Dimensionality Reduction: PCA

High Dimensional Data

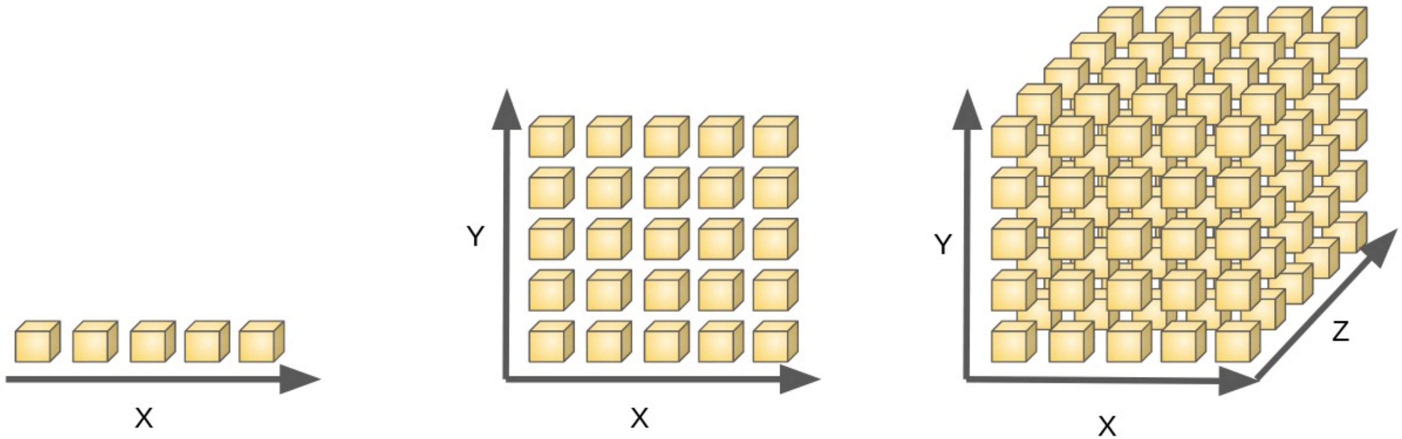
- **With technological advancements, a vast amount of data has been collected, both in terms of the number of samples and variables.**

- **Problems with high dimensional data**
  - There may be many irrelevant variables in the high dimensional data.
  - In some cases, some variables are closely correlated, resulting in multicollinearity.
  - It is difficult to visualize the data and analytic results.
  - As the number of variables increases, the computational costs of models increase.

- **We can consider reducing the dimensionality of data. (Dimensionality reduction)**

The Curse of Dimensionality

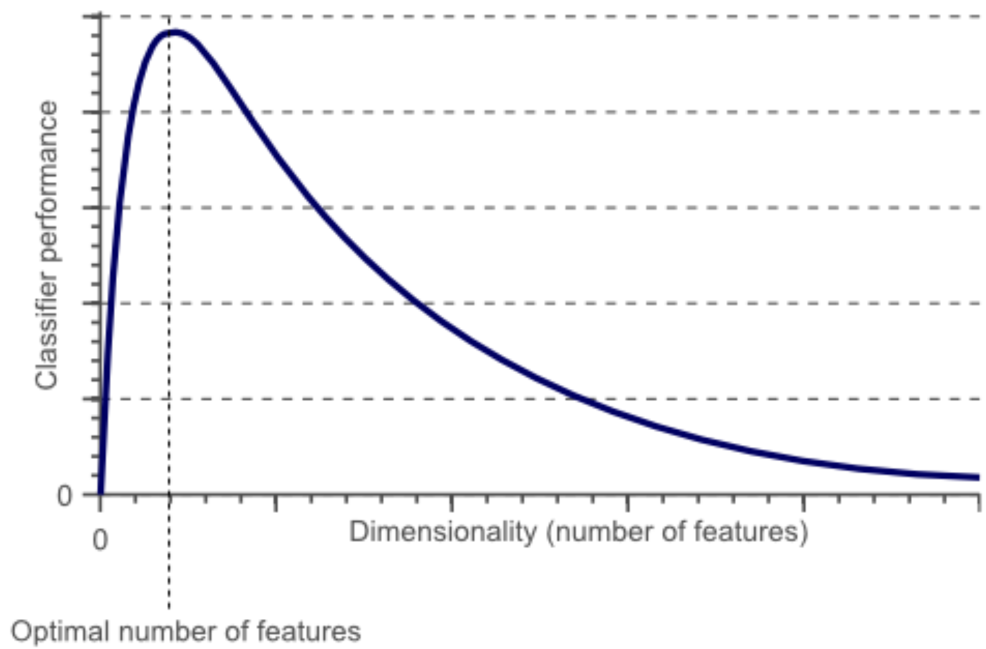
- **Avoid the curse of dimensionality**
  - Curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces that do not occur in low-dimensional space



  - As the dimensionality increases, we need more data to fill the space (filling the space is required to model data patterns both in supervised and unsupervised learning).

