Time Series Econometrics, 2ST111

Lecture 4. Forecasting and Maximum Likelihood Estimation

Yukai Yang

Department of Statistics, Uppsala University

Outline of Today's Lecture

- Forecasting (pp.108-116 in Hamilton)
- Maximum Likelihood Estimation
- Asymptotic Distribution Theory

Forecasting

- Wold's Decomposition Theorem
- The Box-Jenkins Modelling Philosophy
- Model Selection

Wold's Decomposition

Recall that any (covariance) stationary ARMA(p, q) process can be written as

$$Y_t = \mu + \psi(L)\varepsilon_t = \mu + (\psi_0 + \psi_1 L + \psi_2 L^2 + \ldots)\varepsilon_t \tag{1}$$

where ε_t is interpreted as the, white noise, forecast error

$$Y_t - \mathrm{E}(Y_t | y_{t-1}, y_{t-2}, \ldots) \tag{2}$$

and where $\sum_{i=0}^{\infty} \psi_i^2 < \infty$ with $\psi_0 = 1$.

It turns out that the above representation is fundamental for any stationary time series.

Wold's Decomposition

Wold's Decomposition

Any zero-mean covariance stationary process $\{Y_t\}_{t=-\infty}^{\infty}$ can be represented as

$$Y_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i} + \kappa_t \tag{3}$$

where $\psi_0 = 1$, $\sum_{i=0}^{\infty} \psi_i^2 < \infty$ and

$$\varepsilon_t = Y_t - \hat{E}(Y_t | y_{t-1}, y_{t-2}, \ldots) \tag{4}$$

is white noise. The linearly deterministic component

$$\kappa_t = \hat{E}(\kappa_t | y_{t-1}, y_{t-2}, \ldots) \tag{5}$$

is uncorrelated with ε_{t-i} for any i.

Wold's Decomposition

Remarks:

- $\sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$ is called the linearly indeterministic component of Y_t.
- If κ_t is zero for all t, then $\{Y_t\}_{t=-\infty}^{\infty}$ is called purely linearly indeterministic.
- The proposition essentially says that any stationary process can be expressed as the sum of two uncorrelated processes.
- The Wold's representation is named after Herman Wold, who was a professor in Statistics at Uppsala University.
- In practice, we seek a model that gives an adequate approximation to the data with as few parameters as possible. A typical assumption is that $\psi(L)$ can be expressed as

$$\sum_{i=0}^{\infty} \psi_i L^i = \frac{1 + \theta_1 L + \theta_2 L^2 + ... + \theta_q L^q}{1 + \phi_1 L + \phi_2 L^2 + ... + \phi_p L^p} \tag{6}$$

The Box-Jenkins Modelling Philosophy

Box & Jenkins procedure can be broken down into four steps:

- Transform the data, if necessary, so that the assumption of covariance stationarity appears to be satisfied.
- Examine the transformed data to see which ARMA(p, q) process appears to be most appropriate.
- Estimate the parameters in $\phi(L)$ and $\theta(L)$ accompanying the chosen model.
- Assess the chosen models adequacy by checking whether the model assumptions are satisfied. If not, repeat the procedure.

The Box-Jenkins Modelling Philosophy

Remarks:

- They argued that, in practice, the more parameters to estimate, the more room there is to go wrong.
- Transformation: natural logarithm, square root, differencing, and etc.
- Model Identification or Specification: the ARMA form, plausible values of p and q selected by for example sample autocorrelations and partial autocorrelations.
- Estimation: OLS, ML, GMM and etc.
- Diagnostic Checking or Model evaluation: serial correlation test, heteroskedasticity test, Gaussianity test, and etc.

Model Selection

In practice, a time series analyst often ends up with two or more seemingly adequate parsimonious models.

In this case, some model selection information criteria, such as AIC or BIC, can be used to help in choosing an appropriate model.

Maximum Likelihood Estimation

- The Method of Maximum Likelihood
- MLE for a Gaussian AR(1)
- Conditional MLE for a Gaussian AR(1)
- Conditional MLE for a Gaussian MA(1)
- Conditional MLE for a Gaussian ARMA(p, q)
- Statistical Inference with ML Estimation

Maximum Likelihood Estimation

Let us start with the general ARMA(p, q) model

$$Y_t = c + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} \tag{7}$$

where ε_t is white noise with the variance σ^2.

