Parameter Estimation (DHS Ch. 3)

Not all statistics known (remember Case 2 & 3)

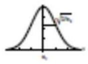
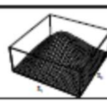
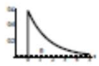
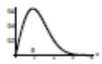
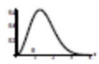
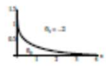
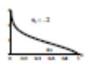
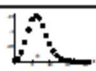
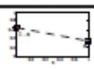
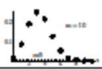
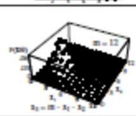
Two techniques for estimating $p(x|S_i)$, assumed not known a priori

1. **Parametric** – Functional form of $p(x|S_i)$ is known or assumed. Estimate parameters. (DHS Ch. 3 & Review Table 3.1 on the next page)
   Example: $p(x|S_i)=N(x,m_i,\Sigma_i)$
   Estimate m_i and Σ_i from training samples
   Two approaches: maximum likelihood – (1) ML estimation and (2) Maximum a Posterior (MAP) estimation (i.e., Bayesian estimation)
2. **Nonparametric**: estimate the density functions themselves. (DHS Ch. 4)

Outline

- Introduction
- Properties of an estimate
- Ad Hoc estimates
- Maximum Likehihood (ML) Estimate
- ML examples
- Estimation of random parameters
- Minimum mean square error (MMSE) estimate
- Maximum a Posteriori (MAP) estimate
- Estimation summary
- Bayes classifier
- Learning the mean of a normal density
  - Multivariate Extension
- General Bayesian learning
  - Recursive Bayesian learning
- Practical problems – Dimensionality
- Component Analysis and Fisher Linear Discriminant (Later)
- Markov Models (?)
- Hidden Markov Models (?)

Table 3.1: Common Exponential Distributions and their Sufficient Statistics.

