Massachusetts Institute of Technology | Guido Kuersteiner
Department of Economics
Time Series 14.384

Lecture Note 4: Prediction and Wold Decomposition

We consider a weakly stationary time series x_t and are interested in obtaining a forecast of x_{t+1} based on past observed values of x_t. It is common to consider forecasts $\hat{x}_{t+1}$ that minimize the mean squared forecast error.

$$E\left(x_{t+1}-\hat{x}_{t+1}\right)^2=\inf_{y\in\mathcal{M}_t}E\left(x_{t+1}-y\right)^2$$

where $\mathcal{M}_t$ is the set of all measurable functions of $\{x_t,...,x_1\}$ such that $y\in\mathcal{M}_t$ iff Ey^2 and includes the constant functions[1]. By the projection theorem

$$\hat{x}_{t+1}=P_{\mathcal{M}_t}(x_{t+1})=E_{\mathcal{M}_t}x_{t+1}$$

where $E_{\mathcal{M}_t}x_{t+1}$ is the conditional expectation defined by

$$EWE_{\mathcal{M}_t}X=EWX\ \ \forall W\in\mathcal{M}_t$$

where X is any random variable defined on the same sample space as x_t.

It follows at once that $x_{t+1}-E_{\mathcal{M}_t}x_{t+1}\in\mathcal{M}_t^{\perp}$ since for all $y\in\mathcal{M}_t$ it has to hold from the definition of the conditional expectations that $E\left(y\left(x_{t+1}-E_{\mathcal{M}_t}x_{t+1}\right)\right)=Eyx_{t+1}-EyE_{\mathcal{M}_t}x_{t+1}=0$. By the projection theorem this establishes that the conditional expectation is a projection. This result is not very useful in practice because the conditional expectation cannot in general be computed. It is therefore useful to restrict attention to the class of best linear predictors. We denote the closed linear span of $\{1,x_t,\dots,x_1\}$ by $\mathcal{M}_t^{\ell}$. Then the best linear predictor satisfies

$$E(x_{t+1}-\hat{x}_{t+1})^2=\inf_{y\in\mathcal{M}_t^{\ell}}E\left(x_{t+1}-y\right)^2$$

and by the projection theorem we have again that $\hat{x}_{t+1}=P_{\mathcal{M}_t^{\ell}}(x_{t+1})=\mu+\sum_{j=0}^{t-1}\phi_{j,t}x_{t-j}$ such that

$$\sum_{j=0}^{t-1}\phi_{j,t}\operatorname{cov}\left(x_{t-j},x_{t-i}\right)=\operatorname{cov}\left(x_{t+1},x_{t-i}\right)\quad i=0,\dots,t-1$$

since $\overline{sp}\{1,x_t,\dots,x_1\}\subset\mathcal{M}_t$ it follows immediately that in general

$$E\left(x_t-E_{\mathcal{M}_t}(x_{t+1})\right)^2\le E\left(x_{t+1}-P_{\mathcal{M}_t^{\ell}}(x_{t+1})\right)^2$$

The only exception is the case where x_t is a Gaussian process. Then it is the case that

$$E_{\mathcal{M}_t}(x_{t+1})=P_{\mathcal{M}_t^{\ell}}(x_{t+1}).$$

In particular we write for the best linear predictor

$$\hat{x}_{t+1}=P_{\mathcal{M}_t^{\ell}}x_{t+1}\qquad\text{for } t\ge 1.$$

Note that $\mathcal{M}_t^{\ell}=\overline{sp}\left\{x_t,...,x_1\right\}=sp\left\{x_t-\hat{x}_t,...,x_1-\hat{x}_1\right\}$. Therefore x_{t+1} can be found by projecting onto the past forecast errors $\{x_t-\hat{x}_t,...,x_1-\hat{x}_1\}$. We can define $\hat{x}_{t+1}$ recursively by setting

$$\hat{x}_1=0$$

[1] More formally, $\mathcal{M}_t$ is the σ-field generated by $\{x_t,....,x_1\}$, i.e. $\mathcal{M}_t$ is the smallest sigma field of the sample space Ω such that $x_t,....,x_1$ are measurable functions.

