

# ARIMA Modelling and Forecasting

Economic time series often appear nonstationary, because of trends, seasonal patterns, cycles, etc.

However, the *differences* may appear stationary.

$$\Delta x_t = x_t - x_{t-1} \text{ (first difference)}$$

$$\Delta^2 x_t = \Delta x_t - \Delta x_{t-1} = x_t - 2x_{t-1} + x_{t-2} \text{ (second difference)}$$

etc.

In the case of a seasonal pattern, the *seasonal differences* could be stationary

$$\Delta_S x_t = x_t - x_{t-S}$$

where  $S = 4$  or  $S = 12$  say.

The ARIMA (autoregressive integrated moving average) model is

$$\lambda(L)\Delta_S^d x_t = \alpha + \theta(L)\varepsilon_t$$

where  $d = 0, 1$  or (occasionally)  $2$ .

Assuming no seasonal pattern ( $S = 1$ ), we speak of the ARIMA( $p, d, q$ ) model where  $p$ ,  $d$ , and  $q$  are integer values to be chosen.

# Box-Jenkins Methodology

Box and Jenkins (*Time Series Analysis: Forecasting and Control*, 1970) advocated a forecasting technique based on the ARIMA model.

The basic steps are as follows:

1. Choose smallest  $d$  so that the series ‘appears’ stationary.
2. Choose values for  $p$  and  $q$ , estimate ARMA model for series  $\Delta^d x_t$  by ML.
3. Check correlogram of residuals for evidence of autocorrelation.
4. Repeat 2 and 3 as necessary to choose the most parsimonious model that accounts for autocorrelation.

The procedure of choosing  $d$ ,  $p$  and  $q$  is called (by B-J) *model identification*.

- B-J advocate using known autocorrelation patterns of AR and MA processes to choose  $p$  and  $q$ . Thus, correlogram of AR ‘dies out’ exponentially as lag increases, while correlogram of MA ‘cuts off’ at  $q+1$  lags.
- If correlogram does not die out fast enough (or estimated AR root close to 1) may need to increase  $d$ .

# Testing for Uncorrelatedness

Recall the correlogram of a stationary process,

$$\rho_j = \frac{\text{Cov}(x_t, x_{t+j})}{\text{Var}(x_t)} \quad j = 1, 2, 3, \dots$$

Suppose  $x_t$  is an *independent* series.

Then,  $\rho_j = 0$  for all  $j > 0$ . The sample counterparts are

$$r_j = \frac{\sum_{t=1}^{T-j} (x_t - \bar{x})(x_{t+j} - \bar{x})}{\sum_{t=1}^T (x_t - \bar{x})^2} \neq 0$$

but these statistics should be ‘small’, on average.

It can be shown that s. e. ( $r_j$ )  $\approx T^{-1/2}$ , and that

$$\sqrt{T} r_j \xrightarrow{d} N(0, 1).$$

When the series is serially independent, it can also be shown that  $r_j$  and  $r_k$  for  $j \neq k$  are asymptotically independent.

Therefore, under the null hypothesis of independence,

$$Q = T \sum_{j=1}^m r_j^2 \xrightarrow{d} \chi^2(m).$$

This is the basis for the *Box-Pierce test* of serial independence.

- Strictly, a test of uncorrelatedness – no assurance of power against other alternatives.
- The problem is to choose  $m$  – large enough to detect all deviations from independence, but should not generally exceed  $T/3$ .
- This is also called the *Portmanteau Test*.

# Testing an ARIMA Specification

Let  $\hat{\varepsilon}_t$  denote the residuals from ARIMA( $p, d, q$ ) estimation. Then let

$$\hat{r}_j = \frac{\sum_{t=1}^{T-j+1} (\hat{\varepsilon}_t - \bar{\hat{\varepsilon}})(\hat{\varepsilon}_{t+j} - \bar{\hat{\varepsilon}})}{\sum_{t=1}^T (\hat{\varepsilon}_t - \bar{\hat{\varepsilon}})^2}$$

Let the null hypothesis be that the specification is 'correct'. That is, if

$$\lambda(L)\Delta_S^d x_t = \alpha + \theta(L)\varepsilon_t$$

where  $\lambda(L) = 1 - \dots - \lambda_p L^p$  and  $\theta(L) = 1 + \dots + \theta_q L^q$ , then  $\varepsilon_t \sim iid(0, \sigma^2)$ .

