


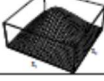








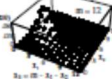
## Parameter Estimation (DHS Ch. 3)

Not all statistics known

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 (Table 3.1). Then estimate parameters.

Table 3.1: Common Exponential Distributions and their Sufficient Statistics.

Name	Distribution	Domain		s	$[g(s, \theta)]^{1/n}$
Normal	$p(x \theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\theta_2(x-\theta_1)^2}$	$\theta_2 > 0$		$\frac{1}{n} \sum_{k=1}^n x_k$ $\frac{1}{n} \sum_{k=1}^n x_k^2$	$\sqrt{\theta_2} e^{-\frac{1}{2}\theta_2(s_2 - 2\theta_1 s_1 + \theta_1^2)}$
Multi-variate Normal	$p(\mathbf{x} \theta) = \frac{ \Theta_2 ^{1/2}}{(2\pi)^{d/2}} e^{-\frac{1}{2}(\mathbf{x}-\theta_1)^t \Theta_2 (\mathbf{x}-\theta_1)}$	$\Theta_2$ positive definite		$\frac{1}{n} \sum_{k=1}^n \mathbf{x}_k$ $\frac{1}{n} \sum_{k=1}^n \mathbf{x}_k \mathbf{x}_k^t$	$ \Theta_2 ^{1/2} e^{-\frac{1}{2}[\text{tr} \Theta_2 s_2 - 2\theta_1^t \Theta_2 s_1 + \theta_1^t \Theta_2 \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	$p(x \theta) = \begin{cases} \frac{\theta^{\theta_1+1}}{\Gamma(\theta_1+1)} x^{\theta_1} e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta_1 > -1$ $\theta_2 > 0$		$\left( \frac{\sum_{k=1}^n x_k}{n} \right)^{1/n}$ $\frac{1}{n} \sum_{k=1}^n x_k$	$\frac{\theta^{\theta_1+1}}{\Gamma(\theta_1+1)} s^{\theta_1} e^{-\theta s}$
Beta	$p(x \theta) = \begin{cases} \frac{\Gamma(\theta_1+\theta_2+2)}{\Gamma(\theta_1+1)\Gamma(\theta_2+1)} x^{\theta_1} (1-x)^{\theta_2} & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$	$\theta_1 > -1$ $\theta_2 > -1$		$\left( \frac{\sum_{k=1}^n x_k}{n} \right)^{1/n}$ $\left( \frac{\sum_{k=1}^n (1-x_k)}{n} \right)^{1/n}$	$\frac{\Gamma(\theta_1+\theta_2+2)}{\Gamma(\theta_1+1)\Gamma(\theta_2+1)} s_1^{\theta_1} s_2^{\theta_2}$
Poisson	$P(x \theta) = \frac{\theta^x}{x!} e^{-\theta} \quad x = 0, 1, 2, \dots$	$\theta > 0$		$\frac{1}{n} \sum_{k=1}^n x_k$	$\theta^s e^{-\theta}$
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	$P(x \theta) = \frac{m!}{x!(m-x)!} \theta^x (1-\theta)^{m-x} \quad x = 0, 1, \dots, m$	$0 < \theta < 1$		$\frac{1}{n} \sum_{k=1}^n x_k$	$\theta^s (1-\theta)^{m-s}$
Multinomial	$P(\mathbf{x} \theta) = \frac{m!}{x_1! \dots x_d!} \theta_1^{x_1} \dots \theta_d^{x_d} \quad \sum_{i=1}^d x_i = m$	$0 < \theta_i < 1$ $\sum_{i=1}^d \theta_i = 1$		$\frac{1}{n} \sum_{k=1}^n \mathbf{x}_k$	$\prod_{i=1}^d \theta_i^{s_i}$

Example:  $p(x|S_i) = N(x, m_i, \Sigma_i)$

Estimate  $m_i$  and  $\Sigma_i$  from training samples

Two approaches: (1) Maximum likelihood (ML) estimation and (2) Maximum a Posterior (MAP) estimation (i.e., Bayesian estimation)

2. **Nonparametric**: estimate the density functions themselves.

=====

## 1) Parametric Models and Estimation

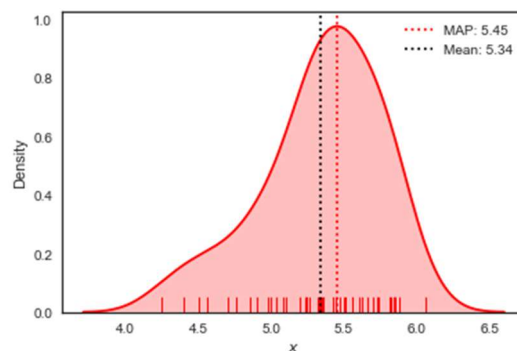
### Preliminaries (DHS 3.1)

