EECS 225A Statistical Signal Processing Lecture 8 - 2/13/2020

Lecture 8: Wold Decomposition and the Kolmogorov–Szego Formula

Lecturer: Jiantao Jiao *Scribe: Shivin Devgon*

We talk about the prediction problem in the previous lecture, and used spectral factorization to derive the optimal predictor as well as the optimal one-step prediction error.

How general are these results? When does spectral factorization exist for general WSS processes? Can we obtain a formula for the one-step prediction error without computing the spectral factorization?

We answer these questions in this lecture through the lens of Wold decomposition and the Kolmogorov–Szego formula. We first define formally a white noise process.

Definition 1 (White noise)**.** *We say WSS process Z_n is a white noise process with variance σ^2, denoted as*

$$Z_n \sim \mathsf{WN}(0, \sigma^2),$$

if and only if

$$\mathbb{E}[Z_n] = 0$$
$$\mathbb{E}[Z_n Z_m^*] = \sigma^2 \delta_{mn}.$$

We define the following Hilbert spaces with inner product $\langle X, Y \rangle \triangleq \mathbb{E}[XY^*]$.

Definition 2 (Some Hilbert Spaces)**.** *[1, Section 2.2.4] Let $\{X_n : n \in \mathbb{Z}\}$ be a WSS process. We define*

$$H(X) \triangleq \overline{\mathsf{sp}}\{X_n : n \in \mathbb{Z}\}$$
$$H_n(X) \triangleq \overline{\mathsf{sp}}\{X_s : s \le n\} \text{ for any } n \in \mathbb{Z}$$
$$H_{-\infty}(X) \triangleq \bigcap_{n \in \mathbb{Z}} H_n(X),$$

where $\overline{\mathsf{sp}}$ denotes the closure of the span of the random variable in this set.

If $H_{-\infty}(X) = H(X)$, we call $\{X_n : n \in \mathbb{Z}\}$ a deterministic *process. The process is called* nondeterministic *if $H_{-\infty}(X) \subset H(X)$ and $H_{-\infty}(X) \neq H(X)$, and* purely nondeterministic *if $H_{-\infty}(X) = \{0\}$.*

We have $H_n(X) \downarrow H_{-\infty}(X)$ as $t \to -\infty$ and $H_n(X) \uparrow H(X)$ as $n \to \infty$.

For any random variable $Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, we use $P_t Y$ to denote the projection of Y on $H_t(X)$.

It follows from standard arguments that the prediction error

$$\delta(\tau) \triangleq \mathbb{E}[|X_n - P_{n-\tau} X_n|^2] \tag{1}$$

is independent from n, and increases with τ, with $\delta(\tau) = 0$ for $\tau \le 0$.

The following theorem gives an alternative description of a process being deterministic or purely nondeterministic.

Theorem 3. *[1, Section 2.2.5] The following statements are true for WSS process X.*

1. *If $\{X_n : n \in \mathbb{Z}\}$ is deterministic, then $\delta(\tau) = 0$ for all $\tau \in \mathbb{Z}$.*
2. *If $\delta(\tau) = 0$ for some $\tau > 0$, then X is deterministic*
3. *The process X is purely nondeterministic if and only if $\delta(\tau) \to \mathbb{E}[|X_0|^2]$ as $\tau \to \infty$.*

For simplify we denote $\delta(1)$ by σ^2, which means the one-step prediction error.

1 Wold's decomposition

Now we are ready to state the the Wold decomposition theorem.