Name	Distribution	Domain		$\mathbf{s}$	$[g(\mathbf{s},\boldsymbol{\theta})]^{1/n}$
Normal	$p(x\|\boldsymbol{\theta}) = \sqrt{\frac{\theta_2}{2\pi}} e^{-(1/2)\theta_2(x-\theta_1)^2}$	$\theta_2 > 0$		$\begin{bmatrix} \frac{1}{n}\sum_{k=1}^n x_k \\ \frac{1}{n}\sum_{k=1}^n x_k^2 \end{bmatrix}$	$\sqrt{\theta_2} e^{-\frac{1}{2}\theta_2(s_2 - 2\theta_1 s_1 + \theta_1^2)}$
Multi-variate Normal	$p(\mathbf{x}\|\boldsymbol{\theta}) = \frac{\|\boldsymbol{\Theta}_2\|^{1/2}}{(2\pi)^{d/2}} e^{-(1/2)(\mathbf{x}-\boldsymbol{\theta}_1)^t \boldsymbol{\Theta}_2 (\mathbf{x}-\boldsymbol{\theta}_1)}$	$\boldsymbol{\Theta}_2$ positive definite		$\begin{bmatrix} \frac{1}{n}\sum_{k=1}^n \mathbf{x}_k \\ \frac{1}{n}\sum_{k=1}^n \mathbf{x}_k \mathbf{x}_k^t \end{bmatrix}$	$\|\boldsymbol{\Theta}_2\|^{1/2} e^{-\frac{1}{2}[\mathrm{tr}\boldsymbol{\Theta}_2 \mathbf{s}_2 - 2\boldsymbol{\theta}_1^t \boldsymbol{\Theta}_2 \mathbf{s}_1 + \boldsymbol{\theta}_1^t \boldsymbol{\Theta}_2 \boldsymbol{\theta}_1]}$
Exponential	$p(x\|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n}\sum_{k=1}^n x_k$	$\theta e^{-\theta s}$
Rayleigh	$p(x\|\theta) = \begin{cases} 2\theta x e^{-\theta x^2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n}\sum_{k=1}^n x_k^2$	$\theta e^{-\theta s}$
Maxwell	$p(x\|\theta) = \begin{cases} \frac{4}{\sqrt{\pi}} \theta^{3/2} x^2 e^{-\theta x^2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n}\sum_{k=1}^n x_k^2$	$\theta^{3/2} e^{-\theta s}$
Gamma	$p(x\|\boldsymbol{\theta}) = \begin{cases} \frac{\theta_2^{\theta_1+1}}{\Gamma(\theta_1+1)} x^{\theta_1} e^{-\theta_2 x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta_1 > -1$, $\theta_2 > 0$		$\begin{bmatrix} \left(\prod_{k=1}^n x_k\right)^{1/n} \\ \frac{1}{n}\sum_{k=1}^n x_k \end{bmatrix}$	$\frac{\theta_2^{\theta_1+1}}{\Gamma(\theta_1+1)} s_1^{\theta_1} e^{-\theta_2 s_2}$
Beta	$p(x\|\boldsymbol{\theta}) = \begin{cases} \frac{\Gamma(\theta_1+\theta_2+2)}{\Gamma(\theta_1+1)\Gamma(\theta_2+1)} x^{\theta_1}(1-x)^{\theta_2} & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$	$\theta_1 > -1$, $\theta_2 > -1$		$\begin{bmatrix} \left(\prod_{k=1}^n x_k\right)^{1/n} \\ \left(\prod_{k=1}^n (1-x_k)\right)^{1/n} \end{bmatrix}$	$\frac{\Gamma(\theta_1+\theta_2+2)}{\Gamma(\theta_1+1)\Gamma(\theta_2+1)} s_1^{\theta_1} s_2^{\theta_2}$
Poisson	$P(x\|\theta) = \frac{\theta^x}{x!} e^{-\theta} \quad x = 0, 1, 2, ...$	$\theta > 0$		$\frac{1}{n}\sum_{k=1}^n x_k$	$\theta^s e^{-\theta}$
Bernoulli	$P(x\|\theta) = \theta^x (1-\theta)^{1-x} \quad x = 0, 1$	$0 < \theta < 1$		$\frac{1}{n}\sum_{k=1}^n x_k$	$\theta^s (1-\theta)^{1-s}$
Binomial	$P(x\|\theta) = \frac{m!}{x!(m-x)!} \theta^x (1-\theta)^{m-x} \quad x = 0, 1, ..., m$	$0 < \theta < 1$		$\frac{1}{n}\sum_{k=1}^n x_k$	$\theta^s (1-\theta)^{m-s}$
Multinomial	$P(\mathbf{x}\|\boldsymbol{\theta}) = \frac{m! \prod_{i=1}^d \theta_i^{x_i}}{\prod_{i=1}^d x_i!} \quad x_i = 0, 1, ..., m, \quad \sum_{i=1}^d x_i = m$	$0 < \theta_i < 1$, $\sum_{i=1}^d \theta_i = 1$		$\frac{1}{n}\sum_{k=1}^n \mathbf{x}_k$	$\prod_{i=1}^d \theta_i^{s_i}$

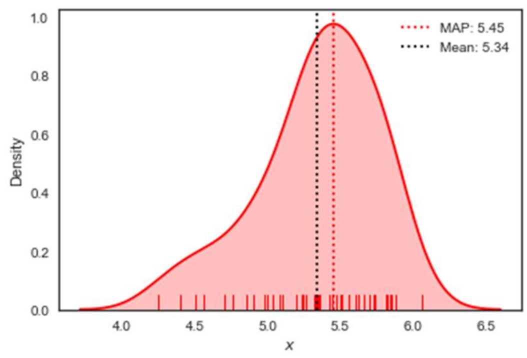
Preliminaries (DHS 3.1)

- Previously, we designed an optimal classifier if the prior probabilities and the class-conditional densities are known.
- However, we rarely have this kind of complete knowledge about the probabilistic structure of the problems
- Now, use the samples to estimate the unknown probabilities and probability densities
- Estimation of the prior probabilities in supervised classification is not a serious problems, but not the class-conditional densities
- At least, assume probability density functions with unknown parameters
- Two common and reasonable procedures: maximum-likelihood (ML) estimation and Bayes estimation
- ML views the parameters as quantities as fixed values, but unknown (Fig. 3.1)
- Bayesian views the parameters as random variables (Fig. 3.2)
- Bayesian learning: observing additional samples sharpens the posteriori densities, causing it to peak near the true values of the parameters (Fig. 3.2)

