10/20/16

Motivation:

- Analyze, model and **predict** data that is observed in a sequential order
- Data is no longer independent, and so standard inferential procedures don't work anymore/are invalid
- Decompose dependent data into independent components
- We care less about finding relationships between a response variable and covariates. We typically want to forecast a response using just its past values.

Regression_Example.R and *ConsIndex.txt*

Definitions

An **observed time series** $\{x_t : \ t \in T\}$ is a collection of observations of a variable of interest over time.

A **time series** is a stochastic process indexed by time. Specifically, we have a sequence of random variables $\{X_t : \ t \in T\}$, where T is an index of time points.

- if T is a discrete set, i.e. $T = \{1, 2, 3, \ldots\}$, then $\{X_t\}$ is a **discrete time series**.
- if T is a continuous interval, i.e. $T = \{t > 0\}$, then $\{X_t\}$ is a **continuous time series**.

A **time series model** is the specification of the joint distribution of the random variables $\{X_t : t \in N\}$: $P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$ for $-\infty < x_1, x_2, \ldots, x_n < \infty$ and $n \in N$. But, in general, we can't hope to estimate all of the parameters in such a model with the data we've observed.

But, most of the information about a distribution is contained in the first two moments:

- First Moments: $E[X_t]$, $t = 1, 2, \ldots$ -> *means*
- Second Moments: $E[X_t X_{t+h}]$, $t = 1, 2, \ldots$ and $h = 0, 1, 2, \ldots$ -> *variances/covariances*

Main take-away: we don't need the whole joint distribution. Our modeling will be based on **second-order properties**.

$\{x_t\}$ —observed from—> $\{X_t\}$

Zero Mean Models

IID Noise

If $\{X_1, X_2, \ldots, X_k\}$ are iid random variables with $E[X_t] = 0, \ t = 1, 2, \ldots, k$, then $P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) = \textit{independence} = \prod_{t=1}^{k} P(X_t \leq x_t) = \textit{identically distributed} = \prod_{t=1}^{k} F(x_t)$. In this special case, the joint distribution is defined by one marginal distribution with zero mean.

White noise

A white noise process is a sequence of **uncorrelated** (not necessarily independent!) random variables $\{X_t\}$ each with mean 0, and finite variance σ^2.

We denote this by $\{X_t\} \sim WN(0, \sigma^2)$.

- $E[X_t] = 0$
- $Var(X_t) = \sigma^2$ finite

- $Cov(X_i, X_j) = 0$ for $i \neq j$

(note: IID noise is a subset of White noise)

Classical Time Series Decomposition

$X_t = m_t + s_t + \epsilon_t$

- m_t: trend term (average change in X_t over time)
- s_t: seasonal term (regular periodic fluctuations)
- ϵ_t: error (unexplained variation in X_t's)

Lecture 1.pptx

Example

Consider average seasonal temperature over many years where we wish to fit a model of the form $X_t = m_t + s_t + \epsilon_t$.

Here, we assume m_t is a polynomial in t, and s_t can be represented with indicator/dummy variables:

- $W_1 = \begin{cases} 1 & \text{if spring} \\ 0 & \text{otherwise} \end{cases}$
- $W_2 = \begin{cases} 1 & \text{if fall} \\ 0 & \text{otherwise} \end{cases}$
- $W_3 = \begin{cases} 1 & \text{if winter} \\ 0 & \text{otherwise} \end{cases}$

$X_t = \sum_{i=0}^{p} \beta_i t^i + \sum_{j=1}^{3} \alpha_j W_j + \epsilon_t$, $\epsilon_t \sim N(0, \sigma^2)$ (iid)

- We typically estimate α's and β's using OLD, which implies that we are making OLS assumptions (which still may not be valid).
- If the assumptions are invalid, then we use the **Box-Jenkins** class of models (i.e. AR, MA, ARMA, SARIMA)

AirPassengers Analysis.R

10/25/16

Recap

- Time series –> $\{X_t : \ t \in N\}$ <– a time series model puts constraints on the first and second moments of these random variables.
- Observed time series –> $\{x_t : \ t \in N\}$.

Stationarity

Strict stationarity

A time series $\{X_t\}$ is said to be **strictly stationary** if the joint distribution of $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ is the same as that of $X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_n+h}$ for all $n, h, t_1, t_2, \ldots, t_n \in N$.

i.e., a strictly stationary time series preserves **all** statistical properties over time shift.

Problems:

- We often can't specify the joint distribution of these random variables and so this assumption is usually impossible to check.
- Also, this assumption tends to be too strict and is not often met.

This motivates the need for a weaker version of stationarity.

But first…

Let $\{X_t\}$ be a time series.

- The **mean function** of $\{X_t\}$ is $\mu_X(t) = E(X_t)$,
- The **covariance function** of $\{X_t\}$ is $\gamma_X(r, s) = Cov(X_r, X_s) = E(X_r X_s) - \mu_X(r)\mu_X(s)$.

Weak stationarity

A time series $\{X_t\}$ is **weakly stationary** if $E(X_t^2) < \infty$ and:

- $\{X_t\}$ is $\mu_X(t) = E(X_t)$ is independent of t,
- $\gamma_X(t, t+h) = Cov(X_t, X_{t+h})$ is independent of t for all h.
  - covariance depends on h but not t

Remarks:

- Strict stationarity ==> weak stationarity
- From now on, "stationarity" means *weak* stationarity
- For a stationary time series $\{X_t\}$:
  - $E(X_t) = \mu_X$
  - $Cov(X_t, X_{t+h}) = \gamma_X(t, t+h) = \gamma_X(0, h) = \gamma_X(h)$

Definitions

Let $\{X_t\}$ be a stationary time series.

- The **autocovariance function** (ACVF) of $\{X_t\}$ at **lag** h is $\gamma_X(h)$.
- The **autocorrelation function** (ACF) of $\{X_t\}$ at **lag** h is $\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = Corr(X_t, X_{t+h})$.
- $\gamma_X(h) = \gamma_X(-h)$.

(**Reminder**: $Corr(X, Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)} = \frac{Cov(X_t, X_{t+h})}{\sqrt{Var(X_t)Var(X_{t+h})}} = \frac{Cov(X_t, X_{t+h})}{\sqrt{Cov(X_t, X_t)Cov(X_{t+h}, X_{t+h})}} = \frac{Cov(X_t, X_{t+h})}{\sqrt{\gamma_X(0)\gamma_X(0)}} = \frac{\gamma_X(h)}{\gamma_X(0)}$)

Examples

First Order Autoregression: AR(1)

Assume $\{X_t\}$ is a stationary time series satisfying the equations

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

for $t \in Z$, $|\phi| < 1$ and $Z_t \sim WN(0, \sigma^2)$. Also assume Z_t and X_s are uncorrelated for all $s < t$. Calculate the ACVF and ACF of $\{X_t\}$.

- $E(X_t) = \phi E(X_{t-1}) + E(Z_t) \to E(X_t) = \phi E(X_{t-1}) \to E(X_t) = 0$ since $\{X_t\}$ is stationary.
- $\gamma_X(h) = Cov(X_t, X_{t-h}) = E(X_t X_{t-h}) = E(\phi X_{t-1} X_{t-h} + Z_t X_{t-h}) = \phi E(X_{t-1} X_{t-h}) + E(Z_t X_{t-h}) = \phi E(X_{t-1} X_{t-h}) = \phi \gamma_X(h-1) = \phi^h \gamma_X(0)$ (assume $h > 0$).

