Exercises and Lecture Notes
STK 4060, Spring 2022

Version 0.79, 5-May-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 or an average of correlated variables

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

(a) Show that

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

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

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

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 $x_1, \ldots, x_n$, with autocorrelation function $\rho(h) = \rho^h$ for $h = 1, 2, 3, \ldots$, and with unknown mean μ, variance σ^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 μ, recommended in 99 statistics books, with s_n the empirical standard deviation, what will be its confidence coverage level?

(d) Give estimators for μ, σ, ρ, 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 $x_1, x_2, \ldots, x_n$ as follows, via i.i.d. $\varepsilon_1, \ldots, \varepsilon_n$ being standard normal. Let $x_1 = \varepsilon_1$ and then $x_{t+1} = \rho x_t + \varepsilon_{t+1}$ for $i = t, 2, \ldots$, where ρ 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 ρ.

(b) Write $\mathcal{F}_t$ for all observed history up to and including time point t. Show that $\mathrm{E}\,(x_{t+1}\,|\,\mathcal{F}_t) = \rho x_t$ and $\mathrm{Var}\,(x_{t+1}\,|\,\mathcal{F}_t) = 1$. Deduce also from this that

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

Show that $\mathrm{E}\,x_t = 0$, 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 $\varepsilon_1, \ldots, \varepsilon_t$. Use this to find and explicit distribution of x_t. Also show

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

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

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

etc.

(f) Show that the real acf (the autocorrelation function) becomes $1, \rho, \rho^2, \rho^3, \ldots$,

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

(h) Then generalise to the case of the new contributions having variance σ_w^2, not necessarily equal to 1. In other words, $x_t = \rho x_{t-1} + w_t$ for $t = 1, 2, \ldots$, where the w_t are i.i.d. with mean zero and variance σ_w^2; Show that for the stationary version of this, where the chain has reached its equilibrium, has

$$\gamma(0) = \mathrm{Var}\,x_t = \frac{\sigma_w^2}{1-\rho^2}, \quad \gamma(h) = \mathrm{cov}(x_t, x_{t+h}) = \frac{\sigma_w^2}{1-\rho^2}\rho^h,$$

which also leads to correlations $\rho(h) = \rho^h$ for $h = 1, 2, \ldots$.

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 $y_1, \ldots, y_n$ 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.

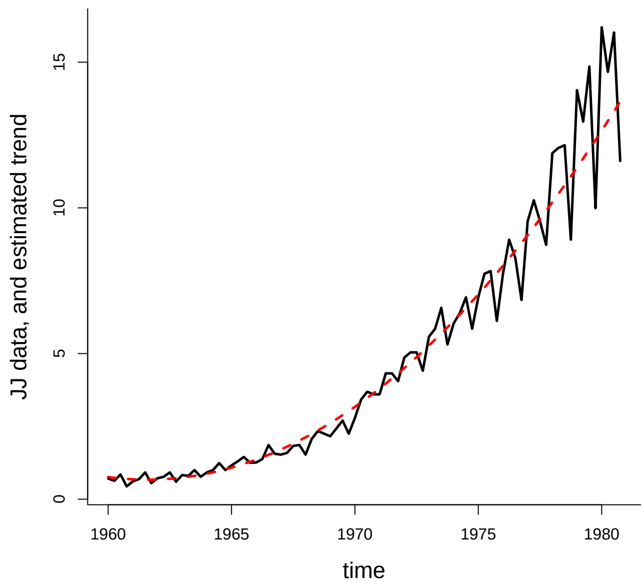

Figure 0.1: The JJ data, with estimated trend, from the five-parameter model.

(a) Write $x_t = t - 1960$, for $t = 1, \ldots, n$. Fit the rather simple classic linear regression model, with $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$, with the ε_t taken i.i.d. $\mathrm{N}(0, \sigma^2)$. Look at the fitted trend $\widehat{m}_1(t) = \widehat{\beta}_0 + \widehat{\beta}_1 x_t$, alongside data, to check that this model is far too simple. For the practice, check also the residuals $r_{1,t} = y_t - \widehat{m}_1(t)$; 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 ε_t taken i.i.d. from the $\mathrm{N}(0, \sigma^2)$. Plot the estimated trend $\widehat{m}_2(t) = \widehat{\beta}_0 + \widehat{\beta}_1 x_t + \widehat{\beta}_2 x_t^2$ alongside data, examine the residuals $r_{2,t} = y_t - \widehat{m}_2(t)$, 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

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

and with the ε_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

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

in terms of the full parameter vector θ.

(d) Find the maximum likelihood (ML) estimates, say $\widehat{\theta}_{\rm ml}$, by numerically maximising the log-likelihood function. Compute also approximate standard errors, for the five parameter estimates, via the general normal approximation theorem for parametric models,

$$\widehat{\theta}_{\rm ml} \approx_d {\rm N}_p(\theta, \widehat{\Sigma}), \quad \text{with } \widehat{\Sigma} = \widehat{J}^{-1}. \tag{0.1}$$

Here $\widehat{J} = -\partial^2 \ell(\widehat{\theta}_{\rm ml})/\partial\theta\partial\theta^{\rm t}$, the Hesse matrix of second order derivatives, computed at the 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 a certain parameter, day γ, which is a function $\gamma = g(\theta)$ of the model parameters, then (i) the ML estimator is $\widehat{\gamma}_{\rm ml} = g(\widehat{\theta}_{\rm ml})$, i.e. via simple plug-in; and (ii) it is approximately a normal, with

$$\widehat{\gamma}_{\rm ml} \approx_d {\rm N}(\gamma, \widehat{\tau}^2),$$

with $\widehat{\tau}^2 = \widehat{c}^{\rm t}\widehat{\Sigma}\widehat{c}$, where $\widehat{c} = \partial g(\widehat{\theta}_{\rm ml})/\partial\theta$ 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 γ, 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 $x_1, x_2, \ldots$ 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 $x_a, \ldots, x_b$.

(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 $A_n \to_{\rm pr} 0$, $B_n \to_{\rm pr} 0$, with '$\to_{\rm pr}$' denoting convergence in probability: $\Pr(|B_n| \geq \varepsilon) \to 0$ for each small ε.

(d) Since A_n is a sum of variables with the same distribution, with mean zero, and $\mathrm{Var}\, A_n \doteq 1/n$, it is natural to expect limiting normality, i.e. $\sqrt{n} A_n \to_d \mathrm{N}(0,1)$. This does *not* follow from the traditional CLTs (central limit theorems), since $x_1 x_2$ is not independent of $x_2 x_3$, etc. Check with the book's Appendix A.2, however, concerning CLTs for m-dependent variables, and verify that indeed $\sqrt{n} A_n \to_d 1$.

(e) From $\sqrt{n} B_n \doteq \sqrt{n} A_n - \sqrt{n} \bar{x}_{1,n}^2$, show that also $\sqrt{n} B_n \to_d \mathrm{N}(0,1)$, 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

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

The main things are that $\widehat{\gamma}(2) \to_{\rm pr} 0$, the true value of $\gamma(2)$ under independence, and that $\sqrt{n} \widehat{\gamma}(2) \to_d \mathrm{N}(0,1)$.

(g) Generalise properly to the result $\sqrt{n} \widehat{\gamma}(h) \to_d \mathrm{N}(0,1)$, for

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

(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 μ and variance σ^2, we have $\sqrt{n} \widehat{\gamma}(h) \to_d \mathrm{N}(0, \sigma^4)$. Finally show that for

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

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

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

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

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

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

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

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

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 ρ.

6. A general stationary normal time series model

Suppose $x_1, \ldots, x_n$ 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 μ, varianec σ^2, and correlation function $\rho(h) = \mathrm{corr}(x_t, x_{t+h})$.

(a) Show that the joint distribution of the full series is a $\mathrm{N}_n(\mu\mathbf{1}, \sigma^2 A)$, where $\mathbf{1} = (1, \ldots, 1)^{\mathrm{t}}$ is the vector of 1s, and A the Â $n \times n$ matrix of $\rho(s-t)$, for $s, t = 1, \ldots, n$; 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

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

wiith θ the parameters involved. If the correlation function is known, then A is known, and θ comprises only μ, σ. For such a case, show that the ML estimators become

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

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 λ, in the correlation function, however, we need also $A = A(\lambda)$, and we have

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

Use the above to find that the log-likelihood profile function, in λ, becomes

$$\ell_{\mathrm{prof}}(\lambda) = -n \log \widehat{\sigma}(\lambda) - \tfrac{1}{2} \log |A(\lambda)| - \tfrac{1}{2} n - \tfrac{1}{2} n \log(2\pi).$$

Here one first computes

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

where

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

(d) Take e.g. $n=100$, generate $x_1,\ldots,x_n$ from the standard normal in your computer, and fit the three-parameter model which has unknown μ, σ, λ, 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 $X=(X_1,\ldots,X_n)$ has the multinormal distribution, with mean ξ and covariance matrix Σ, if its density takes the form

$$f(x)=(2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\{-\tfrac12(x-\xi)^{\rm t}\Sigma^{-1}(x-\xi)\}.$$

We write $X\sim{\rm N}_n(\xi,\Sigma)$ to indicate this; not that the distribution is fully specified by giving the ξ and the Σ.

