The Data Accumulation Method for Symbolic Principal Component Analysis

Ichino, Manabu
Tokyo Denki University, College of Science and Engineering
Hatoyama, Saitama 350-0394, Japan
E-mail: ichino@mail.dendai.ac.jp

Brito, Paula
Universidade do Porto, Faculdade de Economia & LIAAD-INESC Porto LA
Rua Dr. Robert Frias 4200-464 Porto, Portugal
E-mail: mpbrito@fep.up.pt

Several methods have been proposed to realize the principal component analysis for symbolic data tables [1-4, 6]. One of the authors presented the quantile method for symbolic principal component analysis [3-4]. Quantile representation provides a common framework to represent symbolic data described by variables of different types. The principle is to express the observed variable values by some predefined quantiles of the underlying distribution. This common representation then allows for a unified analysis of the data set, taking all variables simultaneously into account [3-5]. The quantile method for symbolic principal component analysis [3-4] transforms the given (N objects)×(d features) symbolic data table to a $\{N\times(m+1)$ sub-objects$\}\times(d$ features) numerical data table, where m is a preselected integer number to determine the number of quantiles. Then, we can execute the standard PCA for this transformed data table. The quantile method for PCA is based on the fact that a monotone property of symbolic objects is characterized by the nesting structure of the Cartesian join regions. In this paper, we present the data accumulation method for symbolic PCA. When we have n symbolic data tables of the form (N objects)×(d features), we first transform each of n data tables to a $\{N\times(m+1)$ sub-objects$\}\times(d$ features) numerical data table for preselected m. Then, the data accumulation method accumulates these n numerical data tables to a $\{N\times(m+1)$ sub-objects$\}\times(d$ features) or a $\{n\times N\times(m+1)$ sub-objects$\}\times(d$ features) numerical data table. We execute the PCA for these transformed data tables. Since, we often encounter periodically summarized data tables, the data accumulation method for symbolic PCA becomes a useful tool to understand n data tables as a whole. We present examples in order to show the usefulness of this method.

n Symbolic Data Tables

Suppose that we have the following n symbolic data tables, $s = 1, 2,\ldots, n$.

***Table* 1** The structure of s-th symbolic data table.

Table s	F_1	F_2	......	F_k	......	F_d
ω_1	E_{11}	E_{12}	......	E_{1k}	......	E_{1d}
ω_2	E_{21}	E_{22}	......	E_{2k}	......	E_{2d}
......	......	......	......	......	......	......
ω_i	E_{i1}	E_{i2}	......	E_{ik}	......	E_{id}
......	......	......	......	......	......	......
ω_N	E_{N1}	E_{N2}	......	E_{Nk}	......	E_{Nd}

This data table is composed of N objects, ω_1, $\omega_2,\ldots, \omega_N$. Each object ω_i is described by d feature values, E_{i1}, $E_{i2},\ldots, E_{id}$ with respect to d features, F_1, $F_2, \ldots, F_d$.

The Quantile Method [3-5]

In the quantile method, we split each object ω_i into (m+1) sub-objects, ω_{i1}, ω_{i2},…, $\omega_{i(m+1)}$, for the preselected integer number m. Each sub-object ω_{ij} is described as follows.

For a numerical feature F_k, each feature value E_{ijk} becomes a numerical value E_{ik} for all j = 1, 2,…, m+1. Then, we have

$$E_{ijk} = E_{ik},\ j = 1, 2,\ldots, m+1.$$

On the other hand, for other feature types such as interval feature, histogram feature, and multinominal feature, each sub-object ω_{ij} is represented, for each feature F_k, by m+1 numerical values corresponding to the minimum value min_{ijk}, (m-1) quantile values, Q_{ijk1}, Q_{ijk2},..., and $Q_{ijk(m-1)}$, and the maximum value max_{ijk}. Then, we have $E_{i1k} = min_{i1k}$, $E_{i2k} = Q_{i2k1}$,..., $E_{imk} = Q_{imk(m-1)}$, and $E_{i(m+1)k} = max_{i(m+1)k}$.

Therefore, we have a new numerical data table shown in Table 2.

