Chapter 3 - Forecasting and Impulse Response for Stationary processes

Often we are interested in time series because we want to answer one of the two following questions:

Impulse response: What is the consequence on Y_t of a shock that took place $t - j$periods ago?

Forecasting: What value do you expect for Y_{t+1} if you observed $Y_1, ..., Y_t$?

We first address these questions in the case of stationary processes.

Wold Decomposition

Any stationary process Y_t may be represented in the form

$$Y_t = \kappa_t + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$$

where

$$\psi_0 = 1, \ \sum_{j=0}^{\infty} \psi_j^2 < \infty$$

and ε_t, the error made in forecasting Y_t on the basis of a linear function of lagged Y (i.e., $\widehat{E}\left(Y_t | Y_{t-1}, ...\right)$)

$$\varepsilon_t = Y_t - \widehat{E}\left(Y_t | Y_{t-1}, ...\right)$$

is such that, for any t,

$$\begin{aligned} E\left(\varepsilon_t\right) &= 0, \, E\left(\varepsilon_t^2\right) = \sigma^2, \\ E\left(\varepsilon_t \varepsilon_\tau\right) &= 0 \text{ if } \tau \neq t \end{aligned}$$

κ_t is the linearly deterministic component of Y_t: it can be predicted arbitrarily well as a linear function of past Y, i.e.

$$\kappa_t = \widehat{E}\left(\kappa_t \middle| Y_{t-1}, ...\right)$$

and it is such that

$$E\left(\kappa_t \varepsilon_{t-j}\right) = 0 \ \forall j$$

Impulse response

The plot of $\frac{\partial Y_t}{\partial \varepsilon_{t-j}}$ (against j) is called Impulse Response Function (IRF): this is the effect on Y_t of a shock that took place $t-j$ periods before.

For a process Y_t that admits

$$Y_t = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$$

for ε_t such that, for any t,

$$\begin{aligned} E(\varepsilon_t) &= 0,\ E(\varepsilon_t^2) = \sigma^2, \\ E(\varepsilon_t \varepsilon_\tau) &= 0 \text{ if } \tau \neq t \end{aligned}$$

notice that

$$\frac{\partial Y_t}{\partial \varepsilon_{t-j}} = \psi_j$$

and the IRF is a plot of ψ_j against j.

Linear filters

From the Wald decomposition, we understand that any stationary process may be seen as at the application of the polynomial $(\psi_0, \psi_1, ...)$ to the white noise process $\{\varepsilon_t\}$.
Using the Lag Operator, we can represent this with the notation

$$\Psi(L) = \left(\psi_0 + \psi_1 L + \psi_2 L^2 + \psi_3 L^3 ...\right)$$

so that

$$\Psi(L)\varepsilon_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$$

This is a *filter*, and it is *linear* because ε_{t-j} is always with power 1.

Sometimes data are treated (by nature or by the researcher) by summing / averaging / differencing ...
For $Y_t = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$, a filter $h(L)$ is applied as

$$X_t = h(L)\, Y_t$$

where

$$h(L) = \sum_{j=-\infty}^{\infty} h_j L^j$$

If

$$\sum_{j=-\infty}^{\infty} |h_j| < \infty, \ \sum_{j=0}^{\infty} |\psi_j| < \infty$$

then

$$X_t = \mu^* + \psi^*(L)\, \varepsilon_t$$

where

$$\mu^* = h(1)\, \mu, \ \psi^*(L) = h(L)\, \psi(L)$$

and $\sum_{j=-\infty}^{\infty} \left| \psi_j^* \right| < \infty$.

Invertibility

Consider again $Y_t = \mu + \Psi(L)\varepsilon_t$: under regularity conditions (known as Invertibility) the polynomial $\Psi(L)^{-1}$ exists.
We can then write

$$\Psi(L)^{-1} Y_t = \Psi(L)^{-1}\mu + \varepsilon_t$$

and notice that $\Psi(L)^{-1}\mu = \Psi(1)^{-1}\mu$. Denote

$$\Pi(L) = \Psi(L)^{-1} \text{ where } \Pi(L) = (\pi_0 - \pi_1 L - \pi_2 L^2 ...)$$

and notice that $\pi_0 = 1$ because $\psi_0 = 1$.

