Financial Data Analysis

Time Series Analysis, Part II

May 17, 2010
**Autoregressive Moving Average (ARMA) Time Series Models**

- ARMA($p, q$) is given by

\[
\phi(L)Y_t = \theta(L)\epsilon_t, \quad (1)
\]

where the lag polynomials

\[
\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p \quad (2)
\]

\[
\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q. \quad (3)
\]

- Assume that $\phi(z)$ and $\theta(z)$ have no roots in common.

- Mixing MA and AR parts often leads to more flexible models with less parameters than using pure MA or AR models.
• As MA processes are stationary, the stationarity of an ARMA model depends on the autoregressive polynomial $\phi(z)$.

• For an ARMA($p,q$) process, both the ACF and the PACF gradually die out (i.e., do not cut off).

• Invertibility: Roots of the MA polynomial outside the unit circle, i.e.,

\[ 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0 \Rightarrow |z| > 1. \]

• For the ARMA(1,1), $Y_t = \phi Y_{t-1} + \theta \epsilon_{t-1} + \epsilon_t$, the ACF is

\[ \rho(\tau) = \phi^{\tau-1} \frac{(\phi + \theta)(1 + \phi \theta)}{1 + 2\theta \phi + \theta^2}, \]

which for the AR(1), i.e., $\theta = 0$, becomes

\[ \rho(\tau) = \phi^\tau. \]
Autocorrelation function of ARMA(1,1) with $\phi = 0.8$

$$\rho(\tau) = \phi^{\tau-1} \frac{(\phi + \theta)(1 + \phi\theta)}{1 + 2\theta\phi + \theta^2}$$
Model Selection

- With forecasting in mind, we want parsimonious models that capture the most pronounced features of the data.
Model Selection: Inspection of the Correlogram

- We plot the sample analogues of the ACF and the PACF and check whether these suggest an appropriate specification.

- To review,

<table>
<thead>
<tr>
<th>Table 1: Behavior of theoretical correlation functions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>ACF</td>
</tr>
<tr>
<td>PACF</td>
</tr>
</tbody>
</table>

- The results in Table 1 suggest that this approach may be useful in particular for pure MA and AR processes.
Estimation of the autocorrelation function (PACF): Sample ACF (SACF)

- With an observed time series of length $T$,

\[
\hat{\mu} = \bar{Y} = \frac{1}{T} \sum_{t=1}^{T} Y_t
\]

\[
\hat{\gamma}(\tau) = \frac{1}{T} \sum_{t=1}^{T-\tau} (Y_{t+\tau} - \bar{Y}_T)(Y_t - \bar{Y}_T)
\]

\[
\hat{\rho}(\tau) = \frac{\hat{\gamma}(\tau)}{\hat{\gamma}(0)}
\]
Identifying pure MA($q$) processes

- If $Y_t$ is MA($q$) with independent white noise, then, for $\tau > q$, the sample autocorrelation $\hat{\rho}(\tau)$ is, for large samples, approximately normal with variance

$$\text{Var}(\hat{\gamma}(\tau)) = \frac{1}{T} \left( 1 + 2 \sum_{i=1}^{q} \rho(i)^2 \right), \quad \tau = q + 1, q + 2, \ldots, \quad (5)$$

where $T$ is the sample size (number of observations), which is known as Bartlett’s formula.

- To use this result, we estimate the autocorrelations in (5) by their sample counterparts and then successively construct confidence intervals for $\hat{\rho}(1), \hat{\rho}(2), \hat{\rho}(3), \ldots$ via $\pm 2/\sqrt{T}$, $\pm 2\sqrt{(1 + 2\hat{\rho}(1)^2)/T}$, $\pm 2\sqrt{(1 + 2\hat{\rho}(1)^2 + 2\hat{\rho}(2)^2)/T}$ etc.

- In practice, the confidence intervals are often approximated by $\pm 2/\sqrt{T}$, which corresponds to a strict white noise process.
That is, if $Y_t$ is strict white noise, then

$$
\hat{\rho}(\tau) \underset{asy}{\sim} \text{Normal} \left(0, \frac{1}{T}\right),
$$

and the sample autocorrelation coefficients at different lags are asymptotically independent.

Result (6) can be used to test whether a time series is strict white noise, but not just uncorrelatedness.

