LINEAR ALGEBRA

MATRIX INVERSE → denoted by A^{-1} and defined as $A^{-1}A = I$

To solve system of linear eqns →

$$Ax = b$$
$$A^{-1}Ax = A^{-1}b$$
$$Ix = A^{-1}b$$

LINEAR DEPENDENCE AND SPAN →

* For A^{-1} to exist $Ax=b$ must have exactly one solution for every value of b.
* It can have no solution, one solution or infinitely many solution. It can not have more than one but less than ∞ solution because if x and y are solutions then so is $z = \alpha x + (1-\alpha)y$ for any real α.
* Essentially b is linear combination of columns of A i.e we are trying to find if b can be formed using linear combination of columns of A.

$$\sum_i x_i \cdot A_{:,i} = b$$

* SPAN → set of all points obtainable by linear combination of original vectors.
  - ⇒ To find solution of $Ax=b$ means whether b lies in column span of A or not.
  - ⇒ For $b \in \mathbb{R}^m$ to lie in column span of A, the column span must be $\mathbb{R}^m$.

* LINEAR INDEPENDENCE → A set of vectors is linearly independent if no vector is linear combination of a subset of vectors.

⇒ For column span of A to be R^m there must exist m linearly independent vectors i.e. for $A_{m \times n}$, $n \geq m$.
Also to have at most one solution $n = m$, making A a square matrix.

* SINGULAR MATRIX → A square matrix with linearly dependent columns.

NORMS →

* L^p norm ⇒ $\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$
* Popular norms →
  1. L^2 norm → $\|x\|_2 = \left(\sum_i |x_i|^2\right)^{1/2}$ → Often we use squared L^2 norm.
  2. L^1 norm → $\|x\|_1 = \sum_i |x_i|$
  3. L^∞ norm → $\|x\|_\infty = \max_i |x_i|$ → As p increases x_i with max value will dominate the sum. (just an intuition)
  3. Forbenius norm → Norm of matrix

$$\|A\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}$$

* Unit Vector → A vector with unit L_2 norm i.e. $\|x\|_2 = 1$
* Orthogonal vectors → Two vectors x and y are orthogonal if $x^T y = 0$
* Orthnormal vectors → if $x^T y = 0$ and $\|x\|_2 = 1$ and $\|y\|_2 = 1$

* Orthogonal matrix → Matrix with orthogonal rows i.e.

$$A^T A = A A^T = I$$

$$\Rightarrow \quad A^{-1} = A^T$$

EIGEN DECOMPOSITION →

* Just like integers can be decomposed into prime factors, matrices can be decomposed into other matrices that reveal properties.
* An eigenvector of a square matrix A is a non zero vector v such that multiplication by A alters only scale of v.

i.e. $Av = \lambda v$

eigen value

eigen vector

* If v is eigenvector of A then so is $s.v$ for $s \in R, s \neq 0$ with same eigen value.
* Eigendecomposition of A →

$$A = V \operatorname{diag}(\lambda) V^{-1}$$

$$V \rightarrow [v^{(1)}, v^{(2)}, \ldots, v^{(n)}]$$

$$\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_n]$$

* Every real symmetric matrix A →

$$A = Q \Lambda Q^T$$

$Q \rightarrow$ Orthogonal matrix of eigenvectors of A

$\Lambda \rightarrow$ diagonal matrix of eigenvalues of A.

⇒ A scales the spaces by λ_i in the direction $v^{(i)}$.

