- The *ARIMA*(p, d, q) class of models as a broad class can describe many real time series.
- *Model specification* for *ARIMA*(p, d, q) models involves
  1. Choosing appropriate values for p, d, and q;
  2. Estimating the parameters (e.g., the ϕ's, θ's, and σ_e^2) of the *ARIMA*(p, d, q) model;
  3. Checking model adequacy, and if necessary, improving the model.
- This process of iteratively proposing, checking, adjusting, and re-checking the model is known as the "Box-Jenkins method" for fitting time series models.

The Sample Autocorrelation Function

- We know that the autocorrelations are important characteristics of our time series models.
- To get an idea of the autocorrelation structure of a process based on observed data, we look at the sample autocorrelation function:
$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$
- We look for patterns in the r_k values that are similar to known patterns of the ρ_k for *ARMA* models that we have studied.
- Since r_k are merely estimates of ρ_k, we cannot expect the r_k patterns to match the ρ_k patterns of a model *exactly*.

The Sampling Distribution of r_k

- Because of its form and the fact that it is a function of possibly correlated variables, r_k does not have a simple sampling distribution.
- For large sample sizes, the approximate sampling distribution of r_k can be found, when the data come from ARMA-type models.
- This sampling distribution is approximately normal with mean ρ_k, so a strategy for checking model adequacy is to see whether the r_k's fall within 2 standard errors of their expected values (the ρ_k's).
- For our purposes, we consider several common models and find the approximate expected values, variances, and correlations of the r_k's under these models.

The Sampling Distribution of r_k Under Common Models

- First, under general conditions, for large n, r_k is approximately normal with expected value ρ_k.
- If $\{Y_t\}$ is white noise, then for large n, $var(r_k) \approx 1/n$ and $corr(r_k, r_j) \approx 0$ for $k \neq j$.
- If $\{Y_t\}$ is $AR(1)$ having $\rho_k = \phi^k$ for $k > 0$, then $var(r_1) \approx (1 - \phi^2)/n$.
- Note that r_1 has smaller variance when ϕ is near 1 or -1.
- And for large k:

$$var(r_k) \approx \frac{1}{n}\left[\frac{1+\phi^2}{1-\phi^2}\right]$$

- This variance tends to be larger for large k than for small k, especially when ϕ is near 1 or -1.
- So when ϕ is near 1 or -1, we can expect r_k to be relatively close to $\rho_k = \phi^k$ for small k, but not especially close to $\rho_k = \phi^k$ for large k.

The Sampling Distribution of r_k Under the $AR(1)$ Model

- In the $AR(1)$ model,

$$corr(r_1, r_2) \approx 2\phi\sqrt{\frac{1-\phi^2}{1+2\phi^2-3\phi^4}}$$

- For example, if $\phi = 0.9$, then $corr(r_1, r_2) = 0.97$. Similarly, if $\phi = -0.9$, then $corr(r_1, r_2) = -0.97$.
- If $\phi = 0.2$, then $corr(r_1, r_2) = 0.38$. Similarly, if $\phi = -0.2$, then $corr(r_1, r_2) = -0.38$.
- Exhibit 6.1 on page 111 of the textbook gives other example values for $corr(r_1, r_2)$ and $var(r_k)$ for selected values of k and ϕ, assuming an $AR(1)$ process.
- To determine whether a certain model (like $AR(1)$ with $\phi = 0.9$) is reasonable, we can examine the sample autocorrelation function and compare the observed values from our data to those we would expect, under that model.

The Sampling Distribution of r_k Under the $MA(1)$ Model

- For the $MA(1)$ model, $var(r_1) = (1 - 3\rho_1^2 + 4\rho_1^4)/n$ and $var(r_k) = (1 + 2\rho_1^2)/n$ for $k > 1$.
- Exhibit 6.2 on page 112 of the textbook gives other example values for $corr(r_1, r_2)$ and $var(r_k)$ for any k and selected values of θ, under the $MA(1)$ model.
- For the $MA(q)$ model, $var(r_k) = (1 + 2\sum_{j=1}^{q} \rho_j^2)/n$ for $k > q$.

The Need to Go Beyond the Autocorrelation Function

- The sample autocorrelation function (ACF) is a useful tool to check whether the lag correlations that we see in a data set match what we would expect under a specific model.
- For example, in an $MA(q)$ model, we know the autocorrelations should be zero for lags beyond q, so we could check the sample ACF to see where the autocorrelations cut off for an observed data set.
- But for an $AR(p)$ model, the autocorrelations don't cut off at a certain lag; they die off gradually toward zero.

