The mathematical structure of ARIMA models

Robert Nau

Fuqua School of Business, Duke University

1. The backshift operator

In the rules for identifying the orders of AR and MA terms in ARIMA models (people.duke.edu/~rnau/411arim3.htm and people.duke.edu/~rnau/arimrule.htm), some of the following claims were made:

- An AR term can "mimic" a 1st difference.
- An MA term can "moderate" a first difference.
- A redundant pair of AR and MA terms may "cancel each other out."
- A first-order MA model is equivalent to a higher-order AR model, and vice versa.

The underlying reasons for these facts--which are by no means obvious--will now be explained. This will require some mathematical sleight-of-hand.

Introducing. . .

B

. . .the *backward shift* ("backshift") operator.

B is defined to perform the following operation: *it causes the observation that it multiplies to be shifted backwards in time by 1 period.* That is, for any time series Y and any period t:

$$BY_t = Y_{t-1} .$$

Multiplication by higher powers of B correspondingly yields a backward shift by more than 1 period:

$$B^2Y_t = B(BY_t) = B(Y_{t-1}) = Y_{t-2}$$

Main web site: people.duke.edu/~rnau/forecasting.htm

…and in general, for any integer n:

$$B^nY_t = Y_{t-n}$$

Thus, multiplying by B-to-the-n^th^-power has the effect of shifting an observation backwards by n periods.

The first-difference operation has a simple representation in terms of B. Suppose that y is the first difference of Y. Then, for any t:

$$y_t = Y_t - Y_{t-1} = Y_t - BY_t = (1-B)Y_t$$

Thus, the differenced series y is obtained from the original series Y by *multiplying by a factor of* $1-B$. Now, if z is the first difference of y, i.e., z is the *second* difference of Y, then we have:

$$z_t = y_t - y_{t-1} = (1-B)\, y_t = (1-B)((1-B)Y_t) = (1-B)^2Y_t\,.$$

The second difference of Y is therefore obtained by multiplying by a factor of $(1-B)^2$, and in general the d^{th} difference of Y would be obtained by multiplying by a factor of $(1-B)^d$.

Note that what we are doing here is manipulating the operation of shifting-in-time as though it were a numeric variable in an equation. This is perfectly legitimate, because B is a *linear operator*, and we will see shortly that such manipulations can be pushed to crazy-but-useful lengths.

Armed with B, let's reconsider the ARIMA(1,1,1) model for the time series Y. For convenience, I will omit the constant term from this model and all the models discussed below. (The presence of a constant term would not change the basic arguments, but it would complicate the details.) The ARIMA(1,1,1) model (sans constant) is defined by the following pair of equations:[2]

$$y_t = Y_t - Y_{t-1}$$

$$y_t = \phi_1\, y_{t-1} + e_t - \theta_1\, e_{t-1}$$

…where e_t is the random shock (noise) occurring at time t. Armed with the backshift operator, we can now rewrite them as follows:

$$y_t = (1-B)Y_t$$

$$y_t = \phi_1 By_t + e_t - \theta_1 Be_t$$

By collecting y's on the left and e's on the right, the second equation can be rewritten as:

$$(1-\phi_1B)y_t = (1-\theta_1B)e_t$$

We can now substitute the expression for y in terms of Y that was given by the first equation, obtaining a single equation involving Y_t and e_t that summarizes the ARIMA(1,1,1) model:

$$\boxed{(1-\phi_1B)(1-B)Y_t = (1-\theta_1B)e_t}$$

---

[2] Some authors define θ_1 (and higher-order MA coefficients, if any) to have the opposite sign, so that this equation would be $y_t = \phi_1\, y_{t-1} + e_t + \theta_1\, e_{t-1}$. The convention followed here is the one used by Box and Jenkins.

To put this in perspective, recall our most basic forecasting model, namely the mean model. In the special case where the mean is assumed to be zero, this model simply asserts that "Y is stationary white noise," i.e.:

$$Y_t = e_t$$

In our new jargon, we could call this model an ARIMA(0,0,0) model. Now, the ARIMA(1,1,1) model is merely obtained by adding bells and whistles to it. Instead of "Y_t equals e_t," the ARIMA(1,1,1) model asserts that "something times Y_t " equals "something times e_t." In particular:

- Including a first difference is equivalent to multiplying Y_t by a factor of $1-B$
- Including an AR(1) term is equivalent to multiplying Y_t by a factor of $1-\phi_1 B$
- Including an MA(1) term is equivalent to multiplying e_t by a factor of $1-\theta_1 B$

This explains some of the facts about ARIMA models that were noted above:

- If the AR(1) coefficient, denoted ϕ_1, is close to 1, then the factor $1-\phi_1$ B on the left side of the ARIMA equation is almost the same as a factor of $1-B$. Now, each factor of $1-B$ appearing on the left side of the equation represents an order of differencing. Hence the AR(1) term is mimicking an additional order of differencing if its estimated coefficient turns out to be close to 1.
- If the MA(1) coefficient, denoted θ_1, is close to 1, then the factor $1-\theta_1 B$ on the right side of the equation is approximately the same as a factor of $1-B$. In this case, the factor of $1-B$ on the left side of the equation (representing a first difference) is "almost cancelled" by factor of $1-\theta_1 B$ on the right. Thus, an MA(1) term can seemingly reduce the order of differencing if its estimated coefficient turns out to be close to 1.
- Suppose that the estimated value of the AR(1) coefficient, ϕ_1, turns out to be almost equal to the estimated value of the MA(1) coefficient, θ_1. Then the factor of $1-\phi_1 B$ on the left side of the ARIMA equation is essentially cancelled by the factor of $1-\theta_1 B$ on the right. This is what you would expect to happen if neither the AR(1) term nor the MA(1) term really belonged in the model in the first place--i.e., if they were both redundant. (More about this below…)

2. Equivalence of pure-AR and pure-MA models

Now, let's try something a little more far-out. Consider the pure MA(1) model. Its equation, in backshift-operator notation, is:

$$Y_t = (1-\theta_1 B)e_t$$

Can this model be rewritten as a pure AR model? In order to do this, we need to rearrange it so as to get e_t by itself on the right side of the equation, and "something times Y_t" on the left. But how do we determine what the "something" is? Well, by our usual method of solving equations, we evidently need to *multiply through by a factor of 1 divided by* $(1-\theta_1 B)$, which may be denoted $(1-\theta_1 B)^{-1}$. We thereby obtain:

$$(1-\theta_1 B)^{-1} Y_t = (1-\theta_1 B)^{-1}(1-\theta_1 B)e_t$$

The factor of $(1-\theta_1 B)^{-1}$ then cancels the factor of $(1-\theta_1 B)$ on the right, leaving:

$$(1-\theta_1 B)^{-1} Y_t = e_t.$$

This is fine, except… how do we interpret the factor of $(1-\theta_1 B)^{-1}$ now appearing on the left? What is the meaning of "1 divided by 1 minus θ_1 times the backshift operator?" Well, not to worry. Perhaps you recall the formula for the infinite geometric series:

$$(1-r)^{-1} = 1 + r + r^2 + r^3 + r^4 + \ldots$$

. . .which is valid for $|r| < 1$. Let's just let $r = \theta_1 B$, and see what happens:

$$(1-\theta_1 B)^{-1} = 1 + \theta_1 B + \theta_1^2 B^2 + \theta_1^3 B^3 + \theta_1^4 B^4 + \ldots$$

Voila! The factor of "1 divided by something-involving-the-backshift-operator" has turned into a sequence of powers of B, which we know how to interpret. The pure-AR form of the MA(1) model can therefore be written as:

$$(1 + \theta_1 B + \theta_1^2 B^2 + \theta_1^3 B^3 + \theta_1^4 B^4 + \ldots) Y_t = e_t$$

. . .which, in view of the definition of B, really means:

$$Y_t = -\theta_1 Y_{t-1} - \theta_1^2 Y_{t-2} - \theta_1^3 Y_{t-3} - \theta_1^4 Y_{t-4} - \ldots + e_t$$

This is *an infinite-order pure-AR model*, with the AR(1) coefficient being $-\theta_1$, the AR(2) coefficient being $-\theta_1^2$, and the AR(n) coefficient being $-\theta_1^n$ for n=3, 4, etc. to infinity.

By the same reasoning, a pure AR(1) model, whose single coefficient is ϕ_1, is equivalent to an infinite-order pure-MA model, in which the MA(1) coefficient is $-\phi_1$, the MA(2) coefficient is $-\phi_1^2$, the MA(3) coefficient is $-\phi_1^3$ and so on.