By stationarity, $\gamma_X(h) = \gamma_X(-h)$ so $\gamma_X(h) = \phi^{|h|} \gamma_X(0)$.

- $\gamma_X(0) = Cov(X_t, X_t) = E(X_t^2) = E(\phi^2 X_{t-1}^2 + 2\phi X_{t-1} Z_t + Z_t^2) = \phi^2 E(X_{t-1}^2) + 2\phi E(X_{t-1} Z_t) + E(Z_t^2) = \phi^2 \gamma_X(0) + \sigma^2 \Rightarrow \gamma_X(0) = \frac{\sigma^2}{1-\phi^2}$

$\therefore \gamma_X(h) = \frac{\phi^{|h|}\sigma^2}{1-\phi^2}$ for $h \in Z$

$\therefore \rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \phi^{|h|}$ for $h \in Z$

ACF signature for AR(1) is exponential decay.

First Order Moving Average: MA(1)

Consider process $X_t = Z_t + \theta Z_{t-1}$ where $t \in N$ and $\{Z_t\} \sim WN(0, \sigma^2)$ and $\theta \in R$. Show $\{X_t\}$ is stationary and derive its ACF.

- $\mu_X = E(X_t) = E(Z_t) + \theta E(Z_{t-1}) = 0$ for all t.
- $\gamma_X(h) = Cov(X_t, X_{t+h}) = Cov(Z_t + \theta Z_{t-1}, Z_{t+h} + \theta Z_{t+h-1}) = Cov(Z_t, Z_{t+h}) + \theta Cov(Z_t, Z_{t+h-1}) + \theta Cov(Z_{t-1}, Z_{t+h}) + \theta^2 Cov(Z_{t-1}, Z_{t+h-1})$

$$\gamma_X(h) = \begin{cases} \sigma^2(1+\theta^2) & \text{if } h = 0 \\ \theta\sigma^2 & \text{if } h = \pm 1 \\ o & \text{otherwise} \end{cases} \quad \leftarrow \text{independent of t.}$$

(Reminder: $Cov(X+Y, W+Z) = Cov(X,W) + Cov(X,Z) + Cov(Y,W) + Cov(Y,Z)$)

$\therefore \{X_t\}$ is stationary.

- $\gamma_X(0) = \sigma^2(1+\theta^2)$

and $\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \{1$ if $h = 0$, $\frac{\theta}{1+\theta^2}$ if $h \pm 1$, 0 otherwise.

ACF signature of MA(1) is a spike for $h = 0, 1$ and then nothing for $h > 1$.

We've seen that the ACF can provide information regarding which model may be appropriate for an observed time series. To do this in practice, we need a sample estimate of the ACF.

Definitions

Let $x_1, x_2, \ldots, x_n$ be our observed time series.

- the **sample mean** is $\hat{\mu}_x = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,

- the **sample autocovariance** is $\hat{\gamma}_x(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x})$,
- the **sample autocorrelation** is $\hat{\rho}_x(h) = \frac{\hat{\gamma}_x(h)}{\hat{\gamma}_x(0)}$

Note:

- θ, a Greek letter, denotes a paramters (unknown number),
- $\hat{\theta}$, is a sample estimate of θ (known number),
- $\tilde{\theta}$, is an estimator, a random variable.

The sample ACF can be used to investigate the "uncorrelatedness" in a time series. For example, we might use this to evaluate the uncorrelated assumption in residuals.

(Reminder: independence $\Rightarrow$ uncorrelated; uncorrelated $\not\Rightarrow$ independence)

For stationary time series, $\tilde{\rho}(h) \dot{\sim} N(0, \frac{1}{n})$ (n = number of data points).

Consequently, an approximate 95% confidence interval for $\rho_x(h)$ is $\pm\frac{1.96}{\sqrt{n}}$.

If $\tilde{\rho}(h)$ falls outside these limits, for any h, we judge this to be significant.

SACF Examples.R

10/27/16

Recap

- Autocovariance function (ACVF): $\gamma_X(h) = Cov(X_t, X_{t-h})$ for all $h \in Z$
- Autocorrelation function (ACF): $\rho_X(h) = Corr(X_t, X_{t-h}) = \frac{\gamma_X(h)}{\gamma_X(0)}$
  - Properties of ACVF:
    * $\gamma_X(0) = Var(X_t)$
    * $\gamma_X(-h) = \gamma_X(h)$
    * $|\rho_X(h)| \leq 1$

Why is stationarity important?

In order to build a model that forecasts with any accuracy, we require an assumptions that something doesn't vary with time. After accounting for deterministic trend and/or seasonality, we hope that the remaining randomness can be described as stationary.

In the Box-Jenkins class of models, we can use AR (autoregressive), MA (moving average), and ARMA models to model stationary time series.

First, notation:

Backshift operator: B, where $BX_t = X_{t-1}$ i.e. $B^2X_t = X_{t-2}$.

Generally, $B^nX_t = X_{t-n}$ and $B^0 = I$

MA(q) Process

A process/time series $\{X_t\}$ is called a **moving average process of order q** if

$$X_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \ldots + \theta_q\epsilon_{t-q}$$

where $\{\epsilon_t\} \sim WN(0, \sigma^2)$ and $\theta_1, \theta_2, \ldots, \theta_q$ are constants.

Remarks:

- MA(q) processes are stationary (*exercise*: prove this!)
- An MA(q) process is **q-correlated** (i.e., $\rho_X(h) = Corr(X_t, X_{t-h}) = 0$ for $h > q$ and not necessarily 0 for $h \leq q$)

Thus, **the ACF signature of an MA(q) process is non-zero spikes for** $h = 0, 1, 2, \ldots, q$ **and then no spikes for ever after**.

- An MA(q) process can be denoted as: $X_t = \epsilon_t + \theta_1 B^1 \epsilon_t + \ldots + \theta_q B^q \epsilon_t = (1 + \sum_{s=1}^{q} \theta_s B^s)\epsilon_t = \theta^q(B)\epsilon_t$

where $\theta^q(z) = 1 + \sum_{s=1}^{q} \theta_s z^s$ is the **generating function**.

- An MA(q) is **invertible** if the complex roots of $\theta^q(z)$ lie outside the unit circle. i.e. For all z such that $\theta^q(z) = 0$, then $|z| > 1$.

Example

$X_t = \epsilon_t + 0.2\epsilon_{t-1} + 0.7\epsilon_{t-2}$

$\theta(z) = 1 + 0.2z + 0.7z$

The roots of $\theta(z)$ are $z = \frac{-0.2 \pm \sqrt{0.2^2 - 4(0.7)(1)}}{2(0.7)} = \frac{-0.2 \pm \sqrt{2.76}i}{1.4} \Rightarrow z = -0.14 \pm 1.19i$

$|z| = \sqrt{(-0.14)^2 + (1.19)^2} = 1.198 > 1$

So $\{X_t\}$ is invertible.

Reminders:

- The zeros of a quadratic of the form $ax^2 + bx + c$ are $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$
- $c = a + ib \Rightarrow |c| = \sqrt{a^2 + b^2}$

AR(p) Process

The process $\{X_t\}$ is called an **autoregressive process of order p** if

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t$$

where $\epsilon_t \sim WN(0, \sigma^2)$ and $\phi_1, \phi_2, \ldots, \phi_p$ are constants.