- Previously, we designed an optimal classifier if the prior probabilities and the class-conditional densities (i.e., likelihood) 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 problem, but not the class-conditional densities
- At least, assume probability density functions with unknown parameters
- Two common and reasonable procedures: (1) Maximum Likelihood (ML) estimation and (2) MAP Bayes estimation
- ML views the parameters as quantities as fixed values, but unknown (Fig. 3.1)
- MAP 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 Likelihood (See Fig. 3.1)
- Maximum A Posteriori and Mode (See right below & Fig. 3.2))
- Fig. 3.1 vs. Fig. 3.2



MAP estimation

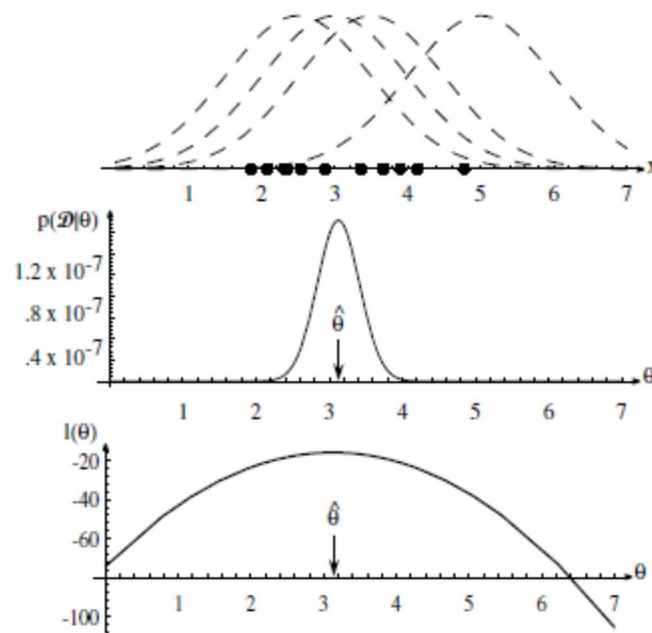


Figure 3.1: The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood  $p(\mathcal{D}|\theta)$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

### Bayesian Learning of Parameters (Fig. 3.2)

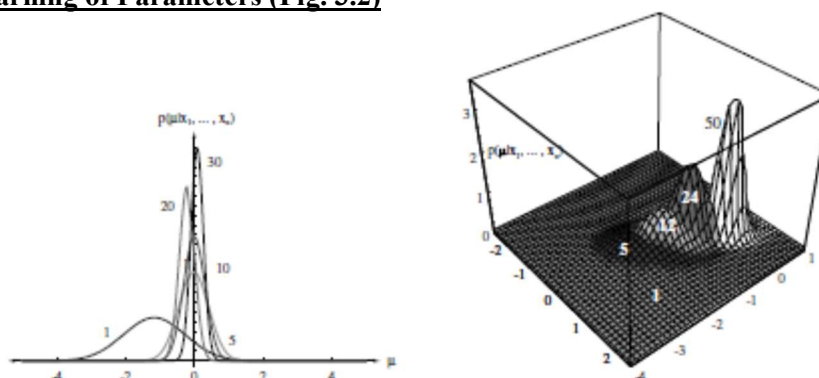


Figure 3.2: Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labelled by the number of training samples used in the estimation.

### Maximum Likelihood Estimate (DHS 3.2.1)

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

The maximum likelihood (ML) estimate  $\hat{\theta}$  of  $\theta$  is that value  $\hat{\theta}$  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.  $\underline{\theta}$ :  $\nabla_{\theta} p(\underline{z} | \underline{\theta})|_{\theta=\hat{\theta}(\underline{z})} = 0$

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

$\nabla_{\theta} = [\partial / \partial \theta_1, \partial / \partial \theta_2, \dots]$

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

J samples, assumed independent.

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

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

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

#### ML Example 1 (DHS 3.2.2)

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

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

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

(Drop vector and matrix notations)

For normal density:

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

$$\nabla_{\underline{m}} [\ln p(x_j | \underline{m})] = \underline{\Sigma}^{-1} (\underline{x}_j - \underline{m})$$

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

$$\sum_{j=1}^J \underline{x}_j = \sum_{j=1}^J \hat{\underline{m}} = J \hat{\underline{m}}$$

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

The sample mean estimate is the ML estimate.

### **ML Example 2 (DHS 3.2.3)**

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

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

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

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

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

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

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

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

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

#### **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.

※ **How to make variance unbiased???**

## Maximum A Posteriori (MAP) Estimate

MAXIMUM A POSTERIORI MODE	We note in passing that a related class of estimators — <i>maximum a posteriori</i> or MAP estimators — find the value of $\theta$ that maximizes $l(\theta)p(\theta)$ . Thus a maximum likelihood estimator is a MAP estimator for the uniform or “flat” prior. As such, a MAP estimator finds the peak, or <i>mode</i> of a posterior density. The drawback of MAP estimators is that if we choose some arbitrary nonlinear transformation of the parameter space (e.g., an overall rotation), the density will change, and our MAP solution need no longer be appropriate (Sec. 3.5.2).
---------------------------------	--

---

\*MAP is beyond the scope of this class

## 2) Nonparametric Models and Estimation

Suppose you don't know the form of the densities

Try to estimate  $p(x|S_k)$  (=likelihood) or  $P(S_k|x)$  (= a posteriori)

→ Estimation of probability density functions.

