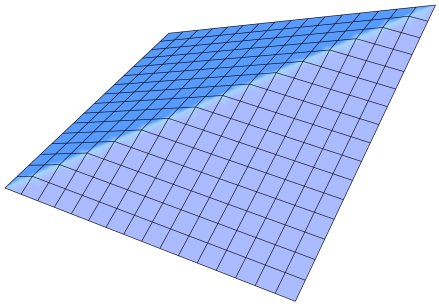


STAT 443: Forecasting

Prof. Reza Remezan • Fall 2014 • University of Waterloo
TR 8:30–9:50AM in RCH 301

Transcribed, LaTeXed, and edited by Rob Zimmerman

Note: These are personal lecture notes, not official course materials. They may contain errors (which by default you should attribute to me, Rob Zimmerman, rather than the course instructor). The permanent home of these — and all of my other lecture notes — is https://rob-zimmerman.github.io/coursenotes.

Contents

Lecture 1 — Tuesday, September 09, 2014

1. Time Series Models

Preliminaries

Definition 1.1 Let T be an index set representing time. A sequence of random variables $\{X_t \mid t \in T\}$ is called a **time series**. If T is discrete, then $\{X_t\}$ is called a **discrete-time time series**; if T is a continuum, then it is called a **continuous-time time series**.

The main focus of the course is on discrete-time models. As a simple motivating example, we might try to model monthly sales data by writing

$$\text{Sales}_t = \beta_0 + \beta_1 t + R_t,$$

where t denotes the month and R_t denotes some kind of random uncertainty. Such a model may be reasonable, but it immediately raises a forecasting question: how do we quantify uncertainty in future values?

Basic Time Series Models

Our interest lies in modelling and analyzing data collected over time. Ideally, given random variables $X_1, X_2, X_3, \ldots$, we would like to specify all joint distributions of the vectors $(X_1, \ldots, X_n)$ for all n; that is,

$$F_n(x_1, \ldots, x_n) := \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

for $-\infty < x_1, \ldots, x_n < \infty$ and $n = 1, 2, 3, \ldots$. In practice, this is usually unrealistic: there is rarely enough information to determine the entire joint distribution. Fortunately, much of the useful information in applications is already captured by the first few moments, such as $\mathbb{E}[X_t]$, $\mathbb{E}\left[X_t^2\right]$, and $\text{Cov}(X_t, X_{t'})$ for $t \neq t'$.

Lecture 2 — Thursday, September 11, 2014

We begin with zero-mean models, then add trend and seasonality.

If the joint distribution is multivariate normal, then the quantities $\mathbb{E}[X_t]$ for all t, $\mathbb{E}\left[X_t^2\right]$ for all t, and $\mathbb{E}[X_t X_{t'}]$ for all t, t' completely determine the distribution. Recall that a p-dimensional **multivariate normal distribution** is written $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the variance-covariance matrix. When $\boldsymbol{\Sigma}$ is positive definite, it has density

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right],$$

where $\boldsymbol{x} = (x_1, \ldots, x_p)$. When $p = 1$, this reduces to the usual univariate normal distribution with density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

Rather than working directly with the full joint distribution, we will mainly study time series models through their means, variances, and covariances.

Definition 2.1 A **time series model** for observed data $\{x_t\}$ is a specification of the joint distributions (or possibly only the means and covariances) of a sequence of random variables $\{X_t\}$, of which $\{x_t\}$ is postulated to be a realization.

In other words, the observed data are treated as one realization of an underlying stochastic process.

Zero-Mean Models

Let us examine some of the most common zero-mean models.

Definition 2.2 A sequence $\{X_t\}$ is called **independent and identically distributed noise** if its terms are independent and identically distributed with $\mathbb{E}[X_t] = 0$.

For independent and identically distributed noise, X_{n+h} is independent of the observed history

$(X_1, \ldots, X_n)$. Consequently,

$$\mathbb{P}(X_{n+h} \leq x \mid X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathbb{P}(X_{n+h} \leq x),$$

whenever the conditioning event has positive probability. The observed past therefore does not help forecast X_{n+h}.

Definition 2.3 Let $S_t = X_1 + X_2 + \cdots + X_t$, where X_t is independent and identically distributed noise for $t = 1, 2, \ldots$. Then $\{S_t\}$ with starting value $S_0 = 0$ is called a **random walk**.

Notice that

$$\mathbb{E}[S_t] = \sum_{i=1}^{t} \mathbb{E}[X_i] = 0,$$

so a random walk is a zero-mean model.

Definition 2.4 A **white noise** is a sequence $\{Z_t\}$ of uncorrelated random variables, each with mean 0 and common variance $\sigma^2 \in (0, \infty)$. We denote this by $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$.

An independent and identically distributed noise with finite variance is thus a white noise. Since

$$\mathrm{Corr}(Z_t, Z_{t'}) = \frac{\mathrm{Cov}(Z_t, Z_{t'})}{\sqrt{\mathrm{Var}(Z_t)\,\mathrm{Var}(Z_{t'})}},$$

uncorrelatedness implies $\mathrm{Cov}(Z_t, Z_{t'}) = 0$. Independence always implies uncorrelatedness, although the converse is false in general.

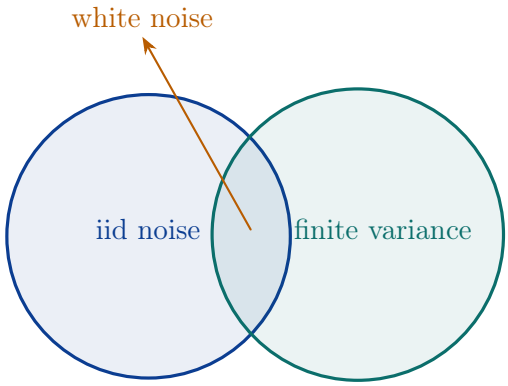


FIGURE 1

Figure 1 summarizes the relation between independent and identically distributed noise, finite variance, and white noise.

Models with Trend and Seasonality

Consider the model $X_t = m_t + Y_t$, where m_t is a slowly varying deterministic function called the **trend** and Y_t has mean zero. Then

$$\mathbb{E}[X_t] = \mathbb{E}[m_t + Y_t] = m_t.$$

Typical examples include a **linear trend** of the form $m_t = \alpha_0 + \alpha_1 t$, and a **quadratic trend** of the form $m_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2$.

More generally, we can consider

$$X_t = m_t + s_t + Y_t,$$

where m_t and Y_t are as above and s_t is periodic with period d, meaning that $s_t = s_{t+d}$ for all t. For example, in discrete time we can take $s_t = \alpha_0 + \alpha_1 \cos(\alpha_2 t)$ whenever $\alpha_2 d$ is an integer multiple of 2π. In these models, the parameters $\alpha_0, \alpha_1, \alpha_2, \ldots$ are usually estimated by maximum likelihood or the method of least squares.

Definition 2.5 A time series model with the representation

$$X_t = m_t + s_t + Y_t,$$

where m_t is a slowly varying deterministic function of t, s_t is a periodic deterministic function of t, and Y_t has mean zero is called a **classical decomposition model**.

Lecture 3 — Tuesday, September 16, 2014

The classical decomposition $X_t = m_t + s_t + Y_t$ separates a deterministic trend m_t, a deterministic seasonal component s_t, and a zero-mean random component Y_t. We can estimate the first two components by regression.

We now discuss how periodic components can be represented in regression models. Since time is discrete in this course, the period of s_t is typically an integer. For example, with monthly data we often expect $s_t = s_{t+12}$. A simple way to represent a seasonal effect is to use indicator variables.

For instance, if

$$s_t = \begin{cases} 1, & t \equiv 1 \pmod{12}, \\ 0, & \text{otherwise}, \end{cases}$$

then the seasonal effect is activated in January and repeats with period 12.

Example 3.1 Suppose we want to model periodic behaviour in quarterly average temperature measurements, so the period is 4. Let

$$x_{jt} := \mathbb{1}\{\text{time } t \text{ is in season } j\}, \qquad j = 1, 2, 3,$$

where seasons 1, 2, and 3 are winter, spring, and summer, respectively; autumn is the reference season. Then the triplet (x_{1t}, x_{2t}, x_{3t}) uniquely determines the season, and

$$s_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} = \begin{cases} \beta_1, & t = 1, 5, 9, \ldots, \\ \beta_2, & t = 2, 6, 10, \ldots, \\ \beta_3, & t = 3, 7, 11, \ldots, \\ 0, & t = 4, 8, 12, \ldots. \end{cases}$$

Hence $s_t = s_{t+4}$. This leads to the regression model

$$X_t = \beta_0 + s_t + Y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + Y_t,$$

where β_0 is the autumn baseline and each β_j is a seasonal contrast with that baseline.

With an intercept and an unrestricted effect for each of d seasonal positions, we use $d - 1$ binary variables and take the omitted position as the reference level.

Model Diagnostics

For a regression model of the form $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{R}$, where $\mathbb{E}[\boldsymbol{R}] = \mathbf{0}$ and $R_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, we must check independence, normality, zero mean, and constant variance.

The **Q-Q plot** checks whether the points lie close to a straight line; if they do, then the normality assumption is reasonable. In R we can generate a reference sample with `a = rnorm(1000)` and plot it using `qqnorm(a)` together with `qqline(a)`.

The **residual-versus-index plot** is the time series analogue of the usual residual diagnostic from ordinary regression. With an intercept in the fitted model, the residuals sum to zero automatically,

so that identity is not itself a diagnostic. What matters is the pattern: residuals should be scattered around zero without a visible trend over time.

To assess constant variance, we can plot the residuals against time or against the fitted values. If the variance is constant, the points should stay within a roughly constant-width band. Fanning out, fanning in, or a diamond-shaped pattern suggests nonconstant variance.

Independence is much harder to assess directly. However, for jointly normal random variables, uncorrelatedness implies independence, so in practice we often combine autocorrelation diagnostics with a normality check.

The built-in `AirPassengers` series makes these modelling choices concrete. It records monthly international airline passenger totals from 1949 through 1960. In the raw series, both the trend and the size of the seasonal swings grow over time; after taking logarithms, the seasonal amplitude is much more nearly constant. This is the visual reason for modelling the logarithm rather than the original counts, as shown in Figure 2.

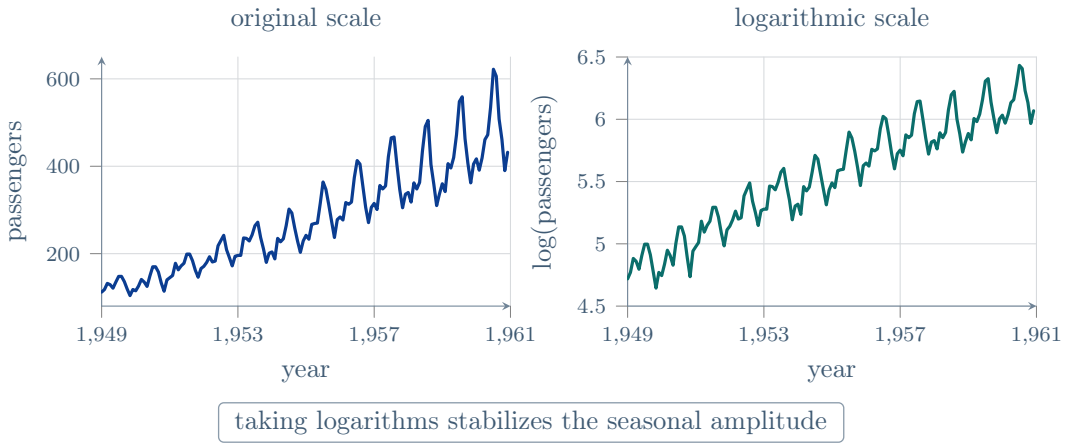


FIGURE 2

We can now fit seasonal regressions with month as a factor. The left panel of Figure 3 overlays linear- and quadratic-trend fits on the logged series. Both explain most of the marginal variation: the adjusted R^2 values are 0.982 and 0.988, respectively. That impressive fit is not the end of the analysis. The residual ACF in the right panel contains sustained positive dependence, including lag-12 correlations of 0.353 for the linear trend and 0.132 for the quadratic trend. The quadratic term helps, but neither regression leaves independent noise.

FIGURE 3

The complete analysis is reproducible with the following base-R code.

```
data(AirPassengers)

passengers <- AirPassengers
time_index <- as.numeric(time(passengers))
month <- factor(cycle(passengers), labels = month.abb)

linear_seasonal <- lm(log(passengers) ~ time_index + month)
quadratic_seasonal <- lm(
  log(passengers) ~ time_index + I(time_index^2) + month
)

summary(linear_seasonal)
summary(quadratic_seasonal)

plot(log(passengers), ylab = "Log passengers")
lines(time_index, fitted(linear_seasonal), col = "steelblue", lwd = 2)
lines(time_index, fitted(quadratic_seasonal), col = "darkorange", lwd = 2)

par(mfrow = c(2, 2))
plot(fitted(linear_seasonal), residuals(linear_seasonal))
qqnorm(residuals(linear_seasonal))
qqline(residuals(linear_seasonal))
plot(residuals(linear_seasonal), type = "l")
acf(residuals(linear_seasonal))
```

Lecture 4 — Thursday, September 18, 2014

The classical decomposition model $X_t = m_t + s_t + Y_t$ can be fitted by representing the seasonal component s_t through regression.

Regression assumes $R_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. The four standard diagnostic plots are therefore central to deciding whether the model is adequate.

In the seasonal regression example above, the seasonal component was represented by 11 binary variables, giving a rather complicated model:

$$\log(Y_t) = \alpha_0 + \alpha_1 t + \beta_1 x_1 + \cdots + \beta_{11} x_{11} + R_t.$$

Even with that model, one of the fundamental regression assumptions failed: the residuals did not behave like an independent and identically distributed $\mathcal{N}(0, \sigma^2)$ sequence. We will return to this example later.

Stationarity and the Autocorrelation Function

Definition 4.1 The time series $\{X_t : t \in T\}$ is called **strictly stationary** (or **strongly stationary**) if the joint distribution of $X_{t_1}, \ldots, X_{t_n}$ is the same as that of $X_{t_1-k}, \ldots, X_{t_n-k}$ for every n and every choice of indices for which both $t_1, \ldots, t_n$ and $t_1 - k, \ldots, t_n - k$ belong to T. In other words, $\{X_t\}$ is strictly stationary if all its statistical properties remain the same under admissible time shifts. We write

$$(X_{t_1}, \ldots, X_{t_n}) \overset{d}{=} (X_{t_1-k}, \ldots, X_{t_n-k}).$$

This is a very strong requirement: every finite-dimensional distribution must remain unchanged under a time shift.

In practice, strict stationarity is often too restrictive. Much of the information we use is already contained in the first two moments: $\mathbb{E}[X_t]$, $\text{Var}(X_t) = \mathbb{E}[(X_t - \mathbb{E}[X_t])^2]$, and $\text{Cov}(X_t, X_{t'}) = \mathbb{E}[X_t X_{t'}] - \mathbb{E}[X_t]\mathbb{E}[X_{t'}]$. This motivates a weaker notion of stationarity based only on these moments.

Definition 4.2 Let $\{X_t\}$ be a time series with $\mathbb{E}[X_t^2] < \infty$ for every t. The **mean function** of $\{X_t\}$ is defined by

$$\mu(t) = \mu_t = \mathbb{E}[X_t].$$

The **covariance function** of $\{X_t\}$ is defined by

$$\gamma_X(r,s) = \text{Cov}(X_r, X_s) = \mathbb{E}[(X_r - \mu_r)(X_s - \mu_s)] \quad \text{for all } r \text{ and } s.$$

If we make the change of index $s = r + h$, then we may also write

$$\gamma_X(r, r+h) = \text{Cov}(X_r, X_{r+h}) = \mathbb{E}[(X_r - \mu_r)(X_{r+h} - \mu_{r+h})] \quad \text{for all } r \text{ and } h.$$

Definition 4.3 The time series $\{X_t\}$ is said to be **weakly stationary** if the following conditions hold:

1. $\mathbb{E}\left[X_t^2\right] < \infty$ for every t.
2. $\mu_X(t) = \mathbb{E}\left[X_t\right]$ does not depend on t.
3. $\gamma_X(t, t+h) = \text{Cov}(X_t, X_{t+h})$ does not depend on t for every lag h.

Example 4.4 Let $\{X_t\}$ satisfy $\mathbb{E}\left[X_t^2\right] < \infty$. Prove that if $\{X_t\}$ is strictly stationary, then it is weakly stationary.

Proof. Assume $\{X_t\}$ is strictly stationary and $\mathbb{E}[X_t^2] < \infty$ for all t. Condition (1) in Definition 4.3 is therefore already satisfied.

For Condition (2), fix t and h. By strict stationarity, the one-dimensional distributions are shift-invariant, so

$$X_t \overset{d}{=} X_{t+h}.$$

Since both variables are integrable, equality in distribution implies equality of means, and hence

$$\mathbb{E}[X_t] = \mathbb{E}[X_{t+h}].$$

Therefore $\mu_X(t) = \mathbb{E}[X_t]$ is constant in t.

For Condition (3), fix a lag h. For any r and t such that $r, r+h, t, t+h \in T$, strict stationarity

with $n = 2$ gives

$$(X_t, X_{t+h}) \stackrel{d}{=} (X_r, X_{r+h}).$$

Because second moments are finite, all mixed moments involved in covariance exist, and equality in distribution yields

$$\text{Cov}(X_t, X_{t+h}) = \text{Cov}(X_r, X_{r+h}),$$

which depends only on h and not on t.

Thus all three conditions in Definition 4.3 hold, so $\{X_t\}$ is weakly stationary. □

Remark 4.5 From now on, whenever we refer to "stationarity" or "stationary" time series, we mean weak stationarity, unless specifically mentioned otherwise.

In view of Condition (3) in Definition 4.3, whenever we refer to the covariance function of a stationary time series $\{X_t\}$, we mean the following function of the single lag h:

$$\gamma_X(h) := \gamma_X(t, t+h) = \gamma_X(t+h, t).$$

The function $\gamma_X(\cdot)$ is very important and has a special name.

Definition 4.6 Let $\{X_t\}$ be a stationary time series. The **autocovariance function** (**ACVF**) at lag h is

$$\gamma_X(h) = \text{Cov}(X_{t+h}, X_t).$$

If $\gamma_X(0) > 0$, the **autocorrelation function** (**ACF**) at lag h is

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \text{Corr}(X_t, X_{t+h}).$$

Tutorial 1 — Thursday, September 18, 2014

We now have two notions of stationarity. Strict stationarity requires

$$(X_{t_1}, \ldots, X_{t_n}) \stackrel{d}{=} (X_{t_1-k}, \ldots, X_{t_n-k}),$$

while weak stationarity requires finite second moments, a mean that does not depend on time, and a covariance structure that depends only on the lag. In that case, $\gamma_X(h)$ is the autocovariance

function and $\rho_X(h) = \gamma_X(h)/\gamma_X(0)$ is the autocorrelation function.

Example 1.1 Investigate the stationarity of white noise. If it is stationary, calculate its ACF.

Solution. Let $\{X_t\} \sim \text{WN}(0, \sigma^2)$ and so $\sigma^2 < \infty$. We check the conditions of Definition 4.3:

1. $\text{Var}(X_t) = \sigma^2 < \infty$.
2. $\mathbb{E}[X_t] = 0$, which is independent of t.
3. $$\text{Cov}(X_t, X_{t+h}) = \begin{cases} \sigma^2, & h = 0 \\ 0, & h \neq 0 \end{cases}$$
   by definition, since white noise is uncorrelated. It is only a function of h, so it does not depend on t.

So Conditions (1)–(3) are satisfied, and so white noise is stationary. Then

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \begin{cases} 1, & h = 0 \\ 0, & h \neq 0 \end{cases}.$$

□

Example 1.2 Investigate the stationarity of a random walk.

Solution. Let $\{X_t\}$ be iid noise with variance $\sigma^2 \in (0, \infty)$, and define $S_t = X_1 + \cdots + X_t$. We know that $\{S_t\}$ with starting point $S_0 = 0$ is a random walk. For $h \geq 0$, observe that

$$\begin{aligned} \text{Cov}(S_{t+h}, S_t) &= \text{Cov}(X_1 + \cdots + X_t + X_{t+1} + \cdots + X_{t+h}, X_1 + \cdots + X_t) \\ &= \text{Cov}(X_1 + \cdots + X_t, X_1 + \cdots + X_t) \\ &= \sum_{i=1}^{t} \text{Var}(X_i) \\ &= t\sigma^2, \end{aligned}$$

which depends on t. Therefore, a random walk is not stationary. □

Example 1.3 Consider the process $X_t = Z_t + \theta Z_{t-1}$, $t = 0, \pm 1, \pm 2, \ldots$ where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. The time series $\{X_t\}$ is called a **first-order moving average process**, denoted MA(1). Show that $\{X_t\}$ is stationary and derive its ACF.

Solution. We check the conditions of Definition 4.3:

1.
$$\begin{aligned}\text{Var}(X_t) &= \text{Var}(Z_t + \theta Z_{t-1}) \\ &= \text{Var}(Z_t) + \theta^2 \,\text{Var}(Z_{t-1}) + 2\theta \,\text{Cov}(Z_t, Z_{t-1}) \\ &= \sigma^2(1 + \theta^2) < \infty.\end{aligned}$$

   Therefore $\mathbb{E}[X_t^2] < \infty$.

2.
$$\mathbb{E}\left[X_t\right] = \mathbb{E}\left[Z_t + \theta Z_{t-1}\right] = \mathbb{E}\left[Z_t\right] + \theta \mathbb{E}\left[Z_{t-1}\right] = 0 + \theta(0) = 0,$$

   which is independent of t.

3. We calculate the covariance function directly:

$$\begin{aligned}\gamma_X(h) &= \text{Cov}(X_t, X_{t+h}) \\ &= \text{Cov}(Z_t + \theta Z_{t-1},\, Z_{t+h} + \theta Z_{t+h-1}) \\ &= \text{Cov}(Z_t, Z_{t+h}) + \theta \,\text{Cov}(Z_t, Z_{t+h-1}) + \theta \,\text{Cov}(Z_{t-1}, Z_{t+h}) + \theta^2 \,\text{Cov}(Z_{t-1}, Z_{t+h-1}).\end{aligned}$$

   Since $\{Z_t\}$ is white noise, $\text{Cov}(Z_i, Z_j) = 0$ for $i \neq j$ and $\text{Cov}(Z_i, Z_i) = \sigma^2$. Hence:

$$\begin{aligned}\gamma_X(0) &= \text{Cov}(Z_t, Z_t) + \theta^2 \,\text{Cov}(Z_{t-1}, Z_{t-1}) = \sigma^2 + \theta^2\sigma^2 = \sigma^2(1 + \theta^2), \\ \gamma_X(1) &= \theta \,\text{Cov}(Z_t, Z_t) = \theta\sigma^2, \\ \gamma_X(-1) &= \theta \,\text{Cov}(Z_{t-1}, Z_{t-1}) = \theta\sigma^2, \\ \gamma_X(h) &= 0, \quad |h| \geq 2.\end{aligned}$$

   Therefore,
$$\text{Cov}(X_t, X_{t+h}) = \gamma_X(h) = \begin{cases} \sigma^2(1 + \theta^2), & h = 0, \\ \theta\sigma^2, & h = \pm 1, \\ 0, & h \neq 0, \pm 1, \end{cases}$$

   and this depends only on h, not on t.

