Global Journals LaTeX JournalKaleidoscope™

Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

CrossRef DOI of original article:

The Quantile Method for Symbolic Principal Component Analysis

Received: 1 January 1970 Accepted: 1 January 1970 Published: 1 January 1970

Abstract

Index terms—

1 I. INTRODUCTION

The generalization of the principal component analysis (PCA) is an important research theme in the symbolic data analysis [1][2][3][4]. The main purpose of the traditional PCA is to transform a number of possibly correlated variables into a small number of uncorrelated variables called principal components. Chouakria [5] proposed the extension of the PCA to interval data as vertices principal component analysis (V-PCA). Chouakria et al. [6] proposed also the centers method of PCA (C-PCA) for interval data, and they presented a comparative example for the V-PCA and the C-PCA. Lauro and Palumbo [7] proposed symbolic object principal component analysis (SO-PCA) as an extended PCA to any numerical data structure. Lauro et al. [8] summarize various methods of SO-PCA for interval data. The author also proposed a general "Symbolic PCA" (S-PCA) based on the quantification method by using the generalized Minkowski metrics [9,10]. In this approach, we first transform the given symbolic data table to a usual numerical data table, and then we execute the traditional PCA on the transformed data table.

In this article, another quantification method for symbolic data tables based on the monotone structures of objects is presented. In Section 2, first we describe the case of point sequences in a d-dimensional Euclidean space. The monotone structures are characterized by the nesting of the Cartesian join regions associated with pairs of objects. If the given point sequence is monotone in the Euclidean d space, the property is also satisfied in any feature axis. In other words, a nesting structure of the given point sequence in the d space confines the orders of points in each feature axis to be similar. Therefore, we can evaluate the degree of similarity between features based on the Kendall or the Spearman's rank correlation coefficients. Then, we can execute a traditional PCA based on the correlation matrix by the selected rank correlation coefficient. Secondly, we describe the "object splitting method" for SO-PCA for interval-valued data [11]. This method splits each of N symbolic objects described by d interval-valued features into the two d-dimensional vertices called the "minimum sub-object" and the "maximum sub object". We should point out the fact that any interval object can be reproduced from the minimum and the maximum sub-objects. Moreover, the nesting structure of interval objects in the d space confines the orders of the minimum and the maximum sub-objects in each feature axis to be similar. Therefore, we can evaluate again the degree of similarity between features based on the Kendall or the Spearman's rank correlation coefficients on the $(2 \times N) \times d$ standard numerical data table. We can execute a traditional PCA based on the correlation matrix by the selected rank correlation coefficient. As a further extension to manipulate histogram data, nominal multi-valued data, and others, we describe the "quantile method" for S-PCA [12] in Section 4.

The problem is how to obtain a common numerical rep resentation of objects described by mixed types of features. For example, in histogram data, the numbers of subinter vals (bins) of the given histograms are mutually different in general. Therefore, we first define the cumulative dis tribution function for each histogram. Then, we select a common integer number m to generate the "quantiles" for all histograms. As the result, for each histogram, we have an (m + 1)-tuple composed of (m -1) quantiles and the minimum and the maximum values of the whole interval of the histogram. Then, we split each object into (m + 1) sub-objects: the minimum sub-object, (m -1) quantile sub objects and the maximum sub-object. By virtue of the monotonic property of the distribution function, (m + 1) sub-objects of an object satisfy automatically a nesting structure. Therefore, the nesting of N objects described by the minimum and the maximum sub-objects in the d space confines the orders of N × (m + 1) sub-objects in each feature axis to be similar. Again, we can evaluate the degree of similarity between features by the Kendall or the Spearman's rank correlation coefficient, and then execute a traditional PCA.

Interval-valued data may be regarded as a special histogram-valued data, where only one bin organizes the histogram. Furthermore, we can also split nominal multi-valued data into (m + 1) sub-objects based on the distribution function associated with rank values attached to categorical values of an object. Therefore, by the quantile method we can transform a given general N × d symbolic data table to an {N × (m + 1)} × d standard numerical data table, and then we can execute a traditional PCA on the transformed data table. In Section 5, we describe several experimental results in order to show the effectiveness of the quantile method. Section 6 is a summary.

2 II. MONOTONE STRUCTURES AND OBJECT SPLITTING METHOD

In this section, we describe some properties of monotone structures for point sequence and for interval objects. Then, we describe the object splitting method for S-PCA.

3 Monotone Structures for Point Sequence

Let a set of N objects U be represented by U = {? 1 , ? 2 , ... , ? N }. Let each object ? i be described by d numerical features, i.e. a vectorx i = (x i1 , x i2 , ... , x id) in a d-dimensional Euclidean space R d .

DEFINITION 1: Rectangular region spanned by x i and x j .

Let J (? i , ? j) be a rectangular region in R d spanned by the vectors x i and x j , and be defined by the following Cartesian product of d closed intervals.J (? i , ? j) = [min(x i1 , x j 1), max(x i1 , x j 1)] × [min(x i2 , x j 2), max(x i2 , x j 2)] ×???× [min(x id , x jd), max(x id , x jd)],**(1)**

where min(a, b) and max(a, b) are the operators to take the minimum value and the maximum value from a and b, respectively. London Journal of Research in Science: Natural and Formal

In the following, we call J (? i , ? j) as the Cartesian join (region) of objects ? i and ? j [9,10, **??**3].

DEFINITION 2: Nesting structure If a series of objects ? 1 , ? 2 , ... , ? N satisifies the nesting propertyJ (? 1 , ? k) ? J (? 1 , ? k+1), k = 1, 2,...,N -1, (**2**)

the series is called a "nesting structure with the starting point ? 1 and the ending point ? N ".

In Fig. **??**, (a) is a monotone increasing series, and (b) is a monotone decreasing series of objects. It should be noted that the two series of objects show the same nesting structures with starting point ? 1 and ending point ? 5 .

PROPOSITION 1: If a series of objects ? 1 , ? 2 , ... , ? N is a nesting structure with the starting point ? 1 and the ending point ? N in the space R d , the series satisfies the same structure in each feature (axis) of the space R d .

Proof: From the definition of rectangular region as in Eq. (1), we have 15J (? 1 , ? k) = [min(x 11 , x k1), max(x 11 , x k1)] × [min(x 12 , x k2), max(x 12 , x k2)] ×???× [min(x 1d , x kd), max(x 1d , x kd)],**(3)**

Therefore, the relations of the Cartesian join regions J (? 1 , ? k) ? J (? 1 , ? k+1), k = 1, 2,... , N -1, in Definition 2, require the following relations for each feature, i.e. for each j (= 1, 2,... , d),

[min(x 1j , x kj), max(x 1j , x kj)] ? [min(x 1j , x k+1,j), max(x 1j , x k+1,j], k = 1, 2,...,N -1.

(5)

Although, there exist several ways to define the mono tone sequences of objects, i.e. monotone structures, we use the following definition.

DEFINITION 3: Monotone structure of a series of points.

A series of objects ? 1 , ? 2 , ... , ? N is called a monotone structure, if the series satisfies the nesting structure of Definition 2.

Since, for a pair of features, we can evaluate the degree of similarity between two sets of orders of objects for the same object set U by using the Kendall or the Spearman's rank correlation coefficient, we have Proposition 2.

PROPOSITION 2: Correlation matrix S .

If a series of objects ? 1 , ? 2 , ... , ? N is a monotone structure in the space R d , the absolute value of each off diagonal element of the d × d correlation matrix S takes the maximum value one in the sense of the Kendall or the Spearman's rank correlation coefficient.

Proof: From Definition 3, any monotone structure must satisfy the nesting property of Definition 2. Then, from Proposition 1, the given series of objects has the identical nesting structure for each feature. This property exactly restricts the order of objects for each feature to be the same way or the reverse way according to the series of objects is monotone increasing or monotone decreasing. Therefore, if a series of objects is a monotone structure in R d , the absolute value of the correlation coefficient for each pair of features takes the maximum value one in the sense of the Kendall or the Spearman's rank correlation coefficient.

From Proposition 2, if many off-diagonal elements of S take highly correlated values, we can expect the existence of a large eigenvalue of S , and that the corresponding eigenvector reproduces well the original nesting property of the set of objects in the space R d .

