7 Fitting ARIMA models

7.1 The Box-Jenkins procedure

A general ARIMA(p, d, q) model is $\phi(B)\nabla(B)^d X = \theta(B)\epsilon$, where $\nabla(B) = I - B$.

The **Box-Jenkins** procedure is concerned with fitting an ARIMA model to data. It has three parts: **identification**, **estimation**, and **verification**.

7.2 Identification

The data may require pre-processing to make it stationary. To achieve stationarity we may do any of the following.

- Look at it.
- Re-scale it (for instance, by a logarithmic or exponential transform.)
- Remove deterministic components.
- Difference it. That is, take $\nabla(B)^d X$ until stationary. In practice $d = 1, 2$ should suffice.

We recognise stationarity by the observation that the autocorrelations decay to zero exponentially fast.

Once the series is stationary, we can try to fit an ARMA(p, q) model. We consider the correlogram $r_k = \hat{\gamma}_k/\hat{\gamma}_0$ and the partial autocorrelations $\hat{\phi}_{k,k}$. We have already made the following observations.

- An MA(q) process has negligible ACF after the qth term.
- An AR(p) process has negligible PACF after the pth term.

As we have noted, very approximately, both the sample ACF and PACF have standard deviation of around $1/\sqrt{T}$, where T is the length of the series. A rule of thumb is that ACF and PACF values are negligible when they lie between $\pm 2/\sqrt{T}$. An ARMA(p, q) process has kth order sample ACF and PACF decaying geometrically for $k > \max(p, q)$.

7.3 Estimation

AR processes

To fit a pure AR(p), i.e., $X_t = \sum_{r=1}^p \phi_r X_{t-r} + \epsilon_t$ we can use the **Yule-Walker equations** $\gamma_k = \sum_{r=1}^p \phi_r \gamma_{|k-r|}$. We fit ϕ by solving $\hat{\gamma}_k = \sum_1^p \phi_r \hat{\gamma}_{|k-r|}$, $k = 1, \dots, p$. These can be solved by a **Levinson-Durbin recursion**, (similar to that used to solve for partial autocorrelations in Section 2.6). This recursion also gives the estimated

residual variance $\hat{\sigma}_p^2$, and helps in choice of p through the approximate log likelihood $-2\log L \simeq T\log(\hat{\sigma}_p^2)$.

Another popular way to choose p is by minimizing **Akaike's AIC** (*an information criterion*), defined as AIC $= -2\log L + 2k$, where k is the number of parameters estimated, (in the above case p). As motivation, suppose that in a general modelling context we attempt to fit a model with parameterised likelihood function $f(X \mid \theta)$, $\theta \in \Theta$, and this includes the true model for some $\theta_0 \in \Theta$. Let $X = (X_1, \dots, X_n)$ be a vector of n independent samples and let $\hat{\theta}(X)$ be the maximum likelihood estimator of θ. Suppose Y is a further independent sample. Then

$$-2n\mathbb{E}_Y\mathbb{E}_X \log f\left(Y \mid \hat{\theta}(X)\right) = -2\mathbb{E}_X \log f\left(X \mid \hat{\theta}(X)\right) + 2k + O\left(1/\sqrt{n}\right) ,$$

where $k = |\Theta|$. The left hand side is $2n$ times the conditional entropy of Y given $\hat{\theta}(X)$, i.e., the average number of bits required to specify Y given $\hat{\theta}(X)$. The right hand side is approximately the AIC and this is to be minimized over a set of models, say $(f_1, \Theta_1), \dots, (f_m, \Theta_m)$.

ARMA processes

Generally, we use the maximum likelihood estimators, or at least squares numerical approximations to the MLEs. The essential idea is prediction error decomposition. We can factorise the joint density of $(X_1, \dots, X_T)$ as

$$f(X_1, \dots, X_T) = f(X_1)\prod_{t=2}^{T} f(X_t \mid X_1, \dots, X_{t-1}) .$$

Suppose the conditional distribution of X_t given $(X_1, \dots, X_{t-1})$ is normal with mean $\hat{X}_t$ and variance P_{t-1}, and suppose also that X_1 is normal $N(\hat{X}_1, P_0)$. Here $\hat{X}_t$ and P_{t-1} are functions of the unknown parameters $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$ and the data.

The log likelihood is

$$-2\log L = -2\log f = \sum_{t=1}^{T}\left[\log(2\pi) + \log P_{t-1} + \frac{(X_t - \hat{X}_t)^2}{P_{t-1}}\right] .$$

We can minimize this with respect to $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$ to fit ARMA(p, q).

Additionally, the second derivative matrix of $-\log L$ (at the MLE) is the observed information matrix, whose inverse is an approximation to the variance-covariance matrix of the estimators.

In practice, fitting ARMA(p, q) the log likelihood $(-2\log L)$ is modified to sum only over the range $\{m+1, \dots, T\}$, where m is small.

Example 7.1
For AR(p), take $m = p$ so $\hat{X}_t = \sum_{r=1}^p \phi_r X_{t-r}$, $t \geq m+1$, $P_{t-1} = \sigma_\epsilon^2$.

Note. When using this approximation to compare models with different numbers of parameters we should always use the same m.

Again we might choose p and q by minimizing the AIC of $-2\log L + 2k$, where $k = p + q$ is the total number of parameters in the model.

