Test big

2016-05-18

Background on Fisher Information is in Part I.

Computing Fisher Information

So how do we compute $I(\hat{\theta})$ or $\mathcal{I}(\hat{\theta}, y)$ (in Part I)? In particular, can we use the analytical derivatives of the full log-likelihood that are part of the EM algorithm? Many researchers have worked on this idea. My notes here were influenced by EM Algorithm: Confidence Intervals which is on the same topic. This blog post is mainly a discussion of the result by Louis (1982) on calculation of the Fisher Information matrix from the 'score' function that one takes the derivative of in the M-step of the EM algorithm.

The 'score' function used in the EM algorithm for a MARSS model is

$$Q(\theta|\theta_j) = E_{X|y,\theta_j}[\log f_{XY}(X, y|\theta)] \tag{1}$$

It is the expected value taken over the hidden random variable X of the full data log-likelihood at $Y = y$ (3); full means it is a function of all the random variables in the model, which includes the hidden or latent variables. x, y is the full 'data', the left side of the x state equation and the y observation equation. We take the expectation of this full data likelihood conditioned on the observed data y and θ_j which is the value of θ at the j-th iteration of the EM algorithm. Although $Q(\theta|\theta_j)$ looks a bit hairy, actually the full-data likelihood may be very easy to write down and considerably easier than the data likelihood $f(y|\theta)$. The hard part is often the expectation step, however for MARSS models the Kalman filter-smoother algorithm computes the expectations involving X and Holmes (2010) shows how to compute the expectations involving Y, which comes up when there are missing values in the dataset (missing time steps, say).

In the M-step of the EM algorithm, we take the derivative of $Q(\theta|\theta_j)$ with respect to θ and solve for the θ where

$$\frac{\partial Q(\theta|\theta_j)}{\partial \theta} = 0. \tag{2}$$

It would be nice if one could use the following to compute the observed Fisher Information

$$-\left.\frac{\partial^2 Q(\theta|\hat{\theta})}{\partial \theta^2}\right|_{\theta=\hat{\theta}} \tag{3}$$

$Q(\theta|\hat{\theta})$ is our score function at the end of the EM algorithm, when $\theta = \hat{\theta}$. Q is a function of θ, the model parameters, and will have terms like $E(X|Y = y, \hat{\theta})$, the expected value of X conditioned on $Y = y$ and the MLE. Those are the expectations coming out of the Kalman filter-smoother. We take the second derivative of Q with respect to θ. That is straight-forward for the MARSS equations. You take the first derivative of Q with respect to θ, which you already have from the update or M-step equations, and take the derivative of that with respect to θ.

Conceptually, this

$$-\left.\frac{\partial^2 Q(\theta|\hat{\theta})}{\partial \theta^2}\right|_{\theta=\hat{\theta}} = \left.\frac{\partial^2 E_{X|y,\hat{\theta}}[\log f(X, y|\theta)]}{\partial \theta^2}\right|_{\theta=\hat{\theta}} \tag{4}$$

looks a bit like the observed Fisher Information:

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

except that instead of the data likelihood $f(y|\theta)$, we use the expected likelihood $E_{X|Y,\hat{\theta}}[\log f_{XY}(X, y|\theta)]$. The expected likelihood is the full likelihood with the X and $XX^\top$ random variables replaced by their expected values assuming $\theta = \hat{\theta}$ and $Y = y$. The problem is that $E_{X|Y,\theta}[\log f(X, y|\theta)]$ is a function of θ and by fixing it at $\hat{\theta}$ we are not accounting for the uncertainty in that expectation. What we need is something like

Information with X fixed at expected value - Information on expected value of X

We account for the fact that we have over-estimated the information from the data by treating the hidden random variable as fixed. The same issue arises when we compute confidence intervals using the estimate of the variance without accounting for the fact that this is an estimate and thus has uncertainty. Louis (1982) and Oakes (1999) are concerned with how to do this correction or adjustment.

