

LECTURE 6: Parameter Estimation

- **Introduction**
- **Parameter Estimation**
- **Maximum Likelihood**
- **Bayesian Estimation**
- **Numerical Examples**



Introduction

- **Previous chapters have shown you how to develop decision regions and classifiers when the underlying density is known**
 - Bayesian Decision Theory introduced the general formulation
 - Quadratic Classifiers covered the case of multi-modal Gaussian data
- **In most situations, however, knowledge of the true distributions is not available and must be determined from experimental data**
 - Two approaches are commonplace
 - Parameter Estimation (this lecture)
 - Non-parametric Density Estimation (the next two lectures)
- **Parameter Estimation**
 - These methods assume a particular form for the density (e.g. Gaussian) so that only the parameters (e.g., mean and variance) need to be determined
 - Maximum Likelihood
 - Bayesian Estimation
- **Non-parametric Density Estimation**
 - These methods do not assume ANY knowledge about the density
 - Kernel Density Estimation
 - Nearest Neighbor Rule



Maximum Likelihood vs. Bayesian Parameter Estimation

■ Maximum Likelihood

- The parameters are assumed to be FIXED but unknown
- The ML solution seeks the solution that “best” explains the dataset X

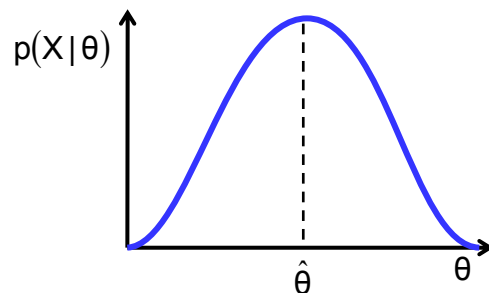
$$\hat{\theta} = \operatorname{argmax}[p(X | \theta)]$$

■ Bayesian Estimation

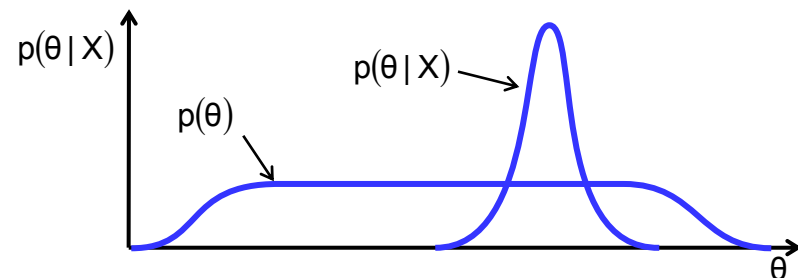
- The parameters are assumed to be random variables with some (assumed) known a priori distribution
- Bayesian methods seeks to estimate the posterior density $p(\theta|X)$
- The final density $p(x|X)$ is obtained by integrating out the parameters:

$$p(x | X) = \int p(x | \theta)p(\theta | X)d\theta$$

MAXIMUM LIKELIHOOD



BAYESIAN



Maximum Likelihood (1)

- Suppose we consider estimating a density function $p(x)$ which depends on a number of parameters $\theta = [\theta_1, \theta_2, \dots, \theta_M]^T$
 - For a Gaussian pdf $\theta_1 = \mu$, $\theta_2 = \sigma$ and $p(x) = N(\mu, \sigma)$
 - To make the dependence on the parameters θ explicit we write $p(x|\theta)$
- Assume that we have a number of examples $X = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$ drawn independently from the distribution $p(x|\theta)$ (an i.i.d. set)

- Then we can write

$$p(X|\theta) = \prod_{k=1}^N p(x^{(k)}|\theta)$$

- The ML estimate of θ is the value that maximizes the likelihood $p(X|\theta)$

$$\hat{\theta} = \operatorname{argmax}[p(X|\theta)]$$

- This corresponds to the intuitively pleasing idea of choosing the value of θ that is most likely to give rise to the data!

