Prediction of stationary processes: general ideas and definitions

In Rome, Monte Mario, we have a Weather Station, that is a facility with instruments to make observations of atmospheric conditions, including temperature, barometric pressure, humidity, wind speed, wind direction, and precipitation amounts. Let us concentrate on temperature.

Two technicians, A and B, are in charge for analyzing the temperature data and making forecasts.

A says that in his experience the formula

$$\hat{x}_{t+1}^{A} = 0.9x_t - .6(x_t - x_{t-1})$$

performs fairly well.

Prediction of stationary processes: general ideas and definitions

A says that in his experience the formula

$$\hat{x}^A_{t+1} = 0.9x_t - .6(x_t - x_{t-1})$$

performs fairly well.

B has a different opinion. She maintains that the influence of the current-day's temperature is weaker, she uses a coefficient of 0.7, and that the change between current-day's and previous-day's temperature has a positive, though small, effect:

$$\hat{x}^B_{t+1} = 0.7x_t + .2(x_t - x_{t-1})$$

Prediction of stationary processes: general ideas and definitions

Both solutions

$$\begin{aligned} \hat{x}^A_{t+1} &= 0.9x_t - .6(x_t - x_{t-1}) \\ \hat{x}^B_{t+1} &= 0.7x_t + .2(x_t - x_{t-1}) \end{aligned}$$

are rules, i.e. functions, that associate a predicted value with observed values of the temperature.

In general a predictor is

$$\hat{x}^f_t = f(x_{t-1}, x_{t-2}, \ldots)$$

that is $\hat{x}^f_t$ is a stochastic process which is a function of

$$x_{t-1},\ x_{t-2},\ \ldots$$

Prediction of stationary processes: general ideas and definitions

Thus in principle we have as many predictors of x_t as many functions. Our task is to select a predictor that is optimal.

But to define optimality we need a criterion. For example:

a. Minimize the absolute value of $x_t - \hat{x}_t^f$, which is called prediction error. More precisely, minimize the expected value of the absolute prediction error

$$E(|x_t - \hat{x}_t^f|)$$

b. Minimize

$$E(x_t - \hat{x}_t^f)^2$$

Prediction of stationary processes: general ideas and definitions

Our criterion will be the second:

$$\min \quad E\left(x_t - \hat{x}_t^f\right)^2$$

But minimum with respect to what?

The answer is

$$\min_{f} \quad E\left(x_t - \hat{x}_t^f\right)^2$$

So we are seeking an element in the set of all functions, such that the expected squared error is minimum. This is a huge set to explore!

Prediction of stationary processes: general ideas and definitions

Now we can pay a second visit to the Weather Station and give advice. We say to technicians A and B that their methods seem no more than rules of thumb, and that they should find a common rule by optimizing with respect to some criterion. They respond that the squared error criterion seems good, but that they are not able to determine the best function f. They feel unequal to the complexity of the problem.

We suggest that they simplify the problem by restricting the set of functions. Precisely, we propose linear functions:

$$a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots$$

Now the problem becomes

$$\min_{a_0, a_1, a_2, \ldots} E\left[x_t - (a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots)\right]^2$$

Prediction of stationary processes: general ideas and definitions

We propose linear functions:

$$a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots$$

The problem becomes

$$\min_{a_0, a_1, a_2, \ldots} E\left[x_t - (a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots)\right]^2$$

This can be restated like this:

$$x_t = [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots] + e_t$$

We look for the coefficients a_j such that

$$\mathrm{E}(e_t^2) \text{ is minimum}$$

and this looks very much like a linear regression of x_t on its lags.

Projections and minimum distance

Consider the stochastic variable y and z. We want the best linear approximation of y by means of z, that is

$$y = az + e$$

where a is such that

$$\mathsf{E}(e^2) = \mathsf{E}(y - az)^2$$

is minimum. Set to zero the derivative with respect to a

$$\frac{d}{da}\left[\mathsf{E}(y^2) + a^2\mathsf{E}(z^2) - 2a\mathsf{E}(yz)\right] = 2a\mathsf{E}(z^2) - 2\mathsf{E}(yz) = 0$$

and you obtain

$$a = \frac{E(yz)}{E(z^2)}$$

Projections and minimum distance

Now consider again y and z. We want to find the number b such that

$$e = y - bz$$

is orthogonal to z, orthogonality between the stochatic variables w_1 and w_2 meaning that the moment $\mathsf{E}(w_1 w_2)$ is equal to zero. We find that

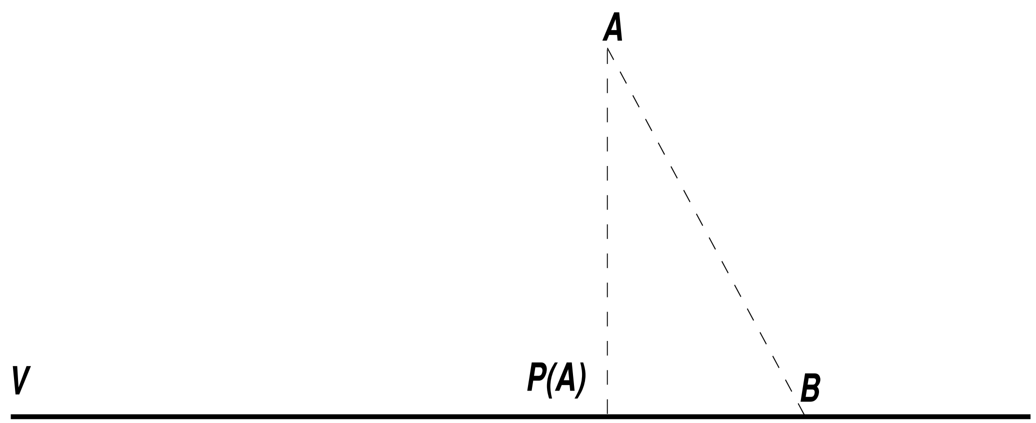
$$\mathsf{E}(ez) = \mathsf{E}(yz) - b\mathsf{E}(z^2)$$

which implies

$$b = \frac{E(yz)}{E(z^2)}$$

which is equal to a.

