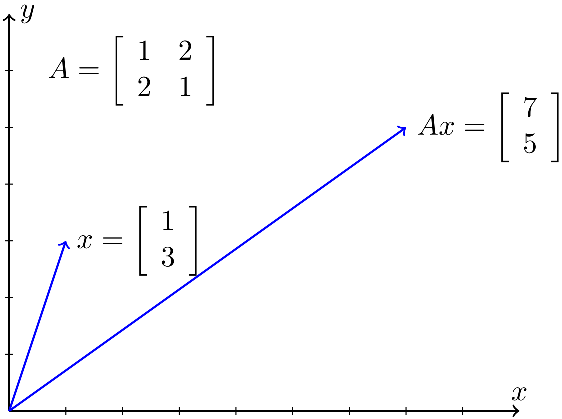
CS7015 (Deep Learning) : Lecture 6

Eigen Values, Eigen Vectors, Eigen Value Decomposition, Principal Component Analysis, Singular Value Decomposition

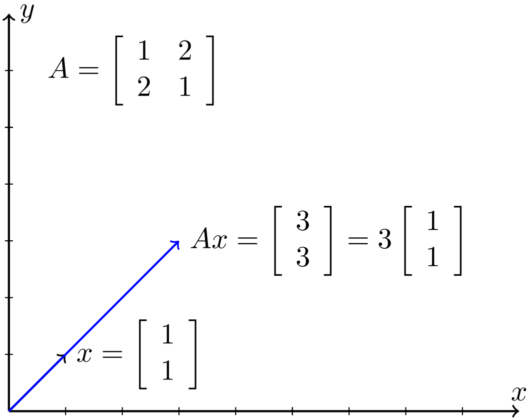
Prof. Mitesh M. Khapra

Department of Computer Science and Engineering
Indian Institute of Technology Madras

Module 6.1 : Eigenvalues and Eigenvectors


y
$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$
$Ax = \begin{bmatrix} 7 \\ 5 \end{bmatrix}$
$x = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$
x

- What happens when a matrix hits a vector?
- The vector gets transformed into a new vector (it strays from its path)
- The vector may also get scaled (elongated or shortened) in the process.



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

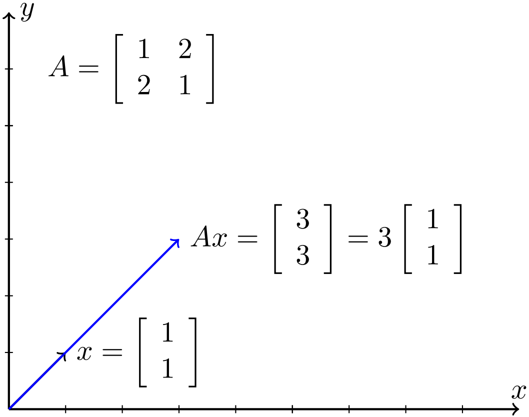
$$Ax = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

- For a given square matrix A, there exist special vectors which refuse to stray from their path.
- These vectors are called eigenvectors.
- More formally,

  $$Ax = \lambda x \text{ [direction remains the same]}$$

- The vector will only get scaled but will not change its direction.

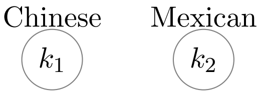


$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

$$Ax = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

- So what is so special about eigenvectors?
- Why are they always in the limelight?
- It turns out that several properties of matrices can be analyzed based on their eigenvalues (for example, see spectral graph theory)
- We will now see two cases where eigenvalues/vectors will help us in this course



$$v_{(0)} = \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}$$

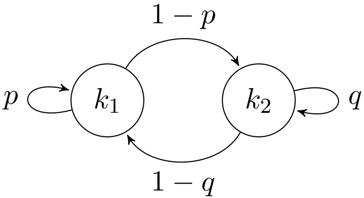
$$\begin{aligned} v_{(1)} &= \begin{bmatrix} pk_1 + (1-q)k_2 \\ (1-p)k_1 + qk_2 \end{bmatrix} \\ &= \begin{bmatrix} p & 1-q \\ 1-p & q \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} v_{(1)} &= Mv_{(0)} \\ v_{(2)} &= Mv_{(1)} \\ &= M^2 v_{(0)} \end{aligned}$$

In general, $v_{(n)} = M^n v_{(0)}$

- Let us assume that on day 0, k_1 students eat Chinese food, and k_2 students eat Mexican food. (Of course, no one eats in the mess!)
- On each subsequent day i, a fraction p of the students who ate Chinese food on day $(i-1)$, continue to eat Chinese food on day i, and $(1-p)$ shift to Mexican food.
- Similarly a fraction q of students who ate Mexican food on day $(i-1)$ continue to eat Mexican food on day i, and $(1-q)$ shift to Chinese food.
- The number of customers in the two restaurants is thus given by the following series:

$$v_{(0)}, Mv_{(0)}, M^2v_{(0)}, M^3v_{(0)}, \ldots$$



- This is a problem for the two restaurant owners.
- The number of patrons is changing constantly.
- Or is it? Will the system eventually reach a steady state? (i.e. will the number of customers in the two restaurants become constant over time?)
- Turns out they will!
- Let's see how?

