Lecture 4: Model Selection Roundup

Francis J. DiTraglia

March 19, 2014

1 Some Time Series Examples

So far we've looked at some completely generic examples (AIC, TIC, cross-validation) and some regression examples (Mallows, AIC corrected). The generic examples can immediately be applied to an ML problem, including time series setting: if you have an ML routine, you already get everything you need for these criteria as a side effect. Could be Kalman filter or conditional likelihood.

We derived the other examples (Mallows, AIC corrected) for a regression problem, but it's easy to adapt these to AR and VAR models. As long as we're willing to use conditional ML (drop some observations) these already *are* regression problems.

We'll take a look at AR and VAR models using conditional likelihood so we can write out explicit formulas for the criteria. We won't do TIC since we can't really unpack the penalty term. We'll treat cross-validation in its own section.

The consistent criteria are ... and the efficient criteria are ...

We won't go through all of the specifics of the derivations for mallows and AICc since they're almost identical to the regression derivation. Some more details can be found in McQuarrie and Tsai (1998).

1.1 Autoregressive Models

For simplicity assume there is no constant term. Then the AR(p) model is

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

where $\epsilon_t \sim$ iid $N(0, \sigma^2)$ and we observe a sample $y_1, \ldots, y_N$. We'll use conditional maximum likelihood, so we lose the first p observations. Thus the *effective sample size* is $T = N - p$. The conditional ML estimator of $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$ is simply the least-squares estimator

$$\widehat{\boldsymbol{\phi}} = (X'X)^{-1}X'\mathbf{y}$$

where $\mathbf{y} = (y_{p+1}, y_{p+2}, \ldots, y_N)'$ and the design matrix is

$$X = \begin{bmatrix} y_p & y_{p-1} & \cdots & y_1 \\ y_{p+1} & y_p & \cdots & y_2 \\ \vdots & \vdots & & \vdots \\ y_{N-1} & y_{N-2} & \cdots & y_{N-p-1} \end{bmatrix}$$

The maximum likelihood estimator of σ^2 is

$$\widehat{\sigma}_p^2 = \frac{\text{RSS}_p}{T}$$

where RSS denotes the residual sum of squares, namely $||\mathbf{y} - X\widehat{\boldsymbol{\phi}}||^2$. Since this is a regression model, it's trivial to adapt both Mallow's C_p and the AIC$_C$ to this case.[1] For Mallow's C_p we have

$$C_p = \frac{\text{RSS}_p}{\widehat{\sigma}^2_{wide}} - T + 2p$$

[1] If you'd like to see all of the details written out, consult McQuarrie & Tsai (1998), Chapter 3.

where $\widehat{\sigma}^2_{wide}$ is the estimate of σ^2 from the model with *maximum order* among those under consideration. For AIC_c we have

$$\text{AIC}_c = \log\left(\widehat{\sigma}^2_p\right) + \frac{T+p}{T-p-2}$$

For both C_p and AIC_c we choose the lag length that *minimizes* the criteiron.

Using an argument essentially identical to the one presented in the notes for Lecture 2, the maximized log-likelihood for the AR(p) model is

$$-\frac{T}{2}\left[\log(2\pi) + \log\left(\widehat{\sigma}^2_p\right) + 1\right]$$

To construct the AIC and BIC, we multiply this quantity by 2 and subtract the appropriate penalty term, ignoring terms that are constant across models. The number of parameters for an AR(p) model is $p+1$, since we estimate σ^2 in addition to the p autoregressive parameters. We'll rescale both AIC and BIC and flip their signs to make them comparable to the C_p and AIC_c expressions from above. Putting everything together for the sake of comparison, we have

$$\begin{aligned}
\text{AIC} &= \log\left(\widehat{\sigma}^2_p\right) + \frac{2(p+1)}{T} \\
\text{AIC}_c &= \log\left(\widehat{\sigma}^2_p\right) + \frac{T+p}{T-p-2} \\
C_p &= \frac{\text{RSS}_p}{\widehat{\sigma}^2_{wide}} + 2p - T \\
\text{BIC} &= \log\left(\widehat{\sigma}^2_p\right) + \frac{\log(T)(p+1)}{T}
\end{aligned}$$

In each case, we choose the model that *minimizes* the criterion. Of these four criteria, only BIC is consistent. The other three criteria, however, are efficient under one-step-ahead squared prediction error loss in an environment in which the true DGP is an infinite-order autoregression. The BIC does not have this property.

