

Automatic Information Criteria - a brief overview

Barry G. Quinn
Statistics Dept
Macquarie University

1 Hypothesis testing

- Suppose the N -dimensional random vector Y has joint pdf or pf $f(y; \theta)$, where $\theta \in \Theta \subset \mathbb{R}^p$.
- We wish to test $H_0 : \theta \in \Theta_0$ against $H_A : \theta \in \Theta_A$, where $\Theta_0 \subset \Theta_A$, and Θ_0 is of lower dimension than Θ_A .
- e.g., $H_0 : \theta_{q+1} = \dots = \theta_p = 0$.
- The log-likelihood function is

$$l(\theta; Y) = \log f(Y; \theta)$$

and the likelihood ratio statistic is

$$\lambda = 2 \sup_{\theta \in \Theta_A} l(\theta; Y) - 2 \sup_{\theta \in \Theta_0} l(\theta; Y).$$

- Often

$$\lim_{N \rightarrow \infty} \Pr(\lambda \leq x) = \Pr(\chi_{p-q}^2 \leq x),$$

where $q = \dim \Theta_0$.

- Rejecting H_0 when λ is too large is known as the likelihood ratio principle.

2 Linear regression

- Model

$$Y = X\beta + \varepsilon$$

where

- β is $p \times 1$ vector of (unknown) parameters;
 - X is $n \times p$ ‘design matrix’;
 - ε is $n \times 1$ vector of uncorrelated ‘error’ random variables.
- We have *one* observation on Y , and must use this to estimate β .

- p, X assumed to be known.
- Problem: which columns of X should be included in the model.
- Equivalent to determining which elements of β are 0.
- Consider two models, where columns of X_1 ($n \times q$) are included as columns of X_2 ($n \times p$).
- Which model is better?

- The ‘larger’ model ought to be better, as more information is used, but would the ‘smaller’ model suffice?
- If we assume $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 I)$, the likelihood ratio principle suggests rejecting the smaller in favour of the larger if

$$T = \frac{S_1 - S_2}{Y'Y - S_2} > k$$

where the ‘residual sum of squares’ for model j is given by

$$S_j = Y' \left\{ I - X_j (X_j' X_j)^{-1} X_j' \right\} Y.$$

- The exact distribution when the smaller model is correct (i.e. the null hypothesis is true) is known.

- k is chosen so that $\Pr(T > k|H_0) = \alpha$, for suitable α (probability of false alarm, $\alpha = 0.05$ is common).
- We thus choose the larger model if S_2 is sufficiently smaller than S_1 .

3 Autoregressive processes

- Model

$$\sum_{j=0}^p \beta_j X_{t-j} = \varepsilon_t,$$

$$\beta_0 = 1, E(\varepsilon_t) = 0, E(\varepsilon_t^2) = \sigma^2$$

$$\sum_{j=0}^p \beta_j z^j \neq 0, |z| \leq 1$$

- We need to impose some conditions on the ε_t .
- Common in literature to see Gaussian i.i.d. conditions assumed.

- Gaussianity and/or independence are actually rarely needed.
- All that is needed is that $\{\varepsilon_t\}$ be stationary and ergodic, with

$$\begin{aligned} E(\varepsilon_t | \varepsilon_{t-1}, \varepsilon_{t-2}, \dots) &= 0 \\ E(\varepsilon_t^2 | \varepsilon_{t-1}, \varepsilon_{t-2}, \dots) &= \sigma^2. \end{aligned}$$

- Fitting autoregressions: If p is known, Yule-Walker relations may be used to estimate $\beta = [\beta_1 \ \dots \ \beta_p]'$.
- The estimator is (asymptotically) equivalent to the least squares estimator of β .

