Chapter 7

Cycles II: The Wold Representation and Its Approximation

This Chapter is a bit more abstract than most, but don't be put off. On the contrary, you may want to read it several times. The material in it is crucially important for time series modeling and forecasting and is therefore central to our concerns. In some parts (finite-ordered autoregressive models) it largely repeats Chapter 6, but that's intentional. It treats much more, including the Wold representation and its approximation and prediction using finite-ordered autoregressions, finite-ordered moving averages, and finite-ordered ARMA processes. Hence even the overlapping material is presented and integrated from a significantly more sohpisticated perspective.

7.1 The Wold Representation and the General Linear Process

7.1.1 The Wold Representation

Many different dynamic patterns are consistent with covariance stationarity. Thus, if we know only that a series is covariance stationary, it's not at all clear what sort of model we might fit to describe its evolution. The trend and seasonal models that we've studied aren't of use; they're models of specific

nonstationary components. Effectively, what we need now is an appropriate model for what's left after fitting the trend and seasonal components – a model for a covariance stationary residual. **Wold's representation theorem** points to the appropriate model.

Theorem:

Let {yt} be any zero-mean covariance-stationary process.[1] Then we can write it as

$$y_t = B(L)\varepsilon_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

where

$$b_0 = 1$$

$$\sum_{i=0}^{\infty} b_i^2 < \infty.$$

In short, the correct "model" for any covariance stationary series is some infinite distributed lag of white noise, called the **Wold representation**. The $\varepsilon_t's$ are often called **innovations**, because (as we'll see) they correspond to the 1-step-ahead forecast errors that we'd make if we were to use a particularly good forecast. That is, the $\varepsilon_t's$ represent that part of the evolution of y that's linearly unpredictable on the basis of the past of y. Note also that the $\varepsilon_t's$, although uncorrelated, are not necessarily independent. Again, it's only for Gaussian random variables that lack of correlation implies independence, and the innovations are not necessarily Gaussian.

In our statement of Wold's theorem we assumed a zero mean. That may seem restrictive, but it's not. Rather, whenever you see y_t, just read $(y_t - \mu)$, so that the process is expressed in deviations from its mean. The deviation from the mean has a zero mean, by construction. Working with zero-mean

[1] Moreover, we require that the covariance stationary processes not contain any deterministic components.

processes therefore involves no loss of generality while facilitating notational economy. We'll use this device frequently.

7.1.2 The General Linear Process

Wold's theorem tells us that when formulating forecasting models for covariance stationary time series we need only consider models of the form

$$y_t = B(L)\varepsilon_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

where the b_i are coefficients with $b_0 = 1$ and $\sum_{i=0}^{\infty} b_i^2 < \infty$.

We call this the **general linear process**, "general" because any covariance stationary series can be written that way, and "linear" because the Wold representation expresses the series as a linear function of its innovations.

The general linear process is so important that it's worth examining its unconditional and conditional moment structure in some detail. Taking means and variances, we obtain the unconditional moments

$$E(y_t) = E(\sum_{i=0}^{\infty} b_i \varepsilon_{t-i}) = \sum_{i=0}^{\infty} b_i E\varepsilon_{t-i} = \sum_{i=0}^{\infty} b_i \cdot 0 = 0$$

and

$$var(y_t) = var(\sum_{i=0}^{\infty} b_i \varepsilon_{t-i}) = \sum_{i=0}^{\infty} b_i^2 var(\varepsilon_{t-i}) = \sum_{i=0}^{\infty} b_i^2 \sigma^2 = \sigma^2 \sum_{i=0}^{\infty} b_i^2.$$

At this point, in parallel to our discussion of white noise, we could compute and examine the autocovariance and autocorrelation functions of the general linear process. Those calculations, however, are rather involved, and not particularly revealing, so we'll proceed instead to examine the conditional mean and variance, where the information set Ω_{t-1} upon which we condition

contains past innovations; that is,

$$\Omega_{t-1} = \varepsilon_{t-1},\ \varepsilon_{t-2},\ ....$$

In this manner we can see how dynamics are modeled via conditional moments.[2] The conditional mean is

$$E(y_t|\Omega_{t-1}) = E(\varepsilon_t|\Omega_{t-1}) + b_1 E(\varepsilon_{t-1}|\Omega_{t-1}) + b_2 E(\varepsilon_{t-2}|\Omega_{t-1}) + ...$$

$$= 0 + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + ... = \sum_{i=1}^{\infty} b_i\varepsilon_{t-i},$$

and the conditional variance is

$$var(y_t|\Omega_{t-1}) = Ey_t - E(y_t|\Omega_{t-1}))^2|\Omega_{t-1}] = E(\varepsilon_t^2|\Omega_{t-1}) = E(\varepsilon_t^2) = \sigma^2.$$

The key insight is that the conditional mean *moves* over time in response to the evolving information set. The model captures the dynamics of the process, and the evolving conditional mean is one crucial way of summarizing them. An important goal of time series modeling, especially for forecasters, is capturing such conditional mean dynamics – the unconditional mean is constant (a requirement of stationarity), but the conditional mean varies in response to the evolving information set.[3]

7.2 Approximating the Wold Representation

When building forecasting models, we don't want to pretend that the model we fit is true. Instead, we want to be aware that we're *approximating* a

[2]Although Wold's theorem guarantees only serially uncorrelated white noise innovations, we shall sometimes make a stronger assumption of independent white noise innovations in order to focus the discussion. We do so, for example, in the following characterization of the conditional moment structure of the general linear process.

[3]Note, however, an embarrassing asymmetry: the conditional variance, like the unconditional variance, is a fixed constant. However, models that allow the conditional variance to change with the information set have been developed recently, as discussed in detail in Chapter **??**.

more complex reality. That's the modern view, and it has important implications for forecasting. In particular, we've seen that the key to successful time series modeling and forecasting is parsimonious, yet accurate, approximation of the Wold representation. Here we consider three approximations: **moving average (MA) models**, **autoregressive (AR) models**, and **autoregressive moving average ($ARMA$) models**. The three models differ in their specifics and have different strengths in capturing different sorts of autocorrelation behavior.

We begin by characterizing the autocorrelation functions and related quantities associated with each model, under the assumption that the model is "true." We do this separately for autoregressive, moving average, and ARMA models.[4] These characterizations have nothing to do with data or estimation, but they're crucial for developing a basic understanding of the properties of the models, which is necessary to perform intelligent modeling and forecasting. They enable us to make statements such as "If the data were really generated by an autoregressive process, then we'd expect its autocorrelation function to have property x." Armed with that knowledge, we use the *sample* autocorrelations and partial autocorrelations, in conjunction with the AIC and the SIC, to suggest candidate forecasting models, which we then estimate.

7.2.1 Rational Distributed Lags

As we've seen, the Wold representation points to the crucial importance of models with infinite distributed lags. Infinite distributed lag models, in turn, are stated in terms of infinite polynomials in the lag operator, which are therefore very important as well. Infinite distributed lag models are not of immediate practical use, however, because they contain infinitely many pa-

[4]Sometimes, especially when characterizing population properties under the assumption that the models are correct, we refer to them as processes, which is short for **stochastic processes**. Hence the terms moving average process, autoregressive process, and ARMA process.

rameters, which certainly inhibits practical application! Fortunately, infinite polynomials in the lag operator needn't contain infinitely many free parameters. The infinite polynomial B(L) may for example be a ratio of finite-order (and perhaps very low-order) polynomials. Such polynomials are called rational polynomials, and distributed lags constructed from them are called **rational distributed lags**.

Suppose, for example, that

$$B(L) = \frac{\Theta\ (L)}{\Phi\ (L)},$$

where the numerator polynomial is of degree q,

$$\Theta(L) = \sum_{i=0}^{q} \theta_i L^i,$$

and the denominator polynomial is of degree p,

$$\Phi(L) = \sum_{i=0}^{p} \phi_i L^i.$$

There are *not* infinitely many free parameters in the $B(L)$ polynomial; instead, there are only $p + q$ parameters (the θ's and the ϕ's). If p and q are small, say 0, 1 or 2, then what seems like a hopeless task – estimation of $B(L)$ – may actually be easy.

More realistically, suppose that $B(L)$ is not exactly rational, but is approximately rational,

$$B(L) \approx \frac{\Theta(L)}{\Phi(L)},$$

Then we can **approximate the Wold representation** using a rational distributed lag. Rational distributed lags produce models of cycles that economize on parameters (they're parsimonious), while nevertheless providing accurate approximations to the Wold representation. The popular ARMA

and ARIMA forecasting models, which we'll introduce shortly, are simply rational approximations to the Wold representation.

7.2.2 Moving Average (MA) Models

The finite-order moving average processes is a natural and obvious approximation to the Wold representation, which is an infinite-order moving average process. Finite-order moving average processes also have direct motivation: the fact that all variation in time series, one way or another, is driven by shocks of various sorts suggests the possibility of modeling time series directly as distributed lags of current and past shocks, that is, as moving average processes.[5]

The $MA(1)$ Process

The first-order moving average, or $MA(1)$, process is

$$y_t = \varepsilon_t + \theta\varepsilon_{t-1} = (1 + \theta L)\varepsilon_t$$

$$\varepsilon_t \sim WN(0,\ \sigma^2).$$

The defining characteristic of the MA process in general, and the $MA(1)$ in particular, is that the current value of the observed series is expressed as a function of current and lagged unobservable shocks – think of it as a regression model with nothing but current and lagged disturbances on the right-hand side.

To help develop a feel for the behavior of the MA(1) process, we show two simulated realizations of length 150 in Figure 7.1. The processes are

$$y_t = \varepsilon_t + .4\varepsilon_{t-1}$$

[5] Economic equilibria, for example, may be disturbed by shocks that take some time to be fully assimilated.

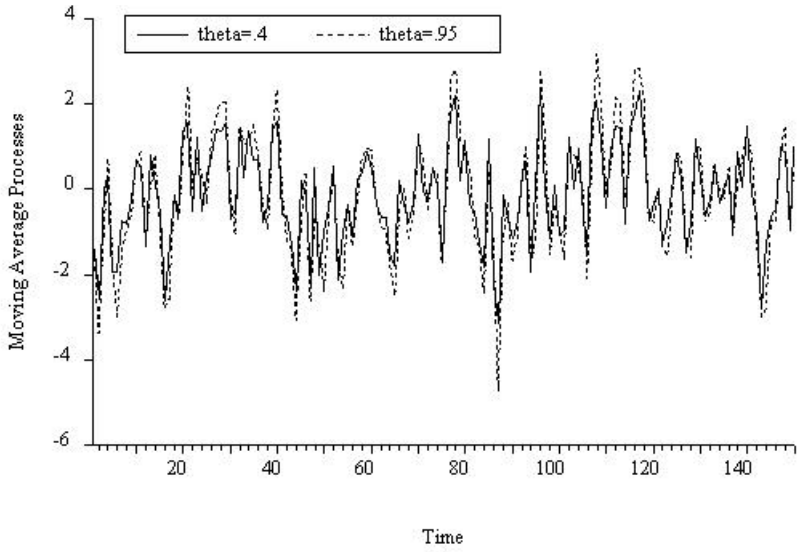


Figure 7.1: Realizations of Two MA(1) Processes

and

$$y_t = \varepsilon_t + .95\varepsilon_{t-1},$$

where in each case

$$\varepsilon_t \sim iid\ N(0,1).$$

To construct the realizations, we used the same series of underlying white noise shocks; the only difference in the realizations comes from the different coefficients. Past shocks feed *positively* into the current value of the series, with a small weight of θ=.4 in one case and a large weight of θ=.95 in the other. You might think that θ=.95 would induce much more persistence than θ=.4, but it doesn't. The structure of the $MA(1)$ process, in which only the first lag of the shock appears on the right, forces it to have a very short memory, and hence weak dynamics, regardless of the parameter value.

The unconditional mean and variance are

$$Ey_t = E(\varepsilon_t) + \theta E(\varepsilon_{t-1}) = 0$$

and

$$var(y_t) = var(\varepsilon_t) + \theta^2 var(\varepsilon_{t-1}) = \sigma^2 + \theta^2\sigma^2 = \sigma^2(1+\theta^2).$$

Note that for a fixed value of σ, as θ increases in absolute value so too does the unconditional variance. That's why the $MA(1)$ process with parameter θ=.95 varies a bit more than the process with a parameter of θ=.4.

The conditional mean and variance of an $MA(1)$, where the conditioning information set is

$$\Omega_{t-1} = \varepsilon_{t-1},\ \varepsilon_{t-2},\ ...,$$

are

$$E(y_t|\Omega_{t-1}) = E(\varepsilon_t + \theta\varepsilon_{t-1}|\Omega_{t-1}) = E(\varepsilon_t|\Omega_{t-1}) + \theta E(\varepsilon_{t-1}|\Omega_{t-1}) = \theta\varepsilon_{t-1}$$

and

$$var(y_t|\Omega_{t-1}) = Ey_t - E(y_t|\Omega_{t-1}))^2|\Omega_{t-1}] = E(\varepsilon_t^2|\Omega_{t-1}) = E(\varepsilon_t^2) = \sigma^2.$$

The conditional mean explicitly adapts to the information set, in contrast to the unconditional mean, which is constant. Note, however, that only the first lag of the shock enters the conditional mean – more distant shocks have no effect on the current conditional expectation. This is indicative of the one-period memory of $MA(1)$ processes, which we'll now characterize in terms of the autocorrelation function.

To compute the autocorrelation function for the $MA(1)$ process, we must first compute the autocovariance function. We have

$$\gamma(\tau) = E(y_t y_{t-\tau}) = E((\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-\tau} + \theta\varepsilon_{t-\tau-1})) = \begin{array}{r} \theta\sigma^2,\ \tau = 1 \\ \\ 0, otherwise. \end{array}.$$

(The proof is left as a problem.) The autocorrelation function is just the

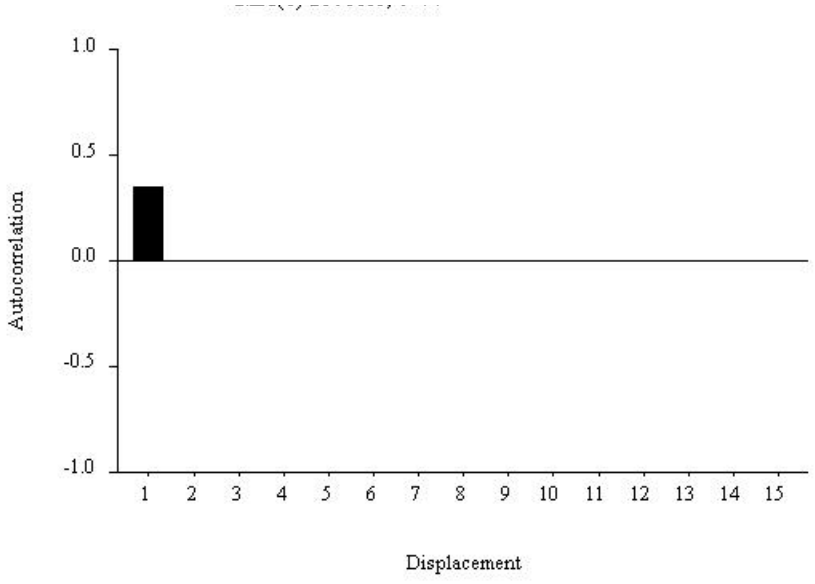


Figure 7.2: MA(1) Population Autocorrelation Function - $\theta = .4$

autocovariance function scaled by the variance,

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} = \begin{array}{l} \frac{\theta}{1+\theta^2}, \ \tau = 1 \\ \\ 0, \ otherwise. \end{array} .$$

The key feature here is the sharp *cutoff in the autocorrelations*. All autocorrelations are zero beyond displacement 1, the order of the MA process. In Figures 7.2 and 7.3, we show the autocorrelation functions for our two $MA(1)$ processes with parameters θ=.4 and θ=.95. At displacement 1, the process with parameter θ=.4 has a smaller autocorrelation (.34) than the process with parameter θ=.95, (.50) but both drop to zero beyond displacement 1.

