332:525 – Homework Set 1

Estimation Problems

1. *Recursive Least-Squares (RLS) Estimators:* Consider a sequence of iid random variables x_n, $n = 0, 1, \ldots$, and form the running average of the first $n+1$ numbers considered as an estimate of the mean $m = E[x_n]$:

$$\hat{m}_n = \frac{x_0 + x_1 + \cdots + x_n}{n+1}$$

   (a) Show that $\hat{m}_n$ is the optimum solution that minimizes the sum of squares:

$$\mathcal{E}_n = \sum_{k=0}^{n} (x_k - \hat{m})^2$$

   What is the minimized value of $\mathcal{E}_n$?

   (b) Moreover show that $\hat{m}_n$ can be re-expressed in the following time-recursive forms, with the second being a Kalman-type predictor/corrector form:

$$\hat{m}_n = \left(\frac{n}{n+1}\right)\hat{m}_{n-1} + \left(\frac{1}{n+1}\right)x_n$$

$$\hat{m}_n = \hat{m}_{n-1} + \left(\frac{1}{n+1}\right)(x_n - \hat{m}_{n-1})$$

   Note that these recursions connect the optimum least-squares solutions of *two* different performance indices. Indeed, $\hat{m}_n$ minimizes the performance index $\mathcal{E}_n$, whereas $\hat{m}_{n-1}$ minimizes $\mathcal{E}_{n-1}$ which runs up to time $k = n-1$.

   (c) Show that $\hat{m}_n$ is an unbiased estimator of the mean. Determine the variance of $\hat{m}_n$, that is, the quantity, $\text{var}(\hat{m}_n) = E[(\hat{m}_n - m)^2]$ and show that $\hat{m}_n$ is a consistent estimator of the mean.
   Hint: Show first that

$$\hat{m}_n - m = \frac{1}{n+1}\sum_{k=0}^{n}(x_k - m)$$

   and use the assumption that the x_n are iid, which implies the decorrelation condition: $E[(x_i - m)(x_j - m)] = \sigma_x^2 \delta_{ij}$.

2. *RLS Estimators with Forgetting Factor:* The RLS estimator $\hat{m}_n$ of the previous problem is appropriate for stationary sequences, that is, whose statistical characteristics don't change over time. Indeed, the performance index $\mathcal{E}_n$ treats all time samples—from the earliest to the latest—on a equal footing.

   Initially, the estimator $\hat{m}_n$ converges very fast to the optimum value m and then gets stuck at that optimum value because the Kalman-type gain factor that appears in the time-update becomes extremely small with increasing n. If there is a non-stationary change in the statistics and the mean m changes to a new value, the estimator $\hat{m}_n$ will have a very hard time tracking this change.

   A more appropriate estimator for tracking non-stationary changes in the statistics would be one that places more emphasis on the more recent data and less on the older data. For example, the following weighted version of $\mathcal{E}_n$ emphasizes the current samples more and forgets the older ones exponentially fast:

$$\mathcal{E}_n = \sum_{k=0}^{n} \lambda^{n-k}(x_k - \hat{m})^2$$

   where the forgetting factor λ must be $0 < \lambda \leq 1$. Note that $\lambda = 1$ recovers the above stationary case.

   (a) Determine the optimum $\hat{m}$ that minimizes $\mathcal{E}_n$ and cast it in a time-recursive form such as:

$$\hat{m}_n = \hat{m}_{n-1} + b_n(x_n - \hat{m}_{n-1})$$

   How does b_n behave in the limit $\lambda \to 1$? Show that $\hat{m}_n$ is an asymptotically unbiased estimator of m.

   (b) Show that for fairly large values of n and for $\lambda \neq 1$, the estimator satisfies the first-order difference equation (otherwise known as a first-order smoother):

$$\hat{m}_n = \lambda \hat{m}_{n-1} + (1-\lambda)x_n \tag{1}$$

3. *RLS Estimators with Forgetting Factor:* The first-order smoother estimator of Eq. (1) was obtained for fairly large values of n. However, it can be thought of as a third-type of estimator in its own right. Assume, therefore, that Eq. (1) defines $\hat{m}_n$ for all $n \geq 0$.

   Show that it is asymptotically unbiased but not consistent. Indeed, show that in the limit of large n, the variance of $\hat{m}_n$ tends to the finite value:

$$\text{var}(\hat{m}_n) = E\big[\big(\hat{m}_n - E[\hat{m}_n]\big)^2\big] \to \frac{1-\lambda}{1+\lambda}\sigma_x^2$$

However, by choosing $\lambda \simeq 1$ it can be made as small as desired, thus providing a good estimator. The trade-off is that the closer λ is to 1, the more sluggish the estimator becomes in tracking non-stationarities.

4. *Least-Mean-Square (LMS) Estimators:* Consider the theoretical performance index

$$\mathcal{E}(\hat{m}) = E[(x-\hat{m})^2] \tag{2}$$

(a) Differentiating it with respect to $\hat{m}$, show that $\mathcal{E}$ is minimized for the optimum value of the parameter $\hat{m} = m = E[x]$.

(b) The LMS algorithm is based on the idea of steepest descent in which $\hat{m}$ is changed iteratively so that at each iteration the performance index $\mathcal{E}$ is decreased and eventually it reaches its minimum value. The key condition is to demand that going from one value of $\hat{m}$ to the next, say, $\hat{m} + \Delta\hat{m}$, will result in a smaller performance index, that is, $\mathcal{E}(\hat{m}+\Delta\hat{m}) \le \mathcal{E}(\hat{m})$. This can be guaranteed by choosing the change $\Delta\hat{m}$ to be proportional to the negative gradient of $\mathcal{E}$, that is, (with $\mu > 0$)

$$\Delta\hat{m} = -\mu\frac{\partial\mathcal{E}}{\partial\hat{m}} \qquad \text{(LMS update)}$$

Replace the theoretical gradient by the instantaneous one:

$$\frac{\partial\mathcal{E}}{\partial\hat{m}} \to \frac{\widehat{\partial\mathcal{E}}}{\partial\hat{m}} = -2(x_n - \hat{m}_n)$$

Apply the LMS update to the instantaneous gradient, that is,

$$\hat{m}_{n+1} = \hat{m}_n + \Delta\hat{m}_n = \hat{m}_n - \mu\frac{\widehat{\partial\mathcal{E}}}{\partial\hat{m}_n}$$

And, show that it can be written in a similar form as the RLS estimator of Eq. (1):

$$\hat{m}_{n+1} = \lambda\hat{m}_n + (1-\lambda)x_n$$

where $\lambda = 1-2\mu$. Thus, the LMS and RLS algorithms for the recursive estimation of the mean are essentially equivalent. Note, however, that in adapting more than just one parameter, the LMS and RLS algorithms are no longer equivalent—the latter having a much faster learning speed at the expense of higher computational cost.

5. Do problems 1.9 and 1.10.

For Problem 1.10, suppose the mixing parameter ϵ is known in advance. Instead of sending x and y into a correlation canceler, you carry out a preprocessing operation, replacing $\{x, y\}$ by the signals $\{x_1, y_1\}$, where $x_1 = x$ and $y_1 = y - \epsilon x$, and then send those into a correlation canceler. Determine the optimum canceler weight H. Show that now the noise component of x can be canceled completely. Draw a block diagram of all the processing operations.

Note: The circumstances of this problem arise in adaptive antenna sidelobe canceling systems that use linearly polarized antennas. Polarization is used as a useful discriminant between signal and interference. In this application, the parameter ϵ is related to the *known* polarization angles of the desired signal. The interference signal is also polarized but with unknown polarization angles with respect to the antennas, but that does not matter because the subsequent adaptive canceler determines them adaptively and cancels the interference completely.

6. (a) Let $\hat{x}$ be the optimum linear estimate of a scalar x based on the random vector $\mathbf{y}$. Show that $\hat{x}$ remains invariant under a linear invertible transformation of the observation vector, $\mathbf{y} = B\mathbf{z}$

(b) Show that $E[e\hat{x}] = 0$ and $E[e^2] = E[ex]$, where $e = x - \hat{x}$.