Now we start to consider the reality! What if all the values in the parameter vector

$$\boldsymbol{\theta} = (c, \phi_1, ..., \phi_p, \theta_1, ..., \theta_q, \sigma^2)' \tag{8}$$

are unknown?

We have to estimate all of them based on the observations $y_1, ..., y_T$.

Maximum Likelihood Estimation

The random numbers $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ are unobservable. But why not assume that they follow some distribution, for example Gaussian? This Gaussian assumption can be checked later.

Then we have the joint probability density function (value) of the sample (at the point $y_1, ..., y_T$)

$$f_{Y_1,...,Y_T}(y_1, y_2, ..., y_T; \boldsymbol{\theta}). \tag{9}$$

Let us suppress the subscript $Y_1, ..., Y_T$ in the density function. We say $f(y_t; \boldsymbol{\theta}) = f_{Y_t}(y_t; \boldsymbol{\theta})$ for random number Y_t and the observation y_t. Similarly we have the conditional density $f(y_t | y_\tau, \boldsymbol{\theta}) = f(y_t | y_\tau; \boldsymbol{\theta}) = f_{Y_t | Y_\tau}(y_t | y_\tau; \boldsymbol{\theta})$.

Maximum Likelihood Estimation

When the sample $\{y_t\}_{t=1}^T$ is *iid*, the joint density can be decomposed as

$$f(y_1, y_2, ..., y_T; \boldsymbol{\theta}) = f(y_1; \boldsymbol{\theta}) f(y_2; \boldsymbol{\theta}) ... f(y_T; \boldsymbol{\theta}). \tag{10}$$

However, in time series, it is normally the case that y_t are dependent.

$f_{Y_1,...,Y_T}(...)$ is the joint density function, whose argument or input is the observations. However, if you regard all the observations as given and the parameters $\boldsymbol{\theta}$ as the argument, it becomes a function of $\boldsymbol{\theta}$. This function is called the likelihood function.

$$L(\boldsymbol{\theta}) = f(y_1, y_2, ..., y_T; \boldsymbol{\theta}) \tag{11}$$

f is a function of y, and L is a function of $\boldsymbol{\theta}$.

Maximum Likelihood Estimation

For a given sample of data with observed values $y_1, ..., y_T$, the maximum likelihood (ML) estimator of $\boldsymbol{\theta}$ is obtained by maximizing the likelihood function (11):

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \tag{12}$$

or, equivalently, by maximizing the log-likelihood function:

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}) \tag{13}$$

Estimator $\hat{\boldsymbol{\theta}}$ means that it is a function of the observations (the observations can be changed).

Estimate $\hat{\boldsymbol{\theta}}$ means that it is the function value of the corresponding estimator at certain observations.

The Likelihood for a Gaussian AR(1) Process

Consider the Gaussian AR(1) process

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t, \quad t = 0, \pm 1, \pm 2, \ldots \tag{14}$$

with $\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$. We have $\boldsymbol{\theta} = (c, \phi, \sigma^2)$.

By assuming stationarity, for Y_1, we have

$$\begin{aligned} \mathrm{E}(Y_1) = \mu &= c/(1-\phi) \\ \mathrm{E}(Y_1 - \mu)^2 &= \sigma^2/(1-\phi^2) \end{aligned}$$

The Likelihood for a Gaussian AR(1) Process

Since ε_t is Gaussian distributed, Y_1 is also Gaussian

$$Y_1 \sim N(c/(1-\phi), \sigma^2/(1-\phi^2))$$

And hence, the unconditional pdf of Y_1 is

$$f(y_1;\boldsymbol{\theta}) = \left(\frac{2\pi\sigma^2}{1-\phi^2}\right)^{-\frac{1}{2}} \exp\left[-\frac{(y_1 - c/(1-\phi))^2}{2\sigma^2/(1-\phi^2)}\right]$$