such that

$$\hat{x}_{t+1} = \sum_{j=0}^{t-1} \theta_{j,t} \left(x_{t-j} - \hat{x}_{t-j} \right) \qquad t > 1$$

where

$$\sum_{j=0}^{t-1} \theta_{j,t} \left\langle x_{t-j} - \hat{x}_{t-j}, x_{t-i} - \hat{x}_{t-i} \right\rangle = \left\langle x_{t+1}, x_{t-i} - \hat{x}_{t-i} \right\rangle.$$

Since $x_{t-j} - \hat{x}_{t-j} \in \mathcal{M}_{t-j-1}^{\ell\perp}$ by the projection theorem, the left-hand side reduces to

$$\theta_{i,t} \left\| x_{t-i} - \hat{x}_{t-i} \right\|^2 = \left\langle x_{t+1}, x_{t-i} - \hat{x}_{t-i} \right\rangle \tag{4.1}$$

Denoting $\left\| x_{t-i} - \hat{x}_{t-i} \right\|^2 = \sigma_{t-i}^2$ and substituting for $\hat{x}_{t-i} = \sum_{j=0}^{t-2-i} \theta_{j,t-i-1} \left(x_{t-i-j-1} - \hat{x}_{t-i-j-1} \right)$ leads to

$$\theta_{i,t} = \sigma_{t-i}^{-2} \left[\gamma_x(i+1) - \sum_{j=0}^{t-2-i} \theta_{j,t-i-1} \left\langle x_{t+1}, x_{t-i-j-1} - \hat{x}_{t-i-j-1} \right\rangle \right] \tag{4.2}$$

where $cov(x_t, x_{t+|h|}) = \gamma_x(h)$. Now, the last term in the sum, $\left\langle x_{t+1}, x_{t-i-j-1} - \hat{x}_{t-i-j-1} \right\rangle$, which is equal to $\sigma_{t-i-j-1}^2 \theta_{j+i+1,t}$ form (4.1), can be substituted in (4.2) to give

$$\theta_{i,t} = \sigma_{t-i}^{-2} \left[\gamma_x(i+1) - \sum_{j=0}^{t-2-i} \theta_{j,t-i-1} \theta_{j+i+1,t} \sigma_{t-i-j-1}^2 \right].$$

Also, $\sigma_t^2 = \left\| x_t - \hat{x}_t \right\|^2 = \left\| x_t \right\|^2 - \left\| \hat{x}_t \right\|^2 = \gamma_x(0) - \sum_{j=0}^{t-1} \theta_{j,t-1}^2 \sigma_{t-j-1}^2$ and $\sigma_1^2 = \gamma_x(0)$. Note that $\theta_{t-1,t} = \sigma_1^{-2} \gamma_x(1)$ assuming that $\gamma_x(h)$ is known or estimated. These equations show that all the coefficients can be calculated recursively.

h-step ahead prediction

We now want to predict x_{t+h} based on $x_t, ..., x_1$. The linear predictor is

$$\begin{aligned} \hat{x}_{t+h} &= P_{\mathcal{M}_t^\ell}(x_{t+h}) \\ &= \sum_{j=0}^{t-1} \theta_{j,t+h-1} \left(x_{t-j} - \hat{x}_{t-j} \right). \end{aligned}$$

We want to apply these general results to forecast x_{t+h} if x_t is assumed to follow an ARMA(p, q) process of the form

$$\phi(L) x_t = \theta(L) \varepsilon_t \qquad \varepsilon_t \sim WN(0, \sigma^2)$$

where $\phi(L) = 1 - \phi_1 L ... - \phi_p L^p$ and $\theta(L) = 1 + \theta_1 L + ... + \theta_q L^q$. Define

$$w_t = \sigma^{-1} \phi(L) x_t \qquad t > \max(p, q)$$

and let

$$w_t = \sigma^{-1} x_t \qquad \text{for } t \leq \max(p, q).$$

Note that $\overline{sp} \{x_t, ..., x_1\} = \overline{sp} \{w_t, ..., w_1\}$. We determine $\hat{w}_t$ recursively as before, i.e.

$$\begin{aligned} \hat{w}_1 &= 0 \\ \hat{w}_{t+1} &= \sum_{j=0}^{t-1} \theta_{j,t} \left(w_{t-j} - \hat{w}_{t-j} \right). \end{aligned}$$