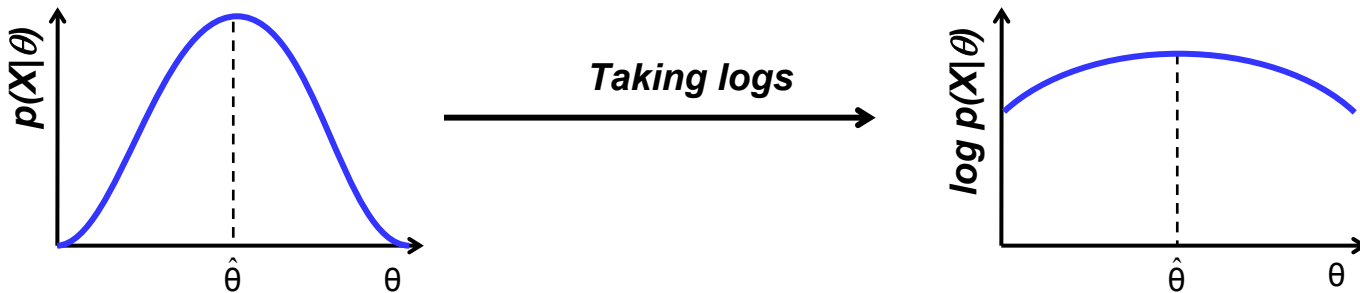


Maximum Likelihood (2)

- For analytical purposes it is convenient to work with the log of the likelihood

- Since the log is a monotonic function

$$\hat{\theta} = \operatorname{argmax}[p(X | \theta)] = \operatorname{argmax}[\log p(X | \theta)]$$



- Then the Maximum Likelihood estimate of the parameter θ can be written as

$$\hat{\theta} = \operatorname{argmax} \left[\log \prod_{k=1}^N p(x^{(k)} | \theta) \right] = \operatorname{argmax} \left[\sum_{k=1}^N \log p(x^{(k)} | \theta) \right]$$

- Maximizing a sum of terms is always an easier task than maximizing a product
 - To convince yourself, think of computing the derivative of a long product of terms!
- An added advantage of taking logs will become very clear when the distribution is Gaussian



Example: Gaussian case, μ unknown

- Assume a dataset $X=\{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$ and a density of the form $p(x)=N(\mu, \sigma)$ where the standard deviation σ is known
- What is the Maximum Likelihood estimate of the mean?

$$\begin{aligned}\theta = \mu &\Rightarrow \hat{\theta} = \operatorname{argmax}_{\theta} \sum_{k=1}^N \log p(x^{(k)} | \theta) \\ &= \operatorname{argmax}_{\mu} \sum_{k=1}^N \log \left(\frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{1}{2\sigma^2} (x^{(k)} - \mu)^2 \right) \right) \\ &= \operatorname{argmax}_{\mu} \sum_{k=1}^N \left\{ \log \left(\frac{1}{\sqrt{2\pi}\sigma} \right) - \frac{1}{2\sigma^2} (x^{(k)} - \mu)^2 \right\}\end{aligned}$$

- The maxima (or minima) of a function are defined by the zeros of its derivative:

$$\frac{\partial \sum_{k=1}^N \log p(x^{(k)} | \theta)}{\partial \theta} = \frac{\partial}{\partial \mu} \sum_{k=1}^N \{\bullet\} = 0 \Rightarrow \mu = \frac{1}{n} \sum_{k=1}^N x^{(k)}$$

- So the ML estimate of the mean is the average value of the training data, a very intuitive result!



Example: Gaussian case, both μ and σ unknown

- This is a more general case when neither the mean nor the standard deviation are known

- Fortunately, the problem can be solved in the same fashion
- In this case, the derivative becomes a gradient since we have two variables

$$\hat{\theta} = \begin{bmatrix} \theta_1 = \mu \\ \theta_2 = \sigma^2 \end{bmatrix} \Rightarrow \nabla_{\theta} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \sum_{k=1}^N \log p(x^{(k)} | \theta) \\ \frac{\partial}{\partial \theta_2} \sum_{k=1}^N \log p(x^{(k)} | \theta) \end{bmatrix} = \sum_{k=1}^N \begin{bmatrix} \frac{1}{\theta_2} (x^{(k)} - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x^{(k)} - \theta_1)^2}{2\theta_2^2} \end{bmatrix} = 0$$