Theorem 4. *[2, Chapter 5.7][1, Chapter 2.2.6] If $\sigma^2 > 0$, then the WSS process $\{X_n : n \in \mathbb{Z}\}$ can be represented uniquely as*

$$X_n = \sum_{j=0}^{\infty} c_j Z_{n-j} + V_n \tag{2}$$

with the following properties:

1. $c_0 = 1, \sum_{j=0}^{\infty} |c_j|^2 < \infty$
2. $Z_n \sim \mathsf{WN}(0, \sigma^2)$ *This is known as the innovation sequence because each Z_n is uncorrelated with previous Z_n.*
3. $Z_n \in H_n, \ \forall n \in \mathbb{Z}$. *This means that Z_n can be constructed from $\{X_t : t \leq n\}$.*
4. $\mathbb{E}[Z_n V_m^*] = 0, \ \forall n, m \in \mathbb{Z}$.
5. $V_n \in H_{-\infty}, \ \forall n \in \mathbb{Z}$
6. V_n *is deterministic.*

Without discussing the proof of Wold's decomposition, we present the constructions of Φ_j, Z_j, V_j below.

$$Z_n = X_n - P_{n-1} X_n \tag{3}$$

$$c_j = \frac{\mathbb{E}[X_n Z_{n-j}^*]}{\sigma^2} \text{ for } j \geq 1 \tag{4}$$

$$V_n = X_n - \sum_{j=0}^{\infty} c_j Z_{n-j} \tag{5}$$

We can think of $\sum_{j=0}^{\infty} c_j Z_{n-j}$ as passing white noise process Z_n through a canonical modeling causal filter c. In other words, the Wold decomposition gives the canonical spectral factorization of the purely nondeterministic part of the process

$$U_n \triangleq \sum_{j=0}^{\infty} c_j Z_{n-j}. \tag{6}$$

We can relate $\delta(\tau)$ to the coefficients c_j computed above:

$$\delta(\tau) = \sigma^2 \sum_{j=0}^{\tau-1} |c_j|^2, \tag{7}$$

Indeed, it follows from the Wold decomposition that

$$P_{n-\tau} X_n = \sum_{j=\tau}^{\infty} c_j Z_{n-j} + V_n \tag{8}$$

$$P_{n-\tau} U_n = \sum_{j=\tau}^{\infty} c_j Z_{n-j}. \tag{9}$$

It also implies that

$$X_n - P_{n-\tau}X_n = U_n - P_{n-\tau}U_n,$$

showing the optimal prediction errors of X_n and U_n are the same.

We define the z-transform of the filter $\{c_j\}_{j=0}^{\infty}$ as

$$C(z) = \sum_{j=0}^{\infty} c_j z^{-j}.$$

Denoting z-spectrum of WSS process U as $S_U(z)$, then the Wold decomposition gives

$$S_U(z) = C(z)\sigma^2 C^*(z^{-*}).$$

However, we would like to emphasize [3, Section 6.4, Remark 5] that under the weak assumptions of Wold decomposition, we can only guarantee that $C(z)$ and $C^{-1}(z)$ are analytic in $|z| > 1$. If $C(z)$ is rational, then we can strengthen the conclusion to state that $C(z)$ is analytic in $|z| \geq 1$, but even for rational $C(z)$ we cannot guarantee that $C^{-1}(z)$ is analytic in $|z| \geq 1$ since $C(z)$ may have isolated unit-circle zeros.

2 How large is σ^2?

The interested readers must have observed that we have assumed $\sigma^2 > 0$ in the Wold decomposition. How would we know whether a WSS has $\sigma^2 > 0$ or not? Can we obtain this information in an elegant way from its power spectral measure?

Recall that in lecture 4, we have showed that for any zero-mean WSS process X_n, there exists a unique right-continuous stochastic process $F(\omega), \omega \in (-\pi, \pi]$ with zero-mean square-integrable orthogonal increments such that

$$X(n) = \int_{-\pi}^{\pi} e^{j\omega n} \mathrm{d}F(\omega), \tag{10}$$

and we define its *power spectral measure* $M_X(\omega)$ as

$$M_X(\omega) \triangleq 2\pi\mathbb{E}|F(\omega) - F(-\pi)|^2.$$

When $M_X(\omega)$ is differentiable, we have

$$M_X'(\omega) = S_X(\omega),$$

where $S_X(\omega)$ is the power spectral density of X. It is clear that generally one needs the power spectral measure instead of the power spectral density to characterize the WSS process. Indeed, for $X(n) = W(n) + A, S_X(\omega) = S_W(\omega) + 2\pi\mathbb{E}[|A|^2]\delta(\omega)$, which is a generalized function, but the power spectral measure is a function with a discontinuous point at zero. Note that M_X is well defined even when the process is not zero mean: one can just add a delta mass $2\pi|\mu_X|^2\delta(\omega)$ to it.

