Chapter 9
Principal Component Analysis

9.1 Introduction

Under the high-dimensional setup with p variables, the problem that often arises is the critical nature of the correlation or covariance matrix. When p is moderately or very large it is generally difficult to identify the true nature of relationship among the variables as well as observations from the covariance or correlation matrix. Under such situations, a very common way to simplify the matter is to reduce the dimension by considering only those variables (components) those are truly responsible for the overall variation.

Principal component analysis (PCA) is a dimension reduction procedure. PCA was developed in 1901 by Karl Pearson, as an analogue of the principal axis theorem in mechanics. It was later independently developed by Harold Hotelling (1933, 1936). Several authors considered PCA in different forms (Joliffe 1982, 2002). There are several case studies and applications (Jeffers 1967; Chattopadhyay and Chattopadhyay 2006). The method is useful when we have a large number of variables, and some variables are of less or no importance. In this case, redundancy means that some of the variables are highly correlated with one another, possibly because they are measuring the same phenomenon. Because of this redundancy, it should be possible to reduce the observed variables into a smaller number of principal components (derived variables) that will account for most of the variance in the observed variables.

Being a dimension reduction technique, principal component analysis has similarities with exploratory factor analysis. The steps followed when conducting a principal component analysis are almost the same as those of exploratory factor analysis. However, there are significant conceptual differences between the reduction procedure that gives a relatively small number of components those account for most of the variance in a set of observed variables. In summary, both factor analysis and principal component analysis have important roles to play in social science research, but their conceptual foundations are quite different.

S. P. Mukherjee et al., *Statistical Methods in Social Science Research*,
https://doi.org/10.1007/978-981-13-2146-7_9

More recently, Independent component analysis (ICA) has been identified as a strong competitor for principal component analysis and factor analysis. ICA finds a set of source data that are mutually independent (not only with respect to the second moment), but PCA finds a set of data that are mutually uncorrelated and the principal components become independent only under Gaussian setup. ICA was primarily developed for non-Gaussian data in order to find independent components responsible for a larger part of the variation. ICA separates statistically independent original source data from an observed set of data mixtures.

9.1.1 Method

In PCA, primarily[1] it is not necessary to make any assumption regarding the underlying multivariate distribution but if we are interested in some inference problems related to PCA then the assumption of multivariate normality is necessary (Chattopadhyay and Chattopadhyay 2014). The eigenvalues and eigenvectors of the covariance or correlation matrix are the main contributors of a PCA. Of course, the eigenvalues of covariance and correlation matrices are different and they coincide when we work with standardized values of the variables. So the decision whether one should start work covariance or correlation matrix is important. Normally, when all the variables are of equal importance, one may start with the correlation matrix. The eigenvectors determine the directions of maximum variability, whereas the eigenvalues specify the variances. In practice, decisions regarding the quality of the principal component approximation should be made on the basis of eigenvalue–eigenvector pairs. In order to study the sampling distribution of their estimates, the multivariate normality assumptions became necessary as otherwise it is too difficult. Principal components are a sequence of projections of the data. The components are constructed in such a way that they are uncorrelated and ordered in variance. The components of a p-dimensional data set provide a sequence of best linear approximations. As only a few of such linear combinations may explain a larger percentage of variation in the data, one can take only those components instead of p variables for further analysis.

A PCA is concerned with explaining the variance–covariance structure through a few linear combinations of the original variables. Its general objectives are data reduction and interpretation. Reduce the number of variables from p to $k < (kp)$. Let the random vector $X' = (X_1 \dots X_p)$ have the covariance matrix Σ (or correlation matrix R) with ordered eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p \geq 0$ and corresponding eigenvectors $e'_1, e'_2, \dots, e'_p$, respectively.

[1] This section draws from one of the authors' published work, 'Statistical Methods for Astronomical Data Analysis,' authored by Asis Kumar Chattopadhyay and Tanuka Chattopadhyay, and published in 2014 by Springer Science+Business Media New York.

Consider the linear combinations

$$
\begin{aligned}
Y_1 &= l_{11}X_1 + l_{21}X_2 + \cdots + l_{p1}X_p = e_1'X \\
Y_2 &= l_{12}X_1 + l_{22}X_2 + \cdots + l_{p2}X_p = e_2'X \\
&\cdot \\
&\cdot \\
Y_p &= l_{1p}X_1 + l_{2p}X_2 + \cdots + l_{pp}X_p = e_p'X
\end{aligned}
$$

Then we have the following result:

Result: Let $X' = (X_1 \ldots X_p)$ have covariance matrix Σ with eigenvalue–eigenvector pairs $(\lambda_1, e_1) \ldots (\lambda_p, e_p)$ where $\lambda_1 \geq \lambda_2 \cdots \geq .\lambda_p \geq 0$.

