Kalman and Levinson Filtering

A&AE 567 Lecture Notes

A. E. Frazho

Contents

Chapter 1

The Projection Theorem

The projection theorem plays a fundamental role in stochastic processes and filtering theory. In this chapter we will introduce some elementary facts concerning Hilbert space and present the projection theorem.

1.1 Notation

In this section we will introduce some notation used throughout these notes. The reader can choose to go directly to the next section and refer back to this section when the appropriate notation is needed. The set of all complex numbers is denoted by $\mathbb{C}$. Throughout all linear spaces are complex vector spaces unless stated otherwise. The inner product on a linear space is denoted by $(\cdot,\cdot)$. To be precise, the inner product on a linear space $\mathcal{K}$ is a complex valued function mapping $\mathcal{K}\times\mathcal{K}$ into $\mathbb{C}$ with the following properties

(i) $(f,h)=\overline{(h,f)}$;

(ii) $(\alpha f+\beta g,h)=\alpha(f,h)+\beta(g,h)$;

(iii) $(h,h)\geq 0$ (for all $h\in\mathcal{K}$);

(iv) $(h,h)=0$ if and only if $h=0$.

Here f,g and h are vectors in $\mathcal{K}$ while α and β are complex numbers. The inner product $(\cdot,\cdot)$ is linear in the first variable and conjugate linear in the second variable. An *inner product space* is simply a linear space with an inner product.

Now assume that $\mathcal{K}$ is an inner product space. The norm of a vector h in $\mathcal{K}$ is defined by $\|h\|=+\sqrt{(h,h)}$. If α is any scalar, then $\|\alpha h\|=|\alpha|\|h\|$. Moreover, $h=0$ if and only if the norm of h is zero. Notice that if f and h are two vectors in $\mathcal{K}$, then

$$\|f+h\|^2=\|f\|^2+2\Re(f,h)+\|h\|^2\,. \tag{1.1}$$

To verify this simply observe that

$$\begin{aligned}\|f+h\|^2 &= (f+h,f+h)=(f,f)+(f,h)+(h,f)+(h,h)\\ &= \|f\|^2+(f,h)+\overline{(f,h)}+\|h\|^2=\|f\|^2+2\Re(f,h)+\|h\|^2\,.\end{aligned}$$

Hence (1.1) holds.

The Cauchy-Schwartz inequality holds for any inner product space. To be precise, assume that $\mathcal{K}$ is an inner product space. The Cauchy-Schwartz inequality states that

$$|(f,h)| \leq \|f\| \, \|h\| \qquad (f,h \in \mathcal{K}) \,. \tag{1.2}$$

Moreover, we have equality $|(f,h)| = \|f\| \, \|h\|$ if and only if f and h are linearly dependent. In this setting, the triangle inequality is given by $\|f+h\| \leq \|f\| + \|h\|$ for all f and h in $\mathcal{K}$. The triangle inequality follows from the Cauchy-Schwartz inequality. To see this simply observe that (1.1) yields

$$\begin{aligned} \|f+h\|^2 &= \|f\|^2 + 2\Re(f,h) + \|h\|^2 \leq \|f\|^2 + 2|(f,h)| + \|h\|^2 \\ &\leq \|f\|^2 + 2\|f\|\|h\| + \|h\|^2 = (\|f\| + \|h\|)^2 \,. \end{aligned}$$

Hence $\|f+h\|^2 \leq (\|f\|+\|h\|)^2$. By taking the square root we arrive at the triangle inequality $\|f+h\| \leq \|f\| + \|h\|$.

Let $\mathcal{K}$ be an inner product space. Then we say that $\{h_i\}_0^\infty$ is a *Cauchy sequence* in $\mathcal{K}$ if $\|h_i - h_j\|$ approaches zero as i and j tend to infinity. An inner product space is *complete* if every Cauchy sequence converges to a vector in $\mathcal{K}$, that is, if $\{h_i\}_0^\infty$ is any Cauchy sequence, then $\{h_i\}_0^\infty$ converges to a vector h in $\mathcal{K}$. A *Hilbert space* is a complete inner product space. A Hilbert space is separable if it is the closure of a countable set. Throughout we only consider separable Hilbert spaces. If $\mathcal{K}$ is a (separable) Hilbert space, then there exists an orthonormal basis $\{\varphi_j\}_1^m$ for $\mathcal{K}$ where m is possibly infinite. In fact, m is the dimension of $\mathcal{K}$. In this case, every h in $\mathcal{K}$ admits a Fourier series expansion of the form

$$h = \sum_{j=1}^{m} (h, \varphi_j)\varphi_j \qquad (h \in \mathcal{K}) \,. \tag{1.3}$$

Moreover, we also have Parseval's relation

$$(h,g) = \sum_{j=1}^{m} (h,\varphi_j)\overline{(g,\varphi_j)} \qquad (h,g \in \mathcal{K}) \,. \tag{1.4}$$

In particular, $\|h\|^2 = \sum_1^m |(h,\varphi_j)|^2$.

If $\mathcal{F}$ is a subset of a linear space $\mathcal{K}$, then $\bigvee \mathcal{F}$ denotes the linear span of $\mathcal{F}$. If $\mathcal{K}$ is a Hilbert space, then $\bigvee \mathcal{F}$ is the closed linear span of $\mathcal{F}$.

Throughout $\mathbb{C}^n$ is the Hilbert space formed by the set of all complex n tuples $[x_1, x_2, \cdots, x_n]^{tr}$ where x_j is in $\mathbb{C}$ for all $j = 1, 2, \cdots, n$ and tr denotes the transpose. The inner product on $\mathbb{C}^n$ is given by

$$(x,y) = \sum_{j=1}^{n} x_j \bar{y}_j$$

where $x = [x_1, x_2, \cdots, x_n]^{tr}$ and $y = [y_1, y_2, \cdots, y_n]^{tr}$. In this setting, the Cauchy-Schwartz inequality becomes

$$\left| \sum_{j=1}^{n} x_j \bar{y}_j \right| \leq \left(\sum_{j=1}^{n} |x_j|^2 \right)^{1/2} \left(\sum_{j=1}^{n} |y_j|^2 \right)^{1/2} .$$

Throughout $L^2 = L^2[0, 2\pi\,]$ is the Hilbert space formed by the set of all square integrable Lebesgue measurable functions over the interval $[0, 2\pi\,]$, that is, f is in L^2 if and only if f is Lebesgue measurable on $[0, 2\pi\,]$ and

$$\int_0^{2\pi} |f(t)|^2\, dt < \infty\,.$$

The inner product on L^2 is given by

$$(f, g) = \frac{1}{2\pi} \int_0^{2\pi} f(t)\overline{g(t)}\, dt\,.$$

In this setting, the Cauchy-Schwartz inequality becomes

$$\left| \int_0^{2\pi} f(t)\overline{g(t)}\, dt \right| \leq \left(\int_0^{2\pi} |f(t)|^2\, dt \right)^{1/2} \left(\int_0^{2\pi} |g(t)|^2\, dt \right)^{1/2}\,.$$

Notice that the fraction $1/2\pi$ cancels out in the Cauchy-Schwartz inequality.

The space of all random variables with finite variance is widely used in filtering theory. To be precise, consider the following space

$$\mathcal{K} = \{x : x \text{ is a random variable satisfying } E|x|^2 < \infty\}\,. \tag{1.5}$$

Here E is the expectation. The inner product on $\mathcal{K}$ is given by

$$(x, y) = Ex\overline{y} \qquad\qquad (x, y \in \mathcal{K})\,.$$

Using arguments from measure theory it is easy to show that $\mathcal{K}$ is a Hilbert space. Notice that in this setting the norm of a random variable x in $\mathcal{K}$ is given by $\|x\| = \sqrt{E|x|^2}$. Recall that the expectation of a constant c is simply c, that is, $Ec = c$. In particular, $\|1\| = 1$.

The Cauchy-Schwartz inequality for the space of all random variables in (1.5) is given by

$$|Ex\overline{y}| \leq \sqrt{E|x|^2}\sqrt{E|y|^2} \qquad\qquad (x, y \in \mathcal{K})\,.$$

Finally, notice that $E|x| \leq \sqrt{E|x|^2}$. In particular, if x is in $\mathcal{K}$, then the mean Ex is finite, and thus, the variance $\sigma_x^2 = E|x|^2 - |Ex|^2$ of x is also finite. To see this observe by applying that Cauchy-Schwartz's inequality to the random variable $|x|$, we have

$$E|x| = E|x|\overline{1} \leq \sqrt{E|x|^2}\sqrt{E|1|^2} = \sqrt{E|x|^2}\,.$$

Hence $E|x| \leq \sqrt{E|x|^2}$.

We say that T is an *operator* if T is a linear map from a linear space $\mathcal{U}$ into a linear space $\mathcal{K}$. Now let $\mathcal{U}$ and $\mathcal{K}$ be Hilbert spaces and T be an operator from $\mathcal{U}$ into $\mathcal{K}$. The norm of T is defined by

$$\|T\| := \sup\{\|Tu\| : u \in \mathcal{U} \text{ and } \|u\| \leq 1\}\,.$$

We say that T is a *bounded operator* if $\|T\|$ is finite. Throughout these notes we deal mainly with bounded operators. Notice that if T, R and S are operators acting between the appropriate Hilbert spaces, then $\|TR\| \leq \|T\|\,\|R\|$ and $\|T + S\| \leq \|T\| + \|S\|$.

As before, let T be a bounded operator mapping $\mathcal{U}$ into $\mathcal{K}$. Then the adjoint T^* of T is the linear operator from $\mathcal{K}$ into $\mathcal{U}$ uniquely determined by $(Tu, k) = (u, T^*k)$ for all u in $\mathcal{U}$ and k in $\mathcal{K}$. It is well known that T and T^* have the same norm, that is, $\|T\| = \|T^*\|$. If $\mathcal{K}$ is finite dimensional, then $\|T\|^2$ equals the largest eigenvalue of T^*T. Finally, it is noted that if T is a matrix from $\mathbb{C}^n$ into $\mathbb{C}^m$, then T^* is the conjugate transpose of T, that is, $T^* = \bar{T}^{tr}$. The transpose of a matrix is denoted by tr.

Now let T be an operator on $\mathcal{X}$. Then we say that T is a self-adjoint operator if $T = T^*$. An operator T on $\mathcal{X}$ is positive, denoted by $T \geq 0$, if $(Tx, x) \geq 0$ for all vectors x in $\mathcal{X}$. It is well known that if T is positive, then T is a self-adjoint operator. Throughout these notes we will use some elementary results from Hilbert spaces. For some references on Hilbert spaces see Akhiezer-Glazman [1], Balakrishnan [3], Conway [7], Gohberg-Goldberg [17], Halmos [18] and Taylor-Lay [30].

1.2 The projection theorem

In this section we will introduce the projection theorem. Then the projection theorem will be used to solve some basic least squares optimization problems.

To establish some notation, let $\mathcal{K}$ be a Hilbert space. Then two vectors f and g in $\mathcal{K}$ are orthogonal, denoted by $f \perp g$, if $(f, g) = 0$. We say that f is orthogonal to a set $\mathcal{H}$, denoted by $f \perp \mathcal{H}$, if $(f, g) = 0$ for all g in $\mathcal{H}$. We say that $\mathcal{H}$ is a *subspace* of $\mathcal{K}$ if $\mathcal{H}$ is a closed linear space contained in $\mathcal{K}$. The subspace $\mathcal{H}$ can be zero $\{0\}$ or the whole space $\mathcal{K}$. Finally, if $\mathcal{H}$ is a subspace, then the orthogonal complement of $\mathcal{H}$ is the subspace of $\mathcal{K}$ defined by $\mathcal{H}^\perp = \{f \in \mathcal{K} : f \perp \mathcal{H}\}$.

Let f be a vector in $\mathcal{K}$ and $\mathcal{H}$ a subspace of $\mathcal{K}$. A basic least squares optimization problem is to find an element $\widehat{f}$ in $\mathcal{H}$, which is closer to f than any other element of $\mathcal{H}$. This naturally leads to the following optimization problem:

$$d(f, \mathcal{H}) = \inf\{\|f - h\| : h \in \mathcal{H}\} . \tag{2.1}$$

The distance from f to the subspace $\mathcal{H}$ is defined as $d(f, \mathcal{H})$. By a slight abuse of terminology we sometimes abbreviate the above optimization problem as $d(f, \mathcal{H}) = \inf \|f - \mathcal{H}\|$. Because $\mathcal{H}$ is closed, it follows that the distance from f to $\mathcal{H}$ is zero if and only if f is a vector in $\mathcal{H}$. The following theorem, known as the projection theorem, shows that there is a unique vector $\widehat{f}$ in $\mathcal{H}$ which achieves the minimum, that is, $d(f, \mathcal{H}) = \|f - \widehat{f}\|$. The projection theorem will play a fundamental role throughout these notes.

THEOREM 1.2.1 (Projection Theorem.) *Let $\mathcal{H}$ be a subspace of a Hilbert space $\mathcal{K}$. Then for every f in $\mathcal{K}$, there exists a unique vector $\widehat{f}$ in $\mathcal{H}$ solving the following optimization problem:*

$$\|f - \widehat{f}\| = \inf\{\|f - h\| : h \in \mathcal{H}\} . \tag{2.2}$$

Moreover, $\widehat{f}$ is the only vector in $\mathcal{H}$ such that $f - \widehat{f}$ is orthogonal to $\mathcal{H}$, that is, if h is any vector in $\mathcal{H}$ and $f - h \perp \mathcal{H}$, then $h = \widehat{f}$ is the unique solution to the optimization problem in (2.2). Finally, if $\widetilde{f} = f - \widehat{f}$, then the distance $d(f, \mathcal{H}) = \|\widetilde{f}\|$ is given by

$$d(f, \mathcal{H})^2 = \|\widetilde{f}\|^2 = \|f - \widehat{f}\|^2 = \|f\|^2 - \|\widehat{f}\|^2 . \tag{2.3}$$

If $\widehat{f}$ is the unique solution to the optimization problem in (2.2), then we say that $\widehat{f}$ is the *orthogonal projection* of f onto the subspace $\mathcal{H}$. The orthogonal projection onto $\mathcal{H}$ is denoted by $P_{\mathcal{H}}$, that is, $\widehat{f} = P_{\mathcal{H}} f$. In other words, $P_{\mathcal{H}} f$ is the unique vector in $\mathcal{H}$ which comes closest to f. Obviously, f is in $\mathcal{H}$, if and only if $f = P_{\mathcal{H}} f$. Finally, it is noted that the range of $P_{\mathcal{H}}$ equals $\mathcal{H}$.

We will not present a proof of the projection theorem. However, we will establish a few important facts concerning this theorem. In many applications, one computes the orthogonal projection $\widehat{f} = P_{\mathcal{H}} f$ by finding the unique vector $\widehat{f}$ in $\mathcal{H}$ such that $f - \widehat{f}$ is orthogonal to $\mathcal{H}$. So let us directly show that if $\widehat{f}$ is in $\mathcal{H}$ and $f - \widehat{f}$ is orthogonal to $\mathcal{H}$, then $\widehat{f}$ is the unique solution to the optimization problem in (2.2), and thus, $\widehat{f} = P_{\mathcal{H}} f$. To see this, recall that if x and y are any vectors in $\mathcal{K}$, then

$$\|x + y\|^2 = \|x\|^2 + 2\Re(x, y) + \|y\|^2 .$$

Let h be any vector in $\mathcal{H}$. Then using $x = f - \widehat{f}$ and $y = \widehat{f} - h$, we have

$$\begin{aligned} \|f - h\|^2 &= \|f - \widehat{f} + \widehat{f} - h\|^2 = \|f - \widehat{f}\|^2 + 2\Re(f - \widehat{f}, \widehat{f} - h) + \|\widehat{f} - h\|^2 \\ &= \|f - \widehat{f}\|^2 + \|\widehat{f} - h\|^2 . \end{aligned} \tag{2.4}$$

Notice that $(f - \widehat{f}, \widehat{f} - h) = 0$ because $\widehat{f} - h$ is in the linear space $\mathcal{H}$ and $f - \widehat{f}$ is orthogonal to $\mathcal{H}$. Equation (2.4), yields

$$\|f - h\|^2 = \|f - \widehat{f}\|^2 + \|\widehat{f} - h\|^2 \geq \|f - \widehat{f}\|^2 . \tag{2.5}$$

This readily implies that

$$\|f - \widehat{f}\|^2 \leq \inf\{\|f - h\|^2 : h \in \mathcal{H}\} = d(f, \mathcal{H})^2 .$$

Because $\widehat{f}$ is in $\mathcal{H}$, it follows that we have equality, that is, $\|f - \widehat{f}\| = d(f, \mathcal{H})$. Equation (2.5) also shows that $\widehat{f}$ is the only solution to the optimization problem in (2.2). If h is another solution to this optimization problem, then $\|f - h\| = \|f - \widehat{f}\|$. By consulting (2.5), this implies that $\|\widehat{f} - h\|^2 = 0$. Hence, $\widehat{f} = h$ which proves our claim.

To establish (2.3), recall that $f - \widehat{f}$ is orthogonal to $\mathcal{H}$. Since $\widehat{f}$ is in $\mathcal{H}$, it follows that $(f - \widehat{f}, \widehat{f})$ is zero. Using this, we have

$$\begin{aligned} \|f - \widehat{f}\|^2 &= \|f\|^2 - 2\Re(f, \widehat{f}) + \|\widehat{f}\|^2 \\ &= \|f\|^2 - 2\Re(f - \widehat{f} + \widehat{f}, \widehat{f}) + \|\widehat{f}\|^2 \\ &= \|f\|^2 - 2\Re(f - \widehat{f}, \widehat{f}) - 2\Re(\widehat{f}, \widehat{f}) + \|\widehat{f}\|^2 \\ &= \|f\|^2 - 2\|\widehat{f}\|^2 + \|\widehat{f}\|^2 = \|f\|^2 - \|\widehat{f}\|^2 . \end{aligned}$$

This and $\widetilde{f} = f - \widehat{f}$ yields (2.3).

As before, let $\widehat{f}$ be the orthogonal projection of f onto the subspace $\mathcal{H}$. Notice that $f = \widehat{f} + \widetilde{f}$ where $\widetilde{f} = f - \widehat{f}$. According to the projection theorem $\widetilde{f}$ is orthogonal to $\mathcal{H}$, that is, $\widetilde{f}$ is a vector in $\mathcal{H}^{\perp}$. Therefore every vector f in $\mathcal{K}$ admits a unique orthogonal

decomposition of the form $f = \widehat{f} + \widetilde{f}$ where $\widehat{f}$ is in $\mathcal{H}$ and $\widetilde{f}$ is in $\mathcal{H}^\perp$. In fact, $\widehat{f}$ is the orthogonal projection of f onto $\mathcal{H}$ and $\widetilde{f}$ is the orthogonal projection of f onto $\mathcal{H}^\perp$. Moreover, $\|f\|^2 = \|\widehat{f}\|^2 + \|\widetilde{f}\|^2$. Motivated by this decomposition, we introduce the notation $\mathcal{K} = \mathcal{H} \oplus \mathcal{N}$. This means that $\mathcal{H}$ and $\mathcal{N}$ are two orthogonal spaces which span $\mathcal{K}$, that is, every vector f in $\mathcal{K}$ admits a unique orthogonal decomposition of the form $f = \widehat{f} + \widetilde{f}$ where $\widehat{f}$ is in $\mathcal{H}$, while $\widetilde{f}$ is in $\mathcal{N}$ and the subspace $\mathcal{H}$ is orthogonal to $\mathcal{N}$. If $\mathcal{K} = \mathcal{H} \oplus \mathcal{N}$, then obviously $\mathcal{N} = \mathcal{H}^\perp$. If $\mathcal{K} = \oplus_1^n \mathcal{H}_j$, then $\mathcal{H}_1, \mathcal{H}_2, \cdots, \mathcal{H}_n$ are n pairwise orthogonal subspaces which span $\mathcal{K}$. If $\mathcal{H}$ and $\mathcal{R}$ are two subspaces satisfying $\mathcal{H} \subset \mathcal{R}$, then $\mathcal{R} \ominus \mathcal{H}$ denotes the orthogonal complement of $\mathcal{H}$ in $\mathcal{R}$, that is, $\mathcal{R} \ominus \mathcal{H} = \{g \in \mathcal{R} : g \perp \mathcal{H}\}$. Obviously, $\mathcal{H}^\perp = \mathcal{K} \ominus \mathcal{H}$.

Recall that $P_\mathcal{H}$ is the orthogonal projection onto the subspace $\mathcal{H}$, that is, $\widehat{f} = P_\mathcal{H} f$ where $\widehat{f}$ is the unique solution to the optimization problem in (2.2). We claim that $P_\mathcal{H}$ is a positive operator on $\mathcal{K}$ satisfying $P_\mathcal{H} = P_\mathcal{H}^2 = P_\mathcal{H}^*$. Moreover, the range of $P_\mathcal{H}$ equals $\mathcal{H}$ and $0 \le P_\mathcal{H} \le I$. To verify this, first notice that $P_\mathcal{H}$ is a mapping from $\mathcal{K}$ into $\mathcal{K}$ whose range equals $\mathcal{H}$. If f is in $\mathcal{H}$, then obviously, $P_\mathcal{H} f = f$. Hence $P_\mathcal{H}^2 = P_\mathcal{H}$. Now let us show that $P_\mathcal{H}$ is a linear map, that is, $P_\mathcal{H}(\alpha f + \beta g) = \alpha P_\mathcal{H} f + \beta P_\mathcal{H} g$ for all vectors f, g in $\mathcal{K}$ and scalars α, β. To this end, let $\widehat{f} = P_\mathcal{H} f$ and $\widehat{g} = P_\mathcal{H} g$. Because both $f - \widehat{f}$ and $g - \widehat{g}$ are orthogonal to $\mathcal{H}$, it follows that $\alpha f + \beta g - (\alpha \widehat{f} + \beta \widehat{g})$ is also orthogonal to $\mathcal{H}$. Clearly, the vector $\alpha \widehat{f} + \beta \widehat{g}$ is in the subspace $\mathcal{H}$. By the projection theorem $\alpha \widehat{f} + \beta \widehat{g}$ must be the orthogonal projection of $\alpha f + \beta g$ onto the subspace $\mathcal{H}$. Therefore $P_\mathcal{H}(\alpha f + \beta g) = \alpha P_\mathcal{H} f + \beta P_\mathcal{H} g$. In other words, $P_\mathcal{H}$ is a linear map. Recall that any vector f in $\mathcal{K}$ admits an orthogonal decomposition of the form $f = \widehat{f} + \widetilde{f}$ where $\widehat{f} = P_\mathcal{H} f$ and $\widetilde{f}$ is in $\mathcal{H}^\perp$. Using this, we obtain

$$\|P_\mathcal{H} f\|^2 = \|\widehat{f}\|^2 \le \|\widehat{f}\|^2 + \|\widetilde{f}\|^2 = \|f\|^2 .$$

So $\|P_\mathcal{H}\| \le 1$. Therefore the orthogonal projection is a bounded operator. Finally, notice that for any f in $\mathcal{K}$, we have

$$(P_\mathcal{H} f, f) = (\widehat{f}, \widehat{f} + \widetilde{f}) = (\widehat{f}, \widehat{f}) \le (f, f) .$$

This readily implies that $0 \le P_\mathcal{H} \le I$. In particular, $P_\mathcal{H}$ is a self-adjoint operator. Therefore the orthogonal projection is an operator satisfying $P_\mathcal{H} = P_\mathcal{H}^2 = P_\mathcal{H}^*$.

An operator P on $\mathcal{K}$ is an *orthogonal projection* if $P = P_\mathcal{H}$ where $P_\mathcal{H}$ is an orthogonal projection onto some subspace $\mathcal{H}$. For example, if $P_\mathcal{H}$ is an orthogonal projection onto $\mathcal{H}$, then it is easy to verify that $I - P_\mathcal{H}$ is the orthogonal projection onto $\mathcal{H}^\perp$, that is, $I - P_\mathcal{H} = P_{\mathcal{H}^\perp}$. We claim that an operator P on $\mathcal{K}$ is an orthogonal projection if and only if $P = P^2 = P^*$. In this case, $P = P_\mathcal{H}$ where $\mathcal{H} = \operatorname{ran} P := P\mathcal{K}$. (The range of an operator is denoted by ran .) To prove this fact it remains to show that if $P = P^2 = P^*$, then the range of P is closed and $P = P_\mathcal{H}$ where $\mathcal{H} = \operatorname{ran} P$. (The range of an operator T is closed if $\operatorname{ran} T$ contains all its limit points.) Notice that for f in $\mathcal{K}$, we have $\|Pf\|^2 = (P^*Pf, f) = (Pf, f) \le \|Pf\| \, \|f\|$. This implies that $\|Pf\| \le \|f\|$. Thus P is a contraction, that is, $\|P\| \le 1$. To show that the range of P is closed, let $\{f_n\}_1^\infty$ be any sequence of vectors in $\mathcal{K}$ such that Pf_n approaches f as n tends towards ∞. Then using the fact that P is a contraction along with $P = P^2$, we have

$$\begin{aligned} \|Pf - f\| &= \|Pf - Pf_n + Pf_n - f\| \le \|Pf - Pf_n\| + \|Pf_n - f\| \\ &= \|P(f - Pf_n)\| + \|Pf_n - f\| \le 2\|Pf_n - f\| \to 0 \end{aligned}$$

as n tends to ∞. Hence $\|Pf - f\| = 0$. Thus $Pf = f$ and f is in the range of P. So the range of P is closed. Finally, using $P = P^2 = P^*$, a simple calculation shows that $f - Pf$ is orthogonal to $P\mathcal{K}$. By the projection theorem, $\widehat{f} = Pf$ is the orthogonal projection onto the range of P, that is, $P = P_{\mathcal{H}}$ where $\mathcal{H} = \operatorname{ran} P$. This completes the proof.

1.3 The Gram matrix

In this section, we use the projection theorem to solve a classical optimization problem via the Gram matrix. To this end, let T be an operator on a finite dimensional Hilbert space $\mathcal{X}$. Recall that T is positive, denoted by, $T \geq 0$ if $(Tx, x) \geq 0$ for all vectors x in $\mathcal{X}$. It is well known that T is positive if and only if $T = T^*$ and all the eigenvalues of T are positive (≥ 0). We say that T is strictly positive, denoted by, $T > 0$ if $(Tx, x) > 0$ for all nonzero vectors x in $\mathcal{X}$. It is well known that T is strictly positive if and only if $T = T^*$ and all the eigenvalues of T are strictly positive (> 0). Moreover, T is strictly positive if and only if T is positive and invertible. If $\mathcal{X} = \mathbb{C}^n$, then T is positive, if and only if $x^*Tx \geq 0$ for all vectors x in $\mathbb{C}^n$. (Recall that $*$ is the conjugate transpose for vectors in $\mathbb{C}^n$.) Furthermore, T is strictly positive, if and only if $x^*Tx > 0$ for all nonzero vectors x in $\mathbb{C}^n$. Finally, it is noted that the notion of a positive operator also extends to an infinite dimensional Hilbert space. To be precise, an operator T on $\mathcal{X}$ is positive $T \geq 0$ if $(Tx, x) \geq 0$ for all vectors x in $\mathcal{X}$. An operator T on $\mathcal{X}$ is strictly positive if there exists a scalar $\delta > 0$ such that $(Tx, x) \geq \delta\|x\|^2$ for all vectors x in $\mathcal{X}$.

Let $\{f_i\}_1^n$ be a set of vectors in a Hilbert space $\mathcal{K}$. Then the Gram matrix associated with $\{f_i\}_1^n$ is defined by

$$G = \begin{bmatrix} (f_1, f_1) & (f_1, f_2) & \cdots & (f_1, f_n) \\ (f_2, f_1) & (f_2, f_2) & \cdots & (f_2, f_n) \\ \vdots & \vdots & & \vdots \\ (f_n, f_1) & (f_n, f_2) & \cdots & (f_n, f_n) \end{bmatrix}. \tag{3.1}$$

Notice that the entries of the Gram matrix G are given by $G_{ij} = (f_i, f_j)$. Moreover, G is a self-adjoint matrix, that is, $G = G^*$. The following result shows that G is strictly positive if and only if $\{f_i\}_1^n$ are linearly independent.

THEOREM 1.3.1 *Let G be the Gram matrix on $\mathbb{C}^n$ determined by a set of vectors $\{f_i\}_1^n$ in a Hilbert space $\mathcal{K}$. Then G is a positive matrix. Moreover, G is strictly positive if and only if $\{f_i\}_1^n$ are linearly independent.*

Proof. For any set of scalars $\{\alpha_k\}_1^n$ we have

$$\begin{aligned} 0 &\leq \|\sum_{i=1}^n \alpha_i f_i\|^2 = (\sum_{i=1}^n \alpha_i f_i, \sum_{j=1}^n \alpha_j f_j) = \sum_{i=1}^n \sum_{j=1}^n \alpha_i (f_i, f_j)\overline{\alpha}_j \\ &= \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{bmatrix} G \begin{bmatrix} \overline{\alpha}_1 & \overline{\alpha}_2 & \cdots & \overline{\alpha}_n \end{bmatrix}^{tr}. \end{aligned} \tag{3.2}$$

Here tr denotes the transpose. Let $x = \begin{bmatrix} \overline{\alpha}_1 & \overline{\alpha}_2 & \cdots & \overline{\alpha}_n \end{bmatrix}^{tr}$. Because the scalars $\{\alpha_i\}_1^n$ are arbitrary $0 \leq x^*Gx$ for all vectors x in $\mathbb{C}^n$. Hence G is positive.

Clearly, a vector h is zero if and only if its norm $\|h\| = 0$. The calculation in (3.2) shows that

$$0 = \sum_{i=1}^{n} \alpha_i f_i \quad \text{if and only if} \quad 0 = \left[\begin{array}{cccc} \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{array} \right] G \left[\begin{array}{cccc} \overline{\alpha}_1 & \overline{\alpha}_2 & \cdots & \overline{\alpha}_n \end{array} \right]^{tr} . \tag{3.3}$$

Recall that the vectors $\{f_i\}_1^n$ are linearly independent if $0 = \sum_1^n \alpha_i f_i$ implies that $\alpha_k = 0$ for $k = 1, 2, \cdots, n$. By consulting (3.3) we see that $\{f_i\}_1^n$ are linearly independent if and only if $x^* G x = 0$ implies that x is zero. Since G is positive, $\{f_i\}_1^n$ are linearly independent if and only if G is strictly positive. This completes the proof.

Let $\{f_i\}_1^n$ be a finite set of vectors in a Hilbert space $\mathcal{K}$ and $\mathcal{H}$ be the space spanned by these vectors. A classical least squares optimization problem is to compute the orthogonal projection $\widehat{f} = P_{\mathcal{H}} f$ for some fixed f in $\mathcal{K}$. In other words, find an element $\widehat{f}$ of $\mathcal{H}$ to solve the following optimization problem

$$\|f - \widehat{f}\| = \inf \left\{ \|f - \sum_{i=1}^{n} \alpha_i f_i\| : \alpha_i \in \mathbb{C} \right\} . \tag{3.4}$$

Without loss of generality we can assume that the set of vectors $\{f_i\}_1^n$ are linearly independent. The following result presents a solution to this optimization problem.

THEOREM 1.3.2 *Let $\{f_i\}_1^n$ be a linearly independent set of vectors in a Hilbert space $\mathcal{K}$. Let $\mathcal{H}$ be the linear subspace spanned by $\{f_i\}_1^n$, and G be the Gram matrix in (3.1) generated by $\{f_i\}_1^n$. Finally, let f be a vector in $\mathcal{K}$. Then the orthogonal projection $\widehat{f}$ of f onto $\mathcal{H}$ is given by*

$$\widehat{f} = P_{\mathcal{H}} f = \sum_{i=1}^{n} \alpha_i f_i \tag{3.5}$$

where the scalars $\{\alpha_i\}_1^n$ are computed by

$$\left[\begin{array}{cccc} \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{array} \right] = \left[\begin{array}{cccc} (f, f_1) & (f, f_2) & \cdots & (f, f_n) \end{array} \right] G^{-1} . \tag{3.6}$$

In particular, the vector $\widehat{f}$ is the unique solution to the optimization problem in (3.4). Moreover, the norm of $\widehat{f}$ is computed by

$$\|\widehat{f}\|^2 = \left[\begin{array}{cccc} (f, f_1) & (f, f_2) & \cdots & (f, f_n) \end{array} \right] G^{-1} \left[\begin{array}{cccc} (f, f_1) & (f, f_2) & \cdots & (f, f_n) \end{array} \right]^* . \tag{3.7}$$

Finally, the optimal error $d(f, \mathcal{H})^2 = \|f\|^2 - \|\widehat{f}\|^2$.

PROOF. According to the projection theorem, $\widehat{f} = P_{\mathcal{H}} f$ is the unique vector in $\mathcal{H}$ such that $f - \widehat{f}$ is orthogonal to $\mathcal{H}$. Since $\mathcal{H}$ is the linear span of $\{f_i\}_1^n$, it follows that $\widehat{f}$ is a vector of the form $\widehat{f} = \sum_1^n \alpha_i f_i$ where $\{\alpha_k\}_1^n$ are constants. Using the fact that $f - \widehat{f}$ is orthogonal to $\mathcal{H}$, we see that $f - \sum_1^n \alpha_i f_i$ is orthogonal to $\{f_j\}_1^n$. In other words,

$$0 = (f - \widehat{f}, f_j) = (f - \sum_{i=1}^{n} \alpha_i f_i, f_j) = (f, f_j) - \sum_{i=1}^{n} \alpha_i (f_i, f_j) \qquad (j = 1, 2, \cdots, n) .$$

By rewriting this in matrix form, we arrive at

$$\left[\begin{array}{cccc} \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{array}\right] G = \left[\begin{array}{cccc} (f,f_1) & (f,f_2) & \cdots & (f,f_n) \end{array}\right] .$$

By inverting the Gram matrix G, we obtain the matrix expression for the coefficients $\{\alpha_i\}_1^n$ in (3.6). This yields the formula for $\widehat{f}$ in (3.5).

To complete the proof it remains to establish (3.7). Let $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_n]$. Using (3.6), we obtain

$$\begin{aligned} \|\widehat{f}\|^2 &= (\sum_{i=1}^n \alpha_i f_i, \sum_{j=1}^n \alpha_j f_j) = \sum_{i=1}^n \sum_{j=1}^n \alpha_i (f_i, f_j) \overline{\alpha}_j = \alpha G \alpha^* \\ &= \left[\begin{array}{cccc} (f,f_1) & (f,f_2) & \cdots & (f,f_n) \end{array}\right] G^{-1} \left[\begin{array}{cccc} (f,f_1) & (f,f_2) & \cdots & (f,f_n) \end{array}\right]^* . \end{aligned}$$

This completes the proof.

REMARK 1.3.3 *Suppose that $\{f_i\}_1^n$ forms an orthonormal set in a Hilbert space $\mathcal{K}$, that is, $(f_i, f_j) = 0$ for all $i \neq j$ and $\|f_i\| = 1$ for $i = 1, 2, \cdots, n$. Then clearly, $G = I$ and $\{f_i\}_1^n$ is linearly independent. Moreover, if $\mathcal{H}$ is the space spanned by $\{f_i\}_1^n$, then formulas (3.5), (3.6) and (3.7) in Theorem 1.3.2 show that for f in $\mathcal{K}$,*

$$P_{\mathcal{H}} f = \sum_{i=1}^n (f, f_i) f_i \qquad \text{and} \qquad \|P_{\mathcal{H}} f\|^2 = \sum_{i=1}^n |(f, f_i)|^2 . \tag{3.8}$$

This is the classical Fourier representation for the orthogonal projection in terms of an orthonormal basis. In particular, if f is in $\mathcal{H}$, then the second equation in (3.8) reduces to

$$\|f\|^2 = \sum_{i=1}^n |(f, f_i)|^2 \qquad (\text{when } f \in \mathcal{H}) . \tag{3.9}$$

1.3.1 An approximation of e^t

In this example will use Theorem 1.3.2 to find an approximation of e^t by a polynomial of degree at most two over the interval $[0, 1]$. Throughout $L^2[0, 1]$ is the Hilbert space formed by the set of all square integrable Lebesgue measurable functions over the interval $[0, 1]$, that is, f is in $L^2[0, 1]$ if and only if f is a Lebesgue measurable function on $[0, 1]$ and

$$\int_0^1 |f(t)|^2 \, dt < \infty .$$

The inner product on $L^2[0, 1]$ is given by

$$(f, g) = \int_0^1 f(t) \overline{g(t)} \, dt .$$

Now consider the problem of approximating the exponential function, $f(t) = e^t$, by a polynomial of degree at most two in the $L^2[0, 1]$ norm. To be precise, we wish to find the optimal polynomial $\widehat{f} = \alpha_0 + \alpha_1 t + \alpha_2 t^2$ to solve the optimization problem

$$\|e^t - \widehat{f}\|^2 = \inf\{ \int_0^1 |e^t - \alpha_0 - \alpha_1 t - \alpha_2 t^2|^2 \, dt : \ \alpha_i \in \mathbb{C} \} . \tag{3.10}$$

To obtain a solution to this problem, let $f_1 = 1$, $f_2 = t$ and $f_3 = t^2$. Clearly, f_1, f_2 and f_3 are linearly independent. Therefore the Gram matrix G corresponding to these vectors is strictly positive. The optimal polynomial $\widehat{f}$ is given by (3.5) and (3.6) in Theorem 1.3.2. In this case, the entries of the Gram matrix are given by

$$G_{ij} = (f_i, f_j) = \int_0^1 t^{i-1} t^{j-1}\, dt = \frac{1}{i+j-1} \qquad (i, j = 1, 2, 3)\,.$$

So the Gram matrix G becomes

$$G = \begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix}. \tag{3.11}$$

Furthermore, the row vector generated by $\{(f, f_j)\}_1^3$ is given by

$$\begin{bmatrix} (e^t, 1) & (e^t, t) & (e^t, t^2) \end{bmatrix} = \begin{bmatrix} e^1 - 1 & 1 & e^1 - 2 \end{bmatrix}. \tag{3.12}$$

By combining (3.11) and (3.12), we obtain

$$\begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 \end{bmatrix} = \begin{bmatrix} e^1 - 1 & 1 & e^1 - 2 \end{bmatrix} G^{-1} \approx \begin{bmatrix} 1.01 & 0.851 & 0.839 \end{bmatrix}.$$

Hence the optimal polynomial $\widehat{f} = \sum_1^3 \alpha_i f_i$ of degree at most two approximating e^t in the $L^2[0, 1]$ norm is given by

$$\widehat{f} \approx 1.01 + 0.85t + 0.84t^2\,. \tag{3.13}$$

Finally, it is noted that the optimal polynomial $\widehat{f}$ in (3.13) does not equal $1 + t/1! + t^2/2!$ which comes from the Taylor series expansion of e^t.

1.4 An application to observability

In this section we will show how the projection theorem can be used to solve a standard observability optimization problem in linear systems. To this end, consider the Hilbert space $L^2([0, t_1], \mathbb{C}^m)$ consisting of the set of all function f with values in $\mathbb{C}^m$ of the form

$$f(t) = \begin{bmatrix} f_1(t) & f_2(t) & \cdots & f_m(t) \end{bmatrix}^{tr}$$

where $f_k(t)$ is a function in $L^2[0, t_1]$ for all $k = 1, 2, \cdots, m$. (Recall that tr denotes the transpose.) The inner product on $L^2([0, t_1], \mathbb{C}^m)$ is defined by

$$(f, g) = \int_0^{t_1} (f(t), g(t))_{\mathbb{C}^m}\, dt = \sum_{k=1}^{m} \int_0^{t_1} f_k(t)\overline{g_k(t)}\, dt$$

where $f = \begin{bmatrix} f_1 & f_2 & \cdots & f_m \end{bmatrix}^{tr}$ and $g = \begin{bmatrix} g_1 & g_2 & \cdots & g_m \end{bmatrix}^{tr}$.

Now consider the following state space system

$$\dot{x} = Ax \qquad \text{and} \qquad y = Cx\,. \tag{4.1}$$

Here A is a matrix on $\mathbb{C}^n$ and C is a matrix mapping $\mathbb{C}^n$ into $\mathbb{C}^m$. The solution to this system is given by $x(t) = e^{At}x_0$ and $y(t) = Ce^{At}x_0$ where $x_0 = x(0)$ is the initial condition. Recall that the pair $\{C, A\}$ is observable if

$$\operatorname{rank} \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} = n\,.$$

Finally, let P be the finite time observability Gramian for the pair $\{C, A\}$ defined by

$$P = \int_0^{t_1} e^{A^*t} C^* C e^{At}\, dt\,. \tag{4.2}$$

Notice that P is a positive matrix on $\mathbb{C}^n$. To see this observe that for any z in $\mathbb{C}^n$, we have

$$(Pz, z) = \int_0^{t_1} (e^{A^*t}C^*Ce^{At}z, z)\, dt = \int_0^{t_1} (Ce^{At}z, Ce^{At}z)\, dt = \int_0^{t_1} \|Ce^{At}z\|^2\, dt \geq 0\,.$$

The second equality follows from the fact that if T is any matrix acting between two Euclidean spaces, then $(T\xi, \varphi) = (\xi, T^*\varphi)$ where ξ and φ are vectors. Hence $(Pz, z) \geq 0$ for all z in $\mathbb{C}^n$. In other words, P is a positive matrix. Finally, it is noted that P is strictly positive if and only if the pair $\{C, A\}$ is observable.

Let f be any function in $L^2([0, t_1], \mathbb{C}^m)$. Then the observability optimization problem is to find an optimal initial condition $\widehat{x}_0$ such that $y(t) = Ce^{At}\widehat{x}_0$ comes as close as possible to the specified function f. In other words, find an initial condition $\widehat{x}_0$ such that

$$d^2 = \int_0^{t_1} \|f(t) - Ce^{At}\widehat{x}_0\|^2\, dt = \inf\{\int_0^{t_1} \|f(t) - Ce^{At}x_0\|^2\, dt : x_0 \in \mathbb{C}^n\}\,. \tag{4.3}$$

Here $d^2 = \int_0^{t_1} \|f(t) - Ce^{At}\widehat{x}_0\|^2\, dt$ is the error in this optimization problem. Finally, if $\widehat{x}_0$ is the optimal initial condition, then $\widehat{x}(t) = e^{At}\widehat{x}_0$ is called the state estimate of $x(t)$ corresponding to $\widehat{x}_0$. Notice that the state estimate $\widehat{x}(t)$ depends upon the interval $[0, t_1]$. The following result provides a solution to the observability optimization problem.

THEOREM 1.4.1 *Consider the observability optimization problem in (4.3) for the pair $\{C, A\}$ where f is a specified function in $L^2([0, t_1], \mathbb{C}^m)$. Let P be the observability Gramian defined in (4.2). Then the following holds.*

(i) There exists a solution $\widehat{x}_0$ in $\mathbb{C}^n$ to the linear equation

$$P\widehat{x}_0 = \int_0^{t_1} e^{A^*t} C^* f(t)\, dt\,. \tag{4.4}$$

(ii) If $\widehat{x}_0$ in $\mathbb{C}^n$ is any solution to (4.4), then $\widehat{x}_0$ is an initial condition solving the observability optimization problem in (4.3). In this case, the error

$$d^2 = \int_0^{t_1} \|f(t)\|^2\, dt - (P\widehat{x}_0, \widehat{x}_0)\,. \tag{4.5}$$

(iii) If the pair $\{C, A\}$ is observable, then there exists a unique solution to the observability optimization problem in (4.3) and this unique solution is given by

$$\widehat{x}_0 = P^{-1} \int_0^{t_1} e^{A^* t} C^* f(t)\, dt\,. \tag{4.6}$$

In this case, the state estimate $\widehat{x}(t) = e^{At}\widehat{x}_0$.

PROOF. Our proof is based on the projection theorem. To this end, let $\mathcal{K} = L^2([0, t_1], \mathbb{C}^m)$. Then the observability optimization problem in (4.3) is equivalent to the following least squares optimization problem

$$\begin{aligned} d &= \inf\{\|f - Ce^{At}x_0\|_{\mathcal{K}} : x_0 \in \mathbb{C}^n\} \\ &= \inf\{\|f - h\|_{\mathcal{K}} : h = Ce^{At}x_0 \text{ and } x_0 \in \mathbb{C}^n\}\,. \end{aligned} \tag{4.7}$$

In this setting the subspace $\mathcal{H}$ of $\mathcal{K}$ is given by

$$\mathcal{H} = \{Ce^{At}x_0 : x_0 \in \mathbb{C}^n \text{ and } 0 \le t \le t_1\}\,. \tag{4.8}$$

According to the projection Theorem 1.2.1, there exists a unique solution to the optimization problem in (4.7). Moreover, this solution is given by the unique function $\widehat{f}$ in $\mathcal{H}$ such that $f - \widehat{f}$ is orthogonal to $\mathcal{H}$. By consulting the form of the subspace $\mathcal{H}$ in (4.8), we see that there exists an initial condition $\widehat{x}_0$ in $\mathbb{C}^n$ such that $\widehat{f} = Ce^{At}\widehat{x}_0$. (The initial condition $\widehat{x}_0$ is uniquely determined by $\widehat{f}$ if and only if the pair $\{C, A\}$ is observable.) Recall that if T is any matrix acting between the appropriate Euclidean spaces, then $(Tg, z) = (g, T^*z)$ where g and z are vectors. Since $\mathcal{H} = \{Ce^{At}x_0 : x_0 \in \mathbb{C}^n\}$, the vector $f - Ce^{At}\widehat{x}_0$ is orthogonal to $Ce^{At}x_0$ for all x_0 in $\mathbb{C}^n$. Thus

$$\begin{aligned} 0 &= (f - Ce^{At}\widehat{x}_0, Ce^{At}x_0)_{\mathcal{K}} = \int_0^{t_1} (f - Ce^{At}\widehat{x}_0, Ce^{At}x_0)_{\mathbb{C}^n}\, dt \\ &= \int_0^{t_1} (e^{A^*t}C^*(f - Ce^{At}\widehat{x}_0), x_0)_{\mathbb{C}^n}\, dt = (\int_0^{t_1} e^{A^*t}C^*(f - Ce^{At}\widehat{x}_0)\, dt, x_0)_{\mathbb{C}^n} = (q, x_0)_{\mathbb{C}^n} \end{aligned}$$

where $q = \int_0^{t_1} e^{A^*t}C^*(f - Ce^{At}\widehat{x}_0)dt$ is simply a vector in $\mathbb{C}^n$. Because $(q, x_0) = 0$ for all vectors x_0 in $\mathbb{C}^n$, the vector $q = 0$. In other words,

$$\begin{aligned} 0 &= \int_0^{t_1} e^{A^*t}C^*(f - Ce^{At}\widehat{x}_0)dt = \int_0^{t_1} e^{A^*t}C^*f(t)dt - \int_0^{t_1} e^{A^*t}C^*Ce^{At}dt\widehat{x}_0 \\ &= \int_0^{t_1} e^{A^*t}C^*f(t)dt - P\widehat{x}_0\,. \end{aligned}$$

This readily implies that

$$P\widehat{x}_0 = \int_0^{t_1} e^{A^*t}C^*f(t)dt\,.$$

Therefore the equation in (4.4) admits a solution. Moreover, if $\widehat{x}_0$ is any solution to (4.4), then $\widehat{f} = Ce^{At}\widehat{x}_0$ is the optimal solution to the observability optimization problem in (4.3).

Now let us compute the error d. By consulting (2.3) in the projection theorem, we obtain

$$\begin{aligned} d^2 &= \|f-\widehat{f}\|^2 = \|f\|^2 - \|\widehat{f}\|^2 = \int_0^{t_1} \|f\|^2\,dt - \int_0^{t_1} \|Ce^{At}\widehat{x}_0\|^2\,dt \\ &= \int_0^{t_1} \|f\|^2\,dt - \int_0^{t_1} (Ce^{At}\widehat{x}_0, Ce^{At}\widehat{x}_0)\,dt = \int_0^{t_1} \|f\|^2\,dt - \int_0^{t_1} (e^{A^*t}C^*Ce^{At}\widehat{x}_0, \widehat{x}_0)\,dt \\ &= \int_0^{t_1} \|f(t)\|^2\,dt - (\int_0^{t_1} e^{A^*t}C^*Ce^{At}\,dt\widehat{x}_0, \widehat{x}_0) = \int_0^{t_1} \|f(t)\|^2\,dt - (P\widehat{x}_0, \widehat{x}_0)\,. \end{aligned}$$

This readily yields (4.5). Therefore Parts (i) and (ii) hold.

Now assume that Part (iii) holds, that is, assume that the pair $\{C, A\}$ is observable. Then P in invertible. In this case, the solution to (4.4) is unique and is given by (4.6). This completes the proof.

Now assume that the pair $\{C, A\}$ is observable. Theorem 1.4.1 shows that the state estimate $\widehat{x}(t)$ for $x(t)$ corresponding to the optimal initial condition $\widehat{x}_0$ is given by

$$\widehat{x}(t) = e^{At}P(t)^{-1}\int_0^t e^{A^*\sigma}C^*f(\sigma)\,d\sigma \qquad \text{where} \qquad P(t) = \int_0^t e^{A^*\sigma}C^*Ce^{A\sigma}\,d\sigma\,. \tag{4.9}$$

If $f(t) = Ce^{At}x_0$, then $\widehat{x}_0 = x_0$ and $\widehat{x}(t) = x(t)$ for all $t > 0$. In many applications $\widehat{x}(t)$ is used to estimate the state $x(t)$ when the output measurement $f(t) = Cx(t) + w(t)$ is corrupted by noise $w(t)$. Riccati differential equations play an important role in Kalman filtering and state estimation. The following result uses a Riccati differential equation to compute the state estimate $\widehat{x}$ for x.

THEOREM 1.4.2 *Consider the observability optimization problem in (4.3) where the pair $\{C, A\}$ is observable. Let $\widehat{x}(t)$ be the state estimate of $x(t)$ defined in (4.9). Then $\widehat{x}(t)$ can be computed by solving the following state space system*

$$\dot{\widehat{x}} = A\widehat{x} + QC^*\left(f(t) - C\widehat{x}(t)\right) \tag{4.10}$$

where $Q(t)$ is the solution to the Riccati differential equation

$$\dot{Q} = AQ + QA^* - QC^*CQ \tag{4.11}$$

subject to the initial condition $Q(t_0) = e^{At_0}P(t_0)^{-1}e^{A^*t_0}$ where $t_0 > 0$.

Proof. Let $Q(t)$ be the operator defined by

$$Q(t) = e^{At}P(t)^{-1}e^{A^*t}\,. \tag{4.12}$$

Because the pair $\{C, A\}$ is controllable, $P(t)$ is invertible for all $t \geq t_0 > 0$. (Notice that in general $Q(0)$ does not exist.) Hence $Q(t)$ is well defined for all $t \geq t_0$. Now let us show that Q satisfies the Riccati differential equation in (4.11). To this end, recall that if an operator $\Omega(t)$ on $\mathcal{X}$ is invertible and differentiable for all t, then

$$\frac{d}{dt}\Omega(t)^{-1} = -\Omega^{-1}\dot{\Omega}\Omega^{-1}\,. \tag{4.13}$$

To see this simply observe that

$$0 = \frac{d}{dt} I = \frac{d}{dt} \Omega \Omega^{-1} = \dot{\Omega} \Omega^{-1} + \Omega \frac{d}{dt} \Omega^{-1} .$$

Multiplying this equation by Ω^{-1} on the left and rearranging the terms yields (4.13). Notice that

$$\dot{P} = e^{A^* t} C^* C e^{At} .$$

Using the definition of Q in (4.12) with the identity in (4.13), we have

$$\begin{aligned} \dot{Q} &= A e^{At} P(t)^{-1} e^{A^* t} + e^{At} P(t)^{-1} e^{A^* t} A^* + e^{At} \left(\frac{d}{dt} P(t)^{-1} \right) e^{A^* t} \\ &= AQ + QA^* - e^{At} P^{-1} \dot{P} P^{-1} e^{A^* t} \\ &= AQ + QA^* - e^{At} P^{-1} e^{A^* t} C^* C e^{At} P^{-1} e^{A^* t} \\ &= AQ + QA^* - QC^* CQ . \end{aligned}$$

This yields the Riccati differential equation in (4.11).

To obtain the state space equation for the estimate $\widehat{x}$ of x observe that

$$\begin{aligned} \dot{\widehat{x}} &= \frac{d}{dt} e^{At} P(t)^{-1} \int_0^t e^{A^* \sigma} C^* f(\sigma) \, d\sigma \\ &= A\widehat{x} + e^{At} \left(\frac{d}{dt} P(t)^{-1} \right) \int_0^t e^{A^* \sigma} C^* f(\sigma) \, d\sigma + e^{At} P(t)^{-1} \frac{d}{dt} \int_0^t e^{A^* \sigma} C^* f(\sigma) \, d\sigma \\ &= A\widehat{x} + e^{At} P(t)^{-1} e^{A^* t} C^* f(t) - e^{At} P^{-1} \dot{P} P^{-1} \int_0^t e^{A^* \sigma} C^* f(\sigma) \, d\sigma \\ &= A\widehat{x} + Q(t) C^* f(t) - e^{At} P^{-1} e^{A^* t} C^* C e^{At} P^{-1} \int_0^t e^{A^* \sigma} C^* f(\sigma) \, d\sigma \\ &= A\widehat{x} + QC^* f(t) - Q(t) C^* C \widehat{x}(t) = A\widehat{x} + QC^* \left(f(t) - C\widehat{x}(t) \right) . \end{aligned}$$

This yields the state equation for $\widehat{x}$ in (4.10) and completes the proof.

1.4.1 The infinite horizon case

In this section we will study the infinite horizon observability optimization problem, that is, the observability optimization problem when $t_1 = \infty$. Recall that the system $\dot{x} = Ax$ is stable if all the eigenvalues of A are contained in the open left half plane $\{s : \Re s < 0\}$.

Now consider the pair $\{C, A\}$ where A is stable. Then the observability Gramian

$$P = \int_0^\infty e^{A^* t} C^* C e^{At} \, dt \tag{4.14}$$

is a well defined positive operator on $\mathbb{C}^n$. Moreover, in this case, P is the unique solution to the Lyapunov equation

$$A^* P + PA + C^* C = 0 . \tag{4.15}$$

To see this set $P = \int_0^\infty e^{A^*t}C^*Ce^{At}\,dt$. Then notice that

$$\begin{aligned} -C^*C &= e^{A^*t}C^*Ce^{At}\big|_0^\infty = \int_0^\infty \frac{d}{dt}\, e^{A^*t}C^*Ce^{At}\,dt \\ &= \int_0^\infty A^*e^{A^*t}C^*Ce^{At}\,dt + \int_0^\infty e^{A^*t}C^*Ce^{At}A\,dt = A^*P + PA\,. \end{aligned}$$

Hence (4.15) holds. To show that the solution to the Lyapunov equation in (4.15) is unique, assume that P is any solution to $A^*P + PA + C^*C = 0$. Then we have

$$\begin{aligned} P &= -e^{A^*t}Pe^{At}\big|_0^\infty = -\int_0^\infty \frac{d}{dt}\, e^{A^*t}Pe^{At}\,dt \\ &= -\int_0^\infty \left(A^*e^{A^*t}Pe^{At} + e^{A^*t}Pe^{At}A\right)\,dt \\ &= -\int_0^\infty e^{A^*t}\left(A^*P + PA\right)e^{At}dt = \int_0^\infty e^{A^*t}C^*Ce^{At}dt\,. \end{aligned}$$

Hence $P = \int_0^\infty e^{A^*t}C^*Ce^{At}dt$. Therefore the solution to the Lyapunov equation in (4.15) is unique. Finally, it is noted that the pair $\{C, A\}$ is observable if and only if P is strictly positive.

To introduce the infinite horizon observability problem, assume that the pair $\{C, A\}$ is stable. Let f be any function in $L^2([0,\infty), \mathbb{C}^m)$. Then the observability optimization problem is to find an optimal initial condition $\widehat{x}_0$ such that $y(t) = Ce^{At}\widehat{x}_0$ comes as close as possible to the specified function f. In other words, find an initial condition $\widehat{x}_0$ such that

$$d^2 = \int_0^\infty \|f(t) - Ce^{At}\widehat{x}_0\|^2\,dt = \inf\{\int_0^\infty \|f(t) - Ce^{At}x_0\|^2\,dt : x_0 \in \mathbb{C}^n\}\,. \tag{4.16}$$

Here d^2 is the error in this optimization problem. By letting $t_1 = \infty$ in Theorem 1.4.1, we readily obtain the following result.

THEOREM 1.4.3 *Consider the observability optimization problem in (4.16) where $\{C, A\}$ is a stable pair and f is a specified function in $L^2([0,\infty), \mathbb{C}^m)$. Let P be the solution to the Lyapunov equation in (4.15). Then the following holds.*

(i) There exists a solution $\widehat{x}_0$ in $\mathbb{C}^n$ to the linear equation

$$P\widehat{x}_0 = \int_0^\infty e^{A^*t}C^*f(t)\,dt\,. \tag{4.17}$$

(ii) If $\widehat{x}_0$ in $\mathbb{C}^n$ is any solution to (4.17), then $\widehat{x}_0$ is an initial condition solving the observability optimization problem in (4.16). In this case, the error

$$d^2 = \int_0^\infty \|f(t)\|^2\,dt - (P\widehat{x}_0, \widehat{x}_0)\,. \tag{4.18}$$

(iii) If the pair $\{C, A\}$ is observable, then there exists a unique solution to the observability optimization problem in (4.16) and this unique solution is given by

$$\widehat{x}_0 = P^{-1}\int_0^\infty e^{A^*t}C^*f(t)\,dt\,. \tag{4.19}$$

1.4.2 Exercise

Problem 1. Consider the system $\dot{x} = Ax$ and $y = Cx$ where

$$A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \qquad \text{and} \qquad C = \begin{bmatrix} 1 & 2 \end{bmatrix}.$$

Is the pair $\{C, A\}$ observable? Find the optimal initial condition $\widehat{x}_0$ and the error d in the observability optimization problem

$$d^2 = \inf\{\int_0^\infty |f(t) - Ce^{At}x_0|^2\, dt : x_0 \in \mathbb{C}^2\}$$

where $f = e^{-t}$. Is your choice of $\widehat{x}_0$ unique? Finally, compute the error d.

Problem 2. Repeat Problem 1 with $f = e^{-3t}$.

Problem 3. Consider the system $\dot{x} = Ax$ and $y = Cx$ where

$$A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \qquad \text{and} \qquad C = \begin{bmatrix} 2 & 1 \end{bmatrix}.$$

Is the pair $\{C, A\}$ observable? Find all optimal initial condition $\widehat{x}_0$ and the error d in the observability optimization problem

$$d^2 = \inf\{\int_0^\infty |f(t) - Ce^{At}x_0|^2\, dt : x_0 \in \mathbb{C}^2\}$$

where $f = e^{-t}$. Is your choice of $\widehat{x}_0$ unique? Finally, compute the error d.

Problem 4. Repeat Problem 3 with $f = e^{-3t}$.

1.5 A least squares optimization problem

Let T be a finite rank operator mapping $\mathcal{U}$ into $\mathcal{Y}$ where $\mathcal{U}$ and $\mathcal{Y}$ are Hilbert spaces. Let y be a vector in $\mathcal{Y}$. The equation $y = Tu$ has a solution if and only if y is in the range of T. If $y = Tu$ does not have a solution, then it makes sense to look for a vector $\widehat{u}$ in $\mathcal{U}$ such that $T\widehat{u}$ is closer to y than any other element in $T\mathcal{U}$, that is, find a vector $\widehat{u}$ in $\mathcal{U}$ which makes $\|y - Tu\|$ as small as possible. This naturally leads to the optimization problem $\inf \|y - T\mathcal{U}\|$. To solve this problem, let $\mathcal{R}$ be the range of T. According to the projection theorem, the distance from y to $\mathcal{R}$ is given by

$$d(y, \mathcal{R}) = \inf\{\|y - Tu\| : u \in \mathcal{U}\}. \tag{5.1}$$

Let $P_{\mathcal{R}}$ be the orthogonal projection onto $\mathcal{R}$ and set $\widehat{y} = P_{\mathcal{R}}y$. Then $d(y, \mathcal{R}) = \|y - \widehat{y}\|$ and

$$\|y - \widehat{y}\| = \inf\{\|y - Tu\| : u \in \mathcal{U}\}. \tag{5.2}$$

Since the rank of T is finite, there exists a vector $\widehat{u}$ in $\mathcal{U}$ such that $\widehat{y} = T\widehat{u}$. In this case, $d(y, \mathcal{R}) = \|y - T\widehat{u}\|$ and $T\widehat{u}$ is the unique vector in $T\mathcal{U}$ which is closer to y than any other element in $T\mathcal{U}$. Furthermore, the vector $\widehat{y} = T\widehat{u}$ achieves the minimum in the optimization problems (5.1) and (5.2). Finally, it is noted that the vector $\widehat{u}$ is not necessarily unique. The vector $\widehat{u}$ is unique if and only if the kernel of T is zero.

Recall that the adjoint T^* is the operator mapping $\mathcal{Y}$ into $\mathcal{U}$ determined by $(Tu, g) = (u, T^*g)$ for all u in $\mathcal{U}$ and y in $\mathcal{Y}$. For example, if T is a matrix mapping $\mathbb{C}^m$ into $\mathbb{C}^n$, then T^* is the complex conjugate transpose of T. We claim that T^*T is positive operator on $\mathcal{U}$. To see this observe that $(T^*Tu, u) = (Tu, Tu) = \|Tu\|^2 \geq 0$ for all u in $\mathcal{U}$. Hence T^*T is positive. Moreover, the kernel of T is zero if and only if T^*T is invertible. To verify this notice that $(T^*Tu, u) = 0$ if and only if $\|Tu\|^2 = 0$. Therefore T^*T is strictly positive if and only if T is one to one.

Now let us use the projection theorem to solve the optimization problem in (5.2). As before, assume the $P_{\mathcal{R}}y = \widehat{y} = T\widehat{u}$. According to the projection theorem, $y - \widehat{y}$ is orthogonal to $\mathcal{R}$, or equivalently, $y - T\widehat{u}$ is orthogonal to $T\mathcal{U}$. By taking the adjoint, we see that $T^*y - T^*T\widehat{u}$ is orthogonal to the whole space $\mathcal{U}$. Hence $T^*y - T^*T\widehat{u}$ must be zero. In other words, $\widehat{u}$ is any vector in $\mathcal{U}$ which solves the equation $T^*y = T^*T\widehat{u}$. The projection theorem guarantees that this equation has a solution. So $\widehat{u} = (T^*T)^{-r}T^*y$ where A^{-r} is the pseudo inverse of A. In particular, if T is one to one, then T^*T is invertible and $\widehat{u} = (T^*T)^{-1}T^*y$. Finally, notice that $P_{\mathcal{R}} = T\widehat{u}$ along with the error formula in (2.3), we obtain

$$d(y, \mathcal{R})^2 = \|y\|^2 - \|P_{\mathcal{R}}y\|^2 = \|y\|^2 - \|T\widehat{u}\|^2 = \|y\|^2 - (T\widehat{u}, T\widehat{u}) = \|y\|^2 - (T^*T\widehat{u}, \widehat{u})\,.$$

In other words, $d(y, \mathcal{R})^2 = \|y\|^2 - (T^*T\widehat{u}, \widehat{u})$. Summing up this analysis readily yields the following result.

THEOREM 1.5.1 *Let T be a finite rank operator from $\mathcal{U}$ into $\mathcal{Y}$ with range $\mathcal{R}$. Consider the optimization problem*

$$d = \|y - T\widehat{u}\| = \inf\{\|y - Tu\| : u \in \mathcal{U}\}\ . \tag{5.3}$$

Then the following results hold.

(i) There exists a solution to the equation $T^*T\widehat{u} = T^*y$.

(ii) If $\widehat{u}$ is any solution to $T^*T\widehat{u} = T^*y$, then $\widehat{u}$ is an optimal solution to the optimization problem in (5.3). Moreover, the cost $d^2 = \|y\|^2 - (T^*T\widehat{u}, \widehat{u})$.

(iii) If T is one to one, then the optimal solution $\widehat{u}$ is uniquely determined and given by $\widehat{u} = (T^*T)^{-1}T^*y$.

1.5.1 An application to curve fitting

In this section, we will use Theorem 1.5.1 to solve a classical least squares polynomial fit problem. To this end, let $\deg p$ denote the degree of a polynomial p. Now let $\{\lambda_1, \lambda_2, \cdots, \lambda_m\}$ be a set of distinct complex numbers and $\{y_1, y_2, \cdots, y_m\}$ be a set of complex numbers. Our

problem is to find a polynomial $\widehat{p}$ of a complex variable λ of degree at most $n-1$ to solve the following classical polynomial curve fitting problem:

$$d^2 := \inf \left\{ \sum_{i=1}^{m} |y_i - p(\lambda_i)|^2 : p \text{ is a polynomial and } \deg p \leq n-1 \right\}. \tag{5.4}$$

Here d is the error for the polynomial fit. Moreover, we say that $\widehat{p}$ is a solution to this curve fitting problem if $\widehat{p}$ is a polynomial of degree at most $n-1$ and

$$d^2 = \sum_{i=1}^{m} |y_i - \widehat{p}(\lambda_i)|^2 . \tag{5.5}$$

Without loss of generality, we assume that $n < m$. If $n \geq m$, then we can find a polynomial p of degree at most $m-1 \leq n-1$ such that $p(\lambda_i) = y_i$ for all $i = 1, 2, \cdots, m$. In fact, one such polynomial is obtained by the classical Lagrange interpolation formula

$$p(\lambda) = \sum_{i=1}^{m} y_i p_i(\lambda) \qquad \text{where} \qquad p_i(\lambda) = \prod_{\substack{j=1 \\ i \neq j}}^{m} \frac{(\lambda - \lambda_j)}{(\lambda_i - \lambda_j)} . \tag{5.6}$$

Notice that $p_i(\lambda_i) = 1$ and $p_i(\lambda_j) = 0$ for $j \neq i$. Using this it follows that $p(\lambda_i) = y_i$ for $i = 1, 2, \cdots, m$. Moreover, this is the only polynomial of degree at most $m-1$ satisfying the interpolation conditions $p(\lambda_i) = y_i$ for $i = 1, 2, \cdots, m$. To see this, assume that q is another polynomial of degree of at most $m-1$ satisfying $q(\lambda_i) = y_i$ for $i = 1, 2, \cdots, m$. Then $p - q$ is a polynomial of degree at most $m-1$ with m roots $\{\lambda_i\}_1^m$. Hence $p - q$ must be zero. This proves our claim.

To solve the optimization problem in (5.4) recall that a Vandermonde matrix is a matrix of the form:

$$V = \begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{n-1} \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{n-1} \\ 1 & \lambda_3 & \lambda_3^2 & \cdots & \lambda_3^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \lambda_m & \lambda_m^2 & \cdots & \lambda_m^{n-1} \end{bmatrix} \tag{5.7}$$

where $\{\lambda_1, \lambda_2, \cdots, \lambda_m\}$ are complex numbers. Notice that, if p is a polynomial of the form $p(\lambda) = \sum_{i=0}^{n-1} \alpha_i \lambda^i$, then V can be used to evaluate p at the points $\{\lambda_1, \lambda_2, \cdots, \lambda_m\}$ in the following fashion:

$$V \begin{bmatrix} \alpha_o \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{bmatrix} = \begin{bmatrix} p(\lambda_1) \\ p(\lambda_2) \\ \vdots \\ p(\lambda_m) \end{bmatrix} . \tag{5.8}$$

We claim that if the $\{\lambda_i\}_1^m$ are distinct and $n \leq m$, then V is one to one. If $V\alpha = 0$ for some α in $\mathbb{C}^n$, then (5.8) shows that $p(\lambda_j) = 0$ for $j = 1, 2, \cdots, m$. However, p is a polynomial of degree at most $n-1$, with m roots. Since $n - 1 < m$, it follows that $p(\lambda) = 0$ for all λ, and thus, $\alpha = 0$. Hence V is one to one when $n \leq m$.

If the $\{\lambda_i\}_1^m$ are distinct and $n = m$, then the Vandermonde matrix is invertible. In this case, (5.8) shows that the unique polynomial of degree at most $m-1$ satisfying the interpolation conditions $p(\lambda_i) = y_i$ for $i = 1, 2, \cdots, m$ is given by

$$p(\lambda) = \sum_{i=0}^{m-1} \alpha_i \lambda^i \quad \text{where} \quad \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_{m-1} \end{bmatrix}^{tr} = V^{-1} \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix}^{tr}.$$

This is precisely the polynomial obtained by the Lagrange interpolation formula (5.6).

As before, consider the polynomial interpolation problem in (5.4) where $n < m$. Using (5.7) and (5.8), it follows that this interpolation problem is equivalent to the following least squares optimization problem:

$$\|y - V\widehat{\alpha}\| = d = \inf\{\|y - V\alpha\| : \alpha \in \mathbb{C}^n\} \tag{5.9}$$

where $y = \begin{bmatrix} y_1 & y_2 & \dots & y_m \end{bmatrix}^{tr}$ and $\|\cdot\|$ is the standard norm on $\mathbb{C}^m$. If $\widehat{\alpha}$ is the solution to this optimization problem, then the corresponding optimal polynomial given by

$$\widehat{p}(\lambda) = \sum_{i=0}^{n-1} \widehat{\alpha}_i \lambda^i \qquad \text{where} \qquad \widehat{\alpha} = \begin{bmatrix} \widehat{\alpha}_0 & \widehat{\alpha}_1 & \cdots \widehat{\alpha}_{n-1} \end{bmatrix}^{tr}$$

is the solution to the least squares optimization in (5.9). According to Theorem 1.5.1, the optimal solution $\widehat{\alpha}$ is unique and is given by $\widehat{\alpha} = (V^*V)^{-1}V^*y$. Therefore the polynomial $\widehat{p}$ which solves the least squares problem in (5.4) is given by

$$\widehat{p}(\lambda) = \sum_{i=0}^{n-1} \widehat{\alpha}_i \lambda^i = \begin{bmatrix} 1 & \lambda & \dots & \lambda^{n-1} \end{bmatrix} (V^*V)^{-1}V^*y\,.$$

The error d is computed by

$$d^2 = \|y\|^2 - ((V^*V)^{-1}V^*y,\, V^*y)\,.$$

Moreover, $d = d(y, \operatorname{ran} V)$, where $d(y, \operatorname{ran} V)$ is the distance from y to the range of V.

REMARK 1.5.2 *The above analysis shows that a square Vandermonde matrix generated by the scalars $\{\lambda_i\}_1^m$ is nonsingular if and only if $\{\lambda_i\}_1^m$ are distinct.*

Chapter 2

Least Squares

In this chapter we develop some standard results concerning the least squares estimation problem for random vectors.

2.1 Random vectors

Let $\mathcal{K}$ be the Hilbert space generated by the set of all random variables g such that $E|g|^2$ is finite. Throughout we always assume that all of our random variables are vectors in $\mathcal{K}$. We say that f is a *random vector* with values in $\mathbb{C}^k$ if f is a vector of the form $f = [f_1, f_2, \cdots, f_k]^{tr}$ where $\{f_j\}_1^k$ are all random variables. (Recall that tr denotes the transpose.) In this case, Ef is the vector in $\mathbb{C}^k$ defined by $Ef = [Ef_1, Ef_2, \cdots, Ef_k]^{tr}$. The correlation matrix R_f is the matrix on $\mathbb{C}^k$ defined by $R_f = Eff^*$. To be precise,

$$R_f = Eff^* = \begin{bmatrix} Ef_1\overline{f}_1 & Ef_1\overline{f}_2 & \cdots & Ef_1\overline{f}_k \\ Ef_2\overline{f}_1 & Ef_2\overline{f}_2 & \cdots & Ef_2\overline{f}_k \\ \vdots & \vdots & \ddots & \vdots \\ Ef_k\overline{f}_1 & Ef_k\overline{f}_2 & \cdots & Ef_k\overline{f}_k \end{bmatrix}. \tag{1.1}$$

Notice that the m-n entry of R_f is given by $Ef_m\overline{f}_n$. The following result shows that R_f is positive.

THEOREM 2.1.1 *Let $f = [f_1, f_2, \cdots, f_k]^{tr}$ be a random vector with values in $\mathbb{C}^k$. Then R_f is a positive matrix on $\mathbb{C}^k$. Moreover, R_f is a strictly positive if and only if the random variables $\{f_j\}_1^k$ are linearly independent.*

PROOF. Let $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_k]^{tr}$ be any vector in $\mathbb{C}^k$. Then

$$(R_f\alpha, \alpha) = (Eff^*\alpha, \alpha) = E(ff^*\alpha, \alpha) = E\|f^*\alpha\|^2 = E\left|\sum_{j=1}^k \overline{f}_j\alpha_j\right|^2 \geq 0\,. \tag{1.2}$$

Hence $(R_f\alpha, \alpha) \geq 0$ for all α in $\mathbb{C}^k$. Therefore R_f is positive.

Equation (1.2) shows that

$$(R_f\alpha,\alpha) = E\left|\sum_{j=1}^{k}\overline{f}_j\alpha_j\right|^2 .$$

Recall that if g is a random variable, then $E|g|^2 = 0$ if and only if $g = 0$. So $(R_f\alpha,\alpha) = 0$ if and only if $\sum_1^k f_j\overline{\alpha}_j = 0$. By definition the positive matrix R_f is strictly positive if $\alpha = 0$ is the only solution to $(R_f\alpha,\alpha) = 0$. On the other hand, $\{f_j\}_1^k$ are linearly independent if the only solution to $\sum_1^k f_j\beta_j = 0$ is $\beta_j = 0$ for all integers $1 \le j \le k$. Therefore R_f is strictly positive if and only if $\{f_j\}_1^k$ are linearly independent. This completes the proof.

2.2 Least squares estimation of a random vector

Recall that $\mathcal{K}$ is the Hilbert space generated by the set of all random variables g such that $\|g\|^2 = E|g|^2$ is finite. Let $f = [f_1, f_2, \cdots, f_k]^{tr}$ be a random vector with values in $\mathbb{C}^k$, that is, $\{f_j\}_1^k$ are all random variables in $\mathcal{K}$. Let $\mathcal{H}$ be a subspace in $\mathcal{K}$ and $P_{\mathcal{H}}$ be the orthogonal projection onto $\mathcal{H}$. Then the orthogonal projection of f onto $\mathcal{H}$ is the random vector $P_{\mathcal{H}}f$ with values in $\mathbb{C}^k$ defined by

$$P_{\mathcal{H}}f = \begin{bmatrix} P_{\mathcal{H}}f_1 \\ P_{\mathcal{H}}f_2 \\ \vdots \\ P_{\mathcal{H}}f_k \end{bmatrix} . \tag{2.1}$$

In this section we will present the classical least squares method to compute $P_{\mathcal{H}}f$ when $\mathcal{H}$ is a finite dimensional subspace.

As before, let $f = [f_1, f_2, \cdots, f_k]^{tr}$ be a random vector with values in $\mathbb{C}^k$. Recall that R_f is the positive matrix on $\mathbb{C}^k$ defined by $R_f = Eff^*$. Moreover, R_f is strictly positive if and only if the vectors $\{f_j\}_1^k$ are linearly independent. Let $g = [g_1, g_2, \cdots, g_m]^{tr}$ a random vector with values in $\mathbb{C}^m$. Obviously, R_g is a positive matrix on $\mathbb{C}^m$. We say that $\mathcal{H}$ is the subspace generated by g if $\mathcal{H}$ equals the linear span of $\{g_j\}_1^m$. Let us use the notation $\bigvee g$ to denote the linear span of $\{g_j\}_1^m$. Clearly, $\mathcal{H}$ is a subspace of $\mathcal{K}$. Recall that R_{fg} is the matrix from $\mathbb{C}^m$ into $\mathbb{C}^k$ defined by $R_{fg} = Efg^*$. The j-ν entry of R_{fg} is $Ef_j\overline{g}_\nu$. By taking the adjoint, it follows that $R_{fg}^* = R_{gf}$. Notice that the subspace $\bigvee f$ generated by f is orthogonal to $\mathcal{H}$ if and only if $Ef_j\overline{g}_\nu = 0$ for all $j = 1, 2, \cdots, k$ and $\nu = 1, 2, \cdots, m$. In other words, $\bigvee f$ is orthogonal to $\bigvee g$ if and only if $R_{fg} = 0$. Motivated by this we say that the random vectors f and g are orthogonal if $Efg^* = 0$. The following classical result provides a formula for computing $P_{\mathcal{H}}f$.

THEOREM 2.2.1 *Let f be a random vector with values in $\mathbb{C}^k$, and g be a random variable with values in $\mathbb{C}^m$. Let $P_{\mathcal{H}}$ be the orthogonal projection onto the subspace $\mathcal{H}$ generated by g, set $\widehat{f} = P_{\mathcal{H}}f$ and the error $\widetilde{f} = f - \widehat{f}$. Then the following holds.*

(i) If M is any matrix from $\mathbb{C}^m$ into $\mathbb{C}^k$, then

$$R_{\widetilde{f}} = E(f-\widehat{f})(f-\widehat{f})^* \le E(f-Mg)(f-Mg)^* . \tag{2.2}$$

(ii) The equality $R_{\widetilde{f}} = E(f - Mg)(f - Mg)^*$ *holds if and only if* $Mg = \widehat{f}$*, or equivalently,* $R_{fg} = MR_g$*. In this case, the estimation error is given by*

$$E(f - \widehat{f})(f - \widehat{f})^* = R_f - MR_gM^* = R_f - R_{fg}M^* . \tag{2.3}$$

(iii) If R_g *is invertible, or equivalently, the set* $\{g_j\}_1^m$ *is linearly independent, then*

$$P_{\mathcal{H}}f = R_{fg}R_g^{-1}g . \tag{2.4}$$

(iv) If R_g *is invertible, then the covariance for the error* $\widetilde{f} = f - \widehat{f}$ *is given by*

$$R_{\widetilde{f}} = E(f - \widehat{f})(f - \widehat{f})^* = R_f - E\widehat{f}\widehat{f}^* = R_f - R_{fg}R_g^{-1}R_{gf} . \tag{2.5}$$

Proof. Let M be any matrix mapping $\mathbb{C}^m$ into $\mathbb{C}^k$. By employing the projection theorem, we obtain

$$\begin{aligned} E(f - Mg)(f - Mg)^* &= E\left(f - \widehat{f} + \widehat{f} - Mg\right)\left(f - \widehat{f} + \widehat{f} - Mg\right)^* \\ &= E(f - \widehat{f})(f - \widehat{f})^* + E(f - \widehat{f})(\widehat{f} - Mg)^* \\ &\quad + E(\widehat{f} - Mg)(f - \widehat{f})^* + E(\widehat{f} - Mg)(\widehat{f} - Mg)^* \\ &= E(f - \widehat{f})(f - \widehat{f})^* + E(\widehat{f} - Mg)(\widehat{f} - Mg)^* \\ &= R_{\widetilde{f}} + E(\widehat{f} - Mg)(\widehat{f} - Mg)^* \geq R_{\widetilde{f}} . \end{aligned}$$

The third equality follows from the fact that $f - \widehat{f}$ is orthogonal to $\mathcal{H}$, that is, $E(f - \widehat{f})h^* = 0$ for all h in $\mathcal{H}$. This readily yields the inequality in (2.2).

To verify Part (ii) notice that the previous calculation also shows that

$$E(f - Mg)(f - Mg)^* = R_{\widetilde{f}} + E(\widehat{f} - Mg)(\widehat{f} - Mg)^* .$$

So we have $E(f - Mg)(f - Mg)^* = R_{\widetilde{f}}$ if and only if $E(\widehat{f} - Mg)(\widehat{f} - Mg)^*$ is zero, or equivalently, $\widehat{f} = Mg$. By the projection theorem, $\widehat{f} = Mg$ if and only if $f - Mg$ is orthogonal to $\mathcal{H}$, or equivalently, $E(f - Mg)g^* = 0$. Therefore $E(f - Mg)(f - Mg)^* = R_{\widetilde{f}}$ if and only if $R_{fg} = MR_g$.

Now let us establish the error formulas in (2.3). By the projection theorem $\widehat{f}$ is orthogonal to $\widetilde{f} = f - \widehat{f}$. Since $f = \widehat{f} + \widetilde{f}$, we obtain

$$R_f = Eff^* = E(\widehat{f} + \widetilde{f})(\widehat{f} + \widetilde{f})^* = E\widehat{f}\widehat{f}^* + E\widetilde{f}\widetilde{f}^* = E\widehat{f}\widehat{f}^* + R_{\widetilde{f}} .$$

Hence $R_f = E\widehat{f}\widehat{f}^* + R_{\widetilde{f}}$. Using $\widehat{f} = Mg$ and $R_{fg} = MR_g$, we have

$$R_{\widetilde{f}} = R_f - E\widehat{f}\widehat{f}^* = R_f - MEgg^*M^* = R_f - MR_gM^* = R_f - R_{fg}M^* .$$

Therefore (2.3) holds.

Clearly, Part (*iii*) follows by inverting R_g in Part (*ii*). Let us directly prove Part (*iii*) from the projection theorem. Since $\mathcal{H}$ equals the span of $\{g_j\}_1^m$, equation (2.1) shows that $\widehat{f} = P_{\mathcal{H}} f = Ag$ where A is a matrix from $\mathbb{C}^m$ into $\mathbb{C}^k$. According to the projection theorem, $f - Ag$ is orthogonal to g, that is, the subspace generated by $f - Ag$ is orthogonal to the subspace generated by g. In other words, the matrix $E(f - Ag)g^* = 0$. Hence $R_{fg} = AEgg^* = AR_g$. This readily implies that $A = R_{fg}R_g^{-1}$. Therefore $\widehat{f} = Ag = R_{fg}R_g^{-1}g$ is given by (2.4).

To complete the proof it remains to establish the error formulas in (2.5). By employing $M = R_{fg}R_g^{-1}$ and $R_{gf} = R_{fg}^*$ in (2.3), we obtain

$$R_{\tilde{f}} = R_f - MR_gM^* = R_f - R_{fg}R_g^{-1}R_gR_g^{-1}R_{gf} = R_f - R_{fg}R_g^{-1}R_{gf}\,.$$

This yields (2.5) and completes the proof.

REMARK 2.2.2 *Let f be a random vector with values in $\mathbb{C}^k$, and g be a random vector with values in $\mathbb{C}^m$. Let $P_{\mathcal{H}}$ be the orthogonal projection onto the subspace $\mathcal{H}$ generated by g and set $\widehat{f} = P_{\mathcal{H}} f$. Let y be a random vector in $\mathbb{C}^\nu$ and let h be the random vector in $\mathbb{C}^{m+\nu}$ defined by $h = \left[\begin{array}{cc} g & y \end{array}\right]^{tr}$ where tr denotes the transpose. Let $P_{\mathcal{K}}$ be the orthogonal projection onto the subspace $\mathcal{K}$ generated by h, that is, $\mathcal{K} = g \bigvee h$. Set $\widehat{f}_1 = P_{\mathcal{K}} f$. Notice that $\mathcal{H}$ is a subspace of $\mathcal{K}$. Therefore $\widehat{f}_1$ is a better estimate for f than $\widehat{f}$. To be precise, we have*

$$E(f - \widehat{f}_1)(f - \widehat{f}_1)^* \leq E(f - \widehat{f})(f - \widehat{f})^*\,. \tag{2.6}$$

Furthermore, $E(f - \widehat{f}_1)(f - \widehat{f}_1)^* = E(f - \widehat{f})(f - \widehat{f})^*$ if and only if $\widehat{f} = \widehat{f}_1$.

To verify that (2.6) holds, let M be the matrix mapping $\mathbb{C}^m$ into $\mathbb{C}^k$ satisfying $Mg = P_{\mathcal{H}} f$. Let N be the block matrix mapping $\mathbb{C}^{m+\nu}$ into $\mathbb{C}^k$ defined by

$$N = \left[\begin{array}{cc} M & 0 \end{array}\right] : \left[\begin{array}{c} \mathbb{C}^m \\ \mathbb{C}^\nu \end{array}\right] \to \mathbb{C}^k\,.$$

Notice that $Nh = Mg$. By applying Theorem 2.2.1 to $\widehat{f}_1 = P_{\mathcal{K}} f$ we have

$$\begin{aligned} E(f - \widehat{f}_1)(f - \widehat{f}_1)^* &\leq E(f - Nh)(f - Nh)^* \\ &= E(f - Mg)(f - Mg)^* = E(f - \widehat{f})(f - \widehat{f})^*\,. \end{aligned} \tag{2.7}$$

Therefore (2.6) holds. Now assume that we have equality in (2.6). Then we have equality in (2.7). In particular, $E(f - \widehat{f}_1)(f - \widehat{f}_1)^* = E(f - Nh)(f - Nh)$. By consulting Part (ii) of Theorem 2.2.1, we see that $\widehat{f}_1 = Nh = Mg = P_{\mathcal{H}} f = \widehat{f}$. Therefore we have equality in (2.6) if and only if $\widehat{f}_1 = \widehat{f}$.

2.2.1 An example with additive noise

Recall that a random variable $\mathbf{z}$ is uniform over the interval $[a, b]$ if its density function $f_{\mathbf{z}}$ is constant over $[a, b]$ and zero otherwise, that is,

$$\begin{aligned} f_{\mathbf{z}}(z) &= \frac{1}{b-a} \qquad \text{if } a \leq z \leq b \\ &= 0 \qquad\qquad \text{otherwise}\,. \end{aligned}$$

In this section we will use a boldface $\mathbf{z}$ to represent the random variable $\mathbf{z}$ while a lower case z is just a point on the real line. The mean of the uniform random variable $\mathbf{z}$ is given by $E\mathbf{z} = (a+b)/2$. This follows from

$$E\mathbf{z} = \int_{-\infty}^{\infty} z f_{\mathbf{z}}(z)\, dz = \frac{1}{b-a}\int_a^b z\, dz = \frac{b+a}{2}\,. \tag{2.8}$$

Moreover, $E\mathbf{z}^2$ is given by

$$E\mathbf{z}^2 = \frac{b^3-a^3}{3(b-a)}\,. \tag{2.9}$$

To see this observe that

$$E\mathbf{z}^2 = \int_{-\infty}^{\infty} z^2 f_{\mathbf{z}}(z)\, dz = \frac{1}{b-a}\int_a^b z^2\, dx = \frac{b^3-a^3}{3(b-a)}\,.$$

Therefore equation (2.9) holds.

