Solutions to the problems of Chapter 2

2.1 Let P_c be the probability of correct classification. Then

$$P_c = \sum_{i=1}^{M} P(\boldsymbol{x} \in R_i, \omega_i) = \sum_{i=1}^{M} P(\omega_i) P(\boldsymbol{x} \in R_i | \omega_i)$$

or

$$P_c = \sum_{i=1}^{M} P(\omega_i) \int_{R_i} P(\boldsymbol{x}|\omega_i) d\boldsymbol{x} = \sum_{i=1}^{M} \int_{R_i} P(\omega_i) p(\boldsymbol{x}|\omega_i) d\boldsymbol{x}$$

or

$$P_c = \sum_{i=1}^{M} \int_{R_i} P(\omega_i|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

For minimum classification error P_e, P_c must be maximum ($P_c + P_e = 1$). Thus P_c is maximized if the regions R_i is chosen so that in each region the corresponding integrals, which are all positive, have the maximum possible value. That is

$$R_i : P(\omega_i|\boldsymbol{x}) p(\boldsymbol{x}) > P(\omega_j|\boldsymbol{x}) p(\boldsymbol{x}) \quad \forall\, i \neq j$$

or

$$R_i : P(\omega_i|\boldsymbol{x}) > P(\omega_j|\boldsymbol{x}) \quad \forall\, i \neq j$$

2.2 From the theory we have

$$\frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} > (<) \frac{P(\omega_2)}{P(\omega_1)} \frac{\lambda_{21}}{\lambda_{12}}$$

Taking the logarithm of both sides

$$\ln p(\boldsymbol{x}|\omega_1) - \ln p(\boldsymbol{x}|\omega_2) \; > (<) \; \ln \frac{P(\omega_2)\lambda_{21}}{P(\omega_1)\lambda_{12}}$$

or

$$\frac{-x^2}{2\sigma^2} + \frac{(x-1)^2}{2\sigma^2} > (<) \ln \frac{P(\omega_2)\lambda_{21}}{P(\omega_1)\lambda_{12}}$$

where constants have been omitted. Hence

$$-2x + 1 \; > (<) \; 2\sigma^2 \ln \frac{P(\omega_2)\lambda_{21}}{P(\omega_1)\lambda_{12}}$$

or

$$x_0 \; = \; \frac{1}{2} \; - \; \sigma^2 \frac{P(\omega_2)\lambda_{21}}{P(\omega_1)\lambda_{12}}$$

2.3 From the respective definitions we have

$$r \; = \; \lambda_{11} \int_{R_1} P(\omega_1) p(\boldsymbol{x}|\omega_1) d\boldsymbol{x} \; + \; \lambda_{21} \int_{R_1} P(\omega_2) p(\boldsymbol{x}|\omega_2) d\boldsymbol{x} \; + \\ \lambda_{12} \int_{R_2} P(\omega_1) p(\boldsymbol{x}|\omega_1) d\boldsymbol{x} \; + \; \lambda_{22} \int_{R_2} P(\omega_2) p(\boldsymbol{x}|\omega_2) d\boldsymbol{x}$$

which from the definitions of ε_1 and ε_2 become

$$r \; = \; \lambda_{11} P(\omega_1)(1-\varepsilon_1) \; + \; \lambda_{21} P(\omega_2)\varepsilon_2 \; + \; \lambda_{12} P(\omega_1)\varepsilon_1 \; + \; \lambda_{22} P(\omega_2)(1-\varepsilon_2)$$

and finally

$$r \; = \; \lambda_{11} P(\omega_1) \; + \; \lambda_{22} P(\omega_2) \; + \; P(\omega_1)(\lambda_{12} - \lambda_{11})\varepsilon_1 + P(\omega_2)(\lambda_{21} - \lambda_{22})\varepsilon_2$$