(a) Check that this becomes the classic formula for ${\rm N}(\xi,\sigma^2)$ in the one-dimensional case. In the general case, show that $Y=AX$ has distribution ${\rm N}_n(A\xi,A\Sigma A^{\rm t})$, 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 Σ_{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:

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

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

$$\begin{pmatrix}X_1\\ X_2\end{pmatrix}\sim{\rm N}_2(\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}1&\rho\\ \rho&1\end{pmatrix}),$$

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 ξ_1,ξ_2 and variances σ_1^2,σ_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

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathrm{N}_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}),$$

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

(b) Suppose $X_1, \ldots, X_n, X_{n+1}$ have a joint multinormal distribution, as for many time series model, and that $x_1, \ldots, x_n$ are observed. Give the distribution for X_{n+1}, given $x_1, \ldots, x_n$. 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 μ, variance σ^2, and correlation function $\rho(h)$ for $h = 1, 2, 3, \ldots$. Again give a prediction, and a prediction interval, for x_{n+1}, assuming that $x_1, \ldots, x_n$ 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 $M_1, \ldots, M_k$, by maximising their likelihoods. Writing $\ell_j(\theta_j)$ for model M_j, we find the ML estimate $\widehat{\theta}_j$ and the maximised log-likelihood value,

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

Then we define

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

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 $2\Delta \approx_d \chi^2_1$, 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 $t = 1, \ldots, n$, with x_t the main outcome (perhaps a time series) and $z_t = (z_{t,1}, \ldots, z_{t,k})^{\rm t}$ a covariate vector of length k. Consider the classical linear regression model, with

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

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

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

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

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

the classic sum of squares.

(b) Show that the ML estimator for β is the least sum of squares estimator, with a formula

$$\widehat\beta = \Sigma_n^{-1} n^{-1}\sum_{t=1}^n z_t x_t = \Big(n^{-1}\sum_{t=1}^n z_t z_t^{\rm t}\Big)^{-1} n^{-1}\sum_{t=1}^n z_t x_t,$$

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

$$\ell_{k,\max} = \max\{\ell_k(\beta,\sigma)\colon \text{all } \beta,\sigma\} = -n\log\widehat\sigma_k - \tfrac12 n - \tfrac12 n\log(2\pi).$$

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

$$\begin{aligned}\mathrm{aic}_k &= 2(k+1) + 2n\log\widehat\sigma_k + n + n\log(2\pi),\\ \mathrm{bic}_k &= (k+1)\log n + 2n\log\widehat\sigma_k + n + n\log(2\pi).\end{aligned}$$

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

(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 aic_k and 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 Frognerseteren, 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 there prefer data to be equidistanced. We may still model and analyse the data, using autocorrelation functions, etc., though.

(a) Let for convenience $z_t = \texttt{year} - 1896$, so that these start out like $1, 2, 3, \ldots$, 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. $\mathrm{N}(0,\sigma_0^2)$, 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. $\mathrm{N}(0,\sigma_1^2)$. Give a 95 percent confidence interval for β, 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, $r_t = \{x_t - \widehat m_1(t)\}/\widehat\sigma_2$, where $\widehat m_2(t)$ 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

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

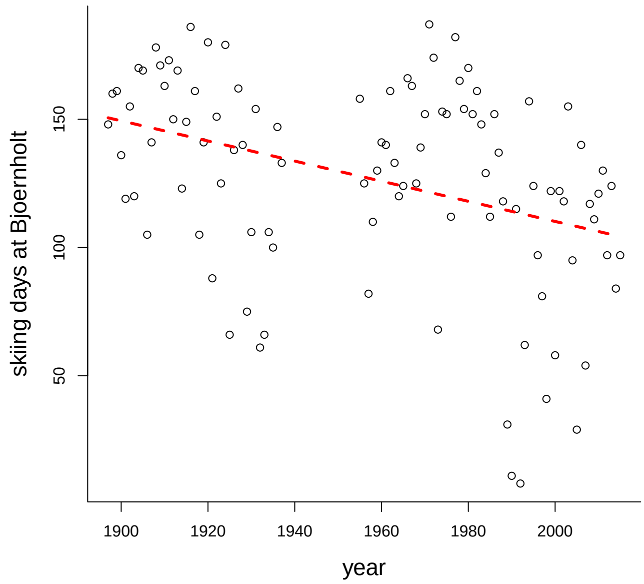

Figure 0.2: The number of skiing days per year, at the location Bjørnholt in Nordmarka, from 1897 to 2015, though with a gap in the series, with no records from 1938 to 1954. The red line is the estimated regression from the four-parameter Model 2.

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

$$x \sim \mathrm{N}_n(\xi, \sigma^2 A_\rho),$$

where $\xi_t = \beta_0 + \beta_1 z_t$, and A_ρ 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_ρ 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

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

where $m_t = \beta_0 + \beta_1 z_t$. It is numerically a bit troublesome to maximise this here (also since we cannot uitilise simplifying formula for the inverse and determinant of A_ρ, 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:

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

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 ρ, you should find that the ML estimate for ρ is $\widehat{\rho} = 0.208$, and that $\ell_{2,\max} = -509.983$. Carry out AIC analysis for comparing Models 0, 1, 2.

(f) Given Model 2, predict the numnber 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.

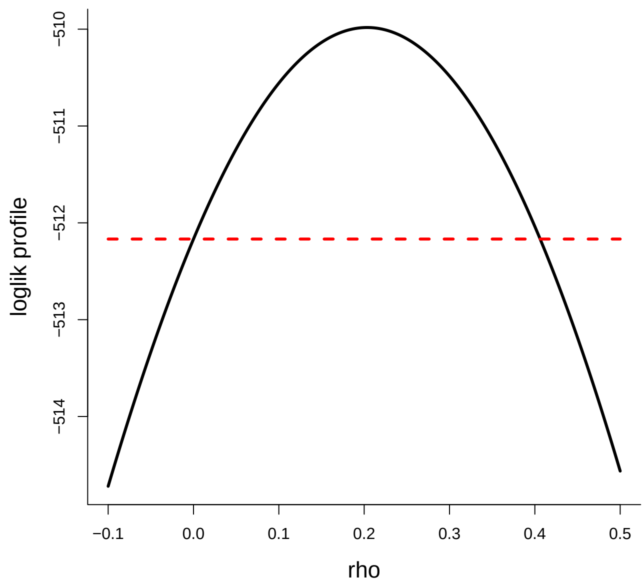

Figure 0.3: The log-likelihood profile funcion $\ell_{\rm prof}(\rho)$, for the Bjørnholt data, for the four-parameter model with linear trend, a constant σ, and correlation function $\rho^{|s-t|}$, for pairs of data with interdistance $|t-s|$. The horizontal dashed line indicates the level $\ell_{1,\max}$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

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

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

(a) 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`.

(b) Then estimate (μ, σ, ρ) from the data you've created, using ML, maximum likelihood. You may do this via the log-likelihood profile function $\ell_{\rm prof}(\rho)$; see earlier R scripts from Nils of this type.

(c) Compare your $\widehat\rho_{\rm ml}$ with two other estimators, both of the form

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

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

(d) 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. Whare are the differences in performance, for the three estimators of ρ? Do your 90 percent confidence intervals manage to capture x_{n+1} anout 90 percent of the time?

13. Estimating cycle length

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

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

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

(a) Simulate such a dataset, for say $n = 200$, and with values you choose yourself for $a_{\rm true}$, $\phi_{\rm true}$, $\sigma_{\rm true}$, and take $\omega_{\rm true} = 7$ (think about seven days a week). First take $\omega_{\rm true}$ to be known, and estimate the parameters a, ϕ. You may use the trick of Example 2.10 in the book, to convert the problem to linear regression in $\cos(2\pi t/\omega_{\rm true})$ and $\sin(2\pi t/\omega_{\rm true})$; 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, ϕ) known, making cycle length ω the single unknown parameter in the mean function, let $\widehat{\phi}$ 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

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

This means that the variance of $\widehat{\omega}$ 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 ω from your simulated dataset, using the profiled log-likelihood function

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

You might find that the cycle length ω 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 autocorrelation parameter.

14. Annual mean temperature at New Haven

Fin 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 $t = 1, 2, \ldots, 60$ for these years, let for numerical convenience $z_t = \mathrm{yr}_t - \bar{\mathrm{yr}}$, travelling through $1912 - \bar{\mathrm{yr}}, \ldots, 1971 - \bar{\mathrm{yr}}$, with $\bar{\mathrm{yr}}$ the averaige 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.

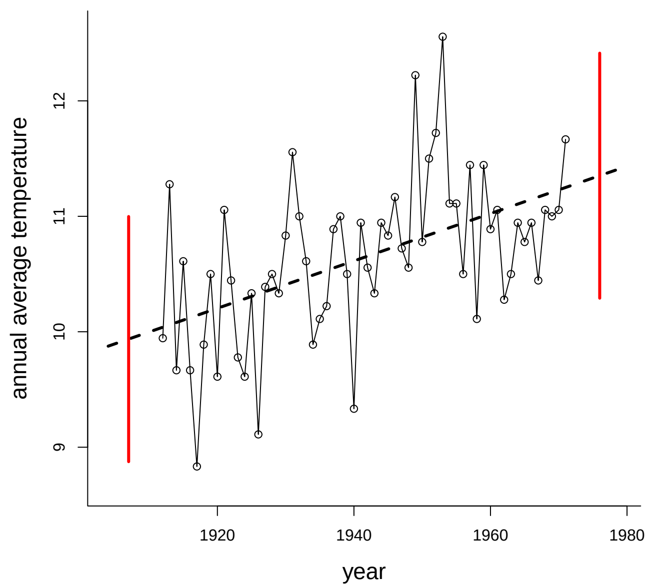

Figure 0.4: Annual average temperature at New Haven, 1912 to 1971, with prediction and 90 percent confidence for 5 years before and 5 years after.

(a) 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 ε are i.i.d. and standard normal. Trusting model M_1, what is a confidence interval for β_1? Give an interpretation of this result.

(b) Then go to model M_2, which takes $x_t = \beta_0 + \beta_1 + \sigma\varepsilon_t$, now with correlations $\mathrm{corr}(\varepsilon_s, \varepsilon_t) = \rho^{s-t}$. Fit this four-parameter model to data. Find a confidence interval for ρ. Find also AIC scores $\mathrm{aic}_0, \mathrm{aic}_1, \mathrm{aic}_2$, and comment.