The Likelihood for a Gaussian AR(1) Process

Since $Y_2 = c + \phi Y_1 + \varepsilon_2$, the distribution of Y_2 conditional on Y_1 is

$$Y_2 | Y_1 \sim N(c + \phi Y_1, \sigma^2)$$

Then

$$f(y_2 | y_1; \boldsymbol{\theta}) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left[-\frac{(y_2 - (c + \phi y_1))^2}{2\sigma^2}\right]$$

The joint density of y_2 and y_1 is

$$f(y_2, y_1; \boldsymbol{\theta}) = f(y_2 | y_1; \boldsymbol{\theta}) \cdot f(y_1; \boldsymbol{\theta})$$

The Likelihood for a Gaussian AR(1) Process

Similarly, since $Y_3 = c + \phi Y_2 + \varepsilon_3$, the distribution of Y_3 conditional on Y_2 and Y_1 is

$$Y_3 | Y_2, Y_1 \sim N(c + \phi Y_2, \sigma^2)$$

Then

$$f(y_3|y_2, y_1; \boldsymbol{\theta}) = f(y_3|y_2; \boldsymbol{\theta}) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left[-\frac{(y_3 - (c + \phi y_2))^2}{2\sigma^2}\right]$$

The joint density of y_3, y_2 and y_1 is

$$\begin{aligned} f(y_3, y_2, y_1; \boldsymbol{\theta}) &= f(y_3|y_2, y_1; \boldsymbol{\theta}) \cdot f(y_2, y_1; \boldsymbol{\theta}) \\ &= f(y_3|y_2; \boldsymbol{\theta}) \cdot f(y_2|y_1; \boldsymbol{\theta}) \cdot f(y_1; \boldsymbol{\theta}) \end{aligned}$$

The Likelihood for a Gaussian AR(1) Process

By induction, the distribution of Y_t conditional on $Y_{t-1}, ..., Y_1$ is

$$Y_t | Y_{t-1}, ..., Y_1 \sim N(c + \phi Y_{t-1}, \sigma^2)$$

Then

$$f(y_t|y_{t-1}, ..., y_1; \boldsymbol{\theta}) = f(y_t|y_{t-1}; \boldsymbol{\theta}) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left[-\frac{(y_t - (c + \phi y_{t-1}))^2}{2\sigma^2}\right]$$

The joint density of $y_t, ..., y_1$ is

$$f(y_t, ..., y_1; \boldsymbol{\theta}) \quad = \quad f(y_1; \boldsymbol{\theta}) \prod_{i=2}^{t} f(y_i|y_{i-1}; \boldsymbol{\theta}).$$

The likelihood function of $\boldsymbol{\theta}$ given the observations $y_T, ..., y_1$ is $f(y_T, ..., y_1; \boldsymbol{\theta})$.

The Likelihood for a Gaussian AR(1) Process

The log-likelihood function for the stationary Gaussian AR(1) process is given by

$$\log L(\boldsymbol{\theta}) = \log f(y_1; \boldsymbol{\theta}) + \sum_{t=2}^{T} \log f(y_t | y_{t-1}; \boldsymbol{\theta}) \tag{15}$$

Replace each f by the corresponding formula, rearrange...

Somewhat complex? There is another expression for that, in matrix form.

The Likelihood for a Gaussian AR(1) Process

Denote the column vector of observations $\mathbf{y} = (y_1, y_2, ..., y_T)'$ and the random vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_T)'$. $\mathbf{Y}$ is multivariate Gaussian distributed, and $\mathbf{y}$ is one realization of it.

Due to stationarity and Gaussian assumption, the unconditional distribution of each element in $\mathbf{Y}$ is identical, but the elements are correlated.

We have $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu} = (\mu, ..., \mu)'$, where $\mu = c/(1-\phi)$.

The covariance matrix of $\mathbf{Y}$ is defined as

$$\mathrm{E}(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})' = \boldsymbol{\Omega}$$

The Likelihood for a Gaussian AR(1) Process

Note that $\boldsymbol{\Omega}$ has elements

$$\begin{pmatrix} \mathrm{E}(Y_1-\mu)^2 & \mathrm{E}(Y_1-\mu)(Y_2-\mu) & \cdots & \mathrm{E}(Y_1-\mu)(Y_T-\mu) \\ \mathrm{E}(Y_2-\mu)(Y_1-\mu) & \mathrm{E}(Y_2-\mu)^2 & \cdots & \mathrm{E}(Y_2-\mu)(Y_T-\mu) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(Y_T-\mu)(Y_1-\mu) & \mathrm{E}(Y_T-\mu)(Y_2-\mu) & \cdots & \mathrm{E}(Y_T-\mu)^2 \end{pmatrix}$$