Maximum-Likelihood vs. Bayesian Maximum A Posteriori (DHS 3.2.1)

Key concepts

- IID = independent and identically distributed random variables
- Likelihood = $p(D|\theta)$ (See Fig. 3.1)
- Log-Likelihood $l(\theta)= \ln\{p(D|\theta)\}$ (See Fig. 3.1)
- Maximum A Posterior and Mode (See right below & Fig. 3.2))
- Fig. 3.1 vs. Fig. 3.2



MAP estimation

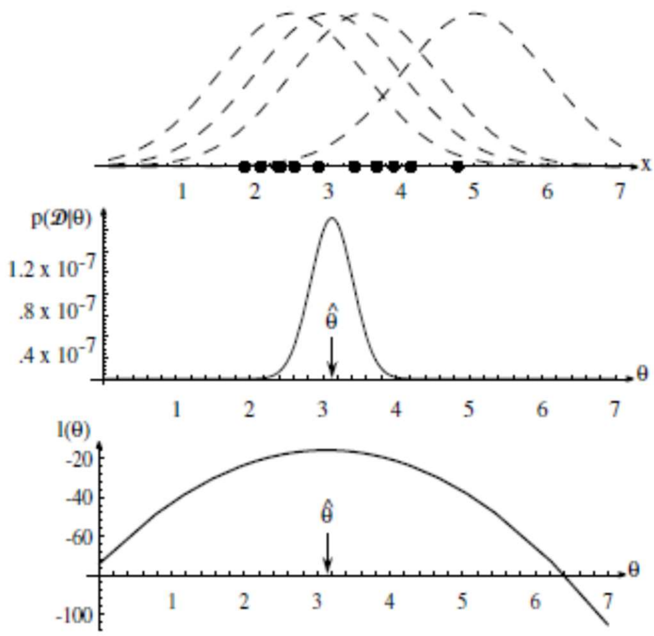


Figure 3.1: The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figures shows the likelihood $p(\mathcal{D}|\theta)$ 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 — i.e., the log-likelihood $l(\theta)$, shown at the bottom. Note especially that the likelihood lies in a different space from $p(x|\hat{\theta})$, and the two can have different functional forms.

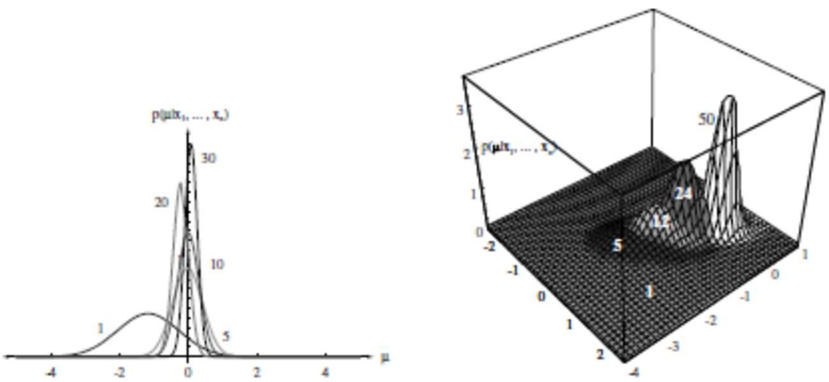


Figure 3.2: Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labelled by the number of training samples used in the estimation.

Introduction

Let $\underline{\theta}$ be a vector of fixed but unknown parameters.

Let $\underline{\hat{\theta}}$ be an estimate of $\underline{\theta}$.

$\underline{\theta}$ is deterministic. $\underline{\hat{\theta}}$ is random

Let $x_1, x_2, \ldots$ be random sample vectors drawn from the density to be estimated.
The x_i's are assumed independent and (usually) identically distributed.

