计算建模

第2章 随机模拟及参数估计(2学时)

刘绍辉

计算机科学与技术学院 哈尔滨工业大学

shliu@hit.edu.cn

2023年秋季

第2章 参数估计

2.0 引言

2.1 最小方差估计

2.1.1 Introduction
2.1.2 Unbiased Estimator
2.1.3 Minimum Variance Criterion
2.1.4 Existence of the MVUE

2.2 最小二乘估计

2.2.1 Outline
2.2.2 One-dimensional regression
2.2.3 Multi-dimentional linear regression

2.3 最大似然估计

2.3.1 Assumption
2.3.2 Problem Formulation
2.3.3 Solution
2.3.4 ML Estimation Using Log-Likelihood
2.3.5 MLE Examples

2.4 贝叶斯参数估计

2.4.1 Bayesian Estimation

2.5 最大后验估计

2.5.1 MAP

2.5.2 Examples

2.6 期望最大化算法

2.6.1 Description
2.6.2 Incomplete Data
2.6.3 Example
2.6.4 Convergence properties of EM

参考文献和站点

2.0计算机模拟

- **图像和信号处理都涉及随机变量的变换**
  - 即使单随机变量的一个函数，也很难根据输入分布来解析描述输出变量的分布
  - 通常会在合成数据上模拟输入分布，并对合成输入产生的输出数据进行统计分析
- **关键：模拟经验分布的数据能很好的拟合感兴趣随机变量的理论分布**
  - 假设X是随机变量，$x_1, x_2, \cdots, x_m$是计算机生成的用来模拟变量X的数据，如果由$x_1, x_2, \cdots, x_m$组成的经验分布能够很好地近似理论分布，则区间概率$P(a < X < b)$应当能够很好地由位于区间(a, b)中的合成数据所近似！
  - 最基本的就是：模拟$(0, 1)$上均匀分布的随机变量U
    - ✓ 思考如何生成？
  - 随机数生成器（random-number generators: pseudorandom）
- **逆变换法**
  - F是一个严格递增的连续概率分布函数，则$X = F^{-1}(U)$就是一个具有概率分布函数F的随机变量

2.0计算机模拟

- **逆变换法**
  - F是一个严格递增的连续概率分布函数，则$X=F^{-1}(U)$就是一个具有概率分布函数F的随机变量（注意U是$(0,1)$上的均匀分布）
  - $F_X(x)=P\left(F^{-1}(U)\leq x\right)=P\left(U\leq F(x)\right)=F(x)$
- **例：模拟参数为b的指数分布随机变量X,其概率分布函数$u=F(x)=1-e^{-bx}$.**
- **解：根据逆变换法，解出$x=-b^{-1}log(1-u)$，从而立即可得$X=-b^{-1}log(1-U)$,该指数分布均值为$\frac{1}{b}$.这里由于$1-U$也是均匀分布，因此进一步简化为：**

$$X=-b^{-1}logU$$

- **是否还有别的计算仿真概率分布的方法?**

2.0 引言

◆参数估计 Parameter Estimation

- 基本概念：我们通过概率分布来建模随机变量，还需要进一步确定分布的参数，从而能够使得概率反应了随机变量所描述的现象的分布！
  - ✓ 如X的密度函数为$p(x;\alpha,\beta)$，则意味着参数α,β未知，需要通过与随机变量X有关的观察来估计参数的值！

- 参数估计是根据从总体中抽取的样本估计总体分布中包含的未知参数的方法。

- 估计理论是统计学和信号处理中的一个分支，主要是通过测量或经验数据来估计概率分布参数的数值。这些参数描述了实质情况或实际对象，它们能够回答估计函数提出的问题。（维基百科）

2.0引言

- **参数估计可分为点估计和区间估计**
  - 点估计：用样本统计量的某一具体数值直接推断未知的参数（总体参数）
  - 区间估计：推断未知参数，并根据统计量的抽样分布的特征，估计参数的一个区间，而不是一个数值，并同时给出总体参数落在这一区间的可能性大小，一般用概率表示
- **点估计**
  - 矩估计法：简单，但无法给出误差大小，样本统计量是一个随机变量，估计一个确定值，肯定有误差！一般用于定性研究中！
  - 最大似然估计法

2.0 引言

◆参数估计 Parameter Estimation

- 估计过程
- 估计理论的全部目的都是获取一个估计函数，最好是一个可以实现的估计函数。估计函数输入测量数据，输出相应参数的估计。
- 最常用的方法：观察随机变量集合$X_1, X_2, \cdots, X_n$，这些都是独立同分布于随机变量X，也称之为X的随机样本。若X的密度函数为$f(x)$，则由其独立同分布特性，样本的联合概率密度函数为：

$$f(x_1, x_2, \cdots, x_n) = \Pi_{k=1}^{n} f(x_k)$$

- 假设X的概率密度函数为具有未知参数θ的$f(x;\theta)$,此时随机样本的函数给出参数的一个估计值：

$$\hat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_n})$$

- $\hat{\theta}$是一个随机变量，也称为参数θ的估计子(estimator)。
- 给定一组样本值$x_1, x_2, \cdots, x_n$，就获得一个参数的估计值，定义这个估计估计函数的规则称为估计子的估计规则。若估计规则是与未知参数无关的，则估计子称为统计估计子。
- 若有两组观测值$x_1, x_2, \cdots, x_n$和$z_1, z_2, \cdots, z_n$则可得两个估计值，是否一个比另一个好？或者两个都好？与估计规则有关！

2.0引言

- **无偏估计(unbiased)**
  - $E[\hat{\theta}] = \theta$，无偏性
  - $E[\hat{\theta}] - \theta$称之为$\hat{\theta}$的偏差
- **渐进无偏**
  - 由于$\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_n)$与样本大小$n$有关，很多情况下估计子是右边的，但是满足：
  - $lim_{(n\to\infty)} E[\hat{\theta}] \to \theta$,则称之为渐进无偏的！
- **无偏性表明平均来说估计子给出了期望的参数值，但如果估计子的方差很大，则也没有实用价值！有效性！**
  - 期望的精度是期望估计子能给出参数在某一容忍度之内的估计
  - 注意，估计子也是一个随机变量，因此对$r > 0$,考虑参数在r之内的概率：
  - $P(|\hat{\theta} - \theta| < r)$,如果对任意$r > 0$，都有$lim_{n\to\infty} P(|\hat{\theta} - \theta| < r) = 1$（依概率收敛），则称之为估计子是参数的一致估计！一致性！

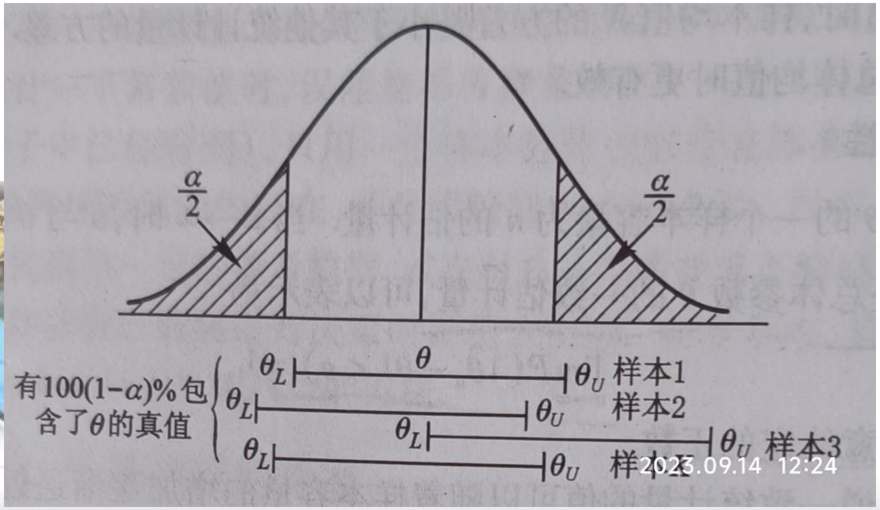
2.0引言

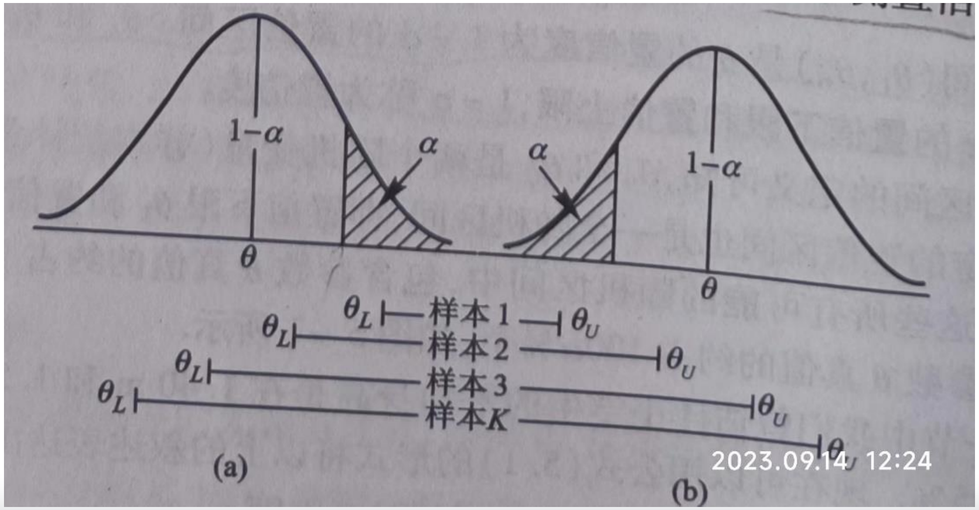
- 如果$\hat{\theta}$是无偏的，则根据Chebyshev不等式：$P\left(\left|\hat{\theta}-\theta\right|<r\right)\ge 1-\frac{Var\left[\hat{\theta}\right]}{r^2}$,因而，如果$\hat{\theta}$是无偏的，且$\lim_{n\to\infty}Var\left[\hat{\theta}\right]=0$,则$\hat{\theta}$是一致估计子！
- 结论可进一步强化为：如果$\hat{\theta}$是渐进无偏估计子，且$\lim_{n\to\infty}Var\left[\hat{\theta}\right]=0$，则$\hat{\theta}$是一致估计子
  - 渐进无偏，只要n足够大，则$\left|E\left[\hat{\theta}\right]-\theta\right|<\frac{r}{2}$，因此
  - $P\left(\left|\hat{\theta}-\theta\right|\ge r\right)\le P\left(\left|\hat{\theta}-E\left[\hat{\theta}\right]\right|+\left|E\left[\hat{\theta}\right]-\theta\right|\ge r\right)\le P\left(\left(\left|\hat{\theta}-E\left[\hat{\theta}\right]\right|\ge\frac{r}{2}\right)\cup\left(\left|\theta-E\left[\hat{\theta}\right]\right|\ge\frac{r}{2}\right)\right)=P\left(\left|\hat{\theta}-E\left[\hat{\theta}\right]\right|\ge\frac{r}{2}\right)\le\frac{4Var\left[\hat{\theta}\right]}{r^2}$

2.0引言

◆区间估计之置信区间

- ➢设$\boldsymbol{\theta}$是待估计的总体参数，$\boldsymbol{\alpha}$是属于$(\mathbf{0},\mathbf{1})$的数值，由样本确定两个统计量$\boldsymbol{\theta}_{L},\boldsymbol{\theta}_{U}$满足
- ➢$\boldsymbol{P}(\boldsymbol{\theta}_{L}<\boldsymbol{\theta}<\boldsymbol{\theta}_{U})=\mathbf{1}-\boldsymbol{\alpha}$
- ➢就称随机区间$(\boldsymbol{\theta}_{L},\boldsymbol{\theta}_{U})$是$\boldsymbol{\theta}$的置信度为$\mathbf{1}-\boldsymbol{\alpha}$的置信区间。$\boldsymbol{\theta}_{L}$和$\boldsymbol{\theta}_{U}$分别称为置信度为$\mathbf{1}-\boldsymbol{\alpha}$的置信下限和置信上限，$\mathbf{1}-\boldsymbol{\alpha}$称为置信度！





1−α
α
θ
α
1−α
θ
θ_L 样本1 θ_U
θ_L 样本2 θ_U
θ_L 样本3 θ_U
θ_L 样本K θ_U
(a)
(b)
2023.09.14 12:24

2.0引言

◆置信区间的求解

- 1.找到一个包含$\boldsymbol{\theta}$的且不包含任何其他未知参数的已知抽样分布的随机变量，如$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$;
- 2.根据给定的$1-\alpha$，在抽样分布中确定两个（单侧置信区间是一个）临界点，如$\left(-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}\right)$，使其满足：$P\left(-z_{\frac{\alpha}{2}}<Z<z_{\frac{\alpha}{2}}\right)=1-\alpha$
- 3.将其进行等价变换，得到

$P(\theta_L<\theta<\theta_U)=1-\alpha$即为参数$\boldsymbol{\theta}$的置信度为$1-\alpha$的置信区间！

- 抽样分布的统计量已经有很多现存的，可以查阅相关资料！

◆置信度与置信区间以及样本量的关系

- 样本容量一定时，置信度越高，置信区间的范围越大，即估计的参数的相对精度越低
- 通过增加样本容量来提高精度，但抽样费用增加！

2.0 引言

◆参数估计 Parameter Estimation

- 参数估计方法
- 最小方差无偏估计（Minimum variance unbiased estimator，简称MVUE）
- 最小均方差（Minimum mean squared error，简称MMSE）
- 最大似然估计（Maximum likelihood estimation，简称MLE）
- 矩估计（Method of moments estimators，简称MME）
- 贝叶斯估计
- 最大后验概率（Maximum a posteriori probability，简称MAP）
- 最佳线性非偏估计（BLUE）
- 期望最大化算法（EM）

2.1 最小方差无偏估计（MVUE）

◆Introduction

- ➢ In this chapter we will begin our search for good estimators of unknown deterministic parameters.
- ➢ We will restrict our attention to estimators which on the average yield the true parameter value.
- ➢ Then, within this class of estimators the goal will be to find the one that exhibits the least variability.
- ➢ The estimator thus obtained will produce values close to the true value most of the time.
- ➢ The notion of a minimum variance unbiased estimator is examined within this chapter.

2.1 最小方差无偏估计（MVUE）

◆Introduction

- For an estimator to be unbiased we mean that on the average the estimator will yield the true value of the unknown parameter.
- Since the parameter value may in general be anywhere in the interval $a<\theta<b$, unbiasedness asserts that no matter what the true value of θ, our estimator will yield it on the average.

$$E\left(\theta\right)=\theta \qquad a<\theta<b \qquad (2.1)$$

2.1 最小方差无偏估计（MVUE）

◆Introduction

- **Example 2.1 (1/2)**
- Consider the observations

$$x[n] = A + w[n] \qquad n = 0, 1, \ldots, N-1$$

- where A is the parameter to be estimated and w[n] is WGN(white gaussian noise). The parameter A can take on any value in the interval $-\infty < A < \infty$.
- The reasonable estimator for the average value of x[n] is

$$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x[n] \qquad (2.2)$$

or the sample mean.

2.1 最小方差无偏估计（MVUE）

- **Example 2.1 (2/2)**
- Due to the linearity properties of the expectation operator

$$
\begin{aligned}
E(\hat{A}) &= E\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] \\
&= \frac{1}{N}\sum_{n=0}^{N-1} E(x[n]) \\
&= \frac{1}{N}\sum_{n=0}^{N-1} A \\
&= A
\end{aligned}
\qquad (2.3)
$$

for all A. The sample mean estimator is unbiased.

2.1 最小方差无偏估计（MVUE）

◆Unbiased Estimators

- The restriction that $E\left(\theta\right)=\theta$ for all θ is an important one.
- It is possible that may hold for some values of θ and not others.

2.1 最小方差无偏估计（MVUE）

- **Example 2.2**
- Consider again Example 2.1 but with the modified sample mean estimator

$$\check{A} = \frac{1}{2N}\sum_{n=0}^{N-1} x[n].$$

- Then,

$$\begin{aligned} E(\check{A}) &= \frac{1}{2}A \\ &= A \text{ if } A = 0 \\ &\neq A \text{ if } A \neq 0. \end{aligned}$$

- It is seen that (2.3) holds for the modified estimator only for A = 0.
- Clearly, it is a biased estimator.

2.1 最小方差无偏估计（MVUE）

◆Unbiased Estimators

- That an estimator is unbiased does not necessarily mean that it is a good estimator.
- It only guarantees that on the average it will attain the true value.
- A persistent bias will always result in a poor estimator.
- As an example, the unbiased property has an important implication when several estimators are combined. A reasonable procedure is to combine these estimates into a better one by averaging them to form

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i. \qquad (2.4)$$

2.1 最小方差无偏估计（MVUE）

◆Unbiased Estimators

- Assuming the estimators are unbiased, with the same variance, and uncorrelated with each other,

$$E(\hat{\theta}) = \theta$$

and

$$\begin{aligned} \mathrm{var}(\hat{\theta}) &= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(\hat{\theta}_i) \\ &= \frac{\mathrm{var}(\hat{\theta}_1)}{n} \end{aligned} \qquad (2.5)$$

so that as more estimates are averaged, the variance will decrease.