- An AR(p) process can be denoted as:

$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \ldots - \phi_p X_{t-p} = \epsilon_t$

$\Leftrightarrow X_t - \phi_1 B^1 X_t - \phi_2 B^2 X_t - \ldots - \phi_p B^p X_t = \epsilon_t$

$\Leftrightarrow (1 - \sum_{r=1}^{p} \phi_r B^r) X_t = \epsilon_t$

$\Leftrightarrow \phi^p(B) X_t = \epsilon_t$

where $\phi^p(z) = 1 - \sum_{r=1}^{p} \phi_r z^r$ is the **generating function**.

- An AR(p) process is **stationary** is the complex roots of $\phi^p(z)$ lie outside the unit circle. i.e. For all z such that $\phi(z) = 0$, we require $|z| > 1$.

Example

$X_t = \phi X_{t-1} + \epsilon_t \Rightarrow (1 - \phi B)X_t = \epsilon_t$

$\phi(z) = 1 - \phi z \Rightarrow \phi(z) = 0$ if $z = \frac{1}{\phi}$

For stationarity, we need $|z| > 1 \Rightarrow |\frac{1}{\phi}| > 1 \Rightarrow |\phi| > 1$.

Partial Autocorrelation Function(PACF)

For a stationary process, the ACF of lag h measures the correlation between X_t and X_{t+h}. This correlation could be due to a direct connection between X_t and X_{t+h}, but it may also be influenced by observations at intermediate lags: $X_{t+1}, X_{t+2}, \ldots, X_{t+h-1}$.

The PACF of lag h measures the correlation between X_t and X_{t+h} once the influence of the intermediate lags has been removed/accounted/controled for.

We remove this effect using **linear predictors**:

$$\hat{X}_t = Pred(X_t | X_{t+1}, X_{t+2}, \ldots, X_{t+h-1})$$

$$\hat{X}_{t+h} = Pred(X_{t+h} | X_{t+1}, X_{t+2}, \ldots, X_{t+h-1})$$

where this prediction is commonly based on a linear regression.

Thus, for a stationary time series $\{X_t\}$, the **partial autocorrelation function of lag h** is: $\alpha_X(h) = \begin{cases} Corr(X_t, X_t) = 1, & \text{if } h = 0 \\ Corr(X_t, X_{t+1}) = \rho_X(1), & \text{if } h = 1 \\ Corr(X_t, X_{t+h}) = Corr(X_t - \hat{X}_t, X_{t+h} - \hat{X}_{t+h}) & \text{if } h > 1 \end{cases}$

(assume without loss of generality that $h \geq 0$)

Example

Derive the PACF of an AR(1) process $X_t = \phi X_{t-1} + \epsilon_t$.

$\alpha_X(h) = \begin{cases} 1 & \text{if } h = 0 \\ \rho(1) = \phi & \text{if } h = 1 \end{cases}$

If $h = 2$:

- $\alpha(2) = Corr[X_t - \hat{X}_t, X_{t+2} - \hat{X}_{t+2}] = Corr[X_t - f(X_{t+1}), X_{t+2} - \phi X_{t+1}] = Corr[X_t - f(X_{t+1}), \epsilon_{t+2}] = Corr[X_t, \epsilon_{t+2}] - Corr[f(X_{t+1}), \epsilon_{t+2}] = 0 - 0 = 0$

We can see that $\alpha(h) = 0$ for any $h \geq 2$.

So **PACF for an AR(1) has non-zero spikes for** $h = 0, 1$ **and is zero for all** $h \geq 2$.

Remarks:

- If $\{X_t\} \sim AR(p)$, then the PACF satisfies $\alpha(h) = 0$ for all $h > p$ and $\alpha(h) \neq 0$ necessarily for $h \leq p$.
- Whereas an ACF can be used to determine the order of an MA process, a PACF can be used to determine the order of an AR process.

ARMA(p,q) Process

$\{X_t\}$ is an **autoregressive moving average process of orders p and q** if

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \ldots - \phi_p X_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}$$

$$\phi^p(B) X_t = \theta^q(B)\epsilon_t$$

where $\{\epsilon_t\} \sim WN(0, \sigma^2)$ and $\phi^p(z)$ and $\theta^q(z)$ are the AR and MA generating functions, and **we require them to have distinct roots**.

Remark:

- ARMA(p, 0) = AR(p)
- ARMA(0, q) = MA(q)

Example: ARMA(1,2)

$\phi^1(B)X_t = \theta^2(B)\epsilon_t \Rightarrow (1 - \phi B)X_t = (1 + \theta_1 B + \theta_2 B^2)\epsilon_t \Rightarrow X_t - \phi X_{t-1} = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}$

We require $\begin{cases} \phi^1(z) = 1 - \phi z \\ \theta^2(z) = 1 + \theta_1 z + \theta_2 z^2 \end{cases}$

	ACF	**PACF**
MA(q)	Spike for $h \leq q$ and negligibly small spikes for $h > q$	Exponential decay
AR(p)	Exponential decay	Spikes for $h \leq p$ and "nothing" for $h > p$
ARMA(p,q)	q spikes then decay	p spikes then decay

$---$

11/01/16

ARMA Stationarity and Invertibility Conditions

ARMA(p,q): $\phi(B)X_t = \theta(B)\epsilon_t$, $\{\epsilon_t\} \sim WN(o, \sigma^2$

where $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p$

and $\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \ldots + \theta_q z^q$.

$\{X_t\} \sim ARMA(p, q)$ is stationary if

- $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p \neq 0$ for all z such that $|z| \leq 1$ i.e., the modulus of all roots/zeros are > 1 (the complex roots of this generating function lie outside the unit circle in the complex plane).

A **causal** ARMA process is stationary.

$\{X_t\} \sim ARMA(p, q)$ is invertible if

- $\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \ldots + \theta_q z^q \neq 0$ for all z such that $|z| \leq 1$ i.e., the modulus of all roots/zeros are > 1 (the complex roots of this generating function lie outside the unit circle in the complex plane).

This criterion is equivalent to requiring that ϵ_t can be written as an infinite weighted sum of the X_t's.

Note: we require an MA(q) process, or the MA component of an ARMA(p,q) process to be invertible so that

- (i) we can estimate the θ's in the model

- (ii) we can forecast with the model

For the sake of usefulness, we'll restrict attention to ARMA(p,q) models that are stationary/causal and invertible.

Example (quiz question)

ARMA(2,1)

$\phi^2(B)X_t = \theta^1(B)\epsilon_t$

- *Represent this in "expanded notation"*

$(1 - \phi_1 B - \phi_2 B^2)X_t = (1 + \theta B)\epsilon_t$

$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = \epsilon_t + \theta\epsilon_{t-1}$

Let $\phi_1 = 0.75$, $\phi_2 = -0.5625$, $\theta = 1.25$

- *Is* $\{X_t\}$ *stationary?*

$\phi^2(z) = 1 - 0.75z + 0.5625z^2$

$\Rightarrow \phi^2(z) = 0 \Leftrightarrow z = \frac{-b \pm \sqrt{b^2-4ac}}{2a} = \frac{0.75 \pm \sqrt{0.75^2 - 4(0.5625)(1))}}{2(0.5625)} = 2(\frac{1 \pm \sqrt{-3}}{3}) = \frac{2 \pm 2i\sqrt{3}}{3}$

$\Rightarrow z_1 = \frac{2}{3} - \frac{2\sqrt{3}}{3}i, z_2 = \frac{2}{3} + \frac{2\sqrt{3}}{3}i$

$\Rightarrow |z_1| = |z_2| = \sqrt{(2/3)^2 + (2\sqrt{3}/3)^2} = \frac{4}{3}$

So $|z_1| = |z_2| > 1$.

