Massachusetts Institute of Technology
Department of Economics
Time Series 14.384

Guido Kuersteiner

Lecture 5: Estimation and Specification of ARMA Models

We first consider estimators for AR(p) models. Assume that x_t is generated by

$$x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + \varepsilon_t$$

where ε_t is a martingale difference sequence, i.e. (or white noise plus mixing plus moment restrictions) $E_{\mathcal{M}_{t-1}}\varepsilon_t = 0$ where $\mathcal{M}_t$ contains all measurable functions of $\{x_s, s \leq t\}$. This assumption is stronger than the WN assumption previously made. Stack $z_t' = (x_{t-1}, ..., x_{t-p})$ then the OLS estimator for ϕ is given by

$$\begin{aligned} \hat{\phi} &= \left(\sum_{t=p+1}^{T} z_t z_t'\right)^{-1} \sum_{t=p+1}^{T} z_t x_t \\ &= \phi + \left(\sum_{t=p+1}^{T} z_t z_t'\right)^{-1} \sum_{t=p+1}^{T} z_t \varepsilon_t \end{aligned}$$

Then by assuming that a WLLN holds, it follows that

$$\frac{1}{T}\sum z_t z_t' \xrightarrow{p} \Gamma$$

where

$$\Gamma = \begin{bmatrix} \gamma_x(0) & \dots & \gamma_x(p-1) \\ & \ddots & \vdots \\ & & \gamma_x(0) \end{bmatrix}.$$

The WLLN holds for example if z_t is strictly stationary, $Ez_t z_t^{'} = \Gamma$ and an additional technical condition, called ergodicity holds.

We turn to the asymptotic distribution next. First note that $z_t\varepsilon_t$ is a martingale difference sequence as well. Then, if $\sup_t E\left\|z_t\varepsilon_t\right\|^{2+\delta} < \infty$ and $\frac{1}{T}\sum(z_t z_t'\varepsilon_t^2 - Ez_t z_t'\varepsilon_t^2) \xrightarrow{p} 0$ then we can apply a martingale difference CLT to show that

$$\frac{1}{\sqrt{T}}\sum_{t=1+p}^{T} z_t\varepsilon_t \xrightarrow{d} N\left(0, \lim_T \frac{1}{T}\sum_{t=1}^{T} E(z_t z_t'\varepsilon_t^2)\right)$$

If in addition $E_{\mathcal{M}_{t-1}}\varepsilon_t^2 = \sigma^2$ then $Ez_t z_t'\varepsilon_t^2 = \sigma^2\Gamma$. Therefore the asymptotic distribution for $\hat{\phi}$ is given by $\sqrt{T}\left(\hat{\phi} - \phi\right) \xrightarrow{d} N\left(0, \sigma^2\Gamma^{-1}\right)$

5.1. ML-Estimation

The maximum likelihood estimator is defined as the value maximizing

$$f\left(x_1, ..., x_T; \psi\right)$$

over ψ, where $f(.\,|\psi)$ is the joint distribution of $\{x_1,...,x_T\}$. If $X_T^{'}=(x_1,...,x_T)$ is a Gaussian time series then the likelihood function is given by

$$f\left(x_1,...,x_T;\psi\right)=\frac{1}{(2\pi)^{T/2}}\det(\Gamma_T(\psi))^{-1/2}\exp(-\frac{1}{2}X_T^{'}\Gamma_T^{-1}(\psi)X_T) \tag{5.1}$$

where $\Gamma_T(\psi)=EX_TX_T^{'}$ is the $T\times T$ covariance matrix of X_T. The covariance matrix is a nonlinear function of the underlying parameters. Maximizing (5.1) directly is therefore a highly nonlinear optimization problem. The problem can be simplified by considering the conditional densities of x_t. We can then write the joint density as a product of conditional densities

$$\begin{aligned} f\left(x_1,...,x_T;\psi\right) &= f(x_1)\cdot f(x_2\,|x_1\,)\cdot f(x_3\,|x_2,x_1\,)... \\ &\quad ...f(x_T\,|x_1...x_{T-1}\,) \end{aligned}$$

If x_t is a Gaussian process the conditional densities are all normal with conditional mean of $x_t=P_{\mathcal{M}_{t-1}^{\ell}}x_t$ and conditional variance of x_t equal to $\sigma_t^2=\left\|x_t-P_{\mathcal{M}_{t-1}^{\ell}}x_t\right\|^2$. We have seen in Lecture 5, how these expressions can be computed recursively. It therefore follows that the exact likelihood, assuming Gaussianity, can be computed in a recursive way for each set of parameter values ψ. In particular we can avoid numerical inversion of the $T\times T$ matrix $\Gamma_T(\psi)$.

