# CS 215 Assignment 3 Report

Diwan Anuj Jitendra - 170070005 Soumya Chatterjee - 170070010

October 21, 2018

## Problem 3

### 0.1 Part 1

We wish to fit a multivariate Gaussian to the data. Thus the likelihood is of the form:

$$f_{\mathbf{X}}(x_1, \ldots, x_k|\boldsymbol{\mu}, \boldsymbol{C}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{C}|}}$$

In this problem, $k = 2$ thus, $\boldsymbol{x}, \boldsymbol{\mu}$ are $2 \times 1$ and $\boldsymbol{C}, \boldsymbol{C^{-1}}$ are $2 \times 2$. Since this is a non-degenerate Gaussian one can assume that $\boldsymbol{C}$ is symmetric. Thus, we get the likelihood:

$$f_{\mathbf{X}}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{C}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^2|\boldsymbol{C}|}}$$

Hence the log likelihood, LL, is:

$$\mathrm{LL}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{C}) = -\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{C}|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

Now, let us find the derivatives of the LL w.r.t. $\mu$ and $C$.

$$\begin{aligned}
\frac{\partial \mathrm{LL}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{C})}{\partial \boldsymbol{\mu}} &= -\frac{1}{2}\frac{\partial(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} \\
&= -\frac{1}{2}\frac{\partial((\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))}{\partial \boldsymbol{x}-\boldsymbol{\mu}} \times \frac{\partial(\boldsymbol{x}-\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}
\end{aligned}$$

Using the reference given in the problem statement, we know that:

$$d(a^TY^TDYb) = (D^TYab^T + DYba^T) :^T dY :$$

Let us take $a$ and $b$ as the identity matrix, $Y = \boldsymbol{x}-\boldsymbol{\mu}$ and $D = \boldsymbol{C}^{-1}$. The above formula is valid whenever $a, b, D$ are independent of the variable being differentiated with respect to(in this case, $\boldsymbol{x}-\boldsymbol{\mu}$ , which is true here. Thus,

$$\begin{aligned}
\frac{\partial a^TY^TDYb}{\partial Y} &= (D^TY + DY) :^T \frac{\partial Y :}{\partial Y} \\
&= (D^TY + DY) :^T \frac{\partial Y}{\partial Y} \quad \text{Since Y is a column vector} \\
&= ((D^T + D)Y)^T \quad D \text{ is } 2 \times 2 \text{ and } Y \text{ is } 2 \times 1 \text{ so their product is } 2 \times 1\text{, already a column vector}
\end{aligned}$$

Thus, putting into original equation:

$$\begin{aligned}
\frac{\partial \text{LL}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{C})}{\partial \boldsymbol{\mu}} &= -\frac{1}{2}((\boldsymbol{C}^{-1^T}+\boldsymbol{C}^{-1})(\boldsymbol{x}-\boldsymbol{\mu}))^T \times \frac{\partial(\boldsymbol{x}-\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} \\
&= -((\boldsymbol{C}^{-1})(\boldsymbol{x}-\boldsymbol{\mu}))^T \times (-1) \\
&= ((\boldsymbol{C}^{-1})(\boldsymbol{x}-\boldsymbol{\mu}))^T \\
&= (\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{C}^{-1})^T \\
&= (\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{C}^{-1}) \quad \text{By symmetry of } \boldsymbol{C}^{-1}
\end{aligned}$$

And for the derivative w.r.t. C :

$$\begin{aligned}
\frac{\partial \text{LL}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{C})}{\partial \boldsymbol{C}} &= -\frac{1}{2}\frac{\partial \log(|\boldsymbol{C}|)}{\partial \boldsymbol{C}} - \frac{1}{2}\frac{\partial(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\partial \boldsymbol{C}} \\
&= -\frac{1}{2|\boldsymbol{C}|}\frac{\partial |\boldsymbol{C}|}{\partial \boldsymbol{C}} - \frac{1}{2}\frac{\partial(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\partial \boldsymbol{C}}
\end{aligned}$$

Using the given reference, we know that:

$$\begin{aligned}
\frac{\mathrm{d}|\boldsymbol{C}|}{\mathrm{d}\boldsymbol{C}} &= |\boldsymbol{C}|(\boldsymbol{C}^{-1})^T \\
&= |\boldsymbol{C}|(\boldsymbol{C}^{-1}) \quad \text{Since it is symmetric}
\end{aligned}$$

thus,

$$= -\frac{1}{2}(\boldsymbol{C}^{-1}) - \frac{1}{2}\frac{\partial(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\partial \boldsymbol{C}}$$