In fact, they are autocovariances

$$\boldsymbol{\Omega} = \begin{pmatrix} \gamma_0 & \gamma_{-1} & \cdots & \gamma_{1-T} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{2-T} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{T-1} & \gamma_{T-2} & \cdots & \gamma_0 \end{pmatrix}$$

The Likelihood for a Gaussian AR(1) Process

For a stationary AR(1) process, $\gamma_j = \sigma^2\phi^j/(1-\phi^2)$, which implies that

$$\boldsymbol{\Omega} = \sigma^2\mathbf{V},$$

where

$$\mathbf{V} = \frac{1}{1-\phi^2}\begin{pmatrix} 1 & \phi & \cdots & \phi^{T-1} \\ \phi & 1 & \cdots & \phi^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \cdots & 1 \end{pmatrix}$$

Thus, we can say that $\mathbf{Y} \sim N_T(\boldsymbol{\mu}, \boldsymbol{\Omega})$.

The Likelihood for a Gaussian AR(1) Process

Since $\mathbf{Y} \sim N_T(\boldsymbol{\mu}, \boldsymbol{\Omega})$, we have

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = (2\pi)^{-T/2} |\boldsymbol{\Omega}|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right] \tag{16}$$

with log-likelihood function

$$\log L(\boldsymbol{\theta}) = -\frac{T}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Omega}| - \frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \boldsymbol{\mu}). \tag{17}$$

Remember that $\boldsymbol{\Omega} = \sigma^2 \mathbf{V}$.

The Likelihood for a Gaussian AR(1) Process

Remarks:

- The two expressions are equivalent.
- However, neither of the two expressions have closed form for the ML estimator.
- The estimate has to be found through numerical optimization.
- Section 5.7 in Hamilton gives an introduction to the numerical optimization methods.

Conditional Likelihood for a Gaussian AR(1) Process

Question: What if we consider y_1 as deterministic and maximize the likelihood function conditional on y_1?

An advantage: Since y_1 is considered as non-stochastic, we do not need to assume that $|\phi| < 1$ (stationarity for y). Since $Y_t = c + \phi Y_{t-1} + \varepsilon_t$

$$Y_t | Y_{t-1}, ..., Y_2, y_1 \sim N(c + \phi y_{t-1}, \sigma^2).$$

The conditional distribution of Y_t given all the available past is Gaussian.

Conditional Likelihood for a Gaussian AR(1) Process

The joint density function of $y_T, ..., y_2$ conditional on y_1 is given by

$$f(y_T, ..., y_2 | y_1, \boldsymbol{\theta}) = \prod_{i=2}^{T} f(y_t | y_{t-1}; \boldsymbol{\theta}), \tag{18}$$

where $f(y_t | y_{t-1}$ is the pdf of $N(c + \phi y_{t-1}, \sigma^2)$. We get rid of the annoying $f(y_1 | \boldsymbol{\theta})$.

The log-likelihood function is

$$\log L(\boldsymbol{\theta}) = -\frac{T-1}{2} \log(2\pi) - \frac{T-1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=2}^{T} (y_t - c - \phi y_{t-1})^2 \tag{19}$$

Conditional Likelihood for a Gaussian AR(1) Process

Denote $\boldsymbol{\beta} = (c, \phi)'$ and $\mathbf{x}_t = (1, y_{t-1})'$. The first order condition (FOC) w.r.t. $\boldsymbol{\beta}$ implies that the ML estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left[\sum_{t=2}^{T} \mathbf{x}_t \mathbf{x}_t'\right]^{-1} \left[\sum_{t=2}^{T} \mathbf{x}_t y_t\right], \tag{20}$$

which coincides with the OLS estimator for $\boldsymbol{\beta}$.