(c) If x is uncorrelated with $\mathbf{y}$, show that $\hat{x} = 0$.

7. Let x be a random variable with mean $E[x] = m$. We wish to estimate x in terms of a zero-mean vector of observations $\mathbf{y}$. Because the mean of x is not zero, we seek an estimate of the form

$$\hat{x} = \mathbf{h}^T\mathbf{y} + b$$

The b-term is called a *bias* term. Assume the correlations $R = E[\mathbf{y}\mathbf{y}^T]$ and $\mathbf{r} = E[x\mathbf{y}]$ are known. Show that the optimum choices for $\mathbf{h}$ and b that minimize the mean square estimation error $\mathcal{E} = E[e^2]$, where $e = x - \hat{x}$, are

$$\mathbf{h} = R^{-1}\mathbf{r} \qquad \text{and} \qquad b = m$$

Note: It is straightforward to reformulate such "biased" estimates adaptively. They are very common especially in neural network applications.

8. (a) Show that the optimum estimate of $\mathbf{y}$ based on itself is itself, that is, $\hat{\mathbf{y}} = \mathbf{y}$.

(b) Let $\mathbf{z} = Q\mathbf{y}$, where Q does not have to be invertible or square. Show that the optimum estimate of $\mathbf{z}$ based on $\mathbf{y}$ is given by $\hat{\mathbf{z}} = Q\mathbf{y}$, that is, $\hat{\mathbf{z}} = \mathbf{z}$.

(c) Suppose $\mathbf{y}$ is divided into two subvectors $\mathbf{y}_1$ and $\mathbf{y}_2$, that is, $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}$. Using the results of the previous part or working directly, show that the optimum estimate of $\mathbf{y}_1$ based on $\mathbf{y}$, that is, $\hat{\mathbf{y}}_1 = E[\mathbf{y}_1\mathbf{y}^T]E[\mathbf{y}\mathbf{y}^T]^{-1}\mathbf{y}$, is given simply by $\hat{\mathbf{y}}_1 = \mathbf{y}_1$.

9. (a) A random variable x is related to the random vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ by

$$x = \mathbf{c}_1^T\mathbf{y}_1 + \mathbf{c}_2^T\mathbf{y}_2 + v = [\mathbf{c}_1^T,\ \mathbf{c}_2^T]\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} + v = \mathbf{c}^T\mathbf{y} + v$$

where v is uncorrelated with $\mathbf{y}_1$ and $\mathbf{y}_2$. Show that the best estimate of x based on the combined observation vector $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}$ is given by $\hat{x} = \mathbf{c}^T\mathbf{y}$. Therefore, the $\mathbf{y}$-dependent part of x is completely canceled from the error output $e = x - \hat{x}$, that is, $e = v$, in this case. (*Hint*: Show that the solution of the normal equations is $\mathbf{h} = \mathbf{c}$.)

(b) Determine the optimum estimate $\hat{x} = \mathbf{h}_1^T\mathbf{y}_1$ of x based only on the first observation vector $\mathbf{y}_1$ and show that in this case the $\mathbf{y}_1$-dependent part of x is still canceled completely from the error output $e = x - \hat{x}$, whereas the $\mathbf{y}_2$-dependent part is canceled *as much as possible*, in the sense that e is given by

$$e = v + \mathbf{c}_2^T(\mathbf{y}_2 - \hat{\mathbf{y}}_{2/1})$$

where

$$\hat{\mathbf{y}}_{2/1} = E[\mathbf{y}_2\mathbf{y}_1^T]E[\mathbf{y}_1\mathbf{y}_1^T]^{-1}\mathbf{y}_1 = R_{21}R_{11}^{-1}\mathbf{y}_1$$

is the best estimate of $\mathbf{y}_2$ based on $\mathbf{y}_1$. (*Hint*: Express $\mathbf{h}_1$ in terms of $\mathbf{c}_1$, $\mathbf{c}_2$, R_{11}, R_{21}.)

(c) Show that the minimized mean square error of the above case is given by:

$$\mathcal{E} = E[e^2] = \sigma_v^2 + \mathbf{c}_2^T(R_{22} - R_{21}R_{11}^{-1}R_{21}^T)\mathbf{c}_2$$

where $R_{22} = E[\mathbf{y}_2\mathbf{y}_2^T]$. Why is the second term in $\mathcal{E}$ non-negative?

Note: The results of this problem will be used later to develop guidelines for picking the *filter order* in adaptive filtering applications.

10. Let $R = \begin{bmatrix} 2 & 4 & 4 \\ 4 & 9 & 10 \\ 4 & 10 & 14 \end{bmatrix}$ be the covariance matrix of $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$, assumed to have zero mean. Determine the innovations representation $\mathbf{y} = B\boldsymbol{\epsilon}$ by carrying out the Gram-Schmidt orthogonalization of the components of $\mathbf{y}$. Then, verify the factorization $R_{yy} = BR_{\epsilon\epsilon}B^T$ by explicit matrix multiplication.

Next consider the estimation of a random variable x in terms of $\mathbf{y}$. The cross correlation between x and $\mathbf{y}$ is known to be $\mathbf{r} = E[x\mathbf{y}] = \begin{bmatrix} 4 \\ 4 \\ 2 \end{bmatrix}$. Determine the optimum estimation weights $\mathbf{h}$ and $\mathbf{g}$ with respect to the correlated basis $\mathbf{y}$ and the innovations basis $\boldsymbol{\epsilon}$, that is,

$$\hat{x} = \mathbf{h}^T\mathbf{y} = \mathbf{g}^T\boldsymbol{\epsilon}$$

Hint: Use $\mathbf{g} = D^{-1}L\mathbf{r}$ and $\mathbf{h} = L^T\mathbf{g}$, where $D = R_{\epsilon\epsilon}$ and $L = B^{-1}$.

11. For the previous problem, compute the optimum estimates of x based on the three successively bigger subspaces $Y_1 = \{y_1\}$, $Y_2 = \{y_1, y_2\}$, $Y_3 = \{y_1, y_2, y_3\}$, in the forms

$$\hat{x}_1 = h_{11}y_1 = g_{11}\epsilon_1$$

$$\hat{x}_2 = h_{21}y_1 + h_{21}y_2 = g_{21}\epsilon_1 + g_{22}\epsilon_2$$

$$\hat{x}_3 = h_{31}y_1 + h_{31}y_2 + h_{33}y_3 = g_{31}\epsilon_1 + g_{32}\epsilon_2 + g_{33}\epsilon_3$$

Show that the g-weights are independent of the order, that is $g_{pi} = g_i$, where g_i was found in the previous problem. Show that the above estimates can be recursively constructed by

$$\hat{x}_1 = g_1\epsilon_1$$

$$\hat{x}_2 = \hat{x}_1 + g_2\epsilon_2$$

$$\hat{x}_3 = \hat{x}_2 + g_3\epsilon_3$$

Assuming $\sigma_x^2 = 30$, use the recursions $\mathcal{E}_i = \mathcal{E}_{i-1} - g_i^2E[\epsilon_i^2]$, where $\mathcal{E}_i = E[e_i^2] = E\big[(x - \hat{x}_i)^2\big]$, to determine the successive estimation errors $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$. Note the gradual improvement of the estimate as the number of observations is increased.

Finally, determine the predictions $\hat{y}_{2/1}$ and $\hat{y}_{3/2}$ of y_2 and y_3 based on the past subspaces Y_1 and Y_2, respectively, write them in the forms,

$$\hat{y}_{2/1} = -a_{21}y_1 = b_{21}\epsilon_1$$

$$\hat{y}_{3/1} = -a_{31}y_1 - a_{32}y_2 = b_{31}\epsilon_1 + b_{32}\epsilon_2$$

and show that the inverse innovations matrix $L = B^{-1}$ can be expressed as:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ b_{21} & 1 & 0 \\ b_{31} & b_{32} & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ a_{21} & 1 & 0 \\ a_{31} & a_{32} & 1 \end{bmatrix}$$

12. Consider the "deterministic" random signal $y_n = 2\cos(\omega_1 n + \phi)$, where $\omega_1 = \pi/3$ and ϕ is a random phase distributed uniformly over the interval $[0, 2\pi]$.

