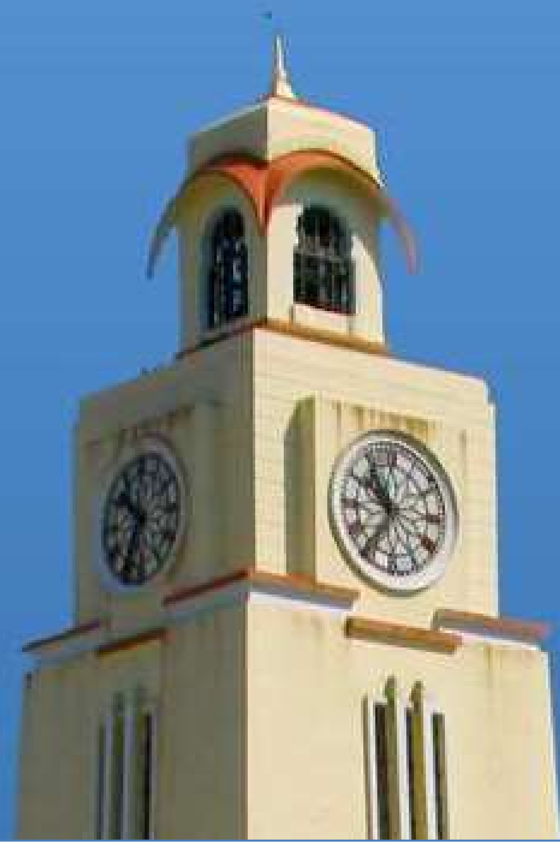

BITS Pilani
Pilani Campus

Principal Component Analysis

Introduction

- Increasing the number of features does not always improve accuracy.
- Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
- The idea of PCA is to reduce the number of variables of a data set, while preserving as much information as possible.

When should I use PCA?

- Do you want to reduce the number of variables, but aren't able to identify variables to completely remove from consideration?

- Do you want to ensure your variables are independent of one another?

- Are you comfortable making your independent variables less interpretable?

If you answered "yes" to all three questions, then PCA is a good method to use. If you answered "no" to question 3, you **should not** use PCA.

Terms

- **Variance :** It is a measure of the variability or it simply measures how spread the data set is.

- **Covariance :** It is a measure of the extent to which corresponding elements from two sets of ordered data move in the same direction.

$$var(x) = \frac{\sum(x_i - \bar{x})^2}{N} \qquad cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Terms

Sparseness

- Data lacks denseness, and its high percentage of the variable’s cells do not contain actual data.
- Fundamentally full of “empty” or “N/A” values.

c1	c2	c3	c4	c5
0	0	0	5	0
0	0	0	0	0
0	0	1	0	0
0	0	0	0	0
3	0	0	0	0
0	0	0	0	0

Example of a problem where PCA is required

- There are 100 students in a class with *m* different features like grade, age, height, weight, hair color, and others.

- Most of the features may not be relevant that describe the student. Therefore, it is vital to find the critical features that characterize a student.

Features to Ignore Vs Features to Keep

innovate achieve lead

Ignore

- Collinear features or linearly dependent features. e.g., leg size and height.
- Noisy features that are constant. e.g., the thickness of hair
- Constant features. e.g., Number of teeth.

Keep

- Non-collinear features or low covariance.
- Features that change a lot, high variance. e.g., grade. hobbies

What does Principal Component Analysis (PCA) do?

- PCA finds a new set of dimensions (or a set of basis of views) such that all the dimensions are orthogonal (and hence **linearly independent**)

- ranked according to the variance of data along them. It means more important principle axis occurs first. (more important = **more variance/more spread out data**)

Steps in PCA

height in cm 170, 190, . . .
weight in kg 60, 70 . . .

STEP 1: STANDARDIZATION

- The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.

- Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

$$z = \frac{value - mean}{\underline{standard\ deviation}}$$

Steps in PCA

STEP 2: COVARIANCE MATRIX COMPUTATION

- The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them.