Such tests are often based on the **Ljung–Box–Pierce** statistic

$$
Q = T(T + 2) \sum_{\tau=1}^{K} \frac{\hat{\rho}(\tau)^2}{T - \tau} \sim \chi^2(K).
$$

The $Q$ statistic is also useful for testing the appropriateness of a fitted ARMA model, but, if applied to the residuals of a fitted model, the number of degrees of freedom of the asymptotic $\chi^2$ distribution needs to be adjusted.
Estimation of the partial autocorrelation function (PACF): Sample PACF (SPACF)

- Fit successively autoregressive models of orders 1, 2, 3, ... by least squares, so that (with demeaned data)

\[ Y_t = \hat{\phi}_{k1}Y_{t-1} + \hat{\phi}_{k2}Y_{t-2} + \cdots + \hat{\phi}_{kk}Y_{t-k} + \hat{\epsilon}_t, \quad (8) \]

where \( \hat{\phi}_{kj} \) is the coefficient of \( Y_{t-j} \) in the fitted autoregressive model with \( k \) lags.

- Then take

\[ \hat{\pi}(k) = \hat{\phi}_{kk}, \quad k = 1, 2, \ldots \quad (9) \]
Identifying pure AR($p$) processes

- To see whether a specific sample partial autocorrelation coefficient can be deemed different from zero, the following result can be used.

- If the true process is an AR($p$), the estimated partial autocorrelations (9) of order $p + 1, p + 2, \ldots$ are approximately normally distributed with variance $1/T$, i.e.,

$$\text{Var}(\hat{\pi}(k)) \approx \frac{1}{T}, \quad k \geq p + 1.$$  

where $T$ is number of observations.
Information Criteria: AIC and BIC

- In practice, often no straightforward “textbook pattern” of the SACF or SPACF shows up, and alternative model selection criteria have to be considered.

- If models are fitted by maximum likelihood, models can be compared by means of the values of their respective maximized likelihood function.

- In general, for any model, the Akaike information criterion (AIC) is given by

\[
AIC = -2 \log L + 2K,
\]

(10)

where \( L \) is the value of the maximized likelihood, and \( K \) is the number of parameters of a model. Whereas increasing the order of the model will always increase the likelihood, the second component is a penalty factor for inclusion of additional parameters.

- Smaller values of AIC are preferred.
• The Bayesian information criterion (BIC) of Schwarz is given by

$$BIC = -2 \log L + K \log T,$$

where $T$ is the sample size.

• The logic is the same as for AIC, but since $\log(T) > 2$ for $T > 8$, BIC is more conservative than AIC in that it penalizes additional parameters more heavily than the AIC, i.e., it tends to select more parsimonious models.

• As for the AIC, smaller values of BIC are preferred.

• To use these criteria, we fix upper bounds for $p$ and $q$, $p_{\text{max}}$ and $q_{\text{max}}$, and then consider all ARMA($p,q$) constellations for $0 \leq p \leq p_{\text{max}}$ and $0 \leq q \leq q_{\text{max}}$. Then we can pick the optimal specification according to either AIC and BIC.

• As for theoretical properties, if the model with the “true” orders $p$ and $q$ is in the set of models considered, BIC is consistent, whereas AIC tends to overestimate the orders, even asymptotically.
Model Selection

- Model selection may also be based on the forecasting performance. To this end, one divides the sample at hand artificially into an “in–sample” and “out–of–sample” period, estimates the parameters of competing specifications over the “in–sample” period and compares their respective MSEs (or any other measure of forecasting accuracy) over the “out–of–sample” period. This requires, of course, that our sample is sufficiently long.
Estimation

- Yule–Walker
- *exact* and *conditional* maximum likelihood
- least squares
Maximum Likelihood Methods

- Let $\theta = (c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)$ denote the parameter vector to be estimated.

- If we have observed a sample of size $T$, $y_1, y_2, \ldots, y_T$, we can calculate the density

$$f(y_T, y_{T-1}, \ldots, y_1; \theta). \quad (12)$$

- The maximum likelihood estimator (MLE) of $\theta$ is the value for which (12) is maximized as a function of $\theta$, given our sample $y_1, y_2, \ldots, y_T$.

- To calculate (12), we need to make a distributional assumption about the innovations $\epsilon_t$ of the process.
Maximum Likelihood Methods

- For Gaussian processes, we have seen that *exact likelihood estimation* is feasible.

