STAT 804: Notes on Lecture 9

Fitting higher order autoregressions

For the model

$$X_t - \mu = \sum_1^p a_i (X_{t-1} - \mu) + \epsilon_t$$

we will use conditional likelihood again. Let ϕ denote the vector $(a_1, \ldots, a_p)^t$. Now we condition on the first p values of X and use

$$\ell_c(\phi, \mu, \sigma) = -\frac{1}{2\sigma^2} \sum_p^{T-1} \left[X_t - \mu - \sum_1^p a_i (X_{t-i} - \mu) \right]^2 - (T-p)\log(\sigma)$$

If we estimate μ using $\bar{X}$ we find that we are trying to maximize

$$-\frac{1}{2\sigma^2} \sum_p^{T-1} \left[X_t - \bar{X} - \sum_1^p a_i (X_{t-i} - \bar{X}) \right]^2 - (T-p)\log(\sigma)$$

To estimate $a_1, \ldots, a_p$ then we merely minimize the sum of squares

$$\sum_p^{T-1} \hat{\epsilon}_t^2 = \sum_p^{T-1} \left[X_t - \bar{X} - \sum_1^p a_i (X_{t-i} - \bar{X}) \right]^2$$

This is a straightforward regression problem. We regress the response vector

$$\begin{bmatrix} X_p - \bar{X} \\ \vdots \\ X_{T-1} - \bar{X} \end{bmatrix}$$

on the design matrix

$$\begin{bmatrix} X_{p-1} - \bar{X} & \cdots & X_0 - \bar{X} \\ \vdots & \vdots & \vdots \\ X_{T-2} - \bar{X} & \cdots & X_{T-p-1} - \bar{X} \end{bmatrix}$$

An alternative to estimating μ by $\bar{X}$ is to define $\alpha = \mu(1 - \sum a_i)$ and then recognize that

$$\ell(\alpha, \phi, \sigma) = -\frac{1}{2\sigma^2} \sum_p^{T-1} \left[X_t - \alpha - \sum_1^p a_i X_{t-i} \right]^2 - (T-p)\log(\sigma)$$

is maximized by regressing the vector

$$\begin{bmatrix} X_p \\ \vdots \\ X_{T-1} \end{bmatrix}$$

on the design matrix

$$\begin{bmatrix} X_{p-1} & \cdots & X_0 \\ \vdots & \vdots & \vdots \\ X_{T-2} & \cdots & X_{T-p-1} \end{bmatrix}$$

From $\hat{\alpha}$ and $\hat{\phi}$ we would get an estimate for μ by

$$\hat{\mu} = \frac{\hat{\alpha}}{1 - \sum \hat{a}_i}$$

Notice that if we put

$$Z = \begin{bmatrix} X_{p-1} - \bar{X} & \cdots & X_0 - \bar{X} \\ \vdots & \vdots & \vdots \\ X_{T-2} - \bar{X} & \cdots & X_{T-p-1} - \bar{X} \end{bmatrix}$$

then

$$Z^t Z \approx T \begin{bmatrix} \hat{C}(0) & \hat{C}(1) & \cdots & \\ \hat{C}(1) & \hat{C}(0) & \cdots & \cdots \\ \vdots & \cdots & \ddots & \cdots \\ \cdots & \cdots & \hat{C}(1) & \hat{C}(0) \end{bmatrix}$$

and if

$$Y = \begin{bmatrix} X_p - \bar{X} \\ \vdots \\ X_{T-1} - \bar{X} \end{bmatrix}$$

then

$$Z^t Y \approx T \begin{bmatrix} \hat{C}(1) \\ \vdots \\ \hat{C}(p) \end{bmatrix}$$

so that the normal equations (from least squares)

$$Z^t Z \phi = Z^T Y$$

are nearly the Yule-Walker equations again.

Full maximum likelihood

To compute a full mle of $\theta = (\mu, \phi, \sigma)$ you generally begin by finding preliminary estimates $\hat{\theta}$ say by one of the conditional likelihood methods above and then iterate via Newton-Raphson or some other scheme for numerical maximization of the log-likelihood.

Fitting $MA(q)$ models

Here we consider the model with known mean (generally this will mean we estimate $\hat{\mu} = \bar{X}$ and subtract the mean from all the observations):

$$X_t = \epsilon_t - b_1\epsilon_{t-1} - \cdots - b_q\epsilon_{t-q}$$

In general X has a $MVN(0, \Sigma)$ distribution and, letting ψ denote the vector of b_is we find

$$\ell(\psi, \sigma) = -\frac{1}{2}\left[\log(\det(\Sigma)) - X^T\Sigma^{-1}X\right]$$

Here X denotes the column vector of all the data. As an example consider $q = 1$ so that

$$\Sigma = \begin{bmatrix} \sigma^2(1+b_1^2) & -b_1\sigma^2 & 0 & \cdots & \cdots \\ -b_1\sigma^2 & \sigma^2(1+b_1^2) & -b_1\sigma^2 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \cdots \\ 0 & \cdots & \cdots & -b_1\sigma^2 & \sigma^2(1+b_1^2) \end{bmatrix}$$

It is not so easy to work with the determinant and inverse of matrices like this. Instead we try to mimic the conditional inference approach above but with a twist; we now condition on something we haven't observed — ϵ_{-1}.

