Feature Dimension Reduction: PCA & LDA

Jin Young Choi
Seoul National University

Outline

- Feature Extraction
- Introduction of PCA & LDA
- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (FLDA)
- Simple Enhancement of PCA/LDA

Feature Extraction

- Features
  - Weight, Height, Width, Volume, Head size, …
  - Edge, Shape, Geometric Relations …
  - RGB Color for each pixel
  - SIFT, SURF, HOG, …

- Feature Extraction from Raw Data
  - Pixel Valued Vector is raw data vector
  - Raw data vector is redundant
  - The dimension should be reduced

Component Analysis and Discriminants

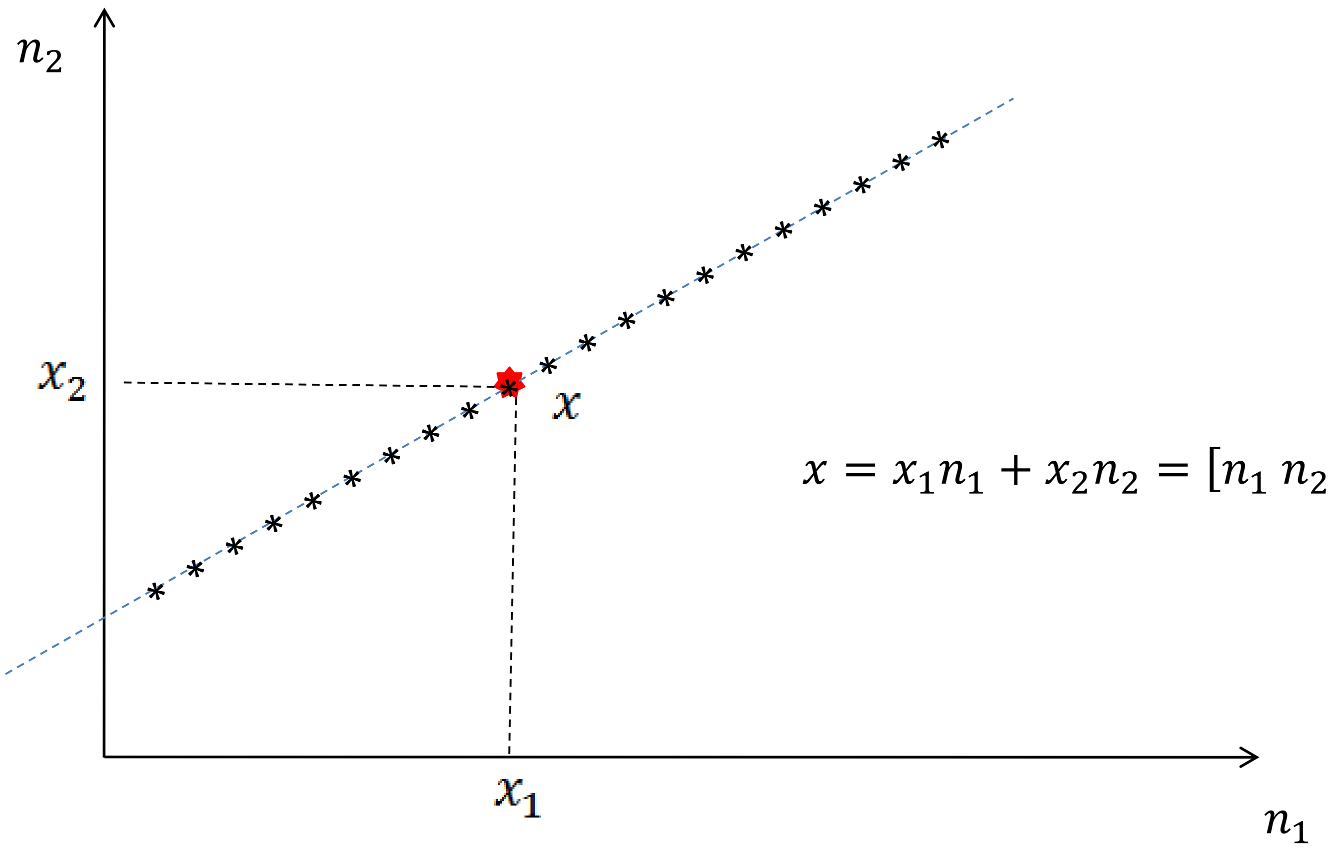
- How to reduce excessive dimensionality?
  - Answer: Combine features highly dependent to each other.

- Linear methods project high-dimensional data onto lower dimensional space.

- Principal Components Analysis (PCA)
  - seeks the projection which best represents the data in a least-square error sense.

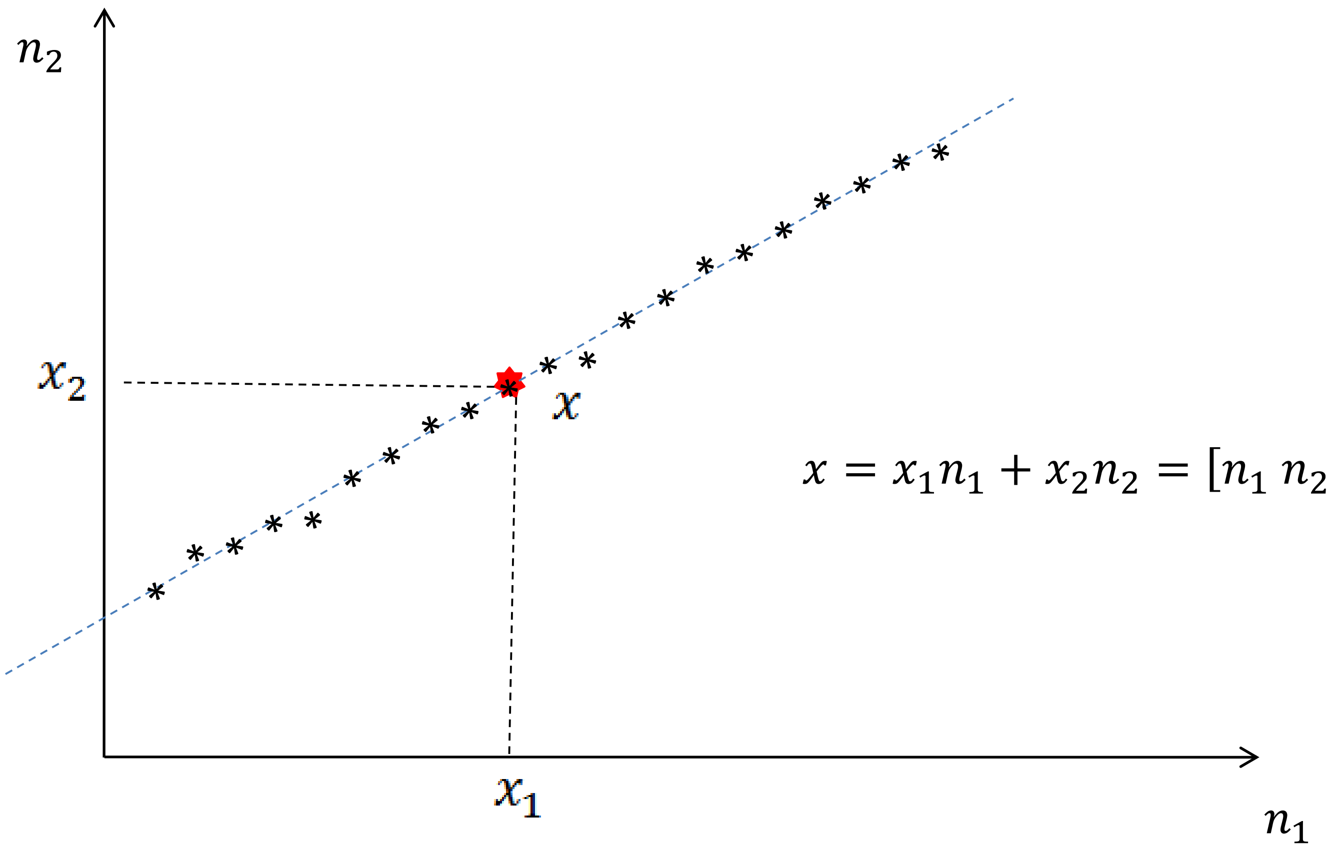
- Linear Discriminant Analysis (LDA) or Fisher Linear Discriminant
  - seeks the projection that best separates the data in a least-square discrimination error sense.

Principal Component Analysis



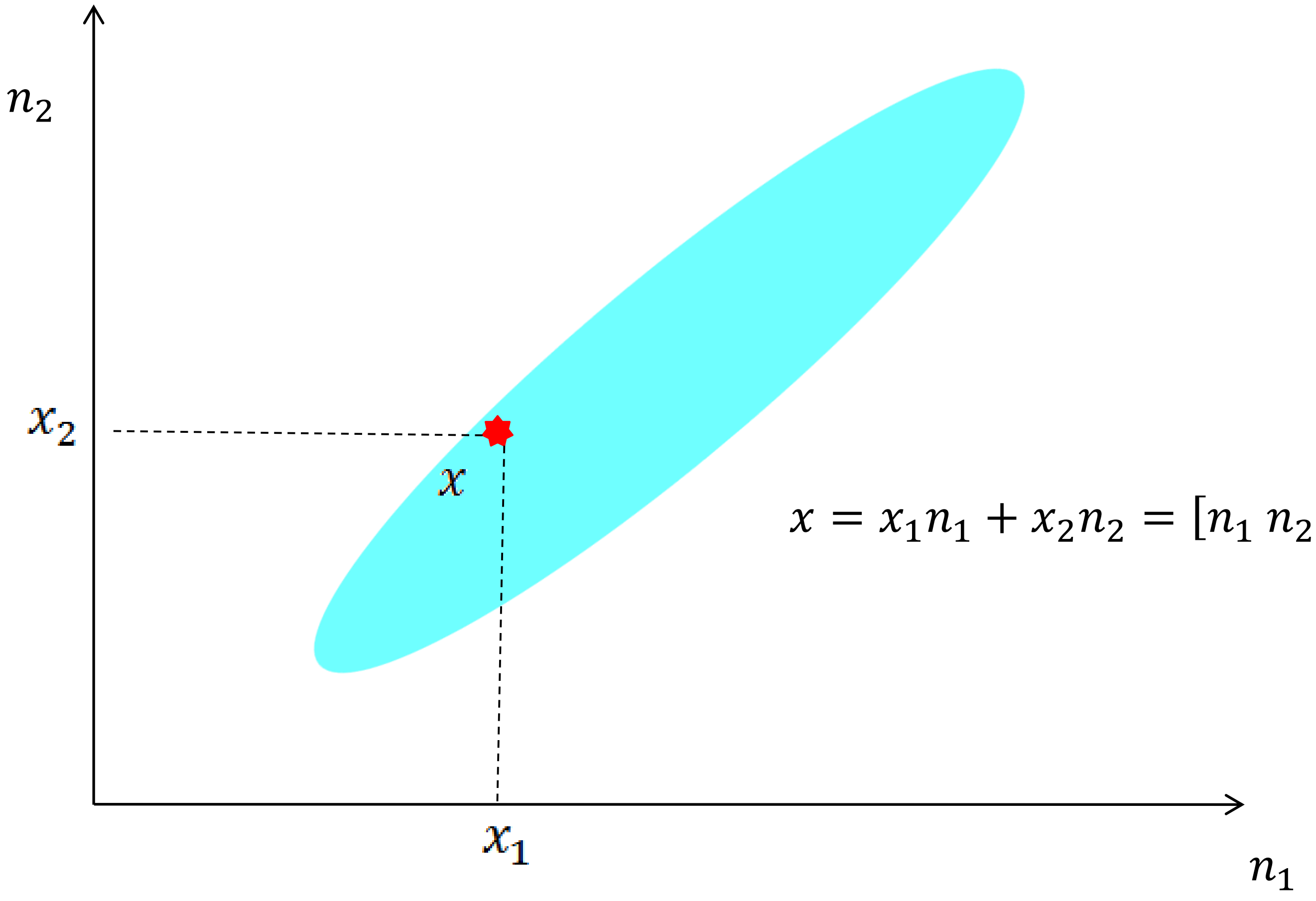
$$x = x_1 n_1 + x_2 n_2 = [n_1\ n_2]x = Nx$$