Now assume that $\mathbf{x}$ is a uniform random variable over $[0, 10]$ and $\mathbf{v}$ is a uniform random variable over $[0, 4]$. In this case, equations (2.8) and (2.9) imply that

$$E\mathbf{x} = 5, \qquad E\mathbf{x}^2 = 100/3, \qquad E\mathbf{v} = 2, \qquad \text{and} \qquad E\mathbf{v}^2 = 16/3\,. \tag{2.10}$$

Moreover, assume $\mathbf{x}$ and $\mathbf{v}$ are independent random variables. Now let $\mathbf{y}$ be the random variable defined by $\mathbf{y} = \mathbf{x} + \mathbf{v}$. For example, assume that $\mathbf{x}$ is a grade on an exam and the one arbitrarily adds $\mathbf{v}$ to the score. The recorded score is $\mathbf{y} = \mathbf{x} + \mathbf{v}$. Notice that the score corresponding to $\mathbf{y}$ is a number between 0 and 14. The problem is to try to estimate the original score $\mathbf{x}$ from the knowledge of $\mathbf{y}$. Of course, the conditional expectation $E(\mathbf{x}|\mathbf{y})$ is the best estimate.

Now let us try to estimate $\mathbf{x}$ using the one dimensional space spanned by $\mathbf{y}$. To be precise, let $\mathcal{H}_1$ be the one dimensional space determined by the linear span of $\mathbf{y}$. Then let us compute $\widehat{\mathbf{x}} = P_{\mathcal{H}_1}\mathbf{x}$ where $P_{\mathcal{H}_1}$ is the orthogonal projection onto $\mathcal{H}_1$. According to Theorem 2.2.1 the orthogonal projection $\widehat{\mathbf{x}} = P_{\mathcal{H}_1}\mathbf{x} = R_{\mathbf{xy}}\mathbf{y}/R_{\mathbf{y}}$. Recall that if $\mathbf{u}$ and $\mathbf{w}$ are independent random variables, then $E\mathbf{uw} = E\mathbf{u}E\mathbf{w}$. Using the fact that $\mathbf{x}$ and $\mathbf{v}$ are independent random variables, we have

$$\begin{aligned} R_{\mathbf{xy}} &= E\mathbf{xy} = E\mathbf{x}(\mathbf{x}+\mathbf{v}) = E\mathbf{x}^2 + E\mathbf{xv} = 100/3 + E\mathbf{x}E\mathbf{v} = 100/3 + 10 = 130/3 \\ R_{\mathbf{y}} &= E\mathbf{y}^2 = E(\mathbf{x}+\mathbf{v})^2 = E\mathbf{x}^2 + 2E\mathbf{x}E\mathbf{v} + E\mathbf{v}^2 = 100/3 + 20 + 16/3 = 176/3\,. \end{aligned}$$

Therefore optimal estimate $\widehat{\mathbf{x}} = P_{\mathcal{H}_1}\mathbf{x} = 130\mathbf{y}/176$ where $\mathcal{H}_1$ is the span of $\mathbf{y}$. Finally, Theorem 2.2.1 shows that the error in this estimation is given by

$$E(\mathbf{x}-\widehat{\mathbf{x}})^2 = R_{\mathbf{x}} - R_{\mathbf{xy}}R_{\mathbf{yx}}/R_{\mathbf{y}} = E\mathbf{x}^2 - R_{\mathbf{xy}}^2/R_{\mathbf{y}} = 100/3 - 130^2/3\times 176 = 175/132\,.$$

So the error $E(\mathbf{x}-\widehat{\mathbf{x}})^2 = 175/132$.

Now let us try to estimate $\mathbf{x}$ using the two dimensional space spanned by $\mathbf{y}$ and the constant vector 1, that is, $\mathcal{H}_2 = \text{span}\{1, \mathbf{y}\}$. In this case, the estimate $\widehat{\mathbf{x}} = P_{\mathcal{H}_2}\mathbf{x}$ where $P_{\mathcal{H}_2}$ is the orthogonal projection onto $\mathcal{H}_2$. Now let $\mathbf{g}$ be the random vector defined by $\mathbf{g} = [1, \mathbf{y}]^{tr}$

where tr denotes the transpose. Clearly, $\mathcal{H}_2$ is the span of $\mathbf{g}$. By applying Theorem 2.2.1 we see that $\widehat{\mathbf{x}} = R_{\mathbf{xg}} R_{\mathbf{g}}^{-1} \mathbf{g}$. Using $E\mathbf{xy} = 130/3$, we obtain

$$R_{\mathbf{xg}} = E\mathbf{x}\mathbf{g}^* = \begin{bmatrix} E\mathbf{x}1 & E\mathbf{xy} \end{bmatrix} = \begin{bmatrix} 5 & 130/3 \end{bmatrix} .$$

Notice that $E\mathbf{y} = E(\mathbf{x}+\mathbf{v}) = E\mathbf{x} + E\mathbf{v} = 7$. Thus $E\mathbf{y} = 7$. By employing $E\mathbf{y}^2 = 176/3$, we have

$$R_{\mathbf{g}} = E\mathbf{g}\mathbf{g}^* = E \begin{bmatrix} 1 \\ \mathbf{y} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{y} \end{bmatrix} = \begin{bmatrix} E1 & E\mathbf{y} \\ E\mathbf{y} & E\mathbf{y}^2 \end{bmatrix} = \begin{bmatrix} 1 & 7 \\ 7 & 176/3 \end{bmatrix} .$$

This readily implies that the optimal estimate is given by

$$\widehat{\mathbf{x}} = R_{\mathbf{xg}} R_{\mathbf{g}}^{-1} \mathbf{g} = \begin{bmatrix} 5 & 130/3 \end{bmatrix} \begin{bmatrix} 1 & 7 \\ 7 & 176/3 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \mathbf{y} \end{bmatrix} = -30/29 + 25\mathbf{y}/29 .$$

In other words, $\widehat{\mathbf{x}} = -30/29 + 25\mathbf{y}/29$. Recall that $R_{\mathbf{gx}} = R_{\mathbf{xg}}^*$. Theorem 2.2.1 shows that the error in this estimation is given by

$$E(\mathbf{x} - \widehat{\mathbf{x}})^2 = R_{\mathbf{x}} - R_{\mathbf{xg}} R_{\mathbf{g}}^{-1} R_{\mathbf{gx}} = 100/3 - \begin{bmatrix} 5 & 130/3 \end{bmatrix} \begin{bmatrix} 1 & 7 \\ 7 & 176/3 \end{bmatrix}^{-1} \begin{bmatrix} 5 \\ 130/3 \end{bmatrix} = \frac{100}{87} .$$

So the error $E(\mathbf{x} - \widehat{\mathbf{x}})^2 = 100/87$. Finally, it is noted that this error is smaller than our previous estimation error 175/132. This follows from the fact that $\mathcal{H}_1$ is a subspace of $\mathcal{H}_2$. Hence the error in estimation is smaller; see Remark 2.2.2.

2.2.2 Exercise

Problem 1. Let $\mathbf{x}$ be a uniform random variable over $[0, 10]$ and $\mathbf{v}$ a uniform random variable over $[0, 4]$. Moreover, assume $\mathbf{x}$ and $\mathbf{v}$ are independent random variables. Now let $\mathbf{y}$ be the random variable defined by $\mathbf{y} = \mathbf{x} + \mathbf{v}$. Let $\mathcal{H}_3$ be the subspace spanned by $\{1, \mathbf{y}, \mathbf{y}^2\}$. Then compute the optimal estimate $\widehat{\mathbf{x}} = P_{\mathcal{H}_3}\mathbf{x}$ and the error in estimation $E(\mathbf{x} - \widehat{\mathbf{x}})^2$. Notice that in this case, $\mathbf{g} = [1, \mathbf{y}, \mathbf{y}^2]^{tr}$. Compute the conditional expectation $E(\mathbf{x}|\mathbf{y})$ and compare your answer $P_{\mathcal{H}_3}\mathbf{x}$ to the solution computed by the conditional expectation.

Hint. According to Lemma 10.3.1 in the Appendix the joint probability density function $f_{\mathbf{x},\mathbf{y}}(x, y) = f_{\mathbf{x}}(x) f_{\mathbf{v}}(y - x)$. (The notation $f_{\mathbf{x}}(x)$ is the density function for the random variable $\mathbf{x}$ evaluated at the point x on the real line.) Moreover, the density $f_{\mathbf{y}}$ for the random variable $\mathbf{y}$ is obtained by convolving $f_{\mathbf{x}}$ with $f_{\mathbf{v}}$. In other words, show that

$$\begin{aligned} f_{\mathbf{y}}(y) &= y/40 && \text{if } 0 \le y \le 4 \\ &= 1/10 && \text{if } 4 \le y \le 10 \\ &= (14 - y)/40 && \text{if } 10 \le y \le 14 . \end{aligned}$$

Verify that the conditional density $f_{\mathbf{x}|\mathbf{y}}$ is given by

$$\begin{aligned} f_{\mathbf{x}|\mathbf{y}}(x|y) &= 1/y && \text{if } 0 \le x \le y \text{ and } 0 \le y \le 4 \\ &= 1/4 && \text{if } y - 4 \le x \le y \text{ and } 4 \le y \le 10 \\ &= 1/(14 - y) && \text{if } y - 4 \le x \le 10 \text{ and } 10 \le y \le 14 . \end{aligned}$$

Finally, show that the conditional expectation is given by

$$\begin{aligned} E(\mathbf{x}|\mathbf{y}=y) &= y/2 && \text{if } 0 \le y \le 4 \\ &= y-2 && \text{if } 4 \le y \le 10 \\ &= (y+6)/2 && \text{if } 10 \le y \le 14\,. \end{aligned}$$

Recall that $P_{\mathcal{H}_1}\mathbf{x} = 130\mathbf{y}/176$ was the best estimate of $\mathbf{x}$ corresponding to the one dimensional subspace $\mathcal{H}_1$ spanned by $\{\mathbf{y}\}$, and $P_{\mathcal{H}_2}\mathbf{x} = -30/29 + 25\mathbf{y}/29$ was the best estimate of $\mathbf{x}$ corresponding to the two dimensional subspace $\mathcal{H}_2$ spanned by $\{1, \mathbf{y}\}$; see Section 2.2.1. Plot $130y/176$ and $-30/29 + 25y/29$ and the conditional expectation $E(\mathbf{x}|\mathbf{y}=y)$ along with your estimate $P_{\mathcal{H}_3}\mathbf{x}$ on the same graph. Finally, comment on the resulting graph.

Problem 2. Let $\mathbf{y}$ be a uniform random variable over $[0,1]$ and $\mathbf{x}$ the random variable defined by $\mathbf{x} = e^{\mathbf{y}}$. Let $\mathcal{H}$ be the subspace spanned by $\{1, \mathbf{y}, \mathbf{y}^2\}$. Then compute the optimal estimate $\widehat{\mathbf{x}} = P_{\mathcal{H}}\mathbf{x}$ and the error in estimation $E(\mathbf{x} - \widehat{\mathbf{x}})^2$. Show that $E(\mathbf{x}|\mathbf{y}=y) = e^y$. Plot your estimate $P_{\mathcal{H}}\mathbf{x}$ and the conditional expectation $E(\mathbf{x}|\mathbf{y}=y) = e^y$ on the same graph, and compare these two estimates.

2.3 An elementary state estimation problem

In this section we will present a simple state estimation problem which is related to some of our observability results studied in Sections 9.6 and 9.8 in Chapter 9. To this end, consider the system

$$x(n+1) = Ax(n) \qquad \text{and} \qquad y(n) = Cx(n) \tag{3.1}$$

where the pair $\{C, A \text{ on } \mathbb{C}^m\}$ is observable. Moreover, let us assume that the initial condition $x(0) = x_0$ is a random vector satisfying $Ex_0x_0^* = I$. An elementary state estimation problem is to compute the best estimate $\widehat{x}(n)$ of $x(n)$ given the past $\{y(k)\}_0^{n-1}$. In other words, let $\mathcal{M}_{n-1}$ be the subspace spanned by $\{y(k)\}_0^{n-1}$. Then find $\widehat{x}(n) = P_{\mathcal{M}_{n-1}}x(n)$ where $P_{\mathcal{M}_{n-1}}$ is the orthogonal projection onto $\mathcal{M}_{n-1}$.

To compute the optimal state estimate $\widehat{x}(n)$, let W_n be the operator matrix and g_n the random vector defined by

$$W_n = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \qquad \text{and} \qquad g_n = \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(n-1) \end{bmatrix}. \tag{3.2}$$

By construction $\mathcal{M}_{n-1}$ equals the span of g_n. Since $y(k) = CA^kx_0$, we see that $g_n = W_nx_0$. According to Part (ii) in Theorem 2.2.1, the optimal estimate $\widehat{x}(n) = Mg_n$ where M is an operator satisfying $R_{x(n)g_n} = MR_{g_n}$. To find M notice that

$$\begin{aligned} R_{g_n} &= Eg_ng_n^* = EW_nx_0(W_nx_0)^* = W_nEx_0x_0^*W_n^* = W_nW_n^* \\ R_{x(n)g_n} &= Ex(n)g_n^* = EA^nx_0(W_nx_0)^* = A^nEx_0x_0^*W_n^* = A^nW_n^*\,. \end{aligned}$$

In other words, $R_{g_n} = W_n W_n^*$ and $R_{x(n)g_n} = A^n W_n^*$. Now assume that the time $n \geq m$. Because the pair $\{C, A \text{ on } \mathbb{C}^m\}$ is observable, $W_n^* W_n$ is invertible. We claim that

$$M = A^n (W_n^* W_n)^{-1} W_n^* .$$

To see this observe that

$$MR_{g_n} = A^n (W_n^* W_n)^{-1} W_n^* W_n W_n^* = A^n W_n^* = R_{x(n)g_n} .$$

Therefore $MR_{g_n} = R_{x(n)g_n}$. In other words, the optimal state estimate is given by

$$\widehat{x}(n) = Mg_n = A^n (W_n^* W_n)^{-1} W_n^* g_n .$$

This is precisely the solution to the observability optimization problem studied in Section 9.8 in Chapter 9. In fact, a recursive solution for computing the optimal state $\widehat{x}(n)$ is given in Theorem 9.8.3.

Now let us compute the error in estimation $E(x(n) - \widehat{x}(n))(x(n) - \widehat{x}(n))^*$ when $n \geq m$. Using $x(n) = A^n x_0$, it follows that $R_{x(n)} = Ex(n)x(n)^* = A^n A^{*n}$. According to Part (ii) of Theorem 2.2.1, the error in estimation is given by

$$E(x(n) - \widehat{x}(n))(x(n) - \widehat{x}(n))^* \;\; = \;\; R_{x(n)} - MR_{g_n} M^* = A^n A^{*n} - A^n A^{*n} = 0 .$$

In other words, the estimation error is zero when $n \geq m$. To see why this result is not surprising, recall that the pair $\{C, A \text{ on } \mathbb{C}^m\}$ is observable. By virtue of Theorem 9.6.2 in Chapter 9, the initial condition $x_0 = (W_n^* W_n)^{-1} W_n^* g_n$. This readily implies that

$$x(n) = A^n x_0 = A^n (W_n^* W_n)^{-1} W_n^* g_n = Mg_n = \widehat{x}(n) .$$

In other words, $x(n) = \widehat{x}(n)$ for all integers $n \geq m$. So obviously, the estimation error $E(x(n) - \widehat{x}(n))(x(n) - \widehat{x}(n))^*$ is zero.

2.4 A matrix inversion lemma

In this section we will present a basic matrix inversion lemma for 2×2 operator matrices. The following matrix inversion lemma will be used in deriving the Kalman filter.

LEMMA 2.4.1 *Let T be an operator matrix mapping $\mathcal{X} \oplus \mathcal{U}$ into $\mathcal{X} \oplus \mathcal{Y}$ of the form*

$$T = \begin{bmatrix} R & M \\ N & Q \end{bmatrix} . \tag{4.1}$$

Assume that R is invertible and set $\Delta = Q - NR^{-1}M$. Then T is invertible if and only if Δ is invertible. Furthermore, in this case,

$$T^{-1} = \begin{bmatrix} R^{-1} + R^{-1}M\Delta^{-1}NR^{-1} & -R^{-1}M\Delta^{-1} \\ -\Delta^{-1}NR^{-1} & \Delta^{-1} \end{bmatrix} . \tag{4.2}$$

The operator $\Delta = Q - NR^{-1}M$ is called the *Schur complement* of T.

PROOF. A simple calculation shows that T admits a factorization of the form

$$T = \begin{bmatrix} I & 0 \\ NR^{-1} & I \end{bmatrix} \begin{bmatrix} R & 0 \\ 0 & Q - NR^{-1}M \end{bmatrix} \begin{bmatrix} I & R^{-1}M \\ 0 & I \end{bmatrix}. \tag{4.3}$$

Because the first and third matrices are invertible, T is invertible if and only if the Schur complement $\Delta = Q - NR^{-1}M$ is invertible.

Now assume that Δ is invertible. Notice that

$$\begin{bmatrix} I & 0 \\ V & I \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ -V & I \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} I & W \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} I & -W \\ 0 & I \end{bmatrix}$$

where V and W are arbitrary operators. Recall that $(FGH)^{-1} = H^{-1}G^{-1}F^{-1}$ where F, G and H are operators acting between the appropriate spaces. By taking the inverses in equation (4.3), we arrive at

$$T^{-1} = \begin{bmatrix} I & -R^{-1}M \\ 0 & I \end{bmatrix} \begin{bmatrix} R^{-1} & 0 \\ 0 & \Delta^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -NR^{-1} & I \end{bmatrix}.$$

This readily yields the formula for the inverse of T in (4.2) and completes the proof.

In many applications the matrix T in (4.1) is a covariance matrix. In this case, T is a positive matrix. Furthermore, R and Q are positive operators and $N = M^*$. This sets the stage for the following result.

LEMMA 2.4.2 *Let T be a self-adjoint operator matrix on $\mathcal{X} \oplus \mathcal{U}$ of the form*

$$T = \begin{bmatrix} R & M \\ M^* & Q \end{bmatrix}. \tag{4.4}$$

Assume that R is a strictly positive operator and let $\Delta = Q - M^*R^{-1}M$ be the Schur complement for T. Then T is positive if and only if Δ is positive. Moreover, T is strictly positive if and only if Δ is strictly positive. Finally, in this case,

$$T^{-1} = \begin{bmatrix} R^{-1} + R^{-1}M\Delta^{-1}M^*R^{-1} & -R^{-1}M\Delta^{-1} \\ -\Delta^{-1}M^*R^{-1} & \Delta^{-1} \end{bmatrix}. \tag{4.5}$$

PROOF. Notice that T in (4.4) is precisely the operator matrix in (4.1) where $N = M^*$. Using $N = M^*$ in (4.3), we see that T admits a factorization of the form

$$T = \begin{bmatrix} I & 0 \\ M^*R^{-1} & I \end{bmatrix} \begin{bmatrix} R & 0 \\ 0 & \Delta \end{bmatrix} \begin{bmatrix} I & R^{-1}M \\ 0 & I \end{bmatrix}. \tag{4.6}$$

This readily shows that $T = U^*\Lambda U$ where T is the diagonal matrix and U is the invertible upper triangular matrix defined by

$$\Lambda = \begin{bmatrix} R & 0 \\ 0 & \Delta \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} I & R^{-1}M \\ 0 & I \end{bmatrix}.$$

So if f is any vector in $\mathcal{X} \oplus \mathcal{U}$, then

$$(Tf, f) = (U^*\Lambda U f, f) = (\Lambda U f, U f)\,.$$

Because U is invertible, this implies that T is positive (respectively strictly positive) if and only if Λ is positive (respectively strictly positive). Since R is strictly positive and Λ is a block diagonal matrix, Λ is positive (respectively strictly positive) if and only if Δ is positive (respectively strictly positive). In other words, T is positive (respectively strictly positive) if and only if its Schur complement Δ is positive (respectively strictly positive). Finally, it is noted that the inverse of T given in (4.5) follows by substituting $N = M^*$ in (4.2). This completes the proof.

2.5 Kalman filtering with no state noise

In this section we will use Theorem 2.2.1 to solve the Kalman filtering problem when there is no additive noise on the state vector. The solution to this problem is motivated by the results on the observability optimization problem in Section 9.8. The general Kalman filtering problem is studied in Chapter 3. We say that $w(n)$ is a *white noise process* if $w(n)$ is a sequence of random vectors in some $\mathbb{C}^m$ space satisfying $Ew(n) = 0$ for all integers n and $Ew(j)w(k)^* = \delta_{jk}I$ where δ_{jk} is the Kronecker delta, that is, $\delta_{jk} = 1$ if $j = k$ and $\delta_{jk} = 0$ if $j \neq k$. In this case, $\{w(n)\}$ is a sequence of mean zero, orthonormal random vectors.

Consider the state space system determined by

$$x(n+1) = Ax(n) \qquad \text{and} \qquad y(n) = Cx(n) + Dv(n)\,. \tag{5.1}$$

Here A is an operator on $\mathcal{X}$ and C maps $\mathcal{X}$ into $\mathcal{Y}$ while D in an operator from $\mathcal{V}$ into $\mathcal{Y}$ where $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{V}$ are all $\mathbb{C}^k$ spaces of the appropriate dimension. The initial condition x_0 is a random vector with values in $\mathcal{X}$. The output disturbance $v(n)$ is a white noise random process. Moreover, we assume that the initial condition x_0 and $v(n)$ are all independent random vectors for all integers $n \geq 0$. In particular, this implies that x_0 and $v(n)$ are orthogonal for all integers n, that is, $Ex_0v(n)^* = 0$ for all n. To see this simply observe that $Ex_0v(n)^* = Ex_0Ev(n)^* = 0$. (Here we used the fact that if f and g are two independent random variables, then $Efg = EfEg$.) Finally, we also assume that the covariance R_{x_0} for the initial state x_0 is known.

The Kalman filtering problem is to compute the best estimate $\widehat{x}(k)$ of the state $x(k)$ given the past output $\{y(j)\}_0^{k-1}$. In other words, find the best estimate of the state $x(k)$ when the output $y(n) = Cx(n) + Dv(n)$ is corrupted by additive white noise. The Kalman filter is an optimal state estimator. To be explicit, let $\mathcal{M}_n$ be the subspace generated by the random vectors $\{y(j)\}_0^n$, that is, $\mathcal{M}_n = \bigvee_{j=0}^n y(j)$. Let $P_{\mathcal{M}_n}$ be the orthogonal projection onto $\mathcal{M}_n$ for all integers $n \geq 0$. Then the best estimate $\widehat{x}(k)$ of the state $x(k)$ is given by the orthogonal projection $\widehat{x}(k) = P_{\mathcal{M}_{k-1}}x(k)$. The following result which is a special case of the Kalman filter provides us with a recursive algorithm to compute $\widehat{x}(k)$.

THEOREM 2.5.1 *Consider the state space system*

$$x(n+1) = Ax(n) \qquad \textit{and} \qquad y(n) = Cx(n) + Dv(n) \tag{5.2}$$

where $v(n)$ is a white noise random processes, and the initial condition $x(0)$ and $v(n)$ are orthogonal random vectors for all integers n. Then the optimal estimate $\widehat{x}(k) = P_{\mathcal{M}_{k-1}} x(k)$ of the state $x(k)$ given the past $\{y(j)\}_0^{k-1}$ is recursively computed by

$$\begin{aligned} \widehat{x}(n+1) &= A\widehat{x}(n) + \Delta_n(y - C\widehat{x}(n)) \qquad (5.3) \\ \Delta_n &= AQ_nC^* \left(CQ_nC^* + DD^*\right)^{-1} . \qquad (5.4) \end{aligned}$$

The state covariance error $Q_k = E(x(k) - \widehat{x}(k))(x(k) - \widehat{x}(k))^*$ is recursively computed by solving the Riccati difference equation

$$Q_{n+1} = AQ_nA^* - AQ_nC^* \left(CQ_nC^* + DD^*\right)^{-1} CQ_nA^* , \qquad (5.5)$$

subject to the initial condition $Q_0 = Ex(0)x(0)^*$.

REMARK 2.5.2 *It is emphasized that when implementing the Kalman filter we always assume that the inverse of $CQ_nC^* + DD^*$ exists for all integers $n \geq 0$. In particular, if DD^* is invertible, then $CQ_nC^* + DD^*$ is strictly positive, and thus, invertible for all n.*

PROOF OF THEOREM 2.5.1. Let g_n be the random vector formed by the first n outputs $\{y(k)\}_0^{n-1}$, that is,

$$g_n = \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(n-1) \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} x_0 + \begin{bmatrix} Dv(0) \\ Dv(1) \\ \vdots \\ Dv(n-1) \end{bmatrix} . \qquad (5.6)$$

Obviously, $\mathcal{M}_{n-1}$ equals the span of g_n. Let Q_n be the state covariance error, that is,

$$Q_n = E(x(n) - \widehat{x}(n))(x(n) - \widehat{x}(n))^* .$$

In a moment we will show that R_{g_n} is invertible. According to Theorem 2.2.1, the optimal state estimate $\widehat{x}(n) = P_{\mathcal{M}_{n-1}} x(n)$ and covariance error Q_n is determined by

$$\begin{aligned} \widehat{x}(n) &= R_{x(n)g_n} R_{g_n}^{-1} g_n \\ Q_n &= R_{x(n)} - R_{x(n)g_n} R_{g_n}^{-1} R_{g_n x(n)} . \qquad (5.7) \end{aligned}$$

To compute $\widehat{x}(n+1)$ we need $R_{g_{n+1}}$. For convenience we will suppress the index n in $R_{x(n)}$, $R_{x(n)g_n}$ and R_{g_n}, that is,

$$R_x = R_{x(n)}, \qquad R_{xg} = R_{x(n)g_n} \qquad \text{and} \qquad R_g = R_{g_n} .$$

However, we will not suppress the index $n+1$. Using the fact that $g_{n+1}^* = \begin{bmatrix} g_n^* & y(n)^* \end{bmatrix}$, we see that the covariance for g_{n+1} is given by

$$R_{g_{n+1}} = E \begin{bmatrix} g_n \\ y(n) \end{bmatrix} \begin{bmatrix} g_n^* & y(n)^* \end{bmatrix} = \begin{bmatrix} R_g & Eg_n y(n)^* \\ Ey(n)g_n^* & Ey(n)y(n)^* \end{bmatrix} . \qquad (5.8)$$

Now let us compute the entries of $R_{g_{n+1}}$. We claim that

$$Ey(n)g_n^* = CR_{x(n)g_n} = CR_{xg}\,. \tag{5.9}$$

Recall that the solution to the difference equation in (5.2) is given by

$$x(k) \;=\; A^k x_0 \qquad \text{and} \qquad y(k) = CA^k x_0 + Dv(k)\,. \tag{5.10}$$

Recall also that $\{v(k)\}$ is a white noise process which is orthogonal to x_0. In particular, equation (5.10) shows that $v(n)$ is orthogonal to $y(k)$ for $k = 0, 1, \cdots, n-1$. Since g_n only contains vectors from $\{y(k)\}_0^{n-1}$, it follows that $v(n)$ is orthogonal to g_n, that is, $Ev(n)g_n^* = 0$. Using this along with $y(n) = Cx(n) + Dv(n)$, we arrive at

$$Ey(n)g_n^* = E(Cx(n) + Dv(n))g_n^* = ECx(n)g_n^* + EDv(n)g_n^* = CEx(n)g_n^* = CR_{xg}\,.$$

Therefore (5.9) holds.

We claim that

$$Ey(n)y(n)^* = CR_{x(n)}C^* + DD^*\,. \tag{5.11}$$

Because $v(n)$ is orthogonal to x_0 and $x(n) = A^n x_0$, we see that $v(n)$ is orthogonal to $x(n)$, that is, $Ex(n)v(n)^* = 0$. This readily implies that

$$\begin{aligned} Ey(n)y(n)^* &= E\left(Cx(n) + Dv(n)\right)\left(Cx(n) + Dv(n)\right)^* \\ &= CEx(n)x(n)^*C^* + CEx(n)v(n)^*D^* + DEv(n)x(n)^*C^* + DEv(n)v(n)^*D^* \\ &= CR_{x(n)}C^* + DD^*\,. \end{aligned}$$

Therefore (5.11) holds.

Recall that $R_{gf} = R_{fg}^*$ where f and g are random vectors. Substituting (5.9) and (5.11) into the expression for $R_{g_{n+1}}$ in (5.8) yields

$$R_{g_{n+1}} = \begin{bmatrix} R_g & R_{gx}C^* \\ CR_{xg} & CR_xC^* + DD^* \end{bmatrix}\,. \tag{5.12}$$

Notice that the Schur complement for $R_{g_{n+1}}$ is given by

$$\Delta = CR_xC^* + DD^* - CR_{xg}R_g^{-1}R_{gx}C^* = CQ_nC^* + DD^*\,. \tag{5.13}$$

The last equality follows from the expression for the covariance error Q_n in (5.7). Throughout we always assume that $CQ_nC^* + DD^*$ is invertible; see Remark 2.5.2. So if R_{g_n} is invertible, then the matrix inversion Lemma 2.4.1 guarantees that $R_{g_{n+1}}$ is invertible. By setting $n = 0$ in (5.11), we see that $R_{g_1} = Ey(0)y(0)^* = CR_{x_0}C^* + DD^*$. In other words, we obtain $R_{g_1} = CQ_0C^* + DD^*$ where $Q_0 = Ex_0x_0^*$. Because we assume that Schur complements $CQ_nC^* + DD^*$ are invertible for all integers $n \geq 0$, an inductive argument shows that R_{g_n} is invertible for all integers $n \geq 1$.

We will also need the cross covariance between $x(n+1)$ and g_{n+1}. We claim that

$$R_{x(n+1)g_{n+1}} = \begin{bmatrix} AR_{xg} & AR_xC^* \end{bmatrix}\,. \tag{5.14}$$

Since $v(n)$ is orthogonal to x_0 and $x(n) = A^n x_0$, we see that $v(n)$ is orthogonal to $x(n)$. Using $x(n+1) = Ax(n)$ and $g_{n+1}^* = \left[\begin{array}{cc} g_n^* & y(n)^* \end{array} \right]$, we obtain

$$\begin{aligned} Ex(n+1)g_{n+1}^* &= EAx(n)g_{n+1}^* = AEx(n)g_{n+1}^* \\ &= \left[\begin{array}{cc} AEx(n)g_n^* & AEx(n)y(n)^* \end{array} \right] \\ &= \left[\begin{array}{cc} AR_{xg} & AEx(n)\left(Cx(n)+Dv(n)\right)^* \end{array} \right] \\ &= \left[\begin{array}{cc} AR_{xg} & AEx(n)x(n)^*C^* \end{array} \right] . \end{aligned}$$

Therefore (5.14) holds.

Using $x(n+1) = Ax(n)$, we obtain $Ex(n+1)x(n+1)^* = AEx(n)x(n)^*A^*$. In other words,

$$R_{x(n+1)} = AR_{x(n)}A^* . \tag{5.15}$$

Notice that the state covariance error Q_{n+1} is obtained by replacing n by $n+1$ in (5.7). By employing the matrix inversion Lemma 4.2 with $M = R_{gx}C^*$ and $N = CR_{xg}$, we see that the state covariance error at time $n+1$ is given by

$$\begin{aligned} &Q_{n+1} = R_{x(n+1)} - R_{x(n+1)g_{n+1}}R_{g_{n+1}}^{-1}R_{g_{n+1}x(n+1)} \\ &= AR_xA^* + \\ &- \left[\begin{array}{cc} AR_{xg} & AR_xC^* \end{array} \right] \left[\begin{array}{cc} R_g^{-1} + R_g^{-1}R_{gx}C^*\Delta^{-1}CR_{xg}R_g^{-1} & -R_g^{-1}R_{gx}C^*\Delta^{-1} \\ -\Delta^{-1}CR_{xg}R_g^{-1} & \Delta^{-1} \end{array} \right] \left[\begin{array}{c} R_{gx}A^* \\ CR_xA^* \end{array} \right] \\ &= AR_xA^* - AR_{xg}R_g^{-1}R_{gx}A^* - AR_{xg}R_g^{-1}R_{gx}C^*\Delta^{-1}CR_{xg}R_g^{-1}R_{gx}A^* \\ &+AR_{xg}R_g^{-1}R_{gx}C^*\Delta^{-1}CR_xA^* + AR_xC^*\Delta^{-1}CR_{xg}R_g^{-1}R_{gx}A^* - AR_xC^*\Delta^{-1}CR_xA^* \\ &= AQ_nA^* - A\left(R_x - R_{xg}R_g^{-1}R_{gx}\right)C^*\Delta^{-1}C\left(R_x - R_{xg}R_g^{-1}R_{gx}\right)A^* \\ &= AQ_nA^* - AQ_nC^*\Delta^{-1}CQ_nA^* . \end{aligned}$$

This is precisely the Riccati difference equation in (5.5).

Notice that the optimal state estimate $\widehat{x}(n+1)$ is obtained by replacing n by $n+1$ in (5.7). In other words, the optimal state estimate $\widehat{x}(n+1)$ is given by

$$\begin{aligned} &\widehat{x}(n+1) = R_{x(n+1)g_{n+1}}R_{g_{n+1}}^{-1}g_{n+1} \\ &= \left[\begin{array}{cc} AR_{xg} & AR_xC^* \end{array} \right] \left[\begin{array}{cc} R_g^{-1} + R_g^{-1}R_{gx}C^*\Delta^{-1}CR_{xg}R_g^{-1} & -R_g^{-1}R_{gx}C^*\Delta^{-1} \\ -\Delta^{-1}CR_{xg}R_g^{-1} & \Delta^{-1} \end{array} \right] \left[\begin{array}{c} g_n \\ y(n) \end{array} \right] \\ &= AR_{xg}R_g^{-1}g_n + AR_{xg}R_g^{-1}R_{gx}C^*\Delta^{-1}CR_{xg}R_g^{-1}g_n - AR_{xg}R_g^{-1}R_{gx}C^*\Delta^{-1}y(n) \\ &-AR_xC^*\Delta^{-1}CR_{xg}R_g^{-1}g_n + AR_xC^*\Delta^{-1}y(n) \\ &= A\widehat{x}(n) - AR_{xg}R_g^{-1}R_{gx}C^*\Delta^{-1}\left(y(n) - C\widehat{x}(n)\right) + AR_xC^*\Delta^{-1}\left(y(n) - C\widehat{x}(n)\right) \\ &= A\widehat{x}(n) + A\left(R_x - R_{xg}R_g^{-1}R_{gx}\right)C^*\Delta^{-1}\left(y(n) - C\widehat{x}(n)\right) \\ &= A\widehat{x}(n) + AQ_nC^*\Delta^{-1}\left(y(n) - C\widehat{x}(n)\right) . \end{aligned}$$

This yields the Kalman filtering recursion for the optimal state estimate $\widehat{x}(n)$ in (5.3) and (5.4). The proof is now complete.

2.5.1 Exercise

Problem 1. Consider the system $\dot{q} = Fq$ and $z = Cq + w$ determined by

$$\begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} \quad \text{and} \quad z(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} + Dw(t)\,. \tag{5.16}$$

For example, this be the equations of motion for a mass ($m = 1$) spring ($k = 2$) and damper ($c = 3$) system where $z(t) = q_1(t) + Dw(t)$ is the measurement of the distance q_1 corrupted by the noise process $w(t)$. Now assume that one samples the output at every $h = .01$ seconds. Set $x(n) = q(nh)$, $y(n) = z(nh)$ and $v(n) = w(nh)$ where n is a positive integer. Since $q(t) = e^{Ft}q(0)$, we see that

$$x(n) = \left(e^{Fh}\right)^n x(0) \qquad \text{and} \qquad y(n) = Cx(n) + Dv(n)\,.$$

In other words, $x(n)$ is the solution to the state space difference system given by

$$x(n+1) = Ax(n) \qquad \text{and} \qquad y(n) = Cx(n) + Dv(n)$$

where $A = e^{Fh}$; see Section 9.3.1 in Chapter 9. Now assume that $v(n)$ is a Gaussian white noise process and the initial state is $x_0 = \begin{bmatrix} 1 & 2 \end{bmatrix}^*$. Implement the Kalman filter in Theorem 2.5.1 to compute the optimal state estimate $\widehat{x}(n)$ for $x(n)$ for $0 \leq n \leq 500$. Notice that $n = 500$ corresponds to five seconds. Compare your estimate $\widehat{x}(n)$ with the actual state $x(n) = A^n x_0$. Assume that $Q_0 = I$ and $\widehat{x}(0) = \begin{bmatrix} 0 & 0 \end{bmatrix}^*$. Try your results for $D = 1/2$ and $D = 1$. Finally, randn is the Matlab command to generate Gaussian white noise.

Chapter 3

Kalman Filtering

This chapter is devoted to the discrete time Kalman filter. A brief introduction to the continuous time Kalman filter is given in Section 3.8.

3.1 The Kalman filter

In this section we will present the discrete time Kalman filter. Recall that $w(n)$ is a discrete time random process with values in $\mathbb{C}^m$ if $w(n)$ is a random vector in $\mathbb{C}^m$ for all integers n. We say that $w(n)$ is a *mean zero process* if $Ew(n) = 0$ for all n. Finally, $w(n)$ is a *white noise process* if $w(n)$ is a mean zero random process and

$$Ew(j)w(k)^* = \delta_{jk} I \,. \tag{1.1}$$

Here δ_{jk} is the Kronecker delta. By definition, $\delta_{jk} = 1$ if $j = k$ and $\delta_{jk} = 0$ if $j \neq k$.

Consider the system given by

$$x(n+1) = Ax(n) + Bu(n) \qquad \text{and} \qquad y(n) = Cx(n) + Dv(n) \,. \tag{1.2}$$

Here A is an operator on $\mathcal{X}$ and B maps $\mathcal{U}$ into $\mathcal{X}$ while C maps $\mathcal{X}$ into $\mathcal{Y}$ and D maps $\mathcal{V}$ into $\mathcal{Y}$ where $\mathcal{X}, \mathcal{U}, \mathcal{Y}$ and $\mathcal{V}$ are all $\mathbb{C}^k$ spaces of the appropriate size. It is emphasized that $\{A, B, C, D\}$ can be time varying matrices, that is, $A = A(n)$, $B = B(n)$, $C = C(n)$ and $D = D(n)$ for all integers n. However, the index n is suppressed in our development. The initial condition x_0 is a random vector with values in $\mathcal{X}$. The disturbance $u(n)$ and $v(n)$ are independent white noise random process. Moreover, we assume that the initial condition x_0, $u(n)$ and $v(m)$ are all independent random vectors for all integers n and m. In particular, this implies that x_0, $u(n)$ and $v(m)$ are orthogonal for all integers n and m. We also assume that the initial condition $\widehat{x}_0$ for the Kalman filter and the initial condition $Q_0 = Ex(0)x(0)^*$ for the discrete time Riccati equation are known. Finally, is it noted that $u(n)$ is called the disturbance or state noise, while $v(n)$ is the measurement noise.

The Kalman filtering problem is to compute the best estimate $\widehat{x}(k)$ of the state $x(k)$ given the past output $\{y(j)\}_0^{k-1}$. The Kalman filter is an optimal state estimator. To be explicit, let $\mathcal{M}_n$ be the subspace generated by the random vectors $\{y(j)\}_0^n$, that is, $\mathcal{M}_n = \bigvee_{j=0}^n y(j)$. Let $P_{\mathcal{M}_n}$ be the orthogonal projection onto $\mathcal{M}_n$ for all integers $n \geq 0$. Then the best estimate

$\widehat{x}(k)$ of the state $x(k)$ is given by the orthogonal projection $\widehat{x}(k) = P_{\mathcal{M}_{k-1}} x(k)$. Finally, we assume that $y(-1) = 0$, or equivalently, $\mathcal{M}_{-1} = \{0\}$. The following result known as the Kalman filter provides us with a recursive algorithm to compute $\widehat{x}(k)$.

THEOREM 3.1.1 *Consider the state space system*

$$x(n+1) = Ax(n) + Bu(n) \qquad \text{and} \qquad y(n) = Cx(n) + Dv(n) \tag{1.3}$$

where $u(n)$ and $v(n)$ are independent white noise random processes, which are independent to the initial condition $x(0)$. Then the optimal estimate $\widehat{x}(k) = P_{\mathcal{M}_{k-1}} x(k)$ of the state $x(k)$ given the past $\{y(j)\}_0^{k-1}$ is recursively computed by

$$\begin{aligned} \widehat{x}(n+1) &= A\widehat{x}(n) + \Delta_n(y(n) - C\widehat{x}(n)) & (1.4)\\ \Delta_n &= AQ_nC^* \left(CQ_nC^* + DD^*\right)^{-1} . & (1.5) \end{aligned}$$

The state covariance error $Q_k = E(x(k) - \widehat{x}(k))(x(k) - \widehat{x}(k))^*$ is recursively computed by solving the Riccati difference equation

$$Q_{n+1} = AQ_nA^* + BB^* - AQ_nC^* \left(CQ_nC^* + DD^*\right)^{-1} CQ_nA^* , \tag{1.6}$$

subject to the initial condition $Q_0 = Ex(0)x(0)^*$.

REMARK 3.1.2 *It is emphasized that when implementing the Kalman filter we always assume that the inverse of $CQ_nC^* + DD^*$ exists for all integers $n \geq 0$. In particular, if DD^* is invertible, then $CQ_nC^* + DD^*$ is strictly positive, and thus, invertible for all n.*

3.1.1 Kalman Prediction

In random processes prediction is trying to estimate the future given the past. This section is devoted to Kalman prediction, that is, predicting the future state given the past output measurements. As before, consider the state space system given in (1.2) where $u(n)$ and $v(n)$ are independent white noise processes which are also independent to the initial state $x(0)$. Now consider any integer $m \geq 0$. The Kalman prediction problem is to compute the best estimate $\widehat{x}(n+m|n-1)$ of the state $x(n+m)$ given the past output $\{y(j)\}_0^{n-1}$. To be precise, let $\mathcal{M}_n$ be the subspace generated by the random vectors $\{y(j)\}_0^n$, that is, $\mathcal{M}_n = \bigvee_{j=0}^n y(j)$. Let $P_{\mathcal{M}_n}$ be the orthogonal projection onto $\mathcal{M}_n$ for all integers $n \geq 0$. Then the optimal state predictor $\widehat{x}(n+m|n-1)$ of the state $x(n+m)$ given $\{y(j)\}_0^{n-1}$ is determined by the orthogonal projection

$$\widehat{x}(n+m|n-1) = P_{\mathcal{M}_{n-1}} x(n+m) . \tag{1.7}$$

Notice that if $m = 0$, then $\widehat{x}(n|n-1) = \widehat{x}(n) = P_{\mathcal{M}_{n-1}} x(n)$ is simply the optimal state estimate $\widehat{x}(n)$ for $x(n)$ computed by the Kalman filter in Theorem 3.1.1. The following result known as the Kalman predictor provides us with a recursive algorithm to compute the optimal state predictor $\widehat{x}(n+m|n-1)$.

THEOREM 3.1.3 *Consider the state space system*

$$x(n+1) = Ax(n) + Bu(n) \qquad \text{and} \qquad y(n) = Cx(n) + Dv(n) \tag{1.8}$$

where $u(n)$ and $v(n)$ are independent white noise random processes, which are independent to the initial condition $x(0)$. Let $\widehat{x}(n) = P_{\mathcal{M}_{n-1}}x(n)$ be the optimal state estimate of $x(n)$ given the past $\{y(j)\}_0^{n-1}$ recursively computed by (1.4), (1.5) and (1.6) in the Kalman filtering Theorem 3.1.1. Then the optimal state predictor $\widehat{x}(n+m|n-1)$ of $x(n+m)$ given the past $\{y(j)\}_0^{n-1}$ is computed by

$$\widehat{x}(n+m|n-1) = A(n+m-1)A(n+m-2)\cdots A(n+1)A(n)\widehat{x}(n) \qquad (\textit{where } m \geq 1)\,. \tag{1.9}$$

In particular, $\widehat{x}(n+1|n-1) = A(n)\widehat{x}(n)$. Finally, if A is time invariant ($A = A(k)$ for all integers k), then the optimal state predictor is given by $\widehat{x}(n+m|n-1) = A^m\widehat{x}(n)$.

Proof. According to the results in Section 9.9 in Chapter 9, the solution to the difference equation in (1.8) is given by

$$x(n+m) = \Psi(n+m-1, n-1)x(n) + \sum_{j=n}^{n+m-1} \Psi(n+m-1, j)B(j)u(j) \tag{1.10}$$

$$y(n) = C(n)\Psi(n-1,-1)x_0 + \sum_{j=0}^{n-1} C(n)\Psi(n-1,j)B(j)u(j) + D(n)v(n). \tag{1.11}$$

Here $\Psi(n,\nu) = A(n)A(n-1)\cdots A(\nu+1)$ and $\Psi(j,j) = I$. In particular,

$$\mathcal{M}_{n-1} = \bigvee_{j=0}^{n-1} y_j \subset \bigvee\{x(0), u(0), u(1), \cdots, u(n-2), v(0), v(1), \cdots, v(n-1)\}\,. \tag{1.12}$$

Recall that $\{v(n)\}$ is a white noise process, which is orthogonal to both x_0 and $\{u(n)\}$. Equation (1.11) or (1.12) shows that $\{u(j)\}_{n-1}^{\infty}$ is orthogonal to $y(j)$ for $j = 0, 1, \cdots, n-1$. This implies that $P_{\mathcal{M}_{n-1}}u(j) = 0$ for all integers $j \geq n-1$. Using this in the expression for $x(m+n)$ in (1.10), we obtain

$$\begin{aligned}\widehat{x}(n+m|n-1) &= P_{\mathcal{M}_{n-1}}x(n+m) = P_{\mathcal{M}_{n-1}}\Psi(n+m-1, n-1)x(n) \\ &= \Psi(n+m-1, n-1)P_{\mathcal{M}_{n-1}}x(n) = \Psi(n+m-1, n-1)\widehat{x}(n)\,.\end{aligned}$$

This yields (1.9) and completes the proof.

3.2 A matrix inversion proof of the Kalman filter

In this section we will use the matrix inversion Lemma 2.4.1 in Chapter 2 to derive the Kalman filtering Theorem 3.1.1. The approach in this section is essentially the same as the proof of Theorem 2.5.1 in Chapter 2, except that we have to incorporate the state noise

in our analysis. To this end, let g_n be the random vector formed by the first n outputs $\{y(k)\}_0^{n-1}$, that is,

$$g_n = \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(n-1) \end{bmatrix}. \tag{2.1}$$

Obviously, $\mathcal{M}_{n-1}$ equals the span of g_n. Let Q_n be the state covariance error, that is,

$$Q_n = E(x(n) - \widehat{x}(n))(x(n) - \widehat{x}(n))^* .$$

In a moment we will show that R_{g_n} is invertible. According to Theorem 2.2.1 in Chapter 2, the optimal state estimate $\widehat{x}(n) = P_{\mathcal{M}_{n-1}} x(n)$ and the corresponding state covariance error Q_n is given by

$$\begin{aligned} \widehat{x}(n) &= R_{x(n)g_n} R_{g_n}^{-1} g_n \\ Q_n &= R_{x(n)} - R_{x(n)g_n} R_{g_n}^{-1} R_{g_n x(n)} . \end{aligned} \tag{2.2}$$

To compute the optimal state estimate $\widehat{x}(n+1)$ we need $R_{g_{n+1}}$. For convenience we will suppress the index n in $R_{x(n)}$, $R_{x(n)g_n}$ and R_{g_n}, that is,

$$R_x = R_{x(n)}, \qquad R_{xg} = R_{x(n)g_n} \qquad \text{and} \qquad R_g = R_{g_n} .$$

However, we will not suppress the index $n+1$. Using the fact that $g_{n+1}^* = \begin{bmatrix} g_n^* & y(n)^* \end{bmatrix}$, we see that the covariance for g_{n+1} is given by

$$R_{g_{n+1}} = E \begin{bmatrix} g_n \\ y(n) \end{bmatrix} \begin{bmatrix} g_n^* & y(n)^* \end{bmatrix} = \begin{bmatrix} R_g & Eg_n y(n)^* \\ Ey(n)g_n^* & Ey(n)y(n)^* \end{bmatrix}. \tag{2.3}$$

Now let us compute the entries of $R_{g_{n+1}}$. We claim that

$$Ey(n)g_n^* = CR_{x(n)g_n} = CR_{xg} . \tag{2.4}$$

According to the results in Section 9.9 in Chapter 9, the solution to the difference equation in (1.3) is given by

$$x(k) = \Psi(k-1,-1)x_0 + \sum_{j=0}^{k-1} \Psi(k-1,j)B(j)u(j) \tag{2.5}$$

$$y(k) = C(k)\Psi(k-1,-1)x_0 + \sum_{j=0}^{k-1} C(k)\Psi(k-1,j)B(j)u(j) + D(k)v(k) . \tag{2.6}$$

Here $\Psi(k,\nu) = A(k)A(k-1)\cdots A(\nu+1)$ and $\Psi(j,j) = I$. In particular,

$$\bigvee x(n) \subset \bigvee \{x(0), u(0), u(1), \cdots, u(n-1)\} \tag{2.7}$$

$$\mathcal{M}_n = \bigvee_{k=0}^{n} y_k \subset \bigvee \{x(0), u(0), u(1), \cdots, u(n-1), v(0), v(1), \cdots, v(n)\} . \tag{2.8}$$

Recall that $\{v(k)\}$ is a white noise process, which is orthogonal to both x_0 and $\{u(k)\}$. In particular, equation (2.6) or (2.8) shows that $v(n)$ is orthogonal to $y(k)$ for $k = 0, 1, \cdots, n-1$. Since g_n only contains vectors from $\{y(k)\}_0^{n-1}$, it follows that $v(n)$ is orthogonal to g_n, that is, $Ev(n)g_n^* = 0$. Using this along with $y(n) = Cx(n) + Dv(n)$, we arrive at

$$Ey(n)g_n^* = E(Cx(n) + Dv(n))g_n^* = ECx(n)g_n^* + EDv(n)g_n^* = CEx(n)g_n^* = CR_{xg}\,.$$

Therefore (2.4) holds.

Next we need an expression for $Ey(n)y(n)^*$. We claim that

$$Ey(n)y(n)^* = CR_{x(n)}C^* + DD^*\,. \tag{2.9}$$

Because $v(n)$ is orthogonal to x_0 and $\{u(k)\}_0^\infty$, equation (2.5) or (2.7) shows that $v(n)$ is orthogonal to $x(n)$, that is, $Ex(n)v(n)^* = 0$. This readily implies that

$$Ey(n)y(n)^* \;=\; E\left(Cx(n) + Dv(n)\right)\left(Cx(n) + Dv(n)\right)^* = CR_{x(n)}C^* + DD^*\,.$$

Therefore (2.9) holds.

Recall that $R_{gf} = R_{fg}^*$ where f and g are random vectors. Substituting (2.4) and (2.9) into the expression for $R_{g_{n+1}}$ in (2.3) yields

$$R_{g_{n+1}} = \begin{bmatrix} R_g & R_{gx}C^* \\ CR_{xg} & CR_xC^* + DD^* \end{bmatrix}\,. \tag{2.10}$$

Notice that the Schur complement for $R_{g_{n+1}}$ is given by

$$\Delta = CR_xC^* + DD^* - CR_{xg}R_g^{-1}R_{gx}C^* = CQ_nC^* + DD^*\,. \tag{2.11}$$

The last equality follows from the expression for the state covariance error Q_n in (2.2). Throughout we always assume that $CQ_nC^* + DD^*$ is invertible; see Remark 3.1.2. So if R_{g_n} is invertible, then the matrix inversion Lemma 2.4.1 in Chapter 2 guarantees that $R_{g_{n+1}}$ is invertible. By setting $n = 0$ in (2.9), we see that $R_{g_1} = Ey(0)y(0)^* = CR_{x_0}C^* + DD^*$. In other words, $R_{g_1} = CQ_0C^* + DD^*$ where $Q_0 = Ex_0x_0^*$. Because we assume that Schur complements $CQ_nC^* + DD^*$ are invertible for all integers $n \geq 0$, an inductive argument shows that R_{g_n} is invertible for all integers $n \geq 1$.

To compute Q_{n+1} we need an expression for $R_{x(n+1)}$. We claim that

$$R_{x(n+1)} = AR_{x(n)}A^* + BB^*\,. \tag{2.12}$$

Recall that $\{u(k)\}$ is a white noise process which is orthogonal to the initial condition x_0. By consulting the solution for $x(n)$ in (2.5) or (2.7), we see that $u(n)$ is orthogonal to $x(n)$, that is, $Ex(n)u(n)^* = 0$. This along with $x(n+1) = Ax(n) + Bu(n)$ yields

$$\begin{aligned} R_{x(n+1)} &= Ex(n+1)x(n+1)^* = E\left(Ax(n) + Bu(n)\right)\left(Ax(n) + Bu(n)\right)^* \\ &= AEx(n)x(n)^*A^* + BEu(n)u(n)^*B^* = AR_{x(n)}A^* + BB^*\,. \end{aligned}$$

Therefore (2.12) holds.

We claim that the cross covariance between $x(n+1)$ and g_n is given by

$$R_{x(n+1)g_{n+1}} = \left[\begin{array}{cc} AR_{xg} & AR_xC^* \end{array} \right] . \tag{2.13}$$

To see this first observe that $u(n)$ is orthogonal to x_0, $\{u(k)\}_0^{n-1}$ and $\{v(k)\}$. By consulting the solution for the output y in (2.6) or (2.8), we see that $u(n)$ is orthogonal to $\{y(k)\}_0^n$. Since $g_{n+1}^* = \left[\begin{array}{cc} g_n^* & y(n)^* \end{array} \right]$, we arrive at $Eu(n)g_{n+1}^* = 0$. Because $v(n)$ is orthogonal to x_0 and $\{u(k)\}$, equation (2.5) shows that $v(n)$ is also orthogonal to $x(n)$. This readily implies that

$$\begin{aligned} Ex(n+1)g_{n+1}^* &= E\left(Ax(n)+Bu(n)\right)g_{n+1}^* = AEx(n)g_{n+1}^* \\ &= \left[\begin{array}{cc} AEx(n)g_n^* & AEx(n)y(n)^* \end{array} \right] \\ &= \left[\begin{array}{cc} AR_{xg} & AEx(n)\left(Cx(n)+Dv(n)\right)^* \end{array} \right] \\ &= \left[\begin{array}{cc} AR_{xg} & AEx(n)x(n)^*C^* \end{array} \right] . \end{aligned}$$

Therefore (2.13) holds.

Notice that the state covariance error Q_{n+1} is obtained by replacing n by $n+1$ in (2.2). By employing the matrix inversion Lemma 2.4.1 in Chapter 2. with $M = R_{gx}C^*$ and $N = CR_{xg}$, we see that the state covariance error at time $n+1$ is given by

$$\begin{aligned} &Q_{n+1} = R_{x(n+1)} - R_{x(n+1)g_{n+1}}R_{g_{n+1}}^{-1}R_{g_{n+1}x(n+1)} \\ &= AR_xA^* + BB^* + \\ &- \left[\begin{array}{cc} AR_{xg} & AR_xC^* \end{array} \right] \left[\begin{array}{cc} R_g^{-1} + R_g^{-1}R_{gx}C^*\Delta^{-1}CR_{xg}R_g^{-1} & -R_g^{-1}R_{gx}C^*\Delta^{-1} \\ -\Delta^{-1}CR_{xg}R_g^{-1} & \Delta^{-1} \end{array} \right] \left[\begin{array}{c} R_{gx}A^* \\ CR_xA^* \end{array} \right] \\ &= BB^* + AR_xA^* - AR_{xg}R_g^{-1}R_{gx}A^* - AR_{xg}R_g^{-1}R_{gx}C^*\Delta^{-1}CR_{xg}R_g^{-1}R_{gx}A^* \\ &+AR_{xg}R_g^{-1}R_{gx}C^*\Delta^{-1}CR_xA^* + AR_xC^*\Delta^{-1}CR_{xg}R_g^{-1}R_{gx}A^* - AR_xC^*\Delta^{-1}CR_xA^* \\ &= BB^* + AQ_nA^* - A\left(R_x - R_{xg}R_g^{-1}R_{gx}\right)C^*\Delta^{-1}C\left(R_x - R_{xg}R_g^{-1}R_{gx}\right)A^* \\ &= BB^* + AQ_nA^* - AQ_nC^*\Delta^{-1}CQ_nA^* . \end{aligned}$$

This is precisely the Riccati difference equation in (1.6). To obtain the initial condition, recall that $\mathcal{M}_{-1} = 0$, that is, $y(-1) = 0$. Hence $\widetilde{x}(0) = x(0) - P_{\mathcal{M}_{-1}}x(0) = x(0)$. Thus $Q_0 = E\widetilde{x}(0)\widetilde{x}(0)^* = Ex(0)x(0)^*$.

Notice that the optimal state estimate $\widehat{x}(n+1)$ is obtained by replacing n by $n+1$ in (2.2). In other words, the optimal state estimate $\widehat{x}(n+1)$ is given by

$$\begin{aligned} &\widehat{x}(n+1) = R_{x(n+1)g_{n+1}}R_{g_{n+1}}^{-1}g_{n+1} \\ &= \left[\begin{array}{cc} AR_{xg} & AR_xC^* \end{array} \right] \left[\begin{array}{cc} R_g^{-1} + R_g^{-1}R_{gx}C^*\Delta^{-1}CR_{xg}R_g^{-1} & -R_g^{-1}R_{gx}C^*\Delta^{-1} \\ -\Delta^{-1}CR_{xg}R_g^{-1} & \Delta^{-1} \end{array} \right] \left[\begin{array}{c} g_n \\ y(n) \end{array} \right] \\ &= AR_{xg}R_g^{-1}g_n + AR_{xg}R_g^{-1}R_{gx}C^*\Delta^{-1}CR_{xg}R_g^{-1}g_n - AR_{xg}R_g^{-1}R_{gx}C^*\Delta^{-1}y(n) \\ &-AR_xC^*\Delta^{-1}CR_{xg}R_g^{-1}g_n + AR_xC^*\Delta^{-1}y(n) \\ &= A\widehat{x}(n) - AR_{xg}R_g^{-1}R_{gx}C^*\Delta^{-1}\left(y(n) - C\widehat{x}(n)\right) + AR_xC^*\Delta^{-1}\left(y(n) - C\widehat{x}(n)\right) \\ &= A\widehat{x}(n) + A\left(R_x - R_{xg}R_g^{-1}R_{gx}\right)C^*\Delta^{-1}\left(y(n) - C\widehat{x}(n)\right) \\ &= A\widehat{x}(n) + AQ_nC^*\Delta^{-1}\left(y(n) - C\widehat{x}(n)\right) . \end{aligned}$$

This yields the Kalman filtering recursion for the optimal state estimate $\widehat{x}(n)$ in (1.4) and (1.5). The proof is now complete.

3.2.1 Exercise

Problem 1. Consider the discrete time system determined by

$$\begin{aligned} \begin{bmatrix} x_1(n+1) \\ x_2(n+1) \end{bmatrix} &= \begin{bmatrix} 0 & 1 \\ -0.98 & 1.94 \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(n) \\ y(n) &= \begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} + v(n)\,. \end{aligned} \tag{2.14}$$

Here $\{u(n)\}$ and $\{v(n)\}$ are independent Gaussian white noise processes which are orthogonal to the initial random variable $x(0) = x_0$. Assume that $Ex_0x_0^* = I$ and the initial state is $x_0 = \begin{bmatrix} 1 & 2 \end{bmatrix}^{tr}$. Implement the Kalman filter in Theorem 3.1.1 to compute the optimal state estimate $\widehat{x}(n)$ for $x(n)$ for $0 \le n \le 500$ where the initial state estimate is $\widehat{x}(0) = \begin{bmatrix} 0 & 0 \end{bmatrix}^{tr}$. Compare your estimate $\widehat{x}(n)$ with the actual state $x(n)$.

3.3 Recursive estimation and Kalman filtering

In this section we will develop a simple recursive estimation result. Then we will use this result to provide another proof of the Kalman filter. As before, let $\mathcal{K}$ be the Hilbert space formed by the set of all random variables x such that $E|x|^2$ is finite. If f in $\mathbb{C}^k$ and g in $\mathbb{C}^m$ are random vectors, then $f \bigvee g$ is the subspace of $\mathcal{K}$ generated by the span of all the components of both f and g. If $\mathcal{H}_1$ and $\mathcal{H}_2$ are two subspaces of random variables, then $\mathcal{H}_1 \bigvee \mathcal{H}_2$ is the subspace formed by the closed linear span of $\mathcal{H}_1$ and $\mathcal{H}_2$. If $\mathcal{M}$ is a subspace of $\mathcal{K}$, then $P_{\mathcal{M}}$ denotes the orthogonal projection onto $\mathcal{M}$. Recall that the notation $\mathcal{F} \oplus \mathcal{G}$ means that $\mathcal{F}$ and $\mathcal{G}$ are two orthogonal subspaces and $\mathcal{F} \oplus \mathcal{G} = \mathcal{F} \bigvee \mathcal{G}$. In this case, $P_{\mathcal{F}\oplus\mathcal{G}} = P_{\mathcal{F}} + P_{\mathcal{G}}$. The following is a simple recursive estimation result.

LEMMA 3.3.1 *Let f be a random vector with values in $\mathbb{C}^k$. Let $\mathcal{M}$ be a subspace of random variables, and $\mathcal{Y}$ be the subspace generated by the random vector y in $\mathbb{C}^m$. Set $\mathcal{H} = \mathcal{M} \bigvee \mathcal{Y}$. Then $\mathcal{H} = \mathcal{E} \oplus \mathcal{M}$ where $\mathcal{E}$ is the subspace generated by the vector $\varphi = y - P_{\mathcal{M}}y$. Moreover, if R_φ is invertible, then*

$$P_{\mathcal{H}}f = P_{\mathcal{M}}f + R_{f\varphi}R_\varphi^{-1}\varphi\,. \tag{3.1}$$

Moreover, this decomposition is orthogonal, that is, $P_{\mathcal{M}}f$ is orthogonal to $R_{f\varphi}R_\varphi^{-1}\varphi = P_{\mathcal{E}}f$. Finally, the error covariance for $f - P_{\mathcal{H}}f$ is given by

$$E(f - P_{\mathcal{H}}f)(f - P_{\mathcal{H}}f)^* = E(f - P_{\mathcal{M}}f)(f - P_{\mathcal{M}}f)^* - R_{f\varphi}R_\varphi^{-1}R_{f\varphi}^*\,. \tag{3.2}$$

Proof. By the projection theorem, the vector $\varphi = y - P_{\mathcal{M}}y$ is orthogonal to $\mathcal{M}$. So the subspace $\mathcal{E}$ generated by φ is orthogonal to $\mathcal{M}$. Since the components of φ are contained in of $\mathcal{M} \bigvee \mathcal{Y}$, we have $\mathcal{M} \oplus \mathcal{E} \subset \mathcal{H}$. Clearly, $\mathcal{Y}$ is contained in $\mathcal{M} \oplus \mathcal{E}$. Therefore $\{\mathcal{M}, \mathcal{Y}\}$

and $\{\mathcal{M}, \mathcal{E}\}$ span the same space. In particular, $\mathcal{H} = \mathcal{M} \oplus \mathcal{E}$. This readily implies that the orthogonal projection $P_{\mathcal{H}} = P_{\mathcal{M}} + P_{\mathcal{E}}$. Because φ generates the subspace $\mathcal{E}$, Theorem 2.2.1 in Chapter 2 shows that $P_{\mathcal{E}} f = R_{f\varphi} R_{\varphi}^{-1} \varphi$. Thus

$$P_{\mathcal{H}} f = P_{\mathcal{M}} f + P_{\mathcal{E}} f = P_{\mathcal{M}} f + R_{f\varphi} R_{\varphi}^{-1} \varphi \,.$$

Hence (3.1) holds. Since $P_{\mathcal{E}} f = R_{f\varphi} R_{\varphi}^{-1} \varphi$ and $\mathcal{M}$ is orthogonal to $\mathcal{E}$, the random vector $P_{\mathcal{M}} f$ is orthogonal to $R_{f\varphi} R_{\varphi}^{-1} \varphi$.

To verify that (3.2) holds, recall that $\mathcal{M}$ is orthogonal to $\mathcal{E}$, and thus, $P_{\mathcal{E}} P_{\mathcal{M}} = 0$. Hence $P_{\mathcal{E}}(f - P_{\mathcal{M}} f) = P_{\mathcal{E}} f$, that is, $P_{\mathcal{E}} f$ equals the orthogonal projection of $f - P_{\mathcal{M}} f$ onto $\mathcal{E}$. By applying Theorem 2.2.1 in Chapter 2 with $f - P_{\mathcal{M}} f$ replacing f and $g = \varphi$, we obtain

$$\begin{aligned}
E(f - P_{\mathcal{H}} f)(f - P_{\mathcal{H}} f)^* &= E((f - P_{\mathcal{M}} f) - P_{\mathcal{E}} f)((f - P_{\mathcal{M}} f) - P_{\mathcal{E}} f)^* \\
&= E((f - P_{\mathcal{M}} f) - P_{\mathcal{E}}(f - P_{\mathcal{M}} f))((f - P_{\mathcal{M}} f) - P_{\mathcal{E}}(f - P_{\mathcal{M}} f))^* \\
&= E(f - P_{\mathcal{M}} f)(f - P_{\mathcal{M}} f)^* - E(P_{\mathcal{E}} f)(P_{\mathcal{E}} f)^* \\
&= E(f - P_{\mathcal{M}} f)(f - P_{\mathcal{M}} f)^* - R_{f\varphi} R_{\varphi}^{-1} E\varphi\varphi^* R_{\varphi}^{-1} R_{f\varphi}^* \\
&= E(f - P_{\mathcal{M}} f)(f - P_{\mathcal{M}} f)^* - R_{f\varphi} R_{\varphi}^{-1} R_{f\varphi}^* \,.
\end{aligned}$$

This yields (3.2) and completes the proof.

Proof of the Kalman filtering Theorem 3.1.1. Now let us present a proof of the Kalman filter by implementing Lemma 3.3.1 with $\mathcal{H} = \mathcal{M}_n = \mathcal{M}_{n-1} \bigvee \mathcal{Y}$. In our setting $\mathcal{Y}$ is the span of $y(n)$, the subspace $\mathcal{M} = \mathcal{M}_{n-1}$ and $\varphi = \varphi(n) = y(n) - P_{\mathcal{M}_{n-1}} y(n)$. Recall that the solution to the difference equation in (1.3) is given by

$$x(n) = \Psi(n-1, -1)x(0) + \sum_{j=0}^{n-1} \Psi(n-1, j) B(j) u(j) \tag{3.3}$$

$$y(n) = C(n)\Psi(n-1, -1)x(0) + \sum_{j=0}^{n-1} C(n)\Psi(n-1, j) B(j) u(j) + D(n) v(n) \,. \tag{3.4}$$

Here $\Psi(n, \nu) = A(n)A(n-1)\cdots A(\nu+1)$ and $\Psi(k, k) = I$. This readily shows that

$$\mathcal{M}_n = \bigvee_{k=0}^{n} y_k \subset \bigvee \{x(0), u(0), u(1), \cdots, u(n-1), v(0), v(1), \cdots, v(n)\} \,. \tag{3.5}$$

Because $u(k)$ and $v(k)$ are independent white noise processes and orthogonal to $x(0)$, the random vector $v(n)$ is orthogonal to $\mathcal{M}_{n-1}$. In particular, $P_{\mathcal{M}_{n-1}} v(n) = 0$. Recall that the optimal state estimate is given by $\widehat{x}(n) = P_{\mathcal{M}_{n-1}} x(n)$. Using this we have

$$\begin{aligned}
\varphi(n) &= y(n) - P_{\mathcal{M}_{n-1}} y(n) = y(n) - P_{\mathcal{M}_{n-1}}(Cx(n) + Dv(n)) \\
&= y(n) - CP_{\mathcal{M}_{n-1}} x(n) = y(n) - C\widehat{x}(n) \,.
\end{aligned}$$

Hence $\varphi(n) = y(n) - C\widehat{x}(n)$. Recall that the state estimation error $\widetilde{x}(n) = x(n) - \widehat{x}(n)$. Since $y(n) = Cx(n) + Dv(n)$, we have $\varphi(n) = C\widetilde{x}(n) + Dv(n)$. This yields the following two useful formulas

$$\varphi(n) = y(n) - C\widehat{x}(n) = C\widetilde{x}(n) + Dv(n)\,. \tag{3.6}$$

By consulting (3.3) and (3.5), we see that $v(n)$ is orthogonal to both $x(n)$ and $\mathcal{M}_{n-1}$. Hence $v(n)$ is orthogonal to $\widetilde{x}(n) = x(n) - \widehat{x}(n)$. This and $\varphi(n) = C\widetilde{x}(n) + Dv(n)$ implies that

$$E\varphi(n)\varphi(n)^* = E\left(C\widetilde{x}(n) + Dv(n)\right)\left(C\widetilde{x}(n) + Dv(n)\right)^* = CE\widetilde{x}(n)\widetilde{x}(n)^*C^* + DD^*\,.$$

By definition $Q_n = E\widetilde{x}(n)\widetilde{x}(n)^*$ is the error covariance. Therefore

$$R_{\varphi(n)} = E\varphi(n)\varphi(n)^* = CQ_nC^* + DD^*\,. \tag{3.7}$$

Equation (3.5) shows that $u(n)$ is orthogonal to $\mathcal{M}_{n-1}$. In other words, $P_{\mathcal{M}_{n-1}}u(n) = 0$. By employing (3.1) in Lemma 3.3.1 with $\varphi = \varphi(n)$ and $y = y(n)$ and $\mathcal{M}_n = \mathcal{M}_{n-1} \bigvee y(n)$, we obtain

$$\begin{aligned}
\widehat{x}(n+1) &= P_{\mathcal{M}_n}x(n+1) = P_{\mathcal{M}_{n-1}}x(n+1) + R_{x(n+1)\varphi(n)}R^{-1}_{\varphi(n)}\varphi(n) \\
&= P_{\mathcal{M}_{n-1}}(Ax(n) + Bu(n)) + R_{x(n+1)\varphi(n)}R^{-1}_{\varphi(n)}\varphi(n) \\
&= A\widehat{x}(n) + R_{x(n+1)\varphi(n)}R^{-1}_{\varphi(n)}\varphi(n)\,.
\end{aligned} \tag{3.8}$$

We need an expression for $R_{x(n+1)\varphi(n)}$. Since $\widehat{x}(n)$ is contained in $\mathcal{M}_{n-1}$, the random vector $\varphi(n) = y(n) - C\widehat{x}(n)$ is contained in $\mathcal{M}_n$. Hence $\varphi(n)$ is orthogonal to $u(n)$; see (3.5). Moreover, $v(n)$ is orthogonal to $x(n)$; see (3.3). Using $\varphi(n) = C\widetilde{x}(n) + Dv(n)$, we have

$$\begin{aligned}
Ex(n+1)\varphi(n)^* &= E(Ax(n) + Bu(n))\varphi(n)^* = AEx(n)\varphi(n)^* \\
&= AEx(n)\left(C\widetilde{x}(n) + Dv(n)\right)^* = AEx(n)\widetilde{x}(n)^*C^* \\
&= AE(\widehat{x}(n) + \widetilde{x}(n))\widetilde{x}(n)^*C^* = AE\widetilde{x}(n)\widetilde{x}(n)^*C^* = AQ_nC^*\,.
\end{aligned}$$

The second from the last equality follows from the fact that $\widehat{x}(n)$ is orthogonal to $\widetilde{x}(n)$. The previous calculation yields the following result

$$R_{x(n+1)\varphi(n)} = Ex(n+1)\varphi(n)^* = AQ_nC^*\,. \tag{3.9}$$

Substituting $R_{x(n+1)\varphi(n)} = AQ_nC^*$ and the expression for $R_{\varphi(n)}$ in (3.7) into (3.8) yields

$$\widehat{x}(n+1) = A\widehat{x}(n) + AQ_nC^*(CQ_n^*C^* + DD^*)^{-1}\varphi(n)\,. \tag{3.10}$$

Finally, using $\varphi(n) = y(n) - C\widehat{x}(n)$ gives the state space formula for $\widehat{x}(n)$ in (1.4).

Now let us use equation (3.2) in Lemma 3.3.1 to derive the discrete time Riccati equation in (1.6). Recall that $u(n)$ is orthogonal to $\mathcal{M}_{n-1}$. Using $P_{\mathcal{M}_{n-1}}u(n) = 0$ along with the optimal state estimate $\widehat{x}(n) = P_{\mathcal{M}_{n-1}}x(n)$, we obtain

$$\begin{aligned}
x(n+1) - P_{\mathcal{M}_{n-1}}x(n+1) &= Ax(n) + Bu(n) - P_{\mathcal{M}_{n-1}}(Ax(n) + Bu(n)) \\
&= Ax(n) + Bu(n) - A\widehat{x}(n) = A\widetilde{x}(n) + Bu(n)\,.
\end{aligned}$$

This readily implies that

$$x(n+1) - P_{\mathcal{M}_{n-1}} x(n+1) = A\widetilde{x}(n) + Bu(n)\,. \tag{3.11}$$

By virtue of (3.3) we see that $u(n)$ is orthogonal to $x(n)$. Since $\widehat{x}(n) = P_{\mathcal{M}_{n-1}} x(n)$ is a vector in $\mathcal{M}_{n-1}$, the random vector $u(n)$ is also orthogonal to $\widehat{x}(n)$; see (3.5). Hence $u(n)$ is orthogonal to the error $\widetilde{x}(n) = x(n) - \widehat{x}(n)$. Using this fact in (3.11) along with $E\widetilde{x}(n)\widetilde{x}(n)^* = Q_n$, we arrive at

$$E\left(x(n+1) - P_{\mathcal{M}_{n-1}} x(n+1)\right)\left(x(n+1) - P_{\mathcal{M}_{n-1}} x(n+1)\right)^* = AQ_nA^* + BB^*\,. \tag{3.12}$$

Recall that $Ex(n+1)\varphi(n)^* = AQ_nC^*$; see (3.9). Finally, by employing equation (3.2) in Lemma 3.3.1 with the expression for $R_{\varphi(n)}$ in (3.7), we have

$$\begin{aligned} Q_{n+1} &= E\left(x(n+1) - P_{\mathcal{M}_n} x(n+1)\right)\left(x(n+1) - P_{\mathcal{M}_n} x(n+1)\right)^* \\ &= E\left(x(n+1) - P_{\mathcal{M}_{n-1}} x(n+1)\right)\left(x(n+1) - P_{\mathcal{M}_{n-1}} x(n+1)\right)^* \\ &\quad -Ex(n+1)\varphi(n)^* R_{\varphi(n)}^{-1} (Ex(n+1)\varphi(n)^*)^* \\ &= AQ_nA^* + BB^* - AQ_nC^*\left(CQ_nC^* + DD^*\right)^{-1} CQ_nA^*\,. \end{aligned}$$

This is precisely the Riccati difference equation in (1.6). To obtain the initial condition, recall that $\mathcal{M}_{-1} = 0$, that is, $y(-1) = 0$. Hence $\widetilde{x}(0) = x(0) - P_{\mathcal{M}_{-1}} x(0) = x(0)$. Thus $Q_0 = E\widetilde{x}(0)\widetilde{x}(0)^* = Ex(0)x(0)^*$. This completes the proof.

3.3.1 Exercise

Problem 1. Let f be a random vector with values in $\mathbb{C}^k$. Let g be a random vector with values in $\mathbb{C}^n$ and $\mathcal{M}$ the subspace spanned by g. Assume that R_g is invertible. According to Theorem 2.2.1 in Chapter 2, we have

$$P_{\mathcal{M}} f = R_{fg} R_g^{-1} g \quad \text{and} \quad E(f - P_{\mathcal{M}} f)(f - P_{\mathcal{M}} f)^* = R_f - R_{fg} R_g^{-1} R_{gf}\,. \tag{3.13}$$

Now let y be a random vector with values in $\mathbb{C}^m$, and set $\varphi = y - P_{\mathcal{M}} y$. Theorem 2.2.1 in Chapter 2 shows that

$$P_{\mathcal{M}} y = R_{yg} R_g^{-1} g \quad \text{and} \quad E(y - P_{\mathcal{M}} y)(y - P_{\mathcal{M}} y)^* = R_\varphi = R_y - R_{yg} R_g^{-1} R_{gy}\,. \tag{3.14}$$

As in Lemma 3.3.1, consider the space $\mathcal{H}$ spanned by g and y, that is, $\mathcal{H} = g \bigvee y$. Let h be the random vector defined by $h = \begin{bmatrix} g & y \end{bmatrix}^{tr}$ where tr denotes the transpose. Notice that $\mathcal{H}$ equals the span of h. Furthermore, R_h and R_{fh} are given by

$$R_h = \begin{bmatrix} R_g & R_{gy} \\ R_{yg} & R_y \end{bmatrix} \quad \text{and} \quad R_{fh} = \begin{bmatrix} R_{fg} & R_{fy} \end{bmatrix}. \tag{3.15}$$

Observe that the Schur complement for R_h is given by $\Delta = R_y - R_{yg} R_g^{-1} R_{gy} = R_\varphi$. In particular, R_h is invertible if and only if R_φ is invertible; see Lemma 2.4.1 in Chapter 2.

Now assume that R_φ is invertible. Then R_h is invertible and Theorem 2.2.1 in Chapter 2 implies that

$$P_{\mathcal{H}} f = R_{fh} R_h^{-1} h \qquad \text{and} \qquad E(f - P_{\mathcal{H}} f)(f - P_{\mathcal{H}} f)^* = R_f - R_{fh} R_h^{-1} R_{hf} \,.$$

Using the matrix inversion Lemma 2.4.1 in Chapter 2, give another proof of equations (3.1) and (3.2) in Lemma 3.3.1, that is, show that

$$\begin{aligned} P_{\mathcal{H}} f &= P_{\mathcal{M}} f + R_{f\varphi} R_\varphi^{-1} \varphi \\ E(f - P_{\mathcal{H}} f)(f - P_{\mathcal{H}} f)^* &= E(f - P_{\mathcal{M}} f)(f - P_{\mathcal{M}} f)^* - R_{f\varphi} R_\varphi^{-1} R_{\varphi f} \,. \end{aligned}$$

Problem 2. As in the Kalman filtering Theorem 3.4.1, consider the state space system

$$x(n+1) = Ax(n) + Bu(n) \qquad \text{and} \qquad y(n) = Cx(n) + Dv(n) \tag{3.16}$$

where $u(n)$ and $v(n)$ are independent white noise random processes, which are independent to the initial condition $x(0)$. Recall that $\mathcal{M}_n$ equals the linear span of $\{y(k)\}_0^n$ and the optimal state estimate in the Kalman filter is given by $\widehat{x}(n) = P_{\mathcal{M}_{n-1}} x(n)$. Find the state estimate $P_{\mathcal{M}_n} x(n)$ for $x(n)$ in terms of $\widehat{x}(n)$ and $y(n)$. Hint, according to Lemma 3.3.1, we have

$$P_{\mathcal{M}_n} f = P_{\mathcal{M}_{n-1}} f + R_{f\varphi(n)} R_{\varphi(n)}^{-1} \varphi(n) \,, \tag{3.17}$$

where f is any random vector, and $\varphi(n) = y(n) - P_{\mathcal{M}_{n-1}} y(n)$.

3.4 An innovations perspective of the Kalman filter

In this section we will present the innovation approach to the Kalman filter developed in Kailath [20]. The innovation derivation involves a Gram-Schmidt orthogonalization procedure, which leads to a proof of the Kalman filter. Finally, recall that if $\mathcal{L}$ is a subspace of a Hilbert space $\mathcal{M}$, then the orthogonal complement of $\mathcal{L}$ in $\mathcal{M}$ is denoted by $\mathcal{M} \ominus \mathcal{L}$, that is, $\mathcal{M} \ominus \mathcal{L} = \{f \in \mathcal{M} : f \perp \mathcal{L}\}$.

3.4.1 Innovations and Gram-Schmidt orthogonalization

Now let us review the Gram-Schmidt orthogonalization procedure in our setting. Let $\{y_k\}_0^\infty$ be a sequence of random vectors with values in $\mathcal{Y}$. Set $\mathcal{M}_{-1} = \{0\}$ and for $\nu \geq 0$, let $\mathcal{M}_\nu$ be the subspace defined by

$$\mathcal{M}_\nu = \bigvee \{y_0, y_1, y_2, \cdots, y_\nu\} \,. \tag{4.1}$$

Notice that $\mathcal{M}_\nu$ is the subspace of all random variables formed by the linear span of all the components of $\{y_k\}_0^\nu$. So the dimension of $\mathcal{M}_\nu$ is at most $(\nu+1)\dim \mathcal{Y}$. Moreover, the subspaces $\mathcal{M}_\nu$ are increasing, that is, $\mathcal{M}_k \subset \mathcal{M}_{k+1}$ for all integers $k \geq 0$. Let $P_{\mathcal{M}_k}$ be the orthogonal projection onto $\mathcal{M}_k$. Now set $\mathcal{E}_0 = \mathcal{M}_0$ and $\mathcal{E}_k = \mathcal{M}_k \ominus \mathcal{M}_{k-1}$ for all integers

$k \geq 1$. Notice that $\widetilde{y}_k = y_k - P_{\mathcal{M}_{k-1}} y_k$ is in $\mathcal{E}_k$. Furthermore, $\mathcal{E}_k = \bigvee \widetilde{y}_k$, that is, $\mathcal{E}_k$ is the subspace of random variables spanned by the components of $\widetilde{y}_k$. We claim that

$$\mathcal{M}_\nu = \bigoplus_{k=0}^{\nu} \mathcal{E}_k = \mathcal{E}_0 \oplus \mathcal{E}_1 \oplus \mathcal{E}_2 \oplus \cdots \oplus \mathcal{E}_\nu \,. \tag{4.2}$$

Because the subspaces $\mathcal{M}_k$ are increasing, we have

$$\mathcal{M}_\nu = (\mathcal{M}_\nu \ominus \mathcal{M}_{\nu-1}) \oplus \mathcal{M}_{\nu-1} = \mathcal{E}_\nu \oplus (\mathcal{M}_{\nu-1} \ominus \mathcal{M}_{\nu-2}) \oplus \mathcal{M}_{\nu-2} = \mathcal{E}_\nu \oplus \mathcal{E}_{\nu-1} \oplus \mathcal{M}_{\nu-2} \,.$$

By continuing in this fashion, we arrive at the orthogonal decomposition of $\mathcal{M}_\nu$ in (4.2).

Set $\varphi_0 = y_0$ and $\varphi_k = \widetilde{y}_k = y_k - P_{\mathcal{M}_{k-1}} y_k$ for all integers $k \geq 1$. The sequence of random vectors $\{\varphi_k\}_0^\infty$ is called the *innovations sequence* for $\{y_k\}_0^\infty$. Notice that $\varphi_k \in \mathcal{E}_k$ for all integers $k \geq 0$. Therefore innovations sequence $\{\varphi_k\}_0^\infty$ is an orthogonal sequence, that is, φ_k is orthogonal to φ_j for all $k \neq j$. However, the sequence $\{\varphi_k\}_0^\infty$ is not necessarily orthonormal. Moreover, by construction $\mathcal{E}_k = \bigvee \varphi_k$ for all $k \geq 0$. So according to (4.2), the innovations $\{\varphi_k\}_0^\infty$ is a sequence of orthogonal random vectors satisfying $\mathcal{M}_\nu = \bigvee_{k=0}^{\nu} \varphi_k$ for all integers $\nu \geq 0$.

The orthogonal decomposition of $\mathcal{M}_\nu$ in (4.2) shows that

$$P_{\mathcal{M}_\nu} = \sum_{k=0}^{\nu} P_{\mathcal{E}_k} \tag{4.3}$$

where $P_{\mathcal{E}_k}$ is the orthogonal projection onto $\mathcal{E}_k$. In particular, if x is any random variable with values in $\mathcal{X}$, then

$$P_{\mathcal{M}_\nu} x = \sum_{k=0}^{\nu} P_{\mathcal{E}_k} x \qquad (x \in \mathcal{X}) \,.$$

Now assume that R_{φ_k} is invertible for $k = 0, 1, \cdots, \nu$. Then Theorem 2.2.1 in Chapter 2 with $f = x$, $\mathcal{H} = \mathcal{E}_k$ and $g = \varphi_k = \widetilde{y}_k = y_k - P_{\mathcal{M}_{k-1}} y_k$ shows that $P_{\mathcal{E}_k} x = R_{x,\,\varphi_k} R_{\varphi_k}^{-1} \varphi_k$. Using this we obtain

$$P_{\mathcal{M}_\nu} x = \sum_{k=0}^{\nu} R_{x,\,\varphi_k} R_{\varphi_k}^{-1} \varphi_k \qquad (x \in \mathcal{X}) \,. \tag{4.4}$$

In particular, $P_{\mathcal{M}_\nu} x = \sum_{k=0}^{\nu} H_k \varphi_k$ where $H_k = R_{x,\,\varphi_k} R_{\varphi_k}^{-1}$ is a linear operator from $\mathcal{Y}$ into $\mathcal{X}$. Finally, it is noted that the expression for $P_{\mathcal{M}_\nu} x$ in (4.4) is a generalization of the method for computing the orthogonal projection by the Gram-Schmidt orthogonalization procedure.

An example of Gram-Schmidt orthogonalization

Let $\mathbf{y}$ be a uniform random variable over the interval $[0, 1]$. Recall that the probability density function for $\mathbf{y}$ is given by

$$\begin{aligned} f_{\mathbf{y}}(y) &= 1 \qquad \text{if } 0 \leq y \leq 1 \\ &= 0 \qquad \text{otherwise} \,. \end{aligned}$$

Notice that we have used a boldface $\mathbf{y}$ for the random variable $\mathbf{y}$. The k-moment $E\mathbf{y}^k$ for $\mathbf{y}$ is given by

$$E\mathbf{y}^k = 1/(k+1) \qquad \text{(for all integers } k \geq 0)\,. \tag{4.5}$$

To see this simply observe that

$$E\mathbf{y}^k = \int_{-\infty}^{\infty} y^k f_{\mathbf{y}}(y)dy = \int_0^1 y^k dy = \frac{1}{k+1}\,.$$

Therefore (4.5) holds.