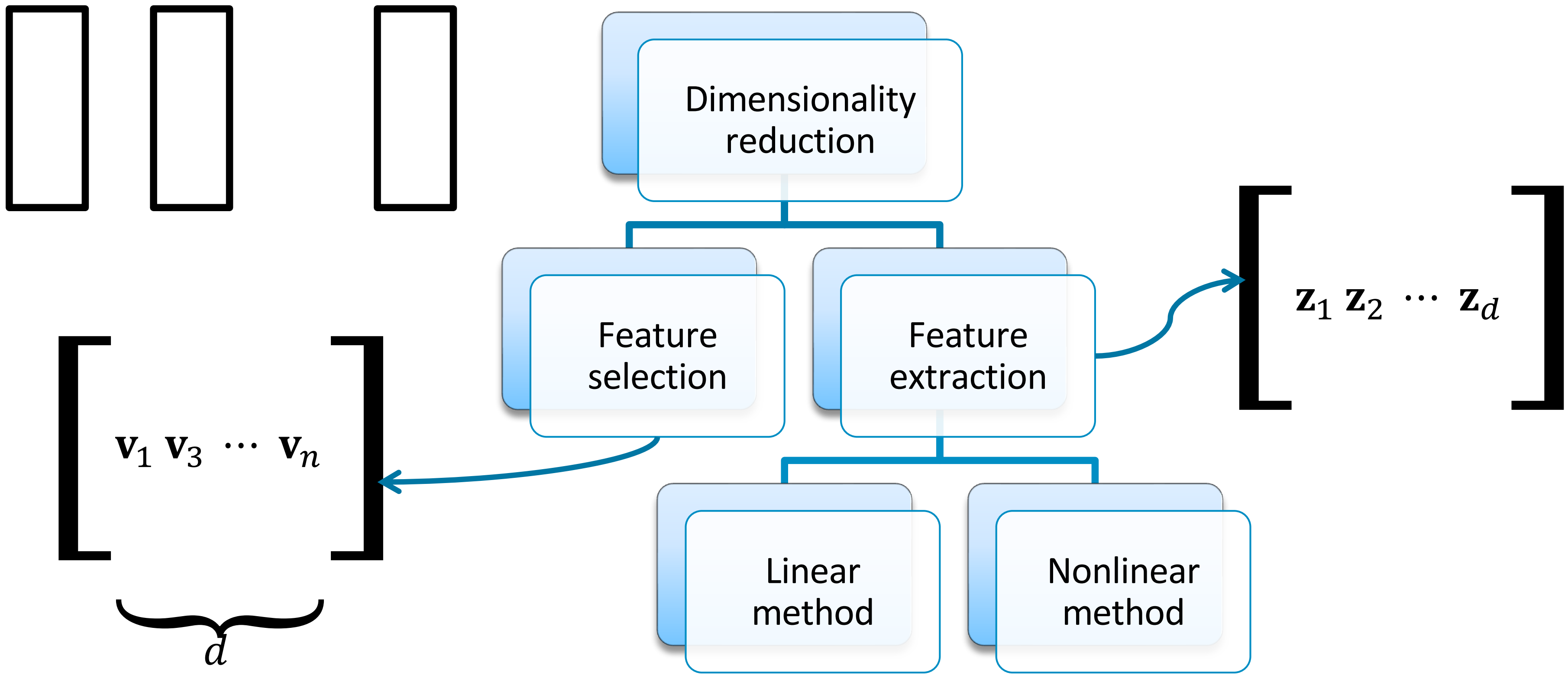
Hughes Phenomenon

- **With a fixed number of training samples, the predictive power of a classifier or regressor first increases as number of dimensions/features used is increased but then decreases**



Getting Rid of the Unnecessary

- **Dimensionality reduction**
  - The process of reducing the number of variables
- **Hierarchy of dimensionality reduction**



Principal Component Analysis

Principal Component Analysis (PCA)

- Which feature is the most helpful to distinguish one house from another?

