

# Exercises and Lecture Notes

STK 4060, Spring 2022

Version 0.32, 8-Mar-2022

**Nils Lid Hjort**

Department of Mathematics, University of Oslo

## Abstract

These are Exercises and Lecture Notes for the course on time series, modelling and analysis, STK 4060 (Master level) or STK 9060 (PhD level), for the spring semester 2022. The collection will grow during the course.

### 1. The variance of an average of correlated variables

A classical and crucial result from traditional statistics is that if  $x_1, \dots, x_n$  are independent with the same distribution, then  $\text{Var } \bar{x}_n = \sigma^2/n$ , for the data average  $\bar{x} = (1/n) \sum_{i=1}^n x_i$ , where  $\sigma^2$  is the variance of a single observation. This is rather different for models with dependence. Suppose now that  $x_1, \dots, x_n$  is a stationary sequence, with  $\text{cov}(x_k, x_{i+h}) = \sigma^2 \rho(|h|)$ , for some correlation function  $\rho(h) = \text{corr}(x_i, x_{i+h})$ .

(a) Show that

$$\text{Var } \bar{x}_n = \frac{\sigma^2}{n} \left\{ 1 + 2 \sum_{h=1}^n (1 - h/n) \rho(h) \right\} = \frac{\sigma^2}{n} \sum_{h=-n}^n (1 - |h|/n) \rho(j).$$

(b) For the special case of  $\rho(h) = \rho^h$ , called autocorrelation of order 1, show that

$$\text{Var } \bar{x}_n = \frac{\sigma^2}{n} \left\{ 1 + 2 \sum_{h=1}^{n-1} \rho^h - (1/n) \sum_{h=1}^{n-1} h \rho^h \right\} = \frac{\sigma^2}{n} \left\{ \frac{1+\rho}{1-\rho} + O(1/n) \right\}.$$

With a positive autocorrelation, therefore, the variance of  $\bar{x}_n$  becomes clearly bigger than under independence.

(c) Suppose you observe such a stationary time series  $x_1, \dots, x_n$ , with autocorrelation function  $\rho(h) = \rho^h$  for  $h = 1, 2, 3, \dots$ , and with unknown mean  $\mu$ , variance  $\sigma^2$ , and autocorrelation parameter  $\rho \in (-1, 1)$ . If you do the traditional  $\bar{x}_n \pm 1.96 s_n / \sqrt{n}$  interval for  $\mu$ , recommended in 99 statistics books, with  $s_n$  the empirical standard deviation, what will be its confidence coverage level?

(d) Give estimators for  $\mu, \sigma, \rho$ , constructed from the observed time series.

- (e) Give a more careful and appropriate 95 percent confidence interval, taking autocorrelation into account. Note in particular that such a confidence interval *is wider* than the traditional one, when the autocorrelation is positive.

## 2. An autoregressive time series model

Construct a time series  $x_1, x_2, \dots, x_n$  as follows, via i.i.d.  $\varepsilon_1, \dots, \varepsilon_n$  being standard normal. Let  $x_1 = \varepsilon_1$  and then  $x_{t+1} = \rho x_t + \varepsilon_{t+1}$  for  $i = t, 2, \dots$ , where  $\rho$  is a value inside  $(-1, 1)$ .

- (a) Take  $n = 100$  and  $\rho = 0.345$ , and simulate such a time series in your computer. Check what the `acf(xdata)` does, playing also a bit with other combinations of  $n$  and  $\rho$ .
- (b) Write  $\mathcal{F}_t$  for all observed history up to and including time point  $t$ . Show that  $E(x_{t+1} | \mathcal{F}_t) = \rho x_t$  and  $\text{Var}(x_{t+1} | \mathcal{F}_t) = 1$ . Deduce also from this that

$$E x_t = \rho E x_{t-1} \quad \text{and} \quad \text{Var } x_t = 1 + \rho^2 \text{Var } x_{t-1}.$$

Show that  $E x_t = 0$ , for all  $t$ , and find a formula for the variance of  $x_t$ .

- (c) Starting from

$$\begin{aligned} x_2 &= \rho \varepsilon_1 + \varepsilon_2, \\ x_3 &= \rho^2 \varepsilon_1 + \rho \varepsilon_2 + \varepsilon_3, \\ x_4 &= \rho^3 \varepsilon_1 + \rho^2 \varepsilon_2 + \rho \varepsilon_3 + \varepsilon_4, \end{aligned}$$

find a general formula for  $x_t$ , expressed in terms of the i.i.d. components  $\varepsilon_1, \dots, \varepsilon_t$ . Use this to find an explicit distribution of  $x_t$ . Also show

$$\text{Var } X_t = 1 + \rho^2 + \rho^4 + \dots + \rho^{2(t-1)} = \frac{1 - \rho^{2t}}{1 - \rho^2},$$

re-proving what you found in (b).

- (d) Find the explicit covariance and correlation between  $x_i$  and  $x_{i-1}$ .
- (e) When the time series has been at work for some time, show that

$$\text{Var } x_i \rightarrow \frac{1}{1 - \rho^2}, \quad \text{cov}(x_i, x_{i+1}) \rightarrow \frac{\rho}{1 - \rho^2}, \quad \text{cov}(x_i, x_{i+2}) \rightarrow \frac{\rho^2}{1 - \rho^2},$$

etc.

- (f) Show that the real acf (the autocorrelation function) becomes  $1, \rho, \rho^2, \rho^3, \dots$ ,
- (g) Simulate a few time series using the above construction, with a few combinations of  $n$  and  $\rho$ . Verify that with  $n$  moderate-to-large, the empirical `acf(xdata)` becomes close to the real  $1, \rho, \rho^2, \rho^3, \dots$

## 3. Using regression modelling for the Johnson & Johnson dataset

Consider the dataset called `jj` in the `astsa` package, giving the quarterly earnings of the J & J company, from quarter 1 1960 to quarter 4 1980. One wishes to study how these  $y_1, \dots, y_n$  evolve over time (with  $n = 84$  quarters over 21 years), e.g. to predict earnings for the coming year. The task here is to go through some regression models, so to speak before factoring in correlations and specific time series aspects.

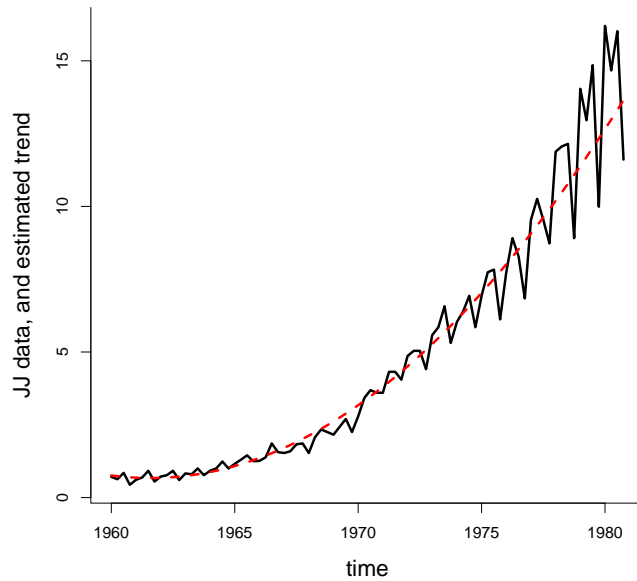


Figure 0.1: The JJ data, with estimated trend, from the five-parameter model.

