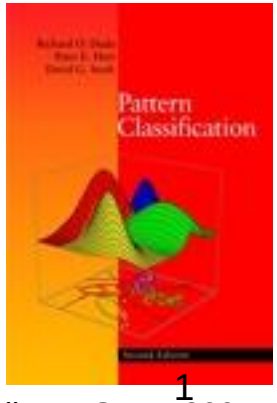
Chapter 3 (part 1): Maximum-Likelihood & Bayesian Parameter Estimation

- Introduction
- Maximum-Likelihood Estimation
  - Example of a Specific Case
  - The Gaussian Case: unknown μ and σ
  - Bias
- Appendix: ML Problem Statement



- Introduction
  - Data availability in a Bayesian framework
    - We could design an optimal classifier if we knew:
      - $P(\omega_i)$ (priors)
      - $p(\mathbf{x}|\omega_i)$ (class-conditional densities)

      Unfortunately, we rarely have this complete information!

  - Design a classifier from a training sample
    - No problem with prior estimation
    - Samples are often too small for class-conditional estimation (large dimension of feature space!)

– A priori information about the problem

– Normality of $p(\mathbf{x}|\omega_i)$

$$p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

- Characterized by 2 parameters

– Estimation techniques

- Maximum-Likelihood (ML), *maximum a posteriori* (MAP) and the Bayesian estimations
- Results are nearly identical, but the approaches are different

- Parameters in ML estimation are fixed but unknown!

- Best parameters are obtained by maximizing the probability of obtaining the samples observed.

- Bayesian methods view the parameters as random variables having some known distribution.

- In the Bayesian case, we shall see that a typical effect of observing additional samples is to sharpen the a posteriori density function, causing it to peak near the true values of the parameters. This phenomenon is known as Bayesian *learning*.

- In either approach, we use $p(\mathbf{x}|\omega_i)$ for our classification rule!

- Supervised vs unsupervised learning
  - In both cases, samples **x** are assumed to be obtained by selecting a state of nature ω_i with probability $P(\omega_i)$, and then independently selecting **x** according to the probability law $p(\mathbf{x}|\omega_i)$.
  - The distinction is that with supervised learning we know the state of nature (class label) for each sample, whereas with unsupervised learning we do not.

Maximum-Likelihood Estimation

- Has good convergence properties as the sample size increases.
- Simpler than any other alternative techniques.

