\[ Y'_1, Y'_2 \] denote a decomposition into two subvectors. Then the mean vector and covariance matrix can be partitioned conformably:

\[
\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.
\]

So \( \Sigma_{21} \) is the covariance between \( Y_2 \) and \( Y_1 \). As already known, both subvectors are multivariate Gaussian, i.e.,

\[
Y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}) \quad Y_2 \sim \mathcal{N}(\mu_2, \Sigma_{22}).
\]

Then using the Schur decomposition (see Proposition 6.5.3 for further details) of \( \Sigma \), and assuming that \( \Sigma_{22} \) is invertible, we obtain the following result (via factorization of the joint pdf of \( Y_1 \) and \( Y_2 \)) on the conditional distribution of \( Y_1 \) given \( Y_2 = y_2 \), namely:

\[
Y_1 | \{ Y_2 = y_2 \} \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2}) \quad \text{(2.1.4)}
\]

\[
\mu_{1|2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (y_2 - \mu_2) \quad \text{(2.1.5)}
\]

\[
\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}. \quad \text{(2.1.6)}
\]

Note that: (i) the independence of \( Y_1 \) and \( Y_2 \) is equivalent to \( \Sigma_{12} \) being a zero matrix, in which case the conditional distribution of \( Y_1 \) given \( Y_2 = y_2 \) is equal to the unconditional distribution of \( Y_1 \), i.e., uncorrelatedness implies independence in the case of joint (multivariate) normality; and (ii) the conditional expectation of \( Y_1 \) given \( Y_2 = y_2 \) is linear/affine as a function of the given quantity \( y_2 \).

**Remark 2.1.15. Decorrelation by Orthogonal Transformation**  Another application of Facts 2.1.7, 2.1.8 and 2.1.12 is to decorrelate random vectors. Suppose \( X \sim \mathcal{N}(0, \Sigma) \) with \( \Sigma \) invertible; applying Fact 2.1.7, we obtain an orthogonal matrix \( P \) such that \( Y = P' X \) has covariance matrix \( \Lambda \), a diagonal matrix; see also Exercise 2.6. Hence the components of \( Y \) are independent. If \( X \) is non-normal, but still has covariance matrix \( \Sigma \), then \( Y \) will have uncorrelated components (but they may be dependent). Furthermore, if we let \( Z = \Lambda^{-1/2} Y \) then the covariance matrix of \( Z \) is \( 1_n \). If \( X \) is Gaussian, then so is \( Z \), and the component of \( Z \) are i.i.d. \( \mathcal{N}(0, 1) \).

There is a converse to Fact 2.1.12, in the sense that the affine property characterizes the Gaussian distribution. To discuss this result, we need the concept of a characteristic function discussed more fully in Definition C.3.5 of Appendix C.

**Proposition 2.1.16. (Cramér-Wold device)**

\[
X \sim \mathcal{N}(\mu, \Sigma) \iff a' X \text{ is univariate normal for any } a \in \mathbb{R}^n \setminus \{0\}.
\]
the sample mean of the time series over each such window (see Paradigm 1.3.1). Hence, estimator (3.1.2) is sometimes called a moving average.\footnote{This is a different notion from the Moving Average process defined in Remark 2.5.7.}

Since \( \mu_t \) changes slowly with \( t \), we can write \( \mu_{t+s} \approx \mu_t \) if \(|s|\) is small. Hence,

\[
E[\hat{\mu}_t] = \frac{1}{2m+1} \sum_{s=-m}^{m} E[X_{t+s}] \approx \mu_t \quad \text{when } m \text{ is small},
\]

i.e., \( \hat{\mu}_t \) is approximately unbiased as an estimator of \( \mu_t \). The weights in equation (3.1.2) are just the reciprocals of \( 2m + 1 \), but they can be made more sophisticated through the device of a kernel.

**Definition 3.1.2.** A kernel is a weighting function \( K(t) \) that is symmetric and attains its maximum value at \( t = 0 \). A kernel estimator of the nonparametric trend \( \mu_t \) in (3.1.1) is a weighted average of the data, with weights determined by a kernel; the estimator is defined as

\[
\hat{\mu}_t = \frac{\sum_{s=1}^{n} K((s-t)/m) X_s}{\sum_{s=1}^{n} K((s-t)/m)}.
\]

The parameter \( m \) is called the bandwidth. Here \( n \) denotes the sample size.