Consider the random variables $y_0 = 1$, $y_1 = \mathbf{y}$ and $y_2 = \mathbf{y}^2$. In this case, $\mathcal{M}_0 = \text{span}\{1\}$ and $\mathcal{M}_1 = \text{span}\{1, y_1\}$ while $\mathcal{M}_2 = \text{span}\{1, y_1, y_2\}$. Now let us compute the innovations $\{\varphi_0, \varphi_1, \varphi_2\}$ corresponding to $\{y_0, y_1, y_2\}$, or equivalently, $\{1, \mathbf{y}, \mathbf{y}^2\}$. First notice that $\varphi_0 = y_0$, that is, $\varphi_0 = 1$. Clearly, $R_{\varphi_0} = E1^2 = 1$. For $\nu = 0$ equation (4.4) reduces to

$$P_{\mathcal{M}_0} f = R_{f\varphi_0} R_{\varphi_0}^{-1} \varphi_0 = (Ef\varphi_0^*)\varphi_0 = Ef$$

where f is any random variable. In other words, $P_{\mathcal{M}_0} f = Ef$. This readily implies that

$$\varphi_1 = y_1 - P_{\mathcal{M}_0} y_1 = y_1 - P_{\mathcal{M}_0}\mathbf{y} = \mathbf{y} - E\mathbf{y} = \mathbf{y} - 1/2\,.$$

In other words, $\varphi_1 = \mathbf{y} - 1/2$. Moreover, we have

$$R_{\varphi_1} = E\varphi_1^2 = E(\mathbf{y} - 1/2)^2 = \int_{-\infty}^{\infty} (y - 1/2)^2 f_{\mathbf{y}}(y)dy = \int_0^1 (y - 1/2)^2 dy = \frac{1}{12}\,.$$

This readily implies that

$$\varphi_1 = \mathbf{y} - 1/2 \qquad \text{and} \qquad R_{\varphi_1} = 1/12\,. \tag{4.6}$$

Recall that $P_{\mathcal{M}_1} f = R_{f\varphi_0} R_{\varphi_0}^{-1} \varphi_0 + R_{f\varphi_1} R_{\varphi_1}^{-1} \varphi_1$ where f is a random variable; see (4.4) with $\nu = 1$. In our setting $R_{f\varphi_0} R_{\varphi_0}^{-1} \varphi_0 = Ef$. Using this and (4.5), we obtain

$$\begin{aligned}
\varphi_2 &= y_2 - P_{\mathcal{M}_1} y_2 = \mathbf{y}^2 - R_{\mathbf{y}^2\varphi_0} R_{\varphi_0}^{-1} \varphi_0 - R_{\mathbf{y}^2\varphi_1} R_{\varphi_1}^{-1} \varphi_1 = \mathbf{y}^2 - E\mathbf{y}^2 - \frac{(E\mathbf{y}^2\varphi_1)}{R_{\varphi_1}}\varphi_1 \\
&= \mathbf{y}^2 - E\mathbf{y}^2 - 12(E\mathbf{y}^2(\mathbf{y} - 1/2))\varphi_1 = \mathbf{y}^2 - 1/3 - 12(1/4 - 1/6)\varphi_1 \\
&= \mathbf{y}^2 - 1/3 - \varphi_1 = \mathbf{y}^2 - 1/3 - (\mathbf{y} - 1/2) = \mathbf{y}^2 - \mathbf{y} + 1/6\,.
\end{aligned}$$

Hence $\varphi_2 = \mathbf{y}^2 - \mathbf{y} + 1/6$. Finally, observe that

$$R_{\varphi_2} = E\varphi_2^2 = E(\mathbf{y}^2 - \mathbf{y} + 1/6)^2 = \int_{-\infty}^{\infty} (y^2 - y + 1/6)^2 f_{\mathbf{y}}(y)dy = \int_0^1 (y^2 - y + 1/6)^2 dy = \frac{1}{180}\,.$$

In other words, $R_{\varphi_2} = 1/180$. The previous analysis readily shows that the innovations $\{\varphi_0, \varphi_1, \varphi_2\}$ corresponding to $\{y_0, y_1, y_2\}$, or equivalently, $\{1, \mathbf{y}, \mathbf{y}^2\}$ are given by

$$\begin{array}{lcll}
\varphi_0 &=& 1 & \text{and} \quad R_{\varphi_0} = 1 \\
\varphi_1 &=& \mathbf{y} - 1/2 & \text{and} \quad R_{\varphi_1} = 1/12 \\
\varphi_2 &=& \mathbf{y}^2 - \mathbf{y} + 1/6 & \text{and} \quad R_{\varphi_2} = 1/180\,.
\end{array}$$

Now consider the random variable $x = e^{\mathbf{y}}$. Let us use the innovations to compute the orthogonal projection $\widehat{x} = P_{\mathcal{M}_2}x$. Recall that $\mathcal{M}_2 = \text{span}\{1, y_1, y_2\} = \text{span}\{\varphi_0, \varphi_1, \varphi_2\}$. According to (4.4), we have

$$\begin{aligned}
\widehat{x} &= P_{\mathcal{M}_2}x = \sum_{k=0}^{2} R_{x\varphi_k} R_{\varphi_k}^{-1} \varphi_k = \sum_{k=0}^{2} \frac{(Ex\varphi_k)}{R_{\varphi_k}} \varphi_k \\
&= (Ee^{\mathbf{y}})\,\varphi_0 + 12\,(Ee^{\mathbf{y}}(\mathbf{y} - 1/2))\,\varphi_1 + 180\left(Ee^{\mathbf{y}}(\mathbf{y}^2 - \mathbf{y} + 1/6)\right)\varphi_2 \\
&= \left(\int_0^1 e^y dy\right)\varphi_0 + 12\left(\int_0^1 e^y(y - 1/2)dy\right)\varphi_1 + 180\left(\int_0^1 e^y(y^2 - y + 1/6)dy\right)\varphi_2 \\
&= 1.72 + 1.69\varphi_1 + 0.84\varphi_2\,.
\end{aligned}$$

Therefore the optimal estimate $\widehat{x} = P_{\mathcal{M}_2}x$ is given by

$$\begin{aligned}
\widehat{x} &= P_{\mathcal{M}_2}x = 1.72 + 1.69\varphi_1 + 0.84\varphi_2 = 1.72 + 1.69(\mathbf{y} - 1/2) + 0.84(\mathbf{y}^2 - \mathbf{y} + 1/6) \\
&= 0.84\mathbf{y}^2 + 0.85\mathbf{y} + 1.01\,. \qquad (4.7)
\end{aligned}$$

Finally, it is noted that $\widehat{x}$ is precisely the same estimate of x computed from Problem 2 in Exercise 2.2.2 in Chapter 2. This is also equivalent to the approximation of e^t in presented in Section 1.3.1 in Chapter 1.

3.4.2 An innovations derivation of the Kalman filter

In this section we will use the innovations process corresponding to the output $y(n)$ to derive the Kalman filter. As before, consider the state space system given by

$$x(n+1) = Ax(n) + Bu(n) \qquad \text{and} \qquad y(n) = Cx(n) + Dv(n)\,. \qquad (4.8)$$

Recall that the input $u(n)$ and $v(n)$ are independent white noise random process. Moreover, we assume that the initial condition x_0, $u(n)$ and $v(m)$ are orthogonal for all integers n and m. As before, let $\mathcal{M}_k$ be the subspace generated by the random vectors $\{y(j)\}_0^k$, that is, $\mathcal{M}_k = \bigvee_{j=0}^k y(j)$ and $\mathcal{M}_{-1} = \{0\}$. The Kalman filtering problem is to compute the best estimate $\widehat{x}(k) = P_{\mathcal{M}_{k-1}}x(k)$ of the state $x(k)$ given the past output $\{y(j)\}_0^{k-1}$. For convenience let us restate the Kalman filtering result.

THEOREM 3.4.1 *Consider the state space system*

$$x(n+1) = Ax(n) + Bu(n) \qquad \textit{and} \qquad y(n) = Cx(n) + Dv(n) \qquad (4.9)$$

where $u(n)$ and $v(n)$ are independent white noise random processes, which are independent to the initial condition $x(0)$. Then the optimal estimate $\widehat{x}(k) = P_{\mathcal{M}_{k-1}}x(k)$ of the state $x(k)$ given the past $\{y(j)\}_0^{k-1}$ is recursively computed by

$$\begin{aligned}
\widehat{x}(n+1) &= A\widehat{x}(n) + \Delta_n(y(n) - C\widehat{x}(n)) \qquad (4.10) \\
\Delta_n &= AQ_nC^*\left(CQ_nC^* + DD^*\right)^{-1}\,. \qquad (4.11)
\end{aligned}$$

The state covariance error $Q_k = E(x(k) - \widehat{x}(k))(x(k) - \widehat{x}(k))^*$ is recursively computed by solving the Riccati difference equation

$$Q_{n+1} = (A - \Delta_n C)Q_n(A - \Delta_n C)^* + BB^* + \Delta_n DD^*\Delta_n^*, \tag{4.12}$$

where $Q_0 = Ex(0)x(0)^*$.

PROOF. Recall that the solution to the difference equation in (4.8) is given by

$$x(n) = \Psi(n-1,-1)x(0) + \sum_{j=0}^{n-1} \Psi(n-1,j)B(j)u(j) \tag{4.13}$$

$$y(n) = C(n)\Psi(n-1,-1)x(0) + \sum_{j=0}^{n-1} C(n)\Psi(n-1,j)B(j)u(j) + D(n)v(n)\,. \tag{4.14}$$

Here $\Psi(n,\nu) = A(n)A(n-1)\cdots A(\nu+1)$ and $\Psi(k,k) = I$. This readily shows that

$$\mathcal{M}_n = \bigvee_{k=0}^{n} y_k \subset \bigvee\{x(0), u(0), u(1), \cdots, u(n-1), v(0), v(1), \cdots, v(n)\}\,. \tag{4.15}$$

Let $\varphi(n)$ be the innovations process for the output defined by

$$\varphi(n) = y(n) - P_{\mathcal{M}_{n-1}} y(n) \qquad (n \geq 0)\,. \tag{4.16}$$

By construction $\mathcal{M}_{n-1} = \oplus_0^{n-1}\mathcal{E}_k$ where $\mathcal{E}_k = \bigvee \varphi(k)$ for $k = 0, 1, 2, \cdots n-1$. Equation (4.15) with $n-1$ replacing n shows that $v(n)$ is orthogonal to $\mathcal{M}_{n-1}$. In other words, $P_{\mathcal{M}_{n-1}} v(n) = 0$. Using $y(n) = Cx(n) + Dv(n)$ we have

$$\varphi(n) = y(n) - P_{\mathcal{M}_{n-1}} y(n) = y(n) - P_{\mathcal{M}_{n-1}}(Cx(n) + Dv(n)) = y(n) - C\widehat{x}(n)\,.$$

Hence $\varphi(n) = y(n) - C\widehat{x}(n)$. Since $\widetilde{x}(n) = x(n) - \widehat{x}(n)$ and $y(n) = Cx(n) + Dv(n)$, we have $\varphi(n) = C\widetilde{x}(n) + Dv(n)$. This yields the following two useful formulas for the innovations

$$\varphi(n) = y(n) - C\widehat{x}(n) = C\widetilde{x}(n) + Dv(n)\,. \tag{4.17}$$

By consulting (4.13) and (4.15) we see that $v(n)$ is orthogonal to both $x(n)$ and $\mathcal{M}_{n-1}$. Hence $v(n)$ is orthogonal to $\widetilde{x}(n) = x(n) - \widehat{x}(n)$. This and $\varphi(n) = C\widetilde{x}(n) + Dv(n)$ implies that

$$E\varphi(n)\varphi(n)^* = E\left(C\widetilde{x}(n) + Dv(n)\right)\left(C\widetilde{x}(n) + Dv(n)\right)^* = CE\widetilde{x}(n)\widetilde{x}(n)^*C^* + DD^*\,.$$

By definition $Q_n = E\widetilde{x}(n)\widetilde{x}(n)^*$ is the error covariance. Thus

$$E\varphi(n)\varphi(n)^* = CQ_nC^* + DD^*\,. \tag{4.18}$$

By consulting (4.4) with $x = x(n)$, the innovation $\varphi(k) = \varphi_k$ and $\nu = n-1$, we see that

$$\widehat{x}(n) = P_{\mathcal{M}_{n-1}} x(n) = \sum_{k=0}^{n-1} H(n,k)\varphi(k) \tag{4.19}$$

where $H(n,k)$ is the operator mapping $\mathcal{Y}$ into $\mathcal{X}$ defined by

$$\begin{aligned} H(n,k) &= E\left(x(n)\varphi(k)^*\right)\left(E\varphi(k)\varphi(k)^*\right)^{-1} \\ &= E\left(x(n)\varphi(k)^*\right)\left(CQ_kC^*+DD^*\right)^{-1} \qquad (k=0,1,\cdots,n-1)\,. \end{aligned} \tag{4.20}$$

The last equality follows from (4.18). By virtue of (4.15), we see that $u(n)$ is orthogonal to $\mathcal{M}_n = \oplus_0^n \mathcal{E}_k$. Since $\mathcal{E}_k = \bigvee \varphi(k)$, the random input $u(n)$ is orthogonal to $\varphi(k)$ for all $k = 0, 1, \cdots, n$. Now let $G_k = (CQ_kC^* + DD^*)^{-1}$. Then using (4.19) with the state space equation $x(n+1) = Ax(n) + Bu(n)$, we obtain

$$\begin{aligned} \widehat{x}(n+1) &= P_{\mathcal{M}_n}x(n+1) = \sum_{k=0}^{n} H(n+1,k)\varphi(k) \\ &= H(n+1,n)\varphi(n) + \sum_{k=0}^{n-1} H(n+1,k)\varphi(k) \\ &= E\left(x(n+1)\varphi(n)^*\right)G_n\varphi(n) + \sum_{k=0}^{n-1} E\left(x(n+1)\varphi(k)^*\right)G_k\varphi(k) \\ &= E\left((Ax(n)+Bu(n))\varphi(n)^*\right)G_n\varphi(n) + \sum_{k=0}^{n-1} E\left((Ax(n)+Bu(n))\varphi(k)^*\right)G_k\varphi(k) \\ &= AE\left(x(n)\varphi(n)^*\right)G_n\varphi(n) + A\sum_{k=0}^{n-1} E\left(x(n)\varphi(k)^*\right)G_k\varphi(k) \\ &= AE\left(x(n)\varphi(n)^*\right)G_n\varphi(n) + A\sum_{k=0}^{n-1} H(n,k)\varphi(k) \\ &= A\widehat{x}(n) + AE\left(x(n)\varphi(n)^*\right)G_n\varphi(n)\,. \end{aligned}$$

In other words,

$$\widehat{x}(n+1) = A\widehat{x}(n) + AE\left(x(n)\varphi(n)^*\right)\left(CQ_nC^*+DD^*\right)^{-1}\varphi(n)\,. \tag{4.21}$$

Using $\varphi(n) = C\widetilde{x}(n) + Dv(n)$ along with the fact that $v(n)$ is orthogonal to $x(n)$, we have

$$\begin{aligned} Ex(n)\varphi(n)^* &= Ex(n)\left(C\widetilde{x}(n)+Dv(n)\right)^* = Ex(n)\widetilde{x}(n)^*C^* \\ &= E(\widehat{x}(n)+\widetilde{x}(n))\widetilde{x}(n)^*C^* = E\widetilde{x}(n)\widetilde{x}(n)^*C^* = Q_nC^*\,. \end{aligned} \tag{4.22}$$

Substituting $Ex(n)\varphi(n)^* = Q_nC^*$ into (4.21) yields

$$\widehat{x}(n+1) = A\widehat{x}(n) + AQ_nC^*(CQ_nC^*+DD^*)^{-1}\varphi(n) = A\widehat{x}(n) + \Delta_n\varphi(n)\,. \tag{4.23}$$

Finally, using $\varphi(n) = y(n) - C\widehat{x}(n)$ gives the state space formula for $\widehat{x}(n)$ in (4.10).

To complete the proof it remains to derive the Riccati difference equation in (4.12). By using $\varphi(n) = C\widetilde{x}(n) + Dv(n)$ in (4.23), we arrive at

$$\widehat{x}(n+1) = A\widehat{x}(n) + \Delta_nC\widetilde{x}(n) + \Delta_nDv(n)\,.$$

Subtracting this from $x(n+1) = Ax(n) + Bu(n)$ yields

$$\widetilde{x}(n+1) = (A - \Delta_n C)\,\widetilde{x}(n) + Bu(n) - \Delta_n Dv(n)\,. \tag{4.24}$$

Using the fact that $\widetilde{x}(n)$, $u(n)$ and $v(n)$ are all orthogonal, we obtain

$$\begin{aligned} Q_{n+1} &= E\widetilde{x}(n+1)\widetilde{x}(n+1)^* \\ &= E\left((A-\Delta_n C)\widetilde{x}(n) + Bu(n) - \Delta_n Dv(n)\right) \\ &\quad \left((A-\Delta_n C)\widetilde{x}(n) + Bu(n) - \Delta_n Dv(n)\right)^* \\ &= (A-\Delta_n C)E\widetilde{x}(n)\widetilde{x}(n)^*(A-\Delta_n C)^* + BB^* + \Delta_n DD^*\Delta_n^* \\ &= (A-\Delta_n C)Q_n(A-\Delta_n C)^* + BB^* + \Delta_n DD^*\Delta_n^*\,. \end{aligned}$$

This is precisely the Riccati difference equation in (4.12). Proposition 3.4.3 below shows that the solutions to the Riccati equations in (4.12) and (1.6) are equivalent. This completes the proof.

REMARK 3.4.2 *As in Theorem 3.4.1, consider the state space system determined by*

$$x(n+1) = Ax(n) + Bu(n) \qquad \text{and} \qquad y(n) = Cx(n) + Dv(n) \tag{4.25}$$

where $u(n)$ and $v(n)$ are independent white noise random processes, which are independent to the initial condition $x(0)$. Our previous analysis shows that the Kalman filter in (4.10) can be used to recursively compute the innovations $\varphi(n) = y(n) - C\widehat{x}(n)$ for the process $y(n)$.

Let us complete this section with the following result.

PROPOSITION 3.4.3 *Let Q_n be the solution for the Riccati difference equation in (1.6) associated with $\{A, B, C, D\}$. Then Q_n is also a solution to the following Riccati difference equation*

$$\begin{aligned} Q_{n+1} &= (A-\Delta_n C)Q_n(A-\Delta_n C)^* + BB^* + \Delta_n DD^*\Delta_n^* \\ \Delta_n &= AQ_nC^*\left(CQ_nC^* + DD^*\right)^{-1}\,. \end{aligned} \tag{4.26}$$

In particular, if the initial condition Q_0 is positive, then this also shows that Q_n is positive for all integers $n \geq 0$.

Proof. By consulting the form of the Riccati difference equation in (1.6), we obtain

$$\begin{aligned} Q_{n+1} &= AQ_nA^* - AQ_nC^*\left(CQ_nC^* + DD^*\right)^{-1}CQ_nA^* + BB^* \\ &= AQ_nA^* - \Delta_n CQ_nA^* + BB^* \\ &= (A-\Delta_n C)Q_nA^* + BB^* \\ &= (A-\Delta_n C)Q_n(A-\Delta_n C)^* + (A-\Delta_n C)Q_nC^*\Delta_n^* + BB^* \\ &= (A-\Delta_n C)Q_n(A-\Delta_n C)^* + AQ_nC^*\Delta_n^* - \Delta_n(CQ_nC^* + DD^*)\Delta_n^* \\ &\quad + \Delta_n DD^*\Delta_n^* + BB^* \\ &= (A-\Delta_n C)Q_n(A-\Delta_n C)^* + \Delta_n DD^*\Delta_n^* + BB^*\,. \end{aligned}$$

This yields (4.26) and completes the proof.

3.4.3 Exercise

Problem 1. Let $\mathbf{y}$ be a uniform random variable over the interval $[0,1]$. Let $\{y_k\}_{k=0}^3$ be the random variables defined by

$$y_0 = 1, \quad y_1 = \mathbf{y}, \quad y_2 = \mathbf{y}^2 \quad \text{and} \quad y_3 = \mathbf{y}^3 .$$

Then compute the innovations $\{\varphi_k\}_{k=0}^3$ for $\{y_k\}_{k=0}^3$. In other words, compute the orthogonal random variables $\varphi_k = y_k - P_{\mathcal{M}_{k-1}} y_k$ for $k = 0, 1, 2, 3$ where $\mathcal{M}_n = \text{span}\{y_k\}_{k=0}^n$. Let $x = e^{\mathbf{y}}$ and let $\widehat{x} = P_{\mathcal{M}_3} x$. Then express $\widehat{x}$ as a linear combination of $\{\varphi_k\}_{k=0}^3$, that is, find the constants $\{\alpha_k\}_{k=0}^3$ such that

$$\widehat{x} = P_{\mathcal{M}_3} x = \sum_{k=0}^{3} \alpha_k \varphi_k .$$

Matlab can be used to compute your answer.

Problem 2. Let $\mathbf{y}$ be a uniform random variable over the interval $[0,1]$. Let $\{y_k\}_{k=0}^\infty$ be the random variables defined by $y_k = \mathbf{y}^k$ for all integers $k \geq 0$. Consider the functions $\{\phi_k\}_{k=0}^\infty$ defined by

$$\phi_k(y) = \frac{k!}{(2k)!} \frac{d^k}{dy^k} y^k (y-1)^k \qquad (k \geq 0) .$$

Then show that $\varphi_k = \phi_k(\mathbf{y})$ is the innovations for $\{y_k\}_{k=0}^\infty$. In other words, show that

$$\phi_k(\mathbf{y}) = y_k - P_{\mathcal{M}_{k-1}} y_k \qquad \text{(for all integers } k \geq 0)$$

where $\mathcal{M}_n = \text{span}\{y_k\}_{k=0}^n$.

3.5 Kalman smoothing

Smoothing is estimating the past of one process given the present and past of another process. This section is devoted to Kalman smoothing, that is, estimating the past state given the present and past output measurements. As before, consider the state space system given by

$$x(n+1) = Ax(n) + Bu(n) \qquad \text{and} \qquad y(n) = Cx(n) + Dv(n) \tag{5.1}$$

where $u(n)$ and $v(n)$ are independent white noise processes which are also independent to the initial state $x(0)$. Now consider any integer $k \geq 0$. The Kalman smoothing problem is to compute the best estimate $\widehat{x}(n-k|n)$ of the state $x(n-k)$ given the past output $\{y(j)\}_0^n$. To be precise, let $\mathcal{M}_n$ be the subspace generated by the random vectors $\{y(j)\}_0^n$, that is, $\mathcal{M}_n = \bigvee_{j=0}^n y(j)$. Let $P_{\mathcal{M}_n}$ be the orthogonal projection onto $\mathcal{M}_n$ for all integers $n \geq 0$. Then the optimal state estimate $\widehat{x}(n-k|n)$ of the past state $x(n-k)$ given $\{y(j)\}_0^n$ is determined by the orthogonal projection

$$\widehat{x}(n-k|n) = P_{\mathcal{M}_n} x(n-k) . \tag{5.2}$$

Throughout this section $\widehat{x}(n) = P_{\mathcal{M}_{n-1}}x(n)$ is the optimal state estimate $\widehat{x}(n)$ for $x(n)$ computed by the Kalman filter in Theorem 3.1.1.

To describe our solution to the Kalman smoothing problem, let Q_n be the solution to the discrete time Riccati equation in (1.6) or (4.12), and set

$$\Delta_n = A(n)Q_nC(n)^*G_n \quad \text{where} \quad G_n = \left(C(n)Q_nC(n)^* + D(n)D(n)^*\right)^{-1} . \tag{5.3}$$

Let F_n be the feedback operator defined by

$$F_n \;=\; A(n) - \Delta_n C(n)\,. \tag{5.4}$$

Finally, set $C_m = C(m)$ for all integers $m \geq 0$, and let $W_{n,\,k}$ be the observability operator defined by

$$W_{n,\,k} = \begin{bmatrix} G_{n-k}C_{n-k} \\ G_{n-k+1}C_{n-k+1}F_{n-k} \\ G_{n-k+2}C_{n-k+2}F_{n-k+1}F_{n-k} \\ G_{n-k+3}C_{n-k+3}F_{n-k+2}F_{n-k+1}F_{n-k} \\ \vdots \\ G_nC_nF_{n-1}F_{n-2}\cdots F_{n-k+1}F_{n-k} \end{bmatrix} . \tag{5.5}$$

The following Kalman smoothing result provides us with a recursive algorithm to compute the optimal state estimate $\widehat{x}(n-k|n)$ of the past state.

THEOREM 3.5.1 *Consider the state space system*

$$x(n+1) = Ax(n) + Bu(n) \qquad \textit{and} \qquad y(n) = Cx(n) + Dv(n) \tag{5.6}$$

where $u(n)$ and $v(n)$ are independent white noise random processes, which are independent to the initial condition $x(0)$. Let $\widehat{x}(n) = P_{\mathcal{M}_{n-1}}x(n)$ be the optimal state estimate of $x(n)$ given the past $\{y(j)\}_0^{n-1}$ recursively computed by (1.4), (1.5) and (1.6) in the Kalman filtering Theorem 3.1.1. Then the optimal state estimate $\widehat{x}(n-k|n)$ of $x(n-k)$ given the past $\{y(j)\}_0^n$ is computed by

$$\widehat{x}(n-k|n) = P_{\mathcal{M}_n}x(n-k) = \widehat{x}(n-k) + Q_{n-k}W_{n,\,k}^* \begin{bmatrix} \varphi(n-k) \\ \varphi(n-k+1) \\ \vdots \\ \varphi(n) \end{bmatrix} \tag{5.7}$$

where $\varphi(m) = y(m) - C\widehat{x}(m)$ and $k \geq 0$ is an integer.

Proof. According to Lemma 3.3.1 or (4.4), we have

$$P_{\mathcal{M}_n}f = P_{\mathcal{M}_{n-1}}f + R_{f\varphi(n)}R_{\varphi(n)}^{-1}\varphi(n)\,, \tag{5.8}$$

where f is any random vector, and $\varphi(n) = y(n) - P_{\mathcal{M}_{n-1}}y(n)$. By consulting (3.7) or (4.18) and the definition of G_n in (5.3), we see that $R_{\varphi(n)}^{-1} = G_n$. By employing (5.8) with $f = x(n)$, we have

$$P_{\mathcal{M}_n}x(n) = P_{\mathcal{M}_{n-1}}x(n) + R_{x(n)\varphi(n)}R_{\varphi(n)}^{-1}\varphi(n) = \widehat{x}(n) + \left(Ex(n)\varphi(n)^*\right)G_n\varphi(n)\,. \tag{5.9}$$

Equation (4.22) shows that $Ex(n)\varphi(n)^* = Q_n C_n^*$. Substituting this into (5.9) yields

$$P_{\mathcal{M}_n} x(n) = \widehat{x}(n) + Q_n C_n^* G_n \varphi(n)\,. \tag{5.10}$$

Since $W_{n,0} = G_n C_n$, we see that (5.7) holds for $k = 0$. Here we used that fact that G_n is a self-adjoint operator.

At this point one can use induction to complete the proof. However, to gain some further insight, let compute the next step $P_{\mathcal{M}_n} x(n-1)$. By employing (5.8) with $f = x(n-1)$, we have

$$P_{\mathcal{M}_n} x(n-1) = P_{\mathcal{M}_{n-1}} x(n-1) + R_{x(n-1)\varphi(n)} R^{-1}_{\varphi(n)} \varphi(n)\,. \tag{5.11}$$

Notice that $P_{\mathcal{M}_{n-1}} x(n-1)$ can be computed by replacing n by $n-1$ in (5.10), that is,

$$P_{\mathcal{M}_{n-1}} x(n-1) = \widehat{x}(n-1) + Q_{n-1} C_{n-1}^* G_{n-1} \varphi(n-1)\,. \tag{5.12}$$

To compute the last term in (5.11), recall that $\varphi(n) = C_n \widetilde{x}(n) + Dv(n)$; see (4.17). Moreover, equation (4.13) shows that $v(n)$, $v(n-1)$ and $u(n)$ are all orthogonal to $x(n-1)$. The difference equation in (4.24) can be written as

$$\widetilde{x}(n+1) = F_n \widetilde{x}(n) + B(n)u(n) - \Delta_n D(n) v(n)\,. \tag{5.13}$$

Using this we have

$$\begin{aligned}
Ex(n-1)\varphi(n)^* &= Ex(n-1)\left(C_n\widetilde{x}(n) + Dv(n)\right)^* = Ex(n-1)\widetilde{x}(n)^* C_n^* \\
&= Ex(n-1)(F_{n-1}\widetilde{x}(n-1))^* C_n^* \\
&+ Ex(n-1)\left(B(n-1)u(n-1) - \Delta_{n-1} D(n-1) v(n-1)\right)^* C_n^* \\
&= E\left(\widehat{x}(n-1) + \widetilde{x}(n-1)\right)\widetilde{x}(n-1)^* F_{n-1}^* C_n^* \\
&= E\widetilde{x}(n-1)\widetilde{x}(n-1)^* F_{n-1}^* C_n^* \\
&= Q_{n-1} F_{n-1}^* C_n^*\,.
\end{aligned}$$

This readily implies that

$$Ex(n-1)\varphi(n)^* = Q_{n-1} F_{n-1}^* C_n^*\,. \tag{5.14}$$

Substituting (5.12) and (5.14) into (5.11) yields the following result

$$P_{\mathcal{M}_n} x(n-1) = \widehat{x}(n-1) + Q_{n-1}\left(C_{n-1}^* G_{n-1}\varphi(n-1) + F_{n-1}^* C_n^* G_n \varphi(n)\right)\,.$$

This is precisely (5.7) when $k = 1$.

To complete the proof let us use induction. Assume that (5.7) holds for some $k > 0$. By employing (5.8) with $f = x(n-k-1)$, we have

$$P_{\mathcal{M}_n} x(n-k-1) = P_{\mathcal{M}_{n-1}} x(n-k-1) + R_{x(n-k-1)\varphi(n)} R^{-1}_{\varphi(n)} \varphi(n)\,. \tag{5.15}$$

Notice that $P_{\mathcal{M}_{n-1}} x(n-1-k)$ can be computed by replacing n by $n-1$ in (5.7), that is,

$$P_{\mathcal{M}_{n-1}} x(n-k-1) = \widehat{x}(n-k-1) + Q_{n-k-1} W_{n-1,k}^* \begin{bmatrix} \varphi(n-k-1) \\ \varphi(n-k) \\ \vdots \\ \varphi(n-1) \end{bmatrix}\,. \tag{5.16}$$

To compute the last term in (5.15), observe that the solution to the difference equation in (5.13) is given by

$$\widetilde{x}(n) = F_{n-1}F_{n-2}\cdots F_{n-k-1}\widetilde{x}(n-k-1) + \sum_{j=n-k-1}^{n-1} L_j\left(B(j)u(j) - \Delta_j D(j)v(j)\right) \tag{5.17}$$

where L_j is a linear operator formed by products containing the appropriate F_i terms. Moreover, (4.13) shows that $u(j)$ and $v(j)$ are orthogonal to $x(n-k-1)$ for all $j \geq n-k-1$. In particular, the sum in (5.17) is orthogonal to $x(n-k-1)$. Using this fact we obtain

$$\begin{aligned}
Ex(n-k-1)\varphi(n)^* &= Ex(n-k-1)\left(C_n\widetilde{x}(n) + Dv(n)\right)^* = Ex(n-k-1)\widetilde{x}(n)^*C_n^* \\
&= Ex(n-k-1)\widetilde{x}(n-k-1)^*\left(F_{n-k-1}^*F_{n-k}^*\cdots F_{n-1}^*\right)C_n^* \\
&= E\widehat{x}(n-k-1)\widetilde{x}(n-k-1)^*\left(F_{n-k-1}^*F_{n-k}^*\cdots F_{n-1}^*\right)C_n^* \\
&\quad + E\widetilde{x}(n-k-1)\widetilde{x}(n-k-1)^*\left(F_{n-k-1}^*F_{n-k}^*\cdots F_{n-1}^*\right)C_n^* \\
&= E\widetilde{x}(n-k-1)\widetilde{x}(n-k-1)^*\left(F_{n-k-1}^*F_{n-k}^*\cdots F_{n-1}^*\right)C_n^* \\
&= Q_{n-k-1}\left(F_{n-k-1}^*F_{n-k}^*\cdots F_{n-1}^*\right)C_n^*.
\end{aligned}$$

This readily implies that

$$Ex(n-k-1)\varphi(n)^* = Q_{n-k-1}\left(F_{n-k-1}^*F_{n-k}^*\cdots F_{n-1}^*\right)C_n^*. \tag{5.18}$$

Substituting (5.16) and (5.18) into (5.15) yields the following result

$$\begin{aligned}
P_{\mathcal{M}_n}x(n-k-1) &= \widehat{x}(n-k-1) + Q_{n-k-1}W_{n-1,\,k}^*\begin{bmatrix} \varphi(n-k-1) \\ \varphi(n-k) \\ \vdots \\ \varphi(n-1) \end{bmatrix} \\
&\quad + Q_{n-k-1}\left(F_{n-k-1}^*F_{n-k}^*\cdots F_{n-1}^*\right)C_n^*G_n\varphi(n).
\end{aligned} \tag{5.19}$$

Now observe that

$$W_{n,\,k+1} = \begin{bmatrix} W_{n-1,\,k} \\ G_nC_nF_{n-1}F_{n-2}\cdots F_{n-k-1} \end{bmatrix}.$$

Using this fact in (5.19) yields the formula in (5.7) when $k+1$ replaces k. This completes the induction and the proof.

3.6 The steady state Kalman filter

In this section we will present the steady state Kalman filter. Throughout we assume that the system in (1.2) is time invariant, that is,

$$x(n+1) = Ax(n) + Bu(n) \qquad \text{and} \qquad y(n) = Cx(n) + Dv(n) \tag{6.1}$$

where A is an operator on $\mathcal{X}$ and B maps $\mathcal{U}$ into $\mathcal{X}$ while C maps $\mathcal{X}$ into $\mathcal{Y}$ and D maps $\mathcal{V}$ into $\mathcal{Y}$. Moreover, the operators $\{A, B, C, D\}$ are all fixed and do not depend upon n. We

also assume that the operator D is onto $\mathcal{Y}$, or equivalently, DD^* is invertible. As before, let Q_n be the solution to the Riccati difference equation in (1.6) associated with $\{A, B, C, D\}$, that is,

$$Q_{n+1} = AQ_nA^* + BB^* - AQ_nC^* \left(CQ_nC^* + DD^*\right)^{-1} CQ_nA^* . \tag{6.2}$$

According to Proposition 3.4.3, this Riccati difference equation can also be rewritten as

$$\begin{aligned} Q_{n+1} &= (A - \Delta_n C) Q_n (A - \Delta_n C)^* + BB^* + \Delta_n DD^* \Delta_n^* \\ \Delta_n &= AQ_nC^* \left(CQ_nC^* + DD^*\right)^{-1} . \end{aligned} \tag{6.3}$$

Moreover, assume that the initial condition Q_0 is positive. The form for the Riccati difference equation in (6.3) shows that Q_n is positive for all integers $n \geq 0$. Since DD^* is invertible and Q_n is positive, $(CQ_nC^* + DD^*)$ is strictly positive. In particular, $(CQ_nC^* + DD^*)$ is invertible. Hence this Riccati difference equation is well defined for all n. The following result shows that if the initial condition $Q_0 = 0$, then the solution Q_n is increasing.

THEOREM 3.6.1 *Consider the time invariant system $\{A, B, C, D\}$ where D is onto. Let Q_n be the solution for the Riccati difference equation in (6.2) where the initial condition $Q_0 = 0$. Then the following holds.*

(i) The solution $\{Q_n\}_0^\infty$ forms an increasing sequence of positive operators. To be precise, $Q_n \leq Q_{n+1}$ for all integers $n \geq 0$.

(ii) If the pair $\{C, A\}$ is observable, then Q_n converges to a positive operator P as n tends to infinity, that is,

$$P = \lim_{n \to \infty} Q_n . \tag{6.4}$$

In this case, P is a positive solution for the algebraic Riccati equation

$$P = APA^* + BB^* - APC^* \left(CPC^* + DD^*\right)^{-1} CPA^* . \tag{6.5}$$

(iii) If $\{A, B, C, D\}$ is controllable and observable, then P is strictly positive.

(iv) If $\{A, B, C, D\}$ is controllable and observable, then $A - K_PC$ is stable where K_P is the operator defined by $K_P = APC^* \left(CPC^* + DD^*\right)^{-1}$.

The steady state Kalman filter, predictor and smoother. Assume that $\{A, B, C, D\}$ is a controllable and observable system. Let P be the positive solution to the algebraic Riccati equation in (6.5) determined by (6.4). (It turns out that there is only one positive solution to the algebraic Riccati equation in (6.5); see Theorem 3.7.3 in Section 3.7.) Notice that Δ_n converges to K_P and n tends to infinity. By passing limits in the Kalman filter (4.10) and (4.11), we arrive at the *steady state Kalman filter* defined by

$$\begin{aligned} \widehat{x}(n+1) &= (A - K_PC)\widehat{x}(n) + K_Py \\ K_P &= APC^* \left(CPC^* + DD^*\right)^{-1} . \end{aligned} \tag{6.6}$$

Theorem 3.6.1 guarantees that $A - K_P C$ is stable. The steady state Kalman filter provides an estimate $\widehat{x}(n)$ for the state $x(n)$ given the past $\{y(j)\}_0^{n-1}$. The Kalman filter converges to the steady state Kalman filter. In other words, the steady state Kalman filter is an optimal state estimator in the limit.

The steady state Kalman predictor is an estimate $\widehat{x}(n+m|n-1)$ for the future state $x(n+m)$ given the past $\{y(j)\}_0^{n-1}$ where $m \geq 0$ is an integer. Motivated by Theorem 3.1.3 in Section 3.1.1, the *Kalman steady state predictor* is defined by $\widehat{x}(n+m|n-1) = A^m \widehat{x}(n)$ where $\widehat{x}(n)$ is the steady state estimate in (6.6). The steady state Kalman predictor is an optimal state predictor in the limit.

To present the steady state smoother, let $W_{n,k}$ be the observability matrix defined by

$$W_{n,k} = \begin{bmatrix} GC \\ GCF \\ GCF^2 \\ \vdots \\ GCF^k \end{bmatrix}.$$

Here $F = A - K_P C$ and $G = (CPC^* + DD^*)^{-1}$ are constant operators. The steady state Kalman smoother is an estimate $\widehat{x}(n-k|n)$ for the past $x(n-k)$ given $\{y(j)\}_0^n$ where $k \geq 0$ is an integer. Motivated by Theorem 3.5.1 in Section 3.5, the *Kalman steady state smoother* is defined by

$$\widehat{x}(n-k|n) = \widehat{x}(n-k) + PW_{n,k}^* \begin{bmatrix} \varphi(n-k) \\ \varphi(n-k+1) \\ \vdots \\ \varphi(n) \end{bmatrix} \tag{6.7}$$

where $\varphi(m) = y(m) - C\widehat{x}(m)$ and $\widehat{x}(n)$ is the steady state estimate in (6.6). The steady state Kalman smoother is an optimal state smoother in the limit.

Now let us return to Theorem 3.6.1. It is noted that if the pair $\{C, A\}$ is not observable, then the solution Q_n to the Riccati difference equation (4.26) may diverge. For example, consider the system $\{2, 1, 0, 1\}$. Then $Q_{n+1} = 4Q_n + 1$ where $Q_0 = 0$. In this case, the solution $Q_n = \sum_{k=0}^{n-1} 4^k = (4^n - 1)/3$ for $n \geq 1$. Clearly, Q_n approaches infinity as n tends to infinity. To prove Theorem 3.6.1 we use the following result known as the minimum principle.

LEMMA 3.6.2 (Minimum principle.) *Consider the time invariant system $\{A, B, C, D\}$ given by (6.1). Let Q_n be the solution to the Riccati difference equation (6.3) subject to the initial condition $Q_0 = 0$. Let V_n be the solution to the Riccati difference equation*

$$V_{n+1} = (A - \Phi_n C)V_n(A - \Phi_n C)^* + BB^* + \Phi_n DD^* \Phi_n^*, \tag{6.8}$$

where $V_0 = 0$ and Φ_n is any operator from $\mathcal{Y}$ into $\mathcal{X}$. Then $Q_n \leq V_n$ for all integers $n \geq 0$.

PROOF. Here we will present a stochastic proof of this result. In Section 3.7 we will give a deterministic proof of the minimum principle. Let $\xi(n)$ be the state for the following system

$$\xi(n+1) = (A - \Phi_n C)\xi(n) + \Phi_n y(n) \qquad (\xi(0) = 0). \tag{6.9}$$

By recursively solving for $\xi(n)$, it follows that

$$\xi(n) = \sum_{k=0}^{n-1} (A - \Phi_{n-1}C)(A - \Phi_{n-2}C) \cdots (A - \Phi_{k+1}C)\Phi_k y(k)\,. \tag{6.10}$$

Recall that $\mathcal{M}_{n-1} = \bigvee_0^{n-1} y(j)$. So the components of $\xi(n)$ are contained in $\mathcal{M}_{n-1}$ for all integers $n \geq 1$. In particular, $\xi(n)$ is orthogonal to $u(n)$ and $v(n)$; see (4.15). Recall that $y(n) = Cx(n) + Dv(n)$. Subtracting the state equation for ξ in equation (6.9) from $x(n+1) = Ax(n) + Bu(n)$ yields

$$\begin{aligned} x(n+1) - \xi(n+1) &= Ax + Bu(n) - (A - \Phi_n C)\xi(n) - \Phi_n(Cx(n) + Dv(n)) \\ &= (A - \Phi_n C)(x(n) - \xi(n)) + Bu(n) - \Phi_n Dv(n)\,. \end{aligned}$$

In other words,

$$x(n+1) - \xi(n+1) = (A - \Phi_n C)(x(n) - \xi(n)) + Bu(n) - \Phi_n Dv(n)\,. \tag{6.11}$$

Let $V_n = E(x(n) - \xi(n))(x(n) - \xi(n))^*$. Notice that $x(n) - \xi(n)$ is orthogonal to both $u(n)$ and $v(n)$. Using this in (6.11) yields the Riccati difference equation in (6.8). Since $Q_n = E(x(n) - \widehat{x}(n))(x(n) - \widehat{x}(n))^*$ and the components of $\xi(n)$ are contained in $\mathcal{M}_{n-1}$, Equation (2.2) in Theorem 2.2.1 in Chapter 2 shows that $Q_n \leq V_n$ for all integers $n \geq 0$. This completes the proof.

Proof of Theorem 3.6.1. First let us prove Part (i). By assumption the initial condition $Q_0 = 0$. (If Q_0 is nonzero, then the corresponding solution Q_n is not necessarily increasing.) We claim that $Q_{n+1} \geq V_n$ where V_n is the solution to the Riccati difference equation in (6.8), subject to the initial condition $V_0 = 0$, and $\Phi_{k-1} = \Delta_k$ for $k \geq 1$. To prove this we use induction. Notice that $Q_1 = BB^* \geq 0 = V_0$, and thus, $Q_1 \geq V_0$. Now assume that $Q_k \geq V_{k-1}$ for some integer $k \geq 2$. By choosing $\Phi_{k-1} = \Delta_k$ in (6.8), we obtain

$$V_k = (A - \Delta_k C)V_{k-1}(A - \Delta_k C)^* + BB^* + \Delta_k DD^*\Delta_k^*\,. \tag{6.12}$$

Subtracting (6.12) from (6.3) yields

$$Q_{k+1} - V_k = (A - \Delta_k C)\,(Q_k - V_{k-1})(A - \Delta_k C)^* \geq 0\,.$$

(The inequality follows from the fact that if M is positive, then NMN^* is also positive.) Thus $Q_{k+1} \geq V_k$. By induction it follows that $Q_{n+1} \geq V_n$ for all integers $n \geq 0$. Using the minimum property, $Q_n \leq V_n \leq Q_{n+1}$. Therefore $\{Q_n\}_0^\infty$ is an increasing sequence of positive operators.

Proof of Part (ii). Now assume that the pair $\{C, A\}$ is observable. Then we claim that the sequence $\{Q_n\}_0^\infty$ is uniformly bounded, that is, there exists a finite constant γ such that $Q_n \leq \gamma I$ for all integers $n \geq 0$. Because the pair $\{C, A\}$ is observable, there exists a feedback L from $\mathcal{Y}$ into $\mathcal{X}$ such that $A - LC$ is stable, that is, all the eigenvalues of $A - LC$

are contained in the open unit disc; see [8, 22, 28]. By setting $\Phi_n = L$ for all n in (6.8), we arrive at the following Riccati difference equation

$$V_{n+1} = (A - LC)V_n(A - LC)^* + BB^* + LDD^*L^* \qquad (V_0 = 0)\,.$$

By recursively solving for V_n, we see that

$$V_n = \sum_{j=0}^{n-1}(A - LC)^j(BB^* + LDD^*L^*)(A - LC)^{*j} \qquad (n \geq 1)\,.$$

Since $BB^* + LDD^*L^*$ is positive, this readily implies that

$$V_n \leq \sum_{j=0}^{\infty}(A - LC)^j(BB^* + LDD^*L^*)(A - LC)^{*j} = V_\infty\,.$$

Here we set V_∞ equal to the infinite sum. Notice that V_∞ is bounded because $A - LC$ is stable. In particular, $V_n \leq V_\infty$ where V_∞ is the solution to the Lyapunov equation

$$V_\infty = (A - LC)V_\infty(A - LC)^* + BB^* + LDD^*L^*\,.$$

By employing the minimum principle, $Q_n \leq V_n \leq V_\infty$. Hence $\{Q_n\}_0^\infty$ is uniformly bounded. Because $\{Q_n\}_0^\infty$ is an increasing uniformly bounded sequence of positive operators, Q_n converges to a positive operator P as n tends to infinity; see Halmos [18]. So by passing limits in the Riccati difference equation (6.2), we arrive at the algebraic Riccati equation in (6.5).

Proof of Parts (*iii*) and (*iv*). Now assume that $\{A, B, C, D\}$ is controllable. Moreover, assume that P is any solution to the algebraic Riccati equation in (6.5). Then we claim that P is strictly positive and $A - K_PC$ is stable. (Notice that we did not assume that P is given by the limit in (6.4). The conclusions in Parts (*iii*) and (*iv*) follow from the hypothesis that P is a positive solution to the algebraic Riccati equation and the system is controllable.) By rearranging terms in (6.5) we see that P is a solution to the following algebraic Riccati equation

$$\begin{aligned} P &= (A - K_PC)P(A - K_PC)^* + \begin{bmatrix} B & K_PD \end{bmatrix}\begin{bmatrix} B & K_PD \end{bmatrix}^* \\ K_P &= APC^*\,(CPC^* + DD^*)^{-1}\,. \end{aligned} \tag{6.13}$$

We claim that the pair $\{A - K_PC, \begin{bmatrix} B & K_PD \end{bmatrix}\}$ is controllable. According to the PBH controllability test, a pair $\{\Lambda \text{ on } \mathcal{X}, \Gamma\}$ is controllable if and only if the rank of the operator $\begin{bmatrix} \Lambda - \lambda I & \Gamma \end{bmatrix}$ equals the dimension of $\mathcal{X}$ for all complex numbers λ; see Lemma 9.4.4 in Chapter 9. Because the range of D equals $\mathcal{Y}$ and the pair $\{A, B\}$ is controllable, we have

$$\operatorname{rank}\begin{bmatrix} (A - K_PC - \lambda I) & B & K_PD \end{bmatrix} = \operatorname{rank}\begin{bmatrix} (A - \lambda I) & B & K_PD \end{bmatrix} = \dim \mathcal{X}$$

for all λ in $\mathbb{C}$. Hence the pair $\{A - K_PC, \begin{bmatrix} B & K_PD \end{bmatrix}\}$ is controllable.

We claim that P is strictly positive. To see this first observe that the kernel or null space of P is an invariant subspace for A^*. If $Px = 0$ for some x in $\mathcal{X}$, then (6.13) implies that

$$0 = (Px, x) = \|P^{1/2}(A - K_PC)^*x\|^2 + \|B^*x\|^2 + \|D^*K_P^*x\|^2\,.$$

(As expected, $P^{1/2}$ denotes the positive square root of P.) Thus $P^{1/2}(A - K_P C)^* x$ and $B^* x$ and $D^* K_P^* x$ are all zero. Since D^* is one to one, $K_P^* x = 0$. This implies that $P^{1/2} A^* x = 0$. In particular, $PA^* x = 0$. So the kernel of P is an invariant subspace for A^*. Let x be any eigenvector for A^* in the kernel of P, that is, assume that $A^* x = \lambda x$ for some nonzero vector x in the kernel of P. Since $Px = 0$, our previous analysis shows that $B^* x = 0$. Hence $B^* A^{*k} x = B^* \lambda^k x = \lambda^k B^* x = 0$. By controllability, x must be zero. This contradicts the fact that an eigenvector is nonzero. In other words, the kernel of P is zero. Therefore P is strictly positive, and Part (*iii*) holds.

The above analysis shows that $\{A - K_P C, \left[\begin{array}{cc} B & K_P D \end{array}\right]\}$ is a controllable pair and P is strictly positive solution to the Lyapunov equation (6.13). By employing Theorem 9.5.4 in Chapter 9, we see that $A - K_P C$ is stable. This completes the proof.

REMARK 3.6.3 *Let $\{A, B, C, D\}$ be a controllable time invariant system where D is onto. Assume that P is a positive solution to the algebraic Riccati equation*

$$P = APA^* + BB^* - APC^* \left(CPC^* + DD^*\right)^{-1} CPA^* . \tag{6.14}$$

Then the proof of Parts (iii) and (iv) of Theorem 3.6.1 show that P is strictly positive. Moreover, the operator $A - APC^* \left(CPC^* + DD^*\right)^{-1} C$ is stable.

3.6.1 Exercise

Problem 1. As in Problem 1 in Exercise 3.2.1, consider the discrete time system given by

$$\begin{aligned} \left[\begin{array}{c} x_1(n+1) \\ x_2(n+1) \end{array}\right] &= \left[\begin{array}{cc} 0 & 1 \\ -0.98 & 1.94 \end{array}\right] \left[\begin{array}{c} x_1(n) \\ x_2(n) \end{array}\right] + \left[\begin{array}{c} 0 \\ 1 \end{array}\right] u(n) \\ y(n) &= \left[\begin{array}{cc} 3 & 4 \end{array}\right] \left[\begin{array}{c} x_1(n) \\ x_2(n) \end{array}\right] + v(n) . \end{aligned} \tag{6.15}$$

Here $\{u(n)\}$ and $\{v(n)\}$ are independent Gaussian white noise processes which are orthogonal to the initial random variable $x(0) = x_0$. Assume that the initial state is $x_0 = \left[\begin{array}{cc} 1 & 2 \end{array}\right]^{tr}$. Implement the steady state Kalman filter in (6.6) to compute the steady state estimate $\widehat{x}(n)$ for $x(n)$ for $0 \leq n \leq 500$ where the initial state estimate is $\widehat{x}(0) = \left[\begin{array}{cc} 0 & 0 \end{array}\right]^{tr}$. Compare your estimate $\widehat{x}(n)$ with the actual state $x(n)$. Discuss the difference between the steady state estimate of the state and the optimal estimate of the state computed in Problem 1 in Exercise 3.2.1.

Problem 2. Consider a mass spring damper system given by

$$bw = m\ddot{q} + c\dot{q} + kq \qquad \text{and} \qquad y = \dot{q} + Dv . \tag{6.16}$$

Assume that the disturbance w and the measurement noise v are independent white noise processes. Notice that the disturbance w is a force on the mass. Now assume that

$$m = 1, \quad c = 1, \quad k = 100, \quad b = 1/2 \quad \text{and} \quad D = 1/4 .$$

Use the results in Section 9.3.1 in Chapter 9, to convert the continuous time system in (6.16) to a discrete time system of the form

$$x(n+1) = Ax(n) + Bu(n) \qquad \text{and} \qquad y(n) = Cx(n) + Dv(n)\,. \tag{6.17}$$

Moreover, assume that $u(n)$ and $v(n)$ are independent white noise random processes, which are independent to the initial condition $x(0)$. Explain how your discrete time state $x(n)$ is related to q and $\dot{q}$. Assume that the initial conditions for the differential equation in (6.16) are given by $q(0) = 1$ and $\dot{q}(0) = 2$. Implement the steady state Kalman filter in (6.6) to compute the steady state estimate for the position q and velocity $\dot{q}$ of the mass. In other words, implement the steady state Kalman filter to compute the optimal state estimate $\widehat{x}(n)$ for $x(n)$ where the initial state estimate is $\widehat{x}(0) = \begin{bmatrix} 0 & 0 \end{bmatrix}^{tr}$. Compare your estimate $\widehat{x}(n)$ with the actual state $x(n)$.

3.7 Discrete time Riccati equations

In this section we will develop some further properties for the discrete time Riccati equation. To this end, recall that Q_n is the solution to the following Riccati difference equation

$$\begin{aligned} Q_{n+1} &= (A - \Delta_n C) Q_n (A - \Delta_n C)^* + BB^* + \Delta_n DD^* \Delta_n^* \\ \Delta_n &= AQ_n C^* \left(CQ_n C^* + DD^*\right)^{-1}\,. \end{aligned} \tag{7.1}$$

We begin with the following generalization of the minimum principle in Lemma 3.6.2.

LEMMA 3.7.1 (Minimum principle.) *Consider the time invariant system $\{A, B, C, D\}$ where D is onto. Let Q_n be the solution to the Riccati difference equation (7.1) subject to the initial condition Q_0. Let V_n be the solution to the Riccati difference equation*

$$V_{n+1} = (A - \Phi_n C) V_n (A - \Phi_n C)^* + BB^* + \Phi_n DD^* \Phi_n^* \tag{7.2}$$

where Φ_n is any operator mapping $\mathcal{Y}$ into $\mathcal{X}$, and V_0 is the initial condition. If $Q_0 \leq V_0$, then $Q_n \leq V_n$ for all integers $n \geq 0$.

Our proof of this minimum principle is based on the following result.

LEMMA 3.7.2 *Let A be an operator on $\mathcal{X}$ and C an operator mapping $\mathcal{X}$ into $\mathcal{Y}$ while R is a strictly positive operator on $\mathcal{Y}$. Let V be any positive operator on $\mathcal{X}$ and Φ an operator mapping $\mathcal{Y}$ into $\mathcal{X}$. Then we have*

$$(A - K_V C) V (A - K_V C)^* + K_V R K_V^* \leq (A - \Phi C) V (A - \Phi C)^* + \Phi R \Phi^* \tag{7.3}$$

where $K_V = AVC^* \left(CVC^* + R\right)^{-1}$. Finally, we have equality in (7.3) if and only if $\Phi = K_V$.

Proof. Let Ω_Φ be the operator defined by the right hand side of (7.3), that is,

$$\Omega_\Phi = (A - \Phi C) V (A - \Phi C)^* + \Phi R \Phi^*\,. \tag{7.4}$$

Notice that the definition of Ω_Φ depends upon the operator Φ. So our task is to show that $\Omega_\Phi \geq \Omega_{K_V}$. To this end, set $F = A - K_V C$ and $Z = K_V - \Phi$. Clearly, $A - \Phi C = F + ZC$ and $\Phi = K_V - Z$. Using these definitions, we obtain

$$\begin{aligned}
\Omega_\Phi &= (A-\Phi C)V(A-\Phi C)^* + \Phi R \Phi^* \\
&= (F+ZC)V(F+ZC)^* + (K_V - Z)R(K_V - Z)^* \\
&= FVF^* + ZCVC^*Z^* + ZCVF^* + FVC^*Z^* \\
&\quad + K_V R K_V^* + ZRZ^* - ZRK_V^* - K_V R Z^* \\
&= FVF^* + Z(CVC^* + R)Z^* + K_V R K_V^* \\
&\quad + (FVC^* - K_V R)Z^* + Z(FVC^* - K_V R)^* \\
&= FVF^* + Z(CVC^* + R)Z^* + K_V R K_V^* \\
&\quad + \left(AVC^* - K_V(CVC^* + R)\right) Z^* + Z\left(AVC^* - K_V(CVC^* + R)\right)^* \\
&= FVF^* + K_V R K_V^* + Z(CVC^* + R)Z^* = \Omega_{K_V} + Z(CVC^* + R)Z^* .
\end{aligned}$$

This readily implies that

$$\Omega_\Phi = \Omega_{K_V} + (K_V - \Phi)(CVC^* + R)(K_V - \Phi)^* \geq \Omega_{K_V} . \tag{7.5}$$

The inequality follows from the fact that $Z(CVC^* + R)Z^*$ is positive. (If M is any positive operator, then NMN^* is also a positive operator.) Because R is invertible, $CVC^* + R$ is strictly positive. Hence $Z(CVC^* + R)Z^* = 0$ if and only if $Z = 0$. Therefore we have equality in (7.5) if and only if $K_V = \Phi$. This complete the proof.

Proof of Lemma 3.7.2 based on the projection theorem. Let $V^{1/2}$ be the positive square root of V and $R^{1/2}$ be the positive square root of R. As before, let Ω_Φ be the operator defined in (7.4). For x in $\mathcal{X}$, we have

$$\begin{aligned}
(\Omega_\Phi x, x) &= ((A-\Phi C)V(A-\Phi C)^*x, x) + (\Phi R \Phi^* x, x) \\
&= (V^{1/2}(A-\Phi C)^*x, V^{1/2}(A-\Phi C)^*x) + (R^{1/2}\Phi^* x, R^{1/2}\Phi^* x) \\
&= \|V^{1/2}(A-\Phi C)^*x\|^2 + \|R^{1/2}\Phi^* x\|^2 \\
&= \| \begin{bmatrix} V^{1/2}A^*x \\ 0 \end{bmatrix} - \begin{bmatrix} V^{1/2}C^*\Phi^* x \\ R^{1/2}\Phi^* x \end{bmatrix} \|^2 .
\end{aligned} \tag{7.6}$$

Now let T be the operator from $\mathcal{Y}$ into $\mathcal{X} \oplus \mathcal{Y}$ and g the vector in $\mathcal{X} \oplus \mathcal{Y}$ defined by

$$T = \begin{bmatrix} V^{1/2}C^* \\ R^{1/2} \end{bmatrix} \qquad \text{and} \qquad g = \begin{bmatrix} V^{1/2}A^*x \\ 0 \end{bmatrix} . \tag{7.7}$$

By consulting (7.6), we have $(\Omega_\Phi x, x) = \|g - T\Phi^* x\|^2$. Hence

$$(\Omega_\Phi x, x) = \|g - T\Phi^* x\|^2 \geq \inf\{\|g - Tu\|^2 : u \in \mathcal{Y}\} . \tag{7.8}$$

Notice that $T^*T = CVC^* + R$. Because R is strictly positive, T^*T is invertible. According to Theorem 1.5.1 in Chapter 1, the solution to the optimization problem in (7.8) is unique and given by

$$\widehat{u} = (T^*T)^{-1}T^*g = (CVC^* + R)^{-1}CVA^*x = K_V^* x .$$

So the optimal solution $\widehat{u} = K_V^* x$. By employing $\widehat{u} = K_V^* x$ in (7.8), we obtain

$$(\Omega_\Phi x, x) = \|g - T\Phi^* x\|^2 \geq \inf\{\|g - Tu\|^2 : u \in \mathcal{Y}\} = \|g - TK_V^* x\|^2 = (\Omega_{K_V} x, x)\,. \quad (7.9)$$

Therefore $\Omega_\Phi \geq \Omega_{K_V}$. Notice that $\Omega_\Phi = \Omega_{K_V}$ if and only if $\|g - T\Phi^* x\| = \|g - TK_V^* x\|$ for all x in $\mathcal{X}$. Because $\widehat{u} = K_V^* x$ is the unique solution to the optimization problem in (7.8), we see that $\Omega_\Phi = \Omega_{K_V}$ if and only if $\Phi^* x = K_V^* x$ for all x, or equivalently, $\Phi = K_V$. This completes the proof.

Proof of the minimum principle Lemma 3.7.1. Let Δ_n and Λ_n be the operators defined by

$$\Delta_n = AQ_nC^*\left(CQ_nC^* + DD^*\right)^{-1} \quad \text{and} \quad \Lambda_n = AV_nC^*\left(CV_nC^* + DD^*\right)^{-1}\,.$$

According to the hypothesis $Q_0 \leq V_0$. Now let us apply induction and assume that $Q_n \leq V_n$. Then by using Lemma 3.7.2 twice with $\Delta_n = K_{Q_n}$ and $\Lambda_n = K_{V_n}$, we have

$$\begin{aligned} Q_{n+1} &= (A - \Delta_n C)Q_n(A - \Delta_n C)^* + \Delta_n DD^*\Delta_n^* + BB^* \\ &\leq (A - \Lambda_n C)Q_n(A - \Lambda_n C)^* + \Lambda_n DD^*\Lambda_n^* + BB^* \\ &\leq (A - \Lambda_n C)V_n(A - \Lambda_n C)^* + \Lambda_n DD^*\Lambda_n^* + BB^* \\ &\leq (A - \Phi_n C)V_n(A - \Phi_n C)^* + \Phi_n DD^*\Phi_n^* + BB^* \\ &= V_{n+1}\,. \end{aligned}$$

Therefore $Q_{n+1} \leq V_{n+1}$ which completes the proof.

Let us complete this section with the following classical result.

THEOREM 3.7.3 *Let $\{A, B, C, D\}$ be a controllable and observable time invariant system where D is onto. Then there exists a unique positive solution to the algebraic Riccati equation*

$$P = APA^* + BB^* - APC^*\left(CPC^* + DD^*\right)^{-1} CPA^*\,. \quad (7.10)$$

In this case, the operator $A - APC^*\left(CPC^* + DD^*\right)^{-1} C$ is stable and P is strictly positive. Moreover, the unique positive solution to this algebraic Riccati equation in (7.10) is given by

$$P = \lim_{n\to\infty} Q_n \quad (7.11)$$

where Q_n is the solution to the Riccati difference equation in (7.1) subject to the initial condition $Q_n = 0$.

Proof. According to Theorem 3.6.1, the algebraic Riccati equation in (7.10) admits a positive solution. In fact, the operator P given by the limit in (7.11) is a positive solution to this algebraic Riccati equation, and $A - K_P C$ is stable. So to complete the proof it remains to show that there is only one positive solution.

Now assume that P and V are two arbitrary positive solutions to the algebraic Riccati equation in (7.10). As before, let

$$K_P = APC^* \left(CPC^* + DD^*\right)^{-1} \quad \text{and} \quad K_V = AVC^* \left(CVC^* + DD^*\right)^{-1} .$$

By implementing Lemma 3.7.2, we have

$$\begin{aligned} P &= (A - K_P C)P(A - K_P C)^* + K_P DD^* K_P^* + BB^* \\ &\leq (A - K_V C)P(A - K_V C)^* + K_V DD^* K_V^* + BB^* . \end{aligned}$$

Subtracting this from $V = (A - K_V C)V(A - K_V C)^* + K_V DD^* K_V^* + BB^*$, we obtain

$$V - P \geq (A - K_V C)(V - P)(A - K_V C)^* . \tag{7.12}$$

Since V is a positive solution to the algebraic Riccati equation (7.10), the operator $A - K_V C$ is stable; see Remark 3.6.3. According to Lemma 3.7.4 below $V - P \geq 0$. In other words, $V \geq P$. Since V and P are two arbitrary positive solutions to the algebraic Riccati equation in (7.10), we can interchange the roles of V and P, which yields $P \geq V$. Therefore $P = V$. This completes the proof.

LEMMA 3.7.4 *Let F be a stable operator on $\mathcal{X}$ and R a self-adjoint operator on $\mathcal{X}$. If $R \geq FRF^*$, then R is positive.*

Proof. By recursively substituting FRF^* for R we have

$$R \geq FRF^* \geq FFRF^*F^* \geq F^3RF^{*3} \geq \cdots \geq F^kRF^{*k}$$

where k is any positive integer. Hence $R \geq F^kRF^{*k}$. Since F is stable, F^k converges to zero as k tends to infinity. Therefore F^kRF^{*k} converges to zero and R is positive. This completes the proof.

For further results and a more detailed discussion of Kalman filtering see [2, 4, 9, 26].

3.8 The continuous time Kalman filter

In this section we will present a brief introduction to the continuous time Kalman filter. For a rigorous presentation of continuous time Kalman filtering see Davis [9]. Recall that $w(t)$ is a random process with values in $\mathbb{C}^m$ if $w(t)$ is a random vector in $\mathbb{C}^m$ for all t. We say that $w(t)$ is a *mean zero process* if $Ew(t) = 0$ for all t. Finally, $w(t)$ is a *white noise process* if $w(t)$ is a mean zero random process and

$$Ew(t)w(\sigma)^* = \delta(t - \sigma)I . \tag{8.1}$$

Here $\delta(t)$ is the delta Dirac function.

To introduce the continuous time Kalman filter, consider the state space system given by

$$\dot{x} = Ax + Bu \qquad \text{and} \qquad y = Cx + Dv . \tag{8.2}$$

Here A is an operator on $\mathcal{X}$ and B maps $\mathcal{U}$ into $\mathcal{X}$ while C maps $\mathcal{X}$ into $\mathcal{Y}$ and D maps $\mathcal{V}$ into $\mathcal{Y}$ where $\mathcal{X}$, $\mathcal{U}$, $\mathcal{Y}$ and $\mathcal{V}$ are all $\mathbb{C}^k$ spaces of the appropriate size. It is emphasized that $\{A, B, C, D\}$ can be time varying matrices, that is, $A = A(t)$, $B = B(t)$, $C = C(t)$ and $D = D(t)$ are matrix valued continuous functions for all time $t \geq 0$. Throughout this section we assume that $D(t)$ is onto, or equivalently, $D(t)D(t)^*$ is invertible for all $t \geq 0$. The initial condition $x(0) = x_0$ is a random vector with values in $\mathcal{X}$. The disturbance $u(t)$ and output measurement noise $v(t)$ are independent white noise random process. Moreover, we assume that the initial condition x_0, $u(t)$ and $v(\sigma)$ are all independent random vectors for all $t \geq 0$ and $\sigma \geq 0$. In particular, this implies that x_0, $u(t)$ and $v(\sigma)$ are orthogonal for all $t \geq 0$ and $\sigma \geq 0$. We also assume that the initial condition $\widehat{x}_0$ for the Kalman filter and the initial condition $Q_0 = Q(0) = Ex_0x_0^*$ for the Riccati differential equation are known. Finally, is it noted that $u(t)$ is called the disturbance or state noise, while $v(t)$ is the measurement noise.

The Kalman filtering problem is to compute the best estimate $\widehat{x}(t)$ of the state $x(t)$ given the past output $\{y(\sigma) : 0 \leq \sigma < t\}$. The Kalman filter is an optimal state estimator. To be precise, let $\mathcal{M}_{t-}$ be the subspace spanned by the random vectors $\{y(\sigma) : 0 \leq \sigma < t\}$, that is,

$$\mathcal{M}_{t-} = \bigvee \{y(\sigma) : 0 \leq \sigma < t\} .$$

Here $\bigvee$ denotes the closed linear span in the space of all random variables z such that $E|z|^2$ is finite. Let $P_{\mathcal{M}_{t-}}$ be the orthogonal projection onto $\mathcal{M}_{t-}$ for all $t \geq 0$. Then the best estimate $\widehat{x}(t)$ of the state $x(t)$ is given by the orthogonal projection $\widehat{x}(t) = P_{\mathcal{M}_{t-}} x(t)$. Finally, we assume that $y(0-) = 0$, or equivalently, $\mathcal{M}_{0-} = \{0\}$. The following result known as the Kalman filter provides us with a recursive algorithm to compute $\widehat{x}$.

THEOREM 3.8.1 *Consider the state space system*

$$\dot{x} = Ax + Bu \qquad \text{and} \qquad y = Cx + Dv \tag{8.3}$$

where u and v are independent white noise processes, which are independent to the initial condition $x(0) = x_0$, and D is onto. Then the optimal estimate $\widehat{x}(t) = P_{\mathcal{M}_{t-}} x(t)$ of the state $x(t)$ given the past $\{y(\sigma) : 0 \leq \sigma < t\}$ is computed by solving the differential equation

$$\dot{\widehat{x}} \;=\; A\widehat{x} + QC^*(DD^*)^{-1}\left(y - C\widehat{x}\right) . \tag{8.4}$$

The state covariance error $Q(t) = E(x(t) - \widehat{x}(t))(x(t) - \widehat{x}(t))^*$ is computed by solving the following Riccati differential equation

$$\dot{Q} \;=\; AQ + QA^* + BB^* - QC^*(DD^*)^{-1}CQ \tag{8.5}$$

subject to the initial condition $Q(0) = Ex_0x_0^*$.

A heuristic proof of Theorem 3.8.1. In this section we will follows the innovations approach in presented in Kailath [20] to present a heuristic proof of Theorem 3.8.1; see also Section 3.4. A precise mathematical proof is given in Davis [9]. Now let us try to mimic the Gram-Schmidt orthogonal procedure presented in Section 3.4.1 for the continuous time setting. To this end, let $\mathcal{M}_t$ and $\mathcal{M}_{t-}$ be the subspaces defined by

$$\mathcal{M}_t = \bigvee \{y(\sigma) : 0 \leq \sigma \leq t\} \quad \text{and} \quad \mathcal{M}_{t-} = \bigvee \{y(\sigma) : 0 \leq \sigma < t\} .$$

Notice that the subspaces $\mathcal{M}_t$ are increasing, that is, $\mathcal{M}_\sigma \subset \mathcal{M}_t$ if $\sigma \le t$. Recall that the solution to the state space system in (8.3) is given by

$$x(t) = \Psi(t,0)x_0 + \int_0^t \Psi(t,\tau)B(\tau)u(\tau)\,d\tau \tag{8.6}$$

$$y(t) = C(t)\Psi(t,0)x_0 + \int_0^t C(t)\Psi(t,\tau)B(\tau)u(\tau)\,d\tau + D(t)v(t)\,. \tag{8.7}$$

Here $\Psi(t,\tau)$ is the state transition matrix for A, that is,

$$\frac{\partial}{\partial t}\Psi(t,\tau) = A(t)\Psi(t,\tau)\,.$$

Moreover, we assume that the integrals in (8.6) and (8.7) do not include the value at time t, that is, $\int_0^t f(\tau)d\tau = \int_0^{t-} f(\tau)d\tau$. The solution for $y(t)$ in (8.7) shows that

$$\begin{aligned} \mathcal{M}_t &\subset \bigvee\{x_0, u(\sigma), v(\tau) : 0 \le \sigma < t \text{ and } 0 \le \tau \le t\} \\ \mathcal{M}_{t-} &\subset \bigvee\{x_0, u(\sigma), v(\tau) : 0 \le \sigma < t \text{ and } 0 \le \tau < t\}\,. \end{aligned} \tag{8.8}$$

Now let $\varphi(t)$ be the random process defined by

$$\varphi(t) = y(t) - P_{\mathcal{M}_{t-}}y(t) \tag{8.9}$$

where $P_{\mathcal{M}_{t-}}$ is the orthogonal projection onto $\mathcal{M}_{t-}$. For convenience we set $R(t) = D(t)D(t)^*$ for all $t \ge 0$. We claim that $\varphi(t)$ a white noise process with intensity $R(t)$. To be precise,

$$E\varphi(t)\varphi(\sigma)^* = R(t)\delta(t-\sigma)I\,. \tag{8.10}$$

Moreover, $\varphi(t)$ and the output process $y(t)$ span the same past, that is,

$$\mathcal{M}_t = \bigvee\{\varphi(\sigma) : 0 \le \sigma \le t\} \quad \text{and} \quad \mathcal{M}_{t-} = \bigvee\{\varphi(\sigma) : 0 \le \sigma < t\}\,. \tag{8.11}$$

The process $\varphi(t)$ is called the *innovations process* for $y(t)$. Notice that $\varphi(t)$ mimics the innovation φ_n for y_n discussed in Section 3.4.1. In fact, the results in Section 3.4.1 readily show that (8.11) holds. So it remains to establish (8.10).

First let us show that $\varphi(t)$ is orthogonal to $\varphi(\sigma)$ when $t \ne \sigma$. Without loss of generality assume that $t > \sigma$. By employing the projection theorem, $\varphi(t) = y(t) - P_{\mathcal{M}_{t-}}y(t)$ is orthogonal to $\mathcal{M}_{t-}$. Since $\mathcal{M}_\sigma \subset \mathcal{M}_{t-}$, we see that $\varphi(t)$ is orthogonal to $\mathcal{M}_\sigma$. However, $\varphi(\sigma) = y(\sigma) - P_{\mathcal{M}_{\sigma-}}y(\sigma)$ is contained in $\mathcal{M}_\sigma$. Therefore $\varphi(t)$ is orthogonal to $\varphi(\sigma)$. In other words, if $t \ne \sigma$, then $\varphi(t)$ is orthogonal to $\varphi(\sigma)$. In particular, (8.10) holds when $t \ne \sigma$.

To show that $E\varphi(t)\varphi(t)^* = R(t)\delta(0)I$, recall that $\widehat{x}(t) = P_{\mathcal{M}_{t-}}x(t)$ is the optimal estimate of the state $x(t)$ given the past $\{y(\sigma) : 0 \le \sigma < t\}$. Let $\widetilde{x}(t) = x(t) - \widehat{x}(t)$ be the error in estimation. Equation (8.8) shows that $v(t)$ is orthogonal to $\mathcal{M}_{t-}$. In other words, $P_{\mathcal{M}_{t-}}v(t) = 0$. Using $y = Cx + Dv$, we have

$$\varphi(t) = y(t) - P_{\mathcal{M}_{t-}}y(t) = y(t) - P_{\mathcal{M}_{t-}}(Cx(t) + Dv(t)) = y(t) - C\widehat{x}(t)\,.$$

Hence $\varphi(t) = y(t) - C\widehat{x}(t)$. Since $\widetilde{x}(t) = x(t) - \widehat{x}(t)$ and $y(t) = Cx(t) + Dv(t)$, we obtain $\varphi(t) = C\widetilde{x}(t) + Dv(t)$. This yields the following two useful formulas for the innovations

$$\varphi(t) = y(t) - C\widehat{x}(t) = C\widetilde{x}(t) + Dv(t)\,. \tag{8.12}$$

By consulting (8.6) and (8.8), we see that $v(t)$ is orthogonal to $\widetilde{x}(t) = x(t) - P_{\mathcal{M}_{t-}}x(t)$. Recall that $Ev(t)v(\sigma)^* = \delta(t-\sigma)I$ and $R = DD^*$. This and $\varphi = C\widetilde{x} + Dv$, yields

$$E\varphi(t)\varphi(t)^* = CE\widetilde{x}(t)\widetilde{x}(t)^*C^* + R(t)\delta(0)\,. \tag{8.13}$$

We claim that the error covariance $E\widetilde{x}(t)\widetilde{x}(t)^*$ is finite. This follows from Equation (2.2) in Theorem 2.2.1 in Chapter 2, that is,

$$\begin{aligned} Q(t) = E\widetilde{x}(t)\widetilde{x}(t)^* &= E\left(x(t) - \widehat{x}(t)\right)\left(x(t) - \widehat{x}(t)\right)^* \\ &\leq E\left(x(t) - 0\right)\left(x(t) - 0\right)^* = Ex(t)x(t)^* < \infty\,. \end{aligned}$$

So using the fact that $Q(t)$ is finite and $\delta(0)$ is infinite in Equation (8.13), we see that $E\varphi(t)\varphi(t)^* = R(t)\delta(0)$. Therefore (8.10) holds.

The second equation in (8.11) with $\widehat{x}(t) = P_{\mathcal{M}_{t-}}x(t)$ implies that there exists a matrix valued function $H(t,\tau)$ such that

$$\widehat{x}(t) = \int_0^t H(t,\tau)\varphi(\tau)\,d\tau\,. \tag{8.14}$$

The integral in (8.14) does not include the value at time t, that is, $\int_0^t f(\tau)d\tau = \int_0^{t-} f(\tau)d\tau$. By the projection theorem $x(t) - \widehat{x}(t)$ is orthogonal to $\mathcal{M}_{t-} = \bigvee\{\varphi(\sigma) : 0 \leq \sigma < t\}$. This readily implies that

$$x(t) - \int_0^t H(t,\tau)\varphi(\tau)\,d\tau \quad \text{is orthogonal to } \varphi(\sigma) \text{ for all } 0 \leq \sigma < t\,.$$

By applying $\varphi(\sigma)^*$ to both sides and taking the expectation, we arrive at

$$\begin{aligned} Ex(t)\varphi(\sigma)^* &= \int_0^t H(t,\tau)E\varphi(\tau)\varphi(\sigma)^*\,d\tau \\ &= \int_0^t H(t,\tau)R(\tau)\delta(\tau-\sigma)\,d\tau = H(t,\sigma)R(\sigma)\,. \end{aligned}$$

The last equality follows from the evaluation property $\int f(\tau)\delta(\tau-\sigma)\,d\tau = f(\sigma)$ of the delta Dirac function $\delta(t)$. By taking the inverse of R, we obtain

$$H(t,\sigma) = Ex(t)\varphi(\sigma)^*R(\sigma)^{-1} \qquad (\text{for } 0 \leq \sigma < t)\,. \tag{8.15}$$

Substituting this in (8.14) yields the following expression for the optimal state

$$\widehat{x}(t) = \int_0^t \left(Ex(t)\varphi(\tau)^*\right)R(\tau)^{-1}\varphi(\tau)\,d\tau\,. \tag{8.16}$$

Let us recall the following version of Leibnitz formula for differentiating the integral

$$\frac{d}{dt}\int_0^t k(t,\tau)\,d\tau = k(t,t) + \int_0^t \frac{\partial}{\partial t}k(t,\tau)\,d\tau\,. \tag{8.17}$$

Here $k(t,\tau)$ is any differentiable function. According to (8.11) the innovations $\varphi(\sigma) \in \mathcal{M}_{t-}$ for all $0 \le \sigma < t$. Hence $u(t)$ is orthogonal to $\varphi(\sigma)$ for $0 \le \sigma < t$; see (8.8). Using this along with Leibnitz's formula we have

$$\begin{aligned}
\dot{\widehat{x}}(t) &= \frac{d}{dt}\int_0^t (Ex(t)\varphi(\tau)^*)\,R(\tau)^{-1}\varphi(\tau)\,d\tau \\
&= (Ex(t)\varphi(t)^*)\,R(t)^{-1}\varphi(t) + \int_0^t \left(\frac{\partial}{\partial t}Ex(t)\varphi(\tau)^*\right) R(\tau)^{-1}\varphi(\tau)\,d\tau \\
&= (Ex(t)\varphi(t)^*)\,R(t)^{-1}\varphi(t) + \int_0^t (E\dot{x}(t)\varphi(\tau)^*)\,R(\tau)^{-1}\varphi(\tau)\,d\tau \\
&= (Ex(t)\varphi(t)^*)\,R(t)^{-1}\varphi(t) + \int_0^t (E(Ax(t)+Bu(t))\varphi(\tau)^*)\,R(\tau)^{-1}\varphi(\tau)\,d\tau \\
&= (Ex(t)\varphi(t)^*)\,R(t)^{-1}\varphi(t) + A\int_0^t (Ex(t)\varphi(\tau)^*)\,R(\tau)^{-1}\varphi(\tau)\,d\tau \\
&= (Ex(t)\varphi(t)^*)\,R(t)^{-1}\varphi(t) + A\widehat{x}(t)\,.
\end{aligned}$$

To obtain the third equality we interchanged the partial derivative with the expectation, that is, $\partial Ex(t)\varphi(\sigma)^*/\partial t = E\partial x(t)\varphi(\sigma)^*/\partial t$. This is justified because the expectation is a weighted integral. Our previous calculation shows that the optimal state $\widehat{x}$ is the solution to the following differential equation

$$\dot{\widehat{x}} = A\widehat{x} + (Ex(t)\varphi(t)^*)\,R(t)^{-1}\varphi(t)\,. \tag{8.18}$$

To simplify this differential equation, we need an expression for $Ex(t)\varphi(t)^*$. To this end, observe that $x(t)$ is orthogonal to $v(t)$; see (8.6). Recall that the error covariance $Q(t) = E\widetilde{x}(t)\widetilde{x}(t)^*$. Using the second expression for φ in (8.12), that is, $\varphi = C\widetilde{x} + Dv$, we obtain

$$\begin{aligned}
Ex(t)\varphi(t)^* &= Ex(t)(C\widetilde{x}(t) + Dv(t))^* = Ex(t)\widetilde{x}(t)^*C^* \\
&= E(\widehat{x}(t) + \widetilde{x}(t))\widetilde{x}(t)^*C^* = E\widetilde{x}(t)\widetilde{x}(t)^*C^* = Q(t)C^*\,.
\end{aligned}$$

Hence $Ex(t)\varphi(t)^* = Q(t)C^*$. Substituting this into (8.18) yields

$$\dot{\widehat{x}} = A\widehat{x} + QC^*R^{-1}\varphi(t) = A\widehat{x} + QC^*R^{-1}\left(y - C\widehat{x}\right)\,. \tag{8.19}$$

The last equality follows from $\varphi = y - C\widehat{x}$; see (8.12). Therefore the optimal state trajectory $\widehat{x}$ is the solution to the differential equation in (8.4).

To complete the proof it remains to derive the Riccati differential equation in (8.5). By employing $\varphi = C\widetilde{x} + Dv$ in (8.19), we have

$$\dot{\widehat{x}} = A\widehat{x}(t) + QC^*R^{-1}C\widetilde{x} + QC^*R^{-1}Dv\,.$$

Subtracting this from $\dot{x} = Ax + Bu$ with $\widetilde{x} = x - \widehat{x}$, we obtain

$$\dot{\widetilde{x}} = \left(A - QC^*R^{-1}C\right)\widetilde{x} + Bu - QC^*R^{-1}Dv\,. \tag{8.20}$$

Now let $\Phi(t,\tau)$ be the state transition matrix for $A - QC^*R^{-1}C$, that is,

$$\frac{\partial}{\partial t}\Phi(t,\tau) = \left(A(t) - Q(t)C(t)^*R(t)^{-1}C(t)\right)\Phi(t,\tau)\,.$$

The solution to the differential equation in (8.20) is given by

$$\begin{aligned}\widetilde{x}(t) &= \Phi(t,0)\widetilde{x}(0) + \int_0^t \Phi(t,\tau)B(\tau)u(\tau)\,d\tau \\ &\quad - \int_0^t \Phi(t,\tau)Q(\tau)C(\tau)^*R(\tau)^{-1}D(\tau)v(\tau)\,d\tau\,.\end{aligned} \tag{8.21}$$

Recall that $\int f(\tau)\delta(\tau-\sigma)\,d\tau = f(\sigma)$. Using the fact that $\widetilde{x}(0)$ and $u(t)$ and $v(\sigma)$ are all orthogonal, with $Eu(t)u(\sigma)^* = \delta(t-\sigma)I$ and $Ev(t)v(\sigma)^* = \delta(t-\sigma)I$, we have

$$\begin{aligned}Q(t) &= E\widetilde{x}(t)\widetilde{x}(t)^* = \Phi(t,0)Q(0)\Phi(t,0)^* + \int_0^t \Phi(t,\tau)B(\tau)B(\tau)^*\Phi(t,\tau)^*\,d\tau \\ &\quad + \int_0^t \Phi(t,\tau)Q(\tau)C(\tau)^*R(\tau)^{-1}C(\tau)Q(\tau)\Phi(t,\tau)^*\,d\tau\,.\end{aligned} \tag{8.22}$$

Finally, the state transition property $\partial\Phi/\partial t = (A - QC^*R^{-1}C)\Phi$ with Leibnitz rule, yields

$$\begin{aligned}\dot{Q} &= \left(A - QC^*R^{-1}C\right)Q + Q\left(A - QC^*R^{-1}C\right)^* + BB^* + QC^*R^{-1}CQ \\ &= AQ + QA^* + BB^* - QC^*R^{-1}CQ\,.\end{aligned}$$

This is precisely the Riccati differential equation in (8.5). Finally, notice that because $\mathcal{M}_{0-} = 0$, we see that $\widetilde{x}(0) = x_0 - P_{\mathcal{M}_{0-}}x_0 = x_0$. Hence $Q(0) = E\widetilde{x}(0)\widetilde{x}(0)^* = Ex_0x_0^*$. This completes the proof.

REMARK 3.8.2 *As in Theorem 3.8.1, consider the state space system determined by*

$$\dot{x} = Ax + Bu \qquad \textit{and} \qquad y = Cx + Dv \tag{8.23}$$

where $u(t)$ and $v(t)$ are independent white noise random processes, which are independent to the initial condition $x(0)$. Our previous analysis shows that the Kalman filter in (8.4) can be used to recursively compute the innovations $\varphi(t) = y(t) - C\widehat{x}(t)$ for the process $y(t)$.

3.8.1 The steady state continuous time Kalman filter

In this section we will present the continuous time steady state Kalman filter. Throughout we assume that the system in (8.2) is time invariant, that is,

$$\dot{x} = Ax + Bu \qquad \text{and} \qquad y = Cx + Dv \tag{8.24}$$

where A is an operator on $\mathcal{X}$ and B maps $\mathcal{U}$ into $\mathcal{X}$ while C maps $\mathcal{X}$ into $\mathcal{Y}$ and D maps $\mathcal{V}$ into $\mathcal{Y}$. Moreover, the operators $\{A, B, C, D\}$ are all fixed and do not depend upon time t. We also assume that the operator D is onto $\mathcal{Y}$, or equivalently, DD^* is invertible. As before, let $Q(t)$ be the solution to the Riccati differential equation in (8.5) associated with $\{A, B, C, D\}$. The following two theorems are classical results in linear quadratic regulator theory and Riccati differential equations; see [8, 23] for a proof of these theorems.

THEOREM 3.8.3 *Consider the linear time invariant system $\{A, B, C, D\}$ where D is onto. Then there exists a solution $Q(t)$ to the Riccati differential equation in (8.5) for all $t \geq 0$. In other words, the Riccati differential equation in (8.5) does not have a finite escape time. Moreover, the following holds.*

(i) The solution $\{Q(t) : t \geq 0\}$ forms an increasing sequence of positive operators, that is, $Q(\sigma) \leq Q(t)$ for all $\sigma \leq t$.

(ii) If the pair $\{C, A\}$ is observable, then $Q(t)$ converges to a positive operator P as t tends to infinity, that is,

$$P = \lim_{t \to \infty} Q(t)\,. \tag{8.25}$$

In this case, P is a positive solution for the algebraic Riccati equation

$$0 = AP + PA^* + BB^* - PC^*(DD^*)^{-1}CP\,. \tag{8.26}$$

(iii) If $\{A, B, C, D\}$ is controllable and observable, then P is strictly positive.

(iv) If $\{A, B, C, D\}$ is controllable and observable, then $A - PC^*(DD^*)^{-1}C$ is a continuous time stable operator, that is, all the eigenvalues of $A - PC^*(DD^*)^{-1}C$ are contained in the open left half plane.