***Table* 2** The structure of n transformed numerical data tables.

Table s	F_1	F_2	……	F_k	……	F_d
ω_{11}	E_{111}	E_{112}	……	E_{11k}	……	E_{11d}
ω_{12}	E_{121}	E_{122}	……	E_{12k}	……	E_{12d}
……	……	……	……	……	……	……
$\omega_{1(m+1)}$	$E_{1\ (m+1)\ 1}$	$E_{1(\mathrm{m}+1)2}$	……	$E_{1(m+1)k}$	……	$E_{1(m+1)d}$
……	……	……	……	……	……	……
ω_{N1}	E_{N11}	E_{N12}	……	E_{N1k}	……	E_{N1d}
ω_{N2}	E_{N21}	E_{N22}	……	E_{N2k}	……	E_{N2d}
……	……	……	……	……	……	……
$\omega_{N(m+1)}$	$E_{N\ (m+1)1}$	$E_{N(\mathrm{m}+1)2}$	……	$E_{N(m+1)k}$	……	$E_{N(m+1)d}$

The quantile method for symbolic PCA executes the standard PCA for this transformed numerical data table. Now, we rewrite all feature values in Table 2 in order to clarify so that they are feature values for the *s-th* data table.

***Table* 3** The structure of *s-th* transformed numerical data table.

Table s	F_1	F_2	……	F_k	……	F_d
ω_{11}	E_{s111}	E_{s112}	……	E_{s11k}	……	E_{s11d}
ω_{12}	E_{s121}	E_{s122}	……	E_{s12k}	……	E_{s12d}
……	……	……	……	……	……	……
$\omega_{1(m+1)}$	$E_{\mathrm{s}1\ (m+1)\ 1}$	$E_{s1(\mathrm{m}+1)2}$	……	$E_{s1(m+1)k}$	……	$E_{s1(m+1)d}$
……	……	……	……	……	……	……
ω_{N1}	E_{sN11}	E_{sN12}	……	E_{sN1k}	……	E_{sN1d}
ω_{N2}	E_{sN21}	E_{sN22}	……	E_{sN2k}	……	E_{sN2d}
……	……	……	……	……	……	……
$\omega_{N(m+1)}$	$E_{sN\ (m+1)1}$	$E_{sN(\mathrm{m}+1)2}$	……	$E_{sN(m+1)k}$	……	$E_{sN(m+1)d}$

The Data Accumulation Methods

In the data accumulation method for symbolic PCA, we accumulate feature values for n data tables and obtain a single numerical data table to which we apply the standard PCA. In the following, we describe two possible methods for the data accumulation.

The Data Accumulation Method I

In the first method, n data tables have the same structure as in Table 3 for the preselected common integer number m. Therefore, each data table is the size of $\{N\times(m+1)$ sub-objects$\}\times\{d$ features$\}$. Then, for each sub-object ω_{ij}, $j = 1, 2,..., (m+1)$; $i = 1, 2,\ldots, N$, we generate new sub-objects and their feature values as follows.

$$\omega_{ij}: E_{ijk} = E_{1ijk} + E_{2ijk} +\cdots+E_{nijk},\ k = 1, 2,..., d.$$

Since $(m+1)$ sub-objects for each object ω_i satisfy the monotone property, $(m+1)$ sub-objects generated by the above data accumulation satisfy again the monotone property. The new accumulated data table is given in the form of Table 4. The data table is again the size of $\{N\times(m+1)$ sub-objects$\}\times\{d$ features$\}$. We apply the standard PCA to this data table. Therefore, in the factor planes, each object is reproduced by an m series of arrow lines, and we can understand the mutual relationships of N objects through the N sets of arrow lines. However, we cannot see the mutual differences between n data tables (see Examples).

***Table* 4** The transformed numerical data table by the data accumulation method I.

Σ	F_1	F_2	……	F_k	……	F_d
ω_{11}	E_{111}	E_{112}	……	E_{11k}	……	E_{11d}
ω_{12}	E_{121}	E_{122}	……	E_{12k}	……	E_{12d}
……	……	……	……	……	……	……
$\omega_{1(m+1)}$	$E_{1\ (m+1)\ 1}$	$E_{1(\mathrm{m}+1)2}$	……	$E_{1(m+1)k}$	……	$E_{1(m+1)d}$
……	……	……	……	……	……	……
ω_{N1}	E_{N11}	E_{N12}	……	E_{N1k}	……	E_{N1d}
ω_{N2}	E_{N21}	E_{N22}	……	E_{N2k}	……	E_{N2d}
……	……	……	……	……	……	……
$\omega_{N(m+1)}$	$E_{N\ (m+1)1}$	$E_{N(\mathrm{m}+1)2}$	……	$E_{N(m+1)k}$	……	$E_{N(m+1)d}$