The ML estimator of σ^2 is found in the similar way (FOC), by inputting $\hat{\boldsymbol{\beta}}$.

$$\hat{\sigma}^2 = \frac{1}{T-1} \sum_{t=2}^{T} (y_t - \hat{c} - \hat{\phi} y_{t-1})^2 \tag{21}$$

which is the average squared residuals from the OLS regression. Note that $\hat{\sigma}^2$ is not unbiased.

Conditional Likelihood for a Gaussian AR(1) Process

Remarks:

- In contrast to the exact MLE, the conditional MLE offers a closed form solution.
- If the sample size T is large enough, the first observation y_1 makes a negligible contribution to the total likelihood.
- The exact and conditional MLE turn out to have the same asymptotic distribution, provided that $|\phi| < 1$.
- The conditional MLE in this case is consistent for both cases $|\phi| < 1$ and $|\phi| \leq 1$.
- The above results can be extended to AR(p) Gaussian processes.

Conditional Likelihood for a Gaussian MA(1) Process

Consider the Gaussian MA(1) process

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}, \quad t = 0, \pm 1, \pm 2, ... \tag{22}$$

where $\varepsilon_t \sim N(0, \sigma^2)$. We have the parameters $\boldsymbol{\theta} = (\mu, \theta, \sigma^2)'$.

Suppose that $\varepsilon_0 = 0$. Then

$$Y_1 | \varepsilon_0 \sim N(\mu, \sigma^2).$$

Moreover, given y_1 and $\varepsilon_0 = 0$, $\varepsilon_1 = y_1 - \mu$ is also known and

$$(Y_2 | y_1, \varepsilon_0 = 0) \sim N(\mu + \theta\varepsilon_1, \sigma^2)$$

Conditional Likelihood for a Gaussian MA(1) Process

Similarly, given y_1, y_2 and $\varepsilon_0 = 0, \varepsilon_2 = y_2 - \mu - \theta\varepsilon_1$ is also known and

$$(Y_3 | y_2, y_1, \varepsilon_0 = 0) \sim N(\mu + \theta\varepsilon_2, \sigma^2)$$

Proceeding in this fasion, it follows that

$$\begin{aligned} f(y_t | y_{t-1}, ..., y_1, \varepsilon_0 = 0, \boldsymbol{\theta}) &= f(y_t | \varepsilon_{t-1}, \boldsymbol{\theta}) \\ &= (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(y_t - \overbrace{\mu - \theta\varepsilon_{t-1}}^{E[Y_t]})^2}{2\sigma^2} \right] \\ &= (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{\varepsilon_t^2}{2\sigma^2} \right], \quad t = 2, 3, ... \end{aligned} \tag{23}$$

Conditional Likelihood for a Gaussian MA(1) Process

The joint density function conditional on $\varepsilon_0 = 0$ is given by

$$f(y_T, ..., y_1|\varepsilon_0 = 0, \boldsymbol{\theta}) = f(y_1|\varepsilon_0, \boldsymbol{\theta}) \prod_{t=2}^{T} f(y_t|\varepsilon_{t-1}, \boldsymbol{\theta})$$

The log-likelihood function is

$$\begin{aligned} \log L(\boldsymbol{\theta}) &= \log f(y_T, ..., y_1|\varepsilon_0 = 0, \boldsymbol{\theta}) \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t^2 \qquad (24) \end{aligned}$$

where $\varepsilon_t = (y_t - \mu) - \theta\varepsilon_{t-1} = (y_t - \mu) - \theta[(y_{t-1} - \mu) - \theta\varepsilon_{t-2}] = ... = (y_t - \mu) - \theta(y_{t-1} - \mu) + \theta^2(y_{t-2} - \mu) - ... + (-1)^{t-1}\theta^{t-1}(y_1 - \mu) + (-1)^t\theta^t\varepsilon_0.$

Conditional Likelihood for a Gaussian MA(1) Process

Remarks:

- The conditional ML estimator for a Gaussian MA(1) process has no closed form.
- If $|\theta|$ is close enough to zero, the effect of imposing $\varepsilon_0 = 0$ will quickly die out, though it may not be a suitable assumption.
- If $|\theta| > 1$, the conditional approach is not reasonable.
- If the numerical optimization results in $|\hat{\theta}| > 1$, the results must be discarded.
- The above results can be extended to a Gaussian MA(q) process.
- The conditional log-likelihood function for a Gaussian MA(q) process is useful only if the process is invertible.