The practical significance of this is that *it can be difficult to tell the difference between an MA(1) model and an AR(2) model, or between and AR(1) model and an MA(2) model, if the first-order coefficients are not large.* For example, suppose that the "true" model for the time series is pure MA(1) with $\theta_1 = 0.3$. This is equivalent to an infinite-order pure-AR model with:

$$\phi_1 = -\theta_1 = -0.3$$

$$\phi_2 = -\theta_1^2 = -0.09$$

$$\phi_3 = -\theta_1^3 = -0.027$$

$$\phi_4 = -\theta_1^4 = -0.0081$$

…and so on. Note that the AR coefficients are all negative, and their magnitudes have an exponentially decreasing pattern. Because of the rapidity of the exponential decrease, ϕ_3 and ϕ_4 and all higher-order coefficients are not significantly different from zero, at least within the precision with which ARIMA coefficients can be estimated from typical-sized data sets, and ϕ_2 is not very large either. Hence, for practical purposes, this is a 2nd-order AR model rather than an infinite-order AR model. But in such a case, you should generally choose the simpler representation (here, MA(1)) on grounds of simplicity.

3. The danger of overfitting a mixed AR-MA model: redundancy and cancellation

Now let's consider the issue of a pair of AR and MA factors cancelling each other out (a so-called "common factor" problem). Suppose the time series Y is really an ARIMA(1,d,0) process, but instead you attempt to fit an ARIMA(2,d,1) model. The ARIMA(2,d,1) model has the equation:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t - \theta_1 e_{t-1}$$

where $y_t = (1\text{–}B)^d Y_t$. In terms of the backshift operator this can be rewritten as:

$$(1 - \phi_1 B - \phi_2 B^2)\, y_t = (1 - \theta_1 B) e_t .$$

Note that the object multiplying y_t on the left side is a second-order polynomial in B, which is called the "AR polynomial" for this model. Like any polynomial, it can be factored into a product of first-order terms:

$$(1 - \phi_1 B - \phi_2 B^2) = (1 - \lambda_1 B)(1 - \lambda_2 B)$$

Here, λ_1 and λ_2 are two numbers that are called the "inverse roots" of the AR polynomial. In practice, these can be determined by applying a specialized form of the quadratic formula:

$$\lambda = (\phi_1 \pm \mathrm{SQRT}(\phi_1^2 + 4\phi_2)))/2$$

For example, if $\phi_1 = 0.8$ and $\phi_2 = -0.15$, we find:

$$\lambda = (0.8 \pm \mathrm{SQRT}(0.8^2 + 4(-0.15)))/2$$

$$= (0.8 \pm \mathrm{SQRT}(0.04))/2$$

$$= (0.8 \pm 0.2)/2$$

...which has solutions $\lambda_1 = 0.5$ and $\lambda_2 = 0.3$. Thus, the AR(2) polynomial in this case can be factored as follows:

$$(1 - 0.8B + 0.15B^2) = (1 - 0.5B)(1 - 0.3B)$$

Note that if the quantity whose square root is computed in the quadratic formula, namely $\phi_1^2 + 4\phi_2$, is *negative*, then the inverse roots will be *complex numbers*, i.e., partly "real" and partly "imaginary." This occurs in AR(2) processes exhibiting sine-wave oscillations.

By factoring the AR polynomial, the ARIMA(2,d,1) equation can be rewritten:

$$(1 - \lambda_1 B)(1 - \lambda_2 B)\, y_t = (1 - \theta_1 B) e_t$$

It is now apparent how cancellation might occur: if $\lambda_2 = \theta_1$, then the AR factor of $1\text{–}\lambda_2 B$ on the left is cancelled by the MA factor of $(1\text{–}\theta_1 B)$ on the right, so the equation reduces to:

$$(1 - \lambda_1 B)\, y_t = e_t$$

which is an ARIMA(1,d,0) model.

For example, suppose the true AR(1,d,0) model has $\phi_1 = 0.5$, i.e.:

$$(1 - 0.5B)y_t = e_t$$

This is equivalent to an ARIMA(2,d,1) model with $\phi_1 = 0.8$, $\phi_2 = -0.15$, and $\theta_1 = 0.3$:

$$(1 - 0.8B + 0.15B^2)y_t = (1 - 0.3B)e_t$$

…since, as we saw above, the AR(2) polynomial on the left can be factored as (1–0.5B)(1–0.3B). Thus, one AR factor (namely 1–0.5B) represents the true autoregressive effect in this case, and the other AR factor (namely 1–0.3B) merely serves to cancel out the superfluous MA(1) factor.

