Equivalent definition of partial ACF

Definition 5.2

An equivalent definition of the *partial ACF* of lag k is the solution to the system of equations

$$\begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \phi_{k1} \\ \vdots \\ \\ \phi_{kk} \end{pmatrix} = \begin{pmatrix} \rho_1 \\ \vdots \\ \\ \rho_k \end{pmatrix},$$

or

$$\mathbf{R}_k \boldsymbol{\phi}_k = \boldsymbol{\rho}_k.$$

Remark.

1). These are Yule-Walker equations for an AR(k) process; refer to Chapter 4.

2). Proof omitted.

PACF of causal AR(p) has cut-off (1/3)

Proposition 5.1. Let $\mathbb{E}[X_s] = 0$ for all s. PACF at lag k, ϕ_{kk}, is in fact the last coefficient of the best/optimal linear prediction of X_t in terms of $\{X_{t-1}, X_{t-2}, \cdots, X_{t-k}\}$.

Remark 1. Optimal linear prediction of X_t in terms of $\{X_{t-1}, X_{t-2}, \cdots, X_{t-k}\}$ is the linear combination of $\{X_{t-1}, X_{t-2}, \cdots, X_{t-k}\}$ such that the MSE is minimized. Denote it by $\hat{X}$.

Proof. Let $\hat{X}_t = \alpha_1 X_{t-1} + \cdots + \alpha_k X_{t-k}$. To determine the coefficients,

$$\min_{\alpha_1, \cdots, \alpha_k} \mathbb{E}[(X_t - \hat{X}_t)^2].$$

First order condition implies that

$$\mathbb{E}[(X_t - \alpha_1 X_{t-1} - \cdots - \alpha_k X_{t-k}) X_{t-i}] = 0, \qquad i = 1, \cdots, k,$$

which is equivalent to $\gamma_i = \alpha_1 \gamma_{i-1} + \cdots + \alpha_k \gamma_{i-k}$, $i = 1, \cdots, k$:

PACF of causal AR(p) has cut-off (2/3)

$$\begin{aligned}
\gamma_i &= (\gamma_{i-1}, \cdots, \gamma_{i-k})(\alpha_1, \cdots, \alpha_k)^\top \quad i = 1, \cdots, k \\
\Leftrightarrow \rho_i &= (\rho_{i-1}, \cdots, \rho_{i-k})(\alpha_1, \cdots, \alpha_k)^\top \quad i = 1, \cdots, k
\end{aligned}$$

$$\Leftrightarrow \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix}$$

■

Remark 2. The best linear prediction of X_t in terms of $\{X_{t-1}, X_{t-2}, \cdots, X_{t-k}\}$ is the projection of X_t onto $\text{span}(X_{t-1}, X_{t-2}, \cdots, X_{t-k})$.

PACF of causal AR(p) has cut-off (3/3)

Proposition 5.2. Let X be a causal AR(p)

$$X_t = \theta_1 X_{t-1} + \theta_2 X_{t-2} + \cdots + \theta_p X_{t-p} + \epsilon_t.$$

If $k > p$, the best linear prediction in terms of $\{X_{t-1}, X_{t-2}, \cdots, X_{t-p}, \cdots, X_{t-k}\}$ is

$$\hat{X}_t = \theta_1 X_{t-1} + \theta_2 X_{t-2} + \cdots + \theta_p X_{t-p}.$$

Proof. By Remark 2, it is sufficient to verify

$$\mathbb{E}\left[\left(X_t - \theta_1 X_{t-1} - \theta_2 X_{t-2} - \cdots - \theta_p X_{t-p}\right) X_{t-i}\right] = 0, \quad i = 1, 2, \cdots, k.$$

Indeed,

$$\mathbb{E}\left[\left(X_t - \theta_1 X_{t-1} - \theta_2 X_{t-2} - \cdots - \theta_p X_{t-p}\right) X_{t-i}\right] = \mathbb{E}\left[\epsilon_t X_{t-i}\right] = 0, \quad i = 1, 2, \cdots, k.$$

Theorem 5.1. The PACF of causal AR(p) cuts off at lag p.

