core and Support data. Results of second stage PCA of both the Core and Support data are presented in Table 4. Both percent and the cumulative percent of the total variance due to each principal component are given in Table 4. For comparison perpose we also have given

Table 2: Fruit juices interval Support data

| Fruit juices | Appearance | Smell | Taste | Naturalness | Sweetness | Density |
|---|---|---|---|---|---|---|
| Apple1 | [ 6.78,7.52 ] | [ 5.47,6.59 ] | [ 7.40,8.40 ] | [ 5.66,7.20 ] | [ 7.27,8.29 ] | [ 5.81,6.74 ] |
| Apple2 | [ 6.60,7.72 ] | [ 6.28,7.40 ] | [ 6.31,7.43 ] | [ 5.72,7.12 ] | [ 6.67,7.65 ] | [ 5.47,6.59 ] |
| Apricot1 | [ 6.82,7.68 ] | [ 7.87,8.68 ] | [ 7.60,8.54 ] | [ 7.35,8.47 ] | [ 7.42,8.40 ] | [ 7.03,8.15 ] |
| Apricot2 | [ 7.32,8.16 ] | [ 7.09,8.19 ] | [ 5.17,6.71 ] | [ 4.66,6.06 ] | [ 4.90,6.31 ] | [ 5.79,6.77 ] |
| Banana1 | [ 4.96,6.37 ] | [ 3.92,5.60 ] | [ 3.64,5.32 ] | [ 4.27,5.95 ] | [ 4.76,6.16 ] | [ 3.62,4.74 ] |
| Banana2 | [ 5.27,6.67 ] | [ 3.68,5.36 ] | [ 3.26,4.94 ] | [ 3.92,5.46 ] | [ 4.23,5.91 ] | [ 3.65,4.77 ] |
| Grapefruit1 | [ 6.28,7.40 ] | [ 6.52,7.65 ] | [ 5.17,6.85 ] | [ 6.00,7.33 ] | [ 2.45,3.39 ] | [ 3.64,4.76 ] |
| Grapefruit2 | [ 6.31,7.43 ] | [ 5.63,6.75 ] | [ 6.35,7.47 ] | [ 6.11,7.23 ] | [ 4.14,5.19 ] | [ 3.06,4.46 ] |
| Orange1 | [ 6.64,7.59 ] | [ 7.12,8.24 ] | [ 6.39,7.44 ] | [ 5.67,6.72 ] | [ 5.75,6.67 ] | [ 3.64,4.97 ] |
| Orange2 | [ 6.89,7.55 ] | [ 6.06,6.90 ] | [ 6.82,7.94 ] | [ 5.60,6.72 ] | [ 5.93,7.13 ] | [ 3.88,4.98 ] |
| Peach1 | [ 7.09,7.93 ] | [ 6.94,7.78 ] | [ 6.42,7.54 ] | [ 5.70,7.10 ] | [ 6.69,7.68 ] | [ 5.03,5.92 ] |
| Peach2 | [ 6.98,7.82 ] | [ 6.22,7.11 ] | [ 7.38,8.38 ] | [ 6.83,7.72 ] | [ 6.83,7.81 ] | [ 4.99,5.85 ] |
| Peer1 | [ 6.89,7.76 ] | [ 7.19,8.24 ] | [ 7.14,8.19 ] | [ 6.44,7.49 ] | [ 7.59,8.54 ] | [ 7.22,8.27 ] |
| Peer2 | [ 7.52,8.20 ] | [ 6.32,7.44 ] | [ 7.69,8.57 ] | [ 6.72,7.63 ] | [ 7.71,8.62 ] | [ 6.72,7.67 ] |
| Pineapple1 | [ 6.61,7.66 ] | [ 5.74,6.66 ] | [ 6.18,7.31 ] | [ 5.45,6.85 ] | [ 5.63,6.75 ] | [ 3.92,5.00 ] |
| Pineapple2 | [ 6.66,7.59 ] | [ 5.90,7.30 ] | [ 5.65,6.98 ] | [ 5.23,6.56 ] | [ 5.52,6.92 ] | [ 3.28,4.69 ] |

