STAT 804: Lecture 8

Simplify maximum likelihood problem several ways:

Usually simply estimate $\widehat{\mu} = \bar{X}$.

Term f_{X_0} in likelihood is different in structure and causes considerable trouble. We drop it.

Result called conditional likelihood:

Consider general statistical inference problem:

If data written in form $X = (Y, Z)$ then can factor density:

$$f_X(x) = f_{Y|Z}(y|z) f_Z(z)$$

First term in factorization, $f_{Y|Z}(y|z)$, is called a **conditional likelihood** (when you think of it as a function of the unknown parameters)

Second term, $f_Z(z)$, is called a **marginal likelihood**.

Sometimes one or the other of the two terms is conveniently simpler than the full likelihood; in these cases people often suggest using the simple piece.

You get less efficient estimates in general but sometimes the loss is not very important.

$AR(1)$ case: Y is $(X_1, \ldots, X_{T-1})$ while Z is X_0. Our conditional log-likelihood is

$$\begin{aligned}\ell(\mu, \rho, \sigma) &= \sum_1^{T-1} \log(f_{X_t|X_0,\ldots,X_{t-1}}) \\ &= \frac{-1}{2\sigma^2} \sum_1^{T-1} \left[X_t - \mu - \rho(X_{t-1} - \mu)\right]^2 \\ &\quad - (T-1)\log(\sigma).\end{aligned}$$

Combining previous two ideas leads to maximization of

$$\ell(\bar{X}, \rho, \sigma) = \frac{-1}{2\sigma^2} \sum_{1}^{T-1} \left[X_t - \bar{X} - \rho(X_{t-1} - \bar{X})\right]^2 - (T-1)\log(\sigma)$$

This may be maximized explicitly to get

$$\widehat{\rho} = \frac{\sum_1^{T-1}(X_t - \bar{X})(X_{t-1} - \bar{X})}{\sum_0^{T-2}(X_t - \bar{X})^2}$$

and

$$\widehat{\sigma}^2 = \frac{\sum_1^{T-1}\left[X_t - \bar{X} - \widehat{\rho}(X_{t-1} - \bar{X})\right]^2}{T-1}$$

Changing range of summation in previous formula for $\widehat{\rho}$ to include all possible terms gives

$$\widehat{\rho} = \frac{\sum_1^{T-1}(X_t - \bar{X})(X_{t-1} - \bar{X})}{\sum_0^{T-1}(X_t - \bar{X})^2} = \frac{\widehat{C}(1)}{\widehat{C}(0)}$$

Notice: many suggestions for simplifications and adjustments.

Typical of statistical research – many ideas, only slightly different from each other, are suggested and compared.

In practice: seems likely there is very little difference between the methods.

Homework problem to investigate differences between several of these methods on a single data set.

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 vector $(a_1, \ldots, a_p)^t$.

Condition on first p values of X; 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 $a_1, \ldots, a_p$ minimize sum of squares

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

Regression problem: regress 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

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

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

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

Notice that if we put (back to $\widehat{\mu} = \bar{X}$)

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

then

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

and if

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

then

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

so 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 $\theta = (\mu, \phi, \sigma)$:

Begin by finding preliminary estimates $\widehat{\theta}$ say by one of the conditional likelihood methods above

Then iterate via say Newton-Raphson or other scheme for numerical maximization.

Fitting $MA(q)$ models

Here we consider the model with known mean (generally this will mean we estimate $\widehat{\mu} = \bar{X}$ 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.

Letting ψ denote vector of b_is get

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

Here X denotes the column vector of all the data.

As an example consider $q = 1$ so that Σ/σ^2 is

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

It is not so easy to work with the determinant and inverse of matrices like this.

Instead: mimic conditional inference approach above but with a twist; we now condition on something we haven't observed — ϵ_{-1}.

Notice that

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

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

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

The parameters are listed in the conditions in this formula merely to indicate which terms depend on which parameters.

Gaussian ϵs: terms in likelihood are squares as usual (plus logarithms of σ) so

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

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

$$-T\log(\sigma) - \frac{1}{2\sigma^2}\sum_{0}^{T-1}\left[X_t + bX_{t-1} + b^2X_{t-2} + \cdots + b^tX_0\right]^2$$

Note: for large T coefficients of ϵ_{-1} are close to 0 for most t; remaining few terms are negligible relatively to total.

Method B: **Backcasting**: process of guessing ϵ_{-1} on basis of data; replace ϵ_{-1} in the log likelihood by

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

Problem: this quantity depends on b and σ.

We will use the **EM algorithm** to solve this problem.