Projections and minimum distance



The point $P(A)$ is: (1) the point on the line V whose distance from A is minimum, (2) the point obtained by orthogonally projecting A on V. You see that as soon as you move away from $P(A)$, like in B, you loose both properties.

Projections and minimum distance

In general, if

$$e = y - (a_0 + a_1 z_1 + \cdots + a_r z_r),$$

the coefficients that give

$$\min \ \mathsf{E}(e^2) \quad \text{(Minimum distance)}$$

satisfy

$$e \perp z_j, \quad j = 1, 2, \ldots, r, \quad \text{and} \quad \mathsf{E}(e) = 0 \quad \text{(Orthogonal projection)}$$

Note that $\mathsf{E}(e) = 0$ means that e is orthogonal to the stochastic variable that is equal to unity with certainty: $\mathsf{E}(e1) = 0$.

Projections and minimum distance

Rewriting the problem as

$$y = a_0 + a_1 z_1 + \cdots + a_r z_r + e \quad \text{that is} \quad y = (a_0 \; a_1 \; \cdots \; a_r) \begin{pmatrix} 1 \\ z_1 \\ z_2 \\ \vdots \\ z_r \end{pmatrix} + e,$$

the solution is $(a_0 \; a_1 \; \cdots \; a_r) = YC^{-1}$, where

$$Y = \mathsf{E}\left[y \; (1 \; z_1 \; z_2 \; \cdots \; z_r)\right], \quad C = \mathsf{E}\left[\begin{pmatrix} 1 \\ z_1 \\ z_2 \\ \vdots \\ z_r \end{pmatrix} (1 \; z_1 \; z_2 \; \cdots \; z_r)\right] \quad (C \text{ is the variance-covariance matrix})$$

Projections and minimum distance

The following statement insists on uniqueness of projection and residual. Suppose that

$$y = p + \epsilon$$

where (1) ϵ is orthogonal to $1,\ z_1,\ z_2,\ \ldots,\ z_r$, (2) p is a linear combination of $1,\ z_1,\ z_2,\ \ldots,\ z_r$. Then p and ϵ are the projection and the residual respectively.

To prove uniqueness just go back to the matrices Y and C and observe that we can assume that the variables $1,\ z_1,\ z_2,\ \ldots,\ z_r$ have a non-singular covariance matrix.

Projections and minimum distance

In empirical situations we do not know the covariances in C and in Y. We observe data that are drawn from the distributions of y and the z_j's. These data are used to estimate the covariances and therefore the coefficients a_h.

For example, the equation is $y = az + e$, and we have observations

$$y_1,\ y_2,\ \ldots,\ y_N, \qquad z_1,\ z_2,\ \ldots,\ z_N$$

The covariances $\mathsf{E}(yz)$ and $\mathsf{E}(z^2)$ are estimated by

$$\hat{\sigma}_{yz} = \frac{1}{N}\sum_{h=1}^{N} y_h z_h, \qquad \hat{\sigma}_z^2 = \frac{1}{N}\sum_{h=1}^{N} z_h^2, \qquad \text{and } \hat{a} = \frac{\hat{\sigma}_{yz}}{\hat{\sigma}_z^2}$$

This your familiar least squares estimation.

Prediction of stationary processes: general ideas and definitions

Now back to our problem:

$$x_t = [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots] + e_t$$

Adopt the simplification

$$x_t = [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_s x_{t-s}] + e_t$$

where e_t is orthogonal to the regressors.

Note that we are using coefficients that are independent of t. But the coefficients depend on the covariances E$(x_t x_{t-k})$. Thus, assuming that the coefficients are time-invariant requires that the covariances are time-invariant, i.e. that x_t is weakly stationary.

Prediction of stationary processes: general ideas and definitions

Discussion: Is temperature stationary? Would you accept that E(x_t) is the same in January and August? I would not. In this case a model could be

$$x_t = S_t + \eta_t$$

where S_t is a non-stochastic function of t, accounting for the seasonal component, while η_t is zero-mean and weakly stationary.

This introduces the general consideration that the theory of stationary processes may require (most often does require), to be applied, that we reduce to stationarity our data. Examples: the price index P_t is not stationary, but its rate of variation

$$\frac{P_t - P_{t-1}}{P_{t-1}}$$

is stationary. The same holds for the GDP.

Prediction of stationary processes: general ideas and definitions

Now back to our problem:

$$x_t = [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots] + e_t$$

What is a regression on an infinite number of regressors? Consider the regression

$$x_t = [a_0^{(r)} + a_1^{(r)} x_{t-1} + a_2^{(r)} x_{t-2} + \cdots + a_s^{(r)} x_{t-r}] + e_t^{(r)} = p_t^{(r)} + e_t^{(r)}.$$

It is possible to prove that as $r \longrightarrow \infty$

$$p_t^{(r)} \longrightarrow p_t, \qquad e_t^{(r)} \longrightarrow e_t$$

where e_t is orthogonal to all the infinite regressors $1,\ x_{t-1},\ \ldots$

Of course in empirical situations, in which only a sample for $t = 1, 2, \ldots, T$ is available, we will estimate a regression on a finite number r of lags, with r determined by some information criterion.