C-PCA, V-PCA and neural networks PCA (NN-PCA) from Giordan and Kiers (2006, Table 5) in Table 4. We suspect that there is a misprint on Component 4 in the C-PCA column in their paper; it should be 97.76 instead of 98.76,. We use *Proc Factor* of *SAS* with Method= Prin Priors=One with Cov and Rotate=Varimax option to get the varimax rotated PCs of the components of the first principal vector with variance-covariance matrix $\mathbf{\Delta}_{2,2}$. Note that the total eigenvalue for both

Table 3: Trace(eigenblocks), percent(%) trace(eigenblocks) and percent(%) cumulative trace(eigenblocks) of Fruit juice data considering it as two-level

| Eigen | Core Data | | | Support Data | | |
|---|---|---|---|---|---|---|
| Block | Trace(Eigenblock) | % | % Cum. | Trace(Eigenblock) | % | % Cum. |
| $\mathbf{\Delta}_{2,2}$ | 13.55237 | 99.70209 | 99.70209 | 14.63206 | 98.87437 | 98.87437 |
| $\mathbf{\Delta}_{2,1}$ | 0.04049 | 0.29791 | 100.00 | 0.16658 | 1.12563 | 100.00 |
| Total | 13.59286 | 100.00 | | 14.79864 | 100.00 | |

Core data and Support data are the respective trace of $\mathbf{\Delta}_{2,2}$ as shown in Table 3. So, *Proc Factor* calculates the percent eigenvalues and cumulative percent eigenvalues based on the total variance 13.55237, which is the trace of $\mathbf{\Delta}_{2,2}$ in the Core data. Note that percent cumulative of our method and C-PCA $\big($calculated by Giordan and Kiers (2006)$\big)$ method are the exact same. Nevertheless,

the total variance is 13.59286, which is the sum of the traces of two eigenblocks in Table 3. So, we must calculate the adjusted percentage eigenvalue as Eigenvalue/13.59286 instead of Eigenvalue/13.55237. Therefore, in Core data the adjusted percentage of total variance accounted for by the first two principal components is $100(9.44793+2.14455)/13.59286 = 85.28359\%$. So, it is actually 85.28% not the apparent 85.54% as shown in Table 4; also found the same by by Giordan and Kiers (2006). Similarly, for the Support Data the adjusted percentage of total variance accounted for by the first two principal components is $100(10.28341 + 2.19060)/14.79864 = 84.29159\%$. So,

Table 4: Eigenvalues, percent(%) eigenvalues and percent(%) cumulative eigenvalues as output of *Proc Factor* of Fruit juice data considering it as Two-level

| No. of comp. | Core Data | | | Support Data | | | Core Data | | |
|---|---|---|---|---|---|---|---|---|---|
| | Midpoint, $\mathbf{\Delta}_{2,2}$ | | | Midpoint, $\mathbf{\Delta}_{2,2}$ | | | C-PCA | V-PCA | NN-PCA |
| ($p$) | Eigenvalue | % | % Cum. | Eigenvalue | % | % Cum. | % Cum. | % Cum. | % Cum. |
| 1 | 9.44793 | 69.71 | 69.71 | 10.28341 | 70.28 | 70.28 | 69.71 | 58.15 | 71.29 |
| 2 | 2.14455 | 15.82 | 85.54 | 2.19060 | 14.97 | 85.25 | 85.54 | 73.58 | 85.34 |
| 3 | 1.14163 | 8.42 | 93.96 | 1.30346 | 8.91 | 94.16 | 93.96 | 84.29 | 91.91 |
| 4 | 0.51418 | 3.79 | 97.76 | 0.56904 | 3.89 | 98.05 | 98.76 | 91.25 | 97.25 |
| 5 | 0.25193 | 1.86 | 99.62 | 0.23386 | 1.60 | 99.65 | 99.62 | 96.43 | 99.04 |
| 6 | 0.05215 | 0.39 | 100.00 | 0.05170 | 0.35 | 100.00 | 100.00 | 100.00 | 100.00 |
| Total | 13.55237 | 100.00 | | 14.63207 | 100.00 | | | | |