Conditional Likelihood for a Gaussian ARMA(p, q) Process

Consider the Gaussian ARMA(p, q) process

$$Y_t = c + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \underline{\varepsilon}_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}, \quad t = 0, \pm 1, \pm 2, \ldots \tag{25}$$

where $\varepsilon_t \sim N(0, \sigma^2)$. We have the parameters $\boldsymbol{\theta} = (c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)'$.

Denote

$$\mathbf{y}_0 = (y_0, y_{-1}, \ldots, y_{-p+1})'$$

and

$$\boldsymbol{\varepsilon}_0 = (\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{-q+1})'$$

Conditional on $\mathbf{y}_0$ and $\boldsymbol{\varepsilon}_0$, the sequence of ε_t can be calculated using the observations y_t by iterating

$$\underline{\varepsilon_t} = y_t - \overbrace{c - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q},}^{E(Y_t)} \tag{26}$$

Conditional Likelihood for a Gaussian ARMA(p, q) Process

Likewise, we have the conditional log-likelihood function for a Gaussian ARMA(p, q) process

$$\log L(\boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t^2$$

where ε_t is given recursively by (26).

In order to maximize the log-likelihood, one approach is to assume, as before, that the initial value of y and ε are equal to their expected values.

$$\mathbf{y}_0 = c/(1 - \phi_1 - \ldots - \phi_p)\mathbf{1}_p, \quad \boldsymbol{\varepsilon}_0 = \mathbf{0}_p.$$

Conditional Likelihood for a Gaussian ARMA(p, q) Process

Another approach, according to Box & Jenkins (1976, pp.211), is to set the initial ε equal to zero, but the initial y to their observed values. This means that first p ys become deterministic, and the sample size becomes $T - p$.

The corresponding log-likelihood function becomes

$$\log L(\boldsymbol{\theta}) = -\frac{T - p}{2} \log(2\pi) - \frac{T - p}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=p+1}^{T} \varepsilon_t^2$$

Statistical Model, Likelihood, ML Estimation

- A statistical model for multivariate data x from a sample space S is given by a parametrized family of probability of joint densities $p(x|\theta)$, where the parameter θ is varying in a parameter set Θ. Compact

  $[a, b]$.

- We define the likelihood function by fixing the data x, and consider the density as a function of the parameter:

$$L(\theta) = p(x|\theta), \quad \theta \in \Theta. \tag{27}$$

- We can also write $L(\theta|x)$.
- The Maximum Likelihood estimator is given as the value of θ that maximize the likelihood function

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta), \quad \text{or } \hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta) \tag{28}$$

- In this way, ML estimator becomes a function of the data, hence a random variable.

Hypothesis testing

- A hypothesis concerning the parameter θ is expressed as a restriction on the parameter set Θ: $H_0 : \theta \in \Theta_0$. The null hypothesis.
- The alternative hypothesis $H_1 : \theta \in \Theta / \Theta_0$.
- In most cases, it is more convenient to express the null hypothesis on the parameter as a restriction on θ: $H_0 : g(\theta) = 0$. In this case, $\Theta_0 = \{\theta : g(\theta) = 0\}$.
- The statistical model under the null becomes $\{p(x|\theta),\ \theta \in \Theta_0\}$, or $\{p(x|\theta),\ g(\theta) = 0\}$. The statistical model under the null is nested by the former statistical model $\theta \in \Theta$.

Hypothesis testing

- The ML estimator for the restricted θ is found by maximizing the likelihood function under the constraint specified by Θ_0. Denote by $\tilde{\theta}$.
- We have the Likelihood Ratio test statistic: $Q(x) = L(\tilde{\theta})/L(\hat{\theta})$.
- p-value is defined to be: $pv = \sup_{\theta \in \Theta_0} \mathrm{Prob}\{Q(x) \leq Q(x_{obs})\}$.
- Sig. level and critical value c_α: $\sup_{\theta \in \Theta_0} \mathrm{Prob}\{Q(x) \leq c_\alpha)\} = \alpha$.
- Power of the test: $\mathrm{Prob}(Q(x) \leq c_\alpha)$, given $\theta \in \Theta/\Theta_0$.

