Time Series Analysis:
Introduction to time series and forecasting

Handout 3: Estimation of ARMA(p,q) models, Impulse response functions and ARCH processes.

Laura Mayoral
IAE and Barcelona GSE
Introduction

- Goal: Given a univariate stationary process \( \{X_t\} \), we want to fit an ARMA\((p, q)\) model.

- 3 STEPS:
  - Identification: select the values for \( p \) and \( q \).
  - Estimation: Maximum likelihood. Exception: if process is AR\((p)\), OLS can be employed.
  - Diagnostic checking: check the accuracy of the proposed model by analyzing the residuals of the estimated model.
Estimation of ARMA(p,q) processes

For now, assume $(p,q)$ are known.

We will consider first estimation of AR processes (it’s simpler) and then we will move to estimation of general ARMA(p,q) processes.

Estimation of AR(p) processes

Two cases:

- $X_t$ follows an AR(p) ($= ARMA(p,0)$), for some finite $p$.
- $X_t$ follows an AR($\infty$) process.
Estimation of AR(p) for finite (and known) $p$

- Let $\{X_t\}$ be a stationary AR(p) process with $p < \infty$, i.e.,

$$X_t = c + \phi_0 + \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \epsilon_t,$$

where $\{\epsilon_t\}$ is a zero-mean innovation (either w.n., m.d.s. or $i.i.d.$) with variance $\sigma^2$.

- Estimation: OLS.

- The resulting estimator has the usual properties: $\sqrt{T}$-consistent and asymptotically normally distributed.
Sketch of the proof for the simplest case: Let $X_t$ be an AR(1) process given by

$$X_t = \phi X_{t-1} + \varepsilon_t,$$

where $\varepsilon_t$ is a m.d.s. and $|\phi| < 1$ (which implies that $X_t$ is stationary and ergodic).

Then,

$$\hat{\phi}_{ols} = \frac{\sum X_t X_{t-1}}{\sum X^2_{t-1}} = \phi + \frac{\sum X_{t-1}\varepsilon_t}{\sum X^2_{t-1}}.$$

Consistency

$$\hat{\phi}_{ols} \xrightarrow{p} \phi \iff \frac{\sum X_{t-1}\varepsilon_t}{\sum X^2_{t-1}} \xrightarrow{p} 0.$$

Under the previous assumptions, $X^2_t$ is also stationary and ergodic, thus by the LLN $\sum X^2_{t-1}/T \xrightarrow{p} \text{var} (X_{t-1}) = \sigma^2 / (1 - \phi^2)$. 
As for the numerator, notice that if $\varepsilon_t$ is a m.d.s., so is $\sum X_{t-1}\varepsilon_t$. Therefore, a LLN also applies to the numerator and $\sum X_{t-1}\varepsilon_t / T \xrightarrow{p} E(X_{t-1}\varepsilon_t) = 0$. Then

$$\hat{\phi}_{ols} = \phi + \frac{o_p(T)}{O_p(T)} = \phi + \frac{o_p(1)}{O_p(1)} = \phi + o_p(1)$$

Asymptotic normality

$$T^{1/2} (\hat{\phi}_{ols} - \phi) = \left( \frac{T^{-1/2} \sum X_{t-1}\varepsilon_t}{T^{-1} \sum X^2_{t-1}} \right) = \frac{T^{-1/2} \sum X_{t-1}\varepsilon_t}{\text{var}(X_t)} + o_p(1)$$

By the CLT for m.d.s,

$$T^{-1/2} \sum X_{t-1}\varepsilon_t \overset{d}{\to} N \left( 0, \text{var}(X_t)\sigma^2 \right)$$
Thus

\[ T^{1/2} \left( \hat{\phi}_{ols} - \phi \right) \xrightarrow{d} N \left( 0, \frac{\sigma^2}{\text{var}(X_t)} \right) = N \left( 0, (1 - \phi^2) \right). \]

But... what happens with the distribution of \( \hat{\phi}_{ols} \) if \( \phi \to 1 \)?

A similar proof can be established for a general stationary AR\((p)\) process.
Estimation of AR$(\infty)$ processes

- We know that any invertible ARMA process can be written as an AR$(\infty)$ process.

- Is it possible to estimate an AR$(\infty)$ process (that involves an infinite number of parameters!) with a finite sample?

- Answer: yes! Berk (1974) showed that under some assumptions one can obtain consistent estimates for the AR coefficients when the underlying process is AR$(\infty)$. 
Estimation of AR(∞) processes, II

Consider a stationary and invertible process \( X_t \) given by

\[
X_t = \sum_{i=1}^{\infty} \psi_i \varepsilon_{t-i}
\]

where \( \sum_{i=1}^{\infty} |\psi_i| < \infty \) and \( \{\varepsilon_t\} \) is an \((0, \sigma^2)\) i.i.d. sequence.

