ESE 524 - Homework 2

Solution Outline
Assigned date: 02/05/19
Due date: 02/19/19
Total Points: 100 + 20 E.C.

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) **MLE and CRB**
Suppose we have N i.i.d. samples $X_1, X_2, \ldots, X_N$, drawn from a Poisson Distribution, i.e.,

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x = 0, 1, 2, \ldots$$

where the rate λ is the parameter that we want to estimate.

a) **(5 pts)** Find the Maximum likelihood estimator (MLE) of λ. In addition, check whether the MLE is unbiased or not.

Solution: The likelihood is written as

$$P(\boldsymbol{X}) = \frac{\lambda^{X_1+X_2+\cdots+X_N} e^{-N\lambda}}{X_1! X_2! \cdots X_N!},$$

By using $\frac{\partial \log P(\boldsymbol{X})}{\partial \lambda} = 0$, we have

$$\lambda_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

Finally, show that

$$E[\lambda_{\text{MLE}}] = \lambda.$$

b) **(5 pts)** Compute the Cramer-Rao lower bound for the estimation of λ.
Solution: The Fisher information is

$$I(\lambda) = -E\left[\frac{\partial^2 \log P(\boldsymbol{X})}{\partial \lambda^2}\right] = \frac{N}{\lambda}.$$

Thus, the CRB is $\frac{\lambda}{N}$.

c) **(10 pts)** Check if the MLE is efficient.

Solution: Since, $X_1, \ldots, X_N$ are N independent samples, thus

$$\text{var}[\lambda_{\text{MLE}}] = \frac{1}{N^2} \sum_{i=1}^{N} \text{var}[X_i]$$

Next, show that $\text{var}[\lambda_{\text{MLE}}]$ is the inverse of the Fisher information, which implies that the MLE if efficient.

2) **Fisher Information**
As learned from the course and slides, we know that under some regularity conditions, an unbiased estimator $T(\boldsymbol{x})$ is efficient for the parameter θ if and only if

$$I(\theta)[T(\boldsymbol{x}) - \theta] = \frac{\partial \log p(\boldsymbol{x}; \theta)}{\partial \theta},$$

where $I(\theta)$ is the Fisher information.

a) **(10 pts)** Use the definition of Fisher information to prove that the Fisher information $I(\theta)$ is non-negative.
Solution: Use the definition of Fisher information matrix and non-negative definiteness to prove this.

$$I(\theta) = E_{\boldsymbol{x}}\left\{\left(\frac{\partial}{\partial\theta}\log p(\boldsymbol{x};\theta)\right)^2\right\} \geq 0$$

b) **(10 pts)** Suppose that Fisher information is positive, and the maximum likelihood estimator can be found by $\frac{\partial \log p(\boldsymbol{x};\theta)}{\partial\theta} = 0$, then prove that if an unbiased efficient estimator $T(\boldsymbol{x})$ exists, then it must be the MLE.

Solution: If such $T(\boldsymbol{x})$ exists, then we have

$$I(\theta)[T(\boldsymbol{x}) - \theta] = \frac{\partial \log p(\boldsymbol{x};\theta)}{\partial\theta}.$$

Pick $\theta = \theta_{\mathrm{ML}}(\boldsymbol{x})$ and show that

$$T(\boldsymbol{x}) = \theta_{\mathrm{ML}}(\boldsymbol{x}).$$

3) **Exponential Family**
Given a probability measure ν (see the note below), the exponential family is defined to have the following general form:

$$p(x|\eta) = h(x)\exp\{\eta^T T(x) - A(\eta)\}$$

for a parameter vector η, often referred to as the canonical parameter. The function $A(\eta)$ is known as the cumulant function.

a) **(10 pts)** Show that with respect to the measure ν,

$$A(\eta) = \log\int h(x)\exp\{\eta^T T(x)\}\nu(dx)$$

Note: A probability measure is a real-valued function defined on a set of events in a probability space that satisfies measure properties such as countable additivity. For this particular problem, you may just need the fact about ν:

$$\int pv(dx) = 1$$

where p is the corresponding probability density.