$\therefore$ this ARMA(2,1) process is stationary.

- *Is* $\{X_t\}$ *invertible?*

$\theta(z) = 1 + 1.25z$

$\Rightarrow \theta(z) = 0 \Leftrightarrow 1 + 1.25z = 0 \Leftrightarrow z = -0.8$

$\Rightarrow |z| = 0.8 < 1.$

$\therefore$ this ARMA(2,1) process is **not** invertible.

Note: also, the roots of the polynomials are **distinct**, so the process is in fact an ARMA(2,1) process, and not a simpler one.

In practice, with an observed time series, we decide whether it is stationary and/or invertible with "**unit root tests**".

"Box-Jenkins Approach"

- **Identification**: identify the orders of the model ← use ACF/PACF plot
- **Estimation**: estimate the parameters of the model identified in step 1
- **Verification**: ensure that the model is appropriate ← residual diagnostics

Estimating ARMA(p,q) Models

Goal: estimate $\phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q, \sigma^2$ in a stationary and invertible ARMA(p,q) process: $\phi^p(B)X_t = \theta^q(B)\epsilon_t$.

- we assume that $\{X_t\}$ has zero mean, or has been "mean-corrected"

These parameters are estimated using the observed time series $\{x_1, x_2, \ldots, x_n\}$.

Many methods (cf. Chap. 5) exist for doing this, but we'll just focus, at a high level, on Maximum Likelihood and Least Squares.

Maximum Likelihood method for ARMA(p,q)

- We have to make distributional assumptions, which may not be valid
- We typically assume $\{X_t\}$ is a "Gaussian" time series i.e., $\vec{X} = (X_1, X_2, \ldots, X_n)^T \sim MVN$
  - This seems limiting, but in practice it's not bad

$L(\phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q, \sigma^2|\vec{x}) = \frac{1}{(2\pi)^{n/2}|\Gamma_n|^{1/2}} \exp\{-\frac{1}{2}\vec{x}^T\Gamma_n^{-1}\vec{x}\}$ where $\vec{x} = (x_1, x_2, \ldots, x_n)^T$ and $\Gamma_n = Cov(\vec{X}, \vec{X}^T)$ (this is the matrix with $\gamma_X(0)$ in the diagonal, $\gamma_X(1)$ in the 1-subdiagonals, ..., $\gamma_X(n-1)$ in the two corners)

We want to find the values of the parameters that maximize this function in light of the observed data. We typically numerically maximize $l(\cdot)$, the log-likelihood function to find $(\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p, \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_q, \hat{\sigma}^2)$.

The usual asymptotic properties of MLE exist here as well. This is the basis for CI calculations.

Least Squares method for ARMA(p,q)

The goal is to minimize $S(\phi, \theta) = S(\phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q)$ rather than maximize $L(\cdot)$, where

$$S(\phi, \theta) = \sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{r_{j-1}}$$

where $E[(X_j - \hat{X}_j)^2] = \sigma^2 r_{j-1} \rightarrow$ LSE of σ^2 is $\hat{\sigma}^2 = \frac{S(\hat{\phi}, \hat{\theta})}{n-p-q}$ $\qquad (r_{j-1} = Var(X_j - \hat{X}_j))$

LSE is good because no distributional assumptions need to be made.

11/03/16

Order Selection

- Choose p and q "optimally".
- Use "goodness of fit" methods to compare different models.
  - $l(\hat{\theta}, \hat{\phi})$ = maximize log-likelihood. $\leftarrow$ we want this to be big
  - $\hat{\sigma}^2$ = estimate of the error variance σ^2. $\leftarrow$ we want this to be small
  - $AIC = -2l(\hat{\theta}, \hat{\phi}) + 2(p + q + 1)$. $\leftarrow$ we want this to be small
- It's sensible to consider all 3 of these, but an "optimal" model for one may not be the "optimal" model according to another.

- A disadvantage to LSE is that we don't have a likelihood function and so $l(\hat{\theta}, \hat{\phi})$ and AIC are not available goodness of fit metrics in this case.

We can compare the fit of two models using a **likelihood ratio test (LRT)**.

$$\begin{cases} H_0: & \text{null and alternative models fit equally well} \\ H_a: & \text{alternative model fits better than the null} \end{cases}$$

Note: the null model is simpler (has fewer parameters) relative to the alternative.

$$D = -2\log\left(\frac{L(null\ model)}{L(alt.\ model)}\right) \sim \chi^2_{(m_A - m_0)}$$

null model has m_o parameters, alt.model has m_A parameters, $m_A > m_0$.

$$D = -2(l(null\ model) - l(alt.\ model))$$

Larger values of D lead to rejection of H_0. p-value $= P(W \geq D) = 1 - P(W < D)$ where $W \sim \mathrm{X}^2_{(m_A - m_0)}$.

ARMA Fitting Example.R

Verification (Residual Diagnostics)

Suppose $\{X_t\}$ is a time series and we believe an ARMA(p,q) model can model it. We'll call the **fitted value** at time t $\hat{X}_t$. We define the residuals as

$$\hat{e}_t = X_t - \hat{X}_t \qquad \text{for } t = 1, 2, \ldots, n$$

If assumptions are met, the residual time series $\{\hat{e}_t\}$ should behave like the white noise sequence that generated the ARMA(p,q) process. In particular, we should find that $\{\hat{e}_t\}$

i. have approximately zero mean
ii. have a constant variance
iii. are uncorrelated (/ independent iff $\{\epsilon_t\} \sim IID(0, \sigma^2)$)
iv. are normally distributed (if $\{\epsilon_t\} \sim N(0, \sigma^2)$) ← only if you are using MLE

We can either work with the residuals, $\hat{e}_t$, or the **standardized residuals**, $\hat{r}_t = \frac{\hat{e}_t}{\hat{\sigma}}$ (expect the variance to be 1).

Informal Diagnostics (plots)

- Plot $\hat{e}_t$ vs. t (or $\hat{r}_t$ vs. t)
  - change of variability with time? i.e. heteroskedasticity
    - this checks ii.
  - check whether points are scattered symmetrically around 0
    - this checks i.
  - systematic trends in the residuals can suggest correlation
    - this checks iii.
  - check for outliers (using $\hat{r}_t$ is sensible)

- ACF of $\hat{e}_t$ (or $\hat{r}_t$)
  - use this to check whether residuals seem to be correlated
    - this checks iii.
  - should see no significant spikes for $h > 0$
- QQ-plot or histogram
  - use this to check whether the residuals seem normally distributed
    - this checks iv.

Formal Diagnostics (hypothesis tests)

- To check $E[\epsilon_t]$, do a one-sample t-test of the residuals
- To check heteroskedasticity, we can use Bartlett's Test or Levene's Test
  - these tests require us to partition the data set (the residuals) into k groups. The goal is to look for homogeneity of variance among these groups
  - $\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 \\ H_a : \sigma_i^2 \neq \sigma_j^2 \qquad \text{for some } i \neq j \end{cases}$
  - Bartlett's test is sensitive to non-normality, but Levene's test isn't. But if the data are normally distributed, Bartlett's test is more powerful
- To check uncorrelatedness, we're interested in testing
  - $\begin{cases} H_0 : \rho(1) = \rho(2) = \ldots = \rho(H) = 0 \\ H_a : \rho(h) \neq 0 \qquad \text{for some } h = 1, 2, \ldots, H \end{cases}$
  - we prefer this test as opposed to using an ACF because it avoids the **multiple hypothesis testing problem**
  - we use "Portmanteau" tests in this case. We consider the **Ljung-Box** test
- To check for normality, use the **Shapiro-Wilk** test where
  - $\begin{cases} H_0 : \{\hat{e}_t\} \sim N(0, \sigma^2) \\ H_a : \{\hat{e}_t\} \nsim N(0, \sigma^2) \end{cases}$

11/08/16

Residual Diagnostics Example.R

Non-stationary models

Idea: transform a non-stationary time series so that the new series is stationary, and hence can be modeled by an ARMA process.

