STATISTICS 679
TIME SERIES ANALYSIS

Doug Wiens*

March 16, 2015

Contents

1 Introduction 5

2 Fitting ARIMA models 19

3 Spectral Theory 39

4 Examples and applications: impulse-response; optimal filtering 52

5 Long memory models 76

6 GARCH models; threshold models 84

7 Regression with autocorrelated errors; transfer function modelling 96

8 Multivariate regression and ARMAX 110

9 Multivariate ARMAX models II 124

10 State-space models - Introduction 136

11 Filtering, Smoothing, Forecasting 140

12 Bayes Iterations; Computing 147

13 Estimation 161

14 Structural models; ARMAX models in state-space form 174

15 Bootstrapping state-space models; nonlinearity and non-normality 190

16 Analysis of longitudinal data 207

17 Multivariate frequency domain methods - Introduction 220

18 Spectral likelihood and frequency domain regression 227

19 Regression for jointly stationary series . . . 241

20 Regression with deterministic inputs 258

21 Random coefficient regression 271

22 Analysis of designed experiments 277

23 Principal Component Analysis 288

24 Discrimination 299

25 Clustering 319

1. Introduction

- A time series $\{X_t\}_{t=1}^n$ is a sequence of random variables observed over time ("t"). It 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 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*.

- The *autocorrelation function* is $\rho(m) = CORR[X_s, X_{s+m}]$, i.e.

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

- 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 . The *cross-covariance function* is

$$\gamma_{XY}(m) = COV[X_{t+m}, Y_t] = \gamma_{YX}(-m).$$

- Cross-correlation function (CCF):

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

- Example. SOI (Southern Oscillation Index - measures changes in air pressure related to sea surface temperatures) and Recruits (numbers of new fish).

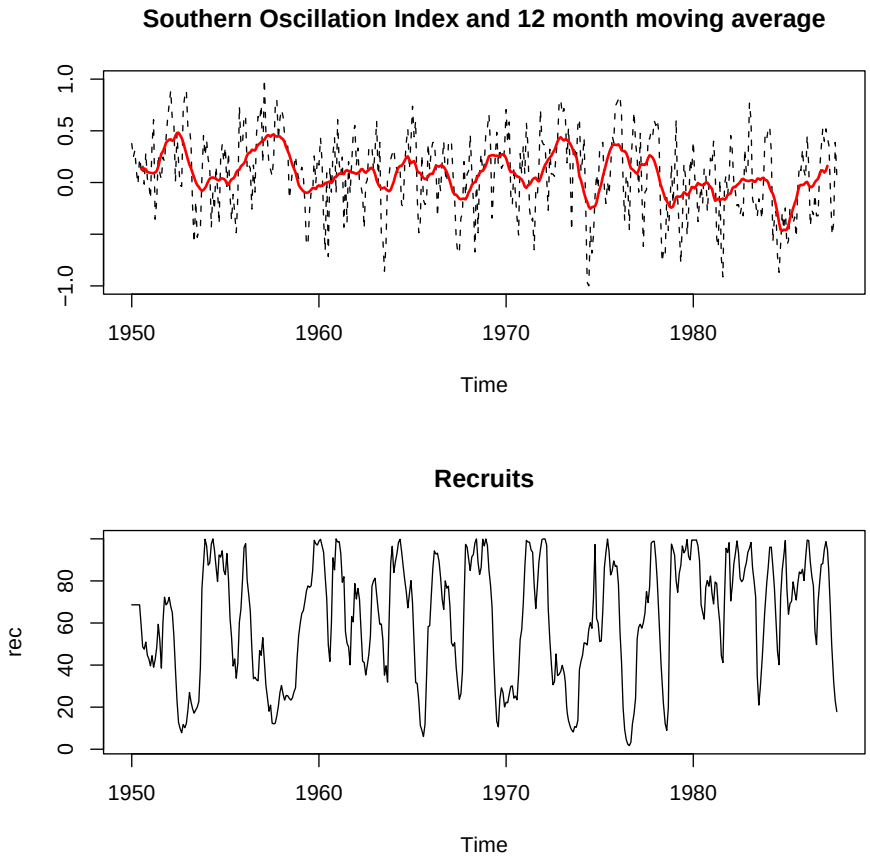

Figure 1.1. SOI, with 12-month moving average (what does this show?), and Recruits.

- 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 $m = 3, ..., 12$.

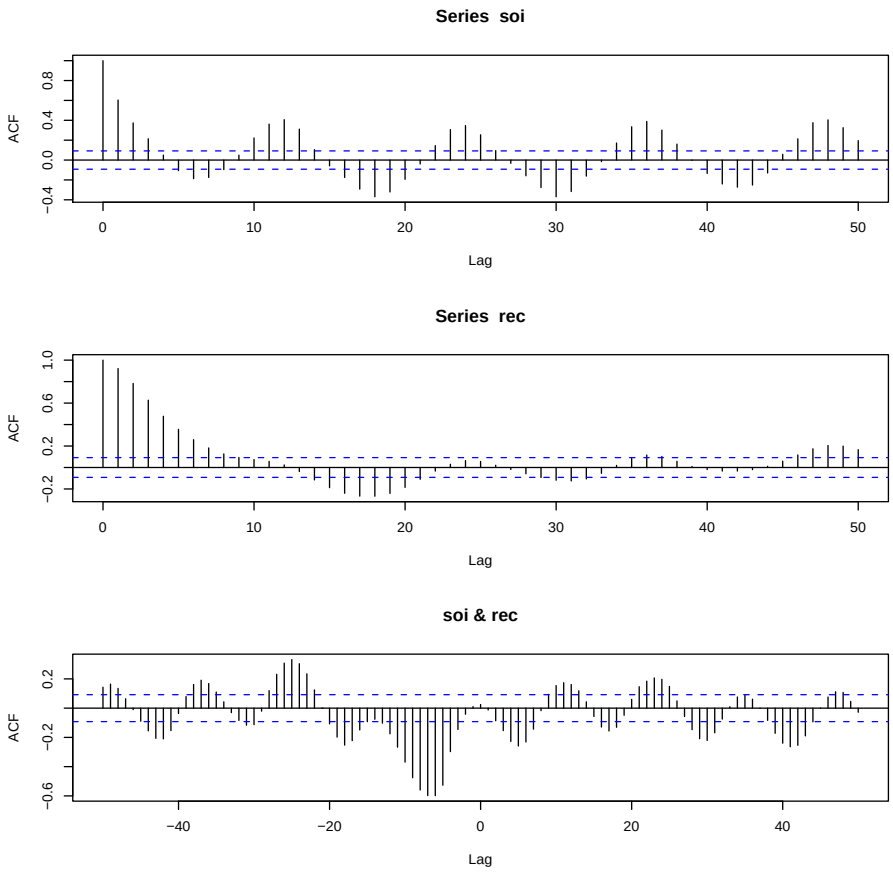

Figure 1.2. ACFs and CCF (`ccf(soi,rec)`$= COV\left[SOI_{t+m}, REC_t\right]$).

- Here we 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; it fits the model

$$Y_t = \beta_1 + \beta_2 t + \sum_{m=3}^{12} \beta_m \dot{X}_{t-m} + \text{error.}$$

with $\dot{X} = X - \bar{X}$. 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 ~trend + L(x,3) + L(x,4)+ L(x,5)
 + L(x,6) + L(x,7) + L(x,8)+ L(x,9) + L(x,10)
 + L(x,11) + L(x,12) )
# Or merely "fit = dynlm(rec ~trend + L(x,3:12))"
summary(fit)
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 d.f.,
    p-value: < 2.2e-16
```

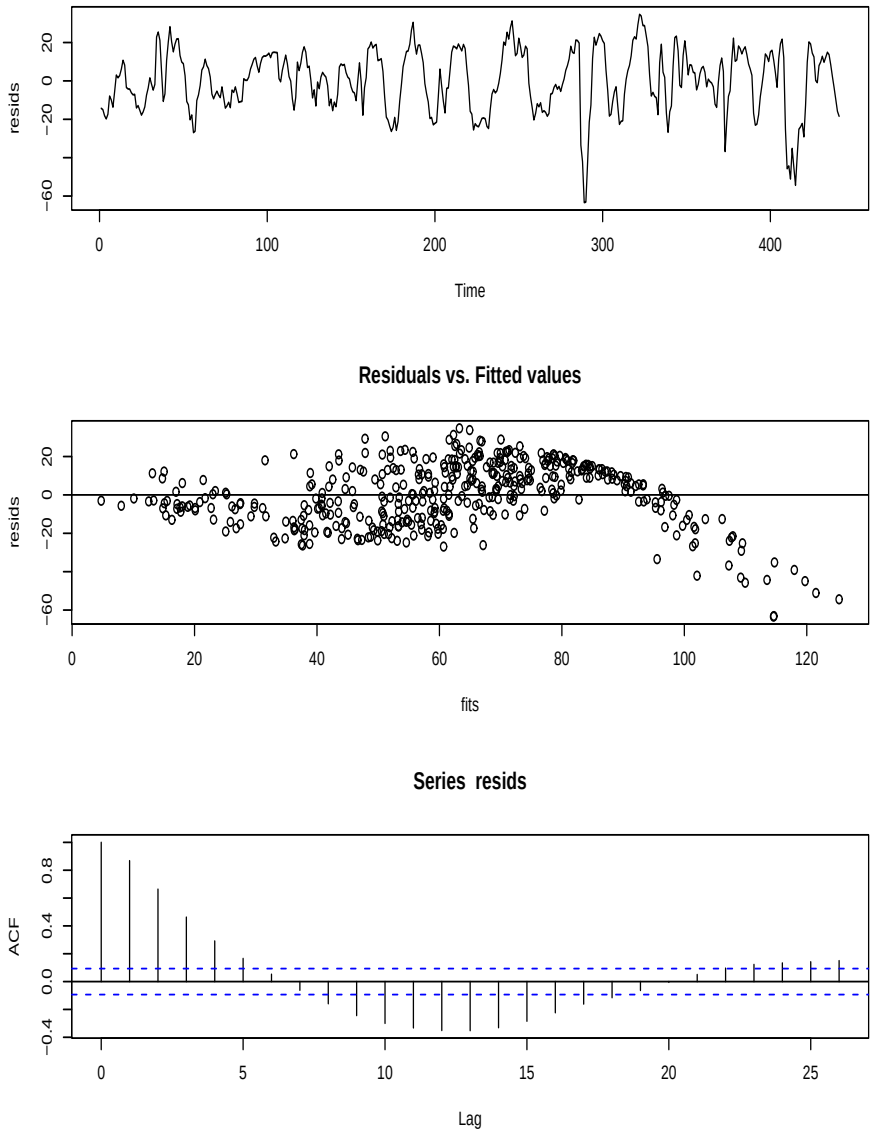

Figure 1.3. Residuals plots from regression of Recruits on time and on lagged SOI. A significant amount of structure remains to be explained.

1.1. Stationarity & Invertibility via Wold's Theorem

- Suppose $\{X_t\}$ is (weakly) stationary. Write $X_t - \mu$ as $\dot{X}_t$. Then (*Wold's Representation Theorem*) we can represent $\dot{X}_t$ as

$$\begin{aligned} \dot{X}_t &= w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + ... + V_t \\ &= \sum_{k=0}^{\infty} \theta_k w_{t-k} + V_t \ (\theta_0 \stackrel{def}{=} 1) \qquad (1.1) \end{aligned}$$

$$\text{and } \sum_{k=1}^{\infty} \theta_k^2 < \infty.$$

Here $\{w_t\}$ is white noise and V_t is "deterministic" - it can be predicted exactly from its own past. We will assume that $\{X_t\}$ is "purely non-deterministic", i.e. that $V_t = 0$.

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 (1.1) holds with $V_t = 0$ 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 as well; this is a simple calculation.

- 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 not really a polynomial, but if it is, i.e. $\theta_k = 0$ for $k > q$, 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 the past history of the process. 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 $\phi_j = 0$ for $j > p$, we say $\{X_t\}$ is an *autoregressive* process of order p, i.e. AR(p). Then

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

- A stationary process is invertible iff all roots of $\theta(B) = 0$ lie outside the unit circle. 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.

- 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 $\psi(z) = \sum_k \psi_k z^k$, $|z| \leq 1$, with square summable coefficients. Then the coefficients θ_k are determined from $\theta(z)/\phi(z) = \psi(z)$. It can be shown that $\psi(z)$ has the required properties only if all zeros of $\phi(z)$ lie outside the unit circle. Similarly an ARMA(p,q) is invertible only if all zeros of $\theta(z)$ lie outside the unit circle. We also require that the polynomials have no common factors.

- A class of nonstationary models is obtained by taking differences, and assuming that the differenced series is 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. The differenced series $\left\{\nabla^d X_t\right\}$ is the one we analyze.

- 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

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

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

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

- Fitting an appropriate model to data begins by looking at the sample ACF (and PACF); these will be discussed next.

2. Fitting ARIMA models

- For an MA(q), the autocovariance function is

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

  Reason: For any stationary process represented as a linear one,

$$\begin{aligned} \gamma(m) &= \text{cov}\left[X_t, X_{t+m}\right] \\ &= \text{cov}\left[\sum_{k=0}^{\infty} \theta_k w_{t-k}, \sum_{l=0}^{\infty} \theta_l w_{t+m-l}\right] \\ &= \sum_{k,l} \theta_k \theta_l \text{cov}\left[w_{t-k}, w_{t+m-l}\right] \\ &= \sum_{k,l} \theta_k \theta_l \sigma_w^2 I\left(t-k = t+m-l\right) \\ &= \sigma_w^2 \sum_{k=0}^{\infty} \theta_k \theta_{k+m}; \end{aligned}$$

  for an MA(q) the last non-zero term corresponds to $k + m = q$.

- The salient feature is that $\gamma(m) = 0$ for $m > q$;

we look for this in the sample ACF:

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

for $m \geq 0$; $\hat{\gamma}(-m) = \hat{\gamma}(m)$; $\hat{\rho}(m) = \hat{\gamma}(m)/\hat{\gamma}(0)$.

- Difficult to identify an AR(p) from its ACF. Assume $\mu_X = 0$; consider the problem of minimizing the function

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

which is the MSE when X_t is forecast by

$$\alpha_{1,m}X_{t-1} + ... + \alpha_{m,m}X_{t-m}.$$

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

$$\phi_{11} = \rho(1).$$

- 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 $j > p$.

- The sample PACF can be obtained by expressing the minimizers in terms of the autocovariances, and then replacing these by their sample estimates. Equivalently, replace the expectation in the definition of f_m by a sample average before doing the minimization.

2.1. Forecasting

- From now on we assume, unless stated otherwise, that the white noise is Gaussian, hence w_t and w_s are independent, rather than merely uncorrelated, if $s \neq t$.

- 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}. The "best" (minimum MSE) 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 $X^t = \{X_s\}^t_{s=-\infty}$.

- 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 $X^t = \{X_s\}^t_{s=-\infty}$. 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 conditioning on X^t is equivalent to conditioning on $w^t = \{w_s\}^t_{s=-\infty}$:

$$\begin{aligned} X^t_{t+l} &= \sum_{k=0}^{\infty} \psi_k E[w_{t+l-k}|w^t] \text{ where} \\ E[w_{t+l-k}|w^t] &= \begin{cases} w_{t+l-k}, & \text{if } l \le k, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

(Note that $E\left[X_{t+l-k}|X^t\right] = X_{t+l-k}$ if $l \le k$.) Thus the forecast is

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

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,

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

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

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

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

- In practice, X_{t+l}^t can often be obtained more directly. Example 1 AR(1): $X_{t+l}^t = \phi^l x_t$, but obtaining the forecast intervals requires writing

the model in linear form. Example 2 MA(1): $\psi_0 = 1$, $\psi_1 = \theta$, $\psi_k = 0$ for $k > 1$; obtaining $X_{t+l}^t = \theta w_t I(l = 1)$ necessitates inverting the model.

- After writing the forecasts in terms of the data and the AR and MA parameters of $\{X_t\}$, we substitute estimates of these parameters, thus obtaining $\hat{X}_{t+l}^t$. Similarly with the interval boundaries, where we also must use an estimate $\hat{\sigma}_w^2$. The *residuals* are

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

and the (adjusted) MLE of σ_w^2 is

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

An easy calculation shows that $\hat{w}_t = \hat{\phi}(B)X_t$, where $w_t = \phi(B)X_t$ is the inverted form of X_t. In other words, the residual can be obtained by writing the white noise in terms of X_t (often recursively) and then estimating the coefficients.

2.2. Estimation

- **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. When evaluated at the numerical data this is a function of $\boldsymbol{\alpha}$ 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*

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

- With $\boldsymbol{\alpha}_0$ denoting the true value, we typically have that $\sqrt{n}\,(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)$ has a limiting (as $n \to \infty$) normal distribution, with mean $\mathbf{0}$ and *covariance matrix*

$$\mathbf{C} = \mathbf{I}^{-1}(\boldsymbol{\alpha}_0)$$

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

- The information matrix is given by

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

a more convenient and equivalent form (for the kinds of models we will be working with) is

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

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

- 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 $\sqrt{n}\left(\hat{\alpha}_j - \alpha_j\right)$ is asymptotically normally distributed with mean zero and variance estimated by $\hat{\mathbf{I}}^{jj}$, so that

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

- For a detailed example (3 parameter AR(1) model with a Normal likelihood) see the STAT 479 notes.

2.3. Example - Varve series

- Box-Jenkins methodology:

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

  2. Estimate the parameters. This allows us to estimate the forecasts X_{t+l}^t, which depend on unknown parameters, by substituting estimates to obtain $\hat{X}_{t+l}^t$.

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

- **Example 3.31**. Apply all of this in the 'varve' series - a set of thicknesses of glacial sedimentary deposits which can be used as indicators of temperature. We work with $Y_t = \nabla \ln(X_t)$ (interpretation?), which appears to be stationary.

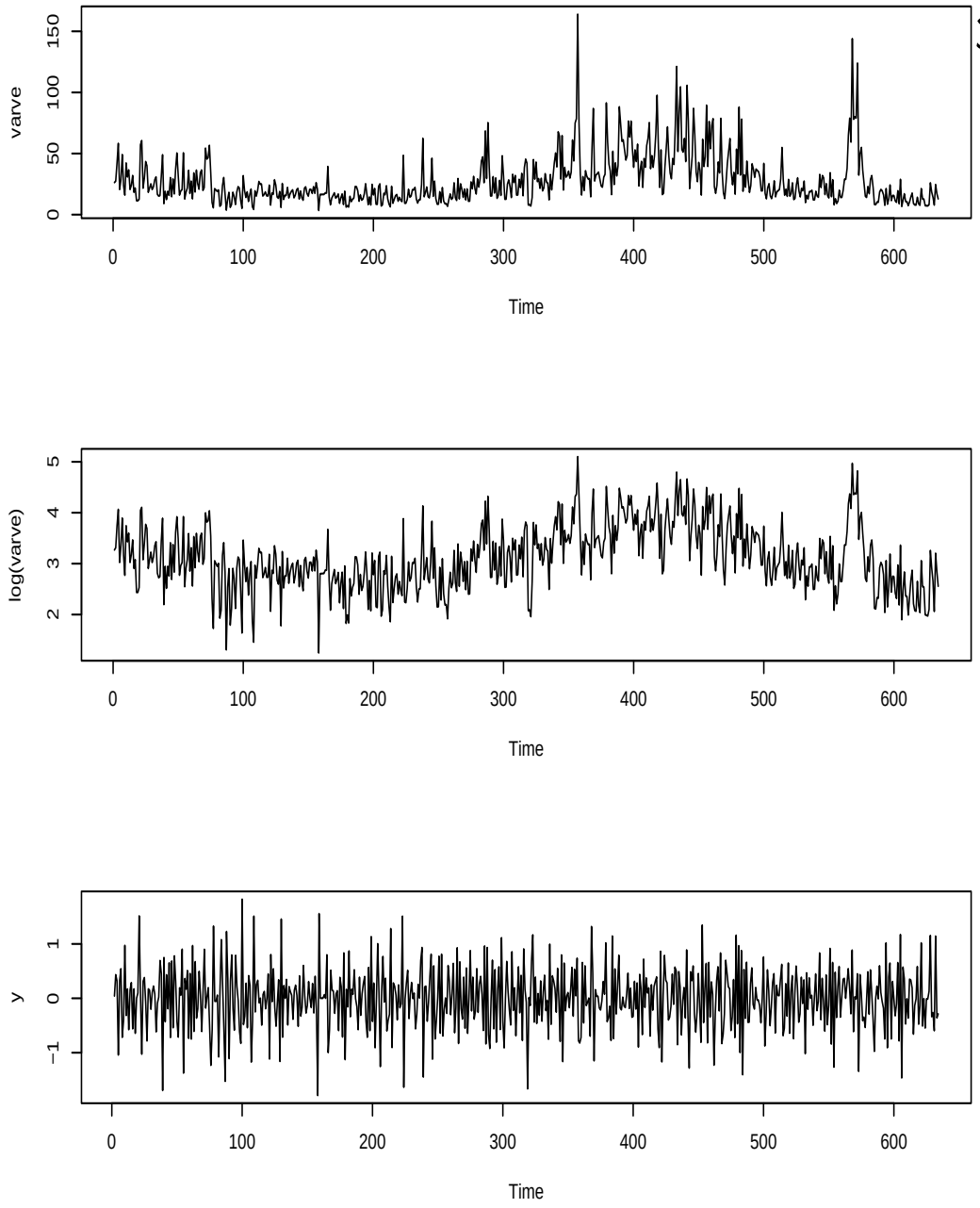

Figure 2.1. Varve series: $X_t, \log(X_t), \nabla \log(X_t)$.

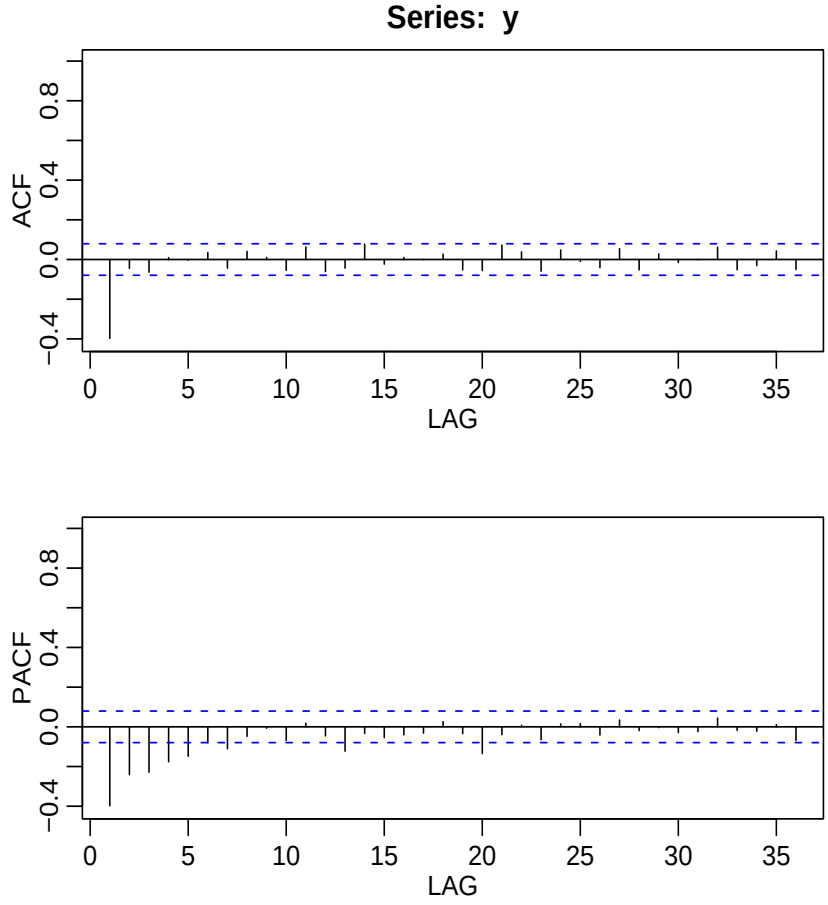

Figure 2.2. ACF and PACF of Y_t. ACF suggests MA(1); PACF suggests AR.

- Start with MA(1): `fit1.y = sarima(data = y, p = 0, d = 0, q = 1):`

```
Coefficients:
          ma1     xmean
      -0.7710  -0.0013
s.e.   0.0341   0.0044
sigma^2 estimated as 0.2353:
log likelihood = -440.68,  aic = 887.36
```

- AIC - one of several Model Selection values. In general, they measure the fit of a proposed model against the number of parameters being fitted; smaller values indicate a superior fit, even after discounting for the increased complexity of the model. All are of the form

$$AIC = \ln \hat{\sigma}_k^2 + \text{penalty for using } k \text{ terms},$$

where k is the number of parameters in the model and $\hat{\sigma}_k^2$ is the estimate of residual variation.

From the R help files:
Generic function calculating the Akaike information criterion for one or several fitted model objects for which a log-likelihood value can be obtained, according to the formula -2*log-likelihood + k*npar, where npar represents the number of parameters in the fitted model, and k = 2 for the usual AIC, or k = log(n)(n the number of observations) for the so-called BIC or SBC (Schwarz's Bayesian criterion).

From the S&S code:

```
BIC=log(fitit$sigma2)+(k*log(n)/n)
AICc=log(fitit$sigma2)+((n+k)/(n-k-2))
AIC=log(fitit$sigma2)+((n+2*k)/n)
```

This gives
AIC = -0.4406366,
AICc = -0.4374168,
BIC = -1.426575.

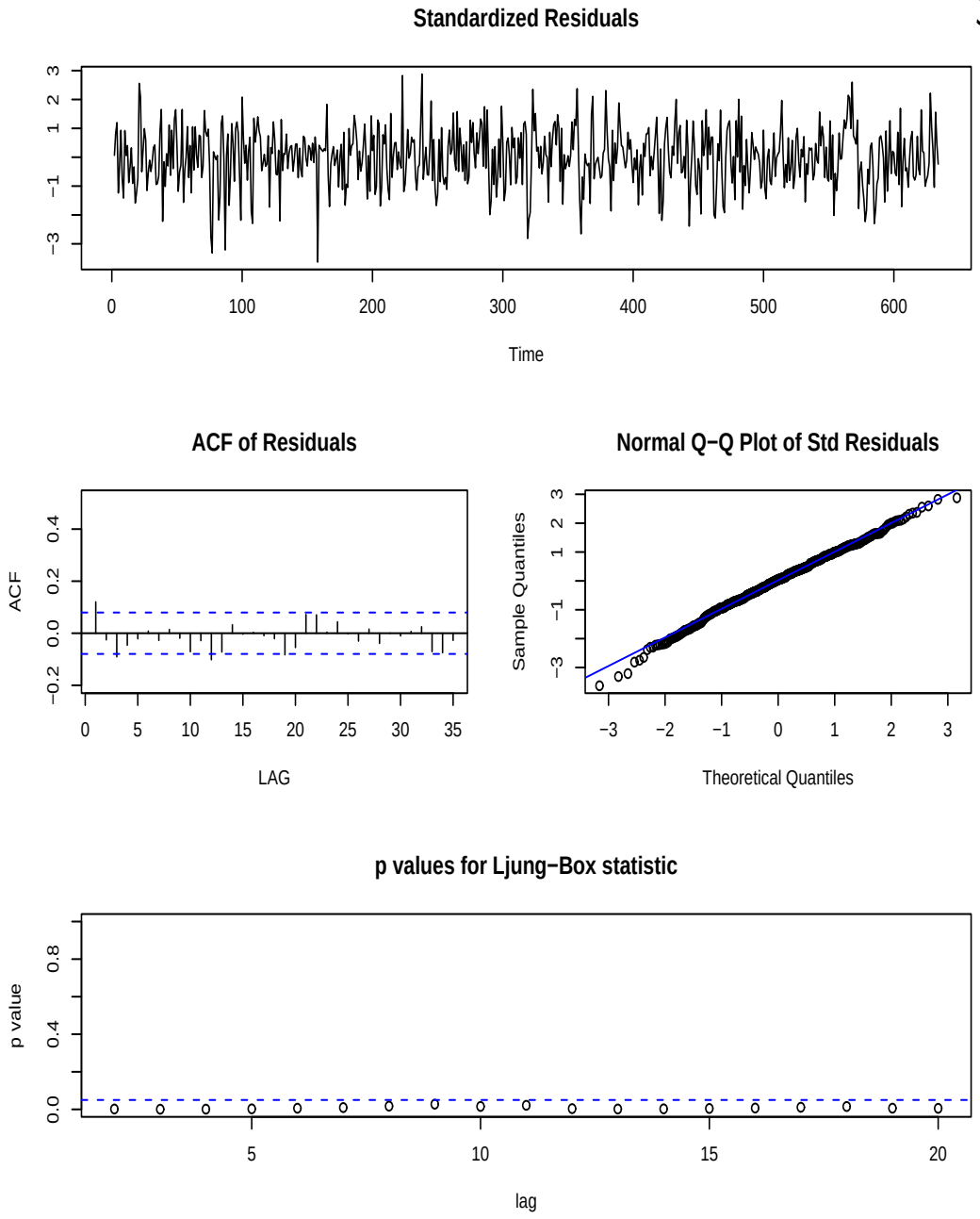

Figure 2.3. MA(1) fit and residual plots.

- **Ljung-Box-Pierce test**. Under the hypothesis of whiteness we expect $\hat{\rho}_w(h)$ to be small in absolute

value for all m; a test can be based on

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

which is approximately $\sim \chi^2_{n-H}$ under the null hypothesis. In R, the p-value is calculated and reported for $H = 1, ..., 20$.

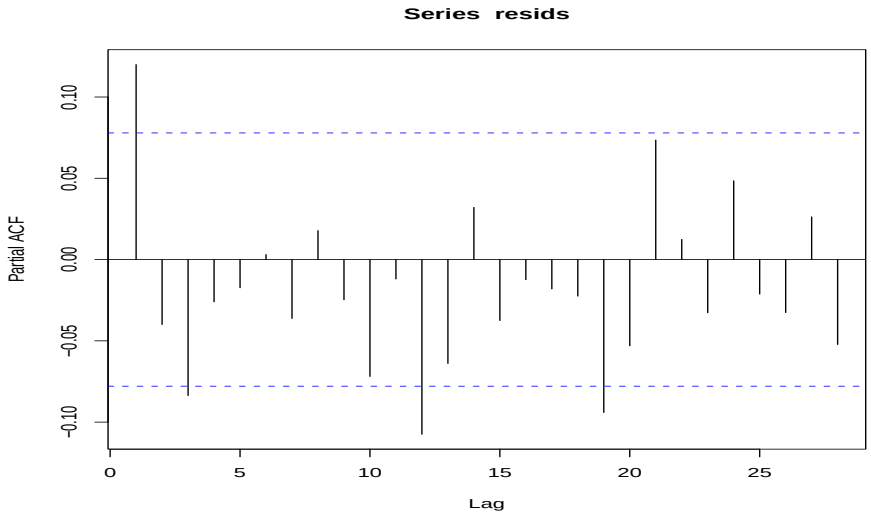

Figure 2.4. Sample pacf of residuals from MA(1) fit.

- Add an AR(1) component:

```
Coefficients:
         ar1      ma1    xmean
      0.2341  -0.8871  -0.0013
s.e.  0.0518   0.0292   0.0028

sigma^2 estimated as 0.2284:
log likelihood = -431.33,  aic = 870.66

$AIC = -0.467376
$AICc = -0.4641159
$BIC = -1.446284
```

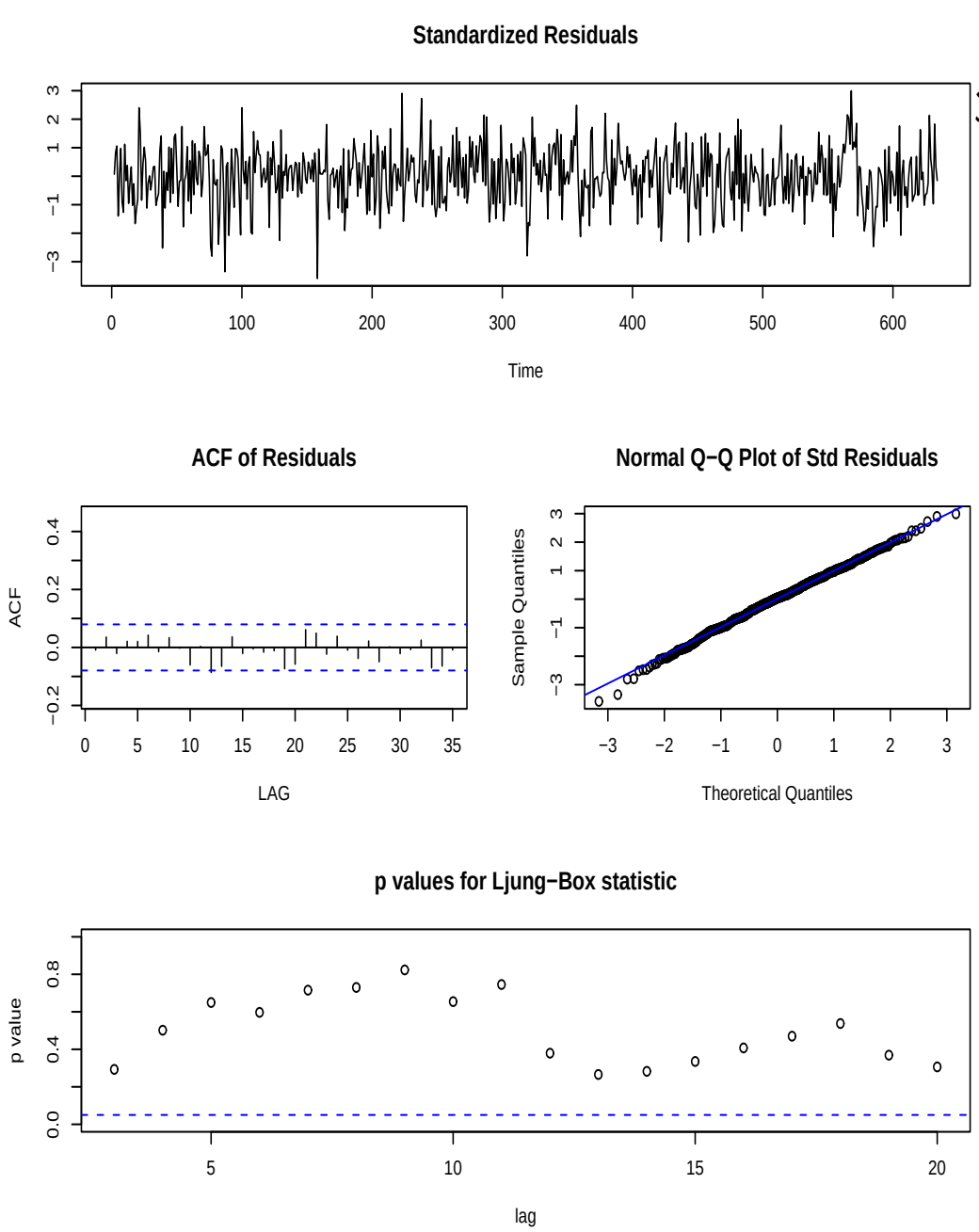

Figure 2.5. ARMA(1,1) fit.

- Forecast the integrated series: `forecasts.varve = sarima.for(log(varve), nahead = 100, p=1, d=1, q=1)`.

Prediction ± 2 $s.e.$ is plotted.

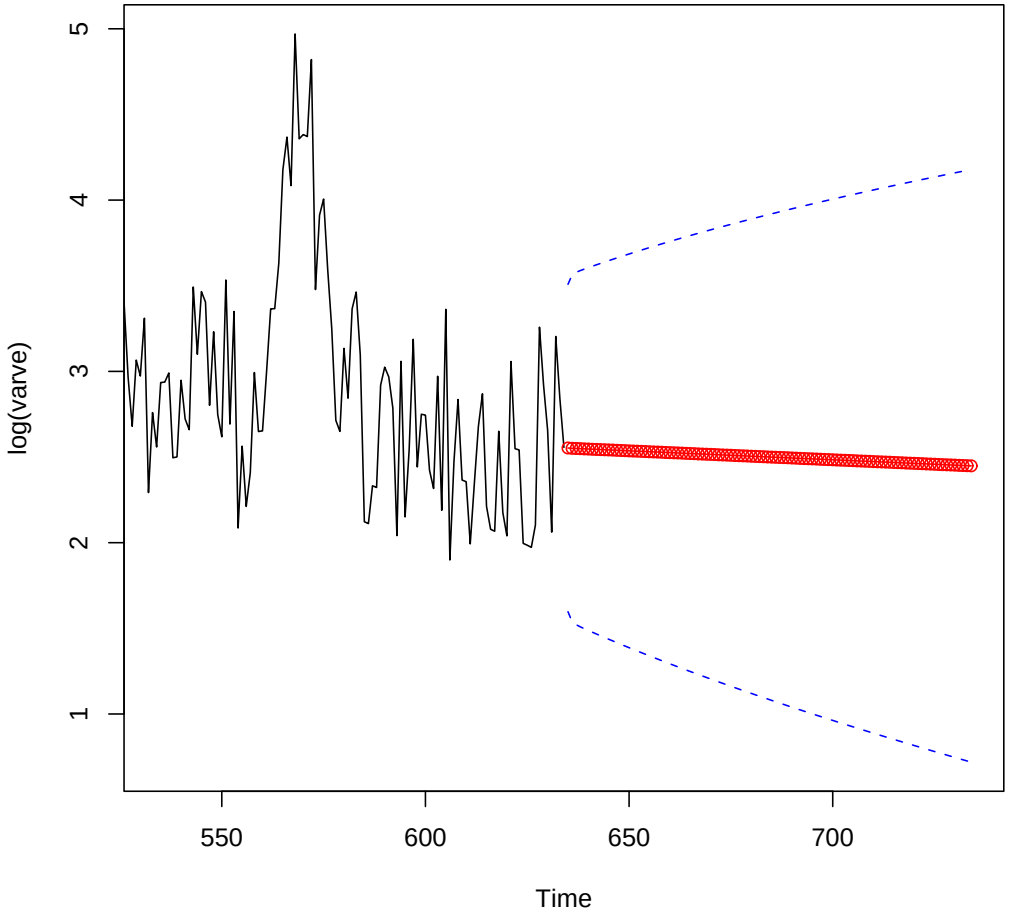

Figure 2.6. Forecasts.

3. Spectral Theory

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

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

  The "total power" is

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

- Motivation: Given any set $x_1, ..., x_n$ of real numbers, we can find "amplitudes" $\{a_k, b_k\}$ for which

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

  with $\nu_k = k/n$ (thus $0 \leq \nu_k \leq 1/2$). More generally, 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, zero-mean r.v.s $\{A_k, B_k\}$ with variances σ_k^2, and "Fourier frequencies" $\lambda_k \in [0, \pi]$. The approximation becomes exact as $N \to \infty$, in the sense of convergence in mean square. For a series as on the rhs above, the variance is $\sigma^2 = \gamma(0) = \sum_{k=0}^{N} \sigma_k^2$ and the ACF is

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

We write this ACF as

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

where the discrete, symmetric r.v. Λ has distribution

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

As $N \to \infty$, the distribution of Λ tends to one with a density (assuming that the ACF is absolutely summable). Now we write $\lambda = 2\pi\nu$, $-1/2 \leq \nu \leq 1/2$, and write the density of ν in the form $f(\nu)/\sigma^2$. The relationship above becomes

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

thus (using the symmetry of f)

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

and so

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

This relationship is invertible :

$$f(\nu) = \sum_{m=-\infty}^{\infty} e^{-2\pi i \nu m} \gamma(m). \tag{3.1}$$

We say that f and γ are "Fourier transform pairs"; we shall refer to f as the Infinite Fourier Transform of γ. Note from (3.1) (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 studied on this interval.

- More generally, 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: $A(\nu) \equiv B(\nu) \Leftrightarrow \{a_t\} \equiv \{b_t\}$.

- 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 also that

$$f_{YX}(\nu) = \bar{f}_{XY}(\nu).$$

- Define the "squared coherence" function by

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

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 if $\gamma_{XY}(m) = 0$ for all m, and 1 attained 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 linear filter of $\{X_t\}$.

- Note that

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

where

$$\tan(\omega) = -q_{YX}(\nu)/c_{YX}(\nu);$$

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,

$$f_{YX}(\nu) = \sqrt{\rho^2_{Y\cdot X}(\nu) f_Y(\nu) f_X(\nu)}\, e^{i\phi_{YX}(\nu)}. \tag{3.2}$$

Estimates of the terms on the rhs of (3.2) are computed by R; then $f_{YX}(\nu)$ can be estimated by plugging in these estimates.

- **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) $\rho^2_{Y\cdot X} = 1$.

3.1. Spectrum estimation

- The key tool in spectrum estimation is the "Discrete Fourier Transform" (DFT) of the data. Given data $\{x_t\}_{t=1}^n$, the DFT is

$$X(k) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} x_t e^{-2\pi i \nu_k t} \ (\nu_k = k/n, \ k = 1, ..., n). \tag{3.3}$$

(This is also written $X(\nu_k)$). 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 $X(k) = X_C(k) - iX_S(k)$. The "periodogram" is

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

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

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

Proof: Substitute (3.3) into the rhs, and apply the identity

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

- The obvious estimate of

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

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

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

After some algebra this reduces to the periodogram:

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

- When $X(k)$ is viewed as a r.v., we have (for $\nu_k < 1$; recall that we're really only interested in estimating $f(\nu_k)$ for $0 \leq \nu_k \leq .5$)

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

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

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

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 $n \to \infty$).

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

$$\frac{\hat{f}(\nu_k)}{f(\nu_k)/2} = \left\{\frac{X_C(k)}{\sqrt{f(\nu_k)/2}}\right\}^2 + \left\{\frac{X_S(k)}{\sqrt{f(\nu_k)/2}}\right\}^2 \overset{d}{\approx} \chi^2_2$$

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

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

- For $k \neq l$, $\hat{f}(\nu_k)$ and $\hat{f}(\nu_l)$ are asymptotically 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 $df = 2$.

- 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 $n' - n$ zeros 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 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. The distributional approximations above continue to hold, with the change

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

4. Examples and applications: impulse-response; optimal filtering

- Example: Southern Oscillation Index and Recruits series. Recall Figure 1.1. Here is the basic R command for periodogram computation:
  `soi.per = spec.pgram(soi, log="??")`
  The data are, by default, detrended before the periodogram is computed.

- See Figure 3.1. Here are some values of the first periodogram ($df = 2 \cdot 453/480 = 1.8875$ is also in the output):

```
              freq        power
                 ...
 [8,] 0.016666667 1.067542e+00
 ***first peak; freq=1/60; period = 5 years
[40,] 0.083333333 1.164058e+01
 ***max power; freq=1/12; period = 12 months
                 ...
[240,] 0.500000000 7.083386e-02
```

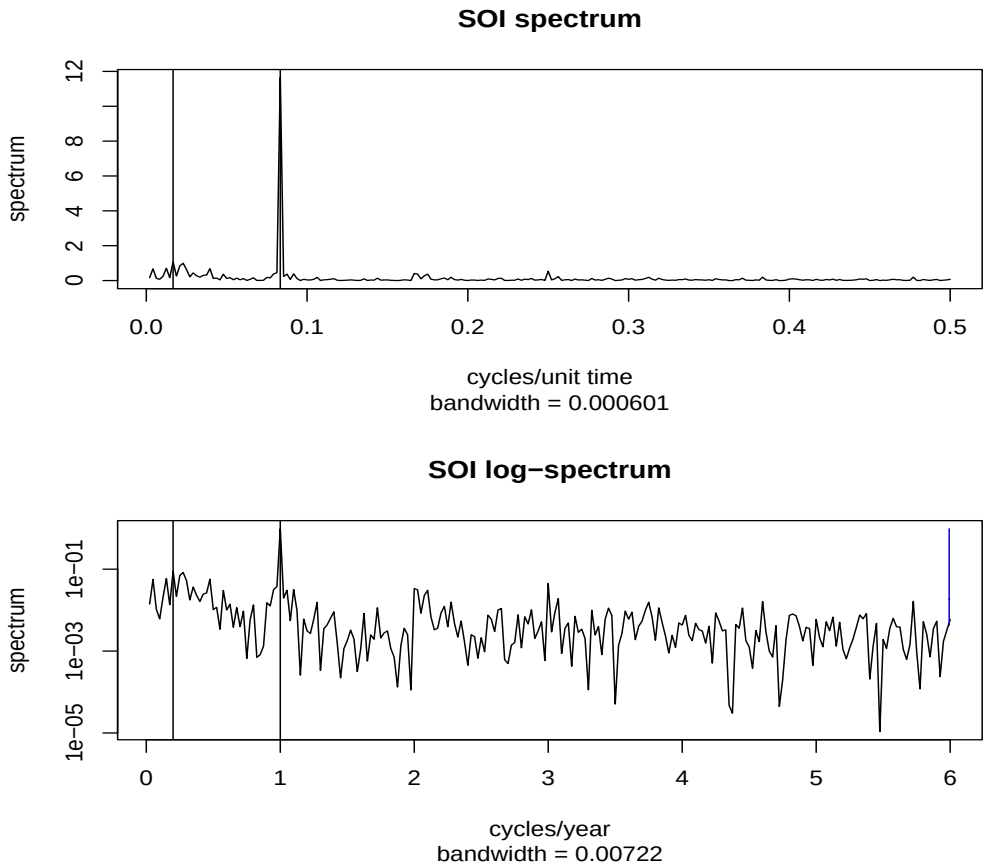

Figure 3.1. SOI periodograms. Note: `log = "yes"` is the default; regardless of this the output is not logged - only the plotted values. As well, the vertical axis is on a log-scale, if `log = "yes"`. In the lower plot the statement `frequency=12` was included when the data were scanned, and the logs are plotted. In this case the (constant) 95% CI width is also shown. Vertical lines mark primary (annual) and secondary (= El Niño) peaks.

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")
```

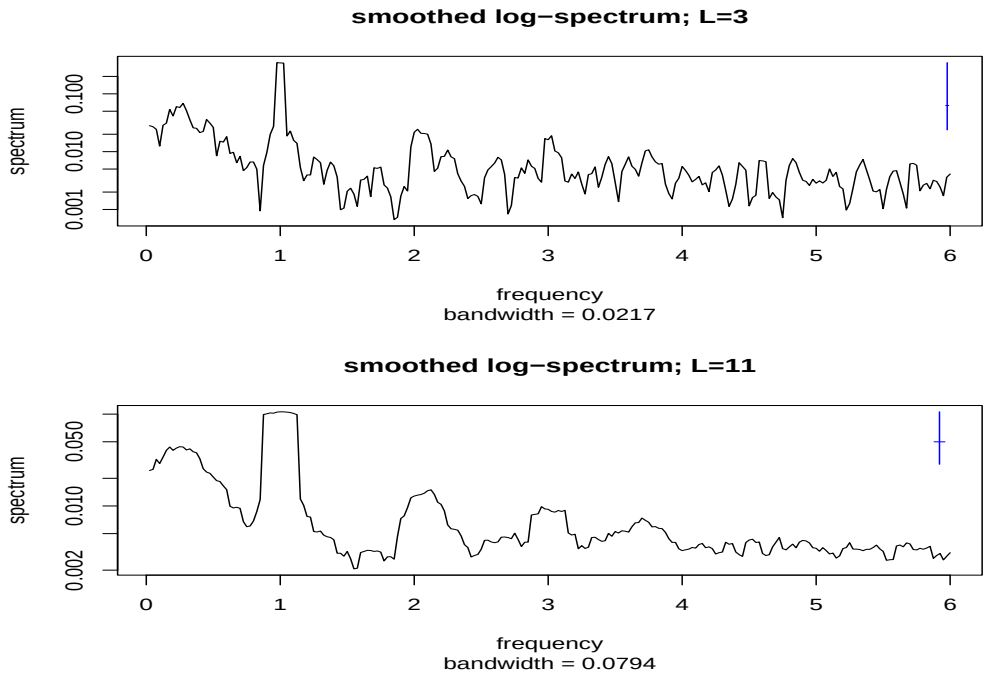

Figure 3.2. Smoothed log-spectra. First uses $L = 3$ ($df = 5.6625$); second uses $L = 11$ ($df = 20.7625$).

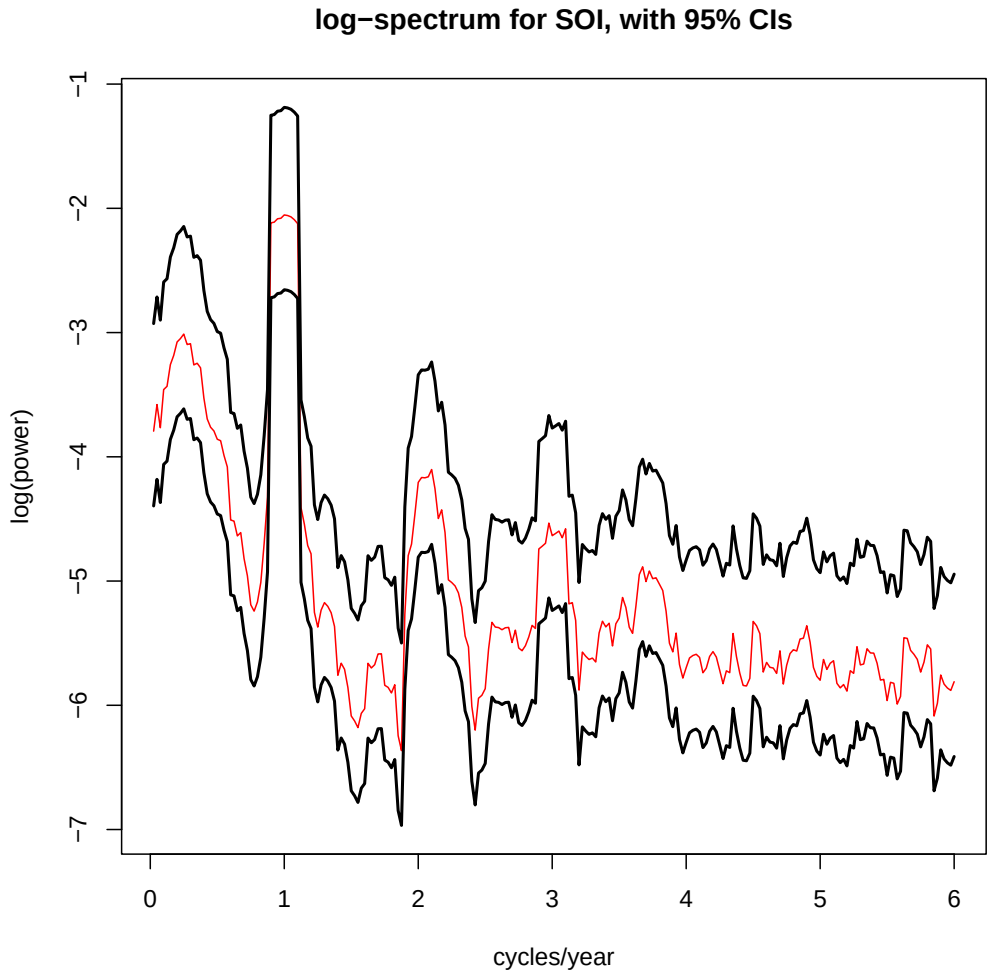

Figure 3.3. Log spectrum with 95% confidence band; $L = 9$. See code on website.