Definition

Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvectors of an $n \times n$ matrix A. λ_1 is called the dominant eigen value of A if

$$|\lambda_1| \geq |\lambda_i| \; i = 2, \ldots, n$$

Theorem

The largest (dominant) eigenvalue of a stochastic matrix is 1.
See proof here

Definition

A matrix M is called a stochastic matrix if all the entries are positive and the sum of the elements in each column is equal to 1.
(Note that the matrix in our example is a stochastic matrix)

Theorem

If A is a $n \times n$ square matrix with a dominant eigenvalue, then the sequence of vectors given by Av_0, A^2v_0, $\ldots$, $A^nv_0, \ldots$ approaches a multiple of the dominant eigenvector of A.
(the theorem is slightly misstated here for ease of explanation)

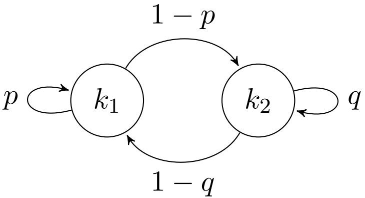
- Let e_d be the dominant eigenvector of M and $\lambda_d = 1$ the corresponding dominant eigenvalue
- Given the previous definitions and theorems, what can you say about the sequence $Mv_{(0)}, M^2v_{(0)}, M^3v_{(0)}, \ldots$?
- There exists an n such that

$$v_{(n)} = M^n v_{(0)} = ke_d \text{ (some multiple of } e_d)$$

- Now what happens at time step $(n+1)$?

$$v_{(n+1)} = Mv_{(n)} = M(ke_d) = k(Me_d) = k(\lambda_d e_d) = ke_d$$

- The population in the two restaurants becomes constant after time step n.
  See Proof Here



- Now instead of a stochastic matrix let us consider any square matrix A
- Let p be the time step at which the sequence $x_0, Ax_0, A^2x_0, \ldots$ approaches a multiple of e_d (the dominant eigenvector of A)

$$\begin{aligned}
A^p x_0 &= ke_d \\
A^{p+1}x_0 &= A(A^p x_0) = kAe_d = k\lambda_d e_d \\
A^{p+2}x_0 &= A(A^{p+1}x_0) = k\lambda_d Ae_d = k\lambda_d^2 e_d \\
A^{p+n}x_0 &= k(\lambda_d)^n e_d
\end{aligned}$$

- In general, if λ_d is the dominant eigenvalue of a matrix A, what would happen to the sequence $x_0, Ax_0, A^2x_0, \ldots$ if
    - $|\lambda_d| > 1$ (will explode)
    - $|\lambda_d| < 1$ (will vanish)
    - $|\lambda_d| = 1$ (will reach a steady state)
- (We will use this in the course at some point)

Module 6.2 : Linear Algebra - Basic Definitions

- We will see some more examples where eigenvectors are important, but before that let’s revisit some basic definitions from linear algebra.

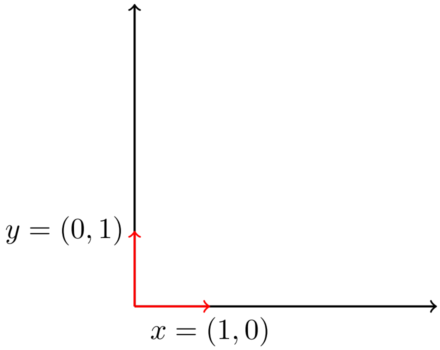
Basis

A set of vectors $\in \mathbb{R}^n$ is called a basis, if they are linearly independent and every vector $\in \mathbb{R}^n$ can be expressed as a linear combination of these vectors.

Linearly independent vectors

A set of n vectors $v_1, v_2, \ldots, v_n$ is linearly independent if no vector in the set can be expressed as a linear combination of the remaining $n-1$ vectors.
In other words, the only solution to

$$c_1 v_1 + c_2 v_2 + \ldots c_n v_n = 0 \text{ is } c_1 = c_2 = \cdots = c_n = 0 (c_i\text{'}s \ are \ scalars)$$



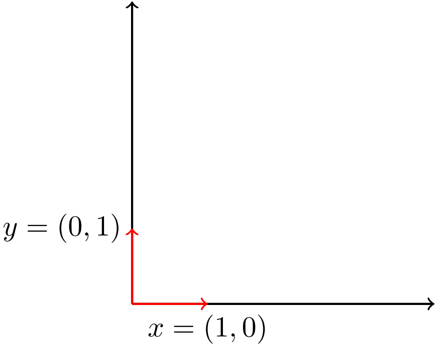
- For example consider the space $\mathbb{R}^2$
- Now consider the vectors

$$x = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ and } y = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

- Any vector $\begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^2$, can be expressed as a linear combination of these two vectors i.e

$$\begin{bmatrix} a \\ b \end{bmatrix} = a \begin{bmatrix} 1 \\ 0 \end{bmatrix} + b \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

- Further, x and y are linearly independent. (the only solution to $c_1x + c_2y = 0$ is $c_1 = c_2 = 0$)

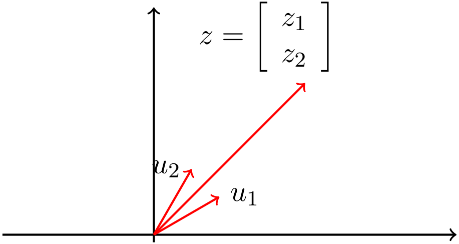


$$\begin{bmatrix} a \\ b \end{bmatrix} = x_1 \begin{bmatrix} 2 \\ 3 \end{bmatrix} + x_2 \begin{bmatrix} 5 \\ 7 \end{bmatrix}$$

$$a = 2x_1 + 5x_2$$

$$b = 3x_1 + 7x_2$$

- In fact, turns out that x and y are unit vectors in the direction of the co-ordinate axes.
- And indeed we are used to representing all vectors in $\mathbb{R}^2$ as a linear combination of these two vectors.
- But there is nothing sacrosanct about the particular choice of x and y.
- We could have chosen any 2 linearly independent vectors in $\mathbb{R}^2$ as the basis vectors.
- For example, consider the linearly independent vectors, $[2,3]^T$ and $[5,7]^T$. See how any vector $[a,b]^T \in \mathbb{R}^2$ can be expressed as a linear combination of these two vectors.
- We can find x_1 and x_2 by solving a system of linear equations.