- (a) Write  $x_t = t - 1960$ , for  $t = 1, \dots, n$ . Fit the rather simple classic linear regression model, with  $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$ , with the  $\varepsilon_t$  taken i.i.d.  $N(0, \sigma^2)$ . Look at the fitted trend  $\hat{m}_1(t) = \hat{\beta}_0 + \hat{\beta}_1 x_t$ , alongside data, to check that this model is far too simple. For the practice, check also the residuals  $r_{1,t} = y_t - \hat{m}_1(t)$ ; these will vary too much, indicating again that this model is too coarse.
- (b) A rather better model is to include a quadratic term for the trend. Fit the regression model  $y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \varepsilon_t$ , again with the  $\varepsilon_t$  taken i.i.d. from the  $N(0, \sigma^2)$ . Plot the estimated trend  $\hat{m}_2(t) = \hat{\beta}_0 + \hat{\beta}_1 x_t + \hat{\beta}_2 x_t^2$  alongside data, examine the residuals  $r_{2,t} = y_t - \hat{m}_2(t)$ , and comment on what you find.
- (c) You learn from the above that the trend function is adequately described by such a parabola, but that that variance of data is not constant; it increases over time. So try the variance heteroscedastic model

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \sigma_t \varepsilon_t, \quad \text{for } t = 1, \dots, n, \quad \text{with } \sigma_t = \exp\{\gamma_0 + \gamma_1(x_t - \bar{x})\},$$

and with the  $\varepsilon_t$  now being i.i.d. and standard normal. The model has three parameters for the mean and two for the variance. Show that the log-likelihood function for this five-parameter model can be expressed as

$$\ell(\theta) = \sum_{t=1}^n \left\{ -\log \sigma_t - \frac{1}{2} (y_t - \beta_0 - \beta_1 x_t - \beta_2 x_t^2)^2 / \sigma_t^2 - \frac{1}{2} \log(2\pi) \right\},$$

in terms of the full parameter vector  $\theta$ .

- (d) Find the maximum likelihood (ML) estimates, say  $\hat{\theta}_{\text{ml}}$ , by numerically maximising the log-likelihood function. Compute also approximate standard errors, for the five parameter esti-

mates, via the general normal approximation theorem for parametric models,

$$\hat{\theta}_{\text{ml}} \approx_d N_p(\theta, \hat{\Sigma}), \quad \text{with } \hat{\Sigma} = \hat{J}^{-1}. \quad (0.1)$$

Here  $\hat{J} = -\partial^2 \ell(\hat{\theta}_{\text{ml}}) / \partial \theta \partial \theta^t$ , the Hesse matrix of second order derivatives, computed at the ML position. Using `nlm` in R you get the Hesse matrix for free, along with the numerical optimisation, using something like

```
hello = nlm(minuslogL,starthere,hessian=T)
```

followed, pretty generically and very usefully, by

```
ML = hello$estimate
Jhat = hello$hessian
se = sqrt(diag(solve(Jhat)))
showme = cbind(ML,se,ML/se)
print(round(showme,4))
```

- (e) Produce a version of Figure 0.1.
- (f) There's at least one more very useful practical thing to learn, following from the general machinery of the Master Theorem (0.1, namely the so-called *delta method*. If one is interested in a certain parameter, say  $\gamma$ , which is a function  $\gamma = g(\theta)$  of the model parameters, then (i) the ML estimator is  $\hat{\gamma}_{\text{ml}} = g(\hat{\theta}_{\text{ml}})$ , i.e. via simple plug-in; and (ii) it is approximately a normal, with

$$\hat{\gamma}_{\text{ml}} \approx_d N(\gamma, \hat{\tau}^2),$$

with  $\hat{\tau}^2 = \hat{c}^t \hat{\Sigma} \hat{c}$ , where  $\hat{c} = \partial g(\hat{\theta}_{\text{ml}}) / \partial \theta$  is the gradient of  $g$ , evaluated at the ML estimate. In R language, if we first programme the  $g$  as a function, we have

```
gammahat = g(ML)
chat = grad(g,ML)
tauhat = sqrt(chat %*% solve(Jhat) %*% chat)
```

I find it practical to include the `numDeriv` package, which has `grad` and `hessian` on board. Now try out such a machinery, by working with  $\gamma$ , the 0.90 quantile of the distribution for the next datapoint, in the JJ estup.

- (g) Once you have the basic code up and running it is relatively easy to try out other variations of such models. Try to put in a cyclic term, perhaps  $\beta_4 \cos(2\pi t/4)$ , and again look at both the residuals and the acf.

#### 4. Understanding the empirical acf, under independence

Suppose  $x_1, x_2, \dots$  are really independent, with mean zero and variance one. What happens then, with the `acd(xdata)`? Below, write  $\bar{x}_{a,b}$  for the average of values  $x_a, \dots, x_b$ .

- (a) Consider first  $A_n = (1/n) \sum_{t=1}^{n-1} x_t x_{t+1}$ . Show that  $A_n$  has mean zero and variance  $(n-1)/n^2$ , i.e. approximately  $1/n$ .

(b) Then go to the proper empirical  $B_n = (1/n) \sum_{t=1}^{n-1} (x_t - \bar{x}_{1,n})(x_{t+1} - \bar{x}_{1,n})$ . Show that

$$B_n = A_n - \frac{n-1}{n} \bar{x}_{1,n} \bar{x}_{1,n-1} - \frac{n-1}{n} \bar{x}_{1,n} \bar{x}_{2,n} + \frac{n-1}{n} \bar{x}_{1,n}^2 \doteq A_n - \bar{x}_{1,n}^2,$$

with  $\doteq$  meaning ‘good approximation, not affecting limits when  $n$  grows’.

(c) Show that  $B_n$ , like the simpler  $A_n$ , has mean zero and variance approximately equal to  $1/n$ . Show then that  $A_n \rightarrow_{\text{pr}} 0$ ,  $B_n \rightarrow_{\text{pr}} 0$ , with ‘ $\rightarrow_{\text{pr}}$ ’ denoting convergence in probability:  $\Pr(|B_n| \geq \varepsilon) \rightarrow 0$  for each small  $\varepsilon$ .

(d) Since  $A_n$  is a sum of variables with the same distribution, with mean zero, and  $\text{Var } A_n \doteq 1/n$ , it is natural to expect limiting normality, i.e.  $\sqrt{n}A_n \rightarrow_d \text{N}(0, 1)$ . This does *not* follow from the traditional CLTs (central limit theorems), since  $x_1x_2$  is not independent of  $x_2x_3$ , etc. Check with the book’s Appendix A.2, however, concerning CLTs for  $m$ -dependent variables, and verify that indeed  $\sqrt{n}A_n \rightarrow_d 1$ .

(e) From  $\sqrt{n}B_n \doteq \sqrt{n}A_n - \sqrt{n}\bar{x}_{1,n}^2$ , show that also  $\sqrt{n}B_n \rightarrow_d \text{N}(0, 1)$ , i.e. the same limit distribution.

(f) Now go from 1-step to 2-step, and work through the details for  $A_n = (1/n) \sum_{t=1}^{n-2} x_t x_{t+2}$  and

$$B_n = \hat{\gamma}(2) = (1/n) \sum_{t=1}^{n-2} (x_t - \bar{x}_{1,n})(x_{t+2} - \bar{x}_{1,n}).$$

The main things are that  $\hat{\gamma}(2) \rightarrow_{\text{pr}} 0$ , the true value of  $\gamma(2)$  under independence, and that  $\sqrt{n}\hat{\gamma}(2) \rightarrow_d \text{N}(0, 1)$ .