Proof. By Proposition 5.2, it holds that for $k > p$
$\hat{X}_t = \theta_1 X_{t-1} + \theta_2 X_{t-2} + \cdots + \theta_p X_{t-p} + 0 X_{t-p-1} + \cdots + 0 X_{t-k}$. By Proposition 5.1, $\phi_{kk} = 0$, $k > p$. ■

Calculation of PACF by 2nd definition: Cramer's rule

The last component ϕ_{kk} of the solution vector of

$$\begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{pmatrix} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{pmatrix}$$

is given by Cramer's rule:

$$\phi_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & 1 & \cdots & \rho_2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & 1 \end{vmatrix}}.$$

Calculation of ϕ_{22} by 2nd definition

In particular, we can find ϕ_{22} by solving

$$\begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix} \begin{pmatrix} \phi_{21} \\ \phi_{22} \end{pmatrix} = \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix}.$$

Applying Cramer's rule we get

$$\phi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2},$$

which is the same as that by Definition 5.1.

Example 5.2

Consider a MA(1) process $\{X_t\}$ with

$$X_t = Z_t + \theta Z_{t-1}, \quad \text{where } Z \sim WN(0, \sigma^2) \text{ and } |\theta| < 1.$$

Recall from Chapter 4 that

$$\gamma_0 = (1+\theta^2)\sigma^2, \quad \gamma_1 = \theta\sigma^2, \quad \gamma_k = 0 \text{ for } k > 1,$$

$$\rho_1 = \frac{\theta}{1+\theta^2} \quad \text{and} \quad \rho_k = 0 \text{ for } k > 1.$$

Since $|\theta| < 1$, it is invertible

$$Z_t = \frac{1}{1+\theta B} X_t = \sum_{n=0}^{\infty} (-\theta)^n X_{t-n} := \sum_{n=0}^{\infty} \pi_n X_{t-n} = X_t + \pi_1 X_{t-1} + \pi_2 X_{t-2} + \cdots$$

which implies that $X_t = Z_t - \pi_1 X_{t-1} - \pi_2 X_{t-2} - \cdots$ For PACF, we have

$$\phi_{kk} = \frac{-(-\theta)^k (1-\theta^2)}{1-\theta^{2(k+1)}} \text{ for } k > 0.$$

Sample PACF (1/2)

If only the observations $\{x_1, \cdots, x_n\}$ of a stationary time series X are given, we can calculate the sample partial autocorrelation function (*sample PACF*).

Sample PACF can be defined in the following way.

- Recall the second definition of PACF

$$\begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{pmatrix} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{pmatrix},$$

- Replace ACF ρ_i by *sample ACF* r_i, $i = 1, \cdots, k$:
- Define $\hat{\phi}_{kk}$ through

$$\begin{pmatrix} 1 & r_1 & \cdots & r_{k-1} \\ r_1 & 1 & \cdots & r_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k-1} & r_{k-2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \hat{\phi}_{k1} \\ \hat{\phi}_{k2} \\ \vdots \\ \hat{\phi}_{kk} \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_k \end{pmatrix}.$$

- $\hat{\phi}_{kk}$ is the sample PACF at lag k.

Sample PACF (2/2)

Remark 1: The k-th sample ACF is given by e.g. Page 2, Chapter 4.
Remark 2:

- For AR(p), the sample PACF based on observations $\{x_1, \cdots, x_n\}$ should reflect the properties of PACF.
- If the sample PACF $\hat{\phi}_{kk}$ is significantly different from zero for $k \leq p$ and negligible for $k > p$, then an AR(p) model might provide a good representation of the data.
- How to decide "negligible"? $\pm \frac{2}{\sqrt{n}}$.

Using Cramer's rule to find $\hat{\phi}_{kk}$ is impractical if lag k is large. The following Durbin-Levinson algorithm may be used:

Recursive forms:

$$\begin{aligned} \hat{\phi}_{kk} &= \frac{r_k - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j} r_j}, \\ \hat{\phi}_{kj} &= \hat{\phi}_{k-1,j} - \hat{\phi}_{kk} \hat{\phi}_{k-1,k-j}, \quad j = 1, 2, \ldots, k-1. \end{aligned}$$

Remark 3: Durbin-Levinson algorithm is also valid from ACF and PACF.