- 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 $\rho^2_{Y\cdot X} = \frac{|f_{YX}(\nu)|^2}{f_Y(\nu)f_X(\nu)}$ is estimated by

$$\hat{\rho}^2_{Y\cdot X} = \frac{|\hat{f}_{YX}(\nu)|^2}{\hat{f}_Y(\nu)\hat{f}_X(\nu)},$$

where

$$\frac{df-2}{2} \cdot \frac{\hat{\rho}^2_{Y\cdot X}}{1-\hat{\rho}^2_{Y\cdot X}} \overset{d}{\approx} F^2_{df-2}$$

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))
```

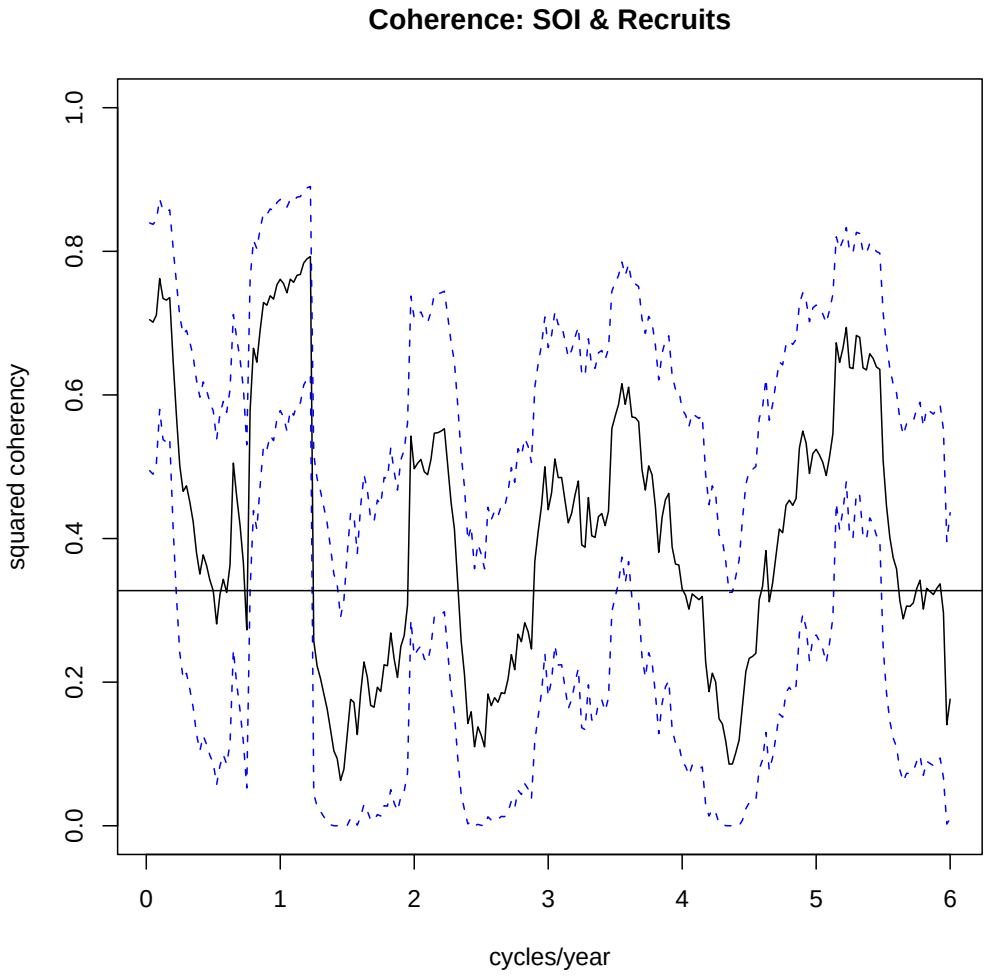

Figure 3.4. Coherence plot with 95% confidence bounds. Horizontal line is at $\alpha = .001$ critical value. Series are strongly coherent at the annual and El Niño frequencies, also at several other higher frequencies, i.e. shorter periods (= seasons?).

- **"Impulse-response" problems**. 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 (detrended) Recruits, X_t is (detrended) SOI, and v_t is zero-mean stationary noise. We will estimate the regression coefficients β_s, and then use only the most significant of them to obtain a finite sum

$$\hat{Y}_t = \sum \hat{\beta}_s X_{t-s},$$

which can then be used to predict Recruits from SOI (if only terms with $s > 0$ are significant).

- 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

$$0 = \frac{\partial MSE}{\partial \beta_r} = E\left[\frac{\partial}{\partial \beta_r}\left\{Y_t - \sum_{s=-\infty}^{\infty} \beta_s X_{t-s}\right\}^2\right]$$

$$= -2E\left[\left\{Y_t - \sum_{s=-\infty}^{\infty} \beta_s X_{t-s}\right\} X_{t-r}\right];$$

$$\text{thus } \gamma_{YX}(r) = \sum_{s=-\infty}^{\infty} \beta_s \gamma_X(r-s)$$

$$(r = 0, \pm 1, \pm 2, \ldots).$$

- Write the equations as

$$\int_{-1/2}^{1/2} e^{2\pi i \nu r} f_{YX}(\nu) d\nu$$

$$= \sum_{s=-\infty}^{\infty} \beta_s \int_{-1/2}^{1/2} e^{2\pi i \nu (r-s)} f_X(\nu) d\nu$$

$$= \int_{-1/2}^{1/2} \left[\sum_{s=-\infty}^{\infty} \beta_s e^{-2\pi i \nu s}\right] e^{2\pi i \nu r} f_X(\nu) d\nu$$

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

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,

$$f_{YX}(\nu) = B_{YX}(\nu) f_X(\nu)$$

and so

$$\begin{aligned} B_{YX}(\nu) &= \frac{f_{YX}(\nu)}{f_X(\nu)}, \\ \beta_s &= \int_{-1/2}^{1/2} B(\nu) e^{2\pi i \nu s} d\nu. \end{aligned}$$

Recall that

$$f_{YX}(\nu) = \sqrt{\rho^2_{Y \cdot X}(\nu) f_Y(\nu) f_X(\nu)}\, e^{i\phi_{YX}(\nu)};$$

this results in

$$B_{YX}(\nu) = \sqrt{\rho^2_{Y \cdot X}(\nu) \frac{f_Y(\nu)}{f_X(\nu)}}\, e^{i\phi_{YX}(\nu)},$$

with estimates of all terms on the right being computed in R.

NOTE: ϕ_{YX} is the phase component of `spec.pgram(cbind(y,x) ... )` - not `cbind(x,y)`.

- In the R programme on the website, for a set of frequencies $\omega_k = k/M$ ($k = 1, 2, ..., M/2$; M should be an even factor of n), we compute $\hat{B}(\omega_k) = \hat{f}_{YX}(\omega_k)/\hat{f}_X(\omega_k)$. Then

$$\beta_s = \int_{-1/2}^{1/2} B(\nu) e^{2\pi i \nu s} d\nu$$

is approximated 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

$$s = -M/2 + 1, ..., -1, 0, 1, ...M/2 - 1.$$

Then we predict $\hat{Y}_t$ by a finite filter

$$\hat{Y}_t = \sum_s \hat{\beta}_s^M X_{t-s}.$$

The authors suggest instead that

$$\beta_s = \int_{-1/2}^{1/2} B(\nu) e^{2\pi i \nu s} d\nu = \int_0^1 B(\nu) e^{2\pi i \nu s} d\nu$$

should be approximated by

$$\hat{\beta}_s^M = \frac{1}{M} \sum_{k=0}^{M-1} \hat{B}(\omega_k) e^{2\pi i \omega_k s}.$$

It can be shown that the difference is negligible:

$$\text{“(1)-(2)”} = \frac{\hat{B}(1/2)(-1)^s - \hat{B}(0)}{M}.$$

(R doesn't return the spectra at $\omega = 0$.)

- Recall $n' = 480$. With $L = 7$ and $M = 80$, some output is shown in Figure 4.1. Other output is:

```
 L = 7 M = 80
 The non-negative lags, at which the coefficients
 are significant, and the betas re-estimated by
 least squares, are
      lag s    beta(s)
 [1,]     5  -19.93
 [2,]     6  -15.91
 [3,]     7  -12.96
 [4,]     8  -10.95
 [5,]     9   -8.54
 [6,]    10   -9.46
 [7,]    11  -12.20
 [8,]    12  -10.99
 [9,]    13   -7.44
[10,]    22    7.01
[11,]    23    7.18
[12,]    24    8.70
The prediction equation is
detrended.output =
   sum( beta(s)*lag(detrended.input, -s)).
 MSE =  229.5924
```

- Interpretation: In order to predict future Recruits from past SOI, we want an equation of the form

$$\hat{Y}_t = \sum_s \hat{\beta}_s X_{t-s}$$

using only positive s. From Figure 4.1 we see that the significant coefficients are at non-negative lags; I have chosen only those which are larger in absolute value than all those at negative lags, i.e. those for which

$$|\hat{\beta}_s| > \max_{s<0} |\hat{\beta}_s| = 4.81.$$

Having determined these lags, I then re-estimated the β_s by least squares (using `dynlm` on R).

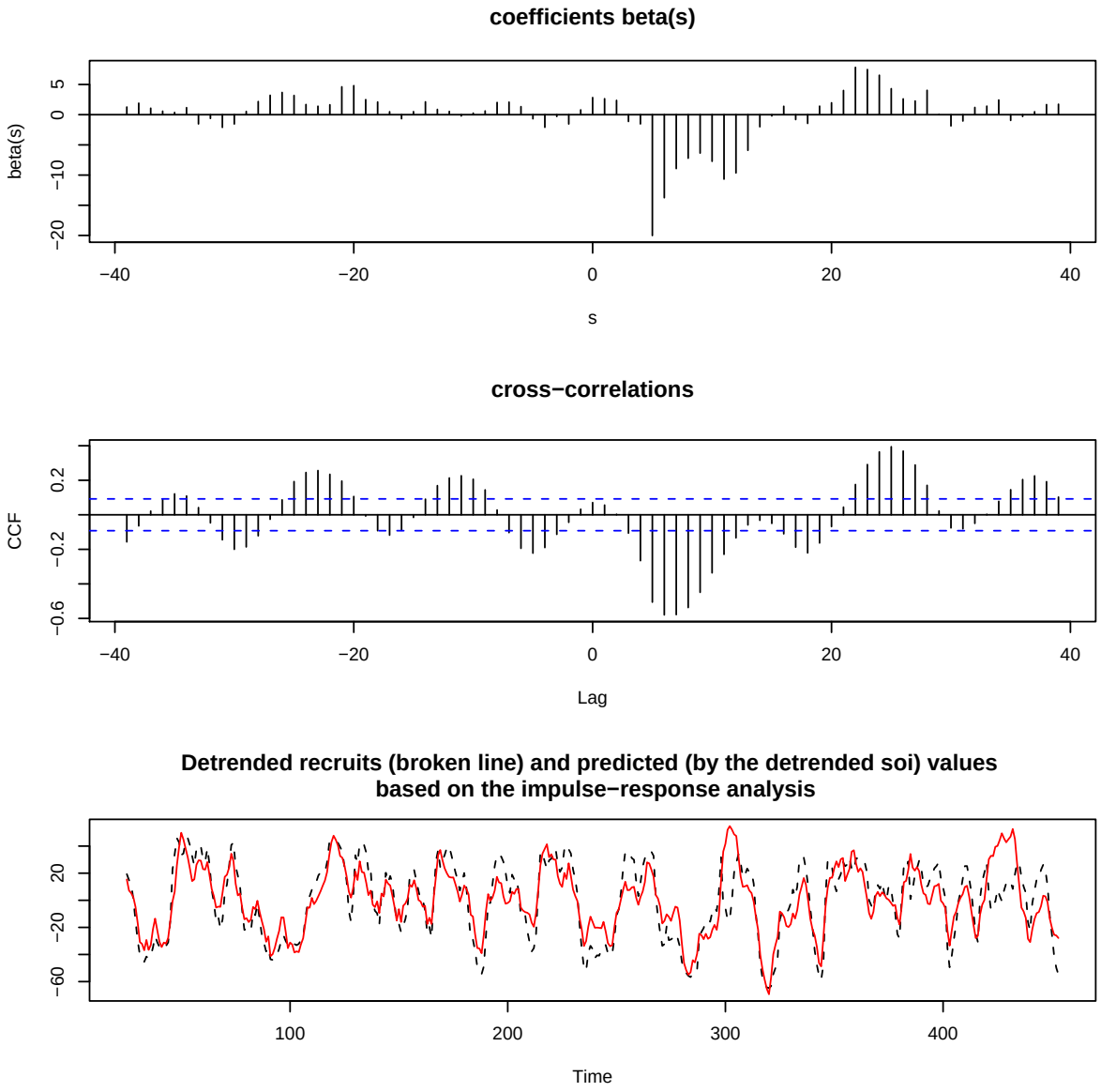

Figure 4.1. Impulse-response analysis of SOI/Recruits. Top: Impulse response function $\left\{\hat{\beta}_s\right\}$. Middle: CCF, for comparison. Bottom: Original Recruits series $\{Y_t\}$ and its prediction $\left\{\hat{Y}_t = \sum \hat{\beta}_s X_{t-s}\right\}$, using only the 'significant' $\hat{\beta}_s$ with $s > 0$.

A more parsimonious model is obtained by first viewing Recruits as the input and using the terms which are significant and which have *non-positive* indices. This gives output

```
 L = 7 M = 80
 The non-positive lags, at which the coefficients
 are significant, and the betas re-estimated by
 least squares, are
     lag s beta(s)
[1,]    -4  0.0179
[2,]    -5 -0.0234
The prediction equation is
detrended.output =
   sum( beta(s)*lag(detrended.input, -s)).
MSE =  0.0679
```

The prediction equation is

$$X_t = \hat{\beta}_{-4} Y_{t+4} + \hat{\beta}_{-5} Y_{t+5};$$

rearranging and shifting the time gives

$$\begin{aligned} Y_t &= \frac{\hat{\beta}_{-4}}{-\hat{\beta}_{-5}} Y_{t-1} + \frac{1}{\hat{\beta}_{-5}} X_{t-5} \\ &= .77 Y_{t-1} - 42.77 X_{t-5}. \end{aligned}$$

- **Optimal filtering**. We estimate $f_X(\nu)$, determine those frequencies which we would like to highlight, and then choose $A(\nu)$ accordingly. Assume that $A(\nu)$ is real and symmetric. 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. Then the sequence $\left\{a_s^M\right\}$ is real and symmetric. 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

$$Y_t^M = \sum_{|s|<M/2} a_s^M X_{t-s} \qquad (4.1)$$

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 programme on the website. The coefficients a_s^M used in (4.1) are replaced by

$$\tilde{a}_s^M = a_s^M h_s,$$

where $h_s = .5\left[1 + \cos\left(2\pi s/\left(M-1\right)\right)\right]$ is a 'cosine bell taper', discussed below. (You should run the script with and without this taper, to see the difference it makes.)

- Example: SOI series. Suppose one wants to study the El Niño signal. An examination of the periodogram ($L = 5$) reveals that to do this we might isolate frequencies in a band $.01 < \nu < .05$. Thus, set $A(\nu) = I\left(.01 < \nu < .05\right)$. With $M = 80$ the output is plotted in Figure 4.2.

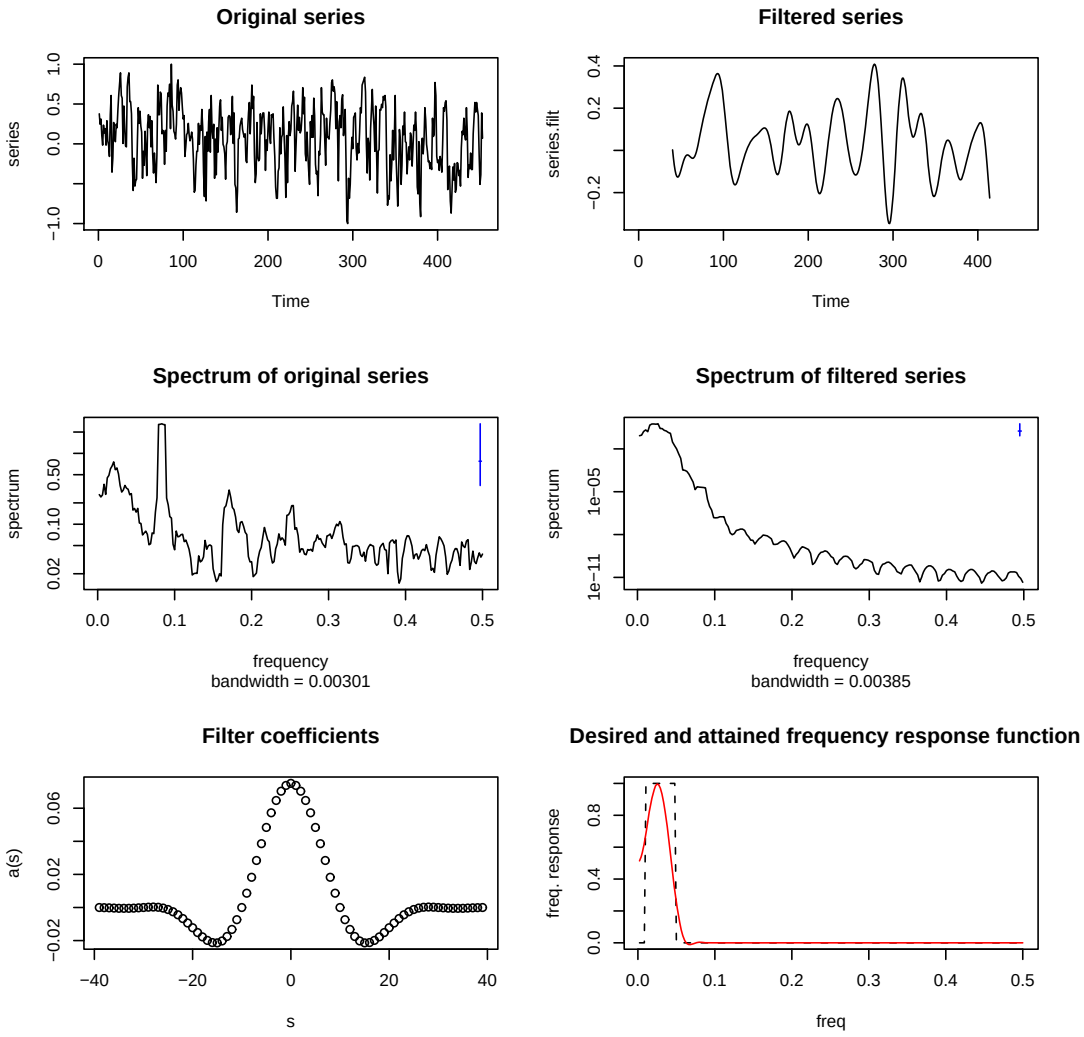

Figure 4.2. Top: Original SOI series $\{X_t\}$ and filtered series $Y_t = \sum_{|s|<M/2} a_s^M X_{t-s}$. Middle: Spectra of these series. Bottom: Filter coefficents $\left\{a_s^M\right\}$; desired and attained frequency responses $A(\nu)$ and $A^M(\nu)$.

- **Tapering**. 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 these raw periodogram values, an alternate approach is to first smooth the data, and to then compute the raw periodogram of the smoothed data. For this, 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 $h_t \equiv 1$; other tapers are also in common use. Typically the taper is chosen to decrease as t moves away from the midpoint $\bar{t}$.

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 $\{h_t\}$. Then $\tilde{f}(\nu_k)$ is

$$\left|\tilde{X}(k)\right|^2 = \frac{1}{n} \sum_{s,t=1}^{n} h_s h_t x_s x_t e^{-2\pi i \nu_k t} e^{2\pi i \nu_k s},$$

with expected value

$$\begin{aligned}
& E\left[\tilde{f}(\nu_k)\right] = E\left[\left|\tilde{X}(k)\right|^2\right] = \\
& \frac{1}{n} \sum_{s,t=1}^{n} h_s h_t \gamma_X(s-t) e^{-2\pi i \nu_k t} e^{2\pi i \nu_k s} \\
= & \frac{1}{n} \sum_{s,t=1}^{n} \left\{ \begin{array}{c} h_s h_t \left[\int_{-1/2}^{1/2} f_X(\omega) e^{2\pi i \omega (s-t)} d\omega\right] \\ \cdot e^{-2\pi i \nu_k t} e^{2\pi i \nu_k s} \end{array} \right\} \\
= & \int_{-1/2}^{1/2} \left\{ \begin{array}{l} \frac{1}{\sqrt{n}} \sum_{s=1}^{n} h_s e^{2\pi i (\nu_k + \omega) s}. \\ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} h_t e^{-2\pi i (\nu_k + \omega) t} \end{array} \right\} f_X(\omega)\, d\omega \\
= & \int_{-1/2}^{1/2} \bar{H}(\nu_k + \omega) H(\nu_k + \omega) f_X(\omega)\, d\omega
\end{aligned}$$

$$= \int_{-1/2}^{1/2} W(\nu_k + \omega) f_X(\omega) d\omega$$
$$= \int_{-1/2}^{1/2} W(\nu_k - \nu) f_X(\nu) d\nu \text{ (how?).}$$

- In particular, our previous approximation of the expected value of the periodogram can now be improved:

$$E\left[\tilde{f}(\nu_k)\right] = \int_{-1/2}^{1/2} W(\nu_k - \nu) f_X(\nu) d\nu. \tag{4.2}$$

- The "window" $W(\omega)$ determines how much of the spectral density $f_X(\omega)$ is "seen" in the computation of the periodogram $\left|\tilde{X}(k)\right|^2$.

- Example 1. If $h_t \equiv 1$ (the unsmoothed periodogram) then (for $\omega \neq 0$)

$$H(\omega) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} e^{-2\pi i \omega t} = \frac{z(1-z^n)}{\sqrt{n}(1-z)}\Big|_{z=e^{-2\pi i\omega}},$$

and so

$$\begin{aligned} W(\omega) &= \frac{1}{n} \cdot \frac{|1-z^n|^2}{|1-z|^2} \\ &= \frac{1}{n} \cdot \frac{2\{1-\mathrm{Re}(z^n)\}}{2\{1-\mathrm{Re}(z)\}} \\ &= \frac{1}{n} \cdot \frac{1-\cos(2\pi n\omega)}{1-\cos(2\pi\omega)}. \end{aligned}$$

Upon using the half-angle formula

$$\cos(2\pi\omega) = 1 - 2\sin^2(\pi\omega),$$

this reduces to

$$W(\omega) = \frac{\sin^2(\pi n\omega)}{n\sin^2(\pi\omega)}.$$

At $\omega = 0$, $H(0) = \sqrt{n}$ and $W(0) = n$. This window is called the *Fejér*, or *modified Bartlett* "kernel".

- Example 2. Let $W_1(\omega)$ and $\tilde{X}_1$ be the window and DFT in Example 1, and put

$$W_L(\omega) = \frac{1}{L} \sum_{l=-(L-1)/2}^{(L-1)/2} W_1(\omega + \nu_l).$$

For this window, (4.2) yields, after a calculation, that

$$E\left[\tilde{f}\left(\nu_k\right)\right] = E\left[\frac{1}{L}\sum_{l=-(L-1)/2}^{(L-1)/2}\left|\tilde{X}_1\left(k+l\right)\right|^2\right].$$

But $\left|\tilde{X}_1\left(k+l\right)\right|^2$ is just the raw periodogram, i.e. is $\left|X\left(k+l\right)\right|^2$; thus

$$E\left[\tilde{f}\left(\nu_k\right)\right] = E\left[\hat{f}_L\left(\nu_k\right)\right] = \int_{-1/2}^{1/2} W_L\left(\nu_k-\nu\right) f_X\left(\nu\right) d\nu.$$

This gives a way to calculate the expected value of the smoothed periodogram exactly. As well, it shows that if one could find a taper $\{h_t\}$ whose DFT $H\left(\omega\right)$ had squared modulus $W_L\left(\omega\right)$, then the resulting spectral estimate $\tilde{f}\left(\nu_k\right)$ would have the same expected value as the smoothed periodogram.

- Example 3. By default, R applies a 'cosine bell' taper, with

$$h_t = .5\left[1+\cos\left(\frac{2\pi\left(t-\bar{t}\right)}{n}\right)\right],\ (\bar{t}=\frac{n+1}{2});$$

this is applied to the first and last 10% of the series (if one keeps the default "`taper=.1`"). Of course, smoothing is done as well.

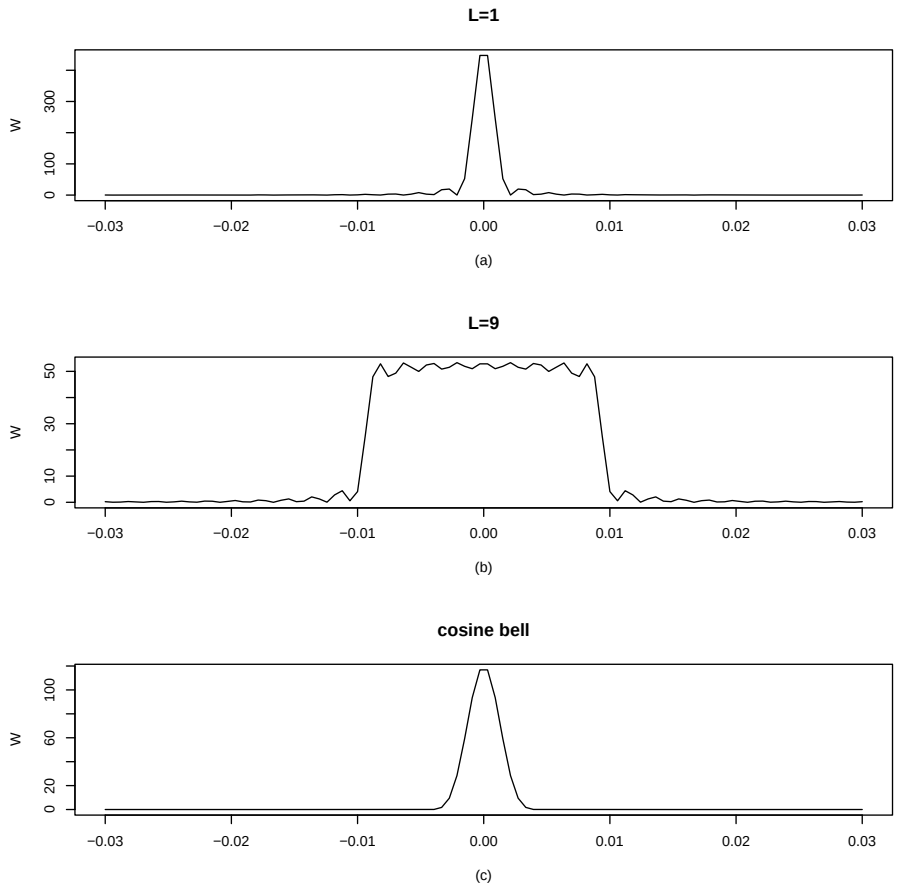

Figure 4.5. Spectral windows, $n = 480$. (a) No smoothing. (b) Smoothing; $L = 9$. (c) Cosine bell taper.

5. Long memory models

- Recall that a ("random walk") model defined by

$$(1 - B)\, x_t = w_t$$

is not stationary. But the "fractionally differenced" series defined by

$$(1 - B)^d\, x_t = w_t$$

is stationary and invertible for certain fractional values of $d \in (-.5, .5)$.

- Binomial expansion, for integer d:

$$(1 - z)^d = \sum_{j=0}^{d} \frac{d\,(d-1)\cdots(d-(j-1))}{j!}\,(-z)^j$$

generalizes (Taylor's Theorem) to

$$(1 - z)^d = \sum_{j=0}^{\infty} \pi_j z^j$$

with

$$\begin{aligned}\pi_j &= \frac{d\,(d-1)\cdots(d-(j-1))}{j!}\,(-1)^j \\ &= \frac{((j-1)-d)\cdots(j-j-d)}{\Gamma\,(j+1)} \\ &= \frac{\Gamma\,(j-d)}{\Gamma\,(-d)\,\Gamma\,(j+1)}.\end{aligned}$$

(By the Ratio Test, convergence is guaranteed for $|z| < 1$.)

- This gives

$$w_t = \sum_{j=0}^{\infty} \pi_j x_{t-j},$$

and

$$x_t = (1-B)^{-d}\, w_t = \sum_{j=0}^{\infty} \psi_j w_{t-j},$$

with

$$\psi_j = \frac{\Gamma\,(j+d)}{\Gamma\,(d)\,\Gamma\,(j+1)}.$$

Stirling's approximation $\Gamma(x) \sim (2\pi)^{1/2} e^{-x} x^{x-1/2}$ as $x \to \infty$ gives $\psi_j \sim e^{-d+1} j^{d-1}/\Gamma(d)$, so that

$$\sum_{j=0}^{\infty} \psi_j^2 < \infty \text{ (stationarity)} \iff d < .5;$$

similarly the process is invertible iff $d > -.5$. Thus we assume $|d| < .5$.

- The ACF may be obtained by integrating the spectrum (which is derived below); this gives

$$\begin{aligned} \gamma_0 &= \sigma_w^2 \frac{\Gamma(1-2d)}{\Gamma^2(1-d)}, \\ \rho_h &= \frac{\Gamma(h+d)\Gamma(1-d)}{\Gamma(h-d+1)\Gamma(d)} = O\left(h^{2d-1}\right). \end{aligned}$$

Thus for $0 < d < .5$ we have $\sum_{h=-\infty}^{\infty} |\rho_h| = \infty$; hence the term "long memory".

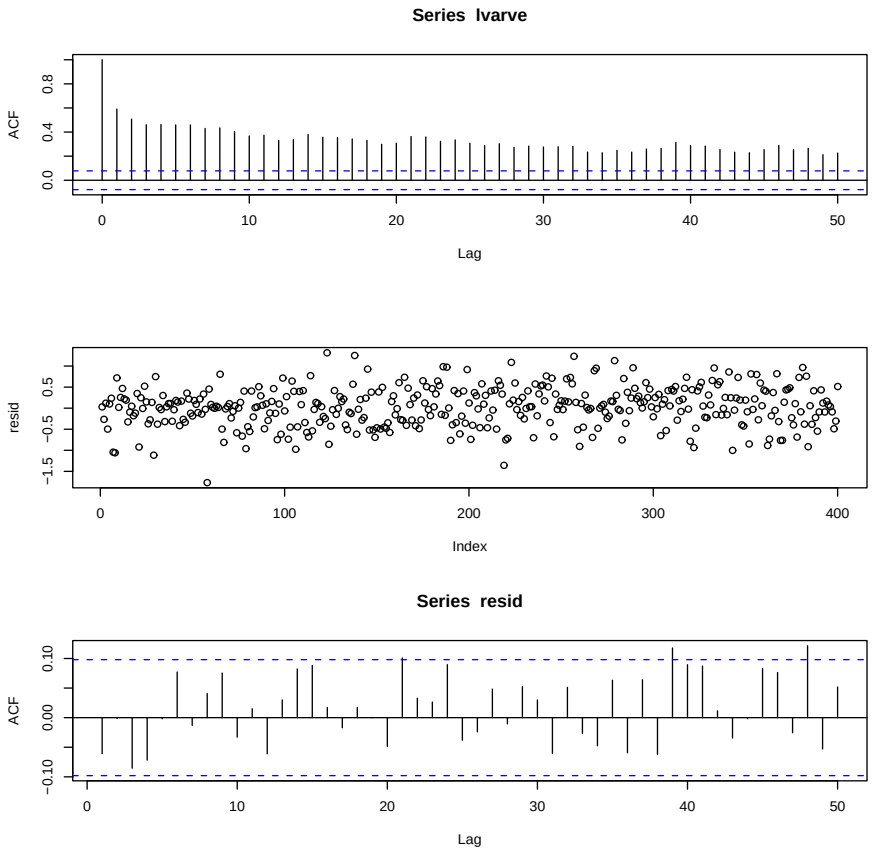

Figure 5.1. Data from Example 5.1 in text. Fit ARFIMA(0,d,0) ("fractionally integrated") to the log(varve) series. R code on website. ARFIMA(0,d,0) with $\hat{d} = .37$ fits about as well as the ARIMA(1,1,1) model fitted earlier.

- Estimation of d is by maximum likelihood. For a Gaussian likelihood this is equivalent to Least Squares, and requires the minimization of $\sum w_t^2(d)$, with $w_t(d)$ approximated by $\sum_{j=0}^{t-1} \pi_j(d)\, x_{t-j}$. Some details in text; R has an implementation which we will use. (`fracdiff` - From the help file: "Calculates the maximum likelihood estimators of the parameters of a fractionally-differenced ARIMA (p,d,q) model, together (if possible) with their estimated covariance and correlation matrices and standard errors, as well as the value of the maximized likelihood. The likelihood is approximated using the fast and accurate method of Haslett and Raftery (1989).")

- Residuals are computed from $\hat{w}_t = \sum_{j=0}^{t-1} \hat{\pi}_j x_{t-j} = \sum_{j=1}^{t} P_j x_{t-j+1}$, where $P_j = \hat{\pi}_{j-1}$.

- Generalization:

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

for characteristic polynomials $\phi(B)$ and $\theta(B)$ of degrees p and q; this is the ARFIMA(p,d,q) model.

- Spectrum. From

$$w_t = (1 - B)^d x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j},$$

by which we can view w_t as a filtered $\{X_t\}$, we get

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

where

$$A(\nu) = \sum_{j=0}^{\infty} \pi_j e^{-2\pi i \nu j} = \left(1 - e^{-2\pi i \nu}\right)^d;$$

thus

$$f_X(\nu) = \frac{\sigma_w^2}{\left|1 - e^{-2\pi i \nu}\right|^{2d}} = \frac{\sigma_w^2}{\left(4 \sin^2(\pi \nu)\right)^d}.$$

The method extends naturally to ARFIMA(p,d,q) models.

- An alternate approach uses the *Whittle likelihood* of the DFT. Assuming a mean of zero, the Gaussian

likelihood of $\{x_t\}_{t=1}^n$ is

$$L(\theta) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} e^{-\frac{\mathbf{x}'\Sigma^{-1}\mathbf{x}}{2}}.$$

The log-likelihood, up to an additive constant, is

$$l(\theta) = -\frac{1}{2}\left\{\log|\Sigma| + \mathbf{x}'\Sigma^{-1}\mathbf{x}\right\}.$$

Write the DFT values as

$$\begin{array}{rcl} \begin{pmatrix} \vdots \\ X(k) \\ \vdots \end{pmatrix} & = & \frac{1}{\sqrt{n}} \begin{pmatrix} & \vdots & \\ e^{-2\pi i\nu_k\cdot 1} & \cdots & e^{-2\pi i\nu_k\cdot n} \\ & \vdots & \end{pmatrix} \mathbf{x}, \\ \text{i.e. } \mathbf{d} & = & \mathbf{U}\mathbf{x}. \end{array}$$

The matrix $\mathbf{U}$ is unitary ($\mathbf{U}\mathbf{U}^* = \mathbf{I}_n$):

$$\begin{aligned} [\mathbf{U}\mathbf{U}^*]_{kl} &= \sum_{j=1}^n \mathbf{U}_{kj}\mathbf{U}^*_{jl} \\ &= \ldots = \frac{1}{n}\sum_{j=1}^n z^j\Big|_{z=e^{-2\pi i\nu_{k-l}}} \\ &= \begin{cases} 1, & k=l, \\ \frac{z(1-z^n)}{1-z} = 0, & k\neq l. \end{cases} \end{aligned}$$

Also $\mathbf{U}\Sigma\mathbf{U}^* = E[\mathbf{d}\mathbf{d}^*] \approx diag(\cdots, f(\nu_k), \cdots)$; this reflects the fact that the $d_k = X(k)$ are as-

ymptotically independent, mean zero, with $E\left[d_k \bar{d}_k\right] = E\left[|X(k)|^2\right] \approx f(\nu_k)$. Thus

$$\begin{aligned} \log|\Sigma| &\approx \sum_{k=1}^{n} \log f(\nu_k), \\ \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} &= \mathbf{d}^*\mathbf{U}\boldsymbol{\Sigma}^{-1}\mathbf{U}^*\mathbf{d} \approx \sum_{k=1}^{n} \frac{|X(k)|^2}{f(\nu_k)}. \end{aligned}$$

This gives, up to an additive constant, "Whittle's likelihood" of $\left\{|X(k)|^2\right\} = \left\{\hat{f}(\nu_k)\right\}$:

$$l(\theta) \approx -\frac{1}{2}\left\{\sum_{k=1}^{n} \log f(\nu_k) + \sum_{k=1}^{n} \frac{\hat{f}(\nu_k)}{f(\nu_k)}\right\}.$$

- Applied to the case above, and with $g_k = 4\sin^2(\pi\nu_k)$, this gives

$$2l(\theta) \approx -n\log\sigma_w^2 + d\sum_{k=1}^{n} \log g_k - \frac{1}{\sigma_w^2}\sum_{k=1}^{n} \hat{f}(\nu_k)\, g_k^d,$$

from which the parameters σ_w^2 and d can be estimated as though the $\left\{\hat{f}(\nu_k)\right\}$ were the data.

6. GARCH models; threshold models

6.1. GARCH models

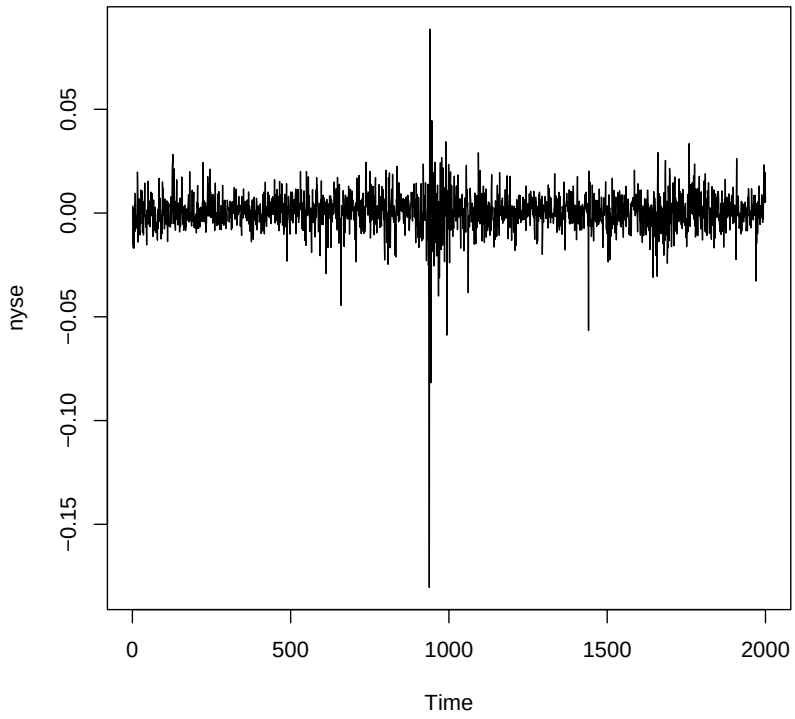

Figure 6.1. New York Stock Exchange returns ($= \nabla z_t / z_{t-1}$ or $\nabla \ln z_t$ for prices z_t). Note burst of activity (high volatility), followed by stretches of less volatility.

- Heteroscedastic models: GARCH (**G**eneralized **A**uto**r**egressive **c**onditionally **h**eteroscedastic) models have been introduced to model changes in 'volatility' (= conditional variation, given the past) over time.

- **ARCH models**. Define

$$y_t = \mu_t + \eta_t,$$

where μ_t is the mean (perhaps with a regression structure) and

$$\eta_t = \sigma_t w_t \tag{6.1}$$

for Gaussian white noise $\{w_t\}$ with unit variance, independent of η^{t-1} (= the history). Furthermore, suppose

$$\sigma_t^2 = \alpha_0 + \alpha_1 \eta_{t-1}^2 + \cdots + \alpha_s \eta_{t-s}^2. \tag{6.2}$$

Note

$$E\left[\eta_t | \eta^{t-1}\right] = \sigma_t E\left[w_t | \eta^{t-1}\right] = 0;$$

we thus call η_t a *martingale.* Then $E[\eta_t] = EE\left[\eta_t|\eta^{t-1}\right] = 0$ and we have $var[y_t] = var[\eta_t] = \sigma_t^2$.

Reason: $var[\eta_t] = E\left\{var\left[\eta_t|\eta^{t-1}\right]\right\}$, with

$$\begin{aligned} var\left[\eta_t|\eta^{t-1}\right] &= E\left\{\eta_t^2|\eta^{t-1}\right\} = E\left\{\sigma_t^2 w_t^2|\eta^{t-1}\right\} \\ &= \sigma_t^2 E\left\{w_t^2|\eta^{t-1}\right\} = \sigma_t^2. \end{aligned}$$

Similarly, $\{\eta_t\}$ is an uncorrelated sequence: For $h > 0$,

$$\begin{aligned} cov[\eta_t, \eta_{t+h}] &= EE\left[\eta_t\eta_{t+h}|\eta^{t+h-1}\right] \\ &= E\left\{\eta_t E\left[\eta_{t+h}|\eta^{t+h-1}\right]\right\} = 0. \end{aligned}$$

- Conditions (6.1) and (6.2) yield that

$$\eta_t^2 = \alpha_0 + \alpha_1\eta_{t-1}^2 + \cdots + \alpha_s\eta_{t-s}^2 + a_t$$

for noise $\{a_t\}$. **Reason**: By (6.1) followed by (6.2),

$$\begin{aligned} \eta_t^2 &= 0 + \sigma_t^2 w_t^2 \\ &= \left[\alpha_0 + \alpha_1\eta_{t-1}^2 + \cdots + \alpha_s\eta_{t-s}^2 - \sigma_t^2\right] + \sigma_t^2 w_t^2 \\ &= \alpha_0 + \alpha_1\eta_{t-1}^2 + \cdots + \alpha_s\eta_{t-s}^2 + a_t, \\ &\quad \text{where } a_t = \sigma_t^2\left(w_t^2 - 1\right). \end{aligned}$$

The parameters and noise must satisfy conditions ensuring that $\eta_t^2 > 0$. The noise a_t is non-Gaussian - it is conditionally a multiple of a centred χ_1^2 r.v. Is it white?

1. $E[a_t] = EE\left[a_t|\eta^{t-1}\right] = E\left\{E\left[\sigma_t^2\left(w_t^2-1\right)|\eta^{t-1}\right]\right\} = E\left\{\sigma_t^2 E\left[w_t^2-1|\eta^{t-1}\right]\right\} = 0.$

2. Suppose $u < t$, then

$$\begin{aligned} & \text{cov}\left[a_u, a_t|\eta^{t-1}\right] \\ = \; & E\left[a_u a_t|\eta^{t-1}\right] \\ = \; & \sigma_u^2\sigma_t^2 E\left[\left(w_u^2-1\right)\left(w_t^2-1\right)|\eta^{t-1}\right] \\ = \; & \sigma_u^2\sigma_t^2 E\left[w_t^2-1\right] E\left[\left(w_u^2-1\right)|\eta^{t-1}\right] \\ = \; & 0, \end{aligned}$$

since w_t is independent of $\left\{\eta^{t-1}, w_u\right\}$. Thus a_u, a_t are uncorrelated if $u < t$. If $u = t$ then the third line becomes

$$\text{var}\left[a_t|\eta^{t-1}\right] = \sigma_t^4 \text{var}\left[w_t^2-1|\eta^{t-1}\right] = 2\sigma_t^4,$$

so that

$$\text{var}\left[a_t\right] = E\left\{\text{var}\left[a_t|\eta^{t-1}\right]\right\} = 2E\left[\sigma_t^4\right].$$

That this be constant imposes certain conditions on the coefficients (discussed below for $s = 1$). If these conditions hold then $\{a_t\}$ is white and $\left\{\eta_t^2\right\}$ is an AR(s) series - stationary, under the usual conditions on the characteristic equation - hence has constant variance.

- To assess these constraints on the coefficients consider as an example the ARCH(1) case (i.e. $s = 1$). We have

$$\begin{aligned}\text{var}\left[\eta_t\right] &= E\left\{var\left[\eta_t|\eta^{t-1}\right]\right\} = E\left\{\sigma_t^2\right\} \\ &= E\left[\alpha_0 + \alpha_1\eta_{t-1}^2\right] = \alpha_0 + \alpha_1\text{var}\left[\eta_{t-1}\right];\end{aligned}$$

if this is to be time independent then

$$\text{var}\left[\eta_t\right] = \frac{\alpha_0}{1-\alpha_1},$$

and it is required that α_0 be positive and $\alpha_1 \in [0, 1)$. That $E\left[\sigma_t^4\right]$ be stationary can be handled in a similar manner and results in the requirement $\alpha_1^2 < 1/3$ - see the text.

- Model selection can be carried out by estimating μ_t (e.g. by Least Squares, in the regression case) and then fitting an AR(s) model to the squared residuals $\hat{\eta}_t^2$.

- The GARCH(r,s) model generalizes (6.2) by replacing it by

$$\beta(B)\,\sigma_t^2 = \alpha_0 + \alpha_1\eta_{t-1}^2 + \cdots + \alpha_s\eta_{t-s}^2,$$

where

$$\beta(B) = 1 - \beta_1 B - \cdots - \beta_r B^r$$

has all zeros outside the unit circle. We assume $\alpha_0 > 0$ and all other coefficients are ≥ 0 (so $\sigma_t^2 > 0$).

- Example: Use R to fit a GARCH(1,1) model to NYSE returns. (Fitting nearby models results in a big deterioration, or obvious overfitting.)

```
Call:
garch(x = nyse, order = c(1, 1))

Coefficient(s):
    Estimate  Std. Error  t value Pr(>|t|)
a0 6.552e-06   6.761e-07    9.691   <2e-16 ***
a1 1.118e-01   4.056e-03   27.554   <2e-16 ***
b1 8.086e-01   1.292e-02   62.566   <2e-16 ***
---
Diagnostic Tests:
Jarque Bera Test data:  Residuals
X-squared = 3983.873, d.f. = 2, p-value < 2.2e-16

Box-Ljung test  data:  Squared.Residuals
X-squared = 1.5874, d.f. = 1, p-value = 0.2077

resids = nyse.g$resid
qqnorm(resids) # Plot not shown
shapiro.test(resids)

Shapiro-Wilk normality test
W = 0.9501, p-value < 2.2e-16
```

- The Jarque Bera Test tests normality by comparing a function of the observed skewness and kurtosis (the normalized third and fourth central moments) of the residuals $\{\hat{w}_t = \hat{\eta}_t/\hat{\sigma}_t\}$ with that expected under normality:

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right), \ (S = \frac{\hat{\mu}_3}{\hat{\sigma}^3}, K = \frac{\hat{\mu}_4}{\hat{\sigma}^4}).$$

- In the R output the "fitted values" are $\pm\hat{\sigma}_t$.

- Various extensions of the GARCH model are discussed in the text.

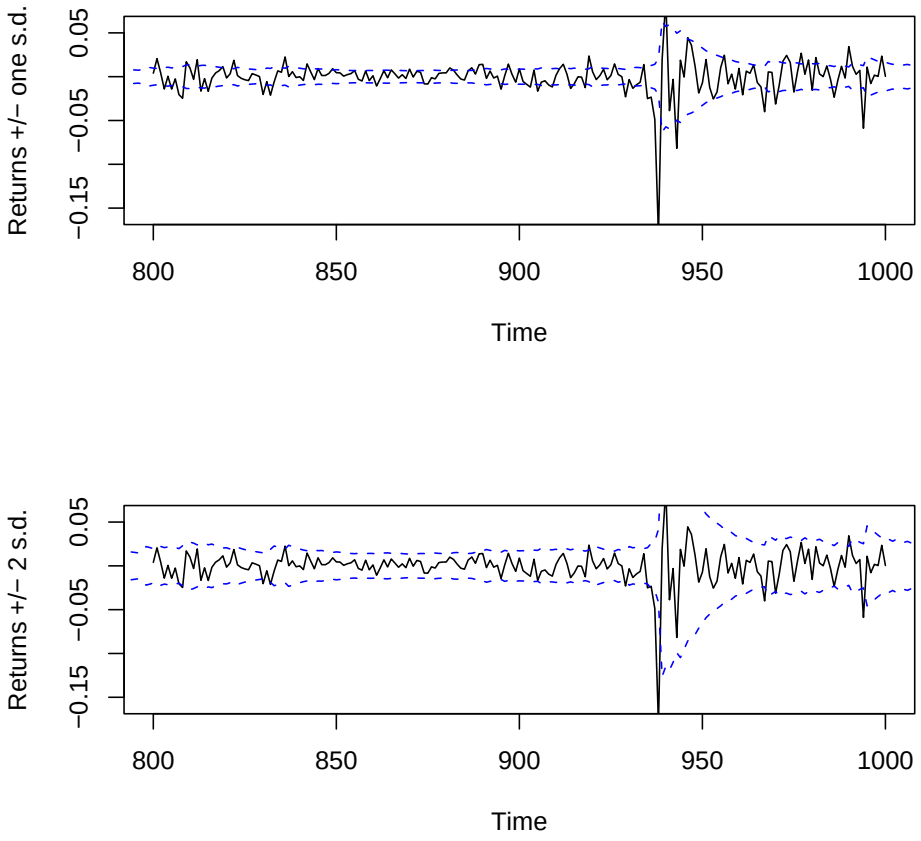

Figure 6.2. Call to R is "`u = predict(nyse.g)`" to get predictions +/- one conditional standard deviation $\hat{\sigma}_t$, then `u + fitted(nyse.g)` adds and subtracts one more standard deviation.

6.2. Threshold models

- The basic idea is that a series $\{x_t\}$ might be well modelled as an AR(p) only for certain values of $\mathbf{x}_{t-1} = \left(x_{t-1}, ..., x_{t-p}\right)'$, say

$$\begin{aligned} x_t &= \phi_0^{(j)} + \mathbf{x}'_{t-1}\phi^{(j)} + w_t^{(j)} \\ \text{for } \mathbf{x}_{t-1} &\in R_j,\ j = 1, ..., r; \\ \text{where } \phi^{(j)} &= \left(\phi_1^{(j)}, ..., \phi_p^{(j)}\right)', \\ \text{and var}\left[w_t^{(j)}\right] &= \sigma_j^2. \end{aligned}$$

  (This allows for an AR(p_j) model on R_j if $p = \max\left\{p_j\right\}$ and $\phi_k^{(j)} = 0$ for $k > p_j$.)

- Example: US monthly pneumonia and influenza deaths, 1969 to 1978. Values increase more slowly than they decrease; peaks not quite evenly spaced; negative trend evident. Time reversal seems to fail. Values of $x_t = \nabla flu_t$ jump to a peak once $x_{t-1} > .05$.

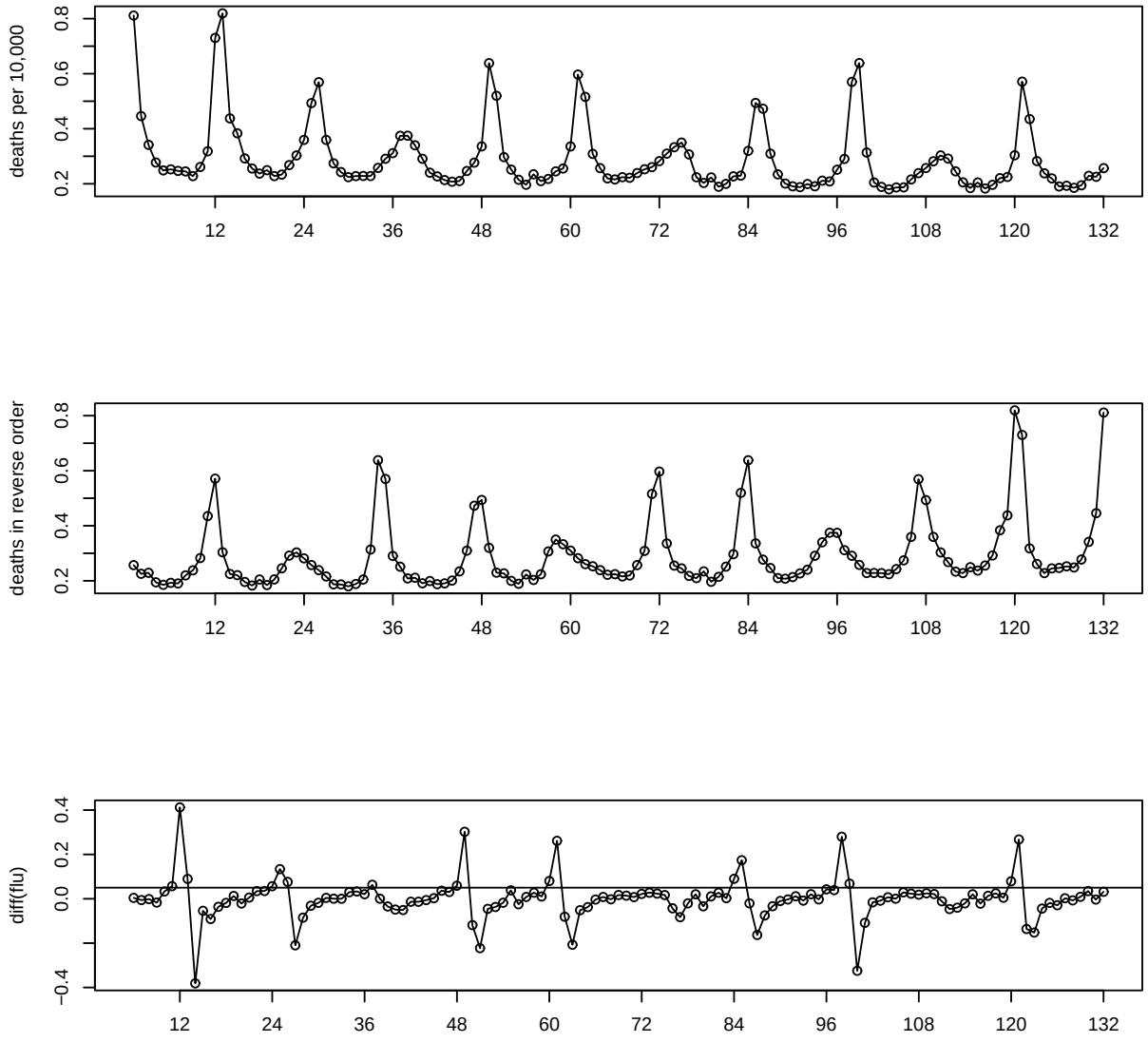

Figure 6.3. Flu deaths in natural and reverse order; first differences.

- Fit a model with $p = 4$, and $R_1 = \{x_t | x_{t-1} < .05\}$, $R_2 = \{x_t | x_{t-1} \geq .05\}$. This is done by using

`dynlm` to regress the values in R_1 on their p predecessors, similarly for R_2. Then the fitted values for the two regressions are pieced together.

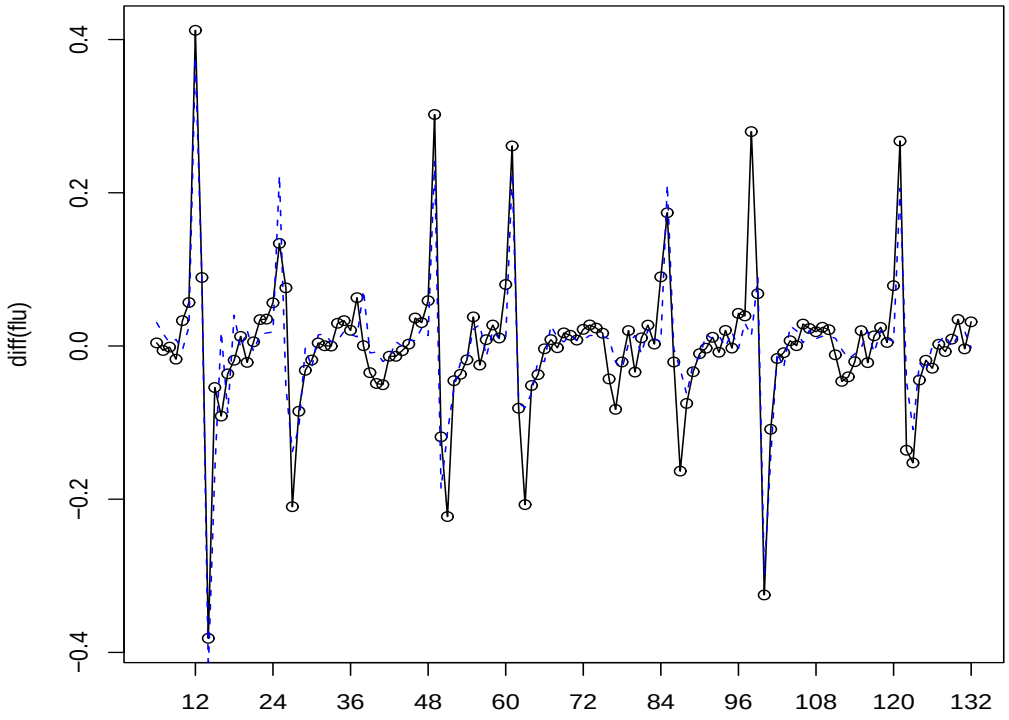

Figure 6.4. Data and fitted values.

See discussion in text regarding the choice of .05 as threshold (.04 was considered but resulted in a poorer fit). Wei (reference on "course information" page on website) discusses formal methods to choose the number of sets R_j, and the thresholds.

7. Regression with autocorrelated errors; transfer function modelling

7.1. Regression with autocorrelated errors

- Consider a time series regression

$$y_t = \beta' \mathbf{z}_t + x_t \tag{7.1}$$

where $\{x_t\}$ is stationary, say AR(p). To fit this we might first fit a regression which assumes uncorrelated errors; then fit an AR model to the residuals. Then (if this works) we have $\phi(B)\, x_t = w_t$; thus

$$\phi(B)\, y_t = \beta' \phi(B)\, \mathbf{z}_t + w_t$$

and one can regress $u_t = \phi(B)\, y_t$ on the regressors $\mathbf{v}_t = \phi(B)\, \mathbf{z}_t$ to obtain an efficient estimate of β. (Then fit an AR model to the residuals which arise from "$u_t = \beta' \mathbf{v}_t + noise$" and iterate; this is the *Cochrane-Orcutt* procedure.)

- This extends, in principle, to ARMA errors. If $\{x_t\}$ is ARMA(p,q) then

$$\phi(B)\, x_t = \theta(B) w_t$$

and the procedure uses $u_t = \frac{\phi(B)}{\theta(B)} y_t$ (compute the u_t's sequentially from $\theta(B) u_t = \phi(B)\, y_t$) and $\mathbf{v}_t = \frac{\phi(B)}{\theta(B)} \mathbf{z}_t$.

- Example: Mortality is related to Temperature and Particulates (in the air). Data (weekly) in Figures 7.1, 7.2. First regress mortality M_t on t, centred temperature $T_t - \bar{T}$, its square and Particulates P_t. ACF and PACF of residuals (Figure 7.3) suggest an AR(2).

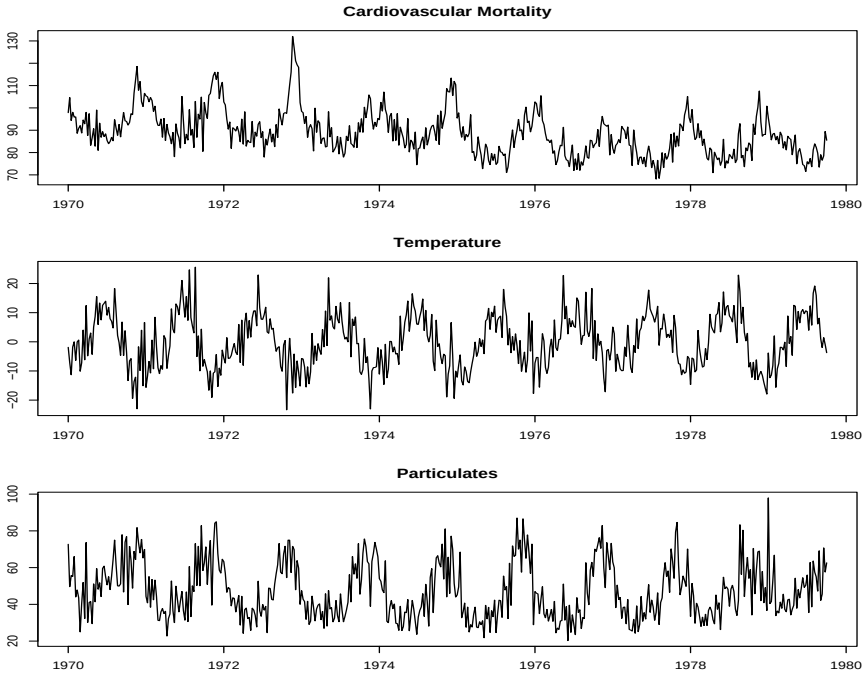

Figure 7.1. Mortality, temperature and particulates.

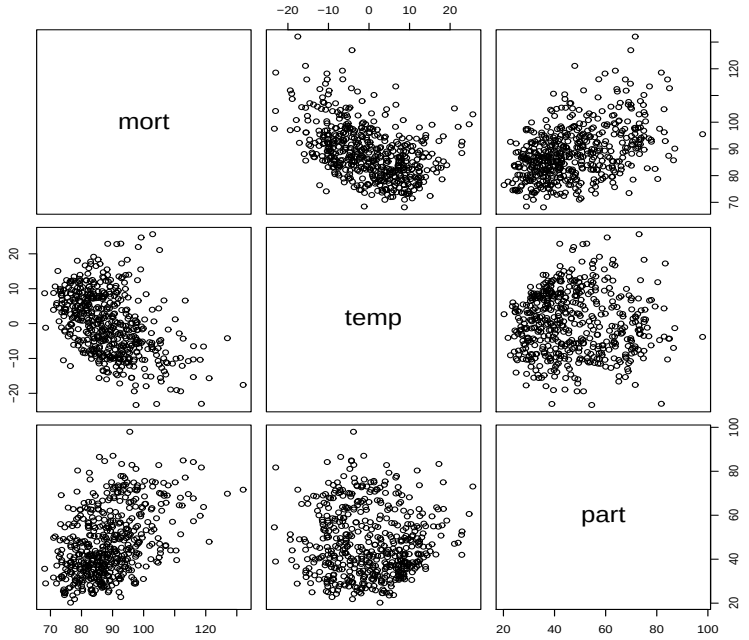

Figure 7.2. Pairs plot.

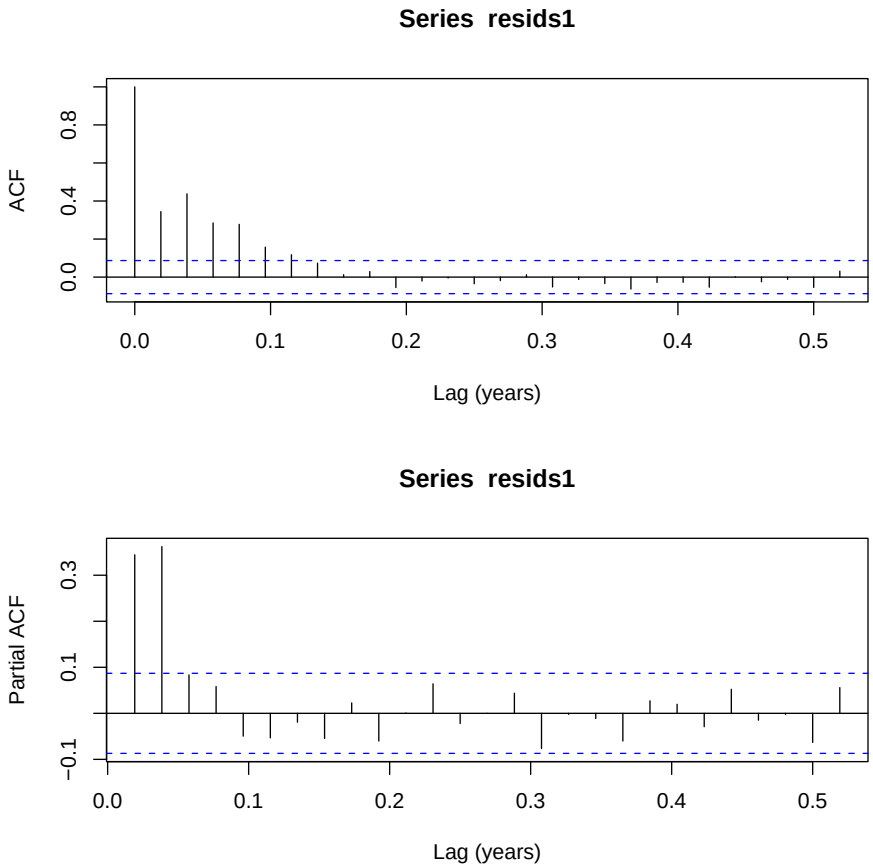

Figure 7.3. ACF and PACF of residuals from first fit of mortality M_t on t, centred temperature $T_t - \bar{T}$, its square and Particulates P_t. Fitting an AR(2) gives $\phi(B) = 1 - 0.2184B - 0.3623B^2$.

Output from initial regression:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Int)	81.592238	1.102148	74.03	< 2e-16	***
trend	-0.026844	0.001942	-13.82	< 2e-16	***
temp	-0.472469	0.031622	-14.94	< 2e-16	***
temp2	0.022588	0.002827	7.99	9.26e-15	***
part	0.255350	0.018857	13.54	< 2e-16	***

Original model

$$M_t = \beta' \mathbf{z}_t + x_t$$

with $\mathbf{z}_t = \left(1, t, \left(T_t - \bar{T}\right), \left(T_t - \bar{T}\right)^2, P_t\right)'$ transforms to

$$\begin{aligned} \phi(B) M_t &= \beta' \phi(B) \mathbf{z}_t + \phi(B) x_t, \text{ i.e.} \\ U_t &= \beta' \mathbf{v}_t + w_t. \end{aligned}$$

Running the regression in the transformed model gives quite a different intercept than that (83.54) reported by S&S:

```
             Estimate Std. Error t value Pr(>|t|)
(Int)      35.018788   0.673906  51.964  < 2e-16 ***
new_trend -0.027768   0.003842  -7.228 1.84e-12 ***
new_temp  -0.197933   0.038662  -5.120 4.37e-07 ***
new_temp2  0.016776   0.002212   7.583 1.65e-13 ***
new_part   0.229610   0.022564  10.176  < 2e-16 ***
```

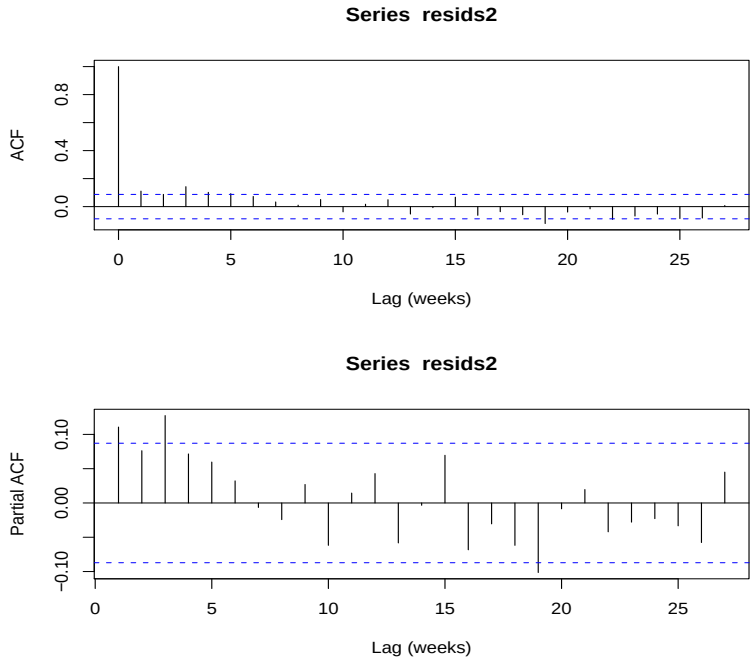

Figure 7.4. ACF and PACF of residuals in transformed model (7.1). They seem at least much closer to being white. Iterating the procedure - fitting AR(1) and repeating - doesn't help.

7.2. Transfer function modelling

- Consider a lagged regression model (e.g. Recruits regressed on SOI) of the form

$$y_t = \alpha(B)\, x_t + \eta_t \text{ for } \alpha(B) = \sum_{j=0}^{\infty} \alpha_j B^j. \tag{7.2}$$

  Here the coefficients of α are assumed absolutely summable, and $\{x_t\}$,$\{\eta_t\}$ are stationary and independent. There may be infinitely many non-zero coefficients α_j, leading to the proposal for a more parsimonious model with

$$\alpha(B) = \frac{\delta(B)\, B^d}{\omega(B)}, \tag{7.3}$$

  with

$$\begin{aligned} \delta(B) &= \delta_0 + \delta_1 B + \cdots + \delta_s B^s, \\ \omega(B) &= 1 - \omega_1 B - \cdots - \omega_r B^r. \end{aligned}$$

- This form of $\alpha(B)$ is called the 'transfer function'.

- S&S propose a sequential approach to fitting. For this, first fit an ARMA model to $\{x_t\}$:

$$\phi(B)\, x_t = \theta(B)\, w_t$$

and apply the operator $\phi(B)/\theta(B)$ (transforming x_t to white noise $\tilde{w}_t$) to both sides of (1):

$$\begin{aligned} \frac{\phi(B)}{\theta(B)} y_t &= \frac{\phi(B)}{\theta(B)} \alpha(B)\, x_t + \frac{\phi(B)}{\theta(B)} \eta_t; \text{ or} \\ \tilde{y}_t &= \alpha(B)\, \tilde{w}_t + \tilde{\eta}_t. \end{aligned}$$

The cross-covariance function of $\{\tilde{y}_t\}$ with $\{\tilde{w}_t\}$ is

$$\gamma_{\tilde{y}\tilde{w}}(h) = cov\left[\alpha(B)\, \tilde{w}_{t+h}, \tilde{w}_t\right] = \sigma^2_{\tilde{w}} \alpha_h;$$

this gives a way to estimate the structure of $\alpha(B)$.

Now use (7.3) in (7.2):

$$\omega(B)\, y_t = \delta(B)\, B^d x_t + \omega(B)\, \eta_t, \tag{7.4}$$

i.e.

$$y_t = \sum_{k=1}^{r} \omega_k y_{t-k} + \sum_{k=0}^{s} \delta_k x_{t-d-k} + u_t$$

for $u_t = \omega(B)\eta_t$. Estimate the parameters by regression. Finally, calculate

$$\hat{\eta}_t = \frac{1}{\omega(B)}\hat{u}_t.$$

and use these to fit an ARMA model to $\{\eta_t\}$:

$$\phi_\eta(B)\eta_t = \theta_\eta(B)z_t,$$

where $\{z_t\}$ is white noise. Thus

$$u_t = \frac{\omega(B)\theta_\eta(B)}{\phi_\eta(B)}z_t$$

gives an ARMA model for $\{u_t\}$.

- Final model is

$$\begin{aligned} y_t &= \alpha(B)x_t + \eta_t \\ &= \frac{\delta(B)B^d}{\omega(B)}x_t + \frac{\theta_\eta(B)}{\phi_\eta(B)}z_t, \end{aligned}$$

where $\{z_t\}$ is white noise.

Example: x_t = SOI (detrended: `soi.detrended = lsfit(time(soi), soi)$resid`) and y_t = Recruits (also detrended). Illustrates the five point sequential process advocated by S&S.

1. **Fit ARMA model to** $\{x_t\}$ Sample ACF/PACF in Figure 7.5 suggest AR(1) (ARMA(1, q=1,2) models were no better); fitting this gives $(1 - \phi B)\, x_t = w_t$ with $\hat{\phi} = .5875$, $\hat{\sigma}_w^2 = .092$.

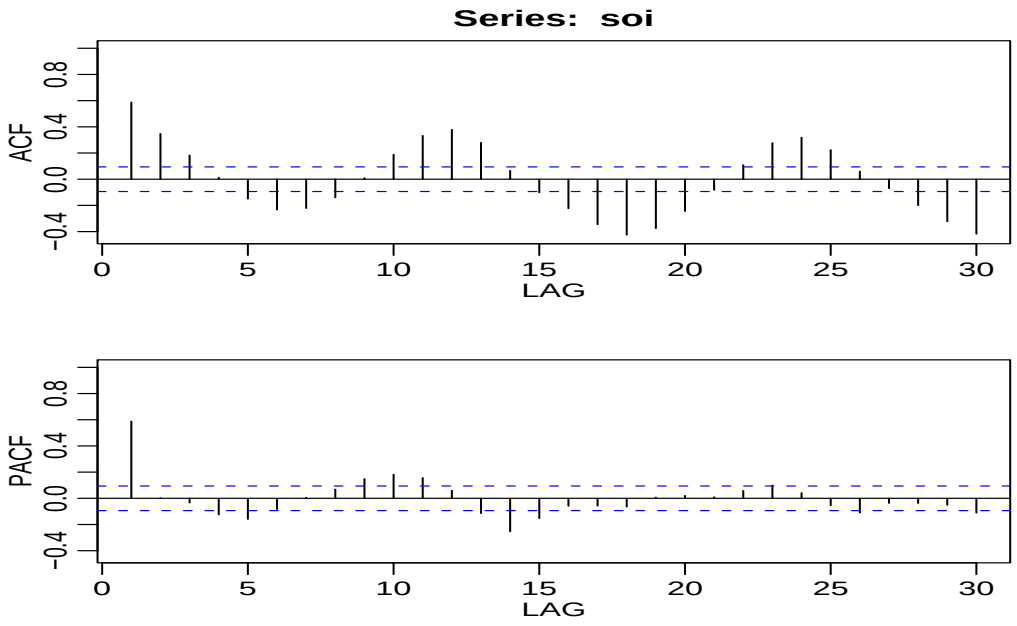

Figure 7.5.

2. Apply to $\{x_t\}$ and $\{y_t\}$ to get (7.3) Compute

$$\left(1-\hat{\phi}B\right)x_t = \tilde{w}_t,\ \left(1-\hat{\phi}B\right)y_t = \tilde{y}_t$$

(note that the $\tilde{w}_t$ are just the residuals from the AR fit).

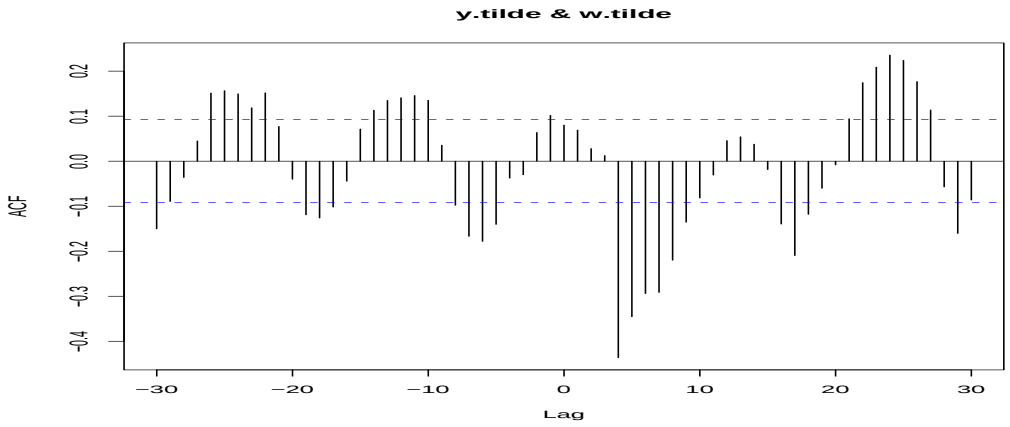

Figure 7.6.

3. Examine the CCF of the transformed series to suggest a transfer function. Here the CCF in Figure

7.6 suggests that $\tilde{w}$ leads $\tilde{y}$ by 5 units (so take $B^5\tilde{w}_t$) and then a roughly exponential decrease:

$$\begin{aligned} \alpha_{h+5} &\propto \omega^h,\ h = 0, 1, ...; \\ \text{i.e. } \alpha(B) &= \delta B^5 \left(1 + \omega B + \omega^2 B^2 + ...\right) = \frac{\delta B^5}{1 - \omega B}. \end{aligned}$$

4. Estimate parameters by regression Equation (7.4) is

$$\begin{aligned} (1 - \omega B) y_t &= \delta B^5 x_t + (1 - \omega B) \eta_t, \text{ i.e.} \\ y_t &= \omega y_{t-1} + \delta x_{t-5} + u_t \end{aligned}$$

with $u_t = (1 - \omega B)\eta_t = \eta_t - \omega\eta_{t-1}$. A regression gives $\hat{\omega} = .8479$, $\hat{\delta} = -20.536$.

5. Fit an ARMA model to $\{\eta_t\}$. We have

$$\frac{1}{\omega(B)} u_t = \eta_t = \frac{\theta_\eta(B)}{\phi_\eta(B)} z_t.$$

Backsolve to get $\eta_t = u_t + \hat{\omega}\eta_{t-1}$; plot the ACF/PACF:

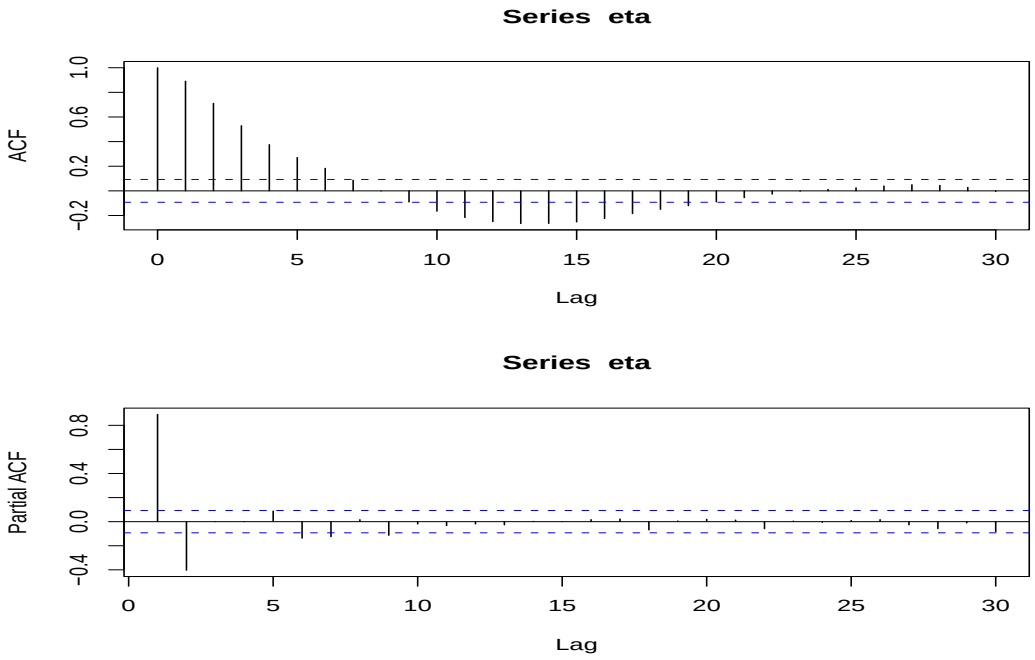

Figure 7.7. ACF/PACF for $\{\eta_t\}$.

Fit an AR(2); i.e. $\theta_\eta(B) = 1$, $\phi_\eta(B) = 1 - \phi_1 B - \phi_2 B^2$:

```
Coefficients:
         ar1      ar2    xmean
      1.258  -0.4099  -0.7902
s.e.  0.043   0.0431   2.1635

sigma^2 estimated as 49.01
```

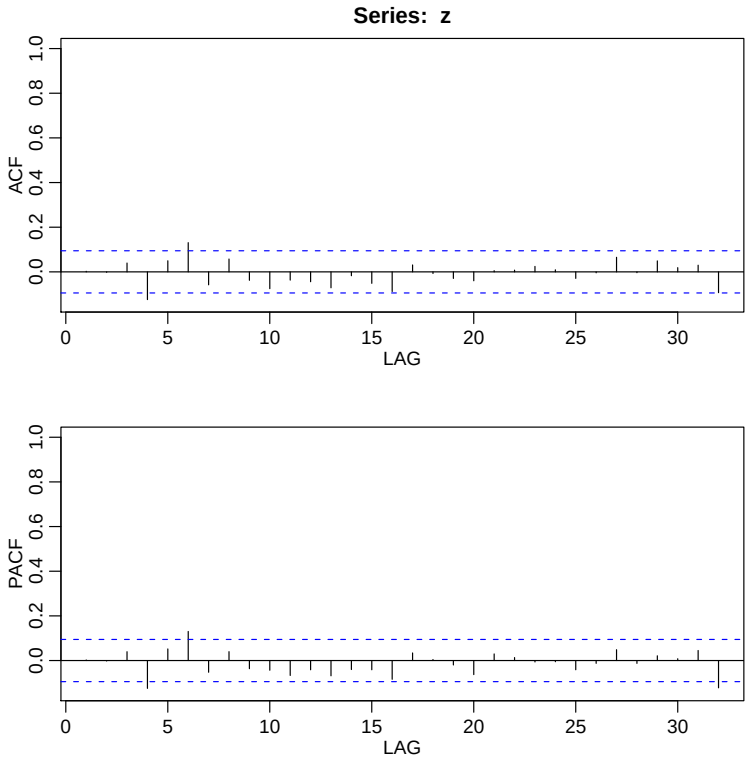

Figure 7.8. Residuals from the fit $\phi_\eta(B)(\eta_t - \bar{\eta}) = z_t$, i.e. $(\eta_t + .79) = 1.258(\eta_{t-1} + .79) - .4099(\eta_{t-2} + .79) + z_t$.

Final model: Since the estimate $\bar{\eta}$ of μ_η is $< .5$ $s.e.$ away from zero, one could take $\mu_\eta = 0$. If a mean is to be fitted, one obtains

$$y_t = \gamma + \omega y_{t-1} + \delta x_{t-5} + \frac{1 - \omega B}{\phi_\eta(B)} z_t,$$

with $\gamma = (1 - \omega)\mu_\eta$ and μ_η estimated by $\bar{\eta} = -.7902$.

8. Multivariate regression and ARMAX

8.1. Multivariate regression

- We extend the univariate linear model

$$y_t = \beta' \mathbf{z}_t + w_t,$$

in which a single variable y_t is expressed in terms of fixed regressors $\mathbf{z}_t = \left(z_{t,1}, \cdots, z_{t,r}\right)'$, parameters $\beta_{r\times 1}$ and white noise $\{w_t\}$, to a multivariate model in which several variables $\mathbf{y}_t = \left(y_{t,1}, \cdots, y_{t,k}\right)'$ are modelled in this way:

$$\begin{pmatrix} y_{t,1} \\ \vdots \\ y_{t,k} \end{pmatrix} = \begin{pmatrix} \beta_1' \\ \vdots \\ \beta_k' \end{pmatrix} \mathbf{z}_t + \begin{pmatrix} w_{t,1} \\ \vdots \\ w_{t,k} \end{pmatrix},$$

$$\text{i.e. } \mathbf{y}_t = \mathbf{B}_{k\times r}\mathbf{z}_t + \mathbf{w}_t.$$

We arrange these as

$$\left(\mathbf{y}_1, \cdots, \mathbf{y}_n\right) = \mathbf{B}\left(\mathbf{z}_1, \cdots, \mathbf{z}_n\right) + \left(\mathbf{w}_1, \cdots, \mathbf{w}_n\right),$$

i.e.

$$\mathbf{Y}' = \mathbf{B}\mathbf{Z}' + \mathbf{W}',$$

for $\mathbf{Y}_{n\times k}, \mathbf{Z}_{n\times r}$ and $\mathbf{W}_{n\times k}$.

- We assume that the disturbances $\mathbf{w}_1, ..., \mathbf{w}_n$ are independently, identically and normally distributed, with means $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_{k\times k}$. Estimation is by maximum likelihood, which is equivalent to least squares for Gaussian errors. Since $\mathbf{y}_t$ is multivariate Normal, with mean $\mathbf{B}\mathbf{z}_t$ and covariance matrix $\mathbf{\Sigma}$, the likelihood is

$$L = \prod_{t=1}^{n} \left\{ \begin{array}{c} (2\pi)^{-k/2} \left|\mathbf{\Sigma}\right|^{-1/2} . \\ \exp -\frac{1}{2} \left(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t\right)' \mathbf{\Sigma}^{-1} \left(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t\right) \end{array} \right\},$$

with log-likelihood (ignoring the constants)

$$\begin{aligned} l &= -\frac{n}{2} \log \left|\mathbf{\Sigma}\right| - \frac{1}{2} \sum_{t=1}^{n} \left(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t\right)' \mathbf{\Sigma}^{-1} \left(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t\right) \\ &= -\frac{n}{2} \log \left|\mathbf{\Sigma}\right| - \frac{1}{2} tr \mathbf{\Sigma}^{-1} \left[\sum_{t=1}^{n} \left(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t\right) \left(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t\right)'\right] \end{aligned}$$

Then

$$-2l = n \log \left|\mathbf{\Sigma}\right| + tr \mathbf{\Sigma}^{-1} \left(\mathbf{Y}' - \mathbf{B}\mathbf{Z}'\right) \left(\mathbf{Y}' - \mathbf{B}\mathbf{Z}'\right)'. \tag{8.1}$$

- We assume that the columns of $\mathbf{Z}$ are linearly independent. Define $\mathbf{H}_{n\times n} = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'$. Note $\mathbf{H}$ is idempotent and projects onto the column space of $\mathbf{Z}$. To maximize the likelihood over $\mathbf{B}$ we are to minimize

$$\begin{aligned} & tr\boldsymbol{\Sigma}^{-1}\left(\mathbf{Y}'-\mathbf{B}\mathbf{Z}'\right)\left(\mathbf{Y}'-\mathbf{B}\mathbf{Z}'\right)' \\ = \; & tr\boldsymbol{\Sigma}^{-1}\left(\mathbf{Y}'-\mathbf{B}\mathbf{Z}'\right)\left[(\mathbf{I}-\mathbf{H})+\mathbf{H}\right]\left(\mathbf{Y}'-\mathbf{B}\mathbf{Z}'\right)' \\ = \; & tr\boldsymbol{\Sigma}^{-1}\mathbf{Y}'\left[\mathbf{I}-\mathbf{H}\right]\mathbf{Y} \\ & +tr\boldsymbol{\Sigma}^{-1}\left(\mathbf{Y}'-\mathbf{B}\mathbf{Z}'\right)\mathbf{H}\left(\mathbf{Y}'-\mathbf{B}\mathbf{Z}'\right)'. \end{aligned}$$

  Only the second trace depends on $\mathbf{B}$; it is nonnegative but can be made $=$ 0 by solving $\mathbf{H}\left(\mathbf{Y}'-\mathbf{B}\mathbf{Z}'\right)' = \mathbf{0}$ to get

$$\hat{\mathbf{B}} = \mathbf{Y}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}. \tag{8.2}$$

  Equivalently, each $\hat{\beta}_i$ is obtained by a univariate regression in the model

$$\begin{pmatrix} y_{1i} \\ \vdots \\ y_{ni} \end{pmatrix} = \mathbf{Z}\beta_i + \begin{pmatrix} w_{1i} \\ \vdots \\ w_{ni} \end{pmatrix},$$

  with i.i.d. errors $w_{ti} \sim N(0, \sigma_i^2)$.

- Now Σ is estimated by minimizing $n \log |\Sigma| + tr\Sigma^{-1}\mathbf{Y}' [\mathbf{I} - \mathbf{H}] \mathbf{Y}$; some calculus shows that the minimizer is

$$\hat{\Sigma}_w = \frac{1}{n}\mathbf{Y}' [\mathbf{I} - \mathbf{H}] \mathbf{Y} = \frac{1}{n} \sum_{t=1}^{n} \left(\mathbf{y}_t - \hat{\mathbf{B}}\mathbf{z}_t\right) \left(\mathbf{y}_t - \hat{\mathbf{B}}\mathbf{z}_t\right)'.$$

We commonly write

$$\begin{aligned} \hat{\mathbf{\Sigma}}_w &= \frac{1}{n} \sum_t \left(\mathbf{x}_t - \hat{\mathbf{B}}\mathbf{z}_t\right) \left(\mathbf{x}_t - \hat{\mathbf{B}}\mathbf{z}_t\right)' \\ &= \frac{1}{n} \sum_t \hat{\mathbf{w}}_t \hat{\mathbf{w}}_t' \\ &= \text{“}\frac{1}{n} RSP\text{”}, \end{aligned}$$

where $\hat{\mathbf{w}}_t = \mathbf{x}_t - \hat{\mathbf{B}}\mathbf{z}_t$ is the t^{th} residual and $RSP = \hat{\mathbf{W}}'\hat{\mathbf{W}}$ is the “residual sum of products” matrix. This is sometimes adjusted for bias by dividing by $rk(\mathbf{I} - \mathbf{H}) = n - r$ rather than by n.

- The measures AIC, BIC etc. are adjusted accordingly:

$$\begin{aligned} AIC &= \ln\left|\hat{\Sigma}_w\right| + \frac{2}{n}\left(kr + \frac{k(k+1)}{2}\right), \\ AICc &= \ln\left|\hat{\Sigma}_w\right| + \frac{k(r+n)}{n-k-r-1}, \\ BIC &= \ln\left|\hat{\Sigma}_w\right| + \frac{k^2 p \ln n}{n} \text{ (for VAR(p)).} \end{aligned}$$

- Standard asymptotic theory of maximum likelihood estimation says that the asymptotic covariance matrix of $\hat{\mathbf{B}}$ is the inverse of the Fisher information matrix:

$$\text{cov}\left[\text{vec}\hat{\mathbf{B}}\right] = \left\{E\left[-\ddot{l}\left(\mathbf{B}\right)\right]\right\}^{-1} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1} \otimes \mathbf{\Sigma}.$$

The variances of the elements of $\hat{\mathbf{B}}$ are on the diagonal of this matrix:

$$\text{var}\left[\hat{\mathbf{B}}_{ij}\right] = \sigma_{ii}\left(\mathbf{Z}'\mathbf{Z}\right)^{jj}, \ i = 1, ..., k; \ j = 1, ..., r.$$

(Compare with S&S p. 302, now corrected)

- Vecs and Kronecker products. If

$$\mathbf{B}_{k\times r} = (\mathbf{b}_1, ..., \mathbf{b}_r),$$

where each $\mathbf{b}_j$ is $k \times 1$, then

$$\text{vec}\mathbf{B} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_r \end{pmatrix} : kr \times 1.$$

If $\mathbf{A}$ is $p \times q$ then

$$\mathbf{A}\otimes\mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1j}\mathbf{B} & \cdots & a_{1q}\mathbf{B} \\ a_{21}\mathbf{B} & & & & & \\ \vdots & & & & & \\ a_{i1}\mathbf{B} & \cdots & \cdots & a_{ij}\mathbf{B} & \cdots & a_{iq}\mathbf{B} \\ \vdots & & & & & \\ a_{p1}\mathbf{B} & \cdots & \cdots & a_{pj}\mathbf{B} & \cdots & a_{pq}\mathbf{B} \end{pmatrix},$$

a $pk \times qr$ matrix. Whenever the matrices involved are conformable, we have

$$\begin{aligned} (\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) &= \mathbf{AC} \otimes \mathbf{BD}; \\ (\mathbf{A} \otimes \mathbf{B})' &= \mathbf{A}' \otimes \mathbf{B}'; \\ (\mathbf{A} \otimes \mathbf{B})^{-1} &= \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}. \end{aligned}$$

- Note in particular

$$\text{vec}\left(\mathbf{ab}'\right) = \mathbf{b} \otimes \mathbf{a};$$

this implies

$$\text{vec}\left(\mathbf{ABC}\right) = \left(\mathbf{C}' \otimes \mathbf{A}\right) \text{vec}\mathbf{B}.$$

(Proof: Write $\mathbf{B} = \sum_{j=1}^{r} \mathbf{b}_j \mathbf{e}_j'$, where $\mathbf{e}_j'$ is the j^{th} unit vector in $\mathbb{R}^r$, etc.).

- Arrangement of partial derivatives: if the elements of $\mathbf{y}_{p\times 1}$ are functions of $\mathbf{x}_{q\times 1}$, then

$$\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right) = \left(\frac{\partial y_i}{\partial x_j}\right) : p \times q.$$

In particular,

$$\left(\frac{\partial \mathbf{Ax}}{\partial \mathbf{x}}\right) = \mathbf{A}, \tag{8.3}$$

$$\left(\frac{\partial \mathbf{x}'\mathbf{Ax}}{\partial \mathbf{x}}\right) = 2\mathbf{x}'\mathbf{A}. \tag{8.4}$$

- We define the gradient and Hessian of a function of matrices by

$$\dot{l}(\mathbf{B}) = \left(\frac{\partial l(\mathbf{B})}{\partial \text{vec}\mathbf{B}}\right)' : kr \times 1$$

and then

$$\ddot{l}(\mathbf{B}) = \left(\frac{\partial \dot{l}(\mathbf{B})}{\partial \text{vec}\mathbf{B}}\right) : kr \times kr.$$

- First, apart from some additive constants we have

$$l(\mathbf{B}) = -\frac{1}{2}\sum_{t=1}^{n} \mathbf{w}_t'\boldsymbol{\Sigma}^{-1}\mathbf{w}_t.$$

where $\mathbf{w}_t = \mathbf{y}_t - \mathbf{B}\mathbf{z}_t$. By (8.4) and the Chain Rule,

$$\frac{\partial l(\mathbf{B})}{\partial \text{vec}\mathbf{B}} = -\sum_{t=1}^{n} \mathbf{w}_t'\boldsymbol{\Sigma}^{-1}\frac{\partial \mathbf{w}_t}{\partial \text{vec}\mathbf{B}}.$$

Now using (8.3),

$$\begin{aligned}\frac{\partial \mathbf{w}_t}{\partial \text{vec}\mathbf{B}} &= \frac{\partial\left(\mathbf{y}_t - \mathbf{B}\mathbf{z}_t\right)}{\partial \text{vec}\mathbf{B}} \\ &= -\frac{\partial\left(\mathbf{z}_t' \otimes \mathbf{I}_k\right)\text{vec}\mathbf{B}}{\partial \text{vec}\mathbf{B}} \\ &= -\mathbf{z}_t' \otimes \mathbf{I}_k; \end{aligned} \tag{8.5}$$

thus

$$\begin{aligned} \dot{l}'\left(\mathbf{B}\right) &= \frac{\partial l\left(\mathbf{B}\right)}{\partial \text{vec}\mathbf{B}} \\ &= \sum_{t=1}^{n} \mathbf{w}_t'\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}_t' \otimes \mathbf{I}_k\right) \\ &= \sum_{t=1}^{n} \left(\mathbf{z}_t' \otimes \mathbf{w}_t'\boldsymbol{\Sigma}^{-1}\right) \\ &= \left(\sum_{t=1}^{n} \left(\mathbf{z}_t \otimes \boldsymbol{\Sigma}^{-1}\mathbf{w}_t\right)\right)' \\ &= \left(\sum_{t=1}^{n} \left(\mathbf{z}_t \otimes \boldsymbol{\Sigma}^{-1}\right)\mathbf{w}_t\right)', \end{aligned}$$

and so

$$\dot{l}\left(\mathbf{B}\right) = \sum_{t=1}^{n} \left(\mathbf{z}_t \otimes \boldsymbol{\Sigma}^{-1}\right)\mathbf{w}_t.$$

Finally

$$\begin{aligned}\ddot{l}(\mathbf{B}) &= \frac{\partial \sum_{t=1}^{n}\left(\mathbf{z}_t \otimes \mathbf{\Sigma}^{-1}\right)\mathbf{w}_t}{\partial \text{vec}\mathbf{B}} \\ &= \sum_{t=1}^{n}\left(\mathbf{z}_t \otimes \mathbf{\Sigma}^{-1}\right)\frac{\partial \mathbf{w}_t}{\partial \text{vec}\mathbf{B}}.\end{aligned}$$

From (8.5), this continues as

$$\begin{aligned}\ddot{l}(\mathbf{B}) &= -\sum_{t=1}^{n}\left(\mathbf{z}_t \otimes \mathbf{\Sigma}^{-1}\right)\left(\mathbf{z}_t' \otimes \mathbf{I}_k\right) \\ &= -\sum_{t=1}^{n}\mathbf{z}_t\mathbf{z}_t' \otimes \mathbf{\Sigma}^{-1} \\ &= -\mathbf{Z}'\mathbf{Z} \otimes \mathbf{\Sigma}^{-1},\end{aligned}$$

with

$$\left\{E\left[-\ddot{l}(\mathbf{B})\right]\right\}^{-1} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1} \otimes \mathbf{\Sigma}.$$

More details on matrix differentiation are in the paper by Wiens on the course website.

8.2. Multivariate ARMAX models I

- Multivariate ARMAX models extend the univariate models in two ways. The dependent variable becomes a *vector* of dependent variables, and new inputs ("exogenous" variables) are allowed to enter the model at each time t.

- A first extension is the vector autoregressive model VAR(1):

$$\mathbf{x}_t = \boldsymbol{\alpha} + \boldsymbol{\Phi}\mathbf{x}_{t-1} + \mathbf{w}_t$$

which has k variables $x_{t,1}, ..., x_{t,k}$ related to their lag-1 values and

$$\begin{aligned} \mathbf{B}_{k\times(k+1)} &= (\boldsymbol{\alpha} \vdots \boldsymbol{\Phi}), \\ \mathbf{Y}' &= (\mathbf{x}_2, \cdots, \mathbf{x}_n), \\ \mathbf{Z}' &= (\mathbf{z}_2, \cdots, \mathbf{z}_n), \end{aligned}$$

for

$$\mathbf{z}_t = \begin{pmatrix} 1 \\ \mathbf{x}_{t-1} \end{pmatrix}.$$

In the notation used above, $r = k + 1$. The VAR(2) model is

$$\mathbf{x}_t = \boldsymbol{\alpha} + \boldsymbol{\Phi}_1 \mathbf{x}_{t-1} + \boldsymbol{\Phi}_2 \mathbf{x}_{t-2} + \mathbf{w}_t.$$

- The VAR(p) is

$$\begin{aligned} \mathbf{x}_t &= \boldsymbol{\alpha} + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \mathbf{x}_{t-j} + \mathbf{w}_t, \text{ where} \\ \mathbf{x}_t &= \left(x_{t,1}, x_{t,2}, \cdots, x_{t,k}\right)' \end{aligned}$$

and each Φ_j is a $k \times k$ matrix. We write

$$\boldsymbol{\alpha} + \sum_{j=1}^{p} \boldsymbol{\Phi}_j \mathbf{x}_{t-j} = \mathbf{B}\mathbf{z}_t,$$

$$\text{for } \mathbf{z}_t = \begin{pmatrix} 1 \\ \mathbf{x}_{t-1} \\ \vdots \\ \mathbf{x}_{t-p} \end{pmatrix} : (kp+1) \times 1,$$

$$\text{and } \mathbf{B} = (\boldsymbol{\alpha}, \boldsymbol{\Phi}_1, \cdots, \boldsymbol{\Phi}_p) : k \times (kp+1).$$

Then

$$\begin{aligned}\left(\mathbf{x}_{p+1},\cdots,\mathbf{x}_n\right) &= \mathbf{B}\begin{pmatrix}1 & 1 & \cdots & 1\\ \mathbf{x}_p & \mathbf{x}_{p+1} & \cdots & \mathbf{x}_{n-1}\\ \vdots & \vdots & \cdots & \vdots\\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_{n-p}\end{pmatrix}+\mathbf{W}',\\ \text{i.e. } \mathbf{Y}' &= \mathbf{B}\mathbf{Z}'+\mathbf{W}',\end{aligned}$$

where

$$\begin{aligned}\mathbf{Y}' &= \left(\mathbf{x}_{p+1},\cdots,\mathbf{x}_n\right):k\times(n-p),\\ \mathbf{Z}' &= \left(\mathbf{z}_{p+1},\cdots,\mathbf{z}_n\right):(kp+1)\times(n-p),\\ \mathbf{W}' &= \left(\mathbf{w}_{p+1},\cdots,\mathbf{w}_n\right):k\times(n-p).\end{aligned}$$

The conditional (on $\mathbf{x}_1,\cdots,\mathbf{x}_p$) likelihood is

$$\begin{aligned}&\prod_{t=p+1}^{n} p\left(\mathbf{x}_t|\mathbf{x}_{t-1},\cdots,\mathbf{x}_{p+1},\mathbf{x}_p,..,\mathbf{x}_1\right)\\ =\ &\prod_{t=p+1}^{n}\left\{\begin{array}{c}(2\pi)^{-k/2}\left|\mathbf{\Sigma}\right|^{-1/2}.\\ \exp-\frac{1}{2}\left(\mathbf{x}_t-\mathbf{B}\mathbf{z}_t\right)'\mathbf{\Sigma}^{-1}\left(\mathbf{x}_t-\mathbf{B}\mathbf{z}_t\right)\end{array}\right\}\end{aligned}$$

with log-likelihood given by (8.1). Then the asymptotic theory above applies: the MLE of $\mathbf{B}$ is given by (8.2), and this is asymptotically normally distributed with mean $\mathbf{B}$ and covariance

$(\mathbf{Z}'\mathbf{Z})^{-1} \otimes \Sigma_w$. The MLE of Σ_w is

$$\begin{aligned}
\hat{\Sigma}_w &= \frac{1}{n-p} \sum_{t=p+1}^{n} \left(\mathbf{x}_t - \hat{\mathbf{B}}\mathbf{z}_t\right)\left(\mathbf{x}_t - \hat{\mathbf{B}}\mathbf{z}_t\right)' \\
&= \frac{1}{n-p} \sum_{t=p+1}^{n} \hat{\mathbf{w}}_t \hat{\mathbf{w}}_t' \\
&= \frac{1}{n-p} \hat{\mathbf{W}}'\hat{\mathbf{W}},
\end{aligned}$$

where $\hat{\mathbf{w}}_t = \mathbf{x}_t - \hat{\mathbf{B}}\mathbf{z}_t = \mathbf{x}_t - \hat{\boldsymbol{\alpha}} - \sum_{j=1}^{p} \hat{\boldsymbol{\Phi}}_j \mathbf{x}_{t-j}$ is the t^{th} residual.

- The default method in R is to first fit

$$\begin{aligned}
\dot{\mathbf{x}}_t &= \sum_{j=1}^{p} \boldsymbol{\Phi}_j \dot{\mathbf{x}}_{t-j} + \mathbf{w}_t, \text{ where} \\
\dot{\mathbf{x}}_t &= \mathbf{x}_t - \bar{\mathbf{x}},
\end{aligned}$$

and then compute

$$\hat{\boldsymbol{\alpha}} = \left(\mathbf{I}_k - \sum_{j=1}^{p} \hat{\boldsymbol{\Phi}}_j\right) \bar{\mathbf{x}}.$$

9. Multivariate ARMAX models II

- Example 5.9: Cardiovascular mortality (x_1), temperature (x_2), and particulate levels (x_3) will be modelled in terms of their own history (and an intercept). Thus

$$\begin{aligned}\mathbf{x}_t &= \begin{pmatrix} x_{t,1} \\ x_{t,2} \\ x_{t,3} \end{pmatrix} = \boldsymbol{\alpha} + \boldsymbol{\Phi} \begin{pmatrix} x_{t-1,1} \\ x_{t-1,2} \\ x_{t-1,3} \end{pmatrix} + \mathbf{w}_t \\ &= \mathbf{B}\mathbf{z}_t + \mathbf{w}_t;\end{aligned}$$

$$\mathbf{B}_{3\times 4} = (\boldsymbol{\alpha} \vdots \boldsymbol{\Phi}),\ \mathbf{x}_t = \begin{pmatrix} x_{t,1} \\ x_{t,2} \\ x_{t,3} \end{pmatrix},\ \mathbf{z}_t = \begin{pmatrix} 1 \\ x_{t-1,1} \\ x_{t-1,2} \\ x_{t-1,3} \end{pmatrix}.$$

  Thus, e.g.

$$x_{t,2} = \alpha_2 + \phi_{21}x_{t-1,1} + \phi_{22}x_{t-1,2} + \phi_{23}x_{t-1,3} + w_{t,2}.$$

- We fit this VAR(1) model to detrended Mortality, centred Temperature and Particulates. See R code on website.

- For a VAR(p), to form the $\mathbf{Z}$ matrix, start with the full $n \times k$ data matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}.$$

More simply, the columns of $\mathbf{X}$ are the k series in the original data. Then, if there is no intercept,

$$\mathbf{Z}_{(n-p)\times pk} = \begin{pmatrix} \mathbf{X}\,[p:(n-1),] & \vdots & \mathbf{X}\,[(p-1):(n-2),] \\ & \vdots \cdots \vdots & \mathbf{X}\,[1:(n-p),] \end{pmatrix}$$

If there is an intercept, add a column of ones at the front. For $p = 2$ see the code on the website.

- The methods described above are "ols" in R, where the default method is instead "yule-walker". Unless p is specified it is chosen automatically by the AIC criterion. In this case using the defaults results in $p = 8$; AIC + "ols" results in $p = 8$. Typically BIC is more parsimonious; here it results in $p = 2$ with either estimation method.

- The next extension is to the VARMA(p,q) model. This is a k-dimensional series $\{\mathbf{x}_t\}$ which is stationary, and

$$\mathbf{x}_t = \boldsymbol{\alpha} + \sum_{j=1}^{p} \Phi_j \mathbf{x}_{t-j} + \mathbf{w}_t + \sum_{l=1}^{q} \Theta_l \mathbf{w}_{t-l}.$$

  The matrices Φ_j and Θ_l are $k \times k$. More generally,

$$\boldsymbol{\alpha} = \left(\mathbf{I}_k - \sum_{j=1}^{p} \Phi_j \right) \mu_x$$

  can be replaced by a vector of "exogenous" inputs of the form $\Gamma \mathbf{u}_t$, where Γ is $k \times r$ and $\mathbf{u}_t$ is $r \times 1$:

$$\mathbf{x}_t = \Gamma \mathbf{u}_t + \sum_{j=1}^{p} \Phi_j \mathbf{x}_{t-j} + \mathbf{w}_t + \sum_{l=1}^{q} \Theta_l \mathbf{w}_{t-l}.$$

  This results in the ARMAX model.

- Example: One might extend a model in which Mortality ($X_{\cdot,1}$ above; now call it M) is AR(2):

$$M_t = \phi_1 M_{t-1} + \phi_2 M_{t-2} + w_t$$

to one in which a linear trend, temperature and particulates enter as well:

$$\begin{aligned} & M_t \\ = & \left\{\gamma_0 + \gamma_1 t + \gamma_2 T_{t-1} + \gamma_3 T_{t-1}^2 + \gamma_4 P_t + \gamma_5 P_{t-4}\right\} \\ & + \phi_1 M_{t-1} + \phi_2 M_{t-2} + w_t, \end{aligned}$$

which is ARMAX(2,0) with

$$\begin{aligned} \Gamma &= (\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5), \\ \mathbf{u}_t' &= \left(1, t, T_{t-1}, T_{t-1}^2, P_t, P_{t-4}\right). \end{aligned}$$

Estimation is as before - the only change is that

$$\begin{aligned} \mathbf{z}_t &= \begin{pmatrix} 1 \\ M_{t-1} \\ M_{t-2} \end{pmatrix} -> \begin{pmatrix} \mathbf{u}_t \\ M_{t-1} \\ M_{t-2} \end{pmatrix}, \\ \mathbf{B} &= (\alpha \vdots \phi_1 \vdots \phi_2) -> (\Gamma \vdots \phi_1 \vdots \phi_2). \end{aligned}$$

- Define AR and MA operators

$$\begin{aligned} \Phi(B) &= \mathbf{I}_k - \sum_{j=1}^{p} \Phi_j B^j, \\ \Theta(B) &= \mathbf{I}_k + \sum_{l=1}^{q} \Theta_l B^l. \end{aligned}$$

In terms of these the zero-mean VARMA(p,q) model is

$$\mathbf{\Phi}(B)\,\mathbf{x}_t = \mathbf{\Theta}(B)\,\mathbf{w}_t.$$

Here B is a scalar, so that these are matrices of polynomials. Thus the determinant $|\mathbf{\Phi}(B)|$ is also a polynomial in B. The condition for stationarity ("causal" model) is that

$$\det \mathbf{\Phi}(B) = 0 \Rightarrow |B| > 1.$$

We use

$$\|\mathbf{\Psi}\| = \sqrt{tr\left(\mathbf{\Psi}'\mathbf{\Psi}\right)} = \sqrt{\sum_{j,k} \psi_{jk}^2}$$

for the matrix norm. (Note this is the Euclidean vector norm $\|vec\mathbf{\Psi}\|$.) Stationarity is equivalent to linearity:

$$\mathbf{x}_t = \mathbf{\Psi}(B)\,\mathbf{w}_t$$

for

$$\mathbf{\Psi}(B) = \sum_{j=0}^{\infty} \mathbf{\Psi}_j B^j, \text{ with } \sum_{j=0}^{\infty} \left\|\mathbf{\Psi}_j\right\|^2 < \infty.$$

Similarly, invertibility requires all roots of $\det \Theta(B) = 0$ to lie outside the unit circle in the complex plane. In this case

$$\mathbf{w}_t = \Pi(B)\,\mathbf{x}_t$$

for

$$\Pi(B) = \mathbf{I}_k - \sum_{l=1}^{\infty} \Pi_l B^l, \text{ with } \sum_{l=1}^{\infty} \|\Pi_l\| < \infty.$$

If the model is causal, then the autocovariance function is

$$\begin{aligned}
\Gamma(h) &\stackrel{def}{=} \text{cov}\left[\mathbf{x}_{t+h}, \mathbf{x}_t\right] \\
&= \text{cov}\left[\sum_{j=0}^{\infty} \Psi_j \mathbf{w}_{t+h-j}, \sum_{l=0}^{\infty} \Psi_l \mathbf{w}_{t-l}\right] \\
&= \sum_{j,l=0}^{\infty} \Psi_j \text{cov}\left[\mathbf{w}_{t+h-j}, \mathbf{w}_{t-l}\right] \Psi_l' \\
&= \sum_{j,l=0}^{\infty} \Psi_j I(j = l+h)\,\Sigma_w \Psi_l' \\
&= \sum_{l=0}^{\infty} \Psi_{l+h} \Sigma_w \Psi_l'.
\end{aligned}$$

Note that $\Gamma(-h) = \Gamma'(h)$.

- For an MA(q) process,

$$\Gamma(h) = \begin{cases} \sum_{l=0}^{q-h} \Theta_{l+h} \Sigma_w \Theta_l', & 0 \leq h \leq q, \\ 0, & h > q. \end{cases}$$

- The Yule-Walker equations for a VAR(p) model are (assigned)

$$\begin{aligned} \Gamma(h) &= \sum_{j=1}^{p} \Phi_j \Gamma(h-j), \ h = 1, 2.3, \cdots \\ \Gamma(0) &= \sum_{j=1}^{p} \Phi_j \Gamma(-j) + \Sigma_w. \end{aligned}$$

  The second of these yields an estimate

$$\hat{\Sigma}_w = \hat{\Gamma}(0) - \sum_{j=1}^{p} \hat{\Phi}_j \hat{\Gamma}(-j),$$

  which R returns as the component $var.pred in the ar() output. This can be useful for model identification - as a function of p it should decrease until the 'correct' p is reached, and then stabilize.

- If the model is ARMA(p,q): $\Phi(B)\,\mathbf{x}_t = \Theta(B)\,\mathbf{w}_t$ then to invert it, i.e. to write $\mathbf{w}_t = \Pi(B)\,\mathbf{x}_t$, we equate

$$\Pi(B) = \Theta(B)^{-1}\,\Phi(B)$$

to a series $\mathbf{I}_k - \sum_{l=1}^{\infty} \Pi_l B^l$ by solving

$$\Theta(B)\left(\mathbf{I}_k - \sum_{l=1}^{\infty} \Pi_l B^l\right) = \Phi(B)$$

and equating coefficients of equal powers of B.

- Parameter identifiability is a serious problem. Here is an example. The VAR(1) model

$$\begin{pmatrix} x_{t,1} \\ x_{t,2} \end{pmatrix} = \begin{pmatrix} 0 & \phi \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_{t-1,1} \\ x_{t-1,2} \end{pmatrix} + \begin{pmatrix} w_{t,1} \\ w_{t,2} \end{pmatrix}$$

can also be written as the ARMA(1,1) model (for *any* θ):

$$\begin{aligned} \begin{pmatrix} x_{t,1} \\ x_{t,2} \end{pmatrix} &= \begin{pmatrix} 0 & \phi+\theta \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_{t-1,1} \\ x_{t-1,2} \end{pmatrix} \\ &\quad + \begin{pmatrix} 0 & -\theta \\ 0 & 0 \end{pmatrix} \begin{pmatrix} w_{t-1,1} \\ w_{t-1,2} \end{pmatrix} + \begin{pmatrix} w_{t,1} \\ w_{t,2} \end{pmatrix}. \end{aligned}$$

The AR form is

$$\mathbf{\Phi}(B)\,\mathbf{x}_t = \mathbf{w}_t;\ \mathbf{\Phi}(B) = \begin{pmatrix} 1 & -\phi B \\ 0 & 1 \end{pmatrix};$$

the ARMA is

$$\begin{aligned} \mathbf{\Phi}_*(B)\,\mathbf{x}_t &= \mathbf{\Theta}_*(B)\,\mathbf{w}_t; \\ \mathbf{\Phi}_*(B) &= \begin{pmatrix} 1 & -(\phi+\theta)B \\ 0 & 1 \end{pmatrix}, \\ \mathbf{\Theta}_*(B) &= \begin{pmatrix} 1 & -\theta B \\ 0 & 1 \end{pmatrix}. \end{aligned}$$

We have

$$\mathbf{\Theta}_*^{-1}(B)\,\mathbf{\Phi}_*(B) = \mathbf{\Phi}(B).$$

There are conditions which will circumvent such problems - see the discussion in the text. A common remedy seems to be to avoid multivariate MA models altogether. Indeed, R doesn't seem to have a function to fit them.

- Similar to the univariate case, one can define the lag h Partial Autocorrelation matrix as the matrix of correlations between the elements of the

residuals $\mathbf{x}_{t+h} - \hat{\mathbf{x}}_{t+h}$ and $\mathbf{x}_t - \hat{\mathbf{x}}_t$, where

$$\begin{aligned} \hat{\mathbf{x}}_{t+h} &= \sum_{j=1}^{h-1} \mathbf{A}_j \mathbf{x}_{t+h-j}, \\ \hat{\mathbf{x}}_t &= \sum_{j=1}^{h-1} \mathbf{B}_j \mathbf{x}_{t+j} \end{aligned}$$

are the best MSE predictors minimizing $E\left[\|\mathbf{x}_{t+h} - \hat{\mathbf{x}}_{t+h}\|^2\right]$ and $E\left[\|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2\right]$, respectively.

- R computes (by the pacf() function) and S&S discuss, a variation of this known as the Partial Autoregression matrix. To compute, fit an AR(h) model to the data, with coefficient matrices $\Phi_{jh}, j = 1, ..., h$. Then the lag h Partial Autoregression matrix is Φ_{hh}. If the true model is AR(p) , then $\Phi_{pp} = \Phi_p$ and $\Phi_{qq} = \mathbf{0}$ for $q > p$. This is as in the univariate case of the Partial Autocorrelation function. However, these are not correlation matrices.

- See the R code for these matrices in the mortality data, for which BIC has previously chosen AR(2). Note that the PAR matrices are negligible past lag 2.

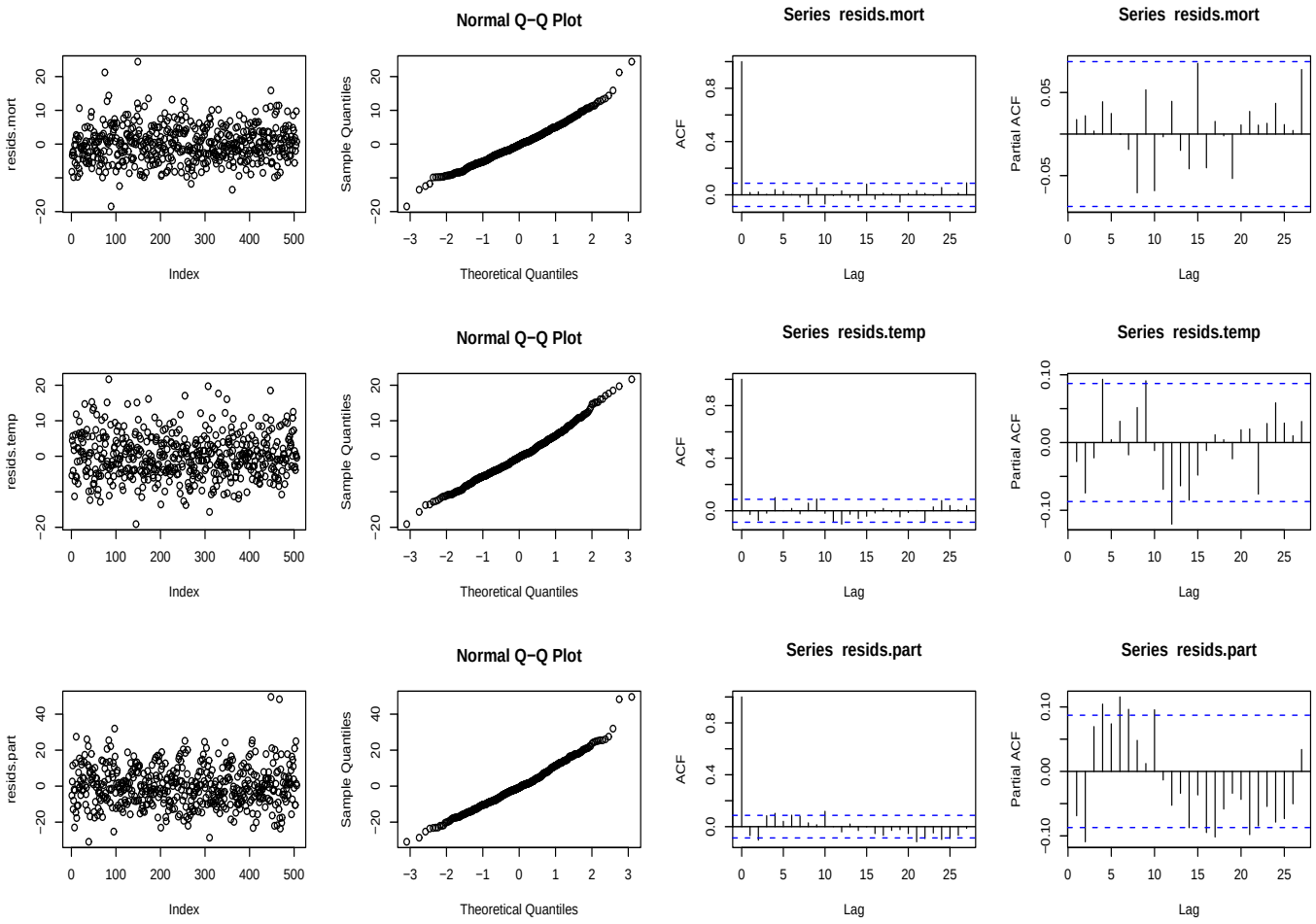

Figure 9.1. Residuals from VAR(2) fit to Mortality data.

- Prediction in VAR(p) models is straightforward. The m-step ahead forecasts are the conditional expectations

$$\begin{aligned}
\mathbf{x}_{t+m}^t &= E\left[\mathbf{x}_{t+m}|\mathbf{x}^t\right] \\
&= E\left[\sum_{j=1}^{p} \mathbf{\Phi}_j \mathbf{x}_{t+m-j} + \mathbf{w}_t|\mathbf{x}^t\right] \\
&= \sum_{j=1}^{p} \mathbf{\Phi}_j \mathbf{x}_{t+m-j}^t,
\end{aligned}$$

  where $\mathbf{x}_{t+m-j}^t = \mathbf{x}_{t+m-j}$ if $j \geq m$; otherwise it is computed recursively. Then

$$\mathbf{x}_{t+m}^t = \sum_{j=1}^{p} \hat{\mathbf{\Phi}}_j \hat{\mathbf{x}}_{t+m-j}^t.$$

- See the R code for 2-step ahead predictions in the Mortality data; compare with S&S Examples 5.10, 5.11. R does not give standard errors of the predictions, but these can be gotten by a state-space approach to VARMA modelling.

10. State-space models - Introduction

10.1. Basic formulation

- The "state space" or "dynamic linear" model provides a unified method for treating a large number of problems in time series. It evolves in the following way. The "state equation" is similar to VAR(1):

$$\mathbf{x}_t = \mathbf{\Phi}\mathbf{x}_{t-1} + \mathbf{w}_t$$

with $\mathbf{x}_t, \mathbf{w}_t : p \times 1$ and $\mathbf{\Phi} : p \times p$. Here

$$\mathbf{w}_1, ..., \mathbf{w}_t, ... \overset{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{Q}).$$

A starting point $\mathbf{x}_0 \sim N(\mu_0, \mathbf{\Sigma}_0)$ (independent of $\{\mathbf{w}_t\}$) is specified. Note that $\mathbf{w}_t$ is independent of $\mathbf{x}^{t-1} = \{\mathbf{x}_s\}_{s<t}$. In fact

$$\mathbf{x}_s = \mathbf{\Phi}^s\mathbf{x}_0 + \sum_{l=1}^{s} \mathbf{\Phi}^{s-l}\mathbf{w}_l$$

depends only on $\{\mathbf{x}_0, \mathbf{w}_l\}_{l \leq s}$, and so is independent of later noise $\mathbf{w}_t$.

- However, we don't observe $\{\mathbf{x}_t\}$; rather we observe

$$\mathbf{y}_t = \mathbf{A}_t \mathbf{x}_t + \mathbf{v}_t,$$

(the "observation equation") where $\mathbf{y}_t$ and $\mathbf{v}_t$ are $q \times 1$ (q can be larger or smaller than p), hence the "observation" or "measurement" matrix $\mathbf{A}_t$ is $q \times p$, and

$$\mathbf{v}_1, ..., \mathbf{v}_t, ... \overset{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{R}).$$

For the moment assume that the $\mathbf{w}$'s and $\mathbf{v}$'s are uncorrelated with each other (hence $\mathbf{v}^t$ is uncorrelated with $\mathbf{x}^t$ and $\mathbf{w}_t$ is independent of $\mathbf{y}^{t-1}$).

  - Motivation: $\mathbf{x}_t$ the true position of an object in space, $\mathbf{y}_t$ the information about the position which is received by a tracking station.

  - Exogenous inputs can be incorporated:

$$\begin{aligned} \mathbf{x}_t &= \mathbf{\Phi} \mathbf{x}_{t-1} + \mathbf{\Upsilon} \mathbf{u}_t + \mathbf{w}_t, \\ \mathbf{y}_t &= \mathbf{A}_t \mathbf{x}_t + \mathbf{\Gamma} \mathbf{u}_t + \mathbf{v}_t, \end{aligned}$$

("Upsilon"; υ) with $\mathbf{u}_t : r \times 1$ *fixed*, i.e. nonrandom.

– "Lag one" is not necessarily implied. For instance

$$\mathbf{X}_t = \begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-1} \\ \vdots \\ \mathbf{x}_{t-m+1} \end{pmatrix}, \mathbf{W}_t = \begin{pmatrix} \mathbf{w}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix} : pm \times 1,$$

$$\Phi = \begin{pmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_{m-1} & \Phi_m \\ \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{pmatrix} : pm \times pm,$$

in the model

$$\mathbf{X}_t = \Phi \mathbf{X}_{t-1} + \mathbf{W}_t$$

yields "$\mathbf{x}_{t-k} = \mathbf{x}_{t-k}$" for $k = 1, ..., m-1$ and

$$\mathbf{x}_t = \sum_{k=1}^{m} \Phi_k \mathbf{x}_{t-k} + \mathbf{w}_t.$$

Now

$$\text{cov}\,[\mathbf{W}_t] = \mathbf{Q}_{p\times p} \oplus \mathbf{0}_{(m-1)p\times(m-1)p}$$

and the observation matrix is $(\mathbf{A}_t \vdots \mathbf{0} \vdots \cdots \vdots \mathbf{0})$.

- Example 1: If $\mathbf{x}_t = \mathbf{\Phi}\mathbf{x}_{t-1} + \mathbf{w}_t$ and $\mathbf{x}_t$ itself is observed with error (yielding $\mathbf{y}_t$), then $\mathbf{A}_t = \mathbf{I}$. The special case

$$\begin{aligned} x_t &= \phi x_{t-1} + w_t, \\ y_t &= x_t + v_t, \end{aligned}$$

gives an AR(1) model with observation noise.

- Example 2: Suppose x_t = global temperature is being studied. One observes y_{t1} = average air temperature at land based stations and y_{t2} = average air temperature at marine based stations. Modelling x_t as a random walk, and assuming that y_{t1} and y_{t2} are, apart from random error, measuring the same thing, gives

$$\begin{aligned} x_t &= x_{t-1} + w_t, \\ \begin{pmatrix} y_{t1} \\ y_{t2} \end{pmatrix} &= \mathbf{y}_t = \begin{pmatrix} 1 \\ 1 \end{pmatrix} x_t + \mathbf{v}_t. \end{aligned}$$

So $\Phi = 1, \mathbf{A}_t = (1,1)'$ $(p = 1, q = 2)$ and the parameters in $Q = \sigma_w^2$ and $\mathbf{R} = \text{cov}[\mathbf{v}_t]$ are to be estimated from the data.

11. Filtering, Smoothing, Forecasting

- Define $\mathbf{y}^s = \{\mathbf{y}_t\}_{1 \leq t \leq s}$, the history of the observed process. Then the minimum mse forecast is

$$\mathbf{x}_t^s = E\left[\mathbf{x}_t | \mathbf{y}^s\right].$$

Define

$$\begin{aligned} \mathbf{P}_{t_1,t_2}^s &= E\left[\left(\mathbf{x}_{t_1} - \mathbf{x}_{t_1}^s\right)\left(\mathbf{x}_{t_2} - \mathbf{x}_{t_2}^s\right)'\right], \\ \mathbf{P}_{t,t}^s &= \mathbf{P}_t^s. \end{aligned}$$

Then $\mathbf{P}_t^s$ is the variance/covariance matrix of the forecast error.

  - We repeatedly use the facts that, for arbitrary r.v.s or r.vecs X, Y: $Y - E\left[Y|X\right]$ has a mean of zero ("Double Expectation Theorem") and is uncorrelated with X. Thus $\mathbf{x}_t - \mathbf{x}_t^s$ has a mean of $\mathbf{0}$ and is uncorrelated with $\mathbf{y}^s$.

  - For Gaussian noise (as is assumed here), $\mathbf{x}_t - \mathbf{x}_t^s$ is *independent* of $\mathbf{y}^s$; thus

$$\mathbf{P}_{t_1,t_2}^s = E\left[\left(\mathbf{x}_{t_1} - \mathbf{x}_{t_1}^s\right)\left(\mathbf{x}_{t_2} - \mathbf{x}_{t_2}^s\right)' | \mathbf{y}^s\right].$$

- The Kalman Filter was derived in order to obtain $\mathbf{x}_t^s$. With $s < t$ this is a *forecasting* or *prediction* problem, with $s = t$ it is a *filtering* problem (since then $\mathbf{x}_t^t$ is presented as a linear combination $\mathbf{x}_t^t = \sum_{s=1}^t \mathbf{B}_s \mathbf{y}_s$) and with $s > t$ it is a *smoothing* problem.

- **Kalman Forecasting**: Suppose that, for $t = 1, ..., n$,

$$\begin{aligned} \mathbf{x}_t &= \Phi \mathbf{x}_{t-1} + \Upsilon \mathbf{u}_t + \mathbf{w}_t, \qquad (11.1) \\ \mathbf{y}_t &= \mathbf{A}_t \mathbf{x}_t + \Gamma \mathbf{u}_t + \mathbf{v}_t, \end{aligned}$$

with given initial values $\mathbf{x}_0^0 = \mu_0$, $P_0^0 = \Sigma_0$. Recall that $\mathbf{u}_t$ is non-random; also $\mathbf{w}_t$ is independent of $\mathbf{y}^{t-1}$. Thus (substituting the state equation (11.1) into "$\mathbf{x}_t^{t-1} = E\left[\mathbf{x}_t | \mathbf{y}^{t-1}\right]$") the one-step ahead forecast is

$$\mathbf{x}_t^{t-1} = \Phi \mathbf{x}_{t-1}^{t-1} + \Upsilon \mathbf{u}_t, \qquad (11.2)$$

with

$$\begin{aligned}
\mathbf{P}_t^{t-1} &= E\left[\left(\mathbf{x}_t - \mathbf{x}_t^{t-1}\right)\left(\mathbf{x}_t - \mathbf{x}_t^{t-1}\right)'\right] \\
&= E\left[\begin{array}{c} \left\{\mathbf{\Phi}\left(\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^{t-1}\right) + \mathbf{w}_t\right\} \\ \cdot\left\{\mathbf{\Phi}\left(\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^{t-1}\right) + \mathbf{w}_t\right\}' \end{array}\right] \\
&= \mathbf{\Phi} E\left[\left(\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^{t-1}\right)\left(\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^{t-1}\right)'\right]\mathbf{\Phi}' \\
&\quad + E\left[\mathbf{w}_t \mathbf{w}_t'\right]; \text{ thus} \\
\mathbf{P}_t^{t-1} &= \mathbf{\Phi}\mathbf{P}_{t-1}^{t-1}\mathbf{\Phi}' + \mathbf{Q}. \qquad (11.3)
\end{aligned}$$

Similarly, forecasts into the future ($t > n$) are given by combining $\mathbf{x}_n^n$, P_n^n (obtained below) with

$$\begin{aligned}
\mathbf{x}_t^n &= \mathbf{\Phi}\mathbf{x}_{t-1}^n + \mathbf{\Upsilon}\mathbf{u}_t \ (t > n), \qquad (11.4) \\
\mathbf{P}_t^n &= \mathbf{\Phi}\mathbf{P}_{t-1}^n\mathbf{\Phi}' + \mathbf{Q} \ (t > n). \qquad (11.5)
\end{aligned}$$

- It is shown below that the filtering problem of obtaining $\left\{\mathbf{x}_t^t, \mathbf{P}_t^t\right\}$ with $t \leq n$ can be reduced to that of obtaining $\left\{\mathbf{x}_t^{t-1}, \mathbf{P}_t^{t-1}\right\}$; by the above this in turn is reduced to that of obtaining $\left\{\mathbf{x}_{t-1}^{t-1}, \mathbf{P}_{t-1}^{t-1}\right\}$. Note that $\left\{\mathbf{x}_t^t, \mathbf{P}_t^t\right\}$ with $t = n$ also initializes (11.4) and (11.5).

- Define innovations

$$\begin{aligned}\varepsilon_t &= \mathbf{y}_t - \mathbf{y}_t^{t-1} \\ &= \mathbf{y}_t - E\left[\mathbf{y}_t|\mathbf{y}^{t-1}\right] \\ &= \mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t^{t-1} - \mathbf{\Gamma u}_t.\end{aligned}$$

Note that for $s < t$,

$$\text{cov}\left[\varepsilon_t, \mathbf{y}_s\right] = E\left[\varepsilon_t \mathbf{y}_s'\right] = E\left[\left(\mathbf{y}_t - \mathbf{y}_t^{t-1}\right)\mathbf{y}_s'\right] = \mathbf{0},$$

since $\mathbf{y}_t - \mathbf{y}_t^{t-1}$ is uncorrelated with $\mathbf{y}^{t-1}$. Thus **the innovations ε_t are independent of y^{t-1} (and hence of ε_s for $s < t$)**.

- We obtain the joint conditional distribution of $\binom{\mathbf{x}_t}{\varepsilon_t}|\mathbf{y}^{t-1}$. The marginal distribution $\mathbf{x}_t|\mathbf{y}^{t-1}$ is $N\left(\mathbf{x}_t^{t-1}, \mathbf{P}_t^{t-1}\right)$ with these given by (11.2) and (11.3). That of $\varepsilon_t|\mathbf{y}^{t-1}$ is $N\left(\mathbf{0}, \mathbf{\Sigma}_t\right)$, with

$$\begin{aligned}\mathbf{\Sigma}_t &= \text{cov}\left[\varepsilon_t|\mathbf{y}^{t-1}\right] = \text{cov}\left[\varepsilon_t\right] \\ &= \text{cov}\left[\mathbf{A}_t\left(\mathbf{x}_t - \mathbf{x}_t^{t-1}\right) + \mathbf{v}_t\right] \\ &= \mathbf{A}_t\text{cov}\left[\mathbf{x}_t - \mathbf{x}_t^{t-1}\right]\mathbf{A}_t' + \text{cov}\left[\mathbf{v}_t\right] \\ &= \mathbf{A}_t\mathbf{P}_t^{t-1}\mathbf{A}_t' + \mathbf{R}.\end{aligned}$$

- Since the data are Gaussian the joint distribution is also normal; all that is left is the determination of the conditional covariance:

$$\begin{aligned}\text{cov}\left[\mathbf{x}_t, \varepsilon_t | \mathbf{y}^{t-1}\right] &= E\left[\mathbf{x}_t \varepsilon_t' | \mathbf{y}^{t-1}\right] \\ &= E\left[\mathbf{x}_t \left(\mathbf{x}_t - \mathbf{x}_t^{t-1}\right)' \mathbf{A}_t' | \mathbf{y}^{t-1}\right] \\ &\quad + E\left[\mathbf{x}_t \mathbf{v}_t' | \mathbf{y}^{t-1}\right] \\ &= \mathbf{P}_t^{t-1} \mathbf{A}_t'.\end{aligned}$$

Thus

$$\begin{aligned}\binom{\mathbf{x}_t}{\varepsilon_t} | \mathbf{y}^{t-1} &\sim N\left(\binom{\mathbf{x}_t^{t-1}}{\mathbf{0}}, \begin{pmatrix} \mathbf{P}_t^{t-1} & \mathbf{P}_t^{t-1}\mathbf{A}_t' \\ \mathbf{A}_t\mathbf{P}_t^{t-1} & \boldsymbol{\Sigma}_t \end{pmatrix}\right), \\ \text{with } \boldsymbol{\Sigma}_t &= \mathbf{A}_t \mathbf{P}_t^{t-1} \mathbf{A}_t' + \mathbf{R}.\end{aligned}$$

Now recall the general multivariate normal distribution theory result: if

$$\binom{\mathbf{x}}{\mathbf{y}} \sim N\left(\binom{\mu_\mathbf{x}}{\mu_\mathbf{y}}, \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}} \\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}} \end{pmatrix}\right),$$

then

$$\mathbf{x}|\mathbf{y} \sim N\left(\begin{array}{c} \mu_\mathbf{x} + \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\left(\mathbf{y} - \mu_\mathbf{y}\right), \\ \boldsymbol{\Sigma}_{\mathbf{xx}} - \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\boldsymbol{\Sigma}_{\mathbf{yx}} \end{array}\right).$$

Using this, we get that

$$\mathbf{x}_t|\varepsilon_t, \mathbf{y}^{t-1} \sim N\left(\begin{array}{c} \mathbf{x}_t^{t-1} + \mathbf{P}_t^{t-1}\mathbf{A}_t'\mathbf{\Sigma}_t^{-1}\varepsilon_t, \\ \mathbf{P}_t^{t-1} - \mathbf{P}_t^{t-1}\mathbf{A}_t'\mathbf{\Sigma}_t^{-1}\mathbf{A}_t\mathbf{P}_t^{t-1} \end{array} \right).$$

Since conditioning on $\varepsilon_t, \mathbf{y}^{t-1}$ is equivalent to conditioning on $\mathbf{y}^t$, we have the **Kalman Filtering** equations:

$$\mathbf{x}_t^t = E\left[\mathbf{x}_t|\varepsilon_t, \mathbf{y}^{t-1}\right] = \mathbf{x}_t^{t-1} + \mathbf{P}_t^{t-1}\mathbf{A}_t'\mathbf{\Sigma}_t^{-1}\varepsilon_t,$$

i.e.

$$\begin{aligned} \mathbf{x}_t^t &= \mathbf{x}_t^{t-1} + \mathbf{K}_t\varepsilon_t, \text{ where} \\ \varepsilon_t &= \mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t^{t-1}, \\ \mathbf{K}_t &= \mathbf{P}_t^{t-1}\mathbf{A}_t'\mathbf{\Sigma}_t^{-1} = \mathbf{P}_t^{t-1}\mathbf{A}_t'\left(\mathbf{A}_t\mathbf{P}_t^{t-1}\mathbf{A}_t' + \mathbf{R}\right)^{-1}, \end{aligned}$$

and

$$\begin{aligned} \mathbf{P}_t^t &= \mathbf{P}_t^{t-1} - \mathbf{P}_t^{t-1}\mathbf{A}_t'\mathbf{\Sigma}_t^{-1}\mathbf{A}_t\mathbf{P}_t^{t-1} \\ &= \left(\mathbf{I} - \mathbf{K}_t\mathbf{A}_t\right)\mathbf{P}_t^{t-1}. \end{aligned}$$

- The determination of $\mathbf{x}_{t-1}^n$ for $t \leq n$ is "smoothing"; the idea is that the data collected at time t or later is used to go back and revise the predictions and filters obtained using incomplete data.

Students should go through the derivation of the following **Kalman Smoother**. For $t = n, n-1, ..., 1$ (so starting with $\mathbf{x}_n^n$ and $\mathbf{P}_n^n$):

$$\begin{aligned}
\mathbf{x}_{t-1}^n &= \mathbf{x}_{t-1}^{t-1} + \mathbf{J}_{t-1}\left(\mathbf{x}_t^n - \mathbf{x}_t^{t-1}\right), \\
\mathbf{P}_{t-1}^n &= \mathbf{P}_{t-1}^{t-1} + \mathbf{J}_{t-1}\left(\mathbf{P}_t^n - \mathbf{P}_t^{t-1}\right)\mathbf{J}_{t-1}', \text{ where} \\
\mathbf{J}_{t-1} &= \mathbf{P}_{t-1}^{t-1}\boldsymbol{\Phi}'\left[\mathbf{P}_t^{t-1}\right]^{-1}.
\end{aligned}$$

- Similarly, a lag-one covariance smoother can be computed. Recall that

$$\mathbf{P}_{t-1,t-2}^n = E\left[\left(\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^n\right)\left(\mathbf{x}_{t-2} - \mathbf{x}_{t-2}^n\right)'\right].$$

We have

$$\mathbf{P}_{n,n-1}^n = \left(\mathbf{I} - \mathbf{K}_n\mathbf{A}_n\right)\boldsymbol{\Phi}\mathbf{P}_{n-1}^{n-1},$$

and then for $t = n, n-1, ..., 2$:

$$\mathbf{P}_{t-1,t-2}^n = \mathbf{P}_{t-1}^{t-1}\mathbf{J}_{t-2}' + \mathbf{J}_{t-1}\left(\mathbf{P}_{t,t-1}^n - \boldsymbol{\Phi}\mathbf{P}_{t-1}^{t-1}\right)\mathbf{J}_{t-2}'.$$

12. Bayes Iterations; Computing

Bayes Iteration scheme. This holds in general; applications to Gaussian densities are in square brackets []. We write

$$\phi_p(\mathbf{z};\Sigma) = (2\pi)^{-p/2}\sqrt{|\Sigma|}\exp\left\{-\frac{1}{2}\mathbf{z}'\Sigma^{-1}\mathbf{z}\right\},$$

for the p-dimensional $N(0,\Sigma)$ density. Some more details are in the Meinhold & Singpurwalla paper on the course website. We repeatedly use the paradigm that

$$\begin{aligned} p(\text{nature}|\text{data}) &= \frac{p(N,D)}{p(D)} = \frac{p(D|N)p(N)}{p(D)} \\ &\propto p(D|N)p(N). \end{aligned}$$

The normalizing constant is obtained by integrating out N from $p(D|N)p(N)$.

- Start with a prior density $p_0(\mathbf{x}_0)$.
  $[= \phi_p(\mathbf{x}_0 - \mu_0; \Sigma_0)]$

- Update after observing $\mathbf{y}_0$:

$$p(\mathbf{x}_0|\mathbf{y}_0) \propto p(\mathbf{y}_0|\mathbf{x}_0)p_0(\mathbf{x}_0)$$

  $[p(\mathbf{y}_t|\mathbf{x}_t) = \phi_q(\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t; \mathbf{R}).]$

- Predict next state from

$$\begin{aligned} p(\mathbf{x}_1|\mathbf{y}_0) &= \int p(\mathbf{x}_1, \mathbf{x}_0|\mathbf{y}_0)d\mathbf{x}_0 \\ &= \int p(\mathbf{x}_1|\mathbf{x}_0)p(\mathbf{x}_0|\mathbf{y}_0)d\mathbf{x}_0. \end{aligned}$$

  Reason:

$$\begin{aligned} p(\mathbf{x}_1, \mathbf{x}_0|\mathbf{y}_0) &= \frac{p(\mathbf{y}_0|\mathbf{x}_1, \mathbf{x}_0)\, p(\mathbf{x}_1, \mathbf{x}_0)}{p(\mathbf{y}_0)} \\ &= \frac{p(\mathbf{y}_0|\mathbf{x}_0)\, p(\mathbf{x}_1|\mathbf{x}_0)\, p(\mathbf{x}_0)}{p(\mathbf{y}_0)} \text{ (why?)} \\ &= p(\mathbf{x}_0|\mathbf{y}_0)\, p(\mathbf{x}_1|\mathbf{x}_0). \end{aligned}$$

  $[p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \phi_p(\mathbf{x}_t - \Phi\mathbf{x}_{t-1}; \mathbf{Q})]$

- Update again, after observing $\mathbf{y}_1$:

$$\begin{aligned} p(\mathbf{x}_1|\mathbf{y}^1) &= p(\mathbf{x}_1|\mathbf{y}_1,\mathbf{y}^0) \\ &= \frac{p(\mathbf{y}_1,\mathbf{y}^0,\mathbf{x}_1)}{p\left(\mathbf{y}_1,\mathbf{y}^0\right)} \\ &= \frac{p(\mathbf{y}_1|\mathbf{y}^0,\mathbf{x}_1)p\left(\mathbf{y}^0,\mathbf{x}_1\right)}{p\left(\mathbf{y}_1,\mathbf{y}^0\right)} \\ &= \frac{p(\mathbf{y}_1|\mathbf{x}_1)p\left(\mathbf{x}_1|\mathbf{y}^0\right)p\left(\mathbf{y}^0\right)}{p\left(\mathbf{y}_1,\mathbf{y}^0\right)} \quad (12.1) \\ &= \frac{p(\mathbf{y}_1|\mathbf{x}_1)p\left(\mathbf{x}_1|\mathbf{y}^0\right)}{p\left(\mathbf{y}_1|\mathbf{y}^0\right)} \\ &\propto p(\mathbf{y}_1|\mathbf{x}_1)p(\mathbf{x}_1|\mathbf{y}^0). \end{aligned}$$

Reason for (12.1): $\mathbf{y}_1 = \mathbf{A}\mathbf{x}_1 + \mathbf{v}_1$ depends on $\left\{\mathbf{y}^0, \mathbf{x}_1\right\}$ only through $\mathbf{x}_1$. Reason for last line: Integrate both sides of second last line w.r.t. $\mathbf{x}_1$.

- Continue. In general, the updating formula - after observing $\mathbf{y}_t$ - is

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{y}^t) &\propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}^{t-1}) \\ [&= \phi_q\left(\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t; \mathbf{R}\right) p(\mathbf{x}_t|\mathbf{y}^{t-1})] \end{aligned}$$

and the one-step ahead prediction is obtained from

$$p(\mathbf{x}_t|\mathbf{y}^{t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})d\mathbf{x}_{t-1}$$

$$[= \int \phi_p\left(\mathbf{x}_t - \mathbf{\Phi}\mathbf{x}_{t-1}; \mathbf{Q}\right) p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})d\mathbf{x}_{t-1}].$$

- Thus, e.g., for Gaussian noise the last [] can be written

$$\begin{aligned} &\phi_p\left(\mathbf{x}_t - \mathbf{x}_t^{t-1}; \mathbf{P}_t^{t-1}\right) \\ = &\int \left\{ \begin{array}{c} \phi_p\left(\mathbf{x}_t - \mathbf{\Phi}\mathbf{x}_{t-1}; \mathbf{Q}\right) \\ \phi_p\left(\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^{t-1}; \mathbf{P}_{t-1}^{t-1}\right) \end{array} \right\} d\mathbf{x}_{t-1}. \end{aligned}$$

Comparing the exponents:

$$e^{-\frac{\left(\mathbf{x}_t-\mathbf{x}_t^{t-1}\right)'\left[\mathbf{P}_t^{t-1}\right]^{-1}\left(\mathbf{x}_t-\mathbf{x}_t^{t-1}\right)}{2}}$$

$$\propto \int e^{\left\{ \begin{array}{c} -\frac{(\mathbf{x}_t-\Phi\mathbf{x}_{t-1})'\left[\mathbf{P}_t^{t-1}\right]^{-1}(\mathbf{x}_t-\Phi\mathbf{x}_{t-1})}{2} \\ -\frac{\left(\mathbf{x}_{t-1}-\mathbf{x}_{t-1}^{t-1}\right)'\left[\mathbf{P}_{t-1}^{t-1}\right]^{-1}\left(\mathbf{x}_{t-1}-\mathbf{x}_{t-1}^{t-1}\right)}{2} \end{array} \right\}} d\mathbf{x}_{t-1}$$

gives the Kalman forecasting equations expressing $\left\{\mathbf{x}_t^{t-1}, \mathbf{P}_t^{t-1}\right\}$ in terms of $\left\{\mathbf{x}_{t-1}^{t-1}, \mathbf{P}_{t-1}^{t-1}\right\}$.

- To compute for Gaussian noise: Suppose that $\mathbf{y}^{t-1} = \{\mathbf{y}_1, ..., \mathbf{y}_{t-1}\}$ have been observed, and on the basis of this we have determined the distribution of

$$\mathbf{x}_{t-1}|\mathbf{y}^{t-1} \sim N\left(\mathbf{x}_{t-1}^{t-1}, \mathbf{P}_{t-1}^{t-1}\right).$$

(This is initialized with $\mathbf{x}_0^0 = \mu_0, \mathbf{P}_0^0 = \Sigma_0$.) Now we obtain another observation $\mathbf{y}_t$ and seek the distribution of

$$\mathbf{x}_t|\mathbf{y}^t \sim N\left(\mathbf{x}_t^t, \mathbf{P}_t^t\right).$$

This is given by the **filter**

$$\begin{aligned}
\mathbf{x}_t^t &= \mathbf{x}_t^{t-1} + \mathbf{K}_t\varepsilon_t, \text{ where} \\
\varepsilon_t &= \mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t^{t-1}, \\
\mathbf{K}_t &= \mathbf{P}_t^{t-1}\mathbf{A}_t'\boldsymbol{\Sigma}_t^{-1} = \mathbf{P}_t^{t-1}\mathbf{A}_t'\left(\mathbf{A}_t\mathbf{P}_t^{t-1}\mathbf{A}_t' + \mathbf{R}\right)^{-1},
\end{aligned}$$

and

$$\mathbf{P}_t^t = \left(\mathbf{I} - \mathbf{K}_t\mathbf{A}_t\right)\mathbf{P}_t^{t-1}.$$

It requires the computation of $\left(\mathbf{x}_t^{t-1}, \mathbf{P}_t^{t-1}\right)$; these are gotten through the **forecasts**

$$\begin{aligned}
\mathbf{x}_t^{t-1} &= \boldsymbol{\Phi}\mathbf{x}_{t-1}^{t-1} + \boldsymbol{\Upsilon}\mathbf{u}_t, \\
\mathbf{P}_t^{t-1} &= \boldsymbol{\Phi}\mathbf{P}_{t-1}^{t-1}\boldsymbol{\Phi}' + \mathbf{Q},
\end{aligned}$$

both of which are available.

- Typically, a **smother** is applied as well, giving $(\mathbf{x}_t^n, \mathbf{P}_t^n)$.

- Example 6.5. Simulate $n = 50$ values of $N(0, 1)$ white noise w_t, and then generate a random walk

$$\mu_t = \mu_{t-1} + w_t,$$

starting with $\mu_0 \sim N(0, 1)$. Then simulate $y_t = \mu_t + v_t$, with $v_t \overset{ind.}{\sim} N(0, 1)$:

```
set.seed(1)
num=50
w = rnorm(num+1,0,1)
v = rnorm(num,0,1)
mu = array0(cumsum(w), offset=0)
    # state: mu[0],mu[1],...,mu[50]
y=mu[1:num]+v
   # obs:  y[1],..., y[50]
# set parameters
mu0=0
sigma0=1
phi=1
cQ=1
cR=1
```

The authors have contributed R code to do prediction, filtering and smoothing - see the course website and the description of the code at http://www.stat.pitt.edu/stoffer/tsa2/chap6.htm. There are 3 'levels' of functions; each calls a smoothing function and a filtering function:

- Level 0 - $\mathbf{A}$ is constant, $\mathbf{w}_t, \mathbf{v}_t$ independent, no exogenous inputs:

$$\begin{aligned} \mathbf{x}_t &= \mathbf{\Phi}\mathbf{x}_{t-1} + \mathbf{w}_t, \\ \mathbf{y}_t &= \mathbf{A}\mathbf{x}_t + \mathbf{v}_t, \\ \mathbf{x}_0 &\sim N\left(\mu_0, \mathbf{\Sigma}_0\right). \end{aligned}$$

  The filtering call is

$$\begin{aligned} \texttt{Kfilter0} \quad = \quad & \texttt{function(num, y, A, mu0, Sigma0,} \\ & \texttt{Phi, cQ, cR),} \end{aligned}$$

  where $num = n$, $\mathbf{y}$ is an $n \times q$ matrix

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix},$$

$\mathbf{A}$ is the $q \times p$ observation matrix, μ_0 is $p \times 1$, $\mathbf{\Sigma}_0$ and $\mathbf{\Phi}$ are $p \times p$ and $\mathbf{cQ}$, $\mathbf{cR}$ are Cholesky decompositions:

$$\mathbf{Q} = \mathbf{cQ}'\mathbf{cQ}$$

with $\mathbf{cQ}$ upper triangular (in R, `cQ = chol(Q)`).

```
> Q # must be p.s.d.
           [,1]       [,2]     [,3]
[1,]  1.2586719 -0.8455278 1.068583
[2,] -0.8455278  4.1006809 1.768944
[3,]  1.0685831  1.7689436 3.368359
> cQ = chol(Q)
> cQ
         [,1]       [,2]      [,3]
[1,] 1.121905 -0.7536533 0.9524716
[2,] 0.000000  1.8795445 1.3230742
[3,] 0.000000  0.0000000 0.8429896
> t(cQ)%*%cQ
           [,1]       [,2]     [,3]
[1,]  1.2586719 -0.8455278 1.068583
[2,] -0.8455278  4.1006809 1.768944
[3,]  1.0685831  1.7689436 3.368359
```

Output:

`list` ($\texttt{xp} = \texttt{xp}, \texttt{Pp} = \texttt{Pp}, \texttt{xf} = \texttt{xf}, \texttt{Pf} = \texttt{Pf}, \texttt{like} = \texttt{like},$
$\texttt{innov} = \texttt{innov}, \texttt{sig} = \texttt{sig}, \texttt{Kn} = \texttt{K}$),

where ("p" for predictions, "f" for filters)

1. xp is a $p \times 1 \times n$ array with $xp\,[, 1, t] = \mathbf{x}_t^{t-1}$,

2. Pp is a $p \times p \times n$ array with $Pp\,[, , t] = \mathbf{P}_t^{t-1}$,

3. xf is a $p \times 1 \times (n+1)$ array with $xf\,[, 1, t] = \mathbf{x}_t^t$
   (for $t = 0, ..., n$ - see 'Oarray')

4. Pf is a $p \times p \times (n+1)$ array with $Pf\,[, , t] = \mathbf{P}_t^t$,

5. $like$ is $-\log(L(\Theta))$,

6. $innov$ is a $q \times 1 \times n$ array with $innov\,[, 1, t] = \varepsilon_t$,

7. sig is a $q \times q \times n$ array with $sig[,,] = \mathbf{\Sigma}_t$,

8. $Kn = \mathbf{K}_n$.

The call to the smoother is identical:

$$\texttt{Ksmooth0} = \texttt{function}(\texttt{num}, \texttt{y}, \texttt{A}, \texttt{mu0}, \texttt{Sigma0}, \texttt{Phi}, \texttt{cQ}, \texttt{cR})$$

and the output is

$$\texttt{list}(\texttt{xs} = \texttt{xs}, \texttt{Ps} = \texttt{Ps}, \texttt{J} = \texttt{J}, \texttt{xp} = \texttt{kf\$xp}, \texttt{Pp} = \texttt{kf\$Pp}, \texttt{xf} = \texttt{kf\$xf}, \texttt{Pf} = \texttt{kf\$Pf}, \texttt{like} = \texttt{kf\$like}, \texttt{Kn} = \texttt{kf\$K}),$$

where

1. xs is a $p \times 1 \times (n+1)$ array with $xs\,[,1,t] = \mathbf{x}_t^n$,

2. Ps is a $p \times p \times (n+1)$ array with $Ps\,[,,t] = \mathbf{P}_t^n$,

3. J is a $p \times p \times n$ array with $J\,[,,t] = \mathbf{J}_t$,

4. all other output is inherited from the filtering output.

- If all three steps - forecasting, filtering, smoothing - are to be done, then only `Ksmooth` need be called. It in turn calls `Kfilter`, which does forecasting and filtering.

- Level 1: $\mathbf{A} = \mathbf{A}_t$ is time dependent, $\mathbf{w}_t, \mathbf{v}_t$ independent, exogenous inputs:

$$\begin{aligned} \mathbf{x}_t &= \mathbf{\Phi x}_{t-1} + \mathbf{\Upsilon u}_t + \mathbf{w}_t, \\ \mathbf{y}_t &= \mathbf{A}_t \mathbf{x}_t + \mathbf{\Gamma u}_t + \mathbf{v}_t, \\ \mathbf{x}_0 &\sim N\left(\mu_0, \mathbf{\Sigma}_0\right). \end{aligned}$$

- Level 2: $\mathbf{w}_t, \mathbf{v}_t$ correlated:

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{\Phi x}_t + \mathbf{\Upsilon u}_t + \mathbf{w}_t, \\ \mathbf{y}_t &= \mathbf{A}_t \mathbf{x}_t + \mathbf{\Gamma u}_t + \mathbf{v}_t, \\ cov\left[\mathbf{w}_s, \mathbf{v}_t\right] &= \mathbf{S} \cdot I\left(s = t\right). \end{aligned}$$

  The details are as for Level 0, except that $\mathbf{\Upsilon}, \mathbf{\Gamma}$, $\mathbf{S}$ and $\mathbf{u}_t$ are included in the inputs.

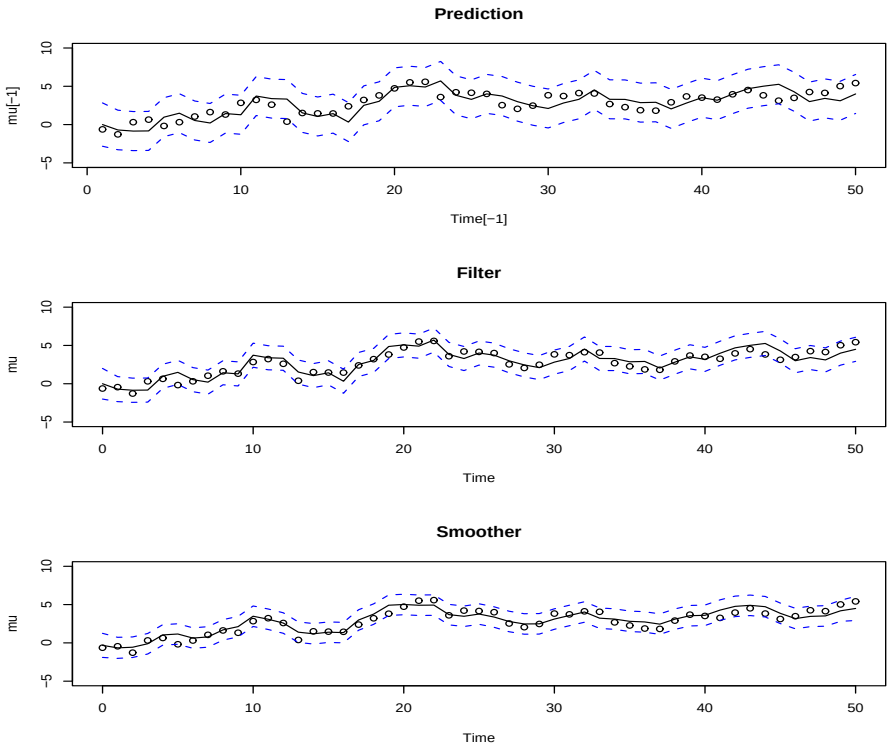

Figure 12.1. Data μ_t plotted as points. Top: Predictions and error bounds $\mu_t^{t-1} \pm 2\sqrt{P_t^{t-1}}$. Middle: Filter and error bounds $\mu_t^t \pm 2\sqrt{P_t^t}$. Bottom: Smoother and error bounds $\mu_t^n \pm 2\sqrt{P_t^n}$. Note that the widths of the error bands decrease: $P_t^{t-1} > P_t^t > P_t^n$.

Now the graphical output in Figure 12.1 is obtained from

```
ks = Ksmooth0(50,y,1,mu0,sigma0,phi,cQ,cR)
#-- pictures (one given here; there are 2 others)
## mu_t^t-1=ks$xp,  P_t^t-1=ks$Pp,
#   t = 1,...,n and "p" for prediction
## mu_t^t=ks$xf,    P_t^t=ks$Pf,
# t = 0,1,...,n and "f" for filter
## mu_t^n=ks$xs,    P_t^n=ks$Ps,
# t = 0,1,...,n and "s" for smoother
  Time=0:num
  par(mfrow=c(3,1))
  plot(Time[-1], mu[-1], main="Prediction",
                 ylim=c(-5,10))
          #[-1] leaves off first component
   lines(ks$xp)
   lines(ks$xp+2*sqrt(ks$Pp),
    lty="dashed", col="blue")
   lines(ks$xp-2*sqrt(ks$Pp),
      lty="dashed", col="blue")
```

13. Estimation

13.1. Maximum likelihood estimation

- Recall

$$\begin{aligned} \mathbf{x}_t &= \mathbf{\Phi x}_{t-1} + \mathbf{\Upsilon u}_t + \mathbf{w}_t, \\ \mathbf{y}_t &= \mathbf{A}_t \mathbf{x}_t + \mathbf{\Gamma u}_t + \mathbf{v}_t, \end{aligned}$$

with given initial parameters μ_0, $\mathbf{\Sigma}_0$. Thus the set of parameters to be estimated is

$$\Theta = \{\mu_0, \mathbf{\Sigma}_0, \mathbf{\Phi}, \mathbf{Q}\,(= \mathrm{cov}\,[\mathbf{w}_t])\,, \mathbf{R}\,(= \mathrm{cov}\,[\mathbf{v}_t])\,, \mathbf{\Upsilon}, \mathbf{\Gamma}\}\,.$$

- Assume all randomness is normally distributed. Here the "innovations form" of the likelihood is considered; this involves viewing the innovations as the r.v.s of interest and obtaining their likelihood. Recall that the innovations are

$$\varepsilon_t = \mathbf{y}_t - E\left[\mathbf{y}_t | \mathbf{y}^{t-1}\right] = \mathbf{y}_t - \left\{\mathbf{A}_t \mathbf{x}_t^{t-1} + \mathbf{\Gamma u}_t\right\};$$

these are independent and Normal with zero means and covariances

$$\mathbf{\Sigma}_t = \mathbf{A}_t \mathbf{P}_t^{t-1} \mathbf{A}_t' + \mathbf{R}.$$

Thus the joint density is

$$\prod_{t=1}^{n} \left\{ (2\pi)^{q/2} \left| \mathbf{\Sigma}_t \right|^{-1/2} \exp - \left(\frac{\varepsilon_t' \mathbf{\Sigma}_t^{-1} \varepsilon_t}{2} \right) \right\},$$

and the log-likelihood $l(\Theta)$ is given by (up to an additive constant)

$$-l(\Theta) = \frac{1}{2} \sum_{t=1}^{n} \log \left| \mathbf{\Sigma}_t(\Theta) \right| + \sum_{t=1}^{n} \varepsilon_t'(\Theta) \mathbf{\Sigma}_t^{-1}(\Theta) \varepsilon_t(\Theta).$$

The MLEs are to minimize $-l(\Theta)$.

- The following algorithm is recommended:

Initialization step Choose $\Theta^{(0)}$.

For $k = 1, 2, \ldots$ to convergence:

Filter step Recall the Kalman filter:

$$\begin{aligned}
\mathbf{x}_t^{t-1} &= \mathbf{\Phi}\mathbf{x}_{t-1}^{t-1} + \mathbf{\Upsilon}\mathbf{u}_t, \\
\mathbf{P}_t^{t-1} &= \mathbf{\Phi}\mathbf{P}_{t-1}^{t-1}\mathbf{\Phi}' + \mathbf{Q} \\
&\ldots \\
\mathbf{x}_t^t &= \mathbf{x}_t^{t-1} + \mathbf{K}_t\varepsilon_t, \text{ where} \\
\mathbf{K}_t &= \mathbf{P}_t^{t-1}\mathbf{A}_t'\mathbf{\Sigma}_t^{-1} = \mathbf{P}_t^{t-1}\mathbf{A}_t'\left(\mathbf{A}_t\mathbf{P}_t^{t-1}\mathbf{A}_t' + \mathbf{R}\right)^{-1}, \\
\mathbf{P}_t^t &= \mathbf{P}_t^{t-1} - \mathbf{P}_t^{t-1}\mathbf{A}_t'\mathbf{\Sigma}_t^{-1}\mathbf{A}_t\mathbf{P}_t^{t-1}.
\end{aligned}$$

Substitute the values $\mathbf{x}_t^{t-1}$ and $\mathbf{P}_t^{t-1}$ obtained in this way into the $\varepsilon_t(\Theta)$ and $\mathbf{\Sigma}_t(\Theta)$, obtaining $\varepsilon_t^{(k)}$, $\mathbf{\Sigma}_t^{(k)}$ and thus a negative likelihood $-l^{(k)}(\Theta)$.

Minimization step Do one iteration of Newton-Raphson:

$$\Theta^{(k+1)} = \Theta^{(k)} - \left[-\ddot{l}\left(\Theta^{(k)}\right)\right]^{-1}\left[-\dot{l}\left(\Theta^{(k)}\right)\right].$$

- Example 6.7. Global Temperature series. Recall x_t = global temperature is being studied. One observes y_{t1} = average air temperature at land based stations and y_{t2} = average air temperature at marine based stations. Modelling x_t as a random walk, and assuming that y_{t1} and y_{t2} are, apart from random error, measuring the same thing, gives

$$\begin{aligned} x_t &= x_{t-1} + w_t, \\ \begin{pmatrix} y_{t1} \\ y_{t2} \end{pmatrix} &= \mathbf{y}_t = \begin{pmatrix} 1 \\ 1 \end{pmatrix} x_t + \mathbf{v}_t. \end{aligned}$$

So $\Phi = 1, \mathbf{A}_t = (1,1)'$ $(p = 1, q = 2)$ and the parameters in $Q = \sigma_w^2$ and $\mathbf{R} = \text{cov}[\mathbf{v}_t]$ are to be estimated from the data. There are no exogenous inputs, and the initial choices of μ_0, Σ_0 are taken to be the final ones. They in turn are the approximate mean and variance of $\{y_{11}, y_{21}\}$. Contributed (by S&S) code on the course website is used here, to estimate only the parameters Q and $\mathbf{R}$. After they are obtained the Kalman smoother is applied, using estimated parameters.

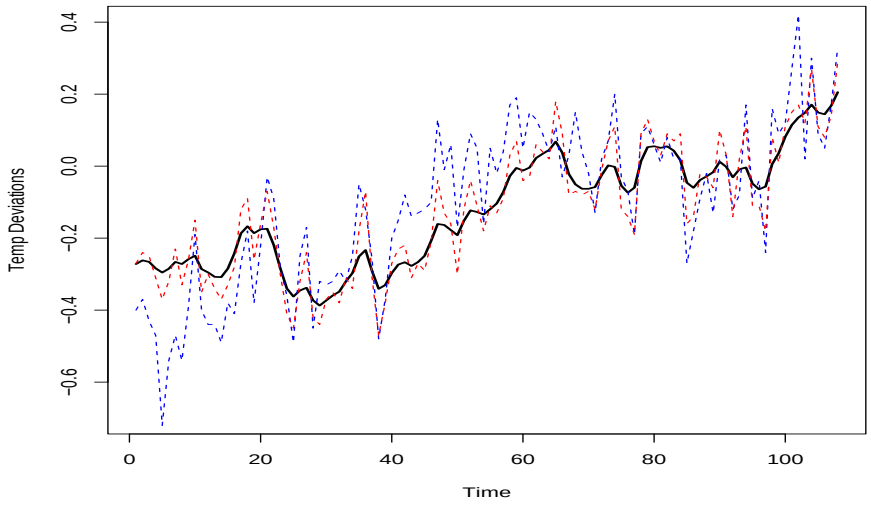

Figure 13.1. Example 6.7. Data (broken lines) and smoothed global temperature deviations.

13.2. EM algorithm

- An alternate approach is based on the EM ("expectation-maximization") algorithm. Suppose, for simplicity, that there are no exogenous inputs:

$$\begin{aligned} \mathbf{x}_t &= \mathbf{\Phi}\mathbf{x}_{t-1} + \mathbf{w}_t, \\ \mathbf{y}_t &= \mathbf{A}_t\mathbf{x}_t + \mathbf{v}_t. \end{aligned}$$

If we could observe all of $\{\mathbf{x}_t, \mathbf{y}_t\}_{t=1}^n$ then the "complete data" likelihood would be

$$
\begin{aligned}
& p_{\mu_0,\Sigma_0}(\mathbf{x}_0) \prod_{t=1}^{n} p(\mathbf{x}_t|\mathbf{x}_{t-1}) \prod_{t=1}^{n} p(\mathbf{y}_t|\mathbf{x}_t) \\
\propto\ & |\Sigma_0|^{-1/2} e^{-\frac{(\mathbf{x}_0-\mu_0)'\Sigma_0^{-1}(\mathbf{x}_0-\mu_0)}{2}} \\
& \cdot \prod_{t=1}^{n} |\mathbf{Q}|^{-1/2} e^{-\frac{(\mathbf{x}_t-\Phi\mathbf{x}_{t-1})'\mathbf{Q}^{-1}(\mathbf{x}_t-\Phi\mathbf{x}_{t-1})}{2}} \\
& \cdot \prod_{t=1}^{n} |\mathbf{R}|^{-1/2} e^{-\frac{(\mathbf{y}_t-\mathbf{A}_t\mathbf{x}_t)'\mathbf{R}^{-1}(\mathbf{y}_t-\mathbf{A}_t\mathbf{x}_t)}{2}},
\end{aligned}
$$

with

$$
\begin{aligned}
& -2l(\Theta) \\
=\ & \ln|\Sigma_0| + (\mathbf{x}_0-\mu_0)'\Sigma_0^{-1}(\mathbf{x}_0-\mu_0) \\
& + \left\{ \begin{array}{c} n\ln|\mathbf{Q}| \\ +\sum_{t=1}^{n}(\mathbf{x}_t-\Phi\mathbf{x}_{t-1})'\mathbf{Q}^{-1}(\mathbf{x}_t-\Phi\mathbf{x}_{t-1}) \end{array} \right\} \\
& + \left\{ \begin{array}{c} n\ln|\mathbf{R}| \\ +\sum_{t=1}^{n}(\mathbf{y}_t-\mathbf{A}_t\mathbf{x}_t)'\mathbf{R}^{-1}(\mathbf{y}_t-\mathbf{A}_t\mathbf{x}_t) \end{array} \right\}.
\end{aligned}
$$

Now for $k = 0, 1, ...$ to convergence, iterate between the following steps.

E step: Since only $\{\mathbf{y}_t\}_{t=1}^n$ is observed, we replace $-2l(\Theta)$ with its expectation $E[-2l(\Theta)|\mathbf{y}^n]$, evaluated at current estimates $\Theta^{(k)}$:

$$E[-2l(\Theta)|\mathbf{y}^n] =$$
$$\left[\begin{array}{c} \ln|\boldsymbol{\Sigma}_0| + \\ tr\boldsymbol{\Sigma}_0^{-1}E\left[(\mathbf{x}_0-\mu_0)(\mathbf{x}_0-\mu_0)'|\mathbf{y}^n\right] \end{array}\right] \quad (13.1)$$
$$+\left[\begin{array}{c} n\ln|\mathbf{Q}| + \\ tr\mathbf{Q}^{-1}E\left[\begin{array}{c}\sum_{t=1}^n(\mathbf{x}_t-\boldsymbol{\Phi}\mathbf{x}_{t-1})\cdot \\ (\mathbf{x}_t-\boldsymbol{\Phi}\mathbf{x}_{t-1})'|\mathbf{y}^n\end{array}\right] \end{array}\right] \quad (13.2)$$
$$+\left[\begin{array}{c} n\ln|\mathbf{R}| + \\ tr\mathbf{R}^{-1}E\left[\begin{array}{c}\sum_{t=1}^n(\mathbf{y}_t-\mathbf{A}_t\mathbf{x}_t)\cdot \\ (\mathbf{y}_t-\mathbf{A}_t\mathbf{x}_t)'|\mathbf{y}^n\end{array}\right]. \end{array}\right] \quad (13.3)$$

(i) Using the Kalman smoother,

$$\begin{aligned} & E\left[(\mathbf{x}_0-\mu_0)(\mathbf{x}_0-\mu_0)'|\mathbf{y}^n\right] \\ = \; & E\left[(\mathbf{x}_0-\mathbf{x}_0^n)(\mathbf{x}_0-\mathbf{x}_0^n)'|\mathbf{y}^n\right] \\ & +(\mathbf{x}_0^n-\mu_0)(\mathbf{x}_0^n-\mu_0)' \\ = \; & P_0^n+(\mathbf{x}_0^n-\mu_0)(\mathbf{x}_0^n-\mu_0)', \end{aligned}$$

hence (13.1) is

$$\ln|\boldsymbol{\Sigma}_0| + tr\boldsymbol{\Sigma}_0^{-1}P_0^n + (\mathbf{x}_0^n-\mu_0)'\boldsymbol{\Sigma}_0^{-1}(\mathbf{x}_0-\mu_0). \quad (13.4)$$

(ii) The expectation in (13.2) is the sum of

$$
\begin{aligned}
& E\left[\left(\mathbf{x}_t - \mathbf{\Phi}\mathbf{x}_{t-1}\right)\left(\mathbf{x}_t - \mathbf{\Phi}\mathbf{x}_{t-1}\right)' |\mathbf{y}^n\right] \\
= \; & E\left[\begin{pmatrix} \left(\mathbf{x}_t - \mathbf{x}_t^n\right) - \mathbf{\Phi}\left(\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^n\right) \\ + \left(\mathbf{x}_t^n - \mathbf{\Phi}\mathbf{x}_{t-1}^n\right) \end{pmatrix}(\cdots)' |\mathbf{y}^n\right] \\
= \; & \mathbf{P}_t^n + \mathbf{\Phi}\mathbf{P}_{t-1}^n\mathbf{\Phi}' + \left(\mathbf{x}_t^n - \mathbf{\Phi}\mathbf{x}_{t-1}^n\right)\left(\mathbf{x}_t^n - \mathbf{\Phi}\mathbf{x}_{t-1}^n\right)' \\
& -\mathbf{P}_{t,t-1}^n\mathbf{\Phi}' - \mathbf{\Phi}\mathbf{P}_{t,t-1}^{n\prime} \\
= \; & \left\{\mathbf{P}_t^n + \mathbf{x}_t^n\mathbf{x}_t^{n\prime}\right\} + \mathbf{\Phi}\left\{\mathbf{P}_{t-1}^n + \mathbf{x}_{t-1}^n\mathbf{x}_{t-1}^{n\prime}\right\}\mathbf{\Phi}' \\
& -\left\{\mathbf{P}_{t,t-1}^n + \mathbf{x}_t^n\mathbf{x}_{t-1}^{n\prime}\right\}\mathbf{\Phi}' - \mathbf{\Phi}\left\{\mathbf{P}_{t,t-1}^{n\prime} + \mathbf{x}_t^n\mathbf{x}_{t-1}^{n\prime}\right\};
\end{aligned}
$$

thus, with

$$
\begin{aligned}
\mathbf{S}_{11} &= \sum_{t=1}^{n}\left\{\mathbf{P}_t^n + \mathbf{x}_t^n\mathbf{x}_t^{n\prime}\right\}, \\
\mathbf{S}_{00} &= \sum_{t=1}^{n}\left\{\mathbf{P}_{t-1}^n + \mathbf{x}_{t-1}^n\mathbf{x}_{t-1}^{n\prime}\right\}, \\
\mathbf{S}_{10} &= \sum_{t=1}^{n}\left\{\mathbf{P}_{t,t-1}^n + \mathbf{x}_t^n\mathbf{x}_{t-1}^{n\prime}\right\},
\end{aligned}
$$

(13.2) is

$$\begin{aligned} & n \ln |\mathbf{Q}| + tr\mathbf{Q}^{-1}\left[\mathbf{S}_{11} + \Phi\mathbf{S}_{00}\Phi' - \mathbf{S}_{10}\Phi' - \Phi\mathbf{S}_{10}'\right] \\ = \; & n \ln |\mathbf{Q}| + \\ & tr\mathbf{Q}^{-1/2}\left[\begin{array}{c} \left(\Phi - \mathbf{S}_{10}\mathbf{S}_{00}^{-1}\right)\mathbf{S}_{00}\left(\Phi - \mathbf{S}_{10}\mathbf{S}_{00}^{-1}\right)' \\ + \mathbf{S}_{11} - \mathbf{S}_{10}\mathbf{S}_{00}^{-1}\mathbf{S}_{10}' \end{array}\right]\mathbf{Q}^{-1/2}. \end{aligned} \tag{13.5}$$

(iii) Similarly, (13.3) is

$$\begin{aligned} & n \ln |\mathbf{R}| + \\ & tr\mathbf{R}^{-1}\sum_{t=1}^{n}\left[\begin{array}{c} \left(\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t^n\right)\left(\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t^n\right)' \\ + \mathbf{A}_t\mathbf{P}_t^n\mathbf{A}_t' \end{array}\right]. \end{aligned} \tag{13.6}$$

The terms $\mathbf{x}_t^n, \mathbf{P}_t^n$ and the S_{ij} are all evaluated at $\Theta^{(k)}$.

M step: The maximization step calls for the minimization of $E\left[-2l\left(\Theta\right)|\mathbf{y}^n\right]$; this gives updated esti-

mates from (13.4), (13.5) and (13.6) respectively:

$$\begin{cases} \mu_0^{(k+1)} = \mathbf{x}_0^n \\ \boldsymbol{\Sigma}_0^{(k+1)} = P_0^n \end{cases}$$

$$\begin{cases} \boldsymbol{\Phi}^{(k+1)} = \mathbf{S}_{10}\mathbf{S}_{00}^{-1} \\ \mathbf{Q}^{(k+1)} = \frac{\mathbf{S}_{11} - \mathbf{S}_{10}\mathbf{S}_{00}^{-1}\mathbf{S}_{10}'}{n} \end{cases}$$

$$\mathbf{R}^{(k+1)} = \frac{\sum_{t=1}^{n} \left[\begin{array}{c} (\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t^n)(\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t^n)' \\ +\mathbf{A}_t\mathbf{P}_t^n\mathbf{A}_t' \end{array} \right]}{n}.$$

Reasons: Clearly, (13.5) is minimized, for fixed $\mathbf{Q}$, at $\Phi = \mathbf{S}_{10}\mathbf{S}_{00}^{-1}$, with minimum value

$$n \ln |\mathbf{Q}| + tr\mathbf{Q}^{-1}\left[\mathbf{S}_{11} - \mathbf{S}_{10}\mathbf{S}_{00}^{-1}\mathbf{S}_{10}'\right].$$

The minimization of this, and the other terms, uses the general result that for $p \times p$ positive definite matrices Σ and $\mathbf{P}$,

$$f(\boldsymbol{\Sigma}) = n \ln |\boldsymbol{\Sigma}| + tr\boldsymbol{\Sigma}^{-1}\mathbf{P}$$

is minimized by $\Sigma = n^{-1}\mathbf{P}$. This is because minimizing over Σ is equivalent to minimizing over

$$\mathbf{M} \stackrel{def}{=} \mathbf{P}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{P}^{1/2}.$$

We have

$$\begin{aligned} f(\boldsymbol{\Sigma}) &= n \ln |\mathbf{P}| + tr\mathbf{M} - n \ln |\mathbf{M}| \\ &= n \ln |\mathbf{P}| + \sum_{i=1}^{p} (\lambda_i - n \ln \lambda_i), \end{aligned}$$

where the $\{\lambda_i\}$ are the (positive) eigenvalues of $\mathbf{M}$. Each summand is minimized by $\lambda_i = n$, so that $\mathbf{M} = n\mathbf{I}_p$, implying $\boldsymbol{\Sigma} = n^{-1}\mathbf{P}$, minimizes $f(\boldsymbol{\Sigma})$.

- Example 6.8. Code to carry out this procedure, using the same simulated data as in Examples 6.3, 6.6 is on the website. The model is AR(1) with observation noise:

$$\begin{aligned} y_t &= x_t + v_t, \\ x_t &= \phi x_{t-1} + w_t \ (t > 0), \\ x_0 &\sim N\left(0, \frac{\sigma_w^2}{1-\phi^2}\right) \text{ (why?); so} \\ \Theta &= \left\{\phi, Q = \sigma_w^2, R = \sigma_v^2, \mu_0, \boldsymbol{\Sigma}_0\right\}, \ A = 1, \end{aligned}$$

and $\{v_t\}, \{w_t\}, x_0$ are independent. Data were simulated ($n = 100, \phi = .8, \sigma_w^2 = \sigma_v^2 = 1$).

Initial estimates - method of moments estimates based on the autocovariance function of y_t - are computed. Then the function EM0 ("zero", not "oh") carries out the EM algorithm. The resulting estimates are fed to `Kfilter0` to obtain the likelihood evaluated at the estimates of the parameters. A further R package (`fdHess` - from the help file: "Evaluates an approximate Hessian and gradient of a scalar function using finite differences") computes the Hessian of the log-likelihood, from which (asymptotic) standard errors are obtained:

```
            estimate       s.e.
phi       0.81052879 0.07846182
sigw      0.85147592 0.16424705
sigv      0.86494130 0.13655175
mu0      -1.93530717         NA
Sigma0    0.02423680         NA
```

- Read the material on 'Asymptotic Distributions' - the end result is that the usual large-sample approximations for the MLE are valid:

$$\sqrt{n}\left(\hat{\Theta}_n - \Theta_0\right) \xrightarrow{L} N\left(\mathbf{0}, I^{-1}\left(\Theta_0\right)\right),$$

where

$$I\left(\Theta_0\right) = \lim \frac{E\left[-\ddot{l}\left(\Theta_0\right)\right]}{n},$$

but some special conditions are required.

- The EM algorithm is most useful when data are missing; in this case they are replaced by their conditional expectations and then one proceeds as if they were observed. Details in the text.

14. Structural models; ARMAX models in state-space form

14.1. Structural models

Figure 14.1. J&J data.

- Here the observed series is assumed to consist of fixed and disturbed trends, and classical autoregressions. Contrary to the Box-Jenkins approach,

in which fixed and seasonal (nonstationary) trends are removed (often by differencing), these are now modelled explicitly. An example is the Johnson and Johnson Quarterly Earnings data - plot (Figure 14.1) reveals an increasing mean trend and a seasonal component with a periodicity of one year (= 4 quarters). The proposed model is

$$y_t = T_t + S_t + v_t$$

where:

$$T_t = \phi T_{t-1} + w_{t1}$$

with $\phi > 1$ to model an exponential increase in the mean,

$$S_t + S_{t-1} + S_{t-2} + S_{t-3} = w_{t2},$$

(with w_{t1}, w_{t2} independent) reflecting that after 4 quarters the seasonal part of the series has returned to the beginning. Thus

$$\begin{aligned} y_t &= \begin{pmatrix} 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} T_t \\ S_t \\ S_{t-1} \\ S_{t-2} \end{pmatrix} + v_t \\ &= \mathbf{A}\mathbf{x}_t + v_t, \end{aligned}$$

with

$$\begin{aligned}
\mathbf{x}_t &= \begin{pmatrix} T_t \\ S_t \\ S_{t-1} \\ S_{t-2} \end{pmatrix} \\
&= \begin{pmatrix} \phi & 0 & 0 & 0 \\ 0 & -1 & -1 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} T_{t-1} \\ S_{t-1} \\ S_{t-2} \\ S_{t-3} \end{pmatrix} + \begin{pmatrix} w_{t1} \\ w_{t2} \\ 0 \\ 0 \end{pmatrix} \\
&= \mathbf{\Phi x}_{t-1} + \mathbf{w}_t.
\end{aligned}$$

Then

$$\begin{aligned}
\text{var}\,[v_t] &= r, \\
\text{cov}\,[\mathbf{w}_t] &= \mathbf{Q} = diag(q_1, q_2, 0, 0).
\end{aligned}$$

Initial values were $\phi = 1.03$ (growth of about 3% per year), $\mu_0 = (.5, .3, .2, .1)'$ (?), $\mathbf{\Sigma}_0 = diag(.01, .01, .01, .01)$, $q_1 = .01, q_2 = .1, r = .04$. The EM algorithm gives

