Lessons 9-10: Multidimensional prediction, Kálmán's filering, and dynamic PCA

Marianna Bolla and Tamás Szabados, BME Math. Inst.

December 7, 2020

In this lesson we deal with the prediction of stochastic processes in general and in the weakly stationary case. We consider one-step and more-step ahead predictions based on finitely many past values or on the infinite past. Actually, the original paper of H. Wold is about 1D, weakly stationary time series, and constructs the famous decomposition via one-step ahead predictions based on the n-length long past with usual multivariate regression techniques, while making use of stationarity as well. Then, at a passage to infinity ($n \to \infty$) he gets the formula for the one-step ahead prediction based on the infinite past. In this way, he decomposes the regular part of a weakly stationary 1D time series as the infinite sum of the innovations that also form a weakly stationary process, namely, a white noise process with the smallest obtainable variance of the linear prediction error. Orthogonality (uncorrelatedness) of the innovation η_t and the past $X_{t-1}, X_{t-2}, \dots$ of X_t is the consequence of the projection principle used in multivariate regression.

The generalization to a multivariate process $\{\mathbf{X}_t\}$ is straightforward with the observation that here parallel multivariate linear regressions are used for the components of $\mathbf{X}_t$ based on all the components of $\mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-n}$. The error terms, $\boldsymbol{\eta}_t$s and their covariance matrices are obtainable by the block Cholesky decomposition of the block Toeplitz matrix $\mathfrak{C}_n$ already used in Chapter 1. At a passage to infinity, we get the multi-dimensional Wold decomposition that is more complicated than the 1D one in that here only the so-called innovation subspaces are unique, the dimension of which is the same as the rank of the spectral density matrix of the weakly stationary $\{\mathbf{X}_t\}$. If this rank r is less than the dimension d of the process, then the error covariance matrix of $\boldsymbol{\eta}_t$ is singular (of rank r), but usually not the zero matrix. In this case, within the innovation subspaces, with usual factor analysis techniques, a $\{\boldsymbol{\xi}_t\} \sim \mathrm{WN}(\boldsymbol{I}_r)$ process can be constructed (up to orthogonal rotation) such that it appears in the multi-dimensional Wold decomposition instead of the innovation subspaces. Actually, this is the task of the dynamic factor analysis when the low-dimensional approximation is not always straightforward but is obtainable with spectral approximations under the conditions of the GDFM (Generalized Dynamic Factor Model).

We also establish asymptotic relations between the spectrum of $\mathfrak{C}_n$ and the spectra of spectral density matrices at the Fourier frequencies, for 'large' n. In this way, the spectra of spectra, that is the spectral decomposition of these matrices plays a crucial rule in dimension reduction (see Chapter 4), and gives rise to computationally more tractable algorithms as the above block Cholesky decomposition.

The technique of the Kálmán's filtering is also introduced together with a recursion to obtain the innovations and the newer and newer predictions for the state variable of a state space system, while using only the newcoming observed variable and the preceding estimate of the state variable. In the heart of the recursion there is the propagation of the error covariance matrices.

1 1D prediction of weakly stationary processes in the time domain

1.1 One-step ahead prediction based on finitely many past values

We have a 1D time series $\{X_t\}$ which is not necessarily stationary, for now; we just assume the existence of the second moments (cross-autocovariances). For simplicity, the state space is $\mathbb{R}$, but the time is discrete ($t \in \mathbb{Z}$).

Assume that $\mathbb{E}(X_t) = 0$ ($t \in \mathbb{Z}$). Select a starting observation X_1 and $H_n := \mathrm{Span}\{X_1, \dots, X_n\}$. We want to linearly predict X_{n+1} based on random past values $X_1, \dots, X_n$. Let $\hat{X}_1 := 0$, and denote by $\hat{X}_{n+1}$ the best linear prediction that minimizes the mean square error $\mathbb{E}(X_{n+1} - \hat{X}_{n+1})^2$, $n = 1, 2, \dots$. If we consider the Hilbert space of the random variables with 0 expectation and finite variance, where the inner product is the covariance (see the background material), then we have $\mathbb{E}(X_{n+1} - \hat{X}_{n+1})^2 = \|X_{n+1} -$

$\hat{X}_{n+1}\|^2$. By the general theory of Hilbert spaces, $\hat{X}_{n+1} = \mathrm{Proj}_{H_n} X_{n+1}$, i.e. the projection of X_{n+1} onto the linear subspace H_n. In the Gaussian case, the solution is $\hat{X}_{n+1} = \mathbb{E}(X_{n+1} \,|\, X_1, \dots, X_n)$, which is the regression plane, but the coefficients of the optimal linear predictor

$$\hat{X}_{n+1} = a_{n1} X_n + \cdots + a_{nn} X_1$$

can be obtained in the non-Gaussian case too, by solving a system of linear equations that contains the second moments and the second cross-moments of the involved random variables as follows by the theory of multivariate linear regression (see also Appendix C).

With the notations $\mathbf{a}_n = (a_{n1}, \dots, a_{nn})^T$, $\boldsymbol{C}_n = [\mathrm{Cov}(X_i, X_j)]_{i,j=1}^n$ and $\mathbf{d}_n = (\mathrm{Cov}(X_{n+1}, X_n), \dots, \mathrm{Cov}(X_{n+1}, X_1))^T$, we have to solve the following system of linear equations (Gauss normal equations):

$$\boldsymbol{C}_n \mathbf{a}_n = \mathbf{d}_n. \tag{1}$$

A solution (the projection) always exists, and it is unique if $\boldsymbol{C}_n$ is positive definite. Then the unique solution is $\mathbf{a}_n = \boldsymbol{C}_n^{-1} \mathbf{d}_n$. Otherwise, there are infinitely many solutions, and we can give them similarly, with any generalized inverse of the positive semidefinite matrix $\boldsymbol{C}_n$. In this case, there are linear relations between $X_1, \dots, X_n$, and so, infinitely many linear combinations of them produce the same projection of X_{n+1} onto the subspace spanned by them. In case of a singular $\boldsymbol{C}_n$ it is customary to use the (unique) Moore–Penrose inverse (see Appendix B) that gives the particular solution $\mathbf{a}_n = \boldsymbol{C}_n^{+} \mathbf{d}_n$.

In particular, if $\{X_t\}$ is stationary, then $\boldsymbol{C}_n = [c(i-j)]_{i,j=1}^n$, so $\boldsymbol{C}_n$ is a Toeplitz matrix, and $d_n(j) = c(j)$, $j = 1, \dots, n$. Therefore, the solution $\mathbf{a}_n$ does not depend on the selection of the starting time of the starting observation X_1. In this case, no double indexing for the coordinates of the vector $\mathbf{a}_n$ is necessary, they can as well be written as $a_1, \dots, a_n$.

Also, when $\{X_t\}$ is stationary, then under very general conditions, there is a unique solution as discussed below. Some remarks are in order.

Remark 1. *Namely, if* $c(0) > 0$ *and* $\lim_{h \to \infty} c(h) = 0$, *then the autocovariance matrix* $\boldsymbol{C}_n = [c(i-j)]_{i,j=1}^n$ *of* $(X_1, \dots, X_n)^T$ *is positive definite for every* $n \in \mathbb{N}$.

Remark 2. *For 'large'* n, *the eigenvalues of* $\boldsymbol{C}_n$ *are asymptotically the same as the union of the values of the spectral density* f *at the Fourier frequencies.*

In Section 3, we will generalize this statement for multidimensional time series.

Note that, in the stationary case we can estimate $\mathbf{a}_n$ if we have a sample, i.e. the set of n-length long windows $X_{1+t}, \dots, X_{n+t}$, $t = 0, \dots, T$. The estimate of $\mathbf{a}_n$ is based on the sample estimates of $\boldsymbol{C}_n$ and $\mathbf{d}_n$, see Section 1.4. There it is discussed that T shuld be 'much larger' than n so that to satisfy ergodicity. For increasing n, there are recursions to find the components of $\mathbf{a}_n$; for example, the Durbin–Levinson algorithm.

Considering the the decomposition

$$X_{n+1} = \hat{X}_{n+1} + \eta_{n+1},$$

where $\hat{X}_{n+1} = \mathbf{a}_n^T \mathbf{X}_n$ and η_{n+1} is the error term, it is easy to see (background material) that the two right-hand side terms are orthogonal (uncorrelated), therefore their variances are added together:

$$\|X_{n+1}\|^2 = \|\hat{X}_{n+1}\|^2 + \|\eta_{n+1}\|^2.$$

With our notation it yields

$$\begin{aligned} c(0) &= \mathrm{Var}(\mathbf{a}_n^T \mathbf{X}_n) + \mathrm{Var}(\eta_{n+1}) = \mathrm{Var}(\mathbf{d}_n^T \boldsymbol{C}_n^{-1} \mathbf{X}_n) + \mathrm{Var}(\eta_{n+1}) \\ &= \mathbf{d}_n^T \boldsymbol{C}_n^{-1} \mathbf{d}_n + \mathrm{Var}(\eta_{n+1}). \end{aligned}$$

Therefore, the prediction error, that is the variance of the error term, is

$$e_n^2 = \|\eta_{n+1}\|^2 = \mathrm{Var}(\eta_{n+1}) = c(0) - \mathbf{d}_n^T \boldsymbol{C}_n^{-1} \mathbf{d}_n. \tag{2}$$

It will be further analyzed in Section 1.2.

Note that equation (1) is exactly the same as the first n Yule-Walker equations for estimating the parameters of a stationary AR(n) process.

The $AR(n)$ process is

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_n X_{t-n} + \eta_t, \quad t = 0, \pm 1, \pm 2, \dots \tag{3}$$

where $\{\eta_t\} \sim \mathrm{WN}(\sigma^2)$ is a white noise process, where σ^2 is also estimated. In case of second order processes, due to the projection principle, it also comes out that η_t (the orthogonal component) is uncorrelated with the regressor, and so with the past values $X_{t-1}, X_{t-2}, \dots$ too.

The Yule-Walker equations based on the first n autocovariances are:

$$c(k) = \begin{cases} a_1 c(1) + \cdots + a_n c(n) + \sigma^2, & k = 0 \\ a_1 c(k-1) + \cdots + a_n c(k-n), & k = 1, \dots, n \end{cases} \tag{4}$$

For real-valued time series, the Yule-Walker equations (4) for $k = 1, \dots, n$ can be written in matrix form:

$$\begin{bmatrix} c(0) & c(1) & \dots & c(n-1) \\ c(1) & c(0) & \dots & c(n-2) \\ \vdots & \vdots & \vdots & \vdots \\ c(n-1) & c(n-2) & \dots & c(0) \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} c(1) \\ c(2) \\ \vdots \\ c(n) \end{bmatrix}. \tag{5}$$

These are the Gauss normal equations, see the background material. If the coefficient matrix is strictly positive definite, (positive semidefiniteness is always true), then we have a unique solution. Substituting this solution in the first equation of (4), which is the same as equation (2), provides the solution for σ^2. Here $\sigma^2 = e_n^2$ if the order n of the AR process is fixed.

Also, in case of a stable AR(n) process, the first n Yule–Walker equations imply the next ones; while in other cases, the solution of the first n Yule–Walker equations just gives the best prediction using n past values, and they are rather called Gauss normal equations.

Remark 3. *If for some $n \geq 1$ the covariance matrix $\boldsymbol{C}_n$ is positive definite, then the nth degree AR polynomial $\alpha(z)$ is causal in the sense that $\alpha(z) \neq 0$ for $z \leq 1$.*

Remark 4. *Comparing Remarks 1 and 3, we can conclude the following. If for the autocovariance function of the process $\{X_t\}$, $c(0) > 0$ and $\lim_{h\to\infty} c(h) = 0$ hold, then the autocovariance matrix $\boldsymbol{C}_n = [c(i-j)]_{i,j=1}^n$ of $(X_1, \dots, X_n)^T$ assigned to the process is positive definite for every $n \in \mathbb{N}$. Consequently, the process $\{X_t\}$ has a (unique) stable AR(n) representation such that the first n autocovariances of it are $c(0), \dots, c(n-1)$, for every $n \in \mathbb{N}$. However, if the sequence $c(h)$ tends to 0, but not exponentially fast, these AR(n) representations based on just $X_1, \dots, X_n$ do not approximate the process at all, and the process is not even necessarily regular.*

On the contrary, if $\boldsymbol{C}_n$ is singular for some n (and consequently, for larger ns too), then using its the generalized inverse, we get an AR(n) solution, but it is not stable.

It is also important, that in case of a stationary process, the h-step ahead prediction, i.e. the prediction of X_{n+h} based on $X_1, \dots, X_n$ can be easily concluded from the one-step ahead prediction, for $h = 1, 2, \dots$. In view of $H_n \subset H_{n+h-1}$,

$$\mathrm{Proj}_{H_n} X_{n+h} = \mathrm{Proj}_{H_n} \mathrm{Proj}_{H_{n+h-1}} X_{n+h} = \mathrm{Proj}_{H_n} \hat{X}_{n+h},$$

we get the equation

$$\boldsymbol{C}_n \mathbf{a}_n = \mathbf{d}_n(h) \tag{6}$$

for the coefficients of the prediction in $\mathbf{a}_n$, where

$$\mathbf{d}_n(h) = [\mathrm{Cov}(X_{n+h}, X_n), \dots, \mathrm{Cov}(X_{n+h}, X_1)]^T,$$

and it is $[c(h), \dots, c(n+h-1)]^T$ in the stationary case. Equation (1) is the special case when $h = 1$ and $\mathbf{d}_n = \mathbf{d}_n(1)$.

1.2 Innovations

Observe that, by the Gram–Schmidt procedure, the prediction error terms form an orthogonal sequence, and they are called *innovations*. In this way, X_ns can as well be expressed in terms of the inovations; in other words, X_n can be written as the linear combination of the normalized error terms that form a complete orthonormal system in H_n. Moreover, this is true in each step of the Gram–Schmidt process, so in the expansion of each X_n only the same and lower index error terms appear. We do it as follows.

First, $\{X_t\}$ is not necessarily stationary. Recall that $H_n := \mathrm{Span}\{X_1, \dots, X_n\}$. Let $\eta_{n+1} := X_{n+1} - \hat{X}_{n+1}$ be the one-step ahead prediction error term, based on the n-length long past, for $n = 0, 1, \dots$. As $\hat{X}_1 = 0$ (see Section 1.1), $\eta_1 = X_1 \in H_1$ and the unique orthogonal decomposition

$$X_2 = \hat{X}_2 + \eta_2$$

works, where $\hat{X}_2 \in H_1$ and $\eta_2 \perp H_1$, whenever $H_1 \subset H_2$ is a proper subspace (disregard the situation $H_1 = H_2$, when $\eta_2 = 0$). Therefore, $\eta_2 \in H_2$ and $\eta_2 \perp \eta_1$. Further,

$$X_2 = l_{21}\eta_1 + \eta_2.$$

With the same considerations,

$$X_{j+1} = \hat{X}_{j+1} + \eta_{j+1}, \quad j = 2, \dots, n$$

with $\hat{X}_{j+1} \in H_j$ and $\eta_{j+1} \perp H_j$ if $H_j \subset H_{j+1}$ is a proper subspace. So $\eta_{j+1} \in H_{j+1}$ and $\eta_{j+1} \perp \eta_j$.

In this way, we get the innovations $\eta_1, \dots, \eta_n$, the linear combination of which produces X_k as

$$X_1 = \eta_1, \quad X_k = \sum_{j=1}^{k-1} l_{kj}\eta_j + \eta_k, \quad k = 2, \dots, n,$$

where the coefficients l_{kj} are obtained recursively, together with the mean square one-step ahead prediction errors $\sigma_k^2 = \|\eta_k\|^2$ $(k = 1, 2, \dots, n)$.

Actually, this is the LDL (variant of the Cholesky) decomposition, see the background material (Complex Matrices). Indeed, with the notation $\boldsymbol{\eta}_n = (\eta_1, \dots, \eta_n)^T$ and $\mathbf{X}_n = (X_1, \dots, X_n)^T$, we have to find an $n \times n$ lower triangular matrix $\boldsymbol{L}_n$ with entries l_{kj}s and all 1's along its main diagonal such that

$$\mathbf{X}_n = \boldsymbol{L}_n \boldsymbol{\eta}_n. \tag{7}$$

Taking the covariance matrices on both sides, yields

$$\boldsymbol{C}_n = \boldsymbol{L}_n \boldsymbol{D}_n \boldsymbol{L}_n^T. \tag{8}$$

If $\boldsymbol{C}_n$ is positive definite, then $\boldsymbol{D}_n = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ is positive definite too. $\boldsymbol{L}_n$ is also nonsingular (with diagonal entries 1s), hence $\boldsymbol{\eta}_n = \boldsymbol{L}_n^{-1}\mathbf{X}_n$, where $\boldsymbol{L}_n^{-1}$ is also lower triangular; therefore, the innovations can as well be written in terms of the same or lower index X_ts. So the LDL decomposition gives the prediction errors (diagonal entries of $\boldsymbol{D}_n$), and the entries of $\boldsymbol{L}_n$ below its main diagonal (the main diagonal is constantly 1). Note that the entries of $\boldsymbol{L}_n$ are obtainable in a nested way, so n does not play an important role here, see the background material. With increasing n, we just extend the rows of $\boldsymbol{L}_n$.

The situation further simplifies in the stationary case, when $\boldsymbol{C}_n$ is a Toeplitz matrix. However, $\boldsymbol{L}_n$ will not be Toeplitz, but asymptotically, it becomes more and more like a Toeplitz one, and the entries of $\boldsymbol{D}_n$ will be more and more similar to each other (the sequence e_n^2 converges) as with $n \to \infty$ the situation resembles the infinite past one (see the Wold decomposition). These issues will be more precisely analyzed in Section 1.3. Actually, this algorithm is also applicable to find MA(q) representation of a process with $n = 1, 2, \dots, q$.

Remark 5. *If the autocovariance function of the zero mean stationary process is such that $c(h) = 0$ for $|h| > q$ and $c(q) \neq 0$, then it is a MA(q) process.*

In the stationary case, the innovation η_n is non-zero if $H_{n-1} \subset H_n$ is a proper subspace. In this case, η_ns are true innovations. Recall that, by Remark 1, this holds true at the same time for any n whenever $c(0) > 0$ and $\lim_{h\to\infty} c(h) = 0$. In this case, we can also standardize the η_ns, and write

$$X_n = \sum_{j=1}^{n} \tilde{l}_{nj}\xi_j, \quad n = 1, 2, \dots$$

where $\xi_j = \eta_j/\sigma_j$ and $\tilde{l}_{nj} = l_{nj}\sigma_j$ for $j = 1, \dots, n$. Here $\xi_1, \dots, \xi_n$ form a complete orthonormal system in H_n. The coefficients $\tilde{l}_{nj}$s are obtainable by the Gram decomposition (see Appendix B)

$$\boldsymbol{C}_n = \boldsymbol{A}_n\boldsymbol{A}_n^T$$

where $\boldsymbol{A}_n = \boldsymbol{L}_n\boldsymbol{D}_n^{1/2}$ can be chosen lower triangular, but it can be post-multiplied with any orthogonal matrix.

1.3 Prediction based on the infinite past

Going farther, in case of a stationary, non-singular process, we can project X_{n+1} onto the infinite past $H_n^- = \overline{\text{span}}\,\{X_t : t \le n\}$ and expand it in terms of an orthonormal system, that is called Wold decomposition. This part will be the regular (causal) part of the process, whereas, the other, singular part, is orthogonal to it. Note that this singular part is of Type (0) deterministic (see Lessons 5-6).