- Solving for θ_1 and θ_2 yields

$$\hat{\theta}_1 = \frac{1}{N} \sum_{k=1}^N x^{(k)}; \quad \hat{\theta}_2 = \frac{1}{N} \sum_{k=1}^N (x^{(k)} - \hat{\theta}_1)^2$$

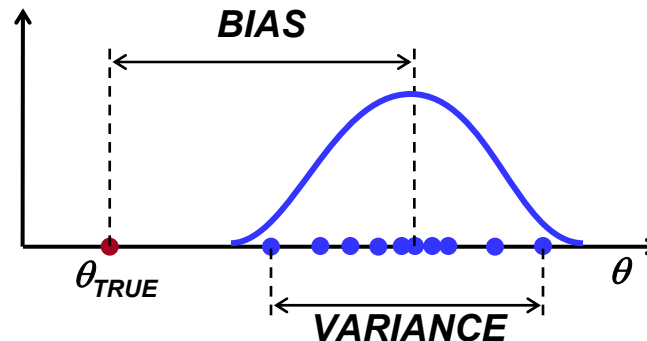
- Therefore, the ML of the variance is the sample variance of the dataset, again a very pleasing result
- Similarly, it can be shown that the Maximum Likelihood parameter estimates for the multivariate Gaussian are also the sample mean vector and sample covariance matrix

$$\hat{\mu} = \frac{1}{N} \sum_{k=1}^N x^{(k)}; \quad \hat{\Sigma} = \frac{1}{N} \sum_{k=1}^N (x^{(k)} - \hat{\mu})(x^{(k)} - \hat{\mu})^T$$



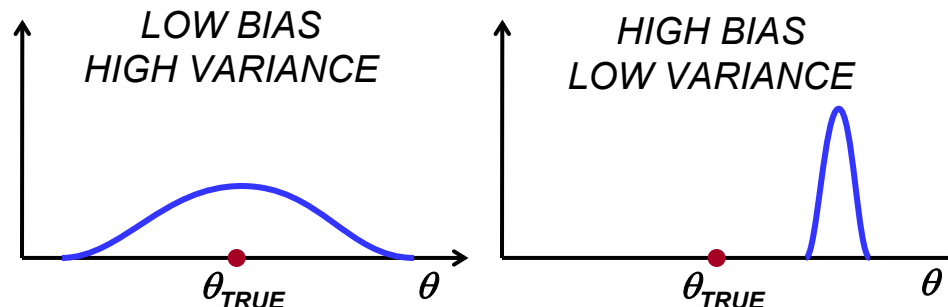
Bias and variance (1)

- How good are these estimates? Two measures of “goodness” are used for statistical estimates
 - **BIAS**: how close is the estimate to the true value?
 - **VARIANCE**: how much does the estimate change for different runs (e.g. different datasets)?



■ The bias-variance tradeoff

- In most cases, you can only decrease one of them at the expense of the other



Bias of the ML estimate of mean and variance

■ What is the bias of the ML estimate of the mean?

$$E[\hat{\mu}] = E\left[\frac{1}{N} \sum_{k=1}^N x^{(k)}\right] = \frac{1}{N} \sum_{k=1}^N E[x^{(k)}] = \mu$$

- Therefore the mean is an unbiased estimate

■ What is the bias of the ML estimate of the variance?

$$E[\hat{\sigma}^2] = E\left[\frac{1}{N} \sum_{k=1}^N (x^{(k)} - \hat{\mu})^2\right] = \frac{N-1}{N} \sigma^2 \neq \sigma^2$$

