5.5: Order Selection

Taylor

University of Virginia

Motivation

So far we have discussed fitting specific ARMA(p,q) models. How we decide which model to fit? That is, how do we decide which numbers to use for p and q?

The FPE Criterion

This is for AR models only!

Two statistically identical and independent time series $\{X_t\}_{t=1}^{n}$ and $\{Y_t\}_{t=1}^{n+1}$.

Estimate parameters with $\{X_t\}$. Look at MSE on $\{Y_t\}$. The idea is to not overfit, but we still assume the true model won't change.

Also, for simplicity we assume $p < n$ (we have enough data).

In general, more complicated models will always shrink $\hat{\sigma}^2$, but we include a penalty that has to do with p.

The FPE Criterion

$$
\begin{aligned}
&E(Y_{n+1} - \hat{\phi}_1 Y_n - \cdots - \hat{\phi}_p Y_{n+1-p})^2 \\
&= E(Y_{n+1} - \phi_1 Y_n - \cdots - \phi_p Y_{n+1-p} - \\
&\qquad\qquad (\hat{\phi}_1 - \phi_1) Y_n - \cdots - (\hat{\phi}_p - \phi_p) Y_{n+1-p})^2 \\
&= E(Z_t + (\hat{\phi}_1 - \phi_1) Y_n - \cdots - (\hat{\phi}_p - \phi_p) Y_{n+1-p})^2 \\
&= \sigma^2 + E[E[((\hat{\phi}_1 - \phi_1) Y_n - \cdots - (\hat{\phi}_p - \phi_p) Y_{n+1-p})^2 | \{X_t\}]] \\
&= \sigma^2 + E\left[\sum_{i=1}^{p}\sum_{j=1}^{p} (\hat{\phi}_i - \phi_i)(\hat{\phi}_j - \phi_j) E\left[Y_{n+1-i} Y_{n+1-j} | \{X_t\}\right]\right] \\
&= \sigma^2 + E[(\hat{\boldsymbol{\phi}}_p - \boldsymbol{\phi}_p)' \Gamma_p (\hat{\boldsymbol{\phi}}_p - \boldsymbol{\phi}_p)] \\
&\approx \sigma^2 (1 + \frac{p}{n}) \qquad \text{(typo in book: } n^{-1/2} \text{ should be } n^{1/2}\text{)} \\
&\approx \frac{n\hat{\sigma}^2}{n-p}(1 + \frac{p}{n}) = \hat{\sigma}^2 \frac{n+p}{n-p} \qquad (n\hat{\sigma}^2/\sigma^2 \text{ approx. } \chi^2_{n-p})
\end{aligned}
$$

The FPE Criterion

```
getFPE <- function(ts, p){
  # note this functon only works on AR models
  n <- length(ts)
  mod <- arma(ts, p=p, q =0)
  sigmaSqdHat <-mod$sigma2
  sigmaSqdHat * ((n+p)/(n-p))
}
getFPE(xt, 0)
getFPE(xt, 1)
getFPE(xt, 2)
getFPE(xt, 3)
getFPE(xt, 4)  #this matches up with the table on page 150
```

AIC and AICc

These are more general. They work for any Gaussian ARMA time series model, not just AR models. However they are still based on the same idea where you assume that you estimate parameters with $\{X_t\}$, and then look at a loss of $\{Y_t\}$.

The loss here is based on the idea of Kullback-Leibler divergence $d(\psi|\theta)$. This is kind of like a distance between probability density functions. Even though it is nonnegative, it isn't technically a distance because it isn't symmetric ($d(\psi|\theta) \neq d(\theta|\psi)$).

However, it is still a loss function in a sense, so we want to pick the model that minimizes it. Also, the math is a little bit more fancy.