Note that the requirements of covariance stationarity (constant unconditional mean, constant and finite unconditional variance, autocorrelation depends only on displacement) are met for any $MA(1)$ process, *regardless* of the values of its parameters. If, moreover, $|\theta| < 1$, then we say that the $MA(1)$ process is **invertible**. In that case, we can "invert" the $MA(1)$ process and express the current value of the series not in terms of a current shock and a

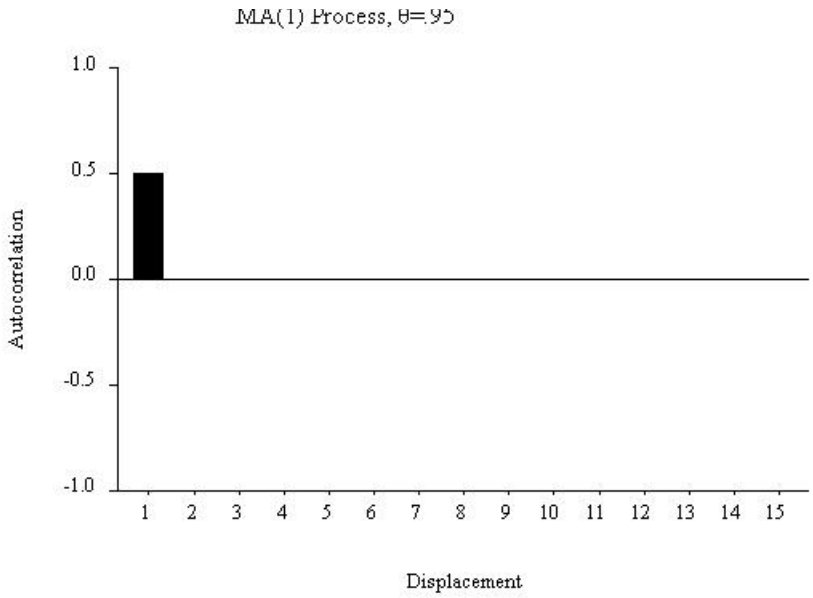


Figure 7.3: MA(1) Population Autocorrelation Function - $\theta = .95$

lagged shock, but rather in terms of a current shock *and lagged values of the series*. That's called an **autoregressive representation**. An autoregressive representation has a current shock and lagged observable values of the series on the right, whereas a moving average representation has a current shock and lagged unobservable shocks on the right.

Let's compute the autoregressive representation. The process is

$$y_t = \varepsilon_t + \theta\varepsilon_{t-1}$$

$$\varepsilon_t \sim WN(0,\ \sigma^2).$$

Thus we can solve for the innovation as

$$\varepsilon_t = y_t - \theta\varepsilon_{t-1}.$$

Lagging by successively more periods gives expressions for the innovations at various dates,

$$\varepsilon_{t-1} = y_{t-1} - \theta\varepsilon_{t-2}$$

$$\varepsilon_{t-2} = y_{t-2} - \theta\varepsilon_{t-3}$$

$$\varepsilon_{t-3} = y_{t-3} - \theta\varepsilon_{t-4},$$

and so forth. Making use of these expressions for lagged innovations we can substitute backward in the $MA(1)$ process, yielding

$$y_t = \varepsilon_t + \theta y_{t-1} - \theta^2 y_{t-2} + \theta^3 y_{t-3} - ...$$

In lag-operator notation, we write the infinite autoregressive representation as

$$\frac{1}{1+\theta L} y_t = \varepsilon_{t.}$$

Note that the back substitution used to obtain the autoregressive representation only makes sense, and in fact a convergent autoregressive representation only exists, if $|\theta| < 1$, because in the back substitution we raise θ to progressively higher powers.

We can restate the invertibility condition in another way: the inverse of the root of the moving average lag operator polynomial $(1 + \theta L)$ must be less than one in absolute value. Recall that a polynomial of degree m has m roots. Thus the $MA(1)$ lag operator polynomial has one root, which is the solution to

$$1 + \theta L = 0.$$

The root is L=-1/θ, so its inverse will be less than one in absolute value if $|\theta| < 1$, and the two invertibility conditions are equivalent. The "inverse root" way of stating invertibility conditions seems tedious, but it turns out to be of greater applicability than the $|\theta| < 1$ condition, as we'll see shortly.

Autoregressive representations are appealing to forecasters, because one way or another, if a model is to be used for real-world forecasting, it's got to link the present observables to the past history of observables, so that we can extrapolate to form a forecast of future observables based on present

and past observables. Superficially, moving average models don't seem to meet that requirement, because the current value of a series is expressed in terms of current and lagged unobservable shocks, not observable variables. But under the invertibility conditions that we've described, moving average processes have equivalent autoregressive representations. Thus, although we want autoregressive representations for forecasting, we don't have to start with an autoregressive model. However, we typically restrict ourselves to invertible processes, because for forecasting purposes we want to be able to express current observables as functions of past observables.

Finally, let's consider the partial autocorrelation function for the $MA(1)$ process. From the infinite autoregressive representation of the MA(1) process, we see that the partial autocorrelation function will decay gradually to zero. As we discussed earlier, the partial autocorrelations are just the coefficients on the last included lag in a sequence of progressively higher-order autoregressive approximations. If $\theta > 0$, then the pattern of decay will be one of damped oscillation; otherwise, the decay will be one-sided.

In Figures 7.4 and 7.5 we show the partial autocorrelation functions for our example MA(1) processes. For each process, $|\theta| < 1$, so that an autoregressive representation exists, and $\theta > 0$, so that the coefficients in the autoregressive representations alternate in sign. Specifically, we showed the general autoregressive representation to be

$$y_t = \varepsilon_t + \theta y_{t-1} - \theta^2 y_{t-2} + \theta^3 y_{t-3} - ...,$$

so the autoregressive representation for the process with θ=.4 is

$$y_t = \varepsilon_t + .4y_{t-1} - .4^2 y_{t-2} + ... = \varepsilon_t + .4y_{t-1} - .16y_{t-2} + ...,$$

and the autoregressive representation for the process with θ=.95 is

$$y_t = \varepsilon_t + .95y_{t-1} - .95^2 y_{t-2} + ... = \varepsilon_t + .95y_{t-1} - .9025y_{t-2} + ...$$

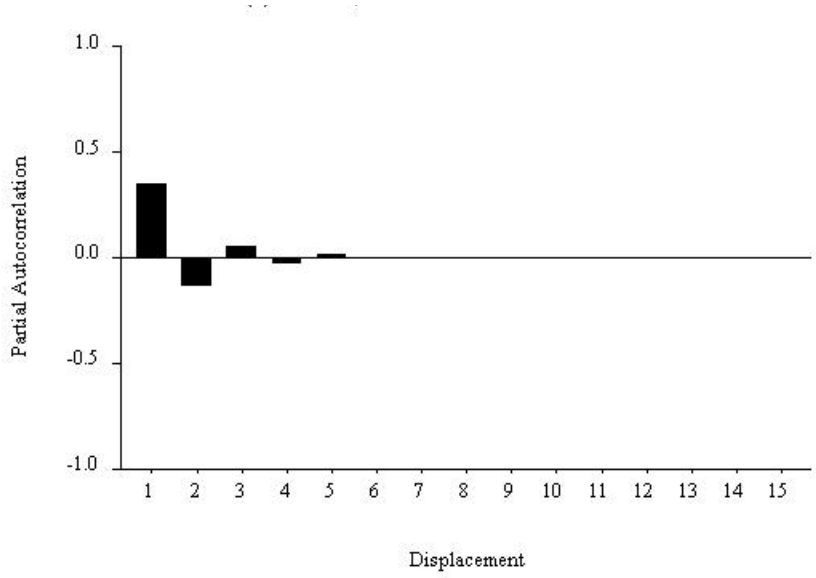


Figure 7.4: MA(1) Population Partial Autocorrelation Function - $\theta = .4$

The partial autocorrelations display a similar damped oscillation.[6] The decay, however, is slower for the θ=.95 case.

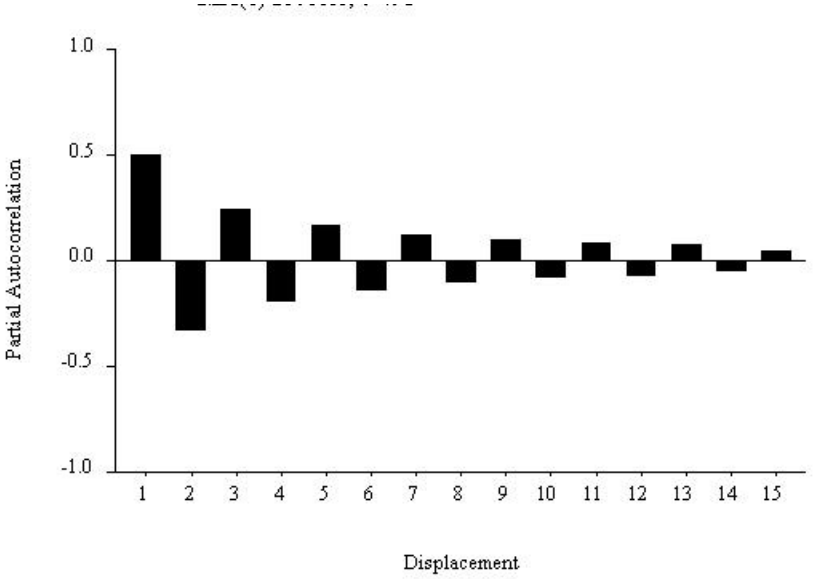


Figure 7.5: MA(1) Population Partial Autocorrelation Function - $\theta = .95$

[6]Note, however, that the partial autocorrelations are *not* the successive coefficients in the infinite autoregressive representation. Rather, they are the coefficients on the last included lag in sequence of progressively longer autoregressions. The two are related but distinct.

The $MA(q)$ Process

Now consider the general finite-order moving average process of order q, or $MA(q)$ for short,

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} = \Theta(L)\varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

where

$$\Theta(L) = 1 + \theta_1 L + ... + \theta_q L^q$$

is a qth-order lag operator polynomial. The $MA(q)$ process is a natural generalization of the $MA(1)$. By allowing for more lags of the shock on the right side of the equation, the $MA(q)$ process can capture richer dynamic patterns, which we can potentially exploit for improved forecasting. The $MA(1)$ process is of course a special case of the $MA(q)$, corresponding to $q = 1$.

The properties of the $MA(q)$ processes parallel those of the $MA(1)$ process in all respects, so in what follows we'll refrain from grinding through the mathematical derivations. Instead we'll focus on the key features of practical importance. Just as the $MA(1)$ process was covariance stationary for any value of its parameters, so too is the finite-order $MA(q)$ process. As with the $MA(1)$ process, the $MA(q)$ process is *invertible* only if a root condition is satisfied. The $MA(q)$ lag operator polynomial has q roots; when $q > 1$ the possibility of complex roots arises. The condition for invertibility of the $MA(q)$ process is that the inverses of all of the roots must be inside the unit circle, in which case we have the convergent autoregressive representation,

$$\frac{1}{\Theta(L)} y_t = \varepsilon_t.$$

The conditional mean of the $MA(q)$ process evolves with the information

set, in contrast to the unconditional moments, which are fixed. In contrast to the $MA(1)$ case, in which the conditional mean depends on only the first lag of the innovation, in the $MA(q)$ case the conditional mean depends on q lags of the innovation. Thus the $MA(q)$ process has the potential for longer memory.

The potentially longer memory of the $MA(q)$ process emerges clearly in its autocorrelation function. In the $MA(1)$ case, all autocorrelations beyond displacement 1 are zero; in the $MA(q)$ case all autocorrelations beyond displacement q are zero. This autocorrelation cutoff is a distinctive property of moving average processes. The partial autocorrelation function of the $MA(q)$ process, in contrast, decays gradually, in accord with the infinite autoregressive representation, in either an oscillating or one-sided fashion, depending on the parameters of the process.

In closing this section, let's step back for a moment and consider in greater detail the precise way in which finite-order moving average processes approximate the Wold representation. The Wold representation is

$$y_t = B(L)\varepsilon_t,$$

where $B(L)$ is of infinite order. The $MA(1)$, in contrast, is simply a first-order moving average, in which a series is expressed as a one-period moving average of current and past innovations. Thus when we fit an $MA(1)$ model we're using the first-order polynomial $1 + \theta L$ to approximate the infinite-order polynomial $B(L)$. Note that $1 + \theta L$ is a rational polynomial with numerator polynomial of degree one and degenerate denominator polynomial (degree zero).

$MA(q)$ process have the potential to deliver better approximations to the Wold representation, at the cost of more parameters to be estimated. The Wold representation involves an infinite moving average; the $MA(q)$ process

approximates the infinite moving average with a *finite-order* moving average,

$$y_t = \Theta(L)\varepsilon_t,$$

whereas the $MA(1)$ process approximates the infinite moving average with a only a *first-order* moving average, which can sometimes be very restrictive.

Soon we shall see that MA processes are absolutely central for understanding forecasting and properties of forecast errors, even if they usually not used directly as forecasting models. Other approximations to the Wold representation are typically more useful for producing forecasts, in particular autoregressive (AR) and mixed autoregressive moving-average ($ARMA$) models, to which we now turn.

7.2.3 Autoregressive (AR) Models

The autoregressive process is also a natural approximation to the Wold representation. We've seen, in fact, that under certain conditions a moving average process has an autoregressive representation, so an autoregressive process is in a sense the same as a moving average process. Like the moving average process, the autoregressive process has direct motivation; it's simply a stochastic difference equation, a simple mathematical model in which the current value of a series is linearly related to its past values, plus an additive stochastic shock. Stochastic difference equations are a natural vehicle for discrete-time stochastic dynamic modeling.

The $AR(1)$ Process

The first-order autoregressive process, $AR(1)$ for short, is

$$y_t = \phi y_{t-1} + \varepsilon_t$$

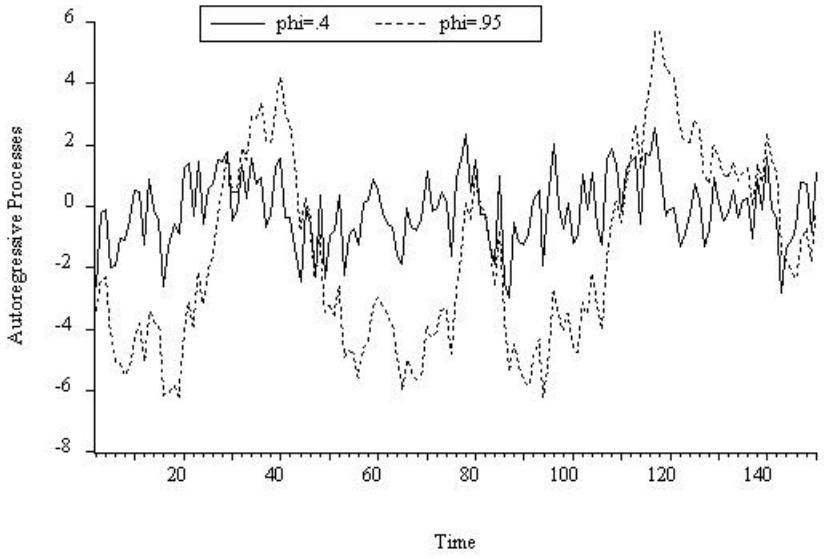


Figure 7.6: Realization of Two AR(1) Processes

$$\varepsilon_t \sim WN(0,\ \sigma^2).$$

In lag operator form, we write

$$(1 - \phi L)y_t = \varepsilon_{t.}$$

In Figure 7.6 we show simulated realizations of length 150 of two AR(1) processes; the first is

$$y_t = .4y_{t-1} + \varepsilon_t,$$

and the second is

$$y_t = .95y_{t-1} + \varepsilon_t,$$

where in each case

$$\varepsilon_t \ iid\ N(0, 1),$$

and the same innovation sequence underlies each realization.