We introduce the concept of Lebesgue decomposition in the special case.

Definition 5 (Lebesgue decomposition)**.** *For any finite measure μ on an interval I, there exists two finite measures μ_0 and μ_1 such that*

1. $\mu = \mu_0 + \mu_1$
2. $\mu_0 \ll \lambda$*, that is, μ_0 is absolutely continuous with respect to the Lebesgue measure on interval I*
3. $\mu_1 \perp \lambda$*, that is, μ_1 and the Lebesgue measure are mutually singular.*

These two measures are uniquely determined by μ.

By absolutely continuity we means that for any set A, if $\lambda(A) = 0$, then $\mu_0(A) = 0$. By mutually singular we mean that there exist two disjoint sets A, B such that A is the complement set of B, and μ_1 is zero on all measurable subsets of A while λ is zero on all measurable subsets of B.

The next theorem characterizes when $\sigma^2 > 0$.

Theorem 6. *[1, Chapter 2.5] Suppose the Lebesgue decomposition of $M_X(\omega)$ is $A(\omega) + B(\omega)$, where A is the absolutely continuous part, and B is the singular part. Then, X is nondeterministic (i.e., $\sigma^2 > 0$) if and only if the* Paley–Wiener *condition (11) holds:*

$$\int_{-\pi}^{\pi} \ln(A'(\omega))\mathrm{d}\omega > -\infty. \tag{11}$$

Since we have assumed X has finite power, it is equivalent to

$$\int_{-\pi}^{\pi} |\ln(A'(\omega))|\mathrm{d}\omega < \infty, \tag{12}$$

where A' denotes the derivative of A.

Moreover, if the Paley–Wiener condition holds, then $A(\omega)$ is the power spectral measure of the purely nondeterministic part of X, and $B(\omega)$ is the power spectral measure of the deterministic part of X.

We call a WSS process regular if in the corresponding Lebesgue decomposition of the power spectral measure has $B = 0$, and the Paley–Wiener condition holds. Clearly, if the WSS process is bandlimited, which means $A'(\omega) = 0$ for a set of positive Lebesgue measure, then the Paley–Wiener condition fails, which implies that any bandlimited process is deterministic.

Can we compute σ^2 as a function of the power spectral measure? Clearly, the deterministic part can be perfectly predicted so it should not contribute to σ^2. It turns out that one can directly compute σ^2 based on the Lebesgue decomposition of the power spectral measure as below.

Theorem 7 (Kolmogorov–Szego formula)**.** *[1, Chapter 2.5.7] The one-step optimal linear prediction error σ^2 of any WSS process X is given by*

$$\sigma^2 = \exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln(A'(\omega))\mathrm{d}\omega\right), \tag{13}$$

where A' is the derivative of $A(\omega)$, and $A(\omega)$ is the absolutely continuous part of the Lebesgue decomposition of $M_X(\omega)$.

As a sanity check, it follows from Jensen's inequality that

$$\begin{aligned}
\sigma^2 &\le \frac{1}{2\pi}\int_{-\pi}^{\pi} A'(\omega)\mathrm{d}\omega \\
&= \frac{1}{2\pi}(A(\pi) - A(-\pi)) \\
&\le \frac{1}{2\pi}(M_X(\pi) - M_X(-\pi)) \\
&= \mathbb{E}[|X_0|^2],
\end{aligned}$$

where the equality is achieved if and only if there is no deterministic part and the the WSS process X is white noise. In this case, no gain in prediction is possible compared to using the trivial estimator of zero.