- Given X_0, \dots, X_{T-1} , minimise

$$\begin{aligned} S(\beta) &= \sum_{t=0}^{T-1} \varepsilon_t^2(\beta) \\ &= \sum_{t=0}^{T-1} \left(\sum_{j=0}^p \beta_j X_{t-j} \right)^2. \end{aligned}$$

- The estimator is, in turn, equivalent to the (conditional) Gaussian maximum likelihood estimator, i.e. the maximiser of the logarithm of

$$l = -T/2 \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} S(\beta).$$

- The Yule-Walker relations are often used to estimate β , (via the Levinson-Durbin

algorithm). Let

$$C_j = T^{-1} \sum_{t=j}^{T-1} (X_t - \bar{X}) (X_{t-j} - \bar{X}).$$

- Then if k is the true order, $\hat{\beta}^k = [\hat{\beta}_1^k \ \dots \ \hat{\beta}_k^k]'$ and $\hat{\sigma}_k^2 = C_0 + \sum_{j=1}^k C_j \hat{\beta}_j^k$, we have

$$\begin{aligned} \hat{\beta}_{k+1}^{k+1} &= \frac{C_{k+1} + \sum_{j=1}^k C_{k+1-j} \hat{\beta}_j^k}{\hat{\sigma}_k^2} \\ \hat{\beta}_j^{k+1} &= \hat{\beta}_j^k + \hat{\beta}_{k+1}^{k+1} \hat{\beta}_{k+1-j}^k, \quad j = 1, \dots, k \\ \hat{\sigma}_{k+1}^2 &= \hat{\sigma}_k^2 \left\{ 1 - \left(\hat{\beta}_{k+1}^{k+1} \right)^2 \right\}. \end{aligned}$$

- This equation is the most useful for understanding the behaviour of AIC and similar procedures.

- The recursion starts with $\hat{\sigma}_0^2 = C_0$.

- Let

$$\sigma_k^2 = \min_{\alpha_1, \dots, \alpha_k} E \left(\sum_{j=0}^k \alpha_j X_{t-j} \right)^2,$$

where $\alpha_0 = 1$.

- Then $\sigma_p^2 = \sigma^2$, and $\sigma_{k+1}^2 = \sigma_k^2 \left\{ 1 - \left(\beta_{k+1}^{k+1} \right)^2 \right\}$, with β_j^k defined in obvious way.

- Thus $\sigma_{k+1}^2 \leq \sigma_k^2$, equality only when $\beta_{k+1}^{k+1} = 0$.

- $\beta_j^k = 0, \forall j > p$, when $k > p$.

- $\forall k$

$$\begin{array}{ccc} \hat{\sigma}_{k+1}^2 & < & \hat{\sigma}_k^2 \\ \hat{\sigma}_k^2 & \xrightarrow{a.s.} & \sigma_k^2 \end{array}$$

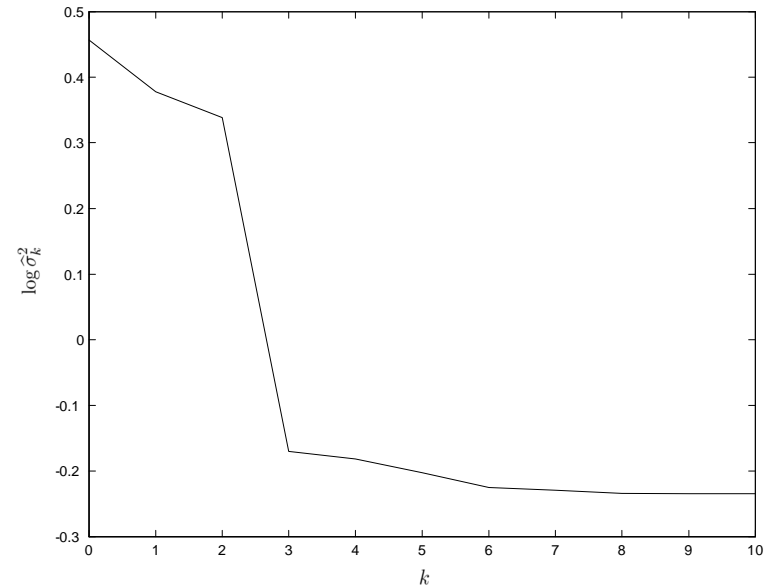
- We know $\beta_p^p = \beta_p \neq 0$. Thus

$$\frac{\hat{\sigma}_p^2}{\hat{\sigma}_{p-1}^2} \xrightarrow{a.s.} \frac{\sigma_p^2}{\sigma_{p-1}^2} = 1 - (\beta_p)^2 < 1.$$

- Also, for $k > p$,

$$\frac{\hat{\sigma}_k^2}{\hat{\sigma}_p^2} \xrightarrow{a.s.} 1$$

- Because of multiplicative nature, and wish for scale invariance, we consider $\log \hat{\sigma}_k^2$.
- Plot of $f(k) = \log \hat{\sigma}_k^2$. True $p = 3$.
- Note jump at $k = 2$ and slow decrease from $k = 3$.