- To illustrate, consider the AR(1) process.

- The process is of the form

\[
Y_t = c + \phi Y_{t-1} + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),
\]

and \( \theta = (c, \phi, \sigma^2) \).

- It is useful to write the joint distribution as

\[
f(y_T, y_{T-1}, \ldots, y_1) = f(y_1) f(y_2|y_1) f(y_3|y_1, y_2) \cdots f(y_T|y_1, y_2, \ldots, y_{T-1}).
\]

\[\text{1}^{\text{Parameters are dropped in the densities to simplify the notation.}}\]
We know that the marginal distribution of the process is also Gaussian, so for $f(y_1)$ we can write

$$Y_1 \sim \text{Normal} \left( \frac{c}{1 - \phi}, \frac{\sigma^2}{1 - \phi^2} \right),$$

i.e.,

$$f(y_1) = \frac{1}{\sqrt{2\pi} \sqrt{\sigma^2/(1 - \phi^2)}} \exp \left\{ -\frac{(y_1 - c/(1 - \phi))^2}{2\sigma^2/(1 - \phi^2)} \right\}.$$

Moreover,

$$Y_2|Y_1 \sim \text{Normal}(c + \phi Y_1, \sigma^2),$$

and more generally,

$$Y_t|Y_{t-1}, Y_{t-2}, \ldots, Y_1 = Y_t|Y_{t-1} = \text{Normal}(c + \phi Y_{t-1}, \sigma^2),$$

so, for $t = 2, \ldots, T$,

$$f(y_t|y_{t-1}, y_{t-2}, \ldots, y_1) = f(y_t|y_{t-1}) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2} \right\}.$$
• The density of the sample is

\[ f(y_T, y_{T-1}, \ldots, y_1; \theta) = f(y_1) \prod_{t=2}^{T} f(y_t|y_{t-1}), \]

and the log-likelihood, which we prefer to optimize, is

\[
\begin{align*}
\log L(\theta) &= \log f(y_T, y_{T-1}, \ldots, y_1; \theta) = \log f(y_1) + \sum_{t=2}^{T} \log f(y_t|y_{t-1}) \\
&= -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log \sigma^2 + \frac{1}{2} \log(1 - \phi^2) - \frac{(y_1 - c/(1 - \phi))^2}{2\sigma^2/(1 - \phi^2)} \\
&\quad - \sum_{t=2}^{T} \frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}.
\end{align*}
\]
• The extension to general AR($p$) processes is conceptually straightforward, and efficient algorithms have been developed for calculating the density of the first $p$ observations.\(^2\)

• Alternatively, we can employ *conditional* maximum likelihood, which conditions on the first $p$ observations.

• This greatly simplifies the calculation of the MLE, and is asymptotically equivalent to the exact MLE (and the difference is negligible for moderately large samples).

• For Gaussian AR($p$) processes, this turns out to be equivalent to least squares estimation.

• For nonnormal innovations, which are ubiquitous in financial applications, conditional maximum likelihood is usually the only way to go, since the marginal distributions of the processes are not known.

• For the AR($p$) process, conditional maximum likelihood estimation, i.e., conditioning on the first $p$ observations $y_1, \ldots, y_p$, amounts to optimizing

$$
\sum_{t=p+1}^{T} \log f(y_t | y_{t-1}, \ldots, y_{t-p})
$$

$$
= -\frac{T-p}{2} \log(2\pi) - \frac{T-p}{2} \log \sigma^2
- \sum_{t=p+1}^{T} \frac{(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p})^2}{2\sigma^2}.
$$

• It is straightforward to see that (13) will be maximized with respect to the parameters of the mean equation when

$$
\sum_{t=p+1}^{T} (y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p})^2,
$$

is minimized.
That is, conditional maximum likelihood is equivalent to ordinary least squares for Gaussian AR\((p)\) processes.

The error variance is estimated via

\[
\hat{\sigma}^2 = \frac{1}{T-p} \sum_{t=p+1}^{T} (y_t - \hat{c} - \hat{\phi}_1 y_{t-1} - \hat{\phi}_2 y_{t-2} - \cdots - \hat{\phi}_p y_{t-p})^2. \tag{14}
\]

The extension of conditional maximum likelihood estimation to AR\((p)\) processes under Gaussianity is straightforward, as it amounts to OLS regression of \(y_t\) on a constant and \(p\) lags.

