STATISTICS 479
TIME SERIES ANALYSIS

Doug Wiens*

April 9, 2015

Contents

Part I

Introduction to Time Series

1. Independence; dependence; stationarity

- What is a time series? Previous STAT courses will have emphasized *independent* data, e.g. $n =$ 5 crops from different farms are studied to determine mean yield; here the crop yields $X_1, .., X_n$ can be viewed as independent (why?). Summary statistics $\bar{x}, S$ can be used to make inferences about $\mu =$ mean yield. Contrast this with: A plant is weighed each week, for n weeks. Put $X_t =$ weight at end of t^{th} week. Then $\{X_1, X_2, ...X_n\}$ are *correlated*, hence *dependent*; for instance $X_1 \leq X_2 \leq ... \leq X_n$.

- Interesting problems:
  - Estimate $\mu_t =$ mean weight at time t.
  - Forecast future size X_{n+t}, given data $\{X_1, ...X_n\}$.

- Crop yield problem - the yields X_i are independent and identically distributed (i.i.d.), and so if, say, we know that they are $N(\mu, \sigma^2)$ then $\bar{x}, S^2$ are the only estimates we need. The plant weight problem would seem to require us to keep track of the entire history of the process, and to accumulate information about the parameters of the distributions, as we go along: information about X_t is contained in $X_1, X_2, ...X_{t-1}$ and so we might need to examine all relationships between these random variables.

- Define

  $$F_{1,...,n}(x_1, x_2, ...x_n) = P(X_1 \leq x_1, X_2 \leq x_2, ...X_n \leq x_n)$$

  the *joint probability distribution* of the weights at times (weeks) 1,...,n. We would have complete knowledge of the probabilistic behaviour of the process if we knew $F_{1,...,n}$. This is of course very intractable. Simplifying assumptions:

- If the variables are independent, then

$$F_{1,\ldots,n}(x_1, x_2, \ldots x_n) = \\ P(X_1 \leq x_1)P(X_2 \leq x_2) \cdots P(X_n \leq x_n);$$

if also identically distributed (i.i.d.) with common (time independent!) distribution function F then

$$F_{1,\ldots,n}(x_1, x_2, \ldots x_n) = \prod_{t=1}^{n} F(x_t).$$

This extreme simplification is what makes the crop yield problem so easy.

- Much less extreme is the assumption of "strict (strong) stationarity":

$F_{t_1,\ldots,t_m} = F_{t_1+s,\ldots,t_m+s}$ for any $t_1, \ldots, t_m$ and s.

Interpretation: the probabilistic information in the process is unchanged if time is shifted s units. e.g. think of the output from an assembly line while the process is "in control"; there will be random fluctuations but the process will "look the same" in any two non-overlapping half-hour intervals.

- Strictly stationary processes are quite special; in time series we work with a weaker form of stationarity which still simplifies things enough to be useful. First recall notions of expectation, covariance and correlation:

  - $E[X]$ = mean of the random variable (r.v.) X. Also written μ_X.

  - $E[h(X,Y)]$ = mean of the r.v. $h(X,Y)$, whose distribution can be determined.

  - These expectation are formally defined in terms of integrals, but in practice *we almost never actually evaluate an integral in Statistics*. There are typically easier ways to compute expectations; most important tool is linearity:

    $$E\left[aX+bY\right] = aE\left[X\right] + bE\left[Y\right],$$

    for constants a, b and r.v.s X, Y.

  - Random variables (X,Y) are independent if for all x, y :

    $$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y);$$

a more useful characterization is that X, Y independent $\Leftrightarrow$ for (almost) all functions f, g :

$$E\left[f(X)g(Y)\right] = E\left[f(X)\right] E\left[g(Y)\right].$$

– Covariance:

$$COV[X,Y] \ = E[(X-\mu_X)\,(Y-\mu_Y)].$$

Note consequence of linearity (assigned):

$$COV[X,Y] \ = E[XY] - \mu_X \mu_Y.$$

– Note that

$$COV[X,X] = E\left[(X-\mu_X)^2\right] = VAR[X].$$

– Correlation:

$$CORR[X,Y] = \frac{COV[X,Y]}{\sqrt{VAR[X]VAR[Y]}}.$$

This is dimensionless, i.e. is just a number, and is always in $[-1,1]$. The bounds ± 1 are attained $\Leftrightarrow Y = aX + b$ (with probability 1), and then the correlation is the sign of a. You should show the "$\Leftarrow$" part of this statement;

it just involves the calculation of the correlation. The other direction is a version of the (incredibly useful) "Cauchy-Schwarz Inequality":

$$(E[XY])^2 \leq E\left[X^2\right] E\left[Y^2\right] \text{ for all } X, Y.$$

– Independent r.v.s are uncorrelated. (Why?) *For jointly normal r.v.s the converse holds as well: uncorrelated* $\Rightarrow$ *independent*.

– Example: $X \sim N(0,1)$ is uncorrelated with X^2 (you should show this) even though dependence is clearly strong. This highlights interpretation of correlation in terms of *linear* dependence, and stresses that uncorrelated $\Rightarrow$ independent only in the presence of joint normality.

- We say a series $\{X_t\}_{t=1}^n$ is "weakly stationary" if:

1. $\mu_t = E[X_t]$ does not depend on t.

2. $COV[X_s, X_t]$ depends on s and t only through the *lag* $|s-t|$, i.e. the covariances (hence correlations) depend only on how far apart the variables are, in time. Hence in particular $\sigma_t^2 = VAR[X_t]$ does not depend on t.

Now if $t = s + m$ we have

$$COV[X_s, X_t] = COV[X_s, X_{s+m}] = \gamma(m)$$

for some function of $|m|$ alone (and not of s); this is the *autocovariance function*. (We sometimes write $\gamma_X(m)$.) By 1. and 2. we have

$$CORR[X_s, X_{s+m}] = \frac{COV[X_s, X_{s+m}]}{\sigma_s \sigma_{s+m}} = \frac{\gamma(m)}{\gamma(0)};$$

the function $\rho(m) = \frac{\gamma(m)}{\gamma(0)}$ is called the *autocorrelation function*.

2. Autocovariance; autocorrelation

- Recall notion of (weak) stationarity. See Figure 2.1 (U.S. births): Under weak stationarity, the mean should be time-independent and correlations between (for instance) January births in one year and in another year should depend only on the number of years and not on the month, so be the same for June births the same number of years apart.

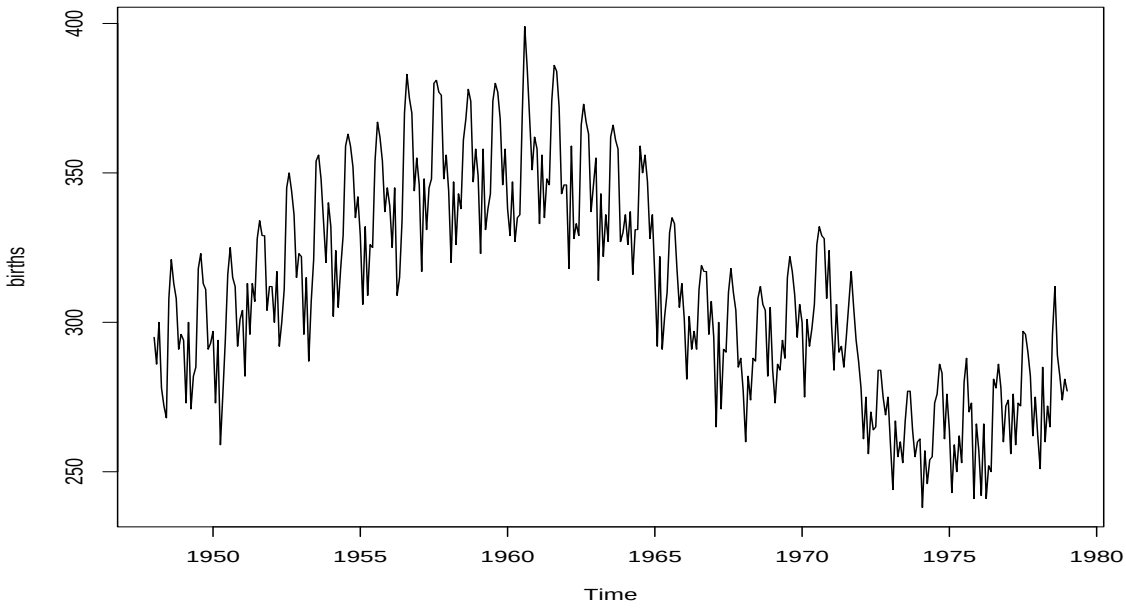

Figure 2.1. X_t = U.S. births, time = Dec. 1947 + t, $t = 1, .., 373$.

- Example: model of plant weight/growth. $X_t =$ weight at end of t^{th} week. Possible model: $X_t = X_{t-1} + \delta + w_t$, where $\{w_1, w_2, ...\}$ is a sequence of uncorrelated r.v.s with mean zero and constant variance σ_w^2. Such a sequence $\{w_t\}$ is called *white noise*. The *growth* is

$$\nabla X_t = X_t - X_{t-1} = \delta + w_t.$$

- The "differenced series" $\{\nabla X_t\}$ is (weakly) stationary:

$$\begin{aligned} E[\nabla X_t] &= \delta + E[w_t] \text{ (why?)} \\ &= \delta, \text{ independent of time.} \\ COV[\nabla X_{t+m}, \nabla X_t] &= E[(\nabla X_{t+m} - \delta)(\nabla X_t - \delta)] \\ &= E[w_{t+m} w_t] \\ &= COV[w_{t+m}, w_t] \\ &= \sigma_w^2 I(m = 0). \end{aligned}$$

  - Notation: $I(E)$ is the "indicator of the event E": it equals 1 if E occurs and 0 otherwise.

- Digression: Some useful facts about covariances. We have

$$COV[aX + b, Y] = aCOV[X, Y].$$

Reason:

$$\begin{aligned} & COV[aX + b, Y] \\ = \; & E\left[\{aX + b - E\left(aX + b\right)\} \cdot \{Y - E(Y)\}\right] \\ = \; & E\left[a\left\{(X - E(X))\right\} \cdot \{Y - E(Y)\}\right] \text{ (why?)} \\ = \; & aE\left[\{(X - E(X))\} \cdot \{Y - E(Y)\}\right] \text{ (why?)} \\ = \; & aCOV[X, Y]. \end{aligned}$$

In particular a constant is uncorrelated with anything: $COV[b, Y] = 0$. As above, we can (and you should) show that the covariance is linear in each argument:

$$COV[aX + b, cY + d] = acCOV[X, Y].$$

This extends to:

$$COV\left[\sum_i a_i X_i, \sum_j c_j Y_j\right] = \sum_i \sum_j a_i c_j COV\left[X_i, Y_j\right].$$

Another consequence:

$$\begin{aligned} & VAR[X+Y] = COV[X+Y, X+Y] \\ = \;& VAR[X] + VAR[Y] + 2COV[X,Y] \text{ (why?)} \end{aligned}$$

and this $= VAR[X] + VAR[Y]$ *if* X, Y *are uncorrelated, not otherwise.*

- Returning to the plant growth example, $\{X_t\}$ is non-stationary:

  1. $\mu_t = E[X_t] = ... = t\delta$ (if $X_0 = 0$); depends on t.

  2. $\sigma_t^2 = ... = t\sigma_w^2$; depends on t. (We assume that X_s and w_t are uncorrelated if $s < t$.)

  3. Let $t > s$. Then (how?)

$$\begin{aligned} COV[X_s, X_t] &= \sigma_s^2 = s\sigma_w^2; \\ CORR[X_s, X_t] &= \frac{\sigma_s^2}{\sigma_s \sigma_t} = \sqrt{\frac{s}{t}}. \end{aligned}$$

  In general $CORR[X_s, X_t] = \sqrt{\frac{\min(s,t)}{\max(s,t)}}$; not a function of $|s-t|$ alone.

- Another example:

$$X_t = \mu + w_t + w_{t-1},$$

where $\{w_t\}$ is white noise. E.g. radio signal = mean signal + noise (atmospheric interference, etc.) from two time periods. Then

$$\begin{aligned}
\mu_t &= \mu, \\
\sigma_t^2 &= 2\sigma_w^2, \\
COV[X_{t-1}, X_t] &= COV[(w_{t-2} + w_{t-1}), \\
&\quad (w_t + w_{t-1})] \\
&= COV[w_{t-1}, w_{t-1}] = \sigma_w^2, \\
\text{For } m &> 1: \\
COV[X_{t-m}, X_t] &= COV[(w_{t-m-1} + w_{t-m}), \\
&\quad (w_t + w_{t-1})] \\
&= 0 \text{ (why?)}.
\end{aligned}$$

Thus

$$\rho(m) = \frac{\gamma(m)}{\gamma(0)} = \begin{cases} 1, & m = 0, \\ 1/2, & m = \pm 1, \\ 0, & |m| > 1. \end{cases}$$

- Birth Series: This is stored as 'birth' in R, once you install and invoke the authors' `astsa` package. See the R-example for Lecture 2, on the course website, to see how to retrieve it and obtain the output leading to Figures 2.1, 2.2. The raw data (Figure 2.1) hints at long term trend and at fluctuations within each year. A 12-month centred moving average eliminates fluctuations within the years and isolates longer trends. A form of smoothing; also known as a filter:

$$Y_t = \frac{\left\{ \begin{array}{c} .5X_{t-6} + X_{t-5} + ... + X_t + \\ X_{t+1} + ... + X_{t+5} + .5X_{t+6} \end{array} \right\}}{12}.$$

  - Can we predict future births? We need to know (or estimate) the mechanism generating the stochastic behaviour - study the random error, dependencies, etc.

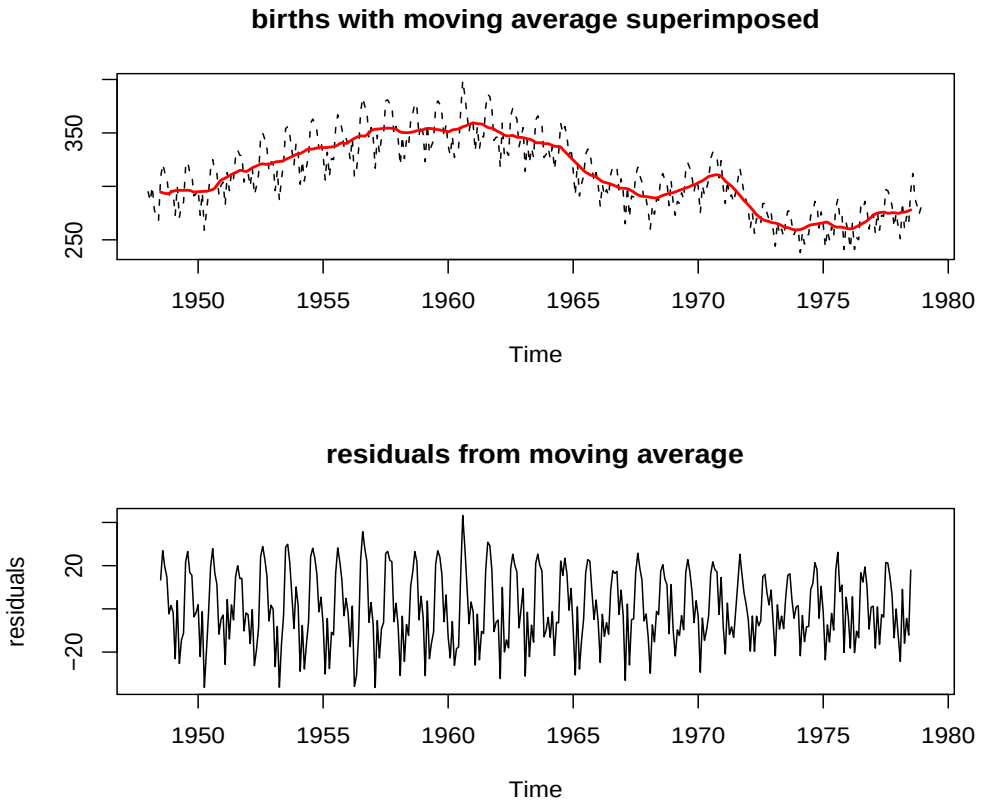

Figure 2.2. Top: Births X_t with $Y_t =$ 12-month centred moving average overlaid. Bottom: Residuals $X_t - Y_t$, monthly fluctuations shown. Peaks in Aug./Sept., troughs in Feb. Note first and last six months are lost in taking moving average; first shown is July 1948.

- Possible approaches

  - Time domain: study a series by, e.g., regressing it on its own past (a common model is AR(1): $X_t = \phi X_{t-1} + w_t$), or by treating it as a sum of uncorrelated errors (for instance MA(2): $X_t = w_t - \theta_1 w_{t-1} - \theta_2 w_{t-2}$); need to see how to estimate parameters, compute forecasts, etc. Chapter 3 of text. In births series we might want to model the long term trend Y_t and the monthly fluctuations $X_t - Y_t$ in very different ways. (Figure 2.2.)

  - Frequency domain: represent the ACF through its Fourier transform (series of sines/cosines); study the periodic variations, at particular frequencies, which produce the series. More on this later - chapter 4 of text. (Note: "ACF" gets used as an abbreviation both for autocovariance function and for autocorrelation function. Since one is just a multiple of the other - the multiple being $\gamma(0) = VAR[X_t]$ (independent of t!!) - there is usually no confusion resulting from this.)

3. Cross-correlation

- Often we study relationships between two time series $\{X_t\}, \{Y_t\}$, possibly to predict (forecast) one from the other. E.g.,

  X_t = unemployment rate in period t,

  Y_t = new houses constructed in time period t.

  One anticipates that Y_t is affected by X_t (but perhaps after a certain lag). Let's entertain a model of the form

  $$Y_{t+D} = AX_t + w_t$$

  for some (unknown) constants A and D; $\{w_t\}$ is white noise uncorrelated with $\{X_t\}$. Presumably $A < 0$ and $D > 0$. Interpretation: Unemployment is a leading indicator of housing starts (and housing starts lag unemployment). The noise serves to explain the effects of other influences on housing.

  - Assume $\{X_t\}, \{Y_t\}$ have had their means subtracted, so that we don't need a constant in this model (analogous to centring the variables in a regression); here we assume stationarity so that these means are independent of t.

- Must impose conditions on the two series analogous to the weak stationarity imposed when studying single series. We say that two series $\{X_t\}, \{Y_t\}$ are jointly stationary if each is stationary, and if $COV[X_{t+m}, Y_t]$ depends on m only, and not on t. We write $\gamma_{XY}(m) = COV[X_{t+m}, Y_t]$, the cross-covariance function.

  - You should show that $\gamma_{XY}(m) = \gamma_{YX}(-m)$; note also that $\gamma_{XX}(m) = \gamma_X(m)$ $(= \gamma_X(-m)$ by the above).

- Cross-correlation function (CCF):

$$\begin{aligned}\rho_{XY}(m) &= \frac{COV[X_{t+m}, Y_t]}{\sqrt{VAR[X_{t+m}]VAR[Y_t]}} \\ &= \frac{\gamma_{XY}(m)}{\sqrt{\gamma_X(0)\gamma_Y(0)}}.\end{aligned}$$

- Example: $\{X_t\}, \{Y_t\}$ as above. Problem: What are the optimal values of A and D? As we will throughout the course, we define optimality in terms of minimum mean squared error. We observe X_t, and predict Y_{t+D} by AX_t. The MSE is

$$\begin{aligned}MSE(A, D) &= E[(\text{Actual r.v. - Predicted value})^2] \\ &= E[(Y_{t+D} - AX_t)^2].\end{aligned}$$

  Assume $\{X_t\}$ is stationary; then (if the model holds) so is $\{Y_t\}$:

$$\begin{aligned}& \gamma_Y(m) \\ = \;& COV\left[AX_{t-D+m} + w_{t-D+m},\ AX_{t-D} + w_{t-D}\right] \\ = \;& A^2 COV\left[X_{t-D+m},\ X_{t-D}\right] \\ & + COV\left[w_{t-D+m}, w_{t-D}\right] \text{ (why?)} \\ = \;& A^2\gamma_X(m) + \sigma_w^2 I(m = 0).\end{aligned}$$

– Similarly $\{X_t\}, \{Y_t\}$ are jointly stationary:

$$\begin{aligned}\gamma_{YX}(m) &= COV\left[Y_{t+m}, X_t\right] \\ &= \ldots = A\gamma_X(m-D).\end{aligned}$$

Now

$$\begin{aligned}& MSE(A, D) \\ = \; & E[\{Y_{t+D} - AX_t\}^2] \\ = \; & \gamma_Y(0) - 2A\gamma_{YX}(D) + A^2\gamma_X(0).\end{aligned}$$

Minimize over A to get the optimal value

$$A^* = A^*(D) = \frac{\gamma_{YX}(D)}{\gamma_X(0)},$$

(note similarity with slope estimate in straight line regression) with

$$MSE(A^*, D) = \gamma_Y(0)\left(1 - \rho^2_{YX}(D)\right),$$

so the optimal D is the value D^* maximizing $|\rho_{YX}(D)|$.

– This calculation also illustrates that, quite generally, differentiation and expectation can be interchanged:

$$\frac{\partial}{\partial A} E\left[\{Y_{t+D} - AX_t\}^2\right] = E\left[\frac{\partial}{\partial A}\{Y_{t+D} - AX_t\}^2\right].$$

- To continue with this problem we must have a way of estimating the ACF and CCF. For a (weakly) stationary series $\{X_t\}$, from which we have data $\{x_t\}_{t=1}^n$, we estimate μ_X by the average $\bar{x}$, and $\gamma_X(m)$ $(m > 0)$ by

$$\hat{\gamma}_X(m) = \frac{\sum_{t=1}^{n-m} (x_{t+m} - \bar{x})(x_t - \bar{x})}{n}.$$

  Then we define $\hat{\gamma}_X(-m) = \hat{\gamma}_X(m)$ and $\hat{\rho}_X(m) = \hat{\gamma}_X(m)/\hat{\gamma}_X(0)$.

  - Note the n in the denominator - neither this nor $n - m$ will remove the bias ($E[\hat{\gamma}_X(m)] \neq \gamma_X(m)$) but n gives a smaller variance and typically smaller MSE $(= E\left[(\hat{\gamma}_X(m) - \gamma_X(m))^2\right])$. You should show that the MSE in estimating a parameter is always expressible as

$$MSE = VAR + BIAS^2;$$

    in this case this takes the form

$$\begin{aligned} & E\left[(\hat{\gamma}_X(m) - \gamma_X(m))^2\right] \\ = \; & E\left[(\hat{\gamma}_X(m) - E\{\hat{\gamma}_X(m)\})^2\right] \\ & + \{E[\hat{\gamma}_X(m)] - \gamma_X(m)\}^2. \end{aligned}$$

- $\hat{\gamma}_X$ and $\hat{\rho}_X$ are the sample autocovariance and sample autocorrelation functions. For jointly stationary series $\{X_t\}, \{Y_t\}$ we estimate the cross-covariance function $\gamma_{XY}(m)$ by the sample cross-covariance function

$$\hat{\gamma}_{XY}(m) = \frac{\sum_{t=1}^{n-m} (x_{t+m} - \bar{x})(y_t - \bar{y})}{n} \quad (m \geq 0)$$

with

$$\begin{aligned} \hat{\gamma}_{XY}(-m) &= \hat{\gamma}_{YX}(m), \\ \hat{\rho}_{XY}(m) &= \hat{\gamma}_{XY}(m)/\sqrt{\hat{\gamma}_X(0)\hat{\gamma}_Y(0)}. \end{aligned}$$

- Inferences about ρ_X can be carried out by using a large sample approximation: If $\{x_t\}_{t=1}^n$ is a series of observations from $\{X_t\}$, and $\{X_t\}$ is white noise (so $\rho_X(m) = I(m = 0)$) then $\sqrt{n}\hat{\rho}_X(m)$ is approximately distributed as $N(0, 1)$. To test $H_0 : \rho_X(m) = 0$ against $\rho_X(m) \neq 0$ at level α we reject if $|\hat{\rho}_X(m)| > z_{\alpha/2}/\sqrt{n}$.

4. Exploratory data analysis; regression

- Example: retrieve the file *sunspotz* (= monthly record of numbers of sunspots; Figure 4.1; not *sunspots*). In R, `qwe = acf(sunspotz,36)` gives Figure 4.2; print out `qwe` to get the values of the sample acf.

```
qwe = acf(sunspotz,36)
mat = cbind(0:36, qwe$acf)
colnames(mat) = c("lag", "acf")
mat
```

```
      lag           acf
 [1,]   0  1.000000000
 [2,]   1  0.943268450
 [3,]   2  0.810296900
 [4,]   3  0.634430886
 [5,]   4  0.432717891
 [6,]   5  0.223065311
 [7,]   6  0.024495760, etc.
```

```
> var(sunspotz) # response: [1] 1538.745
```

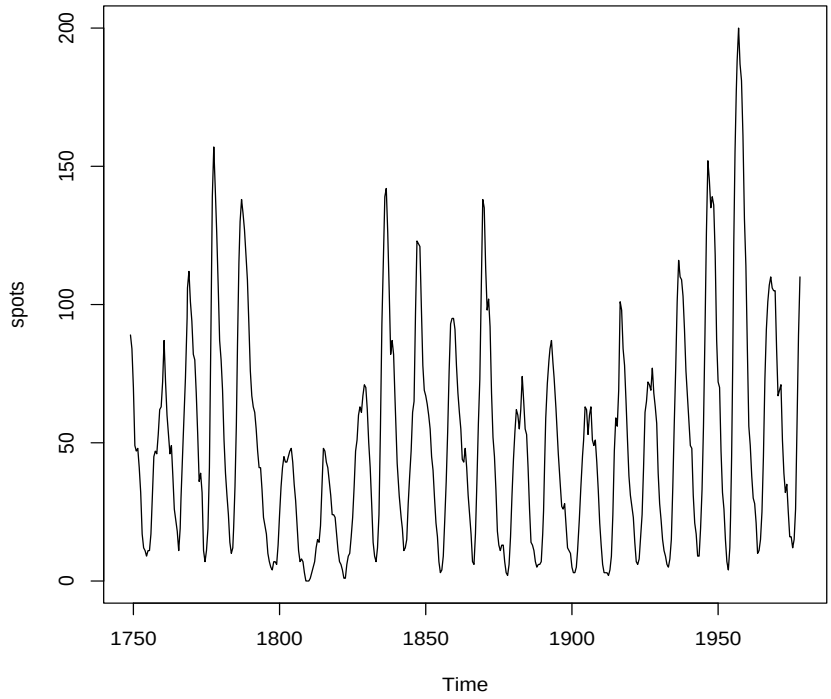

Figure 4.1. Monthly count of sunspots. These are highly correlated with solar storms (intense bursts of magnetic energy, disrupting communications systems on Earth) and (rather mysteriously) with agricultural series.

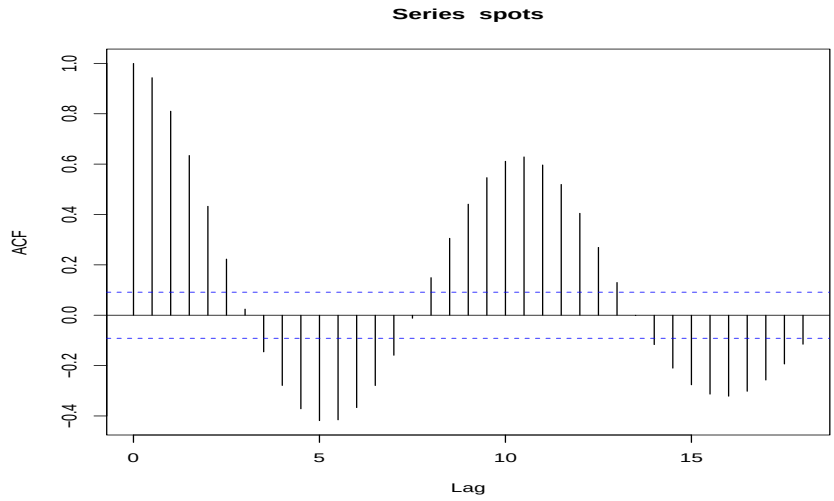

Figure 4.2. Sample ACF, with $\alpha = .05$ (individual) test limits of $\pm .0915$.

- Note $\hat{\rho}_X(6) = .0245$; is there evidence that in fact $\rho_X(6) \neq 0$? If $H_0 : \rho_X(6) = 0$ is true, then $\sqrt{n}\hat{\rho}_X(6)$ is approx. $N(0,1)$ and so $\hat{\rho}_X(6)$ exceeds .0245 *in absolute value* (2 sided alternative!) with probability

$$\begin{aligned} p &= P(|Z| = |\sqrt{n}\hat{\rho}_X(6)| > \sqrt{459} \cdot .0245 = .5248) \\ &= .600. \end{aligned}$$

This is the p-value = probability of observing a value of the test statistic which is at least as extreme as what was in fact observed, *if the null*

hypothesis is indeed true. Since the p-value is so large, we do not have significant evidence against the null hypothesis. To get the p-value down to .05 would require $|\hat{\rho}_X(6)| = \frac{z_{.05/2}}{\sqrt{n}} = .0915$:

$$P\left(|\hat{\rho}_X(6)| > .0915\right) = P(|Z| > 1.96) = .05.$$

- The $\alpha = .05$ band gives the critical values immediately; but *note that this is for one test only*. Even if all $\rho_X(m) = 0$, some of the $\hat{\rho}_X(m)$ will, with high probability, fall outside the testing band purely by chance. Assume that in fact all $\rho_X(m) = 0$. Then we have the *Bonferroni bound*:

$$\begin{aligned} & P\left(\text{at least one false rejection}\right) \\ \leq & \sum_i P(i^{th} \text{ is falsely rejected}) = M\alpha, \end{aligned}$$

where M is number of tests carried out and $\alpha = .05$. Thus we can assert only that

$$P(\text{no false rejections}) \geq 1 - M\alpha,$$

which is a sometimes useful lower bound (at least if M is not too large).

- All this assumes stationarity. Two quick, preliminary checks: (i) ACF should get small reasonably quickly. (ii) ACFs plotted from subsets of the data should look similar.

- Exploratory data analysis: several operations can be carried out, to transform a non-stationary series to stationarity and thus to allow us to employ the methods of this course.

- Differencing:

$$\begin{aligned} \nabla X_t &= X_t - X_{t-1}, \\ \nabla^2 X_t &= \nabla(\nabla X_t) = \nabla X_t - \nabla X_{t-1}, \end{aligned}$$

  etc. E.g. $X_t = at + b + z_t$ is non-stationary even if z_t is stationary (why?); $\nabla X_t = a + \nabla z_t$ is stationary (if ∇z_t is).

  In R use `diff`:

```
> spots[1:5] # response: [1] 89 84 70 49 47
> diff(spots)[1:4]
       # Order 1, with lag = 1, is the default
[1]  -5 -14 -21  -2
> diff(spots, lag = 2)[1:3]
[1] -19 -35 -23
> diff(spots, diff = 2)[1:3]
[1] -9 -7 19
```

- Transformations to stabilize variance: $X_t \to \log X_t$, $\sqrt{X_t}$, etc.

- Time series regression: Regress a series on its history, or an another series. Example: Southern Oscillation Index vs. Recruits. SOI measures changes in air pressure (related to temperature) over the ocean off west coast of S. America. "Recruits" refers to new members of various fish populations. See Figures 4.3, 4.4.

