Introduction to Time Series Analysis. Lecture 15.

Last lecture: Maximum likelihood estimation

1. Diagnostics
2. Model selection
3. Integrated ARMA models

Building ARMA models

1. Plot the time series.

   Look for trends, seasonal components, step changes, outliers.

2. Nonlinearly transform data, if necessary

3. Identify preliminary values of p, and q.

4. Estimate parameters.

5. Use diagnostics to confirm residuals are white/iid/normal.

6. Model selection: Choose p and q.

Diagnostics

How do we check that a model fits well?

The residuals (innovations, $x_t - x_t^{t-1}$) should be white.
Consider the *standardized innovations*,

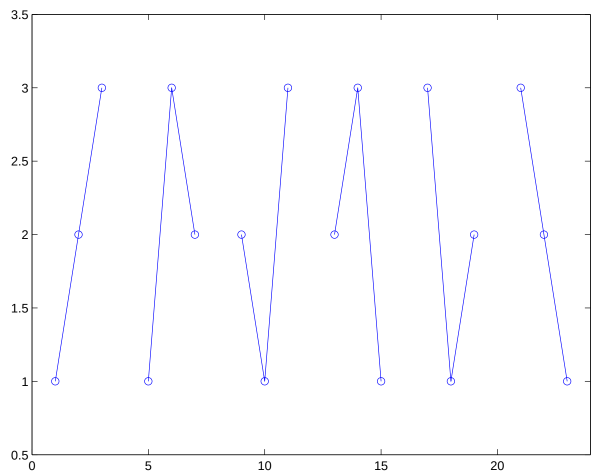
$$e_t = \frac{x_t - \hat{x}_t^{t-1}}{\sqrt{\hat{P}_t^{t-1}}}.$$

This should behave like a mean-zero, unit variance, iid sequence.

- Check a time plot
- Turning point test
- Difference sign test
- Rank test
- Q-Q plot, histogram, to assess normality

Testing i.i.d.: Turning point test

$\{X_t\}$ i.i.d. implies that X_t, X_{t+1} and X_{t+2} are equally likely to occur in any of six possible orders:



(provided X_t, X_{t+1}, X_{t+2} are distinct).

Four of the six are **turning points**.

Testing i.i.d.: Turning point test

Define $T = |\{t : X_t, X_{t+1}, X_{t+2} \text{ is a turning point}\}|$.

$\mathrm{E}T = (n-2)2/3$.

Can show $T \sim AN(2n/3, 8n/45)$.

Reject (at 5% level) the hypothesis that the series is i.i.d. if

$$\left| T - \frac{2n}{3} \right| > 1.96 \sqrt{\frac{8n}{45}}.$$

Tests for positive/negative correlations at lag 1.

Testing i.i.d.: Difference-sign test

$$S = |\{i : X_i > X_{i-1}\}| = |\{i : (\nabla X)_i > 0\}|.$$

$$\mathrm{E}S = \frac{n-1}{2}.$$

Can show $S \sim AN(n/2, n/12)$.

Reject (at 5% level) the hypothesis that the series is i.i.d. if

$$\left| S - \frac{n}{2} \right| > 1.96\sqrt{\frac{n}{12}}.$$

Tests for trend.

(But a periodic sequence can pass this test...)

Testing i.i.d.: Rank test

$$N = |\{(i,j) : X_i > X_j \text{ and } i > j\}|.$$

$$\mathrm{E}N = \frac{n(n-1)}{4}.$$

Can show $N \sim AN(n^2/4, n^3/36)$.

Reject (at 5% level) the hypothesis that the series is i.i.d. if

$$\left| N - \frac{n^2}{4} \right| > 1.96\sqrt{\frac{n^3}{36}}.$$

Tests for linear trend.

Testing if an i.i.d. sequence is Gaussian: qq plot

Plot the pairs $(m_1, X_{(1)}), \dots, (m_n, X_{(n)})$,
where $m_j = \mathrm{E}Z_{(j)}$,
$Z_{(1)} < \cdots < Z_{(n)}$ are order statistics from $N(0,1)$ sample of size n, and
$X_{(1)} < \cdots < X_{(n)}$ are order statistics of the series $X_1, \dots, X_n$.

Idea: If $X_i \sim N(\mu, \sigma^2)$, then

$$\mathrm{E}X_{(j)} = \mu + \sigma m_j,$$

so $(m_j, X_{(j)})$ should be *linear*.

There are tests based on how far correlation of $(m_j, X_{(j)})$ is from 1.

Introduction to Time Series Analysis. Lecture 15.

1. Diagnostics
2. Model selection
3. Integrated ARMA models

Model Selection

We have used the data x to estimate parameters of several models. They all fit well (the innovations are white). We need to choose a single model to retain for forecasting. How do we do it?

If we had access to independent data y from the same process, we could compare the likelihood on the new data, $L_y(\hat{\phi}, \hat{\theta}, \hat{\sigma}_w^2)$.

We could obtain y by leaving out some of the data from our model-building, and reserving it for model selection. This is called *cross-validation*. It suffers from the drawback that we are not using all of the data for parameter estimation.

Model Selection: AIC

We can approximate the likelihood defined using independent data: asymptotically

$$-\ln L_y(\hat{\phi}, \hat{\theta}, \hat{\sigma}_w^2) \approx -\ln L_x(\hat{\phi}, \hat{\theta}, \hat{\sigma}_w^2) + \frac{(p+q+1)n}{n-p-q-2}.$$

AIC_c: corrected Akaike information criterion.

Notice that:

• More parameters incur a bigger penalty.

• Minimizing the criterion over all values of $p, q, \hat{\phi}, \hat{\theta}, \hat{\sigma}_w^2$ corresponds to choosing the optimal $\hat{\phi}, \hat{\theta}, \hat{\sigma}_w^2$ for each p, q, and then comparing the penalized likelihoods.

There are also other criteria: BIC.

Introduction to Time Series Analysis. Lecture 15.

1. Diagnostics
2. Model selection
3. Integrated ARMA models

Integrated ARMA Models: ARIMA(p,d,q)

For $p, d, q \geq 0$, we say that a time series $\{X_t\}$ is an **ARIMA (p,d,q) process** if $Y_t = \nabla^d X_t = (1 - B)^d X_t$ is ARMA(p,q). We can write

$$\phi(B)(1 - B)^d X_t = \theta(B) W_t.$$

Recall the random walk: $X_t = X_{t-1} + W_t$.
X_t is not stationary, but $Y_t = (1 - B)X_t = W_t$ is a stationary process.
In this case, it is white, so $\{X_t\}$ is an ARIMA(0,1,0).

Also, if X_t contains a trend component plus a stationary process, its first difference is stationary.

ARIMA models example

Suppose $\{X_t\}$ is an ARIMA(0,1,1): $X_t = X_{t-1} + W_t - \theta_1 W_{t-1}$.
If $|\theta_1| < 1$, we can show

$$X_t = \sum_{j=1}^{\infty} (1-\theta_1)\theta_1^{j-1} X_{t-j} + W_t,$$

and so
$$\tilde{X}_{n+1} = \sum_{j=1}^{\infty} (1-\theta_1)\theta_1^{j-1} X_{n+1-j}$$
$$= (1-\theta_1)X_n + \sum_{j=2}^{\infty} (1-\theta_1)\theta_1^{j-1} X_{n+1-j}$$
$$= (1-\theta_1)X_n + \theta_1 \tilde{X}_n.$$

Exponentially weighted moving average.