Chapter 4

Analysis of a Single Time Series

Note: The primary reference for these notes is Enders (2004). An alternative and more technical treatment can be found in Hamilton (1994).

Most data used in financial econometrics occur sequentially through time. Interest rates, asset returns, and foreign exchange rates are all examples of time series. This chapter introduces time-series econometrics and focuses primarily on linear models, although some common non-linear models are described in the final section. The analysis of time-series data begins by defining two key concepts in the analysis of time series: stationarity and ergodicity. The chapter next turns to Autoregressive Moving Average models (ARMA) and covers the structure of these models, stationarity conditions, model selection, estimation, inference, and forecasting. Finally, The chapter concludes by examining non-stationary time series.

4.1 Stochastic Processes

A stochastic process is an arbitrary sequence of random data and is denoted

$$\{y_t\} \tag{4.1}$$

where $\{\cdot\}$ is used to indicate that the ys form a sequence. The simplest non-trivial stochastic process specifies that $y_t \overset{\text{i.i.d.}}{\sim} D$ for some distribution D, for example normal. Another simple stochastic process is the random walk,

$$y_t = y_{t-1} + \epsilon_t$$

where ϵ_t is an i.i.d. process.

4.2 Stationarity, Ergodicity, and the Information Set

Stationarity is a probabilistically meaningful measure of regularity. This regularity can be exploited to estimate unknown parameters and characterize the dependence between observa-

tions across time. If the data generating process frequently changed in an unpredictable manner, constructing a meaningful model would be difficult or impossible.

Stationarity exists in two forms, strict stationarity and covariance (also known as weak) stationarity. Covariance stationarity is important when modeling the mean of a process, although strict stationarity is useful in more complicated settings, such as non-linear models.

Definition 4.1 (Strict Stationarity)**.** A stochastic process $\{y_t\}$ is strictly stationary if the joint distribution of $\{y_t, y_{t+1}, \ldots, y_{t+h}\}$ only depends only on h and not on t.

Strict stationarity requires that the *joint* distribution of a stochastic process does not depend on time and so the only factor affecting the relationship between two observations is the gap between them. Strict stationarity is weaker than i.i.d. since the process may be dependent but it is nonetheless a strong assumption and implausible for many time series, including both financial and macroeconomic data.

Covariance stationarity, on the other hand, only imposes restrictions on the first two moments of a stochastic process.

Definition 4.2 (Covariance Stationarity)**.** A stochastic process $\{y_t\}$ is covariance stationary if

$$\begin{aligned} \mathrm{E}[y_t] &= \mu \quad \text{for } t = 1, 2, \ldots \\ \mathrm{V}[y_t] &= \sigma^2 < \infty \quad \text{for } t = 1, 2, \ldots \\ \mathrm{E}[(y_t - \mu)(y_{t-s} - \mu)] &= \gamma_s \quad \text{for } t = 1, 2, \ldots, s = 1, 2, \ldots, t-1. \end{aligned} \tag{4.2}$$

Covariance stationarity requires that both the unconditional mean and unconditional variance are finite and do not change with time. Note that covariance stationarity only applies to *unconditional moments* and not conditional moments, and so a covariance process may have a varying conditional mean (i.e. be predictable).

These two types of stationarity are related although neither nests the other. If a process is strictly stationary *and* has finite second moments, then it is covariance stationary. If a process is covariance stationary and the joint distribution of the studentized residuals (demeaned and standardized by their standard deviation) does not depend on time, then the process is strictly stationary. However, one type can occur without the other, both can occur or neither may be applicable to a particular time series. For example, if a process has higher order moments which depend on time (e.g., time-varying kurtosis), it may be covariance stationary but not strictly stationary. Alternatively, a sequence of i.i.d. Student's t random variables with 2 degrees of freedom is strictly stationary but not covariance stationary since the variance of a t_2 is infinite.

$\gamma_s = \mathrm{E}[(y_t - \mu)(y_{t-s} - \mu)]$ is the covariance of y_t with itself at a different point in time, known as the s$^{\text{th}}$ autocovariance. γ_0 is the lag-0 autocovariance, the same quantity as the *long-run* variance of y_t (i.e. $\gamma_0 = \mathrm{V}[y_t]$).[1]

Definition 4.3 (Autocovariance)**.** The autocovariance of a covariance stationary scalar process $\{y_t\}$ is defined

$$\gamma_s = \mathrm{E}[(y_t - \mu)(y_{t-s} - \mu)] \tag{4.3}$$

where $\mu = \mathrm{E}[y_t]$. Note that $\gamma_0 = \mathrm{E}[(y_t - \mu)(y_t - \mu)] = \mathrm{V}[y_t]$.

[1]The use of long-run variance is used to distinguish $\mathrm{V}[y_t]$ from the innovation variance, $\mathrm{V}[\epsilon_t]$, also known as the short-run variance.

Ergodicity is another important concept in the analysis of time series and is one form of asymptotic independence.

Definition 4.4 (Ergodicity)**.** Let $\{y_t\}$ be a stationary sequence. $\{y_t\}$ is ergodic if for any two bounded functions $f : \mathbb{R}^k \to \mathbb{R}$ $g : \mathbb{R}^l \to \mathbb{R}$

$$\begin{aligned} &\lim_{j\to\infty} \left|\mathrm{E}\left[f\left(y_t,\ldots,y_{t+k}\right)g\left(y_{t+j},\ldots,y_{t+l+j}\right)\right]\right| \\ &\quad = \left|\mathrm{E}\left[f\left(y_t,\ldots,y_{t+k}\right)\right]\right| \left|\mathrm{E}\left[g\left(y_{t+j},\ldots,y_{t+l+j}\right)\right]\right| \end{aligned} \tag{4.4}$$

In essence, if an ergodic stochastic process is sampled at two points far apart in time, these samples will be independent. The ergodic theorem provides a practical application of ergodicity.

Theorem 4.1 (Ergodic Theorem)**.** *If* $\{y_t\}$ *is ergodic and its* r^{th} *moment* μ_r *is finite, then* $T^{-1}\sum_{t=1}^{T} y_t^r \xrightarrow{p} \mu_r$.

The ergodic theorem states that averages will converge to their expectation provided the expectation exists. The intuition for this results follows from the definition of ergodicity since samples far apart in time are (effectively) independent, and so errors average across time.

Not all series are ergodic. Let $y_t = \eta + \epsilon_t$ where $\eta \sim N(0,1)$, $\epsilon_t \overset{\text{i.i.d.}}{\sim} N(0,1)$ and η and ϵ_t are independent for any t. Note that η is drawn only once (not every t). Clearly, $\mathrm{E}[y_t] = 0$. However, $T^{-1}\sum_{t=1}^{T} y_t \xrightarrow{p} \eta \neq 0$, and so even though the average converges it does not converge to $\mathrm{E}[y_t]$ since the effect of the initial draw of η is present in every observation of $\{y_t\}$.

The third important building block of time-series models is white noise. White noise generalizes i.i.d. noise and allows for dependence in a series as long as three conditions are satisfied: the series is mean zero, uncorrelated and has finite second moments.

Definition 4.5 (White Noise)**.** A process $\{\epsilon_t\}$ is known as white noise if

$$\begin{aligned} \mathrm{E}[\epsilon_t] &= 0 \qquad \text{for } t = 1,2,\ldots \\ \mathrm{V}[\epsilon_t] &= \sigma^2 < \infty \qquad \text{for } t = 1,2,\ldots \\ \mathrm{E}\left[\epsilon_t \epsilon_{t-j}\right] &= \mathrm{Cov}(\epsilon_t, \epsilon_{t-j}) = 0 \qquad \text{for } t = 1,2,\ldots, j \neq 0. \end{aligned} \tag{4.5}$$

An i.i.d. series with finite second moments is trivially white noise, but other important processes, such as residuals following an ARCH (Autoregressive Conditional Heteroskedasticity) process, may also be white noise although not independent since white noise only requires linear independence.[2] A white noise process is also covariance stationary since it satisfies all three conditions: the mean, variance, and autocovariances are all finite and do not depend on time.

The final important concepts are conditional expectation and the information set. The information set at time t is denoted $\mathcal{F}_t$ and contains all time t *measurable* events[3], and so the information set includes realization of all variables which have occurred on or before t. For example, the information set for January 3, 2008 contains all stock returns up to an including those which occurred on January 3. It also includes everything else known at this time such as

[2] Residuals generated from an ARCH process have dependence in conditional variances but not mean.

[3] A measurable event is any event that can have probability assigned to it at time t. In general this includes any observed variable but can also include time t beliefs about latent (unobserved) variables such as volatility or the final revision of the current quarter's GDP.

interest rates, foreign exchange rates or the scores of recent football games. Many expectations will often be made conditional on the time-t information set, expressed $\mathrm{E}\left[y_{t+h}|\mathcal{F}_t\right]$, or in abbreviated form as $\mathrm{E}_t\left[y_{t+h}\right]$. The conditioning information set matters when taking expectations and $\mathrm{E}\left[y_{t+h}\right]$, $\mathrm{E}_t\left[y_{t+h}\right]$ and $\mathrm{E}_{t+h}\left[y_{t+h}\right]$ are *not* the same. Conditional variance is similarly defined, $\mathrm{V}\left[y_{t+h}|\mathcal{F}_t\right] = \mathrm{V}_t\left[y_{t+h}\right] = E_t\left[(y_{t+h} - \mathrm{E}_t\left[y_{t+h}\right])^2\right]$.

4.3 ARMA Models

Autoregressive moving average (ARMA) processes form the core of time-series analysis. The ARMA class can be decomposed into two smaller classes, autoregressive (AR) processes and moving average (MA) processes.

4.3.1 Moving Average Processes

The 1st order moving average, written MA(1), is the simplest non-degenerate time-series process,

$$y_t = \phi_0 + \theta_1 \epsilon_{t-1} + \epsilon_t$$

where ϕ_0 and θ_1 are parameters and ϵ_t a white noise series. This process stipulates that the current value of y_t depends on both a new shock and the previous shock. For example, if θ is negative, the current realization will "bounce back" from the previous shock.

Definition 4.6 (First Order Moving Average Process)**.** A first order Moving Average process (MA(1)) has dynamics which follow

$$y_t = \phi_0 + \theta_1 \epsilon_{t-1} + \epsilon_t \tag{4.6}$$

where ϵ_t is a white noise process with the additional property that $E_{t-1}\left[\epsilon_t\right] = 0$.

It is simple to derive both the conditional and unconditional means in this process. The conditional mean is

$$\begin{aligned} \mathrm{E}_{t-1}\left[y_t\right] &= \mathrm{E}_{t-1}\left[\phi_0 + \theta_1 \epsilon_{t-1} + \epsilon_t\right] \\ &= \phi_0 + \theta_1 \mathrm{E}_{t-1}\left[\epsilon_{t-1}\right] + \mathrm{E}_{t-1}\left[\epsilon_t\right] \\ &= \phi_0 + \theta_1 \epsilon_{t-1} + 0 \\ &= \phi_0 + \theta_1 \epsilon_{t-1} \end{aligned} \tag{4.7}$$

where $\mathrm{E}_{t-1}\left[\epsilon_t\right] = 0$ follows by assumption that the shock is unpredictable using the time-$t-1$ information set, and since ϵ_{t-1} is in the time-$t-1$ information set ($\epsilon_{t-1} \in \mathcal{F}_{t-1}$), it passes through the time-$t-1$ conditional expectation. The unconditional mean is

$$\begin{aligned} \mathrm{E}\left[y_t\right] &= \mathrm{E}\left[\phi_0 + \theta_1 \epsilon_{t-1} + \epsilon_t\right] \\ &= \phi_0 + \theta_1 \mathrm{E}\left[\epsilon_{t-1}\right] + \mathrm{E}\left[\epsilon_t\right] \\ &= \phi_0 + \theta_1 0 + 0 \\ &= \phi_0. \end{aligned} \tag{4.8}$$

Comparing these two results, the unconditional mean of y_t, $\mathrm{E}[y_t]$, is ϕ_0 while the conditional mean $\mathrm{E}_{t-1}[y_t] = \phi_0 + \theta_1\epsilon_{t-1}$. This difference reflects the persistence of the previous shock in the current period. The variances can be similarly derived,

$$\begin{aligned}
\mathrm{V}[y_t] &= \mathrm{E}\left[(\phi_0 + \theta_1\epsilon_{t-1} + \epsilon_t - \mathrm{E}[\phi_0 + \theta_1\epsilon_{t-1} + \epsilon_t])^2\right] \\
&= \mathrm{E}\left[(\phi_0 + \theta_1\epsilon_{t-1} + \epsilon_t - \phi_0)^2\right] \\
&= \mathrm{E}\left[(\theta_1\epsilon_{t-1} + \epsilon_t)^2\right] \\
&= \theta_1^2\mathrm{E}\left[\epsilon_{t-1}^2\right] + \mathrm{E}\left[\epsilon_t^2\right] + 2\theta_1\mathrm{E}[\epsilon_{t-1}\epsilon_t] \\
&= \sigma^2\theta_1^2 + \sigma^2 + 0 \\
&= \sigma^2\left(1 + \theta_1^2\right)
\end{aligned} \tag{4.9}$$

where $\mathrm{E}[\epsilon_{t-1}\epsilon_t]$ follows from the white noise assumption. The conditional variance is

$$\begin{aligned}
\mathrm{V}_{t-1}[y_t] &= \mathrm{E}_{t-1}\left[(\phi_0 + \theta_1\epsilon_{t-1} + \epsilon_t - \mathrm{E}_{t-1}[\phi_0 + \theta_1\epsilon_{t-1} + \epsilon_t])^2\right] \\
&= \mathrm{E}_{t-1}\left[(\phi_0 + \theta_1\epsilon_{t-1} + \epsilon_t - \phi_0 - \theta_1\epsilon_{t-1})^2\right] \\
&= \mathrm{E}_{t-1}\left[\epsilon_t^2\right] \\
&= \sigma_t^2
\end{aligned} \tag{4.10}$$

where σ_t^2 is the conditional variance of ϵ_t. White noise does not have to be homoskedastic, although if ϵ_t is homoskedastic then $\mathrm{V}_{t-1}[y_t] = \mathrm{E}\left[\sigma_t^2\right] = \sigma^2$. Like the mean, the unconditional variance and the conditional variance are different. The unconditional variance is unambiguously larger than the average conditional variance which reflects the extra variability introduced by the moving average term.

Finally, the autocovariance can be derived

$$\begin{aligned}
\mathrm{E}[(y_t - \mathrm{E}[y_t])(y_{t-1} - \mathrm{E}[y_{t-1}])] &= \mathrm{E}[(\phi_0 + \theta_1\epsilon_{t-1} + \epsilon_t - \phi_0)(\phi_0 + \theta_1\epsilon_{t-2} + \epsilon_{t-1} - \phi_0)] \\
&= \mathrm{E}\left[\theta_1\epsilon_{t-1}^2 + \theta_1\epsilon_t\epsilon_{t-2} + \epsilon_t\epsilon_{t-1} + \theta_1^2\epsilon_{t-1}\epsilon_{t-2}\right] \\
&= \theta_1\mathrm{E}\left[\epsilon_{t-1}^2\right] + \theta_1\mathrm{E}[\epsilon_t\epsilon_{t-2}] + \mathrm{E}[\epsilon_t\epsilon_{t-1}] + \theta_1^2\mathrm{E}[\epsilon_{t-1}\epsilon_{t-2}] \\
&= \theta_1\sigma^2 + 0 + 0 + 0 \\
&= \theta_1\sigma^2
\end{aligned} \tag{4.11}$$

$$\begin{aligned}
\mathrm{E}[(y_t - \mathrm{E}[y_t])(y_{t-2} - \mathrm{E}[y_{t-2}])] &= \mathrm{E}[(\phi_0 + \theta_1\epsilon_{t-1} + \epsilon_t - \phi_0)(\phi_0 + \theta_1\epsilon_{t-3} + \epsilon_{t-2} - \phi_0)] \\
&= \mathrm{E}[(\theta_1\epsilon_{t-1} + \epsilon_t)(\theta_1\epsilon_{t-3} + \epsilon_{t-2})] \\
&= \mathrm{E}\left[\theta_1\epsilon_{t-1}\epsilon_{t-2} + \theta_1\epsilon_{t-3}\epsilon_t + \epsilon_t\epsilon_{t-2} + \theta_1^2\epsilon_{t-1}\epsilon_{t-3}\right] \\
&= \theta_1\mathrm{E}[\epsilon_{t-1}\epsilon_{t-2}] + \theta_1\mathrm{E}[\epsilon_{t-3}\epsilon_t] + \mathrm{E}[\epsilon_t\epsilon_{t-2}] + \theta_1^2\mathrm{E}[\epsilon_{t-1}\epsilon_{t-3}] \\
&= 0 + 0 + 0 + 0 \\
&= 0
\end{aligned} \tag{4.12}$$

By inspection of eq. (4.12) it follows that $\gamma_s = \mathrm{E}[(y_t - \mathrm{E}[y_t])(y_{t-s} - \mathrm{E}[y_{t-s}])] = 0$ for $s \geq 2$.

The MA(1) can be generalized into the class of MA(Q) processes by including additional lagged errors.

Definition 4.7 (Moving Average Process of Order Q)**.** A Moving Average process of order Q, abbreviated MA(Q), has dynamics which follow

$$y_t = \phi_0 + \sum_{q=1}^{Q} \theta_q \epsilon_{t-q} + \epsilon_t \tag{4.13}$$

where ϵ_t is white noise series with the additional property that $E_{t-1}[\epsilon_t] = 0$.

The following properties hold in higher order moving averages:

- $\mathrm{E}[y_t] = \phi_0$
- $\mathrm{V}[y_t] = (1 + \sum_{q=1}^{Q} \theta_q^2)\sigma^2$
- $\mathrm{E}[(y_t - \mathrm{E}[y_t])(y_{t-s} - \mathrm{E}[y_{t-s}])] = \sigma^2 \sum_{i=0}^{Q-s} \theta_i \theta_{i+s}$ for $s \leq Q$ where $\theta_0 = 1$.
- $\mathrm{E}[(y_t - \mathrm{E}[y_t])(y_{t-s} - \mathrm{E}[y_{t-s}])] = 0$ for $s > Q$

4.3.2 Autoregressive Processes

The other subclass of ARMA processes is the autoregressive process.

Definition 4.8 (First Order Autoregressive Process)**.** A first order autoregressive process, abbreviated AR(1), has dynamics which follow

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t \tag{4.14}$$

where ϵ_t is a white noise process with the additional property that $E_{t-1}[\epsilon_t] = 0$.

Unlike the MA(1) process, y appears on both sides of the equation. However, this is only a convenience and the process can be recursively substituted to provide an expression that depends only on the errors, ϵ_t and an initial condition.

$$\begin{aligned}
y_t &= \phi_0 + \phi_1 y_{t-1} + \epsilon_t \\
y_t &= \phi_0 + \phi_1 (\phi_0 + \phi_1 y_{t-2} + \epsilon_{t-1}) + \epsilon_t \\
y_t &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 y_{t-2} + \epsilon_t + \phi_1 \epsilon_{t-1} \\
y_t &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 (\phi_0 + \phi_1 y_{t-3} + \epsilon_{t-2}) + \epsilon_t + \phi_1 \epsilon_{t-1} \\
y_t &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 \phi_0 + \phi_1^3 y_{t-3} + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_1^2 \epsilon_{t-2} \\
&\vdots \qquad\qquad \vdots \\
y_t &= \sum_{i=0}^{t-1} \phi_1^i \phi_0 + \sum_{i=0}^{t-1} \phi_1^i \epsilon_{t-i} + \phi_1^t y_0
\end{aligned}$$

Using backward substitution, an AR(1) can be expressed as an MA(t). In many cases the initial condition is unimportant and the AR process can be assumed to have begun long ago in the past.

As long as $|\phi_1| < 1$, $\lim_{t\to\infty} \phi^t y_0 \to 0$ and the effect of an initial condition will be small. Using the "infinite history" version of an AR(1), and assuming $|\phi_1| < 1$, the solution simplifies to

$$\begin{aligned} y_t &= \phi_0 + \phi_1 y_{t-1} + \epsilon_t \\ y_t &= \sum_{i=0}^{\infty} \phi_1^i \phi_0 + \sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i} \\ y_t &= \frac{\phi_0}{1-\phi_1} + \sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i} \end{aligned} \tag{4.15}$$

where the identity $\sum_{i=0}^{\infty} \phi_1^i = (1-\phi_1)^{-1}$ is used in the final solution. This expression of an AR process is known as an MA(∞) representation and it is useful for deriving standard properties.

The unconditional mean of an AR(1) is

$$\begin{aligned} \mathrm{E}[y_t] &= \mathrm{E}\left[\frac{\phi_0}{1-\phi_1} + \sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i}\right] \\ &= \frac{\phi_0}{1-\phi_1} + \sum_{i=0}^{\infty} \phi_1^i \mathrm{E}[\epsilon_{t-i}] \\ &= \frac{\phi_0}{1-\phi_1} + \sum_{i=0}^{\infty} \phi_1^i 0 \\ &= \frac{\phi_0}{1-\phi_1}. \end{aligned} \tag{4.16}$$

The unconditional mean can be alternatively derived noting that, as long as $\{y_t\}$ is covariance stationary, that $\mathrm{E}[y_t] = \mathrm{E}[y_{t-1}] = \mu$, and so

$$\begin{aligned} \mathrm{E}[y_t] &= \mathrm{E}[\phi_0 + \phi_1 y_{t-1} + \epsilon_{t-1}] \\ \mathrm{E}[y_t] &= \phi_0 + \phi_1 \mathrm{E}[y_{t-1}] + \mathrm{E}[\epsilon_{t-1}] \\ \mu &= \phi_0 + \phi_1 \mu + 0 \\ \mu - \phi_1 \mu &= \phi_0 \\ \mu(1-\phi_1) &= \phi_0 \\ \mathrm{E}[y_t] &= \frac{\phi_0}{1-\phi_1} \end{aligned} \tag{4.17}$$

The $\mathcal{F}_{t-1}$-conditional expectation is

$$\begin{aligned} \mathrm{E}_{t-1}[y_t] &= \mathrm{E}_{t-1}[\phi_0 + \phi_1 y_{t-1} + \epsilon_t] \\ &= \phi_0 + \phi_1 \mathrm{E}_{t-1}[y_{t-1}] + \mathrm{E}_{t-1}[\epsilon_t] \\ &= \phi_0 + \phi_1 y_{t-1} + 0 \\ &= \phi_0 + \phi_1 y_{t-1} \end{aligned} \tag{4.18}$$

since $y_{t-1} \in \mathcal{F}_{t-1}$. The unconditional and conditional variances are

$$
\begin{aligned}
\mathrm{V}[y_t] &= \mathrm{E}\left[(y_t - \mathrm{E}[y_t])^2\right] \\
&= \mathrm{E}\left[\left(\frac{\phi_0}{1-\phi_1} + \sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i} - \frac{\phi_0}{1-\phi_1}\right)^2\right] \\
&= \mathrm{E}\left[\left(\sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i}\right)^2\right] \\
&= \mathrm{E}\left[\sum_{i=0}^{\infty} \phi_1^{2i} \epsilon_{t-i}^2 + \sum_{i=0}^{\infty} \sum_{j=0, i\neq j}^{\infty} \phi_1^{i+j} \epsilon_{t-i}\epsilon_{t-j}\right] \\
&= \mathrm{E}\left[\sum_{i=0}^{\infty} \phi_1^{2i} \epsilon_{t-i}^2\right] + \mathrm{E}\left[\sum_{i=0}^{\infty} \sum_{j=0, i\neq j}^{\infty} \phi_1^{i+j} \epsilon_{t-i}\epsilon_{t-j}\right] \\
&= \sum_{i=0}^{\infty} \phi_1^{2i} \mathrm{E}\left[\epsilon_{t-i}^2\right] + \sum_{i=0}^{\infty} \sum_{j=0, i\neq j}^{\infty} \phi_1^{i+j} \mathrm{E}\left[\epsilon_{t-i}\epsilon_{t-j}\right] \\
&= \sum_{i=0}^{\infty} \phi_1^{2i} \sigma^2 + \sum_{i=0}^{\infty} \sum_{j=0, i\neq j}^{\infty} \phi_1^{i+j} 0 \\
&= \frac{\sigma^2}{1-\phi_1^2}
\end{aligned}
\tag{4.19}
$$

where the expression for the unconditional variance uses the identity that $\sum_{i=0}^{\infty} \phi_1^{2i} = \frac{1}{1-\phi_1^2}$ and $\mathrm{E}[\epsilon_{t-i}\epsilon_{t-j}] = 0$ follows from the white noise assumption. Again, assuming covariance stationarity and so $\mathrm{V}[y_t] = \mathrm{V}[y_{t-1}]$, the variance can be directly computed,

$$
\begin{aligned}
\mathrm{V}[y_t] &= \mathrm{V}[\phi_0 + \phi_1 y_{t-1} + \epsilon_t] \\
\mathrm{V}[y_t] &= \mathrm{V}[\phi_0] + \mathrm{V}[\phi_1 y_{t-1}] + \mathrm{V}[\epsilon_t] + 2\mathrm{Cov}[\phi_1 y_{t-1}, \epsilon_t] \\
\mathrm{V}[y_t] &= 0 + \phi_1^2 \mathrm{V}[y_{t-1}] + \sigma^2 + 2 \cdot 0 \\
\mathrm{V}[y_t] &= \phi_1^2 \mathrm{V}[y_t] + \sigma^2 \\
\mathrm{V}[y_t] - \phi_1^2 \mathrm{V}[y_t] &= \sigma^2 \\
\mathrm{V}[y_t](1-\phi_1^2) &= \sigma^2 \\
\mathrm{V}[y_t] &= \frac{\sigma^2}{1-\phi_1^2}
\end{aligned}
\tag{4.20}
$$

where $\mathrm{Cov}[y_{t-1}, \epsilon_t] = 0$ follows from the white noise assumption since y_{t-1} is a function of $\epsilon_{t-1}, \epsilon_{t-2}, \ldots$. The conditional variance is

$$\begin{aligned} \mathrm{V}_{t-1}[y_t] &= \mathrm{E}_{t-1}\left[(\phi_1 y_{t-1} + \epsilon_t - \phi_1 y_{t-1})^2\right] \\ &= \mathrm{E}_{t-1}\left[\epsilon_t^2\right] \\ &= \sigma_t^2 \end{aligned} \tag{4.21}$$

Again, the unconditional variance is uniformly larger than the average conditional variance ($\mathrm{E}\left[\sigma_t^2\right] = \sigma^2$) and the variance explodes as $|\phi_1|$ approaches 1 or -1. Finally, the autocovariances can be derived,

$$\mathrm{E}\left[(y_t - \mathrm{E}[y_t])(y_{t-s} - \mathrm{E}[y_{t-s}])\right] = \mathrm{E}\left[\left(\frac{\phi_0}{1-\phi_1} + \sum_{i=0}^{\infty}\phi_1^i\epsilon_{t-i} - \frac{\phi_0}{1-\phi_1}\right)\right. \tag{4.22}$$

$$\left.\times\left(\frac{\phi_0}{1-\phi_1} + \sum_{i=0}^{\infty}\phi_1^i\epsilon_{t-s-i} - \frac{\phi_0}{1-\phi_1}\right)\right] \tag{4.23}$$

$$\begin{aligned} &= \mathrm{E}\left[\left(\sum_{i=0}^{\infty}\phi_1^i\epsilon_{t-i}\right)\left(\sum_{i=0}^{\infty}\phi_1^i\epsilon_{t-s-i}\right)\right] \\ &= \mathrm{E}\left[\left(\sum_{i=0}^{s-1}\phi_1^i\epsilon_{t-i} + \sum_{i=s}^{\infty}\phi_1^i\epsilon_{t-i}\right)\left(\sum_{i=0}^{\infty}\phi_1^i\epsilon_{t-s-i}\right)\right] \\ &= \mathrm{E}\left[\left(\sum_{i=0}^{s-1}\phi_1^i\epsilon_{t-i} + \sum_{i=0}^{\infty}\phi_1^s\phi_1^i\epsilon_{t-s-i}\right)\left(\sum_{i=0}^{\infty}\phi_1^i\epsilon_{t-s-i}\right)\right] \\ &= \mathrm{E}\left[\left(\sum_{i=0}^{s-1}\phi_1^i\epsilon_{t-i}\right)\left(\sum_{i=0}^{\infty}\phi_1^i\epsilon_{t-s-i}\right)\right. \\ &\quad \left.+ \phi_1^s\left(\sum_{i=0}^{\infty}\phi_1^i\epsilon_{t-s-i}\right)\left(\sum_{i=0}^{\infty}\phi_1^i\epsilon_{t-s-i}\right)\right] \end{aligned} \tag{4.24}$$

$$\begin{aligned} &= \mathrm{E}\left[\left(\sum_{i=0}^{s-1}\phi_1^i\epsilon_{t-i}\right)\left(\sum_{i=0}^{\infty}\phi_1^i\epsilon_{t-s-i}\right)\right] \\ &\quad + \mathrm{E}\left[\phi_1^s\left(\sum_{i=0}^{\infty}\phi_1^i\epsilon_{t-s-i}\right)\left(\sum_{i=0}^{\infty}\phi_1^i\epsilon_{t-s-i}\right)\right] \end{aligned} \tag{4.25}$$

$$\begin{aligned} &= 0 + \phi_1^s\mathrm{E}\left[\left(\sum_{i=0}^{\infty}\phi_1^i\epsilon_{t-s-i}\right)\left(\sum_{i=0}^{\infty}\phi_1^i\epsilon_{t-s-i}\right)\right] \\ &= 0 + \phi_1^s\mathrm{V}[y_{t-s}] \\ &= \phi_1^s\frac{\sigma^2}{1-\phi_1^2} \end{aligned}$$

An alternative approach to deriving the autocovariance is to note that $y_t - \mu = \sum_{i=0}^{s-i}\phi_1^i\epsilon_{t-i} + \phi^s(y_{t-s} - \mu)$ where $\mu = \mathrm{E}[y_t] = \mathrm{E}[y_{t-s}]$. Using this identify, the autocovariance can be derived

$$\begin{aligned}
\mathrm{E}\left[(y_t - \mathrm{E}[y_t])(y_{t-s} - \mathrm{E}[y_{t-s}])\right] &= \mathrm{E}\left[\left(\sum_{i=0}^{s-i} \phi_1^i \epsilon_{t-i} + \phi^s(y_{t-s} - \mu)\right)(y_{t-s} - \mu)\right] \\
&= \mathrm{E}\left[\left(\sum_{i=0}^{s-i} \phi_1^i \epsilon_{t-i}\right)(y_{t-s} - \mu) + (\phi^s(y_{t-s} - \mu)(y_{t-s} - \mu))\right] \\
&= \mathrm{E}\left[\left(\sum_{i=0}^{s-i} \phi_1^i \epsilon_{t-i}\right)(y_{t-s} - \mu)\right] + \mathrm{E}\left[(\phi^s(y_{t-s} - \mu)(y_{t-s} - \mu))\right] \\
&= 0 + \phi^s \mathrm{E}\left[((y_{t-s} - \mu)(y_{t-s} - \mu))\right] \\
&= \phi^s \mathrm{V}\left[y_{t-s}\right] \\
&= \phi_1^s \frac{\sigma^2}{1 - \phi_1^2}
\end{aligned} \tag{4.26}$$

where the white noise assumption is used to ensure that $\mathrm{E}\left[\epsilon_{t-u}(y_{t-s} - \mu)\right] = 0$ when $u > s$.

The AR(1) can be extended to the AR(P) class by including additional lags of y_t.

Definition 4.9 (Autoregressive Process of Order P)**.** An Autoregressive process of order P (AR(P)) has dynamics which follow

$$y_t = \phi_0 + \sum_{p=1}^{P} \phi_p y_{t-p} + \epsilon_t \tag{4.27}$$

where ϵ_t is white noise series with the additional property that $E_{t-1}\left[\epsilon_t\right] = 0$.

Some of the more useful properties of general AR process are:

- $\mathrm{E}[y_t] = \frac{\phi_0}{1-\sum_{p=1}^{P}\phi_p}$
- $\mathrm{V}[y_t] = \frac{\sigma^2}{1-\sum_{p=1}^{P}\phi_p\rho_p}$ where ρ_p is the p$^{\text{th}}$ autocorrelation.
- $\mathrm{V}[y_t]$ is infinite if $\sum_{p=1}^{P}\phi_p \geq 1$
- $\mathrm{E}\left[(y_t - \mathrm{E}[y_t])(y_{t-s} - \mathrm{E}[y_{t-s}])\right] \neq 0$ for any s (in general, although certain parameterizations may produce some 0 autocovariances).

These four properties point to some important regularities of AR processes. First, the mean is only finite if $\sum_{p=1}^{P}\phi_p < 1$. Second, the autocovariances are (generally) not zero, unlike those of an MA processes ($\gamma_s = 0$ for $|s| > Q$). This difference in the behavior of the autocovariances plays an important role in model building. Explicit expressions for the variance and autocovariance of higher order AR processes can be found in appendix 4.A.

4.3.3 Autoregressive-Moving Average Processes

Putting these two processes together yields the complete class of ARMA processes.

Definition 4.10 (Autoregressive-Moving Average Process)**.** An Autoregressive Moving Average process with orders P and Q (ARMA(P, Q)) has dynamics which follow

$$y_t = \phi_0 + \sum_{p=1}^{P} \phi_p y_{t-p} + \sum_{q=1}^{Q} \theta_q \epsilon_{t-q} + \epsilon_t \tag{4.28}$$

where ϵ_t is a white noise process with the additional property that $E_{t-1}[\epsilon_t] = 0$.

Again, consider the simplest ARMA(1,1) process that includes a constant term,

$$y_t = \phi_0 + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t$$

To derive the properties of this model it is useful to convert the ARMA(1,1) into its infinite lag representation using recursive substitution,

$$\begin{aligned}
y_t &= \phi_0 + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t \\
y_t &= \phi_0 + \phi_1 (\phi_0 + \phi y_{t-2} + \theta_1 \epsilon_{t-2} + \epsilon_{t-1}) + \theta_1 \epsilon_{t-1} + \epsilon_t \\
y_t &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 y_{t-2} + \phi_1 \theta_1 \epsilon_{t-2} + \phi_1 \epsilon_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t \\
y_t &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 (\phi_0 + \phi y_{t-3} + \theta_1 \epsilon_{t-3} + \epsilon_{t-2}) + \phi_1 \theta_1 \epsilon_{t-2} + \phi_1 \epsilon_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t \\
y_t &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 \phi_0 + \phi_1^3 y_{t-3} + \phi_1^2 \theta_1 \epsilon_{t-3} + \phi_1^2 \epsilon_{t-2} + \phi_1 \theta_1 \epsilon_{t-2} + \phi_1 \epsilon_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t \\
&\vdots \qquad\qquad \vdots \\
y_t &= \sum_{i=0}^{\infty} \phi_1^i \phi_0 + \epsilon_t + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \epsilon_{t-i-1} \\
y_t &= \frac{\phi_0}{1-\phi_1} + \epsilon_t + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \epsilon_{t-i-1}.
\end{aligned} \tag{4.29}$$

Using the infinite lag representation, the unconditional and conditional means can be computed,

$$\begin{aligned}
\mathrm{E}[y_t] &= \mathrm{E}\left[\frac{\phi_0}{1-\phi_1} + \epsilon_t + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \epsilon_{t-i-1} \right] \\
&= \frac{\phi_0}{1-\phi_1} + \mathrm{E}[\epsilon_t] + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \mathrm{E}[\epsilon_{t-i-1}] \\
&= \frac{\phi_0}{1-\phi_1} + 0 + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) 0 \\
&= \frac{\phi_0}{1-\phi_1}
\end{aligned} \tag{4.30}$$

and

$$\begin{aligned} \mathrm{E}_{t-1}[y_t] &= \mathrm{E}_{t-1}[\phi_0 + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t] \\ &= \phi_0 + \phi_1 \mathrm{E}_{t-1}[y_{t-1}] + \theta_1 \mathrm{E}_{t-1}[\epsilon_{t-1}] + \mathrm{E}_{t-1}[\epsilon_t] \\ &= \phi_0 + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + 0 \\ &= \phi_0 + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} \end{aligned} \tag{4.31}$$

Since y_{t-1} and ϵ_{t-1} are in the time-$t-1$ information set, these variables pass through the conditional expectation. The unconditional variance can be tediously derived (see appendix 4.A.3 for the complete derivation)

$$\mathrm{V}[y_t] = \sigma^2 \left(\frac{1 + 2\phi_1\theta_1 + \theta_1^2}{1 - \phi_1^2} \right) \tag{4.32}$$

The conditional variance is identical to that in the AR(1) or MA(1), $\mathrm{V}_{t-1}[y_t] = \sigma_t^2$, and, if ϵ_t is homoskedastic, $\mathrm{V}_{t-1}[y_t] = \sigma^2$.