EXAMPLE 1: As an intuitive example, suppose that the given set of objects in R d organizes an approximate monotone structure which is monotone increasing along each of d features, and the degrees of similarity between two features are the same for all possible pairs. Therefore, all off-diagonal elements of S take an identical value ?, 0 <?< 1. Then, it is known [14] that d eigenvalues of S become? 1 = 1 + (d -1)? and ? 2 = ? 3 =???= ? d = 0,**(6)**

and the eigenvector for ? 1 isa 1 = (1/ ? d, 1/ ? d,... , 1/ ? d). (**7**)

Therefore, the given monotone structure of objects in R d is approximately reproduced around the eigenvector a 1 . As a particular case, when ? = 1, the given set of objects organizes a complete monotone structure in the space R d . Then, the eigenvalue ? 1 becomes d, i.e. its contribution ratio is 100%, and the order of the given object sequence in the space R d is exactly reproduced on the eigenvector a 1 .

4 London Journal of Research in Science: Natural and Formal

In the above, we characterized monotone structures by the nesting property, and obtain the correlation matrix S . The monotone structures include any linear structure as a special case. On the other hand, a monotone structure may be approximated well by an appropriately selected linear structure. This suggests that we can use also the Pearson correlation coefficient to evaluate the degree of similarity between two features instead of the Kendall and the Spearman's rank correlation coefficients.

5 Monotone Structures for Interval Objects

Let each object be described by d interval-valued features. Then, an object ? k ? U becomes a hyper rectangle in R d , i.e. the Cartesian product of d closed intervals:I k = I k1 × I k2 ×???× I kd , (**8**)

where each interval I kp is given byI kp = [x kp(min) , x kp(max)], p = 1, 2,...,d.**(9)**

Then, we can define the minimum vertex x k(min) and the maximum vertex x k(max) by

x k(min) = (x k1(min) , x k2(min) ,...,x kd(min)) and x k(max) = (x k1(max) , x k2(max) ,...,x kd(max)). ()**10**

DEFINITION 4: The minimum sub-object and the max imum sub-object Let the minimum vertex x k(min) and the maximum vertex x k(max) for each object ? k ? U be called the minimum sub object and the maximum sub-object, and be denoted by ? k(min) and ? k(max) , respectively.

6 EXAMPLE 2:

In Table **??**, the minimum and the maxi mum sub-objects of Linseed oil under the first four inter val features are represented by the vertices x Linseed(min) = (0.930, -27, 170, 118) and x Linseed(max) = (0.935, -18, 204, 196), respectively.

PROPOSITION 3: From Definition 1, any interval object ? k ? U is represented in the space R d by the Cartesian join region J (? k(min) , ? k(max)).

Proof: From Eq. (1) in Definition 1 and (8-10), we see thatJ (? k(min) , ? k(max)) = [x k1(min) , x k1(max)] × [x k2(min) , x k2(max)] × ... × [x kd(min) , x kd(max)] = I k1 × I k2 ×???× I kd = I k .

From Eq. (8), d respective intervals for ? i and ? j areI ip = [x ip(min) , x ip(max)], p = 1, 2,...

7 ,d, and

I jp = [x jp(min) , x jp(max)], p = 1, 2,...,d.**(11)**

Thus the closed interval I ijp generated from two intervals I ip and I jp becomesI ijp = [min(x ip(min) , x jp(min)), max(x ip(max) , x jp(max))], p = 1, 2,... ,d. (**12**)

DEFINITION 5: We define the Cartesian join region J (? i , ? j) based on Eq. (12) byJ (? i , ? j) = I ij 1 × I ij 2 ×???× I ijd = [min(x i1(min) , x j 1(min)), max(x i1(max) , x j 1(max))] × [min(x i2(min) , x j 2(min)), max(x i2(max) , x j 2(max))] × ... × [min(x id(min) , x jd(min)), max(x id(max) , x jd(max))]. (**13**)

In this definition, we should note that, for each k,J (? k , ? k) is equivalent to J (? k(min) , ? k(max)). Furthermore,

Table **??**: Fats' and oils' data [10].

8 J (? k(min)

, ? k(min)) and J (? k(max) , ? k(max)) are reduced to the minimum vertex x k(min) and the maximum vertex x k(max) in Eq. (10), respectively.

DEFINITION 6: Nesting structure for interval objects If a series of interval objects ? 1 , ? 2 , ... , ? N satisifies the nesting propertyJ (? 1 , ? k) ? J (? 1 , ? k+1), k = 1, 2,...,N -1, (**14**)

the series is called a "nesting structure with the starting object ? 1 and the ending object ? N ".

Fig. **??** shows a series of five interval objects. It should be noted that the nesting order of objects in each feature axis is the same as that in the two-dimensional space.

Object Specific gravity (g/cm 3), F 1 Freezing point (×???× [min(x 1d(min) , x k+1,d(min)), max(x 1d(max) , x k+1,d(max))].**(16)**

Therefore, the relations of the Cartesian join regionsJ (? 1 , ? k) ? J (? 1 , ? k+1), k = 1, 2,... , N-1, in Definition 5

, require the following relations for each feature, i.e. for each j (= 1, 2, ... , d),[min(x 1j(min) , x kj(min)), max(x 1j(max) , x kj(max))] ? [min(x 1j(min) , x k+1,j(min)), max(x 1j(max) , x k+1,j(max))], k = 1, 2,...,N -1. (**17**)

We define the monotone structure of interval objects by the same way in Definition 3. A series of interval objects ? 1 , ? 2 , ... , ? N is called a monotone structure, if the series satisfies a nesting structure in Definition 6.

According to Definition 7, we assume a series of interval objects ? 1 , ? 2 , ... , ? N is a monotone structure in the space R d . Then, from Proposition 4, the series of objects satisfies the same nesting in each feature axis. However, the nesting in (17) is based on the closed intervals generated from two objects. Therefore, we cannot evaluate the degree of similarity between two features by direct use of the Kendall or the Spearman's rank correlation coefficient. To remove this difficulty, we split each interval object into the minimum sub-object and the maximum sub-object.

PROPOSITION 5: Monotone conditions by sub-objects. Let a series of interval objects ? 1 , ? 2 , ... , ? N be monotone in the space R d . Then, at least one condition of the following must be satisfied.

(1) The series of the minimum sub-objects, ? 1(min) , ? 2(min) , ... , ? N(min) , is monotone in R d .

(2) The series of the maximum sub-objects, ? 1(max) , ? 2(max) , ... , ? N(max) , is monotone in R d .

Proof: Assume that the conditions (1) and (2) are negated simultaneously. Then, there exists a nesting order k in which the object ? k satisfies the nesting property in R d but the corresponding minimum sub-object ? k(min) and the maximum sub-object ? k(max) breaks the nesting property in R d , simultaneously. This contradicts the fact given in Proposition 3. . On the other hand, if the series of objects satisfies only one condition, we call the series of objects as weakly monotone in R d . Fig. **??** shows a case of a strongly monotone struc ture, whereas Fig. 3 illustrates a case of a weakly mono tone structure.

If a series of interval objects ? 1 , ? 2 , ... , ? N in the space R d is given, we can obtain the d × d correlation matrix S by splitting each object into the minimum and the maximum sub-objects and by using the Kendall or the Spearman's rank correlation coefficient. PROPOSITION 6: Property of correlation matrix S by the object splitting.

(1) If the given series of objects is strongly monotone in a pair of features, the corresponding correlation coefficient shows a strictly high score for 2N sub objects by the object splitting.

(2) If the given series of interval objects is weakly monotone, the correlation coefficient shows a degraded score compared to the case (1).

9 R d

and/or the series of the maximum sub-objects in R d also become monotone. Therefore, we have the properties (1) and (2) whether the given series of objects is strongly monotone or weakly monotone.

In the above, we characterized monotone structures of N interval objects in the space R d by the nesting property of 2N sub-objects in R d , i.e. the minimum sub object and the maximum sub-object, and obtained the correlation matrix S based on the Kendall or Spearman's rank correlation coefficient. As noted in the preceding, the monotone structures include any linear structure as a special case. On the other hand, a monotone structure may be approximated well by an appropriately selected linear structure. Therefore, we can use also the Pearson correlation coefficient to evaluate the degree of similarity between two features instead of the Kendall and Spearman's rank correlation coefficients.

