Chapter 11
Prediction Theory for Stationary Random Signals

Abstract Prediction (or forecasting) of future values of the stationary random signals based on the known past depends on the functional analytic tools from Hilbert spaces. Essentially, the optimal predictor is an orthogonal projection of the future values of the signal onto the space spanned by the past values. The chapter presents the relevant Wold decomposition theorem, and an application of the Spectral Representation to the solution of the optimal prediction problem.

11.1 The Wold Decomposition Theorem and Optimal Predictors

In this chapter we will consider prediction problems for discrete time weakly stationary random signals $(X_n), n = \ldots, -2, -1, 0, 1, 2, \ldots$. The assumption is that the second moments are finite, the mean value $\mathbf{E}X_n = 0$, and the span of the "past" of the process in the Hilbert space L_2 will be denoted

$$\mathcal{M}_0 = \overline{\text{span}}\{X_n, n \leq 0\}$$

The optimal predictor $\hat{X}_m$ of the values of the process at time $m > 0$ (in the future) based on the knowledge of the past of the process is, obviously, the orthogonal projection

$$\hat{X}_m := \text{Pred}_0 X_m = \text{Proj}_{\mathcal{M}_0} X_m.$$

In what follows we shall also need the special notation for the following spaces:

$$\mathcal{M}_n = \overline{\text{span}}\{X_k, k \leq n\}, \qquad \mathcal{M}_{-\infty} = \bigcap_n \mathcal{M}_n.$$

W. A. Woyczyński, *A First Course in Statistics for Signal Analysis*, Statistics for Industry, Technology, and Engineering, https://doi.org/10.1007/978-3-030-20908-7_11

We also need to distinguish between two important categories of time series (X_n):

Definition 11.1.1

(a) The process (X_n) is said to be *deterministic* (or, *singular*) if $\mathcal{M}_{-\infty} = \mathcal{M}_{+\infty}$, or, equivalently, in view of the stationarity assumption, if $\mathcal{M}_k = \mathcal{M}_{k+1}$ for all k. In this case the perfect linear prediction is possible because the error

$$\mathbf{E}(\hat{X}_m - X_m)^2 = \|\hat{X}_m - X_m\|_2^2 = 0.$$

(b) The process (X_n) is said to be *regular* if $\mathcal{M}_{-\infty} = \{0\}$. In this case

$$\mathbf{E}(\hat{X}_m - X_m)^2 = \|\hat{X}_m - X_m\|_2^2 > 0.$$

In general,

$$\{0\} \neq \mathcal{M}_{-\infty} \neq \mathcal{M}_{+\infty},$$

so the process is neither deterministic nor regular. However, nondeterministic processes can be decomposed into a regular and deterministic part:

Wold's Decomposition Theorem *If the process (X_n) is regular, then*

$$X_n = Z_n + Y_n, \qquad n = \ldots, -2, -1, 0, 1, 2, \ldots,$$

where (Z_n) is regular, and (Y_n) is deterministic, and, moreover, the two components are orthogonal to each other,

$$(Z_n) \perp (Y_n).$$

The regular process (Z_n) can be expressed in the form

$$Z_n = \sum_{k=0}^{\infty} \gamma_k W_{n-k},$$

where both (W_n) and (Y_n) have zero mean, (W_n) form an uncorrelated sequence with constant variance σ^2, $\gamma_0 = 0$, and $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$. The decomposition is unique.

Proof Let

$$W_k = X_k - \hat{X}_k, \quad k = n, n-1, n-2, \ldots$$

Since $W_k \perp \mathcal{M}_{k-1}$, we see right away that the sequence (W_k) is uncorrelated, that is $\mathbf{E}W_k W_l = 0$, for $l < k$. Define the coefficients γ_k as follows:

$$\gamma_k = \frac{\mathbf{E}X_n W_{n-k}}{\sigma^2}, \quad k = 1, 2, \ldots.$$

Now, we have the obvious inequality

$$0 \leq \mathbf{E}\left(X_n - \sum_{k=0}^{m} \gamma_k W_{n-k}\right)^2 = \mathbf{E}X_n^2 - \sigma^2 \sum_{k=0}^{m} \gamma_k^2,$$

which implies that $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$, and that $\sum_{k=0}^{\infty} \gamma_k W_{n-k}$ converges in L^2 to a random quantity in the subspace spanned by the sequence $W_n, W_{n-1}, W_{n-2}, \ldots$.