Partial Autocorrelation Functions

- The partial autocorrelation function (PACF) can be used to determine the order p of an $AR(p)$ model.
- The PACF at lag k is denoted ϕ_{kk} and is defined as the correlation between Y_t and Y_{t-k} after removing the effect of the variables in between: $Y_{t-1}, \ldots, Y_{t-k+1}$.
- If $\{Y_t\}$ is a normally distributed time series, the PACF can be defined as the correlation coefficient of a *conditional* bivariate normal distribution:

$$\phi_{kk} = corr(Y_t, Y_{t-k} | Y_{t-1}, \ldots, Y_{t-k+1})$$

Partial autocorrelations in the $AR(p)$ and $MA(q)$ Processes

- In the $AR(1)$ process, $\phi_{kk} = 0$ for all $k > 1$.
- So the partial autocorrelation for lag 1 is not zero, but for higher lags, it is zero.
- More generally, in the $AR(p)$ process, $\phi_{kk} = 0$ for all $k > p$.
- Clearly, examining the PACF for an AR process can help us determine the order of that process.
- For an $MA(1)$ process,

$$\phi_{kk} = \frac{\theta^k(1-\theta^2)}{1-\theta^{2(k+1)}}$$

- So the partial autocorrelation of an $MA(1)$ process never equals zero exactly, but it decays to zero quickly as k increases.
- In general, the PACF for a $MA(q)$ process behaves similarly as the ACF for an AR process of the same order.

Expressions for the Partial Autocorrelation Function

- For a stationary process with known autocorrelations $\rho_1, \ldots, \rho_k$, the ϕ_{kk} satisfy the Yule-Walker equations:

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \cdots + \phi_{kk}\rho_{j-k}, \text{ for } j = 1, 2, \ldots, k$$

- For a given k, these equations can be solved for $\phi_{k1}, \phi_{k2}, \ldots, \phi_{kk}$ (though we only care about ϕ_{kk}).
- We can do this for all k.
- If the stationary process is actually $AR(p)$, then $\phi_{pp} = \phi_p$, and the order p of the process is whatever is the highest lag with a nonzero ϕ_{kk}.

The Sample Partial Autocorrelation Function

- By replacing the ρ_k's in the previous set of linear equations by r_k's, we can solve these equations for *estimates* of the ϕ_{kk}'s.
- Equation (6.2.9) on page 115 of the textbook gives a formula for solving recursively for the ϕ_{kk}'s in terms of the ρ_k's.
- Replacing the ρ_k's by r_k's, we get the sample partial autocorrelations, the $\hat{\phi}_{kk}$'s.

Using the Sample PACF to Assess the Order of an AR Process

- If the model of $AR(p)$ is correct, then the sample partial autocorrelations for lags greater than p are approximately normally distributed with means 0 and variances $1/n$.
- So for any lag $k > p$, if the sample partial autocorrelation $\hat{\phi}_{kk}$ is within 2 standard errors of zero (between $-2/\sqrt{n}$ and $2/\sqrt{n}$), then this indicates that we do not have evidence against the $AR(p)$ model.
- If for some lag $k > p$, we have $\hat{\phi}_{kk} < -2/\sqrt{n}$ or $\hat{\phi}_{kk} > 2/\sqrt{n}$, then we may need to change the order p in our model (or possibly choose a model other than AR).
- This is a somewhat informal test, since it doesn't account for the multiple decisions being made across the set of values of $k > p$.

Summary of ACF and PACF to Identify $AR(p)$ and $MA(q)$ Processes

- The ACF and PACF are useful tools for identifying pure $AR(p)$ and $MA(q)$ processes.
- For an $AR(p)$ model, the true ACF will decay toward zero.
- For an $AR(p)$ model, the true PACF will cut off (become zero) after lag p.
- For an $MA(q)$ model, the true ACF will cut off (become zero) after lag q.
- For an $MA(q)$ model, the true PACF will decay toward zero.
- If we propose either an $AR(p)$ model or a $MA(q)$ model for an observed time series, we could examine the sample ACF or PACF to see whether these are close to what the true ACF or PACF would look like for this proposed model.

Extended Autocorrelation Function for Identifying ARMA Models

- For an $ARMA(p, q)$ model, the true ACF and true PACF both have infinitely many nonzero values.
- Neither the true ACF nor the true PACF will cut off entirely after a certain number of lags.
- So it is hard to determine the correct orders of an $ARMA(p, q)$ model simply by using the ACF and PACF.
- The *extended autocorrelation function* (EACF) is one method proposed to assess the orders of a $ARMA(p, q)$ model.
- Other methods for specifying $ARMA(p, q)$ models include the corner method and the smallest canonical correlation (SCAN) method, which we will not discuss here.