10 The Object Splitting Method for SO-PCA

PROCEDURE 1: Object splitting method for SO-PCA. For a set of N objects ? 1 , ? 2 , ... , ? N under d interval valued features, the object splitting method is executed by the following steps.

1. We split each object ? k into the minimum sub object ? k(min) and the maximum sub-object ? k(max) .

As the result, we have a (2N) × d numerical data table. 2. We calculate the d × d correlation matrix S for the (2N) × d data table obtained in (1) based on the selected correlation coefficient, where we can use the Kendall or Spearman's rank correlation coefficient or the Pearson correlation coefficient. 3. We find the principal components based on the correlation matrix in (2). 4. We represent each symbolic object ? k in the factor planes as the arrow line connecting from ? k(min) to ? k(max) , or as the Cartesian join of ? k(min) and ? k(max) , i.e. a rectangular region spanned by ? k(min) and ? k(max) .

EXAMPLE 3: Fats' and oils' data (interval-valued data).

We applied the object splitting method to the Fats' and oils' data of Table **??**. We used only four interval features. The contribution ratios of the first two principal components understanding for the descriptions of symbolic objects in the factor planes compared to the rectangular representation. London Journal of Research in Science: Natural and Formal Chouakria et al. [6] presented a comparative study of the vertices method (V-PCA) and the centers method (C-PCA). The V-PCA is implemented on the numerical data table of the size (N × **??**

d) × d, while the C-PCA is implemented on the size N × d. Therefore, the C-PCA is stronger than the V-PCA in the computational complexity, when the number of descriptive features is large. The contribution ratios of the first two principal components for the fats' and oils' data of Table **??** are 68.29% and 20.23% by the V-PCA, and 75.23% and 15.09% by the C-PCA, respectively. The rectangular representations of objects for these two methods are similar, although their contribution ratios are different. Moreover, their results are also close to the arrow line representations in Figs **??** and **??**.

Lauro et al. [8] presented a comparative study of the V-PCA, the method called spaghetti PCA, and the method based on interval algebra and optimization theory. For the Fats' and oils' data of Table **??**, their results of rectangular representations in the first factor planes are mutually similar. Among them, the spaghetti PCA is especially close to the result in Figs **??** and **??**. The spaghetti PCA uses the main diagonals of the hyper-rectangles to represent multidimensional interval data. The contribution ratios of the first two principal components are 71.33% and 18.09%. In the representation of interval objects in the first factor plane, the lengths and the directions of the main diagonals of the rectangular regions are very similar to those of the arrow lines in Figs **??** and **??**. The spaghetti PCA is a very different method from the object splitting method. However, we should point out the fact that the main diagonal of an object may be described by two end points: the minimum vertex and the maximum vertex.

In this section, we presented the object splitting method of PCA for interval objects. This method transforms the given N × d interval-valued data table into a 2N × d standard numerical data table, then executes the PCA on the transformed data table. We should note that 1. The object splitting method is simple and works as well as other methods for interval objects.

Especially, this method is easily applicable to large data tables. 2. The arrow line representation of objects in the factor planes is useful to provide insights about the mutual relationships of the given interval objects.

In the next section, we present the quantile method, which is an extension of the object splitting method and can manipulate not only interval-valued features but also other type features including histogram features and nominal multi-valued features.

11 III. COMMON REPRESENTATION BY QUANTILES

In the aggregation process of large data sets, the use of histograms is very natural and common to describe the reduced data sets. Billard and Diday [2,4] summarize empirical distribution functions and descriptive statistics for various feature types. Based on knowledge of distribution functions, the quantile method [12] provides a common framework to represent symbolic data described by features of different types. The basic idea is to express the observed feature values by some predefined quantiles of the underlying distribution. In the interval feature case, a distribution is assumed within each interval, e.g., uniform distribution (Bertrand and Goupil [15]). For a histogram feature, quantiles of any histogram may be obtained simply by interpolation, assuming the uniformity in each bin of the histogram [2,4,15]. Although the numbers of bins of the given histograms are mutually different in general, we can obtain the same number of quantiles for each histogram. For nominal multi-valued features, quantiles are determined from ranking defined on the categorical values based on their frequencies. Therefore, when we choose quartiles, for example, we can represent each feature value for different feature types in the same form of a 5-tuple (min,Q 1 , Q 2 , Q 3 , max)

. This common representation then allows for a unified approach to S-PCA. In the following subsections, we describe detail procedures to have quantile values for various feature types.

12 Quantiles for Interval-valued Feature

Let a feature F j be an interval-valued feature and let each object ? k ? U be represented by an interval:I kj = [x kj(min) , x kj(max)], k = 1, 2,...,N.**(18)**

We assume that each interval has a uniform distribu tion [2,4,15]. Then, in the case of m quantiles, the resultant (m -1) quantile values becomeQ kji = x kj(min) + (x kj(max) -x kj(min)) × i/m, i = 1, 2,...,m -1. (**19**)

Therefore, each object ? k ? U for the feature F j is described by an (m + 1)-tuple:(x kj(min) , Q kj1 , Q kj 2 ,...,Q kj(m-1) , x kj(max)), k = 1, 2,...,N. (**20**)

Fig. 6: A histogram-valued data.

13 Quantiles for Histogram-valued Feature

Let a feature F be a histogram feature and let an object ? ? U be represented by a histogram in Fig. 6. Let the histogram be composed of n bins, and let p i be the probability of the ith bin, where it is assumed thatp 1 + p 2 +???+ p n = 1.

Then, under the assumption that n bins (subintervals) have uniform distributions, we define the cumulative distribution function F (x) of the histogram [2,4] as: The Quantile Method for Symbolic Principal Component Analysis Then, in the case of m quantiles, we can find (m + 1) values including (m -1) quantile values from the equations:F (x) = 0 for x ? x 1 F (x) = p 1 (x -x 1)/(x 2 -x 1) for x 1 ? x<x 2 F (x) = F (x 2) + p 2 (x -x 2)/(x 3 -x 2) for x 2 ? x<x 3 ?????? F (x) = F (x n) + p n (x -x n)/(x n+1 -x n) for x n ? x<x n+1 F (x) = 1 for x n+1 ? x. London Journal ofF (min) = 0, (i.e. min = x 1) F (Q 2) = 1/m, F (Q 3) = 2/m . . . , F (Q m)

= (m -1)/m, andF (max) = 1, (i.e. max = x n+1).

Therefore, the object ? k ? U is described by an (m + 1)-tuple(x min , Q 1 , Q 2 ,...,Q m-1 , x max). (**21**)

In general, we can describe each object ? k ? U under a histogram-valued feature F j by an (m + 1)-tuple:(x kj(min) , Q kj1 , Q kj 2 ,...,Q kj(m-1) , x kj(max)), k = 1, 2,...,N.**(22)**

It should be noted that the numbers of bins of the given histograms are mutually different in general. However, we can select an integer number m, and obtain (m + 1)-tuples as the common representation for all histograms.

14 Quantiles for Nominal (categorical) Multi-valued Feature

Let F j be a multi-valued feature which takes n possible categorical values c i , i = 1, 2, ... , n. For each i, let p i be the relative frequency of categorical value c i in terms of N objects [2,4,15]. Then, we sort the relative frequency values. For simplicity, we assume that:p 1 ? p 2 ????? p n . (**23**)

According to this order, we suppose rank values 1, 2, ... , n for the categorical values c 1 , c 2 , ... , c n , respectively. We define the cumulative distribution function for each object ? k ? U based on the rank values.

Let n k be the number of possible categorical values taken by object ? k ? U under F j . Let q ki be the frequency value associated with the category c i and given byq ki = 1/n k if c i is a possible value for ? k ? U under F j , = 0 otherwise.