Also, by stationarity, the one-step ahead prediction error

$$\sigma^2 = \|X_{n+1} - \text{Proj}_{H_n^-}X_{n+1}\| = \mathbb{E}(X_{n+1} - \text{Proj}_{H_n^-}X_{n+1})^2$$

does not depend on n, and it is positive, since the process is non-singular. Again, the Wold decomposition gives

$$X_n = \sum_{j=0}^{\infty} b_j\eta_{n-j} + Y_n,$$

where $\{Y_n\}$ is of Type (0) singular and $\{\eta_t\}$ is white noise with variance σ^2, $b_0 = 1$. If $Y_n = 0$ for all n, the process $\{X_n\}$ is regular. The coefficients b_i/σ are the *impulse responses*. Because of the stationarity and infinite past, b has a single index. Here the coefficients bjs are the limiting values of l_{nj}s when $n \to \infty$ in (**??**). It is in accord with the earlier observation that the matrix $\boldsymbol{L}_n$ will be closer and closer to a Toeplitz one, if we disregard the first finitely many rows of it.

Note that the innovation process is a MA(∞) process, which is a causal TLF. Here η_n is not considered as an error term, but rather than positive information that is not contained in the past of X_n. This is why it is called innovation.

Wold derives his celebrated decomposition theorem for real, univariate stationary time series in the following situation: the one-step ahead prediction of X_t is based on its n-length long past and $n \to \infty$.

More precisely, let us fix X_t and consider its one-step ahead prediction, based on its n-length long past. By stationarity, the mean square prediction error does not depend on t, it only depends on n, and was denoted by e_n^2. It can be written in many equivalent forms, see the theory of multivariate regression:

$$e_n^2 = c(0)(1 - r^2_{X_t,(X_{t-1},\dots,X_{t-n})}) = c(0) - \mathbf{d}_n^T \boldsymbol{C}_n^{-1} \mathbf{d}_n,$$

where $r^2_{X_t,(X_{t-1},\dots,X_{t-n})}$ is the squared multiple correlation coefficient between X_t and $(X_{t-1}, \dots, X_{t-n})$; it does not depend on t either, and obviously increases (does not decrease) with n, i.e. $e_1^2 \geq e_2^2 \geq \dots$. The mean square error can as well be written with the determinants of the consecutive Toeplitz matrices $\boldsymbol{C}_n$ and $\boldsymbol{C}_{n+1}$. The next proposition is also used in the original paper of Wold, but here we give a simple proof by means of the determinants of block matrices.

Proposition 1. *If for some n, $|\boldsymbol{C}_n| \neq 0$, then*

$$e_n^2 = c(0) - \mathbf{d}_n^T \boldsymbol{C}_n^{-1} \mathbf{d}_n = \frac{|\boldsymbol{C}_{n+1}|}{|\boldsymbol{C}_n|}. \tag{9}$$

Proof. We use block matrix techniques for the following partitioned matrix:

$$\boldsymbol{C}_{n+1} = \begin{pmatrix} \boldsymbol{C}_n & \mathbf{d}_n \\ \mathbf{d}_n^T & c(0) \end{pmatrix}.$$

It is known that

$$
\begin{aligned}
|\boldsymbol{C}_{n+1}| &= |\boldsymbol{C}_n - \mathbf{d}_n c^{-1}(0)\mathbf{d}_n^T| \cdot |c(0)| \\
&= c(0)|\boldsymbol{C}_n(\boldsymbol{I}_n - \boldsymbol{C}_n^{-1}\mathbf{d}_n\mathbf{d}_n^T/c(0)| \\
&= c(0)|\boldsymbol{C}_n| \cdot |\boldsymbol{I}_n - \boldsymbol{C}_n^{-1}\mathbf{d}_n\mathbf{d}_n^T/c(0)| \\
&= c(0)|\boldsymbol{C}_n| \cdot (1 - \lambda(\boldsymbol{C}_n^{-1}\mathbf{d}_n\mathbf{d}_n^T/c(0))),
\end{aligned}
$$

where $\lambda(\boldsymbol{C}_n^{-1}\mathbf{d}_n\mathbf{d}_n^T/c(0))$ is the only nonzero eigenvalue of the matrix $\boldsymbol{C}_n^{-1}\mathbf{d}_n\mathbf{d}_n^T/c(0)$, which is of rank 1. Indeed, the rank of the dyad $\mathbf{d}_n\mathbf{d}_n^T$ is 1, and the multiplication with another matrix cannot increase this rank. Therefore, the eigenvalues of $\boldsymbol{I}_n - \boldsymbol{C}_n^{-1}\mathbf{d}_n\mathbf{d}_n^T/c(0)$ are $1 - \lambda(\boldsymbol{C}_n^{-1}\mathbf{d}_n\mathbf{d}_n^T/c(0))$ and 1 (with multiplicity $n-1$). So its determinant is

$$
\begin{aligned}
&1 - \lambda(\boldsymbol{C}_n^{-1}\mathbf{d}_n\mathbf{d}_n^T/c(0)) = 1 - \lambda(\boldsymbol{C}_n^{-1}\mathbf{d}_n\mathbf{d}_n^T)/c(0) = 1 - \operatorname{tr}(\boldsymbol{C}_n^{-1}\mathbf{d}_n\mathbf{d}_n^T)/c(0) \\
&= 1 - \operatorname{tr}(\mathbf{d}_n^T\boldsymbol{C}_n^{-1}\mathbf{d}_n)/c(0) = 1 - \mathbf{d}_n^T\boldsymbol{C}_n^{-1}\mathbf{d}_n/c(0) = \frac{c(0) - \mathbf{d}_n^T\boldsymbol{C}_n^{-1}\mathbf{d}_n}{c(0)},
\end{aligned}
$$

where we used that the only nonzero eigenvalue of a rank 1 matrix is its trace and the cyclic commutativity of the trace operator. Putting things together:

$$
|\boldsymbol{C}_{n+1}| = c(0)|\boldsymbol{C}_n|\frac{c(0) - \mathbf{d}_n^T\boldsymbol{C}_n^{-1}\mathbf{d}_n}{c(0)} = |\boldsymbol{C}_n|(c(0) - \mathbf{d}_n^T\boldsymbol{C}_n^{-1}\mathbf{d}_n),
$$

that proves the statement. □

Remark 6. *If $|\boldsymbol{C}_n| = 0$ for some n, then $|\boldsymbol{C}_{n+1}| = |\boldsymbol{C}_{n+2}| = \cdots = 0$ too. The smallest index n for which this happens indicates that there is a linear relation between n consecutive X_js, but no linear relation between $n-1$ consecutive ones (by stationarity, this property is irrespective of the position of the consecutive random variables). This can happen only if some X_t linearly depends on $n-1$ preceding X_js. In this case $e_{n-1}^2 = 0$ and, of course $e_n^2 = e_{n+1}^2 = \cdots = 0$ too. In any case, $e_1^2 \ge e_2^2 \ge \dots$ is a decreasing (non-increasing) nonnegative sequence, and in view of Equation* (9),

$$
|\boldsymbol{C}_1| = c(0), \quad |\boldsymbol{C}_n| = c(0)e_1^2 \dots e_{n-1}^2, \quad n = 2, 3, \dots,
$$

so, provided $c(0) > 0$, $|\boldsymbol{C}_n| = 0$ holds if and only if $e_{n-1}^2 = 0$. Note that in this stationary case there is no sense of using generalized inverse if $|\boldsymbol{C}_n| = 0$, since then exact one-step ahead prediction with the $n-1$ long past can be done with zero error, and this property is manifested for longer past predictions too.

In the light of these, there are the following possibilities:

- $\boldsymbol{C}_k$ is positive definite up to $k \le h$, but $|\boldsymbol{C}_h| = 0$ for some positive integer h (and so, $|\boldsymbol{C}_k| = 0$ for $k > h$ too). Wold calls such a process singular of rank h. Then, by Remark 6,

  $$e_1^2 \ge e_2^2 \ge \cdots > e_{h-1}^2 = e_h^2 = \cdots = 0.$$

  So X_t can be exactly predicted based on its $(h-1)$length long past. This is caused by periodicities, for instance, in case of the Type (0) singular process of Lessons 5-6. In this case, $c(h)$ cannot tend to 0, otherwise all the $\boldsymbol{C}_h$s were positive definite, in view of Remark 1.

- $|\boldsymbol{C}_n| \ne 0$ for any n, and so, $e_n^2 > 0$ for every n, but still, $\lim_{n\to\infty} e_n^2 = 0$ in a decreasing (non-increasing) way. Wold calls such a process singular of infinite rank. This is caused by hidden periodicities, for instance, the Type (1) and Type (2) singular process of Lessons 5-6. (Then $\lim_{h\to\infty} c(h) = 0$, but not exponentially fast.)

- In the remaining (non-singular) case, $e_n^2 \to \sigma^2$ as $n \to \infty$ in a decreasing (non-increasing) way, where $0 < \sigma^2 < c(0)$. (In case of ARMA processes $\lim_{h\to\infty} c(h) = 0$ exponentially fast.)

Wold shows that the residual process $\eta_{t,n}$ (one-step ahead prediction error term of predicting $\mathbf{X}_t$ with its n-length long past) is stationary for any fixed n. After a passage to the limit, the process $\{\eta_{t,n}\}$ converges in probability to the residual process $\{\eta_t\}$ as $n \to \infty$. We cite the exact theorem (Theorem 6 of the original paper of Wold):

Theorem 1. *A residual process $\{\eta_t\}$ obtained from a non-singular stationary process $\{X_t\}$ is stationary and non-autocorrelated. Further, η_t is non-correlated with $X_{t-1}, X_{t-2}, \dots,$ while*

$$\mathrm{Corr}(X_t, \eta_t) = \frac{\mathrm{Cov}(X_t, \eta_t)}{\sqrt{c(0)}\sqrt{\mathrm{Var}(\eta_t)}} = \frac{\mathrm{Var}(\eta_t)}{\sqrt{c(0)}\sqrt{\mathrm{Var}(\eta_t)}} = \frac{\sqrt{\mathrm{Var}(\eta_t)}}{\sqrt{c(0)}} = \frac{\sigma}{\sqrt{c(0)}}.$$

Wold notes that the arguments used in the proof of this theorem also apply in the singular cases. As the residual variables η_t are here vanishing, their correlation properties will be indeterminate. Accordingly, these cases do not need further comment.

2 Multidimensional prediction

2.1 One-step ahead prediction based on finitely many past values

We have a d-dimensional real time series $\{\mathbf{X}_t\}$ with components $\mathbf{X}_t = [X_t^1, \dots, X_t^d]^T$. It is not necessarily stationary, we just assume the existence of the second moments and cross-moments. For simplicity, the state space is $\mathbb{R}^d$, but the time is discrete ($t \in \mathbb{Z}$).

Assume that $\mathbb{E}(\mathbf{X}_t) = \mathbf{0}$ ($t \in \mathbb{Z}$). Select a starting observation $\mathbf{X}_1$ and

$$H_n := \mathrm{Span}\{X_t^j : t = 1, \dots, n;\ j = 1, \dots, d\},$$

$\dim(H_n) = nd$. (Precisely, it should be denoted by $H_n(\mathbf{X})$, but as the process is fixed, it is briefly denoted by H_n. However, $\dim(H_n)$ is not the same as in the 1D situation.)

We want to linearly predict $\mathbf{X}_{n+1}$ based on past values $\mathbf{X}_1, \dots, \mathbf{X}_n$. Let $\hat{\mathbf{X}}_1 := 0$, and denote by $\hat{\mathbf{X}}_{n+1}$ the best one-step ahead linear prediction that minimizes the mean square error

$$\mathbb{E}(\mathbf{X}_{n+1} - \hat{\mathbf{X}}_{n+1})^2 = \|\mathbf{X}_{n+1} - \hat{\mathbf{X}}_{n+1}\|^2, \quad n = 1, 2, \dots$$

in the Hilbert-space setup (see the background material). Thus, $\hat{\mathbf{X}}_{n+1} = \mathrm{Proj}_{H_n} \mathbf{X}_{n+1}$, i.e. the projection of $\mathbf{X}_{n+1}$ onto the linear subspace H_n. In the Gaussian case, the solution is $\hat{\mathbf{X}}_{n+1} = \mathbb{E}(\mathbf{X}_{n+1} \,|\, \mathbf{X}_1, \dots, \mathbf{X}_n)$, which is the instance of *simultaneous linear regressions* for the components of $\mathbf{X}_{n+1}$ by predictors $\mathbf{X}_1, \dots, \mathbf{X}_n$. In the general case, we have to solve a system of linear equations that resembles (**??**)). Indeed, the projection is looked for in the form

$$\hat{\mathbf{X}}_{n+1} = \boldsymbol{A}_{n1}\mathbf{X}_n + \dots + \boldsymbol{A}_{nn}\mathbf{X}_1 \tag{10}$$

where $\boldsymbol{A}_{n1}, \dots \boldsymbol{A}_{nn}$ are $d \times d$ matrices. But $(\mathbf{X}_{n+1} - \hat{\mathbf{X}}_{n+1}) \perp \mathbf{X}_{n+1-k}$ for $k = 1, \dots, n$ in the sense that

$$\mathbb{E}[(\mathbf{X}_{n+1} - \hat{\mathbf{X}}_{n+1})\mathbf{X}_{n+1-k}^T] = \boldsymbol{O}_d, \quad k = 1, \dots n, \tag{11}$$

where $\boldsymbol{O}_d$ is the $d \times d$ zero matrix. Equations (10) and (11) together yield the following system of linear equations:

$$\sum_{j=1}^{n} \boldsymbol{A}_{nj} \mathrm{Cov}(\mathbf{X}_{n+1-j}, \mathbf{X}_{n+1-k}) = \mathrm{Cov}(\mathbf{X}_{n+1}, \mathbf{X}_{n+1-k}), \quad k = 1, \dots, n, \tag{12}$$

where Cov now denotes an $d \times d$ cross-covariance matrix. This is the extension of the Gauss normal equations for parallel linear predictions with d-dimensional target (see the background material).

When $\{\mathbf{X}_t\}$ is stationary, then Equation (12) simplifies to

$$\sum_{j=1}^{n} \boldsymbol{A}_j \boldsymbol{C}(k-j) = \boldsymbol{C}(k), \quad k = 1, \dots, n,$$

where $\boldsymbol{C}(k)$ is the kth order $d \times d$ autocovariance matrix. This provides a system of $d^2 n$ linear equations with the same number of unknowns that always has a solution. Further, the solution does not depend on the selection of the time of the starting observation $\mathbf{X}_1$, and no double indexing of the coefficient matrices is necessary. For the block matrix version see the background material. The coefficient matrix is just $\mathfrak{C}_n$, which is always positive semidefinite. If positive definite, we have a unique solution; otherwise, with block matrix techniques, reduced rank innovations are obtained.

There are recursions to solve this system (e.g. the Durbin–Levinson algorithm), which resembles the set of the first n Yule–Walker equations for a multidimensional VAR(n) processes.

Proposition 2. *If for some $n \geq 1$ the covariance matrix of $(\mathbf{X}_{n+1}^T, \dots, \mathbf{X}_1^T)^T$ is positive definite, then the matrix polynomial $\boldsymbol{\alpha}(z) = \boldsymbol{I} - \boldsymbol{A}_1 z - \dots - \boldsymbol{A}_n z^n$ is causal in the sense that the determinant $|\boldsymbol{\alpha}(z)| \neq 0$ for $z \leq 1$.*

2.2 Multidimensional innovations

Analogously to the 1D situation, $\mathbf{X}_t$ can again be expanded in terms of the now d-dimensional innovations, i.e. the prediction error terms

$$\boldsymbol{\eta}_{n+1} := \mathbf{X}_{n+1} - \hat{\mathbf{X}}_{n+1}.$$

It can be done step by step as follows. Assume that the $nd \times nd$ covariance matrix $\mathfrak{C}_n$ of the components of $\mathbf{X}_1, \dots, \mathbf{X}_n$ is positive definite for every $n \geq 1$. Let $\hat{\mathbf{X}}_1 := \mathbf{0}$, $\boldsymbol{\eta}_1 := \mathbf{X}_1$ and consider the unique orthogonal decomposition

$$\mathbf{X}_2 = \hat{\mathbf{X}}_2 + \boldsymbol{\eta}_2,$$

where $\hat{\mathbf{X}}_2 \in H_1$ and $\boldsymbol{\eta}_2 \perp H_1$, whenever $H_1 \subset H_2$ is a proper subspace (disregard the situation $H_1 = H_2$, when $\boldsymbol{\eta}_2 = \mathbf{0}$). Therefore, $\boldsymbol{\eta}_2 \in H_2$ and $\boldsymbol{\eta}_2 \perp \boldsymbol{\eta}_1$. With the same considerations,

$$\mathbf{X}_{j+1} = \hat{\mathbf{X}}_{j+1} + \boldsymbol{\eta}_{j+1}, \quad j = 2, \dots, n$$

with $\hat{\mathbf{X}}_{j+1} \in H_j$ and $\boldsymbol{\eta}_{j+1} \perp H_j$ if $H_j \subset H_{j+1}$ is a proper subspace. So $\boldsymbol{\eta}_{j+1} \in H_{j+1}$ and $\boldsymbol{\eta}_{j+1} \perp \boldsymbol{\eta}_j$.

In this way, we get the innovations $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n$ that trivially have $\mathbf{0}$ expectation and form an orthogonal system in the nd-dimensional H_n (their pairwise cross-covariance matrices are zeros). We consider the first n steps, i.e. the recursive equations

$$\mathbf{X}_k = \sum_{j=1}^{k-1} \boldsymbol{B}_{kj} \boldsymbol{\eta}_j + \boldsymbol{\eta}_k, \quad k = 1, 2, \dots, n \tag{13}$$

in the case when the observations $\mathbf{X}_1, \dots, \mathbf{X}_n$ are available.

If our process is stationary, the coefficient matrices are irrespective of the choice of the starting time. The $\boldsymbol{\eta}_j$s are not zeros if $H_n \subset H_{n+1}$ are proper subspaces, i.e. they are true innovations. However, it can be, that though they are not zeros, they span a lower than d-dimensional subspace, i.e. their covariance matrix $\boldsymbol{E}_j = \mathbb{E}\boldsymbol{\eta}_j\boldsymbol{\eta}_j'$ is not zero, but a positive semidefinite matrix of rank $r < d$. By stationarity, this rank is the same for all j, and it is equal to the (constant) rank of the spectral density matrix of the process. When we go to the future, then look back to the 'infinite' past, and obtain the multidimensional Wold decomposition and the forthcoming explanation at the end of this section).

Multiplying the equations in (13) by $\mathbf{X}_j^T$ from the right, and taking expectation, the solution for the matrices $\boldsymbol{B}_{kj}$ and $\boldsymbol{E}_j$ ($k = 1. \dots, n$; $j = 1, \dots, k-1$) can be obtained via the block Cholesky (LDL) decomposition:

$$\mathfrak{C}_n = \boldsymbol{L}_n \boldsymbol{D}_n \boldsymbol{L}_n^T, \tag{14}$$

where $\mathfrak{C}_n$ is $nd \times nd$ positive definite block Toeplitz matrix of general entry $\boldsymbol{C}(i-j)$, see (**??**). $\boldsymbol{D}_n$ is $nm \times nm$ block diagonal and contains the positive semidefinite prediction error matrices $\boldsymbol{E}_1, \dots, \boldsymbol{E}_n$ in its diagonal blocks, whereas $\boldsymbol{L}_n$ is $nd \times nd$ lower triangular with blocks $\boldsymbol{B}_{kj}$s below its diagonal blocks which are $d \times d$ identities, so $\boldsymbol{L}_n$ is non-singular. In matrix form,

$$\boldsymbol{L}_n = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} & \dots & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{B}_{21} & \boldsymbol{I} & \dots & \boldsymbol{O} & \boldsymbol{O} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{B}_{n1} & \boldsymbol{B}_{n2} & \dots & \boldsymbol{B}_{n,n-1} & \boldsymbol{I} \end{bmatrix}, \quad \boldsymbol{D}_n = \begin{bmatrix} \boldsymbol{E}_1 & \boldsymbol{O} & \dots & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{E}_2 & \dots & \boldsymbol{O} & \boldsymbol{O} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{O} & \boldsymbol{O} & \dots & \boldsymbol{O} & \boldsymbol{E}_n \end{bmatrix}. \tag{15}$$

To find the block Cholesky decomposition of (15), the following recursion is at our disposal: for $j = 1, \dots, n$

$$\boldsymbol{E}_j := \boldsymbol{C}(0) - \sum_{k=1}^{j-1} \boldsymbol{B}_{jk} \boldsymbol{E}_k \boldsymbol{B}_{jk}^T, \quad j = 1, \dots, n \tag{16}$$

and for $i = j, \dots, n$

$$\boldsymbol{B}_{ij} := \left(\boldsymbol{C}(i-j) - \sum_{k=1}^{j-1} \boldsymbol{B}_{ik} \boldsymbol{E}_k \boldsymbol{B}_{ik}^T \right) \boldsymbol{E}_j^+, \tag{17}$$

where we take the Moore–Penrose inverse if necessary.

