Basics: Best prediction in Hilbert spaces

Marianna Bolla, BME Math. Inst.

November 4, 2020

Let $L^2(\Omega, \mathcal{A}, \mathbb{P})$ be the Hilbert space of real valued random variables with zero expectation and finite variance; the inner product is the covariance, and $\mathbb{P}$ denotes the joint distribution of all of them. We use subspaces of this, related to multivariate, weakly stationary processes.

Let $\mathbf{X} = \{\mathbf{X}_t\}_{t\in\mathbb{Z}}$ be a weakly stationary, d-dimensional time series, with real-valued coordinates, $\mathbb{E}\mathbf{X}_t = \mathbf{0}$. By weak stationarity, $\mathbf{X}_t$s all have the same covariance matrix $\boldsymbol{C}(0)$. Note that to any weakly stationary process there corresponds a Gaussian one with the same second moments, so it is not a restriction to confine ourselves to Gaussian processes. Sometimes we speak in terms of so-called *second order processes* that are determined by their first and second moments. When the expectations are 0s, the pairwise covariances characterize the process, and predictions can be discussed in terms of projections in Hilbert spaces.

So in the Gaussian case, $\mathbf{X}_t$s all have the same d-variate Gaussian distribution, but they are defined on different d-dimensional marginals of $\mathbb{P}$. In particular, their first, second, etc. autocovariances, $\boldsymbol{C}(1), \boldsymbol{C}(2), \ldots$ characterize their joint distribution. Therefore, this can be regarded as a special random field that extends in space and time (cross-sectionally and longitudinally), i.e., the parameters of the random process contain both (discrete) time and (d-dimensional) space locations.

Corresponding to the above weakly stationary process, throughout the book we consider the following subspaces of $L^2(\Omega, \mathcal{A}, \mathbb{P})$:

$$
\begin{aligned}
H(\mathbf{X}) &= \overline{\text{span}}\{X_k^i \,|\, k \in \mathbb{Z},\, i = 1, \ldots, d\} \\
H_t^-(\mathbf{X}) &= \overline{\text{span}}\{X_k^i \,|\, k \le t,\, i = 1, \ldots, d\} \\
H_t(\mathbf{X}) &= \overline{\text{span}}\{X_k^i \,|\, 1 \le k \le t,\, i = 1, \ldots, d\} \\
&= \text{Span}\,\{X_k^i \,:\, 1 \le k \le t,\, i = 1, \ldots, d\},
\end{aligned}
$$

as the last one is finite dimensional. Obviously, $H_t(\mathbf{X}) \subseteq H_t^-(\mathbf{X}) \subseteq H(\mathbf{X})$.

In any Hilbert space, the following Projection Theorem holds true.

Theorem 1. *Let $\mathcal{M}$ be a closed subspace of the Hilbert space $\mathcal{H}$. Then for any $Y \in \mathcal{H}$, there is a unique element $\hat{Y} \in \mathcal{M}$ such that*

$$\|Y - \hat{Y}\| \leq \|Y - Z\|, \quad \forall Z \in \mathcal{M} \tag{1}$$

and

$$Y - \hat{Y} \perp Z, \quad \forall Z \in \mathcal{M}.$$

This unique $\hat{Y}$ is called the projection of Y onto $\mathcal{M}$ and denoted by $\mathrm{Proj}_{\mathcal{M}} Y$. By the Pythagorean Theorem,

$$\|Y\|^2 = \|\hat{Y}\|^2 + \|Y - \hat{Y}\|^2.$$

The Projection Theorem is widely used in statistics, in the context of multivariate parallel, in other words, simultaneous or joint response regressions. We distinguish between nonparametric and parametric regression, where the former applies to arbitrary multivariate distributions, whereas, the latter to Gaussian ones. More generally, we speak in terms of random vectors which are not necessarily instances of a multidimensional time series. Let $\mathbf{X}$ be q- and $\mathbf{Y}$ be a p-dimensional random vector. In the Gaussian, zero mean case, their joint distribution is characterized by their $q \times p$ cross-covariance matrix $\mathbb{E}\mathbf{X}\mathbf{Y}^T$, and the orthogonality (independence) of them means that each component of $\mathbf{X}$ is uncorrelated with each component of $\mathbf{Y}$, i.e., $\mathbb{E}\mathbf{X}\mathbf{Y}^T = \boldsymbol{O}$. Note that $\mathbb{E}\mathbf{Y}\mathbf{X}^T = [\mathbb{E}\mathbf{X}\mathbf{Y}^T]^T$, whereas $\mathbb{E}\mathbf{X}\mathbf{X}^T$ is the usual (positive semidefinite) covariance matrix of $\mathbf{X}$.