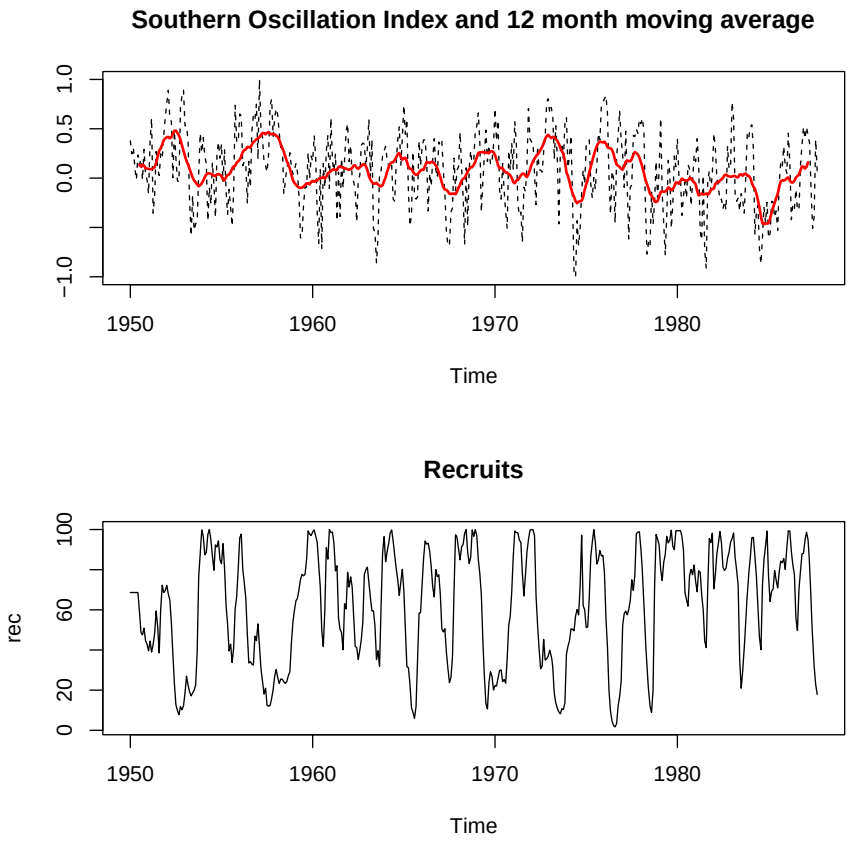

Figure 4.3. SOI, with 12-month moving average, and Recruits. What does the MA(12) filter show?

- CCF suggests that SOI leads Recruits, with most CCF values in the preceding 12 months being significant. Suggests that we might regress $Y_t =$ Recruits on $X_{t-m} =$ lagged SOI for $m = 3, ..., 12$.

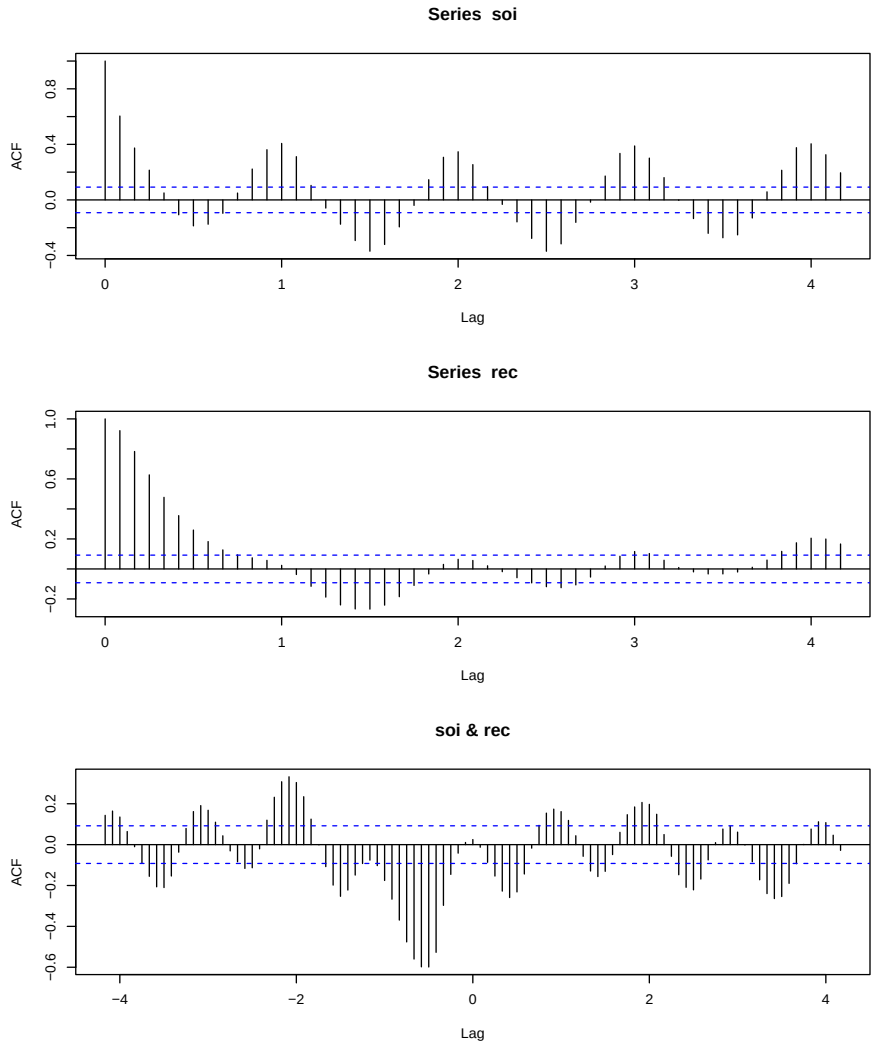

Figure 4.4. ACFs and CCF (`ccf(soi,rec)`$= CORR[SOI_{t+m}, REC_t]$).

We will discuss regression theory in straight line case; then go through multiple regression output.

- $Y_t = \beta_0 + \beta_1 X_t + e_t$; want to fit a linear trend $\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t$. Analogous to the minimum MSE principle introduced earlier we want to minimize $SSE = \sum \left(Y_t - \hat{Y}_t\right)^2$. We hope that the residuals $\hat{e}_t = Y_t - \hat{Y}_t$ behave like a stationary series; we can then apply time series methods to them.

- Regression estimate of σ_e^2 is $S^2 = SSE/(n - 2)$; in general the denominator is $n-$ the number ("p") of regression parameters being estimated.

- t-statistics: each estimate $\hat{\beta} \sim N\left(\beta, \sigma^2_{\hat{\beta}}\right)$ where the variance $\sigma^2_{\hat{\beta}}$ is a certain multiple of σ_e^2; thus

$$\frac{\hat{\beta} - \beta}{\sigma_{\hat{\beta}}} \sim N(0, 1).$$

Replacing σ_e^2 in $\sigma_{\hat{\beta}}^2$ by its estimate S^2 results in an estimated variance $S_{\hat{\beta}}^2$ and

$$\frac{\hat{\beta} - \beta}{S_{\hat{\beta}}} \sim t_{n-p}.$$

Then the reported t-ratio

$$t_{obs} = \frac{\hat{\beta}}{S_{\hat{\beta}}}$$

can be used to test the hypothesis that $\beta = 0$. The t is the number of standard errors that $\hat{\beta}$ is away from the hypothesized value, and the p-value is the probability that it would be (at least) this extreme, if in fact β were $= 0$. Specifically,

$$p - value = 2P\left(t_{n-p} > |t_{obs}|\right).$$

- Example - regress Y_t = Recruits on X_{t-m} = lagged SOI for $m = 3, ..., 12$, and on the time (= 'trend'). See the R-code on the website. Some work goes into forming the lagged SOI values (= `lag(soi, -3)`, etc.); the function 'dynlm' in the R library will do this.

```
x = soi - mean(soi) # Address possible collinearity
trend = time(x)
fit = dynlm(rec~time(x) + L(x, 3:12))
summary(fit)
```

- Example of interpretation: The output that follows reveals that $\hat{\beta}_3$ (= coefficient of X_{t-4}) is .908 standard errors away from 0, and that this would happen by chance 36% of the time if in fact β_3 were = 0. This gives us very little reason to think that $\beta_3 \neq 0$.

```
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Int)    448.69803 158.98688   2.822 0.004991 **
trend     -0.19625  0.08073   -2.431 0.015472 *
L(x, 3)   -1.24458  2.67293   -0.466 0.641722
L(x, 4)   -2.73500  3.01149   -0.908 0.364288
L(x, 5)   -23.25576 3.01318   -7.718 8.38e-14 ***
L(x, 6)   -18.12668 3.02079   -6.001 4.18e-09 ***
L(x, 7)   -13.71622 3.00784   -4.560 6.68e-06 ***
L(x, 8)   -11.22259 3.00780   -3.731 0.000216 ***
L(x, 9)   -8.46734  3.02010   -2.804 0.005282 **
L(x, 10)  -8.19113  3.01584   -2.716 0.006874 **
L(x, 11)  -9.06195  3.01295   -3.008 0.002787 **
L(x, 12) -12.33898  2.68582   -4.594 5.72e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
                0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error:
     16.42 on 429 degrees of freedom
Multiple R-squared: 0.6718
Adjusted R-squared: 0.6633
F-statistic: 79.81 on 11 and 429 DF,
    p-value: < 2.2e-16
```

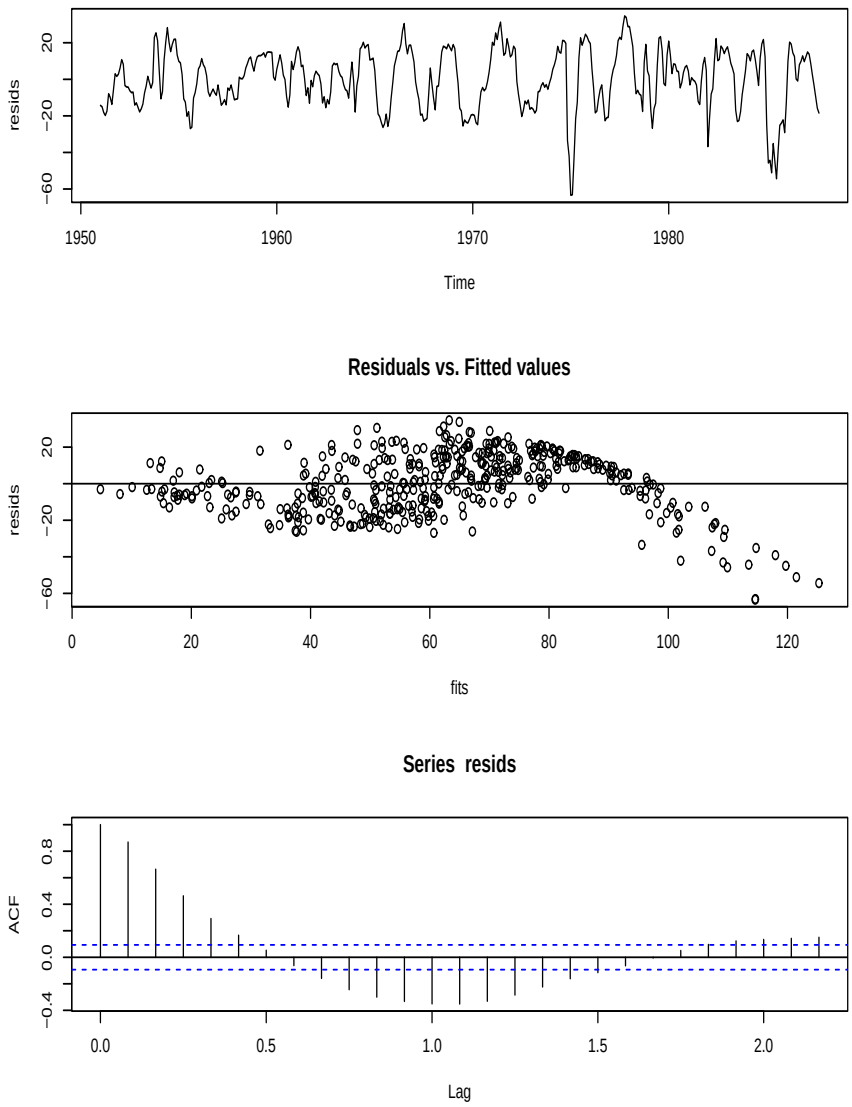

Figure 4.5. Residuals plots from regression of Recruits on time and on lagged SOI. A significant amount of structure remains to be explained.

Part II

Time Domain Analysis

5. Linearity; invertibility

- Suppose $\{X_t\}$ is (weakly) stationary, with mean μ and (finite) variance σ_X^2. Write $X_t - \mu$ as $\dot{X}_t$. Then (*Wold's Representation Theorem*) we can represent $\dot{X}_t$ as

$$\begin{aligned} & \dot{X}_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + ... \\ = & \sum_{k=0}^{\infty} \theta_k w_{t-k} \text{ (with } \theta_0 = 1) \qquad (5.1) \\ & \text{and } \sum_{k=1}^{\infty} \theta_k^2 < \infty. \end{aligned}$$

  Salient feature: *Linear* function of *past and present* (not future) disturbances.

  Interpretation: convergence in mean square; i.e.

$$E\left[\left(\dot{X}_t - \sum_{k=0}^{K} \theta_k w_{t-k}\right)^2\right] \to 0 \text{ as } K \to \infty.$$

  – The conditions ensure that we can take term-by-term expectations of series of the form

$\sum_{k=0}^{\infty} \theta_k X_{t-k}$, if the X_{t-k} have expectations:

$$E\left[\sum_{k=0}^{\infty} \theta_k X_{t-k}\right] = \sum_{k=0}^{\infty} \theta_k E\left[X_{t-k}\right].$$

- If (5.1) holds we say $\{X_t\}$ is a *linear* process (also called *causal* in the text, i.e. doesn't depend on the future). Thus Wold's Representation Theorem can be interpreted as saying that

$$Stationarity \Rightarrow Linearity.$$

The converse holds: Assume $\{X_t\}$ is linear (with a constant mean); then

$$\begin{aligned}
COV[X_t, X_{t+m}] &= E[\dot{X}_t \dot{X}_{t+m}] \\
&= E\left[\sum_{k=0}^{\infty} \theta_k w_{t-k} \sum_{l=0}^{\infty} \theta_l w_{t+m-l}\right] \\
&= \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \theta_k \theta_l E\left[w_{t-k} w_{t+m-l}\right] \\
&= \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \theta_k \theta_l \left\{\sigma_w^2 I\left(l = k+m\right)\right\} \\
&= \sigma_w^2 \sum_{k=0}^{\infty} \theta_k \theta_{k+m}.
\end{aligned}$$

(In particular, $\sigma_X^2 = \sigma_w^2 \sum_{k=0}^{\infty} \theta_k^2 < \infty$.) Thus

$$Stationarity \Leftrightarrow Linearity.$$

- Backshift operator:

$$\begin{aligned} B\left(X_t\right) &= X_{t-1}, \\ B^2\left(X_t\right) &= B \circ B\left(X_t\right) = B\left(X_{t-1}\right) = X_{t-2}, \end{aligned}$$

etc. Then $\{X_t\}$ linear $\Rightarrow \dot{X}_t = \theta(B) w_t$ for the *characteristic polynomial* or *operator*

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + ...$$

This is a "power series" and not really a polynomial unless it terminates. If it does, i.e. if $\theta_k \neq 0$ only for $k \leq q$, we say $\{X_t\}$ is a *moving average* series of order q, written MA(q). Then

$$\begin{aligned} \dot{X}_t &= w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + ... + \theta_q w_{t-q} \\ &= \theta(B) w_t \end{aligned}$$

for

$$\theta(B) = 1 + \theta_1 B + ... + \theta_q B^q,$$

the "MA(q) characteristic polynomial".

- **Invertibility**: $\{X_t\}$ is *invertible* if it can be represented as

  $$\dot{X}_t = \phi_1\dot{X}_{t-1}+\phi_2\dot{X}_{t-2}+...+w_t, \text{ where } \sum_{k=1}^{\infty}|\phi_k| < \infty.$$

  Thus, apart from some noise, X_t is a function of its history. Generally, only invertible processes are of practical interest. In terms of the backshift operator,

  $$\begin{aligned} w_t &= \dot{X}_t - \phi_1\dot{X}_{t-1} - \phi_2\dot{X}_{t-2} - ... \\ &= \phi(B)\dot{X}_t, \end{aligned}$$

  where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ...$ is the characteristic polynomial. If it is a true polynomial, i.e. if $\phi_j \neq 0$ only for $j \leq p$, we say $\{X_t\}$ is an *autoregressive* process of order p, i.e. AR(p). Then

  $$\dot{X}_t = \phi_1\dot{X}_{t-1} + \phi_2\dot{X}_{t-2} + ... + \phi_p\dot{X}_{t-p} + w_t.$$

  - When $\sum_{k=1}^{\infty}|\phi_k| < \infty$ we say the series is *absolutely summable*. The importance of absolute summability is that such series can be

re-arranged - they can be summed in any order. In contrast,

$$\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} = \ln 2 \approx .69,$$

but the series is not absolutely summable: $\sum_{k=1}^{\infty} \frac{1}{k} = \infty$. The original series can be re-arranged to give just about anything; for instance

$$\left(1 + \frac{1}{3} - \frac{1}{2}\right) + \left(\frac{1}{5} + \frac{1}{7} - \frac{1}{4}\right) + \cdots,$$

in which two positive terms are always followed by a negative one, converges to something $> 5/6 \approx .83$ (each term in brackets is positive).

- When is a stationary (i.e. linear) process invertible? Let $\{X_t\}$ be linear, so $\dot{X}_t = \theta(B)w_t$ and $\sum_{k=1}^{\infty} \theta_k^2 < \infty$. Suppose it is invertible. Then $\phi(B)\dot{X}_t = w_t$; thus $\phi(B)\theta(B)w_t = w_t$ and $\phi(B)\theta(B) = 1$. Thus $\phi(B) = 1/\theta(B)$, i.e. $1/\theta(B)$ has a power series expansion with absolutely summable coefficients. This makes $\theta(B)$ quite special.

– Example: MA(1). The operator is $\theta(B) = 1+\theta B$ for some θ. Then if invertible we must have

$$\begin{aligned} & 1/\theta(B) = 1 - \theta B + \theta^2 B^2 - ... \\ = & \sum_{j=0}^{\infty} (-\theta)^j B^j; \text{ AND} \\ & 1 + |\theta| + \left|\theta^2\right| + ... < \infty; \end{aligned}$$

this last point holds iff $|\theta| < 1$. Note that the root of $\theta(B) = 0$ is $B = -1/\theta$, and then $|\theta| < 1 \Leftrightarrow |B| > 1$, i.e. *the MA(1) process with* $\theta(B) = 1 + \theta B$ *is invertible iff the root of* $\theta(B) = 0$ *satisfies* $|B| > 1$.

– In general, a linear process $X_t = \theta(B)w_t$ is invertible iff all roots of the *characteristic equation* $\theta(B) = 0$ satisfy $|B| > 1$ (complex modulus), i.e. they "lie outside the unit circle in the complex plane".

– The *modulus* of a complex number $z = a + ib$ is $|z| = \sqrt{a^2 + b^2}$ (like the norm of a vector with coordinates (a, b)).

- e.g. $X_t = w_t - 2w_{t-1} + 2w_{t-2}$ has characteristic equation

$$\theta(B) = 1 - 2B + 2B^2 = 0.$$

The roots are

$$B = .5 \pm .5i;$$

with $|B| = 1/\sqrt{2} \approx .71 < 1$. Non-invertible.

- Similarly, an invertible process is stationary iff all roots of $\phi(B) = 0$ lie outside the unit circle. This is called the *stationarity condition*. E.g. for an AR(1) the stationarity condition is $|\phi| < 1$.

- A good Primer on Complex Numbers is on the course website.

- From now on we will assume that w_t denotes *Normal* white noise, i.e. $\{w_t\}$ are $N\left(0, \sigma_w^2\right)$ random variables which are uncorrelated, hence *independent*.

6. ARMA models; Yule-Walker equations

Review. Assume now that the mean $\mu_X = 0$. In practice we often subtract the average $\bar{x}$, or a trend $\hat{\beta}_0 + \hat{\beta}_1 t$, from the data values $\{x_t\}$ before starting the analysis ("detrending").

Recall that now $\{w_t\}$ will always represent *Normal* white noise, hence they are *independent* (rather than merely uncorrelated).

- Linear process:

$$\begin{aligned} X_t &= \theta(B) w_t, \\ \theta(B) &= 1 + \theta_1 B + \theta_2 B^2 + ... \end{aligned}$$

with $\sum_{k=0}^{\infty} \theta_k^2 < \infty$. Linear $\Rightarrow$ stationary;

$$\gamma_X(m) = \sigma_w^2 \sum_{k=0}^{\infty} \theta_k \theta_{k+m}.$$

- Linear + "$\theta_k = 0$ for $k > q$": MA(q) process, $\gamma(m) = 0$ for $m > q$. Characteristic polynomial written as $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q$.

- Invertible process:

$$\begin{aligned} \phi(B)X_t &= w_t, \\ \phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - ... \end{aligned}$$

with $\sum_k |\phi_k| < \infty$. Note this is really $\dot{X}_t$; a non-zero mean can be accommodated as follows:

$$\begin{aligned} w_t &= \phi(B)\dot{X}_t = \dot{X}_t - \phi_1 \dot{X}_{t-1} - ... \\ &= (X_t - \mu) - \phi_1 (X_{t-1} - \mu) - ... \\ &= \{X_t - \phi_1 X_{t-1} - ...\} - \mu \{1 - \phi_1 - \phi_2 - ...\} \\ &= \phi(B)X_t - \alpha, \end{aligned}$$

where $\alpha = \mu\phi(1)$.

- Invertible + "$\phi_j = 0$ for $j > p$": AR(p) process.

- Wold's Theorem: Stationary $\Leftrightarrow$ Linear.

- A stationary process is invertible iff all roots of $\theta(B) = 0$ lie outside the unit circle in the complex plane. Thus an MA(q) is stationary (linear), not necessarily invertible.

- An invertible process is stationary iff all roots of $\phi(B) = 0$ lie outside the unit circle. Thus an AR(p) is invertible, not necessarily stationary.

- Example: MA(2).

$$\begin{aligned} X_t &= w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2}, \\ \theta(B) &= 1 + \theta_1 B + \theta_2 B^2. \end{aligned}$$

If $\theta_1^2 - 4\theta_2 < 0$ (so both roots are complex), then invertibility requires $0 < \theta_2 < 1$. Suppose this is so. To invert (why? Prediction!):
we need $\theta(B)\phi(B) = 1$, where
$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots$, so

$$\begin{aligned} 1 &= \left\{ \begin{array}{c} \left[1 + \theta_1 B + \theta_2 B^2\right] \cdot \\ \left[1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_k B^k - \ldots\right] \end{array} \right\} \\ &= 1 - (\phi_1 - \theta_1) B - (\phi_2 + \theta_1 \phi_1 - \theta_2) B^2 - \\ &\quad \ldots - (\phi_k + \phi_{k-1}\theta_1 + \phi_{k-2}\theta_2) B^k + \ldots . \end{aligned}$$

Matching coefficients:

$$\begin{aligned} \phi_1 &= \theta_1, \\ \phi_2 &= \theta_2 - \theta_1 \phi_1 = \theta_2 - \theta_1^2, \\ \phi_k &= -\phi_{k-1}\theta_1 - \phi_{k-2}\theta_2, \ k = 2, 3, \ldots . \end{aligned}$$

- ARMA models are defined in operator notation by $\phi(B)X_t = \theta(B)w_t$; if $\phi(B)$ is an AR(p) characteristic polynomial and $\theta(B)$ an MA(q), we say $\{X_t\}$ is an ARMA(p,q) process. It is stationary (linear, causal) if $X_t = \psi(B)w_t$ for a series $\psi(B) = \sum_k \psi_k B^k$, with square summable coefficients. The coefficients ψ_k are determined from $\theta(B)/\phi(B) = \psi(B)$. It can be shown that $\psi(B)$ has the required (stationarity) properties only if all zeroes of $\phi(B)$ lie outside the unit circle. Similarly an ARMA(p,q) is invertible only if all zeroes of $\theta(B)$ lie outside the unit circle. We also require that the AR and MA polynomials have no common factors.

- Example: (Example 3.7 in text)

$$
\begin{aligned}
& X_t = .4X_{t-1} + .45X_{t-2} + w_t + w_{t-1} + .25w_{t-2} \\
\Rightarrow\ & \left(1 - .4B - .45B^2\right) X_t = \left(1 + B + .25B^2\right) w_t \\
\Rightarrow\ & (1 - .9B)(1 + .5B) X_t = (1 + .5B)(1 + .5B) w_t \\
\Rightarrow\ & (1 - .9B) X_t = (1 + .5B) w_t.
\end{aligned}
$$

Thus series is both stationary and invertible. It is ARMA(1,1), not ARMA(2,2) as it initially appeared. You should verify that the above can be continued as

$$X_t = \left[\sum_{k=0}^{\infty} (.9)^k B^k \cdot (1 + .5B)\right] w_t$$

$$= \left[1 + (.9 + .5)B + ... + (.9)^{k-1}(.9 + .5)B^k + ...\right] w_t$$

$$= \psi(B) w_t$$

where $\psi(B) = \sum_k \psi_k B^k$ and $\psi_0 = 1$, $\psi_k = 1.4(.9)^{k-1}$ for $k > 0$.

- Box-Jenkins methodology:

  1. Determine the theoretical ACF (and PACF, to be defined) for these and other classes of time series models. Use the sample ACF/PACF to match the data to a possible model (MA(q), AR(p), etc.).

2. Estimate parameters using a method appropriate to the chosen model and assess the fit, primarily by studying the residuals. The notion of *residuals* will require a special treatment, for now think of them as $X_t - \hat{X}_t$ where, e.g., in an AR(1) model $X_t = \phi_1 X_{t-1} + w_t$ we have $\hat{X}_t = \hat{\phi}_1 X_{t-1}$. The residuals should then "look like" white noise (why?). If the fit is inadequate revise steps 1. and 2.

3. Finally use model to forecast.

- We treat these three steps in detail. Recall that for an MA(q), the autocovariance function is

$$\gamma(m) = \begin{cases} \sigma_w^2 \sum_{k=0}^{q-m} \theta_k \theta_{k+m}, & 0 \le m \le q, \\ 0 & m > q. \end{cases}$$

The salient feature is that $\underline{\gamma(m) = 0 \text{ for } m > q}$; we look for this in the sample ACF. See Figure 6.1.

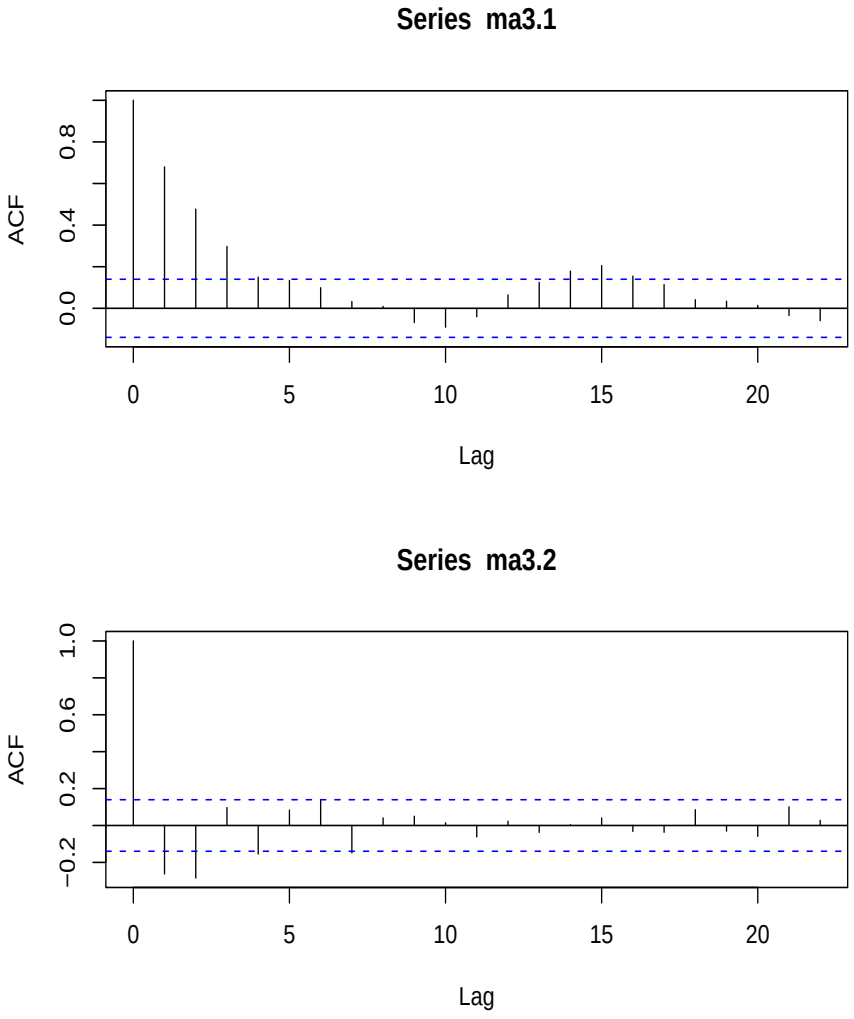

Figure 6.1. Sample ACFs for two simulated MA(3) series. Top: $(\theta_1, \theta_2, \theta_3) = (.5, .5, .2)$. Bottom: $(\theta_1, \theta_2, \theta_3) = (-.5, -.5, .2)$. Would you correctly guess that the underlying model is MA(3)?

- We derive the ACF of a stationary AR(p) process:

$$w_t = X_t - \sum_{i=1}^{p} \phi_i X_{t-i},$$

so that for $m \geq 0$,

$$\begin{aligned} COV\left[w_t,\ X_{t-m}\right] &= COV\left[X_t - \sum_{i=1}^{p} \phi_i X_{t-i},\ X_{t-m}\right] \\ &= \gamma(m) - \sum_{i=1}^{p} \phi_i \gamma(m-i). \end{aligned}$$

Because of stationarity, X_{t-m} is a linear combination $w_{t-m} + \theta_1 w_{t-m-1} + \theta_2 w_{t-m-2} + ..$ with (**this is important!**)

$$\begin{aligned} COV\left[w_t,\ X_{t-m}\right] &= COV\left[\begin{array}{c} w_t,\ w_{t-m} + \theta_1 w_{t-m-1} \\ +\theta_2 w_{t-m-2} + .. \end{array}\right] \\ &= \sigma_w^2 I(m=0), \end{aligned}$$

thus

$$\gamma(m) - \sum_{i=1}^{p} \phi_i \gamma(m-i) = \left\{ \begin{array}{ll} \sigma_w^2, & m = 0, \\ 0 & m > 0. \end{array} \right.$$

These are the "Yule-Walker" equations to be solved to obtain $\gamma(m)$ for $m \geq 0$, then $\gamma(-m) = \gamma(m)$.

- Example: AR(1). Yule-Walker equations are

$$\gamma(m) - \phi\gamma(m-1) = \begin{cases} \sigma_w^2, & m = 0, \\ 0 & m > 0. \end{cases}$$

We get

$$\begin{aligned} \gamma(0) &= \phi\gamma(1) + \sigma_w^2, \\ \gamma(m) &= \phi\gamma(m-1) \text{ for } m > 0. \end{aligned}$$

In particular

$$\begin{aligned} \gamma(0) &= \phi\gamma(1) + \sigma_w^2 \\ &= \phi\left(\phi\gamma(0)\right) + \sigma_w^2, \end{aligned}$$

so

$$\gamma(0) = \frac{\sigma_w^2}{1-\phi^2}.$$

Note that $0 < \gamma(0) = VAR[X_t] < \infty$ by the stationarity condition $|\phi| < 1$.

Iterating $\gamma(m) = \phi\gamma(m-1)$ gives

$$\gamma(m) = \phi^m\gamma(0),\ m = 1, 2, 3, \ldots .$$

Thus

$$\rho(m) = \frac{\gamma(m)}{\gamma(0)} = \phi^{|m|}.$$

7. Partial autocorrelation

- Difficult to identify an AR(p) from its ACF (Figure 7.1).

- Suppose that a series is AR(1), and consider forecasting X_t from two previous values $X_{t-1,}X_{t-2}$:

$$\begin{aligned} X_t &= \phi X_{t-1} + w_t, \\ \hat{X}_t &= \alpha_1 X_{t-1} + \alpha_2 X_{t-2}. \end{aligned}$$

  One suspects that the "best" α's will be $\alpha_1 = \phi, \alpha_2 = 0$. This is in fact true, and is a property of the "Partial Autocorrelation Function" (PACF).

- Assume $\mu_X = 0$; consider the problem of minimizing the MSE when X_t is forecast by

$$\hat{X}_t = \alpha_{1,m} X_{t-1} + ... + \alpha_{m,m} X_{t-m}.$$

  This is

$$\begin{aligned} MSE &= E\left[\{X_t - \alpha_{1,m} X_{t-1} - ... - \alpha_{m,m} X_{t-m}\}^2\right] \\ &= f_m(\alpha_{1,m}, ..., \alpha_{m,m}), \end{aligned}$$

say. Let the minimizers be $\alpha^*_{1,m}, ..., \alpha^*_{m,m}$. The **lag-m PACF value**, written ϕ_{mm}, is defined to be $\alpha^*_{m,m}$.

- It can also be shown that

$$\phi_{mm} = CORR\left[X_t - \hat{X}_t, X_{t-m} - \hat{X}_{t-m}\right],$$

where each $\hat{X}$ denotes the best (i.e. minimum MSE) predictor which is a linear function of $X_{t-1}, ..., X_{t-m+1}$.