Ng & Perron (2005) There are some subtle but important points that we glossed over in the preceding discussion and that are, indeed, rarely mentioned in textbooks or articles on model selection. First there is the question of whether we should use the maximum likelihood estimator $\widehat{\sigma}^2$ or the unbiased estimator that divides by $T - p$ rather than T. In time series applications T may be small enough that it makes a difference. More troubling, however, is the problem of deciding what should count as the sample size, since different lag lengths use a different number of observations in the conditional maximum likelihood setting. Indeed, as they are usually written, expressions for AIC and BIC drop terms that are constant across models in *cross-section regression*, where changing the number of regressors doesn't affect sample size. The situation is of course entirely different for AR models but practicioners *still use the same formulas* in this case. There are numerous different ways to handle these complications. Ng & Perron (2005) review the possibilities and illustrate how each performs in a number of simulation studies.

1.2 Vector Autoregression Models

Again, assume the intercept is zero. Then the VAR(p) model is given by

$$\begin{aligned} \underset{(q\times 1)}{\mathbf{y}_t} &= \underset{(q\times q)}{\Phi_1}\,\mathbf{y}_{t-1} + \ldots + \Phi_p \mathbf{y}_{t-p} + \boldsymbol{\epsilon_t} \\ \boldsymbol{\epsilon}_t &\overset{iid}{\sim} N_q(\mathbf{0}, \Sigma) \end{aligned}$$

where we observe $\mathbf{y}_1, \ldots, \mathbf{y}_N$. Again, if we're content to use conditional maximum likelihood, dropping the first p observations to estimate a VAR(p) model, this is simply a multivariate regression problem and we have an *effective sample size of* $T = N - p$. Written as a multivariate regression model, we have

$$\underset{(T\times q)}{Y} = \underset{(T\times pq)}{X}\,\underset{(pq\times q)}{\Phi} + \underset{(T\times q)}{U}$$

where

$$\underset{(T\times q)}{Y} = \begin{bmatrix} \mathbf{y}'_{p+1} \\ \mathbf{y}'_{p+2} \\ \vdots \\ \mathbf{y}'_{N} \end{bmatrix}, \quad \underset{(pq\times q)}{\Phi} = \begin{bmatrix} \Phi'_1 \\ \Phi'_2 \\ \vdots \\ \Phi'_p \end{bmatrix}, \quad \underset{(T\times q)}{U} = \begin{bmatrix} \boldsymbol{\epsilon}'_{p+1} \\ \boldsymbol{\epsilon}'_{p+2} \\ \vdots \\ \boldsymbol{\epsilon}'_{N} \end{bmatrix}$$

and the design matrix is

$$\underset{(T\times pq)}{X} = \begin{bmatrix} \mathbf{y}'_p & \mathbf{y}'_{p-1} & \cdots & \mathbf{y}'_1 \\ \mathbf{y}'_{p+1} & \mathbf{y}'_p & \cdots & \mathbf{y}'_2 \\ \vdots & \vdots & & \vdots \\ \mathbf{y}'_{N-1} & \mathbf{y}'_{N-2} & \cdots & \mathbf{y}'_{N-p-1} \end{bmatrix}$$

Thus, the conditional maximum likelihood estimator for Φ is

$$\widehat{\Phi} = (X'X)^{-1}X'Y$$

and the maximum likelihood estimator for Σ is

$$\widehat{\Sigma}_p = \frac{\left(Y - X\widehat{\Phi}\right)'\left(Y - X\widehat{\Phi}\right)}{T}$$

The VAR(p) model has a very large number of parameters. First, we have the coefficients of $\Phi_1, \ldots, \Phi_p$. Each of these is an unrestricted $q \times q$ matrix so Φ contains a total of pq^2 parameters. We also need to estimate the variance matrix Σ of the errors ϵ. Although Σ contains q^2 elements, it is a symmetric matrix so there are only $q(q+1)/2$ free parameters. Thus, a VAR(p) model requires us to estimate a total of $pq^2 + (q+1)q/2$ parameters. To calculate the AIC and BIC we also need the maximized log-likelihood, which is given by

$$-\frac{T}{2}\left[q\log(2\pi) + \log\left|\widehat{\Sigma}_p\right| + q\right]$$

