ESE 524 - Homework 3

Solutions Outline
Assigned date: 02/19/19
Due Date: 03/05/19
Total Points: 100 + 20 Extra Credit

These solutions are meant to be sketches. For full solutions we encourage you to fill in the details on your own or ask the TA in the office hours.

1) **Maximum Likelihood Estimation**
Consider samples $\mathbf{x} = [x[0], ..., x[N-1]]^T$ from a model:

$$x[n] = \theta^{1/2} s[n] r[n] + w[n]$$

where $\mathbf{s} = [s[0], ..., s[N-1]]^T$ is a known signal, $\mathbf{r} = [r[0], ..., r[N-1]]^T$ and $\mathbf{w} = [w[0], ..., w[N-1]]^T$ are i.i.d. $\mathcal{N}(0, 1)$ random variables, and $\theta \geq 0$ is an unknown parameter.

(a) **(10 pts)** Suppose $s[n] \in \{-1, 1\}$ is a sequence of +1's and -1's, what is the maximum likelihood estimate (MLE) of θ?
Solution: $\theta^{1/2} s[n] r[n]$ is multiplying $r[n]$ by a constant, so it is distributed as $N(0, \theta s[n]^2) = N(0, \theta(\pm 1)^2) = N(0, \theta)$.
This means that $x[n] \sim N(0, \theta + 1)$.
Then it follows from maximizing the log likelihood that the MLE is:

$$\hat{\theta} = \frac{1}{N} \sum_{n=0}^{N-1} (x[n]^2) - 1$$

(b) **(10 pts)** Compute the bias as well as variance of your estimate from a), and compare the latter with the Cramer-Rao bound (CRB). [*Hint*: You do not need to calculate the CRB of θ].
Solution:
We have

$$\mathbb{E}[\hat{\theta}] = \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}[x[n]^2] - 1 = \frac{1}{N} \sum_{n=0}^{N-1} (1 + \theta) - 1 = \theta$$

Since $x[n]$ are independent,

$$\operatorname{var}(\hat{\theta}) = \frac{2(1+\theta)^2}{N} \tag{1}$$

where we used the variance of a χ^2 random variable to calculate $\operatorname{var}(x[n]^2)$.

2) **Linear Models**
In linear models, we have $\boldsymbol{x} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w}$, where $\boldsymbol{x}$ is our observation, $\boldsymbol{\theta}$ is the parameter vector to be estimated, and $\boldsymbol{w} \sim N(\boldsymbol{0}, \boldsymbol{C})$ is colored Gaussian noise with covariance matrix $\boldsymbol{C}$. Assume that $\boldsymbol{C}$ is positive definite.

(a) **(5 pts)** Compute the CRB of $\boldsymbol{\theta}$.

Hint: recall that Fisher information $\boldsymbol{I}(\boldsymbol{\theta}) = \operatorname{cov}_{p(\boldsymbol{x};\boldsymbol{\theta})}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{x}; \boldsymbol{\theta})\right]$. Further, $\frac{\partial \boldsymbol{x}^T \boldsymbol{S} \boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{S}\boldsymbol{x}$ if $\boldsymbol{S}$ is symmetric, $\frac{\partial \boldsymbol{x}^T \boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}$, and $\frac{\partial \boldsymbol{a}^T \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}$.

(b) **(5 pts)** Show that the estimator $\hat{\boldsymbol{\theta}} = (\boldsymbol{H}^T \boldsymbol{C}^{-1} \boldsymbol{H})^{-1} \boldsymbol{H}^T \boldsymbol{C}^{-1} \boldsymbol{x}$ is unbiased and efficient, i.e., the MVUE, by comparing the covariance matrix and CRB.

