Slides for TS ch 8, p. 238-261

Exercises: 8.4, 8.7, 8.8, 8.14
ARMA modelling and forecasting: preliminary estimation

Useful for

- order identification (requires the fitting of a number of competing models).
- initial parameter estimates for likelihood optimization.

ARMA(p,q) Model: Based on observations $x_1,\ldots, x_n$, from the model

$$
\phi(B) X_t = \theta(B) Z_t, \quad \{Z_t\} \sim WN(0,\sigma^2),
$$

estimate $\phi = (\phi_1, \ldots, \phi_p)$ and $\theta = (\theta_1, \ldots, \theta_q)$, with the orders $p$ and $q$ assumed known.
AR(p) Processes:

Yule-Walker Estimation (moment estimates)

Burg Estimation

ARMA(p,q) Processes:

Innovations Algorithm

Hannan-Rissanen Estimates
Yule-Walker estimation for AR(p) processes

Recall from Chapter 3, that

\[
\gamma(0) - \phi_1 \gamma(1) - \ldots - \phi_p \gamma(p) = \sigma^2
\]
\[
\gamma(k) - \phi_1 \gamma(k-1) - \ldots - \phi_p \gamma(k-p) = 0, \ k=1, \ldots, p
\]

or in matrix form,

\[
\Gamma_p \phi = \gamma_p
\]

where \(\Gamma_p\) is the covariance matrix of \(X_1, \ldots, X_p\), and \(\phi = (\phi_1, \ldots, \phi_p)'\), \(\gamma_p = (\gamma(1), \ldots, \gamma(p))'\). The Y-W estimates are then found by replacing the ACVF \(\gamma(\cdot)\) by it’s estimated value \(\hat{\gamma}(\cdot)\).
The Yule-Walker fitted model is causal.

- The sample ACVF and fitted model ACVF agree at lags $h=0,1,\ldots,p$.
- Estimates are asymptotically efficient (i.e. have the same limit behavior as MLE)

$$\hat{\phi} \approx N(\phi, n^{-1} \sigma^2 \Gamma_p^{-1})$$
One can apply the Durbin-Levinson algorithm to the sample ACVF in order to compute the Yule-Walker estimates recursively.
Ex: (Dow-Jones Utilities Index; DOWJ.DAT).
Plot of $D_1, \ldots, D_{78}$
ACF of differenced series $Y_t = D_t - D_{t-1}$
Preliminary Model for Dow-Jones:

\[ X_t = .4219 \, X_{t-1} + Z_t, \quad \{Z_t\} \sim \text{WN}(0, .1479) \]  
\[ (\pm .094) \]

where

\[ X_t = Y_t - .1336, \quad Y_t = D_t - D_{t-1}, \]
Burg’s algorithm

The Yule-Walker estimates \((\hat{\phi}_1, \ldots, \hat{\phi}_p)\) satisfy

\[
\tilde{P}_p X_{p+1} = \hat{\phi}_1 X_p + \cdots + \hat{\phi}_p X_1
\]

where \(\tilde{P}_p\) is the prediction operator relative to the fitted Y-W model. As can be seen from the Durbin-Levinson algorithm, these coefficients are determined from the sample PACF,

\[
\hat{\alpha}(h) = \hat{\phi}_{hh}, \; h = 0, \ldots, p.
\]

Burg’s algorithm uses a different estimate of the PACF based on minimizing the forward and backward one-step prediction errors.
Properties of Burg’s estimates:

• The fitted model is causal.

• Estimates are asymptotically efficient (i.e. have the same limit behavior as MLE).

• Estimates tend to perform better than the Y-W estimates especially if a zero of the AR polynomial is close to the unit circle.
Model for the Lake Huron data

The sample ACF and PACF’s suggest an AR(2) model for the mean corrected data

\[ X_t = Y_t - 9.0041. \]

*Burg’s model:*

\[ X_t = 1.0449 X_{t-1} - .2456 X_{t-2} + Z_t, \{Z_t\} \sim WN(0, .4705) \]

*Y-W fitted model:*

\[ X_t = 1.0583 X_{t-1} - .2668 X_{t-2} + Z_t, \{Z_t\} \sim WN(0, .4920) \]

Minimum AICCC model using Burg or Y-W is p=2.
Burg’s gives smaller WN variance and larger likelihood
The innovations algorithm

The Yule-Walker estimates were found by applying the Durbin-Levinson Algorithm with the ACF replaced by the sample ACF. The Innovations algorithm estimates are obtained in the same way -- the ACF is replaced by the sample ACF.

