Chapter 3 - Questions and Solutions

Edwin Fennell

3.1 Let $\hat{\boldsymbol{\theta}}_i,\ i = 1, 2, ..., m$ be unbiased estimators of a parameter vector $\boldsymbol{\theta}$, so that $\mathbb{E}[\hat{\boldsymbol{\theta}}_i] = \boldsymbol{\theta},\ i = 1, 2, ..., m$. Moreover, assume that the respective estimators are uncorrelated to each other and that all have the same variance $\boldsymbol{\sigma^2} = \mathbb{E}[(\boldsymbol{\theta_i} - \boldsymbol{\theta})^{\boldsymbol{T}}(\boldsymbol{\theta_i} - \boldsymbol{\theta})]$. Show that by averaging the estimates, e.g.

$$\hat{\boldsymbol{\theta}} = \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$$

the new estimator has total variance $\boldsymbol{\sigma_c^2} = \mathbb{E}[(\boldsymbol{\theta_i} - \boldsymbol{\theta})^{\boldsymbol{T}}(\boldsymbol{\theta_i} - \boldsymbol{\theta})] = \frac{\boldsymbol{\sigma^2}}{\boldsymbol{m}}$.

Trivially, the mean of our unbiased estimators is also an unbiased estimator. The variance of our estimator $\hat{\theta}$ is therefore

$$\mathbb{E}(\hat{\theta}^T\hat{\theta}) - \mathbb{E}(\theta)^T\mathbb{E}(\theta)$$

This expands to

$$\mathbb{E}\left(\sum_{i=1}^{m} \frac{1}{m}\hat{\theta_j}^T \left(\sum_{j=1}^{m} \frac{1}{m}\hat{\theta_j}\right)\right) - \mathbb{E}(\theta)^T\mathbb{E}(\theta)$$

The estimators are all pairwise uncorrelated, which means the the product of the expectations of any two distinct estimators is equal to the expectation of their product. Therefore we can rewrite the above as

$$\sum_{i=1}^{m}\left(\sum_{j=1}^{m} \frac{1}{m^2}\mathbb{E}(\hat{\theta}_i)^T\mathbb{E}(\hat{\theta}_j) + \frac{1}{m^2}(\mathbb{E}(\hat{\theta}_i^T\hat{\theta}_i) - \mathbb{E}(\hat{\theta}_i)^T\mathbb{E}(\hat{\theta}_i))\right) - \mathbb{E}(\theta)^T\mathbb{E}(\theta)$$

The first term with the double sum is actually just

$$\left(\sum_{i=1}^{m} \frac{1}{m}\mathbb{E}(\hat{\theta}_i)\right)^T \left(\sum_{i=1}^{m} \frac{1}{m}\mathbb{E}(\hat{\theta}_i)\right) = E(\theta)^T E(\theta)$$

This just cancels out with the last term, and we are left with the middle term, which is just $\frac{1}{m^2}$ times the sum of the variances of the initial m estimators, as required.

3.2 Let x be random variable uniformly distributed on $[0, \frac{1}{\theta}], \theta > 0$. Assume that g is a Lebesgue measurable function on $[0, \frac{1}{\theta}]$. Show that if $\hat{\theta} = g(x)$ is an unbiased estimator, then

$$\int_0^{\frac{1}{\theta}} g(x)dx = 1$$

Assume that $\hat{\theta}$ is an unbiased estimator. Then regardless of the value of θ

$$\theta = \mathbb{E}(\hat{\theta}) = \mathbb{E}(g(x)) = \int_0^{\frac{1}{\theta}} g(x)\phi(x)dx$$

where ϕ is the p.d.f of x. Since x is uniform, the p.d.f is just constantly θ This directly gives

$$\int_0^{\frac{1}{\theta}} g(x)dx = 1$$

We note that there is no function $g(x)$ s.t. this holds for all θ. We note that our condition gives

$$\int_a^b g(x)dx = 0 \ \forall (0 < a < b)$$

Therefore

$$1 = \int_0^{\frac{1}{\theta}} g(x)dx = \sum_{i=0}^{\infty} \int_{\frac{2^{i+1}}{\theta}}^{\frac{2^i}{\theta}} g(x)dx = 0$$

which is a contradiction.