Solution: Since ν is a probability measure, we have

$$\int h(x)\exp\{\eta^T T(x)\}\exp\{-A(\eta)\}\nu(dx) = 1$$

Take logarithms on both sides and simplify further to arrive at the soluiton.

b) (**10 pts**) Given a distribution in the exponential family:

$$p(\boldsymbol{x};\boldsymbol{\eta}) = h(\boldsymbol{x})\exp\left[\boldsymbol{\eta}^T\boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\eta})\right],$$

Prove that

$$\mathrm{E}_{p(\boldsymbol{x};\boldsymbol{\eta})}\left[\boldsymbol{T}(\boldsymbol{X})\right] = \frac{\partial A(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}}, \qquad \mathrm{var}_{p(\boldsymbol{x};\boldsymbol{\eta})}\left[\boldsymbol{T}(\boldsymbol{X})\right] = \frac{\partial^2 A(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}.$$

Solution:
Same as part (a), consider

$$1 = \int p(\boldsymbol{x};\boldsymbol{\eta})d\boldsymbol{x} = \exp\left[-A(\boldsymbol{\eta})\right]\int \exp\left[\boldsymbol{T}(\boldsymbol{x})^T\boldsymbol{\eta})\right]h(\boldsymbol{x})d\boldsymbol{x}.$$

Taking logarithms on both sides we get:

$$A(\boldsymbol{\eta}) = \log \int \exp\left[\boldsymbol{T}(\boldsymbol{x})^T \boldsymbol{\eta}\right] h(\boldsymbol{x}) d\boldsymbol{x}.$$

Taking the first derivative with respect to $\boldsymbol{\eta}$, we get:

$$\begin{aligned}\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} &= \frac{\int \boldsymbol{T}(\boldsymbol{x}) \exp\left[\boldsymbol{T}(\boldsymbol{x})^T \boldsymbol{\eta}\right] h(\boldsymbol{x}) d\boldsymbol{x}}{\int \exp\left[\boldsymbol{T}(\boldsymbol{x})^T \boldsymbol{\eta}\right] h(\boldsymbol{x}) d\boldsymbol{x}} \\ &= \mathrm{E}_{p(\boldsymbol{x};\boldsymbol{\eta})}[\boldsymbol{T}(\boldsymbol{X})].\end{aligned}$$

Take one more derivative, we have the result,

$$\begin{aligned}\frac{\partial^2 A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} &= \frac{\partial \int \boldsymbol{T}(\boldsymbol{x}) p(\boldsymbol{x};\boldsymbol{\eta}) d\boldsymbol{x}}{\partial \boldsymbol{\eta}^T} \\ &= \mathrm{cov}_{p(\boldsymbol{x};\boldsymbol{\eta})}\left[\boldsymbol{T}(\boldsymbol{X})\right].\end{aligned}$$

The above proof assumes that the exponential family follows a continuous distribution. Note that for the discrete case, the proof is similar.

4) (**Slides, l3.pdf**) **Cramer-Rao Bound**: The measurements $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T$ are independent and identically distributed as Gaussian random variables with unknown mean μ and variance σ^2.

a) **(10 pts)** Show that the sample mean $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ and sample variance $S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$ are unbiased estimators and that they are uncorrelated and independent random variables. [*Hint*: Show that Gaussian variables $X_i - \bar{X}$ and $\bar{X}$ are uncorrelated for $i = 1, \ldots, n$].

