Problem 1

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$, then the mean value of this random vector is $\bar{\boldsymbol{X}} = \left(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n\right)^T$.
The covariance matrix of $\boldsymbol{X}$ is calculated by:

$$\boldsymbol{C}_X = \left\langle \left(X - \bar{X}\right)\left(X - \bar{X}\right)^T \right\rangle$$

Let $\boldsymbol{M}$ be the constant matrix, then :

$$\boldsymbol{Y} = \boldsymbol{M} \cdot \boldsymbol{X}$$

Then, the covariance of $\boldsymbol{Y}$ is calculated by:

$$\begin{aligned}
\boldsymbol{C}_Y &= \left\langle \left(Y - \bar{Y}\right)\left(Y - \bar{Y}\right)^T \right\rangle \\
&= \left\langle \left(MX - M\bar{X}\right)\left(MX - M\bar{X}\right)^T \right\rangle \\
&= \left\langle M\left(X - \bar{X}\right)\left(X - \bar{X}\right)^T M^T \right\rangle \\
&= M\left\langle \left(X - \bar{X}\right)\left(X - \bar{X}\right)^T \right\rangle M^T \\
&= MC_X M^T
\end{aligned}$$

Problem 2

Method 1

The logarithm of the PDF is :

$$\ln p = -\frac{n}{2}\ln|2\pi\boldsymbol{C}| - \frac{1}{2}\left[\left(\boldsymbol{x} - \boldsymbol{\mu}\left(\boldsymbol{\theta}_{true}\right)\right)^T \cdot \boldsymbol{C}^{-1} \cdot \left(\boldsymbol{x} - \boldsymbol{\mu}\left(\boldsymbol{\theta}_{true}\right)\right)\right]$$

Then, the first order derivative w.r.t θ is :

$$\begin{aligned}
\frac{\partial \ln p}{\partial \theta_i} = &\frac{1}{2}\left[\left(\frac{\partial \boldsymbol{\mu}}{\partial \theta_i}\right)^T \cdot \boldsymbol{C}^{-1} \cdot \left(\boldsymbol{x} - \boldsymbol{\mu}\left(\boldsymbol{\theta}_{true}\right)\right)\right] \\
&+ \frac{1}{2}\left[\left(\boldsymbol{x} - \boldsymbol{\mu}\left(\boldsymbol{\theta}_{true}\right)\right)^T \cdot \boldsymbol{C}^{-1} \cdot \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta_i}}\right)\right]
\end{aligned}$$

Then, the second derivative is :

$$
\begin{aligned}
\frac{\partial^2 \ln p}{\partial\theta_i\partial\theta_j} = & \frac{1}{2}\left[\left(\frac{\partial^2 \boldsymbol{\mu}}{\partial\theta_i\partial\theta_j}\right)^T \cdot \boldsymbol{C}^{-1} \cdot \left(\boldsymbol{x} - \boldsymbol{\mu}\left(\boldsymbol{\theta}_{true}\right)\right)\right] \\
& - \frac{1}{2}\left[\left(\frac{\partial \boldsymbol{\mu}}{\partial\theta_i}\right)^T \cdot \boldsymbol{C}^{-1} \cdot \left(\frac{\partial \boldsymbol{\mu}}{\partial\theta_j}\right)\right] \\
& - \frac{1}{2}\left[\left(\frac{\partial \boldsymbol{\mu}}{\partial\theta_j}\right)^T \cdot \boldsymbol{C}^{-1} \cdot \left(\frac{\partial \boldsymbol{\mu}}{\partial\theta_i}\right)\right] \\
& + \frac{1}{2}\left[\left(\boldsymbol{x} - \boldsymbol{\mu}\left(\boldsymbol{\theta}_{true}\right)\right)^T \cdot \boldsymbol{C}^{-1} \cdot \left(\frac{\partial^2 \boldsymbol{\mu}}{\partial\theta_i\partial\theta_j}\right)\right]
\end{aligned}
$$

Then, the $i, j-th$ element in Fisher Matrix is the expectation value of the above equation:

$$
F_{ij} = -\left\langle \frac{\partial^2 \ln p}{\partial\theta_i\partial\theta_j} \right\rangle = 0 - \frac{1}{2}\left\langle\left[\left(\frac{\partial \boldsymbol{\mu}}{\partial\theta_i}\right)^T \cdot \boldsymbol{C}^{-1} \cdot \left(\frac{\partial \boldsymbol{\mu}}{\partial\theta_j}\right)\right]\right\rangle - \frac{1}{2}\left\langle\left[\left(\frac{\partial \boldsymbol{\mu}}{\partial\theta_j}\right)^T \cdot \boldsymbol{C}^{-1} \cdot \left(\frac{\partial \boldsymbol{\mu}}{\partial\theta_i}\right)\right]\right\rangle + 0
$$