The problem here is that if the true data generating process is ARIM(1,d,0) but the model is specified as ARIMA(2,d,1), it is not uniquely identified, because many ARIMA(2,d,1) models can be equivalent to the same ARIMA(1,d,0) model. For example, consider the ARIMA(2,d,1) model with $\phi_1 = 0.9$, $\phi_2 = -0.2$, and $\theta_1 = 0.4$:

$$(1 - 0.9B + 0.2B^2)y_t = (1 - 0.4B)e_t$$

This can be factored as:

$$(1 - 0.5B)(1 - 0.4B)\, y_t = (1 - 0.4B)e_t$$

…which, after cancellation, again reduces to:

$$(1 - 0.5B)\, y_t = e_t$$

So, if you try to fit an ARIMA(2,d,1) model to a time series that is really ARIMA(1,d,0), the least-squares coefficient estimates will not be unique. Sometimes (but not always) this problem will be signaled by the fact that the estimation algorithm will fail to converge in a reasonable number of iterations (say, less than 10) and/or it will yield suspiciously large values for the coefficients.

The same reasoning would apply if you attempted to fit an ARIMA(1,d,2) model to a time series that was really (0,d,1): in this case the MA side of the model would be a 2nd-order polynomial in B, with one superfluous factor. More generally, a common-factor problem may arise whenever the true model is ARIMA(p,d,q) but instead you attempt to fit an ARIMA(p+1,d,q+1) or ARIMA(p,d+1,q+1) model. Note, however, that in a redundant mixed model in which at least one side (AR or MA) is 2nd- or higher-order, *the common factor may not be obvious*. Instead, it may be buried in the roots of the AR and/or MA polynomials.

The moral: *beware of trying to estimate multiple AR coefficients* <u>*and*</u> *multiple MA coefficients simultaneously,* because this may just lead to cancellation of factors on both sides of the equation. Instead, you should (i) try to use pure-AR or pure-MA models, rather than mixed AR-MA models, unless the data clearly indicates otherwise; and (ii) use the "forward stepwise" approach to model-identification, rather than "backwards stepwise." You may sometimes find a series which, after suitable differencing, has spikes at lags 1 and 2 in *both* the ACF and PACF, making it hard to choose between an ARIMA(2,d,0) model and a ARIMA(0,d,2) model. In this case, it may happen that an ARIMA(1,d,1) model gives the best fit. Since there is only one AR coefficient and one MA coefficient, it is easy to detect a problem with cancellation: if the estimated AR and MA coefficients are nearly equal, then cancellation has occurred. Otherwise, it hasn't. In particular, if the estimated AR and MA coefficients turn out to have opposite signs, there is no problem. You sometimes encounter situations in which the best model has 2 or 3 terms of one kind (AR or MA) and 1 of the other, e.g., ARIMA(3,0,1) or ARIMA(1,1,2). (The latter is a damped-trend linear exponential smoothing model.) But you should avoid using models with 2 or more terms of *both* kinds unless you have a good-sized sample of clean data and it was generated by a process that can be reasonably expected to have very stable dynamics (which is often not true of data in business and economics). ARIMA (2,1,2) models are commonly among the suspects tested by automatic forecasting software, and if you are tempted to use such a model, be sure to take a close look at the inverse roots to look for cancellation or nonstationarity or noninvertibility, and see if you get almost as good a fit, as well as plausible-looking forecasts and confidence limits, by reducing q by 1 while simultaneously reducing either p or d by 1. Also think about how you would explain the logic of the model to someone else, and do some more research to see what models others have applied to similar data, but be skeptical about models that are overly complex. (I've seen articles on the web proposing models such as (3,1,3) for economic data. Don't go there.)

4. Unit roots and the Dickey-Fuller tests

The "roots" of an AR(p) polynomial are the real or complex numbers {r} that satisfy $(1 - \phi_1 r - \phi_2 r^2 - \ldots - \phi_p r^p) = 0$, and they must all lie *outside the unit circle* in order for the model for y to be *stationary*, i.e., to yield forecasts that eventually converge to the mean when they are extrapolated far into the future. (Equivalently, the *inverse* roots must all lie *inside* the inside circle.) Furthermore, the roots of an MA(q) polynomial $(1 - \theta_1 r - \theta_2 r^2 \ldots - \theta_q r^q)$ must all lie outside the unit circle in order for the model to be *invertible,* i.e., capable of estimating the "true" errors or shocks that gave rise to the observed series. If the model is not invertible, the residuals cannot be considered as estimates of the true random shocks.

Some software packages (alas, not Statgraphics) will routinely compute and print out the roots (or inverse roots) of any 2nd- or higher-order AR or MA polynomials in the model, so that you can watch out for possible cancellation, non-stationarity and/or non-invertibility. Of course, for an MA(1) or AR(1) polynomial, the (single) inverse root is simply the coefficient, θ_1 or ϕ_1 .

