14.384 Time Series Analysis, Fall 2007
Professor Anna Mikusheva
Paul Schrimpf, scribe
September 20, 2007
revised September 22, 2009

Lecture 5

Spectrum Estimation and Information Criteria

Spectrum Estimation

Same setup as last time. We have a stationary series, $\{z_t\}$ with covariances γ_j and spectrum $S(\omega) = \sum_{j=-\infty}^{\infty} \gamma_j e^{-i\omega j}$. We want to estimate $S(\omega)$.

Naïve approach

We cannot estimate all the covariances from a finite sample. Let's just estimate all the covariances that we can

$$\hat{\gamma}_j = \frac{1}{T} \sum_{j=k+1}^{T} z_j z_{j-k}$$

and use them to form

$$\hat{S}(\omega) = \sum_{j=-(T-1)}^{T-1} \hat{\gamma}_j e^{-i\omega j}$$

This estimator is not consistent. It convergers to a distribution instead of a point. To see this, let $y_\omega = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} e^{-i\omega t} z_t$, so that

$$\hat{S}(\omega) = y_\omega \bar{y}_\omega$$

If $\omega \neq 0$

$$2\hat{S}(\omega) \Rightarrow S(\omega)\chi^2(2)$$

Kernel Estimator

$$\hat{S}(\omega) = \sum_{j=-S_T}^{S_T} \left(1 - \frac{|j|}{S_T}\right) \hat{\gamma}_j e^{-i\omega j}$$

Under appropriate conditions on S_T ($S_T \to \infty$, but more slowly than T), this estimator is consistent[1] This can be shown in a way similar to the way we showed the Newey-West estimator is consistent.

[1] In a uniform sense, *i.e.* $P\left(\sup_{\omega \in [-\pi,\pi]} |\hat{S}(\omega) - S(\omega)| > \epsilon\right) \to 0$

Choosing the right order

Suppose you want to estimate an $AR(p)$, but you don't know the right p. The question arises how to choose the right order? There are two different setting. The first one is when you know that the true p does not exceed a known upper limit $\overline{p} < \infty$. The second case is when we cannot bound p from above, that is, when $\bar{p} = \infty$.

Testing down

We'll begin with the first case. A bad way (a naive way) is *testing down* : We start with a general $AR(\overline{p})$ and test $H_0^{(1)} : AR(\bar{p}-1)$ vs $H_A^{(1)} : AR(\bar{p})$ by testing that the coefficient $a_{\bar{p}}$ is zero. If we accept, then we'll test $\bar{p}-2$ vs $\bar{p}-1$. Continue until rejecting.

There are two issues with this approach:

- If we consider this as a testing procedure, then the size is very difficult to control because we're doing multiple testing. In other words, the individual significance level of each test differs from the probability of type I errors of the entire procedure. For example, to choose p you have to accept $H_0^{\bar{p}}, ..., H_0^p$ and reject H_0^{p-1}. These all are dependent hypothesis. What is the type one error in this procedure?
- If you consider this procedure as estimation, it won't be consistent. Imagine that you are choosing between two models $AR(1)$ and $AR(2)$ by testing $H_0 : a_2 = 0$. We'll incorrectly reject the true null α portion of the time even in the large sample (remember the definition of significance).

Using information criteria

Let $S_p = \ln \hat{\sigma}^2 + pg(T)$, where $\hat{\sigma}_p^2 = \frac{1}{T}\sum_t \hat{e}_t^2$ and $\hat{e}_t$ are the residuals from the model with p terms (in the multivariate case we'd use $\det|\hat{\Sigma}|$ in place of $\hat{\sigma}^2$). $pg(T)$ is a penalty term for having more parameters, where $g(T)$ is a non-random function. We choose

$$\hat{p} = \arg\min_{p \le \bar{p}} S_p$$

Different forms of $g(T)$ have special names

Definition 1. *AIC* (Akaike): $g(T) = \frac{2}{T}$
BIC (Bayesian/Schwartz): $g(T) = \frac{\ln T}{T}$
HQIC (Hannan-Quinn): $g(T) = \frac{2\ln\ln T}{T}$

Consistency

We will approach model selection as an estimation procedure. One of the characteristics of an estimator is consistency. Let p_0 be the true unknown order.

Definition 2. $\hat{p}$ is *weakly consistent* if $\lim P(\hat{p} = p_0) = 1$ (when we just say "consistent", we mean weakly consistent)

Definition 3. $\hat{p}$ is *strongly consistent* if $P(\lim \hat{p} = p_0) = 1$

Theorem 4. *If* $g(T) \to 0$ *and* $Tg(T) \to \infty$*, then* $\hat{p} \xrightarrow{p} p_0$

Proof. As usual, we will only sketch the proof. In this proof we would assume that S_p as a function of p has a U-shape (we would not prove this).