In particular cases the situation simplifies even further. If we specify for example that $\varepsilon_t\sim N(0,\sigma^2)$ and

$$x_t=\phi_1x_{t-1}+\varepsilon_t$$

then

$$f(x_t\,|x_{t-1}\,)=\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left(-\frac{1}{2\sigma^2}\left(x_t-\phi x_{t-1}\right)^2\right)$$

such that

$$\begin{aligned} f(x_1,..,x_T;\psi) &= \frac{1}{(2\pi)^{T/2}}\frac{1}{\sigma^T}\exp\left(-\frac{1}{2\sigma^2}\sum_{t=2}^{T}\left(x_t-\phi x_{t-1}\right)^2\right) \\ &\quad \times\exp\left(-\frac{1}{2}\frac{x_1^2}{\sigma^2(1-\phi^2)^{-1}}\right). \end{aligned}$$

Taking the log of the likelihood function we have

$$\log f(x_1,..,x_T;\psi)=-T\log\sigma-\frac{1}{2\sigma^2}\sum_{t=2}^{T}\left(x_t-\phi x_{t-1}\right)^2-\frac{1}{2}\frac{x_1^2}{\sigma^2(1-\phi^2)^{-1}}. \tag{5.2}$$

If we ignore the last term and maximize $\log f(x_1,..,x_T;\psi)$ with respect to ϕ, we see that the ML estimator is asymptotically equivalent to OLS. This result was derived under the assumption that ε_t is Gaussian. If Gaussianity does not hold we can still use (5.2) as the criterion function. In this case the estimator is called a Quasi Maximum Likelihood estimator. It can be shown that under certain conditions, including that $E_{\mathcal{M}_{t-1}}\varepsilon_t^2=\sigma^2$, the resulting estimator has the same asymptotic distribution as if the errors were indeed normal.

In other cases it is useful to approximate the exact innovations updating formulas. In particular, we have seen that the ARMA(1,1) case can be handled by looking at the limiting behavior of the projection coefficients. For the ARMA(1,1) case parametrized by the polynomials $\phi(L)=(1-\phi L)$ and $\theta(L)=(1-\theta L)$ we use therefore the following approximate formulation for the likelihood function

$$f\left(x_t\,|x_{t-1},...\,\right)\approx\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left(-\frac{1}{2\sigma^2}\left(x_t-\phi x_{t-1}+\theta\varepsilon_{t-1}\right)^2\right)$$

the full likelihood is obtained by setting $\varepsilon_0 = 0$. We have now

$$\begin{aligned} f(x_1, ..., x_T; \psi) &\approx \frac{1}{(2\pi)^{T/2}} \frac{1}{\sigma^T} \exp\left(-\frac{1}{2\sigma^2} \sum_{t=2}^{T} (x_t - \phi x_{t-1} + \theta \varepsilon_{t-1})^2 \right) \\ &\times \exp\left(-\frac{1}{2} \frac{x_1^2}{\sigma^2 c} \right) \end{aligned}$$

where $c = (1 + 2\theta\phi + \theta^2)/(1 - \phi^2)$.

Maximizing $f(x_1, ..., x_T, \psi)$ with respect to ϕ and θ is equivalent to minimizing the sum

$$S_T(\theta, \phi) = \sum_{t=2}^{T} (x_t - \phi x_{t-1} + \theta \varepsilon_{t-1})^2 + \frac{x_1^2}{c^2}$$

The last term is typically left away since it has no effect asymptotically. The sum $S(\phi, \theta)$ can be evaluated for all values ϕ and θ by computing the residuals ε_t recursively, i.e.,

$$\begin{aligned} \hat{\varepsilon}_1 &= x_1 \\ \hat{\varepsilon}_2 &= x_2 - \phi x_1 + \theta x_1 = x_2 + (\theta - \phi) x_1 \\ \hat{\varepsilon}_t &= x_t - \phi x_{t-1} + \theta \widehat{\varepsilon}_{t-1} \end{aligned}$$

We can therefore use numerical algorithms to evaluate $S_T(\phi, \theta)$ at different values of ϕ, θ.