Since \( X_t \) is invertible, it can be written as

\[
X_t = \sum_{i=1}^{\infty} \phi_i X_{t-i} + \varepsilon_t.
\]

Berk showed that if an AR(k) process is fitted to \( X_t \), where \( k \) tends to infinity with the sample size, it is possible to obtain consistent and asymptotically normally distributed estimators of the relevant coefficients using OLS.
To achieve that, $k$ has to verify two conditions:

- an upper bound condition: $k^3/T \to 0$ (that says that $k$ should not increase too quickly),

- and a lower bound one, $T^{1/2} \sum_{j=k+1}^{\infty} \psi_i \to c \neq 0$ (that says that $k$ must not increase too slowly).

The above conditions are quite theoretical, they don’t give us much information about how to choose $k$ in real applications!

How to choose $k$ in practice?
How to choose $k$ in practice?

Standard information criteria (AIC, BIC) tend to choose values of $k$ that are too small (do not satisfy the lower bound condition), (Ng and Perron, 1995). In these cases, estimates are consistent but not asymptotically normal.

The **General-to-Specific** model selection technique verifies Berk’s conditions. Then, the resulting OLS estimators are consistent and asymptotically normal when $k$ is choosing using that criterion (Kuersteiner, 2005).
Estimation of ARMA(p,q) process: Maximum likelihood estimation

Let \( \{X_t\} \) be an ARMA(p,q) process

\[
\phi_p (L) X_t = \alpha + \theta_q (L) \varepsilon_t
\]

and let \( \delta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \alpha, \sigma^2)' \) be the vector containing all the unknown parameters.

Suppose that we have observed a sample of size \( T \): \((x_1, x_2, \ldots)\).

The ML approach will amount to calculate the joint density of \((X_T, \ldots, X_1)\)

\[
f_{X_T, X_{T-1}, \ldots, X_1}(x_T, x_{T-1}, \ldots, x_1; \delta) \tag{1}
\]

which might be loosely interpreted as the probability of having observed this particular sample.
The Maximum likelihood estimator (MLE) of $\delta$ is the value that maximizes (1), i.e., the probability of having observed this sample.

3 steps to obtain the ML estimator

- Specify a distribution for the innovations of the process, $\varepsilon_t$. Typically, $\varepsilon_t$ is a Gaussian white noise, $\varepsilon_t \sim iid \ N(0, \sigma^2)$.

- Compute the log likelihood function (1).

- Obtain the values of $\delta$ that maximize the log likelihood.

  Unless the process is a pure AR process (in which case OLS can be applied), there is no a closed-form solution. Thus, numerical optimization procedures should be applied in order to obtain the values of the estimates.
Some examples: Likelihood function of AR and MA processes

Computing the likelihood function for ARMA processes is complicated.

We will illustrate how this can be done with two simple examples: AR(1) and MA(1) processes.

Some examples: Likelihood function of an AR(1) process

Let \( \{X_t\} \) be an AR(1) process

\[
X_t = c + \phi X_{t-1} + \varepsilon_t,
\]

where \( |\phi| < 1 \) and \( \{\varepsilon_t\} \) is iid \( N(0,\sigma^2) \) and \( \delta' = (c, \phi, \sigma^2) \).
We now compute the density of the first observation $X_1$.

Clearly, since $\varepsilon_t$ is Gaussian, $X_t$ is also Gaussian with $E(X_1) = c/(1 - \phi)$ and $E(X_1 - \mu)^2 = \sigma^2/(1 - \phi^2)$, then

$$f_{X_1}(x_1; \theta) = \frac{1}{\sqrt{2\pi \sigma^2/(1 - \phi^2)}} e^{-\frac{(x_1 - c/(1 - \phi))^2}{2\sigma^2/(1 - \phi^2)}}.$$

Next, we consider the density of $X_2$ conditional on $X_1 = x_1$. It is clear that $X_2|(X_1 = x_1) \sim N(c + \phi x_1, \sigma^2)$ and then

$$f_{X_2|X_1}(x_2|x_1; \delta) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x_2 - c - \phi x_1)^2}{2\sigma^2}}.$$

As for the $X_3$, the distribution of $X_3$ conditional on $X_2 = x_2$ and $X_1 = x_1$ is

$$f_{X_3|X_2,X_1}(x_3|x_2, x_1; \delta) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x_3 - c - \phi x_2)^2}{2\sigma^2}}.$$
The joint distribution of the $X_1, X_2$ and $X_3$ can be written as

$$f_{X_3, X_2, X_1} (x_3, x_2, x_1; \delta)$$

$$= f_{X_3|X_2, X_1} (x_3| x_2, x_1; \delta) f_{X_2, X_1} (x_2, x_1; \delta)$$

$$= f_{X_3|X_2, X_1} (x_3| x_2, x_1; \delta) f_{X_2|X_1} (x_2|x_1; \delta) f_{X_1} (x_1; \delta)$$

Furthermore, notice that the values of $X_1, X_2, ..., X_{t-1}$ matter for $X_t$ only through the value of $X_{t-1}$, and then

$$f_{X_t|X_{t-1},...,X_1} (x_t| x_{t-1},..., x_1; \delta)$$

$$= f_{X_t|X_{t-1}} (x_t| x_{t-1}; \delta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_t-c-\phi x_{t-1})^2}{2\sigma^2}}. \quad (3)$$