2.4 By the definition of the probability

$$\sum_{i=1}^{M} P(\omega_i|\boldsymbol{x}) \; = \; 1$$

since classes cover all space. Thus the maximum $P(\omega_i|\boldsymbol{x})$ has to be larger than $\frac{1}{M}$, otherwise the sum will be smaller than one. Let us now consider the probability of correct classification

$$P_c \; = \; \sum_{i=1}^{M} P(\boldsymbol{x} \in R_i, \omega_i) \; = \; \sum_{i=1}^{M} P(\omega_i) \int_{R_i} p(\boldsymbol{x}|\omega_i) d\boldsymbol{x}$$

or by the definition of R_i,

$$P_c = \sum_{i=1}^{M} \int_{R_i} P(\omega_i|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \geq \frac{1}{M} \sum_{i=1}^{M} \int_{R_i} p(\boldsymbol{x})d\boldsymbol{x}$$

or

$$P_c \geq \frac{1}{M}$$

Hence

$$P_e = 1 - P_c \leq 1 - \frac{1}{M} = \frac{M-1}{M}$$

2.5 The decision boundary point corresponds to

$$\frac{x_0}{\sigma_1^2} \exp(\frac{-x_0^2}{2\sigma_1^2}) = \frac{x_0}{\sigma_2^2} \exp(\frac{-x_0^2}{2\sigma_2^2})$$

or by taking the logarithm

$$\frac{-x_0^2}{2\sigma_1^2} = \ln \frac{\sigma_1^2}{\sigma_2^2} - \frac{x_0^2}{2\sigma_2^2}$$

and finally

$$x_0 = \sqrt{\frac{2\sigma_1^2\sigma_2^2}{\sigma_1^2 - \sigma_2^2} \ln \frac{\sigma_1^2}{\sigma_2^2}}$$

2.6 From the problem requirements we have

$$\varepsilon_1 = \int_{R_2} P(\omega_1)p(\boldsymbol{x}|\omega_1)d\boldsymbol{x} = \varepsilon$$

$$\varepsilon_2 = \int_{R_1} P(\omega_2)p(\boldsymbol{x}|\omega_2)d\boldsymbol{x}$$

Thus, minimizing ε_2 subject to the first constraint equation is equivalent with minimizing

$$\begin{aligned} Q &= \int_{R_1} P(\omega_2)p(\boldsymbol{x}|\omega_2)d\boldsymbol{x} + \theta(\int_{R_2} P(\omega_1)p(\boldsymbol{x}|\omega_1)d\boldsymbol{x} - \varepsilon) \\ &= P(\omega_2) - \theta\varepsilon + \int_{R_2} (\theta P(\omega_1)p(\boldsymbol{x}|\omega_1) - P(\omega_2)p(\boldsymbol{x}|\omega_2))d\boldsymbol{x} \end{aligned}$$

Since the first two terms do not depend on R_2, Q is minimized if R_2 is chosen so that

$$R_2 \quad : P(\omega_2)p(\boldsymbol{x}|\omega_2) > \theta P(\omega_1)p(\boldsymbol{x}|\omega_1)$$

or

$$R_2 \quad : \frac{p(\omega_2|\boldsymbol{x})}{p(\omega_1|\boldsymbol{x})} > \theta$$

It remains to specify θ. In the general case, this is not easy to be computed, however its value must be chosen so that the integral in the constraint equation to be equal to ε.

2.7 a) It suffices to compute the Mahalanobis distance of $[1.6, 1.5]^T$ from mean vectors of the classes. We have

$$\Sigma^{-1} = \begin{bmatrix} 0.9 & 0.2 \\ -0.2 & 0.6 \end{bmatrix}$$

Thus

$$d_1 = 2.05 \quad d_2 = 0.64 \quad d_3 = 3.14$$

Hence $[1.6, 1.5]^T$ is assigned to ω_2.
b) According to theory it suffices to compute the eigenvalues and eigenvectors of Σ. These are

$$\lambda_1 = 1, \ \lambda_2 = 2$$
$$\boldsymbol{v}_1 = [0.89\,,\, -0.45]^T$$
$$\boldsymbol{v}_2 = [0.45\,,\, 0.89]^T$$

Thus the ellipses, centered at $\boldsymbol{\mu}_2$ and axis

$$2\sqrt{\lambda_1}c\boldsymbol{v}_1 \ \texttt{and} \ 2\sqrt{\lambda_2}c\boldsymbol{v}_2$$