Therefore, we define a piecewise linear cumulative distribution function for each object ? k ? U based on uniform densities attached to rank values (see Example 4). Then we find (m + 1) values including quantile values for the selected integer number m. Therefore, we can obtain again the common (m + 1)-tuple representation:(x kj(min) , Q kj 1 , Q kj2 ,...,Q kj(m-1) , x kj(max)), k = 1, 2,...,N.**(24)**

EXAMPLE 4: The fifth feature (Major acids) of Table **??** is an example of nominal multi-valued feature. We suppose the quartile case, i.e. m = 4. For this purpose, we use basically the procedure given in the above. However, in order to prevent ties of rank values, we use the sums of frequency values attached to the category values of each object. where we should note that the interval [9,10] is attached to the maximum rank value nine. The corresponding cumulative distribution function is a piecewise linear function F (x) characterized by:F (x) = 0, 1 ? x < 4; F (x) = 0.2 × (x -4), 4 ? x < 5; F (x) = 0.2 + 0.2 × (x -5), 5 ? x < 6; F (x) = 0.4, 6 ? x < 7; F (x) = 0.4 + 0.2 × (x -7), 7 ? x < 8; F (x) = 0.6 + 0.2 × (x -8), 8 ? x < 9; F (x) = 0.8 + 0.2 × (x -9), 9 ? x ? 10. (**26**)

By solving the equations F (x) = 0.25, F (x) = 0.5, and F (x) = 0.75, we obtain the quartile values Let each object ? k ? U be described with the given d features by (m + 1)-tuples: London Journal of Research in Science: Natural and FormalQ 1 = 5.25,30 Volume 23 | Issue 12 | Compilation 1.0 © 2023 Great Britain Journal Press

The Quantile Method for Symbolic Principal Component Analysis (x kj(min) , Q kj1 , Q kj 2 ,...,Q kj(m-1) , x kj(max)), j = 1, 2,...,d; k = 1, 2,...,N.

(

Then, we define the quantile sub-object ? kQi as:x kQi = (Q k1i , Q k2i ,...,Q kdi), i = 1.2,...,m -1; k = 1, 2,...,N.**(29)**

PROPOSITION 7: For each object ? k ? U , the min imum sub-object ? k(min) , (m -1) quantile sub-objects (? kQ1 , ? kQ2 ,... , ? kQ(m-1)), and the maximum sub-object ? k(max) organize a monotone structure in the space R d .

Proof: From the definition of (m + 1) sub-objects, we can obtain the following nesting relations of the Cartesian join regions:J (? k(min) , ? kQ1) ? J (? k(min) , ? kQ2) ????? J (? k(min) , ? kQ(m-1)) ? J (? k(min) , ? k(max)). (**30**)

Thus, Definition 7 leads the conclusion.

15 PROPOSITION 8: Property of correlation matrix S by the quantile method

Let a series of objects ? k ? U , k = 1, 2,... , N, is mono tone in the space R d and let the d × d correlation matrix S be obtained by applying the Kendall or Spearman's rank correlation coefficients to the N × (m + 1) sub-objects of Definition 9. Then, the absolute value of each off-diagonal element of S is large.

Proof: From Proposition 7, (m + 1) sub-objects for each of N objects organize always a monotone structure in any subspace of R d . Therefore, if the given series of objects is monotone, their nesting property restrict the order of N × (m + 1) sub-objects to be similar in any subspace of R d . This leads to the conclusion. Now, the quantile method for general S-PCA is summa rized as follows.

PROCEDURE 2: The quantile method for S-PCA Let the set of N objects ? 1 , ? 2 , ... , ? N be described by d features, which are a mixture of interval features, histogram features, nominal multi-valued features, and other types. Then, we execute the quantile method by the following steps.

1. We select an integer value m (1 ? m < N). 2. For each feature F j , we find the common represen tation of N objects by the (m + 1)-tuples:(x kj(min) , Q kj 1 , Q kj 2 ,...,Q kj(m-1) ,

x kj(max)), k = 1, 2,...,N.

16 3.

For each object ? k , we find (m + 1) d-dimensional sub-objects: the minimum sub-object ? k(min) , (m -1) quantile sub-objects, ? kQ1 , ? kQ2 , ... , ? kQ(m-1) , and the maximum sub-object ? k(max) . Then we split each object into (m + 1) sub-objects. As the result, we have an {N × (m + 1)} × d numerical data table. 4. We calculate the d × d correlation matrix S for the {N × (m + 1)} × d data table obtained in 3) based on the selected correlation coefficient, where we can use the Kendall or Spearman's rank correlation coefficient, or the Pearson correlation coefficient. 5. We find the principal components based on the correlation matrix in 4).

In the factor planes, we can reproduce each object ? k , k = 1, 2,... , N, as a series of m arrow lines:? k(min) ? ? kQ1 ? ? kQ2 ????? ? kQ(m-1) ? ? k(max) .

(31)

As a different representation, we can use also a series of m rectangles.

In this procedure, if we select as m = 1, the quantile method is reduced to the original "object splitting method".

V. EXAMPLES OF THE QUANTILE METHOD FOR S-PCA EXAMPLE 5: Fats' and oils' data

We illustrate the quartile case, i.e. m = 4. In this case, the common representation of each object under a feature is 5-tuple, i.e. (min, Q 1 , Q 2 , Q 3 , max). For the fifth feature Major acids, we used the quantification in Example 4. For the data in Table **??**, we obtain the necessary 5-tuples for each of the eight objects with respect to five features. Then, we split each object into five sub-objects, i.e. the minimum sub-object, three quantile sub-objects, and the maximum sub-object. Therefore, we have 40 sub-objects for the given eight objects. Table 3 shows a part of our data, where five sub-objects are presented only for Linseed and Perilla. C, and so on. We selected the following eight features to describe objects (hardwoods). The data formats for other features F 2 -F 8 are the same with Table 5, viz., In this example, deciles and quartiles describe each object, where the preselected number m is 6, and the 7-tuple is used as a common representation for the given Ichino: The Quantile Method for Symbolic PCA 195 6 shows a part of the transformed data table.

Table 7 shows the 8 × 8 correlation matrices, where the upper triangular matrix shows the elements of the Pear son correlation matrix, and the lower triangular matrix shows the elements of the Spearman's rank correlation matrix. The Pearson and the Spearman correlation matrices are similar in many elements. However, some differences should be pointed out. Features F 1 (ANNT), F 2 (JANT), F 3 (JULT), and F 7 (GDC5) are highly correlated mutually for the Spearman coefficient. Feature F 4 (ANNP) is strongly correlated with features F 5 (JANP) and F 8 (MITH) for the Spearman coefficient, while F 4 (ANNP) is largely cor related with features F 5 (JANP) and F 6 (JULP) for the Pearson coefficient. We see also a difference between the Pearson and Spearman correlation coefficients concerning feature F 7 (GDC5).

The contribution ratios of the first two principal compo nents are 77.01% and 11.64% for the Pearson correlation matrix, and are 87.41% and 8.38% for the Spear man correlation matrix. 15 line representations of sixteen hardwoods in the factor planes by the Pearson and Spearman correlation matrices, respectively. In the two factor planes, the first principal component plays the role of the size factor, and the given eight features take similar positive weights. In the second principal component, four features concerning precipita tion and moisture, i.e. ANNP, JANP, JULP, and MITH, take positive weights, while other features for temperature and growing degree, i.e. ANNT, JANT, JULT, and GDC5, took negative weights. For the Spearman correlation matrix, moisture (MITH) takes an especially large positive weight for the second principal component. However, for the Pear son correlation matrix, the corresponding weight is very small.

In Fig. 9, many series of arrow lines tend to be slightly right down. Almost all kinds of hardwood in the eastern area of the US organize a large stream of arrow lines. This tendency of the main stream depends on temperature and precipitation. On the other hand, largely fluctuating and mutually separate streams are mainly composed of the hardwoods in the western area. For example, Acer West, Alnus West, Betula, and Fraxinus West most drastically change toward the upper right with the last decile. This change is heavily dependent on precipitation and moisture. In Fig. 10, the main stream of arrow lines has two branches. Each branch initially grows toward the upper right, and then changes direction toward right down. This property is not clear in Fig. 9. Generally, mutual arrow lines are clearly represented in Fig. 10. Therefore, in this example, the Spearman correlation matrix may be better than the Pearson correlation matrix. Since the quantile method is based on the monotonic property of the given set of objects, the use of the Spearman correlation matrix may be natural.