Lemma 5. *If* $g(T) \to 0$*, then* $P(\hat{p} \ge p_0) \to 1$

Proof. Note that

$$\begin{aligned}P(\hat{p} \geq p_0) =& P(S_{p_0} < \min\{S_0, S_1, ..., S_{p_0-1}\}) = \\ =& P\left\{\log \widehat{\sigma}^2_{p_0} + p_0 g(T) < \log \widehat{\sigma}^2_j + j g(T), \quad \forall j = 0, ..., p_0 - 1\right\} \\ =& 1 - P\left\{\log \widehat{\sigma}^2_{p_0} + p_0 g(T) < \log \widehat{\sigma}^2_j + j g(T), \quad \text{for some} \quad j = 0, ..., p_0 - 1\right\} \\ \geq& 1 - \sum_{j=0}^{p_0-1} P\left\{\log \widehat{\sigma}^2_{p_0} + p_0 g(T) < \log \widehat{\sigma}^2_j + j g(T)\right\}\end{aligned}$$

For each $j < p_0$ the true variance of the residual in $AR(j)$ is bigger than the residual in $AR(p_0)$ (since $AR(p_0)$ is the true model). That is, $\frac{\sigma^2_j}{\sigma^2_{p_0}} > 1$. Given that, $\widehat{\sigma_j}^2$ are consistent estimators of σ^2_j, we have $\frac{\hat{\sigma}^2_j}{\hat{\sigma}^2_{p_0}} \to^p \frac{\sigma^2_j}{\sigma^2_{p_0}} > 1$. Since $g(T) \to 0$, we obtain

$$P\left(\log \frac{\hat{\sigma}^2_j}{\hat{\sigma}^2_{p_0}} - (p_0 - j) g(T) > 0\right) \to 1$$

for each $j < p_0$, and since there are only finitely many $j < p_0$, we have

$$P(\hat{p} \geq p_0) \geq 1 - \sum_{j=0}^{p_0-1} P\left\{\log \sigma^2_{p_0} + p_0 g(T) < \log \sigma^2_j + j g(T)\right\} \to 1$$

□

Lemma 6. *If $Tg(T) \to \infty$ then $P(\hat{p} \leq p_0) \to 1$.*

Proof. As above, note that

$$\begin{aligned}P(\hat{p} \leq p_0) =& P(S_{p_0} < \min\{S_{p_0+1}, S_{p_0+1}, ..., S_{\bar{p}}\}) \\ =& P\left(\log \frac{\hat{\sigma}^2_j}{\hat{\sigma}^2_{p_0}} - (p_0 - j) g(T) > 0\,,\, j = p_0 + 1, .., \bar{p}\right)\end{aligned}$$

We can multiply both sides of the inequality by T without changing the probability:

$$P(\hat{p} \leq p_0) = P\left(T \log \frac{\hat{\sigma}^2_j}{\hat{\sigma}^2_{p_0}} - (p_0 - j) T g(T) > 0\,,\, j = p_0 + 1, .., \bar{p}\right) \tag{1}$$

One can notice that the Likelihood ratio (LR) statistic for testing $H_0 : AR(p_0)$ vs the alternative $H_a : AR(j)$ (the same as testing simulteneously $j - p_0$ restrictions $a_j = ... = a_{p_0+1} = 0$) is

$$LR = -T \log \frac{\hat{\sigma}^2_j}{\hat{\sigma}^2_{p_0}} \Rightarrow \chi^2_{j-p_0}$$

if the null is true. That is, $T \log \frac{\hat{\sigma}^2_j}{\hat{\sigma}^2_{p_0}}$ is negative but bounded in probability. If $Tg(T) \to \infty$, then

$$P\{T \log \frac{\hat{\sigma}^2_j}{\hat{\sigma}^2_{p_0}} + (j - p_0) T g(T) > 0\} \to 1$$

for each j. Since we have a finite number of these events, the needed probability (1) converges to 1. □

□

Remark 7. Note in the proof of the previous lemma that if $Tg(T) \to \quad const < \infty$, then $\lim P\{\widehat{p} \leq p_0\} < 1$; that is, we have overestimation of the order.

Remark 8. The proof so far has been for weak consistency. If we additionally have $\frac{g(T)T}{2 \ln \ln T} > 1$, then $\hat{p} \to p$ a.s., so $\hat{p}$ is strongly consistent.

Corollary 9. *BIC is strongly consistent. HQIC is consistent. AIC is not; it has a positive probability of overestimating the number of lags asymptotically.*

Why do we need consistency?

Assume that we are interested in estimating coefficients of $AR(p_0)$. Then what problem do we face if choose wrong order of the model? If we choose p too small, then the estimated coefficients will be biased (think: omitted variable bias). If we choose p too large, then our estimates of coefficients will be inefficient (variance is bigger).