Conditions (1)–(3) imply that X_t is stationary. Our ACF is

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \begin{cases} 1, & h = 0 \\ \dfrac{\theta}{1+\theta^2}, & h = \pm 1 \\ 0, & h \neq 0, \pm 1 \end{cases}.$$

□

We will return to moving average processes later in the course.

Remark 1.4 Observe that covariance is a symmetric operator, which implies that $\gamma_X(h)$ is an even function. This is always true. For this reason, we will calculate $\gamma_X(h)$ only for $h \geq 0$. Also note that $\rho_X(h) \in [-1, 1]$.

At this point, it is worth collecting the basic properties of $\mathbb{E}[X]$, $\mathrm{Var}(X)$, and $\mathrm{Cov}(X, Y)$ that we will use repeatedly.

Proposition 1.5 Whenever the displayed expectations, variances, and covariances exist, the following hold:

1. $$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i\right].$$
2. $\mathbb{E}\left[aX\right] = a\mathbb{E}\left[X\right]$ for $a \in \mathbb{R}$.
3. $$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathbb{E}\left[X_i\right].$$
4. $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$.
5. $\operatorname{Var}(a_1 X_1 \pm a_2 X_2) = a_1^2 \operatorname{Var}(X_1) + a_2^2 \operatorname{Var}(X_2) \pm 2a_1 a_2 \operatorname{Cov}(X_1, X_2)$.
6. $$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \operatorname{Cov}(X_i, X_j).$$
7. $\operatorname{Cov}(a_1 X_1 + b_1, a_2 X_2 + b_2) = \operatorname{Cov}(a_1 X_1, a_2 X_2) = a_1 a_2 \operatorname{Cov}(X_1, X_2)$.
8. $\operatorname{Cov}(X_1 + X_2, X_3 + X_4) = \operatorname{Cov}(X_1, X_3) + \operatorname{Cov}(X_1, X_4) + \operatorname{Cov}(X_2, X_3) + \operatorname{Cov}(X_2, X_4)$.

Lecture 5 — Tuesday, September 23, 2014

Recall that for a stationary series,

$$\gamma_X(h) = \operatorname{Cov}(X_t, X_{t-h}) = \operatorname{Cov}(X_t, X_{t+h}) = \mathbb{E}\left[X_t X_{t-h}\right] - \mathbb{E}\left[X_t\right]\mathbb{E}\left[X_{t-h}\right].$$

Example 5.1 Let $\{X_t\}$ be a stationary time series satisfying

$$X_t = \phi X_{t-1} + Z_t, \quad t \in \mathbb{Z},$$

where $|\phi| < 1$ and $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. Also assume that Z_t and X_s are uncorrelated for all $s < t$. This time series is called a **first-order autoregressive model**, denoted AR(1). Derive its autocorrelation function.

Solution. To begin with,

$$\mathbb{E}[X_t] = \mathbb{E}[\phi X_{t-1} + Z_t] = \phi\mathbb{E}[X_{t-1}] + \mathbb{E}[Z_t] = \phi\mathbb{E}[X_{t-1}].$$

Since X_t is stationary, $\mathbb{E}[X_t] = \mu$. So $\mu = \phi\mu$, and therefore $\mu = 0$, so we are dealing with a zero-mean model. Now, let $h > 0$. Multiply both sides of $X_t = \phi X_{t-1} + Z_t$ by X_{t-h} and take expectations to get

$$\mathbb{E}[X_t X_{t-h}] = \mathbb{E}[\phi X_{t-1} X_{t-h}] + \mathbb{E}[Z_t X_{t-h}] \implies \gamma_X(h) = \phi\gamma_X(h-1), \quad h > 0.$$

When $h = 1$, $\gamma_X(1) = \phi\gamma_X(0)$. When $h = 2$, $\gamma_X(2) = \phi\gamma_X(1) = \phi^2\gamma_X(0)$. In general, by induction, we know that $\gamma_X(h) = \phi^h\gamma_X(0)$ for $h \geq 0$. Since $\gamma_X(h) = \gamma_X(-h)$, we have $\gamma_X(h) = \phi^{|h|}\gamma_X(0)$ for all h. To find $\gamma_X(0)$, note

$$\mathrm{Var}(X_t) = \mathrm{Var}(\phi X_{t-1} + Z_t) = \phi^2\,\mathrm{Var}(X_{t-1}) + \mathrm{Var}(Z_t) = \phi^2\gamma_X(0) + \sigma^2.$$

Thus,

$$\gamma_X(0) = \frac{\sigma^2}{1-\phi^2}.$$

The ACF is

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \phi^{|h|} \quad \text{for } h = 0, \pm 1, \pm 2, \ldots.$$

□

We will also return to autoregressive models later in the course.

It is important to observe that for stationary time series,

$$\gamma_X(0) = \mathrm{Cov}(X_t, X_t) = \mathrm{Var}(X_t)$$

and

$$\rho_X(h) = \mathrm{Corr}(X_t, X_{t+h}).$$

Recall the MA(1) and AR(1) processes. For the MA(1) process $X_t = Z_t + \theta Z_{t-1}$, we had

$$\rho_X(h) = \begin{cases} 1, & h = 0 \\ \dfrac{\theta}{1+\theta^2}, & |h| = 1 \\ 0, & \text{otherwise} \end{cases}.$$

For the AR(1) process $X_t = \phi X_{t-1} + Z_t$ with $|\phi| < 1$, we had $\rho_X(h) = \phi^{|h|}$.

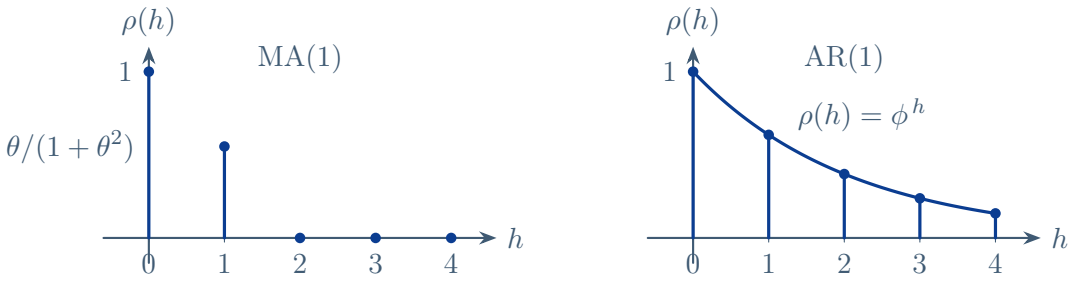


FIGURE 4

Figure 4 contrasts the finite cutoff of the MA(1) autocorrelation function with the geometric decay of the AR(1) autocorrelation function.

Example 5.2 Show that $X_t = 5t + 3 + Z_t$ is not stationary, where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$.

Solution. Since

$$\mathbb{E}[X_t] = 5t + 3,$$

the mean depends on t. Therefore $\{X_t\}$ is not stationary. □

Example 5.2 shows that regression models with a time-varying mean are not stationary.

The Sample Autocorrelation Function

What we have seen so far about the ACF is based on theoretical models. In practical problems, we do not start with a model. Instead, we have the observed data $\{x_1, x_2, \ldots, x_n\}$. We will use the sample ACF to study the dependency structure of the data.

Definition 5.3 Let $x_1, \ldots, x_n$ be observations of a time series. The **sample mean** of $x_1, \ldots, x_n$ is

$$\bar{x} := \frac{1}{n} \sum_{t=1}^{n} x_t.$$

The **sample autocovariance function** is

$$\hat{\gamma}(h) := \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x}), \qquad h = -(n-1), \ldots, n-1.$$

The **sample autocorrelation function** is

$$\hat{\rho}(h) := \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad h = -(n-1), \ldots, n-1.$$

Lecture 6 — Thursday, September 25, 2014

The sample autocorrelation function is

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \frac{\displaystyle\sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x})}{\displaystyle\sum_{t=1}^{n} (x_t - \bar{x})^2}, \quad h = -(n-1), \ldots, n-1.$$

Our convention is that $\hat{\gamma}(h)$ denotes the *numerical estimate* computed from the observed data x_t. The corresponding estimator is

$$\tilde{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (X_{t+|h|} - \bar{X})(X_t - \bar{X}).$$

$\tilde{\gamma}$ is an *estimator*, and hence a random variable. Throughout the course we use a tilde for estimators and a hat for realized estimates.

The sample ACF measures linear dependence among observations, so it is also useful for checking whether regression residuals are uncorrelated. For iid noise with a finite fourth moment and each fixed nonzero lag h, we have the large-sample approximation $\sqrt{n}\,\hat{\rho}(h) \xrightarrow{d} \mathcal{N}(0,1)$. Consequently, the usual pointwise 95% reference interval for a particular nonzero-lag autocorrelation is

approximately $\pm 1.96/\sqrt{n}$:

$$\mathbb{P}\left(\frac{-1.96}{\sqrt{n}} < \hat{\rho}(h) < \frac{1.96}{\sqrt{n}}\right) \approx 0.95.$$

This gives the usual reference bands on a sample ACF plot. They are pointwise rather than simultaneous, so several lags considered together can produce an isolated exceedance by chance.

For observed data $\{x_1, \ldots, x_n\}$, the shape of the sample ACF often gives immediate qualitative information. For example, if the data contain a trend, then $|\hat{\rho}(h)|$ typically decays slowly as h increases. If the data contain a substantial deterministic periodic component, then $\hat{\rho}(h)$ often shows the same periodicity. Here "substantial" means that the deterministic component dominates the random one, as in a model such as $X_t = \sin(2\pi t/12) + Y_t$ with $Y_t \sim \mathcal{N}(0, 1/100)$. These two patterns are compared in Figure 5.

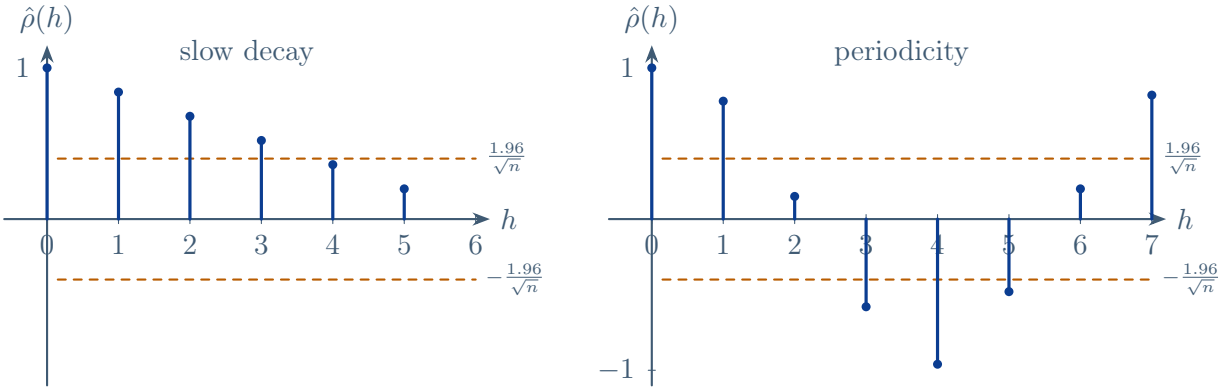


FIGURE 5

Values that remain within the reference bands are usually treated as negligible sample correlations. A few isolated exceedances can occur by chance; a persistent pattern is more informative than a single spike. In particular, periodic behaviour in the ACF is evidence of cyclical dependence, but it does not by itself prove that the mean is periodic or that the series is nonstationary: a stationary process can also have an oscillating ACF.

The built-in `USAccDeaths` series provides a less schematic example. It contains monthly accidental-death totals in the United States from 1973 through 1978. The annual rises and falls visible in the left panel of Figure 6 reappear as large correlations at seasonal lags in the right panel. In particular, the lag-12 sample autocorrelation is 0.629. This agreement between the time plot and the ACF is much stronger evidence of seasonality than either display would provide on its own.

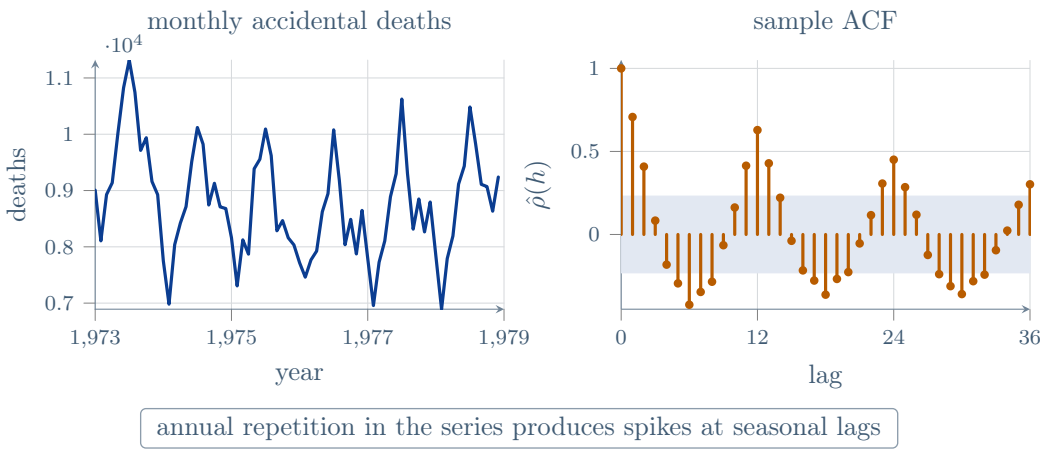


FIGURE 6

The plot and its sample ACF can be reproduced as follows.

```
data(USAccDeaths)

plot(
  USAccDeaths,
  ylab = "Deaths",
  main = "Monthly accidental deaths in the United States"
)
acf(
  USAccDeaths,
  lag.max = 36,
  main = "Sample ACF of USAccDeaths"
)
```

Lecture 7 — Tuesday, September 30, 2014

Linear Regression for Time Series

Stationarity and zero correlation are different ideas. A slowly decaying sample ACF is typical of trend, whereas periodic spikes indicate periodic structure.

We now review the basic regression framework. A model is linear if it is linear in the coefficients. Thus $Y = \beta_0 + \beta_1 x^2 + \varepsilon$ is linear, as is $Y = \beta_0 + \beta_1 x + \varepsilon$ and $Y = \beta_0 + \beta_1 \log(x) + \beta_2 x^2 + \varepsilon$. By contrast, $Y = \beta_0 e^{\beta_1 x} + \varepsilon$ is not linear.

In **simple linear regression** with fixed covariate values $x_1, \ldots, x_n$,

$$Y_i = \alpha + \beta x_i + R_i,$$

where $R_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Here Y is the **response variable** and X is the **explanatory variable**. We use β for the parameter, $\tilde{\beta}$ for the estimator, and $\hat{\beta}$ for the realized estimate. As soon as a line has been fitted, the assumptions $R_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ must be checked.

Suppose a fitted regression has residuals with a conspicuously large lag-1 sample autocorrelation. Then $\text{Cov}(r_t, r_{t-1}) \neq 0$ is a plausible concern, so the residuals should not be treated as noise; an MA(1)-type structure would be a more reasonable possibility.

The essential columns in a simple linear regression summary have the following form.

Coefficient	Estimate	Standard error	Null hypothesis
Intercept	$\hat{\alpha}$	$\widehat{\text{se}}(\hat{\alpha})$	$H_0 : \alpha = 0$
Slope	$\hat{\beta}$	$\widehat{\text{se}}(\hat{\beta})$	$H_0 : \beta = 0$

The reported t statistic is the estimate divided by its standard error, and the associated p-value measures evidence against the corresponding null hypothesis.

When plotting fitted values against observed values, the points should ideally fall near the line $y = x$. Note that the usual p-values rely on the assumption $R_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

As a rule of thumb, one or two unusual data points should not, by themselves, force a change in

the model.

For multiple linear regression, we write $\boldsymbol{y} \mid \mathbf{X} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{R}$. In coordinate form,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + R_i.$$

If $\mathbf{X}$ has full column rank, the least squares estimator is obtained from matrix calculus as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$$

and the fitted response variables are

$$\hat{y}_i = x_i^T\hat{\boldsymbol{\beta}} \implies \hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$$

High correlation among explanatory variables indicates redundancy and is a warning sign for multicollinearity.

Sums of Squares Proofs

For simple linear regression, write

$$Y_i = \alpha + \beta x_i + R_i, \quad i = 1, \ldots, n,$$

and define

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad \text{and} \qquad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Assume throughout this subsection that $S_{xx} > 0$, so the covariate values are not all equal.

Proposition 7.1 (Least squares estimators) The least squares estimators in simple linear regression are

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \qquad \text{and} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}.$$

Proof. We want to minimize

$$Q(a, b) = \sum_{i=1}^{n}(y_i - a - bx_i)^2.$$

Setting partial derivatives to zero gives the normal equations

$$\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0.$$

The first equation gives $a = \bar{y} - b\bar{x}$. Substituting into the second:

$$\sum_{i=1}^{n} x_i[(y_i - \bar{y}) - b(x_i - \bar{x})] = 0.$$

Since $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$, this simplifies to

$$b\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

so $b = S_{xy}/S_{xx}$. Therefore $\hat{\beta} = S_{xy}/S_{xx}$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$. □

Proposition 7.2 (Variances and covariance of $\hat{\alpha}, \hat{\beta}$) If $R_1, \ldots, R_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, then

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2}{S_{xx}}, \qquad \text{Var}(\hat{\alpha}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right), \qquad \text{and} \qquad \text{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\bar{x}\sigma^2}{S_{xx}}.$$

Proof. Because

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{S_{xx}} = \sum_{i=1}^{n} c_i Y_i \qquad \text{and} \qquad c_i = \frac{x_i - \bar{x}}{S_{xx}},$$

we have

$$\text{Var}(\hat{\beta}) = \sum_{i=1}^{n} c_i^2 \,\text{Var}(Y_i) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{S_{xx}}.$$

Also $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$, so

$$\text{Var}(\hat{\alpha}) = \text{Var}(\bar{Y}) + \bar{x}^2 \,\text{Var}(\hat{\beta}) - 2\bar{x}\,\text{Cov}(\bar{Y}, \hat{\beta}).$$

Now

$$\text{Cov}(\bar{Y}, \hat{\beta}) = \text{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \frac{1}{S_{xx}}\sum_{j=1}^{n}(x_j - \bar{x})Y_j\right) = \frac{\sigma^2}{nS_{xx}}\sum_{j=1}^{n}(x_j - \bar{x}) = 0.$$

Hence

$$\mathrm{Var}(\hat{\alpha}) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right).$$

Finally,

$$\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = \mathrm{Cov}(\bar{Y} - \bar{x}\hat{\beta}, \hat{\beta}) = \mathrm{Cov}(\bar{Y}, \hat{\beta}) - \bar{x}\,\mathrm{Var}(\hat{\beta}) = -\frac{\bar{x}\sigma^2}{S_{xx}}.$$

□

Proposition 7.3 (Sums-of-squares identities) With $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$,

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \frac{S_{xy}^2}{S_{xx}},$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}},$$

and therefore

$$S_{yy} = SSR + SSE.$$

Proof. Since $\hat{y}_i - \bar{y} = \hat{\beta}(x_i - \bar{x})$,

$$SSR = \sum_{i=1}^{n} \hat{\beta}^2 (x_i - \bar{x})^2 = \hat{\beta}^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}.$$

Also

$$y_i - \hat{y}_i = (y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x}),$$

so

$$SSE = \sum_{i=1}^{n} \Big[(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x}) \Big]^2 = S_{yy} - 2\hat{\beta} S_{xy} + \hat{\beta}^2 S_{xx} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.$$

Combining the two expressions yields $S_{yy} = SSR + SSE$. □

Lecture 8 — Thursday, October 02, 2014

In the regression model $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{R}$, with $R_1, \ldots, R_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, the least squares estimate is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{y}$. An ACF plot helps us assess whether the residuals are uncorrelated. Independence implies uncorrelatedness, and for jointly Gaussian variables the converse also holds.

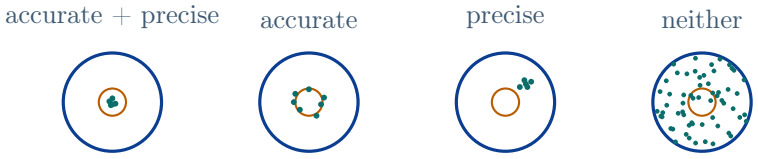


FIGURE 7

Figure 7 gives the usual schematic distinction between accuracy and precision. An estimator is **unbiased** when $\mathbb{E}[\tilde{\theta}] = \theta$, so repeated estimates are centered on the target; it is **precise** when its variance is small. Accuracy is broader and concerns closeness to the target, so it reflects both bias and variability. In the classical regression model, $\boldsymbol{R} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \text{Var}(\boldsymbol{R}) = \sigma^2 \mathbf{I}_{n\times n}$. If the design matrix contains an intercept, the fitted residuals sum to zero, and hence

$$\sum_{i=1}^{n} (\hat{r}_i - \bar{r})^2 = \sum_{i=1}^{n} \hat{r}_i^2 = SSE,$$

so $SSE/(n-p-1)$ estimates σ^2. As usual, σ^2 is the parameter, $\tilde{\sigma}^2$ is the estimator, and $\hat{\sigma}^2$ is its observed value. The least squares procedure minimizes the observed residual sum of squares

$$\sum_{i=1}^{n} \hat{r}_i^2,$$

and hence produces the least squares estimator.

The **likelihood function** $L(\beta_0, \ldots, \beta_p, \sigma^2)$ treats the observed data as fixed and the parameters as variable. Under the model $\boldsymbol{R} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_{n\times n})$, the maximum likelihood estimator of $\boldsymbol{\beta}$ coincides with the least squares estimator; that is, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1}\mathbf{X}^T \boldsymbol{y}$. The maximum likelihood estimator of σ^2 is SSE/n, whereas the unbiased estimator used for classical inference is $\hat{\sigma}^2 = SSE/(n-p-1)$. For the latter,

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1},$$

so a 95% confidence interval for σ^2 can be obtained from the appropriate χ^2 quantiles.

With

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \text{and} \qquad SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,$$

the identity

$$S_{yy} = SSR + SSE$$

decomposes the total variability in $\boldsymbol{y}$ into variability explained by regression and variability attributed to error. Consequently,

$$\frac{SSR}{S_{yy}} = 1 - \frac{SSE}{S_{yy}} = \text{Multiple } R^2,$$

which is the **coefficient of determination**. In simple linear regression, R^2 is equal to the square of the correlation coefficient. Confidence intervals for σ^2 and $\boldsymbol{\beta}$ rely on the assumption $R_1, \ldots, R_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, so that assumption must be checked before carrying out inference.

