CSCI 5521 Homework 2

John Nguyen

October 16, 2019

For this assignment, I will use the notation $(x_i, y_i)_{i=1,...,n}$ instead of using the notation $(x^t, r^t)_{t=1,...,n}$.

When I mention the log-likelihood, I will use the natural log: ln.

Problem 1

(a). Let $\chi = \{x_1, x_2, ..., x_n\}$ be the set of samples. Recall the given probability density function:

$$p(x|\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{x^2}{2\theta}\right)$$

The log likelihood is a monotonically increasing function, so the log-likelihood function maximizes at the same value as the likelihood function. Thus we calculate the log-likelihood:

$$\begin{aligned} L(\theta|\chi) &= \sum_{i=1}^{n} \ln(p(x_i|\theta) \\ &= \sum_{i=1}^{n} \left(\ln(1) - \ln(\theta) + \frac{1}{2}(-\ln(2) - \ln(\pi)) - \frac{x_i^2}{2\theta^2}\right) \\ &= \sum_{i=1}^{n} (\ln(1)) - \sum_{i=1}^{n} (\ln(\theta)) + \sum_{i=1}^{n} \left(\frac{1}{2}(-\ln(2) - \ln(\pi))\right) - \sum_{i=1}^{n} \left(\frac{x_i}{2\theta^2}\right) \\ &= n\ln(1) - n\log(\theta) + \frac{n}{2}(-\ln(2) - \ln(\pi)) - \frac{-\sum_{i=1}^{n} x_i}{2\theta^2} \end{aligned}$$

We now take the partial derivative of the log-likelihood function with respect to θ and simplify. Notice that most terms in the log-likelihood functions are independent of the value of θ, so most of them become 0 after you take the partial derivative.

$$\begin{aligned} \frac{\partial L(\theta|\chi)}{\partial\theta} &= \frac{\partial}{\partial\theta}\left[n\ln(1) - n\log(\theta) + \frac{n}{2}(-\ln(2) - \ln(\pi)) - \frac{-\sum_{i=1}^{n} x_i}{2\theta^2}\right] \\ &= \frac{\partial}{\partial\theta}[n\ln(1)] - \frac{\partial}{\partial\theta}[n\log(\theta)] + \frac{\partial}{\partial\theta}\left[\frac{n}{2}(-\ln(2) - \ln(\pi))\right] - \frac{\partial}{\partial\theta}\left[\frac{-\sum_{i=1}^{n} x_i}{2\theta^2}\right] \\ &= 0 - \frac{n}{\theta} + 0 + \sum_{i=1}^{n} \frac{x_i}{2\theta^3} \\ &= -\frac{n}{\theta} + \sum_{i=1}^{n} \frac{x_i}{\theta^3} \end{aligned}$$

We now set the partial derivative of the log-likelihood function of the given distribution equal to 0 and solve for θ.

$$-\frac{n}{\theta} + \sum_{i=1}^{n} \frac{x_i}{\theta^3} = 0 \implies -\theta^2 n + \sum_{i=1}^{n} x_i = 0 \implies \theta^2 n = \sum_{i=1}^{n} x_i \implies \theta^2 = \frac{\sum_{i=1}^{n} x_i}{n} \implies \theta = \sqrt{\frac{\sum_{i=1}^{n} x_i}{n}}$$

So the maximum likelihood estimator of θ is: $\hat{\theta} = \sqrt{\frac{\sum_{i=1}^{n} x_i}{n}}$.

(b). Let $\chi = \{x_1, x_2, ..., x_n\}$ be the set of samples. Recall the given probability density function:

$$p(x|\theta) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right)$$

The log likelihood is a monotonically increasing function, so the log-likelihood function maximizes at the same value as the likelihood function. Thus we calculate the log-likelihood:

$$\begin{aligned}
L(\theta|\chi) &= \sum_{i=1}^{n} \ln(p(x_i|\theta)) \\
&= \sum_{i=1}^{n} \left(\ln(1) - \ln(\theta) - \frac{x_i}{\theta}\right) \\
&= \sum_{i=1}^{n} \ln(1) - \sum_{i=1}^{n} \ln(\theta) - \sum_{i=1}^{n} \frac{x_i}{\theta} \\
&= n\ln(1) - n\ln(\theta) - \sum_{i=1}^{n} \frac{x_i}{\theta}
\end{aligned}$$