- For example, for a 3-dimensional data set with 3 variables *x*, *y*, and *z*, the covariance matrix is a 3×3 matrix of this from:

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Steps in PCA

STEP 3: COMPUTE THE EIGENVECTORS AND EIGENVALUES OF THE COVARIANCE MATRIX TO IDENTIFY THE PRINCIPAL COMPONENTS

- Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the ***principal components*** of the data.

- If we rank the eigenvalues in descending order, we get $\lambda 1 > \lambda 2$, which means that the eigenvector that corresponds to the first principal component (PC1) is v1 and the one that corresponds to the second principal component (PC2) is v2.

Steps in PCA

STEP 4: FEATURE VECTOR

- The feature vector is simply a matrix that has as columns the eigenvectors of the components that we decide to keep.

- This makes it the first step towards dimensionality reduction, because if we choose to keep only ***p*** eigenvectors (components) out of ***n***, the final data set will have only ***p*** dimensions.

Steps in PCA

LAST STEP: RECAST THE DATA ALONG THE PRINCIPAL COMPONENTS AXES

- In this step, which is the last one, the aim is to use the feature vector formed using the eigenvectors of the covariance matrix, to reorient the data from the original axes to the ones represented by the principal components

- This can be done by multiplying the transpose of the original data set by the transpose of the feature vector.

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

Goal of PCA

- Find linearly independent dimensions (or basis of views) which can losslessly represent the data points.
- Those newly found dimensions should allow us to predict/reconstruct the original dimensions.
- The reconstruction/projection error should be minimized.

Exercise

Let our data matrix **X** be the score of three students :

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

0 to 100

same scale

Solution

- Step 1: Take the whole dataset consisting of d+1 dimensions and ignore the labels such that our new dataset becomes d dimensional.

- Step 2: Compute the mean of every dimension of the whole dataset.

$$\frac{90+90+60+60+30}{5}$$

$$\mathbf{A} = \begin{array}{c} \begin{array}{ccc} M & E & A \end{array} \\ \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix} \end{array}$$

Matrix A

$$\bar{\mathbf{A}} = [\ 66 \quad 60 \quad 60\ ]$$

Mean of Matrix A

- Step3: Compute the covariance matrix of the whole dataset (sometimes also called as the variance-covariance matrix)

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n} (Xi - \bar{x})(Yi - \bar{y})$$

$$\begin{array}{cc} & \begin{array}{ccc} Math & English & Art \end{array} \\ \begin{array}{c} Math \\ English \\ Art \end{array} & \begin{bmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{bmatrix} \end{array}$$

Covariance Matrix of A

a) The covariance between math and English is positive (360), and the covariance between math and art is positive (180). This means the scores tend to covary in a positive way. As scores on math go up, scores on art and English also tend to go up; and vice versa.

b) The covariance between English and art, however, is zero. This means there tends to be no predictable relationship between the movement of English and art scores.

* CAT PCA

- Step 4: Compute Eigenvectors and corresponding Eigenvalues

- The eigenvalues of ***A*** are roots of the characteristic equation

Identity matrix

$$\det(A-\lambda I) = 0$$

$$\det\left(\begin{pmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)$$

- Simplifying the matrix first, we can calculate the determinant later,

$$\begin{pmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{pmatrix} - \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}$$

$$\begin{pmatrix} 504-\lambda & 360 & 180 \\ 360 & 360-\lambda & 0 \\ 180 & 0 & 720-\lambda \end{pmatrix}$$

- Now that we have our simplified matrix, we can find the determinant of the same :

$$\det \begin{pmatrix} 504-\lambda & 360 & 180 \\ 360 & 360-\lambda & 0 \\ 180 & 0 & 720-\lambda \end{pmatrix}$$

$$-\lambda^3 + 1584\lambda^2 - 641520\lambda + 25660800$$

- After solving this equation for the value of ***λ,*** we get the following value

$$\lambda \approx 44.81966\ldots, \lambda \approx 629.11039\ldots, \lambda \approx 910.06995\ldots$$

- So, after solving for *eigenvectors* we would get the following solution for the corresponding *eigenvalues*

$$\begin{pmatrix} -3.75100\ldots \\ 4.28441\ldots \\ 1 \end{pmatrix}, \begin{pmatrix} -0.50494\ldots \\ -0.67548\ldots \\ 1 \end{pmatrix}, \begin{pmatrix} 1.05594\ldots \\ 0.69108\ldots \\ 1 \end{pmatrix}$$

- Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a d × k dimensional matrix W.