The fluctuations in the $AR(1)$ with parameter $\phi = .95$ appear much more persistent that those of the $AR(1)$ with parameter $\phi = .4$. This contrasts sharply with the $MA(1)$ process, which has a very short memory regardless

of parameter value. Thus the $AR(1)$ model is capable of capturing much more persistent dynamics than is the $MA(1)$.

Recall that a finite-order moving average process is always covariance stationary, but that certain conditions must be satisfied for invertibility, in which case an autoregressive representation exists. For autoregressive processes, the situation is precisely the reverse. Autoregressive processes are always invertible – in fact invertibility isn't even an issue, as finite-order autoregressive processes *already are* in autoregressive form – but certain conditions must be satisfied for an autoregressive process to be covariance stationary.

If we begin with the $AR(1)$ process,

$$y_t = \phi y_{t-1} + \varepsilon_t,$$

and substitute backward for lagged y's on the right side, we obtain

$$y_t = \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + ...$$

In lag operator form we write

$$y_t = \frac{1}{1-\phi L}\varepsilon_t.$$

This moving average representation for y is convergent if and only if $|\phi| < 1$; thus, $|\phi| < 1$ is the condition for covariance stationarity in the $AR(1)$ case. Equivalently, the condition for covariance stationarity is that the inverse of the root of the autoregressive lag operator polynomial be less than one in absolute value.

From the moving average representation of the covariance stationary $AR(1)$

process, we can compute the unconditional mean and variance,

$$E(y_t) = E(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + ...)$$

$$= E(\varepsilon_t) + \phi E(\varepsilon_{t-1}) + \phi^2 E(\varepsilon_{t-2}) + ...$$

$$= 0$$

and

$$var(y_t) \;=\; var(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + ...)$$

$$= \sigma^2 + \phi^2\sigma^2 + \phi^4\sigma^2 + ...$$

$$= \sigma^2 \textstyle\sum_{i=0}^{\infty} \phi^{2i}$$

$$= \frac{\sigma^2}{1-\phi^2}.$$

The conditional moments, in contrast, are

$$E(y_t|y_{t-1}) = E(\phi y_{t-1} + \varepsilon_t|y_{t-1})$$

$$= \phi E(y_{t-1}|y_{t-1}) + E(\varepsilon_t|y_{t-1})$$

$$= \phi y_{t-1} + 0$$

$$= \phi y_{t-1}$$

and

$$var(y_t|y_{t-1}) = var((\phi y_{t-1} + \varepsilon_t) \mid y_{t-1})$$

$$= \phi^2 var(y_{t-1}|y_{t-1}) + var(\varepsilon_t|y_{t-1})$$

$$= 0 + \sigma^2$$

$$= \sigma^2.$$

Note in particular that the simple way that the conditional mean adapts to the changing information set as the process evolves.

To find the autocovariances, we proceed as follows. The process is

$$y_t = \phi y_{t-1} + \varepsilon_t,$$

so that multiplying both sides of the equation by $y_{t-\tau}$ we obtain

$$y_t y_{t-\tau} = \phi y_{t-1} y_{t-\tau} + \varepsilon_t y_{t-\tau}.$$

For $\tau \geq 1$, taking expectations of both sides gives

$$\gamma(\tau) = \phi\gamma(\tau - 1).$$

This is called the **Yule-Walker equation**. It is a recursive equation; that is, given $\gamma(\tau)$, for any τ, the Yule-Walker equation immediately tells us how to get $\gamma(\tau + 1)$. If we knew $\gamma(0)$ to start things off (an "initial condition"), we could use the Yule-Walker equation to determine the entire autocovariance sequence. And we *do* know $\gamma(0)$; it's just the variance of the process, which we already showed to be

$$\gamma(0) = \frac{\sigma^2}{1 - \phi^2}.$$

Thus we have

$$\gamma(0) = \frac{\sigma^2}{1-\phi^2}$$

$$\gamma(1) = \phi\frac{\sigma^2}{1-\phi^2}$$

$$\gamma(2) = \phi^2\frac{\sigma^2}{1-\phi^2},$$

and so on. In general, then,

$$\gamma(\tau) = \phi^\tau\frac{\sigma^2}{1-\phi^2},\ \tau = 0,\ 1,\ 2,\ ....$$

Dividing through by $\gamma(0)$ gives the autocorrelations,

$$\rho(\tau) = \phi^\tau,\ \tau = 0,\ 1,\ 2,\ ....$$

Note the gradual autocorrelation decay, which is typical of autoregressive processes. The autocorrelations approach zero, but only in the limit as the displacement approaches infinity. In particular, they don't cut off to zero, as is the case for moving average processes. If ϕ is positive, the autocorrelation decay is one-sided. If ϕ is negative, the decay involves back-and-forth oscillations. The relevant case in business and economics is $\phi > 0$, but either way, the autocorrelations damp gradually, not abruptly. In Figure 7.7 and 7.8 we show the autocorrelation functions for $AR(1)$ processes with parameters $\phi = .4$ and $\phi = .95$. The persistence is much stronger when $\phi = .95$, in contrast to the $MA(1)$ case, in which the persistence was weak regardless of the parameter.

Finally, the partial autocorrelation function for the $AR(1)$ process cuts off abruptly; specifically,

$$p(\tau) = \begin{cases} \phi,\ \tau = 1 \\ \\ 0,\ \tau > 1. \end{cases}.$$

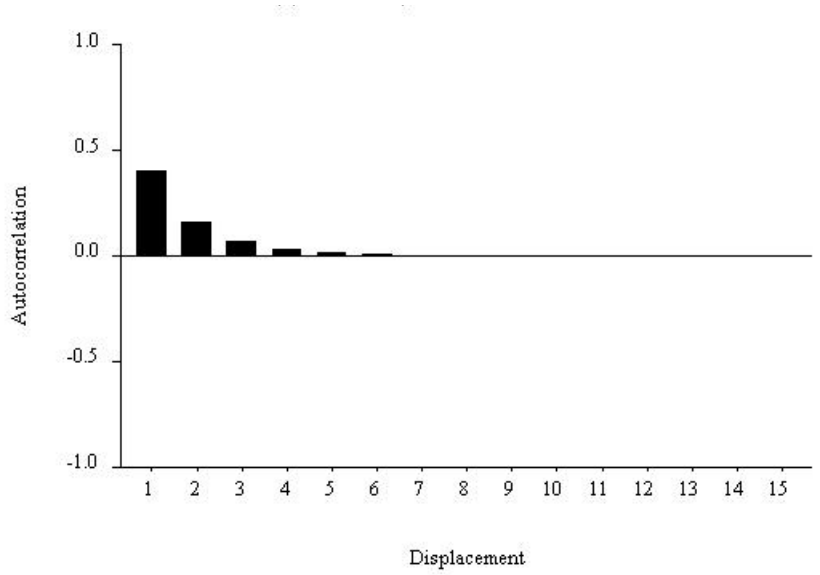


Figure 7.7: AR(1) Population Autocorrelation Function - $\rho = .4$

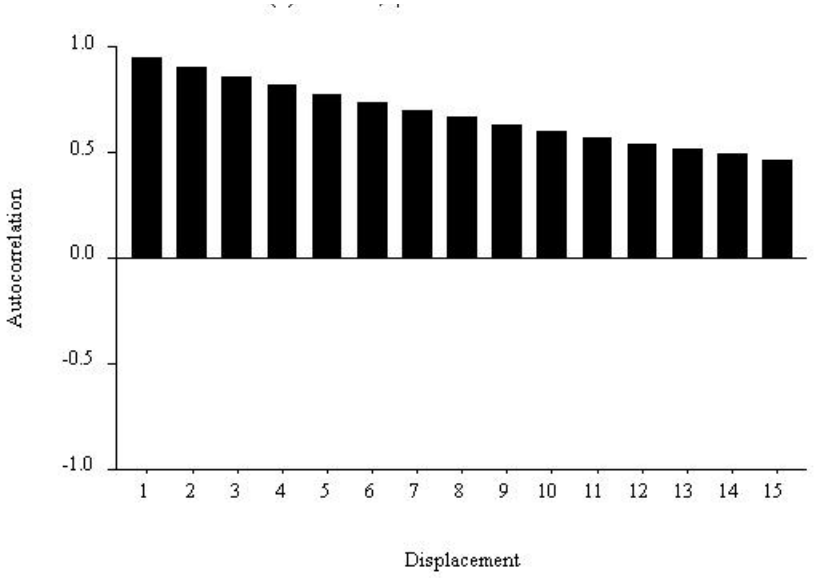


Figure 7.8: AR1) Population Autocorrelation Function - $\rho = .95$

It's easy to see why. The partial autocorrelations are just the last coefficients in a sequence of successively longer population autoregressions. If the true process is in fact an $AR(1)$, the first partial autocorrelation is just the autoregressive coefficient, and coefficients on all longer lags are zero.

In Figures 7.9 and 7.10 we show the partial autocorrelation functions for

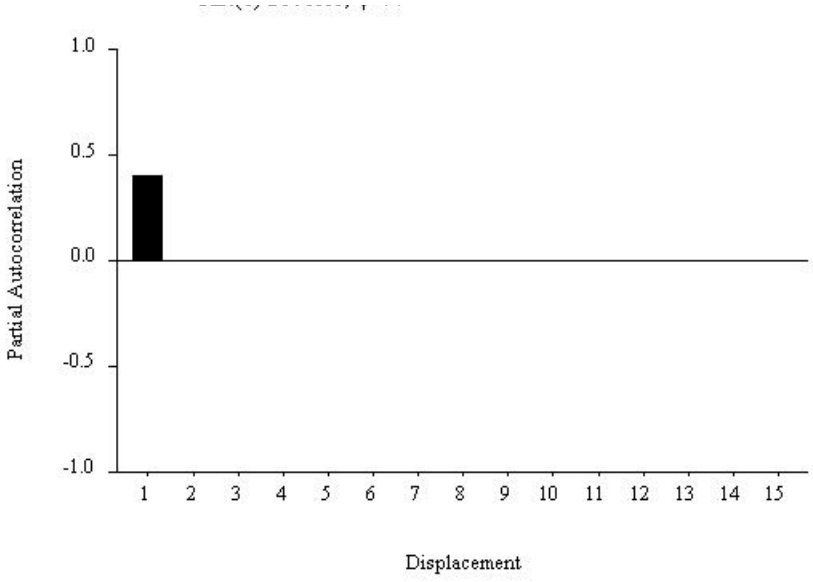


Figure 7.9: AR(1) Population Partial Autocorrelation Function - $\rho = .4$

our two $AR(1)$ processes. At displacement 1, the partial autocorrelations are simply the parameters of the process (.4 and .95, respectively), and at longer displacements, the partial autocorrelations are zero.

The $AR(p)$ Process

The general p-th order autoregressive process, or $AR(p)$ for short, is

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t$$

$$\varepsilon_t \ \sim \ WN(0, \ \sigma^2).$$

In lag operator form we write

$$\Phi(L) y_t = (1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p) \ y_t = \varepsilon_{t.}$$

As with our discussion of the $MA(q)$ process, in our discussion of the $AR(p)$ process we dispense here with mathematical derivations and instead rely on parallels with the $AR(1)$ case to establish intuition for its key properties.

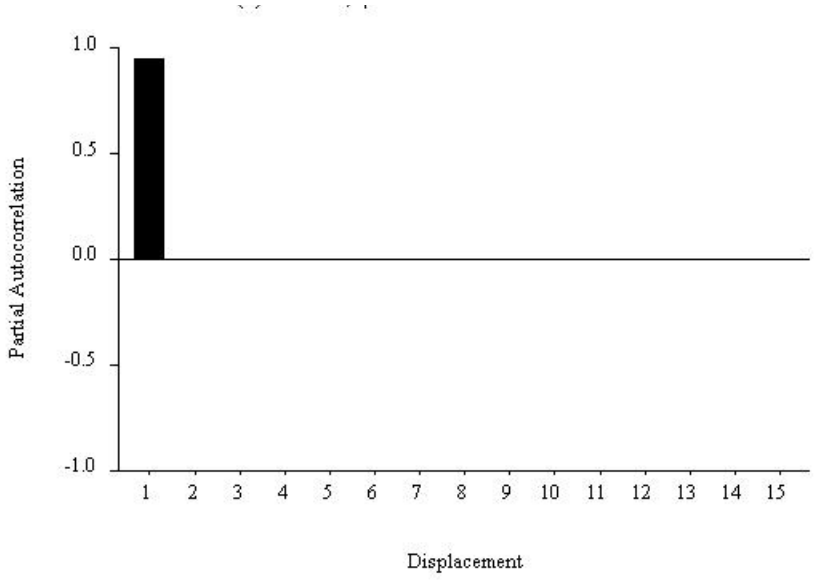


Figure 7.10: AR1) Population Partial Autocorrelation Function - $\rho = .95$

An $AR(p)$ process is covariance stationary if and only if the inverses of all roots of the autoregressive lag operator polynomial $\Phi(L)$ are inside the unit circle.[7] In the covariance stationary case we can write the process in the convergent infinite moving average form

$$y_t = \frac{1}{\Phi(L)}\varepsilon_t.$$

The autocorrelation function for the general $AR(p)$ process, as with that of the $AR(1)$ process, decays gradually with displacement. Finally, the $AR(p)$ partial autocorrelation function has a sharp cutoff at displacement p, for the same reason that the $AR(1)$ partial autocorrelation function has a sharp cutoff at displacement 1.

Let's discuss the $AR(p)$ autocorrelation function in a bit greater depth.

[7] A necessary condition for covariance stationarity, which is often useful as a quick check, is

$$\sum_{i=1}^{p} \phi_i < 1.$$

If the condition is satisfied, the process may or may not be stationary, but if the condition is violated, the process can't be stationary.