- To compute, we solve m equations in m unknowns and retain only the solution for $\alpha_{m,m}$. These equations are

$$\begin{aligned} 0 = -\frac{1}{2}\frac{\partial f_m}{\partial \alpha_{j,m}} = \\ E\left[X_{t-j} \cdot \{X_t - \alpha_{1,m}X_{t-1} - ... - \alpha_{m,m}X_{t-m}\}\right] \\ = \left[\gamma(j) - \alpha_{1,m}\gamma(j-1) - ... - \alpha_{m,m}\gamma(j-m)\right], \end{aligned}$$

for $j = 1, ..., m$, i.e.

$$\sum_{i=1}^{m} \alpha_{i,m}\gamma(j-i) = \gamma(j).$$

Then

$$\begin{aligned} m &= 1 : \phi_{11} = \rho(1), \\ m &= 2 : \phi_{22} = \frac{\rho(2) - \rho^2(1)}{1 - \rho^2(1)}, \text{ etc.} \end{aligned}$$

- Note that for an AR(1), $\rho(j) = \phi^j$ and so $\phi_{11} = \phi$, $\phi_{22} = 0$. See Figure 7.1.

- In general, *if* $\{X_t\}$ *is AR(p) and stationary, then* $\phi_{pp} = \phi_p$ *and* $\phi_{mm} = 0$ *for* $m > p$.

  Proof: Write $X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + w_t$, so for $m \geq p$,

$$\begin{aligned} & f_m(\alpha_{1,m}, ..., \alpha_{m,m}) \\ = \; & E\left[\left\{ \begin{array}{c} w_t + \sum_{j=1}^{p} \left(\phi_j - \alpha_{j,m}\right) X_{t-j} \\ - \sum_{j=p+1}^{m} \alpha_{j,m} X_{t-j} \end{array} \right\}^2 \right] \\ = \; & E\left[\{w_t + Z\}^2\right], \text{ say,} \\ & \text{where } Z \text{ is uncorrelated with } w_t \text{ - why?,} \\ = \; & \sigma_w^2 + E[Z^2]. \end{aligned}$$

  This is minimized if $Z = 0$ with probability 1, i.e. if $\alpha_{j,m} = \phi_j$ for $j \leq p$ and $= 0$ for $j > p$.

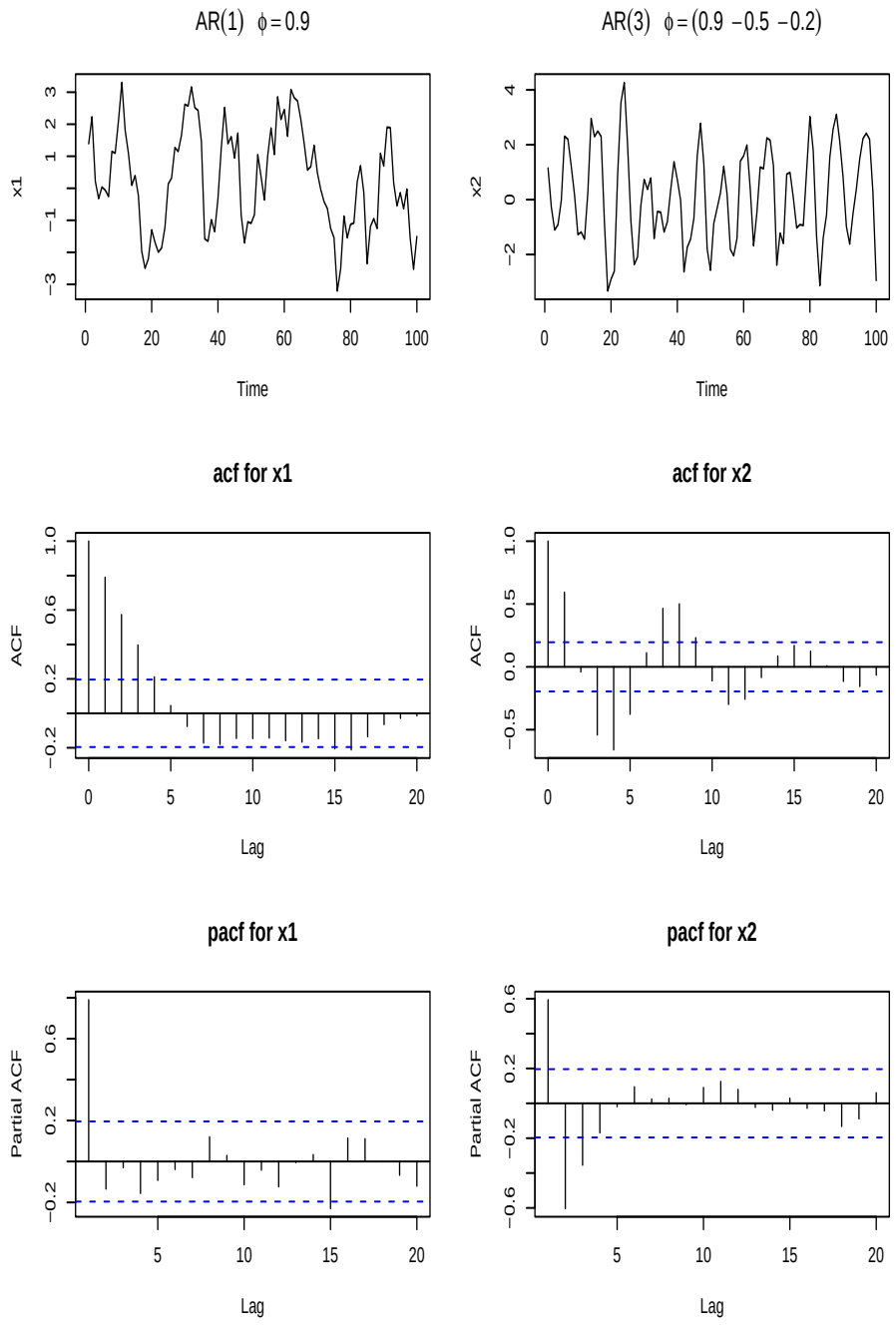

Figure 7.1. Simulated AR series and sample PACFs. Note that there is no 'lag 0' in a PACF plot.

- Forecasting. Given r.v.s $X_t, X_{t-1}, ...$ (into the infinite past, in principle) we wish to forecast a future value X_{t+l}. Let the forecast be X^t_{t+l}. We will later show that the "best" forecast is

$$X^t_{t+l} = E\left[X_{t+l}|X_t, X_{t-1}, ...\right],$$

the conditional expected value of X_{t+l} given $X^t \stackrel{def}{=} \{X_s\}^t_{s=-\infty}$. Our general model identification approach, following Box/Jenkins, is then:

1. Tentatively identify a model, generally by looking at its sample ACF/PACF.

2. Estimate the parameters (generally by the method of Maximum Likelihood, to be covered later). This allows us to estimate the forecasts X^t_{t+l}, which depend on unknown parameters, by substituting estimates to obtain $\hat{X}^t_{t+l}$. We define the *residuals* by

$$\hat{w}_t = X_t - \hat{X}^{t-1}_t.$$

The (adjusted) MLE of σ_w^2 is typically (in $ARMA(p,q)$ models)

$$\hat{\sigma}_w^2 = \frac{1}{n-2p-q} \sum_{t=p+1}^{n} \hat{w}_t^2.$$

The $n-2p-q$ in the denominator is the number of residuals used ($n-p$) minus the number of ARMA parameters estimated ($p+q$).

3. The residuals should "look like" white noise. We study them, and apply various tests of whiteness. To the extent that they are not white, we look for possible alternate models.

4. Iterate; finally use the model to forecast.

8. Forecasting I

- Conditional expectation. Example: Randomly choose a stock from a listing. $Y =$ price in one week, $X =$ price in the previous week. To predict Y, if we have *no* information about X then the best (minimum mse) *constant predictor* of Y is $E[Y]$. (Why? What mathematical problem is being formulated and solved here?) However, suppose we also know that $X = x$. Then we still forecast using the mean of Y, but this now is calculated with respect to the conditional distribution of Y, given that $X = x$. The forecast becomes $E[Y|X = x]$, the mean price of all stocks whose price in the previous week was x.

- In general, if (X, Y) are any r.v.s, then $E[Y|X = x]$ is the expected value of Y, when the population is restricted to those pairs with $X = x$. We use the following facts about conditional expectation.

(i) (X, Y) independent $\Rightarrow E[Y|X = x] = E[Y]$.

(ii) $E[X|X = x] = x$.

(iii) $E[g(X)|X = x] = g(x)$, and more generally

(iv) $E[g(X, Y)|X = x] = E[g(x, Y)|X = x]$.

In particular,

(v) $E[f(X)g(X, Y)|X = x] = f(x)E[g(x, Y)|X = x]$.

- Since the (Normally distributed) white noise terms are independent, we have (**this is important!**):

$$E\left[w_s|w^t\right] \stackrel{def}{=} E\left[w_s|w_t, w_{t-1}, ...\right] = \begin{cases} w_s, & s \leq t, \\ 0, & s > t. \end{cases}$$

- Put $h(x) = E[Y|X = x]$. This is a function of x; as a function of the r.v. X we call it $h(X) = E[Y|X]$ (a r.v. itself). We have the

Double Expectation Theorem:

$$E\left\{E[Y|X]\right\} = E[Y].$$

The inner expectation is with respect to the conditional distribution and is often written $E_{Y|X}[\cdot]$. The outer is with respect to the distribution of X; the theorem can be stated as $E[h(X)] = E[Y]$, where $h(x)$ is as above.

- Example: Y = house values, X = location (neighbourhood) of a house. $E[Y]$ can be obtained by averaging within neighbourhoods, then averaging over neighbourhoods.

- Similarly,

$$E_X\left\{E_{Y|X}\left[g(X,Y)|X\right]\right\} = E\left[g(X,Y)\right].$$

- Minimum MSE forecasting. Consider forecasting a r.v. Y (unobservable) using another r.v. X (or set of r.v.s); e.g. $Y = X_{t+l}$, $X = X^t$. We seek the function $g(X)$ which minimizes the MSE

$$MSE(g) = E\left[\{Y - g(X)\}^2\right].$$

The required function is $g(X) = E[Y|X]$ $(= h(X))$.

- **Proof**: We have to show that for any function g, $MSE(g) \geq MSE(h)$. Write

$$\begin{aligned} & MSE(g) \\ = \; & E\left[\{(Y - h(X)) + (h(X) - g(X))\}^2\right] \\ = \; & E\left[\{Y - h(X)\}^2\right] + E\left[\{h(X) - g(X)\}^2\right] \\ & +2E\left[(h(X) - g(X)) \cdot (Y - h(X))\right]. \end{aligned}$$

We will show that the last term $= 0$; then we have $MSE(g) = MSE(h) + E\left[\{h(X) - g(X)\}^2\right]$ which exceeds $MSE(h)$; equality iff $g(X) = h(X)$ with probability 1. To establish the claim we evaluate the expected value in stages:

$$\begin{aligned} & E\left[(h(X) - g(X)) \cdot (Y - h(X))\right] \\ = \; & E_X\left\{E_{Y|X}\left[(h(X) - g(X)) \cdot (Y - h(X))|X\right]\right\}. \end{aligned}$$

The inner expectation is (why?)

$$\begin{aligned} & (h(X) - g(X) \cdot E_{Y|X}\left[(Y - h(X))|X\right] \\ = \; & (h(X) - g(X) \cdot \left\{E_{Y|X}\left[Y|X\right] - h(X)\right\} \\ = \; & 0. \end{aligned}$$

- The minimum MSE is

$$\begin{aligned} MSE_{\text{min}} &= E\left[\{Y - h(X)\}^2\right] \\ &= E_X\left\{E_{Y|X}\left[\{Y - h(X)\}^2 \,|X\right]\right\} \\ &= E_X\left\{VAR[Y|X]\right\}. \end{aligned}$$

  We will show that

$$VAR[Y] = E_X\left\{VAR[Y|X]\right\} + VAR[E\{Y|X\}],$$

  i.e.

$$VAR[Y] = MSE_{\text{min}} + VAR[h(X)];$$

  thus $MSE_{\text{min}} \leq VAR[Y]$. $VAR[Y]$ is the MSE when Y is forecast by its mean and X is ignored; our result is then that using the information in X never increases MSE, and results in a strict decrease as long as $VAR[h(X)] > 0$, i.e. $h(x)$ is non-constant. (When would it be constant? i.e. when would $E[Y|X]$ not depend on X?)

- Analogous to "within" and "between" breakdown of variation in ANOVA. E.g. variation in house prices within (MSE_{min}) and between ($VAR[h(X)]$) neighbourhoods.

- **Proof of claim**:

$$\begin{aligned} & VAR[Y] = E\left[\{Y - E[Y]\}^2\right] \\ = \; & E\left[\{Y - h(X) + (h(X) - E[Y])\}^2\right] \\ = \; & E\left[\{Y - h(X)\}^2\right] + E\left[\{h(X) - E[Y]\}^2\right] \\ & +2E\left[\{Y - h(X)\}\{h(X) - E[Y]\}\right] \\ = \; & MSE_{\min} + VAR[h(X)] \\ & +2E_X\left\{E_{Y|X}\left[\{Y - h(X)\}\{h(X) - E[Y]\}|X\right]\right\}. \end{aligned}$$

  The inner expectation is

$$\{h(X) - E[Y]\} \cdot E_{Y|X}\left[Y - h(X)|X\right] = 0.$$

- You should (and will, on the current assignment) verify that $Y - E[Y|X]$ is uncorrelated with X.

- Assume $\{X_t\}$ is stationary and invertible. We forecast X_{t+l} by $X^t_{t+l} = E\left[X_{t+l}|X^t\right]$, where $X^t \stackrel{def}{=} \{X_s\}^t_{s=-\infty}$. Note that this forecast is 'unbiased' in that $E[X^t_{t+l}] = E[X_{t+l}]$. By the linearity we have that X_{t+l} can be represented as

$$X_{t+l} = \sum_{k=0}^{\infty} \psi_k w_{t+l-k}, \ (\psi_0 = 1)$$

so that

$$X^t_{t+l} = \sum_{k=0}^{\infty} \psi_k E[w_{t+l-k}|X^t].$$

We have $X_t = \psi(B)w_t$ and (by invertibility) $w_t = \phi(B)X_t$ where $\phi(B)\psi(B) = 1$ determines $\phi(B)$. Thus (**important!**): conditioning on X^t is equivalent to conditioning on w^t:

$$X^t_{t+l} = \sum_{k=0}^{\infty} \psi_k E[w_{t+l-k}|w^t] \text{ where}$$

$$E[w_{t+l-k}|w^t] = \begin{cases} w_{t+l-k}, & \text{if } l \leq k, \\ 0, & \text{otherwise.} \end{cases}$$

[**Of course**: $X^t_{t+l} = X_{t+l}$ if $l \leq 0$.]

Thus the forecast is

$$X_{t+l}^t = \sum_{k=l}^{\infty} \psi_k w_{t+l-k},$$

with forecast error and variance

$$\begin{aligned} X_{t+l} - X_{t+l}^t &= \sum_{k=0}^{l-1} \psi_k w_{t+l-k}, \\ VAR[X_{t+l} - X_{t+l}^t] &= \sigma_w^2 \sum_{k=0}^{l-1} \psi_k^2. \end{aligned}$$

Since $\{w_t\}$ is normal, we have

$$X_{t+l} - X_{t+l}^t \sim N\left(0, \sigma_w^2 \sum_{k=0}^{l-1} \psi_k^2\right)$$

and so a $100(1-\alpha)\%$ prediction (forecast) interval on X_{t+l} is

$$X_{t+l}^t \pm z_{\alpha/2}\sigma_w \sqrt{\sum_{k=0}^{l-1} \psi_k^2}.$$

Interpretation: *the probability that* X_{t+l} *will lie in this interval is* $1-\alpha$.

- If the history X^t was ignored, the interval would be $E[X_{t+l}] \pm z_{\alpha/2}\sqrt{\text{var}[X_{t+l}]} = \mu_X \pm z_{\alpha/2}\sigma_X$, where $\sigma_X = \sigma_w\sqrt{\sum_{k=0}^{\infty}\psi_k^2}$ is generally much larger than the above.

- In practice we must solve for the ψ_k in terms of the AR and MA parameters of $\{X_t\}$, then substitute estimates of these parameters to obtain estimates $\hat{\psi}_k$. Substituting these estimates into the expressions above results in the forecast $\hat{X}_{t+l}^t$; we also must use an estimate $\hat{\sigma}_w^2$. The *residuals*, or *innovations* are

  $\hat{w}_t = X_t - X_t^{t-1}$ with all parameters estimated,

  and typically

  $$\hat{\sigma}_w^2 = \frac{\sum \hat{w}_t^2}{\#\text{ of residuals - }\#\text{ of parameters estimated}}.$$

- An easy calculation (which you should do) shows that $X_t - X_t^{t-1} = w_t = \phi(B)X_t$, so that (if it is easier) $\hat{w}_t = \hat{\phi}(B)X_t$. In other words, *the residual can be obtained by writing the white noise in terms of X_t and then estimating the coefficients*.

9. Forecasting II

Review: a $100(1-\alpha)\%$ prediction (forecast) interval on a future value X_{t+l} is

$$X_{t+l}^t \pm z_{\alpha/2}\sigma_w\sqrt{\sum_{k=0}^{l-1}\psi_k^2}.$$

This assumes the series is stationary:

$$X_{t+l} = \sum_{k=0}^{\infty}\psi_k w_{t+l-k},\ (\psi_0 = 1) \qquad (9.1)$$

and invertible:

$$X_{t+l} = w_{t+l} + \sum_{k=1}^{\infty}\phi_k X_{t+l-k}. \qquad (9.2)$$

- Predicted value: For an autoregression the model specifies how the series depends on its past, and so writing down (a computing algorithm for) the

prediction is trivial. From (9.2):

$$\begin{aligned} X_{t+l}^{t} &= E\left[X_{t+l}|X^{t}\right] = \sum_{k=1}^{\infty} \phi_k E\left[X_{t+l-k}|X^{t}\right] \\ &= \sum_{k=1}^{l-1} \phi_k X_{t+l-k}^{t} + \sum_{k=l}^{\infty} \phi_k X_{t+l-k}. \end{aligned}$$

(For an AR(p) only $\phi_1, ..., \phi_p$ are nonzero.) For a MA $X_t = \theta(B) w_t$ we have to first compute the coefficients of $\phi(B) = 1/\theta(B)$.

- Standard error of the prediction: For a moving average, the coefficients in (9.1) are specified by the model in MA form, and so writing down the s.e. of the prediction is trivial. For an autoregression $w_t = \phi(B) X_t$ we have to first compute the coefficients of $\psi(B) = 1/\phi(B)$.

- In all cases the coefficients must be estimated, generally by Maximum Likelihood.

- Example 1: AR(1) (and stationary).

$$\begin{aligned} X_t &= \phi X_{t-1} + w_t. \\ X_{t+l}^t &= E\left[X_{t+l}|X^t\right] = E\left[\phi X_{t+l-1} + w_{t+l}|X^t\right] \\ &= \phi E\left[X_{t+l-1}|X^t\right] + E\left[w_{t+l}|w^t\right] \\ &= \begin{cases} \phi X_t, & l = 1, \\ \phi X_{t+l-1}^t, & l > 1. \end{cases} \end{aligned}$$

Iterating:

$$X_{t+l}^t = \phi^l X_t \text{ for } l \geq 1.$$

The calculation of the forecast was easy; determining the forecast variance requires us to determine the ψ_k's. Usually this is done numerically; in the present case it can be done explicitly:

$$\begin{aligned} (1 - \phi B) X_t &= w_t, \text{ so} \\ X_t &= (1 - \phi B)^{-1} w_t \\ &= \sum_{k=0}^{\infty} \psi_k w_{t-k}, \end{aligned}$$

for $\psi_k = \phi^k$. Then

$$\sum_{k=0}^{l-1} \psi_k^2 = \frac{1 - \phi^{2l}}{1 - \phi^2},$$

leading to the forecast interval

$$\phi^l X_t \pm z_{\alpha/2}\sigma_w\sqrt{\frac{1-\phi^{2l}}{1-\phi^2}}.$$

Numerically we replace ϕ by its estimate $\hat{\phi}$; then $\hat{X}^t_{t+l} = \hat{\phi}^l X_t$ and the residuals are

$$\hat{w}_t = X_t - \hat{X}^{t-1}_t = X_t - \hat{\phi}X_{t-1} \ (t > 1).$$

Note the similarity with $w_t = X_t - \phi X_{t-1}$. This illustrates (recall the final note in the last class) that *the residual can also be obtained by writing w_t in terms of the data and parameters, and then replacing the parameters with estimates*.

The estimate of the variance of the noise is

$$\hat{\sigma}^2_w = \frac{\sum_{t=2}^n \hat{w}_t^2}{n-2}.$$

- Example 2. AR(p). Similar to Example 1,

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + w_t$$

results in

$$\hat{X}_{t+l}^t = \sum_{i=1}^{p} \hat{\phi}_i X_{t+l-i}^t,$$

where $X_{t+l-i}^t = X_{t+l-i}$ if $l \leq i$. Now solve (numerically) $X_t = (1/\phi(B))w_t = \psi(B)w_t$ to get the ψ_k, then the $\hat{\psi}_k$ and the standard errors of the forecasts. The innovations are obtained from

$$\hat{X}_t^{t-1} = \sum_{i=1}^{p} \hat{\phi}_i X_{t-i}^{t-1} = \sum_{i=1}^{p} \hat{\phi}_i X_{t-i}$$

to get

$$\hat{w}_t = X_t - \sum_{i=1}^{p} \hat{\phi}_i X_{t-i}, \ (t > p),$$

with

$$\hat{\sigma}_w^2 = \frac{\sum_{t=p+1}^{n} \hat{w}_t^2}{n - 2p}.$$

- Example 3. MA(1) (and invertible).

$$\begin{aligned} X_t &= w_t + \theta w_{t-1} = (1+\theta B) w_t \\ \Rightarrow w_t &= \sum_{k=0}^{\infty} (-\theta)^k X_{t-k}. \end{aligned}$$

We make the approximation $X^0 = w^0 = 0$, and then

$$w_t = \sum_{k=0}^{t-1} (-\theta)^k X_{t-k}.$$

Now

$$X_{t+l}^t = E\left[w_{t+l} + \theta w_{t+l-1} | w^t\right] = \begin{cases} \theta w_t, & l = 1, \\ 0, & l > 1, \end{cases}$$

with w_t obtained from the preceding equation. The residuals are

$$\hat{w}_t = \sum_{k=0}^{t-1} \left(-\hat{\theta}\right)^k X_{t-k},$$

and

$$\hat{\sigma}_w^2 = \frac{\sum_{t=1}^n \hat{w}_t^2}{n-1}.$$

Trivially, since $\psi_0 = 1, \psi_1 = \theta$ and $\psi_k = 0$ for $k > 1$, we have

$$\sum_{k=0}^{l-1} \psi_k^2 = \begin{cases} 1, & l = 1, \\ 1 + \theta^2, & l > 1. \end{cases}$$

The prediction intervals are

$$\begin{cases} \hat{\theta}\hat{w}_t \pm z_{\alpha/2}\hat{\sigma}_w, & l = 1, \\ 0 \pm z_{\alpha/2}\hat{\sigma}_w\sqrt{1 + \hat{\theta}^2}, & l > 1. \end{cases}$$

- You should write out the procedure for an invertible MA(q) model.

- Example 4. ARMA(1,1), stationary and invertible. In general, *when there is an MA component we make the approximation* $X^0 = w^0 = 0$. The model is $(1 - \phi B)X_t = (1 + \theta B)w_t$, i.e. $X_t = \phi X_{t-1} + w_t + \theta w_{t-1}$, leading to

$$\begin{aligned} X_{t+l}^t &= \phi X_{t+l-1}^t + w_{t+l}^t + \theta w_{t+l-1}^t \\ &= \begin{cases} \phi X_t + \theta w_t, & l = 1, \\ \phi X_{t+l-1}^t, & l > 1. \end{cases} \end{aligned}$$

To obtain a value for w_t we write

$$\begin{aligned} w_t &= (1-\phi B)(1+\theta B)^{-1} X_t \\ &= (1-\phi B) \sum_{k=0}^{\infty} (-\theta)^k B^k \cdot X_t \\ &= \left(1 - \sum_{k=1}^{\infty} (-\theta)^{k-1} (\theta+\phi) B^k\right) X_t \end{aligned}$$

with approximation

$$w_t \approx X_t - \sum_{k=1}^{t-1} (-\theta)^{k-1} (\theta+\phi) X_{t-k}.$$

For the forecast variance we change ϕ to $-\theta$ and θ to $-\phi$ in the above:

$$X_t = (1+\theta B)(1-\phi B)^{-1} w_t = \sum_{k=0}^{\infty} \psi_k w_{t-k}$$

with $\psi_k = \phi^{k-1}(\phi+\theta)$ (and $\psi_0 = 1$); thus

$$\begin{aligned} & VAR\left[X_{t+l} - X_{t+l}^t\right] \\ &= \sigma_w^2 \sum_{k=0}^{l-1} \psi_k^2 \\ &= \sigma_w^2 \left[1 + (\phi+\theta)^2 \frac{1-\phi^{2(l-1)}}{1-\phi^2}\right]. \end{aligned}$$

The residuals $\hat{w}_t = X_t - \hat{X}_t^{t-1}$ are obtained by substituting estimates into the expression for w_t given above, i.e. $\hat{w}_1 = X_1$ and for $t > 1$:

$$\hat{w}_t = X_t - \sum_{k=1}^{t-1} \left(-\hat{\theta}\right)^{k-1} \left(\hat{\theta} + \hat{\phi}\right) X_{t-k}.$$

Then we take

$$\hat{\sigma}_w^2 = \frac{\sum_{t=2}^{n} \hat{w}_t^2}{n-3}.$$

- This example illustrates another general point - in an ARMA(p, q) model, the assumption $X^0 = 0$ is necessary if we are to be able to calculate *any* of the residuals, but it then allows us to approximate *all* of them. But we typically ignore the first p of them, for agreement with the pure AR(p) model and since their approximations are often quite poor. This gives

$$\hat{\sigma}_w^2 = \frac{\sum_{t=p+1}^{n} \hat{w}_t^2}{n-2p-q}.$$

10. Estimation I

- Estimation. One method is the **method of moments**, in which we take expressions relating parameters to expected values, replace the expected values by series averages, then solve for the unknown parameters. ("Plug-in" estimates.)

  - e.g. $E[X_t] = \mu$ becomes $n^{-1} \sum_{t=1}^{n} x_t = \hat{\mu}$.

  - e.g. For a zero-mean AR(1) model we could replace $\gamma(k)$ by the sample autocovariance $\hat{\gamma}(k)$ in the Yule-Walker equations, then solve them as before to get

$$\begin{aligned} \hat{\gamma}(0) &= \frac{\hat{\sigma}_w^2}{1 - \hat{\phi}^2}, \\ \hat{\gamma}(1) &= \hat{\phi}\hat{\gamma}(0), \end{aligned}$$

    yielding

$$\begin{aligned} \hat{\phi} &= \hat{\rho}(1), \\ \hat{\sigma}_{w,YW}^2 &= \hat{\gamma}(0)\left(1 - \hat{\phi}^2\right). \end{aligned}$$

Recall we previously used the (adjusted Maximum Likelihood) estimate

$$\hat{\sigma}^2_{w,MLE} = \sum_{t=2}^{n} \hat{w}_t^2/(n-2).$$

You should show that the difference between these two estimates is of the order (i.e. a multiple of) $1/n$; in this sense the two estimates are *asymptotically* (i.e. as $n \to \infty$) *equivalent*. In fact

$$\hat{\sigma}^2_{w,MLE} = \frac{n}{n-2}\left[\hat{\sigma}^2_{w,YW} - \frac{x_1^2 + \left(\hat{\phi}x_n\right)^2}{n}\right].$$

- The same technique applied to the MA(1) model starts with $\rho(1) = \theta/\left(1+\theta^2\right)$ ($|\theta| < 1$ for invertibility, and then $|\rho(1)| < 1/2$), then we solve

  $$\hat{\rho}(1) = \hat{\theta}/\left(1+\hat{\theta}^2\right).$$

  If $|\hat{\rho}(1)| < 1/2$ there is a real root $\hat{\theta}$ with $|\hat{\theta}| < 1$ and we use it, obtaining an invertible model. (Otherwise this method fails us.) But even then the estimate can be quite inefficient (highly varied) relative to the MLE, which we consider next.

- **Maximum Likelihood Estimation**. We observe $\mathbf{x} = (x_1, ..., x_n)'$; suppose the joint probability density function (pdf) is $f(\mathbf{x}|\boldsymbol{\alpha})$ for a vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_p)'$ of unknown parameters. E.g. if the X_t are independent $N(\mu, \sigma^2)$ the joint pdf is

$$\begin{aligned} & \prod_{t=1}^{n} \left\{ \left(2\pi\sigma^2\right)^{-1/2} e^{-\frac{(x_t-\mu)^2}{2\sigma^2}} \right\} \\ = \quad & \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{\sum_{t=1}^{n}(x_t-\mu)^2}{2\sigma^2}}. \end{aligned}$$

  When evaluated at the numerical data this is a function of $\boldsymbol{\alpha}$ $(= \left(\mu, \sigma^2\right)'$ in this example) alone, denoted $L(\boldsymbol{\alpha}|\mathbf{x})$ and known as the *Likelihood function.* The value $\hat{\boldsymbol{\alpha}}$ which maximizes $L(\boldsymbol{\alpha}|\mathbf{x})$ is known as the Maximum Likelihood Estimator (MLE). Intuitively, the MLE makes the observed data "most likely to have occurred".

- We put $l(\boldsymbol{\alpha}) = \ln L(\boldsymbol{\alpha}|\mathbf{x})$, the log-likelihood, and typically maximize it (equivalent to maximizing L) by solving the *likelihood equations*

$$\begin{aligned} \dot{l}(\boldsymbol{\alpha}) &= \mathbf{0}, \text{ where} \\ \dot{l}(\boldsymbol{\alpha}) &= \left(\frac{\partial l(\boldsymbol{\alpha})}{\partial \alpha_1}, ..., \frac{\partial l(\boldsymbol{\alpha})}{\partial \alpha_p} \right)'. \end{aligned}$$

  The column vector $\dot{l}(\boldsymbol{\alpha})$ is called the *gradient*.

- The MLE has attractive large sample properties. With $\boldsymbol{\alpha}_0$ denoting the true value, we typically have that for large n,

$$\hat{\boldsymbol{\alpha}} \overset{d}{\approx} N\left(\boldsymbol{\alpha}_0, \frac{1}{n}\mathbf{C}\right) \text{ for } \mathbf{C} = \mathbf{I}^{-1}(\boldsymbol{\alpha}_0),$$

  where $\mathbf{I}(\boldsymbol{\alpha}_0)$ is the *information matrix* defined below.

- **Meaning**: Approximate normality (exact as $n \to \infty$), $E\left[\hat{\alpha}_j\right] = \alpha_{0j}$, cov$\left[\hat{\alpha}_j, \hat{\alpha}_k\right] = C_{jk}/n$; in particular var$\left[\hat{\alpha}_j\right] = C_{jj}/n$.

- The information matrix is given by

$$\mathbf{I}(\boldsymbol{\alpha}_0) = \lim_{n\to\infty} \left\{ \frac{1}{n} E\left[-\ddot{l}(\boldsymbol{\alpha}_0) \right] \right\},$$

where $\ddot{l}(\boldsymbol{\alpha})$ is the *Hessian* matrix with $(j,k)^{th}$ element $\partial^2 l(\boldsymbol{\alpha})/\partial\alpha_j \partial\alpha_k$. (The expected value of a matrix is the matrix of expected values.)

- To apply these results we estimate $\mathbf{I}(\boldsymbol{\alpha}_0)$ by

$$\hat{\mathbf{I}} = \mathbf{I}(\hat{\boldsymbol{\alpha}}).$$

Denote the $(j,k)^{th}$ element of $\hat{\mathbf{I}}^{-1}$ by $\hat{\mathbf{I}}^{jk}$. Then the normal approximation is that $\hat{\alpha}_j$ is asymptotically normally distributed with mean α_{0j} and variance $\mathbf{I}^{jj}/n$, so that

$$\frac{\hat{\alpha}_j - \alpha_{0j}}{s_j} \overset{d}{\approx} N(0,1), \text{ where } s_j = \sqrt{\frac{\hat{\mathbf{I}}^{jj}}{n}}.$$

Then, e.g., the p-value for the hypothesis H_0 : $\alpha_{0j} = 0$ against a two-sided alternative is

$$p = 2P\left(Z > \left| \frac{\hat{\alpha}_j}{s_j} \right| \right).$$

Both $\hat{\alpha}_j$ and s_j are supplied on the R printout.