Principal Component Analysis



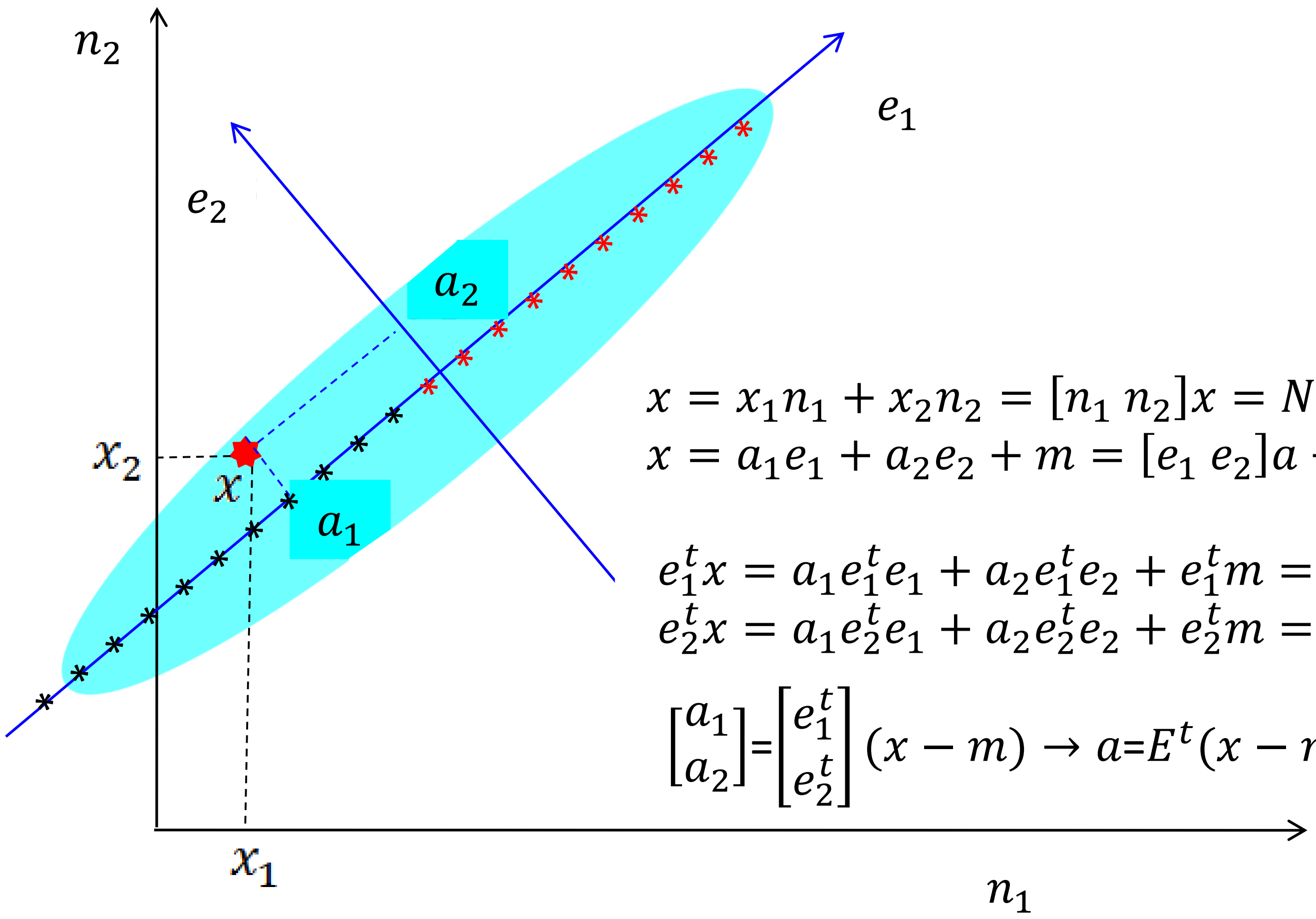
$$x = x_1 n_1 + x_2 n_2 = [n_1\ n_2]x = Nx$$

Principal Component Analysis



$$x = x_1 n_1 + x_2 n_2 = [n_1\ n_2]x = Nx$$

Principal Component Analysis



$x = x_1 n_1 + x_2 n_2 = [n_1\ n_2]x = Nx$

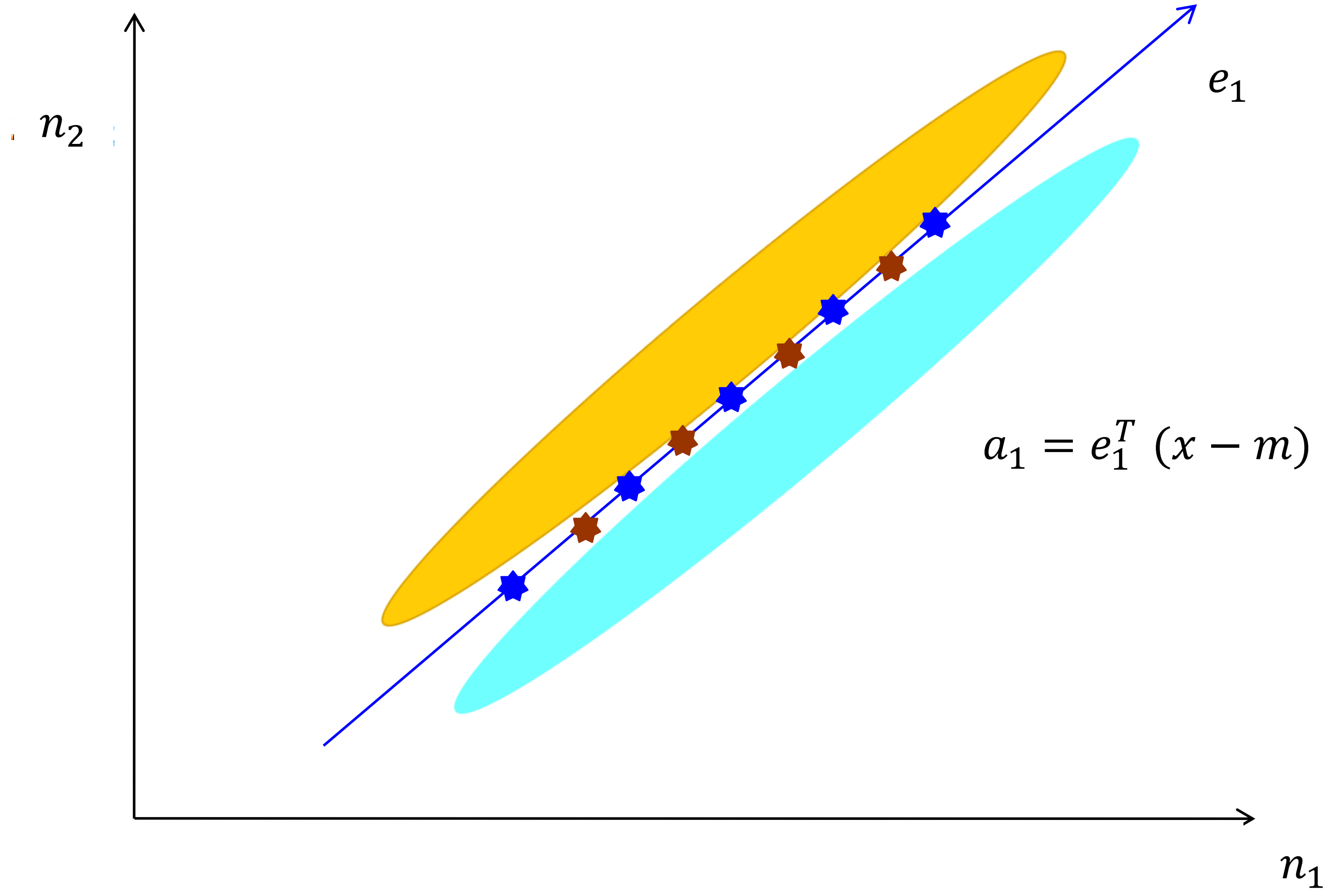
$x = a_1 e_1 + a_2 e_2 + m = [e_1\ e_2]a + m = Ea + m$