The key insight is that, in spite of the fact that its qualitative behavior (gradual damping) matches that of the $AR(1)$ autocorrelation function, it can nevertheless display a richer variety of patterns, depending on the order and parameters of the process. It can, for example, have damped monotonic decay, as in the $AR(1)$ case with a positive coefficient, but it can also have damped oscillation in ways that $AR(1)$ can't have. In the $AR(1)$ case, the only possible oscillation occurs when the coefficient is negative, in which case the autocorrelations switch signs at each successively longer displacement. In higher-order autoregressive models, however, the autocorrelations can oscillate with much richer patterns reminiscent of cycles in the more traditional sense. This occurs when some roots of the autoregressive lag operator polynomial are complex.[8]

Consider, for example, the $AR(2)$ process,

$$y_t = 1.5y_{t-1} - .9y_{t-2} + \varepsilon_t.$$

The corresponding lag operator polynomial is $1 - 1.5L + .9L^2$, with two complex conjugate roots, $.83 \pm .65i$. The inverse roots are $.75 \pm .58i$, both of which are close to, but inside, the unit circle; thus the process is covariance stationary. It can be shown that the autocorrelation function for an $AR(2)$ process is

$$\rho(0) = 1$$

$$\rho(\tau) = \phi_1\rho(\tau - 1) + \phi_2\rho(\tau - 2), \ \tau = 2, \ 3, \ ...$$

$$\rho(1) = \frac{\phi_1}{1 - \phi_2}$$

Using this formula, we can evaluate the autocorrelation function for the process at hand; we plot it in Figure 7.11. Because the roots are complex, the autocorrelation function oscillates, and because the roots are close to the unit circle, the oscillation damps slowly.

[8]Note that complex roots can't occur in the $AR(1)$ case.

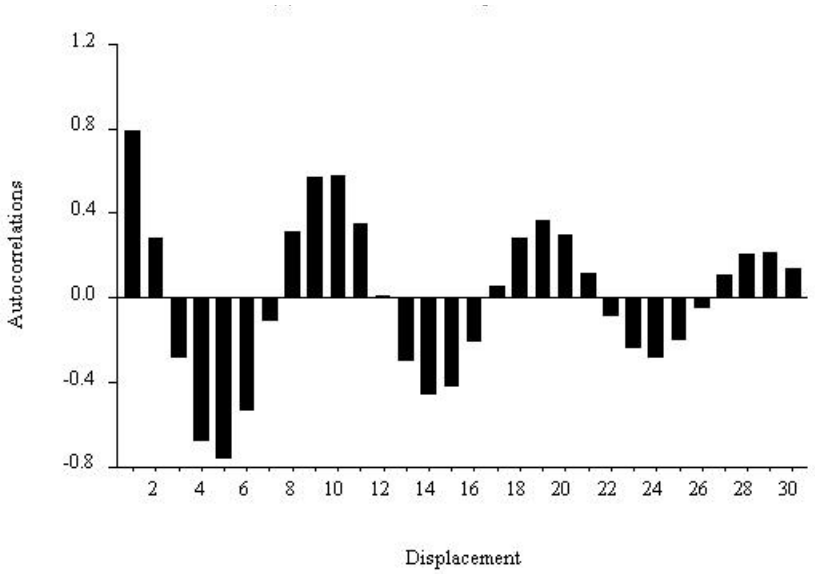


Figure 7.11: Population Autocorrelation Function - AR(2) with Complex Roots

Finally, let's step back once again to consider in greater detail the precise way that finite-order autoregressive processes approximate the Wold representation. As always, the Wold representation is $y_t = B(L)\varepsilon_t$, where $B(L)$ is of infinite order. The $AR(1)$, as compared to the $MA(1)$, is simply a different approximation to the Wold representation. The moving average representation associated with the $AR(1)$ process is $y_t = 1/1 - \phi L\varepsilon_t$. Thus, when we fit an $AR(1)$ model, we're using $1/1 - \phi L$, a rational polynomial with degenerate numerator polynomial (degree zero) and denominator polynomial of degree one, to approximate $B(L)$. The moving average representation associated with the $AR(1)$ process is of infinite order, as is the Wold representation, but it does not have infinitely many free coefficients. In fact, only one parameter, ϕ, underlies it.

The $AR(p)$ is an obvious generalization of the $AR(1)$ strategy for approximating the Wold representation. The moving average representation associated with the $AR(p)$ process is $y_t = 1/\Phi(L)\varepsilon_t$. When we fit an $AR(p)$ model to approximate the Wold representation we're still using a rational polynomial with degenerate numerator polynomial (degree zero), but the de-

nominator polynomial is of higher degree.

7.2.4 Autoregressive Moving Average ($ARMA$) Models

Autoregressive and moving average models are often combined in attempts to obtain better and more parsimonious approximations to the Wold representation, yielding the autoregressive moving average process, **ARMA(p,q)** for short. As with moving average and autoregressive processes, ARMA processes also have direct motivation.[9] First, if the random shock that drives an autoregressive process is itself a moving average process, then it can be shown that we obtain an ARMA process. Second, ARMA processes can arise from aggregation. For example, sums of AR processes, or sums of AR and MA processes, can be shown to be ARMA processes. Finally, AR processes observed subject to measurement error also turn out to be ARMA processes.

The simplest ARMA process that's not a pure autoregression or pure moving average is the ARMA(1,1), given by

$$y_t = \phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

or in lag operator form,

$$(1 - \phi L)\, y_t = (1 + \theta L)\, \varepsilon_t,$$

where $|\phi| < 1$ is required for stationarity and $|\theta| < 1$ is required for invertibility.[10] If the covariance stationarity condition is satisfied, then we have the moving average representation

$$y_t = \frac{(1 + \theta L)}{(1 - \phi L)}\, \varepsilon_t,$$

[9] For more extensive discussion, see Granger and Newbold (1986).

[10] Both stationarity and invertibility need to be checked in the ARMA case, because both autoregressive and moving average components are present.

which is an infinite distributed lag of current and past innovations. Similarly, if the invertibility condition is satisfied, then we have the infinite autoregressive representation,

$$\frac{(1-\phi L)}{(1+\theta L)}\, y_t = \varepsilon_{t.}$$

The ARMA(p,q) process is a natural generalization of the ARMA(1,1) that allows for multiple moving average and autoregressive lags. We write

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q}$$

$$\varepsilon_t \ \sim \ WN(0,\ \sigma^2),$$

or

$$\Phi(L) y_t = \Theta(L) \varepsilon_t,$$

where

$$\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p$$

and

$$\Theta(L) = 1 + \theta_1 L + \theta_2 L^2 + ... \ + \ \theta_q L^q.$$

If the inverses of all roots of Φ(L) are inside the unit circle, then the process is covariance stationary and has convergent infinite moving average representation

$$y_t = \frac{\Theta(L)}{\Phi(L)} \varepsilon_t.$$

If the inverses of all roots of Θ(L) are inside the unit circle, then the process is invertible and has convergent infinite autoregressive representation

$$\frac{\Phi(L)}{\Theta(L)} y_t = \varepsilon_{t.}$$

As with autoregressions and moving averages, ARMA processes have a fixed unconditional mean but a time-varying conditional mean. In contrast to pure moving average or pure autoregressive processes, however, neither the

autocorrelation nor partial autocorrelation functions of ARMA processes cut off at any particular displacement. Instead, each damps gradually, with the precise pattern depending on the process.

ARMA models approximate the Wold representation by a ratio of two finite-order lag-operator polynomials, neither of which is degenerate. Thus ARMA models use ratios of full-fledged polynomials in the lag operator to approximate the Wold representation,

$$y_t = \frac{\Theta(L)}{\Phi(L)} \varepsilon_{t.}$$

ARMA models, by allowing for both moving average and autoregressive components, often provide accurate approximations to the Wold representation that nevertheless have just a few parameters. That is, ARMA models are often both highly accurate and highly parsimonious. In a particular situation, for example, it might take an AR(5) to get the same approximation accuracy as could be obtained with an ARMA(2,1), but the AR(5) has five parameters to be estimated, whereas the ARMA(2,1) has only three.

7.3 Forecasting Cycles From a Moving-Average Perspective: Wiener-Kolmogorov

By now you've gotten comfortable with the idea of an **information set**. Here we'll use that idea extensively. We denote the time-T information set by Ω_T. As first pass it seems most natural to think of the information set as containing the available past history of the series,

$$\Omega_T = \{y_T,\ y_{T-1},\ y_{T-2},\ ...\},$$

where for theoretical purposes we imagine history as having begun in the infinite past.

So long as y is covariance stationary, however, we can just as easily express the information available at time T in terms of current and past shocks,

$$\Omega_T = \{\varepsilon_T,\ \varepsilon_{T-1},\ \varepsilon_{T-2},\ ...\}.$$

Suppose, for example, that the process to be forecast is a covariance stationary $AR(1)$,

$$y_t = \phi y_{t-1} + \varepsilon_{t.}$$

Then immediately,

$$\varepsilon_T = y_T - \phi y_{T-1}$$

$$\varepsilon_{T-1} = y_{T-1} - \phi y_{T-2}$$

$$\varepsilon_{T-2} = y_{T-2} - \phi y_{T-3},$$

and so on. In other words, we can figure out the current and lagged ε's from the current and lagged y's. More generally, for any covariance stationary and invertible series, we can infer the history of ε from the history of y, and the history of y from the history of ε.

Assembling the discussion thus far, we can view the time-T information set as containing the current and past values of either (or both) y and ε,

$$\Omega_T = y_T,\ y_{T-1},\ y_{T-2},\ ...,\ \varepsilon_T,\ \varepsilon_{T-1},\ \varepsilon_{T-2},\ ....$$

Based upon that information set, we want to find the **optimal forecast** of y at some future time $T + h$. The optimal forecast is the one with the smallest loss on average, that is, the forecast that minimizes **expected loss**. It turns out that under reasonably weak conditions the optimal forecast is the **conditional mean**,

$$E(y_{T+h}|\Omega_T),$$

the expected value of the future value of the series being forecast, conditional upon available information.

In general, the conditional mean need not be a linear function of the elements of the information set. Because linear functions are particularly tractable, we prefer to work with linear forecasts – forecasts that are linear in the elements of the information set – by finding the best linear approximation to the conditional mean, called the **linear projection**, denoted

$$P(y_{T+h}|\Omega_T).$$

This explains the common term "**linear least squares forecast**." The linear projection is often very useful and accurate, because the conditional mean is often close to linear. In fact, in the Gaussian case the conditional expectation is exactly linear, so that

$$E(y_{T+h}|\Omega_T) = P(y_{T+h}|\Omega_T).$$

7.3.1 Optimal Point Forecasts for Finite-Order Moving Averages

Our forecasting method is always the same: we write out the process for the future time period of interest, $T + h$, and project it on what's known at time T, when the forecast is made. This process is best learned by example. Consider an $MA(2)$ process,

$$y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}$$

$$\varepsilon_t \ \sim \ WN(0, \ \sigma^2).$$

Suppose we're standing at time T and we want to forecast y_{T+1}. First we write out the process for $T + 1$,

$$y_{T+1} = \varepsilon_{T+1} + \theta_1\varepsilon_T + \theta_2\varepsilon_{T-1}.$$

Then we project on the time-T information set, which simply means that all future innovations are replaced by zeros. Thus

$$y_{T+1,T} = P(y_{T+1}|\Omega_T) = \theta_1\varepsilon_T + \theta_2\varepsilon_{T-1}.$$

To forecast 2 steps ahead, we note that

$$y_{T+2} = \varepsilon_{T+2} + \theta_1\varepsilon_{T+1} + \theta_2\varepsilon_T,$$

and we project on the time-T information set to get

$$y_{T+2,T} = \theta_2\varepsilon_T.$$

Continuing in this fashion, we see that

$$y_{T+h,T} = 0,$$

for all h>2.

Now let's compute the corresponding **forecast errors**.[11] We have:

$$e_{T+1,T} = \varepsilon_{T+1}\ WN$$

$$e_{T+2,T} = \varepsilon_{T+2} + \theta_1\varepsilon_{T+1}\ (MA(1))$$

$$e_{T+h,T} = \varepsilon_{T+h} + \theta_1\varepsilon_{T+h-1} + \theta_2\varepsilon_{T+h-2}\ (MA(2)),$$

for all h>2.

Finally, the **forecast error variances** are:

$$\sigma_1^2 = \sigma^2$$

$$\sigma_2^2 = \sigma^2(1+\theta_1^2)$$

$$\sigma_h^2 = \sigma^2(1+\theta_1^2+\theta_2^2),$$

[11] Recall that the forecast error is simply the difference between the actual and forecasted values. That is, $e_{T+h,T} = y_{T+h} - y_{T+h,T}$.

for all $h > 2$. Moreover, the forecast error variance for h>2 is just the unconditional variance of y_t.

Now consider the general $MA(q)$ case. The model is

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... \; \theta_q \varepsilon_{t-q}.$$

First, consider the forecasts. If $h \leq q$, the forecast has the form

$$y_{T+h,T} = 0 + \text{“}adjustment,\text{”}$$

whereas if $h > q$ the forecast is

$$y_{T+h,T} = 0.$$

Thus, an $MA(q)$ process is not forecastable (apart from the unconditional mean) more than q steps ahead. All the dynamics in the $MA(q)$ process, which we exploit for forecasting, “wash out” by the time we get to horizon q, which reflects the autocorrelation structure of the $MA(q)$ process. (Recall that, as we showed earlier, it cuts off at displacement q.) Second, consider the corresponding forecast errors. They are

$$e_{T+h,T} = MA(h-1)$$

for $h \leq q$ and

$$e_{T+h,T} = MA(q)$$

for $h > q$. The h-step-ahead forecast error for $h > q$ is just the process itself, minus its mean.

Finally, consider the forecast error variances. For $h \leq q$,

$$\sigma_h^2 \leq var(y_t),$$

whereas for $h > q$,

$$\sigma_h^2 = var(y_t).$$

In summary, we've thus far studied the $MA(2)$, and then the general $MA(q)$, process, computing the optimal h-step-ahead forecast, the corresponding forecast error, and the forecast error variance. As we'll now see, the emerging patterns that we cataloged turn out to be quite general.

7.3.2 Optimal Point Forecasts for Infinite-Order Moving Averages

By now you're getting the hang of it, so let's consider the general case of an infinite-order MA process. The infinite-order moving average process may seem like a theoretical curiosity, but precisely the opposite is true. Any covariance stationary process can be written as a (potentially infinite-order) moving average process, and moving average processes are easy to understand and manipulate, because they are written in terms of white noise shocks, which have very simple statistical properties. Thus, if you take the time to understand the mechanics of constructing optimal forecasts for infinite moving-average processes, you'll understand everything, and you'll have some powerful technical tools and intuition at your command.

Recall that the general linear process is

$$y_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i},$$

where

$$\varepsilon_t \sim WN(0, \sigma^2)$$

$$b_0 = 1$$

$$\sigma^2 \sum_{i=0}^{\infty} b_i^2 < \infty.$$

We proceed in the usual way. We first write out the process at the future

time of interest:

$$y_{T+h} = \varepsilon_{T+h} + b_1\varepsilon_{T+h-1} + ... + b_h\varepsilon_T + b_{h+1}\varepsilon_{T-1} + ...$$

Then we project y_{T+h} on the time-T information set. The projection yields zeros for all of the future ε's (because they are white noise and hence unforecastable), leaving

$$y_{T+h,T} = b_h\varepsilon_T + b_{h+1}\varepsilon_{T-1} + ...$$

It follows that the h-step ahead forecast error is serially correlated; it follows an $MA(h-1)$ process,

$$e_{T+h,T} = (y_{T+h} - y_{T+h,T}) = \sum_{i=0}^{h-1} b_i\varepsilon_{T+h-i},$$

with mean 0 and variance

$$\sigma_h^2 = \sigma^2 \sum_{i=0}^{h-1} b_i^2.$$

A number of remarks are in order concerning the optimal forecasts of the general linear process, and the corresponding forecast errors and forecast error variances. First, the 1-step-ahead forecast error is simply ε_{T+1}. ε_{T+1} is that part of y_{T+1} that can't be linearly forecast on the basis of Ω_t (which, again, is why it is called the innovation). Second, although it might at first seem strange that an *optimal* forecast error would be serially correlated, as is the case when $h > 1$, nothing is awry. The serial correlation can't be used to improve forecasting performance, because the autocorrelations of the $MA(h-1)$ process cut off just before the beginning of the time-T information set ε_T, ε_{T-1}, ... This is a general and tremendously important property of the errors associated with optimal forecasts: *errors from optimal forecasts can't be forecast using information available when the forecast was made.* If you can forecast the forecast error, then you can improve the forecast, which means that it couldn't have been optimal. Finally, note that as h approaches

infinity $y_{T+h,T}$ approaches zero, the unconditional mean of the process, and σ_h^2 approaches $\sigma^2 \sum_{i=0}^{\infty} b_i^2$, the unconditional variance of the process, which reflects the fact that as h approaches infinity the conditioning information on which the forecast is based becomes progressively less useful. In other words, the distant future is harder to forecast than the near future!