The Wold Representation Theorem. The Innovation of a stationary process

In conclusion, the best linear predictor of x_t, based on its past, is the projection p_t:

$$x_t = p_t + e_t = [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots] + e_t$$

The process e_t, that is the one-step-ahead prediction error, is also called the innovation of the process x_t.

Looking at the projection equation, the term innovation seems quite appropriate. The only reason why the process x_t is not completely determined by its past values is the presence of the term e_t.

A very important result is that the process e_t is a white noise.

Proof. We have

$$e_t = x_t - [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots]$$

thus e_t is weakly stationary.

The Wold Representation Theorem. The Innovation of a stationary process

Again

$$e_t = x_t - [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots]$$

Remember that e_t is orthogonal to $1,\ x_{t-1},\ x_{t-2},\ \ldots$ But

$$e_{t-1} = x_{t-1} - [a_0 + a_1 x_{t-2} + a_2 x_{t-3} + \cdots]$$

so that e_t is orthogonal to e_{t-1}, etc.

An intuition of the result may be also obtained as follows. Suppose that e_t were not a white noise. For example, the autocovariance $\gamma_1^e \neq 0$. Then in the projection $e_t = \alpha e_{t-1} + \epsilon_t$, the coefficient α is not zero, this implying that E$(\epsilon_t^2) < E(e_t^2)$.
Now

$$\begin{aligned} x_t &= [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots] + e_t = x_t = p_t + e_t = [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots] + \alpha e_{t-1} + \epsilon_t \\ &= [a_0(1-\alpha) + (a_1 + \alpha) x_{t-1} + (a_2 - \alpha a_1) x_{t-2} + \cdots] + \epsilon_t \end{aligned}$$

But this contradicts the assumption that e_t is the residual of the projection of x_t on its past.

The Wold Representation Theorem. The Innovation of a stationary process

Stop for an observation. If y and z are orthogonal then the Pythagorean Theorem holds:

$$\mathsf{E}(y+x)^2 = \mathsf{E}(y^2) + \mathsf{E}(z^2)$$

This is immediately seen computing the left hand side.

Then of course

$$\mathsf{E}(x_t^2) = \mathsf{E}(p_t^2) + \mathsf{E}(e_t^2)$$

so that $\mathsf{E}(e_t^2) \leq \mathsf{E}(x_t^2)$, equality holding if and only if $p_t = 0$, or $x_t = e_t$.

Back to our problem. So $e_t = x_t - [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots]$ is a white noise. On the other hand, $x_t = e_t$ if and only if x_t is a white noise (prove this statement). Therefore a white noise is unpredictable. Better, we can say that stationary processes are predictable in that the pattern of autocorrelation is constant through time. A white noise is the least predictable among stationary processes. Processes whose autocorrelation is not regular through time are absolutely unpredictable.

The Wold Representation Theorem. The Innovation of a stationary process

Examples:

1. $x_t = A$. In this case the projection equation is

$$x_t = x_{t-1} + 0$$

but also $x_t = x_{t-2} + 0$, etc. Thus the innovation is zero. Do not say that there is no innovation, or, say it if you want, but remember what you mean.

2. $x_t = (-1)^t A$. Same as in the previous case, only that here the projection is $x_t = -x_{t-1} + 0 = x_{t-2} + 0$, etc. Zero innovation.

3. The AR(1) process, that is the stationary solution of $z_t = \alpha z_{t-1} + u_t$, $|\alpha| < 1$, which is

$$x_t = u_t + \alpha u_{t-1} + \alpha^2 u_{t-2} + \cdots$$

In this case, using the definition of x_t, firstly

$$u_t \perp x_{t-k} = u_{t-k} + \alpha u_{t-k-1} + \alpha^2 u_{t-k-2} + \cdots$$

for $k \geq 1$. Secondly αx_{t-1} is a linear combination of past values of x_t (too obvious). So

$$x_t = p_t + e_t = \alpha x_{t-1} + u_t$$

This means that the best linear prediction of x_t is αx_{t-1}.

4. The MA(1) process $x_t = u_t - \beta u_{t-1}$. Obviously

$$u_t \perp x_{t-k} = u_{t-k} - \beta u_{t-k-1}$$

for $k \geq 1$. Assume that $|\beta| < 1$. Then, by the same recursive argument used to solve the AR(1) process,

$$u_t = x_t + \beta x_{t-1} + \beta^2 x_{t-2} + \cdots$$

Thus $-\beta u_{t-1}$ is a linear combination of past values of x_t, so that

$$x_t = p_t + e_t = -\beta u_{t-1} + u_t$$

The best linear prediction of x_t is

$$-\beta u_{t-1} = -\beta[x_{t-1} + \beta x_{t-2} + \cdots]$$

The case $|\beta| > 1$ will be discussed later on.

5. The ARMA(p,q) case $a(L)z_t = b(L)u_t$, whose stationary solution is

$$x_t = a(L)^{-1}b(L)u_t = u_t + A_1 u_{t-1} + A_2 u_{t-2} + \cdots$$

This implies that $u_t \perp x_{t-k}$ for $k \geq 1$. If the roots of $b(L)$ are larger than unity in modulus (invertibility), then

$$u_t = b(L)^{-1}a(L)x_t$$

so that u_t is a linear combination of $x_t,\ x_{t-1},\ \cdots$. In that case the projection equation is

$$x_t = p_t + e_t = [\alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} + \beta_1 u_{t-1} + \cdots + \beta_q u_{t-q}] + u_t$$

In conclusion, if the stationarity and invertibility conditions are satisfied, u_t is the innovation of the ARMA process $a(L)z_t = b(L)u_t$.