(g) Generalise properly to the result  $\sqrt{n}\hat{\gamma}(h) \rightarrow_d \text{N}(0, 1)$ , for

$$\hat{\gamma}(h) = (1/n) \sum_{t=1}^{n-h} (x_t - \bar{x}_{1,n})(x_{t+h} - \bar{x}_{1,n}).$$

(h) So far we’ve assumed variance  $\sigma^2 = 1$ , for simplicity of presentation and argumentation. For the general case, show that for a sequence of independent variables, with some mean  $\mu$  and variance  $\sigma^2$ , we have  $\sqrt{n}\hat{\gamma}(h) \rightarrow_d \text{N}(0, \sigma^4)$ . Finally show that for

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = (1/n) \sum_{t=1}^{n-h} \frac{(x_t - \bar{x}_{1,n})(x_{t+h} - \bar{x}_{1,n})}{\hat{\sigma}} = \frac{\sum_{t=1}^{n-h} (x_t - \bar{x}_{1,n})(x_{t+h} - \bar{x}_{1,n})}{\sum_{t=1}^n (x_t - \bar{x}_{1,n})^2},$$

our good friend the acf, we do have the clarifying easy good result  $\sqrt{n}\hat{\rho}(h) \rightarrow_d \text{N}(0, 1)$ .

(i) For such a sequence of i.i.d. variables, show that when one computes the empirical acf, then

$$\Pr\{\hat{\rho}(h) \in [-1.96/\sqrt{n}, 1.96/\sqrt{n}]\} \rightarrow 0.95,$$

for each lag  $h$ . This is the reason for the ‘magical band’  $\pm 1.96/\sqrt{n}$  provided in the standard use of **acf**.

## 5. A simple moving average process

Suppose  $w_0, w_{\pm 1}, w_{\pm 2}, \dots$  are i.i.d., with finite variance  $\sigma^2$ . Then consider the process

$$x_t = aw_{t-1} + (1 - 2a)w_t + aw_{t+1},$$

with  $a$  a tuning parameter. We call this a moving average process, with window length 3.

- (a) Compute the variance of  $x_t$ , and also the covariance function  $\gamma(h)$  and autocorrelation function  $\rho(h)$ . Plot the acf for a few values of  $a$ , including the equal balance case of  $a = 1/3$ .
- (b) Then do a similar analysis for a 5-window moving average process, of the type

$$x_t = aw_{t-2} + aw_{t-1} + (1 - 4a)w_t + aw_{t+1} + aw_{t+2}.$$

Again, plot the acf for a few values of  $a$ , including the balanced case of  $a = 1/5$ .

- (c) Similarly consider the case of

$$x_t = \rho^2 w_{t-2} + \rho w_{t-1} + w_t + \rho w_{t+1} + \rho^2 w_{t+2}.$$

Find the acf, and plot it, for a few values of  $\rho$ .

## 6. A general stationary normal time series model

Suppose  $x_1, \dots, x_n$  is a stationary normal time series, which means that the full vector has a multinormal distribution; this is also equivalent to saying that all linear combinations are normal. Assume it has mean  $\mu$ , variance  $\sigma^2$ , and correlation function  $\rho(h) = \text{corr}(x_t, x_{t+h})$ .

- (a) Show that the joint distribution of the full series is a  $N_n(\mu\mathbf{1}, \sigma^2 A)$ , where  $\mathbf{1} = (1, \dots, 1)^t$  is the vector of 1s, and  $A$  the  $\hat{A}$   $n \times n$  matrix of  $\rho(s - t)$ , for  $s, t = 1, \dots, n$ ; in particular, the diagonal elements are all 1.
- (b) Using the basic definition of the multinormal joint density, show that the log-likelihood function can be written

$$\ell(\theta) = -n \log \sigma - \frac{1}{2} \log |A| - \frac{1}{2} (y - \mu\mathbf{1})^t A^{-1} (y - \mu\mathbf{1}) / \sigma^2 - \frac{1}{2} n \log(2\pi),$$

with  $\theta$  the parameters involved. If the correlation function is known, then  $A$  is known, and  $\theta$  comprises only  $\mu, \sigma$ . For such a case, show that the ML estimators become

$$\hat{\mu} = \frac{\mathbf{1}^t A^{-1} y}{\mathbf{1}^t A^{-1} \mathbf{1}} \quad \text{and} \quad \hat{\sigma}^2 = \frac{Q_0}{n}, \quad \text{with } Q_0 = (y - \hat{\mu}\mathbf{1})^t A^{-1} (y - \hat{\mu}\mathbf{1}).$$

Check that this leads to familiar formulae in the case of i.i.d. observations, where  $A = I_n$ , the identity matrix.

- (c) If there is a parameter, say  $\lambda$ , in the correlation function, however, we need also  $A = A(\lambda)$ , and we have

$$\ell(\mu, \sigma, \lambda) = -n \log \sigma - \frac{1}{2} \log |A(\lambda)| - \frac{1}{2} (y - \mu\mathbf{1})^t A(\lambda)^{-1} (y - \mu\mathbf{1}) / \sigma^2 - \frac{1}{2} n \log(2\pi).$$

Use the above to find that the log-likelihood profile function, in  $\lambda$ , becomes

$$\ell_{\text{prof}}(\lambda) = -n \log \hat{\sigma}(\lambda) - \frac{1}{2} \log |A(\lambda)| - \frac{1}{2} n - \frac{1}{2} n \log(2\pi).$$

Here one first computes

$$\hat{\mu}(\lambda) = \frac{\mathbf{1}^t A(\lambda)^{-1} y}{\mathbf{1}^t A(\lambda)^{-1} \mathbf{1}} \quad \text{and then} \quad \hat{\sigma}^2(\lambda) = (1/n) Q_0(\lambda),$$

where

$$Q_0(\lambda) = \{y - \hat{\mu}(\lambda)\mathbf{1}\}^t A(\lambda)^{-1} \{y - \hat{\mu}(\lambda)\mathbf{1}\}.$$

- (d) Take e.g.  $n = 100$ , generate  $x_1, \dots, x_n$  from the standard normal in your computer, and fit the three-parameter model which has unknown  $\mu, \sigma, \lambda$ , where the correlation function is modelled as  $\rho(h) = \exp(-\lambda h) = \rho^h$ , i.e. with  $\rho = \exp(-\lambda)$  the 1-step correlation. Repeat the experiment a few times, to see how well the ML estimators succeed in coming close to the true values.

### 7. Conditional multinormal distributions

A vector  $X = (X_1, \dots, X_n)$  has the multinormal distribution, with mean  $\xi$  and covariance matrix  $\Sigma$ , if its density takes the form

$$f(x) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x - \xi)^t \Sigma^{-1} (x - \xi)\right\}.$$

We write  $X \sim N_n(\xi, \Sigma)$  to indicate this; note that the distribution is fully specified by giving the  $\xi$  and the  $\Sigma$ .

- (a) Check that this becomes the classic formula for  $N(\xi, \sigma^2)$  in the one-dimensional case. In the general case, show that  $Y = AX$  has distribution  $N_n(A\xi, A\Sigma A^t)$ , if  $A$  is a  $n \times n$  matrix. Show that  $f$  integrates to 1.
- (b) Block  $X$  into  $X_{(1)}$  and  $X_{(2)}$ , of lengths  $p, q$ , with  $p + q = n$ . Write

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

with  $\Sigma_{11}$  of size  $p \times p$ , etc. Try to show that  $X_{(1)} | (X_{(2)} = x_{(2)})$  is multinormal, in dimension  $p$ , with these important formulae for conditional mean and conditional variance:

$$\begin{aligned} E(X_{(1)} | x_{(2)}) &= \xi_{(1)} + \Sigma_{12} \Sigma_{22}^{-1} (x_{(2)} - \xi_{(2)}), \\ \text{Var}(X_{(1)} | x_{(2)}) &= \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}. \end{aligned}$$

In particular, the conditional mean is a linear function of  $x_{(2)}$ , and the conditional variance is constant.