Notice that

$$\begin{aligned} X_0 &= \epsilon_0 - b\epsilon_{-1} \\ X_1 &= \epsilon_1 - b\epsilon_0 \\ &= \epsilon_1 - b(X_0 + \epsilon_{-1}) \\ X_2 &= \epsilon_2 - b\epsilon_1 \\ &= \epsilon_2 - b(X_1 + b(X_0 + \epsilon_{-1})) \\ &\vdots \\ X_{t-1} &= \epsilon_{T-1} - b(X_{T-2} + b(X_{T-3} + \cdots \epsilon_{-1})) \end{aligned}$$

Now imagine that the data were actually

$$\epsilon_{-1}, X_0, \ldots, X_{T-1}$$

Then the same idea we used for an $AR(1)$ would give

$$\begin{aligned} \ell(b,\sigma) &= \log(f(\epsilon_{-1},\sigma)) + \log(f(X_0,\ldots,X_{T-1}|\epsilon_{-1},b,\sigma) \\ &= \log(f(\epsilon_{-1},\sigma)) + \sum_0^{T-1}\log(f(X_t|X_{t-1},\ldots,X_0,\epsilon_{-1},b,\sigma) \end{aligned}$$

The parameters are listed in the conditions in this formula merely to indicate which terms depend on which parameters. For Gaussian ϵs the terms in this likelihood are squares as usual (plus logarithms of σ) leading to

$$\begin{aligned} \ell(b,\sigma) =& \frac{-\epsilon_{-1}}{2\sigma^2} - \log(\sigma) \\ &- \sum_0^{T-1}\left[\frac{1}{2}(X_t - \psi X_{t-1} - \psi^2 X_{t-2} - \cdots - \psi^{t+1}\epsilon_{-1})^2 + \log(\sigma)\right] \end{aligned}$$

We will estimate the parameters by maximizing this function after getting rid of ϵ_{-1} somehow.
Method A: Put $\epsilon_{-1} = 0$ since 0 is the most probable value and maximize

$$-T\log(\sigma) - \frac{1}{2\sigma^2}\sum_0^{T-1}\left[X_t - \psi X_{t-1} - \psi^2 X_{t-2} - \cdots - \psi^t X_0\right]^2$$

Notice that for large T the coefficients of ϵ_{-1} are close to 0 for most t and the remaining few terms are negligible relatively to the total.
Method B: **Backcasting** is the process of guessing ϵ_{-1} on the basis of the data; we replace ϵ_{-1} in the log likelihood by

$$\mathrm{E}(\epsilon_{-1}|X_0,\ldots,X_{T-1}).$$

The problem is that this quantity depends on b and σ.

We will use the **EM algorithm** to solve this problem. The algorithm can be applied when we have (real or imaginary) missing data. Suppose the data we have is X; some other data we didn't get is Y and $Z = (X,Y)$. It often happens that we can think of a Y we didn't observe in such a way that the likelihood for the whole data set Z would be simple. In that case we can try to maximize the likelihood for X by following a two step algorithm first discussed in detail by Dempster, Laid and Rubin. This algorithm has two steps:

1. The **E** or **Estimation** step. We "estimate" the missing data Y by computing $\mathrm{E}(Y|X)$. Technically, we are supposed to estimate the likelihood function based on Z. Factor the density of Z as

   $$f_Z = f_{Y|X} f_X$$

   and take logs to get

   $$\ell(\theta|Z) = \log(f_{Y|X}) + \ell(\theta|X)$$

   We actually estimate the log conditional density (which is a function of θ) by computing

   $$\mathrm{E}_{\theta_0}(\log(f_{Y|X})|X)$$

   Notice the subscript θ_0 on E. This indicates that you have to know the parameter to compute the conditional expectation. Notice too that there is another θ in the conditional expectation – the log conditional density has a parameter in it.

2. We then maximize our estimate of $\ell(\theta|Z)$ to get a new value θ_1 for θ. Go back to step 1 with this θ_1 replacing θ_0 and iterate.

To get started we need a preliminary estimate. Now look at our problem. In our case the quantity Y is ϵ_{-1}. Rather than work with the log-likelihood directly we work with Y. Our preliminary estimate of Y is 0. We use this value to estimate θ as above getting an estimate θ_0. Then we compute $\mathrm{E}_{\theta_0}(\epsilon_{-1}|X)$ and replace ϵ_{-1} in the log-likelihood above by this conditional expectation. Then iterate. This process of guessing ϵ_{-1} is called backcasting.
Summary

- The log likelihood based on $\epsilon_{-1}, X_0, \ldots, X_{T-1}$ is

$$\frac{-\epsilon_{-1}^2}{2\sigma^2} - (T+1)\log(\sigma) - \frac{1}{2}\sum_{0}^{T-1}(X_t - \psi X_{t-1} - \psi^2 X_{t-2} - \cdots - \psi^{t+1}\epsilon_{-1})^2$$

- Put $\epsilon_{-1} = 0$ in this formula and estimate ψ by minimizing

$$\sum \hat{\epsilon}_t^2$$

where

$$\hat{\epsilon}_t = X_t - \psi X_{t-1} - \psi^2 X_{t-2} - \cdots - \psi^t X_0$$

for $t = 0, \ldots, T-1$.

- Now compute $\mathrm{E}(\epsilon_{-1}|X_0, \ldots, X_{T-1})$ Box, Jenkins and Reinsel presents an algorithm to do so based on the fact that there are actually two MA representations of corresponding to a given covariance function (the invertible one and a non-invertible one). The non-invertible representation is

$$X_t = e_t + \frac{1}{\psi}e_{t+1};$$

this form can be used to carry out the computation of the conditional expection.

- Iterate, re-estimating ψ and recomputing the backcast value of ϵ_{-1} if needed.