The Data Accumulation Method II

In the second data accumulation method, we split each sub-object ω_{ij}, $i = 1, 2,\ldots, N$; $j = 1, 2,\ldots, (m+1)$, into n sub-sub-objects, ω_{i1j}, $\omega_{i2j},\ldots$, and ω_{inj}. By using feature values in Table 3, we generate new feature values for each sub-sub-object ω_{isj}, $s = 1, 2,\ldots, n$ as follows.

$$\omega_{i1j}: E_{i1jk} = E_{1ijk},\ k = 1, 2,\ldots, d;\quad j = 1, 2,\ldots, (m+1);\ s = 1.$$

$$\omega_{isj}: E_{isjk} = E_{i(s\text{-}1)(m+1)k} + E_{sijk},\ k = 1, 2,\ldots, d;\quad j = 1, 2,\ldots, (m+1);\ s = 2, 3,\ldots, n.$$

Then, for each object ω_i, $i = 1, 2,\ldots, N$, we have the inequalities:

$$E_{i11k} \le E_{i12k} \le \cdots \le E_{i1(m+1)k} \le E_{i21k} \le E_{i22k} \le \cdots \le E_{i2(m+1)k}$$
$$\cdots \le E_{in1k} \le E_{in2k} \le \cdots \le E_{in(m+1)k},\ k = 1, 2,\ldots, d.$$

Therefore, we have monotone property for $n\times(m+1)$ sub-sub-objects with respect to d features. As the result, we obtain a new numerical data table for each object ω_i as shown in Table 5. In the data accumulation method II, we apply the standard PCA to the numerical data table of the size $\{n\times N\times(m+1)$ sub-sub-objects$\}\times(d$ features). (As a further generalization, we are able to select a different integer number m for each of n data tables. However, we omit here the detail.) In the factor planes, each object is reproduced as a series of $n\times m$ arrow lines. We can obtain mutual relation ships between objects with the difference of n symbolic data tables by the N sets of n×m arrow lines.

***Table* 5** The data table for object ω_i by the data accumulation method II.

$\sum$	F_1	F_2	……	F_k	……	F_d
ω_{i11}	E_{i111}	E_{i112}	……	E_{i11k}	……	E_{i11d}
ω_{i12}	E_{i121}	E_{i122}	……	E_{i12k}	……	E_{i12d}
……	……	……	……	……	……	……
$\omega_{i1(m+1)}$	$E_{i1(m+1)1}$	$E_{i1(m+1)2}$	……	$E_{i1(m+1)k}$	……	$E_{i1(m+1)d}$
……	……	……	……	……	……	……
ω_{i21}	E_{i211}	E_{i212}	……	E_{i21k}	……	E_{i21d}
ω_{i22}	E_{i221}	E_{i222}	……	E_{i22k}	……	E_{i22d}
……	……	……	……	……	……	……
$\omega_{i2(m+1)}$	$E_{i2(m+1)1}$	$E_{i2(m+1)2}$	……	$E_{i2(m+1)k}$	……	$E_{i2(m+1)d}$
……	……	……	……	……	……	……
ω_{in1}	E_{in11}	E_{in12}	……	E_{in1k}		E_{in1d}
ω_{in2}	E_{in21}	E_{in22}	……	E_{in2k}		E_{in2d}
……	……	……	……	……	……	……
$\omega_{in(m+1)}$	$E_{in(m+1)1}$	$E_{in(m+1)2}$	……	$E_{in(m+1)k}$		$E_{in(m+1)d}$

Examples

We use the accommodation and guests data prepared by Japan Tourism Agency (http://www.mlit.go.jp/kankocho/siryou/toukei/shukuhakutoukei.html). We have three annual data tables for 2007, 2008, and 2009. Each data table quarterly (January-March, April-June, July-September, and October-December) summarizes the number of guests for 47 prefectures of Japan according to three accommodation scales by employees (10-29 employees, 30-99 employees, and 100 employees over). We select only 17 prefectures, Hokkaido, Chiba, Tokyo, Kanagawa, Ishikawa, Yamanashi, Nagano, Gifu, Shizuoka, Kyoto, Osaka, Hyogo, Fukuoka, Nagasaki, Kumamoto, Oita, and Okinawa. These prefectures have comparatively large number of guests from foreign countries. For each accommodation scale, we use the total number of guests and the number of guests their main purpose is sightseeing. We use also, for overseas guests, the total number of guests and the number of guests their main purpose is sightseeing. Therefore, each prefecture is described by (four terms)×{(three types of accommodation)×(two numbers of guests) + two numbers of overseas guests} = 32 numerical feature values. We apply the quantile method to each of three data table. Then, the data table for Hokkaido in 2007, for example, is described as Table 6.