The denominator in (3.1.4) ensures that the set of weights in the estimator always add up to unity – this is important in order to claim that estimator \( \hat{\mu}_t \) has negligible bias by analogy to equation (3.1.3).

**Remark 3.1.3.** Rectangular Kernel Recall Definition A.3.2 for the indicator of a set. Utilizing the kernel \( K(x) = 1_{[-1,1]}(x) \) in (3.1.4) yields the simple (unweighted) moving average estimator (3.1.2); this is called the rectangular or “box” kernel. The choice of the kernel \( K \) determines the statistical properties of the kernel estimator, such as bias and variance; however, bandwidth choice is often more crucial.

**Remark 3.1.4.** Role of Bandwidth The role of the bandwidth \( m \) in (3.1.4) is similar to that of \( m \) in (3.1.2): it defines a neighborhood of time values near to the given time \( t \) of interest. Large bandwidth entails a large neighborhood and more smoothing – local features are suppressed. Small bandwidth entails a small neighborhood, so that local features are emphasized. Especially in the rectangular kernel case where \( m \) is just the (half)width of the moving window, it is apparent that less averaging is done when \( m \) is small. If \( m \) is too small, undersmoothing occurs and is often visible in plotting \( \hat{\mu}_t \) as a function of \( t \); e.g., in the extreme case that \( m = 0 \), we simply have \( \hat{\mu}_t = X_t \). If \( m \) is large, there is more averaging but if \( m \) is too large, oversmoothing occurs; in the largest case possible, \( \hat{\mu}_t \) becomes the sample mean which is flat/constant as a function of \( t \). A good bandwidth choice strives for the “sweet spot” between undersmoothing and oversmoothing. There is a lot of literature on optimal bandwidth choice but the usefulness of looking at plots of \( \hat{\mu}_t \) as a function of \( t \) can not be over-emphasized.
CHAPTER 6. TIME SERIES IN THE FREQUENCY DOMAIN

Because the autocovariance sequence is even, i.e., \( \gamma(-k) = \gamma(k) \), and using the fact that \( e^{-i\lambda k} + e^{i\lambda k} = 2\cos(\lambda k) \), it follows that

\[
f(\lambda) = \sum_{k=-\infty}^{\infty} \gamma(k) e^{-i\lambda k} = \gamma(0) + 2 \sum_{k=1}^{\infty} \gamma(k) \cos(\lambda k), \tag{6.1.7}
\]

which implies that the spectral density is always real-valued, and an even function of \( \lambda \). A much less obvious fact – proved in Corollary 6.4.10 in what follows – is that the spectral density of a stationary process is non-negative everywhere, i.e., \( f(\lambda) \geq 0 \) for all \( \lambda \in [-\pi, \pi] \); this is due to the non-negative definite property of the autocovariance sequence.

The action of a filter on a time series has an elegant representation in terms of spectral densities, as shown in the following corollary of Theorem 5.6.6.

Corollary 6.1.9. Suppose that (5.6.2) holds, i.e., \( Y_t = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j} \), and let \( f_x \) and \( f_y \) be the respective spectral densities of the stationary input series \( \{X_t\} \) and the output series \( \{Y_t\} \). Then the following equation gives the relationship between these two spectral densities, in terms of the transfer function:

\[
f_y(\lambda) = |\psi(e^{-i\lambda})|^2 f_x(\lambda) \tag{6.1.8}
\]

for all \( \lambda \in [-\pi, \pi] \), where \( \psi(B) = \sum_{j=-\infty}^{\infty} \psi_j B^j \).

Proof of Corollary 6.1.9. Replace \( z \) by \( e^{-i\lambda} \) and \( z^{-1} \) by \( e^{i\lambda} \) in Theorem 5.6.6, and note that \( \psi(e^{i\lambda}) = \psi(e^{-i\lambda}) \).

Fact 6.1.10. Frequency Response Function  
Evaluating the transfer function of a filter \( \psi(B) \) at \( z = e^{-i\lambda} \), and viewing it as a (complex-valued) function of \( \lambda \in [-\pi, \pi] \) results in what is known as the frequency response function of the filter. The absolute value \( |\psi(e^{-i\lambda})| \) of the frequency response function is called the gain function, and its square \( |\psi(e^{-i\lambda})|^2 \) is called the squared gain function.