17 VI. CONCLUDING REMARKS

We presented the quantile method for the S-PCA. The quantile method can treat not only histogram-valued data, but also nominal and ordinal multi-valued type data, and is simply based on the property of monotone structure of the given objects. By selecting a common integer number m, the quantile method transforms a given N × d complex symbolic data table to a simple (N× (m + 1)) × d numerical data table. An important aspect is that we can select the integer m as a sufficiently small number compared to the number N of objects, and we can apply the traditional PCA simply to the (N× (m + 1)) × d data table. We presented several experimental

11

Figure 1: and J (? 1 Fig. 1 :

Figure 2: L

42

Figure 3: PROPOSITION 4 :Fig. 2 :

Figure 4:

results in order to show the effectiveness of the quantile method. An arrow line representation of objects in the factor plane may be a useful tool to analyze complex symbolic data tables. [1] [2] [3] [4] [5] [6] [7] [8] [9] [10]

[1] Volume 23 | Issue 12 | Compilation 1.0 © 2023 Great Britain Journal PressThe Quantile Method for Symbolic Principal Component Analysis

[2] ©

[3] ©

[4] Volume 23 | Issue 12 | Compilation 1.0 © 2023 Great Britain Journal PressThe Quantile Method for Symbolic Principal Component Analysis

[5] ©

[6] ©

[7] ©

[8] ©

[9] Scientific Research (C) 19500130). The author wishes to thank referees and editors for suggestions leading improvements in this article. The author also acknowledges to Professor Paula Brito for her collaborations.

[10] Volume 23 | Issue 12 | Compilation 1.0 © 2023 Great Britain Journal PressThe Quantile Method for Symbolic Principal Component Analysis

8

Figure 5: DEFINITION 8 :

Figure 6:

3

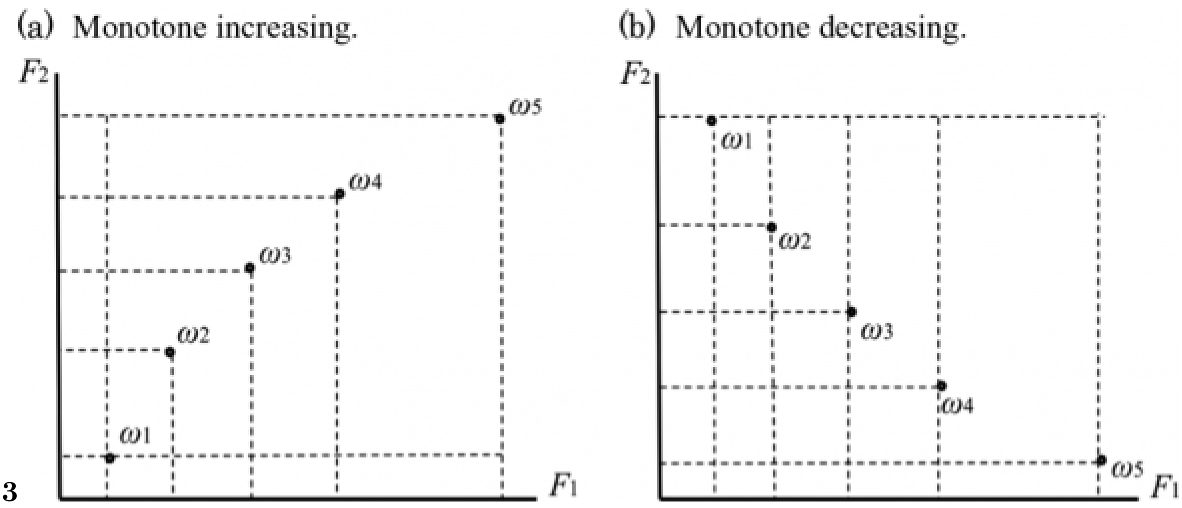


Figure 7: Fig. 3 :

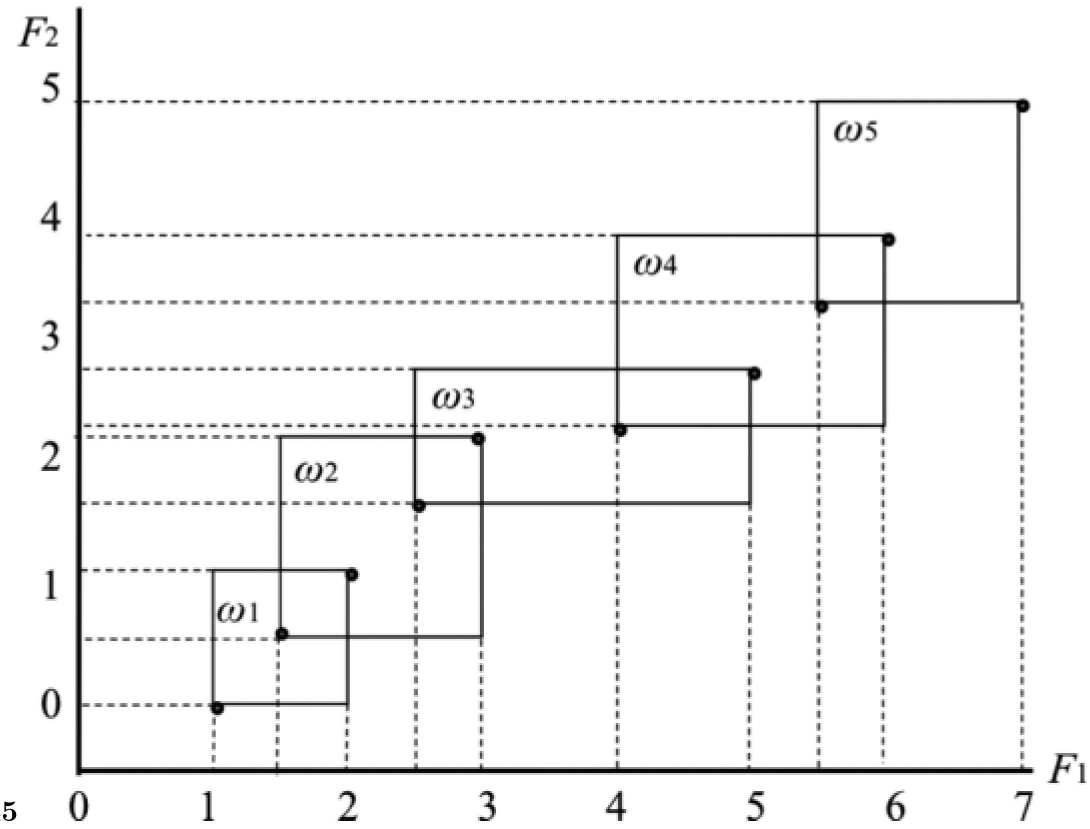


Figure 8: Fig. 4 :Fig. 5 :