- In general, given a set of linearly independent vectors $u_1, u_2, \dots u_n \in \mathbb{R}^n$, we can express any vector $z \in \mathbb{R}^n$ as a linear combination of these vectors.

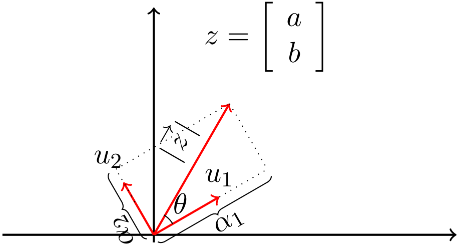
$$z = \alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_n u_n$$

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \alpha_1 \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1n} \end{bmatrix} + \alpha_2 \begin{bmatrix} u_{21} \\ u_{22} \\ \vdots \\ u_{2n} \end{bmatrix} + \dots + \alpha_n \begin{bmatrix} u_{n1} \\ u_{n2} \\ \vdots \\ u_{nn} \end{bmatrix}$$

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} u_{11} & u_{21} & \dots & u_{n1} \\ u_{12} & u_{22} & \dots & u_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ u_{1n} & u_{2n} & \dots & u_{nn} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}$$

(Basically rewriting in matrix form)

- We can now find the α_is using Gaussian Elimination (Time Complexity: $O(n^3)$)



$$\alpha_1 = |\vec{z}|cos\theta = |\vec{z}|\frac{z^T u_1}{|\vec{z}||u_1|} = z^T u_1$$

Similarly, $\alpha_2 = z^T u_2$.
When u_1 and u_2 are unit vectors along the co-ordinate axes

$$z = \begin{bmatrix} a \\ b \end{bmatrix} = a \begin{bmatrix} 1 \\ 0 \end{bmatrix} + b \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

- Now let us see if we have orthonormal basis.
- $u_i^T u_j = 0 \; \forall i \neq j$ and $u_i^T u_i = \|u_i\|^2 = 1$
- Again we have:

$$\begin{aligned} z &= \alpha_1 u_1 + \alpha_2 u_2 + \ldots + \alpha_n u_n \\ u_1^T z &= \alpha_1 u_1^T u_1 + \ldots + \alpha_n u_1^T u_n \\ &= \alpha_1 \end{aligned}$$

- We can directly find each α_i using a dot product between z and u_i (time complexity $O(N)$)
- The total complexity will be $O(N^2)$

Remember

An orthogonal basis is the most convenient basis that one can hope for.

Theorem 1

The eigenvectors of a matrix $A \in \mathbb{R}^{n \times n}$ having distinct eigenvalues are linearly independent.

Proof: See here

Theorem 2

The eigenvectors of a square symmetric matrix are orthogonal.

Proof: See here

- But what does any of this have to do with eigenvectors?
- Turns out that the eigenvectors can form a basis.
- In fact, the eigenvectors of a square symmetric matrix are even more special.
- Thus they form a very convenient basis.
- Why would we want to use the eigenvectors as a basis instead of the more natural co-ordinate axes?
- We will answer this question soon.

Module 6.3 : Eigenvalue Decomposition

Before proceeding let's do a quick recap of eigenvalue decomposition.

- Let $u_1, u_2, \ldots, u_n$ be the eigenvectors of a matrix A and let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the corresponding eigenvalues.
- Consider a matrix U whose columns are $u_1, u_2, \ldots, u_n$.
- Now

$$AU = A \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ u_1 & u_2 & \ldots & u_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ Au_1 & Au_2 & \ldots & Au_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

$$= \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \lambda_1 u_1 & \lambda_2 u_2 & \ldots & \lambda_n u_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

$$= \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ u_1 & u_2 & \ldots & u_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & \lambda_n \end{bmatrix} = U\Lambda$$

- where Λ is a diagonal matrix whose diagonal elements are the eigenvalues of A.

$$AU = U\Lambda$$

- If U^{-1} exists, then we can write,

$$A = U\Lambda U^{-1} \quad \text{[eigenvalue decomposition]}$$
$$U^{-1}AU = \Lambda \quad \text{[diagonalization of A]}$$

- Under what conditions would U^{-1} exist?
  - If the columns of U are linearly independent [**See proof here**]
  - *i.e.* if A has n linearly independent eigenvectors.
  - *i.e.* if A has n distinct eigenvalues [**sufficient condition, proof : Slide 19 Theorem 1**]

- If A is symmetric then the situation is even more convenient.
- The eigenvectors are orthogonal [**proof : Slide 19 Theorem 2**]
- Further let's assume, that the eigenvectors have been normalized [$u_i^T u_i = 1$]

$$Q = U^T U = \begin{bmatrix} \leftarrow u_1 \rightarrow \\ \leftarrow u_2 \rightarrow \\ \dots \\ \leftarrow u_n \rightarrow \end{bmatrix} \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ u_1 & u_2 & \dots & u_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

- Each cell of the matrix, Q_{ij} is given by $u_i^T u_j$

$$\begin{aligned} Q_{ij} = u_i^T u_j \quad &= 0 \text{ if } i \neq j \\ &= 1 \text{ if } i = j \end{aligned}$$

$$\therefore U^T U = \mathbb{I} \text{ (the identity matrix)}$$

- U^T is the inverse of U (very convenient to calculate)

Something to think about

- Given the EVD, $A = U\Sigma U^T$,
  what can you say about the sequence $x_0, Ax_0, A^2x_0, \ldots$ in terms of the eigen values of A.
  (Hint: You should arrive at the same conclusion we saw earlier)

Theorem (one more important property of eigenvectors)

If A is a square symmetric $N \times N$ matrix, then the solution to the following optimization problem is given by the eigenvector corresponding to the largest eigenvalue of A.

$$\max_{x} x^T A x$$

$$\text{s.t } \|x\| = 1$$

and the solution to

$$\min_{x} x^T A x$$

$$\text{s.t } \|x\| = 1$$

is given by the eigenvector corresponding to the smallest eigenvalue of A.