Once the residual assumptions are judged reasonable, we test the overall significance of the regression model through the hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0,$$

using the statistic

$$F = \frac{SSR/p}{SSE/(n-p-1)}.$$

For a hypothetical data set of 25 shipments, suppose we regress shipment volume on price and price squared. If the fitted model gives $\hat{\sigma} = 2.709$, then the residual degrees of freedom are $22 = 25 - 2 - 1$; suppose further that the overall statistic is

$$96.1 = \frac{SSR/p}{SSE/(n-p-1)}.$$

Our working convention is that if the p-value is below 0.05, then we reject H_0 and conclude that at least one slope is nonzero.

Warning It is important to remember that H_0 concerns $\beta_1, \ldots, \beta_p$ and does not include β_0.

Lecture 9 — Tuesday, October 07, 2014

Under the Gaussian linear model, the maximum likelihood estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{y}$, confidence intervals for σ^2 follow from the chi-squared distribution, and $S_{yy} = SSR + SSE$. Overall regression significance is tested through

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0,$$

together with the F-test and the statistic R^2.

Suppose y is the AAPL price and x_1, x_2 are the contemporaneous prices of MSFT and BB. A model of the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ is not useful for forecasting one week ahead because the future values of x_1 and x_2 are themselves unknown. For time series forecasting, it is more natural to model y_t directly as a function of time; for example, $y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \varepsilon_t$.

The **centered form** of the model is

$$y_i = \alpha + \beta(x_i - \bar{x}) + R_i.$$

Here $\tilde{\alpha} = \bar{Y}$, and we have $\tilde{\alpha} \sim \mathcal{N}(\alpha, \sigma^2/n)$ together with $\tilde{\beta} \sim \mathcal{N}(\beta, \sigma^2/S_{xx})$.

Remember that β is the expected change in the average of y when x changes by one unit, with all other explanatory variables held fixed. Why should we centralize? β_0 is supposed to be the intercept when $x = 0$ but often that has no real interpretation.

In the centered parametrization, the intercept estimate is $\hat{\alpha} = \bar{y}$. The estimated mean response at a new covariate value is

$$\tilde{\mu}(x_{\text{new}}) = \tilde{\alpha} + \tilde{\beta}(x_{\text{new}} - \bar{x}) \sim \mathcal{N}\left(\alpha + \beta(x_{\text{new}} - \bar{x}), \sigma^2\left(\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}\right)\right).$$

Our goal is a prediction interval for y_{new}. This is derived in the same way as an ordinary confidence interval, but with the additional variability of the future observation included.

We know that for the model

$$y_i = \beta_0 + \beta_1 x_i + r_i$$

with iid errors $r_i \sim \mathcal{N}(0, \sigma^2)$, we have

$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

If the new error is an independent draw from the same distribution, then $y_{\text{new}} \sim \mathcal{N}(\beta_0+\beta_1 x_{\text{new}}, \sigma^2)$ and y_{new} is independent of $y_1, \ldots, y_n$; therefore, y_{new} is independent of any function of $y_1, \ldots, y_n$, including $\tilde{\alpha}$ and $\tilde{\beta}$. In particular,

$$\mathbb{E}\left[y_{\text{new}} - \tilde{\mu}(x_{\text{new}})\right] = \alpha + \beta(x_{\text{new}} - \bar{x}) - (\alpha + \beta(x_{\text{new}} - \bar{x})) = 0$$

and

$$\text{Var}(y_{\text{new}} - \tilde{\mu}(x_{\text{new}})) = \text{Var}(y_{\text{new}}) + \text{Var}(\tilde{\mu}(x_{\text{new}})) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}\right),$$

so

$$y_{\text{new}} - \tilde{\mu}(x_{\text{new}}) \sim \mathcal{N}\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}\right)\right).$$

Standardizing, we obtain

$$\frac{y_{\text{new}} - \tilde{\mu}(x_{\text{new}})}{\sigma\sqrt{1 + 1/n + \frac{(x_{\text{new}}-\bar{x})^2}{S_{xx}}}} \sim \mathcal{N}(0, 1).$$

Of course σ is rarely known. Replacing it with $\hat{\sigma}$ changes the standardized prediction error from a standard Normal variable to a Student t variable:

$$\frac{y_{\text{new}} - \tilde{\mu}(x_{\text{new}})}{\hat{\sigma}\sqrt{1 + 1/n + (x_{\text{new}} - \bar{x})^2/S_{xx}}} \sim t_{n-2}.$$

The resulting 95% prediction interval is

$$\hat{\mu}(x_{\text{new}}) \pm t_{n-2,0.975}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{xx}}}.$$

In the case of multiple linear regression with $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{R}$, if we have a new explanatory vector $\boldsymbol{x}_{\text{new}}$, the point forecast is

$$\hat{\mu} = \boldsymbol{x}_{\text{new}}^\top \hat{\boldsymbol{\beta}}$$

and the corresponding prediction interval is

$$\left(\hat{\mu} - c\hat{\sigma}\sqrt{1 + \boldsymbol{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{x}_{\text{new}}}, \hat{\mu} + c\hat{\sigma}\sqrt{1 + \boldsymbol{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{x}_{\text{new}}}\right),$$

where c is the appropriate quantile of the t_{n-p-1} distribution.

2. Model Selection and Diagnostics

Model Selection Criteria

If several models all appear to satisfy the required assumptions, we still need criteria for choosing among them.

$R^2 = 1 - SSE/S_{yy}$ measures the proportion of the variability in the response variable y that is explained by the regression model. For example, if $R^2 = 0.86$, then 86% of the variability is explained by the model and the remaining 14% is attributed to error. However, adding arbitrary explanatory variables tends to increase R^2, even when the added terms are not genuinely useful.

Overfitting occurs when we extract too much apparent structure from the data and begin fitting the error itself. A model that interpolates the observed data perfectly, such as an extremely high-degree polynomial, can perform very poorly for prediction. Figure 8 contrasts this behaviour with a simpler linear fit.

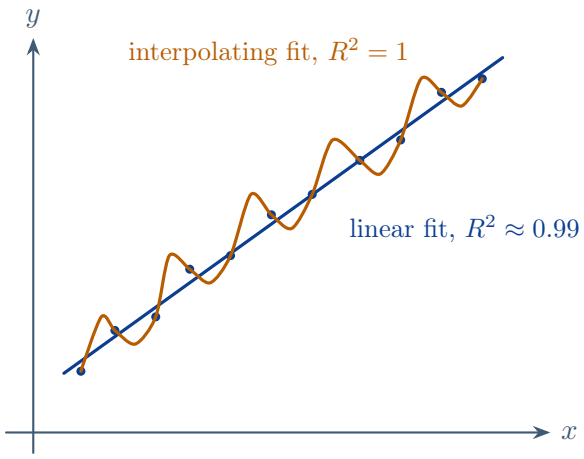


FIGURE 8

Lecture 10 — Thursday, October 09, 2014

Previously, we saw that the coefficient of determination

$$R^2 = \frac{SSR}{S_{yy}} = 1 - \frac{SSE}{S_{yy}}$$

measures the squared correlation between the fitted and observed values. It is also the ratio of the variation in the data captured by the model to the total variation. We noted that R^2 never decreases when additional parameters are added, which is why it cannot be the sole model selection criterion for linear regression. Adding explanatory variates does not always improve forecasting; it can merely add noise.

The estimated standard deviation of a new prediction error,

$$\hat{\sigma}\sqrt{1 + \boldsymbol{x}_{\text{new}}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{x}_{\text{new}}}$$

depends on both the residual variance estimate $\hat{\sigma}^2$ and the uncertainty in the fitted mean, represented by $\boldsymbol{x}_{\text{new}}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{x}_{\text{new}}$. Adding variables never increases SSE, but it reduces the residual degrees of freedom and introduces additional coefficients to estimate; consequently, neither $\hat{\sigma}^2$ nor the prediction standard error is guaranteed to decrease. Thus maximizing R^2 does not necessarily minimize forecasting error.

To determine whether or not it is worthwhile to add an explanatory variate into the model, we can take into account the number of parameters involved. This leads to the **adjusted R-squared** statistic

$$\bar{R}^2 := 1 - (1 - R^2)\frac{n-1}{n-p-1},$$

which penalizes unnecessary explanatory variables, so it can decrease when a weak variable is added. As with R^2, larger values are preferred.

Another important statistic similarly balances goodness of fit (via the likelihood function) with the number of parameters in the model:

Definition 10.1 The **Akaike information criterion** (**AIC**) is defined as

$$\text{AIC} = -2\ell(\hat{\theta}) + 2N_p,$$

where $N_p = p + 2$ counts the intercept, the p slopes, and σ^2; $\theta = (\beta_0, \ldots, \beta_p, \sigma^2)$; and $\ell(\theta)$ is the log-likelihood.

Smaller values of the AIC are preferred. As we add parameters, $-2\ell(\hat{\theta})$ decreases, but we incur a penalty of $2N_p$ for adding parameters which do not improve the fit well enough.

There are other criteria which penalize the log-likelihood function for the number of parameters in different ways, including the corrected AIC and the BIC.

Definition 10.2 The **corrected AIC** (**AICC**) is defined as

$$\text{AICC} = -2\ell(\hat{\theta}) + 2N_p + \frac{2N_p(N_p+1)}{n - N_p - 1}.$$

The AICC penalizes extra parameters more strongly than the AIC, so it is often preferable when the sample size is small.

Definition 10.3 The **Bayesian information criterion** (**BIC**) is defined as

$$BIC = -2\ell(\hat{\theta}) + N_p \log(n).$$

For the sample sizes usually encountered in practice, the BIC penalizes additional parameters more heavily than the AIC and therefore tends to favour simpler models.

If we have the luxury of storing away a portion of the available data to use as a validation dataset (i.e., we do not use it to train the model), then we can compute a validation error.

Definition 10.4 The **validation sum of squares** is defined as

$$\text{VSSE} = \sum_{i \in C} (y_i - \hat{y}_i)^2,$$

where C indexes data points *not* involved in building the model.

For example, if $y_1, y_2, \ldots, y_{100}$ are the available data, we might fit several candidate models on $y_1, \ldots, y_{90}$, and for each candidate model, we produce the forecasts for the validation dataset

$\hat{y}_{91}, \ldots, \hat{y}_{100}$. The validation sums of squares then take the form

$$\sum_{i=91}^{100} (y_i - \hat{y}_i)^2.$$

The model with the smaller validation sum of squares performs better on this particular holdout set (although not necessarily on future data). The term **PRESS** is usually reserved for the leave-one-out sum $\sum_{i=1}^{n} (y_i - \hat{y}_{i,(-i)})^2$, where $\hat{y}_{i,(-i)}$ is fitted without observation i.

In general, there may be many potential covariates which could be used to build a forecasting model. Including every covariate does not result in the best forecasting model; it simply adds noise.

In practice, we usually cannot evaluate all possible subsets of covariates (i.e., **all subsets regression**) when the candidate set is large, so we rely on structured search strategies. In **forward selection**, we start from a small model (often intercept-only) and add the covariate that gives the largest improvement according to a criterion such as AIC, BIC, adjusted R^2, or validation error, repeating until no meaningful gain remains. In **backward elimination**, we start from a larger model and remove the least useful covariate at each step according to the same criterion. In **stepwise regression**, both directions are allowed, so variables can be added and later removed if their contribution changes after other covariates enter the model. Other practical strategies include blockwise or hierarchical selection based on subject-matter structure, and regularization methods such as ridge regression, lasso, or elastic net, which perform selection or shrinkage while controlling overfitting.

Tutorial 2 — Thursday, October 09, 2014

Residual Diagnostic Tests

There are several standard tests for residual diagnostics. The difference sign test and the runs test both use the null hypothesis that the residuals are random, while the Shapiro–Wilk test uses the null hypothesis that the residuals follow a Gaussian distribution. In these tests, we are not testing for independence directly; rather, if the data are independent, they should pass these tests.

For the **Shapiro–Wilk test**, the null hypothesis is that $Y_1, \ldots, Y_n$ are iid observations from a Gaussian distribution. We reject when the p-value is small. In R, this is implemented by `shapiro.test(y)`. For example, a p-value of 0.6877 is large and therefore does not provide evidence

against Gaussianity; it does not establish that the data are Gaussian.

The **difference sign test** is motivated by the fact that independence is stronger than mere randomness. Let S count the indices for which $y_i < y_{i+1}$. An unusually large or small count suggests a systematic trend. For a long iid sequence from a continuous distribution,

$$\mu_S = \mathbb{E}[S] = (n-1)/2 \qquad \text{and} \qquad \sigma_S^2 = \mathrm{Var}(S) = (n+1)/12,$$

so we use the Normal approximation $S \approx \mathcal{N}(\mu_S, \sigma_S^2)$. Equivalently,

$$S' = \frac{S - \mu_S}{\sqrt{\sigma_S^2}} \approx \mathcal{N}(0,1).$$

A large positive (negative) value of $S - \mu_S$ indicates an increasing (decreasing) trend. We reject the null hypothesis that the data are random if $|S'| > z_{1-\alpha/2}$. The test is limited, however: quadratic trends or seasonal patterns can evade it, which is one reason to supplement it with other diagnostics.

Example 2.1 Consider the data

$$4.53,\ 4.21,\ 6.71,\ 7.56,\ 6.31,\ 9.74,\ 10.60,\ 8.92,\ 11.00,\ 13.60.$$

For the difference sign test, let S be the number of indices i such that $y_i < y_{i+1}$. Here $S = 6$. Since $n = 10$, we have

$$\mu_S = \mathbb{E}[S] = \frac{10-1}{2} = 4.5 \qquad \text{and} \qquad \sigma_S^2 = \frac{10+1}{12} \approx 0.92.$$

Therefore,

$$S' = \frac{S - \mu_S}{\sqrt{\sigma_S^2}} = \frac{6 - 4.5}{\sqrt{0.92}} \approx 1.57.$$

At significance level $\alpha = 0.05$, we compare with $z_{0.975} = 1.96$. Since $|1.57| < 1.96$, we fail to reject H_0 and conclude there is no evidence against randomness in this dataset.

The **runs test** is based on drawing the median line and counting the number of runs — that is, maximal consecutive blocks of observations above or below the median, after observations exactly equal to the median are discarded. If that number is R, then

$$\mu_R = \mathbb{E}\,[R] = 1 + \frac{2n_1 n_2}{n_1 + n_2} \qquad \text{and} \qquad \sigma_R^2 = \mathrm{Var}(R) = \frac{(\mu_R - 1)(\mu_R - 2)}{n_1 + n_2 - 1},$$

where n_1 and n_2 are the numbers of observations above and below the median. Conditional on these counts, and when both are sufficiently large, we have the approximation

$$R' = \frac{R - \mu_R}{\sqrt{\sigma_R^2}} \approx \mathcal{N}(0,1),$$

and we again reject the null hypothesis that the data are random if $|R'| > z_{1-\alpha/2}$. In R, the function `runs.test(y)` from the `lawstat` package carries out this test.

Example 2.2 For the same data as in Example 2.1, the median is $m = 8.24$, the observed number of runs is $R = 2$, and $n_1 = 5$, $n_2 = 5$. Therefore,

$$\mu_R = 1 + \frac{2 \times 5 \times 5}{5+5} = 6 \qquad \text{and} \qquad \sigma_R^2 = \frac{(6-1)(6-2)}{5+5-1} = \frac{20}{9} \approx 2.22.$$

The standardized statistic is

$$R' = \frac{R - \mu_R}{\sqrt{\sigma_R^2}} = \frac{2-6}{\sqrt{2.22}} \approx -2.68.$$

At significance level $\alpha = 0.05$, we compare with $z_{0.975} = 1.96$. Since $|-2.68| > 1.96$, we reject H_0 and conclude that the data are not random according to the runs test. Notice that this differs from the conclusion of the difference sign test.

Lecture 11 — Tuesday, October 14, 2014

Trend Removal and Smoothing

Three useful diagnostics are the Shapiro–Wilk test for normality, the difference sign test for randomness, and the runs test for randomness.

In the **classical decomposition model**, $X_t = m_t + s_t + Y_t$, where m_t is a slowly changing trend and s_t is periodic with period d and we require that

$$\sum_{t=1}^{d} s_t = 0.$$

If the periodic component were not centered around zero, it would contribute to the average level

of the data. The centering condition ensures that the average over one period is approximately the trend m_t.

In $X_t = m_t + s_t + Y_t$, the last term Y_t is the error term, assumed stationary and often modelled as either white noise or independent and identically distributed noise (the latter is ideal). For a nonseasonal model with trend,

$$X_t = m_t + Y_t = (m_t + \mathbb{E}[Y_t]) + (Y_t - \mathbb{E}[Y_t]),$$

so it is natural to assume that $\mathbb{E}[Y_t] = 0$.

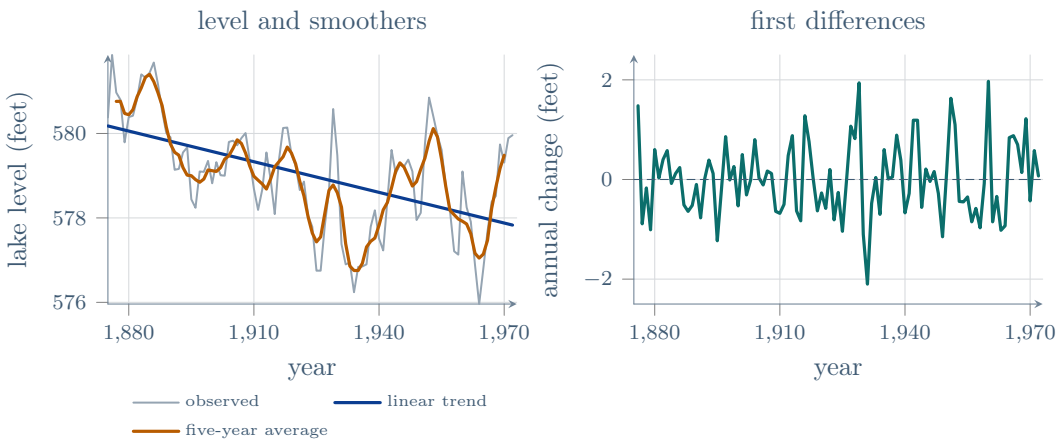
The term **smoothing** refers to the estimation of the trend, or general behaviour, in the data.

The **moving average filter** takes a local window of $2q + 1$ observations, averages them, and places the result at the center of the window. The motivation is that

$$W_{q,t} := \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t-j} = \frac{1}{2q+1} \sum_{j=-q}^{q} m_{t-j} + \frac{1}{2q+1} \sum_{j=-q}^{q} Y_{t-j} \approx m_t.$$

The averaging operation suppresses noise on short scales and reveals the underlying trend. For example, when $q = 2$, we average the current observation together with the two observations on each side. In general, the moving average uses $2q + 1$ observations.

The built-in `LakeHuron` series records annual lake levels, in feet, from 1875 through 1972. In the left panel of Figure 9, a fitted linear trend summarizes the long-run decline, while a centered five-year moving average follows the medium-run fluctuations. The estimated linear slope is about -0.0242 feet per year. The right panel shows the first differences, which fluctuate around a much more stable level after the long-run trend has been removed. The displays also clarify an important distinction: a regression line is a global model for trend, whereas a moving average is a local smoother.



global trend, local smoothing, and differencing answer different questions

FIGURE 9

The analysis uses only the built-in series and base R.

```r
data(LakeHuron)

year <- as.numeric(time(LakeHuron))
linear_trend <- lm(as.numeric(LakeHuron) ~ year)
moving_average <- filter(LakeHuron, rep(1 / 5, 5), sides = 2)

plot(LakeHuron, ylab = "Lake level (feet)")
lines(year, fitted(linear_trend), col = "steelblue", lwd = 2)
lines(moving_average, col = "darkorange", lwd = 2)
legend(
  "topright",
  c("Observed", "Linear trend", "Five-year moving average"),
  col = c("black", "steelblue", "darkorange"),
  lty = 1,
  bty = "n"
)

plot(diff(LakeHuron), ylab = "Annual change (feet)")
abline(h = 0, lty = 2)
```

Oversmoothing means failing to capture enough structure, while **undersmoothing** means capturing too much of the noise. A natural question is whether there is an optimal choice of q.

In **exponential smoothing**, for $\alpha \in [0, 1]$ we define

$$\hat{m}_t = \alpha X_t + (1 - \alpha)\hat{m}_{t-1}$$

with initial condition $\hat{m}_1 = X_1$, which corresponds to a weighted average of today's value (X_t) and yesterday's trend ($\hat{m}_{t-1}$). Iterating the recursion gives

$$\hat{m}_t = \sum_{j=0}^{t-2} \alpha(1 - \alpha)^j X_{t-j} + (1 - \alpha)^{t-1} X_1, \quad t \geq 2.$$

The weights decay exponentially as we move into the past. When $\alpha = 1$, there is effectively no smoothing, whereas very small values of α produce heavy smoothing. Figure 10 compares the two extremes.

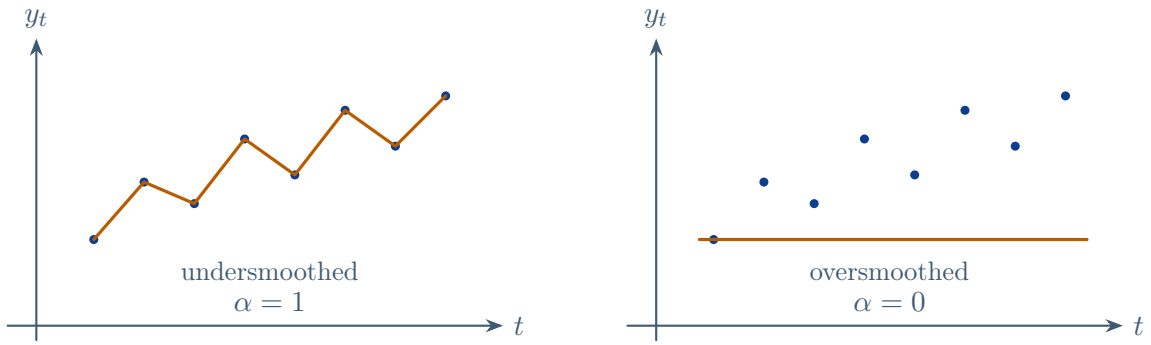


FIGURE 10

For **differencing**, we define two useful operators.

Definition 11.1 The **differencing operator** ∇ and the **backshift operator** B are defined by

$$\nabla X_t := X_t - X_{t-1} \qquad \text{and} \qquad BX_t := X_{t-1},$$

so that $\nabla X_t = (1 - B)X_t$.