Consider the multiple response regression problem, when each component of $\mathbf{Y}$ is projected onto the subspace generated by the coordinates of $\mathbf{X}$. First we deal with the general (nonparamertic) situation.

Lemma 1. *Let $\mathbf{Y}$ be an $\mathbb{R}^p$-valued and $\mathbf{X}$ be an $\mathbb{R}^q$-valued random vector with components of zero expectations. Then for every $f : \mathbb{R}^q \to \mathbb{R}^p$ measurable function for which the expectation below exists and is finite, the conditional expectation of $\mathbf{Y}$ conditioned on $\mathbf{X}$ minimizes the error covariance matrix*

$$\mathbb{E}(\mathbf{Y} - f(\mathbf{X}))(\mathbf{Y} - f(\mathbf{X}))^T \geq E(\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{X}))(\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{X}))^T,$$

in the sense that the difference of the left and right hand side matrices is positive semidefinite.

Proof. With the notation $f^*(\mathbf{X}) = \mathbb{E}(\mathbf{Y}|\mathbf{X})$ we have that

$$\begin{aligned}
&\mathbb{E}(\mathbf{Y} - f(\mathbf{X}))(\mathbf{Y} - f(\mathbf{X}))^T \\
&= \mathbb{E}[(\mathbf{Y} - f^*(\mathbf{X})) + (f^*(\mathbf{X}) - f(\mathbf{X}))][(\mathbf{Y} - f^*(\mathbf{X})) + (f^*(\mathbf{X}) - f(\mathbf{X}))]^T \\
&= \mathbb{E}(\mathbf{Y} - f^*(\mathbf{X}))(\mathbf{Y} - f^*(\mathbf{X}))^T + \mathbb{E}(f^*(\mathbf{X}) - f(\mathbf{X}))(f^*(\mathbf{X}) - f(\mathbf{X}))^T \\
&+ \mathbb{E}(\mathbf{Y} - f^*(\mathbf{X}))(f^*(\mathbf{X}) - f(\mathbf{X}))^T + \mathbb{E}(f^*(\mathbf{X}) - f(\mathbf{X}))(\mathbf{Y} - f^*(\mathbf{X}))^T \\
&= \mathbb{E}(\mathbf{Y} - f^*(\mathbf{X}))(\mathbf{Y} - f^*(\mathbf{X}))^T + \mathbb{E}(f^*(\mathbf{X}) - f(\mathbf{X}))(f^*(\mathbf{X}) - f(\mathbf{X}))^T.
\end{aligned}$$

In the last step we used that $\mathbb{E}(\mathbf{Y} - f^*(\mathbf{X}))(f^*(\mathbf{X}) - f(\mathbf{X}))^T = \boldsymbol{O}$ is the zero matrix, akin to its transpose $\mathbb{E}(f^*(\mathbf{X}) - f(\mathbf{X}))(\mathbf{Y} - f^*(\mathbf{X}))^T$. So it suffices to prove only for the first one:

$$\begin{aligned}
\mathbb{E}(\mathbf{Y} - f^*(\mathbf{X}))(f^*(\mathbf{X}) - f(\mathbf{X}))^T &= \mathbb{E}[\mathbb{E}(\mathbf{Y} - f^*(\mathbf{X}))(f^*(\mathbf{X}) - f(\mathbf{X}))^T \,|\, \mathbf{X}] \\
&= \mathbb{E}[\mathbb{E}(\mathbf{Y} - f^*(\mathbf{X}) \,|\, \mathbf{X})(f^*(\mathbf{X}) - f(\mathbf{X}))^T] \\
&= \mathbb{E}[\mathbb{E}(\mathbf{Y}|\mathbf{X}) - f^*(\mathbf{X}))(f^*(\mathbf{X}) - f(\mathbf{X}))^T] \\
&= \boldsymbol{O}
\end{aligned}$$

since $\mathbb{E}(\mathbf{Y}|\mathbf{X}) - f^*(\mathbf{X}) = \mathbf{0}$. As $\mathbb{E}(f^*(\mathbf{X}) - f(\mathbf{X}))(f^*(\mathbf{X}) - f(\mathbf{X}))^T$ is positive semidefinite (being a covariance matrix), it follows that