If one of the roots (or inverse roots) of the AR polynomial is almost exactly equal to 1, the AR part of the model is said to have a "unit root". In an AR(1) model, this occurs when the single estimated coefficient ϕ_1 is equal to 1. In this case the data is telling you that another difference is what is really needed, not an AR(1) term. At the present order of differencing, it is behaving like a random walk process.

In a 2nd- or higher-order AR model, the same reasoning applies to the roots of the AR polynomial. If one of the roots is almost exactly to 1, then the time series has not been adequately stationarized, and it probably would be better to use one *more* nonseasonal difference and one *less* AR coefficient in the model. For example, if an ARIMA(2,0,0) model has a unit root in the AR polynomial, it would probably be better to use an ARIMA(1,1,0) model instead.

There is an easy way to check for the occurrence of a unit root in the AR polynomial: *just add up all the AR coefficients*. If the sum is very close to 1.0, then the polynomial may have a unit root, indicating that you ought to try increasing the order of nonseasonal differencing and decreasing the number of AR terms. How close? Within 0.05 is a rough rule of thumb for a red flag, but for a bit more rigor, you can perform a *statistical test* for the presence of a unit root by checking to see whether the difference between 1 and the sum of the AR coefficients is *significantly* different from zero. This is known as the *Dickey-Fuller test* in the case of an AR(1) model, and the *augmented Dickey-Fuller test* in the case of an AR(p) model with p>1. Let Y denote the time series of interest and let y denote its first difference. For the case in which Y is believed to be an AR(1) process, the Dickey-Fuller test for the presence of a unit root is to regress y on Y lagged by 1 period. If there is a unit root, then the coefficient of Y lagged by 1 period in this model should not be significantly different from zero, which can be tested in the usual way by looking at its t-statistic. For the case in which Y is believed to be an AR(p) process, p>1, the augmented Dickey-Fuller test for the presence of a unit root is to regress y on itself lagged by 1 period, 2 periods, etc., up to p-1 periods, together with Y lagged by 1 period. (In the AR(2) case, the independent variables would just be y lagged by 1 period and Y lagged by 1 period.) Again, if there is a unit root, then the coefficient of Y lagged by 1 period should not be significantly different from zero, which can be tested by looking at its t-statistic.

In asking whether there is a unit root in the AR polynomial, you should also consider whether it is *logical* that the series should be stationary (mean-reverting) at its current level of differencing. If it *is* logical to think that it is stationary but with only very slow mean reversion, then another difference might not be appropriate, even if there is an AR root that is very close to 1.

The same consideration applies to a regression model fitted to time series data. If Y is the dependent variable and its first two lagged values (LAG(Y,1) and LAG(Y,2) in Statgraphics notation, or Y_LAG1 and Y_LAG2 in RegressIt notation) are included as independent variables, then a unit root is indicated if the sums of their coefficients are not significantly different from 1. In this case it might be better to use the first difference of Y as the dependent variable and to use its lag-1 value but not its lag-2 value as an independent variable.

The *opposite* consideration applies to the MA polynomial in 2nd- or higher-order MA models. If the MA polynomial contains a unit root, it is essentially *cancelling out* one order of nonseasonal differencing. In this situation (which you can test for by adding up the MA coefficients to see if the sum is close to 1.0), you should consider using one *less* nonseasonal difference and one less MA term. Here too, you should stop and think about what is logical. A nonseasonal difference combined with MA terms whose sum is very close to 1 is like an exponential smoothing model that is computing a very long-term moving average without assuming long-run reversion to a global mean. If long-run mean reversion makes more sense, then it would be better to reduce the order of differencing.

5. Stationarity and invertibility of 2nd-order AR or MA models

More generally, stationarity requires the roots of the AR polynomial to be *outside* the unit circle, not merely off of it. (Equivalently, the inverse roots must be inside the unit circle.) The roots of an AR(2) polynomial are outside the unit circle, and hence the model is stationary, if and only if the following constraints are satisfied:

$$\phi_2 + \phi_1 < 1$$

$$\phi_2 - \phi_1 < 1.$$

$$-1 < \phi_2 < 1$$

These inequalities cover the case of complex roots as well as real roots. A unit-root situation is the special case in which $\phi_2 + \phi_1 = 1$ in violation of the first of the inequalities.

Similarly, to determine whether the roots of an MA(2) polynomial are outside the unit circle, in order to verify invertibility, you would apply these conditions with θ_1 and θ_2 in place of ϕ_1 and ϕ_2. (Caution: be sure you know which convention your software uses for the sign of MA coefficients. See footnote 2 on page 2.)

Again, nonstationarity suggests that a higher order of differencing should be considered, while noninvertibility suggests that a lower order should be considered.