ID	Value	Area	Floors	Household
1	148	72	4	20
2	156	76	4	22
3	160	86	4	22
4	165	79	4	24
5	169	88	5	30
6	184	90	5	35

- It's a good thing to have features with high variance, since they will be more informative and more important

→ Maximize variance

- It's a bad thing to have highly correlated features, or high covariance, since they can be deduced from one another with little loss in information, and thus keeping them together is redundant

→ Obtain orthogonal features

Principal Component Analysis (PCA)

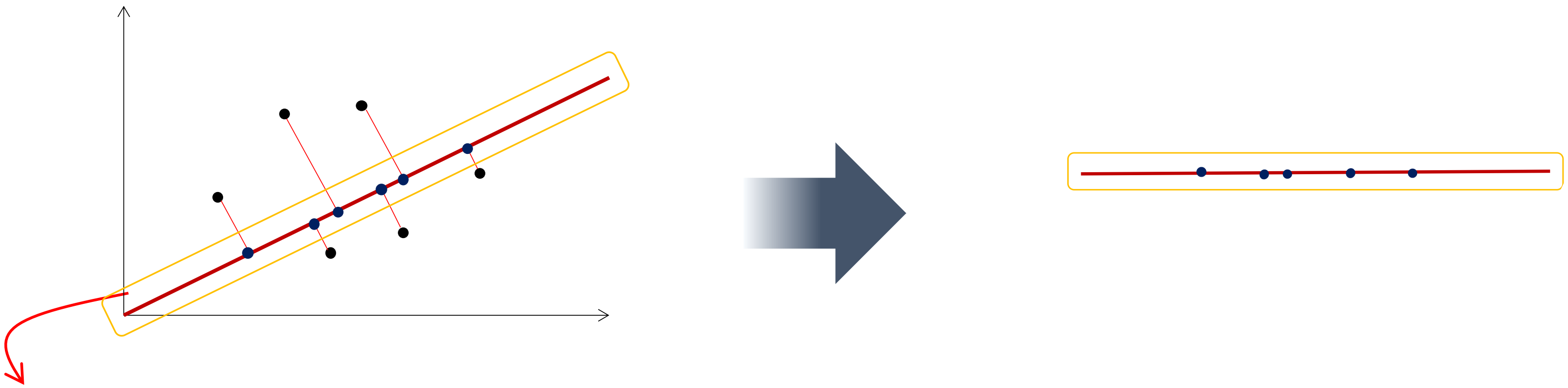
- **PCA is the orthogonal transformation to possibly correlated variables into a set of values into linearly uncorrelated variables**
  - Uncorrelated variables are called principal components

- **The number of principal components is less than or equal to the number of original variables**
  - Even though the number of principal components are equal to the number of original variables, we can select smaller number of components among them and then use for analysis

- **PCA can be understood as finding new axes that preserve variance of original data as much as possible and projecting original data onto these new axes.**
  - If we use small number of axes than the number of variables (= the number of original axes), we can reduce data dimensionality.

Principal Component Analysis (PCA)

- **Criterion to find principal component is to achieve the highest variance**
  - Variance of projected data samples on principal component