Thus, the joint density of the first $t$ observations is

$$f_{X_T, X_{T-1},...,X_1} (x_T, x_{T-1},..., x_1; \delta)$$

$$= f_{X_t|X_{t-1}} (x_t| x_{t-1}; \delta) f_{X_{t-1}|X_{t-2}} (x_{t-1}| x_{t-2}; \delta) ... f_{X_1} (x_1; \delta) \quad (4)$$

$$= f_{X_1} (x_1; \delta)^T = f_{X_t|X_{t-1}} (x_t| x_{t-1}; \delta). \quad (5)$$
Taking logs,

\[ \mathcal{L}(\delta) = \log(f_{X_1}(x_1; \delta)) + \sum_{t=2}^{T} \log f_{X_t|X_{t-1}}(x_t|x_{t-1}; \delta) \]  \hspace{1cm} (6)

and \( \mathcal{L}(\delta) \) is called the log-likelihood function. Clearly, the value of \( \delta \) that maximizes (5) and (6) is the same but the maximization problem is simpler in the latter case, and then the log likelihood function is always preferred.

The next step would be to compute the value of \( \delta \) for which the exact log likelihood in (6) is maximized.

This amounts to deriving the log likelihood and equating the first derivative to zero.
The result is a system of nonlinear equations on $\delta$ and the sample for which there has no close solution in terms of $\left(x_1, ..., x_T\right)$. Then, iterative numerical procedures are needed to obtain $\delta$.

**Exact versus conditional MLE**

An alternative procedure is to regard the value of $x_1$ as deterministic and then we obtain the conditional MLE:

\[
\begin{align*}
\log f_{X_T, X_{T-1}, ..., X_1} (x_T, x_{T-1}, ..., x_1; \delta) &= \log(f_{X_t|X_{t-1}} (x_t|x_{t-1}; \delta) f_{X_{t-1}|X_{t-2}} (x_{t-1}|x_{t-2}; \delta) \\
&\quad \cdots f_{X_1} (x_1; \delta)) \\
&= \sum_{t=2}^{T} \log(f_{X_t|X_{t-1}} (x_t|x_{t-1}; \delta)) \\
&= - \left(\frac{T - 1}{2}\right) \log(2\pi\sigma^2) - \sum_{t=2}^{T} \left(\frac{(x_t - c - \phi x_{t-1})^2}{2\sigma^2}\right) .
\end{align*}
\]
Notice that the conditional MLE of $c$ and $\phi$ are obtained by minimizing

$$\sum_{t=2}^{T} \left( (x_t - c - \phi x_{t-1})^2 \right).$$

It follows that maximization of the log likelihood is equivalent to the minimization of the sum of squared residuals. More generally, the conditional MLE for an AR($p$) process can be obtained from an OLS regression of $y_t$ on a constant and $p$ of its own lagged values.

In contrast to exact MLE, the conditional maximum likelihood is trivial to compute.

Furthermore, if the sample size $T$ is sufficiently large, the first observation makes a negligible contribution to the total likelihood provided that $|\phi| < 1.$
The conditional MLE of the innovation variance is found by differentiating (10) with respect to $\sigma^2$ and setting the result equal to zero. It can be checked that

$$\hat{\sigma}^2 = \sum_{t=2}^{T} \frac{\left(x_t - \hat{c} - \hat{\phi}x_{t-1}\right)^2}{T - 1},$$

and in general, the MLE estimator of $\sigma^2$ of an AR(p) process is given by the sum of squared residuals over $(T - p)$
Example 2: Likelihood function for a MA(1) process

- Let \( \{X_t\} \) be a Gaussian MA(1) process
  \[
  X_t = \mu + \varepsilon_t + \theta \varepsilon_{t-1},
  \]
  where \( \varepsilon_t \) is iid \( N(0, \sigma^2) \).

- Let \( \delta = (\mu, \theta, \sigma^2) \) be the population parameters to be estimated. Then, \( X_t|\varepsilon_{t-1} \sim N(\mu + \theta \varepsilon_{t-1}, \sigma^2) \) and
  \[
  f_{X_t|\varepsilon_{t-1}}(x_t|\varepsilon_{t-1}) = \frac{1}{\sqrt{2\sigma^2 \pi}} e\left(\frac{-(x_t - \mu - \theta \varepsilon_{t-1})^2}{2\sigma^2}\right)
  \]
It is not possible to compute this density since $\varepsilon_{t-1}$ is not directly observed from the data. Assume that $\varepsilon_0 = 0$. Then, all the sequence of innovations could be computed recursively as

\[
\begin{align*}
\varepsilon_1 &= x_1 - \mu, \\
\varepsilon_2 &= x_2 - \mu - \theta x_1, \\
&\vdots \\
\varepsilon_t &= x_t - \mu - \theta x_{t-1}.
\end{align*}
\]

Then, the conditional density is given by

\[
\begin{align*}
f_{X_t|X_{t-1}, \ldots, X_1}(x_t|x_{t-1}, \ldots, x_1, \varepsilon_0 = 0; \delta) &= f_{X_t|\varepsilon_{t-1}}(x_t|\varepsilon_{t-1}; \delta) \\
&= \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{-\varepsilon_t^2}{2\sigma^2}}
\end{align*}
\]
Conditional MLE

- The conditional log likelihood is

\[ L(\delta) = -\frac{T}{2} \log(2\pi\sigma^2) - \sum \frac{\varepsilon_t^2}{2\sigma^2}. \]

- The conditional log likelihood function is a complicated nonlinear function and, in contrast to the AR case, they should be found by numerical optimization.