- Example 1. AR(1) with a constant: $X_t = \phi_0 + \phi_1 X_{t-1} + w_t$. As is commonly done for AR models we will carry out an analysis *conditional on* X_1; i.e. we act as if X_1 is not random, but is the constant x_1. We then carry out the following steps:

  1. Derive the pdf $f(\mathbf{x}|\boldsymbol{\alpha})$ of the "new variables" $X_2, ..., X_n$ from that of the "old variables" $w_2, ..., w_n$.

  2. From 1. the log-likelihood is $l(\boldsymbol{\alpha}) = \ln L(\boldsymbol{\alpha}|\mathbf{x})$, where $\boldsymbol{\alpha} = \left(\phi_0, \phi_1, \sigma_w^2\right)'$.

  3. Maximize $l(\boldsymbol{\alpha})$ to obtain the MLEs $\left(\hat{\phi}_0, \hat{\phi}_1, \hat{\sigma}_w^2\right)$.

  4. Obtain the information matrix and its estimated inverse.

- Step 1. *Transformation of variables*. If the pdf of $w_2, ..., w_n$ is $g(\mathbf{w}|\boldsymbol{\alpha})$ and we write the w's in terms of the X's:

$$w_t = X_t - \phi_0 - \phi_1 X_{t-1}$$

then the pdf of $X_2, ..., X_n$ is

$$f(\mathbf{x}|\boldsymbol{\alpha}) = g(\mathbf{w}|\boldsymbol{\alpha}) \left| \left(\frac{\partial \mathbf{w}}{\partial \mathbf{x}} \right) \right|_+$$

where $\left(\frac{\partial \mathbf{w}}{\partial \mathbf{x}}\right)$ is the ('Jacobian') matrix of partial derivatives, with

$$\left(\frac{\partial \mathbf{w}}{\partial \mathbf{x}} \right)_{jk} = \frac{\partial w_j}{\partial x_k},$$

and $|\cdot|_+$ is the absolute value of the determinant. On the right hand side $g(\mathbf{w}|\boldsymbol{\alpha})$ is evaluated by replacing the w's with their expressions in terms of the x's. In this AR(1) example,

$$g(\mathbf{w}|\boldsymbol{\alpha}) = \left(2\pi\sigma_w^2\right)^{-(n-1)/2} e^{-\frac{\sum_{t=2}^{n} w_t^2}{2\sigma_w^2}}$$

and

$$\frac{\partial \mathbf{w}}{\partial \mathbf{x}} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\phi_1 & 1 & 0 & \ddots & \vdots \\ 0 & -\phi_1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 & 0 \\ 0 & \cdots & 0 & -\phi_1 & 1 \end{pmatrix}$$

with determinant $= 1$ (why?). Thus

$$f(\mathbf{x}|\boldsymbol{\alpha}) = \left(2\pi\sigma_w^2\right)^{-(n-1)/2} e^{-\frac{\sum_{t=2}^{n}(x_t-\phi_0-\phi_1 x_{t-1})^2}{2\sigma_w^2}}.$$

- **Important**: This determinant will always $= 1$ if we can write w_t as X_t+a function of $X_{t-1}, ..., X_1$. But this can always be done for invertible models, if as well we assume that "$X^0 = 0$".

  So *for all models considered in this course*,

$$|\partial \mathbf{w}/\partial \mathbf{x}|_+ = 1.$$

- Step 2.

$$\begin{aligned} & l(\boldsymbol{\alpha}) \\ = \; & \ln\left\{\left(2\pi\sigma_w^2\right)^{-(n-1)/2} e^{-\frac{\sum_{t=2}^{n}(x_t-\phi_0-\phi_1 x_{t-1})^2}{2\sigma_w^2}}\right\} \\ = \; & -\frac{n-1}{2}\ln\sigma_w^2 - \frac{\sum_{t=2}^{n}\left(x_t-\phi_0-\phi_1 x_{t-1}\right)^2}{2\sigma_w^2} \\ & + const. \\ = \; & -\frac{n-1}{2}\ln\sigma_w^2 - \frac{S(\phi_0,\phi_1)}{2\sigma_w^2} + const., \end{aligned}$$

for

$$S(\phi_0, \phi_1) = \sum_{t=2}^{n} (x_t - \phi_0 - \phi_1 x_{t-1})^2 .$$

This "sum of squares" function is, and *always will be for the models in this course*, the sum of squares of the $\{w_t\}$, with each w_t replaced by its expression in terms of the data.

- Step 3. We must now maximize $l(\boldsymbol{\alpha})$ over the parameters $\phi_0, \phi_1, \sigma_w^2$ in $\boldsymbol{\alpha}$. First suppose that maximizers $\hat{\phi}_0$, $\hat{\phi}_1$ have already been obtained. Then the likelihood equation for σ_w^2 becomes

$$0 = \frac{\partial l}{\partial \sigma_w^2} = -\frac{n-1}{2\sigma_w^2} + \frac{S(\hat{\phi}_0, \hat{\phi}_1)}{2\sigma_w^4},$$

satisfied by

$$\begin{aligned} \hat{\sigma}_w^2 &= \frac{S(\hat{\phi}_0, \hat{\phi}_1)}{n-1} \\ &= \frac{\sum_{t=2}^{n} \left(x_t - \hat{\phi}_0 - \hat{\phi}_1 x_{t-1}\right)^2}{n-1} \\ &= \frac{\sum_{t=2}^{n} \hat{w}_t^2}{n-1}. \end{aligned}$$

A usual adjustment for bias is to replace $n-1$ by $n-3=(n-1)-2=\#$ of residuals - number of parameters estimated.

- It is important to note that so far *only the form of* $S(\cdot)$ *depends on the model*. Quite generally, in an ARMA(p,q) model the log-likelihood is

$$-\frac{n-p}{2}\ln\sigma_w^2-\frac{S(\phi,\theta)}{2\sigma_w^2}+const.$$

The AR parameters ϕ and the MA parameters θ appear only in the sum of squares of the $n-p$ noises:

$$S(\phi,\theta)=\sum_{t=p+1}^{n}w_t^2,$$

with each w_t replaced by its expression in terms of the data. Then the MLEs $\hat{\phi},\hat{\theta}$ are the *minimizers* of $S(\phi,\theta)$. Once these are obtained the (adjusted) MLE of σ_w^2 is

$$\hat{\sigma}_w^2=\frac{S\left(\hat{\phi},\hat{\theta}\right)}{df}=\frac{\sum_{t=p+1}^{n}\hat{w}_t^2}{df},$$

where $df=\#$ of residuals - $\#$ of parameters.

11. Estimation II

- Continuing with the AR(1) model ("Step 3"), we must now minimize

$$S(\phi_0, \phi_1) = \sum_{t=2}^{n} (x_t - \phi_0 - \phi_1 x_{t-1})^2 .$$

But this is exactly what is done by the least squares estimates computed from a regression

$$x_t = \phi_0 + \phi_1 x_{t-1} + error,\ t = 2, ..., n.$$

Let $\bar{x}_1$ and $\bar{x}_2$ be the averages of $\{x_{t-1}\}_{t=2}^{n}$ and $\{x_t\}_{t=2}^{n}$. Then the usual formulas for straight line regression give

$$\begin{aligned} \hat{\phi}_0 &= \bar{x}_2 - \hat{\phi}_1 \bar{x}_1, \text{ where} \\ \hat{\phi}_1 &= \frac{\sum_{t=2}^{n} (x_{t-1} - \bar{x}_1)(x_t - \bar{x}_2)}{\sum_{t=2}^{n} (x_{t-1} - \bar{x}_1)^2}. \end{aligned}$$

These are very nearly equal to the method of moment estimates:

$$\hat{\phi}_0 \approx \bar{x}\left(1 - \hat{\phi}_1\right),\ \hat{\phi}_1 \approx \hat{\rho}(1).$$

In general ARMA models, the minimizers of $S(\phi, \theta)$ are still LSEs, but the numerical procedures can get quite involved.

Step 4. Information matrix: From

$$l = -\frac{n-1}{2} \ln \sigma_w^2 - \frac{S(\phi_0, \phi_1)}{2\sigma_w^2} + const.,$$

we get

$$\dot{l} = \begin{pmatrix} \frac{\partial l}{\partial \phi_0} \\ \frac{\partial l}{\partial \phi_1} \\ \frac{\partial l}{\partial \sigma_w^2} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2\sigma_w^2}\frac{\partial S}{\partial \phi_0} \\ -\frac{1}{2\sigma_w^2}\frac{\partial S}{\partial \phi_1} \\ -\frac{n-1}{2\sigma_w^2} + \frac{S}{2\sigma_w^4} \end{pmatrix},$$

and then

$$\begin{aligned} -\ddot{l} &= \frac{\partial - \dot{l}}{\partial\left(\phi_0, \phi_1, \sigma_w^2\right)} \\ &= \begin{pmatrix} \frac{1}{2\sigma_w^2}\frac{\partial^2 S}{\partial \phi_0^2} & \frac{1}{2\sigma_w^2}\frac{\partial^2 S}{\partial \phi_1 \partial \phi_0} & -\frac{1}{2\sigma_w^4}\frac{\partial S}{\partial \phi_0} \\ * & \frac{1}{2\sigma_w^2}\frac{\partial^2 S}{\partial \phi_1^2} & -\frac{1}{2\sigma_w^4}\frac{\partial S}{\partial \phi_1} \\ * & * & -\frac{n-1}{2\sigma_w^4} + \frac{S}{\sigma_w^6} \end{pmatrix}. \end{aligned}$$

('*' indicates that the Hessian is a *symmetric* matrix - it equals its transpose.)

We must now replace the elements of the Hessian by their expectations.

Since $S = \sum_{t=2}^{n} (x_t - \phi_0 - \phi_1 x_{t-1})^2$, the terms in $-\ddot{l}$, and their expectations, are

$$\begin{aligned}
\frac{\partial S}{\partial \phi_0} &= -2 \sum_{t=2}^{n} (x_t - \phi_0 - \phi_1 x_{t-1}) \\
&= -2 \sum_{t=2}^{n} w_t, \text{ with expectation 0,} \\
\frac{\partial S}{\partial \phi_1} &= -2 \sum_{t=2}^{n} x_{t-1} (x_t - \phi_0 - \phi_1 x_{t-1}) \\
&= -2 \sum_{t=2}^{n} x_{t-1} w_t, \text{with expectation } = 0 \text{ (why?),} \\
\frac{\partial^2 S}{\partial \phi_0^2} &= 2(n-1), \text{with exp'n} = 2(n-1), \\
\frac{\partial^2 S}{\partial \phi_1 \partial \phi_0} &= 2 \sum_{t=2}^{n} x_{t-1}, \text{ with exp'n} = 2(n-1)\mu, \\
\frac{\partial^2 S}{\partial \phi_1^2} &= 2 \sum_{t=2}^{n} x_{t-1}^2, \text{ with exp'n} = 2(n-1)\left(\gamma(0) + \mu^2\right),
\end{aligned}$$

and

$$E[S] = E[\sum_{t=2}^{n} w_t^2] = (n-1)\sigma_w^2.$$

Thus

$$\begin{aligned}
& \frac{1}{n}E[-\ddot{l}] \\
= & \frac{n-1}{n}\begin{pmatrix} \frac{1}{\sigma_w^2} & \frac{\mu}{\sigma_w^2} & 0 \\ * & \frac{(\gamma(0)+\mu^2)}{\sigma_w^2} & 0 \\ * & * & -\frac{1}{2\sigma_w^4}+\frac{\sigma_w^2}{\sigma_w^6}=\frac{1}{2\sigma_w^4} \end{pmatrix} \\
\to & \frac{1}{\sigma_w^2}\begin{pmatrix} 1 & \mu & 0 \\ \mu & \gamma(0)+\mu^2 & 0 \\ 0 & 0 & \frac{1}{2\sigma_w^2} \end{pmatrix} \text{ as } n\to\infty \\
= & \frac{1}{\sigma_w^2}\begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0}' & 1/(2\sigma_w^2) \end{pmatrix} = \mathbf{I}(\boldsymbol{\alpha}_0), \text{ where } \mathbf{A} = ...
\end{aligned}$$

The inverse is

$$\begin{aligned}
\mathbf{I}^{-1}(\boldsymbol{\alpha}_0) &= \sigma_w^2\begin{pmatrix} \mathbf{A}^{-1} & \mathbf{0} \\ \mathbf{0}' & 2\sigma_w^2 \end{pmatrix}, \text{ where} \\
\mathbf{A}^{-1} &= \frac{1}{\gamma(0)}\begin{pmatrix} \gamma(0)+\mu^2 & -\mu \\ -\mu & 1 \end{pmatrix}.
\end{aligned}$$

Thus, e.g., the normal approximation for $\hat{\phi}_1$ is that

$$\hat{\phi}_1 - \phi_1 \approx N\left(0, \frac{\mathbf{I}^{22}}{n} = \frac{\sigma_w^2}{n\gamma(0)} = \frac{1-\phi_1^2}{n}\right),$$

with standard error

$$s^2(\hat{\phi}_1) = \frac{1 - \hat{\phi}_1^2}{n}$$

and $\frac{\hat{\phi}_1 - \phi_1}{s(\hat{\phi}_1)} \approx N(0,1)$. A $100(1-\alpha)\%$ confidence interval is $\hat{\phi}_1 \pm z_{\alpha/2} s(\hat{\phi}_1)$.

- In general, for an AR(p):

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + w_t$$

we minimize

$$S(\boldsymbol{\phi}) = \sum_{t=p+1}^{n} \left(x_t - \phi_0 - \sum_{i=1}^{p} \phi_i x_{t-i} \right)^2$$

by fitting a regression model

$$x_t = \phi_0 + \sum_{i=1}^{p} \phi_i x_{t-i} + error$$

for $t = p+1, ..., n$. The resulting LSEs are $\hat{\boldsymbol{\phi}}$ and the associated mean square of the residuals is

$$\hat{\sigma}_w^2 = \frac{S(\hat{\boldsymbol{\phi}})}{n - 2p - 1}.$$

The large-sample standard errors are obtained by R and appear on the printout. More on this later.

- Example 2. ARMA(p,q). Model is

$$X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = w_t + \sum_{k=1}^{q} \theta_k w_{t-k}.$$

Now make the assumption that $X^0 = w^0 = 0$ and invert the model. Equivalently, solve successively for the w_t's in terms of the X_t's:

$$\begin{aligned} w_t &= X_t - \sum_{j=1}^{p} \phi_j X_{t-j} - \sum_{k=1}^{q} \theta_k w_{t-k}; \\ w_1 &= X_1, \\ w_2 &= X_2 - \phi_1 X_1 - \theta_1 w_1, \\ &\text{etc.} \end{aligned}$$

In this way we write $\left(w_{p+1}, ..., w_n\right)$ in terms of $\left(x_{p+1}, ..., x_n\right)$. Thus $|\partial \mathbf{w}/\partial \mathbf{x}| = 1$ and so

$$\begin{aligned} f(\mathbf{x}|\boldsymbol{\alpha}) &= \left(2\pi\sigma_w^2\right)^{-(n-p)/2} e^{-\frac{S(\boldsymbol{\phi},\boldsymbol{\theta})}{2\sigma_w^2}}, \text{ with} \\ l(\boldsymbol{\alpha}) &= -\frac{n-p}{2} \ln \sigma_w^2 - \frac{S(\boldsymbol{\phi},\boldsymbol{\theta})}{2\sigma_w^2} + const. \end{aligned}$$

and $S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=p+1}^{n} w_t^2(\boldsymbol{\phi}, \boldsymbol{\theta})$. Now $S(\boldsymbol{\phi}, \boldsymbol{\theta})$ is minimized numerically to obtain the MLEs $\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}$. The adjusted MLE of σ_w^2 is, as always,

$$\hat{\sigma}_w^2 = \frac{S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})}{df} = \frac{\sum_{t=p+1}^{n} \hat{w}_t^2}{df}.$$

- The matrix

$$\mathbf{I}(\boldsymbol{\alpha}_0) = \lim_{n \to \infty} \left\{ \frac{1}{n} E\left[-\ddot{l}(\boldsymbol{\alpha}_0) \right] \right\}$$

is sometimes estimated by the "observed information matrix" $\frac{1}{n}\left(-\ddot{l}(\hat{\boldsymbol{\alpha}})\right)$ evaluated at the data $\{x_t\}$. This is numerically simpler.

- You should write out the details - as explicitly as possible - for an MA(1) model. When the calculations become too hard to do explicitly, think about how they would be programmed.

- The numerical calculations, when there is an MA component to the model, rely on a modification of least squares regression known as the **Gauss-Newton algorithm**.

- We are to minimize

$$S(\boldsymbol{\psi}) = \sum_t w_t^2(\boldsymbol{\psi}),$$

where $\boldsymbol{\psi}$ is a vector of AR and MA parameters. The idea is to approximate this by the sum of squares in a 'nearby' linear regression model, get the LSEs in this nearby model, and use these to get a closer regression model. Iterate to convergence.

– First choose an initial value $\boldsymbol{\psi}_0$. (R has a default method for this.)

– Now expand $w_t(\boldsymbol{\psi})$ around $\boldsymbol{\psi}_0$ by the Mean Value Theorem:

$$\begin{aligned} w_t(\boldsymbol{\psi}) &\approx w_t(\boldsymbol{\psi}_0) + \dot{w}_t'(\boldsymbol{\psi}_0)\,(\boldsymbol{\psi} - \boldsymbol{\psi}_0) \\ &= \text{“}y_t - \mathbf{z}_t'\boldsymbol{\beta}\text{”}, \end{aligned} \tag{11.1}$$

where $y_t = w_t(\boldsymbol{\psi}_0)$, $\mathbf{z}_t = -\dot{w}_t(\boldsymbol{\psi}_0)$, $\boldsymbol{\beta} = \boldsymbol{\psi} - \boldsymbol{\psi}_0$. Now

$$S(\boldsymbol{\psi}) \approx \sum_t \left(y_t - \mathbf{z}_t'\boldsymbol{\beta}\right)^2$$

is minimized by regressing $\{y_t\}$ on $\{\mathbf{z}_t\}$ to get the LSE

$$\hat{\boldsymbol{\beta}}_1 = \left[\sum_t \mathbf{z}_t\mathbf{z}_t'\right]^{-1} \sum_t \mathbf{z}_t y_t.$$

We now set

$$\boldsymbol{\psi}_1 = \hat{\boldsymbol{\beta}}_1 + \boldsymbol{\psi}_0,$$

expand around $\boldsymbol{\psi}_1$ (i.e. replace $\boldsymbol{\psi}_0$ by $\boldsymbol{\psi}_1$ in (11.1)), and obtain revised estimates $\hat{\boldsymbol{\beta}}_2$ and $\boldsymbol{\psi}_2$. Continue, iterating until the $\boldsymbol{\psi}$'s are no longer changing, i.e. have converged to $\hat{\boldsymbol{\psi}}$.

12. Integrated and seasonal models; example ...

- A class of nonstationary models is obtained by taking differences, and requiring the differenced series to be ARMA(p,q):

$$\begin{aligned} \nabla X_t &= X_t - X_{t-1} = (1-B)X_t, \\ \nabla^2 X_t &= \nabla(\nabla X_t) = (1-B)^2 X_t, \\ &\text{etc.} \end{aligned}$$

  We say $\{X_t\}$ is ARIMA(p,d,q) ("Integrated ARMA") if $\nabla^d X_t$ is ARMA(p,q). If so,

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

  for an AR(p) polynomial $\phi(B)$ and an MA(q) polynomial $\theta(B)$.

- Since $\phi(B)(1-B)^d$ has roots on the unit circle, $\{X_t\}$ cannot be stationary.

- It may happen that the dependence of a series on its past is strongest at multiples of the sampling unit, e.g. monthly economic data may exhibit strong quarterly or annual trends. To model this, define *seasonal* AR(P) and MA(Q) characteristic polynomials

$$\begin{aligned}\Phi(B^s) &= 1 - \Phi_1 B^s - \Phi_2 B^{2s} - ... - \Phi_P B^{Ps}, \\ \Theta(B^s) &= 1 + \Theta_1 B^s + \Theta_2 B^{2s} + ... + \Theta_Q B^{Qs}.\end{aligned}$$

A "seasonal ARMA(P,Q) model, with season s", is defined by

$$\Phi(B^s)X_t = \Theta(B^s)w_t.$$

This can be combined with the hierarchy of ordinary ARMA models, and with differencing, to give the full ARIMA(p,d,q)$\times$(P,D,Q)$_s$ model defined by

$$\Phi(B^s)\phi(B)(1-B^s)^D(1-B)^d X_t = \Theta(B^s)\theta(B)w_t.$$

– Example: the ARIMA(0,1,1)$\times$(0,1,1)$_{12}$ model has $d = D = 1$, $s = 12$ and

$$\begin{aligned} p &= 0, q = 1 \Rightarrow \phi(B) = 1, \theta(B) = 1 + \theta B; \\ P &= 0, Q = 1 \Rightarrow \Phi(B^s) = 1, \Theta(B^s) = 1 + \Theta B^{12}. \end{aligned}$$

Thus

$$(1-B^{12})(1-B)X_t = \left(1 + \Theta B^{12}\right)(1 + \theta B)\, w_t.$$

Expanding:

$$\begin{aligned} X_t &= X_{t-1} + X_{t-12} - X_{t-13} \\ &+ w_t + \theta w_{t-1} + \Theta w_{t-12} + \Theta\theta w_{t-13}. \end{aligned}$$

This model often arises with monthly economic data.

– The analysis of the ACF and PACF proceeds along the same lines as for the previous models, outlined next.

- In economics, to "seasonally adjust" a series means to fit only the seasonal part of the model, and to then study the residuals from that fit.

- Choosing an appropriate model. Some guiding properties of the ACF/PACF:

  - Nonstationarity: ACF drops off very slowly (a root of the AR characteristic equation with $|B|$ near 1 will do this too); PACF large (in absolute value) at 1 (but only at 1 could indicate AR(1)). Try taking differences $\nabla^d X_t$, $d = 1, 2$. Rarely is $d > 2$. *Don't be too hasty to take differences; try ARMA models first.*

  - Seasonal nonstationarity: ACF large at lags $s, 2s, ...$; decays slowly, or PACF very large at s. Try $\nabla_s^D X_t$ for a small D.

  - AR(p) behaviour: PACF zero for $m > p$.

  - MA(q) behaviour: ACF zero for $m > q$.

  - Seasonal AR(P): PACF zero after $m = s, 2s, ..., Ps$.

  - Seasonal MA(Q): ACF zero after $m = s, 2s, ..., Qs$.

To fit these we use the `sarima` function, contributed by the authors. If $d = D = 0$ (no differencing) the AR and MA coefficients reported are for $\dot{X}_t = X_t - \mu_X$, and the MLE of μ_X (= "xmean") is also reported:

```
sarima(log(varve), p=1, d=0, q=0)
        ar1   xmean
      0.591  3.1170
s.e.  0.032  0.0498
```

This means that the fitted model is

$$X_t - 3.117 = .591\,(X_{t-1} - 3.117) + w_t.$$

If $d + D = 1$ then a "constant" is fitted:

```
 sarima(log(varve), p=1, d=1, q=0)
           ar1  constant
      -0.3970   -0.0010
s.e.   0.0365    0.0151
```

This means that the fitted model is

$$\nabla X_t = -.0010 - .3970 \nabla X_{t-1} + w_t.$$

It has the same effect as "detrending" - estimating a linear trend in $E[X_t]$ by $\hat{\alpha} + \hat{\beta}t$ and then fitting a zero-mean ARIMA(1,1,0) model to the residuals $X_t - \left(\hat{\alpha} + \hat{\beta}t\right)$.

When $d + D > 1$ there is no detrending, but trend removal can still be done manually.

- Read the two data analyses in §3.8.

- **Principle of Parsimony**:

  We seek the simplest model that is adequate.

  We can always "improve" the fit by throwing in extra terms, but then the model might only fit these data well.

Example 3.43: U.S. Federal Reserve Board Production Index - an index of economic productivity. You should run the R code from the course website.

- Non-constant mean. Detrending would handle this:
  `resids = ts(lm(x~time(x))$resid)`).
  But ACF (of these resids) decays very slowly, and PACF spikes at $m = 1$ - both signs of nonstationarity.

- Look at the first difference ∇X_t. This will also convert a linear trend to a constant. Nonstationary variance exhibited. ACF indicates seasonal nonstationarity.

- $\nabla_{12}\nabla_1 X_t$. Spike in ACF at $m = 12$ indicates seasonal ($s = 12$) MA(1); MA(q) behaviour for small q indicated as well. PACF indicates AR(1) or AR(2) and (possibly) seasonal AR(2).

- ACF and PACF of $\nabla_{12}\nabla X_t$ shows possible models

$$\begin{aligned} ARIMA(p &= 1-2, d=1, q=1-2) \\ \times(P &= 1-2, D=1, Q=1)_{s=12}. \end{aligned}$$

Where do we go next? There are several "information criteria". All seek to minimize the residual variation while imposing penalties for nonparsimonious models. Let K be the number of AR and MA parameters fitted, and let $\hat{\sigma}_w^2(K)$ be the estimated variance of the residual noise. Each AIC is of the form "$\ln \hat{\sigma}_w^2(K)$+penalty for using K terms". In particular

$$\begin{aligned} AIC(K) &= \ln \hat{\sigma}_w^2(K) + \frac{n+2K}{n}, \\ AIC_c(K) &= \ln \hat{\sigma}_w^2(K) + \frac{n+K}{n-K-2}, \\ & \quad \text{(small sample modification)}, \\ BIC(K) &= \ln \hat{\sigma}_w^2(K) + \frac{K \ln n}{n}, \text{ etc.} \end{aligned}$$

R returns AIC values (computed somewhat differently than described here) by default. We favour the model with minimum AIC, AICc or BIC.

All three are computed by `sarima`. For instance

```
fit = sarima(prod, p=1,d=1,q=1, P=1,D=1,Q=0, S=12)
```

fits a $(1,1,1) \times (1,1,0)_{12}$ model and gives output

```
Coefficients:
         ar1      ma1     sar1
      0.6261  -0.3154  -0.4047
s.e.  0.1006   0.1214   0.0477

sigma^2 estimated as 1.787
log likelihood = -614.75

$AIC
[1] 1.596625

$AICc
[1] 1.602295

$BIC
[1] 0.6282293
```

One can fit a range of models, and check the AIC/BIC values of each as a guide:

```
out = matrix(ncol = 10)
d = 1
D = 1
S = 12
Q = 1
for (p in 1:2) { for (q in 0:2) { for (P in 1:2)  {
fits = sarima(prodn, p, d, q, P, D, Q, S)
out = rbind(out, c(p, d, q, P, D, Q, S,
   fits$AIC, fits$AICc, fits$BIC))
}}}
out = out[-1,]
colnames(out) = c("p", "d", "q", "P",
 "D", "Q", "S", "AIC", "AICc","BIC")
out
```

Note that the default is a non-seasonal model ($P = D = Q = S = 0$), so the seasonal orders P,D,Q,S can be omitted to fit ARIMA(p,d,q) models only (and `sarima` works the same way).

```
      p d q P D Q  S   AIC  AICc   BIC
 [1,] 1 1 0 1 1 1 12 1.380 1.385 0.411
 [2,] 1 1 0 2 1 1 12 1.332 1.337 0.374
 [3,] 1 1 1 1 1 1 12 1.376 1.382 0.418
 [4,] 1 1 1 2 1 1 12 1.327 1.333 0.380
 [5,] 1 1 2 1 1 1 12 1.382 1.388 0.434
 [6,] 1 1 2 2 1 1 12 1.332 1.338 0.395
 [7,] 2 1 0 1 1 1 12 1.376 1.382 0.419
 [8,] 2 1 0 2 1 1 12 1.326 1.332 0.379
 [9,] 2 1 1 1 1 1 12 1.382 1.388 0.434
[10,] 2 1 1 2 1 1 12 1.327 1.333 0.390
[11,] 2 1 2 1 1 1 12 1.383 1.389 0.446
[12,] 2 1 2 2 1 1 12 1.332 1.338 0.406
```

The program gives

```
> out[out[,8] == min(out[,8])] # min AIC
 [1]  2 1 0 2 1 1 12 1.326 1.332 0.379
> out[out[,9] == min(out[,9])] # min AICc
 [1]  2 1 0 2 1 1 12 1.326 1.332 0.379
> out[out[,10] == min(out[,10])] # min BIC
 [1]  1 1 0 2 1 1 12 1.332 1.337 0.374
```

So $(2,1,0) \times (2,1,1)_{12}$ is best according to AIC ($= 1.326$) or AICc ($= 1.332$), while $(1,1,0) \times (2,1,1)_{12}$ is best according to BIC ($= .374$). BIC often picks out a more parsimonious model. The authors chose $(1,1,1) \times (2,1,1)_{12}$ but did not look at too many others. We should look at all three more closely:

```
(i)   1,1,0, 2,1,1, 12  Chosen by BIC
(ii)  2,1,0, 2,1,1, 12  Chosen by AIC and AICc
(iii) 1,1,1, 2,1,1, 12  Chosen by S&S;
      AIC  AICc   BIC = 1.327 1.333 0.380
```

The residual plots are produced by default when `sarima` is used.

- The standardized residuals are the residuals divided by $\hat{\sigma}_w$; a plot of them against time should ideally look like a plot of white noise, i.e. no trend. Here all three plots seem satisfactory, although there are a few outliers in each case.

- The ACF of the residuals should look like those for white noise. A formal test is the **Ljung-Box test.** Under the hypothesis of whiteness we expect $|\hat{\rho}_w(m)|$ to be small for all non-zero m; a test can be based on

$$Q = n(n+2)\sum_{m=1}^{M}\frac{\hat{\rho}_w^2(m)}{n-m},$$

  which is approximately $\sim \chi^2_{M-K}$ under the null hypothesis. The p-value is calculated and plotted for a range of values of M, and compared with .05 (by default). The first model - $(1,1,0)\times(2,1,1)_{12}$ - shows significance at quite a few lags; the others are somewhat better.

- None of the values of the PACF should be significant (in principle!). All three of these PACF plots are very similar and are satisfactory.