(c) **(10 pts)** In order to estimate $\boldsymbol{\theta}$ better, we use two different linear systems to get two independent observations $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$:

$$\boldsymbol{x}_1 = \boldsymbol{H}_1\boldsymbol{\theta} + \boldsymbol{w}_1$$
$$\boldsymbol{x}_2 = \boldsymbol{H}_2\boldsymbol{\theta} + \boldsymbol{w}_2$$

where $\boldsymbol{w}_1 \sim N(\boldsymbol{0}, \boldsymbol{C}_1)$ and $\boldsymbol{w}_2 \sim N(\boldsymbol{0}, \boldsymbol{C}_2)$. Write the MVUE for $\boldsymbol{\theta}$ by using the above two observations. Write the MVUE for $\boldsymbol{\theta}$ if only one observation, for example $\boldsymbol{x}_1$, is available. Compare the covariance matrices of these two MVUEs (Recall that for positive semidefinite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we write $\boldsymbol{A} \geq \boldsymbol{B}$, if $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite and you can assume that $\boldsymbol{H}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{H}_1$ and $\boldsymbol{H}_2^T\boldsymbol{C}_2^{-1}\boldsymbol{H}_2$ are invertible.)

Hint: $(\boldsymbol{A}+\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}(\boldsymbol{B}^{-1}+\boldsymbol{A}^{-1})^{-1}\boldsymbol{A}^{-1}$ if $\boldsymbol{A}$ and $\boldsymbol{B}$ are invertible.

Solution:

i)

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{C}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta})^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta})\right),$$

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\theta}} \log p(\boldsymbol{x};\boldsymbol{\theta}) &= -\frac{1}{2}\left(-\frac{\partial \boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{H}\boldsymbol{\theta}}{\partial\boldsymbol{\theta}} - \frac{\partial \boldsymbol{\theta}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{x}}{\partial\boldsymbol{\theta}} + \frac{\partial \boldsymbol{\theta}^T\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\boldsymbol{\theta}}{\partial\boldsymbol{\theta}}\right) \\
&\vdots \\
&= \boldsymbol{H}^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta})
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{I}(\boldsymbol{\theta}) &= \operatorname{cov}_{p(\boldsymbol{x};\boldsymbol{\theta})}\left[\frac{\partial}{\partial\boldsymbol{\theta}} \log p(\boldsymbol{x};\boldsymbol{\theta})\right] \\
&= \mathbb{E}[\boldsymbol{H}^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta})(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta})^T\boldsymbol{C}^{-1}\boldsymbol{H}] \\
&\vdots \\
&= \boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}
\end{aligned}$$

$$\text{CRB} = \boldsymbol{I}(\boldsymbol{\theta})^{-1} = (\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1}.$$

ii) Unbiased:

$$\mathbb{E}[\hat{\boldsymbol{\theta}}] = (\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{C}^{-1}\mathbb{E}[\boldsymbol{x}] = (\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H}\boldsymbol{\theta} = \boldsymbol{\theta};$$

Efficient:

$$\begin{aligned}
\operatorname{cov}(\hat{\boldsymbol{\theta}}) &= (\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{C}^{-1}\operatorname{cov}[\boldsymbol{x}]\boldsymbol{C}^{-1}\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1} \\
&= (\boldsymbol{H}^T\boldsymbol{C}^{-1}\boldsymbol{H})^{-1} \\
&= \text{CRB}
\end{aligned}$$

As this estimator achieves the CRB, thus it is a MVUE.

iii) Rewrite two observations as:

$$\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{H}_1\\ \boldsymbol{H}_2\end{bmatrix}\boldsymbol{\theta} + \begin{bmatrix}\boldsymbol{w}_1\\ \boldsymbol{w}_2\end{bmatrix}$$