Under Gaussianity, the Yule–Walker, exact maximum likelihood and conditional maximum likelihood estimators are asymptotically equivalent with the same limiting distribution (see the discussion of the Yule–Walker estimator).

For small samples, however, the Yule–Walker estimator is less efficient than its alternatives.
• If the series is not Gaussian, however, then the true MLE (true distribution of the errors is known) will be more efficient than least squares estimation, although the least squares estimator will still produce reasonable results (is consistent, but not efficient).

• For example, if we assume that the innovations in an AR(1) process follow a $t$ distribution with $\nu$ degrees of freedom, then, for $2 \leq t \leq T$,

$$f(y_t|y_{t-1}) = \frac{\Gamma \left( \frac{\nu+1}{2} \right)}{\Gamma(\nu/2)\sqrt{\nu\pi\sigma}} \left\{ 1 + \frac{(y_t - c - \phi_1 y_{t-1})^2}{\nu \sigma^2} \right\}^{-(\nu+1)/2},$$

and the log–likelihood needs to be simultaneously maximized with respect to all parameters

$$\theta = (c, \phi_1, \sigma^2, \nu).$$

(15)

• Student’s $t$ distributions are frequently employed in financial applications.
MA and ARMA Processes

- Exact maximum likelihood estimation for these processes is more involved, so we focus on conditional estimation.

- Conditional maximum likelihood estimation for the MA(1), \( Y_t = \theta \epsilon_{t-1} + \epsilon_t \), works as follows.

- We may assume that \( \epsilon_0 = 0 \) (its expectation) and then recursively calculate (note that the \( Y_t \)'s are what we observe, not the \( \epsilon_t \)s)

\[
\begin{align*}
Y_1 &= \epsilon_1 + \theta \epsilon_0 \Rightarrow \epsilon_1 &= Y_1 - \theta \epsilon_0 \\
Y_2 &= \epsilon_2 + \theta \epsilon_1 \Rightarrow \epsilon_2 &= Y_2 - \theta \epsilon_1 = Y_2 - \theta Y_1 + \theta^2 \epsilon_0 \\
Y_3 &= \epsilon_3 + \theta \epsilon_2 \Rightarrow \epsilon_3 &= Y_3 - \theta \epsilon_2 = Y_3 - \theta Y_2 + \theta^2 Y_1 - \theta^3 \epsilon_0 \\
&\vdots \\
Y_t &= \epsilon_t + \theta \epsilon_{t-1} \Rightarrow \epsilon_t &= Y_t - \theta \epsilon_{t-1} \\
&= Y_t - \theta Y_{t-1} + \theta^2 Y_{t-2} - \cdots + (-1)^{t-1} \theta^{t-1} Y_1 + (-1)^t \theta^t \epsilon_0,
\end{align*}
\]

which shows that this will work only for invertible MA processes, where \(|\theta| < 1\), so that the effect of imposing \( \epsilon_0 = 0 \) quickly dies out.
The conditional log–likelihood (under normality) is

$$\log L(\theta) = -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log \sigma^2 - \sum_{t=1}^{T} \frac{\epsilon_t^2}{2\sigma^2},$$

with the $\epsilon_t$s calculated recursively as above.

This procedure can be extended to (invertible) MA($q$) processes.

Similarly, conditional maximum likelihood estimation of stationary and invertible ARMA($p, q$) processes can be based on the first $p$ sample values $y_1, \ldots, y_p$ and setting the first $q$ innovations equal to zero to start the ARMA recursion, see Hamilton (1994, p. 132) for details.
Diagnostics: Residual–based

- After having estimated an ARMA model, the next step is to check the adequacy of the model.

- Such diagnostic checking may be based on the estimated ARMA errors,

\[ \hat{\epsilon}_t = Y_t - \hat{c} - \hat{\phi}_1 Y_{t-1} - \cdots - \hat{\phi}_p Y_{t-p} - \hat{\theta}_1 \hat{\epsilon}_{t-1} - \cdots - \hat{\theta}_q \hat{\epsilon}_{t-q}. \] (16)

- The residuals of an estimated model can be tested for white noise–like behavior.

