ECE 275A – Homework 7 – Solutions

Solutions

1. For the same specification as in Homework Problem 6.11 we want to determine an estimator for θ using the Method of Moments (MOM). In general, the MOM estimator is neither unbiased nor efficient, so if either of these properties (or any other properties, other than consistency) are found to be the case *it is purely accidental.*

   We have the sample moment

   $$\mathrm{m}_1 = \langle y_k \rangle = \frac{1}{m}\sum_{k=1}^{m} y_k$$

   and the ensemble moment

   $$\alpha_1(\theta) = \mathrm{E}_\theta\left\{y_k\right\} \; .$$

   Equating the sample and ensemble moments yields the estimator

   $$\widehat{\theta} = 2\,\mathrm{m}_1 = 2\,\langle y_k \rangle = \frac{2}{m}\sum_{k=1}^{m} y_k \, .$$

   Although this turns out to be the BLUE, this is a *purely accidental* result. The estimator $\widehat{\theta}$ is easily shown to be unbiased (again, an accidental result). The estimator error variance was computed in Homework Problem 6.11, and goes to zero as $m \to \infty$. This shows that the estimator converges in the mean-square sense to the true parameter, and therefore is consistent (converges in probability). This is as expected, as consistency is the one property that the MOM estimator does generally possess.

2. Here the problem specification common to the previous two problems has been modified to $y_k \sim \mathrm{U}[-\theta, \theta]$. Now we have $\alpha_1(\theta) = 0$ so we cannot obtain a MOM estimate of θ by merely equating the first-order sample and ensemble moments. The second–order sample and ensemble moments are given respectively by

   $$\mathrm{m}_2 = \langle y_k^2 \rangle = \frac{1}{m}\sum_{k=1}^{m} y_k^2$$

   and

   $$\alpha_2(\theta) = \mathrm{E}_\theta\left\{y_k^2\right\} = \frac{(2\,\theta)^2}{12} = \frac{\theta^2}{3} \, .$$

   Of course $\mathrm{E}_\theta\left\{\mathrm{m}_2\right\} = \alpha_2(\theta)$. Equating the second–order sample and ensemble moments yields the MOM estimate

   $$\widehat{\theta} = \sqrt{3\,\mathrm{m}_2} \; .$$

Because the square–root function is strictly concave, Jensen's inequality[1] yields

$$\mathrm{E}_\theta\left\{\widehat{\theta}\right\} = \mathrm{E}_\theta\left\{\sqrt{3\,\mathrm{m}_2}\right\} < \sqrt{\mathrm{E}_\theta\left\{3\,\mathrm{m}_2\right\}} = \theta$$

showing that in general $\widehat{\theta}$ is biased. From the carry–over property of convergence in probability we know that the MOM estimator $\widehat{\theta}$ is consistent because the sample moments are consistent estimators of the ensemble moments. Because the estimator is consistent, it must be asymptotically unbiased.

Unfortunately, it is difficult to demonstrate consistency by directly computing the mean squared-error (mse) of the estimator. One can readily show that

$$0 \le \mathrm{mse}_\theta = \mathrm{E}_\theta\left\{(\widehat{\theta} - \theta)^2\right\} = 2\,\theta\left(\theta - \mathrm{E}_\theta\left\{\widehat{\theta}\right\}\right)$$

which, unfortunately, just leads to the already–verified condition that

$$\mathrm{E}_\theta\left\{\widehat{\theta}\right\} \le \theta\,.$$

Note from the mean squared–error equation that if $\widehat{\theta}$ is asymptotically unbiased,

$$\mathrm{E}_\theta\left\{\widehat{\theta}\right\} \to \theta \quad \text{as} \quad m \to \infty\,,$$

then we have mean-square convergence (and therefore consistency). Of course, if we have convergence we must have asymptotic unbiasedness. We have apparently arrived at a nice little tautological impasse!

