Principal Component Analysis + Singular Value Decomposition

13 April 2023

Sumit Kumar Yadav

Department of Management Studies
Indian Institute of Technology, Roorkee

- PCA theory
- PCA some more theory
- PCA Implementation
- SVD Theory + Implementation

PCA Motivation

- To reduce the number of dimensions in the data
- To visualize the data
- To avoid over fitting

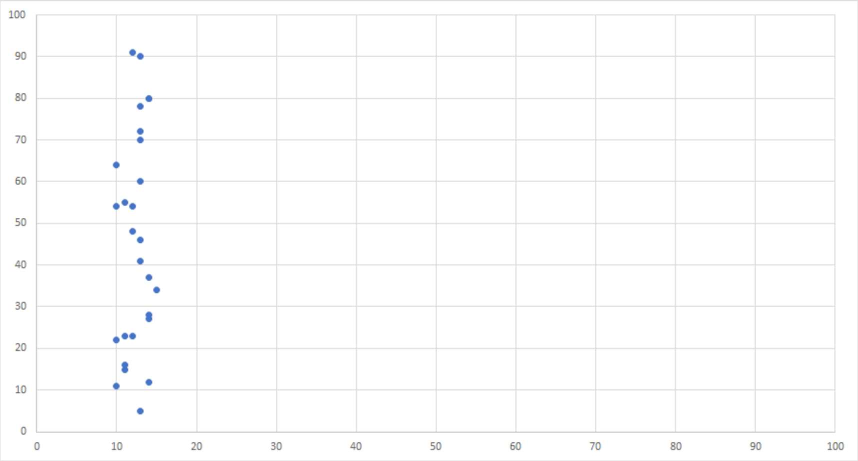
PCA Motivation

- To reduce the number of dimensions in the data
- To visualize the data
- To avoid over fitting

	Location	City	Society	Ambience	Airport
House 1	8	4	9	1	5
House 2	9	6	9	5	5
House 3	10	8	9	7	5
House 4	10	5	9	6	5
House 5	5	4	9	2	5
House 6	2	7	9	9	5
House 7	7	5	9	8	6
House 8	3	4	9	8	5
House 9	4	2	9	7	5
House 10	1	4	9	10	5

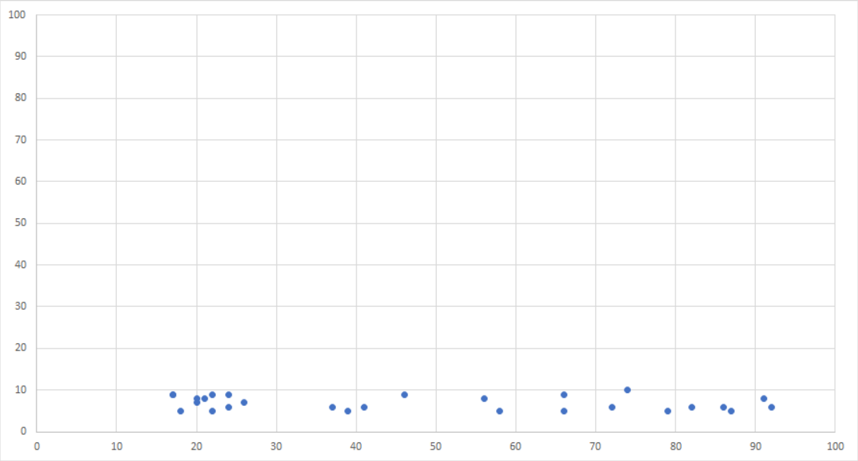
PCA Motivation

Which direction to skip?



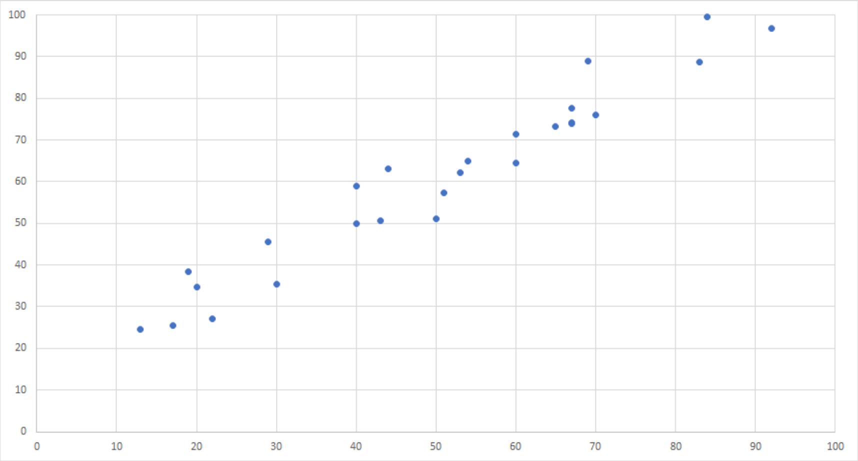
PCA Motivation

Which direction to skip?



