Lecture 5

Prediction/Forecasting

Reminder on Hilbert spaces and projections

Given a measure space $(\Omega, \mathcal{U}, \mu)$, let $L_2(\Omega, \mathcal{U}, \mu)$ (or $L_2(\mu)$ for short) be the set of all measurable functions $f : \Omega \to \mathbb{C}$ (or $f : \Omega \to \mathbb{R}$) such that $\int |f|^2 d\mu < \infty$. We set

$$\langle f_1, f_2 \rangle = \int f_1 \overline{f}_2 d\mu,$$
$$\|f\| = \sqrt{\int |f|^2 d\mu}.$$

Definition 1. *f, g in $L_2(\Omega, \mathcal{U}, \mu)$ are called* orthogonal *if $\langle f, g \rangle = 0$. This is denoted $f \perp g$. Two subsets F and G of $L_2(\Omega, \mathcal{U}, \mu)$ are* orthogonal *if $\langle f, g \rangle = 0$ for every $f \in F$ and for every $g \in G$. This is denoted $F \perp G$.*

Theorem 1 (Projection theorem)**.** *Let $L \subset L_2(\Omega, \mathcal{U}, \mu)$ be a closed linear subspace. For every $f \in L_2(\Omega, \mathcal{U}, \mu)$ there exists a unique element $\Pi f \in L$ that minimizes $l \to \|f - l\|^2$ over $l \in L$. This element is uniquely determined by the requirements $\Pi f \in L$ and $f - \Pi f \perp L$.*

Given a probability space $(\Omega, \mathcal{U}, \mathbb{P})$, the space $L_2(\Omega, \mathcal{U}, \mathbb{P})$ is exactly the set of all complex (real) random variables with finite second moment $\mathbb{E}[|X|^2]$. The inner product is the product expectation, i.e. $\langle X, Y \rangle = \mathbb{E}[X\overline{Y}]$ and the norm is $\|X\| = \sqrt{\mathbb{E}[|X|^2]}$. Let $\mathcal{U}_0$ be a sub σ-field of the σ-field $\mathcal{U}$. The collection L of all $\mathcal{U}_0$-measurable variables $Y \in L_2(\Omega, \mathcal{U}, \mathbb{P})$ is a closed, linear subspace of $L_2(\Omega, \mathcal{U}, \mathbb{P})$. By the projection theorem every square-integrable random variable X possesses a projection onto L and this is the conditional expectation of X given $\mathcal{U}_0$, that is:

Theorem 2. *Let $\mathcal{U}_0$ be a sub σ-field of the σ-field $\mathcal{U}$. If $\mathbb{E}[|X|^2] < \infty$ then $Y = \mathbb{E}[X|\mathcal{U}_0]$ is a version of the orthogonal projection of X onto $L_2(\Omega, \mathcal{U}_0, \mathbb{P})$. In particular, $Y = \mathbb{E}[X|\mathcal{U}_0]$ is the best estimator in the sense of the least squares estimators:*

$$Y \text{ minimizes } \mathbb{E}[|Y' - X|^2] \quad \text{with } Y' \ \mathcal{U}_0\text{-measurable.}$$

Linear and nonlinear prediction

Let $X = (X_t)_{t \in \mathbb{Z}}$ be a weakly stationary process with mean 0 and autocovariance function c. Consider the problem of predicting the value of the process X at time t given a linear combination of the last p values in the past $X_{t-1}, \dots, X_{t-p}$.

Definition 2. *Given a mean zero time series* $X = (X_t)_{t\in\mathbb{Z}}$, *the* best linear predictor of order p *of* X_t *is the linear combination* $\phi_{1,p}X_{t-1} + \phi_{2,p}X_{t-2} + \cdots + \phi_{p,p}X_{t-p}$ *that minimizes* $\mathbb{E}[|X_t - Y|^2]$ *over all linear combinations* Y *of* $X_{t-1}, \ldots, X_{t-p}$. *The minimal value* $\mathbb{E}[|X_t - \phi_{1,p}X_{t-1} - \phi_{2,p}X_{t-2} - \cdots - \phi_{p,p}X_{t-p}|^2]$ *is called the* square prediction error.