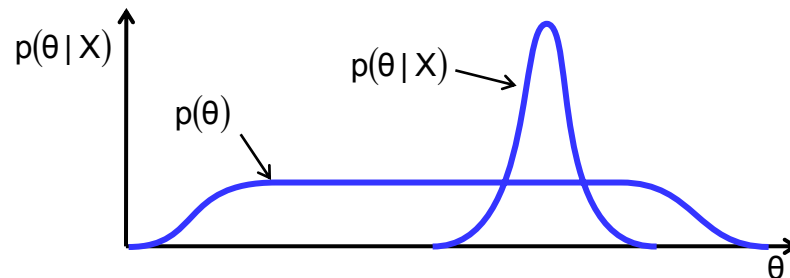
- Thus, the ML estimate of variance is BIASED
 - The problem is that the ML estimate of variance uses the ML estimate of the mean instead of its true value
- How “bad” is this bias?
 - For $N \rightarrow \infty$ the bias becomes zero asymptotically
 - The bias is only noticeable when we have very few samples, in which case we should not be doing statistics in the first place
- Notice that MATLAB uses an unbiased estimate of the co-variance

$$\hat{\Sigma}_{\text{UNBIASED}} = \frac{1}{N-1} \sum_{k=1}^N (x^{(k)} - \hat{\mu})(x^{(k)} - \hat{\mu})^T$$



Bayesian Estimation (1)

- In the Bayesian approach, our uncertainty about the parameters is represented by a pdf
 - Before we observe the data, the parameters are described by a prior density $p(\theta)$ which is typically very broad to reflect the fact that we know little about its true value
 - Once we obtain data, we make use of Bayes theorem to find the posterior $p(\theta|X)$
 - Ideally, we want the data to sharpen the posterior $p(\theta|X)$, in other words, reduce our uncertainty about the parameters



- BUT keep in mind that our goal is to estimate the density $p(x)$ or, more exactly, $p(x|X)$, the density given the evidence provided by the dataset X



Bayesian Estimation (2)

■ Let us derive the theoretical expression of a Bayesian estimate

- From the definition of conditional probability

$$p(x, \theta | X) = p(x | \theta, X)p(\theta | X)$$

- $P(x|\theta, X)$ is independent of X since knowledge of θ completely specifies the (parametric) density. Therefore

$$p(x, \theta | X) = p(x | \theta)p(\theta | X)$$

- and, using the theorem of total probability we can integrate θ out:

$$p(x | X) = \int p(x | \theta)p(\theta | X)d\theta$$

- The only unknown in this expression is $p(\theta|X)$ which, using Bayes rule, becomes:

$$p(\theta | X) = \frac{p(X | \theta)p(\theta)}{p(X)} = \frac{p(X | \theta)p(\theta)}{\int p(X | \theta)p(\theta)d\theta}$$

- Where $p(X|\theta)$ can be computed using the i.i.d. assumption

$$p(X | \theta) = \prod_{k=1}^N p(x^{(k)} | \theta)$$

- *NOTE: The last three expressions suggest a procedure to estimate $p(x|X)$. This is not to say that integration of these expressions is easy!*



Example (1)

- Assume a univariate density where our random variable x is generated from a normal distribution with known standard deviation

- Our goal is to find the mean μ of the distribution given some i.i.d. data points $X = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$
- To capture our knowledge about $\theta = \mu$, we assume that it also follows a normal density with mean μ_0 and standard deviation σ_0

$$p_0(\theta) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right)$$

- We use Bayes rule to develop an expression for the posterior $p(\theta|X)$:

$$\begin{aligned} p(\theta | X) &= \frac{p(X | \theta)p(\theta)}{p(X)} = \frac{p_0(\theta)}{p(X)} \prod_{k=1}^N p(x^{(k)} | \theta) = \\ &\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right) \frac{1}{p(X)} \prod_{k=1}^N \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x^{(k)} - \theta)^2\right) \right] \end{aligned}$$

From [Bishop, 1995]



Example (2)

- To understand how Bayesian Estimation changes the posterior as more data becomes available, we will find the maximum of $p(\theta|X)$

- The partial derivative with respect to $\theta=\mu$ is

$$\frac{\partial}{\partial \theta} \log p(\theta | X) = 0 \Rightarrow \frac{\partial}{\partial \mu} \left(-\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 + \sum_{k=1}^N \frac{-1}{2\sigma^2} (x^{(k)} - \mu)^2 \right) = 0$$