Actually, one can break this impasse by using the statistical linearization technique discussed in Section 9.5 of Kay. However, I do not intend for this problem to be *that* lengthy. Thus, consider the carry-over argument to be a sufficient proof of consistency.

3. Kay 7.3. (a) Unit Variance Gaussian with unknown mean μ. We have that,

$$p(x;\mu) = \phi(x;\mu) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(x-\mu)^2\right\}\,.$$

With iid data $\mathbf{x} = (x[1],\cdots,x[N])^T$, the likelihood function is proportional to

$$p(\mathbf{x};\mu) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\sum_{n=1}^{N}(x[n]-\mu)^2\right\}\,.$$

[1] Jensen's inequality is an important and fundamental mathematical tool used in statistics, information theory, machine learning, and parameter estimation. If you don't know it, you should search in Moon & Stirling and/or Wikipedia and read a description.

To determine a solution to the problem,

$$\widehat{\mu}_{\mathrm{ML}} = \arg\max_{\mu} \ln p(\mathbf{x};\mu)$$

we look at the solutions to the likelihood equation,

$$0 = S(\mu;\mathbf{x}) = \nabla_{\mu} \ln p(\mathbf{x};\mu) = \sum_{n=1}^{N} (x[n] - \mu)\,,$$

where $S(\mu;\mathbf{x})$ denotes the score function. There is only one unique solution to the likelihood equation,

$$\widehat{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x[n]\,.$$

Furthermore the hessian of the log-likelihood has the value,

$$\frac{\partial^2}{\partial\mu^2}\ln p(\mathbf{x};\mu) = \frac{\partial}{\partial\mu} S(\mu;\mathbf{x}) = -N < 0\,,$$

for all μ showing that the log–likelihood is globally concave and hence has a single (unique) maximum. The estimate $\widehat{\mu}_{\mathrm{ML}}$ is unbiased and efficient; it is known that among regular probability families only exponential families (see Kay problem 5.14) can yield finite-sample efficient estimators (as is the case here) and that when an efficient estimator exists, it can be found as a zero to the likelihood equation (as was done here, also see problem 6 below).

(b) Exponential distribution using the so-called "natural parameter" λ where

$$p(x;\lambda) = \lambda \exp\left\{-\lambda x\right\} \chi\left(x \geq 0\right)\,,$$

where $\chi(\cdot)$ is the characteristic (or indicator) function. Here we have,

$$p(\mathbf{x};\lambda) = \prod_{n=1}^{N} p(x[n];\lambda) = \lambda^N \exp\left\{-\lambda \sum_{n=1}^{N} x[n]\right\} \chi\left(x[n] \geq 0,\, n = 1, \cdots, N\right)\,.$$

The maximum likelihood estimate is found as the unique zero of the likelihood equation,

$$\widehat{\lambda}_{\mathrm{ML}} = \frac{N}{\sum_{n=1}^{N} x[n]}\,.$$

Note that the maximum likelihood estimate is biased, and therefore *cannot* be the UMVUE of λ. The hessian of the log-likelihood is given by $-\frac{N}{\lambda^2}$, which is negative for all (positive) λ. This shows that the log–likelihood is globally concave and has only one unique maximum.

4. Kay 7.8. See worked example 12.2.1 on page 547 of Moon & Stirling where it is shown that $\hat{p} = \frac{M}{N}$ where $M = \sum_{n=1}^{N} x[n]$ is the number of times that the outcome "1" occurs.

   Note that it is simple to generalize this result to the multinomial case having $\ell \geq 2$ possible distinct outcomes $O_1, \cdots, O_\ell$ with respective probabilities $p_1, \cdots, p_\ell$. Perhaps the simplest way to do so is to recognize that one can focuss on any one outcome, say the j-th outcome O_j, and examine the binary problem of "j or Not–j". With N_j the number of times that O_j occurs and $N = N_1 + \cdots + N_\ell$ the total number of observations, we can immediately apply the maximum likelihood solution for the binary case to determine the MLE $\hat{p}_j = \frac{N_j}{N}$.