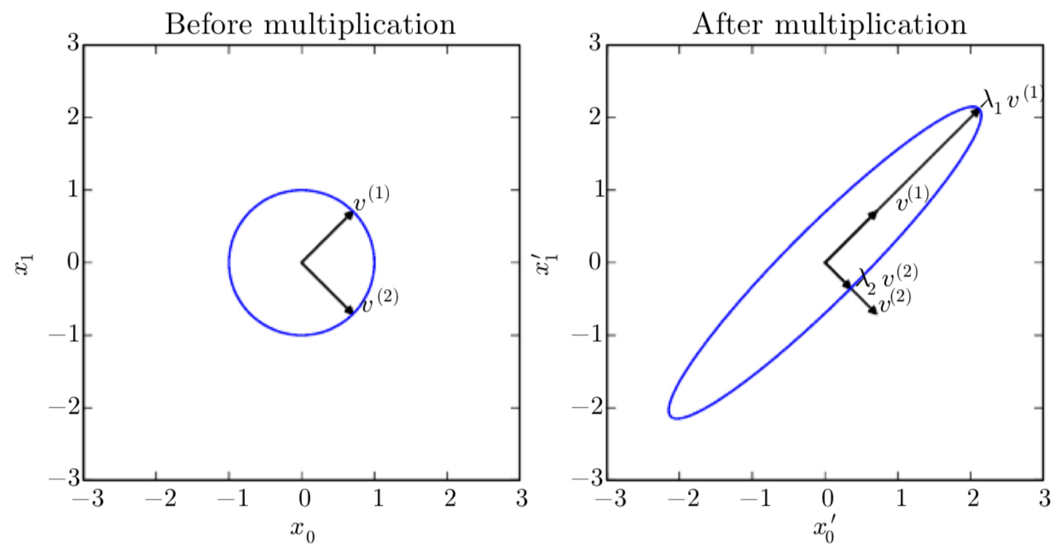


Figure 2.3: An example of the effect of eigenvectors and eigenvalues. Here, we have a matrix $\boldsymbol{A}$ with two orthonormal eigenvectors, $\boldsymbol{v}^{(1)}$ with eigenvalue λ_1 and $\boldsymbol{v}^{(2)}$ with eigenvalue λ_2. *(Left)*We plot the set of all unit vectors $\boldsymbol{u} \in \mathbb{R}^2$ as a unit circle. *(Right)*We plot the set of all points $\boldsymbol{Au}$. By observing the way that $\boldsymbol{A}$ distorts the unit circle, we can see that it scales space in direction $\boldsymbol{v}^{(i)}$ by λ_i.

* A matrix is singular iff any of the eigenvalues are 0.

* A matrix whose eigenvalues are all +ve → Positive Definite
  A matrix whose eigenvalues are all -ve → Negative Definite
  If positive or 0 → Positive semidefinite
  If negative or 0 → Negative semidefinite

* For positive semidefinite matrix →

$$\forall x \quad x^T A x \geq 0$$

→ we can use this property to prove if a matrix A is +ve semi definite or not.

SINGULAR VALUE DECOMPOSITION →

* SVD decomposes a matrix into singular values and vectors.
* $A = UDV^T$

  $U \rightarrow m \times m$ (Left singular vectors)

  $D \rightarrow m \times n$ (Diagonal of D are Singular values of A)

  $V \rightarrow n \times n$ (Right singular vectors)

MOORE-PENROSE PSEUDOINVERSE →

If A is not invertible then we can compute its pseudo inverse using SVD as

$$A^+ = VD^+U^T$$

$D^+ =$ reciprocal of elements of diagonal matrix D

TRACE AND DETERMINANT →

* Trace → $\mathrm{Tr}(A) = \sum_i A_{i,i}$

$$\|A\|_F = \sqrt{\mathrm{Tr}(AA^T)}$$

$$\mathrm{Tr}(A) = \mathrm{Tr}(A^T)$$

$$\mathrm{Tr}(ABC) = \mathrm{Tr}(CAB) = \mathrm{Tr}(BCA)$$

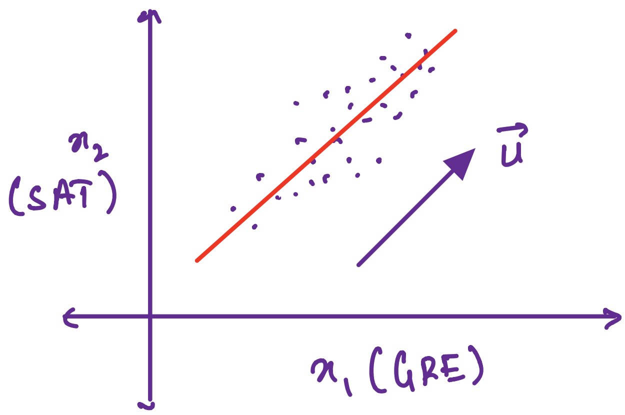
More generally, $\mathrm{Tr}\left(\prod_{i=1}^{n} F^{(i)}\right) = \mathrm{Tr}\left(F^{(n)} \cdot \prod_{i=1}^{n-1} F^{(i)}\right)$