Back to the regression

$$x_t = [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots] + e_t \qquad (*)$$

Thus, as we have observed, x_t is determined by its past plus the innovation e_t. Using

$$x_{t-1} = [a_0 + a_1 x_{t-2} + a_2 x_{t-3} + \cdots] + e_{t-1}$$

to replace x_{t-1} in $(*)$, we obtain

$$x_t = e_t + b_1 e_{t-1} + [f + f_2 x_{t-2} + f_3 x_{t-3} + \cdots]$$

We may hope that iterating the procedure we obtain a result like the one obtained in the AR(1) case:

$$x_t = b + e_t + b_1 e_{t-1} + b_2 e_{t-2} + \cdots$$

The Wold Representation Theorem. The Innovation of a stationary process

We may hope that iterating the procedure we obtain a result like the one obtained in the AR(1) case:

$$x_t = b + e_t + b_1 e_{t-1} + b_2 e_{t-2} + \cdots$$

This is not true in general, as the example $x_t = A$ shows.

The intuition based on the iterative procedure can be given a rigorous version by projecting x_t on $1,\ e_t,\ e_{t-1},\ \ldots$

$$x_t = [b + b_0 e_t + b_1 e_{t-1} + b_2 e_{t-2} + \cdots] + d_t$$

Show that $b_0 = 1$ (use $x_t = [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots] + e_t$) so that the projection is

$$x_t = [b + e_t + b_1 e_{t-1} + b_2 e_{t-2} + \cdots] + d_t$$

This is called the Wold representation of x_t.

The Wold Representation Theorem. The Innovation of a stationary process

The Wold Representation Theorem states that a weakly stationary process x_t has the representation

$$x_t = [b + e_t + b_1 e_{t-1} + b_2 e_{t-2} + \cdots] + d_t$$

where e_t is the innovation of x_t, while d_t is a process with zero innovation, i.e.

$$d_t = D_1 d_{t-1} + D_2 d_{t-2} + \cdots$$

Moreover, d_t is orthogonal to e_s, for all s.
Processes like d_t, with zero innovation, are called linearly deterministic.

In conclusion, a weakly stationary process is the sum of a backward moving average of the innovation, which is a white noise, plus a linearly deterministic process. The two components are orthogonal at all leads and lags.

The Wold Representation Theorem. The Innovation of a stationary process

We have seen that ARMA processes have a Wold representation without the deterministic component:

$$x_t = a(L)^{-1}b(L)u_t$$

On the opposite side, $x_t = A$ has only the deterministic component.

The following is an interesting exercise

$$x_t = u_t + A$$

where u_t is white noise and $u_t \perp A$ for all t. Both A and u_t are zero mean. Prove that u_t is the innovation of x_t and A the deterministic component.

Consider the regression

$$x_t = a_1^{(r)}x_{t-1} + a_2^{(r)}x_{t-2} + \cdots + a_r^{(r)}x_{t-r} + e_t^{(r)}$$

The Wold Representation Theorem. The Innovation of a stationary process

Consider the regression

$$x_t = a_1^{(r)} x_{t-1} + a_2^{(r)} x_{t-2} + \cdots + a_r^{(r)} x_{t-r} + e_t^{(r)}$$

Using the fact that

$$\gamma_k^x = \begin{cases} \sigma_u^2 + \sigma_A^2 & \text{if } k = 0 \\ \sigma_A^2 & \text{if } k \neq 0 \end{cases}$$

we obtain that the coefficients $a_h^{(r)}$ are all equal. Thus the regression is

$$x_t = a^{(r)} [x_{t-1} + x_{t-2} + \cdots + x_{t-r}] + e_t^{(r)}$$

that is

$$e_t + A = a^{(r)} [u_{t-1} + u_{t-2} + \cdots + u_{t-r}] + a^{(r)} r A + u_t^{(r)}$$

The Wold Representation Theorem. The Innovation of a stationary process

Rewrite the last display

$$u_t + A = a^{(r)}[u_{t-1} + u_{t-2} + \cdots + u_{t-r}] + a^{(r)} rA + e_t^{(r)}$$

Using

$$e_t^{(r)} = u_t + A - a^{(r)}[u_{t-1} + u_{t-2} + \cdots + u_{t-r}] - a^{(r)} rA$$

and orthogonality of $e_t^{(r)}$ to $x_{t-1} = u_{t-1} + A$, we obtain

$$a^{(r)} = \frac{\sigma_A^2}{\sigma_u^2 + r\sigma_A^2}$$

that is

$$x_t = p_t^{(r)} + e_t^{(r)} = \left[\frac{\sigma_A^2}{\sigma_u^2 + r\sigma_A^2}[u_{t-1} + u_{t-2} + \cdots + u_{t-r}] + \frac{r\sigma_A^2}{\sigma_u^2 + r\sigma_A^2} A\right] + e_t^{(r)}$$

The Wold Representation Theorem. The Innovation of a stationary process

Rewrite

$$x_t = A + u_t = p_t^{(r)} + e_t^{(r)} = \left[\frac{\sigma_A^2}{\sigma_u^2 + r\sigma_A^2}[u_{t-1} + u_{t-2} + \cdots + u_{t-r}] + \frac{r\sigma_A^2}{\sigma_u^2 + r\sigma_A^2}A\right] + e_t^{(r)}$$

As $r \longrightarrow \infty$

$$p_t^{(r)} \longrightarrow A, \quad e_t^{(r)} \longrightarrow u_t$$

that is

$$\mathsf{E}(A - p_t^{(r)})^2 \longrightarrow 0, \quad \mathsf{E}(u_t - e_t^{(r)})^2 \longrightarrow 0$$

We have to prove that

$$\mathsf{E}\left[\frac{\sigma_A^2}{\sigma_u^2 + r\sigma_A^2}[u_{t-1} + u_{t-2} + \cdots + u_{t-r}]\right]^2 \longrightarrow 0$$

For

$$\left[\frac{\sigma_A^2}{\sigma_u^2 + r\sigma_A^2}\right]^2 r\sigma_u^2 = \left[\frac{\sigma_A^2\sqrt{r\sigma_u^2}}{\sigma_u^2 + r\sigma_A^2}\right]^2 \longrightarrow 0$$

In conclusion, as $p_t^{(r)} \longrightarrow p_t = A$,

$$x_t = p_t + e_t = A + u_t$$

The white noise u_t is the innovation. Of course the projection of $x_t = u_t + A$ on present and past values of the innovation is u_t, so that the Wold representation is

$$x_t = [u_t + b_1 u_{t-1} + \cdots] + d_t = u_t + A$$

The result looks trivial, but obtaining it requires some work.