We now take the partial derivative of the log-likelihood function with respect to θ and simplify. Notice that most terms in the log-likelihood functions are independent of the value of θ, so most of them become 0 after you take the partial derivative.

$$\begin{aligned}
\frac{\partial L(\theta|\chi)}{\partial \theta} &= \frac{\partial}{\partial \theta}\left[n\ln(1) - n\ln(\theta) - \sum_{i=1}^{n} \frac{x_i}{\theta}\right] \\
&= n\frac{\partial}{\partial \theta}\ln(1) - n\frac{\partial}{\partial \theta}\ln(\theta) - \frac{\partial}{\partial \theta}\sum_{i=1}^{n} \frac{x_i}{\theta} \\
&= 0 - \frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} \\
&= -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2}
\end{aligned}$$

We now set the partial derivative of the log-likelihood function of the given distribution equal to 0 and solve for θ.

$$-\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} = 0 \implies -\theta n + \sum_{i=1}^{n} x_i = 0 \implies \theta n = \sum_{i=1}^{n} x_i \implies \theta = \frac{\sum_{i=1}^{n} x_i}{n}$$

So the maximum likelihood estimator of θ is $\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}$.

(c). Let $\chi = \{x_1, x_2, ..., x_n\}$ be the set of samples. Recall the given probability density function:

$$p(x|\theta) = \theta x^{\theta-1}$$

The log likelihood is a monotonically increasing function, so the log-likelihood function maximizes at the same value as the likelihood function. Thus we calculate the log-likelihood:

$$\begin{aligned}
L(\theta|\chi) &= \sum_{i=1}^{n} (\ln(p(x_i|\theta))) \\
&= \sum_{i=1}^{n} (\ln(\theta) + (\theta - 1)\ln(x_i)) \\
&= \sum_{i=1}^{n} \ln(\theta) + (\theta - 1)\sum_{i=1}^{n} \ln(x_i) \\
&= n\ln(\theta) + \theta\sum_{i=1}^{n} \ln(x_i) - \sum_{i=1}^{n} \ln(x_i)
\end{aligned}$$

We now take the partial derivative of the log-likelihood function with respect to θ and simplify. Notice that most terms in the log-likelihood functions are independent of the value of θ, so most of them become 0 after you take the partial derivative.

$$\begin{aligned}
\frac{\partial L(\theta|\chi)}{\partial\theta} &= \frac{\partial}{\partial\theta}\left[n\ln(\theta) + \theta\sum_{i=1}^{n} \ln(x_i) - \sum_{i=1}^{n} \ln(x_i)\right] \\
&= \frac{\partial}{\partial\theta}[n\ln(\theta)] + \frac{\partial}{\partial\theta}\left[\theta\sum_{i=1}^{n} \ln(x_i)\right] - \frac{\partial}{\partial\theta}\left[\sum_{i=1}^{n} \ln(x_i)\right] \\
&= \frac{n}{\theta} + \sum_{i=1}^{n} \ln(x_i)
\end{aligned}$$

We now set the partial derivative of the log-likelihood function of the given distribution equal to 0 and solve for θ.

$$\frac{n}{\theta} + \sum_{i=1}^{n} \ln(x_i) = 0 \implies n + \theta\sum_{i=1}^{n} \ln(x_i) = 0 \implies \theta = -\frac{n}{\sum_{i=1}^{n} \ln(x_i)}$$

So the maximum likelihood estimator of θ is $\hat{\theta} = -\frac{n}{\sum_{i=1}^{n} \ln(x_i)}$.