2.1 最小方差无偏估计（MVUE）

◆Unbiased Estimators

- However, if the estimators are biased or $E(\hat{\theta}_i) = \theta + b(\theta)$, then

$$\begin{aligned} E(\hat{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} E(\hat{\theta}_i) \\ &= \theta + b(\theta) \end{aligned}$$

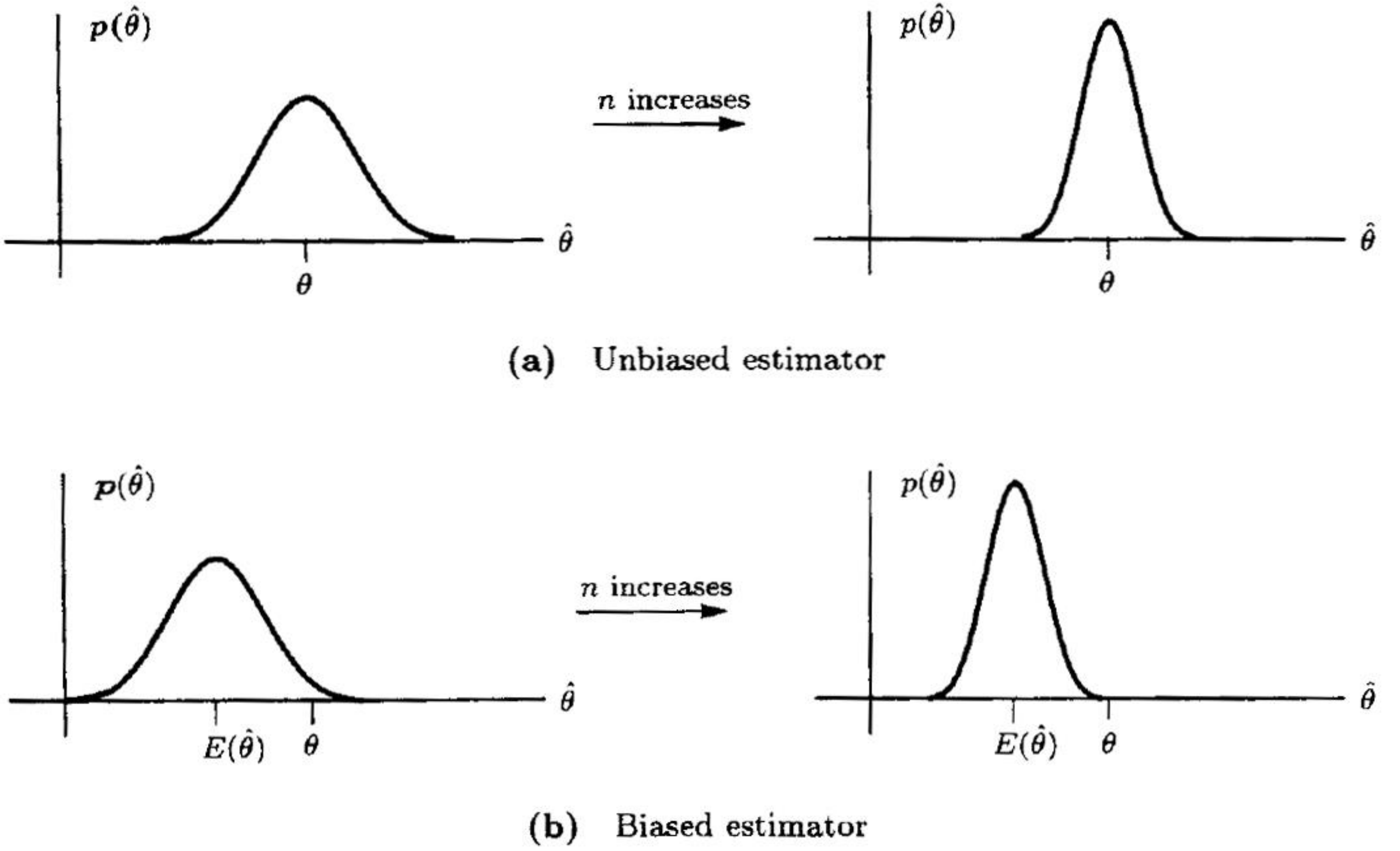
  and no mater how many estimators are averaged, $\hat{\theta}$ will not converge to the true value.

- Note that, in general,

$$b(\theta) = E(\hat{\theta}) - \theta$$

is defined as the bias of the estimator.

2.1 最小方差无偏估计（MVUE）



(a) Unbiased estimator

(b) Biased estimator

Figure 2.2 Effect of combining estimators

2.1 最小方差无偏估计（MVUE）

◆Minimum Variance Criterion

- In searching for optimal estimators we need to adopt some optimality criterion.
- A natural one is the mean square error (MSE), defined as

$$\mathrm{mse}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

- Unfortunately, adoption of this natural criterion leads to unrealizable estimators, ones that cannot be written solely as a function of the data.

2.1 最小方差无偏估计（MVUE）

◆Minimum Variance Criterion

- To understand the problem which arises we first rewrite the MSE as

$$
\begin{aligned}
\operatorname{mse}(\hat{\theta}) &= E\left\{\left[\left(\hat{\theta}-E(\hat{\theta})\right)+\left(E(\hat{\theta})-\theta\right)\right]^2\right\} \\
&= \operatorname{var}(\hat{\theta})+\left[E(\hat{\theta})-\theta\right]^2 \\
&= \operatorname{var}(\hat{\theta})+b^2(\theta) \qquad (2.6)
\end{aligned}
$$

which shows that the MSE is composed of errors due to the variance of the estimator as well as the bias.

2.1 最小方差无偏估计（MVUE）

◆Minimum Variance Criterion

- As an example, for the problem in Example 2.1 consider the modified estimator

$$\check{A} = a\frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

for come constant a.

- We will attempt to find the a which results in the minimum MSE.
- Since $E(\check{A}) = aA$ and $\mathrm{var}(\check{A}) = a^2\sigma^2/N$, we have

$$\mathrm{mse}(\check{A}) = \frac{a^2\sigma^2}{N} + (a-1)^2A^2.$$

2.1 最小方差无偏估计（MVUE）

◆Minimum Variance Criterion

- Differentiating the MSE with respect to a yields

$$\frac{d\,\mathrm{mse}(\breve{A})}{da} = \frac{2a\sigma^2}{N} + 2(a-1)A^2$$

which upon setting to zero and solving yields the optimum value

$$a_{\mathrm{opt}} = \frac{A^2}{A^2 + \sigma^2/N}.$$

- It is seen that the **optimal value of a depends upon the unknown parameter A. The estimator is therefore not realizable**.

2.1 最小方差无偏估计（MVUE）

◆Minimum Variance Criterion

- An alternative approach is to constrain the bias to be zero and find the estimator which minimizes the variance.
- Such an estimator is termed the minimum variance unbiased (MVU) estimator.
- Note that from (2.6) that the MSE of an unbiased estimator is just the variance.
- Minimizing the variance of an unbiased estimator also has the effect of concentrating the PDF of the estimation error about zero.
- The estimation error will therefore be less likely to be large.

2.1 最小方差无偏估计（MVUE）

◆Existence of the Minimum Variance Unbiased Estimator

- The question arises as to whether a MVU estimator exists, i.e., an unbiased estimator with minimum variance for all θ.

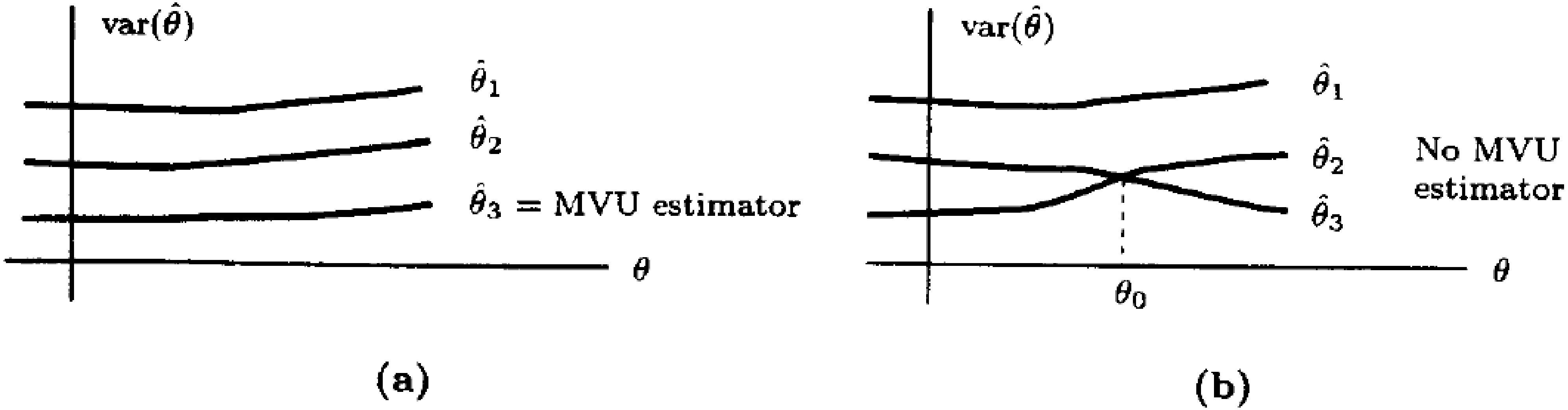


Figure 2.3 Possible dependence of estimator variance with θ

2.1 最小方差无偏估计（MVUE）

◆Existence of the Minimum Variance Unbiased Estimator

- **Example 2.3 (1/3)**
- Assume that we have two independent observations x[0] and x[1] with PDF

$$
\begin{aligned}
x[0] &\sim \mathcal{N}(\theta, 1) \\
x[1] &\sim \begin{cases} \mathcal{N}(\theta, 1) & \text{if } \theta \geq 0 \\ \mathcal{N}(\theta, 2) & \text{if } \theta < 0. \end{cases}
\end{aligned}
$$

- The two estimators

$$
\begin{aligned}
\hat{\theta}_1 &= \frac{1}{2}\left(x[0] + x[1]\right) \\
\hat{\theta}_2 &= \frac{2}{3}x[0] + \frac{1}{3}x[1]
\end{aligned}
$$

can easy be shown to be unbiased.

2.1 最小方差无偏估计（MVUE）

◆Existence of the Minimum Variance Unbiased Estimator

- **Example 2.3 (2/3)**
- To compute the variances we have that

$$\begin{aligned}\operatorname{var}(\hat{\theta}_1) &= \frac{1}{4}\left(\operatorname{var}(x[0]) + \operatorname{var}(x[1])\right)\\ \operatorname{var}(\hat{\theta}_2) &= \frac{4}{9}\operatorname{var}(x[0]) + \frac{1}{9}\operatorname{var}(x[1])\end{aligned}$$

- so that

$$\operatorname{var}(\hat{\theta}_1) = \begin{cases} \frac{18}{36} & \text{if } \theta \geq 0 \\ \frac{27}{36} & \text{if } \theta < 0 \end{cases}$$

- and

$$\operatorname{var}(\hat{\theta}_2) = \begin{cases} \frac{20}{36} & \text{if } \theta \geq 0 \\ \frac{24}{36} & \text{if } \theta < 0. \end{cases}$$

2.1 最小方差无偏估计（MVUE）

◆Existence of the Minimum Variance Unbiased Estimator

- **Example 2.3 (3/3)**
- Clearly, between these two estimators no MVU estimator exists.
- No single estimator can have a variance uniformly less than or equal the minima.

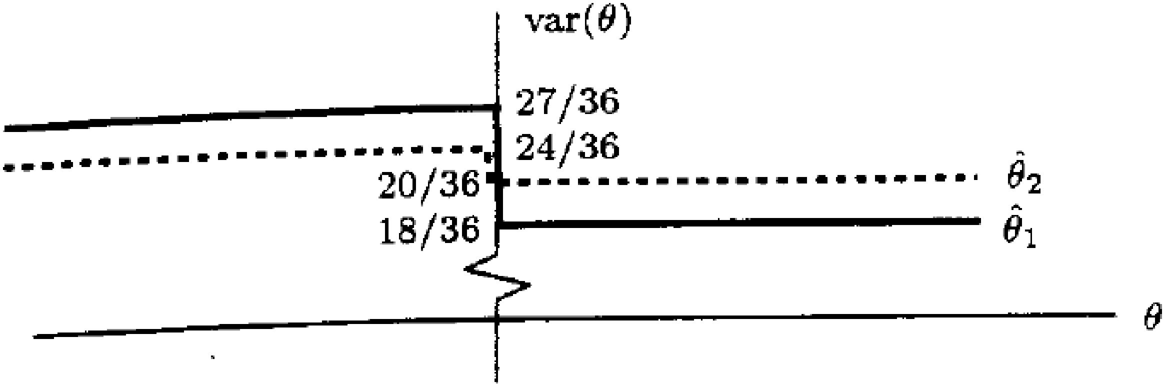


Figure 2.4 Illustration of nonexistence of minimum variance unbiased estimator

◆Let $X_1, X_2, \cdots, X_n$ be a random sample from a random variable X with mean μ. Let C be the class of all linear unbiased estimators of μ, these being of the form $\hat{\mu} = \sum_{i=1}^{n} a_k X_k$, $E[\hat{\mu}] = \mu$

2.1 最小方差无偏估计（MVUE）

- $E[\hat{\mu}] = \left(\sum_{i=1}^{n} a_k\right)\mu \Rightarrow \sum_{i=1}^{n} a_k = 1$.
- The variance of $\hat{\mu}$ is:
- $\sigma_{\hat{\mu}}^2 = \sum_{k=1}^{n} a_k^2 \sigma^2 = \left(\sum_{k=1}^{n-1} a_k^2 + \left(1 - \sum_{k=1}^{n-1} a_k\right)^2\right)\sigma^2$
- For $j = 1, 2, \cdots, n-1$
- $\frac{\partial}{\partial a_j}\sigma_{\hat{\mu}}^2 = \left(a_j - 1 + \sum_{k=1}^{n-1} a_k\right)2\sigma^2 = 0 \Rightarrow a_1 + \cdots + a_{j-1} + 2a_j + a_{j+1} + \cdots + a_{n-1} = 1$,then yields the system

$$\begin{pmatrix} 2 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 1 & \cdots & 1 \\ 1 & 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_{n-1} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

- Solving it, $a_j = \frac{1}{n}, for\ j = 1, 2, \cdots n-1$. The system matrix is nonsingular, this solution is unique. Due to the sum of the coefficients (=1), $a_n = \frac{1}{n}$, and the best linear unbiased estimator is the sample mean！

2.1 最小方差无偏估计（MVUE）

Finding the Minimum Variance Unbiased Estimator

- **Generally difficult! Checking whether a particulat unbiased estimator has minimum variance is relative easy!**
- **Several possible approaches:**

1. Determine the Cramer-Rao lower bound (CRLB) and check to see if some estimator satisfies it.
2. Apply the Rao-Blackwell-Lehmann-Scheffe (RBLS) theorem.
3. Further restrict the class of estimators to be not only unbiased but also linear. Then, find the minimum variance estimator within this restricted class。

CRLB: Under certain regularity conditions, if $\boldsymbol{X_1, X_2, \cdots, X_n}$ comprise a random sample arising from a random variable $\boldsymbol{X}$ and $\boldsymbol{\hat{\theta}}$ is an unbiased estimator of the parameter $\boldsymbol{\theta}$ in the density $\boldsymbol{f(x;\theta)}$ of $\boldsymbol{X}$, then

$$\boldsymbol{Var\left[\hat{\theta}\right] \geq \frac{1}{nE\left[\left(\frac{\partial}{\partial\theta} log f(X;\theta)\right)^2\right]}}$$

2.1 最小方差无偏估计（MVUE）

◆Finding the Minimum Variance Unbiased Estimator

- The CRLB allow us to determine that for any unbiased estimator the variance must be greater than or equal to a given value.
- If an estimator exists whose variance equals the CRLB for each value of θ, then it must be the MVU estimator.

2.1 最小方差无偏估计（MVUE）

◆Extension to a Vector Parameter

- If $\boldsymbol{\theta} = [\theta_1\ \theta_2 \ldots \theta_p]^T$ ı vector of unknown parameters, then we say that an estimator $\hat{\boldsymbol{\theta}} = [\hat{\theta}_1\ \hat{\theta}_2 \ldots \hat{\theta}_p]^T$ is unbiased if

$$E(\hat{\theta}_i) = \theta_i \qquad a_i < \theta_i < b_i \qquad (2.7)$$

for i = 1, 2, …, p.

By defining

$$E(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} E(\hat{\theta}_1) \\ E(\hat{\theta}_2) \\ \vdots \\ E(\hat{\theta}_p) \end{bmatrix}$$

2.1 最小方差无偏估计（MVUE）

◆Extension to a Vector Parameter

- We can equivalently define an unbiased estimator to have the property

$$E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$$

for every θ contained within the space defined in (2.7).

- A MVU estimator has the additional property that $\mathrm{var}(\hat{\theta}_i)$ for i = 1, 2, ..., p is minimum among all unbiased estimators.

