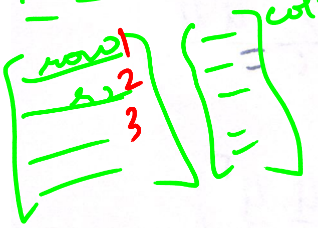
$(AB)^T = B^T A^T$ Variance $= \frac{1}{m}\sum x_i^2 - \mu^2$ $v^T C v \geq 0$

$$= \left[\frac{(\underline{X}\,\underline{v})^T(\underline{X}\,\underline{v})}{m} - (\underline{\mu}^T\underline{v})^T(\underline{\mu}^T\underline{v})\right] \geq 0$$

Variance is always non neg $\Rightarrow$ $\underline{C}$ is psd.

$X\underline{v}$ $\longrightarrow v$



variance of the projection (1D projection) of the row vectors in $\underline{X}$ on the column vector $\underline{v}$. ≥ 0

$\underline{\mu}^T\underline{v}$

$\underline{\mu}^T\underline{v}$ is the projection of the mean vector on the direction $\underline{v}$. This will also be the mean of the 1-D projections of the examples $\underline{x}_i$'s on $\underline{v}$.

The goal of PCA is to one-by-one determine the orthonormal vectors $\underline{v}$ maximizing $\underline{v}^T \underline{C}\, \underline{v}$. variance along direction $\underline{v}$

Because the covariance matrix is symmetric and positive semidefinite, we can obtain its eigen value decomposition

$$C = V \Lambda V^T$$

V: orthonormal eigen vectors of C as columns.

Λ: diagonal matrix of eigen values

Λ_{ii}: eigen value corresponding to i^{th} eigenvector of V.

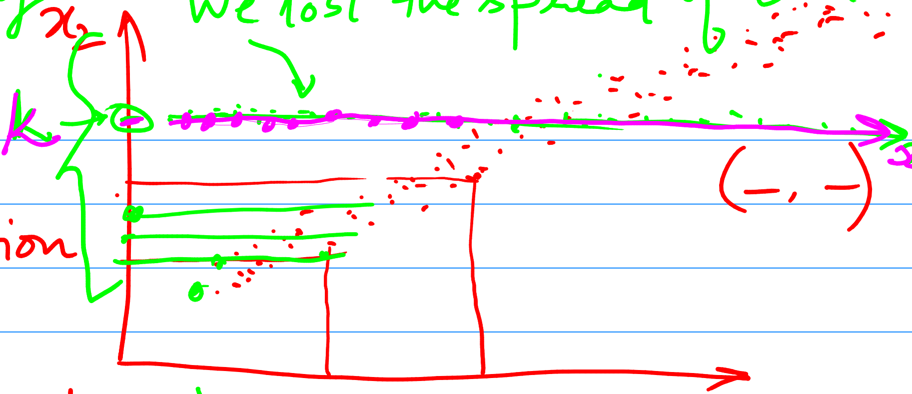
Objective is to maximize $\underbrace{\underline{v}^T \underline{C}\, \underline{v}}_{\text{variance}}$ subject to $\underline{v}$ being a unit vector. $\|v\|^2 = 1$

Formulating the Lagrangian $\boxed{v^T \underline{C} v - \lambda(\|\underline{v}\|^2 - 1)}$ λ: lagrange parameter

Taking derivative w.r.t. $\underline{v}$ and equating it to zero gives

Dropping Axes. Naive way

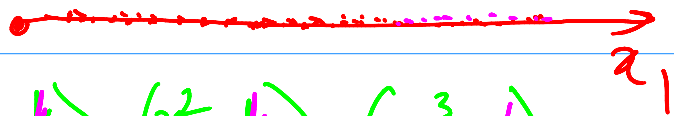
we lost the spread of data along x_2



1 D. loss of Information

Representation using Orthonormal basis (x_1, x_2).

2D to 1D. $x_2 = k$.



1-D.

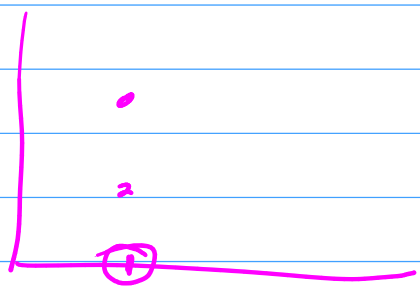
(x_1^1, k) (x_1^2, k) (x_1^3, k).

x_2 is frozen to the mean value

k: mean value of the projections of data points on $\underbrace{\text{axis } x_2}_{\text{(discarded)}}$.