3.3 A family $[p(D, \theta); \theta \in A]$ is called complete if, for any vector function $h(D)$ such that $\mathbb{E}_D[h; D] = 0$, $\forall \theta$, then $h = 0$. Show that if $[p(D; \theta) : \theta \in A]$ in complete. and there exists an MVU estimator, then this estimator is unique.

Suppose we have two MVU estimators, θ_1 and θ_2. Then $\mathbb{E}\left(\frac{\theta_1+\theta_2}{2}\right)$ is also unbiased. This estimator has variance

$$\mathbb{E}\left(\left(\frac{\theta_1 + \theta_2}{2}\right)^T \left(\frac{\theta_1 + \theta_2}{2}\right)\right) = \frac{var(\theta_1)}{4} + \frac{var(\theta_2)}{4} + \mathbb{E}\left(\frac{\theta_1^T \theta_2}{2}\right)$$

$var(\theta_1) = var(\theta_2)$ is a lower bound for this variance, which gives us an inequality. A little rearranging gives

$$\mathbb{E}(2\theta_1^T \theta_2) \geq \mathbb{E}(\theta_1^T \theta_1) + \mathbb{E}(\theta_2^T \theta_2)$$

We also know that

$$\mathbb{E}(2\theta_1^T \theta_2) \leq \mathbb{E}(\theta_1^T \theta_1) + \mathbb{E}(\theta_2^T \theta_2)$$

by examining the expectation of the positive quantity $(\theta_1 - \theta_2)^T(\theta_1 - \theta_2)$. Therefore both of these inequalities hold with equality, and $\mathbb{E}((\theta_1 - \theta_2)^T(\theta_1 - \theta_2)) = 0\ \forall\theta$. By completeness, we have $(\theta_1 - \theta_2)^T(\theta_1 - \theta_2) = 0$ identically, which immediately gives $\theta_1 = \theta_2$, so the MVU estimator must be unique.

3.4 Let $\hat{\boldsymbol{\theta}}_n$ be an unbiased estimator, so that $\mathbb{E}[\hat{\boldsymbol{\theta}}_u] = \boldsymbol{\theta_0}$. Define a biased one by $\hat{\boldsymbol{\theta}}_b = (1 + \boldsymbol{\alpha})\hat{\boldsymbol{\theta}}_u$. Show that the range of $\boldsymbol{\alpha}$ where the MSE of $\hat{\boldsymbol{\theta}}_b$ is smaller than that of $\hat{\boldsymbol{\theta}}_u$ is

$$\boldsymbol{-2 < -\frac{2\text{MSE}(\hat{\theta}_u)}{\text{MSE}(\hat{\theta}_u) + \theta_0^2} < \alpha < 0}$$

We note that the MSE of $\hat{\theta}_u$ is

$$\mathbb{E}((\hat{\theta}_u - \theta_0)^2) = \mathbb{E}(\hat{\theta}_u^2) - \theta_0^2$$

by unbiasedness. Similarly, the MSE of $\hat{\theta}_b$ is

$$\mathbb{E}((\hat{\theta}_b - \theta_0)^2) = (1 + \alpha)^2\mathbb{E}(\hat{\theta}_u^2) - (1 + 2\alpha)\theta_0^2 = (1 + \alpha)^2\text{MSE}(\hat{\theta_u}) + \alpha^2\theta_0^2$$

Therefore the condition we want occurs for exactly

$$(1 + \alpha)^2\text{MSE}(\hat{\theta_u}) + \alpha^2\theta_0^2 < \text{MSE}(\hat{\theta_u})$$

which rearranges to

$$\alpha((\text{MSE}(\hat{\theta_u}) + \theta_0^2)\alpha + 2\text{MSE}(\hat{\theta_u})) < 0$$

This occurs only iff exactly one of α and $(\text{MSE}(\hat{\theta_u}) + \theta_0^2)\alpha + 2 \cdot \text{MSE}(\hat{\theta_u})$ is positive. Note also that $\alpha > 0$ directly implies $(\text{MSE}(\hat{\theta_u}) + \theta_0^2)\alpha + 2 \cdot \text{MSE}(\hat{\theta_u}) > 0$. Thus our proposed condition holds iff

$$-\frac{2\text{MSE}(\hat{\theta_u})}{\text{MSE}(\hat{\theta_u}) + \theta_0^2} < \alpha < 0$$

as required. The final leftmost inequality stems from the fact that

$$\frac{\text{MSE}(\hat{\theta_u})}{\text{MSE}(\hat{\theta_u}) + \theta_0^2} < 1$$