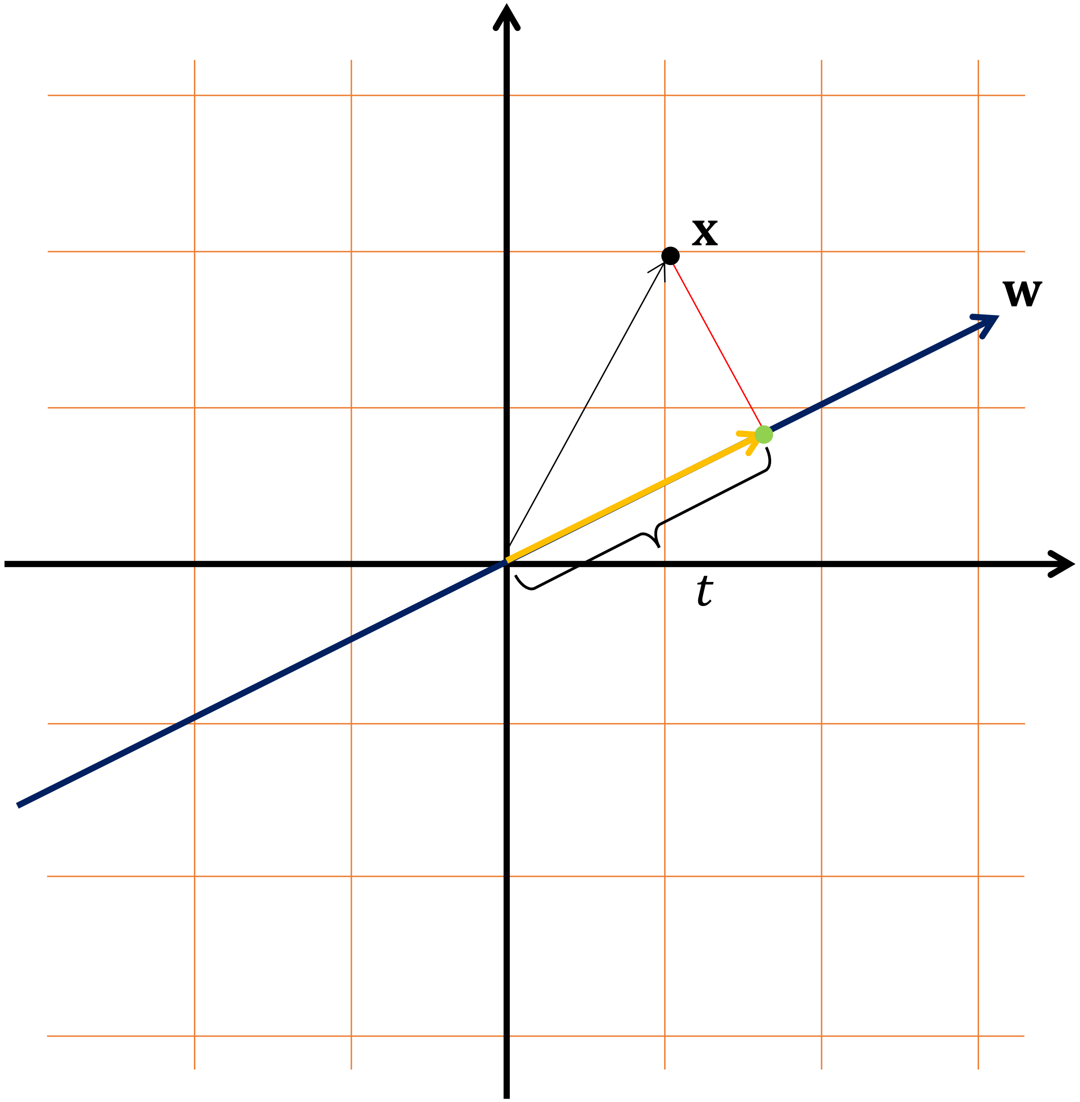


Find the line with the highest variance along the line

1D projected data distributed along the new axis (the first PC)

- **First principal component has the highest possible variance**
- **Each succeeding component in turn has the highest variance possible and it should be orthogonal to the preceding components**

※ Projection on the Line



Projected point on the line of data point $\mathbf{x}$

$t = \mathbf{w} \cdot \mathbf{x}$

Direction of line is defined as $\mathbf{w}$ and $\mathbf{w}$ is unit length vector

$$\mathbf{w} = \left(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}}\right)$$
$$\mathbf{x} = (1,2)$$
$$t = \mathbf{w} \cdot \mathbf{x} = \mathbf{w}^T\mathbf{x} = \frac{2}{\sqrt{5}} + \frac{2}{\sqrt{5}} = \frac{4}{\sqrt{5}}$$

Principal Component Analysis (PCA)

- **Find first component, $\mathbf{w}_1$ for data set that each dimension has zero mean**

$$\mathbf{w}_1 = \underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \sum_i (t_{1i})^2 = \underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \sum_i (\mathbf{x}_i \cdot \mathbf{w})^2$$

  - t_{1i} is the score(projected point on the first component) of i-th data point

- **Define data matrix**

Dot product

$\mathbf{x}_1 \cdot \mathbf{w}$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1{}^T \\ \mathbf{x}_2{}^T \\ \vdots \\ \mathbf{x}_n{}^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \qquad \mathbf{X}\mathbf{w} = \begin{bmatrix} \mathbf{x}_1{}^T \\ \mathbf{x}_2{}^T \\ \vdots \\ \mathbf{x}_n{}^T \end{bmatrix} \mathbf{w} = \begin{bmatrix} \mathbf{x}_1{}^T\mathbf{w} \\ \mathbf{x}_2{}^T\mathbf{w} \\ \vdots \\ \mathbf{x}_n{}^T\mathbf{w} \end{bmatrix}$$

- **Rewrite $\mathbf{w}_1$**

$$\mathbf{w}_1 = \underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \|\mathbf{X}\mathbf{w}\|^2 = \underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}$$

Principal Component Analysis (PCA)

- **Since unit vector constraint**

$$\mathbf{w}_1 = \arg\max\left(\frac{\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}}{\mathbf{w}^T\mathbf{w}}\right)$$