Louis 1982 approach

The following is equations 3.1, 3.2 and 3.3 in Louis (1982) translated to the MARSS case. In the MARSS model, we have two random variables, $X(t)$ and $Y(t)$. The joint distribution of $\{X(t), Y(t)\}$ conditioned on $X(t-1)$ is multivariate normal. Our full data set includes all time steps, $\{X, Y\}$.

Let's call the full state at time t $\{x, y\}$, the value of the X and Y at all times t. The full state can be an unconditional random variable, $\{X, Y\}$ or a conditional random variable $\{X, y\}$ (conditioned on $Y = y$. Page 227 near top of Louis 1982 becomes

$$\lambda(x, y, \theta) = \log\{f_{XY}(x, y|\theta)\} \tag{6}$$

$$\lambda^*(y, \theta) = \log\{f_Y(y|\theta)\} = \log \int_X f_{XY}(x, y|\theta) dx \tag{7}$$

$f(.|\theta)$ is the probability distribution of the random variable conditioned on θ. λ is the full likelihood; 'full' means is includes both x and y. λ^* is the likelihood of y alone. It is defined by the marginal distribution of y (1); the integral over X on the right side of 7. For a MARSS model, the data likelihood can be written easily as a function of the Kalman filter recursions (which is why you can write a recursion for the information matrix based on derivatives of λ^*; see Part III).

Next equation down. Louis doesn't say this and his notation is not totally clear, but the expectation right above section 3 (and in his eqn 3.1) is a conditional expectation. This is critical to know to follow his derivation of equation 3.1 in the appendix. θ_j is his $\theta(0)$; it is the value of θ at the last EM iteration.

$$E_{X|y,\theta_j}[\lambda(X, y, \theta)] = \int_X \lambda(X, y, \theta) f_{X|Y}(x|Y = y, \theta_j) dx \tag{8}$$

My 'expectation' notation is a little different than Louis'. The subscript on the E shows what is being integrated (X) and what are the conditionals.

The term $f_{X|Y}(x|Y = y, \theta_j)$ is the probability of x conditioned on $Y = y$ and $\theta = \theta_j$. The subscript on f indicates that we are using the probability distribution of x conditioned on $Y = y$. For the EM algorithm, we need to distinguish between θ and θ_j because we maximize with respect to θ not θ_j. If we just need the expectation at θ, no maximization step, then we just use θ in $f(.|\theta)$ and the subscript on E.

Before moving on with the derivation, notice that in 8, we fix y, the data. We are not treating that as a random variable. We could certainly treat $E_{\theta_j}[\lambda(\{X, y\}, \theta)]$ as some function $g(y)$ and consider the random variable $g(Y)$. But Louis (1982) will not go that route. y is fixed. Thus we are talking about the observed Fisher Information rather than the expected Fisher Information. The latter would take an expectation over the possible y generated by our model with parameters at the MLE.

Derivation of equation 3.1 in Louis 1982

Now we can derive equation 3.1 in Louis (1982). I am going to combine the info in Louis' section 3.1 and the appendix on the derivation of 3.1. Before proceeding, Louis is using 'denominator' format for his matrix derivations; I normally use denominator format but I will follow his convention here. θ is a column vector of parameters and the likelihood $f(.|\theta)$ is scalar. Under 'denominator format', $f'(.|\theta) = df(.|\theta)/d\theta)$ will be a column vector. $f''(.|\theta) = d^2 f(.|\theta)/d\theta d\theta^\top)$ will be a matrix in Hessian format (the first $d\theta$ goes 1 to n down columns and the second $d\theta$ does 1 to n across rows).

