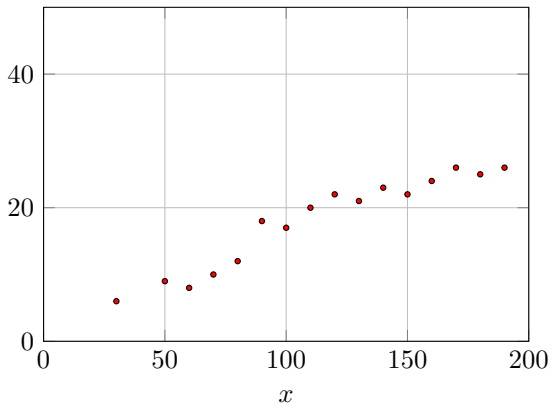
Chapter 13
Principal Component Analysis

1 Problem

Consider this plot

X	Y
30	6
50	9
60	8
70	10
80	12
90	18
100	17
110	20
120	22
130	21
140	23
150	22
160	24
170	26
180	25
190	26

$\Longrightarrow$

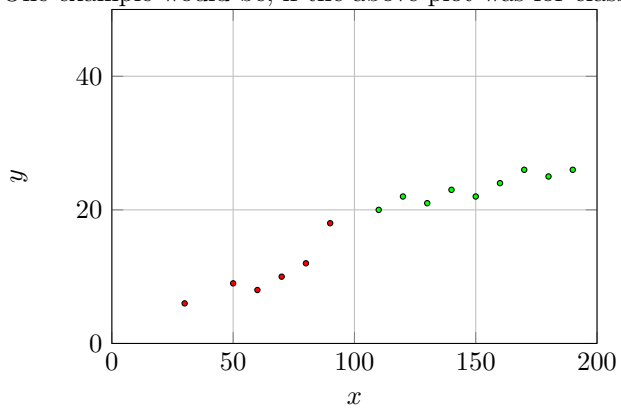


This data uses two dimensions to represent the information about something.
But do we really need 2 dimensions...?
and are these 2 axis(basis) the most informative..?

Ideal Basis Properties

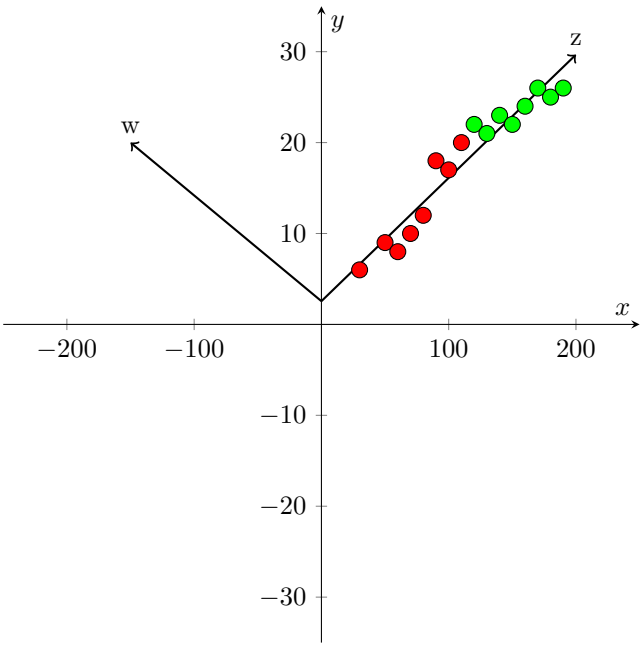
- High Variance along individual basis or axis
- Low Covariance between basis (This means they are linearly independent)
- Non redundant Basis(Less Correlation between basis)

Also we are interested in represting the data using fewer dimensions such that the data has high variance along these dimensions.
One example would be, if the above plot was for classfication.



Green points represent class A
Red points represent class B

Now, what if we chose some other axis instead of these x and y.