PCA Motivation

Which direction to skip?



PCA Motivation

Looks like we are trying to capture as much variance as possible while reducing the number of dimensions.

What is a vector?

- a mathematical object that encodes a length and direction
- A vector is often represented as a 1-dimensional array of numbers, referred to as components and is displayed either in column form or row form
- Represented geometrically, vectors typically represent coordinates within a n-dimensional space

- Vector Space - Closed under addition and multiplication
- Vectors on line $y = 2 * x + 1$ form a vector space??
- Addition/Subtraction (Graphical Representation)
- Multiplication by a scalar
- Dot Product (Inner Product)
- Length/Magnitude
- Angle between two vectors - $\cos\theta = \frac{\vec{x}.\vec{y}}{||\vec{x}||.||\vec{y}||}$
- Dot product of perpendicular vectors $=$??

- concept of zero vector

Consider m vectors in $\mathbb{R}^n$,
If $\alpha_1 \vec{v_1} + \alpha_2 \vec{v_2} + \cdots + \alpha_n \vec{v_n}$ implies
$\alpha_1 = \alpha_2 = \cdots = \alpha_m = \vec{0}$, then
the vectors are said to be linearly independent.

Matrices

In mathematics, a matrix (plural matrices) is a rectangular array or table of numbers, symbols, or expressions, arranged in rows and columns, which is used to represent a mathematical object or a property of such an object. - Wikipedia

- Matrix Size
- Representing any element of a matrix
- (1,n) and (m,1) matrices
- Square Matrix

Operations on Matrices

- Addition/Subtraction
- Multiplication of Matrix by a scalar
- Transpose of a Matrix
- Multiplication of two Matrices

Types of Matrices

- Diagonal
- Identity
- Upper Triangular
- Symmetric
- Singular
- Mirror Matrix

Matrices contd.

- Inverse of a Matrix
- Trace of a Matrix
- Relationship of Eigenvalues with trace and Determinant
- Symmetric Matrix
- Orthogonal Matrix

Eigenvalues and Eigenvectors

Let M be a nxn matrix. A non-zero vector $\vec{X}$ is said to be an eigenvector of M corresponding to eigenvalue λ if

Spectral Decomposition

A n*n matrix can be written in terms of its eigenvalues and eigenvectors as follows -

$$M = \lambda_1 \vec{v_1} \vec{v_1}^T + \lambda_2 \vec{v_2} \vec{v_2}^T + ... + \lambda_n \vec{v_n} \vec{v_n}^T$$

Positive Semi-definite Matrix

$$x^T M x \geq 0$$

All the eigenvalues are greater than zero.

PCA - Objectives

We have a Data Matrix(D), whose dimension is $n * f$.

$$\begin{bmatrix} d_{11} & d_{12} & \dots & d_{1f} \\ d_{21} & d_{22} & \dots & d_{2f} \\ \vdots & \vdots & \dots & \vdots \\ d_{(n-1)1} & d_{(n-1)2} & \dots & d_{(n-1)f} \\ d_{n1} & d_{n2} & \dots & d_{nf} \end{bmatrix}$$

We want to reduce the number of features to f_s.

PCA - Objectives

We have a Data Matrix(D), whose dimension is $n * f$.

$$\begin{bmatrix} d_{11} & d_{12} & \dots & d_{1f} \\ d_{21} & d_{22} & \dots & d_{2f} \\ \vdots & \vdots & \dots & \vdots \\ d_{(n-1)1} & d_{(n-1)2} & \dots & d_{(n-1)f} \\ d_{n1} & d_{n2} & \dots & d_{nf} \end{bmatrix}$$

We want to reduce the number of features to f_s.

Let us solve a simpler problem first. Let us say we want to reduce the number of dimensions/features to just 1. Which is the best way to go about it?

PCA Objective

We are looking for a unit vector $\vec{v} = (v_1, v_2, ..., v_f)$ such that the variance of $D\vec{v}$ is as large as possible.

PCA Objective

We are looking for a unit vector $\vec{v} = (v_1, v_2, ..., v_f)$ such that the variance of $D\vec{v}$ is as large as possible.

Variance($D\vec{v}$) can be written as $\vec{v}^T \Sigma \vec{v}$, where Σ is the variance covariance matrix of D. (Try to prove this or atleast verify this)

PCA Objective

We are looking for a unit vector $\vec{v} = (v_1, v_2, ..., v_f)$ such that the variance of $D\vec{v}$ is as large as possible.