Take the derivative of 6 with respect to θ to define $S(z, \theta)$.

$$S(x, y, \theta) = \lambda'(x, y, \theta) = \frac{d \log\{f_{XY}(x, y|\theta)\}}{d\theta} = \frac{df(x, y|\theta)/d\theta}{f(x, y|\theta)} = \frac{f'(x, y|\theta)}{f(x, y|\theta)} \tag{9}$$

Take the derivative of the far right side of 7 with respect to θ to define $S^*(y, \theta)$. For the last step (far right), I used $f_Y(y|\theta) = \int_X f_{XY}(x, y|\theta)dx$, the definition of the marginal distribution [1], to change the denominator.

$$S^*(y, \theta) = \lambda^{*\prime}(y, \theta) = \frac{d \log \int_X f_{XY}(x, y|\theta)dx}{d\theta} = \frac{\int_X f'_{XY}(x, y|\theta)dx}{\int_X f_{XY}(x, y|\theta)dx} = \frac{\int_X f'_{XY}(x, y|\theta)dx}{f_Y(y|\theta)} \tag{10}$$

Now multiply the integrand in the numerator by $f_{XY}(x, y|\theta)/f_{XY}(x, y|\theta)$. The last step (far right) uses 9.

$$\begin{aligned} \int_X f'_{XY}(x, y|\theta)dx &= \int_X \frac{f'_{XY}(x, y|\theta) f_{XY}(x, y|\theta)}{f_{XY}(x, y|\theta)} dx \\ &= \int_X \frac{f'_{XY}(x, y|\theta)}{f_{XY}(x, y|\theta)} f_{XY}(x, y|\theta)dx \\ &= \int_X S(x, y, \theta) f_{XY}(x, y|\theta)dx \end{aligned} \tag{11}$$

We combine 10 and 11:

$$\begin{aligned} S^*(y, \theta) =& \frac{\int_X f'_{XY}(x, y|\theta)dx}{f_Y(y|\theta)} \\ &= \int_X S(x, y, \theta) \frac{f_{XY}(x, y|\theta)}{f_Y(y|\theta)} dx \\ &= \int_X S(x, y, \theta) f_{X|Y}(x|Y = y, \theta)dx \end{aligned} \tag{12}$$

The second to last step used the fact that $f_Y(y|\theta)$ does not involve x thus we can bring it into the integral. This gives us $f_{XY}(x, y|\theta)/f_Y(y|\theta)$. This is the probability of x conditioned on $Y = y$ (2).

The last step in the derivation of equation 3.1 is to recognize that the far right side of 12 is the conditional expectation in 3.1. Louis does not actually write out the expectation in 3.1 and the notation is rather vague. But the expectation in equation 3.1 is the conditional expectation on the far right side of 12.

$$S^*(y,\theta) = \int_X S(x,y,\theta) f_{X|Y}(x|Y=y,\theta)dx = E_{X|y,\theta}[S(X,y,\theta)] \tag{13}$$

using my notation for a conditional expectation which slightly different than Louis'. At the MLE, $S^*(y,\hat{\theta}) = 0$ since that is how the MLE is defined (it's where the derivative of the data likelihood is zero).

Derivation of equation 3.2 in Louis 1982

The meat of Louis 1982 is equation 3.2. The observed Fisher Information matrix 5 is

$$\mathcal{I}(\theta,y) = B^*(y,\theta) = -S'(x,y,\theta) = -\lambda^{*\prime\prime}(y,\theta) = -\frac{\partial^2 \log f_Y(y|\theta)}{\partial\theta\partial\theta^\top} \tag{14}$$

The first 3 terms on the left are just show that all are notation that refers to the observed Fisher Information. The 4th term is one of the ways we can compute the observed Fisher Information at θ and the far right term shows that derivative explicitly.

We start by taking the second derivative of 6 with respect to θ to define $B(x,y,\theta)$. We use $S'(z,\theta)$ as written in 9.

$$\begin{aligned}\mathcal{I}(\theta,x,y) = B(x,y,\theta) = -\lambda''(x,y,\theta) = -S'(x,y,\theta) = \\ -\frac{d[f'_{XY}(x,y|\theta)/f_{XY}(x,y|\theta)]}{d\theta^\top}\end{aligned} \tag{15}$$

The transpose of $d\theta$ is because we are taking the second derivative $d^2l/d\theta d\theta^\top$ (the Hessian of the log-likelihood); $d\theta d\theta$ wouldn't make sense as that that would be a column vector times a column vector.