- (c) For the most simple but still interesting case of a normalised binormal distribution, show that if

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

then  $X_2 | (X_1 = x_1)$  is normal  $(\rho x_1, 1 - \rho^2)$ . Generalise to the case where  $X_1, X_2$  have means  $\xi_1, \xi_2$  and variances  $\sigma_1^2, \sigma_2^2$ .

### 8. Predicting x2 after having seen x1

Part of the business of time series modelling and analysis is *to predict*: what happens next? If we see  $x_1$ , what can we say about the  $x_2$  of tomorrow? It is useful to learn from the multinormal situation.

(a) Suppose

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

with known  $\rho$ , and that  $x_1$  has been observed. In which sense is  $\hat{x}_2 = \rho x_1$  the best prediction for  $x_2$ ? Give a 95 percent prediction interval for  $x_2$ , and discuss how its length is influenced by  $\rho$ .

(b) Suppose  $X_1, \dots, X_n, X_{n+1}$  have a joint multinormal distribution, as for many time series models, and that  $x_1, \dots, x_n$  are observed. Give the distribution for  $X_{n+1}$ , given  $x_1, \dots, x_n$ . Give also a prediction for  $x_{n+1}$ , and a 95 percent prediction interval.

(c) Specialise the above to the case of a stationary Gaussian time series model, with mean  $\mu$ , variance  $\sigma^2$ , and correlation function  $\rho(h)$  for  $h = 1, 2, 3, \dots$ . Again give a prediction, and a prediction interval, for  $x_{n+1}$ , assuming that  $x_1, \dots, x_n$  have been observed.

(d) Discuss how these formulae hold up outside the multinormal situation.

### 9. The AIC and the BIC

Suppose there are competing parametric models for the same dataset, of size  $n$  (the number of observed data points, or data vectors). One first fits these candidate models, say  $M_1, \dots, M_k$ , by maximising their likelihoods. Writing  $\ell_j(\theta_j)$  for model  $M_j$ , we find the ML estimate  $\hat{\theta}_j$  and the maximised log-likelihood value,

$$\ell_{j,\max} = \ell_j(\hat{\theta}_j) \quad \text{for } j = 1, \dots, k.$$

Then we define

$$\text{aic}_j = 2 \dim(\theta_j) - 2\ell_{j,\max} \quad \text{and} \quad \text{bic}_j = \dim(\theta_j) \log n - 2\ell_{j,\max}, \quad (0.2)$$

with  $\dim(\theta_j)$  the number of parameters estimated in that model. These are *the Akaike Information Criterion* and *the Bayesian Information Criterion*; see Chapters 2, 3 in Claeskens and Hjort (2008) for considerably more information. These two information criteria act as *ranking scores* for the competing models, with small values being preferred over bigger ones. Thus there is an AIC winner and a BIC winner (perhaps the same).

Note that these AIC and BIC recipes are completely general; they may be used with independent data, or for time series models with dependence, we may compare normal with non-normal models, and almost apples with bananas.

(a) Explain, in intuitive terms, why these ranking criteria make sense, balancing complexity with model fit. Explain also that the BIC places a harsher penalty on complexity (well, as long as  $n \geq 8$ ).

(b) Suppose I have two coins, with probabilities  $p_a$  and  $p_b$  for ‘krone’. I flip them 40 times each, and get 17 krone with the first and 23 krone with the second. Model 1 says that  $p_a = p_b$ ; model 2 says that  $p_a$  and  $p_b$  are different. Which of these two models is best, according to the AIC, and to the BIC? – Note that we get answers, of the type ‘model 1 is better than model 2’, etc., without using formal null hypothesis tests, and there’s no ‘0.05’ business going on (well, at least not directly).



- (c) Interestingly, it turns out that I have *three* coins in my skuff. I call their krone probabilities  $p_a, p_b, p_c$ , and the number of times I do get a krone, in 40 flips for each, are 17, 23, 26. Carry out AIC and BIC analysis, to rank as many as five candidate models: (i)  $p_a, p_b, p_c$  are equal; (ii)  $p_a = p_b$  but different from  $p_c$  (iii)  $p_a = p_c$  but different from  $p_b$ ; (iv)  $p_b = p_c$  but different from  $p_a$ ; (v) the three are different.
- (d) Suppose a certain start model has dimension  $k$  and log-likelihood maximum value  $\ell_{0,\max}$ , and that one contemplates extending this start model to a bigger one, with one more parameter. Assume specifically that the narrow model lies inside the bigger model. Argue that

$$\Delta = \ell_{1,\max} - \ell_{0,\max}$$

must be positive. Show that AIC thinks the extended model is a good idea, provided  $\Delta > 1$ . The BIC, however, thinks it's only worth the trouble if  $\Delta > \frac{1}{2} \log n$ . – One may show that if the narrow model holds, then  $2\Delta \approx_d \chi_1^2$ , so this can be used to see how likely it is to ‘incorrectly’, or unnecessarily, choose the bigger model, if the narrow model is already ok.

### 10. The AIC and the BIC for linear regression models

We now apply the general AIC and BIC schemes for comparing and ranking different linear regression models, for the same dataset, perhaps to decide on which covariates to include and which to exclude.

- (a) Suppose we have regression data  $(z_t, x_t)$ , for  $t = 1, \dots, n$ , with  $x_t$  the main outcome (perhaps a time series) and  $z_t = (z_{t,1}, \dots, z_{t,k})^t$  a covariate vector of length  $k$ . Consider the classical linear regression model, with

$$x_t = \beta_1 z_{t,1} + \dots + \beta_k z_{t,k} + \varepsilon_t = z_t^t \beta + \varepsilon_t \quad \text{for } t = 1, \dots, n,$$

with the  $\varepsilon_t$  being i.i.d.  $N(0, \sigma^2)$ . Show that the log-likelihood function can be written

$$\ell_k(\beta, \sigma) = -n \log \sigma - \frac{1}{2} Q(\beta) / \sigma^2 - \frac{1}{2} n \log(2\pi), \quad (0.3)$$

with subscript  $k$  for the number of covariates included in the model. Here

$$Q(\beta) = \sum_{t=1}^n \{x_t - m_t(\beta)\}^2, \quad \text{where } m_t(\beta) = E(x_t | z_t) = z_t^t \beta,$$

the classic sum of squares.

- (b) Show that the ML estimator for  $\beta$  is the least sum of squares estimator, with a formula

$$\hat{\beta} = \Sigma_n^{-1} n^{-1} \sum_{t=1}^n z_t x_t = \left( n^{-1} \sum_{t=1}^n z_t z_t^t \right)^{-1} n^{-1} \sum_{t=1}^n z_t x_t,$$

assuming here that there is no linearity between the covariate vectors, so that  $\Sigma_n$  has full rank. Show then that the ML estimator for  $\sigma$  is  $\hat{\sigma}_k^2 = Q_{\min}/n = Q(\hat{\beta})/n$ . Deduce from this that

$$\ell_{k,\max} = \max\{\ell_k(\beta, \sigma) : \text{all } \beta, \sigma\} = -n \log \hat{\sigma}_k - \frac{1}{2} n - \frac{1}{2} n \log(2\pi).$$

(c) Deduce that for such a linear regression model, with  $k$  covariates on board, we have

$$\begin{aligned} \text{aic}_k &= 2(k+1) + 2n \log \hat{\sigma}_k + n + n \log(2\pi), \\ \text{bic}_k &= (k+1) \log n + 2n \log \hat{\sigma}_k + n + n \log(2\pi). \end{aligned}$$

By omitting factors not depending on the different models, show then, that doing well for AIC is the same as having a small  $k + n \log \hat{\sigma}_k$ , or  $2k + n \log \hat{\sigma}_k^2$ ; and that doing well for BIC is the same as having a small  $k \log n + 2n \log \hat{\sigma}_k$ , or  $k \log n + n \log \hat{\sigma}_k^2$ .