As another sanity check, last lecture shows that if we can do the spectral factorization

$$S_X(\omega) = S_X^+(\omega)S_X^-(\omega),$$

and

$$S_X^+(\omega) = p_0 + \sum_{k=1}^{\infty} p_k e^{-j\omega k}$$

then

$$\sigma^2 = p_0^2.$$

The Kolmogorov–Szego formula seems to suggest that

$$p_0^2 = \exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln(A'(\omega))\mathrm{d}\omega\right).$$

Is this true? Can we use A' to obtain the spectrum factorization of X? The answer turns out to be yes as the next section shows.

3 Kolmogorov Cepstral Method [3, Problem 6.12]

Suppose X is a WSS purely nondeterministic process. Let $S_X(z)$ denote the z-spectrum of X and assume that $\ln(S_X(z))$ is analytic in an annulus that includes the unit circle, so that it can be expanded in a Laurent series

$$\ln(S_X(z)) = \sum_{k=-\infty}^{\infty} \gamma_k z^{-k}.$$

We show that the canonical spectral factorization of $S_X(z) = L(z) r_e L^*(z^{-*})$ is given by

$$L(z) = \exp\left(\sum_{j=1}^{\infty} \gamma_j z^{-j}\right)$$
$$r_e = \exp(\gamma_0).$$

Indeed, for this L we have

$$\begin{aligned} L^*(z^{-*}) &= \exp\left(\sum_{j=1}^{\infty} \gamma_j^* z^j\right) \\ &= \exp\left(\sum_{j=-\infty}^{-1} \gamma_{-j}^* z^{-j}\right) \\ &= \exp\left(\sum_{j=-\infty}^{-1} \gamma_j z^{-j}\right), \end{aligned}$$

where in the last step we used the property that $\gamma_j = \gamma_{-j}^*$, which can be shown when we specialize $z = e^{j\omega}$ and use the property that $S_X(e^{j\omega})$ is real-valued.

As a sanity check, using the IDTFT formula, we know

$$\gamma_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln(S_X(e^{j\omega}))\mathrm{d}\omega,$$

which is consistent with the Kolmogorov–Szego formula.

4 Construct the optimal predictor $P_{n-1}X_n$

Suppose the WSS process has a deterministic part, which makes it impossible to do spectral factorization for the whole spectrum. How can we construct the optimal predictor of X_n using its history $H_{n-\tau}$?

In general, we only have efficient methods to compute the optimal predictor for either purely nondeterministic processes or purely deterministic processes. There does not exist a general formula for predicting a general WSS process.

4.1 Predict a WSS process with line spectrum

Suppose we have a WSS process X with PSD

$$S_X(\omega) = \sum_i \alpha_i \delta(\omega - \omega_i).$$

Construct predictor with causal transfer function

$$H(\omega) = 1 - \prod_i (1 - e^{j\omega_i} e^{-j\omega}),$$

we have the PSD of the prediction error $X_n - (h * X)_n$ given by

$$S_X(\omega)|1 - H(\omega)|^2 = |\prod_i (1 - e^{-j\omega} e^{j\omega_i})|^2 \sum_i \alpha_i \delta(\omega - \omega_i) = 0.$$

4.2 Difficulty for constructing $P_{n-1}X_n$ for general WSS process

Suppose $X_n = W_n + A$, where $A \in \mathbb{R}$ is a random variable independent of W_n process, and W_n is a real-valued white noise process with variance one. Denote the the optimal predictor of X_n using $\{X_{n-1}, X_{n-1}, \ldots, X_{n-k}\}$ as $\hat{X}_{n,k}$. The projection $P_{n-1}X_n$ is given by

$$\lim_{k\to\infty} \hat{X}_{n,k}.$$

We now compute $\hat{X}_{n,k}$. Denote X_n as X, X_{n-i} as Y_i, $1 \le i \le k$. Then, $\hat{X}_{n,k}$ is the optimal linear estimator of X using $Y_1, Y_2, \ldots, Y_k$. We denote the vector $[Y_1, Y_2, \ldots, Y_k]^\top$ as Y.