2.8 The inverse of Σ is

$$\begin{bmatrix} 5 & -2.5 & -2.5 \\ -2.5 & 5 & 2.5 \\ -2.5 & 2.5 & 5 \end{bmatrix}$$

Since the covariance matrix is the same in both classes the discriminant functions are linear given by

$$g_i(\boldsymbol{x}) = \boldsymbol{x}^T\Sigma^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i)$$

where terms independent of the classes have been dropped. For our case we have

$$g_1(\boldsymbol{x}) = \ln P(\omega_1)$$

and

$$g_2(\boldsymbol{x}) = 2.5x_2 + 2.5x_3 - 1.25 + \ln P\omega_2)$$

The decision plane is

$$g_2(\boldsymbol{x}) - g_1(\boldsymbol{x}) = 0$$

Observe that this is basically a 2-dimensional problem due to the specific choice of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. This is not necessarily the case for other choices.

2.9 The Bayesian classifier relies on the test

$$l_{12} = \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} > (<)1$$

Taking the logarithm, assignment to the class depends on the value of

$$u = \ln l_{12} = \ln p(\boldsymbol{x}|\omega_1) - \ln p(\boldsymbol{x}|\omega_2)$$

whether it is positive or negative. For our specific case we have that

$$u = \boldsymbol{x}^T\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

and the probability of error equals

$$P_e = \frac{1}{2}P(u < 0, \boldsymbol{x} \in \omega_1) + \frac{1}{2}P(u > 0, \boldsymbol{x} \in \omega_2)$$

The variable u is also normally distributed, since it is a linear combination of random variables $x_1, \ldots, x_l$, which are themselves jointly Gaussian, (Papoulis). Its mean value depends on whether $\boldsymbol{x}$ originates from ω_1 or ω_2. The respective mean values are.

$$\begin{aligned} E_1[u] &= \boldsymbol{\mu}_1^T\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \equiv \frac{1}{2}d_m^2 \end{aligned}$$

and similarly

$$E_2[u] = -\frac{1}{2}d_m^2$$

The corresponding variances are

$$\sigma_{1,u}^2 = E_1[(u - E_1[u])^2] = E_1[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^T\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]$$
$$= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = d_m^2$$

Similarly

$$\sigma_{2,u}^2 = d_m^2$$

Thus, the probability of error is now given by

$$P_e = \frac{1}{2}\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}d_m}\exp(-\frac{(u - \frac{1}{2}d_m^2)^2}{2d_m^2})du + \frac{1}{2}\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}d_m}\exp(-\frac{(u + \frac{1}{2}d_m^2)^2}{2d_m^2})du$$

which after changing the variables and taking into account the symmetry of Gaussian becomes

$$P_e = \int_{\frac{1}{2}d_m}^{\infty} \frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})dz$$

2.10 We have

$$l_{12} = \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} > (<) \frac{P(\omega_2)}{P(\omega_1)}\frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}} \equiv \theta$$

Taking the logarithm, for the Gaussian pdf's the above becomes

$$-\frac{1}{2}(\boldsymbol{x} - \mu_1)^T\Sigma^{-1}(\boldsymbol{x} - \mu_1) - \frac{1}{2}\ln|\Sigma_1| +$$
$$\frac{1}{2}(\boldsymbol{x} - \mu_2)^T\Sigma^{-1}(\boldsymbol{x} - \mu_2) + \frac{1}{2}\ln|\Sigma_2| > (<) \ln\theta$$

or

$$d_m^2(\boldsymbol{\mu}_1, \boldsymbol{x}|\Sigma_1) - d_m^2(\boldsymbol{\mu}_2, \boldsymbol{x}|\Sigma_2) + \ln\frac{|\Sigma_1|}{|\Sigma_2|} < (>) -2\ln\theta$$

2.11 Rearrangement of the previous equation results in

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma^{-1}\boldsymbol{x} > (<)\,\Theta$$
$$\Theta = \ln\theta + \frac{1}{2}(||\boldsymbol{\mu}_1||_{\Sigma^{-1}} - ||\boldsymbol{\mu}_2||_{\Sigma^{-1}})$$