THEOREM 3.8.4 *Let $\{A, B, C, D\}$ be a controllable and observable time invariant system where D is onto. Then there exists a unique positive solution to the algebraic Riccati equation*

$$0 = AP + PA^* + BB^* - PC^*(DD^*)^{-1}CP\,. \tag{8.27}$$

In this case, $A - PC^*(DD^*)^{-1}C$ is a continuous time stable operator. Moreover, the unique solution to the algebraic Riccati equation in (8.27) is given by

$$P = \lim_{t \to \infty} Q(t) \tag{8.28}$$

where $Q(t)$ is the solution to the Riccati differential equation in (8.5) subject to the initial condition $Q(0) = 0$.

The steady state Kalman filter. Assume that $\{A, B, C, D\}$ is a controllable and observable system. Let P be the unique positive solution to the algebraic Riccati equation in (8.27). By by passing limits in the Kalman filter (1.4), we arrive at the *steady state Kalman filter* defined by

$$\dot{\widehat{x}} \;=\; (A - PC^*(DD^*)^{-1}C)\widehat{x} + PC^*(DD^*)^{-1}y\,. \tag{8.29}$$

Theorem 3.8.3 guarantees that $A - PC^*(DD^*)^{-1}C$ is a continuous time stable operator. The steady state Kalman filter provides an estimate $\widehat{x}(t)$ for the state $x(t)$ given the past $\{y(\sigma) : 0 \le \sigma < t\}$. The Kalman filter converges to the steady state Kalman filter. In other words, the steady state Kalman filter is an optimal state estimator in the limit.

3.8.2 Kalman prediction

In this section we will present a solution to the continuous time Kalman prediction problem. As before, consider the state space system given in (8.2) where $u(t)$ and $v(t)$ are independent white noise processes which are also independent to the initial state $x(0)$. Now consider any future time $t_1 \ge t$. The Kalman prediction problem is to compute the best estimate $\widehat{x}(t_1|t-)$ of the future state $x(t_1)$ given the past output $\{y(\sigma) : 0 \le \sigma < t\}$. To be precise, let $P_{\mathcal{M}_{t-}}$ be the orthogonal projection onto $\mathcal{M}_{t-} = \bigvee\{y(\sigma) : 0 \le \sigma < t\}$. Then the optimal state predictor $\widehat{x}(t_1|t-)$ of the state $x(t_1)$ given $\{y(\sigma) : 0 \le \sigma < t\}$ is determined by

$$\widehat{x}(t_1|t-) = P_{\mathcal{M}_{t-}}x(t_1)\,. \tag{8.30}$$

Notice that if $t_1 = t$, then $\widehat{x}(t|t-) = \widehat{x}(t) = P_{\mathcal{M}_-}x(t)$ is the optimal state estimate $\widehat{x}(t)$ for $x(t)$ computed by the Kalman filter in Theorem 3.8.1. The following result known as the Kalman predictor provides us with an algorithm to compute the optimal state $\widehat{x}(t_1|t-)$.

THEOREM 3.8.5 *Consider the state space system*

$$\dot{x} = Ax + Bu \qquad \textit{and} \qquad y = Cx + Dv \tag{8.31}$$

where $u(t)$ and $v(t)$ are independent white noise random processes, which are independent to the initial condition $x(0)$. Let $\widehat{x}(t) = P_{\mathcal{M}_{t-}}x(t)$ be the optimal state estimate of $x(t)$ given the past $\{y(\sigma) : 0 \le \sigma < t\}$ computed in the Kalman filtering Theorem 3.8.1. Then the optimal state predictor $\widehat{x}(t_1|t-)$ of $x(t_1)$ given the past $\{y(\sigma) : 0 \le \sigma < t\}$ is computed by

$$\widehat{x}(t_1|t-) = \Psi(t_1, t)\widehat{x}(t) \qquad (t_1 \ge t) \tag{8.32}$$

where $\Psi(t, \tau)$ is the state transition matrix for A. Finally, if A is time invariant ($A = A(t)$ for all t), then the optimal state predictor is given by $\widehat{x}(t_1|t-) = e^{A(t_1-t)}\widehat{x}(t)$.

Heuristic Proof. Recall that $x(t_1)$ is given by

$$x(t_1) = \Psi(t_1, t)x(t) + \int_t^{t_1} \Psi(t_1, \tau)B(\tau)u(\tau)\,d\tau\,. \tag{8.33}$$

By consulting (8.8) we see that $u(\tau)$ is orthogonal to $\mathcal{M}_{t-}$ for all $\tau \geq t$. In particular, $P_{\mathcal{M}_{t-}} u(\tau) = 0$ for all $\tau \geq t$. Because the integral is the limit of sums and the orthogonal projection $P_{\mathcal{H}}$ is a linear operator, we formally have $P_{\mathcal{H}} \int f(\tau) d\tau = \int P_{\mathcal{H}} f(\tau) d\tau$ where $f(\tau)$ is a function of random variables and $P_{\mathcal{H}}$ is the orthogonal projection onto the subspace $\mathcal{H}$. Using this in (8.33), we obtain

$$P_{\mathcal{M}_{t-}} x(t_1) = P_{\mathcal{M}_{t-}} \Psi(t_1, t) x(t) + P_{\mathcal{M}_{t-}} \int_t^{t_1} \Psi(t_1, \tau) B(\tau) P_{\mathcal{M}_{t-}} u(\tau)\, d\tau = \Psi(t_1, t) \widehat{x}(t)\,.$$

Therefore $P_{\mathcal{M}_{t-}} x(t_1) = \Psi(t_1, t) \widehat{x}(t)$. This completes the proof.

Steady state Kalman prediction. Now assume that the system $\{A, B, C, D\}$ in (8.31) is controllable, observable and time invariant. As before, u and v are independent white noise processes which are orthogonal to the initial condition x_0. Let $\widehat{x}(t)$ be the steady state estimate computed by the steady state Kalman filter in (8.29), where P is the unique positive solution to the algebraic Riccati equation in (8.26). Motivated by Theorem 3.8.5, the *Kalman steady state predictor* is defined by $\widehat{x}(t_1|t-) = e^{A(t_1-t)} \widehat{x}(t)$ where $\widehat{x}(t)$ is the steady state estimate in (8.29). The steady state Kalman predictor is an optimal state predictor in the limit.

3.8.3 Exercise

Problem 1. Consider a mass spring damper system given by

$$bu = m\ddot{q} + c\dot{q} + kq \qquad \text{and} \qquad y = \dot{q} + Dv\,. \tag{8.34}$$

Assume that the disturbance u and the measurement noise v are independent white noise processes. Notice that the disturbance u is a force on the mass. Now assume that

$$m = 1, \quad c = 1, \quad k = 100, \quad b = 1/2 \quad \text{and} \quad D = 1/4\,.$$

Implement the continuous time steady state Kalman filter in Section 3.8.1 to compute the steady state estimate for the position q and velocity $\dot{q}$ of the mass. In other words, implement the steady state Kalman filter to compute the optimal state estimate $\widehat{x}(t)$ for $x(t)$ where the initial state $x(0) = \begin{bmatrix} 1 & 2 \end{bmatrix}^{tr}$ and $\widehat{x}(0) = \begin{bmatrix} 0 & 0 \end{bmatrix}^{tr}$. Compare your estimate $\widehat{x}(t)$ with the actual state $x(t)$. You may have to use lsim in Matlab to simulate the Kalman filter.

Chapter 4

Wide sense stationary processes

In this chapter we will develop some elementary facts concerning wide sense stationary random processes.

4.1 The autocorrelation function

In this section we define a wide sense stationary random process and its autocorrelation function. Recall that $y(n)$ is a *random process* if $y(n)$ is a random vector in some space $\mathcal{Y}$ for all integers n. In many applications $\mathcal{Y} = \mathbb{C}^k$. A random process $y(n)$ with values in $\mathcal{Y}$ is called *wide sense stationary* if $Ey(n) = c$ is constant for all integers n and $Ey(n)y(m)^* = f(n-m)$ is a function of the time difference $n-m$ for all integers n and m. If $y(n)$ is wide sense stationary, then the *autocorrelation function* $R_y(n)$ for $y(n)$ is defined by $Ey(n)y(0)^* = R_y(n)$. Notice that $R_y(n)$ is an operator on $\mathcal{Y}$ and

$$Ey(n)y(m)^* = R_y(n-m)$$

for all integers n and m. Finally, it is noted that $Ey(n)y(m)^* = f(n-m)$ for all integers n and m if and only if $Ey(n+k)y(k)^* = f(n)$ is a function of just n for all integers n and k. So $y(n)$ is wide sense stationary if and only if $Ey(n) = c$ is constant for all integers n and $Ey(n+k)y(k)^* = f(n)$ is a function of n for all integers n and k. In this case, $R_y(n) = Ey(n+k)y(k)^*$.

If $y(n)$ is a wide sense stationary random process with values in $\mathcal{Y}$, then $R_y(n) = R_y(-n)^*$ for all integers n. To see this imply observe that

$$R_y(n) = Ey(n+k)y(k)^* = (Ey(k)y(n+k)^*)^* = R_y(k-n-k)^* = R_y(-n)^*.$$

Hence $R_y(n) = R_y(-n)^*$ for all n. In particular, if $y(n)$ is a scalar wide sense stationary random process, then $R_y(n) = \overline{R}_y(-n)$ for all integers n. (By scalar valued we mean that $y(n)$ is a random variable in $\mathbb{C}$.)

Let $y(n)$ be a wide sense stationary random process with values in $\mathcal{Y}$. Let g be the random vector defined by

$$g = \begin{bmatrix} y(n) & y(n+1) & \cdots & y(n+\nu-1) \end{bmatrix}^{tr}.$$

Clearly, g is a random vector. Hence $T_\nu = Egg^*$ is a positive matrix. Using $Ey(n)y(m) = R_y(n-m)$, it follows that T_ν is a matrix of the form

$$T_\nu = \begin{bmatrix} R_y(0) & R_y(-1) & \cdots & R_y(1-\nu) \\ R_y(1) & R_y(0) & \cdots & R_y(2-\nu) \\ \vdots & \vdots & \ddots & \vdots \\ R_y(\nu-1) & R_y(\nu-2) & \cdots & R_y(0) \end{bmatrix} \text{ on } \begin{bmatrix} \mathcal{Y} \\ \mathcal{Y} \\ \vdots \\ \mathcal{Y} \end{bmatrix}. \tag{1.1}$$

Notice that the j-k entry of T_ν is given by $\{T_\nu\}_{jk} = R_y(j-k)$. Matrices of the form (1.1) are called *Toeplitz* matrices. The matrix T_ν in (1.1) is referred to as the Toeplitz matrix generated by $\{R_y(k)\}_0^{\nu-1}$. Therefore if $y(n)$ is wide sense stationary, its autocorrelation function uniquely determines a family of a positive Toeplitz matrices T_ν for all integers $\nu \geq 1$. Later we will show how one can construct a wide sense stationary process from a strictly positive Toeplitz matrix.

We say that two random processes $x(n)$ and $y(n)$ are *independent* if the random vectors $x(n)$ and $y(m)$ are independent for all integers n and m. The random processes $x(n)$ and $y(n)$ are *orthogonal* if the random vectors $x(n)$ and $y(m)$ are orthogonal for all integers n and m. Finally, it is noted that if $x(n)$ and $y(n)$ are two mean zero independent processes, then $x(n)$ and $y(n)$ are also orthogonal random processes. To see this observe that in this case $Ex(n)y(m)^* = Ex(n)Ey(m)^* = 0$. Here we used that fact that if two random variables f and g are independent, then $Efg = EfEg$. The following result is useful.

PROPOSITION 4.1.1 *Assume that $y(n) = \sum_{k=1}^{\mu} y_k(n)$ where $y_k(n)$ are mutually orthogonal mean zero wide sense stationary random processes. Then $y(n)$ is a also a mean zero wide sense stationary random process. Moreover, the autocorrelation function for $y(n)$ is given by*

$$R_y(n) = \sum_{k=1}^{\mu} R_{y_k}(n)\,. \tag{1.2}$$

PROOF. Since the mean of $y_k(n)$ is zero, and $y(n) = \sum_{k=1}^{\mu} y_k(n)$, it follows that $Ey(n) = 0$. If $k \neq r$, then $y_k(n)$ is orthogonal to $y_r(m)$ for all integers n and m. Using this orthogonality, we obtain for all integers n and ν

$$Ey(n+\nu)y(\nu)^* = \sum_{k=1}^{\mu}\sum_{r=1}^{\mu} Ey_k(n+\nu)y_r(\nu)^* = \sum_{k=1}^{\mu} Ey_k(n+\nu)y_k(\nu)^* = \sum_{k=1}^{\mu} R_{y_k}(n)\,.$$

Therefore $Ey(n+\nu)y(\nu)^*$ is just a function of n for all integers n and ν. So $y(n)$ is wide sense stationary and $R_y(n)$ is given by (1.2).

4.1.1 A sinusoid process

For an example of a wide sense stationary process, let $\zeta(n)$ be the random process given by $\zeta(n) = a\cos(\omega n + \theta)$ where the amplitude a and the frequency ω are scalars while the phase

θ is a uniform random variable over $[0, 2\pi]$. Recall that the probability density function f_θ for θ is given by

$$\begin{aligned} f_\theta(\phi) &= 1/2\pi \qquad \text{if } 0 \le \phi \le 2\pi \\ &= 0 \qquad\quad \text{otherwise}\,. \end{aligned} \tag{1.3}$$

We claim that $\zeta(n)$ is a mean zero wide sense stationary random process. Moreover, its autocorrelation function $R_\zeta(n) = 2^{-1}|\,a|^2\cos(\omega n)$. To show that the mean of $\zeta(n)$ is zero simply notice that

$$E\zeta(n) = Ea\cos(\omega n + \theta) = \int_{-\infty}^{\infty} a\cos(\omega n + \phi) f_\theta(\phi) d\phi = \frac{1}{2\pi}\int_0^{2\pi} a\cos(\omega n + \phi) d\phi = 0\,.$$

Hence $E\zeta(n) = 0$ for all integers n. Recall for any real α and β, we have

$$\cos(\alpha)\cos(\beta) = \cos(\alpha - \beta)/2 + \cos(\alpha + \beta)/2\,.$$

If n and k are two integers, then this identity yields then

$$\begin{aligned} E\zeta(n+k)\zeta(k)^* &= |\,a|^2 E\cos(\omega(n+k) + \theta)\cos(\omega k + \theta) \\ &= 2^{-1}|\,a|^2 E\cos(\omega n) + 2^{-1}|\,a|^2 E\cos(\omega(n + 2k) + 2\theta) \\ &= 2^{-1}|\,a|^2\cos(\omega n)\,. \end{aligned}$$

Clearly, $E\zeta(n+k)\zeta(k)^*$ is a function of only n for all integers n and k. Thus $\zeta(n)$ is wide sense stationary and $R_\zeta(n) = 2^{-1}|\,a|^2\cos(\omega n)$.

Consider the random process given by

$$y(n) = \sum_{k=1}^{q} a_k \cos(\omega_k n + \theta_k)\,. \tag{1.4}$$

Here we assume that the amplitudes $\{a_k\}_1^q$ and the frequencies $\{\omega_k\}_1^q$ are scalars while the phase $\{\theta_k\}_1^q$ are all independent uniform random variables over $[0, 2\pi]$. Then $y(n)$ is a mean zero wide sense stationary random process whose autocorrelation function is given by

$$R_y(n) = \frac{1}{2}\sum_{k=1}^{q} |\,a_k|^2\cos(\omega_k n)\,. \tag{1.5}$$

To verify this simply observe that

$$y(n) = \sum_{k=1}^{q} y_k(n)$$

where $y_k(n) = a_k\cos(\omega_k n + \theta_k)$ are mean zero wide sense stationary independent random processes for all $1 \le k \le q$. In particular, $y_k(n)$ for $k = 1, 2, \cdots, q$ are mean zero mutually orthogonal wide sense stationary random processes. To see this notice that for $k \ne j$, we have $Ey_k(n)y_j(m) = Ey_k(n)Ey_j(m) = 0$. Here we used the fact that if $f(x)$ and $g(z)$ are functions of two independent random variable x and z, then $Ef(x)g(z) = Ef(x)Eg(z)$. Moreover, our previous analysis with $\zeta(n) = y_k(n)$ shows that $R_{y_k}(n) = |\,a_k|^2\cos(\omega_k n)/2$. By consulting Proposition 4.1.1, it follows that $y(n)$ is a mean zero wide sense stationary processes and its autocorrelation function is given by (1.5).

4.1.2 Exercise

Problem 1. Let $y(n)$ be a scalar valued wide sense stationary random process. Then show that $|R_y(n)| \leq R_y(0)$ for all integers $n \geq 0$. Hint, consider the Cauchy-Schwartz inequality.

Problem 2. Let $y(n)$ be a wide sense stationary random process. Consider the following matrix

$$T = \begin{bmatrix} R_y(0) & R_y(-n) \\ R_y(n) & R_y(0) \end{bmatrix}.$$

Show that T is positive. In particular, T is a self-adjoint operator and $R_y(n)^* = R_y(-n)$.

4.2 A state space realization of sinusoid processes

In this section we will provide a general state space description for a sinusoid process. To motivate our state space representation, consider the wide sense stationary random process $\zeta(n) = a\cos(\omega n + \theta)$ where the amplitude a is nonzero and the frequency ω are scalars while the phase θ is a uniform random variable over $[0, 2\pi]$. Let U on $\mathbb{C}^2$ be the unitary operator, x_0 the random vector in $\mathbb{C}^2$ and C from $\mathbb{C}^2$ into $\mathbb{C}$ be the operator defined by

$$U = \begin{bmatrix} e^{\imath\omega} & 0 \\ 0 & e^{-\imath\omega} \end{bmatrix}, \quad x_0 = \begin{bmatrix} e^{\imath\theta} \\ e^{-\imath\theta} \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} a/2 & a/2 \end{bmatrix}.$$

(Recall that an operator V on a finite dimensional space $\mathcal{V}$ is unitary if $V^*V = I$.) Then using $2\cos(\alpha) = e^{\imath\alpha} + e^{-\imath\alpha}$, we see that $\zeta(n) = CU^n x_0$ for all integers n. Moreover, $Ex_0 = 0$ and $Ex_0x_0^* = I$. In other words, $\zeta(n)$ admits a representation of the form $\zeta(n) = CU^n x_0$ where U is a unitary operator and x_0 is a mean zero random vector satisfying $Ex_0x_0^* = I$. Finally, it is noted that the pair $\{C, U\}$ is observable.

The wide sense stationary random process $y(n)$ in (1.4) also admits a representation of the form $y(n) = CU^n x_0$ for all integers n, where U is a unitary operator on $\mathbb{C}^{2q}$ and x_0 is a mean zero random vector in $\mathbb{C}^{2q}$ satisfying $Ex_0x_0^* = I$. To see this, let U be the diagonal matrix on $\mathbb{C}^{2q}$ and C the operator from $\mathbb{C}^{2q}$ into $\mathbb{C}$ given by

$$\begin{aligned} U &= \operatorname{diag}\{e^{\imath\omega_1}, e^{-\imath\omega_1}, e^{\imath\omega_2}, e^{-\imath\omega_2}, \cdots, e^{\imath\omega_q}, e^{-\imath\omega_q}\} \\ C &= \frac{1}{2}\begin{bmatrix} a_1 & a_1 & a_2 & a_2 & \cdots & a_q & a_q \end{bmatrix} \\ x_0 &= \begin{bmatrix} e^{\imath\theta_1} & e^{-\imath\theta_1} & e^{\imath\theta_2} & e^{-\imath\theta_2} & \cdots & e^{\imath\theta_q} & e^{-\imath\theta_q} \end{bmatrix}^{tr}. \end{aligned}$$

Recall that tr denotes the transpose. Then applying Euler's formula we have $y(n) = CU^n x_0$. Because $\{\theta_k\}_1^q$ are all independent uniform random variables over $[0, 2\pi]$, it follows that the mean of x_0 is zero and $Ex_0x_0^* = I$. Finally, it is noted that if the frequencies $\{\omega_j\}_1^q$ are all distinct and $\{a_j\}_1^q$ are all nonzero, then the pair $\{C, U\}$ is observable; see Proposition 9.6.4 in Chapter 9.

Motivated by this we say that $\{C, U \text{ on } \mathcal{X}\}$ is a unitary pair if U is a unitary operator on $\mathcal{X}$ and C is an operator from $\mathcal{X}$ into $\mathcal{Y}$. Now assume that $\{C, U\}$ is a unitary pair and consider the random process $\xi(n) = CU^n x_0$ where x_0 is a mean zero random vector with

values in $\mathcal{X}$ satisfying $Ex_0x_0^* = I$. Then $\xi(n)$ is a zero mean wide sense stationary random process whose autocorrelation function is given by

$$R_\xi(n) = CU^nC^* \qquad \text{(for all integers } n\text{)}. \tag{2.1}$$

To see this simply notice that $E\xi(n) = CU^nEx_0 = 0$ for all n. So $\xi(n)$ is a mean zero process. Moreover, for any integer ν, we have

$$E\xi(n+\nu)\xi(\nu)^* = ECU^{n+\nu}x_0x_0^*U^{*\nu}C^* = CU^{n+\nu}U^{*\nu}C^* = CU^nC^*\,.$$

Hence $E\xi(n+\nu)\xi(\nu)^*$ is just a function of n. Therefore $\xi(n)$ is a wide sense stationary process and $R_\xi(n) = CU^nC^*$.

We say that $\xi(n)$ is a *sinusoid process* if $\xi(n) = CU^kx_0$ where $\{C, U \text{ on } \mathcal{X}\}$ is a unitary pair and x_0 is a mean zero random vector with values in $\mathcal{X}$ satisfying $Ex_0x_0^* = I$. In this case, $\{C, U; x_0\}$ is called a realization of the process $\xi(n)$. Theorem 4.2.1 below shows that all observable unitary realizations of the same sinusoid process are equivalent up to a unitary transformation.

As before, let $\{C, U \text{ on } \mathcal{X}\}$ be a unitary pair. For a state space realization of the process $\xi(n) = CU^nx_0$, consider the state space system

$$x(n+1) = Ux(n) \qquad \text{and} \qquad \xi(n) = Cx(n) \tag{2.2}$$

where the initial condition $x(0) = x_0$ is a mean zero random vector with values in $\mathcal{X}$ satisfying $Ex_0x_0^* = I$. Notice that $x(n) = U^nx_0$ and $\xi(n) = CU^nx_0$ for all integers $n \geq 0$. Because U is unitary $x(n) = U^*x(n+1)$. So the state space system in (2.2) can also move backwards in time, that is,

$$x(n-1) = U^{-1}x(n) \qquad \text{and} \qquad \xi(n) = Cx(n)\,. \tag{2.3}$$

Therefore the state $x(n) = U^nx_0$ and the output $\xi(n) = Cx(n)$ for all integers n. Finally, it is noted that $x(n)$ is a mean zero wide sense stationary random process whose autocorrelation function is given by $R_x(n) = U^n$.

Two pairs $\{C, U \text{ on } \mathcal{X}\}$ and $\{H, V \text{ on } \mathcal{V}\}$ are *isomorphic* if there exists a unitary operator Φ mapping $\mathcal{V}$ onto $\mathcal{X}$ satisfying $\Phi V = U\Phi$ and $H = C\Phi$. If $\{C, U \text{ on } \mathcal{X}\}$ and $\{H, V \text{ on } \mathcal{V}\}$ are two isomorphic unitary pairs, then $CU^nC^* = HV^nH^*$ for all integers n. So if $\{C, U \text{ on } \mathcal{X}\}$ and $\{H, V \text{ on } \mathcal{V}\}$ are two isomorphic unitary pairs and $\xi(n) = CU^nx_0$ and $z(n) = HV^nv_0$ where x_0 in $\mathcal{X}$ and v_0 in $\mathcal{V}$ are mean zero random vectors satisfying $Ex_0x_0^* = I$ and $Ev_0v_0^* = I$, then $\xi(n)$ and $z(n)$ are two mean zero wide sense stationary process with the same autocorrelation function. This proves part of the following result, which can be viewed as a special case of the Naimark dilation theorem; see [29, 13].

THEOREM 4.2.1 *Let $\{C, U \text{ on } \mathcal{X}\}$ and $\{H, V \text{ on } \mathcal{V}\}$ be two observable unitary pairs where C maps $\mathcal{X}$ into $\mathcal{Y}$ and H maps $\mathcal{V}$ into $\mathcal{Y}$. Let ν any integer satisfying $\nu > \dim \mathcal{X}$. Then $\{C, U\}$ and $\{H, V\}$ are isomorphic if and only if*

$$CU^kC^* = HV^kH^* \qquad (k = 0, 1, \cdots, \nu)\,. \tag{2.4}$$

In particular, if $CU^kC^* = HV^kH^*$ for all integers k, then $\{C, U\}$ and $\{H, V\}$ are isomorphic. Finally, all observable realizations of the same sinusoid process are isomorphic.

PROOF. To complete the proof assume that (2.4) holds. Let μ be the dimension of $\mathcal{X}$, and $\{f_k\}_0^\mu$ be any sequence of vectors in $\mathcal{Y}$. By taking the adjoint in (2.4), we also have $CU^{*k}C^* = HV^{*k}H^*$ for $k = 0, 1, \cdots, \nu$. Using this and (2.4), we obtain

$$\begin{aligned}
\|\sum_{k=0}^{\mu} U^{*k}C^*f_k\|^2 &= (\sum_{k=0}^{\mu} U^{*k}C^*f_k, \sum_{m=0}^{\mu} U^{*m}C^*f_m) = \sum_{k=0}^{\mu}\sum_{m=0}^{\mu}(U^{*k}C^*f_k, U^{*m}C^*f_m) \\
&= \sum_{k=0}^{\mu}\sum_{m=0}^{\mu}(CU^mU^{*k}C^*f_k, f_m) = \sum_{k=0}^{\mu}\sum_{m=0}^{\mu}(CU^{m-k}C^*f_k, f_m) \\
&= \sum_{k=0}^{\mu}\sum_{m=0}^{\mu}(HV^{m-k}H^*f_k, f_m) = \|\sum_{k=0}^{\mu} V^{*k}H^*f_k\|^2 .
\end{aligned}$$

In other words,

$$\|\sum_{k=0}^{\mu} U^{*k}C^*f_k\|^2 = \|\sum_{k=0}^{\mu} V^{*k}H^*f_k\|^2 . \tag{2.5}$$

Because the pair $\{C, U\}$ is observable, $\mathcal{X}$ equals the linear span of $\{U^{*k}C^*\mathcal{Y}\}_0^{\mu-1}$. Equation (2.5) implies that there exists an isometry Ψ mapping $\mathcal{X}$ into $\mathcal{V}$ satisfying

$$\Psi\sum_{k=0}^{\mu} U^{*k}C^*f_k = \sum_{k=0}^{\mu} V^{*k}H^*f_k \qquad (\text{for all } \{f_k\}_0^\mu) . \tag{2.6}$$

In particular, $\Psi U^{*k}C^* = V^{*k}H^*$ for all $k = 0, 1, \cdots, \mu$. According to the Cayley-Hamilton theorem, $U^{*\mu}$ is a linear combination of $\{U^{*k}\}_{k=0}^{\mu-1}$. Since $\Psi U^{*\mu}C^* = V^{*\mu}H^*$, the operator $V^{*\mu}H^*$ is a linear combination of $\{V^{*k}H^*\}_{k=0}^{\mu-1}$. Clearly, $V^{*\mu+1}H^* = V^*V^{*\mu}H^*$. By employing the fact that $V^{*\mu}H^*$ is a linear combination of $\{V^{*k}H^*\}_{k=0}^{\mu-1}$, we see that $V^{*\mu+1}H^*$ is also a linear combination of $\{V^{*k}H^*\}_{k=0}^{\mu-1}$. By continuing in this fashion, $V^{*m}H^*$ is a linear combination of $\{V^{*k}H^*\}_{k=0}^{\mu-1}$ for all integers m. Because the pair $\{H, V\}$ is observable, equation (2.6) shows that Ψ is onto. Therefore Ψ is a unitary operator from $\mathcal{X}$ onto $\mathcal{V}$.

To complete the proof, let $\Phi = \Psi^*$. Then using $\Psi U^{*k}C^* = V^{*k}H^*$, we have $CU^k\Phi = HV^k$ for $k = 0, 1, \cdots, \mu$. In particular, $C\Phi = H$. Moreover,

$$CU^kU\Phi = HV^kV = CU^k\Phi V \qquad (k = 0, 1, \cdots, \mu - 1) .$$

Thus $CU^k(U\Phi - \Phi V) = 0$ for $k = 0, 1, \cdots, \mu - 1$. Because the pair $\{C, U\}$ is observable, $CU^kf = 0$ for all $k = 0, 1, \cdots, \mu - 1$ if and only if $f = 0$. Therefore $U\Phi = \Phi V$. This completes the proof.

4.3 Amplitudes and frequencies of sinusoid processes

In this section we introduce the notion of amplitudes and frequencies associated with a sinusoid process. Throughout this section we assume that the output space $\mathcal{Y} = \mathbb{C}$. Let $\{C, U \text{ on } \mathcal{X}\}$ be a unitary pair and μ the dimension of $\mathcal{X}$. Because U is unitary there exists a unitary operator Φ mapping $\mathbb{C}^\mu$ onto $\mathcal{X}$ satisfying $U\Phi = \Phi V$ where V on $\mathbb{C}^\mu$ is the

diagonal matrix formed by the eigenvalues $\{e^{\imath\omega_k}\}_1^\mu$ of U. (Recall that the eigenvalues of a unitary operator are contained in the unit circle of the complex plane.) Moreover, if H is the operator from $\mathbb{C}^\mu$ into $\mathbb{C}$ defined by $H = C\Phi$, then the pair $\{C, U\}$ is isomorphic to $\{H, V\}$. Clearly, V and H admit matrix representations of the form

$$\begin{aligned} V &= \begin{bmatrix} e^{\imath\omega_1} & 0 & \cdots & 0 \\ 0 & e^{\imath\omega_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\imath\omega_\mu} \end{bmatrix} \text{ on } \mathbb{C}^\mu \\ H &= \begin{bmatrix} a_1 & a_2 & \cdots & a_\mu \end{bmatrix} : \mathbb{C}^\mu \to \mathbb{C}\,. \end{aligned} \tag{3.1}$$

As expected $\{e^{\imath\omega_k}\}_1^\mu$ are the eigenvalues of U and the corresponding angular frequencies $\{\omega_k\}_1^\mu$ are real numbers contained in $[0, 2\pi)$.

As before, let $\xi(n)$ be the sinusoid process determined by $\xi(n) = CU^n x_0$ where x_0 is a mean zero random vector in $\mathcal{X}$ satisfying $Ex_0x_0^* = I$. Recall that $\xi(n)$ is a mean zero wide sense stationary process. Let φ be the random vector with values in $\mathbb{C}^\mu$ defined by $\varphi = \Phi^* x_0$. Since x_0 is a mean zero random vector satisfying $Ex_0x_0^* = I$ and $\Phi^*\Phi = I$, it follows that φ is also a mean zero random vector satisfying $E\varphi\varphi^* = I$. Now let

$$\varphi = \begin{bmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_{\mu-1} & \varphi_\mu \end{bmatrix}^{tr} \tag{3.2}$$

be the components of φ. Because $E\varphi\varphi^* = I$, we have $E\varphi_j\overline{\varphi}_k = \delta_{j,k}$ for all $j, k = 1, 2, \cdots, \mu$. In other words, $\{\varphi_j\}_1^\mu$ is a mean zero orthonormal set of random variables. Using $U\Phi = \Phi V$ and $H = C\Phi$, we obtain

$$\xi(n) = CU^n x_0 = CU^n\Phi\Phi^* x_0 = C\Phi V^n\varphi = HV^n\varphi\,.$$

Thus $\xi(n) = CU^n x_0 = HV^n\varphi$ for all integers n. Combining this (3.1) and (3.2), yields

$$\xi(n) = CU^n x_0 = \sum_{k=1}^{\mu} a_k e^{\imath\omega_k n}\varphi_k\,. \tag{3.3}$$

Therefore any sinusoid process admits a representation of the form $\xi(n) = \sum_{k=1}^{\mu} a_k e^{\imath\omega_k n}\varphi_k$ where $\{\varphi_j\}_1^n$ are mean zero orthonormal random variables. Moreover, the autocorrelation function for $\xi(n)$ is given by

$$R_\xi(n) = CU^n C^* = HV^n H^* = \sum_{k=1}^{\mu} |\,a_k|^2 e^{\imath\omega_k n}\,. \tag{3.4}$$

Finally, it is noted that by multiplying (3.1) by the appropriate unitary diagonal matrix, we can assume without loss of generality that $a_k \geq 0$ for all $k = 1, 2, \cdots, \mu$.

As before, assume that the unitary pair $\{C, U\}$ is isomorphic to the $\{H, V\}$ where V and H are defined in (3.1). Clearly, $\{C, U\}$ is observable if and only if $\{H, V\}$ is observable. Proposition 9.6.4 in Chapter 9 shows that the pair $\{H, V\}$ is observable if and only if the eigenvalues $\{e^{\imath\omega_k}\}_1^\mu$ of U are all distinct and $a_j \neq 0$ for all $j = 1, 2, \cdots, \mu$.

Now let $\{C, U\}$ be an observable unitary pair. Let $\{H, V\}$ be any pair of the form (3.1) which is isomorphic to $\{C, U\}$. Without loss of generality we assume that the frequencies ω_k are in $[0, 2\pi)$ for all $k = 1, 2, \cdots, \mu$. Then we call $\{|\, a_k|, \omega_k\}_1^\mu$ the *amplitudes and frequencies associated* with $\{C, U\}$. If $\xi(n)$ is the wide sense stationary process given by $\xi(n) = CU^n x_0$ where x_0 is a mean zero random vector in $\mathcal{X}$ satisfying $Ex_0x_0^* = I$, then $\{|\, a_k|, \omega_k\}_1^\mu$ is also referred to as the *amplitudes and frequencies associated* with $\xi(n)$. Notice that the amplitudes and frequencies associated with $\{C, U\}$ are unique; see Theorem 4.2.1. Because V is unitarily equivalent to U, it follows that the $\{e^{\imath\omega_k}\}_1^\mu$ are the eigenvalues for U. So the frequencies $\{\omega_k\}_1^\mu$ are uniquely determined by U. By rearranging the terms in $CU^nC^* = \sum_{k=1}^\mu |\, a_k|^2 e^{\imath\omega_k n}$, we arrive at

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ e^{\imath\omega_1} & e^{\imath\omega_2} & \cdots & e^{\imath\omega_\mu} \\ e^{2\imath\omega_1} & e^{2\imath\omega_2} & \cdots & e^{2\imath\omega_\mu} \\ \vdots & \vdots & \cdots & \vdots \\ e^{\imath\omega_1(\mu-1)} & e^{m\imath\omega_2(\mu-1)} & \cdots & e^{m\imath\omega_\mu(\mu-1)} \end{bmatrix} \begin{bmatrix} |\, a_1|^2 \\ |\, a_2|^2 \\ |\, a_3|^2 \\ \vdots \\ |\, a_\mu|^2 \end{bmatrix} = \begin{bmatrix} CC^* \\ CUC^* \\ CU^2C^* \\ \vdots \\ CU^{\mu-1}C^* \end{bmatrix} = \begin{bmatrix} R_\xi(0) \\ R_\xi(1) \\ R_\xi(2) \\ \vdots \\ R_\xi(\mu-1) \end{bmatrix}. \tag{3.5}$$

Notice that the $\mu \times \mu$ matrix in the left hand side of (3.5) is a Vandermonde matrix. Because the pair $\{C, U\}$ is observable all the eigenvalues $\{e^{\imath\omega_k}\}_1^\mu$ of U are distinct. Hence this Vandermonde matrix is invertible. Therefore the amplitudes $\{|\, a_k|\}_1^\mu$ are uniquely determined by the observable unitary pair $\{C, U\}$.

Notice that any sequence $\{|\, a_k|, \omega_k\}_1^\mu$ of amplitudes and frequencies determine a wide sense stationary random process $\xi(n)$. To present a specific sinusoid process associated with $\{|\, a_k|, \omega_k\}_1^\mu$, let φ be the random vector with values in $\mathbb{C}^\mu$ defined by

$$\varphi = \begin{bmatrix} e^{\imath\theta_1} & e^{\imath\theta_2} & e^{\imath\theta_3} & \cdots & e^{\imath\theta_\mu} \end{bmatrix}^{tr} \tag{3.6}$$

where $\{\theta_k\}_1^\mu$ are all independent uniform random variables over $[0, 2\pi]$. Then φ is a mean zero random vector satisfying $E\varphi\varphi^* = I$; see Problem below. As before, let $\{H, V \text{ on } \mathbb{C}^\mu\}$ be the pair unitary pair in (3.1), and set $\xi(n) = HV^n\varphi$. Then $\xi(n)$ is a mean zero wide sense stationary random process of the form

$$\xi(n) = \sum_{k=1}^{\mu} a_k e^{\imath(\omega_k n + \theta_k)}. \tag{3.7}$$

Using $R_\xi(n) = HV^nH^*$, it follows that the autocorrelation function for $\xi(n)$ given by

$$R_\xi(n) = \sum_{k=1}^{\mu} |\, a_k|^2 e^{\imath\omega_k n}. \tag{3.8}$$

Finally, $\{|\, a_k|, \omega_k\}_1^\mu$ are the amplitudes and frequencies associated with $\xi(n)$.

To complete this section, let us directly verify the random process $\xi(n)$ in (3.7) is indeed a mean zero wide sense stationary process and $R_\xi(n)$ is given by (3.8). To this end, let $\zeta(n)$ be the random process given by $\zeta(n) = ae^{\imath(\omega_k n + \theta)}$ where the amplitude a and the frequency ω_k are scalars while the phase θ is a uniform random variable over $[0, 2\pi]$. Recall

that the probability density function f_θ for θ is given by (1.3). We claim that $\zeta(n)$ is a mean zero wide sense stationary random process whose autocorrolation function is given by $R_\zeta(n) = |\, a|^2 e^{\imath\omega_k n}$. To show that the mean of $\zeta(n)$ is zero simply notice that

$$E\zeta(n) = Eae^{\imath(\omega_k n+\theta)} = ae^{\imath\omega_k n}Ee^{\imath\theta} = ae^{\imath\omega_k n}\int_{-\infty}^{\infty} e^{\imath\phi} f_\theta(\phi)d\phi = \frac{ae^{\imath\omega_k n}}{2\pi}\int_0^{2\pi} e^{\imath\phi}d\phi = 0\,. \quad (3.9)$$

Hence $E\zeta(n) = 0$ for all integers n. If n and m are two integers, then

$$E\zeta(n)\zeta(m)^* \;=\; |\, a|^2 Ee^{\imath(\omega_k n+\theta)}e^{-\imath(\omega_k m+\theta)} = |\, a|^2 Ee^{\imath\omega_k(n-m)} = |\, a|^2 e^{\imath\omega_k(n-m)}\,. \quad (3.10)$$

Thus $\zeta(n)$ is wide sense stationary and $R_\zeta(n) = |\, a|^2 e^{\imath\omega_k n}$.

Let $\xi_k(n)$ be the random process given by $\xi_k(n) = a_k e^{\imath(\omega_k n+\theta_k)}$. By consulting (3.9) and (3.10), we see that $\xi_k(n)$ is a mean zero wide sense stationary random processes whose autocorrelation function is given by $R_{\xi_k}(n) = |\, a_k|^2 e^{\imath\omega_k n}$. Moreover,

$$\xi(n) = \sum_{k=1}^{\mu}\xi_k(n) = \sum_{k=1}^{\mu} a_k e^{\imath(\omega_k n+\theta_k)}\,. \quad (3.11)$$

Now let us verify that $\xi(n)$ is a mean zero wide sense stationary random process whose autocorrelation function is given by (3.8).

Since $E\xi_k(n)$ is zero for all k, it follows that $E\xi(n) = 0$. If $k \neq r$, then $\xi_k(n)$ is orthogonal to $\xi_r(m)$ for all integers n and m. To see this observe that $\xi_k(n)$ and $\xi_r(m)$ are independent random variables. Hence

$$E\xi_k(n)\xi_r(m)^* = E\xi_k(n)E\xi_r(m)^* = 0 \qquad (k \neq r)\,.$$

Here we used that fact that if two random variables f and g are independent, then $Efg = EfEg$. So $\xi(n) = \sum_1^\mu \xi_k(n)$ where $\xi_k(n)$ for $k = 1, 2, \cdots, \mu$ are mean zero mutually orthogonal wide sense stationary random processes. According to Proposition 4.1.1, the process $\xi(n)$ is mean zero wide sense stationary and

$$R_\xi(n) = \sum_{k=1}^{\mu} R_{\xi_k}(n) = \sum_{k=1}^{\mu} |\, a_k|^2 e^{\imath\omega_k n}\,.$$

Therefore $R_\xi(n)$ is given by (3.8).

4.3.1 Exercise

Problem 1. Let φ be the random vector with values in $\mathbb{C}^\mu$ defined by

$$\varphi = \begin{bmatrix} e^{\imath\theta_1} & e^{\imath\theta_2} & e^{\imath\theta_3} & \cdots & e^{\imath\theta_\mu} \end{bmatrix}^{tr}$$

where $\{\theta_k\}_1^\mu$ are all independent uniform random variables over $[0, 2\pi]$. Then show that φ is a mean zero random vector satisfying $E\varphi\varphi^* = I$.

Problem 2. Consider the sinusoid process given by $\xi(n) = CU^n x_0$ where

$$U = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad \text{and} \qquad C = \begin{bmatrix} 1 & 2 \end{bmatrix}. \tag{3.12}$$

Here x_0 is a mean zero random vector in $\mathbb{C}^2$ satisfying $Ex_0x_0^* = I$. Find the autocorrelation function $R_\xi(n)$ for $\xi(n)$. Find the amplitudes $\{|a_j|\}_1^2$ and frequencies $\{\omega_j\}_1^2$ for $\xi(n)$.

Problem 3. As in Problem 2, let $\xi(n) = CU^n x_0$ be the sinusoid process determined by U and C in (3.12) while x_0 is a mean zero random vector in $\mathbb{C}^2$ satisfying $Ex_0x_0^* = I$. Consider the sinusoid process $y(n) = a\cos(\omega n + \theta)$ where a and ω are constants, and θ is a uniform random variable over the interval $[0, 2\pi]$. Find a and ω such that $y(n)$ and $\xi(n)$ have the same autocorrelation function.

Problem 4. Consider the sinusoid process given by $\xi(n) = CU^n x_0$ where

$$U = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \text{and} \qquad C = \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}.$$

Here x_0 is a mean zero random vector in $\mathbb{C}^3$ satisfying $Ex_0x_0^* = I$. Find the autocorrelation function $R_\xi(n)$ for $\xi(n)$. Find the amplitudes $\{|a_j|\}_1^2$ and frequencies $\{\omega_j\}_1^2$ for $\xi(n)$.

4.4 State space systems driven by white noise

In this section we will present some results concerning state space systems driven by white noise. We will also give a formula for the spectral density for a wide sense stationary process in state space form.

4.4.1 State space systems with $x(-\infty) = 0$

Let us first review some elementary facts concerning discrete time system. Consider the state space system given by

$$x(n+1) = Ax(n) + Bu(n) \qquad \text{and} \qquad y = Cx(n) + Du(n). \tag{4.1}$$

Here A is an operator on $\mathcal{X}$ and B is an operator mapping $\mathcal{U}$ into $\mathcal{X}$ while C is an operator from $\mathcal{X}$ into $\mathcal{Y}$ and D is an operator mapping $\mathcal{U}$ into $\mathcal{Y}$. The spaces $\mathcal{X}, \mathcal{U}$ and $\mathcal{Y}$ are all $\mathbb{C}^k$ spaces of the appropriate size. The state $x(n)$ is in $\mathcal{X}$, the input $u(n)$ is in $\mathcal{U}$ and the output $y(n)$ is in $\mathcal{Y}$ for all integers n. Now assume that the initial condition $x(m) = x_m$ starts at time m. Obviously, $x(m+1) = Ax_m + Bu(m)$. By recursively solving for the state $x(n)$ in (4.1), we obtain

$$\begin{aligned} x(m+2) &= Ax(m+1) + Bu(m+1) = A^2x_m + ABu(m) + Bu(m+1) \\ x(m+3) &= Ax(m+2) + Bu(m+2) = A^3x_m + A^2Bu(m) + ABu(m+1) + Bu(m+2). \end{aligned}$$

By continuing in this fashion, we see that the solution to (4.1) subject to the initial condition $x(m) = x_m$ is given by

$$x(n) = A^{n-m}x_m + \sum_{k=m}^{n-1} A^{n-k-1}Bu(k) \qquad (x(m) = x_m) \tag{4.2}$$

$$y(n) = CA^{n-m}x_m + Du(n) + \sum_{k=m}^{n-1} CA^{n-k-1}Bu(k)\,. \tag{4.3}$$

Recall that in the discrete time setting an operator A on $\mathcal{X}$ is *stable* if all the eigenvalues of A are contained in the open unit disc $\{z : |z| < 1\}$. Now assume that A is stable and $x_m = f$ is a fixed vector in $\mathcal{X}$. Then $A^{n-m}f$ converges to zero as m tends to minus infinity. Hence as m tends to minus infinity the initial condition x_m does not play a role in the solution to the state space system (4.1); see (4.2) and (4.3). So without loss of generality if the initial condition starts at minus infinity, then we can assume that $x(-\infty) = 0$. Moreover, if $x(-\infty) = 0$, then the solution to the state space system in (4.1) is given by

$$x(n) = \sum_{k=-\infty}^{n-1} A^{n-k-1}Bu(k) \quad \text{and} \quad y(n) = Du(n) + \sum_{k=-\infty}^{n-1} CA^{n-k-1}Bu(k)\,. \tag{4.4}$$

To conclude this section, let us recall that the transfer function for the state space system given in (4.1) is defined by

$$C(zI - A)^{-1}B + D\,. \tag{4.5}$$

Here z is a complex variable.

4.4.2 The spectral density for state space systems

If $y(n)$ is a wide sense stationary random process, then the *spectral density* $S_y(\omega)$ for $y(n)$ is defined by

$$S_y(\omega) = \sum_{k=-\infty}^{\infty} e^{-\imath\omega k}R_y(k) \qquad (\omega \in [0, 2\pi])\,. \tag{4.6}$$

The spectral density $S_y(\omega)$ is simply the Fourier transform of its autocorrelation function $\{R_y(k)\}_{-\infty}^{\infty}$. The spectral density is studied in more detail in Chapter 5.

Recall that $u(n)$ is a white noise random process, if $u(n)$ is a mean zero wide sense stationary process whose autocorrelation function is given by $R_u(n) = \delta_{n,\,0}$ where $\delta_{n,\,m}$ is the Kronecker delta. Now consider the state space system given by

$$x(n+1) = Ax(n) + Bu(n) \qquad \text{and} \qquad y = Cx(n)$$

where $u(n)$ is a white noise random process. Here A, B and C are operators acting between the appropriate spaces. Clearly, the output $y(n)$ in (3.3) is a random process. However, $y(n)$ is not necessarily wide sense stationary.

If A is stable and the initial condition starts at time minus infinity, that is, $x(-\infty) = 0$, then both $x(n)$ and $y(n)$ are wide sense stationary. To be specific, consider the state space system given by

$$x(n+1) = Ax(n) + Bu(n) \qquad (x(-\infty) = 0) \qquad \text{and} \qquad y(n) = Cx(n) \tag{4.7}$$

where A is stable and $u(n)$ is a white noise random process. This system is denoted by $\{A, B, C, x, y\}$. The initial condition $x(-\infty)$ does not have to be zero at $-\infty$ we just wanted to emphasize that the state space system in (4.7) starts at time minus infinity. According to (4.4) with $D = 0$, the solution to (4.7) is given by

$$x(n) = \sum_{k=-\infty}^{n-1} A^{n-k-1}Bu(k) \qquad \text{and} \qquad y(n) = \sum_{k=-\infty}^{n-1} CA^{n-k-1}Bu(k)\,. \tag{4.8}$$

Recall that the controllability Gramian Q for the pair $\{A, B\}$ is the unique solution to the Lyapunov equation

$$Q = AQA^* + BB^*\,. \tag{4.9}$$

The following result shows that y is wide sense stationary.

THEOREM 4.4.1 *Let $\{A, B, C, x, y\}$ be the system given by (4.7) where A is a stable operator on $\mathcal{X}$ and $u(n)$ is a white noise random process. Let Q be the controllability Gramian for the pair $\{A, B\}$. Then $x(n)$ and $y(n)$ are both mean zero wide sense stationary random processes. The state covariance $Ex(n)x(n)^* = Q$ for all integers n. Moreover, the autocorrelation function $R_y(n)$ for $y(n)$ is given by*

$$\begin{aligned} R_y(n) &= CA^nQC^* \qquad \text{if } n \geq 0 \\ &= CQA^{*|n|}C^* \qquad \text{if } n \leq 0\,. \end{aligned} \tag{4.10}$$

Finally, the spectral density for $y(n)$ is given by

$$S_y(\omega) = C(e^{\imath\omega}I - A)^{-1}BB^*(e^{-\imath\omega}I - A^*)^{-1}C^* \qquad (\omega \in [0, 2\pi])\,. \tag{4.11}$$

In other words, $S_y(\omega) = \mathbf{G}(e^{\imath\omega})\mathbf{G}(e^{\imath\omega})^*$ where $\mathbf{G}(z) = C(zI - A)^{-1}B$ is the transfer function for $\{A, B, C\}$.

Proof. Because the mean of $u(n)$ is zero, we have

$$Ex(n) = \sum_{k=-\infty}^{n-1} A^{n-k-1}BEu(k) = 0\,.$$

Hence $Ex(n) = 0$ for all integers n. Recall that $Q = \sum_0^\infty A^kBB^*A^{*k}$ is the solution to the Lyapunov equation in (4.9). Now assume that $n \geq m$. Using the formula for $x(n)$ in (4.8)

and $Eu(k)u(\nu)^* = \delta_{k,\nu}$, we obtain

$$\begin{aligned} Ex(n)x(m)^* &= E\left(\sum_{k=-\infty}^{n-1} A^{n-k-1}Bu(k)\right)\left(\sum_{\nu=-\infty}^{m-1} A^{m-\nu-1}Bu(\nu)\right)^* \\ &= \sum_{k=-\infty}^{n-1}\sum_{\nu=-\infty}^{m-1} A^{n-k-1}BEu(k)u(\nu)^*B^*A^{*(m-\nu-1)} \\ &= \sum_{k=-\infty}^{m-1} A^{n-\nu-1}BB^*A^{*(m-k-1)} = \sum_{k=-\infty}^{m-1} A^{n-m}A^{m-k-1}BB^*A^{*(m-k-1)} \\ &= A^{n-m}\sum_{k=0}^{\infty} A^kBB^*A^{*k} = A^{n-m}Q\,. \end{aligned}$$

Thus $Ex(n)x(m)^* = A^{n-m}Q$ for all integers $n \geq m$. In particular, the state covariance $Ex(n)x(n)^* = Q$ for all integers n. If $n \leq m$, then our previous analysis yields

$$Ex(n)x(m)^* = (Ex(m)x(n)^*)^* = \left(A^{m-n}Q\right)^* = QA^{*|n-m|}\,.$$

Therefore $x(n)$ is a wide sense stationary process and

$$\begin{aligned} R_x(n) &= A^nQ \qquad \text{if } n \geq 0 \\ &= QA^{*|n|} \qquad \text{if } n \leq 0\,. \end{aligned} \tag{4.12}$$

Since the mean of $x(n)$ is zero and $y(n) = Cx(n)$, it follows that $y(n)$ is a zero mean process. Finally, observe that

$$Ey(n+\nu)y(\nu)^* = ECx(n+\nu)x(\nu)^*C^* = CR_x(n)C^*\,.$$

Hence $y(n)$ is a wide sense stationary process and $R_y(n) = CR_x(n)C^*$. By consulting (4.12) we arrive at the formula for $R_y(n)$ in (4.10).

To complete the proof it remains to establish the formula for the spectral density in (4.11). Let $z = e^{\imath\omega}$. Lemma 4.4.2 below shows that $(I - zA)^{-1} = \sum_0^\infty z^kA^k$. Using this Fourier series expansion along with the formula for $R_y(n)$ in (4.10), we have

$$\begin{aligned} S_y(\omega) &= \sum_{k=-\infty}^{\infty} \overline{z}^{\,k}R_y(k) = \sum_{k=0}^{\infty} \overline{z}^{\,k}CA^kQC^* + \sum_{k=0}^{\infty} z^kCQA^{*k}C^* - CQC^* \\ &= C(I-\overline{z}A)^{-1}QC^* + CQ(I-zA^*)^{-1}C^* - CQC^* \\ &= C(I-\overline{z}A)^{-1}\left[Q(I-zA^*) + (I-\overline{z}A)Q - (I-\overline{z}A)Q(I-zA^*)\right](I-zA^*)^{-1}C^* \\ &= C(I-\overline{z}A)^{-1}\left[Q - AQA^*\right](I-zA^*)^{-1}C^* \\ &= C(I-\overline{z}A)^{-1}BB^*(I-zA^*)^{-1}C^* = C(e^{\imath\omega}I-A)^{-1}BB^*(e^{-\imath\omega}I-A^*)^{-1}C^*\,. \end{aligned}$$

Therefore (4.11) holds. This completes the proof.

To conclude this section, let $\{C, U \text{ on } \mathcal{X}\}$ be a unitary pair. Let $\xi(n)$ be the wide sense stationary sinusoid process given by $\xi(n) = CU^nx_0$ where x_0 is a mean zero vector with values in $\mathcal{X}$ satisfying $Ex_0x_0^* = I$. Recall that

$$x(n+1) = Ux(n) \qquad \text{and} \qquad \xi(n) = Cx(n)$$

subject to the initial condition $x(0) = x_0$. Let B be the zero operator from $\mathbb{C}$ into $\mathcal{X}$. Obviously, the pair $\{U, B\}$ is not controllable. However, $Q = I$ is the solution to the Lyapunov equation $Q = UQU^* + BB^*$. Hence $R_\xi(n) = CU^nC^* = CU^nQC^*$. In other words, the formula for the autocorrelation function in (4.10) also works for the process $\xi(n) = CU^nx_0$ generated by a unitary pair $\{C, U\}$ with $B = 0$.

4.4.3 Geometric series

The following lemma generalizes the classical result that if λ is a scalar in the open unit disc, then the geometric series $\sum_0^\infty \lambda^k$ converges to $(1-\lambda)^{-1}$.

LEMMA 4.4.2 *Assume that one is not an eigenvalue for an operator T on $\mathcal{X}$. Then*

$$\sum_{k=0}^{\nu-1} T^k = (I - T^\nu)(I - T)^{-1}\,. \tag{4.13}$$

Moreover, if T is stable, then

$$\sum_{k=0}^{\infty} T^k = (I - T)^{-1}\,. \tag{4.14}$$

In particular, for any complex number $\lambda \neq 1$, we have

$$\sum_{k=0}^{\nu-1} \lambda^k = \frac{1-\lambda^\nu}{1-\lambda} \qquad (\textit{if } \lambda \neq 1) \qquad \textit{and} \qquad \sum_{k=0}^{\infty} \lambda^k = \frac{1}{1-\lambda} \qquad (\textit{if } |\lambda| < 1)\,. \tag{4.15}$$

Proof. Let Z be the operator on $\mathcal{X}$ defined by

$$Z = \sum_{k=0}^{\nu-1} T^k = I + T + T^2 + \cdots + T^{\nu-1}\,.$$

Then $ZT = T + T^2 + T^3 + \cdots + T^\nu$. Subtracting this from the previous equation for Z yields Thus $Z(I-T) = I - T^\nu$. Because one is not an eigenvalue for T, we can invert $I - T$. Hence $Z = (I - T^\nu)(I - T)^{-1}$ and (4.13) holds.

If T is stable, then T^ν converges to zero as ν tends to infinity. So by letting ν approach infinity in (4.13), we obtain $\sum_0^\infty T^k = (I - T)^{-1}$. This completes the proof.

4.4.4 Exercise

One can also use the fast Fourier transform (fft) to compute the autocorrelation function and spectral density for a single input single output state space system. To this end, consider the wide sense stationary system given by

$$x(n+1) = Ax(n) + Bu(n) \qquad (x(-\infty) = 0) \qquad \text{and} \qquad y = Cx(n)$$

where A is a stable operator on $\mathcal{X}$ and the output $\mathcal{U} = \mathcal{Y} = \mathbb{C}$. Let $\mathbf{G}(z) = C(zI - A)^{-1}B$ be the transfer function for $\{A, B, C\}$. (Notice that we are using a boldface $\mathbf{G}$ to represent a transfer function.) According to Theorem 4.4.1, the spectral density for $S_y(\omega)$ is given by

$$\sum_{k=-\infty}^{\infty} e^{-\imath\omega k} R_y(k) = S_y(\omega) = |\mathbf{G}(e^{\imath\omega})|^2 .$$

The following Matlab commands can be used to compute the autocorrelation function and spectral density for $y(n)$:

$$\begin{aligned} [p, q] &= \text{ss2tf}(A, B, C, 0); \\ g &= \text{fft}(p, 4096)./\text{fft}(q, 4096); \\ S &= \text{abs}(g).^2; \\ &\quad \text{plot}(\text{linspace}(0, 2\pi, 4096), S);\ \text{grid} \\ R &= \text{ifft}(S); \end{aligned}$$

Then S is the fast Fourier transform approximation for the spectral density S_y. Moreover, R is the fast Fourier transform approximation for the autocorrelation function R_y, that is, R is a row vector of length 4096 of the form

$$R = \begin{bmatrix} R_y(0) & R_y(1) & R_y(2) & \cdots & R_y(-3) & R_y(-2) & R_y(-1) \end{bmatrix} .$$

Finally, it is noted that in most applications $\{A, B, C\}$ are real matrices. In this case, one may use $R = \text{real}(\text{ifft}(S))$; to compute R. Because $\{A, B, C\}$ is real, the autocorrelation function $R_y(n)$ is real for all n, and computing $\text{real}(\text{ifft}(S))$; eliminates some small imaginary numbers computed by the ifft.

Problem 1. Consider the state space system given by

$$\begin{aligned} \begin{bmatrix} x_1(n+1) \\ x_2(n+1) \\ x_3(n+1) \\ x_4(n+1) \end{bmatrix} &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -0.1 & 0.4 & -1.4 & 2 \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \\ x_4(n) \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} u(n) \\ y(n) &= \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix} x(n) . \end{aligned}$$

The initial condition $x(-\infty) = 0$. Use Theorem 4.4.1 with Matlab to compute the autocorrelation function $R_y(n)$ for the process $y(n)$. Now use the fast Fourier transform to compute the autocorrelation function $R_y(n)$ for $y(n)$. Compare $\{R_y(n)\}_0^5$ by both of these methods. Finally, use the Fast Fourier transform to plot the spectral density $S_y(\omega)$ for $y(n)$.

Problem 2. Consider the time invariant state space system given by

$$\begin{aligned} x(n+1) &= Ax(n) + Bu(n) \qquad (x(-\infty) = 0) \\ y(n) &= Cx(n) + Dv(n) \end{aligned} \tag{4.16}$$

where A on $\mathcal{X}$ is stable. As expected, B is an operator from $\mathcal{U}$ into $\mathcal{X}$ and C is an operator from $\mathcal{X}$ into $\mathcal{Y}$ while D is an operator from $\mathcal{V}$ into $\mathcal{Y}$. Moreover, assume that $u(n)$ and $v(n)$ are independent white noise random processes. Notice that the state space system in (4.16) starts at time minus infinity. The solution to (4.16) is given by

$$x(n) = \sum_{k=-\infty}^{n-1} A^{n-k-1}Bu(k) \quad \text{and} \quad y(n) = \sum_{k=-\infty}^{n-1} CA^{n-k-1}Bu(k) + Dv(n)\,. \tag{4.17}$$

As before, let Q be the controllability Gramian for the pair $\{A, B\}$. Then show that $y(n)$ is a mean zero wide sense stationary process whose autocorrelation function is given by

$$\begin{aligned} R_y(n) &= CA^nQC^* && \text{if } n > 0 \\ &= CQC^* + DD^* && \text{if } n = 0 \\ &= CQA^{*|n|}C^* && \text{if } n < 0\,. \end{aligned} \tag{4.18}$$

Finally, show that the spectral density for y is given by

$$S_y(\omega) = C(e^{\imath\omega}I - A)^{-1}BB^*(e^{-\imath\omega}I - A^*)^{-1}C^* + DD^* \qquad (\omega \in [0, 2\pi])\,. \tag{4.19}$$

Chapter 5

The spectral density

This chapter is devoted to the spectral density for a wide sense stationary random process.

5.1 Fourier series

In this section we review some elementary facts concerning Fourier series. To this end, let ℓ^2 be the Hilbert space consisting of all vectors a of the form

$$a = \begin{bmatrix} \cdots & a_{-2} & a_{-1} & a_0 & a_1 & a_2 & \cdots \end{bmatrix}^{tr} \quad \text{where} \quad \| a\|^2 = \sum_{k=-\infty}^{\infty} | a_k|^2 < \infty \,. \tag{1.1}$$

For convenience the vector a in (1.1) will also be represented as $a = \{a_k\}_{-\infty}^{\infty}$. If $a = \{a_k\}_{-\infty}^{\infty}$ and $b = \{b_k\}_{-\infty}^{\infty}$ are two vectors in ℓ^2, then the inner product between a and b is defined by

$$(a, b) = \sum_{k=-\infty}^{\infty} a_k \bar{b}_k \qquad (a = \{a_k\}_{-\infty}^{\infty} \text{ and } b = \{b_k\}_{-\infty}^{\infty}) \in \ell^2) \,. \tag{1.2}$$

The Cauchy-Schwartz inequality $|(f, g)| \leq \|f\| \|g\|$ guarantees that the infinite sum $\sum_{k=-\infty}^{\infty} a_k \bar{b}_k$ is finite for all a and b in ℓ^2.

Now let L^2 be the Hilbert generated by all Lebesgue measurable function f on $[0, 2\pi]$ satisfying

$$\|f\|^2 = \frac{1}{2\pi} \int_0^{2\pi} |f(\omega)|^2 \, d\omega < \infty \,. \tag{1.3}$$

If f and g are two function in L^2, then the inner product between f and g is defined by

$$(f, g) = \frac{1}{2\pi} \int_0^{2\pi} f(\omega) \overline{g(\omega)} \, d\omega \,. \tag{1.4}$$

Obviously, $e^{-\imath k\omega}$ is a function in L^2 any integer k. Moreover, is easy to show that the set of functions $e^{-\imath k\omega}$ for $k = 0, \pm 1, \pm 2, \cdots$ are orthonormal in L^2, that is,

$$(e^{-\imath k\omega}, e^{-\imath m\omega}) = \frac{1}{2\pi} \int_0^{2\pi} e^{-\imath k\omega} e^{\imath m\omega} \, d\omega = \delta_{k,\,m} \,. \tag{1.5}$$

(Recall that $\delta_{k,\,m}$ is the Kronecker delta.) Furthermore, it is well known that $e^{-\imath k\omega}$ for $k = 0, \pm 1, \pm 2, \cdots$ forms an orthonormal basis for L^2. To be precise, any function f in L^2 admits a Fourier series expansion of the form

$$f = \sum_{k=-\infty}^{\infty} (f, e^{-\imath k\omega}) e^{-\imath k\omega} \qquad (f \in L^2)\,. \tag{1.6}$$

In particular, this implies that

$$\|f\|^2 = \frac{1}{2\pi} \int_0^{2\pi} |f(\omega)|^2 \, d\omega = \sum_{k=-\infty}^{\infty} |(f, e^{-\imath k\omega})|^2\,. \tag{1.7}$$

The Fourier transform $\mathcal{F}$ is the unitary operator from ℓ^2 onto L^2 defined by

$$f(\omega) = \mathcal{F}a = \sum_{k=-\infty}^{\infty} a_k e^{-\imath\omega k} \qquad (a = \{a_k\}_{-\infty}^{\infty} \in \ell^2)\,. \tag{1.8}$$

The inverse Fourier transform is the unitary operator from L^2 into ℓ^2 defined by

$$\mathcal{F}^{-1} f \;\; = \;\; a = \big[\, \cdots \quad a_{-2} \quad a_{-1} \quad a_0 \quad a_1 \quad a_2 \quad \cdots \,\big]^{tr} \qquad (f \in L^2) \tag{1.9}$$

where $\{a_k\}_{-\infty}^{\infty}$ are computed by

$$a_k = (f, e^{-\imath\omega k}) = \frac{1}{2\pi} \int_0^{2\pi} f(\omega) e^{\imath\omega k} \, d\omega \qquad (k = 0, \pm 1, \pm 2, \cdots)\,. \tag{1.10}$$

If f is the Fourier transform of a and g is the Fourier transform of b, then $(f, g)_{L^2} = (\mathcal{F}a, \mathcal{F}b)_{L^2} = (a, b)_{\ell^2}$. In other words,

$$(f, g) = \frac{1}{2\pi} \int_0^{2\pi} f(\omega)\overline{g(\omega)} \, d\omega = \sum_{k=-\infty}^{\infty} a_k \bar{b}_k \qquad (f = \mathcal{F}\{a_k\}_{-\infty}^{\infty} \text{ and } g = \mathcal{F}\{g_k\}_{-\infty}^{\infty})\,. \tag{1.11}$$

In particular, this yields Parseval's equality

$$\|f\|^2 = \frac{1}{2\pi} \int_0^{2\pi} |f(\omega)|^2 \, d\omega = \sum_{k=-\infty}^{\infty} |\, a_k|^2 \qquad (f = \mathcal{F}\{a_k\}_{-\infty}^{\infty})\,. \tag{1.12}$$

To verify that (1.11) holds notice that $f = \sum_{-\infty}^{\infty} a_k e^{-\imath\omega k}$ and $g = \sum_{-\infty}^{\infty} b_k e^{-\imath\omega k}$ yields

$$(f, g) = \sum_{k=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} a_k \bar{b}_m (e^{-\imath\omega k}, e^{-\imath\omega m}) = \sum_{k=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} a_k \bar{b}_m \delta_{k,\,m} = \sum_{k=-\infty}^{\infty} a_k \bar{b}_m\,.$$

Therefore (1.11) holds.

5.1.1 The Fourier transform of sinusoids

In some applications we will have to take the Fourier transform of certain sequences $\{a_k\}_{-\infty}^{\infty}$ not in ℓ^2. To get around this problem we use the delta Dirac function $\delta(\omega)$. Here we only need the delta Dirac function on the interval $[0, 2\pi)$. Recall that if f is any continuous function, then delta Dirac function has the following property

$$\int_{-\infty}^{\infty} f(\omega)\delta(\omega - \omega_0)\, d\omega = f(\omega_0)\,. \tag{1.13}$$

The delta function is formally obtained as the limit of positive functions $g_k(\omega)$ satisfying

$$\lim_{k\to\infty} g_k(\omega) = 0 \text{ if } \omega \neq 0 \qquad \text{and} \qquad \int_{-\infty}^{\infty} g_k(\omega)\, d\omega = 1 \text{ for all } k\,. \tag{1.14}$$

If g_k is a sequence of functions satisfying (1.14), then $g_k(\omega)$ converges weakly to $\delta(\omega)$ as k tends to infinity. By weakly converge we mean that if f is any continuous function, then

$$f(\omega_0) = \lim_{k\to\infty} \int_{-\infty}^{\infty} f(\omega) g_k(\omega - \omega_0)\, d\omega\,.$$

For an example of a sequence positive function g_k satisfying (1.14) consider $g_k(\omega) = 1/k$ if $\omega \in [0, 1/k]$ and $g_k(\omega) = 0$ otherwise. For another example, let g_k be the function determined by the Gaussian density $g_k(\omega) = ke^{-(k\omega)^2/2}/\sqrt{2\pi}$ with mean zero and standard deviation $1/k$.

Now consider the sequence $\{e^{\imath\omega_0 k}\}_{-\infty}^{\infty}$ where ω_0 is a frequency in $[0, 2\pi]$. Clearly, $\{e^{\imath\omega_0 k}\}_{-\infty}^{\infty}$ is not in ℓ^2. In this case, the Fourier transform of $\{e^{\imath\omega_0 k}\}_{-\infty}^{\infty}$ is defined as $2\pi\delta(\omega - \omega_0)$, that is,

$$2\pi\delta(\omega - \omega_0) = \mathcal{F}\{e^{\imath\omega_0 k}\}_{-\infty}^{\infty}\,. \tag{1.15}$$

To verify that this makes sense simply observe that

$$e^{\imath\omega_0 k} = \frac{1}{2\pi}\int_0^{2\pi} e^{\imath\omega k} 2\pi\delta(\omega - \omega_0)\, d\omega \qquad \text{(for all integers } k)\,. \tag{1.16}$$

So at least formally $\mathcal{F}^{-1}2\pi\delta(\omega - \omega_0) = \{e^{\imath\omega_0 k}\}_{-\infty}^{\infty}$. Now let $\{\omega_0\}_1^{\mu}$ be a set of frequencies in $[0, 2\pi)$ and $\{\omega_0\}_1^{\mu}$ a set of scalars. Because the Fourier transform is linear, we also have

$$\sum_{j=1}^{\mu} 2\pi\gamma_j\delta(\omega - \omega_j) = \mathcal{F}\left\{\sum_{j=1}^{\mu} \gamma_j e^{\imath\omega_j k}\right\}_{k=-\infty}^{\infty}\,. \tag{1.17}$$

To gain some further insight, let us try to compute the Fourier transform for the sequence

$$b_k(r) \;=\; e^{\imath\omega_0 k}/r^{|k|} \qquad (r > 1)\,.$$

Here r is a any positive scalar such that $r > 1$. Notice that as r approaches one from above, $b_k(r)$ converges to $e^{\imath\omega_0 k}$. Moreover, for every $r > 1$, the sequence $\{b_k(r)\}_{k=-\infty}^{\infty}$ is in ℓ^2. Hence the Fourier transform $g_r = \mathcal{F}\{b_k(r)\}_{-\infty}^{\infty}$ is a well defined function in L^2. Recall that

$\sum_0^\infty \lambda^k = (1-\lambda)^{-1}$ when $|\lambda| < 1$; see equation (4.15) in Lemma 4.4.2. Let $z = re^{\imath(\omega-\omega_0)}$. Using this we see that g_r is given by

$$\begin{aligned} g_r(\omega) &= \sum_{k=-\infty}^{\infty} e^{-\imath\omega k} e^{\imath\omega_0 k}/r^{|k|} = \sum_{k=0}^{\infty} 1/z^k + \sum_{k=1}^{\infty} 1/\overline{z}^{\,k} \\ &= \frac{1}{1-1/z} + \frac{1/\overline{z}}{1-1/\overline{z}} = \frac{z}{z-1} + \frac{1}{\overline{z}-1} \\ &= \frac{|z|^2-1}{|z-1|^2} = \frac{r^2-1}{|re^{\imath(\omega-\omega_0)}-1|^2} \\ &= \frac{r^2-1}{|r\cos(\omega-\omega_0)-1+\imath r\sin(\omega-\omega_0)|^2} \\ &= \frac{r^2-1}{r^2-2r\cos(\omega-\omega_0)+1}\,. \end{aligned}$$

Therefore

$$g_r(\omega) = \mathcal{F}\{b_k(r)\}_{-\infty}^{\infty} = \frac{r^2-1}{r^2-2r\cos(\omega-\omega_0)+1} := P_r(\omega-\omega_0) \qquad (\text{if } r>1)\,. \tag{1.18}$$

The function $P_r(\omega-\omega_0)$ is called the *Poisson kernel*. It is well known that $P_r(\omega-\omega_0)$ weakly converges to the delta Dirac function $2\pi\delta(\omega-\omega_0)$ as r tends to one from above; see [19]. Notice that if $\omega \neq \omega_0$, then $P_r(\omega-\omega_0)$ converges to zero as r approaches one. Clearly, $P_r(\omega-\omega_0)$ is positive. Moreover,

$$\frac{1}{2\pi}\int_0^{2\pi} P_r(\omega-\omega_0)\,d\omega = \frac{1}{2\pi}\int_0^{2\pi} g_r(\omega)\,d\omega = b_0(r) = 1\,.$$

In other words, the area under $P_r(\omega-\omega_0)/2\pi$ is one. So $P_r(\omega-\omega_0)/2\pi$ converges to the delta Dirac function $\delta(\omega-\omega_0)$ as r tends to one from above.

Let us directly show that $P_r(\omega-\omega_0)/2\pi$ weakly converges to $\delta(\omega-\omega_0)$ as r tends to one. To this end, let f be any function in L^2 with Fourier coefficients $\{a_k\}_{-\infty}^{\infty}$, that is, $f(\omega) = \mathcal{F}\{a_k\}_{-\infty}^{\infty}$. Then using $f_r(\omega) = \mathcal{F}\{b_k(r)\}_{-\infty}^{\infty}$ with $r>1$, we have

$$\frac{1}{2\pi}\int_0^{2\pi} f(\omega)P_r(\omega-\omega_0)\,d\omega = \frac{1}{2\pi}\int_0^{2\pi} f(\omega)\overline{g_r(\omega)}\,d\omega = \sum_{k=-\infty}^{\infty} a_k\overline{b_k(r)} = \sum_{k=-\infty}^{\infty} a_k e^{-\imath\omega_0 k}/r^{|k|}\,.$$

By letting r approach one from above, the previous equation yields

$$\lim_{r\to 1}\frac{1}{2\pi}\int_0^{2\pi} f(\omega)P_r(\omega-\omega_0)\,d\omega = \lim_{r\to 1}\sum_{k=-\infty}^{\infty} a_k e^{-\imath\omega_0 k}/r^{|k|} = f(\omega_0) = \int_0^{2\pi} f(\omega)\delta(\omega-\omega_0)\,d\omega\,.$$

Therefore $P_r(\omega-\omega_0)/2\pi$ weakly converges to the delta Dirac function $\delta(\omega-\omega_0)$ as r tends to one from above.

5.1.2 Exercise

Problem 1. Plot the Poisson kernel in $P_r(\omega - \pi)/2\pi$ in Matlab for several values of r approaching one form above.

5.2 Fourier transforms and positive Toeplitz matrices

The following result develops a fundamental relationship between positive Toeplitz matrices and Fourier transforms. Finally, if $\{G(k)\}_{-\infty}^{\infty}$ is a sequence of operators, then its Fourier transform is defined by $\sum_{-\infty}^{\infty} e^{-\imath\omega k} G(k)$.

THEOREM 5.2.1 *Let $\{R(k)\}_{-\infty}^{\infty}$ be a sequence of operators on $\mathcal{Y}$ and $S(\omega) = \sum_{-\infty}^{\infty} e^{-\imath\omega k} R(k)$ the Fourier transform of $\{R(k)\}_{-\infty}^{\infty}$. Then $S(\omega) \geq 0$ for all ω in $[0, 2\pi]$ if and only if the block Toeplitz matrices*

$$T_\nu = \begin{bmatrix} R(0) & R(-1) & \cdots & R(1-\nu) \\ R(1) & R(0) & \cdots & R(2-\nu) \\ \vdots & \vdots & \ddots & \vdots \\ R(\nu-1) & R(\nu-2) & \cdots & R(0) \end{bmatrix} \text{ on } \begin{bmatrix} \mathcal{Y} \\ \mathcal{Y} \\ \vdots \\ \mathcal{Y} \end{bmatrix} \tag{2.1}$$

are positive for all integers $\nu \geq 1$.

Proof. Let $p(\omega)$ be any polynomial of the form $p(\omega) = \sum_{k=0}^{\nu-1} e^{-\imath\omega k} f_k$ where f_k is in $\mathcal{Y}$ for all integers $0 \leq k \leq \nu - 1$. Let f be the vector given by $f = [f_0, f_1, \ldots, f_{\nu-1}]^{tr}$. Notice that each component f_k of f is in $\mathcal{Y}$. Then we claim that

$$\frac{1}{2\pi} \int_0^{2\pi} (S(\omega)p(\omega), p(\omega))\, d\omega = (T_\nu f, f)\,. \tag{2.2}$$

Using $S(\omega) = \sum_{-\infty}^{\infty} e^{-\imath\omega k} R(k)$, we obtain

$$\begin{aligned} \frac{1}{2\pi} \int_0^{2\pi} (S(\omega)p(\omega), p(\omega))\, d\omega &= \sum_{k=-\infty}^{\infty} \sum_{m=0}^{\nu-1} \sum_{\mu=0}^{\nu-1} \frac{1}{2\pi} \int_0^{2\pi} e^{-\imath\omega(k+m)} e^{\imath\omega\mu} (R(k) f_m, f_\mu)\, d\omega \\ &= \sum_{k=-\infty}^{\infty} \sum_{m=0}^{\nu-1} \sum_{\mu=0}^{\nu-1} (R(k) f_m, f_\mu) \delta_{k+m,\,\mu} \\ &= \sum_{m=0}^{\nu-1} \sum_{\mu=0}^{\nu-1} (R(\mu - m) f_m, f_\mu) = (T_\nu f, f)\,. \end{aligned}$$

Hence (2.2) holds. If $S(\omega) \geq 0$ for all ω in $[0, 2\pi]$, then $(S(\omega)p(\omega), p(\omega))$ is also positive for all ω. This readily implies that the left hand side of (2.2) is positive for all p. Thus $(T_\nu f, f) \geq 0$ for all f and integers $\nu \geq 1$. In other words, the Toeplitz matrices T_ν are positive for all $\nu \geq 1$.

Now assume that T_ν is positive for all integers $\nu \geq 1$. By the Weierstrass approximation theorem any continuous function $g(\omega)$ with values in $\mathcal{Y}$ approximated by a polynomial in $e^{\imath\omega}$, that is, given any $\varepsilon > 0$ there exists a polynomial p of some degree $\nu - 1$ such that $\|g(\omega) - e^{\imath\omega m}p\| < \varepsilon$ for all ω in $[0, 2\pi]$. So by passing limits in (2.2) we see that

$$\int_0^{2\pi} (S(\omega)g(\omega), g(\omega))\, d\omega \geq 0$$

for all continuous function g with values in $\mathcal{Y}$. By employing some standard arguments from measure theory, it follows that $S(\omega) \geq 0$ for all ω in $[0, 2\pi]$. This completes the proof.

Finally, it is noted that if T_ν in (2.1) is positive, then T_ν is self-adjoint. In this case, $R(k) = R(-k)^*$ for all integers $k = 0, 1, 2, \cdots, \nu - 1$. In other words, $R(k) \neq R(-k)^*$ for some integer k, then T_ν is not positive for all $\nu \geq k + 1$. So without loss of generality when implementing Theorem 5.2.1, we can assume that $R(k) = R(-k)^*$ for all integers $k \geq 0$.

REMARK 5.2.2 *Let $S(\omega)$ be the Fourier transform for a sequence of operators $\{R(k)\}_{-\infty}^{\infty}$ on $\mathcal{Y}$. As before, let T_ν be the Toeplitz matrix in (2.1) and assume that $\gamma > 0$. Then the proof of Theorem 5.2.1 shows that $S(\omega) \geq \gamma I$ for almost all ω in $[0, 2\pi]$ if and only if $T_\nu \geq \gamma I$ for all integers $\nu \geq 1$. In particular, if $S(\omega) \geq \gamma I$, then T_ν is invertible for all $\nu \geq 1$.*

5.2.1 Exercise

Let $\{R(k)\}_{-\infty}^{\infty}$ be a sequence of scalars such that $R(k) = \overline{R(-k)}$ for all integers $k \geq 0$. Moreover, assume that $R(k)$ converges to zero as k tends to infinity. Then one can use the fast Fourier transform to determine if the Toeplitz matrix T_ν in (2.1) is positive for all integers $\nu \geq 1$. To see how this works in Matlab, let s be a 2^m fast Fourier transform given by

$$\begin{aligned} r &= \begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(-3) & R(-2) & R(-1) \end{bmatrix}; \\ s &= \text{fft}(r); \quad \text{plot}(\text{linspace}(0, 2\pi, 2^m), s); \end{aligned}$$

Here s is the fast Fourier transform approximation for the Fourier transform S of $\{R(k)\}_{-\infty}^{\infty}$. Then the Toeplitz matrix T_ν in (2.1) is positive for all integers $\nu \geq 1$ if and only if the plot of s in Matlab is positive for all ω in $[0, 2\pi)$.

Problem 1. Let $R(k) = .95^{|k|}$ for all integers k. Then use Matlab to determine if the Toeplitz matrix T_ν in (2.1) is positive for all integers $\nu \geq 1$.

Problem 2. Let A be a stable operator on $\mathcal{X}$ satisfying $\|A\| \leq 1$. Let C be an operator from $\mathcal{X}$ into $\mathcal{Y}$ and $R(k)$ the operators defined by

$$\begin{aligned} R(k) &= CA^kC^* \qquad \text{if } k \geq 0 \\ &= CA^{*|k|}C^* \qquad \text{if } k \leq 0\,. \end{aligned}$$

Then show that the Toeplitz matrix T_ν in (2.1) is positive for all integers $\nu \geq 1$. Hint, consider Theorem 4.4.1 in Chapter 4 with the fact that $AA^* \leq I$.

5.3 Transfer functions for Linear systems

This section is devoted to the transfer function for a linear time invariant system. We say that $\{G(k)\}_{-\infty}^{\infty}$ is in $\ell^2(\mathcal{U},\mathcal{Y})$ if $\{G(k)\}_{-\infty}^{\infty}$ is a sequence of operators mapping $\mathcal{U}$ into $\mathcal{Y}$ and

$$\sum_{k=-\infty}^{\infty} \operatorname{trace} G(k)^* G(k) < \infty \,. \tag{3.1}$$

Recall that the trace of an operator is the sum of its diagonal entries. Assume that $\{G(k)\}_{-\infty}^{\infty}$ is in $\ell^2(\mathcal{U},\mathcal{Y})$. Then the Fourier transform of $\{G(k)\}_{-\infty}^{\infty}$ is the function defined by

$$\mathbf{G}(e^{\imath\omega}) = \mathcal{F}\{G(k)\}_{-\infty}^{\infty} = \sum_{k=-\infty}^{\infty} e^{-\imath\omega k} G(k) \,. \tag{3.2}$$

A boldface $\mathbf{G}$ is used to represent the Fourier transform of $\{G(k)\}_{-\infty}^{\infty}$. Because $\{G(k)\}_{-\infty}^{\infty}$ is in $\ell^2(\mathcal{U},\mathcal{Y})$, the Fourier transform $\mathbf{G}$ is a well defined operator. In fact, $\|\mathbf{G}\|$ is in L^2. Notice that we have chosen to write $e^{\imath\omega}$ in the argument of $\mathbf{G}(e^{\imath\omega})$. This is turns out to be a natural way of viewing the Fourier transforms for the transfer function for state space systems. Finally, if $\{G(k)\}_{-\infty}^{\infty}$ is scalar valued ($\mathcal{U} = \mathcal{Y} = \mathbb{C}$), then we will use a lower case $\{g(k)\}_{-\infty}^{\infty}$ and a lower case boldface $\mathbf{g}$ for its Fourier transform.

Consider the linear time invariant system given by

$$y(n) = \sum_{k=-\infty}^{\infty} G(n-k)u(k) = \sum_{k=-\infty}^{\infty} G(k)u(n-k) \,. \tag{3.3}$$

Here the input $u(n)$ is a sequence with values in $\mathcal{U}$, the output $y(n)$ is a sequence with values in $\mathcal{Y}$ and $\{G(k)\}_{-\infty}^{\infty}$ is in $\ell^2(\mathcal{U},\mathcal{Y})$. The function $\mathbf{G}(e^{\imath\omega}) = \mathcal{F}\{G(k)\}_{-\infty}^{\infty}$ is called the *transfer function* from u to y. Now assume that $\{u(k)\}_{-\infty}^{\infty}$ is in $\ell^2(\mathcal{U})$, that is, assume that $\sum_{-\infty}^{\infty} \|u(k)\|^2$ is finite. Let

$$\mathbf{u}(e^{\imath\omega}) = \mathcal{F}\{u(k)\}_{-\infty}^{\infty} = \sum_{k=-\infty}^{\infty} e^{-\imath\omega k} u(k) \quad \text{and} \quad \mathbf{y}(e^{\imath\omega}) = \mathcal{F}\{y(k)\}_{-\infty}^{\infty} = \sum_{k=-\infty}^{\infty} e^{-\imath\omega k} y(k) \,.$$

Then $\mathbf{y}(e^{\imath\omega}) = \mathbf{G}(e^{\imath\omega})\mathbf{u}(e^{\imath\omega})$. In other words, convolution in the time domain corresponds to multiplication in the frequency domain. To prove this fact let $\lambda = e^{-\imath\omega}$. Then using the second equality in (3.3), we obtain

$$\begin{aligned}
\mathbf{y}(e^{\imath\omega}) &= \sum_{n=-\infty}^{\infty} \lambda^n y(n) = \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \lambda^k G(k) \lambda^{n-k} u(n-k) \\
&= \sum_{\nu=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \lambda^k G(k) \lambda^{\nu} u(\nu) = \left(\sum_{k=-\infty}^{\infty} \lambda^k G(k) \right) \left(\sum_{\nu=-\infty}^{\infty} \lambda^{\nu} u(\nu) \right) \\
&= \mathbf{G}(e^{\imath\omega})\mathbf{u}(e^{\imath\omega}) \,.
\end{aligned}$$

Therefore the Fourier transform $\mathbf{y} = \mathbf{G}\mathbf{u}$.

Now assume that $u(n)$ is an impulse, that is, $u(n) = 0$ if $n \neq 0$ and $u(0) = u_0$ where u_0 is a vector in $\mathcal{U}$. By consulting the input output map in (3.3), we see that $y(n) = G(n)u_0$ for all integers n. Motivated by this we call $\{G(k)\}_{-\infty}^{\infty}$ the *impulse response* for (3.3). Finally, $\{G(k)\}_{-\infty}^{\infty}$ is also referred to as the *impulse response for the transfer function* $\mathbf{G}$.

Recall that a system is causal if the present only depends upon the present and past. The system in (3.3) is *causal* if

$$y(n) = \sum_{k=-\infty}^{n} G(n-k)u(k) = \sum_{k=0}^{\infty} G(k)u(n-k)\,. \tag{3.4}$$

In other words, the system in (3.3) is causal if and only if $G(k) = 0$ for all integers $k < 0$. Motivated by this we say that *the sequence* $\{G(k)\}_{-\infty}^{\infty}$ *is a causal* if $G(k) = 0$ for all integer $k < 0$. Notice that the system in (3.3) is causal if and only if $\mathbf{G}$ admits a Fourier series expansion of the form

$$\mathbf{G} = \sum_{k=0}^{\infty} e^{-\imath\omega k} G(k)\,. \tag{3.5}$$

Motivated by this we say that $\mathbf{G}$ is a *causal transfer function* if $\mathbf{G}$ admits a Fourier series representation of the form (3.5). Throughout $\mathbf{G}(z)$ is the function obtained by replacing $e^{\imath\omega}$ by the complex variable z in $\mathbf{G}(e^{\imath\omega})$. Recall that a transfer function $\mathbf{G}$ is rational if $\mathbf{G}(z) = N(z)/d(z)$ where $N(z)$ is an operator valued polynomial and $d(z)$ is a scalar valued polynomial. Finally, we say that $\mathbf{G}(z)$ is a *proper rational* transfer function if $\mathbf{G}(z) = N(z)/d(z)$ where the degree of the polynomial N is less than or equal to the degree of the polynomial d, that is, $\deg N \leq \deg d$. This sets the stage for the following result.