(c) 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 ε_t now are taken i.i.d. from the t_ν, the t distributiion with degrees of freedom ν. 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.

(d) 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 $x_1, \ldots, x_n$, with the autoregressive property that

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

Here the w_t are seen as i.i.d. white noise terms, with mean zero and variance σ_w^2. We call this an AR(3) model. The parameters in play are $\phi = (\phi_1, \phi_2, \phi_3)^{\rm t}$ for the autoregressive structure and σ_2 for the variability level. It will be clear how to generalise to any AR(p) model, with $p \ge 1$.

(a) It is useful to simulate a few time series realisations from such a model, with different sets of ϕ parameters. One way is as follows: construct a longer chain, say $x_{-50}, x_{-49}, \ldots, x_0, x_1, \ldots, x_n$, 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 $(x_1, \ldots, x_n)$ a sample from the AR(3). – Now do this, with say $n = 250$, and these two choices for ϕ: (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 ϕ_1, ϕ_2, ϕ_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

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

where the w_t are i.i.d. zero mean white noise variables with some variance σ_w^2. Note that this somehow requires an infinite past; that x_t is a function of all these w_s for $s \le t$; 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 $\gamma(1) = \mathrm{cov}(x_t, x_{t+1})$ and $\gamma(2) = \mathrm{cov}(x_t, x_{t+2})$.

(d) Show that $\psi(B)w_t = x_t$, for all t, where $\psi(B) = \psi_0 + \psi_1 B + \psi_2 B^2 + \cdots$ 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 $|z| \le 1$, the unit circle in the complex plane. Argue that we need to have

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

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 ϕ determines all the ψ_j, via the equation above. At least for z small, show that

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

Equating coefficients, perhaps aided by computer algebra code, will then give us all ψ_j, from ϕ_1, ϕ_2, ϕ_3.

(f) Let as elsewhere in the course $\gamma(h) = \mathrm{cov}(x_t, x_{t+h})$, with correlation $\rho(h) = \mathrm{corr}(x_t, x_{t+h})$. 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 ϕ, via the empirical $\gamma(h)$, we also manage to estimate σ_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 $h = 1, 2, \ldots$. 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 ϕ. We may then use the recurrence relations above to read off, or to find values, for say $\rho(1), \ldots, \rho(100)$.

(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 Γ_3 is the 3×3 matrix with elements $\gamma(j-k)$ for $j, k = 1, 2, 3$ and γ_3 the 3×1 vector with elements $\gamma(1), \gamma(2), \gamma(3)$.

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

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

where

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

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

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

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

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

Set up a simulation scheme, to create $x_1, \ldots, x_n$ of length e.g. $n = 250$, from a zero-mean AR(3) model, where you choose your AR parameters ϕ_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 $|z| \le 1$ unit circle, however, to ensure stationarity (and so-called causality) – you may use `polyroot` in R for this, working for any such AR backshift operator polynomial $1 - \phi_1 z - \cdots - \phi_p z^p$. 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 ϕ parameters and also the noise level σ_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

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

in terms of unobserved i.id. variables w_t with mean zero and finite variance, say σ_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 $h \ge 2$. Show from this that $\rho(1) = \theta/(1 + \theta^2)$, and that $|\rho(1)| \le \frac{1}{2}$.

(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 $h \ge 3$. 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 $h \ge 3$.

(c) We saw that for the MA(1), $\rho(1)$ was constrained to be inside $[-\frac{1}{2},\frac{1}{2}]$. For MA(2), check the possible parameter region for $(\rho(1),\rho(2))$. – One easy way to do this is to generate say 1000 values or θ_1,θ_2 from e.g. the standard normal, and plot the resulting $(\rho(1),\rho(2))$. One then discovers the allowed parameter region (in a sense without any mathematical analysis), and may also read off that $|\rho(1)|\le 1/\sqrt{2}$, $|\rho(2)|\le 1/2$.

(d) Assume you actually observe $x_1,\ldots,x_n$, 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)?

(e) Simulate an MA(2) process $x_1,\ldots,x_n$, 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 σ_w. Repeat the experiment say 1000 times, to check the precision of these estimators. – The empirical autocorrelations can be computed from scratch, but are also available via `acf(x)$acf[c(2,3)]`.

(f) Then go on to an MA(3) model, of the type $x_t=w_t+\theta_1w_{t-1}+\theta_2w_{t-2}+\theta_3w_{t-3}$, where the w_t are i.i.d. with zero mean and variance σ_w^2. Show that

$$\rho(1)=\frac{\theta_0\theta_1+\theta_1\theta_2+\theta_2\theta_3}{1+\theta_1^2+\theta_2^2+\theta_3^2},\quad \rho(2)=\frac{\theta_0\theta_2+\theta_1\theta_3}{1+\theta_1^2+\theta_2^2+\theta_3^2},\quad \rho(3)=\frac{\theta_0\theta_3}{1+\theta_1^2+\theta_2^2+\theta_3^2}.$$

Simulate say 10^4 values of the MA(3) parameters, then plot $\rho(1),\rho(2),\rho(3)$, to find wondrous shapes, and read off the implied constraints.

(g) Simulate an AR(3) time series of length $n=500$, with zero mean, θ values $0.55, 0.33, 0.11$, and perhaps $\sigma_w=1$. Then estimate these parameters, by equating the empirical $\widehat\rho(1),\widehat\rho(2),\widehat\rho(3)$ with their theoretical values.

18. Maximum likelihood estimation for the MA(2) model

Consider an MA(2) model for a zero-mean stationary time series, with the representation $x_t=w_t+\theta_1w_{t-1}+\theta_2w_{t-2}$, in terms of w_t being i.i.d. with variance σ_w^2. Above we worked with the correlation fitting estimation method, solving the two equations

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

where $\widehat\rho(1)$ and $\widehat\rho(2)$ are the empirical correlations of orders 1 and 2. After having estimated the θ_j we use $\mathrm{Var}\,x_t=\sigma_w^2(1+\theta_1^2+\theta_2^2)$ to estimate also σ_w. – Now we look into ways of finding the ML estimates. These are expected to have slightly better precision, under model conditions, via general likelihood theory.

(a) Take $x_t=w_t+\theta_1w_{t-1}+\theta_2w_{t-2}$ to be valid for $t\ge3$, supplemented with $x_2=w_2+\theta_1w_1$ and $x_1=w_1$. Show that this may be written $x=Aw$, in linear algebra form, where

$$A=A(\theta_1,\theta_2)=\begin{pmatrix}1&0&0&0&0\ldots\\ \theta_1&1&0&0&0\ldots\\ \theta_2&\theta_1&1&0&0\ldots\\ 0&\theta_2&\theta_1&1&0\ldots\\ 0&0&\theta_2&\theta_1&1\ldots\\ \ldots&&&&\end{pmatrix}$$

Note that A is lower triangular, and show that its determinant is 1.

(b) In addition to assuming that the w_t are i.i.d. with zero mean and variance σ_w^2, assume their distribution is normal. Show that $x \sim \mathrm{N}_n(0, \sigma_w^2 A A^{\mathrm{t}})$, and that the log-likelihood function may be written

$$\ell_n = \ell_n(\theta_1, \theta_2, \sigma_w) = -n \log \sigma_w - \tfrac{1}{2} R_n(\theta_1, \theta_2)/\sigma_w^2 - \tfrac{1}{2} n \log(2\pi),$$

where

$$R_n(\theta_1, \theta_2) = x^{\mathrm{t}}(AA^{\mathrm{t}})^{-1}x = x^{\mathrm{t}}(A^{\mathrm{t}})^{-1}A^{-1}x = zz^{\mathrm{t}} = \sum_{t=1}^n z_t^2,$$

with $z = z(\theta_1, \theta_2) = A^{-1}x$.

(c) Deduce that one way to find the ML estimates is (i) by finding the minimisers $(\widehat\theta_1, \widehat\theta_2)$ of $R_n(\theta_1, \theta_2)$, and then (ii) letting $\widehat\sigma_w^2 = R_{n,\min}/n$.

(d) As in a previous exercise, simulate an MA(2) process, with say $n = 250$, with $(\theta_1, \theta_2) = (0.66, 0.33)$ and $\sigma_w = 1$. Then estimate $(\theta_1, \theta_2, \sigma_w)$, (i) using the correlation fitting method, (ii) using ML, following the lines above.

(e) Do a little simulation experiment, to check the extent to which the ML method beats the correlation fitting method (under model conditions). A very simple seven-minute Nils investigation appears to indicate (i) that differences are not big, but (ii) slightly more noticeable for θ_2 than for θ_1. A bigger investigation would need to look at large-sample theory, and also at different sample sizes and different parts of the parameter domain.

(f) Above we've been in brute force modus, so to speak, using a numerical method with the $n \times n$ matrix A, and needing its inverse to compute $z = A^{-1}x$. These steps might be made more clever and faster, using the structure of A. Briefly look into this. Show that AA^{t} is a band matrix, with elements equal to zero apart from the diagonal and its two diagonal neighbour. Show also that A^{-1} is lower triangular, with a certain structure for its columns.

19. The ARMA(p,q) time series model

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

20. The spectral domain

Here we gently open the door to the spectral of frequency domain for modelling, interpreting, analysing classes of time series.

(a) Show first that

$$x_t = A \cos(2\pi\omega t + \phi) = U_1 \cos(2\pi\omega t) + U_2 \sin(2\pi\omega t),$$

for $U_1 = A\cos\phi$, $U_2 = -A\sin\phi$.