- Of the two models which remain, $(2,1,0)\times(2,1,1)_{12}$ has slightly lower AIC, AICc and BIC values than $(1,1,1)\times(2,1,1)_{12}$, and so I favour it (but wouldn't argue against the other choice).

```
Coefficients:
         ar1     ar2     sar1     sar2     sma1
      0.2992  0.1086  -0.2186  -0.2845  -0.4915
s.e.  0.0526  0.0532   0.0784   0.0620   0.0721
sigma^2 estimated as 1.349
```

13. ... example

- **Normal scores (Q-Q) test**. The quantiles of a distribution $F = F_w$ are the values $F^{-1}(q)$, i.e. the values below which w lies with probability q. The sample versions are the ordered residuals $\hat{w}_{(1)} < \hat{w}_{(2)} < ... < \hat{w}_{(n)}$: the probability of a value w falling at or below $\hat{w}_{(t)}$ is estimated by t/n, so $\hat{w}_{(t)}$ can be viewed as the $(t/n)^{th}$ sample quantile. If the residuals are normal then a plot of the sample quantiles against the standard normal quantiles should be linear, with intercept equal to the mean and slope equal to the standard deviation (follows from $F^{-1}(q) = \mu + \sigma\Phi^{-1}(q)$ if F is the $N(\mu, \sigma^2)$ distribution function; you should verify these statements). The strength of the linearity is measured by the correlation between the two sets of quantiles; values too far below 1 lead to rejection of the hypothesis of normality of the white noise. This is the basis of the "Shapiro-Wilk" test.

```
hist(stdres)
qqnorm(stdres)
shapiro.test(stdres)
```

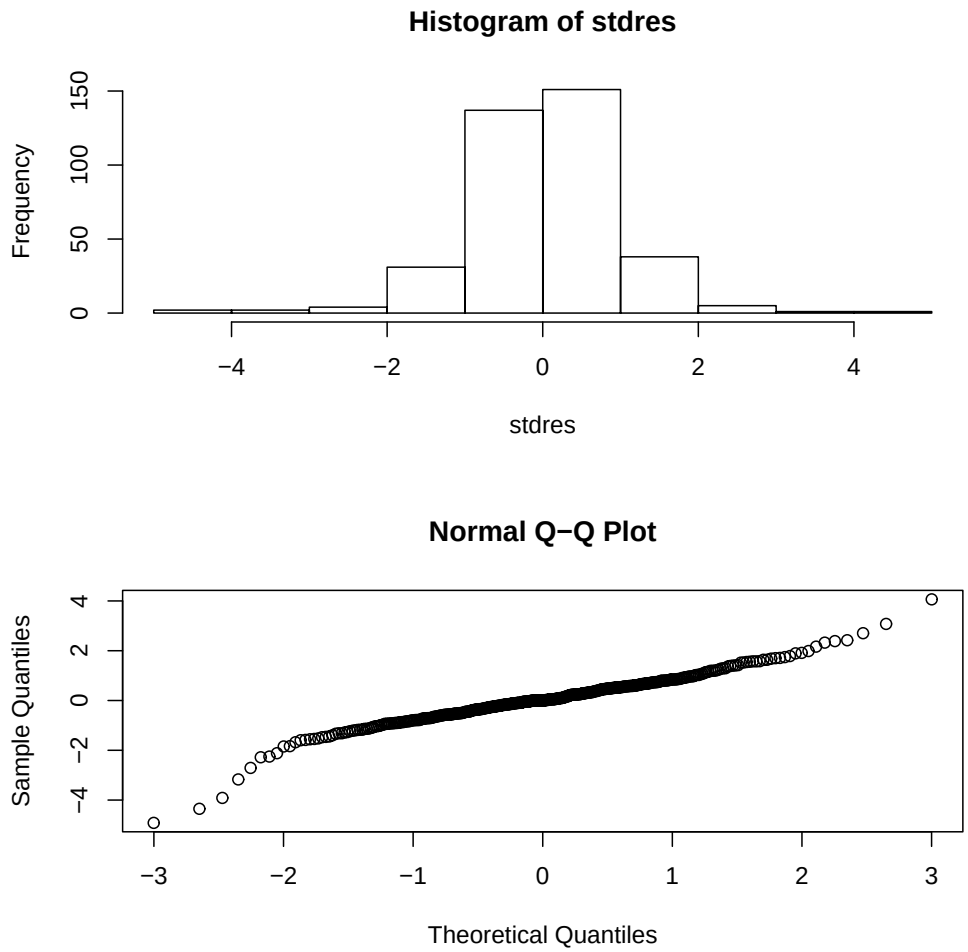

Figure 13.1. Histogram and qq-plot of residuals.

```
        Shapiro-Wilk normality test

data:  stdres
W = 0.956, p-value = 4.168e-09
```

- The residuals appear quite non-normal, primarily because the large, negative residuals are more extreme than the normal law would predict. This is often just something that one must live with. Methods of addressing the problem have been developed, but will not be treated here.

- The residuals from the other two models we looked at were no better with respect to normality.

- A log or square root transformation of the original data sometimes improves normality; logging the data made no difference in this case.

- In contrast, if we intentionally fit inappropriate models we can verify that our errors will show up in the plots. In this example we see the results of replacing the chosen model $(2,1,0) \times (2,1,1)_{12}$ with (i) $(0,1,0) \times (2,1,1)_{12}$ (failing to fit the ordinary AR component; note the PACF and Ljung-Box plot), and (ii) $(2,1,0) \times (0,1,0)_{12}$ (failing to fit the seasonal components).

Forecasting: Fitted model is $(2,1,0)\times(2,1,1)_{12}$:

$$\begin{aligned}&\left(1-\phi_1 B-\phi_2 B^2\right)\left(1-\Phi_1 B^{12}-\Phi_2 B^{24}\right)\nabla\nabla_{12}X_t\\ =\;&\left(1+\Theta B^{12}\right)w_t. \qquad (13.1)\end{aligned}$$

The polynomial on the lhs is expanded as

$$\begin{aligned}&\left(1-\phi_1 B-\phi_2 B^2\right)\left(1-\Phi_1 B^{12}-\Phi_2 B^{24}\right)\cdot\\ &(1-B)\left(1-B^{12}\right)\\ =\;&\alpha(B)=1-\alpha_1 B-\ldots-\alpha_{39}B^{39}, \text{ say.}\end{aligned}$$

This gives the expanded model

$$X_t=\alpha_1 X_{t-1}+\ldots+\alpha_{39}X_{t-39}+w_t+\Theta w_{t-12}. \qquad (13.2)$$

Conditioning on X^t gives

$$\begin{aligned}X^t_{t+l} \;=\;&\alpha_1 X^t_{t+l-1}+\alpha_2 X^t_{t+l-2}+\ldots+\alpha_{39}X^t_{t+l-39}\\ &+w^t_{t+l}+\Theta w^t_{t+l-12}.\end{aligned}$$

We consider the case $l\le 12$ only; $l=13$ left as an exercise.

Recall $X^t_s=X_s$ if $s\le t$. Thus $X^t_{t+l-k}=X_{t+l-k}$ for (at least) $k=12,\ldots,39$ (since then $t+l-k\le t$).

Note that the model is not stationary but seems to be invertible (since $|\hat{\Theta}| = .4915 < 1$). Thus w^t can be expressed in terms of X^t and, if we assume $w^0 = X^0 = 0$, we can express X^t in terms of w^t by iterating (13.2). Then conditioning on X^t is equivalent to conditioning on w^t, hence (as usual)

$$w_{t+l}^t = E\left[w_{t+l}|X^t\right] = E\left[w_{t+l}|w^t\right].$$

As a consequence $w_{t+l-12}^t = w_{t+l-12}$ for $l \leq 12$:

$$X_{t+l}^t = \alpha_1 X_{t+l-1}^t + \alpha_2 X_{t+l-2}^t + ... + \alpha_{39} X_{t+l-39}^t + \Theta w_{t+l-12}^t.$$

These become

$$\begin{aligned}
X_{t+1}^t &= \alpha_1 X_t + \alpha_2 X_{t-1} + ... + \alpha_{39} X_{t-38} + \Theta w_{t-11}. \\
X_{t+2}^t &= \alpha_1 X_{t+1}^t + \alpha_2 X_t + ... + \alpha_{39} X_{t-37} + \Theta w_{t-10}. \\
&\ldots \\
X_{t+12}^t &= \alpha_1 X_{t+11}^t + \alpha_2 X_{t+10}^t + ... + \alpha_{11} X_{t+1}^t \\
&\quad + \alpha_{12} X_t + \alpha_{13} X_{t-1} + ... + \alpha_{39} X_{t-27} + \Theta w_t.
\end{aligned}$$

At each stage, those X_{t+k}^t which are needed have already been obtained at an earlier stage.

To get $w_t, .., w_{t-11}$ we write (13.2) as

$$\begin{aligned} w_t &= -\Theta w_{t-12} + f_t, \text{ where} \\ f_t &= X_t - \{\alpha_1 X_{t-1} + ... + \alpha_{39} X_{t-39}\}, \end{aligned}$$

and calculate successively, using the assumption $w^0 = X^0 = 0$,

$$\begin{aligned} w_1 &= f_1, w_2 = f_2, , ..., w_{12} = f_{12}, \\ w_{13} &= -\Theta w_1 + f_{13},\ w_{14} = -\Theta w_2 + f_{14},\ \text{etc.} \end{aligned}$$

To get the residuals $\hat{w}_t$ replace parameters by estimates in the preceding.

The forecast variances and prediction intervals require us to write $X_t = \psi(B) w_t$, and then

$$PI = \hat{X}^t_{t+l} \pm z_{\alpha/2} \hat{\sigma}_w \sqrt{\sum_{0 \le k < l} \hat{\psi}^2_k}.$$

The model is not stationary and so the coefficients of $\psi(B)$ will not be absolutely summable, however under the assumption that $w^0 = 0$, only finitely many of the ψ_k are needed.

Then from (13.1), i.e. from $\alpha(B)X_t = \left(1 + \Theta B^{12}\right) w_t$, together with $X_t = \psi(B)w_t$, we get

$$\alpha(B)\psi(B) = \left(1 + \Theta B^{12}\right).$$

In expanded form,

$$\begin{aligned} & \left\{ \ 1 - \alpha_1 B - ... - \alpha_{39} B^{39} \ \right\} \cdot \\ & (1 + \psi_1 B + \psi_2 B^2 + ... + \psi_k B^k + ...) \\ = \ & \left(1 + \Theta B^{12}\right). \end{aligned}$$

For $1 \leq k < 12$ the coefficient of B^k is $= 0$ on the rhs; on the lhs it is

$$\begin{aligned} k &= 1 : \psi_1 - \alpha_1, \\ k &= 2 : \psi_2 - \alpha_1\psi_1 - \alpha_2, \\ & \ldots \end{aligned}$$

In general the coefficient is $\psi_k - \sum_{j=0}^{k-1} \psi_j \alpha_{k-j}$, with $\psi_0 = 1$. Thus

$$\begin{aligned} \psi_0 &= 1, \ \psi_1 = \alpha_1, \\ \psi_k &= \alpha_k + \sum_{j=1}^{k-1} \psi_j \alpha_{k-j} \ (k = 2, 3, ..., 11). \end{aligned}$$

To get the forecasts in R, we use `sarima.for`:

```
prod.pr = sarima.for(prod, n.ahead = 12,
     2,1,0, 2,1,1, 12)
abline(v=1979-.05)
```

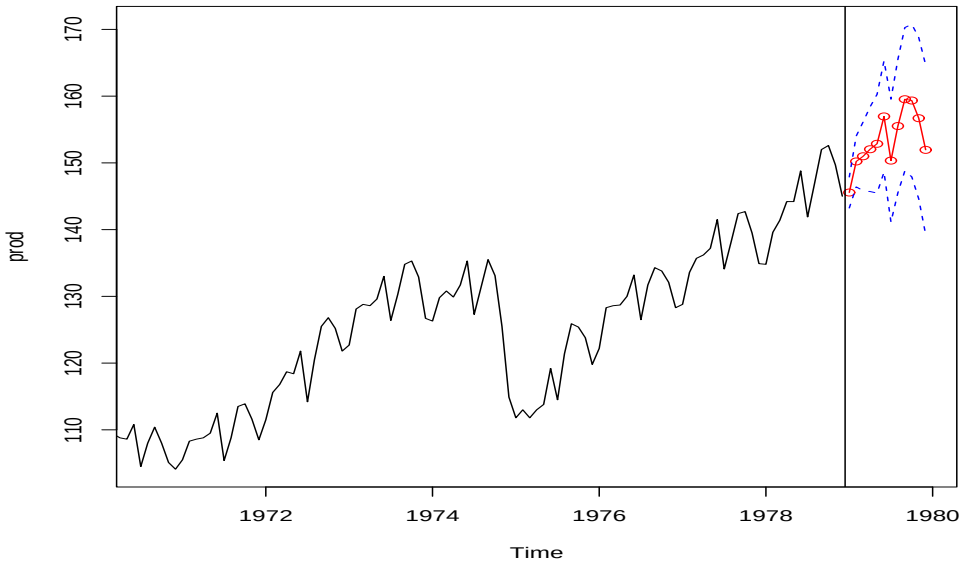

Figure 13.4. Twelve month ahead forecasts from frb series. Broken lines are forecasts ± 2 s.e.

Part III

Frequency Domain Analysis

14. Periodicity; Power spectrum

Recall from the first assignment that the function

$$\rho(m) = \cos(2\pi\nu m) \ (-1/2 < \nu < 1/2)$$

is the ACF of the stationary process

$$X_t = A\cos(2\pi\nu t) + B\sin(2\pi\nu t),$$

where A and B are uncorrelated r.v.s with means $= 0$ and equal variances. We will see that any zero-mean, weakly stationary process can be obtained through generalizing this example. Essentially we will "mix" examples of this type by assuming that ν is itself a r.v., and taking an expectation over it.

- See Figure 14.1. Sunspot data - the periodic behaviour seen in the plot of the series is highlighted in the periodic nature of its sample ACF.

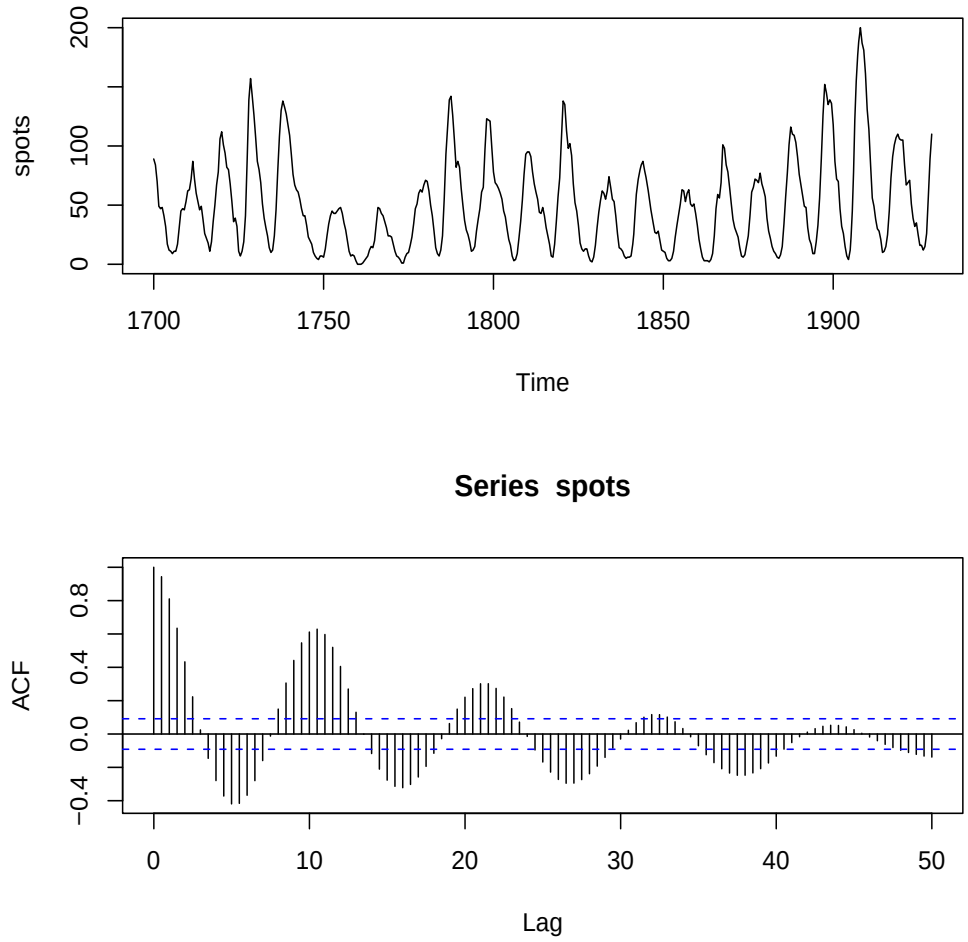

Figure 14.1. Biannual record of sunspot counts, 1700-1950 approximately, with ACF.

- Facts from Fourier analysis.

  - Given any set $x_1, ..., x_n$ of real numbers, we can write (i.e. solve for $\{a_k, b_k)\}$)

    $$x_t = \sum_{k=0}^{N} \{a_k \cos(2\pi\nu_k t) + b_k \sin(2\pi\nu_k t)\},$$

    where the "amplitudes" a_k, b_k are appropriately chosen coefficients, $N = [n/2]$ (integer part) and $\nu_k = k/n$ (thus $0 \leq \nu_k \leq 1/2$).

  - The ν_k are called "frequencies": $\cos(2\pi\nu_k t)$ and $\sin(2\pi\nu_k t)$ are periodic, with frequency ν_k and "period" $1/\nu_k$. (Think of a plot of $\sin(2\pi\nu_k t)$ vs. t - it repeats every $1/\nu_k$ units, hence ν_k times per time period.) We say the frequency is ν_k "cycles per unit time"; the period $1/\nu_k$ is the time required to complete one cycle .

  - We often write $\lambda_k = 2\pi\nu_k \in [0, \pi]$; the "Fourier frequency" λ_k is measured in "radians per unit time".

- More generally, we can consider random variables $\{X_t\}_{t=-\infty}^{\infty}$ defined by

$$X_t = \sum_{k=0}^{N} \{A_k \cos(\lambda_k t) + B_k \sin(\lambda_k t)\}$$

for Fourier frequencies λ_k, where $A_0, ..., A_N, B_0, ..., B_N$ are uncorrelated r.v.s with

$$\begin{aligned} E[A_k] &= E[B_k] = 0, \\ VAR[A_k] &= VAR[B_k] = \sigma_k^2. \end{aligned}$$

Theorem: Any zero-mean, weakly stationary time series may be approximated arbitrarily closely, in this manner.

- We won't attempt to prove the theorem, but will verify that such a series $\{X_t\}$ is weakly stationary. Clearly the mean is 0. We will show that the autocovariance function is

$$\gamma(m) = \sum_{k=0}^{N} \sigma_k^2 \cos(\lambda_k m).$$

To see this write

$$X_t = \sum_{k=0}^{N} X_t(k),$$

where $X_t(k) = A_k \cos(\lambda_k t) + B_k \sin(\lambda_k t)$, and so $X_t(k), X_s(l)$ are uncorrelated if $k \neq l$. Then (using assignment 1),

$$COV\left[X_{t+m}(k), X_t(l)\right] = \begin{cases} 0, & k \neq l, \\ \sigma_k^2 \cos(\lambda_k m), & k = l. \end{cases} \quad (14.1)$$

We then obtain

$$\begin{aligned} COV[X_{t+m}, X_t] &= COV\left[\sum_{k=0}^{N} X_{t+m}(k), \sum_{l=0}^{N} X_t(l)\right] \\ &= \sum_{k=0}^{N}\sum_{l=0}^{N} COV\left[X_{t+m}(k), X_t(l)\right] \\ &= \sum_{k=0}^{N} \sigma_k^2 \cos(\lambda_k m) \text{ (by (14.1))}. \end{aligned}$$

Note that $\gamma(m)$ is a linear combination of periodic functions with the same frequencies as the $\{X_t\}$, suggesting that the series can be studied by studying its ACF.

- To generalize this further, write the variance of X_t as $\sigma^2 = \gamma(0) = \sum_{k=0}^{N} \sigma_k^2$, so that the autocorrelation is

$$\rho(m) = \frac{\gamma(m)}{\gamma(0)} = \sum_{k=0}^{N} \frac{\sigma_k^2}{\sigma^2} \cos(\lambda_k m).$$

This can be written as

$$\rho(m) = E\left[\cos(\Lambda m)\right]$$

for some random variable $\Lambda \in [-\pi, \pi]$ *taking values* $\pm\lambda_k$ *(*$k = 0, ..., N$*) with probabilities*

$$P(\Lambda = \lambda_k) = P(\Lambda = -\lambda_k) \stackrel{def}{=} p_k = \begin{cases} \frac{\sigma_0^2}{\sigma^2}, & k = 0, \\ \frac{\sigma_k^2}{2\sigma^2}, & k \neq 0. \end{cases}$$

Reason: First note that the sum of the probabilities is

$$P\left(\Lambda = \lambda_0\right) + \sum_{k=1}^{N} P\left(\Lambda = \lambda_k\right) + \sum_{k=1}^{N} P\left(\Lambda = -\lambda_k\right)$$

$$\begin{aligned} &= \frac{\sigma_0^2}{\sigma^2} + 2\sum_{k=1}^{N} \frac{\sigma_k^2}{2\sigma^2} \\ &= \frac{\sum_{k=0}^{N} \sigma_k^2}{\sigma^2} \\ &= 1, \end{aligned}$$

so this really is a probability distribution. Next, if the r.v. Λ has this prob. dist'n, then (using $\cos(-x) = \cos(x)$)

$$\begin{aligned} E[\cos(\Lambda m)] &= p_0 \cos(\lambda_0 m) + \sum_{k=1}^{N} p_k \cos(\lambda_k m) \\ &\quad + \sum_{k=1}^{N} p_k \cos(-\lambda_k m) \\ &= \frac{\sigma_0^2}{\sigma^2} \cos(\lambda_0 m) + 2 \sum_{k=1}^{N} \frac{\sigma_k^2}{2\sigma^2} \cos(\lambda_k m) \\ &= \sum_{k=0}^{N} \frac{\sigma_k^2}{\sigma^2} \cos(\lambda_k m) \\ &= \rho(m). \end{aligned}$$

- We say that Λ is *symmetrically* distributed: it is distributed in the same manner as $-\Lambda$.

- What would one expect to happen as $N \to \infty$?

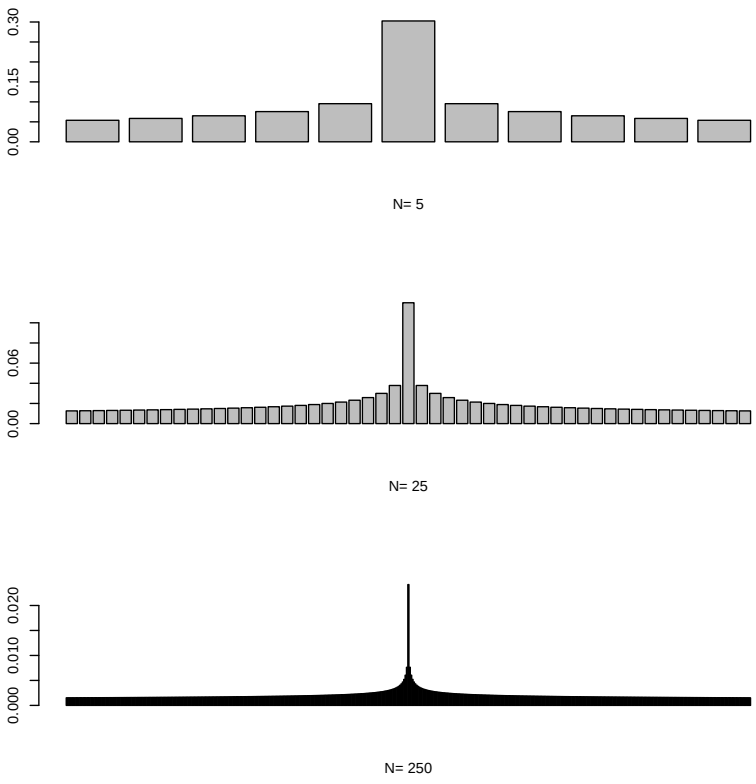

Figure 14.2. Probability distributions $\{p_k\}$ for various values of N; $\sigma_k^2 = 1/\log(2+k)$.

- This suggests studying ACFs of the form

$$\gamma(m) = \sigma^2 E\left[\cos(\Lambda m)\right] = \sigma^2 E\left[\cos(2\pi V m)\right]$$

where the expectation is with respect to a probability distribution which, while it need not be discrete, is still symmetric. We will in fact assume that

$$V = \Lambda/\left(2\pi\right) \in [-1/2, 1/2]$$

has a probability density of the form $f(\nu)/\sigma^2$, so that

$$\gamma(m) = \int_{-1/2}^{1/2} \cos(2\pi\nu m) f(\nu) d\nu.$$

The function $f(\nu)$ is non-negative and symmetric:

$$f(\nu) = f(-\nu).$$

It is called the "spectral density" or "power spectrum". Since V is symmetrically distributed , we have (how?)

$$\begin{aligned} E\left[\sin(2\pi V m)\right] &= E\left[\sin(-2\pi V m)\right] \\ &= -E\left[\sin(2\pi V m)\right] \\ &= 0. \end{aligned}$$

14.1. Notes re projects.

- Some comments on the projects are on the original handout ("course information" on the website). In particular, I'd like to see a discussion of the scientific, or economic, etc. issues around your data and analysis.

- See me right away if you haven't yet gotten started.

- There is a tendency to overdifference. If all you notice in the plot of the data is a (linear) trend in the mean, then this alone is not a reason to difference. Try detrending the series first (to detrend a time series x regress it on its own times: `linear.fit = lm(x~time(x))`; then `linear.fit$resid` is the detrended series, `linear.fit$coef` the coefficients, etc.) and then look at the acf and pacf to see if there are real signs of nonstationarity. (The presence of a few significant autocorrelations among many is perhaps also not a reason to difference.)

- There is also a tendency to rely too much on AIC (or AICc or BIC) to give you the "best" model. It should be used only as a guide. It is a random procedure, and so like any other is prone to error. Even if there were one "right" model, AIC might not find it. But it is a useful starting point. Take the model you get from AIC; if it is a good one (on the basis of the residuals) then fine. Otherwise look at some nearby models to see if they give the required improvements.

- You might have to get data from an Excel file, or some other external source, into R. The basic command to do this is "read.table", and `help(read.table)` will give you lots of details on it. Here is a simple example. Suppose the data are in a certain file, arranged in columns. Open this file, highlight the contents of it (or maybe just the rows and columns that you're interested in), and do CTRL-C to copy these contents to the clipboard. (Just as you would do, if you wanted

to copy these contents to somewhere else.) Then, in R (which you should log into first), enter a command like "`data = read.table("clipboard")`". Now the object "data" should contain your data. It is possible to do things which are much more sophisticated, so as to retain the names of the rows and columns, etc. For this you'll have to consult the help files in R.

- When writing up your projects, include the relevant plots and your arguments for having to take a difference, if you do indeed take one.

- Please also include a copy of your final R script as an appendix to your report.

- Some examples of projects are on the course website.

15. Spectral Representation Theorem

- Review: a zero-mean, weakly stationary series can be approximated as

$$X_t \approx \sum_{k=0}^{N} \{A_k \cos(\lambda_k t) + B_k \sin(\lambda_k t)\}$$

for uncorrelated r.v.s $\{A_k, B_k\}$ and Fourier frequencies $\lambda_k \in [0, \pi]$, with the approximation becoming exact as $N \to \infty$. For a series as on the rhs above, the ACF is

$$\rho(m) = E\left[\cos(\Lambda m)\right]$$

for some discrete, symmetric r.v. Λ with possible values $\lambda_0 = 0, \pm\lambda_1, ..., \pm\lambda_N$. As $N \to \infty$, the distribution of Λ tends to one with a density. Now we write $\lambda = 2\pi\nu$, $-1/2 \leq \nu \leq 1/2$, and write the density of ν in the form $f(\nu)/\sigma_X^2$. The relationship above becomes

$$\frac{\gamma(m)}{\sigma_X^2} = \rho(m) = \int_{-1/2}^{1/2} \cos(2\pi\nu m) \frac{f(\nu)}{\sigma_X^2} d\nu;$$

thus

$$\begin{aligned} \gamma(m) &= \int_{-1/2}^{1/2} \cos(2\pi\nu m) f(\nu) d\nu, \\ 0 &= \int_{-1/2}^{1/2} \sin(2\pi\nu m) f(\nu) d\nu. \end{aligned}$$

- For reasons which will become clear, it is simpler to view the two integrals above as the real and imaginary parts of a certain complex integral. Put $i = \sqrt{-1}$, and note ("Euler's relation"):

$$e^{ix} = \cos x + i \sin x.$$

This is the only feature of complex "analysis" to be used; everything else follows from it: e.g.

  - $|e^{ix}| = \sqrt{\cos^2 x + \sin^2 x} = 1$
  - For k an integer, $e^{i(x+2\pi k)} = e^{ix} e^{i2\pi k}$ $= e^{ix} (\cos 2\pi k + i \sin 2\pi k) = e^{ix}$
  - $e^{ix} + e^{-ix} = [\cos x + i \sin x] + [\cos(-x) + i \sin(-x)]$ $= 2 \cos x$

- See the *Primer on Complex Numbers* on the course website.

Now

$$\begin{aligned} & \int_{-1/2}^{1/2} e^{2\pi i\nu m} f(\nu) d\nu \\ = & \int_{-1/2}^{1/2} \cos(2\pi\nu m) f(\nu) d\nu + i \int_{-1/2}^{1/2} \sin(2\pi\nu m) f(\nu) d\nu \\ = & \gamma(m) + i \cdot 0 \\ = & \gamma(m). \end{aligned}$$

- **Spectral Representation Theorem**: Suppose that $\gamma(m)$ is the ACF of a weakly stationary series with mean 0 and variance σ_X^2. If as well $\sum_{m=-\infty}^{\infty} |\gamma(m)| < \infty$, there exists a symmetric spectral density $f(\nu)$ such that

$$\gamma(m) = \int_{-1/2}^{1/2} e^{2\pi i\nu m} f(\nu) d\nu. \tag{15.1}$$

The "total power" is

$$\sigma_X^2 = \gamma(0) = \int_{-1/2}^{1/2} f(\nu) d\nu.$$

- We can interpret $f(\nu)$ as a decomposition of the total "power" ($=$ variance), in that $f(\nu)d\nu$ is the portion of the variance attributable to the frequencies near ν. This is akin to Analysis of Variance in classical statistics. A peak in f at ν indicates important contributions to the variance at this frequency. A common application is to alter ("filter") a series, so as to accentuate certain frequencies and dampen others.

- The power f is uniquely determined by γ. In fact (15.1) is an integral transform with an "inverse":

$$f(\nu) = \sum_{m=-\infty}^{\infty} e^{-2\pi i\nu m}\gamma(m). \tag{15.2}$$

  We say that f and γ are "Fourier transform pairs"; we shall refer to f as the Infinite Fourier Transform of γ. Note from (15.2) (and $e^{ix} = e^{i(x+2\pi m)}$) that $f(\nu) = f(\nu+1) = f(\nu+2) = \cdots$; f is periodic with period 1. By symmetry, f is completely determined by its behaviour on $[0, 1/2]$ and so is generally only plotted on this interval.

- Example: White noise.

$$\begin{aligned} f(\nu) &= \sum_{m=-\infty}^{\infty} e^{-2\pi i \nu m} \gamma(m) \\ &= \sum_{m=-\infty}^{\infty} e^{-2\pi i \nu m} \left[\sigma_w^2 I(m=0)\right] \\ &= \sigma_w^2. \end{aligned}$$

White noise has a "flat spectrum".

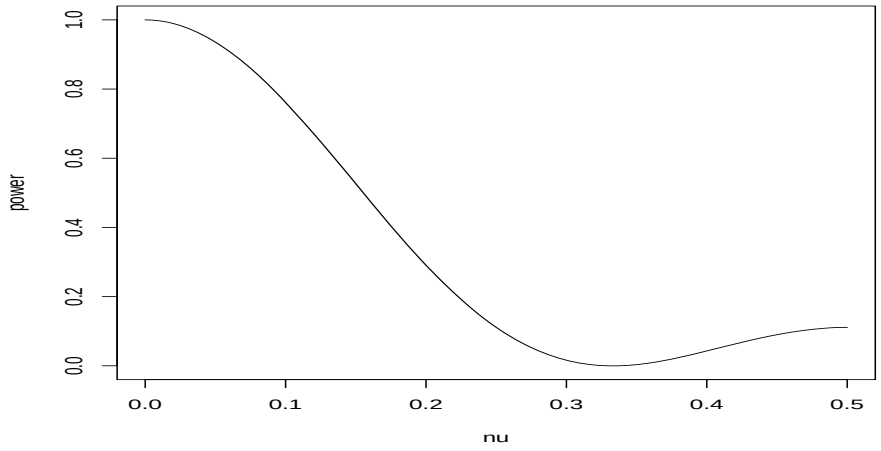

Figure 15.1. Power spectrum of 3-point centred moving average; $\sigma_w^2 = 1$. Frequencies above about $\nu = 1/3$ are "filtered out".

- Example: $X_t = (w_{t-1} + w_t + w_{t+1})/3$, a centred 3-point moving average. **In general**,

$$\begin{aligned} f(\nu) &= \sum_{m=-\infty}^{\infty} e^{-2\pi i \nu m} \gamma(m) \\ &= \gamma(0) + \sum_{m=1}^{\infty} \left(e^{-2\pi i \nu m} + e^{2\pi i \nu m} \right) \gamma(m) \\ &= \gamma(0) + 2 \sum_{m=1}^{\infty} \cos(2\pi \nu m) \cdot \gamma(m). \end{aligned}$$

The rearrangement in the second line above is made possible by the absolute summability of the $\gamma(m)$:

$$\sum_{m=-\infty}^{\infty} \left| e^{-2\pi i \nu m} \gamma(m) \right| = \sum_{m=-\infty}^{\infty} |\gamma(m)| < \infty.$$

Now substitute

$$\gamma(m) = \sigma_w^2 \cdot \begin{cases} 3/9, & m = 0, \\ 2/9, & m = \pm 1, \\ 1/9, & m = \pm 2, \\ 0, & |m| > 2, \end{cases}$$

to obtain

$$f(\nu) = \frac{\sigma_w^2}{9} \left[3 + 4\cos(2\pi\nu) + 2\cos(4\pi\nu) \right].$$

See Figure 15.1: frequencies above about $\nu = 1/3$ (periods < 3) are "filtered out". Recall "births" series (p. 19), and the manner in which a 12 point moving average, applied to monthly data, removed trends with periods < 1 year.