- If \( \theta \) is far from 1 in absolute value, then the effect of imposing \( \varepsilon_0 = 0 \) will quickly die out. However, if \( \theta \) is over 1 in absolute value, the error of imposing this restriction accumulates over time.

- Then, if the output of estimating \( \theta \) is greater than 1 one should discard the results. The numerical optimization should be attempted again with the reciprocal of \( \hat{\theta} \) used as starting value.
Numerical optimization of the log likelihood

Once $\mathcal{L}(\delta)$ has been computed, the next step is to find the value $\delta$ that maximizes it.

With the exception of pure AR processes, for which closed analytical expressions of the estimators are available, numerical optimization procedures should be employed to obtain the estimates.

These procedures make different guesses for $\delta$, evaluate the likelihood at these values and try to infer from these there the value $\hat{\delta}$ for which $\delta$ is largest. See Hamilton, Section 5.7 for a description of these methods.

The search procedure may be greatly accelerated if the optimization algorithm begins with parameter values which are close to the optimum values. For this reason, simple preliminary estimates of $\delta$ are often employed to begin the search. See Brockwell and Davis, Sections 8.2-8.4 for a description of these preliminary estimators.
Asymptotic properties of ML Estimators

ML estimators have very good asymptotic properties.

Under certain conditions it can be shown that they are consistent, asymptotically normal and efficient, since the variance-covariance matrix equals the inverse of the Fisher information matrix.

Let \( \{X_t\} \) be an ARMA(p,q) process, \( \delta_0 \) be the vector containing the true parameter values and \( \hat{\delta} \) is the MLE of \( \delta \). Assuming that neither \( \delta_0 \) nor \( \hat{\delta} \) falls on the boundary of the parameter space then

\[
\sqrt{T} (\hat{\delta} - \delta_0) \overset{d}{\rightarrow} N(0, \mathcal{I}^{-1})
\]

where \( \mathcal{I} \) is the Fisher information matrix

\[
\mathcal{I} = -E \left( \frac{\partial^2 \mathcal{L}(\delta)}{\partial \delta \partial \delta'} \bigg|_{\delta=\delta_0} \right).
\]
Inference

An estimator of the variance-covariance matrix is

\[ \hat{\mathcal{I}} = -T^{-1} \begin{pmatrix} \frac{\partial^2 \mathcal{L}(\hat{\delta})}{\partial \delta \partial \delta'} \bigg|_{\delta=\hat{\delta}} \end{pmatrix}, \]

which is often calculated numerically.

Another popular estimator of the Fisher information matrix is the so-called outer product estimator

\[ \hat{\mathcal{I}} = -T^{-1} \sum_{t=1}^{T} (h_t(\hat{\delta})) (h_t(\hat{\delta}))', \]

where \( h_t(\hat{\delta}) = \frac{\partial \log f(y_t|y_{t-1},y_{t-2},...;\delta)}{\partial \delta} \bigg|_{\delta=\hat{\delta}} \) denotes the vector of derivatives of the log of the conditional density of the t-th observation with respect to the elements of the parameter vector \( \delta \), evaluated at \( \hat{\delta} \).
If the process is correctly specified, both estimators of the variance-covariance matrix should yield similar values.

Thus, if the two estimators differ a great deal, it might be the sign of incorrect specification. See White (1982) for a general test of model specification based on this idea.

These estimates can be used to construct asymptotic standard errors that in turn can be employed to construct confidence intervals and t-tests.

Another popular approaches to testing hypotheses about parameters that are estimated by ML are the Likelihood ratio and the Lagrange Multiplier test. See Hamilton, Chapter 5.8.
Identification of an ARMA process: Selecting $p$ and $q$

- How to select $p$ and $q$?

- ACF and PACF are of help but their usefulness is limited in the general ARMA($p$, $q$) case.

- Other tools: model selection criteria.

- Information criteria: are mechanisms designed for choosing an appropriate number of parameters to be included in the model.

- A model selection criterion that identifies the correct model asymptotically with probability 1 as $T$ tends to infinity is said to be consistent.
Information criteria

Let $k$ be the number of parameters included in a candidate model.

By minimizing a distance between the true and the candidates models (the Kullback-Leibler discrepancy, see Gourieux and Monfort), it is possible to arrive to expressions that usually have the general form:

\[ I_T(k) = \log(\hat{\sigma}_k^2) + k \frac{C(T)}{T}, \]

penalty term

Some popular IC are:

1. Akaike: $C(T) = 2$

2. Schwartz or Bayesian (BIC): $C(T) = \log T$

3. Hannan and Quinn (HQ): $C(T) = 2 \log (\log T)$
The number of parameters \( k \) to be included in the model is chosen as

\[
\hat{k} = \arg \min_{k \leq m} I_T(k)
\]

where \( m \) is some pre-specified maximum number of parameters.