Using the given reference, we know that:

$$d(a^TX^{-1}b) = -(X^{-T}ab^TX^{-T})^T dX$$

thus taking $X = \boldsymbol{C}$ and $b = a = \boldsymbol{x}-\boldsymbol{\mu}$, we get:

$$\begin{aligned}
\frac{\partial(a^TX^{-1}b)}{\partial X} &= -(X^{-T}aa^TX^{-T})^T\frac{\partial X}{\partial X} \\
&= -(X^{-T}aa^TX^{-T})^T \\
&= -((X^{-T})^T a^{T^T} a^T (X^{-T})^T) \\
&= -((X^{-1})aa^T(X^{-1}))
\end{aligned}$$

Thus substituting this result in the original equation,

$$= -\frac{1}{2}(\boldsymbol{C}^{-1}) + \frac{1}{2}(\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{C}^{-1})$$

We must consider the log joint likelihood i.e:

$$\text{JLL}(\{\boldsymbol{x}_i\}|\boldsymbol{\mu},\boldsymbol{C}) = \sum_{i=1}^{n}\text{LL}(\boldsymbol{x}_i|\boldsymbol{\mu},\boldsymbol{C})$$

and set its partial derivatives to zero matrix/vector. Thus,

$$\frac{\partial \text{JLL}(\{\boldsymbol{x_i}\}|\boldsymbol{\mu}, \boldsymbol{C})}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{n} \frac{\partial \text{LL}(\boldsymbol{x_i}|\boldsymbol{\mu}, \boldsymbol{C})}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{n} (\boldsymbol{x_i} - \boldsymbol{\mu})^T (\boldsymbol{C}^{-1}) = 0$$

$$\sum_{i=1}^{n} (\boldsymbol{x_i} - \boldsymbol{\mu})^T (\boldsymbol{C}^{-1})\boldsymbol{C} = 0 \times \boldsymbol{C}$$

$$\sum_{i=1}^{n} (\boldsymbol{x_i} - \boldsymbol{\mu})^T = 0$$

$$\sum_{i=1}^{n} (\boldsymbol{x_i} - \boldsymbol{\mu}) = 0$$

$$\sum_{i=1}^{n} (\boldsymbol{x_i}) = n\boldsymbol{\mu}$$

$$\boldsymbol{\mu} = \frac{\sum_{i=1}^{n} (\boldsymbol{x_i})}{n}$$

This quantity is by definition the sample mean. Each sample point is of the form $x(\theta) = (r\cos(\theta), r\sin(\theta))$ where $r$ is a parameter and $\theta$ is the variable. A uniform distribution on a circle of radius $r$ is equivalent to a uniform distribution of $\theta$ on $[0, 2\pi)$ transformed by the function $f : \mathbb{R} \to \mathbb{R}^2$ as $(r\cos(\theta), r\sin(\theta))$. Thus, the expectation i.e true mean of this uniform distribution on the circle is:

$$\frac{1}{2\pi} \times \int_{\theta=0}^{2\pi} \begin{bmatrix} r\cos(\theta) \\ r\sin(\theta)) \end{bmatrix} \mathrm{d}\theta = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

And for a large sample size $n$, the sample mean as above is very close to the true mean. Thus, the estimated mean should be:

$$\mu \sim \text{E}(\mu) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

And,

$$\frac{\partial \text{JLL}(\{\boldsymbol{x_i}\}|\boldsymbol{\mu}, \boldsymbol{C})}{\partial \boldsymbol{C}} = \sum_{i=1}^{n} \frac{\partial \text{LL}(\boldsymbol{x_i}|\boldsymbol{\mu}, \boldsymbol{C})}{\partial \boldsymbol{C}} = \sum_{i=1}^{n} -\frac{1}{2}(\boldsymbol{C}^{-1}) + \frac{1}{2}(\boldsymbol{C}^{-1}(\boldsymbol{x_i} - \boldsymbol{\mu})(\boldsymbol{x_i} - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}) = 0$$

$$\sum_{i=1}^{n} -\frac{1}{2} + \frac{1}{2}(\boldsymbol{C}^{-1}(\boldsymbol{x_i} - \boldsymbol{\mu})(\boldsymbol{x_i} - \boldsymbol{\mu})^T) = 0 \quad \text{(Multiplying on the right by } C\text{)}$$

$$\sum_{i=1}^{n} -1 + (\boldsymbol{C}^{-1}(\boldsymbol{x_i} - \boldsymbol{\mu})(\boldsymbol{x_i} - \boldsymbol{\mu})^T) = 0$$

