13: THE LINEAR PREDICTION PROBLEM

Suppose we have the weakly stationary zero mean stochastic process \( \{X_t\}_{t=-\infty}^{\infty} \), with known spectrum. Let \( M^X \) be the Hilbert space generated by the linear combination of elements of \( \{X_t\} \), with the inner product \( (X, Y) = E[XY] \). Let \( M^X_t \) be the linear subspace generated by the elements \( X_s \) for \( s \leq t \). We can think of \( M^X_t \) as the linear past of the process. The \textbf{linear prediction problem} is to find the element \( \hat{X}_t \) in \( M^X_t \) which minimizes the mean squared error \( E[X_{t+v} - Y]^2 \). The solution is the best \( v \)-step linear predictor of the process. By the theory on Page 16 of Koopmans, the solution \( \hat{X}_t \) is the orthogonal projection of \( X_{t+v} \) on \( M^X_t \), defined by the following two conditions: 1) \( \hat{X}_t \in M^X_t \), and 2) \( (X_{t+v} - \hat{X}_t) \perp M^X_t \), or in other words, \( X_{t+v} - \hat{X}_t \) is orthogonal to all elements of \( M^X_t \). Ideally, we would like to express the solution \( \hat{X}_t \) explicitly as a (possibly infinite) linear combination of the present and past values of \( \{X_t\} \),

\[
\hat{X}_t = \sum_{k=0}^{\infty} d^{(v)}_k X_{t-k}.
\]

For fixed \( v \), we can think of \( \hat{X}_t \) as a time series which is the output of a linear filter with input \( \{X_t\} \), and transfer function

\[
D_v(\lambda) = \sum_{k=0}^{\infty} d^{(v)}_k e^{-i\lambda k}.
\]

Thus, we would also consider \( D_v(\lambda) \) as a "solution" to the prediction problem, since the \( d^{(v)}_k \) can be recovered from \( D_v(\lambda) \) by Fourier methods.

**Autoregressive Case**

The solution to the prediction problem is simplest if \( X_t \) is autoregressive: \( \sum_{k=0}^{\infty} b_k X_{t-k} = \varepsilon_t \) with \( b_o = 1, \sum b_k^2 < \infty \) and \( \varepsilon_t \perp \varepsilon_s \) for all \( s < t \). In this case, we have

\[
X_{t+1} = -\sum_{k=1}^{\infty} b_k X_{t+1-k} + \varepsilon_{t+1},
\]

and the best one-step linear predictor is
\[
\hat{X}_t(t) = -\sum_{k=1}^{\infty} b_k X_{t-k} .
\]

This follows since clearly \(\hat{X}_t(t) \in M_t^X\), and since \(X_{t+1} - \hat{X}_t(t) = \epsilon_{t+1}\), which is orthogonal to all of the generators \(X_s\) for \(s \leq t\). From an heuristic point of view, the best predictor was obtained from the formula for \(X_{t+1}\) by replacing the unobserved random variable \(\epsilon_{t+1}\) by its expectation, zero.

For \(\nu=1\), we write
\[
X_{t+2} = -b_1 X_{t+1} + -\sum_{k=2}^{\infty} b_k X_{t+2-k} + \epsilon_{t+2} .
\]

In this case, we obtain the best two-step predictor \(\hat{X}_2(t)\) by replacing \(\epsilon_{t+2}\) by its expectation (zero) and substituting for the unobservable \(X_{t+1}\) its best predictor, \(\hat{X}_1(t)\). We obtain
\[
\hat{X}_2(t) = -b_1 \hat{X}_1(t) + -\sum_{k=2}^{\infty} b_k X_{t+2-k} .
\]

This is indeed the best two-step predictor, since
\[
X_{t+2} - \hat{X}_2(t) = b_1 (\hat{X}_1(t) - X_{t+1}) + \epsilon_{t+2} = -b_1 \epsilon_{t+1} + \epsilon_{t+2} ,
\]
which is orthogonal to \(X_s\) for all \(s \leq t\).