### Concept

Bin locations set by samples, bin shape is a parameter.

Each sample  $x_i$  gives rise to a window function centered about  $x_i$ . Estimate  $p(x)$  by summing over window functions.

### Parzen Window Estimation (DHS 4.3)

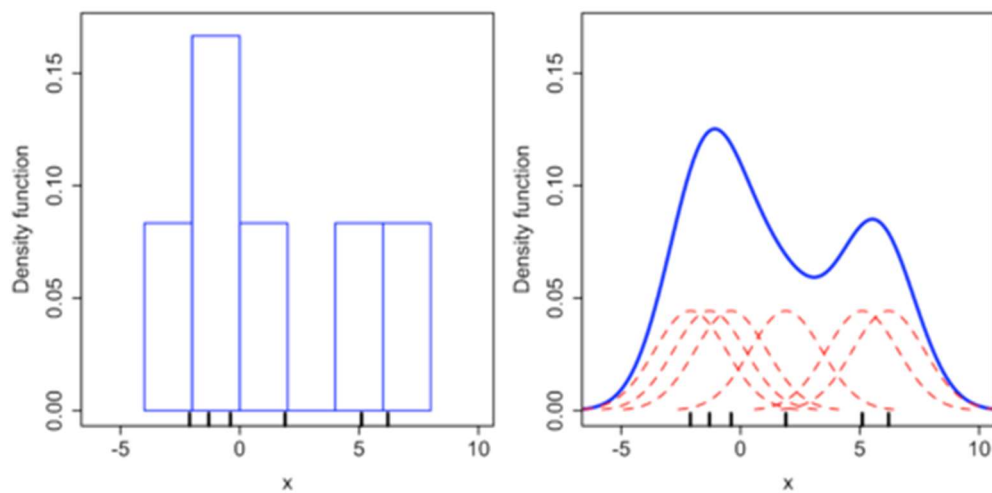
Define a window function  $\Delta(\underline{x}) = \Delta(\underline{x} - \underline{x}_i)$

Estimate  $p(\underline{x})$ . Given a sample  $\underline{x} = \underline{x}_i$ ,  $p(\underline{x}_i)$  is nonzero, and if  $p(\underline{x})$  is continuous,  $p(\underline{x})$  is nonzero for  $\underline{x}$  close to  $\underline{x}_i$

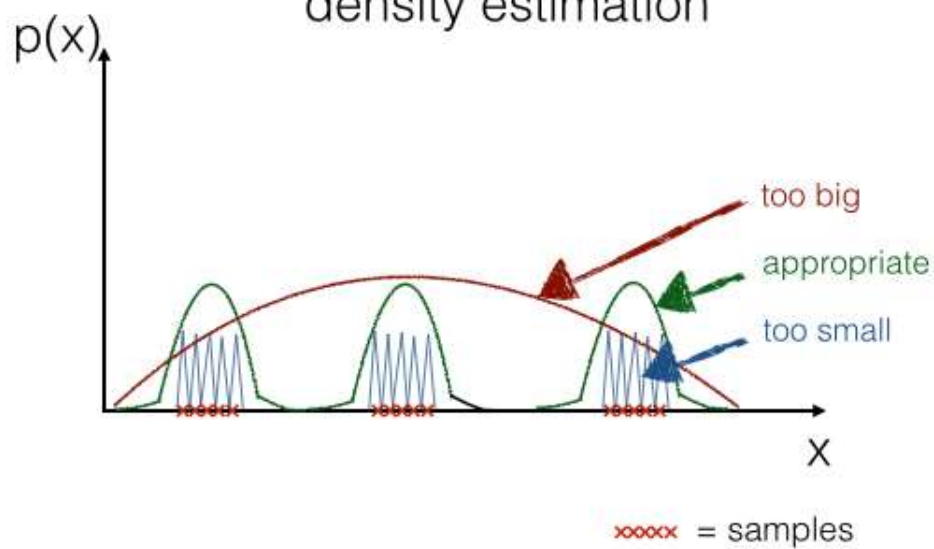
Use window function  $\Delta(\underline{x} - \underline{x}_i)$  centered at  $\underline{x}_i$ .  $\Delta$  should be non-increasing.

Estimate of  $p(\underline{x})$  is

$$p_j(\underline{x}) = (1/j) \sum_{i=1}^j \Delta(\underline{x} - \underline{x}_i) \quad (\text{Parzen window estimate})$$



very simplified illustration of how  
the window width affects the  
density estimation



Check Matlab Handouts for density function estimation