Lecture 4: Time Series Examples

Francis J. DiTraglia

March 20, 2014

1 Some Time Series Examples

Thus far we've looked at a number of model selection criteria. Some of them, namely AIC, BIC and TIC, are completely portable: they can be applied to *any* model that is estimated by maximum likelihood. Each of these can be immediately applied to time series data: if you have a routine to carry out ML estimation, be it conditional ML or the Kalman filter, it already produces all the quantities you need. In contrast, some of the other examples we considered, namely Mallow's C_p and AIC$_c$, were derived for the special case of linear regression. How can we adapt these examples to time series data? Fortunately, if we're willing to use conditional ML estimation, some of the most widely used time series models *are in fact* regression models. In this section we'll take a closer look at model selection for autoregression and vector autoregression models.

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

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

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

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

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.

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

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

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

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

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

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

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

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. The bias correction $2 \times \text{length}(\theta)$ is based on a large-sample argument and fails to provide a good approximation when the number of parameters is too close to the sample size, leading the AIC to choose models that are in general "too large" to acheive our target of minimizing the KL divergence.[2] The idea behind the AIC$_c$ of Hurvich and Tsai (1989) was to provide a better approximation to the AIC bias correction for AR models under a certain set of assumptions. In a similar vein, Cavanaugh & Shumway (1997) propose a refined AIC, the AIC$_b$, for general state space models. Rather than deriving an analytical correction term, they suggest using the bootstrap to approximate the bias of the maximized log-likelihood as an estimator of the expected log likelihood, using the state-space bootstrap procedure proposed by Stoffer and Wall (1991).

[2] Cavanaugh & Shumway (1997) suggest $\text{length}(\theta) \approx T/2$ as a rough approximation of what counts as "too many parameters relative to sample size" for the AIC to work well.

2 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.

2.1 Hannan-Quinn Information Criterion

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 which, in its simplest form, states that

$$\limsup_{T\to\infty} \frac{\sum_{t=1}^{T} Y_t}{\sqrt{T \log\log T}} = \sqrt{2} \text{ a.s.}$$

provided that $Y_t \sim$ iid with mean zero and unit variance. This is essentially the result that "lies in-between" the CLT, which says that a scaling of $T^{-1/2}$ gives a result that has a non-degenerate limit distribution, and the SLLN, which says that a scaling of T^{-1} gives a result that converges to zero, almost surely. Roughly speaking, the idea behind HQ is to find the *weakest* penalty term that will yield consistent model selection. In other words, the point is to find the *slowest growing* function of T that still ensures we select the most parsimonious model among those tha minimize the KL divergence. The answer turns out to be $2c \log\log T \times \text{length}(\theta)$ where $c > 1$, yielding the following criterion:

$$HQ = 2\ell(\widehat{\theta}) - 2c \log\log T \times \text{length}(\theta)$$

Rescaling and reversing the sign of the criterion to make it agree with our other examples we have the following for AR models

$$HQ = \log\left(\widehat{\sigma}_p^2\right) + \frac{2c \log\log T \times (p+1)}{T}$$

and for VAR models we have

$$HQ = \log\left|\widehat{\Sigma}_p\right| + \frac{2c\log\log T \times (pq^2 + q(q+1)/2)}{T}$$

where, again, $c > 1$. The practical problem with HQ is that $\log\log T$ is fairly small unless the sample size is extremely large. This means that different choices of c lead to very different behavior, and the asymptotics provide us with no guidance on this point.

2.2 Final Prediction Error

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

3 More on Cross-Validation

In Lecture 2 we discussed leave-one-out (LOO) cross-validation (CV) and proved that it is asymptotically equivalent to TIC. In fact, there's a lot more that can be said about cross-validation. In this section we'll take a closer look.

3.1 K-Fold Cross-validation

LOO-CV can be viewed as a special case of a more general procedure: **K-fold cross-validation**. The idea behind cross-validation is to create a "stand-in" for the – typically infeasible – ideal of fitting your model on a "training" dataset and assessing its predictive performance out-of-sample on an independent "validation" dataset. In the simplest and most common case we suppose that we have observed an iid sample and K-fold CV proceeds as follows

1. Randomly partition the dataset into K parts of roughly equal size. These

are the "folds."

2. For each $k = 1, \ldots, K$ do the following estimate your model using all observations *except* those contained in the kth fold.

3. For each observation y_t, let $\widehat{y}^{-k(t)}$ denote the predicted value of y_t from the model estimated using all the data *except* the fold in which observation t has been placed.

4. The K-fold cross-validation estimate of the out-of-sample predictive loss is given by

$$CV(K) = \frac{1}{T} \sum_{t=1}^{T} L\left(y_t, \widehat{y}^{-k(t)}\right)$$

   where L is the desired loss function, for example squared error loss.

5. Repeat this procedure for each model under consideration, and select the one that yields the smallest value of $CV(K)$.

So how should we choose K? There is no clear answer to this question, but several points are worth considering. The first is computational complexity. Leave-one-out CV requires us to re-fit each model T times. In contrast 5-fold cross-validation only requires us to re-fit 5 times. As you saw on the recent problem set, however, there are special cases – namely linear smoothers under quadratic loss – in which we don't actually need to re-fit anything to carry out LOO CV. Many interesting models, however, cannot be expressed as linear smoothers so this consideration can be important.

Hastie, Tibshirani & Friedman (2008) phrase the choice of K in terms of a bias-variance tradeoff. When $K = T$, we have as many folds as observations. This is simply leave-one-out CV and it turns out to give an approximately unbiased estimator of the expected out-of-sample prediction error. Using a larger value of K, they argue, introduces a bias but tends to produce a lower variance estimator of the prediction estimator because the partial-sample estimators are less similar to each other when they have fewer observations in common. While this advice is reasonable in certain situations, such as classification and density estimation, it is far from universally applicable as Arlot & Celisse (2010) point out in their comprehensive review article. For example, setting $K = T$ actually *minimizes* the variance of the prediction error estimator in certain settings, such as linear regression.

A third consideration is the asymptotic performance of the cross-validation procedure for different choices of K. When we set $K = T$, K-fold CV reduces to leave-one-out cross-validation and we proved in our second lecture that this is asymptotically equivalent to TIC. Since TIC is an efficient model selection criterion, it follows that LOO-CV is also efficient. But what about more general K-fold CV? This turns out to be a somewhat awkward question to address. It is known that *certain varieties* of cross-validation are consistent. In particular, Shao (1993) showed that "leave-n_v-out" cross-validation is consistent provided that the ratio of n_v/T converges to 1 as $T \to \infty$. In other words, for consistency we need to use a *small* training data set and a *large* validation set. Unlike K-fold CV which splits the data into non-overlapping subsets, "leave-n_v-out" CV, at least in principle, uses *all* subsets of size n_v for validation. K-fold cross-validation is not really amenable to this kind of asymptotic argument, however, since the maximum ratio of n_v/T is 1/2, which corresponds to 2-fold cross-validation.

In practice, many researchers recommend setting K equal to 5 or 10 if the goal is precise estimators.

3.2 LOO-CV for Dependent Data

The way we described it above, CV depended in independence. How can we adapt it for, say, an AR model? 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

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

where

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

and $\hat{\phi}_{j(t)}^h$ 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.

In principle, we could use the h-block idea in other settings as well, for example VAR models. 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 in practice.

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 (2000) who proves the consistency of a particular flavor of cross-validation, so-called "*hv*-block," for dependent data.