(d) Above we've derived AIC and BIC formulae from their general definitions. Check that 'doing well with AIC' is equivalent to what we find by using the book's AIC formula, and the same with BIC, even though the book's AIC and BIC formulae are not fully identical to the  $\text{aic}_k$  and  $\text{bic}_k$  above. – The general AIC and BIC formulae, as laid out in this exercise, are part of the course's active curriculum, and can specifically be used when comparing different time series models for the same dataset.

### 11. Where are the snows of yesteryear?

Figure 0.2 is a dramatic one, for at least my segment of civilisation. It gives the number of skiing days at the location Bjørnholt in Nordmarka, a skiing hour away from tram stations Voksenkollen and Frognersteteren, with skiing day defined as there being at least 25 cm snow on the ground. The linear trend is the estimated regression line using what we call Model 2 below, drastically indicating that the climate has consequences also for the skiing days of the Oslo people. See Heger (2011) and Cunen, Hermansen, and Hjort (2019) for further discussion and details.

The time series goes from 1897 to 2015, but, crucially, there's a big hole in the series, with no data recorded from 1938 to 1954. This spells trouble for classes of traditional time series models, since they prefer data to be equidistant. We may still model and analyse the data, using autocorrelation functions, etc., though.

- (a) Let for convenience  $z_t = \text{year} - 1896$ , so that these start out like  $1, 2, 3, \dots$ , and let  $x_t$  be the skiing days number for year  $t$ , if recorded. Fit first Model 0 and Model 1, using ordinary linear regression, ignoring time dependence. Model 0 takes  $x_t = \beta_0 + \varepsilon_{0,t}$ , with the  $\varepsilon_{0,t}$  i.i.d.  $N(0, \sigma_0^2)$ , i.e. assumes a constant stationary level. Model 1 takes  $x_t = \beta_0 + \beta_1 z_t + \varepsilon_{1,t}$ , with the  $\varepsilon_{1,t}$  i.i.d.  $N(0, \sigma_1^2)$ . Give a 95 percent confidence interval for  $\beta$ , and give an interpretation of this negative trend coefficient. Also carry out AIC analysis. You should find log-likelihood maxima  $\ell_{0,\max} = -519.479$  and  $\ell_{1,\max} = -512.167$ .
- (b) For Model 1, compute and inspect the estimated residuals,  $r_t = \{x_t - \hat{m}_1(t)\}/\hat{\sigma}_2$ , where  $\hat{m}_2(t)$  is the estimated trend under Model 1. Check in particular the acf, and comment.
- (c) Then go to Model 2, which includes autocorrelation. We take

$$x_t = \beta_0 + \beta_1 z_t + \sigma \varepsilon_t \quad \text{for } t = 1, 2, 3, \dots, \quad \text{with } \text{corr}(\varepsilon_s, \varepsilon_t) = \rho^{|s-t|}.$$

So  $\rho$  is the correlation for skiing days numbers for consecutive years;  $\rho^2$  for times two years apart, etc. We also take the  $\varepsilon_t$  to be jointly multinormal with mean zero and variance one. Show that this entails

$$x \sim N_n(\xi, \sigma^2 A_\rho),$$

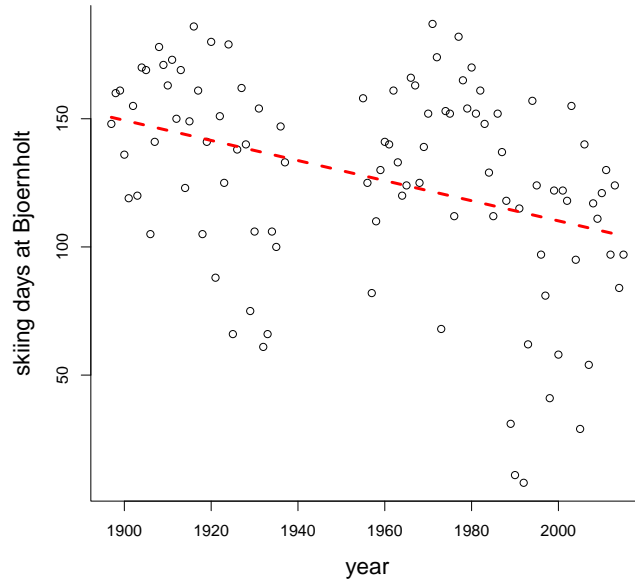


Figure 0.2: The number of skiing days per year, at the location Bjørnholt in Nordmarka, from 1897 to 2015, though with a gap in the series, with no records from 1938 to 1954. The red line is the estimated regression from the four-parameter Model 2.

where  $\xi_t = \beta_0 + \beta_1 z_t$ , and  $A_\rho$  is the  $n \times n$  matrix with 1 on the diagonal and  $\rho^{d_{i,j}}$  in position  $(i, j)$ , with  $d_{i,j}$  the time difference. Note that this  $A_\rho$  is well-defined in spite of the gap in the time series. We have  $n = 102$ , the number of observations.

(d) Show that the log-likelihood function can be written

$$\ell(\beta_0, \beta_1, \sigma, \rho) = -n \log \sigma - \frac{1}{2} \log(\det(A_\rho)) - \frac{1}{2} (x - m_t)^t A_\rho^{-1} (x - m_t) / \sigma^2 - \frac{1}{2} n \log(2\pi),$$

where  $m_t = \beta_0 + \beta_1 z_t$ . It is numerically a bit troublesome to maximise this here (also since we cannot utilise simplifying formula for the inverse and determinant of  $A_\rho$ , due to the gap in the data, which means data not being equidistant). It is practical to compute and display the log-likelihood profile function instead:

$$\ell_{\text{prof}}(\rho) = \max\{\ell(\beta_0, \beta_1, \sigma, \rho) : \text{all } \beta_0, \beta_1, \sigma\} = \ell(\hat{\beta}_0(\rho), \hat{\beta}_1(\rho), \hat{\sigma}(\rho), \rho).$$

Try to reproduce Figure 0.3.

- (e) In particular, by carrying out these computations, involving maximising over parameters  $(\beta_0, \beta_1, \sigma)$  for each  $\rho$ , you should find that the ML estimate for  $\rho$  is  $\hat{\rho} = 0.208$ , and that  $\ell_{2,\text{max}} = -509.983$ . Carry out AIC analysis for comparing Models 0, 1, 2.
- (f) Given Model 2, predict the number of skiing days in 2013, given the data collected up to 2012, and give an approximate 90 percent confidence interval. Do this exercise also trusting Model 1; compare, and discuss.
- (g) Try out one or two more models for these data.