From b), we have that based on two observations:

$$\begin{aligned}\hat{\theta}(\boldsymbol{x}_1,\boldsymbol{x}_2) &= \left(\begin{bmatrix}\boldsymbol{H}_1^T & \boldsymbol{H}_2^T\end{bmatrix}\begin{bmatrix}\boldsymbol{C}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{C}_2\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{H}_1\\ \boldsymbol{H}_2\end{bmatrix}\right)^{-1}\begin{bmatrix}\boldsymbol{H}_1^T & \boldsymbol{H}_2^T\end{bmatrix}\begin{bmatrix}\boldsymbol{C}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{C}_2\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix}\\ &= \left(\boldsymbol{H}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{H}_1+\boldsymbol{H}_2^T\boldsymbol{C}_2^{-1}\boldsymbol{H}_2\right)^{-1}\left(\boldsymbol{H}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{x}_1+\boldsymbol{H}_2^T\boldsymbol{C}_2^{-1}\boldsymbol{x}_2\right)\end{aligned}$$

$$\begin{aligned}\operatorname{cov}(\hat{\theta}(\boldsymbol{x}_1,\boldsymbol{x}_2)) &= \left(\begin{bmatrix}\boldsymbol{H}_1^T & \boldsymbol{H}_2^T\end{bmatrix}\begin{bmatrix}\boldsymbol{C}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{C}_2\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{H}_1\\ \boldsymbol{H}_2\end{bmatrix}\right)^{-1}\\ &= \left(\boldsymbol{H}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{H}_1\right)^{-1} - \left(\boldsymbol{H}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{H}_1\right)^{-1}\left(\left(\boldsymbol{H}_2^T\boldsymbol{C}_2^{-1}\boldsymbol{H}_2\right)^{-1}+\left(\boldsymbol{H}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{H}_1\right)^{-1}\right)^{-1}\left(\boldsymbol{H}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{H}_1\right)\end{aligned}$$

Based on one observation $\boldsymbol{x}_1$, we have

$$\hat{\theta}(\boldsymbol{x}_1) = \left(\boldsymbol{H}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{H}_1\right)^{-1}\boldsymbol{H}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{x}_1$$

$$\operatorname{cov}(\hat{\theta}(\boldsymbol{x}_1)) = \left(\boldsymbol{H}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{H}_1\right)^{-1}$$

Thus,

$$\begin{aligned}&\operatorname{cov}(\hat{\theta}(\boldsymbol{x}_1)) - \operatorname{cov}(\hat{\theta}(\boldsymbol{x}_1,\boldsymbol{x}_2))\\ &\quad= \left(\boldsymbol{H}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{H}_1\right)^{-1}\left(\left(\boldsymbol{H}_2^T\boldsymbol{C}_2^{-1}\boldsymbol{H}_2\right)^{-1}+\left(\boldsymbol{H}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{H}_1\right)^{-1}\right)^{-1}\left(\boldsymbol{H}_1^T\boldsymbol{C}_1^{-1}\boldsymbol{H}_1\right)^{-1}\end{aligned}$$

which is positive semidefinite (positive definite), i.e., $\operatorname{cov}(\hat{\theta}(\boldsymbol{x}_1)) \geq \operatorname{cov}(\hat{\theta}(\boldsymbol{x}_1,\boldsymbol{x}_2))$. From this result, we can see that with more observations, we can reduce the covariance of the estimation.
Note: To recognize that the above matrix is positive semidifinite (PSD), we can use

- if A is PSD, A is invertible, then A is positive definite (PD).
- if A is PD, then A^{-1} is PD, by comparing the eigenvalues of A and A^{-1} .
- if A and B are PD (PSD), then $A+B$ is PD (PSD).

3) **Trace of a Matrix**

(a) **(10 pts)** If $\mathbf{A} \in \mathbb{R}^{m\times n}$, then show that $\|\mathbf{PAQ}\| = \|\mathbf{A}\|$, where $\|\cdot\|$ is the Frobenius norm, $\mathbf{P} \in \mathbb{R}^{m\times m}$ and $\mathbf{Q} \in \mathbb{R}^{n\times n}$ are orthogonal matrices.
Solution: Consider the following:

$$\|\mathbf{PAQ}\|^2 = \operatorname{Tr}\left[(\mathbf{PAQ})^T(\mathbf{PAQ})\right].$$

Expand the above equation, and use the product of trace identity, i.e., $\operatorname{Tr}(\mathbf{AB}) = \operatorname{Tr}(\mathbf{BA})$ and orthogonal property of the matrices $\mathbf{P}$ and $\mathbf{Q}$.