- So, after sorting the eigenvalues in decreasing order, we have

$$\begin{pmatrix} 910.06995 \\ 629.11039 \\ 44.81966 \end{pmatrix}$$

- For our simple example, where we are reducing a 3-dimensional feature space to a 2-dimensional feature subspace, we are combining the two eigenvectors with the highest eigenvalues to construct our d×k dimensional eigenvector matrix W.

- So, *eigenvectors* corresponding to two maximum eigenvalues are :

$$W = \begin{bmatrix} 1.05594 & -0.50494 \\ 0.69108 & -0.67548 \\ 1 & 1 \end{bmatrix}$$

Eigen values & Eigen Vectors

$$A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$$

$$|A - \lambda I| = 0$$

$$\left| \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$\left| \begin{bmatrix} -\lambda & 1 \\ -2 & -3-\lambda \end{bmatrix} \right| = 0$$

$$\lambda^2 + 3\lambda + 2 = 0$$

$$\underbrace{\lambda_1 = -1, \quad \lambda_2 = -2}_{\text{eigen values}}$$

V_1 — Eigen vector for λ_1

$$AV_1 = \lambda_1 . V_1$$

$$(A - \lambda_1) V_1 = 0$$

$$\left(\begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} - \lambda_1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) V_1 = 0$$

$\lambda_1 = -1$

$$\begin{bmatrix} -\lambda_1 & 1 \\ -2 & -3-\lambda_1 \end{bmatrix} V_1 = 0$$

$$\begin{bmatrix} +1 & 1 \\ -2 & -2 \end{bmatrix} V_1 = 0$$

$$V_{11} + V_{12} = 0$$

$$\begin{bmatrix} 1 & 1 \\ -2 & -2 \end{bmatrix} \begin{bmatrix} V_{11} \\ V_{12} \end{bmatrix} = 0$$

$$-2V_{11} - 2V_{12} = 0$$

$$V_{11} = -V_{12}$$

$$V_1 = K_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$\lambda_2 = -2$$

$$A.V_2 = \lambda_2 V_2$$

$$(A - \lambda_2) V_2 = 0$$

$$\begin{bmatrix} -\lambda_2 & 1 \\ -2 & -3-\lambda_2 \end{bmatrix} V_2 = \begin{bmatrix} 2 & 1 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} V_{21} \\ V_{22} \end{bmatrix} = 0$$

$$2V_{21} = -V_{22}$$

$$-2V_{21} = V_{22}$$

$$V_2 = K_2 \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

Self Study

- https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html
- http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
- https://textbooks.math.gatech.edu/ila/eigenvectors.html
- https://towardsdatascience.com/eigenvectors-and-eigenvalues-all-you-need-to-know-df92780c591f

References

- Probability and Statistics for engineering and sciences
- A Tutorial on Principal Component Analysis by Jonathon Shlens∗ Google Research
- https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202
- Book: http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf
- Numerical: https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643, https://www.itl.nist.gov/div898/handbook/pmc/section5/pmc552.htm
- Bivariate: http://personal.kenyon.edu/hartlaub/MellonProject/Bivariate2.html
- https://www.statlect.com/probability-distributions/multivariate-normal-distribution
- Material by Prof. Nathaniel E. Helwig
- https://medium.com/towards-artificial-intelligence/principal-component-analysis-pca-with-python-examples-tutorial-67a917bae9aa