3.1 AIC

- Akaike (1969) proposed that k be estimated by minimising a penalised version of $f(k)$:

$$\phi_{FPE}(k) = \log \hat{\sigma}_k^2 + \log \left(\frac{T+k}{T-k} \right)$$

- This is asymptotically equivalent to

$$\phi(k) = \log \hat{\sigma}_k^2 + \frac{2k}{T}, \quad (1)$$

which has become known as the AIC (Automatic Information Criterion), although AIC is actually more general.

- The idea is that $2k/T$ increases with k , and hopefully more than compensates for the slow decrease in $\log \hat{\sigma}_k^2$ after the true order is reached. We thus find the minimiser of $\phi_{FPE}(k)$ or $\phi(k)$ and use this to estimate p .
- Why might the procedure work? Why $2k/T$?

- From above,

$$\log \hat{\sigma}_{k+1}^2 = \log \hat{\sigma}_k^2 + \log \left\{ 1 - \left(\hat{\beta}_{k+1}^{k+1} \right)^2 \right\}.$$

- But

$$\log \hat{\sigma}_{k+1}^2 - \log \hat{\sigma}_k^2 = \log \left\{ 1 - \left(\hat{\beta}_{k+1}^{k+1} \right)^2 \right\}$$

$$\left\{ \begin{array}{l} \rightarrow \log \left\{ 1 - \left(\beta_{k+1}^{k+1} \right)^2 \right\} \leq 0 \quad ; \quad k < p - 1 \\ \rightarrow \log \left\{ 1 - \left(\beta_p \right)^2 \right\} < 0 \quad ; \quad k = p - 1 \\ \sim - \left(\hat{\beta}_{k+1}^{k+1} \right)^2 \quad ; \quad k \geq p \end{array} \right.$$

- Thus

$$\phi_{FPE}(k+1) - \phi_{FPE}(k) \begin{cases} \rightarrow \log \left\{ 1 - \left(\beta_{k+1}^{k+1} \right)^2 \right\} \leq 0 & ; k < p-1 \\ \rightarrow \log \left\{ 1 - \left(\beta_p \right)^2 \right\} < 0 & ; k = p-1 \\ \sim - \left(\hat{\beta}_{k+1}^{k+1} \right)^2 + \frac{2}{T} & ; k \geq p. \end{cases}$$

- The last equation provides the key to the (asymptotic) behaviour of AIC. It can be shown that under the conditions above, for $k \geq p$,

$$\Pr \left(\sqrt{T} \hat{\beta}_{k+1}^{k+1} \leq x \right) \rightarrow \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}u^2} du.$$

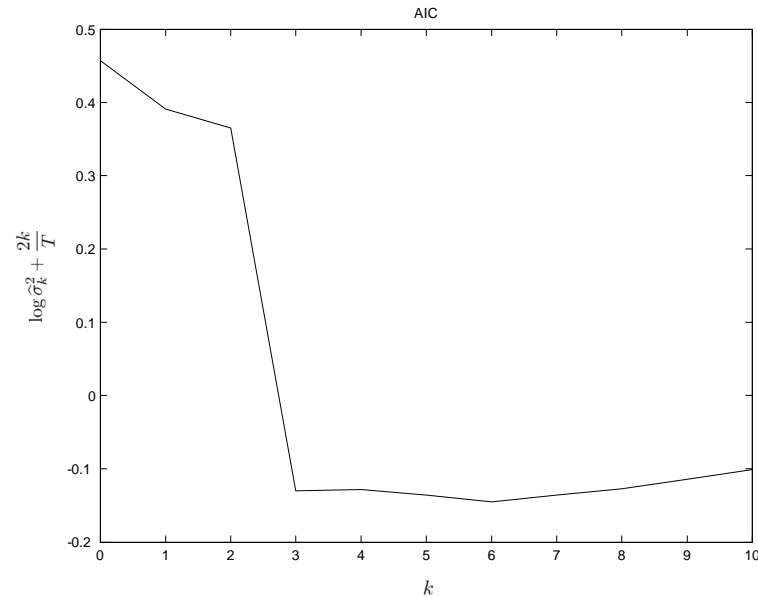
- Note: the exact distribution of $\sqrt{T} \hat{\beta}_{k+1}^{k+1}$ is unknown, even when the ε_t are

Gaussian. As well, the exact moments are unknown.