On this new z axis, all points seem to lie pretty close to it, with negligible variance along the w axis
and since this data was for classification, we can actually represent this 2D data in just 1 Dimension on Z axis

2 Setup

Let $p_1, p_2, p_3, \ldots, p_n$ be a set of linearly independent orthonormal vectors.
These are going to be our new ideal basis vectors.
Let P be n x n matrix such that $p_1, p_2, p_3, \ldots, p_n$ are the columns of P.

$$P = \begin{bmatrix} \uparrow & \uparrow & \ldots & \uparrow \\ p_1 & p_2 & \ldots & p_n \\ \downarrow & \downarrow & \ldots & \downarrow \end{bmatrix}$$

Let $x_1, x_2, \ldots, x_m \in R^n$ be the sample vectors of size (n x 1) from the database of m data points.
X is the matrix such that $x_1, x_2, \ldots, x_m$ are the columns of it.

$$X = \begin{bmatrix} \uparrow & \uparrow & \ldots & \uparrow \\ x_1 & x_2 & \ldots & x_m \\ \downarrow & \downarrow & \ldots & \downarrow \end{bmatrix}$$

Further let us assume that the data is 0-mean and unit-variance.

3 Finding the ideal basis

Let us assume currently the basis of x_{ij} is $\hat{b_j}$
Hence,

$$x_i = x_{i1}\hat{b_1} + x_{i2}\hat{b_2} + \cdots + x_{in}\hat{b_n}$$

Since P is the matrix of new basis.
This means each vector of X can be written as a linear combination of these new basis vectors

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

$$x_i = \begin{bmatrix} \alpha_{i1} \\ \alpha_{i2} \\ \vdots \\ \alpha_{in} \end{bmatrix}$$

and for orthonormal basis we know that we can find α_i's using

$$\alpha_{ij} = x_i^T \cdot p_j$$

$$\alpha_{ij} = \begin{bmatrix} x_{i1} & x_{i2} & \dots & x_{in} \end{bmatrix} \cdot \begin{bmatrix} p_{j1} \\ p_{j2} \\ \vdots \\ p_{jn} \end{bmatrix}$$

Hence,

$$\hat{x_i} = \begin{bmatrix} x_i^T \cdot p_1 \\ x_i^T \cdot p_2 \\ \vdots \\ x_i^T \cdot p_n \end{bmatrix} = \begin{bmatrix} \longleftarrow & x_i & \longrightarrow \end{bmatrix} \cdot \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} \longleftarrow & x_i & \longrightarrow \end{bmatrix} \cdot \begin{bmatrix} \uparrow & \uparrow & \dots & \uparrow \\ p_1 & p_2 & \dots & p_n \\ \downarrow & \downarrow & \dots & \downarrow \end{bmatrix}$$

So, any vector new basis would look like

$$\hat{x_i} = x_i^T \cdot P$$

All the vectors would look like

$$\hat{X} = \begin{bmatrix} \uparrow & \uparrow & \dots & \uparrow \\ \hat{x_1} & \hat{x_2} & \dots & \hat{x_m} \\ \downarrow & \downarrow & \dots & \downarrow \end{bmatrix} = \begin{bmatrix} \uparrow & \uparrow & \dots & \uparrow \\ x_1^T.P & x_2^T.P & \dots & x_m^T.P \\ \downarrow & \downarrow & \dots & \downarrow \end{bmatrix} = \begin{bmatrix} \uparrow & \uparrow & \dots & \uparrow \\ x_1^T & x_2^T & \dots & x_m^T \\ \downarrow & \downarrow & \dots & \downarrow \end{bmatrix} \cdot \begin{bmatrix} \uparrow & \uparrow & \dots & \uparrow \\ p_1 & p_2 & \dots & p_m \\ \downarrow & \downarrow & \dots & \downarrow \end{bmatrix}$$

$$\hat{X} = X \cdot P$$

Theorem:
If X is a matrix such that its column has zero mean and if $\hat{X} = XP$ then columns of $\hat{X}$ would also have zero mean.