The fitted innovations MA(m) model:

\[ X_t = Z_t + \hat{\theta}_{m1} Z_{t-1} + \ldots + \hat{\theta}_{mm} Z_{t-m}, \{Z_t\} \sim WN(0,\hat{\gamma}_m) \]

where \( \hat{\theta} = (\hat{\theta}_{m1}, \ldots, \hat{\theta}_{mm})' \) and \( \hat{\gamma}_m \) are found from the innovations algorithm with the ACVF replaced by the sample ACVF.
Innovations estimates for MA(q) model

\[ X_t = Z_t + \hat{\theta}_{m,1} Z_{t-1} + \ldots + \hat{\theta}_{m,q} Z_{t-q}, \{Z_t\} \sim WN(0,\hat{\nu}_m), \]

where \( m = m_n \) is a sequence of integers,

\[ m_n \to \infty, \quad m_n = o(n^{1/3}) \]
Order selection for AR(p) processes

Usually there is no true AR model. Goal is to find an AR model which represents the data well, in some sense.

Two Techniques:

• If \( \{X_t\} \) is an AR(p) process with \( \{Z_t\} \sim \text{IID}(0, \sigma^2) \), then
  \[
  \hat{\phi}_{mm} = \hat{\alpha}(m) \sim N(0, n^{-1}) \quad \text{for } m > p. \]
  (\( \hat{\alpha}(m) \) is the sample PACF). Estimate \( p \) as the smallest value \( m \) such that
  \[
  |\hat{\alpha}(k)| < 1.96 n^{-0.5} \quad \text{for } k > m.
  \]

• Estimate \( p \) by minimizing the AICC statistic
  \[
  \text{AICC} = -2\ln L(\phi_p) + 2(p+1)n/(n-p-2)
  \]
  where \( L(\cdot) \) denotes the Gaussian likelihood.
Order selection for MA(q) processes

• If \( \{X_t\} \) is an MA(q) process with \( \{Z_t\} \sim \text{IID}(0,\sigma^2) \), then
  \[
  \hat{\rho}(m) \sim N(0, n^{-1}(1+2\rho^2(1)+\ldots+2\rho^2(q)))
  \]
  for all \( m > q \). Estimate \( q \) as the smallest value \( m \) such that \( \hat{\rho}(m) \) is not significantly different from 0, for all \( k > m \).

• Examine the coefficient vector in the fitted innovations MA(m) model to see which are not significantly different from 0.

• Estimate \( q \) by minimizing the AICC statistic
  \[
  \text{AICC} = -2 \ln L(\Theta_q) + 2(q+1)n/(n-q-2)
  \]
Dow-Jones Data: \( X_t = D_t - D_{t-1} - .1336 \)

ACF suggests MA(2).

MA(2) model using innovations estimates: (m=17)

\[
X_t = Z_t + .4269 \ Z_{t-1} + .2704 \ Z_{t-2}, \ \{Z_t\} \sim \text{WN}(0,.1470)
\]

\[
\begin{align*}
(\pm .114) & \quad (\pm .124)
\end{align*}
\]
MA(17) model:

MA Coefficient
.427  .270  .118  .159  .135  .157  .128  -.006
.015  -.003  .197  -.046  .202  .129  -.021  -.258
.076

Coefficient/(1.96*standard error)
1.911  1.113  .473  .631  .533  .613  .497  -.023
.057  -.006  .759  -.176  .767  .480  -.079  -.956
.276
Preliminary estimation for ARMA(p, q) process

Step 1. Use Innovations algorithm to fit MA(m) model

\[ X_t = Z_t + \hat{\theta}_{m1} Z_{t-1} + \cdots + \hat{\theta}_{mm} Z_{t-m}, \{Z_t\} \sim WN(0, \nu_m), \]

where \( m > p + q \).

Step 2. Replace \( \psi_j \) by \( \theta_{mj} \) in the following equations:

\[ \psi_j = \theta_j + \sum_{i=1}^{\min(j, p)} \phi_i \psi_{j-i}, \quad j = 1, \ldots, p+q. \]

Step 3. Solve the equations in Step 2 for \( \phi \) and \( \theta \).
Step 1. Fit a high order AR(m) \((m > \max(p, q))\) using Yule-Walker estimates and compute estimated residuals

\[
\hat{Z}_t = X_t - \hat{\phi}_{m1} X_{t-1} - \ldots - \hat{\phi}_{mm} X_{t-m}, \quad t = m+1, \ldots, n.
\]

Step 2. Estimate \(\phi\) and \(\theta\), by regressing \(X_t\) onto \((X_{t-1}, \ldots, X_{t-p}, \hat{Z}_{t-1}, \ldots, \hat{Z}_{t-q})\), i.e. by minimizing the sum of squares

\[
S(\phi, \theta) = \sum_{t=m+1}^{n} (X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} - \theta_1 \hat{Z}_{t-1} - \ldots - \theta_q \hat{Z}_{t-q})^2
\]
Lake Huron data: ARMA(1,1) model fitted using the Innovations and Hannan-Rissanen estimates for the mean corrected data $X_t = Y_t - 9.0041$.