    (a) Show that y_n satisfies an ordinary 2nd order homogeneous difference equation.

    (b) Using the definition $R(k) = E[y_{n+k}y_n]$, show that $R(k) = 2\cos(\omega_1 k)$.

    (c) Let $\mathbf{y} = [y_0, y_1, y_2]^T$ be three consecutive samples. Using the results in (b), determine the 3×3 autocorrelation matrix $R = E[\mathbf{y}\mathbf{y}^T]$ and show that it has zero determinant.

    (d) Because of the singularity of R, we expect the Cholesky factorization to break down at dimension 3. To see this, carry out the Gram-Schmidt orthogonalization of $\mathbf{y}$ starting with y_0 and ending with y_2, and thereby determine the factorization $R = BR_{\epsilon\epsilon}B^T$. Is the result consistent with part (a)?

13. (a) Let $R(k)$ be the autocorrelation function of a stationary random signal y_n. Express the autocorrelation matrix of the random vector $\mathbf{y} = \begin{bmatrix} y_n \\ y_{n+k} \end{bmatrix}$ in terms of $R(k)$. Then, show the general inequality

$$|R(k)| \le R(0), \qquad \text{for all } k$$

    (b) Let u, v be two random variables. Show the Schwarz inequality:

$$|E[uv]|^2 \le E[u^2]E[v^2]$$

    Hint: $\mathbf{y} = [u, v]^T$.

Supplement – Probability and Statistics Problems

1. (a) Let x be a zero-mean *gaussian* random variable with variance σ_x^2. Show

$$E[x^4] = 3\sigma_x^4$$

    (b) Let $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$ be a block of *mutually uncorrelated* zero-mean gaussian random variables each with variance σ_x^2. Using the above result, show

$$E[x_i x_j x_k x_l] = \sigma_x^4(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$$

    Show also that their covariance matrix is

$$R_{xx} = E[\mathbf{x}\mathbf{x}^T] = \sigma_x^2 I$$

    where I is the $N \times N$ identity matrix.

    (c) Suppose the above N random variables $\mathbf{x}$ are mixed up by an arbitrary invertible linear transformation $\mathbf{y} = B\mathbf{x}$ resulting into the new set of gaussian random variables $\mathbf{y} = [y_1, y_2, \ldots, y_N]^T$. Let $R = E[\mathbf{y}\mathbf{y}^T]$ be their covariance matrix. Show that

$$R = \sigma^2 BB^T$$

    (d) Show the analogous result of part (b):

$$E[y_i y_j y_k y_l] = R_{ij}R_{kl} + R_{ik}R_{jl} + R_{il}R_{jk}$$

2. An estimate of the mean m of N independent identically distributed random variables $\{y_1, y_2, \ldots, y_N\}$ of variance σ^2 can be formed by the weighted sum

$$\hat{m} = h_1 y_1 + h_2 y_2 + \cdots + h_N y_N$$

    Determine expressions for the mean and variance of $\hat{m}$, that is, the quantities $E[\hat{m}]$ and $\mathrm{var}(\hat{m})$. What are the constraints on the weights h_i in order for $\hat{m}$ to be an unbiased estimate of m? What are the optimal choices for these weights, if in addition, it is required that the variance $\mathrm{var}(\hat{m})$ be minimum?

3. The *sample mean* of N independent gaussian random variables $\{y_1, y_2, \ldots, y_N\}$ of mean m and variance σ^2 is given by

$$\hat{m} = \frac{1}{N}(y_1 + y_2 + \cdots + y_N)$$

First, show that $\hat{m}$ is unbiased and its variance is $\mathrm{var}(\hat{m}) = \sigma^2/N$. Then, show that the probability density of $\hat{m}$ is

$$p(\hat{m}) = \frac{N^{1/2}}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{N}{2\sigma^2}(\hat{m} - m)^2\right]$$

Moreover, show that as $N \to \infty$, this density converges to the deterministic delta function density $p(\hat{m}) \to \delta(\hat{m} - m)$.

4. Consider N independent gaussian random variables $\{y_1, y_2, \ldots, y_N\}$ of mean m and variance σ^2. The *sample variance* is defined as

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{m})^2$$

where $\hat{m}$ is the sample mean as defined above. Show that the mean and variance of the sample variance are given by

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

Note: This is somewhat *lower* than the CR lower bound $2\sigma^4/N$. But, this is no contradiction because the CR bound applies to unbiased estimators and the above is slightly biased.

5. Continuing with the previous problem, we can form an *unbiased* estimator for the variance by the *standard deviation*:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{m})^2$$

Therefore, $s^2 = N\hat{\sigma}^2/(N-1)$. Show that its mean and variance are

$$E[s^2] = \sigma^2, \qquad \mathrm{var}(s^2) = \frac{2\sigma^4}{N-1}$$

This does satisfy the CR bound.

6. Next, we determine that the probability distribution of s^2 is a χ^2-distribution with $(N-1)$ degrees of freedom. In the definition of s^2, there are N squared terms $(y_i - \hat{m})^2$, yet we divided by $(N-1)$ not N. But, these terms are not mutually independent because of the presence of $\hat{m}$. Using these dependencies, one can express s^2 as a sum of $(N-1)$ *independent* square terms, as follows.

   (a) Consider the following linear transformation (know as Helmert's transformation) from the set $\{y_1, \ldots, y_N\}$ to a new set $\{z_1, \ldots, z_N\}$:

   $$z_i = c_i(y_1 + y_2 + \cdots + y_i - iy_{i+1}), \qquad i = 1, 2, \ldots, N-1$$
   $$z_N = c_N(y_1 + y_2 + \cdots + y_N)$$

   Determine the scale factors c_i in order for the z_i's to have unit variance.

   (b) Then, show that the z_i have zero mean and are mutually uncorrelated:

   $$E[z_i z_j] = \delta_{ij}, \qquad i, j = 1, 2, \ldots, N$$

   (c) Then, show that the linear transformation preserves the sum of the squares,

   $$\sum_{i=1}^{N} z_i^2 = \frac{1}{\sigma^2}\sum_{i=1}^{N} y_i^2$$

   therefore, it is an orthogonal transformation. Finally, show that the sum of the first $(N-1)$ squared terms is

   $$\chi^2 = \sum_{i=1}^{N-1} z_i^2 = \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - \hat{m})^2$$

   Thus, the sum of the N squared terms in the right-hand-side follows a normalized χ^2-distribution with $(N-1)$ degrees of freedom.

7. The following twenty random numbers come from an unknown probability distribution:

$$\{0.33, -0.52, -2.41, -1.93, 0.46, -0.44, -0.97, -0.38, 0.48, 1.29,$$
$$-1.82, -1.23, -0.21, 2.66, -1.22, -0.41, -0.95, 1.47, -0.83, -0.43\}$$

Test the hypothesis that the underlying distribution is gaussian with zero mean and unit variance. To do this perform the χ^2 test by dividing the range of the gaussian distribution into the following six bins:

$$(-\infty,-1.5),\ (-1.5,-0.5),\ (-0.5,0.0),\ (0.0,0.5),\ (0.5,1.5),\ (1.5,\infty)$$

If the i-th bin is the interval (x_{i-1}, x_i), then the theoretically expected number of observations that will fall into the i-th bin will be

$$\frac{N_i^{\text{th}}}{N} = F(x_i) - F(x_{i-1})$$

where N is the total number of observations and $F(x)$ is the cdf of the *assumed* gaussian distribution, that is,

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2}\, dz$$

Let N_i be the *actual* number of observations that fall into the i-th bin. Then, calculate the χ^2 statistic given by

$$\chi^2 = \sum_{i=1}^{B} \frac{(N_i - N_i^{\text{th}})^2}{N_i^{\text{th}}}$$

where B is the number of bins — here, $B = 6$. This quantity follows a χ^2-distribution with $B-1$ degrees of freedom. Thus, its mean will be equal to the number of degrees of freedom, namely, $B-1$.