- We might plot the SACF and SPACF of the \( \{\hat{\epsilon}_t\} \) and check whether these have any significant spikes.

- The **Ljung–Box–Pierce** statistic (7),

\[ Q = T(T + 2) \sum_{\tau=1}^{K} \frac{\hat{\rho}_\epsilon(\tau)^2}{T - \tau}, \] (17)
is useful for checking the whiteness of the residuals of a fitted ARMA\((p, q)\) model. However, when applied to the autocorrelations of the residuals of an estimated ARMA model, we have to account for the estimated parameters of the conditional mean equation, and the limiting distribution is

\[ Q^{asy} \sim \chi^2(K - p - q). \]

- Statistic (17) tests only for uncorrelatedness of the residuals, but not for independence. If the data are generated by a nonlinear process, an ARMA model may extract all autocorrelation from the data, but more complex temporal dependencies may still be present in the residuals.

- The presence of such dependencies in the absence of significant autocorrelations suggests that the class of ARMA models is inappropriate for the data at hand. Adapting (17), McLeod and Li (1983) proposed to detect (potential) temporal dependencies in higher-order moments by computing the autocorrelations of the squared residuals and considering the statistic

\[ Q^* = T(T + 2) \sum_{\tau=1}^{K} \frac{\hat{\rho}_c^2(\tau)^2}{T - \tau} \overset{asy}{\sim} \chi^2(K). \]  

(18)
• This test will also be useful to check the appropriateness of models for the *conditional volatility* of financial return data.

• We may also want to test for a specific distributional assumption.

• If normality is assumed, the Jarque–Bera test can be applied to the residuals of the model.

• That is,

\[
JB = TS^2/6 + T(\hat{\kappa} - 3)^2/24 \overset{asy}{\sim} \chi^2(2),
\]

where \( \hat{s} \) and \( \hat{\kappa} \) are the sample skewness and kurtosis of the residuals, respectively.

• It should be noted, however, that the \( \chi^2 \) distribution of JB will be valid only for large samples. For small samples, the properties of the sample moments of \( \{\hat{\epsilon}_t\} \) can deviate considerably from those of their theoretical counterparts. Thus, for small samples, results of the Jarque–Bera test have to be interpreted with caution.

• Visual checks such as QQ plots may also be useful.
Diagnostics: Overfitting

- If we want to check an ARMA\((p, q)\) model, we fit a more general model, e.g., ARMA\((p + 1, q)\) or ARMA\((p, q + 1)\).

- The appropriateness of the ARMA\((p, q)\) model may be doubted if
  
  (i) the parameters of the \((p, q)\)-model are unstable, i.e., adding further lags substantially modifies the estimates of these parameters,
  (ii) the new parameters are statistically significant,
  (iii) the inclusion of additional lags considerably reduces the variance of the models’ residuals (16).
Forecasting

• Let the data observed until period $t$ be collected in the information set

$$I_t = \{Y_s : s \leq t\}.$$ 

• Define

  – the one-step-ahead predictor $\hat{Y}_t(1)$
  – the $h$–step–ahead predictor $\hat{Y}_t(h)$

• $h$ is the forecast horizon, and $t$ is the forecast origin.

• The conditional expectation is

$$E_t(Y_{t+h}) := E(Y_{t+h} | I_t) = E(Y_{t+h} | Y_t, Y_{t-1}, \ldots). \quad (20)$$

• The mean–squared prediction error (MSE) is

$$\text{MSE}(\hat{Y}_t(h)) = E[(Y_{t+h} - \hat{Y}_t(h))^2] = E\{[(Y_{t+h} - E_t(Y_{t+h})) + (E_t(Y_{t+h}) - \hat{Y}_t(h))]^2\}. \quad (21)$$
• The MSE is minimized for \( \hat{Y}_t(h) = E_t(Y_{t+h}) \), i.e., the conditional mean is MSE–optimal.