  - The larger $\|\mathbf{w}\|$ is, the larger $\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}$ is
  - $\mathbf{w}^T\mathbf{w}$ is the penalty term on $\|\mathbf{w}\|$ ($\|\mathbf{w}\|$ = $\mathbf{w}^T\mathbf{w}$)

$$\mathbf{x} = (1,2)$$
$$\mathbf{w}_1 = (1,2), \mathbf{w}_2 = (2,4)$$
$$t_1 = \mathbf{w}_1 \cdot \mathbf{x} = {\mathbf{w}_1}^T\mathbf{x} = 1 + 4 = 5$$
$$t_2 = \mathbf{w}_2 \cdot \mathbf{x} = {\mathbf{w}_2}^T\mathbf{x} = 2 + 8 = 10$$
$$\therefore t_1^2 < t_2^2$$

- **Solution of optimization problem**
  - $\mathbf{w}_1$ =eigenvector of $\mathbf{X}^T\mathbf{X}$ with the largest eigenvalue
  - $\mathbf{X}^T\mathbf{X}$ is proportional to covariance matrix of data when mean of each dimension is zero

※ Covariance

- **Variance of a random variable X is the expected value of the squared deviation from the mean($\mu = \mathbb{E}[X]$)**

$$Var(X) = \mathbb{E}[(X-\mu)^2]$$

  - Sample variance is calculated by

$$Var(x) = \frac{\sum_{i=1}^{n}(x-\bar{x})^2}{(n-1)}$$

- **Covariance is a measure of how much two random variables change together**

$$Cov(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$$

  - Variance is the covariance of a random variable with itself

$$Var(X) = Cov(X,X)$$

  - Sample covariance is calculated by

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_j-\bar{y})}{n-1}$$

※ Covariance

- **Covariance matrix, $\mathbf{C}$**
  - Matrix whose elements correspond to possible covariance values between all the different dimensions

$$\mathbf{C} = \begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \cdots & cov(x_1, x_p) \\ cov(x_2, x_1) & cov(x_2, x_2) & \cdots & cov(x_2, x_p) \\ \vdots & \vdots & & \vdots \\ cov(x_p, x_1) & cov(x_p, x_2) & \cdots & cov(x_p, x_p) \end{bmatrix}$$

  - If mean of each dimension is zero in data matrix, $\mathbf{X}$,

$$\mathbf{C} \propto \mathbf{X}^T \mathbf{X}$$

※ Eigenvector and Eigenvalue

- **For some matrix $\mathbf{A}$, the vector $\mathbf{x}$ satisfying following relation is eigenvector of matrix $\mathbf{A}$**

$$\mathbf{Ax} = \lambda\mathbf{x}$$

  - λ is the eigenvalue of eigenvector $\mathbf{x}$
  - The number of eigenvectors depends on matrix

- **Example**

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}$$

  - For vector $\mathbf{x} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$

NOT eigenvector

$$\mathbf{Ax} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 11 \\ 5 \end{bmatrix}$$

  - For vector $\mathbf{y} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$

eigenvector

$$\mathbf{Ay} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4\begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

eigenvalue

※ Eigenvector and Eigenvalue

- **How to get eigenvector and eigenvalue?**

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}$$

- Eigenvector and eigenvalue should satisfy $\mathbf{Ax} = \lambda\mathbf{x}$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} - \lambda \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2-\lambda & 3 \\ 2 & 1-\lambda \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = 0$$

NOT eigenvector

- If there exist nontrivial solution(trivial solution=$\begin{bmatrix} x \\ y \end{bmatrix}$=$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$), determinant of $\begin{bmatrix} 2-\lambda & 3 \\ 2 & 1-\lambda \end{bmatrix}$ should be 0

$$(2-\lambda)(1-\lambda) - 6 = 0 \rightarrow \lambda^2 - 3\lambda + 4 = 0$$

eigenvector

$$\lambda = 4 \; or \; -1$$

- When $\lambda = 4, \mathbf{x} = [3 \quad 2]^T$
- When $\lambda = -1, \mathbf{x} = [1 \quad -1]^T$

eigenvalue

Principal Component Analysis (PCA)

- **Succeeding process**
  - Subtracting preceding components from $\mathbf{X}$ → Create new data matrix

$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{i=1}^{k-1} \mathbf{X}\,\mathbf{w}_i \mathbf{w}_i^T$$

  - Find the principal component that extracts the maximum variance from new data matrix

$$w_k = \underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \left\|\hat{\mathbf{X}}_k \mathbf{w}\right\|^2 = \arg\max\left(\frac{\mathbf{w}^T \hat{\mathbf{X}}_k{}^T \hat{\mathbf{X}}_k \mathbf{w}}{\mathbf{w}^T \mathbf{w}}\right)$$

✓ Solve above equation is the also same as calculated the remaining eigenvectors of $\mathbf{X}^T\mathbf{X}$

✓ $\mathbf{w}_2$ =eigenvector of $\mathbf{X}^T\mathbf{X}$ with the second largest eigenvalue