(b) Then do a little transformation-of-variables analysis, going from (U_1, U_2) in the plane to polar coordinates $U_1 = A\cos\phi$, $U_2 = A\sin\phi$. Find A, ϕ expressed in terms of U_1, U_2. First, assume

U_1, U_2 are independent $\mathrm{N}(0, \sigma^2)$, and work out that $A \sim \sigma^2 \chi^2_2$, that $\phi \sim \mathrm{unif}(-\frac{1}{2}\pi, \frac{1}{2}\pi)$, with these two, length and angle, being independent. Show also the converse, that if (A, ϕ) are given these distributions, then indeed U_1, U_2 are i.i.d. $\mathrm{N}(0, \sigma^2)$. It is instructive to 'verify via simulations'.

(c) With a given cycle length parameter ω, and with U_1, U_2 being independent with zero mean and variance σ^2 (perhaps also normal), show that the $x_t = U_1 \cos(2\pi\omega t) + U_2 \sin(2\pi\omega t)$ time series is stationary, with covariance function $\gamma(h) = \sigma^2 \cos(2\pi\omega h)$; in particular, the variance is σ^2.

(d) Then consider the interesting time series

$$x_t = \sum_{k=1}^{q} \{U_{k,1} \cos(2\pi\omega_k t) + U_{k,2} \sin(2\pi\omega_k t)\}, \tag{0.5}$$

for independent pairs of independent zero-mean $U_{k,1}, U_{k,2}$, with variance σ_k^2, and cycle parameters $\omega_1, \ldots, \omega_q$. Show that its covariance function becomes $\gamma(h) = \sum_{k=1}^q \sigma_k^2 \cos(2\pi\omega_k h)$. In particular, show that the variance is $\sum_{k=1}^q \sigma_k^2$, the sum of the individual variances associated with the pair $(U_{k,1}, U_{k,2})$ at frequency w_k.

(e) Now illustrate the above, in your computer, taking $n = 100$, then $\omega_1 = 5/n, \omega_2 = 10/n, \omega_3 = 15/n$, then

$$\begin{aligned}
x_{1,t} &= 1 \cos(2\pi\omega_1 t) + 2 \sin(2\pi\omega_1 t), \\
x_{2,t} &= 3 \cos(2\pi\omega_2 t) + 4 \sin(2\pi\omega_2 t), \\
x_{3,t} &= 5 \cos(2\pi\omega_3 t) + 6 \sin(2\pi\omega_3 t),
\end{aligned}$$

and finally $x_t = x_{1,t} + x_{2,t} + x_{3,t}$. Check the values of $\max(|x_{1,t}|), \max(|x_{2,t}|), \max(|x_{3,t}|)$, and comment. Generalise to the sum of e.g. six such sub-series; the idea here is that rather complicated time series may be well approximated with those definde in point (d).

21. Spectral representation and the periodogram

For simplicity of presentation of what follows, take the time series length n to be odd. For n even, just a few modifications are required; see the book's page 171.

(a) Let $x_1, \ldots, x_n$ be any numbers. By counting unknowns, argue that there must be a representation in the form of

$$x_t = \bar{x} + \sum_{j=1}^{(n-1)/2} \{a_j \cos(2\pi t j/n) + b_j \sin(2\pi t j/n)\}.$$

(b) In fact these $n - 1$ equations with $n - 1$ unknowns can be nicely solved, with the explicit solutions

$$a_j = (2/n) \sum_{t=1}^{n} x_t \cos(2\pi t j/n), \quad b_j = (2/n) \sum_{t=1}^{n} x_t \sin(2\pi t j/n).$$

First verify that this holds, in a simple simulation, where you generate $x_1, \ldots, x_{99}$ from some distribution. Then attempt to prove it.

(c) Some of the mathematical magic here, leading to the solutions and equations above, involve the following. For each statement, it is instructive to 'check it' numerically (try your own n), and then attempt to prove it:

$$(1/n)\sum_{t=1}^{n}\cos^2(2\pi tj/n)=\tfrac{1}{2},\quad (1/n)\sum_{t=1}^{n}\sin^2(2\pi tj/n)=\tfrac{1}{2},$$

as long as $j \neq 0$ and $j \neq n/2$. Next, for $j \neq k$,

$$\sum_{t=1}^{n}\cos(2\pi tj/n)\cos(2\pi tk/n)=0,\quad \sum_{t=1}^{n}\sin(2\pi tj/n)\sin(2\pi tk/n)=0,$$

and $\sum_{t=1}^{n}\cos(2\pi tj/n)\sin(2\pi tk/n)=0$ for all j,k.

(d) The *scaled periodogram* is then defined by $P(j/n) = a_j^2 + b_j^2$. These are also called the fundamental or the Fourier frequencies. – Simulate any time series, of length say $n = 99$, then compute and display the $P(j/n)$. Discuss the relation of $P(j/n)$ to the representation (0.5).

(e) Just a few additional notes on these identities: with $m = (n-1)/2$, so that $2m = n-1$, consider the $(n-1)\times(n-1)$ matrix K, where row t has $K[t,j] = \cos(2\pi tj/n)$ for $j = 1,\ldots,m$ and then $K[t,m+j] = \sin(2\pi tj/n)$ for $j = 1,\ldots,m$. Show that the first equations, in point (a), correspond to

$$K\begin{pmatrix}a\\ b\end{pmatrix}=x-\bar{x},$$

where $a = (a_1,\ldots,a_m)^{\rm t}$ and $b = (b_1,\ldots,b_m)^{\rm t}$. Show that the squared length of each row of K is m, and that $K[t_1,]\cdot K[t_2,] = -\frac{1}{2}$, for $t_1 \neq t_2$.

(f) Then introduce another $(n-1)\times(n-1)$ matrix M, with rows $M[t,j] = (2/n)\cos(2\pi tj/n)$ for $t = 1,\ldots,m$ and $M[m+t,j] = (2/n)\sin(2\pi tj/n)$ for $t = 1,\ldots,m$, for $j = 1,\ldots,n-1$. Show that K at least partly works as an inverse for K, in that MK is a matrix which has zeroes upper right and lower left; is equal to the identity matrix I_m for lower right, i.e. diagonal with elements 1 on the diagonal for this $m\times m$ submatrix; and finally equal to $I_m - 2/n$ for the upper left $m\times m$ submatrix. [xx finish this suitably. with a few more arguments, this gives (a,b) in terms of $x-\bar{x}$. xx]

22. The Discrete Fourier Transform

Related to the (a_j,b_j) and the scaled periodogram $P(j/n) = a_j^2 + b_j^2$ of the previous exercise is the Discrete Fourier Transform, or DFT, discussed here.

(a) As for Exercise 19(e), generate the time series

$$\begin{aligned}x_{1,t} &= 1\ \cos(2\pi\omega_1 t)+2\ \sin(2\pi\omega_1 t),\\ x_{2,t} &= 3\ \cos(2\pi\omega_2 t)+4\ \sin(2\pi\omega_2 t),\\ x_{3,t} &= 5\ \cos(2\pi\omega_3 t)+6\ \sin(2\pi\omega_3 t),\end{aligned}$$

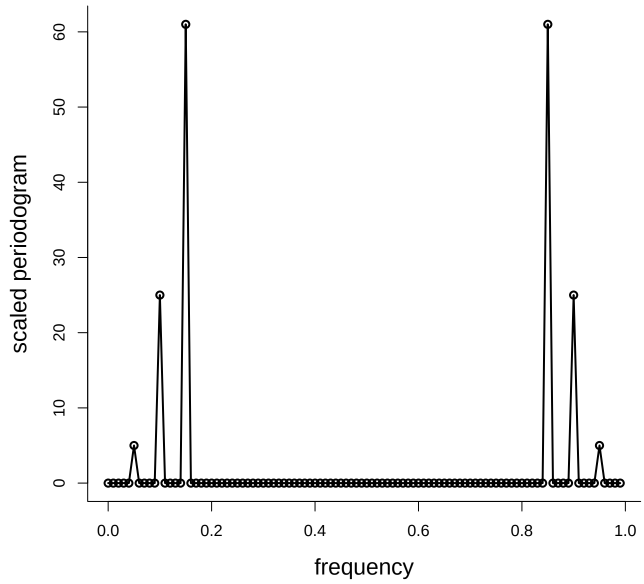

Figure 0.5: Discrete Fourier Transform for data generated according to the model of Exercise 22(a).

of length $n = 100$, with $(\omega_1, \omega_2, \omega_3) = (5/n, 10/n, 15/n)$, followed by $x_t = x_{1,t} + x_{2,t} + x_{3,t}$. Then the DFT is defined as

$$d(j/n) = n^{-1/2} \sum_{t=1}^{n} x_t \exp(-2\pi itj/n) = n^{-1/2} \sum_{t=1}^{n} \{x_t \cos(2\pi tj/n) - x_t \sin(2\pi tj/n)\},$$

for $j = 0, 1, \ldots, n-1$, with $i = \sqrt{-1}$ the famous complex imaginative unit number. Compute the two parts and display them.

(b) Show that

$$|d(j/n)|^2 = \frac{1}{n}\Big\{\sum_{t=1}^{n} x_t \cos(2\pi tj/n)\Big\}^2 + \frac{1}{n}\Big\{\sum_{t=1}^{n} x_t \sin(2\pi tj/n)\Big\}^2.$$

Compute and display these. This is *the periodogram*. Show that $P(j/n) = (4/n)|d(j/n)|^2$.

(c) Construct a version of Figure 0.5. I used

```
P = Mod(2*fft(xx)/nn)^2 ; Fr=0:(nn-1)/nn
plot(Fr,P,type="o",xlab="frequency",ylab="scaled periodogram",lwd=2)
```

(d) Explain why in this case $P(5/n) = 5$, $P(10/n) = 25$, $P(15/n) = 61$, while $P(j/n) = 0$ for the other j.

23. Midnight star gazing

Access the dataset `star` in the `astsa` package, with measurements of the magnitude of a certain star, taken at $n = 600$ consecutive midnights (a hundred years ago; from Whittaker and Robinson, 1923).