LEMMA 5.3.1 *Let* $\mathbf{G}$ *be a rational transfer function and assume that its impulse response* $\{G(k)\}_{-\infty}^{\infty}$ *is in* $\ell^2(\mathcal{U}, \mathcal{Y})$. *Then* $\mathbf{G}$ *is causal if and only if* $\mathbf{G}$ *is a proper rational function and all the poles of* $\mathbf{G}(z)$ *are contained in the open unit disc* $\{z : |z| < 1\}$.

Proof. Assume that $\mathbf{G}$ is a proper rational function and all the poles of $\mathbf{G}(z)$ are contained in the open unit disc $\{z : |z| < 1\}$. Hence there exists a radius $0 < \gamma < 1$ such that all the poles of $\mathbf{G}(z)$ are contained in the open unit disc $\{z : |z| < \gamma\}$. This along with the fact that $\mathbf{G}(z)$ is a proper rational function, implies that $\mathbf{G}(z)$ admits a power series expansion of the form

$$\mathbf{G}(z) = \sum_{k=0}^{\infty} z^{-k} G(k) \qquad (|z| \geq \gamma)\,. \tag{3.6}$$

Moreover, $\|G(k)\| \leq m\gamma^k$ for all integers $k \geq 0$ and some bound $m > 0$. Notice that the series for $\mathbf{G}(z)$ in (3.6) converges uniformly for $e^{\imath\omega} = z$. In other words, $\mathbf{G}(z)$ admits a Fourier series representation of the form $\mathbf{G}(e^{\imath\omega}) = \sum_0^{\infty} e^{-\imath\omega k} G(k)$. Therefore $\mathbf{G}(z)$ is causal.

Now assume that $\mathbf{G}(z)$ is causal and $|z| > 1$. We claim that $\mathbf{G}(z)$ admits a power series expansion of the form

$$\mathbf{G}(z) = \sum_{k=0}^{\infty} z^{-k} G(k) \qquad (|z| > 1)\,. \tag{3.7}$$

By employing the Cauchy-Schwartz inequality and Lemma 4.4.2 in Chapter 4, we have

$$\begin{aligned}\|\mathbf{G}(z)\|^2 &= \|\sum_{k=0}^{\infty} z^{-k}G(k)\|^2 \le \left(\sum_{k=0}^{\infty}|z^{-k}|\|G(k)\|\right)^2 \le \left(\sum_{k=0}^{\infty}|z^{-k}|^2\right)\left(\sum_{k=0}^{\infty}\|G(k)\|^2\right)\\ &\le \frac{1}{1-|z|^{-2}}\sum_{k=0}^{\infty}\operatorname{trace} G(k)^*G(k) < \infty\,.\end{aligned}$$

Hence $\mathbf{G}(z)$ is finite for all $|z| > 1$, and thus, $\mathbf{G}(z)$ admits a power series expansion of the form (3.7). In particular, $\mathbf{G}(z)$ is analytic in $\{z : |z| > 1\}$, or equivalently, all the poles of $\mathbf{G}(z)$ are contained in the closed unit circle $\{z : |z| \le 1\}$. Since $\{G(k)\}_{-\infty}^{\infty}$ is in $\ell^2(\mathcal{U},\mathcal{Y})$, we see that $\|G(k)\|$ converges to zero as k tends to infinity. Hence the transfer function $\mathbf{G}(z)$ cannot have any poles on the unit circle. Therefore all the poles of $\mathbf{G}(z)$ are contained in the open unit disc. Because $\mathbf{G}(z)$ is rational and $\mathbf{G}(z)$ admits a power series expansion of the form (3.7), it follows by long division that $\mathbf{G}(z)$ is a proper rational function. This completes the proof.

5.3.1 State space systems

Now consider the stable state space system defined by

$$\begin{aligned}x(n+1) &= Ax(n) + Bu(n) \qquad (x(-\infty) = 0)\\ y(n) &= Cx(n) + Du(n)\,.\end{aligned} \tag{3.8}$$

Here A is a stable operator on $\mathcal{X}$ and B is an operator from $\mathcal{U}$ into $\mathcal{X}$ while C is an operator mapping $\mathcal{X}$ into $\mathcal{Y}$ and D is an operator from $\mathcal{U}$ into $\mathcal{Y}$. Notice that the initial condition starts at minus infinity, that is, $x(-\infty) = 0$. Clearly, this system is causal. The Transfer function $\mathbf{G}$ from u to y for this state space system is given by

$$\mathbf{G}(z) = C(zI - A)^{-1}B + D \qquad (|z| = 1)\,. \tag{3.9}$$

To verify this recall that the solution to (3.8) is given by

$$y(n) = Du(n) + \sum_{k=-\infty}^{n-1} CA^{n-k-1}Bu(k)\,. \tag{3.10}$$

Let $G(n)$ be the causal sequence of operators defined by

$$\begin{aligned}G(n) &= CA^{n-1}B \qquad \text{if } n \ge 1\\ G(0) &= D\\ G(n) &= 0 \qquad\qquad \text{if } n < -1\,.\end{aligned} \tag{3.11}$$

Equation (3.10) shows that the state space system in (3.8) admits a linear time invariant representation of the form (3.3) where $\{G(n)\}_{-\infty}^{\infty}$ is defined by (3.11). In other words,

$\{G(n)\}_{-\infty}^{\infty}$ in (3.11) is the impulse response for this state space system. Because A is stable, $\{G(k)\}_{-\infty}^{\infty}$ is in $\ell^2(\mathcal{U},\mathcal{Y})$. So according to (3.11) the transfer function is given by

$$\mathbf{G}(e^{\imath\omega}) = \sum_{n=0}^{\infty} e^{-\imath\omega n} G(n) = D + C\sum_{n=1}^{\infty} e^{-\imath\omega n} A^{n-1}B = D + C(e^{\imath\omega}I - A)^{-1}B\,.$$

The last equality follows from $\sum_1^\infty e^{-\imath\omega n}A^{n-1} = e^{-\imath\omega}(I - e^{-\imath\omega}A)^{-1}$; see Lemma 4.4.2 in Chapter 4. Therefore the transfer function for the state space system $\{A, B, C, D\}$ in (3.8) is given by (3.9).

5.3.2 Exercise

One can use the fast Fourier transform to determine if a scalar valued rational transfer function is causal. To see this assume that the transfer function $\mathbf{g}$ is given by

$$\mathbf{g}(z) = \frac{\sum_{k=0}^{\nu} p_k z^k}{\sum_{k=0}^{\nu} q_k z^k}\,. \tag{3.12}$$

Now assume that $|z| = 1$. By multiplying the numerator and denominator of $\mathbf{g}$ by $\overline{z}^{\nu}$, we see that $\mathbf{g}$ can also be written as

$$\mathbf{g}(z) = \frac{p_\nu + p_{\nu-1}\overline{z} + \cdots + p_1\overline{z}^{\nu-1} + p_0\overline{z}^{\nu}}{q_\nu + q_{\nu-1}\overline{z} + \cdots + q_1\overline{z}^{\nu-1} + q_0\overline{z}^{\nu}} \qquad (|z| = 1)\,. \tag{3.13}$$

Now consider the following Matlab commands:

$$\begin{aligned} p &= \text{fft}(\left[\begin{array}{ccccc} p_\nu & p_{\nu-1} & \cdots & p_1 & p_0 \end{array}\right], 2^m) \\ q &= \text{fft}(\left[\begin{array}{ccccc} q_\nu & q_{\nu-1} & \cdots & q_1 & q_0 \end{array}\right], 2^m); \\ G &= p./q; \\ g &= \text{ifft}(G); \end{aligned}$$

(In many application, $m = 11, 12, 13$ or 14.) The form of $\mathbf{g}$ in (3.13) was used to derive the fast Fourier transform expressions for p and q in Matlab. Then G is the fast Fourier transform approximation for $\mathbf{g}$ and g is the fast Fourier transform approximation for the corresponding impulse response $\{g(k)\}_{-\infty}^{\infty}$, that is,

$$g = \left[\begin{array}{cccccc} g(0) & g(1) & g(2) & \cdots & g(-2) & g(-1) \end{array}\right].$$

In practice, $\{p_k\}_0^\nu$ and $\{q_k\}_0^\nu$ are real numbers. In this case, $\{g(k)\}_{-\infty}^{\infty}$ are real numbers. So one can set $g = \text{real}(\text{ifft}(G))$ to eliminate some small imaginary numerical errors computed by the fast Fourier transform computation. Notice that if the computation $g = p./q$ yields a division by zero error warning on the computer, then the polynomial $\sum_{k=0}^{\nu} q_k z^k$ has a zero on the unit circle and $\{g(k)\}_{-\infty}^{\infty}$ is not in ℓ^2, that is, $\sum_{-\infty}^{\infty} |g(k)|^2 = \infty$. Finally, is noted that the fast Fourier transform can be used to determine if a polynomial $q(z) = \sum_{k=0}^{\nu} q_k z^k$ has all its zeros in the open unit disc $\{z : |z| < 1\}$. To accomplish this notice that the transfer function $\mathbf{g}(z) = 1/q(z)$ is causal and $\{g(k)\}_{-\infty}^{\infty} \in \ell^2$ if and only if all the zeros of q

are contained in the open unit disc.

Problem 1. Use the fast Fourier transform to determine the impulse response $\{g(k)\}_{-3}^{3}$ for the following transfer functions

$$\begin{aligned}
\mathbf{g} &= \frac{z^4+z^2+1}{(1-2z)(1-3z)} \\
\mathbf{g} &= \frac{1+2z}{z^4-2z^3+1.4z^2-0.4z+0.1} \\
\mathbf{g} &= \frac{1+2z+z^3}{z^4+2z^3+3z^2+4z+5} \\
\mathbf{g} &= \frac{1+2z+z^3}{1+z+2z^2+3z^3+4z^4+\cdots+98z^{98}+99z^{99}+100z^{100}}\,.
\end{aligned}$$

Determine if any of these transfer functions are causal. Finally, it is noted that the dimpulse command in Matlab will not work unless **g** is causal.

Problem 2. Use the fast Fourier transform to determine if the roots of the following polynomials are all contained in the open unit disc $\{z : |z| < 1\}$

$$\begin{aligned}
q(z) &= z^5-2.5z^4+2.36z^3-1.04z^2+0.21z-0.02 \\
q(z) &= z^4-10z^3+35z^2-50z+24 \\
q(z) &= z^5+2z^4+3z^3-4z^2-3z-2\,.
\end{aligned}$$

5.4 The spectral density

This section introduces and studies the spectral density for a wide sense stationary random process. To this end, let $y(n)$ be a wide sense stationary random process with values in $\mathcal{Y}$. Then the *spectral density* $S_y(\omega)$ for $y(n)$ is the Fourier transform of its autocorrelation function, that is,

$$S_y(\omega) = \mathcal{F}\{R_y(k)\}_{-\infty}^{\infty} = \sum_{n=-\infty}^{\infty} e^{-\imath\omega k} R_y(k)\,. \tag{4.1}$$

The spectral density $S_y(\omega)$ is a positive operator for all ω in $[0, 2\pi]$. To see this recall that the Toeplitz matrix T_ν in (1.1) generated by generated by $\{R_y(k)\}_0^{\nu-1}$ is positive for all integers $\nu \geq 1$; see Section 4.1 in Chapter 4. According to Theorem 5.2.1, the spectral density $S_y(\omega)$ is positive for all ω in $[0, 2\pi]$.

If $w(n)$ is a white noise with values in $\mathcal{W}$, then its spectral density $S_w(\omega) = I$. Recall that the autocorrelation function for a white noise process is $R_w(n) = \delta_{n,0}$ where $\delta_{j,k}$ is the Kronecker delta. Hence $S_w(\omega) = \mathcal{F}\delta_{n,0} = I$.

THEOREM 5.4.1 *Let $y(n)$ be the random process with values in $\mathcal{Y}$ determined by*

$$y(n) = \sum_{k=-\infty}^{\infty} G(n-k)u(k) = \sum_{k=-\infty}^{\infty} G(k)u(n-k) \tag{4.2}$$

where $u(n)$ is a zero mean wide sense stationary process with values in $\mathcal{U}$, and the impulse response $\{G(k)\}_{-\infty}^{\infty}$ is in $\ell^2(\mathcal{U},\mathcal{Y})$. Let $\mathbf{G} = \mathcal{F}\{G(k)\}_{-\infty}^{\infty}$ be the transfer function for (4.2). Then $y(n)$ is a mean zero wide sense stationary random process whose spectral density is given by

$$S_y(\omega) = \mathbf{G}(e^{\imath\omega})S_u(\omega)\mathbf{G}(e^{\imath\omega})^*\,. \tag{4.3}$$

In particular, if $u(n)$ is white noise, then $S_y(\omega) = \mathbf{G}(e^{\imath\omega})\mathbf{G}(e^{\imath\omega})^*$.

PROOF. Because $y(n)$ is a linear combination of $\{u(k)\}_{-\infty}^{\infty}$ and the mean of u is zero, it follows that mean of $y(n)$ is also zero for all integers n. Using the second formula for $y(n)$ in (4.2) yields

$$\begin{aligned} Ey(n+\nu)y(\nu)^* &= E\left(\sum_{m=-\infty}^{\infty} G(m)u(n+\nu-m)\right)\left(\sum_{j=-\infty}^{\infty} G(j)u(\nu-j)\right)^* \\ &= \sum_{m=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} G(m)Eu(n+\nu-m)u(\nu-j)^*G(j)^* \\ &= \sum_{m=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} G(m)R_u(n+j-m)G(j)^*\,. \end{aligned}$$

Hence $Ey(n+\nu)y(\nu)^*$ is just a function of n. So $y(n)$ is wide sense stationary and

$$R_y(n) = \sum_{m=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} G(m)R_u(n+j-m)G(j)^*\,. \tag{4.4}$$

The spectral density S_y is computed by taking the Fourier transform of $\{R_y(n)\}_{-\infty}^{\infty}$, that is,

$$\begin{aligned} S_y(\omega) &= \sum_{n=-\infty}^{\infty} e^{-\imath\omega n}R_y(n) = \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} e^{-\imath\omega n}G(m)R_u(n+j-m)G(j)^* \\ &= \sum_{\nu=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} e^{-\imath\omega(\nu+m-j)}G(m)R_u(\nu)G(j)^* \\ &= \left(\sum_{m=-\infty}^{\infty} e^{-\imath\omega m}G(m)\right)\left(\sum_{\nu=-\infty}^{\infty} e^{-\imath\omega\nu}R_u(\nu)\right)\left(\sum_{j=-\infty}^{\infty} e^{-\imath\omega j}G(j)\right)^* \\ &= \mathbf{G}(e^{\imath\omega})S_u(\omega)\mathbf{G}(e^{\imath\omega})^*\,. \end{aligned}$$

This yields (4.3). Finally, if $u(n)$ is a white noise process, then $S_u(\omega) = I$. In this case, $S_y(\omega) = \mathbf{G}(e^{\imath\omega})\mathbf{G}(e^{\imath\omega})^*$. This completes the proof.

The following result is a generalization of Theorem 4.4.1.

THEOREM 5.4.2 *Consider state space system given by*

$$\begin{aligned} x(n+1) &= Ax(n) + Bu(n) \qquad (x(-\infty) = 0) \\ y(n) &= Cx(n) + Du(n) \end{aligned} \tag{4.5}$$

where $u(n)$ is white noise process. Moreover, assume that A is stable and Q is the controllability Gramian for the pair $\{A, B\}$, that is, $Q = AQA^* + BB^*$. Then following holds.

(i) The output $y(n)$ is a mean zero wide sense stationary random process.

(ii) The spectral density for $y(n)$ is given by $S_y(\omega) = \mathbf{G}(e^{\imath\omega})\mathbf{G}(e^{\imath\omega})^*$ where

$$\mathbf{G}(z) = D + C(zI - A)^{-1}B \qquad (|z| = 1) \tag{4.6}$$

is the transfer function from u to y.

(iii) The autocorrelation for $y(n)$ is given by

$$\begin{aligned} R_y(n) &= CA^{n-1}(BD^* + AQC^*) && \text{if } n \geq 1 \\ &= CQC^* + DD^* && \text{if } n = 0 \\ &= (DB^* + CQA^*)A^{*|n+1|}C^* && \text{if } n \leq -1\,. \end{aligned} \tag{4.7}$$

Proof. By consulting (3.8) to (3.11) we see that

$$y(n) = \sum_{k=-\infty}^{\infty} G(n-k)u(k)$$

where $\{G(n)\}_{-\infty}^{\infty}$ are given by (3.11). Because A is stable, $\{G(n)\}_{-\infty}^{\infty}$ is in $\ell^2(\mathcal{U}, \mathcal{Y})$. Therefore the hypothesis of Theorem 5.4.1 are satisfied. Hence $y(n)$ is a mean zero wide sense stationary process, and part (i) holds. Since $\mathbf{G}(z) = D + C(zI - A)^{-1}B$ is the transfer function from u to y, Theorem 5.4.1 implies that $S_y(\omega) = \mathbf{G}(e^{\imath\omega})\mathbf{G}(e^{\imath\omega})^*$. In other words, part (ii) holds.

To complete the proof it remains to establish the formula for R_y in (4.7). Let $z = e^{\imath\omega}$ and set $\Phi(z) = (I - \overline{z}A)^{-1}$. Using $(zI - A)^{-1} = \overline{z}\,\Phi(z)$, we have

$$\begin{aligned} S_y(\omega) &= \big(D + C(zI - A)^{-1}B\big)\big(D + C(zI - A)^{-1}B\big)^* \\ &= DD^* + \overline{z}\,C\Phi(z)BD^* + zDB^*\Phi(z)^*C^* + C\Phi(z)BB^*\Phi(z)^*C^*\,. \end{aligned} \tag{4.8}$$

By employing the Lyapunov equation $Q = AQA^* + BB^*$, we obtain

$$BB^* = Q - AQA^* = (I - \overline{z}A)Q + Q(I - zA^*) - (I - \overline{z}A)Q(I - zA^*)\,.$$

Multiplying this equation by $C\Phi$ on the left and Φ^*C^* on the right yields

$$C\Phi BB^*\Phi^*C^* = C\Phi QC^* + CQ\Phi^*C^* - CQC^*\,.$$

Substituting this into the last term in (4.8) and using $\Phi(z) = I + \overline{z}\,\Phi(z)A$ implies that

$$\begin{aligned} S_y(\omega) &= DD^* - CQC^* + \overline{z}\,C\Phi BD^* + C\Phi QC^* + zDB^*\Phi^*C^* + CQ\Phi^*C^* \\ &= DD^* + CQC^* + \overline{z}\,C\Phi(BD^* + AQC^*) + z(DB^* + CQA^*)\Phi^*C^*\,. \end{aligned}$$

Since A is stable, $\Phi(z) = (I - \overline{z}\,A)^{-1} = \sum_0^\infty \overline{z}^{\,k}\,A^k$; see Lemma 4.4.2 in Chapter 4. Using this in the previous expression for $S_y(\omega)$ yields

$$S_y = DD^* + CQC^* + \sum_{k=1}^{\infty} e^{-\imath\omega k}CA^{k-1}(BD^* + AQC^*) + \sum_{k=1}^{\infty} e^{\imath\omega k}(DB^* + CQA^*)A^{*k-1}C^*\,.$$

Recall that $S_y(\omega) = \sum_{-\infty}^{\infty} e^{-\imath\omega k}R_y(k)$. By matching the Fourier coefficients of $e^{-\imath\omega k}$ we arrive at the expression for $R_y(n)$ in (4.7). This completes the proof.

5.4.1 Exercise

One can use the fast Fourier transform to compute the spectral density and autocorrelation function corresponding to a scalar valued rational transfer function. For example, consider the wide sense stationary random process $y(n)$ determined by

$$y(n) = \sum_{k=-\infty}^{\infty} g(n-k)u(k) \tag{4.9}$$

where $u(n)$ is a scalar valued white noise process. Moreover, assume that $\{g(k)\}_{-\infty}^{\infty}$ is the impulse response for a rational transfer function

$$\mathbf{g}(z) = \frac{\sum_{k=0}^{\nu} p_k z^k}{\sum_{k=0}^{\nu} q_k z^k}\,. \tag{4.10}$$

Finally, we assume that $\mathbf{g}(z)$ has no poles on the unit circle. According to Theorem 5.4.1, the spectral density for y is given by $S_y(\omega) = |g(e^{\imath\omega})|^2$. To compute this spectral density, consider the following Matlab commands:

$$\begin{array}{rcl} p & = & \mathrm{fft}(\left[\begin{array}{ccccc} p_\nu & p_{\nu-1} & \cdots & p_1 & p_0 \end{array}\right], 2^m) \\ q & = & \mathrm{fft}(\left[\begin{array}{ccccc} q_\nu & q_{\nu-1} & \cdots & q_1 & q_0 \end{array}\right], 2^m); \\ g & = & p./q; \\ s & = & \mathrm{abs}(g).^2 \\ & & \mathrm{plot}(\mathrm{linspace}(0, 2\pi, 2^m), s); \\ r & = & \mathrm{ifft}(s); \end{array}$$

Then s is the fast Fourier transform approximation for S_y, and r is the fast Fourier transform approximation for the autocorrelation function $R_y(n)$, that is,

$$r = \left[\begin{array}{cccccccc} R_y(0) & R_y(1) & R_y(2) & \cdots & R_y(-3) & R_y(-2) & R_y(-1) \end{array}\right].$$

In practice, $\{p_k\}_0^\nu$ and $\{q_k\}_0^\nu$ are real numbers. In this case, $\{R_y(n)\}_{-\infty}^{\infty}$ are real numbers. In particular, $R_y(n) = R_y(-n)$ for all integers n. So one can set $r = \mathrm{real}(\mathrm{ifft}(s))$ to eliminate some small imaginary numerical errors computed by the fast Fourier transform computation. Finally, notice that if the computation $g = p./q$ yields a division by zero error warning on the computer, then the polynomial $\sum_{k=0}^{\nu} q_k z^k$ has a zero on the unit circle and $\{g(k)\}_{-\infty}^{\infty}$ is not in ℓ^2, that is, $\sum_{-\infty}^{\infty} |g(k)|^2 = \infty$.

Problem 1. Let $y(n)$ be the wide sense stationary process determined by (4.9) where $u(n)$ is a white noise process. Then use Matlab to plot the spectral density S_y and compute $\{R_y(k)\}_0^5$ for the transfer function

$$\mathbf{g} = \frac{10z^2 + 4}{z^4 + 2z^3 + 3z^2 + 4z + 5}\,.$$

Finally, it is noted that because all the coefficients of the polynomials are real, we have $R_y(n) = R_y(-n)$ for all integers n.

Problem 2. Consider the state space system given by

$$\begin{bmatrix} x_1(n+1) \\ x_2(n+1) \\ x_3(n+1) \\ x_4(n+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -0.1 & 0.4 & -1.4 & 2 \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \\ x_4(n) \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} u(n)$$
$$y(n) = \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix} x(n) + 2u(n) .$$

The initial condition $x(-\infty) = 0$. Use Theorem 5.4.2 with Matlab to compute the autocorrelation function $R_y(n)$ for the process $y(n)$. Now following the ideas in Section 4.4.4 in Chapter 4, use the fast Fourier transform to compute the autocorrelation function $R_y(n)$ for $y(n)$. Compare $\{R_y(n)\}_0^5$ by both of these methods. Finally, use the Fast Fourier transform to plot the spectral density $S_y(\omega)$ for $y(n)$.

5.5 Jointly wide sense stationary processes.

Let $x(n)$ be a random process with values in $\mathcal{X}$ and $y(n)$ be a random process with values in $\mathcal{Y}$. Then we say that $x(n)$ and $y(n)$ are *jointly wide sense stationary* if the following three conditions hold:

(i) the process $x(n)$ is wide sense stationary;

(ii) the process $y(n)$ is wide sense stationary;

(iii) $Ex(n)y(m)^* = R_{xy}(n-m)$ is a function of $n-m$ for all integers n and m.

In this case, $R_{xy}(n)$ is called the *joint autocorrelation function* for x and y. Notice that $Ex(n)y(m)^* = R_{xy}(n-m)$ for all n and m if and only if $Ex(n+k)y(k)^* = R_{xy}(n)$ is just a function of n for all integers n and k. The processes x and y are jointly wide sense stationary if and only if the processes y and x are jointly wide sense stationary. In this case, $R_{xy}(n) = R_{yx}(-n)^*$. To see this simply observe that

$$R_{xy}(n) = Ex(n+k)y(k)^* = (Ey(k)x(n+k)^*)^* = R_{yx}(k-n-k)^* = R_{yx}(-n)^* .$$

Therefore we have $R_{xy}(n) = R_{yx}(-n)^*$. Finally, it is noted that $R_{xy}(n)$ is an operator from $\mathcal{Y}$ into $\mathcal{X}$ and $R_{yx}(n)$ is an operator from $\mathcal{X}$ into $\mathcal{Y}$ for all integers n.

Now assume that $x(n)$ and $y(n)$ are jointly wide sense stationary. Then the *joint spectral density* S_{xy} for x and y is the Fourier transform of their joint autocorrelation function R_{xy}, that is,

$$S_{xy}(\omega) = \mathcal{F}\{R_{xy}(k)\}_{-\infty}^{\infty} = \sum_{k=-\infty}^{\infty} R_{xy}(k)e^{-\imath\omega k} .$$

Observe that $S_{xy}(\omega)$ is an operator from $\mathcal{Y}$ into $\mathcal{X}$ for almost all ω. Finally, it is noted that $S_{xy}(\omega) = S_{yx}(\omega)^*$ for almost all ω in $[0, 2\pi]$. To verify this notice that $R_{xy}(n) = R_{yx}(-n)^*$ yields

$$S_{xy}(\omega)^* = \sum_{k=-\infty}^{\infty} R_{xy}(k)^* e^{\imath\omega k} = \sum_{k=-\infty}^{\infty} R_{yx}(-k)e^{\imath\omega k} = \sum_{k=-\infty}^{\infty} R_{yx}(k)e^{-\imath\omega k} = S_{yx}(\omega) .$$

In other words, $S_{xy}^* = S_{yx}$.

Let $q(n)$ be the random process with values in $\mathcal{X} \oplus \mathcal{Y}$ determined by

$$q(n) = \begin{bmatrix} x(n) \\ y(n) \end{bmatrix}. \tag{5.1}$$

A simple calculation shows that for all integers n and k, we have

$$Eq(n+k)q(k)^* = \begin{bmatrix} Ex(n+k)x(k)^* & Ex(n+k)y(k)^* \\ Ey(n+k)x(k)^* & Ey(n+k)y(k)^* \end{bmatrix}.$$

Notice that $q(n)$ is wide sense stationary if and only if $x(n)$ and $y(n)$ are jointly wide sense stationary. In this case, the autocorrelation function R_q for the process $q(n)$ is the operator matrix given by

$$R_q(n) = \begin{bmatrix} R_x(n) & R_{xy}(n) \\ R_{yx}(n) & R_y(n) \end{bmatrix}. \tag{5.2}$$

Moreover, the spectral density S_q for the process $q(n)$ is determined by

$$S_q = \begin{bmatrix} S_x & S_{xy} \\ S_{yx} & S_y \end{bmatrix}. \tag{5.3}$$

Because the spectral density is a positive operator, $S_q(\omega)$ is almost everywhere a positive operator on $\mathcal{X} \oplus \mathcal{Y}$. In particular, $S_q(\omega)$ is almost everywhere a self-adjoint operator. This also shows that $S_{xy}(\omega)^* = S_{yx}(\omega)$ for almost all ω. The following result is a generalization of Theorem 5.4.1.

THEOREM 5.5.1 *Let $x(n)$ be the random process with values in $\mathcal{X}$ and $y(n)$ be the random process with values in $\mathcal{Y}$ determined by*

$$x(n) = \sum_{k=-\infty}^{\infty} H(n-k)v(k) \qquad \text{and} \qquad y(n) = \sum_{k=-\infty}^{\infty} G(n-k)u(k)\,. \tag{5.4}$$

Here $u(n)$ and $v(n)$ are jointly wide sense stationary zero mean processes with values in $\mathcal{U}$ and $\mathcal{V}$, respectively. Moreover, assume that the impulse response $\{G(k)\}_{-\infty}^{\infty}$ is in $\ell^2(\mathcal{U},\mathcal{Y})$ and $\{H(k)\}_{-\infty}^{\infty}$ is in $\ell^2(\mathcal{V},\mathcal{X})$. Then $x(n)$ and $y(n)$ are mean zero jointly wide sense stationary random processes whose joint spectral density is given by

$$S_{xy}(\omega) = \mathbf{H}(e^{\imath\omega})S_{vu}(\omega)\mathbf{G}(e^{\imath\omega})^* \tag{5.5}$$

where $\mathbf{H} = \mathcal{F}\{H(k)\}_{-\infty}^{\infty}$ and $\mathbf{G} = \mathcal{F}\{G(k)\}_{-\infty}^{\infty}$. In particular, if $u(n) = v(n)$ is white noise, then $x(n)$ and $y(n)$ are mean zero jointly wide sense stationary and $S_{xy}(\omega) = \mathbf{H}(e^{\imath\omega})\mathbf{G}(e^{\imath\omega})^*$.

PROOF. Let $p(n)$ and $q(n)$ be the random processes determined by

$$p(n) = \begin{bmatrix} v(n) \\ u(n) \end{bmatrix} \quad \text{and} \quad q(n) = \begin{bmatrix} x(n) \\ y(n) \end{bmatrix}.$$

Because $u(n)$ and $v(n)$ are mean zero jointly wide sense stationary, it follows that $p(n)$ is a mean zero wide sense stationary process. Moreover,

$$q(n) = \begin{bmatrix} x(n) \\ y(n) \end{bmatrix} = \sum_{k=-\infty}^{\infty} \begin{bmatrix} H(n-k) & 0 \\ 0 & G(n-k) \end{bmatrix} \begin{bmatrix} v(k) \\ u(k) \end{bmatrix} = \sum_{k=-\infty}^{\infty} J(n-k) \begin{bmatrix} v(k) \\ u(k) \end{bmatrix}.$$

Here $\{J(k)\}_{-\infty}^{\infty}$ is the diagonal impulse response defined by

$$J(k) = \begin{bmatrix} H(k) & 0 \\ 0 & G(k) \end{bmatrix} \quad \text{and its transfer function} \quad \mathbf{J} = \begin{bmatrix} \mathbf{H} & 0 \\ 0 & \mathbf{G} \end{bmatrix}.$$

By consulting Theorem 5.4.1, we see that $q(n)$ is mean zero jointly wide sense stationary processes. Furthermore, its spectral density is given by

$$\begin{bmatrix} S_x & S_{xy} \\ S_{yx} & S_y \end{bmatrix} = S_q = \mathbf{J} S_p \mathbf{J}^* = \mathbf{J} \begin{bmatrix} S_v & S_{vu} \\ S_{uv} & S_u \end{bmatrix} \mathbf{J}^* = \begin{bmatrix} \mathbf{H} S_v \mathbf{H}^* & \mathbf{H} S_{vu} \mathbf{G}^* \\ \mathbf{G} S_{uv} \mathbf{H}^* & \mathbf{G} S_u \mathbf{G}^* \end{bmatrix}.$$

Therefore $S_{xy} = \mathbf{H} S_{vu} \mathbf{G}^*$. This completes the proof.

5.5.1 Exercise

Problem 1. Consider the state space system given by

$$\begin{aligned} x(n+1) &= Ax(n) + Bu(n) \qquad (x(-\infty) = 0) \\ y(n) &= Cx(n) + Du(n) \end{aligned}$$

where $u(n)$ is white noise process. Assume that A is stable and Q is the controllability Gramian for the pair $\{A, B\}$, that is, $Q = AQA^* + BB^*$. Then show that $x(n)$ and $y(n)$ are jointly wide sense stationary processes and their joint spectral density is determined by

$$S_{xy}(\omega) = (zI - A)^{-1} B \left(D + C(zI - A)^{-1} B \right)^* \qquad (\text{where } z = e^{\imath\omega}).$$

Moreover, show that the joint autocorrelation function is given by

$$\begin{aligned} R_{xy}(n) &= A^{n-1}(BD^* + AQC^*) && \text{if } n \geq 1 \\ &= QA^{*|n|}C^* && \text{if } n \leq 0. \end{aligned}$$

Problem 2. Consider the wide sense stationary processes given by

$$x(n) = \sum_{k=-\infty}^{\infty} H(n-k)u(k) \qquad \text{and} \qquad y(n) = \sum_{k=-\infty}^{\infty} G(n-k)u(k)$$

where $u(n)$ is a white noise process. Moreover, assume that the transfer functions $\mathbf{H}$ and $\mathbf{G}$ are given by

$$\begin{aligned} \mathbf{H} &= \frac{z^2 + 2z + 3}{z^4 + 3z^3 - 4z^2 + z + 3} \\ \mathbf{G} &= \frac{z^3 - z + 4}{z^5 + 3z^4 + 4z^3 + 3z^2 + 2z + 4}. \end{aligned}$$

According to Theorem 5.5.1, the processes $x(n)$ and $y(n)$ are jointly wide sense stationary. Use Matlab and the fast Fourier transform to compute $\{R_{xy}(n)\}_{-3}^{3}$.

5.6 Sinusoid processes and linear systems

A classical result in linear systems shows that if the input to a linear time invariant stable system is a sinusoid with frequency ω, then the steady state output is also a sinusoid of the same frequency ω. However, the amplitude and phase of the steady state output is determined by evaluating the corresponding transfer function at ω. In this section, we will present a similar result for sinusoid processes driving linear time invariant systems.

To begin, let $\xi(n)$ be the sinusoid process given by

$$\xi(n) = \sum_{j=1}^{\mu} a_j e^{\imath(\omega_j n + \theta_j)} \,. \tag{6.1}$$

Here the amplitudes $\{a_j\}_1^\mu$ and frequencies $\{\omega_j\}_1^\mu$ are constants while $\{\theta_j\}_1^\mu$ are mutually independent uniform random variables over the interval $[0, 2\pi]$. Recall that $\xi(n)$ is a mean zero wide sense stationary process and its autocorrelation function is given by

$$R_\xi(n) = \sum_{j=1}^{\mu} |\, a_j|^2 e^{\imath\omega_j n} \,; \tag{6.2}$$

see Sections 4.2 and 4.3 in Chapter 4 for further details. Recall that the spectral density for a random process is the Fourier transform of its autocorrelation function. So the spectral density for the sinusoid process ξ in (6.1) is given by

$$S_\xi(\omega) = 2\pi \sum_{j=1}^{\mu} |\, a_j|^2 \delta(\omega - \omega_j) \tag{6.3}$$

where $\delta(\omega)$ is the delta Dirac function. The following is the main result of this section.

THEOREM 5.6.1 *Let $\xi(n)$ be the sinusoid process given in (6.1) where $\{\theta_j\}_1^\mu$ are mutually independent uniform random variables over the interval $[0, 2\pi]$. Let $y(n)$ be the random process determined by the response of the linear time invariant system*

$$y(n) = \sum_{k=-\infty}^{\infty} g(n-k)\xi(k) \tag{6.4}$$

where the corresponding transfer function $\mathbf{g}$ is a scalar valued rational function with no poles on the unit circle. Then $y(n)$ is the sinusoid process given by

$$y(n) = \sum_{j=1}^{\mu} a_j \mathbf{g}(e^{\imath\omega_j}) e^{\imath(\omega_j n + \theta_j)} \,. \tag{6.5}$$

In particular, $y(n)$ is a mean zero wide sense stationary process whose autocorrelation function is given by

$$R_y(n) = \sum_{j=1}^{\mu} |\, a_k \mathbf{g}(e^{\imath\omega_j})|^2 e^{\imath\omega_j n} \quad \textit{and} \quad S_y(\omega) = 2\pi \sum_{j=1}^{\mu} |\, a_j \mathbf{g}(e^{\imath\omega_j})|^2 \delta(\omega - \omega_j) \,. \tag{6.6}$$

PROOF. Using the formula for $\xi(n)$ in (6.1), we obtain

$$\begin{aligned} y(n) &= \sum_{k=-\infty}^{\infty} g(n-k)\xi(k) = \sum_{k=-\infty}^{\infty}\sum_{j=1}^{\mu} g(n-k)a_j e^{\imath(\omega_j k+\theta_j)} \\ &= \sum_{j=1}^{\mu} a_j e^{\imath\theta_j} \sum_{k=-\infty}^{\infty} g(n-k)e^{-\imath\omega_j(n-k)}e^{\imath\omega_j n} \\ &= \sum_{j=1}^{\mu} a_j e^{\imath\omega_j n}e^{\imath\theta_j} \sum_{m=-\infty}^{\infty} g(m)e^{-\imath\omega_j m} \\ &= \sum_{j=1}^{\mu} a_j \mathbf{g}(e^{\imath\omega_j})e^{\imath(\omega_j n+\theta_j)}\,. \end{aligned}$$

This yields the formula for $y(n)$ in equation (6.5). Notice that $y(n)$ is a sinusoid process; see Sections 4.2 and 4.3 in Chapter 4. Therefore $y(n)$ is a mean zero wide sense stationary process whose autocorrelation function is given by (6.6). This completes the proof.

REMARK 5.6.2 *As before, let $\xi(n)$ be the sinusoid process given in (6.1) where $\{\theta_j\}_1^\mu$ are mutually independent uniform random variables over the interval $[0, 2\pi]$. Now consider the state space system given by*

$$x(n+1) = Ax(n) + B\xi(n) \qquad (x(-\infty) = 0) \quad \textit{and} \quad y(n) = Cx(n) + D\xi(n) \tag{6.7}$$

where A is a stable operator on $\mathcal{X}$ and B maps $\mathbb{C}$ into $\mathcal{X}$ while C maps $\mathcal{X}$ into $\mathbb{C}$ and D is a scalar. Then the output $y(n)$ is the process determined by the causal system

$$y(n) = \sum_{k=-\infty}^{n-1} CA^{n-k-1}B\xi(k) + D\xi(n)\,. \tag{6.8}$$

Theorem 5.6.1 shows that $y(n)$ is the sinusoid process given by (6.5) where the transfer function $\mathbf{g}(z) = C(zI - A)^{-1}B + D$. Moreover, $y(n)$ is a mean zero wide sense stationary process whose autocorrelation function R_y and spectral density S_y are given by (6.6).

5.6.1 The sinusoid process $\xi(n) = CU^n x_0$ and linear systems

Let U be a unitary operator on $\mathcal{X}$, and $\{g(k)\}_{-\infty}^{\infty}$ the impulse response corresponding to a scalar valued rational transfer function $\mathbf{g}$ with no poles on the unit circle. Then $\mathbf{g}(U)$ is the operator on $\mathcal{X}$ defined by

$$\mathbf{g}(U) = \sum_{k=-\infty}^{\infty} g(k)U^{-k}\,. \tag{6.9}$$

To complete this section let us present the following version of Theorem 5.6.1.

THEOREM 5.6.3 *Let $\xi(n)$ be the sinusoid process given by $\xi(n) = CU^n x_0$ where U is a unitary operator on $\mathcal{X}$ while C maps $\mathcal{X}$ into $\mathbb{C}$ and x_0 is a mean zero random vector in $\mathcal{X}$*

satisfying $Ex_0x_0^* = I$. Let $y(n)$ be the random process determined by the response of the linear time invariant system

$$y(n) = \sum_{k=-\infty}^{\infty} g(n-k)\xi(k) \tag{6.10}$$

where the corresponding transfer function $\mathbf{g}$ is a scalar valued rational function with no poles on the unit circle. Then $y(n)$ is the sinusoid process given by

$$y(n) = C\mathbf{g}(U)U^n x_0\,. \tag{6.11}$$

In particular, $y(n)$ is a mean zero wide sense stationary process whose autocorrelation function is given by

$$R_y(n) = C\mathbf{g}(U)U^n\mathbf{g}(U)^*C^*\,. \tag{6.12}$$

PROOF. Substituting $\xi(k) = CU^k x_0$ into (6.10), we obtain

$$\begin{aligned} y(n) &= \sum_{k=-\infty}^{\infty} g(n-k)\xi(k) = \sum_{k=-\infty}^{\infty} g(n-k)CU^k x_0 \\ &= \sum_{k=-\infty}^{\infty} g(n-k)CU^{-(n-k)}U^n x_0 = C\sum_{j=-\infty}^{\infty} g(j)U^{-j}U^n x_0 \\ &= C\mathbf{g}(U)U^n x_0\,. \end{aligned}$$

This yields the formula for $y(n)$ in (6.11). Notice that $y(n)$ is a sinusoid process corresponding to the unitary pair $\{C\mathbf{g}(U), U\}$; see Section 4.3 in Chapter 4. Therefore $y(n)$ is a mean zero wide sense stationary process whose autocorrelation function is given by (6.12). This completes the proof.

REMARK 5.6.4 *Let $\xi(n)$ be the sinusoid process given by $\xi(n) = CU^n x_0$ where U is a unitary operator and x_0 is a mean zero vector in $\mathcal{X}$ satisfying $Ex_0x_0^* = I$. Now consider the state space system given by*

$$x(n+1) = Ax(n) + B\xi(n) \qquad (x(-\infty) = 0) \quad \textit{and} \quad y(n) = C_1x(n) + D\xi(n) \tag{6.13}$$

where A is a stable operator on $\mathcal{X}_1$ and B maps $\mathbb{C}$ into $\mathcal{X}_1$ while C_1 maps $\mathcal{X}_1$ into $\mathbb{C}$ and D is a scalar. Then the output $y(n)$ is the process determined by the causal system

$$y(n) = \sum_{k=-\infty}^{n-1} C_1A^{n-k-1}B\xi(k) + D\xi(n)\,. \tag{6.14}$$

Theorem 5.6.3 shows that $y(n)$ is the sinusoid process given by (6.11) where the transfer function $\mathbf{g}(z) = C_1(zI - A)^{-1}B + D$. Moreover, $y(n)$ is a mean zero wide sense stationary process whose autocorrelation function R_y and spectral density S_y are given by (6.12).

5.6.2 Exercise

Problem 1. Let $\xi(n)$ be the sinusoid process given by $\xi(n) = 2\cos(\pi n/4 + \theta)$ where θ is a uniform random variable over the interval $[0, 2\pi]$. Consider the state space system given by

$$\begin{aligned}\begin{bmatrix} x_1(n+1) \\ x_2(n+1) \\ x_3(n+1) \\ x_4(n+1) \end{bmatrix} &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -0.1 & 0.4 & -1.4 & 2 \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \\ x_4(n) \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \xi(n) \\ y(n) &= \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix} x(n) + \xi(n) \,.\end{aligned}$$

The initial condition $x(-\infty) = 0$. Find the autocorrelation function R_y and spectral density S_y for y.

Problem 2. Consider the sinusoid process $\xi(n) = CU^n x_0$ determined by

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} 1 & 2 \end{bmatrix} .$$

As expected, x_0 is a mean zero random vector in $\mathbb{C}^2$ satisfying $Ex_0x_0^*$. Let $y(n)$ be the sinusoid process given by $y(n) = \sum_{k=-\infty}^{\infty} g(n-k)\xi(k)$ where $\{g(k)\}_{-\infty}^{\infty}$ is the impulse response for

$$\mathbf{g} = \frac{40z^2 + 10}{z^4 + 2z^3 + 4z^2 - 5z - 6} \,.$$

Theorem 5.6.3 shows that $y(n) = C\mathbf{g}(U)U^n x_0$ is a sinusoid process. Using Matlab along with the fast Fourier transform compute $C\mathbf{g}(U)$. Find the autocorrelation function $\{R_y(n)\}_{n=0}^{10}$ and spectral density S_y for y.

Chapter 6

Levinson filtering

In this section we will develop and use the Levinson algorithm to solve a simple prediction problem. Recall that a random process $y(n)$ is *wide sense stationary* if $Ey(n) = c$ is constant for all integers n and $Ey(n)y(m)^* = f(n-m)$ is a function of $n-m$ for all integers n and m. If $y(n)$ is wide sense stationary, then the autocorrelation $R_y(n)$ for $y(n)$ is the function defined by $Ey(n)y(0)^* = R_y(n)$. In particular, $Ey(n)y(m)^* = R_y(n-m)$ for all integers n and m. Finally, it is noted that $Ey(n)y(m)^* = f(n-m)$ for all integers n and m if and only if $Ey(n+k)y(k)^* = f(n)$ is a function of just n for all integers n and k. So $y(n)$ is wide sense stationary if and only if $Ey(n) = c$ is constant for all integers n and $Ey(n+k)y(k)^* = f(n)$ is a function of n for all integers n and k. In this case, $R_y(n) = Ey(n+k)y(k)^*$.

6.1 Levinson prediction

Let $y(n)$ be a scalar valued wide sense stationary random process. Let $R_y(n)$ be the autocorrelation function for $y(n)$, that is, $R_y(n-m) = Ey(n)\overline{y(m)}$. Because $y(n)$ is a scalar valued, $R_y(n) = R_y(-n)^* = \overline{R}_y(-n)$ for all integers n. For any integer $\nu \geq 1$, let T_ν be the Toeplitz matrix on $\mathbb{C}^\nu$ generated by $R_y(n)$, that is,

$$T_\nu = \begin{bmatrix} R_y(0) & \overline{R}_y(1) & \cdots & \overline{R}_y(\nu-1) \\ R_y(1) & R_y(0) & \cdots & \overline{R}_y(\nu-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_y(\nu-1) & R_y(\nu-2) & \cdots & R_y(0) \end{bmatrix}. \tag{1.1}$$

Using $R_y(n-m) = Ey(n)\overline{y(m)}$, it follows that $T_\nu = Egg^*$ where g is the random vector with values in $\mathbb{C}^\nu$ defined by $g = [y(n), y(n+1), \cdots, y(n+\nu-1)]^{tr}$. (The transpose is denoted by tr.) Hence T_ν is positive. Moreover, if $k < \nu$, then T_k is the Toeplitz matrix on $\mathbb{C}^k$ contained in the upper left hand corner of T_ν.

Now consider the one step ahead prediction problem of finding the best estimate $\widehat{y}(n+1)$ of $y(n+1)$ given the past $\{y(n), y(n-1), \cdots, y(n-k+1)\}$. To be precise, let $\mathcal{H}$ be the space of all random variables generated by

$$\mathcal{H} = \text{span}\,\{y(n), y(n-1), \cdots, y(n-k+1)\}. \tag{1.2}$$

Notice that $\mathcal{H}$ is simply the space of random variables formed by the linear span of the past k random variables $\{y(n-j)\}_{j=0}^{k-1}$. According to the projection theorem, the best estimate $\widehat{y}(n+1)$ of $y(n+1)$ given the past $\{y(n-j)\}_{j=0}^{k-1}$ is given by $\widehat{y}(n+1) = P_{\mathcal{H}} y(n+1)$ where $P_{\mathcal{H}}$ is the orthogonal projection onto $\mathcal{H}$. The following result provides a solution to this prediction problem.

THEOREM 6.1.1 *Let $y(n)$ be a scalar valued wide sense stationary random process and assume that the Toeplitz matrix T_{k+1} on $\mathbb{C}^{k+1}$ generated by $\{R_y(n)\}_0^k$ is strictly positive. Then there exists a unique solution to*

$$\left[\begin{array}{ccccc} \alpha_1 & \alpha_2 & \cdots & \alpha_k & 1 \end{array}\right] T_{k+1} = \left[\begin{array}{ccccc} 0 & 0 & 0 & \cdots & 0 & \epsilon \end{array}\right]. \tag{1.3}$$

Moreover, $\epsilon > 0$ and the optimal estimate $\widehat{y}(n+1)$ of $y(n+1)$ given the past $\mathcal{H} = \bigvee_{j=0}^{k-1} y(n-j)$ is computed by

$$\widehat{y}(n+1) = P_{\mathcal{H}} y(n+1) = -\sum_{j=0}^{k-1} \alpha_{k-j} y(n-j)\,. \tag{1.4}$$

Finally, the estimation error is given by

$$E|y(n+1) - \widehat{y}(n+1)|^2 = \epsilon\,. \tag{1.5}$$

PROOF. Because T_{k+1} is invertible, there exists a unique solution to

$$\left[\begin{array}{ccccc} \beta_1 & \cdots & \beta_k & \delta \end{array}\right] T_{k+1} = \left[\begin{array}{ccccc} 0 & 0 & \cdots & 0 & 1 \end{array}\right]. \tag{1.6}$$

Since T_{k+1} is strictly positive and $\left[\begin{array}{ccccc} \beta_1 & \cdots & \beta_k & \delta \end{array}\right]$ is nonzero,

$$\begin{aligned} 0 &< \left[\begin{array}{ccccc} \beta_1 & \cdots & \beta_k & \delta \end{array}\right] T_{k+1} \left[\begin{array}{ccccc} \beta_1 & \cdots & \beta_k & \delta \end{array}\right]^* \\ &= \left[\begin{array}{ccccc} 0 & \cdots & 0 & 1 \end{array}\right] \left[\begin{array}{ccccc} \beta_1 & \cdots & \beta_k & \delta \end{array}\right]^* = \delta\,. \end{aligned}$$

Hence $\delta > 0$. By setting $\epsilon = 1/\delta$ and dividing (1.6) by $1/\delta$, we arrive at the Levinson system in (1.3). In particular, there is a unique solution to (1.3) and $\epsilon > 0$.

Clearly, the optimal estimate $\widehat{y}(n+1)$ is a linear combination of $\{y(n-j)\}_{j=0}^{k-1}$, that is,

$$\widehat{y}(n+1) = P_{\mathcal{H}} y(n+1) = -\sum_{j=0}^{k-1} \alpha_{k-j} y(n-j)$$

where $\{\alpha_j\}_1^k$ are scalars. Let α be the row vector and g be the random vector with values in $\mathbb{C}^k$ given by

$$\begin{aligned} \alpha &= \left[\begin{array}{ccccc} \alpha_1 & \alpha_2 & \cdots & \alpha_{k-1} & \alpha_k \end{array}\right] \\ g &= \left[\begin{array}{ccccc} y(n-k+1) & y(n-k+2) & \cdots & y(n-1) & y(n) \end{array}\right]^{tr}. \end{aligned} \tag{1.7}$$

Notice that $\mathcal{H} = \bigvee g$ and $P_{\mathcal{H}} y(n+1) = -\alpha g$. Using $R_y(n-m) = Ey(n)\overline{y(m)}$, it follows that $Egg^* = T_k$. According to the projection theorem, $y(n+1) - \widehat{y}(n+1)$ is orthogonal to $\mathcal{H}$. Hence

$$0 = E(y(n+1) + \alpha g)g^* = Ey(n+1)g^* + \alpha T_k = \left[\begin{array}{cc} \alpha & 1 \end{array}\right] \left[\begin{array}{c} T_k \\ Ey(n+1)g^* \end{array}\right]. \tag{1.8}$$

By employing $R_y(n-m) = Ey(n)\overline{y(m)}$ once again, we have

$$Ey(n+1)g^* = \left[\begin{array}{cccc} R_y(k) & R_y(k-1) & \cdots & R_y(1) \end{array} \right] . \tag{1.9}$$

Substituting this into (1.8) gives

$$0 = \left[\begin{array}{cccc} \alpha_1 & \cdots & \alpha_k & 1 \end{array} \right] \left[\begin{array}{cccc} R_y(0) & \overline{R}_y(1) & \cdots & \overline{R}_y(k-1) \\ R_y(1) & R_y(0) & \cdots & \overline{R}_y(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_y(k-1) & R_y(k-2) & \cdots & R_y(0) \\ R_y(k) & R_y(k-1) & \cdots & R_y(1) \end{array} \right] . \tag{1.10}$$

Because $y(n+1) - \widehat{y}(n+1)$ is orthogonal to $\widehat{y}(n+1)$, we have

$$\begin{aligned} E|y(n+1) - \widehat{y}(n+1)|^2 &= E(y(n+1) - \widehat{y}(n+1))(y(n+1) - \widehat{y}(n+1))^* \\ &= E|\, y(n+1)|^2 - E\widehat{y}(n+1)\overline{y(n+1)} \\ &= R_y(0) + E\alpha g\overline{y(n+1)} \\ &= R_y(0) + \alpha \left[\begin{array}{cccc} R_y(k) & R_y(k-1) & \cdots & R_y(1) \end{array} \right]^* \end{aligned}$$

The last equality follows from $Eg\overline{y(n+1)} = (Ey(n+1)g^*)^*$; see (1.9). In other words,

$$E|y(n+1) - \widehat{y}(n+1)|^2 = \left[\begin{array}{cccccc} \alpha_1 & \cdots & \alpha_k & 1 \end{array} \right] \left[\begin{array}{cccccc} \overline{R}_y(k) & \overline{R}_y(k-1) & \cdots & \overline{R}_y(1) & R_y(0) \end{array} \right]^{tr} .$$

Notice that $\left[\begin{array}{ccc} \overline{R}_y(k) & \cdots & R_y(0) \end{array} \right]^{tr}$ is the last column of the Toeplitz matrix T_{k+1}. By combining this with (1.10), we arrive at the Levinson system of equations in (1.3) where $\epsilon = E|y(n+1) - \widehat{y}(n+1)|^2$. Because the solution to (1.3) is unique, this completes the proof.

REMARK 6.1.2 *Let $y(n)$ be a scalar valued wide sense stationary random process with autocorrelation function $R_y(n)$. Let f be any random vector in $\mathbb{C}^\mu$. Let $\mathcal{H}$ be the subspace of random variables spanned $\{y(n-j)\}_{j=0}^{k-1}$. Furthermore, assume that the Toeplitz matrix T_k on $\mathbb{C}^k$ generated by $\{R_y(j)\}_{j=0}^{k-1}$ is strictly positive. Let g be the random vector in $\mathbb{C}^k$ given by (1.7). Then the orthogonal projection $\widehat{f} = P_{\mathcal{H}} f$ is computed by*

$$\widehat{f} = P_{\mathcal{H}} f = R_{fg} T_k^{-1} \left[\begin{array}{ccccc} y(n-k+1) & y(n-k+2) & \cdots & y(n-1) & y(n) \end{array} \right]^{tr} . \tag{1.11}$$

As expected, $R_{fg} = Efg^*$. To verify that (1.11) holds simply observe that $T_k = Egg^*$. Then the formula in (1.11) follows from Theorem 2.2.1. Moreover, Theorem 2.2.1 shows that the estimation error $E(f-\widehat{f})(f-\widehat{f})^* = R_f - R_{fg}T_k^{-1}R_{gf}$ where $R_f = Eff^*$. Finally, it is noted that Corollary 6.4.3 below presents an efficient algorithm to compute the inverse of T_k.

6.2 Levinson smoothing

In stochastic processes a smoothing problem is estimating the past given the future. Now let us introduce the Levinson smoothing problem which is the dual of the Levinson prediction

problem. As before, let $y(n)$ be a scalar valued wide sense stationary random process. Throughout T_ν on $\mathbb{C}^\nu$ is the Toeplitz matrix in (1.1) generated by $\{R_y(n)\}_0^{\nu-1}$. Now consider the one step smoothing problem of finding the best estimate $\widehat{y}(n)$ of $y(n)$ given the future $\{y(n+1), y(n+2), \cdots, y(n+k)\}$. To be precise, let $\mathcal{G}$ be the space of all random variables generated by

$$\mathcal{G} = \text{span } \{y(n+1), y(n+2), \cdots, y(n+k)\}\,. \tag{2.1}$$

According to the projection theorem, the best estimate $\widehat{y}(n)$ of $y(n)$ given the future space $\{y(n+j)\}_{j=1}^k$ is given by $\widehat{y}(n) = P_{\mathcal{G}} y(n)$ where $P_{\mathcal{G}}$ is the orthogonal projection onto $\mathcal{G}$. The following result provides a solution to this smoothing problem.

THEOREM 6.2.1 *Let $y(n)$ be a scalar valued wide sense stationary random process and assume that the Toeplitz matrix T_{k+1} on $\mathbb{C}^{k+1}$ generated by $\{R_y(n)\}_0^k$ is strictly positive. Then there exists a unique solution to*

$$\begin{bmatrix} 1 & \beta_1 & \beta_2 & \cdots & \beta_k \end{bmatrix} T_{k+1} = \begin{bmatrix} \sigma & 0 & 0 & \cdots & 0 \end{bmatrix}. \tag{2.2}$$

Moreover, $\sigma > 0$ and the optimal estimate $\widehat{y}(n)$ of $y(n)$ given the future $\mathcal{G} = \bigvee_{j=1}^k y(n+j)$ is computed by

$$\widehat{y}(n) = P_{\mathcal{G}} y(n) = -\sum_{j=1}^k \beta_j y(n+j)\,. \tag{2.3}$$

Finally, the estimation error is given by $E|y(n) - \widehat{y}(n)|^2 = \sigma$.

Proof. By adjusting the argument in the first paragraph of the proof of Theorem 6.1.1, it follows that there exists a unique solution to the Levinson system in (2.2) and $\sigma > 0$; see also Lemma 6.2.2 below. Clearly, the optimal estimate $\widehat{y}(n)$ is a linear combination of $\{y(n+j)\}_{j=1}^k$, that is,

$$\widehat{y}(n) = P_{\mathcal{G}} y(n) = -\sum_{j=1}^k \beta_j y(n+j)$$

where $\{\beta_j\}_1^k$ are scalars. Let β be the row vector and g be the random vector with values in $\mathbb{C}^k$ given by

$$\begin{aligned} \beta &= \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_{k-1} & \beta_k \end{bmatrix} \\ g &= \begin{bmatrix} y(n+1) & y(n+2) & \cdots & y(n+k) \end{bmatrix}^{tr}. \end{aligned} \tag{2.4}$$

Notice that $\mathcal{G} = \bigvee g$ and $P_{\mathcal{G}} y(n) = -\beta g$. Using $R_y(n-m) = Ey(n)\overline{y(m)}$, it follows that $Egg^* = T_k$. According to the projection theorem, $y(n) - \widehat{y}(n)$ is orthogonal to $\mathcal{G}$. Hence

$$0 = E(y(n) + \beta g)g^* = Ey(n)g^* + \beta T_k = \begin{bmatrix} 1 & \beta \end{bmatrix} \begin{bmatrix} Ey(n)g^* \\ T_k \end{bmatrix}. \tag{2.5}$$

By employing $R_y(n-m) = Ey(n)\overline{y(m)}$ once again, we have

$$Ey(n)g^* = \begin{bmatrix} \overline{R}_y(1) & \overline{R}_y(2) & \cdots & \overline{R}_y(k) \end{bmatrix}.$$

Substituting this into (2.5) gives

$$0 = \begin{bmatrix} 1 & \beta_1 & \cdots & \beta_k \end{bmatrix} \begin{bmatrix} \overline{R}_y(1) & \overline{R}_y(2) & \cdots & \overline{R}_y(k) \\ R_y(0) & \overline{R}_y(1) & \cdots & \overline{R}_y(k-1) \\ R_y(1) & R_y(0) & \cdots & \overline{R}_y(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_y(k-1) & R_y(k-2) & \cdots & R_y(0) \end{bmatrix}. \tag{2.6}$$

Because $y(n) - \widehat{y}(n)$ is orthogonal to $\widehat{y}(n)$, we have

$$\begin{aligned} E|y(n) - \widehat{y}(n)|^2 &= E(y(n) - \widehat{y}(n))(y(n) - \widehat{y}(n))^* \\ &= E|y(n)|^2 - E\widehat{y}(n)\overline{y(n)} \\ &= R_y(0) + E\beta g\overline{y(n)} \\ &= R_y(0) + \beta \begin{bmatrix} R_y(1) & R_y(2) & \cdots & R_y(k) \end{bmatrix}^{tr} \end{aligned}$$

In other words,

$$E|y(n) - \widehat{y}(n)|^2 = \begin{bmatrix} 1 & \beta_1 & \cdots & \beta_k \end{bmatrix} \begin{bmatrix} R_y(0) & R_y(1) & \cdots & R_y(k) \end{bmatrix}^{tr}.$$

Notice that $\begin{bmatrix} R_y(0) & \cdots & R_y(k) \end{bmatrix}^{tr}$ is the first column of the Toeplitz matrix T_{k+1}. By combining this with (2.6), we arrive at the Levinson system of equations in (2.2) where $\sigma = E|y(n) - \widehat{y}(n)|^2$. Because the solution to (2.2) is unique, this completes the proof.

The prediction problem yields the forward Levinson system in (1.3), while the smoothing problem yields the backward Levinson system in (2.2). Lemma 6.2.2 below shows that $\epsilon = \sigma$ and

$$\begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_k \end{bmatrix} = \begin{bmatrix} \overline{\beta}_k & \overline{\beta}_{k-1} & \cdots & \overline{\beta}_1 \end{bmatrix}. \tag{2.7}$$

We say that T_ν is a Toeplitz matrix on $\mathbb{C}^\nu$ generated by a sequence of scalar $\{r_j\}_0^{\nu-1}$ if T_ν is a matrix of the form

$$T_k = \begin{bmatrix} r_0 & \overline{r}_1 & \cdots & \overline{r}_{k-1} \\ r_1 & r_0 & \cdots & \overline{r}_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k-1} & r_{k-2} & \cdots & r_0 \end{bmatrix} \text{ on } \mathbb{C}^\nu. \tag{2.8}$$

Clearly, T_ν is a self-adjoint operator.

LEMMA 6.2.2 *Let T_ν be a Toeplitz matrix on $\mathbb{C}^\nu$ generated by a sequence of scalar $\{r_j\}_0^{\nu-1}$, that is, T_ν is given by (2.8) with $k = \nu$. Then*

$$\begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{\nu-1} & 1 \end{bmatrix} T_\nu = \begin{bmatrix} 0 & 0 & \cdots & 0 & \epsilon \end{bmatrix} \tag{2.9}$$

if and only if

$$\begin{bmatrix} 1 & \overline{\alpha}_{\nu-1} & \overline{\alpha}_{\nu-2} & \cdots & \overline{\alpha}_1 \end{bmatrix} T_\nu = \begin{bmatrix} \epsilon & 0 & \cdots & 0 & 0 \end{bmatrix}. \tag{2.10}$$

PROOF. First notice that ϵ is a real number. If (2.9) holds, then

$$\epsilon = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{\nu-1} & 1 \end{bmatrix} T_\nu \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{\nu-1} & 1 \end{bmatrix}^* .$$

Because T_ν is a self-adjoint operator, the last term is a real number. Hence ϵ is real. A similar argument shows that if (2.10) holds, then the scalar ϵ in (2.10) is also real.

Let J be the matrix on $\mathbb{C}^\nu$ with one's on the off diagonal and zero's elsewhere, that is,

$$J = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix} . \tag{2.11}$$

It is easy to verify that $J^2 = I$ and $JT_\nu J = T_\nu^{tr}$. Recall that tr denotes the transpose. Now assume that (2.9) holds. Set $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_{\nu-1}, 1]$ and $b = [0, 0, \cdots, 0, \epsilon]$. Then $\alpha T_\nu = b$, and thus, $\alpha J T_\nu J = bJ$. Using $JT_\nu J = T_\nu^{tr}$, we obtain

$$\begin{bmatrix} 1 & \alpha_{\nu-1} & \alpha_{\nu-2} & \cdots & \alpha_1 \end{bmatrix} T_\nu^{tr} = \alpha J T_\nu^{tr} = bJ = \begin{bmatrix} \epsilon & 0 & \cdots & 0 & 0 \end{bmatrix} . \tag{2.12}$$

By taking the complex conjugate, we arrive at (2.10). On the other hand, if (2.10) holds, then multiplying by J on the right hand side of (2.12), yields $\alpha T_\nu = \alpha J T_\nu J = b$, or equivalently, (2.9) holds. This completes the proof.

REMARK 6.2.3 *Assume that T_{k+1} generated by $\{r_j\}_0^k$ is strictly positive. Moreover, let $\{\alpha_j\}_0^k$ with ϵ and $\{\beta_j\}_0^k$ with σ be the unique solutions to the following forward and backward Levinson systems*

$$\begin{aligned} \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_k & 1 \end{bmatrix} T_{k+1} &= \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & \epsilon \end{bmatrix} \\ \begin{bmatrix} 1 & \beta_1 & \beta_2 & \cdots & \beta_k \end{bmatrix} T_{k+1} &= \begin{bmatrix} \sigma & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} . \end{aligned}$$

Then Lemma 6.2.2 shows that $\epsilon = \sigma$ and

$$\begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_k \end{bmatrix} = \begin{bmatrix} \overline{\beta}_k & \overline{\beta}_{k-1} & \cdots & \overline{\beta}_1 \end{bmatrix} . \tag{2.13}$$

Finally, it is noted that $\epsilon > 0$.

6.3 The Levinson algorithm

In this section we will present the Levinson algorithm to recursively solve the Levinson system of equations in (1.3). To this end, let $\{r_j\}_0^\infty$ be a sequence of scalars and T_k the Toeplitz matrix on $\mathbb{C}^k$ generated by $\{r_j\}_0^{k-1}$, that is,

$$T_k = \begin{bmatrix} r_0 & \overline{r}_1 & \cdots & \overline{r}_{k-1} \\ r_1 & r_0 & \cdots & \overline{r}_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k-1} & r_{k-2} & \cdots & r_0 \end{bmatrix} . \tag{3.1}$$

Clearly, T_k is a self-adjoint operator. However, for the moment we have not assumed that T_k is positive. Associated with T_k is the following Levinson system of equations

$$\left[\begin{array}{ccccc} \alpha_{k,1} & \alpha_{k,2} & \cdots & \alpha_{k,k-1} & 1 \end{array}\right] T_k = \left[\begin{array}{ccccc} 0 & 0 & \cdots & 0 & \epsilon_k \end{array}\right] \qquad (k = 1, 2, 3, \cdots). \tag{3.2}$$

If $k = 1$, then $\epsilon_1 = T_1 = r_0$ and the $\alpha_{1,0}$ term is not present. Obviously, one can compute $\{\alpha_{k,m}\}_{m=1}^{k-1}$ and ϵ_k by inverting the Toeplitz matrix T_k. However, this inversion involves approximately k^3 operations. The following result known as the Levinson algorithm provides an efficient method to compute $\{\alpha_{k,m}\}_{m=1}^{k-1}$ and ϵ_k from $\{r_j\}_0^{k-1}$ in approximately k^2 operations.

THEOREM 6.3.1 (Levinson algorithm) *Let T_k on $\mathbb{C}^k$ be the Toeplitz matrix in (3.1) generated by a sequence of scalars $\{r_j\}_0^\infty$. Assume that $\{\alpha_{k,m}\}_{m=1}^{k-1}$ and ϵ_k is a solution to Levinson system (3.2) where $\epsilon_k \neq 0$. Let*

$$\begin{aligned} \left[\begin{array}{cccc} \alpha_{k+1,1} & \alpha_{k+1,2} & \cdots & \alpha_{k+1,k} \end{array}\right] &= \left[\begin{array}{cccc} 0 & \alpha_{k,1} & \alpha_{k,2} & \cdots & \alpha_{k,k-1} \end{array}\right] \\ &\quad -\delta_k \epsilon_k^{-1} \left[\begin{array}{cccc} 1 & \overline{\alpha}_{k,k-1} & \overline{\alpha}_{k,k-2} & \cdots & \overline{\alpha}_{k,1} \end{array}\right] \end{aligned} \tag{3.3}$$

where δ_k and the next ϵ_{k+1} are given by

$$\delta_k = r_k + \sum_{j=1}^{k-1} \alpha_{k,j} r_j \qquad \text{and} \qquad \epsilon_{k+1} = \epsilon_k - |\delta_k|^2/\epsilon_k \,. \tag{3.4}$$

Then $\{\alpha_{k+1,m}\}_{m=1}^{k}$ and ϵ_{k+1} is a solution to the Levinson system in (3.2) for $k+1$, that is,

$$\left[\begin{array}{ccccc} \alpha_{k+1,1} & \alpha_{k+1,2} & \cdots & \alpha_{k+1,k} & 1 \end{array}\right] T_{k+1} = \left[\begin{array}{ccccc} 0 & 0 & \cdots & 0 & \epsilon_{k+1} \end{array}\right]. \tag{3.5}$$

Moreover, if $\epsilon_j \neq 0$ for $j = 1, 2, \cdots, k$, then one can recursively compute $\{\alpha_{j,m}\}_{m=1}^{j-1}$ for $j = 2, 3, \cdots, k+1$ by recursively solving for $\{\alpha_{j,m}\}_{m=1}^{j-1}$ in (3.3) starting with the initial condition $\epsilon_1 = r_0$, or $\alpha_{2,1} = -r_1/r_0$ and $\epsilon_2 = r_0 - |r_1|^2/r_0$.

Proof. Notice that the Toeplitz matrix T_{k+1} admits a decomposition of the form

$$T_{k+1} = \left[\begin{array}{cc} r_0 & z^* \\ z & T_k \end{array}\right] \text{ on } \left[\begin{array}{c} \mathbb{C} \\ \mathbb{C}^k \end{array}\right] \tag{3.6}$$

where $z = [r_1, r_2, \cdots, r_k]^{tr}$. Using this decomposition and the Levinson system in (3.2), we arrive at

$$\left[\begin{array}{ccccc} 0 & \alpha_{k,1} & \alpha_{k,2} & \cdots & \alpha_{k,k-1} & 1 \end{array}\right] T_{k+1} = \left[\begin{array}{ccccc} \delta_k & 0 & \cdots & 0 & 0 & \epsilon_k \end{array}\right]. \tag{3.7}$$

Here $\delta_k = r_k + \sum_{j=1}^{k-1} \alpha_{k,j} r_j$. According to Lemma 6.2.2, the Levinson system in (3.2) can be rewritten as

$$\left[\begin{array}{ccccc} 1 & \overline{\alpha}_{k,k-1} & \overline{\alpha}_{k,k-2} & \cdots & \overline{\alpha}_{k,1} \end{array}\right] T_k = \left[\begin{array}{ccccc} \epsilon_k & 0 & \cdots & 0 & 0 \end{array}\right]. \tag{3.8}$$

The Toeplitz matrix T_{k+1} also admits a matrix decomposition of the form

$$T_{k+1} = \begin{bmatrix} T_k & x \\ x^* & r_0 \end{bmatrix} \text{ on } \begin{bmatrix} \mathbb{C}^k \\ \mathbb{C} \end{bmatrix} \tag{3.9}$$

where $x = [r_k, r_{k-1}, \cdots, r_1]^*$. By employing this decomposition and the Levinson system in (3.8), we obtain

$$\begin{bmatrix} 1 & \overline{\alpha}_{k,\,k-1} & \overline{\alpha}_{k,\,k-2} & \cdots & \overline{\alpha}_{k,\,1} & 0 \end{bmatrix} T_{k+1} = \begin{bmatrix} \epsilon_k & 0 & \cdots & 0 & 0 & \overline{\delta}_k \end{bmatrix} . \tag{3.10}$$

Multiplying (3.10) by $-\delta_k/\epsilon_k$ and adding this to (3.7), yields

$$\begin{aligned} &\{\begin{bmatrix} 0 & \alpha_{k,\,1} & \cdots & \alpha_{k,\,k-1} & 1 \end{bmatrix} - \delta_k/\epsilon_k \begin{bmatrix} 1 & \overline{\alpha}_{k,\,k-1} & \cdots & \overline{\alpha}_{k,\,1} & 0 \end{bmatrix}\} T_{k+1} \\ &= \begin{bmatrix} 0 & 0 & \cdots & 0 & \epsilon_k - |\delta_k|^2/\epsilon_k \end{bmatrix} . \end{aligned}$$

So if the $\{\alpha_{k+1,\,m}\}_{m=1}^{k}$ are given by the recursion in (3.3) and $\epsilon_{k+1} = \epsilon_k - |\delta_k|^2/\epsilon_k$, then

$$\begin{bmatrix} \alpha_{k+1,\,1} & \alpha_{k+1,\,2} & \cdots & \alpha_{k+1,\,k} & 1 \end{bmatrix} T_{k+1} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \epsilon_{k+1} \end{bmatrix} .$$

This is a solution to the Levinson system in (3.2) for $k+1$. This completes the proof.

6.3.1 Reflection coefficients.

As before, let $\{r_j\}_0^\infty$ be a set of scalars. Let T_k on $\mathbb{C}^k$ be the Toeplitz matrix in (3.1) generated by $\{r_j\}_0^{k-1}$. Moreover, assume that T_k is strictly positive for all integers $k \geq 1$. Let $\{\alpha_{k,\,m}\}_{m=1}^{k-1}$ and ϵ_k be the solution to the Levinson system of equations in (3.2). Recall that

$$\delta_k = r_k + \sum_{j=1}^{k-1} \alpha_{k,\,j} r_j \qquad \text{and} \qquad \epsilon_{k+1} = \epsilon_k - |\delta_k|^2/\epsilon_k \,. \tag{3.11}$$

The k-th *reflection coefficient* c_k for $\{r_j\}_0^\infty$ is defined by $c_k = \delta_k/\epsilon_k$ for $k \geq 1$. Equation (3.3) shows that $c_k = -\alpha_{k+1,\,1}$. (Notice that our definition of the reflection coefficient is the complex conjugate of the corresponding Schur number or reflection coefficient in [13]. In most applications the data $\{r_j\}_0^\infty$ is real, and thus, the reflection coefficients $\{c_j\}_1^\infty$ are also real.) We claim that $|c_k| < 1$ and

$$\epsilon_{k+1} = r_0 \prod_{j=1}^{k} (1 - |c_j|^2) \qquad (k \geq 1). \tag{3.12}$$

In particular, this shows that $\{\epsilon_k\}_1^\infty$ is decreasing, that is, $\epsilon_{k+1} \leq \epsilon_k$.

Using $\epsilon_1 = r_0$ and (3.11), we obtain

$$\epsilon_2 = \epsilon_1 - |\delta_1|^2/\epsilon_1 = \epsilon_1(1 - |\delta_1|^2/\epsilon_1^2) = r_0(1 - |c_1|^2) \,.$$

Hence $\epsilon_2 = r_0(1-|c_1|^2)$ and (3.12) holds for $k = 1$. Since $\epsilon_2 > 0$ and $r_0 > 0$, we have $|c_1| < 1$. Now let us use induction and assume (3.12) holds for $k-1$, that is, $\epsilon_k = r_0 \prod_{j=1}^{k-1}(1 - |c_j|^2)$

for some $k \geq 3$. Then

$$\begin{aligned} \epsilon_{k+1} &= \epsilon_k - |\delta_k|^2/\epsilon_k = \epsilon_k(1 - |\delta_k|^2/\epsilon_k^2) = \epsilon_k(1 - |c_k|^2) \\ &= (1 - c_k^2) r_0 \prod_{j=1}^{k-1}(1 - |c_j|^2) = r_0 \prod_{j=1}^{k}(1 - |c_j|^2)\,. \end{aligned}$$

Because $0 < \epsilon_{k+1}$ and $\epsilon_{k+1} = \epsilon_k(1 - |c_k|^2)$, this also shows that $|c_k| < 1$. Therefore by induction $|c_k| < 1$ and (3.12) holds for all k.

In marine seismology the covariance sequence $\{r_k\}_0^\infty$ is the seismic data; see [4, 13]. A fundamental problem in marine seismology is to compute the reflection coefficients $\{c_k\}_1^\infty$ from the seismic data $\{r_k\}_0^\infty$. Therefore the Levinson algorithm provides us with an efficient algorithm to solve this problem, that is the Levinson algorithm can be used to compute the reflection coefficients $\{c_k\}_1^\nu$ from the data $\{r_k\}_0^\nu$ in approximately ν^2 operations.

6.3.2 Exercise

Problem 1. Let T_{20} be the Toeplitz matrix generated by $\{r_j\}_0^{19}$ where $\{r_0, r_1, \cdots, r_{19}\}$ are respectively given by

$$\{600, 420, 250, 200, 136, 120, 110, 100, 91, 84, 75, 70, 60, 55, 50, 48, 44, 40, 38, 35\}\,.$$

Is the Toeplitz matrix T_{20} strictly positive. If so compute the reflection coefficients $\{c_k\}_1^{19}$ from the data $\{r_k\}_0^{19}$.

6.4 Factoring Toeplitz matrices

In this section we will use the Levinson system to obtain a special factorization for the inverse of a strictly positive Toeplitz matrix. If $\{\lambda_k\}_1^\nu$ is a set of complex numbers, then diag $\{\lambda_k\}_1^\nu$ diagonal matrix on $\mathbb{C}^\nu$ defined by

$$\text{diag}\,\{\lambda_k\}_1^\nu = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_\nu \end{bmatrix}. \tag{4.1}$$

The following result shows that the Levinson system of equations can be used to convert T_ν to a diagonal matrix.

LEMMA 6.4.1 *Let T_ν be the Toeplitz matrix on $\mathbb{C}^\nu$ generated by a sequence of scalars $\{r_j\}_0^{\nu-1}$. Moreover, assume that the Levinson system*

$$\begin{bmatrix} \alpha_{k,1} & \alpha_{k,2} & \cdots & \alpha_{k,k-1} & 1 \end{bmatrix} T_k = \begin{bmatrix} 0 & 0 & \cdots & 0 & \epsilon_k \end{bmatrix} \tag{4.2}$$

has a solution for $k = 1, 2, \cdots, \nu$ where $\epsilon_1 = r_0$. Let Δ be the lower triangular matrix on $\mathbb{C}^\nu$ defined by

$$\Delta = \begin{bmatrix} \alpha_{1,1} & 0 & 0 & \cdots & 0 \\ \alpha_{2,1} & \alpha_{2,2} & 0 & \cdots & 0 \\ \alpha_{3,1} & \alpha_{3,2} & \alpha_{3,3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{\nu,1} & \alpha_{\nu,2} & \alpha_{\nu,3} & \cdots & \alpha_{\nu,\nu} \end{bmatrix} \text{ on } \mathbb{C}^\nu . \tag{4.3}$$

where $\alpha_{k,k} = 1$ for all $k = 1, 2, \cdots, \nu$. Then $\Delta T_\nu \Delta^* = \text{diag}\,\{\epsilon_k\}_1^\nu$. In particular, T_ν is strictly positive if and only if $\epsilon_k > 0$ for all $k = 1, 2, \cdots, \nu$. Finally, T_ν is positive and singular if and only if $\{\epsilon_k\}_1^\nu$ are all positive and at least one ϵ_j is zero.

PROOF. Notice that T_ν admits a matrix decomposition of the form

$$T_\nu = \begin{bmatrix} T_k & X^* \\ X & T_{\nu-k} \end{bmatrix} \text{ on } \begin{bmatrix} \mathbb{C}^k \\ \mathbb{C}^{\nu-k} \end{bmatrix} . \tag{4.4}$$

(Here X is an operator from $\mathbb{C}^k$ into $\mathbb{C}^{\nu-k}$.) The k-th row Δ_k of Δ is given by

$$\Delta_k = \begin{bmatrix} \alpha_{k,1} & \alpha_{k,2} & \cdots & \alpha_{k,k} & 0 & 0 & \cdots & 0 \end{bmatrix} . \tag{4.5}$$

To verify that $\Delta T_\nu \Delta^* = \text{diag}\,\{\epsilon_k\}_1^\nu$, it is sufficient to show that $\Delta_k T_\nu \Delta_m^* = \epsilon_k \delta_{k,m}$ where $\delta_{k,m}$ is the Kronecker delta. Recall that T_k is contained in the upper left hand corner of T_ν; see (4.4). Using equation (4.2) and $\alpha_{k,k} = 1$, we have

$$\Delta_k T_\nu \Delta_k^* = \begin{bmatrix} \alpha_{k,1} & \cdots & \alpha_{k,k} \end{bmatrix} T_k \begin{bmatrix} \alpha_{k,1} & \cdots & \alpha_{k,k} \end{bmatrix}^* = \epsilon_k .$$

Hence $\Delta_k T_\nu \Delta_k^* = \epsilon_k$ for all $k = 1, 2, \cdots, \nu$. If $m < k$, then (4.7) implies that

$$\begin{aligned} \Delta_k T_\nu \Delta_m^* &= \begin{bmatrix} \alpha_{k,1} & \cdots & \alpha_{k,k} \end{bmatrix} T_k \begin{bmatrix} \alpha_{m,1} & \cdots & \alpha_{m,m} & 0 & \cdots & 0 \end{bmatrix}^* \\ &= \begin{bmatrix} 0 & 0 & \cdots & 0 & \epsilon_k \end{bmatrix} \begin{bmatrix} \alpha_{m,1} & \cdots & \alpha_{m,m} & 0 & \cdots & 0 \end{bmatrix}^* = 0 . \end{aligned}$$

Thus $\Delta_k T_\nu \Delta_m^* = 0$ when $m < k$. If $m > k$, then $(\Delta_k T_\nu \Delta_m^*)^* = \Delta_m T_\nu \Delta_k^* = 0$ by our previous argument. So $\Delta_k T_\nu \Delta_m^* = \epsilon_k \delta_{k,m}$. To complete the proof notice that the k-m component of $\Delta T_\nu \Delta^*$ is given by $\{\Delta T_\nu \Delta^*\}_{k,m} = \Delta_k T_\nu \Delta_m^* = \epsilon_k \delta_{k,m}$. Therefore $\Delta T_\nu \Delta^* = \text{diag}\,\{\epsilon_k\}_1^\nu$.

Because Δ is invertible, it follows that T_ν is positive (respectively strictly positive) if and only if $\text{diag}\,\{\epsilon_k\}_1^\nu$ is positive (respectively strictly positive). This completes the proof.

COROLLARY 6.4.2 *Let T_ν be the Toeplitz matrix on $\mathbb{C}^\nu$ generated by a sequence of scalars $\{r_j\}_0^{\nu-1}$. Then T_ν is strictly positive if and only if $\epsilon_k > 0$ for all $k = 1, 2, \cdots, \nu$ where ϵ_k is the error in the Levinson system (3.2). In this case, $\{\epsilon_k\}_1^\nu$ are decreasing, that is, $\epsilon_{k+1} \leq \epsilon_k$ for $k = 1, 2, \cdots, \nu - 1$.*

PROOF. If $\{\epsilon_k\}_1^\nu$ are strictly positive, then T_ν is strictly positive; see Lemma 6.4.1. Now assume that T_ν is strictly positive and $k \leq \nu$. Then T_k is also strictly positive. To see this

recall that T_k is the $k \times k$ matrix contained in the upper left hand corner of T_ν; see (4.4). If f is any nonzero vector in $\mathbb{C}^k$, then the decomposition in (4.4) gives

$$(T_k f, f) = (T_\nu \begin{bmatrix} f & 0 & \cdots & 0 \end{bmatrix}^{tr}, \begin{bmatrix} f & 0 & \cdots & 0 \end{bmatrix}^{tr}) > 0\,.$$

Hence T_k is strictly positive. Since T_k is strictly positive, there exists a unique solution to the Levinson system $\begin{bmatrix} \alpha_{k,1} & \cdots & \alpha_{k,k-1} & 1 \end{bmatrix} T_k = \begin{bmatrix} 0 & \cdots & 0 & \epsilon_k \end{bmatrix}$. Because T_k is strictly positive, this implies that

$$0 < \begin{bmatrix} \alpha_{k,1} & \cdots & \alpha_{k,k-1} & 1 \end{bmatrix} T_k \begin{bmatrix} \alpha_{k,1} & \cdots & \alpha_{k,k-1} & 1 \end{bmatrix}^* = \epsilon_k\,.$$

Therefore $\epsilon_k > 0$ for all $k = 1, 2, \cdots, \nu$. The formula for ϵ_k in (3.4) guarantees that $\{\epsilon_k\}_1^\nu$ are decreasing, that is, $\epsilon_{k+1} \le \epsilon_k$. This completes the proof.

Recall that one can apply the Levinson algorithm when $\epsilon_k \neq 0$. So can use the Levinson algorithm along with Corollary 6.4.2 to determine when a Toeplitz matrix T_ν is strictly positive. To be precise, T_ν is strictly positive if and only if the errors in the Levinson algorithm $\epsilon_k > 0$ for all $k = 1, 2, \cdots, \nu$.

Assume that the Toeplitz matrix T_ν is strictly positive. As before, let $\{\alpha_{k,m}\}_{m=1}^{k-1}$ and ϵ_k form the unique solution to the Levinson system in (3.2). Corollary 6.4.2 shows that $\epsilon_k > 0$ for $k = 1, 2, \cdots, \nu$. In this case, the *normalized Levinson coefficients* $\{\beta_{k,m}\}_{m=1}^k$ are defined by

$$\begin{bmatrix} \beta_{k,1} & \cdots & \beta_{k,k-1} & \beta_{k,k} \end{bmatrix} = \begin{bmatrix} \alpha_{k,1} & \cdots & \alpha_{k,k-1} & 1 \end{bmatrix} / \sqrt{\epsilon_k} \qquad (1 \le k \le \nu) \tag{4.6}$$

where $\beta_{1,1} = \epsilon_1^{-1/2} = r_0^{-1/2}$. By dividing by $\epsilon_k^{1/2}$ in (3.2), it follows that the normalized Levinson coefficients $\{\beta_{k,m}\}_{m=1}^k$ are given by the unique solution to the equation

$$\begin{bmatrix} \beta_{k,1} & \beta_{k,2} & \cdots & \beta_{k,k-1} & \beta_{k,k} \end{bmatrix} T_k = \begin{bmatrix} 0 & 0 & \cdots & 0 & \sqrt{\epsilon_k} \end{bmatrix} \qquad (1 \le k \le \nu)\,. \tag{4.7}$$

The following result uses the normalized Levinson coefficients to compute a lower triangular factorization of T_ν^{-1}.

COROLLARY 6.4.3 *Let T_ν be a strictly positive Toeplitz matrix on $\mathbb{C}^\nu$ generated by a sequence of scalars $\{r_j\}_0^{\nu-1}$. Let $\{\beta_{k,m}\}_{m=1}^k$ be the normalized Levinson coefficients defined in (4.6). Let Ω be the lower triangular matrix on $\mathbb{C}^\nu$ defined by*

$$\Omega = \begin{bmatrix} \beta_{1,1} & 0 & 0 & \cdots & 0 \\ \beta_{2,1} & \beta_{2,2} & 0 & \cdots & 0 \\ \beta_{3,1} & \beta_{3,2} & \beta_{3,3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{\nu,1} & \beta_{\nu,2} & \beta_{\nu,3} & \cdots & \beta_{\nu,\nu} \end{bmatrix} \text{ on } \mathbb{C}^\nu\,. \tag{4.8}$$

Then $T_\nu^{-1} = \Omega^* \Omega$.