Note that Equation (14) implies the following:

$$|\mathfrak{C}_n| = |\boldsymbol{D}_n| = \prod_{j=1}^{n} |\boldsymbol{E}_j|.$$

If the prediction is based on the infinite past, then with $n \to \infty$ this procedure (which is a nested one) extends to the multidimensional Wold decomposition. We can construct a causal TLF in this way. Actually, here $n = t$, and as observations arrive, $\mathbf{X}_n$ is predicted based on past values $\mathbf{X}_1, \dots, \mathbf{X}_{n-1}$, and so, $\boldsymbol{\eta}_n$ is in fact, $\boldsymbol{\eta}_{t,n}$. By stationarity, it has the same distribution for all t, especially for $t = n$. Also, if $n \to \infty$, the matrix $\boldsymbol{L}_n$ better and better approaches a Toeplitz one, and the matrices $\boldsymbol{E}_1, \dots, \boldsymbol{E}_n$ are closer and closer to $\boldsymbol{\Sigma}$, the covariance matrix of the innovation process $\{\boldsymbol{\eta}_n\}$. This is supported by Theorem 1, according to which, $\boldsymbol{\eta}_n \to \boldsymbol{\eta}$ in mean square:

$$\|\boldsymbol{E}_n - \boldsymbol{\Sigma}\| = \|\mathbb{E}(\boldsymbol{\eta}_n \boldsymbol{\eta}_n^T) - \mathbb{E}(\boldsymbol{\eta}\boldsymbol{\eta}^T)\| \to 0$$

as $n \to \infty$. Consequently, $\boldsymbol{B}_{nj} \to \boldsymbol{B}_j$ as $n \to \infty$ as it continuously depends on $\boldsymbol{E}_j$s in view of Equations (17).

Going further, when the $\boldsymbol{E}_j$s are of rank $r < d$, we can find a system $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n \in \mathbb{R}^r$ in the d-dimensional innovation subspaces that span the same subspace as $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n$. (Though, in this situation, the block Cholesky decomposition algorithm should be modified by taking generalized inverses.) If the rank is not exactly r (may be full), but the spectral density matrix has $r < d$ structural eigenvalues, then $\boldsymbol{\xi}_j \in \mathbb{R}^r$, $\mathbb{E}\boldsymbol{\xi}_j \boldsymbol{\xi}_j' = \boldsymbol{I}_r$ is the principal component factor of $\boldsymbol{\eta}_j$ obtained from the r-factor model

$$\boldsymbol{\eta}_j = \boldsymbol{A}_j \boldsymbol{\xi}_j + \boldsymbol{\varepsilon}_j,$$

where the columns of $d \times r$ matrix $\boldsymbol{A}_j$ are $\sqrt{\lambda_{j\ell}}\mathbf{u}_{j\ell}$ with the r largest eigenvalues and the corresponding eigenvectors of $\boldsymbol{E}_j$; the vector $\boldsymbol{\varepsilon}_j$ is the error comprised of both the idiosyncratic noise and the error term of the model, but it has a negligible L^2-norm. Note that $\boldsymbol{A}_j$ of the decomposition $\boldsymbol{E}_j = \boldsymbol{A}_j \boldsymbol{A}_j^T$ is far not unique, it can be post-multiplied with an $r \times r$ orthogonal matrix. With this,

$$\mathbf{X}_k \sim \sum_{j=1}^{k} \boldsymbol{B}_{kj} \boldsymbol{A}_j \boldsymbol{\xi}_j, \quad k = 1, 2, \dots, n \tag{18}$$

where $\boldsymbol{B}_{kk} = \boldsymbol{I}_k$. This approaches the following Wold decomposition of the d-dimensional process $\{\mathbf{X}_t\}$ with an r-dimensional ($r \le d$) innovation process $\{\boldsymbol{\xi}_t\}$.

$$\mathbf{X}_t = \sum_{j=0}^{\infty} \boldsymbol{B}_j \boldsymbol{\eta}_{t-j} = \sum_{j=0}^{\infty} \boldsymbol{B}_j \boldsymbol{A} \boldsymbol{\xi}_{t-j},$$

where $\lim_{k\to\infty} \boldsymbol{B}_{kj} = \boldsymbol{B}_j$ is $d \times d$ matrix; $\{\boldsymbol{\eta}_t\}$ is a d-dimensional white-noise sequence with covariance matrix $\boldsymbol{\Sigma}$ of rank r (actually, $\boldsymbol{\Sigma}$ it is the limit of the sequence $\boldsymbol{E}_n$), and $\{\boldsymbol{\xi}_t\}$ is an r-dimensional white-noise sequence with covariance matrix $\boldsymbol{I}_r$. Further, $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$ is the Gram-decomposition of the matrix $\boldsymbol{\Sigma}$ of rank r, where $\boldsymbol{A}$ is $d \times r$ (see the background material). Then the matrix sequence $\boldsymbol{B}_j \boldsymbol{A}$ plays the role of the $d \times r$ coefficient matrices in the multidimensional Wold decomposition of Section 4.4.

Note that here we use $nd \times nd$ block matrices, but the procedure, realized by Equations (16) and (17), iterates only with the $d \times d$ blocks of them, so the computational complexity of this algorithm is not significantly larger than that of the Kálmán's filtering of Section 4. However, in the next Section 3, we can decrease this computational complexity in the frequency domain.

3 Spectra of spectra

Let $\{\mathbf{X}_t\}$ be a d-dimensional, weakly stationary time series with real components and autocovariance matrices $\boldsymbol{C}(h)$, $h \in \mathbb{Z}$, $\boldsymbol{C}(-h) = \boldsymbol{C}^T(h)$. Consider the finite segment $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^d$ of it and the $nd \times nd$ covariance matrix $\mathfrak{C}_n$ of the compounded random vector $[\mathbf{X}_1^T, \dots, \mathbf{X}_n^T]^T \in \mathbb{R}^{nd}$, as introduced in Equation (**??**). As we discussed in Chapter 1, this is a symmetric, positive semidefinite block-Toeplitz matrix, the (i, j) block of which is $\boldsymbol{C}(j - i)$. The symmetry comes from the fact, that the (j, i) entry is $\boldsymbol{C}(i - j) = \boldsymbol{C}^T(j - i)$.

To characterize its eigenvalues, first we need the symmetric block circulant matrix $\mathfrak{C}_n^{(s)}$ that we consider now for odd n, say $n = 2k+1$. The (i,j) block of $\mathfrak{C}_n^{(s)}$ for $1 \le i \le j \le n$ is

$$\mathfrak{C}_n^{(s)}(\text{block}_i, \text{block}_j) = \begin{cases} \boldsymbol{C}(j-i) & j-i \le k \\ \boldsymbol{C}(n-(j-i)), & j-i > k. \end{cases}$$

For $i > j$, it is

$$\mathfrak{C}_n^{(s)}(\text{block}_i, \text{block}_j) = \begin{cases} \boldsymbol{C}^T(i-j) & i-j \le k \\ \boldsymbol{C}^T(n-(i-j)), & i-j > k. \end{cases}$$

In this way, $\mathfrak{C}_n^{(s)}$ is a symmetric block Toeplitz matrix again, and it is the same as $\mathfrak{C}_n$ within the blocks (i,j)s for which $|j-i| \le k$ holds. For example, if $n = 7$ and $k = 3$, then we have

$$\mathfrak{C}_7^{(s)} := \begin{bmatrix} \boldsymbol{C}(0) & \boldsymbol{C}(1) & \boldsymbol{C}(2) & \boldsymbol{C}(3) & \boldsymbol{C}(3) & \boldsymbol{C}(2) & \boldsymbol{C}(1) \\ \boldsymbol{C}^T(1) & \boldsymbol{C}(0) & \boldsymbol{C}(1) & \boldsymbol{C}(2) & \boldsymbol{C}(3) & \boldsymbol{C}(3) & \boldsymbol{C}(2) \\ \boldsymbol{C}^T(2) & \boldsymbol{C}^T(1) & \boldsymbol{C}(0) & \boldsymbol{C}(1) & \boldsymbol{C}(2) & \boldsymbol{C}(3) & \boldsymbol{C}(3) \\ \boldsymbol{C}^T(3) & \boldsymbol{C}^T(2) & \boldsymbol{C}^T(1) & \boldsymbol{C}(0) & \boldsymbol{C}(1) & \boldsymbol{C}(2) & \boldsymbol{C}(3) \\ \boldsymbol{C}^T(3) & \boldsymbol{C}^T(3) & \boldsymbol{C}^T(2) & \boldsymbol{C}^T(1) & \boldsymbol{C}(0) & \boldsymbol{C}(1) & \boldsymbol{C}(2) \\ \boldsymbol{C}^T(2) & \boldsymbol{C}^T(3) & \boldsymbol{C}^T(3) & \boldsymbol{C}^T(2) & \boldsymbol{C}^T(1) & \boldsymbol{C}(0) & \boldsymbol{C}(1) \\ \boldsymbol{C}^T(1) & \boldsymbol{C}^T(2) & \boldsymbol{C}^T(3) & \boldsymbol{C}^T(3) & \boldsymbol{C}^T(2) & \boldsymbol{C}^T(1) & \boldsymbol{C}(0) \end{bmatrix}.$$

In the 1D case, we simply have the $n \times n$ positive semidefinite matrix $\boldsymbol{C}_n$ of (**??**) and the symmetric circulant matrix $\boldsymbol{C}_n^{(s)}$ with the autocovariances $c(h)$s, $h \in \mathbb{Z}$. By Kronecker products (with permutation matrices) it is well known that the jth eigenvalue of $\boldsymbol{C}_n^{(s)}$ is

$$\sum_{h=-k}^{k} c(h)\rho_j^h = c(0) + 2\sum_{h=1}^{k} c(h)\cos(h\omega_j),$$

where $\rho_j = e^{i\omega_j}$ is the jth primitive (complex) nth root of 1 and $\omega_j = \frac{2\pi j}{n}$ is the jth Fourier frequency ($j = 0, 1, \dots, n-1$). Further, the eigenvector corresponding to the jth eigenvalue is $(1, \rho_j, \dots, \rho_j^{n-1})^T$; it has norm $\sqrt{n}$. After normalizing with $\frac{1}{\sqrt{n}}$, we get a complete orthonormal set of eigenvectors (of complex coordinates).

When $\boldsymbol{C}(h)$s are $d \times d$ matrices, by inflation techniques and applying Kronecker products, we use blocks instead of entries and the eigenvectors also follow a block structure. The eigenvalues and eigenvectors of a general symmetric block circulant matrix are characterized in the literature. We apply this result in our situation, when $n = 2k+1$ is odd (for even n similar results hold). Therefore, the spectrum of $\mathfrak{C}_n^{(s)}$ is the union of spectra of the matrices

$$\boldsymbol{M}_j = \boldsymbol{C}(0)+\sum_{h=1}^{k}[\boldsymbol{C}(h)\rho_j^h+\boldsymbol{C}^T(h)\rho_j^{-h}] = \boldsymbol{C}(0)+\sum_{h=1}^{k}[\boldsymbol{C}(h)e^{i\omega_j h}+\boldsymbol{C}^T(h)e^{-i\omega_j h}] \tag{19}$$

for $j = 0, 2, \dots, n-1$, whereas the eigenvectors are obtained by compounding the eigenvectors of these $d \times d$ matrices. So we need the spectral decomposition of the matrices

$$\boldsymbol{M}_0 = \boldsymbol{C}(0) + \sum_{h=1}^{k}[\boldsymbol{C}(h) + \boldsymbol{C}^T(h)]$$

and

$$\boldsymbol{M}_j = \boldsymbol{C}(0) + \sum_{h=1}^{k}[(\boldsymbol{C}(h) + \boldsymbol{C}^T(h))\cos(\omega_j h) + i(\boldsymbol{C}(h) - \boldsymbol{C}^T(h))\sin(\omega_j h)]$$

for $j = 0, 2, \dots, n-1$. Since $\boldsymbol{C}(h) + \boldsymbol{C}^T(h)$ is symmetric and $\boldsymbol{C}(h) - \boldsymbol{C}^T(h)$ is anti-symmetric with 0 diagonal, $\boldsymbol{M}_j$ is self-adjoint for each j and has real eigenvalues with corresponding orthonormal set of eigenvectors of possibly complex coordinates. Indeed, $\boldsymbol{M}_j$ may have complex entries if $j \neq 0$; actually, $\sum_{h=1}^{k}(\boldsymbol{C}(h) + \boldsymbol{C}^T(h))\cos(\omega_j h)$ is the real and $\sum_{h=1}^{k}(\boldsymbol{C}(h) - \boldsymbol{C}^T(h))\sin(\omega_j h)$ is the imaginary part of $\boldsymbol{M}_j$.

It is easy to see that $\boldsymbol{M}_{n-j} = \overline{\boldsymbol{M}_j}$ (entrywise conjugate), therefore, it has the same eigenvalues as $\boldsymbol{M}_j$, but the eigenvectors are the (componentwise) complex conjugates of the eigenvectors of $\boldsymbol{M}_j$. We also need the following form of this matrix:

$$\begin{aligned}\boldsymbol{M}_{n-j} &= \boldsymbol{C}(0) + \sum_{h=1}^{k}[(\boldsymbol{C}(h) + \boldsymbol{C}^T(h))\cos(\omega_j h) - i(\boldsymbol{C}(h) - \boldsymbol{C}^T(h))\sin(\omega_j h)] \\ &= \boldsymbol{C}(0) + \sum_{h=1}^{k}[\boldsymbol{C}(h)e^{-i\omega_j h} + \boldsymbol{C}^T(h)e^{i\omega_j h}], \quad j = 1, \dots, n-1.\end{aligned} \tag{20}$$

Summarizing, for odd $n = 2k+1$, the nd eigenvalues of $\mathfrak{C}_n^{(s)}$ are obtained as the union of the eigenvalues of $\boldsymbol{M}_0$ and those of $\boldsymbol{M}_j$ $(j = 1, \dots, k)$ duplicated. Note that for even n similar arguments hold with the difference that there the spectrum of $\mathfrak{C}_n^{(s)}$ is the union of the eigenvalues of $\boldsymbol{M}_0$ and $\boldsymbol{M}_{n-1}$, whereas the eigenvalues of $\boldsymbol{M}_1, \dots, \boldsymbol{M}_{\frac{n}{2}-1}$ are duplicated.

The eigenvectors of $\mathfrak{C}_n^{(s)}$ are obtainable by compounding the d orthonormal eigenvectors of the $d \times d$ self-adjoint matrices $\boldsymbol{M}_0, \boldsymbol{M}_1, \dots, \boldsymbol{M}_{n-1}$ as follows. For $j = 1, \dots, k$: if $\mathbf{v}$ is an eigenvector of $\boldsymbol{M}_j$ with eigenvalue λ, then the compound vector

$$\mathbf{w} = (\mathbf{v}^T, \rho_j \mathbf{v}^T, \rho_j^2 \mathbf{v}^T, \dots, \rho_j^{n-1} \mathbf{v}^T)^T \in \mathbb{C}^{nd} \tag{21}$$

is an eigenvector of $\mathfrak{C}_n^{(s)}$ with the same eigenvalue λ. Further, if

$$\mathbf{z} = (\mathbf{t}^T, \rho_\ell \mathbf{t}^T, \rho_\ell^2 \mathbf{t}^T, \dots, \rho_\ell^{n-1} \mathbf{t}^T)^T \in \mathbb{C}^{nd}$$

is another eigenvector of $\mathfrak{C}_n^{(s)}$ compounded from an eigenvector $\mathbf{t}$ of another $\boldsymbol{M}_\ell$ $(\ell \neq j)$, then $\mathbf{w}$ and $\mathbf{z}$ are orthogonal, irrespective whether $\boldsymbol{M}_\ell$ has the same eigenvalue λ as $\boldsymbol{M}_j$ or not. Similar construction holds starting with the eigenvectors of $\boldsymbol{M}_0$.

Here for each $j = 0, 1, \dots, n-1$, there are d pairwise orthogonal eigenvectors (potential $\mathbf{v}$s) of $\boldsymbol{M}_j$, and the so obtained $\mathbf{w}$s are also pairwise orthogonal. Assume that the eigenvectors of $\boldsymbol{M}_j$ are enumerated in non-increasing order of its eigenvalues, and the inflated $\mathbf{w}$s also follow this ordering, for $j = 0, 1, \dots, n-1$.

As we saw, if $\mathbf{v}$ is an eigenvector of $\boldsymbol{M}_j$ with real eigenvalue λ, then $\overline{\mathbf{v}}$ is the corresponding eigenvector of $\boldsymbol{M}_{n-j}$ with the same eigenvalue λ; further, the compounded $\mathbf{w}$ and $\overline{\mathbf{w}} \in \mathbb{C}^{nd}$ are orthogonal eigenvectors of $\mathfrak{C}_n^{(s)}$ corresponding to the eigenvalue λ with multiplicity (at least) two; $\mathbf{w}$ and $\overline{\mathbf{w}}$ have the same norm. From them, corresponding to this double eigenvalue λ, the new orthogonal pair of eigenvectors

$$\frac{\mathbf{w} + \overline{\mathbf{w}}}{2} \quad \text{and} \quad i\frac{\mathbf{w} - \overline{\mathbf{w}}}{2} \tag{22}$$

is constructed, but they, in this order, occupy the original positions of $\mathbf{w}$ and $\overline{\mathbf{w}}$. Note that it is necessary to have an orthogonal system of eigenvectors with real coordinates whenever the underlying time series is real, and so, $\mathfrak{C}_n^{(s)}$ is a real symmetric matrix. We do not go in details, neither discuss defective cases.

After normalization, denote by $\mathbf{u}_1, \dots, \mathbf{u}_{nd}$ the so obtained orthonormal set of eigenvectors (of real coordinates) of $\mathfrak{C}_n^{(s)}$ (in the above ordering) and by $\boldsymbol{U} = (\mathbf{u}_1, \dots, \mathbf{u}_{nd})$ the $nd \times nd$ orthogonal matrix containing them columnwise; further, let

$$\mathfrak{C}_n^{(s)} = \boldsymbol{U} \boldsymbol{\Lambda}^{(s)} \boldsymbol{U}^T \tag{23}$$

be the corresponding spectral decomposition. After this preparation, we are able to prove the following theorem.