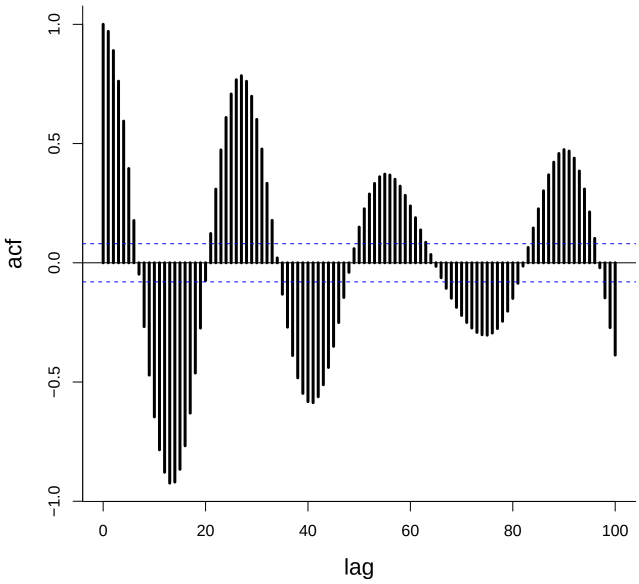

Figure 0.6: Autocorrelation function, for the star gazing dataset, based on measurements of a certain star's magnitude.

(a) Compute the ACF, and produce a version of Figure 0.6. Note its structure and behaviour, certainly different from the typical AR(p) and MA(q).

(b) Compute and display the scaled periodogram; it should have a low number of sharp peaks. Discuss what this might indicate, and fit an appropriate time series model to the data.

24. Spectral representation of an autocovariance function

An important result for the interpretation and analysis of stationary time seires is the following, with further consequences and insights following. If x_t is such a stationary series, with finite $\gamma(h) = \mathrm{cov}(x_t, x_{t+h})$, then there is a *spectral distribution function* F on $[-\frac{1}{2}, \frac{1}{2}]$ such that

$$\gamma(h) = \int_{-1/2}^{1/2} \exp(2\pi i\omega h)\,\mathrm{d}F(\omega) = \int_{-1/2}^{1/2} \cos(2\pi\omega h)\,\mathrm{d}F(\omega)$$

for $h = 0, 1, 2, \ldots$. Details are given in the book's Appendix C.

(a) Choose e.g. $h = 3$. Plot and then integrate the functions $\cos(2\pi\omega h)$ and $\cos^2(2\pi\omega h)$, over $[-\frac{1}{2}, \frac{1}{2}]$, and show that these integrate to 0 and $\frac{1}{2}$ over that interval. Show also that $\int_{-\frac{1}{2}}^{\frac{1}{2}} \cos(2\pi\omega h_1)\cos(2\pi\omega h_2)\,\mathrm{d}w = 0$, for $h_1 \neq h_2$.

(b) For $\mathrm{d}F(\omega)$ insert $f(\omega)\,\mathrm{d}\omega$ above, with

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h)\exp(-2\pi i\omega h) = \gamma(0) + \sum_{h\neq 0}^{\infty} \gamma(h)\exp(-2\pi i\omega h),$$

assuming the series converge absolutely, i.e. that $\sum |\gamma(h)|$ is finite. Show that this works!, i.e. that this $f(\omega)$ leads to the right $\gamma(h)$. – This an instance of general spectral or Fourier analysis, with various methods and results and inversions. This $f(\omega)$ is called the *spectral density* for the time series model.

(c) Show that we also have $f(\omega) = \gamma(0) + 2\sum_{h=1}^{\infty} \gamma(h)\cos(2\pi\omega h)$.

(d) Consider an AR(1) process, with $\rho(h) = \rho^h$ for some $\rho \in (-1,1)$. With $\phi(z) = 1 - \rho z$ the characteristic polynomial (of order 1, in this case), show that

$$|\phi(\exp(-2\pi i\omega)|^2 = 1 - 2\rho\cos(2\pi\omega) + \rho^2.$$

Also, show that

$$f_0(\omega) = \frac{1-\rho^2}{|\phi(\exp(-2\pi i\omega))|^2} = \frac{1-\rho^2}{1 - 2\rho\cos(2\pi\omega) + \rho^2}$$

has the property that

$$\int_{-1/2}^{1/2} \cos(2\phi\omega h) f_0(\omega)\,\mathrm{d}\omega = \rho^h \quad \text{for } h = 0, 1, 2, \ldots.$$

Verify that this is a special case of the book's Property 4.4.

(e) Then consider a general zero-mean AR(1) process, as at the end of Exercise 2, with $x_t = \rho x_{t-1} + w_t$, and the w_t being i.i.d. with zero mean and variance σ_w^2. We know that $\gamma(0) = \sigma_w^2/(1-\rho^2)$. Show that

$$\gamma(h) = \frac{\sigma_w^2}{1-\rho^2}\rho^h = \frac{\sigma_w^2}{1-\rho^2}\int_{-1/2}^{1/2} \exp(2\pi i\omega h)\frac{1-\rho^2}{1-2\rho\cos(2\pi\omega)+\rho^2}\,\mathrm{d}\omega.$$

Show that implies that the spectral density for the AR(1) is

$$f(\omega) = \frac{\sigma_w^2}{|1-\rho\exp(-2\pi i\omega)|^2} = \frac{\sigma_w^2}{1-2\rho\cos(2\pi\omega)+\rho^2}.$$

Check that this is a special case of Property 4.4 in the book.

25. Footnote: probability densities and characteristic functions

It is useful to see just a few facts regarding so-called characteristic functions for probability distributions. There are certain mathematical parallels to the DFT and the inverse DFT, and also to the spectral representation of a covariance function and its inverse. These are all instances of general Fourier Analysis in mathematics. – For a probability density $f(x)$, its characteristic function is

$$\phi(t) = \mathrm{E}\,\exp(itX) = \int \exp(itx) f(x)\,\mathrm{d}x.$$

(a) For the standard normal density, show that $\phi(t) = \exp(-\frac{1}{2}t^2)$. For the general case of $X \sim \mathrm{N}(\xi, \sigma^2)$, show that $\phi(t) = \exp(it\xi - \frac{1}{2}\sigma^2 t^2)$.

(b) Show that when $f(x)$ is symmetric around zero, then $\phi(t)$ is real.

(c) Then there is a nice *inversion theorem*, just as we have such for spectral representations of time series, and for the DFT. A theorem says that if $\phi(t)$ is integrable, then the density f can be found from ϕ, via

$$f(x) = \frac{1}{2\pi}\int \exp(-itx)\phi(t)\,\mathrm{d}t.$$

Show that with $\exp(-\frac{1}{2}t^2)$, one indeed finds $f(x) = (2\pi)^{-1/2}\exp(-\frac{1}{2}x^2)$.

(d) Note that there ought to be cases where f is simple but ϕ is complicated, but also vice versa. A case in point is the characteristic function $\phi(t) = \exp(-|t|^\alpha)$, with $\alpha \in [1, 2]$; here the lower point $\alpha = 1$ corresponds to a Cauchy density and the upper point $\alpha = 2$ to a normal, but anything inside $(1, 2)$ is more difficult. The point is that we nevertheless have an inversion formula, which can be used numerically, etc. Compute and display the density $f_\alpha(x)$ for the distribution with characteristic function $\exp(-|t|^\alpha)$, for $\alpha = 1.0, 1.1, \ldots, 1.9, 2.0$.

26. The DFT and the periodogramme

Consider any sequence $x_1, \ldots, x_n$, and from these define again the Discrete Fourier Transform, now computed at frequencies $\omega_j = j/n$, typically actually for j/n with $j = 0, 1, \ldots, n-1$.

$$d(\omega_j) = (1/\sqrt{n}) \sum_{t=1}^n x_t \exp(-2\pi i \omega_j t).$$

(a) Show that $d(\omega_j) = d_c(\omega_j) - i d_s(\omega_j)$, with the cosine and sine transforms

$$\begin{aligned} d_c(\omega_j) &= (1/\sqrt{n}) \sum_{t=1}^n x_t \cos(2\pi\omega_j t), \\ d_s(\omega_j) &= (1/\sqrt{n}) \sum_{t=1}^n x_t \sin(2\pi\omega_j t). \end{aligned}$$

(b) Create any numbers $x_1, \ldots, x_n$, for say $n = 100$, perhaps simulated from a simple distribution. Then compute and display $d_c(\omega_j)$ and $d_s(\omega_j)$, for $\omega_j = 0/n, \ldots, (n-1)/n$.

(c) For your set of x_t numbers, compute and display also *the periodogramme*, namely $I(\omega_j) = |d(\omega_j)|^2$; here $|z|^2 = z\bar{z} = a^2 + b^2$ for a complex number $z = a + ib$, with complex conjugate $\bar{z} = a - ib$. Thus $|z|^2$ is not the absolute value squared, but the squared modulus, for which one may use `Mod` in R. Show that

$$\begin{aligned} I(\omega_j) &= d_c(\omega_j)^2 + d_s(\omega_j)^2 \\ &= (1/n)\Big\{\sum_{t=1}^n x_t \cos(2\pi\omega_j t)\Big\}^2 + (1/n)\Big\{\sum_{t=1}^n x_t \sin(2\pi\omega_j t)\Big\}^2. \end{aligned}$$

(d) Then comes *the inverse DFT*. For your set of numbers, verify via computations that x_t can be retrieved from the $d(\omega_j)$, via

$$\begin{aligned} x_t &= (1/\sqrt{n}) \sum_{j=0}^{n-1} d(\omega_j) \exp(2\pi i \omega_j t) \\ &= (1/\sqrt{n}) \sum_{j=0}^{n-1} d(\omega_j) \{\cos(2\pi\omega_j t) + i \sin(2\pi\omega_j t)\}. \end{aligned}$$

Note the resemblance to the transform and backtransform for densities and characteristic functions.