PROOF. Because T_ν is strictly positive, $\epsilon_k > 0$ for all $k = 1, 2, \cdots, \nu$. So the normalized Levinson coefficients $\{\beta_{k,m}\}_{m=1}^k$ in (4.6) are well defined for all $k = 1, 2, \cdots, \nu$. According to

Lemma 6.4.1, we have $\Delta T_\nu \Delta^* = \Lambda$ where $\Lambda = \text{diag}\,\{\epsilon_k\}_1^\nu$ and Δ is defined in (4.3). In other words, $\Lambda^{-1/2}\Delta T_\nu \Delta^* \Lambda^{-1/2}$. Notice that $\Omega = \Lambda^{-1/2}\Delta$. Therefore we arrive at $\Omega T_\nu \Omega^* = I$.

We claim that $T_\nu^{-1} = \Omega^*\Omega$. By taking the inverse of $\Omega T_\nu \Omega^* = I$, we obtain $I = \Omega^{-*} T_\nu^{-1} \Omega^{-1}$. Multiplying Ω^* on the left and Ω on the right yields $T_\nu^{-1} = \Omega^*\Omega$. In particular, T_ν^{-1} is strictly positive. To see this, let g be any nonzero vector in $\mathbb{C}^\nu$. Then $(T_\nu^{-1} g, g) = (\Omega^*\Omega g, g) = \|\Omega g\|^2 > 0$. Hence T_ν^{-1} is strictly positive. Therefore T_ν is also strictly positive. This completes the proof.

6.4.1 Exercise

Problem 1. Let $y(n)$ be a scalar valued wide sense stationary random process. Let T_ν be the Toeplitz matrix on $\mathbb{C}^\nu$ generated by $\{R_y(n)\}_0^{\nu-1}$. Let $\{\beta_{k,m}\}_{m=1}^k$ be the normalized Levinson coefficients in (4.8) computed from the Toeplitz matrix T_ν. Let $u_k(n)$ for $k = 1, 2, \cdots, \nu$ be the random processes defined by

$$\begin{bmatrix} u_1(n) \\ u_2(n) \\ \vdots \\ u_\nu(n) \end{bmatrix} = \begin{bmatrix} \beta_{1,1} & 0 & \cdots & 0 \\ \beta_{2,1} & \beta_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{\nu,1} & \beta_{\nu,2} & \cdots & \beta_{\nu,\nu} \end{bmatrix} \begin{bmatrix} y_1(n-\nu+1) \\ y_2(n-\nu+2) \\ \vdots \\ y_\nu(n) \end{bmatrix}.$$

Then show that $u_k(n)$ is a mean zero wide sense stationary random process. Moreover, show that $Eu_j(n)\overline{u_k(n)} = \delta_{j,k}$ where $\delta_{j,k}$ is the Kronecker delta. Hint $\Omega T_\nu \Omega^* = I$.

Problem 2. Let T_ν be the Toeplitz matrix on $\mathbb{C}^\nu$ generated by a sequence of scalars $\{r_k\}_0^{\nu-1}$. Recall that the reflection coefficients $\{c_k\}_1^{\nu-1}$ are defined by $c_k = \delta_k/\epsilon_k$ where δ_k and ϵ_k are computed from the Levinson algorithm; see Section 6.3.1. Show that T_ν is strictly positive if and only if $r_0 > 0$ and $|c_k| < 1$ for all $k = 1, 2, \cdots, \nu - 1$.

6.5 State space and the Levinson algorithm

In this chapter we will study a special state space realization associated with the Levinson algorithm. Let us begin with the following classical realization result in linear systems; see [8, 22, 28] for further results on realization theory.

PROPOSITION 6.5.1 *Let* $\mathbf{g}$ *be a scalar valued rational function of the form*

$$\mathbf{g} = \frac{c_1 + c_2 z + c_3 z^2 + \cdots + c_{k-1} z^{k-2} + c_k z^{k-1}}{\alpha_1 + \alpha_2 z + \alpha_3 z^2 + \cdots + \alpha_k z^{k-1} + z^k}\, b + \gamma \tag{5.1}$$

where b, γ *and* $\{\alpha_j, c_j,\}_1^k$ *are scalars. Let* A *on* $\mathbb{C}^k$ *be the companion matrix,* B *the column*

matrix from $\mathbb{C}$ *into* $\mathbb{C}^k$ *and* C *the row matrix from* $\mathbb{C}^k$ *into* $\mathbb{C}$ *defined by*

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -\alpha_1 & -\alpha_2 & -\alpha_3 & \cdots & -\alpha_{k-1} & -\alpha_k \end{bmatrix} \quad \textit{and} \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ b \end{bmatrix} \tag{5.2}$$

$$C = \begin{bmatrix} c_1 & c_2 & c_3 & \cdots & c_{k-1} & c_k \end{bmatrix}. \tag{5.3}$$

Then $\{A, B, C, \gamma\}$ *is a realization for* $\mathbf{g}$*, that is,*

$$\mathbf{g}(z) = C(zI - A)^{-1}B + \gamma\,. \tag{5.4}$$

Finally, the characteristic polynomial for A *is given by*

$$d(z) = \det[zI - A] = \alpha_1 + \alpha_2 z + \alpha_3 z^2 + \cdots + \alpha_k z^{k-1} + z^k\,. \tag{5.5}$$

In particular, A *is stable if and only if all the roots of* $d(z)$ *are contained in the open unit disc* $\{z : |z| < 1\}$*.*

Proof. By using the form of A, we have

$$(zI - A)\begin{bmatrix} 1 \\ z \\ z^2 \\ \vdots \\ z^{k-2} \\ z^{k-1} \end{bmatrix} = \begin{bmatrix} z & -1 & 0 & \cdots & 0 & 0 \\ 0 & z & -1 & \cdots & 0 & 0 \\ 0 & 0 & z & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & z & -1 \\ \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_{k-1} & z + \alpha_k \end{bmatrix} \begin{bmatrix} 1 \\ z \\ z^2 \\ \vdots \\ z^{k-2} \\ z^{k-1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ d(z) \end{bmatrix}.$$

Here $d(z) = \alpha_1 + \alpha_2 z + \cdots + \alpha_k z^{k-1} + z^k$. If z is not an eigenvalue of A or a zero of $d(z)$, then taking the inverse of $zI - A$ and $d(z)$ in the previous equation yields

$$(zI - A)^{-1}B = (zI - A)^{-1}\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} b = \begin{bmatrix} 1 \\ z \\ \vdots \\ z^{k-2} \\ z^{k-1} \end{bmatrix} \frac{b}{d(z)}\,. \tag{5.6}$$

By consulting the form of B and C in (5.2), we obtain

$$C(zI - A)^{-1}B = \frac{c_1 + c_2 z + \cdots + c_{k-1}z^{k-2} + c_k z^{k-1}}{\alpha_1 + \alpha_2 z + \cdots + \alpha_k z^{k-1} + z^k}\, b\,.$$

Therefore $\mathbf{g}(z) = C(zI - A)^{-1}B + \gamma$ where $\mathbf{g}$ is the proper rational function given in (5.1).

To complete the proof it remains to show that $d(z) = \alpha_1 + \alpha_2 z + \cdots + \alpha_k z^{k-1} + z^k$ is the characteristic polynomial for A. To this end, assume that $\mathbf{g}(z) = 1/d(z)$. The precious part of the proof shows that $1/d(z)$ admits a realization of the form $1/d(z) = C(zI - A)^{-1}B$ where A is a companion matrix of the form (5.2) while B is a column vector and C is a row vector. In particular,

$$\frac{1}{d(z)} = C(zI - A)^{-1}B = \frac{C\text{adj}(zI - A)B}{\det[zI - A]} = \frac{p(z)}{\det[zI - A]}$$

where $\text{adj}(zI - A)$ is the algebraic adjoint of A. Moreover, $p(z) = C\text{adj}(zI - A)B$ is a polynomial of degree at most $k - 1$. Notice that $d(z)$ and $\det[zI - A]$ are both monic polynomial of degree k. (Recall that a monic polynomial is a polynomial whose coefficient of the highest degree is one.) Since $1/d = p/\det[zI - A]$, the polynomials p and $\det[zI - A]$ have no common roots. Because $d(z)$ and $\det[zI - A]$ are both monic polynomial of the same degree, we must have $p(z) = 1$ for all z. Therefore $d(z) = \det[zI - A]$. This completes the proof.

Now let us return to Toeplitz matrices. Consider the Toeplitz matrix T_{k+1} on $\mathbb{C}^{k+1}$ generated by a sequence of scalars $\{r_j\}_0^k$, that is,

$$T_{k+1} = \begin{bmatrix} r_0 & \overline{r}_1 & \cdots & \overline{r}_k \\ r_1 & r_0 & \cdots & \overline{r}_{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_k & r_{k-1} & \cdots & r_0 \end{bmatrix} \text{ on } \mathbb{C}^{k+1}\,. \tag{5.7}$$

Throughout we set $r_{-j} = \overline{r}_j$ for all integers $j \geq 0$. If $y(n)$ is a wide sense stationary random process and $r_n = R_y(n)$ for $n = 0, 1, \cdots, k$, then the Toeplitz matrix T_{k+1} is positive.

As in Section 4.4.1 in Chapter 4, consider the state space system given by

$$x(n+1) = Ax(n) + Bu(n) \qquad (x(-\infty) = 0) \qquad \text{and} \qquad y(n) = Cx(n) \tag{5.8}$$

where A is stable and $u(n)$ is a white noise random process. This system is denoted by $\{A, B, C, x, y\}$. The initial condition is zero at minus infinity. Theorem 4.4.1 shows that the output $y(n)$ to this state space system is a wide sense stationary process. The following result shows that if T_{k+1} is strictly positive, then one can construct a state space realization $\{A, B, C, x, y\}$ such that $r_n = R_y(n)$ for $n = 0, 1, \cdots, k$. In this case, the wide sense stationary random process $y(n)$ can be used as a model for the covariance sequence $\{r_j\}_0^k$.

THEOREM 6.5.2 *Let T_{k+1} be a strictly positive Toeplitz matrix on $\mathbb{C}^{k+1}$ generated by a sequence of scalars $\{r_j\}_0^k$. Let $\{\alpha_m\}_{m=1}^k$ and ϵ be the solution to the Levinson system of equations*

$$\begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_k & 1 \end{bmatrix} T_{k+1} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \epsilon \end{bmatrix}. \tag{5.9}$$

Let A on $\mathbb{C}^k$ be the companion matrix, B the column matrix from $\mathbb{C}$ into $\mathbb{C}^k$ and C the row

matrix from $\mathbb{C}^k$ into $\mathbb{C}$ be defined by

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -\alpha_1 & -\alpha_2 & -\alpha_3 & \cdots & -\alpha_{k-1} & -\alpha_k \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \sqrt{\epsilon} \end{bmatrix} \tag{5.10}$$

$$C = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Then the following holds.

(i) The operator A is stable. Moreover, the characteristic polynomial for A is

$$d(z) = \det[zI - A] = \alpha_1 + \alpha_2 z + \cdots + \alpha_k z^{k-1} + z^k \,. \tag{5.11}$$

In particular, all the zeros of $d(z)$ are contained inside the open unit disc.

(ii) The Toeplitz matrix T_k on $\mathbb{C}^k$ is the controllability Gramian for the pair $\{A, B\}$, that is, $T_k = AT_kA^* + BB^*$.

(iii) Let $y(n)$ be the wide sense stationary process determined by the state space system

$$x(n+1) = Ax(n) + Bu(n) \qquad (x(-\infty) = 0) \qquad \text{and} \qquad y(n) = Cx(n) \tag{5.12}$$

where $u(n)$ is a white noise random process. Then the autocorrelation function

$$R_y(n) = CA^nT_kC^* = r_n \qquad (n = 0, 1, \cdots, k) \,. \tag{5.13}$$

(iv) The transfer function for the state space system in (5.12) is determined by

$$C(zI - A)^{-1}B = \frac{\sqrt{\epsilon}}{\alpha_1 + \alpha_2 z + \cdots + \alpha_k z^{k-1} + z^k} = \frac{\sqrt{\epsilon}}{d(z)} \,. \tag{5.14}$$

(v) The spectral density S_y for $y(n)$ is given by $S_y(\omega) = \epsilon/|d(e^{\imath\omega})|^2$. In particular, Equation (5.13) shows that $\{r_j\}_0^k$ are the first $k+1$ Fourier coefficients for $\epsilon/|d(e^{\imath\omega})|^2$, that is,

$$\frac{1}{2\pi}\int_0^{2\pi} \frac{e^{\imath\omega n}\epsilon}{|d(e^{\imath\omega})|^2}\, d\omega = r_n \qquad (\textit{for all integers } n \in [-k, k]) \,. \tag{5.15}$$

Notice that the state dimension corresponding to the stochastic system $\{A, B, C, x, y\}$ in Theorem 6.5.2 is k. In many applications on can use standard model reduction techniques to find a lower dimensional state space model to approximate the wide sense stationary process $y(n)$ and its spectral density $S_y(\omega) = \epsilon/|d(e^{\imath\omega})|^2$ in Theorem 6.5.2 satisfying (5.13); see Section 6.5.2.

PROOF OF THEOREM 6.5.2. Recall that T_{k+1} admits a matrix decomposition of the form

$$T_{k+1} = \begin{bmatrix} T_k & g_k^* \\ g_k & r_0 \end{bmatrix} \text{ on } \begin{bmatrix} \mathbb{C}^k \\ \mathbb{C} \end{bmatrix} \tag{5.16}$$

where $g_k = \begin{bmatrix} r_k & r_{k-1} & \cdots & r_1 \end{bmatrix}$. Using this decomposition Equation (5.9) implies that

$$\begin{aligned} \begin{bmatrix} -\alpha_1 & -\alpha_2 & \cdots & -\alpha_k \end{bmatrix} T_k &= \begin{bmatrix} r_k & r_{k-1} & \cdots & r_1 \end{bmatrix} = g_k \\ \begin{bmatrix} -\alpha_1 & -\alpha_2 & \cdots & -\alpha_k \end{bmatrix} g_k^* &= r_0 - \epsilon \,. \end{aligned} \tag{5.17}$$

Let V and Λ be the matrices on $\mathbb{C}^k$ defined by

$$V = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \text{ and } \Lambda = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \\ -\alpha_1 & -\alpha_2 & \cdots & -\alpha_k \end{bmatrix} . \tag{5.18}$$

The shift matrix V has ones immediately above the main diagonal and zeros elsewhere. The first $k-1$ rows of the Λ matrix are zero and the last row is $[-\alpha_1, -\alpha_2, \cdots, -\alpha_k]$. Notice that $A = V + \Lambda$. It is easy to show that VT_kV^* admits a matrix representation of the form

$$VT_kV^* = \begin{bmatrix} T_{k-1} & 0 \\ 0 & 0 \end{bmatrix} \text{ on } \begin{bmatrix} \mathbb{C}^{k-1} \\ \mathbb{C} \end{bmatrix} . \tag{5.19}$$

Notice that $g_{k-1} = [r_{k-1}, r_{k-2}, \cdots, r_1]$. By consulting (5.17), we obtain

$$\begin{aligned} AT_kA^* &= (V+\Lambda)T_k(V+\Lambda)^* = VT_kV^* + \Lambda T_kV^* + V^*T_k\Lambda^* + \Lambda T_k\Lambda^* \\ &= \begin{bmatrix} T_{k-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ g_{k-1} & 0 \end{bmatrix} + \begin{bmatrix} 0 & g_{k-1}^* \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & r_0 - \epsilon \end{bmatrix} \\ &= \begin{bmatrix} T_{k-1} & g_{k-1}^* \\ g_{k-1} & r_0 - \epsilon \end{bmatrix} = T_k - BB^*. \end{aligned}$$

Hence $T_k = AT_kA^* + BB^*$. Therefore part (ii) holds.

Because T_{k+1} is strictly positive, the error $\epsilon > 0$. This readily implies the pair $\{A, B\}$ is controllable. Since T_k strictly positive and $T_k = AT_kA^* + BB^*$, the operator A is stable; see Theorem 9.5.4. This and Proposition 6.5.1 proves part (i).

To verify that part (iii) holds, recall that the autocorrelation function for $y(n)$ is given by $R_y(n) = CA^nQC^*$ for all integers $n \geq 0$ where Q is the controllability Gramian for the pair $\{A, B\}$; see Theorem 4.4.1. Here $T_k = Q$. Now observe that $CT_kC^* = r_0$. If $1 \leq n \leq k$, then we have

$$CA^nT_kC^* = CA^{n-1}(V+\Lambda)T_kC^* = CA^{n-1} \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_k \end{bmatrix} = CV^{n-1} \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_k \end{bmatrix} = r_n \,.$$

The r_k term in the right hand side of the second equality follows from the first equation in (5.17). Hence part (iii) holds.

Parts (iv) is a consequence of Proposition 6.5.1. Part (v) follows from Equation (4.11) in Theorem 4.4.1 in Chapter 4. This completes the proof.

6.5.1 An outer spectral factor representation

Let $\mathbf{g}$ be a scalar valued rational transfer function. Then $\mathbf{g}$ is a *an outer function or a minimum phase function* if $\mathbf{g} = p/q$ where p and q are two polynomials of the same degree, all the poles of $\mathbf{g}$ are contained in the open unit disc $\{z : |z| < 1\}$ and all the zeros of $\mathbf{g}$ are contained in the closed unit disc $\{z : |z| \leq 1\}$. For example, the functions

$$\frac{(z+1)(5z+1)}{(1-2z)(1-4z)} \quad \text{and} \quad \frac{(3z+1)(5z+1)}{(1-2z)(1-4z)} \tag{5.20}$$

are outer functions. However, the functions

$$\frac{3-z}{1-2z}, \quad \frac{z^2+1}{1-2z}, \quad \frac{1-2z}{3-z} \quad \text{and} \quad \frac{1}{1-2z}$$

are not outer. Finally, it is noted that all outer functions are causal; see Lemma 5.3.1.

As before, assume that $\mathbf{g}$ is a rational transfer function. Then we say that $\mathbf{g}$ is an *invertible outer function* if the following two conditions hold:

(i) The function $\mathbf{g}(z) = p(z)/q(z)$ where p and q are polynomials of the same degree.

(ii) All the poles and zeros of $\mathbf{g}$ are contained in the open unit disc $\{z : |z| < 1\}$.

An invertible outer function is an outer function. However, the converse is not necessarily true. For example, the first function in (5.20) is just outer, while the second function in (5.20) is an invertible outer function. By consulting Lemma 5.3.1, we see that $\mathbf{g}$ is an invertible outer function if and only if both $\mathbf{g}$ and $1/\mathbf{g}$ are causal functions. Invertible outer functions play a fundamental role in applications. To see this recall that in the z-domain $\mathbf{y}(z) = \mathbf{g}(z)\mathbf{u}(z)$ where u is the input and y is the output. If $\mathbf{g}$ is causal, then one can compute the present output $y(n)$ from the past and present inputs $\{u(k) : k \leq n\}$. Now consider the problem of computing the input u from the output y. Clearly, $\mathbf{u}(z) = \mathbf{y}(z)/\mathbf{g}(z)$. However, if $1/\mathbf{g}$ is not causal, then one needs the future outputs $\{y(k) : k > n\}$ to compute the present input $u(n)$. This is not practical in many applications. However, if $\mathbf{g}$ is an invertible outer function, then the present output $y(n)$ can be computed from the past and present inputs $\{u(k) : k \leq n\}$, and the present input $u(n)$ can be computed from the past and present outputs $\{y(k) : k \leq n\}$.

Notice that the transfer function in (5.14) in Theorem 6.5.2 is not an outer function. The following result computes an invertible outer function to match the entries $\{r_j\}_0^k$ in the Toeplitz matrix T_{k+1}.

THEOREM 6.5.3 *Let T_{k+1} be a strictly positive Toeplitz matrix on $\mathbb{C}^{k+1}$ generated by a sequence of scalars $\{r_j\}_0^k$. Let $\{\alpha_m\}_{m=1}^k$ and ϵ be the solution to the Levinson system of equations*

$$\begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_k & 1 \end{bmatrix} T_{k+1} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \epsilon \end{bmatrix}. \tag{5.21}$$

Let A on $\mathbb{C}^k$ be the companion matrix, B from $\mathbb{C}$ into $\mathbb{C}^k$ and C from $\mathbb{C}^k$ into $\mathbb{C}$ be defined by

$$\begin{aligned} A &= \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -\alpha_1 & -\alpha_2 & -\alpha_3 & \cdots & -\alpha_{k-1} & -\alpha_k \end{bmatrix} \quad \textit{and} \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \sqrt{\epsilon} \end{bmatrix} \\ C &= \begin{bmatrix} -\alpha_1 & -\alpha_2 & -\alpha_3 & \cdots & -\alpha_{k-1} & -\alpha_k \end{bmatrix} \quad \textit{and} \quad D = \sqrt{\epsilon}\,. \end{aligned} \tag{5.22}$$

Then the following holds.

(i) The operator A is stable. Moreover, the characteristic polynomial for A is

$$d(z) = \det[zI - A] = \alpha_1 + \alpha_2 z + \cdots + \alpha_k z^{k-1} + z^k\,. \tag{5.23}$$

In particular, all the zeros of $d(z)$ are contained inside the open unit disc.

(ii) The Toeplitz matrix T_k on $\mathbb{C}^k$ is the controllability Gramian for the pair $\{A, B\}$.

(iii) Let $y(n)$ be the wide sense stationary process determined by the state space system

$$\begin{aligned} x(n+1) &= Ax(n) + Bu(n) \qquad (x(-\infty) = 0) \\ y(n) &= Cx(n) + Du(n) \end{aligned} \tag{5.24}$$

where $u(n)$ is a white noise random process. Then the autocorrelation function

$$\begin{aligned} R_y(0) &= CT_kC^* + DD^* = r_0 \\ R_y(n) &= CA^{n-1}(BD^* + AT_kC^*) = r_n \qquad (n = 1, 2, \cdots, k)\,. \end{aligned} \tag{5.25}$$

(iv) The transfer function $\mathbf{g}$ for the state space system (5.24) is determined by

$$\mathbf{g} = C(zI - A)^{-1}B + D = \frac{z^k\sqrt{\epsilon}}{\alpha_1 + \alpha_2 z + \cdots + \alpha_k z^{k-1} + z^k} = \frac{z^k\sqrt{\epsilon}}{d(z)}\,. \tag{5.26}$$

Moreover, $\mathbf{g}$ is an invertible outer function.

(v) The spectral density $S_y(\omega)$ for $y(n)$ is given by $S_y(\omega) = |\mathbf{g}(e^{\imath\omega})|^2 = \epsilon/|d(e^{\imath\omega})|^2$.

Proof. Parts (i) and (ii) follow from Theorem 6.5.2. Let $\mathbf{g}$ be the transfer function given by $\mathbf{g} = z^k\sqrt{\epsilon}/d(z)$. Clearly, z^k and $d(z)$ are both polynomials of degree k, and zero is the only root of z^k. Since all the roots of d are contained in the open unit disc, $\mathbf{g}$ is an invertible outer function. Notice that $\mathbf{g}$ admits a decomposition of the form

$$\mathbf{g} = \frac{z^k\sqrt{\epsilon}}{d(z)} = \sqrt{\epsilon} + \frac{-\alpha_1 - \alpha_2 z - \cdots - \alpha_{k-1}z^{k-2} - \alpha_k z^{k-1}}{\alpha_1 + \alpha_2 z + \cdots + \alpha_k z^{k-1} + z^k}\sqrt{\epsilon}\,. \tag{5.27}$$

To verify this simply put the last two terms over the common denominator $d(z)$. By consulting Proposition 6.5.1, we see that $\{A, B, C, D\}$ in (5.22) is a state space realization for $\mathbf{g}$, that is,

$$\mathbf{g}(z) = \frac{z^k\sqrt{\epsilon}}{d(z)} = C(zI - A)^{-1}B + D\,. \tag{5.28}$$

Now consider the wide sense stationary random processes $y(n)$ determined by (5.24) where $u(n)$ is white noise. According to Theorem 5.4.2 with $Q = T_k$, the autocorrelation function for y is given by

$$R_y(0) = CT_kC^* + DD^* \quad \text{and} \quad R_y(n) = CA^{n-1}(BD^* + T_kAC^*) \qquad (n \geq 1)\,. \tag{5.29}$$

Furthermore, the spectral density for y is determined by $S_y(\omega) = |\mathbf{g}(e^{\imath\omega})|^2$. By consulting Equation (5.15), we have

$$\begin{aligned} R_y(n) &= \frac{1}{2\pi}\int_0^{2\pi} e^{\imath\omega n} S_y(\omega)\,d\omega = \frac{1}{2\pi}\int_0^{2\pi} e^{\imath\omega n}|\mathbf{g}(e^{\imath\omega})|^2\,d\omega \\ &= \frac{1}{2\pi}\int_0^{2\pi} \frac{e^{\imath\omega n}\epsilon}{|d(e^{\imath\omega})|^2}\,d\omega = r_n \qquad (\text{for all integers } n \in [-k, k])\,. \end{aligned}$$

Combining this with (5.29) yields (5.25) and completes the proof.

6.5.2 Exercise

Let $S(\omega)$ be any continuous scalar valued function satisfying $S(\omega) \geq \gamma > 0$ for all $0 \leq \omega \leq 2\pi$. Let $\{r_k\}_{-\infty}^{\infty}$ be the inverse Fourier transform for S, that is,

$$r_k = \frac{1}{2\pi}\int_0^{2\pi} e^{\imath\omega k} S(\omega)\,d\omega \qquad (\text{for all interers } k)\,.$$

According to Remark 5.2.2 in Chapter 5, the Toeplitz matrix T_ν corresponding to $\{r_k\}_0^{\nu-1}$ is strictly positive for all integers $\nu \geq 1$. In fact, $T_\nu \geq \gamma I$ for all $\nu \geq 1$. So for large ν one can apply Theorem 6.5.3 to construct a random process $y(n)$ such that $S(\omega) \approx S_y$ where S_y is the spectral density for y. Notice that one disadvantage of Theorem 6.5.3 is that the state space model for y corresponding to T_ν has state dimension $\nu - 1$. In other words, if ν is large, then the state space realization in (5.24) is also large. Here we will show how one can use the Levinson algorithm along with the Kalman-Ho algorithm to compute a reduced order state space system of the form

$$x(n+1) = Ax(n) + Bu(n) \qquad (x(-\infty) = 0) \qquad \text{and} \qquad y(n) = Cx(n) + Du(n)$$

where A is stable, $u(n)$ is a white noise process and $S(\omega) \approx S_y(\omega)$. In other words, if $\mathbf{g} = C(zI - A)^{-1}B + D$ is the transfer function for this system, then $S(\omega) \approx |\mathbf{g}(e^{\imath\omega})|^2$. Finally, we also want the transfer function $\mathbf{g}$ to be an invertible outer function.

In most applications the Fourier coefficients $\{r_k\}_{-\infty}^{\infty}$ for S are real. Notice that if $\{r_k\}_{-\infty}^{\infty}$ are real, then

$$S(\omega) = \sum_{k=-\infty}^{\infty} r_k e^{-\imath\omega k} = r_0 + 2\sum_{k=1}^{\infty} r_k \cos(\omega k)\,.$$

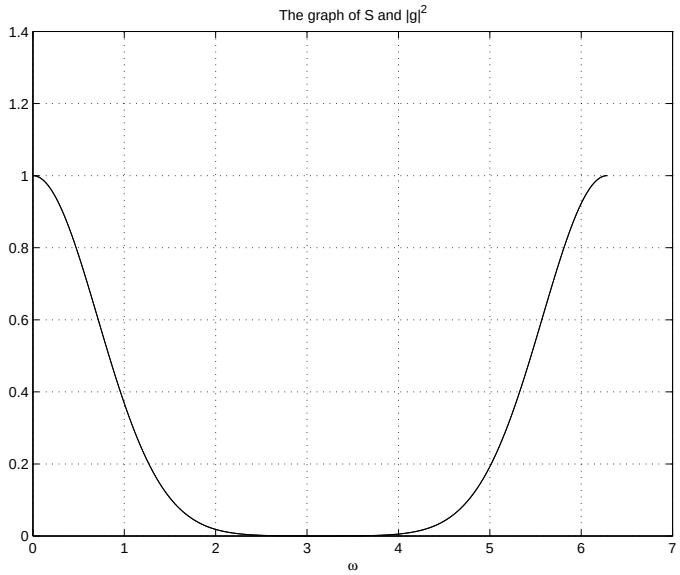

Figure 1

So if $\{r_k\}_{-\infty}^{\infty}$ are real, then $S(\omega)$ is symmetric about π. The converse is also true. In other words, the Fourier coefficients $\{r_k\}_{-\infty}^{\infty}$ are real if and only if $S(\omega)$ symmetric about π.

Now let us present a simple example to demonstrate how one can use the Levinson algorithm along with the Kalman-Ho algorithm, to compute a low order state space model for wide sense stationary process y such that $S(\omega) \approx S_y$. Consider the spectral density defined by

$$\begin{aligned} S(\omega) &= e^{-\omega^2} && \text{if } 0 \leq \omega \leq \pi \\ &= e^{-(\omega-2\pi)^2} && \text{if } \pi < \omega \leq 2\pi\,. \end{aligned} \tag{5.30}$$

The graph of $S(\omega)$ is presented in Figure 1. Notice that S acts like a low pass filter; see Section 5.6 in Chapter 5 for discussion of how sinusoid processes act on linear systems. Moreover, S is symmetric about π. Now let us use the Levinson algorithm to find an invertible outer transfer function $\mathbf{g}$ such that $S(\omega) \approx |\mathbf{g}(e^{\imath\omega})|^2$. In other words, we will find an invertible outer transfer function such that $S(\omega) \approx S_y(\omega)$ where S_y is the spectral density for the wide sense stationary process $y(n) = \sum_{k=-\infty}^{n} g(n-k)u(k)$ and $\{g(k)\}_0^{\infty}$ is the impulse response for $\mathbf{g}$ and u is a white noise process. The following Matlab program was used to compute

the transfer function $\mathbf{g}$:

$$\begin{aligned}
w &= \text{linspace}(0, \pi, 4096); \\
s &= \exp(-w.^2). * (w < \pi) + \exp(-(w - 2\pi).^2). * (w > \pi); \\
&\quad \text{plot}(w, s); \\
r &= \text{real}(\text{ifft}(s)); \\
[al, ep] &= \text{levinson}(r(1:1000)); \\
f &= \text{sqrt}(ep)./\text{fft}(al, 4096); \\
m &= \text{real}(\text{ifft}(f)); \\
[a, b, c, d\,] &= \text{kalho}(m(1:100), 0.0001); \\
[p, q\,] &= \text{ss2tf}(a, b, c, d); \\
g &= \text{fft}(p, 4096)./\text{fft}(q, 4096); \\
&\quad \text{hold on; plot}(w, \text{abs}(g).^2);
\end{aligned}$$

The result of this algorithm is the fourth order transfer function given by

$$\mathbf{g} = \frac{0.1931z^4 + 0.499z^3 + 0.4653z^2 + 0.1863z + 0.0276}{z^4 + 0.584z^3 - 0.258z^2 + 0.0499z - 0.0047}\,. \tag{5.31}$$

Using Matlab is easy to verify that $\mathbf{g}$ is an invertible outer function. Moreover, $\mathbf{g}$ is the transfer function for the following state space system

$$\begin{aligned}
x(n+1) &= \begin{bmatrix} 0.5266 & -0.3389 & 0.0072 & -0.0007 \\ 0.3389 & -0.0293 & -0.1938 & 0.0084 \\ 0.0072 & 0.1938 & -0.1824 & -0.1454 \\ -0.0007 & -0.0084 & -0.1455 & -0.8990 \end{bmatrix} x(n) + \begin{bmatrix} 0.6428 \\ -0.1643 \\ 0.0062 \\ -0.0001 \end{bmatrix} u(n) \\
y(n) &= \begin{bmatrix} 0.6428 & 0.1643 & 0.0062 & -0.0001 \end{bmatrix} x(n) + 0.1931u(n)
\end{aligned}$$

where $u(n)$ is white noise. Here we used the Kalman-Ho algorithm to find a reduced order model $\mathbf{g}$ for $\mathbf{f} = z^{999}\sqrt{\epsilon}/d(z)$ where ϵ and $d(z)$ are defined in Theorem 6.5.3. The graph of both $S(\omega)$ and $|\mathbf{g}(e^{\imath\omega})|^2$ is presented in Figure 1. Notice that one cannot distinguish between $S(\omega)$ and $|\mathbf{g}(e^{\imath\omega})|^2$. Finally, it is noted that the transfer function corresponding to $\mathbf{g}$ has state dimension four, which is clearly less than the 1000 order Toeplitz matrix we inverted to find $z^{999}\sqrt{\epsilon}/d(z)$. Actually, one can use a much smaller Toeplitz matrix in this example. Recall that the errors ϵ_n in the Levinson algorithm are positive and decreasing, that is, $0 \le \epsilon_{n+1} \le \epsilon_n$. Therefore ϵ_n converges to ϵ as n tends to infinity. So in many application one chooses a $n \times n$ Toeplitz matrix where n is the smallest integer such that $\epsilon_n \approx \epsilon$.

The Kalman-Ho algorithm is a method of computing a controllable and observable state space realization $\{A, B, C, D\}$ for a causal transfer function $\mathbf{g}$ from its impulse response $\{g(k)\}_0^\infty$; see [8] for further details. For completeness we included the following simple

Kalman-Ho algorithm in Matlab:

```
function [a, b, c, d, s] = kalho(m, tol)
%  [a, b, c, d, s] = kalho(m, tol)
%  The Kalman-Ho algorithm computes a state space
%  realization from the data m and tol is the tolerance.
%  The default tolerance is 0.000001 and s is the singular
%  values of the Hankel matrix.
%
if  exist('tol'), tol = .000001; end
[x, y] = size(m);
if y > 1.5, m = m'; end
d = m(1); le = length(m) − 1; m = m(2 : le + 1);
n = (le + 1)/2;
h = m(1 : n);
for k = 2 : n; h = [h, m(k : n + k − 1)]; end
[u, s, v] = svd(h);
k = rank(s, tol);
if k == 0
a = 0; b = 0; c = 0;
else
L = u(:, 1 : k) ∗ sqrt(s(1 : k, 1 : k));
r = sqrt(s(1 : k, 1 : k)) ∗ v(:, 1 : k)';
c = L(1, :); b = r(:, 1); L1 = L(1 : n − 1, :); sL = L(2 : n, :);
a = pinv(L1' ∗ L1) ∗ L1' ∗ sL;
s = diag(s); end;
```

Problem 1. Consider the function $S(\omega) = e^{-(\omega-\pi)^2}$ where $0 \leq \omega \leq 2\pi$. Notice that this function acts like a high pass filter. Find a low order state space system of the form

$$x(n+1) = Ax(n) + Bu(n) \qquad (x(-\infty) = 0) \qquad \text{and} \qquad y(n) = Cx(n) + Du(n)$$

where A is stable, $u(n)$ is a white noise random process, $S(\omega) \approx S_y(\omega)$ and the transfer function $\mathbf{g}$ for $\{A, B, C, D\}$ is an invertible outer function. Plot $S(\omega)$ and $S_y(\omega)$ on the same graph. Finally, it is noted that in some problems $S(\omega) \approx 0$ on some interval $[\alpha, \beta]$. So to use the Levinson algorithm to approximate S, one may have to apply the Levinson algorithm to $S + \delta$ where δ is some small positive number.

Problem 2. Consider the wide sense stationary process determined by

$$\xi(n) = \sum_{k=-\infty}^{\infty} h(n-k)v(n)$$

where $v(n)$ is white noise and $\{h(k)\}_{-\infty}^{\infty}$ is the impulse response for the noncausal transfer function given by

$$\mathbf{h} = \frac{10z^2 + 20z + 30}{z^7 - 2z^6 + 3z^5 - 4z^4 + 5z^3 - 6z^2 + 7z + 8}\,.$$

Find a low order state space system of the form

$$x(n+1) = Ax(n) + Bu(n) \qquad (x(-\infty) = 0) \qquad \text{and} \qquad y(n) = Cx(n) + Du(n)$$

where A is stable, $u(n)$ is a white noise random process, $S_\xi(\omega) \approx S_y(\omega)$ and the transfer function $\mathbf{g}$ for $\{A, B, C, D\}$ is an invertible outer function. In other words, find a causal system $y(n) = \sum_{k=-\infty}^{n} g(n-k)u(n)$ such that ξ and y have the same spectral density. Plot S_ξ and S_y on the same graph. Hint, use the fast Fourier transform to find $S_\xi(\omega) = |\mathbf{h}(e^{\imath\omega})|^2$.

Problem 3. Recall that we derived Equation (5.25) in Theorem 6.5.3 by using the spectral density $S_y(\omega) = \epsilon/|d(e^{\imath\omega})|^2$. Give a direct proof of Equation (5.25).

Problem 4. Assume that $\mathbf{g}_o$ is a rational scalar valued function. Recall that $\mathbf{g}_o$ is an *outer function function or a minimum phase function* if $\mathbf{g}_o = p/q$ where p and q are polynomials of the same degree, all the poles of $\mathbf{g}_o$ are contained in the open unit disc $\{z : |z| < 1\}$ and all the zeros of $\mathbf{g}_o$ are contained in the closed unit disc $\{z : |z| \leq 1\}$. Moreover, $\mathbf{g}_i$ is a *rational inner function or a Blaschke product or an all pass transfer function* if $\mathbf{g}_i$ is a proper rational function such that all the poles of $\mathbf{g}_i$ are contained in the open unit disc and $|\mathbf{g}_i(e^{\imath\omega})| = 1$ for all $0 \leq \omega \leq 2\pi$. It is well known that any causal rational transfer function $\mathbf{g}$ admits a unique factorization of the form $\mathbf{g} = \mathbf{g}_i\mathbf{g}_o$ where $\mathbf{g}_o$ is outer and $\mathbf{g}_i$ is inner. By unique we mean that if $\mathbf{g} = \mathbf{h}_i\mathbf{h}_o$ where $\mathbf{h}_o$ is outer and $\mathbf{h}_i$ is inner, then $\mathbf{g}_o = \gamma\mathbf{h}_o$ and $\mathbf{g}_i = \overline{\gamma}\,\mathbf{h}_i$ where γ is a constant of modulus one; see [13, 19, 29] for further details. For example, if $\mathbf{g}(z) = 1/d(z)$ where d is a polynomial of degree k and all the roots of d are contained in the open unit disc, then $\mathbf{g}_o = z^k/d(z)$ and $\mathbf{g}_i = 1/z^k$.

Now consider the causal transfer function given by

$$\mathbf{g} = \frac{z^4 - 5.5833z^3 + 9z^2 - 3.9167z + 0.5}{z^6 + 0.6z^5 - 0.2z^4 - 0.12z^3 + 0.0064z^2 + 0.0038z}$$

Notice that the function $z^k\sqrt{\epsilon}/d(z)$ computed in Theorem 6.5.3 is an invertible outer function. So its reduced order model computed by the Kalman-Ho algorithm is also an invertible outer function. Use the Levinson algorithm to find the inner outer factorization for $\mathbf{g}$. Hint, since $|\mathbf{g}_i(e^{\imath\omega})| = 1$ and $\mathbf{g} = \mathbf{g}_i\mathbf{g}_o$, we must have $|\mathbf{g}(e^{\imath\omega})|^2 = |\mathbf{g}_o(e^{\imath\omega})|^2$. So one can compute $\mathbf{g}_o$ from the Levinson algorithm and Kalman-Ho algorithm. Then $\mathbf{g}_i = \mathbf{g}/\mathbf{g}_o$ follows from an application of the fast Fourier transform with the Kalman-Ho algorithm. Express $\mathbf{g}_i$ and $\mathbf{g}_o$ as rational functions.

Problem 5. Show that $\mathbf{g}_i$ is a rational inner function if and only if $\mathbf{g}_i$ is a proper rational function of the form

$$\mathbf{g}_i(z) = \gamma \prod_{j=1}^{m} \frac{1 - \overline{\alpha}_j z}{z - \alpha_j} \tag{5.32}$$

where $\{\alpha_j\}_j^m$ are scalars contained in the open unit disc and γ is a constant of modulus one. The function $\mathbf{g}_i(z)$ in (5.32) is called a *Blaschke product*. Finally, it is noted that inner functions play an fundamental role in operator theory; see [13, 14, 19, 29].

Chapter 7

Sinusoid estimation

This chapter is concerned with a sinusoid estimation problem, that is, estimating the amplitudes and frequencies of certain sinusoids corrupted by noise.

7.1 Sinusoid processes

In this section we establish some notation concerning the wide sense stationary process $\xi(n)$ in (3.7) presented in Section 4.3. To this end, let $\xi(n)$ be the random process given by

$$\xi(n) = \sum_{k=1}^{\mu} a_k e^{\imath(\omega_k n + \theta_k)} \,. \tag{1.1}$$

Here we assume that the amplitudes $\{a_k\}_1^\mu$ and the frequencies $\{\omega_k\}_1^\mu$ are scalars while the phase $\{\theta_k\}_1^\mu$ are all independent uniform random variables over $[0, 2\pi]$. We call $\xi(n)$ in (1.1) the *standard sinusoid process generated by* the amplitudes and frequencies $\{a_k, \omega_k\}_1^\mu$. Throughout we always assume that the amplitudes $a_k \neq 0$ for all $k = 1, 2, \cdots, \mu$, and the frequencies $\{\omega_k\}_1^\mu$ are all distinct real numbers in $[0, 2\pi)$. According to the discussion in Section 4.3, the process $\xi(n)$ is a mean zero wide sense stationary random process whose autocorrelation function is given by

$$R_\xi(n) = \sum_{k=1}^{\mu} |\, a_k|^2 e^{\imath\omega_k n} \,. \tag{1.2}$$

It is emphasized that in applications the autocorrelation function $R_\xi(n)$ is real valued. This happens because the values on the unit circle in the set $\{e^{\imath\omega_j}\}_1^\mu$ occur in the appropriate complex conjugate pairs; see for example the sinusoid processes in Section 4.1.1.

Now let us present a specific state space representation for $\xi(n)$. Let U be the diagonal

unitary matrix on $\mathbb{C}^\mu$ and C the row operator from $\mathbb{C}^\mu$ into $\mathbb{C}$ defined by

$$U = \begin{bmatrix} e^{\imath\omega_1} & 0 & \cdots & 0 \\ 0 & e^{\imath\omega_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & e^{\imath\omega_\mu} \end{bmatrix} : \mathbb{C}^\mu \to \mathbb{C}^\mu \tag{1.3}$$

$$C = \begin{bmatrix} \overline{a}_1 & \overline{a}_2 & \cdots & \overline{a}_\mu \end{bmatrix} : \mathbb{C}^\mu \to \mathbb{C} \tag{1.4}$$

$$x(0) = \begin{bmatrix} e^{\imath\theta_1} & e^{\imath\theta_2} & \cdots & e^{\imath\theta_\mu} \end{bmatrix}^{tr} : \mathbb{C} \to \mathbb{C}^\mu . \tag{1.5}$$

Because $a_k \neq 0$ for all $k = 1, 2, \cdots, \mu$ and the frequencies $\{\omega_k\}_1^\mu$ are all distinct, the pair $\{C, U\}$ is observable; see Proposition 9.6.4. In other words, $\{C, U\}$ is an observable unitary pair. Moreover, $x(0)$ is a mean zero random vector in $\mathbb{C}^\mu$ satisfying $Ex(0)x(0)^* = I$, and $\xi(n) = CU^n x(0)$. In other words, $\{C, U; x(0)\}$ is a realization of $\xi(n)$. Finally, according to the results in Section 4.2, the autocorrelation $R_\xi(n)$ function is also given by

$$R_\xi(n) = CU^n C^* \qquad \text{(for all integers } n) . \tag{1.6}$$

Let $T_{\xi\nu}$ be the Toeplitz matrix on $\mathbb{C}^\nu$ generated by $R_\xi(n)$, that is,

$$T_{\xi\nu} = \begin{bmatrix} R_\xi(0) & \overline{R}_\xi(1) & \cdots & \overline{R}_\xi(\nu-1) \\ R_\xi(1) & R_\xi(0) & \cdots & \overline{R}_\xi(\nu-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_\xi(\nu-1) & R_\xi(\nu-2) & \cdots & R_\xi(0) \end{bmatrix} . \tag{1.7}$$

Let W be the observability operator from $\mathbb{C}^\mu$ into $\mathbb{C}^\nu$ defined by

$$W = \begin{bmatrix} C \\ CU \\ \vdots \\ CU^{\nu-1} \end{bmatrix} : \mathbb{C}^\mu \to \mathbb{C}^\nu . \tag{1.8}$$

If $\nu \geq \mu$, then the observability of the pair $\{C, U\}$ guarantees that W is one to one. Using $R_\xi(n) = CU^n C^*$ along with the fact that $U^*U = I = UU^*$, we obtain $T_{\xi\nu} = WW^*$. Clearly, $T_{\xi\nu}$ is positive. Furthermore, due to observability $T_{\xi\nu}$ is singular if and only if $\nu > \mu$.

The spectral density for $\xi(n)$. As before, let $\xi(n)$ be the standard sinusoid process generated by the amplitudes and frequencies $\{a_k, \omega_k\}_1^\mu$; see (1.1). Notice that $\{R_\xi(n)\}_{-\infty}^{\infty}$ is not in ℓ^2; see (1.2). Recall that $\delta(\omega)$ is the delta Dirac function. By consulting the results in Section 5.1.1, we see that the spectral density for $\xi(n)$ is given by

$$S_\xi(\omega) = 2\pi \sum_{k=1}^{\mu} |\, a_k|^2 \delta(\omega - \omega_k) . \tag{1.9}$$

7.2 A sinusoid estimation problem

Throughout we use the notation and results established in Section 7.1. As in Section 7.1, let $\xi(n)$ be standard sinusoid process generated by the amplitudes and frequencies $\{a_k, \omega_k\}_1^\mu$, that is,

$$\xi(n) = \sum_{k=1}^{\mu} a_k e^{\imath(\omega_k n + \theta_k)} \,. \tag{2.1}$$

where the amplitudes $\{a_k\}_1^\mu$ are all nonzero, the frequencies $\{\omega_k\}_1^\mu$ are distinct and $\{\theta_k\}_1^\mu$ are independent uniform random variables over $[0, 2\pi]$. Then the sinusoid estimation problem is to find the signal $\xi(n)$ in (2.1) in the presence of noise. In other words, find the amplitudes $\{a_k\}_1^\mu$ and the frequencies $\{\omega_k\}_1^\mu$ for the signal $\xi(n)$ in the presence of noise. To be precise, let $\rho(n)$ be a scalar valued mean zero wide sense stationary noise process, and assume that $\xi(n)$ and $\rho(n)$ are independent random processes. Let $y(n)$ be the wide sense stationary process defined by $y(n) = \xi(n) + \rho(n)$. Then the sinusoid estimation problem is to find the signal process $\xi(n)$ given the output $y(n) = \xi(n) + \rho(n)$, or equivalently, find the amplitudes $\{a_k\}_1^\mu$ and the frequencies $\{\omega_k\}_1^\mu$ for the signal process $\xi(n)$ given $y(n) = \xi(n) + \rho(n)$.

We claim that $y(n) = \xi(n) + \rho(n)$ is a wide sense stationary random process. Moreover, the autocorrelation function $R_y(n)$ for $y(n)$ is given by

$$R_y(n) = R_\xi(n) + R_\rho(n) \,. \tag{2.2}$$

First observe that $Ey(n) = E\xi(n) + E\rho(n) = 0$ for all integers n. Using the fact that $E\xi(n) = 0$ and $\xi(n)$ and $\rho(n)$ are independent, we have

$$\begin{aligned} Ey(n)y(m)^* &= E(\xi(n) + \rho(n))(\xi(m) + \rho(m))^* \\ &= E\xi(n)\xi(m)^* + E\rho(n)\xi(m)^* + E\xi(n)\rho(m)^* + E\rho(n)\rho(m)^* \\ &= R_\xi(n-m) + E\rho(n)E\xi(m)^* + E\xi(n)E\rho(m)^* + R_\rho(n-m) \\ &= R_\xi(n-m) + R_\rho(n-m) \,. \end{aligned}$$

Since $Ey(n)y(m)^*$ is a function of $n - m$, the process $y(n)$ is wide sense stationary and equation (2.2) holds. Finally, it is noted that in applications $R_y(n)$ is a real valued function.

Now let $T_{\rho\nu}$ be the Toeplitz matrix on $\mathbb{C}^\nu$ determined by the noise autocorrelation function $R_\rho(n)$, that is,

$$T_{\rho\nu} = \begin{bmatrix} R_\rho(0) & \overline{R}_\rho(1) & \cdots & \overline{R}_\rho(\nu-1) \\ R_\rho(1) & R_\rho(0) & \cdots & \overline{R}_\rho(\nu-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_\rho(\nu-1) & R_\rho(\nu-2) & \cdots & R_\rho(0) \end{bmatrix} . \tag{2.3}$$

Throughout this chapter we assume that the Toeplitz matrix $T_{\rho\nu}$ for the noise process is invertible for all integers $\nu \geq 1$ and uniformly bounded, that is, there exists a finite positive scalar β such that

$$0 < T_{\rho\nu} \leq \beta I \qquad (\text{for all integers } \nu \geq 1) \,. \tag{2.4}$$

Notice that we do not require $T_{\rho\nu}$ to be uniformly bounded below. In many application $\rho(n) = \gamma u(n)$ where $u(n)$ is a white noise process and γ is a nonzero real number. In this

case, $R_\rho(n) = \gamma^2\delta_{n,0}$ and $T_{\rho\nu} = \gamma^2 I$. (Recall that $\delta_{j,k}$ is the Kronecker delta.) Obviously, condition (2.4) holds in this case.

Let T_ν be the block Toeplitz on $\mathbb{C}^\nu$ formed by $R_y(n)$, that is,

$$T_\nu = \begin{bmatrix} R_y(0) & \overline{R}_y(1) & \cdots & \overline{R}_y(\nu-1) \\ R_y(1) & R_y(0) & \cdots & \overline{R}_y(\nu-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_y(\nu-1) & R_y(\nu-2) & \cdots & R_y(0) \end{bmatrix}. \tag{2.5}$$

By consulting (1.7), (2.2) and (2.3), we have

$$T_\nu = T_{\xi\nu} + T_{\rho\nu} = WW^* + T_{\rho\nu}\,. \tag{2.6}$$

Condition (2.4) implies that $0 < T_{\rho\nu} \le T_{\xi\nu} + T_{\rho\nu} = T_\nu$. Therefore T_ν is invertible for all integers $\nu \ge 1$.

Let $\Phi_\nu(\omega)$ be the row matrix from $\mathbb{C}^\nu$ into $\mathbb{C}$ defined by

$$\Phi_\nu(\omega) = \begin{bmatrix} 1 & e^{-\imath\omega} & e^{-2\imath\omega} & \cdots & e^{-(\nu-1)\imath\omega} \end{bmatrix} \qquad (\omega \in [0, 2\pi])\,. \tag{2.7}$$

For any ω in $[0, 2\pi]$ and integer $\nu \ge 1$, let $M_\nu(\omega)$ be the positive scalar defined by

$$M_\nu(\omega) = \Phi_\nu(\omega) T_\nu^{-1} \Phi_\nu(\omega)^*\,. \tag{2.8}$$

Notice that $M_\nu(\omega) > 0$ for all ω. To see this, recall that T_ν is strictly positive. Hence T_ν^{-1} is also strictly positive. Since $\Phi_\nu(\omega)$ is nonzero, $M_\nu(\omega) = \Phi_\nu(\omega) T_\nu^{-1} \Phi_\nu(\omega)^*$ is strictly positive. This sets the stage for the main result of this section.

THEOREM 7.2.1 *Let $\xi(n)$ be the standard sinusoid process generated by the amplitudes and frequencies $\{a_k, \omega_k\}_1^\mu$. Let $\rho(n)$ be a wide sense stationary noise process independent of $\xi(n)$ whose Toeplitz matrix $T_{\rho\nu}$ satisfies $0 < T_{\rho\nu} \le \beta I$ for all integers $\nu \ge 1$ where β is a finite positive scalar. Finally, let $y(n)$ be the wide sense stationary process defined by $y(n) = \xi(n) + \rho(n)$, and let T_ν be the Toeplitz matrix on $\mathbb{C}^\nu$ corresponding to $R_y(n)$. Then $M_\nu(\omega)^{-1}$ is a decreasing sequence, that is, $M_{\nu+1}(\omega)^{-1} \le M_\nu(\omega)^{-1}$ for all ω in $[0, 2\pi]$. Moreover,*

$$\begin{aligned} \lim_{\nu\to\infty} M_\nu(\omega)^{-1} &= |\,a_j|^2 && \text{if } \omega = \omega_j \text{ for some } j = 1, 2, \cdots, \mu \\ &= 0 && \text{if } \omega \ne \omega_j \text{ for all } j = 1, 2, \cdots, \mu\,. \end{aligned} \tag{2.9}$$

In other words, $M_\nu(\omega)^{-1}$ decreases monotonically to $|\,a_j|^2$ if and only if $\omega = \omega_j$ is a sinusoid frequency of $\xi(n)$, otherwise $M_\nu(\omega)^{-1}$ converges monotonically to zero.

The estimate $M_\nu(\omega)^{-1}$ is known as Capon's maximum likelihood estimator; see [5].

There is an elementary way to implement Theorem 7.2.1 if the ν is not too large. Let Ω be any matrix on $\mathbb{C}^\nu$ such that $\Omega^*\Omega = T_\nu^{-1}$. Then $M_\nu(\omega) = \|\Omega\Phi_\nu(\omega)^*\|^2$. This follows from

$$\begin{aligned} M_\nu(\omega) &= \Phi_\nu(\omega) T_\nu^{-1} \Phi_\nu(\omega)^* = \Phi_\nu(\omega)\Omega^*\Omega\Phi_\nu(\omega)^* \\ &= (\Omega\Phi_\nu(\omega)^*)^*\Omega\Phi_\nu(\omega)^* = \|\Omega\Phi_\nu(\omega)^*\|^2\,. \end{aligned} \tag{2.10}$$

Let $\Omega_k = [\beta_{k,1}, \beta_{k,2}, \cdots, \beta_{k,\nu}]$ be the k-th row of the matrix Ω and set $p_k(\omega) = \Omega_k \Phi_\nu(\omega)^*$. Then

$$p_k(\omega) = \Omega_k \Phi_\nu(\omega)^* = \sum_{m=1}^{\nu} e^{\imath\omega(m-1)} \beta_{k,m}$$

is a polynomial in $e^{\imath\omega}$ of degree at most $\nu - 1$. Notice that

$$M_\nu(\omega) = \|\Omega \Phi_\nu(\omega)^*\|^2 = \sum_{k=1}^{\nu} |\Omega_k \Phi_\nu(\omega)^*|^2 = \sum_{k=1}^{\nu} |p_k(w)|^2 .$$

In other words, $M_\nu(\omega) = \sum_{k=1}^{\nu} |p_k(w)|^2$. It is emphasized that one can use the fast Fourier transform to compute $|p_k(w)|^2$. (In Matlab $|p_k(w)|$ corresponds to abs(fft(Ω_k, 4096)).) Equation (2.9) shows that if ν is large enough, then

$$\left(\sum_{k=1}^{\nu} |p_k(w)|^2 \right)^{-1/2} \begin{array}{ll} \approx & |a_j| \qquad \text{if } \omega = \omega_j \\ \approx & 0 \qquad \text{if } \omega \neq \omega_j \text{ for all } 1 \leq j \leq \mu . \end{array} \tag{2.11}$$

One can use the fast Fourier transform to plot $(\sum_{k=1}^{\nu} |p_k(\omega)|^2)^{-1/2}$ for ω in $[0, 2\pi]$. Then the frequencies where this graph converges to a nonzero value are precisely the frequencies $\{\omega_k\}_1^\mu$ of the signal $\xi(n)$. The amplitude $|a_j|$ is simply the value of the graph at ω_j. Later we will show how one can use the Levinson algorithm to recursively compute a set of polynomials satisfying (2.11).

The proof of Theorem 7.2.1 uses the following result.

LEMMA 7.2.2 *Let $\xi(n)$ be the standard sinusoid process generated by the amplitudes and frequencies $\{a_k, \omega_k\}_1^\mu$, and $\rho(n)$ a wide sense stationary noise process whose block Toeplitz matrix $T_{\rho\nu}$ satisfies $0 < T_{\rho\nu} \leq \beta I$ for all n where β is a finite positive scalar. Let T_ν be the block Toeplitz matrix on $\mathbb{C}^\nu$ formed by the autocorrelation function $R_y(n)$ where $y(n) = \xi(n) + \rho(n)$. Finally, let $\gamma_\nu(\omega)$ be the solution to the optimization problem*

$$\gamma_\nu(\omega) = \inf\{(T_\nu f, f) : \Phi_\nu(\omega) f = 1\} . \tag{2.12}$$

Then $\{\gamma_\nu(\omega)\}_1^\infty$ is a decreasing set of positive numbers, that is, $0 \leq \gamma_{\nu+1}(\omega) \leq \gamma_\nu(\omega)$ for all ω in $[0, 2\pi]$. Furthermore,

$$\gamma_\infty(\omega) = \lim_{\nu\to\infty} \gamma_\nu(\omega) \begin{array}{ll} = & |a_j|^2 \qquad \textit{if } \omega = \omega_j \textit{ for some } j = 1, 2, \cdots, \mu \\ = & 0 \qquad \textit{otherwise} . \end{array} \tag{2.13}$$

PROOF. Let us first verify that $\gamma_\nu(\omega)$ is a positive decreasing sequence. Since T_ν is positive, it follows that $\gamma_\nu(\omega) \geq 0$ for all integers ν. To show that $\gamma_\nu(\omega)$ is decreasing notice that

$$(T_{\nu+1}[f, 0]^{tr}, [f, 0]^{tr}) = (T_\nu f, f)$$

where $[f,0]^{tr}$ is the vector in $\mathbb{C}^{\nu+1}$ whose last component is zero. Using the fact that $\Phi_{\nu+1}(\omega)[f,0]^{tr} = \Phi_\nu(\omega)f$, we obtain

$$\begin{aligned}\gamma_{\nu+1}(\omega) &= \inf\{(T_{\nu+1}g,g) : \Phi_{\nu+1}(\omega)g = 1\} \\ &\le \inf\{(T_{\nu+1}[f,0]^{tr},[f,0]^{tr}) : \Phi_{\nu+1}(\omega)[f,0]^{tr} = 1\} \\ &= \inf\{(T_\nu f,f) : \Phi_\nu(\omega)f = 1\} = \gamma_\nu(\omega)\,.\end{aligned}$$

Therefore $\gamma_{n+1}(\omega) \le \gamma_\nu(\omega)$ and $\{\gamma_\nu(\omega)\}_1^\infty$ is a decreasing sequence of positive numbers. This readily implies that $\gamma_\nu(\omega)$ converges to a positive number $\gamma_\infty(\omega)$ as ν tends to infinity for every ω in $[0,2\pi]$.

Notice that $\Phi_\nu(\omega)\Phi_\nu(\omega)^* = \nu$. Hence $f = \Phi_\nu(\omega)^*/\nu$ satisfies the constraint $\Phi_\nu(\omega)f = 1$. Recall that $T_\nu = T_{\xi\nu} + T_{\rho\nu}$ and $T_{\xi\nu} = WW^*$ where W is the observability matrix from $\mathbb{C}^\mu$ into $\mathbb{C}^\nu$ defined in (1.8). Since $T_{\rho\nu} \le \beta I$, we have

$$\begin{aligned}\gamma_\nu(\omega) &\le (T_\nu f,f) = (T_\nu\Phi_\nu(\omega)^*,\Phi_\nu(\omega)^*)/\nu^2 \\ &= (T_{\xi\nu}f,f) + (T_{\rho\nu}\Phi_\nu(\omega)^*,\Phi_\nu(\omega)^*)/\nu^2 \\ &\le (WW^*f,f) + \beta(\Phi_\nu(\omega)^*,\Phi_\nu(\omega)^*)/\nu^2 \\ &= \|W^*f\|^2 + \beta/\nu = \|W^*\Phi_\nu(\omega)^*\|^2/\nu^2 + \beta/\nu\,.\end{aligned} \tag{2.14}$$

Let us investigate the term $W^*\Phi_\nu(\omega)^*$. First observe that because U^* is a diagonal matrix on $\mathbb{C}^\mu$, we obtain

$$W^* = \begin{bmatrix} C^* & U^*C^* & U^{*2}C^* & \cdots & U^{*\nu-1}C^* \end{bmatrix} = \begin{bmatrix} a_1\Phi_\nu(\omega_1) \\ a_2\Phi_\nu(\omega_2) \\ \vdots \\ a_\mu\Phi_\nu(\omega_\mu) \end{bmatrix}. \tag{2.15}$$

The form of W^* in (2.15) plays an important role in our approach. First observe that $\Phi_\nu(\omega_j)\Phi_\nu(\omega)^* = \sum_{k=0}^{\nu-1} e^{\imath(\omega-\omega_j)k}$. In particular, if $\omega = \omega_j$, then $\Phi_\nu(\omega_j)\Phi_\nu(\omega)^* = \nu$. This proves the first part of the following result

$$\begin{aligned}\Phi_\nu(\omega_j)\Phi_\nu(\omega)^* &= \nu && \text{if } \omega = \omega_j \\ &= (1-e^{\imath(\omega-\omega_j)\nu})/(1-e^{\imath(\omega-\omega_j)}) && \text{if } \omega \ne \omega_j\,.\end{aligned} \tag{2.16}$$

The last equality in (2.16) follows by setting $\lambda = e^{\imath(\omega-\omega_j)}$ in the identity (see Lemma 4.4.2)

$$\sum_{k=0}^{\nu-1}\lambda^k = \frac{1-\lambda^\nu}{1-\lambda} \qquad (\lambda \ne 1)\,. \tag{2.17}$$

If $\omega \ne \omega_j$, then $(1-e^{\imath(\omega-\omega_j)\nu})/(1-e^{\imath(\omega-\omega_j)})$ is uniformly bounded. In fact,

$$\frac{|1-e^{\imath(\omega-\omega_j)\nu}|}{|1-e^{\imath(\omega-\omega_j)}|} \le \frac{2}{|1-e^{\imath(\omega-\omega_j)}|} \qquad (\omega \ne \omega_j)\,. \tag{2.18}$$

Using this in (2.16) readily implies that

$$\begin{aligned} \lim_{\nu\to\infty} a_j\Phi_\nu(\omega_j)\Phi_\nu(\omega)^*/\nu &= a_j \qquad \text{if } \omega=\omega_j \\ &= 0 \qquad \text{if } \omega\neq\omega_j\,. \end{aligned} \tag{2.19}$$

By combining (2.14) and (2.15), we arrive at

$$\gamma_\nu(\omega) \le \beta/\nu + \|W^*\Phi_\nu(\omega)^*/\nu\|^2 = \beta/\nu + \sum_{j=1}^{\mu} |\, a_j\Phi_\nu(\omega_j)\Phi_\nu(\omega)^*/\nu|^2\,. \tag{2.20}$$

If $\omega\neq\omega_j$ for all $j=1,2,\cdots,\mu$, then (2.19) shows that $a_j\Phi_\nu(\omega_j)\Phi_\nu(\omega)^*/\nu$ converges to zero as ν tends to infinity. In this case, the inequality in (2.20) yields

$$\lim_{\nu\to\infty}\gamma_\nu(\omega)=0 \qquad\qquad (\text{if } \omega\neq\omega_j \text{ for all } j=1,2,\cdots,\mu)\,. \tag{2.21}$$

Now assume that $\omega=\omega_j$ for some j. Then using $\Phi_\nu(\omega_j)\Phi_\nu(\omega_j)^*=\nu$, the inequality in (2.20) gives

$$\gamma_\nu(\omega_j) \;\le\; \beta/\nu + |\,a_j|^2 + \sum_{k\neq j}^{\mu} |\,a_k\Phi_\nu(\omega_k)\Phi_\nu(\omega)^*/\nu|^2\,. \tag{2.22}$$

According to (2.19), the terms in the sum converge to zero as ν tends to infinity. This implies that

$$\gamma_\infty(\omega_j) = \lim_{\nu\to\infty}\gamma_\nu(\omega_j) \le |a_j|^2\,. \tag{2.23}$$

Notice that the limit $\gamma_\infty(\omega)$ in (2.23) exists, because $\gamma_\nu(\omega)\ge\gamma_{\nu+1}(\omega)\ge 0$ is a decreasing sequence of positive numbers for each ω. Now let f be any vector in $\mathbb{C}^\nu$ satisfying the constraint $\Phi_\nu(\omega_j)f=1$. Then using $a_j\Phi_\nu(\omega_j)f=a_j$ in (2.15) yields

$$(T_\nu f,f) = (T_{\xi\nu}f,f) + (T_{\rho\nu}f,f) \ge \|W^*f\|^2 = |\,a_j|^2 + \sum_{k\neq j}^{\mu} \|\,a_k^*\Phi_\nu(\omega_k)f\|^2 \ge |\,a_j|^2\,.$$

Therefore

$$|\,a_j|^2 \le \inf\{(T_\nu f,f) : \Phi_\nu(\omega_j)f=1\}\,. \tag{2.24}$$

In other words, $|\,a_j|^2\le\gamma_\nu(\omega_j)$. Combining this with (2.23) implies that

$$\gamma_\infty(\omega_j) = \lim_{\nu\to\infty}\gamma_\nu(\omega_j) = |\,a_j|^2\,. \tag{2.25}$$

This completes the proof.

Proof of Theorem 7.2.1. Recall that $M_\nu(\omega)=\Phi_\nu(\omega)T_\nu^{-1}\Phi_\nu(\omega)^*$. By consulting Lemma 7.2.3 below, the cost $\gamma_\nu(\omega)$ to the optimization problem

$$\gamma_\nu(\omega)=\inf\{(T_\nu f,f) : \Phi_\nu(\omega)f=1\}$$

is given by $\gamma_\nu(\omega) = M_\nu(\omega)^{-1}$. Because $\gamma_\nu(\omega)$ is decreasing, the sequence of scalars $\{M_\nu(\omega)^{-1}\}_1^\infty$ is also decreasing, that is, $M_{\nu+1}(\omega)^{-1} \leq M_\nu(\omega)^{-1}$. Clearly, $M_\nu(\omega)^{-1} = \gamma_\nu(\omega)$ converges to $\gamma_\infty(\omega)$ as ν tends to infinity. Therefore equation (2.9) follows from (2.13). This completes the proof.

It is emphasized that equations (2.16), (2.18), (2.19) and (2.20) can be used to give rates of convergence for $M_\nu(\omega)^{-1}$, that is, $M_\nu(\omega)^{-1}$ converges on the order of $1/\nu$ to its limit in equation (2.9).

LEMMA 7.2.3 *Let T_ν be a strictly positive Toeplitz matrix on $\mathbb{C}^\mu$. Let ω be a fixed frequency in $[0, 2\pi]$. Then $M_\nu(\omega)^{-1} = \gamma_\nu(\omega)$ where*

$$\gamma_\nu(\omega) = \inf\{(T_\nu f, f) : \Phi_\nu(\omega) f = 1\}\,. \tag{2.26}$$

PROOF. Notice that $(T_\nu f, f) = \|T_\nu^{1/2} f\|^2$ where $T_\nu^{1/2}$ is the positive square root of T_ν. By choosing $g = T_\nu^{1/2} f$, we see that the optimization problem in (2.26) is equivalent to

$$\gamma_\nu(\omega) = \inf\{\|g\|^2 : \Phi_\nu(\omega) T_\nu^{-1/2} g = 1\}\,. \tag{2.27}$$

This is a standard least squares optimization problem of the form $\gamma = \inf\{\|g\|^2\}$ subject to the constraint that $Ag = b$ where A is an operator form $\mathcal{X}$ into $\mathcal{Y}$ and b is a vector in $\mathcal{Y}$. If AA^* is invertible, then the optimal solution g_{opt} is given by $g_{opt} = A^*(AA^*)^{-1}b$. Moreover, the optimal cost $\gamma = ((AA^*)^{-1}b, b)$. To see this simply observe that

$$\gamma = \|g_{opt}\|^2 = (A^*(AA^*)^{-1}b, A^*(AA^*)^{-1}b) = ((AA^*)^{-1}AA^*(AA^*)^{-1}b, b) = ((AA^*)^{-1}b, b)\,.$$

Thus $\gamma = ((AA^*)^{-1}b, b)$. In our problem, the operator $A = \Phi_\nu(\omega)T_\nu^{-1/2}$ maps $\mathbb{C}^\nu$ into and $\mathbb{C}$ and $b = 1$. Notice that $AA^* = \Phi_\nu(\omega)T_\nu^{-1}\Phi_\nu(\omega)^* = M_\nu(\omega)$ is a strictly positive number. Obviously, AA^* is invertible. Therefore

$$\gamma_\nu(\omega) = ((AA^*)^{-1}b, b) = (M_\nu(\omega)^{-1}1, 1) = M_\nu(\omega)^{-1}\,.$$

In other words, $\gamma_\nu(\omega) = M_\nu(\omega)^{-1}$. This completes the proof.

7.3 The Levinson algorithm and sinusoid estimation

In this section we will show how the Levinson algorithm plays a fundamental role in Sinusoid estimation. As before, let T_k for $k = 1, 2, \cdots$ be the Toeplitz matrix on $\mathbb{C}^k$ determined by $R_y(n)$; see (2.5). Let $\{\alpha_{k,m}\}_{m=1}^{k-1}$ and ϵ_k be the scalars solving the Levinson system

$$\left[\begin{array}{ccccc} \alpha_{k,1} & \alpha_{k,2} & \cdots & \alpha_{k,k-1} & 1 \end{array}\right] T_k = \left[\begin{array}{ccccc} 0 & 0 & \cdots & 0 & \epsilon_k \end{array}\right] \qquad (k = 1, 2, \cdots)\,. \tag{3.1}$$

If $k = 1$, then $\epsilon_1 = T_1 = R_y(0)$. It is emphasized that $\{\alpha_{k,m}\}_{m=1}^{k-1}$ and ϵ_k can be recursively computed by the Levinson algorithm; see Theorem 6.3.1. Let $\{\beta_{k,m}\}_{m=1}^{k}$ be the normalized Levinson coefficients defined by

$$\left[\begin{array}{ccccc} \beta_{k,1} & \beta_{k,2} & \cdots & \beta_{k,k} & \beta_{k,k} \end{array}\right] = \left[\begin{array}{ccccc} \alpha_{k,1} & \alpha_{k,2} & \cdots & \alpha_{k,k-1} & 1 \end{array}\right] / \sqrt{\epsilon_k} \qquad (k \geq 1)\,. \tag{3.2}$$

Here $\beta_{1,1} = \epsilon_1^{-1/2} = R_y(0)^{-1/2}$. Let $\varphi_k(\omega)$ be the polynomial in $e^{\imath\omega}$ defined by

$$\varphi_k(\omega) = \sum_{m=1}^{k} e^{\imath\omega(m-1)}\beta_{k,m}\,. \tag{3.3}$$

The polynomials $\{\varphi_k(\omega)\}_1^\infty$ are called the *normalized Levinson polynomials* for $R_y(n)$. Notice that the normalized Levinson polynomials can be recursively computed by applying the Levinson algorithm to $R_y(n)$. Because T_ν is strictly positive, $\beta_{k,k} = \epsilon_k^{-1/2}$ is strictly positive. So the degree of the k-th normalized Levinson polynomial $\varphi_k(\omega)$ is $k-1$. The following result uses the normalized Levinson polynomials to compute the amplitudes and frequencies for a wide sense stationary random process.

THEOREM 7.3.1 *Let $\xi(n)$ be the standard sinusoid process generated by the amplitudes and frequencies $\{a_k, \omega_k\}_1^\mu$. Let $\rho(n)$ be a wide sense stationary process independent to $\xi(n)$ whose Toeplitz matrix $T_{\rho\nu}$ satisfies $0 < T_{\rho\nu} \le \beta I$ for all integers $\nu \ge 1$ where β is a finite positive constant. Let $y(n)$ be the wide sense stationary process defined by $y(n) = \xi(n) + \rho(n)$. Finally, let $\{\varphi_k(\omega)\}_1^\infty$ be the normalized Levinson polynomials in (3.3) recursively computed from $\{R_y(n)\}_0^\infty$. Then $M_\nu(\omega) = \sum_{k=1}^{\nu} |\varphi_k(\omega)|^2$. Moreover,*

$$\begin{aligned}\left(\sum_{k=1}^{\infty} |\varphi_k(\omega)|^2\right)^{-1/2} &= |\,a_j| \qquad \text{if } \omega = \omega_j \text{ for some } j = 1, 2, \cdots, \mu \\ &= 0 \qquad \text{otherwise}\,.\end{aligned} \tag{3.4}$$

The Levinson algorithm along with the fast Fourier transform provides us with a numerically efficient method to compute the amplitudes $\{a_k\}_1^\mu$ and frequencies $\{\omega_k\}_1^\mu$ of $\xi(n)$ from $y(n)$. To see this simply use the Levinson algorithm along with the fast Fourier transform to plot

$$\left(\sum_{k=1}^{\nu} |\varphi_k(\omega)|^2\right)^{-1/2} \qquad (\omega \in [0, 2\pi])\,. \tag{3.5}$$

Then the frequencies where this graph converges to a nonzero value are precisely the frequencies $\{\omega_k\}_1^\mu$ of the signal $\xi(n)$. The amplitude $|\,a_j|$ is simply the value of the graph at ω_j. Finally, it is noted that equation (3.4) is due to L. Ya. Geronimus.

By taking the inverse in (3.4), we have

$$\begin{aligned}\sum_{k=1}^{\infty} |\varphi_k(\omega)|^2 &= \frac{1}{|a_j|^2} \qquad \text{if } \omega = \omega_j \text{ for some } j \\ &= \infty \qquad \text{otherwise}\,.\end{aligned} \tag{3.6}$$

If ω_j is a signal frequency, then the sum in (3.6) is finite. Hence $\lim_{\nu\to\infty} |\varphi_\nu(\omega_j)|^2 = 0$. By taking the inverse, we see that

$$\lim_{\nu\to\infty} \frac{1}{|\varphi_\nu(\omega_j)|^2} = +\infty\,. \tag{3.7}$$

In other words, at the signal frequencies, the normalized Levinson polynomials $1/|\varphi_\nu(\omega_j)|^2$ diverge to infinity as ν tends to ∞. Some researchers use this result to find the signal frequencies.

Proof of Theorem 7.3.1. Let Ω be the lower triangular matrix on $\mathbb{C}^\nu$ defined by

$$\Omega = \begin{bmatrix} \beta_{1,1} & 0 & 0 & \cdots & 0 \\ \beta_{2,1} & \beta_{2,2} & 0 & \cdots & 0 \\ \beta_{3,1} & \beta_{3,2} & \beta_{3,3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{\nu,1} & \beta_{\nu,2} & \beta_{\nu,3} & \cdots & \beta_{\nu,\nu} \end{bmatrix}. \tag{3.8}$$

Corollary 6.4.3 shows that $T_\nu^{-1} = \Omega^*\Omega$. Notice that $M_\nu(\omega)^{-1} = \|\Omega\Phi_\nu(\omega)^*\|^2$. This follows from

$$M_\nu(\omega) \;=\; \Phi_\nu(\omega)T_\nu^{-1}\Phi_\nu(\omega)^* = \Phi_\nu(\omega)\Omega^*\Omega\Phi_\nu(\omega)^* = \|\Omega\Phi_\nu(\omega)^*\|^2 .$$

The k-th row Ω_k of Ω is given by

$$\Omega_k = \begin{bmatrix} \beta_{k,1} & \beta_{k,2} & \cdots & \beta_{k,k} & 0 & 0 & \cdots & 0 \end{bmatrix}. \tag{3.9}$$

So the k-th row of $\Omega\Phi_\nu(\omega)^*$ is given by the k-th normalized Levinson polynomial, that is,

$$\varphi_k(\omega) = \Omega_k\Phi_\nu(\omega)^* = \sum_{m=1}^{k} e^{\imath\omega(m-1)}\beta_{k,m} .$$

This readily implies that

$$M_\nu(\omega) = \|\Omega\Phi_\nu(\omega)^*\|^2 = \sum_{k=1}^{\nu} |\Omega_k\Phi_\nu(\omega)^*|^2 = \sum_{k=1}^{\nu} |\varphi_k(\omega)|^2 . \tag{3.10}$$

In other words, $M_\nu(\omega) = \sum_{k=1}^{\nu} |\varphi_k(\omega)|^2$. By passing limits and employing equation (2.9) in Theorem 7.2.1, we obtain (3.4). This completes the proof.

7.3.1 Exercise

Problem 1. Assume that the hypothesis of Theorem 7.3.1 holds. As before, let $T_\nu = T_{\xi\nu} + T_{\rho\nu}$. Then show that

$$\begin{aligned} \lim_{\nu\to\infty} \frac{\Phi_\nu(\omega)T_\nu\Phi_\nu(\omega)^*}{\nu^2} &= |\, a_j|^2 \qquad \text{if } \omega = \omega_j \text{ for some } j \\ &= 0 \qquad \text{otherwise} . \end{aligned}$$

Show also that $\Phi_\nu(\omega)T_\nu\Phi_\nu(\omega)^*$ is a weighted Fourier transform, that is,

$$\begin{aligned} \Phi_\nu(\omega)T_\nu\Phi_\nu(\omega)^* \;=\;& \nu R_y(0) + (\nu-1)\left(R_y(1)e^{-\imath\omega} + R_y(1)^* e^{\imath\omega}\right) \\ &+(\nu-2)\left(R_y(2)e^{-2\imath\omega} + R_y(2)^* e^{2\imath\omega}\right) + \cdots \\ &+R_y(\nu-1)e^{-\imath(\nu-1)\omega} + R_y(\nu-1)^* e^{\imath(\nu-1)\omega} . \end{aligned}$$

Chapter 8

Positive and singular Toeplitz matrices

This chapter is devoted to studying Toeplitz matrices which are positive and singular. This will lead to a solution for a sinusoid estimation problem in white noise.

8.1 A unitary state space model

In this section we will present a unitary state space model for a positive and singular Toeplitz matrix. In Particular, we show that a positive and singular Toeplitz matrix can be represented by a pure sinusoid process. As before, let T_ν be the Toeplitz matrix on $\mathbb{C}^\nu$ generated by a set of scalars $\{r_j\}_0^{\nu-1}$, that is,

$$T_\nu = \begin{bmatrix} r_0 & \overline{r}_1 & \cdots & \overline{r}_{\nu-1} \\ r_1 & r_0 & \cdots & \overline{r}_{\nu-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{\nu-1} & r_{\nu-2} & \cdots & r_0 \end{bmatrix} . \tag{1.1}$$

We say that $\{C, U \text{ on } \mathcal{X}\}$ is a *unitary* pair if U is a unitary operator on $\mathcal{X}$ and C is an operator mapping $\mathcal{X}$ into $\mathbb{C}$. The pair $\{C, U\}$ is a *unitary realization* of $\{r_j\}_0^{\nu-1}$ if U is a unitary operator and

$$CU^kC^* = r_k \qquad (k = 0, 1, 2, \cdots, \nu - 1) . \tag{1.2}$$

Two pairs $\{C, U \text{ on } \mathcal{X}\}$ and $\{E, V \text{ on } \mathcal{V}\}$ are *isomorphic* if there exists a unitary operator Φ mapping $\mathcal{V}$ onto $\mathcal{X}$ satisfying $\Phi V = U\Phi$ and $E = C\Phi$. If $\{C, U \text{ on } \mathcal{X}\}$ and $\{E, V \text{ on } \mathcal{V}\}$ are isomorphic, then $CU^kC^* = EV^kE^*$ for all integers $k \geq 0$. In particular, two unitary isomorphic pairs realize the same sequence. The following result shows that a positive singular Toeplitz matrix is uniquely determined an observable unitary pair.

THEOREM 8.1.1 *Let T_ν be a Toeplitz matrix on $\mathbb{C}^\nu$ generated by a set of scalars $\{r_j\}_0^{\nu-1}$. Then T_ν is positive and singular if and only if there exists an observable unitary realization $\{C, U \text{ on } \mathcal{X}\}$ for $\{r_j\}_0^{\nu-1}$ satisfying $\dim \mathcal{X} < \nu$. In this case, all observable unitary realizations of $\{r_j\}_0^{\nu-1}$ are isomorphic.*

PROOF. Assume that $\{C, U \text{ on } \mathcal{X}\}$ is a unitary realization of $\{r_j\}_0^{\nu-1}$ satisfying $\dim \mathcal{X} < \nu$. Let W_ν mapping $\mathcal{X}$ into $\mathbb{C}^\nu$ be the observability operator defined by

$$W_\nu = \begin{bmatrix} C \\ CU \\ \vdots \\ CU^{\nu-1} \end{bmatrix} : \mathcal{X} \to \mathbb{C}^\nu . \tag{1.3}$$

Then using $CU^kC^* = r_k$ and $U^*U = I$, it follows that $T_\nu = W_\nu W_\nu^*$. In fact, $T_j = W_j W_j^*$ for all $j = 1, 2, \cdots, \nu$. (Here T_j is the Toeplitz matrix on $\mathbb{C}^j$ generated by $\{r_j\}_0^{j-1}$.) In particular, T_ν is positive. If g is in $\mathbb{C}^\nu$, then $(T_\nu g, g) = (W_\nu W_\nu^* g, g) = \|W_\nu^* g\|^2 \geq 0$. Since $\dim \mathcal{X} < \nu$, we have $\text{rank} W_\nu < \nu$. Hence $T_\nu = W_\nu W_\nu^*$ is positive and singular.

Now assume that the Toeplitz matrix T_ν is positive and singular. Let Z be the lower shift on $\mathbb{C}^\nu$ and Γ the column matrix defined by

$$Z = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \text{ on } \mathbb{C}^\nu \quad \text{and} \quad \Gamma = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \mathbb{C} \to \mathbb{C}^\nu . \tag{1.4}$$

In the definition of Z, one's appear immediately below the main diagonal and zero's elsewhere. Notice that if $T_\nu = W_\nu W_\nu^*$, then $W_\nu = ZW_\nu U + \Gamma C$. Because the pair $\{C, U\}$ is observable, W_ν is one to one. So if $M = W_\nu$, then T_ν admits a factorization of the form $T_\nu = MM^*$ and $M = ZMU + \Gamma C$.

Recall that T_ν is a positive and singular Toeplitz matrix. Motivated by the previous paragraph, let M be any one to one operator from a space $\mathcal{X}$ into $\mathbb{C}^\nu$ satisfying $T_\nu = MM^*$. We claim that there is a unique solution $[U, C]^{tr}$ mapping $\mathcal{X}$ into $\mathcal{X} \oplus \mathbb{C}$ satisfying

$$M = \begin{bmatrix} ZM & \Gamma \end{bmatrix} \begin{bmatrix} U \\ C \end{bmatrix} . \tag{1.5}$$

Here U is an operator on $\mathcal{X}$ while C maps $\mathcal{X}$ into $\mathbb{C}$. Let L be the operator from $\mathcal{X} \oplus \mathbb{C}$ into $\mathbb{C}^\nu$ defined by

$$L = \begin{bmatrix} ZM & \Gamma \end{bmatrix} : \begin{bmatrix} \mathcal{X} \\ \mathbb{C} \end{bmatrix} \to \mathbb{C}^\nu . \tag{1.6}$$

We claim that $\text{ran}\, M \subset \text{ran}\, L$. Hence there exists a solution to (1.5).

To show $\text{ran}\, M \subset \text{ran}\, L$, let R be the column operator defined by

$$R = \begin{bmatrix} r_0/2 & r_1 & r_2 & \cdots & r_{\nu-1} \end{bmatrix}^{tr} .$$

Then using the form of the Toeplitz matrix, it follows that

$$T_\nu = ZT_\nu Z^* + \Gamma R^* + R\Gamma^* . \tag{1.7}$$

Obviously, L^* is given by

$$L^* = \begin{bmatrix} M^*Z^* \\ \Gamma^* \end{bmatrix} : \mathcal{X} \to \begin{bmatrix} \mathcal{X} \\ \mathbb{C} \end{bmatrix} . \tag{1.8}$$

We claim that $\ker L^* \subset \ker M^*$. (The kernel or null space of an operator is denoted by ker.) Let x be any vector in the kernel of L^*. Then $\Gamma^* x$ and M^*Z^*x are both zero. Using the Lyapunov equation in (1.7), we obtain

$$\|M^*x\|^2 = (MM^*x, x) = (T_\nu x, x) = (ZT_\nu Z^*x, x) + 2\Re(\Gamma^*x, R^*x) = \|M^*Z^*x\|^2 = 0 .$$

Hence x is in the kernel of M^*. In other words, $\ker L^* \subset \ker M^*$. By taking the orthogonal complement $(\ker M^*)^\perp \subset (\ker L^*)^\perp$. Recall that if J is any operator, then $\operatorname{ran} J = (\ker J^*)^\perp$. Using this we have

$$\operatorname{ran} M = (\ker M^*)^\perp \subset (\ker L^*)^\perp = \operatorname{ran} L . \tag{1.9}$$

Therefore $\operatorname{ran} M \subset \operatorname{ran} L$.

Because $\operatorname{ran} M \subset \operatorname{ran} L$, there exists an operator $[U, C]^{tr}$ mapping $\mathcal{X}$ into $\mathcal{X} \oplus \mathbb{C}$ satisfying

$$M = \begin{bmatrix} ZM & \Gamma \end{bmatrix} \begin{bmatrix} U \\ C \end{bmatrix} = ZMU + \Gamma C . \tag{1.10}$$

By recursively substituting $\Gamma C + ZMU$ for M, we obtain

$$\begin{aligned} M &= \Gamma C + ZMU = \Gamma C + Z\Gamma CU + Z^2MU^2 \\ &= \Gamma C + Z\Gamma CU + Z^2\Gamma CU^2 + Z^3MU^3 = \cdots = \sum_{k=0}^{\nu-1} Z^k \Gamma C U^k . \end{aligned}$$

The last equality follows from the fact that $Z^\nu = 0$. Using the form of Z and Γ in (1.4), it follows that

$$M = \sum_{k=0}^{\nu-1} Z^k \Gamma C U^k = I_\nu \begin{bmatrix} C \\ CU \\ \vdots \\ CU^{\nu-1} \end{bmatrix}$$

where I_ν is the identity operator on $\mathbb{C}^\nu$. Hence $M = W_\nu$ where W_ν is the observability operator defined in (1.3). As before, let μ be the dimension of $\mathcal{X}$. By construction M is one to one. So W_ν is also one to one and $\operatorname{rank} W_\nu = \mu$. Since U is an operator on $\mathcal{X}$ the Cayley-Hamilton shows that $\operatorname{rank} W_\mu = \operatorname{rank} W_k$ for all $k = \mu, \mu+1, \cdots, \nu$. Thus $\operatorname{rank} W_\mu = \mu$ and W_μ is invertible. In particular, the pair $\{C, U\}$ is observable.