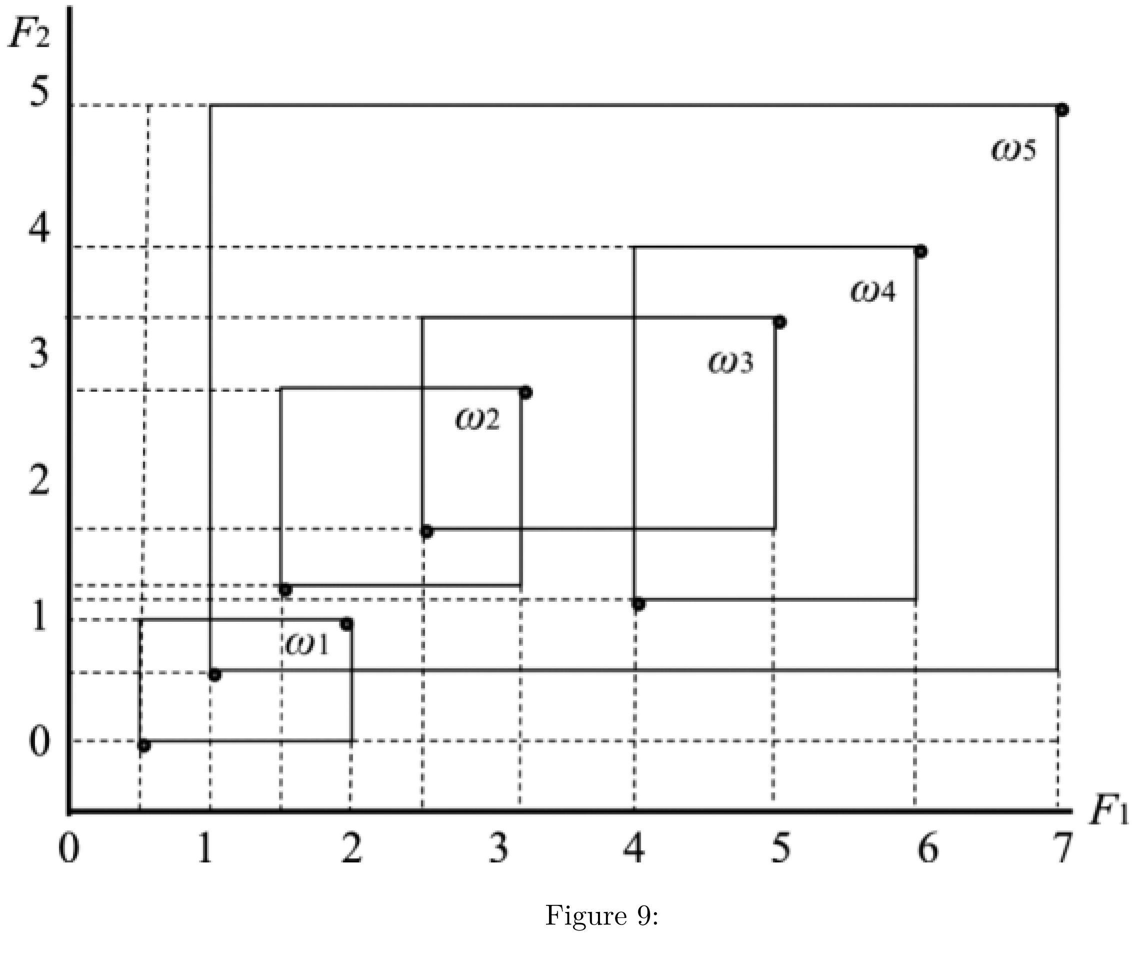
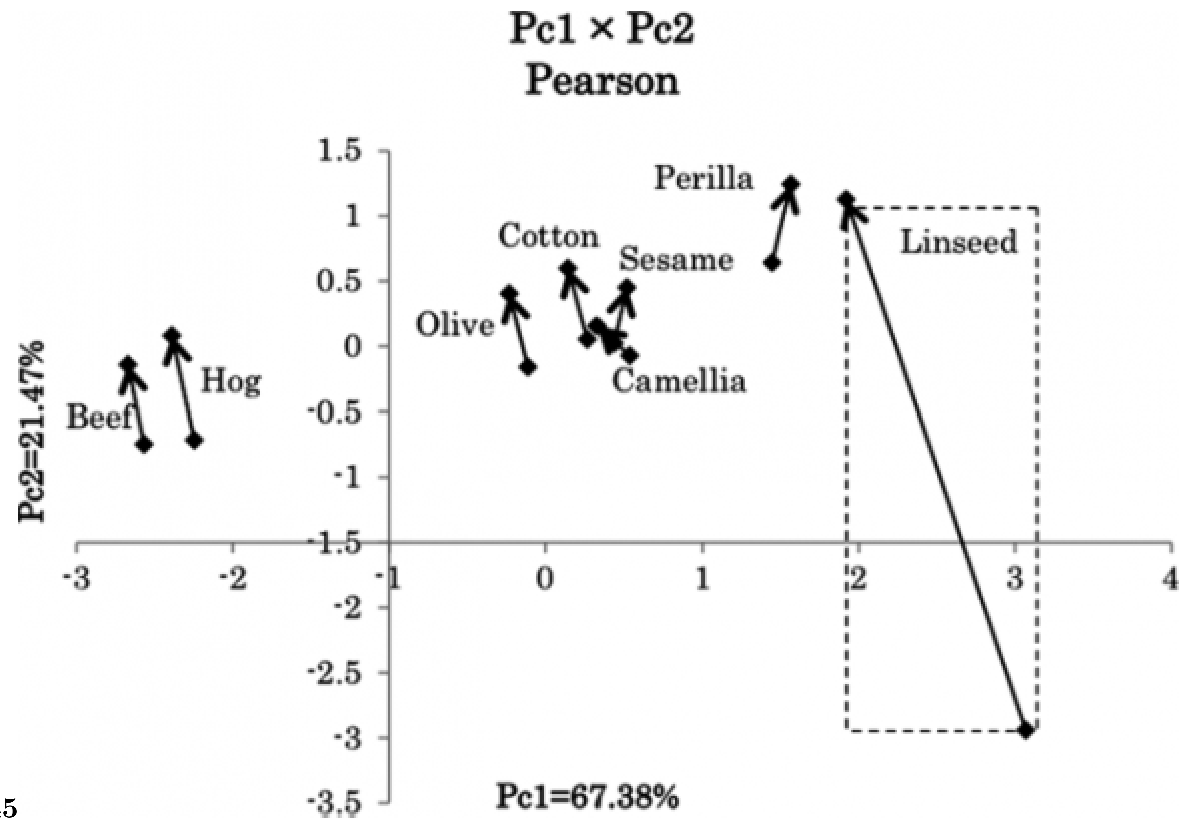
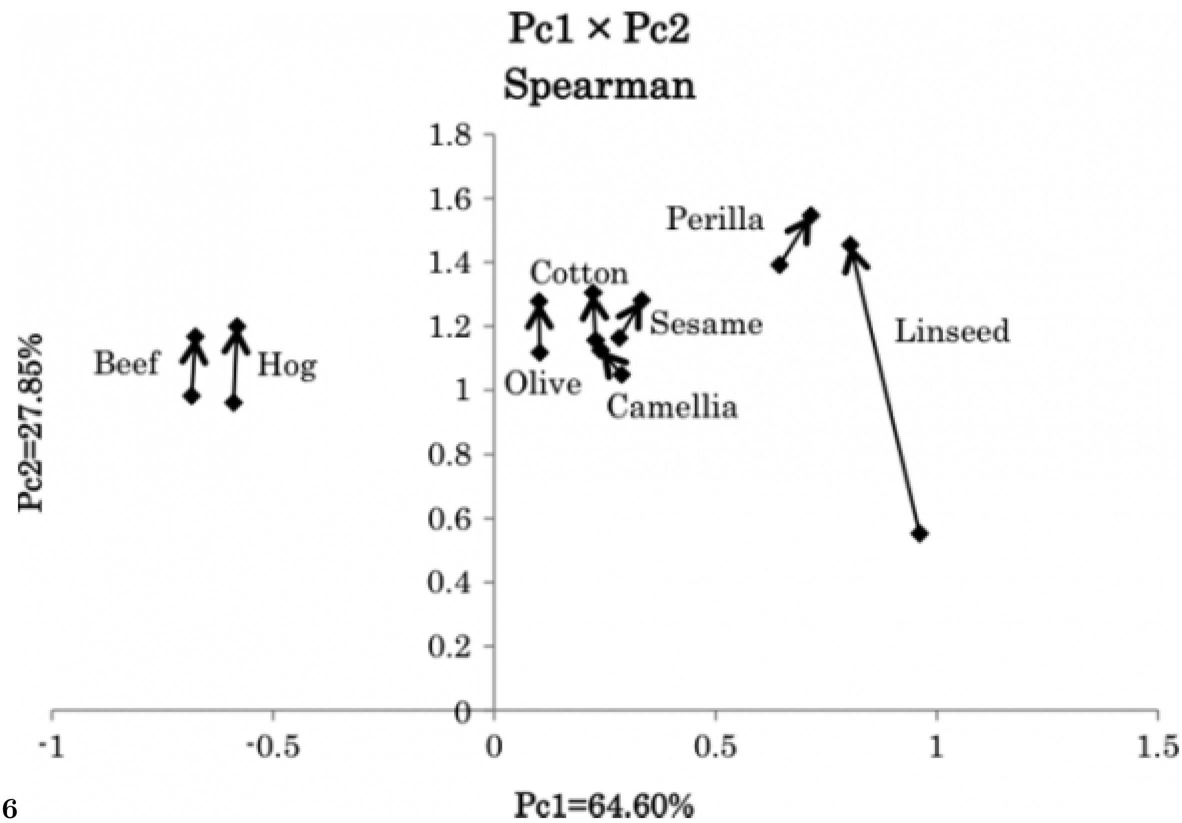


