Principal Subsets Analysis

Maryam Tayefi*and Sharad Gore†
Department of Statistics and Centre for Advanced Studies
University of Pune, Pune 411 007, INDIA

Abstract

Multivariate data are difficult to handle due to the so-called curse of dimensionality. Researchers have developed methods for reducing the dimensionality of multivariate data with help of mathematical transformations. Principal components analysis, factor analysis, and independent components analysis are examples of such methods that reduce the data dimension. One of the major drawbacks of these methods is that the resulting dimensions are difficult to interpret because they are mathematical constructs and are not observed characteristics of the population units.

A new method for reducing the data dimension by forming groups of highly correlated dimensions is introduced in this paper. It uses the correlation coefficient as the measure of association between the dimensions of data elements and forms a partition of these dimensions. Since no mathematical transformation is involved, the dimensions continue to be the same as observed and hence the results are interpretable. How to use the partition for processing the data further is a question addressed in this paper. An illustrative example is given to explain and demonstrate the new method, called the Principal Subsets Analysis (PSA).

Keywords: *Curse of dimensionality; Reduction of dimension; Principal components analysis; Principal subsets analysis*

1 Introduction

Multivariate data usually suffer from the curse of dimensionality. As a consequence, either the data is not utilized to the fullest extent or the analysis is oversimplified in order to make it comprehensive. In either case, the knowledge hidden in the data is hardly discovered to a satisfactory level without any ambiguity. The remedy suggested in such situations usually involves a reduction in data dimension. This may be achieved by principal components analysis (PCA) or factor analysis (FA). A new method of dimension reduction is proposed in this paper that partitions the variables in the dataset in such a way that the variables within every partition set are maximally correlated with one another. The new method selects variables in such a way

*tayefi.maryam@gmail.com
†Corresponding author: sdgore@stats.unipune.ac.in

that successive linear relationships involve mutually exclusive sets of variables. These sets are called principal subsets on the lines of principal components. Each principal subset contains variables that are maximally correlated with each other. Unlike principal components, principal subsets are not uncorrelated with one another. Another advantage of the new method is that it offers an easy interpretation of the results. The principal subsets are obtained by partitioning the entire set of variables. As a consequence, the variables in every subset are in their original and natural form, and not in the form of a transformation.

The principal subsets algorithm is developed from the partitional clustering algorithm of Tayefi and Gore (2013). The only change is that the principal subsets analysis uses the correlation matrix, and not the distance matrix. Section 2 describes all the steps of the principal subsets analysis algorithm. Section 3 contains illustrative examples to show how the principal subsets analysis is carried out and how the results are interpreted.

2 Principal Subsets Analysis

Principal Subsets Analysis (PSA) is a non-iterative and non-recursive method of partitioning the set of variables in a multivariate dataset. The partitioning is done in such a way that, beginning with two variables that have the highest magnitude of correlation coefficient in a set initially, another variable is added to the set if one of the variables already in the set has the largest magnitude of correlation coefficient with the variable under consideration. When none of the variables in the set satisfies this criterion, formation of the set is complete and another set is formed by identifying variables that are not yet included in any set and have the highest magnitude of the correlation coefficient. The new set is formed in the same way, and the process continues until all variables are included in one of the sets. These sets are called principal subsets. It may be noted that the proposed method of forming a set of variables ensures that every set contains at least two variables. Once the principal subsets are identified, there may be different ways of representing these subsets by one variable each, thereby achieving a reduction in the dimension of the data. The major difference between principal components analysis and the proposed principal subsets analysis is regarding the number of variables involved in the transformations. While the former involves all the variables in the computation of every principal component, the latter involves only the variables that belong to the particular subset under consideration.

3 The PSA Algorithm

The data matrix $\boldsymbol{X}$ is formed by arranging the vectors of observations $\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_n$ as its rows. The association between two variables is measured by the correlation coefficient between them. All the correlation coefficients are computed and arranged in the form of the correlation matrix of order $P \times P$. The correlation matrix $\boldsymbol{R}$ is symmetric positive definite and has unity along the main diagonal. We use the correlation matrix to partition the set of variables in to principal subsets. The number and sizes of partition sets are not pre-determined and they emerge as the partition sets are formed sequentially. Even then, these partition sets are unique for a given dataset in the sense that rearranging the data by permuting observations and/or variables does not affect the final result. In this sense, the principal subsets are unique.

The steps of the PSA algorithm are as follows.

Step 1. Initialization.

In every column of the correlation matrix, mark the largest non-diagonal entry, so that every column has at least one marked entry. Multiple marking is allowed in case of a tie. Ignore the sign and use only the magnitude for this purpose. For $j = 1, 2, \cdots, P$, let R_j denote the row that has a marked entry in column number j.

Step 2. Forming a new partition set.

Suppose $r_{i^*,j^*} = \max_{i,j}\{r_{i,j}\}$ so that variables V_{i^*} and V_{j^*} have the highest magnitude of correlation between them. We initiate formation of a new partition set $S = \{i^*, j^*\}$ by including these two variables in the set.

Step 3. Checking completion of partition set.