Then, $\Psi(L)^{-1}Y_t = \Psi(L)^{-1}\mu + \varepsilon_t$ can be written as

$$Y_t = \Pi(1)\mu + \sum_{j=1}^{\infty} \pi_j Y_{t-j} + \varepsilon_t$$

For example, in the MA(1) model, $Y_t = \varepsilon_t + \theta\varepsilon_{t-1}$, if $|\theta| < 1$,

$$Y_t = \Pi(1)\mu + \sum_{j=1}^{\infty} (-\theta_j)^j Y_{t-j} + \varepsilon_t$$

Forecasts based on a linear projection

Assume that

- Y_t is stationary
- $E(Y_t) = 0$

(if $E(Y_t) = \mu \neq 0$, then consider $Y_t - \mu$ instead)

(linear) forecast of Y_{t+1} using Y_t :

$$\widehat{Y}_{t+1|t} = \alpha_1^{(1)} Y_t$$

(linear) forecast of Y_{t+1} using Y_t and Y_{t-1} :

$$\widehat{Y}_{t+1|t,t-1} = \alpha_1^{(2)} Y_t + \alpha_2^{(2)} Y_{t-1}$$

(linear) forecast of Y_{t+1} using Y_t , ..., Y_{t-m+1} :

$$\widehat{Y}_{t+1|t,\ldots,t-m+1} = \alpha_1^{(m)} Y_t + \alpha_2^{(m)} Y_{t-1} + \ldots + \alpha_m^{(m)} Y_{t-m+1}$$

Which values

$$\left(\alpha_1^{(m)}, \alpha_2^{(m)}, ..., \alpha_m^{(m)}\right)'$$

characterise a linear projection?

Let

$$\begin{aligned} X_t &= (Y_t, ..., Y_{t-m+1})' \\ \alpha &= \left(\alpha_1^{(m)}, \alpha_2^{(m)}, ..., \alpha_m^{(m)}\right)' \end{aligned}$$

then α must meet

$$E\left[\left(Y_{t+1} - \alpha' X_t\right) X_t'\right] = 0'$$

(i.e., the forecast error $Y_{t+1} - \alpha' X_t$ is not correlated with X_t)

Then,

$$E\left(Y_{t+1}X_t'\right) - \alpha' E\left(X_t X_t'\right) = 0'$$
$$E\left(X_t Y_{t+1}\right) - E\left(X_t X_t'\right)\alpha = 0$$
$$\alpha = \left[E\left(X_t X_t'\right)\right]^{-1} E\left(X_t Y_{t+1}\right)$$

i.e.

$$\alpha = \begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{m-2} & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{m-3} & \gamma_{m-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma_{m-2} & \gamma_{m-3} & \cdots & \gamma_0 & \gamma_1 \\ \gamma_{m-1} & \gamma_{m-2} & \cdots & \gamma_1 & \gamma_0 \end{pmatrix}^{-1} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \cdots \\ \gamma_{m-1} \\ \gamma_m \end{pmatrix}$$

- Notice that α is exactly the parameter that we would use to predict Y using X in a conditionally gaussian model

α gives the best linear forecast in MSE sense

Proof: Consider another linear forecast $g'X_t$,

$$
\begin{aligned}
& E\left[\left(Y_{t+1} - g'X_t\right)^2\right] \\
= \; & E\left[\left(Y_{t+1} - \alpha'X_t + \alpha'X_t - g'X_t\right)^2\right] \\
= \; & E\left[\left(Y_{t+1} - \alpha'X_t\right)^2\right] \\
& +2E\left[\left(Y_{t+1} - \alpha'X_t\right)\left(\alpha'X_t - g'X_t\right)\right] \\
& +E\left[\left(\alpha'X_t - g'X_t\right)^2\right]
\end{aligned}
$$

and notice that $E\left[\left(\alpha' X_t - g' X_t\right)^2\right] \geq 0$, while