Likelihood, score and information

- Given the pair $x = (x_1, ..., x_T)$ and θ, assuming that θ is a m-dimensional vector, we have the $m \times 1$ score vector:

$$S_T(\theta) = \mathrm{d} \log p(x|\theta)/\mathrm{d}\theta \tag{29}$$

and the $m \times m$ information matrix:

$$I_T(\theta) = -\mathrm{d}S_T/\mathrm{d}\theta = -\mathrm{d}^2 \log p(x|\theta)/\mathrm{d}\theta^2 \tag{30}$$

- We have:

$$\begin{aligned} \mathrm{E}_\theta(S_T(\theta)) &= 0 && (31) \\ \mathrm{E}_\theta(I_T(\theta)) &= \mathrm{E}_\theta(S_T(\theta)S_T(\theta)') = \mathrm{Var}(S_T(\theta)). && (32) \end{aligned}$$

Proof

Obviously we have:

$$\begin{aligned}\int p(x|\theta)\mathrm{d}x &= 1 &(33)\\ \int \frac{\partial p(x|\theta)}{\partial\theta}\mathrm{d}x &= 0 &(34)\\ \int \frac{\partial^2 p(x|\theta)}{\partial\theta^2}\mathrm{d}x &= 0 &(35)\end{aligned}$$

Note that $S_T(\theta) = \frac{\partial \log p(x|\theta)}{\partial\theta} = p(x|\theta)^{-1}\frac{\partial p(x|\theta)}{\partial\theta}$. Then
$\mathrm{E}_\theta(S_T(\theta)) = \int p(x|\theta)^{-1}\frac{\partial p(x|\theta)}{\partial\theta}p(x|\theta)\mathrm{d}x = 0$.
And $I_T(\theta) = -\frac{\partial^2 \log p(x|\theta)}{\partial\theta^2} = -p(x|\theta)^{-1}\frac{\partial^2 p(x|\theta)}{\partial\theta^2} + \frac{\partial \log p(x|\theta)}{\partial\theta}(\frac{\partial \log p(x|\theta)}{\partial\theta})'$.
Thus,
$\mathrm{E}_\theta(I_T(\theta)) = \mathrm{E}_\theta(\frac{\partial \log p(x|\theta)}{\partial\theta}(\frac{\partial \log p(x|\theta)}{\partial\theta})') = E_\theta(S_T(\theta)S_T(\theta)') = \mathrm{Var}(S_T(\theta)$.

The three tests

- Suppose that θ_0 is the true value of the parameter θ.
- The likelihood ratio test:

$$Q(x) = -2\log\frac{L(\theta_0)}{L(\hat{\theta})} = 2(\log(L(\hat{\theta}) - \log(L(\theta_0)) \tag{36}$$

- The Wald test:

$$\hat{\theta} \sim N(\theta_0, I_T(\hat{\theta})^{-1}) \tag{37}$$

- The Lagrange-multiplier or score test:

$$S_T(\theta_0) \sim N(0, I_T(\theta_0)) \tag{38}$$

The relations between the three tests

Stochastic Taylor expansion:

$$\log L(\theta_0) = \log L(\hat{\theta}) + (\theta_0 - \hat{\theta})' S_T(\hat{\theta}) - \frac{1}{2}(\theta_0 - \hat{\theta}) I_T(\hat{\theta})(\theta_0 - \hat{\theta}) + \ldots \tag{39}$$

Since $S_T(\hat{\theta}) = 0$, Taylor again $0 = S_T(\hat{\theta}) = S_T(\theta_0) - I_T(\theta_0)(\hat{\theta} - \theta_0) + \ldots$, and hence

$$S_T(\theta_0) = I_T(\theta_0)(\hat{\theta} - \theta_0) + \ldots \tag{40}$$

We have the relations:

$$-2\log\frac{L(\theta_0)}{L(\hat{\theta})} = (\theta_0 - \hat{\theta}) I_T(\hat{\theta})(\theta_0 - \hat{\theta}) + \ldots = S_T(\theta_0)' I_T(\theta_0)^{-1} S_T(\theta_0) + \ldots \tag{41}$$

$\sim \chi^2(m)$

To be continued! Thank you!