Estimation

- Nonparametric inference:
  - sample mean and sample ACF: the estimators for the first two moments
  - the distributions of the sample moments

- Parametric inference: focus on ARMA models
  - YW estimator for AR models
  - least square method
  - maximum likelihood estimation
Assume \( \{X_t\} \) is stationary with mean \( \mu \) and ACF \( \gamma(\cdot) \) satisfying \( \sum |\gamma(h)| < \infty \).

\[
\bar{X}_n = \frac{1}{n} \sum_{t=1}^{n} X_t, \quad E\bar{X}_n = \mu,
\]

\[
n \text{var}(\bar{X}_n) = nE(\bar{X}_n - \mu)^2 \rightarrow \sum_{h=-\infty}^{\infty} \gamma(h).
\]

- under mild conditions, \( \sqrt{n}( \bar{X}_n - \mu ) \rightarrow N(0, \sum_{h=-\infty}^{\infty} \gamma(h)) \). (Theorem 7.1.1)

- \( \sum_{h=-\infty}^{\infty} \gamma(h) = 2\pi f(0) \) is called long-term variance where \( f(\omega) \) is the spectral density (talk about this later)

- For a linear process \( X_t = \mu + \sum_j \psi_j Z_{t-j}, \sum_{h=-\infty}^{\infty} \gamma(h) = \sigma^2 \left( \sum_j \psi_j \right)^2 \).
Define

\[ \hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (X_t - \bar{X}_n)(X_{t+h} - \bar{X}_n), \quad h = 0, 1, \ldots, n - 1, \]

\[ \tilde{\gamma}(h) = \frac{1}{n-h} \sum_{t=1}^{n-h} (X_t - \bar{X}_n)(X_{t+h} - \bar{X}_n), \]

\[ \hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0), \quad \tilde{\rho}(h) = \tilde{\gamma}(h)/\tilde{\gamma}(0). \]

- \{\hat{\gamma}(h) : h = 0, 1, \ldots, n - 1\} and \hat{\rho}(h) are non-negative definite, but \tilde{\gamma}(\cdot) and \tilde{\rho}(\cdot) are not.
- \hat{\gamma}(h) performs poorly for large \( h \) (few pairs available for computing the estimate).
- use parametric model (constraint on \( \gamma(h) \)) to improve the estimation for ACF.
- under some mild conditions, \( \sqrt{n}(\hat{\rho}_h - \rho_h) \to N(0, W) \) where \( \rho_h = (\rho(1), \ldots, \rho(h))' \), \( W \) is defined by Bartlett’s formula (Theorem 7.2.1).
Bartlett's Formula

\[ w_{ij} = \sum_{k=-\infty}^{\infty} \{ \rho(k+i)\rho(k+j) + \rho(k-i)\rho(k+j) + 2\rho(i)\rho(j)\rho^2(k) \]
\[-2\rho(i)\rho(k)\rho(k+j) - 2\rho(j)\rho(k)\rho(k+i) \}. \]

Special cases:

- WN: \( W = I \) (useful for testing WN)
- MA(\( q \)):
  \[ w_{ii} = 1 + 2 \sum_{h=1}^{q} \rho^2(h), \quad i > q \]
- AR(1):
  \[ w_{ii} = (1 - \phi^{2i})(1 + \phi^2)/(1 - \phi^2) - 2i\phi^{2i}, \quad i \geq 1, \]
  \[ w_{ii} \approx (1 + \phi^2)/(1 - \phi^2), \quad \text{for } i \text{ large}. \]
Test for White Noise

\[ H_0 : \{X_t\} \sim WN \quad \text{against} \quad H_1 : \{X_t\} \text{ is not WN} \]

Individual Test:

Check \( |\hat{\rho}(h)| < 1.96/\sqrt{n} \) since \( \sqrt{n}\hat{\rho}(h) \rightarrow N(0, 1) \) under \( H_0 \).

Portmanteau Test (combining ACF at different lags):

- Box-Pierce (1970):

  \[ Q = n \sum_{j=1}^{h} \hat{\rho}^2(j) \rightarrow \chi^2_h \text{ under } H_0. \]

- Ljung-Box (1978):

  \[ Q^* = n(n + 2) \sum_{j=1}^{h} \hat{\rho}^2(j)/(n - j) \rightarrow \chi^2_h \text{ under } H_0. \quad \text{(provided } EX_t^2 < \infty) \]

- McLeod and Li (1983): test for independence

  \[ Q_{XX}^* = n(n + 2) \sum_{j=1}^{h} \hat{\rho}_{XX}^2(j)/(n - j) \rightarrow \chi^2_h \text{ under } H_0. \quad \text{(provided } EX_t^4 < \infty) \]
Test for White Noise (cont.)