Innovations fitted model:

$$X_t - 0.7234 X_{t-1} = Z_t + 0.3596 Z_{t-1}, \{Z_t\} \sim \text{WN}(0,0.4757)$$

$$\begin{array}{ll}
(\pm 0.115) & (\pm 0.099)
\end{array}$$

Hannan-Rissanen fitted model:

$$X_t - 0.6961 X_{t-1} = Z_t + 0.3788 Z_{t-1}, \{Z_t\} \sim \text{WN}(0,0.4757)$$

$$\begin{array}{ll}
(\pm 0.078) & (\pm 0.147)
\end{array}$$
Ex: Suppose \( \{X_t\} \) is the invertible MA(1) process, 

\[
X_t = Z_t + \theta Z_{t-1}
\]

1. Method of moments (solve \( q/(1 + q^2) = \hat{r}(1) \)). Estimator: \( \hat{\theta}_n^{(1)} \)
2. Innovations estimator \( \hat{\theta}_n^{(2)} = \hat{\theta}_{m,1} \)
3. Maximum likelihood estimator: \( \hat{\theta}_n^{(3)} \)

Asymptotic relative efficiency of \( \hat{\theta}_n^{(1)} \) to \( \hat{\theta}_n^{(2)} \):

\[
e(\theta, \hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}) = \frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}
\]

where \( \sigma_1^2(q) \) and \( \sigma_2^2(q) \) are the respective asymptotic variances of the two estimators.
If the relative efficiency is $< 1$ (>$1$), then we say the second estimator is more (less) efficient than the first. For MA(1),

\[
e(\theta, \hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}) = \frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}
\]

\[
e(\theta, \hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}) = \begin{cases} 
.82, & \theta = .25, \\
.37, & \theta = .50, \\
.06, & \theta = .75.
\end{cases}
\]

\[
e(\theta, \hat{\theta}_n^{(2)}, \hat{\theta}_n^{(3)}) = \begin{cases} 
.94, & \theta = .25, \\
.75, & \theta = .50, \\
.44, & \theta = .75.
\end{cases}
\]

MLE more efficient than innovations algorithm which is more efficient than method of moments.
Suppose \( \{X_t\} \) is a causal ARMA(p,q) process. If the noise is IID \( \mathcal{N}(0,\sigma^2) \), then based on the observations \( X_1, \ldots, X_n \), the Gaussian likelihood is given by

\[
L(\Gamma_n) = (2\pi)^{-n/2} \det(\Gamma_n)^{-1/2} \exp\left\{-\frac{1}{2} X_n' \Gamma_n^{-1} X_n\right\}
\]

Direct calculation of \( \det(\Gamma_n) \) and \( \Gamma_n^{-1} \) can be avoided by expressing this in terms of one-step predictors and their MSEs. To see this, note that

\[
X_t = X_t - \hat{X}_t + \hat{X}_t = X_t - \hat{X}_t + \theta_{t-1,1}(X_{t-1} - \hat{X}_{t-1}) + \cdots + \theta_{t-1,t-1}(X_1 - \hat{X}_1)
\]
thus

\[
\begin{bmatrix}
X_1 \\
X_2 \\
X_3 \\
\vdots \\
X_n
\end{bmatrix} =
\begin{bmatrix}
1 & 0 & 0 & \cdots & 0 \\
\theta_{11} & 1 & 0 & \cdots & 0 \\
\theta_{22} & \theta_{21} & 1 & \ddots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \cdots & 1
\end{bmatrix}
\begin{bmatrix}
X_1 - \hat{X}_1 \\
X_2 - \hat{X}_2 \\
X_3 - \hat{X}_3 \\
\vdots \\
X_n - \hat{X}_n
\end{bmatrix}
\]