$$E\left[\left(Y_{t+1} - \alpha' X_t\right)\left(\alpha' X_t - g' X_t\right)\right] = 0$$

because

$$\begin{aligned} & E\left[\left(Y_{t+1} - \alpha' X_t\right)\left(\alpha' X_t - g' X_t\right)\right] \\ = \; & E\left[\left(Y_{t+1} - \alpha' X_t\right)\left(\alpha - g\right)' X_t\right] \\ = \; & E\left[\left(Y_{t+1} - \alpha' X_t\right) X_t'\left(\alpha - g\right)\right] \\ = \; & E\left[\left(Y_{t+1} - \alpha' X_t\right) X_t'\right]\left(\alpha - g\right) \end{aligned}$$

Thus,

$$E\left[\left(Y_{t+1} - g' X_t\right)^2\right] \geq E\left[\left(Y_{t+1} - \alpha' X_t\right)^2\right]$$

Of course, in some cases a non-linear forecast may be better. However, a linear model is usually easier to use, so it is important that, under regularity conditions (invertibility), a stationary process may be given a representation linear in Y_t.

Additional definitions for stationary processes: Partial Autocorrelation Function

For a stationary Y_t with $E(Y_t) = 0$, consider the linear projection

$$\widehat{Y}_{t+1|t,\ldots,t-m+1} = \alpha_1^{(m)} Y_t + \alpha_2^{(m)} Y_{t-1} + \ldots + \alpha_m^{(m)} Y_{t-m+1}$$

For different values of m,

$$\alpha_1^{(1)}, \alpha_2^{(2)}, \ldots, \alpha_m^{(m)}$$

are the first m partial autocorrelations.

The plot of $\alpha_j^{(j)}$ (against j) is called Partial Autocorrelation Function.

Forecasting with stationary processes: examples.

Let $\{Y_t\}_{t=-\infty}^{\infty}$ be a stationary process with

$$\begin{aligned} E(Y_t) &= \mu \\ Cov(Y_t, Y_{t+j}) &= \gamma_j \end{aligned}$$

Example 1.
Compute the best linear forecast, $\widehat{Y}_{t+1|t,...}$, assuming that

$$\mu = 10,\ \gamma_0 = 2,\ \gamma_1 = 1.2;\ Y_t = 11.$$

$$\widehat{Y}_{t+1|t} = \mu + \alpha_1^{(1)}(Y_t - \mu)$$

where

$$\alpha_1^{(1)} = [\gamma_0]^{-1}\gamma_1$$

so

$$\begin{aligned} \alpha_1^{(1)} &= (2)^{-1}1.2 = 0.6 \\ \widehat{Y}_{t+1|t} &= \mu + \alpha_1^{(1)}(Y_t - \mu) = 10 + 0.6 \times (11 - 10) = 10.6 \end{aligned}$$

Example 2.

Compute the best linear forecast, $\widehat{Y}_{t+1|t,\ldots}$, assuming that

$$\begin{aligned} \mu &= 10,\ \gamma_0 = 2,\ \gamma_1 = 1.2,\ \gamma_2 = 0.6; \\ Y_t &= 11,\ Y_{t-1} = 9 \end{aligned}$$

$$\widehat{Y}_{t+1|t,t-1} = \mu + \alpha_1^{(2)}\,(Y_t - \mu) + \alpha_2^{(2)}\,(Y_{t-1} - \mu)$$

where

$$\begin{pmatrix} \alpha_1^{(2)} \\ \alpha_2^{(2)} \end{pmatrix} = \begin{pmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_0 \end{pmatrix}^{-1} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}$$

so

$$\begin{pmatrix} \alpha_1^{(2)} \\ \alpha_2^{(2)} \end{pmatrix} = \begin{pmatrix} 2 & 1.2 \\ 1.2 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 1.2 \\ 0.6 \end{pmatrix} = \begin{pmatrix} 0.656\,25 \\ -0.093\,75 \end{pmatrix}$$

$$\begin{aligned} & \widehat{Y}_{t+1|t,t-1} \\ = \;& 10 + 0.656\,25 \times (11 - 10) - 0.093\,75 \times (9 - 10) \\ = \;& 10.\,75. \end{aligned}$$

Example 3.