To do the derivative on the far right side of 15, we first need to recognize the form of the equation. $f'_{XY}(x,y|\theta)$ is a column vector and $f(x,y|\theta)$ is a scalar, thus the thing we are taking the derivative of has the form $\overrightarrow{h}(\theta)/g(\theta)$; the arrow over h is indicating that it is a (column) vector while $g()$ is a scalar. Using the chain rule for vector derivatives, we have

$$\frac{d(\overrightarrow{h}(\theta)/g(\theta))}{d\theta^\top} = \frac{d\overrightarrow{h}(\theta)}{d\theta^\top}\frac{1}{g(\theta)} - \frac{\overrightarrow{h}(\theta)}{g(\theta)^2}\frac{g(\theta)}{d\theta^\top} \tag{16}$$

Thus we can write the equation for the negative of $B(x,y,\theta)$ as

$$\begin{aligned}-B(x,y,\theta) =& \frac{d(f'_{XY}(x,y|\theta)/f_{XY}(x,y|\theta))}{d\theta^\top} \\ =& \frac{f''_{XY}(x,y|\theta)}{f_{XY}(x,y|\theta)} - \frac{f'_{XY}(x,y|\theta)f'(z|\theta)^\top}{f_{XY}(x,y|\theta)^2} \\ =& \frac{f''_{XY}(x,y|\theta)}{f_{XY}(x,y|\theta)} - S(x,y|\theta)S(x,y|\theta)^\top\end{aligned} \tag{17}$$

Let's return to 14 and take the derivative of $\lambda^{*\prime}(y,\theta)$ with respect to θ using the form shown in equation 10. I have replaced the integral in the denominator by $f_Y(y|\theta)$ and used the same chain rule used for 17.

$$
\begin{aligned}
\lambda^{*\prime\prime}(y,\theta) =& d\left(\int_X f'_{XY}(x,y|\theta)dx \Big/ f_Y(y|\theta)\right)/d\theta^\top \\
=& \frac{\int_X f''_{XY}(x,y|\theta)dx}{f_Y(y|\theta)} - \frac{\int_X f'_{XY}(x,y|\theta)dx}{f_Y(y|\theta)}\left(\frac{\int_X f'_{XY}(x,y|\theta)dx}{f_Y(y|\theta)}\right) \\
=& \frac{\int_X f''_{XY}(x,y|\theta)dx}{f_Y(y|\theta)} - S^*(y|\theta)S^*(y|\theta)^\top
\end{aligned}
\tag{18}
$$

The last substitution uses 10. Thus,

$$
\lambda^{*\prime\prime}(y,\theta) = \frac{\int_X f''_{XY}(x,y|\theta)dx}{f_Y(y|\theta)} - S^*(y|\theta)S^*(y|\theta)^\top \tag{19}
$$

Let's look at the integral of the second derivative of $f_{XY}(x,y|\theta)$ in 19:

$$
\begin{aligned}
\left(\int_X f''_{XY}(x,y|\theta)dx \Big/ f_Y(y|\theta)\right) &= \int_X \frac{f''_{XY}(x,y|\theta)dx}{f_{XY}(x,y|\theta)}\frac{f_{XY}(x,y|\theta)}{f_Y(y|\theta)}dx \\
&= \int_X \frac{f''_{XY}(x,y|\theta)dx}{f_{XY}(x,y|\theta)} f_{X|Y}(x|Y=y,\theta)dx
\end{aligned}
\tag{20}
$$

This is the conditional expectation $E_{X|Y,\theta}[f''_{XY}(x,y|\theta)dx/f_{XY}(x,y|\theta)]$ that we see 5 lines above the references in Louis (1982). Using 17 we can write this in terms of $B(x,y|\theta)$:

$$
\int_X \frac{f''_{XY}(z|\theta)dx}{f_{XY}(x,y|\theta)} = -B(x,y|\theta) + S(x,y|\theta)S(x,y|\theta)^\top \tag{21}
$$