2.12 a) The Bayesian classifier that minimizes the error probability is given in this case by the minimizing Euclidean distance classifier. Thus, assign $\boldsymbol{x} \in \omega_1$ if

$$||\boldsymbol{x} - \boldsymbol{\mu}_1|| < ||\boldsymbol{x} - \boldsymbol{\mu}_2||$$

b) In this case $\boldsymbol{x}$ is classified to ω_1 if

$$\frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}\frac{\lambda_{21}}{\lambda_{12}}$$

where $\lambda_{12} = 1$ and $\lambda_{21} = 0.5$. Thus following similar arguments as in theory, for Bayesian classification for normal distributions, we conclude that the decision hyperplane is

$$\begin{aligned} g_{12}(\boldsymbol{x}) &= \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}_0) \\ \boldsymbol{w} &= \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \end{aligned}$$

and

$$\boldsymbol{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \sigma^2 \ln \frac{P(\omega_1)\lambda_{21}}{P(\omega_2)\lambda_{12}} \frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2}$$

c) The following MATLAB function takes as input the variance (s), the mean (m) and the number of samples N. The output is a vector 1xN, whose elements are the N samples of the 1-D Gaussian. For 2-D independent variables, combine two samples generated above, in a single vector

$$\boldsymbol{x} = [x_1, x_2]^T$$

Function x=gaussian(m,s,N);

```
x=randn(1,N);
x=x*sqrt(s)+m;
```

2.13 Generate, first, normally distributed random vectors with statistically independent components. That is, each component of x_1, x_2 of the vector $[x_1, x_2]^T$ follows a $\mathcal{N}(1, 1)$ for class ω_1 and $\mathcal{N}(1.5, 1)$ for class ω_2. Then each vector is transformed as

$$\boldsymbol{y} = A\boldsymbol{x}, \quad A = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}$$

Then the covariance matrix of $\boldsymbol{y}$ is

$$E[(\boldsymbol{y}-\boldsymbol{\mu}_y)(\boldsymbol{y}-\boldsymbol{\mu}_y)^T] = AE[(\boldsymbol{x}-\boldsymbol{\mu}_x)(\boldsymbol{x}-\boldsymbol{\mu}_x)^T]A^T = AA^T = \Sigma$$

2.14 The constant Mahalanobis distance curves around, say, $\boldsymbol{\mu}_2$ are

$$d_m(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) = c$$

The gradient with respect to $\boldsymbol{x}$ at $\boldsymbol{x}_0$ is

$$\left.\frac{\partial d_m(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}_0} = \Sigma^{-1}(\boldsymbol{x}_0-\boldsymbol{\mu}_2) \equiv \boldsymbol{y}$$

Let, also , $\boldsymbol{x}-\boldsymbol{x}_0$ be any vector on the hyperplane, which according to the theory is vertical to

$$\boldsymbol{w} = \Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

However

$$\boldsymbol{y} = \Sigma^{-1}(\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)-\boldsymbol{\mu}_2) = \frac{1}{2}\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

is parallel to $\boldsymbol{w}$, hence vertical to $\boldsymbol{x}-\boldsymbol{x}_0$, where for simplicity $P(\omega_1) = P(\omega_2)$ has been assumed.

2.15 The only possible cases for overlap between the two pdf's are shown in the figure 2.1. Other overlap possibilities are not possible, due to the constraint that the area under each pdf must be 1. The error probability is the area of the overlap shaded regions. For the (a) and (b) cases this area is bounded by

$$P_e \leq \int_{-\infty}^{b}\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})dx - \int_{-\infty}^{\alpha}\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})dx$$

$$P_e \leq \int_{-\infty}^{\frac{b-\mu}{\sigma}}\frac{1}{\sqrt{2\pi}}\exp(-z)^2dz - \int_{-\infty}^{\frac{\alpha-\mu}{\sigma}}\frac{1}{\sqrt{2\pi}}\exp(-z)^2dz = G(\frac{b-\mu}{\sigma}) - G(\frac{a-\mu}{\sigma})$$