Now, the sequence (Y_n) can be defined by the equality,

$$Y_n = X_n - \sum_{k=0}^{\infty} \gamma_k W_{n-k},$$

so that

$$\mathbf{E}Y_n W_l = \mathbf{E}X_n W_l - \sigma^2 \gamma_{n-l} = 0, \quad \text{for,} \quad l \leq n,$$

and $\mathbf{E}Y_n W_l = 0$, for $l > n$, because W_l orthogonal to the subspace $\mathcal{M}_n \ni Y_n$. Therefore $W_n \in \mathcal{M}_{n-1}$, and by induction, $W_n \in \mathcal{M}_k$, for all $k \leq n$, so that

$$\mathcal{M}_{-\infty} = \bigcap_{k=0}^{\infty} \mathcal{M}_{n-s}.$$

To finish the proof of the theorem let us make two observations.

(i) If $\mathcal{M}_W^{\perp}$ is the subspace orthogonal to $\mathcal{M}_W^{\perp}$, the subspace spanned by (W_n), then $\mathcal{M}_{-\infty} = \mathcal{M}_W^{\perp}$. Indeed, if $Xin\mathcal{M}_{-\infty}$, then $X \in \mathcal{M}_n$, and is orthogonal to W_{n+1}, for every n. Hence, $X \in \mathcal{M}_W^{\perp}$. Conversely, if $X \in \mathcal{M}_W^{\perp}$, then $X \in \mathcal{M}_n$, for some n. Since $X \perp W_n$ we have $X \in \mathcal{M}_{n-1}$, and, by induction, $X \in \mathcal{M}_k$, for all $k \leq n$. Moreover, $X \in \mathcal{M}_k$, for $k > n$, because $\mathcal{M}_n \subset \mathcal{M}_k$. So the first observation is verified.

(ii) Since $Z_n = \sum_{k=0}^{\infty} \gamma_k W_{n-k}$, the subspace $\mathcal{M}_n^Z$ spanned by $Z_n, Z_{n-1}, \ldots,$ is contained in the subspace $\mathcal{M}_n^W$ spanned by $W_n, W_{n-1}, \ldots$. Conversely, if $W_n \in \mathcal{M}_n = \mathcal{M}_n^Z \oplus \mathcal{M}_n^Y$, and $W_n \perp \mathcal{M}_n^Y$, then $W_n \in \mathcal{M}_n^Z$. So $\mathcal{M}_n^W = \mathcal{M}_n^Z$.

Now we are ready to complete the proof of the Decomposition Theorem. Since, for every n, $Y_n \in \mathcal{M}_{-\infty} =\supseteq \mathcal{M}_n^Y$, the condition $X \in \mathcal{M}_{-\infty}$ implies that $X \in \mathcal{M}_n$ because $X \perp \mathcal{M}_n^W = \mathcal{M}_n^Z$. Thus $X \in \mathcal{M}_n Y$. This proves that $|calM_n^Y = \mathcal{M}_{-\infty}$, and the sequence (Y_n) is deterministic.