We apply the data accumulation method I to 2007, 2008, and 2009 data tables of the form given in Table 6. Figure 1 shows the symbolic representation of 17 prefectures in the first factor plane by the data accumulation method for symbolic PCA based on the Spearman rank correlation matrix. The first principal component plays the role of the size factor, and its contribution ratio is 68.76%. On the other hand, in the second principal component, the second and the forth features take large positive weights, while the fifth, the seventh, and the eighth features take negative weights. The accumulated contribution ratio of the first two principal components is 85.53%. In this figure, Tokyo is reproduced as a longest arrow lines, and receives the largest number of domestic and overseas guests with large-scale accommodations. The purpose of guests is not only sightseeing but also business and others. Therefore, the arrow lines toward right down. Osaka shows the similar direction to Tokyo. However, the size of arrow lines is less than a half. Chiba receives a large number of domestic and overseas guests. Many of guests use large-scale accommodations and their main purpose is sightseeing, for example, to visit tourist facilities such as the Tokyo Disney Land and others. Hokkaido receives the largest number of guests as tourists. Tourists use every type of accommodation. On the other hand, Nagano and Gifu receive main portion of tourists with the first two types of accommodation. As a common property of Hokkaido, Nagano, and Gifu, the most popular season for tourists is the third term, July-September.

Table 6 The quantile representation of Hokkaido.

2007	Accommodation scale by employees							
	10–29 employees		30-99 employees		100 employees over		Number of Overseas guests	
	Number of guests		Number of guests		Number of guests			
	Total	Sight seeing	Total	Sight seeing	Total	Sight seeing	Total	Sight seeing
Hokkaido 2007-1	986,800	438,120	2,088,900	1,407,150	2,719,350	2,367,970	489,160	445,020
Hokkaido 2007-2	2,038,940	861,900	4,223,710	2,910,810	5,083,780	4,370,460	848,260	768,780
Hokkaido 2007-3	3,567,770	1,536,320	7,148,590	4,977,240	8,623,510	7,469,750	1,384,960	1,254,810
Hokkaido 2007-4	4,532,280	1,952,860	9,215,830	6,386,620	11,174,550	9,740,900	1,867,590	1,696,120

Figure 1 Result of the PCA by the Data accumulation method I.