$e_1^t x = a_1 e_1^t e_1 + a_2 e_1^t e_2 + e_1^t m = a_1 + e_1^t m$

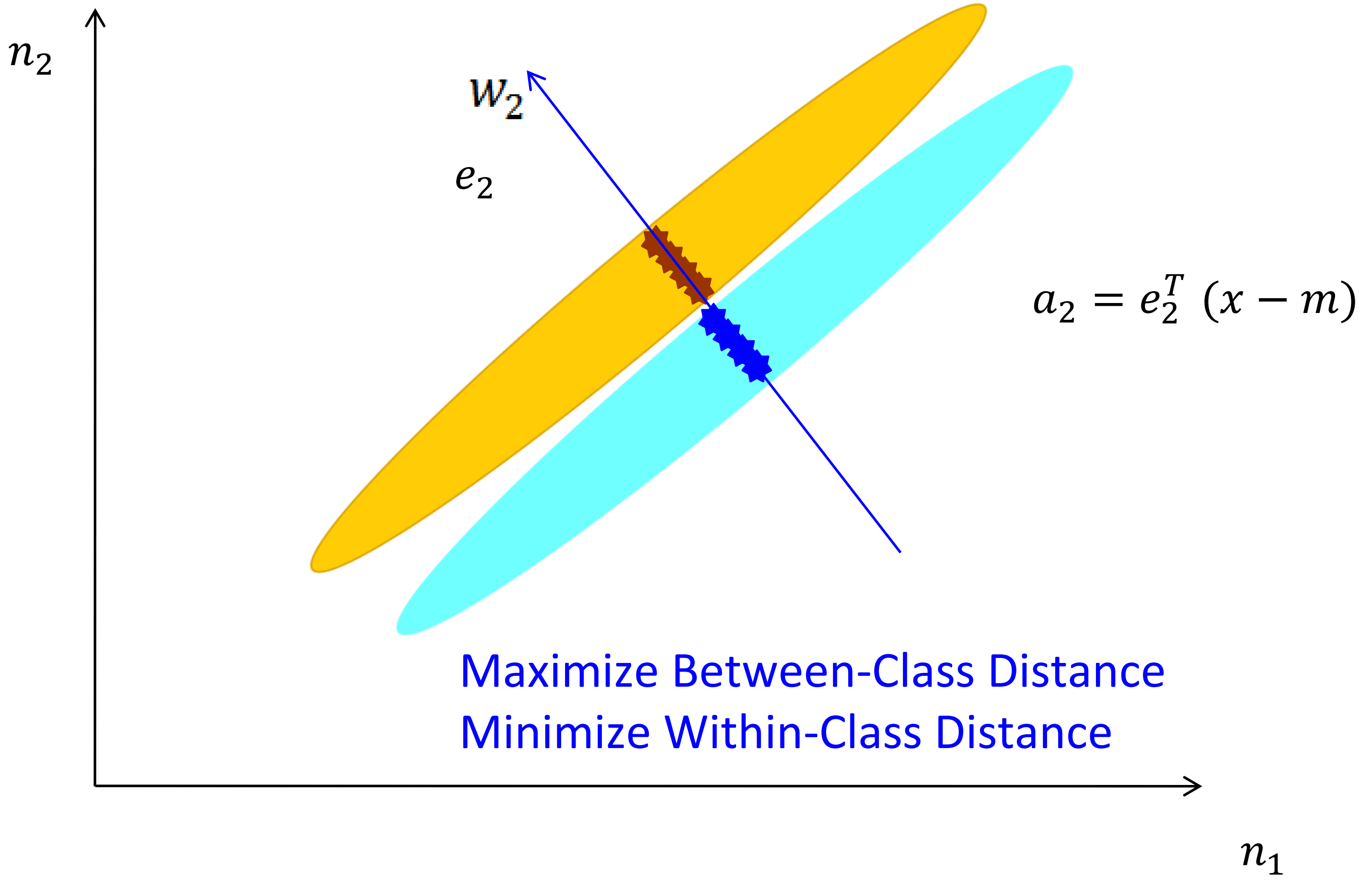
$e_2^t x = a_1 e_2^t e_1 + a_2 e_2^t e_2 + e_2^t m = a_2 + e_2^t m$

$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} e_1^t \\ e_2^t \end{bmatrix} (x - m) \rightarrow a = E^t (x - m)$

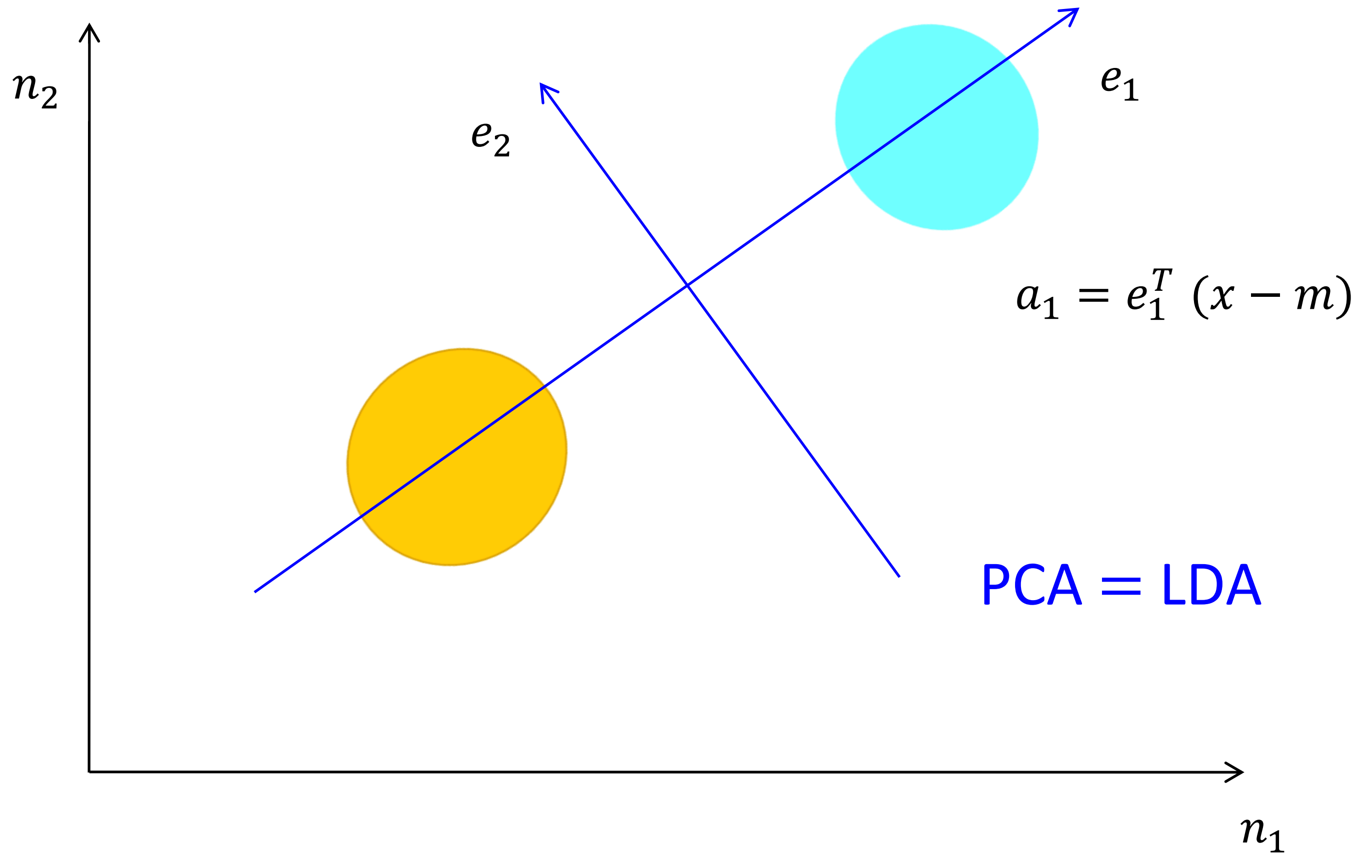
Linear Discriminant Analysis



Linear Discriminant Analysis



PCA & LDA



Principal Components Analysis (PCA)

- How to represent n d-dimensional vector samples $\{\boldsymbol{x}_1, .., \boldsymbol{x}_n\}$ by a single vector $\boldsymbol{x}_0$?
    - Find $\boldsymbol{x}_0$ that minimizes squared error correction function

$$J_0(\mathbf{x}_0) = \sum_{k=1}^{n} ||\mathbf{x}_0 - \mathbf{x}_k||^2.$$

Principal Components Analysis (PCA)

- How to represent n d-dimensional vector samples $\{\boldsymbol{x}_1, .., \boldsymbol{x}_n\}$ by a single vector $\boldsymbol{x}_0$?
  - Find $\boldsymbol{x}_0$ that minimizes squared error correction function

$$J_0(\mathbf{x}_0) = \sum_{k=1}^{n} ||\mathbf{x}_0 - \mathbf{x}_k||^2.$$

- The solution is sample mean

$$\mathbf{x}_0 = \mathrm{m} = \frac{1}{\mathrm{n}} \sum_{k=1}^{n} \mathbf{x}_\mathrm{k}$$

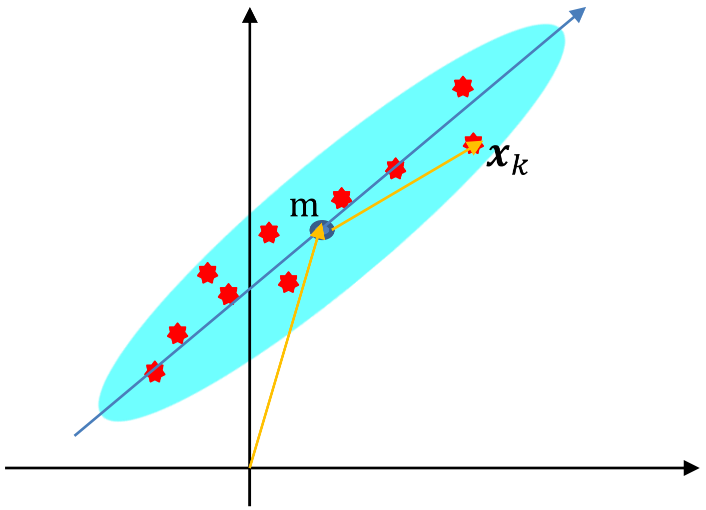
- This is zero-dimensional representation of the data set.
- One-dimensional representation by projecting the data onto a line through the sample mean reveals variability in the data.

Principal Components Analysis (PCA)

- This is zero-dimensional representation of the data set.

$$x_0 = m = \frac{1}{n}\sum_{k=1}^{n} x_k$$

- One-dimensional representation by projecting the data onto a line through the sample mean reveals variability in the data.



PCA ; Projection

- Let $\boldsymbol{e}$ be a unit vector in a direction of the line. The equation of the line

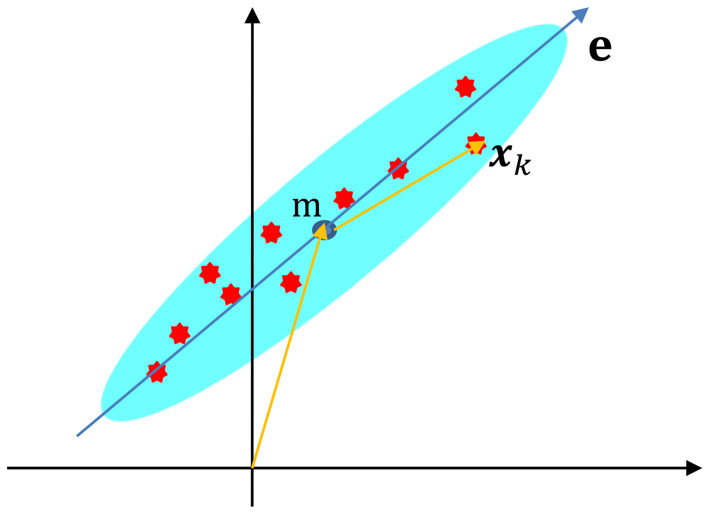
$$\mathbf{x} = \mathbf{m} + a\mathbf{e} \leftrightarrow \mathbf{e}^t(\mathbf{x} - \mathbf{m}) = a$$

- Representing $\boldsymbol{x}_k$ by $\mathbf{m} + a_k\mathbf{e}$, find "optimal" set minimizing criterion function :

$$J_1(a_1, \ldots, a_n, \mathbf{e}) = \sum_{k=1}^{n} ||\mathbf{m} + a_k\mathbf{e} - \mathbf{x}_k||^2.$$

from $\partial J_1 / \partial a_k = 0$

we find $a_k = \mathbf{e}^t(\mathbf{x}_k - \mathbf{m})$



PCA ; Projection

- Representing $\boldsymbol{x}_k$ by $\mathbf{m} + a_k\mathbf{e}$, find "optimal" a_k

  $$a_k = \mathbf{e}^T(\mathbf{x}_k - \mathbf{m})$$

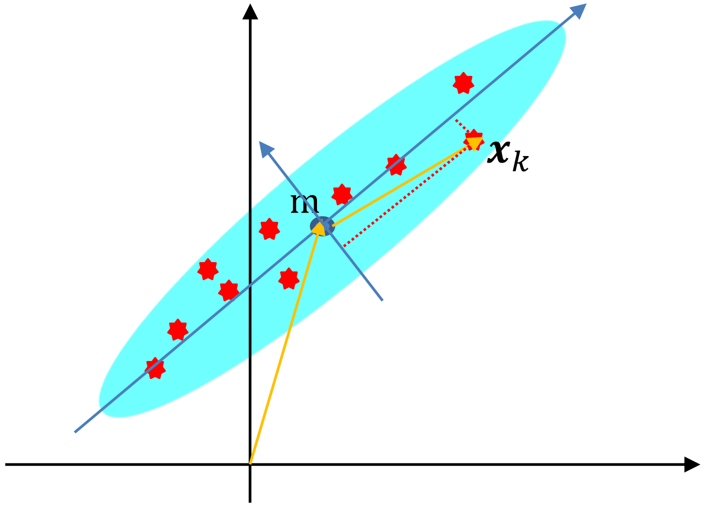
- How to find the *best* direction for ***e*** ?