Now, since $Z_n = W_n + \sum_{k=1}^{\infty} \gamma_k W_{n-k}$, and $W_n \perp \sum_{k=1}^{\infty} \gamma_k W_{n-k} \in \mathcal{M}_{n-1}^W$, the error $\mathbf{E}(Z_n - \hat{Z}_n)^2 = \sigma^2 > 0$, so that the sequence (Z_n) is regular. ∎

Since

$$X_n = Z_n + Y_n = \sum_{k=0}^{\infty} \gamma_k W_{n-k} + Y_n = W_n + \sum_{k=1}^{\infty} \gamma_k W_{n-k} + Y_n,$$

and

$$W_n \perp \sum_{k=1}^{\infty} \gamma_k W_{n-k} + Y_n,$$

the best predictor for X_n is the orthogonal projection of X_n onto $\mathcal{M}_{n-1}$, which is

$$\hat{X}_n = \sum_{k=1}^{\infty} \gamma_k W_{n-k} + Y_n.$$

The square of its error

$$\|X_n - \hat{X}_n\|_{L^2}^2 = \mathbf{E}(X_n - \hat{X}_n)^2 = \mathbf{E}W_n^2 = \sigma^2,$$

because $\gamma_0 = 1$.

11.2 Application of the Spectral Representation to the Solution of the Prediction Problem

In this section we will consider the case of discrete time stationary signal $X(n)$, and assume that $\mathbf{E}X(n) = 0$. The spectral representation theorem of Sect. 10.4 gives rise to a linear isometry

$$L^2([0, 1], dC_{\mathcal{W}}) \ni g \longrightarrow \int_0^1 g(f) d\mathcal{W}(f) \in L^2(\Omega, \mathcal{F}, P),$$

which simply extends the representation,

$$X(n) = \int_0^1 e^{j2\pi nf} d\mathcal{W}(f),$$

where the cumulative control function

$$C_{\mathcal{W}}(f) = \mathbf{E}[\mathcal{W}(f)]^2 = \mathcal{S}_X(f),$$

where $\mathcal{S}_X(f)$ is the cumulative spectral function of the process $X(n)$. Obviously, in the particular case $g(f) = e^{j2\pi nf}$ the isometry is the mapping,

$$e^{j2\pi nf} \longrightarrow X(n).$$

So, the optimal prediction of the value of the signal at the future time $m > 0$, based on the past values $X(n)$, $n \leq 0$, is reduced to finding the function,

$$g(f) \in \overline{\text{span}}_{L^2(dS)}(e^{j2\pi nf}, n \leq 0),$$

such that the error of the prediction is minimal, that is

$$\|e^{j2\pi mf} - g(f)\|_{L^2(dS)} = \min_h \|e^{j2\pi mf} - h(f)\|_{L^2(dS)},$$

where $h \in \text{span}_{L^2(d\mathcal{S})}(e^{j2\pi nf}, n \leq 0)$. Or, equivalently, the optimal choice of g has to be an orthogonal projection in L^2, that is

$$e^{j2\pi mf} - g(f) \perp \overline{\text{span}}_{L^2(d\mathcal{S})}(e^{j2\pi nf}, n \leq 0),$$

that is

$$\int_0^1 \left[e^{j2\pi mf} - g(f)\right] e^{-j2\pi nf} d\mathcal{S}(f) = 0, \quad \text{for} \quad n = 0, -1, -2, \ldots.$$

Remark 11.2.1 Observe that if the cumulative spectral function $\mathcal{S}_X(f)$ does not increase (or, its spectral density $S_X(f) = 0$) over the interval $[a, b] \subset [0, 1]$ of length greater than 1/2, then the signal $X(n)$ is singular.

Indeed, let $e^{-j2\pi f}$ be in the arc of the unit circle in the complex plane corresponding to $f \ni [a, b]$, and let $e^{j2\pi f_0}$ be the midpoint of the arc. Then, for large enough N,

$$\left| e^{j2\pi f_0} - \frac{e^{-j2\pi f}}{N} \right| < 1,$$

because of the above length assumption, so, also,

$$\left| 1 - \frac{e^{-j2\pi f}}{N e^{j2\pi f_0}} \right| < \frac{1}{|e^{j2\pi f_0}|} = 1.$$