$$\mathbb{E}(\mathbf{Y} - f(\mathbf{X}))(\mathbf{Y} - f(\mathbf{X}))^T - \mathbb{E}(\mathbf{Y} - f^*(\mathbf{X}))(\mathbf{Y} - f^*(\mathbf{X}))^T$$

is positive semidefinite that was to be proved. □

Remark 1. *The conditional expectation $f^*(\mathbf{X}) = \mathbb{E}(\mathbf{Y}|\mathbf{X})$ also minimizes*

$$\mathbb{E}\|\mathbf{Y} - f(\mathbf{X})\|^2 = \sum_{i=1}^{p} \mathbb{E}[Y^i - f_i(\mathbf{X})]^2,$$

where f_is are the coordinate functions of f. Indeed, we can minimize the p terms separately. Applying Lemma 1 to the univariate case, we get that the minimizer of the ith term is $f_i^*(\mathbf{X}) = \mathbb{E}(Y^i|\mathbf{X})$. As this is the ith coordinate of $f^*(\mathbf{X}) = \mathbb{E}(\mathbf{Y}|\mathbf{X})$, the minimum of $\mathbb{E}\|\mathbf{Y} - f(\mathbf{X})\|^2$ is attained at the same $f^*(\mathbf{X})$ that minimizes the covariance matrix $\mathbb{E}(\mathbf{Y} - f(\mathbf{X}))(\mathbf{Y} - f(\mathbf{X}))^T$ in the sense of Lemma 1. Note that $\mathbb{E}\|\mathbf{Y} - f(\mathbf{X})\|^2$ is the trace of the error covariance matrix.

Therefore, it suffices to investigate univariate nonparametric regression estimations. If they are mean square consistent, can be constructed with a sequence $f_i^{(n)}$ (for example, with local averaging) such that the mean square error

$$\mathbb{E}[f_i^{(n)}(\mathbf{X}) - f_i^*(\mathbf{X})]^2 \to 0, \quad n \to \infty,$$

for $i = 1, \ldots, p$. This implies that in the p-variate case, for the sequence $f^{(n)}(\mathbf{X}) = (f_1^{(n)}(\mathbf{X}), \ldots, f_p^{(n)}(\mathbf{X}))^T$

$$\mathbb{E}\|f^{(n)}(\mathbf{X}) - f^*(\mathbf{X})\|^2 \to 0, \quad n \to \infty \tag{2}$$

holds, exhibiting a kind of mean square consistency in the multiple target situation. Since $\mathbb{E}\|f^{(n)}(\mathbf{X}) - f^*(\mathbf{X})\|^2$ is the trace of the $p \times p$ symmetric, positive semidefinite error covariance matrix $\boldsymbol{E}_n = \mathbb{E}(f^{(n)}(\mathbf{X}) - f^*(\mathbf{X}))(f^{(n)}(\mathbf{X}) - f^*(\mathbf{X}))^T$, Equation (2) is equivalent to $\|\boldsymbol{E}_n\|_2 \to 0$ as $n \to \infty$ (the spectral norm $\|\boldsymbol{E}_n\|_2$ is the largest eigenvalue of $\boldsymbol{E}_n$). Conversely, if $\|\boldsymbol{E}_n\|_2 \to 0$, then $\mathrm{tr}\boldsymbol{E}_n \to 0$, and by the Cauchy–Schwarz inequality, $\boldsymbol{E}_n \to \boldsymbol{O}$ too.

Now we concentrate on linear estimates that are the best in the above sense too if the underlying distribution is multivariate Gaussian. Therefore, the forthcoming estimation can be called parametric simultaneous (joint response) regression, and can be described by matrices. Here we use the second order property of $\mathbb{P}$: the pairwise inner products of the random variables in $L^2(\Omega, \mathcal{A}, \mathbb{P})$ are determined by their covariances. If we consider subspaces, then the relation of a p- and a q-dimensional subspace can be described by all possible pq pairs of the pairwise covariances, i.e., by the cross-covariance matrices.