The unconditional mean of an ARMA is the same as an AR since the moving average terms, which are all mean zero, make no contribution. The variance is more complicated, and it may be larger or smaller that an AR(1) with the same autoregressive parameter (ϕ_1). The variance will only be smaller if ϕ_1 and θ_1 have opposite signs *and* $2\phi_1\theta_1 < \theta_1^2$. Deriving the autocovariance is straightforward but tedious and is presented in appendix 4.A.

4.4 Difference Equations

Before turning to the analysis of the stationarity conditions for ARMA processes, it is useful to develop an understanding of the stability conditions in a setting without random shocks.

Definition 4.11 (Linear Difference Equation)**.** An equation of the form

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_P y_{t-P} + x_t. \tag{4.33}$$

is known as a P^{th} order linear difference equation where the series $\{x_t\}$ is known as the driving process.

Linear difference equation nest ARMA processes which can be seen by setting x_t equal to the shock plus the moving average component of the ARMA process,

$$x_t = \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_Q \epsilon_{t-Q} + \epsilon_t.$$

Stability conditions depend crucially on the solution to the linear difference equation.

Definition 4.12 (Solution)**.** A solution to a linear difference equation expresses the linear difference equation

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_P y_{t-P} + x_t. \tag{4.34}$$

as a function of only $\{x_i\}_{t=1}^{t}$, a constant and, when y_t has finite history, an initial value y_0.

Consider a first order linear difference equation

$$y_t = \phi_0 + \phi_1 y_{t-1} + x_t.$$

Starting from an initial value y_0,

$$\begin{aligned}
y_1 &= \phi_0 + \phi_1 y_0 + x_1 \\
y_2 &= \phi_0 + \phi_1(\phi_0 + \phi_1 y_0 + x_1) + x_2 \\
&= \phi_0 + \phi_1\phi_0 + \phi_1^2 y_0 + x_2 + \phi_1 x_1 \\
y_3 &= \phi_0 + \phi_1 y_2 + x_2 \\
&= \phi_0 + \phi_1(\phi_0 + \phi_1\phi_0 + \phi_1^2 y_0 + \phi_1 x_1 + x_2) + x_2 \\
&= \phi_0 + \phi_1\phi_0 + \phi_1^2\phi_0 + \phi_1^3 y_0 + x_3 + \phi_1 x_2 + \phi_1^2 x_1
\end{aligned}$$

Continuing these iterations, a pattern emerges:

$$y_t = \phi_1^t y_0 + \sum_{i=0}^{t-1} \phi_1^i \phi_0 + \sum_{i=0}^{t-1} \phi_1^i x_{t-i} \tag{4.35}$$

This is a solution since it expresses y_t as a function of only $\{x_t\}$, y_0 and constants. When no initial condition is given (or the series is assumed to be infinite), the solution can be found by solving backward

$$\begin{aligned}
y_t &= \phi_0 + \phi_1 y_{t-1} + x_t \\
y_{t-1} &= \phi_0 + \phi_1 y_{t-2} + x_{t-1} \Rightarrow \\
y_t &= \phi_0 + \phi_1(\phi_0 + \phi_1 y_{t-2} + x_{t-1}) + x_t \\
&= \phi_0 + \phi_1\phi_0 + \phi_1^2 y_{t-2} + x_t + \phi_1 x_{t-1} \\
y_{t-2} &= \phi_0 + \phi_1 y_{t-3} + x_{t-2} \Rightarrow \\
y_t &= \phi_0 + \phi_1\phi_0 + \phi_1^2(\phi_0 + \phi_1 y_{t-3} + x_{t-2}) + x_t + \phi_1 x_{t-1} \\
&= \phi_0 + \phi_1\phi_0 + \phi_1^2\phi_0 + \phi_1^3 y_{t-3} + x_t + \phi_1 x_{t-1} + \phi_1^2 x_{t-2}
\end{aligned}$$

which leads to the approximate solution

$$y_t = \sum_{i=0}^{s-1} \phi_1^i \phi_0 + \sum_{i=0}^{s-1} \phi_1^i x_{t-i} + \phi_1^s y_{t-s}.$$

To understand the behavior of this solution, it is necessary to take limits. If $|\phi_1| < 1$, $\lim_{s\to\infty} \phi_1^s y_{t-s}$ goes to zero (as long as y_{t-s} is bounded) and the solution simplifies to

$$y_t = \phi_0 \sum_{i=0}^{\infty} \phi_1^i + \sum_{i=0}^{\infty} \phi_1^i x_{t-i}. \tag{4.36}$$

Noting that, as long as $|\phi_1| < 1$, $\sum_{i=0}^{\infty} \phi_1^i = 1/(1-\phi_1)$,

$$y_t = \frac{\phi_0}{1-\phi_1} + \sum_{i=0}^{\infty} \phi_1^i x_{t-i} \tag{4.37}$$

is the solution to this problem with an infinite history. The solution concept is important because it clarifies the relationship between observations in the distant past and the current observation, and if $\lim_{s\to\infty} \phi_1^s y_{t-s}$ does not converge to zero then observations arbitrarily far in the past have an influence on the value of y today.

When $|\phi_1| > 1$ then this system is said to be *nonconvergent* since ϕ_1^t diverges as t grows large and values in the past are not only important, they will dominate when determining the current value. In the special case where $\phi_1 = 1$,

$$y_t = \phi_0 t + \sum_{i=0}^{\infty} x_{t-i},$$

which is a random walk when $\{x_t\}$ is a white noise process, and the influence of a single x_t never diminishes. Direct substitution can be used to find the solution of higher order linear difference equations at the cost of more tedium. A simpler alternative is to focus on the core component of a linear difference equation, the linear homogeneous equation.

4.4.1 Homogeneous Difference Equations

When the number of lags grows large (3 or greater), solving linear difference equations by substitution is tedious. The key to understanding linear difference equations is the study of the homogeneous portion of the equation. In the general linear difference equation,

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_P y_{t-P} + x_t$$

the homogenous portion is defined as the terms involving only y,

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_P y_{t-P}. \tag{4.38}$$

The intuition behind studying this portion of the system is that given the sequence of $\{x_t\}$, all of the dynamics and the stability of the system are determined by the relationship between contemporaneous y_t and its lagged values which allows the determination of the parameter values where the system is stable. Again, consider the homogeneous portions of the simple 1$^{\text{st}}$order system,

$$y_t = \phi_1 y_{t-1} + x_t$$

which has the homogeneous portion

$$y_t = \phi_1 y_{t-1}.$$

To find solutions to this equation, one can try trial and error: one obvious solution is 0 since $0 = \phi \cdot 0$. It is easy to show that

$$y_t = \phi_1^t y_0$$

is also a solution by examining the solution to the linear difference equation in eq. (4.35). But then so is any solution of the form $c\,\phi_1^t$ for an arbitrary constant c. How?

$$\begin{aligned} y_t &= c\,\phi_1^t \\ y_{t-1} &= c\,\phi_1^{t-1} \end{aligned}$$

and

$$y_t = \phi_1 y_{t-1}$$

Putting these two together shows that

$$\begin{aligned} y_t &= \phi_1 y_{t-1} \\ c\,\phi_1^t &= \phi_1 y_{t-1} \\ c\,\phi_1^t &= \phi_1 c\,\phi_1^{t-1} \\ c\,\phi_1^t &= c\,\phi_1^t \end{aligned}$$

and there are many solutions. However, from these it is possible to discern when the solution will converge to zero and when it will explode:

- If $|\phi_1| < 1$ the system converges to 0. If ϕ_1 is also negative, the solution oscillates, while if ϕ_1 is greater than 0, the solution decays exponentially.
- If $|\phi_1| > 1$ the system diverges, again oscillating if negative and growing exponentially if positive.
- If $\phi_1 = 1$, the system is stable and all values are solutions. For example $1 = 1 \cdot 1$, $2 = 1 \cdot 2$, etc.
- If $\phi_1 = -1$, the system is *metastable*. The values, in absolute terms, are unchanged, but it oscillates between + and -.

These categories will play important roles in examining the dynamics of larger equations since they determine how past shocks will affect current values of y_t. When the order is greater than 1, there is an easier approach to examining the stability of the system. Consider the second order linear difference system,

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + x_t$$

and again focus on the homogeneous portion,

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

This equation can be rewritten

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} = 0$$

so any solution of the form

$$\begin{aligned} c z^t - \phi_1 c z^{t-1} - \phi_2 c z^{t-2} &= 0 \\ c z^{t-2}\left(z^2 - \phi_1 z - \phi_2\right) &= 0 \end{aligned} \tag{4.39}$$

will solve this equation.[4] Dividing through by $c z^{t-2}$, this is equivalent to

$$z^2 - \phi_1 z - \phi_2 = 0 \tag{4.40}$$

and he solutions to this quadratic polynomial are given by the quadratic formula,

$$c_1, c_2 = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2} \tag{4.41}$$

The roots of the equation, c_1 and c_2, play the same role as ϕ_1 in the 1st order case.[5] If $|c_1| < 1$ and $|c_2| < 1$, the system is convergent. With two roots both smaller than 1 there are three interesting cases:

Case 1: Both roots are real and positive. In this case, the system will exponentially dampen and not oscillate.

Case 2: Both roots are imaginary (of the form $c + di$ where $i = \sqrt{-1}$) and distinct, or real and at least one negative. In this case, the absolute value of the roots (also called the modulus, defined as $\sqrt{c^2 + d^2}$ for an imaginary number $c + di$) is less than 1, and so the system will be convergent but oscillate.

Case 3: Real but the same. This occurs when $\phi_1^2 + 4\phi_2 = 0$. Since there is only one root, the system is convergent if it is less than 1 in absolute value, which require that $|\phi_1| < 2$.

If either are greater than 1 in absolute terms, the system is divergent.

4.4.2 Lag Operators

Before proceeding to higher order models, it is necessary to define the lag operator. Lag operations are a particularly useful tool in the analysis of time series and are nearly self-descriptive.[6]

Definition 4.13 (Lag Operator)**.** The lag operator is denoted L and is defined as the operator that has the following properties:

$$\begin{aligned} L y_t &= y_{t-1} \\ L^2 y_t &= y_{t-2} \\ L^i y_t &= y_{t-i} \\ L(L(y_t)) = L(y_{t-1}) &= y_{t-2} = L^2 y_t \\ (1 - L - L^2) y_t &= y_t - L y_t - L^2 y_t = y_t - y_{t-1} - y_{t-2} \end{aligned}$$

[4] The solution can only be defined up to a constant, c, since the right hand side is 0. Thus, multiplying both by a constant, the solution will still be valid.

[5] In the first order case, $y_t = \phi_1 y_{t-1}$, so $y_t - \phi_1 y_{t-1} = 0$. The solution has the property that $z^t - \phi_1 z^{t-1} = 0$ so $z - \phi_1 = 0$, which has the single solution $c = \phi_1$.

[6] In some texts, the lag operator is known as the backshift operator, and L is replaced with B.

The last equation above is particularly useful when studying autoregressive processes. One additional property of the lag operator is that the lag of a constant is just the constant, i.e. $Lc = c$.

4.4.3 Higher Order Linear Homogenous Equations

Stability analysis can be applied to higher order systems by forming the characteristic equation and finding the characteristic roots.

Definition 4.14 (Characteristic Equation)**.** Let y_t follow a P$^{\text{th}}$ order linear difference equation

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_P y_{t-P} + x_t \tag{4.42}$$

which can be rewritten as

$$\begin{aligned} y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \ldots - \phi_P y_{t-P} &= \phi_0 + x_t \\ (1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_P L^P) y_t &= \phi_0 + x_t \end{aligned} \tag{4.43}$$

The characteristic equation of this process is

$$z^P - \phi_1 z^{P-1} - \phi_2 z^{P-2} - \ldots - \phi_{P-1} z - \phi_P = 0 \tag{4.44}$$

The characteristic roots are the solutions to this equation and most econometric packages will return the roots of the characteristic polynomial when an ARMA model is estimated.

Definition 4.15 (Characteristic Root)**.** Let

$$z^P - \phi_1 z^{P-1} - \phi_2 z^{P-2} - \ldots - \phi_{P-1} z - \phi_P = 0 \tag{4.45}$$

be the characteristic polynomial associated with a P^{th} order linear difference equation. The P characteristic roots, $c_1, c_2, \ldots, c_P$ are defined as the solution to this polynomial

$$(z - c_1)(z - c_2)\ldots(z - c_P) = 0 \tag{4.46}$$

The conditions for stability are the same for higher order systems as they were for first and second order systems: all roots c_p, $p = 1, 2, \ldots, P$ must satisfy $|c_p| < 1$ (again, if complex, $|\cdot|$ means modulus). If any $|c_p| > 1$ the system is divergent. If one of more $|c_p| = 1$ and none are larger, the system will exhibit unit root (random walk) behavior.

These results are the key to understanding important properties of linear time-series models which turn out to be *stationary if the corresponding linear homogeneous system is convergent*, i.e. $|c_p| < 1$, $p = 1, 2, \ldots, P$.

4.4.4 Example: Characteristic Roots and Stability

Consider 6 linear difference equations, their characteristic equation, and the roots:

- $y_t = 0.9 y_{t-1} + x_t$
  - Characteristic Equation: z-0.9=0

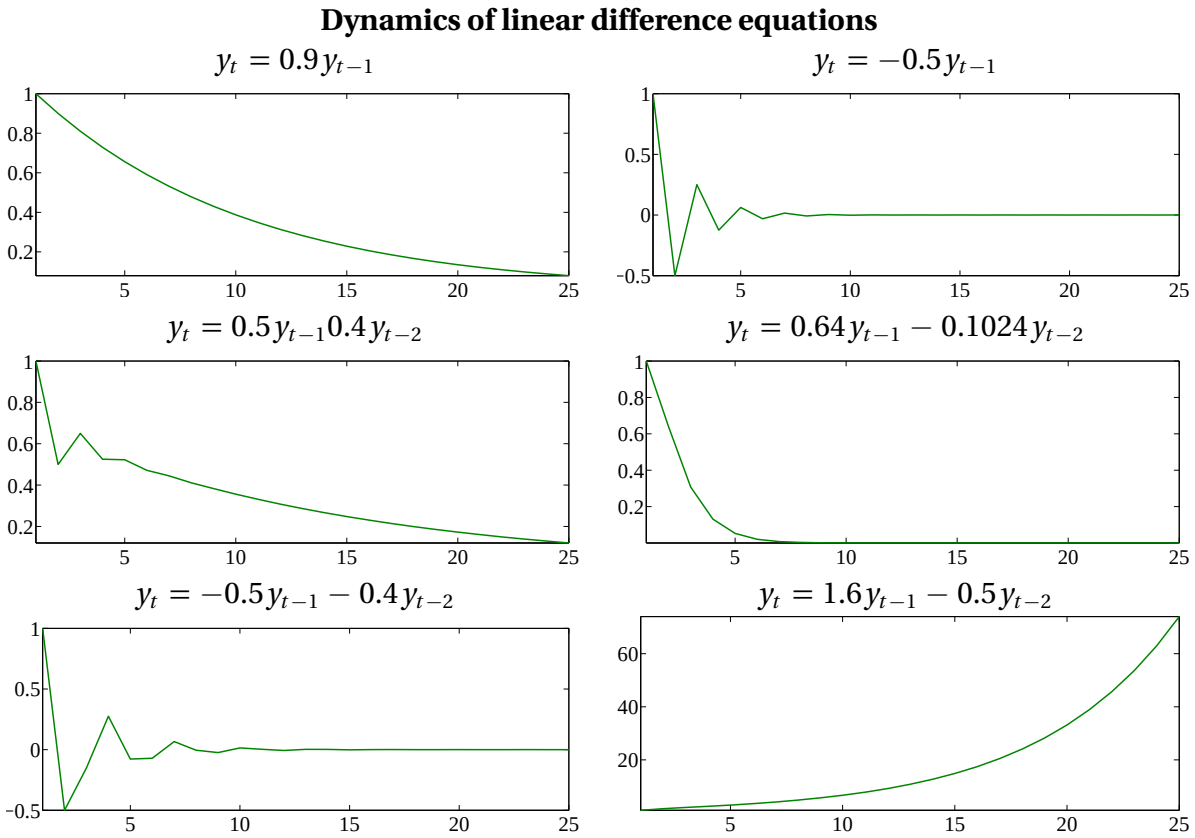

Figure 4.1: These six plots correspond to the dynamics of the six linear homogeneous systems described in the text. All processes received a unit shock at $t = 1$ ($x_1 = 1$) and no other shocks ($x_j = 0,\ j \neq 1$). Pay close attention to the roots of the characteristic polynomial and the behavior of the system (exponential decay, oscillation and/or explosion).

  - Characteristic Root: z=0.9

- $y_t = -0.5y_{t-1} + x_t$
  - Characteristic Equation: z+0.5=0
  - Characteristic Root: z=-0.5

- $y_t = 0.5y_{t-1} + 0.4y_{t-2} + x_t$
  - Characteristic Equation: $z^2 - 0.5z - 0.4 = 0$
  - Characteristic Roots: $z = 0.93, -.43$

- $y_t = 0.64y_{t-1} - 0.1024y_{t-2} + x_t$
  - Characteristic Equation: $z^2 - 0.64z + 0.1024 = 0$
  - Characteristic Roots: $z = 0.32, 0.32$ (identical)

- $y_t = -0.5y_{t-1} - 0.4y_{t-2} + x_t$
  - Characteristic Equation: $z^2 + 0.5z + 0.4 = 0$
  - Characteristic Roots (Modulus): $z = -0.25 + 0.58i(0.63), -0.25 - 0.58i(0.63)$
- $y_t = 1.6y_{t-1} - 0.5y_{t-2} + x_t$
  - Characteristic Equation: $z^2 - 1.6z + 0.5 = 0$
  - Characteristic Roots: $z = 1.17, 0.42$

The plots in figure 4.1 show the effect of a unit (1) shock at $t = 1$ to the 6 linear difference systems above (all other shocks are 0). The value of the root makes a dramatic difference in the observed behavior of the series.

4.4.5 Stationarity of ARMA models

Stationarity conditions for ARMA processes can be determined using the results for the convergence of linear difference equations. First, note that any ARMA process can be written using a lag polynomial

$$\begin{aligned} y_t &= \phi_0 + \phi_1 y_{t-1} + \ldots + \phi_P y_{t-P} + \theta_1 \epsilon_{t-1} + \ldots + \theta_Q \epsilon_{t-Q} + \epsilon_t \\ y_t - \phi_1 y_{t-1} - \ldots - \phi_P y_{t-P} &= \phi_0 + \theta_1 \epsilon_{t-1} + \ldots + \theta_Q \epsilon_{t-Q} + \epsilon_t \\ (1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_P L^P) y_t &= \phi_0 + (1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_Q L^Q) \epsilon_t \end{aligned}$$

This is a linear difference equation, and the stability conditions depend on the roots of the characteristic polynomial

$$z^P - \phi_1 z^{P-1} - \phi_2 z^{P-2} - \ldots - \phi_{P-1} z - \phi_P$$

An ARMA process driven by a white noise shock will be covariance stationary as long as the characteristic roots are less than one in modulus. In the simple AR(1) case, this corresponds to $|z_1| < 1$. In the AR(2) case, the region is triangular with a curved bottom and corresponds to the points $(z_1, z_2) = (-2, -1), (1, 0), (2, -2)$ (see figure 4.2). For higher order models, stability must be checked by numerically solving the characteristic equation.

The other particularly interesting point is that *all* MA processes driven by covariance stationary shocks are stationary since the homogeneous portions of an MA process has no root and thus cannot diverge.

4.5 Data and Initial Estimates

Two series will be used throughout the stationary time-series analysis section: returns on the value weighted market and the spread between the average interest rates on portfolios of Aaa-rated and Baa-rated corporate bonds, commonly known as the default spread or default premium. The VWM returns were taken from CRSP and are available from January 1927 through

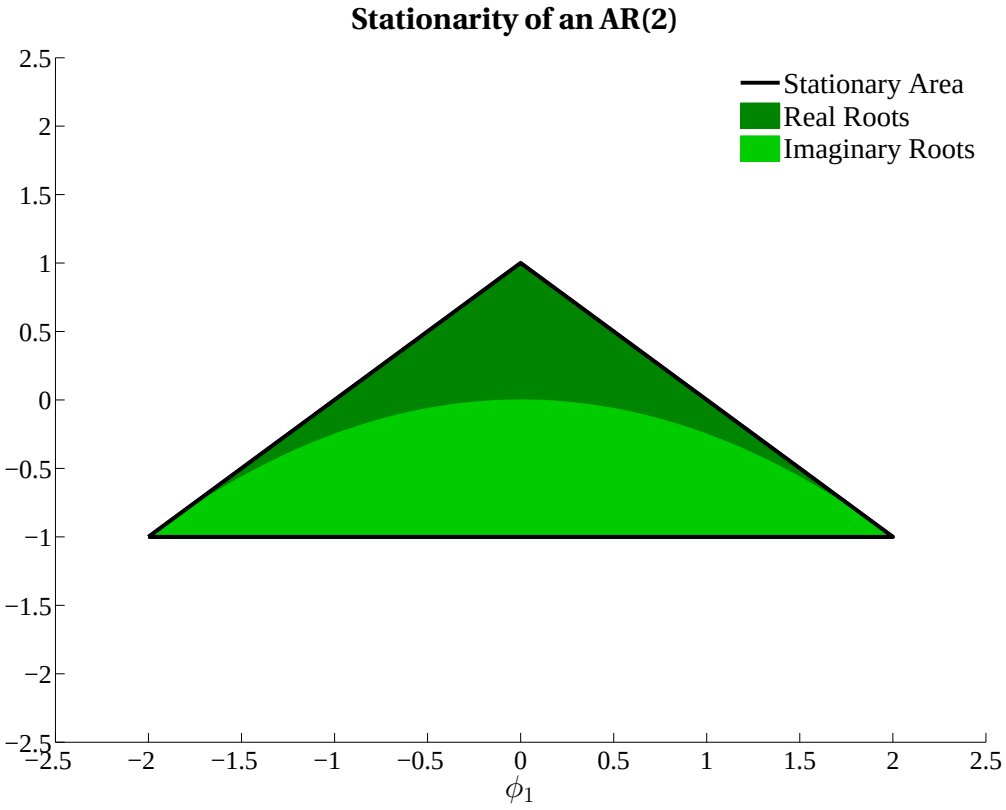

Figure 4.2: The triangular region corresponds to the values of the parameters in the AR(2) $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$. The dark region corresponds to real roots and the light region corresponds to imaginary roots.

July 2008 and the bond yields are available from Moody's via FRED II and are available from January 1919 until July 2008. Both series are monthly.

Figure 4.3 contains plots of the two series. Table 4.1 contains parameter estimates for an AR(1), an MA(1) and an ARMA(1,1) for each series. The default spread exhibits a large autoregressive coefficient (.97) that is highly significant, but it also contains a significant moving average term and in an ARMA(1,1) both parameters are significant. The market portfolio exhibits some evidence of predictability although it is much less persistent than the default spread.[7]

4.6 Autocorrelations and Partial Autocorrelations

Autoregressive processes, moving average processes and ARMA processes all exhibit different patterns in their autocorrelations and partial autocorrelations. These differences can be exploited to select a parsimonious model from the general class of ARMA processes.

[7] For information on estimating an ARMA in MATLAB, see the MATLAB supplement to this course.

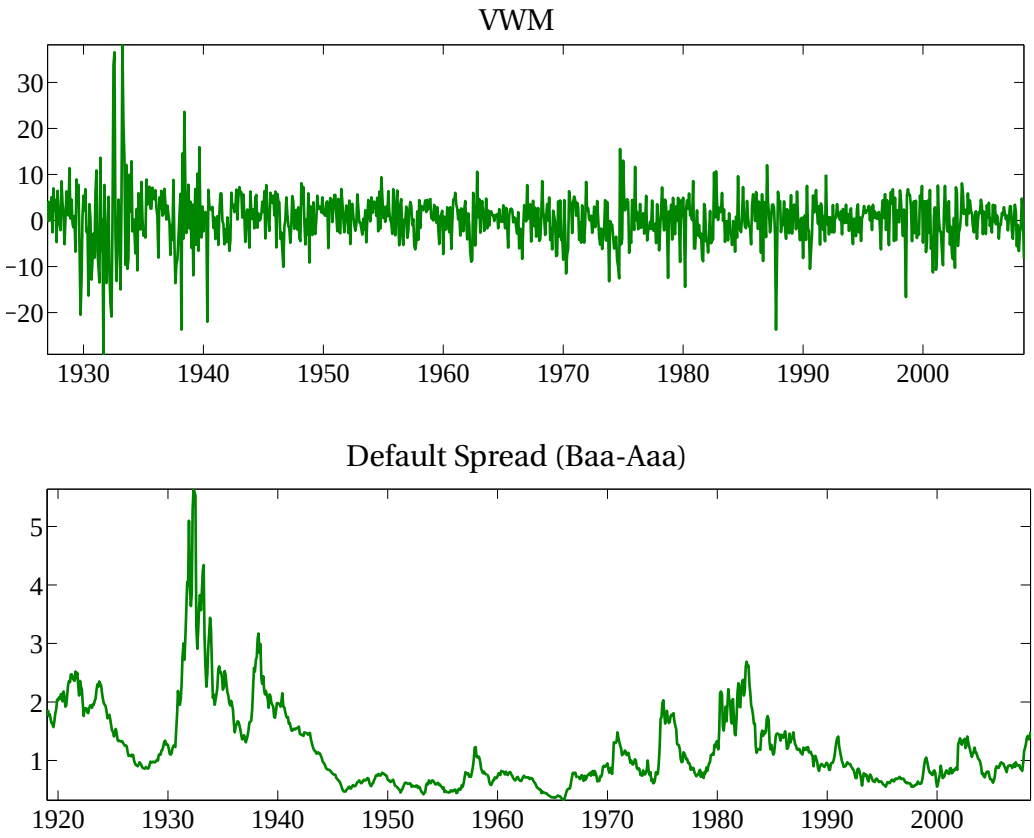

Figure 4.3: Plots of the returns on the VWM and the default spread, the spread between the yield of a portfolio of Baa-rated bonds and the yield of a portfolio of Aaa-rated bonds.

VWM				Baa-Aaa			
$\hat{\phi}_0$	$\hat{\phi}_1$	$\hat{\theta}_1$	$\hat{\sigma}$	$\hat{\phi}_0$	$\hat{\phi}_1$	$\hat{\theta}_1$	$\hat{\sigma}$
0.284 (0.108)	0.115 (0.052)		5.415	0.026 (0.284)	0.978 (0.000)		0.149
0.320 (0.096)		0.115 (0.042)	5.415	1.189 (0.000)		0.897 (0.000)	0.400
0.308 (0.137)	0.039 (0.870)	0.077 (0.724)	5.417	0.036 (0.209)	0.969 (0.000)	0.202 (0.004)	0.146

Table 4.1: Parameter estimates and p-values from an AR(1), MA(1) and ARMA(1,1) for the VWM and Baa-Aaa spread.

4.6.1 Autocorrelations and the Autocorrelation Function

Autocorrelations are to autocovariances as correlations are to covariances. That is, the s$^{\text{th}}$ autocorrelation is the s$^{\text{th}}$ autocovariance divided by the product of the variance of y_t and y_{t-s}, and when a processes is covariance stationary, $\mathrm{V}[y_t] = \mathrm{V}[y_{t-s}]$, and so $\sqrt{\mathrm{V}[y_t]\mathrm{V}[y_{t-s}]} = \mathrm{V}[y_t]$.

Definition 4.16 (Autocorrelation)**.** The autocorrelation of a covariance stationary scalar process

is defined

$$\rho_s = \frac{\gamma_s}{\gamma_0} = \frac{\mathrm{E}[(y_t - E[y_t])(y_{t-s} - E[y_{t-s}])]}{\mathrm{V}[y_t]} \tag{4.47}$$

where γ_s is the s$^{\text{th}}$ autocovariance.

The autocorrelation function (ACF) relates the lag length (s) and the parameters of the model to the autocorrelation.

Definition 4.17 (Autocorrelation Function)**.** The autocorrelation function (ACF), $\rho(s)$, is a function of the population parameters that defines the relationship between the autocorrelations of a process and lag length.

The variance of a covariance stationary AR(1) is $\sigma^2(1-\phi_1^2)^{-1}$ and the s$^{\text{th}}$ autocovariance is $\phi^s\sigma^2(1-\phi_1^2)^{-1}$, and so the ACF is

$$\rho(s) = \frac{\phi^s\sigma^2(1-\phi^2)^{-1}}{\sigma^2(1-\phi^2)^{-1}} = \phi^s. \tag{4.48}$$

Deriving ACFs of ARMA processes is a straightforward, albeit tedious, task. Further details on the derivation of the ACF of a stationary ARMA processes are presented in appendix 4.A.

4.6.2 Partial Autocorrelations and the Partial Autocorrelation Function

Partial autocorrelations are similar to autocorrelations with one important difference: the s$^{\text{th}}$ partial autocorrelation still relates y_t and y_{t-s} but it eliminates the effects of $y_{t-1}, y_{t-2}, \ldots, y_{t-(s-1)}$.

Definition 4.18 (Partial Autocorrelation)**.** The s$^{\text{th}}$ partial autocorrelation (φ_s) is defined as the population value of the regression coefficient on ϕ_s in

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_{s-1} y_{t-(s-1)} + \phi_s y_{t-s} + \epsilon_t.$$

Like the autocorrelation function, the partial autocorrelation function (PACF) relates the partial autocorrelation to population parameters and lag length.

Definition 4.19 (Partial Autocorrelation Function)**.** The partial autocorrelation function (PACF), $\varphi(s)$, defines the relationship between the partial autocorrelations of a process and lag length. The PACF is denoted.

The partial autocorrelations are directly interpretable as population regression coefficients. The s$^{\text{th}}$ partial autocorrelations can be computed using $s+1$ autocorrelations. Recall that the population values of $\phi_1, \phi_2, \ldots, \phi_s$ in

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_{s-1} y_{t-(s-1)} + \phi_s y_{t-s} + \epsilon_t$$

can be defined in terms of the covariance between $y_t, y_{t-1}, y_{t-2}, \ldots, y_{t-s}$. Let $\mathbf{\Gamma}$ denote this covariance matrix,

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \gamma_3 & \cdots & \gamma_{s-1} & \gamma_s \\ \gamma_1 & \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{s-2} & \gamma_{s-1} \\ \gamma_2 & \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{s-3} & \gamma_{s-2} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ \gamma_{s-1} & \gamma_{s-2} & \gamma_{s-3} & \gamma_{s-4} & \cdots & \gamma_0 & \gamma_1 \\ \gamma_s & \gamma_{s-1} & \gamma_{s-2} & \gamma_{s-3} & \cdots & \gamma_1 & \gamma_0 \end{bmatrix}$$

The matrix $\mathbf{\Gamma}$ is known as a Toeplitz matrix which reflects the special symmetry it exhibits which follows from stationarity, and so $\mathrm{E}[(y_t - \mu)(y_{t-s} - \mu)] = \gamma_s = \gamma_{-s} = \mathrm{E}[(y_t - \mu)(y_{t+s} - \mu)]$. $\mathbf{\Gamma}$ can be decomposed in terms of γ_0 (the long-run variance) and the matrix of autocorrelations,

$$\mathbf{\Gamma} = \gamma_0 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \cdots & \rho_{s-1} & \rho_s \\ \rho_1 & 1 & \rho_1 & \rho_2 & \cdots & \rho_{s-2} & \rho_{s-1} \\ \rho_2 & \rho_1 & 1 & \rho_1 & \cdots & \rho_{s-3} & \rho_{s-2} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ \rho_{s-1} & \rho_{s-2} & \rho_{s-3} & \rho_{s-1} & \cdots & 1 & \rho_1 \\ \rho_s & \rho_{s-1} & \rho_{s-2} & \rho_{s-3} & \cdots & \rho_1 & 1 \end{bmatrix}$$

directly by applying the definition of an autocorrelation. The population regression parameters can be computed by partitioning $\mathbf{\Gamma}$ into four blocks, γ_0, the long-run variance of y_t, $\mathbf{\Gamma}_{01} = \mathbf{\Gamma}_{10}'$, the vector of covariances between y_t and $y_{t-1}, y_{t-2}, \ldots, y_{t-s}$, and $\mathbf{\Gamma}_{11}$, the covariance matrix of $y_{t-1}, y_{t-2}, \ldots, y_{t-s}$.

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_0 & \mathbf{\Gamma}_{01} \\ \mathbf{\Gamma}_{10} & \mathbf{\Gamma}_{11} \end{bmatrix} = \gamma_0 \begin{bmatrix} 1 & \mathbf{R}_{01} \\ \mathbf{R}_{10} & \mathbf{R}_{11} \end{bmatrix}$$

where $\mathbf{R}$ are vectors or matrices of autocorrelations. Using this formulation, the population regression parameters $\boldsymbol{\phi} = [\phi_1, \phi_2, \ldots, \phi_s]'$ are defined as

$$\boldsymbol{\phi} = \mathbf{\Gamma}_{11}^{-1}\mathbf{\Gamma}_{10} = \gamma_0^{-1}\mathbf{R}_{11}^{-1}\gamma_0\mathbf{R}_{10} = \mathbf{R}_{11}^{-1}\mathbf{R}_{10}. \tag{4.49}$$

The s$^{\text{th}}$ partial autocorrelation (φ_s) is the s$^{\text{th}}$ element in $\boldsymbol{\phi}$ (when $\mathbf{\Gamma}$ is s by s), $\mathbf{e}_s'\mathbf{R}_{11}^{-1}\mathbf{R}_{10}$ where $\mathbf{e}_s$ is a s by 1 vector of zeros with one in the s$^{\text{th}}$ position.

For example, in a stationary AR(1) model, $y_t = \phi_1 y_{t-1} + \epsilon_t$, the PACF is

$$\begin{aligned} \varphi(s) &= \phi_1^{|s|} \qquad s = 0, 1, -1 \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

That $\varphi_0 = \phi^0 = 1$ is obvious: the correlation of a variable with itself is 1. The first partial autocorrelation is defined as the population parameter of ϕ_1 in the regression $y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t$. Since the data generating process is an AR(1), $\varphi_1 = \phi_1$, the autoregressive parameter. The second partial autocorrelation is defined as the population value of ϕ_2 in the regression

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_2.$$

Since the DGP is an AR(1), once y_{t-1} is included, y_{t-2} has no effect on y_t and the population value of both ϕ_2 and the second partial autocorrelation, φ_2, is 0. This argument holds for any higher order partial autocorrelation.

Note that the first partial autocorrelation and the first autocorrelation are both ϕ_1 in

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t,$$

and at the second (and higher) lag these differ. The autocorrelation at $s = 2$ is the population value of ϕ_2 in the regression

$$y_t = \phi_0 + \phi_2 y_{t-2} + \epsilon$$

while the second partial autocorrelation is the population value of from ϕ_2 in the regression

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon.$$

If the DGP were an AR(1), the second autocorrelation would be $\rho_2 = \phi_1^2$ while the second partial autocorrelation would be $\varphi_2 = 0$.

4.6.2.1 Examples of ACFs and PACFs

The key to understanding the value of ACFs and PACFs lies in the distinct behavior the autocorrelations and partial autocorrelations of AR and MA processes exhibit.

- AR(P)
  - ACF dies exponentially (may oscillate, referred to as sinusoidally)
  - PACF is zero beyond P
- MA(Q)
  - ACF is zero beyond Q
  - PACF dies exponentially (may oscillate, referred to as sinusoidally)

Table 4.2 provides a summary of the ACF and PACF behavior of ARMA models and this difference forms the basis of the Box-Jenkins model selection strategy.