ARIMA (AutoRegressive Integrated Moving Average)

ARIMA processes model non-stationary time series that, after finitely many "differences", become an ARMA.

If d is a non-negative integer, then $\{X_t\}$ is an ARIMA(p,d,q) process if $Y_t = (1-B)^d X_t$ is a stationary ARMA(p,q) process.

This definition implies that $\{X_t\}$ can be written as

$$\phi^*(B)X_t = \theta^q(B)\epsilon_t,\ \{\epsilon_t\} \sim WN(0, \sigma^2)$$

where $\phi^*(B) = (1-B)^d\phi^p(B)$

$\{Y_t\} \sim ARMA(p,q) \Rightarrow \phi^p(B)Y_t = \theta^q(B)\epsilon_t$

$$\Rightarrow \phi^p(B)(1-B)^d X_t = \theta^q(B)\epsilon_t$$

$\phi^*(z) = (1-z)^d\phi^p(z) \Rightarrow z = 1$ is a root of $\phi^*(z)$

$\Rightarrow \{X_t\}$ is not stationary unless $d = 0$, in which case $\{X_t\} \sim ARMA(p,q)$.

What is differencing? (it allows us to model trend)

Notation: $\nabla = (1-B)$ and $\nabla^d = (1-B)^d$

$\nabla X_t = (1-B)X_t = X_t - BX_t = X_t - X_{t-1}$

$\nabla^2 X_t = (1-B)^2 X_t = (1-2B+B^2)X_t = X_t - 2BX_t + B^2X_t = X_t - 2X_{t-1} + X_{t-2}$

Remark: if $\{X_t\}$ exhibits polynomial trend of the form $m_t = \sum_{i=0}^{d} \alpha_i t^i$, the time series $Y_t = \nabla^d X_t = (1-B)^d X_t$ will no longer have a trend component.

Examples

a) $X_t = c + bt + \epsilon_t$

$\nabla X_t = (1-B)X_t = (1-B)(c+bt+\epsilon_t) = c+bt+\epsilon_t - B(c+bt+\epsilon_t) = c+bt+\epsilon_t - c - b(t-1) - \epsilon_{t-1} = b + \nabla\epsilon_t$

b) $X_t = c + bt + at^2 + \epsilon_t$

$\nabla^2 X_t = (1-B)^2 X_t = (1-2B+B^2)X_t = (1-2B+B^2)(c+bt+at^2+\epsilon_t)$

$= (c+bt+at^2) - 2(c+b(t-1)+a(t-1)^2) + (c+b(t-2)+a(t-2)^2) + \nabla^2\epsilon_t$

$= \ldots$

$= 2a + \nabla^2\epsilon_t$

In these cases, and in all cases generally, after an appropriate number of differences d, $\nabla^d X_t$ no longer has any trend.

Once we've identified a suitable value of d, we model the resulting time series with an ARMA(p,q) process.

But how do we choose d?

- Informally: plot of the time series or ACF plot
  - difference until the transformed time series "looks" stationary
  - difference until the transformed time series exhibits rapid decay (as opposed to slow decay)
- Formally: **unit root tests**

- tests that identify whether an observed time series has a root on or "near" the unit circle
- if the AR component has a unit root, then the time series needs to be differenced

Dickey-Fuller Test

Motivate with AR(1): $X_t = \phi X_{t-1} + \epsilon_t$

$\nabla X_t = X_t - X_{t-1} = (\phi X_{t-1} + \epsilon_t) - X_{t-1} = (\phi - 1)X_{t-1} + \epsilon_t = \phi^* X_{t-1} + \epsilon_t$

Using OLS, we test $H_0 : \phi^* = 0$ ($\Leftrightarrow \phi = 1$) vs $H_a : \phi^* \neq 0$ ($\Leftrightarrow \phi \neq 1$)

So the null hypothesis in the DF-test is that the time series is non-stationary. So if we reject H_0 (p-value $< \alpha$), we conclude the time series is stationary.

- if we reject $H_0 \rightarrow$ stop differencing
- if we fail to reject $H_0 \rightarrow$ difference again

The **Augmented Dickey-Fuller (ADF)** generalizes this for any order p in AR(p).

This type of differencing **doesn't remove seasonality**. To remove seasonality, we need **seasonal differencing.**

ARIMA-SARIMA Examples.R

11/10/16

Recap

$\nabla = (1 - B)$

$\nabla X_t = X_t - X_{t-1} \leftarrow$ ordinary differencing removes trend but not seasonality.

Seasonal Differencing

Notation: $(1 - B^k) = \nabla_k \leftarrow$ **this is not the same as** $\nabla^k = (1 - B)^k$.

$\nabla_k X_t = (1 - B^k)X_t = X_t - B^k X_t = X_t - X_{t-k}$

We call this **lag-k differencing**.

$k = 1 \Leftrightarrow$ ordinary differencing.

In R,

- $\nabla^d X_t = (1 - B)^d X_t =$ `diff(x, difference=d)`
- $\nabla_k X_t = (1 - B^k)X_t =$ `diff(x, lag=k)`
- $\nabla_k^D X_t = (1 - B^k)^D X_t =$ `diff(x, lag=k, difference=D)`

Idea: seasonal effects s_t manifest themselves as $s_t = s_{t+m}$, where m is the **period** of the seasonal effect.

Example

$X_t = s_t + \epsilon_t$, $\{\epsilon_t\} \sim WN(0, \sigma^2)$

If we take a lag-m difference (**seasonal difference**), we will remove the seasonal effect:

$\nabla_m X_t = \nabla_m s_t + \nabla_m \epsilon_t = (1 - B^m)s_t + (1 - B^m)\epsilon_t = (s_t - s_{t-m}) + (\epsilon_t - \epsilon_{t-m}) = \epsilon_t - \epsilon_{t-m}$ since $s_t = s_{t-m}$

We hope that after 1 or 2 applications of seasonal differencing, we're left with a stationary time series that can be modeled by an ARMA(p,q) process.

So we use seasonal differencing to eliminate seasonal effects and ordinary differencing to eliminate trend. Order of differencing does not matter.

But how do we choose m?

We can use an ACF plot, which will exhibit seasonality, by counting the number of lags between "peaks". This number corresponds to the period.

SARIMA (Seasonal AutoRegressive Integrated Moving Average)

If $Y_t = (1 - B)^d(1 - B^s)^D X_t$ is stationary, and hence can be modeled by an ARMA(p,q) process, we say $\{X_t\}$ is SARIMA. In particular,

$$\{X_t\} \sim SARIMA(p, d, q) \times (P, D, Q)_s$$

where:

- s is the period of the seasonal effect
- d is the number of ordinary differences necessary to remove trend
- D is the number of seasonal differences necessary to remove seasonality
- p, q are the AR and MA orders of the **within-period** process
- P, Q are the AR and MA orders of the **between-period** process

Typically, $d, D = 1$ **or** 2 **and** $p, P, q, Q < 3$

We can write this in generating function notation as follows:

$$\phi^p(B)\Phi^P(B^s)X_t = \theta^q(B)\Theta^Q(B^s)\epsilon_t$$

where $\{\epsilon_t\} \sim WN(0, \sigma^2)$

Idea:

- the data within a period is a time series,
- the data between periods is a time series,
- these two time series have different ARMA representations.