The differencing operator can be composed with itself:

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

More generally,

$$\nabla^h X_t = (1 - B)^h X_t = \sum_{j=0}^{h} (-1)^j \binom{h}{j} X_{t-j}.$$

Trend elimination via differencing essentially amounts to approximating the trend by a low-order polynomial.

Example 11.2 Consider $Y_t = 5 + 3t + R_t$, where $\{R_t\}$ is iid $\mathcal{N}(0, \sigma^2)$ noise. Investigate the stationarity of Y_t and of ∇Y_t.

Solution. Since $\mathbb{E}[Y_t] = 5 + 3t$ depends on t, the process Y_t is not stationary. On the other hand,

$$\nabla Y_t = Y_t - Y_{t-1} = 5 + 3t + R_t - (5 + 3(t-1) + R_{t-1}) = 3 + R_t - R_{t-1}.$$

This expression no longer depends on t through its mean. In fact, ∇Y_t is an MA(1) process with mean 3 and variance $2\sigma^2$, and its covariance structure depends only on the lag. Hence ∇Y_t is stationary. □

Lecture 12 — Thursday, October 16, 2014

Differencing can remove a deterministic trend. Generalizing Example 11.2, if $X_t = a + bt + Y_t$ with $Y_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, then X_t is not stationary but ∇X_t is.

Example 12.1 Consider the process $X_t = 2 + t + t^2 + Y_t$, where $Y_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Investigate the stationarity of X_t, ∇X_t, and $\nabla^2 X_t$.

Solution. The process X_t is not stationary because its mean depends on t.

For ∇X_t, we calculate

$$\nabla X_t = 2 + t + t^2 + Y_t - (2 + (t-1) + (t-1)^2 + Y_{t-1}) = 2t + Y_t - Y_{t-1},$$

and so $\mathbb{E}[\nabla X_t] = 2t$; hence ∇X_t is still not stationary.

For $\nabla^2 X_t$, we have

$$\begin{aligned} \nabla^2 X_t &= \nabla(\nabla X_t) = \nabla(2t + Y_t - Y_{t-1}) \\ &= 2t + Y_t - Y_{t-1} - (2(t-1) + Y_{t-1} - Y_{t-2}) \end{aligned}$$

$$= Y_t - 2Y_{t-1} + Y_{t-2} + 2.$$

Thus $\mathbb{E}\left[\nabla^2 X_t\right] = 2$, which no longer depends on t. To verify stationarity, define

$$W_t := \nabla^2 X_t - 2 = Y_t - 2Y_{t-1} + Y_{t-2}.$$

Since $\{Y_t\}$ is iid with $\mathrm{Var}(Y_t) = \sigma^2$, we have

$$\gamma_W(h) := \mathrm{Cov}(W_t, W_{t+h}) = \begin{cases} 6\sigma^2, & h = 0, \\ -4\sigma^2, & |h| = 1, \\ \sigma^2, & |h| = 2, \\ 0, & |h| > 2. \end{cases}$$

In particular, $\gamma_W(h)$ depends only on h and not on t. Therefore

$$\mathrm{Cov}(\nabla^2 X_t, \nabla^2 X_{t+h}) = \gamma_W(h)$$

also depends only on the lag, so $\nabla^2 X_t$ is stationary. □

Proposition 12.2 If $\mathbb{E}[X_t]$ is a polynomial in t of degree k, then the mean of $\nabla^k X_t$ no longer depends on time.

Proof. Let $m(t) := \mathbb{E}[X_t]$. By linearity of expectation and the definition of differencing,

$$\mathbb{E}[\nabla X_t] = \mathbb{E}[X_t - X_{t-1}] = m(t) - m(t-1) = (\nabla m)(t).$$

Repeating this argument gives

$$\mathbb{E}[\nabla^j X_t] = (\nabla^j m)(t), \quad j = 1, 2, \ldots.$$

Now assume $m(t)$ is a polynomial of degree k:

$$m(t) = a_k t^k + a_{k-1} t^{k-1} + \cdots + a_0, \quad a_k \neq 0.$$

For each monomial t^r with $r \geq 1$, the first difference

$$\nabla(t^r) = t^r - (t-1)^r$$

is a polynomial of degree $r-1$. Hence applying ∇ once lowers polynomial degree by one. Therefore $(\nabla m)(t)$ has degree $k-1$, $(\nabla^2 m)(t)$ has degree $k-2$, and after k differences, $(\nabla^k m)(t)$ is a constant

(in fact $k!a_k$ for unit lag). Consequently,

$$\mathbb{E}[\nabla^k X_t] = (\nabla^k m)(t)$$

does not depend on t. □

An important principle is that we always begin with first differencing, that is, with ∇X_t. If that is still not stationary, we try $\nabla^2 X_t$, and so on. We should always start with the lowest-order differencing that seems plausible, guided by the time series plot and the ACF.

The reason for this restraint is that differencing can jack up the variance. For example, suppose $X_t = Y_t - 2Y_{t-1}$ with $Y_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Then X_t is already stationary, and $\text{Var}(X_t) = \sigma^2 + 4\sigma^2 = 5\sigma^2$. But if we difference the process, then

$$\nabla X_t = Y_t - 3Y_{t-1} + 2Y_{t-2} \implies \text{Var}(\nabla X_t) = \sigma^2 + 9\sigma^2 + 4\sigma^2 = 14\sigma^2.$$

This illustrates why overdifferencing is undesirable: it can inflate the variance substantially.

For the datasets in this course, finite variance is usually a reasonable working assumption. The practical work is to decide whether the mean and dependence structure are stable over time. We use the time series plot to look for changes in level or variance and the sample ACF to look for trend, seasonality, and remaining serial dependence.

$\hat{Y}_t = X_t - \hat{m}_t - \hat{s}_t$ is the detrended and deseasonalized series (the estimated random component). In practice, the decomposition is carried out in four steps. First, estimate the trend m_t with a moving-average filter chosen to remove seasonality and damp random noise. Second, estimate the seasonal component by averaging observations from the same seasonal position across cycles. Third, remove seasonality by forming $X_t - \hat{s}_t$. Fourth, remove trend from the deseasonalized series using the same approach to estimating the trend.

Estimating Seasonality and Trend in Practice

We work with the additive decomposition

$$X_t = m_t + s_t + Y_t, \quad t = 1, \ldots, n,$$

where $\mathbb{E}[Y_t] = 0$, $s_{t+d} = s_t$ (period d), and

$$\sum_{j=1}^{d} s_j = 0.$$

The practical workflow is:

1. Estimate trend m_t by a moving-average filter chosen to suppress seasonality and noise.
2. Estimate seasonality by averaging detrended values at each seasonal position.
3. Eliminate seasonality via $X_t - \hat{s}_t$.
4. Eliminate trend from the deseasonalized series to obtain the estimated noise sequence.

If $d = 2q + 1$ is odd, use the usual symmetric moving average. If $d = 2q$ is even, use the centered moving average

$$\hat{m}_t = \frac{0.5x_{t-q} + x_{t-q+1} + \cdots + x_{t+q-1} + 0.5x_{t+q}}{d}, \quad q < t \leq n - q.$$

For each seasonal position $k = 1, \ldots, d$, define

$$w_k = \text{average of } \{x_{k+jd} - \hat{m}_{k+jd} : q < k + jd \leq n - q\}.$$

Then normalize:

$$\hat{s}_k = w_k - \frac{\sum_{j=1}^{d} w_j}{d},$$

so that

$$\sum_{j=1}^{d} \hat{s}_j = 0.$$

For monthly data, this is just averaging all detrended January values, all detrended February values, and so on.

Extend the estimates periodically by setting $\hat{s}_t = \hat{s}_k$ whenever $t \equiv k \pmod{d}$. The deseasonalized series is

$$D_t = X_t - \hat{s}_t,$$

and the estimated random component is

$$\hat{Y}_t = X_t - \hat{m}_t - \hat{s}_t.$$

After computing this decomposition, we check whether $\{\hat{Y}_t\}$ behaves like an iid sequence with mean zero.

Graphical checks include inspecting the residual plot and sample ACF. Under the hypothesis of iid noise with a finite fourth moment, $\pm 1.96/\sqrt{n}$ gives an approximate pointwise 95% reference interval at each fixed nonzero lag. Because the intervals are not simultaneous, an isolated exceedance among many inspected lags is unsurprising.

Formal tests include the runs, difference-sign, and Shapiro–Wilk tests; graphical checks include the residual plot, sample ACF, and Q-Q plot. If residuals still show trend or seasonality, reestimate the trend and seasonal components from the residual structure and repeat the diagnostic cycle.

Lecture 13 — Thursday, October 16, 2014

The Holt–Winters Algorithm

Simple exponential smoothing does not accommodate seasonality. The Holt–Winters method extends it to the model

$$X_t = m_t + s_t + Y_t \qquad \text{and} \qquad \mathbb{E}[Y_t] = 0.$$

To generalize exponential smoothing, let us examine the local mean value of the process, namely $\mu_t = a_t + b_t t$, where a_t and b_t vary slowly through time. At each point in time, a_t represents the level and b_t quantifies the trend.

The **Holt–Winters algorithm** combines the level and trend with a seasonal component in an additive or multiplicative fashion. To introduce this method, we use the following notation:

L_t: the level of the process at time t

T_t: the trend of the process at time t

I_t: seasonal effect at time t (period $= p$)

To forecast h steps ahead, Holt and Winters proposed the following algorithms:

1. In the additive case,

$$h\text{-step-ahead forecast} = L_t + hT_t + I_{t-p+1+(h-1) \bmod p}$$

where

$$\begin{aligned} L_t &= \alpha(X_t - I_{t-p}) + (1-\alpha)(L_{t-1} + T_{t-1}) \\ T_t &= \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \\ I_t &= \gamma(X_t - L_t) + (1-\gamma)I_{t-p} \end{aligned}$$

2. In the multiplicative case,

$$h\text{-step-ahead forecast} = (L_t + hT_t)I_{t-p+1+(h-1) \bmod p}$$

where

$$\begin{aligned} L_t &= \alpha\left(\frac{X_t}{I_{t-p}}\right) + (1-\alpha)(L_{t-1} + T_{t-1}) \\ T_t &= \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \\ I_t &= \gamma\left(\frac{X_t}{L_t}\right) + (1-\gamma)I_{t-p} \end{aligned}$$

The term $I_{t-p+1+(h-1) \bmod p}$ selects the seasonal component appropriate to horizon h.

As with any other system of difference equations, we must choose starting values for L_t, T_t, and I_t, and we must also specify α, β, and γ. We will not pursue the theoretical details of the Holt–Winters algorithm here; instead, we emphasize its applied use. In practice, R estimates α, β, and γ by minimizing the one-step forecasting error.

The Holt–Winters method is best understood through worked examples, but the main conceptual point is that it updates level, trend, and seasonality automatically.

Compared with regression, the Holt–Winters method is relatively automatic. In practice, we mainly decide whether an additive or multiplicative seasonal form is more appropriate. In regression analysis, by contrast, we must choose the model and the explanatory variables explicitly.

Lecture 14 — Tuesday, October 21, 2014

Special Cases of the Holt–Winters Algorithm

Figure 11 compares additive and multiplicative seasonal structure.

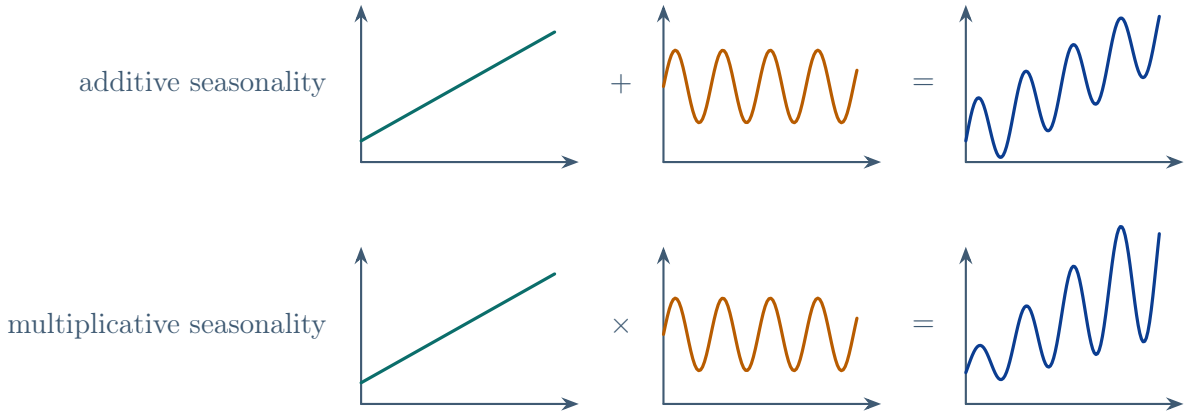


FIGURE 11

In the additive seasonal model, the combined signal is the sum of the trend and the seasonal component. In the multiplicative seasonal model, the combined signal is the product of the trend and the seasonal component.

When there is no seasonal pattern, we omit the seasonal update and use the Holt level–trend recursions

$$L_t = \alpha X_t + (1-\alpha)(L_{t-1} + T_{t-1}),$$
$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}.$$

This is often called **double exponential smoothing**. The parameter α controls the response of the level to new observations, while β controls the response of the trend.

If we also suppress the trend by taking $T_t = 0$, the level recursion reduces to **simple exponential smoothing**:

$$m_t = \alpha Y_t + (1-\alpha)m_{t-1}.$$

Since the one-step-ahead forecast is $\hat{Y}_{t+1} = m_t$, we can write the update in error-correction form

as

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y_t - \hat{Y}_t).$$

At forecast origin t, future observations are unavailable, so simple exponential smoothing produces the flat forecast

$$\hat{Y}_{t+h|t} = m_t, \quad h = 1, 2, 3, \ldots.$$

In R, setting `gamma = FALSE` and `beta = FALSE` in `HoltWinters` selects simple exponential smoothing, and R optimizes α. Setting `gamma = FALSE` while retaining the trend selects Holt's method. For example, with estimated level 256 and estimated slope 6.295, its five-step-ahead forecast is $256 + 5(6.295)$.

With seasonality present, the additive forecast is

$$L_t + hT_t + I_{t-p+1+(h-1) \bmod p},$$

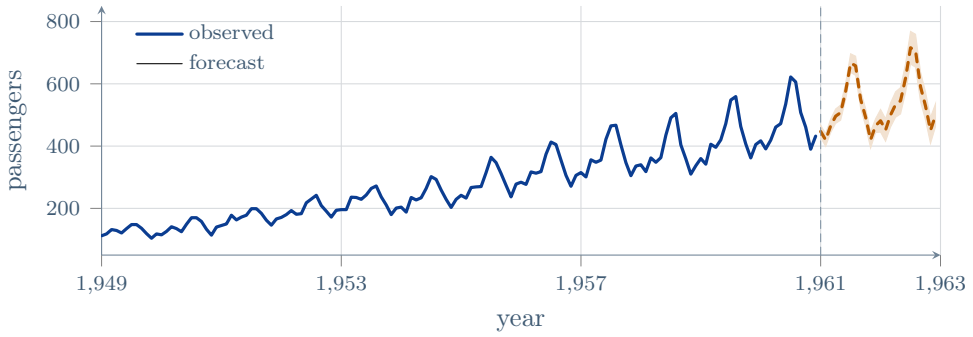
whereas the multiplicative forecast is

$$(L_t + hT_t) I_{t-p+1+(h-1) \bmod p}.$$

For `AirPassengers`, multiplicative Holt–Winters smoothing is appropriate because the seasonal amplitude increases with the level. Base R estimates

$$\hat{\alpha} = 0.276, \qquad \hat{\beta} = 0.033, \qquad \text{and} \qquad \hat{\gamma} = 0.871.$$

Thus the fitted procedure adapts its seasonal factors much more quickly than its trend. The two-year forecast in Figure 12 preserves both the upward movement and the multiplicative annual cycle.



multiplicative Holt–Winters forecasting carries the growing seasonal cycle forward

FIGURE 12

The fitted smoother and its forecast can be reproduced with the following commands.

```
data(AirPassengers)

hw <- HoltWinters(AirPassengers, seasonal = "multiplicative")
hw

forecast <- predict(
  hw,
  n.ahead = 24,
  prediction.interval = TRUE,
  level = 0.95
)
plot(
  hw,
  forecast,
  main = "Multiplicative Holt-Winters forecast"
)
```

Lecture 15 — Thursday, October 23, 2014

3. Moving Average Processes, Autoregressive Processes, and Gaussian Processes

Forecasting requires some aspect of the data-generating mechanism to persist from the past into the future. Stationarity makes that idea precise and provides the framework for the models that follow.

The Moving Average (MA(q)) Process

Definition 15.1 A time series $\{X_t : t \in \mathbb{Z}\}$ is called a **moving average process of order q** if

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, $\theta_1, \ldots, \theta_q$ are constants, and $\theta_q \neq 0$ when $q \geq 1$.

Observe that each observed X_t is a linear combination of the current and previous q **innovations** Z_t, where the innovations are uncorrelated, have constant variance, and have mean zero.

We easily compute the mean and variance:

$$\mathbb{E}[X_t] = \mathbb{E}[Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}] = 0 + \theta_1(0) + \cdots + \theta_q(0) = 0$$

and because the Z_t are uncorrelated,

$$\begin{aligned}
\text{Var}(X_t) &= \text{Var}(Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}) \\
&= \text{Var}(Z_t) + \theta_1^2 \text{Var}(Z_{t-1}) + \cdots + \theta_q^2 \text{Var}(Z_{t-q}) \\
&= \sigma^2(1 + \theta_1^2 + \cdots + \theta_q^2)
\end{aligned}$$

For the autocovariance function,

$$\begin{aligned}
\gamma(h) &= \text{Cov}(X_t, X_{t+h}) \\
&= \text{Cov}\left(\sum_{i=0}^{q} \theta_i Z_{t-i}, \sum_{j=0}^{q} \theta_j Z_{t+h-j}\right) \quad (\text{where } \theta_0 = 1) \\
&= \sum_{i=0}^{q} \sum_{j=0}^{q} \theta_i \theta_j \, \text{Cov}(Z_{t-i}, Z_{t+h-j})
\end{aligned}$$

Since $\text{Cov}(Z_t, Z_s) = \sigma^2$ if $t = s$ and 0 otherwise:

$$\gamma(h) = \begin{cases} \sigma^2 \sum_{i=0}^{q-h} \theta_i \theta_{i+h} & \text{for } 0 \leq h \leq q \\ 0 & \text{for } h > q \end{cases}$$

Note that $\gamma(h) = \gamma(-h)$ by symmetry. The ACF is $\rho(h) = \gamma(h)/\gamma(0)$. These calculations also show that the MA(q) process is stationary.

Definition 15.2 A process $\{X_t\}$ is called ***q*-dependent** if X_t and X_s are independent whenever $|t-s| > q$.

This condition does not require dependence at lag q: a process that is q-dependent is also $(q+1)$-dependent. When we speak of the *exact* dependence order, we mean the smallest nonnegative integer for which the condition holds.

$$\begin{array}{lccccccccc} |t-s|: & 1 & 2 & 3 & \dots & q-1 & & q & q+1 & q+2 & \dots \\ & \multicolumn{5}{c}{\text{no conclusion is available}} & & \downarrow & \multicolumn{3}{c}{\text{independent}} \\ & & & & & & & \text{no conclusion} & & & \end{array}$$

A sequence of iid noise is 0-dependent because distinct terms are independent.

Definition 15.3 Similarly, we say that a stationary time series is ***q*-correlated** if $\gamma(h) = 0$ when $|h| > q$.

Again, this need not be the smallest possible q; the exact correlation order is the smallest such integer.

$$\begin{array}{lccccccccc} |h|: & 1 & 2 & 3 & \dots & q-1 & & q & q+1 & q+2 & \dots \\ & \multicolumn{5}{c}{\text{no conclusion is available}} & & \downarrow & \multicolumn{3}{c}{\gamma(h) = 0} \\ & & & & & & & \text{no conclusion} & & & \end{array}$$

White noise is 0-correlated, but it need not be 0-dependent because uncorrelatedness does not imply independence. By contrast, iid noise is both 0-dependent and 0-correlated.

Example 15.4 Show that an MA(1) process is 1-correlated.

Solution. We have $X_t = Z_t + \theta Z_{t-1}$, $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. In Example 1.3, we showed that

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma^2 & h = 0 \\ \theta\sigma^2 & |h| = 1 \\ 0 & |h| > 1 \end{cases}.$$

So $\gamma(h) = 0$ for all $|h| > 1$, and therefore the MA(1) process is 1-correlated. □

More generally, every MA(q) process is q-correlated. It is also q-dependent when its innovations are independent, but white-noise innovations alone guarantee only the correlation statement. A converse representation theorem is given below. For intuition, Figure 13 shows how adjacent terms of an MA(1) process share innovations:

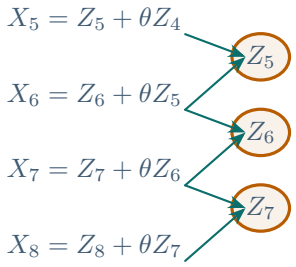


FIGURE 13

The general result is as follows:

Proposition 15.5 If $\{X_t : t \in \mathbb{Z}\}$ is a stationary q-correlated process with mean 0, then it has an MA(q) representation in the mean-square sense.

The proof requires harmonic analysis, which is beyond the scope of our course.

The Autoregressive (AR(p)) Process

Definition 15.6 A time series $\{X_t : t \in \mathbb{Z}\}$ is called an autoregressive process of order p (AR(p)) if

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t, \quad t \in \mathbb{Z},$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, $\phi_1, \ldots, \phi_p$ are constants, and $\phi_p \neq 0$.

We can also express the AR(1) process as $(1-\phi B)X_t = Z_t$ where B is the backward shift operator. More generally, we can show that for the general AR(p) process,

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)X_t = Z_t.$$

If $|\phi| = 1$, the equation $X_t = \phi X_{t-1} + Z_t$ has no weakly stationary solution with positive innovation variance. If $|\phi| < 1$, it has the familiar causal stationary solution. When $|\phi| > 1$, a stationary

solution still exists, but it is noncausal and depends on future innovations:

$$X_t = -\sum_{j=1}^{\infty} \phi^{-j} Z_{t+j}.$$

Thus stationarity and causality are distinct requirements.

There is an important connection between AR(1) and MA(q) processes. Consider an AR(1) process $X_t = \phi X_{t-1} + Z_t$ with $|\phi| < 1$. We have

$$\begin{aligned} X_t &= \phi X_{t-1} + Z_t \\ &= \phi[\phi X_{t-2} + Z_{t-1}] + Z_t \\ &= \phi^2 X_{t-2} + \phi Z_{t-1} + Z_t \\ &= \phi^3 X_{t-3} + \phi^2 Z_{t-2} + \phi Z_{t-1} + Z_t \\ &\vdots \\ &= \phi^m X_{t-m} + \sum_{j=0}^{m-1} \phi^j Z_{t-j}. \end{aligned}$$

For the stationary solution, $\mathbb{E}[X_{t-m}^2] = \gamma(0)$, so

$$\mathbb{E}\big[(\phi^m X_{t-m})^2\big] = \phi^{2m}\gamma(0) \longrightarrow 0.$$

Thus the remainder converges to zero in mean square and

$$X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}.$$

Defining $\theta_j = \phi^j$ for $j = 0, 1, \ldots$, we have thus written X_t as an MA(∞) process.

