Linear time series models

February 18, 2021

©2021 by Christopher A. Sims. ©2021. This document is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
http://creativecommons.org/licenses/by-nc-sa/3.0/

The lag operator

- Given any function of time $f(t)$, Lf is another function of time with $Lf(t) = f(t-1)$, or if we use subscripts to indicate dependence on t, $Ly_t = y_{t-1}$.
- We can take powers of L, so $L^2 y_t = y_{t-2}$, for example.
- We can form polynomials in L, so $(a+bL+cL^2)y_t = ay_t + by_{t-1} + cy_{t-2}$, for example.
- We can take negative powers of L, so $L^{-2}y_t = y_{t+2}$ for example.
- What would we mean by $(1-bL)^{-1}$?

Polynomial long division with lag operators

(blackboard)

$$\frac{1}{1-bL} = 1 + bL + b^2L^2 + b^3L^3 + \dots$$

Also

$$\frac{1}{1-bL} = \frac{-b^{-1}L^{-1}}{1-b^{-1}L} = -b^{-1}L^{-1} - b^{-2}L^{-2} - b^{-3}L^{-3} - \dots$$

If $|b| \neq 1$, just one of these is convergent. If $|b| = 1$, neither is convergent.

Three views of, or ways to specify, a stochastic process

I. A sequence $\{y_t, t \in \mathbb{Z}\}$ of random variables or random vectors, together with a rule for determining the joint distribution of any finite collection $\{y_{t_1}, \ldots, y_{t_k}\}$ of them.

   $\mathbb{Z}$ is the integers. The index set could also be the real line $\mathbb{R}$ or any subset of it. We will stick to $\mathbb{Z}$ or the non-negative integers $\mathbb{Z}^+$. It is also possible to make the index set $\mathbb{R}^2$, which gives you a "spatial process".

II. A rule for making random draws of entire sequences. In this view a random vector is a special case of a stochastic process in which the index set is $\{1, \ldots, n\}$ instead of the infinitely long sequence $\mathbb{Z}^+$.

III. A distribution for y_0 and a distribution for $y_t \mid \{y_{t-s}, s \geq 1\}$ for every $t > 0$. This is obviously a special case of view I.

The innovation in a finite-variance stochastic process

The **innovation** at time t in the stochastic process y is

$$y_t - \left(\text{minimum variance linear predictor of } y_t \text{ based on } \{y_s, s < t\}\right).$$

In other words, the forecast error when using the best linear predictor.

Note that in some cases the best linear predictor involves infinitely many lagged values of y, and may even be something that can't be represented as $\sum_{s=1}^{\infty} a_s y_{t-s}$ for any fixed a_s sequence, but instead only as a limit of a sequence of such $a(L)y$ expressions. (As we'll see below.)

Stationary processes

- A **stationary**R process y is one such that the joint distribution of any collection $\{y_t + \delta_1, y_{t+\delta_2}, \ldots, y_{t+\delta_n}\}$ does not change with t.

- Equivalently, it is a process $\{x_t\}_{t=-\infty}^{\infty}$ such that $L^s x$ has the same distribution as x itself, for any integer s.

Covariance-stationary processes and their ACF's

- A **covariance stationary process** y_t is one such that $\mathrm{Cov}(y_t, y_{t-s}) = R_y(s)$ exists for every s and does not depend on t. Also $E[y_t]$ does not depend on t.

- The function $R_y(s)$, which is an $m \times m$ matrix if y_t is a vector of length m, is called the **autocovariance function**, or ACF.

- A Gaussian process y_t is one such that for any finite collection of dates $\{t_1, \ldots t_k\}$, the joint distribution of $\{y_{t_1}, \ldots, y_{t_k}\}$ is Normal.

- The distribution of a stationary Gaussian process is fully characterized by its mean and its ACF.

The finite MA class of models

$$y_t = \sum_{s=0}^{k} a_s \varepsilon_{t-s} = a(L)\varepsilon_t \, .$$

y may be $m \times 1$, in which case a_s is $m \times m$. $\varepsilon \sim N(0, \Sigma)$, i.i.d. Or sometimes just mean 0, variance Σ, not serially correlated.

- Note that we have not said that ε_t is the innovation in y_t, only that it is serially uncorrelated.
- The same stochastic process can generally be represented as a linear combination of uncorrelated ε_t's in many different ways.
- The representation that makes ε_t the innovation is called the **fundamental** MA representation.