If your calculated χ^2 is near the theoretical mean $B-1$, then you cannot reject the hypothesis that the pdf was gaussian. Alternatively, you can look up the 95 percent confidence interval of the χ^2 distribution with $B-1$ degrees of freedom, that is, the interval $0 \le \chi^2 \le \chi^2_{0.95}$ such that the probability of a χ^2 value falling in it is 0.95 or equivalently, the probability of a χ^2 value falling outside it is only 0.05. Then, if your calculated value of χ^2 falls within that interval you can — with 95 percent confidence — conclude that the gaussian assumption cannot be rejected.

Note: For $B-1=5$ degrees of freedom, we have $\chi^2_{0.95} = 11.07$.

8. Let $F(x)$ be the cdf of a pdf $f(x)$. Show that the random variable u defined by

$$u = F(x)$$

is distributed uniformly over the interval $[0,1)$. Therefore, random variables x following the pdf $f(x)$ can be generated from a uniform random number generator using the inverse function $x = F^{-1}(u)$. This is the *inversion method* for generating random numbers from uniform ones (see Appendix A).

9. The Rayleigh probability density finds application in fading communication channels:

$$p(r) = \frac{r}{\sigma^2} e^{-r^2/2\sigma^2}, \qquad \text{for } r \ge 0$$

Using the inversion method, show how to generate a Rayleigh-distributed random variable r from a uniform variable u.

10. The inversion method may also be applied to the problem of generating *discrete-valued* random variables. Let x be a random variable that can only take one of the discrete values $\{x_1, x_2, \ldots, x_M\}$ with probabilities $\{p_1, p_2, \ldots, p_M\}$, respectively. It is assumed, of course, that the p_i sum up to unity.

    You have available a uniform generator in the interval $[0,1)$. Explain how to generate the discrete random numbers x from a uniform u.

11. You want to simulate a binary experiment in which only two outcomes can occur, one with probability p and the other with probability $1-p$. For example, simulating successive throws of heads or tails, or the transmission of bits 0 or 1, or, an accept/reject decision, etc. This is the same as the previous problem, with $M = 2$.

    The procedure for picking one or the other outcome can be mechanized as follows:

    1. Generate a uniform u.
    2. If $0 \le u < p$, then pick the first outcome.
    3. If $p \le u < 1$, then pick the second outcome.

    Explain why this procedure generates the two outcomes with the correct probabilities p and $1-p$.

    Note: The optimization method of *simulated annealing* uses such two-valued random variables. It is an iterative method of minimizing a performance index $J(\lambda)$, where λ is a vector of parameters with respect to which J must be minimized. Consider two successive choices of the parameter vector, λ_{new} and λ_{old}, and compute the change in the performance index: $\Delta J = J(\lambda_{\text{new}}) - J(\lambda_{\text{old}})$.

    Most iterative minimization algorithms, such as steeliest descent or Newton's method, try to continuously keep decreasing J, that is, they demand that the change in λ always be such that $\Delta J \le 0$. This can easily drive the λ into a local minimum of J and then the algorithm gets stuck there.

    To alleviate this problem, the so-called *Metropolis algorithm* of simulated annealing allows on occasion J to increase, that is, $\Delta J > 0$, in order to

jump over such local minima and continue decreasing towards the absolute minimum. The algorithm is as follows: If $\Delta J \le 0$ then *accept* the change in the parameter vector $\lambda_{\text{old}} \to \lambda_{\text{new}}$. But if $\Delta J > 0$ then accept the change only with probability $p = e^{-\beta \Delta J}$ and *reject* the change with probability $1 - p$, where β is a suitable positive constant.

Using the results of this problem, it should be clear how one will make the decision of whether to accept or reject the change.

12. Consider the Box-Muller transformation

$$x = (-2\ln u)^{1/2}\cos(2\pi v), \qquad y = (-2\ln u)^{1/2}\sin(2\pi v)$$

Show that if $\{u, v\}$ are independent uniform random variables in the interval $[0, 1)$, then $\{x, y\}$ are two independent gaussian random variables with zero mean and unit variance.

13. Consider the generalized Box-Muller transformation

$$x = (-2\ln u)^{1/2}\cos(2\pi v), \qquad y = (-2\ln u)^{1/2}\sin(2\pi v - \phi)$$

where ϕ is a constant angle. Show that if $\{u, v\}$ are independent uniform random variables in the interval $[0, 1)$, then $\{x, y\}$ are two *jointly* gaussian random variables with zero mean, unit variance, and correlation coefficient $E[xy] = \cos\phi$.

14. Let X_1 and X_2 be two independent random variables with cdf's $F_1(x)$ and $F_2(x)$. Show that the random variable $X = \max(X_1, X_2)$ has cdf $F(x) = F_1(x)F_2(x)$. Show also that $X = \min(X_1, X_2)$ has cdf $F(x) = F_1(x) + F_2(x) - F_1(x)F_2(x)$.

15. The inversion method of generating random variables is convenient only when the cdf $F(x)$ is known in closed form or is easily computed.

    An alternative method that works well when the pdf $f(x)$ is known but the cdf $F(x)$ is complicated, like the gaussian case, is the *rejection* method. It requires two conditions that are not difficult to meet: First, there exists a so-called *majorizing* pdf $g(x)$ such that $f(x)$ is bounded from above by

$$f(x) \le c\,g(x), \qquad \text{for all } x$$

where c is a given constant. Second, it is much easier to generate random variables from the distribution $g(x)$ than from $f(x)$. The following algorithm generates an x distributed according to $f(x)$:

1. Generate an x from the distribution $g(x)$.
2. Generate a y which is uniformly distributed over $[0, c\,g(x)]$.
3. If $y \le f(x)$, then output x; else, go to step 1 and repeat.

To show that this procedure correctly generates x's that are distributed according to $f(x)$, we must show that the conditional density of an x generated as above and *given* that $y \le f(x)$, is equal to the desired density $f(x)$, that is,

$$p\big(X = x \mid Y \le f(X)\big) = f(x)$$

(a) Show first that necessarily $c \ge 1$ and that

$$p\big(Y \le f(X) \mid X = x\big) = \frac{f(x)}{c\,g(x)}$$

which follows from the fact that y is uniform.

(b) Then, integrate the above over all x's generated from $g(x)$ to get

$$p\big(Y \le f(X)\big) = \frac{1}{c}$$

(c) Finally, use Bayes' rule to determine the quantity

$$p\big(X = x \mid Y \le f(X)\big) = \frac{p\big(Y \le f(X) \mid X = x\big)\,p(X = x)}{p\big(Y \le f(X)\big)}$$

16. Let $\mathbf{y}$ be an M-dimensional gaussian random vector with zero mean and covariance matrix R. Show that the information content or *entropy* of $\mathbf{y}$ is given by

$$S = -\int p(\mathbf{y})\ln p(\mathbf{y})\,d^M\mathbf{y} = \frac{1}{2}\ln(\det R)$$

up to an unimportant additive constant.

17. Let $\mathbf{y} = B\boldsymbol{\epsilon}$ be the innovations representation of an M-dimensional gaussian zero-mean vector. Show that its entropy can be written, up to an additive constant, as follows

$$S = -\int p(\mathbf{y})\ln p(\mathbf{y})\,d^M\mathbf{y} = \frac{1}{2}\sum_{i=1}^{M}\ln E_i$$

where $E_i = E[\epsilon_i^2]$ are the variances of the innovations.

18. (a) For any two positive real numbers a and b, show the inequality

$$-a\ln\left[\frac{a}{b}\right] \le b-a$$

(b) Let $\mathbf{y}$ be an M-dimensional random vector. For any two probability densities $p(\mathbf{y})$ and $q(\mathbf{y})$, prove the following *information inequality*,

$$-\int p(\mathbf{y})\ln\left[\frac{p(\mathbf{y})}{q(\mathbf{y})}\right] d^M\mathbf{y} \le 0$$

with equality attained when $p(\mathbf{y}) = q(\mathbf{y})$.