Principal Component Analysis (PCA)

- **Finally,**

$$\mathbf{T} = \mathbf{X}\mathbf{W}$$

  - $\mathbf{W}$ is p-by-p matrix whose columns are the eigenvectors of $\mathbf{X}^T\mathbf{X}$ and it is called loading matrix

$$\mathbf{W} = [\mathbf{w}_1 \quad \mathbf{w}_2 \quad \cdots \quad \mathbf{w}_p]$$

  - Principal component is linear combinations of original features and transformation by loading matrix is linear transformation

- **Dimensionality reduction by PCA**
  - Keeping only the first l principal components (where $p > l$)

$$\mathbf{W}_l = [\mathbf{w}_1 \quad \mathbf{w}_2 \quad \cdots \quad \mathbf{w}_l]$$

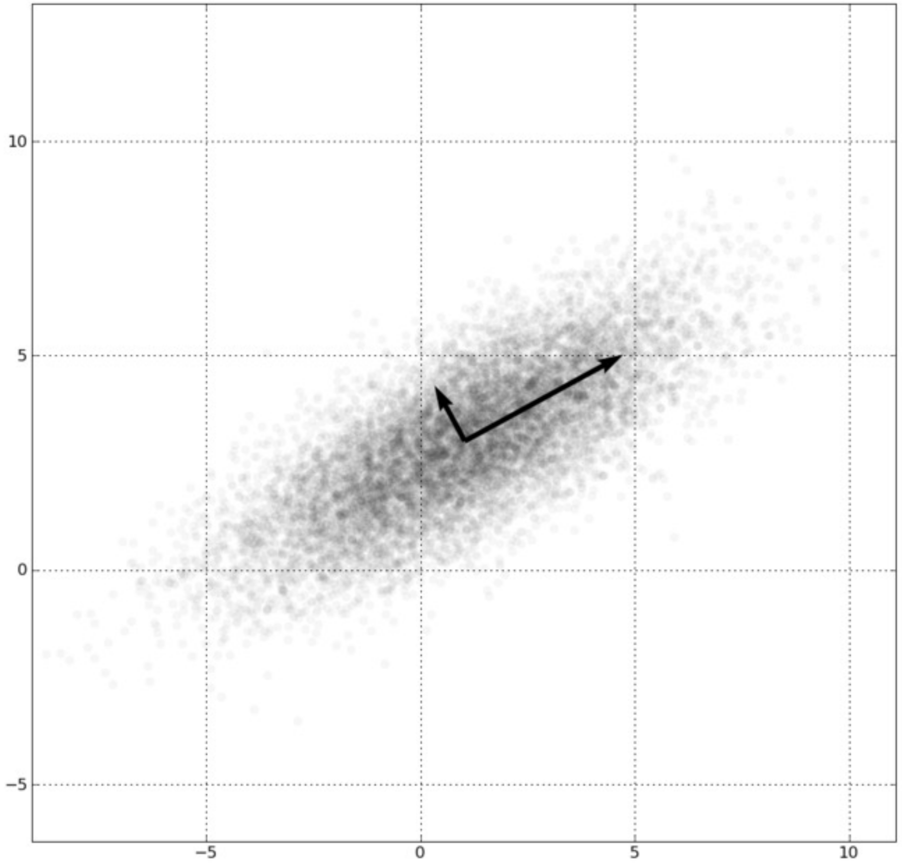
    - ✓ $\mathbf{W}_l$ is $p \times l$ matrix

- **Dimension-reduced data set is obtained by truncated transformation**

$$\mathbf{T}_l = \mathbf{X}\mathbf{W}_l$$

Principal Component Analysis (PCA)

- **2-dimensional data set and its principal components**



- **Web applet**
  - https://setosa.io/ev/principal-component-analysis/

※ Linear transformation

- **Linear transformation is a mapping $f\colon V \rightarrow W$ that preserves the operations of addition and scalar multiplication**