   We can now see that Worked Example 12.1.2 on page 543 of Moon & Stirling requires some clarification. If the samples in this example are drawn from a multinomial distribution,[2] then with nonzero probability the sample values are *not all distinct* and the derived result that all probabilities are $\frac{1}{m}$, where m is the total number of samples, can be *wrong* with nonzero probability. Indeed, if there are ℓ distinct multinomial outcomes and $m > \ell$, for m a non–integer multiple of ℓ, then this answer is always wrong!

   What Moon & Stirling forgot to mention is their key assumption that the random variable X has a *continuous* distribution function. Under this assumption the probability that any two samples have equal value is zero, in which event the result derived by Moon & Stirling that all samples have an estimated probability of $\frac{1}{m}$ of occurring is true with probability one. Note that $\frac{1}{m} \to 0$ as $m \to \infty$ which is consistent with the fact that for a continuously distributed random the probability of occurrence of any particular value is zero.

5. Kay 7.9. See worked example 12.1.3 on page 545 of Moon & Stirling.

6. Kay 7.12. First recall that for appropriately smooth regular probability distributions, MLE's are solutions to the likelihood equation[3]

$$0 = S(y; \hat{\theta}_{\text{mle}}(y)) = S(y; \theta)\Big|_{\theta = \hat{\theta}_{\text{mle}}(y)} .$$

   Note that knowing the maximum likelihood estimate $\hat{\theta}_{\text{mle}}(y)$ for every possible value of y is equivalent to knowing the maximum likelihood estimator $\hat{\theta}_{\text{mle}}(\cdot)$.

[2] Moon & Stirling say that the distribution is unknown, so it *could* be multinomial, or some other discrete or noncontinuous probability distribution.

[3] In other words, a maximum likelihood *estimate* $\hat{\theta}_{\text{mle}}(y)$ (note the dependence on the instantiated measurement value y) is a value of θ which is a zero of the score function.

Now recall that a necessary condition for an *estimator* $\hat{\theta}_{\text{eff}}(\cdot)$ to be efficient is that the score can be written as,[4]

$$S(y;\theta) = J(\theta)\left(\hat{\theta}_{\text{eff}}(y) - \theta\right),$$

where $\hat{\theta}_{\text{eff}}(y)$ is the *estimate* obtained from the efficient estimator given an instantiation value y.

Combining these two conditions,

$$0 = S(y;\hat{\theta}_{\text{mle}}(y)) = J(\hat{\theta}_{\text{mle}}(y))\left(\hat{\theta}_{\text{eff}}(y) - \hat{\theta}_{\text{mle}}(y)\right),$$

we see (assuming that the Fisher information matrix $J(\hat{\theta}_{\text{mle}}(y))$ is full rank so that its nullspace is trivial[5]) that,

$$\hat{\theta}_{\text{mle}}(y) = \hat{\theta}_{\text{eff}}(y).$$

Since the instantiation values (i.e., the two *estimates*) $\hat{\theta}_{\text{mle}}(y)$ and $\hat{\theta}_{\text{eff}}(y)$ are equal for all possible instantiated measurement values y, it is the case that we have equality as *estimators*,[6,7]

$$\hat{\theta}_{\text{mle}}(\cdot) = \hat{\theta}_{\text{eff}}(\cdot).$$

7. Prove the *Invariance Principle of Maximum Likelihood Estimation* (Theorem 7.4) assuming that $g(\cdot)$ is many–to–one.

   The goal is to show that $\widehat{\alpha}_{\text{ML}} = g(\widehat{\theta}_{\text{ML}})$, for any arbitrary many-to-one function $g(\cdot)$.

   To understand the solution to this problem, you need to know that *any* (perhaps many–to–one and/or concontinuous) function induces a disjoint partition of its domain. Thus if $\alpha = g(x)$ is a many to one function with domain Θ, and if $g^{-1}(\alpha) \subset \Theta$ denotes the preimage set of $\alpha \in g(\Theta)$, it is always the case that[8]

$$\Theta = \bigcup_{\alpha \in g(\Theta)} g^{-1}(\alpha) \tag{1}$$

[4] This also a sufficient condition if $\hat{\theta}_{\text{eff}}(\cdot)$ is uniformly unbiased.