We have

$$\begin{aligned} R_{XY} &= \mathbb{E}[XY^\top] \\ &= [\mathbb{E}[XY_1], \mathbb{E}[XY_2], \ldots, \mathbb{E}[XY_k]]. \end{aligned}$$

For each $1 \le i \le k$,

$$\begin{aligned} \mathbb{E}[XY_i] &= \mathbb{E}[X_n X_{n-i}] \\ &= \mathbb{E}[(W_n + A)(W_{n-i} + A)] \\ &= \mathbb{E}[A^2]. \end{aligned}$$

Hence,

$$R_{XY} = \mathbb{E}[A^2]\mathbf{1}^\top,$$

where $\mathbf{1} \in \mathbb{R}^k$ is the column vector with each entry equal 1.

We also compute R_Y as

$$\begin{aligned} R_Y &= \mathbb{E}[YY^\top] \\ &= \mathbb{E}[(Y - A\mathbf{1} + A\mathbf{1})(Y - A\mathbf{1} + A\mathbf{1})^\top] \\ &= \mathbb{E}[(Y - A\mathbf{1})(Y - A\mathbf{1})^\top] + \mathbb{E}[A^2]\mathbf{1}\mathbf{1}^\top \\ &= I + \mathbb{E}[A^2]\mathbf{1}\mathbf{1}^\top. \end{aligned}$$

Now we need the Sherman–Morrison formula, which states that

$$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u},$$

where A is an invertible matrix, and $1 + v^\top A^{-1}u \neq 0$.

It then follows from the Sherman–Morrison formula that

$$R_Y^{-1} = I - \frac{1}{1 + k\mathbb{E}[A^2]}\mathbb{E}[A^2]\mathbf{1}\mathbf{1}^\top.$$

Hence, the optimal linear predictor of X given Y is

$$\begin{aligned} \mathbb{E}[A^2]\mathbf{1}^\top(I - \frac{1}{1 + k\mathbb{E}[A^2]}\mathbb{E}[A^2]\mathbf{1}\mathbf{1}^\top)Y &= \mathbb{E}[A^2]\mathbf{1}^\top Y - \frac{(\mathbb{E}[A^2])^2 n}{1 + k\mathbb{E}[A^2]}\mathbf{1}^\top Y \\ &= \mathbb{E}[A^2]\mathbf{1}^\top Y\left(1 - \frac{\mathbb{E}[A^2]k}{1 + k\mathbb{E}[A^2]}\right) \\ &= \mathbb{E}[A^2]\mathbf{1}^\top Y\frac{1}{1 + k\mathbb{E}[A^2]} \\ &= \frac{\mathbb{E}[A^2]}{1 + k\mathbb{E}[A^2]}\mathbf{1}^\top Y, \end{aligned}$$

which shows that

$$\hat{X}_{n,k} = \frac{\mathbb{E}[A^2]}{1 + k\mathbb{E}[A^2]}\sum_{j=1}^{k} X_{n-j}. \tag{14}$$

4.3 Relations between the optimal predictors for X and its purely nondeterministic and deterministic parts

How can we relate the optimal predictors of X, its purely nondeterministic part, and its purely deterministic part?