$\underline{x^1}$ $(2, 5)$

$\underline{x^2}$ $(2, 3)$

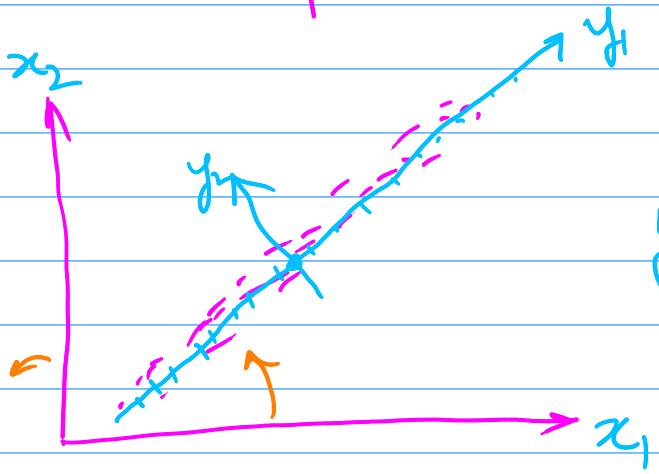


PCA

Rotation of Axes to a new orthonormal basis.

Then drop axes which are not 'important'

Original Representation

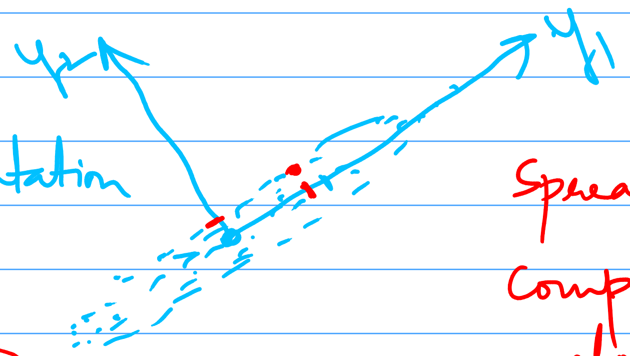


y_2 can be dropped without losing much information

Transformation by Axis rotation

x_1

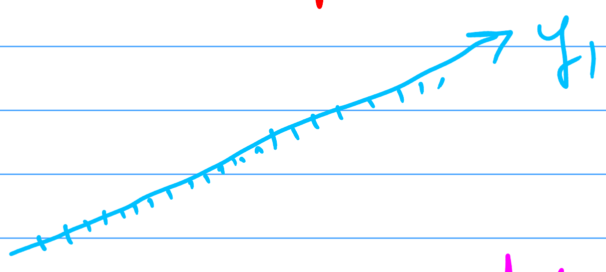
New representation of the data



Spread along y_1 is larger compared to the spread along y_2.

eg1	x_1^1	x_2^1	y_1^1	y_2^1
eg2	x_1^2	x_2^2	y_1^2	y_2^2
eg3	x_1^3	x_2^3	y_1^3	y_2^3