- Here is how the power of the 3-point moving average, and "filtering out periods < 3" are related. Using "$\cos(2x) = 2\cos^2(x) - 1$", the power is

$$f_X(\nu) = \left\{\frac{1}{3}(1 + 2\cos(2\pi\nu))\right\}^2 \sigma^2 = |A(\nu)|^2 f_w(\nu),$$

$$\text{for } A(\nu) = \frac{1}{3}(1 + 2\cos(2\pi\nu)) = \sum_{t=-\infty}^{\infty} a_t e^{-2\pi i\nu t},$$

$$\text{where } a_0 = a_1 = a_{-1} = \frac{1}{3}, \text{ all others } = 0. \qquad (*)$$

- This illustrates the fact that if $X_t = \sum_{s=-\infty}^{\infty} a_s Z_{t-s}$ then $f_X(\nu) = |A(\nu)|^2 f_Z(\nu)$. For the 3-point MA, $\{Z_t\} = \{w_t\}$, (*) holds and $|A(\nu)|^2$ is small for $\nu > 1/3$. Then periodic trends in $\{Z_t\}$ of period < 3 would be removed in the X_t - series.

- Reasoning in reverse, if we decided on the shape of $A(\nu)$, could we calculate $\{a_t\}$ and then do the filtering, to attain a desired aim? This will be the topic of Lecture 20.

- Example: Voice recognition – is Elvis Presley still alive? $\{X_t\}$ a series from a known recording of Elvis; $\{Y_t\}$ from a recent recording alleged to be of Elvis. Obtain $f_X(\nu), f_Y(\nu)$, filter each to accentuate a narrow band $[\nu_0, \nu_1]$. This gives "cleaned" series $\{\tilde{X}_t\}, \{\tilde{Y}_t\}$ with these frequencies highlighted. Compare them; repeat for another frequency band.

- Alternatively we might look for a frequency analogue of the CCF and use it to see if the two series are sufficiently similar that we can conclude they come from the same person. This analogue is the "coherence", to be studied in the next class.

- The relationship between $\gamma(m)$ and $f(\nu)$ in (15.1) and (15.2) holds in more general settings. If $\{a_t\}_{t=-\infty}^{\infty}$ has $\sum_{t=-\infty}^{\infty} |a_t| < \infty$, then

$$A(\nu) = \sum_{t=-\infty}^{\infty} a_t e^{-2\pi i \nu t} \Leftrightarrow a_t = \int_{-1/2}^{1/2} e^{2\pi i \nu t} A(\nu) d\nu.$$

In particular, this implies a uniqueness property. Suppose two spectra $f(\nu)$, $g(\nu)$ have

$$\gamma(m) = \int_{-1/2}^{1/2} e^{2\pi i \nu m} f(\nu) d\nu = \int_{-1/2}^{1/2} e^{2\pi i \nu m} g(\nu) d\nu$$

for all $m = 0, \pm 1, \pm 2, \ldots$. Then

$$f(\nu) = \sum_{m=-\infty}^{\infty} e^{-2\pi i \nu m} \gamma(m) = g(\nu),$$

for all ν.

In other words, *the spectrum is uniquely determined by the ACF*. Conversely, *the ACF is uniquely determined by the spectrum*.

16. Cross-spectrum; filters

- Frequency methods can be used to study relationships between jointly stationary series. Analogous to the case of a single series, if $\{X_t\}$ and $\{Y_t\}$ are jointly stationary, then we can represent their cross-covariance function as

$$\gamma_{XY}(m) = \int_{-1/2}^{1/2} e^{2\pi i \nu m} f_{XY}(\nu) d\nu$$

for a "cross-spectrum" $f_{XY}(\nu)$ satisfying

$$f_{XY}(\nu) = \sum_{m=-\infty}^{\infty} e^{-2\pi i \nu m} \gamma_{XY}(m),$$

provided $\sum_{m=-\infty}^{\infty} |\gamma_{XY}(m)| < \infty$.

- The identity

$$f_X(\nu) = \gamma_X(0) + 2 \sum_{m=1}^{\infty} \cos(2\pi\nu m)\, \gamma_X(m)$$

ensures that $f_X(\nu)$ is real. The same is not the case for the cross-spectrum: it has an imaginary

part. We define the *co-spectrum* $c_{XY}(\nu)$ and *quad-spectrum* $q_{XY}(\nu)$ by $f_{XY}(\nu) = c_{XY}(\nu) - iq_{XY}(\nu)$:

$$\begin{aligned} c_{XY}(\nu) &= \sum_{m=-\infty}^{\infty} \cos(2\pi\nu m)\,\gamma_{XY}(m), \\ q_{XY}(\nu) &= \sum_{m=-\infty}^{\infty} \sin(2\pi\nu m)\,\gamma_{XY}(m). \end{aligned}$$

Note (since $\gamma_{YX}(m) = \gamma_{XY}(-m)$)

$$\begin{aligned} f_{YX}(\nu) &= \sum_{m=-\infty}^{\infty} e^{-2\pi i\nu m}\gamma_{YX}(m) \\ &= \sum_{m=-\infty}^{\infty} e^{-2\pi i\nu m}\gamma_{XY}(-m) \\ &= \sum_{m=-\infty}^{\infty} e^{2\pi i\nu m}\gamma_{XY}(m) \\ &= \bar{f}_{XY}(\nu), \end{aligned}$$

so that

$$f_{YX}(\nu)f_{XY}(\nu) = |f_{XY}(\nu)|^2 = |f_{YX}(\nu)|^2.$$

- Define the "squared coherence" function by

$$\rho^2_{YX}(\nu) = \frac{|f_{YX}(\nu)|^2}{f_Y(\nu) f_X(\nu)}.$$

This looks and behaves like a squared cross-correlation:

$$\rho^2_{YX}(m) = \frac{\gamma^2_{YX}(m)}{\gamma_X(0)\gamma_Y(0)};$$

it is $\in [0, 1]$, with the 0 attained (at all frequencies) if $\gamma_{YX}(m) = 0$ for all m, and 1 attained (at all frequencies) if

$$Y_t = \sum_{s=-\infty}^{\infty} a_s X_{t-s}$$

for constants $\{a_s\}_{s=-\infty}^{\infty}$ (such that $\sum_{s=-\infty}^{\infty} |a_s| < \infty$). In this latter case we say $\{Y_t\}$ is a <u>linear filter</u> of $\{X_t\}$.

- **Theorem**: If $\{Y_t\}$ is a linear filter of $\{X_t\}$, with filter coefficients $\{a_s\}_{s=-\infty}^{\infty}$ satisfying $\sum_{s=-\infty}^{\infty} |a_s| < \infty$ then:
(i) $f_Y(\nu) = |A(\nu)|^2 f_X(\nu)$,
where $A(\nu) = \sum_{s=-\infty}^{\infty} a_s e^{-2\pi i \nu s}$ is the IFT;

(ii) $f_{YX}(\nu) = f_X(\nu)A(\nu)$;
(iii) $\rho^2_{YX}(\nu) = 1$.

Proof: I will prove (i); you do (ii) and then (iii) is immediate. For (i), first obtain

$$\begin{aligned}
\gamma_Y(m) &= COV[Y_{t+m}, Y_t] \\
&= COV\left[\sum_{s=-\infty}^{\infty} a_s X_{t+m-s}, \sum_{r=-\infty}^{\infty} a_r X_{t-r}\right] \\
&= \sum_{s=-\infty}^{\infty}\sum_{r=-\infty}^{\infty} a_s a_r COV\left[X_{t+m-s}, X_{t-r}\right] \\
&= \sum_{s=-\infty}^{\infty}\sum_{r=-\infty}^{\infty} a_s a_r \gamma_X(m-s+r).
\end{aligned}$$

One way to proceed is to substitute this into

$$f_Y(\nu) = \sum_{m=-\infty}^{\infty} e^{-2\pi i\nu m}\gamma_Y(m)$$

and grind through the algebra. Not difficult (try it!), but the following is easier and illustrates a neat trick. We know that

$$\gamma_X(m-s+r) = \int_{-1/2}^{1/2} e^{2\pi i\nu(m-s+r)} f_X(\nu)d\nu,$$

hence in the above

$$\begin{aligned}
& \gamma_Y(m) = \\
& \sum_{s=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} a_s a_r \int_{-1/2}^{1/2} e^{2\pi i\nu(m-s+r)} f_X(\nu) d\nu \\
= & \int_{-1/2}^{1/2} f_X(\nu) e^{2\pi i\nu m} \cdot \\
& \left\{ \sum_{s=-\infty}^{\infty} a_s e^{-2\pi i\nu s} \cdot \sum_{r=-\infty}^{\infty} a_r e^{2\pi i\nu r} \right\} d\nu \\
= & \int_{-1/2}^{1/2} f_X(\nu) e^{2\pi i\nu m} A(\nu)\bar{A}(\nu) d\nu \\
= & \int_{-1/2}^{1/2} f_X(\nu) e^{2\pi i\nu m} |A(\nu)|^2 d\nu.
\end{aligned}$$

But also

$$\gamma_Y(m) = \int_{-1/2}^{1/2} f_Y(\nu) e^{2\pi i\nu m} d\nu.$$

Since these Fourier transforms agree, the functions being transformed are equal:

$$f_X(\nu)|A(\nu)|^2 = f_Y(\nu).$$

- More generally, something that you can (and should) easily show is that if $Y_t = \sum_{s=-\infty}^{\infty} a_s X_{t-s} + w_t$, for white noise uncorrelated with $\{X_t\}$, then

$$\rho^2_{YX}(\nu) = 1 \Big/ \left(1 + \frac{\sigma_w^2}{f_X(\nu)|A(\nu)|^2}\right).$$

- As one application we obtain the spectrum of an AR(p) process. Suppose $\{X_t\}$ is AR(p), so that

$$w_t = X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = \phi(B) X_t.$$

Thus $w_t = \sum_{s=-\infty}^{\infty} a_s X_{t-s}$ with $a_0 = 1, a_1 = -\phi_1, ..., a_p = -\phi_p$ and all other $a_s = 0$. The IFT is

$$\begin{aligned} A(\nu) &= \sum_{s=-\infty}^{\infty} a_s e^{-2\pi i \nu s} \\ &= 1 - \sum_{s=1}^{p} \phi_s e^{-2\pi i \nu s} \\ &= \phi(e^{-2\pi i \nu}), \end{aligned}$$

so $\sigma_w^2 = f_w(\nu) = |A(\nu)|^2 f_X(\nu)$ and

$$f_X(\nu) = \frac{\sigma_w^2}{|\phi(e^{-2\pi i \nu})|^2}. \qquad (16.1)$$

- Example: $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + w_t$. Then

$$\begin{aligned}
& \phi(e^{-2\pi i\nu}) = 1 - \phi_1 e^{-2\pi i\nu} - \phi_2 e^{-4\pi i\nu}, \\
& \bar{\phi}(e^{-2\pi i\nu}) = 1 - \phi_1 e^{2\pi i\nu} - \phi_2 e^{4\pi i\nu}, \\
& |\phi(e^{-2\pi i\nu})|^2 \\
= \; & \left[1 - \phi_1 e^{-2\pi i\nu} - \phi_2 e^{-4\pi i\nu}\right] \\
& + \left[-\phi_1 e^{2\pi i\nu} + \phi_1^2 + \phi_1\phi_2 e^{-2\pi i\nu}\right] \\
& + \left[-\phi_2 e^{4\pi i\nu} + \phi_1\phi_2 e^{2\pi i\nu} + \phi_2^2\right] \\
= \; & 1 + \phi_1^2 + \phi_2^2 \\
& + [\phi_1\phi_2 - \phi_1]\left[e^{-2\pi i\nu} + e^{2\pi i\nu}\right] \\
& - \phi_2\left[e^{-4\pi i\nu} + e^{4\pi i\nu}\right] \\
= \; & 1 + \phi_1^2 + \phi_2^2 \\
& + 2\phi_1(\phi_2 - 1)\cos(2\pi\nu) - 2\phi_2\cos(4\pi\nu)
\end{aligned}$$

and so by (16.1), $f_X(\nu) =$

$$\frac{\sigma_w^2}{1 + \phi_1^2 + \phi_2^2 + 2\phi_1(\phi_2 - 1)\cos(2\pi\nu) - 2\phi_2\cos(4\pi\nu)}.$$

See Figure 16.1, where $f_X(\nu)$ is plotted with $\phi_1 = 1$, $\phi_2 = -.9$, $\sigma_w^2 = 1$. This suggests that if $\{Z_t\}$ is a series whose interesting frequencies are in a

narrow band around .15, then the (recursive) filter

$$X_t = X_{t-1} - .9X_{t-2} + Z_t$$

will highlight these frequencies and largely eliminate others.

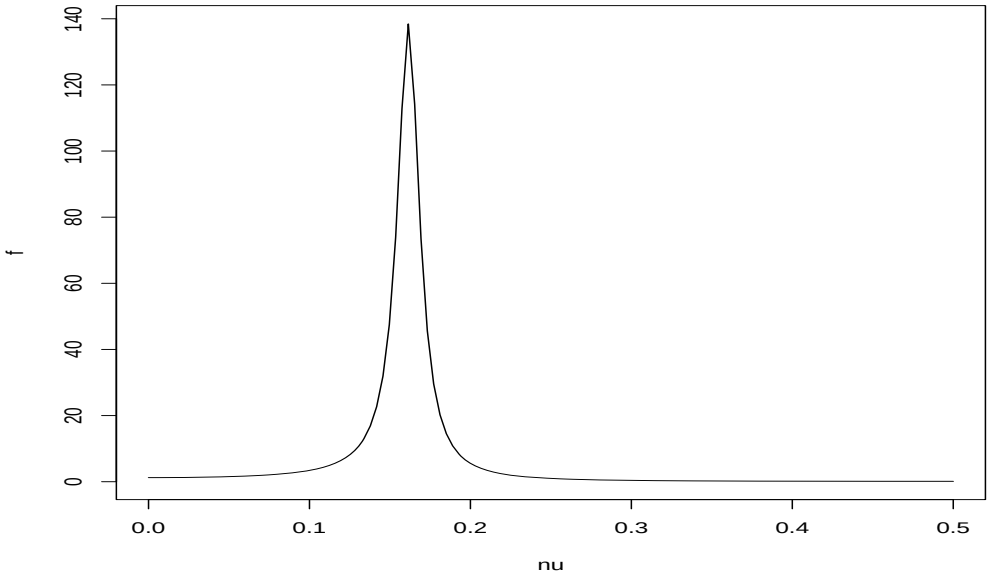

Figure 16.1. Power spectrum

$$f_X(\nu) = \frac{f_w(\nu)}{\left|\phi\left(e^{2\pi i\nu}\right)\right|^2} = \frac{\sigma_w^2}{|A(\nu)|^2}$$

of AR(2) series $X_t = X_{t-1} - .9X_{t-2} + w_t$.

- You should show, using the same method, that if $X_t = \theta(B)w_t$ is MA(q) , then

$$f_X(\nu) = \sigma_w^2 |\theta(e^{-2\pi i\nu})|^2.$$

This extends (how?) to: If $\phi(B)X_t = \theta(B)w_t$, so that X_t is ARMA(p,q) then

$$f_X(\nu) = \sigma_w^2 \frac{|\theta(e^{-2\pi i\nu})|^2}{|\phi(e^{-2\pi i\nu})|^2}. \tag{16.2}$$

This indicates one way to estimate the spectrum - fit an ARMA model to the data in the time domain, and then substitute estimates of the ARMA parameters into (16.2) to get an estimate $\hat{f}_X(\nu)$. Some details are in §4.6 of the text. In the next class we will look at a more common method of estimating the power, using the 'periodogram'.

- Note that

$$\begin{aligned} f_{YX}(\nu) &= |f_{YX}(\nu)| \frac{c_{YX}(\nu) - iq_{YX}(\nu)}{\sqrt{c_{YX}^2(\nu) + q_{YX}^2(\nu)}} \\ &= |f_{YX}(\nu)| \, e^{i\omega}, \end{aligned}$$

where

$$\begin{aligned} \cos(\omega) &= c_{YX}(\nu)/\sqrt{c_{YX}^2(\nu) + q_{YX}^2(\nu)}, \\ \sin(\omega) &= -q_{YX}(\nu)/\sqrt{c_{YX}^2(\nu) + q_{YX}^2(\nu)}, \\ \tan(\omega) &= -q_{YX}(\nu)/c_{YX}(\nu); \end{aligned}$$

these are summarized by writing

$$\begin{aligned} f_{YX}(\nu) &= |f_{YX}(\nu)| \, e^{i\phi_{YX}(\nu)}, \text{ where} \\ \phi_{YX}(\nu) &= \tan^{-1}\left(-\frac{q_{YX}(\nu)}{c_{YX}(\nu)}\right) \text{ is the } \textit{phase}. \end{aligned}$$

In terms of the coherence,

$$f_{YX}(\nu) = \sqrt{\rho_{YX}^2(\nu) f_Y(\nu) f_X(\nu)} e^{i\phi_{YX}(\nu)}. \tag{16.3}$$

Estimates of the terms on the rhs of (16.3) are computed by R; then $f_{YX}(\nu)$ can be estimated by plugging in these estimates.

17. Discrete Fourier Transform

The theorem regarding filters suggests another possible application. Briefly, given (data from) a series $\{X_t\}_{t=-\infty}^{\infty}$ we will consider a filtered series $Y_t = \sum_{s=-\infty}^{\infty} a_s X_{t-s}$, where the coefficients $\{a_s\}_{s=-\infty}^{\infty}$ are to be chosen by us, in order that Y_t exhibit certain desired properties. As above, $f_Y(\nu) = |A(\nu)|^2 f_X(\nu)$. Thus we might:

1. Estimate $f_X(\nu)$ from the data. Determine from this which frequencies are the "interesting" ones, which we would like to have highlighted.

2. Choose the desired form of $|A(\nu)|^2$ appropriately, e.g. choose it to be large at the interesting frequencies, small elsewhere.

3. Invert $A(\nu)$ to obtain $a_t = \int_{-1/2}^{1/2} e^{2\pi i \nu t} A(\nu) d\nu$.

4. Compute the Y_t's.

- Here is an outline of these steps; details to follow. Recall the SOI series, related to the El Niño effect. Figure 17.1 gives estimates of the spectrum, and log-spectrum. Most of the power is at frequencies of about .25, 1, 2, 3 cycles/year (periods of about 4 years, 1 year, etc.). The first of these arises from the El Niño effect, the second is an annual periodic trend, as expected. The others are harder to interpret – seasons? .

- Since observations are made monthly, the El Niño frequency is about .02 cycles/month, the annual about .08 cycles/month. To isolate the El Niño effect we might 'filter out' frequencies $>$ .05, for instance:
  `out = SigExtract(soi, L=9, M=64, max.freq=.05).`
  See Figures 17.2, 17.3 for the results.

- All this will be discussed in more detail in Lecture 20.

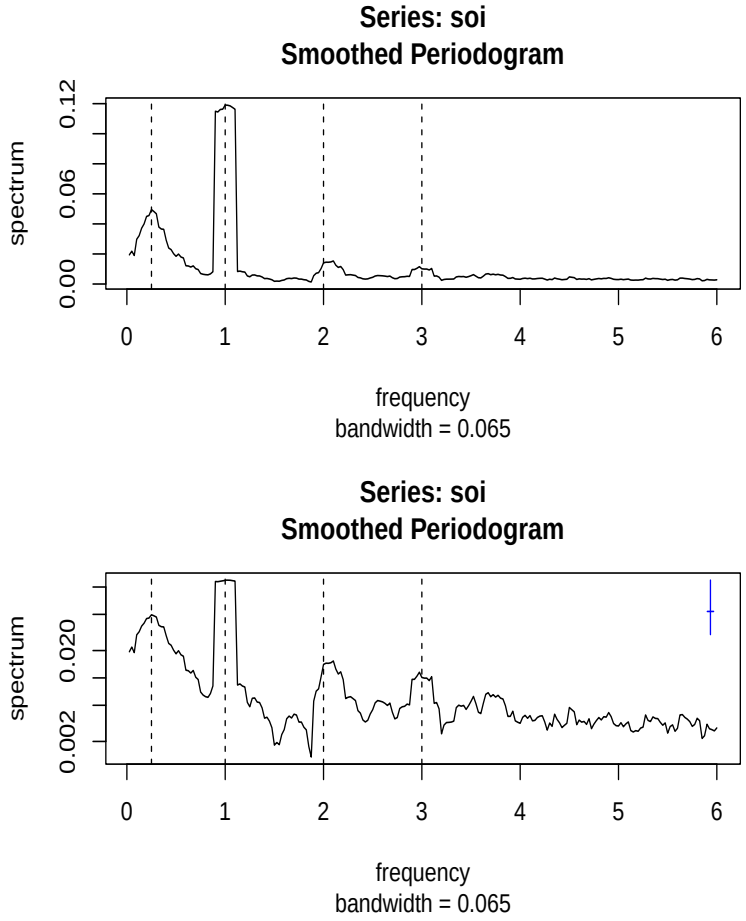

Figure 17.1. Southern oscillation index; estimated spectrum and log-spectrum.

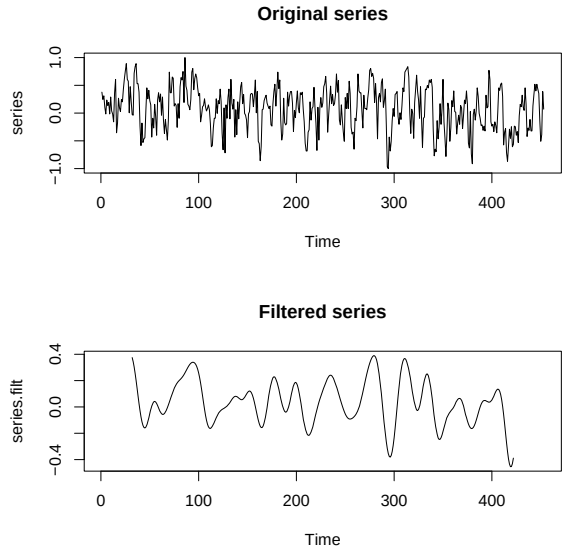

Figure 17.2. Original and filtered SOI series.

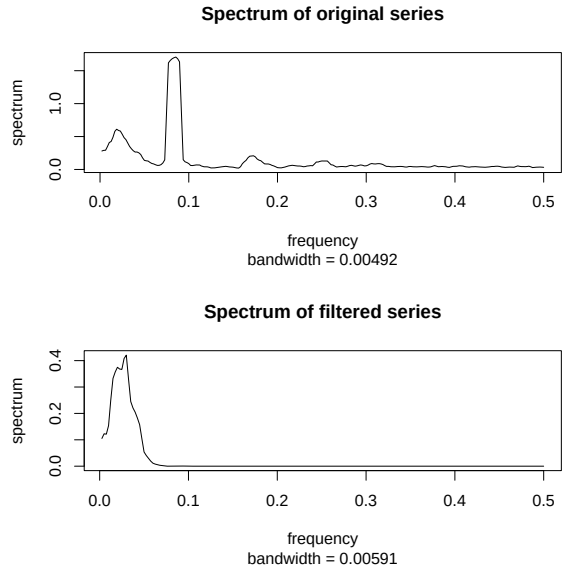

Figure 17.3. Spectrum of original and filtered SOI series.

- A filter with IFT similar to that here, or of the MA(3) plotted earlier (Figure 15.1) is called a "low-pass" filter - only the low frequencies (long period trends) are "passed". A filter whose IFT is similar to that of the AR(2) plotted earlier (Figure 16.1) is a "band-pass" filter - only frequencies in a narrow band (around $\nu = .15$ in this case) are passed.

- To apply any of this we require a means of estimating the spectrum and/or cross spectrum. The obvious estimate of

$$f(\nu) = \sum_{m=-\infty}^{\infty} e^{-2\pi i \nu m} \gamma(m),$$

using data $\{x_t\}_{t=1}^{n}$, is

$$\hat{f}(\nu) = \sum_{m=-(n-1)}^{n-1} e^{-2\pi i \nu m} \hat{\gamma}(m).$$

With $\nu_k = k/n$, $k = 1, ..., n$, and after some algebra this reduces to the "periodogram":

$$\hat{f}(\nu_k) = |X(k)|^2,$$

where $X(k)$ is the "Discrete Fourier Transform" (DFT) of the data:

$$X(k) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} x_t e^{-2\pi i \nu_k t}. \tag{17.1}$$

(This is also written $X(\nu_k)$, and $|X(k)|^2$ is sometimes written as $I(\nu_k)$).

- The real and imaginary parts

$$\begin{aligned} X_C(k) &= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} x_t \cos(2\pi\nu_k t), \\ X_S(k) &= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} x_t \sin(2\pi\nu_k t) \end{aligned}$$

are the *cosine* and *sine* transforms of the data, and then $X(k) = X_C(k) - iX_S(k)$. Thus

$$\hat{f}(\nu_k) = |X(k)|^2 = X_C^2(k) + X_S^2(k) = I(\nu_k).$$

- **Inversion Theorem**. The DFT contains all of the information in the data, in that the data can be recovered via

$$x_t = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} X(k) e^{2\pi i \nu_k t}. \tag{17.2}$$

  - Verifying this requires an important preliminary result. For any integer t,

$$\sum_{k=1}^{n} e^{2\pi i \nu_k t} = \begin{cases} n, & \text{if } \frac{t}{n} \text{ is an integer,} \\ 0, & \text{otherwise.} \end{cases}$$

    Reason: If t/n is an integer then $\nu_k t = kt/n$ is an integer and the result follows. Otherwise

$$\sum_{k=1}^{n} e^{2\pi i \nu_k t} = \sum_{k=1}^{n} z^k_{|z=e^{2\pi i t/n} \neq 1} = \frac{z(1-z^n)}{1-z} = 0,$$

    since $z^n = e^{2\pi i t} = 1$.

    Note: Remembering this derivation is easier than remembering the precise form of the result itself.

– **Proof of Inversion Theorem**: The RHS of (17.2) is

$$\begin{aligned}
& \frac{1}{\sqrt{n}} \sum_{k=1}^{n} X(k) e^{2\pi i \nu_k t} \\
= & \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \left[\frac{1}{\sqrt{n}} \sum_{s=1}^{n} x_s e^{-2\pi i \nu_k s} \right] e^{2\pi i \nu_k t} \\
= & \frac{1}{n} \sum_{s=1}^{n} x_s \sum_{k=1}^{n} e^{2\pi i \nu_k (t-s)} \\
= & \frac{1}{n} \sum_{s=1}^{n} \left\{ x_s \cdot \left[n \cdot I(\frac{t-s}{n} \text{ an integer}) \right] \right\} \\
= & \frac{1}{n} x_t n \text{ (how?)} \\
= & x_t.
\end{aligned}$$

- A practical concern is that non-stationary trends (e.g. linear trends $\beta_0 + \beta_1 t$) should be removed from the data before estimating the spectrum. In R the data are, by default, detrended before the periodogram is computed - a linear regression on t is carried out and the rest of the analysis is carried out on the residuals from this regression.

- Removing a non-zero average alone will not affect $\hat{f}(\nu_k)$ for $\nu_k < 1$:

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \bar{x} e^{-2\pi i \nu_k t} = \frac{\bar{x}}{\sqrt{n}} \sum_{t=1}^{n} e^{-2\pi i \nu_k t} \\ = & \frac{\bar{x}}{\sqrt{n}} \sum_{t=1}^{n} \left(e^{-2\pi i (k/n)} \right)^t = 0, \end{aligned}$$

  for $k < n$. (Recall that we're really only interested in estimating $f(\nu_k)$ for $0 \leq \nu_k \leq .5$)

- Similarly, when $X(k)$ is viewed as a r.v., we have (for $\nu_k < 1$)

$$E\left[X(k)\right] = \frac{\mu_X}{\sqrt{n}} \sum_{t=1}^{n} e^{-2\pi i \nu_k t} = 0,$$

so that

$$X_C(k) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_t \cos(2\pi\nu_k t)$$

and $X_S(k)$ both have means of 0. One can show that

$$VAR\left[X_C(k)\right] \text{ and } VAR\left[X_S(k)\right] \approx \frac{f\left(\nu_k\right)}{2},$$

and that $\frac{X_C(k)}{\sqrt{f(\nu_k)/2}}$ and $\frac{X_S(k)}{\sqrt{f(\nu_k)/2}}$ are approximately $N(0,1)$ (we write $\overset{d}{\approx} N(0,1)$) and approximately independent (exact as $n \to \infty$). This is a consequence of an appropriate version of the Central Limit Theorem, which asserts the approximate normality of averages:

$$\frac{X_C(k)}{\sqrt{f(\nu_k)/2}} = \sqrt{n} \cdot \left(\begin{array}{c} \text{average of the zero mean,} \\ \text{unit variance r.v.s } X_t \frac{\cos(2\pi\nu_k t)}{\sqrt{f(\nu_k)/2}} \end{array} \right).$$

- Since these r.v.s are approximately normal and approximately independent:

$$\frac{\hat{f}(\nu_k)}{f(\nu_k)/2} = \left\{\frac{X_C(k)}{\sqrt{f(\nu_k)/2}}\right\}^2 + \left\{\frac{X_S(k)}{\sqrt{f(\nu_k)/2}}\right\}^2 \overset{d}{\approx} \text{ what?}$$

  By this, $\hat{f}(\nu_k) \overset{d}{\approx} (f(\nu_k)/2)\,\chi^2_2$, with

$$\begin{aligned} E\left[\hat{f}(\nu_k)\right] &\approx \frac{f(\nu_k)}{2} E\left[\chi^2_2\right] = f(\nu_k), \quad (17.3) \\ VAR\left[\hat{f}(\nu_k)\right] &\approx \left\{\frac{f(\nu_k)}{2}\right\}^2 VAR\left[\chi^2_2\right] = f^2(\nu_k). \end{aligned}$$

- For $k \neq l$, $\hat{f}(\nu_k)$ and $\hat{f}(\nu_l)$ are approximately independent - this causes problems to be dealt with later.

- Write the above as

$$\hat{f}(\nu_k) \overset{d}{\approx} \frac{f(\nu_k)}{df}\chi^2_{df}$$

  with $df = 2$. (Modifications to come.)

18. Computing the periodogram and cross-periodogram

- The algorithm used to compute $\hat{f}(\nu_k)$ ("Fast Fourier Transform") works best when $n = 2^m$ for some integer m. It still works well if n has many factors of $2, 3$ or 5. Thus, let n' (= `nextn(n)` in R) be the next 'good' value of n, and add $n' - n$ zeroes to the end of the data. This is done by default in R, and has no effect on the value of

$$X(k) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n'} x_t e^{-2\pi i \nu_k t}$$

(note the $\sqrt{n}$ remains) but $\nu_k = k/n'$, and also there is a modification required in the degrees of freedom:

$$df = 2n/n' (\leq 2).$$

- Note that the variance of $\hat{f}(\nu_k)$, hence the width of CIs, does not decrease as the number of observations increases. This is contrary to the common procedures (t-intervals, etc.) and results in highly varied, "jiggly" periodograms. A remedy is to "smooth" the periodogram by averaging adjacent values. Let L be an odd integer, typically much less than n. Let $\hat{f}(\nu_k)$ now be the average of the L values

$$I(\nu_k + \frac{l}{n}),\ l = 0, \pm 1, ..., \pm\frac{L-1}{2},$$

i.e.

$$\hat{f}(\nu_k) = \frac{1}{L} \sum_{l=-\frac{L-1}{2}}^{\frac{L-1}{2}} |X(k+l)|^2.$$

If $L = 1$ this is the "raw" periodogram used earlier. In general it is a centred moving average, of length L, applied to the raw periodogram. The distributional approximations above continue to hold, with the change

$$df = 2L\frac{n}{n'}.$$

We still have unbiasedness:

$$E\left[\hat{f}(\nu_k)\right] \approx \left[\frac{f\left(\nu_k\right)}{df}\chi^2_{df}\right] = f\left(\nu_k\right),$$

but now

$$\begin{aligned} VAR\left[\hat{f}(\nu_k)\right] &\approx \left\{\frac{f(\nu_k)}{df}\right\}^2 VAR\left[\chi^2_{df}\right] \\ &= f^2(\nu_k)\frac{2}{df} \\ &= f^2(\nu_k)\frac{n'}{Ln}; \end{aligned}$$

this decreases at the rate $1/L$. The "bandwidth" L (this is <u>not</u> what R calls the bandwidth) is typically chosen by trial and error. A value too small results in highly varied estimates $\hat{f}(\nu_k)$ ("jiggly" plots), and a value too large smooths $\hat{f}(\nu_k)$ too much - important features of the data may be lost. Common values of L - for the series of moderate length used in this course - are $3, 5, ..., 21$. The value is typically chosen by trial and error, upon examining the plots.