Lemma 2. *Let $\mathbf{Y} \in \mathbb{R}^p$ and $\mathbf{X} \in \mathbb{R}^q$ be random vectors on a joint probability space with existing second moments and zero expectation. Then the $q \times p$ matrix $\boldsymbol{A}$ minimizing $\mathbb{E}\|\mathbf{Y} - \boldsymbol{A}^T\mathbf{X}\|^2$ is*

$$\boldsymbol{A} = [\mathbb{E}\mathbf{X}\mathbf{X}^T]^-[\mathbb{E}\mathbf{X}\mathbf{Y}^T], \tag{3}$$

where we use generalized inverse $^-$ if the covariance matrix $\mathbb{E}\mathbf{X}\mathbf{X}^T$ of $\mathbf{X}$ is singular (see Appendix B). If it is positive definite, then we get a unique minimizer $\boldsymbol{A}$ with the unique inverse matrix $[\mathbb{E}\mathbf{X}\mathbf{X}^T]^{-1}$.

Proof. Observe that minimizing

$$\mathbb{E}\|\mathbf{Y} - \boldsymbol{A}^T\mathbf{X}\|^2 = \sum_{i=1}^{p} \mathbb{E}(Y^i - \mathbf{a}_i^T\mathbf{X})^2$$

with respect to $\boldsymbol{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_p]$ falls apart into the following p minimization tasks, with respect to the q-dimensional column vectors of $\boldsymbol{A}$:

$$\min_{\mathbf{a}_i} \mathbb{E}(Y^i - \mathbf{a}_i^T \mathbf{X})^2, \quad i = 1, \ldots, p.$$

The ith task, with the coordinates of $\mathbf{X} = (X^1, \ldots, X^q)^T$ and $\mathbf{a}_i^T = (a_{1i}, \ldots, a_{qi})$, is equivalent to

$$\mathbb{E}(Y^i - \sum_{k=1}^{q} a_{ki} X^k)^2 \to \min .$$

Take the derivative with respect to a_{ji} and make it equal to 0. (We assume regularity, i.e. that the differentiation and taking the expectation can be interchanged, which is true if the underlying distribution is Gaussian.)

$$2\mathbb{E}[(-X_j)(Y^i - \sum_{k=1}^{q} a_{ki} X^k)] = 0, \quad j = 1, \ldots, q.$$

After rearranging, we have the system of equations

$$\sum_{k=1}^{q} a_{ki} \mathbb{E}(X^j X^k) = \mathbb{E}(X^j Y^i), \quad j = 1, \ldots, q.$$

This can be condensed into the well-known system of *Gauss normal equations* from the classical theory of multivariate regression:

$$[\mathbb{E}\mathbf{X}\mathbf{X}^T]\mathbf{a}_i = [\mathbb{E}\mathbf{X}Y^i], \quad i = 1, \ldots, p.$$

(Actually, the original equations of Gauss apply to the sample version, and do not contain expectations.) Since this system of linear equations is consistent (the vector $\mathbb{E}\mathbf{X}Y^i$ is in the column space of $\mathbb{E}\mathbf{X}\mathbf{X}^T$), it always has a solution in the general form:

$$\mathbf{a}_i = [\mathbb{E}\mathbf{X}\mathbf{X}^T]^-[\mathbb{E}\mathbf{X}Y^i], \quad i = 1, \ldots, p.$$

Here $[\mathbb{E}\mathbf{X}\mathbf{X}^T]^-$ is the generalized inverse of the matrix in brackets.

Therefore the matrix $\boldsymbol{A}$ giving the optimum is

$$\boldsymbol{A} = [\mathbb{E}\mathbf{X}\mathbf{X}^T]^-[\mathbb{E}\mathbf{X}\mathbf{Y}^T],$$

that is unique only if $\mathbb{E}\mathbf{X}\mathbf{X}^T$ is invertible (positive definite), otherwise (if $\mathbb{E}\mathbf{X}\mathbf{X}^T$ is singular, positive semidefinite) infinitely many versions of the generalized inverse give infinitely many convenient $\boldsymbol{A}$s (see Appendix B). Albeit, with different linear combinations of the coordinates of $\mathbf{X}_i$s, these always provide the same optimal linear prediction for $\mathbf{Y}$ as follows:

$$\widehat{\mathbf{Y}} = \begin{bmatrix} \widehat{Y}^1 \\ \widehat{Y}^2 \\ \vdots \\ \widehat{Y}^p \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^T\mathbf{X} \\ \mathbf{a}_2^T\mathbf{X} \\ \vdots \\ \mathbf{a}_p^T\mathbf{X} \end{bmatrix}.$$