Re-scaling as we did for the AR model, we have

$$\begin{aligned} \text{AIC} &= \log\left|\widehat{\Sigma}_p\right| + \frac{2pq^2 + q(q+1)}{T} \\ \text{BIC} &= \log\left|\widehat{\Sigma}_p\right| + \frac{\log(T)pq^2}{T} \end{aligned}$$

The multivariate generalization of AIC_c is

$$\text{AIC}_c = \log\left|\widehat{\Sigma}_p\right| + \frac{(T+qp)q}{T-qp-q-1}$$

as explained in Chapter 5 of McQuarrie and Tsai (1998). For each of the preceding three expressions, we choose the model that *minimizes* the given criterion. Of these criteria, both AIC and its corrected version are efficient while BIC is consistent.

1.3 Corrected AIC for State Space Models

As the lag length p grows, the number of parameters in a VAR(p) model explodes, and can easily come close to the effective sample size. In situations like this, AIC is known to perform poorly

Cavanaugh & Shumway (1997)

2 More on Cross-Validation

How to extend it to time series. Varieties other than leave-one-out. Efficiency versus consistency. Racine (2000) and Burman, Chow & Nolan (1994).

2.1 K-Fold Cross-validation

2.2 How to handle dependent observations

AR example.

Cross-Validation for AR The way we described it above, CV depended in independence. How can we adapt it for AR models? Roughly speaking, the idea is to use the fact that dependence dies out over time and treat observations that are "far enough apart" as *approximately* independent. Specifically, we choose an integer value h and assume that y_t and y_s can be treated as independent as long as $|s-t| > h$. This idea is called "h-block cross-validation" and was introduced by Burman, Chow & Nolan (1994). As in the iid version of leave-one-out cross-validation, we still evaluate a loss function by predicting *one* witheld observation at a time using a model estimated without it. The difference is that we also omit the h neighboring observations *on each side* when fitting the model. For example, if we choose to evaluate squared-error loss, the criterion is

$$CV_h(1) = \frac{1}{T-p} \sum_{t=p+1}^{T} \left(y_t - \hat{y}^h_{(t)}\right)^2$$

where

$$\hat{y}^h_{(t)} = \hat{\phi}^h_{1(t)} y_{t-1} + \ldots + \hat{\phi}^h_{1(t)} y_{t-p}$$

and $\hat{\phi}^h_{j(t)}$ denotes the jth parameter estimate from the conditional least-squares estimator with observations $y_{t-h}, \ldots, y_{t+h}$ removed. We still have the question of what h to choose. Here there is a trade-off between making the assumption of independence more plausible and leaving enough observations to get precise model estimates. Intriguingly, the simulation evidence presented in McQuarrie and Tsai (1998) suggests that setting $h = 0$, which yields plain-vanilla leave-one-out CV, works well even in settings with dependence.

The idea of h-block cross-validation can also be adapted to versions of

cross-validation other than leave-one-out. For details, see Racine (1997, 2000).

Cross-Validation for VARs In principle we could use the same h-block idea here as we did for for the AR example above. However, given the large number of parameters we need to estimate, the sample sizes witholding $2h+1$ observations at a time may be too small for this to work well.

3 Two Additional Criteria

We've already covered the most commonly used model selection criteria, but there are two others that come up from time to time: Akaike's Final Prediction Error (FPE), and the Hannan-Quinn Information Criterion (HQ). Roughly speaking FPE behaves like AIC while HQ behaves like BIC: while FPE is efficient, HQ is consistent.

3.1 Final Prediction Error

3.2 Hannan-Quinn

In our last lecture we examined a consistency result based on the central limit theorem. It is also possible to construct a consistency proof by appealing to the law of the iterated logarithm. This is how HQ is derived.

For VAR models

$$\begin{aligned} FPE &= \left|\widehat{\Sigma}_p\right| \left(\frac{T+qp}{T-qp}\right)^q \\ HQ &= \log\left|\widehat{\Sigma}_p\right| + \frac{c\log\log(T)pq^2}{T} \end{aligned}$$

where $c > 2$. A commonly-used value is 2.01.