```
          estimate      stderr
Phi11 1.035084e+00 0.002523754
sigw1 1.390357e-01 0.021329986
sigw2 2.199824e-01 0.023568936
sigv  4.993932e-07 0.239357104
```

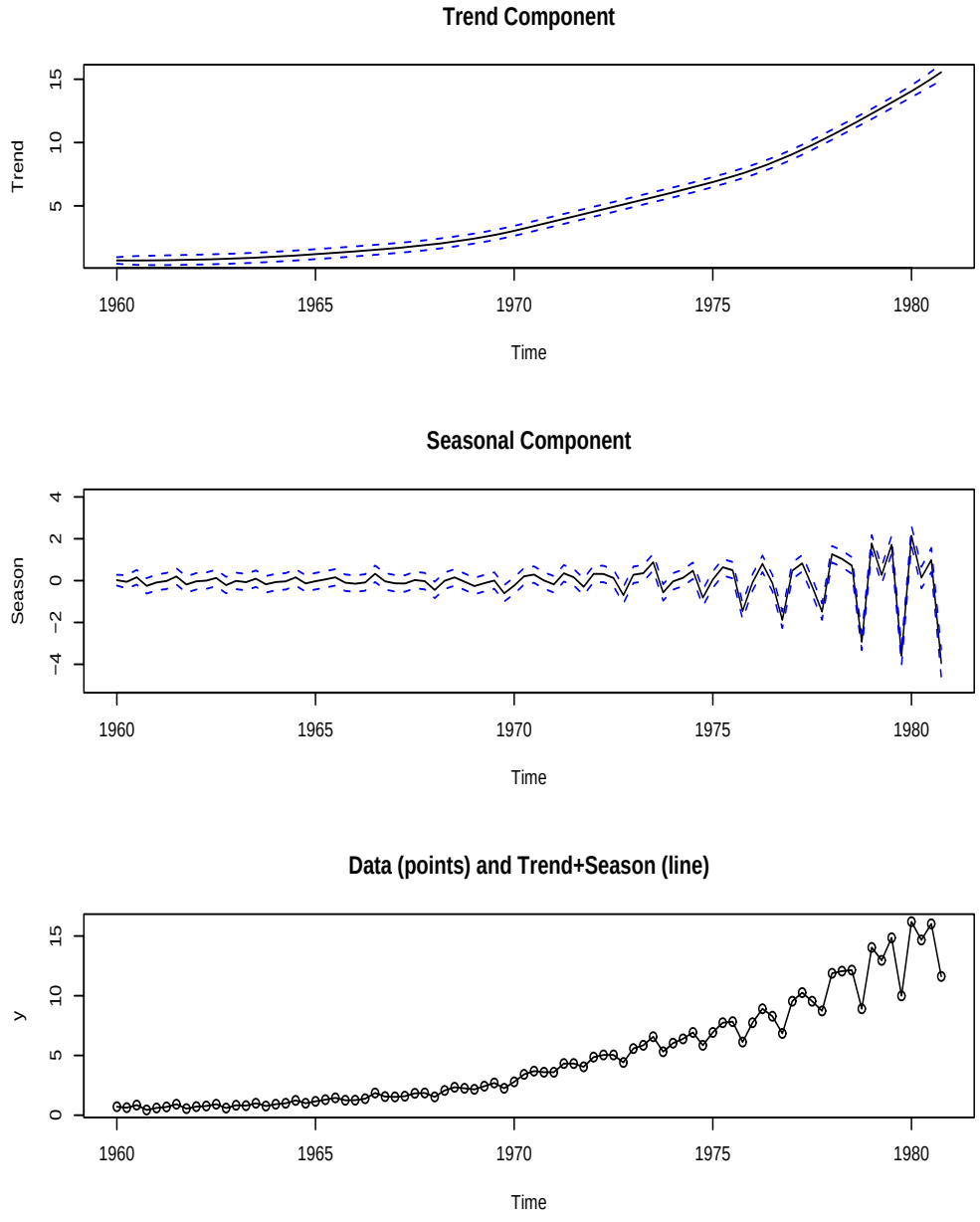

Figure 14.2. J&J Quarterly Earnings output.

14.2. ARMAX models

- Recall the ARMA model from Lecture 9. Here we write it in terms of $(\mathbf{y}, \mathbf{v})$ rather than $(\mathbf{x}, \mathbf{w})$:

$$\mathbf{y}_t = \sum_{j=1}^{p} \Phi_j \mathbf{y}_{t-j} + \mathbf{v}_t + \sum_{l=1}^{q} \Theta_l \mathbf{v}_{t-l}, \ t = 1, ..., n.$$

First put $s = \max(p, q+1)$ and write the model as

$$\mathbf{y}_t = \sum_{j=1}^{s} \Phi_j \mathbf{y}_{t-j} + \mathbf{v}_t + \sum_{l=1}^{s-1} \Theta_l \mathbf{v}_{t-l}, \tag{14.1}$$

with $\Phi_{p+1} = \cdots = \Phi_s = \mathbf{0}$ if $s = q+1 > p$, and $\Theta_{q+1} = \cdots = \Theta_{s-1} = \mathbf{0}$ if $s = p > q+1$. Here Φ_j, Θ_l are $k \times k$, $\mathbf{y}_t$ and $\mathbf{v}_t$ are $k \times 1$.

- Following S&S, we try to put this into state-space form

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \mathbf{v}_t, \tag{14.2}$$
$$\mathbf{x}_{t+1} = \mathbf{\Phi}\mathbf{x}_t + \mathbf{w}_t, \tag{14.3}$$
$$\mathbf{w}_t = \mathbf{\Psi}\mathbf{v}_t, \tag{14.4}$$

with

$$\mathbf{A}_{k\times ks} = \begin{pmatrix} \mathbf{I}_k & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix}, \ \mathbf{x}_t = \begin{pmatrix} \mathbf{x}_{t,1} \\ \vdots \\ \mathbf{x}_{t,s} \end{pmatrix} : ks \times 1,$$

$$\mathbf{\Phi}_{ks\times ks} = \begin{pmatrix} \mathbf{\Phi}_1 & \mathbf{I}_k & & & \\ \mathbf{\Phi}_2 & & \mathbf{I}_k & \mathbf{0} & \\ \vdots & & \mathbf{0} & \ddots & \\ \mathbf{\Phi}_{s-1} & & & & \mathbf{I}_k \\ \mathbf{\Phi}_s & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{pmatrix},$$

$$\mathbf{\Psi}_{ks\times k} = \begin{pmatrix} \Theta_1 + \mathbf{\Phi}_1 \\ \Theta_2 + \mathbf{\Phi}_2 \\ \vdots \\ \Theta_{s-1} + \mathbf{\Phi}_{s-1} \\ \mathbf{\Phi}_s \end{pmatrix}.$$

- If these equations are to hold, what is $\mathbf{x}_t$? First note that in these terms the state equation (14.3) and noise equation (14.4) result in

$$\begin{aligned}\mathbf{x}_{t+1,j} &= \mathbf{\Phi}_j \mathbf{x}_{t,1} + \mathbf{x}_{t,j+1} + \left(\mathbf{\Theta}_j + \mathbf{\Phi}_j\right)\mathbf{v}_t,\ 1 \le j < s,\\ \mathbf{x}_{t+1,s} &= \mathbf{\Phi}_s \mathbf{x}_{t,1} + \mathbf{\Phi}_s \mathbf{v}_t.\end{aligned}$$

The observation equation (14.2) forces

$$\mathbf{x}_{t,1} = \mathbf{y}_t - \mathbf{v}_t.$$

Then the first equation above ($j = 1$) gives

$$\begin{aligned}&\mathbf{x}_{t,2}\\ =\ &\mathbf{x}_{t+1,1} - \mathbf{\Phi}_1 \mathbf{x}_{t,1} - \left(\mathbf{\Theta}_1 + \mathbf{\Phi}_1\right)\mathbf{v}_t\\ =\ &\left(\mathbf{y}_{t+1} - \mathbf{v}_{t+1}\right) - \mathbf{\Phi}_1\left(\mathbf{y}_t - \mathbf{v}_t\right) - \left(\mathbf{\Theta}_1 + \mathbf{\Phi}_1\right)\mathbf{v}_t\\ =\ &\mathbf{y}_{t+1} - \mathbf{\Phi}_1 \mathbf{y}_t - \left(\mathbf{v}_{t+1} + \mathbf{\Theta}_1 \mathbf{v}_t\right).\end{aligned}$$

Continuing in this manner we obtain

$$\begin{aligned}\mathbf{x}_{t,k} &= \mathbf{y}_{t+k-1} - \sum_{j=1}^{k-1} \mathbf{\Phi}_j \mathbf{y}_{t+k-1-j}\\ &\quad - \left(\mathbf{v}_{t+k-1} + \sum_{l=1}^{k-1} \mathbf{\Theta}_l \mathbf{v}_{t+k-1-l}\right),\end{aligned} \tag{14.5}$$

for $k = 1, ..., s$, with $\sum_{j=1}^{0}$ defined to be zero.

- This shows that if the state-space representation of (14.1), given by the observation, state and noise equations (14.2), (14.3) and (14.4), are to hold for this choice of $\mathbf{A}$, Φ and Ψ, then it is necessary that $\mathbf{x}_t$ be defined by (14.5). Conversely, it is easily verified that if $\mathbf{y}_t$ is given by (14.1) and $\mathbf{x}_t$ is given by (14.5) then (14.2) – (14.4) hold for the specified $\mathbf{A}$, Φ and Ψ.

- If $\mathbf{R} = \text{cov}[\mathbf{v}_t]$ we have

$$\begin{aligned} \mathbf{Q} &= \text{cov}\left[\mathbf{w}_t\right] = \Psi\mathbf{R}\Psi', \\ \mathbf{S} &= \text{cov}\left[\mathbf{w}_t, \mathbf{v}_t\right] = E\left[\mathbf{w}_t\mathbf{v}_t'\right] = \Psi\mathbf{R}. \end{aligned}$$

  Thus $\mathbf{w}_t$ and $\mathbf{v}_t$ are correlated, and so the 'level 2' R routines contributed by S&S are used to do the filtering and smoothing. The smoothing equations of Lecture 11 continue to hold but the prediction/filtering equations must be modified - see the text ("Property P6.5").

- Example: Univariate ARMA(1,1) model. This is

$$y_t = \phi y_{t-1} + v_t + \theta v_{t-1}. \tag{14.6}$$

A formal application of the development above gives $s = \max(p, q+1) = 2$, $\phi_2 = 0$,

$$\begin{aligned} y_t &= (1, 0) \begin{pmatrix} x_{t,1} \\ x_{t,2} \end{pmatrix} + v_t, \\ \begin{pmatrix} x_{t+1,1} \\ x_{t+1,2} \end{pmatrix} &= \begin{pmatrix} \phi & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_{t,1} \\ x_{t,2} \end{pmatrix} + \begin{pmatrix} \phi + \theta \\ 0 \end{pmatrix} v_t. \end{aligned}$$

More succinctly,

$$\begin{aligned} y_t &= x_t + v_t, & (14.7) \\ x_{t+1} &= \phi x_t + (\phi + \theta) v_t. & (14.8) \end{aligned}$$

Then $w_s = (\phi + \theta) v_s$ is uncorrelated with v_t unless $s = t$, in which case the covariance is $(\phi + \theta) \sigma_v^2$. To recover (14.6) from (14.7), (14.8):

$$\begin{aligned} y_t &= x_t + v_t \\ &= \{\phi x_{t-1} + (\phi + \theta) v_{t-1}\} + v_t \text{ (from (14.8))} \\ &= \phi (x_{t-1} + v_{t-1}) + v_t + \theta v_{t-1} \\ &= \phi y_{t-1} + v_t + \theta v_{t-1} \text{ (from (14.7))}; \end{aligned}$$

this is (14.6).

- Suppose now that there are exogenous inputs:

$$\mathbf{y}_t = \Gamma \mathbf{u}_t + \sum_{j=1}^{s} \Phi_j \mathbf{y}_{t-j} + \mathbf{v}_t + \sum_{l=1}^{s-1} \Theta_l \mathbf{v}_{t-l} \tag{14.9}$$

for $\Gamma : k \times r$. S&S claim(ed) that there is still a representation of the form

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \Gamma \mathbf{u}_t + \mathbf{v}_t, \tag{14.10}$$

$$\mathbf{x}_{t+1} = \Phi \mathbf{x}_t + \mathbf{w}_t, \tag{14.11}$$

with this choice of $\mathbf{A}$,Φ and Ψ. In fact it can be shown that this is possible, for some choice of $\{\mathbf{x}_t\}$, iff

$$\sum_{j=1}^{p} \Phi_j \Gamma \mathbf{u}_{t+p-j} = \mathbf{0},\ t = 0, 1, \ldots .$$

Equivalently, with $\Phi(B)$ representing the VAR(p) characteristic polynomial $\mathbf{I}_k - \sum_{j=1}^{p} \Phi_j B^j$,

$$\left(\mathbf{I}_k - \Phi(B)\right) \Gamma \mathbf{u}_t = \mathbf{0},\ t = 0, 1, \ldots .$$

- Here is an interpretation of the case of exogenous inputs which preserves the structure of (14.10) and (14.11). Define α_t by

$$\Phi(B)\,\alpha_t = \Gamma \mathbf{u}_t,$$

and set

$$\dot{\mathbf{y}}_t = \mathbf{y}_t - \alpha_t.$$

Consider the model, without inputs,

$$\dot{\mathbf{y}}_t = \sum_{j=1}^{s} \Phi_j \dot{\mathbf{y}}_{t-j} + \mathbf{v}_t + \sum_{l=1}^{s-1} \Theta_l \mathbf{v}_{t-l}. \tag{14.12}$$

In operator notation this is

$$\Phi(B)\,\dot{\mathbf{y}}_t = \Theta(B)\,\mathbf{v}_t.$$

Equivalently,

$$\Phi(B)\,\mathbf{y}_t = \Gamma \mathbf{u}_t + \Theta(B)\,\mathbf{v}_t,$$

which is (14.9). The development above, applied to (14.12), results in

$$\dot{\mathbf{y}}_t = \mathbf{A}\mathbf{x}_t + \mathbf{v}_t,$$

and (14.3), (14.4), i.e.

$$\begin{aligned} \mathbf{y}_t &= \mathbf{A}\mathbf{x}_t + \alpha_t + \mathbf{v}_t, \\ \mathbf{x}_{t+1} &= \mathbf{\Phi}\mathbf{x}_t + \mathbf{w}_t. \end{aligned}$$

- The authors have now posted the following correction. In state-space form one can (assigned) write (14.9) as

$$\begin{aligned} \mathbf{y}_t &= \mathbf{A}\mathbf{z}_t + \mathbf{v}_t, \\ \mathbf{z}_{t+1} &= \mathbf{\Phi}\mathbf{z}_t + \mathbf{\Upsilon}\mathbf{u}_{t+1} + \mathbf{w}_t, \end{aligned}$$

for some choice of $\{\mathbf{z}_t\}$, where

$$\mathbf{\Upsilon} = \begin{pmatrix} \mathbf{\Gamma} \\ 0 \\ \vdots \\ 0 \end{pmatrix} : ks \times r.$$

- More generally, consider the state-space model

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{\Phi}\mathbf{x}_t + \mathbf{\Upsilon}\mathbf{u}_t + \mathbf{w}_t, \ t = 0, 1, ..., n, \\ \mathbf{y}_t &= \mathbf{A}_t\mathbf{x}_t + \mathbf{\Gamma}\mathbf{u}_t + \mathbf{v}_t, \ t = 1, ..., n, \end{aligned}$$

with $\mathbf{x}_0 \sim N(\mu_0, \Sigma_0)$. Here Φ is $p \times p$, Υ is $p \times r$, $\mathbf{A}_t$ is $q \times p$ and Γ is $q \times r$. We still assume

$$\mathbf{w}_1, ..., \mathbf{w}_t, ... \overset{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{Q}),$$
$$\mathbf{v}_1, ..., \mathbf{v}_t, ... \overset{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{R}).$$

However we now allow $\mathbf{w}_s, \mathbf{v}_t$ to be correlated if $s = t$; in this case cov$[\mathbf{w}_t, \mathbf{v}_t] = E[\mathbf{w}_t \mathbf{v}_t']$ is given by a non-zero matrix $\mathbf{S}$.

14.3. Introduction to bootstrapping

- First the 'bootstrap' method is described. For more details see the Bootstrappng lecture in my STAT 665 notes, at www.stat.ualberta.ca/~wiens/stat665.

- Suppose that we estimate a functional $\theta = h(F)$ by $\hat{\theta}_n = h(\hat{F}_n)$ and then assess the performance through some measure $\lambda_n(F)$ which we estimate

by $\lambda_n(\hat{F}_n)$. (F is the population d.f., $\hat{F}_n$ the empirical d.f.) Examples are

(i) $\lambda_n(F) = P_F\left(\sqrt{n}\left(\hat{\theta}_n - h(F)\right) \leq a\right)$,

(ii) bias $\lambda_n(F) = E_F\left[\hat{\theta}_n\right] - h(F)$,

(iii) variance $\lambda_n(F) = E_F\left[\left(\hat{\theta}_n - E_F\left[\hat{\theta}_n\right]\right)^2\right]$.

The "plug-in estimator" is obtained by replacing *every* occurrence of F by $\hat{F}_n$. In (i), F is replaced by $\hat{F}_n$ in P_F and in $h(F)$. But also $\hat{\theta}_n$ depends on F since the sample values are i.i.d. $\sim F$. We must then now sample from $\hat{F}_n$. Thus $\hat{\theta}_n$ is replaced by

$$\theta_n^* = \hat{\theta}\left(X_1^*, ..., X_n^*\right),$$

where the X_i^* are a random sample drawn with replacement from the data values $x_1, ..., x_n$, i.e. independently drawn from the distribution

$$P\left(X^* = x_j\right) = n^{-1},\ j = 1, ..., n.$$

In (i) then we write $\lambda_n(\hat{F}_n)$ as

$$\begin{aligned} & \lambda_n(\hat{F}_n) = P_{\hat{F}_n}\left(\sqrt{n}\left(\theta_n^* - \hat{\theta}_n\right) \leq a\right) \\ = \; & P_{\hat{F}_n}\left((X_1^*, ..., X_n^*) \in S\right), \text{ where} \\ & S = \left\{(X_1^*, ..., X_n^*) \mid \sqrt{n}\left(\theta_n^* - \hat{\theta}_n\right) \leq a\right\}. \end{aligned}$$

This probability can sometimes be calculated exactly. Generally it must be approximated. For this we draw a large number (B) of "bootstrap" samples $\left(X_{b,1}^*, ..., X_{b,n}^*\right),\ b = 1, ..., B$ from $\hat{F}_n$ and approximate $\lambda_n(\hat{F}_n)$ by the relative frequency of those in S. This requires calculating θ_n^* each time.

- In each of the examples above, we can write

$$\lambda_n(F) = E_F\left[g_n\left(\hat{\theta}_n; F\right)\right]$$

for some function $g_n(\cdot; F)$. This is to be estimated by

$$\lambda_n(\hat{F}_n) = E_{\hat{F}_n}\left[g_n\left(\theta_n^*; \hat{F}_n\right)\right],$$

which is in turn approximated by

$$\lambda^*_{B,n} = \frac{1}{B}\sum_{b=1}^{B} g_n\left(\theta^*_{b,n}\hat{F}_n\right).$$

Typically the WLLN applies:

$$\lambda^*_{B,n} \xrightarrow{pr} \lambda_n(\hat{F}_n) \text{ as } B \to \infty.$$

- We are still estimating $\lambda_n(F)$ by $\lambda_n(\hat{F}_n)$, but the latter is being approximated by $\lambda^*_{B,n}$. We use *approximate* rather than *estimate* because $\lambda_n(\hat{F}_n)$ is not an unknown parameter. Rather, it is a number which can in principle but not in practice be computed.

15. Bootstrapping state-space models; nonlinearity and non-normality

15.1. Bootstrapping

- Consider the state-space model:

$$\begin{aligned}
\mathbf{x}_{t+1} &= \mathbf{\Phi x}_t + \mathbf{\Upsilon u}_t + \mathbf{w}_t,\ t = 0, 1, ..., n, \\
\mathbf{y}_t &= \mathbf{A}_t\mathbf{x}_t + \mathbf{\Gamma u}_t + \mathbf{v}_t,\ t = 1, ..., n, \\
&\mathbf{w}_1, ..., \mathbf{w}_t, ... \overset{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{Q}), \\
&\mathbf{v}_1, ..., \mathbf{v}_t, ... \overset{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{R}), \\
\text{cov}[\mathbf{w}_t, \mathbf{v}_t] &= E\left[\mathbf{w}_t\mathbf{v}_t'\right] = \mathbf{S}.
\end{aligned}$$

  We seek approximations to the distributions of the MLEs which are more accurate, at least in short series, than those obtained from the asymptotic results.

- Notation: The innovations and their covariance matrices are

$$\begin{aligned}
\varepsilon_t &= \mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t^{t-1} - \mathbf{\Gamma u}_t, \qquad (15.1) \\
\mathbf{\Sigma}_t &= \mathbf{A}_t\mathbf{P}_t^{t-1}\mathbf{A}_t' + \mathbf{R}.
\end{aligned}$$

The one-step ahead forecasts, with their conditional covariance matrices, are

$$\begin{aligned} \mathbf{x}_{t+1}^{t} &= \mathbf{\Phi}\mathbf{x}_{t}^{t-1} + \mathbf{\Upsilon}\mathbf{u}_t + \mathbf{K}_t\varepsilon_t, \quad (15.2) \\ \text{where } \mathbf{K}_t &= \left[\mathbf{\Phi}\mathbf{P}_{t}^{t-1}\mathbf{A}_t' + \mathbf{S}\right]\mathbf{\Sigma}_t^{-1}, \\ \text{and } \mathbf{P}_{t+1}^{t} &= \mathbf{\Phi}\mathbf{P}_{t}^{t-1}\mathbf{A}_t' + \mathbf{Q} - \mathbf{K}_t\mathbf{\Sigma}_t\mathbf{K}_t'. \end{aligned}$$

(This is "Property P6.5", valid when $\mathbf{w}_t, \mathbf{v}_t$ are correlated, in a more compact form.) By (15.1),

$$\mathbf{y}_t = \mathbf{A}_t\mathbf{x}_t^{t-1} + \mathbf{\Gamma}\mathbf{u}_t + \varepsilon_t; \quad (15.3)$$

this and (15.2) are the "innovations form" of the model and are the basis of the resampling algorithm to be discussed. Define standardized innovations

$$\mathbf{e}_t = \mathbf{\Sigma}_t^{-1/2}\varepsilon_t$$

and, as usual, denote by Θ the set of all parameters, with 'true' value Θ_0. So Φ, Υ, $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{\Sigma}_t$ and Γ are all viewed as functions of Θ, evaluated at Θ_0. Then (up to an additive constant) the log-likelihood $l(\Theta)$ is given by

$$-2l(\Theta) = \sum_{t=1}^{n}\left[\log|\mathbf{\Sigma}_t(\Theta)| + \mathbf{e}_t'(\Theta)\,\mathbf{e}_t(\Theta)\right],$$

and is maximized by the MLE $\hat{\Theta}$.

- Bootstrapping algorithm:

1. Construct $\mathbf{e}_t\left(\hat{\Theta}\right)$, $t = 1, ..., n$.

2. "Bootstrap the residuals" - sample, with replacement, from $\left\{\mathbf{e}_t\left(\hat{\Theta}\right)\right\}_{t=1}^{n}$; thus obtaining a "bootstrap sample" $\left\{\mathbf{e}_t^*\left(\hat{\Theta}\right)\right\}_{t=1}^{n}$.

3. Obtain a bootstrap sample $\{\mathbf{y}_t^*\}_{t=1}^{n}$ as follows. By (15.2) and (15.3),

$$\begin{aligned}
\xi_t &\stackrel{def}{=} \begin{pmatrix} \mathbf{x}_{t+1}^{t} \\ \mathbf{y}_t \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{\Phi x}_t^{t-1} + \mathbf{\Upsilon u}_t + \mathbf{K}_t\mathbf{\Sigma}_t^{1/2}\mathbf{e}_t \\ \mathbf{A}_t\mathbf{x}_t^{t-1} + \mathbf{\Gamma u}_t + \mathbf{\Sigma}_t^{1/2}\mathbf{e}_t \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{\Phi} & \mathbf{0} \\ \mathbf{A}_t & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_t^{t-1} \\ \mathbf{y}_{t-1} \end{pmatrix} + \begin{pmatrix} \mathbf{\Upsilon} \\ \mathbf{\Gamma} \end{pmatrix} \mathbf{u}_t + \begin{pmatrix} \mathbf{K}_t\mathbf{\Sigma}_t^{1/2} \\ \mathbf{\Sigma}_t^{1/2} \end{pmatrix} \mathbf{e}_t;
\end{aligned}$$

i.e.

$$\xi_t = \mathbf{F}_t \xi_{t-1} + \mathbf{G}\mathbf{u}_t + \mathbf{H}_t \mathbf{e}_t.$$

Use this, with $\mathbf{e}_t$ replaced by $\mathbf{e}_t^* \left(\hat{\Theta}\right)$ and all parameters evaluated at $\hat{\Theta}$, to generate $\{\xi_t\}_{t=1}^n$ and hence a bootstrap sample $\{\mathbf{y}_t^*\}_{t=1}^n$. (The procedure starts with $\mathbf{x}_1^0 = \Phi\mu_0 + \Upsilon\mathbf{u}_0$ and $\mathbf{y}_0 = \mathbf{0}$.)

4. From the bootstrap sample construct a likelihood and obtain an MLE $\hat{\Theta}^*$.

5. Repeat 2-4 a large number (B) of times. Then the finite sample distribution of $\hat{\Theta} - \Theta_0$ is approximated by the empirical distribution of $\left\{\hat{\Theta}_b^* - \hat{\Theta}\right\}_{b=1}^B$. The bootstrap estimate of Θ_i is $\bar{\Theta}_i^*$, the average of the bootstrapped values. The bootstrap estimate of its standard deviation is the sample standard deviation of the B values Θ_{bi}^*.

- Example - stochastic regression. Quarterly interest rates (y_t) and inflation rates (z_t) for 100 quarters are plotted in Figure 15.1.

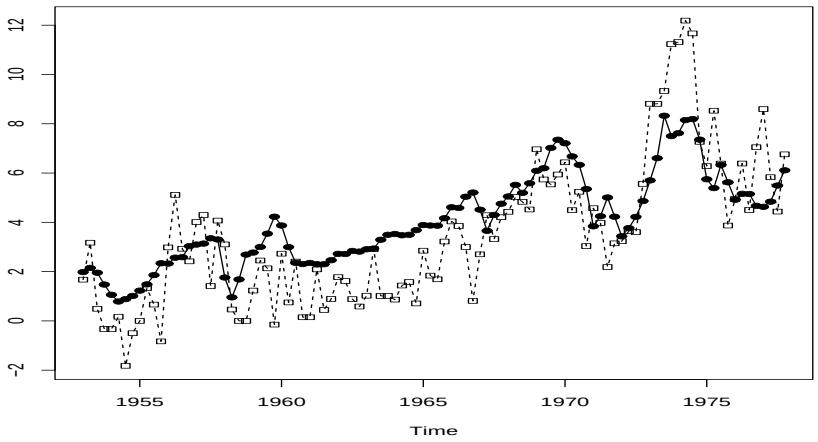

Figure 15.1. Interest rates y_t (solid line) and inflation rates z_t (dashed line).

The model is

$$y_t = \alpha + \beta_t z_t + v_t$$

for a stochastic regression coefficient β_t and constant α. The regression coefficient is modelled as AR(1):

$$\beta_t - b = \phi\left(\beta_{t-1} - b\right) + w_t$$

for a constant b. The white noise processes $\{w_t\}$ and $\{v_t\}$ are assumed uncorrelated. This is then

of the form

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{\Phi x}_t + \mathbf{\Upsilon u}_t + \mathbf{w}_t \\ \text{with } \mathbf{x}_t &= \beta_t, \mathbf{\Phi} = \phi, \mathbf{\Upsilon} = b\left(1-\phi\right), \mathbf{u}_t = 1; \\ \mathbf{y}_t &= \mathbf{A}_t\mathbf{x}_t + \mathbf{\Gamma u}_t + \mathbf{v}_t \\ \text{with } \mathbf{y}_t &= y_t, \mathbf{A}_t = z_t, \mathbf{\Gamma} = \alpha, \end{aligned}$$

and $Q = \sigma_w^2$, $R = \sigma_v^2$, $S = 0$. Then the parameters are

$$\mathbf{\Theta} = \{\phi, \alpha, b, \sigma_w, \sigma_v\}.$$

The R code is on the website. With $B = 200$ and the tolerance parameter set at .01, the results are:

Asymptotic and bootstrapped ($B = 200$) means and standard errors

Parameter	MLE	asym. s.e.	Boot. mean	Boot. s.e.
ϕ	0.877	.072	0.783	0.206
α	2.179	.154	2.034	0.205
b	0.490	.085	0.889	0.536
σ_w	0.079	.022	0.103	0.110
σ_v	0.823	.070	0.950	0.134

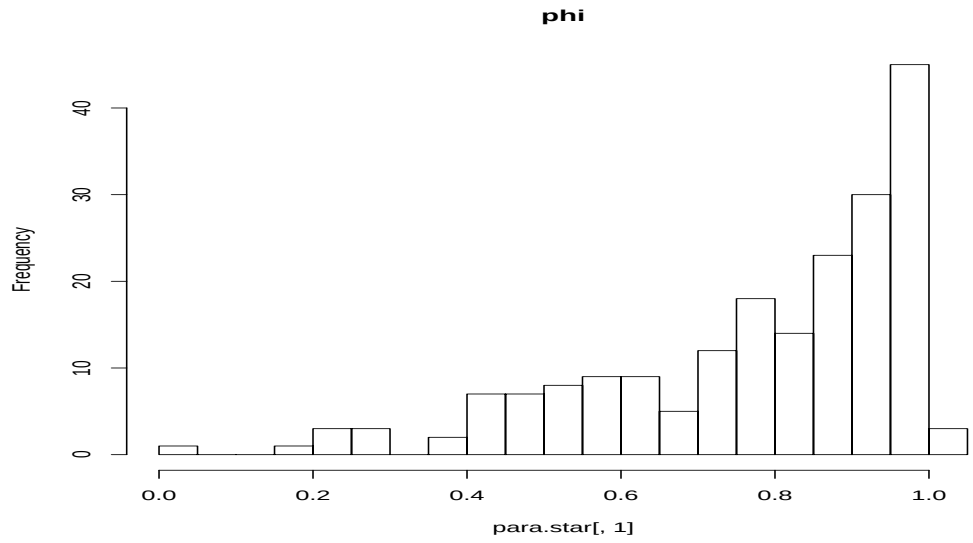

Figure 15.2. $B = 200$ bootstrapped values $\left\{\phi_b^*\right\}_{b=1}^{B}$.

Figure 15.2 gives the $B = 200$ values of ϕ_b^*. The lower and upper 5% values of this empirical distribution are $[0.401, 0.996]$, so this is a bootstrap 90% confidence interval on ϕ. Similarly, a 90% confidence interval on σ_w is $[-.166, .251]$, so that $\sigma_w = 0$, implying that β_t is fixed and not stochastic, is plausible.

15.2. Nonlinearity and non-normality

- Replace

$$\begin{aligned} \mathbf{x}_t &= \mathbf{\Phi}\mathbf{x}_{t-1} + \mathbf{w}_t, \\ \mathbf{y}_t &= \mathbf{A}_t\mathbf{x}_t + \mathbf{v}_t, \end{aligned}$$

  with

$$\begin{aligned} \mathbf{x}_t &= F_t\left(\mathbf{x}_{t-1}, \mathbf{w}_t\right), \\ \mathbf{y}_t &= H_t\left(\mathbf{x}_t, \mathbf{v}_t\right). \end{aligned}$$

  Here F_t and H_t need not be linear, and $\{\mathbf{w}_t, \mathbf{v}_t\}$ need not be Gaussian. But the procedure is still Markovian. Here, several possible approaches are outlined.

- One possibility is Bayesian updating, considered earlier:
  - Start with a prior density $p_0(\mathbf{x}_0)$.

  - Update $\mathbf{x}_t$ after observing $\mathbf{y}^t$:

$$p(\mathbf{x}_t|\mathbf{y}^t) \propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}^{t-1}).$$

  - Predict next state from

$$p(\mathbf{x}_t|\mathbf{y}^{t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})d\mathbf{x}_{t-1}.$$

- MLEs are computed from the likelihood

$$L_{X,Y}(\Theta) = p_\Theta(\mathbf{x}_0)\prod_{t=1}^{n} p_\Theta(\mathbf{x}_t|\mathbf{x}_{t-1})p_\Theta(\mathbf{y}_t|\mathbf{x}_t).$$

- This updating scheme requires many numerical integrations. Recently, Markov Chain Monte Carlo (MCMC) methods have been derived to address this problem. In general, the aim is to simulate a sample (i.i.d.) from a particular density $p_\Theta(\mathbf{z})$. For this, one simulates a sequence $\{\mathbf{z}_t\}$, where, after $\mathbf{z}_{t-1}$ is obtained, $\mathbf{z}_t$ is simulated from a "transition density" $\pi\left(\mathbf{z}_t|\mathbf{z}_{t-1}\right)$. Although the r.v.s so obtained are not independent, the dependence can be tailored to decay quickly (approximate m-independence), so that for large m the

$\{\mathbf{z}_{t+lm}|l = 1, 2, \cdots\}$ are close enough to being independent. Or, one might generate a number of sequences $\left\{\mathbf{z}_t^{(g)}\right\}$ and take, for a large m, the sequence $\left\{\mathbf{z}_m^{(g)}|g = 1, 2, \cdots\right\}$ as having the density $p_\Theta(\mathbf{z})$.

- A popular method of implementing MCMC is the "Gibbs Sampler". Suppose that we seek information about the joint density $p_\Theta(\mathbf{z}_1, ..., \mathbf{z}_k)$ of several r.vecs. This joint density is of high dimension, and intractable, but we have some method of sampling from the conditionals $p_\Theta(\mathbf{z}_j|\mathbf{z}_i, i \neq j)$. These generally determine the joint density $p_\Theta(\mathbf{z}_1, ..., \mathbf{z}_k)$ and hence the marginals $p_\Theta(\mathbf{z}_j)$.

- The Gibbs sampler proceeds as follows:
  - Initialize: $\left\{\mathbf{z}_{1[0]}, ..., \mathbf{z}_{k[0]}\right\}$ (arbitrary).

– Draw

$$\begin{aligned}
&\mathbf{z}_{1[1]} \text{ from } p_\Theta(\mathbf{z}_1|\mathbf{z}_{2[0]}, ..., \mathbf{z}_{k[0]}),\\
&\mathbf{z}_{2[1]} \text{ from } p_\Theta(\mathbf{z}_2|\mathbf{z}_{1[1]}, \mathbf{z}_{3[0]}, ..., \mathbf{z}_{k[0]}),\\
&\ldots\\
&\mathbf{z}_{k[1]} \text{ from } p_\Theta(\mathbf{z}_k|\mathbf{z}_{1[1]}, \mathbf{z}_{2[1]}, ..., \mathbf{z}_{k-1[1]}).
\end{aligned}$$

– Repeat; at step l one has a collection $\left\{\mathbf{z}_{1[l]}, ..., \mathbf{z}_{k[l]}\right\}$. Under appropriate conditions, as $l \longrightarrow \infty$ these converge in distribution to $p_\Theta(\mathbf{z}_1, ..., \mathbf{z}_k)$.

– For a suitably large l, denote this collection by $\left\{\mathbf{z}_1^{(1)}, ..., \mathbf{z}_k^{(1)}\right\}$. Repeat the entire process G times, obtaining $\left\{\mathbf{z}_1^{(g)}, ..., \mathbf{z}_k^{(g)}|g = 1, ..., G\right\}$. These are used to estimate functionals of the joint density. For instance the marginal $p_\Theta(\mathbf{z}_j)$ can be estimated by

$$\hat{p}_\Theta(\mathbf{z}_j) = \frac{1}{G}\sum_{g=1}^{G} p_\Theta(\mathbf{z}_j|\mathbf{z}_i^{(g)}),$$

if the conditional density is available for some

$i \neq j$. The rationale for this is that

$$p_\Theta(\mathbf{z}_j) = \int p_\Theta(\mathbf{z}_j|\mathbf{z}_i) p_\Theta(\mathbf{z}_i)\, d\mathbf{z}_i = E_{\mathbf{Z}_i}\left[p_\Theta(\mathbf{z}_j|\mathbf{Z}_i)\right].$$

- In the above, $\mathbf{z}_t$ can represent parameter values θ, state values $\mathbf{x}_t$ or future observations $\mathbf{y}_{n+l}$. As an example, consider the Gaussian DLM with univariate observations:

$$\begin{aligned} \mathbf{x}_t &= \mathbf{\Phi}\mathbf{x}_{t-1} + \mathbf{w}_t, \mathbf{x}_0 \sim N(\mu_0, \mathbf{\Sigma}_0), \mathbf{w}_t \sim N(\mathbf{0}, \mathbf{Q}); \\ y_t &= \boldsymbol{\alpha}_t' \mathbf{x}_t + v_t, v_t \sim N(0, r). \end{aligned}$$

  In the notation of the development above, let $\mathbf{z}$ represent the parameter vector Θ, so that we want the posterior density $p(\Theta|y^n = (y_1, ..., y_n))$ (w.r.t. to a prior $\pi(\Theta)$). With $\mathbf{x}^n = (\mathbf{x}_0, ..., \mathbf{x}_n)$ this is

$$p(\Theta|y^n) = \int\int p(\Theta|\mathbf{x}^n, y^n)\, p(\mathbf{x}^n, \Theta^*|y^n)\, d\mathbf{x}^n d\Theta^*.$$

Reason:

$$\begin{aligned} p(\Theta|y^n) &= \frac{p(\Theta, y^n)}{p(y^n)} \\ &= \int \frac{p(\Theta, y^n, \mathbf{x}^n)}{p(y^n)} d\mathbf{x}^n \\ &= \int \frac{p(\Theta|y^n, \mathbf{x}^n)}{p(y^n)} p(y^n, \mathbf{x}^n)\, d\mathbf{x}^n \\ &= \int\int p(\Theta|y^n, \mathbf{x}^n) \frac{p(y^n, \mathbf{x}^n, \Theta^*)}{p(y^n)} d\mathbf{x}^n d\Theta^* \\ &= \int\int p(\Theta|\mathbf{x}^n, y^n)\, p(\mathbf{x}^n, \Theta^*|y^n)\, d\mathbf{x}^n d\Theta^*. \end{aligned}$$

- Assume that $p(\Theta|\mathbf{x}^n, y^n)$ is available. Simulate draws $X_n^{(g)} \stackrel{def}{=} (\mathbf{x}^n, \Theta^*)^{(g)}$ from $p(\mathbf{x}^n, \Theta^*|y^n)$ and write the above as

$$p(\Theta|y^n) = E_{X_n}\left[p(\Theta|\mathbf{x}^n, y^n)\right];$$

approximate it by

$$\hat{p}(\Theta|y^n) = \frac{1}{G}\sum_{g=1}^{G} p\left(\Theta|X_n^{(g)}, y^n\right).$$

- Simulating the $X_n^{(g)}$ requires two different MCMC procedures. One alternates between sampling $X_{n[l]}$ given $\Theta^*_{[l-1]}$ - i.e. sampling $\mathbf{x}^n$ with l suitably large - from $p\left(\mathbf{x}^n|\Theta^*_{[l-1]}, y^n\right)$, and sampling $\Theta^*_{[l]}$ from $p\left(\Theta|\mathbf{x}^n_{[l]}, y^n\right)$.

1. To sample $X_{n[l]}$ given $\Theta^*_{[l-1]}$ from $p\left(\mathbf{x}^n|\Theta^*_{[l-1]}, y^n\right)$, apply Gibbs sampling as follows. At this stage Θ is fixed and known, and we are to sample the entire sequence $\mathbf{x}^n$ from $p_{\Theta^*_{[l-1]}}(\mathbf{x}^n|y^n)$, where

$$\begin{aligned} & p_\Theta\left(\mathbf{x}^n|y^n\right) \\ = \; & p_\Theta\left(\mathbf{x}_n|y^n\right) p_\Theta\left(\mathbf{x}_{n-1}|\mathbf{x}_n, y^{n-1}\right) \cdots p_\Theta\left(\mathbf{x}_0|\mathbf{x}_1\right). \end{aligned}$$

Reason:

$$\begin{aligned}
p_{\Theta}\left(\mathbf{x}^n|y^n\right) &= p\left(\mathbf{x}_n, \mathbf{x}^{n-1}|y^n\right) \\
&= \frac{p\left(\mathbf{x}_n, \mathbf{x}^{n-1}, y^n\right)}{p\left(y^n\right)} \\
&= \frac{p\left(\mathbf{x}^{n-1}|\mathbf{x}_n, y^n\right) p\left(\mathbf{x}_n, y^n\right)}{p\left(y^n\right)} \\
&= p\left(\mathbf{x}_n|y^n\right) p\left(\mathbf{x}^{n-1}|\mathbf{x}_n, y^n\right) \\
&= p\left(\mathbf{x}_n|y^n\right) p\left(\mathbf{x}^{n-1}|\mathbf{x}_n, y^{n-1}\right) \\
&\quad ... \\
&= p\left(\mathbf{x}_n|y^n\right) p\left(\mathbf{x}_{n-1}|\mathbf{x}_n, y^{n-1}\right) \\
&\quad \cdot p\left(\mathbf{x}^{n-2}|\mathbf{x}_{n-1}, y^{n-2}\right) \\
&\quad ...
\end{aligned}$$

This allows for a backwards simulation scheme. These conditionals are all Gaussian. The first - $p_{\Theta}\left(\mathbf{x}_n|y^n\right)$ - has mean $\mathbf{x}_n^n$ and covariance $\mathbf{P}_n^n$. The others have means and covariances

$$\begin{aligned}
\mathbf{m}_t &= E\left[\mathbf{x}_t|\mathbf{x}_{t+1}, y^t\right] = \mathbf{x}_t^t + \mathbf{J}_t\left(\mathbf{x}_{t+1} - \mathbf{x}_{t+1}^t\right), \\
\mathbf{V}_t &= \text{cov}\left[\mathbf{x}_t|\mathbf{x}_{t+1}, y^t\right] = \mathbf{P}_t^t - \mathbf{J}_t\mathbf{P}_{t+1}^t\mathbf{J}_t',
\end{aligned}$$

for $\mathbf{J}_t = \mathbf{P}_t^t \Phi' \left[\mathbf{P}_{t+1}^t\right]^{-1}$.

Reason: From the Kalman smoothing equations, with $n = t$,

$$E\left[\mathbf{x}_{t-1}|y^t\right] = \mathbf{x}_{t-1}^t = \mathbf{x}_{t-1}^{t-1} + \mathbf{J}_{t-1}\left(\mathbf{x}_t^t - \mathbf{x}_t^{t-1}\right);$$

replacing t by $t+1$ gives

$$E\left[\mathbf{x}_t|y^{t+1}\right] = \mathbf{x}_t^t + \mathbf{J}_t\left(\mathbf{x}_{t+1}^{t+1} - \mathbf{x}_{t+1}^t\right).$$

Thus

$$E\left[\mathbf{x}_t|\mathbf{x}_{t+1}, y^t, y_{t+1}\right] = \mathbf{x}_t^t + \mathbf{J}_t\left(\mathbf{x}_{t+1} - \mathbf{x}_{t+1}^t\right),$$

which does not depend on y_{t+1}. Verification of the expression for $\mathbf{V}_t$ is similar. Thus, one first samples $\mathbf{x}_n$ from a $N\left(\mathbf{x}_n^n, \mathbf{P}_n^n\right)$ density (this mean and covariance are obtained from the Kalman filter), then samples $\mathbf{x}_t$ from a $N\left(\mathbf{m}_t, \mathbf{V}_t\right)$ density for $t = n-1, ..., 0$.

2. Now sample $\Theta_{[l]}^*$ from $p\left(\Theta|\mathbf{x}_{[l]}^n, y^n\right)$ - some suggestions for this are in the text.

- In the case of nonlinear state and observation equations, Carlin, Polson & Stoffer (paper on course website) take nonlinear equations with additive noise:

$$\begin{aligned} \mathbf{x}_t &= F_t\left(\mathbf{x}_{t-1}\right) + \mathbf{w}_t, \\ \mathbf{y}_t &= H_t\left(\mathbf{x}_t\right) + \mathbf{v}_t. \end{aligned}$$

  The distributions of $\mathbf{w}_t, \mathbf{v}_t$ are scale mixtures of normals, e.g.

$$p\left(\mathbf{w}_t\right) = \int_0^\infty \left|2\pi\lambda\mathbf{Q}\right|^{-1/2} e^{-\frac{\mathbf{w}_t'\mathbf{Q}^{-1}\mathbf{w}_t}{2\lambda}} \pi\left(\lambda\right) d\lambda,$$

  for a density $\pi\left(\lambda\right)$. The interpretations is that, given λ, $\mathbf{w}_t \sim N\left(\mathbf{0}, \lambda\mathbf{Q}\right)$. Similarly, given ω, $\mathbf{v}_t \sim N\left(\mathbf{0}, \omega\mathbf{R}\right)$. The class of scale mixtures of normals includes many common distributions. For instance, in one dimension if λ^{-1} has a Gamma distribution then w has a t-distribution, $\lambda \sim$ Exponential results in $w \sim$ Laplace; a more complicated mixing density for λ results in $w \sim$ logistic.

- For this class of distributions, conditioning on λ and ω allows much of the methodology for normal models to be applied - see the text; more details are in the paper.

16. Analysis of longitudinal data

- Suppose one makes k-dimensional observations $\mathbf{y}_t$, over time, on each of N independent individuals. If each follows an ARMAX model then the data are

$$\mathbf{y}_{tl} = \Gamma_{k\times g}\mathbf{u}_{tl} + \sum_{i=1}^{p} \Phi_i \mathbf{y}_{t-i,l} + \mathbf{w}_{tl} + \sum_{j=1}^{q} \Theta_j \mathbf{w}_{t-j,l},$$

for $l = 1, ..., N$ and $t = 1, ..., n$ and cov$[\mathbf{w}_{tl}] = \Sigma_w$. As before, this can be represented in state-space form.

- Note that this is more than analyzing the N models individually, since the models have parameters in common.

- Here the concentration is on ARX modelling:

$$\begin{aligned}
\mathbf{y}_{tl} &= \mathbf{B}\mathbf{z}_{tl} + \mathbf{w}_{tl},\ l = 1, ..., N,\ t = p+1, ..., n \\
\text{for } \mathbf{B} &= \left(\begin{array}{cccc} \Gamma & \Phi_1 & \cdots & \Phi_p \end{array} \right) : k \times (pk + g)\,, \\
\mathbf{z}_{tl} &= \left(\begin{array}{c} \mathbf{u}_{tl} \\ \mathbf{y}_{t-1,l} \\ \vdots \\ \mathbf{y}_{t-p,l} \end{array} \right).
\end{aligned}$$

This can be viewed as N independent replicates of the multivariate regression model considered in Lecture 8. For each l, the (conditional) log-likelihood derived there (times -2) was

$$\begin{aligned}
& (n-p) \log |\mathbf{\Sigma}| \\
& + tr \mathbf{\Sigma}^{-1} \left[\sum_{t=p+1}^{n} \left(\mathbf{y}_{tl} - \mathbf{B}\mathbf{z}_{tl} \right) \left(\mathbf{y}_{tl} - \mathbf{B}\mathbf{z}_{tl} \right)' \right] \\
= \ & (n-p) \log |\mathbf{\Sigma}| + tr \mathbf{\Sigma}^{-1} \left(\mathbf{Y}_l' - \mathbf{B}\mathbf{Z}_l' \right) \left(\mathbf{Y}_l' - \mathbf{B}\mathbf{Z}_l' \right)',
\end{aligned}$$

where

$$\mathbf{Y}_l = \left(\begin{array}{c} \mathbf{y}_{p+1,l}' \\ \vdots \\ \mathbf{y}_{n,l}' \end{array} \right), \mathbf{Z}_l = \left(\begin{array}{c} \mathbf{z}_{p+1,l}' \\ \vdots \\ \mathbf{z}_{n,l}' \end{array} \right) : (n-p) \times (pk + g)\,.$$

Then

$$\begin{aligned}
\hat{\mathbf{B}} &= \mathbf{Y}_l'\mathbf{Z}_l\left(\mathbf{Z}_l'\mathbf{Z}_l\right)^{-1} \\
&= \left[\sum_{t=p+1}^{n}\mathbf{y}_{t,l}\mathbf{z}_{t,l}'\right]\left[\sum_{t=p+1}^{n}\mathbf{z}_{t,l}\mathbf{z}_{t,l}'\right]^{-1}, \\
\hat{\mathbf{\Sigma}}_w &= \frac{1}{n-p-1}\sum_{t=p+1}^{n}\left(\mathbf{y}_{tl}-\hat{\mathbf{B}}\mathbf{z}_{tl}\right)\left(\mathbf{y}_{tl}-\hat{\mathbf{B}}\mathbf{z}_{tl}\right)'.
\end{aligned}$$

In the current framework the log-likelihood above is summed over $l = 1, ..., N$. Equivalently, the $\mathbf{Y}_l$ are stacked, as are the $\mathbf{Z}_l$. This results in

$$\begin{aligned}
\hat{\mathbf{B}} &= \left[\sum_{l=1}^{N}\sum_{t=p+1}^{n}\mathbf{y}_{t,l}\mathbf{z}_{t,l}'\right]\left[\sum_{l=1}^{N}\sum_{t=p+1}^{n}\mathbf{z}_{t,l}\mathbf{z}_{t,l}'\right]^{-1}, \\
\hat{\mathbf{\Sigma}}_w &= \frac{\sum_{l=1}^{N}\sum_{t=p+1}^{n}\left(\mathbf{y}_{tl}-\hat{\mathbf{B}}\mathbf{z}_{tl}\right)\left(\mathbf{y}_{tl}-\hat{\mathbf{B}}\mathbf{z}_{tl}\right)'}{N\left(n-p-1\right)}.
\end{aligned}$$

- A less parsimonious model allows for time-varying parameters:

$$\begin{aligned}
\mathbf{y}_{tl} &= \Gamma_t \mathbf{u}_{tl} + \sum_{i=1}^{p_t} \Phi_{ti} \mathbf{y}_{t-i,l} + \mathbf{w}_{tl}, \\
&= \mathbf{B}_t \mathbf{z}_{tl} + \mathbf{w}_{tl}, \\
(l &= 1, ..., N, t = p_t + 1, ..., n), \text{ and} \\
\text{cov}\left[\mathbf{w}_{tl}\right] &= \Sigma_t.
\end{aligned}$$

The data then consist of n regressions, each with its own parameters, and so estimation of the regression parameter $(\mathbf{B})$ is exactly as in Lecture 8, repeated n times (holding t fixed each time):

$$\begin{aligned}
\hat{\mathbf{B}}_t &= \left[\sum_{l=1}^{N} \mathbf{y}_{t,l} \mathbf{z}'_{t,l}\right] \left[\sum_{l=1}^{N} \mathbf{z}_{t,l} \mathbf{z}'_{t,l}\right]^{-1}, \\
\hat{\Sigma}_t &= \frac{1}{N - p_t - 1} \sum_{l=1}^{N} \left(\mathbf{y}_{tl} - \hat{\mathbf{B}}_t \mathbf{z}_{tl}\right) \left(\mathbf{y}_{tl} - \hat{\mathbf{B}}_t \mathbf{z}_{tl}\right)'.
\end{aligned}$$

- Example: S&S discuss the following study (but the data are not given, so only the models are presented here). There were $N = 318$ children examined at $n = 4$ times - at ages 8, 18, 36 and 48 months ($t = 1, 2, 3, 4$; $t = 0$ is birth). Growth indices y_{tl} ($t = 1, 2, 3, 4$; $l = 1, ..., 318$) were calculated - a growth index is weight adjusted for age, gender and height. The purpose of the study is to examine the effect of prenatal smoking. The model is

$$\begin{aligned} y_{tl} &= \gamma_{0t} + \gamma_{1t} S_l + \gamma_{2t} R_l + \gamma_{3t} S_l R_l \\ &\quad + \sum_{j=1}^{t} \phi_{tj} \left(y_{t-j,l} - \hat{y}_{t-j,l} \right) + w_{tl}, \end{aligned}$$

for $t = 0, ..., 4$ with var$[w_{tl}] = \sigma_t^2$. Here

$$\begin{aligned} \mathbf{\Gamma}_t &= (\gamma_{0t}, \gamma_{1t}, \gamma_{2t}, \gamma_{3t}), \\ \mathbf{u}'_{tl} &= (1, S_l, R_l, S_l R_l) \end{aligned}$$

($\mathbf{u}_{tl}$ does not vary with time) with exogenous variables

$$\begin{aligned} S_l &= \text{average \# of cig's smoked daily by mother } l, \\ R_l &= I\,(\text{race of mother is white, rather than black}). \end{aligned}$$

Then $\hat{y}_{t,l}$ is the fitted value of y_{tl} in the regression model

$$\begin{aligned} E[y_{tl}] &= \gamma_{0t} + \gamma_{1t}S_l + \gamma_{2t}R_l + \gamma_{3t}S_lR_l \\ &= \begin{cases} \gamma_{0t} + \gamma_{1t}S_l, & \text{if black,} \\ (\gamma_{0t} + \gamma_{2t}) + (\gamma_{1t} + \gamma_{3t})S_l, & \text{if white,} \end{cases} \end{aligned}$$

based on data $\{y_{tl}\}_{l=1}^{n}$. (Thus $\hat{y}_{t,l}$ is a non-linear function of the exogenous inputs.) The purpose of including the residual $y_{t-j,l} - \hat{y}_{t-j,l}$ is to eliminate effects of smoking or race on *previous* growth, while retaining the effect of previous growth on current growth.

For example, at birth ($t = 0$) the model for birthweight is

$$y_{0l} = \gamma_{00} + \gamma_{10}S_l + \gamma_{20}R_l + \gamma_{30}S_lR_l + w_{0l},$$

and the fitted values are

$$\hat{y}_{0l} = \hat{\gamma}_{00} + \hat{\gamma}_{10}S_l + \hat{\gamma}_{20}R_l + \hat{\gamma}_{30}S_lR_l.$$

At eight months ($t = 1$) the model is

$$\begin{aligned} y_{1l} &= \gamma_{01} + \gamma_{11}S_l + \gamma_{21}R_l + \gamma_{31}S_lR_l \\ &\quad + \phi_{11}\left(y_{0,l} - \hat{y}_{0,l}\right) + w_{1l}. \end{aligned}$$

The effect of smoking and race on birthweight has been removed, and the parameters now measure their effect in the current time period. At 18 months ($t = 2$) the model is

$$\begin{aligned} y_{2l} &= \gamma_{02} + \gamma_{12} S_l + \gamma_{22} R_l + \gamma_{32} S_l R_l \\ &\quad + \phi_{21} \left(y_{1,l} - \hat{y}_{1,l} \right) + \phi_{22} \left(y_{0,l} - \hat{y}_{0,l} \right) + w_{2l}, \end{aligned}$$

where

$$\begin{aligned} y_{1l} &= \gamma_{01} + \gamma_{11} S_l + \gamma_{21} R_l + \gamma_{31} S_l R_l + w_{1l}, \\ \hat{y}_{1l} &= \hat{\gamma}_{01} + \hat{\gamma}_{11} S_l + \hat{\gamma}_{21} R_l + \hat{\gamma}_{31} S_l R_l, \end{aligned}$$

etc.

The S&S analysis results in the following. Dropping terms deemed to be statistically insignificant, they report

$$\begin{aligned} \hat{y}_{0l} &= \hat{\gamma}_{00} - .011_{(.002)} S_l + \hat{\gamma}_{20} R_l, \\ \hat{y}_{1l} &= \hat{\gamma}_{11} S_l + \hat{\gamma}_{21} R_l + \hat{\gamma}_{31} S_l R_l \\ &\quad + .214_{(.127)} \left(y_{0,l} - \hat{y}_{0,l} \right), \end{aligned}$$

$$
\begin{aligned}
\hat{y}_{2l} &= \hat{\gamma}_{02} + .278_{(.125)} R_l + \hat{\phi}_{21}\left(y_{1,l} - \hat{y}_{1,l}\right) \\
&\quad + \hat{\phi}_{22}\left(y_{0,l} - \hat{y}_{0,l}\right), \\
\hat{y}_{3l} &= \hat{\gamma}_{03} + .008_{(.004)} S_l + \hat{\phi}_{31}\left(y_{2,l} - \hat{y}_{2,l}\right) \\
&\quad + \hat{\phi}_{32}\left(y_{1,l} - \hat{y}_{1,l}\right) + \hat{\phi}_{33}\left(y_{0,l} - \hat{y}_{0,l}\right), \\
\hat{y}_{4l} &= \hat{\gamma}_{04} + \hat{\phi}_{41}\left(y_{3,l} - \hat{y}_{3,l}\right) + \hat{\phi}_{42}\left(y_{2,l} - \hat{y}_{2,l}\right) \\
&\quad + \hat{\phi}_{43}\left(y_{1,l} - \hat{y}_{1,l}\right).
\end{aligned}
$$

The first of these indicates that smoking significantly decreases birthweight. In the second the interaction term is significant and so the effect of smoking depends on race. As well, current growth is significantly affected by birthweight. At $t = 2$ (18 months) the effect of smoking is gone and white children tend to be larger. At $t = 3$ the effect of race is gone but an effect of smoking returns, indicating an (obesity?) effect of smoking. At $t = 4$ there is no effect of smoking or race, and the effect of birthweight is gone while that of growth in the first three time periods remains.

16.1. Mixed linear models

- A mixed model for longitudinal data: the vector $\mathbf{y}_l$ of responses for individual l is modelled as

$$\begin{aligned}
\mathbf{y}_l &= \mathbf{X}_l\boldsymbol{\beta} + \mathbf{Z}_l\boldsymbol{\gamma}_l + \varepsilon_l, \\
\mathbf{X}_l &: n_l \times b \text{ a fixed design matrix,} \\
\boldsymbol{\beta} &: b \times 1 \text{ fixed parameters,} \\
\mathbf{Z}_l &: n_l \times g \text{ a fixed design matrix,} \\
\boldsymbol{\gamma}_l &: g \times 1 \text{ random effects;} \\
\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_N &\sim \text{i.i.d. } N(\mathbf{0}, \mathbf{D}), \\
\varepsilon_l &\sim \text{ind. } N(\mathbf{0}, \boldsymbol{\Sigma}_l) \text{ (ind. of } \{\boldsymbol{\gamma}_l\}_{l=1}^N).
\end{aligned}$$

  e.g. a patient's health profile may be modelled in terms of fixed effects (treatments) and random effects (blood pressure).

- A consequence is that

$$\mathbf{y}_l \overset{ind.}{\sim} N\left(\mathbf{X}_l\boldsymbol{\beta}, \mathbf{Z}_l\mathbf{D}\mathbf{Z}_l' + \boldsymbol{\Sigma}_l \overset{def}{=} \mathbf{V}_l\right).$$

e.g.1. "compound symmetry": $g = 1$ with

$$\begin{aligned} \mathbf{Z}_l &= \mathbf{1}_{n_l}, \\ \mathbf{D} &= \sigma_\gamma^2, \\ \mathbf{\Sigma}_l &= \sigma^2 \mathbf{I}_{n_l}; \end{aligned}$$

thus

$$\begin{aligned} \mathbf{V}_l &= \sigma_\gamma^2 \mathbf{1}_{n_l} \mathbf{1}'_{n_l} + \sigma^2 \mathbf{I}_{n_l} \\ &= \begin{pmatrix} \sigma^2 + \sigma_\gamma^2 & \sigma_\gamma^2 & \cdots & \sigma_\gamma^2 \\ \sigma_\gamma^2 & \sigma^2 + \sigma_\gamma^2 & \cdots & \sigma_\gamma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\gamma^2 & \sigma_\gamma^2 & \cdots & \sigma^2 + \sigma_\gamma^2 \end{pmatrix} \\ &= \left(\sigma^2 + \sigma_\gamma^2\right) \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}, \end{aligned}$$

with $\rho = \frac{\sigma_\gamma^2}{\sigma^2 + \sigma_\gamma^2} = \frac{1}{1 + \frac{\sigma^2}{\sigma_\gamma^2}}$.

e.g.2. $g = 0$ (no random effects) with autoregressive structure:

$$\mathbf{V}_l = \mathbf{\Sigma}_l = \sigma^2 \left\{ \rho^{|i-j|} \right\}_{i,j=1}^{n_l}. \tag{16.1}$$

- Write $\mathbf{y}_l = \left(y_{1l}, ..., y_{tl}, ...y_{n_l,l}\right)'$, so y_{tl} is the measurement on individual l at time t. If $\mathbf{x}'_{tl}$ and $\mathbf{z}'_{tl}$ are the t^{th} rows of $\mathbf{X}_l$ and $\mathbf{Z}_l$, and $\mathbf{\Sigma}_l = \left\{\sigma_{l;t,s}\right\}_{t,s=1}^{n_l}$, then

$$\begin{aligned} y_{tl} &= \mathbf{x}'_{tl}\boldsymbol{\beta} + \mathbf{z}'_{tl}\boldsymbol{\gamma}_l + \varepsilon_{tl} \\ &\sim N\left(\mathbf{x}'_{tl}\boldsymbol{\beta},\ \mathbf{z}'_{tl}\mathbf{D}\mathbf{z}_{tl} + \sigma_{l;t,t}\right); \\ \text{cov}\left[y_{tl}, y_{sl}\right] &= \mathbf{z}'_{tl}\mathbf{D}\mathbf{z}_{sl} + \sigma_{l;t,s}, \\ \text{cov}\left[y_{tl}, y_{sk}\right] &= 0 \text{ if } k \neq l. \end{aligned}$$

- State space formulation. For simplicity take $\mathbf{\Sigma}_l = \sigma^2\mathbf{I}_{n_l}$ (a common choice). Consider the state-space model

$$\begin{aligned} y_{tl} &= \mathbf{x}'_{tl}\boldsymbol{\beta} + \mathbf{z}'_{tl}\mathbf{s}_{tl} + \varepsilon_{tl}, \\ \mathbf{s}_{tl} &= \mathbf{s}_{t-1,l} + \mathbf{w}_{tl}, \\ \mathbf{s}_{0l} &\sim N\left(\mathbf{0}, \mathbf{D}\right), \\ \mathbf{w}_{tl} &\sim N\left(\mathbf{0}, \mathbf{Q}\right) \end{aligned}$$

  with $\mathbf{Q} = \mathbf{0}$. Then $\mathbf{w}_{tl}$ is the zero vector, the 'state-space vector' $\mathbf{s}_{tl}$ is $\mathbf{s}_{0l}$, i.e. is $\boldsymbol{\gamma}_l$, and this state-space model is the same as the model above.

– If there are no random effects and an autoregressive structure is assumed, then this becomes

$$\begin{aligned} y_{tl} &= \mathbf{x}'_{tl}\boldsymbol{\beta} + \varepsilon_{tl}, \\ \varepsilon_{tl} &= \rho\varepsilon_{t-1,l} + w_{tl}. \end{aligned}$$

Following S&S, we now write s_{tl} in place of ε_{tl}:

$$\begin{aligned} y_{tl} &= \mathbf{x}'_{tl}\boldsymbol{\beta} + s_{tl}, \\ s_{tl} &= \rho s_{t-1,l} + w_{tl}; \end{aligned}$$

this is a state-space model with no observation errors ($R = 0$) but $Q = \sigma_w^2 > 0$. The correct AR(1) autocorrelation structure (16.1) is obtained if $s_{0l} \sim N\left(0, \frac{\sigma_w^2}{1-\rho^2}\right)$.

– If there <u>are</u> random effects and an autoregressive structure is assumed, then take

$$\begin{aligned} y_{tl} &= \mathbf{x}'_{tl}\boldsymbol{\beta} + \mathbf{A}_t\mathbf{s}_{tl}, \\ \mathbf{s}_{tl} &= \boldsymbol{\Phi}\mathbf{s}_{t-1,l} + \mathbf{w}_{tl}, \end{aligned}$$

with $\mathbf{s}_{tl} : (g+1) \times 1$ and

$$\begin{aligned}
\mathbf{\Phi} &= \begin{pmatrix} \mathbf{I}_g & \mathbf{0} \\ \mathbf{0}' & \rho \end{pmatrix}, \ \mathbf{A}_t = \begin{pmatrix} \mathbf{z}'_{tl} & 1 \end{pmatrix}, \\
\mathbf{Q} &= \text{cov}\left[\mathbf{w}_{tl}\right] = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & \sigma_w^2 \end{pmatrix}, \ R = 0.
\end{aligned}$$

Reason: Partition $\mathbf{s}_{tl}$ as

$$\mathbf{s}_{tl} = \begin{pmatrix} \tilde{\mathbf{s}}_{tl} \\ \varepsilon_{tl} \end{pmatrix} \begin{matrix} \leftarrow g \\ \leftarrow 1 \end{matrix};$$

then the above is

$$\begin{aligned}
y_{tl} &= \mathbf{x}'_{tl}\boldsymbol{\beta} + \mathbf{z}'_{tl}\tilde{\mathbf{s}}_{tl} + \varepsilon_{tl} \\
\tilde{\mathbf{s}}_{tl} &= \tilde{\mathbf{s}}_{t-1,l} \ (= \tilde{\mathbf{s}}_{0l} = \boldsymbol{\gamma}_l), \\
\varepsilon_{tl} &= \rho\varepsilon_{t-1,l} + w_{tl};
\end{aligned}$$

this gives the required model if we impose the initial conditions

$$\begin{aligned}
\tilde{\mathbf{s}}_{0l} &\sim N(\mathbf{0}, \mathbf{D}), \\
\varepsilon_{0l} &\sim N\left(0, \frac{\sigma_w^2}{1-\rho^2}\right).
\end{aligned}$$

- Example 6.24 - assigned.

17. Multivariate frequency domain methods - Introduction

- Possible applications:
  - SOI and Recruits - we earlier did an impulse-response analysis to identify a lagged regression relationship (conditional on the input process) between these series; here we might extend this type of analysis to multiple inputs.
  - Similar analyses might view the inputs as fixed (non-random) series - indicators, for instance - as in a designed experiment. This is akin to ANOVA in classical statistics, but carried out in the frequency domain.
  - Random coefficients, discrimination, clustering, principal component analysis, ... ; all can be extended to the frequency domain.

17.1. The Complex Normal distribution

- For any complex r.v. $U = X - iY$ (X, Y real) we define $E[U] = E[X] - iE[Y]$ and

$$\begin{aligned} \text{var}[U] &= E\left[(U-\mu_U)\left(\overline{U-\mu_U}\right)\right] \\ = \quad & E\left[|U-\mu_U|^2\right] = \text{var}[X] + \text{var}[Y]. \end{aligned}$$

Similarly the complex random vector $\mathbf{z} = \mathbf{x}_1 - i\mathbf{x}_2$ has mean vector

$$E[\mathbf{z}] = \mu_{\mathbf{z}} = \mu_{\mathbf{x}_1} - i\mu_{\mathbf{x}_2}$$

and covariance matrix

$$E\left[(\mathbf{z}-\mu_{\mathbf{z}})(\mathbf{z}-\mu_{\mathbf{z}})^*\right] = \Sigma_{\mathbf{z}} = \mathbf{C} - i\mathbf{Q},$$

with $\mathbf{C}$, $\mathbf{Q}$ real. A consequence is that $\Sigma_{\mathbf{z}} = \Sigma_{\mathbf{z}}^*$, i.e. $\Sigma_{\mathbf{z}}$ is *Hermitian*; thus $\mathbf{C}$ is symmetric and $\mathbf{Q}$ is skew-symmetric: $\mathbf{Q} = -\mathbf{Q}'$. Then a further consequence is that, for any real vector $\mathbf{a}$ we have (since $\mathbf{a}^* = \mathbf{a}'$ and $\mathbf{a}'\mathbf{Q}\mathbf{a} \equiv 0$)

$$\begin{aligned} 0 \quad \leq \quad & \text{var}[\mathbf{a}^*\mathbf{z}] = E\left[|\mathbf{a}^*(\mathbf{z}-\mu_{\mathbf{z}})|^2\right] = \mathbf{a}^*\Sigma_{\mathbf{z}}\mathbf{a} \\ = \quad & \mathbf{a}'\mathbf{C}\mathbf{a} - i\mathbf{a}'\mathbf{Q}\mathbf{a} = \mathbf{a}'\mathbf{C}\mathbf{a}; \end{aligned}$$

thus $\mathbf{C}$ is positive semi-definite.

- Recall that a real r.vec. has a multivariate normal distribution iff all linear combinations have univariate normal distributions. To see the implications of complex normality we start with the definition "A complex random vector $\mathbf{z} = \mathbf{x}_1 - i\mathbf{x}_2$ with mean vector $\mu_{\mathbf{z}}$ and covariance matrix $\Sigma_{\mathbf{z}}$ has the complex normal distribution (we write $\mathbf{z} \sim CN_p(\mu_{\mathbf{z}}, \Sigma_{\mathbf{z}})$) if *the variances of the real and imaginary parts of every linear combination* $\mathbf{a}^*\mathbf{z}$ *are equal, and these linear combinations are normally distributed*."
  To see what this entails, set

  $$\Sigma_{11} = \mathrm{cov}[\mathbf{x}_1], \Sigma_{22} = \mathrm{cov}[\mathbf{x}_2], \Sigma_{12} = \mathrm{cov}[\mathbf{x}_1, \mathbf{x}_2].$$

  Then for every real $\mathbf{a}_1, \mathbf{a}_2$ and with $\mathbf{a} = \mathbf{a}_1 + i\mathbf{a}_2$ we have

  $$\mathbf{a}^*\mathbf{z} = \left(\mathbf{a}_1'\mathbf{x}_1 - \mathbf{a}_2'\mathbf{x}_2\right) - i\left(\mathbf{a}_2'\mathbf{x}_1 + \mathbf{a}_1'\mathbf{x}_2\right)$$

  and so $\mathrm{var}\left[\mathbf{a}_1'\mathbf{x}_1 - \mathbf{a}_2'\mathbf{x}_2\right] = \mathrm{var}\left[\mathbf{a}_2'\mathbf{x}_1 + \mathbf{a}_1'\mathbf{x}_2\right]$, i.e.

  $$\mathbf{a}_1'\Sigma_{11}\mathbf{a}_1 + \mathbf{a}_2'\Sigma_{22}\mathbf{a}_2 - 2\mathbf{a}_1'\Sigma_{12}\mathbf{a}_2 = \mathbf{a}_2'\Sigma_{11}\mathbf{a}_2 + \mathbf{a}_1'\Sigma_{22}\mathbf{a}_1 + 2\mathbf{a}_2'\Sigma_{12}\mathbf{a}_1.$$

With $\mathbf{a}_1 = \mathbf{a}_2$ we get that $\mathbf{a}_1' \boldsymbol{\Sigma}_{12} \mathbf{a}_1 \equiv 0$, so that the real, symmetric matrix $\boldsymbol{\Sigma}_{12} + \boldsymbol{\Sigma}_{12}'$ has all eigenvalues $= 0$; thus $\boldsymbol{\Sigma}_{12}$ is skew-symmetric. With $\mathbf{a}_2 = 0$ we get

$$\mathbf{a}_1' \left(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{22} \right) \mathbf{a}_1 \equiv 0;$$

thus $\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22}$.
Now note that

$$\begin{aligned} \text{var}\left[\mathbf{a}^* \mathbf{z}\right] &= \text{var}\left[\mathbf{a}_1' \mathbf{x}_1 - \mathbf{a}_2' \mathbf{x}_2\right] + \text{var}\left[\mathbf{a}_2' \mathbf{x}_1 + \mathbf{a}_1' \mathbf{x}_2\right] \\ &= 2\left(\mathbf{a}_1' \boldsymbol{\Sigma}_{11} \mathbf{a}_1 + \mathbf{a}_2' \boldsymbol{\Sigma}_{11} \mathbf{a}_2 - 2\mathbf{a}_1' \boldsymbol{\Sigma}_{12} \mathbf{a}_2\right), \end{aligned}$$

since the two variances are by definition equal. But also

$$\begin{aligned} \text{var}\left[\mathbf{a}^* \mathbf{z}\right] &= \mathbf{a}^* \boldsymbol{\Sigma}_{\mathbf{z}} \mathbf{a} \\ &= \left(\mathbf{a}_1' - i\mathbf{a}_2'\right) \left(\mathbf{C} - i\mathbf{Q}\right) \left(\mathbf{a}_1 + i\mathbf{a}_2\right) \\ &= \ldots \\ &= \mathbf{a}_1' \mathbf{C} \mathbf{a}_1 + \mathbf{a}_2' \mathbf{C} \mathbf{a}_2 + 2\mathbf{a}_1' \mathbf{Q} \mathbf{a}_2; \end{aligned}$$

thus

$$\begin{aligned} & \mathbf{a}_1' \boldsymbol{\Sigma}_{11} \mathbf{a}_1 + \mathbf{a}_2' \boldsymbol{\Sigma}_{11} \mathbf{a}_2 - 2\mathbf{a}_1' \boldsymbol{\Sigma}_{12} \mathbf{a}_2 \\ \equiv \; & \frac{\mathbf{a}_1' \mathbf{C} \mathbf{a}_1 + \mathbf{a}_2' \mathbf{C} \mathbf{a}_2}{2} + \mathbf{a}_1' \mathbf{Q} \mathbf{a}_2. \qquad (17.1) \end{aligned}$$

- As above, (17.1) implies that

$$\begin{aligned} \text{cov}\left[\mathbf{x}_1\right] &= \Sigma_{11} = \Sigma_{22} = \text{cov}\left[\mathbf{x}_2\right] = \frac{1}{2}\mathbf{C}, \\ \text{cov}\left[\mathbf{x}_1, \mathbf{x}_2\right] &= \Sigma_{12} = -\Sigma_{12}' = -\frac{1}{2}\mathbf{Q}. \end{aligned}$$

  The real part of $\mathbf{a}^*\mathbf{z}$ is $\left(\mathbf{a}_1'\mathbf{x}_1 - \mathbf{a}_2'\mathbf{x}_2\right)$, and is normally distributed. Thus all linear combinations of $\mathbf{x} = \binom{\mathbf{x}_1}{\mathbf{x}_2}$ are normally distributed, and so another consequence of complex normality is that

$$\mathbf{x} = \binom{\mathbf{x}_1}{\mathbf{x}_2} \sim N_{2p}\left(\mu_{\mathbf{x}} = \binom{\mu_{\mathbf{x}_1}}{\mu_{\mathbf{x}_2}}, \Sigma_{\mathbf{x}} = \frac{1}{2}\begin{pmatrix} \mathbf{C} & -\mathbf{Q} \\ \mathbf{Q} & \mathbf{C} \end{pmatrix}\right), \tag{17.2}$$

  where $\mathbf{C}$ is symmetric and $\mathbf{Q}$ is skew-symmetric.

- Conversely, from (17.2) it follows that

$$\mathbf{z} = \mathbf{x}_1 - i\mathbf{x}_2 \sim CN_p\left(\mu_{\mathbf{z}}, \Sigma_{\mathbf{z}}\right).$$

- If $\Sigma_{\mathbf{z}}$ is nonsingular then the density of $\mathbf{z}$, which is identified with that of $\mathbf{x}$, exists and is

$$p_{\mathbf{z}}\left(\mathbf{z}\right) = \pi^{-p}\left|\Sigma_{\mathbf{z}}\right|^{-1} e^{-(\mathbf{z}-\mu_{\mathbf{z}})^*\Sigma_{\mathbf{z}}^{-1}(\mathbf{z}-\mu_{\mathbf{z}})}. \tag{17.3}$$

Reason: This starts as

$$\begin{aligned} p_{\mathbf{Z}}(\mathbf{z}) &= (2\pi)^{-\frac{2p}{2}} |\mathbf{\Sigma_X}|^{-1/2} \\ &\qquad \exp -\frac{\left(\mathbf{x}_1' - \mu_{\mathbf{x}_1}', \mathbf{x}_2' - \mu_{\mathbf{x}_2}'\right) \mathbf{\Sigma_X}^{-1} \begin{pmatrix} \mathbf{x}_1 - \mu_{\mathbf{x}_1} \\ \mathbf{x}_2 - \mu_{\mathbf{x}_2} \end{pmatrix}}{2}. \end{aligned}$$

Now

$$|\mathbf{C} - i\mathbf{Q}| = |\mathbf{\Sigma_Z}| = \left|\mathbf{\Sigma_Z'}\right| = |\mathbf{C} + i\mathbf{Q}|,$$

so

$$\begin{aligned} |\mathbf{\Sigma_Z}|^2 &= |\mathbf{C} - i\mathbf{Q}| \, |\mathbf{C} + i\mathbf{Q}| \\ &= |\mathbf{C}| \left|\mathbf{I} - i\mathbf{C}^{-1}\mathbf{Q}\right| |\mathbf{C}| \left|\mathbf{I} + i\mathbf{C}^{-1}\mathbf{Q}\right| \\ &= |\mathbf{C}|^2 \left|\mathbf{I} + \mathbf{C}^{-1}\mathbf{Q}\mathbf{C}^{-1}\mathbf{Q}\right| \\ &= \begin{pmatrix} \mathbf{C} & -\mathbf{Q} \\ \mathbf{Q} & \mathbf{C} \end{pmatrix} \\ &= |2\mathbf{\Sigma_X}|; \end{aligned}$$

thus $|\mathbf{\Sigma_Z}|^{-1} = |2\mathbf{\Sigma_X}|^{-1/2} = 2^{-\frac{2p}{2}} |\mathbf{\Sigma_X}|^{-1/2}$. Now solve

$$\mathbf{I} = \mathbf{\Sigma_Z}^{-1}\mathbf{\Sigma_Z} = (\mathbf{A} - i\mathbf{B})(\mathbf{C} - i\mathbf{Q})$$

to get

$$\begin{aligned}
\mathbf{A} &= \left(\mathbf{C}+\mathbf{Q}\mathbf{C}^{-1}\mathbf{Q}\right)^{-1}, \\
\mathbf{B} &= -\left(\mathbf{C}+\mathbf{Q}\mathbf{C}^{-1}\mathbf{Q}\right)^{-1}\mathbf{Q}\mathbf{C}^{-1}\ (=-\mathbf{B}'), \\
\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} &= \left(\mathbf{C}+\mathbf{Q}\mathbf{C}^{-1}\mathbf{Q}\right)^{-1}\left(\mathbf{I}+i\mathbf{Q}\mathbf{C}^{-1}\right).
\end{aligned}$$

These also satisfy

$$\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} = 2\begin{pmatrix} \mathbf{A} & -\mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{pmatrix},$$

and so

$$\begin{aligned}
&(\mathbf{z}-\mu_{\mathbf{z}})^{*}\,\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\,(\mathbf{z}-\mu_{\mathbf{z}}) = \ldots \\
= \ &\left(\mathbf{x}_1'-\mu_{\mathbf{x}_1}',\mathbf{x}_2'-\mu_{\mathbf{x}_2}'\right)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\begin{pmatrix}\mathbf{x}_1-\mu_{\mathbf{x}_1}\\ \mathbf{x}_2-\mu_{\mathbf{x}_2}\end{pmatrix}/2.
\end{aligned}$$

- Note the **useful identity**: if $(\mathbf{C}-i\mathbf{Q})^{-1}=(\mathbf{A}-i\mathbf{B})$ with $\mathbf{C}$ symmetric and $\mathbf{Q}$ skew-symmetric, then $\mathbf{A}$ is symmetric, $\mathbf{B}$ is skew-symmetric and

$$\begin{pmatrix} \mathbf{C} & -\mathbf{Q} \\ \mathbf{Q} & \mathbf{C} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A} & -\mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{pmatrix}.$$

18. Spectral likelihood and frequency domain regression

18.1. Spectral density matrix

- Let $\gamma(h) = E[x_{t+h}\overline{x_t}]$ be the autocovariance function of a stationary, possibly complex, zero-mean series $\{x_t\}$. This is necessarily "Hermitian non-negative definite": for any set of complex constants $\{a_t\}$ we have that

$$\sum_{s=1}^{n}\sum_{t=1}^{n}\overline{a}_s\gamma(s-t)a_t \geq 0,$$

since the lhs is $E\left[\left|\sum_{s=1}^{n}\overline{a}_s x_s\right|^2\right]$.

Theorem 18.1. *A function $\{\gamma(h)\}_{h=-\infty}^{\infty}$ is Hermitian non-negative definite iff*

$$\gamma(h) = \int_{-1/2}^{1/2} e^{2\pi i\omega h} dF(\omega), \tag{18.1}$$

where $F(\omega)/\gamma(0)$ *is a c.d.f. on* $[-1/2, 1/2]$*: non-decreasing, right continuous,* $F(-1/2)/\gamma(0) = 0$, $F(1/2)/\gamma(0) = 1$.

Proof of sufficiency: If (18.1) holds then

$$\sum_{s=1}^{n}\sum_{t=1}^{n}\overline{a}_s\gamma(s-t)a_t = \int_{-1/2}^{1/2}\left|\sum_{s=1}^{n}\overline{a}_s e^{2\pi i\omega s}\right|^2 dF(\omega) \geq 0.$$

Proof of necessity: Suppose $\{\gamma(h)\}$ is n.n.d. and define

$$f_n(\omega) = \frac{1}{n}\sum_{s=1}^{n}\sum_{t=1}^{n} e^{-2\pi i\omega s}\gamma(s-t)e^{2\pi i\omega t} \geq 0.$$

The sum is the sum of all elements of the Hermitian, Toeplitz matrix $W_{n\times n}$ with elements

$$w_{st} = e^{-2\pi i\omega s}\gamma(s-t)e^{2\pi i\omega t}$$

satisfying

$$w_{s+m,s} = e^{-2\pi i\omega m}\gamma(m) \stackrel{def}{=} r(m)$$

and so

$$\begin{aligned} f_n(\omega) &= \frac{1}{n}\left[\begin{array}{c} nr(0)+(n-1)r(1)+...+r(n-1) \\ +(n-1)r(-1)+...+r(-(n-1)) \end{array}\right] \\ &= \sum_{u=-(n-1)}^{n-1}\left(1-\frac{|u|}{n}\right)e^{-2\pi i\omega u}\gamma(u). \end{aligned}$$

The function

$$F_n(\omega)=\int f_n(\nu)\, I_{(-1/2,\omega]}d\nu$$

is, by the definition of the integral, monotonic and right continuous, with $F_n(-1/2)=0$ and

$$\begin{aligned} F_n(1/2) &= \int_{-1/2}^{1/2} f_n(\nu)\, d\nu \\ &= \int_{-1/2}^{1/2} \sum_{u=-(n-1)}^{n-1}\left(1-\frac{|u|}{n}\right)e^{-2\pi i\nu u}\gamma(u)\, d\nu \\ &= \sum_{u=-(n-1)}^{n-1}\left(1-\frac{|u|}{n}\right)\gamma(u)\int_{-1/2}^{1/2} e^{-2\pi i\nu u}d\nu \\ &= \gamma(0); \end{aligned}$$

here we use that

$$\int_{-1/2}^{1/2} e^{-2\pi i\nu u}d\nu = I(u=0).$$

Similarly

$$\int_{-1/2}^{1/2} e^{2\pi i\omega u} dF_n(\omega) = \int_{-1/2}^{1/2} e^{2\pi i\omega u} f_n(\omega)\, d\omega$$

$$= \sum_{s=-(n-1)}^{n-1} \left(1 - \frac{|s|}{n}\right) \gamma(s) \int_{-1/2}^{1/2} e^{2\pi i\omega(u-s)} d\omega$$

$$= \left(1 - \frac{|u|}{n}\right) \gamma(u)\, I(|u| < n).$$

By the *Helly selection theorem* the sequence $\{F_n\}$ of (constant multiples of) d.f.s contains a subsequence $\{F_{n_k}\}$ converging weakly to a c.d.f. F, and then by the definition of weak convergence, since the complex exponentials are bounded and continuous, we have that

$$\int_{-1/2}^{1/2} e^{2\pi i\omega u} dF_{n_k}(\omega) \to \int_{-1/2}^{1/2} e^{2\pi i\omega u} dF(\omega).$$

But the lhs is

$$\left(1 - \frac{|u|}{n_k}\right) \gamma(u)\, I(|u| < n_k),$$

which $\to \gamma(u)$; thus

$$\int_{-1/2}^{1/2} e^{2\pi i\omega u} dF(\omega) = \gamma(u),$$

as required. □

- It is more convenient to be able to work with the spectral density; for this we assume that

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty.$$

It follows that

$$f_n(\omega) = \sum_{u=-\infty}^{\infty} I(|u| < n) \left(1 - \frac{|u|}{n}\right) e^{-2\pi i \omega u} \gamma(u)$$

is absolutely convergent, hence convergent, and that the limit function is the "spectral density"

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

with

$$\lim F_n(\omega) = F(\omega) = \int f(\nu) I_{(-1/2,\omega]} d\nu.$$

- All of this can be extended to the vector case by working with linear combinations of vector-valued

series - see the text. The end result is that if $\{\mathbf{x}_t\}$ is a vector process with autocovariance function (matrix)

$$\Gamma(h) = \text{cov}\left[\mathbf{x}_{t+h}, \mathbf{x}_t\right],$$

then

$$\sum_{s=1}^{n}\sum_{t=1}^{n} \mathbf{a}_s^* \Gamma(s-t)\,\mathbf{a}_t \geq 0$$

for vectors $\{\mathbf{a}_t\}_{t=1}^n$ and a spectral representation exists, as in the theorem above.

- If $\sum_{h=-\infty}^{\infty} \left|\gamma_{jk}(h)\right| < \infty$ for all j, k, then there is a spectral density (matrix)

$$\mathbf{f}(\omega) = \sum_{h=-\infty}^{\infty} e^{-2\pi i\omega h}\Gamma(h)$$

for which

$$\Gamma(h) = \int_{-1/2}^{1/2} e^{2\pi i\omega h}\mathbf{f}(\omega)\,d\omega.$$

18.2. Spectral likelihood

- Let $\{\mathbf{x}_t\}_{t=1}^n$ be a p-dimensional time series, with mean $E[\mathbf{x}_t] = \mu_t$ (not necessarily constant) and DFT (for $\omega_k = \frac{k}{n}, 0 < |\omega_k| < \frac{1}{2}$)

$$\begin{aligned}
\mathbf{X}(\omega_k) &= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{x}_t e^{-2\pi i \omega_k} \\
&= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{x}_t \cos(2\pi\omega_k) - i \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{x}_t \sin(2\pi\omega_k) \\
&= \mathbf{X}_C(\omega_k) - i\mathbf{X}_S(\omega_k).
\end{aligned}$$

This has mean

$$\begin{aligned}
\mathbf{M}(\omega_k) &= E[\mathbf{X}(\omega_k)] \\
&= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mu_t e^{-2\pi i \omega_k} \\
&= \mathbf{M}_C(\omega_k) - i\mathbf{M}_S(\omega_k).
\end{aligned}$$

- If $\{\mathbf{x}_t\}_{t=1}^n$ is Gaussian then $\mathbf{X}_C(\omega_k)$ and $\mathbf{X}_S(\omega_k)$ are jointly Normal and (see Appendix C) $\mathbf{X}(\omega_k)$ is asymptotically complex Normal:

$$\mathbf{X}(\omega_k) \overset{d}{\approx} CN_p(\mathbf{M}(\omega_k), \mathbf{C}(\omega_k) - i\mathbf{Q}(\omega_k)),$$

where

$$\begin{aligned}
\frac{\mathbf{C}(\omega_k)}{2} &= \text{cov}[\mathbf{X}_C(\omega_k)] = \text{cov}[\mathbf{X}_S(\omega_k)], \\
\frac{\mathbf{Q}(\omega_k)}{2} &= \text{cov}[\mathbf{X}_C(\omega_k), \mathbf{X}_S(\omega_k)] \\
&= -\text{cov}[\mathbf{X}_S(\omega_k), \mathbf{X}_C(\omega_k)].
\end{aligned}$$

