Maximized Likelihood Ratio Tests

BS2 Statistical Inference, Lecture 13
Michaelmas Term 2004

Steffen Lauritzen, University of Oxford; November 23, 2004

The maximized likelihood ratio test

Recall that the MLRT (or LRT for short), has critical region of the form

$$\Lambda = \lambda(X) = -2 \log \frac{L(\theta_0; X)}{L(\hat{\theta}; X)} > K.$$

One particular advantage of this test statistic is that it immediately extends to the multiparameter case, as the definition of this statistic as above has no reference to the dimension of the parameter space.

Indeed, the LRT also has easy extension to the case where the null hypothesis $H_0 : \theta \in \Theta_0$ is *composite*, i.e. where Θ_0 has more than one value.

For a composite hypothesis we define the likelihood ratio test statistic as

$$\lambda(x) = -2\log\frac{\sup_{\theta\in\Theta_0} L(\theta;x)}{\sup_{\theta\in\Theta} L(\theta;x)} = -2\log\frac{L(\hat{\hat{\theta}};x)}{L(\hat{\theta};x)},$$

where $\hat{\theta} = \arg\max_{\theta\in\Theta}$ and $\hat{\hat{\theta}} = \arg\max_{\theta\in\Theta_0}$.

As $\Theta_0 \subseteq \Theta$, we would generally have that

$$\hat{\theta} \in \Theta_A, \quad \sup_{\theta\in\Theta} L(x;\theta) = \sup_{\theta\in\Theta_A} L(x;\theta) = L(x;\hat{\theta}).$$

The MLRT can thus bee seen as a standard LRT, just *comparing the most likely value of θ within the hypothesis with the most likely value within the alternative.*

Example

Consider the case where $X = (X_1, \ldots, X_n)$ is a sample from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, where $\mu > 0$ and $\sigma^2 > 0$ are unknown and consider the hypothesis

$$H_0 : \sigma^2 = \mu^2.$$

In this case we have, if $\bar{x} > 0$ that

$$\hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_i (x_i - \hat{\mu})^2 = SS/n - \bar{x}^2.$$

If $\bar{x} \leq 0$ the MLE does not exist as we have restricted the parameter space to have $\mu > 0$.

We get

$$\log L(\hat{\mu}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \hat{\sigma}^2 - n/2$$

and

$$\log L(\hat{\hat{\mu}}) = -\frac{n}{2}\log(2\pi) - n\log\hat{\hat{\mu}} - \frac{SS}{2\hat{\hat{\mu}}^2} + \frac{S}{\hat{\hat{\mu}}} - n/2$$

where

$$\hat{\hat{\mu}} = \frac{-S + \sqrt{S^2 + 4nSS}}{2n}, \quad \hat{\hat{\sigma}}^2 = \hat{\hat{\mu}}^2$$

so

$$\Lambda = n\log\frac{\hat{\hat{\sigma}}^2}{\hat{\sigma}^2} + \left(\frac{SS}{\hat{\hat{\mu}}^2} - \frac{2S}{\hat{\hat{\mu}}}\right).$$

The first term reflects differences in the variance estimates whereas the second term directly reflects deviations from the hypothesis.

Asymptotic properties

The LRT is extremely useful and therefore widely used. Not only does it apply to a wide range of testing problems, but its large sample distribution is generally very simple.

Indeed if Θ is an open and connected subset of $\mathcal{R}^d$ and Θ_0 is specified by restriction of the parameter space as

$$\theta \in \Theta_0 \iff h(\theta) = 0$$

where the function

$$h(\theta) = (h_1(\theta), \ldots, h_k(\theta))$$

is twice continuously differentiable and its matrix of derivatives

$$H(\theta) = \left\{ \frac{\partial h_r(\theta)}{\partial \theta_s} \right\}$$

has constant and full rank k for $\theta \in \Theta_0$ and if further the usual regularity conditions are fulfilled, *it holds that*

$$\Lambda \xrightarrow{D} Y$$

where Y follows a χ^2-distribution with k degrees of freedom.

The case of a simple hypothesis has $k = d$ and $h(\theta) = \theta - \theta_0$ and *the asymptotic distribution of Λ is then $\chi^2(d)$*.

The proof of this general result is a bit involved so we shall only look at the case of a simple hypothesis.

In the example just considered, we could let $h(\mu, \sigma^2) = \sigma^2 - \mu^2$ so here Λ is asymptotically distributed as $\chi^2(1)$.

The case of a simple hypothesis

Here is a sketch of the proof in the case of a simple hypothesis $\Theta_0 = \{\theta_0\}$.

First we use Taylor's theorem to expland $\log L(\theta_0)$ around the MLE $\hat{\theta}$:

$$\begin{aligned} \log L(\theta_0) &= \log L(\hat{\theta}) + S(\hat{\theta})(\theta_0 - \hat{\theta}) \\ &\quad - \frac{1}{2}(\theta_0 - \hat{\theta})^\top j(\hat{\theta})(\theta_0 - \hat{\theta}) + R(\theta_0, \hat{\theta}). \end{aligned}$$

Now ignore the remainder term and use that $S(\hat{\theta}) = 0$:

$$\log L(\theta_0) \approx \log L(\hat{\theta}) - \frac{1}{2}(\theta_0 - \hat{\theta})^\top j(\hat{\theta})(\theta_0 - \hat{\theta}).$$

Move $\log L(\hat{\theta})$ to the other side of this equation and multiply by -2 to obtain

$$\Lambda \approx (\theta_0 - \hat{\theta})^\top j(\hat{\theta})(\theta_0 - \hat{\theta}).$$

Slutsky's theorem in combination with the facts that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} \mathcal{N}\{0, i(\theta_0)^{-1}\}, \quad j(\hat{\theta})/n \xrightarrow{P} i(\theta_0)$$

now yields the result.

Approximate likelihood ratio tests

The proof of the asymptotic result for the likelihood ratio test, also in the general case, essentially relies upon approximating Λ by a quadratic form as

$$\Lambda \approx n(\hat{\theta} - \hat{\hat{\theta}})^\top C(\hat{\theta} - \hat{\hat{\theta}})$$

where C is a consistent estimate of the information matrix at the true value $i(\theta_0)$.

As in the one-parameter case it also makes sense to use the right hand side directly as a test statistic.

There are essentially four possibilities for the choice of C:

$$C_1 = i(\hat{\theta}), \quad C_2 = i(\hat{\hat{\theta}}), \quad C_3 = j_n(\hat{\theta})/n, \quad C_4 = j_n(\hat{\hat{\theta}})/n,$$

all leading to test statistics which have an asymptotic $\chi^2(k)$ distribution if the null hypothesis is true.

This leads to the *Wald statistics*

$$W = n(\hat{\theta} - \hat{\hat{\theta}})^\top i(\hat{\theta})(\hat{\theta} - \hat{\hat{\theta}}), \quad \tilde{W} = (\hat{\theta} - \hat{\hat{\theta}})^\top j_n(\hat{\theta})(\hat{\theta} - \hat{\hat{\theta}})$$

or the χ^2 *statistics*

$$X^2 = n(\hat{\theta} - \hat{\hat{\theta}})^\top i(\hat{\hat{\theta}})(\hat{\theta} - \hat{\hat{\theta}}), \quad \tilde{X}^2 = (\hat{\theta} - \hat{\hat{\theta}})^\top j_n(\hat{\hat{\theta}})(\hat{\theta} - \hat{\hat{\theta}}).$$

As in the one-parameter case, care should be taken with the Wald statistic as it may have very undesirable properties when the hypothesis is far from being true.