Compute the best linear forecast $\widehat{Y}_{t+1|t,\ldots}$ assuming

$$\begin{aligned} \mu &= 10,\ \gamma_0 = 2,\ \gamma_1 = 1.2,\ \gamma_2 = 0.6,\ \gamma_3 = 0.3; \\ Y_t &= 11,\ Y_{t-1} = 9,\ Y_{t-2} = 9.5 \end{aligned}$$

$$\widehat{Y}_{t+1|t,t-1,t-2} = \mu + \alpha_1^{(3)}\,(Y_t - \mu) + \alpha_2^{(3)}\,(Y_{t-1} - \mu) + \alpha_3^{(3)}\,(Y_{t-9} - \mu)$$

where

$$\begin{pmatrix} \alpha_1^{(3)} \\ \alpha_2^{(3)} \\ \alpha_3^{(3)} \end{pmatrix} = \begin{pmatrix} \gamma_0 & \gamma_1 & \gamma_2 \\ \gamma_1 & \gamma_0 & \gamma_1 \\ \gamma_2 & \gamma_1 & \gamma_0 \end{pmatrix}^{-1} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix}$$

Since

$$\begin{pmatrix} 2 & 1.2 & 0.6 \\ 1.2 & 2 & 1.2 \\ 0.6 & 1.2 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 1.2 \\ 0.6 \\ 0.3 \end{pmatrix} = \begin{pmatrix} 0.657\,64 \\ -0.103\,45 \\ 0.014\,778 \end{pmatrix}$$

then

$$\begin{aligned} &\widehat{Y}_{t+1|t,t-1,t-2} = 10 + 0.657\,64 \times (11 - 10) \\ &- 0.103\,45 \times (9 - 10) + 0.014\,778 \times (9.5 - 10) = 10.\,754. \end{aligned}$$

Forecasting with stationary processes, further comments

In practice, we do not know the autocovariances γ_j, so to use this formula we must replace them with estimates.

Nonparametric: Proceeding as we did for the inference on the mean, we could use a non-parametric method.
Noticing that in stationary and ergodic procesess the contribution of observations that are very far in the past to the forecast is very small, and the autocovariances $|\gamma_j|$ and the weights $\left|\alpha_j^{(T)}\right|$ drop towards 0 as $j \to \infty$, we may for example replace γ_j by its sample moment when $j \leq M$ and by 0 when $j > M$, for example for $M = \sqrt{T}$ (quite like when we used the rectangular kernel to estimate the long run matrix).

Parametric: When we have a parametric model, we may use it to compute all the autocovariances.
For example, if we know that Y_t is generated by the AR(1) model $Y_t = \phi Y t - 1 + \varepsilon_t$, for independent, identically distributed ε_t, with $E(\varepsilon_t) = 0$, $Var(\varepsilon_t) = \sigma^2$, and $|\phi| < 1$ then we can compute all the autocovariances, and find
$\gamma_0 = \frac{\sigma^2}{(1-\phi^2)}$, $\gamma_1 = \phi\frac{\sigma^2}{(1-\phi^2)}$, ... $\gamma_j = \phi^j\frac{\sigma^2}{(1-\phi^2)}$ (for $j \geq 0$),
and therefore $\rho_j = \phi^j$ (for $j \geq 0$).

Thus, if we have an estimate $\widehat{\phi}$, we can estimate $\widehat{\rho_j} = \widehat{\phi}^j$

- advantage: we use all the autocovariances, not only a fraction of them. So, we do not incurr in the bias due to setting to 0 an autocovariance that is close to 0 but not exactly equal to it.
- advantage: we estimate just a small number of parameters (in the example above only one, that is, ϕ), as opposed to many autocovariances. The variance of the estimate of all the autocovariances may be much smaller.
- disadvantage: we may have inconsistent estimates of the autocovariances if we specified the wrong model for Y_t

The parametric approach seems to be more frequent in practice.