Lecture 17 — Tuesday, October 28, 2014

An MA(q) process is always q-correlated and is q-dependent when the innovations are independent. An AR(1) process with $|\phi| < 1$ has an MA(∞) representation.

Gaussian Processes

Definition 17.1 A time series $\{X_t\}$ is called a **Gaussian process** if all of its finite-dimensional joint distributions are multivariate normal. That is, for any collection of times $t_1, t_2, \ldots, t_n$, the random vector $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ has a multivariate normal distribution.

Equivalently, every finite-dimensional random vector extracted from the process is multivariate normal.

Recall that if $\sigma_1^2 > 0$, $\sigma_2^2 > 0$, and

$$\boldsymbol{X} = (X_1, X_2)^\top \sim \mathcal{N}_2\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}\right),$$

then

$$\begin{aligned} X_1|X_2 = x_2 &\sim \mathcal{N}\left(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}\right) \\ &= \mathcal{N}\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2 - \mu_2), \sigma_1^2(1-\rho^2)\right) \end{aligned}$$

where $\rho = \sigma_{12}/(\sigma_1\sigma_2)$.

Example 17.2 Consider a stationary Gaussian process $\{X_t\}$ with $\sigma^2 = \text{Var}(X_t) > 0$. Suppose that X_n has been observed and that we wish to forecast X_{n+h}, where $h > 0$, using a function $m(X_n)$ of X_n. We measure the quality of the forecast by the conditional mean squared error

$$MSE = \mathbb{E}\left[(X_{n+h} - m(X_n))^2 \mid X_n\right],$$

which we wish to minimize. Derive $m(X_n)$.

Solution. It can be shown that the function which minimizes the MSE in the general case (not necessarily Gaussian) is $m(X_n) = \mathbb{E}[X_{n+h}|X_n]$. In the case of this example, stationarity implies that $\mathbb{E}[X_{n+h}] = \mathbb{E}[X_n] = \mu$. Also, $\text{Cov}(X_{n+h}, X_n) = \gamma(h)$ and so

$$\text{Corr}(X_{n+h}, X_n) = \frac{\text{Cov}(X_{n+h}, X_n)}{\sqrt{\text{Var}(X_{n+h})\,\text{Var}(X_n)}} = \frac{\gamma(h)}{\sigma^2} = \rho(h).$$

By the conditional distribution formula for multivariate normal vectors,

$$(X_{n+h}, X_n)^\top \sim \mathcal{N}_2\left(\begin{bmatrix}\mu\\ \mu\end{bmatrix}, \begin{bmatrix}\sigma^2 & \gamma(h)\\ \gamma(h) & \sigma^2\end{bmatrix}\right).$$

Hence

$$X_{n+h}|X_n = x_n \sim \mathcal{N}\left(\mu + \frac{\gamma(h)}{\sigma^2}(x_n - \mu), \sigma^2(1-\rho^2(h))\right).$$

Thus, $m(X_n) = \mathbb{E}\left[X_{n+h}|X_n\right] = \mu + \rho(h)(X_n - \mu)$ minimizes the MSE, giving

$$MSE = \mathbb{E}\left[(X_{n+h} - m(X_n))^2|X_n\right] = \mathrm{Var}(X_{n+h}|X_n) = \sigma^2(1-\rho^2(h)).$$

The mean squared error is equal to 0 if and only if $\rho^2(h) = 1$, which means that $\rho(h) = \pm 1$. Thus the precision of prediction depends on how strongly correlated X_n and X_{n+h} are. □

Warning In general, Gaussian processes need not be stationary.

Remark 17.3 For a Gaussian process, weak stationarity implies (and is thus equivalent to) strong stationarity.

Predictors of the form $aX_n + b$ remain useful even without Gaussianity. We choose a and b to minimize $\mathbb{E}[(X_{n+h} - aX_n - b)^2]$. The tradeoff is clear: we drop the Gaussian-process assumption but restrict attention to the best linear predictor.

Linear Prediction

We now consider the problem of predicting X_{n+h}, where $h > 0$, for a stationary time series with known mean μ and autocovariance function $\gamma(\cdot)$, based on the previous values $X_1, \ldots, X_n$. Denote the linear predictor of X_{n+h} by P_nX_{n+h}. We seek a predictor of the form

$$P_nX_{n+h} = a_0 + a_1X_n + a_2X_{n-1} + \cdots + a_nX_1,$$

which minimizes

$$S(a_0, a_1, \ldots, a_n) = \mathbb{E}\left[(X_{n+h} - P_nX_{n+h})^2\right].$$

To determine $a_0, a_1, \ldots, a_n$, we solve the system of equations $(\partial S)/(\partial a_j) = 0$ for $j = 0, 1, \ldots, n$. If the covariance matrix of $(X_n, \ldots, X_1)^\top$ is nonsingular, the solution is unique. Doing so, we obtain

$$a_0 = \mu \left(1 - \sum_{i=1}^{n} a_i\right),$$

and

$$\boldsymbol{\Gamma}_n \boldsymbol{a} = \boldsymbol{\gamma}_n,$$

where

$$\boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\Gamma}_n = \begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \ldots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \ldots & \gamma(n-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \ldots & \ldots & \gamma(0) \end{bmatrix},$$

and

$$\boldsymbol{\gamma}_n = \begin{bmatrix} \gamma(h) \\ \gamma(h+1) \\ \vdots \\ \gamma(h+n-1) \end{bmatrix}.$$

Thus,

$$P_n X_{n+h} = \mu + \sum_{j=1}^{n} a_j (X_{n+1-j} - \mu).$$

We record two useful derivations for linear prediction.

Proposition 17.4 (Best linear predictor based on X_n) For a stationary time series $\{X_t\}$ with $\gamma(0) > 0$, the coefficients a, b minimizing

$$\mathbb{E}\left[(X_{n+h} - aX_n - b)^2\right]$$

are

$$a = \frac{\gamma(h)}{\gamma(0)} = \rho(h) \qquad \text{and} \qquad b = \mu(1-a) = \mu\big(1 - \rho(h)\big).$$

Hence the best linear predictor based on X_n is

$$aX_n + b = \mu + \rho(h)(X_n - \mu).$$

Proof. Let

$$M(a,b) = \mathbb{E}\left[(X_{n+h} - aX_n - b)^2\right].$$

From $\partial M/\partial b = 0$,

$$\mathbb{E}[X_{n+h}] - a\mathbb{E}[X_n] - b = 0.$$

By stationarity, $\mathbb{E}[X_t] = \mu$, so $b = (1-a)\mu$.

From $\partial M/\partial a = 0$,

$$\mathbb{E}[X_n X_{n+h}] - a\mathbb{E}[X_n^2] - b\mathbb{E}[X_n] = 0.$$

Substitute $b = (1-a)\mu$:

$$\mathbb{E}[X_n X_{n+h}] - \mu^2 - a\big(\mathbb{E}[X_n^2] - \mu^2\big) = 0.$$

Since $\mathrm{Cov}(X_n, X_{n+h}) = \gamma(h)$ and $\mathrm{Var}(X_n) = \gamma(0)$,

$$\gamma(h) - a\gamma(0) = 0,$$

so $a = \gamma(h)/\gamma(0) = \rho(h)$, and therefore $b = \mu(1 - \rho(h))$. □

Proposition 17.5 (Normal equations for the full linear predictor) Let

$$P_n X_{n+h} = a_0 + a_1 X_n + a_2 X_{n-1} + \cdots + a_n X_1$$

minimize

$$S(a_0, \ldots, a_n) = \mathbb{E}\left[(X_{n+h} - P_n X_{n+h})^2\right].$$

Then

$$a_0 = \mu\left(1 - \sum_{i=1}^{n} a_i\right),$$

and for $j = 1, \ldots, n$,

$$\gamma(h + j - 1) = \sum_{i=1}^{n} a_i \gamma(i - j).$$

Equivalently,

$$\boldsymbol{\Gamma}_n \boldsymbol{a} = \boldsymbol{\gamma}_n,$$

where

$$\boldsymbol{a} = (a_1, \ldots, a_n)^\top, \qquad \boldsymbol{\Gamma}_n = [\gamma(i-j)]_{i,j=1}^{n},$$
$$\text{and} \qquad \boldsymbol{\gamma}_n = (\gamma(h), \gamma(h+1), \ldots, \gamma(h+n-1))^\top.$$

Therefore,

$$P_n X_{n+h} = \mu + \sum_{i=1}^{n} a_i (X_{n+1-i} - \mu).$$

Proof. Set the normal equations $\partial S / \partial a_j = 0$, $j = 0, 1, \ldots, n$. For $j = 0$,

$$\frac{\partial S}{\partial a_0} = 0 \implies \mathbb{E}[X_{n+h} - a_0 - a_1 X_n - \cdots - a_n X_1] = 0,$$

hence

$$a_0 = \mu\left(1 - \sum_{i=1}^{n} a_i\right).$$

For $j = 1, \ldots, n$,

$$\frac{\partial S}{\partial a_j} = 0 \implies \mathbb{E}\big[X_{n+1-j}\big(X_{n+h} - a_0 - a_1 X_n - \cdots - a_n X_1\big)\big] = 0.$$

Substitute the formula for a_0 and expand:

$$\mathbb{E}[X_{n+1-j}X_{n+h}] - \mu^2 - \sum_{i=1}^{n} a_i\big(\mathbb{E}[X_{n+1-j}X_{n+1-i}] - \mu^2\big) = 0.$$

By stationarity this becomes

$$\gamma(h+j-1) - \sum_{i=1}^{n} a_i\gamma(i-j) = 0,$$

that is,

$$\gamma(h+j-1) = \sum_{i=1}^{n} a_i\gamma(i-j), \quad j = 1,\dots,n.$$

These equations are exactly $\boldsymbol{\Gamma}_n\boldsymbol{a} = \boldsymbol{\gamma}_n$. Finally,

$$P_nX_{n+h} = a_0 + \sum_{i=1}^{n} a_iX_{n+1-i} = \mu + \sum_{i=1}^{n} a_i(X_{n+1-i} - \mu).$$

□

We now justify this formula in a standard case.

Example 17.6 Suppose $X_t = \phi X_{t-1} + Z_t$, where $|\phi| < 1$ and $\{Z_t\} \sim \text{WN}(0,\sigma^2)$. We derive the one-step predictor for this AR(1) process. We know that $\gamma(h) = \phi^{|h|}\gamma(0)$ for $h = 0,1,2,\dots$, and that $\mu = \mathbb{E}[X_t] = 0$. To find the best linear predictor, we solve $\boldsymbol{\Gamma}_n\mathbf{a} = \gamma_n(h)$:

$$\begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \dots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \dots & \gamma(n-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \dots & \dots & \gamma(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \gamma(h) \\ \gamma(h+1) \\ \vdots \\ \gamma(h+n-1) \end{bmatrix}$$

One-step prediction: $h = 1$. So we have:

$$\begin{bmatrix} \gamma(0) & \phi\gamma(0) & \dots & \phi^{n-1}\gamma(0) \\ \phi\gamma(0) & \gamma(0) & \dots & \phi^{n-2}\gamma(0) \\ \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1}\gamma(0) & \dots & \dots & \gamma(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \phi\gamma(0) \\ \phi^2\gamma(0) \\ \vdots \\ \phi^n\gamma(0) \end{bmatrix}$$

Divide by $\gamma(0)$:

$$\begin{bmatrix} 1 & \phi & \dots & \phi^{n-1} \\ \phi & 1 & \dots & \phi^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \dots & \dots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \phi \\ \phi^2 \\ \vdots \\ \phi^n \end{bmatrix}$$

Because $|\phi| < 1$, the covariance matrix is positive definite, and the unique solution to the system is

$$a_1 = \phi, a_2 = 0, \dots, a_n = 0.$$

Also, since $\mu = 0$,

$$a_0 = \mu \left(1 - \sum_{i=1}^{n} a_i\right) = 0 \implies P_n X_{n+1} = a_0 + a_1 X_n + \cdots + a_n X_1 = \phi X_n.$$

So the best linear predictor is $P_n X_{n+1} = \phi X_n$.

Lecture 18 — Thursday, October 30, 2014

For a stationary process, the best linear predictor based on $X_1, \dots, X_n$ has the form

$$P_n X_{n+h} = a_0 + a_1 X_n + \cdots + a_n X_1,$$

where the coefficients satisfy

$$\boldsymbol{\Gamma}_n \boldsymbol{a}_n = \boldsymbol{\gamma}_n(h) \qquad \text{and} \qquad a_0 = \mu \left(1 - \sum_{i=1}^{n} a_i\right).$$

Proposition 18.1 (Orthogonality properties of linear prediction) The linear predictor $P_n X_{n+h}$ has the following properties:

(1) It is determined by μ and $\gamma(\cdot)$.

(2) The prediction error has mean zero: $\mathbb{E}[X_{n+h} - P_n X_{n+h}] = 0$.

(3) For each $j = 1, 2, \ldots, n$, we have $\mathbb{E}[(X_{n+h} - P_n X_{n+h}) X_j] = 0$.

(4) Consequently, the prediction error is uncorrelated with each X_j, for $j = 1, 2, \ldots, n$.

Exercise 18.2 Prove the four properties listed in the proposition above.

The same argument works for arbitrary predictors. Let $X, W_1, \ldots, W_n$ have finite second moments, and define

$$\begin{aligned} \boldsymbol{W} &= (W_1, \ldots, W_n)^\top, \\ \boldsymbol{\mu}_W &= \mathbb{E}[\boldsymbol{W}], \\ \boldsymbol{\gamma}_X &= \operatorname{Cov}(\boldsymbol{W}, X), \\ \boldsymbol{\Gamma} &= \operatorname{Cov}(\boldsymbol{W}, \boldsymbol{W}). \end{aligned}$$

If $\boldsymbol{\Gamma}$ is nonsingular, the best linear predictor of X based on $\boldsymbol{W}$ is

$$\operatorname{Pred}(X \mid \boldsymbol{W}) = \mathbb{E}[X] + \boldsymbol{a}_X^\top (\boldsymbol{W} - \boldsymbol{\mu}_W), \quad \boldsymbol{a}_X = \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma}_X.$$

Its mean squared error is

$$\mathbb{E}\left[(X - \operatorname{Pred}(X \mid \boldsymbol{W}))^2\right] = \operatorname{Var}(X) - \boldsymbol{\gamma}_X^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma}_X.$$

We record the standard projection properties without proof.

Proposition 18.3 Suppose X, Y, and the components of $\boldsymbol{W}$ have finite second moments and $\boldsymbol{\Gamma} = \operatorname{Cov}(\boldsymbol{W}, \boldsymbol{W})$ is nonsingular. Then:

1. $\operatorname{Pred}(X \mid \boldsymbol{W}) = \mathbb{E}[X] + \boldsymbol{a}_X^\top(\boldsymbol{W} - \boldsymbol{\mu}_W)$, where $\boldsymbol{a}_X = \boldsymbol{\Gamma}^{-1} \operatorname{Cov}(\boldsymbol{W}, X)$.
2. The prediction error has mean zero and is uncorrelated with every component of $\boldsymbol{W}$.
3. $$\mathbb{E}\left[(X - \operatorname{Pred}(X \mid \boldsymbol{W}))^2\right] = \operatorname{Var}(X) - \operatorname{Cov}(X, \boldsymbol{W})\boldsymbol{\Gamma}^{-1} \operatorname{Cov}(\boldsymbol{W}, X).$$
4. For constants α_1, α_2, and β,
$$\operatorname{Pred}(\alpha_1 X + \alpha_2 Y + \beta \mid \boldsymbol{W}) = \alpha_1 \operatorname{Pred}(X \mid \boldsymbol{W}) + \alpha_2 \operatorname{Pred}(Y \mid \boldsymbol{W}) + \beta.$$
5. $$\operatorname{Pred}\left(\sum_{i=1}^n \alpha_i W_i + \beta \,\middle|\, \boldsymbol{W}\right) = \sum_{i=1}^n \alpha_i W_i + \beta.$$
6. If $\operatorname{Cov}(\boldsymbol{W}, X) = \boldsymbol{0}$, then $\operatorname{Pred}(X \mid \boldsymbol{W}) = \mathbb{E}[X]$.

Exercise 18.4 What is the best linear predictor for X_{n+h} in an AR(p) process based on $X_1, \ldots, X_n$?

Let us calculate the mean squared prediction error in Example 17.6. For one-step prediction in the AR(1) model, we have $\boldsymbol{a}_n = (\phi, 0, 0, \ldots, 0)$, so $P_n X_{n+1} = \phi X_n$. There are at least two ways to do this:

$$\text{MSE} = \mathbb{E}\left[(X_{n+1} - P_n X_{n+1})^2\right] = \mathbb{E}\left[(X_{n+1} - \phi X_n)^2\right] = \mathbb{E}\left[Z_{n+1}^2\right] = \operatorname{Var}(Z_{n+1}) = \sigma^2.$$

Alternatively,

$$\begin{aligned}
\text{MSE} &= \gamma(0) - \boldsymbol{a}_n^\top \boldsymbol{\gamma}(h) \\
&= \gamma(0) - (\phi, 0, \ldots, 0) \cdot (\gamma(0)\phi, \gamma(0)\phi^2, \ldots, \gamma(0)\phi^n) \\
&= \gamma(0)[1 - \phi^2] \\
&= \frac{\sigma^2}{1 - \phi^2}(1 - \phi^2) \\
&= \sigma^2.
\end{aligned}$$

Linear Processes

We've discussed linear predictors, in which future values of the time series are predicted by a linear combination of "historical values." We now focus on the class of linear processes (time series), which provides a general framework for studying stationary processes.

Definition 18.5 The time series $\{X_t\}$ is called a **linear process** if

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$$

for all t, where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ and $\{\psi_j\}$ is a sequence of constants that converges absolutely, in the sense that

$$\sum_{j=-\infty}^{\infty} |\psi_j| < \infty.$$

Example 18.6 Show that an AR(1) process with $|\phi| < 1$ is a linear process.

Solution. We have already shown that if $X_t = \phi X_{t-1} + Z_t$, then

$$X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}.$$

Thus, if we define $\psi_j = \phi^j$ for $j \geq 0$, then

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}.$$

Moreover,

$$\sum_{j=0}^{\infty} |\psi_j| = \sum_{j=0}^{\infty} |\phi|^j = \frac{1}{1-|\phi|} < \infty.$$

□

In theory, a linear process at time t depends on both future and historical values relative to time

t. We are interested in dependence on the past only. However,

$$\sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$$

involves future innovations such as $Z_{t+1}, Z_{t+2}, \ldots$.

Definition 18.7 A linear process

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$$

is called **causal** if $\psi_j = 0$ for all $j < 0$.

Proposition 18.8 If $\{X_t\}$ is an AR(1) process

$$X_t = \phi X_{t-1} + Z_t, \quad |\phi| < 1,$$

then $\{X_t\}$ is causal.

Proof. Iterating the recursion gives, for each $m \geq 1$,

$$X_t = \phi^m X_{t-m} + \sum_{j=0}^{m-1} \phi^j Z_{t-j}.$$

For the stationary AR(1) solution with $|\phi| < 1$,

$$\mathbb{E}[(\phi^m X_{t-m})^2] = \phi^{2m}\gamma(0) \longrightarrow 0,$$

so $\phi^m X_{t-m} \to 0$ in mean square and

$$X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}.$$

Hence X_t is a linear process with coefficients

$$\psi_j = \phi^j \ (j \geq 0) \qquad \text{and} \qquad \psi_j = 0 \ (j < 0).$$

Also,

$$\sum_{j=0}^{\infty} |\psi_j| = \sum_{j=0}^{\infty} |\phi|^j = \frac{1}{1-|\phi|} < \infty.$$

Therefore the representation is absolutely summable and uses only present/past innovations, so the process is causal. □

The MA(q) process is also causal. For the MA(q) process,

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

so we can set

$$\psi_0 = 1, \qquad \psi_j = \theta_j \text{ for } j = 1, 2, \ldots, q, \qquad \text{and} \qquad \psi_j = 0 \text{ otherwise.}$$

Then

$$X_t = \sum_{j=0}^{q} \psi_j Z_{t-j} \qquad \text{and} \qquad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty.$$

4. ARMA Models

Box–Jenkins Models

The **Box–Jenkins methodology** uses ARMA and ARIMA models for forecasting. It tries to balance goodness-of-fit with a limited number of parameters. The class of ARIMA models (ARMA with differencing) is used when the time series is not stationary.

When seasonal behavior exists, the more general SARIMA models are used.

All of these models rely on two key functions: ACF and PACF (to be discussed).

Tutorial 3 — Thursday, October 30, 2014

Definition 3.1 $\{X_t\}$ is an **ARMA**(p, q) process if the following hold:

1. $\{X_t\}$ is stationary.
2. $$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$
   for all t, where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$.
3. The autoregressive and moving-average polynomials
   $$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \qquad \text{and} \qquad \theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$$
   have no common factors.
4. $\phi_p \neq 0$ when $p \geq 1$, and $\theta_q \neq 0$ when $q \geq 1$, so that the displayed orders are exact.

We say that $\{X_t, t \in \mathbb{T}\}$ is an ARMA process with mean μ if $\{X_t - \mu\}$ is an ARMA process.

Recall the backshift operator: $BX_t = X_{t-1}$ and $B^j X_t = X_{t-j}$. Using this operator, we can represent an ARMA(p, q) process as

$$\phi(B)X_t = \theta(B)Z_t,$$

where

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \qquad \text{and} \qquad \theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q.$$

This general model has a unique stationary solution for X_t if $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0$ for every $z \in \mathbb{C}$ with $|z| = 1$. In this case, there exists $\delta > 0$ such that the Laurent series

$$\frac{1}{\phi(z)} = \sum_{j=-\infty}^{\infty} x_j z^j,$$

converges for $|z| \in (1-\delta, 1+\delta)$, and

$$\sum_{j=-\infty}^{\infty} |x_j| < \infty.$$

Under this condition,

$$\frac{1}{\phi(B)} = \sum_{j=-\infty}^{\infty} x_j B^j$$

is a **linear filter**. Substituting this into the model, we get

$$\phi(B)X_t = \theta(B)Z_t \implies X_t = \frac{\theta(B)}{\phi(B)} Z_t \implies X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j},$$

a linear process, since

$$\sum_{j=-\infty}^{\infty} |x_j| < \infty \implies \sum_{j=-\infty}^{\infty} |\psi_j| < \infty.$$

An ARMA(p,q) process $\phi(B)X_t = \theta(B)Z_t$ is said to be **causal** if there exists a sequence of constants $\{\psi_j\}$ such that

$$\sum_{j=0}^{\infty} |\psi_j| < \infty$$

and

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$$

for all t. This condition is equivalent to $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0$ when $|z| \leq 1$. In other words, the process is causal if the roots of $\phi(z)$ lie outside the unit circle.