$$\Rightarrow \mathrm{Tr}(AB) = \mathrm{Tr}(BA)$$

* Determinant → Product of all eigen values of A. For singular matrix, $\det(A) = 0$

PRINCIPAL COMPONENT ANALYSIS

* Say we have data $\{x^{(i)}\}_{i=1}^{m}$



- There is high correlation b/w x_1 and x_2, i.e. even if data is in 2D it is more or less in 1D.

* Underlying question is can we do dimensionality reduction i.e. we project $x^{(i)}$ on $\vec{u}$ to get $\{z^{(i)}\}_{i=1}^{m}$ where $z^{(i)} = u^T x^{(i)}$ such that most of the "information" in data is still captured.
* Unless the data is perfectly correlated there will be some loss.
* We want to find best set of $\{u_1 \dots u_{k}\}$

Q Why project the data?
1. Discovering hidden patterns in the data (correlation etc)
2. Projecting onto a lower dimensional space makes things tractable. For $x^{(i)} \in R^n$ $k << n$.
3. Helps in reducing noise in data.

How to quantify loss of information → Variance

$$Var(\{x^{(i)}\}_{i=1}^{m}) = \sum_{j=1}^{n} \left[\frac{1}{m} \sum_{i=1}^{m} (x_j^{(i)} - \mu_j^{(i)})^2 \right]$$

Let $\{z^{(i)}\}_{i=1}^{m}$ → projection over $\{u_1, \dots u_k\}$ then $Var(\{z^{(i)}\}_{i=1}^{m})$ should be close to variance of original data.

<u>Problem Statement</u> → Given $\{x^{(i)}\}_{i=1}^{m}$ $x^{(i)} \in \mathbb{R}^n$.

Find $(u_1, \dots, u_k)$ $u_l \in \mathbb{R}^n \ \forall l$, $u_l^T u_l = 1$

s.t.

$x^{(i)} \sim z^{(i)}$ → projection of $x^{(i)}$ on $\{u_l\}$

$\Rightarrow z_l^{(i)} = x^{(i)T} u_l$

s.t $\mathrm{var}(\{z^{(i)}\}_{i=1}^{m})$ is maximized

① We first normalize the points so that resulting data has 0 mean and unit variance. (This is so that all features are on same "scale" eg height and weight are on same scale)

$$x_j^{(i)} \leftarrow \left(\frac{x_j^{(i)} - \mu_j}{\sigma_j}\right)$$

mean, $\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)} \ \forall j = 0$

variance, $\mathrm{var} = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)2} = 1$

After transformation.

mean of z along dimension l, $\mu_l^z = \frac{1}{m} \sum_{i=1}^{m} x^{(i)T} u_l$

$$= \left[\frac{1}{m} \sum_{i=1}^{m} x^{(i)}\right]^T u_l$$

$$= \mu^T u_l = 0 \cdot u_l = 0$$

$Var(\{z^{(i)}\}_{i=1}^{m}) = \sum_{l=1}^{K} \frac{1}{m} \sum_{i=1}^{m} (x^{(i)T} u_l)^2$ ($\because$ mean is 0)

$$= \sum_{l=1}^{K} \frac{1}{m} \sum_{i=1}^{m} (u_l^T x^{(i)})(x^{(i)T} u_l)$$

$$= \sum_{l=1}^{K} u_l^T \underbrace{\left[\frac{1}{m} \sum_{i=1}^{m} x^{(i)} \cdot x^{(i)T} \right]}_{\text{Emperical covariance matrix}} u_l$$

$$= \sum_{l=1}^{K} u_l^T \Sigma u_l \qquad \left(\text{Here } \Sigma_{jk} = \underbrace{\frac{1}{m} \sum_{i=1}^{m} x_j^{(i)} x_k^{(i)}}_{\text{emperical\_covariance}(x_j,\, x_k)} \right)$$