For general lead time \(\nu\), the best \(\nu\)-step predictor is
\[
\hat{X}_\nu(t) = -\sum_{k=1}^{\nu-1} b_k \hat{X}_{\nu-k}(t) + -\sum_{k=\nu}^{\infty} b_k X_{t+\nu-k} .
\]

We now study the transfer function \(D_\nu(\lambda)\) of the best predictor in the autoregressive case. (The main reason for doing this is to motivate the solution for general processes.) Since the autoregression has a one-sided MA representation \(X_t = \sum_{k=0}^{\infty} a_k \epsilon_{t-k}\), every linear function of \(\{X_s : s \leq t\}\) can be expressed as a linear function of \(\{\epsilon_s : s \leq t\}\). Thus, \(M_t^X \subset M_t^\epsilon\), where the Hilbert space \(M_t^\epsilon\) is the linear past of \(\{\epsilon_s : s \leq t\}\). Also, by definition of an AR process, \(\epsilon_t = \sum_{k=0}^{\infty} b_k X_{t-k}\), so that \(M_t^\epsilon \subset M_t^X\). Thus, \(M_t^X = M_t^\epsilon\) for all \(t\).

Thus, the best \(\nu\)-step predictor \(\hat{X}_\nu(t)\) can be computed as the projection of \(X_{t+\nu}\) on \(M_t^\epsilon\). Since
\[
X_{t+\nu} = \sum_{k=0}^{\infty} a_k \epsilon_{t+\nu-k} = \sum_{s=\nu}^{\infty} a_{\nu+s} \epsilon_{t-s} ,
\]
the best \( n \)-step predictor is obtained by setting the future values \( \epsilon_{t+1}, \ldots, \epsilon_{t+n} \) to zero:

\[
\hat{X}_n(t) = \sum_{s=0}^{\infty} a_{n+s} \epsilon_{t-s} .
\]

Using the spectral representation, we have

\[
\hat{X}_n(t) = \int_{-\pi}^{\pi} \exp(i \lambda t) A_n(\lambda) dZ_\epsilon(\lambda) ,
\]

where

\[
A_n(\lambda) = \sum_{s=0}^{\infty} a_{n+s} e^{-i \lambda t} .
\]

By the moving average representation,

\[
X_t = \int_{-\pi}^{\pi} \exp(i \lambda t) A(\lambda) dZ_\epsilon(\lambda) ,
\]

where \( A(\lambda) = A_n(\lambda) \). Thus,

\[
dZ_X(\lambda) = A(\lambda) dZ_\epsilon(\lambda) ,
\]

and (1) yields

\[
\hat{X}_n(t) = \int_{-\pi}^{\pi} \exp(i \lambda t) [A_n(\lambda)/A(\lambda)] dZ_X(\lambda) ,
\]

so the best predictor is the result of passing \( X_t \) through a linear filter with transfer function \( D_n(\lambda) = A_n(\lambda)/A(\lambda) \). This is the required solution. It can be shown that the resulting prediction error is

\[
E [X_{t+n} - \hat{X}_n(t)]^2 = \sigma^2 \sum_{k=0}^{n-1} a_k^2 ,
\]

where \( \sigma^2 = \text{Var} [\epsilon_t] \).

**Transfer Function Of Best Predictor, General Case**

Using the same technique, the transfer function of the best predictor can be obtained for any weakly stationary process \( X_t \) with a one-sided moving average representation such that \( M_{n} = M_{n} \). In practice, aside from a perfectly predictable component, \( X_t \) can be expressed in this way, according to
Wold’s Theorem:

Any zero-mean weakly stationary process $X_t$ which is not perfectly predictable can be expressed as $X_t = U_t + V_t$, where

1) $\{U_t\}$ and $\{V_t\}$ are uncorrelated processes,

2) $\{U_t\}$ has a one-sided moving average representation $U_t = \sum_{k=0}^{\infty} a_k \varepsilon_{t-k}$, $a_o = 1$, $M^U_t = M^U_t$, and

3) $M^V_s = M^V_t$ for all $s, t$, so that $\{V_t\}$ is perfectly predictable.

