An Algorithm for Principal Points Considering External Criterion for Multivariate Binary Distributions

Haruka YAMASHITA

Department of Information and Communication, Sophia University, 4 Yonbancho, Chiyoda-ku Tokyo, 102-0081 Japan

Abstract

The concept of principal points has been the subject of numerous theoretical and application studies. There are also the studies of principal points for multivariate binary distributions. Moreover, recently, principal points for multivariate binary distribution considering two classes using an external criterion has studied. For example binary questionnaire data for products or services are classified based on the external criterion of customer satisfaction. In previous studies, principal points were located by minimizing the expected squared distance between the principal points and a multivariate binary distribution within a class, and maximizing the squared distance between the two classes. However, this approach to finding principal points was very time-consuming. In this study, we propose an algorithm for finding principal points considering an external criterion for multivariate binary distributions. Moreover, we apply our method to real binary data that is divided into two classes by the binary criterion, and we demonstrate the effectiveness of our approach.

Keywords: Classification Model, Discriminant Model, Mixture Distribution, Multivariate Data Analysis, Heuristics

1. Introduction

Principal points, as defined by Flury [1], are widely studied in the field of statistics. Principal points are representative points of the probability density function of a dataset in terms of the expected squared distance between these points and the realizations of the distribution.

Analyzing binary (0 or 1) data, particularly multivariate binary data, is one of the difficult and important problems. This type of analysis is required in areas such as questionnaire surveys and qualitative data analyses. When we analyze multivariate binary data, correspondence analysis is most suitable. However, the objects of this analysis method are continuous values; thus, interpreting the analysis results can be quite difficult. For this reason, Yamashita and Suzuki [2] proposed using principal points for binary data. In particular, they proposed representing the data using a small number of realizations (i.e., a small number of binary points) of the multivariate binary distribution. Moreover, certain algorithms were considered suitable in terms of avoiding computational complexity for identifying principal points in multivariate binary distributions [3].

However, in many cases, the multivariate binary distribution comprises a mixture of two different distributions. For example, the distribution of a questionnaire dataset may be comprised of the following two distributions: one of the male respondents and one of the female respondents. Also, Yamashita and Suzuki [4] proposed an approach for finding principal points that minimizes the expectation of the squared distance between the principal points and a multivariate binary distribution within a class and maximizes the squared distance between the two classes, i.e., the points are not only representative of each class but also representative of the difference between the features of the classes. This method is considered as a discriminant model, and it is effective in terms of the presentation of the representative points. However, in practice, finding principal points requires a large computational time.

In this study, we propose an algorithm for finding principal points considering an external criterion for multivariate binary distributions that is based on a greedy algorithm. Moreover, we apply our method to real binary data that is divided into two classes by the binary criterion, and we demonstrate the effectiveness of the proposed approach.

2. Preliminaries

2.1 Principal points

Flury [1] defined the problem of finding k-principal points as follows:

\[(i) \text{ Let } y_j \in \mathbb{R}^p (j = 1, \ldots, k) \text{ be a set of } k \text{ points in a } p\text{-dimensional space, and let the distance } d(x | y_1, \ldots, y_k) \text{ between } x \in \mathbb{R}^p \text{ and } y_j \text{ be defined as follows:} \]

\[d(x | y_1, \ldots, y_k) = \min_{1 \leq j \leq k} \{ (x - y_j)^T(x - y_j) \}^{1/2} \quad (1)\]
principal points of a random vector \( X \) with a probability distribution \( F \) if a set of \( j \) satisfies the following:

\[
E_F \{ d^2(X | \xi_1, ..., \xi_j) \} = \min_{y_j \in \mathbb{R}^n} E_F \{ d^2(X | y_1, ..., y_j) \}
\] (2)

The points \( \xi_1, ..., \xi_j \) represent \( k \) “principal” points of the random vector \( X \), and these \( k \) points also divide \( X \) into \( k \) areas. Namely, \( X \) is described by a set of \( k \) points \( \xi_j \) \((j = 1, ..., k)\) obtained in a manner such that the expected squared distance between the random vector \( X \) and the nearest of the \( k \) points is minimized. Let us discuss the analysis of the multivariate binary distribution according to the concept of principal points, as proposed by Flury (1990). For example, given five variant data points \((10001), (01100), (10001), (11100), \text{and} (00001)\), two principal points can be obtained from these five variant data points, \((1/3 \ 0001)\) \text{and} \((1/2 \ 1100)\), when we assume that there is a probability mass of 1/5 on each data point. It should be noted that the two principal points obtained here take continuous values and do not represent realizations (binary values).