3.5 Show that for the setting of Problem 3.4, the optimal value of $\boldsymbol{\alpha}$ is equal to

$$\boldsymbol{\alpha_* = -\frac{1}{1 + \frac{\theta_0^2}{var(\hat{\theta}_u)}}}$$

We note that our MSE for the biased estimator is a quadratic in α. Therefore we can just pick the unique value of α for which the derivative is 0, and we obtain the minimum possible MSE. The derivative is

$$2(\text{MSE}(\hat{\theta_u}) + \theta_0^2)\alpha + 2\text{MSE}(\hat{\theta_u})$$

which rearranges to the required result.

3.6 Show that the regularity condition for the Cramer-Rao lower bound holds true if the order of differentiation and integration can be interchanged.

Note that $\frac{\partial\ \log(p(x|\theta))}{\partial\theta} = \frac{1}{p(x|\theta)} \cdot \frac{\partial\, p(x|\theta)}{\partial\theta}$. This gives

$$\mathbb{E}\left(\frac{\partial\ \log(p(x|\theta))}{\partial\theta}\right) = \int_{x\in X\cap\theta} p(x|\theta)\cdot\frac{1}{p(x|\theta)}\cdot\frac{\partial\, p(x|\theta)}{\partial\theta}\ dx = \int_{x\in X\cap\theta} \frac{\partial\, p(x|\theta)}{\partial\theta}\ dx$$

where $X\cap\theta$ is the probability space created by intersecting the probability space X with the event θ, and μ is its associated probability measure. (We fix θ to be a single event - we are using the frequentist concept of a "true" value for θ)

Now, if we assume that differentiation and integration can be interchanged, then this is equal to

$$\frac{\partial}{\partial\theta}\int_{x\in X\cap\theta} p(x|\theta)\ d\mu$$

The expression being differentiated is just the integral of the probability of all events over the probability space, and so so is just equal to 1. Therefore the derivative is equal to 0, and thus the regularity condition holds, as required.

3.7 Derive the Cramer-Rao bound for the LS estimator, when the training data result from the linear model

$$\boldsymbol{y_n = \theta x_n + \eta_n, \quad n = 1, 2, ...,}$$

where $\boldsymbol{x_n}$ and $\boldsymbol{\eta_n}$ are i.i.d. samples of a zero mean random variable with variance $\boldsymbol{\sigma_x^2}$, and a Gaussian random variable with zero mean and variance $\boldsymbol{\sigma_\eta^2}$ respectively. Assume also that $\boldsymbol{x}$ and $\boldsymbol{\eta}$ are independent. Then show that the LS simulator achieves the CR bound only asymptotically.

We here make the sensible assumption that both x and y are observable, and that η is unobserved and the only source of uncertainty in our system. This gives us our likelihood function as

$$f(\mathbf{a}, \mathbf{b}; \theta) = p(\mathbf{x} = \mathbf{a}, \boldsymbol{\eta} = \mathbf{b} - \theta\mathbf{a})$$

where p is our joint likelihood function. By assumption of mutual independence of the all elements of $\mathbf{x}$ and $\boldsymbol{\eta}$, this is just equal to

$$\prod_{n=1}^{N} p(x_n = a_n) \cdot p(\eta_n = b_n - \theta a_n)$$

(via an abuse of notation, the p here correspond to the individual likelihood functions to the relevant quantities inside the brackets) In order to obtain the Fisher information, we take the log, differentiate twice w.r.t θ, multiply by -1, and take the expectation. This yields

$$\frac{1}{\sigma_\eta^2} \sum_{n=1}^{N} E(x_0^2)$$

which is just equal to

$$N \cdot \frac{\sigma_x^2}{\sigma_\eta^2}$$

Therefore the Cramer-Rao lower bound is

$$\frac{\sigma_\eta^2}{N \cdot \sigma_x^2}$$

Given a set of observations $\mathbf{x}, \mathbf{y}$, the value of $\hat{\theta}$ that minimises least squares is

$$\hat{\theta} = \frac{\mathbf{x} \cdot \mathbf{y}}{\mathbf{x} \cdot \mathbf{x}} = \frac{\mathbf{x} \cdot (\theta \mathbf{x} + \boldsymbol{\eta})}{\mathbf{x} \cdot \mathbf{x}} = \theta + \frac{\mathbf{x} \cdot \boldsymbol{\eta}}{\mathbf{x} \cdot \mathbf{x}}$$