Denoting $\sigma_t^2 = \|w_t - \hat{w}_t\|^2$ we can now determine the coefficients $\theta_{j,t}$ as

$$\theta_{j,t}\sigma_{t-j}^2 = \langle w_{t+1}, w_{t-j} - \hat{w}_{t-j}\rangle . \tag{4.3}$$

Now for $t > \max(p,q)$ and $j > q$ it follows that $w_{t-j} - \hat{w}_{t-j} \in \mathcal{M}_{t-j}^{\ell}$ and $w_{t+1} \in \mathcal{M}_{t-j}^{\ell\perp}$. Therefore $\theta_{j,t} = 0$ for $t > \max(p,q)$ and $j > q$. From (4.3) we have now for $t > \max(p,q)$ and $j < q$

$$\theta_{j,t} = \sigma_{t-j}^{-2}\left[\langle w_{t+1}, w_{t-j}\rangle - \sum_{k=0}^{q-j-1} \theta_{k,t-j-1}\theta_{k+j+1,t}\sigma_{t-k-j-1}^2\right]$$

and $\theta_{q,t} = \sigma_{t-q}^{-2}\langle w_{t+1}, w_{t-q}\rangle$. Also,

$$\sigma_{t-j}^2 = \|w_{t-j}\|^2 - \sum_{k=0}^{q-1} \theta_{k,t-j-1}^2\sigma_{t-j-1-k}^2$$

and noting that $Ew_iw_j = \gamma_w(i-j)$ with

$$\gamma_w(i-j) = \begin{cases} \sigma^{-2}\gamma_x(i-j) & 1 \le i,j \le m \\ \sigma^{-2}\left[\gamma_x(i-j) - \sum_{r=1}^{p}\phi_r\gamma_x(r-|i-j|)\right] & \min(i,j) \le m < \max(i,j) \le 2m \\ \sum_{r=0}^{q}\theta_r\theta_{r+|i-j|} & \min(i,j) > m \\ 0 & \text{otherwise.} \end{cases}$$

These relationships allow for recursive estimation of $\theta_{j,t}$, σ_t^2 and $\hat{w}_t$. For large t and $\theta(L)$ invertible $\hat{w}_t$ can be approximately determined by using the parameters θ_j of the lag polynomial $\theta(L)$ instead of the optimal projection parameters. The predictions for x_t are now obtained from

$$\begin{aligned} \hat{w}_t &= \sigma^{-1}P_{\mathcal{M}_{t-1}^{\ell}}x_t \\ &= \sigma^{-1}\hat{x}_t \qquad t < \max(p,q) \end{aligned}$$

and

$$\begin{aligned} \hat{w}_t &= \sigma^{-1}P_{\mathcal{M}_{t-1}^{\ell}}\phi(L)x_t \\ &= \sigma^{-1}(\hat{x}_t - \phi_1x_{t-1} - ... - \phi_px_{t-p}) \qquad t \ge \max(p,q). \end{aligned}$$

It follows that $\sigma(w_t - \hat{w}_t) = x_t - \hat{x}_t$. Therefore

$$\begin{aligned} \hat{x}_{t+1} &= \sum_{j=0}^{t}\theta_{j,t}\left(x_{t-j} - \hat{x}_{t-j}\right) \qquad t < \max(p,q) \\ \hat{x}_{t+1} &= \phi_1x_t + ... + \phi_px_{t-p+1} + \sum_{j=0}^{q}\theta_{j,t}\left(x_{t-j} - \hat{x}_{t-j}\right) \qquad t \ge \max(p,q) \end{aligned}$$

It follows immediately that for $x_t \sim$ARMA$(p,0)$

$$\hat{x}_{t+1} = \phi_1x_t + ... + \phi_px_{t-p+1}.$$

The h-step ahead predictor can be found iteratively

$$\begin{aligned} \hat{x}_{t+2} &= P_{\mathcal{M}_t^{\ell}}x_{t+2} \\ &= \phi_1P_{\mathcal{M}_t^{\ell}}x_{t+1} + \phi_2x_t + ... + \phi_px_{t-p+2} \\ &\quad + \sum_{j=h}^{q}\theta_{j,t+1}\left(x_{t+2-j} - \widehat{x}_{t+2-j}\right), \end{aligned}$$

so in particular for the AR(p) model

$$\hat{x}_{t+h} = \phi_1\hat{x}_{t+h-1} + \phi_2\hat{x}_{t+h-2} + ... + \phi_p\hat{x}_{t+h-p+1} \qquad h > p-1.$$

4.1. The Wold Decomposition

We show that a mean zero stationary process x_t can be decomposed into a perfectly predictable component and a $MA(\infty)$ process with white noise innovations.