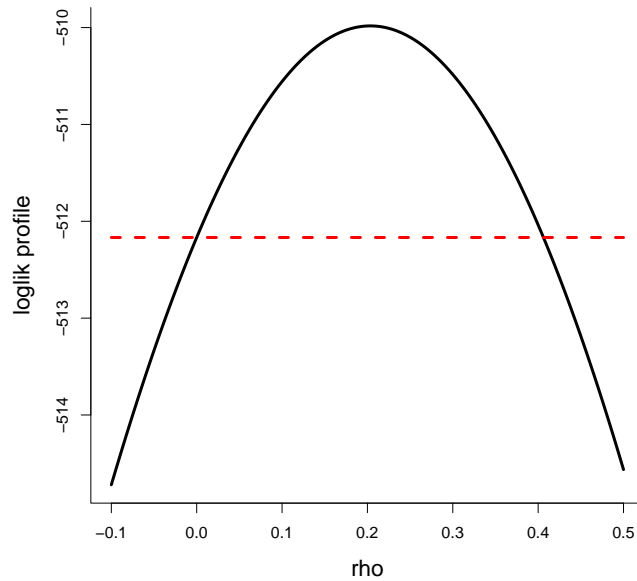


Figure 0.3: The log-likelihood profile function  $\ell_{\text{prof}}(\rho)$ , for the Bjørnholt data, for the four-parameter model with linear trend, a constant  $\sigma$ , and correlation function  $\rho^{|s-t|}$ , for pairs of data with interdistance  $|t-s|$ . The horizontal dashed line indicates the level  $\ell_{1,\text{max}}$  obtained for the submodel of independence, where  $\rho = 0$ .

## 12. Estimating the three parameters in stationary AR(1)

Consider the stationary Gaussian AR(1) model, with

$$x_t = \mu + \sigma \varepsilon_t \quad \text{for } t = 1, \dots, n,$$

where the  $\varepsilon_t$  are standard normal, but correlated with  $\text{corr}(\varepsilon_s, \varepsilon_t) = \rho^{|t-s|}$ .

- Take  $n = 100$ ,  $\mu = 0$ ,  $\sigma = 1$ ,  $\rho = 0.555$  in your computer, and simulate a dataset from this model. Use results and insights from Exercise 2 to do this. – There are also other general simulation schemes, for simulating from a general multinormal distribution, which I will briefly come back to in my teaching. You may also use `library(MASS)` and then use `mvrnorm`.
- Then estimate  $(\mu, \sigma, \rho)$  from the data you've created, using ML, maximum likelihood. You may do this via the log-likelihood profile function  $\ell_{\text{prof}}(\rho)$ ; see earlier R scripts from Nils of this type.
- Compare your  $\hat{\rho}_{\text{ml}}$  with two other estimators, both of the form

$$\rho^* = \frac{1}{n} \sum_{t=2}^n \frac{x_{t-1} - \hat{\mu}}{\hat{\sigma}} \frac{x_t - \hat{\mu}}{\hat{\sigma}}.$$

Version (i) uses the simple classic estimates for  $(\mu, \sigma)$ , trusting independence; version (ii) uses the more elaborate  $(\hat{\mu}_{\text{ml}}, \hat{\sigma}_{\text{ml}})$ , from ML in the three-parameter model.

- Construct both an estimator and an (approximate) 90 percent confidence interval for the next point, i.e.  $x_{n+1}$ , based on having observed the first  $n$  datapoints.

- (e) When your code works, for a single simulated dataset, to a loop on top, to simulate the full thing e.g. `sim = 1000` times, to learn how the estimators perform. Where are the differences in performance, for the three estimators of  $\rho$ ? Do your 90 percent confidence intervals manage to capture  $x_{n+1}$  about 90 percent of the time?

### 13. Estimating cycle length

A model used a few places in the book for capturing cyclic behaviour is

$$x_t = a \cos(2\pi t/\omega + \phi) + \varepsilon_t \quad \text{for } t = 1, \dots, n,$$

with different natural assumptions for the the  $\varepsilon_t$ . We will do fuller time series versions of this later, but on this occasion we make life simple by taking the  $\varepsilon_t$  i.i.d.  $N(0, \sigma^2)$ . The model has three parameters for the mean, including the crucial cycle length parameter  $\omega$ , and so far one for the variability. It turns out that estimation of  $\omega$  can be carried out with remarkable precision.

- (a) Simulate such a dataset, for say  $n = 200$ , and with values you choose yourself for  $a_{\text{true}}$ ,  $\phi_{\text{true}}$ ,  $\sigma_{\text{true}}$ , and take  $\omega_{\text{true}} = 7$  (think about seven days a week). First take  $\omega_{\text{true}}$  to be known, and estimate the parameters  $a, \phi$ . You may use the trick of Example 2.10 in the book, to convert the problem to linear regression in  $\cos(2\pi t/\omega_{\text{true}})$  and  $\sin(2\pi t/\omega_{\text{true}})$ ; or why not attack the problem directly, minimising

$$Q_n(a, \phi) = \sum_{t=1}^n \{x_t - a \cos(2\pi t/\omega + \phi)\}^2$$

by throwing it to the clever `nlm` minimisation algorithm. Check that these two computational methods give the same answers.

- (b) For the case of  $(a, \phi)$  known, making cycle length  $\omega$  the single unknown parameter in the mean function, let  $\hat{\omega}$  be the minimiser of  $Q_n(\omega) = \sum_{t=1}^n \{x_t - a \cos(2\pi t/\omega + \phi)\}^2$ . Attempt to prove the miraculous result that

$$n^{3/2}(\hat{\omega} - \omega) \rightarrow_d \frac{\sqrt{6}}{2\pi} \omega^2 \frac{\sigma}{a} N(0, 1).$$

This means that the variance of  $\hat{\omega}$  is surprisingly small. I haven't seen this in the literature, and I might write up a paper about such themes.

- (c) Then estimate also  $\omega$  from your simulated dataset, using the profiled log-likelihood function

$$\ell_{\text{prof}}(\omega) = \max\{\ell(a, \phi, \sigma, \rho) : \text{over all } a, \phi, \sigma\}.$$

You might find that the cycle length  $\omega$  is rather sharply estimated, with good precision.

- (d) How can you set approximate 90 percent confidence intervals for the parameters? Play with your code a bit, setting different values for  $(a, \phi, \omega, \sigma)$ , and also  $n$ . Check how your estimates work.
- (e) Try to extend your model and estimation schemes to the case where there also is an auto-correlation parameter.

#### 14. Annual mean temperature at New Haven

Find the dataset `nhtemp` of annual average temperature at New Haven, from 1912 to 1971, and then please translate these to the Celsius scale; this is  $x_t = (x_{t,F} - 32)/(5(9))$ , I think. Writing  $t = 1, 2, \dots, 60$  for these years, let for numerical convenience  $z_t = yr_t - \bar{yr}$ , travelling through  $1912 - \bar{yr}, \dots, 1971 - \bar{yr}$ , with  $\bar{yr}$  the average of these  $n = 60$  year. It is easiest and best to write down and work with models in terms of such a  $z_t$ , rather than with the high numbers 1912-1971.

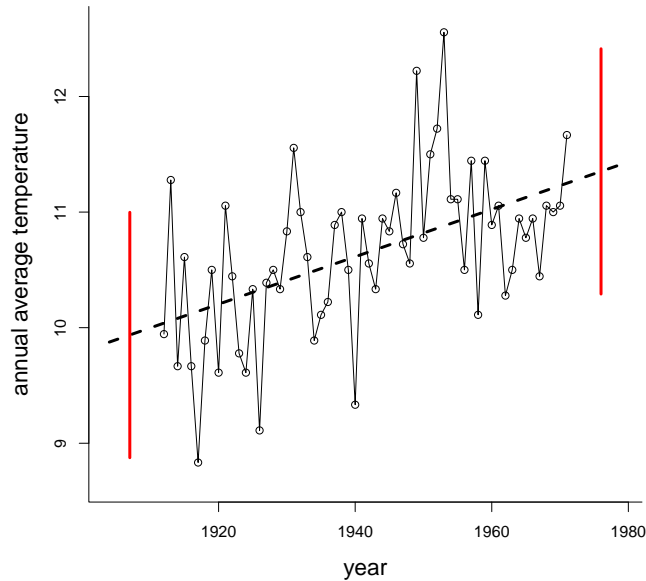


Figure 0.4: Annual average temperature at New Haven, 1912 to 1971, with prediction and 90 percent confidence for 5 years before and 5 years after.