Since this two terms are symmetric, we then have:

$$
F_{ij} = -\left\langle \frac{\partial^2 \ln p}{\partial\theta_i\partial\theta_j} \right\rangle = -\left\langle\left[\left(\frac{\partial \boldsymbol{\mu}}{\partial\theta_i}\right)^T \cdot \boldsymbol{C}^{-1} \cdot \left(\frac{\partial \boldsymbol{\mu}}{\partial\theta_j}\right)\right]\right\rangle
$$

Method 2

Some properties that would be used in calculating the gradient of matrix:

> $$
> \begin{gathered}
> x = \operatorname{tr}(x) \\
> \operatorname{tr}\left(c_1\boldsymbol{A} + c_2\boldsymbol{B}\right) = c_1\operatorname{tr}(\boldsymbol{A}) + c_2\operatorname{tr}(\boldsymbol{B}) \\
> \operatorname{tr}(\boldsymbol{A}) = \operatorname{tr}\left(\boldsymbol{A}^T\right) \\
> \langle\boldsymbol{A}, \boldsymbol{B}\rangle = \operatorname{tr}\left(\boldsymbol{A}^T\boldsymbol{B}\right) \\
> \operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA}) \\
> \operatorname{tr}(\boldsymbol{ABC}) = \operatorname{tr}(\boldsymbol{BCA}) = \operatorname{tr}(\boldsymbol{CAB}) \\
> \mathrm{d}\left(\boldsymbol{F}(\boldsymbol{X})\boldsymbol{G}(\boldsymbol{X})\right) = \mathrm{d}\left(\boldsymbol{F}(\boldsymbol{X})\right)\boldsymbol{G}(\boldsymbol{X}) + \boldsymbol{F}(\boldsymbol{X})\mathrm{d}\left(\boldsymbol{G}(\boldsymbol{X})\right) \\
> \mathrm{d}\boldsymbol{F}^T{}_{p\times q}(\boldsymbol{X}) = \left(\mathrm{d}\boldsymbol{F}_{p\times q}(\boldsymbol{X})\right)^T
> \end{gathered}
> $$

> And the most important one:
>
> $$
> \operatorname{tr}\left(\mathrm{d}\boldsymbol{F}_{p\times q}\right) = \operatorname{tr}\left(\left(\frac{\mathrm{d}\boldsymbol{F}_{p\times q}}{\mathrm{d}\boldsymbol{X}}\right)^T \mathrm{d}\boldsymbol{X}\right)
> $$
>
> So, the derivative of the matrix is then the transpose of the left part.

$$\begin{aligned}\operatorname{tr}(\mathrm{d}\ln p) &= \operatorname{tr}\left(\frac{1}{2}\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathrm{d}\boldsymbol{\theta}\right)^T\cdot\boldsymbol{C}^{-1}\cdot[\boldsymbol{X}-\boldsymbol{\mu}(\boldsymbol{\theta})]+\frac{1}{2}[\boldsymbol{X}-\boldsymbol{\mu}(\boldsymbol{\theta})]^T\cdot\boldsymbol{C}^{-1}\cdot\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathrm{d}\boldsymbol{\theta}\right)\right)\\
&= \operatorname{tr}\left(\frac{1}{2}\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathrm{d}\boldsymbol{\theta}\right)^T\cdot\boldsymbol{C}^{-1}\cdot[\boldsymbol{X}-\boldsymbol{\mu}(\boldsymbol{\theta})]\right)+\operatorname{tr}\left(\frac{1}{2}[\boldsymbol{X}-\boldsymbol{\mu}(\boldsymbol{\theta})]^T\cdot\boldsymbol{C}^{-1}\cdot\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathrm{d}\boldsymbol{\theta}\right)\right)\\
&= \frac{1}{2}\operatorname{tr}\left([\boldsymbol{X}-\boldsymbol{\mu}(\boldsymbol{\theta})]^T\cdot\left(\boldsymbol{C}^{-1}\right)^T\cdot\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathrm{d}\boldsymbol{\theta}\right)\right)+\frac{1}{2}\operatorname{tr}\left([\boldsymbol{X}-\boldsymbol{\mu}(\boldsymbol{\theta})]^T\cdot\boldsymbol{C}^{-1}\cdot\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathrm{d}\boldsymbol{\theta}\right)\right)\end{aligned}$$