we see that Core data accounts for little larger variance. Therefore, we only consider the component loading matrices of the first two PCs for the Core data and they are given in Table 7. The components are varimax rotated as we have used Rotate=Varimax option in *Proc Factor* of *SAS*. We see that these component loadings are different from the reported varimax rotated component loadings in Giordan and Kiers (2006). However, the interpretation of the principal components are the exact same. We see that the first PC especially related to Smell, Taste and Naturalness, and the second PC refers to the features Sweetness and Density. The first two PCs which truly account for 85.28% of the variance does not even refer Appearance as an important feature. This is somehow not expected as all the judges evaluated Appearance and Smell first before tasting and the remaining characteristics later, even though Giordan and Kiers (2006) commented Appearance is less important than the other attributes, and fruit juice can be very nice but its appearance may be unpleasant. Since all the judges evaluated the attributes Appearance and Smell first, we strongly believe that these two features, Appearance and Smell, bring together the first impression of the fruit juices. Anyway, Appearance seems to be not an important attribute as Smell is, according to our analysis. Nonetheless, it does not seem to be right somehow. In the following section we analyze the Fruit juice data as three-level data and see whether Appearance

emerges as an important feature as it should be.

### 5.2 Results considering the Fruit juice data as three-level

The PCA of Fruit juice data considering it as three-level is conducted in this section. The trace(eigenblocks), the percent and the percent cumulative trace(eigenblocks) at the first stage PCA are presented in Table 5. We see the first eigenblock $\mathbf{\Delta}_{3,3}$ accounts for 86.5471% of the total

Table 5: Trace(eigenblocks), percent(%) trace(eigenblocks) and percent(%) cumulative trace(eigenblocks) of Fruit juice data considering it as three-level

| Eigen | Core Data | | | Support Data | | |
|---|---|---|---|---|---|---|
| Block | Trace(Eigenblock) | % | % Cum. | Trace(Eigenblock) | % | % Cum. |
| $\mathbf{\Delta}_{3,3}$ | 24.81989 | 86.54707 | 86.54707 | 26.72301 | 85.67737 | 85.67737 |
| $\mathbf{\Delta}_{3,2}$ | 3.77795 | 13.17375 | 99.72082 | 4.12342 | 13.22022 | 98.89759 |
| $\mathbf{\Delta}_{3,1}$ | 2 ×0.04003 | 0.27917 | 100.00 | 2 ×0.17192 | 1.10240 | 100.00 |
| Total | 28.67790 | 100.00 | | 31.19027 | 100.00 | |

variation and the second eigenblock $\mathbf{\Delta}_{3,2}$ accounts for 13.1738% of the total variation of the Core data, while for the Support data first eigenblock $\mathbf{\Delta}_{3,3}$ accounts for 85.6774% and the second eigenblock $\mathbf{\Delta}_{3,2}$ accounts for 13.2202% of the total variation. In other words, the first two eigenblocks account for 99.7208% of the total variance for the Core data, while the first two eigenblocks account for 98.8976% of the total variance for the Support data. So, for the second

Table 6: Eigenvalues, percent(%) eigenvalues and percent(%) cumulative eigenvalues as output of *Proc Factor* of Fruit juice data considering it as three-level

| No. of | Core Data | | | | | | Support Data | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| comp. | Brand Difference, $\mathbf{\Delta}_{3,2}$ | | | Midpoint, $\mathbf{\Delta}_{3,3}$ | | | Brand Difference, $\mathbf{\Delta}_{3,2}$ | | | Midpoint, $\mathbf{\Delta}_{3,3}$ | | |
| ($p$) | Eigen Value | % | % Cum. | Eigen Value | % | % Cum. | Eigen Value | % | % Cum. | Eigen Value | % | % Cum. |
| 1 | 2.82965 | 74.90 | 74.90 | 18.37705 | 74.04 | 74.04 | 3.16187 | 76.68 | 76.68 | 19.86196 | 74.33 | 74.33 |
| 2 | 0.63535 | 16.82 | 91.72 | 4.33446 | 17.46 | 91.51 | 0.60070 | 14.57 | 91.25 | 4.44492 | 16.63 | 90.96 |
| 3 | 0.19595 | 5.19 | 96.90 | 1.63450 | 6.59 | 98.09 | 0.25411 | 6.16 | 97.41 | 1.83128 | 6.85 | 97.81 |
| 4 | 0.08500 | 2.25 | 99.15 | 0.39946 | 1.61 | 99.70 | 0.07423 | 1.80 | 99.21 | 0.49909 | 1.87 | 99.68 |
| 5 | 0.02838 | 0.75 | 99.90 | 0.05788 | 0.23 | 99.93 | 0.02747 | 0.67 | 99.88 | 0.06954 | 0.26 | 99.94 |
| 6 | 0.00363 | 0.10 | 100.00 | 0.01654 | 0.07 | 100.00 | 0.00504 | 0.12 | 100.00 | 0.01622 | 0.06 | 100.00 |
| Total | 3.77795 | | | 24.81988 | | | 4.12342 | | | 26.72301 | | |