Ordinary Least-Squares

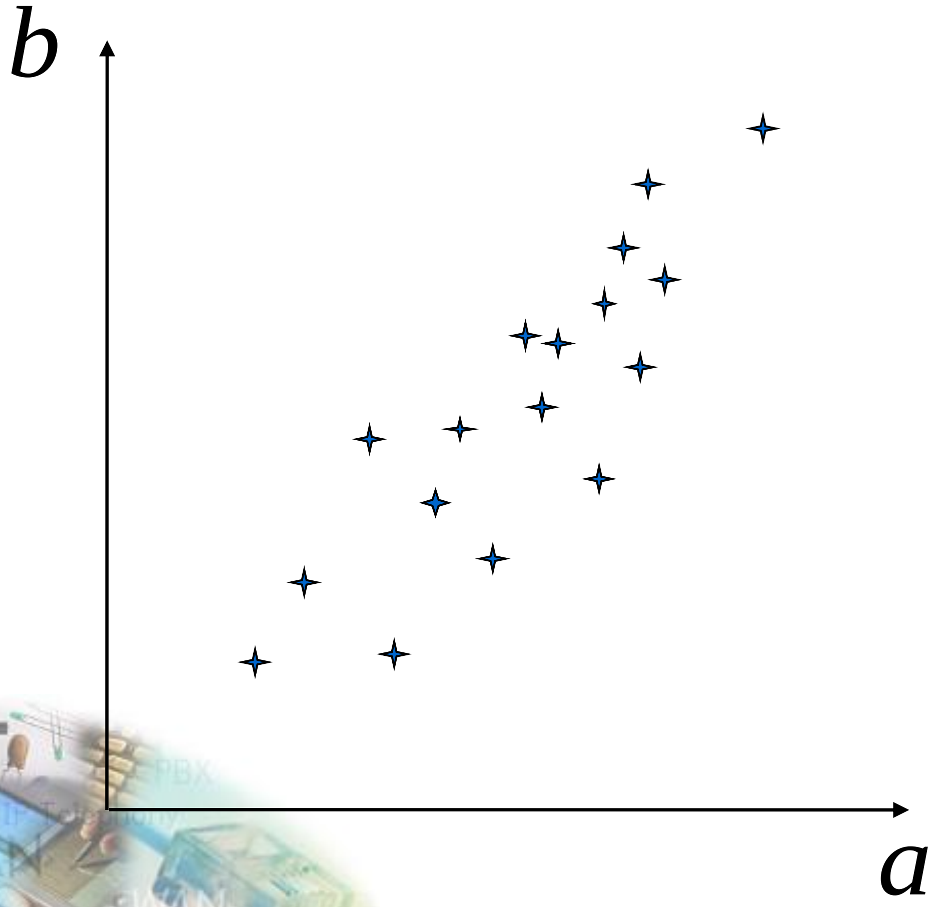
2.2 最小二乘估计Least-Squares

◆Outline

- ➢Linear regression
- ➢Geometry of least-squares
- ➢Discussion of the Gauss-Markov theorem

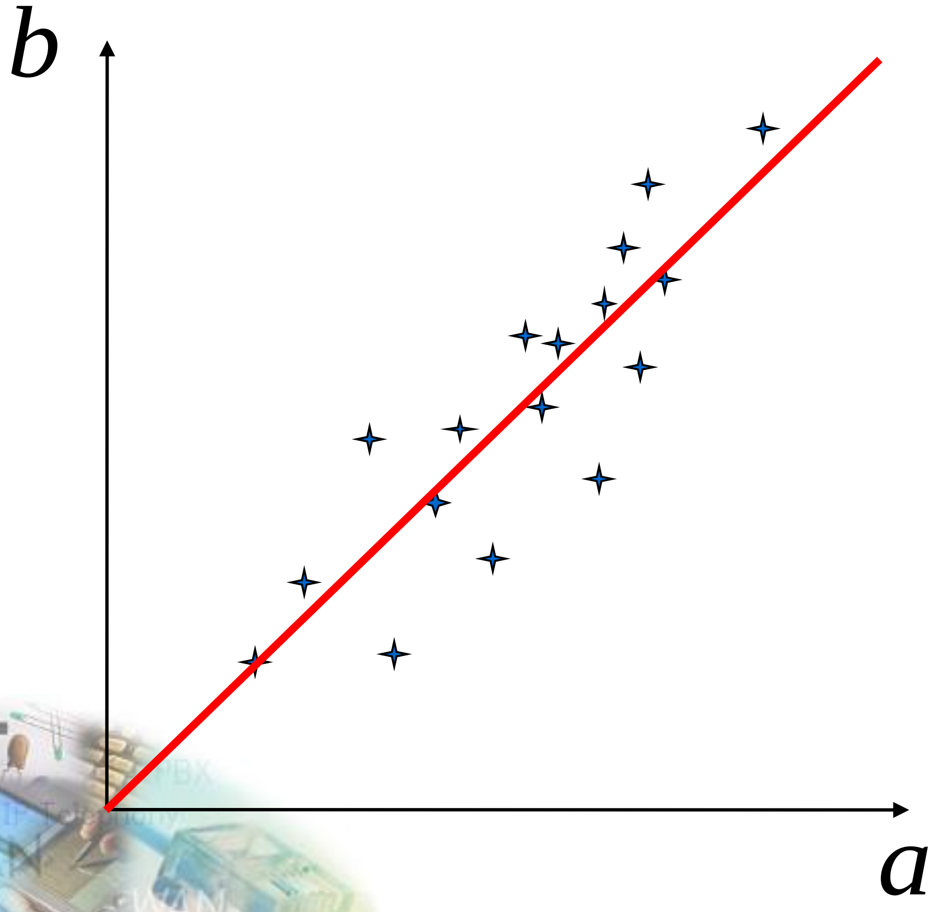
2.2 最小二乘估计Least-Squares

◆**One-dimensional regression**



2.2 最小二乘估计Least-Squares

◆One-dimensional regression

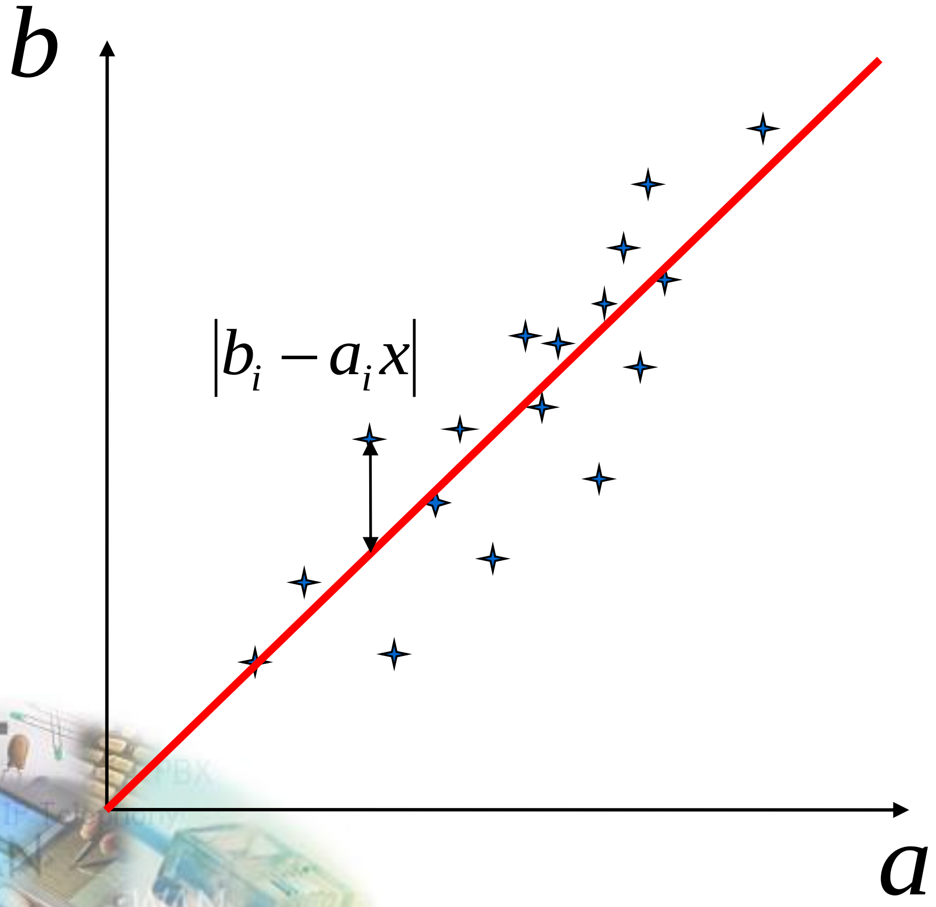


Find a line that represent the "best" linear relationship:

$$b = ax$$

2.2 最小二乘估计Least-Squares

◆One-dimensional regression

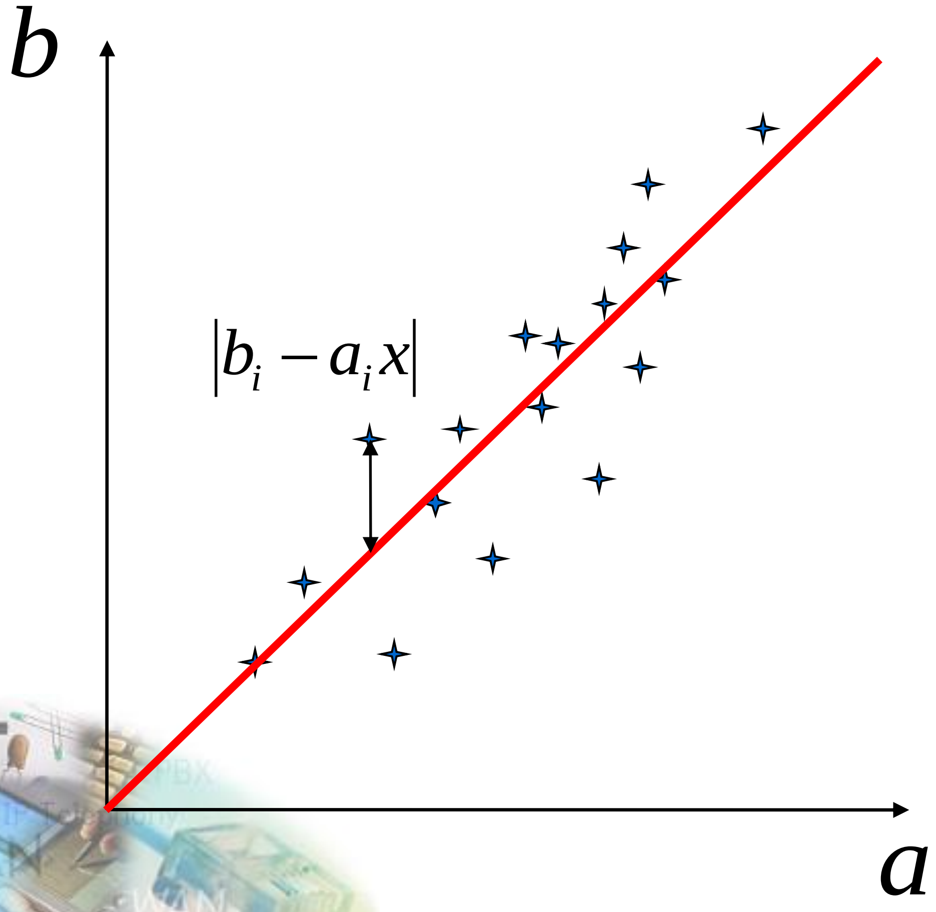


- Problem: the data does not go through a line

$$e_i = b_i - a_i x$$

2.2 最小二乘估计Least-Squares

◆One-dimensional regression



- Problem: the data does not go through a line

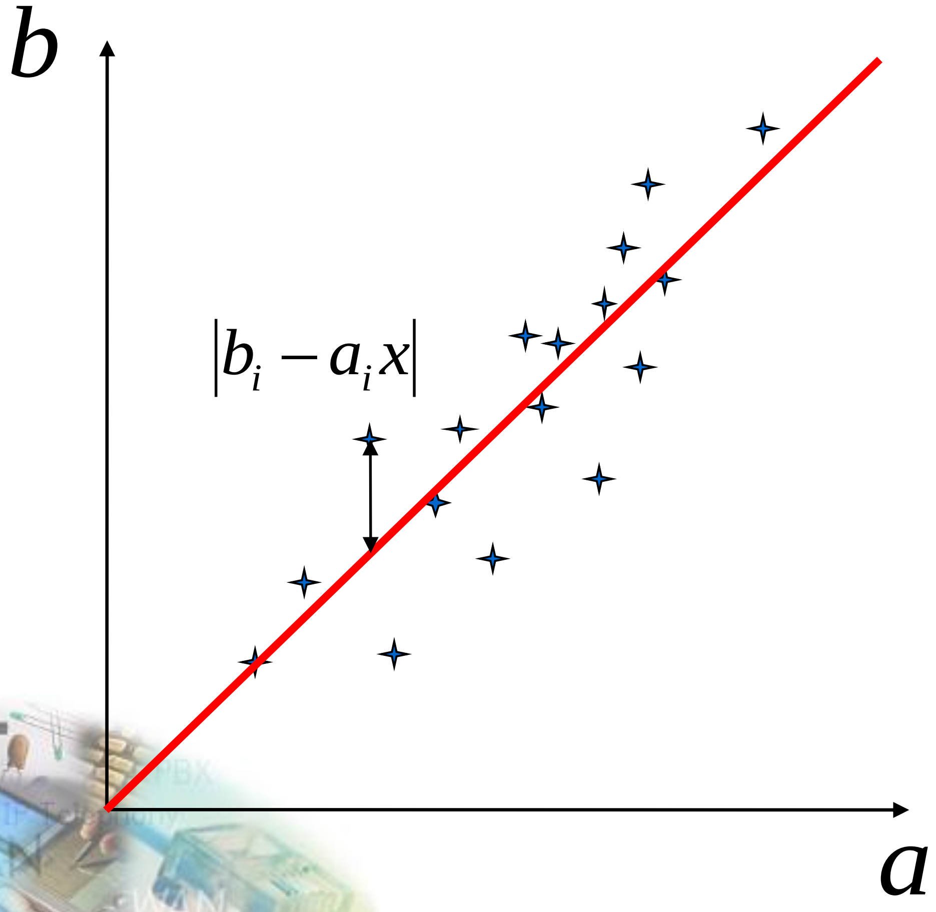
$$e_i = b_i - a_i x$$

- Find the line that minimizes the sum:

$$\sum_i (b_i - a_i x)^2$$

2.2 最小二乘估计Least-Squares

◆One-dimensional regression



- Problem: the data does not go through a line

$$e_i = b_i - a_i x$$

- Find the line that minimizes the sum:

$$\sum_i (b_i - a_i x)^2$$

- We are looking for $\hat{x}$ that minimizes

$$e(x) = \sum_i (b_i - a_i x)^2$$

2.2 最小二乘估计Least-Squares

◆Matrix notation

Using the following notations

$$\boldsymbol{a} = \begin{bmatrix} a_1 \\ : \\ a_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ : \\ b_n \end{bmatrix}$$

2.2 最小二乘估计Least-Squares

◆Matrix notation

Using the following notations

$$\boldsymbol{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$

we can rewrite the error function using linear algebra as:

$$e(x) = \sum_i (b_i - a_i x)^2$$

$$= (\boldsymbol{b} - x\boldsymbol{a})^T (\boldsymbol{b} - x\boldsymbol{a})$$

$$e(x) = \|\boldsymbol{b} - x\boldsymbol{a}\|^2$$

2.2 最小二乘估计Least-Squares

◆Matrix notation

Using the following notations

$$\boldsymbol{a} = \begin{bmatrix} a_1 \\ : \\ a_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ : \\ b_n \end{bmatrix}$$

we can rewrite the error function using linear algebra as:

$$e(x) = \sum_i (b_i - a_i x)^2$$

$$= (\boldsymbol{b} - x\boldsymbol{a})^T (\boldsymbol{b} - x\boldsymbol{a})$$

$$\boxed{e(x) = \|\boldsymbol{b} - x\boldsymbol{a}\|^2}$$

2.2 最小二乘估计Least-Squares

◆Multidimentional linear regression

Using a model with *m* parameters

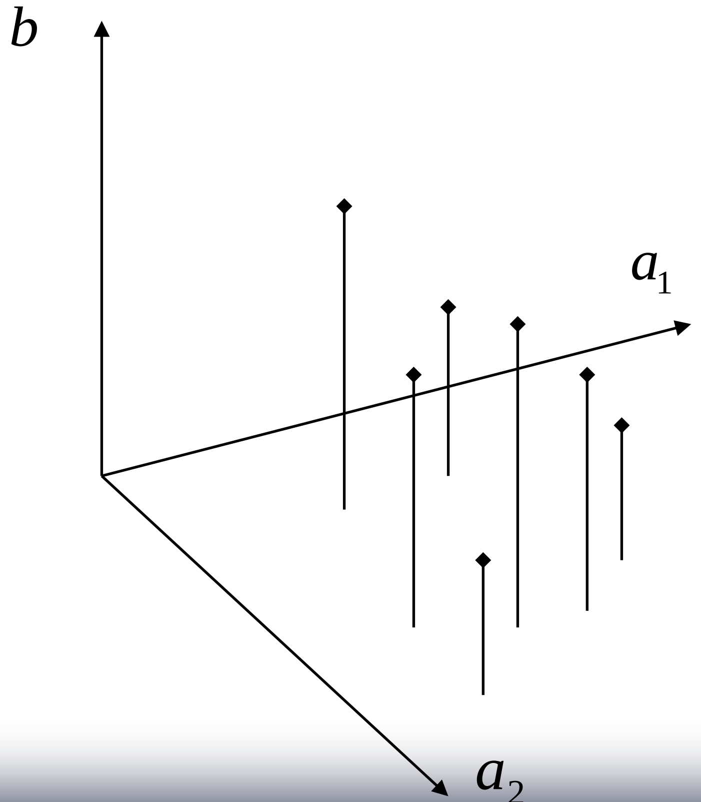
$$b = a_1 x_1 + \ldots + a_m x_m = \sum_j a_j x_j$$