(d). Notice the bounds on x and θ:

$$0 \leq x \leq \theta \text{ and } 0 < \theta$$

Notice that the data $\chi = \{x_1, x_2, x_3, ..., x_n\}$ is a finite set of real numbers, so there exists an ordering of elements. Define a function $f : \{1, 2, 3, ..., n\} \to \{1, 2, 3, ..., n\}$ such that for the data $\chi = \{x_1, x_2, x_3, ..., x_n\}$,

$$x_{f(1)} \leq x_{f(2)} \leq x_{f(3)} \leq ... \leq x_{f(n)}$$

So $\max \chi = x_{f(n)}$. Lets calculate the likelihood function:

$$l(\theta|\chi) = \frac{1}{\theta^n}$$

So, $0 \leq x_{f(1)}$ and $\theta \geq x_{f(n)}$. Now lets compute the log-likelihood function:

$$L(\theta|\chi) = \sum_{i=1}^{n} \ln(\frac{1}{\theta})$$
$$= n \ln(\frac{1}{\theta})$$

We now find the derivative of $L(\theta|\chi)$ with respect to θ:

$$\frac{\partial L}{\partial \theta} = -\frac{n}{\theta} < 0$$

$-\frac{n}{\theta} < 0$ because $\theta > 0$ and $n > 0$. This derivative tells us $L(\theta|\chi) = \theta^{-n}$ is a decreasing function of θ, where $\theta \geq x_{f(n)}$. So $L(\theta|\chi)$ and $l(\theta|\chi)$ are maximized when $\theta = x_{f(n)}$. So the maximum likelihood estimator of θ is $\hat{\theta} = x_{f(n)}$.

Problem 2

(a). Let $\chi = \{x_1, x_2, x_3, ..., x_n\}$ be the samples, where $x_i \in \mathbb{R}^d$ for $i = 1, 2, 3, .., n$. Recall the given probability density function:

$$p(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

We calculate the log-likelihood function:

$$\begin{aligned}
L(\mu|\chi) &= \ln\left[\prod_{i=1}^{n}\left(\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]\right)\right] \\
&= \sum_{i=1}^{n}\left(\ln(1) - \frac{d}{2}(\ln(2)+\ln(\pi)) -]frac12\ln(|\Sigma|) - \frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)\right) \\
&= \sum_{i=1}^{n}\ln(1) - \frac{d}{2}\sum_{i=1}^{n}[\ln(2)+\ln(\pi)] - \frac{1}{2}\sum_{i=1}^{n}\ln(|\Sigma|) - \sum_{i=1}^{n}\left(-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)\right) \\
&= n\ln(1) - \frac{dn}{2}(\ln(2)+\ln(\pi)) - \frac{n}{2}\ln(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{n}\left((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)
\end{aligned}$$

We now take the derivative with respect to μ, causing all but the last term to go to 0:

$$
\begin{aligned}
\frac{\partial L(\mu|\chi)}{\partial \mu} &= \frac{\partial}{\partial \mu}\left[-\frac{1}{2}\sum_{i=1}^{n}\big((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\big)\right] \\
&= -\frac{1}{2}\sum_{i=1}^{n}\frac{\partial}{\partial \mu}\big((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\big)
\end{aligned}
$$

Recall that the covaraince matrix Σ is symmetric, so Σ^{-1} is also symmetric. Using this fact and Formula (86) from the Matrix Cookbook, we find that:

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

Therefore:

$$
\begin{aligned}
\frac{\partial L(\mu|\chi)}{\partial \mu} &= -\frac{1}{2}\sum_{i=1}^{n}\frac{\partial}{\partial \mu}\big((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\big) \\
&= -\frac{1}{2}\left(-2\sum_{i=1}^{n}\Sigma^{-1}(x_i-\mu)\right) \\
&= \Sigma^{-1}\sum_{i=1}^{n}(x_i-\mu) \\
&= \Sigma^{-1}(n\mu-\sum_{i=1}^{n}x_i)
\end{aligned}
$$

We now set the partial derivative of the log-likelihood function to 0, and solve for μ:

$$
\begin{aligned}
&\Sigma^{-1}(n\mu-\sum_{i=1}^{n}x_i) = 0 \\
&\implies n\mu-\sum_{i=1}^{n}x_i = 0 \\
&\implies n\mu = \sum_{i=1}^{n}x_i \\
&= \mu = \frac{1}{n}\sum_{i=1}^{n}x_i
\end{aligned}
$$

So the maximum likelihood estimator of μ is $\hat{\mu_n} = \frac{1}{n}\sum_{i=1}^{n}x_i$.

