Likelihood Ratio for Linear Models

April 9, 2019

Likelihood ratio for Linear Models

- Recall the linear model:

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w}$$

- $\boldsymbol{w} = [w[0] ... w[N-1]]^T$ is a vector of i.i.d. samples with joint pdf $\boldsymbol{w} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$
- $\boldsymbol{H}$ is the known model matrix.
- We want to test the simple hypotheses:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{0}$$

$$H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1, \text{ known}$$

- Here we assume we know $\boldsymbol{\theta}$, whereas previously in class we have attempted to estimate it.

Likelihood Ratio

- The likelihood ratio is:

$$\frac{p(x|H_1)}{p(x|H_0)} = \frac{\frac{1}{(2\pi)^{N/2}\sqrt{\det \sigma^2 \boldsymbol{I}}} \exp[-1/(2\sigma^2)(\boldsymbol{x} - \boldsymbol{H\theta_1})^T(\boldsymbol{x} - \boldsymbol{H\theta_1})]}{\frac{1}{(2\pi)^{N/2}\sqrt{\det \sigma^2 \boldsymbol{I}}} \exp[-1/(2\sigma^2)\boldsymbol{x}^T\boldsymbol{x}]} \underset{}{\overset{H_1}{\gtrless}} \lambda$$

where $\lambda = \frac{\pi_0 L(1|0)}{\pi_1 L(0|1)}$ is the bayesian threshold for the likelihood ratio test.

Likelihood Ratio

- The likelihood ratio is:

$$\frac{p(x|H_1)}{p(x|H_0)} = \frac{\frac{1}{(2\pi)^{N/2}\sqrt{\det \sigma^2 \boldsymbol{I}}}\exp[-1/(2\sigma^2)(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta_1})^T(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\theta_1})]}{\frac{1}{(2\pi)^{N/2}\sqrt{\det \sigma^2 \boldsymbol{I}}}\exp[-1/(2\sigma^2)\boldsymbol{x}^T\boldsymbol{x}]} \overset{H_1}{\gtrless} \lambda$$

  where $\lambda = \frac{\pi_0 L(1|0)}{\pi_1 L(0|1)}$ is the bayesian threshold for the likelihood ratio test.
- This simplifies, after some work to:

$$\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{H}\boldsymbol{\theta_1}/\sigma^2 \overset{H_1}{\gtrless} \ln\lambda + \frac{\boldsymbol{\theta_1^T H^T H\theta_1}}{2\sigma^2}$$

Likelihood Ratio

- The likelihood ratio is:

$$\frac{p(x|H_1)}{p(x|H_0)} = \frac{\frac{1}{(2\pi)^{N/2}\sqrt{\det \sigma^2 \boldsymbol{I}}} \exp[-1/(2\sigma^2)(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta_1})^T(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta_1})]}{\frac{1}{(2\pi)^{N/2}\sqrt{\det \sigma^2 \boldsymbol{I}}} \exp[-1/(2\sigma^2)\boldsymbol{x}^T\boldsymbol{x}]} \overset{H_1}{\gtrless} \lambda$$

  where $\lambda = \frac{\pi_0 L(1|0)}{\pi_1 L(0|1)}$ is the bayesian threshold for the likelihood ratio test.
- This simplifies, after some work to:

$$\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{H}\boldsymbol{\theta_1}/\sigma^2 \overset{H_1}{\gtrless} \ln \lambda + \frac{\boldsymbol{\theta_1^T H^T H \theta_1}}{2\sigma^2}$$

- $\boldsymbol{T}(\boldsymbol{x})$ is a normal random variable, so for the NP test we need to find the means and variances under each hypothesis to get P_D and P_F.

Mean and Variance under H_0

- For H_0 we have $\boldsymbol{x} = \boldsymbol{H}\boldsymbol{0} + \boldsymbol{w} = \boldsymbol{w}$:

$$\mathbb{E}(\boldsymbol{T}(\boldsymbol{x})|H_0) = \mathbb{E}(\frac{\boldsymbol{w}^T\boldsymbol{H}\boldsymbol{\theta_1}}{\sigma^2}) = \boldsymbol{0}$$

$$\mathrm{var}(\boldsymbol{T}(\boldsymbol{X})|H_0) = \mathbb{E}[(\frac{\boldsymbol{w}^T\boldsymbol{H}\boldsymbol{\theta_1}}{\sigma^2})^T\frac{\boldsymbol{w}^T\boldsymbol{H}\boldsymbol{\theta_1}}{\sigma^2})] = \mathbb{E}(\frac{\boldsymbol{\theta_1^T}\boldsymbol{H^T}}{\sigma^2}\boldsymbol{w}\boldsymbol{w^T}\frac{\boldsymbol{H}\boldsymbol{\theta_1}}{\sigma^2}) =$$