In other words, the best linear predictor of order p of X_t, denoted by $\Pi_p X_t$, is the projection of X_t onto the linear subspace $\mathcal{H}_{t-1,p}$ spanned by $X_{t-1}, \ldots, X_{t-p}$, i.e.

$$\mathcal{H}_{t-1,p} = \mathrm{Vect}(X_{t-1}, \ldots, X_{t-p}).$$

Thanks to Theorem 1,

$$\Pi_p X_t = \sum_{k=1}^{p} \phi_{k,p} X_{t-k},$$

where the coefficients $(\phi_{k,p})_{1\le k\le p}$ satisfy

$$\langle X_t - \phi_{1,p}X_{t-1} - \cdots - \phi_{p,p}X_{t-p}, X_{t-j}\rangle = 0, \quad j = 1, \ldots, p, \tag{1}$$

where $\langle\cdot,\cdot\rangle$ is the inner product in $L_2(\Omega, \mathcal{U}, \mathbb{P})$. Equation (1) can be written as

$$\langle X_t, X_{t-j}\rangle = \sum_{k=1}^{p} \phi_{k,p}\langle X_{t-k}, X_{t-j}\rangle, \quad j = 1, \ldots, p$$

and thus, using the stationarity of X,

$$\sum_{k=1}^{p} \phi_{k,p} c(k-j) = c(j), \quad j = 1, \ldots, p. \tag{2}$$

Let $\mathbf{C}_p$ be the autocovariance matrix of the vector $(X_{t-1}, \ldots, X_{t-p})$, i.e.

$$\mathbf{C}_p = \begin{pmatrix} c(0) & c(1) & \ldots & c(p-1) \\ c(1) & \ddots & \ddots & c(p-2) \\ \vdots & \ddots & \ddots & \vdots \\ c(p-1) & \ldots & c(1) & c(0) \end{pmatrix}.$$

Then, (2) can be rewritten as

$$\mathbf{C}_p \boldsymbol{\phi}_p = \mathbf{c}_p, \tag{3}$$

where $\boldsymbol{\phi}_p = (\phi_{1,p}, \ldots, \phi_{p,p})'$ and $\mathbf{c}_p = (c(1), \ldots, c(p))'$. If $\mathbf{C}_p$ is nonsingular, then $\phi_{1,p}, \ldots, \phi_{p,p}$ can be solved uniquely. Otherwise there are multiple solutions of (3), but any solution will give the best linear predictor, as this is uniquely determined by the projection theorem.

Proposition 1. *If* $c(0) > 0$ *and if* $c(h) \to 0$ *as* $h \to \infty$ *then* $\mathbf{C}_n = (c(i-j))_{i,j=1,\ldots,n}$, *is invertible for every* n.

Proof. Admitted. □

The square prediction error can be expressed via the coefficients $\phi_{1,p}, \dots, \phi_{p,p}$ by Pythagoras' rule, which gives, for a weakly stationary process X,

$$\mathbb{E}[|X_t - \Pi_p X_t|^2] = \mathbb{E}[|X_t|^2] - \mathbb{E}[|\Pi_p X_t|^2] = c(0) - \boldsymbol{\phi}_p' \mathbf{C}_p \boldsymbol{\phi}_p. \tag{4}$$

Example 1. *Let X be a causal $AR(m)$ solution of*

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_m X_{t-m} + Z_t, \tag{5}$$

where $Z \sim WN(0, \sigma^2)$ and $\phi(z) = 1 - \sum_{k=1}^m \phi_k z^k \neq 0$ on $\{z \in \mathbb{C} : |z| \leq 1\}$. Then, the best linear predictor of order p of X, with $p \geq m$, is given by $\sum_{k=1}^p \phi_{k,p} X_{t-k}$ with

$$\phi_{k,p} = \begin{cases} \phi_k & 1 \leq k \leq m, \\ 0 & m < k \leq p. \end{cases}$$

Indeed, since X is causal, X is of the form

$$X_t = \sum_{k=0}^{\infty} \eta_k Z_{t-k}, \quad \sum_{k \in \mathbb{N}} |\eta_k| < \infty.$$