2.2 最小二乘估计Least-Squares

◆Multidimentional linear regression

Using a model with *m* parameters

$$b = a_1 x_1 + \ldots + a_m x_m = \sum_j a_j x_j$$

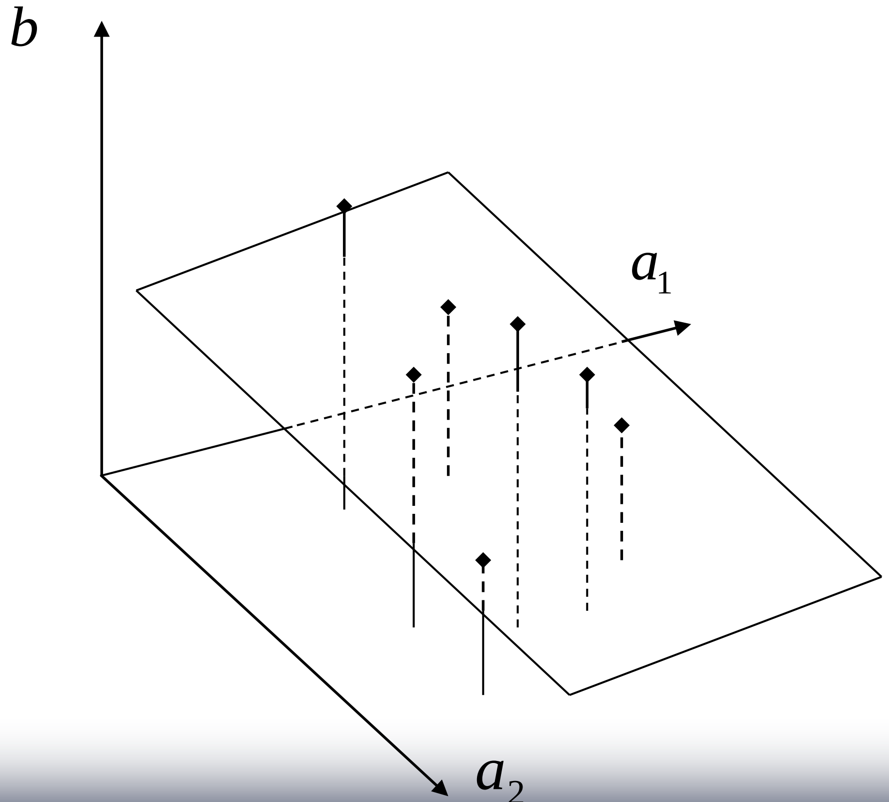


2.2 最小二乘估计Least-Squares

◆Multidimentional linear regression

Using a model with *m* parameters

$$b = a_1 x_1 + \ldots + a_m x_m = \sum_j a_j x_j$$



2.2 最小二乘估计Least-Squares

◆Multidimentional linear regression

Using a model with *m* parameters

$$b = a_1 x_1 + \ldots + a_m x_m = \sum_j a_j x_j$$

and *n* measurements

$$e(\boldsymbol{x}) = \sum_{i=1}^{n} \left(b_i - \sum_{j=1}^{m} a_{i,j} x_j \right)^2$$

$$= \left\| \boldsymbol{b} - \left[\sum_{j=1}^{m} a_{i,j} x_j \right] \right\|^2$$

$$= \left\| \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \right\|^2$$

2.2 最小二乘估计Least-Squares

◆Multidimentional linear regression

Using a model with *m* parameters

$$b = a_1 x_1 + \ldots + a_m x_m = \sum_j a_j x_j$$

and *n* measurements

$$e(\boldsymbol{x}) = \sum_{i=1}^{n} \left(b_i - \sum_{j=1}^{m} a_{i,j} x_j\right)^2$$

$$= \left\| \boldsymbol{b} - \left[\sum_{j=1}^{m} a_{i,j} x_j \right] \right\|^2$$

$$\boxed{e(\boldsymbol{x}) = \left\| \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \right\|^2}$$

2.2 最小二乘估计Least-Squares

- **Multidimentional linear regression** $\boldsymbol{b}-\boldsymbol{Ax}$

$$\boldsymbol{b}-\boldsymbol{Ax}=\begin{bmatrix} b_1 \\ : \\ b_n \end{bmatrix}-\begin{bmatrix} a_{1,1} & .. & a_{1,m} \\ : & & : \\ a_{n,1} & .. & a_{n,m} \end{bmatrix}.\begin{bmatrix} x_1 \\ : \\ x_m \end{bmatrix}$$

2.2 最小二乘估计Least-Squares

◆Multidimentional linear regression

$$\boldsymbol{b}-\boldsymbol{Ax}=\begin{bmatrix} b_1 \\ : \\ b_n \end{bmatrix}-\begin{bmatrix} a_{1,1} & \cdot\cdot & a_{1,m} \\ : & & : \\ a_{n,1} & \cdot\cdot & a_{n,m} \end{bmatrix}\cdot\begin{bmatrix} x_1 \\ : \\ x_m \end{bmatrix}$$

$$=\begin{bmatrix} b_1-(a_{1,1}x_1+\ldots+a_{1,m}x_m) \\ : \\ b_n-(a_{n,1}x_1+\ldots+a_{n,m}x_m) \end{bmatrix}$$

2.2 最小二乘估计Least-Squares

parameter 1

$$\boldsymbol{b}-\boldsymbol{Ax}=\begin{bmatrix} b_1 \\ : \\ b_n \end{bmatrix}-\begin{bmatrix} \boxed{\begin{matrix} a_{1,1} \\ : \\ a_{n,1} \end{matrix}} & \begin{matrix} \cdot\cdot \\ \\ \cdot\cdot \end{matrix} & \begin{matrix} a_{1,m} \\ : \\ a_{n,m} \end{matrix} \end{bmatrix}.\begin{bmatrix} x_1 \\ : \\ x_n \end{bmatrix}$$

$$=\begin{bmatrix} b_1-(a_{1,1}x_1+\ldots+a_{1,m}x_m) \\ : \\ b_n-(a_{n,1}x_1+\ldots+a_{n,m}x_m) \end{bmatrix}$$

2.2 最小二乘估计Least-Squares

parameter 1

$$\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}=\begin{bmatrix} b_1\\ : \\ b_n\end{bmatrix}-\begin{bmatrix} a_{1,1} & .. & a_{1,m}\\ : & & :\\ a_{n,1} & .. & a_{n,m}\end{bmatrix}.\begin{bmatrix} x_1\\ :\\ x_n\end{bmatrix}$$

measurement n

$$=\begin{bmatrix} b_1-(a_{1,1}x_1+\ldots+a_{1,m}x_m)\\ : \\ b_n-(a_{n,1}x_1+\ldots+a_{n,m}x_m)\end{bmatrix}$$

2.2 最小二乘估计Least-Squares

◆Multidimentional linear regression

- ➢Minimizing $e(\boldsymbol{x})$

$$\boldsymbol{x}_{\min}\ \text{minimizes}\ e(\boldsymbol{x})\ \text{if}\ \left\{ \vphantom{\begin{matrix} \\ \\ \\ \\ \\ \end{matrix}} \right.$$

2.2 最小二乘估计Least-Squares

◆Multidimentional linear regression

- ➤Minimizing $e(\boldsymbol{x})$

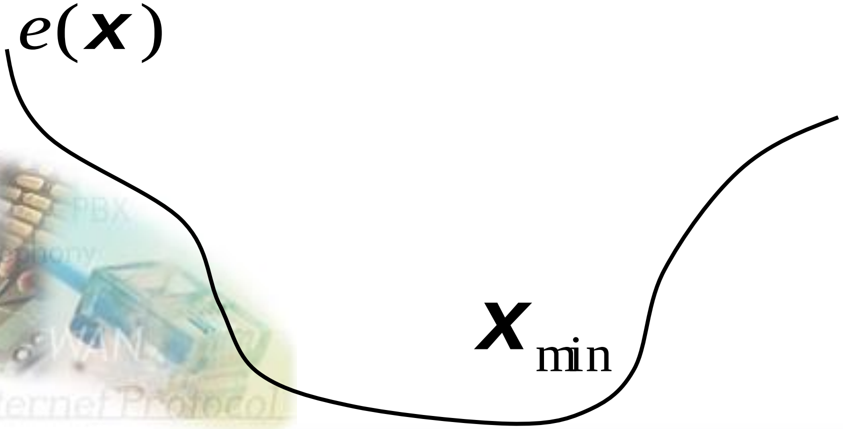
$\boldsymbol{x}_{\min}$ minimizes $e(\boldsymbol{x})$ if

$e(\boldsymbol{x})$ is flat at $\boldsymbol{x}_{\min}$

$$\nabla e(\boldsymbol{x}_{\min}) = \boldsymbol{0}$$

$e(\boldsymbol{x})$ does not go down around $\boldsymbol{x}_{\min}$

$H_e(\boldsymbol{x}_{\min})$ is positive semi-definite



2.2 最小二乘估计Least-Squares

◆Multidimentional linear regression

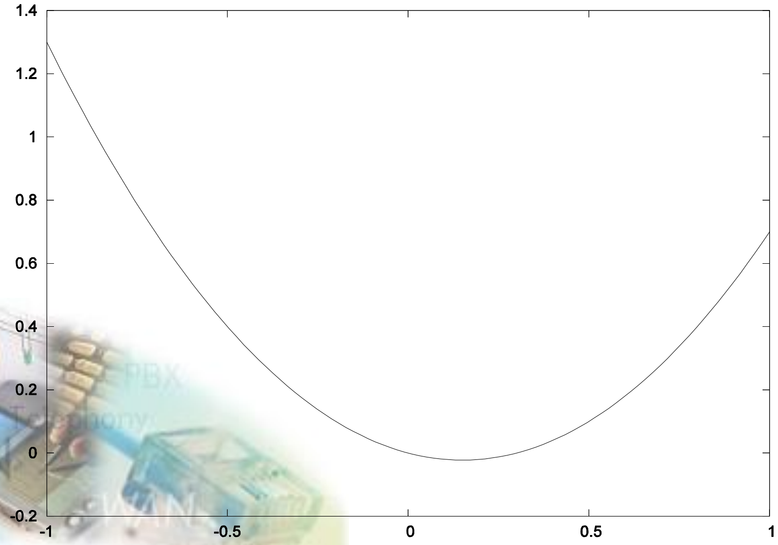
- ➢Positive semi-definite

$$\boldsymbol{A} \text{ is positive semi-definite}$$

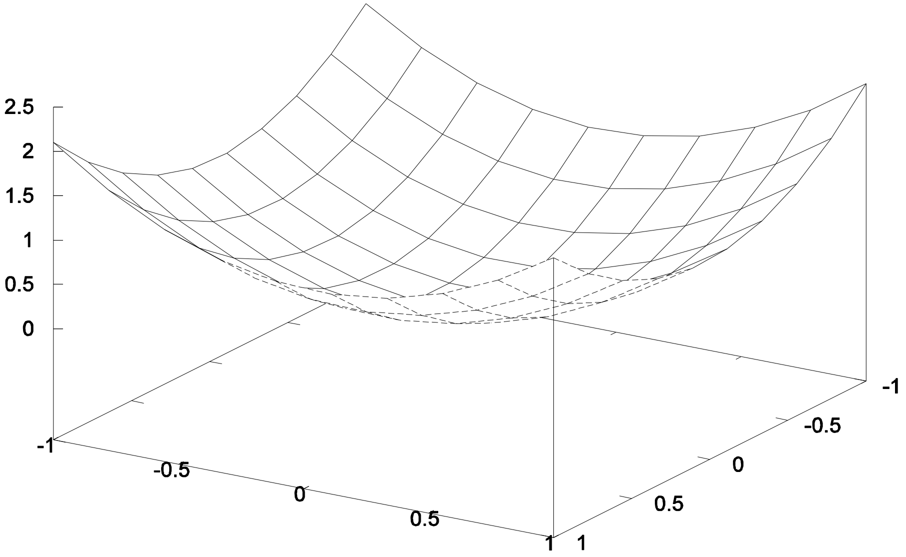
$$\Leftrightarrow$$

$$\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{0}, \text{ for all } \boldsymbol{x}$$

In 1-D



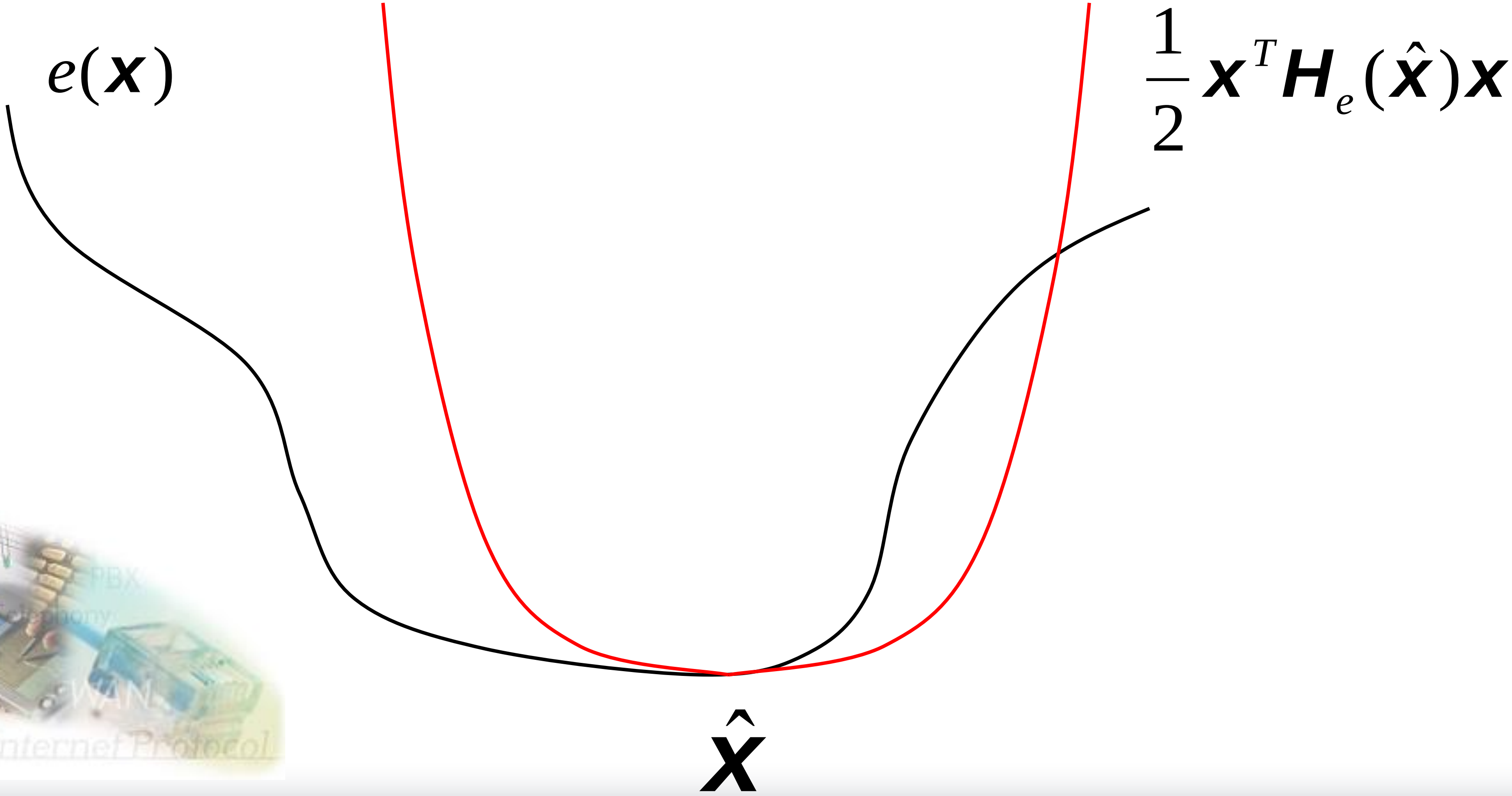
In 2-D



2.2 最小二乘估计Least-Squares

◆Multidimentional linear regression

- ➤Minimizing $e(\boldsymbol{x})$

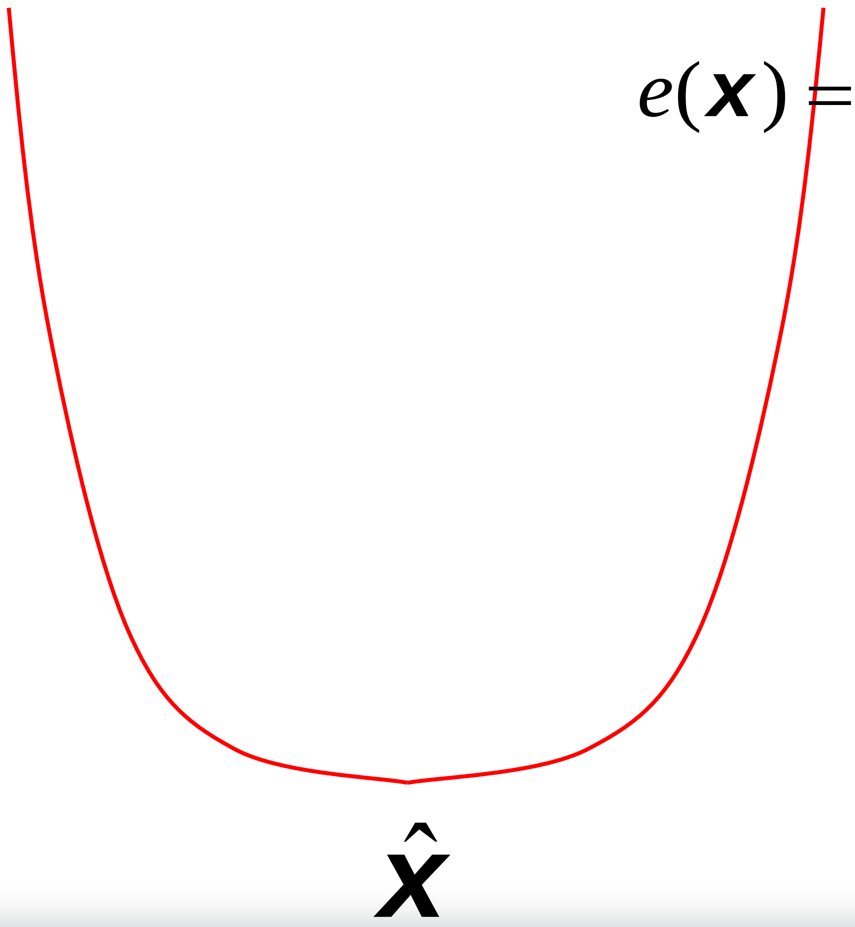

$e(\boldsymbol{x})$
$\frac{1}{2}\boldsymbol{x}^T\boldsymbol{H}_e(\hat{\boldsymbol{x}})\boldsymbol{x}$
$\hat{\boldsymbol{x}}$

2.2 最小二乘估计Least-Squares

◆Multidimentional linear regression

- Minimizing $e(\boldsymbol{x}) = \left\| \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \right\|^2$

$$e(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T \boldsymbol{H}_e(\hat{\boldsymbol{x}})\boldsymbol{x}$$



2.2 最小二乘估计Least-Squares

◆Multidimentional linear regression

- ➢ Minimizing $e(\boldsymbol{x}) = \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2$

$$\hat{\boldsymbol{x}} \text{ minimizes } e(\boldsymbol{x}) \text{ if } \begin{cases} \boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{A}^T\boldsymbol{b} \\ 2\boldsymbol{A}^T\boldsymbol{A} \text{ is positive semi-definite} \end{cases}$$