- Fit first the two simple classical models  $M_0$ , where  $x_t = \beta_0 + \sigma\varepsilon_t$ , and  $M_1$ , where  $x_t = \beta_0 + \beta_1 z_t + \sigma\varepsilon_t$ , where the  $\varepsilon$  are i.i.d. and standard normal. Trusting model  $M_1$ , what is a confidence interval for  $\beta_1$ ? Give an interpretation of this result.
- Then go to model  $M_2$ , which takes  $x_t = \beta_0 + \beta_1 z_t + \sigma\varepsilon_t$ , now with correlations  $\text{corr}(\varepsilon_s, \varepsilon_t) = \rho^{s-t}$ . Fit this four-parameter model to data. Find a confidence interval for  $\rho$ . Find also AIC scores  $\text{aic}_0, \text{aic}_1, \text{aic}_2$ , and comment.
- For fun & profit, investigate one more model, namely the four-parameter model  $x_t = \beta_0 + \beta_1 z_t + \sigma\varepsilon_t$ , where the  $\varepsilon_t$  now are taken i.i.d. from the  $t_\nu$ , the t distribution with degrees of freedom  $\nu$ . The point is that this allows fatter tails than the normal, with outcomes say 2 standard deviations away from the mean less strange than under normal conditions. Fit the model, and find the AIC score.
- Go 5 years into the future, and also 5 years into the past, to provide both a point estimate and a 90 percent prediction interval for the average temperature at New Hanven, for the years 1907 and 1976. Try to construct a version of Figure 0.4. Play with your code to learn a bit more.

### 15. The AR(p) model: definition, presentation, backshift polynomials, estimation

Consider a stationary zero-mean time series model for  $x_1, \dots, x_n$ , with the autoregressive property that

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + w_t \quad \text{for } t = 4, 5, \dots, n. \quad (0.4)$$

Here the  $w_t$  are seen as i.i.d. white noise terms, with mean zero and variance  $\sigma_w^2$ . We call this an AR(3) model. The parameters in play are  $\phi = (\phi_1, \phi_2, \phi_3)^t$  for the autoregressive structure and  $\sigma_w^2$  for the variability level. It will be clear how to generalise to any AR(p) model, with  $p \geq 1$ .

- (a) It is useful to simulate a few time series realisations from such a model, with different sets of  $\phi$  parameters. One way is as follows: construct a longer chain, say  $x_{-50}, x_{-49}, \dots, x_0, x_1, \dots, x_n$ , with an extra burn-in phase, starting at perhaps even strange values, and then letting (0.4) decide on the rest. After this, trusting that the chain has reached its equilibrium after the burn-in, discard this burn-in portion, and consider  $(x_1, \dots, x_n)$  a sample from the AR(3). – Now do this, with say  $n = 250$ , and these two choices for  $\phi$ : (i) (0.60, 0.30, 0.05); (ii) (0.30, 0.40, 0.50).
- (b) You are supposed to learn from simple simulations above that not all  $\phi_1, \phi_2, \phi_3$  are OK, but OK means stationarity and stability; not-OK might mean explosions and eruptions. We shall find a clear criterion for OK-ness below. Start by showing that

$$\phi(B)x_t = (1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3)x_t = w_t \quad \text{for all } t,$$

where  $B$  is the backshift operator, with  $Bx_t = x_{t-1}$ ,  $B^2 x_t = B(x_{t-1}) = x_{t-2}$ , etc.

- (c) Then we allude to a general stationarity lemma (not made precise here, and not proven): for any zero-mean stationary sequence, with finite variance, it can be presented in the form

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} = \psi_0 w_t + \psi_1 w_{t-1} + \psi_2 w_{t-2} + \dots,$$

where the  $w_t$  are i.i.d. zero mean white noise variables with some variance  $\sigma_w^2$ . Note that this somehow requires an infinite past; that  $x_t$  is a function of all these  $w_s$  for  $s \leq t$ ; but that  $x_t$  is not allowed to depend on the future. – Show that  $x_t$  has variance  $(\sum_{j=0}^{\infty} \psi_j^2) \sigma_w^2$ , so convergence of this series is assumed. Find also expressions for  $\gamma(1) = \text{cov}(x_t, x_{t+1})$  and  $\gamma(2) = \text{cov}(x_t, x_{t+2})$ .

- (d) Show that  $\psi(B)w_t = x_t$ , for all  $t$ , where  $\psi(B) = \psi_0 + \psi_1 B + \psi_2 B^2 + \dots$  is the psi representation infinite-degree polynomial. Show from this that

$$\phi(B)\psi(B)w_t = w_t, \quad \psi(B)\phi(B)x_t = x_t,$$

for all  $t$ . Consider  $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \phi_3 z^3$  and  $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j$ , and note that the natural domain for  $z$ , inside which there is convergence of the power series, is  $|z| \leq 1$ , the unit circle in the complex plane. Argue that we need to have

$$\psi(z)\phi(z) = 1 \quad \text{for all } z \text{ with } |z| \leq 1.$$

A criterion for OK-ness is clearly that  $\phi(z) \neq 0$  for  $|z| \leq 1$ ; all the roots, of this 3rd order polynomial, need to lie outside. This is actually a necessary and sufficient condition for (0.4) to determine a well-defined stationary mean-zero process.

- (e) In principle, the  $\phi$  determines all the  $\psi_j$ , via the equation above. At least for  $z$  small, show that

$$\begin{aligned}\psi(z) &= \frac{1}{1 - \phi_1 z - \phi_2 z^2 - \phi_3 z^3} \\ &= 1 + (\phi_1 z + \phi_2 z^2 + \phi_3 z^3) + (\phi_1 z + \phi_2 z^2 + \phi_3 z^3)^2 + (\phi_1 z + \phi_2 z^2 + \phi_3 z^3)^3 + \dots \\ &= 1 + \phi_1 z + (\phi_2 + \phi_1^2) z^2 + (\phi_3 + 2\phi_1 \phi_2 + \phi_1^3) z^3 + \dots.\end{aligned}$$

Equating coefficients, perhaps aided by computer algebra code, will then give us all  $\psi_j$ , from  $\phi_1, \phi_2, \phi_3$ .