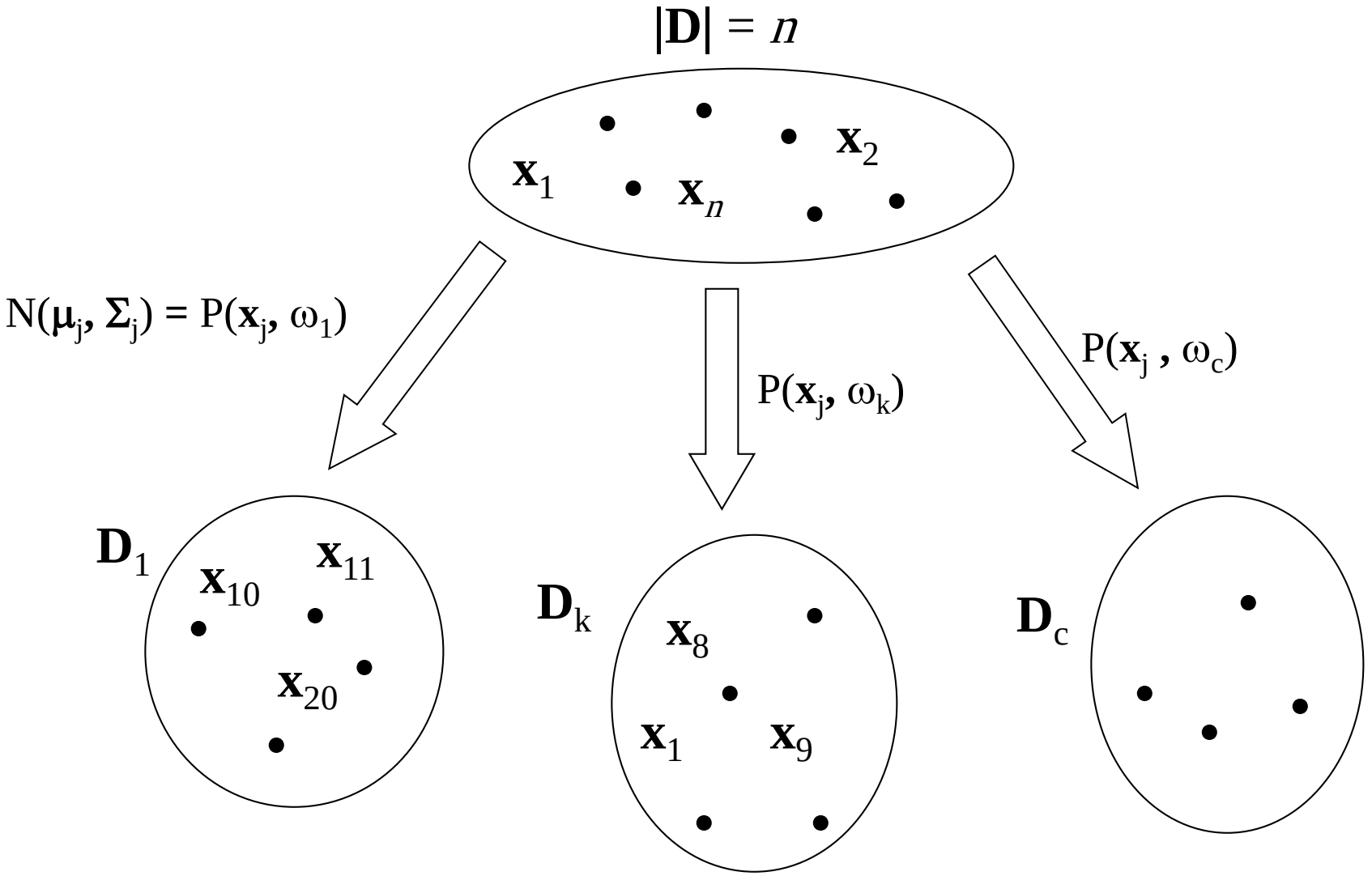
– General principle

- Assume we have c classes and

  $p(\mathbf{x}|\omega_j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$

  $p(\mathbf{x}|\omega_j) \equiv p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)$ where:

$$\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = (\mu_j^1, \mu_j^2, ..., \sigma_j^{11}, \sigma_j^{22}, ..., \text{cov}(x_j^m, x_j^n)...)$$


|D| = n
x1
x2
xn
N(μj, Σj) = P(xj, ω1)
P(xj, ωk)
P(xj , ωc)
D1
x10
x11
x20
Dk
x8
x1
x9
Dc

- Use the information provided by the training samples to estimate $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_c)$. $\boldsymbol{\theta}_i$ ($i = 1, 2, \ldots, c$) is associated with each category.

- Suppose that D contains n samples, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ were drawn independently

$$p(D \mid \boldsymbol{\theta}) = \prod_{k=1}^{k=n} p(\mathbf{x}_k \mid \boldsymbol{\theta}) = F(\boldsymbol{\theta})$$

$p(D \mid \boldsymbol{\theta})$ is called the likelihood of $\boldsymbol{\theta}$ w.r.t. the set of samples)

- ML estimate of $\boldsymbol{\theta}$ is, by definition the value that $\hat{\boldsymbol{\theta}}$ maximizes $p(D \mid \boldsymbol{\theta})$

  "It is the value of $\boldsymbol{\theta}$ that best agrees with the actually observed training sample"