Theorem 2. *Let $\{\mathbf{X}_t\}$ be d-dimensional weakly stationary time series of real components. Denoting by $\boldsymbol{C}(h) = [c_{ij}(h)]$ the $d \times d$ autocovariance matrices $(\boldsymbol{C}(-h) = \boldsymbol{C}^T(h),\ h \in \mathbb{Z})$ in the time domain, assume that their entries are absolutely summable, i.e., $\sum_{h=0}^{\infty} |c_{pq}(h)| < \infty$ for $p, q = 1, \dots, d$. Then, the self-adjoint, positive semidefinite spectral density matrix $\boldsymbol{f}(\omega)$ exists in the frequency domain, and it is defined by*

$$\boldsymbol{f}(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \boldsymbol{C}(h) e^{-ih\omega}, \quad \omega \in [0, 2\pi].$$

For odd $n = 2k+1$, consider $\mathbf{X}_1, \dots \mathbf{X}_n$ with the block Toeplitz matrix $\mathfrak{C}_n$; further, the Fourier frequencies $\omega_j = \frac{2\pi j}{n}$ for $j = 0, \dots, n-1$. Let $\boldsymbol{D}_n$ be the $dn \times dn$ diagonal matrix that contains the spectra of the matrices $\boldsymbol{f}(0), \boldsymbol{f}(\omega_1), \boldsymbol{f}(\omega_2), \dots, \boldsymbol{f}(\omega_k), \boldsymbol{f}(\omega_k), \dots, \boldsymbol{f}(\omega_2), \boldsymbol{f}(\omega_1)$ in its main diagonal, i.e.,

$$\boldsymbol{D}_n = \operatorname{diag}(\operatorname{spec} \boldsymbol{f}(0), \operatorname{spec} \boldsymbol{f}(\omega_1), \dots, \operatorname{spec} \boldsymbol{f}(\omega_k), \operatorname{spec} \boldsymbol{f}(\omega_k), \dots, \operatorname{spec} \boldsymbol{f}(\omega_1)).$$

Here spec *contains the eigenvalues of the affected matrix in non-increasing order if not otherwise stated. (The duplication is due to the fact that $\boldsymbol{f}(\omega_j) = \boldsymbol{f}(\omega_{n-j})$, $j = 1, \dots, k$, for real time series). Then, with the spectral decomposition (23),*

$$\boldsymbol{U}^T \mathfrak{C}_n \boldsymbol{U} - 2\pi \boldsymbol{D}_n \to \boldsymbol{O}, \quad n \to \infty,$$

i.e., the entries of the matrix $\boldsymbol{U}^T \mathfrak{C}_n \boldsymbol{U} - 2\pi \boldsymbol{D}_n$ tend to 0 uniformly as $n \to \infty$.

Proof. We saw that $\boldsymbol{U}^T \mathfrak{C}_n^{(s)} \boldsymbol{U} = \boldsymbol{\Lambda}^{(s)}$. Recall that the eigenvalues in the diagonal of $\boldsymbol{\Lambda}^{(s)}$ comprise the union of spectra of the matrices $\boldsymbol{M}_0$ and those of $\boldsymbol{M}_1, \dots, \boldsymbol{M}_{n-1}$, which are the same as the eigenvalues of $\boldsymbol{M}_0$ and those of $\boldsymbol{M}_{n-1}, \dots, \boldsymbol{M}_{n-k}$ of (20), duplicated. But these matrices are finite sub-sums (for $|h| \le k$) of the infinite summations

$$2\pi \boldsymbol{f}(\omega_j) = \sum_{h=-\infty}^{\infty} \boldsymbol{C}(h) e^{-ih\omega} = \boldsymbol{C}(0) + \sum_{h=1}^{\infty} [\boldsymbol{C}(h) e^{-i\omega_j h} + \boldsymbol{C}^T(h) e^{i\omega_j h}],$$

so (by the continuity of the spectra), the pairwise distances between the eigenvalues of $\boldsymbol{M}_j$ and the corresponding eigenvalues of $2\pi \boldsymbol{f}(\omega_j)$ (both in non-increasing order) tend to 0 as $n \to \infty$, for $j = 0, 1, \dots, k$. Here we used the absolute summability of the entries of $\boldsymbol{C}(h)$s, which fact implies that the diagonal entries of the diagonal matrix $\boldsymbol{\Lambda}^{(s)} - 2\pi \boldsymbol{D}_n$ are bounded in absolute value by

$$\max_{p,q \in \{1,\dots,d\}} \sum_{|h|>k} |c_{pq}(h)| \to 0, \quad n = 2k+1 \to \infty.$$

So the matrix $\boldsymbol{\Lambda}^{(s)} - 2\pi \boldsymbol{D}_n$ tends to the zero matrix entrywise as $n \to \infty$. Therefore, it remains to show that the entries of $\mathbf{U}^T \mathfrak{C}_n \mathbf{U} - \mathbf{U}^T \mathfrak{C}_n^{(s)} \mathbf{U}$ tend to 0 uniformly as $n \to \infty$.

Before doing this, some facts should be clarified.

- The pth row sum of $\boldsymbol{M}_j$ is bounded by

$$\sum_{q=1}^{d} |c_{pq}(0)| + \sum_{q=1}^{d} \sum_{h=1}^{k} |c_{pq}(h)| + \sum_{q=1}^{d} \sum_{h=1}^{k} |c_{qp}(h)| \le d c_{pp}(0) + 2dL,$$

  for $p \in \{1, \dots, d\}$ with $L = \max_{p,q \in \{1,\dots,d\}} \sum_{h=1}^{\infty} |c_{pq}(h)| > 0$, independently of n, because of the absolute summability of the entries of $\boldsymbol{C}(h)$. This is true for any $j \in \{0, 1, \dots, n-1\}$. For simplicity, consider (any) one of the $\boldsymbol{M}_j$s, and denote it by $\boldsymbol{M} = [m_{pq}]_{p,q=1}^{d}$. Then

$$\|\boldsymbol{M}\|_\infty = \max_{p \in \{1,\dots,d\}} \sum_{q=1}^{d} |m_{pq}| \le d \max_{p \in \{1,\dots,d\}} c_{pp}(0) + 2dL = K.$$

  As the spectral radius of $\boldsymbol{M}$ is at most $\|\boldsymbol{M}\|_\infty$, any eigenvalue λ of $\boldsymbol{M}$ is bounded in absolute value by K (independenty of n).

- Let $\mathbf{v}$ be an eigenvector of $\boldsymbol{M}_j$ with eigenvalue λ, we can assume that $\|\mathbf{v}\| = \sqrt{\mathbf{v}^* \mathbf{v}} = 1$. Then the vector $\mathbf{w}$ in Equation (21) is an eigenvector of $\mathfrak{C}_n^{(s)}$. Since

$$\mathbf{w}^* \mathbf{w} = \mathbf{v}^* \mathbf{v} (1 + \rho_j \rho_j^{-1} + \rho_j^2 \rho_j^{-2} + \dots + \rho_j^{n-1} \rho_j^{-(n-1)}) = n,$$

the (complex) vector $\frac{1}{\sqrt{n}}\mathbf{w}$ will have unit norm. Further, by transformation (22), the coordinates of any (real) unit-norm eigenvector $\mathbf{u}$ are bounded by $\sqrt{\frac{2}{n}}$ in absolute value.

Now we are ready to show that

$$|\mathbf{u}_i^T \mathfrak{C}_n^{(s)} \mathbf{u}_j - \mathbf{u}_i^T \mathfrak{C}_n \mathbf{u}_j| \to 0, \quad n \to \infty$$

uniformly in $i, j \in \{1, \dots, nd\}$. Recall that in the $nd \times nd$ matrices $\mathfrak{C}_n^{(s)}$ and $\mathfrak{C}_n$ the (m, ℓ) blocks are the same if $|m - \ell| \le k$. Denote by $\mathbf{u}_{i,m}$ and $\mathbf{u}_{j,\ell}$ the mth and ℓth blocks of the unit-norm eigenvectors $\mathbf{u}_i$ and $\mathbf{u}_j$, respectively. Recall (see their description preceding the theorem) that they both were compounded from n vectors of length d, and their coordinates are bounded by $\sqrt{\frac{2}{n}}$ in absolute value. Then

$$\begin{aligned}
&|\mathbf{u}_i^T(\mathfrak{C}_n^{(s)} - \mathfrak{C}_n)\mathbf{u}_j| \\
&= \left| \sum_{m=1}^{k} \sum_{\ell=1}^{m} [\mathbf{u}_{i,\ell}^T (\boldsymbol{C}(m) - \boldsymbol{C}(n-m))\mathbf{u}_{j,n-m+\ell} + \mathbf{u}_{i,n-m+\ell}^T (\boldsymbol{C}(n-m) - \boldsymbol{C}(m))\mathbf{u}_{j,\ell}] \right| \\
&\le \left| \sqrt{\frac{2}{n}} \sum_{m=1}^{k} \mathbf{1}_d^T (\boldsymbol{C}(m) - \boldsymbol{C}(n-m)) \sum_{\ell=1}^{m} \mathbf{u}_{j,n-m+\ell} \right| \\
&+ \left| \sqrt{\frac{2}{n}} \sum_{m=1}^{k} \sum_{\ell=1}^{m} \mathbf{u}_{i,n-m+\ell}^T (\boldsymbol{C}(m) - \boldsymbol{C}(n-m))\mathbf{1}_d \right| \\
&\le 2\sqrt{\frac{2}{n}}\sqrt{\frac{2}{n}} \left| \sum_{m=1}^{k} m \mathbf{1}_d^T (\boldsymbol{C}(m) - \boldsymbol{C}(n-m))\mathbf{1}_d \right| \\
&\le \frac{4}{n} \left(\sum_{m=1}^{k} m \sum_{p=1}^{d} \sum_{q=1}^{d} |c_{pq}(m)| + \sum_{m=1}^{k} m \sum_{p=1}^{d} \sum_{q=1}^{d} |c_{pq}(n-m)| \right) \\
&\le 4d^2 \left(\max_{p,q \in \{1,\dots,d\}} \sum_{m=1}^{k} \frac{m}{n} |c_{pq}(m)| + \max_{p,q \in \{1,\dots,d\}} \sum_{m=1}^{k} \frac{m}{n} |c_{pq}(n-m)| \right) \\
&\le 4d^2 \left(\max_{p,q \in \{1,\dots,d\}} \sum_{m=1}^{k} \frac{m}{n} |c_{pq}(m)| + \max_{p,q \in \{1,\dots,d\}} \sum_{m=n-k}^{n-1} \frac{k}{n} |c_{pq}(m)| \right),
\end{aligned}$$

where $\mathbf{1}_d \in \mathbb{R}^d$ is the vector of all 1 coordinates and so, the quadratic form $\mathbf{1}_d^T(\boldsymbol{C}(m) - \boldsymbol{C}(n-m))\mathbf{1}_d$ is the sum of the entries of $\boldsymbol{C}(m) - \boldsymbol{C}(n-m)$.

In the last line, the second term converges to 0, since it is bounded by $\sum_{m=k}^{\infty} |c_{pq}(m)|$ (indeed, $\sum_{m=n-k}^{n-1} \frac{k}{n} |c_{pq}(m)| \le \sum_{m=k}^{\infty} |c_{pq}(m)|$ as $k < n-k$), and together with n, k tends to ∞ too; further, it holds uniformly for all $p, q \in \{1, \dots, d\}$. The first term for every p, q pair also tends to 0 as $n \to \infty$ by the discrete version of the dominated convergence theorem (for series), see the forthcoming Lemma 1. Indeed, the summand is dominated by $|c_{pq}(m)|$ and $\sum_{m=1}^{\infty} |c_{pq}(m)| < \infty$; further, $\frac{m}{n} |c_{pq}(m)| \to 0$ as $n \to \infty$, for any fixed m. Consequently, $\sum_{m=1}^{\infty} \frac{m}{n} |c_{pq}(m)|$ tends to 0, and so does $\sum_{m=1}^{k} \frac{m}{n} |c_{pq}(m)|$ as $n \to \infty$. It holds uniformly for all p, q, and also for all i, j, so the proof is complete. □

Lemma 1 (Dominated convergence theorem for sums, discrete version)**.** *Consider $\sum_{m=1}^{\infty} f_n(m)$ and Assume that $|f_n(m)| \le g(m)$ with $\sum_{m=1}^{\infty} g(m) < \infty$. If $\lim_{n\to\infty} f_n(m) = f(m)$ exists $\forall m \in \mathbb{N}$, then*

$$\lim_{n\to\infty} \sum_{m=1}^{\infty} f_n(m) = \sum_{m=1}^{\infty} f(m).$$

Some consequences of the above theorem are to be discussed.

3.1 Bounds for the eigenvalues of $\mathfrak{C}_n$

Proposition 3. *The above theorem implies the following. Assume that for the spectra of the spectral densities $\boldsymbol{f}$ of the d-dimensional weakly stationary process $\{\mathbf{X}_t\}$ of real coordinates the following hold:*

$$m := \inf_{\omega\in[0,2\pi], q\in\{1,\dots,d\}} \lambda_q(\boldsymbol{f}(\omega)) > 0,$$
$$M := \sup_{\omega\in[0,2\pi], q\in\{1,\dots,d\}} \lambda_q(\boldsymbol{f}(\omega)) < \infty.$$

(Note that under the conditions of Theorem 2, $\boldsymbol{f}(\omega) > 0$ and it is continuous almost everywhere on $[0, 2\pi]$, so the above conditions are readily satisfied.)

Then for the eigenvalues $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_{nd}$ of the block Toeplitz matrix $\mathfrak{C}_n$ the following holds:

$$2\pi m \le \lambda_1 \le \lambda_{nd} \le 2\pi M.$$

Proof. Let λ be an arbitrary eigenvalue of $\mathfrak{C}_n$ with a corresponding eigenvector $\mathbf{x} \in \mathbb{C}^{nd}$, $\mathbf{x}^* = [\mathbf{x}_1^*, \dots, \mathbf{x}_n^*]$, $\mathbf{x}_j \in \mathbb{C}^d$: $\mathfrak{C}_n \mathbf{x} = \lambda \mathbf{x}$. Take the spectral decomposition of the spectral density matrix $\boldsymbol{f}$:

$$\boldsymbol{f}(\omega) = \sum_{\ell=1}^{d} \lambda_\ell(\boldsymbol{f}(\omega)) \cdot \mathbf{u}_\ell(\boldsymbol{f}(\omega)) \cdot \mathbf{u}_\ell^*(\boldsymbol{f}(\omega)).$$

Then we can write that

$$\begin{aligned}
\lambda|\mathbf{x}|^2 &= \lambda\mathbf{x}^*\mathbf{x} = \mathbf{x}^*\mathfrak{C}_n\mathbf{x} \\
&= \mathbf{x}^* \cdot \int_{-\pi}^{\pi} \left[e^{-i(j-k)\omega}\boldsymbol{f}(\omega)\right]_{j,k=1}^{n} d\omega \cdot \mathbf{x} \\
&= \int_{-\pi}^{\pi} \sum_{j,k=1}^{n} e^{-i(j-k)\omega}\mathbf{x}_j^*\boldsymbol{f}(\omega)\mathbf{x}_k d\,\omega \\
&= \int_{-\pi}^{\pi} \sum_{j,k=1}^{n} e^{-i(j-k)\omega} \sum_{\ell=1}^{d} \lambda_\ell(\boldsymbol{f}(\omega)) \cdot \mathbf{x}_j^* \cdot \mathbf{u}_\ell(\boldsymbol{f}(\omega)) \cdot \mathbf{u}_\ell^*(\boldsymbol{f}(\omega)) \cdot \mathbf{x}_k \, d\omega \\
&= \int_{-\pi}^{\pi} \sum_{\ell=1}^{d} \lambda_\ell(\boldsymbol{f}(\omega)) \sum_{j,k=1}^{n} e^{-ij\omega}\mathbf{x}_j^* \cdot \mathbf{u}_\ell(\boldsymbol{f}(\omega)) \cdot \mathbf{u}_\ell^*(\boldsymbol{f}(\omega)) \cdot \mathbf{x}_k \cdot e^{ik\omega} \, d\omega \\
&= \int_{-\pi}^{\pi} \sum_{\ell=1}^{d} \lambda_\ell(\boldsymbol{f}(\omega)) \left| \sum_{j=1}^{n} e^{-ij\omega} \cdot \mathbf{x}_j^* \cdot \mathbf{u}_\ell(\boldsymbol{f}(\omega)) \right|^2 d\omega \\
&\le M \sum_{j,k=1}^{n} \mathbf{x}_j^* \cdot \int_{-\pi}^{\pi} e^{-i(j-k)\omega} \sum_{\ell=1}^{d} \mathbf{u}_\ell(\boldsymbol{f}(\omega)) \cdot \mathbf{u}_\ell^*(\boldsymbol{f}(\omega)) \, d\omega \cdot \mathbf{x}_k \\
&= 2\pi M \sum_{j=1}^{n} \mathbf{x}_j^*\mathbf{x}_j = 2\pi M|\mathbf{x}|^2.
\end{aligned}$$

This proves that $\lambda \le 2\pi M$ for any eigenvalue of $\mathfrak{C}_n$. The proof of the fact that $\lambda \ge 2\pi m$ is similar. □

3.2 Principal Component transformation as discrete Fourier transformation

The complex Principal Component (PC) transform of the collection of random vectors $\mathbf{X} = (\mathbf{X}_1^T, \dots \mathbf{X}_n^T)^T$ of real coordinates is the random vector $\mathbf{Z} = (\mathbf{Z}_1^T, \dots, \mathbf{Z}_n^T)^T$ of complex coordinates obtained by

$$\mathbf{Z} = \boldsymbol{W}^*\mathbf{X}.$$

Here, analogously to (**??**), $\mathfrak{C}_n^{(s)}$ also has the spectral decomposition

$$\mathfrak{C}_n^{(s)} = \boldsymbol{W}\boldsymbol{\Lambda}^{(s)}\boldsymbol{W}^*, \tag{24}$$

where the unitary matrix $\boldsymbol{W} = (\mathbf{w}_1, \dots, \mathbf{w}_{nd})$, contains a complete orthonormal set of eigenvectors of $\mathfrak{C}_n^{(s)}$, columnwise. They usually have complex coordinates.

To relate the PC transformation to a discrete Fourier transformation, we also make PC transformations within the blocks. For this purpose we use the eigenvectors in the columns of $\boldsymbol{W}$ (of complex coordinates) in the ordering described in the preparation of Theorem 2. We utilize their block structure and also assume that they are already normalized to have a complete orthonormal system in $\mathbb{C}^{nd}$.