- which, after some algebraic manipulation, becomes:

$$\mu_N = \underbrace{\frac{\sigma^2}{\sigma^2 + N\sigma_0^2} \mu_0}_{\text{PRIOR}} + \underbrace{\frac{N\sigma_0^2}{\sigma^2 + N\sigma_0^2} \sum_{k=1}^N x^{(k)}}_{\text{MAXIMUM LIKELIHOOD}}$$

- Therefore, as the number N of available data points increases, the estimate of the mean μ_N moves from the initial prior μ_0 to the maximum likelihood solution
- Similarly, the standard deviation σ_N can be found to be:

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}$$

From [Bishop, 1995]



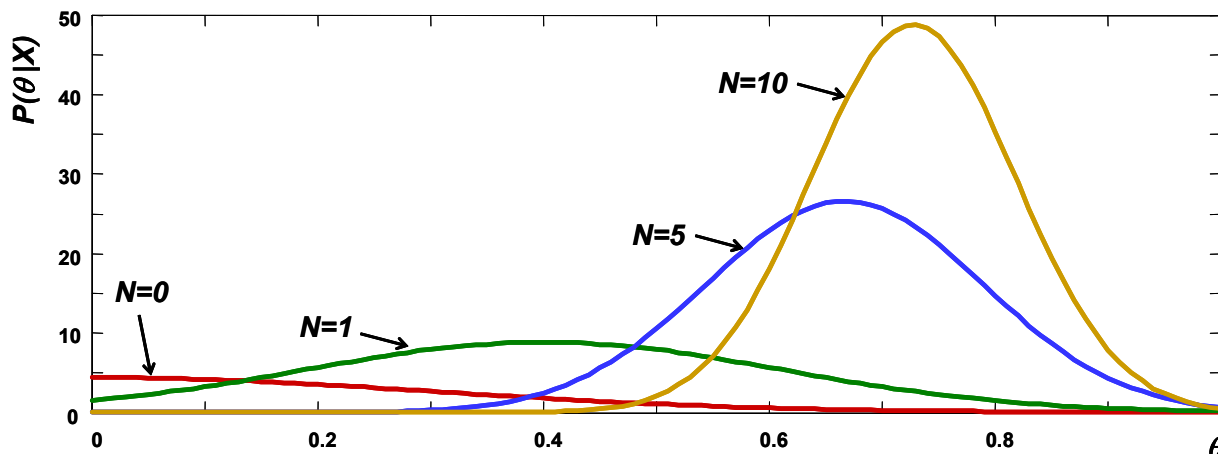
Example (3)

■ A numerical example will allow us to develop a better insight about the process of Bayesian Estimation

- Assume that the true mean of the distribution $p(x)$ is $\mu=0.8$ with standard deviation $\sigma=0.3$
 - In reality we would not know the true mean, $\mu=0.8$, we are just “playing God”
- We generate a number of examples from this distribution
- To capture our lack of knowledge about the mean, we assume a normal prior $p_0(\theta)$ $\mu_0=0.0$ and $\sigma_0=0.3$

■ The figure below shows the posterior $p(\mu|X)$

- As the number of training examples increases, the estimate μ_N approaches its true value ($\mu=0.8$) and the spread σ_N (or uncertainty in the estimate) decreases



Maximum Likelihood vs. Bayesian Estimation

■ What is the relationship between these two estimates?

- By definition, $p(X|\theta)$ peaks at the ML estimate. If this peak is relatively sharp and the prior is broad, then the integral below will be dominated by the region around the ML estimate

$$p(x | X) = \int p(x | \theta) p(\theta | X) d\theta \cong p(x | \hat{\theta}) \underbrace{\int p(\theta | X) d\theta}_1 = p(x | \hat{\theta})$$

- Therefore, the Bayesian estimate will approximate the ML solution
- As we have seen in the previous example, when the number of available data increases, the posterior $p(\theta|X)$ tends to sharpen
 - Therefore, the Bayesian estimate of $p(x)$ will approach the ML solution as $N \rightarrow \infty$
 - In practice, only when we have a limited number of observations will the two approaches yield different results