Proof: Next slide.

- This is a constrained optimization problem that can be solved using Lagrange Multipliers:

$$L = x^T Ax - \lambda(x^T x - 1)$$
$$\frac{\partial L}{\partial x} = 2Ax - \lambda(2x) = 0 => Ax = \lambda x$$

- Hence x must be an eigenvector of A with eigenvalue λ.
- Multiplying by x^T:

$$x^T Ax = \lambda x^T x = \lambda (\text{since } x^T x = 1)$$

- Therefore, the critical points of this constrained problem are the eigenvalues of A.
- The maximum value is the largest eigenvalue, while the minimum value is the smallest eigenvalue.

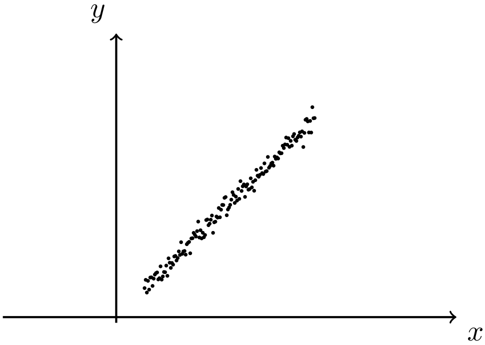
The story so far...

- The eigenvectors corresponding to different eigenvalues are linearly independent.
- The eigenvectors of a square symmetric matrix are orthogonal.
- The eigenvectors of a square symmetric matrix can thus form a convenient basis.
- We will put all of this to use.

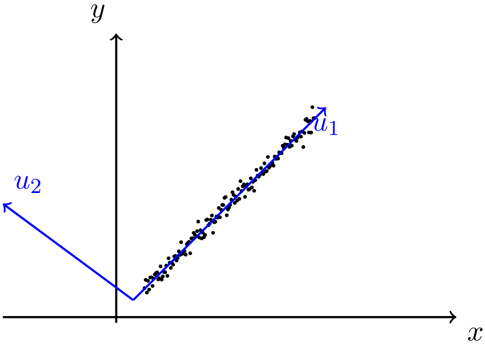
Module 6.4 : Principal Component Analysis and its Interpretations

The story ahead...

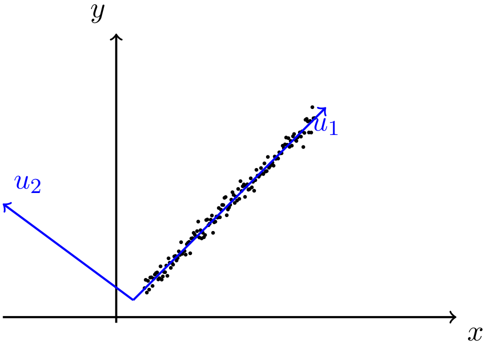
- Over the next few slides we will introduce Principal Component Analysis and see three different interpretations of it

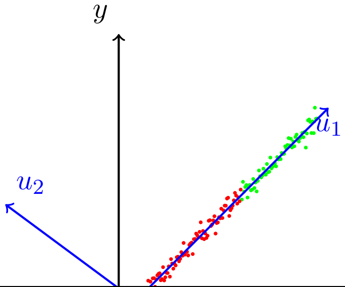


- Consider the following data
- Each point (vector) here is represented using a linear combination of the x and y axes (i.e. using the point's x and y co-ordinates)
- In other words we are using x and y as the basis
- What if we choose a different basis?

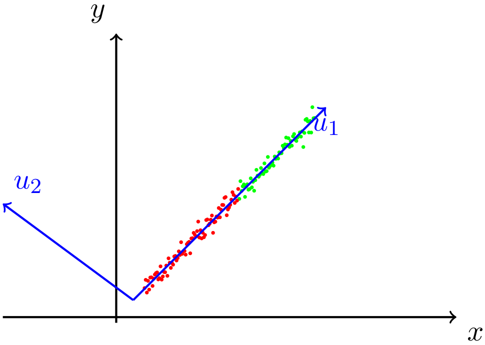


- For example, what if we use u_1 and u_2 as a basis instead of x and y.
- We observe that all the points have a very small component in the direction of u_2 (almost noise)
- It seems that the same data which was originally in $\mathbb{R}^2(x, y)$ can now be represented in $\mathbb{R}^1(u_1)$ by making a smarter choice for the basis





- Let's try stating this more formally
- Why do we not care about u_2?
- Because the variance in the data in this direction is very small (all data points have almost the same value in the u_2 direction)
- If we were to build a classifier on top of this data then u_2 would not contribute to the classifier as the points are not distinguishable along this direction

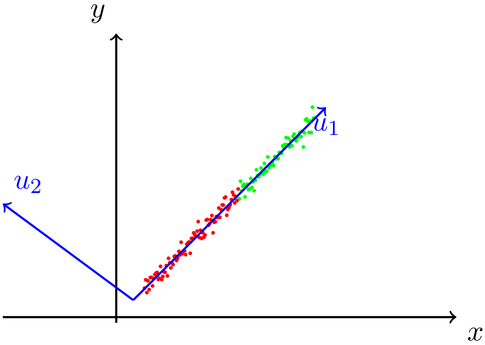


- In general, we are interested in representing the data using fewer dimensions such that the data has high variance along these dimensions
- Is that all?
- No, there is something else that we desire. Let's see what.

x	**y**	**z**
1	1	1
0.5	0	0
0.25	1	1
0.35	1.5	1.5
0.45	1	1
0.57	2	2.1
0.62	1.1	1
0.73	0.75	0.76
0.72	0.86	0.87

$$\rho_{yz} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(z_i - \overline{z})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}(z_i - \overline{z})^2}}$$

- Consider the following data
- Is z adding any new information beyond what is already contained in y?
- The two columns are highly correlated (or they have a high covariance)
- In other words the column z is redundant since it is linearly dependent on y.