By Theorem 2, $\mathbb{E}\mathbf{Z}\mathbf{Z}^* \sim 2\pi\boldsymbol{D}_n$, so the coordinates of $\mathbf{Z}$ are asymptotically uncorrelated, for 'large' n. Instead, we consider the blocks $\mathbf{Z}_j$s of it, and perform a 'partial principal component transformation' (in d-dimension) of them. Let $\mathbf{w}_{1j}, \dots, \mathbf{w}_{dj}$ be the columns of $\mathbf{W}$ corresponding to the coordinates of $\mathbf{Z}_j$. In view of (**??**), $\mathbf{Z}_j$ can be written as

$$\mathbf{Z}_j = \frac{1}{\sqrt{n}}(\boldsymbol{V}_j^* \otimes \mathbf{r}^*)\mathbf{X},$$

where $\mathbf{r}^* = (1, \rho_j^{-1}, \rho_j^{-2}, \dots, \rho_j^{-(n-1)})$ and $\boldsymbol{V}_j$ is the $d \times d$ unitary matrix in the spectral decomposition $\boldsymbol{M}_j = \boldsymbol{V}_j\boldsymbol{\Lambda}_j\boldsymbol{V}_j^*$. Because of $\mathbb{E}\mathbf{Z}_j\mathbf{Z}_j^* = \boldsymbol{\Lambda}_j$ (apparently from the proof of Theorem 2), we have that

$$\mathbb{E}(\boldsymbol{V}_j\mathbf{Z}_j)(\boldsymbol{V}_j\mathbf{Z}_j)^* = \boldsymbol{V}_j\boldsymbol{\Lambda}_j\boldsymbol{V}_j^* = \boldsymbol{M}_j.$$

At the same time,

$$\boldsymbol{V}_j\mathbf{Z}_j = \frac{1}{\sqrt{n}}\boldsymbol{V}_j(\boldsymbol{V}_j^*\otimes\mathbf{r}^*)\mathbf{X} = \frac{1}{\sqrt{n}}(\boldsymbol{I}_d\otimes\mathbf{r}^*)\mathbf{X} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\mathbf{X}_t e^{-it\omega_j}, \quad j = 1, \dots, n.$$

This is the discrete Fourier transform of $\mathbf{X}_1, \dots, \mathbf{X}_n$. It is in accord with the existence of the orthogonal increment process $\{\mathbf{Z}_\omega\}$ (see Lessons 1-2) of which $\boldsymbol{V}_j\mathbf{Z}_j \sim \mathbf{Z}_{\omega_j}$ is the discrete analogue. Also, $\mathbf{Z}_1, \dots \mathbf{Z}_n$ are asymptotically pairwise orthogonal akin to $\boldsymbol{V}_1\mathbf{Z}_1, \dots, \boldsymbol{V}_n\mathbf{Z}_n$. Further,

$$\mathbb{E}(\boldsymbol{V}_j\mathbf{Z}_j)(\boldsymbol{V}_j\mathbf{Z}_j)^* \sim 2\pi\boldsymbol{f}(\omega_j),$$

and it is in accord with the fact that

$$\mathbb{E}\mathbf{Z}_j\mathbf{Z}_j^* \sim 2\pi\,\mathrm{diag\,spec}\;\boldsymbol{f}(\omega_j),$$

for $j = 0, 1, \dots, n-1$ when n is 'large'.

4 Kálmán's filtering

Given a linear dynamical system, with state equations and specified matrices, R. E. Kálmán gave a recursive algorithm, how to find prediction for the state variable $\mathbf{X}_t$ in the possession of newer and newer observations for the observable variable $\mathbf{Y}_t$. Starting at time 0, estimates $\widehat{\mathbf{X}}_{t+1|t}$ are found, while observing $\mathbf{Y}_t$, $t = 1, 2, \dots$. The point is that we only use the last observation $\mathbf{Y}_t$ and the preceding estimate $\widehat{\mathbf{X}}_{t|t-1}$. During the recursion, we use the linearity of the state equations and the predictions, for which either normality is assumed, or we confine ourselves to the second moments of the underlying distributions, see the background material.

This so-called filtering technique is widely used in the engineering practice, when we can 'get rid' of the noise, and also possess an algorithm to find the innovations of the observed $\{\mathbf{Y}_t\}$ process (we need not perform the block Cholesky decomposition of Section 2.2, but get the innovations recursively). The problem is that we merely have the output of a linear system that is burdened with noise, and usually not invertible; e.g., in case of telecommunication systems, when sensors can sense only noisy signals. It is not by accident that the research of R. E. Kálmán and R. S. Bucy followed the era of the information theoretical breakthroughs, e.g., the intensive use of the Shannon entropy.

Here we follow the discussion of R. E. Kálmán's original paper, where stationarity is not assumed, but the random vectors are Gaussian. (Sometimes we use simpler notation in accordance with the one used in the previous sections of this chapter.) Here the linear dynamical system is

$$\begin{aligned} \mathbf{X}_{t+1} &= \boldsymbol{A}_t \mathbf{X}_t + \mathbf{U}_t \\ \mathbf{Y}_t &= \boldsymbol{C}_t \mathbf{X}_t, \end{aligned} \tag{25}$$

where $\boldsymbol{A}_t$ and $\boldsymbol{C}_t$ are specified matrices; $\boldsymbol{A}_t$ is an $n \times n$ matrix, called *phase transition matrix*, and $\boldsymbol{C}_t$ is $p \times n$; further, $\mathbf{U}_t$ is an orthogonal noise process with $\mathbb{E}\mathbf{U}_t\mathbf{U}_s^T = \delta_{st}\boldsymbol{Q}_{\mathbf{U}}(t)$ and $\mathbb{E}\mathbf{X}_s^T\mathbf{U}_t = \mathbf{0}$ for $s \le t$. All the expectations are zeros, and all the random vectors have real components. Sometimes $\mathbf{U}_t$ is called *random excitation*, $\mathbf{X}_t$ is the n-dimensional hidden *state variable*, while $\mathbf{Y}_t$ is the p-dimensional *observable variable*. In the paper of Kálmán, $p \le n$ is assumed, but it is not a restriction. Even if $p = n$, the matrix $\boldsymbol{C}_t$ is not invertible, otherwise the process $\mathbf{X}_t$ is trivially observable, unless a noise term is added to $\boldsymbol{C}_t\mathbf{X}_t$ in the second equation (we will touch upon this possibility at the end of this section).

So the problem is the following: starting the observations at time 0, given $\mathbf{Y}_0, \dots, \mathbf{Y}_{t-1}$, we want to estimate $\mathbf{X}$ component-wise, with minimum mean square error. More precisely, if $\widehat{\mathbf{X}}$ denotes this estimate, then $\widehat{\mathbf{X}} = \mathrm{Proj}_{H_{t-1}(\mathbf{Y})}\mathbf{X}$, where $H_{t-1}(\mathbf{Y}) = \mathrm{Span}\,(\mathbf{Y}_0, \dots, \mathbf{Y}_{t-1})$ consists of the linear combinations of all the components of $\mathbf{Y}_0, \dots, \mathbf{Y}_{t-1}$ (with the notation of the background material, but here the indexing starts at 0). If we minimize the mean square error, the minimizer is the conditional expectation $\mathbb{E}(\mathbf{X} \,|\, \mathbf{Y}_0, \dots, \mathbf{Y}_{t-1})$, which is the linear function of the coordinates of the random vectors in the condition, whenever the underlying distribution is Gaussian.

If $\mathbf{X} = \mathbf{X}_t$, this is the prediction problem and we denote the optimal *one-step ahead prediction* of $\mathbf{X}_t$ by $\widehat{\mathbf{X}}_{t|t-1}$. In a similar vein, $\widehat{\mathbf{X}}_{t|t}$ solves the *filtration* problem, when we project onto $H_t(\mathbf{Y})$ for the prediction; finally, $\widehat{\mathbf{X}}_{t|t+h}$ solves the *smoothing* problem, when we project onto $H_{t+h}(\mathbf{Y})$ with $h > 0$ integer. The first one-step ahead prediction problem can be generalized to the h-step ahead prediction of $\widehat{\mathbf{X}}_{t+h|t-1}$, $h > 0$ integer (not the same as the smoothing problem). The first problem is sometimes called extrapolation, whereas the second two interpolation, respectively. Note that the problem itself is originated in the Wiener–Hopf problem.

As for the one-step ahead prediction problem, if $\mathbf{Y}_0, \dots, \mathbf{Y}_{t-1}$ are observed, i.e. $H_{t-1}(\mathbf{Y})$ is known, then the newly observed (measured) $\mathbf{Y}_t$ can be orthogonally decomposed as

$$\mathbf{Y}_t = \mathrm{Proj}_{H_{t-1}(\mathbf{Y})}\mathbf{Y}_t + \widetilde{\mathbf{Y}}_{t|t-1} = \overline{\mathbf{Y}}_{t|t-1} + \widetilde{\mathbf{Y}}_{t|t-1}, \tag{26}$$

where the orthogonal component $\widetilde{\mathbf{Y}}_{t|t-1} \in I_t(\mathbf{Y})$, and $I_t(\mathbf{Y})$ is the so-called *innovation subspace.* (Actually, the components of $\widetilde{\mathbf{Y}}_{t|t-1}$ span $I_t(\mathbf{Y})$). We shall make intensive use of this innovation. Assume that $I_t(\mathbf{Y})$ is not the sole $\mathbf{0}$ vector, otherwise observing $\mathbf{Y}_t$ does not give any additional information to $H_{t-1}(\mathbf{Y})$. If $\{\mathbf{Y}_t\}$ is weakly stationary, it means that the process is *regular.*

Equation (26) implies the decomposition of the corresponding subspaces like

$$H_t(\mathbf{Y}) = H_{t-1}(\mathbf{Y}) \oplus I_t(\mathbf{Y}) \tag{27}$$

that is the analogue of the multidimensional Wold decomposition in the case when the prediction is based on finite past measurements. The multidimensional Wold decomposition applies to the stationary and infinite past case. When $t \to \infty$, i.e. going to the future, we approach this situation.

Assume that $\widehat{\mathbf{X}}_{t|t-1}$ is already known. We shall give a recursion to find $\widehat{\mathbf{X}}_{t+1|t}$ by using the new value of $\mathbf{Y}_t$. In view of equation (27), we proceed as follows:

$$\begin{aligned}\widehat{\mathbf{X}}_{t+1|t} &= \mathrm{Proj}_{H_t(\mathbf{Y})}\mathbf{X}_{t+1} = \mathrm{Proj}_{H_{t-1}(\mathbf{Y})}\mathbf{X}_{t+1} + \mathrm{Proj}_{I_t(\mathbf{Y})}\mathbf{X}_{t+1} \\ &= \boldsymbol{A}_t\mathrm{Proj}_{H_{t-1}(\mathbf{Y})}\mathbf{X}_t + \mathrm{Proj}_{H_{t-1}(\mathbf{Y})}\mathbf{U}_t + \boldsymbol{K}_t\widetilde{\mathbf{Y}}_{t|t-1} \\ &= \boldsymbol{A}_t\widehat{\mathbf{X}}_{t|t-1} + \boldsymbol{K}_t\widetilde{\mathbf{Y}}_{t|t-1},\end{aligned} \tag{28}$$

where we utilized the background material and the fact that $\mathbf{U}_t \perp H_{t-1}(\mathbf{Y})$; we also used the first (state) equation of (25). Since $\mathrm{Proj}_{I_t(\mathbf{Y})}\mathbf{X}_{t+1}$ is a linear operation and results in a vector of $I_t(\mathbf{Y})$ (linear combination of the components of $\widetilde{\mathbf{Y}}_{t|t-1}$), its effect can be written as a matrix $\boldsymbol{K}_t$ times $\widetilde{\mathbf{Y}}_{t|t-1}$. This $n \times p$ matrix $\boldsymbol{K}_t$ is called *Kálmán gain* matrix. (In fact, the notation $\boldsymbol{K}$ is first used in the paper of Kálmán and Bucy.) In another context, when $\widehat{\mathbf{X}}_{t|t}$ is produced, then a strongly related matrix $\boldsymbol{L}_t$ emerges that in some places is also called gain matrix; however, $\boldsymbol{K}_t = \boldsymbol{A}_t\boldsymbol{L}_t$ as it will be shown later in this section.) In the stationary case, the rank $r(\leq p)$ of the spectral density matrix of the process $\{\mathbf{Y}_t\}$ is equal to the dimension of the innovation subspace if we predict based on the infinite past. In this stationary, infinite past case, $\boldsymbol{K}_t$ does not depend on t, and it is unique only if the spectral density matrix of the process $\{\mathbf{Y}_t\}$ is of full rank $(r = p)$, or equivalently, if the $p \times p$ covariance matrix of the innovations is non-singular. If $t \to \infty$, we approach the infinite past based prediction, and so, if $\mathbb{E}[\widetilde{\mathbf{Y}}_{t|t-1}\widetilde{\mathbf{Y}}_{t|t-1}^T]$ has near zero eigenvalues, this is an indication of a reduced rank spectral density matrix of $\{\mathbf{Y}_t\}$, see Section 2.2. In the nonstatonary case too, even if there are innovations (the innovations are not zeros), the innovation subspace can be of reduced rank, in which case $\mathbb{E}[\widetilde{\mathbf{Y}}_{t|t-1}\widetilde{\mathbf{Y}}_{t|t-1}^T]$ is not invertible (we shall take its generalized inverse later if necessary).

To specify the Kálmán gain matrix $\boldsymbol{K}_t$, we have to write $\widetilde{\mathbf{Y}}_{t|t-1}$ in terms of $\widehat{\mathbf{X}}_{t|t-1}$ and $\mathbf{Y}_t$. For this purpose, let us project both sides of the second (observation) equation of (25), i.e. of $\mathbf{Y}_t = \boldsymbol{C}_t\mathbf{X}_t$, onto $H_{t-1}(\mathbf{Y})$. We get that

$$\overline{\mathbf{Y}}_{t|t-1} = \boldsymbol{C}_t\widehat{\mathbf{X}}_{t|t-1}.$$

Taking the orthogonal decomposition (26) of $\mathbf{Y}_t$ into consideration yields that

$$\widetilde{\mathbf{Y}}_{t|t-1} = \mathbf{Y}_t - \overline{\mathbf{Y}}_{t|t-1} = \mathbf{Y}_t - \boldsymbol{C}_t\widehat{\mathbf{X}}_{t|t-1}. \tag{29}$$

We substitute this into the last line of equation (28) and obtain that

$$\widehat{\mathbf{X}}_{t+1|t} = \boldsymbol{A}_t\widehat{\mathbf{X}}_{t|t-1} + \boldsymbol{K}_t\widetilde{\mathbf{Y}}_{t|t-1} = (\boldsymbol{A}_t - \boldsymbol{K}_t\boldsymbol{C}_t)\widehat{\mathbf{X}}_{t|t-1} + \boldsymbol{K}_t\mathbf{Y}_t. \tag{30}$$

With the notation

$$\boldsymbol{A}_t^* = \boldsymbol{A}_t - \boldsymbol{K}_t\boldsymbol{C}_t \tag{31}$$

for the updated transition matrix, we get the new linear dynamics:

$$\widehat{\mathbf{X}}_{t+1|t} = \boldsymbol{A}_t^*\widehat{\mathbf{X}}_{t|t-1} + \boldsymbol{K}_t\mathbf{Y}_t. \tag{32}$$

It is also important that equations (28) and (31) give two equivalent formulas for the prediction of $\widehat{\mathbf{X}}_{t+1|t}$:

$$\widehat{\mathbf{X}}_{t+1|t} = \boldsymbol{A}_t\widehat{\mathbf{X}}_{t|t-1} + \boldsymbol{K}_t(\mathbf{Y}_t - \boldsymbol{C}_t\widehat{\mathbf{X}}_{t|t-1}) = \boldsymbol{A}_t^*\widehat{\mathbf{X}}_{t|t-1} + \boldsymbol{K}_t\mathbf{Y}_t. \tag{33}$$

We shall intensively use this equivalence.

The estimation error is also governed by the linear dynamical system. This error term is

$$\begin{aligned}\widetilde{\mathbf{X}}_{t+1|t} &= \mathbf{X}_{t+1} - \widehat{\mathbf{X}}_{t+1|t} = \boldsymbol{A}_t\mathbf{X}_t + \mathbf{U}_t - \boldsymbol{A}_t^*\widehat{\mathbf{X}}_{t|t-1} - \boldsymbol{K}_t\boldsymbol{C}_t\mathbf{X}_t \\ &= \boldsymbol{A}_t^*(\mathbf{X}_t - \widehat{\mathbf{X}}_{t|t-1}) + \mathbf{U}_t = \boldsymbol{A}_t^*\widetilde{\mathbf{X}}_{t|t-1} + \mathbf{U}_t,\end{aligned} \tag{34}$$

so $\boldsymbol{A}_t^*$ is not only the transition matrix in (32), but it is also the transition matrix of the linear dynamical system governing the error. By the equivalence, stated in equation (33), we get another expression for the same error term:

$$\begin{aligned}\widetilde{\mathbf{X}}_{t+1|t} &= \mathbf{X}_{t+1} - \widehat{\mathbf{X}}_{t+1|t} \\ &= \boldsymbol{A}_t\mathbf{X}_t + \mathbf{U}_t - \boldsymbol{A}_t\widehat{\mathbf{X}}_{t|t-1} - \boldsymbol{K}_t(\mathbf{Y}_t - \boldsymbol{C}_t\widehat{\mathbf{X}}_{t|t-1}) \\ &= \boldsymbol{A}_t\widetilde{\mathbf{X}}_{t|t-1} + \mathbf{U}_t - \boldsymbol{K}_t(\mathbf{Y}_t - \boldsymbol{C}_t\widehat{\mathbf{X}}_{t|t-1}).\end{aligned} \tag{35}$$

In the heart of the algorithm there is a recursion for the propagation of the the covariance matrix of the above error term, which is defined as

$$\boldsymbol{P}(t) = \mathbb{E}[\widetilde{\mathbf{X}}_{t|t-1}\widetilde{\mathbf{X}}_{t|t-1}^T]. \tag{36}$$

Then we shall write $\boldsymbol{P}(t+1)$ in terms of $\boldsymbol{P}(t)$ with the help of the two alternative equations (34) and (35) for the same error term:

$$\begin{aligned}&\boldsymbol{P}(t+1) = \mathbb{E}[\widetilde{\mathbf{X}}_{t+1|t}\widetilde{\mathbf{X}}_{t+1|t}^T] \\ &= \mathbb{E}[(\boldsymbol{A}_t^*\widetilde{\mathbf{X}}_{t|t-1} + \mathbf{U}_t)(\boldsymbol{A}_t\widetilde{\mathbf{X}}_{t|t-1} + \mathbf{U}_t - \boldsymbol{K}_t(\mathbf{Y}_t - \boldsymbol{C}_t\widehat{\mathbf{X}}_{t|t-1}))^T] \\ &= \boldsymbol{A}_t^*\mathbb{E}[\widetilde{\mathbf{X}}_{t|t-1}\widetilde{\mathbf{X}}_{t|t-1}^T]\boldsymbol{A}_t^T + \boldsymbol{Q}_{\mathbf{U}}(t) = \boldsymbol{A}_t^*\boldsymbol{P}(t)\boldsymbol{A}_t^T + \boldsymbol{Q}_{\mathbf{U}}(t),\end{aligned} \tag{37}$$

where recall that $\boldsymbol{Q}_{\mathbf{U}}(t) = \mathbb{E}[\mathbf{U}_t\mathbf{U}_t^T]$ and we used that $\mathbf{U}_t$ is uncorrelated with $\mathbf{X}_t$ and, therefore, with $\widetilde{\mathbf{X}}_{t|t-1}$ too; further, that $\mathbf{Y}_t - \boldsymbol{C}_t\widehat{\mathbf{X}}_{t|t-1}$ is within the innovation subspace $I_t(\mathbf{Y})$.