Another easy proof uses the Projection Theorem 1 simultaneously as follows. We know that the $q\times p$ matrix $\boldsymbol{A}$, giving the minimum of $\mathbb{E}\|\mathbf{Y}-\boldsymbol{A}^T\mathbf{X}\|^2$, is such that $\boldsymbol{A}^T\mathbf{X} = \mathrm{Proj}_{\mathcal{M}}\mathbf{Y}$, where $\mathcal{M}$ is spanned by p-tuples of linear combinations of the coordinates of the vector $\mathbf{X} \in \mathbb{R}^q$, while $\mathbf{Y} \in \mathbb{R}^p$. But $\mathbf{Y} - \mathrm{Proj}_{\mathcal{M}}\mathbf{Y}$ is orthogonal to any vector in $\mathcal{M}$, which has the form $\boldsymbol{B}^T\mathbf{X}$ with a $q \times p$ matrix $\boldsymbol{B}$. Therefore,

$$\mathbb{E}[(\mathbf{Y} - \boldsymbol{A}^T\mathbf{X})(\boldsymbol{B}^T\mathbf{X})^T] = \boldsymbol{O}, \quad \forall \boldsymbol{B}_{q\times p}.$$

Equivalently,

$$[\mathbb{E}(\mathbf{Y}\mathbf{X}^T) - \boldsymbol{A}^T\mathbb{E}(\mathbf{X}\mathbf{X}^T)]\boldsymbol{B} = \boldsymbol{O}, \quad \forall \boldsymbol{B}_{q\times p}.$$

This implies that the matrix in brackets is the zero matrix, which fact after transposing and using that $\mathbb{E}(\mathbf{X}\mathbf{X}^T)$ is symmetric (it is the usual covariance matrix of $\mathbf{X}$) gives again the system of Gauss normal equations in concise form:

$$[\mathbb{E}(\mathbf{X}\mathbf{X}^T)]\boldsymbol{A} = \mathbb{E}(\mathbf{X}\mathbf{Y}^T).$$

□

Remark 2. *We can also estimate the attainable minimum error. When $p = 1$, then with the notations*

$$\boldsymbol{C} := \mathbb{E}\mathbf{X}\mathbf{X}^T \quad \textit{and} \quad \mathbf{d} := \mathbb{E}\mathbf{X}Y \quad (i = 1, \ldots, p)$$

by the theory of multivariate regression we have that

$$\mathbb{E}(Y - \widehat{Y})^2 = \mathrm{Var}(Y)(1 - r^2_{Y\mathbf{X}}) = \mathrm{Var}(Y) - \mathbf{d}^T\boldsymbol{C}^{-1}\mathbf{d},$$

where we assumed that $\boldsymbol{C}$ is positive definite and $r_{Y\mathbf{X}}$ denotes the multiple correlation between Y and the components of $\mathbf{X}$.

Adapting this for a p-dimensional $\mathbf{Y}$ we get that

$$\mathbb{E}\|\mathbf{Y}-\widehat{\mathbf{Y}}\|^2=\sum_{i=1}^{p}\mathbb{E}(Y^i-\widehat{Y}^i)^2=\|\mathbf{Y}\|^2-\sum_{i=1}^{p}\mathbf{d}_i^T\boldsymbol{C}^{-1}\mathbf{d}_i,$$

where $\mathbf{d}_i=\mathbb{E}\mathbf{X}Y^i$, $i=1,\dots,p$.

Lemma 3. *Let $\mathbf{Y}\in\mathbb{R}^p$ and $\mathbf{X}\in\mathbb{R}^q$ be random vectors on a joint probability space with existing second moments and zero expectation, and let $\mathrm{Proj}_{\mathcal{M}}\mathbf{Y}$ denote the best linear prediction of $\mathbf{Y}$ based on p-tuples of linear combinations of the coordinates of $\mathbf{X}$, denoted by $\mathcal{M}$, as in Lemma 2. Then with any $p\times p$ matrix $\boldsymbol{\Phi}$,*

$$\mathrm{Proj}_{\mathcal{M}}(\boldsymbol{\Phi}\mathbf{Y})=\boldsymbol{\Phi}\mathrm{Proj}_{\mathcal{M}}\mathbf{Y}.$$

