Notes on computing the Fisher Information matrix for MARSS models. Part I Background

EE Holmes, National Marine Fisheries Service & University of Washington

2016-05-18

This is part of a series on computing the Fisher Information for Multivariate Autoregressive State-Space Models. Part I: Background, Part II: Louis 1982, Part III: Harvey 1989, Background, Part IV: Harvey 1989, Implementation.

Citation: Holmes, E. E. 2016. Notes on computing the Fisher Information matrix for MARSS models. Part I Background. Technical Report. https://doi.org/10.13140/RG.2.2.27306.11204/1

(Expected) Fisher Information

The Fisher Information is defined as

$$I(\theta) = E_{Y|\theta}\{[\partial \log L(\theta|Y)/\partial\theta]^2\} = \int_x [\partial \log L(\theta|y)/\partial\theta]^2 f(y|\theta)dy \tag{1}$$

In words, it is the expected value (taken over all possible data) of the square of the gradient (first derivative) of the log likelihood surface at θ. It is a measure of how much information data (from our experiment or monitoring) have about θ. The log-likelihood surface is for a fixed set of data and the θ vary. The peak is at the MLE, which is not θ, so the surface has some gradient (slope) at θ since the peak is at the MLE not θ. The Fisher Information is the expected value (over possible data) of those gradients (squared).

It can be shown[1] that the Fisher Information can also be written as

$$I(\theta) = -E_{Y|\theta}\{\partial^2 \log L(\theta|Y)/\partial\theta^2\} = -\int_y [\partial^2 \log L(\theta|y)/\partial\theta^2 f(y|\theta)dy \tag{2}$$

So the Fisher Information is the average (over possible data) convexity of the log-likelihood surface at θ. That doesn't quite make sense to me. When I imagine the surface, that the convexity at a non-peak value θ is not intuitively the information. The gradient squared, I understand, but the convexity at a non-peak?

Note, my y should be understood to be some multi-dimensional data set (multiple sites over multiple time points, say), and is comprised of multiple samples. Often in this case Fisher Information is written $I_n(\theta)$ and if the data points are all independent, $I(\theta) = \frac{1}{n} I_n(\theta)$. However I'm not using that notation. My $I(\theta)$ is referring to the Fisher Information for a dataset not individual data points within that data set.

We do not know θ so we need to use an estimator for the Fisher Information. A common approach is to use $I(\hat{\theta})$, the Fisher Information at the MLE θ as an estimator of $I(\theta)$ because:

$$I(\hat{\theta}) \xrightarrow{P} I(\theta) \tag{3}$$

This is called the *expected* Fisher Information and is computed at the MLE:

$$I(\hat{\theta}) = -E_{Y|\hat{\theta}}\{\partial^2 \log L(\theta|Y)/\partial\theta^2\}|_{\theta=\hat{\theta}} \tag{4}$$

That $|_{\theta=\hat{\theta}}$ at the end means that after doing the derivative with respect to θ, we replace θ with $\hat{\theta}$. It would not make sense to do the substitution before since $\hat{\theta}$ is a fixed value and so you cannot take the derivative with respect to it.

This is a viable approach if you can take the derivative of the log-likelihood with respect to θ and can take the expectation over the data. You could always do that expectation using simulation of course. You just need to be able to simulate data from your model with $\hat{\theta}$.

Observed Fisher Information

Another approach is to drop the expectation. This is termed the *observed* Fisher Information:

$$\mathcal{I}(\hat{\theta}, y) = - \left. \frac{\partial^2 \log L(\theta|y)}{\partial \theta^2} \right|_{\theta=\hat{\theta}} \tag{5}$$

where y is the one dataset we collected. The observed Fisher Information is the curvature of the log-likelihood function around the MLE. When you estimate the variance of the MLEs from the Hessian of the log-likelihood (output from say some kind of Newton method or any other algorithm that uses the Hessian of the log-likelihood), then you are using the observed Fisher Information matrix. Efron and Hinkley (1978) (and Fisher they say in their article) say that the observed Fisher Information is a better estimate of the variance of $\hat{\theta}$[2][3], while Cavanaugh and Shumway (1996) show results from MARSS models that indicate that the expected Fisher Information has lower mean squared error (though may be more biased; mean squared error measures both bias and precision).

Computing the Fisher Information

So how do we compute $I(\hat{\theta})$ or $\mathcal{I}(\hat{\theta}, y)$? In particular, I am interested in whether I can use the analytical derivatives of the full log-likelihood that are part of the EM algorithm to compute the Fisher Information. Notes on computing the Fisher Information matrix for MARSS models. Part II.

Endnotes

1. See any detailed write-up on Fisher Information. For example page 2 of these Lecture Notes on Fisher Information. ↩

2. The motivation for computing the Fisher Information is to get an estimate of the variance of $\hat{\theta}$ for standard errors on the parameter estimates, say. $var(\hat{\theta}) \xrightarrow{P} \frac{1}{I(\theta)}$. ↩

3. Note I am using the notation of Cavanaugh and Shumway (1996). Efron and Hinkley (1978) use $\mathscr{I}(\theta)$ for the expected Fisher Information and $I(\theta)$ for the observed Fisher Information. Cavanaugh and Shumway (1996) use $I(\theta)$ for the expected Fisher Information and $\mathcal{I}(\theta, Y)$ for the observed Fisher Information. I use the same notation as Cavanaugh and Shumway (1996) except that they use $I_n()$ and $\mathcal{I}_n$ to be explicit that the data have n data points. I drop the n since I am interested in the Fisher Information of the dataset not individual data points and if I need to use the information of the j-th data point, I would just write $I_j()$. The other difference is that I use y to refer to the data. In my notation, Y is the random variable 'data' and y is a particular realization of that random variable. In some cases, I use $y(1)$. That is when the random variable Y is only partially observed (meaning there are missing data points or time steps); $y(1)$ is the observed portion of Y. ↩

References I looked at while working on this

Fisher Information Lectures and Background

- http://people.missouristate.edu/songfengzheng/Teaching/MTH541/Lecture%20notes/Fisher_info.pdf
- http://www.math.umt.edu/patterson/Information.pdf
- http://www.stat.umn.edu/geyer/old03/5102/notes/fish.pdf
- Wikipedia Fisher Information page.
- Cavanaugh and Shumway (1996) have a succinct summary of Fisher Information in their introduction and I adopted their notation.

Papers

- Efron and Hinkley 1978. Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher Information. *This paper argues that the observed Fisher Information is better than expected Fisher Information in many/some cases. The same paper argues for using the likelihood ratio method for CIs.* PDF

- Cavanaugh and Shumway 1996. On computing the expected Fisher Information Matrix for state-space model parameters.