Variance$(D\vec{v})$ can be written as $\vec{v}^T \Sigma \vec{v}$, where Σ is the variance covariance matrix of D. (Try to prove this or atleast verify this)

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1f} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2f} \\ \vdots & \vdots & \dots & \vdots \\ \sigma_{(f-1)1} & \sigma_{(f-1)2} & \dots & \sigma_{(f-1)f} \\ \sigma_{1f} & \sigma_{f2} & \dots & \sigma_f^2 \end{bmatrix}$$

PCA Optimization problem

Maximize - $\vec{v}^T \Sigma \vec{v}$

such that $||\vec{v}|| = 1$

PCA Optimization problem

Maximize - $\vec{v}^T \Sigma \vec{v}$

such that $||\vec{v}|| = 1$

Σ can be shown to be positive semi-definite. Thus, all the eigenvalues are greater than zero, and all eigenvectors are orthogonal.

PCA Optimization problem

Maximize - $\vec{v}^T \Sigma \vec{v}$

such that $||\vec{v}|| = 1$

Σ can be shown to be positive semi-definite. Thus, all the eigenvalues are greater than zero, and all eigenvectors are orthogonal.

$$\Sigma = \lambda_1 \vec{v_1} \vec{v_1}^T + \lambda_2 \vec{v_2} \vec{v_2}^T + \ldots + \lambda_n \vec{v_n} \vec{v_n}^T$$

$$\lambda_1 \geq \lambda_2 \geq \ldots . \geq \lambda_n$$

Maximize - $\vec{v}^T(\lambda_1 \vec{v_1}\vec{v_1}^T + \lambda_2 \vec{v_2}\vec{v_2}^T + ... + \lambda_n \vec{v_n}\vec{v_n}^T)\vec{v}$

such that $||\vec{v}|| = 1$

PCA Optimization problem

Maximize - $\vec{v}^T(\lambda_1 \vec{v_1} \vec{v_1}^T + \lambda_2 \vec{v_2} \vec{v_2}^T + ... + \lambda_n \vec{v_n} \vec{v_n}^T)\vec{v}$

such that $||\vec{v}|| = 1$

All $\vec{v_i}'s$ are unit vectors and orthogonal to each other, so the best choice for $\vec{v}$ is ??

What about reducing to 2 dimensions?

Variance is no longer a scalar which can be compared across all possible 2 dimensions where data can be projected.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdot & \\ \sigma_{12} & \sigma_2^2 & \cdot & - \\ - & & \ddots & \\ | & \vdots & & \sigma_f^2 \end{bmatrix}$$

$$\lambda_1 \geq \lambda_2 \; \text{- - -} \geq \lambda_f$$

$$\vec{v}_1 \quad \vec{v}_2 \quad \vec{v}_f$$

What about reducing to 2 dimensions?

Variance is no longer a scalar which can be compared across all possible 2 dimensions where data can be projected.

We have an additional constraint that the co-variance terms of the reduced dimensions should be zero.

$D_{n \times f}$ | $D_{trans} = D \underbrace{\begin{bmatrix} \vec{v_1} & \vec{v_2} & \vec{v_3} & \cdots & \vec{v_{f_s}} \\ | & | & | & & | \end{bmatrix}}_{}$ $+ E_{f \times f_s}$

$(n \times f_s)$

$D_{\text{(contrans)}} = D_{trans} E^T$

What about reducing to 2 dimensions?

Variance is no longer a scalar which can be compared across all possible 2 dimensions where data can be projected.

We have an additional constraint that the co-variance terms of the reduced dimensions should be zero.

Fortunately, this happens if we keep going in the same sequence of the principal components.

What about reducing to 2 dimensions?

Variance is no longer a scalar which can be compared across all possible 2 dimensions where data can be projected.

We have an additional constraint that the co-variance terms of the reduced dimensions should be zero.

Fortunately, this happens if we keep going in the same sequence of the principal components.

We can now try to maximize the sum of the variance terms in the 2 projected directions.
So, it looks like that the most logical thing to do is to pick up the eigenvector corresponding to the second largest eigenvector. This is what we do in PCA.

Reconstructing the data

One could start with the objective that we want to minimize the reconstruction error, and we would get the same result as what we have described above.
Basic Ideas - Any k-dimensional vector can be represented in terms of k-orthonormal vectors.

Reconstructing the data

One could start with the objective that we want to minimize the reconstruction error, and we would get the same result as what we have described above.

Basic Ideas - Any k-dimensional vector can be represented in terms of k-orthonormal vectors.

Eg - Consider the vector (2,3) in 2-D space and two orthonormal vectors (1,0) and (0,1)

$(2,3) = ((2,3).(1,0))(1,0) + ((2,3).(0,1))(0,1)$

Reconstructing the data

One could start with the objective that we want to minimize the reconstruction error, and we would get the same result as what we have described above.

Basic Ideas - Any k-dimensional vector can be represented in terms of k-orthonormal vectors.