Tarpey and Petkova [5] proposed a method for finding principal points considering an external criterion (e.g., distinguishing male from female-participant data). Moreover, in this discussion, it is important to consider the application efficiency of the placebo effect.

2.2 The definition of \( k \)-principal points for binary distribution

Yamashita and Suzuki [2] proposed the following definition of principal points for binary distributions:

Let \( x_i \in \{0, 1\}^p \ (i = 1, ..., 2^p) \) be a set of realization vectors of the multivariate binary distribution \( F \), whose random vector is \( X \), and \( k \) binary vectors be \( y_j \in \{0, 1\}^p \ (j = 1, ..., k) \). Then, the distance between \( \xi_j \in \{0, 1\}^p \) \((j = 1, ..., k) \) and \( y_j \in \{0, 1\}^p \ (j = 1, ..., k) \) is defined as in Section 2.1. Thus, \( k \)-principal points for binary distributions are given as the set of vectors \( \xi_j \) \((j = 1, ..., k) \), which satisfies the following:

\[
E_F \{ d^2(X | \xi_1, ..., \xi_j) \} = \min_{y_j \in \{0, 1\}^p, \xi_j \in \mathbb{R}^n} E_F \{ d^2(X | y_1, ..., y_j) \}
\]

\[
= \min_{y_j \in \{0, 1\}^p, \xi_j \in \mathbb{R}^n} \sum_{i=1}^{2^p} P[X=x_i] \{ d^2(x_i | y_1, ..., y_j) \}
\] (3)

Here, the computational complexity in finding the \( k \)-principal points for binary data is represented as follows:

\[
\binom{2^p}{k} = \frac{2^p!}{k!(2^p-k)!}
\]

If the combinations of \( k \) candidates out of \( 2^p \) are considered, the number of their combinations increases exponentially with an increase in \( p \). Therefore, this problem is regarded as a NP-hard problem. The algorithms proposed by Yamashita and Suzuki [2,3] were based on a k-means algorithm and a greedy algorithm, respectively.

2.3 The definition of \( l \)-k-principal points for multivariate binary distributions

In analysis of multivariate binary data, there may be many occurrences \( l \) when a response “1” is given. For example, when we analyze multiple choice questionnaire survey data in which the number of selected items is predetermined, dummy variables data including categories of sex, blood type, and occupation, it is useful to find binary “principal” points where the number of occurrences \( l \) of response “1” is set. Yamashita and Suzuki [2] proposed a special definition for \( k \)-principal points of binary distributions when the number of occurrences \( l \) of response “1” is given, and these points were referred to as “\( l \)-principal points.”

Let \( x_i \in \mathcal{S} \subseteq \mathbb{R}^p \ (i = 1, ..., \pi) \) be a set of realization vectors of distribution \( F \), whose random vector is \( X \), and \( F^r \) is the distribution, whose number of occurrences is set, where

\[
\pi = \left( \frac{p}{l} \right)
\]

\[
\frac{p!}{l!(p-l)!}.
\]

\( \mathcal{S} \) denotes the set of realizations of \( F^r(\mathcal{S}=\{x \in \{0,1\}^p | x^T 1 = l\}) \).

Next, let \( y_j \ (j = 1, ..., k) \) be a set of \( k \) binary vectors containing \( l \) elements with a value of 1. We measure the distance between \( x_i \) and \( y_j \) according to (3). Then, \( \xi_j \in \mathcal{S} \) \((j = 1, ..., k) \) can be a set of \( l \)-k-principal points of distribution \( F^r \), if \( \xi_j \) satisfies the following:

\[
E_F \{ d^2(X | \xi_1, ..., \xi_j) \} = \min_{y_j \in \{0, 1\}^p, \xi_j \in \mathbb{R}^n} \sum_{i=1}^{2^p} P[X=x_i] \{ d^2(x_i | y_1, ..., y_j) \}.
\]

2.4 Principal points for binary distribution considering external criterion

2.4.1 The definition of \( k \)-principal points for binary distribution considering external criterion

Here, we show the definition of principal points for multivariate binary distribution, where an external criterion is used to characterize the data. Yamashita and Suzuki [4] proposed the definition by restricting the number of distributions mixed to two (i.e., \( F_1 \) and \( F_2 \)); \( w_1 \) and \( w_2 \) are the weights of the distributions \( F_1 \) and \( F_2 \) in the mixed distribution, respectively.