It remains to find an explicit formula for $\boldsymbol{K}$, and thus, also for $\boldsymbol{A}^*$. Recall that $\boldsymbol{K}_t$ is the matrix of the linear operation $\mathrm{Proj}_{I_t(\mathbf{Y})}\mathbf{X}_{t+1}$, therefore by the geometry of projections:

$$\boldsymbol{K}_t = [\mathbb{E}\mathbf{X}_{t+1}\widetilde{\mathbf{Y}}^T_{t|t-1}][\mathbb{E}(\widetilde{\mathbf{Y}}_{t|t-1}\widetilde{\mathbf{Y}}^T_{t|t-1}]^+,$$

where $^+$ denotes the Moore–Penrose generalized inverse. Now we calculate the matrices in brackets. By the second equation of (25), that extends to $\widetilde{\mathbf{Y}}_t = \boldsymbol{C}_t\widetilde{\mathbf{X}}_t$, we get that

$$\mathbb{E}\widetilde{\mathbf{Y}}_{t|t-1}\widetilde{\mathbf{Y}}^T_{t|t-1} = \mathbb{E}(\boldsymbol{C}_t\widetilde{\mathbf{X}}_{t|t-1})(\boldsymbol{C}_t\widetilde{\mathbf{X}}_{t|t-1})^T = \boldsymbol{C}_t\boldsymbol{P}(t)\boldsymbol{C}_t^T.$$

By the first and second equation of (25) and the orthogonality of $\widehat{\mathbf{X}}_{t|t-1}$ and $\widetilde{\mathbf{X}}_{t|t-1}$:

$$\begin{aligned}\mathbb{E}\mathbf{X}_{t+1}\widetilde{\mathbf{Y}}^T_{t|t-1} &= \boldsymbol{A}_t\mathbb{E}\mathbf{X}_t\widetilde{\mathbf{Y}}^T_{t|t-1} = \boldsymbol{A}_t\mathbb{E}(\widehat{\mathbf{X}}_{t|t-1} + \widetilde{\mathbf{X}}_{t|t-1})(\boldsymbol{C}_t\widetilde{\mathbf{X}}^T_{t|t-1}) \\ &= \boldsymbol{A}_t\boldsymbol{P}(t)\boldsymbol{C}_t^T.\end{aligned} \tag{38}$$

Therefore,

$$\boldsymbol{K}_t = \boldsymbol{A}_t\boldsymbol{P}(t)\boldsymbol{C}_t^T[\boldsymbol{C}_t\boldsymbol{P}(t)\boldsymbol{C}_t^T]^+. \tag{39}$$

Instead of the Moore–Penrose generalized inverse, we use the regular inverse provided the matrix in brackets is invertible, i.e. the innovation subspace $I_t(\mathbf{Y})$ is of full dimension p, and $\boldsymbol{C}_t$ is of full rank p.

Then the recursion starts at $t = 1$, when the systems of p linear equations $\boldsymbol{C}_1\widehat{\mathbf{X}}_{1|0} = \overline{\mathbf{Y}}_{1|0}$ and $\boldsymbol{C}_1\mathbf{X}_1 = \mathbf{Y}_1$ should be solved for the coordinates of $\widehat{\mathbf{X}}_{1|0}$ and $\mathbf{X}_1$, respectively (the n coordinates are the unknowns). They obviously have a solution if $\boldsymbol{C}_1$ is of full rank. Here $\overline{\mathbf{Y}}_{1|0} = \mathbb{E}(\mathbf{Y}_1)$ if $\mathbf{Y}_0$ is a constant vector. Even if it is $\mathbf{0}$, the system has a nontrivial solution in the $p \le n$ case. Then $\widetilde{\mathbf{X}}_{1|0} = \mathbf{X}_1 - \widehat{\mathbf{X}}_{1|0}$. In the original paper of Kálmán, the following starting is suggested: $\widehat{\mathbf{X}}_{1|0} := \mathbf{0}$; $\widetilde{\mathbf{X}}_{1|0} := \mathbf{X}_1$; $\boldsymbol{P}(1) := \mathbb{E}[\mathbf{X}_1\mathbf{X}_1^T]$. This can be the product moment estimate from the training sample (possible past).

We can summarize the above results and recursion as follows.

Proposition 4. *The optimal estimate $\widehat{\mathbf{X}}_{t+1|t}$ of $\mathbf{X}_{t+1}$ given $\mathbf{Y}_1, \dots, \mathbf{Y}_t$ is generated by the linear dynamical system*

$$\widehat{\mathbf{X}}_{t+1|t} = \boldsymbol{A}_t^*\widehat{\mathbf{X}}_{t|t-1} + \boldsymbol{K}_t\mathbf{Y}_t.$$

The estimation error is given by

$$\widetilde{\mathbf{X}}_{t+1|t} = \boldsymbol{A}_t^* \widetilde{\mathbf{X}}_{t|t-1} + \mathbf{U}_t$$

and the propagated covariance matrix of the estimation error is

$$\boldsymbol{P}(t) = \mathbb{E}[\widetilde{\mathbf{X}}_{t|t-1} \widetilde{\mathbf{X}}_{t|t-1}^T],$$

while the expected quadratic loss is $\mathrm{tr}\boldsymbol{P}(t)$. *The matrices involved are generated by the following recursion. Starting with*

$$\widehat{\mathbf{X}}_{1|0} = \mathrm{Proj}_{\mathbf{Y}_0}\mathbf{X}_1,\ \widetilde{\mathbf{X}}_1 = \mathbf{X}_1 - \widehat{\mathbf{X}}_{1|0},\ \boldsymbol{P}(1) = \mathbb{E}[\widetilde{\mathbf{X}}_{1|0}\widetilde{\mathbf{X}}_{1|0}^T] = \mathbb{E}[\mathbf{X}_1\mathbf{X}_1^T] - \mathbb{E}[\widehat{\mathbf{X}}_{1|0}\widehat{\mathbf{X}}_{1|0}^T],$$

for $t = 1, 2, \dots,$ *the steps of the following recursion are uniquely defined:*

- *Evaluate* $\boldsymbol{K}_t$ *by (39):* $\boldsymbol{K}_t = \boldsymbol{A}_t \boldsymbol{P}(t) \boldsymbol{C}_t^T [\boldsymbol{C}_t \boldsymbol{P}(t) \boldsymbol{C}_t^T]^+$.
- *Input* $\mathbf{Y}_t$. *Output*

  $$\widehat{\mathbf{X}}_{t+1|t} = (\boldsymbol{A}_t - \boldsymbol{K}_t \boldsymbol{C}_t)\widehat{\mathbf{X}}_{t|t-1} + \boldsymbol{K}_t \mathbf{Y}_t.$$

- *Evaluate* $\boldsymbol{A}_t^*$ *by (31):* $\boldsymbol{A}_t^* = \boldsymbol{A}_{t+1|t} - \boldsymbol{K}_t \boldsymbol{C}_t$.
- *Eventually, calculate* $\boldsymbol{P}(t+1)$ *by (37) that completes the cycle:*

  $$\boldsymbol{P}(t+1) = \boldsymbol{A}_t^* \boldsymbol{P}(t) \boldsymbol{A}_t^T + \boldsymbol{Q}_{\mathbf{U}}(t).$$

  Note that $\boldsymbol{Q}_{\mathbf{U}}(t)$ *is known/given or estimated from a training sample. In the forthcoming Remark 11, a symmetric expression is also given for the matrix* $\boldsymbol{P}(t+1)$.

Some remarks are in order.

Remark 7. *As for the starting,* $\widehat{\mathbf{X}}_{1|0} = \mathrm{Proj}_{\mathbf{Y}_0}\mathbf{X}_1 = \hat{\boldsymbol{\Sigma}}_{\mathbf{XY}} \hat{\boldsymbol{\Sigma}}_{\mathbf{YY}}^+ \mathbf{Y}_0$, *where the last training sample entry can be chosen for* $\mathbf{Y}_0$. *To initialize* $\boldsymbol{P}(1)$, *the whole training sample can be used. Another possibility is to start with* $\widehat{\mathbf{X}}_{1|0} = \mathbf{0}$.

Remark 8. *As a byproduct, the algorithm is able to get the innovations via equation (29).*

Remark 9. *In some situations, the observation equation also contains a noise term, for example; Kálmán and Bucy consider the continuous-time case, but they write that even in this case, the assumption that every observed signal contains a white noise term, "is unnecessary when the random processes in question are sampled (discrete-time parameter)"; even in the continuous-time case, it "is no real restriction since it can be removed in various ways". However, the random excitation in the state (message) process "is quite basic; it is analogous to but somewhat less restrictive than the assumption of rational spectra in the conventional theory". Indeed, Kálmán uses only the regularity (causality) of the process if stationary, but not the rational spectral density. He mostly considers Gaussian processes that is not a restriction in the possession of second order processes, when we confine ourselves to the second moments.*

In this case, the state equations have the form

$$\begin{aligned} \mathbf{X}_{t+1} &= \boldsymbol{A}_t \mathbf{X}_t + \mathbf{U}_t \\ \mathbf{Y}_t &= \boldsymbol{C}_t \mathbf{X}_t + \mathbf{W}_t, \end{aligned} \tag{40}$$

where $\mathbf{W}_t$ is independent of $\mathbf{X}_t$ and $\mathbf{U}_t$ (latter condition can be relaxed by introducing the covariance matrix between $\mathbf{U}_t$ and $\mathbf{W}_t$ as a given parameter; further the covariance matrix of the zero expectation $\mathbf{W}_t$ is $\boldsymbol{Q}_{\mathbf{W}} = \mathbb{E}\mathbf{W}_t\mathbf{W}_t^T$ is also given.

The only difference in the calculations is that now

$$\begin{aligned} \mathbb{E}\widetilde{\mathbf{Y}}_{t|t-1}\widetilde{\mathbf{Y}}_{t|t-1}^T = \mathbb{E}(\boldsymbol{C}_t\widetilde{\mathbf{X}}_{t|t-1} + \mathbf{W}_t)(\boldsymbol{C}_t\widetilde{\mathbf{X}}_{t|t-1} + \mathbf{W}_t)^T = \boldsymbol{C}_t\boldsymbol{P}(t)\boldsymbol{C}_t^T \\ + \boldsymbol{Q}_{\mathbf{W}}(t), \end{aligned}$$

and so,

$$\boldsymbol{K}_t = \boldsymbol{A}_t\boldsymbol{P}(t)\boldsymbol{C}_t^T[\boldsymbol{C}_t\boldsymbol{P}(t)\boldsymbol{C}_t^T + \boldsymbol{Q}_{\mathbf{W}}(t)]^+.$$

Instead of $\mathbf{U}_t$ we may write $\boldsymbol{B}_t\mathbf{V}_t$ with some $n \times q$ matrix $\boldsymbol{B}_t$ with $q \le n$ and q-dimensional orthogonal noise $\mathbf{V}_t$, i.e. $\mathbb{E}\mathbf{V}_t\mathbf{V}_t^T = \boldsymbol{Q}_{\mathbf{V}}(t)$ is a given diagonal matrix. Here instead of $\boldsymbol{Q}(t)$ the matrix $\boldsymbol{B}_t\boldsymbol{Q}_{\mathbf{V}}(t)\boldsymbol{B}_t^T$ enters into the equation (37). This approach mainly used in the stationary case, when a lower rank driving force (excitation) is assumed, but this is the topic of Dynamic Factor Analysis, see Section 5.

In the same vein, instead of $\mathbf{W}_t$ we may write $\boldsymbol{D}_t\mathbf{Z}_t$ with some $p \times s$ matrix $\boldsymbol{B}_t$ with $s \le p$ and s-dimensional orthogonal noise $\mathbf{Z}_t$, i.e. $\mathbb{E}\mathbf{Z}_t\mathbf{Z}_t^T = \boldsymbol{Q}_{\mathbf{Z}}(t)$ is a given diagonal matrix. Here instead of $\boldsymbol{Q}_{\mathbf{W}}(t)$ the, possibly reduced rank, matrix $\boldsymbol{D}_t\boldsymbol{Q}_{\mathbf{Z}}(t)\boldsymbol{D}_t^T$ enters into the calculations.

Remark 10. *Equation (32) gives rise to a predictive filtering, in the possession of the gain matrix $\boldsymbol{K}_t$. After this, the algorithm is also applicable to filtering. Indeed,*

$$\widehat{\mathbf{X}}_{t+1|t} = \mathrm{Proj}_{H_t(\mathbf{Y})}\mathbf{X}_{t+1} = \mathrm{Proj}_{H_t(\mathbf{Y})}(\boldsymbol{A}_t\mathbf{X}_t + \mathbf{U}_t) = \boldsymbol{A}_t\widehat{\mathbf{X}}_{t|t}.$$

Now, provided $\boldsymbol{A}_t$ is invertible,

$$\widehat{\mathbf{X}}_{t|t} = \boldsymbol{A}_t^{-1}\widehat{\mathbf{X}}_{t+1|t}.$$

If $\boldsymbol{A}_t$ is not invertible, then we proceed as follows:

$$\widehat{\mathbf{X}}_{t|t} = \mathrm{Proj}_{H_t(\mathbf{Y})}\mathbf{X}_t = \mathrm{Proj}_{H_{t-1}(\mathbf{Y})}\mathbf{X}_t + \mathrm{Proj}_{I_t(\mathbf{Y})}\mathbf{X}_t = \widehat{\mathbf{X}}_{t|t-1} + \boldsymbol{L}_t\widetilde{\mathbf{Y}}_{t|t-1}.$$

Now the gain matrix is $\boldsymbol{L}_t$, which is not the same as $\boldsymbol{K}_t$ (though, sometimes this is what called Kálmán gain matrix), can be determined with a similar calculation:

$$\boldsymbol{L}_t = [\mathbb{E}\mathbf{X}_t\widetilde{\mathbf{Y}}_{t|t-1}^T][\mathbb{E}(\widetilde{\mathbf{Y}}_{t|t-1}\widetilde{\mathbf{Y}}_{t|t-1}^T]^+.$$

The only difference between the formula for $\boldsymbol{K}_t$ and $\boldsymbol{L}_t$ that here we calculate the covariance between $\mathbf{X}_t$ and $\widetilde{\mathbf{Y}}_{t|t-1}^T$, but equation (38) is at our disposal in this situation too. We get that

$$\mathbb{E}\mathbf{X}_t\widetilde{\mathbf{Y}}_{t|t-1}^T = \boldsymbol{P}(t)\boldsymbol{C}_t^T,$$

and so,

$$\boldsymbol{L}_t = \boldsymbol{P}(t)\boldsymbol{C}_t^T[\boldsymbol{C}_t\boldsymbol{P}(t)\boldsymbol{C}_t^T]^+.$$

Consequently,

$$\boldsymbol{K}_t = \boldsymbol{A}_t\boldsymbol{L}_t,$$

so we could first find

$$\boldsymbol{L}_t = \boldsymbol{P}(t)\boldsymbol{C}_t^T[\boldsymbol{C}_t\boldsymbol{P}(t)\boldsymbol{C}_t^T]^+$$

and then, $\boldsymbol{K}_t$. Therefore, in course of the iteration, the filtered process $\{\widehat{\mathbf{X}}_{t|t}\}$ can as well be obtained.

Remark 11. *If we write the expression for $\boldsymbol{K}_t$, $\boldsymbol{L}_t$, and $\boldsymbol{A}_t^*$ into equation (37), then we get*

$$\begin{aligned}
&\boldsymbol{P}(t+1) = \boldsymbol{A}_t^*\boldsymbol{P}(t)\boldsymbol{A}_t{}^T + \boldsymbol{Q}_{\mathbf{U}}(t)\\
&= (\boldsymbol{A}_t - \boldsymbol{A}_t\boldsymbol{L}_t\boldsymbol{C}_t)\boldsymbol{P}(t)\boldsymbol{A}_t{}^T + \boldsymbol{Q}(t)\\
&= (\boldsymbol{A}_t(\boldsymbol{I} - \boldsymbol{L}_t\boldsymbol{C}_t))\boldsymbol{P}(t)\boldsymbol{A}_t{}^T + \boldsymbol{Q}_{\mathbf{U}}(t)\\
&= \boldsymbol{A}_t\boldsymbol{P}(t)\boldsymbol{A}_t{}^T - \boldsymbol{A}_t\boldsymbol{L}_t\boldsymbol{C}_t\boldsymbol{P}(t)\boldsymbol{A}_t{}^T + \boldsymbol{Q}_{\mathbf{U}}(t)\\
&= \boldsymbol{A}_t\boldsymbol{P}(t)\boldsymbol{A}_t{}^T - \boldsymbol{A}_t\boldsymbol{P}(t)\boldsymbol{C}_t^T[\boldsymbol{C}_t\boldsymbol{P}(t)\boldsymbol{C}_t^T]^{-1}\boldsymbol{C}_t\boldsymbol{P}(t)\boldsymbol{A}_t{}^T\\
&+ \boldsymbol{Q}_{\mathbf{U}}(t)
\end{aligned}\tag{41}$$

which final formula shows that $\boldsymbol{P}(t+1)$ is indeed a symmetric matrix.

Remark 12. *Assume that the underlying process is weakly stationary, and put $\boldsymbol{A}$ for $\boldsymbol{A}_t$, $\boldsymbol{C}$ for $\boldsymbol{C}_t$, and $\boldsymbol{Q}_{\mathbf{U}}$ for $\boldsymbol{Q}_{\mathbf{U}}(t)$. In this case, instead of the recursion, we get the fixed point iteration*

$$\boldsymbol{P}_{t+1} = \boldsymbol{A}\boldsymbol{P}_t\boldsymbol{A}^T - \boldsymbol{A}\boldsymbol{P}_t\boldsymbol{C}^T[\boldsymbol{C}\boldsymbol{P}_t\boldsymbol{C}^T]^+\boldsymbol{C}\boldsymbol{P}_t\boldsymbol{A}^T + \boldsymbol{Q}_{\mathbf{U}},$$

where now $\boldsymbol{P}_t$ just denotes the t-th step of the iteration. Note that some authors consider the question when the discrete matrix Riccati equation

$$\boldsymbol{P} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T - \boldsymbol{A}\boldsymbol{P}\boldsymbol{C}^T[\boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^T]^+\boldsymbol{C}\boldsymbol{P}\boldsymbol{A}^T + \boldsymbol{Q}_{\mathbf{U}},\tag{42}$$

has a unique solution and so, the method of successive approximation, resembling the recursion in (41), is able to find it. (Actually, the Riccati operator is concave and has a unique fixed point under very general conditions.) With this $\boldsymbol{P}$, the limit of the sequence $\boldsymbol{K}_t$ is $\boldsymbol{K} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{C}^T[\boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^T]^+$ as $t \to \infty$, and $\boldsymbol{L}_t$ also has a limiting $\boldsymbol{L} = \boldsymbol{P}\boldsymbol{C}^T[\boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^T]^+$, when our sequence is weakly stationary.

The paper of Kálmán gives guidance to the solution, mainly considers the continuous time case, and contains many applications in engineering and telecommunication. The authors also discuss the relation to differential equations and the Fisher information matrix.

We remark that in the possession of another error term $\mathbf{W}$ (but $\boldsymbol{Q}_{\mathbf{W}}$ does not depend on the time), equation (42) has the slightly modified form

$$\boldsymbol{P} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T - \boldsymbol{A}\boldsymbol{P}\boldsymbol{C}^T[\boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^T + \boldsymbol{Q}_{\mathbf{W}}]^+\boldsymbol{C}\boldsymbol{P}\boldsymbol{A}^T + \boldsymbol{Q}_{\mathbf{U}},$$

though it does not change the type of the matrix Riccati equation.

In the stationary case, the stability of the matrix $\boldsymbol{A}$ should be assumed, as well as that of the new transition matrix, corresponding to $\boldsymbol{A}^*$, which is $\boldsymbol{A} - \boldsymbol{K}\boldsymbol{C}$.

5 Dynamic Principal Component and Factor Analysis

Here we confine ourselves to high dimensional weakly stationary processes that are usually of lower rank than their dimension or can be approximated with a lower rank process. In the time domain, we are looking for the convenient filters and for the matrices in the state equations too. In the frequency domain, we use the low rank approximation of the spectral density matrix at the Fourier frequencies. We summarize the findings based of the previous sections.

5.1 Time domain approach via innovations

First we use the method of innovations. If $\mathbf{X}_t$s have different dimensions, then denoting by d the minimal dimension, first we perform a static factor analysis on them, and start with the so obtained d-dimensional static factor process. We also deprive the process from trend and seasonality, and assume that it has a spectral density matrix of constant rank. If the process is also deprived of the singular part, then a regular process is at our disposal.

If $\mathbf{X}_t$ is regular, we learned that it can be expanded in terms of the d-dimensional innovations

$$\boldsymbol{\eta}_{t+1} := \mathbf{X}_{t+1} - \hat{\mathbf{X}}_{t+1},$$

where $\hat{\mathbf{X}}_{t+1}$ is the projection of $\mathbf{X}_{t+1}$ onto the subspace spanned by $\mathbf{X}_1, \dots, \mathbf{X}_t$, denoted by H_t. It can be done step by step as described in Section 2.2. If not regular, the prediction process gives the regular part of it.