Order Selection

Step 1: Choose d, D, s such that $Y_t = (1 - B)^d(1 - B^s)^D X_t$ is stationary.

Step 2: Examine the ACF and PACF plots of $\{Y_t\}$ for p, P, q, Q.

- p and q are chosen such that $\rho(1), \rho(2), \ldots, \rho(s-1)$ and $\alpha(1), \alpha(2), \ldots, \alpha(s-1)$ are compatible with ARMA(p,q).
- P and Q are chosen such that $\rho(ks)$ and $\alpha(ks)$ where $k = 1, 2, \ldots$ are compatible with ARMA(P,Q).

Step 3: Fit models in the neighborhood of these choices and use goodness of fit metrics to choose an optimal model.

Full Box-Jenkins Approach

1. Check for non-constance variance and apply transformation if necessary.
2. Check for seasonal and non-seasonal trends.
3. Use differencing to make the time series stationary.
4. Identify p, P, q, Q from the ACF and PACF plots of the differenced data.
5. Fit the proposed model and iterate to an optimal one.
6. Check residuals to assess assumptions.
7. Forecast into the future.

ARIMA-SARIMA Examples.R

11/15/16

chemical.txt

RMSE = $\sqrt{\frac{\sum_{t=1}^{n}(x_t-\hat{x}_t)^2}{n}}$

wine.txt

11/17/16

Forecasting

Based on the "history" of the process up to time n, i.e. $X_1, X_2, \ldots, X_n$, we are interested in deriving a prediction for X_{n+h}, $h > 0$, denoted $\hat{X}_{n+h}$ which minimizes $MSE = E[(X_{n+h} - \hat{X}_{n+h})^2 | X_1, X_2, \ldots, X_n]$.

The value that minimizes this (recall *Assignment 1*) is $\hat{X}_{n+h} = E[X_{n+h}|X_1, X_2, \ldots, X_n]$.

Suppose $\{X_t\} \sim SARIMA(p,d,q) \times (P,D,Q)_s$. Forecasts $\hat{X}_{t+h}$ are found by substituting estimates of each parameter into the model, and taking $\hat{X}_{t+h} = E[X_{t+h}|X_1, X_2, \ldots, X_t]$.

Example

$\{X_t\} \sim ARIMA(2,1,1)$

$\Leftrightarrow \phi^2(B)(1-B)X_t = \theta(B)\epsilon_t$

$\Leftrightarrow (1-\phi_1 B - \phi_2 B^2)(1-B)X_t = (1+\theta B)\epsilon_t$

$\Leftrightarrow (1-\phi_1 B - \phi_2 B^2 - B + \phi_1 B^2 + \phi_2 B^3)X_t = \epsilon_t + \theta B \epsilon_t$

$\Leftrightarrow (1-(\phi_1+1)B - (\phi_2-\phi_1)B^2 + \phi_2 B^3)X_t = \epsilon_t + \theta B \epsilon_t$

$\Leftrightarrow X_t - (\phi_1+1)X_{t-1} - (\phi_2-\phi_1)X_{t-2} + \phi_2 X_{t-3} = \epsilon_t + \theta\epsilon_{t-1}$

$\Leftrightarrow X_t = (\phi_1+1)X_{t-1} + (\phi_2-\phi_1)X_{t-2} - \phi_2 X_{t-3} + \epsilon_t + \theta\epsilon_{t-1}$

But we care about predicting X_{t+h}.

$\Rightarrow X_{t+h} = (\phi_1 + 1)X_{t+h-1} + (\phi_2 - \phi_1)X_{t+h-2} - \phi_2 X_{t+h-3} + \epsilon_{t+h} + \theta\epsilon_{t+h-1}$

In order to obtain $\hat{X}_{t+h}$, we substitute estimates of our parameters into the equation above, and we replace ϵ's with residuals for time points in the past, and zeros otherwise.

For example, $\hat{X}_{t+1} = (\hat{\phi}_1 + 1)X_t + (\hat{\phi}_2 - \hat{\phi}_1)X_{t-1} - \hat{\phi}_2 X_{t-2} + \hat{\theta}e_t$

$$\hat{X}_{t+2} = (\hat{\phi}_1 + 1)X_{t+1} + (\hat{\phi}_2 - \hat{\phi}_1)X_t - \hat{\phi}_2 X_{t-1} + \hat{\theta}e_{t+1}$$

We can calculate $SE(\hat{X}_{t+h})$ with which we calculate prediction intervals. Assuming the error process is normally distributed, a $100(1-\alpha)\%$ prediction interval is given by $\hat{X}_{t+h} \pm Z_{1-\alpha/2}SE(\hat{X}_{t+h})$.

$SE(\hat{X}_{t+h})$ is a function of $\hat{\sigma}^2$ and h, where the SE increases as h increases.

Forecasting Examples.R (wine)

In practice, when performing model selection, we should divide the observed time series into **training** and **test** sets. As a general rule-of-thumb, we use 80% training vs 20% test. The test set should correspond to the final 20% of the data.

When prediction is the primary goal, we can define a model as "optimal" if it minimizes the prediction **root mean squared error** (RMSE). For a test set $\{x_1.x_2.\ldots,x_n\}$, the RMSE is:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(x_t - \hat{x}_t)^2}{n}}$$

where $\hat{x}_i$ is the prediction of x_i.

Forecasting Examples.R (chemical)

Practical Considerations

Since one-step-ahead predictions are most accurate, let's minimize h.

- **Rolling Window** (RW): rolls forward including all past data (length increases). Constant starting point, with ever increasing size.
- **Moving Window** (MW): rolls forward keeping a fixed length (doesn't include all past data). Constant size, with ever increasing starting point.

The window size is context-specific and can be determined by trial and error.

RW Advantages and Disadvantages

- Learns from the past and it "doesn't forget"
  - can reduce variance σ^2
- Sensitive to shocks in the system

MW Advantages and Disadvantages

- Not as affected by shocks in the system, so it can adapt
- May "lose"/"forget" valuable information

11/22/16

Exponential Smoothing (ES)

The objective is to predict X_{n+h} given the history $\{x_1, x_2, \ldots, x_n\}$ of observations up to time n. Using ES techniques, we do so by using a set of recursive equations that do not require any distributional assumptions.

We'll use different sets of equations depending on whether the observed time series has

1) no trend + no seasonality (Single ES)

2) trend + no seasonality (Double ES)

3) trend + seasonality (Triple ES)

1) Simple (Single) Exponential Smoothing (SES)

Here, $\hat{X}_{n+h} = a_n$ for $h = 1, 2, 3, \ldots$ where $a_t = \alpha X_t + (1-\alpha)a_{t-1}$.

a_t is called the **level** of smoothing.

$\rightarrow$ this is commonly referred to as **exponentially weighted moving average** (EWMA).

$\rightarrow$ $0 \leq \alpha \leq 1$ is a "smoothing constant":

- If $\alpha = 0$, $a_t = a_0$ $\forall t$, where a_0 is the starting value of the recursion.
- If $\alpha = 1$, $a_t = X_t$ $\forall t$, and no smoothing has occurred at all.

$\rightarrow$ so small α gives more smoothing, and large α gives less.