Let $Y_1 = e_1'X, Y_2 = e_2'X \ldots Y_p = e_p'X$.

Then

$$\text{var}(Y_i) = \lambda_i (i = 1, 2, \ldots p) \text{ and}$$

$$
\begin{aligned}
\sigma_{11} + \sigma_{22} \cdots + \sigma_{pp} &= \sum\nolimits_1^p \text{var}(Xi) \\
&= \lambda_1 + \cdots + \lambda_p \\
&= \sum\nolimits_1^p \text{var}(Yi)
\end{aligned}
$$

Here $Y_1, Y_2, \ldots, Y_p$ are called **principal components**. In particular, Y_1 is the **first principal component** (having the largest variance), Y_2 is the second principal component (having the second largest variance), and so on.

(For proof of the above result, one may consult any standard textbook.)

Here instead of original p variables $X_1 \ldots X_p$, only a few principal components $Y_1, Y_2, \ldots, Y_k (k < p)$ are used which explains maximum part of the total variation. There are several methods to find the optimum value of k.

The specific aim of the analysis is to reduce a large number of variables to a smaller number of components by retaining the total variance (sum of the diagonal components of the covariance matrix) almost same among the observations. The analysis therefore helps us to determine the optimum set of artificial variables (viz. linear combinations) explaining the overall variations in the nature of objects.

Many criteria have been suggested by different authors for deciding how many principal components (k) to retain. Some of these criteria are as follows:

1. Include just enough components to explain some arbitrary amount (say 80%) of the total variance which is the sum of the variances (diagonal elements of the covariance matrix) of all the variables.
2. Exclude those principal components with eigenvalues below the average. For principal components calculated from the correlation matrix, this criterion excludes components with eigenvalues less than 1.

3. Use of the screen plot (plotting eigenvalues against components) technique.[2]

Example 9.1.1 (http://openmv.net/) The following data set gives the relative consumption of certain food items in European and Scandinavian countries. The numbers represent the percentage of the population consuming that food type corresponding to 15 countries and 9 food types. As there are 9 food types corresponding to only 15 countries, it is necessary to reduce the dimension in order to search for major food types.

	Instant			Powder
Country	coffee	Tea	Biscuits	soup
Germany	49	88	57	51
Italy	10	60	55	41
France	42	63	76	53
Holland	62	98	62	67
Belgium	38	48	74	37
Luxembourg	61	86	79	73
England	86	99	91	55
Portugal	26	77	22	34
Austria	31	61	29	33
Switzerland	72	85	31	69
Denmark	17	92	66	32
Norway	17	83	62	51
Finland	12	84	64	27
Spain	40	40	62	43
Ireland	52	99	80	75

		Frozen			
Country	Potatoes	fish	Apples	Oranges	Butter
Germany 21	27	81	75	91	
Italy	2	4	67	71	66
France	23	11	87	84	94
Holland	7	14	83	89	31
Belgium	9	13	76	76	84
Luxembourg	7	26	85	94	94
England	17	20	76	68	95
Portugal	5	20	22	51	65
Austria	5	15	49	42	51
Switzerland	17	19	79	70	82
Denmark	11	51	81	72	92
Norway	17	30	61	72	63
Finland	8	18	50	57	96
Spain	14	23	59	77	44
Ireland	2	5	57	52	97

[2] A significant part of 'Chattopadhyay and Chattopadhyay (2014). Statistical methods for Astronomical Data Analysis, Springer Series in Astrostatistics, Springer' is reproduced in this part.

Table 9.1 Eigen analysis of the correlation matrix

Components	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Eigenvalue	3.4129	1.5591	1.3412	1.0164	0.7587	0.3294	0.2633	0.2027	0.1162
Proportion	0.379	0.173	0.149	0.113	0.084	0.037	0.029	0.023	0.013
Cumulative	0.379	0.552	0.701	0.814	0.899	0.935	0.965	0.987	1.000

Table 9.2 Coefficients of 15 variables in 9 principal components

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Instant	0.383	−0.367	−0.035	−0.247	−0.335	−0.360	−0.573	0.140	0.258
Tea	0.241	−0.229	0.596	−0.326	−0.274	0.000	0.441	0.304	0.256
Biscuits	0.368	0.062	0.069	0.571	−0.249	−0.650	0.135	−0.058	−0.155
Powder s	0.389	−0.467	−0.053	−0.170	−0.047	0.225	0.119	−0.561	−0.467
Potatoes	0.284	0.412	−0.078	−0.219	0.667	−0.144	0.476	−0.058	−0.032
Frozen f	0.079	0.575	0.305	−0.452	−0.298	−0.155	−0.377	−0.248	−0.221
Apples	0.465	0.174	−0.205	0.057	−0.108	0.360	−0.123	0.632	−0.389
Oranges	0.394	0.208	−0.424	−0.013	−0.340	0.201	0.058	−0.258	0.629
Butter	0.224	0.134	0.561	0.471	0.297	0.427	−0.240	−0.197	0.169