- The least square solution: project the vector $\boldsymbol{x}_k$ onto the line in the direction of $\boldsymbol{e}$, passing through the sample mean.

  $$J_1(a_1, \ldots, a_n, \mathbf{e}) = \sum_{k=1}^{n} ||\mathbf{m} + a_k\mathbf{e} - \mathbf{x}_k||^2.$$

  $$a_k = \mathbf{e}^t(\mathbf{x}_k - \mathbf{m})$$

- Minimize J w.r.t ***e*.**



PCA ; Scatter matrix

- Substituting a_k into $J_1(a, \mathbf{e})$ we find

$$J_1(a, \mathbf{e}) = \sum_{k=1}^{n} a_k{}^2 ||\mathbf{e}||^2 - 2\sum_{k=1}^{n} a_k \mathbf{e}^T(\mathbf{x}_k - \mathbf{m}) + \sum_{k=1}^{n} ||\,\mathbf{x}_k - \mathbf{m}||^2$$

$$= \sum_{k=1}^{n} a_k{}^2 - 2\sum_{k=1}^{n} a_k{}^2 + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2$$

$$= -\sum_{k=1}^{n}[\mathbf{e}^T(\mathbf{x}_k - \mathbf{m})]^2 + \sum_{k=1}^{n} ||\,\mathbf{x}_k - \mathbf{m}||^2$$

$$= -\sum_{k=1}^{n} \mathbf{e}^T(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^T\mathbf{e} + \sum_{k=1}^{n} ||\,\mathbf{x}_k - \mathbf{m}||^2$$

$$= -\mathbf{e}^T\mathbf{S}\mathbf{e} + \sum_{k=1}^{n} ||\,\mathbf{x}_k - \mathbf{m}||^2$$

where a *scatter matrix* $\boldsymbol{S}$ which is $(n-1)$ times of sample covariance matrix given as

$$\mathbf{S} = \sum_{k=1}^{n}(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^T.$$

PCA ; Scatter matrix

- $J_1(a, \mathbf{e}) = -\mathbf{e}^T\mathbf{S}\mathbf{e} + \sum_{k=1}^{n} ||\mathbf{x}_k - \mathbf{m}||^2$
- Vector $\mathbf{e}$ that minimizes J_1 also maximizes $\mathbf{e}^T\mathbf{S}\mathbf{e}$.
- So we find $\mathbf{e}$, which maximize $\mathbf{e}^T\mathbf{S}\mathbf{e}$

  subject to constraint $||\mathbf{e}||=1$
- Let λ be Lagrange multiplier.
- Differentiating L with respect to $\mathbf{e}$: $L = \mathbf{e}^T\mathbf{S}\mathbf{e} - \lambda(\mathbf{e}^T\mathbf{e} - 1)$

$$\nabla_{\mathbf{e}} L(\mathbf{e}) = 2\mathbf{S}\mathbf{e} - 2\lambda\mathbf{e}$$

- By setting to zero we see that $\mathbf{e}$ is an eigenvector of $\mathbf{S}$:

$$\mathbf{S}\mathbf{e} = \lambda\mathbf{e} \rightarrow \mathbf{e}^T\mathbf{S}\mathbf{e} = \lambda$$

- So to maximize $\mathbf{e}^T\mathbf{S}\mathbf{e}$ takes maximal λ_i. $\mathbf{e}_i$ should be normalized to each other.

PCA ; Scatter matrix

- The result is easily extended to d' dimensional projection:

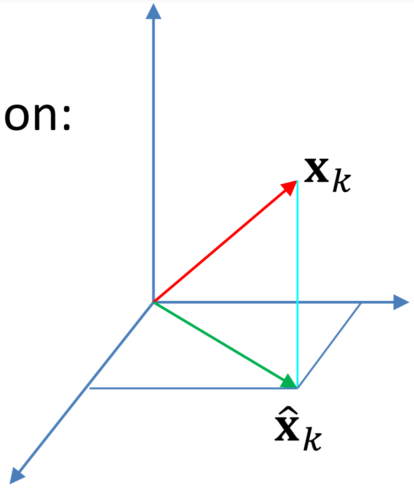
$$\hat{\mathbf{x}}_k = \mathbf{m} + \sum_{i=1}^{d'} a_k^i \mathbf{e}_i \quad \text{where} \quad d' \le d$$

- The criterion function

$$J_{d'} = \sum_{k=1}^{n} \left\| \left(\mathbf{m} + \sum_{i=1}^{d'} a_k^i \mathbf{e}_i \right) - \mathbf{x}_k \right\|^2$$

is minimized when $\mathbf{e_1}, \mathbf{e_2}, \dots, \mathbf{e_{d'}}$ are the eigenvectors having the largest eigenvalues.

- The coefficients $a_k^i = \mathbf{e}_i^T(\mathbf{x}_k - \mathbf{m})$ are *principal components.*



Error function

- If $d' < d$ error which is made by dropping the last terms is

$$J_{d'} = \sum_{k=1}^{n} \left\| \sum_{i=d'+1}^{d} a_k^i \mathbf{e}_i \right\|^2$$
$$= \sum_{i=d'+1}^{d} \mathbf{e}_i^T \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^T \mathbf{e}_i$$
$$= \sum_{i=d'+1}^{d} \mathbf{e}_i^T \mathbf{S} \mathbf{e}_i = \sum_{i=d'+1}^{d} \lambda_i$$

$$\mathbf{x}_k = \mathbf{m} + \sum_{i=1}^{d'} a_k^i \mathbf{e}_i$$

$$a_k^i = \mathbf{e}_i^T (\mathbf{x}_k - \mathbf{m})$$

- This is a sum of lowest eigenvalues.

PCA – the algorithm

- Input: $X^{(n)} = \{\mathbf{x}_1, .., \mathbf{x}_n\}, \quad \mathbf{x}_k = (x_1^k, \cdots, x_d^k)$
- Take $d' < d$
- Output: $A^{(n)} = \{\mathbf{a}_1, .., \mathbf{a}_n\} \;\; \mathbf{a}_k = (a_k^1, .., a_k^{d'})$
- Algorithm:
  - Compute the mean of the training set $\mathbf{m} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$.
  - Compute the scatter matrix $\mathbf{S}$.
  - Find eigenvectors of $\mathbf{S}$ and corresponding eigenvalues:

$$S\{\mathbf{e}_i, \lambda_i\}_{i=1}^{d} \;, \qquad \forall i: \mathbf{S}\mathbf{e}_i = \lambda \mathbf{e}_i \,, \lambda_1 \geq \lambda_2 \geq \ldots \lambda_d$$

  - Choose d' eigenvectors, and for each sample $\mathbf{x}_k$ point compute

$$a_k^i = \mathbf{e}_i{}^T(\mathbf{x}_k - \mathbf{m}), \qquad i = 1, \cdots, d'$$

Interim Summary

- Principal Component Analysis
  - ✓Feature Extraction
  - ✓Dimension Reduction

$$J_{d'} = \sum_{k=1}^{n} \left\| \left(\mathbf{m} + \sum_{i=1}^{d'} a_k^i \mathbf{e}_i \right) - \mathbf{x}_k \right\|^2$$

$$\mathbf{S} = \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^T.$$

$$\mathbf{S}\mathbf{e} = \lambda \mathbf{e}$$

$$\mathbf{e}^T \mathbf{S}\mathbf{e} = \lambda$$

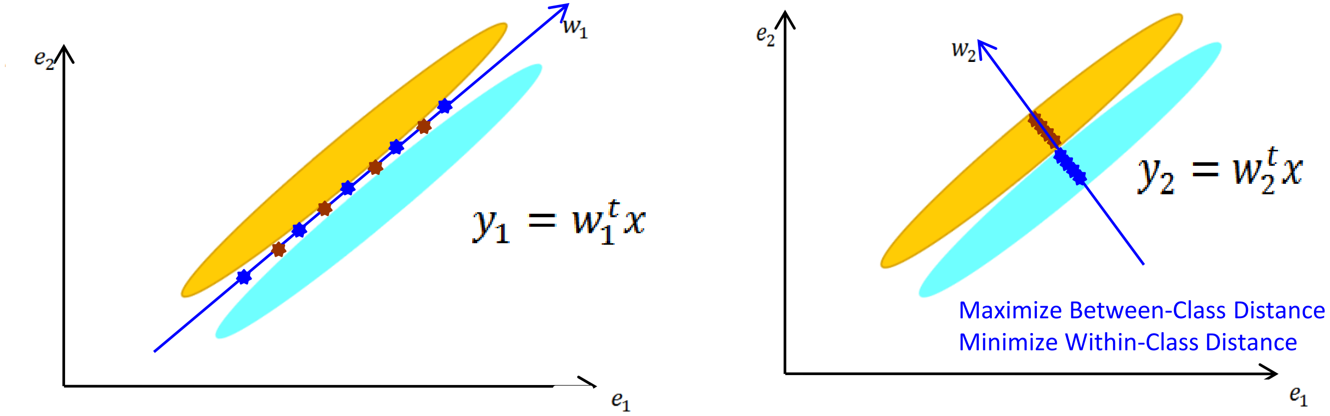
$$a_k^i = \mathbf{e}_i{}^T (\mathbf{x}_k - \mathbf{m}), i = 1, \cdots, d'$$

$$\begin{bmatrix} a_k^1 \\ a_k^2 \\ \vdots \\ a_k^{d'} \end{bmatrix} = \begin{bmatrix} e_1^T \\ e_2^T \\ \vdots \\ e_{d'}^T \end{bmatrix} (\mathbf{x}_k - \mathbf{m})$$