There are a few insightful observations. Note that we may not be able to represent the optimal predictor of X_n given its causal history as

$$\hat{X}_n = \sum_{j=1}^{\infty} h_j X_{n-j}, \tag{15}$$

but we are always able to write it as the limit [4, Theorem 2]

$$\hat{X}_n = \lim_{k\to\infty} \hat{X}_{n,k} \tag{16}$$

$$= \lim_{k\to\infty} \sum_{j=1}^{k} h_j^{(k)} X_{n-j}, \tag{17}$$

where $\hat{X}_{n,k}$ is the optimal linear predictor of X_n given $\{X_{n-1}, X_{n-2}, \ldots, X_{n-k}\}$. Here the convergence is in the mean squared sense. We also have

$$\mathbb{E}|X_n - \hat{X}_n|^2 = \lim_{k\to\infty} \mathbb{E}|X_n - \hat{X}_{n,k}|^2 \tag{18}$$

$$= \sigma^2. \tag{19}$$

Now we decompose $X_n = U_n + V_n$. Then, for any $k \geq 1$,

$$\mathbb{E}|X_n - \hat{X}_{n,k}|^2 = \mathbb{E}|U_n - (h^{(k)} * U)_n + V_n - (h^{(k)} * V)_n|^2 \tag{20}$$

$$= \mathbb{E}|U_n - (h^{(k)} * U)_n|^2 + \mathbb{E}|V_n - (h^{(k)} * V)_n|^2 \tag{21}$$

$$\geq \sigma^2 + 0, \tag{22}$$

In the last step we used the orthogonality of U_n and V_n. Then we take the limit $k \to \infty$. Since $\lim_{k\to\infty} \mathbb{E}|X_n - \hat{X}_{n,k}|^2 = \sigma^2$, it implies that

$$\lim_{k\to\infty} \mathbb{E}|U_n - (h^{(k)} * U)_n|^2 = \sigma^2 \tag{23}$$

$$\lim_{k\to\infty} \mathbb{E}|V_n - (h^{(k)} * V)_n|^2 = 0. \tag{24}$$

Indeed, the filter in (14) applied to the white noise process X_n converges to 0 as $k \to \infty$.

We just showed that the optimal predictor for X can be used for optimal linear prediction for its purely nondeterministic and purely deterministic parts. However, the example above shows that the optimal predictor for the purely nondeterministic part of X, which is constant zero, is clearly a bad predictor for the process X. When is it in fact good?

Now assume that the optimal linear predictor of X_n indeed can be written as

$$\hat{X}_n = \sum_{j=1}^{\infty} h_j X_{n-j}. \tag{25}$$

Then we have

$$\sigma^2 = \mathbb{E}|X_n - \hat{X}_n|^2 \tag{26}$$

$$= \mathbb{E}|U_n - (h * U)_n + V_n - (h * V)_n|^2 \tag{27}$$

$$= \mathbb{E}|U_n - (h * U)_n|^2 + \mathbb{E}|V_n - (h * V)_n|^2 \tag{28}$$

$$\geq \sigma^2 + 0. \tag{29}$$

It implies that equality must hold for all the steps, showing that the filter h is in fact the optimal predictor simultaneously for X_n, U_n, V_n.

It implies that h is also an optimal linear predictor for both the purely nondeterministic part and purely deterministic part. In general many filters could predict the purely deterministic part perfectly, but under some conditions the optimal predictor for the purely nondeterministic part is unique. Indeed, if both h and g are optimal predictor for U_n, then we have

$$\mathbb{E}|(h * U)_n - (g * U)_n|^2 = 0, \tag{30}$$

which is equivalent to

$$S_U(\omega)|H(\omega) - G(\omega)|^2 = 0 \tag{31}$$

for all $\omega \in (-\pi, \pi]$. If $S_U(\omega)$ is non-zero everywhere, then it implies $H(\omega) = G(\omega)$ everywhere, showing that h and g are the same filter.

References

[1] R. Ash and M. Gardner, "Topics in stochastic processes," 1975.

[2] P. J. Brockwell and R. A. Davis, *Time series: theory and methods: theory and methods.* Springer Science & Business Media, 1991.

[3] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear estimation.* Prentice Hall, 2000, no. BOOK.

[4] H. Bierens, "The wold decomposition," 2012. [Online]. Available: \url{http://www.personal.psu.edu/hxb11/WOLD.PDF}