7.3.3 Interval and Density Forecasts

Now we construct interval and density forecasts. Regardless of whether the moving average is finite or infinite, we proceed in the same way, as follows. The definition of the h-step-ahead forecast error is

$$e_{T+h,T} = y_{T+h} - y_{T+h,T}.$$

Equivalently, the h-step-ahead realized value, y_{T+h} , equals the forecast plus the error,

$$y_{T+h} = y_{T+h,T} + e_{T+h,T}.$$

If the innovations are normally distributed, then the future value of the series of interest is also normally distributed, conditional upon the information set available at the time the forecast was made, and so we have the 95% h-step-ahead interval forecast $y_{T+h,T} \pm 1.96\sigma_h$.[12] In similar fashion, we construct the h-step-ahead density forecast as

$$N(y_{T+h,T},\ \sigma_h^2).$$

The mean of the conditional distribution of y_{T+h} is $y_{T+h,T}$, which of course must be the case because we constructed the point forecast as the conditional mean, and the variance of the conditional distribution is σ_h^2, the variance of

[12] Confidence intervals at any other desired confidence level may be constructed in similar fashion, by using a different critical point of the standard normal distribution. A 90% interval forecast, for example, is $y_{T+h,T} \pm 1.64\sigma_h$. In general, for a Gaussian process, a $(1-\alpha)\cdot 100\%$ confidence interval is $y_{T+h,T} \pm z_{\alpha/2}\sigma_h$, where $z_{\alpha/2}$ is that point on the $N(0,1)$ distribution such that $prob(z > z_{\alpha/2}) = \alpha/2$.

the forecast error.

As an example of interval and density forecasting, consider again the $MA(2)$ process,

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

Assuming normality, the 1-step-ahead 95% interval forecast is

$$y_{T+1,T} = (\theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1}) \pm 1.96\sigma,$$

and the 1-step-ahead density forecast is

$$N(\theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1}, \sigma^2).$$

7.3.4 Making the Forecasts Operational

So far we've assumed that the parameters of the process being forecast are known. In practice, of course, they must be estimated. To make our forecasting procedures operational, we simply replace the unknown parameters in our formulas with estimates, and the unobservable innovations with residuals.

Consider, for example, the $MA(2)$ process,

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}.$$

As you can readily verify using the methods we've introduced, the 2-step ahead optimal forecast, assuming known parameters, is

$$y_{T+2,T} = \theta_2 \varepsilon_T,$$

with corresponding forecast error

$$e_{T+2,T} = \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1},$$

and forecast-error variance

$$\sigma_2^2 = \sigma^2(1 + \theta_1^2).$$

To make the forecast operational, we replace unknown parameters with estimates and the time-T innovation with the time-T residual, yielding

$$\hat{y}_{T+2,T} = \hat{\theta}_2 \hat{\varepsilon}_T$$

and forecast error variance

$$\hat{\sigma}_2^2 = \hat{\sigma}^2(1 + \hat{\theta}_1^2).$$

Then, if desired, we can construct operational 2-step-ahead interval and density forecasts, as

$$\hat{y}_{T+2,T} \pm z_{\alpha/2}\hat{\sigma}_2$$

and

$$N(\hat{y}_{T+2,T},\ \hat{\sigma}_2^2).$$

The strategy of taking a forecast formula derived under the assumption of known parameters, and replacing unknown parameters with estimates, is a natural way to operationalize the construction of point forecasts. However, using the same strategy to produce operational interval or density forecasts involves a subtlety that merits additional discussion. The forecast error variance estimate so obtained can be interpreted as one that ignores parameter estimation uncertainty, as follows. Recall once again that the actual future value of the series is

$$y_{T+2} = \varepsilon_{T+2} + \theta_1\varepsilon_{T+1} + \theta_2\varepsilon_T,$$

and that the operational forecast is

$$\hat{y}_{T+2,T} = \hat{\theta}_2 \varepsilon_T.$$

Thus the exact forecast error is

$$\hat{e}_{T+2,T} = y_{T+2} - \hat{y}_{T+2,T} = \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1} + (\theta_2 - \hat{\theta}_2)\varepsilon_T,$$

the variance of which is very difficult to evaluate. So we make a convenient approximation: we ignore parameter estimation uncertainty by assuming that estimated parameters equal true parameters. We therefore set

$$(\theta_2 - \hat{\theta}_2)$$

to zero, which yields

$$\hat{e}_{T+2,T} = \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1},$$

with variance

$$\sigma_2^2 = \sigma^2(1 + \theta_1^2),$$

which we make operational as

$$\hat{\sigma}_2^2 = \hat{\sigma}^2(1 + \hat{\theta}_1^2).$$

7.4 Forecasting Cycles From an Autoregressive Perspective: Wold's Chain Rule

7.4.1 Point Forecasts of Autoregressive Processes

Because any covariance stationary $AR(p)$ process can be written as an infinite moving average, there's no need for specialized forecasting techniques for autoregressions. Instead, we can simply transform the autoregression into a moving average, and then use the techniques we developed for forecasting

moving averages. It turns out, however, that a very simple recursive method for computing the optimal forecast is available in the autoregressive case.

The recursive method, called the **chain rule of forecasting**, is best learned by example. Consider the $AR(1)$ process,

$$y_t = \phi y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

First we construct the optimal 1-step-ahead forecast, and then we construct the optimal 2-step-ahead forecast, which depends on the optimal 1-step-ahead forecast, which we've already constructed. Then we construct the optimal 3-step-ahead forecast, which depends on the already-computed 2-step-ahead forecast, which we've already constructed, and so on.

To construct the 1-step-ahead forecast, we write out the process for time $T+1$,

$$y_{T+1} = \phi y_T + \varepsilon_{T+1}.$$

Then, projecting the right-hand side on the time-T information set, we obtain

$$y_{T+1,T} = \phi y_T.$$

Now let's construct the 2-step-ahead forecast. Write out the process for time $T+2$,

$$y_{T+2} = \phi y_{T+1} + \varepsilon_{T+2}.$$

Then project directly on the time-T information set to get

$$y_{T+2,T} = \phi y_{T+1,T}.$$

Note that the future innovation is replaced by 0, as always, and that we have directly replaced the time $T+1$ value of y with its earlier-constructed optimal forecast. Now let's construct the 3-step-ahead forecast. Write out the process

for time $T+3$,

$$y_{T+3} = \phi y_{T+2} + \varepsilon_{T+3}.$$

Then project directly on the time-T information set,

$$y_{T+3,T} = \phi y_{T+2,T}.$$

The required 2-step-ahead forecast was already constructed.

Continuing in this way, we can recursively build up forecasts for any and all future periods. Hence the name "chain rule of forecasting." Note that, for the $AR(1)$ process, only the most recent value of y is needed to construct optimal forecasts, for any horizon, and for the general $AR(p)$ process only the p most recent values of y are needed.

7.4.2 Point Forecasts of ARMA processes

Now we consider forecasting covariance stationary ARMA processes. Just as with autoregressive processes, we could always convert an ARMA process to an infinite moving average, and then use our earlier-developed methods for forecasting moving averages. But also as with autoregressive processes, a simpler method is available for forecasting ARMA processes directly, by combining our earlier approaches to moving average and autoregressive forecasting.

As always, we write out the $ARMA(p,q)$ process for the future period of interest,

$$y_{T+h} = \phi_1 y_{T+h-1} + ... + \phi_p y_{T+h-p} + \varepsilon_{T+h} + \theta_1 \varepsilon_{T+h-1} + ... + \theta_q \varepsilon_{T+h-q}.$$

On the right side we have various future values of y and ε, and perhaps also past values, depending on the forecast horizon. We replace everything on the right-hand side with its projection on the time-T information set. That is, we replace all future values of y with optimal forecasts (built up recursively

using the chain rule) and all future values of ε with optimal forecasts (0), yielding

$$y_{T+h,T} = \phi_1 y_{T+h-1,T} + ... + \phi_p y_{T+h-p,T} + \varepsilon_{T+h,T} + \theta_1 \varepsilon_{T+h-1,T} + ... + \theta_q \varepsilon_{T+h-q,T}.$$

When evaluating this formula, note that the optimal time-T "forecast" of any value of y or ε dated time T or earlier is just y or ε itself.

As an example, consider forecasting the $ARMA(1,1)$ process,

$$y_t = \phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$

$$\varepsilon_t \sim WN(0,\ \sigma^2).$$

Let's find $y_{T+1,T}$. The process at time $T+1$ is

$$y_{T+1} = \phi y_T + \varepsilon_{T+1} + \theta \varepsilon_T.$$

Projecting the right-hand side on Ω_T yields

$$y_{T+1,T} = \phi y_T + \theta \varepsilon_T.$$

Now let's find $y_{T+2,T}$. The process at time $T+2$ is

$$y_{T+2} = \phi y_{T+1} + \varepsilon_{T+2} + \theta \varepsilon_{T+1}.$$

Projecting the right-hand side on Ω_T yields

$$y_{T+2,T} = \phi y_{T+1,T}.$$

Substituting our earlier-computed 1-step-ahead forecast yields

$$\begin{aligned} y_{T+2,T} &= \phi \left(\phi y_T + \theta \varepsilon_T \right) && (7.1) \\ &= \phi^2 y_T + \phi \theta \varepsilon_T. && (7.2) \end{aligned}$$

Continuing, it is clear that

$$y_{T+h,T} = \phi y_{T+h-1,T},$$

for all $h > 1$.

7.4.3 Interval and Density Forecasts

The chain rule, whether applied to pure autoregressive models or to ARMA models, is a device for simplifying the computation of point forecasts. Interval and density forecasts require the h-step-ahead forecast error variance, which we get from the moving average representation, as discussed earlier. It is

$$\sigma_h^2 = \sigma^2 \sum_{i=0}^{h-1} b_i^2,$$

which we operationalize as

$$\hat{\sigma}_h^2 = \hat{\sigma}^2 \sum_{i=0}^{h-1} \hat{b}_i^2.$$

Note that we don't actually estimate the moving average representation; rather, we solve backward for as many b's as we need, in terms of the original model parameters, which we then replace with estimates.

Let's illustrate by constructing a 2-step-ahead 95% interval forecast for the $ARMA(1,1)$ process. We already constructed the 2-step-ahead point forecast, $y_{T+2,T}$; we need only compute the 2-step-ahead forecast error variance. The process is

$$y_t = \phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$

Substitute backward for y_{t-1} to get

$$\begin{aligned} y_t &= \phi(\phi y_{t-2} + \varepsilon_{t-1} + \theta\varepsilon_{t-2}) + \varepsilon_t + \theta\varepsilon_{t-1} && (7.3) \\ &= \varepsilon_t + (\phi + \theta)\varepsilon_{t-1} + ... && (7.4) \end{aligned}$$

We need not substitute back any farther, because the 2-step-ahead forecast error variance is

$$\sigma_2^2 = \sigma^2(1 + b_1^2),$$

where b_1 is the coefficient on ε_{t-1} in the moving average representation of the ARMA(1,1) process, which we just calculated to be ($\phi + \theta$). Thus the 2-step-ahead interval forecast is $y_{T+2,T} \pm 1.96\sigma_2$, or $(\phi^2 y_T + \phi\theta\varepsilon_T) \pm 1.96\sigma\sqrt{1 + (\phi + \theta)^2}$. We make this operational as $(\hat{\phi}^2 y_T + \hat{\phi}\hat{\theta}\varepsilon_T) \pm 1.96\hat{\sigma}\sqrt{1 + (\hat{\phi} + \hat{\theta})^2}$.

7.5 Canadian Employment

We earlier examined the correlogram for the Canadian employment series, and we saw that the sample autocorrelations damp slowly and the sample partial autocorrelations cut off, just the opposite of what's expected for a moving average. Thus the correlogram indicates that a finite-order moving average process would not provide a good approximation to employment dynamics. Nevertheless, nothing stops us from fitting moving average models, so let's fit them and use the AIC and the SIC to guide model selection.

Moving average models are nonlinear in the parameters; thus, estimation proceeds by nonlinear least squares (numerical minimization). The idea is the same as when we encountered nonlinear least squares in our study of nonlinear trends – pick the parameters to minimize the sum of squared residuals – but finding an expression for the residual is a little bit trickier. To understand why moving average models are nonlinear in the parameters, and to get a feel for how they're estimated, consider an invertible MA(1) model, with a

nonzero mean explicitly included for added realism,

$$\mathrm{y_t} \;=\; \mu \;+\; \varepsilon_\mathrm{t} \;+\; \theta\varepsilon_{\mathrm{t}-1}.$$

Substitute backward m times to obtain the autoregressive approximation

$$\mathrm{y_t} \;\approx\; \frac{\mu}{1+\theta} \;+\; \theta\mathrm{y}_{\mathrm{t}-1} \;-\; \theta^2\mathrm{y}_{\mathrm{t}-2} \;+\; ... + \;(-1)^{\mathrm{m}+1}\;\theta^{\mathrm{m}}\mathrm{y}_{\mathrm{t}-\mathrm{m}} \;+\; \varepsilon_{\mathrm{t}.}$$

Thus an invertible moving average can be approximated as a finite-order autoregression. The larger is m, the better the approximation. This lets us (approximately) express the residual in terms of observed data, after which we can use a computer to solve for the parameters that minimize the sum of squared residuals,

$$\hat{\mu},\hat{\theta} \;=\; \underset{\mu,\theta}{argmin} \;\sum_{\mathrm{t}=1}^{\mathrm{T}}\left[\mathrm{y_t} \;-\; \left(\frac{\mu}{1+\theta} \;+\; \theta\mathrm{y}_{\mathrm{t}-1} \;-\; \theta^2\mathrm{y}_{\mathrm{t}-2} \;+\; ... \;+\; (-1)^{\mathrm{m}+1}\theta^{\mathrm{m}}\mathrm{y}_{\mathrm{t}-\mathrm{m}}\right)\right]^2$$

$$\hat{\sigma}^2 \;=\; \frac{1}{\mathrm{T}}\sum_{\mathrm{t}=1}^{\mathrm{T}}\left[\mathrm{y_t} \;-\; \left(\frac{\hat{\mu}}{1+\hat{\theta}} \;+\; \hat{\theta}\mathrm{y}_{\mathrm{t}-1} \;-\; \hat{\theta}^2\mathrm{y}_{\mathrm{t}-2} \;+\; ... \;+\; (-1)^{\mathrm{m}+1}\hat{\theta}^{\mathrm{m}}\mathrm{y}_{\mathrm{t}-\mathrm{m}}\right)\right]^2.$$

The parameter estimates must be found using numerical optimization methods, because the parameters of the autoregressive approximation are restricted. The coefficient of the second lag of y is the square of the coefficient on the first lag of y, and so on. The parameter restrictions must be imposed in estimation, which is why we can't simply run an ordinary least squares regression of y on lags of itself.