Let $\underline{z}=[x_1, x_2, \ldots, x_J]$

The estimate $\underline{\hat{\theta}}$: $\underline{\hat{\theta}}=G(x_1, x_2, \ldots, x_J)=G(\underline{z})=\underline{\hat{\theta}}(z)$ => $\underline{\hat{\theta}}$=random.

Properties of an estimate $\hat{\theta}$

Unbiased estimate (most important)

If $E\{\underline{\hat{\theta}}\}=\int G(\underline{z})p(\underline{z})d\underline{z}=\underline{\theta}$ then $\underline{\hat{\theta}}$ is an unbiased estimate of $\underline{\theta}$. Otherwise $\underline{\hat{\theta}}$ is biased.

Consistent estimate

$P\lim \breve{\theta} = \theta^*$ Probability limit of $\breve{\theta}$ => $\lim_{k\to\infty} P\{|\breve{\theta}-\theta^*| \geq \varepsilon\} \to 0$

$\breve{\theta}$ is a consistent estimate of θ if $P\lim \breve{\theta} = \theta$

Efficient estimate

Unbiased and have the smallest possible error variance.

Var($\underline{\hat{\theta}}$)≤ Var($\underline{\hat{\hat{\theta}}}$), then $\underline{\hat{\theta}}$ is more efficient than $\underline{\hat{\hat{\theta}}}$

Sufficient estimate

An estimate is called sufficient for $\underline{\theta}$ if it contains all information about $\underline{\theta}$ which is contained in $\underline{z}$.

$\underline{\hat{\theta}}_1(z)$ is a sufficient estimate iff for any other estimates $\underline{\hat{\theta}}_2(z), \ldots, \underline{\hat{\theta}}_N(z)$,

the conditional density function of $\underline{\hat{\theta}}_2(z), \ldots, \underline{\hat{\theta}}_N(z)$ given $\hat{\theta}_1(z)$ does not depend on θ.

$p(\underline{\hat{\theta}}_2, \underline{\hat{\theta}}_3, \ldots, \underline{\hat{\theta}}_N | \hat{\theta}_1, \theta) = f(\underline{\hat{\theta}}_1, \underline{\hat{\theta}}_2, \ldots, \underline{\hat{\theta}}_N)$

∴ The best estimate would be: unbiased, consistent, efficient, and sufficient.

Ad Hoc estimates

Moment estimates:

Sample mean vector

$$\underline{\hat{m}} = 1/J \sum_{j=1}^{J} x_j$$

$E\{\underline{\hat{m}}\} = \underline{m}$ => unbiased.

Sample is unbiased.

Sample correlation estimate (no mean removed)

$$\underline{\hat{S}} = 1/J \sum_{j=1}^{J} \underline{x}_j \underline{x}_j^T$$

This is unbiased, consistent.

Sample covariance estimate (mean removed)

$$\underline{\hat{\Sigma}} = 1/J \sum_{j=1}^{J} (\underline{x}_j - \underline{\hat{m}})(\underline{x}_j - \underline{\hat{m}})^T$$

Is $\hat{\Sigma}$ unbiased? It is a biased estimate.

An unbiased estimate can be obtained:

$$\underline{\hat{\Sigma}} = 1/(J-1) \sum_{j=1}^{J} (\underline{x}_j - \underline{\hat{m}})(\underline{x}_j - \underline{\hat{m}})^T$$

If $\underline{x}_j$ are normal, $\underline{\hat{m}}$ is normal.

If $\underline{x}_j$ are arbitrary, $\underline{\hat{m}}$ tends to normal by the central limit theorem.

Maximum Likelihood Estimate (DHS 3.2.1)

Estimate $\underline{\hat{\theta}}$ ($\underline{\theta}$ = fixed but unknown)

The maximum likelihood (ML) estimate $\underline{\hat{\theta}}$ of θ is that value $\underline{\hat{\theta}}$ which maximizes $p(\underline{z}|\underline{\theta})$.

Can find this by maximizing $\ln(p(\underline{z}|\underline{\theta}))$ w.r.t. $\underline{\theta}$.

The ML estimate is the est. that maximizes the probability of obtaining the samples actually observed.