From the screen plot and Table 9.1, it is clear that **4 components** have variances (i.e., eigenvalues of the correlation matrix) **greater than one** and these four components explain **more than 80 % of the total variation, i.e., the sum of the variances of all the variables**. Hence, one can work with four principal components instead of the original nine variables.

From Table 9.2, it is clear that most of the variables have similar importance in all the first four components so that it is difficult to associate a particular component to a subset of variables. So here it is not possible to identify the physical nature of the components. This feature is generally true for principal component analysis. In order to find inherent factors, one can take help of factor analysis if the nature of the covariance matrix admits (Figs. 9.1 and 9.2).

9.1.2 The Correlation Vector Diagram (Biplot, Gabriel 1971)

A matrix of rank 2 can be displayed as a biplot consisting of a vector for each row and a vector for each column, chosen so that each element of the matrix is exactly the inner product of the vectors corresponding to its row and its column (Gabriel 1971). If a matrix is of higher rank, one may display it approximately by a biplot of a matrix of rank 2 that approximates the original matrix. In PCA, a biplot can show inter-unit distances and indicate the clustering of units, as well as displaying the variances and correlations of the variables.

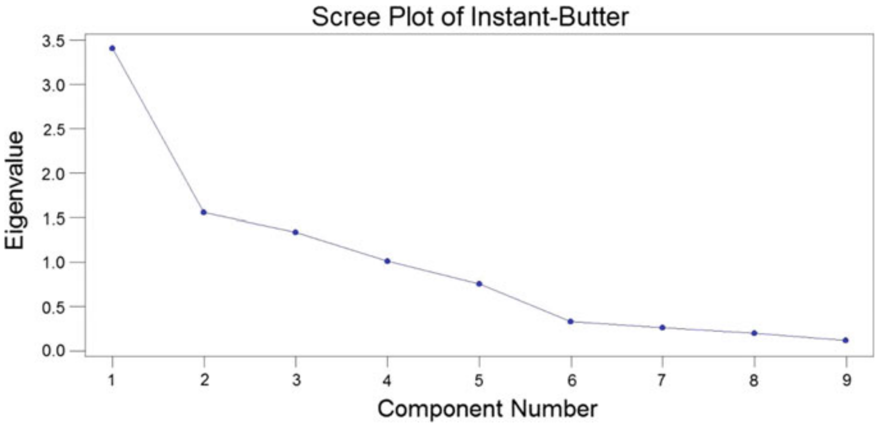

Fig. 9.1 Screen plot used to decide about the number of significant principal components. The components with eigenvalues greater than 1 are usually taken as significant

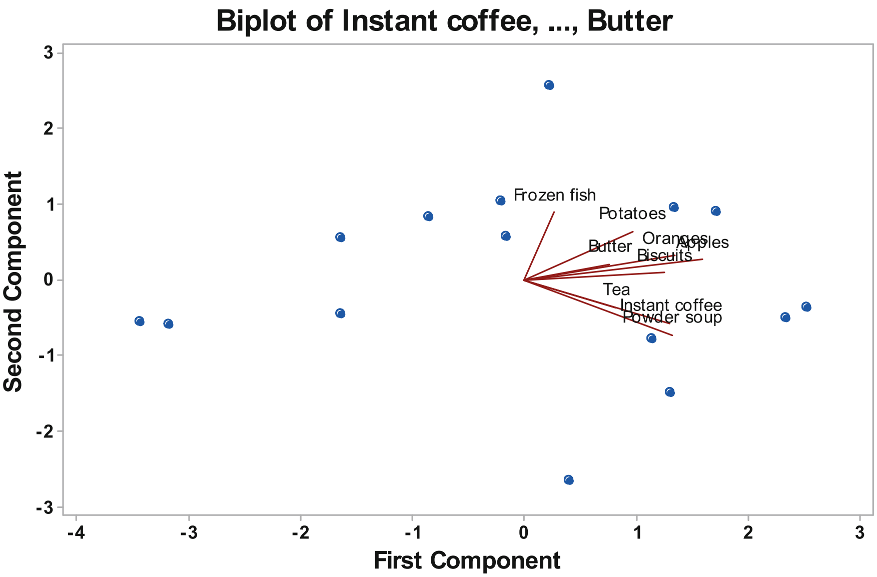

Fig. 9.2 Biplot for the data used in Example 9.1.1. The vector lengths represent variances of corresponding variables, and the angles show correlations of the variables (smaller angles indicate higher correlations). Dot points indicate the positions of the 15 countries with respect to their first and second component values. The origin represents the average value for each variable; that is, it represents the object that has an average value in each variable