4.6.3 Sample Autocorrelations and Partial Autocorrelations

Sample autocorrelations are computed using sample analogues of the population moments in the definition of an autocorrelation. Define $y_t^* = y_t - \bar{y}$ to be the demeaned series where $\bar{y} = T^{-1}\sum_{t=1}^{T} y_t$. The s$^{\text{th}}$ sample autocorrelation is defined

$$\hat{\rho}_s = \frac{\sum_{t=s+1}^{T} y_t^* y_{t-s}^*}{\sum_{t=1}^{T} (y_t^*)^2} \tag{4.50}$$

although in small samples one the the corrected versions

Process	ACF	PACF
White Noise	All 0	All 0
AR(1)	$\rho_s = \phi^s$	0 beyond lag 2
AR(P)	Decays toward zero exponentially	Non-zero through lag P, 0 thereafter
MA(1)	$\rho_1 \neq 0, \rho_s = 0, s > 1$	Decays toward zero exponentially
MA(Q)	Non-zero through lag Q, 0 thereafter	Decays toward zero exponentially
ARMA(P,Q)	Exponential Decay	Exponential Decay

Table 4.2: Behavior that the ACF and PACF for various members of the ARMA family.

$$\hat{\rho}_s = \frac{\frac{\sum_{t=s+1}^{T} y_t^* y_{t-s}^*}{T-S}}{\frac{\sum_{t=1}^{T}\left(y_t^*\right)^2}{T}} \tag{4.51}$$

or

$$\hat{\rho}_s = \frac{\sum_{t=s+1}^{T} y_t^* y_{t-s}^*}{\sqrt{\sum_{t=s+1}^{T}\left(y_t^*\right)^2 \sum_{t=1}^{T-s}\left(y_t^*\right)^2}}. \tag{4.52}$$

may be more accurate.

Definition 4.20 (Sample Autocorrelogram)**.** A plot of the sample autocorrelations against the lag index in known as a sample autocorrelogram.

Inference on estimated autocorrelation coefficients depends on the null hypothesis tested and whether the data are homoskedastic. The most common assumptions are that the data are homoskedastic and that *all* of the autocorrelations are zero. In other words, $y_t - \mathrm{E}[y_t]$ is white noise process. Under the null $H_0 : \rho_s = 0,\ s \neq 0$, inference can be made noting that $\mathrm{V}[\hat{\rho}_s] = T^{-1}$ using a standard t-test,

$$\frac{\hat{\rho}_s}{\sqrt{\mathrm{V}[\hat{\rho}_s]}} = \frac{\hat{\rho}_s}{\sqrt{T^{-1}}} = T^{\frac{1}{2}}\hat{\rho}_s \xrightarrow{d} N(0,1). \tag{4.53}$$

A alternative null hypothesis is that the autocorrelations on lags s and above are zero but that the autocorrelations on lags $1, 2, \ldots, s-1$ are unrestricted, $H_0 : \rho_j = 0,\ j \geq s$. Under this null, and again assuming homoskedasticity,

$$\begin{aligned} \mathrm{V}[\hat{\rho}_s] &= T^{-1} && \text{for } s = 1 \\ &= T^{-1}(1 + 2\sum_{j=1}^{s-1}\hat{\rho}_j^2) && \text{for } s > 1 \end{aligned} \tag{4.54}$$

If the null is $H_0 : \rho_s = 0$ with no further restrictions on the other autocorrelations, the variance of the s$^{\text{th}}$ autocorrelation is (assuming homoskedasticity)

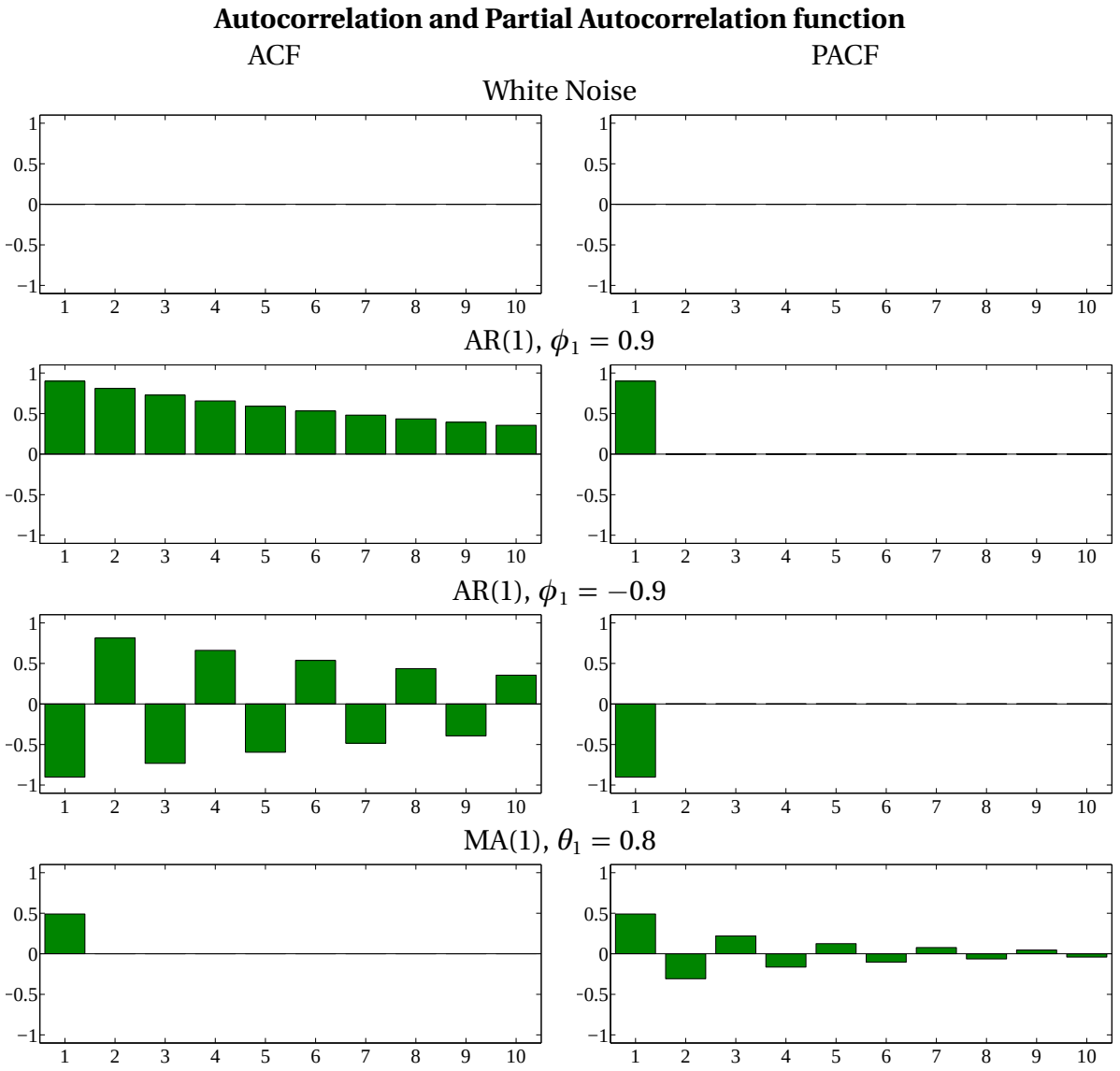

Figure 4.4: Autocorrelation function and partial autocorrelation function for 4 processes. Note the difference between how the ACF and PACF respond in AR and MA models.

$$\mathrm{V}[\hat{\rho}_s] = T^{-1}(1 + 2\sum_{j=1, j\neq s}^{\infty} \hat{\rho}_j^2) \tag{4.55}$$

which is infeasible. The usual practice is to truncate the variance estimator at some finite lag L where L is a function of the sample size, often assumed that $L \propto T^{\frac{1}{3}}$ (if L is not an integer, rounding to the nearest one).[8]

[8]The choice of $L \propto T^{\frac{1}{3}}$ is motivated by asymptotic theory where $T^{\frac{1}{3}}$ has been shown to be the optimal rate in the sense that it minimizes the asymptotic mean square error of the variance estimator.

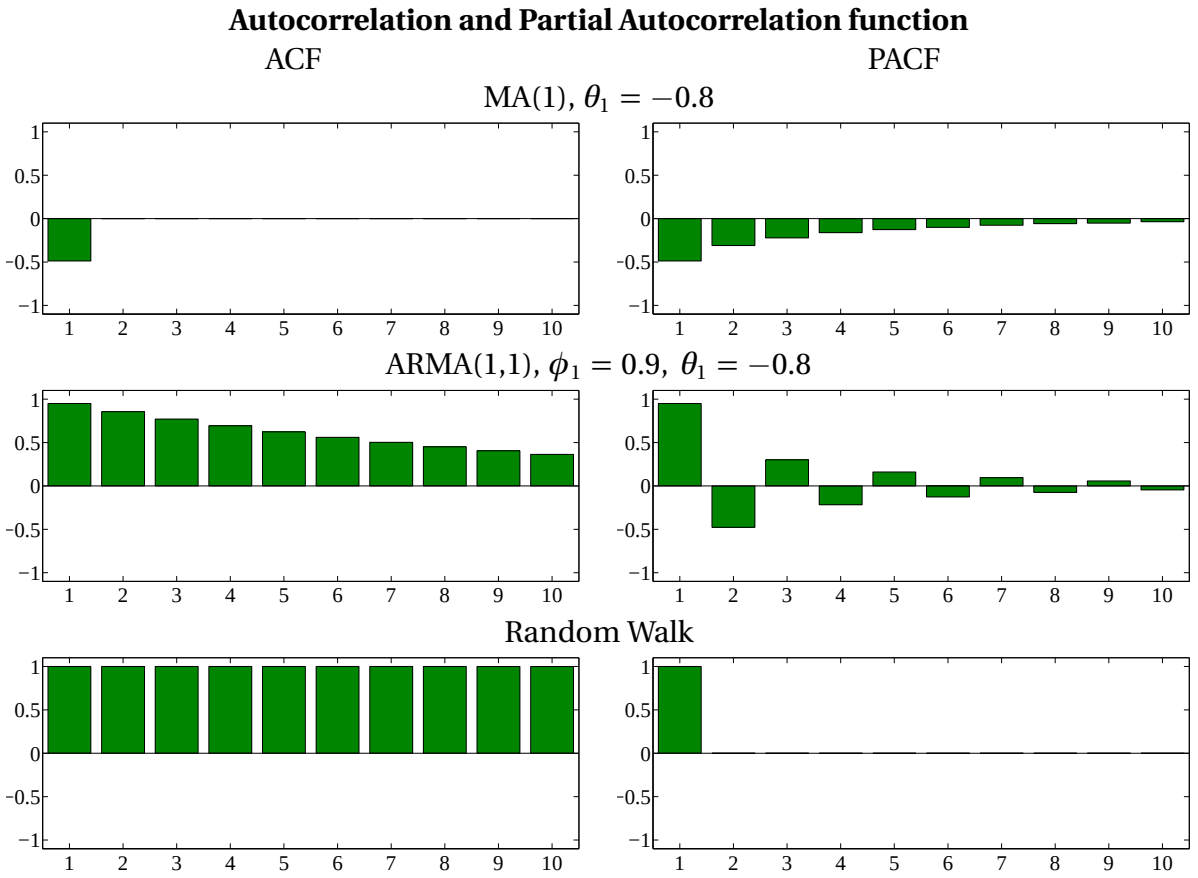

Figure 4.5: Autocorrelation function and partial autocorrelation function for 3 processes, an MA(1), and ARMA(1,1) and a random walk. Note the difference between how the ACF and PACF respond in AR and MA models.

Once the assumption of homoskedasticity is relaxed inference becomes more complicated. First consider the most restrictive null $H_0 : \rho_s = 0,\ s \neq 0$. If $\{y_t\}$ is a heteroskedastic white noise process (plus possibly a non-zero mean), inference can be made using White's heteroskedasticity robust covariance estimator (see chapter 3) so that

$$\begin{aligned} \mathrm{V}[\hat{\rho}_s] &= T^{-1}\left(T^{-1}\sum_{t=1}^{T} y_{t-s}^{*2}\right)^{-1}\left(T^{-1}\sum_{t=1}^{T} y_t^{*2}y_{t-s}^{*2}\right)\left(T^{-1}\sum_{t=1}^{T} y_{t-s}^{*2}\right)^{-1} \\ &= \frac{\sum_{t=s+1}^{T} y_t^{*2}y_{t-s}^{*2}}{\left(\sum_{t=s+1}^{T} y_{t-s}^{*2}\right)^2}. \end{aligned} \tag{4.56}$$

This covariance estimator is identical to White's covariance estimator for the regression

$$y_t = \rho_s y_{t-s} + \epsilon_t$$

since under the null that $\rho_s = 0$, $y_t = \epsilon_t$.

To test one of the more complicated null hypotheses a Heteroskedasticity-Autocorrelation Consistent (HAC) covariance estimator is required, the most common of which is the Newey-West covariance estimator.

Definition 4.21 (Newey-West Variance Estimator)**.** Let z_t be a series that may be autocorrelated and define $z_t^* = z_t - \bar{z}$ where $\bar{z} = T^{-1}\sum_{t=1}^{T} z_t$. The L-lag Newey-West variance estimator for the variance of $\bar{z}$ is

$$\begin{aligned}\hat{\sigma}_{NW}^2 &= T^{-1}\sum_{t=1}^{T} z_t^{*2} + 2\sum_{l=1}^{L} w_l T^{-1}\sum_{t=l+1}^{T} z_t^* z_{t-l}^* \\ &= \hat{\gamma}_0 + 2\sum_{l=1}^{L} w_l \hat{\gamma}_l\end{aligned} \tag{4.57}$$

where $\hat{\gamma}_l = T^{-1}\sum_{t=l+1}^{T} z_t^* z_{t-l}^*$ and $w_l = \frac{L+1-l}{L+1}$.

The Newey-West estimator has two important properties. First, it is always greater than 0. This is a desirable property of any variance estimator. Second, as long as $L \to \infty$, the $\hat{\sigma}_{NW}^2 \xrightarrow{p} \mathrm{V}[y_t]$. The only remaining choice is which value to choose for L. Unfortunately this is problem dependent and it is important to use as small a value for L as the data will permit. Newey-West estimators tend to perform poorly in small samples and are worse, often substantially, than simpler estimators such as White's heteroskedasticity-consistent covariance estimator. This said, they also work in situations where White's estimator fails: when a sequence is autocorrelated White's estimator is not consistent.[9] Long-run variance estimators are covered in more detail in the Multivariate Time Series chapter (chapter 5).

When used in a regression, the Newey-West estimator extends White's covariance estimator to allow $\{y_{t-s}\epsilon_t\}$ to be both heteroskedastic and autocorrelated, setting $z_t^* = y_t^* y_{t-s}^*$,

$$\begin{aligned}\mathrm{V}[\hat{\rho}_s] = {} & T^{-1}\left(T^{-1}\sum_{t=s+1}^{T} y_{t-s}^{*2}\right)^{-1} \\ & \times \left(T^{-1}\sum_{t=s+1}^{T} y_t^{*2} y_{t-s}^{*2} + 2\sum_{j=1}^{L} w_j T^{-1}\sum_{t=s+j+1}^{T} y_t^* y_{t-s}^* \left(y_{t-j}^* y_{t-s-j}^*\right)\right) \\ & \times \left(T^{-1}\sum_{t=s+1}^{T} y_{t-s}^{*2}\right)^{-1} \\ = {} & \frac{\sum_{t=s+1}^{T} y_t^{*2} y_{t-s}^{*2} + 2\sum_{j=1}^{L} w_j \sum_{t=s+j+1}^{T} y_t^* y_{t-s}^* \left(y_{t-j}^* y_{t-s-j}^*\right)}{\left(\sum_{t=s+1}^{T} y_{t-s}^{*2}\right)^2}.\end{aligned} \tag{4.58}$$

Note that only the center term has been changed and that L must diverge for this estimator to be consistent – even if $\{y_t\}$ follows an MA process, and the efficient choice sets $L \propto T^{\frac{1}{3}}$.

Tests that multiple autocorrelations are simultaneously zero can also be conducted. The standard method to test that s autocorrelations are zero, $H_0 = \rho_1 = \rho_2 = \ldots = \rho_s = 0$, is the Ljung-Box Q statistic.

[9] The Newey-West estimator nests White's covariance estimator as a special case by choosing $L = 0$.

Definition 4.22 (Ljung-Box Q statistic)**.** The Ljung-Box Q statistic, or simply Q statistic, tests the null that the first s autocorrelations are all zero against an alternative that at least one is non-zero: $H_0 : \rho_k = 0$ for $k = 1, 2, \ldots, s$ versus $H_1 : \rho_k \neq 0$ for $k = 1, 2, \ldots, s$. The test statistic is defined

$$Q = T(T+2)\sum_{k=1}^{s} \frac{\hat{\rho}_k^2}{T-k} \tag{4.59}$$

and Q has a standard χ_s^2 distribution.

The Q statistic is only valid under an assumption of homoskedasticity so caution is warranted when using it with financial data. A heteroskedasticity robust version of the Q-stat can be formed using an LM test.

Definition 4.23 (LM test for serial correlation)**.** Under the null, $\mathrm{E}[y_t^* y_{t-j}^*] = 0$ for $1 \leq j \leq s$. The LM-test for serial correlation is constructed by defining the score vector $\mathbf{s}_t = y_t^* \begin{bmatrix} y_{t-1}^* & y_{t-2}^* & \ldots & y_{t-s}^* \end{bmatrix}'$,

$$LM = T\bar{\mathbf{s}}'\hat{\mathbf{S}}\bar{\mathbf{s}} \xrightarrow{d} \chi_s^2 \tag{4.60}$$

where $\bar{\mathbf{s}} = T^{-1}\sum_{t=1}^{T} \mathbf{s}_t$ and $\hat{\mathbf{S}} = T^{-1}\sum_{t=1}^{T} \mathbf{s}_t\mathbf{s}_t'$.[10]

Like the Ljung-Box Q statistic, this test has an asymptotic χ_s^2 distribution with the added advantage of being heteroskedasticity robust.

Partial autocorrelations can be estimated using regressions,

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \hat{\varphi}_s y_{t-s} + \epsilon_t$$

where $\hat{\varphi}_s = \hat{\phi}_s$. To test whether a partial autocorrelation is zero, the variance of $\hat{\phi}_s$, under the null and assuming homoskedasticity, is approximately T^{-1} for any s, and so a standard t-test can be used,

$$T^{\frac{1}{2}}\hat{\phi}_s \xrightarrow{d} N(0,1). \tag{4.61}$$

If homoskedasticity cannot be assumed, White's covariance estimator can be used to control for heteroskedasticity.

Definition 4.24 (Sample Partial Autocorrelogram)**.** A plot of the sample partial autocorrelations against the lag index in known as a sample partial autocorrelogram.

4.6.3.1 Example: Autocorrelation, partial autocorrelation and Q Statistic

Figure 4.6 contains plots of the first 20 autocorrelations and partial autocorrelations of the VWM market returns and the default spread. The market appears to have a small amount of persistence and appears to be more consistent with a moving average than an autoregression. The default spread is highly persistent and appears to be a good candidate for an AR(1) since the autocorrelations decay slowly and the partial autocorrelations drop off dramatically after one lag, although an ARMA(1,1) cannot be ruled out.

[10] Refer to chapters 2 and 3 for more on LM-tests.

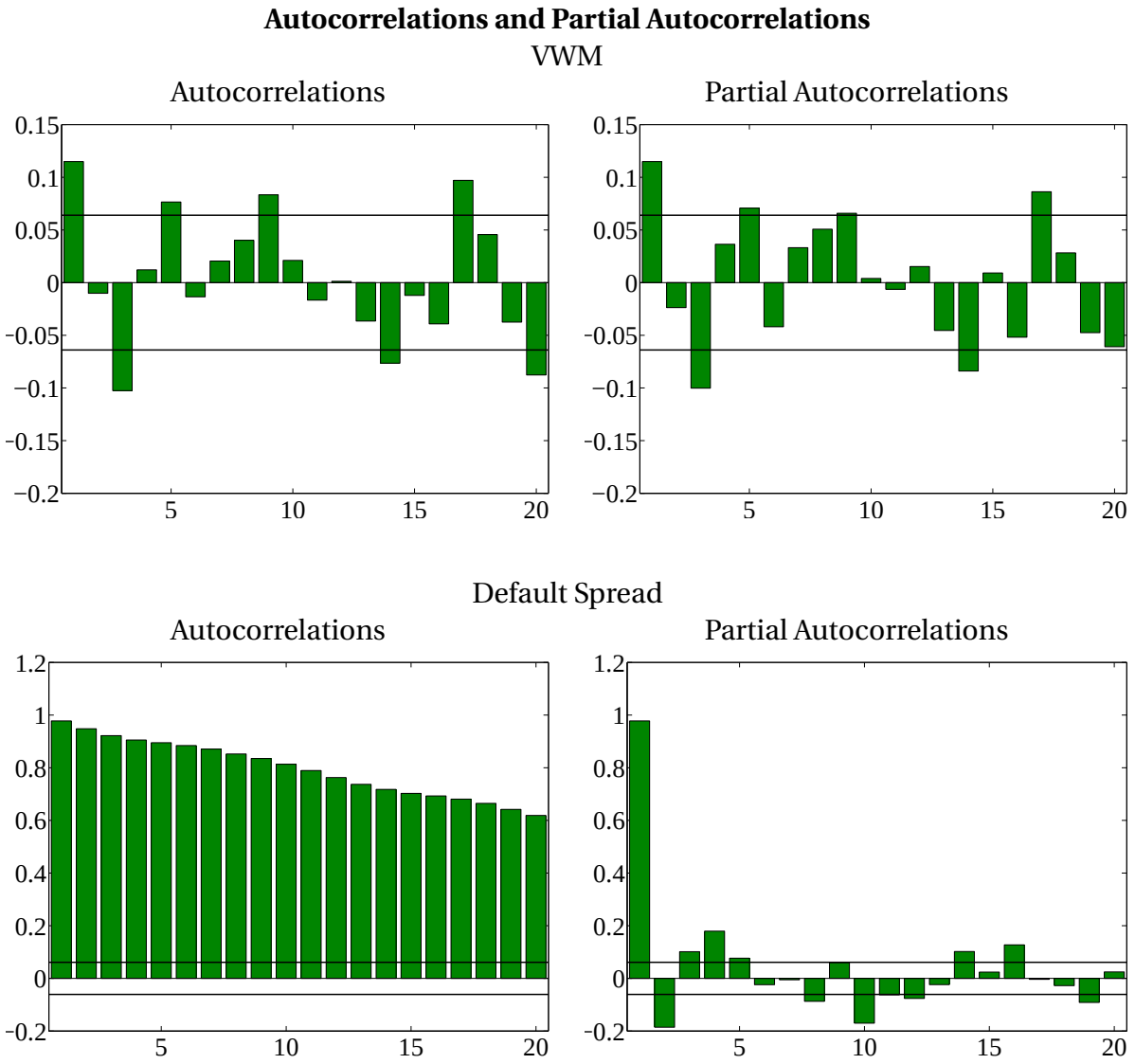

Figure 4.6: These for pictures plot the first 20 autocorrelations (left) and partial autocorrelations (right) of the VWM (top) and the Baa-Aaa spread (bottom). Approximate standard errors, assuming homoskedasticity, are in parenthesis.

4.6.4 Model Selection: The Box-Jenkins Methodology

The Box and Jenkins methodology is the most common approach for time-series model selection. It consists of two stages:

- Identification: Visual inspection of the series, the autocorrelations and the partial autocorrelations.
- Estimation: By relating the sample autocorrelations and partial autocorrelations to the ACF and PACF of ARMA models, candidate models are identified. These candidates are estimated and the residuals are tested for neglected dynamics using the residual autocorrela-

tions, partial autocorrelations and Q statistics or LM-tests for serial correlation. If dynamics are detected in the residuals, a new model is specified and the procedure is repeated.

The Box-Jenkins procedure relies on two principles: parsimony and invertibility.

Definition 4.25 (Parsimony)**.** Parsimony is a property of a model where the specification with the fewest parameters capable of capturing the dynamics of a time series is preferred to other representations equally capable of capturing the same dynamics.

Parsimony is an intuitive principle and using the smallest model has other benefits, particularly when forecasting. One consequence of the parsimony principle is that parameters which are not needed are excluded. For example, if the data generating process were and AR(1), selecting an AR(2) would adequately describe the process. The parsimony principle indicates the AR(1) should be preferred to an AR(2) since both are equally capable of capturing the dynamics of the data. Further, recall that an AR(1) can be reformulated as an MA(T) where $\theta_s = \phi_1^s$. Both the AR(1) and MA(T) are capable of capturing the dynamics of the data if the DGP is an AR(1), although the number of parameters in each is very different. The parsimony principle provides guidance on selecting the AR(1) over the MA(T) since it contains (many) fewer parameters yet provides an equivalent description of the relationship between current and past values of the data.

Definition 4.26 (Invertibility)**.** A moving average is invertible if it can be written as a finite or convergent autoregression. Invertibility requires the roots of

$$(1 - \theta_1 z - \theta_2 z^2 - \ldots - \theta_Q z^Q) = 0$$

to be greater than one in modulus (absolute value).

Invertibility is a technical requirement stemming from the use of the autocorrelogram and partial autocorrelogram to choose the model, and it plays an important role in achieving unique identification of the MA component of a model. For example, the ACF and PACF of

$$y_t = 2\epsilon_{t-1} + \epsilon_t$$

and

$$y_t = .5\epsilon_{t-1} + \epsilon_t$$

are identical. The first autocorrelation is $\theta_1/(1+\theta_1^2)$, and so in the first specification $\rho_1 = 2/(1+2^2) = .4$ and in the second $\rho_1 = .5/(1+.5^2) = .4$ while all other autocorrelations are zero. The partial autocorrelations are similarly identical – partial correlation are functions of autocorrelations – and so two processes are indistinguishable. Invertibility rules out the first of these two models since the root of $1 - 2z = 0$ is $\frac{1}{2} < 1$.

Information criteria such as the AIC or S/BIC can also be used to choose a model. Recall the definitions of the AIC and BIC:

Definition 4.27 (Akaike Information Criterion)**.** The Akaike Information Criteria (AIC) is

$$AIC = \ln \hat{\sigma}^2 + \frac{2k}{T} \tag{4.62}$$

where $\hat{\sigma}^2$ is the estimated variance of the regression error and k is the number of parameters in the model.

Definition 4.28 (Schwarz/Bayesian Information Criterion)**.** The Schwarz Information Criteria (SIC), also known as the Bayesian Information Criterion (BIC) is

$$BIC = \ln \hat{\sigma}^2 + \frac{k \ln T}{T} \tag{4.63}$$

where $\hat{\sigma}^2$ is the estimated variance of the regression error and k is the number of parameters in the model.

ICs are often applied by estimating the largest model which is thought to correctly capture the dynamics and then dropping lags until the AIC or S/BIC fail to decrease. Specific-to-General (StG) and General-to-Specific (GtS) are also applicable to time-series modeling and suffer from the same issues as those described in chapter 3, section 3.13.

4.7 Estimation

ARMA models are typically estimated using maximum likelihood (ML) estimation assuming that the errors are normal, using either conditional maximum likelihood, where the likelihood of y_t given $y_{t-1}, y_{t-2}, \ldots$ is used, or exact maximum likelihood where the joint distribution of $[y_1, y_2, \ldots, y_{t-1}, y_t]$ is used.

4.7.1 Conditional Maximum Likelihood

Conditional maximum likelihood uses the distribution of y_t given $y_{t-1}, y_{t-2}, \ldots$ to estimate the parameters of an ARMA. The data are assumed to be conditionally normal, and so the likelihood is

$$\begin{aligned} f(y_t | y_{t-1}, y_{t-2}, \ldots; \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) &= (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{\epsilon_t^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(y_t - \phi_0 - \sum_{i=1}^{P} \phi_i y_{t-i} - \sum_{j=1}^{Q} \theta_j \epsilon_{t-j})^2}{2\sigma^2}\right) \end{aligned} \tag{4.64}$$

Since the $\{\epsilon_t\}$ series is assumed to be a white noise process, the joint likelihood is simply the product of the individual likelihoods,

$$f(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2} \ldots; \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = \prod_{t=1}^{T} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{\epsilon_t^2}{2\sigma^2}\right) \tag{4.65}$$

and the conditional log-likelihood is

$$l(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2; \mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2} \ldots) = -\frac{1}{2} \sum_{t=1}^{T} \ln 2\pi + \ln \sigma^2 + \frac{\epsilon_t^2}{\sigma^2}. \tag{4.66}$$

Recall that the first-order condition for the mean parameters from a normal log-likelihood does not depend on σ^2 and that given the parameters in the mean equation, the maximum likelihood estimate of the variance is

$$\hat{\sigma}^2 = T^{-1}\sum_{t=1}^{T}(y_t - \phi_0 - \phi_1 y_{t-1} - \ldots - \phi_P y_{t-P} - \theta_1\epsilon_{t-1} - \ldots - \theta_Q\epsilon_{t-Q})^2 \tag{4.67}$$

$$= T^{-1}\sum_{t=1}^{T}\epsilon_t^2. \tag{4.68}$$

This allows the variance to be concentrated out of the log-likelihood so that it becomes

$$\begin{aligned} l(\mathbf{y}_t|\mathbf{y}_{t-1},\mathbf{y}_{t-2}\ldots;\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2) &= -\frac{1}{2}\sum_{t=1}^{T}\ln 2\pi + \ln(T^{-1}\sum_{t=1}^{T}\epsilon_t^2) + \frac{\epsilon_t^2}{T^{-1}\sum_{t=1}^{T}\epsilon_t^2} \\ &= -\frac{1}{2}\sum_{t=1}^{T}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T}\ln(T^{-1}\sum_{t=1}^{T}\epsilon_t^2) - \frac{T}{2}\sum_{t=1}^{T}\frac{\epsilon_t^2}{\sum_{t=1}^{T}\epsilon_t^2} \\ &= -\frac{1}{2}\sum_{t=1}^{T}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T}\ln(T^{-1}\sum_{t=1}^{T}\epsilon_t^2) - \frac{T}{2}\frac{\sum_{t=1}^{T}\epsilon_t^2}{\sum_{t=1}^{T}\epsilon_t^2} \\ &= -\frac{1}{2}\sum_{t=1}^{T}\ln 2\pi - \frac{1}{2}\sum_{t=1}^{T}\ln(T^{-1}\sum_{t=1}^{T}\epsilon_t^2) - \frac{T}{2} \\ &= -\frac{1}{2}\sum_{t=1}^{T}\ln 2\pi - \frac{T}{2} - \frac{1}{2}\sum_{t=1}^{T}\ln(T^{-1}\sum_{t=1}^{T}\epsilon_t^2) \\ &= -\frac{1}{2}\sum_{t=1}^{T}\ln 2\pi - \frac{T}{2} - \frac{T}{2}\ln\hat{\sigma}^2. \end{aligned} \tag{4.69}$$

Eliminating terms that do not depend on model parameters shows that maximizing the likelihood is equivalent to minimizing the error variance,

$$\max_{\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2} l(\mathbf{y}_t|\mathbf{y}_{t-1},\mathbf{y}_{t-2}\ldots;\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2) = -\frac{T}{2}\ln\hat{\sigma}^2. \tag{4.70}$$

where $\hat{\epsilon}_t = y_t - \phi_0 - \phi_1 y_{t-1} - \ldots - \phi_P y_{t-P} - \theta_1\epsilon_{t-1} - \ldots - \theta_Q\epsilon_{t-Q}$, and so . estimation using conditional maximum likelihood is equivalent to least squares, although unlike linear regression the objective is nonlinear due to the moving average terms and so a nonlinear maximization algorithm is required. If the model does not include moving average terms ($Q = 0$), then the conditional maximum likelihood estimates of an AR(P) are identical the least squares estimates from the regression

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_P y_{t-P} + \epsilon_t. \tag{4.71}$$

Conditional maximum likelihood estimation of ARMA models requires either backcast values or truncation since some of the observations have low indices (e.g., y_1) that depend on observations not in the sample (e.g., y_0, y_{-1}, ϵ_0, ϵ_{-1}, etc.). Truncation is the most common and the

likelihood is only computed for $t = P+1, \ldots, T$, and initial values of ϵ_t are set to 0. When using backcasts, missing values of y can be initialized at the long-run average, $\bar{y} = T^{-1}\sum_{t=1}^{T} y_t$, and the initial values of ϵ_t are set to their unconditional expectation, 0. Using unconditional values works well when data are not overly persistent and T is not too small. The likelihood can then be recursively computed where estimated errors $\hat{\epsilon}_t$ used are using in moving average terms,

$$\hat{\epsilon}_t = y_t - \phi_0 - \phi_1 y_{t-1} - \ldots - \phi_P y_{t-P} - \theta_1 \hat{\epsilon}_{t-1} - \ldots - \theta_Q \hat{\epsilon}_{t-Q}, \tag{4.72}$$

where backcast values are used if any index is less than or equal to 0. The estimated residuals are then plugged into the conditional log-likelihood (eq. (4.69)) and the log-likelihood value is computed. The numerical maximizer will search for values of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ that produce the largest log-likelihood. Once the likelihood optimizing values have been found, the maximum likelihood estimate of the variance is computed using

$$\hat{\sigma}^2 = T^{-1}\sum_{t=1}^{T}(y_t - \hat{\phi}_0 - \hat{\phi}_1 y_{t-1} - \ldots - \hat{\phi}_P y_{t-P} - \hat{\theta}_1 \hat{\epsilon}_{t-1} - \ldots - \hat{\theta}_Q \hat{\epsilon}_{t-Q})^2 \tag{4.73}$$

or the truncated version which sums from $P+1$ to T.

4.7.2 Exact Maximum Likelihood

Exact maximum likelihood directly utilizes the autocorrelation function of an ARMA(P,Q) to compute the correlation matrix of *all* of the y data, which allows the joint likelihood to be evaluated. Define

$$\mathbf{y} = [y_t \; y_{t-1} \; y_{t-2} \; \ldots \; y_2 \; y_1]'$$

and let $\boldsymbol{\Gamma}$ be the T by T covariance matrix of $\mathbf{y}$. The joint likelihood of $\mathbf{y}$ is given by

$$f(\mathbf{y}|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = (2\pi)^{-\frac{T}{2}} |\boldsymbol{\Gamma}|^{-\frac{T}{2}} \exp\left(-\frac{\mathbf{y}'\boldsymbol{\Gamma}^{-1}\mathbf{y}}{2}\right). \tag{4.74}$$

The log-likelihood is

$$l(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2; \mathbf{y}) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln|\boldsymbol{\Gamma}| - \frac{1}{2}\mathbf{y}'\boldsymbol{\Gamma}^{-1}\mathbf{y}. \tag{4.75}$$

where $\boldsymbol{\Gamma}$ is a matrix of autocovariances,

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \gamma_3 & \cdots & \gamma_{T-1} & \gamma_T \\ \gamma_1 & \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{T-2} & \gamma_{T-1} \\ \gamma_2 & \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{T-3} & \gamma_{T-2} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ \gamma_{T-1} & \gamma_{T-2} & \gamma_{T-3} & \gamma_{T-4} & \cdots & \gamma_0 & \gamma_1 \\ \gamma_T & \gamma_{T-1} & \gamma_{T-2} & \gamma_{T-3} & \cdots & \gamma_1 & \gamma_0 \end{bmatrix}$$

and that are determined by the model parameters (excluding the constant), $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, and σ^2. A nonlinear maximization algorithm can be used to search for the vector of parameters that maximizes this log-likelihood. The exact maximum likelihood estimator is generally believed to be more precise than conditional maximum likelihood and does note require backcasts of data or errors.

4.8 Inference

Inference on ARMA parameters from stationary time series is a standard application of maximum likelihood theory. Define $\boldsymbol{\psi} = [\boldsymbol{\phi}\ \boldsymbol{\theta}\ \sigma^2]'$ as the parameter vector. Recall from 2 that maximum likelihood estimates are asymptotically normal,

$$\sqrt{T}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}^{-1}) \tag{4.76}$$

where

$$\mathcal{I} = -\mathrm{E}\left[\frac{\partial^2 l(\mathbf{y};\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'}\right].$$

where $\partial^2 l(\mathbf{y};\boldsymbol{\psi})/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'$ is the second derivative matrix of the log-likelihood (or Hessian). In practice $\mathcal{I}$ is not known and it must be replaced with a consistent estimate,

$$\hat{\mathcal{I}} = T^{-1}\sum_{t=1}^{T} -\frac{\partial^2 l(y_t;\hat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'}.$$

Wald and t-tests on the parameter estimates can be computed using the elements of $\mathcal{I}$, or likelihood ratio tests can be used by imposing the null on the model and comparing the log-likelihood values of the constrained and unconstrained estimators.