We can as well reduce the dimension of the innovation process to $k < d$. This k-dimensional innovation process can be considered as a dynamic factor process, where $k \le r$, and r is the rank of the spectral density matrix of the process. As an alternative to the block Cholesky decomposition, the Kálmán filtering is also able to find the innovations, see equation (29). In this way, instead of the decomposition of a huge block matrix, we operate with matrices of size comparable to the dimension of the process.

The above is also related to the minimal phase spectral factor. To find this and a reduced rank causal approximation of a process of rational spectral density,.

We saw that a d-dimensional regular process $\{\mathbf{X}_t\}$, whose spectral density matrix $\boldsymbol{f}$ is of rank $r \le d$ has the variant of the multidimensional Wold decomposition:

$$\mathbf{X}_t = \sum_{j=0}^{\infty} \boldsymbol{B}_j \boldsymbol{\nu}_{t-j}, \tag{43}$$

where $\boldsymbol{B}_j$s are $d \times r$ matrices (like dynamic factor loadings), and $\{\boldsymbol{\nu}_t\} \sim \mathrm{WN}(\boldsymbol{\Sigma})$ is r-dimensional white noise (like non-standardized minimal dynamic factors).

It is important that there is a one-to-one correspondence between $\boldsymbol{f}$ (frequency domain) and the $\boldsymbol{B}(z), \boldsymbol{\Sigma}$ pair (time domain):

$$\boldsymbol{B}(z) = \sum_{j=0}^{\infty} \boldsymbol{B}_j z^j, \quad |z| \le 1 \tag{44}$$

and $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{\nu}_t$. This correspondence is given by

$$\boldsymbol{f}(z) = \frac{1}{2\pi} \boldsymbol{B}(z) \boldsymbol{\Sigma} \boldsymbol{B}^*(z).$$

We can as well write $\boldsymbol{f}(z) = \frac{1}{2\pi} \boldsymbol{H}(z) \boldsymbol{H}^*(z)$, where $\boldsymbol{H}(z) = \boldsymbol{B}(z) \boldsymbol{\Sigma}^{1/2}$ is the transfer function and it is unique only up to unitary transformation. At the same time, the matrices $\boldsymbol{B}_j \boldsymbol{\Sigma}^{-1/2}$ are the impulse responses. So by performing the expansion (44) at the Fourier frequencies, we can estimate the transfer function.

In Section 2.2, we give an algorithm to this in the time domain, via block Cholesky decomposition. Then we can perform a static PCA on $\boldsymbol{\Sigma}$ with $k \le r$ principal components, that results in dynamic factors of dimension k. The choice of k is such that there are $n(r-k)$ negligible eigenvalues in the spectrum of $\mathfrak{C}_n$. By Theorem 2 , for 'large' n, this is in accord with the existence of $r-k$ negligible eigenvalues of the spectral density matrix at all the n Fourier frequencies. Therefore, we proceed in the frequency domain.

5.2 Frequency domain approach

Let $\{\mathbf{X}_t\}$ be discrete time, d-dimensional, weakly stationary time series of zero expectation and spectral density matrix of constant rank. For given $0 < k \le d$ we are looking for the k-dimensional time series $\mathbf{Y}_t$ such that

$$\mathbf{Y}_t = \sum_j \boldsymbol{b}_{t-j} \mathbf{X}_j, \quad t \in \mathbf{Z},$$

where $\boldsymbol{b}_j$s are $k \times d$ matrices and $\boldsymbol{b}$ is the corresponding transfer function. (Here k is less than the rank of the process itself.)

Then approximate $\mathbf{X}_t$ with

$$\hat{\mathbf{X}}_t = \sum_j \boldsymbol{c}_{t-j} \mathbf{Y}_j, \quad t \in \mathbf{Z},$$

where the impulse responses $\boldsymbol{c}_j$s are $d \times k$ matrices, and $\boldsymbol{c}$ is the transfer function.

So $\hat{\mathbf{X}}$ is obtained from $\mathbf{X}$ with the time invariant filter

$$\boldsymbol{a}(\omega) = \boldsymbol{c}(\omega)\boldsymbol{b}(\omega).$$

The error of approximation is measured with

$$\mathbb{E}(\mathbf{X}_t - \hat{\mathbf{X}}_t)^*(\mathbf{X}_t - \hat{\mathbf{X}}_t).$$

Then Brillinger in his book states that the minimum is attained with the impulse responses

$$\boldsymbol{b}_j = \frac{1}{2\pi} \int_0^{2\pi} \boldsymbol{b}(\omega) e^{ij\omega}\, d\omega$$

and

$$\boldsymbol{c}_j = \frac{1}{2\pi} \int_0^{2\pi} \boldsymbol{c}(\omega) e^{ij\omega}\, d\omega,$$

where

$$\boldsymbol{c}(\omega) = (\mathbf{u}_1(\omega), \dots, \mathbf{u}_k(\omega))$$

contains columnwise the orthonormal eigenvectors corresponding to the k largest eigenvalues of the spectral density matrix $\boldsymbol{f}$ of $\{\mathbf{X}_t\}$. Further, $\boldsymbol{b}(\omega) = \boldsymbol{c}^*(\omega)$. The approximation error is

$$\int_0^{2\pi} \sum_{j=k+1}^{d} \lambda_j(\omega)\, d\omega.$$

This is in Frobenius norm, in spectral norm it only depends on λ_{k+1}, but the best k-rank approximation is the same in any unitary invariant norm (that depends only on the eigenvalues). The larger the gap in the spectrum between the k largest and the other eigenvalues, the better the approximation is.

$\{\mathbf{Y}_t\}$ is called principal component process. Its spectral density matrix is diagonal with diagonal entries $\lambda_1(\omega), \dots \lambda_k(\omega)$. If the original process is regular, then its best k-rank approximation is regular too.

5.3 Best low-rank approximation in the frequency domain, and low-dimensional approximation in the time domain

Let $\{\mathbf{X}_t\}_{t=1}^n$ be the finite part of a d-dimensional process of real coordinates and constant rank $1 \le r \le d$. Its discrete Fourier transform, discussed in Section 3.2, is

$$\mathbf{T}_j = \mathbf{V}_j \mathbf{Z}_j = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{X}_t e^{-it\omega_j}, \quad j = 0, \dots, n-1.$$

More precisely, $\mathbf{T}_0 = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{X}_t$,

$$\mathbf{T}_j = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{X}_t [\cos(t\omega_j) - i \sin(t\omega_j)],$$

and $\mathbf{T}_{n-j} = \overline{\mathbf{T}}_j$, for $j = 1, \dots, k$ $(n = 2k+1)$. Therefore,

$$\mathbf{Z}_j = \mathbf{V}_j^{-1} \mathbf{T}_j = \mathbf{V}_j^* \mathbf{T}_j, \quad j = 0, \dots, n-1.$$

It can easily be seen that $\mathbf{Z}_{n-j} = \overline{\mathbf{Z}}_j$.

To find the best m-rank approximation $(1 < m \le r)$ of the process, we project the d-dimensional vector $\mathbf{T}_j$ onto the subspace spanned by the m leading eigenvectors of $\mathbf{V}_j$ for the linear algebra justification for this). Important that the eigenvalues in $\mathbf{\Lambda}_j$ are in non-increasing order. Let us denote the eigenvectors corresponding to the m largest eigenvalues by $\mathbf{v}_{j1}, \dots, \mathbf{v}_{jm}$. Then

$$\widehat{\mathbf{T}}_j := \mathrm{Proj}_{\mathrm{Span}\{\mathbf{v}_{j1},\dots,\mathbf{v}_{jm}\}} \mathbf{T}_j = \sum_{\ell=1}^{m} (\mathbf{v}_{j\ell}^* \mathbf{V}_j \mathbf{Z}_j) \mathbf{v}_{j\ell} = \sum_{\ell=1}^{m} Z_{j\ell} \mathbf{v}_{j\ell},$$

and $\widehat{\mathbf{T}}_{n-j} = \overline{\widehat{\mathbf{T}}}_j$, for for $j = 1, \dots, k$ (by the previous considerations), were $n = 2k+1$. Further,

$$\widehat{\mathbf{T}}_0 := \sum_{\ell=1}^{m} Z_{0\ell} \mathbf{v}_{0\ell}.$$

So, for each j, the resulting vector is the linear combination of the vectors $\mathbf{v}_{j\ell}$s with the corresponding coordinates $Z_{j\ell}$s of $\mathbf{Z}_j$, $\ell = 1, \dots, m$.

Eventually, we find the m-rank approximation of $\mathbf{X}_t$ by inverse Fourier transformation:

$$\begin{aligned}\widehat{\mathbf{X}}_t &:= \frac{1}{\sqrt{n}}\sum_{j=0}^{n-1}\widehat{\mathbf{T}}_j e^{it\omega_j} = \\ &= \frac{1}{\sqrt{n}}\left\{\widehat{\mathbf{T}}_0 + \sum_{j=1}^{k}[(\widehat{\mathbf{T}}_j + \overline{\widehat{\mathbf{T}}}_j)\cos(t\omega_j) + i(\widehat{\mathbf{T}}_j - \overline{\widehat{\mathbf{T}}}_j)\sin(t\omega_j)]\right\} \\ &= \frac{1}{\sqrt{n}}\left\{\widehat{\mathbf{T}}_0 + \sum_{j=1}^{k}[(2\mathrm{Re}(\widehat{\mathbf{T}}_j)\cos(t\omega_j) + i\cdot 2i\cdot \mathrm{Im}(\widehat{\mathbf{T}}_j)\sin(t\omega_j)]\right\} \\ &= \frac{1}{\sqrt{n}}\left\{\widehat{\mathbf{T}}_0 + 2\sum_{j=1}^{k}[\mathrm{Re}(\widehat{\mathbf{T}}_j)\cos(t\omega_j) - \mathrm{Im}(\widehat{\mathbf{T}}_j)\sin(t\omega_j)]\right\}.\end{aligned}$$

Apparently, the vectors $\widehat{\mathbf{X}}_t$ ($t = 1, \dots, n$) all have real coordinates ($n = 2k+1$).

In this way, we have a lower rank process with spectral density of rank $m \le r$. Note that if the process is regular (e.g. it has a rational spectral density), then so is its low-rank approximation. The theory guarantees that the 'larger' the gap between the mth and $(m+1)$th eigenvalues (in non-increasing order) of the spectral density matrix, the 'smaller' the approximation error is.

To back-transform the PC process into the time domain, note that

$$Z_{j\ell} = \mathbf{v}_{j\ell}^*\mathbf{T}_j, \quad \ell = 1, \dots m$$

defines the coordinates of an m-dimensional approximation of $\mathbf{T}_j$, $m \le r \le d$. This is the m-dimensional vector $\tilde{\mathbf{T}}_j = (Z_{j1}, \dots, Z_{jm})^T$. That is, we take the first m complex PCs in each blocks (it is important that the entries in the diagonal of each $\mathbf{\Lambda}_j$ are in non-increasing order). The other $d-m$ coordinates of $\mathbf{Z}_j$ are disregarded (they are taken zeros in the new coordinate system $\mathbf{v}_{j1}, \dots, \mathbf{v}_{jd}$). The proportion of the total variance explained by the first m principal components at the jth Fourier frequency is $\sum_{\ell=1}^m \lambda_{j\ell}/\sum_{\ell=1}^d \lambda_{j\ell}$.

Then the m-dimensional approximation of $\mathbf{X}_t$ by the PC process is as follows:

$$\begin{aligned}\tilde{\mathbf{X}}_t &:= \frac{1}{\sqrt{n}}\sum_{j=0}^{n-1}\tilde{\mathbf{T}}_j e^{it\omega_j} = \\ &= \frac{1}{\sqrt{n}}\left\{\tilde{\mathbf{T}}_0 + \sum_{j=1}^{k}[(\tilde{\mathbf{T}}_j + \overline{\tilde{\mathbf{T}}}_j)\cos(t\omega_j) + i(\tilde{\mathbf{T}}_j - \overline{\tilde{\mathbf{T}}}_j)\sin(t\omega_j)]\right\} \\ &= \frac{1}{\sqrt{n}}\left\{\tilde{\mathbf{T}}_0 + \sum_{j=1}^{k}[(2\mathrm{Re}(\tilde{\mathbf{T}}_j)\cos(t\omega_j) + i\cdot 2i\cdot \mathrm{Im}(\tilde{\mathbf{T}}_j)\sin(t\omega_j)]\right\} \\ &= \frac{1}{\sqrt{n}}\left\{\tilde{\mathbf{T}}_0 + 2\sum_{j=1}^{k}[\mathrm{Re}(\tilde{\mathbf{T}}_j)\cos(t\omega_j) - \mathrm{Im}(\tilde{\mathbf{T}}_j)\sin(t\omega_j)]\right\}\end{aligned}$$

that again results in real coordinates. Equivalently, the m-dimensional PC process is:

$$\tilde{\mathbf{X}}_t = \frac{1}{\sqrt{n}}(\sum_{j=0}^{n-1} Z_{j1}e^{it\omega_j}, \dots, \sum_{j=0}^{n-1} Z_{jm}e^{it\omega_j})^T. \tag{45}$$

5.4 Dynamic factor analysis

Standard factor analysis can be generalized to the case of a d-dimensional, real valued, vector stochastic process $\{\mathbf{X}_t\}$. Here $t \geq 0$ is the time, and our sample usually consists of observations at discrete time instances $t = 1, \dots, T$. In the classical factor analysis approach, the data come from i.i.d. observations, and the dimension reduction happens in the so-called cross-sectional dimension, i.e. the number d of variables is decreased. In dynamic factor analysis, the observations $\mathbf{X}_t$'s are not independent, and we want to compress the information, embodied by them, in the cross-sectional and the time dimension as well. Sometimes even the cross-sectional dimension d is large compared to the time span T.

Assume that $\{\mathbf{X}_t\}$ is *weakly stationary* with an absolutely continuous spectral distribution, i.e. it has the $d \times d$ spectral density matrix $\boldsymbol{f}_{\mathbf{X}}$. With the integer $1 \leq k < d$, the *dynamic k-factor model* for $\mathbf{X}_t$ is

$$\mathbf{X}_t = \boldsymbol{\mu} + \boldsymbol{B}(L)\mathbf{Z}_t + \mathbf{e}_t = \boldsymbol{\mu} + \boldsymbol{\chi}_t + \mathbf{e}_t \tag{46}$$

or with components,

$$X_t^i = \mu^i + b_{i1}(L)Z_t^1 + \cdots + b_{ik}(L)Z_t^k + e_t^i$$

where the k-dimensional stochastic process $\mathbf{Z}_t = (Z_t^1, \dots Z_t^k)^T$ is the *dynamic factor*, $\boldsymbol{\chi}_t$ is called *common component*, the d-dimensional stochastic process $\mathbf{e}_t = (e_t^1, \dots, e_t^d)^T$ is called *noise component*, and the $d \times k$ matrix $\boldsymbol{B}(L) = (b_{ij}(L))$, $i = 1, \dots, d$, $j = 1, \dots, k$, is the *transfer function*. Here L is the *lag operator* (backward shift) and $b_{ij}(L)$ is a square-summable one-sided filter, i.e. $b_{ij}(L) = b_{ij}(0) + b_{ij}(1)L + b_{ij}(2)L^2 + \dots$ with $\sum_{\ell=0}^{\infty} b_{ij}^2(\ell) < \infty$. Further, the components of (46) satisfy the following requirements:

$$\begin{aligned} \mathbb{E}(\mathbf{Z}_t) &= \mathbf{0}, \quad \mathbb{E}(\mathbf{e}_t) = \mathbf{0}, \quad t \in \mathbb{Z} \\ \mathrm{Cov}(e_t^i, Z_s^j) &= 0, \quad i = 1, \dots, d, \quad j = 1, \dots, k, \quad t, s \in \mathbb{Z},\, s \le t. \\ \mathrm{Cov}(e_t^i, e_s^j) &= 0, \quad i, j = 1, \dots d, \quad i \neq j, \quad t, s \in \mathbb{Z},\, s < t. \end{aligned} \tag{47}$$

If $\mathbf{Z}_t$ and $\mathbf{e}_t$ are also weakly stationary and they have rational spectral densities $\boldsymbol{f}_{\mathbf{Z}}$ and $\boldsymbol{f}_{\mathbf{e}}$, the model equation (46) extends to the spectral density matrices:

$$\boldsymbol{f}_{\mathbf{X}}(\omega) = \boldsymbol{f}_{\boldsymbol{\chi}}(\omega) + \boldsymbol{f}_{\mathbf{e}}(\omega) = \boldsymbol{B}(e^{-i\omega})\boldsymbol{f}_{\mathbf{Z}}(\omega)\boldsymbol{B}(e^{-i\omega})^* + \boldsymbol{f}_{\mathbf{e}}(\omega), \quad \omega \in [-\pi, \pi]. \tag{48}$$

Very frequently, $\mathbf{Z}_t$ is assumed to be orthonormal WN($\boldsymbol{I}_k$) process. Then equation (48) simplifies to

$$\boldsymbol{f}_{\mathbf{X}}(\omega) = \frac{1}{2\pi}\boldsymbol{B}(e^{-i\omega})\boldsymbol{B}(e^{-i\omega})^* + \boldsymbol{f}_{\mathbf{e}}(\omega). \tag{49}$$

The so-called *static* case occurs if, in addition, $\boldsymbol{B}$ is constant. Otherwise, equation (46) is dynamic in that the latent variables Z_t^js can affect the observables X_t^is both contemporaneously and with lags.

Like in the standard factor model, neither $\boldsymbol{B}(L)$ nor $\mathbf{Z}_t$ are identified uniquely; and given the spectral density $\boldsymbol{f}_{\mathbf{X}}$, the spectra $\boldsymbol{f}_{\boldsymbol{\chi}}$ and $\boldsymbol{f}_{\mathbf{e}}$ are generically can be determined for $k \le n - \sqrt{n}$ (reminiscent of the Lederman bound).

5.5 General Dynamic Factor Model (GDFM)

Let $\mathbf{X}_t$ be a *weakly stationary* time series ($t = 1, 2, \dots$) with an absolutely continuous spectral measure and the positive semidefinite *spectral density* matrix $\boldsymbol{f}_{\mathbf{X}}$.

Assume that $\boldsymbol{f}_{\mathbf{X}}(\omega)$ has *constant rank* r for a.e. $\omega \in [-\pi, \pi]$. If $\mathbf{X}_t$ is also regular (it always holds if $\boldsymbol{f}_{\mathbf{X}}$ is a rational spectral density matrix), then the multidimensional Wold decomposition is able to make it a one-sided

$VMA(\infty)$ process. It is important that the dimension of the *innovation subspaces* is also r.

With the integer $1 \le k \le r$, the *k-factor GDFM*:

$$\mathbf{X}_t = \boldsymbol{\chi}_t + \mathbf{e}_t, \quad t = 1, 2, \dots$$

where now $\boldsymbol{\chi}_t$ denotes the common component, $\mathbf{e}_t$ is the *idiosyncratic noise*, and all the expectations are zeros, for simplicity. Here $\boldsymbol{\chi}_t$ is subordinated to $\mathbf{X}_t$, but has spectral density matrix of rank $k \le r$. For example, there are k uncorrelated signals (given by k distinct sources) detected by r sensors. Opposed to the static factors, this is not a low-rank approximation of the (zero-lag) auto-covariance matrix that provides the *static factors.*

Forni, Lippi, and Deistler et al. gave necessary and sufficient conditions for the existence of an underlying GDFM in terms of the expanding sequence of $n \times n$ spectral density matrices $\boldsymbol{f}^n_{\mathbf{X}}(\omega)$, $n \in \mathbb{N}$.