Some rearrangement reveals the variance of this variable to be

$$\sigma_\eta^2 \cdot \mathbb{E}\left(\frac{1}{\mathbf{x} \cdot \mathbf{x}}\right)$$

We note that as $N \to \infty$, the variance of $\frac{\mathbf{x} \cdot \mathbf{x}}{N}$ decays to 0, and since this is a non-negative quantity, by Jensen's inequality we obtain

$$\mathbb{E}\left(\frac{N}{\mathbf{x} \cdot \mathbf{x}}\right) \to \frac{N}{\mathbb{E}(\mathbf{x} \cdot \mathbf{x})} = \frac{1}{\sigma_x^2}$$

from above as $N \to \infty$. The desired result follows directly.

3.8 Let us consider the regression model

$$\boldsymbol{y_n = \theta^T x_n + \eta_n, \quad n = 1, 2, 3, ..., n}$$

where the noise samples $\boldsymbol{\eta = [\eta_1, ..., \eta_N]^T}$ come from a zero-mean Gaussian random vector with covariance matrix $\boldsymbol{\Sigma_\eta}$. If X is our input matrix and y is our output vector, show that

$$\boldsymbol{\hat{\theta} = (X^T \Sigma_\eta^{-1} X)^{-1} X^T \Sigma_\eta^{-1} y}$$

is a sufficient estimate.

Note that the above expression may be written as

$$\theta + (X^T\Sigma_\eta^{-1}X)^{-1}X^T\Sigma_\eta^{-1}\eta$$

The covariance of this expression is just

$$(X^T\Sigma_\eta^{-1}X)^{-1}$$

The likelihood function here is given by

$$p(y=a) = p(\eta = a - \theta^T X) = \frac{1}{((2\pi)^d|\Sigma_\eta|)^{\frac{1}{2}}}\prod_{n=1}^{N} e^{\frac{-(a-\theta^T X)^T\Sigma_\eta^{-1}(a-\theta^T X)}{2}}$$

From this, we obtain the Fisher information matrix as simply $(X^T\Sigma_\eta^{-1}X)$, and parameter efficiency via Cramer-Rao follows trivially.

3.9 Assume a set of i.i.d. $\boldsymbol{X = \{x_1, x_2, ..., x_N\}}$ samples of a random variable[sic] with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2}$. Define also the quantities

$$\boldsymbol{S_\mu := \frac{1}{N}\sum_{n=1}^{N} x_n}$$

$$\boldsymbol{S_{\sigma^2} := \frac{1}{N}\sum_{n=1}^{N}(x_n - S_\mu)^2}$$

$$\boldsymbol{\bar{S}_{\sigma^2} := \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2}$$

Show that is $\boldsymbol{\mu}$ is considered to be known, a sufficient statistic for $\boldsymbol{\sigma^2}$ is $\boldsymbol{\bar{S}_{\sigma^2}}$. Moreover in the case where both $\boldsymbol{(\mu, \sigma^2)}$ are unknown, then a sufficient statistic is the pair $\boldsymbol{(S_\mu, S_{\sigma^2})}$

This is not well-posed in its current form. It does however make sense if we constrain the x_i to be drawn from a Gaussian distribution.

We note the the p.d.f of X is

$$p(X=x) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}e^{\frac{-\sigma^2(x-\mu)^T(x-\mu)}{2}}$$

This can be written as

$$\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}e^{\frac{-\sigma^2 N\bar{S}_{\sigma^2}}{2}}$$

which is a function of σ and $\bar{S}_{\sigma^2}$ only. Therefore by Fisher-Neyman we have the first result. For the second part, we note that the p.d.f may alternatively be written as

$$\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{\frac{-N\sigma^2(S_{\sigma^2}-2\mu S_\mu+\mu^2+S_\mu^2)}{2}}$$

which is solely a function of μ, σ^2, S_μ, and S_{σ^2}. Therefore we have the second result by Fisher-Neyman.