Example 5.3: Cramer's rule

Given the following sample autocorrelations for a set of data ($n = 159$):

k	1	2	3	4	5	6	7	8	9	10
r_k	0.841	0.683	0.584	0.515	0.457	0.427	0.405	0.386	0.361	0.321
k	11	12	13	14	15	16	17	18	19	20
r_k	0.329	0.338	0.337	0.294	0.231	0.166	0.126	0.062	0.047	0.042

Then $\hat{\phi}_{11} = r_1 = 0.841$,

$$\hat{\phi}_{22} = \frac{\begin{vmatrix} 1 & r_1 \\ r_1 & r_2 \end{vmatrix}}{\begin{vmatrix} 1 & r_1 \\ r_1 & 1 \end{vmatrix}} = \frac{0.683 - (0.841)^2}{1 - (0.841)^2} = \frac{-0.0243}{0.293} = -0.083,$$

$$\hat{\phi}_{33} = \frac{\begin{vmatrix} 1 & r_1 & r_1 \\ r_1 & 1 & r_2 \\ r_2 & r_1 & r_3 \end{vmatrix}}{\begin{vmatrix} 1 & r_1 & r_2 \\ r_1 & 1 & r_1 \\ r_2 & r_1 & 1 \end{vmatrix}} = \frac{0.00945}{0.0851} = 0.111, \ldots$$

Example 5.3: Durbin-Levinson algorithm

Recursive forms:

$$\begin{aligned}\hat{\phi}_{kk} &= \frac{r_k - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j} r_j}, \\ \hat{\phi}_{kj} &= \hat{\phi}_{k-1,j} - \hat{\phi}_{kk} \hat{\phi}_{k-1,k-j}, \quad j = 1, 2, \ldots, k-1.\end{aligned}$$

The first 20 sample PACF are given by:

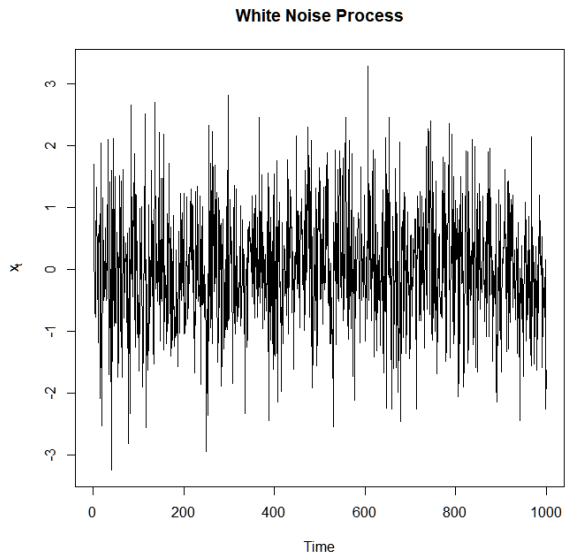
k	1	2	3	4	5	6	7	8	9	10
$\hat{\phi}_{kk}$	0.841	−0.083	0.111	0.036	0.018	0.091	0.025	0.035	0.003	−0.044
k	11	12	13	14	15	16	17	18	19	20
$\hat{\phi}_{kk}$	0.168	0.001	0.027	−0.110	−0.080	−0.057	0.007	−0.152	0.122	−0.071

5.2 Model Identification

Model identification

Three iterative procedure is used to build an ARIMA(p, d, q) model (Box-Jenkins method):

1. Identify plausible choices for p, d, q.
2. Estimate parameters.
3. Check that residuals are consistent with white noise:



Major tools in (i) are the acf, pacf and *extended* acf, AIC and BIC. Before that, we can use time plots and differencing (refer to Chapter 2) to check stationarity.

Model identification (1/2)

- **Time plot** of stationary time series spreads randomly and there is no special pattern.
  - Test stationarity by ADF test (Unit root test): if 1 is a solution of the characteristic equation, we say the time series has a unit root or integrated of order 1.

$$H_0 : \text{unit-root exits} \qquad \text{against} \qquad H_1 : \text{stationarity.}$$