(e) Attempt also to prove the inversion formula mathematically. Start from

$$\begin{aligned} x_t^* &= \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} d(\omega_j) \exp(2\pi i \omega_j t) \\ &= \frac{1}{n} \sum_{j=0}^{n-1} \sum_{t'=1}^n x_{t'} \exp(-2\pi i \omega_j t') \exp(2\pi i \omega_j t), \end{aligned}$$

and sort this into $t' = t$ and $t' \neq t$. For z any complex number, different from 1, show

$$1 + z + z^2 + \cdots + z^{n-1} = \frac{1 - z^n}{1 - z}.$$

Use this to prove $\sum_{j=0}^{n-1} \exp(2\pi icj/n) = 0$, for any $c = \pm 1, \pm 2, \ldots$, and show that this leads to $x_t^* = x_t$, thus proving the inverse DFT formula. As a byproduct, congratulations with having proven that

$$\sum_{j=0}^{n-1} \cos(2\pi cj/n) = 0, \quad \sum_{j=0}^{n-1} \sin(2\pi cj/n) = 0,$$

for any non-zero integer c.

(f) Verify, for your $x_1, \ldots, x_n$ numbers, that their periodogramme may be found simply and quickly, using the Fast Fourier Transform, i.e. without doing $d_c(\omega_j)$ and $d_s(\omega_j)$ from scratch: show that

```
peri = Mod( fft(xx) )^2/nn ; Fr = jj/nn
```

gives the right answer, with `jj` the numbers $0, 1, \ldots, n-1$. Use `plot(Fr,peri,type="o")`.

(g) When you have code for the above, experiment with $x_1, \ldots, x_n$ drawn from a few time series models, to check how the periodogramme looks like.

27. The Whittle likelihood

Suppose $x_1, \ldots, x_n$ stem from a stationary, Gaussian time series with zero mean and finite variances.

(a) The joint distribution of the data is then multinormal. Show that its log-density can be written

$$\ell = -\tfrac{1}{2} \log |\Sigma| - \tfrac{1}{2} x^{\rm t} \Sigma^{-1} x - \tfrac{1}{2} n \log(2\pi),$$

where Σ is the $n \times n$ matrix with elements $\gamma(i-j)$. – Using general matrix theory for this particular form of circular variance matrices, along with the spectral represention $\gamma(h) = \int \exp(2\pi i\omega h) f(\omega)\,{\rm d}\omega$, certain mathematical approximations to the eigenvalues and eigenvectors of Σ are given on the book's page 185. This again leads to the very useful approximation

$$\ell^w = -\sum_{0<\omega_j<1/2} \Big\{ \log f(\omega_j) + \frac{I(\omega_j)}{f(\omega_j)} \Big\},$$

with $\omega_j = j/n$.

(b) In particular, in case there is a parametric model $f(\omega, \theta)$ for the spectral density, show that this, almost by definition, leads to the log-likelihood approximation

$$\ell^w(\theta) = -\sum_{0<\omega_j<1/2} \Big\{ \log f(\omega_j, \theta) + \frac{I(\omega_j)}{f(\omega_j, \theta)} \Big\}.$$

This is called the *Whittle log-likelihood* (going all the way back to Peter Whittle's PhD in Uppsala 1951, I believe). Its maximiser $\widehat\theta$ is the the Whittle maximum likelihood estimator, and one may prove, cf. details in the book's Appendix C, that

$$\widehat\theta \approx_d {\rm N}(\theta_0, \widehat J_w^{-1}),$$

with θ_0 being the true paramete values, and where $\widehat{J}_w = -\partial^2\ell^w(\widehat{\theta})/\partial\theta\partial\theta^{\rm t}$ is the Hessian matrix, computed at the Whittle ML position.

28. Whittle ML estimation for MA(q) data

Here we look into estimation of MA(q) process parametes, using the Whittle log-likelihood.

(a) Consider first a simple zero-mean MA(1) procecss, with $x_t = w_t + \theta w_{t-1}$, where the w_t are i.i.d. $\mathrm{N}(0, \sigma_w^2)$. We have seen earlier, but please show it again, that $\gamma(0) = (1+\theta^2)\sigma_w^2$, $\gamma(1) = \theta\sigma_w^2$, with $\gamma(h) = 0$ for $h = 2, 3, \ldots$. Show that

$$f(\omega) = \sigma_w^2\{1 + \theta^2 + 2\theta\cos(2\pi\omega)\}.$$

Generate a dataset of size $n = 100$ from this model, taking $(\sigma_w, \theta) = (0.77, 0.44)$. Take first σ_w to be known, and compute and display the Whittle log-likelihood $\ell^w(\theta)$. Compute the Whittle ML estimat, and repeat the experiment a few times. Then generalise to the case of both parameters unknown.

(b) For the MA(2) model, with $x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2}$, show that

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

From this, show that the spectral density can be written

$$f(\omega) = \sigma_w^2\{1 + \theta_1^2 + \theta_2^2 + 2(\theta_0\theta_1 + \theta_1\theta_2)\cos(2\pi\omega) + 2\theta_0\theta_2\cos(4\pi\omega)\}.$$

Choose parameters $(\sigma_w, \theta_1, \theta_2)$, generate a dataset of size $n = 200$, and estimate the parameters using Whittle ML. Test the hypothesis that $\theta_2 = 0$, i.e. that the data come from the simpler MA(1).

(c) Generalise to MA(3), and further. Try out how the Whittle ML works.

29. Whittle ML estimation for AR(p) data

Here we look into Whittle ML estimation for AR(p) parameters. This can be seen as a viable and practical alternative to a few other methods, including the Yule–Walker equations looked at in e.g. Exercises 15, 16.

(a) Consider first a simple zero-mean AR(1) procecss, with $x_t = \rho x_{t-1} + w_t$, where the w_t are i.i.d. $\mathrm{N}(0, \sigma_w^2)$. We have seen in Exercise 24 that the spectral density can be written

$$f(\omega) = \frac{\sigma_w^2}{1 - 2\cos(2\pi\omega) + \rho^2}.$$

Generate a dataset from such an AR(1) model, compute the periodogram $I(\omega_j)$, for $\omega_j = j/n$, and use the Whittle log-likelihood to estimate the two parameters. Repeat the experiment a few times to see the variability of $(\widehat{\sigma}_w, \widehat{\rho})$. Test $\rho = 0$.

(b) Then consider the AR(2) model, with $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$, assumed to have the roots of $\phi(z) = 1 - \phi_1 z - \phi_2 z^2$ outside the unit circle. Show that the spectral density may be written

$$f(\omega) = \frac{\sigma_w^2}{|1 - \phi_1\exp(-2\pi i\omega) - \phi_2\exp(-4\pi i\omega)|^2},$$

and the the denumerator may be written

$$\begin{aligned} g(\omega) &= |1-\phi_1\cos(2\pi\omega)-\phi_2\cos(4\pi\omega)+i\phi_1\sin(2\pi\omega)+i\phi_2\sin(4\pi\omega)|^2 \\ &= \{1-\phi_2\cos(2\pi\omega)-\phi_2\cos(4\pi\omega)\}^2+\{\phi_1\sin(2\pi\omega)+\phi_2\sin(4\pi\omega)\}^2. \end{aligned}$$

Take $(\sigma_w,\phi_1,\phi_2)=(1.00,0.44,0.11)$, and display the $f(\omega)$ function.

(c) Generate data from the model, with these parameters, and estimate them, using the Whittle ML method.

30. The spectral density with an impulse function

A few useful facts for the spectral density are as follows.

(a) Note, again, that if $z=a+ib$ is a complex number, then $\bar z=a-ib$ is its complex conjugate, and $|z|^2=z\bar z=a^2+b^2$ is its squared modulus. Show that $|\sum_j c_jz_j|^2=\sum_{j,k}c_jc_kz_j\bar z_k$, and that

$$\Big|\sum_j c_j\exp(2\pi i\omega_j)\Big|^2=\sum_{j,k}c_jc_k\exp(2\pi i(\omega_j-\omega_k)).$$

(b) Consider a stationary time series x_t with covariance sequence $\gamma(h)$, which then may be represented as $\int\exp(2\pi i\omega h)\,\mathrm{d}F(\omega)$. Show that

$$\begin{aligned} \mathrm{Var}\Big(\sum_j c_jx_j\Big) &= \sum_{j,k}c_jc_k\int\exp(2\pi i\omega(j-k))\,\mathrm{d}F(\omega) \\ &= \int\Big|\sum c_j\exp(2\pi i\omega j)\Big|^2\,\mathrm{d}F(\omega). \end{aligned}$$

Of course this is nonnegative; a nice mathematical result, going back to Cramér and Wold, and other probabilists in other branches, is that the nonnegativity of all such variances implies the existence of a unique measure F on $[-\pi/2,\pi/2]$ such that $\gamma(h)=\int\exp(2\pi i\omega h)\,\mathrm{d}F(\omega)$ for all h; check the book's Appendix.

(c) Assume x_t has spectral density $f(\omega)$, and consider the linear filter $y_t=\sum a_jx_{t-j}$. Show that this y_t series has covariances

$$\gamma^*(h)=\int\exp(2\pi i\omega h)|A(\omega)|^2f(\omega)\,\mathrm{d}\omega,$$

where $A(\omega)=\sum_j a_j\exp(-2\pi i\omega j)$ is the so-called *impulse function* associated with the linear filter.

(d) To illustrate consider an AR(1) series x_t, with $x_t=\rho x_{t-1}+w_t$, and then let $y_t=0.1\,x_{t-1}+0.8\,x_t+0.1\,x_{t+1}$. Compute and display the $|A(\omega)|$ function, and the spectral density function for y_t.