Box and Pierce (1970) show that in this case,

$$Q = T \sum_{j=1}^m \hat{r}_j^2 \stackrel{d}{\rightarrow} \chi^2(m - p - q)$$

Ljung and Box (1978) suggest a small-sample correction:

$$Q^* = T(T + 2) \sum_{j=1}^m \frac{r_j^2}{T - j}$$

- Since  $T \rightarrow \infty$  while  $m$  is fixed, note that this has the same asymptotic distribution as  $Q$ .
- Simulations suggest that the small-sample distribution is closer to the limit case.

# Testing for Nonlinear Model Features

Suppose model has uncorrelated disturbances. Does this mean that they are independent?

Nonlinear dependence (given uncorrelateness of levels) can be tested by the  $Q$  statistic computed for the *squares* of the series.

$$Q = T \sum_{j=1}^m \hat{r}_j^2$$

where

$$\hat{r}_j = \frac{\sum_{t=1}^{T-j+1} (\hat{\varepsilon}_t^2 - \overline{(\hat{\varepsilon}^2)}) (\hat{\varepsilon}_{t+j}^2 - \overline{(\hat{\varepsilon}^2)})}{\sum_{t=1}^T (\hat{\varepsilon}_t^2 - \overline{(\hat{\varepsilon}^2)})^2}.$$

This has  $\chi^2(m)$  distribution in the limit when the disturbances are independent. (see McLeod and Li (1983), Li and Mak (1994))

- Note: If the series is not autocorrelated, this test will tend to reject whether or not there is nonlinear dependence.
- The test should be performed only if the usual  $Q$  test does *not* reject.

# Consistent Model Selection Criteria

A *consistent* criterion is one that selects the correct model with probability 1 as  $T \rightarrow \infty$ , when this is one of the alternatives examined.

Suppose there are  $M$  models under consideration, indexed by  $k = 1, \dots, M$ .

Let the maximized (quasi-) log likelihood be

$$\hat{L}_T = -\frac{T}{2} \log\left(T^{-1} \sum_t \hat{\varepsilon}_t^2\right).$$

Let  $\hat{L}_{kT}$  = maximized log-likelihood for model  $k$ , and  $p_k$  = number of fitted parameters in model  $k$ .

- If we simply selected model  $k$  giving the largest  $\hat{L}_{kT}$  we should tend to over-fit, by including too many parameters.
- Some penalty for allowing extra parameters is necessary.

Instead, choose  $k$  to maximize

$$A_{kT} = \hat{L}_{kT} - p_k r_T$$

where  $r_T \rightarrow \infty$  as  $T \rightarrow \infty$ , but  $r_T/T \rightarrow 0$ .

Let

$$\bar{L}_k = \text{plim}_{T \rightarrow \infty} T^{-1} \hat{L}_{kT} < \infty.$$

If the true model (case  $k = m$ ) is among those considered, should have  $\bar{L}_k \leq \bar{L}_m$ , all  $k$ .

A consistent criterion should choose the case  $k = m$  with probability  $\rightarrow 1$  as  $T \rightarrow \infty$ .