Example of fundamental and non-fundamental MA's

- A: $y_t = \varepsilon_t + .5\varepsilon_{t-1}$, B: $y_t = \eta_t + 2\eta_{t-1}$, $\mathrm{Var}(\varepsilon_t) = 4\,\mathrm{Var}(\eta_t)$
- If $\mathrm{Var}(\varepsilon_t) = 4$, $R_y(0) = 5$, $R_y(1) = 2$, $R_y(s) = 0$ for $|s| > 1$. Both MA representations give the same R_y.
- Which is fundamental? I.e., which is the innovation, ε_t or η_t? Or, which is the best linear predictor, $.5\varepsilon_{t-1}$ or $2\eta_{t-1}$?

Properties of finite order MA processes.

- Dense in the space of **linearly regular** stationary processes. These are processes such that the best linear forecast converges to the mean of the process as the forecast horizon grows to infinity.
- Closed under taking linear combinations.
- Closed under taking subvectors.
- For forecasting, we need the fundamental MA.

ACF and MA related via polynomial multiplication

- It is not hard to verify that, if $y_t = a(L)\varepsilon_t$ with ε_t i.i.d. $N(0, \Sigma)$, $R_y(s)$ is the coefficient on L^s in

$$a(L)\,\Sigma\,a'(L)\,,$$

  where we define $a'(L) = \sum_s a'_s L^{-s}$, i.e. the $'$ operator on $a(L)$ both transposes matrix coefficients and changes the sign on the exponents of L.

- We often abuse notation by writing $R_y(L) = a(L)\,\Sigma\,a'(L)$.

Univariate special case; examples

$$R_y(L) = a(L)a(L^{-1})\sigma_\varepsilon^2$$

$$y_t = (1 + bL)\varepsilon_t \Rightarrow R_y(L) = (bL^{-1} + 1 + b^2 + bL)\sigma_\varepsilon^2$$

$$y_t = \left(I + \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} L \right) \varepsilon_t$$

$$\Rightarrow R_y(L) = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} L^{-1} + \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} L$$

Roots and fundamentalness: first-order scalar case

Suppose $y_t = (1 + bL)\varepsilon_t$. Multiply both sides by $(1 + bL)^{-1}$:

$$\varepsilon_t = (1 + bL)^{-1} y_t = y_t - by_{t-1} + b^2 y_{t-2} - b^3 y_{t-3} + \dots .$$

- Does this expression make any sense? If $|b| < 1$, it does, because then the coefficients in the infinite sum converge.
- If y_t has bounded variance for all t, the partial sums $\sum_0^T (-b)^s y_{t-s}$ converge to a well-defined limit. (Technically, they form a Cauchy sequence when we use the norm $\|X\| = \sqrt{\mathrm{Var}(X)}$.)
- In that case we have expressed ε_t as a linear combination of current and past y's.

- Then ε_t is the innovation: it is uncorrelated with past ε and therefore also uncorrelated with past y values. Thus there is no way to predict it with linear combinations of past y's.

- And of course if $|b| > 1$, this argument doesn't work because the partial sums don't converge to anything.

- Note that the variance of the serially uncorrelated shocks in a moving average representation is *largest* for the fundamental MAR. (Why?)

- The ACF here is $Ry(L) = \sigma_\varepsilon^2(1 + bL)(1 + b^{-1}L)$. There are just two MA representations, one with $|b| < 1$ and the the root b^{-1} of $1 - bz$ therefore larger than one in absolute value, the other with root b smaller than one in absolute value.

What if $|b| = 1$?

Suppose $b = 1$. In that case $(1+L)^{-1} = 1 - L + L^2 - L^3 + \dots$ does not converge, so it might seem that ε_t is not the innovation. However it turns out that there are sequences of linear combinations of current and past y's that do converge to ε_t If we set $a^n(L)$ as

$$a_s^n = (1 - s/n)(-1)^s, s \in \{0, \dots, n\}, 0 \text{ otherwise},$$

$$a^n(L)y_t = \varepsilon_t + \frac{1}{n}\sum_{s=1}^{n}(-1)^s\varepsilon_{t-s}$$

$$\mathrm{Var}\big((\varepsilon_t - a^n(L)y_t\big) = \frac{\sigma_\varepsilon^2}{n} \xrightarrow[n\to\infty]{} 0$$

A similar argument works for $b = -1$.

Roots and fundamentalness

- The z-transform of $a(L)$ is $a(z) = \sum_s a_s z^s$. I.e., we replace the lag operator L in the polynomial with a complex number z and consider the function on the complex plane that this defines.

- Note that the roots of the polynomial $|a(z)|$ are just the inverses of the roots of $|a'(z)| = \left|a(z^{-1})\right|$.