- Remarks about Portmanteau test:
  - choose $h$ to grow with $n$
  - if testing data are residuals, then the asymptotic distribution is $\chi^2_{h-df}$ where $df$ is the number of parameters in the fitted model
  - better use for eliminating bad models than for selecting good models
  - similar test can be used for testing normality (McLeod-Li)

- Other testing methods: turning point test, difference-sign test, rank test
YW Estimation for AR($p$)

\[ \Gamma_p \phi = \gamma_p \quad \text{(YW equation)} \]
\[ \sigma^2 = \gamma(0) - \gamma'_p \Gamma_p^{-1} \gamma_p = \gamma(0) - \gamma'_p \phi. \]

- solve $\hat{\phi}$ and $\hat{\sigma}^2$ by plugging the sample ACF $\hat{\gamma}(h)$ in $\Gamma_p$, $\gamma_p$ and $\gamma(0)$, i.e.,

\[ \hat{\phi} = \hat{\Gamma}_p^{-1} \hat{\gamma}_p, \quad \hat{\sigma}^2 = \hat{\gamma}(0) \left( 1 - \hat{\rho}_p \hat{R}_p^{-1} \hat{\rho}_p \right). \]

- YW estimator is a moment estimator but it is efficient and asymptotically equivalent to the MLE for AR processes.

- as $n \to \infty$, $\sqrt{n}(\hat{\phi} - \phi) \to N \left( 0, \sigma^2 \Gamma_p^{-1} \right)$ and $\hat{\sigma}^2 \to \sigma^2$. 
Gaussian Likelihood and MLE

Assume \( \{X_t\} \) is a zero-mean stationary process with ACF \( \gamma(h) \), the Gaussian likelihood is

\[
\mathcal{L}_n(\theta) = f(X_n) = (2\pi)^{-n/2}|\Gamma_n|^{-1/2} \exp\left\{-X_n'\Gamma_n^{-1}X_n/2\right\}.
\]

Another representation based on one-step prediction:

\[
\mathcal{L}_n(\theta) = f(X_n) = \prod_{t=1}^{n} f(X_t|X_{t-1})
\]

\[
= (2\pi)^{-n/2} \left(\prod_{t=1}^{n} v_{t-1}\right)^{-1/2} \exp\left\{-\frac{1}{2} \sum_{t=1}^{n} (X_t - \hat{X}_t)^2 / v_{t-1}\right\},
\]

where \( \hat{X}_t \) is the best linear predictor of \( X_t \) given \( X_{t-1} \) and \( v_{t-1} \) is the corresponding PMSE, both are functions of \( \theta \).
For $t > p$,

$$\hat{X}_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} \equiv \phi X_{t-1}(p), \quad v_{t-1} = \sigma^2.$$ 

The Gaussian likelihood becomes

$$\mathcal{L}_n(\theta) = f(X_p) \prod_{t=p+1}^{n} f(X_t|X_{t-1}) = f(X_p) \prod_{t=p+1}^{n} f(X_t|X_{t-1}(p))$$

$$= (2\pi)^{-n/2} |\Gamma_p|^{-1/2} \exp\left\{-X_p'\Gamma_p^{-1} X_p/2\right\}$$

$$\times \left(\sigma^2\right)^{-(n-p)/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{t=p+1}^{n} (X_t - \phi' \hat{X}_{t-1}(p))^2 \right\}.$$ 

Re-parametrize $v_t \equiv \sigma^2 r_t$, the log Gaussian likelihood is

$$\ell_n(\theta) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2} \sum_{t=1}^{n} \log r_{t-1} - \frac{1}{2\sigma^2} \sum_{t=1}^{n} (X_t - \hat{X}_t)^2/r_{t-1}.$$
Gaussian Likelihood (cont.)

- scale parameter $\sigma^2$ can be concentrated out from $\ell_n(\theta)$ first:

\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^{n} (X_t - \hat{X}_t)^2 / r_{t-1}
\]

- concentrated (or reduced) log-likelihood:

\[
\ell^*_n(\theta^*) = \ell_n(\theta^*, \hat{\sigma}^2) = -\frac{n}{2} \log \hat{\sigma}^2 - \frac{1}{2} \sum_{t=1}^{n} \log r_{t-1}
\]

- most of $r_t$ is exact 1, the number of terms that $r_t \neq 1$ compared to $n$ is relatively small

- ignore log $r_t$, to maximize $\ell^*_n(\theta^*)$ is equivalent to minimize $\hat{\sigma}^2$ (weighted least square)

- most of $r_t$ is 1, weighted least square (WLS) is equivalent to OLS which minimize $\sum_{t=1}^{n} (X_t - \hat{X}_t)^2$
WLS: minimize $\sum_{t=1}^{p}(X_t - \hat{X}_t)^2/r_{t-1} + \sum_{t=p+1}^{n}(X_t - \phi' \hat{X}_{t-1}(p))^2$

- ignore the first $p$ terms, the estimator to minimize $\sum_{t=p+1}^{n}(X_t - \phi' \hat{X}_{t-1}(p))^2$ becomes the conditional MLE (conditional on the first $p$ observations)

- as $n \to \infty$, WLS, OLS, conditional MLE, YW estimator are all equivalent to MLE.