Discard y_2



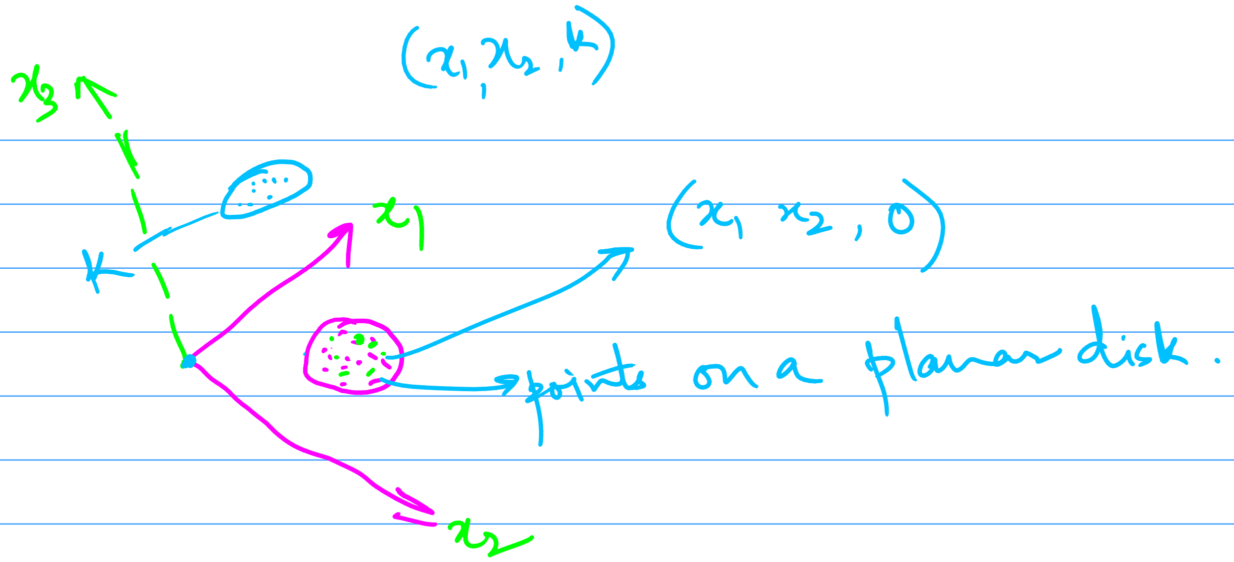
input space basis vectors

linear transformation

rotated basis vectors

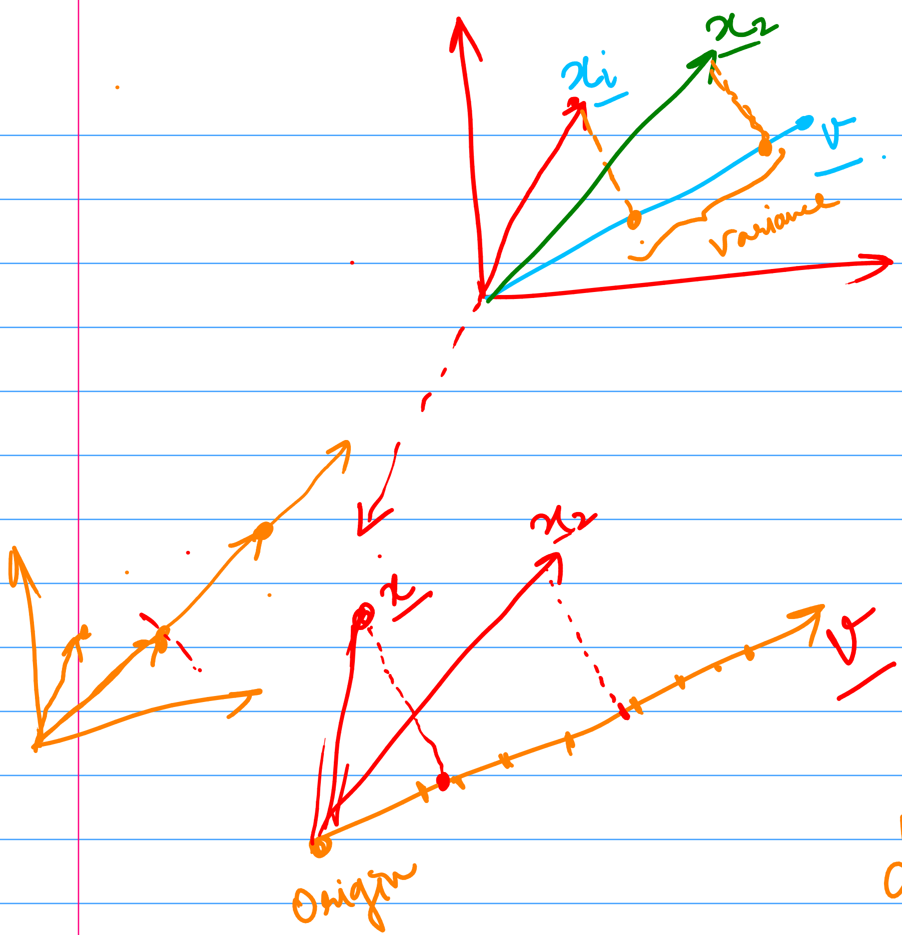
PCA.

dim reduction



$\underbrace{\text{planar disk}}_{\text{2 features}}$ embedded in a $\underbrace{\text{3D space}}$.

1D projection



$v^T C v$

variance of all the points (examples) along direction $\underline{v}$

$\underline{v} = (- - -)$

direction

$\underline{v}$ defines a one-dim space

projecting any vector on $\underline{v}$ $\Rightarrow$ 1-D projection

$v^T C v - \lambda (\|v\|^2 - 1)$

$-\lambda(v^T v - 1)$ $\quad C \underline{v} - \lambda \underline{v} = 0$

$$\boxed{C \underline{v} = \lambda \underline{v}}$$

This is the Eigen vector equation.

The solution $\underline{v}$ is an eigenvector of matrix C.

The variance along the eigen vector $\underline{v}$ is

$$\underline{v}^T \underbrace{C \underline{v}}_{} = \underline{v}^T \underbrace{\lambda \underline{v}}_{} = \lambda \qquad \because \underline{v}^T \underline{v} = 1.$$

$\lambda v^T v = \lambda$

$\therefore$ the eigen value corresponding to the eigenvector $\underline{v}$ gives the variance.

The orthonormal set of eigenvectors are obtained by the eigenvalue decomposition $C = V \Lambda V^T$ where the columns of the matrix V are the eigenvectors and Λ is a diagonal matrix with eigenvalues $\lambda_1, \lambda_2 \ldots \lambda_d$ corresponding to the eigenvectors in V.

Projecting the data onto the eigenvectors is equivalent to transforming the data to a new representation. $x^T \underline{v}$

The eigenvectors form the orthonormal basis of the new representation.

The effect of the transformation is a rotation of the axis system (basis vectors) of the original data representation to ~~the~~ ~~are~~ align with the new basis vectors, which are the eigenvectors.