31. The spectral density for the AR(p) model

The spectral density can be expressed as $f(\omega)=\sum\gamma(h)\exp(-2\pi i\omega h)$, and can hence be computed, with a formula or numerically, if the covariance sequence $\gamma(h)$ is known. This can be done from first principles for an AR(1), since $\gamma(h)=\{\sigma_w^2/(1-\rho^2)\}\rho^h$ is not too complicated, see Exercise xx, but is harder for AR(p) for $p\ge 2$.

(a) Consider first the simple case of a white noise sequence w_t, with zero mean and variance σ_2^2. Show that its spectral density is $f_0(\omega) = \sigma_w^2$, i.e. constant. Then work with $x_t = \sum_j \psi_j w_{t-j}$, where ψ_j for $j = 0, \pm 1, \pm 2, \ldots$ form a sequence with $\sum_j |\psi_j| < \infty$, and show that that process has spectral density

$$f(\omega) = |\psi(\exp(-2\pi i\omega))|^2\sigma_w^2 \quad \text{for } \omega \in [-\tfrac{1}{2}, \tfrac{1}{2}],$$

where $\psi(z) = 1 + \psi_1 z + \psi_2 z^2 + \cdots$.

(b) For an AR(p) process we have seen in Exercise 15 that x_t really can be represented as $\sum_{j=0}^{\infty} \psi_j w_{t-j}$, with the coefficients ψ_j being determined by the equation $\phi(z)\psi(z) = 1$, in which $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ and $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j$. Prove hence that

$$f(\omega) = \frac{\sigma_w^2}{|\phi(\exp(-2\pi i\omega))|^2}.$$

(c) For an AR(1) process, with $x_t = \phi x_{t-1} + w_t$, show, again, that

$$f(\omega) = \frac{\sigma_w^2}{1 - 2\phi\cos(2\pi\omega) + \phi^2}.$$

Draw this curve, over $[-\frac{1}{2}, \frac{1}{2}]$, for a few values of ϕ.

(d) We've seen the form of f for an AR(2) process above. For an AR(3), where $\phi(z) = 1 - \phi_1 z - \phi_2 z - \phi_3 z^3$ is assumed never to be zero insider the unit circle, show that $f(\omega) = \sigma_w^2/g(\omega)$, with

$$\begin{aligned} g(\omega) &= |1 - \phi\exp(-2\pi i\omega) - \phi_2\exp(-4\pi i\omega) - \phi_3\exp(-6\pi i\omega)|^2 \\ &= \{1 - \phi_1\cos(2\pi\omega) - \phi_2\cos(4\pi\omega) - \phi_3\cos(6\pi\omega)\}^2 \\ &\quad + \{\phi_1\sin(2\pi\omega) + \phi_2\sin(4\pi\omega) + \phi_3\sin(6\pi\omega)\}^2 \end{aligned}$$

Compute and plot this function, for a few values of (ϕ_1, ϕ_2, ϕ_3). To check that the values you choose are ok, in the sense of leading to a well-defined stationary process with values not depending on the future, you may use `polyroot` in R, which gives the roots of a given polynomial. For example, `Mod(polyroot(c(1,0.66,0.44,0.88)))` sows that one of the three roots is inside the unit circle, so the model $x_t = 0.66\,x_{t-1} - 0.44\,x_{t-2} - 0.88\,x_{t-3} + w_t$ is not ok.

32. Estimating the spectral density

Suppose $x_1, \ldots, x_n$ are observed, from a stationary time series model, with covariances $\gamma(h)$ and spectral density $f(\omega) = \sum \gamma(h)\exp(-2\pi i\omega h)$. How can we estimate this f from data, without further assumptions?

(a) A simple idea is as follows. From the data, compute $\widehat{\gamma}(h)$, for say $h = 0, 1, \ldots, h_0$, where h_0 is relatively small compared to n; a simple default choice might be $h_0 = [n^{1/3}]$. This is easily done as say `acf$acf[1:(h0+1)]`, multiplied by `var(x)`. From these, form

$$\widehat{f}(\omega) = \sum_{h=-h_0}^{h_0} \widehat{\gamma}(h)\exp(-2\pi i\omega h) = \widehat{\gamma}(0) + 2\sum_{h=1}^{h_0}\cos(2\pi h\omega).$$

Do this, for some simulated data, e.g. an AR(1) or MA(2) process.

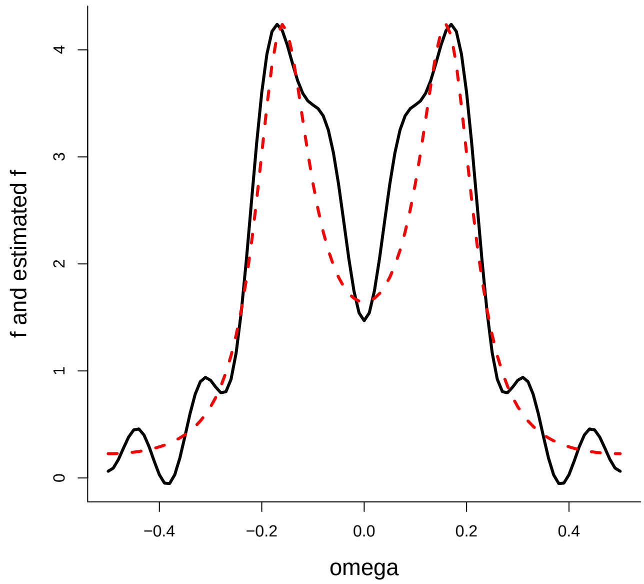

Figure 0.7: The true spectral density f (red, dashed) along with the estimated $\widehat{f}$ (black, full), for an AR(2) case with $n = 500$, $\phi_1 = 0.66$, $\phi_2 = -0.44$.

(b) Construct a version of Figure 0.7, using simulated data from an AR(2), with $n = 500, \phi_1 = 0.66, \phi_2 = -0.44$ and $\sigma_w = 1$. The task is to draw both the real underlying $f(\omega)$ and the estimated $\widehat{f}(\omega)$.

33. The periodogramme as an estimate of the spectral density

It is of interest to see that the squared modulus $I(\omega) = |d(\omega)|^2$ of the DFT can be seen as an estimate of the spectral density $f(\omega) = \sum \gamma(h)\exp(-2\pi i\omega h)$. In particular, the periodogramme $I(\omega_j)$ is an estimate of $f(\omega_j)$ at position $\omega_j = j/n$.

(a) Start from the DFT, and show that

$$d(\omega) = (1/\sqrt{n})\sum_{t=1}^{n} x_t \exp(-2\pi i\omega t) = (1/\sqrt{n})\sum_{t=1}^{n}(x_t - \bar{x})\exp(-2\pi i\omega t)$$

Show then that

$$\begin{aligned} I(\omega) = |d(\omega)|^2 &= \frac{1}{n}\sum_{r,s}(x_r - \bar{x})(x_s - \bar{x})\exp(-2\pi i\omega(r-s)) \\ &= \sum_{|h|\le n-1}\sum_{t\le n-|h|}\frac{1}{n}(x_t - \bar{x})(x_{t+h} - \bar{x})\exp(-2\pi i\omega h), \end{aligned}$$

which is then to be recognised as $\sum_{|h|\le n-1}\widehat{\gamma}(h)\exp(-2\pi i\omega h)$.

(b) Since $\widehat{\gamma}(h)$ is sum over $n-h$ product terms, divided by n, there is a bias involved. Show that

$$\mathrm{E}\, I(\omega) = \sum_{|h|\le n-1}(1 - |h|/n)\gamma(h)\exp(-2\pi i\omega h).$$

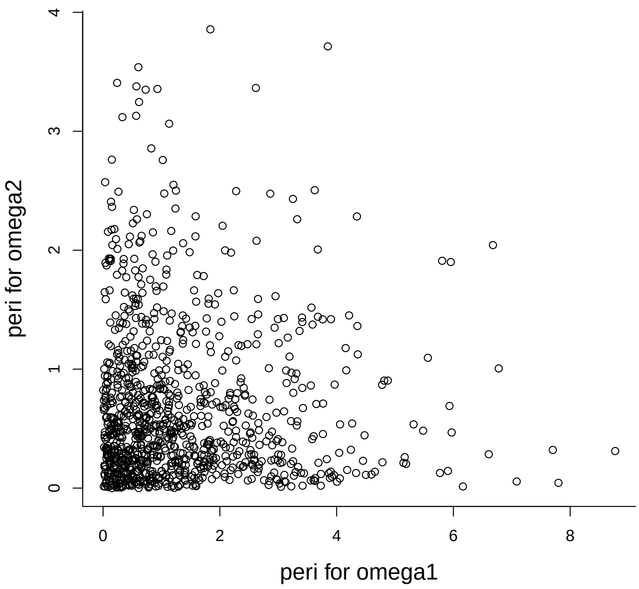

Figure 0.8: With 1000 simulations from the MA(1) model with $\sigma_w = 1$ and $\theta = 0.55$, for $n = 200$, the plot shows the resulting periodogramme pairs $(I(\omega_{j_1}), I(\omega_{j_2}))$, for $j_1/n \doteq 0.25$ and $j_2/n \doteq 0.35$.

(c) Via arguments in the book's Section 4.3, along with finer technical details in Appendix C, one may prove the following, valid for growing sample size n. First, $\mathrm{E}\, I(\omega) \doteq f(\omega)$ and $\mathrm{Var}\, I(\omega) \doteq f(\omega)^2$. Second, for fixed ω, with j chosen so that $j/n \to \omega$,

$$I(\omega_{j,n}) \to_d f(\omega)\, V, \quad \text{with } V \sim \mathrm{Expo}(1).$$

This is convergence in distribution, so

$$\Pr(a \le I(\omega_{j,n})/f(\omega) \le b) \to \Pr(a \le V \le b) = \int_a^b \exp(-v)\,\mathrm{d}v = \exp(-a) - \exp(-b)$$

for all $a < b$. Use this to construct an approximate 90 percent confidence interval for $f(\omega)$.