stage PCA we analyze the first two principal vectors with the variance-covariance matrices $\mathbf{\Delta}_{3,3}$ and $\mathbf{\Delta}_{3,2}$ respectively for both Core and Support data as discussed in Example 2. The first principal vector represents the grand midpoints and the second principal vector represents the difference between the two brands. As before, we use *Proc Factor* of *SAS* with Method= Prin,

Priors=One, Cov and Rotate=Varimax option to get the varimax rotated PCs. Eigenvalues, percent eigenvalues and percent cumulative eigenvalues for both Core and Support data are given in Table 6. Note that *Proc Factor* calculates the percent eigenvalue and percent cumulative eigenvalue based on the total variance 24.8199 and 3.7780, which are the traces of $\mathbf{\Delta}_{3,3}$ and $\mathbf{\Delta}_{3,2}$ for the Core data as shown in Table 5. *Proc Factor* also calculates the same for the Support data. Nevertheless, the total variance is 28.6779 for the Core data, which is the sum of the traces of all three eigenblocks in Table 5. So, we should calculate the adjusted percent eigenvalue as Eigenvalue/28.6779 instead of Eigenvalue/24.81989 and Eigenvalue/3.77795 for midpoints and Brabd difference. Therefore, in Core data the adjusted percentage of total variance accounted for by the first two principal components in midpoint is 100(18.37705 + 4.33446)/28.677904= 79.195153%, not the apparent 91.51% as shown in Table 6. Now, if we calculate the adjusted percentage eigenvalues we see that

Table 7: Component loading matrices for Two-level as well as Three-level of the Fruit juice data

| | Tow-level | | Three-level | | |
|---|---|---|---|---|---|
| | Midpoint, $\mathbf{\Delta}_{2,2}$ | | Midpoint, $\mathbf{\Delta}_{3,3}$ | | Brand Difference, $\mathbf{\Delta}_{3,2}$ |
| | PC1 | PC2 | PC1 | PC2 | PC3 |
| Appearance | 0.6878 | 0.2707 | **0.8692** | 0.4176 | -0.4936 |
| Smell | **0.8503** | -0.0464 | **0.9309** | 0.0984 | 0.0480 |
| Taste | **0.8155** | 0.5498 | **0.8712** | 0.4446 | **0.6167** |
| Naturalness | **0.8564** | 0.3033 | **0.9254** | 0.1395 | **0.8758** |
| Sweetness | 0.2220 | **0.9055** | 0.2332 | **0.9085** | 0.3870 |
| Density | 0.2960 | **0.4911** | 0.3374 | **0.4738** | **0.7410** |

we can take only one PC in brand difference along with two PCs in Midpoint. Therefore, in Core data the first three principal components explain a percentage

$$100(18.37705 + 4.33446 + 2.82965)/28.677904 = 89.062157\%$$

of the total population variance. And, in Support data the first three principal components explain a percentage

$$100(19.86196 + 4.44492 + 3.16187)/31.19027 = 88.06833\%$$

of the total population variance. Since Core data accounts for larger variance we only discuss component loadings for PCs of Core data and the results are given in Table 7. We see this time the first PC clearly refers Appearance, Smell, Taste and Naturalness all together. As expected Appearance and Smell have come together as important attributes in the first PC. The second PC is same as before, like it refers to Sweetness and Density. The third PC which represents the

brand difference refers to Taste, Naturalness and Density. So, the attributes Taste, Naturalness and Density are vital in differentiating the brands. Note that Appearance and Smell do not play any role in differentiating the brands; as Appearance and smell are same for both the brands of a particular fruit juice, and thus these two attributes cannot differentiate the brands. Results seem promising when we treat Fruit juice data as three-level, since it gives all the relevant information of the data as anticipated and guessed.