The Wold Representation Theorem. The Innovation of a stationary process

Now consider again the MA(1) process

$$x_t = u_t - \beta u_{t-1}$$

In order to prove that u_t is the innovation of x_t we argue that

(1) $u_t \perp x_{t-k}$ for $k \geq 1$

(2) u_t is a linear combination of $x_t, \ x_{t-1}, \ \ldots$

To prove (2)

$$u_t = x_t + \beta x_{t-1} + \beta^2 x_{t-2} + \cdots$$

But this requires that $|\beta| < 1$. What if $|\beta| > 1$?

The Wold Representation Theorem. The Innovation of a stationary process

Rewrite the MA(1) as

$$x_t = (1 - \beta L)u_t$$

We know the trick to obtain u_t as a moving average of the x's.

$$u_t = \frac{1}{1-\beta L}x_t = \frac{-\beta^{-1}F}{1-\beta^{-1}F}x_t = -\beta^{-1}[x_{t+1} + \beta^{-1}x_{t+2} + \beta^{-2}_{t+3} + \ldots]$$

Thus when $|\beta| > 1$, u_t is a linear combination of future values of x_t and is not the innovation of x_t.

The Wold Representation Theorem. The Innovation of a stationary process

To find the innovation of $x_t = (1-\beta L)u_t$, for $|\beta| > 1$, we use the following statement. There exists a white noise v_t such that

$$x_t = (1-\beta L)u_t = (1-\beta^{-1}L)v_t$$

Then v_t is the innovation of x_t.

Determining v_t is easy

$$\begin{aligned} v_t &= \frac{1-\beta L}{1-\beta^{-1}L}u_t = (1-\beta L)(1+\beta^{-1}L+\beta^{-2}L^2+\cdots)u_t \\ &= [1+(\beta^{-1}-\beta)L+\beta^{-1}(\beta^{-1}-\beta)L^2+\beta^{-2}(\beta^{-1}-\beta)L^3+\cdots]u_t \end{aligned}$$

But we have to prove the v_t, though being a moving average of a white noise, is a white noise. This is an interesting exercise, requiring only sums of geometric series. Note that v_t is an infinite moving average. A finite moving average of a white noise cannot be a white noise.

The Wold Representation Theorem. The Innovation of a stationary process

Consider now the case $\beta = 1$:

$$x_t = u_t - u_{t-1}$$

We can prove that u_t is the innovation of x_t (not very difficult).

An important observation. Consider the regression

$$x_t = p_t^{(r)} + u_t^{(r)} = a_1^{(r)} x_{t-1} + a_2^{(r)} x_{t-2} + \cdots + a_r^{(r)} x_{t-r} + e_t^{(r)}$$

In this case, although

$$p_t^{(r)} \longrightarrow p_t = -u_{t-1}$$

the projection cannot be represented as

$$x_t = p_t + u_t = [a_1 x_{t-1} + a_2 x_{t-2} + \cdots] + u_t$$

The reason is that the polynomial $1 - L$ is not invertible, so that all the coefficients $a_2^{(r)}$ tend to 1.

The Wold Representation Theorem. The Innovation of a stationary process

Rewrite:

In this case, although

$$p_t^{(r)} \longrightarrow p_t = -u_{t-1}$$

the projection cannot be represented as

$$x_t = p_t + u_t = [a_1 x_{t-1} + a_2 x_{t-2} + \cdots] + u_t$$

The reason is that the polynomial $1 - L$ is not invertible, so that all the coefficients $a_2^{(r)}$ tend to 1.

Therefore, though convenient, writing

$$x_t = a_1^{(r)} x_{t-1} + a_2^{(r)} x_{t-2} + \cdots + a_r^{(r)} x_{t-r} + e_t^{(r)} = [a_1 x_{t-1} + a_2 x_{t-2} + \cdots] + e_t$$

is not completely rigorous. If x_t is a moving average we must add the assumption that no root has unit modulus.

The Wold Representation Theorem. The Innovation of a stationary process

Given the ARMA

$$\begin{aligned} a(L)x_t &= b(L)u_t = u_t + \beta_1 u_{t-1} + \cdots + \beta_q u_{t-q} \\ &= (1-\delta_1 L)(1-\delta_2 L)\cdots(1-\delta_q L)u_t \end{aligned} \qquad (*)$$

we can apply the technique shown above for the MA(1) to replace all the roots δ_j whose modulus is smaller than 1 with their reciprocals. Thus given an MA(q), this can be transformed into an invertible MA(q).

As we have seen, if $b(L)$ is invertible, i.e. if the roots of $b(L)$ lie outside of the unit circle, then u_t is the innovation of x_t. We also say that u_t is fundamental for x_t or that representation $(*)$ is a fundamental representation for x_t.

For example, $x_t = u_t - 2u_{t-1}$ is not fundamental, but we know that x_t has also the representation $x_t = v_t - 0.5v_{t-1}$, which is fundamental.

The Wold Representation Theorem. The Innovation of a stationary process

If our aim is prediction, then only fundamental representations are important. However, non fundamental representations may arise in structural analysis. Consider the following stylized example.