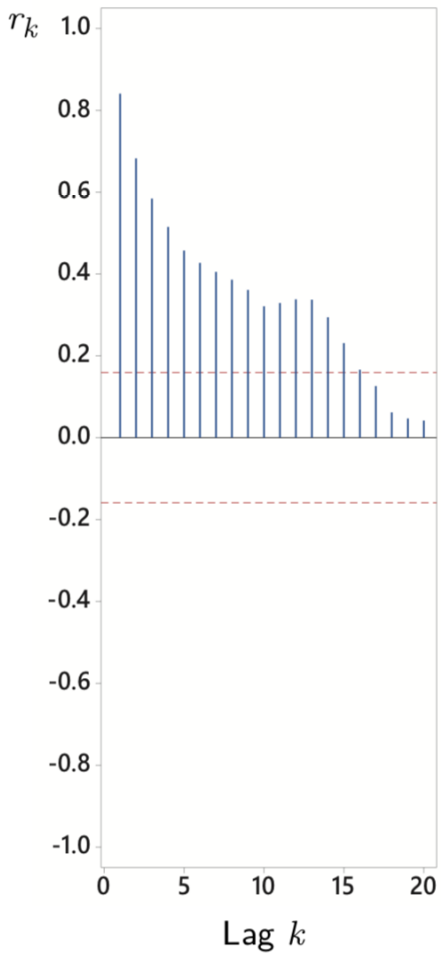
  - If nonstationarity or unit root is found, difference should be considered.
- **Differencing** can eliminate both deterministic and stochastic trends.
  - Over differencing tends to increase variability, e.g. consider $X_t = a + bt + \epsilon_t$, then ∇X is stationary, as well as $\nabla^2 X$ (why?). However,

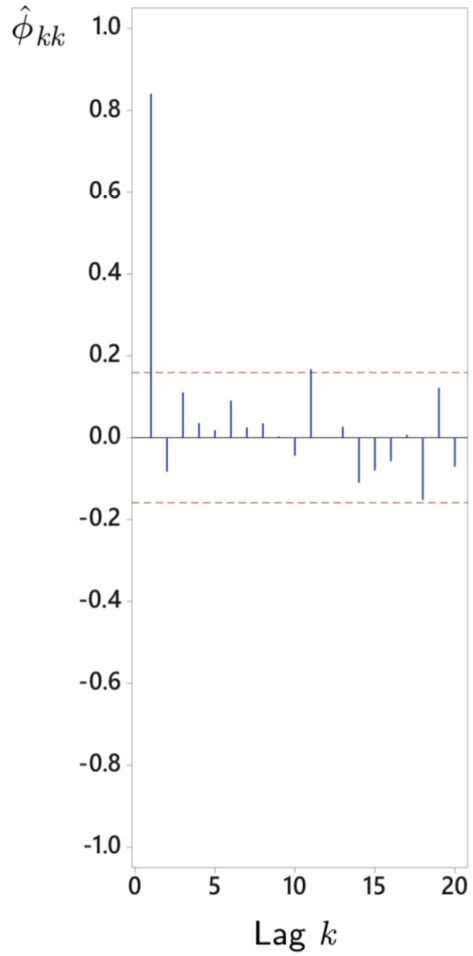
$$\text{Var}(\nabla^2 X_t) > \text{Var}(\nabla X_t).$$

  - Avoid under differencing.
- **(Partial) correlogram** can be used to determined the order or pure AR(p) or MA(q). Conventional error band is $\pm\frac{2}{\sqrt{n}}$.

Example 5.3, *cont'd*

Now $n = 159$ means $2/\sqrt{159} \approx 0.159$. The sample correlogram and partial correlogram are given as follows.





They show that an AR(1) model may be suitable.

Model identification (2/2)

- ACF and PACF of AR and MA models

Model	ACF	Partial ACF
MA(q)	Cuts off after lag q	Dies down exponentially and/or sinusoidally
AR(p)	Dies down exponentially and/or sinusoidally	Cuts off after lag p
ARMA(p, q)	Dies down exponentially and/or sinusoidally	Dies down exponentially and/or sinusoidally

- The form of decay (exponential or sinusoidal) depends on the sign/relation of the coefficients.

- Determination of the order in mixed model is more complicated. One can use e.g. *extended ACF*, AIC and BIC.