[5] This is usually the case because $J(\theta)$ can happen to have full rank for all possible values of θ, which is equivalent to the identifiability assumption on the statistical family

[6] I.e., if for two functions f and g we have equality of the *function values*, $f(y) = g(y)$, for all domain elements y, then *merely by definition* of "function", we have equality of the two *functions*. In this situation f and g are merely *two different names* for the *same function*.

[7] When reading the proof of the result in textbooks and/or on-line, you should note the common overloaded use of the notation $\hat{\theta}_{\text{mle}}$ to denote both the estimate $\hat{\theta}_{\text{eff}}(y)$ and the estimator $\hat{\theta}_{\text{mle}}(\cdot)$. Similarly, $\hat{\theta}_{\text{eff}}$ is used to denote both $\hat{\theta}_{\text{eff}}(y)$ and $\hat{\theta}_{\text{eff}}(\cdot)$. Furthermore, θ is to be understood as a *deterministic* parameter vector, *not* as a realization value (much less a function) because *not* being a random variable the concept of a realization *makes no sense*. It is necessary to keep these tacit and context-dependent interpretations straight in one's head in order to avoid confusion. As in Law, the fine-print matters.

[8] This is due to the definition of what it means to be a function; specifically that a function *must be single-valued* by definition.

and

$$g^{-1}(\alpha) \bigcap g^{-1}(\alpha') = \emptyset\,, \quad \forall \alpha \neq \alpha'\,.$$

The set $g^{-1}(\alpha)$ is a so–called *equivalence class* indexed by α. The elements of $g^{-1}(\alpha)$ are all “equivalent” in the sense that they each produce the same (i.e., equivalent) value α under the action of $g(\cdot)$.[9]

Because of the fact (**??**), it is the case that the maximum value of the likelihood function $\ell(\theta) = p_\theta(x)$ can be determined in two equivalent ways,

$$\max_{\alpha \in g(\Theta)} \max_{\theta \in g^{-1}(\alpha)} p_\theta(x) = \max_{\theta \in \Theta} p_\theta(x)\,.$$

This expression is true because in *both* sides of the equality *all* possible values of θ in Θ are considered when searching for the maximum of $\ell(\theta) = p_\theta(x)$ over Θ.

For $\alpha = g(x)$, the so-called induced (or modified) likelihood function $\ell(\alpha)$ is defined by

$$\ell(\alpha) = f_\alpha(x) \triangleq \max_{\theta \in g^{-1}(\alpha)} p_\theta(x)\,.$$

Note that $f_\alpha(\cdot)$ determines a family of probability distribution parameterized by α.

By definition

$$\widehat{\theta}_{\text{ML}} = \arg \max_{\theta \in \Theta} p_\theta(x) \qquad \text{and} \qquad \widehat{\alpha}_{\text{ML}} = \arg \max_{\alpha \in g(\Theta)} f_\alpha(x)\,.$$

Based on the above discussion, it should be clear that the maximum values attained in the optimizations are equal:

$$f_{\widehat{\alpha}_{\text{ML}}}(x) = \max_{\alpha \in g(\Theta)} f_\alpha(x) = \max_{\alpha \in g(\Theta)} \max_{\theta \in g^{-1}(\alpha)} p_\theta(x) = \max_{\theta \in \Theta} p_\theta(x) = p_{\widehat{\theta}_{\text{ML}}}(x)\,.$$