Therefore, for any $h \geq 1$, using the continuity of the scalar product in L_2, we obtain

$$\mathbb{E}[Z_t X_{t-h}] = \mathbb{E}\left[\sum_{k=0}^{\infty} \eta_k Z_t Z_{t-h-k}\right] = 0.$$

Hence, from (5), we deduce that

$$\mathbb{E}\left[\left(X_t - \sum_{k=1}^m \phi_k X_{t-k}\right) X_{t-h}\right] = \mathbb{E}[Z_t X_{t-h}] = 0, \quad \forall h \geq 1,$$

so that, for any $p \geq m$, $\sum_{k=1}^m \phi_k X_{t-k} \in \mathcal{H}_{t-1,p}$ and $(X_t - \sum_{k=1}^m \phi_k X_{t-k}) \perp \mathcal{H}_{t-1,p}$.

Linear prediction is very common in time series analysis, especially because it is very simple to use. Indeed, the linear predictor depends on the mean and autocovariance only, and in a simple way. On the other hand, utilization of general functions $f(X_{t-1}, \dots, X_{t-p})$ of the observations as predictors may decrease the prediction error. That's why sometimes, non-linear predictors are used rather than linear predictors.

Definition 3. *The* best predictor *of X_t based on $X_{t-1}, \dots, X_{t-p}$ is the function $f_p(X_{t-1}, \dots, X_{t-p})$ that minimizes $\mathbb{E}[|X_t - f(X_{t-1}, \dots, X_{t-p})|^2]$ over all measurable functions $f : \mathbb{R}^p \to \mathbb{R}$.*

In other words, the best predictor of X_t is the conditional expectation of X_t given the variables $X_{t-1}, \dots, X_{t-p}$.

Yule-Walker estimators and least square estimators for $AR(p)$

Suppose that we observe n realizations $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$ from a weakly stationary time series X with mean 0 and autocovariance function c. More precisely, suppose that $X = (X_t)_{t\in\mathbb{Z}}$ is a centered and causal autoregressive process of order p with unknown parameters $\phi_1, \ldots, \phi_p$ and σ^2, i.e.

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t, \quad Z \sim WN(0, \sigma^2). \tag{6}$$

Our goal is the estimation of the parameters $\phi_1, \ldots, \phi_p$ and σ^2 from the data.

Since X is causal, thanks to Theorem 3 in Lecture 3 we have

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \tag{7}$$

where $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{1}{\phi(z)}$, $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$, $|z| \leq 1$. Therefore, for any $j = 0, \ldots, p$,

$$\mathbb{E}\left[\left(X_t - \sum_{i=1}^{p} \phi_i X_{t-i}\right) X_{t-j}\right] = \mathbb{E}[Z_t X_{t-j}] = \mathbb{E}\left[\sum_{k=0}^{\infty} \psi_k Z_{t-j-k} Z_t\right].$$

Thus, observing that $\psi_0 = 1$ (and using the continuity of the scalar product il L_2), we get

$$c(j) - \sum_{i=1}^{p} \phi_i c(j-i) = \sum_{k=0}^{\infty} \psi_k \mathbb{E}[Z_{t-j-k} Z_t] = \begin{cases} \sigma^2 & \text{if } j = k = 0 \\ 0 & \text{otherwise.} \end{cases}$$

To sum up we have

$$\mathbf{C}_p \boldsymbol{\phi}_p = \mathbf{c}_p$$

and

$$\sigma^2 = c(0) - \boldsymbol{\phi}_p' \mathbf{c}_p,$$

where C_p is the autocovariance matrix $(c(i-j))_{i,j=1,\ldots,p}$, $\boldsymbol{\phi}_p = (\phi_1, \ldots, \phi_p)'$ and $\mathbf{c}_p = (c(1), \ldots, c(p))'$. These equations, known as the *Yule-Walker equations*, express the parameters via the second moments of the observations. The *Yule-Walker estimators* $\widehat{\boldsymbol{\phi}}_p$ and $\widehat{\sigma}^2$ are defined by replacing the true autocovariances c by their sample versions $\widehat{c}_n$, namely

$$\widehat{\mathbf{C}}_p \widehat{\boldsymbol{\phi}}_p = \widehat{\mathbf{c}}_p$$

and

$$\widehat{\sigma}^2 = \widehat{c}_n(0) - \widehat{\boldsymbol{\phi}}_p' \widehat{\mathbf{c}}_p,$$

where $\widehat{\mathbf{C}}_p = (\widehat{c}_n(i-j))_{i,j=1}^{p}$ and $\widehat{\mathbf{c}}_p = (\widehat{c}_n(1), \ldots, \widehat{c}_n(p))'$.