Model identification: extended ACF (1/4)

- Idea: in ARMA(p, q), if we filter out the AR component, we have a pure MA(q) model, which can be identified by ACF.
- Consider ARMA(1,1): $X_t = \theta X_{t-1} + Z_t + \phi Z_{t-1}$.
- Filter out AR component:
  - Regress X_t on X_{t-1}: $X_t = \hat{\theta} X_{t-1} + \epsilon_t$, and collect the residual R_t.
  - Regress X_t on X_{t-1} and R_{t-1}: $X_t = \hat{\theta} X_{t-1} + \hat{\theta} R_{t-1} + \epsilon_t$. Here, R_{t-1} is expected to be a proxy of Z_{t-1}.
  - In the second regression $X_t - \hat{\theta} X_{t-1}$ should satisfy MA(1).
- For a general ARMA(p, q) $\Theta(B) X_t = \Phi(B) Z_t$, we need to run the regressions for more rounds (q extra regressions).

Model identification: extended ACF (2/4)

- However, in practice, both p and q are unknown. One needs to run regressions for a series of conjectured (k, j), $k, j = 0, 1, \cdots$
  - Let (p, q) be the true order:

$$X_t = \theta_1 X_{t-1} + \cdots + \theta_p X_{t-p} + \phi_1 Z_{t-1} + \cdots + \phi_q Z_{t-q}.$$

  - Let the hypothetical model be ARMA(k, j).
  - Let the estimated coefficients of the AR component be $(\hat{\theta}_1, \cdots, \hat{\theta}_k)$.
  - Then $W_{t,k,j} := X_t - \hat{\theta}_1 X_{t-1} - \cdots - \hat{\theta}_k X_{t-k}$.
  - The *extended (sample) ACF (EACF)* is the (sample) ACF of $W_{t,k,j}$.

1. If the hypothesized AR order k equals the true AR order p, and if the hypothesized MA order $j \geq q$, then $W_{\cdot,k,j}$ is an MA(q) process.
   In this case, ACF of $W_{\cdot,k,j}$ at lag $q+1$, $q+2$, $\cdots$ should be zero.
2. If $k < p$, then $W_{\cdot,k,j}$ ~ARMA. No cutoff at any lags.
3. If $k > p$, then ACF of $W_{\cdot,k,j}$ has cutoff at $q+1$, if $j \geq q+1$.

Model identification: extended ACF (3/4)

Generation of tables for extended ACF.

- If the ACF of $W_{\cdot,k,j}$ at lag $j+1$ is not zero, then set the position (k, j) as "X".
- If the ACF of $W_{\cdot,k,j}$ at lag $j+1$ is zero, set the position (k, j) as "o".
- Consider the table for an ARMA(1,1)

Model identification: extended ACF (4/4)

AR/MA	0	1	2	3	4	5	6	7	8	9	10
0	x	x	x	x	x	x	x	x	x	x	x
1	x	x	x	x	x	x	x	x	x	x	x
2	x	x	x	o	o	o	o	o	o	o	o
3	x	x	x	x	o	o	o	o	o	o	o
4	x	x	x	x	x	o	o	o	o	o	o
5	x	x	x	x	x	x	o	o	o	o	o
6	x	x	x	x	x	x	x	o	o	o	o
7	x	x	x	x	x	x	x	x	o	o	o

Table: Extended Autocorrelation Function (EACF) Table

Model identificaiton: AIC and BIC

- Akaike information criterion (**AIC**):

$$AIC = 2k - 2\log L,$$

  where k is the number of parameters, L is the likelihood function.

  - The smaller of AIC, the better
  - k is incorporated to avoid overfitting.

- Bayesian information criterion (**BIC**):

$$BIC = k * \log n - 2\log L,$$

  where n is the number of observations.

- Software reports the result of both.

5.3 Parameter Estimation

Parameter estimation

Three iterative procedure is used to build an ARIMA(p, d, q) model (Box-Jenkins method):

(i) Identify plausible choices for p, d, q.

(ii) Estimate parameters:

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \quad Z \sim WN(0, \sigma^2).$$

Unknown parameters are: ϕ_j, $j = 1, \cdots, p$, θ_i, $i = 1, \cdots, q$ and σ^2.

- Yule-Walker equations.
- Maximum likelihood estimation (MLE).

(iii) Check that residuals are consistent with white noise.

Parameter estimation by Yule-Walker equations: AR(p)

- Consider the general *causal* AR(p) process

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t, \quad Z \sim WN(0, \sigma^2).$$

- Recall the Yule-Walker equations in Chapter 4

$$\begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \cdots & \rho_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix},$$

  where ρ_k is the ACF of X at lag k.