The distribution \( F \) of the given data is assumed to be the mixture of distributions \( F_n \), as follows.

International Journal of Japan Association for Management Systems
F = w_1 F_1 + w_2 F_2,  w_1 + w_2 = 1

Here, the weights of the distributions \( w_1 \) and \( w_2 \) are given.

Now, the definition of the \( k \)-principal points for a multivariate binary distribution considering an external criterion is shown as follows:

We denote the random variable of the random distribution \( F_i \) by \( X^{[i]}(s = 1,2) \), and the realizations of each distribution by \( X^{[i]}(s = 0,1)^r \). Subsequently, we denote the candidate principal points of the random distribution \( F_i \) by \( \xi^{[i]}(s = 0,1)^r \), and the candidate principal points of \( F_2 \) by \( \eta^{[i]}(s = 0,1)^r \). Moreover, we assume that there is at least one principal point in each distribution in the model.

Then, the difference of the two squared distances between the realization and the \( k \) candidate vectors for the principal points \( \Delta d^2 \) is:

\[
\Delta d^2(x^{[1]}_1, ..., y^{[1]}_1, ..., y^{[2]}_1, ..., y^{[2]}_1) = \min_{a \in \{0,1\}} (x^{[1]}_i - y^{[1]}_i)^2 - \min_{a \in \{0,1\}} (x^{[2]}_i - y^{[2]}_i)^2
\]

Then, the principal points \( \xi^{[1]}_1, ..., \xi^{[1]}_1, \xi^{[2]}_1, ..., \xi^{[2]}_1 \) of distribution \( F \) comprising two distributions \( F_1 \) and \( F_2 \) is defined as follows:

\[
E[\Delta d^2(x^{[1]}_1, ..., y^{[1]}_1, ..., y^{[2]}_1, ..., y^{[2]}_1)] = w_1 E[\Delta d^2(x^{[1]}_1, ..., \xi^{[1]}_1, \xi^{[2]}_1, ..., \xi^{[2]}_1)] + w_2 E[\Delta d^2(x^{[1]}_1, ..., \xi^{[1]}_1, \xi^{[2]}_1, ..., \xi^{[2]}_1)]
\]

The details are described in subsection 2.4.2. Moreover, if there are more than two principal points that have the same distance from a realization, we do not classify the realization.

2.4.2 The evaluation of \( k \)-principal points for binary distribution considering external criterion

Conditional ratio of correct discrimination

Yamashita and Suzuki [4] proposed the evaluation of \( k \)-principal points for binary distribution considering an external criterion.

Let \( \xi^{[1]}_1, ..., \xi^{[1]}_1, \xi^{[2]}_1, ..., \xi^{[2]}_1 \) be the set of \( k \)-principal points of the multivariate binary distribution \( F \) defined in the equation (4). Then, let \( p_u(s = u) \) be the ratio that the \( j^{th} \) realization vector included in class \( s \) be the same as the decision of the class \( u \) of the proposed model, which is the class decided by the division of the distribution using principal points.

Consequently, let \( p_u(s = u) \) be the ratio of the probability of the region divided by \( j^{th} \) principal points in class \( s \) is different from the decision of the class \( u \) of the proposed model (4), which is the class decided by the division of the distribution using principal points. Then, the conditional ratio of correct discrimination is formulated as \( p_u(s = u) \), and we determine the principal points representing greatly or not based on this conditional probability. The conditional ratio of correct discrimination and misclassification is shown in Table 1.