Theorem 3. *The nested sequence* $\{\mathbf{X}^n_t : n \in \mathbb{N}, t = 1, 2, \dots\}$ *can be represented by a sequence of k-factor GDFMs if and only if*

- *the k largest eigenvalues,* $\lambda^n_{\mathbf{X},1}(\omega) \ge \cdots \ge \lambda^n_{\mathbf{X},k}(\omega)$ *of* $\boldsymbol{f}^n_{\mathbf{X}}(\omega)$ *diverge almost everywhere in* $[-\pi, \pi]$ *as* $n \to \infty$*;*
- *the* $(k{+}1)$*-th largest eigenvalue* $\lambda^n_{\mathbf{X},k+1}(\omega)$ *of* $\boldsymbol{f}^n_{\mathbf{X}}(\omega)$ *is uniformly bounded for* $\omega \in [-\pi, \pi]$ *(almost everywhere) and for all* $n \in \mathbb{N}$.

The theorem is rather theoretical; its message is that for large n and T (T is not necessarily larger than n) we can conclude for k from the spectral gap of the constant rank spectral density matrix. The estimate $\boldsymbol{\chi}^n_t$ is consistent if $n, T \to \infty$. The idiosyncratic noise is less and less important when $n, T \to \infty$, and it may have slightly correlated components. Also, the largest eigenvalue of $\boldsymbol{f}^n_{\mathbf{e}}(\omega)$ is uniformly bounded for $\omega \in [-\pi, \pi]$ and for all $n \in \mathbb{N}$. As we learned in the preceding lessons, a stationary process with a not full rank spectral density matrix may have some singular components. All these parts are included in the weakly dependent idiosyncratic noise.

Dynamic factor analysis is an unsupervised learning method, and with the lag-dependent factor loading matrices we are able to give meaning to the dynamic factors that embody the comovements between the components at different lags. For example, when we use a parametric method, we are also able to give predictions for the dynamic factors (via autoregression) and, in turn, for the components of the time series too. There are also state space models that are able to estimate the parameter matrices via singular autoregression. The reduced rank approximation in Section 5.3 offers a first step, and the Yule–Walker equations can be solved for the reduced rank process.

6 Summary

First we have a 1D real valued time series $\{X_t\}$ which is not necessarily stationary, $\mathbb{E}(X_t) = 0$ $(t \in \mathbb{Z})$. Selecting a starting observation X_1 and with the notation $H_n = \mathrm{Span}\{X_1, \dots, X_n\}$, X_{n+1} is predicted linearly based on random past values $X_1, \dots, X_n$ such that $\hat{X}_1 := 0$ and $\mathbb{E}\eta_{n+1}^2 = \mathbb{E}(X_{n+1} - \hat{X}_{n+1})^2$ is minimized, $n = 1, 2, \dots$. By the general theory of Hilbert spaces, $\hat{X}_{n+1}$ is the projection of X_{n+1} onto the linear subspace H_n. The coefficients of the optimal linear predictor

$$\hat{X}_{n+1} = a_{n1}X_n + \cdots + a_{nn}X_1$$

can be obtained by solving the system of linear equations

$$\boldsymbol{C}_n \mathbf{a}_n = \mathbf{d}_n,$$

where $\mathbf{a}_n = (a_{n1}, \dots, a_{nn})^T$, $\boldsymbol{C}_n = [\mathrm{Cov}(X_i, X_j)]_{i,j=1}^n$ and $\mathbf{d}_n = (\mathrm{Cov}(X_{n+1}, X_n), \dots, \mathrm{Cov}(X_{n+1}, X_1))^T$. A solution (the projection) always exists, and it is unique if $\boldsymbol{C}_n$ is positive definite; then the unique solution is $\mathbf{a}_n = \boldsymbol{C}_n^{-1}\mathbf{d}_n$, otherwise, the generalized inverse of $\boldsymbol{C}_n$ comes into existence. However, in case of stationary processes, this is not an issue. The h-step ahead prediction is obtained from $\boldsymbol{C}_n\mathbf{a}_n = \mathbf{d}_n(h)$, where $\mathbf{d}_n(h) = (c(h), \dots, c(n+h-1))^T$ in the stationary case.

As for the innovation $\boldsymbol{\eta}_n = (\eta_1, \dots, \eta_n)^T$, we have to find an $n \times n$ lower triangular matrix $\boldsymbol{L}_n$ such that $\mathbf{X}_n = \boldsymbol{L}_n\boldsymbol{\eta}_n$. Taking the covariance matrices on both sides, yields $\boldsymbol{C}_n = \boldsymbol{L}_n\boldsymbol{D}_n\boldsymbol{L}_n^T$. In this way, the LDL decomposition (a variant of the Cholesky decomposition) gives the prediction errors (diagonal entries of $\boldsymbol{D}_n$), and the entries of $\boldsymbol{L}_n$ below its main diagonal (the main diagonal is constantly 1). The situation further simplifies in the stationary case, when $\boldsymbol{C}_n$ is a Toeplitz matrix. However, $\boldsymbol{L}_n$ will not be Toeplitz, but asymptotically, it becomes more and more like a Toeplitz one, and the entries of $\boldsymbol{D}_n$ will be more and more similar to each other, i.e. to the limit $\sigma^2 = \lim_{n\to\infty} e_n^2$.

In particular, if $\{X_t\}$ is stationary, then $\boldsymbol{C}_n = [c(i-j)]_{i,j=1}^n$, so $\boldsymbol{C}_n$ is a Toeplitz matrix, and $d_n(j) = c(j)$, $j = 1, \dots, n$. Therefore, no double indexing is necessary, but $\mathbf{a}_n = (a_1, \dots, a_n)^T$. With it, the defining equation is exactly the same as the first n Yule–Walker equations for estimating the parameters of a stationary AR(n) process. The prediction error is $e_n^2 = \mathrm{Var}(\eta_{n+1})$. It can be written in many equivalent forms, e.g.

$$e_n^2 = c(0)(1 - r^2_{X_t,(X_{t-1},\dots,X_{t-n})}) = c(0) - \mathbf{d}_n^T \boldsymbol{C}_n^{-1} \mathbf{d}_n,$$

where $r^2_{X_t,(X_{t-1},\dots,X_{t-n})}$ is the squared multiple correlation coefficient between X_t and $(X_{t-1}, \dots, X_{t-n})$; it does not depend on t either, and obviously increases (does not decrease) with n, i.e. $e_1^2 \geq e_2^2 \geq \dots$. The mean square error can as well be written with the determinants of the consecutive Toeplitz matrices $\boldsymbol{C}_n$ and $\boldsymbol{C}_{n+1}$. If for some n, $|\boldsymbol{C}_n| \neq 0$, then

$$e_n^2 = c(0) - \mathbf{d}_n^T \boldsymbol{C}_n^{-1} \mathbf{d}_n = \frac{|\boldsymbol{C}_{n+1}|}{|\boldsymbol{C}_n|}.$$

If $|\boldsymbol{C}_n| = 0$ for some n, then $|\boldsymbol{C}_{n+1}| = |\boldsymbol{C}_{n+2}| = \dots = 0$ too. The smallest index n for which this happens indicates that there is a linear relation between n consecutive X_js, but no linear relation between $n-1$ consecutive ones (by stationarity, this property is irrespective of the position of the consecutive random variables). This can happen only if some X_t linearly depends on $n-1$ preceding X_js. In this case $e_{n-1}^2 = 0$ and, of course $e_n^2 = e_{n+1}^2 = \dots = 0$ too. In any case, $e_1^2 \geq e_2^2 \geq \dots$ is a decreasing (non-increasing) nonnegative sequence, and in view of Equation (9),

$$|\boldsymbol{C}_1| = c(0), \quad |\boldsymbol{C}_n| = c(0) e_1^2 \dots e_{n-1}^2, \quad n = 2, 3, \dots,$$

so, provided $c(0) > 0$, $|\boldsymbol{C}_n| = 0$ holds if and only if $e_{n-1}^2 = 0$. Note that in this stationary case there is no sense of using generalized inverse if $|\boldsymbol{C}_n| = 0$, since then exact one-step ahead prediction with the $n-1$ long past can be done with zero error, and this property is manifested for longer past predictions too. Note that the previous LDL decomposition also implies that $|\boldsymbol{C}_n| = |\boldsymbol{D}_n| = c(0) e_1^2 \dots e_{n-1}^2$, $n = 2, 3, \dots$.

In case of a stationary, non-singular process, we can project X_{n+1} onto the infinite past $H_n^- = \overline{\mathrm{span}}\,\{X_t : t \leq n\}$ and expand it in terms of an orthonormal system, see the Wold decomposition. This part will be the regular (causal) part of the process, whereas, the other, singular part, is orthogonal to it. Also, by stationarity, the one-step ahead prediction error σ^2 does not depend on n, and it is positive, since the process is non-singular.

Then we have a d-dimensional time series $\{\mathbf{X}_t\}$ with components $\mathbf{X}_t = (X_t^1, \dots, X_t^d)^T$, the state space is $\mathbb{R}^d$ and $\mathbb{E}(\mathbf{X}_t) = \mathbf{0}$. Select a starting observation $\mathbf{X}_1$ and

$$H_n := \operatorname{span}\{X_t^j : t = 1, \dots, n;\ j = 1, \dots, d\},$$

$\dim(H_n) = dn$. We want to linearly predict $\mathbf{X}_{n+1}$ based on random past values $\mathbf{X}_1, \dots, \mathbf{X}_n$. Analogously to the 1D situation, $\hat{\mathbf{X}}_1 := 0$ and $\hat{\mathbf{X}}_{n+1}$ is the best one-step ahead linear predictor that minimizes $\mathbb{E}(\mathbf{X}_{n+1} - \hat{\mathbf{X}}_{n+1})^2$. Now we solve the system of linear equations

$$\sum_{j=1}^{n} \boldsymbol{A}_{nj} \operatorname{Cov}(\mathbf{X}_{n+1-j}, \mathbf{X}_{n+1-k}) = \operatorname{Cov}(\mathbf{X}_{n+1}, \mathbf{X}_{n+1-k}). \quad k = 1, \dots, n.$$

When $\{\mathbf{X}_t\}$ is stationary, then it simplifies to

$$\sum_{j=1}^{n} \boldsymbol{A}_j \boldsymbol{C}(k-j) = \boldsymbol{C}(k). \quad k = 1, \dots, n,$$

This provides a system of $d^2 n$ linear equations with the same number of unknowns that always has a solution. Further, the solution does not depend on the selection of the time of the starting observation $\mathbf{X}_1$, and no double indexing of the coefficient matrices is necessary. If for some $n \geq 1$ the covariance matrix of $(\mathbf{X}_{n+1}^T, \dots, \mathbf{X}_1^T)^T$ is positive definite, then the matrix polynomial (VAR polynomial) $\boldsymbol{\alpha}(z) = \boldsymbol{I} - \boldsymbol{A}_1 z - \cdots - \boldsymbol{A}_n z^n$ is causal in the sense that $|\boldsymbol{\alpha}(z)| \neq 0$ for $z \leq 1$; otherwise, block matrix techniques and reduction in the innovation subspaces is needed.

$\mathbf{X}_t$ can again be expanded in terms of the now d-dimensional innovations, i.e. the prediction error terms $\boldsymbol{\eta}_{n+1} = \mathbf{X}_{n+1} - \hat{\mathbf{X}}_{n+1}$. In this way, we get the innovations $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n$ that trivially have $\mathbf{0}$ expectation and form an orthogonal system in the nd-dimensional H_n. Actually, we have the recursive equations

$$\mathbf{X}_k = \sum_{j=1}^{k-1} \boldsymbol{B}_{kj} \boldsymbol{\eta}_j + \boldsymbol{\eta}_k, \quad k = 1, 2, \dots, n.$$

Here the covariance matrix $\boldsymbol{E}_j = \mathbb{E}\boldsymbol{\eta}_j \boldsymbol{\eta}_j^T$ is a positive semidefinite matrix, but can be of reduced rank. At a passage to infinity, we obtain the multidimensional Wold decomposition. At the end, we have to perform the block

Cholesky (LDL) decomposition:

$$\mathfrak{C}_n = \boldsymbol{L}_n \boldsymbol{D}_n \boldsymbol{L}_n^T,$$

where $\mathfrak{C}_n$ is $nd \times nd$ positive definite block Toeplitz matrix, $\boldsymbol{D}_n$ is $nm \times nm$ block diagonal and contains the positive semidefinite prediction error matrices $\boldsymbol{E}_1, \dots, \boldsymbol{E}_n$ in its diagonal blocks, whereas $\boldsymbol{L}_n$ is $nd \times nd$ lower triangular with blocks $\boldsymbol{B}_{kj}$s below its diagonal blocks which are $d \times d$ identities. In view of this,

$$|\mathfrak{C}_n| = |\boldsymbol{D}_n| = \prod_{j=1}^{n} |\boldsymbol{E}_j|,$$

analogously to the 1D situation. We also prove that if the entries of the autocovariance matrices are absolutely summable, then the eigenvalues of $\mathfrak{C}_n$ asymptotically comprise the union of the spectra of the spectral density matrices at the n Fourier frequencies as $n \to \infty$.

When the $\boldsymbol{E}_j$s are of reduced rank, we can find a system $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n \in \mathbb{R}^r$ in the d-dimensional innovation subspaces that span the same subspace as $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n$. If the the spectral density matrix has $r < d$ structural eigenvalues in a General Dynamic Factor Model, then $\boldsymbol{\xi}_j \in \mathbb{R}^r$ is the principal component factor of $\boldsymbol{\eta}_j$ obtained from an r-factor model.

Note that here we use $d \times d$ block matrices in the calculations, so the computational complexity of the procedure is not significantly larger than that of the subsequent Kálmán's filtering for which we use the notation of R. E. Kálmán's original paper [?], where stationarity is not assumed, but the random vectors are Gaussian. The linear dynamical system is

$$\begin{aligned} \mathbf{X}_{t+1} &= \boldsymbol{A}_t \mathbf{X}_t + \mathbf{U}_t \\ \mathbf{Y}_t &= \boldsymbol{C}_t \mathbf{X}_t, \end{aligned}$$

where $\boldsymbol{A}_t$ and $\boldsymbol{C}_t$ are specified matrices; $\boldsymbol{A}_t$ is an $n \times n$ matrix, called phase transition matrix, and $\boldsymbol{C}_t$ is $p \times n$; further, $\mathbf{U}_t$ (random excitation) is an orthogonal noise process with $\mathbb{E}\mathbf{U}_t \mathbf{U}_s^T = \delta_{st} \boldsymbol{Q}_{\mathbf{U}}(t)$ and $\mathbb{E}\mathbf{X}_s^T \mathbf{U}_t = \mathbf{0}$ for $s \le t$. All the expectations are zeros, and all the random vectors have real components. $\mathbf{X}_t$ is the n-dimensional hidden state variable, while $\mathbf{Y}_t$ is the p-dimensional observable variable. In the paper [?], $p \le n$ is assumed, but it is not a restriction. Even if $p = n$, the matrix $\boldsymbol{C}_t$ is not invertible, otherwise the process $\mathbf{X}_t$ is trivially observable, unless a noise term is added to $\boldsymbol{C}_t \mathbf{X}_t$ in the second equation.

The problem is the following: starting the observations at time 0, given $\mathbf{Y}_0, \dots, \mathbf{Y}_{t-1}$, we want to estimate $\mathbf{X}$ component-wise, with minimum mean square error. If $\mathbf{X} = \mathbf{X}_t$, this is the prediction problem and we denote the optimal one-step ahead prediction of $\mathbf{X}_t$ by $\widehat{\mathbf{X}}_{t|t-1}$. If $\mathbf{Y}_0, \dots, \mathbf{Y}_{t-1}$ is observed and $\widehat{\mathbf{X}}_{t|t-1}$ is already known, then we give a recursion to find $\widehat{\mathbf{X}}_{t+1|t}$ by using the new value of $\mathbf{Y}_t$:

$$\widehat{\mathbf{X}}_{t+1|t} = \boldsymbol{A}_t \widehat{\mathbf{X}}_{t|t-1} + \boldsymbol{K}_t(\mathbf{Y}_t - \boldsymbol{C}_t \widehat{\mathbf{X}}_{t|t-1}),$$

where $\boldsymbol{K}_t$ is the Kálmán gain matrix:

$$\boldsymbol{K}_t = \boldsymbol{A}_t \boldsymbol{P}(t) \boldsymbol{C}_t^T [\boldsymbol{C}_t \boldsymbol{P}(t) \boldsymbol{C}_t^T]^{-}.$$

Here

$$\boldsymbol{P}(t) = \mathbb{E}\widetilde{\mathbf{X}}_{t|t-1}\widetilde{\mathbf{X}}_{t|t-1}^T$$

is the error covariance matrix that drives the process. For it, the recursion

$$\boldsymbol{P}(t+1) = \boldsymbol{A}_t \boldsymbol{P}(t) \boldsymbol{A}_t{}^T - \boldsymbol{A}_t \boldsymbol{P}(t) \boldsymbol{C}_t^T [\boldsymbol{C}_t \boldsymbol{P}(t) \boldsymbol{C}_t^T]^{-} \boldsymbol{C}_t \boldsymbol{P}(t) \boldsymbol{A}_t{}^T + \boldsymbol{Q}_{\mathbf{U}}(t)$$

holds, which makes rise to an iteration. The above equation results in a matrix Riccati equation for $\boldsymbol{P} = \boldsymbol{P}(t) = \boldsymbol{P}(t+1)$ if the process is stationary.

With the integer $1 \le k < r$, the dynamic k-factor mode (GDFM) is:

$$\mathbf{X}_t = \boldsymbol{\chi}_t + \mathbf{e}_t, \quad t = 1, 2, \dots$$

where now $\boldsymbol{\chi}_t$ denotes the common component, $\mathbf{e}_t$ is the n-dimensional idiosyncratic noise, and all the expectations are zeros, for simplicity. Here $\boldsymbol{\chi}_t$ is subordinated to $\mathbf{X}_t$, but has spectral density matrix of rank $k < r$. For example, there are k uncorrelated signals (given by k distinct sources), detected by d sensors. Opposed to the static factors, this is not a low-rank approximation of the (zero-lag) auto-covariance matrix that provides the static factors.

Forni and Lippi [**?**] and Deistler et al. [**?**, **?**] gave necessary and sufficient conditions for the existence of an underlying GDFM in terms of the observable $n \times n$ spectral densities $\boldsymbol{f}_{\mathbf{X}}^n(\omega)$, $n \in \mathbb{N}$. The nested sequence $\{\mathbf{X}_t^n : n \in \mathbb{N}, t = 1, 2, \dots\}$ can be represented by a sequence of GDFMs if and only if

- the first k eigenvalues, $\lambda_{\mathbf{X},1}^n(\omega) \ge \dots \ge \lambda_{\mathbf{X},k}^n(\omega)$ (in non-increasing order), of $\boldsymbol{f}_{\mathbf{X}}^n(\omega)$ diverge almost everywhere in $[-\pi, \pi]$ as $n \to \infty$;

- the $(k+1)$-th eigenvalue $\lambda^n_{\mathbf{X},k+1}(\omega)$ of $\boldsymbol{f}^n_{\mathbf{X}}(\omega)$ is uniformly bounded for $\omega \in [-\pi, \pi]$ almost everywhere and for all $n \in \mathbb{N}$.

So we can conclude for k from the spectral gap. The estimate $\boldsymbol{\chi}^n_t$ is consistent if $n, T \to \infty$. The idiosyncratic noise is less and less important when n, T get larger and larger, and it may have slightly correlated components.