One important assumption in the above distribution theory is that the estimator is a maximum likelihood estimator; this requires the likelihood to be correctly specified, or, in other words, for the data to be homoskedastic and normally distributed. This is generally an implausible assumption when using financial data and a modification of the above theory is needed. When one likelihood is specified for the data but they actually have a different distribution the estimator is known as a Quasi Maximum Likelihood estimator (QML). QML estimators, like ML estimators, are asymptotically normal under mild regularity conditions on the data but with a different asymptotic covariance matrix,

$$\sqrt{T}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}^{-1}\mathcal{J}\mathcal{I}^{-1}) \tag{4.77}$$

where

$$\mathcal{J} = \mathrm{E}\left[\frac{\partial\, l(\mathbf{y};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\frac{\partial\, l(\mathbf{y};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}\right]$$

$\mathcal{J}$ must also be estimated and the usual estimator is

$$\hat{\mathcal{J}} = T^{-1}\sum_{t=1}^{T} \frac{\partial\, l(y_t;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\frac{\partial\, l(y_t;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}$$

where $\frac{\partial\, l(y_t;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}$ is the score of the log-likelihood. $\mathcal{I}^{-1}\mathcal{J}\mathcal{I}^{-1}$ is known as a sandwich covariance estimator, White's covariance estimator.

A sandwich covariance estimator is needed when the model for the data is not completely specified or is misspecified, and it accounts for the failure of Information Matrix Inequality to hold (see chapters 2and 3). As was the case in linear regression, a sufficient condition for the IME to fail in ARMA estimation is heteroskedastic residuals. Considering the prevalence of conditionally heteroskedasticity in financial data, this is nearly a given.

4.9 Forecasting

Forecasting is a common objective of many time-series models. The objective of a forecast is to minimize a loss function.

Definition 4.29 (Loss Function)**.** A loss function is a function of the observed data, y_{t+h} and the time-t constructed forecast, $\hat{y}_{t+h|t}$, $L(y_t, \hat{y}_{t+h|t})$, that has the three following properties:

- Property 1: The loss of any forecast is non-negative, so $L(y_{t+h}, \hat{y}_{t+h|t}) \geq 0$.
- Property 2: There exists a point, y^*_{t+h}, known as the optimal forecast, where the loss function takes the value 0. That is $L(y_{t+h}, y^*_{t+h}) = 0$.
- Property 3: The loss is non-decreasing away from y^*_{t+h}. That is if $y^B_{t+h} > y^A_{t+h} > y^*_{t+h}$, then $L(y_{t+h}, y^B_{t+h}) > L(y_{t+h}, y^A_{t+h}) > L(y_{t+h}, y^*_{t+h})$. Similarly, if $y^D_{t+h} < y^C_{t+h} < y^*_{t+h}$, then $L(y_{t+h}, y^D_{t+h}) > L(y_{t+h}, y^C_{t+h}) > L(y_{t+h}, y^*_{t+h})$.

The most common loss function is Mean Square Error (MSE) which chooses the forecast to minimize

$$\mathrm{E}[L(y_{t+h}, \hat{y}_{t+h|t})] = \mathrm{E}[(y_{t+h} - \hat{y}_{t+h|t})^2] \tag{4.78}$$

where $\hat{y}_{t+h|t}$ is the time-t forecast of y_{t+h}. Notice that this is just the optimal projection problem and the optimal forecast is the conditional mean, $y^*_{t+h|t} = \mathrm{E}_t[y_{t+h}]$ (See chapter 3). It is simple to verify that this loss function satisfies the properties of a loss function. Property 1 holds by inspection and property 2 occurs when $y_{t+h} = \hat{y}^*_{t+h|t}$. Property 3 follows from the quadratic form. MSE is far and away the most common loss function but others, such as Mean Absolute Deviation (MAD), Quad-Quad and Linex are used both in practice and in the academic literature. The MAD loss function will be revisited in chapter 6 (Value-at-Risk). The Advanced Financial Econometrics elective will study non-MSE loss functions in more detail.

The remainder of this section will focus exclusively on forecasts that minimize the MSE loss function. Fortunately, in this case, forecasting from ARMA models is an easy exercise. For simplicity consider the AR(1) process,

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t.$$

Since the optimal forecast is the conditional mean, all that is needed is to compute $\mathrm{E}_t[y_{t+h}]$ for any h. When $h = 1$,

$$y_{t+1} = \phi_0 + \phi_1 y_t + \epsilon_{t+1}$$

so the conditional expectation is

$$\begin{aligned} \mathrm{E}_t[y_{t+1}] &= \mathrm{E}_t[\phi_0 + \phi_1 y_t + \epsilon_{t+1}] \\ &= \phi_0 + \phi_1 \mathrm{E}_t[y_t] + \mathrm{E}_t[\epsilon_{t+1}] \\ &= \phi_0 + \phi_1 y_t + 0 \\ &= \phi_0 + \phi_1 y_t \end{aligned} \tag{4.79}$$

which follows since y_t is in the time-t information set ($\mathcal{F}_t$) and $\mathrm{E}_t[\epsilon_{t+1}] = 0$ by assumption.[11]

[11] This requires a sightly stronger assumption than ϵ_t is a white noise process.

The optimal forecast for $h = 2$ is given by $\mathrm{E}_t[y_{t+2}]$,

$$\begin{aligned}\mathrm{E}_t[y_{t+2}] &= \mathrm{E}_t[\phi_0 + \phi_1 y_{t+1} + \epsilon_{t+2}] \\ &= \phi_0 + \phi_1 \mathrm{E}_t[y_{t+1}] + \mathrm{E}_t[\epsilon_{t+1}] \\ &= \phi_0 + \phi_1(\phi_0 + \phi_1 y_t) + 0 \\ &= \phi_0 + \phi_1\phi_0 + \phi_1^2 y_t\end{aligned}$$

which follows by substituting in the expression derived in eq. (4.79) for $\mathrm{E}_t[y_{t+1}]$. The optimal forecast for any arbitrary h uses the recursion

$$\mathrm{E}_t[y_{t+h}] = \phi_0 + \phi_1 \mathrm{E}_t[y_{t+h-1}] \tag{4.80}$$

and it is easily shown that $\mathrm{E}_t[y_{t+h}] = \phi_0 \sum_{i=0}^{h-1} \phi_1^i + \phi_1^h y_t$. If $|\phi_1| < 1$, as $h \to \infty$, the forecast of y_{t+h} and $E_t[y_{t+h}]$ converges to $\phi_0/(1-\phi_1)$, the unconditional expectation of y_t. In other words, for forecasts in the distant future there is no information about the location of y_{t+h} other than it will return to its unconditional mean. This is not surprising since y_t is covariance stationary when $|\phi_1| < 1$.

Next consider forecasts from an MA(2),

$$y_t = \phi_0 + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \epsilon_t.$$

The one-step-ahead forecast is given by

$$\begin{aligned}\mathrm{E}_t[y_{t+1}] &= \mathrm{E}_t[\phi_0 + \theta_1\epsilon_t + \theta_2\epsilon_{t-1} + \epsilon_{t+1}] \\ &= \phi_0 + \theta_1\mathrm{E}_t[\epsilon_t] + \theta_2\mathrm{E}_t[\epsilon_{t-1}] + \mathrm{E}_t[\epsilon_{t+1}] \\ &= \phi_0 + \theta_1\epsilon_t + \theta_2\epsilon_{t-1} + 0\end{aligned}$$

which follows since ϵ_t and ϵ_{t-1} are in the $\mathcal{F}_t$ information set and $\mathrm{E}_t[\epsilon_{t+1}] = 0$ by assumption. In practice the one step ahead forecast would be given by

$$\mathrm{E}_t[y_{t+1}] = \hat{\phi}_0 + \hat{\theta}_1\hat{\epsilon}_t + \hat{\theta}_2\hat{\epsilon}_{t-1}$$

where both the unknown parameters *and* the unknown residuals would be replaced with their estimates.[12] The 2-step ahead forecast is given by

$$\begin{aligned}\mathrm{E}_t[y_{t+2}] &= \mathrm{E}_t[\phi_0 + \theta_1\epsilon_{t+1} + \theta_2\epsilon_t + \epsilon_{t+2}] \\ &= \phi_0 + \theta_1\mathrm{E}_t[\epsilon_{t+1}] + \theta_2\mathrm{E}_t[\epsilon_t] + \mathrm{E}_t[\epsilon_{t+2}] \\ &= \phi_0 + \theta_1 0 + \theta_2\epsilon_t + 0 \\ &= \phi_0 + \theta_2\epsilon_t.\end{aligned}$$

The 3 or higher step forecast can be easily seen to be ϕ_0. Since all future residuals have zero expectation they cannot affect long horizon forecasts. Like the AR(1) forecast, the MA(2) forecast

[12] The residuals are a natural by-product of the parameter estimation stage.

is mean reverting. Recall the unconditional expectation of an MA(Q) process is simply ϕ_0. For any $h > Q$ the forecast of y_{t+h} is just this value, ϕ_0.

Finally consider the 1 to 3-step ahead forecasts from an ARMA(2,2),

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \epsilon_t.$$

Conditioning on the information set $\mathcal{F}_t$, the expectation of y_{t+1} is

$$\begin{aligned}\mathrm{E}_t[y_{t+1}] &= \mathrm{E}_t[\phi_0 + \phi_1 y_t + \phi_2 y_{t-1} + \theta_1 \epsilon_t + \theta_2 \epsilon_{t-1} + \epsilon_{t+1}] \\ &= \mathrm{E}_t[\phi_0] + \mathrm{E}_t[\phi_1 y_t] + \mathrm{E}_t[\phi_2 y_{t-1}] + \mathrm{E}_t[\theta_1 \epsilon_t] + \mathrm{E}_t[\theta_2 \epsilon_{t-1}] + \mathrm{E}_t[\epsilon_{t+1}].\end{aligned}$$

Noting that all of the elements are in $\mathcal{F}_t$ except ϵ_{t+1}, which has conditional expectation 0,

$$\mathrm{E}_t[y_{t+1}] = \phi_0 + \phi_1 y_t + \phi_2 y_{t-1} + \theta_1 \epsilon_t + \theta_2 \epsilon_{t-1}$$

Note that in practice, the parameters and errors will all be replaced by their estimates (i.e. $\hat{\phi}_1$ and $\hat{\epsilon}_t$). The 2-step ahead forecast is given by

$$\begin{aligned}\mathrm{E}_t[y_{t+2}] &= \mathrm{E}_t[\phi_0 + \phi_1 y_{t+1} + \phi_2 y_t + \theta_1 \epsilon_{t+1} + \theta_2 \epsilon_t + \epsilon_{t+2}] \\ &= \mathrm{E}_t[\phi_0] + \mathrm{E}_t[\phi_1 y_{t+1}] + \mathrm{E}_t[\phi_2 y_t] + \theta_1 \mathrm{E}_t[\epsilon_{t+1}] + \theta_2 \epsilon_t + \mathrm{E}_t[\epsilon_{t+2}] \\ &= \phi_0 + \phi_1 \mathrm{E}_t[y_{t+1}] + \phi_2 y_t + \theta_1 \mathrm{E}_t[\epsilon_{t+1}] + \theta_2 \epsilon_t + \mathrm{E}_t[\epsilon_{t+2}] \\ &= \phi_0 + \phi_1 (\phi_0 + \phi_1 y_t + \phi_2 y_{t-1} + \theta_1 \epsilon_t + \theta_2 \epsilon_{t-1}) + \phi_2 y_t + \theta_1 0 + \theta_2 \epsilon_t + 0 \\ &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 y_t + \phi_1 \phi_2 y_{t-1} + \phi_1 \theta_1 \epsilon_t + \phi_1 \theta_2 \epsilon_{t-1} + \phi_2 y_t + \theta_2 \epsilon_t \\ &= \phi_0 + \phi_1 \phi_0 + (\phi_1^2 + \phi_2) y_t + \phi_1 \phi_2 y_{t-1} + (\phi_1 \theta_1 + \theta_2) \epsilon_t + \phi_1 \theta_2 \epsilon_{t-1}.\end{aligned}$$

In this case, there are three terms which are not known at time t. By assumption $\mathrm{E}_t[\epsilon_{t+2}] = \mathrm{E}_t[\epsilon_{t+1}] = 0$ and $\mathrm{E}_t[y_{t+1}]$ has been computed above, so

$$\mathrm{E}_t[y_{t+2}] = \phi_0 + \phi_1 \phi_0 + (\phi_1^2 + \phi_2) y_t + \phi_1 \phi_2 y_{t-1} + (\phi_1 \theta_1 + \theta_2) \epsilon_t + \phi_1 \theta_2 \epsilon_{t-1}$$

In a similar manner,

$$\begin{aligned}\mathrm{E}_t[y_{t+3}] &= \phi_0 + \phi_1 \mathrm{E}_t[y_{t+2}] + \phi_2 \mathrm{E}_t[y_{t+1}] + \theta_1 \epsilon_{t+2} + \theta_2 \epsilon_{t+1} + \epsilon_{t+3} \\ \mathrm{E}_t[y_{t+3}] &= \phi_0 + \phi_1 \mathrm{E}_t[y_{t+2}] + \phi_2 \mathrm{E}_t[y_{t+1}] + 0 + 0 + 0\end{aligned}$$

which is easily solved by plugging in the previously computed values for $\mathrm{E}_t[y_{t+2}]$ and $\mathrm{E}_t[y_{t+1}]$. This pattern can be continued by iterating forward to produce the forecast for an arbitrary h.

Two things are worth noting from this discussion:

- If there is no AR component, all forecast for $h > Q$ will be ϕ_0.
- For large h, the optimal forecast converges to the unconditional expectation given by

$$\lim_{h \to \infty} \mathrm{E}_t[y_{t+h}] = \frac{\phi_0}{1 - \phi_1 - \phi_2 - \ldots - \phi_P} \tag{4.81}$$

4.9.1 Forecast Evaluation

Forecast evaluation is an extensive topic and these notes only cover two simple yet important tests: Mincer-Zarnowitz regressions and Diebold-Mariano tests.

4.9.1.1 Mincer-Zarnowitz Regressions

Mincer-Zarnowitz regressions (henceforth MZ) are used to test for the optimality of the forecast and are implemented with a standard regression. If a forecast is correct, it should be the case that a regression of the realized value on its forecast and a constant should produce coefficients of 1 and 0 respectively.

Definition 4.30 (Mincer-Zarnowitz Regression)**.** A Mincer-Zarnowitz (MZ) regression is a regression of a forecast, $\hat{y}_{t+h|t}$ on the realized value of the predicted variable, y_{t+h} and a constant,

$$y_{t+h} = \beta_1 + \beta_2 \hat{y}_{t+h|t} + \eta_t. \tag{4.82}$$

If the forecast is optimal, the coefficients in the MZ regression should be consistent with $\beta_1 = 0$ and $\beta_2 = 1$.

For example, let $\hat{y}_{t+h|t}$ be the h-step ahead forecast of y constructed at time t. Then running the regression

$$y_{t+h} = \beta_1 + \beta_2 \hat{y}_{t+h|t} + \nu_t$$

should produce estimates close to 0 and 1. Testing is straightforward and can be done with any standard test (Wald, LR or LM). An augmented MZ regression can be constructed by adding time-t measurable variables to the original MZ regression.

Definition 4.31 (Augmented Mincer-Zarnowitz Regression)**.** An Augmented Mincer-Zarnowitz regression is a regression of a forecast, $\hat{y}_{t+h|t}$ on the realized value of the predicted variable, y_{t+h}, a constant and any other time-t measurable variables, $\mathbf{x}_t = [x_{1t}\ x_{2t}\ \ldots\ x_{Kt}]$,

$$y_{t+h} = \beta_1 + \beta_2 \hat{y}_{t+h|t} + \beta_3 x_{1t} + \ldots + \beta_{K+2} x_{Kt} + \eta_t. \tag{4.83}$$

If the forecast is optimal, the coefficients in the MZ regression should be consistent with $\beta_1 = \beta_3 = \ldots = \beta_{K+2} = 0$ and $\beta_2 = 1$.

It is crucial that the additional variables are time-t measurable and are in $\mathcal{F}_t$. Again, any standard test statistic can be used to test the null $H_0 : \beta_2 = 1 \cap \beta_1 = \beta_3 = \ldots = \beta_{K+2} = 0$ against the alternative $H_1 : \beta_2 \neq 1 \cup \beta_j \neq 0,\ j = 1, 3, 4, \ldots, K-1, K-2$.

4.9.1.2 Diebold-Mariano Tests

A Diebold-Mariano test, in contrast to an MZ regression, examines the relative performance of two forecasts. Under MSE, the loss function is given by $L(y_{t+h}, \hat{y}_{t+h|t}) = (y_{t+h} - \hat{y}_{t+h|t})^2$. Let A and B index the forecasts from two models $\hat{y}^A_{t+h|t}$ and $\hat{y}^B_{t+h|t}$, respectively. The losses from each can be defined as $l^A_t = (y_{t+h} - \hat{y}^A_{t+h|t})^2$ and $l^B_t = (y_{t+h} - \hat{y}^B_{t+h|t})^2$. If the models were equally good (or bad), one would expect $\bar{l}^A \approx \bar{l}^B$ where $\bar{l}$ is the average loss. If model A is better, meaning it

has a lower expected loss $\mathrm{E}[L(y_{t+h}, \hat{y}^A_{t+h|t})] < \mathrm{E}[L(y_{t+h}, \hat{y}^B_{t+h|t})]$, then, on average, it should be the case that $\bar{l}^A < \bar{l}^B$. Alternatively, if model B were better it should be the case that $\bar{l}^B < \bar{l}^A$. The DM test exploits this to construct a simple t-test of equal predictive ability.

Definition 4.32 (Diebold-Mariano Test)**.** Define $d_t = l^A_t - l^B_t$. The Diebold-Mariano test is a test of equal predictive accuracy and is constructed as

$$DM = \frac{\bar{d}}{\sqrt{\widehat{\mathrm{V}[\bar{d}]}}}$$

where M (for modeling) is the number of observations used in the model building and estimation, R (for reserve) is the number of observations held back for model evaluation and $\bar{d} = R^{-1}\sum_{t=M+1}^{M+R} d_t$. Under the null that $\mathrm{E}[L(y_{t+h}, \hat{y}^A_{t+h|t})] = \mathrm{E}[L(y_{t+h}, \hat{y}^B_{t+h|t})]$, and under some regularity conditions on $\{d_t\}$, $DM \xrightarrow{d} N(0,1)$. $\mathrm{V}[d_t]$ is the *long-run variance* of d_t and must be computed using a HAC covariance estimator.

If the models are equally accurate, one would expect that $\mathrm{E}[d_t] = 0$ which forms the null of the DM test, $H_0 : \mathrm{E}[d_t] = 0$. To test the null, a standard t-stat is used although the test has two alternatives: $H^A_1 : \mathrm{E}[d_t] < 0$ and $H^B_1 : \mathrm{E}[d_t] > 0$ which correspond to the superiority of model A or B, respectively. DM is asymptotically normally distributed. Large negative values (less than -2) indicate model A produces less loss on average and hence is superior, while large positive values indicate the opposite. Values close to zero indicate neither is statistically superior.

In Diebold-Marino tests the variance must be estimated using a Heteroskedasticity-Autocorrelation Consistent variance estimator.

Definition 4.33 (Heteroskedasticity Autocorrelation Consistent Covariance Estimator)**.** Covariance estimators which are robust to both ignored autocorrelation in residuals and to heteroskedasticity are known as Heteroskedasticity-Autocorrelation Consistent (HAC) covariance. The most common example of an HAC estimator is the Newey-West (or Bartlett) covariance estimator.

The typical variance estimator cannot be used in DM tests and a kernel estimator must be substituted (e.g., Newey-West).

Despite all of these complications, implementing a DM test is very easy. The first step is to compute the series of losses, $\{l^A_t\}$ and $\{l^B_t\}$, for both forecasts. Next compute $d_t = l^A_t - l^B_t$. Finally, regress d_t on a constant and use Newey-West errors,

$$d_t = \beta_1 + \epsilon_t.$$

The t-stat on β_1 is the DM test statistic and can be compared to critical values of a normal distribution.

4.10 Nonstationary Time Series

Nonstationary time series present some particular difficulties and standard inference often fails when a process depends explicitly on t. Nonstationarities can be classified into one of four categories:

- Seasonalities
- Deterministic Trends (also known as Time Trends)
- Unit Roots (also known as Stochastic Trends)
- Structural Breaks

Each type has a unique feature. Seasonalities are technically a form of deterministic trend, although their analysis is sufficiently similar to stationary time series that little is lost in treating a seasonal time series as if it were stationary. Processes with deterministic trends have unconditional means which depend on time while unit roots processes have unconditional variances that grow over time. Structural breaks are an encompassing class which may result in either or both the mean and variance exhibiting time dependence.

4.10.1 Seasonality, Diurnality, and Hebdomadality

Seasonality, diurnality and hebdomadality are pervasive in economic time series. While many data series have been *seasonally adjusted* to remove seasonalities, particularly US macroeconomic series, there are many time-series where no seasonally adjusted version is available. Ignoring seasonalities is detrimental to the precision of parameters and forecasting and model estimation and selection is often more precise when both seasonal and nonseasonal dynamics are simultaneously modeled.

Definition 4.34 (Seasonality)**.** Data are said to be seasonal if they exhibit a non-constant deterministic pattern with an annual frequency.

Definition 4.35 (Hebdomadality)**.** Data which exhibit day-of-week deterministic effects are said to be hebdomadal.

Definition 4.36 (Diurnality)**.** Data which exhibit intra-daily deterministic effects are said to be diurnal.

Seasonal data are non-stationary, although seasonally de-trended data (usually referred to as deseasonalized data) may be stationary. Seasonality is common in macroeconomic time series, diurnality is pervasive in ultra-high frequency data (tick data) and hebdomadality is often believed to be a feature of asset prices. Seasonality is, technically, a form of non-stationarity since the mean of a process exhibits explicit dependence on t through the seasonal component, and the Box-Jenkins methodology is not directly applicable. However, a slight change in time scale, where the seasonal pattern is directly modeled along with any non-seasonal dynamics produces a residual series which is stationary and so the Box-Jenkins methodology may be applied.

For example, consider a seasonal quarterly time series. Seasonal dynamics may occur at lags $4, 8, 12, 16, \ldots$, while nonseasonal dynamics can occur at any lag $1, 2, 3, 4, \ldots$. Note that multiples of 4 appear in both lists and so the identification of the seasonal and nonseasonal dynamics may be difficult (although separate identification makes little practical difference).

The standard practice when working with seasonal data is to conduct model selection over two sets of lags by choosing a maximum lag to capture the seasonal dynamics and by choosing a maximum lag to capture nonseasonal ones. Returning to the example of a seasonal quarterly

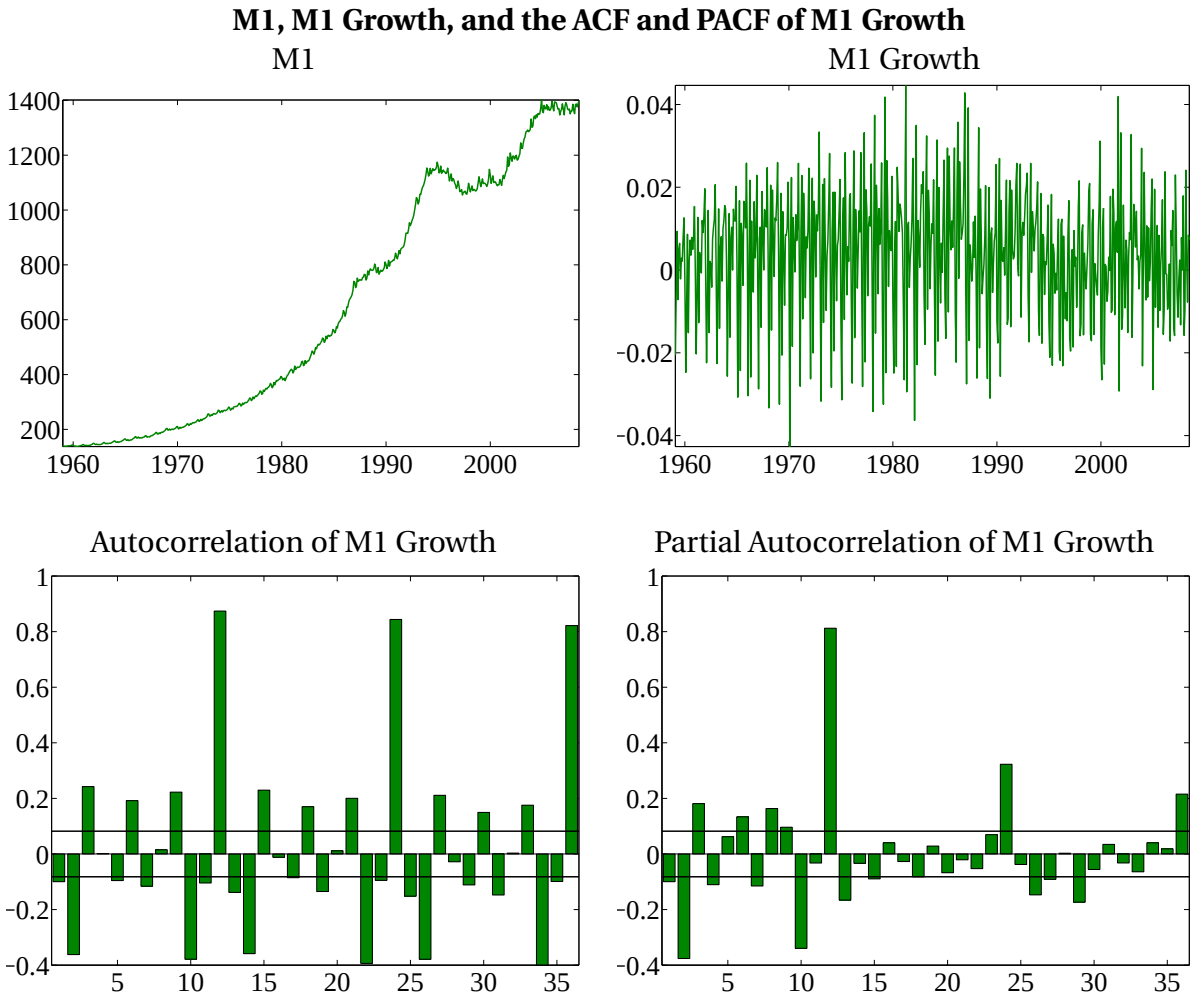

Figure 4.7: Plot of the money supply (M1), M1 growth (log differences), and the sample autocorrelogram and sample partial autocorrelogram of M1 growth. There is a clear seasonal pattern at 12 months which appears consistent with a seasonal ARMA(1,1).

time series, a model may need to examine up to 4 lags to capture nonseasonal dynamics and up to 4 lags of the seasonal component, and if the seasonal component is annual, these four seasonal lags correspond to regressors as $t-4$, $t-8$, $t-12$, and $t-16$.

4.10.1.1 Example: Seasonality

Most U.S. data series are available seasonally adjusted, something that is not true for data from many areas of the world, including the Euro zone. This example makes use of monthly data on the U.S. money supply, M1, a measure of the money supply that includes all coins, currency held by the public, travelers' checks, checking account balances, NOW accounts, automatic transfer service accounts, and balances in credit unions.

Figure 4.10.1.1 contains a plot of monthly M1, the growth of M1 (log differences), and the sample autocorrelogram and sample partial autocorrelogram of M1. These figures show evi-

Modeling seasonalities in M1 growth

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_{12} y_{t-12} + \theta_{12}\epsilon_{t-12} + \epsilon_t$$

$\hat{\phi}_0$	$\hat{\phi}_1$	$\hat{\phi}_{12}$	$\hat{\theta}_{12}$	SIC
0.000 (0.245)	−0.014 (0.000)	0.984 (0.000)	−0.640 (0.000)	−9.989
0.001 (0.059)	−0.011 (0.000)	0.873 (0.000)		−9.792
0.004 (0.002)	−0.008 (0.000)		0.653 (0.000)	−9.008

Table 4.3: Estimated parameters, p-values and SIC for three models with seasonalities. The SIC prefers the larger specification with both seasonal AR and MA terms. Moreover, correctly modeling the seasonalities frees the AR(1) term to model the oscillating short run dynamics (notice the significant negative coefficient).

dence of an annual seasonality (lags 12, 24 and 36), and applying the Box-Jenkins methodology, the seasonality appears to be a seasonal AR, or possibly a seasonal ARMA. The short run dynamics oscillate and appear consistent with an autoregression since the partial autocorrelations are fairly flat (aside from the seasonal component). Three specifications which may be appropriate to model the process were fit: a 12 month seasonal AR, a 12 month seasonal MA and a 12-month seasonal ARMA, all combined with an AR(1) to model the short run dynamics. Results are reported in table 4.3

4.10.2 Deterministic Trends

The simplest form of nonstationarity is a deterministic trend. Models with deterministic time trends can be decomposed into three components:

$$y_t = \text{deterministictrend} + \text{stationarycomponent} + \text{noise} \tag{4.84}$$

where $\{y_t\}$ would be stationary if the trend were absent. The two most common forms of time trends are polynomial (linear, quadratic, etc) and exponential. Processes with polynomial time trends can be expressed

$$y_t = \phi_0 + \delta_1 t + \delta_2 t^2 + \ldots + \delta_S t^S + \text{stationarycomponent} + \text{noise},$$

and linear time trend models are the most common,

$$y_t = \phi_0 + \delta_1 t + \text{stationarycomponent} + \text{noise}.$$

For example, consider a linear time trend model with an MA(1) stationary component,

$$y_t = \phi_0 + \delta_1 t + \theta_1 \epsilon_{t-1} + \epsilon_t$$

The long-run behavior of this process is dominated by the time trend, although it may still exhibit persistent fluctuations around $\delta_1 t$.

Exponential trends appear as linear or polynomial trends in the log of the dependent variable, for example

$$\ln y_t = \phi_0 + \delta_1 t + \text{stationarycomponent} + \text{noise}.$$

The trend is the permanent component of a nonstationary time series, and so any two observation are permanently affected by the trend line irrespective of the number of observations between them. The class of deterministic trend models can be reduced to a stationary process by detrending.

Definition 4.37 (Trend Stationary)**.** A stochastic process, $\{y_t\}$ is trend stationary if there exists a nontrivial function $g(t, \boldsymbol{\delta})$ such that $\{y_t - g(t, \boldsymbol{\delta})\}$ is stationary.

Detrended data may be strictly or covariance stationary (or both).

4.10.2.1 Modeling the time trend in GDP

U.S. GDP data was taken from FRED II from Q1 1947 until Q2 July 2008. To illustrate the use of a time trend, consider two simple models for the level of GDP. The first models the level as a quadratic function of time while the second models the natural log of GDP in an exponential trend model.

$$GDP_t = \phi_0 + \delta_1 t + \delta_2 t^2 + \epsilon_t$$

and

$$\ln GDP_t = \phi_0 + \delta_1 t + \epsilon_t.$$

Figure 4.8 presents the time series of GDP, the log of GDP and errors from two models that include trends. Neither time trend appears to remove the extreme persistence in GDP which may indicate the process contains a unit root.

4.10.3 Unit Roots

Unit root processes are generalizations of the classic random walk. A process is said to have a unit root if the distributed lag polynomial can be factored so that one of the roots is exactly one.

Definition 4.38 (Unit Root)**.** A stochastic process, $\{y_t\}$, is said to contain a unit root if

$$(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_P L^P) y_t = \phi_0 + (1 - \theta_1 L - \theta_2 L^2 - \ldots - \theta_Q L^Q)\epsilon_t \tag{4.85}$$

can be factored

$$(1 - L)(1 - \tilde{\phi}_1 L - \tilde{\phi}_2 L^2 - \ldots - \tilde{\phi}_{P-1} L^{P-1}) y_t = \phi_0 + (1 - \theta_1 L - \theta_2 L^2 - \ldots - \theta_Q L^Q)\epsilon_t. \tag{4.86}$$

The simplest example of a unit root process is a random walk.

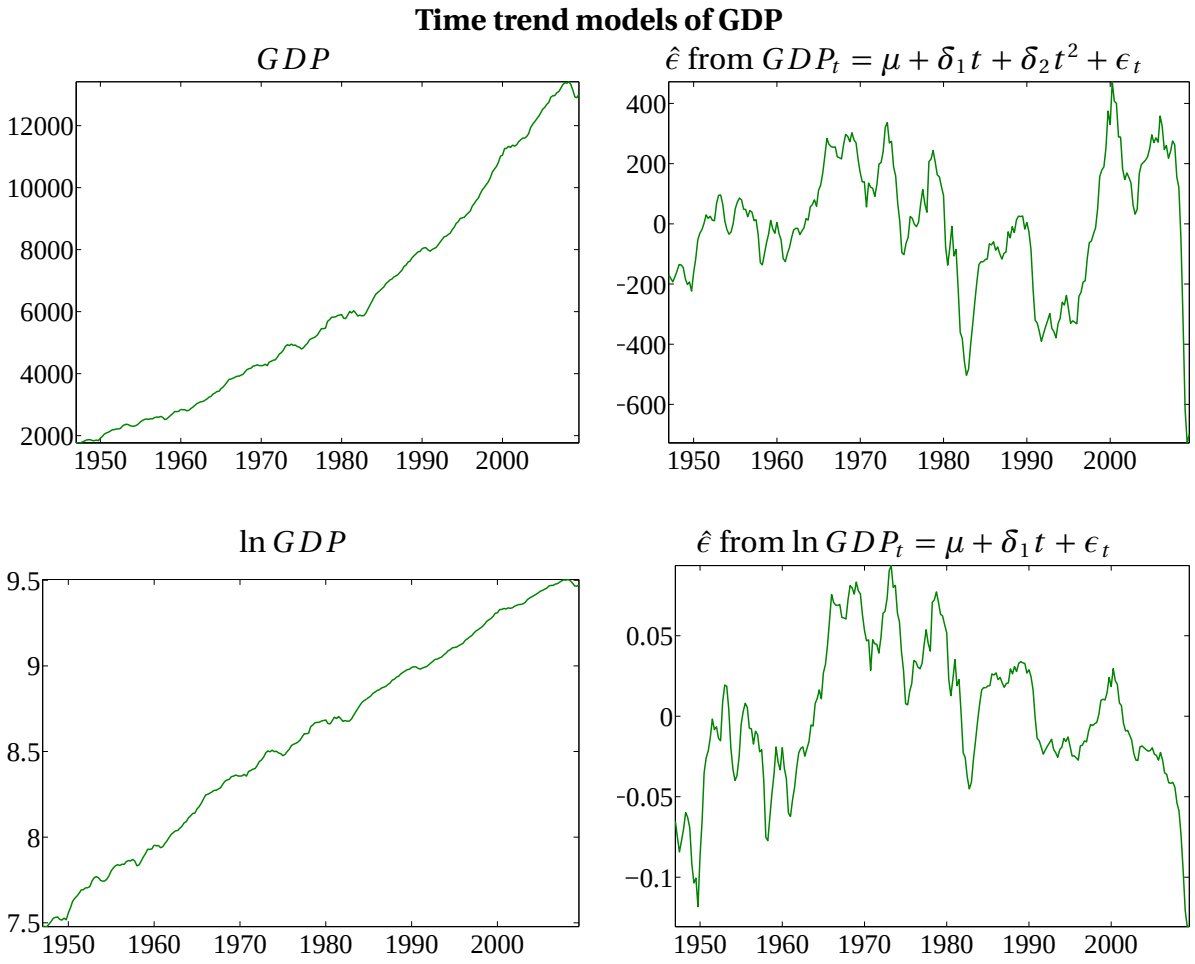

Figure 4.8: Two time trend models are presented, one on the levels of GDP and one on the natural log. Note that the detrended residuals are still highly persistent. This is a likely sign of a unit root.