2.2 最小二乘估计Least-Squares

◆Multidimentional linear regression

- ➢Minimizing $e(\boldsymbol{x}) = \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2$

$\hat{\boldsymbol{x}}$ minimizes $e(\boldsymbol{x})$ if

$$\begin{cases} \boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{A}^T\boldsymbol{b} \\ 2\boldsymbol{A}^T\boldsymbol{A} \text{ is positive semi-definite} \end{cases}$$

Always true

2.2 最小二乘估计Least-Squares

◆Multidimentional linear regression

- Minimizing $e(\boldsymbol{x}) = \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|^2$

$\hat{\boldsymbol{x}}$ minimizes $e(\boldsymbol{x})$ if

$$\boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{A}^T\boldsymbol{b}$$

The *normal equation*

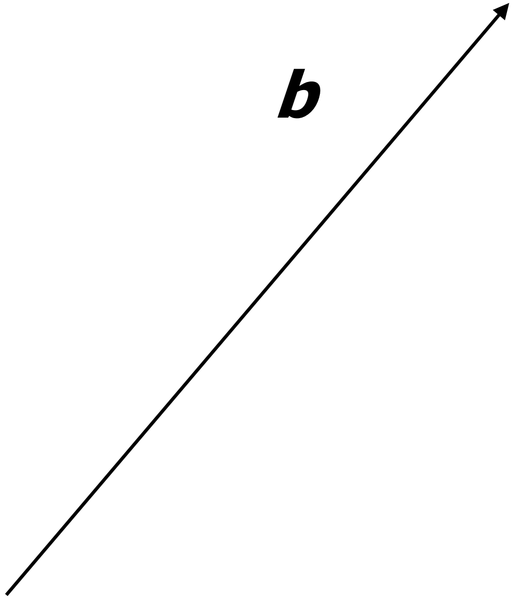
$2\boldsymbol{A}^T\boldsymbol{A}$ is positive semi-definite

Always true

2.2 最小二乘估计Least-Squares

◆Geometric interpretation

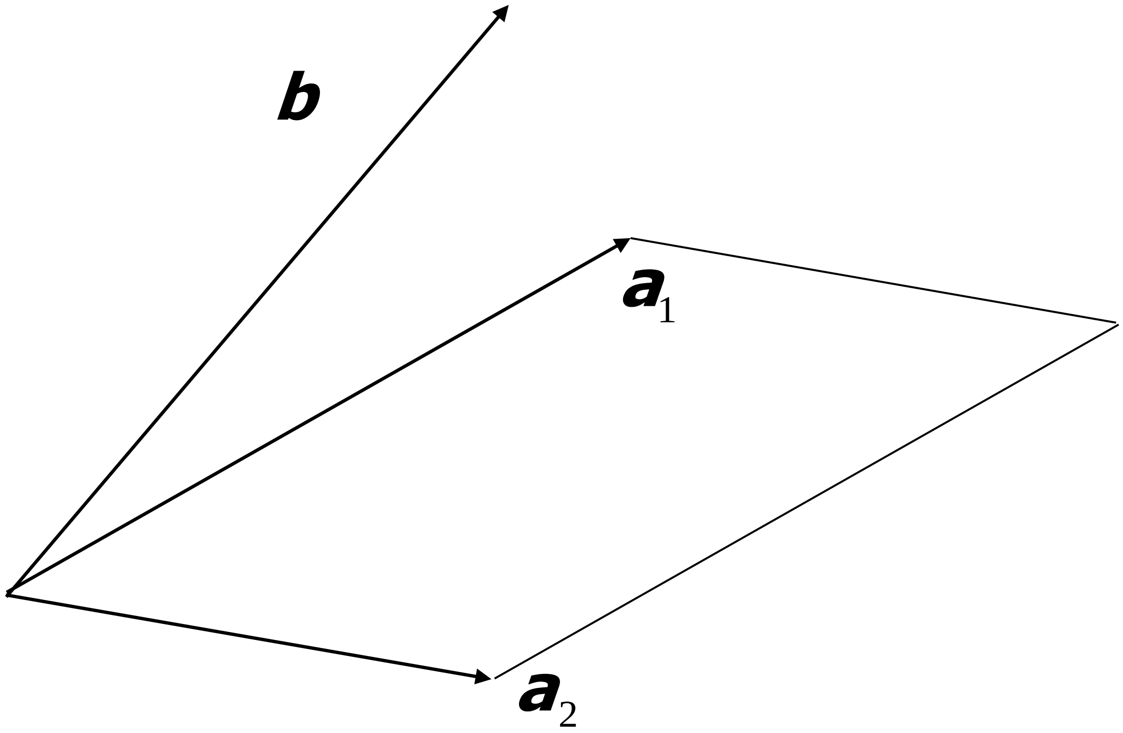
- ➢ $\boldsymbol{b}$ is a vector in R^n



2.2 最小二乘估计Least-Squares

◆Geometric interpretation

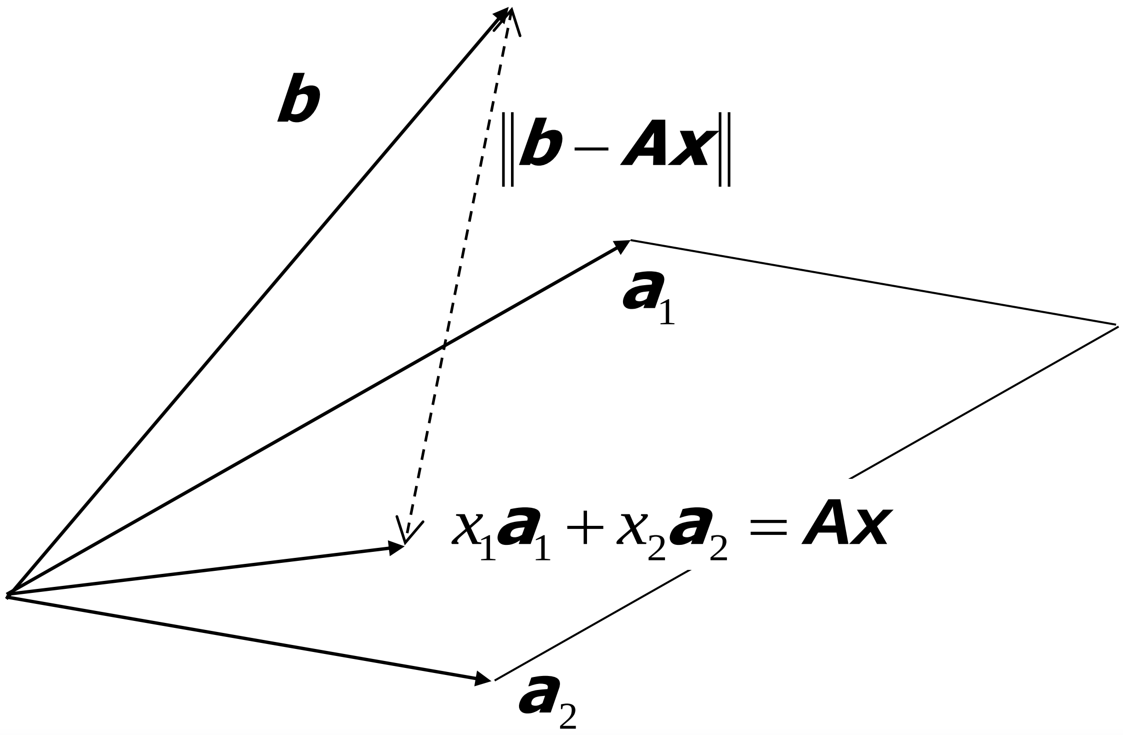
- ➢ ***b*** is a vector in $\boldsymbol{R}^n$
- ➢ The columns of A define a vector space range(A)



2.2 最小二乘估计Least-Squares

◆Geometric interpretation

- ***b*** is a vector in ***R****^n^*
- The columns of ***A*** define a vector space ***range(A)***
- ***Ax*** is an arbitrary vector in ***range(A)***

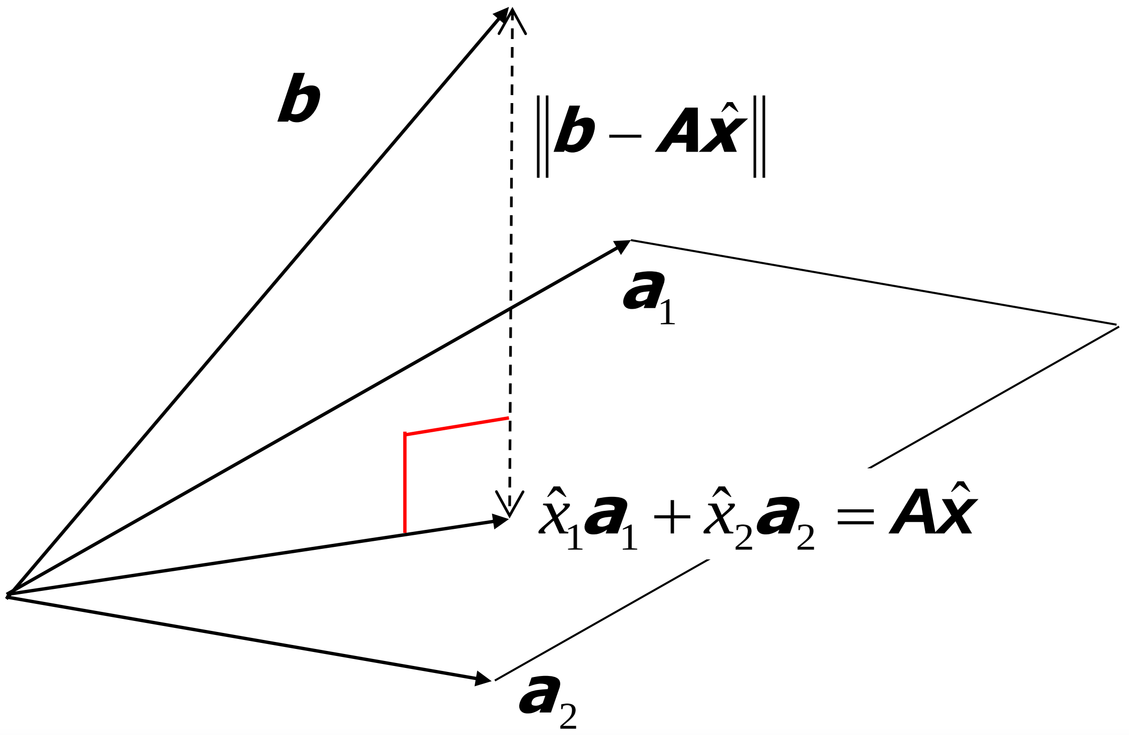

$\boldsymbol{b}$
$\|\boldsymbol{b}-\boldsymbol{Ax}\|$
$\boldsymbol{a}_1$
$x_1\boldsymbol{a}_1+x_2\boldsymbol{a}_2=\boldsymbol{Ax}$
$\boldsymbol{a}_2$

2.2 最小二乘估计Least-Squares

Geometric interpretation

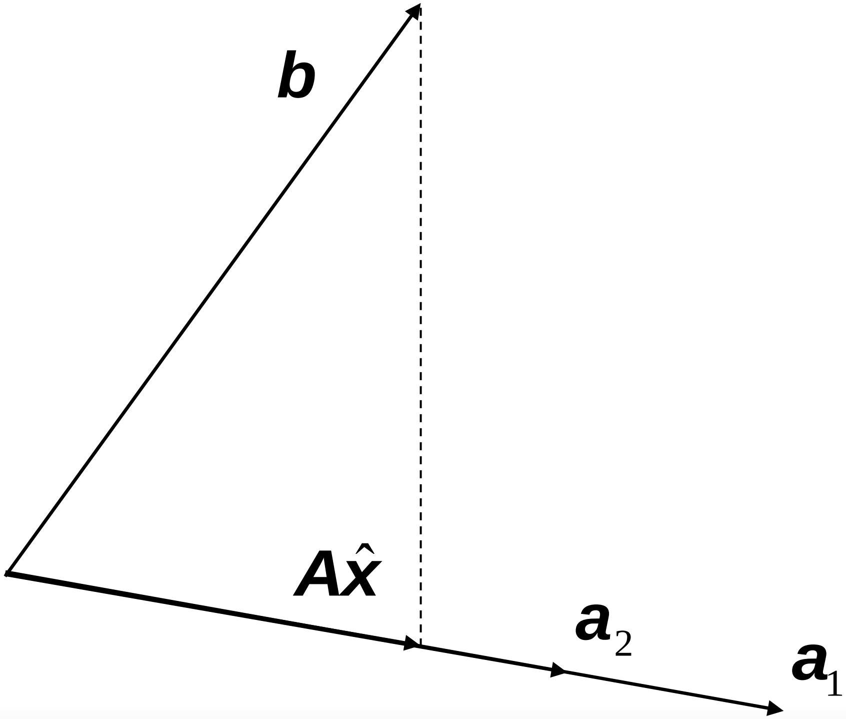
- $A\hat{x}$ is the orthogonal projection of *b* onto *range(A)*

$$\Leftrightarrow A^T\left(b - A\hat{x}\right) = 0 \Leftrightarrow A^T A\hat{x} = A^T b$$


b
$\left\| b - A\hat{x} \right\|$
a_1
$\hat{x}_1 a_1 + \hat{x}_2 a_2 = A\hat{x}$
a_2

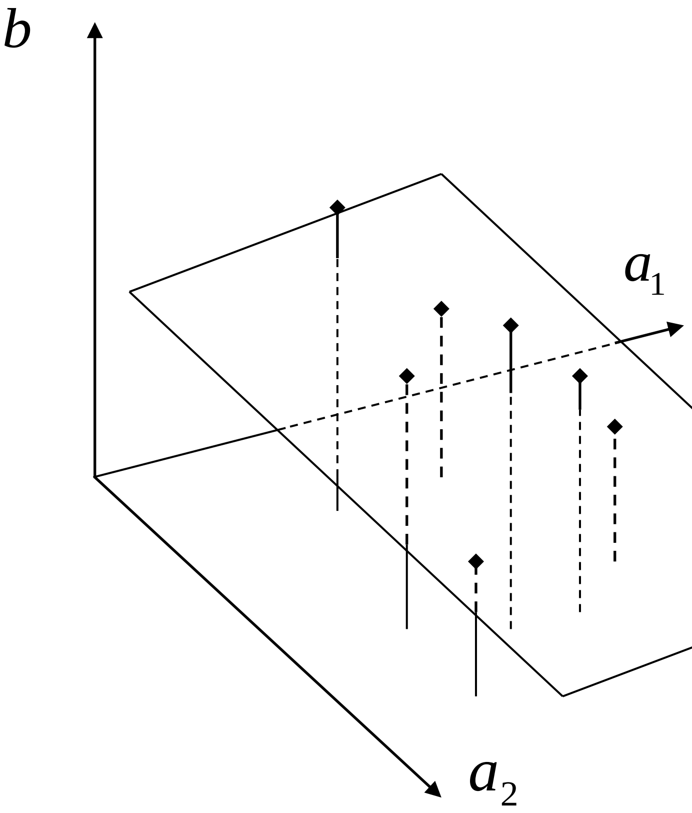
2.2 最小二乘估计Least-Squares

- **The normal equation:** $\boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{A}^T\boldsymbol{b}$
  - **Existence:** $\boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{A}^T\boldsymbol{b}$ **has always a solution**
  - **Uniqueness: the solution is unique if the columns of $\boldsymbol{A}$ are linearly independent**