• To see this, write (21) as

\[
MSE(\hat{Y}_t(h)) = MSE(E_t(Y_{t+h})) + E[(E_t(Y_{t+h}) - \hat{Y}_t(h))^2] \\
\quad - 2E[(Y_{t+h} - E_t(Y_{t+h}))(E_t(Y_{t+h}) - \hat{Y}_t(h))].
\]

• Now

- \( Y_{t+h} - E_t(Y_{t+h}) \) depends only on \( \epsilon_{t+1}, \epsilon_{t+2}, \ldots, \epsilon_{t+h} \), and
- \( E_t(Y_{t+h}) - \hat{Y}_t(h) \) depends only on \( \epsilon_t, \epsilon_{t-1}, \ldots \),

so

\[
E[(Y_{t+h} - E_t(Y_{t+h}))(E_t(Y_{t+h}) - \hat{Y}_t(h))] = 0,
\]

since

\[
E[(Y_{t+h} - E_t(Y_{t+h}))] = 0
\]

irrespective of \( \epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}, \ldots \)
Calculation of Forecasts

- Consider the AR(1) process, $Y_t = c + \phi_1 Y_{t-1} + \epsilon_t$.

- Forecasts can be calculated recursively by setting all future shocks equal to their expected value zero and for future $Y_{t+\tau}$, $\tau > 0$, we substitute predicted values.

- Thus, with $\mu = \frac{c}{1-\phi_1} = E(Y_t)$,

$$\hat{Y}_t(1) = c + \phi_1 Y_t$$
$$\hat{Y}_t(2) = c + \phi_1 \hat{Y}_t(1) = c + \phi_1 c + \phi_1^2 Y_t$$
$$\vdots$$
$$\hat{Y}_t(h) = c + \phi_1 \hat{Y}_t(h-1) = \cdots = c \sum_{i=0}^{h-1} \phi_1^i + \phi_1^h Y_t$$

$$= \frac{1 - \phi_1^h}{1 - \phi_1} + \phi_1^h Y_t$$
$$= \mu + \phi_1^h (Y_t - \mu) = \mu (1 - \phi_1^h) + \phi_1^h Y_t.$$
• Note that $\hat{Y}_t(h) \to \mu$ as $h$ becomes larger, so there is little predictability at large forecast horizons.

• To find the prediction error variance,

$$E[(Y_{t+h} - E_t(Y_{t+h}))^2],$$

we use

$$Y_{t+h} = \mu + \sum_{i=0}^{\infty} \phi_i^1 \epsilon_{t+h-i},$$

so

$$E_t(Y_{t+h}) = \mu + \sum_{i=h}^{\infty} \phi_i^1 \epsilon_{t+h-i},$$

and the $h$–step prediction error is

$$Y_{t+h} - E_t(Y_{t+h}) = \sum_{i=0}^{h-1} \phi_i^1 \epsilon_{t+h-i},$$
with prediction error variance

\[
E[(Y_{t+h} - E_t(Y_{t+h}))^2] = \sigma^2 \sum_{i=0}^{h-1} \phi^2_i = \frac{\sigma^2(1 - \phi^{2h})}{1 - \phi^2}.
\]  

(24)

- If we add the assumption of normality, we have

\[
Y_{t+h} \sim \text{Normal} \left( E_t(Y_{t+h}), \frac{\sigma^2(1 - \phi^{2h})}{1 - \phi^2} \right),
\]  

(25)

which, in order to assess to precision of the forecasts, can be used to construct (e.g., 95%) prediction intervals based on the conditional normal distribution of \( Y_{t+h} \),

\[
E_t(Y_{t+h}) \pm 1.96\sigma \sqrt{\frac{1 - \phi^{2h}}{1 - \phi^2}}.
\]  

(26)

- Note that these calculations do not account for the additional uncertainty introduced by estimation error.
• As the forecast horizon increases, the length of the prediction interval also increases, and the forecast error converges to the unconditional variance.

• So uncertainty increases with increasing forecasting horizon, but the added uncertainty from forecasting farther into the future becomes negligible as $h$ increases.
$Y_t = 0.95Y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim N(0,1)$, value at forecast origin: $Y_t = E(Y_t) = 0$
The extension of the AR(1) case to the AR($p$) case is straightforward:

For forecast horizon $h = 1$, point forecast $\hat{Y}_t(1)$ is formed just by substituting estimates into the model definition and setting $\epsilon_{t+1}$ equal to its expectation zero,

$$\hat{Y}_t(1) = c + \sum_{i=1}^{p} \hat{\phi}_i Y_{t-i}.$$ 

For calculating $h$–step forecasts, $\epsilon_{t+1}, \ldots, \epsilon_{t+h}$ are replaced by their expected values of zero, and