- In the first-order scalar case we verified that the fundamental MA representation is the only one satisfying $|a(z)| = 0$ only at values of z on or outside the unit circle (i.e. greater than or equal to one in absolute value). (Though for the scalar, one-lag case complex roots can't occur.) It turns out this rule applies also to models with more lags and with y a vector (so that $|a(z)|$ is the determinant of a matrix).

Inference for finite order MA's

- AR models, which we consider next, are more commonly used in applied work and easier to estimate.

- One approach to estimating an MA. Form the likelihood $p(y_1, \ldots, y_T \mid a, \sigma_\varepsilon^2)$ and maximize it or sample from it as a flat-prior posterior.

Likelihood for a finite order MA

$$Y_T = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} \sim N(\mathbf{1} \otimes \mu_y, \Sigma)$$

$$\Sigma = \begin{bmatrix} R_y(0) & R_y(1) & R_y(2) & \dots & R_y(T-1) \\ R_y(-1) & R_y(0) & R_y(1) & \dots & R_y(T-2) \\ R_y(-2) & R_y(-1) & R_y(0 & \dots & R_y(T-3)) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ R_y(1-T) & \dots & R_y(-2) & R_y(-1) & R_y(0) \end{bmatrix}$$

$R_y(t)$ can be computed from a and σ_ε^2 for any t. $\mathbf{1}$ is a column or vector of 1's and μ_y is the constant mean of y_t. Note that $R_y(-t) = (R_y(t))'$.

Likelihood for a finite order MA

The expression for the Normal pdf for Y_T, treated as a function of $\mu_y, a, \sigma_\varepsilon^2$, is then the likelihood, and its log is

$$-\tfrac{1}{2}\log(|\Sigma|) - \frac{1}{2}(Y_T - \mu_y)'\Sigma^{-1}(Y_T - \mu_y)\,.$$

Whether using the likelihood to find a MLE or to do MCMC, you would want to restrict the parameters to lie in the region of the parameter space that makes the MA representation fundamental.

You should be sure you can code a likelihood function for, e.g., the case $Y_T = (1, 0, -1)'$, $a(L) = 1 + bL$, $\mu_y = 0$ as a function of b and σ_ε^2.

Detecting non-fundamental MAR's

- If $y_t = A(L)\varepsilon_t$ and $y_t = B(L)\eta_t$, with both ε_t and η_t i.i.d. $N(0, I)$, then since there is only one $R_y(t)$, we must have $R_y(L) = A(L)A'(L) = B(L)B'(L)$.

- If $A(L)$ is finite order, $B(L)$ is finite-order of the same order.

- The roots of $A'(L)$ are the inverses of the roots of $A(L)$, and same for $B(L), B'(L)$.

- Therefore the roots of $A(L)$ are either roots of $B(L)$ or inverses of roots of $B(L)$.

Flipping roots

Since in the univariate case the roots of a polynomial fully characterize it (up to a scale factor), we can in that case convert a finite-order non-fundamental moving average operator $A(L)$ to its fundamental counterpart by "flipping" all its roots that lie inside the unit circle, replacing them with their inverses. We then generally need to rescale to get variances right. E.g.:

$$y_t = \varepsilon_t + 2\varepsilon_{t-1} + .75\varepsilon_{t-2} = (1 + 2L + .75L^2)\varepsilon_t = (1 + 1.5L)(1 + .5L)\varepsilon_t \, .$$

$$y_t = (1 + \tfrac{2}{3}L)(1 + .5L)\eta_t = (1 + 1.16667L + .3333L^2)\eta_t \, .$$

The rescaling comes in because to make R_y the same for these two representations, we need $\mathrm{Var}(\eta_t) = 2.25\,\mathrm{Var}(\varepsilon_t)$. If we normalize instead by making $\mathrm{Var}(\varepsilon_t) = \mathrm{Var}(\eta_t) = 1$, then the coefficients in the fundamental polynomial have to be multiplied by 1.5.

Multivariate flipping

(*much* harder)

Some qualitative rules

- If $A(L)$ is fundamental and $B(L)$ is not, and if they are normalized so disturbance variance is the identity, then $A_0A_0' - B_0B_0'$ is positive semi-definite.

- Therefore if $A_0A_0' \succ B_0B_0'$, B is not fundamental. If neither $A_0A_0' \succ B_0B_0'$ nor $B_0B_0' \succ A_0A_0'$, neither is fundamental. Here $X \succ Y$ means "$X - Y$ is positive semi-definite".