Let M_k be the operator mapping $\mathcal{X}$ into $\mathbb{C}$ determined by the k-th row of M. Then using $M = W_\nu$, we arrive at

$$\begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_\nu \end{bmatrix} = M = W_\nu = \begin{bmatrix} C \\ CU \\ \vdots \\ CU^{\nu-1} \end{bmatrix} .$$

In other words, $M_k = CU^{k-1}$ for $k = 1, 2, \cdots, \nu$. In particular, the operator $C = M_1$ is uniquely determined by the first row of M. Notice that

$$\begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_\mu \end{bmatrix} U = W_\mu U = \begin{bmatrix} M_2 \\ M_3 \\ \vdots \\ M_{\mu+1} \end{bmatrix}. \tag{1.11}$$

Because the matrix $[M_1, M_2, \cdots, M_\mu]^{tr} = W_\mu$ formed by the first μ rows of M is invertible, this shows that U is uniquely determined. In fact,

$$U = \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_\mu \end{bmatrix}^{-1} \begin{bmatrix} M_2 \\ M_3 \\ \vdots \\ M_{\mu+1} \end{bmatrix}.$$

Therefore there is a unique solution $[U, C]^{tr}$ to the equation $M = L[U, C]^{tr}$. In particular, the operator L in (1.6) is one to one. Moreover, the pair $\{C, U\}$ is observable.

Let us show that U is unitary. Since L is one to one its adjoint L^* is onto $\mathcal{X} \oplus \mathbb{C}$. So there exists a μ dimensional subspace $\mathcal{G}$ such that $L^*\mathcal{G} = \mathcal{X} \oplus \{0\}$. If g is in $\mathcal{G}$, then $\Gamma^* g = 0$; see (1.8). Thus $M^* g = [U^*, C^*] L^* g$ yields $M^* g = U^* M^* Z^* g$. Using this and the Lyapunov equation in (1.7), we obtain

$$\begin{aligned} \|U^* M^* Z^* g\|^2 &= \|M^* g\|^2 = (T_\nu g, g) = (Z T_\nu Z^* g, g) + 2\Re(\Gamma^* g, R^* g) \\ &= (T_\nu Z^* g, Z^* g) = \|M^* Z^* g\|^2 . \end{aligned}$$

Hence $\|U^* M^* Z^* g\| = \|M^* Z^* g\|$ for all g in $\mathcal{G}$. Because $M Z^* \mathcal{G} = \mathcal{X}$, we have $\|U^* x\| = \|x\|$ for all x in $\mathcal{X}$. In other words, U^* is unitary. So U is also unitary. Using $T_\nu = MM^* = W_\nu W_\nu^*$, it follows that (1.2) holds, that is, $\{C, U\}$ is an observable unitary realization of $\{r_k\}_0^{\nu-1}$.

Theorem 4.2.1 shows that all observable unitary realizations of $\{r_k\}_0^{\nu-1}$ are isomorphic. This completes the proof.

COROLLARY 8.1.2 *Let T_ν be a singular and positive Toeplitz matrix on $\mathbb{C}^\nu$ generated by a set of scalars $\{r_j\}_0^{\nu-1}$. Then an observable unitary realization $\{C, U\}$ of $\{r_j\}_0^{\nu-1}$ is given by*

$$\begin{bmatrix} U \\ C \end{bmatrix} = \begin{bmatrix} M^* Z^* Z M & M^* Z^* \Gamma \\ \Gamma^* Z M & \Gamma^* \Gamma \end{bmatrix}^{-1} \begin{bmatrix} M^* Z^* M \\ \Gamma^* M \end{bmatrix}. \tag{1.12}$$

Here M is any one to one operator from $\mathcal{X}$ into $\mathbb{C}^\nu$ satisfying $T_\nu = MM^*$, while Z on $\mathbb{C}^\nu$ is the lower shift and Γ is the column vector in $\mathbb{C}^\nu$ defined in (1.4).

Proof. By consulting the proof of Theorem 8.1.1, we see that an observable unitary realization $\{C, U\}$ of $\{r_j\}_0^{\nu-1}$ is given by the unique solution to the equation $M = L[U, C]^{tr}$ where $L = [ZM, \Gamma]$. Multiplying by L^* on the left, gives $L^* M = L^* L[U, C]^{tr}$. Since L is one to one, $L^* L$ is invertible. Hence $[U, C]^{tr} = (L^* L)^{-1} L^* M$. This readily yields the formula for $[U, C]^{tr}$ in (1.12) and completes the proof.

8.2 Sinusoid processes and singular Toeplitz matrices.

In this section we will develop a sinusoid wide sense stationary model for a positive and singular Toeplitz matrix. As before, let T_ν on $\mathbb{C}^\nu$ be a positive and singular Toeplitz matrix generated by a set of scalars $\{r_j\}_0^{\nu-1}$. Let $\{C, U \text{ on } \mathcal{X}\}$ be an observable unitary realization for $\{r_j\}_0^{\nu-1}$ and μ the dimension of $\mathcal{X}$. Because U is unitary there exists a unitary operator Φ mapping $\mathbb{C}^\mu$ onto $\mathcal{X}$ satisfying $U\Phi = \Phi V$ where V on $\mathbb{C}^\mu$ is the diagonal matrix formed by the eigenvalues of U. Moreover, if E is the operator from $\mathbb{C}^\mu$ into $\mathbb{C}$ defined by $E = C\Phi$, then the pair $\{C, U\}$ is isomorphic to $\{E, V\}$. In particular, $\{E, V\}$ is an observable unitary realization for $\{r_j\}_0^{\nu-1}$. Clearly, V and E admit matrix representations of the form

$$\begin{aligned} V &= \begin{bmatrix} e^{\imath\omega_1} & 0 & \cdots & 0 \\ 0 & e^{\imath\omega_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\imath\omega_\mu} \end{bmatrix} \text{ on } \mathbb{C}^\mu \\ E &= \begin{bmatrix} a_1 & a_2 & \cdots & a_\mu \end{bmatrix} : \mathbb{C}^\mu \to \mathbb{C}\,. \end{aligned} \tag{2.1}$$

Here $\{e^{\imath\omega_k}\}_1^\mu$ are the eigenvalues of U. Because the pair $\{E, V\}$ is observable all the eigenvalues of U are distinct and $a_j \neq 0$ for all $j = 1, 2, \cdots, \mu$; see Proposition 9.6.4. Since $r_k = EV^kE^*$ for $k = 0, 1, \cdots, \nu-1$, we obtain the following expression for $\{r_j\}_0^{\nu-1}$

$$r_n = \sum_{k=1}^{\mu} |\, a_k|^2 e^{\imath\omega_k n} \qquad (n = 0, 1, \cdots, \nu-1)\,. \tag{2.2}$$

Motivated by this we call $\{|\, a_k|, \omega_k\}_1^\mu$ the *amplitudes and frequencies associated* with $\{r_k\}_1^\nu$ or the Toeplitz matrix T_ν. So given any positive and singular Toeplitz matrix T_ν, we can compute the amplitudes and frequencies associated with T_ν. Because all observable unitary realization of $\{r_k\}_1^\nu$ are isomorphic, the amplitudes and frequencies $\{|\, a_k|, \omega_k\}_1^\mu$ associated with the Toeplitz matrix T_ν are also unique; see Section 4.3.

Now consider the standard sinusoid process $\xi(n)$ generated by the amplitudes and frequencies $\{a_k, \omega_k\}_1^\mu$, that is,

$$\xi(n) = \sum_{k=1}^{\mu} a_k e^{\imath(\omega_k n + \theta_k)} \tag{2.3}$$

where the phase $\{\theta_k\}_1^\mu$ are all independent uniform random variables over $[0, 2\pi]$. By consulting the results in Section 7.1, it follows that $\xi(n)$ is a mean zero wide sense stationary random process whose autocorrelation function is given by

$$R_\xi(n) = \sum_{k=1}^{\nu} |\, a_k|^2 e^{\imath\omega_k n}\,.$$

So $R_\xi(n) = r_n$ for $n = 0, 1, \cdots, \nu-1$. Therefore the wide sense stationary random process $\xi(n)$ can be used as a model for matching the entries $\{r_j\}_0^{\nu-1}$ in a positive and singular Toeplitz matrix T_ν. Motivated by this, we call $\xi(n)$ in (2.3) the *sinusoid process associated* with T_ν or $\{r_k\}_0^{\nu-1}$.

8.3 Unique positive Toeplitz expansions

In this section we will show that a positive and singular Toeplitz matrix has only one positive Toeplitz expansion. Let S be an operator on $\mathcal{M}$ and $\mathcal{H}$ a subspace of $\mathcal{M}$. Then we say that T on $\mathcal{H}$ is the *compression* of S to $\mathcal{H}$ if T is the operator on $\mathcal{H}$ defined by $T = P_{\mathcal{H}} S|\mathcal{H}$ where $P_{\mathcal{H}}$ is the orthogonal projection onto $\mathcal{H}$ and the $|\mathcal{H}$ means that the operator is restricted to $\mathcal{H}$. Notice that T is the compression of S to $\mathcal{H}$ if and only if S admits an operator matrix representation of the form

$$S = \begin{bmatrix} T & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \text{ on } \begin{bmatrix} \mathcal{H} \\ \mathcal{H}^{\perp} \end{bmatrix}. \tag{3.1}$$

As before, let T be an operator on $\mathcal{H}$. Then we say that an operator S on $\mathcal{M}$ is an *expansion* of T if $\mathcal{H} \subset \mathcal{M}$ and T is the compression of S to $\mathcal{H}$. In other words, S is an expansion of T if S admits an operator matrix representation of the form (3.1).

Now let T_ν be a Toeplitz matrix on $\mathbb{C}^\nu$ generated by $\{r_j\}_0^{\nu-1}$, and S_m be the Toeplitz matrix on $\mathbb{C}^m$ generated by $\{b_j\}_0^{m-1}$. If $\nu \leq m$, then we say that T_ν is the compression of S_m to $\mathbb{C}^\nu$ if T_ν is the $\nu \times \nu$ matrix contained in the upper right corner of S_m, that is, S_m admits a matrix representation of the form

$$S_m = \begin{bmatrix} T_\nu & * \\ * & * \end{bmatrix} \text{ on } \begin{bmatrix} \mathbb{C}^\nu \\ \mathbb{C}^{m-\nu} \end{bmatrix}. \tag{3.2}$$

Notice that T_ν is the compression of S_m to $\mathbb{C}^\nu$ if and only if $r_j = b_j$ for $j = 0, 1, \cdots, \nu - 1$. As expected, we say that S_m is an expansion of T_ν if S_m admits a matrix representation of the form (3.2). In other words, S_m is an expansion of T_ν if and only if $r_j = b_j$ for $j = 0, 1, \cdots, \nu - 1$. The following result shows that if T_ν is positive and singular, then there is only one positive Toeplitz expansion on $\mathbb{C}^m$ of T_ν.

THEOREM 8.3.1 *Let T_ν on $\mathbb{C}^\nu$ be a positive and singular Toeplitz matrix generated by a sequence of scalars $\{r_j\}_0^{\nu-1}$ and m any integer satisfying $m \geq \nu$. Then there is only one positive Toeplitz matrix T_m on $\mathbb{C}^m$ expanding T_ν. Moreover, if $\{C, U\}$ is an observable unitary realization for $\{r_j\}_0^{\nu-1}$, then this positive expansion T_m of T_ν is the Toeplitz matrix on $\mathbb{C}^m$ generated by*

$$r_j = CU^j C^* \qquad (j = 0, 1, \cdots, m-1). \tag{3.3}$$

In other words, the Toeplitz matrix T_m on $\mathbb{C}^m$ generated by $\{CU^jC^*\}_{j=0}^{m-1}$ is the only positive Toeplitz matrix on $\mathbb{C}^m$ expanding T_ν.

PROOF. If T_m is the Toeplitz on $\mathbb{C}^m$ generated by (3.3), then clearly, T_m is a Toeplitz extension of T_ν. Theorem 8.1.1 guarantees that this Toeplitz extension T_m is positive and singular. Now let us show that this is the only positive Toeplitz extension of T_ν. Let T_m on $\mathbb{C}^m$ be any positive Toeplitz extension of T_ν. Then T_m must be singular. To see this assume that T_m is strictly positive. Since T_ν is the compression of T_m to $\mathbb{C}^\nu$, it follows that T_ν is also strictly positive. This contradicts the fact that T_ν is positive and singular. Thus T_m must be singular. Because T_m is singular, there exist an observable unitary realization

$\{E, V\}$ for $\{b_j\}_0^{m-1}$ where $\{b_j\}_0^{m-1}$ are the unique scalars which generate T_m. Since T_m is an extension of T_ν, we have $b_j = r_j$ for $j = 1, 2, \cdots, \nu-1$. In particular, $\{E, V\}$ is an observable unitary realization for $\{r_j\}_0^{\nu-1}$. Theorem 8.1.1 shows that all observable unitary realizations for $\{r_j\}_0^{\nu-1}$ are isomorphic. So without loss of generality we can assume $E = C$ and $V = U$. Therefore T_m is the Toeplitz matrix generated by $\{CU^kC^*\}_0^{m-1}$. This completes the proof.

8.4 The Levinson algorithm and the singular case

The following result shows how one can use the Levinson algorithm to determine when a Toeplitz matrix is positive and singular. To this end recall that the Levinson system associated with T_ν is given by

$$\begin{bmatrix} \alpha_{k,1} & \alpha_{k,2} & \cdots & \alpha_{k,k-1} & 1 \end{bmatrix} T_k = \begin{bmatrix} 0 & 0 & \cdots & 0 & \epsilon_k \end{bmatrix} \qquad (k = 1, 2, \cdots, \nu). \tag{4.1}$$

If $k = 1$, then $\epsilon_1 = T_1 = r_0$ and the $\alpha_{1,0}$ term is not present.

THEOREM 8.4.1 *Let T_ν on $\mathbb{C}^\nu$ be the Toeplitz matrix generated by a set of scalars $\{r_j\}_0^{\nu-1}$. Then T_ν is positive and singular if and only if the following two conditions hold.*

(i) There exists an integer $\mu < \nu$ such that the Levinson system (4.1) has a solution for $k = 1, 2, \cdots, \mu+1$ and the errors $\{\epsilon_j\}_{j=1}^{\mu+1}$ satisfy

$$\epsilon_1 \geq \epsilon_2 \geq \cdots \geq \epsilon_\mu > \epsilon_{\mu+1} = 0. \tag{4.2}$$

(ii) The scalars $\{r_k\}_\mu^{\nu-1}$ recursively satisfy the difference equations

$$r_k = -\sum_{j=1}^{\mu} \alpha_{\mu+1,j} r_{k+j-(\mu+1)} \qquad (k = \mu, \mu+1, \cdots, \nu). \tag{4.3}$$

In this case, the polynomial

$$\alpha(z) = z^\mu + \sum_{k=1}^{\mu} \alpha_{\mu+1,k} z^{k-1} \tag{4.4}$$

has μ distinct roots on the unit circle, where $\{\alpha_{\mu+1,k}\}_{k=1}^{\mu}$ is the $\mu+1$ solution to the Levinson system in (4.1) .

PROOF. First assume that conditions (i) and (ii) hold. Since $\epsilon_1 \geq \epsilon_2 \geq \cdots \geq \epsilon_\mu > 0$, Corollary 6.4.2 shows that T_μ is strictly positive and the Levinson system in (4.1) has a unique solution for $k = 1, 2, \cdots, \mu$. Furthermore, (4.3) implies that

$$\begin{bmatrix} \alpha_{k,1} & \alpha_{k,2} & \cdots & \alpha_{k,k-1} & 1 \end{bmatrix} T_k = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \qquad (k = \mu+1, \mu+2, \cdots, \nu)$$

where

$$\begin{bmatrix} \alpha_{k,1} & \alpha_{k,2} & \cdots & \alpha_{k,k-1} & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \alpha_{\mu+1,1} & \alpha_{\mu+1,2} & \cdots & \alpha_{\mu+1,\mu} & 1 \end{bmatrix}.$$

In other words, with this definition of $\{\alpha_{k,\,m}\}_{m=1}^{k-1}$ for $k = \mu+1, \mu+2, \cdots, \nu$ the Levinson system in (4.1) has a solution for all $k = 1, 2, \cdots, \nu$. Moreover, the error terms $\epsilon_k = 0$ for $k = \mu+1, \mu+2, \cdots, \nu$. Lemma 6.4.1 shows that T_ν is positive and singular.

Now assume that T_ν is positive and singular. Let $\{C, U\}$ be an observable unitary realization for $\{r_j\}_1^{\nu-1}$. Recall that $T_\nu = W_\nu W_\nu^*$ where W_ν is the observability operator from $\mathcal{X}$ into $\mathbb{C}^\nu$ defined in (1.3). Moreover, if $\mu = \dim \mathcal{X}$, then W_μ is invertible. Using $T_j = W_j W_j^*$, it follows that μ is the largest integer such that T_j is invertible. Since T_μ is strictly positive, Corollary 6.4.2 guarantees that there exists a unique solution to the Levinson system (4.1) for $k = 1, 2, \cdots, \mu$ and $\epsilon_1 \geq \epsilon_2 \geq \cdots \geq \epsilon_\mu > 0$. Since $\epsilon_\mu \neq 0$ the Levinson algorithm shows that there exists a solution to

$$\left[\begin{array}{cccccc} \alpha_{\mu+1,\,1} & \alpha_{\mu+1,\,2} & \cdots & \alpha_{\mu+1,\,\mu} & 1 \end{array}\right] T_{\mu+1} = \left[\begin{array}{ccccc} 0 & 0 & \cdots & 0 & \epsilon_{\mu+1} \end{array}\right]; \tag{4.5}$$

see Theorem 6.3.1. Moreover, the solution to this equation is unique. To see this recall that T_μ is the $\mu \times \mu$ matrix contained in the upper left hand corner of $T_{\mu+1}$. So (4.5) yields

$$\left[\begin{array}{cccc} \alpha_{\mu+1,\,1} & \alpha_{\mu+1,\,2} & \cdots & \alpha_{\mu+1,\,\mu} \end{array}\right] T_\mu = \left[\begin{array}{cccc} r_\mu & r_{\mu-1} & \cdots & r_1 \end{array}\right].$$

Because T_μ is invertible, the $\{\alpha_{\mu+1,\,k}\}_{k=1}^{\mu}$ are uniquely determined by this equation. By consulting (4.5) this readily implies that $\epsilon_{\mu+1}$ is also unique. Hence there is a unique solution to (4.5).

Let $\alpha(z)$ be the characteristic polynomial for U, that is,

$$\alpha(z) = \det[zI - U] = \alpha_1 + \alpha_2 z + \alpha_3 z^2 + \cdots + \alpha_\mu z^{\mu-1} + z^\mu.$$

According to the Cayley-Hamilton theorem

$$\alpha_1 + \alpha_2 U + \alpha_3 U^2 + \cdots + \alpha_\mu U^{\mu-1} + U^\mu = 0.$$

This readily implies that $[\alpha_1, \alpha_2, \cdots, \alpha_\mu, 1] W_{\mu+1} = 0$. Since $T_{\mu+1} = W_{\mu+1} W_{\mu+1}^*$, we obtain the following Levinson system of equations

$$\left[\begin{array}{cccccc} \alpha_1 & \alpha_2 & \cdots & \alpha_\mu & 1 \end{array}\right] T_{\mu+1} = 0. \tag{4.6}$$

So the row vector $[\alpha_1, \alpha_2, \cdots, \alpha_\mu]$ corresponding to the characteristic polynomial for U forms a solution to this Levinson system. Because the solution to solution to (4.5) is unique, we obtain $\alpha_{\mu+1,\,j} = \alpha_j$ for $j = 1, 2, \cdots, \mu$ and $\epsilon_{\mu+1} = 0$. This proves part (i).

Recall that $r_m = CU^m C^*$ for $m = 0, 1, \cdots, \nu$. By employing the Cayley-Hamilton Theorem with $k \geq \mu$, we obtain

$$r_k = CU^k C^* = CU^{k-\mu} U^\mu C^* = -\sum_{j=1}^{\mu} \alpha_j CU^{k-\mu} U^{j-1} C^* = -\sum_{j=1}^{\mu} \alpha_j r_{k+j-(\mu+1)}.$$

This yields the formula in (4.3). Our previous analysis shows that $\alpha(z)$ in (4.4) is the characteristic polynomial for U. Because U is unitary and the pair $\{C, U\}$ is observable, all the eigenvalues of U are distinct and on the unit circle; see Section 8.2. Therefore $\alpha(z)$ has μ distinct roots on the unit circle. This completes the proof.

The Levinson algorithm can be used to determine if a Toeplitz matrix T_ν on $\mathbb{C}^\nu$ generated by $\{r_k\}_1^\nu$ is positive and singular. To accomplish simply run the Levinson algorithm on $\{r_k\}_1^\nu$. Then T_ν is positive and singular if and only if there exists and integer $\mu < \nu$ satisfying parts (i) and (ii) in Theorem 8.4.1.

8.4.1 Computing sinusoids from the Levinson algorithm

Now assume that T_ν on $\mathbb{C}^\nu$ is a positive and singular Toeplitz matrix generated by $\{r_k\}_1^\nu$. The results in Section 8.2 show that $\{r_k\}_1^\nu$ admit a representation of the form

$$r_n = \sum_{k=1}^{\mu} |\, a_k|^2 e^{\imath\omega_k n} \qquad (n = 0, 1, \cdots, \nu - 1)\,. \tag{4.7}$$

Recall that $\{|\, a_k|, \omega_k\}_1^\mu$ are called the amplitudes and frequencies associated with T_ν. Let $\xi(n)$ be the standard sinusoid process generated by the amplitudes and frequencies $\{a_k, \omega_k\}_1^\mu$. Then the autocorrelation function for $\xi(n)$ is given by $R_\xi(n) = r_n$ for $n = 0, 1, \cdots, \nu-1$. To obtain the amplitudes and frequencies $\{|\, a_k|, \omega_k\}_1^\mu$ of T_ν directly from the Levinson algorithm, let $\mu < \nu$ be the integer satisfying parts (i) and (ii) in Theorem 8.4.1. Let $\{\alpha_{\mu+1,\,k}\}_1^\mu$ be the solution to (4.1) obtained form the Levinson algorithm with $k = \mu + 1$ and $\epsilon_{\mu+1} = 0$. Then $\{e^{\imath\omega_k}\}_1^\mu$ are the roots of the polynomial $\alpha(z)$ in (4.4). Moreover, the amplitudes $\{|\, a_k|\}_1^\mu$ are obtained by solving the following equation for any $\mu - 1 \le m \le \nu$

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ e^{\imath\omega_1} & e^{\imath\omega_2} & \cdots & e^{\imath\omega_\mu} \\ e^{2\imath\omega_1} & e^{2\imath\omega_2} & \cdots & e^{2\imath\omega_\mu} \\ \vdots & \vdots & \cdots & \vdots \\ e^{m\imath\omega_1} & e^{m\imath\omega_2} & \cdots & e^{m\imath\omega_\mu} \end{bmatrix} \begin{bmatrix} |\, a_1|^2 \\ |\, a_2|^2 \\ |\, a_3|^2 \\ \vdots \\ |\, a_\mu|^2 \end{bmatrix} = \begin{bmatrix} r_0 \\ r_1 \\ r_2 \\ \vdots \\ r_{m-1} \end{bmatrix}. \tag{4.8}$$

To see this let $\{C, U\}$ be any observable unitary realization for $\{r_j\}_0^{\nu-1}$. Recall that $\alpha(z)$ is the characteristic polynomial for U. So the roots $\{e^{\imath\omega_k}\}_1^\mu$ of $\alpha(z)$ are the eigenvalues of U. By consulting Section 8.2, we see that $\{r_k\}_0^{\nu-1}$ is given by (4.7). Rearranging the summation in (4.7) readily yields the linear system in (4.8). In particular, there exists a solution $\{|\, a_k|\}_1^\mu$ to (4.8). Notice that the matrix in (4.8) is a $(m+1) \times \mu$ Vandermonde matrix. Because the roots of $\alpha(z)$ are distinct, and $m + 1 \ge \mu$, it follows that this matrix is one to one. So there exists a unique solution to (4.8).

8.5 Sinusoids plus white noise

In this section we study the sinusoid estimation problem when the noise is white. As in Section 7.1, let $\xi(n)$ be the standard sinusoid process generated by the amplitudes and frequencies $\{a_k, \omega_k\}_1^\mu$, that is,

$$\xi(n) = \sum_{k=1}^{\mu} a_k e^{\imath(\omega_k n + \theta_k)} \qquad \text{and} \qquad R_\xi(n) = \sum_{k=1}^{\mu} |\, a_k|^2 e^{\imath\omega_k n}\,. \tag{5.1}$$

Let $y(n)$ be the wide sense stationary random process determined by

$$y(n) = \xi(n) + d\varrho(n) = \sum_{k=1}^{\mu} a_k e^{\imath(\omega_k n + \theta_k)} + d\varrho(n)\,. \tag{5.2}$$

Here $\varrho(n)$ is a white noise random process independent to $\xi(n)$ and d is a constant. In particular, $\varrho(m)$ is orthogonal to $\xi(n)$ for all integers m and n. Because $\varrho(n)$ is white noise, $R_\varrho(n) = \delta_{n,0}$ where $\delta_{j,k}$ is the Kronecker delta. Hence the autocorrelation function for $y(n)$ is given by

$$R_y(n) = R_\xi(n) + d^2\delta_{n,0} = \sum_{k=1}^{\mu} |\, a_k|^2 e^{\imath\omega_k n} + d^2\delta_{n,0}\,. \tag{5.3}$$

The following result shows that any strictly positive Toeplitz matrix can be represented by a sinusoid plus a constant times a white noise process.

THEOREM 8.5.1 *Let T_ν on $\mathbb{C}^\nu$ be a strictly positive Toeplitz matrix generated by $\{r_k\}_1^\nu$. Let λ_ν be the smallest eigenvalue of T_ν and $\{|\, a_k|, \omega_k\}_1^\mu$ the amplitudes and frequencies associated with the positive and singular Toeplitz matrix $T_\nu - \lambda_\nu I$. Finally, let $y(n) = \xi(n) + d\varrho(n)$ be the wide stationary process defined by (5.2) where $d = \lambda_\nu^{1/2}$, and $\xi(n)$ is the standard sinusoid process generated by $\{a_k, \omega_k\}_1^\mu$ while $\varrho(n)$ is an independent white noise processes. Then $r_n = R_y(n)$ for $n = 0, 1, \cdots, \nu - 1$.*

Proof. Notice that $T_\nu - \lambda_\nu I$ is a positive and singular Toeplitz matrix. So there exists $\mu < \nu$ amplitudes and frequencies $\{|\, a_k|, \omega_k\}_1^\mu$ associated with $T_\nu - \lambda_\nu I$; see Sections 8.2 and 8.4.1. Clearly, $T_\nu - \lambda_\nu I$ is generated by $\{r_k - \lambda_\nu\delta_{k,0}\}_1^{\nu-1}$. Hence

$$r_n - \lambda_\nu\delta_{n,0} = \sum_{k=1}^{\mu} |\, a_k|^2 e^{\imath\omega_k n} = R_\xi(n) \qquad (n = 0, 1, \cdots, \nu - 1)\,.$$

Since $R_y(n) = R_\xi(n) + d^2\delta_{n,0}$, we arrive at $r_n = R_y(n)$ for $n = 0, 1, \cdots, \nu - 1$. This completes the proof.

8.5.1 Sinusoid estimation in white noise

As before, let $\xi(n)$ be the standard sinusoid process generated by the amplitudes and frequencies the $\{a_k, \omega_k\}_1^\mu$. Let $T_{\xi\nu}$ on $\mathbb{C}^\nu$ be the Toeplitz matrix generated by $\{R_\xi(n)\}_0^{\nu-1}$. Let U on $\mathbb{C}^\nu$ be the unitary diagonal matrix and C the operator from $\mathbb{C}^\nu$ into $\mathbb{C}$ defined by

$$U = \mathrm{diag}\,\{e^{\imath\omega_k}\}_1^\mu \qquad \text{and} \qquad C = \begin{bmatrix} a_1 & a_2 & \cdots & a_\mu \end{bmatrix}. \tag{5.4}$$

Then $\{C, U\}$ is an observable unitary realization of $\{R_\xi(n)\}_0^{\nu-1}$. Moreover, $T_{\xi\nu} = W_\nu W_\nu^*$ where $W_\nu = [C, CU, \cdots, CU^{\nu-1}]^{tr}$. Finally, recall that

$$\begin{aligned} \mathrm{rank} T_{\xi k} &= k \qquad \text{if } k \le \mu \\ \mathrm{rank} T_{\xi k} &= \mu \qquad \text{if } k > \mu\,. \end{aligned} \tag{5.5}$$

Notice that $\mathrm{rank} T_{\xi k}$ is positive and singular if and only if $k > \mu$. In particular, the rank of $T_{\xi k}$ can be used to determine the number of sinusoids in the random process $\xi(n)$.

As before, let $y(n)$ be the wide sense stationary process given by $y(n) = \xi(n) + d\varrho(n)$, where $\varrho(n)$ is a white noise random process independent to $\xi(n)$ and d is a constant. In this setting, our sinusoid estimation problem is to find the signal $\xi(n)$ in (5.1) given the

process $y(n)$. In other words, find the amplitudes and frequencies $\{|\,a_k|, \omega_k\}_1^\mu$ for the signal $\xi(n)$ from the process $y(n)$. Clearly, this is a special case of the sinusoid estimation problem introduced in Section 7.2.

Because $\varrho(n)$ is white noise, the autocorrelation function $R_y(n) = R_\xi(n) + d^2\delta_{n,0}$. Let T_ν on $\mathbb{C}^\nu$ be the Toeplitz matrix generated by $\{R_y(n)\}_0^{\nu-1}$. Then $T_\nu = T_{\xi\nu} + d^2 I_\nu$ where I_ν is the identity operator on $\mathbb{C}^\nu$. Now assume that $\nu > \mu$. Then $T_\nu - d^2 I_\nu = T_{\xi\nu}$ is a positive and singular Toeplitz matrix of rank μ. Because T_ν is positive and $\nu > \mu$, it follows that d^2 is the smallest eigenvalue for T_ν. In other words, let λ_ν be the smallest eigenvalue of T_ν. Then the rank of $T_\nu - \lambda_\nu I_\nu$ equals μ for all $\nu > \mu$. On the other hand, if $\nu \leq \mu$, then clearly, $\text{rank}(T_\nu - \lambda_\nu I_\nu) < \mu$. Therefore the number of sinusoids μ in the process $\xi(n)$ is given by

$$\mu = \max\{\text{rank}(T_k - \lambda_k I_k) : k \geq 1\}\,. \tag{5.6}$$

This observation leads to the following method to estimate the sinusoid process $\xi(n)$ from $y(n)$: To estimate the amplitudes and frequencies $\{|\,a_k|, \omega_k\}_1^\mu$ from $y(n)$, let λ_k be the smallest eigenvalue for T_k and compute μ in (5.6). Once μ is determined, choose any $\nu > \mu$. Then $d = \lambda_\nu^{1/2}$ and $T_{\xi\nu} = T_\nu - \lambda_\nu I_\nu$ is positive and singular. Now the amplitudes and frequencies of $\xi(n)$ are computed by applying the techniques in Section 8.2 or 8.4.1 to compute the amplitudes and frequencies $\{|\,a_k|, \omega_k\}_1^\mu$ associated with the $T_{\xi\nu}$.

If μ is infinite, then this algorithm does not work. In this case, the process $y(n)$ does not admit a representation of the form $y(n) = \xi(n) + d\varrho(n)$ where the noise is white.

Finally, it is noted that one can also use the Levinson algorithm to determine the smallest eigenvalue for a positive Toeplitz matrix T_ν. To see this choose any $\lambda > 0$. Then apply the Levinson algorithm to determine if the Toeplitz matrix $T_\nu - \lambda I$ is positive. If $T_\nu - \lambda I$ not positive, then apply the Levinson algorithm to see if $T_\nu - (\lambda/2)I$ is positive. If $T_\nu - \lambda I$ is positive, then apply the Levinson algorithm to determine if $T_\nu - 2\lambda I$ is positive. By continuing in this fashion one can use a bisection method to converge to the smallest eigenvalue of T_ν.

Chapter 9

Appendix: Discrete Time Systems

This chapter presents some elementary facts concerning discrete time state space systems. Controllability and observability for discrete systems is studied.

9.1 Discrete time invariant systems

In this section we will review some elementary facts concerning discrete time state space systems. To this end, consider the discrete time system

$$x(n+1) = Ax(n) \tag{1.1}$$

subject to the initial condition $x(0) = x_0$. The time index is an integer $n \geq 0$. Moreover, A is an operator on a finite dimensional space $\mathcal{X}$ and $x(n)$ is a vector in $\mathcal{X}$ for all integers $n \geq 0$. The vector $x(n)$ is called the *state* and $\mathcal{X}$ is the *state space.* The solution to the state space system in (1.1) is given by

$$x(n) = A^n x_0 \qquad (n \geq 0)\,. \tag{1.2}$$

To see this notice that by recursively computing $x(n+1) = Ax(n)$, we obtain

$$\begin{aligned} x(1) &= Ax_0 \\ x(2) &= Ax(1) = A^2 x_0 \\ x(3) &= Ax(2) = A^3 x_0\,. \end{aligned}$$

By continuing in this fashion, we arrive at $x(n) = A^n x_0$ for all integers $n \geq 0$. Therefore the solution to the state space system $x(n+1) = Ax(n)$ is given by $x(n) = A^n x_0$.

One can use the Jordan form to compute A^n. To see this recall that $A = PJP^{-1}$ where J is a Jordan matrix containing the eigenvalues of A and P is an invertible matrix consisting of the eigenvectors and generalized eigenvectors for A. In most problems A is a matrix on $\mathbb{C}^m$ and the eigenvalues $\{\lambda_j\}_1^m$ of A are distinct. In this case, J is a diagonal matrix consisting of the eigenvalues for A. To be precise, if all the eigenvalues of $\{\lambda_j\}_1^m$ of A are distinct and $\{f_j\}_1^m$ are the corresponding eigenvectors ($Af_j = \lambda_j f_j$ for $j = 1, 2, \cdots, m$), then $A = PJP^{-1}$

where J and P are the matrices given by

$$\begin{aligned} J &= \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_m \end{bmatrix} \\ P &= \begin{bmatrix} f_1 & f_2 & f_3 & \cdots & f_m \end{bmatrix}. \end{aligned} \tag{1.3}$$

Finally, it is well known that the eigenvectors $\{f_j\}_1^m$ corresponding to distinct eigenvalues are linearly independent. Therefore P is invertible.

We claim that $A^n = PJ^nP^{-1}$ for all integers $n \geq 0$. Using $A = PJP^{-1}$, we obtain

$$\begin{aligned} A^2 &= AA = (PJP^{-1})(PJP^{-1}) = PJ^2P^{-1} \\ A^3 &= A^2A = (PJ^2P^{-1})(PJP^{-1}) = PJ^3P^{-1}. \end{aligned}$$

By continuing in this fashion, we see that $A^n = PJ^nP^{-1}$ for all integers $n \geq 0$.

An example. Consider the state space system given by

$$\begin{bmatrix} x_1(n+1) \\ x_2(n+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1/6 & 5/6 \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix}. \tag{1.4}$$

In this case, the matrix A is given by

$$A = \begin{bmatrix} 0 & 1 \\ -1/6 & 5/6 \end{bmatrix}.$$

Notice that the characteristic polynomial for A is

$$\det[\lambda I - A] = \det \begin{bmatrix} \lambda & -1 \\ 1/6 & \lambda - 5/6 \end{bmatrix} = \lambda^2 - \frac{5}{6}\lambda + \frac{1}{6} = \left(\lambda - \frac{1}{3}\right)\left(\lambda - \frac{1}{2}\right).$$

Therefore the eigenvalues for A are $\lambda_1 = 1/3$ and $\lambda_2 = 1/2$. The eigenvectors f_1 and f_2 respectively corresponding to the eigenvalues $1/3$ and $1/2$ are given by

$$f_1 = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

This readily implies that

$$J = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} \qquad \text{and} \qquad P = \begin{bmatrix} 3 & 2 \\ 1 & 1 \end{bmatrix}.$$

Using $A^n = PJ^nP^{-1}$, we obtain

$$\begin{aligned} A^n &= PJ^nP^{-1} = \begin{bmatrix} 3 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1/3^n & 0 \\ 0 & 1/2^n \end{bmatrix} \begin{bmatrix} 1 & -2 \\ -1 & 3 \end{bmatrix} = \begin{bmatrix} 3 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1/3^n & -2/3^n \\ -1/2^n & 3/2^n \end{bmatrix} \\ &= \begin{bmatrix} 1/3^{n-1} - 1/2^{n-1} & 3/2^{n-1} - 2/3^{n-1} \\ 1/3^n - 1/2^n & 3/2^n - 2/3^n \end{bmatrix}. \end{aligned}$$

This readily shows that

$$A^n = \begin{bmatrix} 1/3^{n-1} - 1/2^{n-1} & 3/2^{n-1} - 2/3^{n-1} \\ 1/3^n - 1/2^n & 3/2^n - 2/3^n \end{bmatrix}.$$

In particular, the solution $x(n)$ to the discrete time system in (1.4) is given by $x(n) = A^n x(0)$, that is,

$$\begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} = \begin{bmatrix} 1/3^{n-1} - 1/2^{n-1} & 3/2^{n-1} - 2/3^{n-1} \\ 1/3^n - 1/2^n & 3/2^n - 2/3^n \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}.$$

Finally, it is noted that A^n converges to zero as n tends to infinity. This follows from the fact that the eigenvalues of A are contained in the open unit disc $\{z : |z| < 1\}$.

Discrete time approximation of continuous time systems. Now let us show how one can approximate a continuous time system by a discrete time system. To this end, consider the continuous time system given by $\dot{q} = Fq$ where F is an operator on a finite dimensional state $\mathcal{X}$. Recall that the solution to this system is given by $q(t) = e^{Ft}q(0)$. Now consider any small time increment $h > 0$. Let A be the operator on $\mathcal{X}$ defined by $A = e^{Fh}$. Set $x(n) = q(nh)$ and $x_0 = q(0)$. Notice that $x(n)$ is the solution to the differential equation $\dot{q} = Fq$ at time $t = nh$. Thus

$$x(n) = q(nh) = e^{Fnh}q(0) = \left(e^{Fh}\right)^n q(0) = A^n x_0 .$$

In other words, $x(n) = A^n x_0$. Hence $x(n)$ is the solution to the difference equation

$$x(n+1) = Ax(n) \qquad (x_0 = q(0)) . \tag{1.5}$$

Therefore the solution $x(n)$ to the difference equation (1.5) yields the solution to the differential equation $\dot{q} = Fq$ at times nh, that is, $q(nh) = x(n)$ for all integers $n \geq 0$.

9.1.1 The gambler's ruin problem

A discrete time system naturally occurs in the classical gambler's ruin problem. To recall this problem consider a gambler playing a card game or tossing a coin with a probability of p of winning the game and a probability of q of losing the game where $p + q = 1$. Without loss of generality, we assume that $p > 0$, that is, the probability of winning any game is strictly positive. Moreover, assume that the player bets one dollar in each game. So the gambler makes one dollar for a winning hand and looses one dollar for a loosing hand. Now assume that the strategy of the gambler is to start with n dollars and keep playing until the gambler achieves $m \geq n$ dollars or goes broke. The profit is $m - n$ dollars. To solve this problem, let $W_{n,m}$ be the event that the gambler makes m dollars starting with n dollars. Set $p_n = P(W_{n,m})$ where $P(E)$ denoted the probability that the event E occurs. Recall that

$$P(E) = P(E|F)P(F) + P(E|F^c)P(F^c) \tag{1.6}$$

where $P(E|F) = P(E \bigcap F)/P(F)$ is the conditional probability and F^c is the complement of the event F. Now let F be the probability that the gambler wins the current game. Clearly, $P(F) = p$ and $P(F^c) = q$. Then using (1.6) we obtain the following difference equation

$$p_n = P(W_{n,m}) = P(W_{n,m}|F)P(F) + P(W_{n,m}|F^c)P(F^c) = p_{n+1}p + p_{n-1}q .$$

This readily yields the following difference equation

$$p_n = p_{n+1}p + p_{n-1}q\,. \tag{1.7}$$

This difference equation is subject to an initial condition and final condition. Notice that $p_0 = P(W_{0,m})$ is the probability that the gambler wins m dollars starting with no money. Hence $p_0 = 0$. On the other hand, observe that $p_m = P(W_{m,m})$ is the probability that the gambler achieves m dollars starting with m dollars. In other words, the gambler already has m dollars and is not going to play the game. Thus $p_m = 1$. Therefore the solution to the gambler's ruin problem is obtained by solving the difference equation in (1.7) subject to the initial condition $p_0 = 0$ and final condition $p_m = 1$. This is a two point boundary value problem.

There is a trick to solving the difference equation in (1.7). However, this method does not work in general. To present this method first observe that applying $p + q = 1$ in (1.7) yields

$$(p_{n+1} - p_n)p = (p_n - p_{n-1})q\,.$$

Now let $\Delta_n = p_n - p_{n-1}$ and set $r = q/p$. Then we obtain the difference equation $\Delta_{n+1} = r\Delta_n$ where the initial condition $\Delta_1 = p_1$. Notice that $\Delta_2 = r\Delta_1 = rp_1$ and $\Delta_3 = r\Delta_2 = r^2p_1$. By continuing in this fashion, we see that $\Delta_n = r^{n-1}p_1$ for all integers $n \geq 1$. By using a telescoping series we arrive at

$$\begin{aligned} p_n &= (p_1 - p_0) + (p_2 - p_1) + (p_3 - p_2) + \cdots + (p_{n-1} - p_{n-2}) + (p_n - p_{n-1}) \\ &= \sum_{k=1}^{n} \Delta_k = \sum_{k=1}^{n} r^{k-1}p_1 = p_1 \sum_{k=0}^{n-1} r^k\,. \end{aligned}$$

Thus $p_n = p_1 \sum_0^{n-1} r^k$. Notice that $r = 1$ if and only if $p = q = 1/2$, that is, the game is fair. To see this observe that if $r = 1$, then $p = q$. Using $p + q = 1$, we have $2p = 1$, or equivalently, $p = 1/2$. By applying the classical geometric series Lemma 4.4.2 in Chapter 4 to the series $p_n = p_1 \sum_0^{n-1} r^k$, we obtain

$$\begin{aligned} p_n &= \frac{1 - r^n}{1 - r}\, p_1 \qquad \text{if } p \neq 1/2 \\ &= np_1 \qquad\qquad\quad \text{if } p = 1/2\,. \end{aligned} \tag{1.8}$$

By employing the final condition $p_m = 1$ in (1.8), we arrive at $p_1 = (1 - r)/(1 - r^m)$ if $p \neq 1/2$, and $p_1 = 1/m$ if $p = 1/2$. Substituting this into (1.8) yields the following solution to the gambler's ruin problem

$$\begin{aligned} P(W_{n,m}) &= \frac{1 - r^n}{1 - r^m} \qquad \text{if } p \neq 1/2 \\ &= \frac{n}{m} \qquad\quad \text{if } p = 1/2\,. \end{aligned} \tag{1.9}$$

The solution to the gambler's ruin problem shows that the probability of doubling one's money $m = 2n$ in a fair game $p = 1/2$ is 50%. This result is not surprising. However, the

solution to the gambler's ruin problem also shows that if $p < 1/2$, then it is better to bet all your n dollars in one game rather than play one game at a time. On the other hand, if $p > 1/2$, then it is better to bet one dollar on each game rather than bet all your money in one game. To be more explicit, for the moment assume that $p > 1/2$. In this case, $r = q/p < 1$. In particular, r^m converges to zero as m tends to infinity. So according to (1.9), the probability of making an infinite amount of money $m = \infty$ starting with n dollars is given by

$$P(W_{n,\infty}) = 1 - r^n \qquad (p > 1/2)\,.$$

For example, if $p = 0.51$ and $n = 100$, then the probability that a gambler will make an infinite amount of money is $1 - (49/51)^{100} \approx 0.98$. This is why a casino will not let the players count cards or use a computer. In this case, $p > 1/2$ and a player can bankrupt the casino. This is also why casino's make a tremendous amount of money. The p for a casino for many games is greater than or equal to 0.55.

Now assume that $p < 1/2$. In this case, $r = q/p > 1$. So if n and m are large, then

$$P(W_{n,m}) = \frac{1 - r^n}{1 - r^m} \approx \frac{r^n}{r^m} = \left(\frac{p}{q}\right)^{m-n}\,.$$

If $p = 1/2$, then there is a 50% probability of achieving $m = 200$ dollars starting with $n = 100$ dollars. Now assume that $p = 0.49$ and the gambler starts out with $n = 100$ dollars and $m = 200$ dollars. Then $P(W_{n,m}) = 0.018$. So in this case, it is better to bet the one hundred dollars in the first game which yields a 49% chance of achieving 200 dollars, rather than playing one game at a time which has only a 1.8% chance of doubling the original 100 dollars. In fact, there is only a 36.4% chance of achieving $m = 125$ dollars starting with 100 dollars playing one game at a time. The situation is even worse the as p becomes smaller. For example, if $p = 0.45$ and $n = 100$, then there is only a 37% chance of achieving $m = 105$ dollars, and 0.66% chance of achieving $m = 125$ dollars.

A state space interpretation of the gambler's ruin problem. Now let us use state space techniques to gain some further insight into the difference equation (1.7) arising in the gambler's ruin problem. To convert (1.7) to a second order state space system, let $x_1(n) = p_n$ and $x_2(n) = x_1(n+1) = p_{n+1}$. By consulting (1.7) with $r = q/p$, we see that

$$x_2(n+1) = p_{n+2} = -qp_n/p + p_{n+1}/p = -rx_1(n) + x_2(n)/p\,.$$

Therefore the difference equation in (1.7) admits a state space representation of the form

$$\begin{bmatrix} x_1(n+1) \\ x_2(n+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -r & 1/p \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} \tag{1.10}$$

subject to the initial condition $x_1(0) = 0$ and final condition $x_1(m) = 1$. Here $p_n = x_1(n)$ is the probability of winning m dollars starting with n dollars.

In this case, the state matrix A given by

$$A = \begin{bmatrix} 0 & 1 \\ -r & 1/p \end{bmatrix}\,. \tag{1.11}$$

Notice that the characteristic polynomial for A is determined by

$$\det[\lambda I - A] = \det \begin{bmatrix} \lambda & -1 \\ r & \lambda - 1/p \end{bmatrix} = \lambda^2 - \lambda/p + r = (\lambda - r)(\lambda - 1) .$$

The last equality follows from the fact that $r + 1 = q/p + p/p = (p + q)/p = 1/p$. Therefore $(\lambda - r)(\lambda - 1)$ is the characteristic polynomial for A. In particular, the eigenvalues for A are given by r and 1.

Now let us compute A^n when $r \neq 1$, or equivalently, $p \neq 1/2$. In this case, the eigenvalues r and 1 for A are distinct. Moreover, the eigenvectors f_1 and f_2 corresponding to r and 1 are respectively given by

$$f_1 = \begin{bmatrix} 1 \\ r \end{bmatrix} \quad \text{and} \quad f_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} .$$

Thus $A^n = QJ^nQ^{-1}$ where $Q = [f_1, f_2]$ and $J = \operatorname{diag}[r, 1]$. A simple calculation shows that

$$\begin{aligned} A^n &= \frac{1}{1-r} \begin{bmatrix} 1 & 1 \\ r & 1 \end{bmatrix} \begin{bmatrix} r^n & 0 \\ 0 & 1^n \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -r & 1 \end{bmatrix} = \frac{1}{1-r} \begin{bmatrix} 1 & 1 \\ r & 1 \end{bmatrix} \begin{bmatrix} r^n & -r^n \\ -r & 1 \end{bmatrix} \\ &= \frac{1}{1-r} \begin{bmatrix} r^n - r & 1 - r^n \\ r^{n+1} - r & 1 - r^{n+1} \end{bmatrix} . \end{aligned}$$

Therefore A^n is given by

$$A^n = \frac{1}{1-r} \begin{bmatrix} r^n - r & 1 - r^n \\ r^{n+1} - r & 1 - r^{n+1} \end{bmatrix} \qquad (\text{if } p \neq 1/2) . \tag{1.12}$$

Now let us compute A^n when $p = 1/2$. In this case, the matrix A is given by

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix} . \tag{1.13}$$

An easy way to compute A^n is simply to apply L'Hospital's rule to the entries of A^n in (1.12), that is,

$$A^n = \lim_{r \to 1} \frac{1}{-1} \begin{bmatrix} nr^{n-1} - 1 & -nr^{n-1} \\ (n+1)r^n - 1 & -(n+1)r^n \end{bmatrix} = \begin{bmatrix} 1 - n & n \\ -n & n + 1 \end{bmatrix} .$$

In other words,

$$A^n = \begin{bmatrix} 1 - n & n \\ -n & 1 + n \end{bmatrix} \qquad (\text{if } p = 1/2) . \tag{1.14}$$

To obtain this result using Jordan forms, recall that r and 1 are the eigenvalues for A. So if $r = 1$, then one is the only eigenvalue for A. The eigenvector f_1 corresponding to A is given by $f_1 = [1, 1]^*$. Recall that the generalized eigenvector f_2 for A satisfies $(A - \lambda I)f_2 = f_1$. Since $\lambda = 1$, we have

$$f_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} f_2 = (A - I)f_2 .$$

Thus $f_2 = [0, 1]^*$ is a generalized eigenvector for A. In this case, $A^n = QJ^nQ^{-1}$, where

$$Q = \begin{bmatrix} f_1 & f_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \qquad \text{and} \qquad J = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} .$$

Notice that

$$J^n = \begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix}.$$

Using this along with $A^n = QJ^nQ^{-1}$, we obtain

$$A^n = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1-n & n \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1-n & n \\ -n & 1+n \end{bmatrix}.$$

Therefore A^n is given by (1.14) when $p = 1/2$.

Recall that the solution to the state space difference equation $x(n+1) = Ax(n)$ is given by $x(n) = A^n x(0)$. By consulting the formulas for A^n in (1.12) and (1.14), we see the general solution to the difference equation in (1.10) corresponding to the gambler,s ruin problem is given by

$$\begin{aligned} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} &= \frac{1}{1-r} \begin{bmatrix} r^n - r & 1 - r^n \\ r^{n+1} - r & 1 - r^{n+1} \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} \qquad (\text{if } p \neq 1/2) \\ \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} &= \begin{bmatrix} 1-n & n \\ -n & 1+n \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} \qquad (\text{if } p = 1/2)\,. \end{aligned} \tag{1.15}$$

To solve the gambler's ruin problem, recall that $p_n = x_1(n)$ where $x(n)$ is computed according to the state space equation in (1.10) subject to the initial condition $x_1(0) = 0$ and the final condition $x_1(m) = 1$. By employing these initial and final conditions in (1.15) with $m = n$, we arrive at

$$\begin{aligned} \begin{bmatrix} 1 \\ x_2(m) \end{bmatrix} &= \frac{1}{1-r} \begin{bmatrix} r^m - r & 1 - r^m \\ r^{m+1} - r & 1 - r^{m+1} \end{bmatrix} \begin{bmatrix} 0 \\ x_2(0) \end{bmatrix} \qquad (\text{if } p \neq 1/2) \\ \begin{bmatrix} 1 \\ x_2(m) \end{bmatrix} &= \begin{bmatrix} 1-m & m \\ -m & 1+m \end{bmatrix} \begin{bmatrix} 0 \\ x_2(0) \end{bmatrix} \qquad (\text{if } p = 1/2) \end{aligned}$$

where $x_2(0)$ is the initial condition and $x_2(m)$ is the final condition, which are not specified. Notice that both of these equations have a unique solution for $x_2(0)$ and $x_2(m)$. In fact,

$$\begin{aligned} x_2(0) &= (1-r)/(1-r^m) \qquad \text{if } p \neq 1/2 \\ &= 1/m \qquad \text{if } p = 1/2\,. \end{aligned} \tag{1.16}$$

By substituting the initial conditions $x_1(0) = 0$ and $x_2(0)$ in (1.16) into equation (1.15) with $x_1(n) = p_n$, we obtain the following classical solution to the gambler's ruin problem

$$\begin{aligned} P(W_{n,m}) &= \frac{1-r^n}{1-r^m} \qquad \text{if } p \neq 1/2 \\ &= \frac{n}{m} \qquad \text{if } p = 1/2\,. \end{aligned} \tag{1.17}$$

9.2 Stable discrete time systems

This section is devoted to the stability of discrete time systems. As before, consider the discrete time system given by $x(n+1) = Ax(n)$ where A is an operator on a finite dimensional space $\mathcal{X}$. The discrete time system $x(n+1) = Ax(n)$ is *stable* if given any initial condition $x(0) = x_0$, then

$$\lim_{n\to\infty} x(n) = 0\,.$$

Since $x(n) = A^n x_0$, we see that $x(n+1) = Ax(n)$ is stable if and only if

$$\lim_{n\to\infty} A^n = 0\,. \tag{2.1}$$

The operator A on $\mathcal{X}$ is *discrete time stable* if the state space system $x(n+1) = Ax(n)$ is stable, or equivalently, A^n converges to zero as n tends to infinity. Notice that A^n converges to zero if and only if A^{*n} converges to zero. Therefore A is stable if and only if A^* is stable. The following is a fundamental stability result for discrete time systems.

THEOREM 9.2.1 *Let A be an operator on a finite dimensional space. Then the system $x(n+1) = Ax(n)$ is stable if and only if all the eigenvalues of A are contained in the open unit disc $\{z : |z| < 1\}$.*

PROOF. Assume that A is stable. Let λ be an eigenvalue for A with eigenvector f, that is, $Af = \lambda f$ where f is nonzero. Notice that $A^n f = \lambda^n f$ for all integers $n \geq 0$. Using the fact that A^n converges to zero as n tends to infinity, we obtain

$$0 = \lim_{n\to\infty} \|A^n f\| = \lim_{n\to\infty} \|\lambda^n f\| = \|f\| \lim_{n\to\infty} |\lambda|^n\,.$$

Because f is nonzero, $|\lambda|^n$ converges to zero as n tends to infinity. Therefore $|\lambda| < 1$. In other words, all the eigenvalues of A are contained in the open unit disc $\{z : |z| < 1\}$.

Now assume that all the eigenvalues of A are contained in the open unit disc. Recall that $A = PJP^{-1}$ where J is a Jordan matrix containing the eigenvalues of A on the diagonal and P is an invertible matrix consisting of the eigenvectors and generalized eigenvectors for A. For example, if all the eigenvalues of A are distinct, then J and P are given by (1.3) where $\{\lambda_k\}_1^m$ and $\{f_k\}_1^m$ are the eigenvalues and corresponding eigenvectors for A. Recall that $A^n = PJ^nP^{-1}$ for all integers $n \geq 0$. Because all the eigenvalues of A are contained in the open unit disc, J^n converges to zero as n tends to infinity. Since $A^n = PJ^nP^{-1}$, we see that A^n converges to zero as n tends to infinity. In other words, A is stable. This completes the proof.

Finally, it is noted that the matrix A in (1.11) associated with the gamblers ruin problem is always unstable. This follows from the fact that one is always an eigenvalue for this A.

9.2.1 Exercise

Problem 1. Compute A^n for the matrix A given by

$$A = \begin{bmatrix} 0 & -1 \\ 2 & 2 \end{bmatrix}\,.$$

Is A a discrete time stable matrix?

Problem 2. Compute A^n for the matrix A given by

$$A = \begin{bmatrix} 0 & -1/8 \\ 1 & 3/4 \end{bmatrix}.$$

Is A a discrete time stable matrix?

Problem 3. Assume that F is an operator on $\mathcal{X}$ and $A = e^{Fh}$ where $h > 0$. Then show that A is a discrete time stable operator if and only if all the eigenvalues of F are contained in the open left half plane $\{s : \Re s < 0\}$.

9.3 The general state space system

The general discrete time state space system is given by

$$x(n+1) = Ax(n) + Bu(n) \qquad \text{and} \qquad y = Cx(n) + Du(n)\,. \tag{3.1}$$

Here A is an operator on a finite dimensional space $\mathcal{X}$ and B is an operator mapping $\mathcal{U}$ into $\mathcal{X}$ while C is an operator from $\mathcal{X}$ into $\mathcal{Y}$ and D is an operator mapping $\mathcal{U}$ into $\mathcal{Y}$. The spaces $\mathcal{X}, \mathcal{U}$ and $\mathcal{Y}$ are all $\mathbb{C}^k$ spaces of the appropriate size. The state $x(n)$ is in $\mathcal{X}$, the input $u(n)$ is in $\mathcal{U}$ and the output $y(n)$ is in $\mathcal{Y}$ for all integers n. The initial condition is $x(0) = x_0$.

To obtain a solution to the state space system in (3.1), observe that $x(1) = Ax_0 + Bu(0)$. By recursively solving for the state $x(n)$ in (3.1), we obtain

$$\begin{aligned} x(2) &= Ax(1) + Bu(1) = A^2x_0 + ABu(0) + Bu(1) \\ x(3) &= Ax(2) + Bu(2) = A^3x_0 + A^2Bu(0) + ABu(1) + Bu(2)\,. \end{aligned}$$

By continuing in this fashion we see that the solution to (3.1) subject to the initial condition $x(0) = x_0$ is given by

$$x(n) = A^n x_0 + \begin{bmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{bmatrix} \begin{bmatrix} u(n-1) \\ u(n-2) \\ u(n-3) \\ \vdots \\ u(0) \end{bmatrix}. \tag{3.2}$$

By rewriting this equation along with $y(n) = Cx(n) + Du(n)$, we arrive at the following solution to the state space system in (3.1)

$$x(n) = A^n x_0 + \sum_{k=0}^{n-1} A^{n-k-1}Bu(k) \qquad (x(0) = x_0) \tag{3.3}$$

$$y(n) = CA^n x_0 + Du(n) + \sum_{k=0}^{n-1} CA^{n-k-1}Bu(k)\,. \tag{3.4}$$

9.3.1 A Discrete time approximation for a continuous time system

In this section we will construct a discrete time approximation for a continuous time system. To this end, consider the continuous time state space system given by

$$\dot{q} = Fq + Hw \qquad \text{and} \qquad z = Cq + Dw\,. \tag{3.5}$$

Here F is an operator on a finite dimensional state space $\mathcal{X}$ and H maps $\mathcal{U}$ into $\mathcal{X}$ while C maps $\mathcal{X}$ into $\mathcal{Y}$ and D maps $\mathcal{U}$ into $\mathcal{Y}$. Moreover, the state $q(t)$ is a vector in $\mathcal{X}$ while the input w a continuous function with values in $\mathcal{U}$ and the output z a vector in $\mathcal{Y}$. Now let $h > 0$ be a small increment of time. Let A on $\mathcal{X}$ and B mapping $\mathcal{U}$ into $\mathcal{X}$ be the operators defined by

$$A = e^{Fh} \qquad \text{and} \qquad B = \int_0^h e^{F(h-\tau)} H\, d\tau\,. \tag{3.6}$$

Let $\{u(n)\}_0^\infty$ be the sequence of vectors with values in $\mathcal{U}$ defined by

$$u(n) \;=\; w(nh + h/2) \qquad (n \geq 0)\,. \tag{3.7}$$

We claim that the discrete time system

$$x(n+1) = Ax(n) + Bu(n) \qquad \text{and} \qquad y(n) = Cx(n) + Du(n) \tag{3.8}$$

can be used as an approximation for the solution to the differential equation in (3.5). To be precise, if h is sufficiently small, then

$$q(nh) \approx x(n) \quad \text{and} \quad z(nh) \approx y(n) \qquad (\text{for all integers } n \geq 0)\,. \tag{3.9}$$

To derive the approximation in (3.9), let $\delta_k(t)$ be the operator valued function defined by

$$\begin{aligned} \delta_k(t) &= I_{\mathcal{U}} && \text{if} \quad kh \leq t < (k+1)h \\ &= 0 && \text{otherwise} \end{aligned}$$

where $k \geq 0$ is a positive integer. Here $I_{\mathcal{U}}$ is the identity operator on $\mathcal{U}$. Using fact that w is a continuous function, we see that $\sum_0^\infty \delta_k(t)u(k)$ is a step function which is approximately equal to $w(t)$, that is,

$$w(t) \approx \sum_{k=0}^{\infty} \delta_k(t)u(k)\,.$$

Recall that the solution to the state space system $\dot{q} = Fq + Hw$ is given by

$$q(t) = e^{Ft}q(0) + \int_0^t e^{F(t-\tau)} Hw(\tau)\, d\tau\,. \tag{3.10}$$

Using this along with $x_0 = q(0)$, we have

$$\begin{aligned} q(nh) &= e^{Fnh}x_0 + \int_0^{nh} e^{F(nh-\tau)}Hw(\tau)\,d\tau \\ &\approx \left(e^{Fh}\right)^n x_0 + \int_0^{nh} e^{F(nh-\tau)}H\left(\sum_{k=0}^{\infty}\delta_k(\tau)u(k)\right)d\tau \\ &= A^n x_0 + \sum_{k=0}^{n-1}\int_{kh}^{(k+1)h} e^{F(nh-\tau)}Hu(k)d\tau \\ &= A^n x_0 + \sum_{k=0}^{n-1} e^{F(nh-(k+1)h)}\int_{kh}^{(k+1)h} e^{F((k+1)h-\tau)}Hu(k)d\tau \\ &= A^n x_0 + \sum_{k=0}^{n-1} A^{n-k-1}\int_0^h e^{F(h-\sigma)}Hd\sigma u(k) \\ &= A^n x_0 + \sum_{k=0}^{n-1} A^{n-k-1}Bu(k)\,. \end{aligned}$$

The second from the last equality follows by a preforming the change of variable $-\sigma = kh - \tau$ in the integral. This readily shows that

$$q(nh) \approx A^n x_0 + \sum_{k=0}^{n-1} A^{n-k-1}Bu(k) = x(n)$$

where $x(n)$ is the solution to the difference equation in (3.8). In other words, $q(nh) \approx x(n)$. Since

$$z(t) = Cq(t) + Dw(t) \quad \text{and} \quad y(n) = Cx(n) + Du(n)$$

we also see that $z(nh) \approx y(n)$. Therefore if h is sufficiently small, then the difference equation in (3.8) can be used as an approximation for the continuous time system in (3.5).

9.3.2 Exercise

Problem 1. Use Matlab with $h = 0.01$ to find a discrete time approximation for the following state space system:

$$\begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}\begin{bmatrix} \dot{q} \\ \dot{q} \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} w(t) \quad \text{and} \quad z(t) = \begin{bmatrix} 1 & 2 \end{bmatrix}\begin{bmatrix} \dot{q} \\ \dot{q} \end{bmatrix}.$$

9.4 Controllability of discrete time systems

In this section we review some elementary facts concerning the notion of controllability for discrete time systems.

9.4.1 Some fundamental results from linear algebra

First let us recall the following result known as the Cayley-Hamilton theorem.

THEOREM 9.4.1 (Cayley-Hamilton) *Let A be an operator on a m dimensional space $\mathcal{X}$ and*

$$d(\lambda) = \det[\lambda I - A] = \alpha_0 + \alpha_1\lambda + \alpha_2\lambda^2 + \cdots + \alpha_{m-1}\lambda^{m-1} + \lambda^m$$

be the characteristic polynomial for A. Then $d(A) = 0$, that is,

$$0 = \alpha_0 I + \alpha_1 A + \alpha_2 A^2 + \cdots + \alpha_{m-1}A^{m-1} + A^m\,.$$

In particular, if n is any integer greater than or equal to m, then A^n is a linear combination of $\{A^k\}_{k=0}^{m-1}$.

Let T be an operator from $\mathbb{C}^k$ into $\mathbb{C}^m$. Recall that T is onto if the range of T equals $\mathbb{C}^m$. The null space or kernel of T is defined by $\{f \in \mathbb{C}^k : Tf = 0\}$. Moreover, T is one to one if the null space of T is zero. In other words, T is one to one if $Tf = 0$ implies that $f = 0$. The following result from linear algebra will also be useful in studying controllability.

LEMMA 9.4.2 *Let T be an operator from $\mathbb{C}^k$ into $\mathbb{C}^m$. Then the following statements are equivalent.*

(i) The range of T equals $\mathbb{C}^m$.

(ii) The rank of T equals m.

(iii) The kernel or null space of T^* is zero.

(iv) The operator TT^* is invertible.

(v) The operator TT^* is strictly positive.

(vi) The equation $g = Tf$ has a solution for all g in $\mathbb{C}^m$.

Finally, assume that any one of the above conditions hold and g is a specified vector in $\mathbb{C}^m$. Then a solution to the equation $g = Tf$ is given by

$$f = T^*(TT^*)^{-1}g\,. \tag{4.1}$$

Here we will not provide a proof of this lemma. However, if any one of the six conditions in Lemma 9.4.2 hold, then it is easy to verify that $f = T^*(TT^*)^{-1}g$ is a solution to the equation $g = Tf$. To see this simply observe that $T(T^*(TT^*)^{-1}g) = g$. Finally, it noted that the solution to equation $Tf = g$ in Lemma 9.4.2 may not be unique. The solution is unique if and only if the null space of T is zero.

An application of Lemma 9.4.2. For an example of how to use Lemma 9.4.2, consider the system

$$\begin{bmatrix} 40 \\ 20 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 2 \\ 2 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix} . \tag{4.2}$$

In this setting

$$T = \begin{bmatrix} 1 & 1 & 0 & 2 \\ 2 & 0 & 1 & 1 \end{bmatrix} \quad \text{and} \quad g = \begin{bmatrix} 40 \\ 20 \end{bmatrix} .$$

Notice that

$$TT^* = \begin{bmatrix} 6 & 4 \\ 4 & 6 \end{bmatrix} \quad \text{and} \quad (TT^*)^{-1} = \frac{1}{10} \begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix} .$$

Obviously, TT^* is invertible. Because TT^* is invertible, the system of equations in (4.2) has a solution. Moreover, a solution is given by

$$\begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix} = T^*(TT^*)^{-1}g = \frac{1}{10} \begin{bmatrix} 1 & 2 \\ 1 & 0 \\ 0 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix} \begin{bmatrix} 40 \\ 20 \end{bmatrix} = \begin{bmatrix} 4 \\ 8 \\ -2 \\ 14 \end{bmatrix} .$$

Hence $f = [4, 8, -2, 14]^*$ is a solution to the equation $g = Tf$. Finally, it is noted that the solution to this equation $g = Tf$ is not unique. The set of all solutions to this equations are given by $f + \xi$ where ξ is a vector in the null space of T.

9.4.2 Controllability

In this section we introduce the concept of controllability for discrete time systems. Consider the system given by

$$x(n+1) = Ax(n) + Bu(n) \qquad (x(0) = x_0) . \tag{4.3}$$

Here A is an operator on $\mathcal{X}$ and B is an operator from $\mathcal{U}$ into $\mathcal{X}$. Assume that n is an integer greater than or equal to the dimension of $\mathcal{X}$. Then this system or the pair $\{A, B\}$ is *controllable* over the interval $[0, n]$ if given any x_1 in $\mathcal{X}$, then there exists an input sequence $\{u(k)\}_0^{n-1}$ such that $x_1 = x(n)$. In other words, the system is controllable, if given any initial condition $x(0) = x_0$ and any x_1 in $\mathcal{X}$, then there exists a control sequence $\{u(k)\}_{k=0}^{n-1}$ which drives the initial condition $x(0) = x_0$ to the terminal state $x(n) = x_1$ at time n.

Let x_1 be any specified vector in $\mathcal{X}$. According to (3.2), the state

$$x(n) = A^n x_0 + \begin{bmatrix} B & AB & A^2B & \cdots A^{n-1}B \end{bmatrix} \begin{bmatrix} u(n-1) \\ u(n-2) \\ \vdots \\ u(0) \end{bmatrix} . \tag{4.4}$$

Notice that $x(n) = x_1$ if and only if $x_1 - A^n x_0$ is in the range of the operator

$$T_n = \begin{bmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{bmatrix} . \tag{4.5}$$

In other words, $x(n) = x_1$ if and only if the equation $x_1 - A^n x_0 = T_n f$ has a solution. Recall that the system is controllable over the interval $[0, n]$ if given any x_0 and x_1, then there exists a control sequence $\{u(k)\}_0^{n-1}$ such that $x(n) = x_1$. Because x_1 can be arbitrary, it follows that the pair $\{A, B\}$ is controllable if and only if the equation $x_1 - A^n x_0 = T_n f$ has a solution for all x_1 in $\mathcal{X}$, or equivalently, $g = T_n f$ has a solution for all g in $\mathcal{X}$. In other words, the pair $\{A, B\}$ is controllable over the interval $[0, n]$ if and only if the range of the operator T_n in (4.5) equals the whole space $\mathcal{X}$. By employing Lemma 9.4.2, we see that the system is controllable if and only if $T_n T_n^*$ is invertible. Moreover, in this case a control sequence $\{u(k)\}_0^{n-1}$ which drives the system from the initial condition $x(0) = x_0$ to a final state $x_1 = x(n)$ at time n is given by

$$\begin{bmatrix} u(n-1) \\ u(n-2) \\ \vdots \\ u(0) \end{bmatrix} = T_n^*(T_n T_n^*)^{-1}(x_1 - A^n x_0) . \tag{4.6}$$

To verify that this control works simply substitute $T_n^*(T_n T_n^*)^{-1}(x_1 - A^n f)$ into (4.4).

Now let us show that our definition of controllability is independent of the interval $[0, n]$ as long as $n \geq m$ where m is the dimension of $\mathcal{X}$. So assume that $n \geq m$. Notice that T_m is contained in the first m block columns of T_n, that is, T_n is a matrix of the form

$$T_n = \begin{bmatrix} T_m & A^m B & A^{m+1}B & \cdots & A^{n-1}B \end{bmatrix} .$$

By the Cayley-Hamilton theorem for all integers $\nu \geq m$ the operator A^ν is a linear combination of $\{A^k\}_{k=0}^{m-1}$. In particular, this implies that the last $n - m$ columns of T_n are linear combinations of the first m columns of T_n. So according to the Cayley-Hamilton theorem, T_n and T_m have the same rank. In other words, the pair $\{A, B\}$ is controllable over the interval $[0, n]$ if and only if the rank of T_m equals m, or equivalently, the pair $\{A, B\}$ is controllable over the interval $[0, m]$. In other words, the notion of controllability is independent of the interval as long as $n \geq m$. So from now on we will drop the interval when referring to controllability. Summing up this analysis along with Lemma 9.4.2, we obtain the following basic controllability result.

THEOREM 9.4.3 *Consider the pair $\{A$ on $\mathcal{X}, B\}$ where m is the dimension of $\mathcal{X}$. Let T_n be the controllability matrix defined in (4.5). Then the following statements are equivalent.*

(i) The pair $\{A, B\}$ is controllable.

(ii) The range of T_m equals $\mathcal{X}$.

(iii) The kernel or null space of the operator T_m^* is zero.

(iv) The operator $T_m T_m^*$ is invertible.

(v) The rank of T_m equals m.

(vi) The null space of the operator T_n^* is zero for any integer $n \geq m$.

(vii) The rank of T_n equals m for any integer $n \geq m$.

(viii) The operator $T_n T_n^*$ is invertible for any integer $n \geq m$.

Finally, if the pair $\{A, B\}$ is controllable, $n \geq m$ and x_1 is any specified vector in $\mathcal{X}$, then a control sequence $\{u(k)\}_{k=0}^{n-1}$ which drives the initial condition $x(0) = x_0$ to the terminal state $x(n) = x_1$ over the interval $[0, n]$ is given by (4.6).

The following result known as the PBH test due to Popov, Belevitch, and Hautus is useful in applications.

LEMMA 9.4.4 *The pair $\{A$ on $\mathbb{C}^m, B\}$ is controllable if and only if*

$$rank \begin{bmatrix} A - \lambda I & B \end{bmatrix} = m \tag{4.7}$$

for all λ in $\mathbb{C}$.

For a proof of Lemma 9.4.4 see [8, 22, 28]

An example. Consider the discrete time system

$$\begin{bmatrix} x_1(n+1) \\ x_2(n+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(n)\,. \tag{4.8}$$

In this case, the matrix A and B are given by

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Let us compute a control sequence $\{u(k)\}_0^3$ which drives the system from $x(0) = x_0 = [1, 1]^*$ to the final state $x(4) = x_1 = [2, 1]^*$. To this end, observe that

$$T_2 = \begin{bmatrix} B & AB \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}.$$

Notice that T_2 is invertible. So the rank of T_2 equals 2. In other words, the pair $\{A, B\}$ is controllable.

Now let us compute a sequence $\{u(k)\}_0^3$ which drives the initial condition x_0 to $x(4) = x_1$. To this end, notice that

$$T_4 = \begin{bmatrix} B & AB & A^2B & A^3B \end{bmatrix} = \begin{bmatrix} 1 & 1 & 3 & 7 \\ 1 & 3 & 7 & 17 \end{bmatrix}.$$

A simple calculation in Matlab shows that

$$\begin{bmatrix} u(3) \\ u(2) \\ u(1) \\ u(0) \end{bmatrix} = T_4^*(T_4T_4^*)^{-1}(x_1 - A^4x_0) = \begin{bmatrix} 2.08 \\ -1.25 \\ -0.42 \\ -2.08 \end{bmatrix}.$$

In other words, the input sequence

$$u(0) = -2.08, \quad u(1) = -0.42, \quad u(2) = -1.25 \quad \text{and} \quad u(3) = 2.08$$

drives the initial state $x(0) = x_0 = [1, 1]^*$ to the final state $x(4) = [2, 1]^*$.

9.4.3 Exercise

Problem 1. Consider the state space system given by

$$\begin{bmatrix} x_1(n+1) \\ x_2(n+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1/6 & 5/6 \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} + \begin{bmatrix} 6 \\ 6 \end{bmatrix} u(n)\,.$$

Is this system controllable? Compute a control sequence $\{u(k)\}_0^3$ which drives this system from the initial condition $x(0) = x_0 = [1, 2]^*$ to the final state $x(4) = x_1 = [3, 4]^*$. Is this control sequence the only control input which drives this system from $x(0) = x_0 = [1, 2]^*$ to $x(4) = x_1 = [3, 4]^*$?

9.5 Discrete time Lyapunov equations

In this section we will introduce and study discrete time Lyapunov equations. In the discrete time setting a Lyapunov equation is an equation of the form

$$P = APA^* + Q \tag{5.1}$$

where A and Q are specified operators on $\mathcal{X}$. If there exists an operator P on $\mathcal{X}$ satisfying $P = APA^* + Q$, then P is a solution to this Lyapunov equation. The Matlab command to solve the discrete time Lyapunov equation in (5.1) is dlyap.

PROPOSITION 9.5.1 *Let A be a discrete time stable operator on $\mathcal{X}$ and Q an operator on $\mathcal{X}$. Then there exists a unique solution P on $\mathcal{X}$ to the Lyapunov equation in (5.1). Moreover, this solution P is given by*

$$P = \sum_{k=0}^{\infty} A^k Q A^{*k}\,. \tag{5.2}$$

PROOF. Because A is stable, the series in (5.2) converges. Hence $P = \sum_0^\infty A^k Q A^{*k}$ is a well defined operator on $\mathcal{X}$. Using this P, we obtain

$$\begin{aligned} P &= \sum_{k=0}^{\infty} A^k Q A^{*k} = Q + \sum_{k=1}^{\infty} A^k Q A^{*k} \\ &= Q + A\left(\sum_{k=0}^{\infty} A^k Q A^{*k}\right) A^* = Q + APA^* . \end{aligned}$$

Therefore $P = \sum_0^\infty A^k Q A^{*k}$ is a solution to the Lyapunov equation $P = APA^* + Q$. In particular, there exists a solution to the Lyapunov equation in (5.1).

To show that the solution to the Lyapunov equation (5.1) is unique, let P be any solution to the Lyapunov equation (5.4). Then by repeatedly replacing P by $Q + APA^*$, we have

$$\begin{aligned} P &= Q + APA^* = Q + A\left(Q + APA^*\right) A^* \\ &= Q + AQA^* + A^2 P A^{*2} = Q + AQA^* + A^2\left(Q + APA^*\right) A^{*2} \\ &= Q + AQA^* + A^2 Q A^{*2} + A^3 P A^{*3} . \end{aligned}$$

By continuing in this fashion, we obtain

$$P = \sum_{k=0}^{n} A^k Q A^{*k} + A^{n+1} P A^{*n+1} , \tag{5.3}$$

where n is any positive integer. Since A is stable, $A^{n+1} P A^{*n+1}$ converges to zero as n tends to infinity. So by taking the limit as n approach infinity in (5.3), we see that $P = \sum_0^\infty A^k Q A^{*k}$ is the unique solution to the Lyapunov equation in (5.1). This completes the proof.

The following result plays a fundamental role in studying stable controllable systems.

THEOREM 9.5.2 *Consider the pair $\{A, B\}$ where A is a discrete time stable operator on $\mathcal{X}$ and B is an operator from $\mathcal{U}$ into $\mathcal{X}$. Then there is a unique solution P on $\mathcal{X}$ to the Lyapunov equation*

$$P = APA^* + BB^* . \tag{5.4}$$

Moreover, this solution P is given by

$$P = \sum_{k=0}^{\infty} A^k B B^* A^{*k} . \tag{5.5}$$

Finally, P is a positive operator.

PROOF. By employing $Q = BB^*$ in Proposition 9.5.1, we see that there exists a unique solution P to the Lyapunov equation in (5.4). Moreover, this solution is given by $P = \sum_0^\infty A^k B B^* A^{*k}$. To complete the proof, it remains to show that P is a positive operator. To this end, let f be any vector in $\mathcal{X}$. Using $P = \sum_0^\infty A^k B B^* A^{*k}$, we have

$$(Pf, f) = \sum_{k=0}^{\infty} (A^k B B^* A^{*k} f, f) = \sum_{k=0}^{\infty} \|B^* A^{*k} f\|^2 \geq 0 . \tag{5.6}$$

Therefore P is positive. This completes the proof.

Let $\{A, B\}$ be a stable pair, that is, assume that A is a stable operator on $\mathcal{X}$ and B maps $\mathcal{U}$ into $\mathcal{X}$. Then the unique solution to the Lyapunov equation in (5.4) is called the *controllability Gramian* for the pair $\{A, B\}$.

COROLLARY 9.5.3 *Let P be the controllability Gramian for a stable pair $\{A, B\}$. Then P is strictly positive if and only if the pair $\{A, B\}$ is controllable.*

PROOF. Let f be any vector in $\mathcal{X}$. By consulting (5.6), we have

$$(Pf, f) = \sum_{k=0}^{\infty} \|B^* A^{*k} f\|^2 \qquad (f \in \mathcal{X}) .$$

Hence $(Pf, f) = 0$ if and only if $\|B^* A^{*k} f\| = 0$ for all integers $k \geq 0$. This readily implies that $(Pf, f) = 0$ if and only if $B^* A^{*k} f = 0$ for all $k \geq 0$. According to the Cayley-Hamilton theorem $B^* A^{*k} f = 0$ for all $k \geq 0$ if and only if $B^* A^{*k} f = 0$ for all $k = 0, 1, 2, \cdots, m-1$ where m is the dimension of $\mathcal{X}$. Notice that the adjoint T_m^* of the controllability matrix T_m in (4.5) is given by

$$T_m^* = \begin{bmatrix} B^* \\ B^* A^* \\ B^* A^{*2} \\ \vdots \\ B^* A^{*m-1} \end{bmatrix} . \tag{5.7}$$

Hence $(Pf, f) = 0$ if and only if $T_m^* f = 0$. Recall that P is strictly positive if $(Pg, g) = 0$ implies that $g = 0$. Thus P is strictly positive if and only if T_m^* is one to one. By consulting Part (*iii*) of Theorem 9.4.3, we see that the pair $\{A, B\}$ is controllable if and only if P is strictly positive. This completes the proof.

THEOREM 9.5.4 *Let $\{A, B\}$ be a controllable pair. Then A is stable if and only if there exists a strictly positive solution P to the Lyapunov equation*

$$P = APA^* + BB^* . \tag{5.8}$$

In this case, the solution P to the Lyapunov equation (5.8) is unique.

PROOF. If A is stable, then Theorem 9.5.2 and Corollary 9.5.3 show that the solution to the Lyapunov equation (5.8) is unique and strictly positive.

Now assume that P is a strictly positive solution to the Lyapunov equation in (5.8). Let λ be an eigenvalue of A^* with corresponding eigenvector f, that is, $A^* f = \lambda f$ where f is a nonzero vector. Using (5.8), we have

$$(Pf, f) = (APA^* f, f) + (BB^* f, f) = (PA^* f, A^* f) + \|B^* f\|^2 = |\lambda|^2 (Pf, f) + \|B^* f\|^2 .$$

Hence $(Pf, f) = |\lambda|^2 (Pf, f) + \|B^* f\|^2$, or equivalently,

$$0 \leq \|B^* f\|^2 = (1 - |\lambda|^2)(Pf, f) .$$

Because (Pf, f) is strictly positive, $|\lambda| \leq 1$. We claim that $|\lambda| < 1$. Let us proceed by contradiction and assume that $|\lambda| = 1$. Then $\|B^* f\|^2 = 0$, or equivalently, $B^* f = 0$. Hence $B^* A^{*k} f = B^* \lambda^k f = \lambda^k B^* f = 0$ for all integers $0 \leq k \leq m-1$ where m is the dimension of $\mathcal{X}$. By consulting the formula for T_m^* in (5.7), we see that $T_m^* f = 0$. Because the pair $\{A, B\}$ is controllable, $f = 0$; see Part (iii) of Theorem 9.4.3. However, f is an eigenvector and cannot be zero by definition. This is a contradiction which arose from our assumption that $|\lambda| = 1$. Thus $|\lambda| < 1$ and A^* is stable. Therefore A is stable. This completes the proof.

9.6 Observability of discrete time systems

In this section we review some elementary facts concerning the notion of observability for discrete time systems.

9.6.1 A fundamental lemma from linear algebra

Recall that an operator T is one to one if the null space of T is zero. The operator T is onto if the range of T equals the whole space. The following result from linear algebra is the dual of Theorem 9.4.2 and will also be useful in studying observability.

LEMMA 9.6.1 *Let W be an operator from $\mathbb{C}^m$ into $\mathbb{C}^k$. Then the following statements are equivalent.*

(i) The kernel or null space of W is zero.

(ii) The rank of W equals m.

(iii) The range of W^* equals $\mathbb{C}^m$.

(iv) The operator W^*W is invertible.

(v) The operator W^*W is strictly positive.

(vi) If the equation $g = Wf$ has a solution for a specified vector g in $\mathbb{C}^k$, then the solution f is unique.

Finally, assume that any one of the above conditions hold, and the equation $g = Wf$ has a solution for a specified vector g in $\mathbb{C}^k$. Then the solution to the equation $g = Wf$ is unique and given by

$$f = (W^*W)^{-1} W^* g \,. \tag{6.1}$$

Here we will not provide a proof of this lemma. However, if any one of the six conditions in Lemma 9.6.1 hold and the equation $g = Wf$ has a solution, then it is easy to verify that $f = (W^*W)^{-1}W^*g$ is the solution to the equation $g = Wf$. To see this notice that multiplying by W^* on left of $g = Wf$, yields $W^*g = W^*Wf$. Therefore $f = (W^*W)^{-1}W^*g$. Finally, it noted that the equation $Wf = g$ in Lemma 9.6.1 does not always have a solution. The equation $g = Wf$ has a solution for all g in $\mathbb{C}^k$ if and only if the range of W equals $\mathbb{C}^k$.

9.6.2 Observability

In this section we introduce the concept of observability for discrete time systems. Consider the system given by

$$x(n+1) = Ax(n) \qquad \text{and} \qquad y(n) = Cx(n) \tag{6.2}$$

where the initial condition $x(0) = x_0$. Here A is an operator on $\mathcal{X}$ and C maps $\mathcal{X}$ into $\mathcal{Y}$. Assume that n is an integer greater than or equal to the dimension of $\mathcal{X}$. Then this system or the pair $\{C, A\}$ is *observable* over the interval $[0, n)$ if given the output $\{y(k)\}_0^{n-1}$, then one can uniquely determine the state trajectory $\{x(k)\}_0^{n-1}$. Since $x(k) = A^k x_0$, we see that the pair $\{C, A\}$ is *observable* over the interval $[0, n)$ if given the output $\{y(k)\}_0^{n-1}$, then one can uniquely determine the initial state $x(0) = x_0$.

Recall that the solution to the state space system in (6.2) is given by $y(k) = CA^k x_0$. In other words,

$$\begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(n-1) \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} x_0 \,. \tag{6.3}$$

Now let W_n be the operator defined by

$$W_n = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} . \tag{6.4}$$

Hence the equation in (6.3) is equivalent to $g = W_n x_0$ where g is the transpose of the vector $[y(0), y(1), \cdots, y(n-1)]$. Clearly, the equation $g = W_n x_0$ has a solution. Notice that the pair $\{C, A\}$ is observable if and only if the equation $g = W_n x_0$ has a unique solution. In other words, $\{C, A\}$ is observable over the interval $[0, n)$ if and only if the null space of W_n is zero. So according to Lemma 9.6.1, the pair $\{C, A\}$ is observable over the interval $[0, n)$ if and only if the rank of W_n equals m which is the dimension of $\mathcal{X}$. In this case, Lemma 9.6.1 shows that the initial condition is uniquely determined by

$$x_0 = (W_n^* W_n)^{-1} W_n^* \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(n-1) \end{bmatrix} = (W_n^* W_n)^{-1} \sum_{k=0}^{n-1} A^{*k} C^* y(k) \,. \tag{6.5}$$

Now let us show that our definition of observability is independent of the interval $[0, n)$ as long as $n \geq m$ where m is the dimension of $\mathcal{X}$. So assume that $n \geq m$. Notice that W_m

is contained in the first m block rows of W_n, that is, W_n is a matrix of the form

$$W_n = \begin{bmatrix} W_m \\ CA^m \\ \vdots \\ CA^{n-1} \end{bmatrix}.$$

By the Cayley-Hamilton theorem for all integers $\nu \geq m$ the operator A^ν is a linear combination of $\{A^k\}_{k=0}^{m-1}$. In particular, this implies that the last $n-m$ rows of W_n are linear combinations of the first m rows of W_n. So according to the Cayley-Hamilton theorem, W_n and W_m have the same rank. In other words, the pair $\{C, A\}$ is observable over the interval $[0, n)$ if and only if the rank of W_m equals m, or equivalently, the pair $\{C, A\}$ is observable over the interval $[0, m)$. In other words, the notion of observability is independent of the interval as long as $n \geq m$. So from now on we will drop the interval when referring to observability. Summing up this analysis along with Lemma 9.6.1 we obtain the following basic observability result.