In general, we are interested in representing the data using fewer dimensions such that

- the data has high variance along these dimensions
- the dimensions are linearly independent (uncorrelated)
- (even better if they are orthogonal because that is a very convenient basis)

Let $p_1, p_2, \cdots, p_n$ be a set of such n linearly independent orthonormal vectors. Let P be a $n \times n$ matrix such that $p_1, p_2, \cdots, p_n$ are the columns of P.

Let $x_1, x_2, \cdots, x_m \in \mathbb{R}^n$ be m data points and let X be a matrix such that $x_1, x_2, \cdots, x_m$ are the rows of this matrix. Further let us assume that the data is 0-mean and unit variance.

We want to represent each x_i using this new basis P.

$$x_i = \alpha_{i1}p_1 + \alpha_{i2}p_2 + \alpha_{i3}p_3 + \cdots + \alpha_{in}p_n$$

For an orthonormal basis we know that we can find these $\alpha_i' s$ using

$$\alpha_{ij} = x_i^T p_j = \begin{bmatrix} \leftarrow & x_i & \rightarrow \end{bmatrix}^T \begin{bmatrix} \uparrow \\ p_j \\ \downarrow \end{bmatrix}$$

In general, the transformed data $\hat{x}_i$ is given by

$$\hat{x}_i = \begin{bmatrix} \leftarrow & x_i^T & \rightarrow \end{bmatrix} \begin{bmatrix} \uparrow & & \uparrow \\ p_1 & \cdots & p_n \\ \downarrow & & \downarrow \end{bmatrix} = x_i^T P$$

and

$$\hat{X} = XP \qquad (\hat{X} \text{ is the matrix of transformed points})$$

Theorem:

If X is a matrix such that its columns have zero mean and if $\hat{X} = XP$ then the columns of $\hat{X}$ will also have zero mean.

Proof: For any matrix A, $\mathbf{1}^T A$ gives us a row vector with the i^{th} element containing the sum of the i^{th} column of A. (this is easy to see using the row-column picture of matrix multiplication).

Consider

$$\mathbf{1}^T \hat{X} = \mathbf{1}^T XP = (\mathbf{1}^T X)P$$

But $\mathbf{1}^T X$ is the row vector containing the sums of the columns of X. Thus $\mathbf{1}^T X = 0$. Therefore, $\mathbf{1}^T \hat{X} = 0$.

Hence the transformed matrix also has columns with sum $= 0$.

Theorem:

$X^T X$ is a symmetric matrix.

Proof: We can write $(X^T X)^T = X^T (X^T)^T = X^T X$

Definition:

If X is a matrix whose columns are zero mean then $\Sigma = \frac{1}{m}X^TX$ is the covariance matrix. In other words each entry Σ_{ij} stores the covariance between columns i and j of X.

Explanation: Let C be the covariance matrix of X. Let μ_i, μ_j denote the means of the i^{th} and j^{th} column of X respectively. Then by definition of covariance, we can write :

$$\begin{aligned} C_{ij} &= \frac{1}{m}\sum_{k=1}^{m}(X_{ki}-\mu_i)(X_{kj}-\mu_j) \\ &= \frac{1}{m}\sum_{k=1}^{m}X_{ki}X_{kj} & (\because \mu_i = \mu_j = 0) \\ &= \frac{1}{m}X_i^TX_j = \frac{1}{m}(X^TX)_{ij} \end{aligned}$$

$$\hat{X} = XP$$

- Using the previous theorem & definition, we get $\frac{1}{m}\hat{X}^T\hat{X}$ is the covariance matrix of the transformed data. We can write :

$$\frac{1}{m}\hat{X}^T\hat{X} = \frac{1}{m}(XP)^T XP = \frac{1}{m}P^T X^T XP = P^T\left(\frac{1}{m}X^T X\right)P = P^T \Sigma P$$

- Each cell i, j of the covariance matrix $\frac{1}{m}\hat{X}^T\hat{X}$ stores the covariance between columns i and j of $\hat{X}$.

- Ideally we want,

$$\left(\frac{1}{m}\hat{X}^T\hat{X}\right)_{ij} = 0 \qquad\qquad i \neq j \text{ (covariance} = 0)$$

$$\left(\frac{1}{m}\hat{X}^T\hat{X}\right)_{ij} \neq 0 \qquad\qquad i = j \text{ (variance} \neq 0)$$

In other words, we want

$$\frac{1}{m}\hat{X}^T\hat{X} = P^T \Sigma P = D \qquad\qquad [\text{ where D is a diagonal matrix }]$$

- We want,

$$P^T \Sigma P = D$$

- But Σ is a square matrix and P is an orthogonal matrix
- Which orthogonal matrix satisfies the following condition?

$$P^T \Sigma P = D$$

- In other words, which orthogonal matrix P diagonalizes Σ?
- **Answer:** A matrix P whose columns are the eigen vectors of $\Sigma = X^T X$ [By Eigen Value Decomposition]
- Thus, the new basis P used to transform X is the basis consisting of the eigen vectors of $X^T X$

- Why is this a good basis?
- Because the eigen vectors of X^TX are linearly independent (**proof : Slide 19 Theorem 1**)
- And because the eigen vectors of X^TX are orthogonal ($\because$ X^TX is symmetric - **saw proof earlier**)
- This method is called Principal Component Analysis for transforming the data to a new basis where the dimensions are non-redundant (low covariance) & not noisy (high variance)
- In practice, we select only the top-k dimensions along which the variance is high (this will become more clear when we look at an alternalte interpretation of PCA)