- Thus, in particular

$$\Pr \{ \phi_{FPE}(p+1) < \phi_{FPE}(k) \} \sim \Pr \left\{ T \left(\hat{\beta}_{k+1}^{k+1} \right)^2 > 2 \right\}$$
$$\rightarrow 1 - \Phi(\sqrt{2}) > 0.$$

- Thus AIC cannot estimate p consistently.
- Noticed by simulation by econometricians, and proved by Shibata (1977).
- Here is a plot of AIC for the above example



- Note that AIC starts to increase at $k = 3$ but then decreases, and the minimiser is actually 6.
- In fact, if we denote the minimiser of AIC or FPE over $k \leq K$ by \hat{p} , then

(Quinn(1980))

$$\begin{aligned} \lim_{K \rightarrow \infty} \Pr \{ \hat{p} = p \} &= \exp \left\{ - \sum_{k=1}^{\infty} k^{-1} \Pr \left(\chi_k^2 > 2k \right) \right\} \\ &\sim 0.71 \end{aligned}$$

- How can we 'fix' AIC?

3.2 HQIC

- Let

$$\phi_g(k) = \log \hat{\sigma}_k^2 + \frac{kg(T)}{T},$$

where $g(T) \uparrow \infty$, but $T^{-1}g(T) \rightarrow 0$.

- Then

$$\phi_g(k+1) - \phi_g(k) \begin{cases} \rightarrow \log \left\{ 1 - \left(\beta_{k+1}^{k+1} \right)^2 \right\} \leq 0 & ; \quad k < p-1 \\ \rightarrow \log \left\{ 1 - \left(\beta_p \right)^2 \right\} < 0 & ; \quad k = p-1 \\ \sim - \left(\hat{\beta}_{k+1}^{k+1} \right)^2 + \frac{g(T)}{T} & ; \quad k \geq p, \end{cases}$$

and we have, for $k \geq p$,

$$\begin{aligned} T \left\{ \phi_g(k+1) - \phi_g(k) \right\} &= g(T) - T \left(\hat{\beta}_{k+1}^{k+1} \right)^2 \\ &\rightarrow \infty. \end{aligned}$$

- Hence the minimiser of $\phi_g(k)$ converges in probability to p .
- What about convergence almost surely? Does \hat{p} converge to p on a set of probability measure 1?
- We need more than the asymptotic distribution of the $\sqrt{T} \hat{\beta}_{k+1}^{k+1}$.

- Hannan and Quinn (1979) proved the following result.
- For $k \geq p$, the limit points of the sequence (in T)

$$\sqrt{\frac{T}{2 \log(\log T)}} \hat{\beta}_{k+1}^{k+1}$$

are precisely $[-1, 1]$.

- Thus the sample partial autocorrelations satisfy a LIL (law of the iterated logarithm).
- Hence

$$\limsup \sqrt{\frac{T}{2 \log(\log T)}} \left| \hat{\beta}_{k+1}^{k+1} \right| = 1, a.s.$$

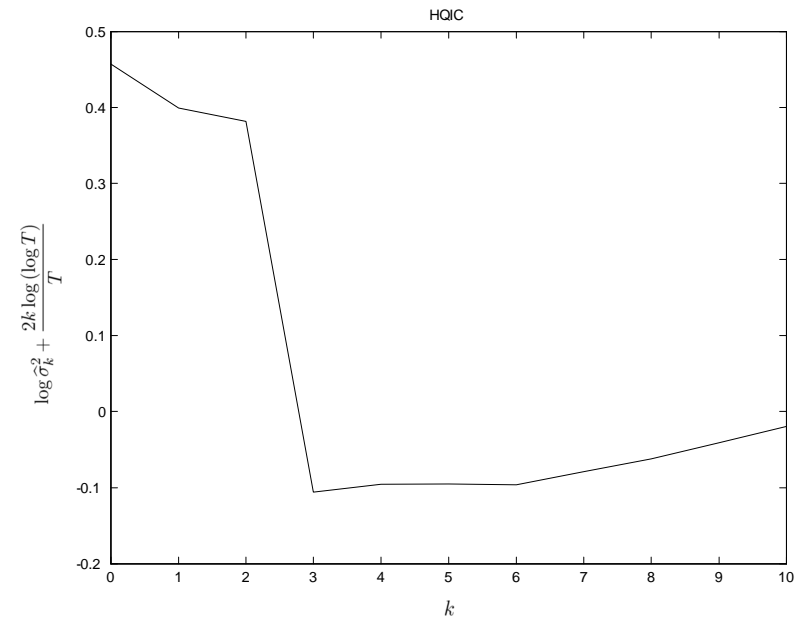
- Using the same argument as above, it therefore follows that \hat{p} converges to p almost surely if

