Principal Component Analysis with Application to Credit Card Data

Eleanor Cain
eleanor.cain@jacks.sdstate.edu
South Dakota State University
Department of Mathematics and Statistics

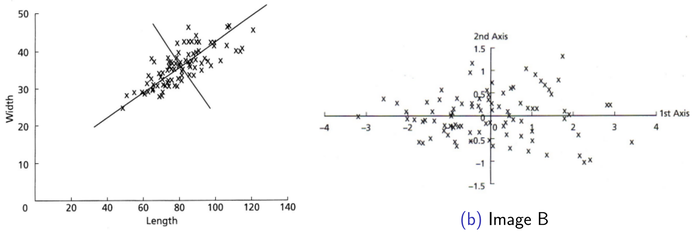
Introduction

Principal Component Analysis (PCA) is a method commonly used for dimension reduction or visualization and is the process by which principal components are computed then used to reduce the dimension of a data set. PCA was first invented by Karl Pearson in 1901, then later developed and named by Harold Hotelling in the 1930s.

Overview of PCA

Principal Components (PCs)

- Each PC is a normalized linear function of the original variables
- They explain as much variance as possible from the original data set
- All PCs are uncorrelated with each other
- Those PCs that contribute significantly to the variability should be kept when reducing the data set
- PCs can be interpreted geometrically



(a) Image A

(b) Image B

Figure: PCs Interpreted Geometrically

- Image A in the figure above displays the scatter plot of sepal length vs width from the Iris data set and the first two PCs extracted from the data
- Image B in the figure above displays the rotated data using PCs as axes

Principal Component Derivations

We will compute the first and second PCs. Given $\mathbf{x}$, a vector of length p,

$$y_k = \mathbf{a}_k'\mathbf{x} = a_{k1}x_1 + a_{k2}x_2 + ... + a_{kp}x_p.$$

Goal: obtain $\mathbf{a}$ such that we maximize

$$\text{var}(\mathbf{a}_k'\mathbf{x}) = \mathbf{a}_k'\Sigma\mathbf{a}_k$$

subject to the normalization constraint $\mathbf{a}_k'\mathbf{a}_k = 1$ with Σ being the covariance matrix of $\mathbf{x}$. For the first PC, we will maximize the function

$$f(\text{a}) = \mathbf{a}_1'\Sigma\mathbf{a}_1 - \lambda_1(\mathbf{a}_1'\mathbf{a}_1 - 1),$$

where λ_1 is a Lagrange multiplier. Taking the derivative with respect to $\mathbf{a}_1$, we get $\Sigma\mathbf{a}_1 = \lambda_1\mathbf{a}_1$. Hence, λ_1 is an eigenvalue of the covariance matrix Σ and $\mathbf{a}_1$ is the corresponding eigenvector.
Then,

$$\text{var}(\mathbf{a}_1'\mathbf{x}) = \mathbf{a}_1'\Sigma\mathbf{a}_1 = \mathbf{a}_1'\lambda_1\mathbf{a}_1 = \lambda_1\mathbf{a}_1'\mathbf{a}_1$$

where $\mathbf{a}_1'\mathbf{a}_1 = 1$. Therefore, we have

$$\text{var}(\mathbf{a}_1'\mathbf{x}) = \lambda_1.$$

For the second PC, in addition to maximizing variability, we want

$$\text{cov}[\mathbf{a}_j'\mathbf{x}, \mathbf{a}_k'\mathbf{x}] = 0$$

when $j \neq k$. Once again, we must maximize

$$\text{var}(\mathbf{a}_2'\mathbf{x}) = \mathbf{a}_2'\Sigma\mathbf{a}_2$$

subject to $\mathbf{a}_2'\mathbf{a}_2 = 1$ and $\mathbf{a}_2'\mathbf{a}_1 = 0$. In the end, we observe

$$\Sigma\mathbf{a}_2 = \lambda_2\mathbf{a}_2$$

which implies that λ_2 is once more an eigenvalue of the covariance matrix and $\text{var}(\mathbf{a}_2'\mathbf{x}) = \lambda_2$. These processes continue to find all PCs [2].

PCA on Credit Card Data

Credit Card Data Set

- 18,000 observations and 17 variables
- Consider only the numerical variables, except customer number (9 total variables)
- 24 missing values

Using the prcomp function from the stats package in R Studio [1], we can

- compute the PCs
- create graphs and tables to interpret the PCs

The cumulative percent variance for each PC from the credit card data are presented in the figure below.

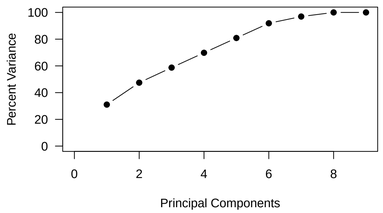


Figure: Scree Plot of Cumulative Variance Explained by PCs

From a scree plot, we can determine the number of PCs to consider. The first 6 PC loadings for each variable are shown in the table below.

Table: Table of Loading Vectors

	PC1	PC2	PC3	PC4	PC5	PC6
# Bank Accounts Open	0.00	0.01	0.57	-0.32	0.35	-0.67
# Credit Cards Held	0.00	0.03	-0.29	-0.94	-0.06	0.17
# Homes Owned	-0.01	-0.01	0.47	-0.09	-0.88	-0.01
Household Size	0.00	0.02	0.60	-0.08	0.32	0.73
Average Balance	-0.59	-0.07	0.00	-0.01	0.01	0.00
Q1 Balance	-0.22	-0.70	0.01	-0.03	0.01	0.01
Q2 Balance	-0.49	-0.31	-0.01	-0.01	0.01	0.00
Q3 Balance	-0.48	0.32	0.00	0.02	0.01	0.00
Q4 Balance	-0.35	0.55	0.00	0.00	-0.01	-0.01

PCA on Credit Card Data Cont.

A biplot is a specific type of graph used when performing PCA to help interpret the loading vectors.

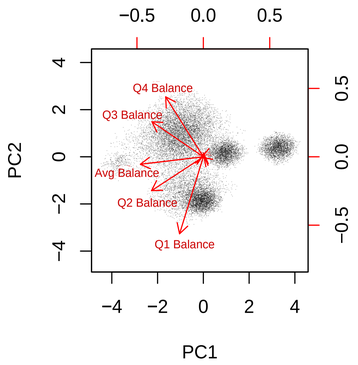


Figure: Biplot for Credit Card Data

Conclusion and Future Work

- Dimension reduction techniques such as PCA reduce the dimension of a data set while retaining as much information as possible
- If the first few PCs explain nearly all of the variance, the dimension can be significantly reduced
- Our future work aims to compare other PCs and investigate clustered observations in our biplot

References

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Jolliffe, I.T. (2002). *Principal Component Analysis, Second Edition*. Springer, New York. 59-61, 65-67.