The variable x_t is the quarterly rate of change of aggregate productivity

The white noise u_t is a shock to technical knowledge.

The shock to technical knowledge takes two quarters to be completely absorbed by a change in productivity:

$$
\begin{aligned}
x_t &= a_0 u_t + a_1 u_{t-1}, \quad a_0 + a_1 = 1 \\
x_t &= w_t + \alpha w_{t-1}, \quad w_t = a_0 u_t, \;\; \alpha = a_1/a_0
\end{aligned}
$$

The shock u_t is fundamental for x_t if and only if $a_0 > a_1$. But this is not necessarily true. If the coefficients a_j represent a learning-by-doing process, or diffusion of technical innovations among firms, then why should the first impact be more important than the lagged effect?

The Wold Representation Theorem. The Innovation of a stationary process

Rewrite the display:

$$
\begin{aligned}
x_t &= a_0 u_t + a_1 u_{t-1}, \quad a_0 + a_1 = 1 \\
x_t &= w_t + \alpha w_{t-1}, \quad w_t = a_0 u_t, \;\; \alpha = a_1/a_0
\end{aligned}
$$

The shock u_t is fundamental for x_t if and only if $a_0 > a_1$. But this is not necessarily true. If the coefficients a_j represent a learning-by-doing process, or diffusion of technical innovations among firms, then why should the first impact be more important than the lagged effect?

Now, if the econometrician only observes x_t, the rate of change of productivity, he/she is not able to choose which MA(1) representation is the structural representation, the fundamental or the other. This identification problem is know as the fundamentalness problem. But if you are interested only in prediction then no identification problem arises. You just choose the fundamental representation.

The Wold Representation Theorem. The Innovation of a stationary process

Summing up, every stationary process has the representation

$$x_t = [b + e_t + b_1 e_{t-1} + b_2 e_{t-2} + \cdots] + d_t$$

where d_t is predictable without error using its past values.

ARMA processes do not contain the term d_t.

Can we say that only processes without d_t are interesting for economists? Yes and no.

The Wold Representation Theorem. The Innovation of a stationary process

Remember the space of trajectories? Consider $\mathbb{R}^{\mathbb{Z}}$ and the four trajectories

$$\begin{array}{ccccccccccc} \cdots & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & \cdots & \text{time} \\ \cdots & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & \cdots & g_1 \\ \cdots & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & \cdots & g_2 \\ \cdots & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & \cdots & g_3 \\ \cdots & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & \cdots & g_4 \end{array}$$

Interpret time as quarters. The trajectories g_j represent an event occurring every year in the j-th quarter. Now, the probability space Ω is $\mathbb{R}^{\mathbb{Z}}$ with probability $1/4$ assigned to each of the trajectories g_j, and zero for the set of all other trajectories.

Lastly, define the stochastic process

$$d_t(g_j) = g_{j,t}$$

and call d_t the Independence Day process.

Interpretation. Many years ago, a war has been fought for independence of our country. The decisive battle took place in the first quarter, so ever since we celebrate the day of that battle. This is why the number of working days in the first quarter must be corrected to take the Independence Day into account. But that battle might have been fought in a different quarter, or maybe it was decided that that battle has been decisive against the opinion that another was the most important. This is why we interpret Independence Day as a stochastic process: it might have been different. (With a different outcome of the battle there would not be an Independence Day.)

Remember that linearly deterministic processes do not look like stochastic processes. Remember $x_t = A$, or $x_t = (-1)^t$, or, now, the Independence Day process. This point will be touched upon again when talking of estimation.

The Wold Representation Theorem. The Innovation of a stationary process

Usually, effects due to special celebrations like Independence Day, Easter, Christmas, etc. are removed from economic time series within preliminary analysis.

Preliminary analysis should also remove:

(1) Outliers, that is values of the time series that are likely not to belong to the stationary distribution, like an earthquake, or a very important strike, etc.

(2) Seasonal components, i.e. sizable oscillations in output, hour worked, etc., that are due to atmospheric variations and are not interesting from an economic point of view. Italian industrial production falls dramatically in August, but this has no economic meaning.

(3) Trend. This will be dealt with later on.

The Wold Representation Theorem. The Innovation of a stationary process

In conclusion, after preliminary analysis we can say that the Wold Representation of an economic time series is

$$x_t = b + e_t + b_1 e_{t-1} + b_2 e_{t-2} + \cdots$$

where e_t is the innovation of x_t.

Prediction of ARMA processes

Given the ARMA

$$a(L)x_t = b(L)u_t$$

with all the roots in the right place, we have seen that the projection equation is

$$x_t = p_t + e_t = [\alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} + \beta_1 u_{t-1} + \cdots + \beta_q u_{t-q}] + u_t$$

Now, replacing x_{t-1} we obtain

$$x_t = \left[(\alpha_1^2 + \alpha_2)x_{t-2} + \cdots + \alpha_1\alpha_p x_{t-p-1} + (\alpha_1\beta_1 + \beta_2)u_{t-2} + \cdots + \alpha_1\beta_q u_{t-q-1}\right] + [u_t + (\alpha_1 + \beta_1)u_{t-1}]$$

This is the projection of x_t on the space spanned by $x_{t-2},\ x_{t-3},\ \ldots$. You see that the two-step-ahead prediction error is no longer a white noise (the argument used proving the Wold Theorem does not apply here; are you convinced?). Further replacements provide the h-step-ahead prediction error for all h.