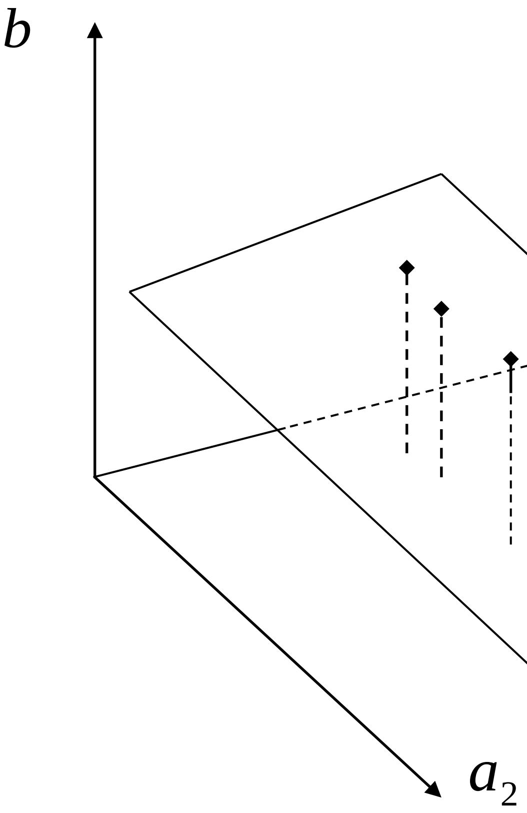
2.2 最小二乘估计Least-Squares

◆Under-constrained problem



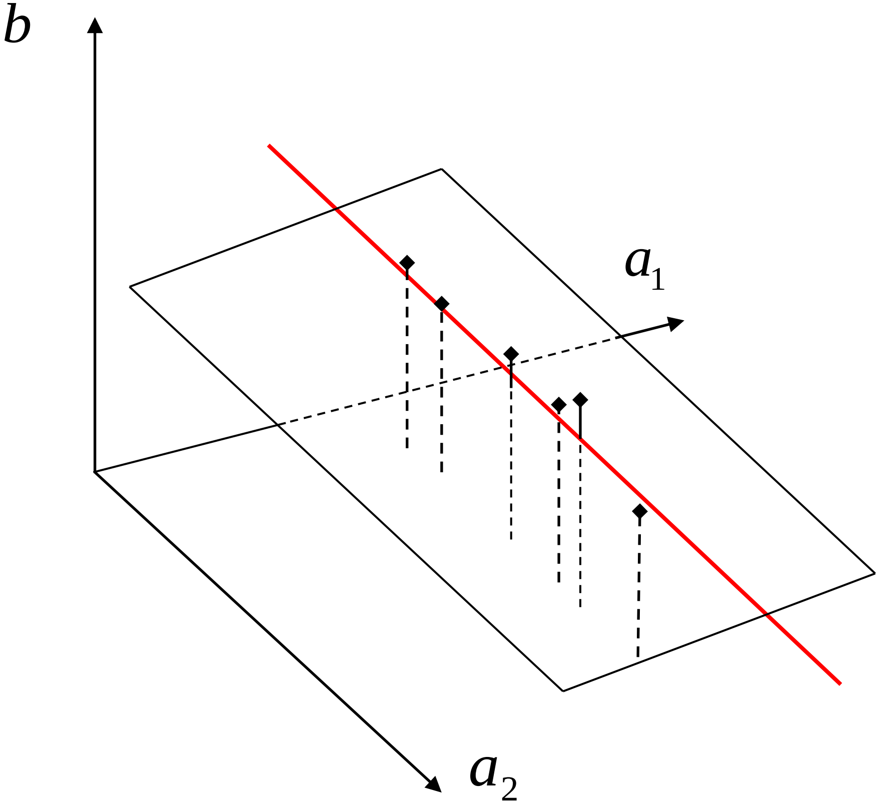
2.2 最小二乘估计Least-Squares

◆**Under-constrained problem**



2.2 最小二乘估计Least-Squares

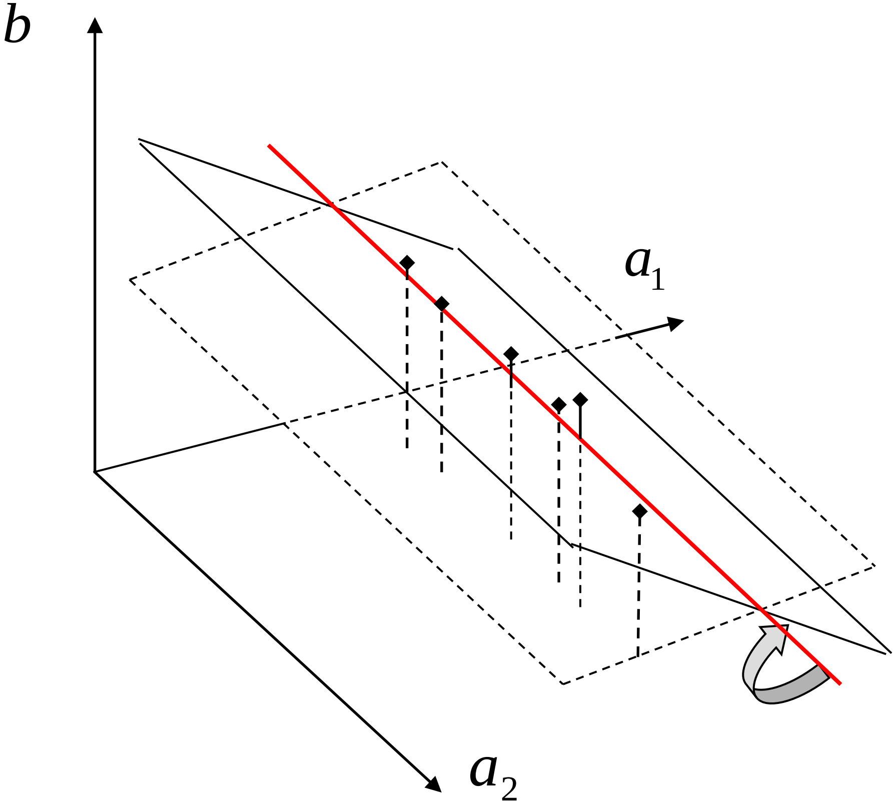
◆Under-constrained problem



2.2 最小二乘估计Least-Squares

◆Under-constrained problem

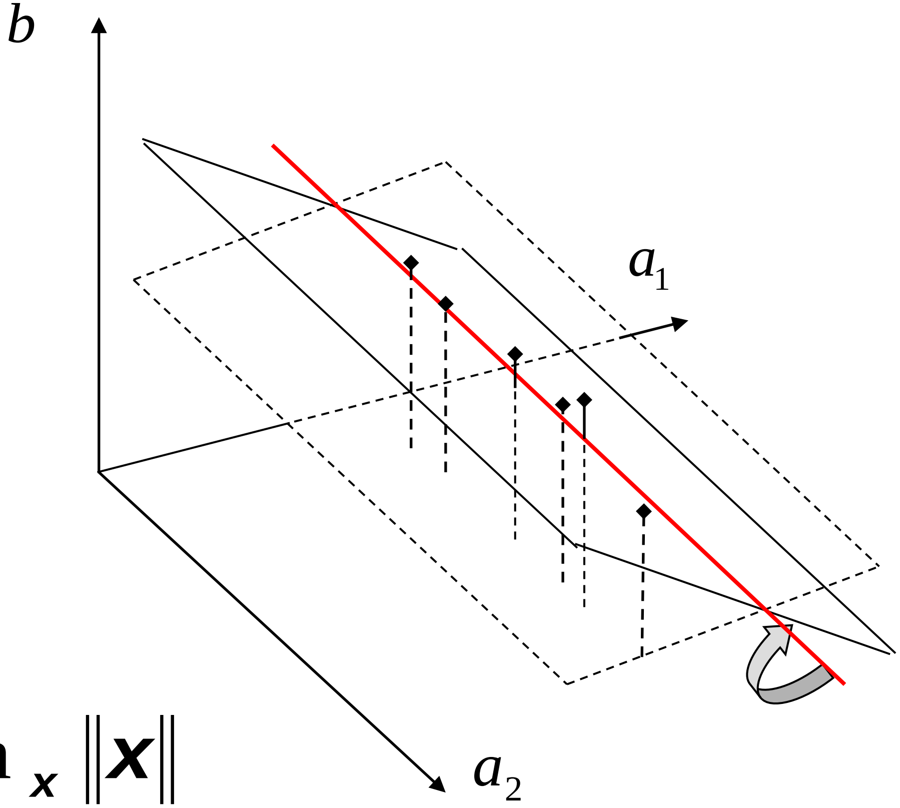
- ➢ Poorly selected data
- ➢ One or more of the parameters are redundant



2.2 最小二乘估计Least-Squares

◆ **Under-constrained problem**

- **Poorly selected data**
- **One or more of the parameters are redundant**



Add constraints

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b} \quad \text{with} \min_{\boldsymbol{x}} \|\boldsymbol{x}\|$$

2.2 最小二乘估计Least-Squares

◆How good is the least-squares criteria?

➢ **Optimality: the Gauss-Markov theorem**

Let $\{b_i\}$ **and** $\{x_j\}$ **be two sets of random variables and define:**

$$e_i = b_i - a_{i,1}x_1 - \ldots - a_{i,m}x_m$$

If

A1: $\{a_{i,j}\}$ are not random variables,

A2: $E(e_i) = 0$, for all i,

A3: $\mathrm{var}(e_i) = \sigma$, for all i,

A4: $\mathrm{cov}(e_i, e_j) = 0$, for all i and j,

Then $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \sum e_i^2$ **is the best unbiased linear estimator**

Maximum Likelihood Estimation

2.3 最大似然估计 MLE

◆ Parameter Estimation

- Bayesian Decision Theory allows us to design an optimal classifier given that we have estimated $P(\omega_j)$ and $p(x|\omega_j)$ first:

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

- Estimating $P(\omega_j)$ is usually not very difficult.
- Estimating $p(x|\omega_j)$ could be more difficult:
  - ✓ Dimensionality of feature space is large.
  - ✓ Number of samples is often too small.

2.3 最大似然估计 MLE

◆Parameter Estimation (cont’d)

- We will make the following assumptions:
- A set of training samples D ={x$_1$, x$_2$, ..., x$_n$} is given, where the samples were drawn according to $p(\mathrm{x}|\omega_j)$.
- $p(\mathrm{x}|\omega_j)$ has some known parametric form:

$$\text{e.g.,}\quad p(\mathbf{x}|\omega_j) \sim \mathrm{N}(\boldsymbol{\mu}_j, \Sigma_j)$$

also denoted as $p(\mathbf{x}|\theta)$ where $\boldsymbol{\theta}=(\boldsymbol{\mu}_j, \Sigma_j)$

◆Parameter estimation problem:

Given *D*, find the best possible θ

2.3 最大似然估计 MLE

◆Main Methods in Parameter Estimation

- ➢ **Maximum Likelihood Estimation (MLE)**

- ➢ **Bayesian Estimation (BE)**

2.3 最大似然估计 MLE

◆Maximum Likelihood Estimation (MLE)

- Best estimation $\hat{\boldsymbol{\theta}}$ is obtained by maximizing the probability of obtaining the samples D ={x_1, x_2, …., x_n} actually observed:

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n | \boldsymbol{\theta}) = p(D | \boldsymbol{\theta})$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(D | \boldsymbol{\theta})$$

- MLE assumes that θ is fixed and makes a point estimate:

$$p(\mathbf{x} | \theta) \to p(\mathbf{x} | \hat{\theta})$$

2.3 最大似然估计 MLE

◆ML Estimation - Assumptions

- Consider c classes and c training data sets (i.e., one for each class):

- Samples in D_j are drawn independently according to $p(\mathrm{x}|\omega_j)$.

$$D_1, D_2, \ldots, D_c$$

- Problem: given $D_1, D_2, \ldots, D_c$ and a model for $p(\mathrm{x}|\omega_j) \sim \mathrm{p}(\mathrm{x}|\theta)$, estimate:

$$\theta_1, \ \theta_2, \cdots, \ \theta_c$$

2.3 最大似然估计 MLE

◆ML Estimation - Problem Formulation

- If the samples in D_j provide no information about θ_i ($i \neq j$), we need to solve c independent problems (i.e., one for each class).

- The ML estimate for $D=\{x_1,x_2,..,x_n\}$ is the value $\hat{\boldsymbol{\theta}}$ that maximizes $p(D \mid \theta)$ (i.e., best supports the training data).

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} \; p(D|\boldsymbol{\theta})$$

- Using independence assumption, we can simplify $p(D \mid \theta)$:

$$p(D|\boldsymbol{\theta}) = p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta})$$

2.3 最大似然估计 MLE

◆ML Estimation - Solution

- ➢ How can we find the maximum of *p(D| θ)* ?

$$\nabla_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta}) = 0$$

where $$\nabla_{\boldsymbol{\theta}} \equiv \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\theta}_1} \\ \vdots \\ \frac{\partial}{\partial \boldsymbol{\theta}_p} \end{bmatrix}$$ (gradient)

2.3 最大似然估计 MLE

◆ML Estimation Using Log-Likelihood

- ➢ Taking the log for simplicity:

$$p(D|\boldsymbol{\theta}) = p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta})$$

$$\Downarrow$$

$$\ln p(D|\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln p(\mathbf{x}_k|\boldsymbol{\theta})$$

log-likelihood

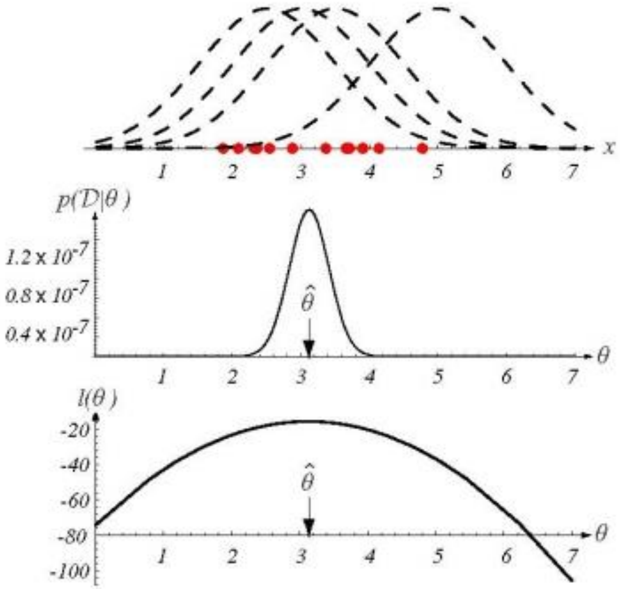
- ➢ Maximizes ln p(D| θ):

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} \ \ln p(D|\boldsymbol{\theta})$$

$$\nabla_{\boldsymbol{\theta}} \ln p(D|\boldsymbol{\theta}) = 0 \quad or \quad \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k|\boldsymbol{\theta}) = 0$$

2.3 最大似然估计 MLE

- **Example:training data: unknown mean, known variance**



p(D|θ)

ln p(D|θ)

$$\hat{\theta}=\mu$$

FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

2.3 最大似然估计 MLE

◆MLE for Multivariate Gaussian Density: Case of Unknown θ=μ

- Assume

$$p(\mathbf{x}/\boldsymbol{\mu}) \sim N(\boldsymbol{\mu},\Sigma)$$

$$\ln p(\mathbf{x}/\boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma|$$

- Computing the gradient, we have:

$$\nabla_{\boldsymbol{\mu}} \ln p(D\backslash\boldsymbol{\mu}) = \sum^{k} \nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k \backslash \boldsymbol{\mu}) = \sum^{k} \Sigma^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

2.3 最大似然估计 MLE

◆ML for Multivariate Gaussian Density: Case of Unknown θ=μ (cont’d)

- Setting $\nabla_{\boldsymbol{\mu}} \ln p(D|\boldsymbol{\mu}) = 0$ we have:

$$\sum_{k=1}^{n} \Sigma^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) = 0 \ or \ \sum_{k=1}^{n} \mathbf{x}_k - n\boldsymbol{\mu} = 0$$

- The solution $\hat{\boldsymbol{\mu}}$ is given by

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k$$

- The ML estimate is simply the “sample mean”.

2.3 最大似然估计 MLE

◆ML for Univariate Gaussian Density: Case of Unknown θ=(μ,σ^2)

- Assume $$p(x|\boldsymbol{\theta}) \sim N(\mu, \sigma^2)$$ $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$

$$\ln p(\mathrm{x}_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(\mathrm{x}_k - \mu)^2 \quad or$$

$$\ln p(\mathrm{x}_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(\mathrm{x}_k - \theta_1)^2$$

- Computing $\nabla_\theta \ln p(x_k/\theta)$ we have:

$$\frac{\partial \ln p(x_k/\theta)}{\theta_1} = \frac{1}{\theta_2}(x_k - \theta_1)$$

$$\frac{\partial \ln p(x_k/\theta)}{\theta_2} = -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2}$$

2.3 最大似然估计 MLE

◆ ML for Univariate Gaussian Density: Case of Unknown θ=(μ,σ²) (cont'd)

- Setting $\nabla_\theta \ln p(\mathbf{x}/\theta) = 0$ we have:

$$\sum_{k=1}^{n} \frac{1}{\theta_2}(x_k - \theta_1) = 0$$

$$-\sum_{k=1}^{n} \frac{1}{2\theta_2} + \sum_{k=1}^{n} \frac{(x_k - \theta_1)^2}{2\theta_2^2} = 0$$

➢ The solutions are given by:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$ sample mean

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$$ sample variance

2.3 最大似然估计 MLE

◆ML for Multivariate Gaussian Density: Case of Unknown θ=(μ,Σ)

- In the general case (i.e., multivariate Gaussian) the solutions are:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

sample mean

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})(x_k - \hat{\mu})^t$$

sample covariance

2.3 最大似然估计 MLE

◆Comments

- ➢ ML estimation is simpler than alternative methods.
- ➢ ML provides more accurate estimates as the number of training samples increases.
- ➢ If the model for $p(x|\ \theta)$ is correct, and the independence assumptions among samples are true, ML will work well.

2.4 贝叶斯参数估计 BE

◆Bayesian Estimation (BE)

- ➢ Assumes that θ is a set of random variables that have some known a-priori distribution p(θ).
- ➢ Estimates a distribution rather than making a point estimate (i.e., like ML):

$$p(\mathbf{x} \mid D) = \int p(\mathbf{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$$

Note: The BE solution p(**x**|D) might not be of the parametric form assumed (e.g., p(**x**|**θ**)).

2.5 最大后验估计 MAP

◆Special Case: Maximum A-Posteriori Estimator (MAP)

- Assume that θ is a random variable with known *p*(θ).