The next step would be to estimate $MA(q)$ models, $q = 1, 2, 3, 4$. Both the AIC and the SIC suggest that the $MA(4)$ is best. To save space, we

report only the results of $MA(4)$ estimation in Table 7.12a. The results of the $MA(4)$ estimation, although better than lower-order MAs, are nevertheless poor. The R^2 of .84 is rather low, for example, and the Durbin-Watson statistic indicates that the $MA(4)$ model fails to account for all the serial correlation in employment. The residual plot, which we show in Figure 7.12b, clearly indicates a neglected cycle, an impression confirmed by the residual correlogram (Table 7.13, Figure 7.14).

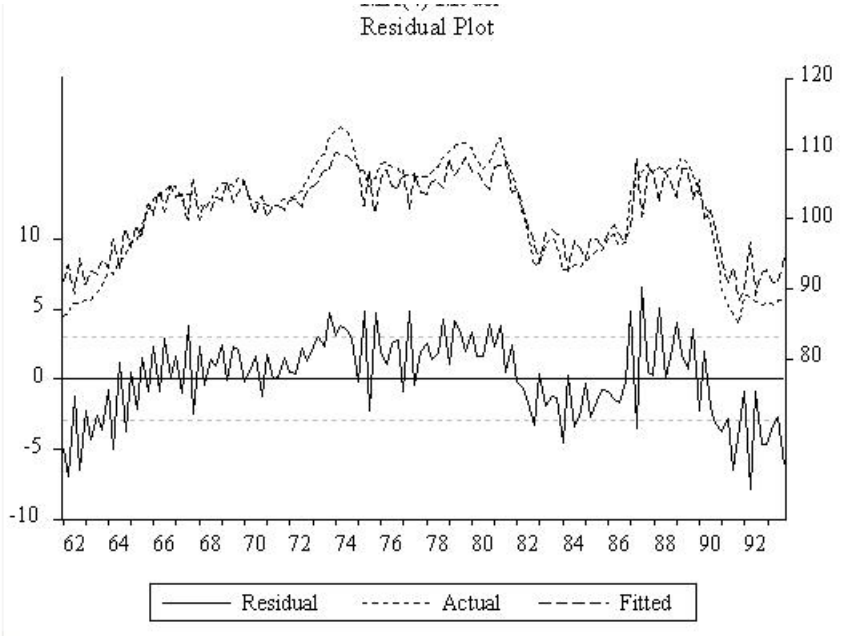
LS // Dependent Variable is CANEMP
Sample: 1962:1 1993:4
Included observations: 128
Convergence achieved after 49 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	100.5438	0.843322	119.2234	0.0000
MA(1)	1.587641	0.063908	24.84246	0.0000
MA(2)	0.994369	0.089995	11.04917	0.0000
MA(3)	-0.020305	0.046550	-0.436189	0.6635
MA(4)	-0.298387	0.020489	-14.56311	0.0000

R-squared	0.849951	Mean dependent var	101.0176
Adjusted R-squared	0.845071	S.D. dependent var	7.499163
S.E. of regression	2.951747	Akaike info criterion	2.203073
Sum squared resid	1071.676	Schwarz criterion	2.314481
Log likelihood	-317.6208	F-statistic	174.1826
Durbin-Watson stat	1.246600	Prob(F-statistic)	0.000000

Inverted MA Roots	.41	-.56+.72i	-.56 -.72i	-.87

(a) Employment MA(4) Regression



(b) Employment MA(4) Residual Plot

Figure 7.12: Employment: MA(4) Model

Sample: 1962:1 1993:4
Included observations: 128
Q-statistic probabilities adjusted for 4 ARMA term(s)

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	0.345	0.345	.088	15.614	
2	0.660	0.614	.088	73.089	
3	0.534	0.426	.088	111.01	
4	0.427	-0.042	.088	135.49	
5	0.347	-0.398	.088	151.79	0.000
6	0.484	0.145	.088	183.70	0.000
7	0.121	-0.118	.088	185.71	0.000
8	0.348	-0.048	.088	202.46	0.000
9	0.148	-0.019	.088	205.50	0.000
10	0.102	-0.066	.088	206.96	0.000
11	0.081	-0.098	.088	207.89	0.000
12	0.029	-0.113	.088	208.01	0.000

Figure 7.13: Employment MA(4) Residual Correlogram

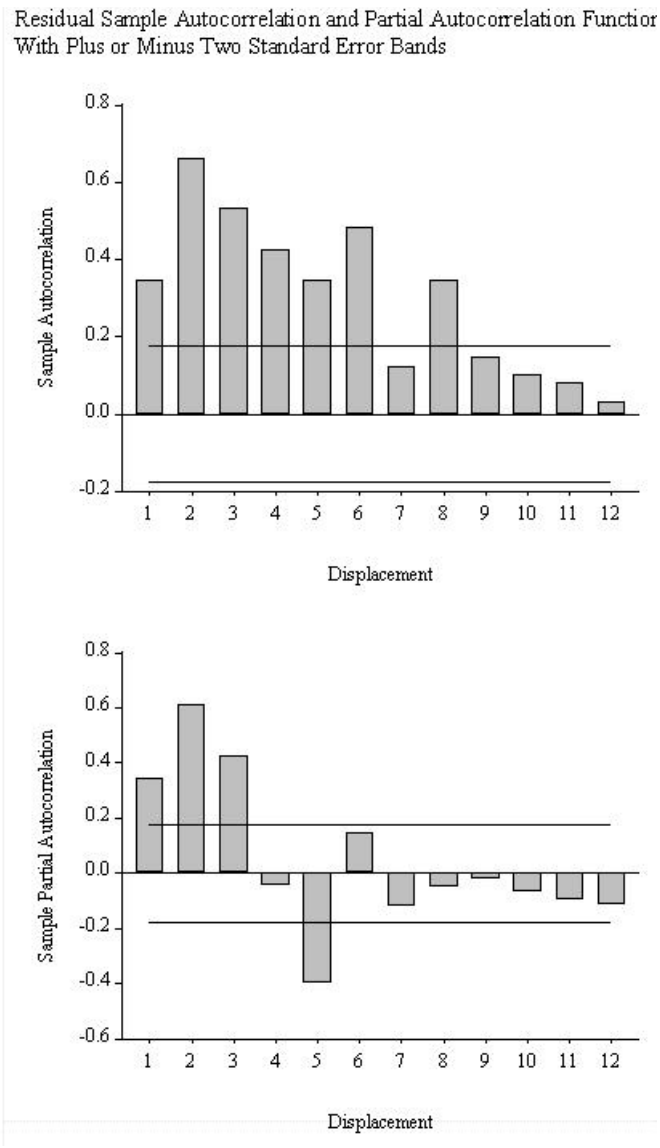


Figure 7.14: Employment MA(4) Residual Sample Autocorrelation and Partial Autocorrelation

If we insist on using a moving average model, we'd want to explore orders greater than four, but all the results thus far indicate that moving average processes don't provide good approximations to employment dynamics. Thus let's consider alternative approximations, such as autoregressions. Autoregressions can be conveniently estimated by ordinary least squares regression. Consider, for example, the $AR(1)$ model,

$$\begin{aligned}(\mathrm{y_t} - \mu) \;&=\; \phi(\mathrm{y_{t-1}} - \mu) \;+\; \varepsilon_\mathrm{t} \\ \varepsilon_\mathrm{t} \;&\sim\; (0, \sigma^2)\end{aligned}$$

We can write it as

$$\mathrm{y_t} \;=\; \mathrm{c} \;+\; \phi \mathrm{y_{t-1}} \;+\; \varepsilon_\mathrm{t}$$

where $c \;=\; \mu(1-\phi)$. The least squares estimators are

$$\begin{aligned}\hat{\mathrm{c}}, \hat{\phi} \;&=\; \underset{c,\,\phi}{argmin} \;\sum_{\mathrm{t}=1}^{\mathrm{T}} [\mathrm{y_t} \;-\; \mathrm{c} \;-\; \phi \mathrm{y_{t-1}}]^2 \\ \hat{\sigma}^2 \;&= \frac{1}{\mathrm{T}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \left[\mathrm{y_t} \;-\; \hat{\mathrm{c}} \;-\; \hat{\phi} \mathrm{y_{t-1}}\right]^2 .\end{aligned}$$

The implied estimate of μ is

$$\hat{\mu} \;=\; \hat{\mathrm{c}}/(1-\hat{\phi}).$$

Unlike the moving average case, for which the sum of squares function is nonlinear in the parameters, requiring the use of numerical minimization methods, the sum of squares function for autoregressive processes is linear in the parameters, so that estimation is particularly stable and easy. In the $AR(1)$ case, we simply run an ordinary least squares regression of y on one

lag of y; in the $AR(p)$ case, we regress y on p lags of y.

We estimate $AR(p)$ models, $p = 1, 2, 3, 4$. Both the AIC and the SIC suggest that the $AR(2)$ is best. To save space, we report only the results of $AR(2)$ estimation in Table 7.15a. The estimation results look good, and the residuals (Figure 7.15b) look like white noise. The residual correlogram (Table 7.16, Figure 7.17) supports that conclusion.

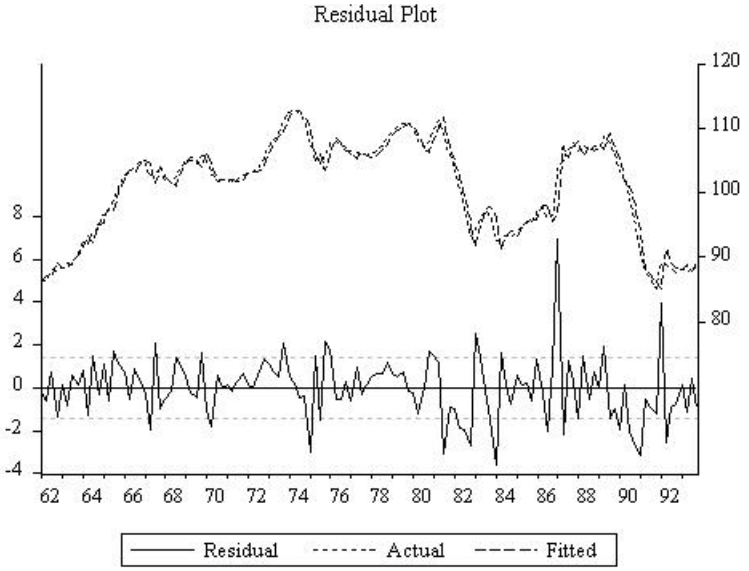
LS // Dependent Variable is CANEMP
Sample: 1962:1 1993:4
Included observations: 128
Convergence achieved after 3 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	101.2413	3.399620	29.78017	0.0000
AR(1)	1.438810	0.078487	18.33188	0.0000
AR(2)	-0.476451	0.077902	-6.116042	0.0000

R-squared	0.963372	Mean dependent var	101.0176
Adjusted R-squared	0.962786	S.D. dependent var	7.499163
S.E. of regression	1.446663	Akaike info criterion	0.761677
Sum squared resid	261.6041	Schwarz criterion	0.828522
Log likelihood	-227.3715	F-statistic	1643.837
Durbin-Watson stat	2.067024	Prob(F-statistic)	0.000000

Inverted AR Roots .92 .52

(a) Employment AR(2) Model



(b) Employment AR(2) Residual Plot

Figure 7.15: Employment: MA(4) Model

Sample: 1962:1 1993:4
Included observations: 128
Q-statistic probabilities adjusted for 2 ARMA term(s)

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	-0.035	-0.035	.088	0.1606	
2	0.044	0.042	.088	0.4115	
3	0.011	0.014	.088	0.4291	0.512
4	0.051	0.050	.088	0.7786	0.678
5	0.002	0.004	.088	0.7790	0.854
6	0.019	0.015	.088	0.8272	0.935
7	-0.024	-0.024	.088	0.9036	0.970
8	0.078	0.072	.088	1.7382	0.942
9	0.080	0.087	.088	2.6236	0.918
10	0.050	0.050	.088	2.9727	0.936
11	-0.023	-0.027	.088	3.0504	0.962
12	-0.129	-0.148	.088	5.4385	0.860

Figure 7.16: Employment AR(2) Residual Correlogram

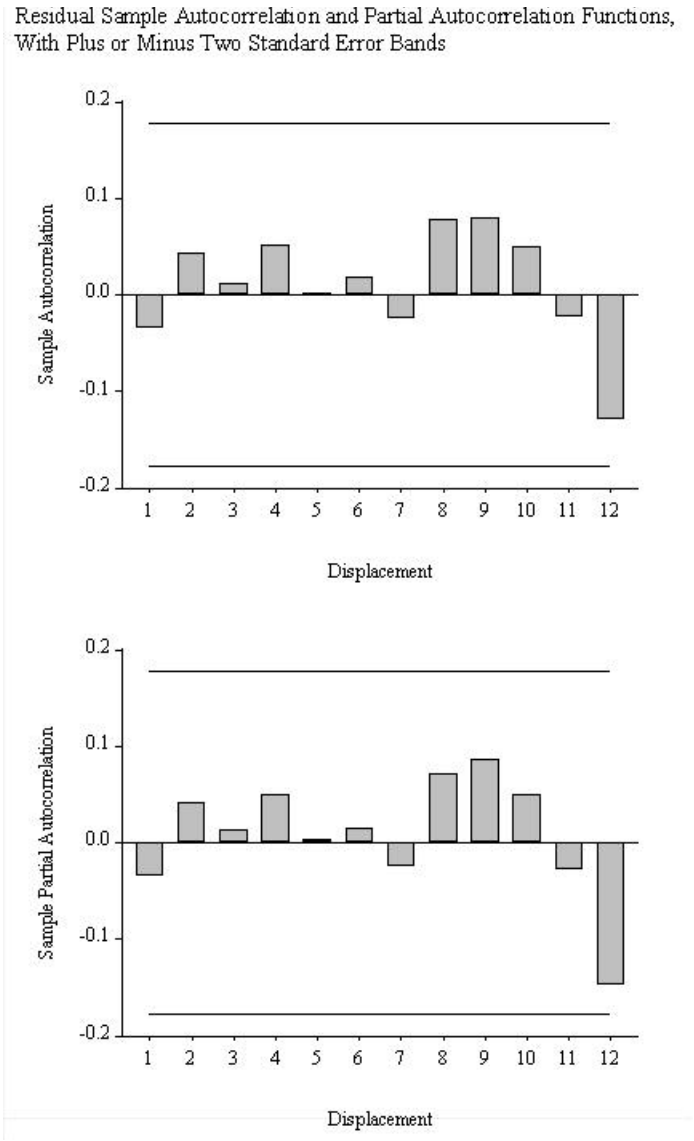


Figure 7.17: Employment AR(2) Residual Sample Autocorrelation and Partial Autocorrelation

				MA Order		
		0	1	2	3	4
	0		2.86	2.32	2.47	2.20
	1	1.01	.83	.79	.80	.81
AR Order	2	.762	.77	.78	.80	.80
	3	.77	.761	.77	.78	.79
	4	.79	.79	.77	.79	.80

(a) Employment AIC Values

Various ARMA Models

				MA Order		
		0	1	2	3	4
	0		2.91	2.38	2.56	2.31
	1	1.05	.90	.88	.91	.94
AR Order	2	.83	.86	.89	.92	.96
	3	.86	.87	.90	.94	.96
	4	.90	.92	.93	.97	1.00

(b) Employment SIC Values

Figure 7.18: Employment - Information Criterion for ARMA Models

Finally, we consider $ARMA(p, q)$ approximations to the Wold representation. $ARMA$ models are estimated in a fashion similar to moving average models; they have autoregressive approximations with nonlinear restrictions on the parameters, which we impose when doing a numerical sum of squares minimization. We examine all $ARMA(p, q)$ models with p and q less than or equal to four; the SIC and AIC values appear in Tables 7.18a and 7.18b. The SIC selects the $AR(2)$ (an $ARMA(2, 0)$), which we've already discussed. The AIC, which penalizes degrees of freedom less harshly, selects an $ARMA(3, 1)$ model. The $ARMA(3, 1)$ model looks good; the estimation results appear in Table 7.19a, the residual plot in Figure 7.19b, and the residual correlogram in Table 7.20 and Figure fig: employment arma(3,1) residual sample autocorrelation and partial autocorrelation.