The finite order AR class of models

$$y_t = \sum_{s=1}^{k} b_s y_{t-s} + \varepsilon_t, \qquad \text{or } b(L)y_t = \varepsilon_t \, .$$

$\varepsilon \sim N(0, \Sigma)$, or sometimes just mean 0, variance Σ, not correlated with past y's, and therefore not serially correlated.

Properties:

- Dense in the space of LR stationary processes, plus includes some types of non-stationary processes
- *Not* closed under taking linear combinations

- *Not* closed under taking subvectors.

- No uniqueness problem. Every set of real numbers used to populate b_s, $s = 1, \ldots, k$ results in a distinct model. Restrictions like that to obtain fundamental MA's are needed if we want to consider only stationary models. But this restriction is not needed to prevent redundancy.

Finite-order ARMA models

$$B(L)y_t = A(L)\varepsilon_t \, , B(0) = A(0) = I \, ,$$

where $\varepsilon_t \perp \{y_s, \, s < t\}$ (and ε_t is therefore the innovation in y at t) and B and A are finite-order polynomials in L, perhaps with matrix-valued coefficients.

Properties:

- Contains MA and AR models, so is also dense in the LR class of models.
- Closed under taking linear combinations.
- Like the finite-order AR class, contains non-stationary as well as stationary models.

- Has the same problems as the MA class with possible redundancy in the $A(L)$ parameter space.

- Has the same problem as the AR class with the restrictions on $B(L)$ needed if we want to restrict to stationary models.

- Has its own special, severe problem of non-uniqueness, because of possible cancellation between AR and MA roots.

ARMA root cancellation

This is easiest to see in the univariate case. Then a model $A(L)y = B(L)\varepsilon$ can be written

$$(1-\rho_1 L)(1-\rho_2 L)\cdots(1-\rho_m L)y = \alpha\cdot(1-\nu_1 L)(1-\nu_2 L)\cdots(1-\nu_k L)\varepsilon\,,$$

where the ρ's are the inverses of the roots of the A polynomial and the ν's are the inverses of the roots of the B polynomial. If for some i, j $\rho_i = \nu_j$, the corresponding terms in the equation cancel, implying lower-order A and B polynomials. But if we fix the orders m and k and try to optimize fit over the coefficients in the two polynomials the cancellation will show up in indeterminacy of the optimal coefficients, so attempts at MLE will run in to numerical problems.

The practical implication is that one usually chooses only m or k freely to optimize fit, keeping the other polynomial of low order.

AR and MA representations

- When B is a finite-order polynomial and there is no root z of $|B(z)| = 0$ with $|z| = 1$, and when ε_t are i.i.d. with finite variance,

$$y_t = B^{-1}(L)\varepsilon_t \Rightarrow B(L)y_t = \varepsilon_t$$
$$y_t = B(L)\varepsilon_t \Rightarrow B^{-1}(L)y_t = \varepsilon_t$$

  Here we are allowing both B and B^{-1} to have non-zero coefficients on negative, as well as positive, powers of L, and insisting that both be convergent. This implies there is a unique inverse for $B(L)$ under our assumption of no roots on the unit circle.

- In fact these relations are more general. $B(z)$ can be defined even for infinite-order z, so long as the coefficients go to zero sufficiently fast, and therefore the finite-order requirement can be dropped.

AR and MA representations, cont.

When B is a finite-order polynomial in non-negative powers of L and y_t has finite variance, ε_t are of finite variance, and ε_t is uncorrelated with y_{t-s}, all $s > 0$,

$$B(L)y_t = \varepsilon_t \Rightarrow y_t = B^{-1}(L)\varepsilon_t \,,$$

where $B^{-1}(L)$ is the inverse of B in non-negative powers of L, which might not converge. (Its coefficients converge if all the roots of $|B(z)| = 0$ are outside the unit circle.)

Obviously y can be stationary, with the ε_t process stationary, only if $B^{-1}(L)$ does converge. If not, the variances of the ε_t's must shrink as we go back in the past. Most commonly, in non-stationary models we think of ε_t as zero for all t less than some initial date t_0, and ε_t i.i.d. for $t > t_0$.

Stacking an AR model into first-order form

$$y_t = B(L)y_t + \varepsilon_t \text{ , } \varepsilon_t \text{ the innovation in } y_t$$

$$y_t = \sum_{s=1}^{k} B_s y_{t-s} + \varepsilon_t$$

$$\begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-k+1} \end{bmatrix} = \begin{bmatrix} B_1 & B_2 & \dots & B_{k-1} & B_k \\ & & I & & 0 \\ & & {\scriptstyle n(k-1)\times n(k-1)} & & \end{bmatrix} \begin{bmatrix} y_{t-1} \\ \vdots \\ y_{t-k} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$Y_t = BLY_t + \nu_t \text{ , } \qquad (I - BL)Y_t = \nu_t$$

The eigenvalues of the $nk \times nk$ B matrix are the inverses of the roots of $|I - B(z)|$.