$$\limsup \{g(T) - 2 \log(\log T)\} > 0$$

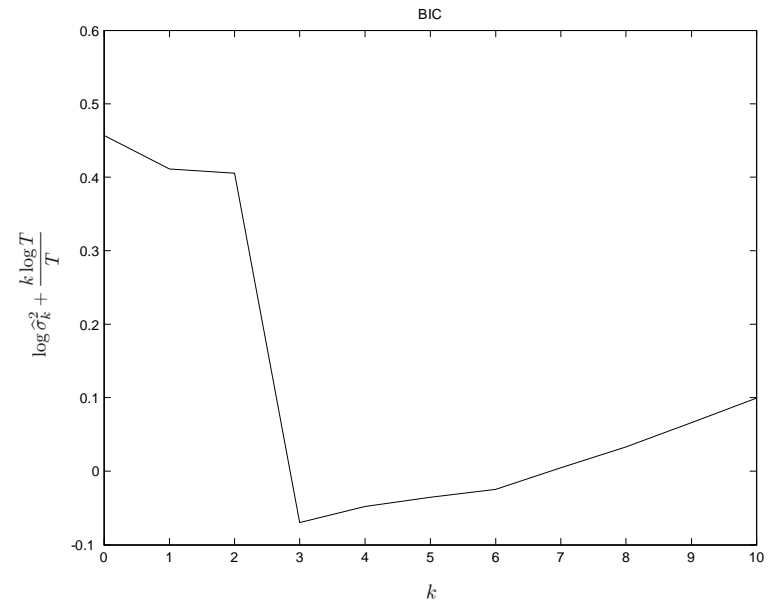
and only if

$$\limsup \{g(T) - 2 \log(\log T)\} \geq 0.$$

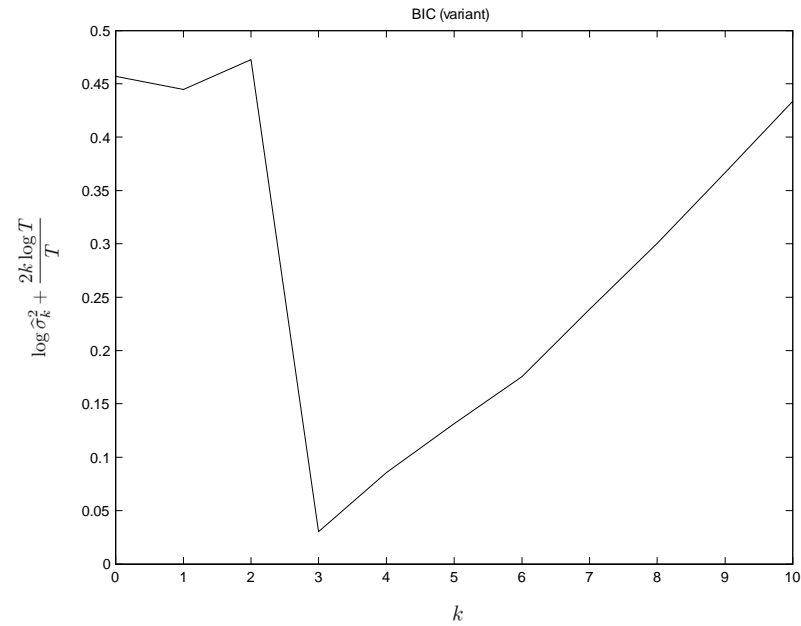
- Plot of ϕ_g with $g(T) = 2 \log(\log T)$.



- Plot of ϕ_g with $g(T) = \log T$ (BIC, Rissanen's MDL and Schwartz).