Module 6.5 : PCA : Interpretation 2

Given n orthogonal linearly independent vectors $P = p_1, p_2, \cdots, p_n$ we can represent x_i exactly as a linear combination of these vectors.

$$x_i = \sum_{j=1}^{n} \alpha_{ij} p_j \quad \text{[we know how to estimate } \alpha'_{ij}s \text{ but we will come back to that later]}$$

But we are interested only in the top-k dimensions (we want to get rid of noisy & redundant dimensions)

$$\hat{x}_i = \sum_{j=1}^{k} \alpha_{ik} p_k$$

We want to select $p'_i s$ such that we minimise the reconstructed error

$$e = \sum_{i=1}^{m} (x_i - \hat{x}_i)^T (x_i - \hat{x}_i)$$

$$
\begin{aligned}
e &= \sum_{i=1}^{m} (x_i - \hat{x}_i)^T (x_i - \hat{x}_i) \\
&= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \alpha_{ij} p_j - \sum_{j=1}^{k} \alpha_{ij} p_j \right)^2 \\
&= \sum_{i=1}^{m} \left(\sum_{j=k+1}^{n} \alpha_{ij} p_j \right)^2 = \sum_{i=1}^{m} \left(\sum_{j=k+1}^{n} \alpha_{ij} p_j \right)^T \left(\sum_{j=k+1}^{n} \alpha_{ij} p_j \right) \\
&= \sum_{i=1}^{m} \left(\alpha_{i,k+1} p_{k+1} + \alpha_{i,k+2} p_{k+2} + \ldots + \alpha_{i,n} p_n \right)^T \left(\alpha_{i,k+1} p_{k+1} + \alpha_{i,k+2} p_{k+2} + \ldots + \alpha_{i,n} p_n \right) \\
&= \sum_{i=1}^{m} \sum_{j=k+1}^{n} \alpha_{ij} p_j^T p_j \alpha_{ij} + \sum_{i=1}^{m} \sum_{j=k+1}^{n} \sum_{L=k+1, L \neq k}^{n} \alpha_{ij} p_j^T p_L \alpha_{iL} \\
&= \sum_{i=1}^{m} \sum_{j=k+1}^{n} \alpha_{ij}^2 \qquad (\because p_j^T p_j = 1, p_i^T p_j = 0 \quad \forall i \neq j) \\
&= \sum_{i=1}^{m} \sum_{j=k+1}^{n} \left(x_i^T p_j \right)^2
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{m} \sum_{j=k+1}^{n} \left(p_j^T x_i \right) \left(x_i^T p_j \right) \\
&= \sum_{j=k+1}^{n} p_j^T \left(\sum_{i=1}^{m} x_i x_i^T \right) p_j \\
&= \sum_{j=k+1}^{n} p_j^T m C p_j \qquad \left[\because \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T = \frac{X^T X}{m} = C \right]
\end{aligned}
$$

We want to minimize e

$$\min_{p_{k+1},p_{k+2},\cdots,p_n} \sum_{j=k+1}^{n} p_j^T mCp_j \qquad s.t. \quad p_j^T p_j = 1 \quad \forall j = k+1, k+2, \cdots, n$$

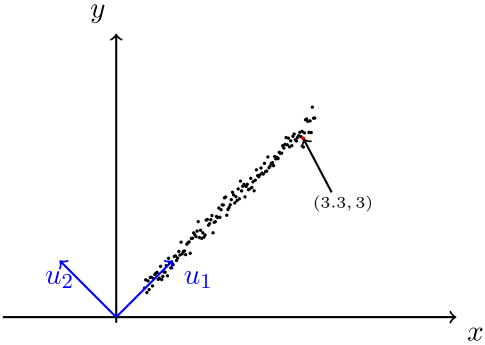
The solution to the above problem is given by the eigen vectors corresponding to the smallest eigen values of C (**Proof : refer Slide 26**).

Thus we select $P = p_1, p_2, \cdots, p_n$ as eigen vectors of C and retain only top-k eigen vectors to express the data [or discard the eigen vectors $k+1, \cdots, n$]

Key Idea

Minimize the error in reconstructing x_i after projecting the data on to a new basis.

Let's look at the ***'Reconstruction Error'*** *in the context of our toy example*



- $u_1 = [1, 1]$ and $u_2 = [-1, 1]$ are the new basis vectors
- Let us convert them to unit vectors $u_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$ & $u_2 = \begin{bmatrix} \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$

- Consider the point $x = [3.3, 3]$ in the original data
- $\alpha_1 = x^T u_1 = 6.3/\sqrt{2}$
  $\alpha_2 = x^T u_2 = -0.3/\sqrt{2}$
- the perfect reconstruction of x is given by (using $n = 2$ dimensions)

$$x = \alpha_1 u_1 + \alpha_2 u_2 = \begin{bmatrix} 3.3 & 3 \end{bmatrix}$$

- But we are going to reconstruct it using fewer (only $k = 1 < n$ dimensions, ignoring the low variance u_2 dimension)

$$\hat{x} = \alpha_1 u_1 = \begin{bmatrix} 3.15 & 3.15 \end{bmatrix}$$

(reconstruction with minimum error)

Recap

- The eigen vectors of a matrix with distinct eigenvalues are linearly independent
- The eigen vectors of a square symmetric matrix are orthogonal
- PCA exploits this fact by representing the data using a new basis comprising only the top-k eigen vectors
- The $n - k$ dimensions which contribute very little to the reconstruction error are discarded
- **These are also the directions along which the variance is minimum**