$\{U_t\}$ is said to be nondeterministic (i.e., not perfectly predictable), and $\{V_t\}$ is said to be deterministic. The best predictor is

$$\hat{X}_v(t) = \hat{U}_v(t) + \hat{V}_v(t).$$

In practice, the deterministic part $V_t$ is usually due to the discrete component of the spectrum. In this case, every realization of $V_t$ is just a superposition of (a countable number of) pure sine and cosine waves, which are relatively easy to detect and remove. In this way, we can recover $U_t$ from the raw data $X_t$. Then we can predict $U_t$, and add this to a (perfect) prediction for $V_t$ to yield a prediction for $X_t$. So the main problem is how to predict $U_t$.

**Determination of the Moving Average Transfer Function, $A(\lambda)$**

We assume that $\{X_t\}$ is purely nondeterministic (i.e., $V_t = 0$), so that $\{X_t\}$ has a one-sided moving average representation. Then we can determine the transfer function $D_v(\lambda)$ of the best predictor exactly as before. To carry out this procedure, however, we need to know the coefficients of the one-sided moving average representation, $X_t = \sum_{k=0}^{\infty} a_k \varepsilon_{t-k}$, the transfer function $A(\lambda) = \sum_{k=0}^{\infty} a_k \exp(-i \omega k)$ and the white noise variance $\sigma^2 = \text{Var}[\varepsilon_t]$. Since there is no deterministic component, $X_t$ must have a continuous spectrum with spectral density $f_X(\lambda)$. The problem, then, is to find constants $\{a_k\}$ and $\sigma^2$ such that

$$f_X(\lambda) = \frac{\sigma^2}{2\pi} \left| \sum_{k=0}^{\infty} a_k \exp(-i \lambda k) \right|^2 = \frac{\sigma^2}{2\pi} |A(\lambda)|^2,$$

where $a_o = 1$ and $\sum_{k=0}^{\infty} a_k^2 < \infty$. This procedure is called spectral factorization, and if it can be carried out, then the one-sided moving average is $X_t = \sum_{k=0}^{\infty} a_k \varepsilon_{t-k}$ with respect to some white noise process $\varepsilon_t$. 
Furthermore, if $A(\lambda)>0$ for all $\lambda$, $X_t$ will have the autoregressive representation

$$
\epsilon_t = \int_{-\pi}^{\pi} \exp(i\lambda t)[1/A(\lambda)]dZ_{\lambda}(\lambda) \ .
$$

The innovation variance is equal to the minimum mean squared error of prediction, and is given by Kolmogorov’s Formula as

$$
\sigma^2 = 2\pi \exp\left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \log f_X(\lambda) d\lambda \right\} \ .
$$

This formula is very important. It holds even if the process does not have a moving average representation. In this case, $\sigma^2$ is taken to represent the minimum mean squared error of linear prediction, based on the infinite past. We will not pursue the details of spectral factorization here, but we note that Kolmogorov showed how to carry out the factorization as long as

$$
\int_{-\pi}^{\pi} \log f_X(\lambda) > -\infty \ ,
$$

in other words, as long as $\sigma^2>0$. But the condition that $\sigma^2>0$ is a consequence of the assumption, which we made at the outset, that $\{X_t\}$ has a one-sided moving average representation.

We have now shown how to predict any nondeterministic process whose deterministic component (if any) has a purely discrete spectrum. Unfortunately, not all deterministic processes have purely discrete spectra. In fact, it is possible for a deterministic process to have a spectral density! For example, suppose that $\{X_t\}$ has spectral density $f(\lambda) = e^{-1/|\lambda|}$ for $\lambda \in [-\pi, \pi]$. Then

$$
\int_{-\pi}^{\pi} \log f(\lambda) d\lambda = -\infty ,
$$

so that $\sigma^2=0$, by Kolmogorov’s Formula. Thus, $\{X_t\}$ is deterministic. Since it has no discrete components in its spectral distribution, $\{X_t\}$ will not be just a superposition of (a countable number of) sine and cosine waves. Furthermore, $\{X_t\}$ cannot be written as a one-sided moving average.