Figure 9:



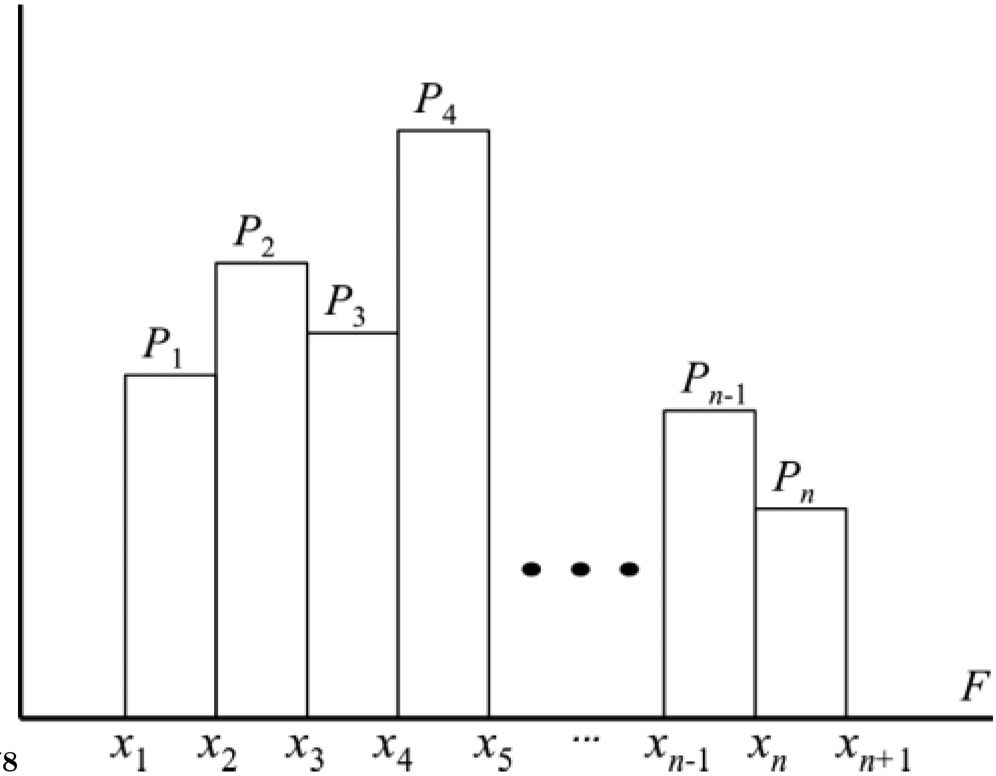
12345

Figure 10: F 1 :F 2 :F 3 :F 4 :F 5 :



6

Figure 11: F 6 :



78

Figure 12: F 7 :F 8 :

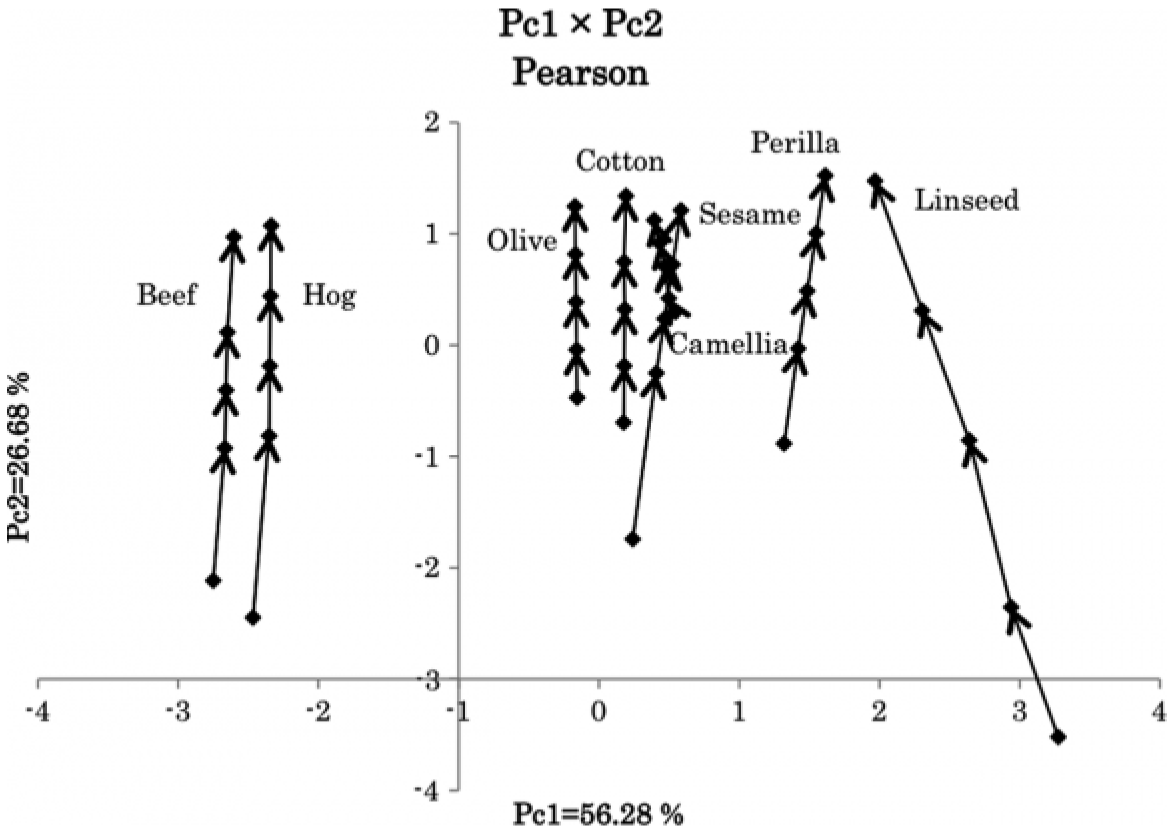
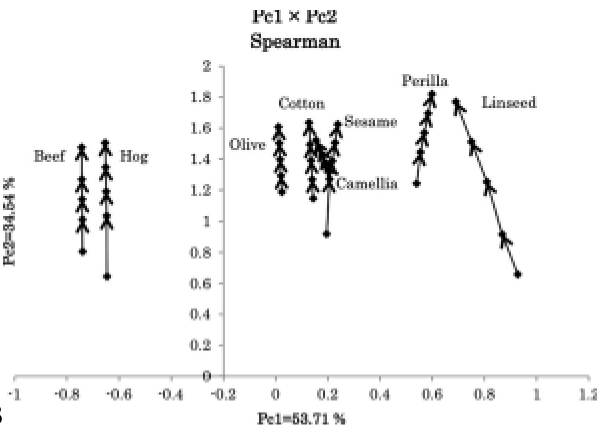


Figure 13:



8

Figure 14: Fig. 8 :

9

Figure 15: Fig. 9 :

10

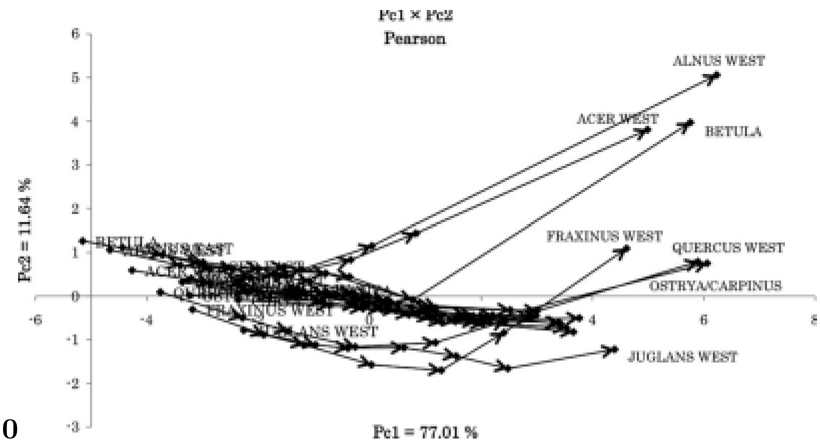


Figure 16: Fig. 10 :