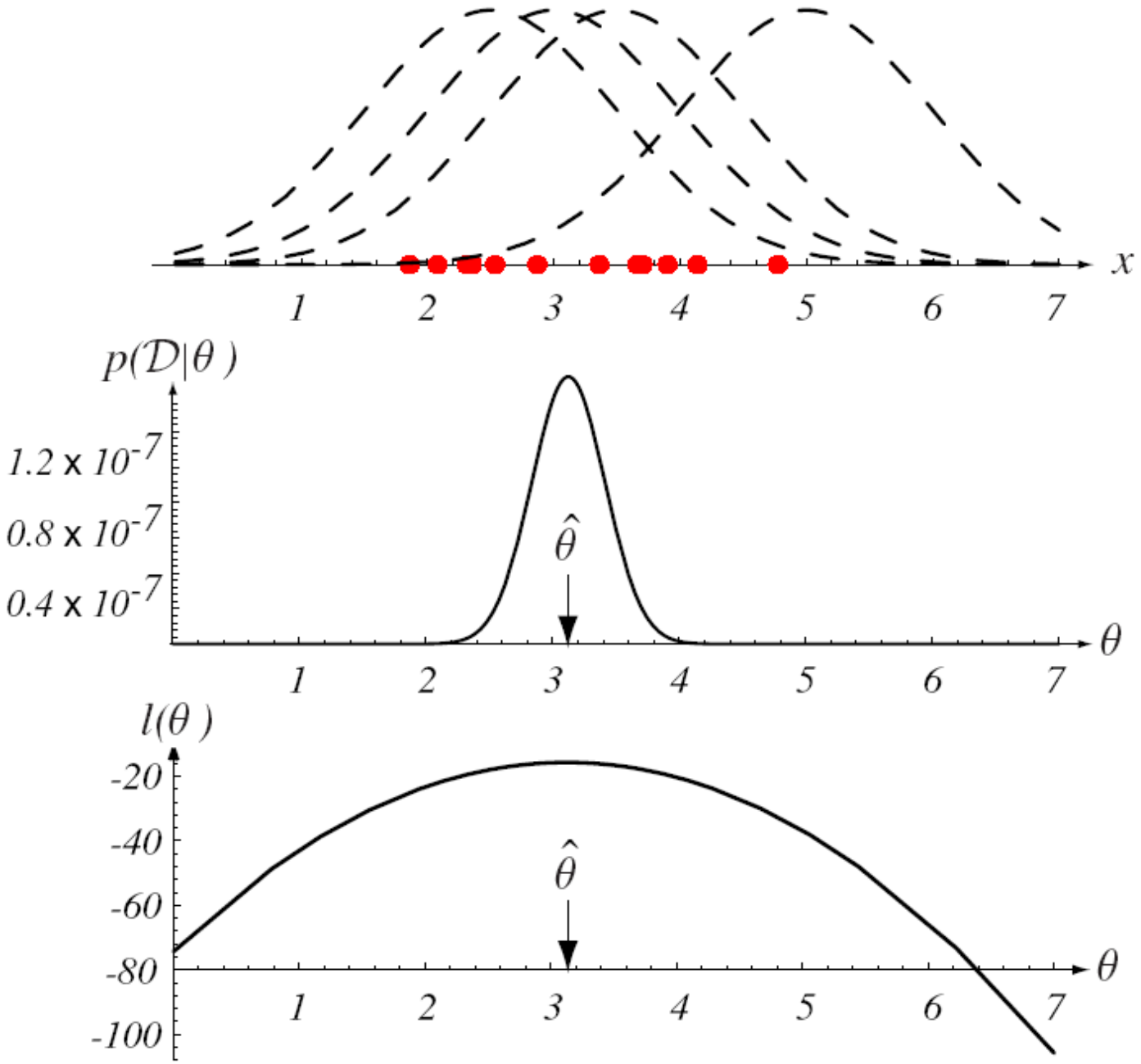

x
1
2
3
4
5
6
7
p(D|θ)
1.2 x 10-7
0.8 x 10-7
0.4 x 10-7
θ̂
θ
l(θ)
-20
-40
-60
-80
-100

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.

- For analytical purposes, it is usually easier to work with the logarithm of the likelihood than with the likelihood itself.
- If $p(D/\boldsymbol{\theta})$ is a well behaved, differentiable function of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$ can be found by the standard methods of differential calculus.
- Optimal estimation (number of parameters to be set is p)
  - Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^t$ and let $\nabla_{\boldsymbol{\theta}}$ be the gradient operator

$$\nabla_{\boldsymbol{\theta}} = \left[\frac{\partial}{\partial \theta_1} \quad \frac{\partial}{\partial \theta_2} \quad \cdots \quad \frac{\partial}{\partial \theta_p} \right]^t$$

- We define $l(\boldsymbol{\theta})$ as the log-likelihood function

$$l(\boldsymbol{\theta}) = \ln p(D|\boldsymbol{\theta})$$

  - New problem statement:

    determine $\boldsymbol{\theta}$ that maximizes the log-likelihood

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$$

where the dependence on the data set D is implicit

Thus $p(D|\boldsymbol{\theta}) = \prod_{k=1}^{k=n} p(\mathbf{x}_k|\boldsymbol{\theta})$ we have

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

Set of necessary conditions for an optimum is:

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

$$\nabla_{\boldsymbol{\theta}} l = \mathbf{0}$$

A solution of $\hat{\boldsymbol{\theta}}$ could represent a true global maximum, a *local* maximum or minimum, or (rarely) an inflection point of $l(\boldsymbol{\theta})$.

One must be careful, too, to check if the extremum occurs at a boundary of the parameter space, which might not be apparent from the solution to this Eq.

- We note in passing that a related class of estimators — *maximum a posteriori* or MAP estimators — find the value of $\boldsymbol{\theta}$ that maximizes $l(\boldsymbol{\theta})+\ln p(\boldsymbol{\theta})$, where $p(\boldsymbol{\theta})$ describes the prior probability of different parameter values. ($l(\boldsymbol{\theta})+\ln p(\boldsymbol{\theta})=ln\, p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})=\ln p(\boldsymbol{\theta}|D)\, p(D)$)
- Thus a ML estimator is a MAP estimator for the uniform or "flat" prior.
- A MAP estimator finds the peak, or *mode* 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.