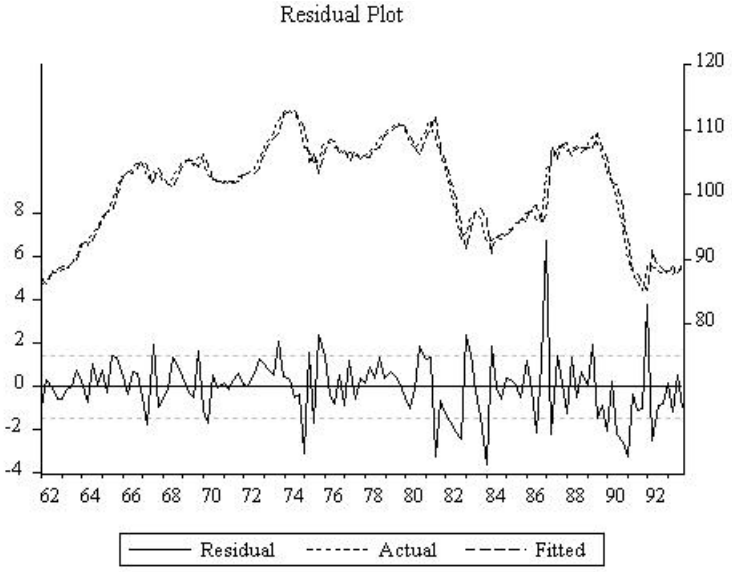
LS // Dependent Variable is CANEMP
Sample: 1962:1 1993:4
Included observations: 128
Convergence achieved after 17 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	101.1378	3.538602	28.58130	0.0000
AR(1)	0.500493	0.087503	5.719732	0.0000
AR(2)	0.872194	0.067096	12.99917	0.0000
AR(3)	-0.443355	0.080970	-5.475560	0.0000
MA(1)	0.970952	0.035015	27.72924	0.0000

R-squared	0.964535	Mean dependent var	101.0176
Adjusted R-squared	0.963381	S.D. dependent var	7.499163
S.E. of regression	1.435043	Akaike info criterion	0.760668
Sum squared resid	253.2997	Schwarz criterion	0.872076
Log likelihood	-225.3069	F-statistic	836.2912
Durbin-Watson stat	2.057302	Prob(F-statistic)	0.000000

Inverted AR Roots	.93	.51	-.94
Inverted MA Roots	-.97		

(a) Employment ARMA(3,1) Model



(b) Employment ARMA(3,1) Residual Plot

Figure 7.19: Employment: MA(4) Model

Residual Correlogram

Sample: 1962:1 1993:4
Included observations: 128
Q-statistic probabilities adjusted for 4 ARMA term(s)

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	-0.032	-0.032	.09	0.1376	
2	0.041	0.040	.09	0.3643	
3	0.014	0.017	.09	0.3904	
4	0.048	0.047	.09	0.6970	
5	0.006	0.007	.09	0.7013	0.402
6	0.013	0.009	.09	0.7246	0.696
7	-0.017	-0.019	.09	0.7650	0.858
8	0.064	0.060	.09	1.3384	0.855
9	0.092	0.097	.09	2.5182	0.774
10	0.039	0.040	.09	2.7276	0.842
11	-0.016	-0.022	.09	2.7659	0.906
12	-0.137	-0.153	.09	5.4415	0.710

Figure 7.20: Employment ARMA(3,1) Correlogram

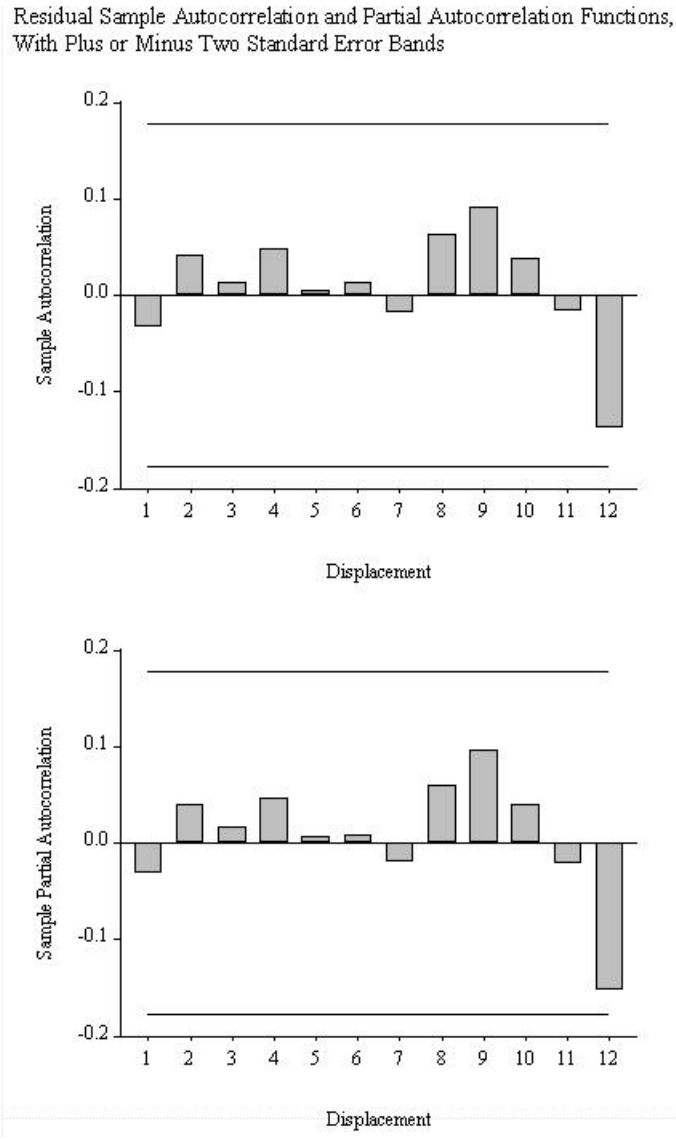


Figure 7.21: Employment ARMA(3,1) Residual Sample Autocorrelation and Partial Autocorrelation

Although the $ARMA(3,1)$ looks good, apart from its lower AIC it looks no better than the $AR(2)$, which basically seemed perfect. In fact, there are at least three reasons to prefer the $AR(2)$. First, for the reasons that we discussed in Chapter 15 , when the AIC and the SIC disagree we recommend using the more parsimonious model selected by the SIC. Second, if we consider a model selection strategy involving not just examination of the AIC and SIC, but also examination of autocorrelations and partial autocorrelations, which we advocate, we're led to the $AR(2)$. Finally, and importantly, the impression that the $ARMA(3,1)$ provides a richer approximation to employment dynamics is likely spurious in this case. The $ARMA(3,1)$ has a inverse autoregressive root of -.94 and an inverse moving average root of -.97. Those roots are of course just *estimates* and are likely to be statistically indistinguishable from one another, in which case we can *cancel* them, which brings us down to an $ARMA(2,0)$, or $AR(2)$, model with roots virtually indistinguishable from those of our earlier-estimated $AR(2)$ process! We refer to this situation as one of common factors in an $ARMA$ model. Look out for such situations, which can lead to substantial model simplification.

Now we put our forecasting technology to work to produce point and interval forecasts for Canadian employment. Recall that the best moving average model was an MA(4), while the best autoregressive model, as well as the best ARMA model and the best model overall, was an AR(2).

Consider forecasting with the MA(4) model. Figure 7.22 shows employment history together with operational 4-quarter-ahead point and interval extrapolation forecasts. The 4-quarter-ahead extrapolation forecast reverts quickly to the mean of the employment index. In 1993.4, the last quarter of historical data, employment is well below its mean, but the forecast calls for a quick rise. The forecasted quick rise seems unnatural, because employment dynamics are historically very persistent. If employment is well below its mean in 1993.4, we'd expect it to stay below its mean for some time.

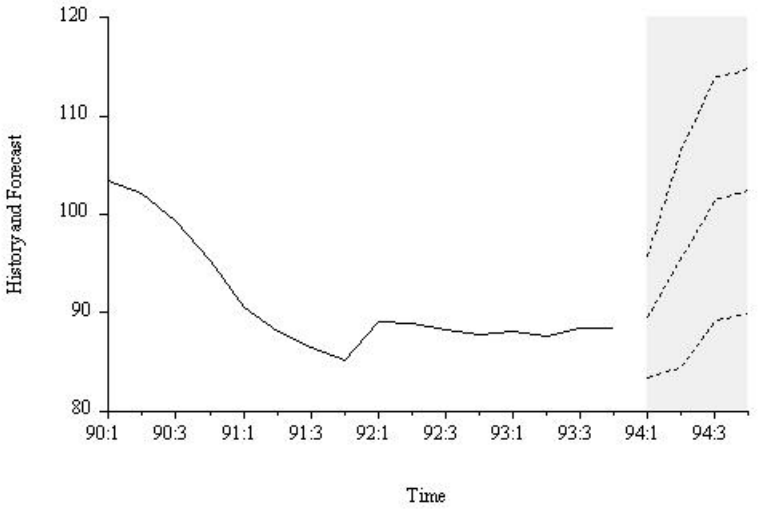


Figure 7.22: Employment History and Forecast - MA(4)

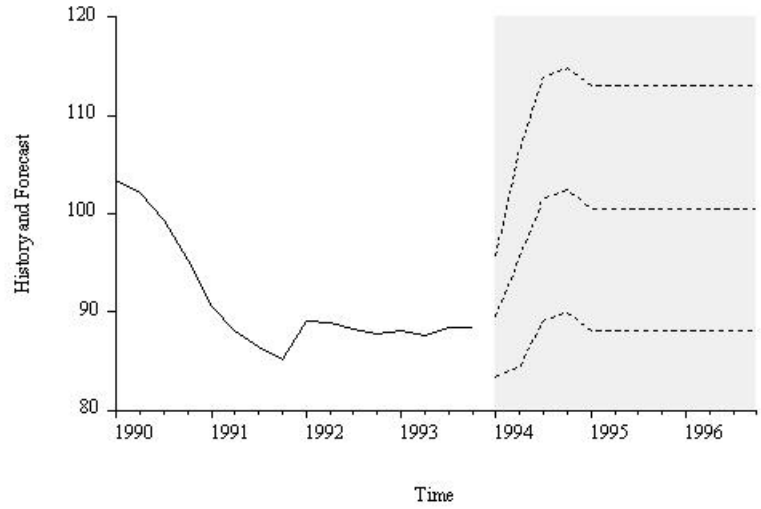


Figure 7.23: Employment History and Long-Horizon Forecast - MA(4)

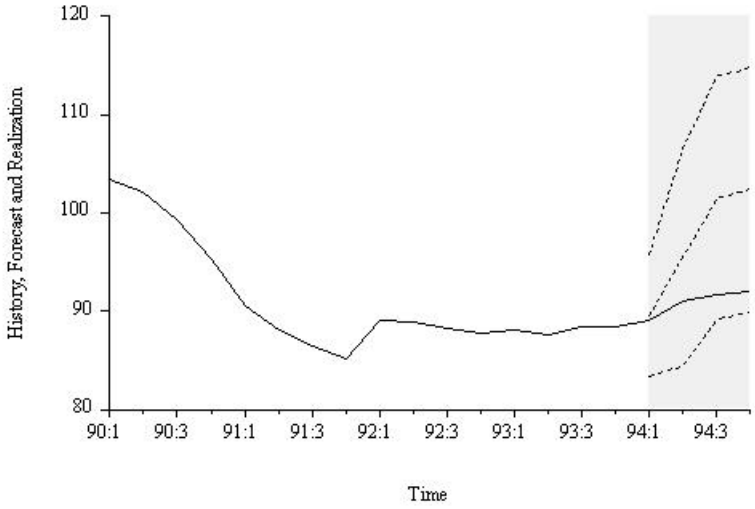


Figure 7.24: Employment History, Forecast, and Realization - MA(4)

The MA(4) model is unable to capture such persistence. The quick reversion of the MA(4) forecast to the mean is a manifestation of the short memory of moving average processes. Recall, in particular, that an MA(4) process has a 4-period memory – all autocorrelations are zero beyond displacement 4. Thus, all forecasts more than four steps ahead are simply equal to the unconditional mean (100.2), and all 95% interval forecasts more than four steps ahead are plus or minus 1.96 unconditional standard deviations. All of this is made clear in Figure 7.23, in which we show the employment history together with 12-step-ahead point and interval extrapolation forecasts.

In Figure 7.24 we show the 4-quarter-ahead forecast and realization. Our suspicions are confirmed. The actual employment series stays well below its mean over the forecast period, whereas the forecast rises quickly back to the mean. The mean squared forecast error is a large 55.9.

Now consider forecasting with the AR(2) model. In Figure 7.25 we show the 4-quarter-ahead extrapolation forecast, which reverts to the unconditional mean much less quickly, as seems natural given the high persistence of employment. The 4-quarter-ahead point forecast, in fact, is still well below the mean. Similarly, the 95% error bands grow gradually and haven't approached their long-horizon values by four quarters out.

Figures 7.26 and 7.28 make clear the very different nature of the autoregressive forecasts. Figure 7.26 presents the 12-step-ahead extrapolation forecast, and Figure 7.28 presents a much longer-horizon extrapolation forecast. Eventually the unconditional mean *is* approached, and eventually the error bands do go flat, but only for very long-horizon forecasts, due to the high persistence in employment, which the AR(2) model captures.

In Figure 7.27 we show the employment history, 4-quarter-ahead AR(2) extrapolation forecast, and the realization. The AR(2) forecast appears quite accurate; the mean squared forecast error is 1.3, drastically smaller than that of the MA(4) forecast.