Ratio of correct classification and misclassification

The ratio of correct classifications to misclassifications enables the evaluation of how well a set of principal points for multivariate binary distribution considering an external criterion represents the distribution. Here, the total correct classification ratio, i.e., a set of principal points that discriminates each realization correctly is formulated as follows:

\[
\eta = w_1(p_{11} + \cdots + p_{aa}) + w_2(p_{a+12} + \cdots + p_{k2}) (5)
\]

Moreover, the whole misclassification ratio is formulated as follows:

\[
\delta = w_1(p_{11} + \cdots + p_{aa}) + w_2(p_{a+12} + \cdots + p_{k2}) (6)
\]

Note that, if there are more than two principal points that are same distance from a realization, we do not classify the realization; therefore, the sum of the ratio of correct classification and misclassification is not always 1.
3. An algorithm for principal points considering external criterion for multivariate binary distributions

As mentioned in subsection 2.2, find principal points for multivariate binary distributions, we must solve an NP-hard problem. Moreover, the problem of principal points considering external criterion for multivariate binary distributions requires heavy computation.

For the algorithm of principal points for a multivariate binary distribution, Yamashita and Suzuki [2] proposed the greedy algorithm, and they showed that the approximations feature submodularity [6], based on the greedy algorithm, and guarantee \((1 - 1/e)\) times the optimal solution. This algorithm can be adopted to solve the problem of principal points considering external criterion for multivariate binary distribution. The simple strategy is as follows: first, we find a principal point for each distribution and set all the identified points as principal points. Then, we add a principal point that minimizes the function with the set principal points one by one. The detailed steps are described as algorithm 1.

**STEP1** We define \(\mathcal{E}_t = \{e_1^{(t)}, \ldots, e_2^{(t)}\}\) as a set of \(t\)-principal points considering the external criterion for a multivariate binary distribution, and set \(t = 2\). One principal point is taken from each distribution.

**STEP2** We calculate the following optimization problem

\[
\min_{x^{(t)}} \left\{ \sum_{i=1}^{k} w_i \mathbb{P}(x^{(t)} = x^{(i)}) \Delta d^2(X^{(t)}|s_1^{(i)}, \ldots, s_k^{(i)}, y_1^{(i)}, \ldots, y_r^{(i)}) + \sum_{i=1}^{k} w_i \mathbb{P}(x^{(t)} = x^{(i)}) \Delta d^2(X^{(t)}|s_1^{(i)}, \ldots, s_k^{(i)}, y_1^{(i)}, \ldots, y_r^{(i)}) \right\}
\]

Where,

\[
\Delta d^2(X^{(t)}|s_1^{(i)}, \ldots, s_k^{(i)}, y_1^{(i)}, \ldots, y_r^{(i)}) = \min_{\alpha_{t+2}} \left\{ \min_{\alpha_{t+1}} \left( \min_{\alpha_{t+1}} (s_j^{(i)} - y_j^{(i)}) \right) \right\} - \min_{\alpha_{t+2}} \left\{ \min_{\alpha_{t+1}} \left( \min_{\alpha_{t+1}} (s_j^{(i)} - y_j^{(i)}) \right) \right\} \]

**STEP3** Add \(y^{(t)}\) to a set of \(\mathcal{E}_t = \{e_1^{(t)}, \ldots, e_2^{(t)}\}\), and let \(t = t + 1\). Then, we arrange \(\mathcal{E}_{t+1} = \{e_1^{(t+1)}, \ldots, e_2^{(t+1)}\}\). For example, if \(t = 4\), \(\mathcal{E}_4 = \{e_1^{(4)}, \ldots, e_2^{(4)}\}\), and the class of the fourth principal point considering external criterion is 1, then we have to arrange \(\mathcal{E}_t = \{e_1^{(t)}, \ldots, e_3^{(t)}\}\).

**STEP4** We repeat STEPs 2 and 3 until \(t = k\).

This algorithm enables us to reduce the number of combinations to be considered as follows:

\[k2^{p} + (1 + k)/2\]

Note that since the problem of principal points considering an external criterion does not have submodularity, we cannot guarantee the optimality of the approximations.

4. An example of application of principal points considering external criterion for multivariate distributions

4.1 Data description and analysis

In this section, we present a real-world data analysis based on \(k\)-principal points considering an external criterion for multivariate binary distributions. The data were obtained from Japanese local railway records ([7]-[12]).