Definition 4.39 (Random Walk)**.** A stochastic process $\{y_t\}$ is known as a random walk if

$$y_t = y_{t-1} + \epsilon_t \tag{4.87}$$

where ϵ_t is a white noise process with the additional property that $\mathrm{E}_{t-1}[\epsilon_t] = 0$.

The basic properties of a random walk are simple to derive. First, a random walk is a martingale since $\mathrm{E}_t[y_{t+h}] = y_t$ for any h.[13] The variance of a random walk can be deduced from

$$\begin{aligned} \mathrm{V}[y_t] &= \mathrm{E}[(y_t - y_0)^2] \\ &= \mathrm{E}[(\epsilon_t + y_{t-1} - y_0)^2] \\ &= \mathrm{E}[(\epsilon_t + \epsilon_{t-1} + y_{t-2} - y_0)^2] \end{aligned} \tag{4.88}$$

[13] Since the effect of an innovation never declines in a unit root process, it is not reasonable to consider the infinite past as in a stationary AR(1).

$$\begin{aligned}
&= \mathrm{E}[(\epsilon_t + \epsilon_{t-1} + \ldots + \epsilon_1)^2] \\
&= \mathrm{E}[\epsilon_t^2 + \epsilon_{t-1}^2 + \ldots + \epsilon_1^2] \\
&= t\sigma^2
\end{aligned}$$

and this relationship holds for any time index, and so $\mathrm{V}[y_s] = s\sigma^2$. The s^{th} autocovariance (γ_s) of a unit root process is given by

$$\begin{aligned}
\mathrm{V}[(y_t - y_0)(y_{t-s} - y_0)] &= \mathrm{E}[(\epsilon_t + \epsilon_{t-1} + \ldots + \epsilon_1)(\epsilon_{t-s} + \epsilon_{t-s-1} + \ldots + \epsilon_1)] \\
&= \mathrm{E}[(\epsilon_{t-s}^2 + \epsilon_{t-s-1}^2 + \ldots + \epsilon_1^2] \\
&= (t-s)\sigma^2
\end{aligned} \tag{4.89}$$

and the s^{th} autocorrelation is then

$$\rho_s = \frac{t-s}{t} \tag{4.90}$$

which tends to 1 for large t and fixed s. This is a useful property of a random walk process (and any unit root process): The autocorrelations will be virtually constant at 1 with only a small decline at large lags. Building from the simple unit root, one can define a unit root plus drift model,

$$y_t = \delta + y_{t-1} + \epsilon_t$$

which can be equivalently expressed

$$y_t = \delta t + \sum_{i=1}^{t} \epsilon_i + y_0$$

and so the random walk plus drift process consists of both a deterministic trend and a random walk. Alternatively, a random walk model can be augmented with stationary noise so that

$$y_t = \sum_{i=1}^{t} \epsilon_i + \eta_t$$

which leads to the general class of random walk models plus stationary noise processes

$$\begin{aligned}
y_t &= \sum_{i=1}^{t} \epsilon_i + \sum_{j=1}^{t-1} \theta_j \eta_{t-j} + \eta_t \\
&= \sum_{i=1}^{t} \epsilon_i + \Theta(L)\eta_t
\end{aligned}$$

where $\Theta(L)\eta_t = \sum_{j=1}^{t-1} \theta_j \eta_{t-j} + \eta_t$ is a compact expression for a lag polynomial in θ. Since $\Theta(L)\eta_t$ can include any covariance stationary process, this class should be considered general. More importantly, this process has two components: a permanent one, $\sum_{i=1}^{t} \epsilon_i$ and a transitory one $\Theta(L)\eta_t$. The permanent behaves similarly to a deterministic time trend, although unlike the

deterministic trend model, the permanent component of this specification depends on random increments. For this reason, it is known as a *stochastic trend*.

Like the deterministic model, where the process can be detrended, a process with a unit root can be stochastically detrended, or *differenced*, $\Delta y_t = y_t - y_{t-1}$. Differencing a random walk produces a stationary series,

$$\begin{aligned} y_t - y_{t-1} &= \sum_{i=1}^{t} \epsilon_i + \Theta(L)\eta_t - \sum_{i=1}^{t-1} \epsilon_i + \Theta(L)\eta_{t-1} \\ \Delta y_t &= \epsilon_t + (1-L)\Theta(L)\eta_t \end{aligned}$$

Over-differencing occurs when the difference operator is applied to a stationary series. While over-differencing cannot create a unit root, it does have negative consequences such as increasing the variance of the residual and reducing the magnitude of possibly important dynamics. Finally, unit root processes are often known as I(1) processes.

Definition 4.40 (Integrated Process of Order 1)**.** A stochastic process $\{y_t\}$ is integrated of order 1, written $I(1)$, if $\{y_t\}$ is non-covariance-stationary and if $\{\Delta y_t\}$ is covariance stationary. Note: A process that is already covariance stationary is said to be $I(0)$.

The expression integrated is derived from the presence of $\sum_{i=1}^{t} \epsilon_i$ in a unit root process where the sum operator is the discrete version of an integrator.

4.10.4 Difference or Detrend?

Detrending removes nonstationarities from deterministically trending series while differencing removes stochastic trends from unit roots. What happens if the wrong type of detrending is used? The unit root case is simple, and since the trend is stochastic, no amount of detrending can eliminate the permanent component. Only knowledge of the stochastic trend at an earlier point in time can transform the series to be stationary.

Differencing a stationary series produces another series which is stationary but with a larger variance than a detrended series.

$$\begin{aligned} y_t &= \delta t + \epsilon_t \\ \Delta y_t &= \delta + \epsilon_t - \epsilon_{t-1} \end{aligned}$$

while the properly detrended series would be

$$y_t - \delta t = \epsilon_t$$

If ϵ_t is a white noise process, the variance of the differenced series is twice that of the detrended series with a large negative MA component. The parsimony principle dictates that the correctly detrended series should be preferred even though differencing is a viable method of transforming a nonstationary series to be stationary. Higher orders of time trends can be eliminated by re-differencing at the cost of even higher variance.

4.10.5 Testing for Unit Roots: The Dickey-Fuller Test and the Augmented DF Test

Dickey-Fuller tests (DF), and their generalization to augmented Dickey-Fuller tests (ADF) are the standard test for unit roots. Consider the case of a simple random walk,

$$y_t = y_{t-1} + \epsilon_t$$

so that

$$\Delta y_t = \epsilon_t.$$

Dickey and Fuller noted that if the null of a unit root were true, then

$$y_t = \phi_1 y_{t-1} + \epsilon_t$$

can be transformed into

$$\Delta y_t = \gamma y_{t-1} + \epsilon_t$$

where $\gamma = \phi - 1$ and a test could be conducted for the null $H_0 : \gamma = 0$ against an alternative $H_1 : \gamma < 0$. This test is equivalent to testing whether $\phi = 1$ in the original model. $\hat{\gamma}$ can be estimated using a simple regression of Δy_t on y_{t-1}, and the t-stat can be computed in the usual way. If the distribution of $\hat{\gamma}$ were standard normal (under the null), this would be a very simple test. Unfortunately, it is non-standard since, under the null, y_{t-1} is a unit root and the variance is growing rapidly as the number of observations increases. The solution to this problem is to use the Dickey-Fuller distribution rather than the standard normal to make inference on the t-stat of $\hat{\gamma}$.

Dickey and Fuller considered three separate specifications for their test,

$$\begin{aligned} \Delta y_t &= \gamma y_{t-1} + \epsilon_t \\ \Delta y_t &= \phi_0 + \gamma y_{t-1} + \epsilon_t \\ \Delta y_t &= \phi_0 + \delta_1 t + \gamma y_{t-1} + \epsilon_t \end{aligned} \tag{4.91}$$

which correspond to a unit root, a unit root with a linear time trend, and a unit root with a quadratic time trend. The null and alternative hypotheses are the same: $H_0 : \gamma = 0$, $H_1 : \gamma < 0$ (one-sided alternative), and the null that y_t contains a unit root will be rejected if $\hat{\gamma}$ is sufficiently negative, which is equivalent to $\hat{\phi}$ being significantly less than 1 in the original specification.

Unit root testing is further complicated since the inclusion of deterministic regressor(s) affects the asymptotic distribution. For example, if $T = 200$, the critical values of a Dickey-Fuller distribution are

	No trend	Linear	Quadratic
10%	-1.66	-2.56	-3.99
5%	-1.99	-2.87	-3.42
1%	-2.63	-3.49	-3.13

The Augmented Dickey-Fuller (ADF) test generalized the DF to allow for short-run dynamics in the differenced dependent variable. The ADF is a DF regression augmented with lags of the differenced dependent variable to capture short-term fluctuations around the stochastic trend,

$$\Delta y_t = \gamma y_{t-1} + \sum_{p=1}^{P} \phi_p \Delta y_{t-p} + \epsilon_t \tag{4.92}$$

$$\Delta y_t = \phi_0 + \gamma y_{t-1} + \sum_{p=1}^{P} \phi_p \Delta y_{t-p} + \epsilon_t$$

$$\Delta y_t = \phi_0 + \delta_1 t + \gamma y_{t-1} + \sum_{p=1}^{P} \phi_p \Delta y_{t-p} + \epsilon_t$$

Neither the null and alternative hypotheses nor the critical values are changed by the inclusion of lagged dependent variables. The intuition behind this result stems from the observation that the Δy_{t-p} are "less integrated" than y_t and so are asymptotically less informative.

4.10.6 Higher Orders of Integration

In some situations, integrated variables are not just I(1) but have a higher order or integration. For example, the log consumer price index (ln CPI) is often found to be I(2) (integrated of order 2) and so *double differencing* is required to transform the original data into a stationary series. As a consequence, both the level of ln CPI and its difference (inflation) contain unit roots.

Definition 4.41 (Integrated Process of Order d)**.** A stochastic process $\{y_t\}$ is integrated of order d, written $I(d)$, if $\{(1-L)^d y_t\}$ is a covariance stationary ARMA process.

Testing for higher orders of integration is simple: repeat the DF or ADF test on the differenced data. Suppose that it is not possible to reject the null that the level of a variable, y_t, is integrated and so the data should be differenced (Δy_t). If the differenced data rejects a unit root, the testing procedure is complete and the series is consistent with an I(1) process. If the differenced data contains evidence of a unit root, the data should be double differenced ($\Delta^2 y_t$) and the test repeated. The null of a unit root should be rejected on the double-differenced data since no economic data are thought to be I(3), and so if the null cannot be rejected on double-differenced data, careful checking for omitted deterministic trends or other serious problems in the data is warranted.

4.10.6.1 Power of Unit Root tests

Recall that the power of a test is 1 minus the probability Type II error, or simply the probability that the null is rejected when the null is false. In the case of a unit root, power is the ability of a test to reject the null that the process contains a unit root when the largest characteristic root is less than 1. Many economic time-series have roots close to 1 and so it is important to maximize the power of a unit root test so that models have the correct order of integration.

DF and ADF tests are known to be very sensitive to misspecification and, in particular, have very low power if the ADF specification is not flexible enough to account for factors other than

the stochastic trend. Omitted deterministic time trends or insufficient lags of the differenced dependent variable both lower the power by increasing the variance of the residual. This works analogously to the classic regression testing problem of having a low power test when the residual variance is too large due to omitted variables.

A few recommendations can be made regarding unit root tests:

- Use a loose model selection criteria to choose the lag length of the included differenced dependent variables (e.g., AIC).
- Including extraneous deterministic regressors lowers power, but failing to include relevant deterministic regressors produces a test with no power, even asymptotically, and so be conservative when excluding deterministic regressors.
- More powerful tests than the ADF are available. Specifically, DF-GLS of Elliott, Rothenberg, and Stock (1996) is increasingly available and it has maximum power against certain alternatives.
- Trends tend to be obvious and so always plot both the data and the differenced data.
- Use a general-to-specific search to perform unit root testing. Start from a model which should be too large. If the unit root is rejected, one can be confident that there is not a unit root since this is a low power test. If a unit root cannot be rejected, reduce the model by removing insignificant deterministic components first since these lower power without affecting the t-stat. If all regressors are significant, and the null still cannot be rejected, then conclude that the data contains a unit root.

4.10.7 Example: Unit root testing

Two series will be examined for unit roots: the default spread and the log U.S. consumer price index. The $\ln CPI$, which measure consumer prices index less energy and food costs (also known as core inflation), has been taken from FRED, consists of quarterly data and covers the period between August 1968 and August 2008. Figure 4.9 contains plots of both series as well as the first and second differences of $\ln CPI$.

$\ln CPI$ is trending and the spread does not have an obvious time trend. However, deterministic trends should be over-specified and so the initial model for $\ln CPI$ will include both a constant and a time-trend and the model for the spread will include a constant. The lag length used in the model was automatically selected using the BIC.

Results of the unit root tests are presented in table 4.4. Based on this output, the spreads reject a unit root at the 5% level but the $\ln CPI$ cannot. The next step is to difference the $\ln CPI$ to produce Δ lnCPI. Rerunning the ADF test on the differenced CPI (inflation) and including either a constant or no deterministic trend, the null of a unit root still cannot be rejected. Further differencing the data, $\Delta^2 \ln CPI_t = \delta \ln CPI_t - \ln CPI_{t-1}$, strongly rejects, and so $\ln CPI$ appears to be I(2). The final row of the table indicates the number of lags used in the ADF and was selected using the BIC with a maximum of 12 lags for $\ln CPI$ or 36 lags for the spread (3 years).

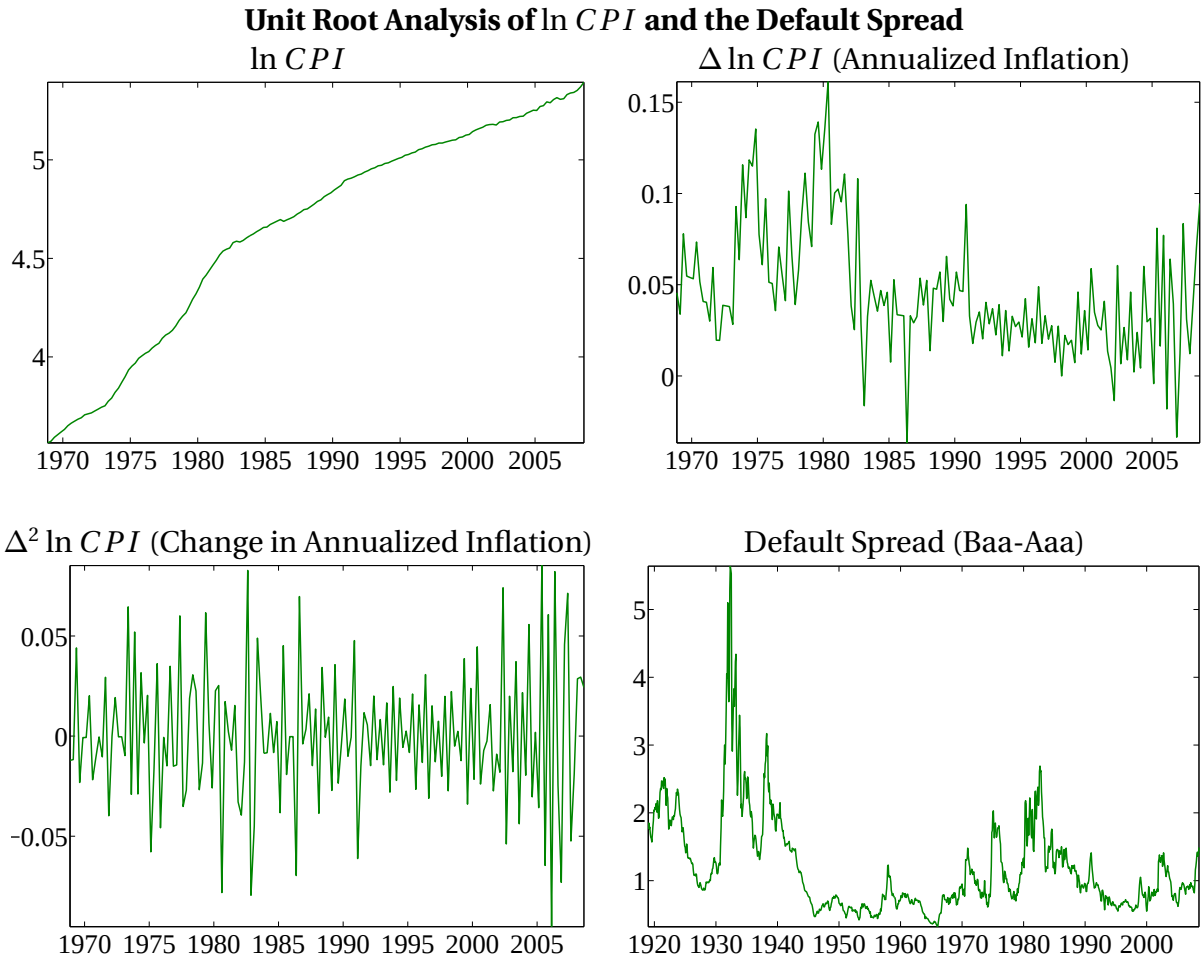

Figure 4.9: These four panels plot the log consumer price index (ln CPI), $\Delta \ln CPI$, $\Delta^2 \ln CPI$ and the default spread. Both $\Delta^2 \ln CPI$ and the default spread reject the null of a unit root.

	ln CPI	ln CPI	ln CPI	$\Delta \ln CPI$	$\Delta \ln CPI$	$\Delta^2 \ln CPI$	Default Sp.	Default Sp.
t-stat	-2.119	-1.541	1.491	-2.029	-0.977	-13.535	-3.130	-1.666
p-val	0.543	0.517	0.965	0.285	0.296	0.000	0.026	0.091
Deterministic	Linear	Const.	None	Const.	None	None	Const.	None
# lags	4	4	4	3	3	2	15	15

Table 4.4: ADF results for tests that ln CPI and the default spread have unit roots. The null of an unit root cannot be rejected in ln CPI , nor can the null that Δ lnCPI contains a unit root and so CPI appears to be an I(2) process. The default spread rejects the null of a unit root although clearly highly persistent.

4.11 Nonlinear Models for Time-Series Analysis

While this chapter has exclusively focused on linear time-series processes, many non-linear time-series processes have been found to parsimoniously describe important dynamics in financial data. Two which have proven particularly useful in the analysis of financial data are Markov Switching Autoregressions (MSAR) and Threshold Autoregressions (TAR), especially the subclass of Self-Exciting Threshold Autoregressions (SETAR).[14]

4.12 Filters

The ultimate goal of most time-series modeling is to forecast a time-series in its entirety, which requires a model for both permanent and transitory components. In some situations, it may be desirable to focus on either the short-run dynamics or the long-run dynamics exclusively, for example in technical analysis where prices are believed to be long-run unpredictable but may have some short- or medium-run predictability. Linear filters are a class of functions which can be used to "extract" a stationary cyclic component from a time-series which contains both short-run dynamics and a permanent component. Generically, a filter for a time series $\{y_t\}$ is defined as

$$x_t = \sum_{i=-\infty}^{\infty} w_i y_{t-i} \tag{4.93}$$

where x_t is the filtered time-series or filter output. In most applications, it is desirable to assign a label to x_t, either a trend, written τ_t, or a cyclic component, c_t.

Filters are further categorized into *causal* and *non-causal*, where causal filters are restricted to depend on only the past and present of y_t, and so as a class are defined through

$$x_t = \sum_{i=0}^{\infty} w_i y_{t-i}. \tag{4.94}$$

Causal filters are more useful in isolating trends from cyclical behavior for forecasting purposes while non-causal filters are more useful for historical analysis of macroeconomic and financial data.

4.12.1 Frequency, High- and Low-Pass Filters

This text has exclusively dealt with time series in the *time domain* – that is, understanding dynamics and building models based on the time distance between points. An alternative strategy for describing a time series is in terms of *frequencies* and the magnitude of the cycle at a given frequency. For example, suppose a time series has a cycle that repeats every 4 periods. This series could be equivalently described as having a cycle that occurs with a frequency of 1 in 4, or

[14] There are many nonlinear models frequently used in financial econometrics for modeling quantities *other* than the conditional mean. For example, both the ARCH (conditional volatility) and CaViaR (conditional Value-at-Risk) models are nonlinear in the original data.

.25. A frequency description is relatively compact – it is only necessary to describe a process from frequencies of 0 to 0.5, the latter of which would be a cycle with a period of 2.[15]

The idea behind filtering is to choose a set of frequencies and then to isolate the cycles which occur within the selected frequency range. Filters that eliminate high-frequency cycles are known as *low-pass filters*, while filters that eliminate low-frequency cycles are known as *high-pass filters*. Moreover, high- and low-pass filters are related in such a way that if $\{w_i\}$ is a set of weights corresponding to a high-pass filter, $v_0 = 1 - w_0$, $v_i = -w_i$ $i \neq 0$ is a set of weights corresponding to a low-pass filter. This relationship forms an identity since $\{v_i + w_i\}$ must correspond to an *all-pass* filter since $\sum_{i=-\infty}^{\infty}(v_i + w_i)y_{t-1} = y_t$ for any set of weights.

The goal of a filter is to select a particular frequency range and nothing else. The *gain function* describes the amount of attenuations which occurs at a given frequency.[16] A gain of 1 at a particular frequency means any signal at that frequency is passed through unmodified while a gain of 0 means that the signal at that frequency is eliminated from the filtered data. Figure 4.10 contains a graphical representation of the gain function for a set of *ideal filters*. The four panels show an all-pass (all frequencies unmodified), a low-pass filter with a cutoff frequency of $\frac{1}{10}$, a high-pass with a cutoff frequency of $\frac{1}{6}$, and a band-pass filter with cutoff frequencies of $\frac{1}{6}$ and $\frac{1}{32}$.[17] In practice, only the all-pass filter (which corresponds to a filter with weights $w_0 = 1$, $w_i = 0$ for $i \neq 0$) can be constructed using a finite sum, and so applied filtering must make trade-offs.

4.12.2 Moving Average and Exponentially Weighted Moving Average (EWMA)

Moving averages are the simplest filters and are often used in technical analysis. Moving averages can be constructed as both causal and non-causal filters.

Definition 4.42 (Causal Moving Average)**.** A causal moving average (MA) is a function which takes the form

$$\tau_t = \frac{1}{n}\sum_{i=1}^{n} y_{t-i+1}.$$

Definition 4.43 (Centered (Non-Causal) Moving Average)**.** A centered moving average (MA) is a function which takes the form

$$\tau_t = \frac{1}{2n+1}\sum_{i=-n}^{n} y_{t-i+1}.$$

Note that the centered MA is an average over $2n+1$ data points.

[15] The frequency $\frac{1}{2}$ is known as the *Nyquist* frequency since it is not possible to measure any cyclic behavior at frequencies above $\frac{1}{2}$ since these would have a cycle of 1 period and so would appear constant.

[16] The gain function for any filter of the form $x_t = \sum_{i=-\infty}^{\infty} w_i y_{t-i}$ can be computed as

$$G(f) = \left| \sum_{k=-\infty}^{\infty} w_j \exp(-ik2\pi f) \right|$$

where $i = \sqrt{-1}$.

[17] Band-pass filters are simply the combination of two low-pass filters. Specifically, if $\{w_i\}$ is set of weights from a low-pass filter with a cutoff frequency of f_1 and $\{v_i\}$ is a set of weights from a low-pass filter with a cutoff frequency of f_2, $f_2 > f_1$, then $\{v_i - w_i\}$ is a set of weights which corresponds to a band-pass filter with cutoffs at f_1 and f_2.

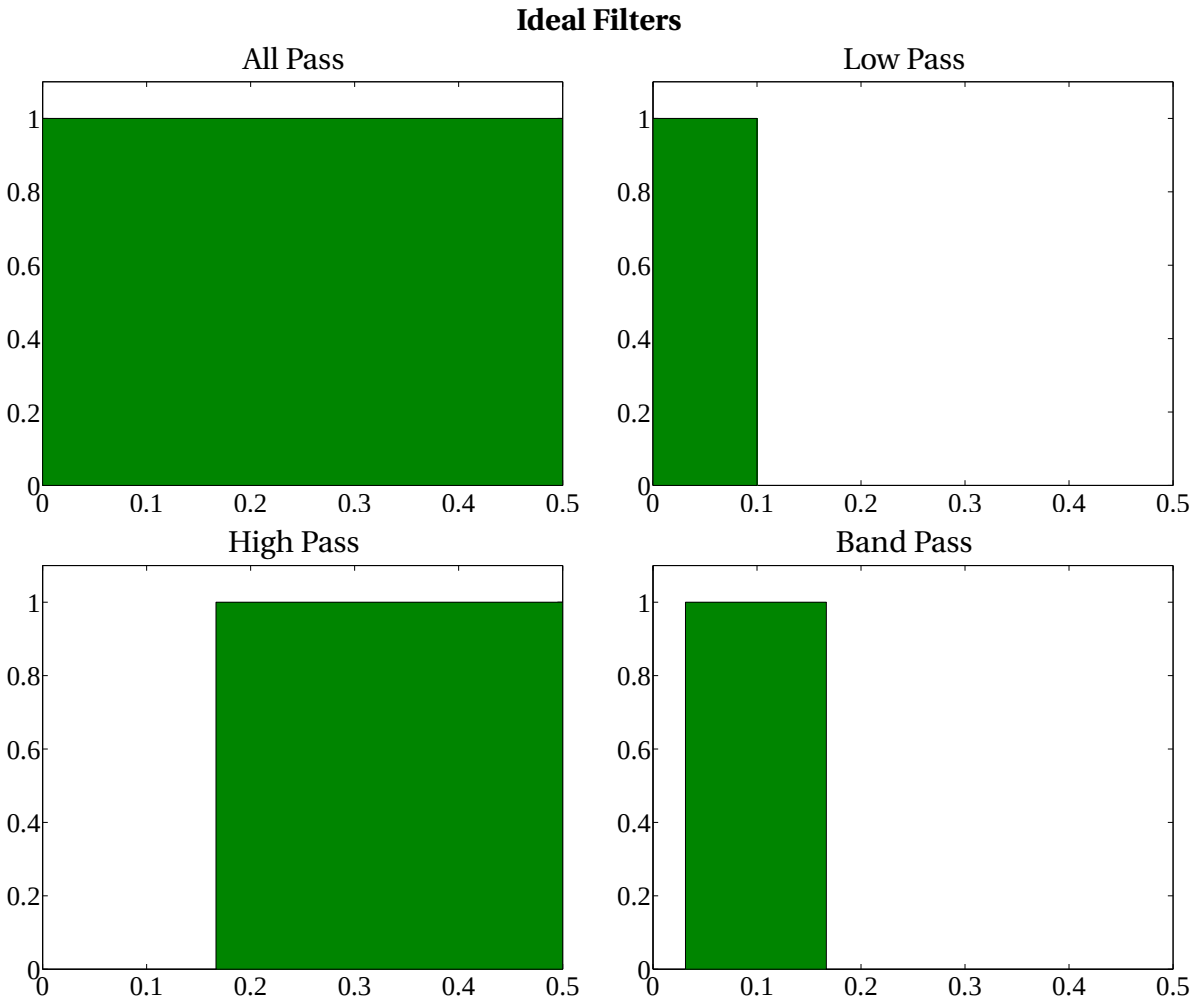

Figure 4.10: These four panels depict the gain functions from a set of *ideal filters*. The all-pass filter allows all frequencies through. The low-pass filter cuts off at $\frac{1}{10}$. The high-pass cuts off below $\frac{1}{6}$ and the band-pass filter cuts off below $\frac{1}{32}$ and above $\frac{1}{6}$.

Moving averages are low-pass filters since their weights add up to 1. In other words, the moving average would contain the permanent component of $\{y_t\}$ and so would have the same order of integration. The cyclic component, $c_t = y_t - \tau_t$, would have a lower order of integration that y_t. Since moving averages are low-pass filters, the difference of two moving averages must be a band-pass filter. Figure 4.11 contains the gain function from the difference between a 20-day and 50-day moving average which is commonly used in technical analysis.

Exponentially Weighted Moving Averages (EWMA) are a close cousin of the MA which places greater weight on recent data than on past data.

Definition 4.44 (Exponentially Weighted Moving Average)**.** A exponentially weighed moving average (EWMA) is a function which takes the form

$$\tau_t = (1-\lambda)\sum_{i=0}^{\infty}\lambda^i y_{t-i}$$

for some $\lambda \in (0,1)$.

The name EWMA is derived from the exponential decay in the weights, and EWMAs can be equivalently expressed (up to an initial condition) as

$$\tau_t = (1-\lambda)\lambda y_t + \lambda \tau_{t-1}.$$

Like MAs, EWMAs are low-pass filters since the weights sum to 1.

EWMAs are commonly used in financial applications as volatility filters, where the dependent variable is chosen to be the squared return. The difference between two EWMAs is often referred to as a Moving Average Convergence Divergence (MACD) filter in technical analysis. MACDs are indexed by two numbers, a fast period and a slow period, where the number of data in the MACD can be converted to λ as $\lambda = (n-1)/(n+1)$, and so an MACD(12,26) is the difference between two EWMAs with parameters .842 and .926. 4.11 contains the gain function from an MACD(12,26) (the difference between two EWMAs), which is similar to, albeit smoother than, the gain function from the difference of a 20-day and a 50-day MAs.

4.12.3 Hodrick-Prescott Filter

The Hodrick and Prescott (1997) (HP) filter is constructed by balancing the fitting the trend to the data and the smoothness of the trend. The HP filter is defined as the solution to

$$\min_{\tau_t} \sum_{t=1}^{T} (y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1} ((\tau_{t-1} - \tau_t) - (\tau_t + \tau_{t-1}))$$

where $(\tau_{t-1} - \tau_t) - (\tau_t + \tau_{t-1})$ can be equivalently expressed as the second-difference of τ_t, $\Delta^2 \tau_t$. λ is a smoothness parameter: if $\lambda = 0$ then the solution to the problem is $\tau_t = y_t \; \forall t$, and as $\lambda \to \infty$, the "cost" of variation in $\{\tau_t\}$ becomes arbitrarily high and $\tau_t = \beta_0 + \beta_1 t$ where β_0 and β_1 are the least squares fit of a linear trend model to the data.

It is simple to show that the solution to this problem must have

$$\mathbf{y} = \mathbf{\Gamma}\boldsymbol{\tau}$$

where $\mathbf{\Gamma}$ is a band-diagonal matrix (all omitted elements are 0) of the form

$$\mathbf{\Gamma} = \begin{bmatrix}
1+\lambda & -2\lambda & \lambda & & & & & & & \\
-2\lambda & 1+5\lambda & -4\lambda & \lambda & & & & & & \\
\lambda & -4\lambda & 1+6\lambda & -4\lambda & \lambda & & & & & \\
 & \lambda & -4\lambda & 1+6\lambda & -4\lambda & \lambda & & & & \\
 & & \ddots & \ddots & \ddots & \ddots & \ddots & & & \\
 & & & & \lambda & -4\lambda & 1+6\lambda & -4\lambda & \lambda & \\
 & & & & & \lambda & -4\lambda & 1+6\lambda & -4\lambda & \lambda \\
 & & & & & & \lambda & -4\lambda & 1+5\lambda & -2\lambda \\
 & & & & & & & \lambda & -2\lambda & 1+\lambda
\end{bmatrix}$$

and The solution to this set of T equations in T unknowns is

$$\boldsymbol{\tau} = \boldsymbol{\Gamma}^{-1}\mathbf{y}.$$

The cyclic component is defined as $c_t = y_t - \tau_t$.

Hodrick and Prescott (1997) recommend values of 100 for annual data, 1600 for quarterly data and 14400 for monthly data. The HP filter is non-causal and so is not appropriate for prediction. The gain function of the cyclic component of the HP filter with $\lambda = 1600$ is illustrated in figure 4.11. While the filter attempts to eliminate components with a frequency below ten years of quarterly data ($\frac{1}{40}$), there is some gain until about $\frac{1}{50}$ and the gain is not unity until approximately $\frac{1}{25}$.

4.12.4 Baxter-King Filter

Baxter and King (1999) consider the problem of designing a filter to be close to the ideal filter subject to using a finite number of points.[18] They further argue that extracting the cyclic component requires the use of both a high-pass and a low-pass filter – the high-pass filter is to cutoff the most persistent components while the low-pass filter is used to eliminate high-frequency noise. The BK filter is defined by a triple, two-period lengths (inverse frequencies) and the number of points used to construct the filter (k), and is written as $BK_k(p,q)$ where $p < q$ are the cutoff frequencies.

Baxter and King suggest using a band-pass filter with cutoffs at $\frac{1}{32}$ and $\frac{1}{6}$ for quarterly data. The final choice for their approximate ideal filter is the number of nodes, for which they suggest 12. The number of points has two effects. First, the BK filter cannot be used in the first and last k points. Second, a higher number of nodes will produce a more accurate approximation to the ideal filter.

Implementing the BK filter is simple. Baxter and King show that the optimal weights for a low-pass filter at particular frequency f, satisfy

$$\tilde{w}_0 = 2f \tag{4.95}$$

$$\tilde{w}_i = \frac{\sin(2i\pi f)}{i\pi}, \quad i = 1, 2, \ldots, k \tag{4.96}$$

$$\theta = [2k+1]^{-1}\left(1 - \sum_{i=-k}^{k} \tilde{w}_i\right) \tag{4.97}$$

$$w_i = \tilde{w}_i + \theta, \quad i = 0, 1, \ldots, k \tag{4.98}$$

$$w_i = w_{-i}. \tag{4.99}$$

The BK filter is constructed as the difference between two low-pass filters, and so

[18] Ideal filters, except for the trivial all-pass, require an infinite number of points to implement, and so are infeasible in applications.

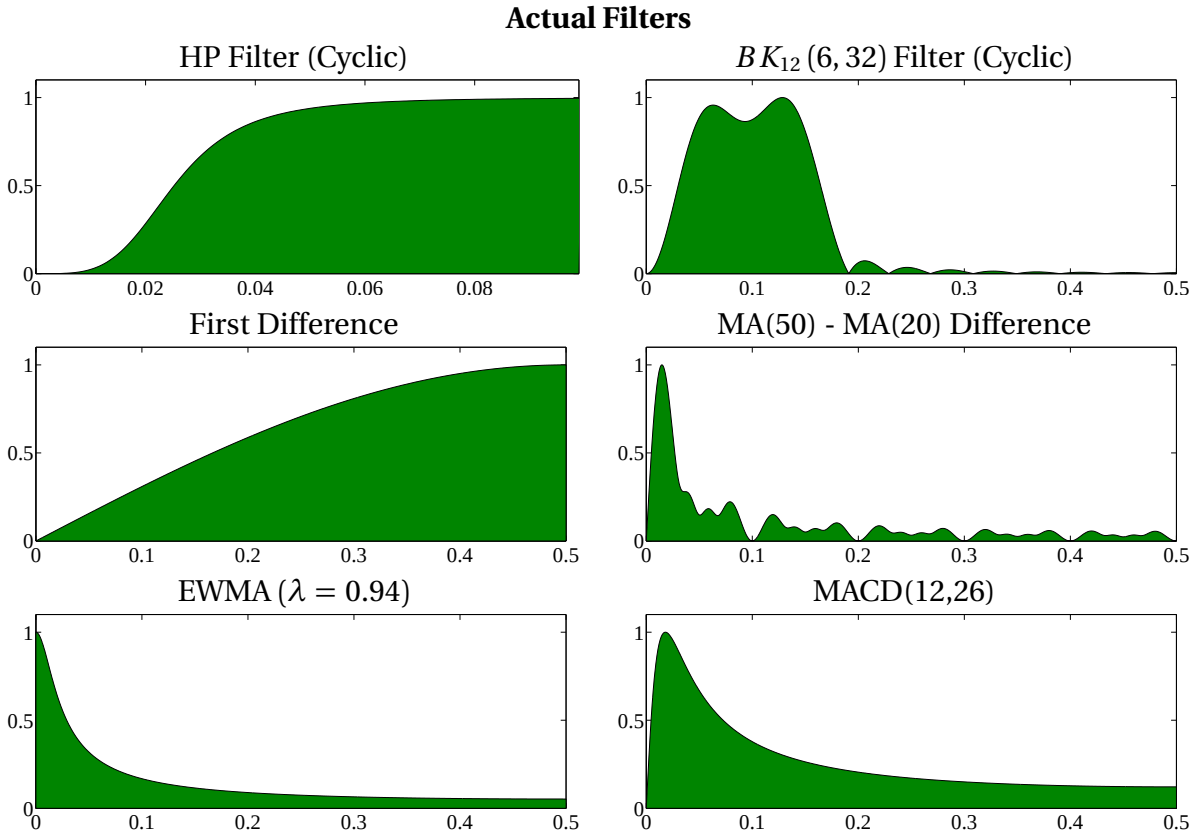

Figure 4.11: These six panels contain the standard HP filter, the $BK_{12}(6,32)$ filter, the first difference filter, an EWMA with $\lambda = .94$, a MACD(12,26) and the difference between a 20-day and a 50-day moving average. The gain functions in the right hand column have been normalized so that the maximum weight is 1. The is equivalent to scaling all of the filter weights by a constant, and so is simple a change in variance of the filter output.

$$\tau_t = \sum_{i=-k}^{k} w_i y_{t-i} \tag{4.100}$$

$$c_t = \sum_{i=-k}^{k} (v_i - w_i) y_{t-i} \tag{4.101}$$

where $\{w_i\}$ and $\{v_i\}$ are both weights from low-pass filters where the period used to construct $\{w_i\}$ is longer than the period used to construct $\{v_i\}$. The gain function of the $BK_{12}(6,32)$ is illustrated in the upper right panel of figure 4.11. The approximation is reasonable, with near unit gain between $\frac{1}{32}$ and $\frac{1}{6}$ and little gain outside.