$$\mathbf{a}_k = E^T (\mathbf{x}_k - \mathbf{m})$$

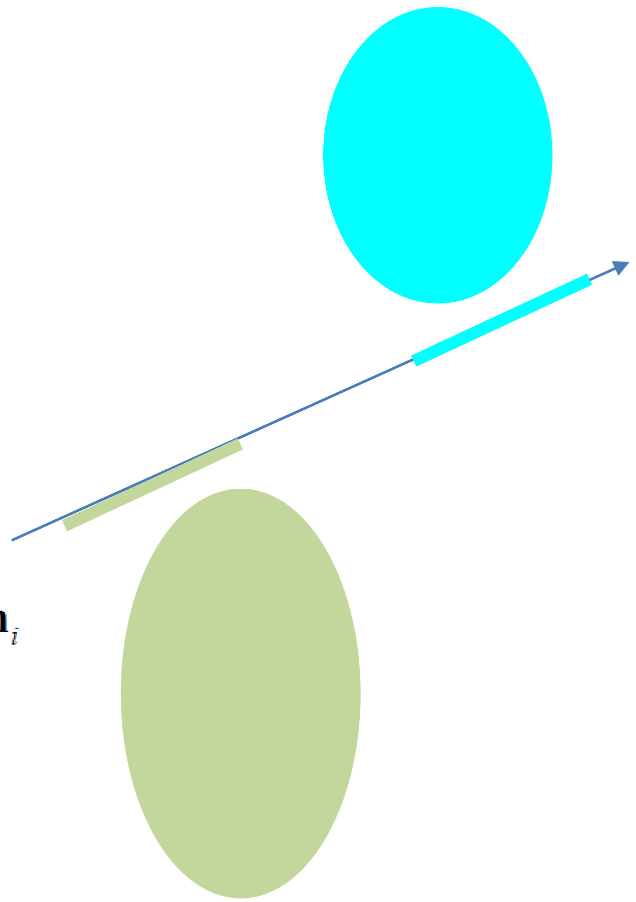
Linear Discriminant Analysis: LDA

- We have n d-dimensional samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, n_1 in a subset D_1, labeled w_1 and n_2 in a subset D_2 , labeled w_2.
- Find direction of line $\boldsymbol{w}$, that maximally separate the data.



- Let a difference between sample means be a measure of separation of projected points

Fisher Linear Discriminant



- Project samples $\mathbf{x}_k$ onto $\boldsymbol{w}$.

$$y_k = \mathbf{w}^t \mathbf{x}_k$$

- n samples y_k are divided into the subsets Y_1 and Y_2
- Let $\mathbf{m}_i$ be the sample mean

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$$

- The sample mean for projected points

$$\tilde{m}_i = \frac{1}{n_i} \sum_{\mathrm{y} \in \mathrm{Y}_i} \mathrm{y} = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{w}^t \mathbf{x} = \mathbf{w}^t \mathbf{m}_i$$

- Distance between the projected means is

$$|\tilde{m}_1 - \tilde{m}_2| = |\mathbf{w}^t (\mathbf{m}_1 - \mathbf{m}_2)|$$

Fisher Linear Discriminant

- A scatter for projected samples labeled ω_i

$$\tilde{s}_i^2 = \sum_{y \in Y_i} (y - \tilde{m}_i)^2$$

$(1/n)(\tilde{s}_1^2 + \tilde{s}_2^2)$ is an estimate of the variance of the pooled data.
$\tilde{s}_1^2 + \tilde{s}_2^2$ is called total within-class scatter of the projected samples.

- The Fisher discriminant employs $\mathbf{w}^t\mathbf{x}$ for which criterion

$$J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

is maximum

Fisher Linear Discriminant

- Define scatter matrices $\mathbf{S}_i$ and $\mathbf{S}_w$ by

$$S_i = \sum_{x \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$$

and

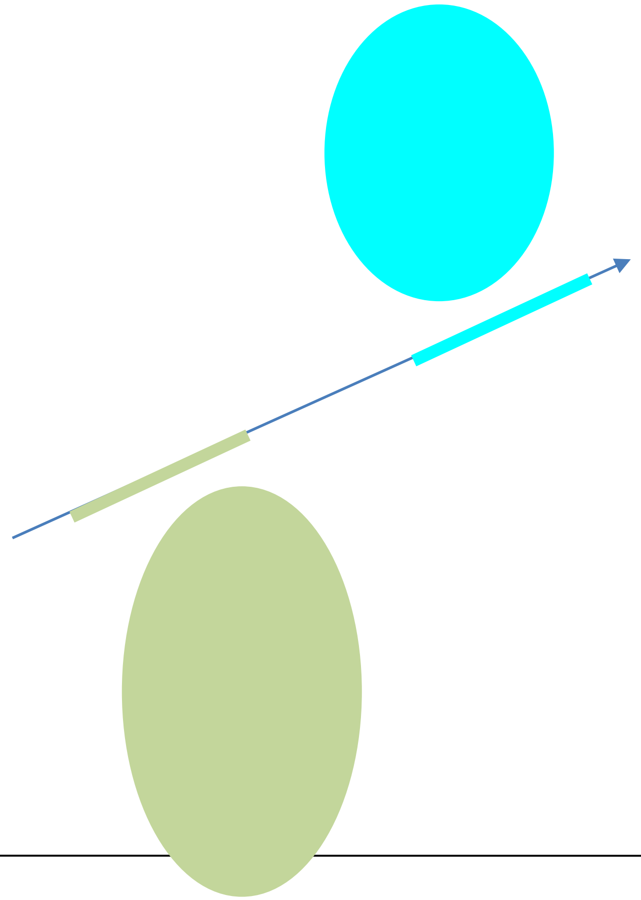
$$S_w = S_1 + S_2$$

- Then

$$\tilde{s}_i^2 = \sum_{\mathbf{x} \in D_i} (\mathbf{w}^t\mathbf{x} - \mathbf{w}^t m_i)^2 = \sum_{\mathbf{x} \in D_i} \mathbf{w}^t(\mathbf{x} - m_i)(\mathbf{x} - m_i)^t\mathbf{w} = \mathbf{w}^t\mathbf{S}_i\mathbf{w}$$

- Thus

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^t\mathbf{S}_w\mathbf{w}$$



Fisher Linear Discriminant

- Similarly,

$$(\tilde{m}_1 - \tilde{m}_2)^2 = (\mathbf{w}^t\mathbf{m}_1 - \mathbf{w}^t\mathbf{m}_2)^2 = \mathbf{w}^t(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{w} = \mathbf{w}^t\mathbf{S}_B\mathbf{w}$$

  $\mathbf{S}_w$ is called within-class scatter matrix (proportional to sample covariance matrix)

  $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$ is called between-class scatter matrix.

- This gives the equivalent expression for Fisher's discriminant

$$J(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}} = \lambda$$

- Which vector $\boldsymbol{w}$ maximizes it?

$$\nabla_{\mathbf{w}}J(\mathbf{w}) = \frac{2\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}} - \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}}\frac{2\mathbf{S}_W\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}} = 0 \rightarrow \mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w}$$

Fisher Linear Discriminant

- Hence one gets

$$\mathbf{S}_B\mathbf{w} = \lambda \mathbf{S}_W\mathbf{w}, \quad \lambda = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}},$$

  or equivalently

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \lambda\mathbf{w},$$

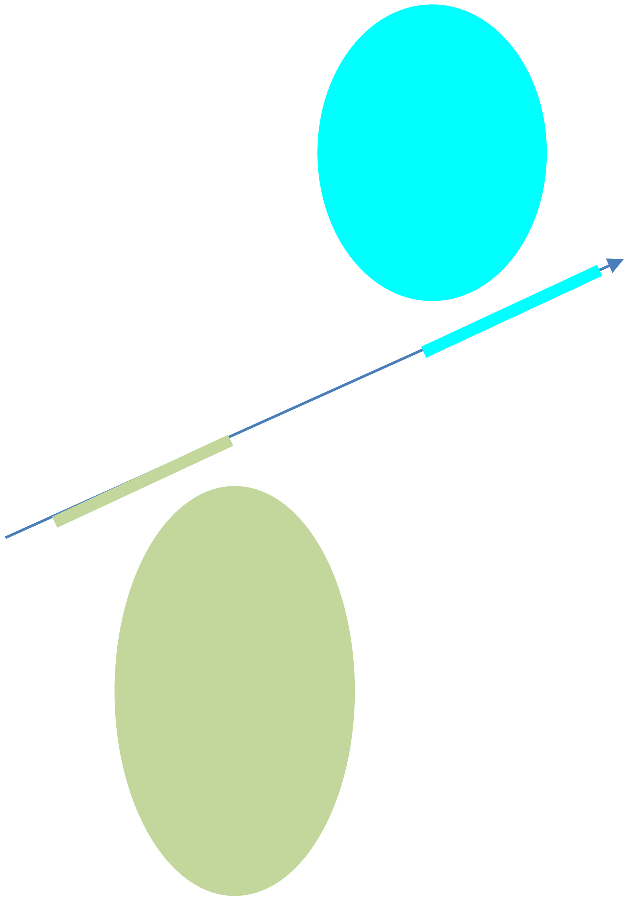
- Since for any $\boldsymbol{w}$, $\mathbf{S}_B\mathbf{w}$ is always in the direction of $\boldsymbol{m}_1 - \boldsymbol{m}_2$:

$$\mathbf{S}_B\mathbf{w} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{w} = \alpha(\mathbf{m}_1 - \mathbf{m}_2)$$

- It is not necessary to determine the eigenvalues of $\mathbf{S}_W^{-1}\mathbf{S}_B$.
- One simply gets

$$\mathbf{w} \propto S_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

- Scale factor for $\boldsymbol{w}$ is unimportant (why?).
- FLDA is one-dimensional projection



Summary

- Feature Extraction
- Introduction of PCA & LDA
- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (FLDA)
- Simple Enhancement of PCA/LDA

Regression Analysis I

Jin Young Choi
Seoul National University

Outline

- linear regression
  - simple linear regression
  - multiple linear regression
- nonlinear regression
  - logistic regression
  - high-order regression
  - basis-function regression
- matrix form for regression
  - recursive least squares
- partial least squares
  - over-fitting and underfitting
  - bias/variance
  - principle component regression
  - partial least squares algorithm
  - ridge regression
  - lasso, elastic regression
- Gaussian process regression
- Kalman filtering

LINEAR REGRESSION

JIN YOUNG CHOI

ECE, SEOUL NATIONAL UNIVERSITY

http://3.droppdf.com/files/pjxkI/regression-analysis-by-example-5th-edition.pdf
https://github.com/jwangjie/Gaussian-Processes-Regression-Tutorial

Regression Analysis

- For independent random variable X, and dependent random variable Y, assume they have a functional correlation between them, i.e.

$$Y = f(X)$$

- Regression: a process to find a parametric model $\hat{f}$ that gives the best fit of f for the observed samples

$$Y = \hat{f}(X) + \epsilon, \quad X\text{: predictor r.v., } Y\text{: response r.v.}$$