(b) **(10 pts)** Show that expectation and trace operators commute, i.e., if $\mathbf{X} \in \mathbb{R}^{n\times n}$ is a random matrix, then

$$\mathbb{E}[\operatorname{Tr}(\mathbf{X})] = \operatorname{Tr}[\mathbb{E}(\mathbf{X})]$$

Solution: Consider

$$\begin{aligned}\mathbb{E}[\operatorname{Tr}(\mathbf{X})] &= \sum_{\mathbf{A}}\Pr(\mathbf{X}=\mathbf{A})\sum_i A_{i,i}\\ &= \sum_i\sum_{\mathbf{A}}\Pr(\mathbf{X}=\mathbf{A})A_{i,i}\\ &= \sum_i\sum_a\Pr(\mathbf{X}_{i,i}=a)a\\ &= \sum_i\mathbb{E}(\mathbf{X}_{i,i}) = \operatorname{Tr}[\mathbb{E}(\mathbf{X})].\end{aligned}$$

4) **Maximum Likelihood Estimation**
Let $X_1, \ldots, X_n$ be independently identically distributed random variables with probability density function

$$f(x;\sigma,\lambda) = \frac{\sigma^{1/\lambda}}{\lambda} \exp\left[-(1+\frac{1}{\lambda})\log(x)\right] \mathbb{I}(x \geq \sigma),$$

where $x \geq \sigma$, $\sigma > 0$, and $\lambda > 0$. The indicator function $\mathbb{I}(x \geq \sigma) = 1$ if $x \geq \sigma$ and 0 otherwise. Find the maximum likelihood estimator (MLE) of (σ, λ) denoted as $(\hat{\sigma}, \hat{\lambda})$ with the following steps

(a) **(10 pts)** Show that $\hat{\sigma} = X_{(1)} = \min\{X_1, \ldots, X_n\}$ is the MLE of σ regardless of the value of $\lambda > 0$.
Solution:
The likelihood function can be written as

$$\begin{aligned} L(\sigma,\lambda;\boldsymbol{X}) &= \left[\prod_{i=1}^{n} f(X_i;\sigma,\lambda)\right] \\ &= \sigma^{n/\lambda}\frac{1}{\lambda^n}\exp\left[-(1+\frac{1}{\lambda})\sum_{i=1}^{n}\log(X_i)\right]\mathbb{I}(X_{(1)} \geq \sigma) \end{aligned}$$

where $X_{(1)} = \min\{X_1, \ldots, X_n\}$. Clearly, the likelihood is maximized by making σ as large as possible. Hence, $\hat{\sigma} = X_{(1)}$.

Solution:
The likelihood function can be written as

$$\begin{aligned} L(\sigma,\lambda;\boldsymbol{X}) &= \left[\prod_{i=1}^{n} f(X_i;\sigma,\lambda)\right] \\ &= \sigma^{n/\lambda}\frac{1}{\lambda^n}\exp\left[-(1+\frac{1}{\lambda})\sum_{i=1}^{n}\log(X_i)\right]\mathbb{I}(X_{(1)} \geq \sigma) \end{aligned}$$

where $X_{(1)} = \min\{X_1, \ldots, X_n\}$. Clearly, the likelihood is maximized by making σ as large as possible. Hence, $\hat{\sigma} = X_{(1)}$.