Any matrix of observations y of order mXn can be written by singular value decomposition as

$$y = \Sigma_1^r \lambda_i piqi'(\lambda_1 \geq \lambda_2 \cdots \geq \lambda_r)$$

where r is the rank of the matrix y and λ_i, p_i, and q_i' are the singular value, singular column, and singular row, respectively. Then by the method of least-squares fitting a matrix of rank 2, an approximation of y is given by

$$y = \Sigma_1^2 \lambda_i piqi'$$

and the corresponding measure of goodness of fit is given by

$$\rho(2) = \frac{\lambda_1^2 + \lambda_2^2}{\sigma_1^r \lambda i^2}$$

If $\rho(2)$ is near to 1, then such a biplot will give a good approximation to y. If we denote by

$$\begin{aligned} \boldsymbol{S}^{mXm} &= (1/n)\boldsymbol{y}'\boldsymbol{y} = (s_{ij}) = \text{variance–covariance matrix and} \\ \boldsymbol{R}^{mXm} &= (r_{ij}) = \text{correlation matrix} \end{aligned}$$

then it can be shown that

$$\boldsymbol{y}^{nXm} \sim \boldsymbol{G}^{nX2} \boldsymbol{H}'^{2Xm}$$

where

$$\boldsymbol{G}^{nX2} = (p1'p2')\sqrt{n} = (g_1^{nX1} g_2^{nX1})$$

and

$$\boldsymbol{H}^{mX2} = \left(\frac{1}{\sqrt{n}}\right)(\lambda_{1q1}\lambda_{2q2}) = (h_1^{mX1} h_2^{mX1}).$$

Further,

$$s_{ij} \sim h_i' h_j$$

$$s_j^2 \sim ||h_j||^2$$

$$r_{ij} \sim \cos(h_i h_j).$$

9.2 Properties of Principal Components

In PCA, the first component extracted explains the maximum amount of total variance in the observed variables. Under some conditions, this means that the first component will be correlated with at least some of the observed variables. It may be correlated with many. The second component will have two important characteristics. First, this component explains a maximum amount of variance in the data set that was not accounted for by the first component. Again under some conditions, this means that the second component will be correlated with some of the observed variables that did not display strong correlations with the first component.

The second characteristic of the second component is that it will be uncorrelated (orthogonal) with the first component. The remaining components that are extracted in the analysis display the same two characteristics: Each component accounts for a maximum amount of variance in the observed variables which was not accounted for by the preceding components, and is uncorrelated with all of the preceding components. A principal component analysis proceeds in this fashion, with each new component accounting for progressively smaller and smaller amounts of variance (this is why only the first few components are usually retained and interpreted). When the analysis is complete, the resulting components will display varying degrees of correlation with the observed variables (https://support.sas.com/publishing/pubcat/chaps/55129.pdf), but are completely uncorrelated with one another.

Since no correlation does not generally imply that the components are independent, principal components are not generally independent except for normal distribution under which zero correlation implies independence. This is the reason why PCA works more successfully for Gaussian data. For non-Gaussian data, the independent component analysis is a better option.

References and Suggested Readings

Chattopadhyay, A. K. (2013). Independent component analysis for the objective classification of globular clusters of the galaxy NGC 5128. *Computational Statistics and Data Analysis*, *57*, 17–32.

Chattopadhyay, A. K., & Chattopadhyay, T. (2014). *Statistical methods for astronomical data analysis*., Springer series in astrostatistics New York: Springer.

Chattopadhyay, T., & Chattopadhyay, A. K. (2006). Objective classification of spiral galaxies having extended rotation curves beyond the optical radius. *Astronomical Journal*, *131*, 2452.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*(3), 453.

Hotelling H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441; *24*(7), 498–520.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*, 321–77.

Jeffers, J. N. R. (1967). Two case studies in the application of principal component analysis. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, *16*(3), 225–236.

Joliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, *31*(3), 300–303.

Jolliffe, I. T. (2002). Principal component analysis. Springer series in statistics (2$^{\text{nd}}$ ed., XXIX, 487 p. illus.). New York: Springer. ISBN 978-0-387-95442-4.