We now calculate the maximum likelihood estimator of the covariance matrix Σ. Recall the log-likelihood of the given probability density function:

$$L(\Sigma|\chi) = n\ln(1) - \frac{dn}{2}\left(\ln(2)+\ln(\pi)\right) - \frac{n}{2}\ln(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{n}\left((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$$

We now take the partial derivative of the log-likelihood function with respect to the covariance matrix:

$$\frac{\partial L(\Sigma|\chi)}{\partial \Sigma} = -\frac{1}{2}\left[\sum_{i=1}^{n}\left(\frac{\partial}{\partial \Sigma}\ln(|\Sigma|)\right) + \sum_{i=1}^{n}\left(\frac{\partial}{\partial \Sigma}\left[(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right]\right)\right]$$

Using Formula (57) from the Matrix Cookbook, we find that: $\frac{\partial}{\partial \Sigma}\ln(|\Sigma|) = (\Sigma^T)^{-1}$. Since Σ is symmetric, $(\Sigma^T)^{-1} = \Sigma^{-1}$.

Using Formula (61) in the Matrix Cookbook, we find that $\frac{\partial}{\partial \Sigma}\left[(x_i-\mu)\Sigma^{-1}(x_i-\mu)\right] = -\Sigma^{-T}(x_i-\mu)(x_i-\mu)^T\Sigma^{-T}$. Again, since Σ is symmetric, Σ^{-1} is symmetric too, so: $-\Sigma^{-T}(x_i-\mu)(x_i-\mu)^T\Sigma^{-T} = -\Sigma^{T}(x_i-\mu)(x_i-\mu)^T\Sigma^{-1}$. Thus:

$$\begin{aligned}
\frac{\partial L(\Sigma|\chi)}{\partial \Sigma} &= -\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\partial}{\partial \Sigma}\ln(|\Sigma|)\right) + \sum_{i=1}^{n}\left(\frac{\partial}{\partial \Sigma}\left[(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right]\right) \\
&= -\frac{1}{2}\sum_{i=1}^{n}\Sigma^{-1} + \sum_{i=1}^{n}\left(-\Sigma^{T}(x_i-\mu)(x_i-\mu)^T\Sigma^{-1}\right) \\
&= -\frac{1}{2}\left[n\Sigma^{-1} - \Sigma^{-1}\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)\Sigma^{-1}\right]
\end{aligned}$$

Now, we set our formula equal to 0 and solve for Σ to find $\hat{\Sigma}$:

$$\begin{aligned}
&-\frac{1}{2}\left[n\Sigma^{-1} - \Sigma^{-1}\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)\Sigma^{-1}\right] = 0 \\
&\implies n\Sigma^{-1} - \Sigma^{-1}\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)\Sigma^{-1} = 0 \\
&\implies n\Sigma^{-1} = \Sigma^{-1}\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)\Sigma^{-1} \\
&\implies \Sigma^{-1} = \frac{1}{n}\Sigma^{-1}\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)\Sigma^{-1} \\
&\implies \Sigma\Sigma^{-1}\Sigma = \frac{1}{n}\Sigma\Sigma^{-1}\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)\Sigma^{-1}\Sigma \\
&\implies \Sigma = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)
\end{aligned}$$

So the maximum likelihood estimator of the covariance matrix Σ is $\hat{\Sigma} = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T\right)$.

(b). We call the maximum likelihood estimator of μ, $\hat{\mu_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$, is called a biased estimate of μ if:

$$Bias(\hat{\mu_n}) = E\left[\hat{\mu_n}\right] - \mu \neq 0$$

We calculate this value:

$$\begin{aligned}
Bias(\hat{\mu_n}) &= E\left[\hat{\mu_n}\right] - \mu \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] - \mu \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n} x_i\right] - \mu \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left[x_i\right] - \mu
\end{aligned}$$

Recall that for all $i = 1, 2, 3, ..., n$, x_i is sampled from a multivariate gaussian distribution with mean μ and covariance matrix Σ. Since the samples are iid (independent, identically distributed), $E\left[x_i\right] = \mu$. Therefore:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} E\left[x_i\right] - \mu &= \frac{1}{n}\sum_{i=1}^{n} \mu - \mu \\
&= \frac{1}{n}(n\mu) - \mu \\
&= \mu - \mu \\
&= 0
\end{aligned}$$

Therefore by definition of bias, the maximum likelihood estimator $\hat{\mu}$ is an unbiased estimator of μ.