\(X_n = C_n (X_n - \hat{X}_n)\)

hence

\(\Gamma_n = \text{cov}(X_n) = C_n \text{cov}(X_n - \hat{X}_n)C_n' = C_n D_n C_n'\)

where

\(D_n = \text{diag}(\nu_0, \ldots, \nu_{n-1})\)
From
\[ \Gamma_n = C_n D_n C_n' \quad \text{and} \quad D_n = \text{diag}(v_0, \ldots, v_{n-1}) \]
we find that
\[ |\Gamma_n| = |D_n| = v_0 \cdots v_{n-1} \quad \text{and} \quad \Gamma_n^{-1} = C_n^{-1} D_n^{-1} C_n^{-1} \]
Hence,
\[ X_n' \Gamma_n^{-1} X_n = (X_n - \hat{X}_n)' C_n^{-1} D_n^{-1} C_n^{-1} C_n (X_n - \hat{X}_n) \]
\[ = (X_n - \hat{X}_n)' D_n^{-1} (X_n - \hat{X}_n) \]
\[ = \sum_{t=1}^{n} (X_t - \hat{X}_t)^2 / v_{t-1} \]
so that the Gaussian likelihood can be calculated as
\[ L(\Gamma_n) = (2\pi)^{-n/2} (v_0 \cdots v_{n-1})^{-1/2} \exp\{-1/2 \sum_{t=1}^{n} (X_t - \hat{X}_t)^2 / v_{t-1}\} \]
Maximum likelihood and least squares

Suppose \( \{X_t\} \) is a causal ARMA(p,q) process.

The Gaussian Likelihood:

\[
L(\phi, \theta, \sigma^2) = (2\pi\sigma^2)^{-n/2} (r_0 \cdots r_{n-1})^{-1/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{t=1}^{n} (X_t - \hat{X}_t)^2 / r_{t-1}\right\}
\]

where \( \hat{X}_t = P_{sp\{X_1, \ldots, X_{t-1}\}} X_t \), \( \sigma^2 r_{t-1} = E(X_t - \hat{X}_t)^2 \)

The maximum likelihood estimators are found by maximizing the likelihood, or equivalently, the log-likelihood. Maximizing first wrt to \( \sigma^2 \) gives that

\[
\hat{\sigma}^2 = n^{-1} \sum_{t=1}^{n} (X_t - \hat{X}_t)^2 / r_{t-1} =: n^{-1} S(\phi, \theta)
\]

Where \( \phi \) and \( \theta \) minimize

\[
l(\phi, \theta) = \ln(n^{-1} S(\phi, \theta)) + \sum_{t=1}^{n} \ln r_{t-1}
\]
The quantity

\[ l(\phi, \theta) = \ln(n^{-1} S(\phi, \theta)) + \sum_{t=1}^{n} \ln r_{t-1} \]

is called the reduced (or profile) likelihood and does not depend on \( \sigma^2 \).

Alternatively, one could minimize the sum of squares,

\[ S(\phi, \theta) = \sum_{t=1}^{n} \frac{(X_t - \hat{X}_t)^2}{r_{t-1}} \]

Such estimators are called least squares estimates.
Order selection

AICC Criterion:

Choose \( p, q, \phi_p \) and \( \theta_q \) to minimize

\[
\text{AICC} = -2 \ln L(\phi_p, \theta_q) + 2(p+q+1)n/(n-p-q-2).
\]

For fixed \( p, q \), AICC is minimized when \( \phi_p, \theta_q \) are the maximum likelihood estimates.

Large Sample Behavior of MLE:

For a large sample,

\[
(\hat{\phi}, \hat{\theta})' \approx \mathcal{N}((\phi, \theta)', n^{-1}V(\phi, \theta))
\]
Ex: Lake Huron Data

Preliminary ARMA(1,1) model:

\[ X_t - .7234 \ X_{t-1} = Z_t + .3596 \ Z_{t-1}, \ \{Z_t\} \sim WN(0,.4757) \]

(Model fitted to the mean-corrected data, \( X_t = Y_t - 9.004 \) using the innovations algorithm)

Maximum likelihood ARMA(1,1) model:

\[ X_t - .7453 \ X_{t-1} = Z_t + .3208 \ Z_{t-1}, \ \{Z_t\} \sim WN(0,.4750) \]

\((\pm .0771)\) \hspace{1cm} \((\pm .1116)\)
Maximum Likelihood AR(2) Model:

\[ X_t - 1.0415 X_{t-1} + 0.2494 X_{t-2} = Z_t, \{Z_t\} \sim WN(0, 0.4790) \]

AICC = 213.54 (AR(2) Model)

AICC = 212.77 (ARMA(1,1) model)

Based on AICC, ARMA(1,1) model is slightly better.