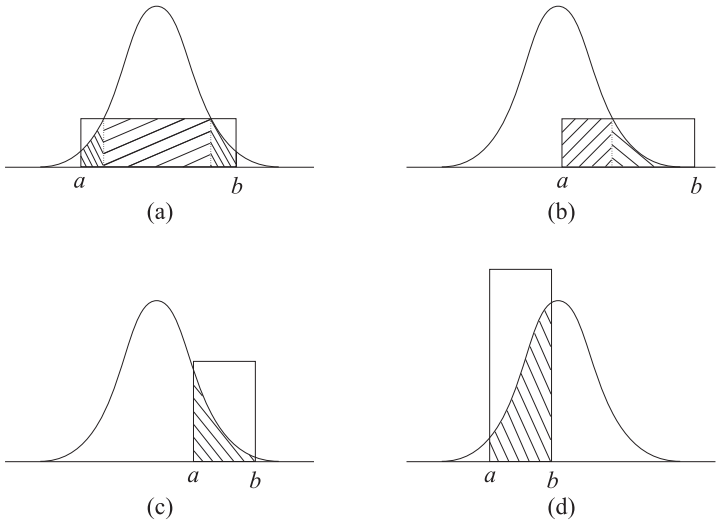

Figure 2.1: Problem 2.15

For the (c) and (d) cases the probability of error is equal to $G(\frac{b-\mu}{\sigma}) - G(\frac{a-\mu}{\sigma})$

2.16

$$\frac{\partial \ln p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{1}{p(\boldsymbol{x};\boldsymbol{\theta})}$$

Thus, by definition of the mean value

$$\boldsymbol{\mu} = \int_{-\infty}^{\infty} \frac{\partial p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{p(\boldsymbol{x};\boldsymbol{\theta})}{p(\boldsymbol{x};\boldsymbol{\theta})} d\boldsymbol{x} = \int_{-\infty}^{\infty} \frac{\partial p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} d\boldsymbol{x} =$$
$$= \frac{\partial}{\partial \boldsymbol{\theta}} \int_{-\infty}^{\infty} p(\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x} = \frac{\partial 1}{\partial \boldsymbol{\theta}} = 0$$

2.17 The likelihood function is

$$P(X;q) = \prod_{i=1}^{N} q^{x_i}(1-q)^{(1-x_i)}$$

or

$$P(X;q) = q^{\sum_{i=1}^{N} x_i}(1-q)^{(N-\sum_{i=1}^{N} x_i)}$$

$$\frac{\partial P(X;q)}{\partial q} = (\sum_{i=1}^{N} x_i) q^{(\sum_{i=1}^{N} x_i - 1)}(1-q)^{(N-\sum_{i=1}^{N} x_i)}$$
$$-(N - \sum_{i=1}^{N} x_i)(1-q)^{(N-\sum_{i=1}^{N} x_i - 1)} q^{\sum_{i=1}^{N} x_i} = 0$$

or

$$q^{\sum_{i=1}^{N} x_i}(1-q)^{(N-\sum_{i=1}^{N} x_i)}\left(\frac{\sum_{i=1}^{N} x_i}{q} - \frac{N-\sum_{i=1}^{N} x_i}{1-q}\right) = 0$$

The solutions $q = 0, 1$ result in a minimum of $P(X;q)$. The maximum comes from

$$\frac{\sum_{i=1}^{N} x_i}{q} - \frac{N-\sum_{i=1}^{N} x_i}{1-q} = 0 \Rightarrow q = \frac{1}{N}\sum_{i=1}^{N} x_i$$

2.18

$$L(\mu) = \sum_{i=1}^{N} \ln p(x_k;\mu) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_k-\mu)^2$$

Where constants have been omitted

$$\frac{\partial L(\mu)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_k-\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{N} x_k - \frac{N}{\sigma^2}\mu$$

$$\frac{\partial^2 L(\mu)}{\partial \mu^2} = -\frac{N}{\sigma^2}$$

Thus, the Cramer - Rao bound for the variance of the ML estimate is $\frac{\sigma^2}{N}$. We know from theory, that the ML estimate of the mean is