$E[y]$ and $R_y(0)$ for a stationary AR

- Whereas for an MA model computing the full R_y function is straightforward polynomial matrix multiplication, finding R_y is more work for an AR model.

- In the first-order case, once we know $R_y(0)$, we can find $R_y(s) = B^s\, R_y(0)$ for all $s > 0$, and use the fact that $R_y(-s) = (R_y(s))'$. But for $R_y(0)$ we need to solve

$$R_y(0) = \mathrm{Cov}(By_{t-1} + \varepsilon_t) = B\, R_y(0) B' + \Sigma_\varepsilon\,,$$

  which is a system of linear equations in the elements of $R_y(0)$.

$E[y]$ and $R_y(0)$ for a stationary AR

- It's a big system, and there's a literature on ways to solve it speedily. One name for it is "Lyapunov equation".

- For small models,

$$(I - B \otimes B)\overrightarrow{R_y(0)} = \overrightarrow{\Sigma_\varepsilon}\,,$$

  where the arrows indicate stacking of the columns of a matrix to form a vector, can be solved. Because $R_y(0)$ is symmetric the system as written above has more equations than unknowns, so the system can be made somewhat smaller by dropping redundant rows and columns.

- Another approach uses the fact that for a stationary system we can derive

$$R_y(0) = \sum_{s=0}^{\infty} B^s \Sigma_\varepsilon (B')^s \, .$$

If we let $\Omega_0 = \Sigma_\varepsilon$, $W_0 = B$ and calculate recursively

$$\Omega_j = \Omega_{j-1} + W_j \Omega_{j-1} W_j' \, , \qquad W_{j+1} = W_j^2 \, ,$$

then $\Omega_j \to R_y(0)$ rapidly, unless the largest root of B is very close to one.

$E[y]$ and $R_y(0)$ for a stationary AR

- If the system has a constant term c, the mean $\bar{y}$ of the stationary process is found by solving

$$\bar{y} = B\bar{y} + c \Rightarrow \bar{y} = (I - B)^{-1}c\,.$$

Inference for stationary AR models

- The same idea we have discussed for MA inference, forming the Gaussian pdf for Y_T, can apply, now that we know how to form the mean and covariance matrix.

Differencing to allow for the possibility that the model is non-stationary?

- By replacing y_t with Δy_t we may make the series stationary, even though some components of the original y_t might be non-stationary. Even if y_t is stationary, Δy_t is just a new stationary process. So why not just difference to be sure of stationarity?

- One problem with this: If y_t was a finite order AR, Δy_t will not be. It will have an MA component, because of the differencing of the error term.

- Still AR's are dense, so if the differencing induces stationarity, a finite order AR may still give a good approximate fit.

Differencing discards information

- If we are differencing because the data are *possibly* non-stationary, we have thrown out all information about mean values. In general, information about low-frequency (i.e. slowly varying) components of the data is lost or de-emphasized when we fit to differenced data.

- Low-frequency connections between variables may be of central interest, and differencing makes it hard to recover them. (We return to this issue under the heading of "cointegration".)

Conditioning on initial values

The pdf of $y_{k+1}, \ldots, y_T$, conditional on $y_1, \ldots, y_k$, when y is a k'th order AR, is given by

$$\prod_{s=k+1}^{T} p(y_s \mid y_{s-k}, \ldots y_{s-1}) \, .$$

If the disturbances are normal with constant variance, we know how to give this density function an explicit form, and it is the same form we get for a standard normal linear regression model with the constant and lagged y variables treated as right-hand side "X_t" variables. The MLE is OLS, and the usual OLS formulas for standard errors correctly characterize the posterior distribution.

Note that this is true, under normality assumptions, *regardless of whether the model is stationary or not*. This is why inference in time series with possible nonstationarity makes frequentist and Bayesian approaches different, even asymptotically. See the "Helicopter Tour" paper for intuition on this point. We will return to models with non-stationarity in later lectures.

Problems with conditioning on initial conditions

- This approach also discards information. It amounts to acting as if the distribution of the initial conditions $y_1, \ldots, y_k$ has nothing to tell us about the model parameters.

- But if the model is stationary, the distribution of the initial conditions depends strongly on the parameters and thus has a lot to tell us.