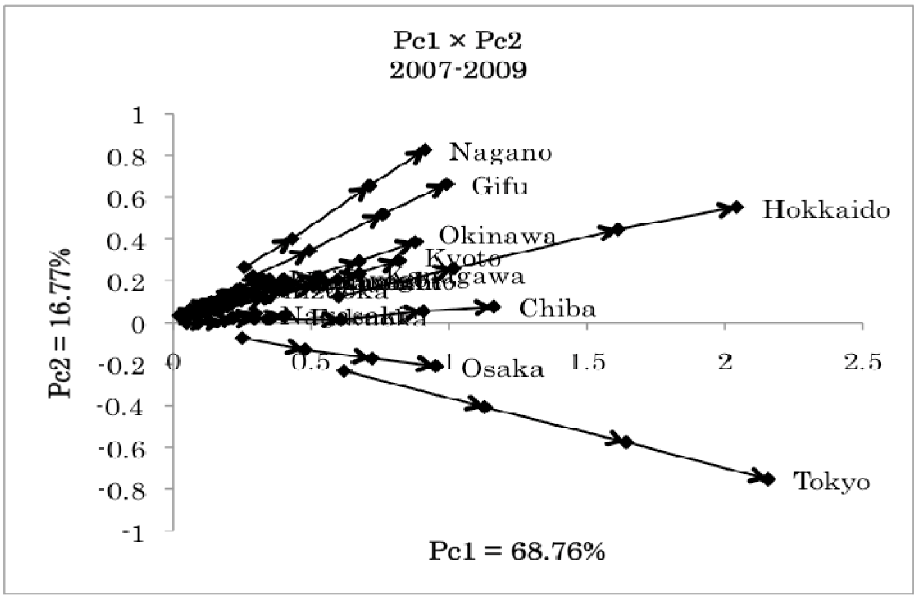
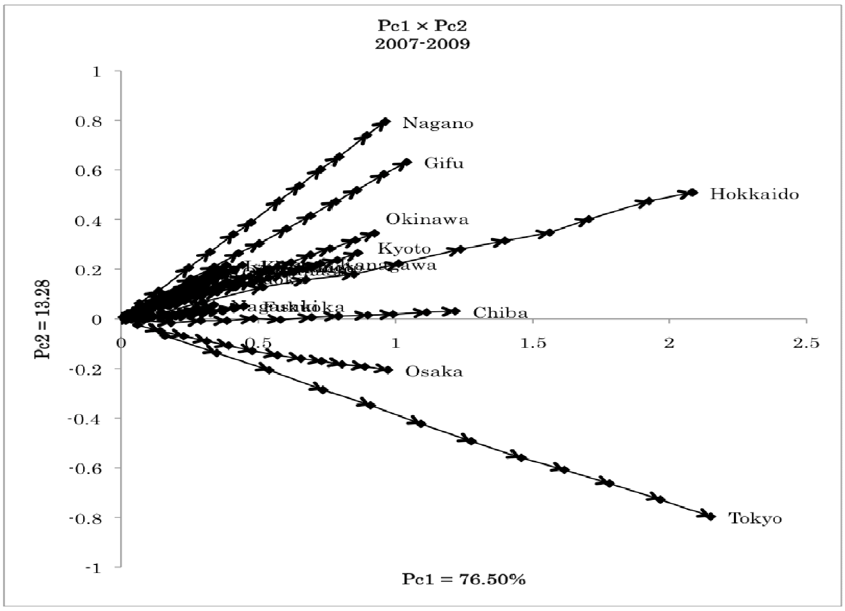
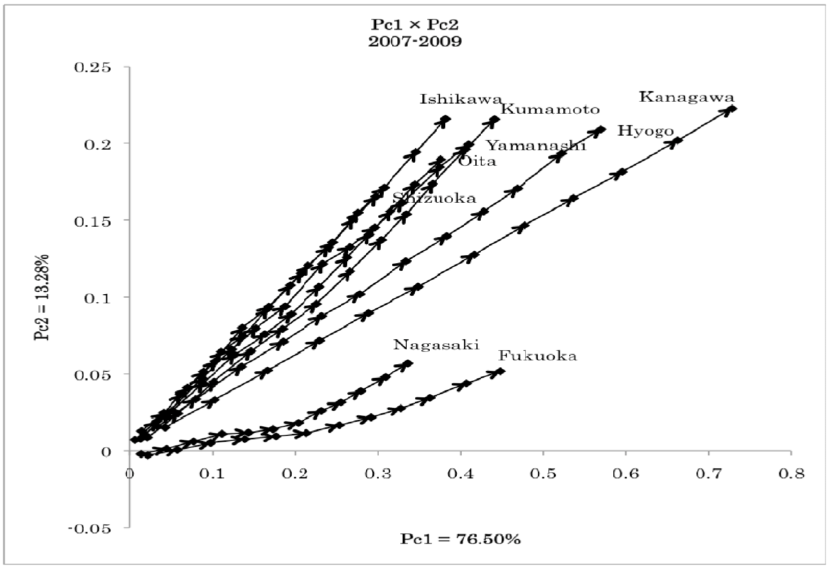


Figure 2 shows the result of the data accumulation method II. In this method, each prefecture is reproduced as an eleven series of arrow lines for 2007, 2008 and 2009. The accumulated contribution ratio of the first principal component is 89.78%. The meanings of these principal components are the same with Figure 1. Figure 2(b) is a zoomed representation. Arrow line representation figures out well the annual changes of the number of guests.

***Figure* 2** (a) The result of the PCA by the data accumulation II.



(b)



REFERENCES

[1] BILLARD, L. and DIDAY, E. (2006): Symbolic Data Analysis; Conceptual Statistics and Data Mining, Chichester, Wiley.

[2] DIDAY, E. and NOIRHOMME-FRAITURE, M. (2008): Symbolic Data Analysis and the SODAS Software, Chichester, Wiley.

[3] ICHINO, M. (2008): Symbolic PCA for histogram-valued data, In: Proc. IASC 2008, Dec.5-8, Yokohama, Japan.

[4] ICHINO, M. (2011); The quantile method for symbolic Principal component analysis, Statistical Analysis and Data Mining 4, 184-198, Wiley (in press).

[5] BRITO, P. and ICHINO, M. (2010): Symbolic clustering based on quantile representation, In: Proc. COMPSTAT2010, August 22-27, Paris, France.

[6] DOUZAL-CHOUAKRIA A., BILLARD, L., DIDAY, E. (2011): Principal component analysis for interval-valued observations, Statistical Analysis and Data Mining 4, 229-246, Wiley (in press).