The density of $\mathbf{z} = \mathbf{X}(\omega_k)$ is given by (17.3) with

$$\begin{aligned}
\mathbf{\Sigma}_\mathbf{z} &= E[(\mathbf{z} - \mu_\mathbf{z})(\mathbf{z} - \mu_\mathbf{z})^*] \\
&= E[(\mathbf{X}(\omega_k) - \mathbf{M}(\omega_k))(\mathbf{X}(\omega_k) - \mathbf{M}(\omega_k))^*].
\end{aligned}$$

- If $\sum_{h=-\infty}^{\infty} |\gamma_{jk}(h)| < \infty$ for each j, k, then there is a spectral density matrix

$$\mathbf{f}(\omega_k) = \sum_{m=-\infty}^{\infty} e^{-2\pi i \omega_k m} \mathbf{\Gamma}_X(m).$$

If each $\sum_{h=-\infty}^{\infty} |h\gamma_{jk}(h)| < \infty$ we have that, as $n \to \infty$:

1. the correlation between values of the DFT at $\omega_k \neq \omega_l$ tends to 0;

2. the covariance matrix $\mathbf{\Sigma_z} \longrightarrow \mathbf{f}(\omega_k)$.

- Thus the density of $\mathbf{z}$ is

$$\begin{array}{rl} & p(\omega_k) \\ \approx & \pi^{-p} |\mathbf{f}(\omega_k)|^{-1} e^{-\left\{ \begin{array}{c} (\mathbf{X}(\omega_k) - \mathbf{M}(\omega_k))^* \mathbf{f}^{-1}(\omega_k) \\ (\mathbf{X}(\omega_k) - \mathbf{M}(\omega_k)) \end{array} \right\}} \end{array}$$

If values

$\mathbf{X}_l \stackrel{def}{=} \mathbf{X}(\omega_{k+l}), l = -(L-1)/2, ..., (L-1)/2$

are available for which the power is approximately constant, and if the means $\mathbf{M}_l$ are known, then the log-likelihood is approximately

$$= -L \ln |\mathbf{f}(\omega_k)| - \sum_{|l| < \frac{L-1}{2}} (\mathbf{X}_l - \mathbf{M}_l)^* \mathbf{f}^{-1}(\omega_k) (\mathbf{X}_l - \mathbf{M}_l)$$

and the MLE of $\mathbf{f}(\omega_k)$ based on these data is the mean-adjusted smoothed periodogram

$$\hat{\mathbf{f}}(\omega_k) = \frac{1}{L} \sum_{|l| < \frac{L-1}{2}} (\mathbf{X}_l - \mathbf{M}_l)(\mathbf{X}_l - \mathbf{M}_l)^*.$$

18.3. Frequency domain regression

- Similar to the "impulse-response" problem considered earlier, suppose that the (detrended) series y_t is related to several input series $\{x_{t1}\}, ..., \{x_{tq}\}$:

$$y_t = \sum_{r=-\infty}^{\infty} \beta_r' \mathbf{x}_{t-r} + w_t$$

where $\mathbf{x}_t = \left(x_{t1}, ..., x_{tq}\right)'$ and w_t is zero-mean stationary noise uncorrelated with $\mathbf{x}_t$. We will estimate the regression coefficient vectors β_r' by first minimizing the MSE:

$$MSE = E\left[\left\{y_t - \sum_{r=-\infty}^{\infty} \beta_r' \mathbf{x}_{t-r}\right\}^2\right].$$

Differentiating w.r.t. β_s gives

$$\begin{aligned} \mathbf{0}' &= E\left[\left\{y_t - \sum_{r=-\infty}^{\infty} \beta_r' \mathbf{x}_{t-r}\right\} \mathbf{x}_{t-s}'\right] \quad (18.2) \\ &= \gamma_{yx}'(s) - \sum_{r=-\infty}^{\infty} \beta_r' \Gamma_{xx}(s-r), \quad (18.3) \end{aligned}$$

for $s = 0, \pm 1, \pm 2, \ldots$. Here

$$\begin{aligned} \gamma_{yx}(s) &= \left(\text{cov}\left[y_{t+s}, x_{t1}\right], \ldots, \text{cov}\left[y_{t+s}, x_{tq}\right]\right)', \\ \Gamma_{xx}(s) &= \left(\text{cov}\left[x_{t+s,i}, x_{t,j}\right]\right)_{i,j=1,\ldots,q}; \end{aligned}$$

these are the vector of covariances between the output and all inputs, at lag s, and the matrix of covariances between all inputs, at lag s. Now define the joint spectral density matrix:

$$\mathbf{f}(\omega) = \begin{pmatrix} \mathbf{f}_{xx}(\omega) & \mathbf{f}_{xy}(\omega) \\ \mathbf{f}'_{yx}(\omega) & f_{yy}(\omega) \end{pmatrix} \begin{matrix} \leftarrow q \\ \leftarrow 1 \end{matrix}.$$

For instance

$$\begin{aligned} \mathbf{f}_{xy}(\omega) &= \sum_{s=-\infty}^{\infty} e^{-2\pi i \omega s} \gamma_{xy}(s), \\ \mathbf{f}_{xx}(\omega) &= \sum_{s=-\infty}^{\infty} e^{-2\pi i \omega s} \Gamma_{xx}(s). \end{aligned}$$

We frequently use the identities

$$\begin{aligned} \mathbf{f}_{yx}(\omega) &= \bar{\mathbf{f}}_{xy}(\omega), \\ \mathbf{f}^*_{xx}(\omega) &= \mathbf{f}_{xx}(\omega). \end{aligned}$$

Then

$$\begin{aligned}\gamma_{yx}(s) &= \int_{-1/2}^{1/2} e^{2\pi i\omega s}\mathbf{f}_{yx}(\omega)d\omega,\\ \Gamma_{xx}(s) &= \int_{-1/2}^{1/2} e^{2\pi i\omega s}\mathbf{f}_{xx}(\omega)d\omega.\end{aligned}$$

Substituting these into (18.3) gives

$$\begin{aligned}&\int_{-1/2}^{1/2} e^{2\pi i\omega s}\mathbf{f}'_{yx}(\omega)d\omega\\ =\ &\sum_{r=-\infty}^{\infty}\beta'_r\int_{-1/2}^{1/2} e^{2\pi i\omega(s-r)}\mathbf{f}_{xx}(\omega)d\omega\\ =\ &\int_{-1/2}^{1/2}\left[\sum_{r=-\infty}^{\infty}\beta'_r e^{-2\pi i\omega r}\right]e^{2\pi i\omega s}\mathbf{f}_{xx}(\omega)d\omega\\ =\ &\int_{-1/2}^{1/2}\mathbf{B}'(\omega)e^{2\pi i\omega s}\mathbf{f}_{xx}(\omega)d\omega,\end{aligned}$$

where

$$\mathbf{B}(\omega)_{q\times 1}=\sum_{r=-\infty}^{\infty}\beta_r e^{-2\pi i\omega r}$$

is the IFT of $\{\beta_r\}$. By uniqueness of Fourier transforms,

$$\mathbf{B}'(\omega)\mathbf{f}_{xx}(\omega)=\mathbf{f}'_{yx}(\omega)=\mathbf{f}^*_{xy}(\omega),$$

hence

$$\mathbf{B}'(\omega) = \mathbf{f}_{xy}^*(\omega)\mathbf{f}_{xx}^{-1}(\omega).$$

The coefficients $\{\beta_r\}$ form what is called the "impulse response function", and $\mathbf{B}(\omega)$ is the "frequency response function".

- To approximate

$$\beta_r = \int_{-1/2}^{1/2} e^{2\pi i\omega r}\mathbf{B}(\omega)d\omega = \int_0^1 e^{2\pi i\omega r}\mathbf{B}(\omega)d\omega,$$

S&S suggest discretizing the integrand and computing

$$\begin{aligned}\beta_r^M &= \frac{1}{M}\sum_{k=0}^{M-1}\mathbf{B}(\nu_k)e^{2\pi i\nu_k r} \\ &= \left\{\frac{1}{M}\sum_{k=0}^{M-1}\mathbf{f}_{xy}^*(\nu_k)\mathbf{f}_{xx}^{-1}(\nu_k)e^{2\pi i\nu_k r}\right\}',\end{aligned}$$

where $\nu_k = k/M$ and M is even and $\ll n'$. This is done for the M values

$$r = -M/2+1, ..., -1, 0, 1, ...M/2-1.$$

Similar to an exercise in Asst. 1,

$$\beta_r^M = \beta_r + \sum_{|l|>0} \beta_{r+lM},$$

and so the coefficients are recovered exactly if M is large enough that $\beta_s = \mathbf{0}$ for $|s| \geq M/2$. I use instead

$$\hat{\beta}_s^M = \mathrm{Re}\left\{ \frac{1}{M/2} \sum_{k=1}^{M/2} \hat{\mathbf{B}}(\omega_k) e^{2\pi i \omega_k s} \right\}.$$

This is easier to compute than the authors' suggestion, since R doesn't return the spectra at $\omega = 0$. The difference is negligible: this minus the suggestion of S&S $= \frac{\hat{B}(1/2)(-1)^s - \hat{B}(0)}{M}$ in each component.

- In practice $\mathbf{f}$ is estimated (`spec.pgram`):

$$\hat{\mathbf{f}}(\nu_k) = \begin{pmatrix} \hat{\mathbf{f}}_{xx}(\nu_k) & \hat{\mathbf{f}}_{xy}(\nu_k) \\ \hat{\mathbf{f}}'_{yx}(\nu_k) = \hat{\mathbf{f}}^*_{xy}(\nu_k) & \hat{\mathbf{f}}_{yy}(\nu_k) \end{pmatrix}.$$

19. Regression for jointly stationary series

- 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 in the frequency domain. The minimum MSE, playing the role of the unexplained variation (of Y_t around its best predictor $\sum_{r=-\infty}^{\infty} \beta_r' \mathbf{x}_{t-r}$) is $MSE =$

$$E\left[\left\{y_t - \sum_{r=-\infty}^{\infty} \beta_r' \mathbf{x}_{t-r}\right\}\left\{y_t - \sum_{s=-\infty}^{\infty} \mathbf{x}_{t-s}' \beta_s\right\}\right]$$

$$= E\left[\left\{y_t - \sum_{r=-\infty}^{\infty} \beta_r' \mathbf{x}_{t-r}\right\} y_t\right],$$

using (18.2). Recall that

$$\mathbf{B}(\omega)_{q\times 1} = \sum_{r=-\infty}^{\infty} \beta_r e^{-2\pi i \omega r}$$

is the IFT of $\{\beta_r\}$. Then (assuming that with this best predictor the residuals have mean zero) the calculation continues as $MSE =$

$$\begin{aligned}
& \gamma_y(0) - \sum_{r=-\infty}^{\infty} \beta_r' \boldsymbol{\gamma}_{yx}(r) \\
= \; & \gamma_y(0) - \sum_{r=-\infty}^{\infty} \int_{-1/2}^{1/2} e^{2\pi i \omega r} \mathbf{B}'(\omega) d\omega \cdot \boldsymbol{\gamma}_{yx}(r) \\
= \; & \gamma_y(0) - \int_{-1/2}^{1/2} \mathbf{B}'(\omega) \sum_{r=-\infty}^{\infty} e^{2\pi i \omega r} \boldsymbol{\gamma}_{yx}(r) d\omega \\
= \; & \gamma_y(0) - \int_{-1/2}^{1/2} \mathbf{B}'(\omega) \overline{\mathbf{f}}_{yx}(\omega) d\omega, \\
= \; & \int_{-1/2}^{1/2} f_{yy}(\omega) d\omega - \int_{-1/2}^{1/2} \mathbf{f}_{xy}^*(\omega) \mathbf{f}_{xx}^{-1}(\omega) \mathbf{f}_{xy}(\omega) d\omega \\
= \; & \int_{-1/2}^{1/2} f_{yy}(\omega) \left[1 - \rho_{y \cdot x}^2(\omega)\right] d\omega,
\end{aligned}$$

where

$$\rho_{y \cdot x}^2(\omega) = \frac{\mathbf{f}_{xy}^*(\omega) \mathbf{f}_{xx}^{-1}(\omega) \mathbf{f}_{xy}(\omega)}{f_{yy}(\omega)}$$

is the "squared multiple coherence". The reduction in MSE over merely using μ_y to forecast y_t is $\int_{-1/2}^{1/2} f_{yy}(\omega) \rho_{y \cdot x}^2(\omega) d\omega$, so that the method is

most effective when applied to strongly coherent series.

- The above can also be written as

$$MSE = \int_{-1/2}^{1/2} f_{y\cdot x}(\omega) d\omega,$$

where

$$f_{y\cdot x}(\omega) = f_{yy}(\omega) - \mathbf{f}_{xy}^{*}(\omega)\mathbf{f}_{xx}^{-1}(\omega)\mathbf{f}_{xy}(\omega)$$

denotes the "error spectrum", or "residual spectrum".

- Suppose that one is interested in testing the hypothesis that $\mathbf{B}(\omega) = \mathbf{0}$ at a particular frequency $\omega = \omega_k = k/n$. Equivalently, $\rho_{y\cdot x}^2(\omega) = 0$. For this, first note that if

$$y_t = \sum_{r=-\infty}^{\infty} \beta_r' \mathbf{x}_{t-r} + w_t$$

is substituted into

$$f_y\left(\omega + \frac{l}{n}\right) = Y\left(\omega + \frac{l}{n}\right) = \sum_{t=-\infty}^{\infty} y_t e^{-2\pi i\left(\omega + \frac{l}{n}\right)t},$$

the end result is that

$$Y\left(\omega+\frac{l}{n}\right)=\mathbf{B}'(\omega+\frac{l}{n})\mathbf{X}\left(\omega+\frac{l}{n}\right)+V\left(\omega+\frac{l}{n}\right),$$

with

$$f_w\left(\omega+\frac{l}{n}\right)=V\left(\omega+\frac{l}{n}\right)=\sum_{t=-\infty}^{\infty}w_t e^{-2\pi i\left(\omega+\frac{l}{n}\right)t}.$$

It is important to note here that $f_w(\omega)$ depends on the model - it will change depending on how well $\mathbf{X}(\omega)$ 'explains' the power in the Y-series. If there is no explanatory power (so that the hypothesis is true) then $f_w(\omega)=f_y(\omega)$; if however the hypothesis is false and $\mathbf{B}(\omega)$ is chosen in an optimal manner then $f_w(\omega)$ can be as small as $f_{y.x}(\omega)$.

- If $\mathbf{B}(\omega + \frac{l}{n}) = \mathbf{b}(\omega)_{q\times 1}$ is treated as constant (reasonable for $|l| \leq m = (L-1)/2$, small) and

$$\mathbf{y}_{L\times 1} = \begin{pmatrix} \vdots \\ Y\left(\omega + \frac{l}{n}\right) \\ \vdots \end{pmatrix}_{l=-m,\ldots,m},$$

$$\mathbf{Z}_{L\times q} = \begin{pmatrix} \vdots \\ \mathbf{X}'\left(\omega + \frac{l}{n}\right) \\ \vdots \end{pmatrix}_{l=-m,\ldots,m},$$

$$\mathbf{v}_{L\times 1} = \begin{pmatrix} \vdots \\ V\left(\omega + \frac{l}{n}\right) \\ \vdots \end{pmatrix}_{l=-m,\ldots,m},$$

then we can consider a regression model

$$\mathbf{y} = \mathbf{Z}\mathbf{b}(\omega) + \mathbf{v}. \tag{19.1}$$

- Now assume model (19.1), but with Y, $\mathbf{X}$ and V denoting the DFTs based on the sampled data. By the theory of the previous lecture the joint distribution of the cosine and sine transforms

$\left(V_C\left(\omega+\frac{l}{n}\right), V_S\left(\omega+\frac{l}{n}\right)\right)$ tends to that of L independent bivariate normal r.vecs. with covariance structure

$$\mathbf{\Sigma}(\omega_k) = \frac{1}{2}\begin{pmatrix} c(\omega_k) & -q(\omega_k) \\ q(\omega_k) & c(\omega_k) \end{pmatrix}.$$

Furthermore

$$c(\omega_k) = E\left[|V(\omega_k)|^2\right] = f_w(\omega_k) \stackrel{def}{=} \sigma_v^2$$

and (verified below) $q(\omega_k) = 0$. Thus $\mathbf{\Sigma}(\omega_k) = diag(\sigma_v^2/2, \sigma_v^2/2)$. The L values $V_C\left(\omega+\frac{l}{n}\right)$ (denoted $\mathbf{v}_c : L\times 1$) and the L values $V_S\left(\omega+\frac{l}{n}\right)$ (denoted $\mathbf{v}_S : L\times 1$) are asymptotically i.i.d. Normal:

$$\begin{pmatrix}\mathbf{v}_c \\ \mathbf{v}_s\end{pmatrix} \stackrel{d}{\approx} N_{2L}\left(\mathbf{0}, \frac{\sigma_v^2}{2}\mathbf{I}_{2L}\right).$$

This will allow us to apply standard Normal theory procedures, e.g. compare estimates of σ_v^2 under competing models. Note that, at the moment, we are holding ω fixed.

- Verification that $q(\omega_k) = 0$:

$$\begin{aligned} q(\omega_k) &= E[V_C(\omega_k) V_S(\omega_k)] \\ &= E\left[\begin{array}{c} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} w_t \cos(2\pi\omega_k t) \cdot \\ \frac{1}{\sqrt{n}} \sum_{s=1}^{n} w_s \sin(2\pi\omega_k s) \end{array} \right] \\ &= \frac{1}{n} \sum_{s,t=1}^{n} E[w_t w_s] \cos(2\pi\omega_k t) \sin(2\pi\omega_k s) \\ &= \frac{\sigma_w^2}{n} \sum_{t=1}^{n} \cos(2\pi\omega_k t) \sin(2\pi\omega_k t) \\ &= 0 \text{ (how?).} \end{aligned}$$

- Write (19.1) as

$$\mathbf{y}_c - i\mathbf{y}_s = \mathbf{y} = (\mathbf{Z}_c - i\mathbf{Z}_s)(\mathbf{b}_c - i\mathbf{b}_s) + (\mathbf{v}_c - i\mathbf{v}_s)$$

where, e.g.

$$\mathbf{Z}_c = \frac{1}{\sqrt{n}} \begin{pmatrix} \vdots \\ \sum_{t=1}^{n} \mathbf{x}'_t \cos\left(2\pi\left(\omega + \frac{l}{n}\right)t\right) \\ \vdots \end{pmatrix} \stackrel{def}{=} \mathbf{X}'_c;$$

similarly $\mathbf{Z}_s \stackrel{def}{=} \mathbf{X}'_s$.

Expanding this gives

$$\begin{aligned}\begin{pmatrix}\mathbf{y}_c\\ \mathbf{y}_s\end{pmatrix} &= \begin{pmatrix}\mathbf{X}_c' & -\mathbf{X}_s'\\ \mathbf{X}_s' & \mathbf{X}_c'\end{pmatrix}\begin{pmatrix}\mathbf{b}_c\\ \mathbf{b}_s\end{pmatrix} + \begin{pmatrix}\mathbf{v}_c\\ \mathbf{v}_s\end{pmatrix}\\ &\overset{d}{\approx} N_{2L}\left(\begin{pmatrix}\mathbf{X}_c' & -\mathbf{X}_s'\\ \mathbf{X}_s' & \mathbf{X}_c'\end{pmatrix}\begin{pmatrix}\mathbf{b}_c\\ \mathbf{b}_s\end{pmatrix}, \frac{\sigma_v^2}{2}\mathbf{I}_{2L}\right).\end{aligned}$$

One can now go on to consider the usual F-test of a regression relationship. For this, write

$$\mathbf{u} = \begin{pmatrix}\mathbf{y}_c\\ \mathbf{y}_s\end{pmatrix}, \mathbf{P} = \begin{pmatrix}\mathbf{X}_c' & -\mathbf{X}_s'\\ \mathbf{X}_s' & \mathbf{X}_c'\end{pmatrix}, \boldsymbol{\theta} = \begin{pmatrix}\mathbf{b}_c\\ \mathbf{b}_s\end{pmatrix},$$

so that

$$\mathbf{u} \overset{d}{\approx} N_{2L}\left(\mathbf{P}\boldsymbol{\theta}, \frac{\sigma_v^2}{2}\mathbf{I}_{2L}\right).$$

The sum of squares function is

$$SS(\boldsymbol{\theta}) = \|\mathbf{u} - \mathbf{P}\boldsymbol{\theta}\|^2,$$

and the test of the hypothesis that $\mathbf{B}(\omega_k) = \mathbf{0}$ rejects for large values of

$$F = \frac{MS_{drop}}{MSE} = \frac{\left(SS(\mathbf{0}) - SS\left(\hat{\boldsymbol{\theta}}\right)\right)/(2q)}{SS\left(\hat{\boldsymbol{\theta}}\right)/(2L-2q)} \overset{d}{\approx} F^{2q}_{2(L-q)}.$$

We have

$$SS\left(\mathbf{0}\right) = \left\|\mathbf{u}\right\|^2 = \mathbf{y}_c'\mathbf{y}_c + \mathbf{y}_s'\mathbf{y}_s = \mathbf{y}^*\mathbf{y} = L\hat{f}_{yy}(\omega),$$

and

$$\begin{aligned} SS\left(\hat{\boldsymbol{\theta}}\right) &= SS\left(\left(\mathbf{P}'\mathbf{P}\right)^{-1}\mathbf{P}'\mathbf{u}\right) \\ &= \mathbf{u}'(\mathbf{I}_{2L} - \mathbf{P}\left(\mathbf{P}'\mathbf{P}\right)^{-1}\mathbf{P}')\mathbf{u} \\ &= L\hat{f}_{yy}(\omega) - \mathbf{u}'\mathbf{P}\left(\mathbf{P}'\mathbf{P}\right)^{-1}\mathbf{P}'\mathbf{u}. \end{aligned}$$

It can be shown (assigned) that

$$\mathbf{u}'\mathbf{P}\left(\mathbf{P}'\mathbf{P}\right)^{-1}\mathbf{P}'\mathbf{u} = L\left(\hat{\mathbf{f}}_{xy}^*(\omega)\hat{\mathbf{f}}_{xx}^{-1}(\omega)\hat{\mathbf{f}}_{xy}(\omega)\right);$$

thus

$$\begin{aligned} SS\left(\hat{\boldsymbol{\theta}}\right) &= L\hat{f}_{y\cdot x}(\omega) \\ &= L\hat{f}_{yy}(\omega)\left(1 - \hat{\rho}_{y\cdot x}^2(\omega)\right), \\ SS\left(\mathbf{0}\right) - SS\left(\hat{\boldsymbol{\theta}}\right) &= L\hat{f}_{yy}(\omega)\hat{\rho}_{y\cdot x}^2(\omega) \\ &= L\left(\hat{f}_{yy}(\omega) - \hat{f}_{y\cdot x}(\omega)\right), \end{aligned}$$

and so

$$MS_{drop} = \frac{L\hat{f}_{yy}(\omega)\hat{\rho}^2_{y\cdot x}(\omega)}{2q}, \quad (19.2)$$

$$MSE = \frac{L\hat{f}_{yy}(\omega)\left(1-\hat{\rho}^2_{y\cdot x}(\omega)\right)}{2\left(L-q\right)}, \quad (19.3)$$

$$F = \frac{L-q}{q}\frac{\hat{\rho}^2_{y\cdot x}(\omega)}{1-\hat{\rho}^2_{y\cdot x}(\omega)}. \quad (19.4)$$

- Similarly, one can write

$$\beta_r = \begin{matrix}\beta_{1,r} \\ \beta_{2,r}\end{matrix} \begin{matrix}\leftarrow q_1 \\ \leftarrow q_2\end{matrix}$$

so that the model is

$$y_t = \sum_{r=-\infty}^{\infty} \beta'_{1,r}\mathbf{x}_{t-r,1} + \sum_{r=-\infty}^{\infty} \beta'_{2,r}\mathbf{x}_{t-r,2} + w_t.$$

Now consider the hypothesis that $\beta_{2,r} = \mathbf{0}$ for all r, at frequency ω. For this, partition $\hat{\mathbf{f}}_{xy}(\omega)$ as

$$\hat{\mathbf{f}}_{xy}(\omega) = \begin{matrix}\hat{\mathbf{f}}_{1y}(\omega) \\ \hat{\mathbf{f}}_{2y}(\omega)\end{matrix} \begin{matrix}\leftarrow q_1 \\ \leftarrow q_2\end{matrix},$$

and partition the spectral input matrix as

$$\hat{\mathbf{f}}_{xx}(\omega) = \begin{pmatrix} \hat{\mathbf{f}}_{11}(\omega) & \hat{\mathbf{f}}_{12}(\omega) \\ \hat{\mathbf{f}}_{12}^{*}(\omega) & \hat{\mathbf{f}}_{22}(\omega) \end{pmatrix}.$$

The development is analogous to that leading to (19.2) - (19.4). The residual power under the full model is estimated as

$$\hat{f}_{y\cdot x}(\omega) = \hat{f}_{yy}(\omega) - \hat{\mathbf{f}}_{xy}^{*}(\omega)\hat{\mathbf{f}}_{xx}^{-1}(\omega)\hat{\mathbf{f}}_{xy}(\omega);$$

under the reduced model it is

$$\hat{f}_{y\cdot 1}(\omega) = \hat{f}_{yy}(\omega) - \hat{\mathbf{f}}_{1y}^{*}(\omega)\hat{\mathbf{f}}_{11}^{-1}(\omega)\hat{\mathbf{f}}_{1y}(\omega).$$

The F-test is then based on

$$F = \frac{MS_{drop}}{MSE} \overset{d}{\approx} F_{2(L-q)}^{2q_2},$$

where

$$\begin{aligned} MS_{drop} &= \frac{L\left(\hat{f}_{y\cdot 1}(\omega) - \hat{f}_{y\cdot x}(\omega)\right)}{2q_2}, \\ MSE &= \frac{L\hat{f}_{y\cdot x}(\omega)}{2(L-q)}. \end{aligned}$$

- These F statistics can be computed by obtaining the residual power in each model - with and without those predictor series which are dropped when β_2 is dropped - and comparing them.

- The elements of the vector $\hat{\mathbf{f}}_{xy}(\omega)$ and the matrix $\hat{\mathbf{f}}_{xx}$ are all either spectra of univariate series (on the diagonal of $\hat{\mathbf{f}}_{xx}$), or cross-spectra of two series, determined in R via the relationship

$$f_{yx}(\nu) = \sqrt{\rho^2_{yx}(\nu) f_y(\nu) f_x(\nu)} e^{i\phi_{yx}(\nu)},$$

applied to each pair of variables. The coherences $\rho^2_{yx}(\nu)$, phases $\phi_{yx}(\nu)$ and marginal spectra $f_y(\nu)$ and $f_x(\nu)$ are in turn returned by `spec.pgram(cbind(y,x), kernel("daniell",m))` See my function `stoch.regr` on the course website.

- Example 7.1.

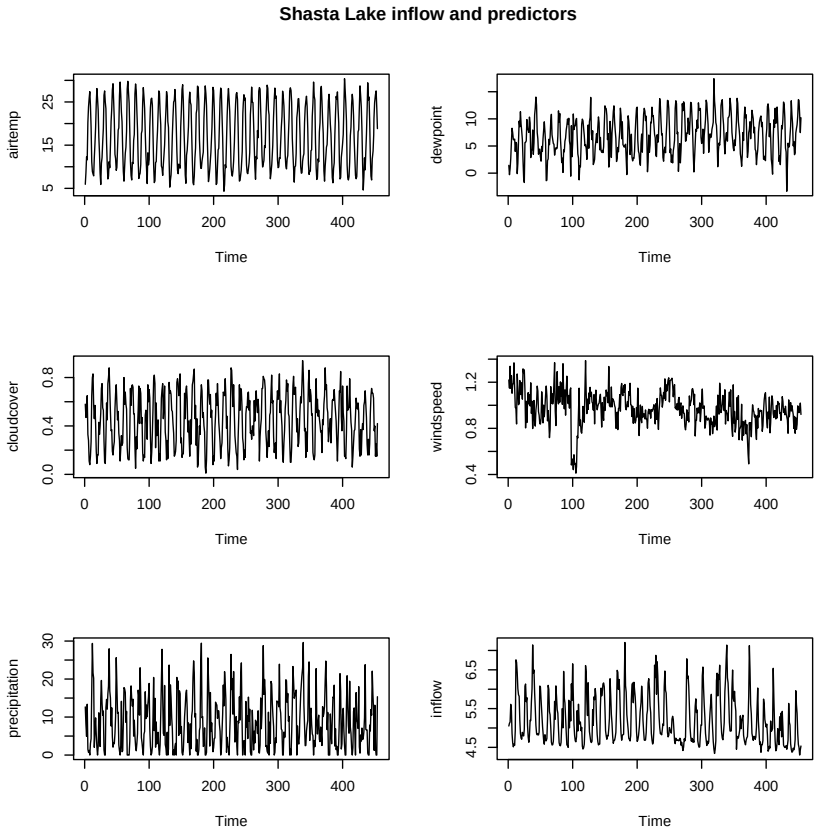

Figure 19.1. Shasta Lake data - relate log(inflow) to sqrt(precipitation), air temperature, dewpoint, cloudcover and windspeed.

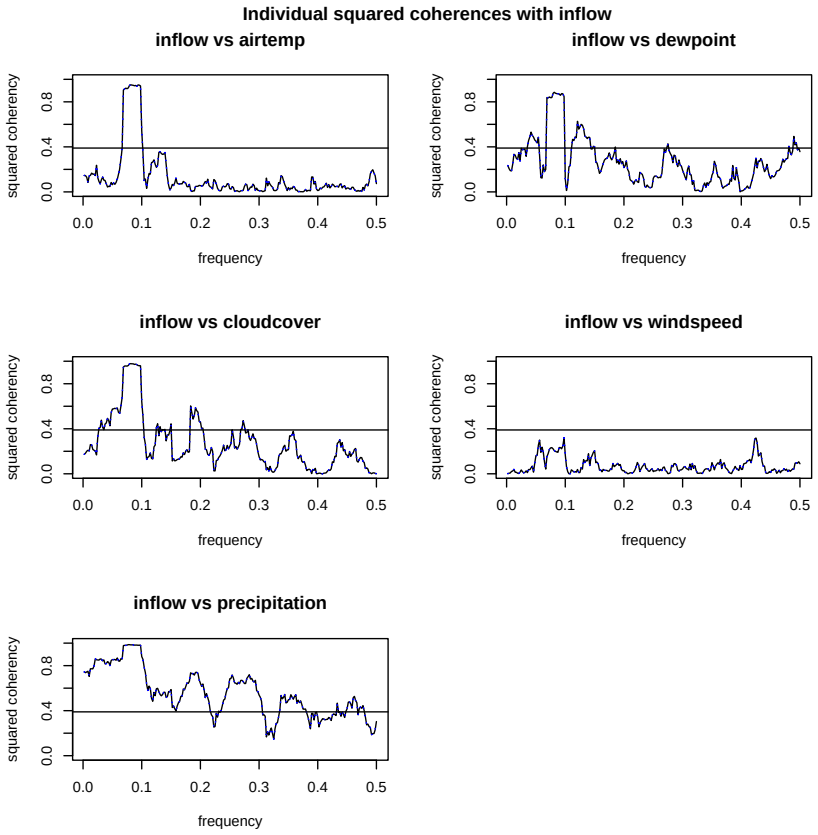

Figure 19.2. Individual (squared) coherences with Inflow. Best single predictor is Precipitation. Horizontal lines at $\alpha = .001$ critical value.

These individual coherences are produced along with the output from `spec.pgram`, when the input is a matrix.

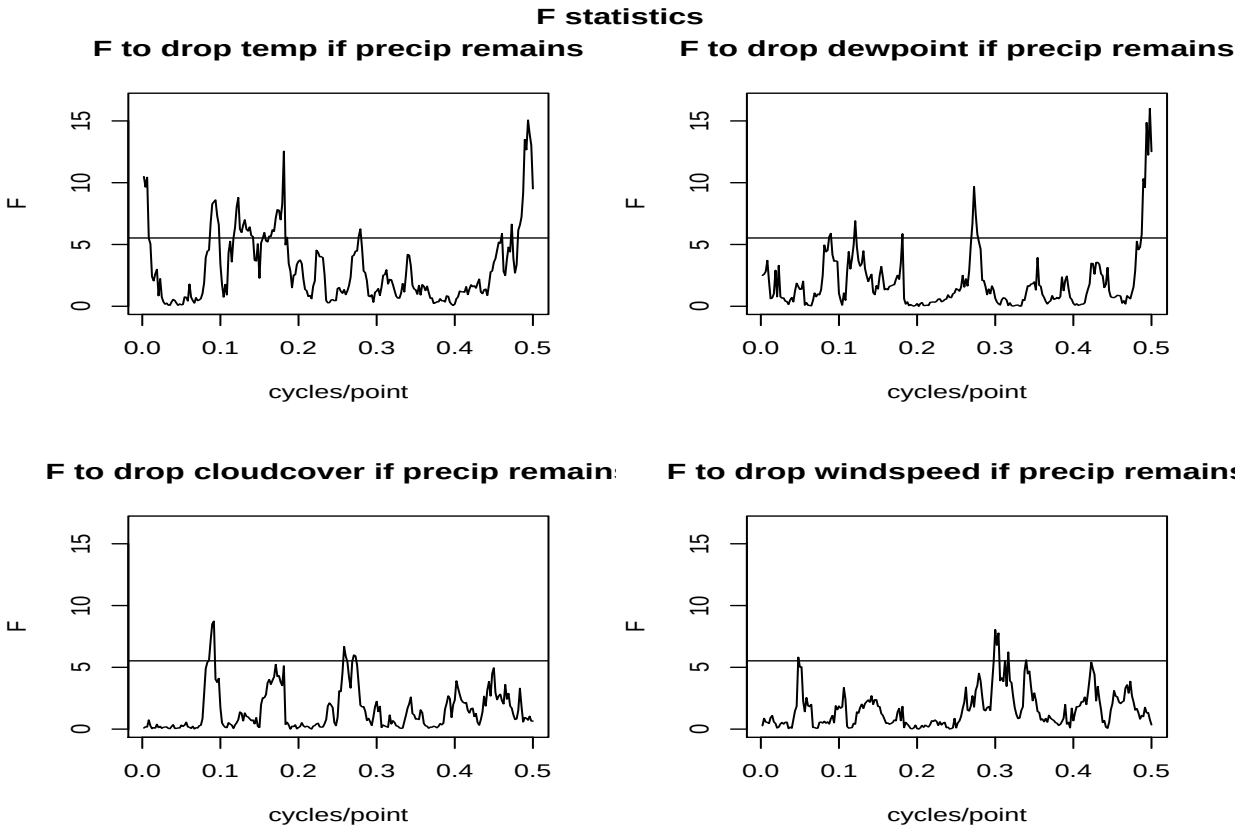

Figure 19.3. F statistics comparing models with one predictor besides Precipitation, with the reduced model containing Precipitation alone. None seem to add much, but Temp is better than the others.

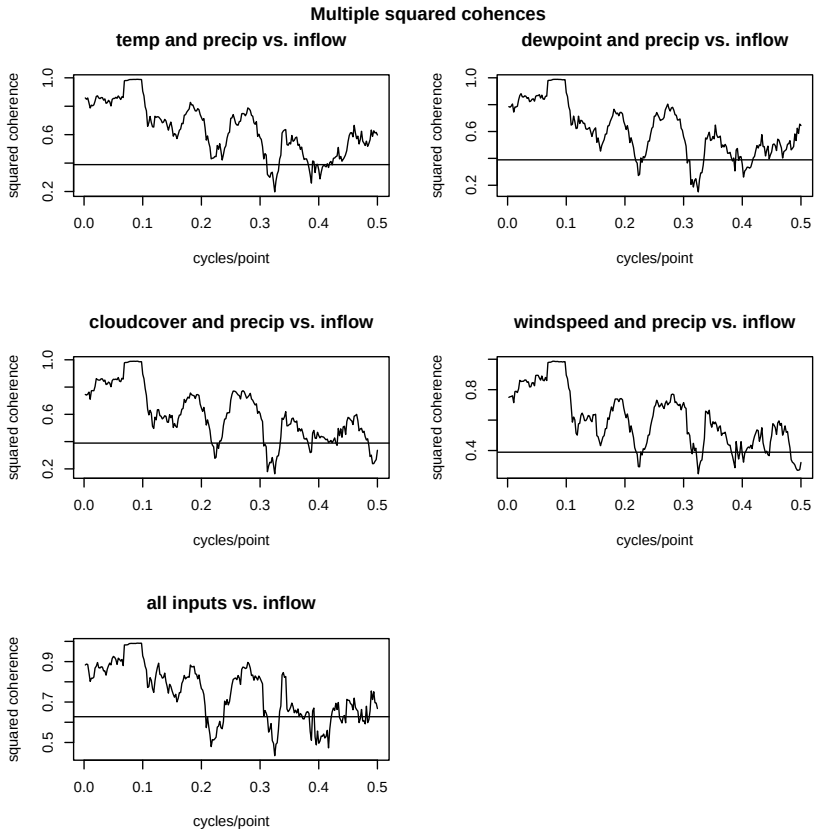

Figure 19.4. Multiple (squared) coherences for all models with Precipitation and one other predictor. That with temperature doesn't seem any better than that for Precipitation alone.

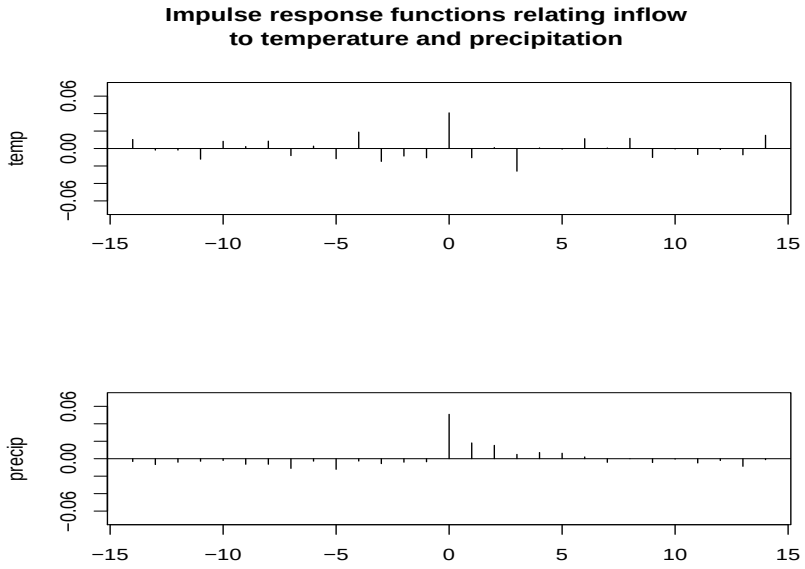

Figure 19.5. Multiple impulse reponse functions relating Inflow to Precipitation and Temperature.

20. Regression with deterministic inputs

- Here the input series are non-random, but the experiments are replicated:

$$y_{jt} = \sum_{r=-\infty}^{\infty} \beta_r' \mathbf{z}_{j,t-r} + v_{jt},$$

for $j = 1, ..., N$. The inputs $\left\{\mathbf{z}_{j,t-r}\right\}$ are deterministic; each $\mathbf{z}_{j,t-r}$ is a $q \times 1$ vector of regressors. So the same model is observed N times, although the values of the regressors might change each time. In matrix form,

$$\begin{pmatrix} y_{1t} \\ \vdots \\ y_{Nt} \end{pmatrix} = \begin{pmatrix} \mathbf{z}_{1,t-r}' \\ \vdots \\ \mathbf{z}_{N,t-r}' \end{pmatrix} \beta_r + \begin{pmatrix} v_{1t} \\ \vdots \\ v_{Nt} \end{pmatrix}, \text{ i.e.}$$

$$\mathbf{y}_t = \sum_{r=-\infty}^{\infty} \mathbf{z}_{t-r} \beta_r + \mathbf{v}_t.$$

Thus $\mathbf{z}_t$ is an $N \times q$ matrix with rows $\left\{\mathbf{z}_{j,t}'\right\}_{j=1}^{N}$. For fixed t the v_{jt} are i.i.d., with mean zero and variance σ_v^2. It is also assumed that $\{v_{1t}\}_{t=-\infty}^{\infty}, ...,$ $\{v_{Nt}\}_{t=-\infty}^{\infty}$ are i.i.d. series, each with spectral

density $\sigma_v^2 = f_v(\omega)$. Thus the spectral matrix of $\{\mathbf{v}_t\}_{t=-\infty}^{\infty}$ is

$$\mathbf{f}(\omega) = \sum_{h=-\infty}^{\infty} e^{-2\pi i\omega h}\Gamma(h) = \Gamma(0) = f_v(\omega)\mathbf{I}_N.$$

- Example 7.2. Signals from an explosion arrive at one of $N = 3$ sensors. At time t the signal sent is β_t, but because of certain (predictable) delays, the arrival times are slightly different. If y_{jt} is the signal received at sensor j at time t, then the model is

$$y_{jt} = \beta_{t-\tau_j} + v_{jt} = \sum_{r=-\infty}^{\infty} \beta_r z_{j,t-r} + v_{jt},$$

for $z_{j,t} = I\left(t = \tau_j\right)$, where τ_j is the time delay. We will show that the "Best Linear Unbiased Estimate" (BLUE) is

$$\hat{\beta}_t = \frac{1}{N}\sum_{j=1}^{N} y_{j,t+\tau_j}.$$

- **Estimation of the regression relationship**. A BLUE is an estimate of the form

$$\hat{\beta}_t = \sum_{r=-\infty}^{\infty} H_r \mathbf{y}_{t-r},$$

where H_r is a $q \times N$ matrix of filter coefficients. Within this class of linear filters the estimate is to be minimum variance unbiased, in that for any choice $\{\mathbf{a}_r\}_{r=-\infty}^{\infty}$, the scalar estimate

$$\hat{\psi}_t = \sum_{r=-\infty}^{\infty} \mathbf{a}_r' \hat{\beta}_{t-r}$$

is to be unbiased and have minimum variance.

Unbiasedness requires

$$\begin{aligned}
\sum_{r=-\infty}^{\infty} \mathbf{a}_r' \beta_{t-r} &= E\left[\hat{\psi}_t\right] \\
&= \sum_{r=-\infty}^{\infty} \mathbf{a}_r' E\left[\sum_{s=-\infty}^{\infty} H_s \mathbf{y}_{t-r-s}\right] \\
&= \sum_{r,s,u=-\infty}^{\infty} \mathbf{a}_r' H_s \mathbf{z}_{t-r-s-u} \beta_u.
\end{aligned}$$

Let $A(\omega)_{q\times 1}, B(\omega)_{q\times 1}, \mathbf{H}(\omega)_{q\times N}$ and $\mathbf{Z}(\omega)_{N\times q}$ be the IFTs; then calculating the IFTs of each side above yields

$$A'(\omega) B(\omega) = A'(\omega) \mathbf{H}(\omega) \mathbf{Z}(\omega) B(\omega)$$

for all ω. Equivalently,

$$\mathbf{H}(\omega) \mathbf{Z}(\omega) = \mathbf{I}_q. \tag{20.1}$$

A particular solution to (20.1) is

$$\mathbf{H}(\omega) = \left(\mathbf{Z}^*(\omega) \mathbf{Z}(\omega)\right)^{-1} \mathbf{Z}^*(\omega). \tag{20.2}$$

(What are the others?) We now claim that if $\hat{\psi}_t$ is the linear filter corresponding to (20.2), it is the unique BLUE. To see this, suppose that

$$\tilde{\psi}_t = \sum_{r=-\infty}^{\infty} \mathbf{a}_r' \tilde{\beta}_{t-r} = \sum_{r=-\infty}^{\infty} \mathbf{a}_r' \sum_{s=-\infty}^{\infty} G_s \mathbf{y}_{t-r-s}$$

is any other unbiased estimate (necessarily satisfying (20.1): $\mathbf{G}(\omega) \mathbf{Z}(\omega) = \mathbf{I}_q$). We will show that

$$\text{cov}\left[\left(\tilde{\psi}_t - \hat{\psi}_t\right), \hat{\psi}_t\right] = 0, \tag{20.3}$$

so that

$$\begin{aligned} \text{var}\left[\tilde{\psi}_t\right] &= \text{var}\left[\left(\tilde{\psi}_t - \hat{\psi}_t\right) + \hat{\psi}_t\right] \\ &= \text{var}\left[\tilde{\psi}_t - \hat{\psi}_t\right] + \text{var}\left[\hat{\psi}_t\right] \\ &\geq \text{var}\left[\hat{\psi}_t\right]. \end{aligned}$$

This last inequality is strict unless $\text{var}\left[\tilde{\psi}_t - \hat{\psi}_t\right] = 0$; since $E\left[\tilde{\psi}_t - \hat{\psi}_t\right] = 0$ this requires

$$\tilde{\psi}_t = \hat{\psi}_t \text{ a.e.}$$

To verify (20.3) we calculate

$$\begin{aligned} & \text{cov}\left[\tilde{\psi}_t - \hat{\psi}_t, \hat{\psi}_t\right] \\ = & \text{cov}\left[\begin{array}{c} \sum_{r=-\infty}^{\infty} \mathbf{a}_r' \sum_{s=-\infty}^{\infty} \left(G_s - H_s\right) \mathbf{y}_{t-r-s}, \\ \sum_{u=-\infty}^{\infty} \mathbf{a}_u' \sum_{w=-\infty}^{\infty} H_w \mathbf{y}_{t-u-w} \end{array}\right] \\ = & \sum_{r,s,u,w} \mathbf{a}_r' \left(G_s - H_s\right) \text{cov}\left[\mathbf{y}_{t-r-s}, \mathbf{y}_{t-u-w}\right] H_w' \mathbf{a}_u \\ = & \sigma_v^2 \sum_{r,s,u=-\infty}^{\infty} \mathbf{a}_r' \left(G_s - H_s\right) H_{r+s-u}' \mathbf{a}_u. \end{aligned}$$

Substitute

$$\begin{aligned} H'_{r+s-u} &= \int_{-1/2}^{1/2} \mathbf{H}'(\omega)\, e^{2\pi i\omega(r+s-u)} d\omega \\ &= \int_{-1/2}^{1/2} \mathbf{H}^*(\omega)\, e^{-2\pi i\omega(r+s-u)} d\omega \\ &\quad \text{(since } H \text{ is real)}, \end{aligned}$$

obtaining

$$\begin{aligned} &\text{cov}\left[\tilde{\psi}_t - \hat{\psi}_t, \hat{\psi}_t\right] \\ &= \sigma_v^2 \sum_{r,s,u=-\infty}^{\infty} \left\{ \begin{array}{c} \mathbf{a}'_r (G_s - H_s) \\ \int_{-1/2}^{1/2} \mathbf{H}^*(\omega)\, e^{-2\pi i\omega(r+s-u)} d\omega \mathbf{a}_u \end{array} \right\} \\ &= \sigma_v^2 \int_{-1/2}^{1/2} \left\{ \begin{array}{c} \sum_r \mathbf{a}'_r e^{-2\pi i\omega r} \sum_s (G_s - H_s)\, e^{-2\pi i\omega s} \\ \mathbf{H}^*(\omega) \sum_u \mathbf{a}_u e^{2\pi i\omega u} d\omega \end{array} \right\} \\ &= \sigma_v^2 \int_{-1/2}^{1/2} A'(\omega) \left(\mathbf{G}(\omega) - \mathbf{H}(\omega)\right) \mathbf{H}^*(\omega) A(-\omega)\, d\omega \\ &= 0, \end{aligned}$$

since by (20.2) and (20.1),

$$\begin{aligned} &\left(\mathbf{G}(\omega) - \mathbf{H}(\omega)\right) \mathbf{H}^*(\omega) \\ &= \left(\mathbf{G}(\omega) - \mathbf{H}(\omega)\right) \mathbf{Z}(\omega) \left(\mathbf{Z}^*(\omega)\, \mathbf{Z}(\omega)\right)^{-1} \\ &= \mathbf{0}. \end{aligned}$$

Summary: a BLUE is given by

$$\begin{aligned}\hat{\beta}_t &= \sum_{r=-\infty}^{\infty} H_r \mathbf{y}_{t-r} \text{ and} \\ H_r &= \int_{-1/2}^{1/2} \mathbf{H}(\omega)\, e^{2\pi i \omega r} d\omega,\end{aligned}$$

where

$$\begin{aligned}\mathbf{H}(\omega) &= \mathbf{S}_z^{-1}(\omega)\, \mathbf{Z}^*(\omega) : q \times N, \\ \mathbf{S}_z &= \mathbf{Z}^*(\omega)\, \mathbf{Z}(\omega) : q \times q, \\ \mathbf{Z}(\omega) &= \sum_{t=-\infty}^{\infty} \mathbf{z}_t e^{-2\pi i \omega t} : N \times q \\ &= \begin{pmatrix} \vdots \\ \mathbf{Z}'_j(\omega) = \sum_{t=-\infty}^{\infty} \mathbf{z}'_{t,j} e^{-2\pi i \omega t} \\ \vdots \end{pmatrix}.\end{aligned}$$

- Example 7.3 (Example 7.2 continued). We have $q = 1$ and

$$\begin{aligned}
\mathbf{Z}_j(\omega) &= \sum_{t=-\infty}^{\infty} z_{j,t} e^{-2\pi i\omega t} \\
&= \sum_{t=-\infty}^{\infty} I\left(t = \tau_j\right) e^{-2\pi i\omega t} \\
&= e^{-2\pi i\omega\tau_j},
\end{aligned}$$

so that

$$\begin{aligned}
\mathbf{S}_z &= \sum_{j=1}^{N} \overline{\mathbf{Z}_j(\omega)}\mathbf{Z}_j(\omega) = N, \\
\mathbf{H}(\omega) &= \frac{1}{N}\left(e^{2\pi i\omega\tau_1}, e^{2\pi i\omega\tau_2}, e^{2\pi i\omega\tau_3}\right), \\
h_{j,r} &= \int_{-1/2}^{1/2} \mathbf{H}_j(\omega) e^{2\pi i\omega r} d\omega, \\
&= \frac{1}{N}\int_{-1/2}^{1/2} e^{2\pi i\omega\left(r+\tau_j\right)} d\omega \\
&= \frac{1}{N} I\left(r + \tau_j = 0\right), \ j = 1, 2, 3.
\end{aligned}$$

(Both real and imaginary parts of the integral vanish unless $r+\tau_j = 0$, in which case the exponential

is $\equiv 1$.) Thus

$$H_r = \frac{1}{N}\left(I\left(r+\tau_1=0\right),...,I\left(r+\tau_N=0\right)\right)$$

and so

$$\begin{aligned}\hat{\beta}_t &= \sum_{r=-\infty}^{\infty} H_r \mathbf{y}_{t-r} \\ &= \frac{1}{N}\sum_{r=-\infty}^{\infty}\sum_{j=1}^{N} I\left(r+\tau_j=0\right)\mathbf{y}_{j,t-r} \\ &= \frac{1}{N}\sum_{j=1}^{N} y_{j,t+\tau_j},\end{aligned}$$

as claimed.

- **Hypothesis testing**. The distributional approximation from the previous lecture can be used here. Write

$$\mathbf{Y}_j\left(\omega_k\right) = \mathbf{B}'(\omega_k)\mathbf{Z}_j\left(\omega_k\right) + \mathbf{V}_j\left(\omega_k\right),$$

where $\mathbf{Y}_j\left(\omega_k\right)$ is the DFT $n^{-1/2}\sum_{t=-\infty}^{\infty} y_{j,t}e^{-2\pi i\omega_k t}$ of $\left\{y_{j,t}\right\}_{t=-\infty}^{\infty}$ and $\mathbf{Z}_j\left(\omega_k\right)$ is the IFT, or the DFT

(approximating the IFT). This results, as before, in

$$\begin{aligned} \mathbf{Y}(\omega_k) &= \mathbf{Z}(\omega_k)\mathbf{b}(\omega_k) + \mathbf{V}(\omega_k), \text{ i.e.} \\ \mathbf{y} &= \mathbf{Z}\mathbf{b}(\omega_k) + \mathbf{v}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{y} &= \mathbf{Y}(\omega_k) = \begin{pmatrix} \mathbf{Y}_1(\omega_k) \\ \vdots \\ \mathbf{Y}_N(\omega_k) \end{pmatrix}, \\ \mathbf{v} &= \mathbf{V}(\omega_k) = \begin{pmatrix} \mathbf{V}_1(\omega_k) \\ \vdots \\ \mathbf{V}_N(\omega_k) \end{pmatrix}, \end{aligned}$$

and $\mathbf{Z} = \mathbf{Z}(\omega) : N \times q$ is as above. The $L \times 1$ error vector $\mathbf{v}$ is considered to have i.i.d. components with variance σ_v^2.

- The MLE of $\mathbf{b} = \mathbf{B}(\omega_k)$ is the minimizer of the quadratic form in the exponent in the density of $\mathbf{y}$, i.e. the minimizer of $(\mathbf{y} - \mathbf{Z}\mathbf{b})^*(\mathbf{y} - \mathbf{Z}\mathbf{b})$. This can be determined exactly as in the real regression case - introduce

$$\mathbf{P} = \mathbf{Z}(\mathbf{Z}^*\mathbf{Z})^{-1}\mathbf{Z}^*;$$

note that $\mathbf{P} = \mathbf{P}^* = \mathbf{P}^2$; verify that

$$\begin{aligned} & (\mathbf{y} - \mathbf{Z}\mathbf{b})^* (\mathbf{y} - \mathbf{Z}\mathbf{b}) \\ = \; & \mathbf{y}^* (\mathbf{I} - \mathbf{P}) \mathbf{y} + (\mathbf{y} - \mathbf{Z}\mathbf{b})^* \mathbf{P} (\mathbf{y} - \mathbf{Z}\mathbf{b}) \\ \geq \; & \mathbf{y}^* (\mathbf{I} - \mathbf{P}) \mathbf{y}, \end{aligned}$$

with equality iff

$$\hat{\mathbf{b}} = \hat{\mathbf{B}}(\omega_k) = (\mathbf{Z}^*\mathbf{Z})^{-1} \mathbf{Z}^*\mathbf{y} = \mathbf{S}_z^{-1}(\omega_k) \mathbf{s}_{zy}(\omega_k),$$

where

$$\mathbf{s}_{zy}(\omega_k) = \mathbf{Z}^*\mathbf{y} = \sum_{j=1}^{N} \overline{\mathbf{Z}_j(\omega_k)} \mathbf{Y}_j(\omega_k).$$

The MLE of the error power (i.e. σ_v^2) is

$$\begin{aligned} & \mathbf{y}^* (\mathbf{I} - \mathbf{P}) \mathbf{y} / N \\ = \; & N^{-1} \left(\sum_{j=1}^{N} \left| \mathbf{Y}_j(\omega_k) \right|^2 - \mathbf{s}_{zy}^*(\omega_k) \mathbf{S}_z^{-1}(\omega_k) \mathbf{s}_{zy}(\omega_k) \right) \\ = \; & N^{-1} \left(s_y^2(\omega_k) - \mathbf{s}_{zy}^*(\omega_k) \mathbf{S}_z^{-1}(\omega_k) \mathbf{s}_{zy}(\omega_k) \right) \\ = \; & N^{-1} s_{y \cdot z}^2(\omega_k), \end{aligned}$$

where

$$s_y^2(\omega_k) = \sum_{j=1}^{N} \left| \mathbf{Y}_j(\omega_k) \right|^2$$

is the estimate of the error power under the hypothesis $H : \mathbf{B}(\omega_k) = \mathbf{0}$. Assuming that some smoothing is done, these competing estimates of the power are replaced by averages over the frequencies $\omega_k + \frac{l}{n}$ ($|l| \leq m = (L-1)/2$). This gives

$$\begin{aligned} SSR(\omega_k) &= \sum_{|l| \leq m} \left\{ \begin{array}{c} \mathbf{s}_{zy}^{*}\left(\omega_k + \frac{l}{n}\right) \mathbf{S}_z^{-1}\left(\omega_k + \frac{l}{n}\right) \\ \mathbf{s}_{zy}\left(\omega_k + \frac{l}{n}\right) \end{array} \right\}, \\ SSE(\omega_k) &= \sum_{|l| \leq m} s_{y \cdot z}^2\left(\omega_k + \frac{l}{n}\right), \\ F &= \frac{SSR(\omega_k)/q}{SSE(\omega_k)/(N-q)} \overset{d}{\approx} F_{2L(n-q)}^{2Lq} \\ & \quad \text{under } H. \end{aligned}$$

- In the partitioned case, we write

$$\beta_r = \begin{pmatrix} \beta_{1,r} \\ \beta_{2,r} \end{pmatrix} \begin{array}{l} \leftarrow q_1 \\ \leftarrow q_2 \end{array}$$

and consider the hypothesis that $\beta_{2,r} = \mathbf{0}$ for all r, at frequency $\omega = \omega_k$. For this, partition

$\mathbf{S}_z(\omega_k)$ and the cross-spectral vector $\mathbf{s}_{zy}(\omega_k)$ compatibly as

$$\begin{aligned}\mathbf{S}_z(\omega_k) &= \begin{pmatrix} \mathbf{S}_{11}(\omega_k) & \mathbf{S}_{12}(\omega_k) \\ \mathbf{S}_{21}(\omega_k) & \mathbf{S}_{22}(\omega_k) \end{pmatrix}, \\ \mathbf{s}_{zy}(\omega_k) &= \begin{pmatrix} \mathbf{s}_{1y}(\omega_k) \\ \mathbf{s}_{2y}(\omega_k) \end{pmatrix}.\end{aligned}$$

Then the residual power under the reduced model is

$$s^2_{y\cdot 1}(\omega_k) = s^2_y(\omega_k) - \mathbf{s}^*_{1y}(\omega_k)\,\mathbf{S}^{-1}_{11}(\omega_k)\,\mathbf{s}_{1y}(\omega_k)$$

and

$$SS_{drop} = \sum_{|l|\le m}\left[s^2_{y\cdot 1}\left(\omega_k + \frac{l}{n}\right) - s^2_{y\cdot z}\left(\omega_k + \frac{l}{n}\right)\right];$$

the test rejects for large values of

$$F = \frac{SS_{drop}/q_2}{MSE} \overset{d}{\approx} F^{2Lq_2}_{2L(N-q)}.$$

A derivation of these F statistics, in a similar framework, is given in Lecture 22.

- Example 7.4 (Example 7.2 continued). Code (contributed by Fraser Newton) on course website.

21. Random coefficient regression

- We consider models

$$\mathbf{y}_t = \sum_{r=-\infty}^{\infty} \mathbf{z}_{t-r}\beta_r + \mathbf{v}_t,$$

where $\mathbf{y}_t$ is $N \times 1$, $\mathbf{z}_{t-r}$ is an $N \times q$ non-random matrix, $\{\beta_r\}$ is a zero mean, uncorrelated (and uncorrelated with $\{\mathbf{v}_t\}$), $q \times 1$ stationary series with spectral matrix $f_\beta(\omega)\mathbf{I}_q$. The zero-mean error series $\{\mathbf{v}_t\}_{t=-\infty}^{\infty}$ has spectral density $f_v(\omega)\mathbf{I}_N$. Thus, with

$$\begin{aligned}
\mathbf{Z}(\omega) &= \begin{pmatrix} \mathbf{Z}_1'(\omega) \\ \vdots \\ \mathbf{Z}_N'(\omega) \end{pmatrix} : N \times q, \\
\mathbf{Z}_j(\omega) &= \sum_{t=-\infty}^{\infty} \mathbf{z}_{j,t} e^{-2\pi i\omega t} : q \times 1,
\end{aligned}$$

the spectral matrix of $\{\mathbf{y}_t\}_{t=-\infty}^{\infty}$ is

$$\mathbf{f}_y(\omega) = f_\beta(\omega)\mathbf{Z}(\omega)\mathbf{Z}^*(\omega) + f_v(\omega)\mathbf{I}_N.$$

- Reason: Put

$$\mu_t = \sum_{r=-\infty}^{\infty} \mathbf{z}_{t-r}\beta_r = \sum_{r=-\infty}^{\infty} \mathbf{z}_r\beta_{t-r},$$

then $\mathbf{y}_t = \mu_t + \mathbf{v}_t$ and since these are uncorrelated, $\Gamma_y(h) = \Gamma_\mu(h) + \Gamma_v(h)$; thus

$$\mathbf{f}_y(\omega) = \mathbf{f}_\mu(\omega) + f_v(\omega)\mathbf{I}_N.$$

Then since

$$\begin{aligned}
\Gamma_\mu(h) &= \text{cov}\left[\sum_{r=-\infty}^{\infty} \mathbf{z}_r\beta_{t+h-r}, \sum_{s=-\infty}^{\infty} \mathbf{z}_s\beta_{t-s}\right] \\
&= \sum_{r,s=-\infty}^{\infty} \mathbf{z}_r \text{cov}\left[\beta_{t+h-r}, \beta_{t-s}\right] \mathbf{z}_s' \\
&= \sigma_\beta^2 \sum_{r=-\infty}^{\infty} \mathbf{z}_r\mathbf{z}_{r-h}',
\end{aligned}$$

we obtain

$$\begin{aligned}
\mathbf{f}_\mu(\omega) &= \sum_{h=-\infty}^{\infty} e^{-2\pi i h\omega}\Gamma_\mu(h) \\
&= \sigma_\beta^2 \sum_{r,h=-\infty}^{\infty} e^{-2\pi i h\omega}\mathbf{z}_r\mathbf{z}_{r-h}' \\
&= f_\beta(\omega)\mathbf{Z}(\omega)\mathbf{Z}^*(\omega).
\end{aligned}$$

- A very special case is the additive noise model $y_t = \beta_t + v_t$.

- In general, we seek to estimate the regression function β_t through "deconvolution". Consider estimators

$$\hat{\beta}_t = \sum_{r=-\infty}^{\infty} H_r \mathbf{y}_{t-r}$$

for $q \times N$ matrices H_r. The minimum MSE estimator is then determined (assigned) from the orthogonality requirement

$$E\left[\left(\beta_t - \hat{\beta}_t\right)\mathbf{y}'_{t-r}\right] = \mathbf{0},\ r = 0, \pm 1, \pm 2, ...$$

and has

$$\mathbf{H}(\omega) = \left[\mathbf{S}_z(\omega) + \theta(\omega)\mathbf{I}_q\right]^{-1}\mathbf{Z}^*(\omega), \tag{21.1}$$

as the Fourier transform of $\{H_r\}$. Here

$$\theta(\omega) = \frac{f_v(\omega)}{f_\beta(\omega)}$$

is the inverse of the signal-to-noise (frequency) ratio. Then (assigned) the minimum mse is

$$E\left[\left(\beta_t-\hat{\beta}_t\right)\left(\beta_t-\hat{\beta}_t\right)'\right] = \int_{-1/2}^{1/2} f_v(\omega)\left[S_Z(\omega)+\theta(\omega)\mathbf{I}_q\right]^{-1}d\omega. \quad (21.2)$$

- Example 7.5 (Example 7.2 continued). The model considered in Example 7.2 was

$$\mathbf{y}_t=\sum_{r=-\infty}^{\infty}\mathbf{z}_{t-r}\beta_r+v_t, \text{ for } \mathbf{z}_t=\begin{pmatrix} I(t=\tau_1)\\ I(t=\tau_2)\\ I(t=\tau_3)\end{pmatrix}.$$

Then

$$\mathbf{Z}(\omega)=\sum_{t=-\infty}^{\infty}e^{-2\pi i t\omega}\mathbf{z}_t=\begin{pmatrix} e^{-2\pi i\tau_1\omega}\\ e^{-2\pi i\tau_2\omega}\\ e^{-2\pi i\tau_3\omega}\end{pmatrix}$$

and (derived earlier) $\mathbf{S}_z=N\ (=3)$; thus

$$\mathbf{H}(\omega)=\left(e^{2\pi i\tau_1\omega},e^{2\pi i\tau_2\omega},e^{2\pi i\tau_3\omega}\right)/\left(N+\theta(\omega)\right),$$

with mse given by (21.2), i.e. $f_v(\omega)/\left(N+\theta(\omega)\right)$. This requires a knowledge of $\theta(\omega)$, which is typically unavailable. That suggests the following approach.

- As in the previous section, we consider a linear model

$$\mathbf{Y}(\omega_k) = \mathbf{Z}(\omega_k)\,\mathbf{b}(\omega_k) + \mathbf{V}(\omega_k),$$

where

$$\mathbf{Y}(\omega_k) = \begin{pmatrix} \mathbf{Y}_1(\omega_k) \\ \vdots \\ \mathbf{Y}_N(\omega_k) \end{pmatrix}, \mathbf{V}(\omega_k) = \begin{pmatrix} \mathbf{V}_1(\omega_k) \\ \vdots \\ \mathbf{V}_N(\omega_k) \end{pmatrix},$$

and $\mathbf{Z} = \mathbf{Z}(\omega) : N \times q$ is as above. Now, however, $\mathbf{b}(\omega_k)$ is viewed as random, with spectral matrix $f_\beta(\omega_k)\,\mathbf{I}_q$. As in Lecture 20, the regression SS ($= SS_{drop}$, measuring the effect of dropping all predictors from the regression model) is

$$\begin{aligned} SSR(\omega_k) &= s_y^2(\omega_k) - s_{y\cdot z}^2(\omega_k) \\ &= \mathbf{s}_{zy}^*(\omega_k)\,\mathbf{S}_z^{-1}(\omega_k)\,\mathbf{s}_{zy}(\omega_k) \\ &= \mathbf{Y}^*(\omega_k)\,\mathbf{Z}(\omega_k)\,\mathbf{S}_z^{-1}(\omega_k)\,\mathbf{Z}^*(\omega_k)\,\mathbf{Y}(\omega_k) \\ &= \mathbf{Y}^*(\omega_k)\,\mathbf{P}\mathbf{Y}(\omega_k) \end{aligned}$$

(recall $\mathbf{P} = \mathbf{Z}(\mathbf{Z}^*\mathbf{Z})^{-1}\mathbf{Z}^* = \mathbf{Z}(\omega_k)\,\mathbf{S}_z^{-1}(\omega_k)\,\mathbf{Z}^*(\omega_k)$)

with expected value

$$\begin{aligned} E\left[SSR\left(\omega_k\right)\right] &= E\left[trSSR\left(\omega_k\right)\right] \\ &= tr\mathbf{P}E\left[\mathbf{Y}\left(\omega_k\right)\mathbf{Y}^*\left(\omega_k\right)\right] \\ &= tr\left(\mathbf{P}\cdot\mathbf{f}_y\left(\omega_k\right)\right) \\ &= tr\left(\mathbf{P}\left\{\begin{array}{c} f_\beta\left(\omega_k\right)\mathbf{Z}\left(\omega_k\right)\mathbf{Z}^*\left(\omega_k\right) \\ +f_v\left(\omega_k\right)\mathbf{I}_N \end{array}\right\}\right) \\ &= f_\beta\left(\omega_k\right)tr\mathbf{S}_z\left(\omega_k\right)+qf_v\left(\omega_k\right). \end{aligned}$$

Smooth $\mathbf{S}_z\left(\omega_k\right)$ and $\mathbf{s}_{zy}\left(\omega_k\right)$ (smoothing parameter $= L$), thus obtaining a smoothed estimate $\widehat{SSR}\left(\omega_k\right)$ with

$$E\left[\widehat{SSR}\left(\omega_k\right)\right] = Lf_\beta\left(\omega_k\right)tr\mathbf{S}_z\left(\omega_k\right)+qLf_v\left(\omega_k\right).$$

Similarly $E\left[SSE\left(\omega_k\right)\right] = E\left[\mathbf{s}^2_{y\cdot z}\left(\omega_k\right)\right] = (N-q)f_v\left(\omega_k\right)$; smoothing $\mathbf{s}^2_{y\cdot z}\left(\omega_k\right)$ gives $\widehat{SSE}\left(\omega_k\right)$ with

$$E\left[\widehat{SSE}\left(\omega_k\right)\right] = L\left(N-q\right)f_v\left(\omega_k\right).$$

Now equate $\widehat{SSR}$ and $\widehat{SSE}$ to their expectations and solve to get estimates of $f_v\left(\omega_k\right)$ and $f_\beta\left(\omega_k\right)$.

- Example 7.6 (Example 7.2 continued). Code (contributed by Junfeng Ma) on course website.

22. Analysis of designed experiments

- Recall the regression model with deterministic inputs and coefficients:

$$\mathbf{y}_t = \sum_{r=-\infty}^{\infty} \mathbf{z}_{t-r}\beta_r + \mathbf{v}_t.$$

Assume now that $\mathbf{z}_s = I(s = 0)\mathbf{Z}$ is a constant $N \times q$ matrix and write this model as

$$\mathbf{y}_t = \mathbf{Z}\beta_t + \mathbf{v}_t.$$

The corresponding frequency domain model is

$$\begin{aligned} \mathbf{Y}(\omega_k) &= \mathbf{Z}\mathbf{b}(\omega_k) + \mathbf{V}(\omega_k), \quad (22.1) \\ \text{i.e. } \mathbf{y} &= \mathbf{Z}\mathbf{b}(\omega_k) + \mathbf{v}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{y} &= \mathbf{Y}(\omega_k) = \begin{pmatrix} \mathbf{Y}_1(\omega_k) \\ \vdots \\ \mathbf{Y}_N(\omega_k) \end{pmatrix}, \\ \mathbf{v} &= \mathbf{V}(\omega_k) = \begin{pmatrix} \mathbf{V}_1(\omega_k) \\ \vdots \\ \mathbf{V}_N(\omega_k) \end{pmatrix}. \end{aligned}$$

The $L \times 1$ error vector $\mathbf{v}$ is considered to have i.i.d. components with variance $f_v(\omega_k)$. The results from the previous treatment hold in this special case - merely replace $\mathbf{Z}(\omega_k)$ by the constant matrix $\mathbf{Z}$.

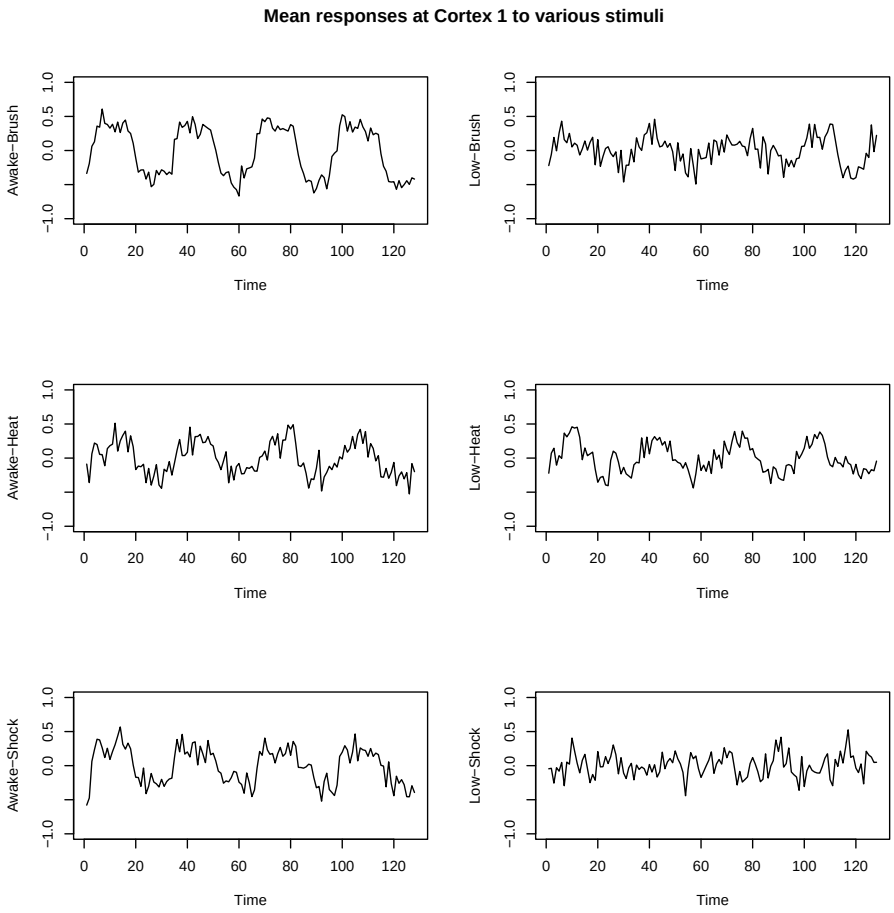

Figure 22.1.

- Example - Figure 22.1 gives the mean responses of subjects to various (six) stimuli in one particular location ("cortex 1") in the brain. The full data set `brainm.dat` has 234 series, each of length 128. There are 26 series at each of 9 locations. These 26 series correspond to 6 types of stimuli administered to a total of 26 subjects (the group sizes differ, and are described in Table 7.4). The stimuli are periodic in nature, applied alternately for 32 seconds, then stopped for 32 seconds. There is an observation made every 2 seconds. If y_{ijt} denotes the response of subject j to stimulus i at time t ($i = 1, ..., I = 6$, $j = 1, ..., N_i$) then we might adopt the model

$$y_{ijt} = \mu_t + \alpha_{it} + v_{ijt},$$

where μ_t is an overall mean and α_{it} is the effect of the i^{th} stimulus, subject as usual to $\sum_i \alpha_{it} = 0$. We can write

$$y_{ijt} = \mathbf{z}'_{ij}\beta_t + v_{ijt},$$

where

$$\beta_t = \left(\tilde{\mu}_t, \tilde{\alpha}_{1t}, ..., \tilde{\alpha}_{I-1,t}\right)'$$

and $\mathbf{z}'_{ij} = (1, I(S_2), ...I(S_6))$, with

$$I(S_k) = I\,(\text{subject } i \text{ gets stimulus } k)\,.$$

(Relate the $\tilde{\mu}$, $\tilde{\alpha}$ to μ, α.) These are then stacked to give $\mathbf{y}_t$ and $\mathbf{Z}$.

- Equality of means: In the framework of the preceding example, the previous theory, for testing that all $\alpha_{it} = 0$, i.e. that the subvector $\beta_{2t} = \left(\tilde{\alpha}_{1t}, ..., \tilde{\alpha}_{I-1,t}\right)' = \mathbf{0}$ results in (assigned)

$$\begin{aligned} SSR\left(\omega_k\right) &= \sum_i \sum_j \left|Y_{i\cdot}\left(\omega_k\right) - Y_{\cdot\cdot}\left(\omega_k\right)\right|^2, \\ SSE\left(\omega_k\right) &= \sum_i \sum_j \left|Y_{ij}\left(\omega_k\right) - Y_{i\cdot}\left(\omega_k\right)\right|^2, \end{aligned}$$

if there is no smoothing. If these DFTs have been smoothed then the test is based on

$$F = \frac{SSR\left(\omega_k\right) / \left(2L\left(I-1\right)\right)}{SSE\left(\omega_k\right) / \left(2L\left(N-I\right)\right)} \overset{d}{\approx} F^{2L(I-1)}_{2L(N-I)},$$

where $N = \sum_i N_i$.

- Other ANOVA models can be considered. For instance in the preceding example the responses might be classified by state of consciousness (awake/anesthetized) as well as by location, leading to the model

$$y_{ijkt} = \mu_t + \alpha_{it} + \beta_{jt} + \gamma_{ijt} + v_{ijkt},$$

when the k^{th} individual receives a stimulus to location i while in state j ($i = 1, ..., I; j = 1, ..., J; k = 1, ..., n_{ij}$). The usual restrictions are

$$\sum_i \alpha_{it} = \sum_j \beta_{jt} = \sum_i \gamma_{ijt} = \sum_j \gamma_{ijt} = 0.$$

One can test for the absence of either effect, i.e. $\alpha_{it} \equiv 0$ or $\beta_{jt} \equiv 0$; presumably one would first test the interaction effects.

- **Derivation of these Fs and the degrees of freedom**. When smoothing is applied, (22.1) becomes (I'm using $\mathbf{v}$ generically; it has i.i.d. elements)

$$\begin{aligned}\begin{pmatrix} \mathbf{Y}(\omega_{k-m}) \\ \vdots \\ \mathbf{Y}(\omega_{k+m}) \end{pmatrix} &= \begin{pmatrix} \mathbf{Zb}(\omega_{k-m}) \\ \vdots \\ \mathbf{Zb}(\omega_{k+m}) \end{pmatrix} + \mathbf{v} \\ &= (\mathbf{I}_L \otimes \mathbf{Z}) \begin{pmatrix} \mathbf{b}(\omega_{k-m}) \\ \vdots \\ \mathbf{b}(\omega_{k+m}) \end{pmatrix} + \mathbf{v}.\end{aligned}$$

Each

$$\mathbf{Y}(\omega_{k+l}) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{y}_t e^{-2\pi i \omega_{k+l} t}$$

is $N \times 1$. Split into the real and imaginary parts:

$$\begin{aligned}\mathbf{Y} &= \mathbf{y}_C - i\mathbf{y}_S,\ \mathbf{b} = \mathbf{b}_C - i\mathbf{b}_S, \\ \mathbf{y}_C &= \begin{pmatrix} \mathbf{Y}_C(\omega_{k-m}) \\ \vdots \\ \mathbf{Y}_C(\omega_{k+m}) \end{pmatrix} : LN \times 1, \\ \mathbf{b}_C &= \begin{pmatrix} \mathbf{b}_C(\omega_{k-m}) \\ \vdots \\ \mathbf{b}_C(\omega_{k+m}) \end{pmatrix} : Lq \times 1,\end{aligned}$$

etc.

Then

$$\begin{pmatrix} \mathbf{y}_C \\ \mathbf{y}_S \end{pmatrix} = \begin{pmatrix} \mathbf{I}_L \otimes \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_L \otimes \mathbf{Z} \end{pmatrix} \begin{pmatrix} \mathbf{b}_C \\ \mathbf{b}_S \end{pmatrix} + \mathbf{v},$$

i.e.

$$\mathbf{u} = \mathbf{P}\boldsymbol{\theta} + \mathbf{v},$$

with

$$\begin{aligned} \mathbf{u} &= \begin{pmatrix} \mathbf{y}_C \\ \mathbf{y}_S \end{pmatrix} : 2LN \times 1, \\ \mathbf{P} &= \mathbf{I}_{2L} \otimes \mathbf{Z} : 2LN \times 2Lq, \\ \boldsymbol{\theta} &= \begin{pmatrix} \mathbf{b}_C \\ \mathbf{b}_S \end{pmatrix} : 2Lq \times 1. \end{aligned}$$

The "reduced" model has this same structure but a smaller $\mathbf{Z}$, say $\mathbf{Z}_{red} : N \times q_1$. Then the corresponding $\boldsymbol{\theta}_{red}$ has $2Lq_1$ elements - we drop $2Lq_2$ parameters ($q = q_1 + q_2$) - and so

$$\begin{aligned} F &= \frac{\left(SS_{red} - SS_{full}\right) / (2Lq_2)}{SS_{full}/(2LN - 2Lq)} \\ &= \frac{\left(SS_{red} - SS_{full}\right) / q_2}{SS_{full}/(N - q)} \\ &\overset{d}{\approx} F^{2Lq_2}_{2L(N-q)}. \end{aligned}$$

- One way to carry out the analysis is to merely treat $\mathbf{u}$ as the data and carry out two regressions - one in the full model and one in the reduced model - and obtain the residual sums of squares (numerically) in each. The elements of $\mathbf{u}$ can in turn be gotten by using the R function `fft()` to compute all of the terms $\sum_{t=0}^{n-1} \mathbf{y}_t e^{-2\pi i \omega t}$ at the desired frequencies. Note the range of t in the sum.

- Here the sums of squares are obtained algebraically. In the full model the minimum SS is

$$\begin{aligned}
\min_{\boldsymbol{\theta}} \|\mathbf{u} - \mathbf{P}\boldsymbol{\theta}\|^2 &= \left\|\mathbf{u} - \mathbf{P}\left(\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{P}^T\mathbf{u}\right\|^2 \\
&= \mathbf{u}^T\left(\mathbf{I} - \mathbf{P}\left(\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{P}^T\right)\mathbf{u} \\
&= \mathbf{u}^T\mathbf{u} - \left(\mathbf{u}^T\mathbf{P}\right)\left(\mathbf{P}^T\mathbf{P}\right)^{-1}\left(\mathbf{P}^T\mathbf{u}\right).
\end{aligned}$$

We have

$$\begin{aligned}
\mathbf{u}^T\mathbf{u} &= \sum_{|l|\leq m} \left[\begin{array}{c} \mathbf{Y}_C^T(\omega_{k+l})\,\mathbf{Y}_C(\omega_{k+l}) \\ +\mathbf{Y}_S^T(\omega_{k+l})\,\mathbf{Y}_S(\omega_{k+l}) \end{array} \right] \\
&= \sum_{|l|\leq m} \mathbf{Y}^*(\omega_{k+l})\,\mathbf{Y}(\omega_{k+l}) \\
&= \sum_{|l|\leq m} s_y^2(\omega_{k+l}) \\
&= Ls_y^2(\omega_k),
\end{aligned}$$

where

$$s_y^2(\omega_k) = \frac{1}{L}\sum_{|l|\leq m} s_y^2(\omega_{k+l}) \tag{22.2}$$

is the smoothed estimate. Then

$$\begin{aligned}
\left(\mathbf{P}^T\mathbf{P}\right)^{-1} &= \mathbf{I}_{2L}\otimes\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}, \\
\mathbf{P}^T\mathbf{u} &= \left(\begin{array}{c} \mathbf{Z}^T\mathbf{Y}_C(\omega_{k-m}) \\ \vdots \\ \mathbf{Z}^T\mathbf{Y}_C(\omega_{k+m}) \\ \mathbf{Z}^T\mathbf{Y}_S(\omega_{k-m}) \\ \vdots \\ \mathbf{Z}^T\mathbf{Y}_S(\omega_{k+m}) \end{array} \right),
\end{aligned}$$

and so

$$\begin{aligned}
& \left(\mathbf{u}^T\mathbf{P}\right)\left(\mathbf{P}^T\mathbf{P}\right)^{-1}\left(\mathbf{P}^T\mathbf{u}\right) \\
= & \sum_{|l|\leq m}\left[\begin{array}{c} \mathbf{Y}_C^T\left(\omega_{k+l}\right)\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{Y}_C\left(\omega_{k+l}\right) \\ +\mathbf{Y}_S^T\left(\omega_{k+l}\right)\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{Y}_S\left(\omega_{k+l}\right) \end{array}\right] \\
= & \sum_{|l|\leq m}\mathbf{Y}^*\left(\omega_{k+l}\right)\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{Y}\left(\omega_{k+l}\right) \\
= & \sum_{|l|\leq m}\mathbf{s}_{zy}^*\left(\omega_{k+l}\right)\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{s}_{zy}\left(\omega_{k+l}\right) \\
= & L\mathbf{s}_{zy}^*\left(\omega_k\right)\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{s}_{zy}\left(\omega_k\right),
\end{aligned}$$

where

$$\begin{aligned}
& \mathbf{s}_{zy}^*\left(\omega_k\right)\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{s}_{zy}\left(\omega_k\right) \qquad (22.3) \\
= & \frac{1}{L}\sum_{|l|\leq m}\mathbf{s}_{zy}^*\left(\omega_{k+l}\right)\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{s}_{zy}\left(\omega_{k+l}\right)
\end{aligned}$$

is the smoothed estimate.

- Putting this all together gives

$$\begin{aligned} SS_{full} &= L\left(s_y^2(\omega_k) - \mathbf{s}_{zy}^*(\omega_k)\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{s}_{zy}(\omega_k)\right) \\ &= Ls_{y.z}^2(\omega_k), \end{aligned}$$

  where these smoothed estimates of the power are defined by (22.2) and (22.3). Similarly,

$$\begin{aligned} SS_{red} &= L\left(s_y^2(\omega_k) - \mathbf{s}_{1y}^*(\omega_k)\left(\mathbf{Z}_{red}^T\mathbf{Z}_{red}\right)^{-1}\mathbf{s}_{1y}(\omega_k)\right) \\ &= Ls_{y.1}^2(\omega_k) \end{aligned}$$

  is computed in exactly the same way, with $\mathbf{Z}$ replaced everywhere by $\mathbf{Z}_{red}$. Then

$$F = \frac{\frac{SS_{red}-SS_{full}}{2Lq_2}}{\frac{SS_{full}}{2L(N-q)}} = \frac{\left(s_{y.1}^2(\omega_k) - s_{y.z}^2(\omega_k)\right)/q_2}{s_{y.z}^2(\omega_k)/(N-q)}.$$

- Example 7.7/7.8. Code (contributed by Junfeng Ma) on course website.

23. Principal Component Analysis

- PCA - a method of reducing the complexity of high dimensional problems, through looking only at certain important linear combinations of the variables of interest.

- We must at times work with complex-valued time series

$$\mathbf{x}_t = \mathbf{x}_{1t} - i\mathbf{x}_{2t};$$

such a series is stationary if $E[\mathbf{x}_t] = \mu$ is time-independent and

$$\Gamma_{xx}(h) = \text{cov}[\mathbf{x}_{t+h}, \mathbf{x}_t] = E[(\mathbf{x}_{t+h} - \mu)(\mathbf{x}_t - \mu)^*]$$

depends only on h. Note that this is also

$$\Gamma_{xx}(h) = E[\mathbf{x}_{t+h}\mathbf{x}_t^*] - E[\mathbf{x}_{t+h}]E[\mathbf{x}_t^*].$$

Properties are similar to real-valued time series: $\Gamma_{xx}(0)$ is Hermitian and non-negative definite, with a real, non-negative diagonal (the variances).

Note that S&S state that $\Gamma_{xx}(h)$ is Hermitian, but in fact

$$\Gamma_{xx}^{*}(h) = \Gamma_{xx}(-h). \tag{23.1}$$

The spectral density matrix is given by

$$\mathbf{f}_{xx}(\omega) = \sum_{h=-\infty}^{\infty} \Gamma_{xx}(h) e^{-2\pi i h \omega}.$$

By (23.1), $\mathbf{f}_{xx}(\omega)$ is Hermitian.

- **Classical PCA**. Suppose that a p-dimensional r.vec. $\mathbf{x}$ has covariance matrix $\mathbf{\Sigma}_{xx}$, so that a linear combination $\mathbf{c}'\mathbf{x}$ has variance $\mathbf{c}'\mathbf{\Sigma}_{xx}\mathbf{c}$. We seek the linear combination y_1 with maximum variance (subject to $\mathbf{c}'\mathbf{c} = 1$); this is the "first principal component". To obtain it, let

$$\lambda_1(\mathbf{\Sigma}_{xx}) \geq \cdots \geq \lambda_p(\mathbf{\Sigma}_{xx}) \geq 0$$

be the eigenvalues, with corresponding orthogonal eigenvectors $\mathbf{e}_1, \cdots, \mathbf{e}_p$, scaled to have unit norms. Then

$$\mathbf{\Sigma}_{xx} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}',$$

where

$$\begin{aligned} \mathbf{E} &= (\mathbf{e}_1, \cdots, \mathbf{e}_p), \\ \Lambda &= diag(\lambda_1, \cdots, \lambda_p) \end{aligned}$$

and

$$\max_{\mathbf{c} \neq \mathbf{0}} \frac{\mathbf{c}'\boldsymbol{\Sigma}_{xx}\mathbf{c}}{\mathbf{c}'\mathbf{c}} = \mathbf{e}_1'\boldsymbol{\Sigma}_{xx}\mathbf{e}_1 = \lambda_1(\boldsymbol{\Sigma}_{xx}).$$

Thus $y_1 = \mathbf{e}_1'\mathbf{x}$ is the first p.c., with variance $\lambda_1(\boldsymbol{\Sigma}_{xx})$. Similarly the second p.c. is defined as the linear combination $y_2 = \mathbf{c}'\mathbf{x}$ with maximum variance, subject to $\mathbf{c}'\mathbf{c} = 1$ and cov$[y_1, y_2] = 0$. To obtain this define $\mathbf{f} = \mathbf{E}'\mathbf{c}$. The side conditions force $\mathbf{f}'\mathbf{f} = 1$ and $f_1 = 0$. Then

$$\text{var}[y_2] = \mathbf{c}'\boldsymbol{\Sigma}_{xx}\mathbf{c} = \mathbf{c}'\mathbf{E}\Lambda\mathbf{E}'\mathbf{c} = \sum_{j=2}^{p} \lambda_j f_j^2$$

is maximized by $f_2 = 1, f_3 = \cdots = f_p = 0$ and so $\mathbf{c} = \mathbf{E}\mathbf{f} = \mathbf{e}_2$, with $y_2 = \mathbf{e}_2'\mathbf{x}$ and var$[y_2] = \lambda_2(\boldsymbol{\Sigma}_{xx})$. In general, the k^{th} p.c. is the linear combination $y_k = \mathbf{c}'\mathbf{x}$ with maximum variance, subject to $\mathbf{c}'\mathbf{c} = 1$ and cov$\Big[y_j, y_k\Big] = 0$ for $j < k$. It is given by

$$y_k = \mathbf{e}_k'\mathbf{x}, \text{ with var}[y_k] = \lambda_k(\boldsymbol{\Sigma}_{xx}).$$

The p principal components thus partition the total variance

$$\sum_{j=1}^{p} \text{var}\left[x_j\right] = \sum_{j=1}^{p} \sigma_{jj} = tr\boldsymbol{\Sigma}_{xx} = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} \text{var}\left[y_j\right]$$

into the variances of p uncorrelated contributors, each "explaining" less of the variance than the preceding one. The proportion of the total variance attributed to the k^{th} p.c. is

$$\frac{\text{var}\left[y_k\right]}{\sum_{j=1}^{p} \text{var}\left[y_j\right]} = \frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j}.$$

The sample p.c.s are defined similarly, but with $\boldsymbol{\Sigma}_{xx}$ replaced by the sample covariance matrix.

- **Time series PCA**. Here we decompose $\mathbf{f}_{xx}(\omega)$ in a similar manner as above. We can write

$$\mathbf{f}_{xx}(\omega) = \mathbf{E}(\omega)\,\Lambda(\omega)\,\mathbf{E}^{*}(\omega),$$

where $\mathbf{E}$ and Λ are respectively the matrix whose columns are the orthonormal eigenvectors, and the diagonal matrix of eigenvalues, which satisfy:

$$\mathbf{f}_{xx}(\omega)\,\mathbf{e}_j(\omega) = \lambda_j \mathbf{e}_j(\omega).$$

These eigenvalues are necessarily real since $\mathbf{f}_{xx}(\omega)$ is Hermitian (how?). At a fixed frequency ω we seek complex-valued, univariate time series of the form

$$y_{t,k} = \mathbf{c}_k^*(\omega)\,\mathbf{x}_t$$

such that $\mathbf{c}_k^*(\omega)\,\mathbf{f}_{xx}(\omega)\,\mathbf{c}_k(\omega)$ is maximized, subject to $\mathbf{c}_k^*(\omega)\,\mathbf{c}_k(\omega) = 1$ and $\mathbf{c}_j^*(\omega)\,\mathbf{f}_{xx}(\omega)\,\mathbf{c}_k(\omega) = 0$ for $j < k$. Note that

$$\mathbf{c}_k^*(\omega)\,\mathbf{f}_{xx}(\omega)\,\mathbf{c}_k(\omega) = \sum_{h=-\infty}^{\infty} \text{cov}\left[y_{t+h,k}, y_{t,k}\right] e^{-2\pi i h\omega}$$

is the spectrum of the univariate series $\left\{y_{t,k}\right\}_{t=-\infty}^{\infty}$ and

$$\mathbf{c}_j^*(\omega)\,\mathbf{f}_{xx}(\omega)\,\mathbf{c}_k(\omega) = \sum_{h=-\infty}^{\infty} \text{cov}\left[y_{t+h,j}, y_{t,k}\right] e^{-2\pi i h\omega}$$

is the cross-spectrum; thus $\left\{y_{t,j}\right\}_{t=-\infty}^{\infty}$ and $\left\{y_{t,k}\right\}_{t=-\infty}^{\infty}$ are uncorrelated at all lags (and the squared coherency is zero at frequency ω) , if $j \neq k$. Exactly as above,

$$y_{t,k} = \mathbf{e}_k^*(\omega)\,\mathbf{x}_t;$$

this is the "k^{th} p.c. at frequency ω".

- PCA in the frequency domain can be motivated differently. Suppose that we observe $\{\mathbf{x}_t\}$ but seek to transmit a univariate approximation

$$y_t = \sum_{j=-\infty}^{\infty} \mathbf{c}_j^* \mathbf{x}_{t-j}.$$

The receiver of the signal hopes to reconstruct the multivariate $\{\mathbf{x}_t\}$ via

$$\hat{\mathbf{x}}_t = \sum_{j=-\infty}^{\infty} \mathbf{b}_j y_{t-j},$$

in such a manner that the mse

$$mse\,(\mathbf{b}) = E\left[(\mathbf{x}_t - \hat{\mathbf{x}}_t)^* (\mathbf{x}_t - \hat{\mathbf{x}}_t)\right]$$

is minimized. The solution to this problem is stated in Brillinger's book (see "Course Information") and proven in his 1969 paper "The canonical analysis of stationary time series", pp. 331-350 in Multivariate Analysis - II (ed. P.R. Krishnaiah). It is similar to that given above: the IFT of $\left\{\mathbf{c}_j\right\}$ is $\mathbf{c}\,(\omega) = \mathbf{e}_1\,(\omega)$, and that of $\left\{\mathbf{b}_j\right\}$ is $\mathbf{b}\,(\omega) = \overline{\mathbf{e}_1\,(\omega)}$; then

$$\mathbf{c}_j = \int_{-1/2}^{1/2} \mathbf{e}_1\,(\omega)\, e^{2\pi i \omega} d\omega = \bar{\mathbf{b}}_j.$$

- In practice one replaces $\mathbf{f}_{xx}(\omega)$ by a (smoothed) periodogram; S&S generalize to

$$\hat{\mathbf{f}}_{xx}(\omega_k) = \sum_{|l| \leq (L-1)/2} h_l \mathbf{I}(\omega_{k+l})$$

where $\mathbf{I}(\omega_k) = \mathbf{X}(\omega_k)\mathbf{X}^*(\omega_k)$ and $\{h_l\}$ are symmetric, non-negative weights summing to one. They discuss the asymptotic properties of the sample p.c.s.

- Example 7.14. Signals are observed every two seconds for 256 seconds, at each of eight locations in the brain, after a stimulus is applied. The first four and second four series are plotted in Figure 23.1. Series 1 – 4 represent locations in the cortex, 5 and 6 represent locations in the thalamus, and 7 and 8 represent locations in the cerebellum. The stimulus was applied for 32 seconds and then stopped for 32 seconds, for a "signal period" of 64 seconds; this translates to a frequency of 4 cycles in 256 seconds (128 time periods): $\omega = 4/128 = 1/32$ cycles per unit time.

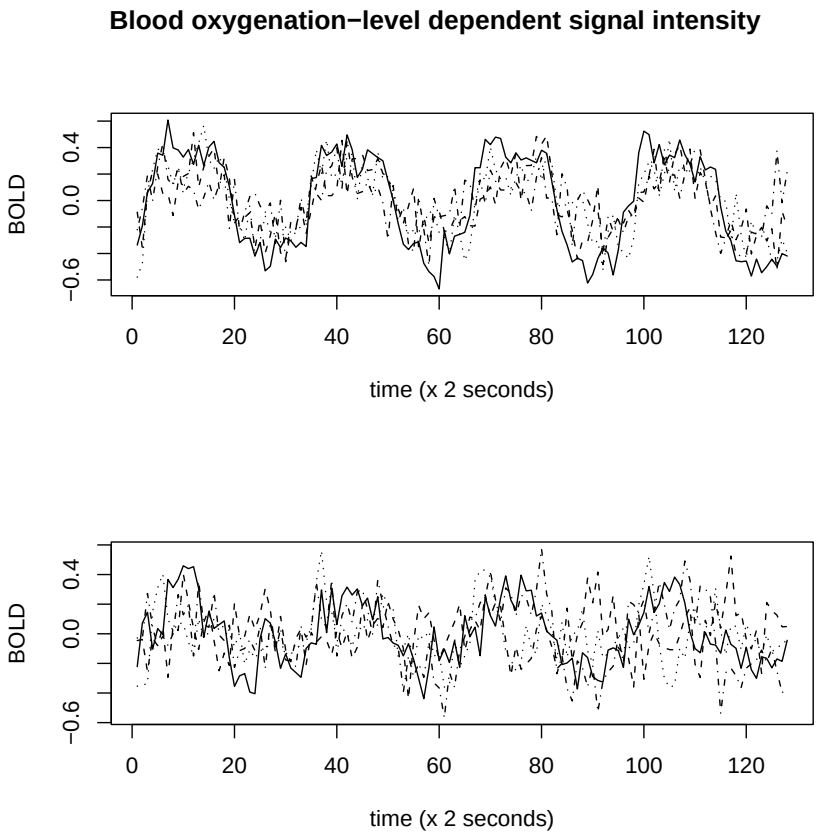

Figure 23.1. fMRI data; first four series (top) and second four series (bottom).

The 8 series were first standardized by dividing each by its sample standard deviation. Then the spectral density matrix was computed at each frequency $\omega_k = k/128$, $k = 1, ..., 64$. The 8 individual spectra are in Figure 23.2. The eigenvalues were computed at each of these frequencies; the largest at each frequency are plotted in Figure 23.3.

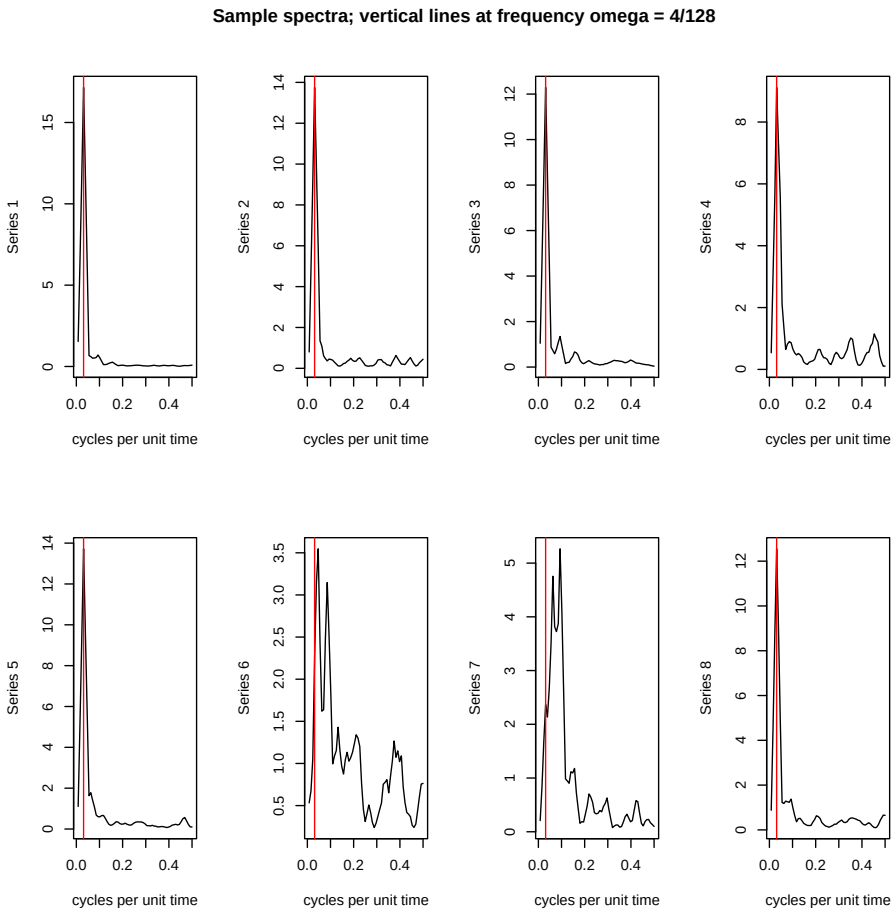

Figure 23.2. Spectra of the eight individual fmri series; all but series 6,7 peak at about $\omega = 4/128$.

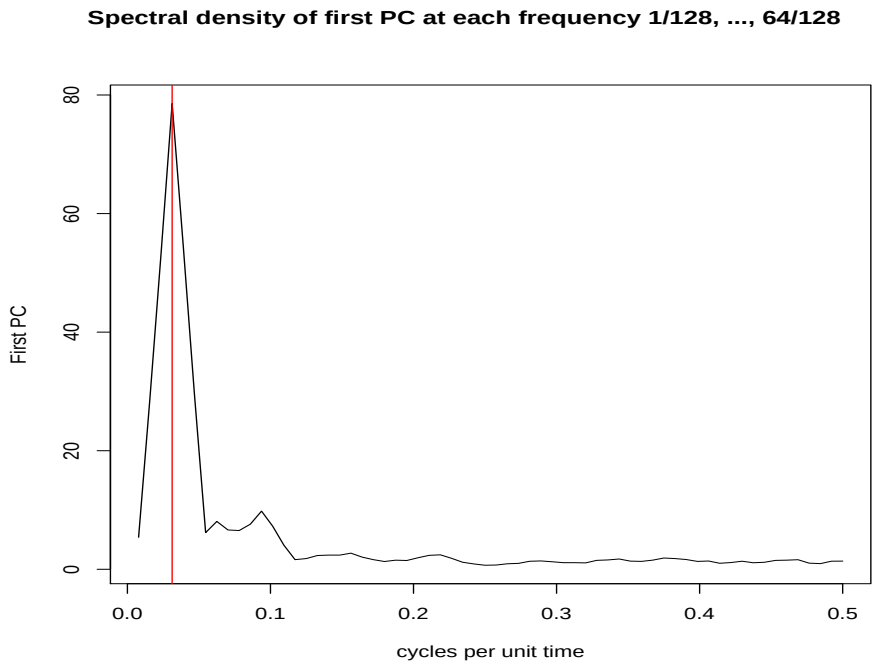

Figure 23.3. Spectral density of first principal component $\lambda_1(\omega_k)$, $\omega_k = 1/128, ...64/128$.

The spectra matrix $\hat{\mathbf{f}}_{xx}(4/128)$ was then decomposed:

```
Eigenvalues at frequency 4/128 are
  78.57 2.99 1.25 0.3 0.1 0 0 0
Contribution of the first PC is 94.42%
Coefficients of first PC are
 -0.47+0i -0.4+0.11i -0.38+0.1i -0.32-0.04i
 -0.41-0.04i 0.1+0.04i -0.13-0.06i -0.39-0.03i
with moduli 0.47 0.42 0.39 0.32 0.42 0.1 0.14 0.39
```

The analysis suggests that locations 1-5 and 8 are the major contributors to the power at this frequency.

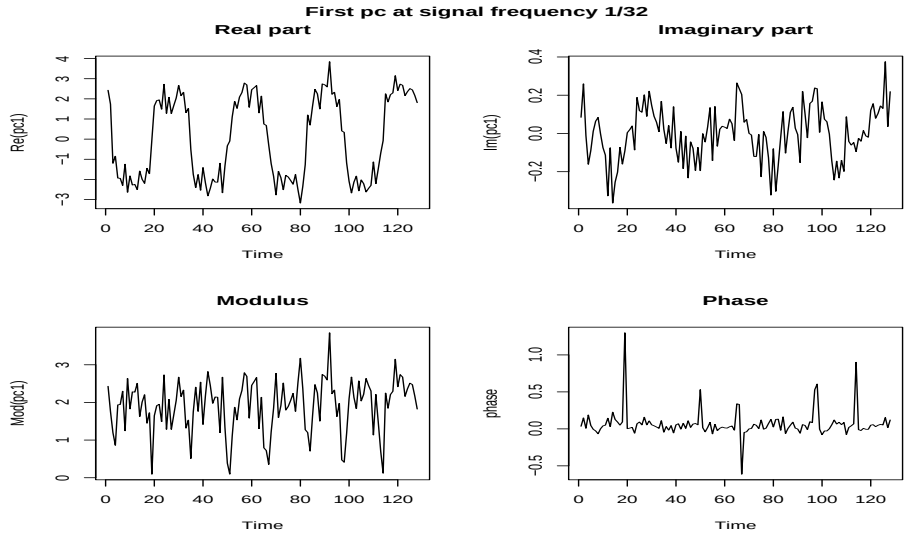

Figure 23.4.

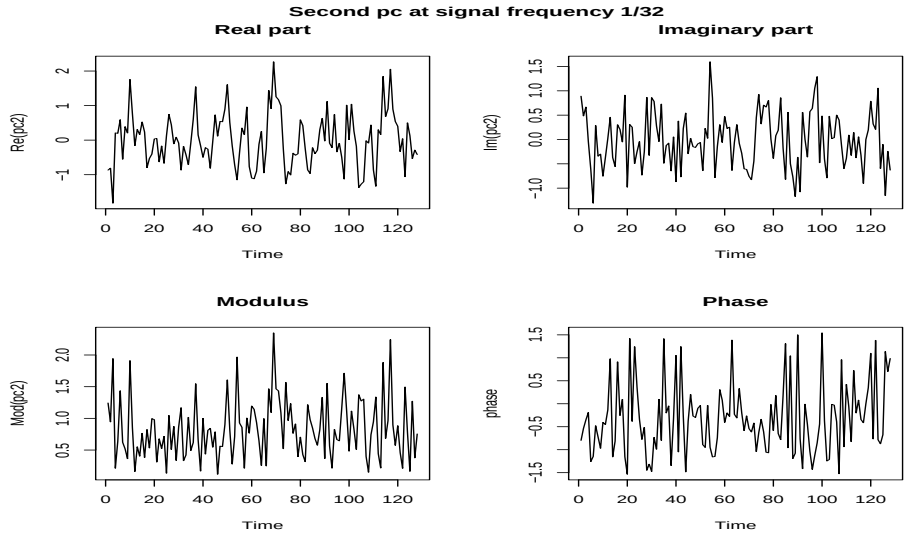

Figure 23.5.

24. Discrimination

24.1. Time Domain Discrimination

- Example: Seismic signals arising from either earthquakes or mining explosions (8 of each) at locations in Scandinavia are each split into initial (P) signals and later (S) signals. Those labelled 5 and 6 are plotted, for comparison with the Novaya Zembla ("NZ"; in Russia) event, in Figure 24.1. (Earthquake i and explosion i are unrelated - only the labels are the same.) Mining explosions are used as surrogates for nuclear explosions, and the purpose of the study is to classify the NZ event as an earthquake or an explosion. More details are in Kakizawa et al. on the course website.

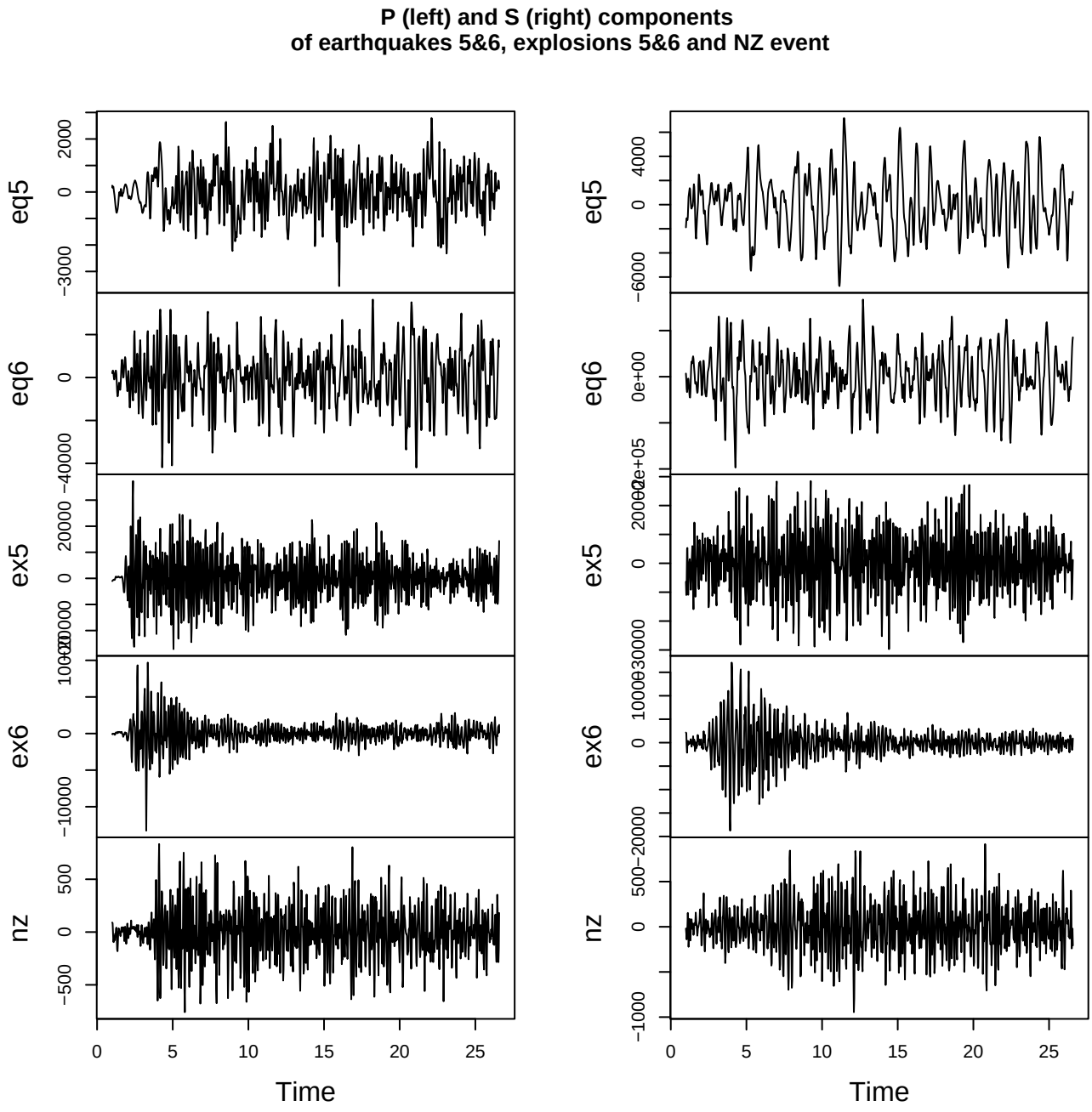

Figure 24.1. P and S components of events 5, 6 and NZ event.

- We are given a p-dimensional data vector $\mathbf{x}$ known to have arisen from one of g populations Π_1, ..., Π_g. We wish to classify it in such way as to minimize the misclassification probabilities. This calls for a partition $\cup_{j=1}^{g} R_j$ of the sample space; we then assign $\mathbf{x}$ to Π_j iff $\mathbf{x} \in R_j$.

- Let $p_i(\mathbf{x})$ be the density corresponding to population i, so that the probability of misclassifying an observation from Π_i into Π_j is

$$P(j|i) = \int_{R_j} p_i(\mathbf{x})\, d\mathbf{x}.$$

  Let π_i be the prior probability of Π_i, then the probability that an observation arises from Π_i and is misclassified into Π_j is $\pi_i P(j|i)$, and the total error probability is

$$P_e = \sum_i \pi_i \sum_{j \neq i} P(j|i).$$

  This is minimized by the rule "classify $\mathbf{x}$ into Π_i if $\pi_i p_i(\mathbf{x}) = \max$".

- **Proof**: Write $P(j|i) = \int_\Omega p_i(\mathbf{x}) I\left(\mathbf{x} \in R_j\right) d\mathbf{x}$, then

$$\begin{aligned} P_e &= \sum_i \int_\Omega \pi_i p_i(\mathbf{x}) \sum_{j \neq i} I\left(\mathbf{x} \in R_j\right) d\mathbf{x} \\ &= \sum_i \int_\Omega \pi_i p_i(\mathbf{x}) \left[1 - I(\mathbf{x} \in R_i)\right] d\mathbf{x} \\ &= 1 - \int_\Omega \sum_i \pi_i p_i(\mathbf{x}) I(\mathbf{x} \in R_i) d\mathbf{x}; \end{aligned}$$

  this is minimized by choosing the R_i so as to maximize the integrand at each point $\mathbf{x}$.

- The posterior probability of membership in Π_i, conditional on observing $\mathbf{x}$, is

$$P(\Pi_i|\mathbf{x}) = \frac{\pi_i p_i(\mathbf{x})}{\sum_j \pi_j p_j(\mathbf{x})};$$

  thus another interpretation of the optimal rule is that it classifies $\mathbf{x}$ into the population with the greatest posterior probability.

- If $p_i(\mathbf{x})$ is the $N(\mu_i, \Sigma_i)$ density then the optimal classification rule reduces to the "quadratic

discrimination rule" based on

$$\begin{aligned}\ln\left(\pi_i p_i\left(\mathbf{x}\right)\right) &= g_i\left(\mathbf{x}\right) - \frac{p}{2}\ln 2\pi, \text{ where} \\ g_i\left(\mathbf{x}\right) &= -\frac{1}{2}\ln\left|\Sigma_i\right| - \frac{1}{2}\left(\mathbf{x}-\mu_i\right)'\Sigma_j^{-1}\left(\mathbf{x}-\mu_i\right) \\ &\quad + \ln\pi_i;\end{aligned}$$

then

$$P\left(\Pi_i|\mathbf{x}\right) = \frac{e^{g_i(\mathbf{x})}}{\sum_j e^{g_j(\mathbf{x})}}.$$

- In the case of only two populations (the most common case; e.g. "earthquake" or "explosion") and with equal covariance matrices the rule is based on the "linear discriminant function"

$$\begin{aligned}d\left(\mathbf{x}\right) &= g_1\left(\mathbf{x}\right) - g_2\left(\mathbf{x}\right) \\ &= -\frac{1}{2}\left[\begin{array}{c}\left(\mathbf{x}-\mu_1\right)'\Sigma^{-1}\left(\mathbf{x}-\mu_1\right) \\ -\left(\mathbf{x}-\mu_2\right)'\Sigma^{-1}\left(\mathbf{x}-\mu_2\right)\end{array}\right] + \ln\frac{\pi_1}{\pi_2} \\ &= \left(\mu_1-\mu_2\right)'\Sigma^{-1}\left(\mathbf{x}-\frac{\mu_1+\mu_2}{2}\right) + \ln\frac{\pi_1}{\pi_2};\end{aligned}$$

this rule classifies $\mathbf{x}$ into Π_1 if $d\left(\mathbf{x}\right) \geq 0$. Then

$$P\left(\Pi_1|\mathbf{x}\right) = \frac{e^{d(\mathbf{x})}}{1+e^{d(\mathbf{x})}} = 1 - P\left(\Pi_2|\mathbf{x}\right).$$

- These rules depend on the unknown parameters; one typically substitutes estimates of these computed from the "training samples" - e.g. the 16 series of known origin in the example described above. The performance of the rule can be estimated by applying it to the training data, to get the sample proportions of observations misclassified. This tends to underestimate the true misclassification probabilities, since the method is tested on the same sample for which it was optimized. But cross-validation ("leave one out", for instance) improves on this.

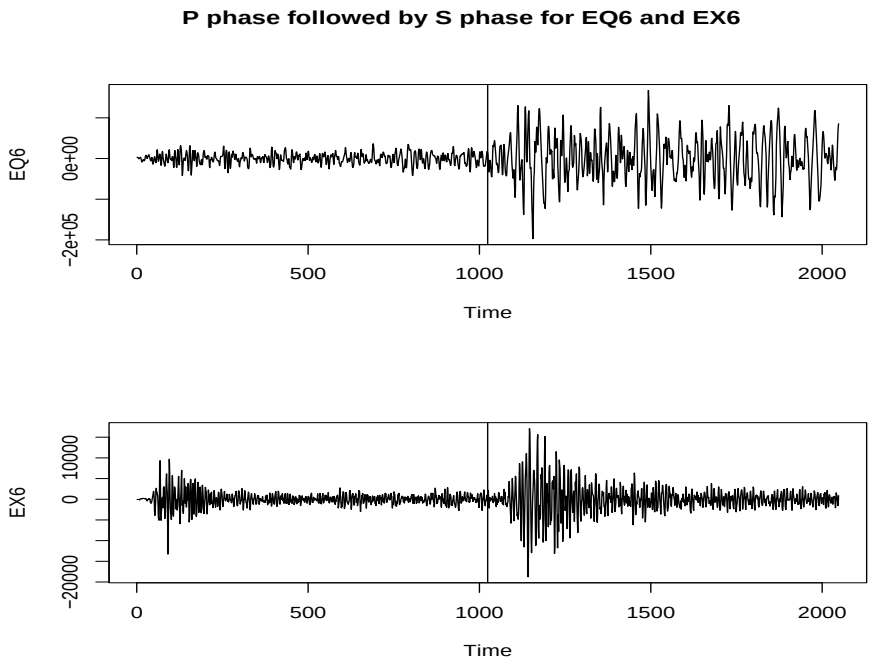

Figure 24.2. Relative (S vs. P) fluctuations tend to be larger for earthquakes than for explosions.

- Example 7.11. It has been noticed that earthquakes and explosions can sometimes be distinguished on the basis of the relative magnitudes of the S and P fluctuations - see Figure 24.2. Thus, for each of the 17 events, we compute

$$\mathbf{x} = \begin{pmatrix} x_P \\ x_S \end{pmatrix} = \begin{pmatrix} \log_{10} range\,(P \text{ series}) \\ \log_{10} range\,(S \text{ series}) \end{pmatrix};$$

where "range" is max − min; S&S call this the "maximum peak-to-peak amplitude". This gives eight points $\mathbf{x}_{eq,i}$ and eight points $\mathbf{x}_{ex,i}$ in the training samples, and $\mathbf{x}_{nz}$ to be classified. The data are:

```
> training data:
     eq.P eq.S ex.P ex.S
[1,] 3.91 4.67 4.55 4.88
[2,] 4.87 5.71 4.74 4.43
[3,] 3.98 4.86 4.93 5.09
[4,] 3.78 4.14 4.71 4.86
[5,] 3.80 4.14 4.81 4.76
[6,] 4.89 5.56 4.36 4.55
[7,] 5.10 6.03 5.05 5.07
[8,] 3.81 4.45 4.09 4.15
> new datapoint:
     NZ.P NZ.S
[1,] 3.20 3.27
```

The sample means and covariance matrices, computed from these 2 bivariate samples $\left\{\mathbf{x}_{eq,i}\right\}_{i=1}^{8}$ and $\left\{\mathbf{x}_{ex,i}\right\}_{i=1}^{8}$ are:

$$\bar{x}_{eq} = \begin{pmatrix} 4.27 \\ 4.95 \end{pmatrix}, \bar{x}_{ex} = \begin{pmatrix} 4.66 \\ 4.73 \end{pmatrix},$$

and

```
> cov.eq = var(cbind(eq.P,eq.S))
> cov.ex = var(cbind(ex.P, ex.S))
> print(round(cov.eq, 4))
       eq.P   eq.S
eq.P 0.3327 0.4109
eq.S 0.4109 0.5394
> print(round(cov.ex, 4))
       ex.P   ex.S
ex.P 0.0973 0.0830
ex.S 0.0830 0.1058
```

and so the "pooled covariance matrix", used to estimate the population covariance matrix $\mathbf{\Sigma}$ (assumed to be the same for both populations) is

$$S_{pool} = \frac{S_{eq} + S_{ex}}{2} = \begin{pmatrix} .2150 & .2470 \\ * & .3226 \end{pmatrix}.$$

The linear discriminant is (assuming equal prior probabilities)

$$\begin{aligned}\hat{d}(\mathbf{x}) &= (\bar{x}_{eq} - \bar{x}_{ex})' S_{pool}^{-1} \left(\mathbf{x} - \frac{\bar{x}_{eq} + \bar{x}_{ex}}{2}\right) \\ &= -21.397 x_P + 17.0601 x_S + 12.9766,\end{aligned}$$

with $\hat{d}(\mathbf{x}_{nz}) = .249$ and $P(EQ|\mathbf{x}) = .562$. Classify NZ - barely - as an earthquake.

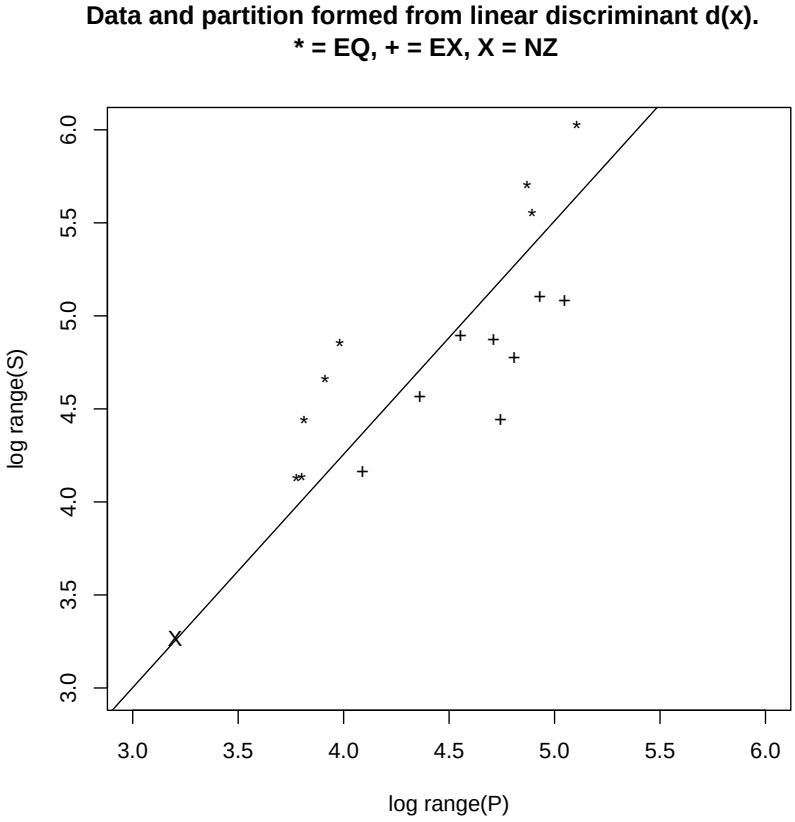

Figure 24.3. Linear function is $\hat{d}(\mathbf{x}) = 0$, i.e. $x_S = 1.25 x_P - .76$.

To assess the classification rule we apply a "leave one out" cross-validation procedure: drop one member of the training sample, determine the classification rule based on those remaining, classify the holdout. Repeat for all 16 possible holdouts. The results are:

```
     EQ P(EQ|data) EX P(EX|data)
[1,]  1      1.000  1      0.718
[2,]  2      0.996  2      1.000
[3,]  3      1.000  3      0.994
[4,]  4      0.856  4      0.990
[5,]  5      0.781  5      1.000
[6,]  6      0.917  6      0.920
[7,]  7      0.997  7      1.000
[8,]  8      0.999  8      0.953
```

All observations are correctly classified by cross-validation.

24.2. Frequency domain discrimination

- We base discrimination in the frequency domain on the comparative densities of the DFTs in the competing populations. Recall the complex normal distribution of the DFT

$$\mathbf{X}(\omega_k) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{x}_t e^{-2\pi i \omega_k}$$

from §18.2:

$$\mathbf{X}(\omega_k) \sim CN_p(\mathbf{M}(\omega_k), \mathbf{C}(\omega_k) - i\mathbf{Q}(\omega_k)),$$

with density

$$\begin{aligned} & p(\omega_k) \\ \approx\ & \pi^{-p} |\mathbf{f}(\omega_k)|^{-1} e^{-\left\{ \begin{array}{c} (\mathbf{X}(\omega_k) - \mathbf{M}(\omega_k))^* \mathbf{f}^{-1}(\omega_k) \\ (\mathbf{X}(\omega_k) - \mathbf{M}(\omega_k)) \end{array} \right\}} \end{aligned}$$

where $\mathbf{f}(\omega_k)$ is the spectral density matrix

$$\mathbf{f}(\omega_k) = \sum_{m=-\infty}^{\infty} \Gamma_{xx}(m) e^{-2\pi i \omega_k m}.$$

If the means and spectral matrices are $\mathbf{M}_j(\omega_k)$ and $\mathbf{f}_j(\omega_k)$ in population Π_j, then the quadratic

discrimination rule, as before, is based on the relative likelihoods $\pi_j \prod_{0<\omega_k<1/2} p_j(\omega_k)$, or equivalently on the relative values of

$$\begin{aligned} & \ln \pi_j + \sum_{0<\omega_k<1/2} \ln p_j(\omega_k) \\ = \; & \ln \pi_j + \sum_{0<\omega_k<1/2} \left[\begin{array}{c} -p \ln \pi - \log \left| \mathbf{f}_j(\omega_k) \right| \\ - \text{ "exponent"} \end{array} \right] \\ = \; & g_j(\mathbf{X}) - \sum_{0<\omega_k<1/2} p \ln \pi, \end{aligned}$$

where

$$\begin{aligned} & g_j(\mathbf{X}) = \ln \pi_j \\ & - \sum_{0<\omega_k<1/2} \left[\begin{array}{c} \log \left| \mathbf{f}_j(\omega_k) \right| \\ + \left\{ \begin{array}{c} \left(\mathbf{X}(\omega_k) - \mathbf{M}_j(\omega_k) \right)^* \\ \mathbf{f}_j^{-1}(\omega_k) \left(\mathbf{X}(\omega_k) - \mathbf{M}_j(\omega_k) \right) \end{array} \right\} \end{array} \right] \end{aligned}$$

(Frequencies at which $\mathbf{f}(\omega_k)$ is singular are also excluded.) The rule is to classify into Π_j if $g_j(\mathbf{X}) = \max$.

- If the populations have time-independent mean vectors μ_j, then we have $\mathbf{M}_j(\omega_k) \equiv \mathbf{0}$ and

$$\begin{aligned} & g_j(\mathbf{X}) = \ln \pi_j \\ & - \sum_{0<\omega_k<1/2} \left[\begin{array}{c} \log \left|\mathbf{f}_j(\omega_k)\right| \\ + tr\left\{\mathbf{X}(\omega_k)\mathbf{X}^*(\omega_k)\mathbf{f}_j^{-1}(\omega_k)\right\} \end{array} \right] \\ = \; & \ln \pi_j \\ & - \sum_{0<\omega_k<1/2} \left[\begin{array}{c} \log \left|\mathbf{f}_j(\omega_k)\right| \\ + tr\left\{\mathbf{I}(\omega_k)\mathbf{f}_j^{-1}(\omega_k)\right\} \end{array} \right], \end{aligned} \quad (24.1)$$

where $\mathbf{I}(\omega_k) = \mathbf{X}(\omega_k)\mathbf{X}^*(\omega_k)$ is the unsmoothed periodogram estimate of the spectral matrix.

- The discrimination rule can be expressed in a variety of ways, using certain measures of disparity. One such measure is the Kullback-Leibler Information, measuring the information which is lost when a density $p_1(x)$ is approximated by $p_2(x)$:

$$I(p_1; p_2) = E_{p_1}\left[\log \frac{p_1(X)}{p_2(X)}\right].$$

Note that, by Jensen's Inequality,

$$\begin{aligned} I(p_1; p_2) &= E_{p_1}\left[-\log \frac{p_2(X)}{p_1(X)}\right] \\ &\geq -\log E_{p_1}\left[\frac{p_2(X)}{p_1(X)}\right] \\ &= 0; \end{aligned}$$

the inequality is strict unless $p_1(x) = p_2(x)$ a.e.

- We now write the K-L information loss in terms of the Whittle likelihood derived above (and call it $I(\mathbf{f}_1; \mathbf{f}_2)$ rather than $I(p_1; p_2)$). The log of the likelihood is

$$\log \prod_{0<\omega_k<1/2} p_j(\omega_k) = g_j(\mathbf{X}) - \ln \pi_j - \sum_{\omega_k} p \ln \pi,$$

so that if $\pi_1 = \pi_2$,

$$\begin{aligned}
& I\left(\mathbf{f}_1; \mathbf{f}_2\right) \\
= \; & E_{p_1}\left[g_1\left(\mathbf{X}\right) - g_2\left(\mathbf{X}\right)\right] \\
= \; & E_{p_1}\left[\sum_{\omega_k}\left\{\begin{array}{c}\left[\begin{array}{c}\log\left|\mathbf{f}_2\left(\omega_k\right)\right| \\ +tr\left\{\mathbf{I}\left(\omega_k\right)\mathbf{f}_2^{-1}\left(\omega_k\right)\right\}\end{array}\right] \\ -\left[\begin{array}{c}\log\left|\mathbf{f}_1\left(\omega_k\right)\right| \\ +tr\left\{\mathbf{I}\left(\omega_k\right)\mathbf{f}_1^{-1}\left(\omega_k\right)\right\}\end{array}\right]\end{array}\right\}\right] \\
= \; & E_{p_1}\left[\sum_{\omega_k}\left\{\begin{array}{c}-\log\frac{\left|\mathbf{f}_1(\omega_k)\right|}{\left|\mathbf{f}_2(\omega_k)\right|} \\ -tr\left[\mathbf{I}\left(\omega_k\right)\left[\mathbf{f}_1^{-1}\left(\omega_k\right) - \mathbf{f}_2^{-1}\left(\omega_k\right)\right]\right]\end{array}\right\}\right] \\
\approx \; & \sum_{\omega_k}\left\{\begin{array}{c}-\log\frac{\left|\mathbf{f}_1(\omega_k)\right|}{\left|\mathbf{f}_2(\omega_k)\right|} \\ -tr\left[\mathbf{f}_1\left(\omega_k\right)\left[\mathbf{f}_1^{-1}\left(\omega_k\right) - \mathbf{f}_2^{-1}\left(\omega_k\right)\right]\right]\end{array}\right\} \\
= \; & \sum_{0<\omega_k<1/2}\left\{\begin{array}{c}tr\left[\mathbf{f}_1\left(\omega_k\right)\mathbf{f}_2^{-1}\left(\omega_k\right)\right] \\ -\log\frac{\left|\mathbf{f}_1(\omega_k)\right|}{\left|\mathbf{f}_2(\omega_k)\right|} - p\end{array}\right\}.
\end{aligned}$$

S&S divide this by n and discuss its asymptotic properties.

- We use this as a measure of disparity between a spectral matrix $\mathbf{f}$ (based on a new series) and its

approximation $\mathbf{f}_j$ (the spectrum for Π_j). The method is to compute (an estimate of)

$$I\left(\mathbf{f};\mathbf{f}_j\right) = \sum_{0<\omega_k<1/2} \left\{ \begin{array}{c} tr\left[\mathbf{f}\left(\omega_k\right)\mathbf{f}_j^{-1}\left(\omega_k\right)\right] \\ -\log\frac{|\mathbf{f}(\omega_k)|}{|\mathbf{f}_j(\omega_k)|} - p \end{array} \right\}$$

$$= \sum_{0<\omega_k<1/2} \left\{ \begin{array}{c} tr\left[\mathbf{f}\left(\omega_k\right)\mathbf{f}_j^{-1}\left(\omega_k\right)\right] \\ +\log\left|\mathbf{f}_j\left(\omega_k\right)\right| \end{array} \right\} + const.$$

Small values indicate close agreement and so we classify into Π_j if

$$\sum_{0<\omega_k<1/2} \left\{tr\left[\mathbf{f}\left(\omega_k\right)\mathbf{f}_j^{-1}\left(\omega_k\right)\right] + \log\left|\mathbf{f}_j\left(\omega_k\right)\right|\right\} = \min;$$

this is exactly the rule based on (24.1) above, if $\mathbf{f}\left(\omega_k\right)$ is estimated by $\mathbf{I}\left(\omega_k\right)$ (or if both are replaced by smoothed periodograms).

- In practice the $\mathbf{f}_j\left(\omega_k\right)$ are themselves estimated; S&S suggest

$$\hat{\mathbf{f}}_j\left(\omega_k\right) = \frac{\sum_{l=1}^{N_j}\left(\mathbf{X}_{jl}\left(\omega_k\right) - \mathbf{X}_{j\cdot}\left(\omega_k\right)\right)^*\left(\mathbf{X}_{jl}\left(\omega_k\right) - \mathbf{X}_{j\cdot}\left(\omega_k\right)\right)}{N_j - 1},$$

where N_j is the number of series in the j^{th} "training series". It seems though that since we are assuming a time-independent mean in each group, one should instead use

$$\begin{aligned}\hat{\mathbf{f}}_j(\omega_k) &= \frac{\sum_{l=1}^{N_j} \mathbf{X}_{jl}(\omega_k)^* \mathbf{X}_{jl}(\omega_k)}{N_j} \\ &= \frac{1}{N_j}\sum_{l=1}^{N_j} \hat{\mathbf{f}}_{jl}(\omega_k),\end{aligned}$$

and also smooth the $\hat{\mathbf{f}}_{jl}$. This is what is done in the following example (and is what it appears S&S do, in the Matlab programme on their website).

- Example 7.12. The eight spectral matrices $\left\{\hat{\mathbf{f}}_{eq,l}\right\}_{l=1}^{8}$ from the training series of earthquakes, and the eight spectral matrices $\left\{\hat{\mathbf{f}}_{ex,l}\right\}_{l=1}^{8}$ from the training series of explosions, were computed and averaged, to get estimates $\hat{\mathbf{f}}_{eq}$ and $\hat{\mathbf{f}}_{ex}$. Then

$$\begin{aligned}I.eq &= \sum_{\omega_k}\left\{tr\left[\hat{\mathbf{f}}_{nz}(\omega_k)\,\hat{\mathbf{f}}_{eq}^{-1}(\omega_k)\right] + \log\left|\hat{\mathbf{f}}_{eq}(\omega_k)\right|\right\}, \\ I.ex &= \sum_{\omega_k}\left\{tr\left[\hat{\mathbf{f}}_{nz}(\omega_k)\,\hat{\mathbf{f}}_{ex}^{-1}(\omega_k)\right] + \log\left|\hat{\mathbf{f}}_{ex}(\omega_k)\right|\right\},\end{aligned}$$

were computed, yielding the output

KL discrepancies($\div n$) between f = f.NZ and the spectra based on the training samples are

```
          I.eq     I.ex
[1,] -5.910203 -7.660834
```

On the basis of this, NZ is classified as an explosion (in agreement with S&S and Kakizawa et al.). Again one can test the method by cross-validation:
Differences in KL discrepancies($\div n$); negative values indicate misclassifications

```
     EQ I.ex-I.eq EX I.eq-I.ex
[1,]  1     -2.18  1     -0.26
[2,]  2      0.60  2      2.01
[3,]  3      1.03  3      0.02
[4,]  4      1.30  4      1.07
[5,]  5      1.28  5      3.81
[6,]  6      1.06  6      0.55
[7,]  7      0.50  7      0.76
[8,]  8      0.76  8     -1.67
```

Relative to S&S, the same two explosions were misclassified but, as well, earthquake 1 was misclassified as an explosion. As in S&S the data were first scaled by dividing each of the 34 series by its range; this greatly improved the performance, as judged by cross-validation.

25. Clustering

- Given a collection $\{\mathbf{X}_i\}_{i=1}^N$ of (possibly multivariate) series, we seek to "cluster" them into similar groups. This is more exploratory in nature than discrimination, where we classify into *known* groups.

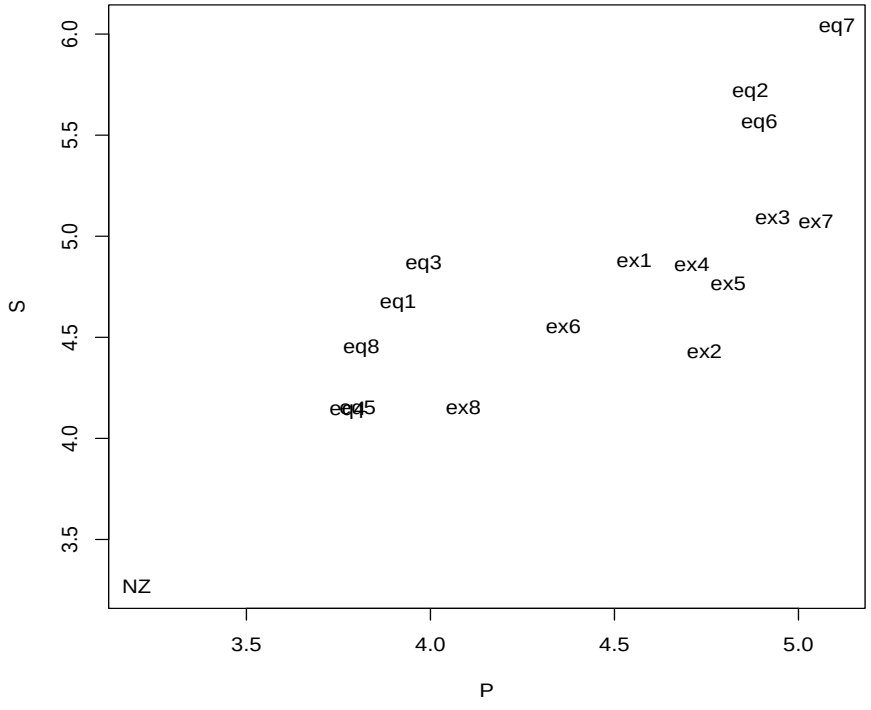

Figure 25.1. Explosion/earthquake $\log_{10} P$ and $\log_{10} S$ values.

- We need a measure by which we can judge pairs of series to be more or less similar than other pairs. In the time domain the usual divergence measure is the Euclidean distance $J(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$. (The "Manhattan" distance $J(\mathbf{x}, \mathbf{y}) = \sum_i |x_i - y_i|$ is also popular.) In the frequency domain we will use the measure

$$J(\mathbf{f}_1; \mathbf{f}_2) = I(\mathbf{f}_1; \mathbf{f}_2) + I(\mathbf{f}_1; \mathbf{f}_2),$$

which can be expressed as

$$J(\mathbf{f}_1; \mathbf{f}_2) = \sum_{\omega_k} tr\left[\mathbf{F}_{12}(\omega_k) + \mathbf{F}_{21}(\omega_k)\right] + const.$$

for $\mathbf{F}_{12}(\omega_k) = \mathbf{f}_1(\omega_k)\mathbf{f}_2^{-1}(\omega_k)$ (and so $\mathbf{F}_{21} = \mathbf{F}_{12}^{-1}$). The constant can be ignored. The divergence between a cluster $\{\mathbf{X}_i | i \in I\}$ and a cluster $\{\mathbf{X}_k | k \in K\}$ is defined to be one of

(i) "average":

$$J(\mathbf{f}_I; \mathbf{f}_K) = aver_{i \in I} aver_{k \in K} J(\mathbf{f}_i; \mathbf{f}_k),$$

(ii) "nearest neighbour" ("single linkage"):

$$J(\mathbf{f}_I; \mathbf{f}_K) = \min_{i \in I, k \in K} J(\mathbf{f}_i; \mathbf{f}_k),$$

or

(iii) $J(\mathbf{f}_I; \mathbf{f}_K) = J(aver_{i \in I} \mathbf{f}_i; aver_{k \in K} \mathbf{f}_k)$.

- **Hierarchical clustering**. Begin by considering each series to be a cluster, and form a cluster of size 2 by grouping together the pair (i, j) for which $J\left(\mathbf{f}_i; \mathbf{f}_j\right) = \min$. At each step the two least divergent clusters are merged, until at the final step there is only one cluster. Either (i) or (ii) is used. The R function `agnes` (first load the `cluster` library) will carry this out; it is contributed by Anja Struyf, Mia Hubert and Peter J. Rousseeuw and described in their paper on the course website.

- Example 7.13. Using the earthquake/explosion data, we first carry out hierarchical clustering in the time domain, using the P and S values.

```
cluster1 = agnes(PSdata, method = "single")
pltree(cluster1, labels = colnames(data.combined))
```

carries out hierarchical clustering using distance (ii). Using method = "average" resulted in the same "dendogram" as in Figure 25.2.

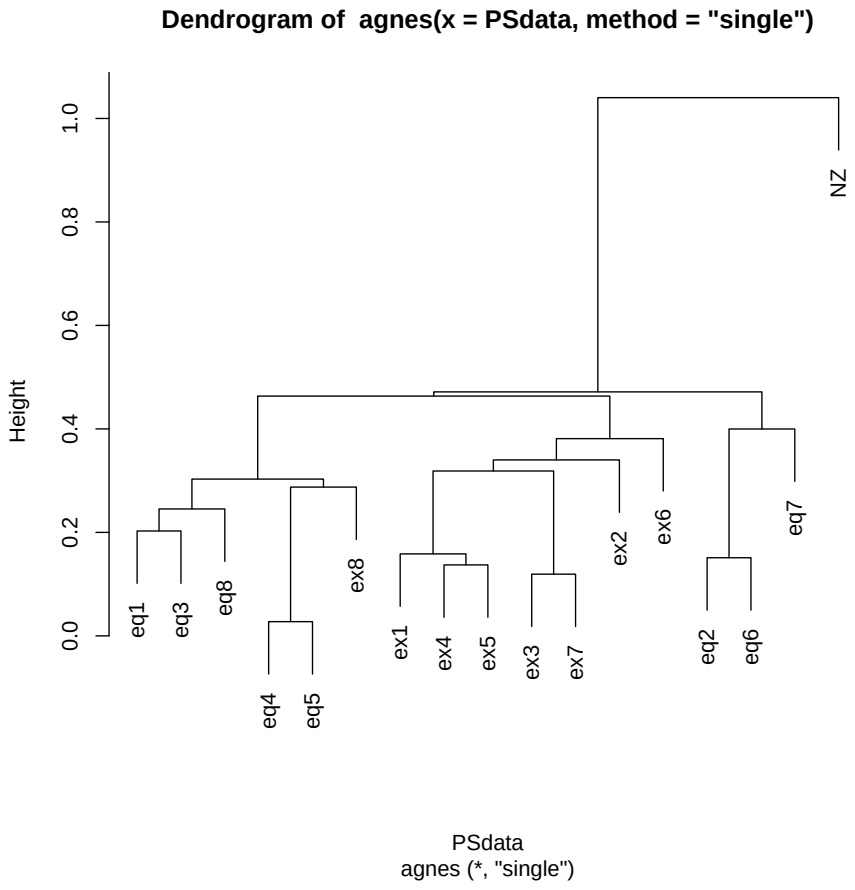

Figure 25.2. Hierarchical clustering using P/S values.

If four groups, they are:

Group	EQ	EX	
1	1,3,4,5,8	8	
2		1,2,3,4,5,6,7	
3	2,6,7		
4			NZ

If three groups, then groups 2 and 3 above are merged.

- In the frequency domain, we first compute all values $J\left(\mathbf{f}_i; \mathbf{f}_j\right)$ of the "dissimilarity matrix" $\mathbf{J}$. (The smallest of these is the divergence between EQ3 and EQ8, with $J = 3.88$.) Then

```
library(cluster)
cluster2 = agnes(J, diss = T, method = "average")
 # Clusters on the basis of the measures in the
 # dissimilarity matrix J
summary(cluster1)
pltree(cluster1, labels = colnames(data.combined))
```

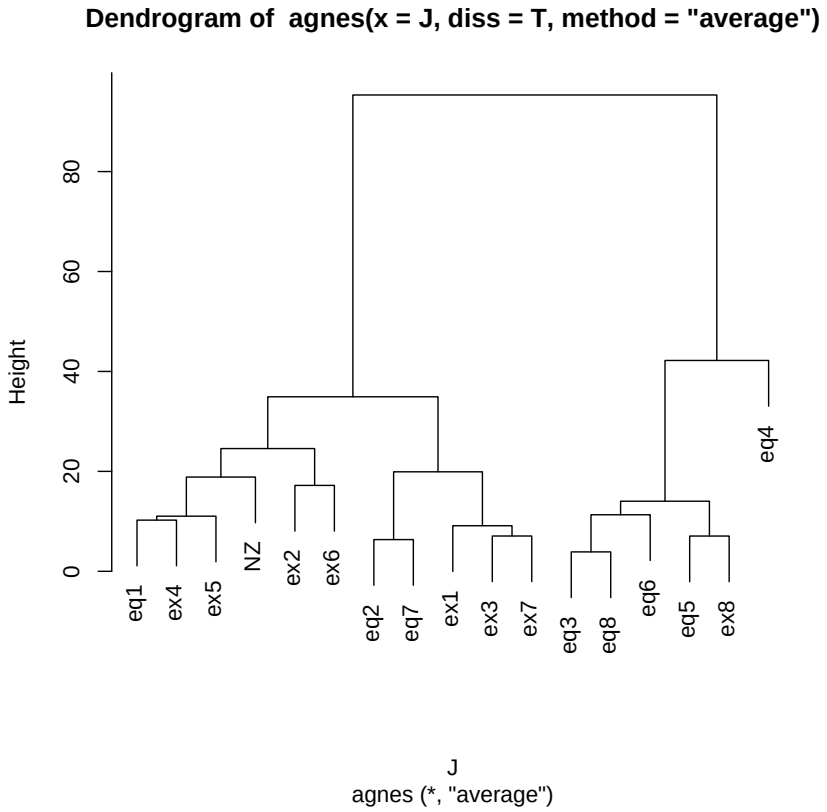

Figure 25.3.

If three groups, they are:

Group	EQ	EX	
1	1,2,7	1,2,3,4,5,6,7	NZ
2	3,5,6,8	8	
3	4		

If two groups, then groups 2 and 3 are merged.

- **Partitioned clustering**. Here one begins with a prespecified number of groups, and an initial partition. Then one computes the divergence between the i^{th} series and each group for $i = 1, ..., N$ and assigns one series to its closest cluster. Repeat, until all series remain in their clusters. In this method the divergence between $\mathbf{X}_i$ and a cluster $\{\mathbf{X}_k | k \in K\}$ is defined by (iii). This requires many recomputations of divergences, and so is unsuitable for this problem in the frequency domain. An alternative (and more robust) method `pam`, also contributed by Struyf et al., will be used.

- The call to `pam` is

```
cluster3 = pam(J, k = 2, diss = T)
summary(cluster3)
plot(cluster3)
```

where `k` is the desired number of clusters. The plot is a "silhouette plot"; roughly speaking the values (scaled to lie in $(-1, 1)$) are interpreted as:

$s \approx 1 \Rightarrow$ the object is well-classified in the indicated group;

$s \approx 0 \Rightarrow$ the object is intermediate between the indicated group and the 'second-best' group;

$s \approx -1 \Rightarrow$ the object is badly classified (closer to the second-best group).

See Figures 25.4, 25.5 for robust partitioned clustering into $k = 2$ and $k = 3$ groups.

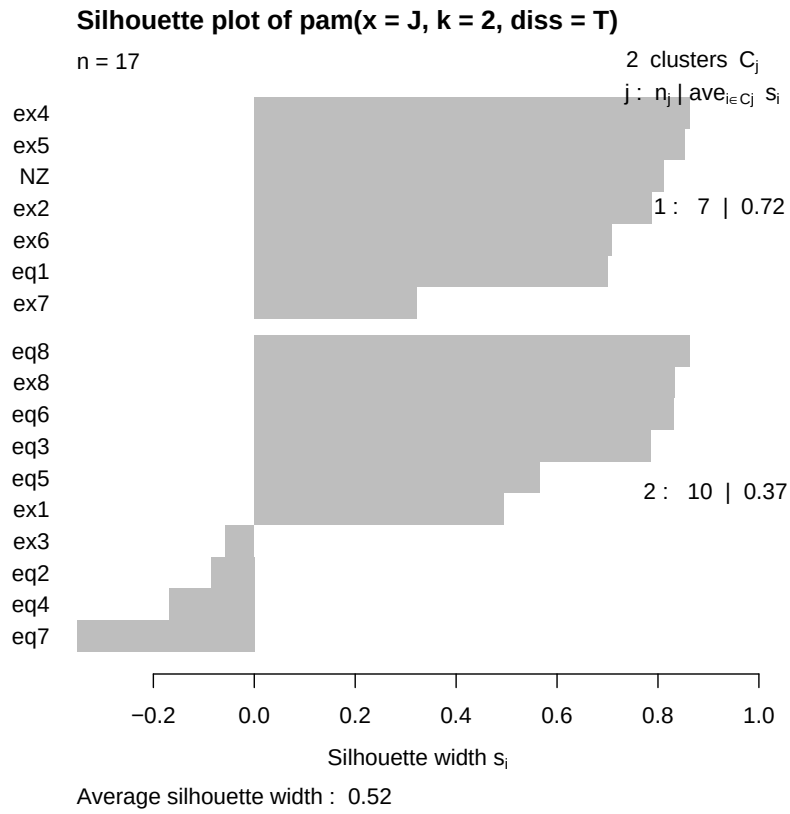

Figure 25.4. Silhouette plot for 2-group partitioned clustering.

Group	EQ	EX	
1	1	2,4,5,6,7	NZ
2	2,3,4,5,6,7,8	1,3,8	

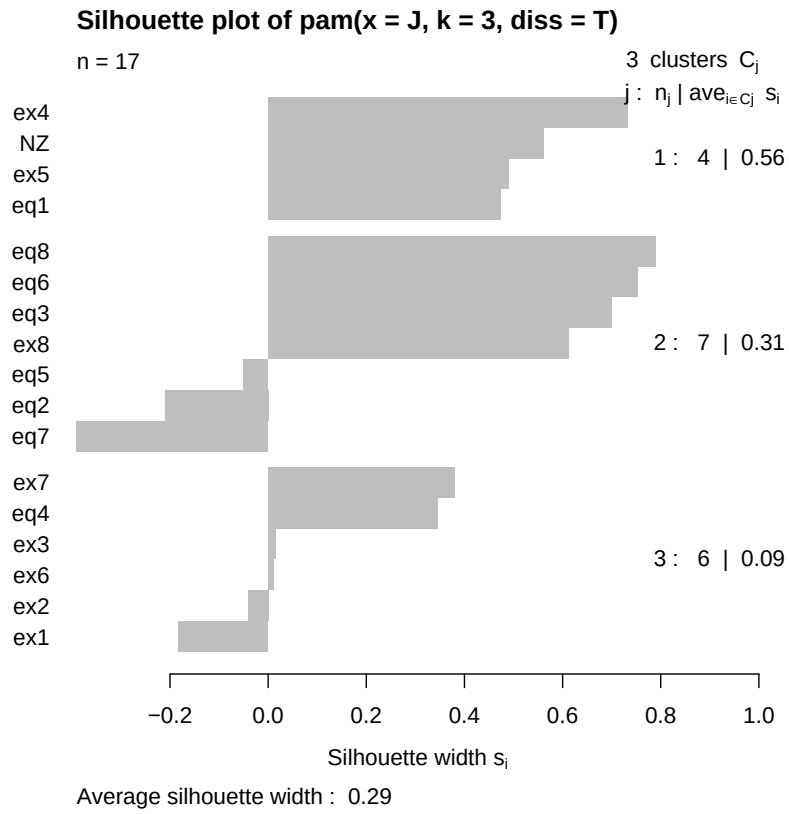

Figure 25.5. Silhouette plot for 3-group partitioned clustering.

Group	EQ	EX	
1	1	4,5	NZ
2	2,3,5,6,7,8	8	
3	4	1,2,3,6,7	

- Clustering in the time domain gives more definitive clusters:

```
cluster5 = pam(PSdata, k = 2, diss = F)
summary(cluster5)
plot(cluster5, labels = 2)
```

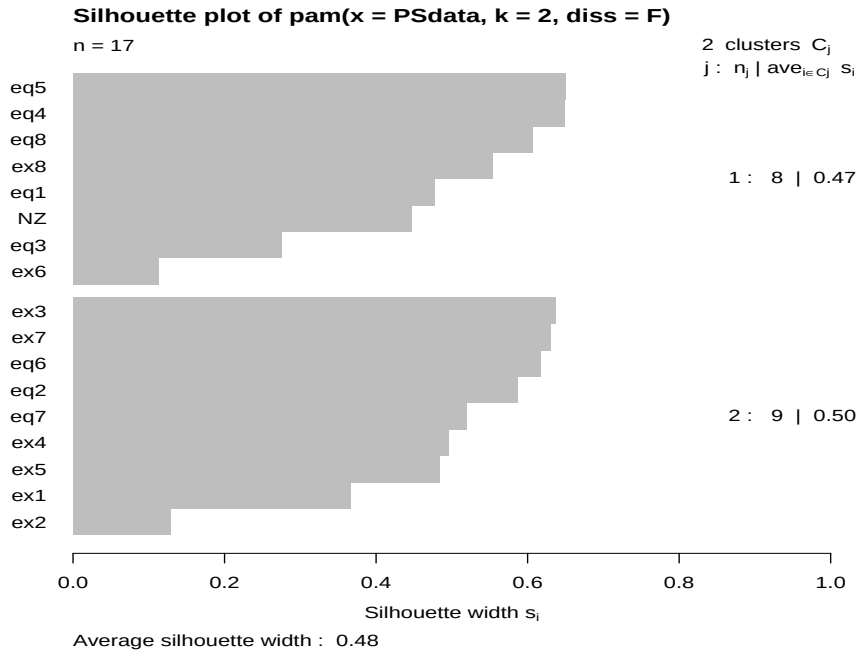

Figure 25.6.

Group	EQ	EX	
1	1,3,4,5,8	6,8	NZ
2	2,6,7	1,2,3,4,5,7	