19. Consider the subset of all M-dimensional probability densities $p(\mathbf{y})$ that have a *given* mean $\mathbf{m}$ and covariance Σ. Show that the density from this subset that has *maximum entropy*,

$$S = -\int p(\mathbf{y})\ln p(\mathbf{y})\, d^M\mathbf{y} = \max$$

is the gaussian. *Hint*: Use Lagrange multipliers to enforce the given constraints. Alternatively, use the information inequality of the previous problem.

20. Let $R\mathbf{e}_i = \lambda_i\mathbf{e}_i$, $i = 1,2,\ldots,M$ be the M eigenvalues and *orthonormal* eigenvectors of the covariance matrix of an M-dimensional random vector $\mathbf{y}$. Define the M transformed random variables:

$$z_i = \mathbf{e}_i^T\mathbf{y}, \qquad i = 1,2,\ldots,M$$

(a) Show that they are mutually uncorrelated with variances λ_i, that is,

$$E[z_iz_j] = \lambda_i\delta_{ij}$$

(b) Show that $\mathbf{y}$ can be expanded in terms of the z_i as follows:

$$\mathbf{y} = \sum_{i=1}^{M} z_i\mathbf{e}_i$$

Thus, the randomness of $\mathbf{y}$ arises only from the randomness of the z_i's which are uncorrelated. If the eigenvalues are arranged in decreasing order and the first L largest eigenvalues are dominant, then the sum may be approximated by

$$\mathbf{y} \simeq \sum_{i=1}^{L} z_i\mathbf{e}_i$$

Thus, the M-vector $\mathbf{y}$ is represented by only $L < M$ parameters, namely, $z_1, z_2, \ldots, z_L$. This approximation forms the basis of *data compression* using the Karhunen-Loeve transform.

(c) Show the equality of quadratic forms

$$\mathbf{y}^TR^{-1}\mathbf{y} = \sum_{i=1}^{M}\frac{z_i^2}{\lambda_i}$$

(d) Determine the pdf $p_z(\mathbf{z})$ of the vector $\mathbf{z} = [z_1, z_2, \ldots, z_M]^T$ in terms of the pdf $p_y(\mathbf{y})$ (do not assume gaussian distributions). Show that the information content of $\mathbf{y}$ is the same as that of $\mathbf{z}$, in the sense that they have equal entropies.

(e) If we denote by B the modal matrix of R, that is, the matrix whose columns are the eigenvectors $\mathbf{e}_i$, then show that $\mathbf{y}$ is related to the z-basis as

$$\mathbf{y} = B\mathbf{z}, \qquad \text{where} \quad B = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_M]$$

Show also that B satisfies $BB^T = B^TB = I$, and that

$$R = BDB^T, \qquad D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_M)$$

332:525 – Solutions

1. Differentiating $\mathcal{E}_n$ with respect to $\hat{m}$ and setting the gradient to zero gives:

$$\frac{\partial \mathcal{E}_n}{\partial \hat{m}} = -2 \sum_{k=0}^{n} (x_k - \hat{m}) = 0$$

which has solution:

$$\hat{m}_n = \frac{\sum_{k=0}^{n} x_k}{\sum_{k=0}^{n} 1} = \frac{\sum_{k=0}^{n} x_k}{n+1}$$

In part (b), the required recursions were shown in class. For part (c), we take expectations of both sides of the definition of $\hat{m}_n$ to get:

$$E[\hat{m}_n] = \frac{1}{n+1} \sum_{k=0}^{n} E[x_k] = \frac{1}{n+1} \sum_{k=0}^{n} m = m$$

Next, we have:

$$\hat{m}_n - E[\hat{m}_n] = \hat{m}_n - m = \frac{1}{n+1} \sum_{k=0}^{n} x_k - \frac{1}{n+1} \sum_{k=0}^{n} m = \frac{1}{n+1} \sum_{k=0}^{n} (x_k - m)$$

The variance of $\hat{m}_n$ will be then

$$E[(\hat{m}_n - m)^2] = \frac{1}{(n+1)^2} \sum_{k=0}^{n} \sum_{j=0}^{n} E[(x_k - m)(x_j - m)]$$

And, using the iid assumption, we have $E[(x_k - m)(x_j - m)] = \sigma_x^2 \delta_{kj}$ which gives for the variance of $\hat{m}_n$:

$$E[(\hat{m}_n - m)^2] = \frac{1}{(n+1)^2} \sum_{k=0}^{n} \sum_{j=0}^{n} \sigma_x^2 \delta_{kj} = \frac{1}{(n+1)^2} \sigma_x^2 \sum_{k=0}^{n} 1 = \frac{\sigma_x^2}{n+1}$$

2 . The gradient of the performance index is now:

$$\frac{\partial \mathcal{E}_n}{\partial \hat{m}} = -2 \sum_{k=0}^{n} \lambda^{n-k} (x_k - \hat{m}) = 0$$

with optimum solution:

$$\hat{m}_n = \frac{\sum_{k=0}^{n} \lambda^{n-k} x_k}{\sum_{k=0}^{n} \lambda^{n-k}} = \frac{x_n + \lambda x_{n-1} + \lambda^2 x_{n-2} + \cdots + \lambda^n x_0}{1 + \lambda + \lambda^2 + \cdots + \lambda^n}$$

Using the finite geometric series, we may write the denominator as

$$\sum_{k=0}^{n} \lambda^{n-k} = 1 + \lambda + \lambda^2 + \cdots + \lambda^n = \frac{1 - \lambda^{n+1}}{1 - \lambda}$$

which gives for the estimator $\hat{m}_n$:

$$\hat{m}_n = \frac{(1-\lambda) \sum_{k=0}^{n} \lambda^{n-k} x_k}{1 - \lambda^{n+1}}$$

Replacing n by $n-1$ and multiplying by a factor of λ, gives:

$$\lambda \hat{m}_{n-1} = \frac{(1-\lambda)\lambda \sum_{k=0}^{n-1} \lambda^{n-1-k} x_k}{1 - \lambda^n} = \frac{(1-\lambda) \sum_{k=0}^{n-1} \lambda^{n-k} x_k}{1 - \lambda^n}$$

Thus, we can express the sum up to $k = n-1$ in terms of $\hat{m}_{n-1}$:

$$(1-\lambda) \sum_{k=0}^{n-1} \lambda^{n-k} x_k = \lambda (1 - \lambda^n) \hat{m}_{n-1}$$

Therefore, we obtain the recursion for $\hat{m}_n$

$$\hat{m}_n = \frac{(1-\lambda)\left(\sum_{k=0}^{n-1} \lambda^{n-k} x_k + x_n\right)}{1 - \lambda^{n+1}} = \frac{\lambda - \lambda^{n+1}}{1 - \lambda^{n+1}} \hat{m}_{n-1} + \frac{1-\lambda}{1-\lambda^{n+1}} x_n$$

which can be written in the predictor/corrector form:

$$\hat{m}_n = \hat{m}_{n-1} + \left(\frac{1-\lambda}{1-\lambda^{n+1}}\right)(x_n - \hat{m}_{n-1})$$

In the limit $\lambda \to 1$, the Kalman gain coefficient tends to the expected limit:

$$\lim_{\lambda \to 1} \left(\frac{1-\lambda}{1-\lambda^{n+1}}\right) = \frac{1}{n+1}$$

On the other hand, if λ is strictly less than one, then the term λ^{n+1} can be ignored after a few iterations, and therefore, the recursion becomes essentially the first-order smoother:

$$\hat{m}_n = \hat{m}_{n-1} + (1-\lambda)(x_n - \hat{m}_{n-1}) = \lambda \hat{m}_{n-1} + (1-\lambda) x_n$$