Solution: To show $\bar{X}$ is unbiased:

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \mu$$

To show uncorrelatedness, show that $\mathrm{var}(\bar{X})$ is given as

$$\mathrm{var}(\bar{X}) = \mathbb{E}((\bar{X} - \mu)^2) = \frac{\sigma^2}{n}.$$

To show S^2 is unbiased, consider

$$\begin{aligned}(n-1)S^2 &= \sum_{k=1}^{n}(X_k - \bar{X})^2 \\ &= \sum_{k=1}^{n}(X_k - \mu)^2 - n(\mu - \bar{X})^2\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}((n-1)S^2) &= \sum_{k=1}^{n}\mathbb{E}((X_k - \mu)^2) - n\mathbb{E}((\mu - \bar{X})^2) \\ &= (n-1)\sigma^2.\end{aligned}$$

To show that $\bar{X}$ and S^2 are uncorrelated, consider

$$\begin{aligned}\mathrm{cov}(\bar{X}, X_k - \bar{X}) &= \mathrm{cov}(\bar{X}, X_k) - \mathrm{var}(\bar{X}) \\ &= 0.\end{aligned}$$

b) **(5 pts)** Using the results in (i), derive the covariance matrix for the estimator $\hat{\boldsymbol{\theta}} = [\bar{X}, S^2]^T$.

Solution: The covarinace matrix of the estimate $\hat{\boldsymbol{\theta}}$ can be written as

$$\text{cov}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} \text{var}(\bar{X}) & \text{cov}(\bar{X}, S^2) \\ \text{cov}(S^2, \bar{X}) & \text{var}(S^2) \end{bmatrix}$$

Next, show that the variance of $\text{var}(S^2)$ is given as

$$\text{var}(S^2) = \frac{2\sigma^4}{n-1}.$$

using the chi-square distribution property.

Therefore,

$$\text{cov}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/(n-1) \end{bmatrix}$$

c) **(5 pts)** Derive the Cramer-Rao bound on the covariance matrix of any unbiased estimator $\hat{\theta}$ of $\theta = [\mu, \sigma^2]^T$. Compare to the results obtained in part (ii).

Solution: As X_k's are drawn from the same distribution, $\mathcal{I}(\boldsymbol{\theta}) = n\mathcal{I}_1(\boldsymbol{\theta})$, where

$$\mathcal{I}_1(\boldsymbol{\theta}) = \begin{bmatrix} \mathcal{I}_{11}(\boldsymbol{\theta}) & \mathcal{I}_{12}(\boldsymbol{\theta}) \\ \mathcal{I}_{21}(\boldsymbol{\theta}) & \mathcal{I}_{22}(\boldsymbol{\theta}) \end{bmatrix}$$

and

$$\mathcal{I}_{i,k}(\boldsymbol{\theta}) = -\mathbb{E}_{X|\boldsymbol{\theta}} \left[\frac{\partial^2}{\partial\theta_i \partial\theta_k} \ln f_{X|\boldsymbol{\theta}}(X|\boldsymbol{\theta}) | \boldsymbol{\theta} \right] \tag{1}$$

.

On computing, we get

$$\mathcal{I}(\boldsymbol{\theta}) = n \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 0.5/(\sigma^2)^2 \end{bmatrix}.$$

And the CRB is given as the inverse of the Fisher information matrix.

$$\text{CRB}(\boldsymbol{\theta}) = \mathcal{I}^{-1}(\boldsymbol{\theta}) = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2(\sigma^4)/n \end{bmatrix} \tag{2}$$

Note that the CRB and the covariance matrix converge in asymptoic sense, that is, when $n \to \infty$.

5) **Logistic Regression (MATLAB Problem)**: In cases involving binary data (e.g. heads/tails coin flip, passing/failing a test) the assumptions required for linear models do not always hold. This problem will use data from space shuttle o-ring tests to provide a short introduction to Generalized Linear Models and Logistic Regression. Download the file named "challenger.mat" from the blackboard Assignments-¿ , which contains the variables "Temperature" and "Failure".

a) **(5 pts)** Let $\mathbf{Y} \in \{0, 1\}$ be a Bernoulli random variable with mean p representing the outcome of interest (in our case whether the o-ring fails) , $\mathbf{x} = [x_1 ... x_n]$ be a vector of "explanatory" variables, and $\theta = [\theta_0 ... \theta_n]$ be a set of unknown parameters. Logistic regression defines the following relationship between the probability of the outcome and the explanatory variables:

$$\ln \frac{\Pr(\mathbf{Y} = 1|\mathbf{x}, \theta)}{1 - \Pr(\mathbf{Y} = 1|\mathbf{x}, \theta)} = \theta_0 + \theta_1 x_1 + ... + \theta_n x_n$$

Derive the expression for the mean $p = \Pr(\mathbf{Y} = 1|\mathbf{x}, \theta)$ in terms of $\mathbf{x}$ and θ.

