5.2: Maximum Likelihood Estimation

Taylor

University of Virginia

Maximum Likelihood Estimation

We now assume further that our time series is a Multivariate Normal random vector. Now that we have a specific probability density to work with, we can talk about maximizing a likelihood. There is not always a closed form solution for this task, so we may have to use an iterative scheme (e.g. Newton-Raphson, gradient descent, etc.). In the previous chapter, the techniques we learned about fell under the heading of "Preliminary Estimation" because they can provide us with starting parameter values for an iterative scheme.

Maximum Likelihood Estimation

Let $\mathbf{X}_n = (X_1, \ldots, X_n)'$ be the vector of our zero-mean univariate time series. Let $\Gamma_n = E[\mathbf{X}_n \mathbf{X}_n']$ be the covariance matrix for this random vector. Our likelihood is

$$L(\Gamma_n) = (2\pi)^{-n/2} (\det \Gamma_n)^{-1/2} \exp\left(-\frac{1}{2} \mathbf{X}_n' \Gamma_n^{-1} \mathbf{X}_n \right).$$

Evaluating this (or the log of this) may be difficult for long time series (large n) because we have to invert and find the determinant of a very large covariance matrix.

Maximum Likelihood Estimation

Solution: instead of looking at the density for the data, make a transformation and look at the density for the innovations. Recall from 2.5.4 that

$$\mathbf{X}_n = \mathbf{C}_n(\mathbf{X}_n - \hat{\mathbf{X}}_\mathbf{n}) = \mathbf{C}_n\mathbf{U}_n.$$

$\mathbf{C}_n$ was lower-diagonal with 1s on the diagonal. Also recall the transformation theorem from STAT 3120 (I drop the bold face):

$$\begin{aligned} f_{U_n}(u_n) &= f_{X_n}(x_n[u_n])\,|\det C_n| \\ &= (2\pi)^{-n/2}(\det \Gamma_n)^{-1/2} \exp\left(-\frac{1}{2}U_n' C_n' \Gamma_n^{-1} C_n U_n\right) |\det C_n| \\ &= (2\pi)^{-n/2}(v_0 \cdots v_{n-1})^{-1/2} \exp\left(-\frac{1}{2}U_n' D_n^{-1} U_n\right) \end{aligned}$$

Where $D_n = \text{diag}(v_0, \ldots, v_{n-1})$.

Maximum Likelihood Estimation

The likelihood factors up into pieces, so it's usually calculated with the innovations algorithm

$$
\begin{aligned}
L(\phi, \boldsymbol{\theta}, \sigma^2) &= (2\pi)^{-n/2}(v_0 \cdots v_{n-1})^{-1/2} \exp\left(-\frac{1}{2} U_n' D_n^{-1} U_n\right) \\
&= \frac{1}{\sqrt{(2\pi\sigma^2)^n r_0 r_1 \cdots r_{n-1}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\frac{(X_j - \hat{X}_j)^2}{r_{j-1}}\right\} \\
&= \prod_{j=1}^{n}\left[(2\pi\sigma^2 r_{j-1})^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\frac{(X_j - \hat{X}_j)^2}{r_{j-1}}\right\}\right]
\end{aligned}
$$

where $\sigma^2 r_j = v_j$.

Maximum Likelihood Estimation

A homework question: take the partial derivative of the log-likelihood $\ell(\phi, \theta)$, set it equal to 0, and deduce that:

MLE

$$\hat{\sigma}^2 = n^{-1} S(\hat{\phi}, \hat{\theta})$$

is the estimate for σ^2, where $S(\hat{\phi}, \hat{\theta}) = \sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{r_{j-1}}$, and the estimators $\hat{\phi}$ and $\hat{\theta}$ are the minimizers of

$$\ell(\phi, \theta) = \ln(n^{-1} S(\phi, \theta)) + n^{-1} \sum_{j=1}^{n} \ln r_{j-1}.$$

PS: this technique is sometimes called using a "profile" or "concentrated" log-likelihood.

Maximum Likelihood Estimation

As we mentioned earlier, minimization of $\ell(\boldsymbol{\phi}, \boldsymbol{\theta})$ is sometimes accomplished numerically. What this means is that the statistician first picks an initial set of values for the parameters (using a technique from the previous section is advised). Then the procedure will revise these parameters over and over again. The parameters get "better and better" in the sense that every subsequent evaluation of the likelihood will be bigger and bigger with these new parameters. The algorithm stops when the likelihood no longer improves.

The specifics of these optimization algorithms will not be discussed further in this class.

Least Squares Estimation

Least Squares

The least squares estimates $\tilde{\phi}, \tilde{\theta}$ for ϕ, θ are obtained by minimizing

$$S(\hat{\phi}, \hat{\theta}) = \sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{r_{j-1}}.$$

This ignores the other part of the likelihood, which does involve ϕ, θ.

Model Selection with AICC

How do we pick which model to fit? That is, how do we pick p and q? Introducing the corrected Akaike Information Criterion:

AICC Criterion

Choose p, q, ϕ_p, θ_q to minimize

$$-2 \ln L(\boldsymbol{\phi}_p, \boldsymbol{\theta}_q) + 2(p+q+1)n/(n-p-q-2).$$

- $2(p+q+1)n/(n-p-q-2)$ is the penalty portion for large p and q.
- This is discussed further in 5.5

Confidence Intervals

The MLE for $\boldsymbol{\beta} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$ is asymptotically normal.

MLE Sampling Distribution

$$\hat{\boldsymbol{\beta}} \overset{\text{approx.}}{\sim} \text{Normal}(\boldsymbol{\beta}, n^{-1} V(\boldsymbol{\beta})).$$

- $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' n V^{-1}(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_{p+q}$
- This is good for confidence intervals and hypothesis testing.

Confidence Intervals

From last slide

$$\hat{\beta} \overset{\text{approx.}}{\sim} \text{Normal}(\beta, n^{-1}V(\beta)).$$

$$n(\hat{\beta}-\beta)'V^{-1}(\beta)(\hat{\beta}-\beta) \sim \chi^2_{p+q}$$

1. confidence intervals and hypothesis tests for individual β elements are easy
2. $\left\{\beta : n(\hat{\beta}-\beta)'V^{-1}(\beta)(\hat{\beta}-\beta) \leq \chi^2_{p+q,\alpha}\right\}$ is a confidence ellipsoid

Confidence Intervals

MLE Sampling Distribution

$$\hat{\boldsymbol{\beta}} \overset{\text{approx.}}{\sim} \text{Normal}(\boldsymbol{\beta}, n^{-1}V(\boldsymbol{\beta})).$$

- $n^{-1}V(\boldsymbol{\beta})$ is approximated by $2H^{-1}$ where H is the Hessian matrix
- in R, a Hessian can be returned by the optimization function
- ```
  est <- optim(initialParams,
      negTwiceLogLikeFunc, NULL,
      method="BFGS",
      hessian=TRUE)
  ```
- $H_{ij} = \frac{\partial^2 \ell(\beta)}{\partial \beta_i \partial \beta_j}$ for $i, j = 1, \ldots, p+q$
- this is only justified if you are *minimizing* $\ell(\beta) = -2\log L(\boldsymbol{\beta})$
- this is called the *observed* Fisher information
  ```