Properties of the different IC

- The AIC is not consistent (i.e., it might not choose asymptotically the correct model with probability one).

- The AIC tends to overestimate the number of parameters to be included in the model. However, this does not mean that this method is useless. It minimizes the MSE for one-step ahead forecasting.
The BIC and the HQ criteria are consistent, that is
\[ \lim_{T \to \infty} P(\hat{k} = k_0) = 1. \] (12)

In fact, the consistency result (12) holds for any criterion of the type (11) with \( \lim_{T \to \infty} C(T)/T = 0 \) and \( \lim_{T \to \infty} C(T) = \infty \).

Notice, however, that BIC and HQ are more likely to underestimate \( k_0 \) than the AIC.

Underestimating \( k_0 \) has worse consequences than overestimating it!
Other model selection approaches

- General-to-specific criterion, (Ng and Perron, 1995). This method amounts to

  - set the maximum number of parameters to be estimated. For instance, an AR(k) model with \( k = 7 \).

  - Estimate the 'general' model, (an AR(7) in this case).

  - test for the statistical significance of the coefficients associated with the higher order lags—with either a \( t \) or an F test—(\( y_{t-7} \) in this case).

  - Exclude the nonsignificant parameters, reestimate the smaller model and test again for significance. Repeat until all the remaining estimates are significantly different from zero.
Summarizing: steps you should follow to fit an ARMA\((p,d)\) process

- You should always start by plotting the data, ACFs and PACFs.
- What are the main patterns you see in the time plot, in the ACFs and the PACFs?
- Most likely it will look nonstationary: the first step will be to find a stationary transformation. We’ll see how to do this.
- Plot again the transformed data, as well as its ACF and PACF.
- Use information criteria to choose \(p\) and \(q\).
- Estimate the resulting model by maximum likelihood.
Summarizing: steps you should follow to fit an ARMA(p,d) process, II

You can use different software to do this: MATLAB (but you need the econometrics toolbox), STATA, EVIEWS, GRETL...
Diagnostic checking

If the fitted model is appropriate, the residuals should behave in a similar way as the true innovations of the process.

Thus a way of checking that the model is correctly specified is to look at the residuals and see whether they behave in a similar way as the innovations of the model.

Innovations are typically assumed to be

i.) zero-mean

ii) homokedastic (ie. they should have constant variance)

iii) uncorrelated.

Thus the residuals should also look like this if the model is correct.
Signs of misspecification: cyclic or trended behavior, non-constant variance, etc.

Among the conditions above: i) is not very restrictive (i.e., models that can be very “wrong” can satisfy this property), ii) is more stringent, iii) lag of correlation for any lag is central for ensuring that the model is suitable.
Autocorrelation tests

- Goal: we want to test whether the residuals are autocorrelated.
- Let \( \{e_t\} \) be the sequence of residuals, given by

\[
e_t = X_t - \hat{X}_t,
\]

where \( \hat{X}_t \) are the fitted values.
- Then, the autocorrelation function of the residuals is given by

\[
\hat{\rho}_e(h) = \frac{\sum_{t=1}^{T-h} (e_t - \bar{e})(e_{t+h} - \bar{e})}{\sum_{t=1}^{T} (e_t - \bar{e})^2}, \quad h=1,2,...
\]
Box-Pierce statistic $Q$-statistic

Under the null hypothesis that the model is correctly specified, then

$$Q_e(H) = T \sum_{i=1}^{H} \hat{\rho}_e^2(i) \overset{d}{\to} \chi^2_{(H-p-q)}.$$ 

Thus, we reject that the model is correctly specified at level $\alpha$ if

$$Q_e(H) = T \sum_{i=1}^{H} \hat{\rho}_e^2(i) > \chi^2_{1-\alpha(H-p-q)}.$$
Ljung-Box test

- Ljung and Box (1978) suggest replacing $Q_e$ by

$$Q_e^*(H) = n(n + 2) \sum_{i=1}^{H} \hat{\rho}_e^2(i) / (n - i),$$

- This statistic offers a better approximation to the $\chi^2$ distribution.

$$Q_e^*(H) = T \sum_{i=1}^{H} \hat{\rho}_e^2(i) \xrightarrow{d} \chi^2_{(H-p-q)}.$$
Tests for zero-mean

- The estimated residuals are not subject to the restriction that the average of the residuals is equal to zero.

- This condition is only imposed when estimating AR models including a constant.

- We can test this condition using the statistic

\[
\frac{\bar{e}}{\bar{\sigma}_e / \sqrt{T}}
\]

- And will reject the null of zero-mean is the value of this statistic is large when compared to the critical values of a normal distribution.
Examples

We next examine several residual processes obtained after fitting ARMA (or ARIMA) models to some real datasets.
Example 89

Let's begin with a classical example: The airline time series. The figure gives the graph of the residuals of the series estimated using an ARIMA(0,1,1) × (0,1,1)_{12} model.
Example 89

The figure gives ACF of the residuals.