Theorem:
If X is a matrix whose columns have zero mean then $\Sigma = \frac{X^T X}{m}$ is the Covariance matrix.
$\Sigma_{i,j}$ stores the covariance between x_i and x_j

$$\Sigma = \frac{\begin{bmatrix} \longleftarrow & x_1 & \longrightarrow \\ \longleftarrow & x_2 & \longrightarrow \\ \vdots & & \\ \longleftarrow & x_m & \longrightarrow \end{bmatrix} \cdot \begin{bmatrix} \uparrow & \uparrow & \dots & \uparrow \\ x_1 & x_2 & \dots & x_m \\ \downarrow & \downarrow & \dots & \downarrow \end{bmatrix}}{m} = \frac{\begin{bmatrix} \longleftarrow x_1.x_1 \longrightarrow & \longleftarrow x_1.x_2 \longrightarrow & \dots & \longleftarrow x_1.x_m \longrightarrow \\ \longleftarrow x_2.x_1 \longrightarrow & \longleftarrow x_2.x_2 \longrightarrow & \dots & \longleftarrow x_2.x_m \longrightarrow \\ \vdots & \vdots & \ddots & \vdots \\ \longleftarrow x_m.x_1 \longrightarrow & \longleftarrow x_m.x_2 \longrightarrow & \dots & \longleftarrow x_m.x_m \longrightarrow \end{bmatrix}}{m}$$

$$\Sigma = \begin{bmatrix} Cov(x_1, x_1) & Cov(x_1, x_2) & \dots & Cov(x_1, x_m) \\ Cov(x_2, x_1) & Cov(x_2, x_2) & \dots & Cov(x_2, x_m) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(x_m, x_1) & Cov(x_m, x_2) & \dots & Cov(x_m, x_m) \end{bmatrix}$$

Since $\hat{X}$ is the data with new basis

Covariance matrix of $\hat{X} = \frac{\hat{X}^T.\hat{X}}{m}$

$$\frac{\hat{X}^T.\hat{X}}{m} = \frac{(X.P)^T.(X.P)}{m} = P^T.\frac{(X^T.X)}{m}.P = P^T\Sigma P$$

Ok, we want basis vectors to be linearly independent, this means all basis vectors should have 0 covariance and we also want vectors to have non-zero variance with itself.

Ideal Basis Properties

- $\left(\frac{\hat{X}^T.\hat{X}}{m}\right)_{(i,j)}$ =0 if i≠ j, meaning low covariance
- $\left(\frac{\hat{X}^T.\hat{X}}{m}\right)_{(i,j)}$ ≠0 if i=j, meaning non-zero variance

Hence,

$$\frac{\hat{X}^T.\hat{X}}{m} = \begin{bmatrix} Cov(\hat{x_1},\hat{x_1}) & Cov(\hat{x_1},\hat{x_2}) & \dots & Cov(\hat{x_1},\hat{x_m}) \\ Cov(\hat{x_2},\hat{x_1}) & Cov(\hat{x_2},\hat{x_2}) & \dots & Cov(\hat{x_2},\hat{x_m}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\hat{x_m},\hat{x_1}) & Cov(\hat{x_m},\hat{x_2}) & \dots & Cov(\hat{x_m},\hat{x_m}) \end{bmatrix}$$

will become

$$\frac{\hat{X}^T.\hat{X}}{m} = \begin{bmatrix} Cov(x_1,x_1) & 0 & \dots & 0 \\ 0 & Cov(x_2,x_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Cov(x_m,x_m) \end{bmatrix}$$

This means

$\frac{\hat{X}^T.\hat{X}}{m} = P^T \Sigma P$ is a Diagonal matrix.

And we know,
Σ is a Square matrix
P is an orthogonal matrix

The questions is, Which orthogonal matrix can diagonalise Σ?
Answer is, a matrix whose columns are eigenvectors of Σ.

Also,

$$P^T \Sigma P = \Lambda$$

Here Λ is a diagonal matrix containing Eigen-values of Σ matrix.

$$P = \begin{bmatrix} \uparrow & \uparrow & \dots & \uparrow \\ \vec{v_1} & \vec{v_2} & \dots & \vec{v_n} \\ \downarrow & \downarrow & \dots & \downarrow \end{bmatrix}$$

Here, $\vec{v_1}, \vec{v_2}, \dots, \vec{v_n}$ are the eigenvectors of Σ.

Hence, we found the ideal basis vector values with these properties.

Ideal Basis Properties

- High Variance along individual basis or axis
- Low Covariance between basis (This means they are linearly independent)
- Non redundant Basis(Less Correlation between basis)

That is, the new basis P used to transform X is the basis consisting of Eigenvectors of $X^T X$

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).