- (f) Let as elsewhere in the course  $\gamma(h) = \text{cov}(x_t, x_{t+h})$ , with correlation  $\rho(h) = \text{corr}(x_t, x_{t+h})$ . Multiply the start equation  $x_t - \phi_1 x_{t-1} - \phi_2 x_{t-2} - \phi_3 x_{t-3} - w_t = 0$  with  $x_{t-h}$ , to get

$$x_{t-h} x_t - \phi_1 x_{t-h} x_{t-1} - \phi_2 x_{t-h} x_{t-2} - \phi_3 x_{t-h} x_{t-3} - x_{t-h} w_t = 0.$$

First, for  $h = 0$ , deduce that

$$\gamma(0) - \phi_1 \gamma(1) - \phi_2 \gamma(2) - \phi_3 \gamma(3) = \sigma_w^2,$$

important in its own right; if we manage to estimate the  $\phi$ , via the empirical  $\gamma(h)$ , we also manage to estimate  $\sigma_w$ . Secondly, show that

$$\begin{aligned}\gamma(h) - \phi_1 \gamma(h-1) - \phi_2 \gamma(h-2) - \phi_3 \gamma(h-3) &= 0, \\ \rho(h) - \phi_1 \rho(h-1) - \phi_2 \rho(h-2) - \phi_3 \rho(h-3) &= 0,\end{aligned}$$

valid for  $h = 1, 2, \dots$ . These are called the Yule–Walker equations (and though modern in outlook and use, they’re astoundingly old, in essence from these two scholars’ papers from 1927 and 1931). From this find

$$\begin{aligned}\rho(1) - \phi_1 \rho(0) - \phi_2 \rho(1) - \phi_3 \rho(2) &= 0, \\ \rho(2) - \phi_1 \rho(1) - \phi_2 \rho(0) - \phi_3 \rho(1) &= 0.\end{aligned}$$

and find from these equations  $\rho(-2), \rho(-1)$  expressed via the  $\phi$ . We may then use the recurrence relations above to read off, or to find values, for say  $\rho(1), \dots, \rho(100)$ .

- (g) From equations above, show that

$$\begin{aligned}\gamma(0)\phi_1 + \gamma(1)\phi_2 + \gamma(2)\phi_3 &= \gamma(1), \\ \gamma(1)\phi_1 + \gamma(0)\phi_2 + \gamma(1)\phi_3 &= \gamma(2), \\ \gamma(2)\phi_1 + \gamma(1)\phi_2 + \gamma(0)\phi_3 &= \gamma(3),\end{aligned}$$

or in matrix form

$$\Gamma_3 \phi = \gamma_3, \quad \text{or} \quad \phi = \Gamma_3^{-1} \gamma_3.$$

Here  $\Gamma_3$  is the  $3 \times 3$  matrix with elements  $\gamma(j-k)$  for  $j, k = 1, 2, 3$  and  $\gamma_3$  the  $3 \times 1$  vector with elements  $\gamma(1), \gamma(2), \gamma(3)$ .



(h) Argue that all of this leads to the estimator

$$\hat{\phi} = \begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \hat{\phi}_3 \end{pmatrix} = \begin{pmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \hat{\gamma}(2) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \hat{\gamma}(1) \\ \hat{\gamma}(2) & \hat{\gamma}(1) & \hat{\gamma}(0) \end{pmatrix}^{-1} \begin{pmatrix} \hat{\gamma}(1) \\ \hat{\gamma}(2) \\ \hat{\gamma}(3) \end{pmatrix},$$

where

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

The estimation for the AR(3) process is completed by setting

$$\hat{\sigma}_w^2 = \hat{\gamma}(0) - \hat{\phi}_1 \hat{\gamma}(1) - \hat{\phi}_2 \hat{\gamma}(2) - \hat{\phi}_3 \hat{\gamma}(3).$$

(i) (xx briefly. using appendix. need  $\sqrt{n}(\hat{\gamma} - \gamma)$ . then read off  $\sqrt{n}(\hat{\phi} - \phi)$ . also need  $\sigma_w^2$ . xx)

### 16. Simulating AR(p) processes and estimating their parameters

Set up a simulation scheme, to create  $x_1, \dots, x_n$  of length e.g.  $n = 250$ , from a zero-mean AR(3) model, where you choose your AR parameters  $\phi_j$  as you wish. Make sure that the AR polynomial  $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \phi_3 z^3$  is not touching zero for the  $|z| \leq 1$  unit circle, however, to ensure stationarity (and so-called causality). Do the simulation by starting ‘somewhere in the past’, with a burn-in phase, then throw away the burn-in part afterwards.

For your simulated chain of  $x_t$ , estimate the  $\phi$  parameters and also the noise level  $\sigma_w$ , using the methods of Exercise 15. Play with your code a bit, to learn how well the estimators work.

### 17. The MA(q) time series model

The MA(q) model for a zero-mean time series holds that

$$x_t = w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q},$$

in terms of unobserved i.i.d. variables  $w_t$  with mean zero and finite variance, say  $\sigma_w^2$ . One sometimes writes the first term as  $\theta_0 w_t$ , for notational symmetry, with  $\theta_0 = 1$ .

(a) Consider the MA(1) model, with  $x_t = w_t + \theta w_{t-1}$ . Show that

$$\gamma(0) = (1 + \theta^2)\sigma_w^2, \quad \gamma(1) = \theta\sigma_w^2, \quad \gamma(2) = 0,$$

indeed with  $\gamma(h) = 0$  for  $h \geq 2$ . Show from this that  $\rho(1) = \theta/(1 + \theta^2)$ , and that  $|\rho(1)| \leq \frac{1}{2}$ .

(b) For the MA(2) model, show that

$$\gamma(0) = (1 + \theta_1^2 + \theta_2^2)\sigma_w^2, \quad \gamma(1) = (\theta_0\theta_1 + \theta_1\theta_2)\sigma_w^2, \quad \gamma(2) = \theta_0\theta_2\sigma_w^2,$$

with  $\gamma(h) = 0$  for  $h \geq 3$ . Also show that

$$\rho(1) = \frac{\theta_0\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2}, \quad \rho(2) = \frac{\theta_0\theta_2}{1 + \theta_1^2 + \theta_2^2},$$

with  $\rho(h) = 0$  for  $h \geq 3$ .

- (c) Assume you actually observe  $x_1, \dots, x_n$ , for a reasonably high  $n$ , and compute its acf. What behaviour would you expect this to have, if the model behind the data is an MA(2)?
- (d) Simulate an MA(2) process  $x_1, \dots, x_n$ , for say  $n = 500$ . with  $\theta_1 = 0.66$  and  $\theta_2 = 0.33$ . Compute the autocorrelations  $\widehat{\rho}(1), \widehat{\rho}(2)$  from the data, and equate these to the population parameters  $\rho(1), \rho(2)$ , to find autocorrelation based estimators  $\widehat{\theta}_1, \widehat{\theta}_2$ . From these also estimate the underlying  $\sigma_w$ . Repeat the experiment say 1000 times, to check the precision of these estimators.

### 17. The ARMA(p,q) time series model

(xx after all of this: put them together, to form an AR(2,2) process, more generally an AR(p,q) process. backshift operator, polynomials, estimation, approximate log-likelihood, AIC. i also find a few real data examples. xx)

### References

- Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging*. Cambridge University Press.
- Cunen, C., Hermansen, G., and Hjort, N.L. (2019). Confidence distributions for change-points and regime shifts. *Journal of Statistical Planning and Inference* 195, 14–34.
- Heger, A. (2011). Jeg og jordkloden. Dagsavisen.  
[aheger.blogspot.com/2007/03/jeg-og-jordkloden.html](http://aheger.blogspot.com/2007/03/jeg-og-jordkloden.html)
- Hjort, N.L. (2022). On estimating the cycle length. Manuscript.
- Schweder, T. and Hjort, N.L. (2008). *Confidence, Likelihood, Probability*. Cambridge University Press.
- Shumway, R.H. and Stoffer, D.S. (2016). *Time Series Analysis and Its Applications* (4th ed.). Springer, Berlin.
- Walker, G. (1931). On Periodicity in Series of Related Terms. *Proceedings of the Royal Society of London, Series A*, Vol. 131, 518–532.
- Yule, G.U. (1927). On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers. *Philosophical Transactions of the Royal Society of London, Series A*, Vol 226, 267–298.