Since $X_t = (\theta(B)/\phi(B))Z_t$, we write $\psi(z) = \theta(z)/\phi(z)$. Equivalently,

$$\theta(z) = \phi(z)\psi(z),$$

so for all z we have

$$1 + \theta_1 z + \cdots + \theta_q z^q = (1 - \phi_1 z - \cdots - \phi_p z^p)(\psi_0 + \psi_1 z + \psi_2 z^2 + \dots).$$

If $\phi(z) = 1$, then $\phi(B)X_t = \theta(B)Z_t$ reduces to

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

which is an MA(q) process. If $\theta(z) = 1$, then the equation reduces to

$$\phi(B)X_t = Z_t,$$

which is an AR(p) process. Thus, MA(q) and AR(p) are both special cases of ARMA processes.

Definition 3.2 An ARMA(p, q) process is **invertible** if there exist constants $\{\pi_j\}$ such that

$$\sum_{j=0}^{\infty} |\pi_j| < \infty$$

and

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$$

for all t.

Invertibility is equivalent to the condition that $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \neq 0$ for all z such that $|z| \leq 1$.

To obtain an invertible solution, we write

$$\frac{\phi(z)}{\theta(z)} = \pi(z),$$

or equivalently

$$\phi(z) = \theta(z)\pi(z) \implies (1 - \phi_1 z - \cdots - \phi_p z^p) = (1 + \theta_1 z + \cdots + \theta_q z^q)(\pi_0 + \pi_1 z + \dots).$$

Example 3.3 Consider $X_t - 0.5X_{t-1} = Z_t + 0.4Z_{t-1}$, where $Z_t \sim \text{WN}(0, \sigma^2)$. Investigate causality and invertibility of X_t.

Solution. Observe that $\phi(z) = 1 - 0.5z$ has only the root $z = 2$. Since $|z| = 2 > 1$, a causal solution exists.

Moreover, $\theta(z) = 1 + 0.4z$ has the root $z = -1/0.4 = -2.5$. Since $|z| = 2.5 > 1$, an invertible

solution exists here too. □

Lecture 19 — Tuesday, November 04, 2014

For an ARMA process $\phi(B)X_t = \theta(B)Z_t$, stationarity, causality, and invertibility correspond to three distinct root conditions:

1. The process is stationary if all roots of $\phi(z)$ lie off the unit circle, that is, if $\phi(z_i) = 0$ implies $|z_i| \neq 1$ for all i.
2. The process is causal if all roots of $\phi(z)$ lie outside the unit circle, that is, if $\phi(z_i) = 0$ implies $|z_i| > 1$ for all i.
3. The process is invertible if all roots of $\theta(z)$ lie outside the unit circle, that is, if $\theta(z_i) = 0$ implies $|z_i| > 1$ for all i.

The root-location test is pictured in Figure 14. The plotted roots are schematic: causality concerns the AR roots, invertibility concerns the MA roots, and the unit circle is the boundary in both tests.

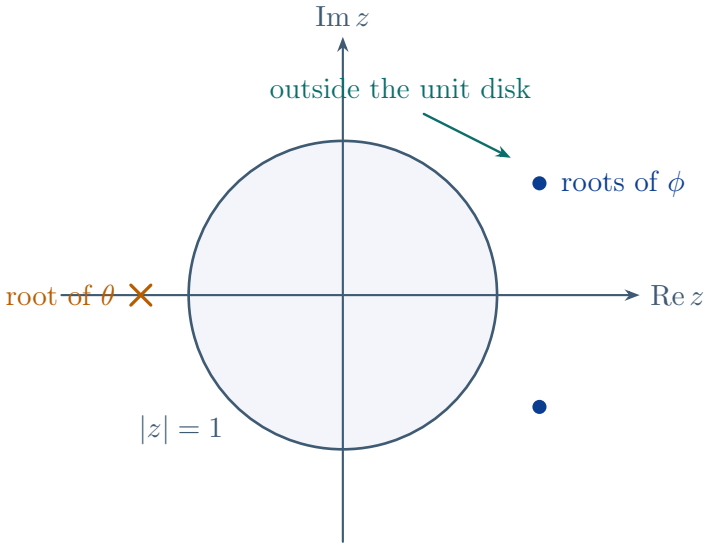


FIGURE 14

Example 19.1 Consider $X_t - 0.5X_{t-1} = Z_t + 0.4Z_{t-1}$, where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. Show that this model is both causal and invertible, and find the solutions for each.

Solution. We begin by checking causality and invertibility through the roots of the AR and MA polynomials:

$$\phi(z) = 1 - 0.5z \qquad \text{and} \qquad \theta(z) = 1 + 0.4z.$$

The unique root of ϕ is $z = 2$, so $|z| > 1$, and therefore the model is causal. The unique root of θ is $z = -2.5$, so $|z| > 1$, and therefore the model is invertible.

Next, we derive the causal MA(∞) representation by writing

$$\psi(z) = \frac{\theta(z)}{\phi(z)} \qquad \text{and} \qquad \theta(z) = \phi(z)\psi(z),$$

which gives

$$1 + 0.4z = (1 - 0.5z)(\psi_0 + \psi_1 z + \psi_2 z^2 + \cdots).$$

Matching coefficients yields

$$\psi_0 = 1, \qquad \psi_1 - 0.5\psi_0 = 0.4, \qquad \text{and} \qquad \psi_j - 0.5\psi_{j-1} = 0 \ (j \geq 2).$$

Hence

$$\psi_1 = 0.9 \qquad \text{and} \qquad \psi_j = 0.5\,\psi_{j-1} = 0.9(0.5)^{j-1} \quad (j \geq 1).$$

Therefore, the causal solution is

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} = Z_t + \sum_{j=1}^{\infty} 0.9(0.5)^{j-1} Z_{t-j}.$$

Finally, we derive the invertible AR(∞) representation by writing

$$\pi(z) = \frac{\phi(z)}{\theta(z)} \qquad \text{and} \qquad \phi(z) = \theta(z)\pi(z),$$

which gives

$$1 - 0.5z = (1 + 0.4z)(\pi_0 + \pi_1 z + \pi_2 z^2 + \cdots).$$

Matching coefficients yields

$$\pi_0 = 1, \qquad \pi_1 + 0.4\pi_0 = -0.5, \qquad \text{and} \qquad \pi_j + 0.4\pi_{j-1} = 0 \ (j \geq 2).$$

Hence

$$\pi_1 = -0.9 \qquad \text{and} \qquad \pi_j = -0.4\,\pi_{j-1} = -0.9(-0.4)^{j-1} \quad (j \geq 1).$$

Therefore, the invertible solution is

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} = X_t + \sum_{j=1}^{\infty} \big[-0.9(-0.4)^{j-1}\big] X_{t-j}.$$

□

Example 19.2 Consider $X_t - 0.7X_{t-1} + 0.1X_{t-2} = Z_t$, where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. Investigate the causality and invertibility of this process.

Solution. We first examine causality and invertibility through the AR and MA polynomials:

$$\phi(z) = 1 - 0.7z + 0.1z^2 = (1 - 0.5z)(1 - 0.2z) \qquad \text{and} \qquad \theta(z) = 1.$$

The roots of ϕ are $z = 2$ and $z = 5$, and both satisfy $|z| > 1$, so the process is causal. Since $\theta(z) = 1 \neq 0$ for every z, the process is also invertible.

To obtain the causal representation, write

$$\theta(z) = \phi(z)\psi(z) \implies 1 = (1 - 0.7z + 0.1z^2)(\psi_0 + \psi_1 z + \psi_2 z^2 + \cdots).$$

Comparing coefficients gives

$$\psi_0 = 1, \qquad \psi_1 - 0.7\psi_0 = 0, \qquad \text{and} \qquad \psi_j - 0.7\psi_{j-1} + 0.1\psi_{j-2} = 0 \quad (j \geq 2).$$

Hence $\psi_1 = 0.7$, $\psi_2 = 0.39$, and in general

$$\psi_j = 0.7\psi_{j-1} - 0.1\psi_{j-2} \quad (j \geq 2).$$

Therefore a causal solution is

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}.$$

To obtain the invertible representation, write

$$\phi(z) = \theta(z)\pi(z) \implies 1 - 0.7z + 0.1z^2 = 1(\pi_0 + \pi_1 z + \pi_2 z^2 + \cdots).$$

Thus

$$\pi_0 = 1, \qquad \pi_1 = -0.7, \qquad \pi_2 = 0.1, \qquad \text{and} \qquad \pi_j = 0 \text{ for all } j \geq 3.$$

Hence

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} = X_t - 0.7X_{t-1} + 0.1X_{t-2},$$

which is exactly the given model equation. □

Pure AR models are automatically invertible because $\theta(z) = 1$, and pure MA models are automatically causal because $\phi(z) = 1$. For a general ARMA model, causality and invertibility are separate properties and must be checked through the roots of ϕ and θ, respectively.

Example 19.3 Consider $X_t - 0.75X_{t-1} + 0.5625X_{t-2} = Z_t + 1.25Z_{t-1}$, where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. Investigate the causality and invertibility of this process.

Solution. To determine causality, we examine the AR polynomial

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

Its roots are complex conjugates, and each has modulus $4/3 > 1$. Therefore all roots of ϕ lie outside the unit circle, so the process is causal.

To determine invertibility, we examine the MA polynomial

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

Solving $\theta(z) = 0$ gives $z = -0.8$, so $|z| = 0.8 < 1$. Therefore a root of θ lies inside the unit circle, and the process is not invertible. □

Proposition 19.4 Suppose $\phi(B)X_t = \theta(B)Z_t$ is a causal ARMA(p, q) model with $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, and let

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$$

be its causal representation. Then $\mathbb{E}[X_t] = 0$, and the autocovariance function satisfies

$$\gamma(h) = \mathrm{Cov}(X_t, X_{t+h}) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}, \quad h \in \mathbb{Z}.$$

Equivalently,

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h} \quad (h \geq 0),$$

and

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j-h} = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|} \quad (h < 0).$$

Proof. Because $\mathbb{E}[Z_t] = 0$, the causal representation gives

$$\mathbb{E}[X_t] = \sum_{j=0}^{\infty} \psi_j \mathbb{E}[Z_{t-j}] = 0.$$

For any integer h,

$$\gamma(h) = \mathbb{E}[X_t X_{t+h}] = \mathbb{E}\left[\sum_{j=0}^{\infty} \psi_j Z_{t-j} \sum_{k=0}^{\infty} \psi_k Z_{t+h-k}\right] = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \psi_j \psi_k \, \mathbb{E}[Z_{t-j} Z_{t+h-k}].$$

Since $\{Z_t\}$ is white noise,

$$\mathbb{E}[Z_i Z_\ell] = \begin{cases} \sigma^2, & i = \ell, \\ 0, & i \neq \ell. \end{cases}$$

Hence only terms with $t - j = t + h - k$, or equivalently $k = j + h$, contribute.

If $h \geq 0$, then $k = j + h \geq 0$ for every $j \geq 0$, so

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}.$$

If $h < 0$, write $m = -h > 0$. The condition $k = j - m \geq 0$ implies $j \geq m$, so

$$\gamma(h) = \sigma^2 \sum_{j=m}^{\infty} \psi_j \psi_{j-m}.$$

Reindexing with $r = j - m$ gives

$$\gamma(h) = \sigma^2 \sum_{r=0}^{\infty} \psi_{r+m}\psi_r = \sigma^2 \sum_{r=0}^{\infty} \psi_r \psi_{r+|h|}.$$

Combining the two cases yields

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}.$$

□

Example 19.5 Find the ACVF for $X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}$, where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$.

Solution. Since $\phi(z) = 1 - \phi z$, the AR polynomial has root $z = 1/\phi$, and the causal case is $|\phi| < 1$. We therefore work with the causal representation $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$.

To find $\{\psi_j\}$, write

$$\theta(z) = \phi(z)\psi(z) \implies 1 + \theta z = (1 - \phi z)(\psi_0 + \psi_1 z + \psi_2 z^2 + \cdots).$$

Comparing coefficients gives

$$\psi_0 = 1, \qquad \psi_1 = \phi + \theta, \qquad \text{and} \qquad \psi_j = \phi^{j-1}(\phi + \theta) \text{ for } j \geq 1.$$

Using $\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}$, we first compute $\gamma(0)$:

$$\gamma(0) = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 = \sigma^2 \left[1 + \sum_{j=1}^{\infty} \phi^{2(j-1)}(\phi + \theta)^2\right] = \sigma^2 \left[1 + (\phi + \theta)^2 \frac{1}{1 - \phi^2}\right].$$

Now let $h \geq 1$. Then

$$\begin{aligned}\gamma(h) &= \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h} = \sigma^2 \left[\psi_h + \sum_{j=1}^{\infty} \psi_j \psi_{j+h}\right] \\ &= \sigma^2 \left[\phi^{h-1}(\phi + \theta) + \sum_{j=1}^{\infty} \phi^{j-1}(\phi + \theta)\, \phi^{j+h-1}(\phi + \theta)\right].\end{aligned}$$

which simplifies to

$$\gamma(h) = \sigma^2 \left[\phi^{h-1}(\phi + \theta) + \frac{(\phi + \theta)^2 \phi^h}{1 - \phi^2} \right].$$

In particular,

$$\gamma(1) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+1} = \sigma^2 \left[\phi + \theta + (\phi + \theta)^2 \frac{\phi}{1 - \phi^2} \right] = \sigma^2 \left[\phi + \theta + \frac{(\phi + \theta)^2 \phi}{1 - \phi^2} \right].$$

Therefore $\gamma(h) = \phi^{h-1}\gamma(1)$ for $h \geq 1$, and the ACVF is

$$\gamma(h) = \begin{cases} \sigma^2 \left[1 + \dfrac{(\phi + \theta)^2}{1 - \phi^2} \right], & h = 0, \\ \sigma^2 \left[\phi + \theta + \dfrac{(\phi + \theta)^2 \phi}{1 - \phi^2} \right] \phi^{|h|-1}, & |h| \geq 1. \end{cases}$$

□

Example 19.6 Derive the ACVF of an AR(1) process using an ARMA process.

Solution. Write the AR(1) model as

$$X_t - \phi X_{t-1} = Z_t, \qquad |\phi| < 1, \qquad \text{and} \qquad \{Z_t\} \sim \text{WN}(0, \sigma^2).$$

This is an ARMA$(1, 0)$ model with $\phi(B) = 1 - \phi B$ and $\theta(B) = 1$. From $\theta(z) = \phi(z)\psi(z)$, we obtain

$$1 = (1 - \phi z)(\psi_0 + \psi_1 z + \psi_2 z^2 + \cdots),$$

so $\psi_0 = 1$ and $\psi_j = \phi^j$ for $j \geq 1$.

Using the general causal ARMA formula

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|},$$

we get

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \phi^j \phi^{j+|h|} = \sigma^2 \phi^{|h|} \sum_{j=0}^{\infty} \phi^{2j} = \frac{\sigma^2 \phi^{|h|}}{1 - \phi^2}, \quad h \in \mathbb{Z}.$$

Therefore the ACVF of AR(1) is

$$\gamma(h) = \frac{\sigma^2}{1-\phi^2}\,\phi^{|h|}.$$

□

Example 19.7 Derive the ACF of an MA(q) process using the ACF of an ARMA process.

Solution. Consider

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \qquad \text{and} \qquad \{Z_t\} \sim \text{WN}(0, \sigma^2).$$

This is an ARMA$(0, q)$ model, so its causal coefficients are finite:

$$\psi_0 = 1, \qquad \psi_j = \theta_j \ (1 \le j \le q), \qquad \text{and} \qquad \psi_j = 0 \ (j > q).$$

Using

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|},$$

we see immediately that $\gamma(h) = 0$ for $|h| > q$, because one factor is always 0.

For $0 \le h \le q$,

$$\gamma(h) = \sigma^2 \sum_{j=0}^{q-h} \psi_j \psi_{j+h}.$$

Hence

$$\gamma(0) = \sigma^2 \sum_{j=0}^{q} \psi_j^2,$$

and the ACF is

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\sum_{j=0}^{q-|h|} \psi_j \psi_{j+|h|}}{\sum_{j=0}^{q} \psi_j^2}, \quad |h| \le q,$$

with

$$\rho(h) = 0, \quad |h| > q.$$

Therefore the ACF of an MA(q) process cuts off after lag q. □

Since an AR(1) process is an ARMA$(1, 0)$ process, the general ARMA autocovariance calculation recovers the earlier formula $\gamma(h) = \phi^{|h|}\gamma(0)$. Likewise, for an MA($q$) process, the autocovariance

function vanishes for $|h| > q$, as shown earlier.

An AR(p) process and an ARMA(p, q) process with $p > 0$ can both produce ACFs that tail off, decaying to 0 without reaching it, rather than cut off at a finite lag. The ACF alone is therefore not enough to separate the two model classes in practice. We need another function that removes the linear effect of intermediate lags and is more informative about autoregressive order.

Tutorial 4 — Thursday, November 06, 2014

For a causal ARMA(p, q) process, the autocovariance function can be written

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}.$$

For an ARMA$(1, 1)$ process satisfying $X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}$, we also saw that

$$\rho(h) = \rho(1)\phi^{h-1} \quad \text{for} h \geq 1,$$

provided that $|\phi| < 1$.

Recall that, for a stationary process, $\rho(h) = \gamma(h)/\gamma(0) = \text{Corr}(X_t, X_{t+h})$ for all h. For the MA(1) model $X_t = Z_t + \theta Z_{t-1}$, the variables X_t and X_{t+1} are both functions of Z_t, so the dependence comes from a shared innovation. For the AR(1) model $X_t = \phi X_{t-1} + Z_t$ with $0 < |\phi| < 1$, we have $\rho(h) = \phi^{|h|}$ for $h = 0, 1, 2, \ldots$, so nonzero correlation persists at every lag through the autoregressive recursion.

The Partial Autocorrelation Function

Definition 4.1 The **partial autocorrelation function** (**PACF**), denoted $\alpha(h)$, is defined to be:

$$\alpha(h) = \begin{cases} 1, & h = 0, \\ \rho(1), & h = 1, \\ \text{Corr}(X_t - \text{Pred}(X_t | X_{t+1}, \ldots, X_{t+h-1}), & \\ \qquad X_{t+h} - \text{Pred}(X_{t+h} | X_{t+1}, \ldots, X_{t+h-1})), & h \geq 2. \end{cases}$$

The ACF measures the correlation between X_t and X_{t+h}. This correlation can be due to a direct connection, or through the intermediate steps $X_{t+1}, X_{t+2}, \ldots, X_{t+h-1}$. The PACF, on the other hand, looks at the correlation between X_t and X_{t+h} once the effect of the intermediate steps is removed. We measure the effect of intermediate steps by $\text{Pred}(X_{t+h}|X_{t+1}, \ldots, X_{t+h-1})$ and $\text{Pred}(X_t|X_{t+1}, \ldots, X_{t+h-1})$, both linear predictors.

Example 4.2 Derive the PACF of the AR(1) process

$$X_t = \phi X_{t-1} + Z_t, \qquad |\phi| < 1, \qquad \text{and} \qquad \{Z_t\} \sim \text{WN}(0, \sigma^2).$$

Solution. For $h = 0$, the definition gives $\alpha(0) = 1$. For $h = 1$, we have

$$\alpha(1) = \text{Corr}(X_t, X_{t+1}) = \rho(1) = \phi.$$

Now consider $h \geq 2$. By definition,

$$\alpha(h) = \text{Corr}\Big(X_t - \text{Pred}(X_t \mid X_{t+1}, \ldots, X_{t+h-1}),\, X_{t+h} - \text{Pred}(X_{t+h} \mid X_{t+1}, \ldots, X_{t+h-1})\Big).$$

For an AR(1) process, the one-step linear predictor is $\text{Pred}(X_{t+h} \mid X_{t+1}, \ldots, X_{t+h-1}) = \phi X_{t+h-1}$, so

$$X_{t+h} - \text{Pred}(X_{t+h} \mid X_{t+1}, \ldots, X_{t+h-1}) = X_{t+h} - \phi X_{t+h-1} = Z_{t+h}.$$

Therefore

$$\alpha(h) = \text{Corr}\Big(X_t - \text{Pred}(X_t \mid X_{t+1}, \ldots, X_{t+h-1}),\, Z_{t+h}\Big).$$

The first term is a linear function of $X_t, \ldots, X_{t+h-1}$, and because the AR(1) is causal, it is a function of innovations up to time $t+h-1$. Since Z_{t+h} is uncorrelated with all earlier innovations, this correlation is 0. Hence $\alpha(h) = 0$ for all $h \geq 2$.

We conclude that

$$\alpha(h) = \begin{cases} 1, & h = 0, \\ \phi, & h = 1, \\ 0, & h \geq 2. \end{cases}$$

□

Theorem 4.3 Suppose the finite covariance matrices of the weakly stationary process $\{X_t\}$ are nonsingular. Then $\{X_t\}$ is a causal AR process of exact order p if and only if its PACF satisfies

$$\alpha(h) = 0 \quad \text{for all } h > p \qquad \text{and} \qquad \alpha(p) \neq 0.$$

Moreover, if

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t$$

with $\phi_p \neq 0$, then

$$\alpha(p) = \phi_p.$$

The proof is beyond the scope of the course.

	ACF	PACF
AR(p)	tails off	cuts off at lag p
MA(q)	cuts off at lag q	tails off
ARMA(p, q)	tails off	tails off

Lecture 20 — Tuesday, November 11, 2014

Model Identification with PACFs

The PACF of a causal AR(p) process cuts off after lag p: $\alpha(p) \neq 0$, $\alpha(h) = 0$ for $h > p$, and $\alpha(p) = \phi_p$.

Individual coefficient magnitudes do not by themselves settle stationarity for an AR(p) model when $p > 1$; we must check the roots of $\phi(z)$. For an AR(1) model, this reduces to the familiar condition $|\phi_1| < 1$.

Theorem 4.3 shows that the PACF is a powerful tool for identifying AR(p) processes. In visual terms, the ACF plays for MA(q) the same role that the PACF plays for AR(p). The summary table above captures this analogy.

Like the ACF, the PACF is an even function, that is, $\alpha(-h) = \alpha(h)$. Therefore, we only need to consider $h \geq 0$.

When reading ACF and PACF plots, values inside the confidence bands are usually treated as negligible. The basic heuristics are the familiar ones: an ACF that cuts off after lag 1 with a gradually decaying PACF suggests MA(1); a cutoff after lag 2 suggests MA(2); a cutoff after lag 3 suggests MA(3); and an exponentially decaying ACF together with a PACF that cuts off after lag 1 suggests AR(1). When both ACF and PACF decay gradually, an ARMA$(1,1)$ model is often a reasonable first candidate. As always, these graphical rules are heuristics rather than proofs.