Let $\mathcal{M}_t = \overline{sp}\{x_s, s \leq t\}$ and define the one-step mean square prediction error as

$$\sigma^2 = E\left(x_t - P_{\mathcal{M}_{t-1}} x_t\right)^2. \tag{4.4}$$

Also let $\mathcal{M}_{-\infty} = \bigcap_{t=-\infty}^{\infty} \mathcal{M}_t$ such that $\mathcal{M}_{-\infty}$ is a closed linear subspace of $\mathcal{M} = \overline{sp}\{x_t, t \in \mathbb{Z}\}$. We call a process x_t deterministic if $x_t \in \mathcal{M}_{-\infty}$. For a deterministic process the forecast error variance is

$$E\left(x_t - P_{\mathcal{M}_{t-1}} x_t\right)^2 = E\left(x_t - x_t\right)^2 = 0$$

since $x_t \in \mathcal{M}_{-\infty} \subset \mathcal{M}_{t-1}$. We prove the Wold decomposition theorem.

Theorem 4.1 (Wold Decomposition). *If x_t is weakly stationary and mean zero with $\sigma^2 > 0$ as defined in (4.4) then*

$$x_t = \sum \psi_j \varepsilon_{t-j} + v_t \tag{4.5}$$

with

i) $\varepsilon_t \sim WN\left(0, \sigma^2\right)$,

ii) $E\left(\varepsilon_t v_s\right) = 0 \quad \forall\, t, s$,

iii) $v_t \in \mathcal{M}_{-\infty}$,

iv) $\sum \psi_j^2 < \infty$.

v) v_t is deterministic.

Proof. Let

$$\begin{aligned} \varepsilon_t &= x_t - P_{\mathcal{M}_{t-1}} x_t \\ \sigma^2 \psi_j &= \langle x_t, \varepsilon_{t-j} \rangle \\ v_t &= x_t - \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} \end{aligned}$$

We have $\varepsilon_t \in \mathcal{M}_t$ and $\varepsilon_t \in \mathcal{M}_{t-1}^{\perp}$ by the projection theorem. Therefore

$$E(\varepsilon_t \varepsilon_s) = 0 \;\forall\; t \neq s.$$

Also $E\varepsilon_t = 0$ by linearity of $P_{\mathcal{M}_{t-1}} x_t$ and stationarity of x_t. Again by linearity of $P_{\mathcal{M}_{t-1}} x_t$ and weak stationarity of x_t we have $E\varepsilon_t^2 = E(x_t - P_{\mathcal{M}_{t-1}} x_t)^2 = \sigma^2$ independent of t. This shows that ε_t is $WN(0, \sigma^2)$.

Also let $\mathcal{H}_t = \overline{sp}\{\varepsilon_t, \varepsilon_{t-1}, ...\}$. Then $\mathcal{H}_t$ has a countably infinite orthogonal basis ε_t. The projection of x_t onto $\mathcal{H}_t$ is then given by

$$P_{\mathcal{H}_t} x_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}.$$

To show this let $y_t = P_{\mathcal{H}_t} x_t$ such that by the definition of the projection operator $y_t \in \mathcal{H}_t$. It now follows that for every $\epsilon > 0$ and some $k < \infty$

$$\left\| y_t - \sum_{j=0}^{k} \langle y_t, \varepsilon_{t-j} \rangle \varepsilon_{t-j} \right\|^2 = \sum_{j=k+1}^{\infty} \left|\langle y_t, \varepsilon_{t-j} \rangle\right|^2 < \epsilon.$$