Proof. We saw that $\mathrm{Proj}_{\mathcal{M}}\mathbf{Y}=\boldsymbol{A}^T\mathbf{X}$, where by (3) $\boldsymbol{A}=[\mathbb{E}\mathbf{X}\mathbf{X}^T]^-[\mathbb{E}\mathbf{X}\mathbf{Y}^T]$, and we use generalized inverse $^-$ if the covariance matrix $\mathbb{E}\mathbf{X}\mathbf{X}^T$ of $\mathbf{X}$ is singular. Then

$$\begin{aligned}\mathrm{Proj}_{\mathcal{M}}(\boldsymbol{\Phi}\mathbf{Y})&=\{[\mathbb{E}\mathbf{X}\mathbf{X}^T]^-[\mathbb{E}\mathbf{X}(\boldsymbol{\Phi}\mathbf{Y})^T]\}^T\mathbf{X}=[\mathbb{E}(\boldsymbol{\Phi}\mathbf{Y}\mathbf{X}^T)][\mathbb{E}\mathbf{X}\mathbf{X}^T]^-\mathbf{X}\\&=\boldsymbol{\Phi}[\mathbb{E}(\mathbf{Y}\mathbf{X}^T)][\mathbb{E}\mathbf{X}\mathbf{X}^T]^-\mathbf{X}=\boldsymbol{\Phi}\mathrm{Proj}_{\mathcal{M}}\mathbf{Y}.\end{aligned}$$

□

The above lemma shows that this projection is linear in $\mathbf{Y}$ and it commutes with $\boldsymbol{\Phi}$. In the Gaussian case, obviously, we have that

$$\mathrm{Proj}_{\mathcal{M}}(\boldsymbol{\Phi}\mathbf{Y})=\mathbb{E}(\boldsymbol{\Phi}\mathbf{Y}\,|\,\mathbf{X})=\boldsymbol{\Phi}\mathbb{E}(\mathbf{Y}\,|\,\mathbf{X})=\boldsymbol{\Phi}\mathrm{Proj}_{\mathcal{M}}(\mathbf{Y})$$

by the properties of the conditional expectation.

Now go back to time series. In particular, if we look for the best linear prediction of the p-dimensional random vector $\mathbf{Y}$ based on the segment $\mathbf{X}_1,\dots,\mathbf{X}_t$ of a d-dimensional time series, then with the $q=dt$ dimensional vector $\mathbf{X}=[\mathbf{X}_1^T,\dots,\mathbf{X}_t^T]^T$, the above formula adapts as

$$\widehat{\mathbf{Y}}=\mathrm{Proj}_{H_t(\mathbf{X})}\mathbf{Y}=\boldsymbol{A}^T\mathbf{X}=\begin{bmatrix}\mathbf{a}_{11}^T&\dots&\mathbf{a}_{1t}^T\\\mathbf{a}_{21}^T&\dots&\mathbf{a}_{2t}^T\\\vdots&\vdots&\vdots\\\mathbf{a}_{p1}^T&\dots&\mathbf{a}_{pt}^T\end{bmatrix}\begin{bmatrix}\mathbf{X}_1\\\mathbf{X}_2\\\vdots\\\mathbf{X}_t\end{bmatrix}=\begin{bmatrix}\mathbf{a}_{11}^T\mathbf{X}_1+\dots+\mathbf{a}_{1t}^T\mathbf{X}_t\\\mathbf{a}_{21}^T\mathbf{X}_1+\dots+\mathbf{a}_{2t}^T\mathbf{X}_t\\\vdots\\\mathbf{a}_{p1}^T\mathbf{X}_1+\dots+\mathbf{a}_{pt}^T\mathbf{X}_t\end{bmatrix},$$

where the columns of the $q \times p$ matrix $\boldsymbol{A}$ are partitioned into t segments of length d; or equivalently, the columns of the $p \times q$ matrix $\boldsymbol{A}^T$ are partitioned into $p \times d$ matrices $\boldsymbol{A}_1^T, \dots, \boldsymbol{A}_t^T$ like

$$\boldsymbol{A} = \begin{bmatrix} \mathbf{a}_{11} & \dots & \mathbf{a}_{p1} \\ \mathbf{a}_{12} & \dots & \mathbf{a}_{p2} \\ \vdots & \vdots & \vdots \\ \mathbf{a}_{1t} & \dots & \mathbf{a}_{pt} \end{bmatrix}, \quad \boldsymbol{A}^T = \begin{bmatrix} \boldsymbol{A}_1^T & \dots & \boldsymbol{A}_t^T \end{bmatrix}, \quad \boldsymbol{A}_j^T = \begin{bmatrix} \mathbf{a}_{1j}^T \\ \mathbf{a}_{2j}^T \\ \vdots \\ \mathbf{a}_{pj}^T \end{bmatrix} \ (j = 1, \dots, t).$$