The complementary cutoff patterns for first-order autoregressive and moving-average models are shown in Figure 15. The shaded bands indicate values that would ordinarily be treated as sampling noise.

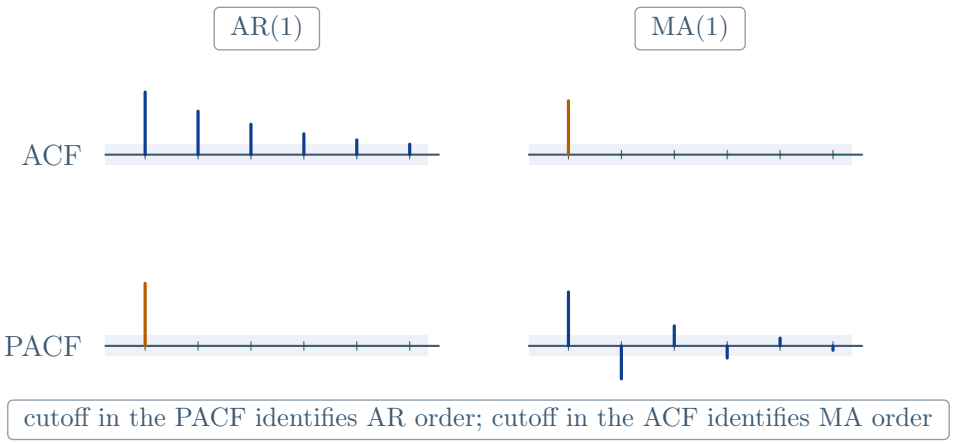


FIGURE 15

The theoretical cutoff rules are easiest to recognize after seeing how sampling variability perturbs them. Figure 16 uses 200 simulated observations from an MA(2) process and an AR(1) process. The MA(2) sample ACF becomes negligible after lag 2, while its PACF tails off; for the AR(1) series, the roles are reversed. A fixed random seed makes the exercise exactly reproducible.

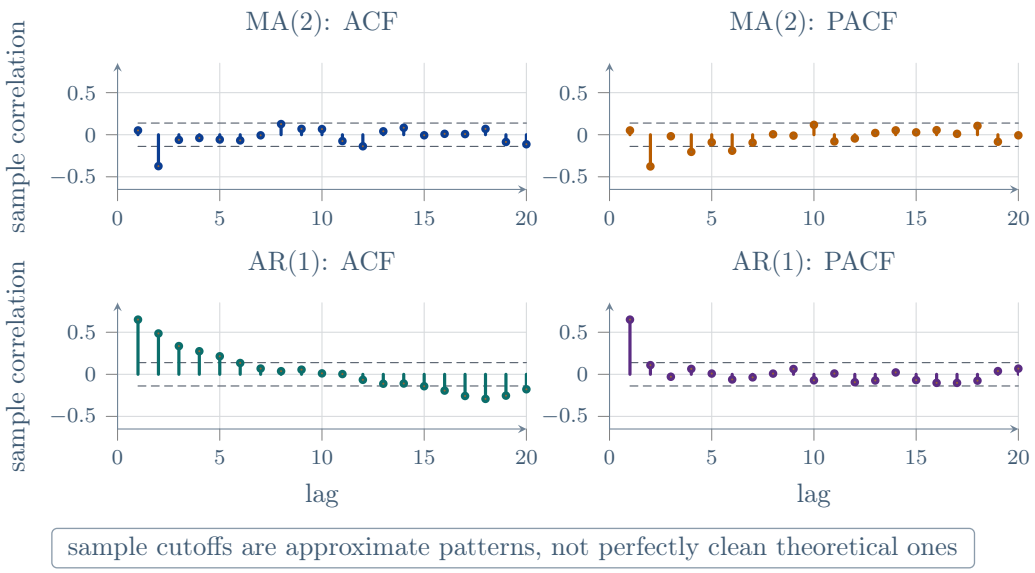


FIGURE 16

Here is the code that generates the two series and their diagnostic plots.

```
set.seed(443)

ma2 <- arima.sim(model = list(ma = c(0.7, -1)), n = 200)
ar1 <- arima.sim(model = list(ar = 0.7), n = 200)

par(mfrow = c(2, 2))
acf(ma2, lag.max = 20, main = "MA(2): sample ACF")
pacf(ma2, lag.max = 20, main = "MA(2): sample PACF")
acf(ar1, lag.max = 20, main = "AR(1): sample ACF")
pacf(ar1, lag.max = 20, main = "AR(1): sample PACF")
```

In the general case of ARMA processes, the PACF is defined as

$$\alpha(h) = \begin{cases} 1, & h = 0 \\ \phi_{hh}, & h \geq 1 \end{cases}$$

where ϕ_{hh} is the last component of the vector $\boldsymbol{\phi}_h = \boldsymbol{\Gamma}_h^{-1}\boldsymbol{\gamma}_h$, in which

$$\boldsymbol{\Gamma}_h = \begin{pmatrix} \gamma(0) & \dots & \gamma(h-1) \\ \vdots & \ddots & \vdots \\ \gamma(h-1) & \dots & \gamma(0) \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\gamma}_h = \begin{pmatrix} \gamma(1) \\ \vdots \\ \gamma(h) \end{pmatrix}.$$

Notice the similarity between the calculation above and the best linear predictor $\boldsymbol{\Gamma}\boldsymbol{\phi} = \boldsymbol{\gamma}$.

Definition 20.1 Given observations $\{x_1, \dots, x_n\}$, define the sample autocovariances $\hat{\gamma}(h)$, and form

$$\hat{\boldsymbol{\Gamma}}_h = \begin{pmatrix} \hat{\gamma}(0) & \cdots & \hat{\gamma}(h-1) \\ \vdots & \ddots & \vdots \\ \hat{\gamma}(h-1) & \cdots & \hat{\gamma}(0) \end{pmatrix} \qquad \text{and} \qquad \hat{\boldsymbol{\gamma}}_h = \begin{pmatrix} \hat{\gamma}(1) \\ \vdots \\ \hat{\gamma}(h) \end{pmatrix}.$$

When $\hat{\boldsymbol{\Gamma}}_h$ is nonsingular, let $\hat{\boldsymbol{\phi}}_h = \hat{\boldsymbol{\Gamma}}_h^{-1}\hat{\boldsymbol{\gamma}}_h$, and write $\hat{\phi}_{hh}$ for its last component. The **sample PACF** is then defined by

$$\hat{\alpha}(h) = \begin{cases} 1, & h = 0 \\ \hat{\phi}_{hh}, & h \geq 1 \end{cases}$$

Example 20.2 Calculate $\alpha(2)$ for an MA(1) process.

Solution. Let

$$X_t = Z_t + \theta Z_{t-1} \qquad \text{and} \qquad \{Z_t\} \sim \text{WN}(0, \sigma^2).$$

From the MA(1) calculations, we have

$$\gamma(0) = (1+\theta^2)\sigma^2, \qquad \gamma(1) = \theta\sigma^2, \qquad \text{and} \qquad \gamma(h) = 0 \text{ for } h \geq 2.$$

For $h = 2$, the PACF is the last component of

$$\boldsymbol{\phi}_2 = \boldsymbol{\Gamma}_2^{-1}\boldsymbol{\gamma}_2,$$

where

$$\boldsymbol{\Gamma}_2 = \begin{pmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\gamma}_2 = \begin{pmatrix} \gamma(1) \\ \gamma(2) \end{pmatrix} = \begin{pmatrix} \gamma(1) \\ 0 \end{pmatrix}.$$

Using

$$\boldsymbol{\Gamma}_2^{-1} = \frac{1}{\gamma(0)^2 - \gamma(1)^2} \begin{pmatrix} \gamma(0) & -\gamma(1) \\ -\gamma(1) & \gamma(0) \end{pmatrix},$$

we obtain

$$\boldsymbol{\phi}_2 = \frac{1}{\gamma(0)^2 - \gamma(1)^2} \begin{pmatrix} \gamma(0)\gamma(1) \\ -\gamma(1)^2 \end{pmatrix}.$$

Therefore

$$\alpha(2) = \phi_{22} = \frac{-\gamma(1)^2}{\gamma(0)^2 - \gamma(1)^2} = \frac{-\theta^2}{(1+\theta^2)^2 - \theta^2} = \frac{-\theta^2}{1+\theta^2+\theta^4}.$$

Continuing in the same way gives

$$\alpha(h) = \phi_{hh} = \frac{-(-\theta)^h}{1+\theta^2+\theta^4+\cdots+\theta^{2h}}.$$

□

Lecture 21 — Thursday, November 13, 2014

ARIMA and SARIMA Models

For moving-average models, the PACF tails off rather than cutting off. In the MA(1) case, for example, we obtained

$$\alpha(h) = \frac{-(-\theta)^h}{1+\theta^2+\cdots+\theta^{2h}}.$$

ARMA(p,q) is a class of models for stationary time series. A generalization of this class is the ARIMA(p,d,q) process.

Definition 21.1 Let d be a nonnegative integer. The process $\{X_t\}$ is called an **ARIMA(p,d,q) process** if $Y_t := (1-B)^d X_t$ is a causal ARMA(p,q) process and d is the smallest nonnegative integer with this property.

Equivalently, $\{X_t\}$ satisfies an equation of the form $\phi^*(B)X_t = \theta(B)Z_t$, where $\phi^*(B) = \phi(B)(1-B)^d$, $\{Z_t\} \sim \text{WN}(0,\sigma^2)$, and no factor $1-B$ is removable by reducing the differencing order.

Notice that if $d \neq 0$, then $\phi^*(z) = \phi(z)(1-z)^d = 0$ at $z = 1$. Under the minimal-order convention in the definition (so that this unit root is not an artefact of overdifferencing or cancellation), $\{X_t\}$ is stationary if and only if $d = 0$, in which case ARIMA(p,d,q) reduces to ARMA(p,q). In

particular, ARMA(p, q) is the same as ARIMA$(p, 0, q)$.

If $\{X_t\}$ has a polynomial mean of degree d, say $m_t = a_0 + a_1 t + \cdots + a_d t^d$, then $(1 - B)^d X_t$ has a constant mean. If the differenced random fluctuations are stationary as well, an ARIMA model is therefore appropriate.

The first-difference picture in Figure 17 shows the basic effect: a drifting level becomes a series fluctuating around a stable mean.

FIGURE 17

Recall that a smooth, "well-behaved" function or trend can be well-approximated by a polynomial.

Example 21.2 Consider the process $X_t = 0.8X_{t-1} + 2t - Z_t$, where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. Write it as an ARIMA$(p, d, q)$ model.

Solution. Rearranging gives

$$X_t - 0.8X_{t-1} = 2t - Z_t \implies (1 - 0.8B)X_t = 2t - Z_t.$$

Define $W_t := -Z_t$, so $\{W_t\} \sim \text{WN}(0, \sigma^2)$. Then

$$X_t = 0.8X_{t-1} + 2t + W_t.$$

Now difference once:

$$\begin{aligned}\nabla X_t &= X_t - X_{t-1} \\ &= 0.8(X_{t-1} - X_{t-2}) + \big(2t - 2(t-1)\big) + (W_t - W_{t-1}) \\ &= 0.8\nabla X_{t-1} + 2 + W_t - W_{t-1}.\end{aligned}$$

If $Y_t := \nabla X_t$, then

$$Y_t - 0.8Y_{t-1} = 2 + W_t - W_{t-1}.$$

So Y_t is an ARMA$(1,1)$-type process with nonzero mean. Its mean $\mu_Y = \mathbb{E}[Y_t]$ satisfies

$$\mu_Y - 0.8\mu_Y = 2,$$

hence $\mu_Y = 10$. Define $Y_t^* := Y_t - 10$. Then

$$(1 - 0.8B)Y_t^* = (1 - B)W_t,$$

so Y_t^* is a causal ARMA$(1,1)$ process. Therefore X_t is ARIMA$(1,1,1)$. □

Ordinary differencing can remove a slowly varying trend. A separate operation, **seasonal differencing**, removes a pattern that repeats at a fixed seasonal lag.

Recall that the backward shift operator is defined by $BX_t = X_{t-1}$, so $B^k X_t = X_{t-k}$. Note that $(1-B)^2 X_t = X_t - 2X_{t-1} + X_{t-2}$ corresponds to two successive first differences, whereas $(1 - B^2)X_t = X_t - X_{t-2}$ is a single difference at lag 2. These are fundamentally different filters. More generally, $(1-B)^k$ means k successive differences, while $(1-B^k)$ means one difference taken k periods apart.

This distinction is exactly what we need for seasonality. If the data are monthly and the seasonal effect repeats every 12 months, then the seasonal component should be similar at times t and $t - 12$. That motivates the seasonal difference

$$Y_t := X_t - X_{t-12} = (1 - B^{12})X_t,$$

which is designed to remove seasonal trend. Fitting an ARMA(p,q) model to the seasonally differenced series is therefore equivalent to fitting

$$\phi(B)(1 - B^s)X_t = \theta(B)Z_t,$$

which is the simplest seasonal extension of the ARMA framework.

Definition 21.3 If $d, D \in \mathbb{N}_0$, then $\{X_t\}$ is called a **SARIMA$(p,d,q) \times (P,D,Q)_s$ process** if the differenced series

$$Y_t = \nabla^d \nabla_s^D X_t = (1 - B)^d (1 - B^s)^D X_t$$

is a causal ARMA process satisfying

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t,$$

where

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \qquad \text{and} \qquad \Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_P z^P,$$

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \qquad \text{and} \qquad \Theta(z) = 1 + \Theta_1 z + \cdots + \Theta_Q z^Q.$$

The specification is understood to be reduced: no ordinary or seasonal differencing factor can be cancelled, and the differencing orders are not larger than necessary.

One way to think about a SARIMA model is that it simultaneously accounts for short-range dependence within the series and longer-range dependence across matching seasons.

Tutorial 5 — Thursday, November 13, 2014

If $d, D \in \mathbb{N}_0$, then $\{X_t\}$ is a seasonal ARIMA$(p,d,q) \times (P,D,Q)_s$ model with period s when the differenced series

$$Y_t = \nabla^d \nabla_s^D X_t = (1-B)^d (1-B^s)^D X_t$$

is a causal ARMA process satisfying

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t,$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ and

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \qquad \text{and} \qquad \Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_P z^P,$$

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \qquad \text{and} \qquad \Theta(z) = 1 + \Theta_1 z + \cdots + \Theta_Q z^Q.$$

Observe that the process $\{Y_t\}$ is causal if and only if $\phi(z) \neq 0$ and $\Phi(z) \neq 0$ for all z such that $|z| \leq 1$. In practice, D is rarely larger than 1, and P and Q are typically less than 3.

Example 5.1 Write down the equations for an ARMA$(1,1)_{12}$ process. By convention, any component not explicitly mentioned is taken to be 0, so ARMA$(1,1)_{12}$ means

SARIMA$(0,0,0) \times (1,0,1)_{12}$.

Solution. Starting from

$$\phi(B)\Phi(B^s)\nabla^d\nabla_s^D X_t = \theta(B)\Theta(B^s)Z_t,$$

and substituting $d = D = 0$, $\phi(B) = \theta(B) = 1$, $\Phi(B^{12}) = 1 - \Phi_1 B^{12}$, $\Theta(B^{12}) = 1 + \Theta_1 B^{12}$, we get

$$(1 - \Phi_1 B^{12})X_t = (1 + \Theta_1 B^{12})Z_t.$$

Since $d = D = 0$, no differencing is present, so this specification is stationary (under the usual root conditions). □

Example 5.2 Derive the ACF of ARMA$(0,0,1)_{12}$.

Solution. First note that ARMA$(0,0,1)_{12}$ = SARIMA$(0,0,0) \times (0,0,1)_{12}$. For ARMA$(0,0,1)_{12}$, we have $d = D = 0$, $\phi(B) = \Phi(B^s) = \theta(B) = 1$, and $\Theta(B^{12}) = 1 + \Theta_1 B^{12}$. Thus

$$X_t = (1 + \Theta_1 B^{12})Z_t = Z_t + \Theta_1 Z_{t-12}.$$

This process is stationary. Its autocovariance is

$$\gamma(h) = \mathrm{Cov}(X_t, X_{t+h}) = \mathrm{Cov}(Z_t + \Theta_1 Z_{t-12},\, Z_{t+h} + \Theta_1 Z_{t+h-12}),$$

so only shared innovations contribute. Therefore

$$\gamma(h) = \begin{cases} (1 + \Theta_1^2)\sigma^2, & h = 0, \\ \Theta_1\sigma^2, & |h| = 12, \\ 0, & \text{otherwise.} \end{cases}$$

Hence

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \begin{cases} 1, & h = 0, \\ \dfrac{\Theta_1}{1 + \Theta_1^2}, & |h| = 12, \\ 0, & \text{otherwise.} \end{cases}$$

This is analogous to the MA(1) ACF, but at seasonal lag 12. □

Example 5.3 Write down the equations for SARIMA$(0,1,1)\times(0,1,1)_{12}$.

Solution. For SARIMA$(0,1,1)\times(0,1,1)_{12}$, we have

$$p=0,\ d=1,\ q=1, \qquad P=0,\ D=1,\ Q=1, \qquad \text{and} \qquad s=12.$$

Therefore,

$$\phi(B)=1, \qquad \Phi(B^{12})=1, \qquad \theta(B)=1+\theta_1 B, \qquad \text{and} \qquad \Theta(B^{12})=1+\Theta_1 B^{12}.$$

Substituting into

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D X_t=\theta(B)\Theta(B^s)Z_t$$

gives

$$(1-B)(1-B^{12})X_t=(1+\theta_1 B)(1+\Theta_1 B^{12})Z_t.$$

Expanding both sides yields

$$X_t-X_{t-1}-X_{t-12}+X_{t-13}=Z_t+\theta_1 Z_{t-1}+\Theta_1 Z_{t-12}+\theta_1\Theta_1 Z_{t-13},$$

which is the time-domain equation for SARIMA$(0,1,1)\times(0,1,1)_{12}$. □

To use the Box–Jenkins framework, we first assess stationarity, then difference as needed, next identify plausible values of p, q, P, and Q from the ACF and PACF of the differenced series, and finally fit the resulting model for forecasting.

In the worked seasonal example, the data first show nonconstant variance, so a variance-stabilizing transformation such as a logarithm is natural. After that, first differencing removes the ordinary trend, and seasonal differencing removes the remaining yearly pattern. Once both sources of nonstationarity are addressed, the ACF and PACF of the transformed series are used to propose candidate values for p, q, P, and Q:

p	q
1	1
0	1
0	5

P	Q
0	1
1	0

These six combinations give a manageable list of candidate models to compare.

Lecture 22 — Tuesday, November 18, 2014

5. Parameter Estimation and Forecasting

Seasonal differencing leads naturally to models of the form SARIMA$(p,d,q)\times(P,D,Q)_s$. We now turn from model identification to parameter estimation.

Parameter Estimation in ARMA Models and the Yule–Walker Equations

Suppose $\phi(B)X_t = \theta(B)Z_t$, where $\{Z_t\} \sim \text{WN}(0,\sigma^2)$. After proposing a model from the ACF and PACF, we usually assume that the orders p and q are correctly identified and that $\mathbb{E}[X_t]=0$. If the mean is not zero, we first center the series by writing $X_t^* = X_t - \mu$. Standard estimation methods include **likelihood-based methods**, the **Yule–Walker equations**, the **innovations algorithm**, and the **Durbin–Levinson algorithm**. We focus here on the first two.

The likelihood method begins with observed data $x_1, x_2, \dots, x_n$ and treats the joint density as a function of the unknown parameter. For example, if $x_1,\dots,x_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$, then

$$f(x_1,\dots,x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}.$$

The key distinction is that the joint density is naturally written as a function of the observations, whereas the likelihood treats the observations as fixed and varies the parameter. This requires a distributional model for the vector of observations $(X_1,\dots,X_n)$.

For a zero-mean Gaussian process satisfying $\phi(B)X_t = \theta(B)Z_t$, let $\boldsymbol{x} = (x_1,\dots,x_n)^\top$ and let $\boldsymbol{\Gamma}_n$ be its covariance matrix. The likelihood for the AR parameters, MA parameters, and innovation variance is

$$L(\phi,\theta,\sigma^2) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Gamma}_n|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Gamma}_n^{-1}\boldsymbol{x}\right),$$

where

$$\boldsymbol{\Gamma}_n = \begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \dots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \dots & \gamma(n-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \dots & \dots & \gamma(0) \end{bmatrix}.$$

We maximize L with respect to the parameters in ϕ, the parameters in θ, and σ^2. The calculation is nonlinear, and the dimension of $\mathbf{\Gamma}_n$ grows with n.

For a causal AR(p) process,

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t \qquad \text{and} \qquad \{Z_t\} \sim \text{WN}(0, \sigma^2),$$

the **Yule–Walker equations** are obtained by multiplying both sides by X_{t-k} and taking expectations for $k = 0, 1, \ldots, p$:

$$\mathbb{E}[X_t X_{t-k}] - \phi_1 \mathbb{E}[X_{t-1} X_{t-k}] - \cdots - \phi_p \mathbb{E}[X_{t-p} X_{t-k}] = \mathbb{E}[X_{t-k} Z_t].$$

By stationarity, $\mathbb{E}[X_t X_{t-k}] = \gamma(k)$. Also, for a causal AR process,

$$\mathbb{E}[X_{t-k} Z_t] = \begin{cases} \sigma^2, & k = 0, \\ 0, & k \geq 1. \end{cases}$$

Hence,

$$\gamma(0) - \phi_1 \gamma(1) - \cdots - \phi_p \gamma(p) = \sigma^2,$$

and for $k = 1, \ldots, p$,

$$\gamma(k) - \phi_1 \gamma(k-1) - \cdots - \phi_p \gamma(k-p) = 0.$$

These are the Yule–Walker equations.