$\rightarrow$ $\alpha = 0.2$ is typically a good choice, but an optimal α can be determined by minimizing the sum of squared one-step-ahead prediction errors: $SSE_{PI} = \sum_{i=2}^{n} e_i^2$ where $e_i = x_i - \hat{x}_i = x_i - a_{i-1}$.

So why is it called exponential?

$$
\begin{aligned}
a_t &= \alpha X_t + (1-\alpha)a_{t-1} \\
&= \alpha X_t + (1-\alpha)(\alpha X_{t-1} + (1-\alpha)a_{t-2}) \\
&= \alpha X_t + \alpha(1-\alpha)X_{t-1} + (1-\alpha)^2 a_{t-2} \\
&= \alpha X_t + \alpha(1-\alpha)X_{t-1} + (1-\alpha)^2(\alpha X_{t-2} + (1-\alpha)a_{t-3}) \\
&= \alpha X_t + \alpha(1-\alpha)X_{t-1} + \alpha(1-\alpha)^2 X_{t-2} + (1-\alpha)^3 a_{t-3} \\
&= \ldots \\
&= \textstyle\sum_{i=0}^{t-1} \alpha(1-\alpha)^i X_{t-1} + (1-\alpha)^t a_0
\end{aligned}
$$

Thus a_t is literally an exponentially weighted moving average of $\{X_t\}$.

We also see that as t increases, the dependence on $a_0 \rightarrow 0$, so the value we choose for a_0 is not important, though $a_0 = x_1$ or $a_0 = \bar{x}$ are sensible choices.

Smoothing Examples.R (Single Exponential Smoothing)

We can see that any long-term dependency on time, through trend and/or seasonality, for example, is not going to be accounted for here.

2) Double Exponential Smoothing (DES)

Here, $\hat{X}_{n+h} = a_n + hb_n$ for $h = 1, 2, 3, \ldots$ where

- $a_t = \alpha X_t + (1 - \alpha)(a_{t-1} + b_{t-1})$
  - a_t is called the **level** of smoothing.
  - weighted average of observed value at time $t - 1$ and the predicted value at time $t - 1$.
- $b_t = \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1}$
  - b_t is called the **trend** of smoothing.
  - weighted average of previous changes in level.

$\rightarrow 0 < \alpha < 1$ and $0 < \beta < 1$ are "smoothing constants".

Smoothing Examples.R (Double Exponential Smoothing)

We can see that DES works fine if the non-stationarity is due to trend. If seasonality exists, we'll want TES.

3) Triple Exponential Smoothing (TES)

Additive case

Here, $\hat{X}_{n+h} = a_n + hb_n + s_{n+h-m}$ for $h = 1, 2, 3, \ldots$ where m is the period and

- $a_t = \alpha(X_t - s_{t-m}) + (1 - \alpha)(a_{t-1} + b_{t-1})$
  - a_t is called the **level** of smoothing.
  - weighted average of the seasonally adjusted observation and the non-seasonal forecast at time t.
- $b_t = \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1}$
  - b_t is called the **trend** of smoothing.
  - weighted average of previous changes in level.
- $s_t = \gamma(X_t - a_t) + (1 - \gamma)s_{t-m}$
  - s_t is called the **seasonality** of smoothing.
  - weighted average of the current seasonal index and the seasonal index from the previous season.

$\rightarrow 0 < \alpha < 1$, $0 < \beta < 1$, and $0 < \gamma < 1$ are "smoothing constants" that can be chosen or estimated by minimizing the sum of squared one-step-ahead prediction errors.

Smoothing Examples.R (Triple Exponential Smoothing - Additive)

Multiplicative case

Here, $\hat{X}_{n+h} = (a_n + hb_n)s_{n+h-m}$ for $h = 1, 2, 3, \ldots$ where m is the period and

- $a_t = \alpha(\frac{X_t}{s_{t-m}}) + (1 - \alpha)(a_{t-1} + b_{t-1})$
- $b_t = \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1}$
- $s_t = \gamma(\frac{X_t}{a_t}) + (1 - \gamma)s_{t-m}$

This is useful when variability increases over time.

Smoothing Examples.R (Triple Exponential Smoothing - Multiplicative)

11/29/16

Multivariate Time Series

Until now, we've only considered univariate time series; that is, we've been interested in forecasting the time series of a single variable using only its own history.

If we observe other time series that are correlated with the one of primary interest (the **response**), we'd like to incorporate this additional information as it may improve the accuracy of forecasts.

Depending on how we treat the relationship between these variables, we will take different approaches.

1) If we treat these external variables as **exogenous**, i.e., they influence the response but not the other way around, we can fit **ARIMAX** models to account for this relationship.
2) If we treat these external variables as **endogenous**, i.e., they influence the response and the response influences them, we can fit **vector autoregression (VAR)** models to simultaneously account for all of these dependencies.

ARIMAX Models

An ARIMA model can be thought of as "ARIMA plus explanatory variables". Although, in principle, this generalizes to time series with seasonality.

For illustration, we consider a stationary time series $\{Y_t\}$ and demonstrate the $ARMAX(p,q)$ approach:

$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \ldots + \beta_r X_{r,t}$$
$$= \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{k=1}^{r} \beta_k X_{k,t} + \epsilon_t$$

where $X_{k,t}$ represents exogenous variable k ($k = 1, 2, \ldots, r$) at time t. We now have parameters β_k that quantify the relationship between the response time series $\{Y_t\}$ and the exogenous time series $\{X_{k,t}\}$.

Note: we do not interpret these β's in the same way as we do in linear regression, because the deterministic component of the model also depends on the Y's.

If $\{Y_t\}$ is not stationary, we simply difference it (ordinarily and/or seasonally) until it can be modeled by an ARMAX model. Notice that this implicitly requires differencing the exogenous variables as well. Fortunately, R takes care of this automatically with the `xreg` input to the `arima()` function.

In order to forecast an ARIMAX model, you need future values of the exogenous variables, or predicted values of them. The added uncertainty associated with this prediction is not reflected in prediction intervals for $\{Y_t\}$.

Multivariate Examples.R (ARIMAX)

Vector Autoregression (VAR)

In this framework, all variables are treated symmetrically or as if the are endogenous. In this setting, we alter notiation and let $Y_{1,t}$ denote the t^{th} observation of the first variable and, in general, $Y_{k,t}$ denotes the t^{th} observation of the k^{th} variable.

For r endogenous variables, the model has one equation per variable and each equation has a constant and a linear combination of lags of every other variable in the system. The number of lags p is known as the **order** of the model. We write $VAR(p)$:

$$\begin{cases} Y_{1,t} = c_1 + \sum_{i=1}^{p} \phi_{11,i} Y_{1,t-i} + \sum_{i=1}^{p} \phi_{12,i} Y_{2,t-i} + \ldots + \sum_{i=1}^{p} \phi_{1r,i} Y_{r,t-i} + \epsilon_{1,t} \\ Y_{2,t} = c_2 + \sum_{i=1}^{p} \phi_{21,i} Y_{1,t-i} + \sum_{i=1}^{p} \phi_{22,i} Y_{2,t-i} + \ldots + \sum_{i=1}^{p} \phi_{2r,i} Y_{r,t-i} + \epsilon_{2,t} \\ \ldots \\ Y_{r,t} = c_r + \sum_{i=1}^{p} \phi_{r1,i} Y_{1,t-i} + \sum_{i=1}^{p} \phi_{r2,i} Y_{2,t-i} + \ldots + \sum_{i=1}^{p} \phi_{rr,i} Y_{r,t-i} + \epsilon_{r,t} \end{cases}$$

where $\{\epsilon_{k,t}\} \sim WN(0, \sigma_k^2)$ for $k = 1, 2, \ldots, r$. Note that these error processes may be **contemporaneously** correlated.