3. The difference equation

$$\hat{m}_n = \lambda \hat{m}_{n-1} + (1-\lambda) x_n$$

can be solved assuming zero initial conditions, by convolving the x_n sequence with the filter sequence $(1-\lambda)\lambda^n$. This gives:

$$\hat{m}_n = (1-\lambda) \sum_{k=0}^{n} \lambda^{n-k} x_k$$

Taking expectations of both sides and using the finite geometric series, we obtain:

$$E[\hat{m}_n] = (1-\lambda) \sum_{k=0}^{n} \lambda^{n-k} m = \frac{1-\lambda^{n+1}}{1-\lambda} m$$

which tends to m for large n. Thus, $\hat{m}_n$ is asymptotically unbiased. Subtracting the mean $E[\hat{m}_n]$ from $\hat{m}_n$ gives also:

$$\hat{m}_n - E[\hat{m}_n] = (1-\lambda) \sum_{k=0}^{n} \lambda^{n-k} (x_k - m)$$

Using the same sort of calculation as in Problem 1, we obtain for the variance of $\hat{m}_n$:

$$E\big[\,(\hat{m}_n - E[\hat{m}_n])^2\,\big] = (1-\lambda)^2 \sum_{k=0}^{n} \sum_{j=0}^{n} \lambda^{n-k} \lambda^{n-j} E\big[\,(x_k - m)(x_j - m)\,\big]$$

$$= (1-\lambda)^2 \sum_{k=0}^{n} \sum_{j=0}^{n} \lambda^{n-k} \lambda^{n-j} \sigma_x^2 \delta_{kj} = \sigma_x^2 (1-\lambda)^2 \sum_{k=0}^{n} \lambda^{2(n-k)}$$

$$= \sigma_x^2 (1-\lambda)^2 \frac{1-\lambda^{2(n+1)}}{1-\lambda^2} = \frac{1-\lambda}{1+\lambda} \sigma_x^2 (1-\lambda^{2(n+1)})$$

which in the limit of large n converges to the required result.

4. The theoretical gradient is:

$$\frac{\partial \mathcal{E}}{\partial \hat{m}} = -2E\big[\,(x_n - \hat{m})\,\big] = -2(m - \hat{m})$$

Thus, it vanishes when $\hat{m} = m$. The instantaneous gradient is obtained by dropping the expectation value, that is,

$$\frac{\widehat{\partial \mathcal{E}}}{\partial \hat{m}} = -2(x_n - \hat{m})$$

Putting this into the LMS updating equation gives:

$$\hat{m}_{n+1} = \hat{m}_n + \Delta \hat{m}_n = \hat{m}_n - \mu \frac{\widehat{\partial \mathcal{E}}}{\partial \hat{m}_n} = \hat{m}_n + 2\mu (x_n - \hat{m}_n)$$

Setting $2\mu = 1 - \lambda$, we rewrite the difference equation as

$$\hat{m}_{n+1} = \hat{m}_n + 2\mu (x_n - \hat{m}_n) = (1-2\mu)\hat{m}_n + 2\mu x_n = \lambda \hat{m}_n + (1-\lambda) x_n$$

5. *Problem 1.9:* Using $x = s + n_1 = s + Fn_2$ and $y = n_2$, we find $R_{yy} = E[y^2] = E[n_2^2]$ and $R_{xy} = E[xy] = E[\,(x+Fn_2)\,n_2\,] = FE[n_2^2]$. The optimal canceler will be $H = R_{xy} R_{yy}^{-1} = FE[n_2^2] E[n_2^2]^{-1} = F$. The corresponding optimum estimate will be $\hat{x} = Hy = Fn_2$, and the estimation error $e = x - \hat{x} = (s + Fn_2) - Fn_2 = s$.

Problem 1.10: First determine H. Noting that

$$y = n_2 + \epsilon s = \frac{1}{F} n_1 + \epsilon s$$

and using the definition of the gain G, we find R_{yy} and R_{xy}:

$$R_{yy} = E[yy] = \frac{1}{F^2} E[n_1^2] + \epsilon^2 E[s^2] = \left(\frac{1}{F^2} + \epsilon^2 G \right) E[n_1^2]$$

$$R_{xy} = E[xy] = \frac{1}{F} E[n_1^2] + \epsilon E[s^2] = \left(\frac{1}{F} + \epsilon G \right) E[n_1^2]$$

Therefore,

$$H = R_{xy} R_{yy}^{-1} = \frac{\frac{1}{F} + \epsilon G}{\frac{1}{F^2} + \epsilon^2 G} = \frac{F(1 + \epsilon F G)}{1 + \epsilon^2 F^2 G}$$

The error output will be

$$e = x - \hat{x} = x - Hy = s + n_1 - H\left(\frac{1}{F} n_1 + \epsilon s\right) = (1 - \epsilon H) s + \left(1 - \frac{H}{F}\right) n_1$$

Thus, the coefficients a and b will be

$$a = 1 - \epsilon H = 1 - \frac{\epsilon F(1+\epsilon FG)}{1+\epsilon^2F^2G} = \frac{1-\epsilon F}{1+\epsilon^2F^2G}$$

$$b = 1 - \frac{H}{F} = 1 - \frac{1+\epsilon FG}{1+\epsilon^2F^2G} = -\epsilon FG\frac{1-\epsilon F}{1+\epsilon^2F^2G} = -\epsilon FGa$$

If the coefficient ϵ is known in advance, then the pre-processed signals will be

$$x_1 = x = s + n_1 = s + Fn_2$$

$$y_1 = y - \epsilon x = n_2 + \epsilon s - \epsilon s - \epsilon Fn_2 = (1-\epsilon F)n_2$$

Thus, y_1 is correlated only with the noise part of x_1. We find

$$\begin{aligned} E[x_1y_1] &= F(1-\epsilon F)E[n_2^2] \\ E[y_1y_1] &= (1-\epsilon F)^2E[n_2^2] \end{aligned} \quad \Rightarrow \quad H_1 = E[x_1y_1]E[y_1y_1]^{-1} = \frac{F}{1-\epsilon F}$$

and, therefore,

$$\hat{x}_1 = H_1y_1 = \frac{F}{1-\epsilon F}(1-\epsilon F)n_2 = Fn_2$$

$$e_1 = x_1 - \hat{x}_1 = s + Fn_2 - Fn_2 = s$$

6. For part (a), we have

$$E[\mathbf{y}\mathbf{y}^T] = BE[\mathbf{z}\mathbf{z}^T]B^T \qquad \Rightarrow \qquad E[\mathbf{y}\mathbf{y}^T]^{-1} = B^{-T}E[\mathbf{z}\mathbf{z}^T]^{-1}B^{-1}$$

And, similarly,

$$E[x\mathbf{y}] = BE[x\mathbf{z}]$$

The optimal Wiener weights with respect to the two bases are:

$$\mathbf{h} = E[\mathbf{y}\mathbf{y}^T]^{-1}E[x\mathbf{y}], \qquad \mathbf{g} = E[\mathbf{z}\mathbf{z}^T]^{-1}E[x\mathbf{z}]$$

Therefore, they are related by

$$\mathbf{h} = E[\mathbf{y}\mathbf{y}^T]^{-1}E[x\mathbf{y}] = B^{-T}E[\mathbf{z}\mathbf{z}^T]^{-1}B^{-1}BE[x\mathbf{z}] = B^{-T}\mathbf{g}$$

or, $\mathbf{h}^T = \mathbf{g}^TB^{-1}$. It follows that the optimal estimate $\hat{x}$ will be invariant under a change of basis:

$$\hat{x} = \mathbf{h}^T\mathbf{y} = \mathbf{g}^TB^{-1}B\mathbf{z} = \mathbf{g}^T\mathbf{z}$$

Parts (b) and (c) were done in class.

