Problem sheet 2, solutions.

1(a). From the definition of covariance:

$$\begin{aligned}\mathrm{COV}(X,Y) &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

1(b). Since X and Y are independent, $P(X,Y) = P(X)P(Y)$. Let $\mathcal{X}$ be the set of all possible values of X and $\mathcal{Y}$ the set of all possible values of Y. Then:

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} xyP(X=x,Y=y) \\ &= \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} xyP(X=x)P(Y=y) \\ &= \sum_{x\in\mathcal{X}} xP(X=x) \times \sum_{y\in\mathcal{Y}} yP(Y=y) \\ &= \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

Substituting into (1) above, gives $\mathrm{COV}(X,Y) = 0$.

1(c). Since $Y = X^2$, $\mathbb{E}[XY] = \mathbb{E}[X^3] = 0$. Also, we can see that $\mathbb{E}[X] = 0$ and therefore $\mathbb{E}[X]\mathbb{E}[Y] = 0$. This gives $\mathrm{COV}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$.

2. Let $p(x)$ represent the pdf of RV X. Then:

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty x\,p(x)\,\mathrm{d}x \\ &= \int_0^a x\,p(x)\,\mathrm{d}x + \int_a^\infty x\,p(x)\,\mathrm{d}x\end{aligned}$$

Since $p(x)$ is a pdf, it is everywhere non-negative, so the first term on the RHS must be non-negative. This means:

$$\mathbb{E}[X] \geq \int_a^\infty x\,p(x)\,\mathrm{d}x$$

Since a is the lower bound on the integral above, we can write

$$\int_a^\infty x\,p(x)\,\mathrm{d}x \quad \geq \quad \int_a^\infty a\,p(x)\,\mathrm{d}x$$

which gives

$$\begin{aligned} \mathbb{E}[X] &\geq \int_a^\infty a\,p(x)\,\mathrm{d}x \\ &= a\int_a^\infty p(x)\,\mathrm{d}x \\ &= a\,P(X \geq a) \end{aligned}$$

from which the required result follows.

3. First, note that

$$P(|X - \mu_X| \geq a) \quad = \quad P((X-\mu_X)^2 \geq a^2)$$

Here, $(X - \mu_X)^2$ is a non-negative RV. Using the Markov inequality, we get:

$$\begin{aligned} P((X-\mu_X)^2 \geq a^2) &\leq \frac{\mathbb{E}[(X-\mu_X)^2]}{a^2} \\ &= \frac{\sigma_X^2}{a^2} \end{aligned}$$

as required.

4. If $\hat{\theta}_n$ is unbiased, we can write

$$P(|\hat{\theta}_n - \theta| \geq \epsilon) \quad = \quad P(|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]| \geq \epsilon)$$

Applying the Chebyshev inequality to the RHS, we get:

$$P(|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]| \geq \epsilon) \quad \leq \quad \frac{\mathrm{VAR}(\hat{\theta}_n)}{\epsilon^2}$$

From the RHS above we can see that if

$$\lim_{n\to\infty} \mathrm{VAR}(\hat{\theta}_n) \quad = \quad 0$$

the estimator converges in probability to θ, that is, it is consistent.

5. Let $\bar{X}_n$ denote the sample mean derived from n observations. This is easily shown to be unbiased. Using the Chebyshev inequality:

$$P(|\bar{X}_n - \mu_X| \geq \epsilon) \quad \leq \quad \frac{\mathrm{VAR}(\bar{X}_n)}{\epsilon^2}$$

But:

$$\begin{aligned}\mathrm{VAR}(\bar{X}_n) &= \mathrm{VAR}\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) \\ &= \frac{\sigma_X^2}{n}\end{aligned}$$

Therefore

$$P(|\bar{X}_n - \mu_X| \geq \epsilon) \quad \leq \quad \frac{\sigma_X^2}{n\epsilon^2}$$

and

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu_X| \geq \epsilon) = 0$$

which means $\bar{X}_n$ converges in probability to the true mean μ_X, as required.

6(a). Log-likelihood:

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad = \quad -\frac{dn}{2}\log(2\pi) - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \boldsymbol{\mu})$$

6(b). We proceed in two steps: we first treat $\boldsymbol{\Sigma}$ as fixed, and maximize $\mathcal{L}$ to get a value $\hat{\boldsymbol{\mu}}(\boldsymbol{\Sigma})$ which maximizes $\mathcal{L}$ for a given matrix parameter $\boldsymbol{\Sigma}$. Taking the derivative of the $\mathcal{L}$ wrt vector $\boldsymbol{\mu}$, we get:

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\mu}}\mathcal{L} \quad = \quad (\Sigma^{-1}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu}))^T$$

Setting the derivative to zero, taking the transpose of both sides and pre-multiplying by $\boldsymbol{\Sigma}$, we get:

$$\mathbf{0} \quad = \quad \sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})$$

Solving for $\boldsymbol{\mu}$:

$$\begin{aligned} \hat{\boldsymbol{\mu}}(\boldsymbol{\Sigma}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \\ &= \bar{\mathbf{X}} \end{aligned}$$

Since this solution does not depend on $\boldsymbol{\Sigma}$, $\bar{\mathbf{X}}$ is the maximum likelihood estimator of $\boldsymbol{\mu}$ for any $\boldsymbol{\Sigma}$. To obtain $\hat{\boldsymbol{\Sigma}}$ we plug $\hat{\boldsymbol{\mu}}(\boldsymbol{\Sigma}) = \bar{\mathbf{X}}$ into the log-likelihood to obtain

$$-\frac{dn}{2}\log(2\pi) - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})^T \boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}) \tag{1}$$

and maximize this function wrt $\boldsymbol{\Sigma}$.

We first introduce a sample covariance matrix $\mathbf{S}$ defined as follows:

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$$

This allows us to re-write the quadratic form in (1) as a matrix trace:

$$\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})^T \boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}) = n\,\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of its matrix argument.

This in turn allows us to write the derivative of (1) wrt $\boldsymbol{\Sigma}$ as follows:

$$-\frac{n}{2}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Sigma}}\log(|\boldsymbol{\Sigma}|) - \frac{n}{2}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Sigma}}\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})$$

At this point we make use of two useful matrix derivatives (these can be found in Appendix C of Bishop and the note "Matrix Identities" by Roweis, available on the course website):

$$\begin{aligned} \frac{\partial}{\partial \mathbf{A}}\log(|\mathbf{A}|) &= (\mathbf{A}^{-1})^T \\ \frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{X}^{-1}\mathbf{A}) &= -\mathbf{X}^{-1}\mathbf{A}^T\mathbf{X}^{-1} \end{aligned}$$

This gives the derivative (2) in the following form (where we make use of the fact that both $\boldsymbol{\Sigma}^{-1}$ and $\mathbf{S}$ are symmetric):

$$-\frac{n}{2}\boldsymbol{\Sigma}^{-1} + \frac{n}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}$$

Setting to zero and solving, we get:

$$
\begin{aligned}
\hat{\mathbf{\Sigma}} &= \mathbf{S} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T
\end{aligned}
$$