Our objective: $\underset{u_1, \ldots, u_l}{\operatorname{argmax}} \sum_{l=1}^{K} u_l^T \Sigma u_l$

subject to ① $\|u_l\|_2^2 = 1$

② $u_{l_1}^T u_{l_2} = 0 \quad \forall\, l_1 \neq l_2$

Assume $K=1$, result then generalizes

Objective $\to \max_{u_l} u_l^T \Sigma u_l$ subject to $u_l^T u_l = 1$ } constrained optimization problem

We will use Langragian to solve above problem.

$$\mathcal{L}(u_l, \lambda) = u_l^T \Sigma u_l + \lambda [1 - u_l^T u_l]$$

Primal $\to$ $\max_{u_l}\left(\min_{\lambda} L(u_l,\lambda)\right) = \max_{u_l}\min_{\lambda} u_l^T \Sigma u_l + \lambda[1-u_l^T u_l]$

Dual $\to$ $\min_{\lambda}\left(\max_{u_l} L(u_l,\lambda)\right) = \min_{\lambda}\max_{u_l} u_l^T \Sigma u_l + \lambda[1-u_l^T u_l]$

Maximizing wrt u_l gradient should vanish $\to$

$$\nabla_{u_l}\left[u_l^T \Sigma u_l - \lambda(1+u_l^T u_l)\right] = 2\Sigma u_l - 2\lambda u_l = 0$$

$$\Rightarrow \quad \Sigma u_l = \lambda u_l$$

Σ: $n\times n$ matrix, $u_l \in \mathbb{R}^n$, λ: scalar

$\Rightarrow$ λ is eigenvalue of Σ and u_l is the eigenvector

$$\underset{u_1,\dots,u_K}{\operatorname{argmax}} \sum_{l=1}^{K} u_l^T \Sigma u_l \qquad \text{constraints} \to \quad u_l^T u_l = 1 \quad \forall l$$
$$u_{l_1}^T u_{l_2}^T = 0 \quad l_1 \neq l_2$$

Solution to above problem :-
- u_l are eigen vectors of Σ
- λ_l are eigen values of Σ that capture amount of variance among u_l.

\# Principal components are vectors corresponding to largest k eigen values of Σ

$\Rightarrow$ Objective reduces to find eigen values and eigen vectors of Σ and keep $u_1,\dots,u_k$ corresponding to top k eigenvalues of Σ.

$\Rightarrow \lambda_l$ is proportional to amount of variance captured in u_l.

To find eigen value and eigenvectors $\rightarrow$ ① Eigen Decomposition
② Singular Value Decomposition.

We can write $\Sigma = \frac{1}{m} X^T X$ where X is the data matrix.

$\Rightarrow$ Problem reduces to finding Top k eigenvalues of $\frac{1}{m} X^T X$

Let $A = \frac{1}{m} X^T X$

$$A = Q \Lambda Q^{-1} = Q \Lambda Q^T \quad \text{as } Q \text{ are orthonormal}$$

Challenge is that computing $A = \frac{1}{m} X^T X \rightarrow O(n^2 m)$
and eigendecomposition will be $\rightarrow O(n^3)$
if num of examples $n >> m \rightarrow$ too expensive.

Using SVD $\rightarrow$

$$\underset{(m \times m)}{U} \; \underset{(m \times n)}{D} \; \underset{(n \times n)}{V^T} = A$$

U and V are orthonormal

- Columns of V are eigenvectors of A
- Entries of D are square root of eigen values of A.

Complexity of SVD $\rightarrow O(\min(m^2 n, n^2 m))$

Q $X = UDV^T$ show that X^TX, show that columns of V are eigenvector of X^TX and Diagonal entries of D^TD are eigenvalues of X^TX.

$$X^TX = (UDV^T)^T(UDV^T)$$
$$= VD^TU^TUDV^T$$
$$= VD^TDV^T \qquad (\because U^TU = I)$$