$$\frac{\boldsymbol{\theta_1^T}\boldsymbol{H^T}}{\sigma^2}\mathbb{E}(\boldsymbol{w}\boldsymbol{w^T})\frac{\boldsymbol{H}\boldsymbol{\theta_1}}{\sigma^2} = \frac{\boldsymbol{\theta_1^T}\boldsymbol{H^T}}{\sigma^2}\sigma^2\boldsymbol{I}\frac{\boldsymbol{H}\boldsymbol{\theta_1}}{\sigma^2} = \frac{\boldsymbol{\theta_1^T}\boldsymbol{H^T}\boldsymbol{H}\boldsymbol{\theta_1}}{\sigma^2}$$

- So $p(T(x)|H_0) \sim \mathcal{N}(\boldsymbol{0}, \frac{\boldsymbol{\theta_1^T}\boldsymbol{H^T}\boldsymbol{H}\boldsymbol{\theta_1}}{\sigma^2})$

Probability of False Alarm

- Let $\lambda' = \ln\lambda + \frac{\boldsymbol{\theta_1^T H^T H \theta_1}}{2\sigma^2}$
- Then the probablity of false alarm is:

$$P_F = Pr(\boldsymbol{T}(\boldsymbol{x}) > \lambda' | H_0) =$$

$$Pr(\underbrace{\frac{(\boldsymbol{T}(\boldsymbol{x}) - 0)\sigma^2}{\boldsymbol{\theta_1^T H^T H \theta_1}}}_{N(0,1)} > \frac{(\lambda' - 0)\sigma^2}{\boldsymbol{\theta_1^T H^T H \theta_1}}) =$$

$$1 - \Phi(\frac{(\ln\lambda + \frac{\boldsymbol{\theta_1^T H^T H \theta_1}}{2\sigma^2})\sigma^2}{\boldsymbol{\theta_1^T H^T H \theta_1}}) = Q(\frac{(\ln\lambda + \frac{\boldsymbol{\theta_1^T H^T H \theta_1}}{2\sigma^2})\sigma^2}{\boldsymbol{\theta_1^T H^T H \theta_1}})$$

where Φ is the cumulative distribution function for the standard normal distribution.

Mean and Variance under H_1

- For H_1 use $\boldsymbol{x} = \boldsymbol{H\theta_1} + \boldsymbol{w}$:

$$\mathbb{E}(\boldsymbol{T}(\boldsymbol{x})|H_1) = \mathbb{E}(\frac{(\boldsymbol{H\theta_1} + \boldsymbol{w})^T \boldsymbol{H\theta_1}}{\sigma^2}) = \frac{\boldsymbol{\theta_1^T H^T H\theta_1}}{\sigma^2}$$

- The variance is a bit more complicated:

$$\mathrm{var}(\boldsymbol{T}(\boldsymbol{x})|H_1) = \mathbb{E}((\frac{\boldsymbol{x^T H\theta_1}}{\sigma^2} - \mathbb{E}(\frac{\boldsymbol{x^T H\theta_1}}{\sigma^2}))^T(\frac{\boldsymbol{x^T H\theta_1}}{\sigma^2} - \mathbb{E}(\frac{\boldsymbol{x^T H\theta_1}}{\sigma^2}))) =$$

$$\mathbb{E}((\boldsymbol{x} - \mathbb{E}(x))^T \frac{\boldsymbol{H\theta_1}}{\sigma^2})^T (\boldsymbol{x} - \mathbb{E}(x))^T \frac{\boldsymbol{H\theta_1}}{\sigma^2})) =$$

$$\mathrm{cov}(\boldsymbol{x})(\frac{\boldsymbol{H\theta_1}}{\sigma^2})^T(\frac{\boldsymbol{H\theta_1}}{\sigma^2}) = \frac{\boldsymbol{\theta_1^T H^T H\theta_1}}{\sigma^2} = \mathrm{var}(\boldsymbol{T}(\boldsymbol{x})|H_0)$$

- So under H_1, $\boldsymbol{T}(\boldsymbol{x}) \sim N(\frac{\boldsymbol{\theta_1^T H^T H\theta_1}}{\sigma^2}, \frac{\boldsymbol{\theta_1^T H^T H\theta_1}}{\sigma^2})$, which has a different mean and the same variance as H_0