Example of a specific case: unknown $\boldsymbol{\mu}$

– $p(\mathbf{x}_i|\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
(Samples are drawn from a multivariate normal population)

$$\ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d \left|\boldsymbol{\Sigma}\right|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

and $$\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

$\boldsymbol{\theta} = \boldsymbol{\mu}$ therefore:

– The ML estimate for $\boldsymbol{\mu}$ must satisfy:

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

- Multiplying by $\boldsymbol{\Sigma}$ and rearranging, we obtain:

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

(Just the arithmetic average of the samples of the training samples!)

Conclusion:

"If $p(\mathbf{x}_k|\omega_j)$ (j = 1, 2, …, *c*) is supposed to be Gaussian in a *d* dimensional feature space; then we can estimate the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_c)^t$ and perform an optimal classification !"

ML Estimation:

Gaussian Case: *unknown* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$

$\boldsymbol{\theta} = (\theta_1, \theta_2)^t = (\mu, \sigma^2)^t$ single point

$$l = \ln p(x_k \mid \boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} p(x_k \mid \boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial\theta_1}(\ln p(x_k \mid \boldsymbol{\theta})) \\ \frac{\partial}{\partial\theta_2}(\ln p(x_k \mid \boldsymbol{\theta})) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

Summation (Applying above eq. to the full log-likelihood leads to the conditions):

$$\begin{cases} \sum_{k=1}^{k=n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 & (1) \\ -\sum_{k=1}^{k=n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{k=n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 & (2) \end{cases}$$

Combining (1) and (2), one obtains (By substituting $\hat{\mu} = \hat{\theta}_1, \ \hat{\sigma}^2 = \hat{\theta}_2$ and doing a little rearranging):

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

The multivariate case

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

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

The maximum likelihood estimate for the mean vector is the sample mean.

The maximum likelihood estimate for the covariance matrix is the arithmetic average of the n matrices

$$(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t.$$

Convergence of the Mean

- Does the maximum likelihood estimate of the variance converge to the true value of the variance? Let's start with a few simple results we will need later.
- Expected value of the ML estimate of the mean:

$$E[\hat{\mu}] = E\left[\frac{1}{n}\sum_{k=1}^{n} x_k\right]$$

$$= \frac{1}{n}\sum_{k=1}^{n} E[x_k]$$

$$= \frac{1}{n}\sum_{k=1}^{n} \mu$$

$$= \mu$$

$$\text{var}[\hat{\mu}] = E[\hat{\mu}^2] - (E[\hat{\mu}])^2$$

$$= E[\hat{\mu}^2] - \mu^2$$

$$= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)\right] - \mu^2$$

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} E[x_i x_j]\right) - \mu^2$$

$$= \frac{\sigma^2}{n} \quad \text{why?}$$

From Dr. Joseph Picone

Variance of the ML Estimate of the Mean

- The expected value of $x_i x_j$ will be μ^2 for $i \neq j$ since the two random variables are independent.
- The expected value of x_i^2 will be $\mu^2 + \sigma^2$.
- Hence, in the summation above, we have n^2-n terms with expected value μ^2 and n terms with expected value $\mu^2 + \sigma^2$.
- Thus,

$$\text{var}[\hat{\mu}] = \frac{1}{n^2}\left(\left(n^2 - n\right)\mu^2 + n\left(\mu^2 + \sigma^2\right)\right) - \mu^2 = \frac{\sigma^2}{n}$$

which implies:

$$E[\hat{\mu}^2] = \text{var}[\hat{\mu}] + (E[\hat{\mu}])^2 = \frac{\sigma^2}{n} + \mu^2$$

- We see that the variance of the estimate goes to zero as n goes to infinity, and our estimate converges to the true estimate (error goes to zero).

- Bias

  – ML estimate for σ^2 is biased because:

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right] = \frac{n-1}{n}.\sigma^2 = \sigma^2 - \frac{\sigma^2}{n} \neq \sigma^2$$

Why?