- No coefficient is significant and also Q statistic is not significant for all lags.
- Thus we conclude that, with this test, we find no evidence of serial dependence in the residuals.
The mean of the residuals is not significantly different from zero and the variability of the residuals, except for one possible outlier, seems constant over time.

Is normally distributed? Yes
Example 90

The figure gives the graph of the residuals of the vehicle registration series estimated using an ARIMA\((0, 1, 1) \times (1, 1, 1)_{12}\) model. Some noticeable outliers are observed.
Example 90

The figure gives ACF of the residuals.

<table>
<thead>
<tr>
<th>Auto-correlation</th>
<th>AC</th>
<th>Q-Stat</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.049</td>
<td>1.1092</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.004</td>
<td>1.1163</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.075</td>
<td>3.7297</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.098</td>
<td>8.1915</td>
<td>0.004</td>
</tr>
<tr>
<td>5</td>
<td>0.005</td>
<td>8.2039</td>
<td>0.017</td>
</tr>
<tr>
<td>6</td>
<td>0.029</td>
<td>8.5645</td>
<td>0.035</td>
</tr>
<tr>
<td>7</td>
<td>0.003</td>
<td>11.793</td>
<td>0.019</td>
</tr>
<tr>
<td>8</td>
<td>0.042</td>
<td>13.607</td>
<td>0.027</td>
</tr>
<tr>
<td>9</td>
<td>0.031</td>
<td>13.066</td>
<td>0.042</td>
</tr>
<tr>
<td>10</td>
<td>0.081</td>
<td>16.169</td>
<td>0.024</td>
</tr>
<tr>
<td>11</td>
<td>0.060</td>
<td>17.839</td>
<td>0.022</td>
</tr>
<tr>
<td>12</td>
<td>0.025</td>
<td>18.135</td>
<td>0.034</td>
</tr>
<tr>
<td>13</td>
<td>0.023</td>
<td>18.363</td>
<td>0.049</td>
</tr>
<tr>
<td>14</td>
<td>0.007</td>
<td>18.409</td>
<td>0.073</td>
</tr>
<tr>
<td>15</td>
<td>0.005</td>
<td>18.423</td>
<td>0.103</td>
</tr>
<tr>
<td>16</td>
<td>0.008</td>
<td>20.032</td>
<td>0.081</td>
</tr>
<tr>
<td>17</td>
<td>0.029</td>
<td>21.019</td>
<td>0.101</td>
</tr>
<tr>
<td>18</td>
<td>0.053</td>
<td>22.371</td>
<td>0.098</td>
</tr>
<tr>
<td>19</td>
<td>0.011</td>
<td>22.430</td>
<td>0.130</td>
</tr>
<tr>
<td>20</td>
<td>0.029</td>
<td>22.526</td>
<td>0.155</td>
</tr>
<tr>
<td>21</td>
<td>0.001</td>
<td>24.092</td>
<td>0.152</td>
</tr>
<tr>
<td>22</td>
<td>0.000</td>
<td>24.092</td>
<td>0.193</td>
</tr>
<tr>
<td>23</td>
<td>0.071</td>
<td>26.540</td>
<td>0.149</td>
</tr>
<tr>
<td>24</td>
<td>0.024</td>
<td>26.822</td>
<td>0.177</td>
</tr>
</tbody>
</table>

- No coefficient is clearly significant but the Q statistic is significant for lags 4 to 14.
- Thus, we conclude that, with this test, we reject the serial independence of the residuals.
The mean of the residuals is not significantly different from zero and the variability of the residuals, except for some possible outliers, seems constant over time.
Is normally distributed? **No.** But, if we omit three atypical observations, we obtain the following residual statistics:

![Histogram with statistical summary](image)

Thus we conclude that outliers can influence the autocorrelation and normality test’s results.
Summarizing

By now we have learnt how to identify, estimate and check the accuracy of a model for a stationary time series process $X_t$.

But what can we do with these models?

Many things!

For instance

- Compute Impulse response functions
- Compute forecasts
- Use what we’ve learn to fit models to 1) nonstationary processes, 2) multivariate processes.
A few references

Brockwell and Davis, (1991), Chapters 8, 9
Hamilton, 1994, Chapter 5.
Wei, Chapters 6, 7.
Definition

Let \( \{y_t\} \) be a time series process. The IRF is the path \( y \) follows if it is kicked by a unitary shock at time \( t \) i.e., \( \varepsilon_t = 1 \), assuming all other posterior shocks are equal to zero, i.e., \( \varepsilon_{t+j} = 0 \), for all \( j > 1 \).

The IRF is a first step to start thinking about "causes" and "effects".

For univariate processes: How does \( y_t \) evolve over time when it is hit at time \( t \) by a particular shock, \( \varepsilon_t \). Useful to see how persistent a process is.