(c). Like in Problem 2(b), we call the maximum likelihood estimator of the covariance matrix, $\hat{\Sigma}$ a biased estimator if:

$$Bias(\hat{\Sigma}) = E\left[\hat{\Sigma}\right] - \Sigma \neq 0$$

Thus we evaluate $E\left[\hat{\Sigma}\right]$. Using the calculation from Problem 2(a),

$$
\begin{aligned}
E\left[\hat{\Sigma}\right] &= E\left[\frac{1}{n}\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right] \\
&= \Sigma + \frac{1}{n}E\left[\sum_{i=1}^{n}(\hat{\mu}-\mu)(\hat{\mu}-\mu)^T - (x_i-\mu)(\hat{\mu}-\mu)^T - (\hat{\mu}-\mu)(x_i-\mu)^T\right] \\
&= \Sigma + E\left[\hat{\mu}\hat{\mu}^T\right] - \frac{1}{n}\sum_{i=1}^{n}E\left[x_i\hat{\mu}^T\right] - 1n\sum_{i=1}^{n}E\left[\hat{\mu}x_i^T\right] + \mu\mu^T \\
&= \Sigma + \mu\mu^T - E\left[\hat{\mu}\hat{\mu}^T\right]
\end{aligned}
$$

Notice that for all $i \neq j$,

$$
E\left[x_i x_j^T\right] = E\left[x_i\right]E\left[x_j^T\right] = \mu\mu^T
$$

because each sample is iid (independent, identically distributed).

A sample x_i is a random variable, which follows the distribution it is sampled from. Recall the definition of covariance in a gaussian:

$$
\Sigma = E\left[(x_i-\mu)(x_i-\mu)^T\right] = E\left[x_i x_i^T\right] - \mu\mu^T
$$

which implies $E\left[x_i x_i^T\right] = \Sigma + \mu\mu^T$.

Thus,

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

Expanding the multiplication, there are n terms that are the same and $n(n-1)$ terms that are different, so $E\left[\hat{\mu}\hat{\mu}^T\right] = \frac{1}{n^2}(n(\Sigma+\mu\mu^T) + n(n-1)\mu\mu^T)$ which implies $E\left[\hat{\mu}\mu^T\right] = \frac{1}{n}\Sigma + \mu\mu^T$.

Thus,

$$
\begin{aligned}
E\left[\hat{\Sigma}\right] &= \Sigma + \mu\mu^T - \frac{1}{n}\Sigma - \mu\mu^T \\
&= \frac{n-1}{n}\Sigma
\end{aligned}
$$

So,

$$
Bias(\hat{\Sigma}) = E\left[\hat{\Sigma}\right] - \Sigma = \frac{n-1}{n}\Sigma - \Sigma = -\frac{1}{n}\Sigma \neq 0
$$

So $\hat{\Sigma}$ is a biased estimator.

Problem 3

(a). (EXPLAIN HERE)

(b).

Error Rates for MultiGaussClassify with Full Covariance Matrix on Boston50						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.34	0.25	0.10	0.33	0.12	0.23	0.10

Error Rates for MultiGaussClassify with Full Covariance Matrix on Boston75						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.38	0.13	0.35	0.40	0.04	0.26	0.14

Error Rates for MultiGaussClassify with Full Covariance Matrix on Digits						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.06	0.11	0.10	0.06	0.09	0.08	0.02

Error Rates for MultiGaussClassify with Diagonal Covariance Matrix on Boston50						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.43	0.26	0.15	0.35	0.15	0.27	0.11

Error Rates for MultiGaussClassify with Diagonal Covariance Matrix on Boston75						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.33	0.13	0.26	0.42	0.08	0.24	0.13

Error Rates for MultiGaussClassify with Diagonal Covariance Matrix on Digits						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.55	0.52	0.47	0.52	0.53	0.52	0.03

Error Rates for Logistic Regression on Boston50						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.13	0.12	0.09	0.28	0.22	0.17	0.07

Error Rates for Logistic Regression on Boston75						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.10	0.08	0.13	0.11	0.05	0.09	0.03

Error Rates for Logistic Regression on Digits						
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
0.07	0.11	0.05	0.04	0.10	0.08	0.03