Combining 19, 20, and 21, we can write the equation above the references in Louis:

$$
\lambda^{*\prime\prime}(y,\theta) = E_{X|y,\theta}[-B(X,y|\theta) + S(X,y|\theta)S(X,y|\theta)^\top] - S^*(y|\theta)S^*(y|\theta)^\top \tag{22}
$$

The negative of this is the observed Fisher Information (14) which gives us equation 3.2 in Louis (1982):

$$
\mathcal{I}(\theta,y) = E_{X|y,\theta}[B(X,y|\theta)] - E_{X|y,\theta}[S(X,y|\theta)S(X,y|\theta)^\top] + S^*(y|\theta)S^*(y|\theta)^\top \tag{23}
$$

Derivation of equation 3.3 in Louis 1982

Louis states that "The first term in (3.2) is the conditional expected full data observed information matrix, while the last two produce the expected information for the conditional distribution of X given $X \in R$." His X is my $\{X,Y\}$ and $X \in R$ means $Y = y$ in my context. He writes this in simplified form with X replaced by XY:

$$
I_Y = I_{XY} - I_{X|Y} \tag{24}
$$

$$
\mathcal{I}(\theta,y) = E_{X|y,\theta}[\mathcal{I}(\theta,X,y)] - I_{X|Y} \tag{25}
$$

Let's see how this is the case.

The full data observed information matrix is

$$\mathcal{I}(\theta, x, y) = -\lambda''(x, y|\theta) = B(x, y, \theta) \tag{26}$$

This is simply the definition that Louis gives to $B(x, y, \theta)$. We do not know x so we do not know the full data observed Information matrix. But we have the distribution of x conditioned on our data y.

$$E_{X|y,\theta}[B(X, y|\theta)] \tag{27}$$

is thus the expected full data observed information matrix conditioned on our observed data y. So this is the first part of his statement. The second part of his statement takes a bit more effort to work out. First we substitute $S^*(y|\theta)$ with $E_{X|Y,\theta}[S(X, y|\theta)]$ from 13. This gives us:

$$\begin{aligned} E_{X|y,\theta}[S(X, y|\theta)S(X, y|\theta)^\top] - S^*(y|\theta)S^*(y|\theta)^\top = \\ E_{X|y,\theta}[S(X, y|\theta)S(X, y|\theta)^\top] - E_{X|y,\theta}[S(X, y|\theta)]E_{X|y,\theta}[S(X, y|\theta)^\top] \end{aligned} \tag{28}$$

Using the computational form of the variance, $var(X) = E(XX) - E(X)E(X)$, we can see that 28 is the conditional variance of $S(X, y|\theta)$.

$$var_{X|y,\theta}(S(X, y|\theta)) \tag{29}$$

But the variance of the first derivative of $f'(X|\theta)$ is the expected Fisher Information of X (4). In this case, it is the expected Fisher Information of the hidden state X, where we specify that X has the conditional distribution $f_{X|Y}(X|Y = y, \theta)$. Thus we have the second part of Louis' statement.

Relating Louis 1982 to the update equations in the MARSS EM algorithm

The main result in Louis (1982) (23) can be written

$$\mathcal{I}(\theta, y) = E_{X|y,\theta}[B(X, y|\theta)] - var_{X|y,\theta}[S(X, y|\theta)] \tag{30}$$

The M-step of the EM algorithm involves the first derivative of the log-likelihood with respect to θ, $S(X, y|\theta)$, since it involves setting this derivative to zero:

$$\begin{aligned} Q'(\theta|\theta_j) =& d(E_{X|y,\theta_j}[\log f_{XY}(X, y|\theta)])/d\theta \\ =& E_{X|y,\theta_j}[\log f'_{XY}(X, y|\theta)] \\ =& E_{X|y,\theta_j}[S(X, y|\theta)] \end{aligned} \tag{31}$$

With the MARSS model, $S(X, y|\theta)$ is analytical and we can also compute $B(X, y|\theta)$, the second derivative, analytically.