(b) **(10 pts)** Find the MLE of $\hat{\lambda}$ if $\sigma = \hat{\sigma}$ (that is assume σ is known).
Solution:
From part (a), we get

$$L(\hat{\sigma},\lambda;\boldsymbol{X}) = \hat{\sigma}^{n/\lambda}\frac{1}{\lambda^n}\exp\left[-(1+\frac{1}{\lambda})\sum_{i=1}^{n}\log(X_i)\right]\mathbb{I}(X_{(1)} \geq \hat{\sigma}) \tag{2}$$

The loglikelihood is given as

$$\log L(\hat{\sigma},\lambda;\boldsymbol{X}) = \frac{n}{\lambda}\log\hat{\sigma} - n\log(\lambda) - (1+\frac{1}{\lambda})\sum_{i=1}^{n}\log(X_i)$$

Taking the derivative of the loglikelihood with respect to λ and setting it zero we get

$$\frac{d}{d\lambda}\log L(\hat{\sigma},\lambda;\boldsymbol{X}) = -\frac{n}{\lambda^2}\log\hat{\sigma} - \frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{n}\log(X_i) = 0 \tag{3}$$

On further simplifying, we get

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\log(X_i/\hat{\sigma}) \tag{4}$$

Take the second derivative of the loglikelihood function and substitute $\hat{\lambda}$ and show that it is always less than zero. We avoid the proof as it involves simple algebra.

5) **Maximum Likelihood Estimation for US Household Income - MATLAB Problem**

(a) **(5 pts)** Given a parameter vector $\theta = [\alpha, c, k]$, the *Burr Distribution* has density function:

$$p(x[n];\theta) = \frac{\frac{kc}{\alpha}\left(\frac{x[n]}{a}\right)^{c-1}}{\left(1+\left(\frac{x[n]}{a}\right)^{c}\right)^{k+1}} \;\forall\, \alpha, c, k, x[n] > 0.$$

Find the gradient of the log likelihood for N samples.

Solution The joint likelihood function is

$$p(\mathbf{x};\theta) = \prod_{n=0}^{N-1} \frac{\frac{kc}{\alpha}\left(\frac{x}{\alpha}\right)^{c-1}}{\left(1+\left(\frac{x}{\alpha}\right)^{c}\right)^{k+1}}$$

The Log Likelihood is:

$$L(\mathbf{x},\theta) = \sum_{n=0}^{N-1} \ln(k) + \ln(c) - \ln(\alpha) + (c-1)\ln(x) - (c-1)\ln(\alpha) - (k+1)\ln(1+\left(\frac{x}{\alpha}\right)^{c})$$

The Gradient is given as:

$$(\nabla L)^T = \begin{bmatrix} \sum_{n=0}^{N-1} -\frac{1}{\alpha} - \frac{c-1}{\alpha} + \frac{(k+1)c}{(1+\left(\frac{x[n]}{\alpha}\right)^{c})\alpha^{c+1}} \\ \sum_{n=0}^{N-1} \frac{1}{c} + \ln x - \ln\alpha - \frac{(k+1)}{(1+\left(\frac{x[n]}{\alpha}\right)^{c})}\left(\frac{x}{\alpha}\right)^{c}\ln(\frac{x}{\alpha}) \\ \sum_{n=0}^{N-1} \frac{1}{k} - \ln(1+\left(\frac{x[n]}{\alpha}\right)^{c}) \end{bmatrix}$$

(b) **(5 pts)** Download “income.mat” from Canvas. Use gradient descent to find the MLE estimates for α, c, and k based on the data in the income file.

Solutions: See HW3.m

(c) **(5 pts)** Use the `mle` or `fmincon` commands in MATLAB to find the parameter estimates, and compare them to the estimates from part b).

(d) **(5 pts)** Plot a histogram of the income data, as well as the Burr distribution you fitted in either b) or c). [*Hint*: The `yyaxis` command in MATLAB can be used to set multiple axes on the same plot.]

6) (**20 pts**) **Extra Credit**: Come up with an example and solution illustrating one or more concepts from class so far. This example should be something you believe would be good to present in class to help other students understand a concept from the lectures. MATLAB (or other software) simulations are encouraged. Problems can be inspired by or explore applications from literature, but should not just copy the results of a paper.