T-61.3040 Statistical Signal Modeling Lendasse, Miche, Eirola
Exercise 2, September 22, 2011, Solutions

1. In this problem we consider a probability distribution that consists of both point masses and a continuous density function. These are combined into a single probability density function by representing the point masses in terms of Dirac delta function (actually a distribution). Then the expectations can be mechanically evaluated by integrating over this distribution.

   (a) K = total amount of water. The density function of K is

   $$p_K(k) = 0.5\delta(k) + 0.5p_U(k)$$

   where $\delta(k)$ is the Dirac delta function and $p_U(k)$ is the pdf of the uniform distribution in the interval $[0, 1]$, i.e. $p_U(k) = 1$, when $0 \leq k \leq 1$ and zero otherwise.

   (b)

   $$\begin{aligned} \mathrm{E}(K) &= \int k p_K(k)\, dk = 0.5 \int k\delta(k)\, dk + 0.5 \int_0^1 k\, dk \\ &= 0 + 0.5 \times 0.5 = \frac{1}{4} \end{aligned}$$

   (c)

   $$\begin{aligned} \mathrm{Var}(K) &= \mathrm{E}((K - 1/4)^2) = \int (k - 1/4)^2 p_K(k)\, dk \\ &= 0.5 \int \delta(k)(k - 1/4)^2\, dk + 0.5 \int_0^1 (k - 1/4)^2\, dk \end{aligned}$$

   According to the integration rules of the Dirac delta function, the value of the first integral is the value of the function $(k - 1/4)^2$ evaluated at he point $k = 0$. So we get

   $$\mathrm{Var}(K) = 0.5 \times (1/4)^2 + 0.5 \int_0^1 (k - 1/4)^2 dk = \frac{1}{32} + \frac{7}{96} = \frac{5}{48}$$

2. (a) The observations originate from a normal distribution with the density function

   $$p(x(i)|\mu, \sigma^2) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x(i) - \mu)^2}{2\sigma^2}\right)$$

   The logarithm of the density function gives the log-likelihood function J. Because the observations are independent, we get

   $$\begin{aligned} J &= \log p(x(1), \ldots, x(N)|\mu, \sigma^2) = \log\left(\prod_{i=1}^{N} p(x(i)|\mu, \sigma^2)\right) \\ &= \sum_{i=1}^{N} \log p(x(i)|\mu, \sigma^2) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2}\sigma^{-2}\sum_{i=1}^{N}(x(i) - \mu)^2. \end{aligned}$$

J can be maximised simultaneously in terms of both mean and variance by requiring the partial derivatives w.r.t. these variables to be zero:

$$\begin{cases} \frac{\partial J}{\partial \mu} = \sigma^{-2} \sum_{i=1}^{N} (x(i) - \mu) = 0 \\ \frac{\partial J}{\partial \sigma^2} = -\frac{N}{2}\sigma^{-2} + \frac{1}{2}\sigma^{-4} \sum_{i=1}^{N} (x(i) - \mu)^2 = 0 \end{cases}$$

The solution to the above pair of equations is

$$\begin{cases} \mu = \frac{1}{N} \sum_{i=1}^{N} x(i) \\ \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x(i) - \mu)^2 \end{cases}$$

By inserting the formula for μ in the first of the equations into the second one, one obtains the maximum likelihood (ML) estimator for the variance:

$$\hat{\sigma}^2_{ML} = \frac{1}{N} \sum_{i=1}^{N} (x(i) - \hat{\mu})^2$$

where $\hat{\mu}$ is the sample mean

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x(i).$$

(b) An estimator of variance $\hat{\sigma}^2$ is unbiased if

$$E[\hat{\sigma}^2 | \sigma^2] = \sigma^2$$

One unbiased estimator can be formed by first calculating the bias of the ML estimator in the part (a) and then multiplying the estimator with a scaling factor that neutralises the bias.
The ML estimator in part (a) is