7. The estimation error is $e = x - \hat{x} = x - \mathbf{h}^T\mathbf{y} - b$. The minimization conditions for the performance index $\mathcal{E} = E[e^2]$ are

$$\frac{\partial\mathcal{E}}{\partial\mathbf{h}} = 2E[e\frac{\partial e}{\partial\mathbf{h}}] = -2E[e\mathbf{y}] = 0$$

$$\frac{\partial\mathcal{E}}{\partial b} = 2E[e\frac{\partial e}{\partial b}] = -2E[e] = 0$$

which are equivalent to

$$E[e\mathbf{y}] = E[(x - \mathbf{y}^T\mathbf{h})\mathbf{y}] = E[x\mathbf{y}] - E[\mathbf{y}\mathbf{y}^T]\mathbf{h} = \mathbf{r} - R\mathbf{h} = 0$$

$$E[e] = E[x - \mathbf{h}^T\mathbf{y} - b] = E[x] - \mathbf{h}^TE[\mathbf{y}] - b = m - b = 0$$

8. Part (a) follows from part (b) with the choice $Q = I$. For part (b), we have

$$R_{zy} = E[\mathbf{z}\mathbf{y}^T] = QE[\mathbf{y}\mathbf{y}^T] = QR_{yy} \qquad \Rightarrow \qquad H = R_{zy}R_{yy}^{-1} = Q$$

It follows that $\hat{\mathbf{z}} = H\mathbf{y} = Q\mathbf{y} = \mathbf{z}$. Part (c) can be shown as follows: Note that the subvector $\mathbf{y}_1$ can be obtained from the full vector $\mathbf{y}$ by the projection matrix

$$\mathbf{y}_1 = \begin{bmatrix} I_1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = Q\mathbf{y}$$

where I_1 is the identity matrix with the same dimension as $\mathbf{y}_1$. Using part (b) with $\mathbf{z} = \mathbf{y}_1$, we find $\hat{\mathbf{y}}_1 = \mathbf{y}_1$. This result can also be shown directly, as follows. Using the notation $R_{ij} = E[\mathbf{y}_i\mathbf{y}_j^T]$, for $i,j = 1,2$, we have

$$E[\mathbf{y}_1\mathbf{y}^T] = E[\mathbf{y}_1[\mathbf{y}_1^T, \mathbf{y}_2^T]] = [E[\mathbf{y}_1\mathbf{y}_1^T], E[\mathbf{y}_1\mathbf{y}_2^T]] = [R_{11}, R_{12}]$$

$$E[\mathbf{y}\mathbf{y}^T] = E[\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}[\mathbf{y}_1^T, \mathbf{y}_2^T]] = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

But noting that

$$[R_{11}, R_{12}] = [I_1, 0]\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

we obtain

$$H = E[\mathbf{y}_1\mathbf{y}^T]E[\mathbf{y}\mathbf{y}^T]^{-1} = [R_{11}, R_{12}]\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}^{-1} = [I_1, 0]$$

Thus, $\hat{\mathbf{y}}_1 = H\mathbf{y} = [I_1, 0]\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \mathbf{y}_1$.

9. Using part (a) of the previous problem, we have $\hat{\mathbf{y}} = \mathbf{y}$. Therefore,

$$\hat{x} = \widehat{\mathbf{c}^T\mathbf{y}} = \mathbf{c}^T\hat{\mathbf{y}} = \mathbf{c}^T\mathbf{y}$$

and $e = x - \hat{x} = \mathbf{c}^T\mathbf{y} + v - \mathbf{c}^T\mathbf{y} = v$. If the estimation is based only on the subvector $\mathbf{y}_1$, then we have $\hat{\mathbf{y}}_1 = \mathbf{y}_1$, and therefore,

$$\hat{x} = \mathbf{c}_1^T\hat{\mathbf{y}}_1 + \mathbf{c}_2^T\hat{\mathbf{y}}_2 = \mathbf{c}_1^T\mathbf{y}_1 + \mathbf{c}_2^T\hat{\mathbf{y}}_{2/1}$$

and for the error output

$$e = x - \hat{x} = \mathbf{c}_1^T\mathbf{y}_1 + \mathbf{c}_2^T\mathbf{y}_2 + v - \mathbf{c}_1^T\mathbf{y}_1 - \mathbf{c}_2^T\hat{\mathbf{y}}_{2/1} =$$

Setting $\mathbf{e}_2 = \mathbf{y}_2 - \hat{\mathbf{y}}_2$, we have $e = v + \mathbf{c}_2^T\mathbf{e}_2$. And,

$$\mathcal{E} = E[e^2] = \sigma_v^2 + \mathbf{c}_2^T E[\mathbf{e}_2\mathbf{e}_2^T]\mathbf{c}_2$$

But, $E[\mathbf{e}_2\mathbf{e}_2^T] = R_{22} - R_{21}R_{11}^{-1}R_{12}$, which also shows the non-negativity property.

10. Going through the Gram-Schmidt orthogonalization procedure, we find the matrices B and $D = R_{\epsilon\epsilon}$:

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & 2 & 1 \end{bmatrix}, \qquad D = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

We also need the inverses

$$B^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 2 & -2 & 1 \end{bmatrix}, \qquad R^{-1} = B^{-T}D^{-1}B^{-1} = \begin{bmatrix} 6.5 & -4 & 1 \\ -4 & 3 & -1 \\ 1 & -1 & 0.5 \end{bmatrix}$$

Thus, the innovations basis is

$$\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_2 \end{bmatrix} = \boldsymbol{\epsilon} = B^{-1}\mathbf{y} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 2 & -2 & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 - 2y_1 \\ y_3 - 2y_2 + 2y_1 \end{bmatrix}$$

and conversely,

$$\begin{bmatrix} y_1 \\ y_2 \\ y_2 \end{bmatrix} = \mathbf{y} = B\boldsymbol{\epsilon} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & 2 & 1 \end{bmatrix}\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 + 2\epsilon_1 \\ \epsilon_3 + 2\epsilon_2 + 2\epsilon_1 \end{bmatrix}$$

For the estimation part, we calculate the $\mathbf{h}$ and $\mathbf{g}$ weights using the formulas

$$\mathbf{g} = D^{-1}B^{-1}\mathbf{r} = \begin{bmatrix} 2 \\ -4 \\ 1 \end{bmatrix}, \qquad \mathbf{h} = B^{-T}\mathbf{g} = R^{-1}\mathbf{r} = \begin{bmatrix} 12 \\ -6 \\ 1 \end{bmatrix}$$

11. The three $\mathbf{g}$ weights are the optimal weights for the lower order estimation problems, that is,

$$\hat{x}_1 = g_1\epsilon_1 = 2\epsilon_1$$
$$\hat{x}_2 = g_1\epsilon_1 + g_2\epsilon_2 = 2\epsilon_1 - 4\epsilon_2$$
$$\hat{x}_3 = g_1\epsilon_1 + g_2\epsilon_2 + g_3\epsilon_3 = 2\epsilon_1 - 4\epsilon_2 + \epsilon_3$$

Replacing the ϵ_i in terms of the y_i, we get

$$\hat{x}_1 = 2y_1$$
$$\hat{x}_2 = 2y_1 - 4(y_2 - 2y_1) = 10y_1 - 4y_2$$
$$\hat{x}_3 = 10y_1 - 4y_2 + (y_3 - 2y_2 + 2y_1) = 12y_1 - 6y_2 + y_3$$

For the mean square errors, using the variances of the ϵ_i, $\{E_1, E_2, R_3\} = \{2, 1, 2\}$, and starting with $\mathcal{E}_0 = \sigma_x^2 = 30$, we get

$$\mathcal{E}_1 = \mathcal{E}_0 - g_1^2E_1 = 30 - 2^2 \cdot 2 = 22$$
$$\mathcal{E}_2 = \mathcal{E}_1 - g_2^2E_2 = 22 - (-4)^2 \cdot 1 = 6$$
$$\mathcal{E}_3 = \mathcal{E}_2 - g_3^2E_3 = 6 - 1^2 \cdot 2 = 4$$

For the prediction, we want to show that the a_{ij} coefficients are the matrix elements of B^{-1}. This can be seen in general by writing the expression of the ϵ_i in terms of the y_i, as follows:

$$\begin{aligned} \epsilon_1 &= y_1 \\ \epsilon_2 &= y_2 - \hat{y}_{2/1} = y_2 + a_{21}y_1 \\ \epsilon_3 &= y_3 - \hat{y}_{3/2} = y_3 + a_{32}y_2 + a_{31}y_1 \end{aligned} \quad \Rightarrow \quad \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ a_{21} & 1 & 0 \\ a_{31} & a_{32} & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

which is equivalent to $\boldsymbol{\epsilon} = B^{-1}\mathbf{y}$.