With this, $\widehat{\mathbf{Y}}$ is the linear combination of $\mathbf{X}_1, \dots, \mathbf{X}_t$ with matrices $\boldsymbol{A}_1^T, \dots, \boldsymbol{A}_t^T$, i.e.

$$\widehat{\mathbf{Y}} = \boldsymbol{A}_1^T \mathbf{X}_1 + \dots + \boldsymbol{A}_t^T \mathbf{X}_t.$$

Remark 3. *Observe that the pdt-dimensional linear space generated by the linear combinations $\boldsymbol{A}_1^T \mathbf{X}_1 + \dots + \boldsymbol{A}_t^T \mathbf{X}_t$ of $\mathbf{X}_1, \dots, \mathbf{X}_t$ with $p \times d$ matrices $\boldsymbol{A}_1^T, \dots, \boldsymbol{A}_t^T$ is the p-tuple Cartesian product of $H_t(\mathbf{X})$, which is dt-dimensional and contains scalar linear combinations of all the d coordinates of $\mathbf{X}_1, \dots, \mathbf{X}_t$. So, in case of p simulteneous regressions, $\mathbf{X}_1, \dots, \mathbf{X}_t$ are linearly combined with $p \times d$ matrices, the rows of which give scalar linear combintions that define ithe individual regressions. Just the solution is organized in matrix form, which is more suitable for our purposes.*

So far, the time series was not necessarily stationary. When it is so, then the covariance matrix of the compounded vector $\mathbf{X} = [\mathbf{X}_1^T, \dots, \mathbf{X}_t^T]^T$ is

$$\mathbb{E}\mathbf{X}\mathbf{X}^T = \begin{bmatrix} \boldsymbol{C}(0) & \boldsymbol{C}(1) & \dots & \boldsymbol{C}_(t-1) \\ \boldsymbol{C}(-1) & \boldsymbol{C}(0) & \dots & \boldsymbol{C}(t-2) \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{C}(1-t) & \boldsymbol{C}(2-t) & \dots & \boldsymbol{C}(0) \end{bmatrix}$$

which is the symmetric (due to $\boldsymbol{C}(-k) = \boldsymbol{C}^T(k)$), positive semidefinite block Toeplitz matrix discussed in the context of VARMA processes. It is also positive definite whenever the process $\{\mathbf{X}_t\}$ is regular (see the Wold decomposition).

Going further, with $\mathbf{Y} = \mathbf{X}_{t+1}$, we get the one-step ahead prediction $\widehat{\mathbf{X}}_{t+1} = \boldsymbol{A}^T \mathbf{X}$, where the optimal $dt \times d$ matrix $\boldsymbol{A}$ is

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{C}(0) & \boldsymbol{C}(1) & \dots & \boldsymbol{C}_(t-1) \\ \boldsymbol{C}(-1) & \boldsymbol{C}(0) & \dots & \boldsymbol{C}(t-2) \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{C}(1-t) & \boldsymbol{C}(2-t) & \dots & \boldsymbol{C}(0) \end{bmatrix}^{-} \begin{bmatrix} \boldsymbol{C}(1) \\ \boldsymbol{C}(2) \\ \vdots \\ \boldsymbol{C}(t) \end{bmatrix}.$$

When the above block Toeplitz matrix is not singular, $\boldsymbol{A}$ is the unique solution of the system of equations

$$\begin{bmatrix} \boldsymbol{C}(0) & \boldsymbol{C}(1) & \dots & \boldsymbol{C}_(t-1) \\ \boldsymbol{C}(-1) & \boldsymbol{C}(0) & \dots & \boldsymbol{C}(t-2) \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{C}(1-t) & \boldsymbol{C}(2-t) & \dots & \boldsymbol{C}(0) \end{bmatrix} \boldsymbol{A} = \begin{bmatrix} \boldsymbol{C}(1) \\ \boldsymbol{C}(2) \\ \vdots \\ \boldsymbol{C}(t) \end{bmatrix}$$

that are exactly the first t Yule-Walker equations introduced in the context of VAR processes.