Addition: $f(\mathbf{x}+\mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$

Scalar multiplication: $f(\alpha\mathbf{x}) = \alpha f(\mathbf{x})$

- **Example**
  - Identity map $f(\mathbf{x}) = \mathbf{x}$ is linear transformation
    - ✓ $f(\mathbf{x}+\mathbf{y})$= $\mathbf{x}$+ $\mathbf{y}$=$f(\mathbf{x}) + f(\mathbf{y})$
    - ✓ $f(\alpha\mathbf{x})$=$\alpha\mathbf{x}$ = $\alpha f(\mathbf{x})$
  - map $f\colon x \rightarrow x^2$ is not linear transformation
    - ✓ $f(\mathbf{x}+\mathbf{y})$= $\mathbf{x^2} + 2\mathbf{x}\,\mathbf{y} + \mathbf{y}^2 \neq f(\mathbf{x}) + f(\mathbf{y})$
    - ✓ $f(\alpha\mathbf{x})$=$\alpha^2\mathbf{x}^2 \neq \alpha f(\mathbf{x})$
  - map $f\colon x \rightarrow x+1$ is not linear transformation
    - ✓ $f(\mathbf{x}+\mathbf{y}) = \mathbf{x}+\mathbf{y}+1 \neq f(\mathbf{x}) + f(\mathbf{y})$ $(f(\mathbf{x}) + f(\mathbf{y}) = \mathbf{x}+1+\mathbf{y}+1)$
    - ✓ $f(\alpha\mathbf{x}) = \alpha\mathbf{x} + 1 \neq \alpha f(\mathbf{x})$ $(\alpha f(\mathbf{x}) = \alpha\mathbf{x} + \alpha)$
  - If $\mathbf{A}$ is $m{\times}n$ matrix, then map $f\colon \mathbf{x} \rightarrow \mathbf{Ax}$ is linear transformation

※ Linear Combination

- **If $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$ are vectors and $a_1, a_2, \cdots, a_n$ are scalars, then linear combination of those vectors with those scalars as coefficient is**

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n$$

- **3-dimensional vector (a_1, a_2, a_3) is linear combination of $e_1 = (1,0,0), e_2 = (0,1,0), e_3 = (0,0,1)$**

$$\begin{aligned}(a_1, a_2, a_3) &= (a_1, 0,0) + (0, a_2, 0) + (0,0, a_3) \\ &= a_1(1,0,0) + a_2(0,1,0) + a_3(0,0,1) = a_1e_1 + a_2e_2 + a_3e_3\end{aligned}$$

- **Principal component $\mathbf{w} = (w_1, w_2, \ldots, w_p)$ is linear combination of unit vectors on each dimension representing by each variables**

Example: PCA

- **Find principal components of given data**

x	2.5	0.5	2.2	1.9	3.1	2.3	2.0	1.0	1.5	1.1
y	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

- Step1) substract from of the data dimensions for each dimension to have zero mean
  - ✓ $\bar{x} = 1.81, \bar{y} = 1.91$

x	0.69	-1.31	0.39	0.09	1.29	0.49	0.19	-0.81	-0.31	-0.71
y	0.49	-1.21	0.99	0.29	1.09	0.79	-0.31	-0.81	-0.31	-1.01

- Step 2) Calculate covariance matrix of new data($\mathbf{X}^T\mathbf{X}$)

$$C = \begin{bmatrix} 0.617 & 0.615 \\ 0.615 & 0.717 \end{bmatrix}$$

Example: PCA

- **Find principal components of given data**

x	2.5	0.5	2.2	1.9	3.1	2.3	2.0	1.0	1.5	1.1
y	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

- Step 3) Calculate the eigenvectors and eigenvalues of the covariance matrix
  - ✓ The largest eigenvalue is 1.28 and corresponding eigenvector is

$$\mathbf{w}_1 = \begin{bmatrix} -0.678 \\ -0.735 \end{bmatrix}$$

  - ✓ The second largest eigenvalue is 0.049 and corresponding eigenvector is

$$\mathbf{w}_2 = \begin{bmatrix} -0.735 \\ 0.678 \end{bmatrix}$$

※ Check eigenvector is unit vector!

Example: PCA

- **Find principal components of given data**

x	2.5	0.5	2.2	1.9	3.1	2.3	2.0	1.0	1.5	1.1
y	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

- Step 4) Choosing components and forming a loading matrix
  - ✓ If you choose two principal components both

$$\mathbf{W} = \begin{bmatrix} -0.678 & -0.735 \\ -0.735 & 0.678 \end{bmatrix}$$

  - ✓ If you want to reduce dimensionality

$$\mathbf{W} = \begin{bmatrix} -0.678 \\ -0.735 \end{bmatrix}$$

- Step 5) Derive the new data set

$$\mathbf{T} = \mathbf{XW}$$

x'	-0.83	1.78	-0.99	-0.27	-1.68	-0.91	0.99	1.14	0.44	1.22
y'	-0.18	0.14	0.38	0.13	-0.21	0.18	-0.35	0.46	0.02	-0.16

Feature Scaling

- **PCA finds principal component to achieve the highest variance**
  - The variable with large scale is dominated on principal component
    ex) When distance measure is change from m to cm, variance increases
    10000(100^2) times

Length(m)	Length(cm)
1.5	150
1.7	170
2.3	230
3.3	330
2.7	270
1.9	190

Sample variance(m)=0.46
Sample variance(cm)=4586

- **Before apply PCA to data samples, standardization is applied**
  - Transform each dimension to have unit variance