Hence, we get the following uniformly convergent expansion on the complement of the interval $[a, b]$:

$$e^{j2\pi f} = \frac{1}{e^{-j2\pi f}} = \frac{1}{Ne_0^{j2\pi f}} \cdot \frac{1}{e^{-j2\pi f}/(Ne^{j\pi f_0})}$$

$$= \frac{1}{Ne^{j2\pi f_0}} \cdot \frac{1}{1-(1-e^{-j2\pi f}/(Ne^{j2\pi f_0}))}$$

$$\frac{1}{Ne^{j2\pi f_0}} \cdot \sum_{n=0}^{\infty}(1-e^{-jf2\pi}/(Ne^{j2\pi f_0}))^n \in \overline{\text{span}}_{L^2(dS_X)}(e^{j2\pi fn}, n \le 0) = \mathcal{M}_0,$$

which completes the justification of the above statement. On the other hand, on the set $[a, b]$, where the spectral density is 0, the approximation is trivial.

Remark 11.2.2 It turns out that the Wold decomposition is equivalent to decomposition of the spectral measure into the absolutely continuous (with density) and singular components[1]

In the reminder of this section we will just consider the absolutely continuous case when

$$\mathcal{S}(f) = S(f)df$$

with the spectral density $S(f)$ satisfying the condition,

$$0 < C_1 \le S(f) \le C_2 < \infty, \tag{11.2.1}$$

in which case $L^2(d\mathcal{S}) = L^2(df)$ and the convergences in those two spaces are equivalent.

In this case the best predictor $g(f)$ satisfies the following two conditions:

$$\int_0^1 [e^{j2\pi mf} - g(f)]S(f)e^{-j2\pi nf}df = 0, \quad \text{for} \quad n \le 0, \tag{11.2.2}$$

and

$$[e^{j2\pi mf} - g(f)]S(f) \in \overline{\text{span}}_{L^2(S(f)df)}(e^{j2\pi nf}, n \ge 0) \equiv \mathcal{M}_{>0}. \tag{11.2.3}$$

Now, assume that we can factor the spectral density,

$$S(f) = S_1(f) \cdot S_1^*(f),$$

[1]For more details see, U. Grenander and M. Rosenblatt, *Statistical Analysis of Stationary Time Series*, Almqvist and Wiksell, Stockholm 1956, and P. Bremaud, *Fourier Analysis and Stochastic Processes*, Springer 2014.

with both

$$S_1(f),\, S_1^{-1}(f) \in \text{span}_{\mathcal{C}}(e^{j2\pi nf},\, n \le 0) =: \mathcal{C}_{\le 0}$$

where $\mathcal{C}$ denotes the space of continuous functions. Then the condition (11.2.3) can be rewritten in the form

$$[e^{j2\pi mf} - g(f)]S_1(f)S_1^*(f) \in \overline{\text{span}}_{L^2(S(f)df)}(e^{j2\pi nf},\, n \ge 0) \equiv \mathcal{M}_{>0}. \tag{11.2.4}$$

with

$$(S_1^{-1}(f))^* \in \text{span}_{\mathcal{C}}(e^{j2\pi nf},\, n \ge 0).$$

Hence,

$$h(f) := [e^{j2\pi mf} - g(f)]S_1(f) \in \overline{\text{span}}_{L^2(S(f)df)}(e^{j2\pi nf},\, n > 0),$$

and the condition for the best linear prediction can be reformulated as follows:

$$e^{j2\pi mf}S_1(f) = g(f)S_1(f) + h(f), \qquad g \in \mathcal{M}_{\le 0},\ h \in \mathcal{M}_{>0}. \tag{11.2.5}$$

Since $S_1,\, S_1^{-1} \in \mathcal{C}_{\le 0}$,

$$g \in \mathcal{M}_{\le 0} \iff gS_1 \in \mathcal{M}_{\le 0},$$

so, what needs to be done at this point is to split the Fourier series of $e^{j2\pi mf}S_1(f)$ into the $\mathcal{M}_{\le 0}$, and $\mathcal{M}_{>0}$ parts.