- Some authors use $g(T) = 2 \log T$. Here is a plot



- Thoughts: Small sample considerations? Is there a point? No. It is only the asymptotic theory which leads anywhere theoretically. Maybe simulations may lead to better results.

- Should we really be looking at consistency? Maybe we wish to control the probability of overestimating p ?

4 Generalisations

- Hannan generalised results to univariate ARMA and Quinn to multivariate AR processes.
- AIC is really a generalisation of FPE to all statistical models.
- The Gaussian log-likelihood is often of the form

$$l = -\frac{T}{2} \log (2\pi\sigma^2) - \frac{1}{2\sigma^2} S_T (\theta).$$

- The maximum of this over θ and σ^2 is

$$-\frac{T}{2} \log (2\pi\hat{\sigma}^2) - \frac{T}{2},$$

where

$$\begin{aligned}\hat{\sigma}^2 &= S_T(\hat{\theta}) \\ \hat{\theta} &= \arg \min S_T(\theta).\end{aligned}$$

- We can thus write AIC above as

$$\phi_{AIC}(k) = -\frac{2(l-k)}{T} + c,$$

so that

$$\phi_{AIC}(k+1) - \phi_{AIC}(k) = -\frac{\lambda-2}{T},$$

where λ is the likelihood ratio statistic for testing $H_0 : \beta_{k+1}^{k+1} = 0$ when Θ comprises the autoregressions of order $k + 1$.

- NOTE: The equation above makes sense only when the models under H_0 and H_A are “nested”.

5 Philosophical Considerations

- Akaike presupposes that there is no true model, and that we are trying to find the best fit from some subset of all possible models. Thus consistency is irrelevant.
- Rissanen's point of view is that we are trying to describe data with as few bits as possible – his criterion is motivated by this minimal description and by entropy considerations.
- Shibata has introduced the concept of efficiency of models (this should not be confused with the efficiency of estimators).

6 Some generalisations to nonstationary processes

- Quinn (1989) examined AIC for sinusoids in white noise. The usual likelihood ratio theory breaks down.

- Need

$$\phi(k) = \log \hat{\sigma}_k^2 + \frac{ck \log T}{T},$$

where $c > 2$ and

$$T \hat{\sigma}_k^2 = \sum_{t=0}^{T-1} (X_t - \bar{X})^2 - S_k,$$

where S_k is the sum of the k largest of

$$\left\{ I_X \left(\frac{2\pi j}{T} \right) ; 1 \leq j \leq \left\lfloor \frac{T-1}{2} \right\rfloor \right\},$$

and

$$I_X(\omega) = \frac{2}{T} \left| \sum_{t=0}^{T-1} X_t e^{-it\omega} \right|^2.$$

7 Modern applications

- AIC can be used when we are not so interested in the actual structure, but need to approximate it adequately for other reasons.
- Kavalieris and Hannan (1994) extended the above approach to the non-white case and removed the need for canonical frequencies.
- They did this by incorporating the estimation of the order of the best autoregressive fit.
- Another example: suppose we wish to test if two stationary processes $\{X_t\}$ and $\{Y_t\}$ have the same spectral structure, without specifying that structure.

- i.e. if $f_X(\omega)$ and $f_Y(\omega)$ are the two spectral densities, then we wish to test if $f_Y(\omega) = kf_X(\omega)$ for some $k, \forall \omega$.
- Various authors have used functions of $\left\{ I_Y(2\pi j/T) / I_X(2\pi j/T); 1 \leq j \leq \left\lfloor \frac{T-1}{2} \right\rfloor \right\}$, but the distributions, exact or asymptotic, are known only when $\{X_t\}$ and $\{Y_t\}$ are Gaussian and white.
- My approach. Use the likelihood ratio statistic to carry out the test, but use AIC to fit autoregressions to $\{X_t\}$ and $\{Y_t\}$.
- Under H_A , we use AIC to fit autoregressions separately to the two series.

- Under H_0 , we must specify a special bivariate autoregressive model:

$$X_t + \sum_{j=1}^p \beta_j X_{t-j} = \varepsilon_t$$
$$Y_t + \sum_{j=1}^p \beta_j Y_{t-j} = u_t,$$

where $\{\varepsilon_t\}$ and $\{u_t\}$ are white, but possibly correlated with each other.

- We then need a technique for estimating p , as well as the β_j .