4.12.5 First Difference

Another very simple filter to separate a "trend" from a "cyclic" component is the first difference. Note that if y_t is an I(1) series, then the first difference which contains the "cyclic" component,

$c_t = \frac{1}{2}\Delta y_t$, is an I(0) series and so the first difference is a causal filter. The "trend" is measured using an MA(2), $\tau_t = \frac{1}{2}(y_t + y_{t-1})$ so that $y_t = c_t + \tau_t$. The FD filter is not sharp – it allows for most frequencies to enter the cyclic component – and so is not recommended in practice.

4.12.6 Beveridge-Nelson Decomposition

The Beveridge and Nelson (1981) decomposition extends the first order difference decomposition to include any predictable component in the future trend as part of the current trend. The idea behind the BN decomposition is simple: if the predictable part of the long-run component places the long-run component above its current value, then the cyclic component should be negative. Similarly, if the predictable part of the long-run component expects that the long run component should trend lower then the cyclic component should be positive. Formally the BN decomposition if defined as

$$\begin{aligned} \tau_t &= \lim_{h\to\infty} \hat{y}_{t+h|t} - h\mu \\ c_t &= y_t - \tau_t \end{aligned} \tag{4.102}$$

where μ is the drift in the trend, if any. The trend can be equivalently expressed as the current level of y_t plus the expected increments minus the drift,

$$\tau_t = y_t + \lim_{h\to\infty} \sum_{i=1}^{h} \mathrm{E}\left[\Delta \hat{y}_{t+i|t} - \mu\right] \tag{4.103}$$

where μ is the unconditional expectation of the increments to y_t, $\mathrm{E}[\Delta \hat{y}_{t+j|t}]$. The trend component contains the persistent component and so the filter applied must be a low-pass filter, while the cyclic component is stationary and so must be the output of a high-pass filter. The gain of the filter applied when using the BN decomposition depends crucially on the forecasting model for the short-run component.

Suppose $\{y_t\}$ is an I(1) series which has both a permanent and transitive component. Since $\{y_t\}$ is I(1), Δy_t must be I(0) and so can be described by a stationary ARMA(P,Q) process. For simplicity, suppose that Δy_t follows an MA(3) so that

$$\Delta y_t = \phi_0 + \theta_1 \epsilon_{t-1} + \theta_1 \epsilon_{t-2} + \theta_3 \epsilon_{t-3} + \epsilon_t$$

In this model $\mu = \phi_0$, and the h-step ahead forecast is given by

$$\begin{aligned} \Delta \hat{y}_{t+1|t} &= \mu + \theta_1 \epsilon_t + \theta_2 \epsilon_{t-1} + \theta_3 \epsilon_{t-2} \\ \Delta \hat{y}_{t+2|t} &= \mu + \theta_2 \epsilon_t + \theta_3 \epsilon_{t-1} \\ \Delta \hat{y}_{t+3|t} &= \mu + \theta_3 \epsilon_t \\ \Delta \hat{y}_{t+h|t} &= \mu \quad h \geq 4, \end{aligned}$$

and so

$$\tau_t = y_t + (\theta_1 + \theta_2 + \theta_3)\epsilon_t + (\theta_2 + \theta_3)\epsilon_{t-1} + \theta_3 \epsilon_{t-2}$$

and

$$c_t = -(\theta_1 + \theta_2 + \theta_3)\epsilon_t - (\theta_2 + \theta_3)\epsilon_{t-1} - \theta_3\epsilon_{t-2}.$$

Alternatively, suppose that Δy_t follows an AR(1) so that

$$\Delta y_t = \phi_0 + \phi_1 \Delta y_{t-1} + \epsilon_t.$$

This model can be equivalently defined in terms of deviations around the long-run mean, $\Delta\tilde{y}_t = \Delta y_t - \phi_0/(1-\phi_1)$, as

$$\begin{aligned}
\Delta y_t &= \phi_0 + \phi_1 \Delta y_{t-1} + \epsilon_t \\
\Delta y_t &= \phi_0 \frac{1-\phi_1}{1-\phi_1} + \phi_1 \Delta y_{t-1} + \epsilon_t \\
\Delta y_t &= \frac{\phi_0}{1-\phi_1} - \phi_1 \frac{\phi_0}{1-\phi_1} + \phi_1 \Delta y_{t-1} + \epsilon_t \\
\Delta y_t - \frac{\phi_0}{1-\phi_1} &= \phi_1 \left(\Delta y_{t-1} - \frac{\phi_0}{1-\phi_1} \right) + \epsilon_t \\
\Delta \tilde{y}_t &= \phi_1 \Delta \tilde{y}_{t-1} + \epsilon_t.
\end{aligned}$$

In this transformed model, $\mu = 0$ which simplifies finding the expression for the trend. The h-step ahead forecast if $\Delta\tilde{y}_t$ is given by

$$\Delta \hat{\tilde{y}}_{t+h|t} = \phi_1^h \Delta \tilde{y}_t$$

and so

$$\begin{aligned}
\tau_t &= y_t + \lim_{h\to\infty} \sum_{i=1}^{h} \Delta \hat{\tilde{y}}_{t+i|t} \\
&= y_t + \lim_{h\to\infty} \sum_{i=1}^{h} \phi_1^i \Delta \tilde{y}_t \\
&= y_t + \lim_{h\to\infty} \Delta \tilde{y}_t \sum_{i=1}^{h} \phi_1^i \\
&= y_t + \lim_{h\to\infty} \Delta \tilde{y}_t \frac{\phi_1}{1-\phi_1}
\end{aligned}$$

which follows since $\lim_{h\to\infty} \sum_{i=1}^{h} \phi_1^i = -1 + \lim_{h\to\infty} \sum_{i=0}^{h} \phi_1^i = 1/(1-\phi_1) - 1$. The main criticism of the Beveridge-Nelson decomposition is that the trend and the cyclic component are perfectly (negatively) correlation.

4.12.7 Extracting the cyclic components from Real US GDP

To illustrate the filters, the cyclic component was extracted from log real US GDP data taken from the Federal Reserve Economics Database. Data was available from 1947 Q1 to Q2 2009. Figure 4.12 contains the cyclical component extracted using 4 methods. The top panel contains the standard HP filter with $\lambda = 1600$. The middle panel contains $BK_{12}(6,32)$ (solid) and $BK_{12}(1,32)$ (dashed) filters, the latter of which is a high pass-filter since the faster frequency is 1. Note that the first and last 12 points of the cyclic component are set to 0. The bottom panel contains the cyclic component extracted using a Beveridge-Nelson decomposition based on an AR(1) fit to GDP growth. For the BN decomposition, the first 2 points are zero which reflects the loss of data due to the first difference and the fitting of the AR(1) to the first difference.[19]

The HP filter and the $BK_{12}(1,32)$ are remarkably similar with a correlation of over 99%. The correlation between the $BK_{12}(6,32)$ and the HP filter was 96%, the difference being in the lack of a high-frequency component. The cyclic component from the BN decomposition has a small negative correlation with the other three filters, although choosing a different model for GDP growth would change the decomposition.

4.12.8 Markov Switching Autoregression

Markov switching autoregression, introduced into econometrics in Hamilton (1989), uses a composite model which evolves according to both an autoregression and a latent state which determines the value of the autoregressive parameters. In financial applications using low-frequency asset returns, an MSAR that allows the mean and the variance to be state-dependent has been found to outperform linear models (Perez-Quiros and Timmermann, 2000).

Definition 4.45 (Markov Switching Autoregression)**.** A k-state Markov switching autoregression (MSAR) is a stochastic process which has dynamics that evolve through both a Markovian state process and an autoregression where the autoregressive parameters are state dependent. The states, labeled $1, 2, \ldots, k$, are denoted s_t and follow a k-state latent Markov Chain with transition matrix $\mathbf{P}$,

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{bmatrix} \tag{4.104}$$

where $p_{ij} = Pr(s_{t+1} = i | s_t = j)$. Note that the columns must sum to 1 since $\sum_{i=1}^{k} Pr(s_{t+1} = i | s_t = j) = 1$. Data are generated according to a P^{th} order autoregression,

$$y_t = \phi_0^{(s_t)} + \phi_1^{(s_t)} y_{t-1} + \ldots + \phi_P^{(s_t)} y_{t-p} + \sigma^{(s_t)} \epsilon_t \tag{4.105}$$

where $\boldsymbol{\phi}^{(s_t)} = [\phi_0^{(s_t)} \phi_1^{(s_t)} \ldots \phi_P^{(s_t)}]'$ are state-dependent autoregressive parameters, $\sigma^{(s_t)}$ is the state-dependent standard deviation and $\epsilon_t \overset{\text{i.i.d.}}{\sim} N(0,1)$.[20] The unconditional state probabilities ($\Pr(s_t = i)$), known as the ergodic probabilities, are denoted $\boldsymbol{\pi} = [\pi_1 \, \pi_2 \, \ldots \, \pi_k]'$ and are the solution to

$$\boldsymbol{\pi} = \mathbf{P}\boldsymbol{\pi}. \tag{4.106}$$

[19] The AR(1) was chosen from a model selection search of AR models with an order up to 8 using the SBIC.

[20] The assumption that $\epsilon_t \overset{\text{i.i.d.}}{\sim} N(0,1)$ can be easily relaxed to include other i.i.d. processes for the innovations.

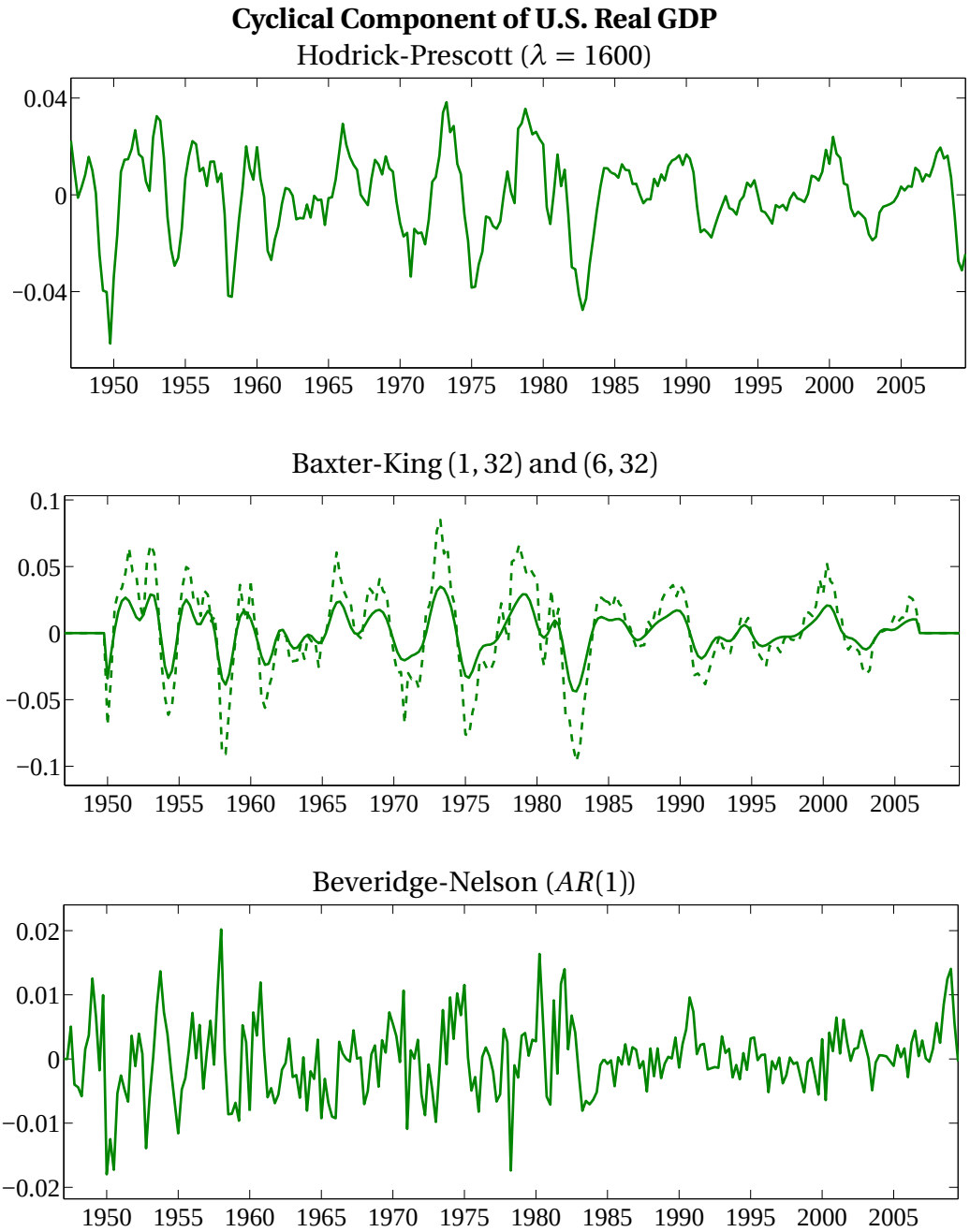

Figure 4.12: The top panel contains the filtered cyclic component from a HP filter with $\lambda = 1600$. The middle panel contains the cyclic component from $BK_{12}(6,32)$ (solid) and $BK_{12}(1,32)$ (dashed) filters. The bottom panel contains the cyclic component from a Beveridge-Nelson decomposition based on an AR(1) model for GDP growth rates.

The ergodic probabilities can also be computed as the normalized eigenvector of **P** corresponding to the only unit eigenvalue.

Rather than attempting to derive properties of an MSAR, consider a simple specification with

two states, no autoregressive terms, and where only the mean of the process varies[21]

$$y_t = \begin{cases} \phi^H + \epsilon_t \\ \phi^L + \epsilon_t \end{cases} \tag{4.107}$$

where the two states are indexed by H (high) and L (low). The transition matrix is

$$\mathbf{P} = \begin{bmatrix} p_{HH} & p_{HL} \\ p_{LH} & p_{LL} \end{bmatrix} = \begin{bmatrix} p_{HH} & 1 - p_{LL} \\ 1 - p_{HH} & p_{LL} \end{bmatrix} \tag{4.108}$$

and the unconditional probabilities of being in the high and low state, π_H and π_L, respectively, are

$$\pi_H = \frac{1 - p_{LL}}{2 - p_{HH} - p_{LL}} \tag{4.109}$$

$$\pi_L = \frac{1 - p_{HH}}{2 - p_{HH} - p_{LL}}. \tag{4.110}$$

This simple model is useful for understanding the data generation in a Markov Switching process:

1. At $t = 0$ an initial state, s_0, is chosen according to the ergodic (unconditional) probabilities. With probability π_H, $s_0 = H$ and with probability $\pi_L = 1 - \pi_H$, $s_0 = L$.

2. The state probabilities evolve independently from the observed data according to a Markov Chain. If $s_0 = H$, $s_1 = H$ with probability p_{HH}, the probability $s_{t+1} = H$ given $s_t = H$ and $s_1 = L$ with probability $p_{LH} = 1 - p_{HH}$. If $s_0 = L$, $s_1 = H$ with probability $p_{HL} = 1 - p_{LL}$ and $s_1 = L$ with probability p_{LL}.

3. Once the state at $t = 1$ is known, the value of y_1 is chosen according to

$$y_1 = \begin{cases} \phi^H + \epsilon_1 & \text{if } s_1 = H \\ \phi^L + \epsilon_t & \text{if } s_1 = L \end{cases}.$$

4. Steps 2 and 3 are repeated for $t = 2, 3, \ldots, T$, to produce a time series of Markov Switching data.

4.12.8.1 Markov Switching Examples

Using the 2-state Markov Switching Autoregression described above, 4 systems were simulated for 100 observations.

- Pure mixture
  - $\mu_H = 4, \mu_L = -2, \mathrm{V}[\epsilon_t] = 1$ in both states
  - $p_{HH} = .5 = p_{LL}$

[21] See Hamilton (1994, chapter 22) or Krolzig (1997) for further information on implementing MSAR models.

  - $\pi_H = \pi_L = .5$
  - Remark: This is a "pure" mixture model where the probability of each state does not depend on the past. This occurs because the probability of going from high to high is the same as the probability of going from low to high, 0.5.

- Two persistent States
  - $\mu_H = 4, \mu_L = -2, \mathrm{V}[\epsilon_t] = 1$ in both states
  - $p_{HH} = .9 = p_{LL}$ so the average duration of each state is 10 periods.
  - $\pi_H = \pi_L = .5$
  - Remark: Unlike the first parameterization this is not a simple mixture. Conditional on the current state being H, there is a 90% chance that the next state will remain H.

- One persistent state, on transitory state
  - $\mu_H = 4, \mu_L = -2, \mathrm{V}[\epsilon_t] = 1$ if $s_t = H$ and $\mathrm{V}[\epsilon_t] = 2$ if $s_t = L$
  - $p_{HH} = .9,\ p_{LL} = .5$
  - $\pi_H = .83,\ \pi_L = .16$
  - Remark: This type of model is consistent with quarterly data on U.S. GDP where booms (H) typically last 10 quarters while recessions die quickly, typically in 2 quarters.

- Mixture with different variances
  - $\mu_H = 4, \mu_L = -2, \mathrm{V}[\epsilon_t] = 1$ if $s_t = H$ and $\mathrm{V}[\epsilon_t] = 16$ if $s_t = L$
  - $p_{HH} = .5 = p_{LL}$
  - $\pi_H = \pi_L = .5$
  - Remark: This is another "pure" mixture model but the variances differ between the states. One nice feature of mixture models (MSAR is a member of the family of mixture models) is that the unconditional distribution of the data may be non-normal even though the shocks are conditionally normally distributed.[22]

Figure 4.13 contains plots of 100 data points generated from each of these processes. The first (MSAR(1)) produces a mixture with modes at -2 and 4 each with equal probability and the states (top panel, bottom right) are i.i.d.. The second process produced a similar unconditional distribution but the state evolution is very different. Each state is very persistent and, conditional on the state being high or low, it was likely to remain the same. The third process had one very persistent state and one with much less persistence. This produced a large skew in the unconditional distribution since the state where $\mu = -2$ was visited less frequently than the state with $\mu = 4$. The final process (MSAR(4)) has state dynamics similar to the first but produces a very different unconditional distribution. The difference occurs since the variance depends on the state of the Markov process.

[22] Mixtures of finitely many normals, each with different means and variances, can be used approximate many non-normal distributions.

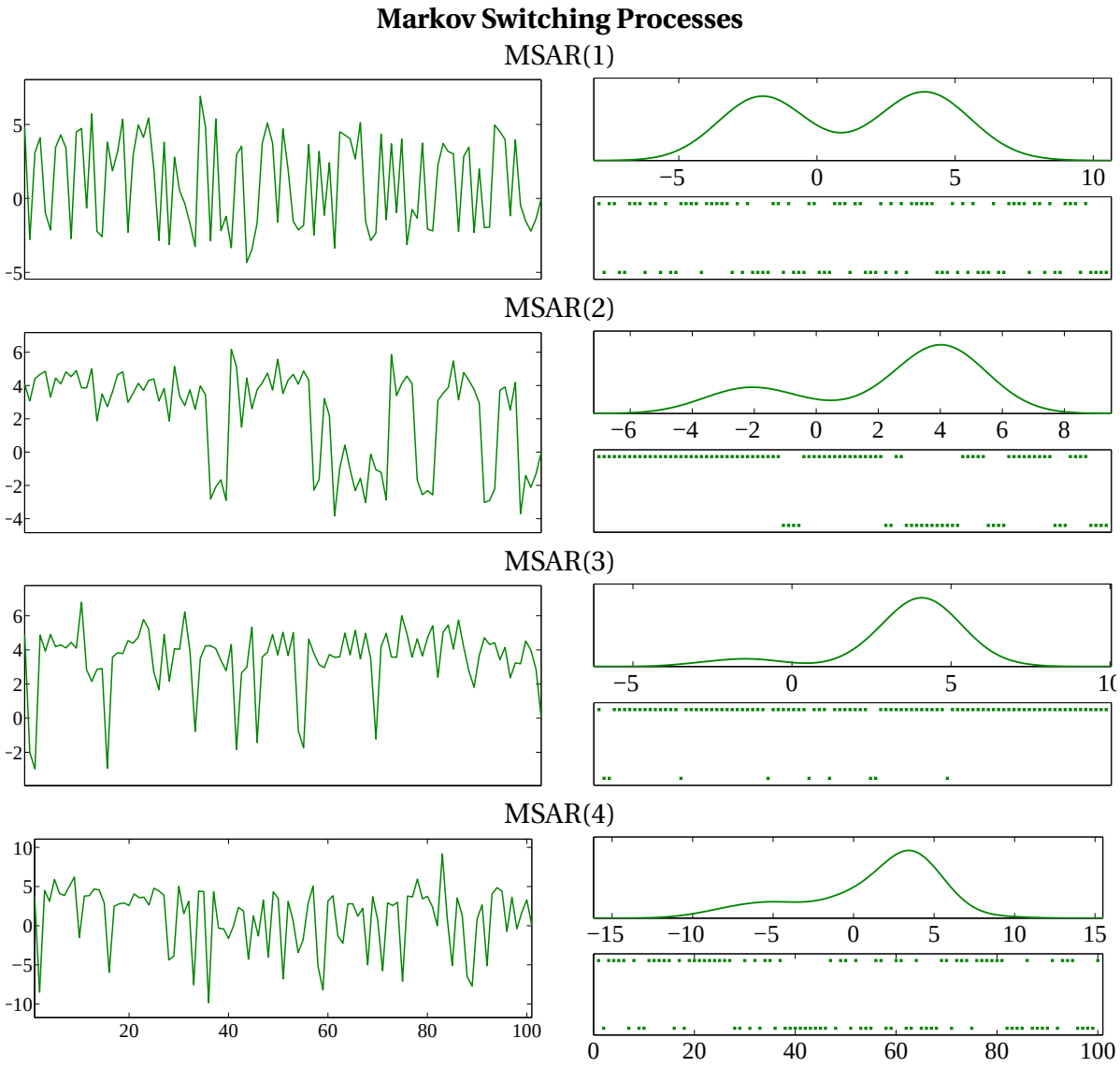

Figure 4.13: The four panels of this figure contain simulated data generated by the 4 Markov switching processes described in the text. In each panel, the large subpanel contains the generated data, the top right subpanel contains a kernel density estimate of the unconditional density and the bottom right subpanel contains the time series of the state values (high points correspond to the high state).

4.12.9 Threshold Autoregression and Self-Exciting Threshold Autoregression

A second class of nonlinear models that have gained considerable traction in financial applications are Threshold Autoregressions (TAR), and in particular, the subfamily of Self-Exciting Threshold Autoregressions (SETAR).[23]

Definition 4.46 (Threshold Autoregression)**.** A threshold autoregression is a Pth Order autore-

[23] See Fan and Yao (2005) for a comprehensive treatment of non-linear time-series models.

gressive process with state-dependent parameters where the state is determined by the lagged level of an exogenous variable x_{t-k} for some $k \geq 1$.

$$y_t = \phi_0^{(s_t)} + \phi_1^{(s_t)} y_{t-1} + \ldots + \phi_P^{(s_t)} y_{t-p} + \sigma^{(s_t)} \epsilon_t \tag{4.111}$$

Let $-\infty = x_0 < x_1 < x_2 < \ldots < x_N < x_{N+1} = \infty$ be a partition of x in to $N+1$ distinct bins. $s_t = j$ if $x_{t-k} \in (x_j, x_{j+1})$.

Self-exciting threshold autoregressions, introduced in Tong (1978), are similarly defined. The only change is in the definition of the threshold variable; rather than relying on an exogenous variable to determine the state, the state in SETARs is determined by lagged values of the dependent variable.

Definition 4.47 (Self Exciting Threshold Autoregression)**.** A self exciting threshold autoregression is a P$^{\text{th}}$ Order autoregressive process with state-dependent parameters where the state is determined by the lagged level of the dependent variable, y_{t-k} for some $k \geq 1$.

$$y_t = \phi_0^{(s_t)} + \phi_1^{(s_t)} y_{t-1} + \ldots + \phi_P^{(s_t)} y_{t-p} + \sigma^{(s_t)} \epsilon_t \tag{4.112}$$

Let $-\infty = y_0 < y_1 < y_2 < \ldots < y_N < y_{N+1} = \infty$ be a partition of y in to $N+1$ distinct bins. $s_t = j$ is $y_{t-k} \in (y_j, y_{j+1})$.

The primary application of SETAR models in finance has been to exchange rates which often exhibit a behavior that is difficult to model with standard ARMA models: many FX rates exhibit random-walk-like behavior in a range yet remain within the band longer than would be consistent with a simple random walk. A symmetric SETAR is a parsimonious model that can describe this behavior and is parameterized

$$\begin{aligned} y_t &= y_{t-1} + \epsilon_t \text{ if } C - \delta < y_t < C + \delta \\ y_t &= C(1-\phi) + \phi y_{t-1} + \epsilon_t \text{ if } y_t < C - \delta \text{ or } y_t > C + \delta \end{aligned} \tag{4.113}$$

where C is the "target" exchange rate. The first equation is a standard random walk, and when y_t is within the target band it behaves like a random walk. The second equation is only relevant when y_t is outside of its target band and ensures that y_t is mean reverting towards C as long as $|\phi| < 1$.[24] ϕ is usually assumed to lie between 0 and 1 which produces a smooth mean reversion back towards the band.

To illustrate the behavior of this process and the highlight the differences between it and a random walk, 200 data points were generated with different values of ϕ using standard normal innovations. The mean was set to 100 and the used $\delta = 5$, and so y_t follows a random walk when between 95 and 105. The lag value of the threshold variable (k) was set to one. Four values for ϕ were used: 0, 0.5, 0.9 and 1. The extreme cases represent a process which is immediately mean reverting ($\phi = 0$), in which case as soon as y_t leaves the target band it is immediately returned to C, and a process that is a pure random walk ($\phi = 1$) since $y_t = y_{t-1} + \epsilon_t$ for any value of y_{t-1}. The two interior cases represent smooth reversion back to the band; when $\phi = .5$ the reversion is quick and when $\phi = .9$ the reversion is slow. When ϕ is close to 1 it is very difficult to differentiate a band SETAR from a pure random walk, which is one of the explanations for the poor performance of unit root tests where tests often fail to reject a unit root despite clear economic theory predicting that a time series should be mean reverting.

[24] Recall the mean of an AR(1) $y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t$ is $\phi_0/(1-\phi_1)$ where $\phi_0 = C(1-\phi)$ and $\phi_1 = \phi$ in this SETAR.

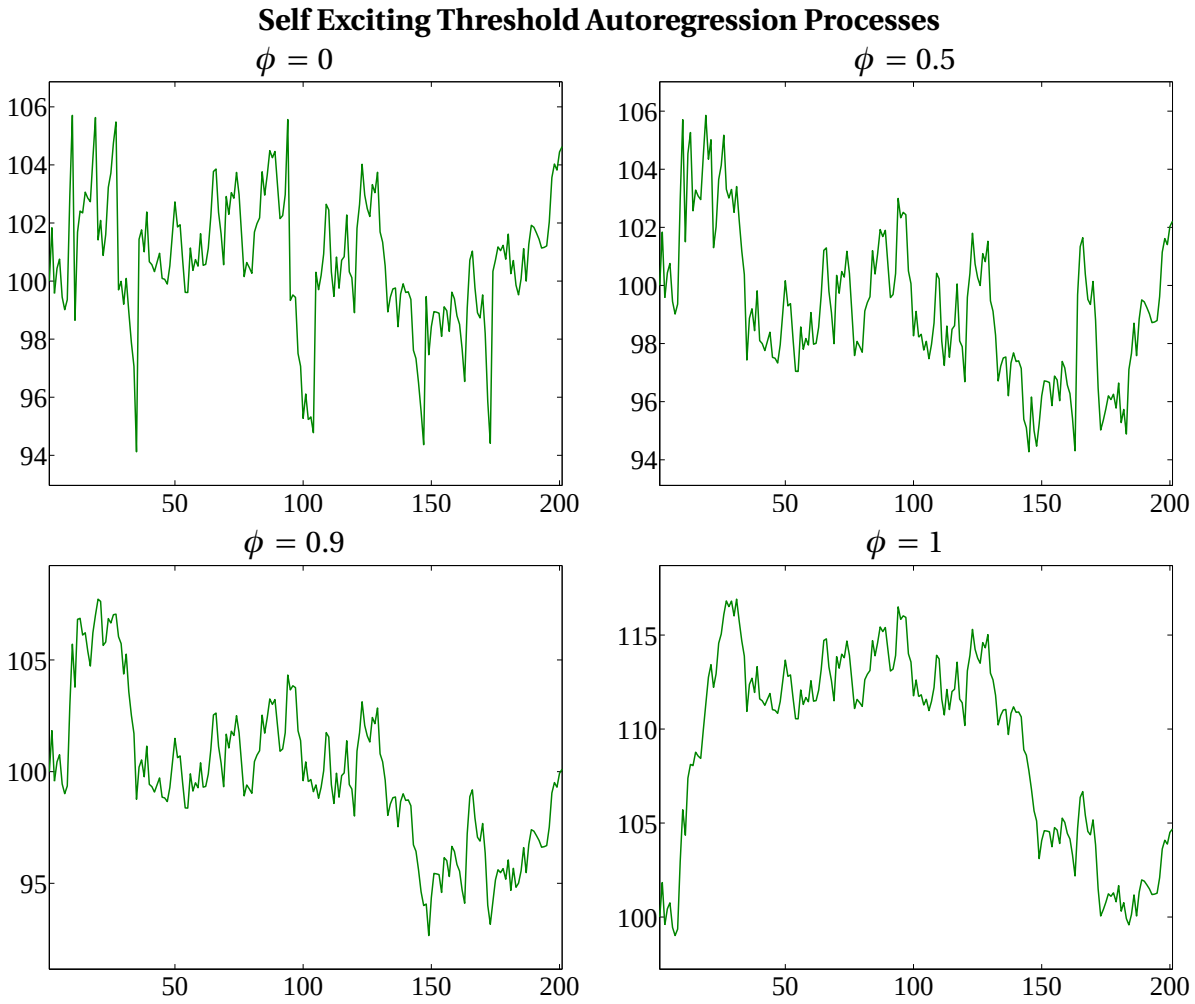

Figure 4.14: The four panels of this figure contain simulated data generated by a SETAR with different values of ϕ. When $\phi = 0$ the process is immediately returned to its unconditional mean $C = 100$. Larger values of ϕ increase the amount of time spent outside of the "target band" (95–105) and when $\phi = 1$, the process is a pure random walk.

4.A Computing Autocovariance and Autocorrelations

This appendix covers the derivation of the ACF for the MA(1), MA(Q), AR(1), AR(2), AR(3) and ARMA(1,1). Throughout this appendix, $\{\epsilon_t\}$ is assumed to be a white noise process and the processes parameters are always assumed to be consistent with covariance stationarity. All models are assumed to be mean 0, an assumption made without loss of generality since autocovariances are defined using demeaned time series,

$$\gamma_s = \mathrm{E}[(y_t - \mu)(y_{t-s} - \mu)]$$

where $\mu = \mathrm{E}[y_t]$. Recall that the autocorrelation is simply the of the s^{th} autocovariance to the variance,

$$\rho_s = \frac{\gamma_s}{\gamma_0}.$$

This appendix presents two methods for deriving the autocorrelations of ARMA processes: backward substitution and the Yule-Walker equations, a set of k equations with k unknowns where $\gamma_0, \gamma_1, \ldots, \gamma_{k-1}$ are the solution.

4.A.1 Yule-Walker

The Yule-Walker equations are a linear system of $\max(P, Q) + 1$ equations (in an ARMA(P,Q)) where the solution to the system are the long-run variance and the first $k-1$ autocovariances. The Yule-Walker equations are formed by equating the definition of an autocovariance with an expansion produced by substituting for the contemporaneous value of y_t. For example, suppose y_t follows an AR(2) process,

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$$

The variance must satisfy

$$\begin{aligned} \mathrm{E}[y_t y_t] &= \mathrm{E}[y_t(\phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t)] \\ \mathrm{E}[y_t^2] &= \mathrm{E}[\phi_1 y_t y_{t-1} + \phi_2 y_t y_{t-2} + y_t \epsilon_t] \\ \mathrm{V}[y_t] &= \phi_1 \mathrm{E}[y_t y_{t-1}] + \phi_2 \mathrm{E}[y_t y_{t-2}] + \mathrm{E}[y_t \epsilon_t]. \end{aligned} \tag{4.114}$$

In the final equation above, terms of the form $\mathrm{E}[y_t y_{t-s}]$ are replaced by their population values, γ_s and $\mathrm{E}[y_t \epsilon_t]$ is replaced with its population value, σ^2.

$$\mathrm{V}[y_t y_t] = \phi_1 \mathrm{E}[y_t y_{t-1}] + \phi_2 \mathrm{E}[y_t y_{t-2}] + \mathrm{E}[y_t \epsilon_t] \tag{4.115}$$

becomes

$$\gamma_0 = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \sigma^2 \tag{4.116}$$

and so the long run variance is a function of the first two autocovariances, the model parameters, and the innovation variance. This can be repeated for the first autocovariance,

$$\mathrm{E}[y_t y_{t-1}] = \phi_1 \mathrm{E}[y_{t-1} y_{t-1}] + \phi_2 \mathrm{E}[y_{t-1} y_{t-2}] + \mathrm{E}[y_{t-1} \epsilon_t]$$

becomes

$$\gamma_1 = \phi_1 \gamma_0 + \phi_2 \gamma_1, \tag{4.117}$$

and for the second autocovariance,

$\mathrm{E}[y_t y_{t-2}] = \phi_1 \mathrm{E}[y_{t-2} y_{t-1}] + \phi_2 \mathrm{E}[y_{t-2} y_{t-2}] + \mathrm{E}[y_{t-2} \epsilon_t]$ becomes

becomes

$$\gamma_2 = \phi_1 \gamma_1 + \phi_2 \gamma_0. \tag{4.118}$$

Together eqs. (4.116), (4.117) and (4.118) form a system of three equations with three unknowns. The Yule-Walker method relies heavily on the covariance stationarity and so $\mathrm{E}[y_t y_{t-j}] = E[y_{t-h} y_{t-h-j}]$ for any h. This property of covariance stationary processes was repeatedly used in forming the producing the Yule-Walker equations since $\mathrm{E}[y_t y_t] = \mathrm{E}[y_{t-1} y_{t-1}] = \mathrm{E}[y_{t-2} y_{t-2}] = \gamma_0$ and $\mathrm{E}[y_t y_{t-1}] = \mathrm{E}[y_{t-1} y_{t-2}] = \gamma_1$.

The Yule-Walker method will be demonstrated for a number of models, starting from a simple MA(1) and working up to an ARMA(1,1).