Some Details about the Extended Autocorrelation Function Method

- In an $ARMA(p, q)$ model, if we "filter out" (i.e., subtract off) the autoregressive component(s), we are left with a pure $MA(q)$ process that can be specified using the ACF approach.
- For example, consider an $ARMA(1, 1)$ model:

$$Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1}.$$

- If we regress Y_t on Y_{t-1}, we get an inconsistent estimator of ϕ, but this regression's residuals do tell us about the behavior of the error process $\{e_t\}$.
- So then let's regress Y_t on Y_{t-1} AND the lag 1 of the first regression's residuals, which stand in for e_{t-1}.
- In this second regression, the estimated coefficient of Y_{t-1} (call it $\tilde{\phi}$) is a consistent estimator of ϕ.
- Then $W_t = Y_t - \tilde{\phi} Y_{t-1}$, having "filtered out" the autoregressive part, should be approximately $MA(1)$.

More Details about the Extended Autocorrelation Function Method

- For higher-order ARMA processes, we would need more of these sequential regressions to consistently estimate the AR coefficients (we'd need q extra regressions for an $AR(p, q)$ model).
- In practice, both AR order p and MA order q are unknown, so we need to do this iteratively, considering grids of values for p and q.
- This iterative estimation of AR coefficients, assuming a hypothetical $ARMA(k, j)$ model, produces "filtered" values

$$W_{t,k,j} = Y_t - \tilde{\phi}_1 Y_{t-1} - \cdots - \tilde{\phi}_k Y_{t-k}$$

Extended Sample Autocorrelations

- The *extended sample autocorrelations* are the sample autocorrelations of $W_{t,k,j}$.
- If the hypothesized AR order k is actually the correct AR order, p, and if the hypothesized MA order $j \geq q$, then $\{W_{t,k,j}\}$ is an $MA(q)$ process.
- In that case, the true autocorrelations of $W_{t,k,j}$ of lag $q + 1$ or higher should be zero.
- We can try finding the extended sample autocorrelations for a grid of values of $k = 0, 1, 2, \ldots$ and a grid of values of $j = 0, 1, 2, \ldots$.

The EACF in Table Form

- We can summarize the EACFs by creating a table with an "X" in the k-th row and j-th column if the lag $j+1$ sample autocorrelation of $W_{t,k,j}$ is significantly different from zero.
- Since the sample autocorrelations are approximately $N(0, 1/(n-k-j))$ under the $MA(j)$ process, the sample autocorrelation is significantly different from zero if its absolute value exceeds $1.96/\sqrt{n-j-k}$.
- The table gets an "0" in its k-th row and j-th column if the lag $j+1$ sample autocorrelation of $W_{t,k,j}$ is NOT significantly different from zero.

The EACF in Table Form (Continued)

- The EACF table for an $ARMA(p, q)$ process should theoretically have a triangular pattern of zeroes with the top-left zero occurring in the p-th row and q-th column (with the row and column labels both starting from 0).
- (In reality, the sample EACF table will not be as clear-cut as the examples that follow, since the sample EACF values have sampling variability.)

Theoretical EACF Table for an $ARMA(1, 1)$ Process

AR/MA	0	1	2	3	4	5	6	7	8	9	10
0	x	x	x	x	x	x	x	x	x	x	x
1	x	0	0	0	0	0	0	0	0	0	0
2	x	x	0	0	0	0	0	0	0	0	0
3	x	x	x	0	0	0	0	0	0	0	0
4	x	x	x	x	0	0	0	0	0	0	0
5	x	x	x	x	x	0	0	0	0	0	0
6	x	x	x	x	x	x	0	0	0	0	0
7	x	x	x	x	x	x	x	0	0	0	0

Theoretical EACF Table for an $ARMA(2,3)$ Process

AR/MA	0	1	2	3	4	5	6	7	8	9	10
0	x	x	x	x	x	x	x	x	x	x	x
1	x	x	x	x	x	x	x	x	x	x	x
2	x	x	x	0	0	0	0	0	0	0	0
3	x	x	x	x	0	0	0	0	0	0	0
4	x	x	x	x	x	0	0	0	0	0	0
5	x	x	x	x	x	x	0	0	0	0	0
6	x	x	x	x	x	x	x	0	0	0	0
7	x	x	x	x	x	x	x	x	0	0	0