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_k$$

which is on unbiased estimate of the true mean and the variance of the estimate (See chapter 5) is $\frac{\sigma^2}{N}$. Thus, for this case the ML estimate is efficient. If the unknown is the variance we have

$$L(\sigma^2) = \sum_{i=1}^{N} \ln p(x_k;\sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_k-\mu)^2 - \frac{N}{2}\ln\sigma^2$$

$$\frac{\partial L(\sigma^2)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{N}(x_k-\mu)^2$$

Equating the above to zero gives the estimator

$$\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(x_k-\mu)^2$$

which is the ML estimate of the variance if the mean is known. This is also an unbiased estimator

$$E[\hat{\sigma}^2] = \frac{1}{N}\sum_{i=1}^{N} E[(x_k-\mu)^2] = \sigma^2$$

For the Cramer-Rao bound we have

$$-E[\frac{\partial^2 L(\sigma^2)}{\partial^2 \sigma^2}] = -E[\frac{N}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{N}(x_k-\mu)^2]$$
$$-\frac{N}{2\sigma^4} + \frac{N}{\sigma^4} = \frac{N}{2\sigma^4}$$

Thus, the variance of any estimator of σ^2 is lower bounded by $\frac{2\sigma^4}{N}$. It turns out that the ML estimate of the variance of a normal distribution is also efficient. Indeed, we have

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_k-\mu)^2 \Longrightarrow$$
$$\frac{N\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^{N}(\frac{x_k-\mu}{\sigma})^2$$

However, each of the random variables $\frac{x_k-\mu}{\sigma}$ follows a $\mathcal{N}(0,1)$ and they are independent. Hence the random variable $\sum_{i=1}^{N}(\frac{x_k-\mu}{\sigma})^2$ follows a chi-square distribution with N degrees of freedom and its variance is equal to [Papoulis p117] $2N$. That is

$$var(\frac{N\hat{\sigma}^2}{\sigma^2}) = 2N$$

or

$$var(\hat{\sigma}^2) = \frac{2\sigma^4}{N}$$

Which is equal to the Cramer-Rao bound.

2.19 We shall focus on the simplest case where $\Sigma = \sigma^2 I$. The likelihood function is

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\mu}, \sigma^2) = \sum_{k=1}^{N} \ln p(\boldsymbol{x}_k; \boldsymbol{\mu}, \sigma^2)$$

$$= -\frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{k=1}^{N}(\boldsymbol{x}_k-\boldsymbol{\mu})^T(\boldsymbol{x}_k-\boldsymbol{\mu})$$

The unknown parameter is now $\boldsymbol{\theta}^T = [\boldsymbol{\mu}^T, \sigma^2]$

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} \\ \frac{\partial L(\boldsymbol{\theta})}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}\sum_{k=1}^{N}(\boldsymbol{x}_k - \boldsymbol{\mu}) \\ -\frac{N}{2\sigma^2} + \frac{\sum_{k=1}^{N}(\boldsymbol{x}_k-\boldsymbol{\mu})^T(\boldsymbol{x}_k-\boldsymbol{\mu})}{2\sigma^4} \end{bmatrix} = 0$$

Solving the above system w.r. to $\boldsymbol{\mu}$ and σ^2 results in

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

$$\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(\boldsymbol{x}_k - \boldsymbol{\mu})^T(\boldsymbol{x}_k - \boldsymbol{\mu})$$