4.A.2 MA(1)

The MA(1) is the simplest model to work with.

$$y_t = \theta_1 \epsilon_{t-1} + \epsilon_t$$

The Yule-Walker equation are

$$\begin{aligned} \mathrm{E}[y_t y_t] &= \mathrm{E}[\theta_1 \epsilon_{t-1} y_t] + \mathrm{E}[\epsilon_t y_t] \\ \mathrm{E}[y_t y_{t-1}] &= \mathrm{E}[\theta_1 \epsilon_{t-1} y_{t-1}] + \mathrm{E}[\epsilon_t y_{t-1}] \\ \mathrm{E}[y_t y_{t-2}] &= \mathrm{E}[\theta_1 \epsilon_{t-1} y_{t-2}] + \mathrm{E}[\epsilon_t y_{t-2}] \end{aligned} \tag{4.119}$$

$$\begin{aligned} \gamma_0 &= \theta_1^2 \sigma^2 + \sigma^2 \\ \gamma_1 &= \theta_1 \sigma^2 \\ \gamma_2 &= 0 \end{aligned} \tag{4.120}$$

Additionally, both γ_s and ρ_s, $s \geq 2$ are 0 by the white noise property of the residuals, and so the autocorrelations are

$$\begin{aligned} \rho_1 &= \frac{\theta_1 \sigma^2}{\theta_1^2 \sigma^2 + \sigma^2} \\ &= \frac{\theta_1}{1 + \theta_1^2}, \\ \rho_2 &= 0. \end{aligned}$$

4.A.2.1 MA(Q)

The Yule-Walker equations can be constructed and solved for any MA(Q), and the structure of the autocovariance is simple to detect by constructing a subset of the full system.

$$\begin{aligned}
\mathrm{E}[y_t y_t] &= \mathrm{E}[\theta_1 \epsilon_{t-1} y_t] + \mathrm{E}[\theta_2 \epsilon_{t-2} y_t] + \mathrm{E}[\theta_3 \epsilon_{t-3} y_t] + \ldots + \mathrm{E}[\theta_Q \epsilon_{t-Q} y_t] \qquad (4.121)\\
\gamma_0 &= \theta_1^2\sigma^2 + \theta_2^2\sigma^2 + \theta_3^2\sigma^2 + \ldots + \theta_Q^2\sigma^2 + \sigma^2\\
&= \sigma^2(1 + \theta_1^2 + \theta_2^2 + \theta_3^2 + \ldots + \theta_Q^2)\\
\mathrm{E}[y_t y_{t-1}] &= \mathrm{E}[\theta_1 \epsilon_{t-1} y_{t-1}] + \mathrm{E}[\theta_2 \epsilon_{t-2} y_{t-1}] + \mathrm{E}[\theta_3 \epsilon_{t-3} y_{t-1}] + \ldots + \mathrm{E}[\theta_Q \epsilon_{t-Q} y_{t-1}] \qquad (4.122)\\
\gamma_1 &= \theta_1\sigma^2 + \theta_1\theta_2\sigma^2 + \theta_2\theta_3\sigma^2 + \ldots + \theta_{Q-1}\theta_Q\sigma^2\\
&= \sigma^2(\theta_1 + \theta_1\theta_2 + \theta_2\theta_3 + \ldots + \theta_{Q-1}\theta_Q)\\
\mathrm{E}[y_t y_{t-2}] &= \mathrm{E}[\theta_1 \epsilon_{t-1} y_{t-2}] + \mathrm{E}[\theta_2 \epsilon_{t-2} y_{t-2}] + \mathrm{E}[\theta_3 \epsilon_{t-3} y_{t-2}] + \ldots + \mathrm{E}[\theta_Q \epsilon_{t-Q} y_{t-2}] \qquad (4.123)\\
\gamma_2 &= \theta_2\sigma^2 + \theta_1\theta_3\sigma^2 + \theta_2\theta_4\sigma^2 + \ldots + \theta_{Q-2}\theta_Q\sigma^2\\
&= \sigma^2(\theta_2 + \theta_1\theta_3 + \theta_2\theta_4 + \ldots + \theta_{Q-2}\theta_Q)
\end{aligned}$$

The pattern that emerges shows,

$$\gamma_s = \theta_s\sigma^2 + \sum_{i=1}^{Q-s} \sigma^2\theta_i\theta_{i+s} = \sigma^2(\theta_s + \sum_{i=1}^{Q-s} \theta_i\theta_{i+s}).$$

and so , γ_s is a sum of $Q - s + 1$ terms. The autocorrelations are

$$\begin{aligned}
\rho_1 &= \frac{\theta_1 + \sum_{i=1}^{Q-1} \theta_i\theta_{i+1}}{1 + \theta_s + \sum_{i=1}^{Q} \theta_i^2} \qquad (4.124)\\
\rho_2 &= \frac{\theta_2 + \sum_{i=1}^{Q-2} \theta_i\theta_{i+2}}{1 + \theta_s + \sum_{i=1}^{Q} \theta_i^2}\\
\vdots &= \qquad \vdots\\
\rho_Q &= \frac{\theta_Q}{1 + \theta_s + \sum_{i=1}^{Q} \theta_i^2}\\
\rho_{Q+s} &= 0, \quad s \geq 0
\end{aligned}$$

4.A.2.2 AR(1)

The Yule-Walker method requires be $\max(P, Q) + 1$ equations to compute the autocovariance for an ARMA(P,Q) process and in an AR(1), two are required (the third is included to establish this point).

$$y_t = \phi_1 y_{t-1} + \epsilon_t$$

$$\begin{aligned}
\mathrm{E}[y_t y_t] &= \mathrm{E}[\phi_1 y_{t-1} y_t] + \mathrm{E}[\epsilon_t y_t] \qquad (4.125)\\
\mathrm{E}[y_t y_{t-1}] &= \mathrm{E}[\phi_1 y_{t-1} y_{t-1}] + \mathrm{E}[\epsilon_t y_{t-1}]\\
\mathrm{E}[y_t y_{t-2}] &= \mathrm{E}[\phi_1 y_{t-1} y_{t-2}] + \mathrm{E}[\epsilon_t y_{t-2}]
\end{aligned}$$

These equations can be rewritten in terms of the autocovariances, model parameters and σ^2 by taking expectation and noting that $\mathrm{E}[\epsilon_t y_t] = \sigma^2$ since $y_t = \epsilon_t + \phi_1\epsilon_{t-1} + \phi_1^2\epsilon_{t-2} + \ldots$ and $\mathrm{E}[\epsilon_t y_{t-j}] = 0$, $j > 0$ since $\{\epsilon_t\}$ is a white noise process.

$$\begin{aligned}\gamma_0 &= \phi_1\gamma_1 + \sigma^2 \\ \gamma_1 &= \phi_1\gamma_0 \\ \gamma_2 &= \phi_1\gamma_1\end{aligned} \tag{4.126}$$

The third is redundant since γ_2 is fully determined by γ_1 and ϕ_1, and higher autocovariances are similarly redundant since $\gamma_s = \phi_1\gamma_{s-1}$ for any s. The first two equations can be solved for γ_0 and γ_1,

$$\begin{aligned}\gamma_0 &= \phi_1\gamma_1 + \sigma^2 \\ \gamma_1 &= \phi_1\gamma_0 \\ \Rightarrow \gamma_0 &= \phi_1^2\gamma_0 + \sigma^2 \\ \Rightarrow \gamma_0 - \phi_1^2\gamma_0 &= \sigma^2 \\ \Rightarrow \gamma_0(1-\phi_1^2) &= \sigma^2 \\ \Rightarrow \gamma_0 &= \frac{\sigma^2}{1-\phi_1^2}\end{aligned}$$

and

$$\begin{aligned}\gamma_1 &= \phi_1\gamma_0 \\ \gamma_0 &= \frac{\sigma^2}{1-\phi_1^2} \\ \Rightarrow \gamma_1 &= \phi_1\frac{\sigma^2}{1-\phi_1^2}.\end{aligned}$$

The remaining autocovariances can be computed using the recursion $\gamma_s = \phi_1\gamma_{s-1}$, and so

$$\gamma_s = \phi_1^2\frac{\sigma^2}{1-\phi_1^2}.$$

Finally, the autocorrelations can be computed as ratios of autocovariances,

$$\begin{aligned}\rho_1 &= \frac{\gamma_1}{\gamma_0} = \phi_1\frac{\sigma^s}{1-\phi_1^2}\Big/\frac{\sigma^2}{1-\phi_1^2} \\ \rho_1 &= \phi_1\end{aligned}$$

$$\begin{aligned}\rho_s &= \frac{\gamma_s}{\gamma_0} = \phi_1^s\frac{\sigma^2}{1-\phi_1^2}\Big/\frac{\sigma^2}{1-\phi_1^2} \\ \rho_s &= \phi_1^s.\end{aligned}$$

4.A.2.3 AR(2)

The autocorrelations in an AR(2)

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$$

can be similarly computed using the $\max(P, Q) + 1$ equation Yule-Walker system,

$$\begin{aligned} \mathrm{E}[y_t y_t] &= \phi_1 \mathrm{E}[y_{t-1} y_t] + \phi_2 \mathrm{E}[y_{t-2} y_t] + \mathrm{E}\epsilon_t y_t] \\ \mathrm{E}[y_t y_{t-1}] &= \phi_1 \mathrm{E}[y_{t-1} y_{t-1}] + \phi_2 \mathrm{E}[y_{t-2} y_{t-1}] + \mathrm{E}[\epsilon_t y_{t-1}] \\ \mathrm{E}[y_t y_{t-2}] &= \phi_1 \mathrm{E}[y_{t-1} y_{t-2}] + \phi_2 \mathrm{E}[y_{t-2} y_{t-2}] + \mathrm{E}[\epsilon_t y_{t-2}] \end{aligned} \tag{4.127}$$

and then replacing expectations with their population counterparts, γ_0,γ_1, γ_2 and σ^2.

$$\begin{aligned} \gamma_0 &= \phi_1 \gamma_1 + \phi_2 \gamma_2 + \sigma^2 \\ \gamma_1 &= \phi_1 \gamma_0 + \phi_2 \gamma_1 \\ \gamma_2 &= \phi_1 \gamma_1 + \phi_2 \gamma_0 \end{aligned} \tag{4.128}$$

Further, it must be the case that $\gamma_s = \phi_1 \gamma_{s-1} + \phi_2 \gamma_{s-2}$ for $s \geq 2$. To solve this system of equations, divide the autocovariance equations by γ_0, the long run variance. Omitting the first equation, the system reduces to two equations in two unknowns,

$$\begin{aligned} \rho_1 &= \phi_1 \rho_0 + \phi_2 \rho_1 \\ \rho_2 &= \phi_1 \rho_1 + \phi_2 \rho_0 \end{aligned}$$

since $\rho_0 = \gamma_0/\gamma_0 = 1$.

$$\begin{aligned} \rho_1 &= \phi_1 + \phi_2 \rho_1 \\ \rho_2 &= \phi_1 \rho_1 + \phi_2 \end{aligned}$$

Solving this system,

$$\begin{aligned} \rho_1 &= \phi_1 + \phi_2 \rho_1 \\ \rho_1 - \phi_2 \rho_1 &= \phi_1 \\ \rho_1 (1 - \phi_2) &= \phi_1 \\ \rho_1 &= \frac{\phi_1}{1 - \phi_2} \end{aligned}$$

and

$$\begin{aligned}\rho_2 &= \phi_1\rho_1 + \phi_2 \\ &= \phi_1\frac{\phi_1}{1-\phi_2} + \phi_2 \\ &= \frac{\phi_1\phi_1 + (1-\phi_2)\phi_2}{1-\phi_2} \\ &= \frac{\phi_1^2 + \phi_2 - \phi_2^2}{1-\phi_2}\end{aligned}$$

Since $\rho_s = \phi_1\rho_{s-1} + \phi_2\rho_{s-2}$, these first two autocorrelations are sufficient to compute the other autocorrelations,

$$\begin{aligned}\rho_3 &= \phi_1\rho_2 + \phi_2\rho_1 \\ &= \phi_1\frac{\phi_1^2 + \phi_2 - \phi_2^2}{1-\phi_2} + \phi_2\frac{\phi_1}{1-\phi_2}\end{aligned}$$

and the long run variance of y_t,

$$\begin{aligned}\gamma_0 &= \phi_1\gamma_1 + \phi_2\gamma_2 + \sigma^2 \\ \gamma_0 - \phi_1\gamma_1 - \phi_2\gamma_2 &= \sigma^2 \\ \gamma_0(1 - \phi_1\rho_1 - \phi_2\rho_2) &= \sigma^2 \\ \gamma_0 &= \frac{\sigma^2}{1 - \phi_1\rho_1 - \phi_2\rho_2}\end{aligned}$$

The final solution is computed by substituting for ρ_1 and ρ_2,

$$\begin{aligned}\gamma_0 &= \frac{\sigma^2}{1 - \phi_1\frac{\phi_1}{1-\phi_2} - \phi_2\frac{\phi_1^2+\phi_2-\phi_2^2}{1-\phi_2}} \\ &= \frac{1-\phi_2}{1+\phi_2}\left(\frac{\sigma^2}{(\phi_1+\phi_2-1)(\phi_2-\phi_1-1)}\right)\end{aligned}$$

4.A.2.4 AR(3)

Begin by constructing the Yule-Walker equations,

$$\begin{aligned}\mathrm{E}[y_t y_t] &= \phi_1\mathrm{E}[y_{t-1}y_t] + \phi_2\mathrm{E}[y_{t-2}y_t] + \phi_3\mathrm{E}[y_{t-3}y_t] + \mathrm{E}[\epsilon_t y_t] \\ \mathrm{E}[y_t y_{t-1}] &= \phi_1\mathrm{E}[y_{t-1}y_{t-1}] + \phi_2\mathrm{E}[y_{t-2}y_{t-1}] + \phi_3\mathrm{E}[y_{t-3}y_{t-1}] + \mathrm{E}[\epsilon_t y_{t-1}] \\ \mathrm{E}[y_t y_{t-2}] &= \phi_1\mathrm{E}[y_{t-1}y_{t-2}] + \phi_2\mathrm{E}[y_{t-2}y_{t-2}] + \phi_3\mathrm{E}[y_{t-3}y_{t-2}] + \mathrm{E}[\epsilon_t y_{t-2}] \\ \mathrm{E}[y_t y_{t-3}] &= \phi_1\mathrm{E}[y_{t-1}y_{t-3}] + \phi_2\mathrm{E}[y_{t-2}y_{t-3}] + \phi_3\mathrm{E}[y_{t-3}y_{t-3}] + \mathrm{E}[\epsilon_t y_{t-4}].\end{aligned}$$

Replacing the expectations with their population values, $\gamma_0, \gamma_1, \ldots$ and σ^2, the Yule-Walker equations can be rewritten

$$\begin{aligned}\gamma_0 &= \phi_1\gamma_1 + \phi_2\gamma_2 + \phi_3\gamma_3 + \sigma^2 \\ \gamma_1 &= \phi_1\gamma_0 + \phi_2\gamma_1 + \phi_3\gamma_2 \\ \gamma_2 &= \phi_1\gamma_1 + \phi_2\gamma_0 + \phi_3\gamma_1 \\ \gamma_3 &= \phi_1\gamma_2 + \phi_2\gamma_1 + \phi_3\gamma_0\end{aligned} \tag{4.129}$$

and the recursive relationship $\gamma_s = \phi_1\gamma_{s-1} + \phi_2\gamma_{s-2} + \phi_3\gamma_{s-3}$ can be observed for $s \geq 3$.

Omitting the first condition and dividing by γ_0,

$$\begin{aligned}\rho_1 &= \phi_1\rho_0 + \phi_2\rho_1 + \phi_3\rho_2 \\ \rho_2 &= \phi_1\rho_1 + \phi_2\rho_0 + \phi_3\rho_1 \\ \rho_3 &= \phi_1\rho_2 + \phi_2\rho_1 + \phi_3\rho_0.\end{aligned}$$

leaving three equations in three unknowns since $\rho_0 = \gamma_0/\gamma_0 = 1$.

$$\begin{aligned}\rho_1 &= \phi_1 + \phi_2\rho_1 + \phi_3\rho_2 \\ \rho_2 &= \phi_1\rho_1 + \phi_2 + \phi_3\rho_1 \\ \rho_3 &= \phi_1\rho_2 + \phi_2\rho_1 + \phi_3\end{aligned}$$

Following some tedious algebra, the solution to this system is

$$\begin{aligned}\rho_1 &= \frac{\phi_1 + \phi_2\phi_3}{1 - \phi_2 - \phi_1\phi_3 - \phi_3^2} \\ \rho_2 &= \frac{\phi_2 + \phi_1^2 + \phi_3\phi_1 - \phi_2^2}{1 - \phi_2 - \phi_1\phi_3 - \phi_3^2} \\ \rho_3 &= \frac{\phi_3 + \phi_1^3 + \phi_1^2\phi_3 + \phi_1\phi_2^2 + 2\phi_1\phi_2 + \phi_2^2\phi_3 - \phi_2\phi_3 - \phi_1\phi_3^2 - \phi_3^3}{1 - \phi_2 - \phi_1\phi_3 - \phi_3^2}.\end{aligned}$$

Finally, the unconditional variance can be computed using the first three autocorrelations,

$$\begin{aligned}\gamma_0 &= \phi_1\gamma_1 + \phi_2\gamma_2 + \phi_3\gamma_3\sigma^2 \\ \gamma_0 - \phi_1\gamma_1 - \phi_2\gamma_2 - \phi_3\gamma_3 &= \sigma^2 \\ \gamma_0(1 - \phi_1\rho_1 + \phi_2\rho_2 + \phi_3\rho_3) &= \sigma^2 \\ \gamma_0 &= \frac{\sigma^2}{1 - \phi_1\rho_1 - \phi_2\rho_2 - \phi_3\rho_3} \\ \gamma_0 &= \frac{\sigma^2\left(1 - \phi_2 - \phi_1\phi_3 - \phi_3^2\right)}{(1 - \phi_2 - \phi_3 - \phi_1)\left(1 + \phi_2 + \phi_3\phi_1 - \phi_3^2\right)(1 + \phi_3 + \phi_1 - \phi_2)}\end{aligned}$$

4.A.2.5 ARMA(1,1)

Deriving the autocovariances and autocorrelations of an ARMA process is slightly more difficult than for a pure AR or MA process. An ARMA(1,1) is specified as

$$y_t = \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t$$

and since $P = Q = 1$, the Yule-Walker system requires two equations, noting that the third or higher autocovariance is a trivial function of the first two autocovariances.

$$\begin{aligned} \mathrm{E}[y_t y_t] &= \mathrm{E}[\phi_1 y_{t-1} y_t] + \mathrm{E}[\theta_1 \epsilon_{t-1} y_t] + \mathrm{E}[\epsilon_t y_t] \\ \mathrm{E}[y_t y_{t-1}] &= \mathrm{E}[\phi_1 y_{t-1} y_{t-1}] + \mathrm{E}[\theta_1 \epsilon_{t-1} y_{t-1}] + \mathrm{E}[\epsilon_t y_{t-1}] \end{aligned} \tag{4.130}$$

The presence of the $\mathrm{E}[\theta_1 \epsilon_{t-1} y_t]$ term in the first equation complicates solving this system since ϵ_{t-1} appears in y_t directly though $\theta_1 \epsilon_{t-1}$ and indirectly through $\phi_1 y_{t-1}$. The non-zero relationships can be determined by recursively substituting y_t until it consists of only ϵ_t, ϵ_{t-1} and y_{t-2} (since y_{t-2} is uncorrelated with ϵ_{t-1} by the WN assumption).

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t \\ &= \phi_1(\phi_1 y_{t-2} + \theta_1 \epsilon_{t-2} + \epsilon_{t-1}) + \theta_1 \epsilon_{t-1} + \epsilon_t \\ &= \phi_1^2 y_{t-2} + \phi_1 \theta_1 \epsilon_{t-2} + \phi_1 \epsilon_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t \\ &= \phi_1^2 y_{t-2} + \phi_1 \theta_1 \epsilon_{t-2} + (\phi_1 + \theta_1)\epsilon_{t-1} + \epsilon_t \end{aligned} \tag{4.131}$$

and so $\mathrm{E}[\theta_1 \epsilon_{t-1} y_t] = \theta_1(\phi_1 + \theta_1)\sigma^2$ and the Yule-Walker equations can be expressed using the population moments and model parameters.

$$\begin{aligned} \gamma_0 &= \phi_1 \gamma_1 + \theta_1(\phi_1 + \theta_1)\sigma^2 + \sigma^2 \\ \gamma_1 &= \phi_1 \gamma_0 + \theta_1 \sigma^2 \end{aligned}$$

These two equations in two unknowns which can be solved,

$$\begin{aligned} \gamma_0 &= \phi_1 \gamma_1 + \theta_1(\phi_1 + \theta_1)\sigma^2 + \sigma^2 \\ &= \phi_1(\phi_1 \gamma_0 + \theta_1 \sigma^2) + \theta_1(\phi_1 + \theta_1)\sigma^2 + \sigma^2 \\ &= \phi_1^2 \gamma_0 + \phi_1 \theta_1 \sigma^2 + \theta_1(\phi_1 + \theta_1)\sigma^2 + \sigma^2 \\ \gamma_0 - \phi_1^2 \gamma_0 &= \sigma^2(\phi_1 \theta_1 + \phi_1 \theta_1 + \theta_1^2 + 1) \\ \gamma_0 &= \frac{\sigma^2(1 + \theta_1^2 + 2\phi_1 \theta_1)}{1 - \phi_1^2} \end{aligned}$$

$$\begin{aligned}
\gamma_1 &= \phi_1\gamma_0 + \theta_1\sigma^2 \\
&= \phi_1\left(\frac{\sigma^2(1+\theta_1^2+2\phi_1\theta_1)}{1-\phi_1^2}\right) + \theta_1\sigma^2 \\
&= \phi_1\left(\frac{\sigma^2(1+\theta_1^2+2\phi_1\theta_1)}{1-\phi_1^2}\right) + \frac{(1-\phi_1^2)\theta_1\sigma^2}{1-\phi_1^2} \\
&= \frac{\sigma^2(\phi_1+\phi_1\theta_1^2+2\phi_1^2\theta_1)}{1-\phi_1^2} + \frac{(\theta_1-\theta_1\phi_1^2)\sigma^2}{1-\phi_1^2} \\
&= \frac{\sigma^2(\phi_1+\phi_1\theta_1^2+2\phi_1^2\theta_1+\theta_1-\phi_1^2\theta_1)}{1-\phi_1^2} \\
&= \frac{\sigma^2(\phi_1^2\theta_1+\phi_1\theta_1^2+\phi_1+\theta_1)}{1-\phi_1^2} \\
&= \frac{\sigma^2(\phi_1+\theta_1)(\phi_1\theta_1+1)}{1-\phi_1^2}
\end{aligned}$$

and so the 1st autocorrelation is

$$\rho_1 = \frac{\frac{\sigma^2(\phi_1+\theta_1)(\phi_1\theta_1+1)}{1-\phi_1^2}}{\frac{\sigma^2(1+\theta_1^2+2\phi_1\theta_1)}{1-\phi_1^2}} = \frac{(\phi_1+\theta_1)(\phi_1\theta_1+1)}{(1+\theta_1^2+2\phi_1\theta_1)}.$$

Returning to the next Yule-Walker equation,

$$\mathrm{E}[y_t y_{t-2}] = \mathrm{E}[\phi_1 y_{t-1} y_{t-2}] + \mathrm{E}[\theta_1\epsilon_{t-1} y_{t-2}] + \mathrm{E}[\epsilon_t y_{t-2}]$$

and so $\gamma_2 = \phi_1\gamma_1$, and, dividing both sized by γ_0, $\rho_2 = \phi_1\rho_1$. Higher order autocovariances and autocorrelation follow $\gamma_s = \phi_1\gamma_{s-1}$ and $\rho_s = \phi_1\rho_{s-1}$ respectively, and so $\rho_s = \phi_1^{s-1}\rho_1$, $s \geq 2$.

4.A.3 Backward Substitution

Backward substitution is a direct but tedious method to derive the ACF and long run variance.

4.A.3.1 AR(1)

The AR(1) process,

$$y_t = \phi_1 y_{t-1} + \epsilon_t$$

is stationary if $|\phi_1| < 1$ and $\{\epsilon_t\}$ is white noise. To compute the autocovariances and autocorrelations using backward substitution, $y_t = \phi_1 y_{t-1} + \epsilon_t$ must be transformed into a pure MA process by recursive substitution,

$$
\begin{aligned}
y_t &= \phi_1 y_{t-1} + \epsilon_t \\
&= \phi_1(\phi_1 y_{t-2} + \epsilon_{t-1}) + \epsilon_t \\
&= \phi_1^2 y_{t-2} + \phi_1 \epsilon_{t-1} + \epsilon_t \\
&= \phi_1^2(\phi_1 y_{t-3} + \epsilon_{t-2}) + \phi_1 \epsilon_{t-1} + \epsilon_t \\
&= \phi_1^3 y_{t-3} + \phi_1^2 \epsilon_{t-2} + \phi_1 \epsilon_{t-1} + \epsilon_t \\
&= \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_1^2 \epsilon_{t-2} + \phi_1^3 \epsilon_{t-3} + \dots \\
y_t &= \sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i}.
\end{aligned} \tag{4.132}
$$

The variance is the expectation of the square,

$$
\begin{aligned}
\gamma_0 &= \mathrm{V}[y_t] = \mathrm{E}[y_t^2] \\
&= \mathrm{E}[(\sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i})^2] \\
&= \mathrm{E}[(\epsilon_t + \phi_1 \epsilon_{t-1} + \phi_1^2 \epsilon_{t-2} + \phi_1^3 \epsilon_{t-3} + \dots)^2] \\
&= \mathrm{E}[\sum_{i=0}^{\infty} \phi_1^{2i} \epsilon_{t-i}^2 + \sum_{i=0}^{\infty} \sum_{j=0, i \neq j}^{\infty} \phi_1^i \phi_1^j \epsilon_{t-i} \epsilon_{t-j}] \\
&= \mathrm{E}[\sum_{i=0}^{\infty} \phi_1^{2i} \epsilon_{t-i}^2] + \mathrm{E}[\sum_{i=0}^{\infty} \sum_{j=0, i \neq j}^{\infty} \phi_1^i \phi_1^j \epsilon_{t-i} \epsilon_{t-j}] \\
&= \sum_{i=0}^{\infty} \phi_1^{2i} \mathrm{E}[\epsilon_{t-i}^2] + \sum_{i=0}^{\infty} \sum_{j=0, i \neq j}^{\infty} \phi_1^i \phi_1^j \mathrm{E}[\epsilon_{t-i} \epsilon_{t-j}] \\
&= \sum_{i=0}^{\infty} \phi_1^{2i} \sigma^2 + \sum_{i=0}^{\infty} \sum_{j=0, i \neq j}^{\infty} \phi_1^i \phi_1^j 0 \\
&= \sum_{i=0}^{\infty} \phi_1^{2i} \sigma^2 \\
&= \frac{\sigma^2}{1 - \phi_1^{2i}}
\end{aligned} \tag{4.133}
$$

The difficult step in the derivation is splitting up the ϵ_{t-i} into those that are matched to their own lag (ϵ_{t-i}^2) to those which are not ($\epsilon_{t-i}\epsilon_{t-j}$, $i \neq j$). The remainder of the derivation follows from the assumption that $\{\epsilon_t\}$ is a white noise process, and so $\mathrm{E}[\epsilon_{t-i}^2] = \sigma^2$ and $\mathrm{E}[\epsilon_{t-i}\epsilon_{t-j}]$=0, $i \neq j$. Finally, the identity that $\lim_{n\to\infty} \sum_{i=0}^{n} \phi_1^{2i} = \lim_{n\to\infty} \sum_{i=0}^{n} \left(\phi_1^2\right)^i = \frac{1}{1-\phi_1^2}$ for $|\phi_1| < 1$ was used to simplify the expression.

The 1st autocovariance can be computed using the same steps on the MA(∞) representation,

$$
\begin{aligned}
\gamma_1 &= \mathrm{E}[y_t y_{t-1}] \\
&= \mathrm{E}[\sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i} \sum_{i=1}^{\infty} \phi_1^{i-1} \epsilon_{t-i}] \\
&= \mathrm{E}[(\epsilon_t + \phi_1\epsilon_{t-1} + \phi_1^2\epsilon_{t-2} + \phi_1^3\epsilon_{t-3} + \ldots)(\epsilon_{t-1} + \phi_1\epsilon_{t-2} + \phi_1^2\epsilon_{t-3} + \phi_1^3\epsilon_{t-4} + \ldots)] \\
&= \mathrm{E}[\sum_{i=0}^{\infty} \phi_1^{2i+1}\epsilon_{t-1-i}^2 + \sum_{i=0}^{\infty}\sum_{j=1,i\neq j}^{\infty} \phi_1^i\phi_1^{j-1}\epsilon_{t-i}\epsilon_{t-j}] \\
&= \mathrm{E}[\phi_1 \sum_{i=0}^{\infty} \phi_1^{2i}\epsilon_{t-1-i}^2] + \mathrm{E}[\sum_{i=0}^{\infty}\sum_{j=1,i\neq j}^{\infty} \phi_1^i\phi_1^{j-1}\epsilon_{t-i}\epsilon_{t-j}] \\
&= \phi_1 \sum_{i=0}^{\infty} \phi_1^{2i}\mathrm{E}[\epsilon_{t-1-i}^2] + \sum_{i=0}^{\infty}\sum_{j=1,i\neq j}^{\infty} \phi_1^i\phi_1^{j-1}\mathrm{E}[\epsilon_{t-i}\epsilon_{t-j}] \\
&= \phi_1 \sum_{i=0}^{\infty} \phi_1^{2i}\sigma^2 + \sum_{i=0}^{\infty}\sum_{j=1,i\neq j}^{\infty} \phi_1^i\phi_1^{j-1}0 \\
&= \phi_1 \left(\sum_{i=0}^{\infty} \phi_1^{2i}\sigma^2 \right) \\
&= \phi_1 \frac{\sigma^2}{1-\phi_1^2} \\
&= \phi_1\gamma_0
\end{aligned}
\tag{4.134}
$$

and the s[th] autocovariance can be similarly determined.

$$
\begin{aligned}
\gamma_s &= \mathrm{E}[y_t y_{t-s}] \\
&= \mathrm{E}[\sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i} \sum_{i=s}^{\infty} \phi_1^{i-s} \epsilon_{t-i}] \\
&= \mathrm{E}[\sum_{i=0}^{\infty} \phi_1^{2i+s}\epsilon_{t-s-i}^2 + \sum_{i=0}^{\infty}\sum_{j=s,i\neq j}^{\infty} \phi_1^i\phi_1^{j-s}\epsilon_{t-i}\epsilon_{t-j}] \\
&= \mathrm{E}[\phi_1^s \sum_{i=0}^{\infty} \phi_1^{2i}\epsilon_{t-s-i}^2] + \mathrm{E}[\sum_{i=0}^{\infty}\sum_{j=s,i\neq j}^{\infty} \phi_1^i\phi_1^{j-s}\epsilon_{t-i}\epsilon_{t-j}] \\
&= \phi_1^s \sum_{i=0}^{\infty} \phi_1^{2i}\sigma^2 + \sum_{i=0}^{\infty}\sum_{j=s,i\neq j}^{\infty} \phi_1^i\phi_1^{j-s}0 \\
&= \phi_1^s \left(\sum_{i=0}^{\infty} \phi_1^{2i}\sigma^2 \right) \\
&= \phi_1^s\gamma_0
\end{aligned}
\tag{4.135}
$$

Finally, the autocorrelations can be computed from rations of autocovariances, $\rho_1 = \gamma_1/\gamma_0 = \phi_1$ and $\rho_s = \gamma_s/\gamma_0 = \phi_1^s$.

4.A.3.2 MA(1)

The MA(1) model is the simplest non-degenerate time-series model considered in this course,

$$y_t = \theta_1 \epsilon_{t-1} + \epsilon_t$$

and the derivation of its autocorrelation function is trivial since there no backward substitution is required. The variance is

$$\begin{aligned} \gamma_0 &= \mathrm{V}[y_t] = \mathrm{E}[y_t^2] \\ &= \mathrm{E}[(\theta_1 \epsilon_{t-1} + \epsilon_t)^2] \\ &= \mathrm{E}[\theta_1^2 \epsilon_{t-1}^2 + 2\theta_1 \epsilon_t \epsilon_{t-1} + \epsilon_t^2] \\ &= \mathrm{E}[\theta_1^2 \epsilon_{t-1}^2] + \mathrm{E}[2\theta_1 \epsilon_t \epsilon_{t-1}] + \mathrm{E}[\epsilon_t^2] \\ &= \theta_1^2 \sigma^2 + 0 + \sigma^2 \\ &= \sigma^2 (1 + \theta_1^2) \end{aligned} \tag{4.136}$$

and the 1st autocovariance is

$$\begin{aligned} \gamma_1 &= \mathrm{E}[y_t y_{t-1}] \\ &= \mathrm{E}[(\theta_1 \epsilon_{t-1} + \epsilon_t)(\theta_1 \epsilon_{t-2} + \epsilon_{t-1})] \\ &= \mathrm{E}[\theta_1^2 \epsilon_{t-1} \epsilon_{t-2} + \theta_1 \epsilon_{t-1}^2 + \theta_1 \epsilon_t \epsilon_{t-2} + \epsilon_t \epsilon_{t-1}] \\ &= \mathrm{E}[\theta_1^2 \epsilon_{t-1} \epsilon_{t-2}] + \mathrm{E}[\theta_1 \epsilon_{t-1}^2] + \mathrm{E}[\theta_1 \epsilon_t \epsilon_{t-2}] + \mathrm{E}[\epsilon_t \epsilon_{t-1}] \\ &= 0 + \theta_1 \sigma^2 + 0 + 0 \\ &= \theta_1 \sigma^2 \end{aligned} \tag{4.137}$$

The 2nd(and higher) autocovariance is

$$\begin{aligned} \gamma_2 &= \mathrm{E}[y_t y_{t-2}] \\ &= \mathrm{E}[(\theta_1 \epsilon_{t-1} + \epsilon_t)(\theta_1 \epsilon_{t-3} + \epsilon_{t-2})] \\ &= \mathrm{E}[\theta_1^2 \epsilon_{t-1} \epsilon_{t-3} + \theta_1 \epsilon_{t-1} \epsilon_{t-2} + \theta_1 \epsilon_t \epsilon_{t-3} + \epsilon_t \epsilon_{t-2}] \\ &= \mathrm{E}[\theta_1^2 \epsilon_{t-1} \epsilon_{t-3}] + \mathrm{E}[\theta_1 \epsilon_{t-1} \epsilon_{t-2}] + \mathrm{E}[\theta_1 \epsilon_t \epsilon_{t-3}] + \mathrm{E}[\epsilon_t \epsilon_{t-2}] \\ &= 0 + 0 + 0 + 0 \\ &= 0 \end{aligned} \tag{4.138}$$

and the autocorrelations are $\rho_1 = \theta_1/(1 + \theta_1^2)$, $\rho_s = 0$, $s \geq 2$.

4.A.3.3 ARMA(1,1)

An ARMA(1,1) process,

$$y_t = \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t$$

is stationary if $|\phi_1| < 1$ and $\{\epsilon_t\}$ is white noise. The derivation of the variance and autocovariances is more tedious than for the AR(1) process.It should be noted that derivation is longer and more complex than solving the Yule-Walker equations.