For multivariate processes: if we are modelling several variables, the IRF describes the evolution of one variable, say GDP, when is perturbed by another variable at time \( t \), say interest rates.
Computation

- The IRF can be computed as the derivative of $y_{t+h}$ w.r.t to the shock that has been perturbed, $\varepsilon_t$

$$IRF(h) = \frac{\partial y_{t+h}}{\partial \varepsilon_t}, \ h \geq 0.$$ 

If $y_t$ is stationary

$$y_t = \mu + \varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2\varepsilon_{t-2} + ...$$

and then,

$$IRF(h) = \frac{\partial y_{t+h}}{\partial \varepsilon_t} = \psi_h.$$ 

- Therefore, for stationary processes, the coefficients of the Wold representation are the values of the IRF.
Computation, II

For nonstationary processes the method above cannot be employed, since the Wold representation does not exist.

An alternative way of computing this function is as follows:

$$\text{IRF}(h) = E(y_{t+h} | \varepsilon_t = 1) - E(y_{t+h} | \varepsilon_t = 0)$$

i.e., the IRF at lag $h$ is the difference of two conditional expectations.

This is usually the simplest way of computing the IRF.
Examples

Compute the IRF(h) of the following processes.

1) $y_t = \varepsilon_t$
2) $y_t = \varepsilon_t + \psi \varepsilon_{t-1}$
3) $y_t = \phi y_{t-1} + \varepsilon_t$, with $|\phi| < 1$.
4) $y_t = y_{t-1} + \varepsilon_t$
Estimation and inference of the IRF

- **Estimation.**

- The IRF is a function of the parameters of the model.

- To estimate this function, you need first to estimate the model, then you use these values to compute the IRF, simply by replacing the unknown parameter by the estimate you obtained.
Inference

- for stationary processes, since the estimates of the parameters are asymptotically normal, estimates of the IRF are also asymptotically normal (by the delta method).

- However, the normal approximation works poorly for highly persistent processes.

- This is an important problem since economic data is usually very persistent.

- Solutions: Monte carlo confidence bands and bootstrap confidence bands.

- We’ll see these methods in more details when deriving IRFs for multivariate processes.
ARCH models

So far we have developed models for the **conditional mean** of a process \( \{y_t\} \).

Consider, for instance, a stationary AR(1) process

\[
y_t = c + \phi y_{t-1} + \varepsilon_t,
\]

Its mean is constant \((c/(1-\phi))\)

It’s conditional mean it’s not: \( E(y_t/y_{t-1}, y_{t-2}, \ldots) = c + \phi y_{t-1} \).

The conditional mean (=optimal forecast) changes over time, depending on the observed data.
Volatility clustering

Now we are going to look at models where the unconditional variance is constant (as in the previous case) but the conditional variance is not.

Many financial and macroeconomic variables have variances that seem to change over time.

Some periods are tranquil while others are very turbulent: volatility comes in clusters.

Features of typical behavior of financial data: volatility clustering and think tails
If there is volatility clustering there is some dependence in the variance that we can model.

Being able to obtain good models for the (time-changing) variance, it’s very useful since

The variance is a measure of uncertainty—for financial data is the risk of owning that asset.

The price of some financial derivatives (options) depends on the variance of the underlying asset.

Forecasting variances makes it possible to have accurate forecast intervals.
ARCH MODELS

Autoregressive conditionally heterokedastic models (ARCH) were developed by Engle (1982).

Consider a stationary process $y_t$

$$\Phi(L) y_t = C(L)\varepsilon_t, \text{ where } \varepsilon_t \sim (0, \sigma^2).$$

$\{\varepsilon_t\}$ is a white noise sequence (uncorrelated) but it does not have to be independent.

ARCH model: is a model for the variance of a white noise process, $\varepsilon_t$. 
\{\varepsilon_t\} is ARCH(p) if

\[ \varepsilon_t^2 = \alpha_0 + \alpha_1^2 \varepsilon_{t-1}^2 + \alpha_1^2 \varepsilon_{t-2}^2 + \ldots + \alpha_p^2 \varepsilon_{t-p}^2 + \omega_t, \]

where \{\omega_t\} is white noise and \( \alpha_i \) are coefficients to be estimated.

An ARCH(p) model is simply an AR(p) model for the squares of the innovations.

Since \( \varepsilon_t^2 \) is always positive, we need to impose some restrictions

- \( \alpha_0 > 0, \omega_t > -\alpha_0 \) and \( \alpha_j \geq 0. \)

Stationarity condition: roots of \( \alpha(L) \) larger than 1 (which implies: \( \sum_{i=1}^{p} \alpha_i < 1. \))
In ARCH models the (unconditional) variance of $\varepsilon_t$ is constant but the conditional one is not.

- **Unconditional variance of $\varepsilon_t$**

\[
E\left(\varepsilon_t^2\right) = \frac{\alpha_0}{1 - \alpha_1 - \ldots - \alpha_p} > 0.
\]

- **However, the conditional mean of $\varepsilon_t^2$ changes over time.**

\[
E\left(\varepsilon_t^2 | \psi_{t-1}\right) = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \ldots + \alpha_p \varepsilon_{t-p}^2,
\]

where $\psi_{t-1}$ represents all the information available up to $t-1$. 
Estimation and Inference in ARCH models

- ARCH models are estimated using Maximum Likelihood.