The difficulty arises with this term: $var_{X|Y,\theta}[S(X, y|\theta)]$. The $S(X, y|\theta)$ is a summation from $t = 1$ to T that involves X_t or $X_t X_{t-1}^\top$ for some parameters. When we do the cross-product, we will end up with terms like $E[X_t X_{t+k}^\top]$ and $E[X_t X_t^\top X_{t+k} X_{t+k}^\top]$. The latter is not a problem; all the random variables in a MARSS models are multivariate normal and the k-th central moments can be expressed in terms of the first and second moments (5), but that will still leave us with terms like $E[X_t X_{t+k}^\top]$, which are the smoothed covariance between X at time t and $t + k$ conditioned on all the data ($t = 1 : T$).

Computing these is not hard. These are the the n-step apart smoothed covariances. Harvey (1989), page 148, discusses how to use the Kalman filter to get the n-step ahead prediction covariances and a similar

approach can be used (presumably) to get the $V(t, t+k)$ smoothed covariances. However this will end up being computationally expensive because we will need all of the $t, t+k$ combinations, i.e., {1,3}, {1,4}, ..., {2,3}, {2,4}, ... etc.. That will be a lot: $T + T - 1 + T - 2 + T - 3 + \ldots$, i.e. $T(T+1)/2$, smoothed covariances.

Lystig and Hughes (2012) and Duan and Fulop (2011) discuss this issue for in a related application of the approach in Louis (1982). They suggest that you do not need to include covariances with a large time separation because the covariance goes to zero. You just need to include enough time-steps.

Conclusion

I think the approach of Louis (1982) is not viable for MARSS models. The derivatives $B(x, y|\theta)$ and $S(x, y|\theta)$ are straight-forward (if tedious) to compute analytically following the approach in Holmes (2010). But the computing all the n-step smoothed covariances is going to be very slow and each computation involves many matrix multiplications. However, one could compute $\mathcal{I}(\theta, y)$ via simulation using 30. It is easy enough to simulate X using the MLEs and then you compute $B(x_b, y|\theta)$ and $S(x_b, y|\theta)$ for each where x_b is the bootstrapped x time series and y is the data. I don't think it makes sense to do that for MARSS models since there are two recursion approaches for computing the observed and expected Fisher Information using $f(y|\theta)$ and the Kalman filter equations (Harvey 1989, pages 140-142; Cavanaugh and Shumway 1996).

Footnotes

1. Given a joint probability distribution of $\{X, Y\}$, the marginal distribution of Y is $\int_X f(X, Y)dx$. Discussions of the estimators for MARSS models often use the property of the marginal distributions of a multivariate normal without actually stating that this property is being used. The step in the derivation will just say, 'Thus' with no indication of what property was just used. Reviewed here: http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html

   If you have a joint likelihood of some random variables, and you want the likelihood of a subset of those random variables, then you compute the marginal distribution. i.e. you integrate over the random variables you want to get rid of:

$$L(\theta|y)] = \int_X L(\theta|X, Y)p(x|Y = y, \theta_j)dx|_{Y=y}. \tag{32}$$

   So we integrate out X from the full likelihood and then set $Y = y$ to get the likelihood.

The marginal likelihood is a little different. The marginal likelihood is used when you want to get rid of some of the parameters, nuisance parameters. The integral you use is different:

$$L(\theta_1|y) = \int_{\theta_2} p(y|\theta_1, \theta_2)p(\theta_2|\theta_1)d\theta_2 \tag{33}$$

This presumes that you have $p(\theta_2|\theta_1)$. The expected likelihood is different yet again:

$$E_{X,Y|Y=y,\theta_j}[L(\theta|X, Y)] = \int_X L(\theta|X, Y)p(x|Y = y, \theta_j)dx. \tag{34}$$

On the surface it looks like the equation for $L(\theta|y)$ but it is different. θ_j is not θ. It is the parameter value at which we are computing the expected value of X. Maximizing the $E_{X,Y|Y=y,\theta_j}[L(\theta|X, Y)]$ will increase the

likelihood but will not take you to the MLE. You have to imbed this maximization in the EM algorithm that walks up the likelihood surface.