**Cases:**

1. If the  $k$ th model incorrect in the sense  $\bar{L}_k < \bar{L}_m$  then

$$\frac{A_{mT}}{A_{kT}} \xrightarrow{pr} \frac{\bar{L}_m}{\bar{L}_k} > 1.$$

2. Suppose  $\bar{L}_k = \bar{L}_m$  but  $p_k > p_m$  ( $k$ th model correct but over-parameterized).  
Since  $\hat{L}_{kT} = O_p(T)$  and  $|\hat{L}_{kT} - \hat{L}_{mT}| = O_p(1)$ ,

$$\begin{aligned} \frac{A_{mT}}{A_{kT}} &= \frac{\hat{L}_{mT} - p_m r_T}{\hat{L}_{kT} - p_k r_T} = 1 + r_T \left( \frac{p_k - p_m}{\hat{L}_{kT} - p_k r_T} \right) + O_p(T^{-1}) \\ &> 1 \text{ with probability } \rightarrow 1 \text{ as } T \rightarrow \infty. \end{aligned}$$

Note: the second r.h.s. term is  $O_p(r_T/T)$  and dominates the last one as  $T \rightarrow \infty$ .

In each case, the false model is rejected w.p.  $\rightarrow 1$  as  $T \rightarrow \infty$ .  $P(\text{model } m \text{ selected})$  can be made as near 1 as desired by taking  $T$  large enough.

## Three popular criteria:

1. *Akaike criterion*: “ maximize  $\hat{L}_T - k$  ”
  2. *Schwarz criterion*: “ maximize  $\hat{L}_T - \frac{1}{2}k \log T$  ”
  3. *Hannan-Quinn criterion*: “ maximize  $\hat{L}_T - k \log \log T$  ”
- When the number of parameters in the true model is finite,
    - The Akaike criterion is not consistent.
    - The Schwarz and H-Q are consistent.
    - Schwarz criterion favours the most parsimonious models.
  - **Caution**: some books/software packages reverse the signs, so the rule is to choose the model that *minimises* the criterion.
    - There is no accepted convention in this respect - read the small print!

# Some Guidelines for Model Choice

1. Decide on  $d$  – either by ‘eye’ (B-J), or by a unit root test.
2. Decide on the maximum values of  $p$  and  $q$  (and  $p + q$ ) compatible with the available sample, and feasible estimation. (Too many parameters may take too long to compute, and cause numerical instability.)
3. Use a consistent selection criterion to choose the ‘best’ ARIMA specification out of the (as many as)  $(p + 1)(q + 1)$  possible cases.
4. Remember that significance tests on the AR and MA coefficients ( $t$ -ratios) are valid *only* in context of a ‘correct’ (i.e. possibly over parameterized) model. Examine with caution!
5. Run the  $Q$  or  $Q^*$  test on the residuals of the model. – as a guideline, choose  $m < \min\{T/3, 20\}$ .
6. Final selection: best model on the criterion, subject to side condition of not rejecting on the portmanteau test.

In my experience, it’s rare to have a conflict at the final stage unless there are model features the ARIMA cannot describe, e.g., seasonality.

# Forecasting: General Principles

Given information represented by sigma field  $\mathcal{F}_t$ , we wish to forecast a variable  $Y$  (e.g.  $Y = x_{t+m}$ ).

If  $Z_t$  is a  $\mathcal{F}_t$ -measurable forecast function, let

$$E(Y - Z_t)^2$$

the *mean squared error of forecast* (MSE) be the preferred loss function. The aim is to choose  $Z_t$  to make the loss as small as possible.

- Note, other loss functions are possible - this one is symmetric, weighting positive and negative errors equally.

As we know, setting  $Z_t$  equal to  $E(Y|\mathcal{F}_t)$  gives the *minimum MSE* forecast.

Therefore, consider the ARMA model:

$$x_{t+m} = \alpha + \lambda_1 x_{t+m-1} + \cdots + \lambda_p x_{t+m-p} + \varepsilon_{t+m} + \theta_1 \varepsilon_{t+m-1} + \cdots + \theta_q \varepsilon_{t+m-q}$$

Note that

$$E(x_{t+m}|\mathcal{F}_t) = \alpha + \lambda_1 E(x_{t+m-1}|\mathcal{F}_t) + \cdots + \lambda_p E(x_{t+m-p}|\mathcal{F}_t) + \theta_m \varepsilon_t + \cdots + \theta_q \varepsilon_{t+m-q}$$

- The last terms vanish if  $m > q$ .
- $E(x_{t+m-j}|\mathcal{F}_t) = x_{t+m-j}$  if  $j > m$ .