Prediction of stationary processes

A more general way to analyze the h-step-ahead prediction error is the following. Consider the Wold representation

$$x_{t+h} = b + e_{t+h} + b_1 e_{t+h-1} + b_2 e_{t+h-2} + \cdots$$

Rewrite this as

$$x_{t+h} = [e_{t+h} + b_1 e_{t+h-1} + \cdots + b_{h-1} e_{t+1}] + [b + b_h e_t + b_{h+1} e_{t-1} + \cdots] = e_{t+h|t} + p_{t+h|t}$$

Since $e_t = x_t - (a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots)$, then e_t belongs to the space spanned by $1,\ x_t,\ x_{t-1},\ x_{t-2},\ \ldots$. On the other hand, since $x_t = b + e_t + b_1 e_{t-1} + a_2 e_{t-2} + \cdots$, then x_t belongs to he space spanned by $1,\ e_t,\ e_{t-1},\ e_{t-2},\ \ldots$ so that the two spaces coincide. Thus $p_{t+h|t}$ and $e_{t+h|t}$ are the projection of x_{t+h} on the space spanned by $1,\ x_t,\ x_{t-1},\ \ldots$ and the residual respectively.

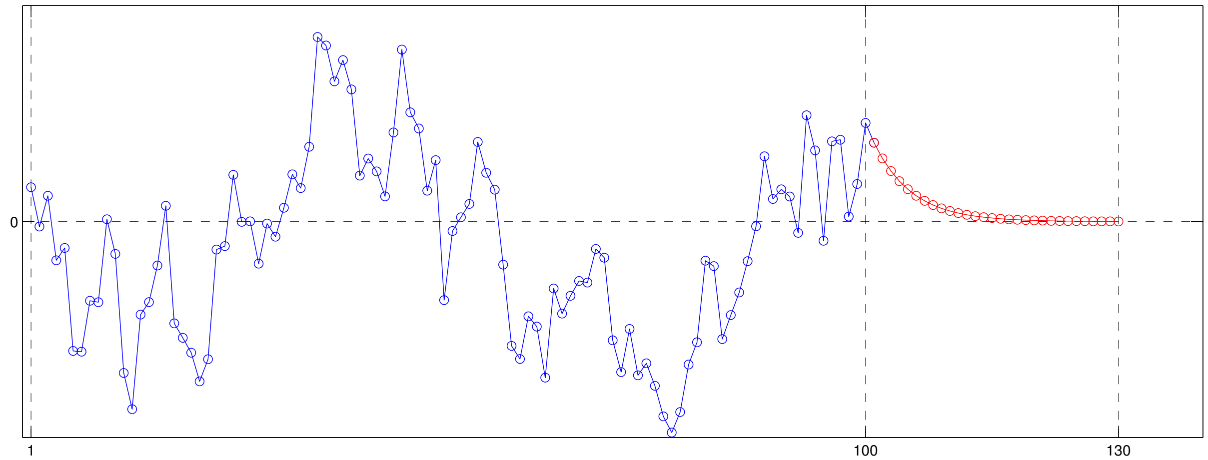
Prediction of stationary processes

$$x_{t+h} = [e_{t+h} + b_1 e_{t+h-1} + \cdots + b_{h-1} e_{t+1}] + [b + b_h e_t + b_{h+1} e_{t-1} + \cdots] = e_{t+h|t} + p_{t+h|t}$$

This also shows that as $h \longrightarrow \infty$,

$$p_{t+h|t} \longrightarrow b, \quad e_{t+h|t} - (x_{t+h} - b) \longrightarrow 0$$

This implies that the variance of the prediction error has a finite bound as $h \longrightarrow \infty$, namely σ_x^2.

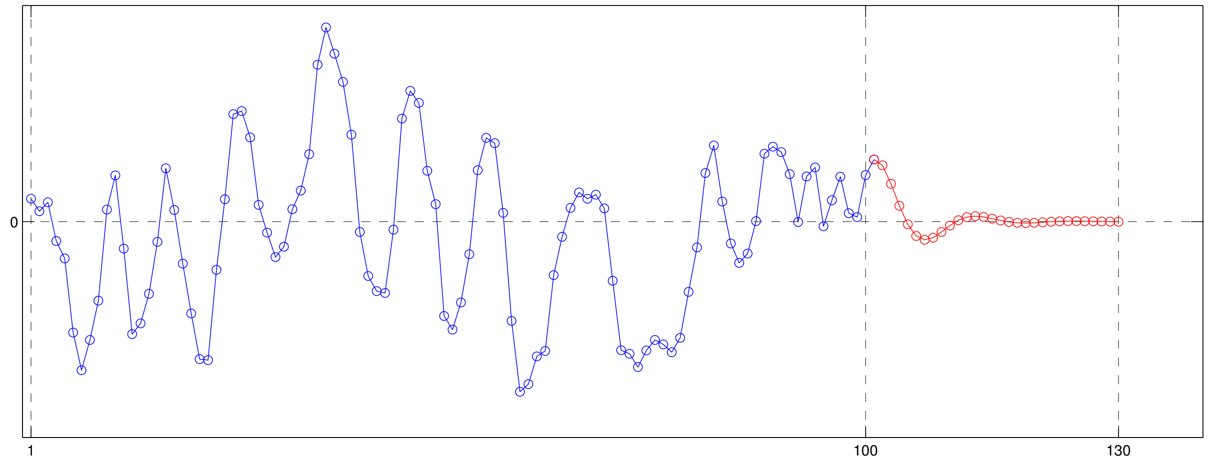


Prediction of stationary processes

The plot has

$$x_t = 0.8x_{t-1} + u_t$$

between 1 and 100, followed by 30 predicted values (red circled).