Module 6.6 : PCA : Interpretation 3

- We started off with the following wishlist
- We are interested in representing the data using fewer dimensions such that
  - the dimensions have low covariance
  - the dimensions have high variance
- So far we have paid a lot of attention to the covariance
- It has indeed played a central role in all our analysis
- But what about variance? Have we achieved our stated goal of high variance along dimensions?
- To answer this question we will see yet another interpretation of PCA

The i^{th} dimension of the transformed data $\hat{X}$ is given by

$$\hat{X}_i = Xp_i$$

The variance along this dimension is given by

$$\begin{aligned}
\frac{\hat{X}_i^T \hat{X}_i}{m} &= \frac{1}{m} p_i^T \underbrace{X^T X p_i} \\
&= \frac{1}{m} p_i^T \lambda_i p_i \qquad\qquad [\because p_i \text{ is the eigen vector of } X^T X] \\
&= \frac{1}{m} \lambda_i \underbrace{p_i^T p_i}_{=1} \\
&= \frac{\lambda_i}{m}
\end{aligned}$$

- Thus the variance along the i^{th} dimension (i^{th} eigen vector of $X^T X$) is given by the corresponding (scaled) eigen value.
- Hence, we did the right thing by discarding the dimensions (eigenvectors) corresponding to lower eigen values!

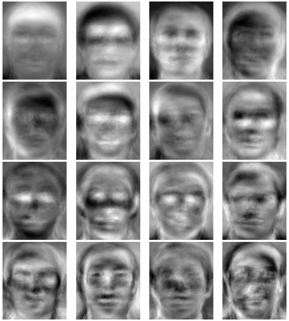
A Quick Summary

We have seen 3 different interpretations of PCA

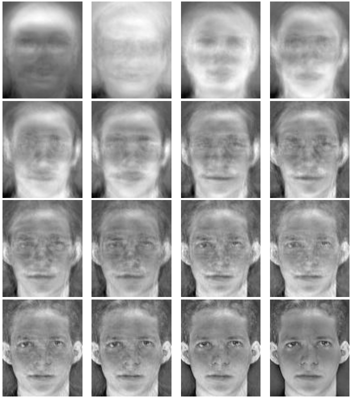
- It ensures that the covariance between the new dimensions is minimized
- It picks up dimensions such that the data exhibits a high variance across these dimensions
- It ensures that the data can be represented using less number of dimensions

Module 6.7 : PCA : Practical Example

- Consider we are given a large number of images of human faces (say, m images)
- Each image is 100×100 [10K dimensions]
- We would like to represent and store the images using much fewer dimensions (around 50-200)
- We construct a matrix $X \in \mathbb{R}^{m \times 10K}$
- Each row of the matrix corresponds to 1 image
- Each image is represented using 10K dimensions



- $X \in \mathbb{R}^{m \times 10K}$ (as explained on the previous slide)
- We retain the top 100 dimensions corresponding to the top 100 eigen vectors of X^TX
- Note that X^TX is a $n \times n$ matrix so its eigen vectors will be n dimensional ($n = 10K$ in this case)
- We can convert each eigen vector into a 100×100 matrix and treat it as an image
- Let's see what we get
- What we have plotted here are the first 16 eigen vectors of X^TX (basically, treating each 10K dimensional eigen vector as a 100×100 dimensional image)



$$\sum_{i=1}^{1}\alpha_{1i}p_i \sum_{i=1}^{2}\alpha_{1i}p_i \sum_{i=1}^{4}\alpha_{1i}p_i \sum_{i=1}^{8}\alpha_{1i}p_i \sum_{i=1}^{12}\alpha_{1i}p_i \sum_{i=1}^{16}\alpha_{1i}p_i$$

- These images are called eigenfaces and form a basis for representing any face in our database
- In other words, we can now represent a given image (face) as a linear combination of these eigen faces
- In practice, we just need to store $p_1, p_2, \cdots, p_k$ (one time storage)
- Then for each image i we just need to store the scalar values $\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{ik}$
- This significantly reduces the storage cost without much loss in image quality

Module 6.8 : Singular Value Decomposition

Let us get some more perspective on eigen vectors before moving ahead

- Let $v_1, v_2, \cdots, v_n$ be the eigen vectors of A and let $\lambda_1, \lambda_2, \cdots, \lambda_n$ be corresponding eigen values

$$Av_1 = \lambda_1 v_1, Av_2 = \lambda_2 v_2, \cdots, Av_n = \lambda_n v_n$$

- If a vector x in $\mathbb{R}^n$ is represented using $v_1, v_2, \cdots, v_n$ as basis then

$$x = \sum_{i=1}^{n} \alpha_i v_i$$

$$\text{Now, } Ax = \sum_{i=1}^{n} \alpha_i A v_i = \sum_{i=1}^{n} \alpha_i \lambda_i v_i$$

- The matrix multiplication reduces to a scalar multiplication if the eigen vectors of A are used as a basis.