- for $Y_s, s \leq t$, the observed value is used,
- otherwise the forecast $\hat{Y}_t(s)$ is used.
To illustrate, for \( p = 3 \) \( (Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \epsilon_t) \),

\[
\hat{Y}_t(1) = \hat{\phi}_1 Y_t + \hat{\phi}_2 Y_{t-1} + \hat{\phi}_3 Y_{t-2}, \\
\hat{Y}_t(2) = \hat{\phi}_1 \hat{Y}_t(1) + \hat{\phi}_2 Y_t + \hat{\phi}_3 Y_{t-1}, \\
\hat{Y}_t(3) = \hat{\phi}_1 \hat{Y}_t(2) + \hat{\phi}_2 \hat{Y}_t(1) + \hat{\phi}_3 Y_t,
\]

\[
\vdots
\]

\[
\hat{Y}_t(h) = \hat{\phi}_1 \hat{Y}_t(h-1) + \hat{\phi}_2 \hat{Y}_t(h-2) + \hat{\phi}_3 \hat{Y}_t(h-3), \quad h > p.
\]

For the ARMA\((p, q)\),

\[
Y_{t+h} = \phi_1 Y_{t+h-1} + \cdots + \phi_p Y_{t+h-p} + \epsilon_{t+h} + \theta_1 \epsilon_{t+h-1} + \cdots + \theta_q \epsilon_{t+h-q},
\]

we then set

- \( Y_s, s \leq t \), equal to their observed values,
- \( Y_s, s > t \), equal to their predictions \( \hat{Y}_t(s) \),
- \( \epsilon_s, s \leq t \), equal to the model residuals \( \hat{\epsilon}_s \) defined in (16),
- \( \epsilon_s, s > t \), equal to their expectation zero.
• Calculation of the prediction error is similar to the AR(1) case:

• We have the MA($\infty$) representation

$$ y_{t+h} = \sum_{i=0}^{\infty} \psi_i \epsilon_{t+h-i}, $$

so

$$ E_t(y_{t+h}) = \sum_{i=h}^{\infty} \psi_i \epsilon_{t+h-i}, $$

and the $h$–step prediction error is

$$ y_{t+h} - E_t(y_{t+h}) = \sum_{i=0}^{h-1} \psi_i \epsilon_{t+h-i} $$

with prediction error variance

$$ E[(y_{t+h} - E_t(y_{t+h}))^2] = \sigma^2 \sum_{i=0}^{h-1} \psi_i^2. $$
• For a Gaussian process, we have

\[ y_{t+h} \sim \text{Normal} \left( E_t(y_{t+h}), \sigma^2 \sum_{i=0}^{h-1} \psi_i^2 \right), \]

which can again be used to construct prediction intervals.

• As in the AR(1) case, these calculations do not account for estimation error.

• As the forecast horizon increases, the length of the prediction interval also increases, and the forecast error converges to the unconditional variance,

\[
E[(y_{t+h} - E_t(y_{t+h}))^2] = \sigma^2 \sum_{i=0}^{h-1} \psi_i^2 \xrightarrow{h \to \infty} \sigma^2 \sum_{i=0}^{\infty} \psi_i^2 = \text{Var}(Y_t), \quad (27)
\]

and \( E_t(y_{t+h}) \) converges to the unconditional mean \( c/(1 - \phi_1 - \phi_2 - \cdots - \phi_p) \).
Processes With a Unit Root

- Consider the AR(1) process, $Y_t = \phi Y_{t-1} + \epsilon_t$, with ACF $\rho(\tau) = \phi^\tau$.

- What happens if $\phi$ approaches unity?

- Put $\phi = 1 - \delta$ for small $\delta$; then since $(1 - \delta)^\tau \approx 1 - \tau \delta$ for moderate $\tau$, $\phi$ approaching unity will be reflected in an ACF declining rather slowly and in an almost linear fashion.
\[ \delta = 0.001 \]

\[ 1 - \delta \tau \]

\[ (1 - \delta)^\tau \]
log of exchange rate Euro/USD, 1990–2009

SACF of log exchange rate
• Consider the ARMA process

\[ \phi(L)Y_t = c + \theta(L)\epsilon_t. \] (28)

• Suppose that the autoregressive polynomial \( \phi(L) \) has a unit root, then we may write (28) as

\[ \phi^*(L)(1 - L)Y_t = c + \theta(L)\epsilon_t, \] (29)

where \( \phi^*(L) = \phi(L)/(1 - L) \).