[1,4[:0;[4,5[:0.2;[5,6[:0.2;[6,7[:0.2;[7,8[:0.2; [8,9[:0.2;[9,10]:0.2, (25)

Figure 17:

2

Object	Lu	A	C	Ln	M	S	P	L	O
Linseed	0	0	0	0.2	0 .2	0	0 .2	0 .2	0 .2
Perilla	0	0	0	0.2	0	0 .2	0 .2	0 .2	0 .2
Cotton	0	0	0	0	0.2	0 .2	0 .2	0 .2	0 .2
Sesame	0	0.2	0	0	0	0 .2	0 .2	0 .2	0 .2
Camellia	0	0	0	0	0	0	0	0.5	0 .5
Olive	0	0	0	0	0	0.25	0.25	0.25	0.25
Beef	0	0	0.2	0	0 .2	0 .2	0 .2	0	0 .2
Hog	0.167	0	0	0	0.167	0.167	0.167	0.167	0.167
q ij	0.167	0.2	0 .2	0 .4	0 .767	1.217	1.417	1.717	1.917
R a n k	1	2	2	4	5	6	7	8	9

Q 2 = 7.5, and Q 3 = 8.75, respectively. Finally, we have the desired 5-tuple:

(4, 5.25, 7.5, 8.75, 10). (27)

IV. THE QUANTILE METHOD FOR S-PCA

DEFINITION 9: Quantile sub-objects.

Figure 18: Table 2 :

3

London Journal of Research in Science: Natural and Formal	F 1	F 2	F 3	F 4	F 5
Linseed					
1	0.93000	-27	170	118	4
2	0.93125	-24.75	178.5	137.5	5 .25
3	0.93250	-22.5	187	157	7.5
4	0.93375	-20.25	195.5	176.5	8 .75
5	0.93500	-18	204	196	10
Perilla					
1	0.93000	-5	192	188	4
2	0.93175	-4.75	196	190.25	6.25
3	0.93350	-4.5	200	192.5	7 .5
4	0.93525	-4.25	204	194.75	8.75
5	0.93700	-4	208	197	10
32 Volume 23 | Issue 12 | Compilation 1.0					© 2023 Great Britain Journal Press

Figure 19: Table 3 :

4

S	Spec.	Freez.	Iodine	Sapon.	M. acids
Spec.	1.0000 -0.8923		0.7682 -0.3187		0.2432
Freez.	-0.6309	1.0000 -0.6368		0.4968 -0.1138	
Iodine	0.9582 -0.6142		1.0000 -0.3834		0.1107
Saponi.	-0.2044	0.6437 -0.1980		1.0000	0.3634
M. acids	0.2558	0.0398	0.1805	0.6428	1.0000

Fig. 7: The result of the S-PCA for Fats' and oils' data (Pearson). 16 hardwoods. According to the Procedure 2 for S-PCA in Section 4, we transform the given (16 objects) × (8 features) symbolic data table to a (16 × 7 sub-objects) × (8 features) standard numerical data table.

Figure 20: Table 4 :

Figure 21: Table

5

Taxon name	N 6 865	0% -2.3 -3.9 -10.2 -12.2 -	50% 9 .2	75% 1	90% 1 7	100% 2
Acer East	1 954 10	13.4 3.6 Histogram 10% 0	4 .2 0 .6	4 .4 7	.9 1 0 .3	3 .8 2 0
Acer West	144 4	.6 0 .2 -4.4 -4.6 -8.4 data	0 .3 -1.0	.5 6 .1	1 5 .0 7	.6 2 0 .9
Alnus East	761 16	(annual 25% 3 .8 1 .9 -2.3	tempera-	3 .2 3	.6 1 2 .6	1 8 .7 2 0
Alnus West	815 4	-3.0 -5.1	ture).	.9		.3
Betula Carya	638					

London Journal of Research in Science: Natural and FormalThe Quantile Method for Symbolic Principal Component Analysis

Figure 22: Table 5 :

6

Figure 23: Table 6 :

7

London Journal of Research in Science: Natural and Formal 36 Volume 23 | Issue 12 | Compilation 1.0 © 2023 Great Britain Journal PressThe Quantile Method for Symbolic Principal Component Analysis

Figure 24: Table 7 :

.1 ACKNOWLEDGMENTS

This research was partly supported by Japan Society of the Promotion of Science (Grant-in Aid for

[London Journal of Research in Science: Natural and Formal] , *London Journal of Research in Science: Natural and Formal*

[Bock and Diday (eds.) ()] *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*, H H Bock , E Diday (eds.) (Berlin) 2000. Springer-Verlag.

[Billard and Diday ()] L Billard , E Diday . *Symbolic Data Analysis: Concep tual Statistics and Data Mining*, (Chichester, Wiley) 2006.

[Chatfield et al. ()] 'Descriptive statistics for sym bolic data'. C Chatfield , A J Collins , P Bertrand , F Goupil . *Analysis of Symbolic Data*, H.-H Bock , E Diday (eds.) (New York; Berlin) 1984. 2000. Springer-Verlag. (Introduction to Multivariate Analysis)

[Chouakria ()] *Extension de l'analyse en composantes principales a des donnees de type intervalle*, A Chouakria . 1998. University of Paris IX Dauphine (Doctoral ´Thesis)

[Billard and Diday ()] 'From the statistics of data to the statistics of knowledge: symbolic data analysis'. L Billard , E Diday . *J Am Stat Assoc* 2003. 98 (462) p. .

[Ichino ()] 'General metrics for mixed features-the Carte sian space theory for pattern recognition'. M Ichino . *Proceedings on International Conference on Systems, Man, and Cybernet ics*, (on International Conference on Systems, Man, and Cybernet icsChaina, Beijin) 1988.

[Ichino and Yaguchi ()] 'Generalized Minkowski metrics for mixed feature type data analysis'. M Ichino , H Yaguchi . *IEEE Trans Syst Man Cybern* 1994. 24 (4) p. .

[Histogram data by the U.S. Geological Survey, Climate Vegetation Atlas of North America (2008)] `http://pubs.usgs.gov/pp/p1650-b/` *Histogram data by the U.S. Geological Survey, Climate Vegetation Atlas of North America*, October 2, 2008.

[Lauro and Palumbo ()] 'Principal component analysis of interval data: a symbolic data analysis approach'. C Lauro , F Palumbo . *Comput Stat* 2000. 15 (1) p. .

[Lauro et al. ()] 'Principal component analysis of symbolic data described by intervals'. C Lauro , R Verde , A Irpino . *Sym bolic Data Analysis and the SODAS Software*, E Diday , M Noirhomme-Fraiture (eds.) 2008. Chichester, Wiley. p. .

[Diday and Noirhomme-Fraiture (eds.) ()] *Symbolic Data Analysis and the SODAS Software*, E Diday , M Noirhomme-Fraiture (eds.) (Chichester, Wiley) 2008.

[Ichino et al. ()] 'Symbolic pattern classifiers 13. based on the Cartesian system model'. M Ichino , M Ichino , H Yaguchi . *Proceedings of IASC 2008*, Data Science , Related Classification , C Methods , Hayashi (eds.) (IASC 2008Japan, Yokohama; Tokyo) 2008. 1998. Springer-Verlag. p. . (Symbolic PCA for histogram-valued data)

[Chouakria et al. ()] 'Symbolic princi pal component analysis'. A Chouakria , P Cazes , E Diday . *Analysis of Symbolic Data*, H.-H Bock , E Diday (eds.) (Berlin) 2000. Springer-Verlag.

[Ichino ()] 'Symbolic principal component analysis based on the nested covering'. M Ichino . *Proceedings ISI2007*, (ISI2007Portugal, Lisbon) 2007.