- The Probability of Detection is given by:

$$P_D = Pr(\boldsymbol{T}(\boldsymbol{x}) > \lambda' | H_1) = Pr(\underbrace{\frac{(\boldsymbol{T}(\boldsymbol{x}) - \frac{\boldsymbol{\theta_1^T H^T H \theta_1}}{\sigma^2})\sigma^2}{\boldsymbol{\theta_1^T H^T H \theta_1}}}_{N(0,1)} > \frac{(\lambda' - \frac{\boldsymbol{\theta_1^T H^T H \theta_1}}{\sigma^2})\sigma^2}{\boldsymbol{\theta_1^T H^T H \theta_1}}) =$$

$$1 - \Phi(\frac{(\ln\lambda - \frac{\boldsymbol{\theta_1^T H^T H \theta_1}}{2\sigma^2})\sigma^2}{\boldsymbol{\theta_1^T H^T H \theta_1}}) = Q(\frac{(\ln\lambda - \frac{\boldsymbol{\theta_1^T H^T H \theta_1}}{2\sigma^2})\sigma^2}{\boldsymbol{\theta_1^T H^T H \theta_1}})$$

- This is almost the same as P_F, with one subtraction instead of addition.
- Large λ leads to higher P_D and P_F
- $P_D = \Phi^{-1}(\Phi(P_F) - \sqrt{\frac{\boldsymbol{\theta_1^T \theta_1}}{\sigma^2}})$ is another what to express P_D in this case - proof is in Kay CH. 3 and 4.

Example: Sum of Sinusoids, Kay ex. 4.9

- $H_0 : x[n] = w[n]$
- $H_1 : \boldsymbol{x}[n] = a\cos(2\pi f_0 n) + b\sin(2\pi f_0 n) + w[n]$
- $\theta_1 = \begin{bmatrix} a \\ b \end{bmatrix}$
- $\boldsymbol{H} = \begin{bmatrix} 1 & 0 \\ \cos(2\pi f_0) & \sin(2\pi f_0) \\ \vdots & \vdots \\ \cos(2\pi f_0(N-1)) & sin(2\pi f_0(N-1)) \end{bmatrix}$
- For this example, let $N = 2$.

Example: Sum of Sinusoids, Kay ex. 4.9

- The test statistic is given by:

$$\frac{1}{\sigma^2}\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{H}\boldsymbol{\theta_1} = \frac{1}{\sigma^2}\begin{bmatrix} x[0] & x[1] \end{bmatrix}\begin{bmatrix} 1 & 0 \\ \cos(2\pi f_0) & \sin(2\pi f_0) \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} =$$

$$\frac{1}{\sigma^2}(ax[0] + x[1](a\cos(2\pi f_0) + b\sin(2\pi f_0))) =$$

$$\frac{1}{\sigma^2}(a(x[0] + x[1]\cos(2\pi f_0)) + b(\sin(2\pi f_0))) =$$

$$a\hat{a} + b\hat{b}$$

- Here $\hat{a}$ and $\hat{b}$ are the estimators of a and b based on the fourier series formula!
- Under H_0, $\hat{a}$ and $\hat{b}$ are very small, so $\boldsymbol{T}(x)$ is small.
- Under H_1, $\boldsymbol{T}(\boldsymbol{x}) \approx a^2 + b^2$. This is proportional to the signal power.

Comments

- What happens when $\boldsymbol{\theta_1}$ is unknown?
- We have to use the MLE estimator in something called the Generalized Likelihood Ratio Test.
- The hypotheses are slightly different:

$$H_0 : \boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{b}$$

$$H_1 : \boldsymbol{A}\boldsymbol{\theta} \neq \boldsymbol{b}$$

- This hypothesis test whether or not $\boldsymbol{\theta}$ lives in a particular subspace (e.g. line, plane, hyperplane)
- $\boldsymbol{A}$, $\boldsymbol{b}$ are known and form a consistent (solveable) set of equations.
- Then you apply similar formulas, but the covariance matrix of $\boldsymbol{T}$ changes because $\hat{\boldsymbol{\theta}}$ is a function of $\boldsymbol{x}$ and therefore a random variable.
- P_D and P_F become χ^2 distributions.

Comments on Testing Coefficients

- Suppose $H_0 = \boldsymbol{\theta}_0 = [0, \theta_1, \theta_2, ..., \theta_p]$
- And $H_1 = \boldsymbol{\theta}_1 = [\theta_1, \theta_2, \theta_3, ..., \theta_p]$ is almost exactly the same.
- The formulas are slightly more complicated since the mean of H_0 is no longer 0, but we can still apply the LRT. This is testing whether or not the data in the first column of $\mathbf{H}$ improves the model or not, and is a simple version of where the p-values next to coefficients in statistics software comes from.
- This is often done in Linear Regression - I will talk more about this case when we get to the Generalized LRT.
- In machine learning this is a type of Feature Selection - there are many ways to do this for linear regression.