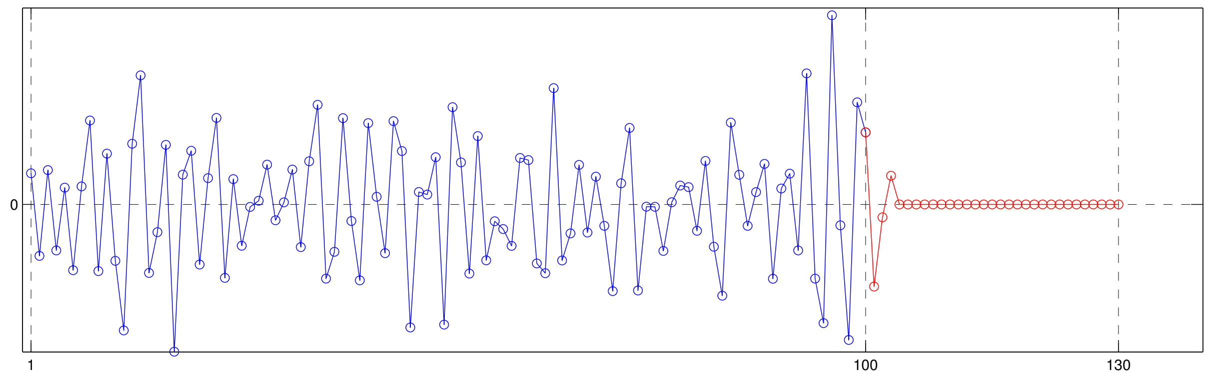
Prediction of stationary processes

The plot has

$$x_t = 1.4x_{t-1} - 0.66x_{t-2} + u_t$$

between 1 and 100, followed by 30 predicted values (red circled). The roots of the polynomial $1 - 1.4L + 0.66L^2$ are complex:

$$r = 0.9\left(\cos\frac{2\pi}{12} \pm i\sin\frac{2\pi}{12}\right)$$



Prediction of stationary processes

The plot has

$$x_t = u_t - 0.7u_{t-1} - 0.33u_{t-2} + 0.46u_{t-3}$$

between 1 and 100, followed by 30 predicted values (red circled). The roots of the polynomial $1 - 0.7L - 0.33L^2 + 0.46L^3$ are all outside the unit circle. Note that all predicted values after the third are zero, consistently with

$$x_t = [e_t + b_1 e_{t-1} + \cdots + b_{h-1} e_{t-h+1}] + [b_h e_{t-h} + b_{h+1} e_{t-h-1} + \cdots] = e_{t+h|t} + p_{t+h|t}$$

Prediction of stationary processes

Rewrite

$$x_{t+h} = [e_{t+h} + b_1 e_{t+h-1} + \cdots + b_{h-1} e_{t+1}] + [b + b_h e_t + b_{h+1} e_{t-1} + \cdots] = e_{t+h|t} + p_{t+h|t}$$

Remember that the space spanned by $e_t,\ e_{t-1},\ e_{t-2},\ \ldots$ and the space spanned by $x_t,\ x_{t-1},\ x_{t-2},\ \ldots$ coincide. Thus

$$p_{t+h|t} = a_h^h x_t + a_{h+1}^h x_{t-1} + \cdots$$

Also

$$x_{t+h} = [a_h^h x_t + a_{h+1}^h x_{t-1} + \cdots] + e_{t+h|t} = [a_h^h x_t + a_{h+1}^h x_{t-1} + \cdots] + [e_{t+h} + b_1 e_{t+h-1} + \cdots + b_{h-1} e_{t+1}]$$

If e_t and e_s are independent for $t \neq s$, white noise in the strict sense, then e_t and x_{t-k}, for $k > 0$ are independent. As a consequence $a_h^h x_t + a_{h+1}^h x_{t-1} + \cdots$ is the conditional expectation of x_{t+h}, given $x_t,\ x_{t-1},\ \cdots$.
Conditional expectation is often used for $a_h^h x_t + a_{h+1}^h x_{t-1} + \cdots$ even without assuming that e_t is white noise in the strict sense.

Prediction of stationary processes

Consider the following example

$$x_t = u_t + \beta u_{t-1} u_{t-2}$$

where $|\beta| < 1$ and u_t is white noise in the strict sense. A simple exercise shows that x_t is a white noise, i.e. it has zero mean and zero autocovariances γ_k^x for $k \neq 0$. Thus

$$x_{t+h|t} = 0$$

for all $h > 0$.

However, it is possible to prove that u_t can be recovered as limit of non-linear functions of $x_t,\ x_{t-1},\ \ldots$ Therefore, the best prediction of x_{t+1}, based on $x_t,\ x_{t-1},\ \ldots$, is $\beta u_t u_{t-1}$, not 0. Thus the non-linear prediction has a smaller prediction error compared with that of the linear prediction. In other words, there is something you can learn about x_{t+1} if you consider non-linear combinations of $x_t,\ x_{t-1},\ \ldots$, is $\beta u_t u_{t-1}$.

Prediction of stationary processes

Rewrite

$$x_t = u_t + \beta u_{t-1} u_{t-2} = u_t + G(x_{t-1}, x_{t-2}, \ \ldots)$$

Since u_t is independent of x_{t-k}, $k > 0$, then $\beta u_{t-1} u_{t-2} = G(x_{t-1}, x_{t-2}, \ \ldots)$ is the conditional expectation of x_t given past values of x_t.

This is a particular case of a general theorem: given the stochastic process x_t, the best prediction of x_t, given x_{t-k}, $k > 0$, with respect to the minimum mean square error criterion, is the conditional expectation of x_t, given x_{t-k}, $k > 0$.

As observed above, if x_t is stationary with one-step-ahead prediction error e_t, then if e_t is white noise in the strict sense, the best prediction and the best linear prediction of x_t coincide. Thus the conditional expectation of x_{t+h} given x_{t-k}, $k \geq 0$, is a linear combination of $x_t, \ x_{t-1}, \ \ldots$