Reconstruct to Original Space

- **Transformed data by PCA can be reconstructed to original space**
  - Recall the final transformation

$$\mathbf{T} = \mathbf{XW}$$

  - Old data can be written as

$$\mathbf{X} = \mathbf{TW}^{-1}$$

    ✓ If $\mathbf{W}$ consists of unit vectors which are orthogonal to each other, inverse matrix of $\mathbf{W}$ is the same as the transpose of $\mathbf{W}$, $\mathbf{W}^T$

$$\therefore \mathbf{X} = \mathbf{TW}^T$$

    ✓ If you substract mean of each dimension from original data

$$\mathbf{X} = \mathbf{TW}^T + \boldsymbol{\mu}$$

      – $\boldsymbol{\mu}$ is mean vector of $\mathbf{X}$ ($\boldsymbol{\mu} = \left[\bar{x}_1 \bar{x}_2 \ \cdots \bar{x}_p\right]^T$)

    ✓ Actually, if $\mathbf{W}$ is not square matrix(if you choose the smaller number of principal components than original dimension), $\mathbf{W}^{-1}$ does not exist. However, in this case, reconstruction is performed through $\mathbf{W}^T$

Application: PCA

- **Extract important features through PCA for face recognition**
  - For image recognition, simple way to represent each image is to vectorization

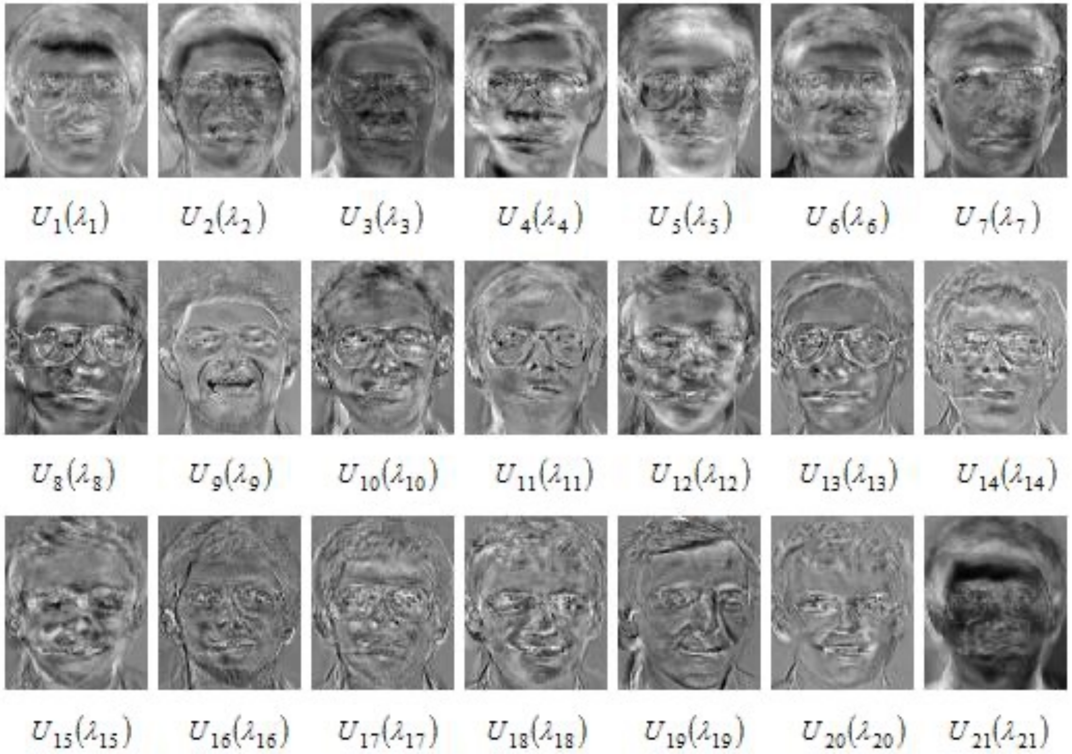
16×16 image

➡ Each pixel represents one dimension ➡ Transform to vector with 256 dimensions

  - Apply PCA to the set of image vectors and obtain principal components
    → transform image vector to lower dimensional space by loading matrix
    - ✓ Usually image is high-dimensional data
    - ✓ Through PCA, image can be compressed to low-dimensional data

Application: PCA

- **Eigenfaces**
  - A set of eigenvectors when they are used in the computer vision problem of human face recognition



✓ Eigenface can be viewed as a sort of map of the variations between faces

✓ PCA analysis has identified the statistical patterns in the data

Thank you!