In the correlation matrix, check if there are any marked entries in rows numbered i^* and j^*. If there is no marked entry in either of these rows, then follow Step 4. Otherwise follow Step 6.

Step 4. Completion of a partition set.

The partition set S is complete in the sense that there are no more variables that can be entered in it. Update the correlation matrix $\boldsymbol{R}$ by removing rows and columns corresponding to variables in S.

Step 5. Checking for termination of the algorithm.

Check if the correlation matrix has at least two rows and at least two columns. If so, follow Step 2 to form a new partition set. Otherwise, there are no more variables to partition. Hence, the algorithm terminates

Step 6. Entering variables in the partition set S.

Let C_{i^*} denote the column of the correlation matrix having a marked entry in row numbered

i^*. Add the variable $V_{C_{i^*}}$ to the partition set S. This procedure is followed for every marked entry in row numbered i^*.

Step 7. Entering more variables in the partition set S.

Let C_{j^*} denote the column of the correlation matrix having a marked entry in row numbered j^*. Enter the variable $V_{C_{j^*}}$ in the partition set S. This procedure is followed for every marked entry in row numbered j^*.

Step 8. Entering more variables in S on the basis of variables entered in Step 6 and Step 7. For every variable numbered k^* that is entered in the partition set S in step 6 and step 7, check if there is any marked entry in row numbered k^* and enter the corresponding variable in the partition set S.

Step 9. Repeat Step 8 as long as at least one marked entry is found in the correlation matrix in any row corresponding to a variable already in the partition set S. When there are no more marked elements, declare completion of the partition set S.

Step 10. Update the correlation matrix $\boldsymbol{R}$ by removing rows and columns corresponding to variables in the partition set S.

Step 11. The algorithm is complete and has partitioned the set of variables.

4 An Illustrative Example: The `Chekin` data from Yelp Dataset Challenge

This example relates to the dataset `Chekin` from the Yelp Dataset Challenge. The dataset contains information on the hourly number of chekins. The format of data is such that there are three entries for every chekin, namely the day of the week, the hour of the day and the number of chekins. There is no entry if there is no chekin in a specific day-hour combination.

The data is first reformatted to avoid ambiguities and inconsistencies. In the original form, the dataset has rows of unequal length. This causes a problem in handling data for statistical analysis. The maximum number of columns is found to be 501. This is reduced to 168 hours in a week. As a result, the dataset is reduced to have 8282 rows and 168 columns. Every column denotes a fixed hour of the specific day of the week. Moreover, the successive columns are also in a chronological order, making it easy to interpret data and results of the analysis.

The principal subsets analysis algorithm partitions the 168 variables into a total of 35 partition sets. These 35 partition sets are called the principal subsets of the given set of variables.

For comparison, we carry out PCA and select the first 35 principal components. The variances of the first 35 principal components are compared with the variances of the first principal

component of each of the 35 principal subsets.

The amount of information in the first 35 principal components is 152.1542. In comparison, the sum of the information contained in the first principal components of the 35 principal subsets is 139.8152.

In principal subsets, the variables within a subset are correlated. We retain correlations between variables that belong to the same subset and ignore correlations between variables that belong to different subsets. This results in a block diagonal matrix as described below.

The correlations between variables belonging to different principal subsets are not present in this matrix. It is therefore interesting to compare this matrix with the original correlation matrix in order to find out the effect of defining principal subsets.

The determinant of correlation matrix contains information on mutual dependence among the variables. More the dependence between variables, smaller is the determinant. The determinant of the original correlation matrix is 1.046126e-133 and the determinant of the block diagonal matrix is 4.322087e-91 .

It is interesting to note that the principal subsets analysis has reduced the data dimension from 168 to 35, implying that there is a reduction of almost 80 percent in the data dimension. However, the amount of information contained in the first 35 principal components is 152.1542, indicating that the loss of information is not even 10 percent. These two numbers indicate the effectiveness of the principal components analysis. Further, the principal subsets analysis produces 35 principal subsets and the sum of the variances of the first principal components of these 35 principal subsets is 139.8152, which makes it 83.2233 percent of the total information. Again, even though the principal subsets analysis contains smaller amount of information in comparison to the principal components analysis, the amount of work in computing principal subsets is substantially smaller than that required for computing principal components.

References

1. Tayefi, M., Gore, S. D. (2013). *Distance-based partition of multivariate data.* In Proceedings ICCS-12, Vol. 23, pp. 535-542. Proceedings of the 12th Islamic Countries Conference on Statistical Sciences, December 19-22, 2012 at Qatar University, Doha, Qatar. ISBN: 978-969-8858-11-7.

2. Keogh, E., zhu, Q., Hu, B., Hao, Y., Xi, X., Wei, L., Ratanamahatana, C. A. (2011). The UCR Classification/Clustering Homepage: `www.cs.ucr.edu/~eamonn/time_series_data/`.

3. Olszewski, R. T. (2001). *Generalized feature extraction for structural pattern recognition in time series data*. Ph. D thesis Carnegie Mellon University.

4. Yelp Dataset Challenge (2013). `https://www.yelp.com/dataset_challenge/dataset`