To compute the autocovariance of the output \( Y_t = \psi(B)X_t \), we can determine the Fourier coefficients of the squared gain function \( |\psi(e^{-i\lambda})|^2 \), and convolve these with the acvf of \( \{X_t\} \); this is an application of the convolution formula, given below (see Exercise 6.2 for the proof).

Proposition 6.1.11. Convolution Formula  
Consider two functions \( f(\lambda) \) and \( g(\lambda) \) belonging to \( L_2[-\pi, \pi] \); expand them in Fourier series to obtain

\[
f(\lambda) = \sum_{k=-\infty}^{\infty} \langle f \rangle_k e^{-i\lambda k} \quad \text{and} \quad g(\lambda) = \sum_{k=-\infty}^{\infty} \langle g \rangle_k e^{-i\lambda k}. \tag{6.1.9}
\]

The Fourier coefficients of the product \( f(\lambda)g(\lambda) \) are given by the discrete convolution of the Fourier coefficients of \( f(\lambda) \) and \( g(\lambda) \) respectively, i.e.,

\[
\langle fg \rangle_k = \sum_{h=-\infty}^{\infty} \langle f \rangle_{h-k} \langle g \rangle_k. \tag{6.1.10}
\]


Pairing completeness with the notion of inner product yields a so-called *Hilbert space*.

**Definition 4.3.4.** *An inner product space that is complete is called a Hilbert space.*

**Fact 4.3.5. Inner Product Space Completeness** *An inner product space is complete if and only if it is closed.*

**Example 4.3.6. A Hilbert space on \( \mathbb{R} \)** Consider the vector space \( \mathbb{R} \) with inner product given by the scalar product, and let \( x_n = 1/n \) for \( n \geq 1 \) be a sequence; this is clearly a Cauchy sequence that converges to 0, which lies in \( \mathbb{R} \). It can be shown that Euclidean vector spaces are complete.

**Example 4.3.7. Not a Hilbert Space** Consider the vector space \( (0, 1] \) with scalar product for inner product. Then, the sequence \( x_n = 1/n \) is Cauchy; it tends to 0 \( \not\in (0, 1] \), so the sequence does not converge to an element of the space. Hence \( (0, 1] \) is not complete, and is not a Hilbert space. Note that this is consistent with Fact 4.3.5, since \( (0, 1] \) is not closed.

**Fact 4.3.8. Common Hilbert spaces** The spaces \( \mathbb{R}^n \), \( \ell_2 \), and \( L_2 \) (see Example 4.1.9 and Definition 4.2.1) with their associated inner products, are all Hilbert spaces.

We now list the main properties of a Hilbert space \( \mathcal{H} \) with an inner product denoted by \( \langle x, y \rangle \), and norm \( \|x\| = \sqrt{\langle x, x \rangle} \) for \( x, y \in \mathcal{H} \).

**Theorem 4.3.9.** Let \( \mathcal{H} \) be a Hilbert space, and let \( x, y, z \in \mathcal{H} \) and \( a \in \mathbb{R} \). Then:

1. \( \langle x, y \rangle = \langle y, x \rangle \) (symmetry)
2. \( \langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle \) (linearity in the first argument)
3. \( \langle ax, z \rangle = a \langle x, z \rangle \) (linearity in the first argument)
4. \( \|x\| \geq 0 \) with equality\(^1\) if and only if \( x = 0 \).
5. **Cauchy-Schwarz inequality:** \( |\langle x, y \rangle| \leq \|x\| \cdot \|y\| \) with equality if \( x = ay + b \) for some \( a \in \mathbb{R} \) and \( b \in \mathcal{H} \).
6. **Triangle inequality:** \( \|x + y\| \leq \|x\| + \|y\| \)
7. \( \|ax\| = |a| \|x\| \)
8. **Parallelogram law:** \( \|x + y\|^2 + \|x - y\|^2 = 2 \|x\|^2 + 2 \|y\|^2 \)
9. **Continuity of the inner product:** if \( \|x_n - x\| \to 0 \) and \( \|y_n - y\| \to 0 \) as \( n \to \infty \), then \( \|x_n\| \to \|x\| \) and \( \langle x_n, y_n \rangle \to \langle x, y \rangle \) as \( n \to \infty \).
10. **Completeness:** if \( x_n \) is Cauchy, then there exists some \( x \in \mathcal{H} \) such that \( x_n \to x \) in norm.

\(^1\) Caveat: in \( L_2(\Omega, \mathcal{F}, \mathbb{P}) \) this is weakened to \( x = 0 \) with probability one.