So, the gradient of $\ln p$ is :

$$\frac{\mathrm{d}\ln p}{\mathrm{d}\boldsymbol{\theta}}=\frac{1}{2}\left(\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T\cdot\left(\left(\boldsymbol{C}^{-1}\right)^T+\boldsymbol{C}^{-1}\right)\cdot[\boldsymbol{X}-\boldsymbol{\mu}(\boldsymbol{\theta})]\right)$$

then, take the second derivative of this :

$$\begin{aligned}\mathrm{d}\left(\frac{\mathrm{d}\ln p}{\mathrm{d}\boldsymbol{\theta}}\right) &= \frac{1}{2}\mathrm{d}\left(\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T\right)\cdot\left[\left(\boldsymbol{C}^{-1}\right)^T+\boldsymbol{C}^{-1}\right]\cdot[\boldsymbol{X}-\boldsymbol{\mu}(\boldsymbol{\theta})]\\
&\quad-\frac{1}{2}\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T\cdot\left[\left(\boldsymbol{C}^{-1}\right)^T+\boldsymbol{C}^{-1}\right]\cdot\mathrm{d}\boldsymbol{\mu}(\boldsymbol{\theta})\end{aligned}$$

Then take the expectation value of this equation, the first term is 0, so, we have:

$$\begin{aligned}\left\langle\mathrm{d}\left(\frac{\mathrm{d}\ln p}{\mathrm{d}\boldsymbol{\theta}}\right)\right\rangle &= \left\langle-\frac{1}{2}\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T\cdot\left[\left(\boldsymbol{C}^{-1}\right)^T+\boldsymbol{C}^{-1}\right]\cdot\mathrm{d}\boldsymbol{\mu}(\boldsymbol{\theta})\right\rangle\\
&= \left\langle-\frac{1}{2}\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T\cdot\left[\left(\boldsymbol{C}^{-1}\right)^T+\boldsymbol{C}^{-1}\right]\cdot\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathrm{d}\boldsymbol{\theta}\right\rangle\end{aligned}$$

Therefore:

$$\left\langle\mathrm{d}\left(\frac{\mathrm{d}^2\ln p}{\mathrm{d}\boldsymbol{\theta}^2}\right)\right\rangle=\left\langle-\frac{1}{2}\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T\cdot\left[\left(\boldsymbol{C}^{-1}\right)^T+\boldsymbol{C}^{-1}\right]\cdot\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\rangle$$

Since the covariance matrix is symmetric, the final result is then:

$$\left\langle\frac{\mathrm{d}^2\ln p}{\mathrm{d}\boldsymbol{\theta}^2}\right\rangle=-\left\langle\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T\cdot\boldsymbol{C}^{-1}\cdot\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\rangle$$

Then, we have:

$$\boldsymbol{F}_{ij}=\left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\theta_i}\right)^T\cdot\boldsymbol{C}^{-1}\cdot\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\theta_j}$$

When $n=1=m$,

$$F=\left(\frac{\partial\mu}{\partial\theta}\right)^2\frac{1}{\sigma^2}$$

So, higher value of F means that is contains more information. There are two explanations:

1. when the derivative is large, that means μ is sensitive to θ, a small deviation from the true value could lead to a big difference.

2. when the σ variance is small, that means the distribution is more localized around the mean value.

Problem 3

Take the derivative of $\chi^2(a)$:

$$\begin{aligned}
\chi^2(a) &= [\boldsymbol{x} - a\boldsymbol{v}]^T \boldsymbol{C}^{-1} [\boldsymbol{x} - a\boldsymbol{v}] \\
\mathrm{d}\left(\chi^2(a)\right) &= [-\boldsymbol{v}\mathrm{d}a]^T \boldsymbol{C}^{-1} [\boldsymbol{x} - a\boldsymbol{v}] + [\boldsymbol{x} - a\boldsymbol{v}]^T \boldsymbol{C}^{-1} [-\boldsymbol{v}\mathrm{d}a] \\
&= \left\{[-\boldsymbol{v}]^T \boldsymbol{C}^{-1} [\boldsymbol{x} - a\boldsymbol{v}]\mathrm{d}a + [\boldsymbol{x} - a\boldsymbol{v}]^T \boldsymbol{C}^{-1} [-\boldsymbol{v}]\mathrm{d}a\right\} \\
&= -2\left[\boldsymbol{v}^T \boldsymbol{C}^{-1} [\boldsymbol{x} - a\boldsymbol{v}]\right] da \\
\frac{\mathrm{d}\left(\chi^2(a)\right)}{\mathrm{d}a} &= -2\left[\boldsymbol{v}^T \boldsymbol{C}^{-1} [\boldsymbol{x} - a\boldsymbol{v}]\right]
\end{aligned}$$

Since the transpose of a scalar is still itself, we can combine the two term in the third line.