- it is easy to solve the conditional MLE for AR models since it only need to solve a linear equation

- In splus, “arima.mle” computes conditional MLE, “ar.yw” computes YW estimator and “ar.burg” computes estimate using Burg’s algorithm. (missing data are allowed for some functions)

- In R, “arima” computes exact MLE.
Asymptotic Distribution of MLE

Consider an ARMA(p,q) satisfying

\[(1 - \phi_1 B - \cdots - \phi_p B^p) X_t = (1 + \theta_1 B + \cdots + \theta_q B^q) Z_t, \quad Z_t \sim WN(0, \sigma^2).\]

Let \(\beta' = (\phi', \theta').\) The MLE has the limit distribution

\[n^{1/2}(\hat{\beta} - \beta) \rightarrow N(0, V(\beta)),\]

\[V(\beta) = \sigma^2 \left( EU_t U'_t \quad EU_t V'_t \right)^{-1},\]

where

\[U_t = (U_t, \ldots, U_{t+1-p})',\]
\[V_t = (V_t, \ldots, V_{t+1-q})',\]
\[\phi(B) U_t = Z_t,\]
\[\theta(B) V_t = Z_t.\]
AR(1) Likelihood and Estimation (MLE: solid line; YW: dash line)

AR(1): MLE and YW estimate; n= 25 ; phi= 0.5

AR(1): MLE and YW estimate; n= 25 ; phi= 0.9

AR(1): MLE and YW estimate; n= 100 ; phi= 0.5

AR(1): MLE and YW estimate; n= 100 ; phi= 0.9
MA(1) Likelihood and Estimation

MA like & css; n= 25 ; theta= 0.5

MA like & css; n= 100 ; theta= 0.5

MA like & css; n= 25 ; theta= 0.9

MA like & css; n= 100 ; theta= 0.9
Unconstrained Likelihood and Estimation for MA(1)

MA(1) like ; n= 100 ; theta= 0.5
Remarks about Estimation and Prediction

1. **Non-identifiability issue:**
   - Common component: e.g., $$(1 - \phi B)X_t = (1 - \theta B)Z_t, \phi = \theta.$$  
   - Roots of AR and MA polynomials are in side or out side the unit circle: e.g., 
     $$(1 - \phi B)X_t = Z_t \text{ v.s. } (1 - \phi^{-1}B)X_t = U_t.$$ 

2. **Unit root problem:**
   - The asymptotic properties of estimators and test statistics are different from the non-unit cases: e.g., when $\phi = 1$ in an AR(1), $n(\hat{\phi} - 1) = O_p(1)$.
   - Nearly unit case: finite sample distribution for the estimators can be corrected by using bootstrap methods. E.g., for an AR(1) with $\phi \approx 1$, generate $y_t^* = \hat{\phi}y_{t-1}^* + Z_t$, $Z_t \sim N(0, \sigma^2)$; and estimate $\phi$ based on the bootstrap samples $\{y_t^*: t = 1, ..., n\}$ for $m$ times (large $m$) to obtain $\{\hat{\phi}_i^*: i = 1, ..., m\}$. Construct the finite-sample
distribution for \( \hat{\phi} \) by \( \{\hat{\phi}_i^*: i = 1, \ldots, m\} \).

3. **Relationship between IMA and EWMA:**

- EWMA (exponentially weighted moving averages): a commonly used forecasting method in engineering. \( \bar{X}_{n+1} = (1 - \lambda)X_n + \lambda \bar{X}_n, 0 < \lambda < 1. \)
  \( (\lambda: \text{smoothing parameter to be selected}) \)

- Forecasts using an EWMA are **optimal** (BLP) if the model is a particular IMA.

**Proof:** For an IMA: \( (1 - B)X_t = \epsilon_t - \lambda \epsilon_{t-1}, |\lambda| < 1, \)
\( \epsilon_t = (1 - \lambda B)^{-1}(1 - B)X_t = \left(\sum_{j=0}^{\infty} \lambda^j B^j\right)(1 - B)X_t = X_t - \sum_{j=1}^{\infty} (1 - \lambda)\lambda^{j-1}X_{t-j}. \)

Therefore, given the infinite past of \( \{X_t\} \), the best 1-step prediction satisfies
\[ \bar{X}_{n+1} = \sum_{j=1}^{\infty} (1 - \lambda)\lambda^{j-1}X_{n+1-j} = (1 - \lambda)X_n + \lambda\left\{\sum_{j=1}^{\infty} (1 - \lambda)\lambda^{j-1}X_{n-j}\right\} \]
\[ = (1 - \lambda)X_n + \lambda \bar{X}_n. \]