Consider: $$p(\boldsymbol{\theta} / D) = \frac{p(D / \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(D)}$$

- Maximize p(θ|D) or p(D|θ)p(θ) or *ln p(D| θ)p(θ)*:

$$\prod_{k=1}^{n} p(\mathbf{x}_k / \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \qquad \sum_{k=1}^{n} \ln p(\mathbf{x}_k / \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})$$

2.5 最大后验估计 MAP

◆Special Case: Maximum A-Posteriori Estimator (MAP) (cont’d)

- What happens when *p(*θ*)* is uniform?

$$\sum_{k=1}^{n} \ln p(\mathbf{x}_k / \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})$$

$$\Downarrow$$

$$\sum_{k=1}^{n} \ln p(\mathbf{x}_k / \boldsymbol{\theta})$$

MAP is equivalent to MLE

2.5 最大后验估计 MAP

◆MAP for Multivariate Gaussian Density: Case of Unknown θ=μ

- ➢ Assume $p(\mathbf{x}/\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \Sigma = Diag(\sigma_{\boldsymbol{\mu}}))$
- ➢ and $p(\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}_{\mathbf{0}}, \Sigma_{\mu} = Diag(\sigma_{\boldsymbol{\mu}_0}))$

(both $\boldsymbol{\mu}_{\mathbf{0}}$ and $\sigma_{\boldsymbol{\mu}_0}$ are known)

- ➢ Maximize *ln p(μ /D) = ln p(D/ μ)p(μ):*

$$\sum_{k=1}^{n} \ln p(\mathbf{x}_k / \boldsymbol{\mu}) + \ln p(\boldsymbol{\mu})$$

$$\nabla_{\boldsymbol{\mu}}(\sum_{k=1}^{n} \ln p(\mathbf{x}_k / \boldsymbol{\mu}) + \ln p(\boldsymbol{\mu})) = 0$$

2.5 最大后验估计 MAP

◆**MAP for Multivariate Gaussian Density: Case of Unknown θ=μ (cont’d)**

$$\sum_{k=1}^{n}\frac{1}{\sigma_{\boldsymbol{\mu}}^{2}}(\mathbf{x}_k-\boldsymbol{\mu})-\frac{1}{\sigma_{\boldsymbol{\mu}_0}^{2}}(\boldsymbol{\mu}-\boldsymbol{\mu}_{\mathbf{0}})=0 \qquad or \qquad \hat{\boldsymbol{\mu}}=\frac{\boldsymbol{\mu}_{\mathbf{0}}+\frac{\sigma_{\boldsymbol{\mu}_0}^{2}}{\sigma_{\boldsymbol{\mu}}^{2}}\sum_{k=1}^{n}\mathbf{x}_k}{1+\frac{\sigma_{\boldsymbol{\mu}_0}^{2}}{\sigma_{\boldsymbol{\mu}}^{2}}n}$$

- If $\frac{\sigma_{\boldsymbol{\mu}_0}^{2}}{\sigma_{\boldsymbol{\mu}}^{2}}>>1$, then $\hat{\boldsymbol{\mu}}=\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$

- What happens when $\sigma_{\boldsymbol{\mu}_0}\rightarrow 0$?

$$\hat{\boldsymbol{\mu}}=\boldsymbol{\mu}_{\mathbf{0}}$$

2.5 最大后验估计 MAP

◆Biased and Unbiased Estimates

- An estimate $\hat{\boldsymbol{\theta}}$ is unbiased when

$$E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$$

- The ML estimate $\hat{\boldsymbol{\mu}}$ is unbiased, i.e.,

$$E[\hat{\boldsymbol{\mu}}] = \boldsymbol{\mu}$$

- The ML estimates $\hat{\boldsymbol{\sigma}}$ and $\hat{\Sigma}$ are biased:

$$E[\hat{\boldsymbol{\sigma}}^2] = \frac{n-1}{n}\sigma^2 \qquad E[\hat{\Sigma}] = \frac{n-1}{n}\Sigma$$

2.5 最大后验估计 MAP

◆Biased and Unbiased Estimates (cont’d)

- The following are unbiased estimates of $\hat{\boldsymbol{\sigma}}$ and $\hat{\Sigma}$

$$\hat{\sigma} = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\mu})^2$$

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\mu})(\mathbf{x}_k - \hat{\mu})^t$$

Expectation-Maximization (EM)

2.6 期望最大化方法 EM

◆Description

- EM is an iterative ML estimation method:

- Starts with an initial estimate for θ.

- Refines the current estimate iteratively to increase the likelihood of the observed data:

$$p(D/ \theta)$$

2.6 期望最大化方法 EM

◆Description

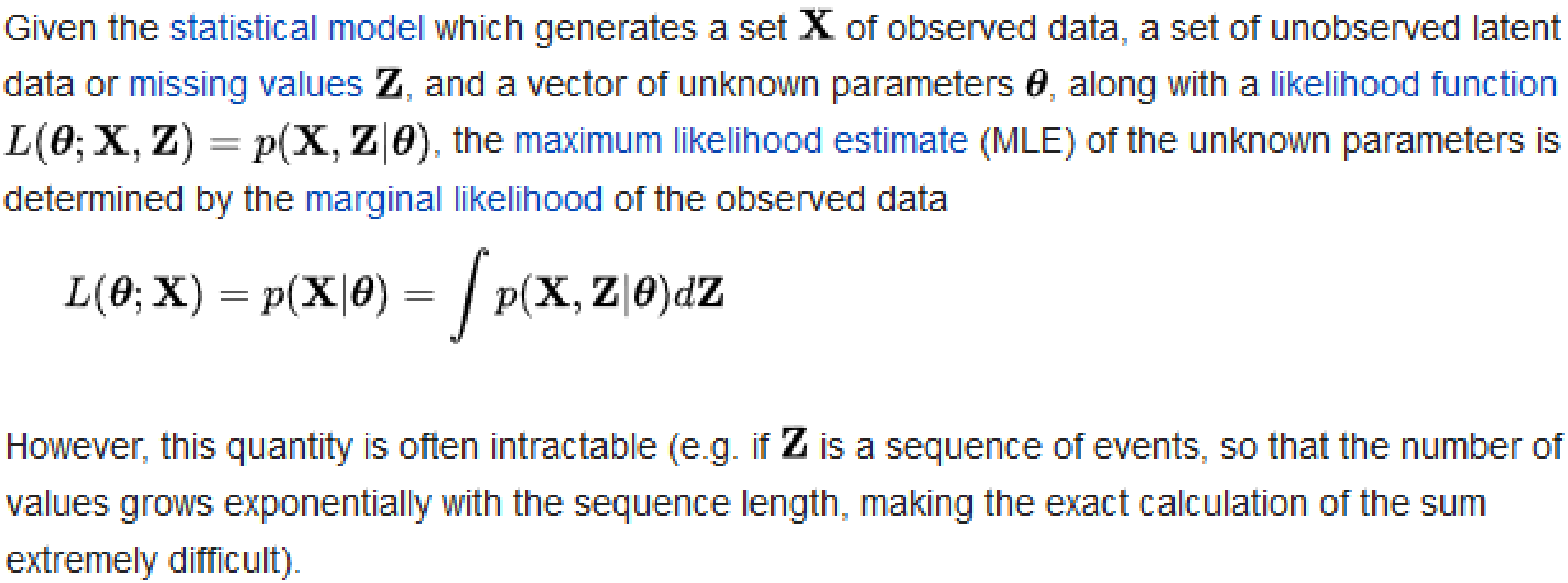
- ➢ EM represents a **general framework** – works best in situations where the data is incomplete (or can be thought as being incomplete)

- ➢ Some creativity is required to recognize where the EM algorithm can be used.

- ➢ Standard method for estimating the parameters of Mixtures of Gaussians (MoG).

2.6 期望最大化方法 EM

◆Description



Given the statistical model which generates a set $\mathbf{X}$ of observed data, a set of unobserved latent data or missing values $\mathbf{Z}$, and a vector of unknown parameters $\boldsymbol{\theta}$, along with a likelihood function $L(\boldsymbol{\theta};\mathbf{X},\mathbf{Z})=p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$, the maximum likelihood estimate (MLE) of the unknown parameters is determined by the marginal likelihood of the observed data

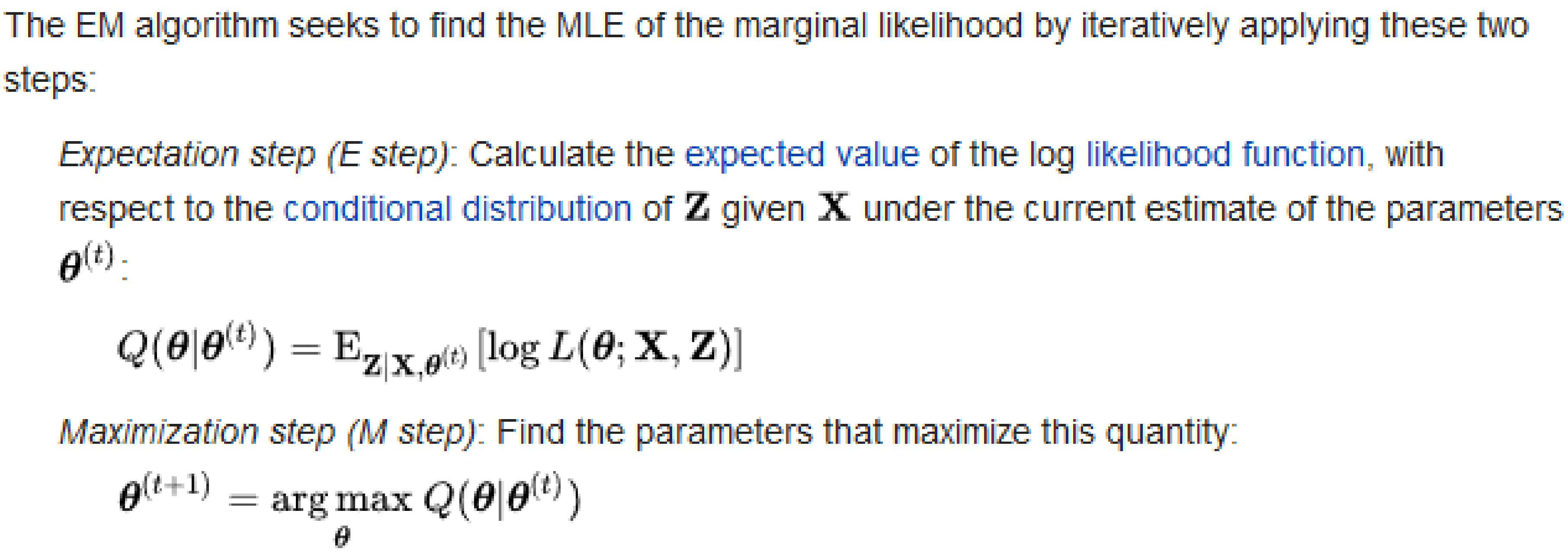
$$L(\boldsymbol{\theta};\mathbf{X})=p(\mathbf{X}|\boldsymbol{\theta})=\int p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})d\mathbf{Z}$$

However, this quantity is often intractable (e.g. if $\mathbf{Z}$ is a sequence of events, so that the number of values grows exponentially with the sequence length, making the exact calculation of the sum extremely difficult).

https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

2.6 期望最大化方法 EM

◆Description



The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying these two steps:

Expectation step (E step): Calculate the expected value of the log likelihood function, with respect to the conditional distribution of $\mathbf{Z}$ given $\mathbf{X}$ under the current estimate of the parameters $\boldsymbol{\theta}^{(t)}$:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \operatorname{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(t)}}\left[\log L(\boldsymbol{\theta};\mathbf{X},\mathbf{Z})\right]$$

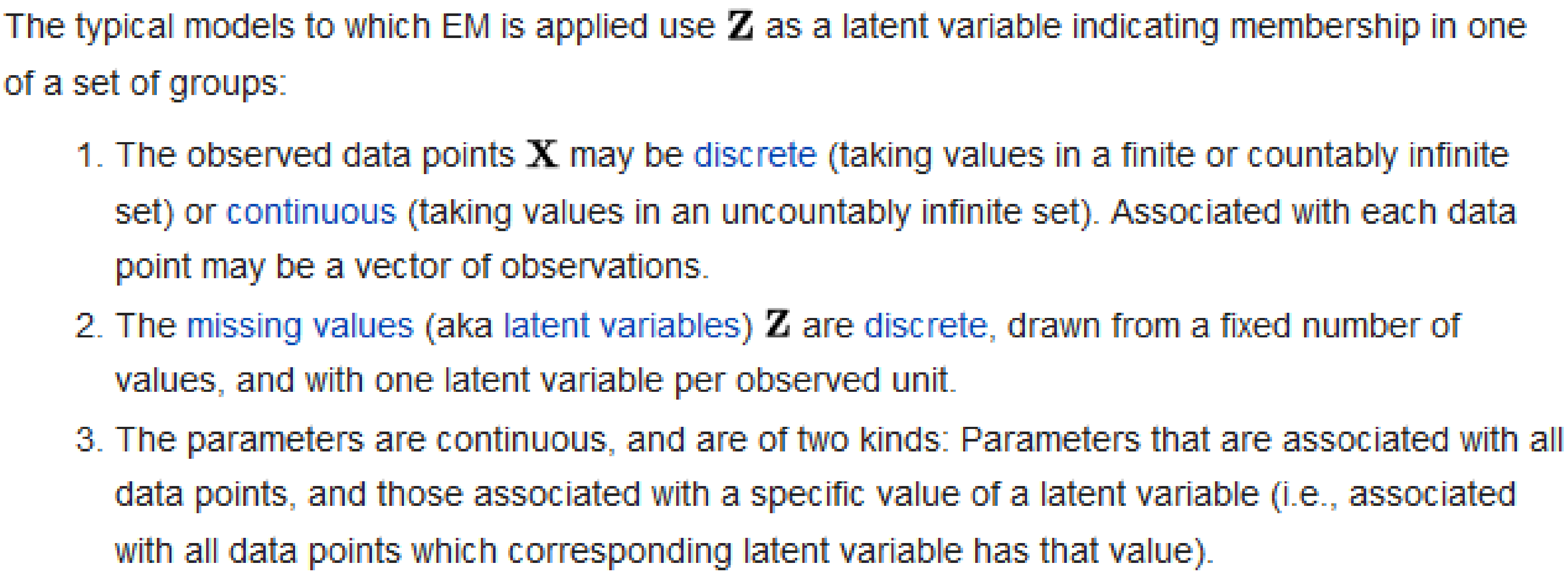
Maximization step (M step): Find the parameters that maximize this quantity:

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{arg\,max}}\ Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

2.6 期望最大化方法 EM

◆Description



The typical models to which EM is applied use $\mathbf{Z}$ as a latent variable indicating membership in one of a set of groups:

1. The observed data points $\mathbf{X}$ may be discrete (taking values in a finite or countably infinite set) or continuous (taking values in an uncountably infinite set). Associated with each data point may be a vector of observations.
2. The missing values (aka latent variables) $\mathbf{Z}$ are discrete, drawn from a fixed number of values, and with one latent variable per observed unit.
3. The parameters are continuous, and are of two kinds: Parameters that are associated with all data points, and those associated with a specific value of a latent variable (i.e., associated with all data points which corresponding latent variable has that value).

https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

2.6 期望最大化方法 EM

◆Incomplete Data

- Many times, it is impossible to apply ML estimation because certain features cannot be measured directly.

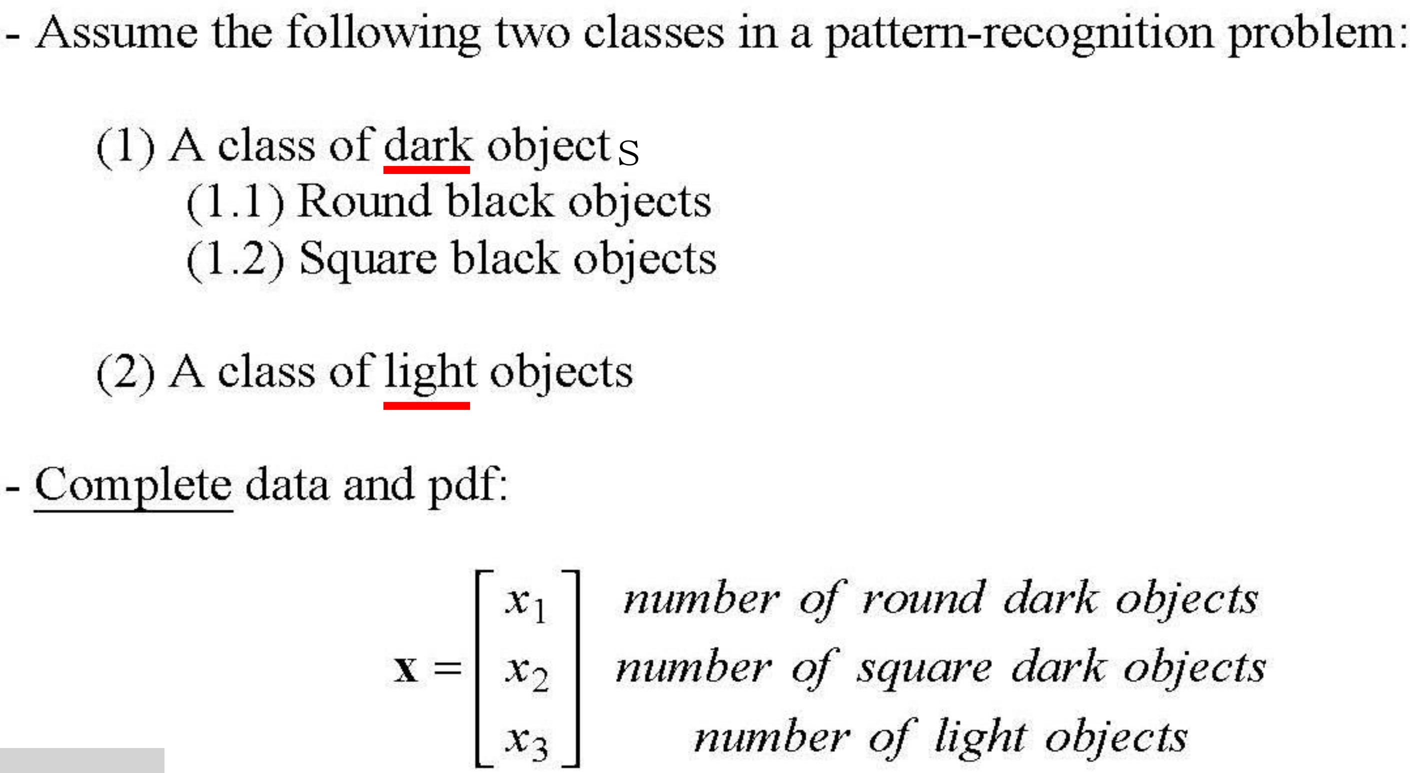
- The EM algorithm is ideal for problems with unobserved (missing) data.

$$\text{Actual data: } \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \text{Observed data: } \mathbf{y} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\underline{\text{Complete pdf:}}\ p(\mathbf{x}/\theta), \quad \underline{\text{Incomplete pdf:}}\ p(\mathbf{y}/\theta)$$

2.6 期望最大化方法 EM

◆Example (Moon, 1996)



- Assume the following two classes in a pattern-recognition problem:

(1) A class of dark objects

(1.1) Round black objects

(1.2) Square black objects

(2) A class of light objects

- Complete data and pdf:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \begin{array}{c} \textit{number of round dark objects} \\ \textit{number of square dark objects} \\ \textit{number of light objects} \end{array}$$

Assume a trinomial distribution:

$$p(x_1, x_2, x_3/\theta) = (\frac{k!}{x_1!x_2!x_3})(1/4)^{x_1}(1/4+\theta/4)^{x_2}(1/2-\theta/4)^{x_3}$$

!