Define

$$\boldsymbol{\phi} = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \end{pmatrix}, \qquad \boldsymbol{\gamma}_p = \begin{pmatrix} \gamma(1) \\ \vdots \\ \gamma(p) \end{pmatrix}, \qquad \text{and} \qquad \mathbf{\Gamma}_p = \begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(0) \end{pmatrix}.$$

Then

$$\mathbf{\Gamma}_p \boldsymbol{\phi} = \boldsymbol{\gamma}_p \qquad \text{and} \qquad \sigma^2 = \gamma(0) - \boldsymbol{\phi}^\top \boldsymbol{\gamma}_p.$$

In practice, replace $\gamma(\cdot)$ by $\hat{\gamma}(\cdot)$. The **sample Yule–Walker equations** are

$$\hat{\mathbf{\Gamma}}_p \hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\gamma}}_p \qquad \text{and} \qquad \hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\phi}}^\top \hat{\boldsymbol{\gamma}}_p.$$

Equivalently, in ACF form,

$$\hat{\boldsymbol{\phi}} = \hat{\mathbf{R}}_p^{-1} \hat{\boldsymbol{\rho}}_p \qquad \text{and} \qquad \hat{\sigma}^2 = \hat{\gamma}(0) \left(1 - \hat{\boldsymbol{\phi}}^\top \hat{\boldsymbol{\rho}}_p\right),$$

where

$$\hat{\boldsymbol{\rho}}_p = \begin{pmatrix} \hat{\rho}(1) \\ \vdots \\ \hat{\rho}(p) \end{pmatrix} \qquad \text{and} \qquad \hat{\mathbf{R}}_p = \begin{pmatrix} 1 & \hat{\rho}(1) & \hat{\rho}(2) & \cdots & \hat{\rho}(p-1) \\ \hat{\rho}(1) & 1 & \hat{\rho}(1) & \cdots & \hat{\rho}(p-2) \\ \hat{\rho}(2) & \hat{\rho}(1) & 1 & \cdots & \hat{\rho}(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}(p-1) & \hat{\rho}(p-2) & \hat{\rho}(p-3) & \cdots & 1 \end{pmatrix}.$$

Under standard regularity conditions, such as iid innovations with a finite fourth moment, asymptotic theory gives

$$\hat{\boldsymbol{\phi}} \overset{\text{a}}{\sim} \mathcal{N}_p\left(\boldsymbol{\phi}, \frac{\sigma^2}{n}\boldsymbol{\Gamma}_p^{-1}\right),$$

which can be used to construct approximate confidence intervals for $\phi_1, \phi_2, \ldots, \phi_p$.

Example 22.1 For an AR(p) process, find the approximate distribution of $\hat{\boldsymbol{\phi}}$ and a 95% confidence interval for $\phi_1, \ldots, \phi_p$, given $n = 200$, $\widehat{\text{Var}}(X_t) = 3.69$, $\hat{\mathbf{R}}_2 = \begin{pmatrix} 1 & 0.821 \\ 0.821 & 1 \end{pmatrix}$, $\hat{\boldsymbol{\rho}}_2 = \begin{pmatrix} 0.821 \\ 0.764 \end{pmatrix}$, and sample autocovariance information suggesting an AR(2) fit.

Solution. Because the ACF decays and the PACF cuts off after lag 2, we take $p = 2$. Then

$$\hat{\boldsymbol{\phi}} = \hat{\mathbf{R}}_2^{-1}\hat{\boldsymbol{\rho}}_2 = \begin{pmatrix} 1 & 0.821 \\ 0.821 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0.821 \\ 0.764 \end{pmatrix} \approx \begin{pmatrix} 0.5944 \\ 0.2760 \end{pmatrix}.$$

Thus $\hat{\phi}_1 \approx 0.5944$ and $\hat{\phi}_2 \approx 0.2760$. Next, since $\hat{\gamma}(0) = \widehat{\text{Var}}(X_t) = 3.69$,

$$\hat{\sigma}^2 = \hat{\gamma}(0)\left[1 - \hat{\boldsymbol{\phi}}^\top \hat{\boldsymbol{\rho}}_2\right] = 3.69\left[1 - (0.5944 \cdot 0.821 + 0.2760 \cdot 0.764)\right] \approx 1.1112.$$

So the fitted model is

$$X_t - 0.5944X_{t-1} - 0.2760X_{t-2} = Z_t \qquad \text{and} \qquad \{Z_t\} \sim \text{WN}(0, 1.1112).$$

Also,

$$\hat{\boldsymbol{\Gamma}}_2^{-1} = \frac{1}{\hat{\gamma}(0)}\hat{\mathbf{R}}_2^{-1} = \frac{1}{3.69}\begin{pmatrix} 1 & 0.821 \\ 0.821 & 1 \end{pmatrix}^{-1},$$

which gives

$$\frac{\hat{\sigma}^2}{n}\hat{\mathbf{\Gamma}}_2^{-1} = \begin{pmatrix} 0.00462 & -0.00379 \\ -0.00379 & 0.00462 \end{pmatrix}.$$

Hence

$$\hat{\boldsymbol{\phi}} \approx \mathcal{N}_2\left(\begin{pmatrix} 0.5944 \\ 0.2760 \end{pmatrix}, \begin{pmatrix} 0.00462 & -0.00379 \\ -0.00379 & 0.00462 \end{pmatrix}\right).$$

Using marginal normal approximations,

$$\phi_1 : 0.5944 \pm 1.96\sqrt{0.00462} \approx (0.461,\, 0.728),$$
$$\phi_2 : 0.2760 \pm 1.96\sqrt{0.00462} \approx (0.143,\, 0.409).$$

□

Warning The rectangle $(0.461,\, 0.728) \times (0.143,\, 0.409)$ is *not* a joint confidence region for $\boldsymbol{\phi}$.

Lecture 23 — Thursday, November 20, 2014

Once a model has been identified and estimated, we can turn it into forecasts.

Forecasting in ARMA Models

Among all functions of the observed history, the minimum-mean-square predictor is

$$\mathbb{E}\left[X_{n+h} \mid X_n, X_{n-1}, \ldots, X_1\right].$$

In the ARMA calculations below, $P_n X_{n+h}$ denotes the best *linear* predictor based on $X_1, \ldots, X_n$. For Gaussian processes it equals the conditional expectation above. With only white-noise assumptions, the linear-projection statement is the one we can guarantee. We write $\hat{X}_{n+h} = P_n X_{n+h}$.

Example 23.1 Derive the h-step-ahead predictor for a causal AR(p) process of the form $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t$ with $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$.

Solution. For a causal AR(p) model,

$$X_{n+h} = \sum_{j=1}^{p} \phi_j X_{n+h-j} + Z_{n+h}.$$

Linearity of projection and the orthogonality of the future innovation Z_{n+h} to the observed history give

$$\hat{X}_{n+h} = \sum_{j=1}^{p} \phi_j P_n X_{n+h-j}.$$

Using the convention $P_n X_k = X_k$ for $k \leq n$, this single recursion covers every horizon. When $h = 1$, all lagged terms X_{n+1-j} are observed, so

$$\hat{X}_{n+1} = \sum_{j=1}^{p} \phi_j X_{n+1-j}.$$

For $h = 2, \ldots, p$, part of the right-hand side uses future values and part uses observed values. Splitting those terms yields

$$\hat{X}_{n+h} = \sum_{j=1}^{h-1} \phi_j \hat{X}_{n+h-j} + \sum_{j=h}^{p} \phi_j X_{n+h-j}.$$

For $h > p$, all terms involve future values only, so the recursion becomes

$$\hat{X}_{n+h} = \sum_{j=1}^{p} \phi_j \hat{X}_{n+h-j}.$$

Therefore AR(p) forecasts are obtained recursively from the model coefficients, recent observations, and previously computed forecasts. □

The main point is that AR(p) forecasting is recursive and linear in the most recent observations and forecasts.

Example 23.2 Suppose annual sales data for a chain are adequately described by an AR(2) model. In particular,

$$\hat{\phi}_1 = 1, \qquad \hat{\phi}_2 = -0.21, \qquad \text{and} \qquad \hat{\sigma}^2 = 1.78.$$

If sales were 10 million dollars in 2010, 11 million dollars in 2011, and 9 million dollars in 2012, find the forecast for 2015.

Solution. The fitted model is

$$X_t = X_{t-1} - 0.21X_{t-2} + Z_t.$$

Using the AR(2) recursion,

$$\begin{aligned}
\hat{X}_{2013} &= \hat{\phi}_1 X_{2012} + \hat{\phi}_2 X_{2011} = 1 \cdot 9 - 0.21 \cdot 11 = 6.69, \\
\hat{X}_{2014} &= \hat{\phi}_1 \hat{X}_{2013} + \hat{\phi}_2 X_{2012} = 1 \cdot 6.69 - 0.21 \cdot 9 = 4.80, \\
\hat{X}_{2015} &= \hat{\phi}_1 \hat{X}_{2014} + \hat{\phi}_2 \hat{X}_{2013} = 1 \cdot 4.80 - 0.21 \cdot 6.69 \approx 3.40.
\end{aligned}$$

Therefore the 2015 forecast is approximately 3.4 million dollars. □

Example 23.3 Derive the h-step-ahead predictor for an invertible MA(q) process

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \qquad \text{and} \qquad \{Z_t\} \sim \text{WN}(0, \sigma^2).$$

Solution. Consider first the ideal linear predictor based on the complete history $\mathcal{H}_n = \overline{\text{span}}\{1, X_n, X_{n-1}, \dots\}$. Denote its projection operator by $P_{\mathcal{H}_n}$. By linearity of projection,

$$\hat{X}_{n+h} = P_{\mathcal{H}_n} Z_{n+h} + \sum_{j=1}^{q} \theta_j P_{\mathcal{H}_n} Z_{n+h-j}.$$

If $h > q$, every innovation in this display is in the future and is orthogonal to the observed past. Therefore

$$\hat{X}_{n+h} = 0, \quad h > q.$$

If $1 \le h \le q$, the terms with $j < h$ still involve future innovations and drop out. Invertibility lets us recover the remaining innovations from the complete past, so the ideal predictor is

$$\hat{X}_{n+h} = \sum_{j=h}^{q} \theta_j Z_{n+h-j}, \quad 1 \le h \le q.$$

This is an infinite-history formula, not the exact projection based only on $X_1, \dots, X_n$. With a finite

sample, presample innovations are unavailable. A standard practical approximation initializes them at zero and computes

$$U_t = X_t - \sum_{j=1}^{q} \theta_j U_{t-j} \qquad \text{and} \qquad U_0 = U_{-1} = U_{-2} = \cdots = 0.$$

Replacing Z_{n+h-j} by these recursively estimated innovations gives

$$\hat{X}_{n+h} = \begin{cases} \sum_{j=h}^{q} \theta_j U_{n+h-j}, & 1 \le h \le q, \\ 0, & h > q. \end{cases}$$

□

Tutorial 6 — Thursday, November 20, 2014

Example 6.1 Let $X_t = Z_t + 0.5Z_{t-1}$, where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. If

$$x_1 = 0.3, x_2 = -0.1, x_3 = 0.1,$$

find the forecasts of x_4 and x_5.

Solution. This is an MA(1) process, so forecasts beyond one step are 0. With current time $n = 3$, we have

$$\hat{X}_5 = \hat{X}_{3+2} = 0.$$

For the one-step forecast,

$$\hat{X}_4 = \theta U_3 = 0.5\, U_3.$$

Using $U_0 = 0$ and the recursion $U_t = X_t - 0.5U_{t-1}$,

$$U_1 = 0.3, \qquad U_2 = -0.1 - 0.5(0.3) = -0.25, \qquad \text{and} \qquad U_3 = 0.1 - 0.5(-0.25) = 0.225.$$

Therefore,

$$\hat{X}_4 = 0.5(0.225) = 0.1125 \qquad \text{and} \qquad \hat{X}_5 = 0.$$

□

Example 6.2 For the MA(1) process $X_t = Z_t + \theta Z_{t-1}$, where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ and $|\theta| < 1$, consider the two predictors

$$\hat{X}_{n+1} = \theta U_n, \qquad U_0 = 0, \qquad \text{and} \qquad U_t = X_t - \theta U_{t-1},$$

and

$$\hat{X}^*_{n+1} = -\sum_{i=1}^{n} (-\theta)^i X_{n+1-i}.$$

show that these two predictors are identical.

Solution. The recursion starts as

$$U_0 = 0, \qquad U_1 = X_1, \qquad U_2 = X_2 - \theta X_1, \qquad \text{and} \qquad U_3 = X_3 - \theta X_2 + \theta^2 X_1,$$

which suggests the pattern

$$U_n = X_n + \sum_{i=1}^{n-1} (-\theta)^i X_{n-i}.$$

This is proved by induction. Multiplying by θ,

$$\theta U_n = \theta X_n + \sum_{i=1}^{n-1} (-\theta)^{i+1} X_{n-i} = -\sum_{j=1}^{n} (-\theta)^j X_{n+1-j}.$$

Hence

$$\hat{X}_{n+1} = \theta U_n = -\sum_{j=1}^{n} (-\theta)^j X_{n+1-j} = \hat{X}^*_{n+1}.$$

Therefore the two predictor formulas are identical. □

This highlights an important contrast between AR and MA forecasting. For AR(p) models, only the most recent p values are needed, whereas forecasting an MA(q) model depends on the entire past through the innovation recursion. This is closely related to the invertible representation of an MA model as an AR(∞) process.

Lecture 24 — Tuesday, November 25, 2014

The prediction formulas for AR(p) and MA(q) processes extend naturally to fitted SARIMA models.

In a SARIMA model, the innovation sequence is assumed to be white noise with mean 0 and constant variance. Forecasts can be formed under these second-moment conditions, but Gaussian prediction intervals require the additional assumption of Gaussian innovations. Ignoring parameter-estimation uncertainty, a generic approximate 95% interval has the form

$$\hat{X}_{n+h} \pm 1.96\sqrt{\operatorname{Var}(X_{n+h} - \hat{X}_{n+h})}.$$

A typical forecast fan is shown in Figure 18: the fitted conditional mean continues beyond the last observation while uncertainty accumulates with the horizon.

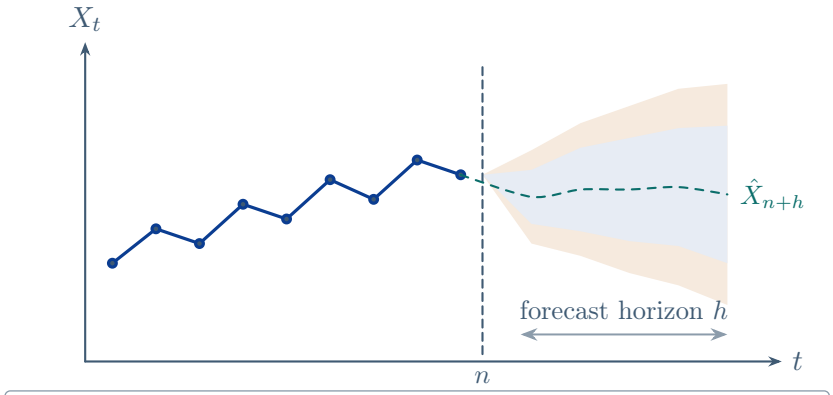


FIGURE 18

Applied Case Studies

We finish by carrying one data set through transformation, identification, estimation, diagnostics, and forecasting. Let

$$X_t = \log(\text{AirPassengers}_t).$$

As Figure 2 showed, the logarithm makes the seasonal swings roughly constant in size. The logged series still has trend and seasonality, however, so we examine both the ordinary difference ∇X_t and the ordinary-seasonal difference $\nabla\nabla_{12}X_t = (1-B)(1-B^{12})X_t$.

The upper row of Figure 19 shows that ordinary differencing removes the long-run rise but leaves a regular annual pattern. Seasonal differencing removes that pattern as well. The lower row tells the same story through autocorrelations: pronounced seasonal structure remains after ordinary differencing, whereas the twice-differenced series has only a few short-lag and seasonal spikes. These diagnostics suggest

$$d = 1, \qquad D = 1, \qquad \text{and} \qquad s = 12.$$

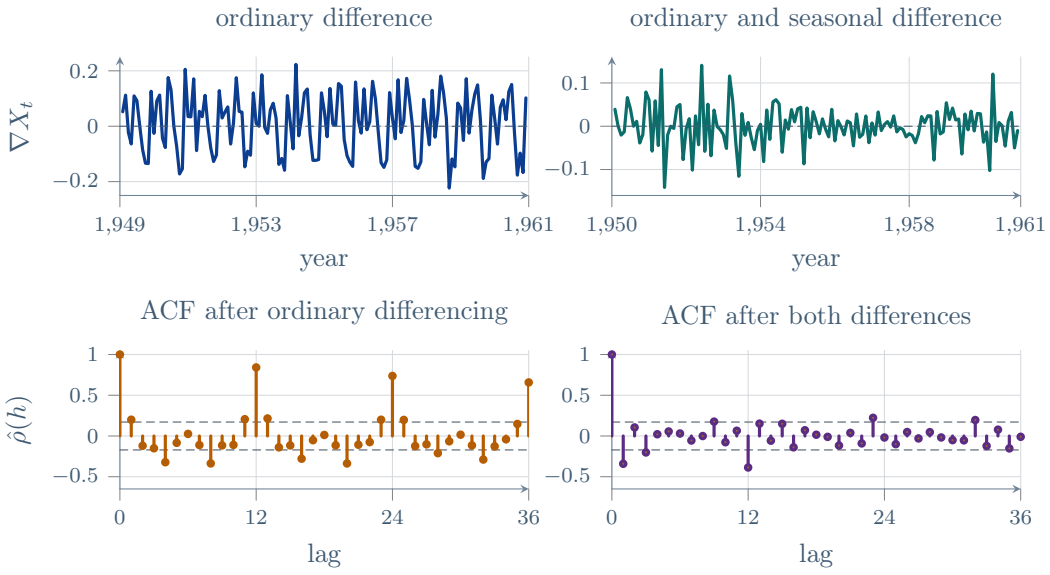


FIGURE 19

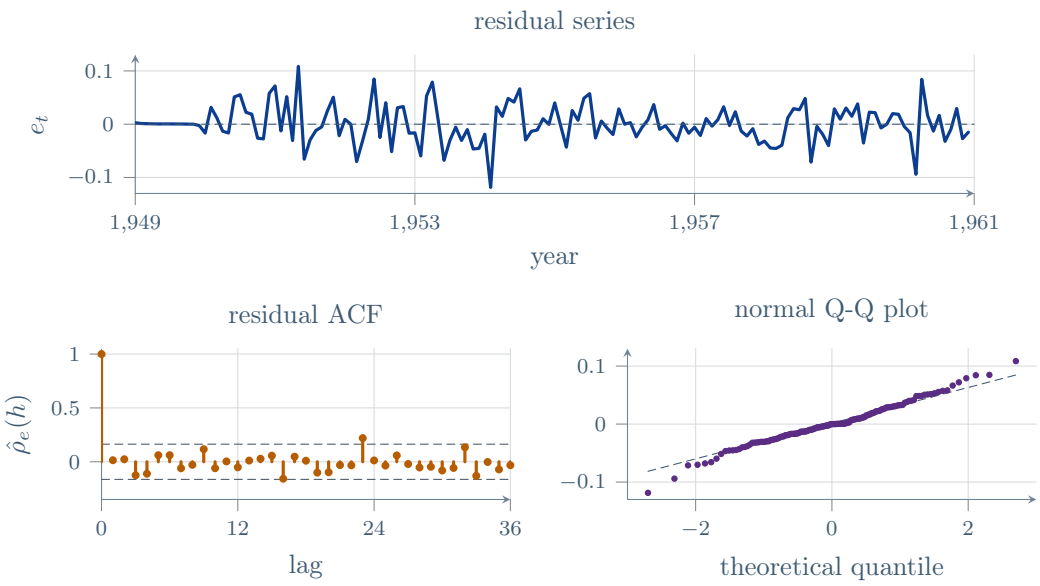
The nonseasonal and seasonal ACF spikes motivate the candidate model

$$\text{SARIMA}(0,1,1) \times (0,1,1)_{12},$$

or, equivalently,

$$(1-B)(1-B^{12})X_t = (1+\theta B)(1+\Theta B^{12})Z_t, \qquad \{Z_t\} \sim \text{WN}(0,\sigma^2).$$

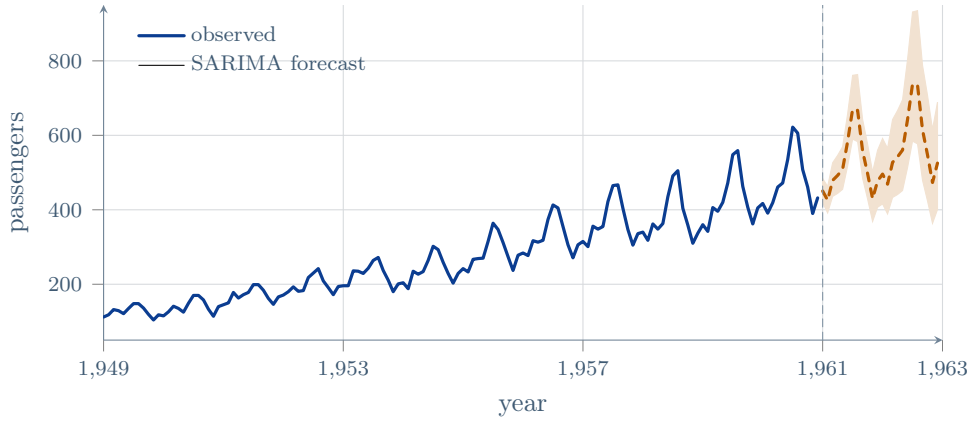
Maximum likelihood gives $\hat{\theta} = -0.402$, $\hat{\Theta} = -0.557$, $\hat{\sigma}^2 = 0.00135$, and AIC $= -483.4$. The residual series in Figure 20 has no persistent change in level or variance, and the Q-Q plot is close to linear. The residual ACF is also generally quiet. One isolated spike at lag 23 crosses the approximate reference band, but it is not accompanied by a broader pattern of residual dependence.



diagnostics check stability, remaining dependence, and Gaussian shape

FIGURE 20

Finally, Figure 21 shows the fitted two-year forecast on the original passenger scale. The centre line is obtained by exponentiating the forecast on the log scale, and the shaded region is the corresponding pointwise 95% prediction interval. The interval widens with horizon while retaining the annual pattern.



the seasonal forecast continues the annual cycle while uncertainty grows

FIGURE 21

Selected original-scale forecasts are listed below. The point forecast does not rise monotonically with the horizon because the fitted seasonal cycle is retained; comparing horizons 12 months apart is therefore more informative than comparing adjacent months.

Horizon (months)	Point forecast	Lower 95% limit	Upper 95% limit
1	450	419	484
6	583	517	658
12	477	407	560
18	642	513	804
24	525	401	689

The entire identification, fitting, diagnostic, and forecasting sequence is reproduced below.

```
data(AirPassengers)

x <- log(AirPassengers)
ordinary_difference <- diff(x)
ordinary_seasonal_difference <- diff(ordinary_difference, lag = 12)

par(mfrow = c(2, 2))
plot(ordinary_difference)
acf(ordinary_difference, lag.max = 36)
plot(ordinary_seasonal_difference)
acf(ordinary_seasonal_difference, lag.max = 36)

fit <- arima(
  x,
  order = c(0, 1, 1),
  seasonal = list(order = c(0, 1, 1), period = 12),
  method = "ML"
)
fit

par(mfrow = c(2, 2))
plot(residuals(fit), type = "l")
acf(residuals(fit), lag.max = 36)
pacf(residuals(fit), lag.max = 36)
qqnorm(residuals(fit))
qqline(residuals(fit))

forecast <- predict(fit, n.ahead = 24)
forecast_passengers <- exp(forecast$pred)
lower <- exp(forecast$pred - 1.96 * forecast$se)
upper <- exp(forecast$pred + 1.96 * forecast$se)
```

Appendix: Lecture and Tutorial Index