7.4 Verification

The third stage in the Box-Jenkins algorithm is to check whether the model fits the data. There are several tools we may use.

- Overfitting. Add extra parameters to the model and use likelihood ratio test or t-test to check that they are not significant.
- Residuals analysis. Calculate the residuals from the model and plot them. The autocorrelation functions, ACFs, PACFs, spectral densities, estimates, etc., and confirm that they are consistent with white noise.

7.5 Tests for white noise

Tests for white noise include the following.

(a) The turning point test (explained in Lecture 1) compares the number of peaks and troughs to the number that would be expected for a white noise series.

(b) The **Box–Pierce test** is based on the statistic

$$Q_m = T\sum_{k=1}^{m} r_k^2\,,$$

where r_k is the kth sample autocorrelation coefficient of the residual series, and $p + q < m \ll T$. It is called a 'portmanteau test', because it is based on the all-inclusive statistic. If the model is correct then $Q_m \sim \chi^2_{m-p-q}$ approximately.

In fact, r_k has variance $(T-k)/(T(T+2))$, and a somewhat more powerful test uses the Ljung-Box statistic quoted in Section 2.7,

$$Q'_m = T(T+2)\sum_{k=1}^{m}(T-k)^{-1}r_k^2\,,$$

where again, $Q'_m \sim \chi^2_{m-p-q}$ approximately.

(c) Another test for white noise can be constructed from the periodogram. Recall that $I(\omega_j) \sim (\sigma^2/\pi)\chi^2_2/2$ and that $I(\omega_1),\dots,I(\omega_m)$ are mutually independent.

Define $C_j = \sum_{k=1}^{j} I(\omega_k)$ and $U_j = C_j/C_m$. Recall that χ^2_2 is the same as the exponential distribution and that if $Y_1,\dots,Y_m$ are i.i.d. exponential random variables, then $(Y_1+\dots+Y_j)/(Y_1+\dots+Y_m)$, $j = 1,\dots,m-1$, have the distribution of an ordered sample of $m-1$ uniform random variables drawn from $[0,1]$. Hence under the hypothesis that $\{X_t\}$ is Gaussian white noise U_j, $j = 1,\dots,m-1$ have the distribution of an ordered sample of $m-1$ uniform random variables on $[0,1]$. The standard test for this is the Kolomogorov-Smirnov test, which uses as a test statistic, D, defined as the maximum difference between the theoretical distribution function for $U[0,1]$, $F(u) = u$, and the empirical distribution $\hat{F}(u) = \{\#(U_j \le u)\}/(m-1)$. Percentage points for D can be found in tables.

7.6 Forecasting with ARMA models

Recall that $\phi(B)X = \theta(B)\epsilon$, so the power series coefficients of $C(z) = \theta(z)/\phi(z) = \sum_{r=0}^{\infty} c_r z^r$ give an expression for X_t as $X_t = \sum_{r=0}^{\infty} c_r\epsilon_{t-r}$.

But also, $\epsilon = D(B)X$, where $D(z) = \phi(z)/\theta(z) = \sum_{r=0}^{\infty} d_r z^r$ — as long as the zeros of θ lie strictly outside the unit circle and thus $\epsilon_t = \sum_{r=0}^{\infty} d_r X_{t-r}$.

The advantage of the representation above is that given $(\dots,X_{t-1},X_t)$ we can calculate values for $(\dots,\epsilon_{t-1},\epsilon_t)$ and so can forecast X_{t+1}.

In general, if we want to forecast X_{T+k} from $(\dots,X_{T-1},X_T)$ we use

$$\hat{X}_{T,k} = \sum_{r=k}^{\infty} c_r\epsilon_{T+k-r} = \sum_{r=0}^{\infty} c_{k+r}\epsilon_{T-r}\,,$$

which has the least mean squared error over all linear combinations of $(\dots,\epsilon_{T-1},\epsilon_T)$. In fact,

$$\mathbb{E}\left((\hat{X}_{T,k} - X_{T+k})^2\right) = \sigma_\epsilon^2\sum_{r=0}^{k-1} c_r^2\,.$$

In practice, there is an alternative recursive approach. Define

$$\hat{X}_{T,k} = \begin{cases} X_{T+k}, & -(T-1) \le k \le 0\,, \\ \text{optimal predictor of } T_{T+k} \text{ given } X_1,\dots,X_T, & 1 \le k\,. \end{cases}$$

We have the recursive relation

$$\hat{X}_{T,k} = \sum_{r=1}^{p}\phi_r\hat{X}_{T,k-r} + \hat{\epsilon}_{T+k} + \sum_{s=1}^{q}\theta_s\hat{\epsilon}_{T+k-s}$$

For $k = -(T-1), -(T-2),\dots,0$ this gives estimates of $\hat{\epsilon}_t$ for $t = 1,\dots,T$.

For $k > 0$, this gives a forecast $\hat{X}_{T,k}$ for X_{T+k}. We take $\hat{\epsilon}_t = 0$ for $t > T$.

But this needs to be started off. We need to know $(X_t,\ t \le 0)$ and ϵ_t, $t \le 0$. There are two standard approaches.

1. Conditional approach: take $X_t = \epsilon_t = 0$, $t \le 0$.
2. Backcasting: we forecast the series in the reverse direction to determine estimators of $X_0, X_{-1},\dots$ and $\epsilon_0,\epsilon_{-1},\dots$.