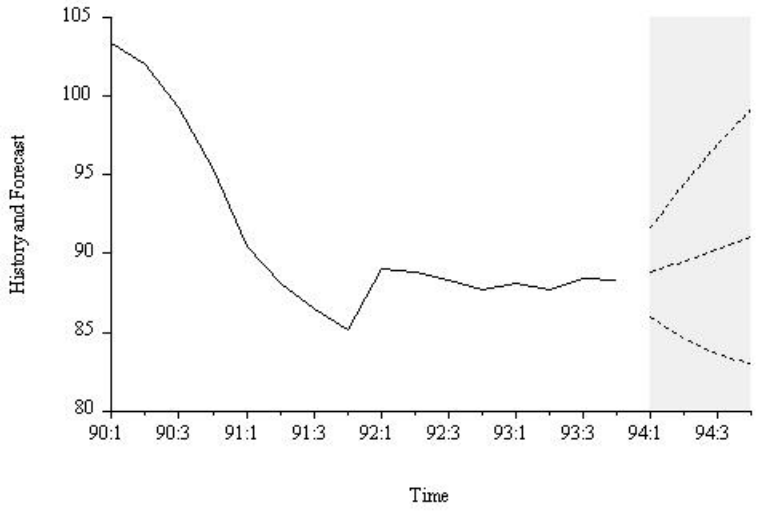


Figure 7.25: Employment History and Forecast - AR(2)

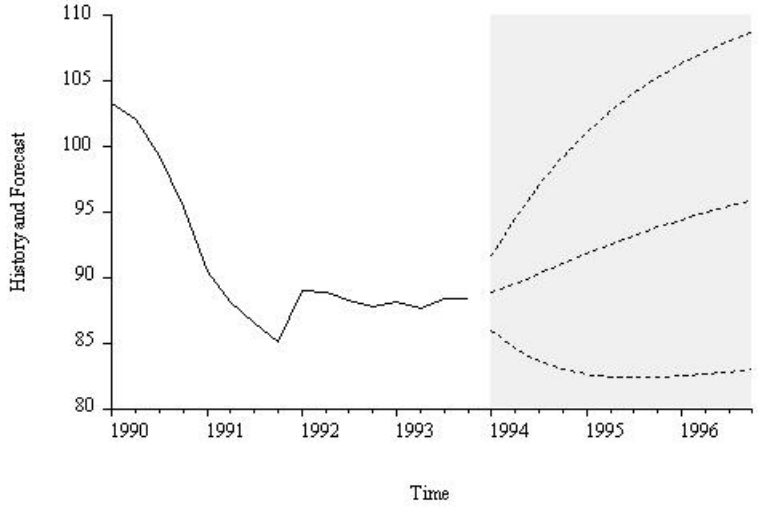


Figure 7.26: Employment History and Forecast, 12-step ahead - AR(2)

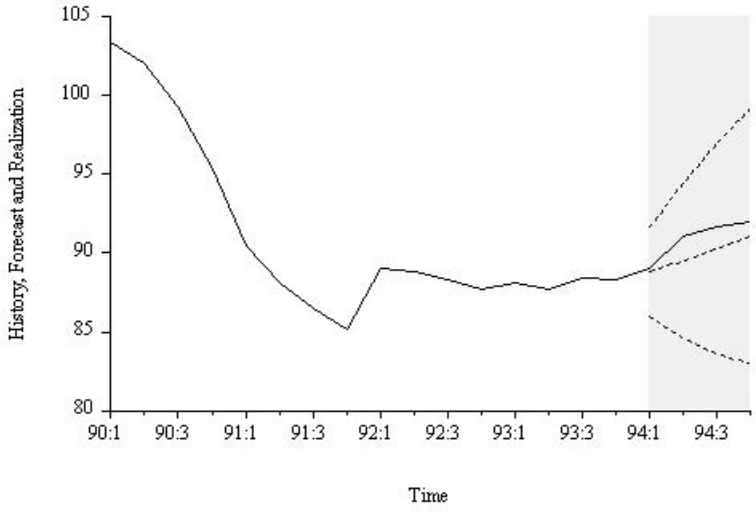


Figure 7.27: Employment History, Forecast, and Realization - AR(2)

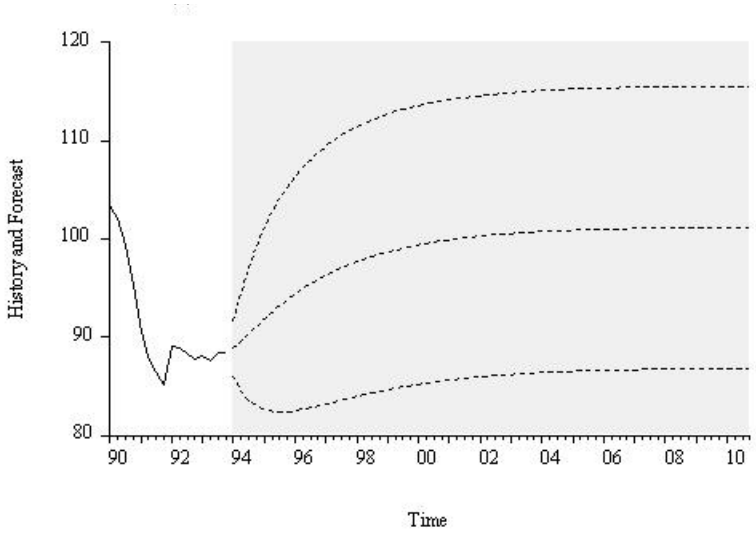


Figure 7.28: Employment History and Long-Horizon Forecast - AR(2)

7.6 Exercises, Problems and Complements

1. Shapes of correlograms.

   Given the following ARMA processes, sketch the expected forms of the autocorrelation and partial autocorrelation functions. (Hint: examine the roots of the various autoregressive and moving average lag operator polynomials.)

   (a) $y_t = \left(\frac{1}{1-1.05L-.09L^2}\right)\varepsilon_t$

   (b) $y_t = (1 - .4L)\varepsilon_t$

   (c) $y_t = \left(\frac{1}{1-.7L}\right)\varepsilon_t$.

2. The autocovariance function of the $MA(1)$ process, revisited.

   In the text we wrote

$$\gamma(\tau) = E(y_t y_{t-\tau}) = E((\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-\tau} + \theta\varepsilon_{t-\tau-1})) = \begin{cases} \theta\sigma^2, \ \tau = 1 \\ \\ 0, \ otherwise \end{cases}.$$

   Fill in the missing steps by evaluating explicitly the expectation

$$E((\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-\tau} + \theta\varepsilon_{t-\tau-1})).$$

3. ARMA algebra.

   Derive expressions for the autocovariance function, autocorrelation function, conditional mean, unconditional mean, conditional variance and unconditional variance of the following processes:

   (a) $y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}$

   (b) $y_t = \phi y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$.

4. Mechanics of fitting ARMA models.

   You have data for daily transfers over BankWire, a financial wire transfer system in a country responsible for much of the world's finance, over a recent span of 200 business days.

   (a) Is trend or seasonality operative? Defend your answer.

   (b) Find a parsimonious $ARMA(p, q)$ model that fits well, and defend its adequacy.

   (c) In item 4b above, you were asked to find a parsimonious ARMA(p,q) model that fits the transfer data well, and to defend its adequacy. Repeat the exercise, this time using only the first 175 days for model selection and fitting. Is it necessarily the case that the selected ARMA model will remain the same as when all 200 days are used? Does yours?

   (d) Use your estimated model to produce point and interval forecasts for days 176 through 200. Plot them and discuss the forecast pattern.

   (e) Compare your forecasts to the actual realizations. Do the forecasts perform well? Why or why not?

5. A different way to estimate autoregressive models.

   We discussed estimation of autoregressive models using ordinary least squares. We could also write the model as a regression on an intercept, with a serially correlated disturbance. Thus the autoregressive model is

$$y_t = \mu + \varepsilon_t$$

$$\Phi(L)\varepsilon_t = v_t$$

$$v_t \sim WN(0,\ \sigma^2).$$

We can estimate the model using nonlinear least squares. Eviews and other forecasting packages proceed in precisely that way.[13]

This framework – regression on a constant with serially correlated disturbances – has a number of attractive features. First, the mean of the process is the regression constant term.[14] Second, it leads us naturally toward regression on more than just a constant, as other right-hand side variables can be added as desired.

6. **Aggregation** and **disaggregation**: **top-down vs. bottom-up forecasting models**.

   Related to the issue of methods and complexity discussed in Chapter 2 is the question of aggregation. Often we want to forecast an aggregate, such as total sales of a manufacturing firm, but we can take either an aggregated or disaggregated approach.

   Suppose, for example, that total sales is composed of sales of three products. The aggregated, or top-down, or macro, approach is simply to model and forecast total sales. The disaggregated, or bottom- up, or micro, approach is to model and forecast separately the sales of the individual products, and then to add them together.

   (a) Perhaps surprisingly, it's impossible to know in advance whether the aggregated or disaggregated approach is better. It all depends on the specifics of the situation; the only way to tell is to try both approaches and compare the forecasting results.

   (b) However, in real-world situations characterized by likely model misspecification and parameter estimation uncertainty, there are reasons to suspect that the aggregated approach may be preferable.

[13] That's why, for example, information on the number of iterations required for convergence is presented even for estimation of the autoregressive model.

[14] Hence the notation "μ" for the intercept.

First, standard (e.g., linear) models fit to aggregated series may be less prone to specification error, because aggregation can produce approximately linear relationships even when the underlying disaggregated relationships are not linear. Second, if the disaggregated series depend in part on a common factor (e.g., general business conditions) then it will emerge more clearly in the aggregate data. Finally, modeling and forecasting of one aggregated series, as opposed to many disaggregated series, relies on far fewer parameter estimates.

(c) Of course, if our interest centers on the disaggregated components, then we have no choice but to take a disaggregated approach.

(d) Sometimes, even if interest centers on an aggregate, there may no data available for it, but there may be data for relevant components. Consider, for example, forecasting the number of pizzas eaten next year by Penn students. There's no annual series available for "pizzas eaten by Penn students," but there may be series of Penn enrollment, annual U.S. pizza consumption, U.S. population, etc. from which a forecast could be built. This is called "Fermi-izing" the problem, after the great Italian physicist Enrico Fermi. See Tetlock and Gardner (2015), chapter 5.

(e) It is possible that an aggregate forecast may be useful in forecasting disaggregated series. Why? (Hint: See Fildes and Stekler, 2000.)

7. Forecasting an $ARMA(2,2)$ process.

Consider the $ARMA(2,2)$ process:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}.$$

a. Verify that the optimal 1-step ahead forecast made at time T is

$$y_{T+1,T} = \phi_1 y_T + \phi_2 y_{T-1} + \theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1}.$$

b. Verify that the optimal 2-step ahead forecast made at time T is

$$y_{T+2,T} = \phi_1 y_{T+1,T} + \phi_2 y_T + \theta_2 \varepsilon_T,$$

and express it purely in terms of elements of the time-T information set.

c. Verify that the optimal 3-step ahead forecast made at time T is

$$y_{T+3,T} = \phi_1 y_{T+2,T} + \phi_2 y_{T+1,T},$$

and express it purely in terms of elements of the time-T information set.

d. Show that for any forecast horizon h greater than or equal to three,

$$y_{T+h,T} = \phi_1 y_{T+h-1,T} + \phi_2 y_{T+h-2,T}.$$

8. ARMA lag inclusion.

In our MA model fitting for employment, why did we leave the $MA(3)$ term in the preferred $MA(4)$ model, despite the insignificant p-value? Discuss costs and benefits of dropping the insignificant $MA(3)$ term.

9. Modeling cyclical dynamics.

As a research analyst at the U.S. Department of Energy, you have been asked to model non-seasonally-adjusted U.S. imports of crude oil.

(a) Find a suitable time series on the web.

(b) Create a model that captures the trend in the series.

(c) Adding to the model from part 9b, create a model with trend and a full set of seasonal dummy variables.

(d) Observe the residuals of the model from part b and their correlogram. Is there evidence neglected dynamics? If so, what to do?

10. Applied $ARMA$ modeling.

Nile.com, a successful on-line bookseller, monitors and forecasts the number of "hits" per day to its web page. You have daily hits data for 1/1/98 through 9/28/98.

a. Fit and assess the standard linear, quadratic, and log linear trend models.

b. For a few contiguous days roughly in late April and early May, hits were much higher than usual during a big sale. Do you find evidence of a corresponding group of outliers in the residuals from your trend models? Do they influence your trend estimates much? How should you treat them?

c. Model and assess the significance of day-of-week effects in Nile.com web page hits.

d. Select a final model, consisting only of trend and seasonal components, to use for forecasting.

e. Use your model to forecast Nile.com hits through the end of 1998.

f. Generalize your earlier trend + seasonal model to allow for cyclical dynamics, if present, via $ARMA(p, q)$ disturbances. Write the full specification of your model in general notation (e.g., with p and q left unspecified).

g. Estimate all models, corresponding to $p = 0, 1, 2, 3$ and $q = 0, 1, 2, 3$, while leaving the original trend and seasonal specifications intact, and select the one that optimizes SIC.

h. Using the model selected in part 10g, write theoretical expressions for the 1- and 2-day- ahead point forecasts and 95% interval forecasts, using estimated parameters.

i. Calculate those point and interval forecasts for Nile.com for 9/29 and 9/30.

11. Mechanics of fitting ARMA models.

On the book's web page you will find data for daily transfers over BankWire, a financial wire transfer system in a country responsible for much of the world's finance, over a recent span of 200 business days.

a. Is trend or seasonality operative? Defend your answer.

b. Find a parsimonious $ARMA(p, q)$ model that fits well, and defend its adequacy.

c. Repeat the exercise 11b, this time using only the first 175 days for model selection and fitting. Is it necessarily the case that the selected ARMA model will remain the same as when all 200 days are used? Does yours?

d. Use your estimated model to produce point and interval forecasts for days 176 through 200. Plot them and discuss the forecast pattern.

e. Compare your forecasts to the actual realizations. Do the forecasts perform well? Why or why not?

f. Discuss precisely how your software constructs point and interval forecasts. It should certainly match our discussion in spirit, but it may differ in some of the details. Are you uncomfortable with any of the assumptions made? How, if at all, could the forecasts be improved?

7.7 Notes

Our discussion of estimation was a bit fragmented; we discussed estimation of moving average and ARMA models using nonlinear least squares, whereas we discussed estimation of autoregressive models using ordinary least squares. A more unified approach proceeds by writing each model as a regression on an intercept, with a serially correlated disturbance. Thus the moving average model is

$$\begin{aligned} y_t &= \mu + \varepsilon_t \\ \varepsilon_t &= \Theta(L)v_t \\ v_t &\sim WN(0,\ \sigma^2), \end{aligned}$$

the autoregressive model is

$$\begin{aligned} y_t &= \mu + \varepsilon_t \\ \Phi(L)\varepsilon_t &= v_t \\ v_t &\sim WN(0,\ \sigma^2), \end{aligned}$$

and the ARMA model is

$$\begin{aligned} y_t &= \mu + \varepsilon_t \\ \Phi(L)\varepsilon_t &= \Theta(L)v_t \\ v_t &\sim WN(0,\ \sigma^2). \end{aligned}$$

We can estimate each model in identical fashion using nonlinear least squares. Eviews and other forecasting packages proceed in precisely that way.[15]

This framework – regression on a constant with serially correlated disturbances – has a number of attractive features. First, the mean of the process is the regression constant term.[16] Second, it leads us naturally toward re-

[15] That's why, for example, information on the number of iterations required for convergence is presented even for estimation of the autoregressive model.

[16] Hence the notation "μ" for the intercept.

gression on more than just a constant, as other right-hand side variables can be added as desired. Finally, it exploits the fact that because autoregressive and moving average models are special cases of the ARMA model, their estimation is also a special case of estimation of the ARMA model.

Our description of estimating ARMA models – compute the autoregressive representation, truncate it, and estimate the resulting approximate model by nonlinear least squares – is conceptually correct but intentionally simplified. The actual estimation methods implemented in modern software are more sophisticated, and the precise implementations vary across software packages. Beneath it all, however, all estimation methods are closely related to our discussion, whether implicitly or explicitly. You should consult your software manual for details.