NB: Since $\mathbb{E}[X_t] = 0$ we will consider as estimator of $c(h)$ the estimator defined as

$$\widehat{c}_n(h) = \frac{1}{n} \sum_{i=1}^{n-h} X_i X_{i+h}$$

and not the estimator $\frac{1}{n} \sum_{i=1}^{n-h} (X_i - \widehat{\mu}_n)(X_{i+h} - \widehat{\mu}_n)$.

Remark 1. *The Yule-Walker estimators come from the comparison between the empirical autocovariance and the true autocovariance function and therefore are examples of* moment estimators*, that is estimators that are defined by matching empirical and true moments or functionals of them.*

Proposition 2. *If $\widehat{c}_n(0) > 0$ then $\widehat{\mathbf{C}}_p$ is not singular.*

Proof. Admitted. □

Thanks to Proposition 2, if $\widehat{c}_n(0) > 0$ we can write

$$\widehat{\boldsymbol{\phi}}_p = \widehat{\mathbf{C}}_p^{-1}\widehat{\mathbf{c}}_p \tag{8}$$

and

$$\widehat{\sigma}^2 = \widehat{c}_n(0) - \widehat{\mathbf{c}}_p'\widehat{\mathbf{C}}_p^{-1}\widehat{\mathbf{c}}_p. \tag{9}$$

Remark 2. *Another way to obtain (9) is the following. From (4) and Example 1 (taking $m = p$) we know that $\Pi_p X_t = \sum_{k=1}^{p} \phi_k X_{t-k}$ and*

$$c(0) - \boldsymbol{\phi}_p'\mathbf{C}_p\boldsymbol{\phi}_p = \mathbb{E}[|X_t - \Pi_p X_t|^2] = \mathbb{E}[Z_t^2] = \sigma^2.$$

This suggests to take as an estimator of σ^2 the quantity

$$\widehat{c}_n(0) - \widehat{\boldsymbol{\phi}}_p'\widehat{\mathbf{C}}_p\widehat{\boldsymbol{\phi}}_p$$

that coincides with $\widehat{\sigma}^2$ as defined in (9). Indeed, from (8), we have

$$\widehat{c}_n(0) - \widehat{\boldsymbol{\phi}}_p'\widehat{\mathbf{C}}_p\widehat{\boldsymbol{\phi}}_p = \widehat{c}_n(0) - \widehat{\boldsymbol{\phi}}_p'\widehat{\mathbf{C}}_p\widehat{\mathbf{C}}_p^{-1}\widehat{\mathbf{c}}_p = \widehat{c}_n(0) - \widehat{\boldsymbol{\phi}}_p'\widehat{\mathbf{c}}_p = \widehat{c}_n(0) - \widehat{\mathbf{c}}_p'\widehat{\mathbf{C}}_p^{-1}\widehat{\mathbf{c}}_p = \widehat{\sigma}^2.$$

Remark 3. *Suppose that the data set that we have at our disposal consists of n observations, $x_1, \ldots, x_n$, (assumed to come) from a centered weakly stationary time series with autocovariance function c. Provided that $\widehat{c}_n(0) > 0$ we can propose as a model to fit the data an autoregressive process of order $m < n$ of the form*

$$X_t - \widehat{\phi}_1 X_{t-1} - \cdots - \widehat{\phi}_m X_{t-m} = Z_t, \quad Z \sim WN(0, \widehat{\sigma}_m^2),$$

where from (8) and (9),

$$\widehat{\boldsymbol{\phi}}_m := (\widehat{\phi}_1, \ldots, \widehat{\phi}_m)' = \widehat{\mathbf{C}}_m^{-1}\widehat{\mathbf{c}}_m$$

and

$$\widehat{\sigma}_m^2 = \widehat{c}_n(0) - \widehat{\boldsymbol{\phi}}_m'\widehat{\mathbf{c}}_m.$$

A natural question is then how to efficiently compute the vector $\widehat{\boldsymbol{\phi}}_m$ and $\widehat{\sigma}_m^2$, that is how to bypass the matrix inversion required in the direct computation of the Yule-Walker estimators. There are different options: for instance, the Durbin Levinson algorithm or the innovations algorithm (a possible reference is Chapter 8 in [1]).