- Sample ACF r_k is considered to be an estimate of ACF $\rho(k)$, thus, the estimate of $\phi_1, \cdots, \phi_p$, denoted by $\widehat{\phi}_1, \cdots, \widehat{\phi}_p$, is determined by

$$\begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{pmatrix} = \begin{pmatrix} 1 & r_1 & \cdots & r_{p-1} \\ r_1 & 1 & \cdots & r_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p-1} & r_{p-2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \widehat{\phi}_1 \\ \widehat{\phi}_2 \\ \vdots \\ \widehat{\phi}_p \end{pmatrix},$$

Parameter estimation by Yule-Walker equations: AR(1)

AR(1): $X_t - \phi X_{t-1} = Z_t$, $Z \sim WN(0, \sigma^2)$.

Parameter estimation by Yule-Walker equations: MA(2)

MA(2): $X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2}$, $Z \sim WN(0, \sigma^2)$.

Exercise!

Parameter estimation by YW equations: ARMA(1,1)

Consider *causal* **ARMA(1,1)**: $X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}, \quad Z \sim WN(0, \sigma^2)$.

Exercise!

Maximum likelihood estimation (1/2)

- Why MLE? In large samples, MLE is consistent, efficient and normally distributed. In small samples, it satisfies an invariance property, is a function of sufficient statistics and in some cases is unbiased.
- We begin with the case of cross-sectional data. Suppose that we have a sample of independent r.v.s. X_t, $t = 1, 2, \ldots, N$ each with pdf $f(x_t)$. Then the joint density function is

$$f(x_1, x_2, \ldots, x_N) = \prod_{t=1}^{N} f(x_t). \tag{2}$$

  Let X_t contain an unknown parameter θ (it can also be a vector). The likelihood function is the joint density (2), but is regarded as a function of the parameter θ:

$$L(\theta) = \prod_{t=1}^{N} f(x_t; \theta).$$

- $\widehat{\theta} = \arg\max_\theta L(\theta) = \arg\max_\theta \log L(\theta) = \arg\max_\theta \sum_{t=1}^{N} \log f(x_t; \theta)$.

Maximum likelihood estimation (2/2)

- The problem with time series is that the observations are NOT independent, so that (2) is not applicable.
- Denote all observations up to and including t by $\boldsymbol{x}_t = (x_1, \cdots, x_t)$. By Bayes rule

$$\begin{aligned} f(x_1, x_2, \ldots, x_N) &= f(x_N|\boldsymbol{x}_{N-1}) f(x_1, x_2, \ldots, x_{N-2}, x_{N-1}) \\ &= f(x_N|\boldsymbol{x}_{N-1}) f(x_{N-1}|\boldsymbol{x}_{N-2}) f(x_1, x_2, \ldots, x_{N-2}), \end{aligned}$$

and so on until

$$f(x_1, x_2, \ldots, x_N) = \prod_{t=2}^{N} f(x_t|\boldsymbol{x}_{t-1}) f(x_1),$$

where $f(x_t|\boldsymbol{x}_{t-1})$ is the conditional pdf of x_t given all observations prior to t. So the likelihood $L(\theta)$ can be written as the product of these conditional distributions:

$$L(\theta) = \prod_{t=2}^{n} f(x_t|\boldsymbol{x}_{t-1}) f(x_1).$$

Example 5.5

Consider a causal AR(1) model

$$X_t = \phi X_{t-1} + \epsilon_t, \quad \epsilon \sim NID(0, \sigma^2).$$

Assume $X_0 = 0$. Estimate ϕ and σ^2 by MLE.

Example 5.6 (1/2)

Let $\{x_1, x_2, \ldots, x_n\}$ be observations of an *invertible* MA(1) model, i.e.,

$$X_t = \theta a_{t-1} + a_t, \quad a \sim NID(0, \sigma^2).$$

Estimate θ and σ^2 by MLE.

Conditioned on $\{x_1, \cdots, x_{t-1}\}$ = conditioned on $\{a_1, \cdots, a_{t-1}\}$:
For each s, x_s is a linear combination of $\{a_1, \cdots, a_{s-1}, a_s\}$.
For each s, a_s is a linear combination of $\{x_1, \cdots, x_s\}$ by invertibility and assuming the observations before zero are well known.