12. The difference equation is

$$y_n - 2\cos\omega_1 y_{n-1} + y_{n-2} = 0$$

Indeed, using $y_n = A\cos(\omega_1 n + \phi)$, we have

$$\begin{aligned} 2\cos\omega_1 y_{n-1} &= 2A\cos\omega_1\cos\big(\omega_1(n-1)+\phi\big) \\ &= A\cos(\omega_1 n + \phi) + A\cos(\omega_1 n - 2\omega_1 + \phi) = y_n + y_{n-2} \end{aligned}$$

where we used the trig identity

$$2\cos a\cos b = \cos(a+b) + \cos(a-b)$$

Using this trig identity again, we obtain for the autocorrelation function:

$$\begin{aligned} R(k) &= E[y_{n+k}y_n] = A^2 E[\cos(\omega_1 n + \omega_1 k + \phi)\cos(\omega_1 n + \phi)] \\ &= \frac{1}{2}A^2 E[\cos(2\omega_1 n + \omega_1 k + \phi) + \cos(\omega_1 k)] \\ &= \frac{1}{2}A^2\cos(\omega_1 k) \end{aligned}$$

where the first expectation value is zero, as follows from the property

$$E[\cos(\phi+\theta)] = \int_0^{2\pi}\cos(\phi+\theta)\,\frac{d\phi}{2\pi} = 0$$

for ϕ uniform over $[0, 2\pi)$ and θ deterministic. The 3×3 autocorrelation matrix will be $R_{ij} = E[y_i y_j] = R(i-j)$. Noting that $R(i-j) = R(j-i)$, we find

$$R = \begin{bmatrix} R(0) & R(1) & R(2) \\ R(1) & R(0) & R(1) \\ R(2) & R(1) & R(0) \end{bmatrix} = \frac{1}{2}A^2\begin{bmatrix} 1 & \cos\omega_1 & \cos 2\omega_1 \\ \cos\omega_1 & 1 & \cos\omega_1 \\ \cos 2\omega_1 & \cos\omega_1 & 1 \end{bmatrix}$$

Its determinant is

$$\det R = \frac{1}{8}A^6\big[1 + 2\cos^2\omega_1\cos 2\omega_1 - \cos^2 2\omega_1 - 2\cos^2\omega_1\big]$$

Using the trig identity $\cos 2\omega_1 = 2\cos^2\omega_1 - 1$, we can verify that the expression in the brackets vanishes. The same result also follows from the observation that the rank of R is two not three because each column

can be expressed as a linear combination of the other two, for example, the first column is expressible as

$$\begin{bmatrix} 1 \\ \cos\omega_1 \\ \cos 2\omega_1 \end{bmatrix} = 2\cos\omega_1\begin{bmatrix} \cos\omega_1 \\ 1 \\ \cos\omega_1 \end{bmatrix} - \begin{bmatrix} \cos 2\omega_1 \\ \cos\omega_1 \\ 1 \end{bmatrix}$$

The Gram-Schmidt construction proceeds as follows:

$$\begin{aligned} \epsilon_0 &= y_0 \\ \epsilon_1 &= y_1 - b_{10}\epsilon_0 \\ \epsilon_2 &= y_2 - b_{20}\epsilon_0 - b_{21}\epsilon_1 \end{aligned}$$

where

$$b_{10} = \frac{E[y_1\epsilon_0]}{E[y_0 y_0]} = \frac{R(1)}{R(0)} = \cos\omega_1$$

The quantity $E_1 = E[\epsilon_1^2]$ is calculated by squaring the expression $y_1 = \epsilon_1 + b_{10}\epsilon_0$, taking expectations of both sides, and using $E_0 = E[\epsilon_0^2] = R(0)$:

$$R(0) = E[y_1^2] = E_1 + b_{10}^2 E_0 \qquad \Rightarrow \qquad E_1 = R_0 - b_{10}^2 E_0 = R(0)\,(1 - b_{10}^2)$$

or,

$$E_1 = R(0)\,(1 - \cos^2\omega_1) = R(0)\sin^2\omega_1$$

Similarly, we find

$$\begin{aligned} b_{20} &= \frac{E[y_2\epsilon_0]}{E_0} = \frac{R(2)}{R(0)} = \cos 2\omega_1 \\ b_{21} &= \frac{E[y_2\epsilon_1]}{E_1} = \frac{E[y_2(y_1 - b_{10}y_0)]}{E_1} = \frac{R(1) - b_{10}R(2)}{E_1} \\ &= \frac{\cos\omega_1 - \cos\omega_1\cos 2\omega_1}{\sin^2\omega_1} = \frac{\cos\omega_1(1-\cos 2\omega_1)}{\sin^2\omega_1} = \frac{\cos\omega_1(2\sin^2\omega_1)}{\sin^2\omega_1} \\ &= 2\cos\omega_1 \end{aligned}$$

Thus, the B matrix will be

$$B = \begin{bmatrix} 1 & 0 & 0 \\ b_{10} & 1 & 0 \\ b_{20} & b_{21} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \cos\omega_1 & 1 & 0 \\ \cos 2\omega_1 & 2\cos\omega_1 & 1 \end{bmatrix}$$

The prediction error E_2 is expected to be zero because the y_2 can be predicted exactly from $\{y_0, y_1\}$, as follows from the difference equation applied with $n = 2$:

$$y_2 - 2\cos\omega_1 y_1 + y_0 = 0$$

Indeed, squaring the equation $y_2 = \epsilon_2 + b_{20}\epsilon_0 + b_{21}\epsilon_1$ and taking expectations of both sides, we get

$$R(0) = E[y_2^2] = E_2 + b_{20}^2 E_0 + b_{21}^2 E_1$$

and solving for E_2

$$\begin{aligned} E_2 &= R(0) - b_{20}^2 E_0 - b_{21}^2 E_1 = R(0) - \cos^2 2\omega_1 E_0 - 4\cos^2\omega_1 E_1 \\ &= R(0) - \cos^2 2\omega_1 R(0) - 4\cos^2\omega_1 \sin^2\omega_1 R(0) \\ &= (1 - \cos^2 2\omega_1) R(0) - 4\cos^2\omega_1 \sin^2\omega_1 R(0) \\ &= \sin^2 2\omega_1 R(0) - \sin^2 2\omega_1 R(0) = 0 \end{aligned}$$

Thus, the D matrix will be

$$D = \begin{bmatrix} E_0 & 0 & 0 \\ 0 & E_1 & 0 \\ 0 & 0 & E_2 \end{bmatrix} = R(0) \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sin^2\omega_1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Finally, one should be able to verify the Cholesky factorization $R = BDB^T$, which in this case reads as follows (we removed an overall factor of $R(0)$):

$$\begin{bmatrix} 1 & \cos\omega_1 & \cos 2\omega_1 \\ \cos\omega_1 & 1 & \cos\omega_1 \\ \cos 2\omega_1 & \cos\omega_1 & 1 \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ \cos\omega_1 & 1 & 0 \\ \cos 2\omega_1 & 2\cos\omega_1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sin^2\omega_1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & \cos\omega_1 & \cos 2\omega_1 \\ 0 & 1 & 2\cos\omega_1 \\ 0 & 0 & 1 \end{bmatrix}$$

13. Part (a) follows from part (b) and stationarity. Indeed,

$$|E[y_{n+k}y_n]|^2 \le E[y_{n+k}^2]E[y_n^2] \qquad \Rightarrow \qquad |R(k)|^2 \le R(0)R(0)$$

or, $|R(k)| \le R(0)$. Part (b) can be derived as follows: The autocorrelation matrix of $\mathbf{y} = \begin{bmatrix} u \\ v \end{bmatrix}$ is

$$R = E[\mathbf{y}\mathbf{y}^T] = E\left[\begin{bmatrix} u \\ v \end{bmatrix}[u, v]\right] = \begin{bmatrix} E[u^2] & E[uv] \\ E[vu] & E[v^2] \end{bmatrix}$$

Because this matrix is positive semi-definite, its determinant will be non-negative, that is,

$$\det R = E[u^2]E[v^2] - E[uv]^2 \ge 0$$