In Japan, there is a problem of the doughnut phenomenon, that is, the population in countryside is continuously decreasing. Consequently, the number of users of the local railroad is decreasing, and many railroads have been closed; however, there are also local railroads that remain functional. The dataset includes the features of the abolished railroads and the functioning railroads. This dataset includes information (i.e., 6 variables) on each 38 railroads (25 railroads are still functioning; class1, and others have abolished; class2) as follows:

(i) Are there more than two railroads connected to a local railroad?
(ii) Is there a city (or cities) where only the railroad exists?
(iii) Are there sight(s) in the area across which the railroad runs?
(iv) Is the railroad run by a big company?
(v) Is there another railroad competing with this railroad?
Does the railroad run within 50 km from the urban city? We analyze the data based on the Tarpey and Petkova’s method [5], principal points considering external criterion \([4]\) and the approximation of \([4]\) using proposed algorithm shown in Section 3. We also compare the results in terms of accuracy of the solution, computational time, and the conditional probability, and we discuss each method based on its results. Note that since the proposal is an approximation method of \([4]\), the value of expected error of principal points found by the proposed method should be worse than the optimal solutions.

4.2 Results of an analysis
We describe the results obtained by each algorithm in Tables 2, 3, 4, and 5.

Table 2 Principal points and expected squared distance

<table>
<thead>
<tr>
<th>Tarpey and Petkova’s method [5]</th>
<th>(E[\Delta d^2])</th>
</tr>
</thead>
<tbody>
<tr>
<td>(k)</td>
<td>Class</td>
</tr>
<tr>
<td>4</td>
<td>Abolished</td>
</tr>
<tr>
<td></td>
<td>Keep running</td>
</tr>
<tr>
<td>5</td>
<td>Abolished</td>
</tr>
<tr>
<td></td>
<td>Keep running</td>
</tr>
</tbody>
</table>

Table 3 Conditional correct probability of each principal point

<table>
<thead>
<tr>
<th>Tarpey and Petkova’s method [5]</th>
</tr>
</thead>
<tbody>
<tr>
<td>(k)</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>Conditional</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>Conditional</td>
</tr>
</tbody>
</table>

Table 4 Accurate and error ratio of three methods

<table>
<thead>
<tr>
<th>Tarpey’s method</th>
</tr>
</thead>
<tbody>
<tr>
<td>(k)</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

Table 5 Computational time for three methods

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>9m48s</td>
<td>2h28m10s</td>
<td>4.25s</td>
</tr>
<tr>
<td>5</td>
<td>15h45s</td>
<td>8h32m08s</td>
<td>5.13s</td>
</tr>
</tbody>
</table>
Table 4 shows that the accuracy of discrimination based on the proposed method (i.e., the solution is an approximation of principal points considering an external criterion for multivariate binary distribution) is the best of the three methods. Indeed, we were able to identify principal points showing the feature of each distribution (class) well. In Table 5, we found the efficiency of the proposed method in terms of the computational complexity. Overall, our proposal found preferable solutions, and it enabled us to show the effectiveness of the proposed approximation method.

Note that, the interpretation of principal points found by proposed method for the keep running rail roads (1,1,1,1,1,1), (1,1,1,0,0,1) and (1,0,0,1,1,1) is as follows: the first principal point is the good conditioned local rail roads. The second and third points are the principal points of rail roads that make the effort for keeping them functional. For example, there is a rail road building the anime character painted train car and make the rail road itself to be sightseeing places. The principal points for the abolished railroad are also fit to our experimental knowledge, and then we conclude the adequacy of our proposal for analyzing this data.

5. Conclusion

In this study, we proposed an algorithm for finding principal points by considering an external criterion for multivariate binary distributions based on the greedy algorithm.

Our future work is to analyze more datasets than the datasets analyzed in this study and show the adequacy of our proposal and to derive the numerical simulation. Moreover, we did not discuss the theoretical feature of our proposal. These topics are going to be our future work.

References
[7] Horiguchi, S. : Abolishment of Rail Road and Bus as the Alternative Transportation , Publisher of Tokyodo (2010), in Japanese

Acknowledgments
The authors would like to express our thanks to Professor Suzuki in Keio University for his helpful comments for this research. This study was supported by JSPS KAKENHI, Grant Number 16K16361.