More generally the ML estimator for the ARMA(p,q) class parametrized by $\phi(L) = (1 - \phi_1 L - ... - \phi_p L^p)$ and $\theta(L) = (1 - \theta_1 L - ... - \theta_q L^q)$of models can be written as

$$f(x_1, ..., x_T; \psi) \approx \frac{1}{(2\pi)^{T/2}} \frac{1}{\sigma^T} \exp\left(-\frac{1}{2} \sum_{t=m+1}^{T} \varepsilon_t^2 \right) \exp\left(-\frac{1}{2} X_m' \Gamma_m^{-1} X_m \right)$$

where $m = \max(p, q)$, $X_m' = (x_1, ..., x_m)$ and $\Gamma_m = E X_m X_m'$. The errors can be approximated again by

$$\begin{aligned} \hat{\varepsilon}_1 &= x_1 \\ \hat{\varepsilon}_2 &= x_2 - \phi_1 x_1 + \theta_1 \hat{\varepsilon}_1 \\ \hat{\varepsilon}_t &= x_t - \phi_1 x_{t-1} - ... - \phi_p x_{t-p} + \theta_1 \widehat{\varepsilon}_{t-1} + ... + \theta_q \widehat{\varepsilon}_{t-q} \end{aligned}$$

A further approximation step then uses

$$f(x_1, ..., x_T; \psi) \approx \frac{1}{(2\pi)^{T/2}} \frac{1}{\sigma^T} \exp\left(-\frac{1}{2} \sum_{t=m+1}^{T} \hat{\varepsilon}_t^2 \right)$$

to estimate the parameters.

5.2. Asymptotic Distribution of ML-estimators

It can be shown that estimators minimizing the criterion function $S_T(\phi, \theta)$ are consistent and asymptotically normal. More generally, let $Q_T(\beta) = \log f(x_1, ..., x_T; \psi)$. Then consistency follows if for some set C,

$$\sup_{\psi \in C} |Q_T(\psi) - Q(\psi)| \to 0 \tag{5.3}$$

in probability, where $Q(\psi)$ is a nonstochastic function. Moreover, we also need that for any $\delta > 0$ and a neighborhood $N(\psi_0, \delta)$

$$\sup_{\psi \in C \backslash N(\psi_0, \delta)} Q(\psi) < Q(\psi_0) \tag{5.4}$$

For the case of an $ARMA$ model we have $\psi = (\beta, \sigma)$ where $\beta = \left(\phi_1, ..., \phi_p \theta_1, ..., \theta_q\right)$. It can be shown that for the ARMA class the set C that satisfies (5.4) when $Q_T(\psi)$ is the Gaussian likelihood is

$$C = \left\{\beta \in \mathbb{R}^{p+q} \mid \phi(z)\theta(z) \neq 0 \text{ for } |z| \leq 1, \, \phi_p \neq 0, \, \theta_q \neq 0, \, \phi(z) \text{ and } \theta(z) \text{ have no common zeros}\right\}$$

Note that σ is identified once $\beta \in C$.

In words this condition means, that the AR and MA polynomials should have no common zeros, should both have roots outside the unit circle and should be of order p and q respectively in a nontrivial way. In particular this means that the coefficient on the highest order lag in both polynomials should be nonzero. We give an example of an MA model that has the same autocovariance function for two different parameter values for θ. You should check that only one of the models is contained in C.

Example 5.1. *The MA(1) models*

$$x_t = \varepsilon_t + \theta \varepsilon_{t-1}$$

and

$$x_t = \varepsilon_t + \frac{1}{\theta} \varepsilon_{t-1}$$

are observationally equivalent in the sense that they imply the same autocovariance function.

If conditions (5.3) and (5.4) are satisfied and if

$$\hat{\psi}_T = \underset{\psi \in C}{\arg\min} \, Q_T(\psi) + o_p(1)$$

then it follows that $\hat{\psi} \to \psi_0$ in probability. Conditions (5.3) and (5.4) can be shown to hold for the $ARMA$ model with Gaussian criterion function. The proofs are somewhat complicated because C is not a compact set. We will omit them here.