The conditional distribution of X_t given $\{X_1, \cdots, X_{t-1}\}$ is $\mathcal{N}(\theta a_{t-1}, \sigma^2)$. The conditional density is $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_t - \theta a_{t-1})^2}{2\sigma^2}\right]$. The likelihood function is

$$\prod_{t=2}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_t - \theta a_{t-1})^2}{2\sigma^2}\right].$$

and hence the conditional log-likelihood is

$$l(\theta, \sigma^2) = -\frac{n-1}{2}\log(2\pi\sigma^2) - \sum_{t=1}^{n-1} \frac{(x_{t+1} - \theta a_t)^2}{2\sigma^2}.$$

Example 5.6 (2/2)

The first order condition implies

$$\begin{cases} \dfrac{\partial l}{\partial \sigma^2} = 0 \Rightarrow \sigma^2 = \displaystyle\sum_{t=1}^{n-1} \frac{(x_{t+1} - \theta a_t)^2}{n-1} \\ \dfrac{\partial l}{\partial \theta} = 0 \Rightarrow f(\theta) := \displaystyle\sum_{t=1}^{n-1} (x_{t+1} - \theta a_t) \left(a_t + \theta \frac{\partial a_t}{\partial \theta} \right) = 0. \end{cases}$$

Use the second equation to solve θ then take $\widehat{\theta}$ into the first equation to get $\widehat{\sigma}^2$.
Note that a is not observable, but we do have $a_t = \sum_{j=0}^{\infty} (-\theta)^j x_{t-j}$.

5.4 Residual Diagnosis

Residual diagnosis

Box-Jenkins method:

(i) Identify plausible choices for p, d, q.

(ii) Estimate parameters.

(iii) Check that residuals are consistent with white noise:

- After a (several) tentative model(s) has (have) been fit to the data, we need examine its (their) adequacy. This is done through residual analysis/diagnosis checking/residual diagnosis.
- The residual of ARMA(p, q) can be obtained from

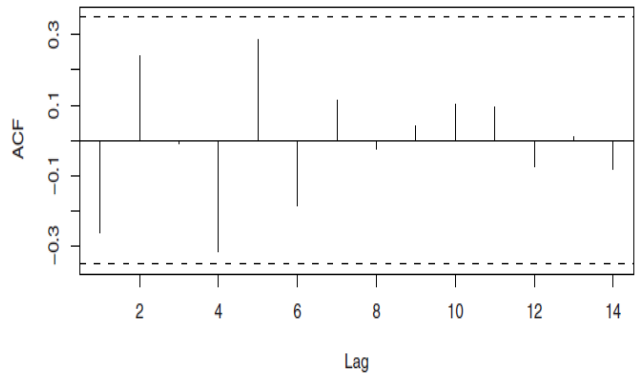
$$\widehat{\epsilon}_t = y_t - \left(\widehat{\mu} + \sum_{i=1}^{p} \widehat{\phi}_i y_{t-i} - \sum_{i=1}^{q} \widehat{\theta}_i \widehat{\epsilon}_{t-i} \right).$$

- If the specified model is adequate, the residuals should behave like white noise:
  - uncorrelated; if there is correlation between residuals, then there is information left in residuals which should be used in the forecasts.
  - zero mean; otherwise the forecast is biased.

Techniques of residual diagnosis

The diagnosis technique includes but not limited to

1. correlogram of residuals: sample ACF of $\{\widehat{\epsilon_t}\}$, denoted by $r_\epsilon(k)$.



2. Portmanteau test, or χ^2 test: Ljung-Box test ↷

Ljung-Box test

H_0: the residuals are independently distributed;
H_1: residuals are not independently distributed.

Ljung-Box statistic:

$$Q_{\text{LB}} = n\,(n+2) \sum_{k=1}^{h} \frac{r_\epsilon(k)^2}{n-k}.$$

- n is the sample size,
- $r_\epsilon(k)$ is the sample ACF of the residuals at lag k, and
- h is the number of lags being tested.

Remark. 1) Rather than considering ACF $r_\epsilon(k)$ individually, we may obtain an indication of whether the first h residual ACF together indicate adequacy of the model.

2) The statistic Q_{LB} follows $\chi^2(h-m)$, where m is the number of parameters in the model. If the specified model is ARMA(p,q), $Q_{LB} \sim \chi^2(h-p-q)$.

3) A popular choice of h is $\min\{20, n-1\}$, suggested by Box and Jenkins.