To get the a_{best} that makes the χ^2 minimized, make the derivative 0 :

$$\begin{aligned}
\frac{\mathrm{d}\left(\chi^2(a)\right)}{\mathrm{d}a} &= -2\left[\boldsymbol{v}^T \boldsymbol{C}^{-1} [\boldsymbol{x} - a\boldsymbol{v}]\right] = 0 \\
a &= \frac{\boldsymbol{v}^T \boldsymbol{C}^{-1} \boldsymbol{x}}{\boldsymbol{v}^T \boldsymbol{C}^{-1} \boldsymbol{v}}
\end{aligned}$$

For a random variable A_{best}, its variance is :

$$\begin{aligned}
A_{best} &= \frac{\boldsymbol{X}^T \boldsymbol{C}^{-1} \boldsymbol{V}}{\boldsymbol{V}^T \boldsymbol{C}^{-1} \boldsymbol{V}} \\
\mathrm{Cov}\left(A_{best}\right) &= \mathrm{Cov}\left(\frac{\boldsymbol{X}^T \boldsymbol{C}^{-1} \boldsymbol{V}}{\boldsymbol{V}^T \boldsymbol{C}^{-1} \boldsymbol{V}}\right) \\
&= \frac{1}{\left(\boldsymbol{V}^T \boldsymbol{C}^{-1} \boldsymbol{V}\right)^2} \mathrm{Cov}\left(\boldsymbol{X}^T \boldsymbol{C}^{-1} \boldsymbol{V}\right) \\
&= \frac{1}{\left(\boldsymbol{V}^T \boldsymbol{C}^{-1} \boldsymbol{V}\right)^2} \mathrm{Cov}\left(\boldsymbol{V}^T \boldsymbol{C}^{-1} \boldsymbol{X}\right) \\
&= \frac{1}{\left(\boldsymbol{V}^T \boldsymbol{C}^{-1} \boldsymbol{V}\right)^2} \left(\boldsymbol{V}^T \boldsymbol{C}^{-1}\right) \mathrm{Cov}\left(\boldsymbol{X}\right) \left(\boldsymbol{V}^T \boldsymbol{C}^{-1}\right)^T \\
&= \frac{1}{\left(\boldsymbol{V}^T \boldsymbol{C}^{-1} \boldsymbol{V}\right)^2} \left(\boldsymbol{V}^T \boldsymbol{C}^{-1}\right) \boldsymbol{C} \boldsymbol{C}^{-1} \boldsymbol{V} \\
&= \frac{1}{\boldsymbol{V}^T \boldsymbol{C}^{-1} \boldsymbol{V}}
\end{aligned}$$

So, the variance of A_{best} is then $\mathrm{Var}\left(A_{best}\right) = \boldsymbol{V}^{-1} \boldsymbol{C} \left(\boldsymbol{V}^T\right)^{-1}$.

Fisher matrix for the gaussian distribution is :

$$\boldsymbol{F} = \left(\frac{\partial \boldsymbol{\mu}\left(\vec{\theta}\right)}{\partial \vec{\theta}}\right)^T \cdot \boldsymbol{C}^{-1} \cdot \left(\frac{\partial \boldsymbol{\mu}\left(\vec{\theta}\right)}{\partial \vec{\theta}}\right)$$

Here, the variable is a_{best}, so the Fisher matrix becomes:

$$\begin{aligned} \boldsymbol{F} &= \left(\frac{\partial a \boldsymbol{v}}{\partial a}\right)^T \cdot \boldsymbol{C}^{-1} \cdot \left(\frac{\partial a \boldsymbol{v}}{\partial a}\right) \\ &= \boldsymbol{v}^T \cdot \boldsymbol{C}^{-1} \cdot \boldsymbol{v} \end{aligned}$$

So, the inverse of Fisher matrix is :

$$\boldsymbol{F}^{-1} = \boldsymbol{V}^{-1} \cdot \boldsymbol{C} \cdot \left(\boldsymbol{V}^T\right)^{-1}$$

So, this is the same as the inverse of Fisher matrix.

> Cramer-Rao-bound :
>
> $$\operatorname{var}\left(\hat{\theta}\right) \geq \frac{1}{F\left(\theta\right)}$$
>
> The precision of any unbiased estimation is at most the Fisher information matrix.

The estimator is statistically efficient, making the best use of the available information in estimating the parameter of interest.