How to Maximize Likelihood

Maximize p($\underline{z}|\underline{\theta}$) w.r.t $\underline{\theta}$.

Gradient w.r.t. $\underline{\theta}$: $\nabla_\theta p(z|\theta)|_{\theta=\hat{\theta}(z)}=0$

Or $\nabla_\theta \ln p(z|\theta)|_{\theta=\hat{\theta}(z)}=0$

$$\nabla_\theta = [\partial/\partial\theta_1, \partial/\partial\theta_2, ...]$$

$$p(\underline{z}|\underline{\theta}) = \prod_{j=1}^{J} p(\underline{x}_i|\underline{\theta})$$

J samples, assumed independent.

ln p($\underline{z}|\underline{\theta}$)= $\sum_{j=1}^{J}$ ln p($\underline{x}_j|\underline{\theta}$)

$$\nabla_\theta[\ln p(z|\theta)] = \sum_{j=1}^{J}\nabla_\theta\{\ln p(x_j|\theta)\} = 0$$

Solution $\underline{\hat{\theta}}$ = maximum likelihood.

ML Example 1 (DHS 3.2.2)

Multivariate normal, **unknown mean**, **known variance**. N($\underline{x}_j$,$\underline{m}$,$\underline{\Sigma}$)

$$p(\underline{z}|\underline{\theta}) = \prod_{j=1}^{J} p(\underline{x}_i|\underline{\theta})$$

ln p($\underline{z}|\underline{\theta}$)= $\sum_{j=1}^{J}$ ln p($\underline{x}_j|\underline{\theta}$)

(Drop vector and matrix notations)

For normal density:

ln p(x_j|m) = -1/2 ln{$(2\pi)^J|\Sigma|$} – 1/2 $(x_j-m)^T\Sigma^{-1}(x_j-m)$

$$\nabla_m[\ln p(x_j|m)] = \Sigma^{-1}(x_j - m)$$

$$\nabla_m[\ln p(z|m)]|_{m=\hat{m}} = \sum_{j=1}^{J}\Sigma^{-1}(x_j - \hat{m}) = 0$$

$\sum_{j=1}^{J}$ x_j= $\sum_{j=1}^{J}$ $\hat{m} = J\hat{m}$

$\hat{m}$=1/J $\sum_{j=1}^{J}$ x_j The sample mean estimate is the ML estimate.

ML Example 2 (DHS 3.2.3)

Univariate normal, **unknown mean**, **unknown variance**.

$\theta=[\theta_1, \theta_2]= [m, \sigma^2]$

$\ln[p(x_j|\theta)]=-1/2 \ln[2\pi\theta_2]-1/2\theta_2(x_j-\theta_1)^2$

$$\nabla_\theta[\ln p(x_j \mid \theta)] = [1/\theta_2(x_j-\theta_1), -1/2\theta_2 + 1/2\theta_2^2(x_j-\theta_1)^2]$$

$$\nabla_\theta[\ln p(z \mid \theta)]_{\theta=\hat{\theta}} = 0$$

$$\sum_{j=1}^{J} \frac{1}{\hat{\theta}_2}(x_j - \hat{\theta}_1) = 0$$

$$\sum_{j=1}^{J} -\frac{1}{2\hat{\theta}_2} + \frac{1}{2\hat{\theta}_2^2} \sum_{j=1}^{J} (x_j - \hat{\theta}_1)^2 = 0$$

Univariate case

$$\hat{\theta}_1 = \hat{m} = \frac{1}{J}\sum x_j \quad \text{(sample mean)}$$

$$\hat{\theta}_2 = \hat{\sigma}^2 = \frac{1}{J}\sum (x_j - \hat{m})^2 \quad \text{(sample variance)}$$

Note: $\hat{\sigma}^2$ is a biased estimate.

Multivariate case yields:

$$\hat{m} = \frac{1}{J}\sum x_j$$

$$\hat{\Sigma} = \frac{1}{J}\sum_{j=1}^{J} (x_j - \hat{m})(x_j - \hat{m})^T$$

Note: $\hat{\Sigma}$ is a biased estimate.