2.20 Prove that the covariance estimate

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

is an unbiased one, where

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

We have that

$$\begin{aligned} E[\hat{\Sigma}] &= \frac{1}{N-1}\sum_{k=1}^{N} E\left[((\boldsymbol{x}_k - \boldsymbol{\mu}) - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}))((\boldsymbol{x}_k - \boldsymbol{\mu}) - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}))^T\right] \\ &= \frac{1}{N-1}\sum_{k=1}^{N} E\left[(\boldsymbol{x}_k - \boldsymbol{\mu})(\boldsymbol{x}_k - \boldsymbol{\mu})^T\right] + \frac{1}{N-1}\sum_{k=1}^{N} E\left[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T\right] - \\ &\quad \frac{1}{N-1}\sum_{k=1}^{N} E\left[(\boldsymbol{x}_k - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T\right] - \\ &\quad \frac{1}{N-1}\sum_{k=1}^{N} E\left[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\boldsymbol{x}_k - \boldsymbol{\mu})^T\right] \end{aligned} \tag{2.1}$$

However

$$\begin{aligned} E\left[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T\right] &= E\left[\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})\frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{x}_j - \boldsymbol{\mu})^T\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{N} E\left[(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T\right] \\ &= \frac{1}{N^2}N\Sigma = \frac{1}{N}\Sigma \end{aligned} \tag{2.2}$$

where independence among the samples has been assumed, i.e., $E\left[(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_j-\boldsymbol{\mu})^T\right]=\delta_{ij}\Sigma$. Following a similar path we end up with

$$E\left[(\boldsymbol{x}_k-\boldsymbol{\mu})(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^T\right]=E\left[(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})(\boldsymbol{x}_k-\boldsymbol{\mu})^T\right]=\frac{1}{N}\Sigma \tag{2.3}$$

Also

$$E\left[(\boldsymbol{x}_k-\boldsymbol{\mu})(\boldsymbol{x}_k-\boldsymbol{\mu})^T\right]=\Sigma \tag{2.4}$$

Combining Eqs (2.1)-(2.4) we get

$$E[\hat{\Sigma}]=\frac{N}{N-1}\Sigma+\frac{1}{N-1}\Sigma-\frac{1}{N-1}\Sigma-\frac{1}{N-1}\Sigma=\Sigma$$

Hence the estimate is an unbiased one.

2.21 Prove that the ML estimates of the mean value and the covariance matrix (Problem 2.19) can be computed recursively, i.e.,

$$\hat{\boldsymbol{\mu}}_{N+1}=\hat{\boldsymbol{\mu}}_N+\frac{1}{N+1}(\boldsymbol{x}_{N+1}-\hat{\boldsymbol{\mu}}_N)$$

and

$$\hat{\Sigma}_{N+1}=\frac{N}{N+1}\hat{\Sigma}_N+\frac{N}{(N+1)^2}(\boldsymbol{x}_{N+1}-\hat{\boldsymbol{\mu}}_N)(\boldsymbol{x}_{N+1}-\hat{\boldsymbol{\mu}}_N)^T$$

where the subscript in the notation of the estimates, $\hat{\boldsymbol{\mu}}_N$, $\hat{\Sigma}_N$ indicates the number of samples used for their computation.

From Problem 2.19 we know that

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{N+1} &= \frac{1}{N+1}\sum_{k=1}^{N+1}\boldsymbol{x}_k=\frac{1}{N+1}\sum_{k=1}^{N}\boldsymbol{x}_k+\frac{1}{N+1}\boldsymbol{x}_{N+1}\\ &= \frac{N}{N+1}\hat{\boldsymbol{\mu}}_N+\frac{1}{N+1}\boldsymbol{x}_{N+1}=\hat{\boldsymbol{\mu}}_N+\frac{1}{N+1}(\boldsymbol{x}_{N+1}-\hat{\boldsymbol{\mu}}_N)\end{aligned}$$

For the covariance matrix we get

$$\begin{aligned}\Sigma_{N+1} &= \frac{1}{N+1}\sum_{k=1}^{N+1}(\boldsymbol{x}_k-\hat{\boldsymbol{\mu}}_{N+1})(\boldsymbol{x}_k-\hat{\boldsymbol{\mu}}_{N+1})^T\\ &= \frac{1}{N+1}\sum_{k=1}^{N+1}\boldsymbol{x}_k\boldsymbol{x}_k^T-\frac{1}{N+1}\hat{\boldsymbol{\mu}}_{N+1}\sum_{k=1}^{N+1}\boldsymbol{x}_k^T-\end{aligned}$$