(d) Third, is is also the case that with $\omega_1 < \cdots < \omega_k$ different fixed values, we not only have

$$I(\omega_{j_1,n}) \to_d f(\omega_1)V_1, \cdots, I(\omega_{j_k,n}) \to_d f(\omega_k)V_k,$$

where $j_1/n \to \omega_1, \ldots, j_k/n \to \omega_k$, where $V_1, \ldots, V_k$ are unit exponentials, but these are fully independent in the limit.

(e) Consider an MA(1) process, $x_t = w_t + \theta w_{t-1}$, for some θ. Show, again, that the spectral density becomes $f(\omega) = \sigma_w^2\{1 + \theta^2 + 2\theta\cos(2\pi\omega)\}$. Then simulate say $n = 200$ data points, and compute the periodogramme at j_1 and j_2, where $j_1/n \doteq 0.25$ and $j_2/n \doteq 0.35$. Do this say 1000 times, and plot the resuling $(I(\omega_{j_1}), I(\omega_{j_2}))$ in a diagram, as with Figure 0.8 (where I've used $\theta = 0.55$). Check that the two histograms look exponential, with means close to $f(\omega_1)$ and $f(\omega_2)$, and that the correlation between them is low.

34. Estimating the cumulative spectral distribution

Consider $x_1, \ldots, x_n$ from a stationary time series model, with spectral density $f(\omega) = \sum_h \gamma(h) \exp(-2\pi i\omega h)$. We have seen that the periodogramme $I(\omega) = |d(\omega)|^2$ estimates $f(\omega)$, but with noticeable variability, also for large n. Some smoothing is required to estimate $f(\omega)$ well. It is useful to estimate *the cumulative spectral distribution* separately, $F(t) = \int_0^t f(\omega)\,\mathrm{d}\omega$, for $0 \le t \le \frac{1}{2}$.

(a) Study the cumulative estimator

$$\widehat{F}(t) = (1/n) \sum_{\omega_j \le t} I(\omega_j), \quad \text{for } 0 \le t \le \tfrac{1}{2},$$

again with $\omega_j = j/n$. Show that $\mathrm{E}\,\widehat{F}(t) \doteq (1/n) \sum_{j/n \le t} f(\omega_j)$, and that this tends to the cumulative $F(t)$ with increasing n.

(b) Show also that

$$\mathrm{Var}\,\widehat{F}(t) \doteq (1/n^2) \sum_{\omega_j \le t} f(\omega_j)^2 \doteq (1/n) \int_0^t f(\omega)^2\,\mathrm{d}\omega.$$

(c) Using the Lindeberg theorem, from large-sample theory, show that

$$Z_n(t) = \sqrt{n}\{\widehat{F}(t) - F(t)\} \to_d Z(t) \sim \mathrm{N}(0, \tau(t)^2), \quad \text{with } \tau(t)^2 = \int_0^t f(\omega)^2\,\mathrm{d}\omega.$$

There is actually full process convergence here, to a Gaußian martingale, i.e. with independent increments, $\mathrm{Var}\,\mathrm{d}Z(\omega) = f(\omega)^2\,\mathrm{d}\omega$.

(d) Simulate an AR(2) process of length say $n = 250$, for some ϕ_1, ϕ_2, compute the periodogramme and its cumulative sum, and then plot the resulting $\widehat{F}(t)$ alongside the real $F(t)$. Try with a few other models too.

(e) Look into a kernel smoothing operation, like

$$\widehat{f}(\omega) = \int K_b(\omega - u)\,\mathrm{d}\widehat{F}(u) = (1/n) \sum_{\omega_j} K_b(\omega - \omega_j) I(\omega_j),$$

with $K_b(v) = b^{-1} K(b^{-1} v)$ and K e.g. the standard normal kernel. Here b is the bandwidth parameter, which needs to be such that $b \to 0$ and $nb \to \infty$ as n increases, in order to secure consistency of $\widehat{f}$, i.e. $\widehat{f}(\omega) \to_{\mathrm{pr}} f(\omega)$.

35. A simple state-space model

The following is meant to serve as a simple illustration of the general theory of Ch. 6 (where sections 6.1, 6.2, 6.3, along with relevant exercises, are our curriculum).

(a) Let x_t be a simple zero-mean stationary AR(1) process, with $x_t = \rho x_{t-1} + w_t$ in terms of a white noise process w_t with variance σ_w^2. Recall that

$$\gamma(h) = \mathrm{cov}(x_t, x_{th}) = \frac{\sigma_w^2}{1 - \rho^2} \rho^h \quad \text{for } h = 0, 1, 2, \ldots.$$

Then suppose the x_t are not observed, they are 'hidden' in a layer below the observations, which we here take to be $y_t \,|\, x_t \sim \mathrm{N}(x_t, \sigma_v^2)$; we may think of σ_v in terms of measurement error. Find

$$\gamma^*(h) = \mathrm{cov}(y_t, y_{t+h}).$$

(b) Assume also that the x_t process is normal. Set up the joint multinormal distribution for $(x_1, \ldots, x_n, y_1, \ldots, y_n)$. You should find that this is a $(2n)$-dimensional multinormal, with zero means, and variance matrix

$$\begin{pmatrix} B, & B \\ B, & B + \sigma_v^2 I \end{pmatrix}, \quad \text{where } B = \sigma_w^2/(1-\rho^2)\, A_\rho,$$

with A_ρ the matrix with elements $\rho^{|j-i|}$.

(c) Find expressions for the conditional distribution of the hidden $x_1, \ldots, x_n$ given $y_1, \ldots, y_n$. (We will learn certain extra methods and formulae and algorithms in a little while, but here you may do it 'brute force'.)

(d) Using this simple state space model, construct predictions and confidence intervals for (i) the next state, i.e. x_{n+1}, and (ii) the next observation, i.e. y_{n+1}. So far, take the parameters of the model, i.e. ρ, σ_w for the states and σ_v for the observations, as known.

(e) Via the above you'll already found that $y \sim \mathrm{N}_n(0, \Sigma)$, with $\Sigma = B + \sigma_v^2 I = \sigma_w^2/(1-\rho^2)A_\rho + \sigma_v^2 I$. Show that this leads to the log-likelihood function

$$\ell = -\tfrac{1}{2} \log |\sigma_w^2/(1-\rho^2)A_\rho + \sigma_v^2 I| - \tfrac{1}{2} y^{\mathrm{t}} (\sigma_w^2/(1-\rho^2)A_\rho + \sigma_v^2 I)^{-1} y.$$

(f) Make an illustration, where you for given values of ρ, σ_w, σ_v simulate $x_1, \ldots, x_n$, followed by simulating $y_1, \ldots, y_n$ given $x_1, \ldots, x_n$, e.g. takin $\sigma_w = 1$ and a suitable value for σ_v (which may be played with when you have the code). In a diagram, put up three curves: (i) the x_t; (ii) the y_t; (iii) the conditional means $\widehat{x}_t = \mathrm{E}\,(x_t \,|\, y_1, \ldots, y_n)$, which may be seen as the best estimates of the hidden layer variables. Also attempt to estimate the three parameters based on the $y_1, \ldots, y_n$ data.

(g) Think through what the above means, in a Bayesian setup, where the $x_1, \ldots, x_n$ might be seen as unknown parameters, following the prior model of an AR(1).

36. A setup for a class of state-space models

Assume $x_1, \ldots, x_n$ form a Markov chain, with densities $f(x_t \,|\, x_{t-1})$, and that the y_t given $x_1, \ldots, x_n$ are independent, with density $g(y_t \,|\, x_t)$.

(a) Show that one in general terms has

$$f(x_t \,|\, x_{t-1}) g(y_t \,|\, x_t) = f(x_t \,|\, x_{t-1}, y_t) g(y_t \,|\, x_{t-1}),$$

in obvious notation. For the normal-normal case, where $x_t \,|\, x_{t-1} \sim \mathrm{N}(\xi_t, \tau_t^2)$ and $y_t \,|\, x_t \sim \mathrm{N}(x_t, {\sigma_v}^2)$, show that conditional on x_{t-1} we have

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} \sim \mathrm{N}_2(\begin{pmatrix} \xi_t \\ \xi_t \end{pmatrix}, \begin{pmatrix} \tau_t^2, & \tau_t^2 \\ \tau_t^2, & \tau_t^2 + \sigma_v^2 \end{pmatrix}).$$

(b)

(c)

(d)

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.

Cunen, C., Hjort, N.L., and Nygård, H. (2020). Statistical sightings of better angels. *Journal of Peace Research*, 57, 221–234.

Heger, A. (2011). Jeg og jordkloden. Dagsavisen.
`aheger.blogspot.com/2007/03/jeg-og-jordkloden.html`

Hermansen, G.H. and Hjort, N.L. Бернштейн–von Mises theorems for nonparametric function analysis via locally constant modelling: A unified approach. *Journal of Statistical Planning and Inference*, 166, 138–157.

Hjort, N.L. (2022). On estimating the cycle length. Manuscript.

Hjort, N.L. and Stoltenberg, E.Aa. (2022). *Statistical Inference: 444 Exercises, 55 Stories, and Solutions to All.* Cambridge University Press; book manuscript in progress.

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, 131, 518–532.

Whittaker, E.T. and Robinson, G. (1923). *The Calculus of Observations, a Treatise on Numerical Mathematics.* Blackie and Son, Ltd.

Whittle, P. (1963). *Prediction and Regulation.* English Universities Press, London.

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.