- Assume $E(\epsilon) = 0$, var$(\epsilon) = \sigma^2$, then $E(Y|x) = \hat{f}(x)$ for an observed non-random value x
- $\hat{f}$ can be estimated from the sample pairs $\{(y_i, x_i)|\ i = 1, 2, \cdots, n\}$

$$y_i = \hat{f}(x_i) + \epsilon_i,\ \ i = 1,\ \cdots,\ n,$$

where ϵ_i are i.i.d. zero mean and variance σ^2

Simple Linear Regression

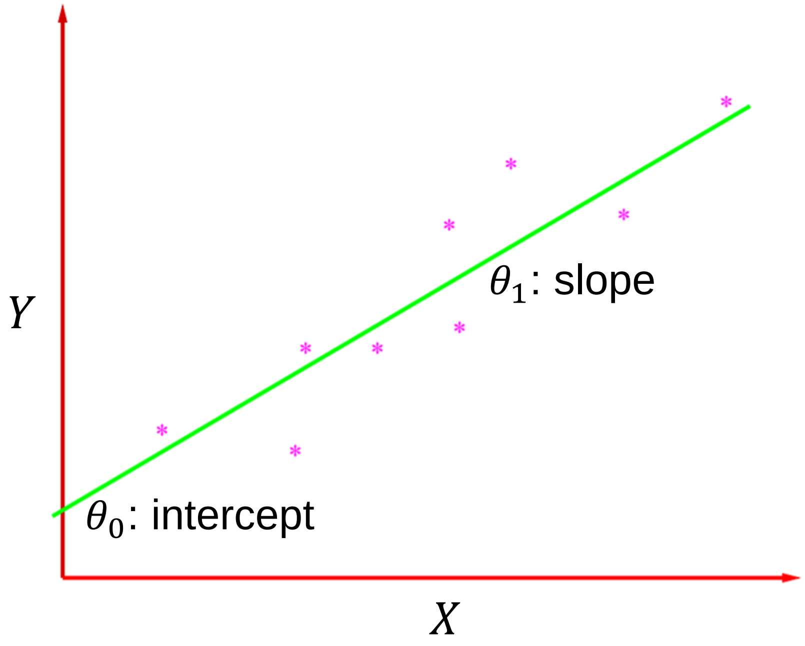
- Simple linear regression model

$$Y = \theta_0 + \theta_1 X + \epsilon$$

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i,\ i = 1,\ \cdots,\ n,$$

where θ_0: intercept, θ_1: slope

Observation Number	Response Y	Predictor X
1	y_1	x_1
2	y_2	x_2
3	y_3	x_3
⋮	⋮	⋮
n	y_n	x_n



Simple Linear Regression

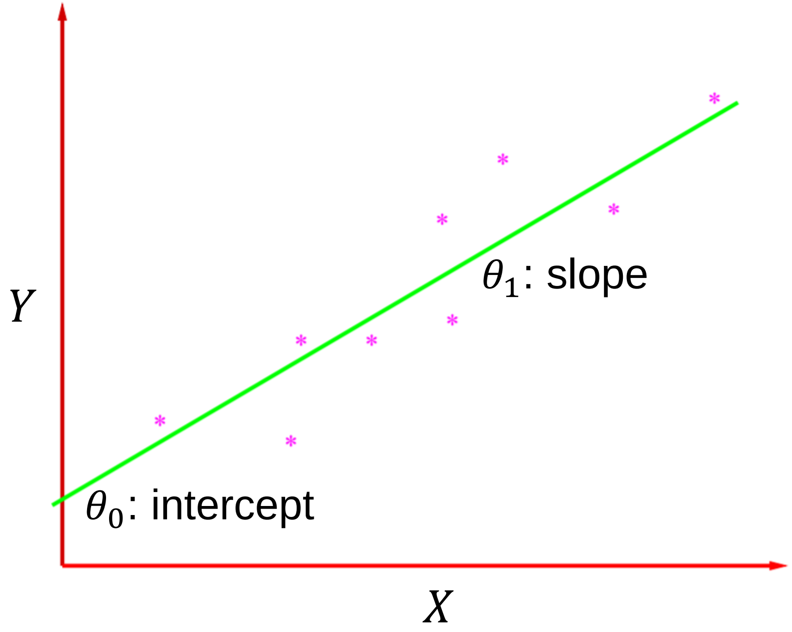
- Correlation of Y & X

$$Y = \theta_0 + \theta_1 X + \epsilon$$

$$\text{Cov}(Y, X) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})\ (x_i - \bar{x})$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$, $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Observation Number	Response Y	Predictor X
1	y_1	x_1
2	y_2	x_2
3	y_3	x_3
⋮	⋮	⋮
n	y_n	x_n



Simple Linear Regression

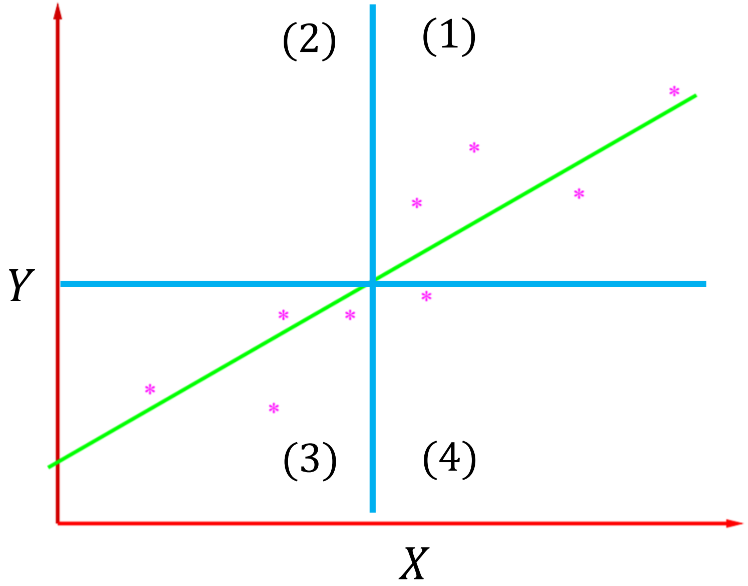
- Correlation of Y & X

$$Y = \theta_0 + \theta_1 X + \epsilon$$

$$\text{Cov}(Y,X) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})\,(x_i - \bar{x})$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Q	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})(x_i - \bar{x})$
(1)	+	+	+
(2)	+	−	−
(3)	−	−	+
(4)	−	+	−



$\theta_1 \geq 0$ → $\text{Cor}(Y,X) \geq 0$

$\theta_1 < 0$ → $\text{Cor}(Y,X) < 0$

Simple Linear Regression

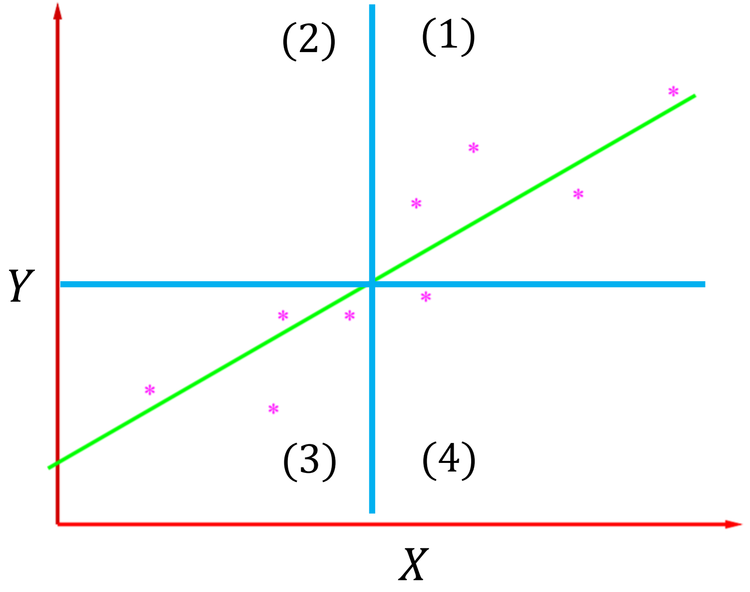
- Correlation Coefficient of Y & X

$$Y = \theta_0 + \theta_1 X + \epsilon$$

$$\rho(Y,X) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i-\bar{y}}{\sigma_y}\right)\left(\frac{x_i-\bar{x}}{\sigma_x}\right)$$

where $\sigma_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$, $\sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

Q	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})(x_i - \bar{x})$
(1)	+	+	+
(2)	+	−	−
(3)	−	−	+
(4)	−	+	−



$\theta_1 \geq 0 \longrightarrow 1 \geq \rho\,(Y,X) \geq 0$

$\theta_1 < 0 \longrightarrow -1 \leq \rho\,(Y,X) < 0$

Parameter Estimation

- Least Squares Estimation

Parameters are estimated by maximum likelihood estimation (MLE)

$$\epsilon_i = y_i - \theta_0 + \theta_1 x_i,\ i = 1,\ \cdots,\ n,\quad \epsilon_i \sim N(0, \sigma^2)$$

MLE:

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{{\epsilon_i}^2}{2\sigma^2})$$

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmax}} \ln \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{{\epsilon_i}^2}{2\sigma^2})$$

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmin}} \sum_{i=1}^{n} {\epsilon_i}^2 = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

LSE:

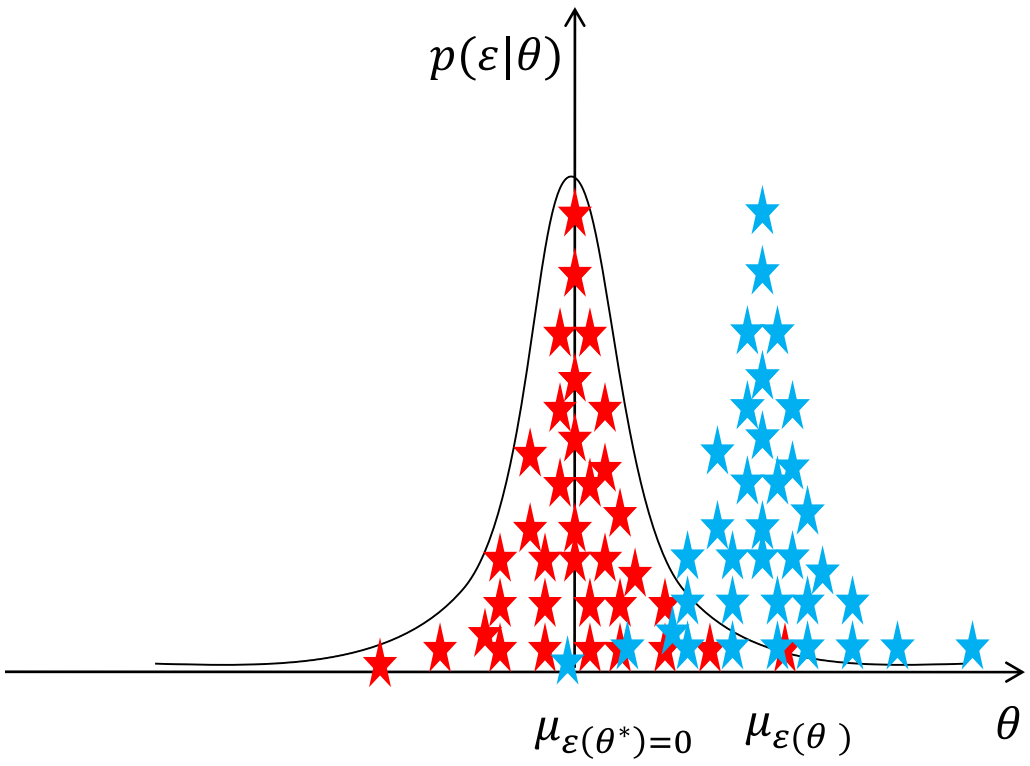
minimizing $S(\theta_0,\ \theta_1) = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$.