$$\hat{\sigma}^2_{ML} = \frac{1}{N} \sum_{i=1}^{N} (x(i) - \hat{\mu})^2.$$

The expected value of the estimator is

$$\begin{aligned} E[\hat{\sigma}^2_{ML}] &= E\left[\frac{1}{N} \sum_{i=1}^{N} (x(i) - \hat{\mu})^2 \right] = E\left[\frac{1}{N} \sum_{i=1}^{N} (x(i) - \mu + \mu - \hat{\mu})^2 \right] \\ &= E\left[\frac{1}{N} \sum_{i=1}^{N} (x(i) - \mu)^2 - 2(\hat{\mu} - \mu) \frac{1}{N} \sum_{i=1}^{N} (x(i) - \mu) + \frac{1}{N} \sum_{i=1}^{N} (\hat{\mu} - \mu)^2 \right] \\ &= E\left[\frac{1}{N} \sum_{i=1}^{N} (x(i) - \mu)^2 - 2(\hat{\mu} - \mu) \Bigg(\underbrace{\left[\frac{1}{N} \sum_{i=1}^{N} x(i) \right]}_{=\hat{\mu}} - \mu \Bigg) + (\hat{\mu} - \mu)^2 \right] \\ &= E\left[\frac{1}{N} \sum_{i=1}^{N} (x(i) - \mu)^2 - 2(\hat{\mu} - \mu)^2 + (\hat{\mu} - \mu)^2 \right] \\ &= \frac{1}{N} \sum_{i=1}^{N} E\left[(x(i) - \mu)^2 \right] - E\left[(\hat{\mu} - \mu)^2 \right] \end{aligned}$$

In the last form

$$E\left[(x(i)-\mu)^2\right] = \sigma^2.$$

The expectation $E\left[(\hat{\mu}-\mu)^2\right]$ in turn is the variance of the estimator $\hat{\mu}$ because $\hat{\mu}$ is unbiased. By employing the calculation rules for variance, one obtains

$$E\left[(\hat{\mu}-\mu)^2\right] = \mathrm{Var}(\hat{\mu}) = N\,\mathrm{Var}\left(\frac{1}{N}x(i)\right) = \frac{N}{N^2}\,\mathrm{Var}(x(i)) = \frac{\sigma^2}{N}.$$

Using these results, one gets

$$E[\hat{\sigma}^2_{ML}] = \sigma^2 - \frac{\sigma^2}{N} = \frac{N-1}{N}\sigma^2$$

Therefore, an unbiased estimator can be obtained by multiplying the ML estimator with the number $\frac{N}{N-1}$, i.e.

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x(i)-\hat{\mu})^2.$$

(c) We are in search for an estimator of the form

$$\hat{\sigma}^2_{MSE} = c\hat{\sigma}^2$$

where $\hat{\sigma}^2$ is the unbiased estimator from part (b).
The MSE of on the formulated estimator is

$$J_{MSE} = \mathrm{E}[(c\hat{\sigma}^2-\sigma^2)^2] = c^2\,\mathrm{E}[(\hat{\sigma}^2)^2] - 2c\sigma^2\,\mathrm{E}[\hat{\sigma}^2] + (\sigma^2)^2$$

Let's differentiate the MSE w.r.t. c and set the derivative to zero. This gives

$$c = \frac{\sigma^2\,\mathrm{E}[\hat{\sigma}^2]}{\mathrm{E}[(\hat{\sigma}^2)^2]}.$$

From this point on, the details of the solution do not belong to the essential contents of this course. They are shown here, however, for the sake of completeness and also because the end result is interesting.
Of the required expectations, $\mathrm{E}[\hat{\sigma}^2]$ is easily evaluated as $\mathrm{E}[\hat{\sigma}^2] = \sigma^2$ since $\hat{\sigma}^2$ is an unbiased estimator. It is furthermore known that the unbiased estimator $\hat{\sigma}^2$ in the case of Gaussian observations is distributed according to the χ^2-distribution with $N-1$ degrees of freedom[1] in the sense that

$$\frac{N-1}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{N-1}$$

A variable distributed according to the χ^2-distribution with $N-1$ degrees of freedom has expectation $N-1$ and variance $2(N-1)$. Thus

$$\mathrm{E}[\hat{\sigma}^2] = \sigma^2 \quad \text{and} \quad \mathrm{Var}[\hat{\sigma}^2] = \frac{2\sigma^4}{N-1}$$

[1] This can be shown using Cochran's theorem.