Using the definition of the induced likelihood this is equivalent to

$$f_{\widehat{\alpha}_{\text{ML}}}(x) = \max_{\theta \in g^{-1}(\widehat{\alpha}_{\text{ML}})} p_\theta(x) = p_{\widehat{\theta}_{\text{ML}}}(x)$$

showing that it must be the case that

$$\widehat{\theta}_{\text{ML}} \in g^{-1}(\widehat{\alpha}_{\text{ML}})\,.$$

Of course, this last statement is equivalent to our desired result:

$$\widehat{\alpha}_{\text{ML}} = g(\widehat{\theta}_{\text{ML}})\,.$$

[9] For a discussion of equivalence classes and their relationship to function–induced partitions see *Linear Operator Theory in Engineering and Science,* A.W. Naylor and G.R. Sell, or *Some Modern Mathematics for Physicists and other Outsiders, Vol. 1: Algebra, Topology, and Measure Theory,* P. Roman.

8. Kay 7.17. Let x_k, $k = 1, \cdots, n$ be n iid samples drawn from a $N(0, \frac{1}{\theta})$ pdf. If you happen to know that the MLE of the variance is $\frac{1}{n}\sum_{k=1}^{n} x_k^2$, then the invariance principle immediately gives $\hat{\theta}_{\text{ML}} = \frac{n}{\sum_{k=1}^{n} x_k^2}$. However, we also want to determine the CRLB, which is the asymptotic variance of the asymptotically efficient MLE, so a more complete analysis is required.

   With
$$p_\theta(\mathbf{x}) = \left(\frac{\theta}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{\theta}{2}\sum_{k=1}^{n} x_k^2\right\},$$
   we have
$$S(\mathbf{x};\theta) = \frac{\partial \log p_\theta(\mathbf{x})}{\partial\theta} = \frac{\partial\left(\frac{n}{2}\log\theta - \frac{\theta}{2}\sum_{k=1}^{n} x_k^2\right)}{\partial\theta} = \frac{n}{2\theta} - \frac{1}{2}\sum_{k=1}^{n} x_k^2.$$
   Solving the likelihood equation $S(\mathbf{x};\theta) = 0$ yields
$$\hat{\theta}_{\text{ML}} = \frac{n}{\sum\limits_{k=1}^{n} x_k^2}$$
   as expected. The Fisher information is determined as
$$J(\theta) = -\mathrm{E}\left\{\frac{\partial S(\mathbf{x};\theta)}{\partial\theta}\right\} = \frac{n}{2\theta^2}.$$
   In the limit as $n \to \infty$ we have
$$\hat{\theta}_{\text{ML}} \overset{a}{\sim} N\left(\theta, J^{-1}(\theta)\right) = N\left(\theta, \frac{2\theta^2}{n}\right),$$
   where θ is the true unknown parameter and we use a notation which emphasizes the fact that this is an asymptotic result. Note, in particular, that the asymptotic error covariance depends on the unknown parameter θ. If the Fisher information is a continuous function of the unknown parameter (as in our case) it can be shown that $J(\hat{\theta}_{\text{ML}})$ is a consistent estimator of $J(\theta)$. In this case we have that
$$\hat{\theta}_{\text{ML}} \overset{a}{\sim} N\left(\theta, J^{-1}(\hat{\theta}_{\text{ML}})\right) = N\left(\theta, \frac{2\hat{\theta}_{\text{ML}}^2}{n}\right)$$
   enabling us to empirically construct confidence intervals[10] about our estimate $\hat{\theta}_{\text{ML}}$ in the limit as $n \to \infty$.

[10] Also known as "error bars."

9. Kay 7.21. Also determine the MLE of the standard deviation.

   The maximum likelihood estimates of the mean, $\hat{A}_{\text{ML}}$ and the variance $\widehat{\sigma^2}_{\text{ML}}$ are derived in Example 7.12 on page 183 of Kay. Invoking the principle of invariance we have

$$\hat{\alpha}_{\text{ML}} = \frac{\hat{A}^2_{\text{ML}}}{\widehat{\sigma^2}_{\text{ML}}} \quad \text{and} \quad \hat{\sigma}_{\text{ML}} = \sqrt{\widehat{\sigma^2}_{\text{ML}}}\,.$$