Solution:

by $\partial S/\partial\theta_0 = 0, \partial S/\partial\theta_1 = 0$ at $\hat{\theta}_0$ & $\hat{\theta}_1$,

Maximum Likelihood Estimation

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, p(\{\varepsilon_i\}|\theta) \qquad \varepsilon(\theta) = Y - \theta X$$



$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{{\epsilon_i}^2}{2\sigma^2})$$

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmax}} \ln \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{{\epsilon_i}^2}{2\sigma^2})$$

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmin}} \sum_{i=1}^{n} {\epsilon_i}^2 = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y} - \Phi\theta\|^2 \cong S(\theta)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \phi_1^T \\ \phi_2^T \\ \vdots \\ \phi_n^T \end{bmatrix} \theta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad \Phi_k = \begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1p} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{n1} & \phi_{n2} & \cdots & \phi_{np} \end{bmatrix}$$

$$y_i = \phi_i^T \theta + \epsilon_i$$

$$y_i = \theta_0 + \theta_1 \phi_{i1} + \theta_2 \phi_{i2} + \cdots + \theta_p \phi_{i(p-1)} + \epsilon_i,$$

$$i = 1, \cdots, n$$

Parameter Estimation

- Least Squares Estimation

  Parameters are estimated by maximum likelihood estimation (MLE)

$$\epsilon_i = y_i - \theta_0 + \theta_1 x_i,\ i = 1,\ \cdots,\ n,\quad \epsilon_i \sim N(0, \sigma^2)$$

MLE:

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{{\epsilon_i}^2}{2\sigma^2})$$

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmax}} \ln \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{{\epsilon_i}^2}{2\sigma^2})$$

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmin}} \sum_{i=1}^{n} {\epsilon_i}^2 = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

LSE:

minimizing $S(\theta_0,\ \theta_1) = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$.

Solution:

by $\partial S / \partial \theta_0 = 0, \partial S / \partial \theta_1 = 0$ at $\hat{\theta}_0$ & $\hat{\theta}_1$,

Parameter Estimation

- Least Squares Estimation

$$\epsilon_i = y_i - \theta_0 + \theta_1 x_i, \qquad i = 1, \qquad \cdots, \qquad n.$$

LSE:

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmin}}\ S(\theta_0,\ \theta_1) = \sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)^2 .$$

Solution:

by $\partial S/\partial\theta_0 = 0, \partial S/\partial\theta_1 = 0$ at $\hat{\theta}_0$ & $\hat{\theta}_1$,

$$\sum_{i=1}^{n}\left(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i\right) = 0, \quad \rightarrow \boxed{\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}}$$

$$\sum_{i=1}^{n}\left(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i\right)x_i = 0, \rightarrow \sum_{i=1}^{n}\left(y_i - \bar{y} - \hat{\theta}_1(x_i - \bar{x})\right)(x_i - \bar{x} + \bar{x}) = 0,$$

$$\rightarrow \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - \hat{\theta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 = 0 \rightarrow \boxed{\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

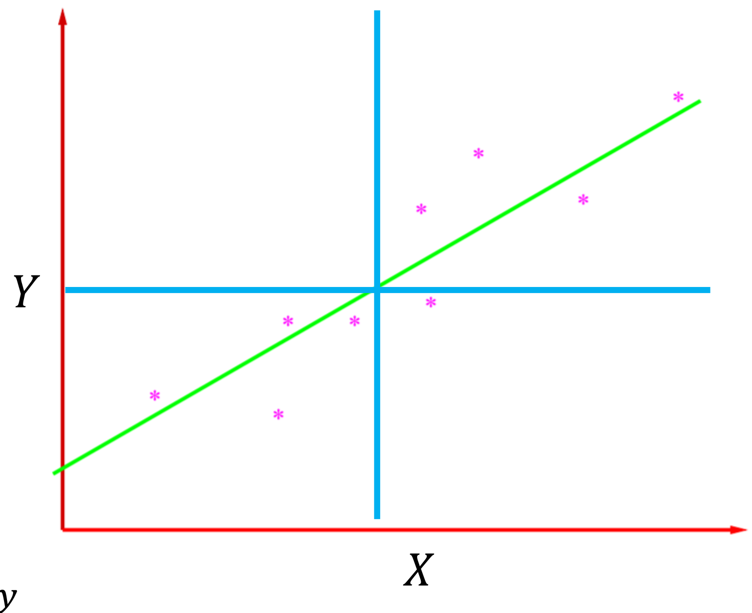
Parameter Estimation

- Least squares regression line

$$\hat{Y} = \hat{\theta}_0 + \hat{\theta}_1 X.$$

Fitted values:

$$\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i, \qquad i = 1, \cdots, n.$$



Error to the i-th observation:

$$e_i = y_i - \hat{y}_i, \; i = 1, \cdots, n.$$

Alternative formula for $\hat{\theta}_1$:

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(Y, X)}{Var(X)} = \rho(Y, X)\frac{\sigma_y}{\sigma_x}$$

→ slope has the same sign with the correlation (covariance)

Measuring the Quality of Fit

- Original Model:

$$Y = \theta_0 + \theta_1 X + \epsilon.$$

Least squares regression line:

$$\hat{Y} = \hat{\theta}_0 + \hat{\theta}_1 X.$$

- Correlation between Y & $\hat{Y}$:

$$\rho(Y,\hat{Y}) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}_i)}{sqrt\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}}_i)^2\right)}$$

Note that $Cor(Y,\hat{Y})$ can not be negative. Why?

Note that $Cor(Y,\hat{Y}) = 1$ implies the perfect fit.

Measuring the Quality of Fit

- Goodness-of-fit index:

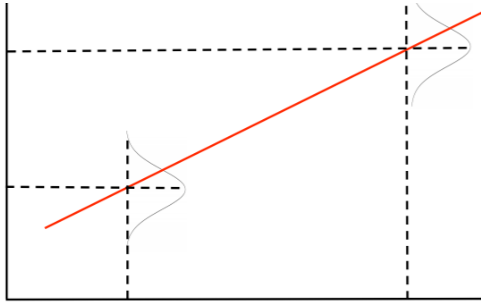
  SST: $\sum_{i=1}^{n}(y_i - \bar{y})^2$, SST: Total sum of squares

  SSR: $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, SSR: Regression (explained) sum of squares

  SSE: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, SSE: Residual (error) sum of squares

- Interpretation:

y_i	$=$	$\hat{y}_i$	$+$	$y_i - \hat{y}_i$
Observed	=	Fit	+	Error
$y_i - \bar{y}$	$=$	$\hat{y}_i - \bar{y}$	$+$	$y_i - \hat{y}_i$
Deviation		Deviation to Fit		Residual
SST	$\approx$	SSR	$+$	SSE



$SST \approx SSR + SSE \quad \because \sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) \approx 0$

- R^2: Coefficient of determination

  $$R^2 = \frac{SSR}{SST} \approx 1 - \frac{SSE}{SST}$$ (R = 1 implies the perfect fit)

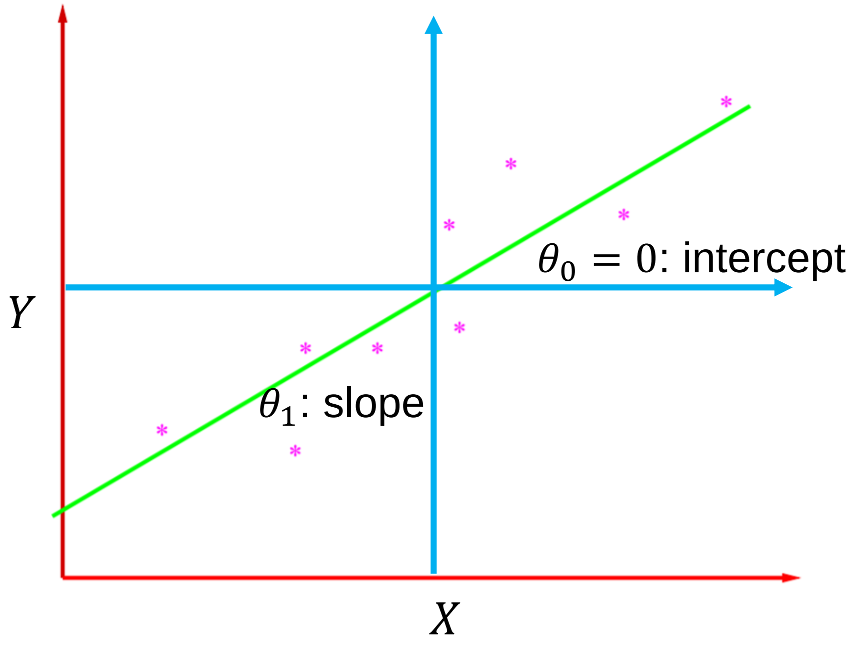
Regression Line through Origin

- Simple linear regression model

$$Y = \theta_0 + \theta_1 X + \epsilon$$

$$Y = \theta_1 X + \epsilon,$$ *no-intercept* model, $\bar{y} = \bar{x} = 0$

Observation Number	Response Y	Predictor X
1	$y_1 - \bar{y}$	$x_1 - \bar{x}$
2	$y_2 - \bar{y}$	$x_2 - \bar{x}$
3	$y_3 - \bar{y}$	$x_3 - \bar{x}$
⋮	⋮	⋮
n	$y_n - \bar{y}$	$x_n - \bar{x}$



Regression Line through Origin

- *no-intercept* model

$$y_i = \theta_1 x_i + \epsilon,$$
$$\hat{y}_i = \hat{\theta}_1 x_i,\ \ i = 1, \cdots, n$$
$$e_i = y_i - \hat{y}_i.$$

$$Cov(Y, X) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})\ (x_i - \bar{x}) \rightarrow Cov(Y, X) = \frac{1}{n-1}\sum_{i=1}^{n} y_i x_i$$

$$\rho(Y, X) = \frac{1}{n-1}\sum_{i=1}^{n}\frac{y_i x_i}{\sigma_y \sigma_x}\ ,\quad \sigma_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} {y_i}^2\ , \sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} {x_i}^2$$

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \rightarrow \hat{\theta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} {x_i}^2} = \frac{Cov(Y,X)}{\sigma_x^2} = \rho(Y, X)\frac{\sigma_y}{\sigma_x}$$

$$R^2 = \frac{\sum_{i=1}^{n} {\hat{y}_i}^2}{\sum_{i=1}^{n} {y_i}^2} = 1 - \frac{\sum_{i=1}^{n} {e_i}^2}{\sum_{i=1}^{n} {y_i}^2}$$