# Forecasting using the ARIMA(p,1,q)

Observations  $1, \dots, T$  are used to identify and estimate the model

$$\Delta x_t = \hat{\alpha} + \hat{\lambda}_1 \Delta x_{t-1} + \dots + \hat{\lambda}_p \Delta x_{t-p} + \hat{\varepsilon}_t + \hat{\theta}_1 \hat{\varepsilon}_{t-1} + \dots + \hat{\theta}_q \hat{\varepsilon}_{t-q}$$

– ‘hats’ denote estimates and residuals.

- To forecast  $x_{T+1}$ , compute

$$\Delta \hat{x}_{T+1} = \hat{\alpha} + \hat{\lambda}_1 \Delta x_T + \dots + \hat{\lambda}_p \Delta x_{T+1-p} + \hat{\theta}_1 \hat{\varepsilon}_T + \dots + \hat{\theta}_q \hat{\varepsilon}_{T+1-q}$$

and so  $\hat{x}_{T+1} = x_T + \Delta \hat{x}_{T+1}$  .

To forecast  $x_{T+2}$  , compute recursively,

$$\begin{aligned} \Delta \hat{x}_{T+2} &= \hat{\alpha} + \hat{\lambda}_1 \Delta \hat{x}_{T+1} + \hat{\lambda}_2 \Delta x_T + \dots + \hat{\lambda}_p \Delta x_{T+2-p} \\ &\quad + \hat{\theta}_2 \hat{\varepsilon}_T + \dots + \hat{\theta}_q \hat{\varepsilon}_{T+2-q} \end{aligned}$$

and so  $\hat{x}_{T+2} = \hat{x}_{T+1} + \Delta \hat{x}_{T+2}$  . etc. etc.

- Put post-sample residuals to 0 and post-sample observations to forecasts.
- Note that when  $d = 1$ , the intercept is the coefficient of linear trend.

# Forecast Confidence Intervals

Ignoring errors in parameter estimates, the difference between the  $m$ -period-ahead forecast and the 'out-turn' is found by putting the unknown future shocks to zero.

**Case 1:**  $d = 0$ .

$$\begin{aligned}x_{t+m} &= \frac{\alpha}{\lambda(1)} + \frac{\theta(L)}{\lambda(L)} \varepsilon_{t+m} \\ &= a + \varepsilon_{t+m} + b_1 \varepsilon_{t+m-1} + b_2 \varepsilon_{t+m-2} + \cdots + b_m \varepsilon_t + \cdots\end{aligned}$$

(say).

The forecast of  $x_{t+m}$  (assuming parameters known) is

$$\hat{x}_{t+m|t} = a + b_m \varepsilon_t + b_{m+1} \varepsilon_{t-1} + b_{m+2} \varepsilon_{t-2} + \cdots$$

Hence, the forecast error is

$$f_{t+m|t} = x_{t+m} - \hat{x}_{t+m|t} = \varepsilon_{t+m} + b_1 \varepsilon_{t+m-1} + \cdots + b_{m-1} \varepsilon_{t+1}.$$

The forecast error variance, and confidence intervals, can therefore be calculated from the MA( $\infty$ ) form of the model.