When $\hat{p}$ is consistent, estimating an AR(p) model in this 2-step way has the same asymptotic distribution as if we knew p_0. Let's sketch this argument. Let $\widehat{\beta}(M_p)$ is estimate if model M_p is true. The two-step estimate then is

$$\widetilde{\beta} = \sum_{p=0}^{\bar{p}} \widehat{\beta}(M_p) \mathbb{I}_{\{\widehat{p}=p\}}.$$

What do we know about the distribution of $\widetilde{\beta}$?

$$P\left\{\widetilde{\beta} \leq s\right\} = P\left\{\widehat{\beta}(M_p) \leq s | \widehat{p} = p\right\} P\left\{\widehat{p} = p\right\}$$

If we have consistency $P\{\widehat{p} = p_0\} \to 1$, then asymptotically

$$P\left\{\widetilde{\beta} \leq s\right\} \cong P\left\{\widehat{\beta}(M_{p_0}) \leq s\right\}$$

A word of causion, in finite samples, not knowing p can make a difference, and the finite sample distribution of our estimates is influenced by model selection.

AIC and forecasting

Despite its inconsistency for picking the number of lags, the AIC is not useless. The choice of info criterion should depend on the problem at hand. AIC minimizes MSE for one-step ahead forecast.

If the goal is to do a good forecasting one may come up with the following idea: we want to forecast y based on x: $y_t = \beta x_t + u_t$, and are thinking what to include into x. You may think of AR(p) as predicting y_t from $y_{t-1}, ..., y_{t-p}$. Let's do this forecasting exercise as a real-time process. That is, assume you to ok AR(p) as a model. When you have data from time 1 to time $t-1$ you can calculate estimate $\hat{\beta}_{t-1}$ based on this observations only and then predict y_t and check how well you did. That is,

$$\hat{\beta}_{t-1} = (X'_{t-1} X_{t-1})^{-1} X'_{t-1} y_{t-1}$$
$$\hat{y}_{t|t-1} = \hat{\beta}_{t-1} x_t$$
$$PLS(p) = \sum_{t=m+1}^{T} (y_t - y_{t|t-1})^2$$

PLS stays for predicted least squares. You may wish to choose the order of the model based on how well it does on this predicting task: $\widehat{p} = PLS(p)$. Here m is the first observation for which you start doing

predictions.
Wei (1992) shows that for m fixed

$$\frac{1}{T-m} PLS(p) = S_p^{BIC} + o(\frac{1}{T})$$
$$\hat{p}_{PLS} = \hat{p}_{BIC} + o_p(1)$$

Innoue and Killian (2002) shows that if $m = cT$

$$\frac{1}{T-m} PLS(p) = S_p^{AIC} + o(\frac{1}{T^2})$$

Discussion

- A larger point is that the best model selection criteria depends on what your goal is. For the estimation it may be different than for forecasting.
- Information criterion can be thought of as "sequential testing" but with changing critical values
- Consistent estimation guarantees asymptotic efficiency of estimation, but can behave poorly in finite samples.
- The main problem however is that we don't know $\bar{p}(\bar{p} = \infty)$

On the case of $\bar{p} = \infty$.

Example 10. Suppose we want to estimate the spectrum by fitting an $AR(\infty)$ model.

$$a(L)y_t = e_t a(L) = 1 + \sum_{j=1}^{\infty} a_j L^j$$

The spectrum is

$$S(\omega) = \frac{\sigma^2}{|1 + \sum_{j=1}^{\infty} a_j e^{-i\omega j}|^2}$$

From this, we see that what we really want to estimate is the infinite sum in the denominator. There are two approaches

1. Estimate $AR(p_T)$, let $p_T \to \infty$, but not too fast, say $\frac{p_T^3}{T} \to \infty$. This procedure will give us a consistent estimate of the spectrum. To see this, consider:

   $$|\sum_{j=1}^{\infty} a_j e^{-i\omega j} - \sum_{j=1}^{p_T} \hat{a}_j e^{-i\omega j}| \leq \sum_{j=1}^{\infty} |a_j - \hat{a}_j|$$

   By a similar argument to what we used to show the consistency of HAC, we could show that $\sum_{j=1}^{\infty} |a_j - \hat{a}_j| \to 0$

2. Same p_T, but choose $\hat{p} = \min_{p \leq p_T} BIC(p)$. Then estimate $AR(\hat{p})$ and compute the spectral density. If $p_T = (\ln T)^a$, $0 < a < \infty$, and if (i) $p_0 < \infty$ then $P(\hat{p} = p_0) \to 1$, or if (ii) $p_0 = \infty$, then $\frac{\hat{p}}{\ln T} \to 1$ and $\sup |S(\omega) - \hat{S}(\omega)| \to^p 0$.

   Remark 11. Of course, the condition that $p_T = (\ln T)^a$ is a bit difficult to interprete in practice. For example,

T	$4\left(\frac{T}{50}\right)^{1/3}$	$\frac{1}{4}(\ln T)^2$
50	4	4
250	7	8
2500	15	15

14.384 Time Series Analysis

Fall 2013