## 6 Concluding Remarks

Due to recent development of cheaper and more manageable way to store large amounts of digital data, big data are registered continuously in almost all scientific fields. Appropriate statistical methods need to be developed that is suitable for multi-level interval data. In this article we have proposed a new approach to modeling and analyzing three-level interval data analytically and systematically and deriving principal components in two stages.

In this article we have discussed why one should not just analyze the midpoint variables and just get percent and percent cumulative eigenvalues for only midpoint variables, rather should get adjusted percent and adjusted percent cumulative eigenvalues. Our approach can be easily extended to more than three levels, i.e., to multi-level or multi-dimensional interval data where each component variable is bounded within hyper-rectangles. We are working on this and report it in a future correspondence.

## References


[1] Billard L. and Diday, E. (2006) Symbolic Data Analysis: Conceptual Statistics and Data Mining, John Wiley & Sons Ltd. Chichester, West Sussex, England.

[2] Bock, H. H. and Diday, E. (2000) Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data, Springer, Heidelberg.

[3] Cazes, P., Chouakria, A., Diday, E., and Schektman, Y. (1997) Extensions de l'Analyse en Composantes Principales a des Donnees de Type Intervalle, Revue de Statistique Appliquee, 24, 5-24.

[4] Chouakria, A. (1998) Extension des Methodes d'Analyse Factorielle a des Donees de Type Intervalle, unpublished doctoral thesis, Universit6 Paris Dauphine.

[5] Chouakria, A., Diday, E., and Cazes, P. (1999) An Improved Factorial Representation of Symbolic Objects, In Knowledge Extraction From Statistical Data, Luxembourg: European Commission Eurostat, 301-305.

[6] Lauro C. and Palumbo, F. (2000) Principal component analysis of interval data: a symbolic data analysis approach. Computational Statistics. 15. 73-87.

[7] Giordani, P. and Kiers, H. A. L. (2006) A comparison of three methods for principal component analysis of fuzzy interval data. Computational Statatistics and Data Analysis, 51:379397.

[8] Harville, D. A. (1997) *Matrix algebra from a Statistician's perspective*, Springer-Verlag NY.

[9] Leiva, R. and Roy, A. (2011) Linear Discrimination for Multi-level Multivariate Data with Separable Means and Jointly Equicorrelated Covariance Structure. Journal of Statistical Planning and Inference, 141(5), 1910-1924.

[10] Mardia, K. V., Kent, J. T., Bibby, J. M. (1979) Multivariate Analysis. New York: Academic Press Inc.

[11] Palumbo, F., Lauro, C., (2003) A PCA for interval-valued data based on midpoints and radii. In: Yanai, H., Okada, A., Shigemasu, K., Kano, Y., Meulman, J. (Eds.), New Developments in Psychometrics. Springer, Tokyo, pp. 641648.

[12] Roy, A. (2014a) Principal Component Analyses of Symbolic Data using Patterned Covariance Structures, The Paper Presented at the Multivariate models 2014, Bedlewo, Poland, March 24-28.

[13] Roy, A. (2014b) Two-stage Principal Component Analyses of Symbolic Data using Patterned Covariance Structures, The Paper Presented at the Joint Statistical Meetings 2014, Boston, August 2-7.

[14] Roy, A. and Fonseca M. (2012) Linear Models with Doubly Exchangeable Distributed Errors, Communications in Statistics - Theory and Methods, 41(13), 2545-2569.

[15] Roy A. and Leiva R. (2011) Estimating and Testing a Structured Covariance Matrix for Three-level Multivariate Data, Communications in Statistics - Theory and Methods, 40(11), 1945-1963.

[16] Roy A. and Leiva R. (2007) Discrimination with Jointly Equicorrelated Multi-level Multivariate Data, Advances in Data Analysis and Classification, 1(3), 175-199.

[17] SAS Institute Inc, (2012) SAS/STAT Users Guide Version 9.3, SAS Institute Inc., Cary, NC.