2. $P(A|B) = P(A \cup B)/P(B)$
3. I normally think about Y as being partially observed (missing values) so I also take the expectation over $Y(2)$ conditioned on $Y(1)$, where (1) means observed and (2) means missing. In Holmes (2010), this is done in order to derive general EM update equations for the missing values case. But my notation is getting hairy, so for this write-up, I'm treating Y as fully observed; so no $Y(2)$ and I've dropped the integrals (expectations) over $Y(2)$.
4. http://people.missouristate.edu/songfengzheng/Teaching/MTH541/Lecture%20notes/Fisher_info.pdf
5. https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Higher_moments

Papers and online references

- Ng, Krishnan and McLachlan 2004 The EM algorithm. Section 3.5 discusses standard errors approaches https://www.econstor.eu/dspace/bitstream/10419/22198/1/24_tk_gm_skn.pdf http://hdl.handle.net/10419/22198
- Efron and Hinkley 1978 (argues that the observed Fisher Information is better than expected Fisher Information in many/some cases. The same paper argues for the likelihood ratio method for CIs) Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher Information https://www.stat.tamu.edu/~suhasini/teaching613/expected_observed_information78.pdf
- Hamilton 1994 http://web.pdx.edu/~crkl/readings/Hamilton94.pdf
- Hamilton's exposition assumes you know the marginal distribution of a multivariate normal. Scroll down to the bottom. http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html
- Meilijson 1989 Fast improvement to the EM algorithm on its own terms http://www.jstor.org/stable/pdf/2345847.pdf
- Oakes 1999 Direct calculation of the information matrix via the EM algorithm http://www.jstor.org/stable/pdf/2680653.pdf?_=1463187953783
- Ho, Shumway and Ombao 2006 (this has a brief statement that Oakes 1999 derivatives are hard to compute. It doesn't say why. It says nothing of Louis 1982.) Chapter 7, The state-space approach to modeling dynamic processes Models for Intensive Longitudinal Data https://books.google.com/books?hl=en&lr=&id=Semo20xZ_M8C
- Louis 1982 (so elegant. alas, MARSS deals with time series data...) Finding the observed information matrix when using the EM algorithm http://www.jstor.org/stable/pdf/2345828.pdf http://www.markirwin.net/stat220/Refs/louis1982.pdf
- Lystig and Hughes 2012 (helped me better understand why Louis 1982 is hard for MARSS models) Exact computation of the observed information matrix for hidden Markov models http://www.tandfonline.com.offcampus.lib.washington.edu/doi/abs/10.1198/106186002402
- Duan and Fulop 2011 (also helped me better understand why Louis 1982 is hard for MARSS models) A stable estimator for the information matrix under EM for dependent data http://www.rmi.nus.edu.sg/DuanJC/index_files/files/EM_Variance_March%205%202007.pdf http://link.springer.com/article/10.1007/s11222-009-9149-4
- Naranjo 2007 (didn't use) State-space models with exogenous variables and missing data, PhD U of FL http://etd.fcla.edu/UF/UFE0021568/naranjo_a.pdf

- Dempster, Laird, Rubin 1977 (didn't really use but looked up more info on the 'score' function Q) Maximum likelihood for incomplete data via the EM algorithm http://web.mit.edu/6.435/www/Dempster77.pdf
- van Dyk, Meng and Rubin 1995 (this looks promising) Maximum likelihood estimation via the ECM algorithm: computing the asymptotic variance http://wwwf.imperial.ac.uk/~dvandyk/Research/95-sinica-secm.pdf
- Cavanaugh and Shumway 1996 On computing the expected Fisher Information Matrix for state-space model parameters
- Harvey 1989, pages 140-143, Section 3.4.5 Information matrix Forecasting, structural time series models and the Kalman filter