Since for any estimator $\mathrm{E}[\theta^2] = \mathrm{Var}[\theta] + \mathrm{E}[\theta]^2$, we get that

$$\mathrm{E}[(\hat{\sigma}^2)^2] = \mathrm{Var}[\hat{\sigma}^2] + \mathrm{E}[\hat{\sigma}^2]^2 = \frac{2\sigma^4}{N-1} + (\sigma^2)^2 = \frac{N+1}{N-1}\sigma^4$$

We can now evaluate c to be

$$c = \frac{\sigma^2 \,\mathrm{E}[\hat{\sigma}^2]}{\mathrm{E}[(\hat{\sigma}^2)^2]} = \frac{(N-1)\sigma^2\sigma^2}{(N+1)\sigma^4} = \frac{N-1}{N+1}$$

One thus gets

$$\hat{\sigma}^2_{MSE} = \frac{N-1}{N+1}\hat{\sigma}^2 = \frac{1}{N+1}\sum_{i=1}^{N}(x(i) - \hat{\mu})^2.$$

This is not an unbiased estimator for the variance. Thus, the ML and MSE procedures result in different estimtors, neither of which is unbiased. We can additionally observe that $\hat{\sigma}^2_{\text{unbiased}} > \hat{\sigma}^2_{ML} > \hat{\sigma}^2_{MSE}$.
In a general case forming an MSE estimator is difficult, if not impossible, if one does not restrict oneself to estimators of a particular form. It has been shown that there exist estimators with a smaller MSE than the one above, but they are problematic to apply in practice.

3. (a)

$$\begin{aligned}\mathrm{E}(\hat{\theta}_3) &= \mathrm{E}[\alpha\hat{\theta}_1 + (1-\alpha)\hat{\theta}_2] = \alpha\,\mathrm{E}(\hat{\theta}_1) + (1-\alpha)\,\mathrm{E}(\hat{\theta}_2) \\ &= \alpha\theta + (1-\alpha)\theta = \theta.\end{aligned}$$

Thus also $\hat{\theta}_3$ is unbiased.

(b)

$$\begin{aligned}J_{MSE}(\hat{\theta}_3) &= \mathrm{E}\{[\hat{\theta}_3 - \theta]^2\} \\ &= \mathrm{E}\{[\alpha\hat{\theta}_1 + (1-\alpha)\hat{\theta}_2 - \theta]^2\} \\ &= \mathrm{E}\{[\alpha(\hat{\theta}_1 - \theta) + (1-\alpha)(\hat{\theta}_2 - \theta)]^2\} \\ &= \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 + 2\alpha(1-\alpha)\,\mathrm{E}\{[\hat{\theta}_1 - \theta][\hat{\theta}_2 - \theta]\}.\end{aligned}$$

$\hat{\theta}_1$ and $\hat{\theta}_2$ are independent $\Longrightarrow \mathrm{E}(\hat{\theta}_1\hat{\theta}_2) = \mathrm{E}(\hat{\theta}_1)\,\mathrm{E}(\hat{\theta}_2)$, so that

$$\begin{aligned}\mathrm{E}\{[\hat{\theta}_1 - \theta][\hat{\theta}_2 - \theta]\} &= \mathrm{E}(\hat{\theta}_1\hat{\theta}_2) - \mathrm{E}(\theta)\,\mathrm{E}(\hat{\theta}_1) - \mathrm{E}(\theta)\,\mathrm{E}(\hat{\theta}_2) + \mathrm{E}^2(\theta) \\ &= \mathrm{E}^2(\theta) - \mathrm{E}^2(\theta) - \mathrm{E}^2(\theta) + \mathrm{E}^2(\theta) = 0\end{aligned}$$

Therefore,

$$J_{MSE}(\hat{\theta}_3) \;=\; \underline{\alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2}.$$

(c) We set the derivative of MSE w.r.t. α to zero:

$$\frac{\partial\,\mathrm{E}\{[\hat{\theta}_3 - \theta]^2\}}{\partial\alpha} = 2\alpha\sigma_1^2 - 2(1-\alpha)\sigma_2^2 = 0$$

$$\Longrightarrow \alpha = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

4. The observations are independent, so that their joint density is

$$f(\mathbf{z}|\lambda) = f(z_1, \ldots, z_M|\lambda) = \prod_{i=1}^{M} f(z_i|\lambda) = \prod_{i=1}^{M} e^{-\lambda}\lambda_i^z/z_i!$$