- Confidence intervals: Let χ^2_L and χ^2_U be the lower and upper $\alpha/2$ points in the χ^2_{df} distribution; then

$$\begin{aligned} 1-\alpha &= P\left(\chi^2_L \le \frac{\hat{f}(\nu_k)}{f(\nu_k)/df} \le \chi^2_U\right) \\ &= P\left(\frac{df\cdot\hat{f}(\nu_k)}{\chi^2_U} \le f(\nu_k) \le \frac{df\cdot\hat{f}(\nu_k)}{\chi^2_L}\right), \end{aligned}$$

so that

$$\left[\frac{df\cdot\hat{f}(\nu_k)}{\chi^2_U}, \frac{df\cdot\hat{f}(\nu_k)}{\chi^2_L}\right]$$

is a $100(1-\alpha)\%$ (approximately) confidence interval. Note that the width is a multiple of $\hat{f}(\nu_k)$. A remedy, leading to fixed width CIs, is to write

$$1-\alpha = P\left(\ln\frac{df\cdot\hat{f}(\nu_k)}{\chi^2_U} \le \ln f(\nu_k) \le \ln\frac{df\cdot\hat{f}(\nu_k)}{\chi^2_L}\right),$$

so that a $100(1-\alpha)\%$ CI on $\ln f(\nu_k)$ is

$$\left[\ln\hat{f}(\nu_k) - \ln\left(\frac{\chi^2_U}{df}\right), \ln\hat{f}(\nu_k) + \ln\left(\frac{df}{\chi^2_L}\right)\right].$$

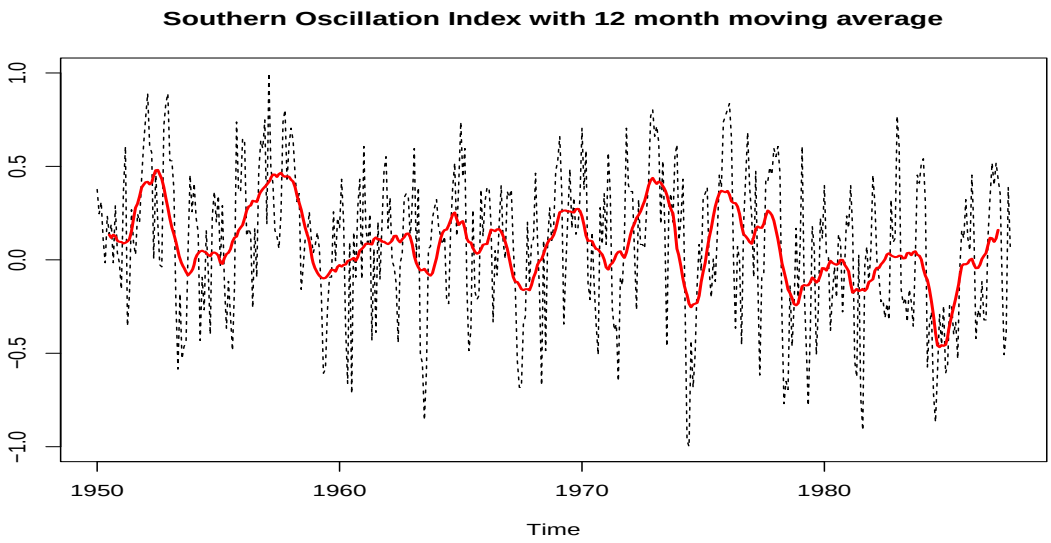

Figure 18.1. SOI series with 12-month centred moving average superimposed.

Example: Southern Oscillation Index (SOI). Here is the basic R command for periodogram computation (with no smoothing):

```
soi.per = spec.pgram(soi, log="??").
```

Notes: `log = "yes"` is the default; regardless of this the output is not logged - only the plotted values. The vertical axis is on a log-scale, if `log = "yes"`.

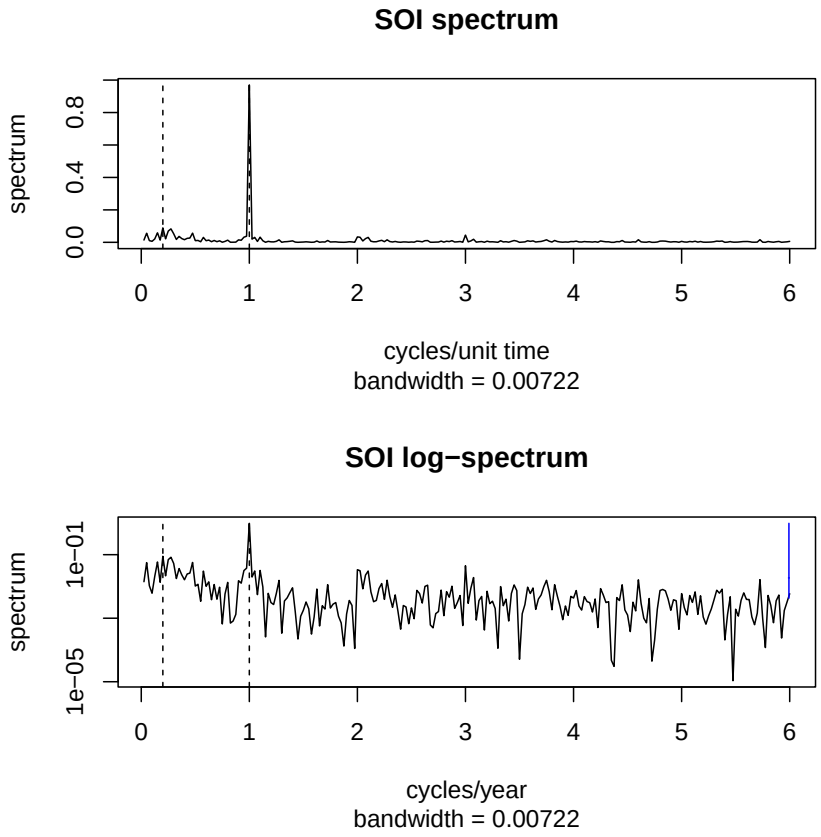

Figure 18.2. SOI periodograms. The top plot uses `log="no"`. In the lower plot `log="yes"` was specified. In this case the (constant) 95% CI width is also shown. Vertical lines - added by me - mark primary (annual) and secondary (= what?) peaks.

Here are some values of the first periodogram:

```
              freq        power
                 ...
   [7,] 0.014583333 1.352532e-02
  [8,] 0.016666667 8.896183e-02
  ***first peak; freq=1/60; period = 5 years
  [9,] 0.018750000 2.157570e-02
                 ...
  0.079166667 3.163875e-02
 [39,] 0.081250000 3.761849e-02
  ***max power; freq=1/12; period = 12 months
 [40,] 0.083333333 9.700486e-01
                  ...
[239,] 0.497916667 3.196408e-03
[240,] 0.500000000 5.902822e-03
```

Notes:

1. In R the series is, by default, detrended before the periodogram is computed.

2. $df = 2 \cdot 453/480 = 1.8875$ is also in the output if '`taper=0`' is specified; tapering will be discussed later. Use the default of '`taper=0.1`'; this results in a further small change (to 1.6908) in the d.f.

To smooth the series, decide on a value of L (an odd integer) and apply a centred moving average filter with coefficients `rep(1,L)/L` to the periodogram. Equivalently, set $m = (L-1)/2$, and specify that a "Daniell kernel of order m" be used:

```
k = kernel("daniell",1)  # L = 3, m = 1
soi.ave = spec.pgram(soi, k, log="yes")
```

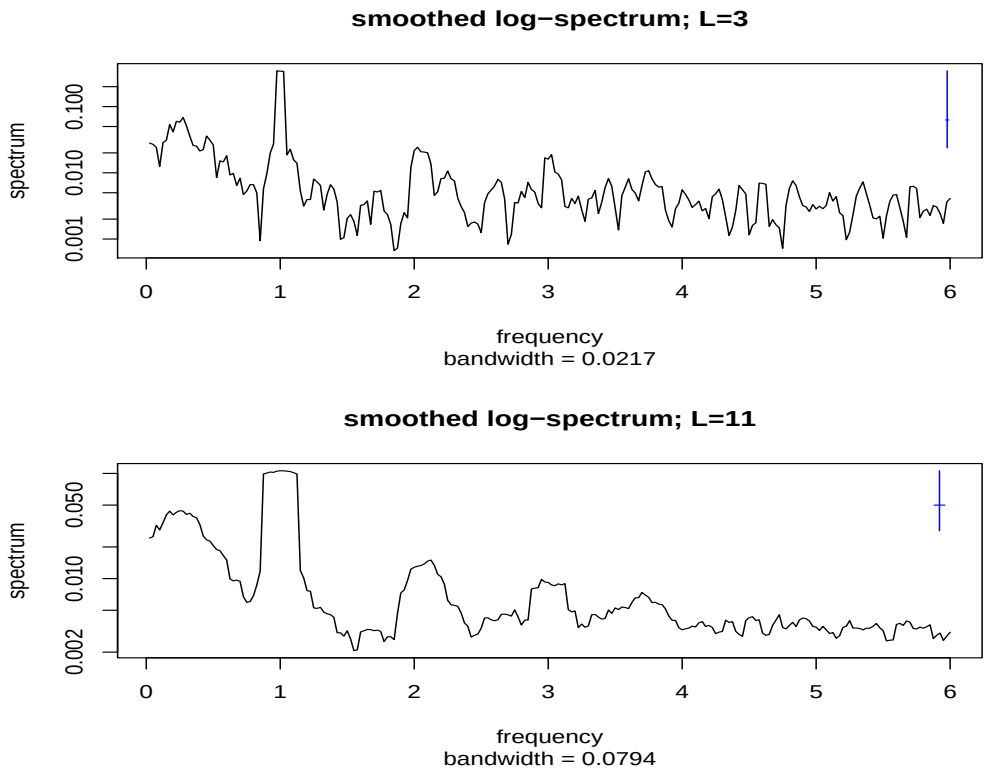

Figure 18.3. Smoothed log-spectra. First uses $L = 3$ ($df = 5.072$); second uses $L = 11$ ($df = 18.599$). The estimates of the El Niño period are 44 months and 48 months, respectively.

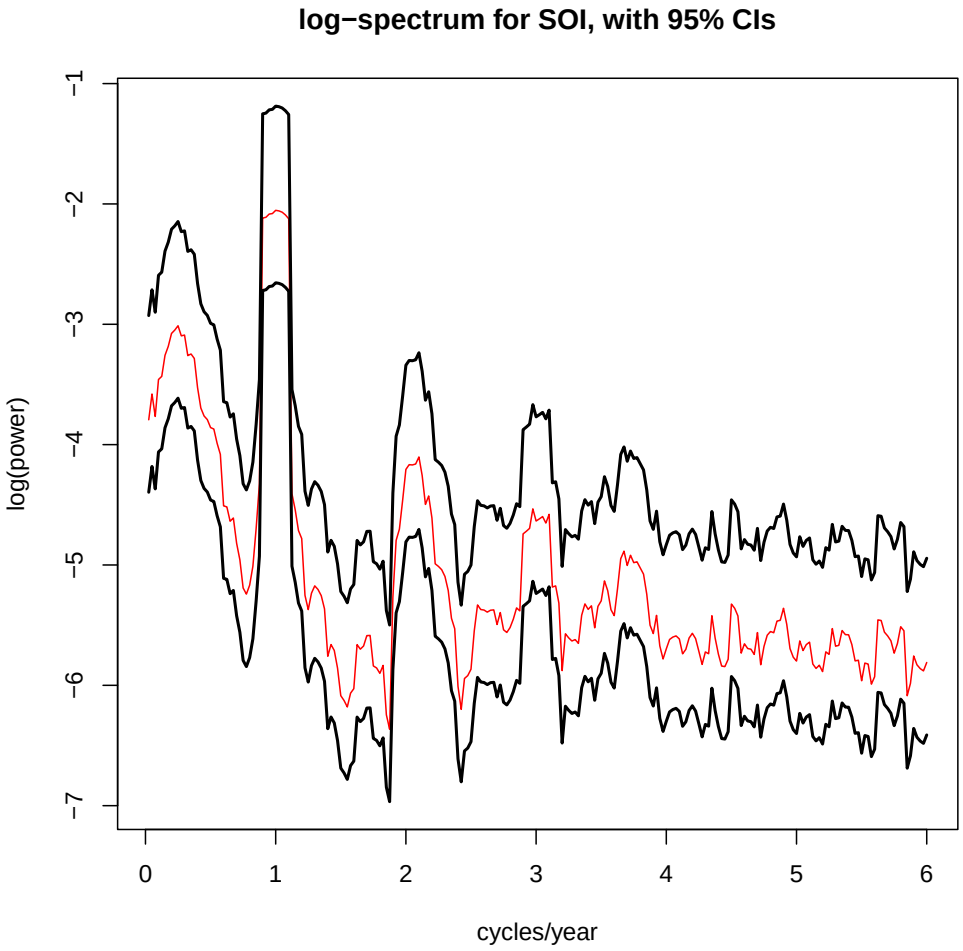

Figure 18.4. Log spectrum with 95% confidence band; $L = 9$. See code on website.

Similar to estimating the power, the cross spectrum $f_{XY}(\nu_k)$ is estimated by the smoothed *cross-periodogram*

$$\hat{f}_{XY}(\nu_k) = \frac{1}{L} \sum_{l=-\frac{L-1}{2}}^{\frac{L-1}{2}} X(k+l)\bar{Y}(k+l).$$

Then the squared coherence $\rho^2_{YX}(\nu) = \frac{|f_{YX}(\nu)|^2}{f_Y(\nu)f_X(\nu)}$ is estimated by

$$\hat{\rho}^2_{YX}(\nu) = \frac{|\hat{f}_{YX}(\nu)|^2}{\hat{f}_Y(\nu)\hat{f}_X(\nu)},$$

where (under the hypothesis that $\rho^2_{YX}(\nu) = 0$)

$$\frac{df-2}{2} \cdot \frac{\hat{\rho}^2_{YX}(\nu)}{1-\hat{\rho}^2_{YX}(\nu)} \overset{d}{\approx} F^2_{df-2}$$

and $df = 2Ln/n'$ as before.

- Example: Southern Oscillation Index and Recruits series.

```
x = ts(cbind(soi,rec))
s = spec.pgram(x, kernel("daniell",9),
    plot.type = "coh", ...)
```

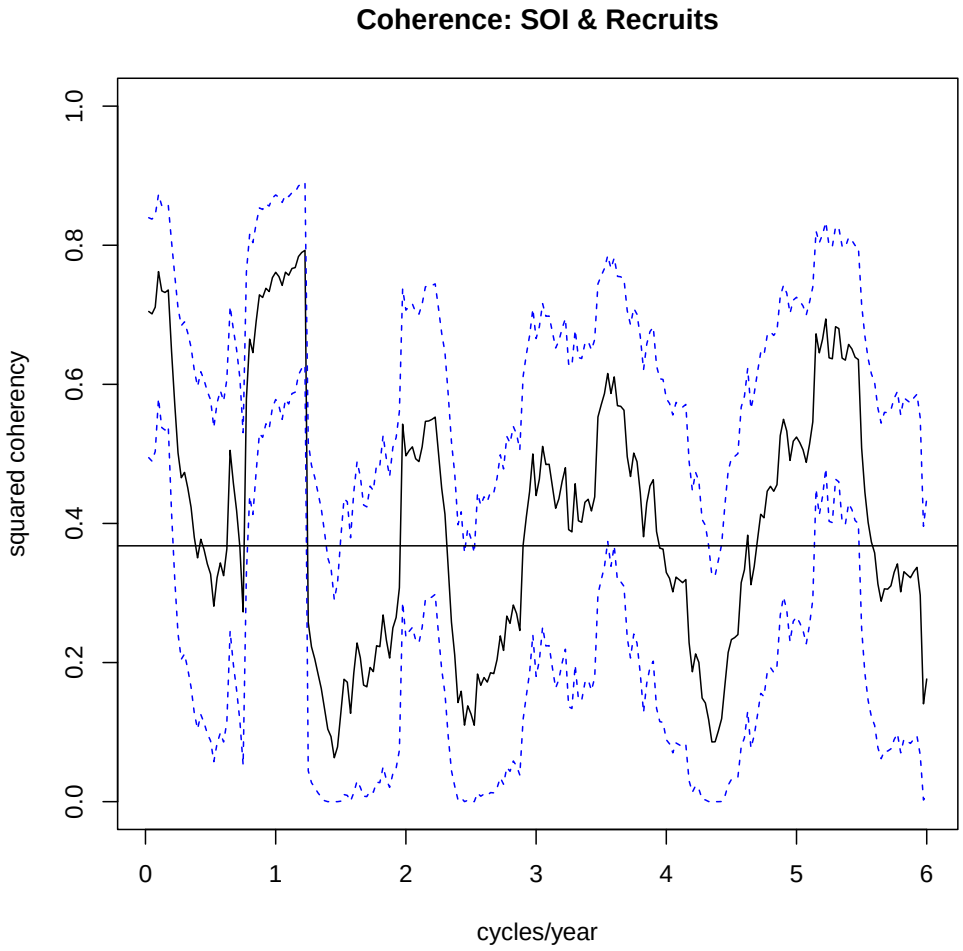

Figure 18.5. Coherence plot with 95% confidence bounds. Horizontal line is at $\alpha = .001$ critical value ($= \frac{f}{f + \frac{df-2}{2}}$, where $\alpha = P\left(F^2_{df-2} > f\right)$.) Series are strongly coherent at the annual and El Niño frequencies, also at several other higher frequencies, i.e. shorter periods (= seasons?).

- Consider the SOI series, or any other in which observations are made monthly (we say the "sampling interval" is 1 month). We have $\nu \leq .5$, so that only periods of length ≥ 2 (months) can be recognized, or "resolved". *Only periods exceeding twice the sampling interval can be resolved*. Equivalently, the "maximum sampling interval" is $1/2$ the length of the shortest period we wish to be able to recognize: for a monthly period we would need to sample every half-month (or more frequently).

- This should be intuitively clear. If, e.g. there is a weekly period then we won't recognize it if we make observations only monthly. At least 2 observations per week will be necessary if we are to observe changes during the week and see that they reoccur weekly.

- Since the power $f(\nu)$ is periodic, with period 1, the power at frequencies $\nu + j$, $j = \pm 1, \pm 2, ...$ is indistinguishable from that at ν. We say these frequencies are *aliased*. The problem of aliasing is avoided if the sampling interval is small enough that frequencies outside of $[-.5, .5]$ are of no interest, i.e. if the sampling interval is less than the maximum sampling interval.

19. Impulse-response problems

- Lagged regression. In the SOI/Recruits example (recall the series were strongly coherent at many frequencies), we might posit a time series regression model of the form

$$Y_t = \sum_{s=-\infty}^{\infty} \beta_s X_{t-s} + v_t$$

where Y_t is Recruits (minus the mean), X_t is SOI (minus the mean), and v_t is zero-mean stationary noise uncorrelated with $\{X_t\}$. We will estimate finitely many of the regression coefficients $\{\beta_s\}$ by $\left\{\hat{\beta}_s^M\right\}_{|s|<M/2}$, and then use these, or only the most significant of them, to obtain a finite sum

$$\hat{Y}_t = \sum \hat{\beta}_s^M X_{t-s}.$$

This can then be used to study the relationship between the series or to predict future Recruits from SOI (if only terms with $s > 0$ are used).

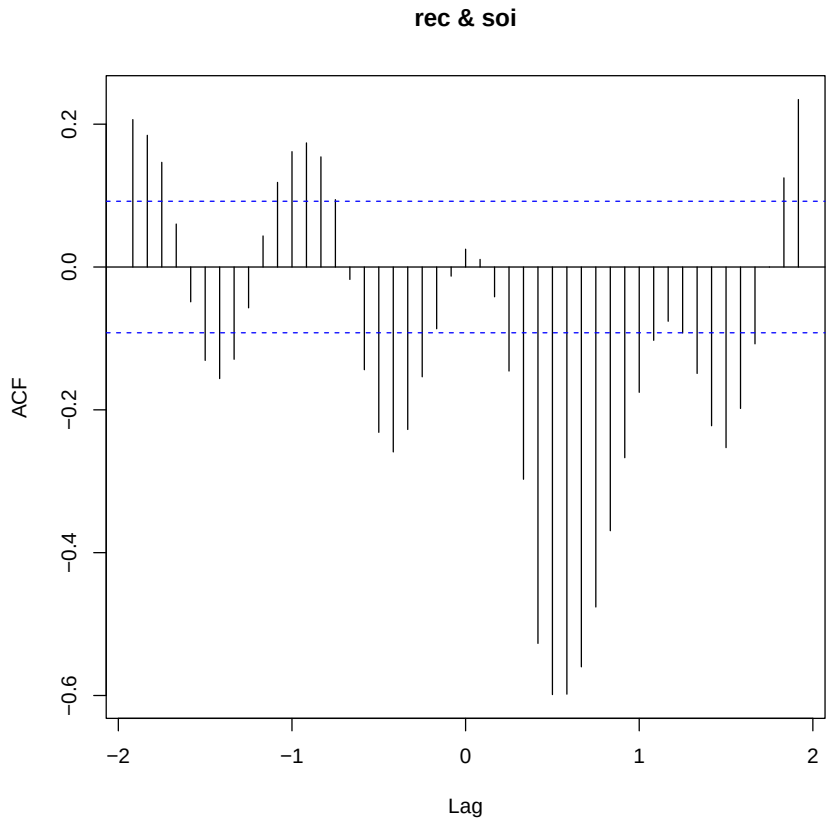

Figure 19.1. $CCF =$corr(REC_{t+m}, SOI_t): SOI leads Recruits.

- Also called an "impulse-response" problem - SOI the impulse, Recruits the response.

- First obtain the "best" coefficients β_s by minimizing the MSE:

$$MSE = E\left[\left\{Y_t - \sum_{s=-\infty}^{\infty} \beta_s X_{t-s}\right\}^2\right].$$

Differentiating w.r.t. β_r gives

$$\begin{aligned} 0 &= \frac{\partial MSE}{\partial \beta_r} = E\left[\frac{\partial}{\partial \beta_r}\left\{Y_t - \sum_{s=-\infty}^{\infty} \beta_s X_{t-s}\right\}^2\right] \\ &= -2E\left[\left\{Y_t - \sum_{s=-\infty}^{\infty} \beta_s X_{t-s}\right\} X_{t-r}\right]; \quad (19.1) \end{aligned}$$

i.e. (since the means have been removed)

$$\gamma_{YX}(r) = \sum_{s=-\infty}^{\infty} \beta_s \gamma_X(r-s) \quad (19.2)$$

for $r = 0, \pm 1, \pm 2, \ldots$. There are infinitely many equations in infinitely many unknowns. But things simplify (a lot!) in the frequency domain.

- In (19.2), replace $\gamma_{YX}(r)$ and $\gamma_X(r-s)$ by their expressions in terms of f_{YX} and f_X:

$$\begin{aligned}
& \int_{-1/2}^{1/2} e^{2\pi i\nu r} f_{YX}(\nu) d\nu \\
= & \sum_{s=-\infty}^{\infty} \beta_s \int_{-1/2}^{1/2} e^{2\pi i\nu(r-s)} f_X(\nu) d\nu \\
= & \int_{-1/2}^{1/2} \left[\sum_{s=-\infty}^{\infty} \beta_s e^{-2\pi i\nu s} \right] e^{2\pi i\nu r} f_X(\nu) d\nu \\
= & \int_{-1/2}^{1/2} B(\nu) e^{2\pi i\nu r} f_X(\nu) d\nu,
\end{aligned}$$

  where $B(\nu)$ is the IFT of $\{\beta_s\}$. These coefficients $\{\beta_s\}$ form what is called the "impulse response function", and $B(\nu)$ is the "frequency response function".

- By uniqueness of Fourier transforms,

$$f_{YX}(\nu) = B(\nu) f_X(\nu)$$

  and so $B(\nu) = f_{YX}(\nu) / f_X(\nu)$; then

$$\beta_s = \int_{-1/2}^{1/2} B(\nu) e^{2\pi i\nu s} d\nu.$$

- Note that $\beta_s = \bar{\beta}_s$, and so is real. **Reason**:

$$\bar{B}(\nu) = \frac{\bar{f}_{YX}(\nu)}{\bar{f}_X(\nu)} = \frac{f_{YX}(-\nu)}{f_X(\nu)} = \frac{f_{YX}(-\nu)}{f_X(-\nu)} = B(-\nu),$$

so

$$\begin{aligned}
\bar{\beta}_s &= \int_{-1/2}^{1/2} \overline{B(\nu) e^{2\pi i \nu s}} d\nu \\
&= \int_{-1/2}^{1/2} \bar{B}(\nu) e^{-2\pi i \nu s} d\nu \\
&= \int_{-1/2}^{1/2} B(-\nu) e^{-2\pi i \nu s} d\nu \\
&= \int_{-1/2}^{1/2} B(\omega) e^{2\pi i \omega s} d\omega \\
&= \beta_s.
\end{aligned}$$

Recall from Lecture 16 that

$$f_{YX}(\nu) = \sqrt{\rho^2_{YX}(\nu) f_Y(\nu) f_X(\nu)} e^{i\phi_{YX}(\nu)};$$

this results in

$$B(\nu) = \sqrt{\rho^2_{YX}(\nu) \frac{f_Y(\nu)}{f_X(\nu)}} e^{i\phi_{YX}(\nu)}, \tag{19.3}$$

with estimates of all terms on the right being computed in R.

- In practice, for a set of frequencies $\omega_k = k/M$ ($k = 1, 2, ..., M/2$; M should be even, and much smaller than n), we compute $\hat{B}(\omega_k)$. Then

$$\hat{\beta}_s = \int_{-1/2}^{1/2} \hat{B}(\nu) e^{2\pi i \nu s} d\nu$$

is evaluated by discretizing the integral:

$$\begin{aligned}
\hat{\beta}_s^M &= \frac{1}{M} \sum_{k=1}^{M/2} \left[\hat{B}(\omega_k) e^{2\pi i \omega_k s} + \hat{B}(-\omega_k) e^{-2\pi i \omega_k s} \right] \\
&= \frac{1}{M} \sum_{k=1}^{M/2} \left[\hat{B}(\omega_k) e^{2\pi i \omega_k s} + \overline{\hat{B}(\omega_k) e^{2\pi i \omega_k s}} \right] \\
&= \text{Re} \left\{ \frac{1}{M/2} \sum_{k=1}^{M/2} \hat{B}(\omega_k) e^{2\pi i \omega_k s} \right\}.
\end{aligned}$$

This is done for the $M - 1$ values s with $|s| < M/2$. Then we predict Y_t (now including the means) by the finite filter $\hat{Y}_t - \bar{Y} = \sum_{|s|<M/2} \hat{\beta}_s^M \left(X_{t-s} - \bar{X} \right)$, i.e.

$$\begin{aligned}
\hat{Y}_t &= \hat{\alpha} + \sum_{|s|<M/2} \hat{\beta}_s^M X_{t-s}, \text{ with} \\
\hat{\alpha} &= \bar{Y} - \sum_{|s|<M/2} \hat{\beta}_s^M \bar{X}.
\end{aligned}$$

- In the SOI and Recruits series $n = 453$, so that $n' = \texttt{nextn}(453) = 480$. I took $L = 15$ and $M = 32$. The relevant function to compute and invert $\hat{B}$ is 'LagReg':
  `out.ir.1=LagReg(input=soi,output=rec,L=15, M=32,threshold=6.9,inverse=FALSE).`
  This function first calls the R function `spec.pgram` to estimate $B(\nu)$ as at (19.3). The estimate is then evaluated at the frequencies ω_k described above, yielding estimates $\hat{B}(\omega_k)$. From these, the $\hat{\beta}_s^M$ are computed.

- See help(LagReg) in R - the program first uses '`threshold = 0`' to compute and exhibit all of the coefficients. It then uses only those which are larger in absolute value than the user-chosen 'threshold'. In this example the threshold of 6.9 was chosen so that all the coefficients fitted, with $s \geq 0$, will exceed (in absolute value) all those with $s < 0$. Then the intercept $\hat{\alpha}$ is fitted.

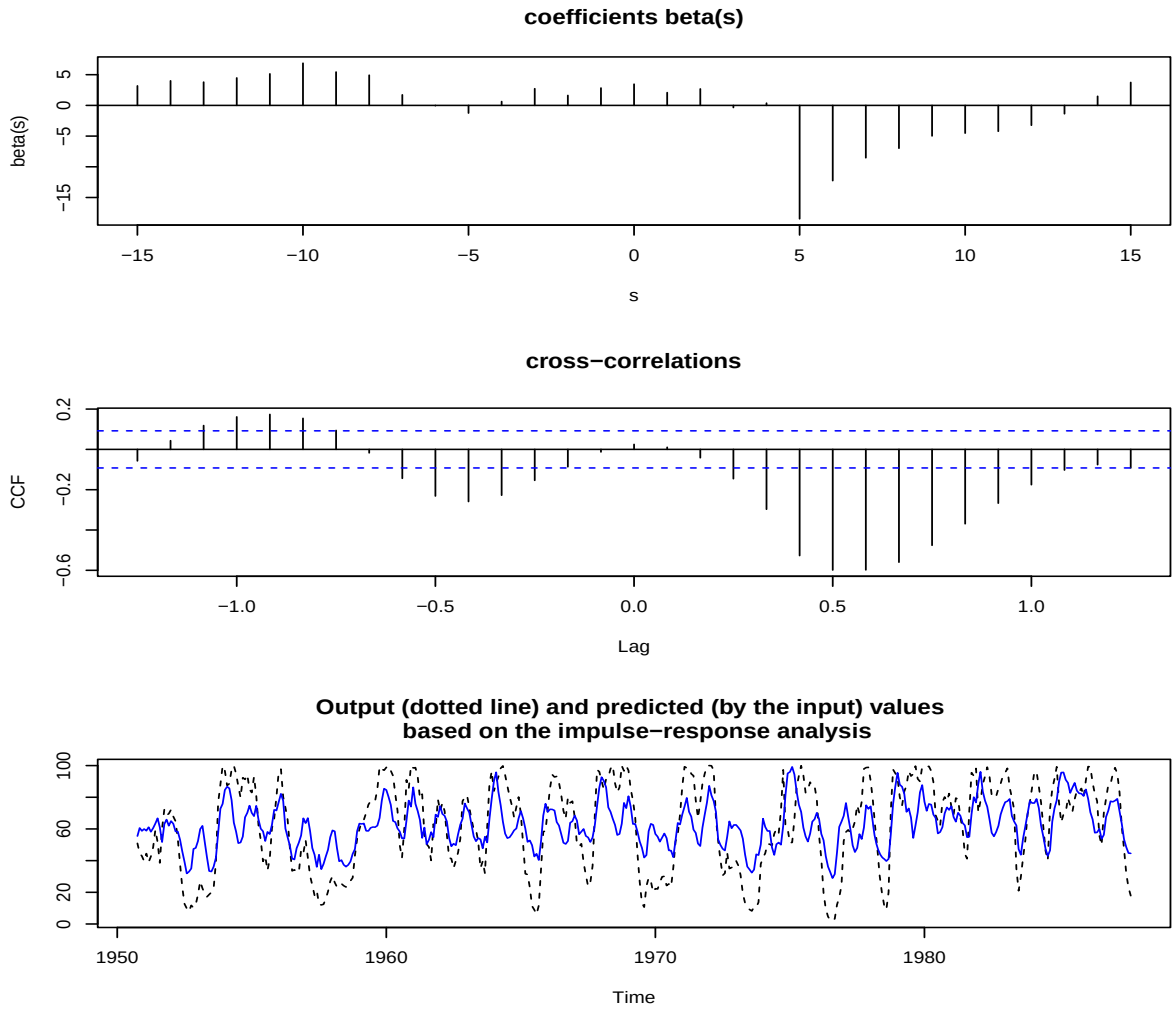

Figure 19.2. Impulse-response analysis of SOI/Recruits. Top: Impulse response function $\left\{\hat{\beta}_s^M\right\}$. Middle: CCF(REC$_{t+m}$, SOI$_t$). Bottom: Recruits $\{Y_t\}$ and predictions $\left\{\hat{Y}_t = \hat{\alpha} + \sum \hat{\beta}_s^M X_{t-s}\right\}$, using the backward (threshold $= 6.9$) impulse response function.