Multivariate Linear Regression

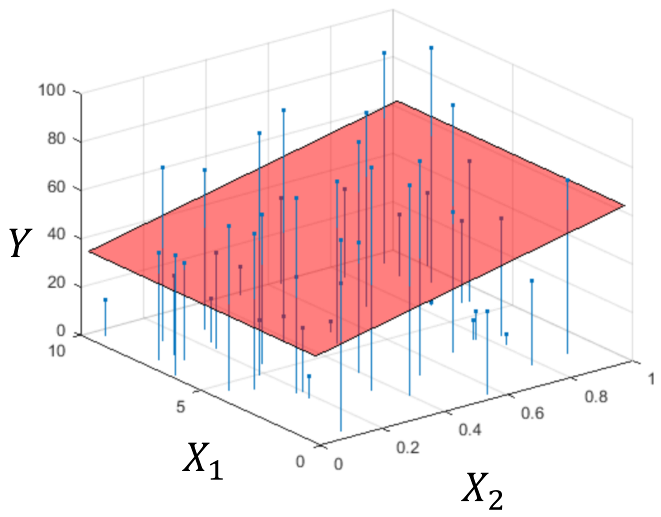
- Multivariate linear regression model: p predictor (explanatory) variables

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_p X_p + \epsilon$$

$$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_p x_{ip} + \epsilon_i,\ i = 1,\ \cdots,\ n,$$

where θ_0: intercept, $(\theta_1, \theta_2, \cdots, \theta_p)$: normal vector

i	Y	Predictor			
		X_1	X_2	$\cdots$	X_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
3	y_3	x_{31}	x_{32}	$\cdots$	x_{3p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}



Multivariate Linear Regression

- Multivariate linear regression model: p predictor (explanatory) variables

  $$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_p X_p + \epsilon$$

  $$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_p x_{ip} + \epsilon_i,\ i = 1,\ \cdots,\ n,$$

  where θ_0: intercept, $(\theta_1, \theta_2, \cdots, \theta_p)$: normal vector
- Fitted model by LSE: $n - p - 1$; degree of freedom (df)

  $$\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_{i1} + \hat{\theta}_2 x_{i2} + \cdots + \hat{\theta}_p x_{ip}, i = 1, \cdots, n,$$

  $$e_i = y_i - \hat{y}_i.$$
- Measuring Quality of Fit:

  $$\rho(Y, \hat{Y}) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2\right)}}$$

  $$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- Adjusted R^2 : $R_a^2 = 1 - \frac{1/(n-p-1) \sum_{i=1}^{n} e_i^2}{1/(n-1) \sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{n-1}{n-p-1}(1 - R^2)$

Multivariate Linear Regression

- Tests of Hypotheses for Multivariate linear model

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_p X_p + \epsilon$$

$$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_p x_{ip} + \epsilon_i,\ \ i = 1,\ \cdots,\ n,$$

where θ_0: intercept, $(\theta_1, \theta_2, \cdots, \theta_p)$: normal vector

- Hypotheses: H_0: Reduced model (RM), H_1: Full model (FM)
  1. All the regression coefficients associated with the predictor variables are zero.
  2. Some of the regression coefficients are zero.
  3. Some of the regression coefficients are equal to each other.
  4. The regression parameters satisfy certain specified constraints.
- Sum of Squares: $SSE(RM) \geq SSE(FM)$

$$SSE(FM) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SSE(RM) = \sum_{i=1}^{n} (y_i - \hat{y}_i^*)^2$$

- F-test: $F = \dfrac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$ (F is large → RM is inadequate[†])

[†] The critical values are given in Table A.4 and A.5 in “Regression Analysis by Example”, S. Chatterjee et.al., Wiley.

NONLINEAR REGRESSION

JIN YOUNG CHOI

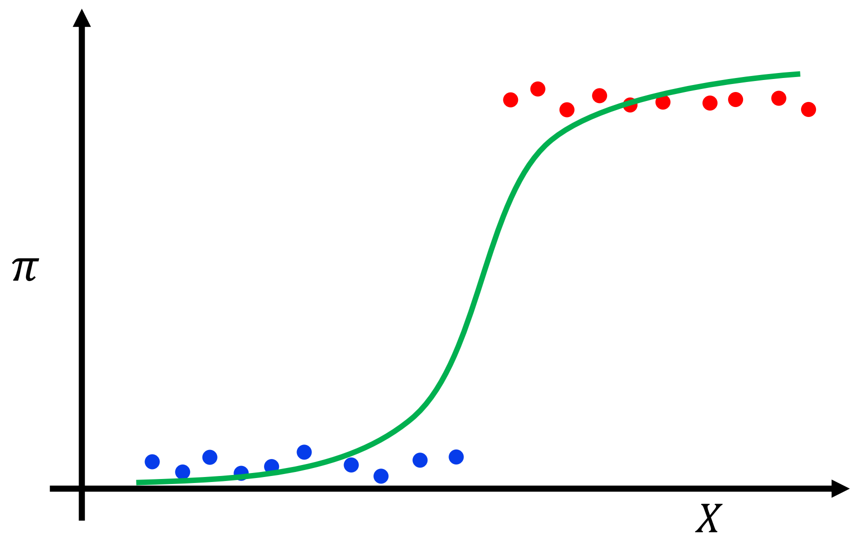
ECE, SEOUL NATIONAL UNIVERSITY

https://github.com/jwangjie/Gaussian-Processes-Regression-Tutorial

Logistic Regression

- Logistic response function representing the relation between the probability π and $X_1, X_2, \cdots, X_p$

$$\pi = p\left(Y = 1 \middle| X_1 = x_1, \cdots, X_p = x_p\right) = \frac{\exp(\theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p)}{1 + \exp(\theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p)}$$



Logistic Regression

- Logistic response function

$$\pi(X_1 = x_1, \cdots, X_p = x_p) = p\left(Y = 1 \middle| X_1 = x_1, \cdots, X_p = x_p\right) = \frac{\exp(\theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p)}{1 + \exp(\theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p)}$$

$$1 - \pi(X_1 = x_1, \cdots, X_p = x_p) = p\left(Y = 0 \middle| X_1 = x_1, \cdots, X_p = x_p\right) = \frac{1}{1 + \exp(\theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p)}$$

$$\frac{\pi}{1 - \pi} = \exp(\theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p)$$

$$f\left(X_1 = x_1, \cdots, X_p = x_p\right) = \ln \frac{\pi}{1 - \pi} = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p$$

$$f(X) = \ln \frac{\pi}{1 - \pi} = \theta_0 + \theta_1 X_1 + \cdots + \theta_p X_p$$

High-order Regression

- High-order polynomial regression model

$$Y = \theta_0 + \theta_1 X + \theta_2 X^2 + \cdots + \theta_m X^m + \epsilon$$

$$y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \cdots + \theta_m x_i^m + \epsilon_i, \; i = 1, \; \cdots, \; n.$$

- High-order multivariate regression model

$$Y = \theta_0 + \theta_1 X_1 + \cdots + \theta_p X_p + \cdots + \theta_{p+k} X_1 X_k + \cdots + \theta_M X_p^m + \epsilon$$

$$y_i = \theta_0 + \theta_1 x_{i1} + \cdots + \theta_p x_{ip} + \cdots + \theta_{p+k} x_{i1} x_{ik} + \cdots + \theta_M x_{ip}^m + \epsilon_i$$

- Matrix-vector form

Let $\theta = [\theta_0 \;\; \theta_1 \;\; \cdots \; \theta_M]^T, \quad \phi_i = [1 \;\; \phi_{i1} \cdots \phi_{iM}]^T$

$y = [y_1 \;\; y_2 \;\; \cdots \;\; y_n]^T, \quad \epsilon = [\epsilon_1 \;\; \epsilon_2 \;\; \cdots \;\; \epsilon_n]^T$

$$\Phi = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ 1 & x_3 & x_3^2 & \cdots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix}$$

Then $y_i = \phi_i^T \theta + \epsilon_i, \; i = 1, \; \cdots, \; n.$

$y = \Phi\theta + \epsilon, \quad \Phi = [\phi_1 \; \phi_2 \;\; \cdots \;\; \phi_n]^T$

Basis-function Regression

- Matrix-vector form of General Regression

  Let $\theta = [\theta_0 \;\; \theta_1 \;\; \cdots \; \theta_M]^T, \quad \phi_i = [1 \;\; \phi_{i1} \; \cdots \; \phi_{iM}]^T$

  $\boldsymbol{y} = [y_1 \;\; y_2 \;\; \cdots \;\; y_n]^T, \quad \boldsymbol{\epsilon} = [\epsilon_1 \;\; \epsilon_2 \;\; \cdots \;\; \epsilon_n]^T$

  Then $y_i = \phi_i^T \theta + \epsilon_i, \;\; i = 1, \; \cdots, \; n.$

  $\boldsymbol{y} = \Phi\theta + \boldsymbol{\epsilon}, \quad \Phi = [\phi_1 \, \phi_2 \;\; \cdots \;\; \phi_n]^T$

- Basis for General Regression
  - sin, cos basis: $\phi_{im} = \sin \omega_m x_i$ or $\cos \omega_m x_i$
  - radial basis: $\phi_{im} = \exp \frac{-\|x_i - \mu_m\|^2}{\sigma_m^2}$
  - sigmoid basis: $\phi_{im} = \frac{1}{1+\exp(-w_m^T x_i - b_m)}$ or $\frac{\exp(w_m^T x_i + b_m)}{1+\exp(w_m^T x_i + b_m)}$

Logistic Regression

Parameter Estimation in Matrix form

- Least Squares Estimation

$$\boldsymbol{y} = \Phi\theta + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$$

MLE:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|\epsilon\|^2}{2\sigma^2}\right)$$

LSE:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y} - \Phi\theta\|^2 \cong S(\theta)$$

Solution:

by $\nabla_\theta S(\theta) = 0$ at $\hat{\theta}$.

Parameter Estimation in Matrix form

- Least Squares Estimation

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y} - \Phi\theta\|^2 \cong S(\theta)$$

Solution:

$$\nabla_\theta S(\theta) = 0 \text{ at } \hat{\theta}$$

$$\nabla_\theta (\boldsymbol{y} - \Phi\theta)^T (\boldsymbol{y} - \Phi\theta) = 0 \text{ at } \hat{\theta}$$

$$2\Phi^T \left(\boldsymbol{y} - \Phi\hat{\theta}\right) = 0$$

$$\Phi^T \boldsymbol{y} - \Phi^T\Phi\hat{\theta} = 0$$

$$\hat{\theta} = (\Phi^T\Phi)^{-1} \Phi^T \boldsymbol{y}$$

Interim Summary

- linear regression
  - simple linear regression
  - multiple linear regression
- nonlinear regression
  - logistic regression
  - high-order regression
  - basis-function regression
- matrix form for regression
  - recursive least squares
- partial least squares
  - over-fitting and underfitting
  - bias/variance
  - principle component regression
  - partial least squares algorithm
  - ridge regression
  - lasso, elastic regression
- Gaussian process regression
- Kalman filtering