To illustrate, consider a $VAR(1)$ model with 2 variables:

$$\begin{cases} Y_{1,t} = c_1 + \phi_{11,1} Y_{i,t-1} + \phi_{12,1} Y_{2,t-1} + \epsilon_{1,t} \\ Y_{2,t} = c_2 + \phi_{21,1} Y_{i,t-1} + \phi_{22,1} Y_{2,t-1} + \epsilon_{2,t} \end{cases}$$

- In general, we estimate $r(rp+1)$ parameters, so to avoid overfitting, we like to keep p and r small;
- We choose p in accordance with information criteria.

Multivariate Examples.R (VAR)

12/01/16

ARCH/GARCH

Motivation

Until now, we've dealt with heteroskedasticity via a suitable variance stabilizing transformation. This works well if variability *increases* with time, but perhaps not so well if we have periods of increased variability intersperced with periods of decreased variability. This type of **volatility clustering** is often exhibited in financial/economic time series where periods of **volatility** alternate with periods of **tranquility**. In this setting, we may rather model/explain the heteroskedasticity than remove it with a transformation. This is what we do with ARCH/GARCH models.

With volatility clustering, observations may be uncorrelated themselves, but their magnitudes are correlated, i.e., $Corr(X_t, X_{t+h}) = 0$ but $Corr(|X_t|, |X_{t+h}|) > 0$ or $Corr(X_t^2, X_{t+h}^2) > 0$.

ARCH(l): AutoRegressive Conditional Heteroskedasticity

$\{X_t\} \sim ARCH(l)$ if $X_t = \sigma_t \epsilon_t$ where $\sigma_t^2 = \omega + \alpha_1 X_{t-1}^2 + \alpha_2 X_{t-2}^2 + \ldots + \alpha_l X_{t-l}^2$ and $\{\epsilon_t\} \sim IID(0,1)$.

- $\omega > 0,\ \alpha_i \geq 0,$ and $\epsilon_t \perp X_s,\ s < t.$

Properties

- $E(X_t) = 0$
- $Var(X_t | X_{t-1}, X_{t-2}, \ldots, X_{t-l}) = \sigma_t^2$
- $Cov(X_t, X_{t+h}) = 0$

Proofs

- $E(X_t) = E[E(X_t|X_{t-1}, \ldots, X_{t-l})]$
  $= E[E(\sigma_t\epsilon_t|X_{t-1}, \ldots, X_{t-l})]$
  $= E[\sigma_t E(\epsilon_t|X_{t-1}, \ldots, X_{t-l})] \leftarrow$ since given the history $X_{t-1}, \ldots, X_{t-l}$, σ_t is constant
  $= \sigma_t E(\epsilon_t) \leftarrow$ since $\epsilon_t \perp X_s \ \forall s < t$
  $= \sigma_t \times 0$
  $= 0$
- $Var(X_t|X_{t-1}, X_{t-2}, \ldots, X_{t-l}) = \ldots$
  $= \sigma_t^2$
- $Cov(X_t, X_{t+h}) = \ldots$
  $= 0$

So, thinking of $\{X_t\}$ as an error process, the corresponding residuals may not be correlated themselves, but their variability clearly is dependent upon their history. This is why we check whether the ARCH approach is appropriate by checking ACF plots of $|X_t|$ or X_t^2.

A generalization of the $ARCH(l)$ process is the **generalized ARCH**, $GARCH(k, l)$ process

GARCH(k, l)

$\{X_t\} \sim GARCH(k, l)$ if $X_t = \sigma_t\epsilon_t$ where $\sigma_t^2 = \omega + \alpha_1 X_{t-1}^2 + \alpha_2 X_{t-2}^2 + \ldots + \alpha_l X_{t-l}^2 + \beta_1\sigma_{t-1}^2 + \beta_2\sigma_{t-2}^2 + \ldots + \beta_k\sigma_{t-k}^2 = \omega + \sum_{i=1}^{l} \alpha_i X_{t-i}^2 + \sum_{j=1}^{k} \beta_j\sigma_{t-j}^2$ and $\{\epsilon_t\} \sim IID(0, 1)$.

- $\omega > 0,\ \alpha_i, \beta_j \geq 0,$ and $\epsilon_t \perp X_s,\ s < t.$

We choose the orders k and l by examining the ACF/PACF plots of $|X_t|$ or X_t^2, where l is chosen from the ACF and k from the PACF. In general, small values of k and l are sufficient, with $GARCH(1, 1)$ working well in practice.

Assess model fit using usual diagnostic and goodness-of-fit techniques, i.e., AIC, $\log L(\hat{\theta})$ (LRT), residual diagnostics.

Estimation is carried out via the maximum likelihood approach where the distributional assumption is placed on $\{\epsilon_t\}$. We typically assume $\{\epsilon_t\} \overset{iid}{\sim} N(0, 1)$, but if a QQ-plot of the residuals suggests this assumption is invalid, we can specify other distributions so long as $E(\epsilon_t) = 0$, $Var(\epsilon_t) = 1$.

We can assess model adequacy using residual diagnostics to check whether the residuals behave like the $IID(0, 1)$ noise process:

- Zero mean
- Uncorrelatedness
- Homoskedasticity
- Distributionl assumption is valid

ARCH-GARCH Examples.R

12/06/16

SARIMA with GARCH Errors

SARIMA defines the "mean model", and GARCH defines the "variance model".

$\{X_t\} \sim SARIMA(p,d,q) \times (P,D,Q)_s$ if

$$(1-B)^d(1-B^s)^D\phi^p(B)\Phi^P(B^s)X_t = \theta^q(B)\Theta^Q(B^s)\epsilon_t$$

Previously, we assumed $\{\epsilon_t\} \sim WN(0,\sigma^2)$.

If we fit such a model, but the residuals appear to be heteroskedastic – and in particular **volatility clustering** is evident –, then we can consider modeling the residuals with a GARCH(k,l) process. This corresponds to "SARIMA with GARCH errors".

Mathematically, rather than assuming $\{\epsilon_t\} \sim WN(0,\sigma^2)$, we assume

- $\epsilon_t = \sigma_t e_t$ where $\{e_t\} \sim IID(0,1)$
- $\sigma_t^2 = \omega + \alpha_1\epsilon_{t-1}^2 + \ldots + \alpha_l\epsilon_{t-l}^2 + \beta_1\sigma_{t-1}^2 + \ldots + \beta_k\sigma_{t-k}^2$

where $\omega > 0$, $\alpha_i, \beta_i \geq 0$.

In principle. the mean model can be any Box-Jenkins model(AR, MA, ARMA, ARIMA, SARIMA) and the variance model can be either ARCH or GARCH.

$\rightarrow$ We choose the Box-Jenkins model (i.e., p,q,P,Q,d,D) as we always do, i.e., with ACF/PACF plots of the raw time series.

$\rightarrow$ We choose the orders k,l by examining ACF/PACF plots of |residuals| or (residuals)^2, where the residuals come from the Box-Jenkins fit.

Modeling the residuals in this way does not change the forecasts of the Box-Jenkins model (i.e., the mean model), but it impacts the associated prediction intervals, as these depend on an estimate of σ^2, which is now a function of time.

ARMA GARCH Example.R