Solution:
Exponentiating both sides and solving yields $\frac{\exp(\theta^T \mathbf{x})}{1+\exp(\theta^T \mathbf{x})} = \frac{1}{1+\exp(-\theta^T \mathbf{x})}$.

b) **(5 pts)** Let $\mathbf{Y^i}$ be independent observations of $\mathbf{Y}$ and $\mathbf{x}^i$ be the corresponding explanatory vectors. Derive the gradient of the log likelihood function with respect to $[\theta_0, ..., \theta_n]$:

$$\ln(L(\theta)) = \ln(\prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y^i})$$

(Hint: the function $f(x) = \frac{1}{1-\exp(-x)}$ is called the logistic function and its derivative is given by $f(x)(1 - f(x))$.)

Solution:
Let $f(x)$ be the logistic function as defined above The probability distribution of a Bernoulli random variable can be written as

$$\Pr(\mathbf{Y}^i|\mathbf{x}^i, \theta) = \Pr(\mathbf{Y}^i = 1|\mathbf{x}^i, \theta)^{Y^i} \Pr(\mathbf{Y}^i = 0|\mathbf{x}^i, \theta)^{1-Y^i} = f(\theta^T\mathbf{x})^{Y^i}(1 - f(\theta^T\mathbf{x}))^{1-Y^i}$$

Therefore the log-likelihood is

$$\ln(L(\theta)) = \ln(\prod_{i=1}^{n} \Pr(\mathbf{Y}^i|\mathbf{x}^i, \theta)) = \sum_{i=1}^{n} \ln\Pr(\mathbf{Y}^i|\mathbf{x}^i, \theta) =$$

$$\sum_{i=1}^{n} \ln(f(\theta^T\mathbf{x^i})^{Y^i}) + \ln((1 - f(\theta^T\mathbf{x^i})^{1-Y^i})$$

Using the chain rule and the hint the gradient is given by:

$$\nabla_\theta(\sum_{i=1}^{n} \ln(f(\theta^T\mathbf{x^i})^{Y^i}) + \ln((1 - f(\theta^T\mathbf{x^i})^{1-Y^i})) =$$

$$\sum_{i=1}^{n} Y^i \frac{1}{f(\theta^T\mathbf{x^i})}(f(\theta^T\mathbf{x^i})(1 - f(\theta^T\mathbf{x^i}))x^i + (1 - Y^i)\frac{1}{1 - f(\theta^T\mathbf{x^i})}(-1)(f(\theta^T\mathbf{x^i})(1 - f(\theta^T\mathbf{x^i}))x^i =$$

$$\sum_{i=1}^{n}(Y^i - f(\theta^T\mathbf{x^i}))x^i$$

c) **(5 pts)** Maximum Likelihood is usually done numerically. In the case of the o-rings data Y^i is the vector "Failure" and the samples X^i are found in the "Temperature" vector. Construct the likelihood function and it's gradient in MATLAB as a function of θ_0 and θ_1 and find the optimal weights using simple gradient descent (Note: 0.0001 works as a step size for gradient descent. You will need to multiply the log-likelihood by -1 to minimize using gradient descent). Compare your values to the coefficients generated using the `glmfit` function in MATLAB with "binomial" as the distribution.

Solution: See the HW3 file in "Course Documents/Homework Solutions" for a matlab coding solution. The coefficients that should be reached by both methods should be $\theta_0 = 15.043$ and $\theta_1 = -0.232$.

d) **(5 pts)** Plot $(X^i, P(Y^i = 1|\hat{\theta_1}, \hat{(}\theta_0))$ using the estimators you found in part (iii). On the same plot create a scatter plot of the actual values (X^i, Y^i). Do the failed o-rings have a high probability of failure?

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.