A consistency result is usually the first step in deriving the asymptotic distribution of an estimator. A second step consists in showing that $\hat{\psi}_T$ is contained in a $1/\sqrt{T}$ neighborhood of the true parameter with high probability. A Taylor expansion in the neighborhood of the true parameter ψ_0 is then used to obtain the asymptotic distribution of the estimator. It can be shown that

$$\sqrt{T} \left(\hat{\beta}_T - \beta\right) \xrightarrow{d} N\left(0, V(\beta)\right)$$

where

$$V(\beta) = \sigma^2 E \begin{bmatrix} U_t U_t' & U_t V_t' \\ V_t U_t' & V_t V_t' \end{bmatrix}^{-1}$$

with $u_t = \phi(L)^{-1} \varepsilon_t$ and $v_t = \theta(L)^{-1} \varepsilon_t$. The limiting covariance matrix $V(\beta)$ can then be expressed in terms of $U_t = [u_t, u_{t-1}, ..., u_{t-p+1}]$and $V_t = [v_t, v_{t-1}, ..., v_{t-q+1}]$. Note that the limiting covariance matrix does not depend on σ^2.

Example 5.2 ($ARMA(1,1)$)**.**

$$x_t = \phi_1 x_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$

so $u_t = \sum_{j=0}^{\infty} \phi_1^j \varepsilon_{t-j}$ and $v_t = \sum_{j=0}^{\infty} \theta_1^j \varepsilon_{t-j}$. From this it follows that $Ev_t^2 = \sigma^2 \frac{1}{1-\phi_1^2}$, $Ev_t^2 = \sigma^2 \frac{1}{1-\theta_1^2}$ and $Eu_t v_t = \sigma^2 \frac{1}{1-\phi_1 \theta_1}$.

5.3. Order Selection

Before an ARMA(p,q) model can be estimated we need to select the order p and q of the AR and MA-polynomial. We have seen in Lecture 2 that in principle the autocorrelation and partial autocorrelation function characterize pure $AR(p)$ and $MA(q)$ models.

For a pure $MA(q)$ process the variance of the jth autocorrelation coefficient is given by

$$\operatorname{var}\left(\hat{\rho}_T(j)\right) = \frac{1}{T}\left(1 + 2\sum_{i=1}^{q} \rho^2(i)\right)$$

which can be estimated by

$$\hat{\sigma}\left(\hat{\rho}(j)\right) = \frac{1}{\sqrt{T}}\left((1+2)\sum_{i=1}^{q} \hat{\rho}(i)^2\right)^{1/2}$$

where

$$\rho(i) = \frac{\operatorname{cov}(x_{t_1} x_{t+|j|})}{\sqrt{var(x_t)}\sqrt{var(x_{t-|j|})}}.$$

In order to identify the degree of the MA polynomial we can check for which value of h the estimated autocorrelation coefficient $\hat{\rho}(h)$ stays within

$$\pm\frac{1.96}{\sqrt{T}}\left(1 + 2\widehat{\rho}(1) + ... + \widehat{\rho}(h-1)\right)$$

In the same way we can identify a pure $AR(p)$ model from its partial autocorrelation function. It can be shown that for a pure $AR(p)$ model the partial autocorrelations $\hat{\phi}(n)$ for $n > p$ have variance $\frac{1}{T}$. We can thus check for which value of j the estimated coefficient $\hat{\phi}(j)$ lies within $\pm\frac{1.96}{\sqrt{T}}$.

If the model is likely to be a mixed $ARMA(p,q)$ model then the above identification procedure runs into difficulties. We can still look at autocorrelation and partial autocorrelation functions of the data to gain some insight into the maximal degree of the two lag polynomials. There is however a more formal procedure, based on information criteria, that can be used to determine the best model in an automated way.