THEOREM 9.6.2 *Consider the pair $\{C, A$ on $\mathcal{X}\}$ where m is the dimension of $\mathcal{X}$. Let W_n be the observability operator defined in (6.4). Then the following statements are equivalent*

(i) The pair $\{C, A\}$ is observable.

(ii) The kernel or null space of the operator W_m is zero.

(iii) The range of W_m^* equals $\mathcal{X}$.

(iv) The operator $W_m^* W_m$ is invertible.

(v) The rank of W_m equals m.

(vi) The null space of the operator W_n is zero for any integer $n \geq m$.

(vii) The rank of W_n equals m for any integer $n \geq m$.

(viii) The operator $W_n^* W_n$ is invertible for any integer $n \geq m$.

Finally, if the pair $\{C, A\}$ is observable, $n \geq m$, and the output $y(k) = Cx(k)$ for $k = 0, 1, 2, \cdots, n-1$, then the initial condition $x(0) = x_0$ uniquely determined by $\{y(k)\}_{k=0}^{n-1}$ is given by (6.5).

By comparing Theorems 9.4.3 and 9.6.2, we see that the pair $\{A, B\}$ is controllable if and only if the pair $\{B^*, A^*\}$ is observable. Motivated by this we say that controllability is the dual of observability.

Clearly, the pair $\{C, A\}$ is observable if and only if the pair $\{A^*, C^*\}$ is controllable. Recall that rank of a matrix equals the rank of its transpose. By employing this duality in the PBH Lemma 9.4.4, we obtain the following PBH observability result due to Popov, Belevitch, and Hautus.

LEMMA 9.6.3 *Consider the pair $\{C, A \text{ on } \mathcal{X}\}$ and let Γ_λ be the operator defined by*

$$\Gamma_\lambda = \begin{bmatrix} A - \lambda I \\ C \end{bmatrix} : \mathcal{X} \to \begin{bmatrix} \mathcal{X} \\ \mathcal{Y} \end{bmatrix}.$$

Then the pair $\{C, A\}$ is observable if and only if the null space of Γ_λ is zero for all complex numbers λ.

The following result is a simple application of the PBH test.

PROPOSITION 9.6.4 *Consider the pair of matrices*

$$A = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix} \text{ on } \mathbb{C}^m \quad \text{and} \quad C = \begin{bmatrix} c_1 & c_2 & \cdots & c_m \end{bmatrix} : \mathbb{C}^m \to \mathbb{C} \tag{6.6}$$

where $\{\lambda_k\}_1^m$ and $\{c_k\}_1^m$ are scalars. Then $\{C, A\}$ is observable if and only if $\{\lambda_k\}_1^m$ are distinct and $c_i \neq 0$ for $i = 1, 2, \cdots, m$.

Proof. For any scalar λ the matrix Γ_λ in the PBH Lemma is given by

$$\Gamma_\lambda = \begin{bmatrix} \lambda_1 - \lambda & 0 & \cdots & 0 \\ 0 & \lambda_2 - \lambda & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m - \lambda \\ c_1 & c_2 & \cdots & c_m \end{bmatrix}.$$

If $c_i = 0$, then with $\lambda = \lambda_i$, the i-th column of Γ_λ is zero. Hence, Γ_λ has a nontrivial kernel and by the PBH Lemma, $\{C, A\}$ is unobservable. If $\lambda_i = \lambda_j$, then with $\lambda = \lambda_i$, the i-th and j-th columns are linearly dependent. By the PBH Lemma, $\{C, A\}$ is unobservable.

On the other hand, assume that $\{\lambda_j\}_1^m$ are distinct and $c_i \neq 0$ for $i = 1, 2, \cdots, m$. Clearly, the columns of Γ_λ are linearly independent when λ is not an eigenvalue of A. Now consider $\lambda = \lambda_i$. Because $c_i \neq 0$, the i-th column of Γ_λ is linearly independent of the other $m - 1$ linearly independent columns. So the kernel of Γ_λ is zero for all λ. According the PBH Lemma, $\{C, A\}$ is observable. This completes the proof.

9.6.3 Exercise

Problem 1. Consider the state space system given by

$$\begin{bmatrix} x_1(n+1) \\ x_2(n+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} \quad \text{and} \quad y(n) = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix}.$$

Is this system observable? Now assume that the output is given by

$$y(0) = 3, \quad y(1) = 5, \quad y(2) = 13 \quad \text{and} \quad y(3) = 31\,.$$

Then find the initial condition $x(0) = x_0$ and the state trajectory $\{x(k)\}_0^3$ corresponding to this output sequence. Is the state $\{x(k)\}_0^3$ uniquely determined by the output $\{y(k)\}_0^3$.

Problem 2. Consider the following state space system determined by the Gambler's ruin problem

$$\begin{bmatrix} x_1(n+1) \\ x_2(n+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -q/p & 1/p \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} \quad \text{and} \quad y(n) = \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix}. \tag{6.7}$$

Recall that $p + q = 1$ and $p > 0$. Notice that in this case the output $y(n) = p_{n+1} - p_n$ where $p_n = P(W_{n,m})$. Show that the system in (6.7) is not observable.

9.7 Lyapunov equations and observability

.

Recall that the pair $\{C, A\}$ is observable if and only if $\{A^*, C^*\}$ is controllable. So by replacing A by A^* and B by C^* in Theorem 9.5.2, we readily obtain the following result.

THEOREM 9.7.1 *Consider the pair $\{C, A\}$ where A is a discrete time stable operator on $\mathcal{X}$ and C is an operator from $\mathcal{X}$ into $\mathcal{Y}$. Then there is a unique solution P on $\mathcal{X}$ to the Lyapunov equation*

$$P = A^*PA + C^*C\,. \tag{7.1}$$

Moreover, this solution P is given by

$$P = \sum_{k=0}^{\infty} A^{*k}C^*CA^k\,. \tag{7.2}$$

Finally, P is a positive operator.

Assume that A is a stable operator on $\mathcal{X}$ and C maps $\mathcal{X}$ into $\mathcal{Y}$. Then the unique solution to the Lyapunov equation in (7.1) is called the *observability Gramian* for the pair $\{C, A\}$. By replacing A by A^* and B by C^* in Corollary 9.5.3 we readily obtain the following result.

COROLLARY 9.7.2 *Let P be the observability Gramian for a stable pair $\{C, A\}$. Then P is strictly positive if and only if the pair $\{C, A\}$ is observable.*

Recall that an operator A is stable if and only if its adjoint A^* is stable. By replacing A by A^* and B by C^* in Theorem 9.5.4, we readily obtain the following result.

THEOREM 9.7.3 *Let $\{C, A\}$ be an observable pair. Then A is stable if and only if there exists a strictly positive solution P to the Lyapunov equation*

$$P = A^*PA + C^*C\,. \tag{7.3}$$

In this case, the solution P to the Lyapunov equation (7.3) is unique.

9.8 An observability optimization problem

In this section we will introduce an observability optimization problem. The results in this section are the discrete time version of the continuous time observability optimization results in Section 1.4 in Chapter 1. To begin, consider the discrete time system given by

$$x(n+1) = Ax(n) \qquad \text{and} \qquad y(n) = Cx(n) \tag{8.1}$$

where the initial condition $x(0) = x_0$. Here A is an operator on $\mathcal{X} = \mathbb{C}^m$ and C maps $\mathcal{X}$ into $\mathcal{Y}$. Recall that the output $y(n) = CA^n x_0$. Let $\{f_k\}_0^{n-1}$ be any sequence of vectors such that $f_k \in \mathcal{Y}$ for all $k = 0, 1, 2, \cdots, n-1$. Then the observability optimization problem is to find an optimal initial condition $\widehat{x}_0$ such that $y(k) = CA^k\widehat{x}_0$ comes as close as possible to the specified sequence of vectors $\{f_k\}_0^{n-1}$. To be precise, let f and h be the vectors defined by

$$f = \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \end{bmatrix} \quad \text{and} \quad h = \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(n-1) \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} x_0 = W_n x_0 \,. \tag{8.2}$$

Here W_n is the observability operator defined in (6.4). Then the observability optimization problem is to find an initial condition $\widehat{x}_0$ to minimize the distance between f and $W_n x_0$. In other words, find an initial condition $\widehat{x}_0$ such that

$$d = \|f - W_n\widehat{x}_0\| = \inf\{\|f - W_n x_0\| : x_0 \in \mathbb{C}^m\} \,. \tag{8.3}$$

Here $d = \|f - W_n\widehat{x}_0\|$ is the error in this optimization problem. Finally, if $\widehat{x}_0$ is an optimal initial condition, then the estimate $\widehat{x}(n)$ of the state $x(n)$ at time n is defined by $\widehat{x}(n) = A^n\widehat{x}_0$. The following result provides a solution to this problem.

THEOREM 9.8.1 *Consider the observability optimization problem in (8.3) for the pair $\{C, A\}$ where f is a specified vector of the form in (8.2). Let W_n be the observability operator defined in (6.4). Then the following holds.*

(i) There exists a solution $\widehat{x}_0$ in $\mathbb{C}^m$ to the linear equation

$$(W_n^* W_n)\widehat{x}_0 = W_n^* f = \sum_{k=0}^{n-1} A^{*k} C^* f_k \,. \tag{8.4}$$

(ii) If $\widehat{x}_0$ in $\mathbb{C}^m$ is any solution to (8.4), then $\widehat{x}_0$ is an initial condition solving the observability optimization problem in (8.3). In this case, the error

$$d^2 = \|f\|^2 - (W_n^* W_n \widehat{x}_0, \widehat{x}_0) \,. \tag{8.5}$$

(iii) If the pair $\{C, A\}$ is observable and $n \geq m$, then there exists a unique solution to the observability optimization problem in (8.3) and this unique solution is given by

$$\widehat{x}_0 = (W_n^* W_n)^{-1} W_n^* f = (W_n^* W_n)^{-1} \sum_{k=0}^{n-1} A^{*k} C^* f_k \,. \tag{8.6}$$

In this case, the state estimate $\widehat{x}(n) = A^n \widehat{x}_0$.

PROOF. Parts (i) and (ii) follow by applying Theorem 1.5.1 in Chapter 1 to the optimization problem in (8.3). Theorem 9.6.2 shows that if the pair $\{C, A\}$ is observable and $n \geq m$, then $W_n^* W_n$ is invertible. Hence Part (iii) is also a consequence of Theorem 1.5.1 in Chapter 1. This completes the proof.

The following result, known as the matrix inversion lemma, is used to obtain a recursive solution to the observability optimization problem in Theorem 9.8.1.

LEMMA 9.8.2 *Let* T *be an operator on* $\mathcal{X}$ *and* M *an operator from* $\mathcal{U}$ *into* $\mathcal{X}$ *and* N *an operator from* $\mathcal{X}$ *into* $\mathcal{U}$. *Then*

$$(T + MN)^{-1} = T^{-1} - T^{-1}M\left(I + NT^{-1}M\right)^{-1} NT^{-1}. \tag{8.7}$$

PROOF. We will use the following matrix identity

$$Q\,(I + PQ)^{-1} = (I + QP)^{-1}\,Q \tag{8.8}$$

where Q and P are operators acting between the appropriate spaces. To verify that this identity holds simply observe that $(I + QP)Q = Q(I + PQ)$. By taking inverses we arrive at formula in (8.8). Now using (8.8), we obtain

$$\begin{aligned}
(T + MN)^{-1} &= T^{-1}\left(I + MNT^{-1}\right)^{-1} \\
&= T^{-1}\left(I + MNT^{-1} - MNT^{-1}\right)\left(I + MNT^{-1}\right)^{-1} \\
&= T^{-1} - T^{-1}MNT^{-1}\left(I + MNT^{-1}\right)^{-1} \\
&= T^{-1} - T^{-1}M\left(I + NT^{-1}M\right)^{-1} NT^{-1}.
\end{aligned}$$

The fourth equality follows from the operator identity in (8.8) with $Q = NT^{-1}$. This yields (8.7) and completes the proof.

Now assume that the pair $\{C, A \text{ on } \mathbb{C}^m\}$ is observable. Theorem 9.8.1 shows that the state estimate $\widehat{x}(n)$ for $x(n)$ corresponding to the optimal initial condition $\widehat{x}_0$ is given by

$$\widehat{x}(n) = A^n (W_n^* W_n)^{-1} \sum_{k=0}^{n-1} A^{*k} C^* f_k \qquad (\text{if } n \geq m)\,. \tag{8.9}$$

If $f_k = CA^k x_0$, then $\widehat{x}_0 = x_0$ and $\widehat{x}(k) = x(k)$ for all $k \geq 0$. In many applications $\widehat{x}(k)$ is used to estimate the state $x(k)$ when the output measurement $f_k = Cx(k) + w_k$ is corrupted by noise w_k. Riccati difference equations play an important role in Kalman filtering and state estimation. The following result uses a Riccati difference equation to compute the state estimate $\widehat{x}$ for x.

THEOREM 9.8.3 *Consider the observability optimization problem in (8.3) where the pair* $\{C, A \text{ on } \mathbb{C}^m\}$ *is observable and* f *is a vector of the form (8.2). Let* W_n *be the observability*

operator defined in (6.4). Then the state estimate $\widehat{x}(n) = A^n\widehat{x}_0$ for $n \geq m$ can be computed recursively by the following formula

$$\begin{aligned} \widehat{x}(n+1) &= A\widehat{x}(n) + \Delta_n\left(f_n - C\widehat{x}(n)\right) \\ \Delta_n &= AQ_nC^*(I + CQ_nC^*)^{-1} \end{aligned} \tag{8.10}$$

where Q_n is the solution to the discrete time Riccati equation

$$Q_{n+1} = AQ_nA^* - AQ_nC^*(I + CQ_nC^*)^{-1}CQ_nA^* \tag{8.11}$$

subject to the initial condition $Q_m = A^m(W_m^*W_m)^{-1}A^{*m}$.

PROOF. Set $g_n = f = [f_0, f_1, \cdots, f_{n-1}]^{tr}$ where tr denotes the transpose. According to Theorem 9.8.1, the state estimate $\widehat{x}(n)$ is given by

$$\widehat{x}(n) = A^n(W_n^*W_n)^{-1}W_n^*g_n\,.$$

Now assume that $n \geq m$ and let Q_n be the matrix defined by

$$Q_n = A^n(W_n^*W_n)^{-1}A^{*n} \qquad (n \geq m)\,.$$

Because the pair $\{C, A\}$ is observable, the inverse of $W_n^*W_n$ exists. Set $G_n = (W_n^*W_n)^{-1}$. Notice that $Q_n = A^nG_nA^{*n}$. Using $W_{n+1} = [W_n, CA^n]^{tr}$, we have

$$W_{n+1}^*W_{n+1} = W_n^*W_n + A^{*n}C^*CA^n \quad \text{and} \quad W_{n+1}^*g_{n+1} = W_n^*g_n + A^{*n}C^*f_n\,.$$

By employing (8.7) in the matrix inversion Lemma 9.8.2, we obtain

$$\begin{aligned} \widehat{x}(n+1) &= A^{n+1}(W_{n+1}^*W_{n+1})^{-1}W_{n+1}^*g_{n+1} = A^{n+1}(W_n^*W_n + A^{*n}C^*CA^n)^{-1}W_{n+1}^*g_{n+1} \\ &= A^{n+1}G_nW_{n+1}^*g_{n+1} - A^{n+1}G_nA^{*n}C^*\left(I + CA^nG_nA^{*n}C^*\right)^{-1}CA^nG_nW_{n+1}^*g_{n+1} \\ &= A^{n+1}G_n\left(W_n^*g_n + A^{*n}C^*f_n\right) \\ &\quad -AQ_nC^*\left(I + CQ_nC^*\right)^{-1}CA^nG_n\left(W_n^*g_n + A^{*n}C^*f_n\right) \\ &= A\widehat{x}(n) + AQ_nC^*f_n - \Delta_nC\widehat{x}(n) - AQ_nC^*\left(I + CQ_nC^*\right)^{-1}CQ_nC^*f_n \\ &= A\widehat{x}(n) - \Delta_nC\widehat{x}(n) + AQ_nC^*\left[I - \left(I + CQ_nC^*\right)^{-1}CQ_nC^*\right]f_n \\ &= A\widehat{x}(n) - \Delta_nC\widehat{x}(n) + AQ_nC^*\left(I + CQ_nC^*\right)^{-1}\left(\left(I + CQ_nC^*\right) - CQ_nC^*\right)f_n \\ &= A\widehat{x}(n) - \Delta_nC\widehat{x}(n) + AQ_nC^*\left(I + CQ_nC^*\right)^{-1}f_n \\ &= A\widehat{x}(n) + \Delta_n\left(f_n - C\widehat{x}(n)\right)\,. \end{aligned}$$

This yields the state space equation in (8.10).

To complete the proof it remains to obtain the Riccati equation in (8.11). To this end, notice that the matrix inversion Lemma 9.8.2 gives

$$\begin{aligned} Q_{n+1} &= A^{n+1}(W_{n+1}^*W_{n+1})^{-1}A^{*n+1} = A^{n+1}(W_n^*W_n + A^{*n}C^*CA^n)^{-1}A^{*n+1} \\ &= A^{n+1}G_nA^{*n+1} - A^{n+1}G_nA^{*n}C^*\left(I + CA^nG_nA^{*n}C^*\right)^{-1}CA^nG_nA^{*n+1} \\ &= A^nQ_nA^* - AQ_nC^*\left(I + CQ_nC^*\right)^{-1}CQ_nA^*\,. \end{aligned}$$

This yields the Riccati difference equation in (8.11) and completes the proof.

9.8.1 The infinite horizon case

In this section we will study the infinite horizon observability optimization problem, that is, the observability optimization problem when $n = \infty$. In this case, f is a vector in $\ell^2_+(\mathcal{Y})$, that is, f is a vector of the form

$$f = \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \end{bmatrix} \qquad \text{and} \qquad \|f\|^2 = \sum_{k=0}^{\infty} \|f_k\|^2 < \infty \,. \tag{8.12}$$

Here $f_k \in \mathcal{Y}$ for all integers $k \geq 0$. Now consider the pair $\{C, A\}$ where A is stable. In this setting the infinite horizon observability optimization problem is given by

$$d^2 = \sum_{k=0}^{\infty} \|f_k - CA^k \widehat{x}_0\|^2 = \inf\{\sum_{k=0}^{\infty} \|f_k - CA^k x_0\|^2 : x_0 \in \mathbb{C}^m\} \,. \tag{8.13}$$

Here d^2 is the error in this optimization problem.

Recall that the observability Gramian P for $\{C, A\}$ is given by the unique solution to the discrete time Lyapunov equation

$$P = A^* P A + C^* C \,. \tag{8.14}$$

Moreover, P is also given by the series formula

$$P = \sum_{k=0}^{\infty} A^{*k} C^* C A^k \,. \tag{8.15}$$

Furthermore, that the pair $\{C, A\}$ is observable if and only if P is strictly positive; see Theorem 9.7.1 and Corollary 9.7.2. Notice that $W_n^* W_n = \sum_{k=0}^{n-1} A^{*k} C^* C A^k$. Because A is stable, we have

$$\lim_{n\to\infty} W_n^* W_n = \lim_{n\to\infty} \sum_{k=0}^{n-1} A^{*k} C^* C A^k = P \,.$$

In other words, $W_n^* W_n$ converges to the observability Gramian P as n tends to infinity. By letting $n = \infty$ in Theorem 9.8.1, we readily obtain the following result.

THEOREM 9.8.4 *Consider the observability optimization problem in (8.13) where $\{C, A\}$ is a stable pair and f in (8.12) is a specified function in $\ell^2_+(\mathcal{Y})$. Let P be the solution to the Lyapunov equation in (8.14). Then the following holds.*

(i) There exists a solution $\widehat{x}_0$ in $\mathbb{C}^n$ to the linear equation

$$P\widehat{x}_0 = \sum_{k=0}^{\infty} A^{*k} C^* f_k \,. \tag{8.16}$$

(ii) If $\widehat{x}_0$ in $\mathbb{C}^m$ is any solution to (8.16), then $\widehat{x}_0$ is an initial condition solving the observability optimization problem in (8.13). In this case, the error

$$d^2 = \sum_{k=0}^{\infty} \|f_k\|^2 - (P\widehat{x}_0, \widehat{x}_0)\,. \tag{8.17}$$

(iii) If the pair $\{C, A\}$ is observable, then there exists a unique solution to the observability optimization problem in (8.13) and this unique solution is given by

$$\widehat{x}_0 = P^{-1} \sum_{k=0}^{\infty} A^{*k} C^* f_k\,. \tag{8.18}$$

In this case, the state estimate $\widehat{x}(n) = A^n \widehat{x}_0$.

9.8.2 Exercise

Problem 1. Consider the system $x(n+1) = Ax(n)$ and $y(n) = Cx(n)$ where

$$A = \begin{bmatrix} 0 & 1 \\ -1/6 & 5/6 \end{bmatrix} \qquad \text{and} \qquad C = \begin{bmatrix} 1 & 6 \end{bmatrix}.$$

Is the pair $\{C, A\}$ observable? Find the optimal initial condition $\widehat{x}_0$ and the error d in the observability optimization problem

$$d^2 = \inf\{\sum_{k=0}^{\infty} |f_k - CA^k x_0|^2 : x_0 \in \mathbb{C}^2\}$$

where $f_k = (3/4)^k$ for all integers $k \geq 0$. Is your choice of $\widehat{x}_0$ unique? Finally, compute the error d. Finally, it is noted that dlyap is the Matlab command used to compute the solution to the discrete time Lyapunov equation.

Problem 2. Repeat Problem 1 with $f_k = 1/2^k$ for all integers $k \geq 0$.

Problem 3. Consider the system $x(n+1) = Ax(n)$ and $y(n) = Cx(n)$ where

$$A = \begin{bmatrix} 0 & 1 \\ -1/6 & 5/6 \end{bmatrix} \qquad \text{and} \qquad C = \begin{bmatrix} 3 & -6 \end{bmatrix}.$$

Is the pair $\{C, A\}$ observable? Find the all optimal initial conditions $\widehat{x}_0$ and the error d in the observability optimization problem

$$d^2 = \inf\{\sum_{k=0}^{\infty} |f_k - CA^k x_0|^2 : x_0 \in \mathbb{C}^2\}$$

where $f_k = (3/4)^k$ for all integers $k \geq 0$. Is your choice of $\widehat{x}_0$ unique? Finally, compute the error d.

9.9 Time varying state space systems

To complete this chapter we will introduce discrete time varying systems. Consider the time varying state space system

$$x(n+1) = A(n)x(n) + B(n)u(n) \qquad \text{and} \qquad y = C(n)x(n) + D(n)v(n)\,. \tag{9.1}$$

For every integer $n \geq 0$, the operator $A(n)$ is on $\mathcal{X}$ and $B(n)$ is an operator mapping $\mathcal{U}$ into $\mathcal{X}$ while $C(n)$ is an operator from $\mathcal{X}$ into $\mathcal{Y}$ and $D(n)$ is an operator mapping $\mathcal{V}$ into $\mathcal{Y}$. The state $x(n)$ is in $\mathcal{X}$, the input $u(n)$ is in $\mathcal{U}$, the input $v(n)$ is in $\mathcal{V}$ and the output $y(n)$ is in $\mathcal{Y}$ for all integers n. The initial condition $x(0) = x_0$. Throughout $\Psi(n,\nu)$ is the *state transition matrix* for $A(n)$ defined by

$$\begin{aligned} \Psi(n,\nu) &= A(n)A(n-1)\cdots A(\nu+1) && \text{if } n > \nu \\ &= I && \text{if } n = \nu\,. \end{aligned} \tag{9.2}$$

For example, $\Psi(4,1) = A(4)A(3)A(2)$ and $\Psi(3,-1) = A(3)A(2)A(1)A(0)$, while $\Psi(3,3) = I$. Finally, it is noted that $\Psi(n+1,\nu) = A(n+1)\Psi(n,\nu)$.

To obtain a solution to the time varying state space system in (9.1), observe that the state $x(1) = A(0)x_0 + B(0)u(0)$. By recursively solving for the state $x(n)$ in (9.1), we obtain

$$\begin{aligned} x(2) &= A(1)x(1) + B(1)u(1) = A(1)A(0)x_0 + A(1)B(0)u(0) + B(1)u(1) \\ &= \Psi(1,-1)x_0 + \Psi(1,0)B(0)u(0) + \Psi(1,1)B(1)u(1)\,; \end{aligned}$$

$$\begin{aligned} x(3) &= A(2)x(2) + B(2)u(2) \\ &= A(2)A(1)A(0)x_0 + A(2)A(1)B(0)u(0) + A(2)B(1)u(1) + B(2)u(2) \\ &= \Psi(2,-1)x_0 + \Psi(2,0)B(0)u(0) + \Psi(2,1)B(1)u(1) + \Psi(2,2)B(2)u(2)\,. \end{aligned}$$

By recursively solving for the state $x(n)$, it follows that the solution to the time varying state space system in (9.1) is given by

$$\begin{aligned} x(n) &= \Psi(n-1,-1)x_0 + \sum_{j=0}^{n-1} \Psi(n-1,j)B(j)u(j) \\ y(n) &= C(n)\Psi(n-1,-1)x_0 + \sum_{j=0}^{n-1} C(n)\Psi(n-1,j)B(j)u(j) + D(n)v(n)\,. \end{aligned} \tag{9.3}$$

Finally, it is noted that time varying systems play a basic role in linear systems.

Chapter 10

Appendix: A Review of Probability

In this chapter we review some elementary properties concerning probability, expectation and conditional expectation.

10.1 The probability density function

In this section we review and establish some notation concerning random variables. Let $\mathcal{K}$ be the Hilbert space generated by the set of all random variables $\mathbf{z}$ such that $\|\mathbf{z}\|^2 = E|\mathbf{z}|^2$ is finite. Throughout we always assume that all of our random variables are vectors in $\mathcal{K}$. Moreover, in the Appendix and Section 2.2.1 in Chapter 2 we will use a boldface $\mathbf{z}$ to represent the random variable $\mathbf{z}$. This is done mainly to distinguish between the random variable $\mathbf{z}$ and the real number z in the density function $f_{\mathbf{z}}(z)$. This boldface notation for a random variable is not used in the rest of the notes. Let $\mathbf{x}$ be a real random variable and $f_{\mathbf{x}}(x)$ the probability density function associated with $\mathbf{x}$. Recall that $f_{\mathbf{x}}$ is positive and has area one, that is, $f_{\mathbf{x}}(x) \geq 0$ for all x in $\mathbb{R}$ and

$$1 = \int_{-\infty}^{\infty} f_{\mathbf{x}}(x)dx\,.$$

The random variable $\mathbf{x}$ and its probability density function $f_{\mathbf{x}}$ uniquely determine each other. The probability that the random variable $\mathbf{x}$ lives in a certain measurable set Δ is given by

$$\text{Probability}\{\mathbf{x} \in \Delta\} = \int_{\Delta} f_{\mathbf{x}}(x)dx\,.$$

Now let us recall some examples of a random variable. We say that $\mathbf{x}$ is *uniform random variable* over the interval $[a, b]$ if its probability density function is given by

$$\begin{aligned} f_{\mathbf{x}}(x) &= \frac{1}{b-a} \qquad \text{if } a \leq x \leq b \\ &= 0 \qquad\quad \text{otherwise}\,. \end{aligned} \tag{1.1}$$

Obviously, $f_{\mathbf{x}}(x)$ is positive and has area one.

We say that $\mathbf{x}$ is an *exponential random variable* if its probability density function is given by

$$\begin{aligned} f_{\mathbf{x}}(x) &= e^{-x/\mu}/\mu \qquad \text{if } x \geq 0 \\ &= 0 \qquad\qquad\quad \text{if } x < 0\,. \end{aligned} \tag{1.2}$$

Here μ is a strictly positive constant. It is easy to verify that $f_{\mathbf{x}}(x)$ is positive and has area one.

For our final example of a random variable, recall that $\mathbf{x}$ is a *Gaussian random variable* with mean μ and variance $\sigma > 0$ if its its probability density function is given by

$$f_{\mathbf{x}}(x) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-(x-\mu)^2/2\sigma^2} \qquad (-\infty < x < \infty)\,. \tag{1.3}$$

As before, let $\mathbf{x}$ be a random variable. Let h be any measurable function from $\mathbb{R}$ into $\mathbb{C}$, then $h(\mathbf{x})$ is a random variable. Moreover, the expected value of $h(\mathbf{x})$ is given by

$$Eh(\mathbf{x}) = \int_{-\infty}^{\infty} h(x) f_{\mathbf{x}}(x) dx\,.$$

In particular, if γ is any constant, then $E\gamma = \gamma$. To see this simply observe that

$$E\gamma = \int_{-\infty}^{\infty} \gamma f_{\mathbf{x}}(x) dx = \gamma \int_{-\infty}^{\infty} f_{\mathbf{x}}(x) dx = \gamma\,.$$

The *mean* $\mu_{\mathbf{x}}$ and *standard deviation* $\sigma_{\mathbf{x}}$ of $\mathbf{x}$ are defined by

$$\mu_{\mathbf{x}} = E\mathbf{x} = \int_{-\infty}^{\infty} x f_{\mathbf{x}}(x) dx \quad \text{and} \quad \sigma_{\mathbf{x}}^2 = E|\mathbf{x} - \mu_{\mathbf{x}}|^2 = \int_{-\infty}^{\infty} |x - \mu_{\mathbf{x}}|^2 f_{\mathbf{x}}(x) dx\,.$$

The *variance* of $\mathbf{x}$ is $\sigma_{\mathbf{x}}^2$, the square of the standard deviation.

For example, let $\mathbf{x}$ be a uniform random variable over the interval $[a, b]$, then the mean $\mu_{\mathbf{x}}$ and variance $\sigma_{\mathbf{x}}^2$ for $\mathbf{x}$ is given by

$$\mu_{\mathbf{x}} = \frac{a+b}{2} \qquad \text{and} \qquad \sigma_{\mathbf{x}}^2 = \frac{(b-a)^3}{12}\,. \tag{1.4}$$

To see this recall that the probability density function $f_{\mathbf{x}}(x)$ for $\mathbf{x}$ is given by (1.1). Hence

$$\mu_{\mathbf{x}} = E\mathbf{x} = \int_{-\infty}^{\infty} x f_{\mathbf{x}}(x) dx = \frac{1}{b-a}\int_a^b x dx = \frac{x^2}{2(b-a)}\bigg|_a^b = \frac{a+b}{2}\,.$$

This yields the first equation in (1.4). To obtain the variance observe that

$$\begin{aligned} \sigma_{\mathbf{x}}^2 &= E(\mathbf{x} - \mu_{\mathbf{x}})^2 = \int_{-\infty}^{\infty} (x - \mu_{\mathbf{x}})^2 f_{\mathbf{x}}(x) dx = \frac{1}{b-a}\int_a^b (x - \mu_{\mathbf{x}})^2 dx \\ &= \frac{(x-\mu_{\mathbf{x}})^3}{3(b-a)}\bigg|_a^b = \frac{(x-(a+b)/2)^3}{3(b-a)}\bigg|_a^b = \frac{(b-a)^3}{12}\,. \end{aligned}$$

Therefore the variance $\sigma_{\mathbf{x}}^2$ is given by the second equation in (1.4).

Now let $\mathbf{x}$ be the exponential random variable whose probability density function $f_{\mathbf{x}}(x)$ is given by (1.2) where $\mu > 0$. Then μ is the mean and standard deviation for $\mathbf{x}$, that is, $\mu_{\mathbf{x}} = \mu$ and $\sigma_{\mathbf{x}} = \mu$. To obtain the mean simply observe that

$$E\mathbf{x} = \int_{-\infty}^{\infty} x f_{\mathbf{x}}(x)dx = \frac{1}{\mu}\int_0^{\infty} xe^{-x/\mu}dx = -\left[xe^{-x/\mu} + \mu e^{-x/\mu}\right]_0^{\infty} = \mu\,.$$

Using $\mu_{\mathbf{x}} = \mu$, we obtain

$$\begin{aligned}\sigma_{\mathbf{x}}^2 &= E(\mathbf{x}-\mu_{\mathbf{x}})^2 = \int_{-\infty}^{\infty}(x-\mu_{\mathbf{x}})^2 f_{\mathbf{x}}(x)dx = \frac{1}{\mu}\int_0^{\infty}(x-\mu)^2 e^{-x/\mu}dx\\ &= -\left[(x-\mu)^2 e^{-x/\mu} + 2\mu(x-\mu)e^{-x/\mu} + 2\mu^2 e^{-x/\mu}\right]_0^{\infty}\\ &= \mu^2 - 2\mu^2 + 2\mu^2 = \mu^2\,.\end{aligned}$$

Therefore the standard deviation $\sigma_{\mathbf{x}} = \mu_{\mathbf{x}} = \mu$, the mean for an exponential random variable.

Now let $\mathbf{x}$ be the Gaussian random variable whose probability density function $f_{\mathbf{x}}(x)$ is given by (1.3) where $-\infty < \mu < \infty$ and $\sigma > 0$. As expected, the mean $\mu_{\mathbf{x}} = \mu$ and the standard deviation $\sigma_{\mathbf{x}} = \sigma$. In other words,

$$\begin{aligned}\mu_{\mathbf{x}} &= E\mathbf{x} = \int_{-\infty}^{\infty} x f_{\mathbf{x}}(x)dx = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} xe^{-(x-\mu)^2/2\sigma^2}dx = \mu\\ \sigma_{\mathbf{x}} &= E(\mathbf{x}-\mu_{\mathbf{x}})^2 = \int_{-\infty}^{\infty}(x-\mu_{\mathbf{x}})^2 f_{\mathbf{x}}(x)dx = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}(x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2}dx = \sigma^2\,.\end{aligned}$$

The details are left to the reader as an exercise.

To provide an interpretation for the mean and standard deviation, consider the optimization problem

$$\sigma^2 = \inf\{E|\mathbf{x}-\mu|^2 : \mu \in \mathbb{C}\} = \inf\{\|\mathbf{x}-\mu\|_{\mathcal{K}}^2 : \mu \in \mathbb{C}\}\,. \tag{1.5}$$

We claim that then mean $\widehat{\mathbf{x}} = \mu_{\mathbf{x}}$ is the unique solution to this optimization problem, and the error $\sigma = \sigma_{\mathbf{x}}$ the standard deviation. In other words, the mean $\mu_{\mathbf{x}}$ is the constant which comes closest to the random variable $\mathbf{x}$ in the norm $\|\mathbf{z}\|_{\mathcal{K}}^2 = E|\mathbf{z}|^2$. The standard deviation is the distance from $\mathbf{x}$ to the mean $\mu_{\mathbf{x}}$. Of course, one can solve the optimization problem in (1.5) by elementary calculus. However, let us apply the projection theorem to solve this problem. To this end, let $\mathcal{H}$ be the one dimensional space spanned by the constant random variables, that is, $\mathcal{H} = \mathbb{C}$. By the projection theorem there exists a unique solution to (1.5). Furthermore, the optimal solution $\widehat{\mathbf{x}} = P_{\mathcal{H}}\mathbf{x}$ where $P_{\mathcal{H}}$ is the orthogonal projection onto $\mathcal{H}$. In particular, $\widehat{\mathbf{x}}$ is a constant, and thus, $E\widehat{\mathbf{x}} = \widehat{\mathbf{x}}$. Moreover, $\mathbf{x} - \widehat{\mathbf{x}}$ is orthogonal to $\mathcal{H}$, or equivalently, $\mathbf{x} - \widehat{\mathbf{x}}$ is orthogonal to 1. This readily implies that

$$0 = E(\mathbf{x}-\widehat{\mathbf{x}})1^* = E\mathbf{x} - E\widehat{\mathbf{x}} = E\mathbf{x} - \widehat{\mathbf{x}}\,.$$

Hence $\widehat{\mathbf{x}} = E\mathbf{x} = \mu_{\mathbf{x}}$. Therefore the error $\sigma^2 = E|\mathbf{x}-\widehat{\mathbf{x}}|^2 = \sigma_{\mathbf{x}}^2$.

Let $\mathbf{y}$ be a random vector with values in $\mathbb{R}^\nu$, that is, $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_\nu]^{tr}$ where $\mathbf{y}_k$ is a real valued random variable for $k = 1, 2, \cdots, \nu$. Let $f_{\mathbf{y}}(y)$ be the probability density function associated with $\mathbf{y}$. Recall that $f_{\mathbf{y}}(y)$ is a positive function mapping $\mathbb{R}^\nu$ into $\mathbb{R}$ of the form

$$f_{\mathbf{y}}(y) = f_{\mathbf{y}}(y_1, y_2, \cdots, y_\nu)\,.$$

Moreover, $f_{\mathbf{y}}(y)$ is positive and has area one, that is, $f_{\mathbf{y}}(y) \geq 0$ for all y in $\mathbb{R}^\nu$ and

$$1 = \int_{\mathbb{R}^\nu} f_{\mathbf{y}}(y)dy\,.$$

Here $dy = dy_1 dy_2, \cdots, dy_\nu$. The random variable $\mathbf{y}$ and its probability density function $f_{\mathbf{y}}$ uniquely determine each other. The probability that the random variable $\mathbf{y}$ lives in a certain measurable set Δ in $\mathbb{R}^\nu$ is given by

$$\text{Probability}\{\mathbf{y} \in \Delta\} = \int_{\Delta} f_{\mathbf{y}}(y)dy\,.$$

If $\mathbf{y}_k$ is the random variable in the k-component of $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_\nu]^{tr}$, then the probability density function $f_{\mathbf{y}_k}(y_k)$ is given by

$$f_{\mathbf{y}_k}(y_k) = \int_{\mathbb{R}^{\nu-1}} f_{\mathbf{y}}(y_1, y_2, \cdots, y_\nu)dy_1 dy_2 \cdots dy_{k-1}\; dy_{k+1} dy_{k+2} \cdots dy_\nu\,. \tag{1.6}$$

In this equation all the dy_j terms are present except when $j = k$.

Let g be any measurable function from $\mathbb{R}^\nu$ into $\mathbb{C}$, then $g(\mathbf{y})$ is a random variable. Moreover, the expected value of $g(\mathbf{y})$ is given by

$$Eg(\mathbf{y}) = \int_{\mathbb{R}^\nu} g(y) f_{\mathbf{y}}(y)dy\,. \tag{1.7}$$

In particular, the expectation E is a linear operator, that is, if $\{c_j\}_1^\nu$ are scalars, then

$$E\left(\sum_{k=1}^{\nu} c_k \mathbf{y}_k\right) = \sum_{k=1}^{\nu} c_k E\mathbf{y}_k\,. \tag{1.8}$$

To see this, let g be the function from $\mathbb{R}^\nu$ into $\mathbb{C}$ defined by $g(y) = \sum_{k=1}^{\nu} c_k y_k$ where $y = [y_1, y_2, \cdots, y_\nu]^{tr}$. Then (1.6) and (1.7) give

$$\begin{aligned} E\left(\sum_{k=1}^{\nu} c_k \mathbf{y}_k\right) &= Eg(\mathbf{y}) = \int_{\mathbb{R}^\nu} g(y) f_{\mathbf{y}}(y)dy = \sum_{k=1}^{\nu} c_k \int_{\mathbb{R}^\nu} y_k f_{\mathbf{y}}(y)dy \\ &= \sum_{k=1}^{\nu} c_k \int_{\mathbb{R}} y_k dy_k \int_{\mathbb{R}^{\nu-1}} f_{\mathbf{y}}(y)dy_1 \cdots dy_{k-1}\; dy_{k+1} \cdots dy_\nu \\ &= \sum_{k=1}^{\nu} c_k \int_{\mathbb{R}} y_k f_{\mathbf{y}_k}(y)dy_k = \sum_{k=1}^{\nu} c_k E\mathbf{y}_k\,. \end{aligned}$$

Therefore (1.8) holds.

To develop the conditional expectation we will need the joint density function between a random variable $\mathbf{x}$ and a random vector $\mathbf{y}$. To this end, let $\mathbf{x}$ be a real valued random variable and $\mathbf{y}$ a random vector with values in $\mathbb{R}^\nu$. The space $\mathbb{R}^\nu$ could simply be $\mathbb{R}^1$, and then $\mathbf{y}$ is a random variable. Throughout $f_{\mathbf{x},\mathbf{y}}$ denotes the joint probability density function

for $\mathbf{x}$ and $\mathbf{y}$. In other words, $f_{\mathbf{x},\mathbf{y}}$ is simply the probability density function for the random variable $[\mathbf{x},\mathbf{y}]^{tr}$ with values in $\mathbb{R}^{\nu+1}$. In particular, $f_{\mathbf{x},\mathbf{y}}$ is a function mapping $\mathbb{R}^{\nu+1}$ into $\mathbb{R}$ of the form

$$f_{\mathbf{x},\mathbf{y}}(x,y) = f_{\mathbf{x},\mathbf{y}}(x,y_1,y_2,\cdots,y_\nu) \tag{1.9}$$

where x is in $\mathbb{R}$ and $y = [y_1,y_2,\cdots,y_\nu]^{tr}$ is a vector in $\mathbb{R}^\nu$. The joint probability density function is uniquely determined by the random variable $\mathbf{x}$ and random vector $\mathbf{y}$. Recall that $f_{\mathbf{x},\mathbf{y}}$ is positive with area one, that is,

$$\begin{aligned} 0 &\leq f_{\mathbf{x},\mathbf{y}}(x,y) \qquad \text{(for all } x\in\mathbb{R} \text{ and } y\in\mathbb{R}^\nu) \\ 1 &= \int_{\mathbb{R}}\int_{\mathbb{R}^\nu} f_{\mathbf{x},\mathbf{y}}(x,y)dydx\,. \end{aligned} \tag{1.10}$$

Here $dy = dy_1dy_2\cdots dy_\nu$. Moreover, the joint density function has the following property

$$f_{\mathbf{y}}(y) = \int_{\mathbb{R}} f_{\mathbf{x},\mathbf{y}}(x,y)dx \qquad \text{and} \qquad f_{\mathbf{x}}(x) = \int_{\mathbb{R}^\nu} f_{\mathbf{x},\mathbf{y}}(x,y)dy\,. \tag{1.11}$$

If $c(x,y)$ is a measurable function from $\mathbb{R}\oplus\mathbb{R}^\nu$ into $\mathbb{C}$, then $c(\mathbf{x},\mathbf{y})$ is a random variable. Moreover, the mean of $c(\mathbf{x},\mathbf{y})$ is given by

$$Ec(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}}\int_{\mathbb{R}^\nu} c(x,y)f_{\mathbf{x},\mathbf{y}}(x,y)dydx\,. \tag{1.12}$$

Recall that $\mathbf{x}$ and $\mathbf{y}$ are independent random variables if and only if $f_{\mathbf{x},\mathbf{y}}(x,y) = f_{\mathbf{x}}(x)f_{\mathbf{y}}(y)$ for all x in $\mathbb{R}$ and y in $\mathbb{R}^\nu$. Now assume that $\mathbf{x}$ and $\mathbf{y}$ are independent random variables. Let $h(x)$ be a measurable function from $\mathbb{R}$ into $\mathbb{C}$ and $g(y)$ a measurable function from $\mathbb{R}^\nu$ into $\mathbb{C}$. Then $h(\mathbf{x})$ and $g(\mathbf{y})$ are independent random variables. Moreover,

$$Eh(\mathbf{x})g(\mathbf{y}) = Eh(\mathbf{x})Eg(\mathbf{y}) \qquad \text{(if } \mathbf{x} \text{ and } \mathbf{y} \text{ are independent)}\,. \tag{1.13}$$

In particular, $E\mathbf{x}\mathbf{y}^* = E\mathbf{x}E\mathbf{y}^*$. To verify that (1.13) holds simply observe that

$$\begin{aligned} Eh(\mathbf{x})g(\mathbf{y}) &= \int_{\mathbb{R}}\int_{\mathbb{R}^\nu} h(x)g(y)f_{\mathbf{x},\mathbf{y}}(x,y)dydx = \int_{\mathbb{R}}\int_{\mathbb{R}^\nu} h(x)g(y)f_{\mathbf{x}}(x)f_{\mathbf{y}}(y)dydx \\ &= \int_{\mathbb{R}} h(x)f_{\mathbf{x}}(x)dx\int_{\mathbb{R}^\nu} g(y)f_{\mathbf{y}}(y)dy = Eh(\mathbf{x})Eg(\mathbf{y})\,. \end{aligned}$$

Therefore (1.13) holds.

10.2 Conditional expectation

Let $\mathcal{K}$ be the Hilbert space generated by the set of all random variables $\mathbf{z}$ such that $E|\mathbf{z}|^2$ is finite. Throughout we always assume that all of our random variables are vectors in $\mathcal{K}$.

As before, let $f_{\mathbf{x},\mathbf{y}}(x,y)$ denote the joint density function for a real valued random variable $\mathbf{x}$ and a random vector $\mathbf{y}$ with values in $\mathbb{R}^\nu$. Notice that if $f_{\mathbf{y}}(y) = 0$ for some y in $\mathbb{R}^\nu$, then

$f_{\mathbf{x},\mathbf{y}}(x,y) = 0$ for all x in $\mathbb{R}$. To see this observe that the first equality in (1.11) shows that the area under $f_{\mathbf{x},\mathbf{y}}(x,y)$ with respect to x is zero. Because $f_{\mathbf{x},\mathbf{y}}$ is positive, this implies that $f_{\mathbf{x},\mathbf{y}}(x,y) = 0$ for all x. The *conditional density* $f_{\mathbf{x}|\mathbf{y}}$ is defined by

$$f_{\mathbf{x}|\mathbf{y}}(x|y) = \frac{f_{\mathbf{x},\mathbf{y}}(x,y)}{f_{\mathbf{y}}(y)}\,. \tag{2.1}$$

The conditional density $f_{\mathbf{x}|\mathbf{y}}$ is only defined when $f_{\mathbf{y}}(y)$ is nonzero. If $f_{\mathbf{y}}(y)$ is zero, then we set $f_{\mathbf{x}|\mathbf{y}}(x|y)$ equal to zero. Obviously, $f_{\mathbf{x},\mathbf{y}}(x,y) = f_{\mathbf{x}|\mathbf{y}}(x|y)f_{\mathbf{y}}(y)$. It is important to emphasize that the conditional density $f_{\mathbf{x}|\mathbf{y}}$ is a density function when y is fixed. In other words, $f_{\mathbf{x}|\mathbf{y}}(x|y)$ is positive and has area one. To be precise,

$$0 \le f_{\mathbf{x}|\mathbf{y}}(x|y) \qquad \text{and} \qquad 1 = \int_{\mathbb{R}} f_{\mathbf{x}|\mathbf{y}}(x|y)dx\,.$$

Clearly, $f_{\mathbf{x}|\mathbf{y}} = f_{\mathbf{x},\mathbf{y}}/f_{\mathbf{y}}$ is positive. The last equality follows from the definition of the conditional density and equation (1.11), that is,

$$\int_{\mathbb{R}} f_{\mathbf{x}|\mathbf{y}}(x|y)dx = \int_{\mathbb{R}} \frac{f_{\mathbf{x},\mathbf{y}}(x,y)}{f_{\mathbf{y}}(y)}\,dx = \frac{f_{\mathbf{y}}(y)}{f_{\mathbf{y}}(y)} = 1\,.$$

Finally, if $\mathbf{x}$ and $\mathbf{y}$ are independent, then $f_{\mathbf{x}|\mathbf{y}}(x|y) = f_{\mathbf{x}}(x)$. This follows from the fact that $f_{\mathbf{x},\mathbf{y}}(x,y) = f_{\mathbf{x}}(x)f_{\mathbf{y}}(y)$ when $\mathbf{x}$ and $\mathbf{y}$ are independent.

The *condition expectation* $E(\mathbf{x}|\mathbf{y})$ is the real valued random variable defined by

$$E(\mathbf{x}|\mathbf{y} = y) = \int_{\mathbb{R}} x f_{\mathbf{x}|\mathbf{y}}(x|y)\,dx\,. \tag{2.2}$$

In this notation for the condition expectation $E(\mathbf{x}|\mathbf{y})$ is defined by computing the expectation when the random variable $\mathbf{y} = y$. For another interpretation of the conditional expectation, let $\widehat{g}$ be the function mapping $\mathbb{R}^\nu$ into $\mathbb{C}$ defined by

$$\widehat{g}(y) = \int_{-\infty}^{\infty} x f_{\mathbf{x}|\mathbf{y}}(x|y)\,dx\,. \tag{2.3}$$

Then the conditional expectation is simply the random variable defined by

$$E(\mathbf{x}|\mathbf{y}) = \widehat{g}(\mathbf{y}) = \widehat{g}(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_\nu) \qquad (\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_\nu]^{tr})\,. \tag{2.4}$$

Since $\widehat{g}(\mathbf{y})$ is a function of $\mathbf{y}$ it is clear that the conditional expectation $E(\mathbf{x}|\mathbf{y}) = \widehat{g}(\mathbf{y})$ is a random variable. Finally, it is noted that the expected value of the random variable $E(\mathbf{x}|\mathbf{y})$ is simply $E\mathbf{x}$, that is,

$$EE(\mathbf{x}|\mathbf{y}) = E\widehat{g}(\mathbf{y}) = E\mathbf{x}\,. \tag{2.5}$$

To verify this simply observe that

$$\begin{aligned} EE(\mathbf{x}|\mathbf{y}) &= E\widehat{g}(\mathbf{y}) = \int_{\mathbb{R}^\nu} \widehat{g}(y) f_{\mathbf{y}}(y)dy = \int_{\mathbb{R}^\nu}\int_{\mathbb{R}} x f_{\mathbf{x}|\mathbf{y}}(x|y) f_{\mathbf{y}}(y)dxdy \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}^\nu} x f_{\mathbf{x},\mathbf{y}}(x,y)dydx = \int_{\mathbb{R}} x f_{\mathbf{x}}(x)dx = E\mathbf{x}\,. \end{aligned}$$

The fifth equality follows from the second equation in (1.11).

Consider the optimization problem

$$\gamma = \inf\{E|\mathbf{x} - g(\mathbf{y})|^2 : g \text{ is a measurable function from } \mathbb{R}^\nu \text{ into } \mathbb{C}\}\,. \tag{2.6}$$

Let us use the projection theorem to solve this problem. To this end, let $\mathcal{G}$ be the subspace of $\mathcal{K}$ defined by

$$\mathcal{G} = \{g(\mathbf{y}) \in \mathcal{K} : g \text{ is a measurable function from } \mathbb{R}^\nu \text{ into } \mathbb{C}\}\,.$$

Clearly, $\mathcal{G}$ is a subspace of $\mathcal{K}$. According to the projection theorem, there is a unique solution $\widehat{\mathbf{x}}$ to the optimization problem (2.6). Moreover, this solution is given by $\widehat{\mathbf{x}} = P_{\mathcal{G}}\mathbf{x}$ where $P_{\mathcal{G}}$ is the orthogonal projection onto $\mathcal{G}$.

We claim that the optimal solution is given by the conditional expectation, that is, $P_{\mathcal{G}}\mathbf{x} = \widehat{g}(\mathbf{y}) = E(\mathbf{x}|\mathbf{y})$ where $\widehat{g}(y)$ is the function defined in (2.3). According to the projection theorem there is a unique solution $\widehat{g}(\mathbf{y})$ to the optimization problem in (2.6). Moreover, $\mathbf{x} - \widehat{g}(\mathbf{y})$ is orthogonal to $h(\mathbf{y})$ where $h(y)$ is any measurable function from $\mathbb{R}^\nu$ into $\mathbb{C}$ such that $h(\mathbf{y})$ is in $\mathcal{K}$. Hence

$$\begin{aligned}
0 &= E(\mathbf{x} - \widehat{g}(\mathbf{y}))\overline{h}(\mathbf{y}) = \int_{\mathbb{R}^\nu}\int_{\mathbb{R}} (x - \widehat{g}(y))\overline{h}(y) f_{\mathbf{x},\mathbf{y}}(x,y)\,dxdy \\
&= \int_{\mathbb{R}^\nu}\left[\int_{\mathbb{R}} (x - \widehat{g}(y)) f_{\mathbf{x},\mathbf{y}}(x,y)\,dx\right]\overline{h}(y)dy \\
&= \int_{\mathbb{R}^\nu}\left[\int_{\mathbb{R}} x f_{\mathbf{x},\mathbf{y}}(x,y)\,dx - \widehat{g}(y)\int_{\mathbb{R}} f_{\mathbf{x},\mathbf{y}}(x,y)\,dx\right]\overline{h}(y)dy \\
&= \int_{\mathbb{R}^\nu}\left[\int_{\mathbb{R}} x f_{\mathbf{x},\mathbf{y}}(x,y)\,dx - \widehat{g}(y) f_{\mathbf{y}}(y)\right]\overline{h}(y)dy\,.
\end{aligned}$$

Because this integral is zero for all functions $h(y)$, the term in the brackets $[\cdots]$ must also be zero. Hence $\widehat{g}(y)f_{\mathbf{y}}(y) = \int_{\mathbb{R}} x f_{\mathbf{x},\mathbf{y}}(x,y)\,dx$, or equivalently,

$$\widehat{g}(y) = \frac{1}{f_{\mathbf{y}}(y)}\int_{\mathbb{R}} x f_{\mathbf{x},\mathbf{y}}(x,y)dx = \int_{\mathbb{R}} x f_{\mathbf{x}|\mathbf{y}}(x|y)dx\,. \tag{2.7}$$

Therefore $P_{\mathcal{G}}\mathbf{x} = \widehat{g}(\mathbf{y}) = E(\mathbf{x}|\mathbf{y})$.

Since the constant random variable 1 is in $\mathcal{G}$, the vector $\mathbf{x} - \widehat{g}(\mathbf{y})$ is orthogonal to 1. In other words, $E(\mathbf{x} - \widehat{g}(\mathbf{y}))1 = 0$. Hence $E\mathbf{x} = E\widehat{g}(\mathbf{y}) = EE(\mathbf{x}|\mathbf{y})$. This yields another proof of (2.5). Because the orthogonal projection is a linear operator and $P_{\mathcal{G}}\mathbf{x} = E(\mathbf{x}|\mathbf{y})$, the conditional expectation $E(\mathbf{x}|\mathbf{y})$ is also linear in $\mathbf{x}$, that is, if $\mathbf{x}$ and $\mathbf{z}$ are scalar valued random variables and α and β are scalars, then

$$E(\alpha\mathbf{x} + \beta\mathbf{z}|\mathbf{y}) = \alpha E(\mathbf{x}|\mathbf{y}) + \beta E(\mathbf{z}|\mathbf{y})\,. \tag{2.8}$$

Summing up the previous analysis yields the following result.

THEOREM 10.2.1 *Let* $\mathbf{x}$ *be a real valued random variable and* $\mathbf{y}$ *a random vector with values in* $\mathbb{R}^\nu$*. Then the conditional expectation* $\widehat{g}(\mathbf{y}) = E(\mathbf{x}|\mathbf{y})$ *is the unique random variable solving the optimization problem*

$$E|\mathbf{x} - \widehat{g}(\mathbf{y})|^2 = \inf\{E|\mathbf{x} - g(\mathbf{y})|^2 : g \text{ is a measurable function from } \mathbb{R}^\nu \text{ into } \mathbb{C}\}\,. \qquad (2.9)$$

Finally, the conditional expectation $E(\mathbf{x}|\mathbf{y})$ *is linear in* $\mathbf{x}$ *and* $E\mathbf{x} = EE(\mathbf{x}|\mathbf{y})$*.*

Of course, one can also directly verify that $P_{\mathcal{G}}\mathbf{x} = E(\mathbf{x}|\mathbf{y})$. In other words, once we have the formula for the conditional expectation it is easy to verify that $P_{\mathcal{G}}\mathbf{x} = E(\mathbf{x}|\mathbf{y})$. To prove this simply notice that $\widehat{g}(\mathbf{y})$ is in $\mathcal{G}$. Recall that $P_{\mathcal{G}}\mathbf{x} = \widehat{g}(\mathbf{y})$ if and only if $\mathbf{x} - \widehat{g}(\mathbf{y})$ is orthogonal to $\mathcal{G}$. Now let $h(\mathbf{y})$ be any vector in $\mathcal{G}$, that is, assume that $h(y)$ is a measurable function from $\mathbb{R}^\nu$ into $\mathbb{C}$ such that $h(\mathbf{y})$ is in $\mathcal{K}$. Then

$$\begin{aligned}
E(\mathbf{x} - \widehat{g}(\mathbf{y}))\overline{h}(\mathbf{y}) &= E\mathbf{x}\overline{h}(\mathbf{y}) - \int_{\mathbb{R}}\int_{\mathbb{R}^\nu} \widehat{g}(y)\overline{h}(y) f_{\mathbf{x},\mathbf{y}}(x,y)dydx \\
&= E\mathbf{x}\overline{h}(\mathbf{y}) - \int_{\mathbb{R}^\nu} \widehat{g}(y)\overline{h}(y)\int_{\mathbb{R}} f_{\mathbf{x},\mathbf{y}}(x,y)dxdy \\
&= E\mathbf{x}\overline{h}(\mathbf{y}) - \int_{\mathbb{R}^\nu} \widehat{g}(y)\overline{h}(y) f_{\mathbf{y}}(y)dy \\
&= E\mathbf{x}\overline{h}(\mathbf{y}) - \int_{\mathbb{R}^\nu}\int_{\mathbb{R}} x f_{\mathbf{x}|\mathbf{y}}(x|y)\overline{h}(y) f_{\mathbf{y}}(y)dxdy \\
&= E\mathbf{x}\overline{h}(\mathbf{y}) - \int_{\mathbb{R}^\nu}\int_{\mathbb{R}} x\overline{h}(y) f_{\mathbf{x},\mathbf{y}}(x,y)dxdy \\
&= E\mathbf{x}\overline{h}(\mathbf{y}) - E\mathbf{x}\overline{h}(\mathbf{y}) = 0\,.
\end{aligned}$$

Hence $\mathbf{x} - \widehat{g}(\mathbf{y})$ is orthogonal to $\mathcal{G}$. Therefore $P_{\mathcal{G}}\mathbf{x} = \widehat{g}(\mathbf{y}) = E(\mathbf{x}|\mathbf{y})$.

In many application the conditional expectation is hard to compute. To approximate the conditional expectation consider the space of polynomials given by

$$\mathcal{G}_k = \{p(\mathbf{y}) \in \mathcal{K} : p \text{ is a polynomial from of degree at most } k \text{ from } \mathbb{R}^\nu \text{ into } \mathbb{C}\}\,.$$

If $\mathbf{x}$ and $\mathbf{y}$ have a joint density, then the closed linear span of $\{\mathcal{G}_k\}_0^\infty$ equals $\mathcal{G}$. Thus $P_{\mathcal{G}_k}\mathbf{x}$ converges to $P_{\mathcal{G}}\mathbf{x} = E(\mathbf{x}|\mathbf{y})$ as k tends to infinity. Here $P_{\mathcal{G}_k}$ is the orthogonal projection onto $\mathcal{G}_k$. In many instances it is easier to compute $P_{\mathcal{G}_k}\mathbf{x}$ than the conditional expectation. So in applications one can compute $P_{\mathcal{G}_k}\mathbf{x}$ to approximate the conditional expectation.

10.2.1 Gaussian random vectors and estimation

We say that $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_n \end{bmatrix}^{tr}$ is a *Gaussian random vector* if every component $\mathbf{y}_k$ of $\mathbf{y}$ is a linear combination of independent mean zero Gaussian random variables plus a constant. To be precise, we say that $\mathbf{y}$ is a Gaussian random vector if

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_m \end{bmatrix} + \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_n \end{bmatrix}$$

where $\{u_k\}_1^m$ is a set of linearly independent mean zero Gaussian random variables while $\{\gamma_k\}_1^n$ and $\{a_{jk}\}_{11}^{nm}$ are constants. Finally, we say that $\mathbf{x}$ and $\mathbf{y}$ are *jointly Gaussian random* vectors if

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$$

is a Gaussian random vector.

The following result shows that for jointly Gaussian random vectors the linear estimate equals the conditional expectation.

THEOREM 10.2.2 *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be jointly Gaussian random vectors and let* $\mathcal{M}$ *be the linear space spanned by the components of* $\begin{bmatrix} 1 & \mathbf{y} \end{bmatrix}^{tr}$. *Let* $\widehat{\mathbf{x}} = P_{\mathcal{M}}\mathbf{x}$ *be the orthogonal projection of* $\mathbf{x}$ *onto the subspace* $\mathcal{M}$. *Then*

$$\widehat{\mathbf{x}} = P_{\mathcal{M}}\mathbf{x} = E(\mathbf{x}|\mathbf{y})\,. \tag{2.10}$$

In other words, for jointly Gaussian random vectors the linear estimate of $\mathbf{x}$ *given* $\begin{bmatrix} 1 & \mathbf{y} \end{bmatrix}^{tr}$ *equals the nonlinear estimate* $E(\mathbf{x}|\mathbf{y})$.

A proof of Theorem 10.2.2 is given in Doob [10].

10.3 The sum of two exponential random variables.

In this section we will compute the conditional expectation $E(\mathbf{x}|\mathbf{y})$ where $\mathbf{y} = \mathbf{x} + \mathbf{v}$ is the sum of two independent exponential random variables $\mathbf{x}$ and $\mathbf{v}$. Recall that $\mathbf{z}$ is an exponential random variable if its probability density function is given by

$$\begin{aligned} f_{\mathbf{z}}(z) &= e^{-z/\mu}/\mu \qquad \text{if } z \geq 0 \\ &= 0 \qquad\qquad\ \ \text{if } z < 0\,. \end{aligned} \tag{3.1}$$

The parameter μ is strictly positive and uniquely determines the density $f_{\mathbf{z}}$. Moreover, μ is the mean of the exponential random variable $\mathbf{z}$, that is,

$$E\mathbf{z} = \int_{-\infty}^{\infty} z f_{\mathbf{z}}(z)dz = \frac{1}{\mu}\int_0^{\infty} z e^{-z/\mu}dz = \mu\,.$$

Hence $\mu = \mu_{\mathbf{z}}$. So we say that $\mathbf{z}$ is an exponential random variable with mean μ if the probability density function for $\mathbf{z}$ is given by (3.1). For a physical interpretation of the exponential random variable, recall that the exponential density can be used to compute the failure time of a light bulb or certain electrical equipment. So the probability that a certain component will fail in a time interval $[z_1, z_2]$ is given by

$$\text{Probability}\{z_1 \leq \mathbf{z} \leq z_2\} = \frac{1}{\mu}\int_{z_1}^{z_2} e^{-z/\mu}dz = e^{-z_1/\mu} - e^{-z_2/\mu}\,.$$

Let $\mathbf{x}$ and $\mathbf{v}$ be two independent exponential random variables. Let $\mathbf{y} = \mathbf{x} + \mathbf{v}$. Then consider the problem of finding the best estimate of $\mathbf{x}$ given $\mathbf{y}$, that is, compute $E(\mathbf{x}|\mathbf{y})$. For

a physical interpretation of this problem recall that the exponential density can be used to compute the failure time of a light bulb or certain electrical equipment. In this case, $\mathbf{y}$ is the failure time of the sum of two components. For example, suppose that component $\mathbf{x}$ is turned on and when $\mathbf{x}$ fails then component $\mathbf{v}$ is turned on, then $\mathbf{y}$ is the time recorded when component $\mathbf{v}$ fails, that is, $\mathbf{y} = \mathbf{x} + \mathbf{v}$. So our problem is to find the best estimate of the failure time of $\mathbf{x}$ given $\mathbf{y}$. Finally, without loss of generality we can normalize the problem so that the mean of $\mathbf{x}$ is one.

10.3.1 The case when $E\mathbf{v} = 1$

For the moment assume that $\mathbf{x}$ and $\mathbf{v}$ are independent exponential random variables with mean one. Notice that $f_{\mathbf{x}}(x) = 0$ when $x < 0$ and $f_{\mathbf{v}}(y - x) = 0$ when $x > y$; see (3.1). So $f_{\mathbf{x}}(x) f_{\mathbf{v}}(y - x)$ is zero outside the interval $[0, y]$. Recall that $\mathbf{y} = \mathbf{x} + \mathbf{v}$. By employing Lemma 10.3.1 below the density function for $f_{\mathbf{y}}(y)$ when $y \geq 0$ is given by the convolution formula

$$f_{\mathbf{y}}(y) = \int_{-\infty}^{\infty} f_{\mathbf{x}}(x) f_{\mathbf{v}}(y - x) dx = \int_0^y e^{-x} e^{-(y-x)} dx = e^{-y} \int_0^y dx = y e^{-y} .$$

Furthermore, if $y < 0$, then $f_{\mathbf{y}}(y) = 0$. So the probability density function for $\mathbf{y}$ is given by

$$\begin{aligned} f_{\mathbf{y}}(y) &= y e^{-y} \qquad \text{if } y \geq 0 \\ &= 0 \qquad\quad \text{if } y < 0 . \end{aligned} \tag{3.2}$$

Since $f_{\mathbf{x}}(x) f_{\mathbf{v}}(y - x) = e^{-y}$ when $0 \leq x \leq y$ and zero otherwise, Lemma 10.3.1 also shows that the conditional expectation $f_{\mathbf{x}|\mathbf{y}}$ is given by

$$\begin{aligned} f_{\mathbf{x}|\mathbf{y}}(x|y) &= 1/y \qquad \text{if } 0 \leq x \leq y \\ &= 0 \qquad\quad \text{otherwise} . \end{aligned} \tag{3.3}$$

The optimal function $\widehat{g}(y) = \int_0^y x f_{\mathbf{x}|\mathbf{y}}(x|y) dx$ is given by

$$\widehat{g}(y) = y/2 \qquad (\text{if } E\mathbf{v} = 1) . \tag{3.4}$$

Therefore the conditional expectation $E(\mathbf{x}|\mathbf{y}) = \widehat{g}(\mathbf{y}) = \mathbf{y}/2$. In other words, if $\mathbf{x}$ and $\mathbf{v}$ are independent exponential random variables with mean one, then the best estimate of the failure time of $\mathbf{x}$ given $\mathbf{y}$ is simply $\mathbf{y}/2$. Finally, it is noted that in this case the conditional expectation is a $E(\mathbf{x}|\mathbf{y})$ linear function of $\mathbf{y}$.

LEMMA 10.3.1 *Let $\mathbf{y} = \mathbf{x} + \mathbf{v}$ where $\mathbf{x}$ and $\mathbf{v}$ are two independent random variables. Then the density functions $f_{\mathbf{x},\mathbf{y}}$ and $f_{\mathbf{y}}$ are given by*

$$f_{\mathbf{x},\mathbf{y}}(x, y) = f_{\mathbf{x}}(x) f_{\mathbf{v}}(y - x) \qquad \textit{and} \qquad f_{\mathbf{y}}(y) = \int_{-\infty}^{\infty} f_{\mathbf{x}}(x) f_{\mathbf{v}}(y - x) dx . \tag{3.5}$$

In particular, the conditional density

$$f_{\mathbf{x}|\mathbf{y}}(x|y) = \frac{f_{\mathbf{x}}(x) f_{\mathbf{v}}(y - x)}{f_{\mathbf{y}}(y)} . \tag{3.6}$$

PROOF. To prove this result we need that following classical result from probability theory. Let $g(x,y)$ and $h(x,y)$ be two differentiable functions from $\mathbb{R}^2$ into $\mathbb{R}$. Moreover, assume that $[g(x,y),h(x,y)]^{tr}$ is an invertible mapping from $\mathbb{R}^2$ onto $\mathbb{R}^2$. Furthermore, let ξ and η be two random variables. Let $\mathbf{u}$ and $\mathbf{w}$ the random variables defined by $\mathbf{u} = g(\xi,\eta)$ and $\mathbf{w} = h(\xi,\eta)$. Then the density function

$$f_{\xi,\eta}(x,y) = f_{\mathbf{u},\mathbf{w}}(g(x,y),h(x,y))|\det J| \tag{3.7}$$

where J is the Jacobian, that is,

$$J = \begin{bmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \\ \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} \end{bmatrix} \tag{3.8}$$

Now assume that $\mathbf{y} = \mathbf{x}+\mathbf{v}$ where $\mathbf{x}$ and $\mathbf{v}$ are independent random variables. Obviously, $\mathbf{x} = \mathbf{x}$ and $\mathbf{v} = \mathbf{y}-\mathbf{x}$. Let $g(x,y) = x$ and $h(x,y) = y-x$. Clearly, $\mathbf{x} = g(\mathbf{x},\mathbf{y})$ and $\mathbf{v} = h(\mathbf{x},\mathbf{y})$. Furthermore,

$$\begin{bmatrix} g(x,y) \\ h(x,y) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

is an invertible map on $\mathbb{R}^2$. In this case, the Jacobian is given by

$$J = \begin{bmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \\ \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}.$$

Obviously, the determinant of J is one. Because $\mathbf{x}$ and $\mathbf{v}$ are independent random variables, $f_{\mathbf{x},\mathbf{v}}(x,v) = f_{\mathbf{x}}(x)f_{\mathbf{v}}(v)$. According to (3.7) with $\mathbf{x} = \xi = \mathbf{u}$, the random variable $\mathbf{y} = \eta$ and $\mathbf{v} = \mathbf{w}$, the joint density function $f_{\mathbf{x},\mathbf{y}}$ is given by

$$f_{\mathbf{x},\mathbf{y}}(x,y) = f_{\mathbf{x},\mathbf{v}}(g(x,y),h(x,y))|\det J| = f_{\mathbf{x}}(g(x,y))f_{\mathbf{v}}(h(x,y)) = f_{\mathbf{x}}(x)f_{\mathbf{v}}(y-x).$$

Hence $f_{\mathbf{x},\mathbf{y}}(x,y) = f_{\mathbf{x}}(x)f_{\mathbf{v}}(y-x)$ and the first equality in (3.5) holds. The second equality follows from

$$f_{\mathbf{y}}(y) = \int_{-\infty}^{\infty} f_{\mathbf{x},\mathbf{y}}(x,y)dx = \int_{-\infty}^{\infty} f_{\mathbf{x}}(x)f_{\mathbf{v}}(y-x)dx.$$

This completes the proof.

10.3.2 The case when $E\mathbf{v} \neq 1$

Now assume that $\mathbf{x}$ is an exponential random variable with mean one and $\mathbf{v}$ is an exponential random variable with mean $\mu = \mu_{\mathbf{v}} \neq 1$. As before, the product $f_{\mathbf{x}}(x)f_{\mathbf{v}}(y-x)$ is zero outside the interval $[0,y]$. Recall that $\mathbf{y} = \mathbf{x}+\mathbf{v}$. According to Lemma 10.3.1, the density function for $f_{\mathbf{y}}(y)$ when $y \geq 0$ is given by the convolution formula

$$\begin{aligned} f_{\mathbf{y}}(y) &= \int_{-\infty}^{\infty} f_{\mathbf{x}}(y-x)f_{\mathbf{v}}(x)dx = \frac{1}{\mu}\int_0^y e^{-(y-x)}e^{-x/\mu}dx \\ &= \frac{e^{-y}}{\mu}\int_0^y e^{x(\mu-1)/\mu}dx = \left.\frac{e^{-y}e^{x(\mu-1)/\mu}}{\mu-1}\right|_0^y = \frac{e^{-y/\mu}-e^{-y}}{\mu-1}. \end{aligned}$$

Furthermore, if $y < 0$, then $f_{\mathbf{y}}(y) = 0$. Thus the density for $\mathbf{y}$ is given by

$$\begin{aligned} f_{\mathbf{y}}(y) &= \frac{e^{-y/\mu} - e^{-y}}{\mu - 1} \qquad \text{if } y \geq 0 \\ &= 0 \qquad\qquad\qquad\quad \text{if } y < 0\,. \end{aligned} \tag{3.9}$$

Since $f_{\mathbf{x}}(x) f_{\mathbf{v}}(y-x) = \mu^{-1} e^{-x} e^{-(y-x)/\mu}$ when $0 \leq x \leq y$ and zero otherwise, Lemma 10.3.1 also shows that the conditional expectation $f_{\mathbf{x}|\mathbf{y}}$ is given by

$$\begin{aligned} f_{\mathbf{x}|\mathbf{y}}(x|y) &= \frac{e^{-y/\mu} e^{-(\mu-1)x/\mu}(\mu - 1)}{(e^{-y/\mu} - e^{-y})\mu} \qquad \text{if } 0 \leq x \leq y \\ &= 0 \qquad\qquad\qquad\qquad\qquad\quad \text{otherwise}\,. \end{aligned} \tag{3.10}$$

Hence the function $\widehat{g}(y) = \int_0^y x f_{\mathbf{x}|\mathbf{y}}(x|y)dx$ is given by

$$\widehat{g}(y) = \frac{\mu e^{-y/\mu} + (y - y\mu - \mu)e^{-y}}{(e^{-y/\mu} - e^{-y})(\mu - 1)} \qquad (\mu = E\mathbf{v} \neq 1)\,. \tag{3.11}$$

In this case, the conditional expectation is given by the nonlinear function $E(\mathbf{x}|\mathbf{y}) = \widehat{g}(\mathbf{y})$. Finally, it is noted that by applying L'Hospital's rule twice to the optimal function $\widehat{g}(y)$ in (3.11), it follows that $\widehat{g}(y)$ converges to $y/2$ as μ tends to one.

10.3.3 The linear estimate

Now consider the problem of finding the best estimate of $\mathbf{x}$ of the form $c_0 + c_1\mathbf{y}$ where c_0 and c_1 are constants, that is,

$$\gamma = \inf\{E|\mathbf{x} - c_0 - c_1\mathbf{y}|^2 : c_0, c_1 \in \mathbb{C}\}\,. \tag{3.12}$$

To solve this problem, let $\mathcal{H}$ be the two dimensional subspace spanned by $\{1, \mathbf{y}\}$. According to the projection theorem, the solution to the optimization problem in (3.12) is given by $P_{\mathcal{H}}\mathbf{x}$ where $P_{\mathcal{H}}$ is the orthogonal projection onto $\mathcal{H}$. Moreover, the estimation error $\gamma = E|\mathbf{x} - P_{\mathcal{H}}\mathbf{x}|^2$. Obviously, the conditional expectation $E(\mathbf{x}|\mathbf{y})$ is a better estimate of $\mathbf{x}$ given $\mathbf{y}$ than $P_{\mathcal{H}}\mathbf{x}$. This follows because the conditional expectation $E(\mathbf{x}|\mathbf{y})$ computes $P_{\mathcal{G}}\mathbf{x}$, where $P_{\mathcal{G}}$ is the orthogonal projection onto $\mathcal{G}$ the subspace of $\mathcal{K}$ spanned by all functions of $\mathbf{y}$. Clearly, $\mathcal{H}$ is subspace of $\mathcal{G}$. So the estimation error in the conditional expectation is smaller, that is,

$$E|\mathbf{x} - P_{\mathcal{G}}\mathbf{x}|^2 \leq E|\mathbf{x} - P_{\mathcal{H}}\mathbf{x}|^2\,.$$

However, in application computing $P_{\mathcal{H}}\mathbf{x}$ is easier and can provide a useful estimate of $\mathbf{x}$. The following result computes $P_{\mathcal{H}}\mathbf{x}$ for our exponential random variable problem.

PROPOSITION 10.3.2 *Let $\mathbf{y} = \mathbf{x}+\mathbf{v}$ where $\mathbf{x}$ and $\mathbf{v}$ are independent exponential random variables with mean one and $\mu = \mu_{\mathbf{v}}$, respectively. Let $\mathcal{H}$ be the subspace spanned by $\{1, \mathbf{y}\}$. Then the orthogonal projection of $\mathbf{x}$ onto $\mathcal{H}$ is given by*

$$P_{\mathcal{H}}\mathbf{x} = \frac{\mu(\mu - 1)}{1 + \mu^2} + \frac{\mathbf{y}}{1 + \mu^2} \qquad \textit{and} \qquad E|\mathbf{x} - P_{\mathcal{H}}\mathbf{x}|^2 = \frac{\mu^2}{1 + \mu^2}\,. \tag{3.13}$$

In particular, if the mean of $\mathbf{v}$ is one, then $P_{\mathcal{H}}\mathbf{x} = \mathbf{y}/2$, and $P_{\mathcal{H}}\mathbf{x} = E(\mathbf{x}|\mathbf{y})$.

If μ is large, then Proposition 10.3.2 shows that $P_{\mathcal{H}}\mathbf{x}$ is approximately equal to 1 the mean of $\mathbf{x}$ and the estimation error $E|\mathbf{x} - P_{\mathcal{H}}\mathbf{x}|^2 \approx 1$. On the other hand, if μ is small, then $P_{\mathcal{H}}\mathbf{x}$ is approximately equal to $\mathbf{y}$ and the estimation error $E|\mathbf{x} - P_{\mathcal{H}}\mathbf{x}|^2 \approx 0$, that is, $\mathbf{x} \approx \mathbf{y}$.

Proof of Proposition 10.3.2. Let ξ be the random variable with values in $\mathbb{R}^2$ given by $\xi = [1, \mathbf{y}]^{tr}$. Clearly, 1 and $\mathbf{y}$ are linearly independent. According to (2.4) in Theorem 2.2.1 in Chapter 2, the orthogonal projection $P_{\mathcal{H}}\mathbf{x} = R_{\mathbf{x}\xi}R_{\xi}^{-1}[1, \mathbf{y}]^{tr}$. If $\mathbf{z}$ is a exponential random variables with mean λ, then a simple calculation shows that

$$E\mathbf{z}^k = \frac{1}{\lambda}\int_0^\infty z^k e^{-z/\lambda}\, dz = k!\lambda^k \qquad (k = 1, 2, 3, \cdots)\,. \tag{3.14}$$

In particular, $E\mathbf{x}^2 = 2$ and $E\mathbf{v}^2 = 2\mu^2$. Using the fact that $\mathbf{x}$ and $\mathbf{v}$ are independent, we obtain

$$R_{\mathbf{x}\xi} = E\mathbf{x}\xi^* = \begin{bmatrix} E\mathbf{x}1 & E\mathbf{x}\mathbf{y} \end{bmatrix} = \begin{bmatrix} 1 & E\mathbf{x}(\mathbf{x}+\mathbf{v}) \end{bmatrix} = \begin{bmatrix} 1 & 2 + E\mathbf{x}E\mathbf{v} \end{bmatrix} = \begin{bmatrix} 1 & 2+\mu \end{bmatrix}.$$

Hence $R_{\mathbf{x}\xi} = \begin{bmatrix} 1 & 2+\mu \end{bmatrix}$. Notice that

$$E|\mathbf{y}|^2 = E(\mathbf{x}+\mathbf{v})^2 = E\mathbf{x}^2 + 2E\mathbf{x}E\mathbf{v} + E\mathbf{v}^2 = 2 + 2\mu + 2\mu^2\,.$$

This readily implies that

$$R_{\xi} = \begin{bmatrix} E1^2 & E1\mathbf{y} \\ E\mathbf{y}1 & E\mathbf{y}^2 \end{bmatrix} = \begin{bmatrix} 1 & 1+\mu \\ 1+\mu & 2+2\mu+2\mu^2 \end{bmatrix}.$$

The determinant of R_ξ is $1 + \mu^2$. The inverse of R_ξ is given by

$$R_{\xi}^{-1} = \frac{1}{1+\mu^2}\begin{bmatrix} 2+2\mu+2\mu^2 & -(1+\mu) \\ -(1+\mu) & 1 \end{bmatrix}.$$

By combining this with $R_{\mathbf{x}\xi} = \begin{bmatrix} 1 & 2+\mu \end{bmatrix}$, we arrive at

$$R_{\mathbf{x}\xi}R_{\xi}^{-1} = (1+\mu^2)^{-1}\begin{bmatrix} \mu^2 - \mu & 1 \end{bmatrix}.$$

Using $P_{\mathcal{H}}\mathbf{x} = R_{\mathbf{x}\xi}R_{\xi}^{-1}[1, \mathbf{y}]^{tr}$, yields the first equation in (3.13). Finally, equation (2.5) in Theorem 2.2.1 in Chapter 2 shows that

$$E|\mathbf{x} - P_{\mathcal{H}}\mathbf{x}|^2 = E\mathbf{x}^2 - R_{\mathbf{x}\xi}R_{\xi}^{-1}R_{\mathbf{x}\xi}^* = 2 - \frac{2+\mu^2}{1+\mu^2} = \frac{\mu^2}{1+\mu^2}\,.$$

This completes the proof.

10.3.4 Exercise

Problem 1. Let $\mathbf{y} = \mathbf{x} + \mathbf{v}$ where $\mathbf{x}$ and $\mathbf{v}$ are independent exponential random variables with mean one and 1/2, respectively. Let $\mathcal{H}$ be the subspace spanned by $\{1, \mathbf{y}\}$, and $\mathcal{G}_2$ the subspace spanned by $\{1, \mathbf{y}, \mathbf{y}^2\}$. Let $P_{\mathcal{H}}$ and $P_{\mathcal{G}_2}$ be respectively be the orthogonal projections

onto $\mathcal{H}$ and $\mathcal{G}_2$. Then compute the estimate $P_{\mathcal{H}}\mathbf{x}$ of $\mathbf{x}$ given $\mathcal{H}$, and the estimate $P_{\mathcal{G}_2}\mathbf{x}$ of $\mathbf{x}$ given $\mathcal{G}_2$. Notice that $P_{\mathcal{H}}\mathbf{x} = g_1(\mathbf{y})$ and $P_{\mathcal{G}_2}\mathbf{x} = g_2(\mathbf{y})$ where g_1 and g_2 are polynomials of the form $g_1(y) = c_0 + c_1 y$ and $g_2(y) = \gamma_0 + \gamma_1 y + \gamma_2 y^2$. Compute the conditional expectation $\widehat{g}(\mathbf{y}) = E(\mathbf{x}|\mathbf{y})$. Compare these three estimates by graphing $g_1(y)$, $g_2(y)$ and $\widehat{g}(y)$. Finally, compute the error estimates

$$E|\mathbf{x} - \widehat{g}(\mathbf{y})|^2 \leq E|\mathbf{x} - P_{\mathcal{G}_2}\mathbf{x}|^2 \leq E|\mathbf{x} - P_{\mathcal{H}}\mathbf{x}|^2 \,.$$

Bibliography

[1] Akhiezer, N.I. and I.M. Glazman, *Theory of Linear Operators in Hilbert Space*, Dover Publishing, New York, 1993.

[2] B.D.O. Anderson and J.B. Moore, *Optimal Filtering*, Prentice Hall, New Jersey, 1979.

[3] Balakrishnan, A.V., *Applied Functional Analysis*, Springer-Verlag, New York, 1976.

[4] P.E. Caines, *Linear Stochastic Systems*, Wiley, New York, 1988.

[5] J. Capon, High-resolution frequency-wave number spectrum analysis, *Proceedings of the IEEE*, **57** (1969) pp. 1408-1418.

[6] J. F. Claerbout, *Fundamentals of Geophysical Data Processing,* McGraw-Hill, New York, 1976.

[7] Conway, J. B., *A Course in Functional Analysis*, Springer-Verlag, Berlin, 1985.

[8] M. J Corless and A.E. Frazho, *Linear Systems and Control: An Operartor Perspective*, Marcel Deckker, New York, 2003.

[9] M. H. A. Davis, *Linear Estimation and Stochastic Control*, Chapman and Hall, New York, 1977.

[10] J. L. Doob, *Stochastic Processes,* John Wiley and Sons, New York, 1953.

[11] C. Foias, A. E. Frazho and P. J. Sherman, A geometric approach to the maximum likelihood spectral estimator for sinusoids in noise, *IEEE Transactions on Information Theory*, **34** (1988) pp. 1066-1070.

[12] C. Foias, A. E. Frazho and P. J. Sherman, A new approach for determining the spectral data of multichannel harmonic signals in noise, *Mathematics of Control, Signals, and Systems*, **3** (1990) pp. 31-43.

[13] C. Foias and A. E. Frazho, *The Commutant Lifting Approach to Interpolation Problems,* Operator Theory: Advances and Applications, **44**, Birkhäuser-Verlag, Basel, 1990.

[14] C. Foias, A.E. Frazho, I. Gohberg and M. A. Kaashoek, *Metric Constrained Interpolation, Commutant Lifting and Systems,* Operator Theory: Advances and Applications, vol 100, Birkhäuser, 1998.

[15] L. Ya. Geronimus, Polynomials orthogonal on a circle and their application, *Zapiski Nauchno-issledovatel'skogo in-ta matematiki i mekhaniki i KhMo*, **19** (1948) pp. 35, 1948.

[16] L. Ya. Geronimus, *Orthogonal Polynomials*, Consultants Bureau, New York, 1961.

[17] Gohberg, I. and S. Goldberg, *Basic Operator Theory*, Birkhäuser, Basel, 1981.

[18] P. Halmos, *A Hilbert Space Problem Book,* Springer- Verlag, New York Inc, 1982.

[19] K. Hoffman, *Banach Spaces of Analytic Functions,* Prentice Hall, New Jersey, 1962.

[20] T. Kailath, An innovations approach to least squares estimation, Part I, *IEEE Transaction on Automatic Control*, **AC-13** (1968) pp. 646-655.

[21] T. Kailath, *Lectures on Linear Least-Squares Estimation*, CISM Courses and Lectures, **140**, Springer-Verlag, New York, 1978.

[22] T. Kailath, *Linear Systems*, Prentice Hall, New Jersey, 1980.

[23] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*, Wiley Interscience, New York, 1972.

[24] R. T. Lacoss, Data adaptive spectral analysis methods, *Geophysics*, **36** (1971) pp. 661-675.

[25] N. Levinson, The Wiener RMS (root mean square) error criterion in filter design and prediction, *J. Math. Physics*, **25** (1947) pp. 261-278.

[26] D.G. Luenberger, *Opimization by Vector Space Methods*, Wiley, New York, 1969.

[27] S. L. Marple, *Digital Spectral Analysis with Applications*, Prentice Hall, 1987.

[28] W. J. Rugh, *Linear System Theory*, Prentice Hall, New Jersey, 1993.

[29] B. Sz.-Nagy and C. Foias, *Harmonic Analysis of Operators on Hilbert Space,* North Holland Publishing Co., Amsterdam-Budapest, 1970.

[30] Taylor, A.E. and D.C. Lay, *Introduction to Functional Analysis*, John Wiley and Sons, New York, 1980.