Assuming that  $\varepsilon_t \sim iid(0, \sigma^2)$ , note that

- $E(f_{t+m|t}) = 0$  (unbiased forecasts)
- $\text{Var}(f_{t+m|t}) = \sigma^2(1 + b_1^2 + \cdots + b_{m-1}^2)$

**Case 2:**  $d = 1$ .

$$x_{t+m} = x_t + \Delta x_{t+1} + \cdots + \Delta x_{t+m}.$$

and hence

$$x_{t+m} - \hat{x}_{t+m|t} = f_{t+1} + \cdots + f_{t+m} = \varepsilon_{t+m} + (1 + b_1)\varepsilon_{t+m-1} + \cdots + (1 + b_1 + \cdots + b_{m-1})\varepsilon_{t+1}$$

and

$$\text{Var}(f_{t+m|t}) = \sigma^2(1 + c_1^2 + \cdots + c_{m-1}^2)$$

where

$$c_j = 1 + \sum_{i=1}^j b_i$$

Notice the difference between the two cases.

- If  $d = 0$ , the forecast error variance tends as  $m \rightarrow \infty$  to

$$\text{Var}(f_{t+m|t}) \rightarrow \sigma^2 \sum_{j=0}^{\infty} b_j^2 < \infty.$$

- If  $d = 1$ , then

$$\text{Var}(f_{t+m|t}) = O(m).$$

# State Space Modelling

A formalized representation of model dynamics.

Let  $\mathbf{x}_t$  ( $m \times 1$ ) represent a vector of observed variables, and  $\boldsymbol{\xi}_t$  ( $r \times 1$ ) an unobserved *state vector*.

The evolution of  $\mathbf{x}_t$  is described by two equations:

## State equation

$$\boldsymbol{\xi}_t = \mathbf{F}\boldsymbol{\xi}_{t-1} + \mathbf{v}_t, \quad E(\mathbf{v}_t) = \mathbf{0}, E(\mathbf{v}_t\mathbf{v}_t') = \mathbf{Q}$$

## Measurement equation

$$\mathbf{x}_t = \boldsymbol{\mu} + \mathbf{H}\boldsymbol{\xi}_t + \mathbf{w}_t \quad E(\mathbf{w}_t) = \mathbf{0}, E(\mathbf{w}_t\mathbf{w}_t') = \mathbf{R}$$

Let it also be assumed that  $E(\mathbf{u}_t\mathbf{w}_t') = \mathbf{0}$ .

- Optionally, the matrices  $\mathbf{F}$ ,  $\mathbf{H}$ ,  $\mathbf{Q}$  and  $\mathbf{R}$  can be time-dependent, and receive  $t$  subscripts.
- Optionally, the measurement equation can include explanatory exogenous variables. as well as an intercept.
- The motivation for the state-space form is that almost any linear time series model can be cast into this form.

Example: the ARMA( $p, q$ ).

Let  $m = 1$ ,  $r = \max(p, q)$ , and consider the ARMA( $r, r$ ) case: - extend the AR or MA orders as required, by specifying zero coefficients

Put

$$\mathbf{F} = \begin{bmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_{r-1} & \lambda_r \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & \vdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \quad \mathbf{v}_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

(basically, an application of the companion form) and

$$\mathbf{H} = \begin{bmatrix} 1 & \theta_1 & \cdots & \theta_{r-1} \end{bmatrix}, \quad R = 0, \text{ hence } w_t = 0$$

Note that this system resolves as

$$\mathbf{x}_t - \boldsymbol{\mu} = \theta(L)\xi_{1t} = \frac{\theta(L)}{\lambda(L)}\varepsilon_t.$$

# The Kalman Filter

This is a computer algorithm for solving a state space model recursively.

- It can be used for model simulation and forecasting.
- It provides a convenient vehicle for evaluating the likelihood function in dynamic models, generating the sequence of residuals from the data.
- It is popular in applied work, since numerous software packages are available to compute it, having a wide range of applications.

On the other hand:

- There is nothing the Kalman filter can do that cannot be done by more direct means, coded for the problem at hand.
- It cannot handle nonlinear time series models, or long memory models - it is ultimately limited in scope.
- The Kalman filter is just a tool for doing certain recursive calculations, especially in ARIMA-type models. It's not a modelling paradigm of itself.
- I'm not sure that it deserves the prominence it usually receives in time series syllabuses. (A personal view!)