$x_1+x_2+x_3=k$

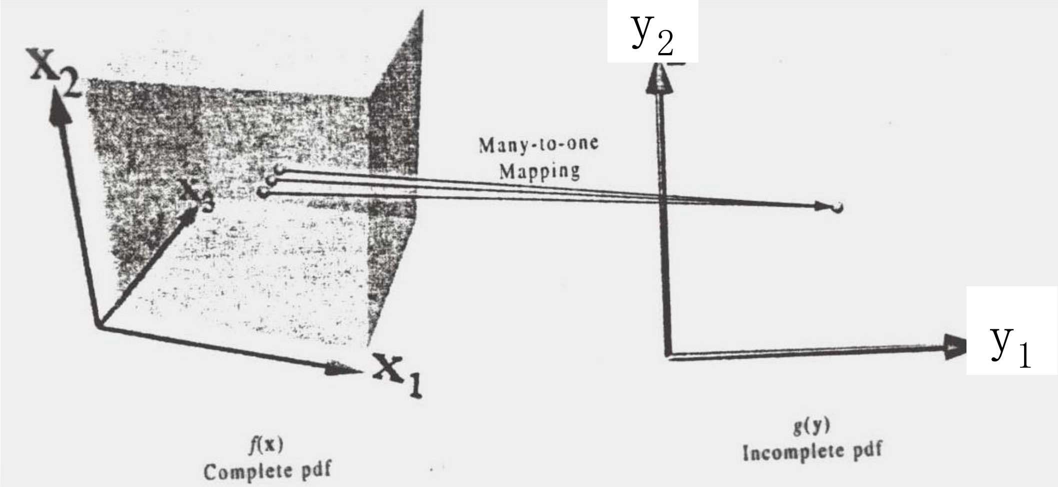
2.6 期望最大化方法 EM

◆Example (Moon, 1996) (cont’d)

- Observed (incomplete) data and pdf:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 \\ x_3 \end{bmatrix} \begin{matrix} \textit{number of dark objects} \\ \textit{number of light objects} \end{matrix}$$

(many-to-one mapping !!)



2.6 期望最大化方法 EM

◆EM: Main Idea

- If **x** was available, we could estimate θ using ML:

$$\hat{\theta} = \arg\max_{\theta} \ln p(D_x / \theta)$$

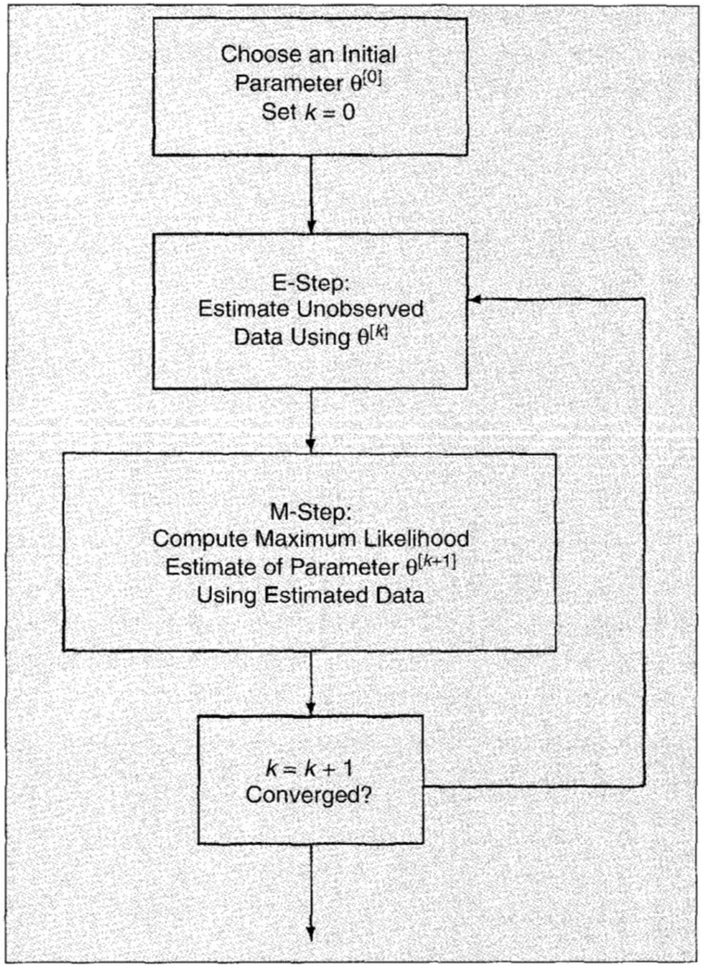
- Given that only **y** is available, estimate θ by:

Maximizing the expectation of ln *p(D*

2.6 期望最大化方法 EM

◆EM Steps

- (1) Initialization
- (2) E-Step: Expectation
- (3) M-Step: Maximization
- (4) Test for convergence



1. An overview of the EM algorithm. After initialization, the E-step and the M-step are alternated until the parameter estimate has converged (no more change in the estimate).

2.6 期望最大化方法 EM

◆EM Steps (cont'd)

- (1) Initialization Step: initialize the algorithm with a guess θ^0
- (2) Expectation Step: it is performed with respect to the unobserved variables, using the current estimate of parameters and conditioned upon the observations:

$$Q(\theta;\theta^t) = E_{x_{unobserved}}(\ln p(D_x/\theta)/D_y,\theta^t)$$

–Note: if ln $p(D_x \mid \theta)$ is a linear function of the unobserved variables, the expectation step is equivalent to:

$$Q(\theta;\theta^t) = E(x_{unobserved}/D_y,\theta^t)$$

2.6 期望最大化方法 EM

◆EM Steps (cont'd)

- (3) Maximization Step: provides a new estimate of the parameters:

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta;\theta^{t})$$

- (4) Test for Convergence:

$$\text{if} \quad |\theta^{t+1} - \theta^{t}| < \varepsilon$$

stop; otherwise, go to Step 2.

2.6 期望最大化方法 EM

◆Example (Moon, 1996) (cont’d)

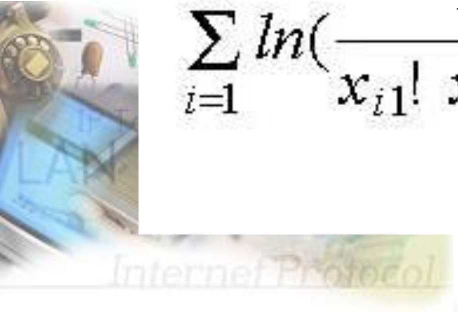
Expectation step: compute $E(ln\ p(D_x/\theta)\ /\ D_y, \theta^t))$

$$p(x_1, x_2, x_3/\theta) = \left(\frac{k!}{x_1!x_2!x_3!}\right)(1/4)^{x_1}(1/4+\theta/4)^{x_2}(1/2-\theta/4)^{x_3}$$

$$p(D_x/\theta) = \prod_{i=1}^{n} p(\mathbf{x}_i/\theta) \quad ==> \quad \ln\ p(D_x/\theta) = \sum_{i=1}^{n} \ln\ p(\mathbf{x}_i/\theta) =$$

$$\sum_{i=1}^{n} ln\left(\frac{k!}{x_{i1}!\ x_{i2}!\, x_{i3}!}\right) + x_{i1}\ ln(1/4) + x_{i2}\ ln(1/4+\theta/4) + x_{i3}\ ln(1/2-\theta/4)$$

where $\mathbf{x}_i=(x_{i1},x_{i2},x_{i3})$



2.6 期望最大化方法 EM

◆Example (Moon, 1996) (cont’d)

Expectation step: compute $E(ln\ p(D_x/\theta)\ /\ D_y, \theta^t))$

- Take the expected value:

$$E[ln\ p(D_x/\theta)/D_y, \theta^t] = \sum_{i=1}^{n} E[ln(\frac{k!}{x_{i1}!\ x_{i2}! x_{i3}!})/D_y, \theta^t] + E[x_{i1}/D_y, \theta^t]\ ln(1/4) +$$

$$E[x_{i2}/D_y, \theta^t]\ ln(1/4 + \theta/4) + x_{i3} ln(1/2 - \theta/4)$$

Let’s look at the M-step before completing the E-step …

2.6 期望最大化方法 EM

◆Example (Moon, 1996) (cont’d)

Maximization step: compute θ^{t+1} by maximizing $E(ln\ p(D_x/\theta)\ /\ D_y, \theta^t)$

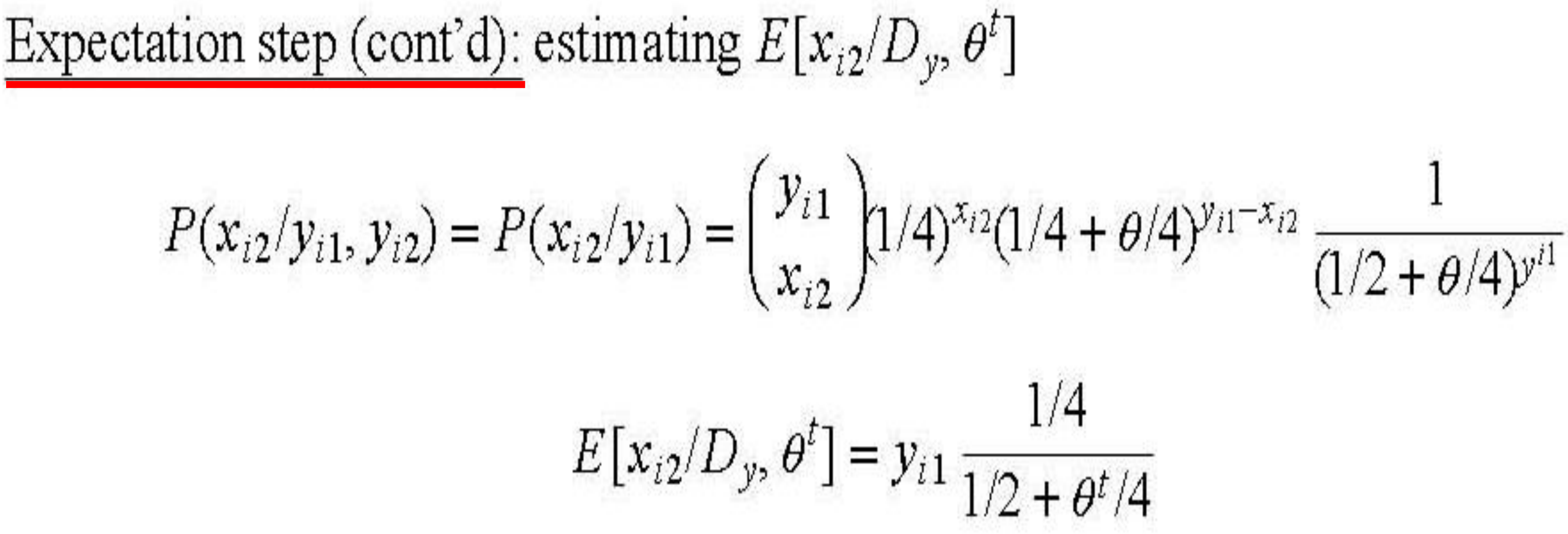
$$\frac{d}{d\theta}E[ln\ p(D_x/\theta)/D_y, \theta^t] = 0 ==> \theta^{t+1} = \frac{2\Sigma_i\ [x_{i2}/D_y, \theta^t] - x_{i3}}{\Sigma_i\ [x_{i2}/D_y, \theta^t] + x_{i3}}$$

- We only need to estimate: $E[x_{i2}/D_y, \theta^t]$

Let’s complete the E-step now …

2.6 期望最大化方法 EM

◆Example (Moon, 1996) (cont’d)



Expectation step (cont’d): estimating $E[x_{i2}/D_y, \theta^t]$

$$P(x_{i2}/y_{i1}, y_{i2}) = P(x_{i2}/y_{i1}) = \binom{y_{i1}}{x_{i2}}(1/4)^{x_{i2}}(1/4+\theta/4)^{y_{i1}-x_{i2}}\frac{1}{(1/2+\theta/4)^{y_{i1}}}$$

$$E[x_{i2}/D_y, \theta^t] = y_{i1}\frac{1/4}{1/2+\theta^t/4}$$

(see Moon’s paper, page 53)

2.6 期望最大化方法 EM

◆Example (Moon, 1996) (cont'd)

- Initialization: $\boldsymbol{\theta}^0$
- Expectation Step:

$$E[x_{i2}/D_y, \theta^t] = y_{i1}\,\frac{1/4}{1/2+\theta^t/4}$$

- Maximization Step:

$$\theta^{t+1} = \frac{2\Sigma_i\ [x_{i2}/D_y, \theta^t] - x_{i3}}{\Sigma_i\ [x_{i2}/D_y, \theta^t] + x_{i3}}$$

- Convergence Step:

$$|\theta^{t+1} - \theta^t| < \varepsilon$$

2.6 期望最大化方法 EM

◆Example (Moon, 1996) (cont’d)

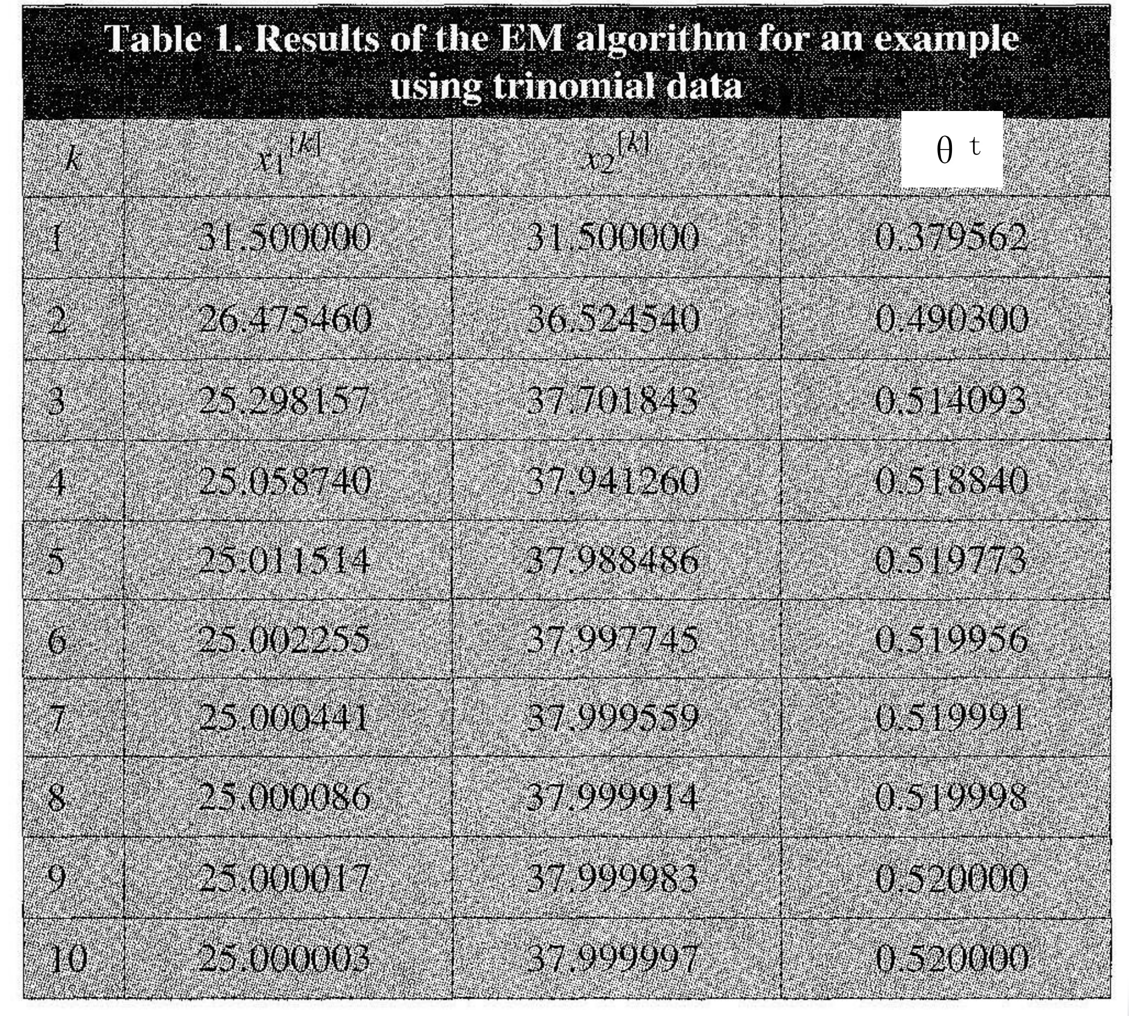


Table 1. Results of the EM algorithm for an example using trinomial data

k	$x_1^{[k]}$	$x_2^{[k]}$	θ^t
1	31.500000	31.500000	0.379562
2	26.475460	36.524540	0.490300
3	25.298157	37.701843	0.514093
4	25.058740	37.941260	0.518840
5	25.011514	37.988486	0.519773
6	25.002255	37.997745	0.519956
7	25.000441	37.999559	0.519991
8	25.000086	37.999914	0.519998
9	25.000017	37.999983	0.520000
10	25.000003	37.999997	0.520000

2.6 期望最大化方法 EM

◆Convergence properties of EM

- The solution depends on the initial estimate θ_0
- At each iteration, a value of θ is computed so that the likelihood function does not decrease.
- The algorithm is guaranteed to be **stable** (i.e., does not oscillate).
- There is **no guarantee** that it will convergence to a global maximum.

◆参考文献和站点

- Fundamentals of Statistical Signal Processing: Estimation Theory by Steven M. Kay (ISBN 0-13-345711-7)
- An Introduction to Signal Detection and Estimation by H. Vincent Poor (ISBN 0-38-794173-8)
- Detection, Estimation, and Modulation Theory, Part 1 by Harry L. Van Trees (ISBN 0-47-109517-6; website)
- Duda et al., Pattern Classification (Second Edition), Chapter 3
- T. K. Moon, The Expectation-Maximization Algorithm, IEEE Signal Processing Magazine, pp. 47-60, November 1996
- https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

END

第2章 参数估计