Eg - Consider the vector (2,3) in 2-D space and two orthonormal vectors (1,0) and (0,1)

$(2,3) = ((2,3).(1,0))(1,0) + ((2,3).(0,1))(0,1)$

Too trivial, isn't it??

Reconstructing the data

Eg - Consider the vector (2,3) in 2-D space and two orthonormal vectors (0.6,0.8) and (-0.8,0.6)

$$(2,3) = ((2,3).(0.6,0.8))(0.6,0.8) + ((2,3).(-0.8,0.6))(-0.8,0.6)$$

Reconstructing the data

Eg - Consider the vector (2,3) in 2-D space and two orthonormal vectors (0.6,0.8) and (-0.8,0.6)

(2,3) = ((2,3).(0.6,0.8))(0.6,0.8) + ((2,3).(-0.8,0.6))(-0.8,0.6)

Thus, in general for a k-dimensional vector $\vec{d}$, and k-orthonormal vectors $\vec{v_1}, \vec{v_2}, \vec{v_3}, ..., \vec{v_k}$

$$\vec{d} = (\vec{d}.\vec{v_1})\vec{v_1} + (\vec{d}.\vec{v_2})\vec{v_2} + (\vec{d}.\vec{v_3})\vec{v_3} + ... + (\vec{d}.\vec{v_k})\vec{v_k}$$

$\vec{v}_1, \vec{v}_2, \vec{v}_3, \ldots \ldots, \vec{v}_f$

In PCA, once we have found out the eigenvalues and eigenvectors and projected the data in the lower dimensional space(f_s), the way to reconstruct for a data point back to the original dimensions (f) is -

$$\vec{d}_1 = (\vec{d}_1 \cdot \vec{v}_1)\vec{v}_1 + (\vec{d}_1 \cdot \vec{v}_2)\vec{v}_2 \ldots + (\vec{d}_1 \cdot \vec{v}_f)\vec{v}_f$$

$$\vec{d}_2 = (\vec{d}_2 \cdot \vec{v}_1)\vec{v}_1 + (\vec{d}_2 \cdot \vec{v}_2)\vec{v}_2 + \ldots + (\vec{d}_2 \cdot \vec{v}_f)\vec{v}_f$$

$$\vdots$$

$$\vec{d}_n = (\vec{d}_n \cdot \vec{v}_1)\vec{v}_1 + (\vec{d}_n \cdot \vec{v}_2)\vec{v}_2 + \ldots + (\vec{d}_n \cdot \vec{v}_f)\vec{v}_f$$

PCA reconstruction

In PCA, once we have found out the eigenvalues and eigenvectors and projected the data in the lower dimensional space(f_s), the way to reconstruct for a data point back to the original dimensions (f) is -

$$\vec{d_{inv}} = (\vec{d}.\vec{v_1})\vec{v_1} + (\vec{d}.\vec{v_2})\vec{v_2} + (\vec{d}.\vec{v_3})\vec{v_3} + ... + (\vec{d}.\vec{v_{f_s}})\vec{v_{f_s}}$$

PCA reconstruction

In PCA, once we have found out the eigenvalues and eigenvectors and projected the data in the lower dimensional space(f_s), the way to reconstruct for a data point back to the original dimensions (f) is -

$$\vec{d_{inv}} = (\vec{d}.\vec{v_1})\vec{v_1} + (\vec{d}.\vec{v_2})\vec{v_2} + (\vec{d}.\vec{v_3})\vec{v_3} + ... + (\vec{d}.\vec{v_{f_s}})\vec{v_{f_s}}$$

This is essentially what we get after the matrix manipulations we saw in the previous session.

Percentage of Variance Captured in PCA

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1f} \\ \sigma_{12} & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & \\ & & & \sigma_f^2 \\ \sigma_{1f} & \sigma_{2f} & \cdots & \end{bmatrix} \qquad \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_f^2$$

$$\lambda_1 + \lambda_2 + \cdots + \lambda_f = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_f^2$$

SVD

Singular Value Decomposition - A way to approximate a general matrix of dimension m*n.

$$M = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \ldots \ldots + \lambda_{1000} v_{1k} v_{1k}^T$$

Singular Value Decomposition - A way to approximate a general matrix of dimension m*n.
Can we use PCA to approximate the matrix of size n*n?

$$(1000) \times 5 + 5 = 5005$$

$$1000 \times 1000 = 10 \text{ lakh}$$

SVD

Singular Value Decomposition - A way to approximate a general matrix of dimension m*n.
Can we use PCA to approximate the matrix of size n*n? Can we do it always??

SVD

Let A be a general matrix of size m*n.
What can be said about AA^T and A^TA

Let A be a general matrix of size m*n.
What can be said about AA^T and A^TA

SVD

Thank you for your attention