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

$$\underline{\mu} = \begin{bmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

So Σ_{21} *is the covariance between* $\underline{Y}_2$ *and* $\underline{Y}_1$*. As already known, both subvectors are multivariate Gaussian, i.e., :*

$$\underline{Y}_1 \sim \mathcal{N}\left(\underline{\mu}_1, \Sigma_{11}\right) \qquad \underline{Y}_2 \sim \mathcal{N}\left(\underline{\mu}_2, \Sigma_{22}\right).$$

Then using the Schur decomposition (see **Proposition 6.5.3** *for further details) of* Σ*, and assuming that* Σ_{22} *is invertible, we obtain the following result (via factorization of the joint pdf of* $\underline{Y}_1$ *and* $\underline{Y}_2$*) on the conditional distribution of* $\underline{Y}_1$ *given* $\underline{Y}_2 = \underline{y}_2$*, namely:*

$$\underline{Y}_1 | \{\underline{Y}_2 = \underline{y}_2\} \sim \mathcal{N}\left(\underline{\mu}_{1|2}, \Sigma_{1|2}\right) \tag{2.1.4}$$

$$\underline{\mu}_{1|2} = \underline{\mu}_1 + \Sigma_{12}\, \Sigma_{22}^{-1} \left(\underline{y}_2 - \underline{\mu}_2\right) \tag{2.1.5}$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\, \Sigma_{22}^{-1}\, \Sigma_{21}. \tag{2.1.6}$$

Note that: (i) the independence of $\underline{Y}_1$ *and* $\underline{Y}_2$ *is equivalent to* Σ_{12} *being a zero matrix, in which case the conditional distribution of* $\underline{Y}_1$ *given* $\underline{Y}_2 = \underline{y}_2$ *is equal to the unconditional distribution of* $\underline{Y}_1$*, i.e., uncorrelatedness implies independence in the case of joint (multivariate) normality; and (ii) the conditional expectation of* $\underline{Y}_1$ *given* $\underline{Y}_2 = \underline{y}_2$ *is linear/affine as a function of the given quantity* $\underline{y}_2$*.*

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 $\underline{X} \sim \mathcal{N}(\underline{0}, \Sigma)$ with Σ invertible; applying Fact 2.1.7, we obtain an orthogonal matrix P such that $\underline{Y} = P'\,\underline{X}$ has covariance matrix Λ, a diagonal matrix; see also Exercise 2.6. Hence the components of $\underline{Y}$ are independent. If $\underline{X}$ is non-normal, but still has covariance matrix Σ, then $\underline{Y}$ will have uncorrelated components (but they may be dependent). Furthermore, if we let $\underline{Z} = \Lambda^{-1/2}\,\underline{Y}$ then the covariance matrix of $\underline{Z}$ is 1_n. If $\underline{X}$ is Gaussian, then so is $\underline{Z}$, and the component of $\underline{Z}$ 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)*

$$\underline{X} \sim \mathcal{N}(\underline{\mu}, \Sigma) \;\Leftrightarrow\; \underline{a}'\underline{X} \text{ is univariate normal for any } \underline{a} \in \mathbb{R}^n \setminus \{0\}.$$

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*.[1]

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

$$\mathbb{E}[\widehat{\mu}_t] = \frac{1}{2m+1} \sum_{s=-m}^{m} \mathbb{E}[X_{t+s}] \approx \mu_t \quad \text{when } m \text{ is small,} \tag{3.1.3}$$

i.e., $\widehat{\mu}_t$ is approximately unbiased as an estimator of μ_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 μ_t in (3.1.1) is a weighted average of the data, with weights determined by a kernel; the estimator is defined as*

$$\widehat{\mu}_t = \frac{\sum_{s=1}^{n} K((s-t)/m)\, X_s}{\sum_{s=1}^{n} K((s-t)/m)}. \tag{3.1.4}$$

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 $\widehat{\mu}_t$ 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 $K(x) = \mathbf{1}_{[-1,1]}(x)$ 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 $\widehat{\mu}_t$ as a function of t; e.g., in the extreme case that $m = 0$, we simply have $\widehat{\mu}_t = X_t$. If m is large, there is more averaging but if m is too large, *oversmoothing* occurs; in the largest case possible, $\widehat{\mu}_t$ 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 $\widehat{\mu}_t$ as a function of t can not be overemphasized.

[1] This is a different notion from the Moving Average *process* defined in Remark 2.5.7.

Fact 6.1.8. Further Properties of the Spectral Density *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 λ. 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:*

$$f_y(\lambda) = \left|\psi(e^{-i\lambda})\right|^2 f_x(\lambda) \tag{6.1.8}$$

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 $\psi(e^{i\lambda}) = \overline{\psi(e^{-i\lambda})}$. □

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 $\left|\psi(e^{-i\lambda})\right|^2$ 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 $\left|\psi(e^{-i\lambda})\right|^2$, 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 $\mathbb{L}_2[-\pi, \pi]$; expand them in Fourier series to obtain*

$$f(\lambda) = \sum_{k=-\infty}^{\infty} \langle f \rangle_k\, e^{-i\lambda k} \quad \textit{and} \quad g(\lambda) = \sum_{k=-\infty}^{\infty} \langle g \rangle_k\, e^{-i\lambda k}. \tag{6.1.9}$$

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.,

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

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 $0 \notin (0,1]$, 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 $\mathbb{R}^n$, ℓ_2, and $\mathbb{L}_2$ (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 $\langle \underline{x}, \underline{y} \rangle$, and norm $\|\underline{x}\| = \sqrt{\langle \underline{x}, \underline{x} \rangle}$ for $\underline{x}, \underline{y} \in \mathcal{H}$.

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

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

[1] Caveat: in $\mathbb{L}_2(\Omega, \mathbb{P}, \mathcal{F})$ this is weakened to $\underline{x} = 0$ with probability one.