Given the expansion

$$S_1(f) = c_0 + c_{-1}e^{-j2\pi f} + c_{-2}e^{-j2\pi 2f} + \ldots$$

we can write (11.2.5) with

$$h(f) = c_0 e^{j2\pi mf} + c_{-1}e^{k2\pi(m-1)f} + \ldots + c_{-m+1}e^{j2\pi f},$$

and

$$g(f)S_1(f) = c_{-m} + c_{-m-1}e^{-j2\pi f} + c_{-m-2}e^{-j2\pi 2f} + \ldots.$$

Hence,

$$g(f) = \left[c_{-m} + c_{-m-1}e^{-j2\pi f} + c_{-m-2}e^{-j2\pi 2f} + \ldots\right] \cdot S_1^{-1}(f),$$

which expands as follows:

$$g(f) = b_0 + b_{-1}e^{-j2\pi f} + b_{-2}e^{-j2\pi 2f} + \ldots,$$

with the predictor

$$\hat{X}_m = b_0 X_0 + b_{-1} X_{-1} + b_{-2} X_{-2} + \dots$$

The prediction error can then be calculated as follows:

$$\|\hat{X}_m - X_m\|^2_{L^2(S(f)df)} = \int_0^1 |e^{j2\pi mf} - g(f)|^2 S(f) df$$

$$= \int_0^1 |e^{j2\pi mf} - g(f) S_1(f)|^2 df = \int_0^1 |h(f)|^2 df \tag{11.2.6}$$

$$= \int_0^1 |c_0 e^{j2\pi mf} + c_{-1} e^{k2\pi(m-1)f} + \dots + c_{-m+1} e^{j2\pi f}|^2 df = |c_0|^2 + \dots |c_{-m+1}|^2.$$

When $m \to \infty$,

$$\sum_{n=0}^{\infty} |c_{-n}|^2 = \int_0^1 |S_1(f)|^2 df = \int_0^1 S(f) df = \mathbf{E}|X_k|^2, \qquad \forall k,$$

so that the signal $(X(k))$ is regular.

Remark 11.2.3 Let us take a look at the one step predictor $\hat{X}_1$ in the case $\log S(f)$ satisfies some smoothness conditions to permit the following expansion of its logarithm, $\log S(f)$:

$$\left(\dots + a_{-2} e^{-j2\pi 2f} + a_1 e^{-j2\pi} + \frac{a_0}{2}\right) + \left(\frac{a_0}{2} + a_1 e^{+j2\pi} + a_2 e^{+j2\pi 2f} + \dots\right).$$

Substituting

$$S_1(f) = \exp\left(\dots + a_{-2} e^{-j2\pi 2f} + a_1 e^{-j2\pi} + \frac{a_0}{2}\right),$$

we see that both S_1 and S_1^{-1} are functions from $\mathcal{C}_{\leq 0}$. Using the standard expansion $e^z = 1 + z + z^2/2 + \dots$, one obtains the equality

$$c_0 = 1 \frac{a_0}{2} + \frac{(a_0/2}{2!} + \dots = e^{a_0/2}.$$

Hence, the one step error

$$\|\hat{X}_1 - X_1\|^2_{L^2(S(f)df)} = |c_0|^2 = e^{a_0} = \exp\left(\int_0^1 \log S(f) df\right).$$

Notice that, in general, this error is nonzero if, and only if,

$$\int_0^1 \log S(f)df > -\infty,$$

which is the general condition for the regularity of the random stationary signal X_n.[2]

11.3 Examples of Linear Prediction for Stationary Time Series

In this section we will consider a simple example of stationary time series where the calculation of the optimal predictor is not very difficult.