  – An elementary unbiased estimator for $\boldsymbol{\Sigma}$ is:

$$\underbrace{C = \frac{1}{n-1}\sum_{k=1}^{k=n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t}_{\textit{Sample}\ \text{covariance matrix}} \quad , \quad E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right] = \sigma^2$$

Derivation of Expectation of ML estimate for σ^2

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right]$$

$$= E[\frac{1}{n}\sum_{i=1}^{n} x_i^2] - 2E[\frac{1}{n}\sum_{i=1}^{n} x_i\hat{\mu}] + E[\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}^2]$$

$$= \mu^2 + \sigma^2 - 2(\mu^2 + \frac{\sigma^2}{n}) + (\mu^2 + \frac{\sigma^2}{n})$$

$$= \frac{n-1}{n}.\sigma^2 = \sigma^2 - \frac{\sigma^2}{n} \neq \sigma^2$$

Unbiased Estimator

- If an estimator is unbiased for *all* distributions, as for example the variance estimator, then it is called *absolutely unbiased*.
- If the estimator tends to become absolutely unbiased as the number of samples becomes very large, then the estimator is *asymptotically unbiased*.
- Clearly, $\hat{\boldsymbol{\Sigma}} = \left[(n-1)/n\right]C$, and $\hat{\boldsymbol{\Sigma}}$ is asymptotically unbiased.
- What the existence of two actually shows is that no single estimate possesses all of the properties we might desire.

Model error

- If we have a reliable model for the underlying distributions and their dependence upon the parameter vector $\boldsymbol{\theta}$, the maximum likelihood classifier will give excellent results.
- But what if our model is wrong? For instance, what if we assume that a distribution comes from $\mathcal{N}(\mu, 1)$ but instead it actually comes from $\mathcal{N}(\mu, 10)$?
- Will the value we find for $\theta = \mu$ by maximum likelihood yield the best of all classifiers of the form derived from $\mathcal{N}(\mu, 1)$?
- **No.** This points out the need for reliable information concerning the models — if the assumed model is very poor, we cannot be assured that the classifier we derive is the best, even among our model set.

Maximum *a Posteriori* Probability Estimation†

- We consider $\boldsymbol{\theta}$ as a random vector, and we will estimate its value on the condition that samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ have occurred.

$$p(\boldsymbol{\theta})p(X|\boldsymbol{\theta}) = p(X)p(\boldsymbol{\theta}|X) \Rightarrow p(\boldsymbol{\theta}|X) = \frac{p(\boldsymbol{\theta})p(X|\boldsymbol{\theta})}{p(X)}$$

- The *maximum a posteriori probability* (MAP) estimate $\hat{\boldsymbol{\theta}}_{MAP}$ is defined at the point where $p(\boldsymbol{\theta}|X)$ becomes maximum

$$\hat{\boldsymbol{\theta}}_{MAP} : \frac{\partial}{\partial \boldsymbol{\theta}} p(\boldsymbol{\theta}|X) = 0 \quad \text{or} \quad \frac{\partial}{\partial \boldsymbol{\theta}} (p(\boldsymbol{\theta})p(X|\boldsymbol{\theta})) = 0$$

- The difference between the ML and the MAP estimates lies in the involvement of $p(\boldsymbol{\theta})$ in the latter case.

† Pattern Recognition.4th Ed.-Theodoridis-Koutroumbas

Example 2.4 (Maximum a Posteriori Probability Estimation) Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ be vectors stemmed from a normal distribution with known covariance matrix and unknown mean, that is,

$$p(\boldsymbol{x}_k;\boldsymbol{\mu}) = \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_k-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_k-\boldsymbol{\mu})\right)$$

And the unknown mean vector $\boldsymbol{\mu}$ is known to be normally distributed as

$$p(\boldsymbol{\mu}) = \frac{1}{(2\pi)^{l/2}\sigma_\mu^l} \exp\left(-\frac{1}{2}\frac{\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|^2}{\sigma_\mu^2}\right)$$

The MAP estimate is given by the solution of

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln\left(\prod_{k=1}^{N} p(\boldsymbol{x}_k|\boldsymbol{\mu})p(\boldsymbol{\mu})\right) = \mathbf{0}$$

or, for $\Sigma = \sigma^2 I$

$$\sum_{k=1}^{N} \frac{1}{\sigma^2}(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}) - \frac{1}{\sigma_\mu^2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0) = \mathbf{0} \Rightarrow$$

$$\hat{\boldsymbol{\mu}}_{MAP} = \frac{\boldsymbol{\mu}_0 + \frac{\sigma_\mu^2}{\sigma^2}\sum_{k=1}^{N} \boldsymbol{x}_k}{1 + \frac{\sigma_\mu^2}{\sigma^2} N}$$

We observe that if $\frac{\sigma_\mu^2}{\sigma^2} >> 1$ that is, the variance σ_μ^2 is very large and the corresponding Gaussian is very wide with little variation over the range of interest, then

$$\hat{\boldsymbol{\mu}}_{MAP} \approx \hat{\boldsymbol{\mu}}_{ML} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_k$$