Another classical way to estimate the parameters $\phi_1, \ldots, \phi_p$ and σ^2 in (6) is to use the fact that the true values $\phi_1, \ldots, \phi_p$ minimize the expectation

$$(\beta_1, \ldots, \beta_p) \to \mathbb{E}[(X_t - \beta_1 X_{t-1} - \cdots - \beta_p X_{t-p})^2].$$

The *least squares estimators* are the empirical version of this criterion, namely we define $\widehat{\phi}_1, \ldots, \widehat{\phi}_p$ as the minimizing of the function

$$(\beta_1, \ldots, \beta_p) \to \frac{1}{n} \sum_{t=p+1}^{n} \big(X_t - \beta_1 X_{t-1} - \cdots - \beta_p X_{t-p}\big)^2. \tag{10}$$

The minimum value itself is a reasonable estimator of $\mathbb{E}[Z_t^2] = \sigma^2$. The least squares estimators $\widehat{\phi}_j$ obtained in this way are not identical to the Yule-Walker estimators but the difference is small. Indeed, let $\boldsymbol{\beta}_p = (\beta_1, \ldots, \beta_p)'$, $Y_n = (X_n, \ldots, X_{p+1})'$ and

$$D_n = \begin{pmatrix} X_{n-1} & X_{n-2} & \ldots & X_{n-p} \\ X_{n-2} & X_{n-3} & \ldots & X_{n-p-1} \\ \vdots & \vdots & & \vdots \\ X_p & X_{p-1} & \ldots & X_1 \end{pmatrix}.$$

Then, the right hand side of (10) is equal to $\frac{1}{n}\|Y_n - D_n \beta_p\|^2$ (here $\|\cdot\|$ stands for the Euclidian norm, i.e. if $A \in \mathbb{R}^p$, then $\|A\| = \sqrt{\sum_{i=1}^{p} |a_i|^2}$) which is minimized by the vector $\boldsymbol{\beta}_p$ such that $D_n \boldsymbol{\beta}_p$ is the projection of Y_n onto the range of the matrix D_n. Therefore, by the projection theorem, $\boldsymbol{\beta}_p$ is such that $D_n'(Y_n - D_n \boldsymbol{\beta}_p) = 0$. Solving in $\boldsymbol{\beta}_p$ one finds that the minimizing vector is given by

$$\widehat{\boldsymbol{\phi}}_p = \left(\frac{1}{n} D_n' D_n\right)^{-1} \frac{1}{n} D_n' Y_n.$$

Observe that, for any $s, t \in \{1, \ldots, p\}$,

$$\left(\frac{1}{n} D_n' D_n\right)_{s,t} = \frac{1}{n} \sum_{j=p+1}^{n} X_{j-s} X_{j-t} \approx \widehat{c}_n(s-t) = (\widehat{\mathbf{C}}_p)_{s,t},$$

$$\left(\frac{1}{n} D_n' Y_n\right)_{t} = \frac{1}{n} \sum_{j=p+1}^{n} X_{j-t} X_j \approx (\widehat{\mathbf{c}}_p)_t,$$

that is the least square estimators are nearly identical to the Yule-Walker estimators. More precisely, they possess the same (normal) limit distribution.

Theorem 3. *Let X be a centered causal $AR(p)$ weakly stationary process with $Z \sim IID(0, \sigma^2)$. Then both the Yule-Walker and the least squares estimators satisfy*

$$\sqrt{n}(\widehat{\boldsymbol{\phi}}_p - \boldsymbol{\phi}_p) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2 \mathbf{C}_p^{-1}),$$

where $\mathbf{C}_p$ is the covariance matrix of $(X_1, \ldots, X_p)$.

References

[1] Brockwell, P. and Davis, R. Time Series: Theory and Methods, Springer, 2006.