The output accompanying Figure 19.2 includes:

```
 L = 15 M = 32

The positive lags, at which the coefficients
are large in absolute value, and the coefficients
themselves, are:
     lag s    beta(s)
[1,]     5 -18.479306
[2,]     6 -12.263296
[3,]     7  -8.539368
[4,]     8  -6.984553

The prediction equation is
rec(t) = alpha + sum_s[ beta(s)*soi(t-s) ],
where alpha = 65.96584
MSE = 414.0847
```

The output is a bit sensitive to the choices of L and M; I experimented with them and made a final choice on the basis of the $MSE = \sum \left(Y_t - \hat{Y}_t\right)^2 / (n - M + 2)$.

- All this is only possible for $M/2 \leq t \leq n+1-M/2$ - the first and last $M/2-1$ values of $\hat{Y}_t$ can't be computed since, for these, $t-s$ becomes < 1 or $> n$.

- Setting inverse=TRUE will fit a forward-lagged regression (only $s \leq 0$ used); the default is to run a backward-lagged regression.

- A more parsimonious model is obtained by first viewing Recruits as the input and using inverse = TRUE to get a forward regression:
  `out.ir.2 = LagReg(rec, soi, L=15, M=32, inverse=TRUE, threshold=.005)`
  This gives output:

```
     lag s     beta(s)
[1,]     3  0.01593167
[2,]     4 -0.02120013
The prediction equation is
soi(t) = alpha + sum_s[ beta(s)*rec(t+s) ],
where alpha = 0.4080661
```

- The model at this point is

$$X_t = \alpha + \beta_4 Y_{t+4} + \beta_5 Y_{t+5},$$

with $\hat{\alpha} = .4081$, $\hat{\beta}_4 = .0159$, $\hat{\beta}_5 = -.0212$ and $Y = Rec, X = SOI$. (**Important**: for a forward lagged-regression get the lags from the plot, not the printed output). Re-arranging and shifting the time gives

$$Y_t + \frac{\beta_4}{\beta_5} Y_{t-1} = -\frac{\alpha}{\beta_5} + \frac{1}{\beta_5} X_{t-5},$$

resulting in

$$\hat{Y}_t = 19.2483 + .7515 Y_{t-1} - 47.1695 X_{t-5}. \tag{19.4}$$

Alternatively one might re-estimate the coefficients by regression (using the R function `dynlm` to regress Y_t on Y_{t-1} and X_{t-5}: `dynlm(rec~L(rec,1) + L(soi,5))`), yielding

$$\hat{Y}_t = 11.3136 + .8434 Y_{t-1} - 20.3004 X_{t-5}. \tag{19.5}$$

See the bottom plot in Figure 19.3 - this is the best fitting model among those we have tried.

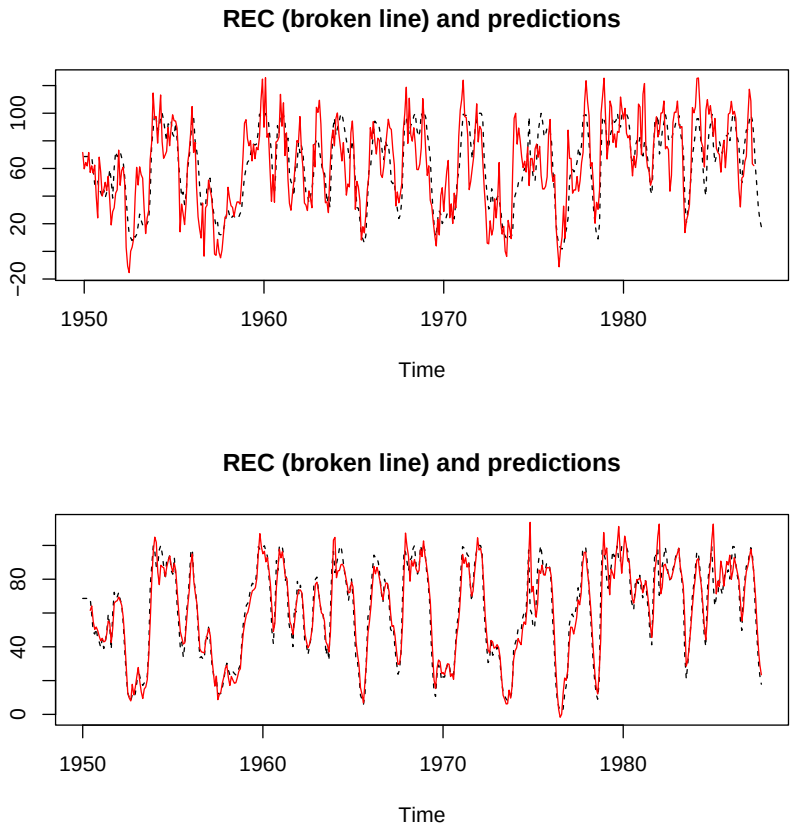

Figure 19.3. Recruits Y_t and $\hat{Y}_t$ predicted using (19.4) (top) and (19.5) (bottom); MSE values 370.7 and 59.4 respectively.

- This method - reversing the roles of input and output, and then shifting and re-arranging - is something you might keep in mind for your projects.

- In ordinary regression, the SS of Y around $\bar{Y}$ ("total SS") is decomposed as that of Y around $\hat{Y}$ and that of $\hat{Y}$ around $\bar{Y}$ ("unexplained + explained" variation; note $\bar{Y}$ is also the average of the fitted values). The second of these ("explained") is r^2 times the total SS, and so the unexplained is $1-r^2$ times the total SS . A similar relationship holds for the impulse response problem. The minimum MSE, playing the role of the unexplained variation (of Y_t around its best predictor $\sum_{s=-\infty}^{\infty} \beta_s X_{t-s}$) is (calculation on next page)

$$MSE = \int_{-1/2}^{1/2} f_Y(\nu) \left[1 - \rho_{YX}^2(\nu\right]) d\nu.$$

The reduction in MSE over merely using μ_Y to forecast Y_t is $\int_{-1/2}^{1/2} f_Y(\nu)\rho_{YX}^2(\nu)d\nu$, so that the method is most effective when applied to strongly coherent series.

$$
\begin{aligned}
MSE = \quad & E\left[\left\{Y_t - \sum_{s=-\infty}^{\infty} \beta_s X_{t-s}\right\}\left\{Y_t - \sum_{r=-\infty}^{\infty} \beta_r X_{t-r}\right\}\right. \\
= \quad & E\left[\left\{Y_t - \sum_{s=-\infty}^{\infty} \beta_s X_{t-s}\right\} Y_t\right] \text{ by (19.1)} \\
= \quad & \gamma_Y(0) - \sum_{s=-\infty}^{\infty} \beta_s \gamma_{YX}(s) \\
= \quad & \gamma_Y(0) - \sum_{s=-\infty}^{\infty} \int_{-1/2}^{1/2} B(\nu) e^{2\pi i \nu s} d\nu \cdot \gamma_{YX}(s) \\
= \quad & \gamma_Y(0) - \int_{-1/2}^{1/2} B(\nu) \sum_{s=-\infty}^{\infty} e^{2\pi i \nu s} \gamma_{YX}(s) d\nu \\
= \quad & \gamma_Y(0) - \int_{-1/2}^{1/2} B(\nu) \bar{f}_{YX}(\nu) d\nu, \\
= \quad & \int_{-1/2}^{1/2} f_Y(\nu) d\nu - \int_{-1/2}^{1/2} \frac{f_{YX}(\nu)}{f_X(\nu)} \bar{f}_{YX}(\nu) d\nu \\
= \quad & \int_{-1/2}^{1/2} f_Y(\nu) \left[1 - \frac{f_{YX}(\nu)\bar{f}_{YX}(\nu)}{f_X(\nu) f_Y(\nu)}\right] d\nu \\
= \quad & \int_{-1/2}^{1/2} f_Y(\nu) \left[1 - \rho^2_{YX}(\nu)\right] d\nu.
\end{aligned}
$$

20. Signal extraction; optimal filtering

- Filters. Recall that for the filter

$$Y_t = \sum_{s=-\infty}^{\infty} a_s X_{t-s}$$

we have $f_Y(\nu) = |A(\nu)|^2 f_X(\nu)$, where $A(\nu)$ is the IFT of $\{a_s\}$ - the "frequency response function". Note the distinction between this and the impulse-response problem. In that problem we were *estimating* the coefficients β_s after observing the series $\{X_t, Y_t = \sum_{s=-\infty}^{\infty} \beta_s X_{t-s} + v_t\}$. In the current setup we consider the problem of *choosing* $\{a_s\}$, and then *computing* $\{Y_t\}$ so that it has certain desirable properties.

  - Example: 3-point centred moving average ($a_{-1} = a_0 = a_1 = 1/3$, all others $= 0$) was low-pass, filtering out frequencies beyond about $1/3$. Figure 15.1.

  - Example: AR(2) filter ($a_1 = 1$, $a_2 = -1$, $a_3 = .9$) was band-pass. Figure 16.1.

– Example: $\nabla X_t = X_t - X_{t-1}$ - a filter with $a_0 = 1,\ a_1 = -1$, hence

$$\begin{aligned} A(\nu) &= \sum_{s=-\infty}^{\infty} a_s e^{-2\pi i \nu s} = 1 - e^{-2\pi i \nu}, \\ \bar{A}(\nu) &= 1 - e^{2\pi i \nu}, \\ |A(\nu)|^2 &= 2(1 - \cos(2\pi\nu)). \end{aligned}$$

See Figure 20.1. This is a 'high pass' filter.

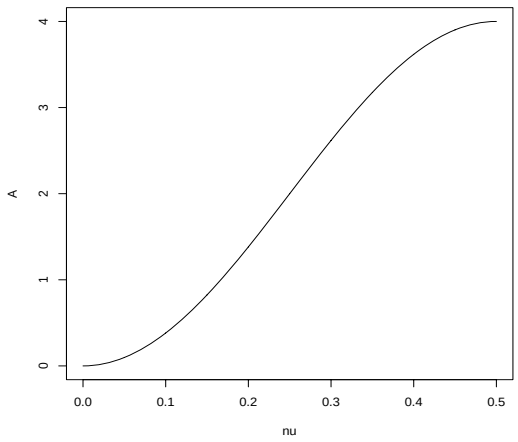

Figure 20.1. IFT ($|A(\nu)|^2$) for $\nabla X_t = X_t - X_{t-1}$ as a filter.

We estimate $f_X(\nu)$, determine those frequencies which we would like to highlight, and then choose $A(\nu)$ accordingly. We proceed as in the impulse-response problem. Assume that $A(\nu)$ is symmetric ($A(\nu) = A(-\nu)$), with a period of 1 (the R function will make it so). Then $\{a_s\}$ is recovered from

$$\begin{aligned} a_s &= \int_{-1/2}^{1/2} A(\nu) e^{2\pi i \nu s} d\nu = \int_0^1 A(\nu) e^{2\pi i \nu s} d\nu \\ &\approx \frac{1}{M} \sum_{k=0}^{M-1} A(\omega_k) e^{2\pi i \omega_k s} \stackrel{def}{=} a_s^M, \end{aligned}$$

for frequencies $\omega_k = k/M$ and M even. You should verify that the sequence $\left\{a_s^M\right\}$ is real and symmetric ($a_s^M = a_{-s}^M$). This is done for $|s| < M/2$. We write

$$A^M(\nu) = \sum_{|s|<M/2} a_s^M e^{-2\pi i \nu s}$$

for the IFT of $\left\{a_s^M\right\}$. Then the filtered series is

$$Y_t^M = \sum_{|s|<M/2} a_s^M X_{t-s} \qquad (20.1)$$

with spectrum

$$f_Y^M(\nu) = |A^M(\nu)|^2 f_X(\nu).$$

- In fact, the above formulas are modified a bit, in their implementation in the R program in astsa. The coefficients a_s^M used in (20.1) are (by default) replaced by

$$\tilde{a}_s^M = h_s a_s^M,$$

where the 'taper' $\{h_s\}$ will be discussed in the next class.

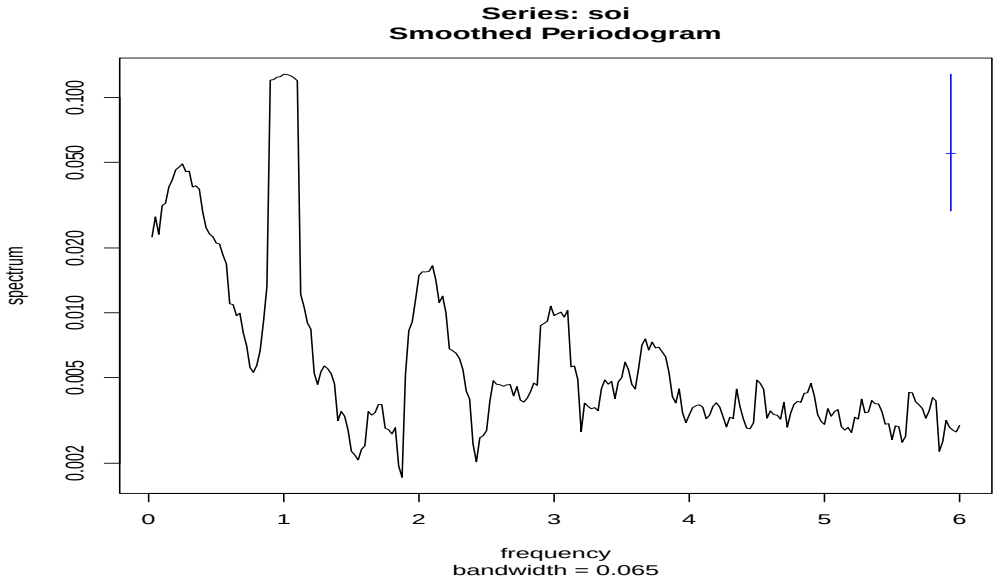

Figure 20.2. SOI periodogram.

Example: SOI series. Suppose one wants to study the El Niño signal. An examination of the periodogram in

Figure 20.2 ($L = 9$) reveals that to do this we might isolate frequencies in a band $\nu < .6/12 = .05$. Thus, set $A(\nu) = I(\nu < .05)$ and compute $\tilde{a}_s^M$ as above ($M = 64$) . This is done via

```
out = SigExtract(soi, L=9, M=64, max.freq=.05);
```

this returns the filtered series and graphical output: plots of the $\left\{\tilde{a}_s^M\right\}$ and of

$$\tilde{A}^M(\nu) = \sum_{|s|<M/2} \tilde{a}_s^M e^{-2\pi i \nu s},$$

as well as the spectra and values of the original and filtered series.

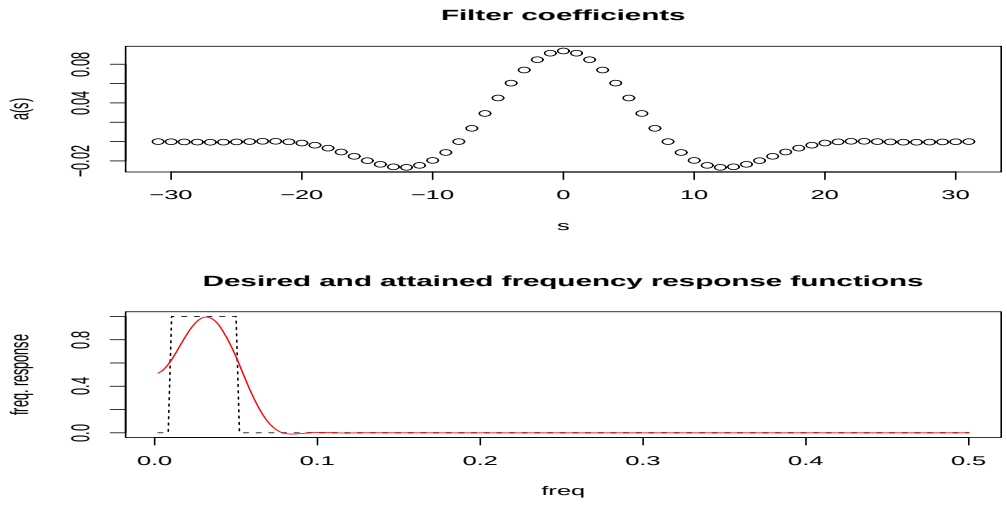

Figure 20.3. Filter coefficents $\left\{\tilde{a}_s^M\right\}$; desired and attained frequency responses $A(\nu)$ and $\tilde{A}^M(\nu)$.

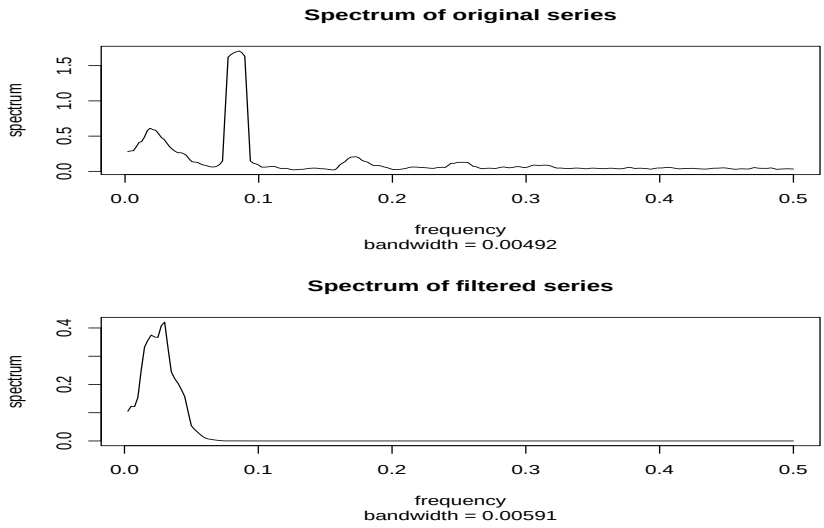

Figure 20.4. Spectra of original and filtered SOI series.

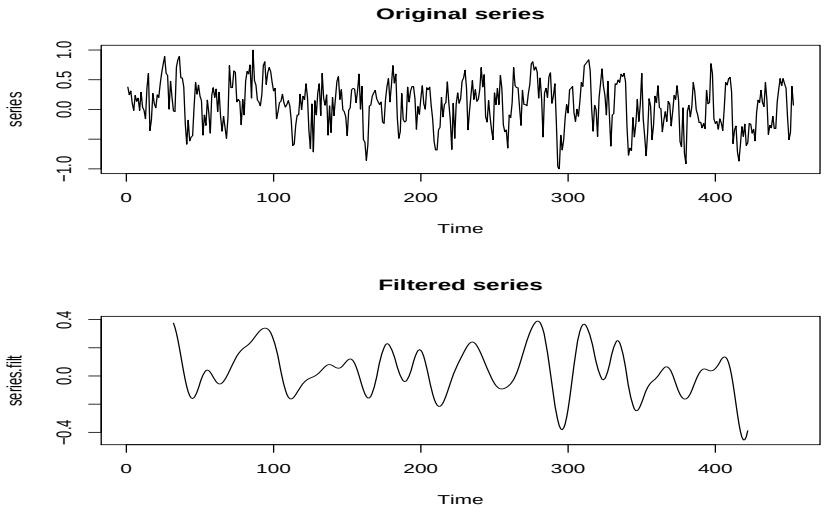

Figure 20.5. Original SOI series $\{X_t\}$ and filtered series $Y_t^M = \sum_{|s|<M/2} \tilde{a}_s^M X_{t-s}$

- How does the choice of M affect things? To see, first let $\{a_t\}$ be the "correct" filter coefficients:

$$a_t = \int_{-1/2}^{1/2} A(\nu) e^{2\pi i \nu t} d\nu.$$

We are approximating these by

$$a_s^M = \frac{1}{M} \sum_{k=0}^{M-1} A(\omega_k) e^{2\pi i \omega_k s};$$

since $A(\omega_k) = \sum_{t=-\infty}^{\infty} a_t e^{-2\pi i \omega_k t}$ the approximations are

$$\begin{aligned} a_s^M &= \frac{1}{M} \sum_{k=0}^{M-1} \left\{ \sum_{t=-\infty}^{\infty} a_t e^{-2\pi i \omega_k t} \right\} e^{2\pi i \omega_k s} \\ &= \frac{1}{M} \sum_{t=-\infty}^{\infty} a_t \left\{ \sum_{k=0}^{M-1} e^{-2\pi i \omega_k (t-s)} \right\}. \end{aligned}$$

The inner sum equals M when $(t-s)/M$ is an integer, and otherwise is zero, and so

$$\begin{aligned} a_s^M &= \sum_{t=-\infty}^{\infty} a_t \left[I(t - s = lM \text{ for an integer } l) \right] \\ &= a_s + \sum_{l \neq 0} a_{s+lM}. \end{aligned}$$

- How large should M be? Ideally, we would like to arrange things so that, for each s with $|s| < M/2$, we have

  $$a_{s+lM} = 0 \text{ if } l \neq 0.$$

  This will hold (approximately) if M is so large that

  $$a_s \approx 0 \text{ for } |s| \geq M/2. \tag{20.2}$$

  Why? Because $|s + lM| \geq M/2$ if $|s| < M/2$ and $l \neq 0$.
  Then our filter

  $$Y_t^M = \sum_{|s|<M/2} a_s^M X_{t-s}$$

  and the desired filter

  $$Y_t = \sum_{s=-\infty}^{\infty} a_s X_{t-s} = \sum_{|s|<M/2} a_s X_{t-s}$$

  will agree (approximately).

- Note that since $\sum_{t=-\infty}^{\infty} |a_t| < \infty$, (20.2) must eventually hold.

- This suggests choosing M as large as possible, so that (20.2) will hold.

- Values of M which are too small result in error messages, due to numerical problems.

- However, recall now that $\{X_t\}$ is being replaced by the data $\{x_t\}_{t=1}^n$. As M increases the range of t, for which $y_t^M = \sum_{|s|<M/2} a_s^M x_{t-s}$ can be computed, shrinks - the first and last $M/2 - 1$ values of the filtered series cannot be computed.

- Thus a balance must be struck - choose M as large as you can, without losing too many values from the filtered series.

21. Special topics

- **Recursive filtering**. An alternate approach to filtering is to construct a *recursive* filter, in which the filter depends on its own previous values as well as on the series being filtered.

  - Example: Given a series $\{X_t\}$, set $Y_0 = 0$, $Y_t = \phi Y_{t-1} + X_t$ ($|\phi| < 1$). Iterating this gives

    $$\begin{aligned} Y_t &= \phi^{t-1} X_1 + \phi^{t-2} X_2 + ... + \phi X_{t-1} + X_t \\ &= \sum_{s=0}^{t-1} \phi^s X_{t-s}. \end{aligned}$$

    This is called *exponential smoothing*. Note that, unlike our earlier filters, it is not of the form $Y_t = \sum_{s=-\infty}^{\infty} a_s X_{t-s}$:

    $$a_s = \phi^s I(0 \leq s \leq t-1)$$

    depends on t as well as s.

  - Example: Recall the AR(2), band-pass filter plotted in Figure 16.1.

– More generally, we could take the filter

$$Y_t = \sum_{k=1}^{p} \phi_k Y_{t-k} + X_t + \sum_{j=1}^{q} \theta_j X_{t-j}, \tag{21.1}$$

or

$$\phi(B)Y_t = \theta(B)X_t$$

in an obvious notation. If we assume that all values $\{X_t\}_{t=-\infty}^{\infty}$ are available (so that (21.1) defines a series $\{Y_t\}_{t=-\infty}^{\infty}$), this can be analyzed using our usual methods. Put first $Z_t = \phi(B)Y_t$, then $= \theta(B)X_t$ and obtain

$$f_Z(\nu) = |\phi(e^{-2\pi i\nu})|^2 f_Y(\nu) = |\theta(e^{-2\pi i\nu})|^2 f_X(\nu),$$

whence

$$f_Y(\nu) = \frac{|\theta(e^{-2\pi i\nu})|^2}{|\phi(e^{-2\pi i\nu})|^2} f_X(\nu).$$

– In the case

$$\begin{aligned} \phi(B) &= 1 - \phi B, \\ \theta(B) &\equiv 1, \end{aligned}$$

(exponential smoothing) $f_Y(\nu)$ is being derived and plotted for $\phi = .1, .8$ in Assignment 3.

- **Parametric spectral estimation**. If $\{X_t\}$ is AR(p), with characteristic polynomial $\phi(B) = 1 - \sum_{s=1}^{p} \phi_s B^s$, then the spectrum is

$$f_X(\nu) = \frac{\sigma_w^2}{|\phi(e^{-2\pi i \nu})|^2}.$$

This can be estimated by fitting the AR model to the time domain data, estimating the AR coefficients and the noise variance, and plugging these in:

$$\hat{f}_X(\nu) = \frac{\hat{\sigma}_w^2}{|\hat{\phi}(e^{-2\pi i \nu})|^2}, \tag{21.2}$$

where $\hat{\phi}(z) = 1 - \sum_{s=1}^{p} \hat{\phi}_s z^s$.

  - One can also do this with MA and ARMA processes; interest is however focussed on the AR case because of the following: If $g(\nu)$ is the spectral density of a stationary process, and ε is any (small) positive number, then there is a stationary AR(p) process, with spectrum $f_X(\nu)$ satisfying

$$|f_X(\nu) - g(\nu)| < \varepsilon, \text{ for all } \nu \in [-1/2, 1/2].$$

Of course choosing a smaller ε entails requiring a larger p. A drawback is that even the asymptotic properties of this estimate are more complicated than those of the periodogram.

– Example - SOI series (`sunspots` series – Ass't 3). Basic command is `spaic = spec.ar(soi, log="??")`. This fits a series of models, chooses the best using AIC (see the R code on course web site for AIC plots), then computes (21.2).

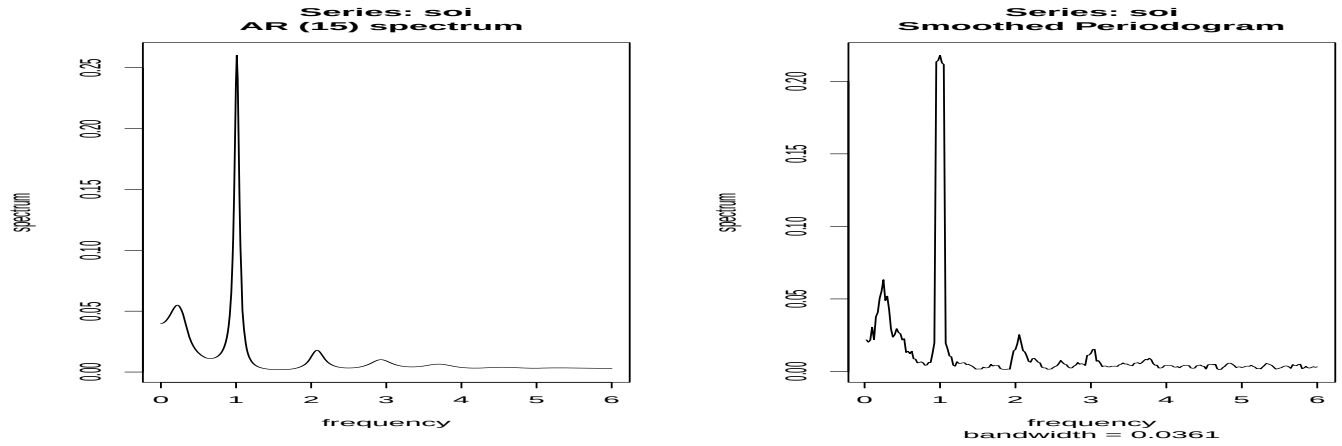

Figure 21.1. Left: Parametric estimate of SOI spectrum; log = "no". The default is "yes". Right: Smoothed ($L = 5$) periodogram estimate.

- **Tapering and windows**. (Not on exam, but ...) We have used the smoothed periodogram (which I will write here as $\hat{f}_L(\nu_k)$), which is an average of the values of $|X(k+l)|^2$ for values of l near zero: $|l| \leq (L-1)/2$. Rather than smoothing in this way, an alternate approach is to first smooth the data, and to then compute the raw periodogram of the smoothed data. Or, one can do both. First, one replaces x_t by $\tilde{x}_t = x_t h_t$; the function $\{h_t\}$ is called the *taper*. Then one computes the DFT

$$\tilde{X}(k) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \tilde{x}_t e^{-2\pi i \nu_k t},$$

and spectral estimate

$$\tilde{f}(\nu_k) = \left|\tilde{X}(k)\right|^2.$$

The unsmoothed periodogram $\hat{f}_1(\nu_k)$ corresponds to $h_t \equiv 1$. Typically the taper is chosen to decrease as t moves away from the midpoint $\bar{t} = (n+1)/2$ of $\{1, 2, ..., n\}$. To see the effect that a given taper will have, define

$$H(\omega) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} h_t e^{-2\pi i \omega t} \text{ and } W(\omega) = |H(\omega)|^2.$$

These are the DFT and its squared modulus, derived from $\{h_t\}$. A calculation yields (an improvement on (17.3)):

$$E\left[\tilde{f}\left(\nu_k\right)\right] = \int_{-1/2}^{1/2} W\left(\nu_k - \nu\right) f_X\left(\nu\right) d\nu. \tag{21.3}$$

- As is evident from Figure 21.2, the "window" $W\left(\omega\right)$ determines how much of the spectral density $f_X\left(\omega\right)$ is "seen" in the computation of the periodogram $\left|\tilde{X}\left(k\right)\right|^2$.

- Example 1. If $h_t \equiv 1$ (the unsmoothed periodogram) then $W\left(\omega\right) = \sin^2\left(\pi n \omega\right) / \left(n \sin^2\left(\pi\omega\right)\right)$.

- Example 2. By default, R applies a 'cosine bell' taper, with

  $$h_t = .5\left[1 + \cos\left(\frac{2\pi\left(t - \bar{t}\right)}{n}\right)\right];$$

  this is applied to the first and last 10% of the series (if one keeps the default "`taper=.1`"). Then periodogram smoothing is done as well, if $L > 1$.

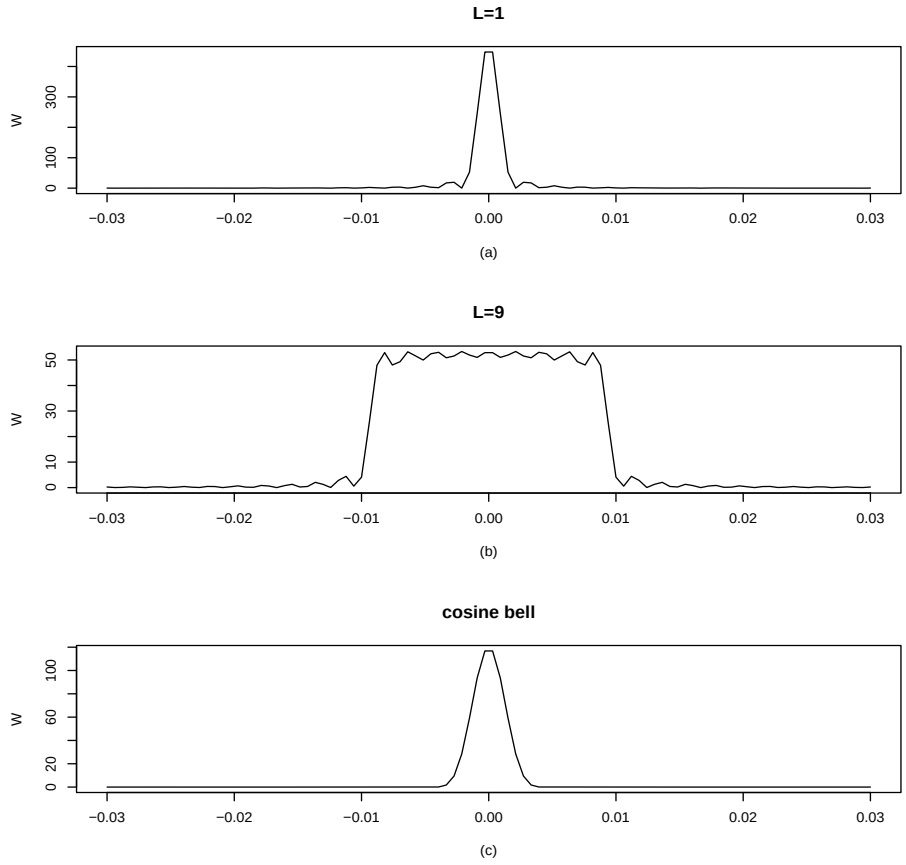

Figure 21.2. Spectral windows $W(\omega)$, $n = 480$. (a) No smoothing (Example 1). (b) Window corresponding to smoothing; $L = 9$. (c) Cosine bell taper (Example 2).