Begin by computing the MA(∞) representation,

$$
\begin{aligned}
y_t &= \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t \\
y_t &= \phi_1(\phi_1 y_{t-2} + \theta_1 \epsilon_{t-2} + \epsilon_{t-1}) + \theta_1 \epsilon_{t-1} + \epsilon_t \\
y_t &= \phi_1^2 y_{t-2} + \phi_1\theta_1 \epsilon_{t-2} + \phi_1 \epsilon_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t \\
y_t &= \phi_1^2(\phi_1 y_{t-3} + \theta_1 \epsilon_{t-3} + \epsilon_{t-2}) + \phi_1\theta_1 \epsilon_{t-2} + (\phi_1 + \theta_1)\epsilon_{t-1} + \epsilon_t \\
y_t &= \phi_1^3 y_{t-3} + \phi_1^2\theta_1 \epsilon_{t-3} + \phi_1^2 \epsilon_{t-2} + \phi_1\theta_1 \epsilon_{t-2} + (\phi_1 + \theta_1)\epsilon_{t-1} + \epsilon_t \\
y_t &= \phi_1^3(\phi_1 y_{t-4} + \theta_1 \epsilon_{t-4} + \epsilon_{t-3}) + \phi_1^2\theta_1 \epsilon_{t-3} + \phi_1(\phi_1 + \theta_1)\epsilon_{t-2} + (\phi_1 + \theta_1)\epsilon_{t-1} + \epsilon_t \\
y_t &= \phi_1^4 y_{t-4} + \phi_1^3\theta_1 \epsilon_{t-4} + \phi_1^3 \epsilon_{t-3} + \phi_1^2\theta_1 \epsilon_{t-3} + \phi_1(\phi_1 + \theta_1)\epsilon_{t-2} + (\phi_1 + \theta_1)\epsilon_{t-1} + \epsilon_t \\
y_t &= \phi_1^4 y_{t-4} + \phi_1^3\theta_1 \epsilon_{t-4} + \phi_1^2(\phi_1 + \theta_1)\epsilon_{t-3} + \phi_1(\phi_1 + \theta_1)\epsilon_{t-2} + (\phi_1 + \theta_1)\epsilon_{t-1} + \epsilon_t \\
y_t &= \epsilon_t + (\phi_1 + \theta_1)\epsilon_{t-1} + \phi_1(\phi_1 + \theta_1)\epsilon_{t-2} + \phi_1^2(\phi_1 + \theta_1)\epsilon_{t-3} + \dots \\
y_t &= \epsilon_t + \sum_{i=0}^{\infty} \phi_1^i(\phi_1 + \theta_1)\epsilon_{t-1-i}
\end{aligned}
\tag{4.139}
$$

The primary issue is that the backward substitution form, unlike in the AR(1) case, is not completely symmetric. Specifically, ϵ_t has a different weight than the other shocks and does not follow the same pattern.

$$
\begin{aligned}
\gamma_0 = \mathrm{V}[y_t] &= \mathrm{E}\left[y_t^2\right] \\
&= \mathrm{E}\left[\left(\epsilon_t + \sum_{i=0}^{\infty} \phi_1^i(\phi_1 + \theta_1)\epsilon_{t-1-i}\right)^2\right] \\
&= \mathrm{E}\left[\left(\epsilon_t + (\phi_1 + \theta_1)\epsilon_{t-1} + \phi_1(\phi_1 + \theta_1)\epsilon_{t-2} + \phi_1^2(\phi_1 + \theta_1)\epsilon_{t-3} + \dots\right)^2\right] \\
&= \mathrm{E}\left[\epsilon_t^2 + 2\epsilon_t \sum_{i=0}^{\infty} \phi_1^i(\phi_1 + \theta_1)\epsilon_{t-1-i} + \left(\sum_{i=0}^{\infty} \phi_1^i(\phi_1 + \theta_1)\epsilon_{t-1-i}\right)^2\right] \\
&= \mathrm{E}\left[\epsilon_t^2\right] + \mathrm{E}\left[2\epsilon_t \sum_{i=0}^{\infty} \phi_1^i(\phi_1 + \theta_1)\epsilon_{t-1-i}\right] + \mathrm{E}\left[\left(\sum_{i=0}^{\infty} \phi_1^i(\phi_1 + \theta_1)\epsilon_{t-1-i}\right)^2\right] \\
&= \sigma^2 + 0 + \mathrm{E}\left[\left(\sum_{i=0}^{\infty} \phi_1^i(\phi_1 + \theta_1)\epsilon_{t-1-i}\right)^2\right] \\
&= \sigma^2 + \mathrm{E}\left[\sum_{i=0}^{\infty} \phi_1^{2i}(\phi_1 + \theta_1)^2 \epsilon_{t-1-i}^2 + \sum_{i=0}^{\infty}\sum_{j=0, j\neq i}^{\infty} \phi_1^i\phi_1^j(\phi_1 + \theta_1)^2 \epsilon_{t-1-i}\epsilon_{t-1-j}\right] \\
&= \sigma^2 + \sum_{i=0}^{\infty} \phi_1^{2i}(\phi_1 + \theta_1)^2 \mathrm{E}\left[\epsilon_{t-1-i}^2\right] + \sum_{i=0}^{\infty}\sum_{j=0, j\neq i}^{\infty} \phi_1^i\phi_1^j(\phi_1 + \theta_1)^2 \mathrm{E}\left[\epsilon_{t-1-i}\epsilon_{t-1-j}\right]
\end{aligned}
\tag{4.140}
$$

$$
\begin{aligned}
&= \sigma^2 + \sum_{i=0}^{\infty} \phi_1^{2i} (\phi_1 + \theta_1)^2 \sigma^2 + \sum_{i=0}^{\infty} \sum_{j=0, j\neq i}^{\infty} \phi_1^i \phi_1^j (\phi_1 + \theta_1)^2 0 \\
&= \sigma^2 + \sum_{i=0}^{\infty} \phi_1^{2i} (\phi_1 + \theta_1)^2 \sigma^2 \\
&= \sigma^2 + \frac{(\phi_1 + \theta_1)^2 \sigma^2}{1 - \phi_1^2} \\
&= \sigma^2 \frac{1 - \phi_1^2 + (\phi_1 + \theta_1)^2}{1 - \phi_1^2} \\
&= \sigma^2 \frac{1 + \theta_1^2 + 2\phi_1\theta_1}{1 - \phi_1^2}
\end{aligned}
$$

The difficult step in this derivations is in aligning the ϵ_{t-i} since $\{\epsilon_t\}$ is a white noise process. The autocovariance derivation is fairly involved (and presented in full detail).

$$
\begin{aligned}
\gamma_1 &= \mathrm{E}[y_t y_{t-1}] \qquad (4.141) \\
&= \mathrm{E}\left[\left(\epsilon_t + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \epsilon_{t-1-i}\right)\left(\epsilon_{t-1} + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \epsilon_{t-2-i}\right)\right] \\
&= \mathrm{E}\left[\left(\epsilon_t + (\phi_1 + \theta_1)\epsilon_{t-1} + \phi_1(\phi_1 + \theta_1)\epsilon_{t-2} + \phi_1^2(\phi_1 + \theta_1)\epsilon_{t-3} + \ldots\right) \times \right. \\
&\qquad \left. \left(\epsilon_{t-1} + (\phi_1 + \theta_1)\epsilon_{t-2} + \phi_1(\phi_1 + \theta_1)\epsilon_{t-3} + \phi_1^2(\phi_1 + \theta_1)\epsilon_{t-4} + \ldots\right)\right] \\
&= \mathrm{E}\left[\epsilon_t\epsilon_{t-1} + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1)\epsilon_t\epsilon_{t-2-i} + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1)\epsilon_{t-1}\epsilon_{t-1-i}\right. \\
&\qquad \left. + \left(\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1)\epsilon_{t-1-i}\right)\left(\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1)\epsilon_{t-2-i}\right)\right] \\
&= \mathrm{E}[\epsilon_t\epsilon_{t-1}] + \mathrm{E}\left[\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1)\epsilon_t\epsilon_{t-2-i}\right] + \mathrm{E}\left[\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1)\epsilon_{t-1}\epsilon_{t-1-i}\right] \\
&\qquad + \mathrm{E}\left[\left(\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1)\epsilon_{t-1-i}\right)\left(\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1)\epsilon_{t-2-i}\right)\right] \\
&= 0 + 0 + (\phi_1 + \theta_1)\sigma^2 + \mathrm{E}\left[\left(\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1)\epsilon_{t-1-i}\right)\left(\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1)\epsilon_{t-2-i}\right)\right] \\
&= (\phi_1 + \theta_1)\sigma^2 + \mathrm{E}\left[\sum_{i=0}^{\infty} \phi_1^{2i+1} (\phi_1 + \theta_1)^2 \epsilon_{t-2-i}^2 + \sum_{i=0}^{\infty} \sum_{j=0, i\neq j+1}^{\infty} \phi_1^i \phi_1^j (\phi_1 + \theta_1)^2 \epsilon_{t-1-i}\epsilon_{t-2-i}\right] \\
&= (\phi_1 + \theta_1)\sigma^2 + \mathrm{E}\left[\sum_{i=0}^{\infty} \phi_1^{2i+1} (\phi_1 + \theta_1)^2 \epsilon_{t-2-i}^2\right] + \mathrm{E}\left[\sum_{i=0}^{\infty} \sum_{j=0, i\neq j+1}^{\infty} \phi_1^i \phi_1^j (\phi_1 + \theta_1)^2 \epsilon_{t-1-i}\epsilon_{t-2-i}\right] \\
&= (\phi_1 + \theta_1)\sigma^2 + \mathrm{E}\left[\phi_1 \sum_{i=0}^{\infty} \phi_1^{2i} (\phi_1 + \theta_1)^2 \epsilon_{t-2-i}^2\right] + 0
\end{aligned}
$$

$$
\begin{aligned}
&= (\phi_1 + \theta_1)\sigma^2 + \phi_1 \sum_{i=0}^{\infty} \phi_1^{2i} (\phi_1 + \theta_1)^2 \mathrm{E}\left[\epsilon_{t-2-i}^2\right] \\
&= (\phi_1 + \theta_1)\sigma^2 + \phi_1 \sum_{i=0}^{\infty} \phi_1^{2i} (\phi_1 + \theta_1)^2 \sigma^2 \\
&= (\phi_1 + \theta_1)\sigma^2 + \phi_1 \frac{(\phi_1 + \theta_1)^2 \sigma^2}{1 - \phi_1^2} \\
&= \frac{\sigma^2 \left[\left(1 - \phi_1^2\right)(\phi_1 + \theta_1) + \phi_1 (\phi_1 + \theta_1)^2\right]}{1 - \phi_1^2} \\
&= \frac{\sigma^2 \left(\phi_1 + \theta_1 - \phi_1^3 - \phi_1^2\theta_1 + \phi_1^3 + 2\phi_1^2\theta_1 - \phi_1\theta_1^2\right)}{1 - \phi_1^2} \\
&= \frac{\sigma^2 \left[\phi_1 + \theta_1 + \phi_1^2\theta_1 - \phi_1\theta_1^2\right]}{1 - \phi_1^2} \\
&= \frac{\sigma^2 (\phi_1 + \theta_1)(\phi_1\theta_1 + 1)}{1 - \phi_1^2}
\end{aligned}
$$

The most difficult step in this derivation is in showing that $\mathrm{E}[\sum_{i=0}^{\infty} \phi_1^i(\phi_1 + \theta_1)\epsilon_{t-1}\epsilon_{t-1-i}] = \sigma^2(\phi_1 + \theta_1)$ since there is one ϵ_{t-1-i} which is aligned to ϵ_{t-1} (i.e. when $i = 0$), and so the autocorrelations may be derived,

$$
\begin{aligned}
\rho_1 &= \frac{\frac{\sigma^2(\phi_1+\theta_1)(\phi_1\theta_1+1)}{1-\phi_1^2}}{\frac{\sigma^2(1+\theta_1^2+2\phi_1\theta_1)}{1-\phi_1^2}} \\
&= \frac{(\phi_1 + \theta_1)(\phi_1\theta_1 + 1)}{(1 + \theta_1^2 + 2\phi_1\theta_1)}
\end{aligned}
\tag{4.142}
$$

and the remaining autocorrelations can be computed using the recursion, $\rho_s = \phi_1\rho_{s-1}$, $s \geq 2$.

Shorter Problems

Problem 4.1. What is the optimal 3-step forecast from the ARMA(1,2), $y_t = \phi_0 + \phi_1 y_{t-1} + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \epsilon_t$, where ϵ_t is a mean 0 white noise process?

Problem 4.2. What are the expected values for α, β and γ when a forecasting model is well specified in the Mincer-Zarnowitz regression,

$$y_{t+h} = \alpha + \beta \hat{y}_{t+h|t} + \gamma x_t + \eta_{t+h}.$$

Provide an explanation for why these values should be expected.

Problem 4.3. What are the consequences of using White or Newey-West to estimate the covariance in a linear regression when the errors are serially uncorrelated and homoskedastic?

Problem 4.4. What are the 1-step and 2-step optimal forecasts for the conditional mean when $y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t$ where $\epsilon_t \overset{\text{i.i.d.}}{\sim} N(0, 1)$?

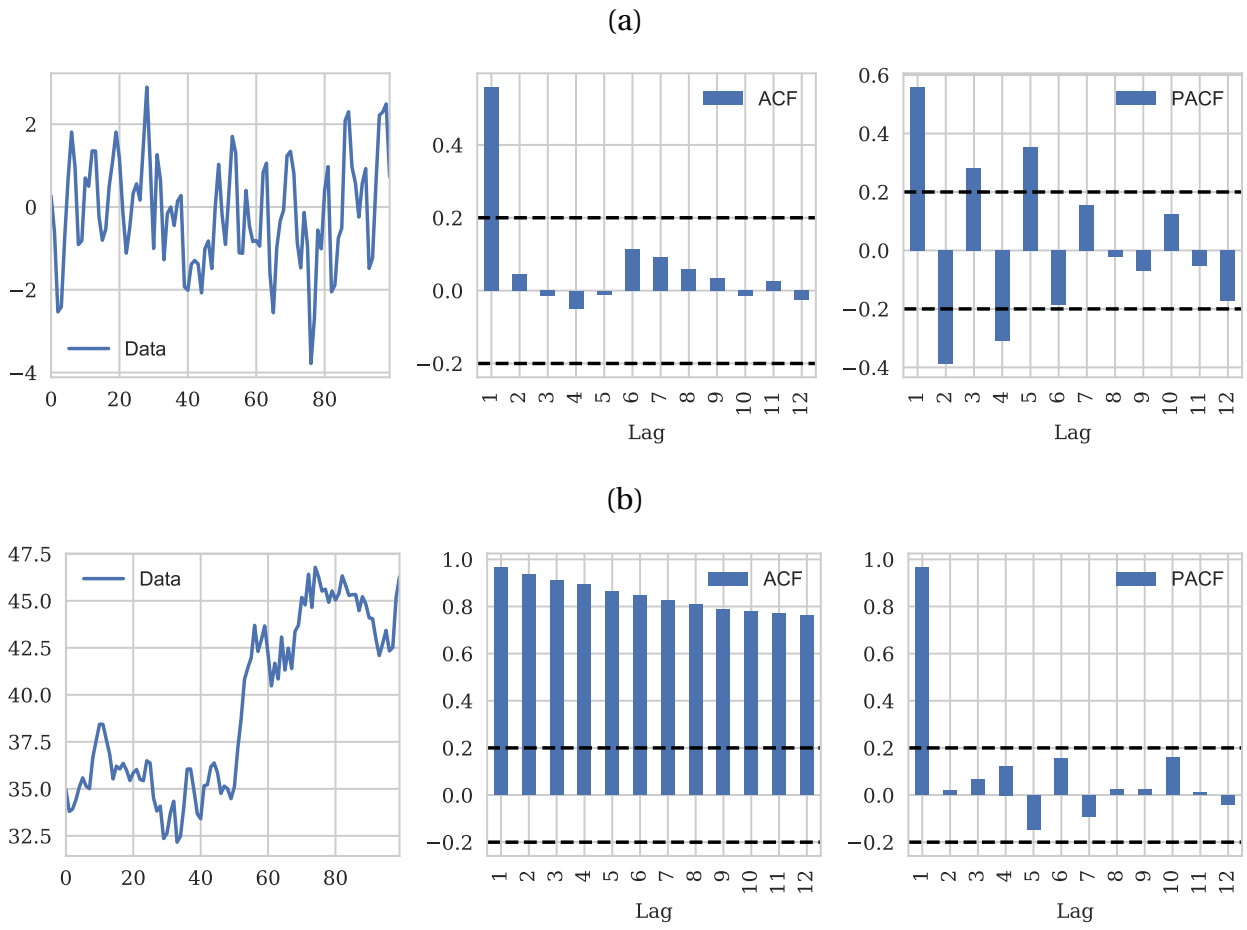

Figure 4.15: Plots for question 2(b).

Problem 4.5. Is the sum of two white noise processes, $\epsilon_t = \eta_t + \nu_t$ necessarily a white noise process?

Problem 4.6. What are the 1-step and 2-step optimal mean square forecast errors when $y_t = \phi_0 + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \epsilon_t$ where $\epsilon_t \overset{\text{i.i.d.}}{\sim} N(0,1)$?

Problem 4.7. Outline the steps needed to perform a Diebold-Mariano test that two models for the conditional mean are equivalent (in the MSE sense).

Problem 4.8. Justify a reasonable model for each of these time series in Figure 4.15 using information in the autocorrelation and partial autocorrelation plots. In each set of plots, the left most panel shows that data ($T = 100$). The middle panel shows the sample autocorrelation with 95% confidence bands. The right panel shows the sample partial autocorrelation for the data with 95% confidence bands.

Longer Exercises

Exercise 4.1. Answer the following questions:

1. Under what conditions on the parameters and errors are the following processes covariance stationary?

   (a) $y_t = \phi_0 + \epsilon_t$

   (b) $y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t$

   (c) $y_t = \phi_0 + \theta_1 \epsilon_{t-1} + \epsilon_t$

   (d) $y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$

   (e) $y_t = \phi_0 + \phi_2 y_{t-2} + \epsilon_t$

   (f) $y_t = \phi_0 + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t$

2. Is the sum of two white noise processes, $\nu_t = \epsilon_t + \eta_t$, necessarily a white noise process? If so, verify that the properties of a white noise are satisfied. If not, show why and describe any further assumptions required for the sum to be a white noise process.

Exercise 4.2. Consider an AR(1)

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t$$

1. What is a minimal set of assumptions sufficient to ensure $\{y_t\}$ is covariance stationary if $\{\epsilon_t\}$ is an i.i.d. sequence?

2. What are the values of the following quantities?

   (a) $\mathrm{E}[y_{t+1}]$

   (b) $\mathrm{E}_t[y_{t+1}]$

   (c) $\mathrm{V}[y_{t+1}]$

   (d) $\mathrm{V}_t[y_{t+1}]$

   (e) ρ_{-1}

   (f) ρ_2

Exercise 4.3. Consider an MA(1)

$$y_t = \phi_0 + \theta_1 \epsilon_{t-1} + \epsilon_t$$

1. What is a minimal set of assumptions sufficient to ensure $\{y_t\}$ is covariance stationary if $\{\epsilon_t\}$ is an i.i.d. sequence?

2. What are the values of the following quantities?

   (a) $\mathrm{E}[y_{t+1}]$

   (b) $\mathrm{E}_t[y_{t+1}]$

   (c) $\mathrm{V}[y_{t+1}]$

   (d) $\mathrm{V}_t[y_{t+1}]$

   (e) ρ_{-1}

(f) ρ_2

3. Suppose you were trying to differentiate between an AR(1) and an MA(1) but could not estimate any regressions. What would you do?

Exercise 4.4. Consider an MA(2)

$$y_t = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \epsilon_t$$

1. What is a minimal set of assumptions sufficient to ensure $\{y_t\}$ is covariance stationary if $\{\epsilon_t\}$ is an i.i.d. sequence?

2. What are the values of the following quantities?

   (a) $\mathrm{E}[y_{t+1}]$
   (b) $\mathrm{E}_t[y_{t+1}]$
   (c) $\mathrm{V}[y_{t+1}]$
   (d) $\mathrm{V}_t[y_{t+1}]$
   (e) ρ_{-1}
   (f) ρ_2
   (g) ρ_3

Exercise 4.5. Answer the following questions:

1. For each of the following processes, find $\mathrm{E}_t[y_{t+1}]$. You can assume $\{\epsilon_t\}$ is a mean zero i.i.d. sequence.

   (a) $y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t$
   (b) $y_t = \phi_0 + \theta_1 \epsilon_{t-1} + \epsilon_t$
   (c) $y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$
   (d) $y_t = \phi_0 + \phi_2 y_{t-2} + \epsilon_t$
   (e) $y_t = \phi_0 + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t$

2. For (a), (c) and (e), derive the h-step ahead forecast, $\mathrm{E}_t[y_{t+h}]$. What is the long run behavior of the forecast in each case?

3. The forecast error variance is defined as $\mathrm{E}[(y_{t+h} - \mathrm{E}_t[y_{t+h}])^2]$. Find an explicit expression for the forecast error variance for (a) and (c).

Exercise 4.6. Answer the following questions:

1. What are the characteristic equations for the above systems?

   (a) $y_t = 1 + .6 y_{t-1} + x_t$

(b) $y_t = 1 + .8y_{t-2} + x_t$

(c) $y_t = 1 + .6y_{t-1} + .3y_{t-2} + x_t$

(d) $y_t = 1 + 1.2y_{t-1} + .2y_{t-2} + x_t$

(e) $y_t = 1 + 1.4y_{t-1} + .24y_{t-2} + x_t$

(f) $y_t = 1 - .8y_{t-1} + .2y_{t-2} + x_t$

2. Compute the roots for the characteristic equation? Which are convergent? Which are explosive? Are any stable or metastable?

Exercise 4.7. Suppose that y_t follows a random walk then $\Delta y_t = y_t - y_{t-1}$ is stationary.

1. Is $y_t - y_{t-j}$ for and $j \geq 2$ stationary?
2. If it is and $\{\epsilon_t\}$ is an i.i.d. sequence of standard normals, what is the distribution of $y_t - y_{t-j}$?
3. What is the joint distribution of $y_t - y_{t-j}$ and $y_{t-h} - y_{t-j-h}$ (Note: The derivation for an arbitrary h is challenging)?
   Note: If it helps in this problem, consider the case where $j = 2$ and $h = 1$.

Exercise 4.8. Outline the steps needed to perform a unit root test on as time-series of FX rates. Be sure to detail the any important considerations that may affect the test.

Exercise 4.9. Answer the following questions:

1. How are the autocorrelations and partial autocorrelations useful in building a model?
2. Suppose you observe the three sets of ACF/PACF in figure 4.16. What ARMA specification would you expect in each case. Note: Dashed line indicates the 95% confidence interval for a test that the autocorrelation or partial autocorrelation is 0.
3. Describe the three methods of model selection discussed in class: general-to-specific, specific-to-general and the use of information criteria (Schwarz/Bayesian Information Criteria and/or Akaike Information Criteria). When might each be preferred to the others?
4. Describe the Wald, Lagrange Multiplier (Score) and Likelihood ratio tests. What aspect of a model does each test? What are the strengths and weaknesses of each?

Exercise 4.10. Answer the following questions about forecast errors.

1. Let $y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t$ with the usual assumptions on $\{\epsilon_t\}$. Derive an explicit expression for the 1-step and 2-step ahead forecast errors, $e_{t+h|t} = y_{t+h} - \hat{y}_{t+h|t}$ where $\hat{y}_{t+h|t}$ is the MSE optimal forecast where $h = 1$ or $h = 2$ (what is the MSE optimal forecast?).
2. What is the autocorrelation function of a time-series of forecast errors $\{e_{t+h|t}\}$, $h = 1$ or $h = 2$. (Hint: Use the formula you derived above)
3. Can you generalize the above to a generic h? (In other words, leave the solution as a function of h).

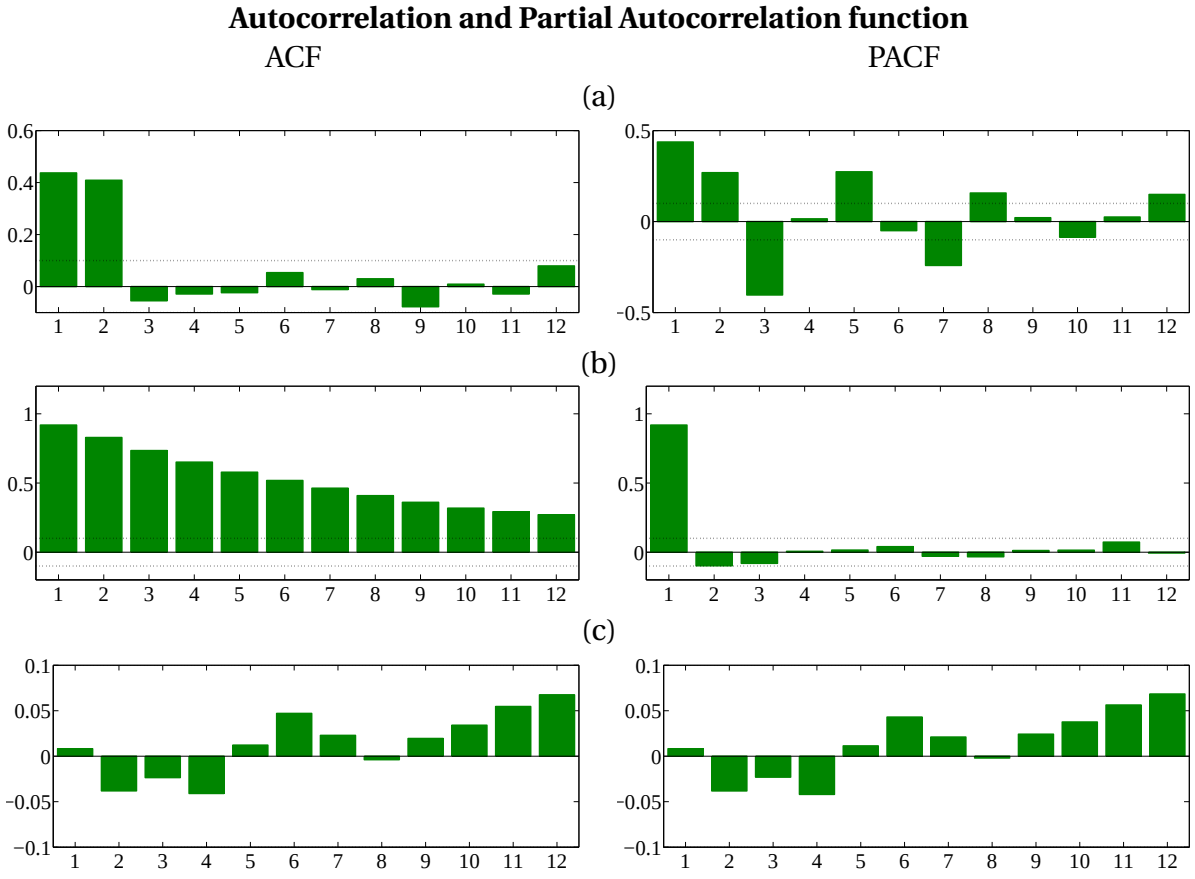

Figure 4.16: The ACF and PACF of three stochastic processes. Use these to answer question 4.9.

4. How could you test whether the forecast has excess dependence using an ARMA model?

Exercise 4.11. Answer the following questions.

1. Outline the steps needed to determine whether a time series $\{y_t\}$ contains a unit root. Be certain to discuss the important considerations at each step, if any.
2. If y_t follows a pure random walk driven by whit noise innovations then $\Delta y_t = y_t - y_{t-1}$ is stationary.
   (a) Is $y_t - y_{t-j}$ for and $j \geq 2$ stationary?
   (b) If it is and $\{\epsilon_t\}$ is an i.i.d. sequence of standard normals, what is the distribution of $y_t - y_{t-j}$?
   (c) What is the joint distribution of $y_t - y_{t-j}$ and $y_{t-h} - y_{t-j-h}$?
3. Let $y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t$ where $\{\epsilon_t\}$ is a WN process.
   (a) Derive an explicit expression for the 1-step and 2-step ahead forecast errors, $e_{t+h|t} = y_{t+h} - \hat{y}_{t+h|t}$ where $\hat{y}_{t+h|t}$ is the MSE optimal forecast where $h = 1$ or $h = 2$.
   (b) What is the autocorrelation function of a time-series of forecast errors $\{e_{t+h|t}\}$ for $h = 1$ and $h = 2$?
   (c) Generalize the above to a generic h? (In other words, leave the solution as a function of h).
   (d) How could you test whether the forecast has excess dependence using an ARMA model?

Exercise 4.12. Suppose

$$y_t = \phi_0 + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t$$

where $\{\epsilon_t\}$ is a white noise process.

1. Precisely describe the two types of stationarity.
2. Why is stationarity a useful property?
3. What conditions on the model parameters are needed for $\{y_t\}$ to be covariance stationary?
4. Describe the Box-Jenkins methodology for model selection.
   Now suppose that $\phi_1 = 1$ and that ϵ_t is homoskedastic.
5. What is $\mathrm{E}_t[y_{t+1}]$?
6. What is $\mathrm{E}_t[y_{t+2}]$?
7. What can you say about $\mathrm{E}_t[y_{t+h}]$ for $h > 2$?
8. What is $\mathrm{V}_t[y_{t+1}]$?
9. What is $\mathrm{V}_t[y_{t+2}]$?

10. What is the first autocorrelation, ρ_1?

Exercise 4.13. Which of the following models are covariance stationary, assuming $\{\epsilon_t\}$ is a mean-zero white noise process. If the answer is conditional, explain the conditions required. In any case, explain your answer:

1. $y_t = \phi_0 + 0.8y_{t-1} + 0.2y_{t-2} + \epsilon_t$
2. $y_t = \phi_0 + \phi_1 I_{[t>200]} + \epsilon_t$
3. $y_t = \alpha t + 0.8\epsilon_{t-1} + \epsilon_t$
4. $y_t = 4\epsilon_{t-1} + 9\epsilon_{t-2} + \epsilon_t$
5. $y_t = \epsilon_t + \sum_{j=1}^{\infty} \gamma_j \epsilon_{t-j}$

Exercise 4.14. Answer the following questions:

1. Consider the AR(2)
$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$$
   (a) Rewrite the model with Δy_t on the left-hand side and y_{t-1} and Δy_{t-1} on the right-hand side.
   (b) What restrictions are needed on ϕ_1 and ϕ_2 for this model to collapse to an AR(1) in the first differences?
   (c) When the model collapses, what does this tell you about y_t?
2. Discuss the important issues when testing for unit roots in economic time-series.

Exercise 4.15. In which of the following models are the$\{y_t\}$covariance stationary, assuming $\{\epsilon_t\}$ is a mean-zero white noise process. If the answer is conditional, explain the conditions required. In any case, explain your answer:

1. $\Delta y_t = -0.2y_{t-1} + \epsilon_t$
2. $y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$
3. $y_t = \phi_0 + 0.1x_{t-1} + \epsilon_t$, $x_t = x_{t-1} + \epsilon_t$
4. $y_t = 0.8y_{t-1} + \epsilon_t$

Exercise 4.16. Suppose
$$y_t = \phi_0 + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t$$
where $\{\epsilon_t\}$ is a white noise process.

1. Precisely describe the two types of stationarity.
2. Why is stationarity a useful property?

3. What conditions on the model parameters are needed for $\{y_t\}$ to be covariance stationary?
4. Describe the Box-Jenkins methodology for model selection.
5. Now suppose that $\phi_1 = 1$ and that ϵ_t is homoskedastic.
6. What is $\mathrm{E}_t[y_{t+1}]$?
7. What is $\mathrm{E}_t[y_{t+2}]$?
8. What can you say about $\mathrm{E}_t[y_{t+h}]$ for $h > 2$?
9. What is $\mathrm{V}_t[y_{t+1}]$?
10. What is $\mathrm{V}_t[y_{t+2}]$?
11. What is the first autocorrelation, ρ_1?

Exercise 4.17. Answer the following questions.

1. Suppose $y_t = \phi_0 + \phi_1 y + \phi_2 y_{t-2} + \epsilon_t$ where $\{\epsilon_t\}$ is a white noise process.
2. Write this model in companion form.
   (a) Using the companion form, derive expressions for the first two autocovariances of y_t. (It is not necessary to explicitly solve them in scalar form)
   (b) Using the companion form, determine the formal conditions for ϕ_1 and ϕ_2 to for $\{y_t\}$ to be covariance stationary. You can use the result that when $\mathbf{A}$ is a 2 by 2 matrix, its eigenvalues solve the two equations

$$\begin{aligned} \lambda_1\lambda_2 &= a_{11}a_{22} - a_{12}a_{21} \\ \lambda_1 + \lambda_2 &= a_{11} + a_{22} \end{aligned}$$

Exercise 4.18. Justify a reasonable model for each of these time series in Figure 4.17 using information in the autocorrelation and partial autocorrelation plots. In each set of plots, the left most panel shows that data ($T = 100$). The middle panel shows the sample autocorrelation with 95% confidence bands. The right panel shows the sample partial autocorrelation for the data with 95% confidence bands.

1. Panel (a)
2. Panel (b)
3. Panel (c)

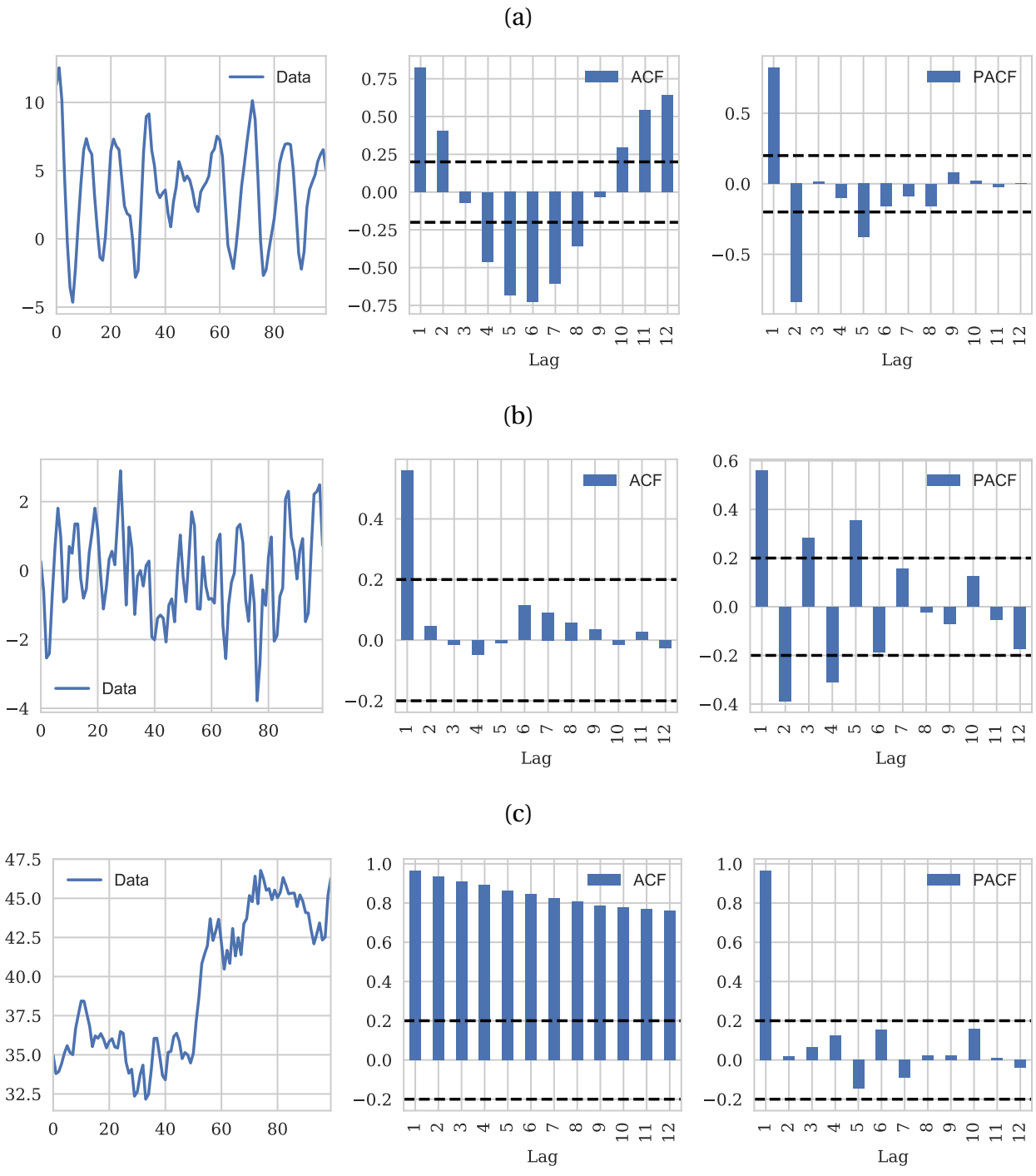

Figure 4.17: Plots for question 2(b).