Let $X(t)$ be a stationary time series, $t = \ldots, -1, 0, 1, \ldots$, with the autocovariance function

$$\gamma_X(t) = a^{|t|}, \qquad -1 < a < 1.$$

The corresponding spectral density, assuming the representation $\gamma_X(t) = \int S_X(f)e^{-jft}df$, is

$$S_x(f) = \frac{1 - a^2}{2\pi(e^{jf} - a)(e^{-jf} - a)},$$

which can be rewritten in the form

$$S_X(f) = \hat{S}_X(e^{jf}),$$

where

$$\hat{S}_X(z) = \frac{(1 - a^2)z}{2\pi(z - a)(1 - az)}.$$

Finding the optimal predictor m steps ahead requires finding a function

$$\Phi_m(f) = a_1e^{-jf} + a_2e^{-j2f} + a_3e^{-j3f},$$

satisfying the condition

$$\int_{-\pi}^{\pi} e^{jkf}[e^{jmf} - \Phi_m(f)]S_X(f)df = 0, \qquad k = 1, 2, 3, \ldots$$

[2] Again, see, Grenander and Rosenblatt, and Bremaud's books cited on page 284, for more details.

In other words, the Fourier expansion of the function

$$\Psi_m(f) = [e^{jmf} - \Phi_m(f)]S_X(f) = \sum_{k=0}^{\infty} c_k e^{jkf}$$

contains only nonnegative powers of e^{jf}.

In the case of rational $\hat{S}_X(z)$, the function

$$\hat{\Phi}_m(z) = \sum_{k=1}^{\infty} a_k z^{-k}$$

is an analytic function of z for $|z| \geq 1$, with $\hat{\Phi}_m(\infty) = 0$, and

$$\hat{\Psi}_m(z) = [z^m - \hat{\Psi}_m(z)]\hat{S}_X(z),$$

is analytic for $|z| \leq 1$.

So, if in our case we are attempting to make a prediction one time step ahead, that is, assuming $m = 0$, we need to find a function $\hat{\Phi}_0(z)$ with no singularities for $|z| \geq 1$, vanishing at infinity, and such that the function

$$\hat{\Psi}_0(z) = \frac{(1-a^2)[1 - \hat{\Phi}_0(z)]]z}{2\pi(z-a)(1-az)}$$

has no singularities for $|z| \leq 1$. Since $|a| < 1$ we must have $\hat{\Phi}_0(a) = 1$. The above formula implies that $\hat{\Phi}_0(z)$ has no singularities other that a simple pole at $z = 0$. Thus,

$$\hat{\Phi}_0(z) = g_0(z)z^{-1},$$

where $g_0(z)$ is analytic in the whole complex plane, and $g_0(a) = a$. So the only function satisfying the above conditions is

$$\hat{\Phi}_0(a) = az^{-1}, \qquad with \qquad \Phi(f) = ae^{-jf}.$$

Therefore the optimal predictor for $X(t)$ is $aX(t-1)$. So, in this case the best predictor just depends on the value of the process one step back and does not depend on the whole past of the process.[3]

[3]For more details and analysis of more complicated rational spectral densities see *An Introduction to the Theory of Random Stationary Functions*, by A.M. Yaglom, Dover Publications. New York, 1973.

11.4 Problems and Exercises

1 Verify that in the case considered in Remark 11.2.1 the best predictor one time-step ahead $\hat{X}_1$ is expressed by the formula

$$\hat{X}_1 = \sum_{n=0}^{\infty} (Ne^{jf_0})^{-n-1} \sum_{k=0}^{n} \binom{n}{k} (e^{jf_0})^{n-k} (-1)^k X_{-k}.$$

2 Prove that if the spectral density $S(f)$ is satisfying the condition (11.2.1),

$$0 < C_1 \leq S(f) \leq C_2 < \infty,$$

then $L^2(dS) = L^2(df)$, and the convergences in those two spaces are equivalent.

3 Show that in the case analyzed in Sect. 11.3 the optimal prediction m time steps ahead, that is at time $t + m$, also depends only on the single value of the process in the past and is of the form

$$a^{m+1} X(t-1).$$

4 Show that in the case of the spectral density of the form

$$S_x(f) = \frac{1}{|e^{jf} - a_1|^2 |e^{jf} - a_2|^2}, \qquad |a_1|, |a_2| < 1,$$

the optimal prediction one time step ahead depends only on the two values of the process in the past, and is of the form

$$(a_1 + a_2) X(t-1) + a_1 a_2 X(t-2).$$