The maximum likelihood estimator is obtained by maximising the log-likelihood function

$$\begin{aligned}\log f(\mathbf{z}|\lambda) &= \sum_{i=1}^{M} \log f(z_i|\lambda) = \sum_{i=1}^{M} \log(e^{-\lambda}\lambda^{z_i}/z_i!) \\ &= -\lambda M + \sum_{i=1}^{M} z_i \log(\lambda) - \sum_{i=1}^{M} \log(z_i!).\end{aligned}$$

The derivative w.r.t. the estimated parameter λ is set to zero:

$$\frac{\partial}{\partial\lambda} \log f(\mathbf{z}|\lambda) = -M + \frac{1}{\lambda}\sum_{i=1}^{M} z_i = 0,$$

resulting in

$$\hat{\lambda}_{\mathrm{ML}} = \frac{1}{M}\sum_{i=1}^{M} z_i$$

5. Describing the orthogonality principle is simpler in terms of vectors and matrices. Denote

$$\begin{aligned}\mathbf{y} &= [y(0)\, y(1)\, \ldots\, y(M)]^T \\ \mathbf{b} &= [b(1)\, b(2)\, \ldots\, b(N)]^T \\ \mathbf{X} &= \begin{bmatrix} x_1(0) & \cdots & x_N(0) \\ \vdots & \ddots & \vdots \\ x_1(M) & \cdots & x_N(M) \end{bmatrix} = [\mathbf{x}_1\, \ldots\, \mathbf{x}_N] \\ \mathbf{e} &= [e(0)\, e(1)\, \ldots\, e(M)]^T\end{aligned}$$

Then the model can be written as

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \sum_{i=1}^{N} \mathbf{x}_i b(i)$$

(a) The goal is to minimise the mean squared error w.r.t. the coefficients $b(i)$. Equivalently, we can minimise the quadratic form $\mathbf{e}^T\mathbf{e}$ of the error $\mathbf{e}$. The necessary condition for the minimum is that the partial derivatives w.r.t. each parameter $b(i)$ are all zero:

$$\frac{\partial}{\partial b(i)}\mathbf{e}^T\mathbf{e} = 0, \quad \forall i = 1, \ldots, N$$

Differentiate:

$$\begin{aligned}\frac{\partial}{\partial b(i)}\mathbf{e}^T\mathbf{e} &= \left(\frac{\partial}{\partial b(i)}\mathbf{e}\right)^T \mathbf{e} + \mathbf{e}^T\left(\frac{\partial}{\partial b(i)}\mathbf{e}\right) = 2\mathbf{e}^T\frac{\partial}{\partial b(i)}\mathbf{e} \\ &= 2\mathbf{e}^T\frac{\partial}{\partial b(i)}\left(\mathbf{y} - \sum_{j=1}^{N}\mathbf{x}_j b(j)\right) = -2\mathbf{e}^T\mathbf{x}_i\end{aligned}$$

The orthogonality principle is obtained by setting the derivative to zero:

$$\mathbf{e}^T \mathbf{x}_i = 0, \quad \forall i = 1, \ldots, N$$

(b) We will assume that the columns of $\mathbf{X}$ are linearly independent. This implies that $M + 1 \geq N$, i.e., there are at least as many observations (years) of each signal as there are signals (measuring stations).

By the orthogonality principle, the solution must satisfy

$$\mathbf{x}_i^T \mathbf{e} = 0, \quad \forall i = 1, \ldots, N$$

In matrix form this can be written as $\mathbf{X}^T \mathbf{e} = \mathbf{0}$, that is:

$$\begin{aligned} \mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{b}) &= \mathbf{0} \\ \mathbf{X}^T\mathbf{y} &= \mathbf{X}^T\mathbf{X}\mathbf{b} \\ \mathbf{b} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \end{aligned}$$

The inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ exists because $\mathbf{X}$ was assumed to be full rank.

If $\mathbf{X}$ has linearly dependent columns, there are several solutions which give the minimum MSE. One solution is obtained by removing sufficiently many linearly dependent columns from $\mathbf{X}$ until it is full rank, and proceeding as above.