- As a result, estimators of ARCH coefficients are efficient and normally distributed.

- Inference on the estimated coefficients can be carried out using standard techniques.
Testing for ARCH effects

- Test whether the correlations of \( \{ \varepsilon_t^2 \} \) are equal to zero, as we did for ARMA processes.

- Alternatively, Engle (1982) proposed:

1) For a given process, \( y_t \) estimate a model and obtain the residuals, \( \hat{\varepsilon}_t \).

2) Fit an AR(p) process to \( \hat{\varepsilon}_t^2 \).

3) The \( R_{\varepsilon}^2 \) statistic from the previous regression times the sample size, \( T \), follows a \( \chi_p^2 \) under the null hypothesis that \( \varepsilon_t \) is i.i.d.

4) Thus, reject the null of i.i.d. (no ARCH effects) if \( TR_{\varepsilon}^2 \) is larger than the corresponding critical value.
Example 1: Price of IBM Stock

(A) IBM stock, percent month-on-month

(B) Squared returns

(C) ACF - Returns

(D) ACF - Squared returns
Example II: Danish Stock Market Index (KFX)
Empirical Example

- Daily data for the **Nasdaq index** 31/1-2000 till 26/2-2004, 1042 observations. We consider the log returns

\[ y_t = \log(\text{NASDAQ}_t) - \log(\text{NASDAQ}_{t-1}). \]

- We want to estimate a GARCH(1,1) model based on an AR(1), i.e.

\[
\begin{align*}
y_t & = \theta_0 + \theta_1 y_{t-1} + \epsilon_t \\
\sigma_t^2 & = \omega + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2
\end{align*}
\]

with normal innovations.
EQ( 1) Modelling DlogNasdaq by OLS (using Nasdaq.in7)

The estimation sample is: 3 to 1042

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std.Error</th>
<th>t-value</th>
<th>t-prob</th>
<th>Part.R^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DlogNasdaq_1</td>
<td>-0.00243025</td>
<td>0.03096</td>
<td>-0.0785</td>
<td>0.937</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.000628696</td>
<td>0.0007396</td>
<td>-0.850</td>
<td>0.395</td>
</tr>
</tbody>
</table>

ARCH coefficients:

<table>
<thead>
<tr>
<th>Lag</th>
<th>Coefficient</th>
<th>Std.Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.080823</td>
<td>0.03109</td>
</tr>
<tr>
<td>2</td>
<td>0.1732</td>
<td>0.03106</td>
</tr>
<tr>
<td>3</td>
<td>0.068651</td>
<td>0.03144</td>
</tr>
<tr>
<td>4</td>
<td>0.087759</td>
<td>0.03104</td>
</tr>
<tr>
<td>5</td>
<td>0.080797</td>
<td>0.03105</td>
</tr>
</tbody>
</table>

RSS = 0.00118881  sigma = 0.00107537

Testing for error ARCH from lags 1 to 5
ARCH 1-5 test:  F(5,1028) = 19.666 [0.0000]**
WOL(1) Modelling DlogNasdaq by restricted GARCH(1,1) (Nasdaq.in7)

The estimation sample is: 3 to 1042

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std.Error</th>
<th>robust-SE</th>
<th>t-value</th>
<th>t-prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>DlogNasdaq_1</td>
<td>Y</td>
<td>-0.0121906</td>
<td>0.03246</td>
<td>-0.393</td>
<td>0.695</td>
</tr>
<tr>
<td>Constant</td>
<td>X</td>
<td>0.000595005</td>
<td>0.0005788</td>
<td>0.0006317</td>
<td>0.942</td>
</tr>
<tr>
<td>alpha_0</td>
<td>H</td>
<td>3.30480e-006</td>
<td>2.113e-006</td>
<td>1.872e-006</td>
<td>1.77</td>
</tr>
<tr>
<td>alpha_1</td>
<td>H</td>
<td>0.0746032</td>
<td>0.01588</td>
<td>0.01610</td>
<td>4.63</td>
</tr>
<tr>
<td>beta_1</td>
<td>H</td>
<td>0.919901</td>
<td>0.01578</td>
<td>0.01423</td>
<td>64.7</td>
</tr>
</tbody>
</table>

log-likelihood 2523.8308  HMSE 2.16818
mean(h_t) 0.000576232 var(h_t) 1.90665e-007
no. of observations 1040 no. of parameters 5
AIC.T -5037.6616 AIC -4.84390539
mean(DlogNasdaq) -0.000627028 var(DlogNasdaq) 0.000567281
alpha(1)+beta(1) 0.994504 alpha_1+beta_1>=0, alpha(1)+beta(1)<1
(A) Residuals and 95% confidence band

2 conditional standard errors
Example 1: Price of IBM Stock

(A) IBM stock, percent month-on-month

(B) Squared returns

(C) ACF - Returns

(D) ACF - Squared returns