To see this first note that $\|y_t\|^2 \leq \|x_t\|^2$ by the projection theorem. Then by Bessel's inequality

$$\sum_{j=0}^{k} |\langle y_t, \varepsilon_{t-j}\rangle|^2 \leq \|y_t\|^2 \text{ for all } k$$

which proves the above inequality. Next note that

$$\langle y_t, \varepsilon_{t-j}\rangle = \langle y_t - x_t, \varepsilon_{t-j}\rangle + \langle x_t, \varepsilon_{t-j}\rangle = \langle x_t, \varepsilon_{t-j}\rangle = \sigma^2\psi_j$$

since $y_t - x_t$ is orthogonal to $\mathcal{H}_t$. We have therefore established that

$$\left\| y_t - \sum_{j=0}^{k} \psi_j \varepsilon_{t-j} \right\|^2 < \epsilon$$

and that

$$\sum_{j=0}^{\infty} \psi_j^2 < \infty.$$

Then

$$\begin{aligned} Ev_t\varepsilon_s &= E\left(x_t - \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \varepsilon_s\right) \\ &= Ex_t\varepsilon_s - \psi_j E\varepsilon_s^2 \\ &= Ex_t\varepsilon_s - \frac{Ex_t\varepsilon_s}{\sigma^2}\sigma^2 = 0 \end{aligned}$$

for $s \leq t$ and for $s > t$ $\varepsilon_s \in \mathcal{M}_{s-1}^{\perp} \subset \mathcal{M}_t^{\perp}$, but $v_t \in \mathcal{M}_t$ so $Ev_t\varepsilon_s = 0$ again. From $v_t \in \mathcal{M}_t = \mathcal{M}_{t-1} \oplus \overline{sp}\{\varepsilon_t\}$ and $Ev_t\varepsilon_t = 0$ it follows $v_t \in \mathcal{M}_{t-1}$. Repeating the same argument and using $Ev_t\varepsilon_{t-j} = 0$ leads to $v_t \in \mathcal{M}_{t-j}$ thus $v_t \in \bigcap_{j=0}^{\infty} \mathcal{M}_{t-j} \subseteq \mathcal{M}_{-\infty}$. Then

$$\overline{sp}\{v_j, j \leq t\} \subseteq \mathcal{M}_{-\infty}.$$

From $x_t = v_t + \sum \psi_j \varepsilon_{t-j}$ we have

$$\mathcal{M}_t = \mathcal{H}_t \oplus \overline{sp}\{v_j, j \leq t\}. \tag{4.6}$$

Finally, if $z \in \mathcal{M}_{-\infty} \cap \mathcal{M}_t = \mathcal{M}_{-\infty}$ then $z \in \mathcal{M}_{s-1}$ such that $\langle z, \varepsilon_s\rangle = 0$ for all s. But this shows that $z \in \mathcal{H}_t^{\perp}$ or $z \in \overline{sp}\{v_j, j \leq t\}$ by (4.6) such that$\mathcal{M}_{-\infty} \subseteq \overline{sp}\{v_j, j \leq t\}$ implying that

$$\mathcal{M}_{-\infty} = \overline{sp}\{v_j, j \leq t\} \text{ for all } t.$$

This means that v_j is deterministic, i.e. the prediciton error variance is zero.

■

A process is said to be purely non-deterministic if $\mathcal{M}_{-\infty} = \{0\}$. In this case the Wold decomposition is of the form

$$x_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$$

where ψ_j and ε_{t-j} are as defined before. Processes in this class include the $ARMA(p, q)$ model introduced before.

The h-step ahead predictor of (4.5) is given by

$$P_{\mathcal{M}_t} x_{t+h} = \sum_{j=h}^{\infty} \psi_j \varepsilon_{t+h-j} + v_{t+h}$$

since $\varepsilon_{t+k} \perp \mathcal{M}_t$ for $k > 0$. It also follows that the variance of the prediction error is given by

$$\begin{aligned} \|x_{t+h} - P_{\mathcal{M}_t} x_{h+1}\|^2 &= \left\| \sum_{j=0}^{h-1} \psi_j \varepsilon_{t+h-j} \right\|^2 \\ &= \sigma^2 \sum_{j=0}^{h-1} \psi_j^2. \end{aligned}$$

As $h \to \infty$ the variance of the prediction error tends to the variance of x_t.