$$\sum_{i=1}^{n} -\boldsymbol{C} + ((\boldsymbol{x_i} - \boldsymbol{\mu})(\boldsymbol{x_i} - \boldsymbol{\mu})^T) = 0$$

$$\sum_{i=1}^{n} ((\boldsymbol{x_i} - \boldsymbol{\mu})(\boldsymbol{x_i} - \boldsymbol{\mu})^T) = n\boldsymbol{C}$$

$$\boldsymbol{C} = \frac{\sum_{i=1}^{n} ((\boldsymbol{x_i} - \boldsymbol{\mu})(\boldsymbol{x_i} - \boldsymbol{\mu})^T)}{n}$$

Here, we must set $\boldsymbol{\mu} = \frac{\sum_{i=1}^{n} \boldsymbol{x_i}}{n}$ Similar to the previous, let us estimate $\boldsymbol{C}$ by taking its expectation(since sample size is large, this is very close to the true value.)

Clearly the quantity above is the biased sample covariance matrix and its expectation is hence equal to the covariance matrix of the uniform distribution on the circle off by a factor of $\frac{n-1}{n}$. Because $n$ is very large, this factor can be ignored. For the latter, we can compute it as follows:

$$
\begin{aligned}
\text{Cov(Uniform on circle)} &= \begin{bmatrix} E((x_1-0)(x_1-0)) & E((x_1-0)(x_2-0)) \\ E((x_1-0)(x_2-0)) & E((x_2-0)(x_2-0)) \end{bmatrix} \\
&= \int_{\theta=0}^{2\pi} \begin{bmatrix} {x_1}^2 & x_1x_2 \\ x_1x_2 & {x_2}^2 \end{bmatrix} \times \frac{1}{2\pi} \mathrm{d}\theta \\
&= \int_{\theta=0}^{2\pi} \begin{bmatrix} r^2\cos^2\theta & r^2\cos\theta\sin\theta \\ r^2\cos\theta\sin\theta & r^2\sin^2\theta \end{bmatrix} \times \frac{1}{2\pi} \mathrm{d}\theta \\
&= \frac{r^2}{2\pi} \int_{\theta=0}^{2\pi} \begin{bmatrix} \cos^2\theta & \cos\theta\sin\theta \\ \cos\theta\sin\theta & \sin^2\theta \end{bmatrix} \mathrm{d}\theta \\
&= \frac{r^2}{2\pi} \begin{bmatrix} \pi & 0 \\ 0 & \pi \end{bmatrix} \\
&= \begin{bmatrix} \frac{r^2}{2} & 0 \\ 0 & \frac{r^2}{2} \end{bmatrix}
\end{aligned}
$$

Thus, finally, we have:

$$
\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and } \boldsymbol{C} = \begin{bmatrix} \frac{r^2}{2} & 0 \\ 0 & \frac{r^2}{2} \end{bmatrix}
$$

## 0.2 Part 2

The mode of the Gaussian is at $(0,0)$ since the exponential is a decreasing function with the maxima only at $x = \mu = (0,0)$.(For a Gaussian, mean and mode are the same). Therefore I conclude that the Gaussian **does not** fit the data well and isn't a good model for the uniform distribution on a circle. This is because the uniform distribution on a circle has probability mass only on the circle and zero everywhere else(and its mode is the whole circle), whereas the Gaussian claims that the probability mass is maximum at the centre and much lower at the circle, which is grossly incorrect. One can also see that for large $r$, the Gaussian approximately gives the uniform distribution over a disk(including the interior) which is very different from the uniform distribution over a circle.

## 0.3 Part 3

How to run: Run `p3.m` on MATLAB. To change value of $r$, change it at the top of the code and while outputting it.
Using a MATLAB script, the following results were obtained for $N = 10^7$:

```
Taking r=5
Experimental Mean =
    0.0003    0.0019

Covariance Matrix =
   12.5001   -0.0001
   -0.0001   12.4999
```

which fits the theoretically predicted values very well:

```
Taking r=5
Theoretical Mean =
    0    0

Covariance Matrix =
   12.5         0
     0       12.5
```

Similarly,

```
Taking r=4
Experimental Mean =
    -0.0011    0.0007

Covariance Matrix =
    7.9996    0.0029
    0.0029    8.0004
```

which fits the theoretically predicted values very well:

```
Taking r=4
Theoretical Mean =
    0    0

Covariance Matrix =
     8         0
     0         8
```