• If there is no further unit root in \( \phi^*(L) \), then (29) implies that

\[ (1 - L)Y_t = Y_t - Y_{t-1} =: \Delta Y_t \]

follows a stationary ARMA process, where \( \Delta := 1 - L \) is the difference operator.

• Thus, the first difference of an ARMA process with a single unit root is a stationary ARMA process, where the order of the autoregressive polynomial is reduced by one.
• If there are \( d \) unit roots in \( \phi(L) \), then we can write
\[
\phi^*(L)(1 - L)^dY_t = c + \theta(L)\epsilon_t,
\]
and
\[
(1 - L)^dY_t = \Delta^dY_t \tag{30}
\]
is a stationary ARMA process, i.e., we have to difference \( d \) times to obtain stationarity.

• Such a process is referred to as an \textit{autoregressive integrated moving average} process, denoted ARIMA\((p, d, q)\), where \( p \) refers to the number of autoregressive lags \textit{not} counting the unit roots, \( d \) is the number of unit roots (or the \textit{order of integration}), and \( q \) is the number of moving average lags.

• \( Y_t \) is also referred to as being \textit{integrated of order} \( d \), written \( Y_t \sim I(d) \).

• We will mainly be interested in the case \( d = 1 \).
- Somewhat loosely speaking, the *Efficient Market Hypothesis* states that financial markets incorporate all information relevant in evaluating an asset instantaneously into its price.

- Only new information, which is then independent of the current price, can lead to further price changes. Consequently, the best prediction of the price in the future (e.g., tomorrow) is just the price today.

- We may then write the current price $p_t$ as a *random walk*,

\[ p_t = p_{t-1} + \epsilon_t, \]

where $\epsilon_t$ is the unpredictable shock that represents the impact of new information.

- Often the *random walk with drift*,

\[ p_t = c + p_{t-1} + \epsilon_t, \]

is more appropriate, since we expect stocks to generate a positive return.
• The random walk (with or without drift) is obviously an example of a process which is integrated of order 1.

• To illustrate the consequences random walk behavior, consider the random walk with drift,

\[
Y_t = c + Y_{t-1} + \epsilon_t. \quad (31)
\]

Assume that the process starts at \( t = 0 \) at some finite value \( Y_0 \). Then, by iterating equation (31),

\[
Y_t = Y_0 + ct + \sum_{i=1}^{t} \epsilon_i. \quad (32)
\]

• The solution (32) of the random walk with drift has two nonstationary components.
• The first is a linear deterministic trend, $ct$, and the second is given by the partial sum time series

$$S_t := \sum_{i=1}^{t} \epsilon_i,$$

which is called the stochastic trend.

• Alternatively, we might suppose that a time series exhibiting a linear trend is generated by a process which can be described by a sum of a linear trend component and a stationary component, i.e.,

$$Y_t = a + bt + \epsilon_t.$$  \hspace{1cm} (33)

• Process (33) can be made stationary by taking deviations from a fitted trend line, and it is therefore called a trend–stationary process.

• On the other hand, it is clear that taking deviations from a linear trend of a random walk with drift does not lead to a stationary process.
If a trend is fitted to an integrated process, all this does is to remove the drift in the random walk. Fitting a trend and taking deviations does not make an integrated process stationary. The deviations from the fitted trend are still a random walk, because all we have done is to remove the drift.

We have seen that, in this case, stationarity is obtained by taking first differences, and thus these processes are also referred to as difference–stationary processes.

Statistical tests of the null hypothesis that a time series has a unit root versus the alternative that it is stationary are called unit root tests.

E.g., the Dickey–Fuller tests are still frequently applied.

We will not go into this, but note that (log–)prices of e.g. stock returns are in all likelihood characterized by a unit root rather than a deterministic trend.
Is there any chance that this has been generated by a random walk?
\begin{itemize}
  \item $Y_t = c + Y_{t-1} + \epsilon_t$, where $c$ is the mean of DAX returns, and $\epsilon_t \sim t(3)$ (Student’s $t$ with 3 degrees of freedom) with standard deviation of the DAX returns.
\end{itemize}