- So far all the discussion was centered around square matrices ($A \in \mathbb{R}^{n \times n}$)
- What about rectangular matrices $A \in \mathbb{R}^{m \times n}$? Can they have eigen vectors?
- Is it possible to have $A_{m \times n} x_{n \times 1} = x_{n \times 1}$? Not possible !
- The result of $A_{m \times n} x_{n \times 1}$ is a vector belonging to $\mathbb{R}^m$ (whereas $x \in \mathbb{R}^n$)
- So do we miss out on the advantage that a basis of eigen vectors provides for square matrices (i.e. converting matrix multiplications into scalar multiplications)?
- We will see the answer to this question over the next few slides

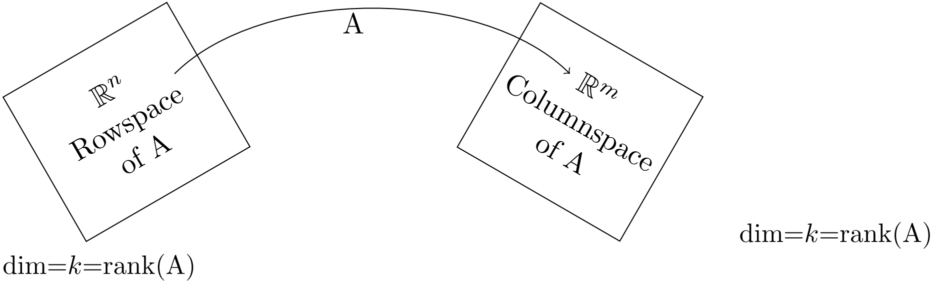
- Note that matrix $A_{m\times n}$ provides a transformation $\mathbb{R}^n \rightarrow \mathbb{R}^m$
- What if we could have pairs of vectors $(v_1, u_1), (v_2, u_2), \cdots, (v_k, u_k)$ such that $v_i \in \mathbb{R}^n$, $u_i \in \mathbb{R}^m$ and $Av_i = \sigma_i u_i$
- Further let's assume that $v_1, \cdots, v_k, \cdots, v_n$ are orthogonal & thus form a basis V in $\mathbb{R}^n$
- Similarly let's assume that $u_1, \cdots, u_k, \cdots, u_m$ are orthogonal & thus form a basis U in $\mathbb{R}^m$
- Now what if every vector $x \in \mathbb{R}^n$ is represented using the basis V

$$x = \sum_{i=1}^{k} \alpha_i v_i \qquad \text{[note we are using } k \text{ instead of } n \text{ ; will clarify this in a minute]}$$

$$\begin{aligned} Ax &= \sum_{i=1}^{k} \alpha_i A v_i \\ &= \sum_{i=1}^{k} \alpha_i \sigma_i u_i \end{aligned}$$

- Once again the matrix multiplication reduces to a scalar multiplication

Let's look at a geometric interpretation of this



- $\mathbb{R}^n$ - Space of all vectors which can multiply with A to give Ax [this is the space of inputs of the function]
- $\mathbb{R}^m$ - Space of all vectors which are outputs of the function Ax
- We are interested in finding a basis U, V such that
  - V - basis for inputs
  - U - basis for outputs
- such that if the inputs and outputs are represented using this basis then the operation Ax reduces to a scalar operation

- What do we mean by saying that dimension of rowspace is k? If $x \in \mathbb{R}^n$ then why is the dimension not n.
- It means that of all the possible vectors in $\mathbb{R}^n$ only a subspace of vectors lying in $\mathbb{R}^k$ can act as inputs to Ax and produce a non-zero output. The remaining vectors in $\mathbb{R}^{n-k}$ will produce a zero output
- Hence we need only k dimensions to represent x

$$x = \sum_{i=1}^{k} \alpha_i v_i$$

- Let's look at a way of writing this as a matrix operation

$$Av_1 = \sigma_1 u_1, Av_2 = \sigma_2 u_2, \cdots, Av_k = \sigma_k u_k$$

$$A_{m\times n} V_{n\times k} = U_{m\times k} \underbrace{\Sigma_{k\times k}}_{\text{diagonal matrix}}$$

- If we have k orthogonal vectors $(V_{n\times k})$ then using Gram Schmidt orthogonalization, we can find $n-k$ more orthogonal vectors to complete the basis for $\mathbb{R}^n$ [We can do the same for U]

$$A_{m\times n} V_{n\times n} = U_{m\times m} \Sigma_{m\times n}$$

$$U^T AV = \Sigma \qquad [U^{-1} = U^T] \qquad A = U\Sigma V^T \qquad [V^{-1} = V^T]$$

- Σ is a diagonal matrix with only the first k diagonal elements as non-zero
- Now the question is how do we find V, U and Σ

- Suppose V, U and Σ exist, then

$$\begin{aligned} A^TA &= (U\Sigma V^T)^T(U\Sigma V^T) \\ &= V\Sigma^T U^T U\Sigma V^T \\ A^TA &= V\Sigma^2 V^T \end{aligned}$$

- What does this look like? Eigen Value decomposition of A^TA
- Similarly we can show that

$$AA^T = U\Sigma^2 U^T$$

- Thus U and V are the eigen vectors of AA^T and A^TA respectively and $\Sigma^2 = \Lambda$ where Λ is the diagonal matrix containing eigen values of A^TA

$$\begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix}_{m\times n} = \begin{bmatrix} \uparrow & \cdots & \uparrow \\ & & \\ u_1 & \cdots & u_k \\ \downarrow & \cdots & \downarrow \end{bmatrix}_{m\times k} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix}_{k\times k} \begin{bmatrix} \leftarrow & v_1 & \rightarrow \\ & \vdots & \\ \leftarrow & v_k & \rightarrow \end{bmatrix}_{k\times n}$$

$$= \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

Theorem:

$\sigma_1 u_1 v_1^T$ is the best rank-1 approximation of the matrix A. $\sum_{i=1}^{2} \sigma_i u_i v_i^T$ is the best rank-2 approximation of matrix A. In general, $\sum_{i=1}^{k} \sigma_i u_i v_i^T$ is the best rank-k approximation of matrix A. In other words, the solution to

$$\min \|A - B\|_F^2 \quad \text{is given by :}$$

$$B = U_{.,k} \Sigma_{k,k} V_{k,.}^T \quad \text{(minimizes reconstruction error of A)}$$

$$\sigma_i = \sqrt{\lambda_i} = \text{singular value of A}$$
$$U = \text{left singular matrix of A}$$
$$V = \text{right singular matrix of A}$$