AIC and AICc

(Twice) the Kullback-Leibler divergence between your model's density $f(\mathbf{x};\psi)$ and the true model $f(\mathbf{x};\theta)$ is

$$\begin{aligned} d(\psi|\theta) &= 2E_\theta\left[-\ln\frac{f(\mathbf{x};\psi)}{f(\mathbf{x};\theta)}\right] \\ &= \int -2\ln\frac{f(\mathbf{x};\psi)}{f(\mathbf{x};\theta)} f(\mathbf{x};\theta)d\mathbf{x} \\ &= E_\theta\left[-2\ln f(\mathbf{x};\psi)\right] - E_\theta\left[-2\ln f(\mathbf{x};\theta)\right] \\ &= \Delta(\psi|\theta) - \Delta(\theta|\theta) \end{aligned}$$

Since the second part has nothing to do with our model, minimizing KL is the same as minimizing $\Delta(\psi|\theta)$.

NB: θ here does not denote the MA parameters; it's all parameters.

AIC and AICc

Say we estimate our parameters (ϕ, θ, σ^2) with $\hat{\beta} = (\hat{\phi}, \hat{\theta}, \hat{\sigma}^2)$ using the dataset $\mathbf{X}$. Also, the true parameters might not be the same dimension.

$$\begin{aligned}
\Delta(\hat{\theta}|\theta) &= E_\theta\left[-2\ln f(\mathbf{y};\hat{\beta})\right] \\
&= E_\theta\left[-2\ln\left\{(2\pi\prod_{j=1}^{n}\hat{\sigma}^2 r_{j-1})^{-1/2}\exp\left[-\frac{1}{2\hat{\sigma}^2}\sum_{j=1}^{n}\frac{(y_j-\hat{y}_j)^2}{r_{j-1}}\right]\right\}\right] \\
&= E_\theta\left[\ln 2\pi + n\log\hat{\sigma}^2 + \sum_{j=1}^{n}\ln r_{j-1} + \frac{S_Y(\hat{\beta})}{\hat{\sigma}^2} + n\frac{\hat{\sigma}^2}{\hat{\sigma}^2} - n\right] \\
&= E_\theta\left[-2\ln f(\mathbf{x};\hat{\beta})\right] + E\left[\frac{S_Y(\hat{\beta})}{\hat{\sigma}^2}\right] - n \\
&\approx E_\theta\left[-2\ln f(\mathbf{x};\hat{\beta})\right] + \frac{2n(p+q+1)}{n-p-q-2} \quad \text{(we won't prove this part)}
\end{aligned}$$

AIC and AICc

An approximately unbiased estimate for the quantity

$$\Delta(\hat{\theta}|\theta) \approx E_\theta\left[-2\ln f(\mathbf{x};\hat{\beta})\right] + \frac{2n(p+q+1)}{n-p-q-2}$$

is given by

AICc

the AICc statistic

$$-2\ln f(\mathbf{x};\hat{\beta}) + \frac{2n(p+q+1)}{n-p-q-2}$$

or we can use

AIC

the AIC statistic:

$$-2\ln f(\mathbf{x};\hat{\beta}) + 2(p+q+1)$$

AIC and AICc

Both of these select the model that has the best fit, but penalizes larger models. From the book:

"For fitting autoregressive models, Monte Carlo studies (Jones 1975; Shibata 1976) suggest that the AIC has a tendency to overestimate p. The penalty factors $2(p+q+1)n/(n-p-q-2)$ and $2(p+q+1)$ for the AICC and AIC statistics are asymptotically equivalent as $n \to \infty$. The AICC statistic, however, has a more extreme penalty for large-order models, which counteracts the overfitting tendency of the AIC. "

AIC and AICc

The book mentions BIC as well, but we will not. Here is some R code:

```
> arma(xt, p = 0, q = 1)$aicc
[1] 253.4228
> arma(xt, p = 0, q = 2)$aicc
[1] 229.1882
> arma(xt, p = 1, q = 0)$aicc
[1] 217.3914
> arma(xt, p = 1, q = 1)$aicc #this one!
[1] 212.7675
> arma(xt, p = 1, q = 2)$aicc
[1] 214.9143
> arma(xt, p = 2, q = 0)$aicc
[1] 213.5388
> arma(xt, p = 2, q = 1)$aicc
[1] 214.9269
> arma(xt, p = 2, q = 2)$aicc
[1] 217.083
```