If the process $\{x_t\}_{t=1}^{T}$ has a true density $f(x,\psi_0)$ and the ARMA-class has densities $f(x,\theta)$ then the Kullback-Leibler distance is

$$\begin{aligned} d\left(\psi_0 \,|\theta\,\right) &= \int_{\mathbb{R}^T} -2\ln\left(\frac{f(x,\psi_0)}{f(x,\theta)}\right) f(x,\theta)dx \\ &\geq -2\ln\int \frac{f(x,\psi_0)}{f(x,\theta)} f(x,\theta)dx \\ &= -2\ln\int f(x,\psi_0)dx = 0, \end{aligned}$$

where $d(\psi_0\,|\theta\,) = 0$ if and only if $f(x,\psi_0) = f(x,\theta)$ a.e. The distance measure $d(\psi_0\,|\theta\,)$ can be approximated by

$$\text{AIC}(p,q) = \ln\hat{\sigma}^2 + 2(p+q)/T,$$

where $\hat{\sigma}^2 = \frac{1}{T}\sum\hat{\varepsilon}_t^2$ is the maximum likelihood estimator of σ^2. The best model specification is that found by calculating $\text{AIC}(p,q)$ for different values of p and q and picking the combination (p^*,q^*) such that $\text{AIC}(p,q)$ is minimized.

Increasing the number p,q reduces the value of $\hat{\sigma}^2$. This comes at a cost of overparametrizing the model which is captured in the term $2(p+q)/T$. It can be shown that AIC is inconsistent in the sense that it asymptotically picks p and q too large.

A modified criterion, called BIC, does not suffer from this problem. It is defined as

$$\mathrm{BIC}(p,q) = \ln \widehat{\sigma}^2 + (p+q)\ln T/T.$$

5.4. Diagnostic Checking

Once we have determined the values for p and q the model can be estimated with the methods of the previous section. It is a sensible strategy to start with low-order models and then test against increasing the order of an AR or MA polynomial by one. Note that one should never increase the AR and MA polynomial at the same time.

Assume we have already estimated an ARMA(1,1) model, and we want to test whether an ARMA(2,1) or ARMA(1,2) is more appropriate. One way to proceed is to estimate both the ARMA(2,1) and ARMA(1,2) model and then test whether the additional coefficient is significantly different from zero. In particular, we choose the ARMA(1,1) if

$$\frac{\left|\hat{\phi}_2\right|}{\sqrt{\mathrm{var}(\hat{\phi}_1)}} < 1.96 \text{ and } \frac{|\theta_2|}{\sqrt{\mathrm{var}(\hat{\theta}_2)}} < 1.96.$$

Note that the variances of the parameter estimates should be determined from the null distribution, i.e. under the assumption that the true parameter value is zero.

An alternative procedure is to test if the residuals are white noise. If the estimated model is correctly specified then the time dependence in the data should be captured by the model and the residuals should be uncorrelated. If we obtain residuals from

$$\hat{\varepsilon}_t = x_t - \hat{\phi}_1 x_{t-1} - \ldots - \hat{\phi}_p x_{t-p} + \theta_1 \hat{\varepsilon}_{t-1} + \ldots + \theta_q \hat{\varepsilon}_{t-q}$$

and calculate

$$\hat{\gamma}_\varepsilon \varepsilon(j) = \frac{1}{T}\sum_{t=1}^{T-j} \hat{\varepsilon}_t \hat{\varepsilon}_{t+|j|}$$

then $\rho_\varepsilon(j) = \hat{\gamma}_\varepsilon(j)/\hat{\gamma}_\varepsilon(0)$ should be close to zero for all j. A popular test of this hypothesis is the Portmanteau or Box-Pierce test. It is based on

$$Q = T\sum_{j=1}^{K} \widehat{\rho}_\varepsilon^2(j).$$

It can be shown that under $H_0 = \rho(j) = 0 \; \forall j$ the limit distribution of Q is $Q \sim \chi^2_{K-(p+q)}$. The reduction in degrees of freedom by $p+q$ in the asymptotic distribution is determined by the number of parameters used in the estimation of the model. In practical applications K should be chosen at least 15 to 20. There is however a trade off between low power of the test for K too large and inconsistency of the test for K too small. In practice it is therefore advisable to look at a plot of $\widehat{\rho}_\varepsilon(j)$ before applying the test. Box and Ljung show that the statistic

$$\tilde{Q} = T(T+2)\sum_{j=1}^{K} \frac{\hat{\rho}_\varepsilon(j)^2}{T-j} \xrightarrow{d} \chi^2_{K-p-q}$$

has less bias than Q relative to the asymptotic χ^2 distribution in small samples.